

MÉTODOS ESTATÍSTICOS APLICADOS À CIÊNCIA FLORESTAL

Análise de regressão

ISBN: 978-65-80261-37-6

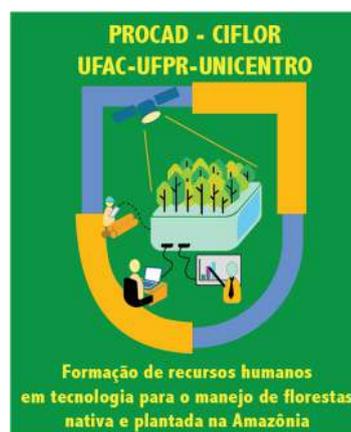
2024

Thiago Augusto da Cunha

Afonso Figueiredo Filho

Métodos Estatísticos Aplicados à Ciência Florestal: Análise de regressão

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Código de Financiamento 001, no âmbito do Programa Nacional de Cooperação Acadêmica na Amazônia – Procad Amazônia edital 2018 coordenado pelo Programa de Pós-graduação em Ciência Florestal da Universidade Federal do Acre:



Rio Branco, Acre

Stricto Sensu Editora

CNPJ: 32.249.055/001-26

Prefixos Editorial: ISBN: 80261 – 86283 / DOI: 10.35170

Editora Geral: Profa. Dra. Naila Fernanda Sbsczk Pereira Meneguetti

Editor Científico: Prof. Dr. Dionatas Ulises de Oliveira Meneguetti

Bibliotecária: Tábata Nunes Tavares Bonin – CRB 11/935

Avaliação: Foi realizada pelos editores chefe e científico da Stricto Sensu Editora

Revisão: Realizada pela Stricto Sensu Editora e autores

Conselho Editorial

Prof^a. Dr^a. Ageane Mota da Silva (Instituto Federal de Educação Ciência e Tecnologia do Acre)

Prof. Dr. Amilton José Freire de Queiroz (Universidade Federal do Acre)

Prof. Dr. Benedito Rodrigues da Silva Neto (Universidade Federal de Goiás – UFG)

Prof. Dr. Edson da Silva (Universidade Federal dos Vales do Jequitinhonha e Mucuri)

Prof^a. Dr^a. Denise Jovê Cesar (Instituto Federal de Educação Ciência e Tecnologia de Santa Catarina)

Prof. Dr. Francisco Carlos da Silva (Centro Universitário São Lucas)

Prof. Dr. Humberto Hissashi Takeda (Universidade Federal de Rondônia)

Prof. Msc. Herley da Luz Brasil (Juiz Federal – Acre)

Prof. Dr. Jader de Oliveira (Universidade Estadual Paulista Júlio de Mesquita Filho - UNESP - Araraquara)

Prof. Dr. Jesus Rodrigues Lemos (Universidade Federal do Piauí – UFPI)

Prof. Dr. Leandro José Ramos (Universidade Federal do Acre – UFAC)

Prof. Dr. Luís Eduardo Maggi (Universidade Federal do Acre – UFAC)

Prof. Msc. Marco Aurélio de Jesus (Instituto Federal de Educação Ciência e Tecnologia de Rondônia)

Prof^a. Dr^a. Mariluce Paes de Souza (Universidade Federal de Rondônia)

Prof. Dr. Paulo Sérgio Bernarde (Universidade Federal do Acre)

Prof. Dr. Romeu Paulo Martins Silva (Universidade Federal de Goiás)

Prof. Dr. Renato Abreu Lima (Universidade Federal do Amazonas)

Prof. Dr. Rodrigo de Jesus Silva (Universidade Federal Rural da Amazônia)

Ficha Catalográfica

Dados Internacionais de Catalogação na Publicação (CIP)

M593

Métodos estatísticos aplicados à ciência florestal : análise de regressão / Thiago Augusto da Cunha, Afonso Figueiredo Filho (Org.). – Rio Branco : Stricto Sensu, 2024.
335 p. : il

ISBN: 978-65-80261-37-6

DOI: 10.35170/ss.ed.9786580261376

1. Engenharia. 2. Estatística. 3. Ciência florestal. I. Título. II. Cunha, Thiago Augusto da. III. Figueiredo Filho, Afonso.

CDD 22. ed. 628.021

Bibliotecária Responsável: Tábata Nunes Tavares Bonin / CRB 11-935

O conteúdo dos capítulos do presente livro, correções e confiabilidade são de responsabilidade exclusiva dos autores.

É permitido o download deste livro e o compartilhamento do mesmo, desde que sejam atribuídos créditos aos autores e a editora, não sendo permitido à alteração em nenhuma forma ou utilizá-la para fins comerciais.

www.sseditora.com.br

APRESENTAÇÃO

Este manual tem por objetivo disponibilizar aos estudantes de graduação e pós-graduação das ciências florestais, um instrumental estatístico básico para a realização de trabalhos científicos de monografia, dissertação ou tese de doutorado. Foi elaborado inicialmente para apoiar as análises estatísticas das pesquisas científicas realizadas no Programa de Pós-Graduação em Ciência Florestal da Universidade Federal do Acre.

Com o avanço na edição do mesmo, buscou-se ampliar e aprimorar a abordagem do tema a todos aqueles que se dedicam à pesquisa na área dos recursos florestais.

Trata-se de uma orientação prática e aplicada de como realizar a análise de dados de campo até a apresentação dos resultados para fins de elaboração de trabalhos científicos.

O objetivo desse manual não é apenas querer que os estudantes e pesquisadores realizem análises estatísticas, e sim ensinar o pensamento estatístico a fim de desenvolver conceitos próprios para tomada de decisão na análise de dados.

Desta forma, o foco do manual se concentra em determinar qual análise estatística é a adequada para extrair o máximo de informações úteis dos dados no campo florestal de acordo ao objetivo da pesquisa.

Com respeito ao uso de software estatístico, o manual inclui orientações de como usar uma das ferramentas analítica da SAS Institute, o SAS Studio® disponível de forma totalmente gratuita na plataforma SAS On Demand for Academics.

Quanto ao método de ensino, o manual propõe exercícios considerando que o aprendizado deve ser ativo de acordo ao método desenvolvido por Willian Glasser (Psiquiatra norte americano, 1925-2013). O método propõe que para alcançar o máximo do aprendizado o estudante deve, inclusive, ensinar os demais. O método é descrito na Figura 1.

Portanto, para enriquecer ainda mais o aprendizado e melhor assimilar a teoria de estatística, todos os capítulos desse manual possuem exercícios práticos com uso aplicado em software estatístico.

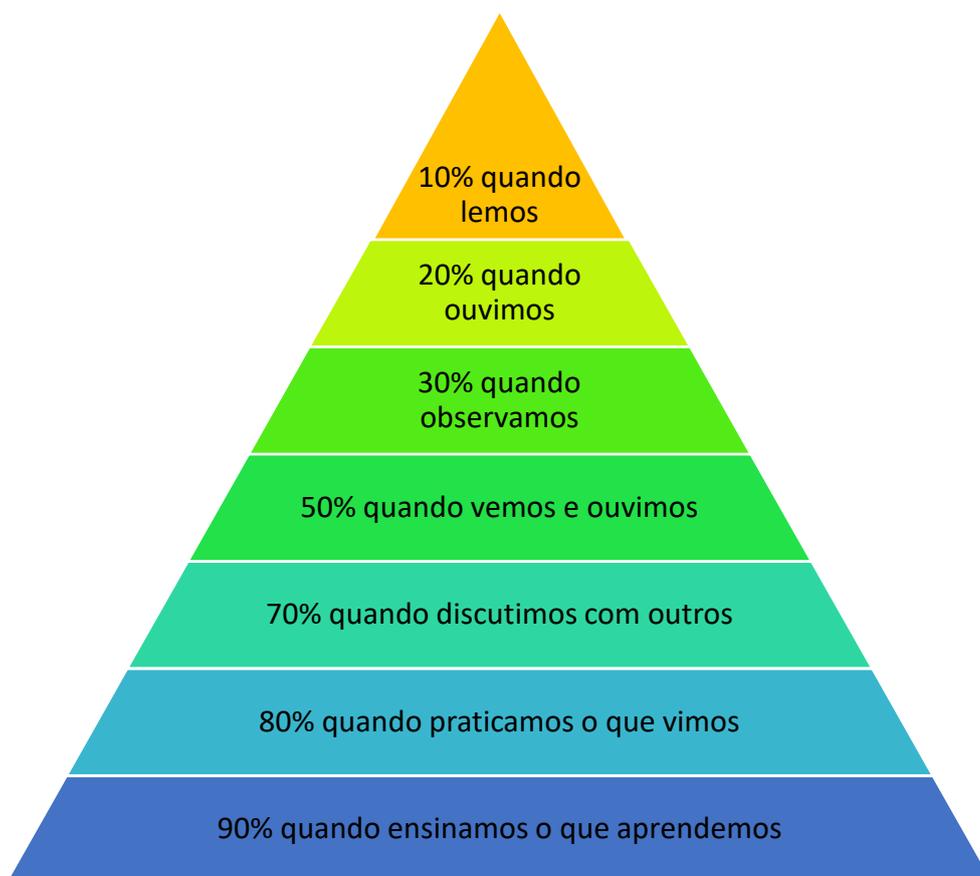


Figura 1. Níveis de aprendizado proposto por Willian Glasser de acordo a atividade realizada durante o ensino.

Padronização da simbologia

Todas as variáveis de mensuração e manejo florestal utilizadas no manual foram padronizadas considerando os símbolos propostos por Silva et al. (2022).

Uso de software no ensino de estatística

O uso de software para o ensino de estatística é bastante discutido entre os professores da área. Vários são os motivos para grande parte dos docentes questionarem a eficiência do aprendizado de estatística combinado com o uso de software durante as aulas.

De fato, o uso de software para o ensino de estatística deve ser avaliado do ponto de vista do objetivo do curso. Caso o objetivo é que os discentes realizem algumas análises

estatísticas a partir de dados coletados em campo utilizando ferramentas computacionais em nível básico, um software básico como Excel produz gráficos e realiza análises necessárias. Caso for realizar análise mais complexas o uso de softwares com linguagem de programação fornece análises e gráficos mais elaborados. A exemplo temos os softwares R, Python, o SAS, entre outros.

Softwares programáveis são mais adequados pois requerem lidar com a lógica dos métodos usados para escrever a programação. Essa forma de realizar análise estatística exige do discente um conhecimento prévio do que se deseja analisar e qual análise está sendo realizada, pois precisam escrever certa quantidade de códigos em vez de apenas inserir certa quantidade de dados, selecionar menu de opções e pressionar alguns botões.

Essa forma de realizar análise estatística não apenas os apresenta às ferramentas que os pesquisadores usam, mas o faz de uma forma que os força a realmente lidar com os métodos estatísticos que usam.

O fato é que devemos ter em mente que o objetivo do ensino de estatística não pode ser apenas o de saber usar um software para realizar análises e construir gráficos (Técnicos em estatística). É preciso reconhecer que o ensino na Universidade, mais especificamente em curso de pós-graduação, deve formar cientistas e, portanto, deve priorizar o pensamento estatístico de forma que o discente entenda os dados e seja criativo o suficiente para encontrar boas maneiras para deixar os dados contarem sua história de maneira que possamos entendê-los.

Adicionalmente, deve-se considerar que ensinar estatística mostrando como executar cálculos por meio de softwares consiste em apenas ensinar como calcular. O importante é que o discente entenda o significado dos resultados e sua aplicação para discuti-los.

Neste sentido, este manual visa o ensino de estatística aplicada na área florestal por meio de métodos e exemplos práticos comumente utilizados na área.

Para a parte prática, serão utilizados diversos exemplos com dados secundários coletados em campo de forma a aplicar o processamento dos dados de pesquisa utilizando software programável.

O manual está organizado com as seguintes fases de acordo à Figura 2 e resume a sequência a ser seguida para realizar análise de dados considerando que estes devem ser coletados em campo ou até mesmo em laboratório. Caso o pesquisador já tenha disponível dados coletados, é possível iniciar diretamente na fase de exploração dos dados.



Figura 2. Sequência simples para realizar uma análise estatística de dados de campo/laboratório utilizando o software estatístico SAS Studio.

SUMÁRIO

1. Software SAS Studio.....	13
1.1. Versão acadêmica do SAS System.....	14
1.1.1. Acesso à plataforma ODA.....	15
1.1.2. O ambiente SAS Studio (Versão ODA)	21
1.2. Conjunto de dados SAS (SAS Data Set)	26
1.2.1. Atributos de variáveis SAS.....	27
1.3. Upload de dados no SAS Studio.....	29
1.3.1. Inserindo os dados diretamente no editor do SAS.....	29
1.3.2. Criando conjunto de dados SAS a partir de arquivo externo.....	31
1.3.2.1. Usando o procedimento PROC IMPORT para importar dados para o SAS.....	31
1.3.2.2. Usando o comando LIBNAME para ler arquivos MS Excel.....	32
1.4 Programação SAS.....	33
1.4.1. Erros comuns em programação SAS.....	35
2. Processo para obtenção de dados para pesquisa.....	37
2.1. População e amostra.....	37
2.1.1. Amostragem ou censo?	39
2.2. Tipos de variáveis e escala de medição.....	42
2.2.1. Variável dependente e variável independente.....	42
2.2.2. Organização dos dados.....	43
2.2.2.1. Dados estruturados.....	43
2.2.2.2. Dados não-estruturados.....	45
2.2.3. Tipos de variáveis e escala de medição.....	45

2.3. Desenhos de estudo estatísticos.....	49
2.3.1. Estudo experimental (Manipulativo).....	50
2.3.2. Estudo observacional.....	51
2.4. Obtenção de amostras.....	52
2.4.1. Amostragem não-probabilística.....	53
2.4.2. Amostragem probabilística.....	56
2.4.3. Métodos de amostragem.....	59
2.4.4. Seleção de amostras aleatórias no SAS Studio.....	60
2.4.4.1. Preparação dos dados da população.....	62
2.4.5. Amostragem aleatória simples.....	63
2.4.5.1. Amostragem simples sem reposição.....	63
2.4.5.2. Amostragem simples com reposição.....	72
2.4.6. Amostragem aleatória estratificada.....	72
2.4.7. Amostragem aleatória estratificada com alocação otimizada.....	79
2.4.7.1. Método de alocação Proporcional (Proportional).....	80
2.4.7.2. Método de alocação de Neyman (Neyman).....	82
2.4.7.3. Método de alocação Ótima (Optimal).....	82
2.4.8. Teorema Central do Limite e o tamanho da amostra.....	87
2.4.9. Cálculo do tamanho da amostra.....	96
2.4.9.1. Tamanho da amostra: Uma média.....	96
2.4.10. Poder da Análise (Power Analysis)	99
2.4.10.1. Nível de significância.....	100
2.4.10.2. Poder do teste e Tamanho da amostra.....	100
2.4.10.3. Tamanho efetivo.....	105

2.4.10.4. Aplicação da Análise de Poder no SAS Studio.....	107
2.4.10.5. Utilizando o Task and Utilities do SAS Studio.....	113
3. Análise de regressão.....	117
3.1. Regressão Linear.....	117
3.1.1 Regressão Linear Simples.....	119
3.1.2. Regressão Linear Múltipla	120
3.2. Avaliação preliminar para análise de regressão.....	123
3.3. Ajuste de modelos de regressão linear.....	136
3.3.1. Estimativa dos coeficientes de regressão linear.....	137
3.3.2. Procedimentos SAS para ajuste de regressão linear.....	141
3.3.2.1. Ajuste de modelo de regressão com variáveis preditoras contínuas no PROC IML.....	142
3.3.2.2. Ajuste de modelo de regressão com variáveis preditoras contínua no PROC REG.....	146
3.3.3. Ajuste de modelo incluindo variável qualitativa no PROC REG.....	151
3.3.4. Ajuste de modelo incluindo variável qualitativa no PROC GLM.....	156
3.3.5. Ajuste de modelo incluindo variável qualitativa no PROC GLMSELECT.....	159
3.3.6. Comparação de ajuste entre os procedimentos REG, GLM E GLMSELECT.....	161
3.3.6.1. O que acontece se aparece a letra “B” na tabela de parâmetros estimados?.....	162
3.3.7. Outros procedimentos SAS para ajuste de modelos geral de regressão linear.....	165
3.4. Testes de hipóteses e ANOVA em análise de regressão.....	166
3.5. Avaliação de modelos de regressão após o ajuste.....	170
3.5.1. Critérios de bondade de ajuste.....	170

3.5.1.1. Erro padrão da estimativa.....	170
3.5.1.2. Coeficiente de variância.....	171
3.5.1.3. Coeficiente de determinação.....	172
3.5.1.4. Critérios de informação para comparação entre modelos.....	173
3.5.2. Análise de resíduos e observações influentes.....	174
3.5.2.1. Avaliação das condicionantes da regressão.....	176
3.5.2.1.1. Avaliação da condicionante de “Normalidade” para os resíduos.....	178
3.5.2.1.2. Avaliação das condicionantes de independência e homoscedasticidade para os resíduos.....	184
3.5.2.2. Avaliação de observações influentes.....	189
3.5.2.3. Como remediar não atendimento as condicionantes de regressão?.....	200
3.5.3. Colinearidade?.....	205
3.5.3.1. Efeitos da multicolinearidade.....	207
3.5.3.2. Diagnóstico da multicolinearidade.....	207
3.5.3.3. Alternativas para reparar o efeito da multicolinearidade.....	210
3.6. Seleção de variáveis para construção de modelos de regressão.....	211
3.6.1. Aplicação dos métodos de controle de seleção de variáveis.....	217
3.6.1.1. Seleção de variáveis considerando nível de significância.....	217
3.6.1.2. Seleção de variáveis considerando critérios de informação.....	229
3.6.2. Validação do modelo de regressão.....	236
3.6.2.1. Avaliação do desempenho de um modelo.....	237
3.6.2.2. Particionamento de dados.....	241

3.6.2.3. Validação um modelo de regressão linear.....	243
3.7. Regressão não-linear.....	247
3.7.1. Modelos de curva de tamanho-idade.....	250
3.7.2. Características e formas das curvas de tamanho-idade.....	253
3.7.3. Características e variações dos modelos de crescimento.....	255
3.7.4. Ajuste de regressão não-linear no SAS System.....	259
3.7.4.1. Aplicação com o modelo de Mitscherlich.....	262
3.7.4.2. Aplicação com o modelo de Chapman-Richards para dados de crescimento.....	269
3.7.4.2.1. Inclusão da covariável sítio no modelo de Chapman-Richards.....	274
3.7.4.3. Modelo para a descrição da forma do tronco de árvores.....	284
3.7.4.4. Avaliação de modelos não-lineares.....	286
3.8. Regressão Logística.....	290
3.8.1. Ajuste da regressão logística no SAS System.....	297
3.8.1.1. Aplicação para variável dependente binária.....	299
3.8.2. Medidas de ajuste da regressão logística.....	310
3.8.3. Modelagem em regressão logística.....	314
3.8.3.1. Modelagem com o PROC LOGISTIC.....	314
3.8.3.2. Modelagem com o PROC HPLOGISTIC.....	322
REFERÊNCIAS.....	325
ORGANIZADORES.....	329
ÍNDICE REMISSIVO	330

1. Software SAS Studio

O SAS (*Statistical Analysis System*) é um conjunto de softwares estatísticos desenvolvido inicialmente para análise de dados agrícolas criado por Jim Goodnight e seus colegas de classe da North Carolina State University na década de 1970. Em seguida, no ano de 1976 a empresa SAS foi criada com um total de 120 clientes (SAS Institute).

Atualmente a empresa é responsável por diversas ferramentas analíticas desenvolvidas especificamente para diferentes tipos de clientes como bancos, empresas farmacêuticas, universidades entre outros.

O Sistema SAS consiste de vários produtos que permitem, entre outros a realização de análise estatística avançada incluindo soluções analíticas para demanda de *business intelligence* e para processamento de *big data*.

Os módulos do sistema SAS foram desenvolvidos considerando a sequência de demanda para análise de dados conforme a seguir:

- Acessar dados;
- Explorar dados;
- Preparar dados;
- Analisar e apresentar dados e
- Exportar resultados.

Cada módulo consiste em um pacote de softwares com diferentes capacidades de análise. O módulo SAS/STAT possui várias ferramentas e procedimentos de análise estatística. Outro módulo bastante utilizado é o SAS/GRAPH com grande capacidade de elaboração de gráficos diversos. O Quadro 1 lista alguns dos diferentes módulos disponíveis no sistema SAS. A quantidade de módulos depende da licença de uso.

Quadro 1. Alguns módulos SAS e suas utilidades para o cliente.

Módulo SAS	Função
SAS/STAT	Um dos módulos mais utilizados, pois, possui softwares específicos para realizar análise estatística.
SAS/GRAPH	Utilizado para construção de gráficos.
SAS/ETS	Para análise de séries temporais e econometria.
SAS IML	Específico para análises utilizando manipulação de matrizes
SAS/QC	Para trabalhar com ferramentas de controle de qualidade
SAS/GIS	Para elaboração de mapas e outras demandas no contexto de sistema de informações geográficas.
SAS/ACCESS	Módulo para leitura de arquivos externos de outros softwares como MS Excel, MS Access e outros.

A SAS Institute oferece diversos cursos gratuitos para a comunidade acadêmica desde o aprendizado em programação básica e avançada, análise estatística, cientista de dados, aprendizado de máquinas e outros. Aos interessados em acessar os cursos interativos e gratuitos, devem realizar um cadastro por meio do sítio https://www.sas.com/pt_br/training/offers/free-training.html.

1.1. Versão acadêmica do SAS System

Até o ano de 2013 para acessar o SAS era necessário comprar uma licença individual ou para vários usuários de uma empresa, por exemplo.

Entretanto, no ano de 2014 a SAS lançou a **versão acadêmica totalmente gratuita** chamada de SAS University Edition criada para ser utilizada por estudantes, faculdades e interessados em realizar qualquer análise estatística no SAS. O acesso ao software poderia ser realizado por qualquer pessoa individual por meio da instalação de uma máquina virtual para abrir o software.

Recentemente (julho 2021) o SAS University Edition foi descontinuado dando lugar a uma nova plataforma de serviço que possibilita os usuários de acessar algumas das ferramentas analíticas SAS, por meio de processamento em nuvem (cloud-based software), denominado de SAS On Demand for Academics (ODA) que disponibiliza o software SAS Studio.

Trata-se de um serviço disponibilizado de forma gratuita para usuário individual, universidades, governo ou outra modalidade desde que não seja comercial.

Para acessar a plataforma ODA é necessário um navegador de internet com acesso à rede e um simples cadastro de perfil na SAS Institute. Portanto, não é necessário realizar nenhum download para acessar o software SAS Studio.

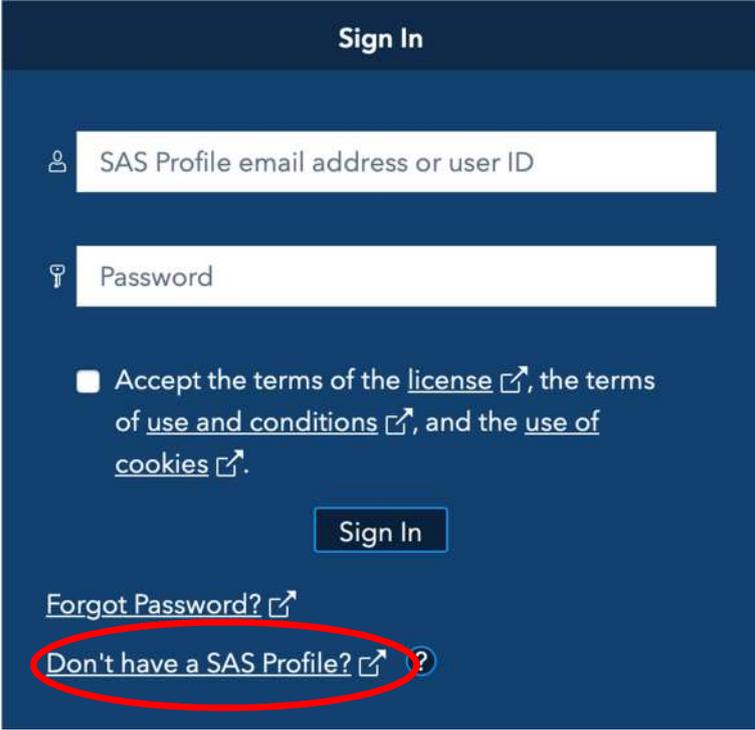
É importante notar que a versão gratuita possui as mesmas funcionalidades da versão paga. Ademais, possui amplo suporte técnico gratuito, comunidades de discussão para realização de processamento de dados bem como milhares de códigos disponíveis na internet de forma gratuita.

1.1.1. Acesso à plataforma ODA

Para os usuários que ainda não possuem um perfil SAS para acesso à plataforma ODA, basta seguir os seguintes passos:

Passo 1: Criar seu perfil para acessar cursos, softwares gratuitos e outras oportunidades na SAS:

Para criar seu perfil SAS basta acessar o seguinte endereço eletrônico: <https://welcome.oda.sas.com>. Em seguida irá aparecer uma janela solicitando as credenciais SAS. Clicar em “Don't have a SAS Profile?”.



Sign In

SAS Profile email address or user ID

Password

Accept the terms of the [license](#), the terms of [use and conditions](#), and the [use of cookies](#).

Sign In

[Forgot Password?](#)

[Don't have a SAS Profile?](#)

Preencher as informações solicitadas e, em seguida, aceitar os termos e condições e clicar em “Create profile”.

SAS Profile

Step 1 of 2: Tell us about yourself.

Preferred Language

First Name *

Last Name *

Email *

Country/Region *

Affiliation With SAS *

Organization/University *

*Required

Yes, I would like to receive occasional emails from SAS Institute Inc. and its affiliates about SAS products and services. I understand that I can withdraw my consent at any time by clicking the opt-out link in the emails.

I agree to the [terms of use and conditions](#). *

All personal information will be handled in accordance with the [SAS Privacy Statement](#).

Create profile

After clicking "Create profile," you will receive a verification email with instructions for setting your password and activating your profile.

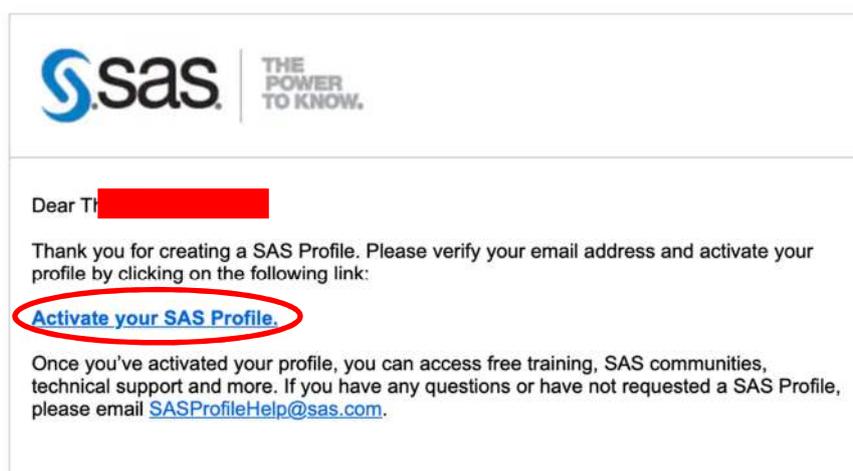
[Privacy Statement](#) | [Terms of Use and Legal Information](#)

Copyright © SAS Institute Inc. All rights reserved.

Após preencher as informações, aparecerá uma mensagem informando que um e-mail foi enviado no endereço informado.



Verifique sua caixa de entrada no e-mail e siga as instruções para ativar seu perfil SAS clicando em “Activate your profile”.



SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration. © SAS Institute Inc. All rights reserved.

No navegador escolha uma senha com os requisitos mínimos de 8 caracteres, 1 letra maiúscula, 1 símbolo e 1 número. Anote essa senha para utilizar toda vez que for abrir a plataforma. Clique em Set password.

SAS

SAS Profile

Step 2 of 2: Please set your new password.

Password (show)

bitkoq-Tojky6-havken Strong Password ✓

Do not use #, \, --, first name, last name, or part of your email address.

Your password cannot be any of your previous 4 passwords.
Your password must contain at least:

- ✓ 8 characters
- ✓ 1 lowercase letter
- ✓ 1 uppercase letter
- ✓ 1 symbol
- ✓ 1 number

Confirm password

bitkoq-Tojky6-havken Strong Password ✓

Set password

[Privacy Statement](#) | [Terms of Use and Legal Information](#)

Copyright © SAS Institute Inc. All rights reserved.

Após clicar em Set password uma nova janela irá aparecer informando que seu perfil foi ativado e receberá um e-mail com a confirmação.

SAS

SAS Profile

Your profile is now active! You can now sign in using your new SAS Profile credentials.

Continue

[Privacy Statement](#) | [Terms of Use and Legal Information](#)

Copyright © SAS Institute Inc. All rights reserved.

[Cookie Preferences](#)

Passo 2: Registro na plataforma SAS On Demand for Academics (ODA).

Agora que você tem um perfil SAS é possível se registrar-se na plataforma ODA para acessar o software SAS Studio. Retorne ao endereço eletrônico: <https://welcome.oda.sas.com> e preencha as informações solicitadas de e-mail e senha registrados no passo 1. Aceite os termos de licença e clique em “Sign In”.

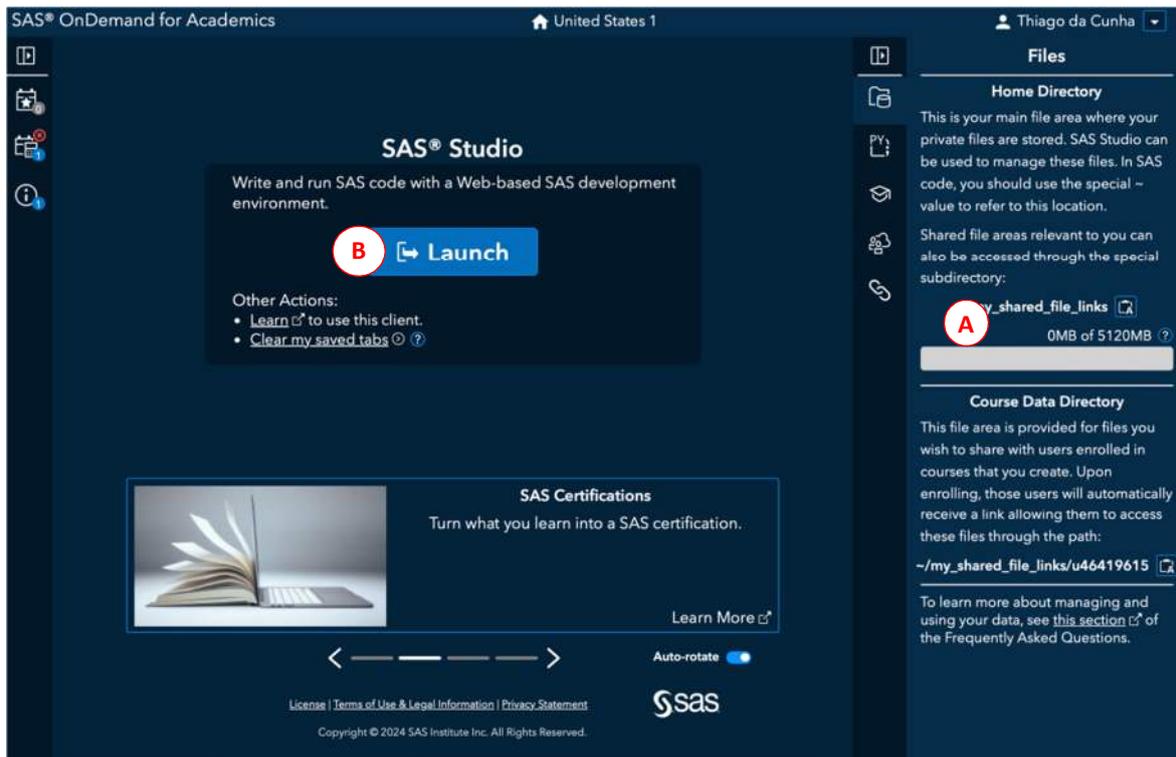


Escolha uma das regiões de servidores SAS disponíveis (Europa, Ásia ou Estados Unidos). Uma dica é escolher a mais próxima da sua região. A escolha de uma das regiões é necessária para poder realizar o processamento dos dados solicitados no SAS Studio.

Logo, você receberá um e-mail de confirmação contendo seu número de usuário da plataforma ODA. Você pode utilizar o número de usuário ou o e-mail cadastrado para acesso à plataforma.

Passo 3: Fazer o login na plataforma ODA para acessar o Software SAS Studio.

Abriu novamente o site a partir do endereço eletrônico: <https://welcome.oda.sas.com> e clicar em Sign In. Logo, preencher as informações de usuário e senha. O Dashboard da plataforma ODA irá aparecer conforme espelho da tela a seguir:



A Armazenamento disponível

Mostra a quantidade de espaço disponível para armazenamento a ser utilizado na plataforma ODA. Cada usuário tem disponível 5GB onde é possível armazenar bases de dados de diferentes aplicativos a serem importados bem como arquivos SAS.

Para usuários cadastrados como professores com turmas para ensino a SAS disponibiliza mais 3GB de espaço.

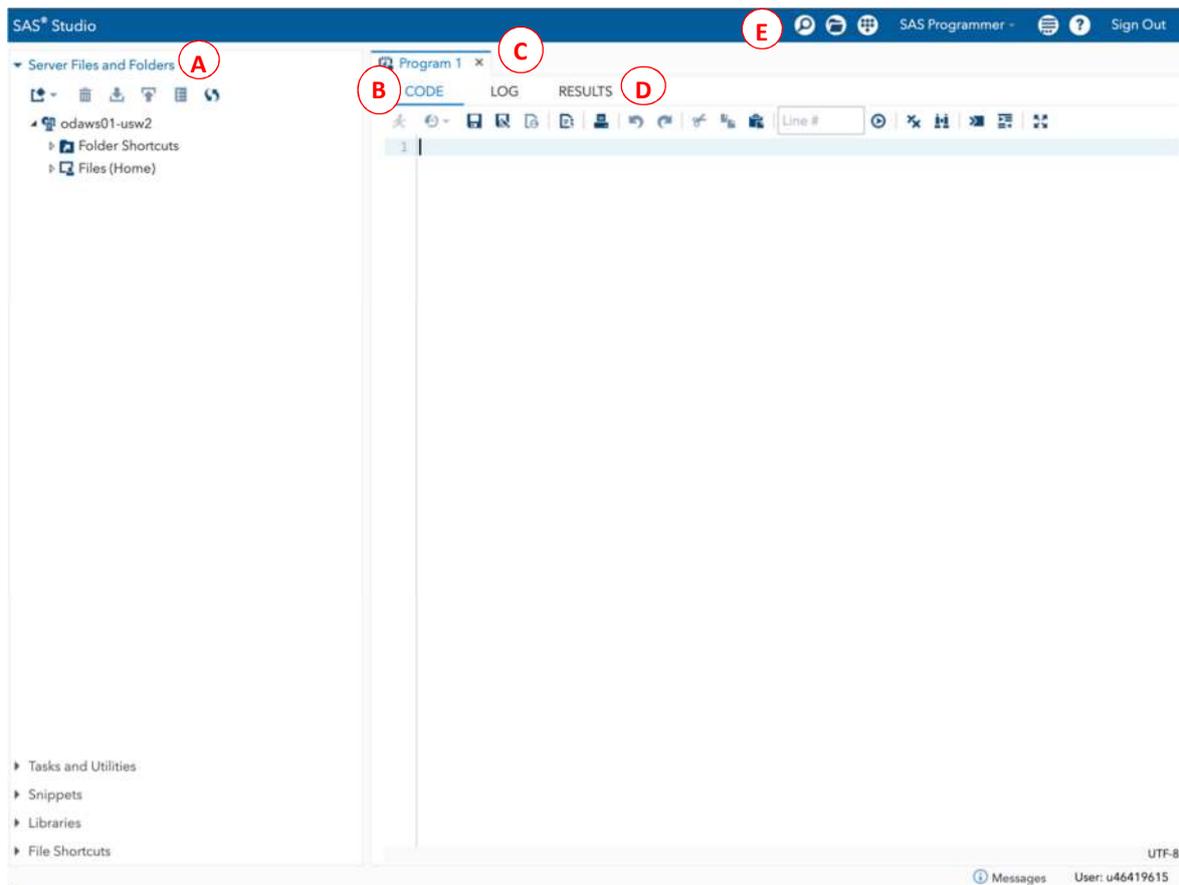
B Aplicações

A aba de aplicativos mostra os softwares disponíveis para o usuário. Neste caso, apenas o software SAS Studio está disponível. É possível incluir mais aplicativos como o SAS Enterprise Miner desde que o usuário seja convidado a partir de algum curso SAS, por exemplo.

Ao clicar em “Launch” o software SAS Studio abre em uma nova aba do navegador de internet.

1.1.2. O ambiente SAS Studio (Versão ODA)

O layout do SAS Studio é constituído de duas partes principais: i) painel de navegação à esquerda da tela e ii) a área de trabalho à direita da tela conforme imagem a seguir da tela do SAS Studio.



A Painel de navegação

O painel de navegação está dividido em cinco itens. O primeiro é o “Server Files and Folders” utilizado para realizar upload de dados no SAS. É possível criar novas pastas e arquivar uma quantidade de dados desde que não exceda 5 GB de espaço disponibilizado na nuvem.

Outro item importante é o “Task and Utilities”. Essa opção é utilizada para ensinar a programação SAS para iniciantes em programação. Ao expandir será mostrado várias opções para realizar análises na interface de “point-and-click” do SAS Studio:



Essas opções predefinidas no SAS Studio gera as sintaxes do SAS de forma automática para uma determinada análise selecionada. É possível também realizar a importação de dados para dentro do SAS utilizando a opção “Import Data”.

O item “Libraries” contém várias pastas de arquivos por padrão. Quando expandida, apresenta os seguintes ícones:



A pasta SASHELP possui várias tabelas de dados SAS a serem utilizadas como exemplos durante o ensino. A pasta WORK armazena de forma temporária as tabelas SAS criadas durante um processamento de dados. Ao expandir a livreria SASHELP aparecerá vários arquivos contendo tabelas SAS prontas para uso. Como exemplo, a tabela SAS BMIMEN foi expandida mostrando que a mesma contém duas variáveis Age e BMI (Body Mass Index) como mostrado a seguir:



Para abrir a tabela e verificar as observações de cada variável basta clicar duas vezes sob o nome do arquivo.

B CODE

A aba CODE é destinada à criação e edição de programas (Sintaxe) SAS. Um programa é uma série de comandos ou declarações que informa ao SAS qual ação será realizada e como essa ação será realizada.

Nesta aba há uma variedade de ícones sendo que um deles executa a programação SAS escrita no CODE (✖). Outro ícone muito útil formata o programa SAS em uma aparência organizada como espaçamento entre os códigos (≡).

C LOG

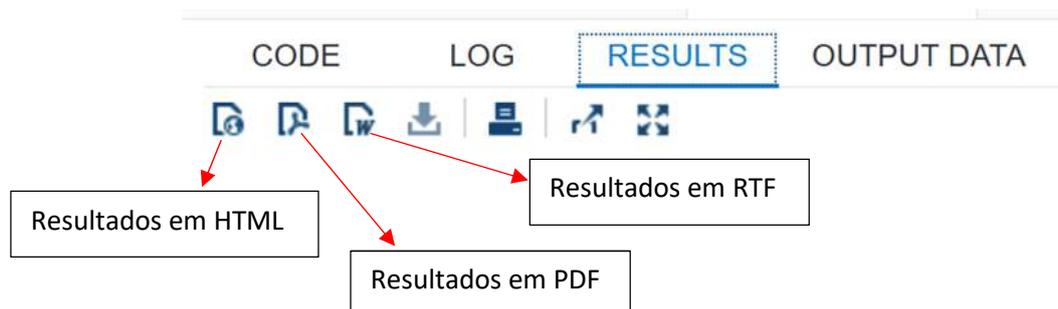
É uma das mais importantes ferramentas se algum problema ocorrer durante um processamento de dados. Esta aba mostra uma mensagem (crítica) a cada submissão de processamento no SAS utilizando cores para comunicar o andamento do processamento. A LOG contém três tipos de mensagens: Notas, Cuidado e Erro descritos no Quadro 2.

Quadro 2. Significado das cores nas mensagens da aba LOG do SAS Studio.

Cor da mensagem	Significado
Azul (Notas)	<p>Mensagens com informações sobre o andamento e status do processamento solicitado. Contém informações como o número de observações em um conjunto de dados, tempo de processamento e outras. As notas são escritas na cor azul conforme exemplo a seguir:</p> <pre>NOTE: There were 159 observations read from the data set SASHELP.FISH. NOTE: PROCEDURE PRINT used (Total process time): real time 0.21 seconds user cpu time 0.21 seconds system cpu time 0.00 seconds memory 4849.43k OS Memory 30372.00k Timestamp 06/10/2024 05:41:40 PM</pre>
Verde (Cuidado)	<p>Mensagem de atenção necessária à sintaxe submetida com possíveis problemas. Neste caso, a mensagem mostra o código/dado introduzido com erro conforme exemplo a seguir:</p> <pre>78 proc print dada=teste; <u>1</u> WARNING 1-322: Assuming the symbol DATA was misspelled as dada. 79 run;</pre> <p>Observe que o SAS interpreta que o código deveria ser DATA e não DADA, mas mesmo assim o SAS realiza o processamento e mostra os resultados na aba Results.</p>
Vermelho (Erro)	<p>Mensagem de erro no processamento aparecem na cor vermelha. Neste caso, o SAS não realiza o processamento dos dados sendo necessário corrigir o erro na sintaxe ou nos dados. O texto inicia com a palavra "ERROR:" seguido da mensagem indicado a localização do erro com um sinal de traço baixo (Underlined) conforme exemplo a seguir:</p> <pre>77 proc print data teste; <u>73</u> ERROR 73-322: Expecting an =. 78 run;</pre> <p>Neste caso, o SAS indica a localização do erro na sintaxe do PROC PRINT que é a falta de um símbolo de igualdade logo após a declaração data.</p>

D RESULTS

A aba “RESULTS” apresenta os resultados de qualquer análise solicitada no SAS Studio por meio de um programa SAS escrito na aba CODE. Essa aba possui alguns ícones úteis que possibilitam o download dos resultados em três diferentes formatos: HTML, PDF e RTF. Portanto, caso desejar inserir um gráfico específico criado no SAS para dentro do texto Word de um trabalho acadêmico, basta clicar no ícone RTF que o arquivo será salvo em formato Rich Text Format que pode ser aberto no Word:



E BARRA DE FERRAMENTAS SAS

A barra de ferramentas SAS possui ícones que possibilitam acessar outros componentes do software SAS Studio como descritos no Quadro 3.

Quadro 3. Ícones disponíveis na barra de ferramentas do software SAS Studio.

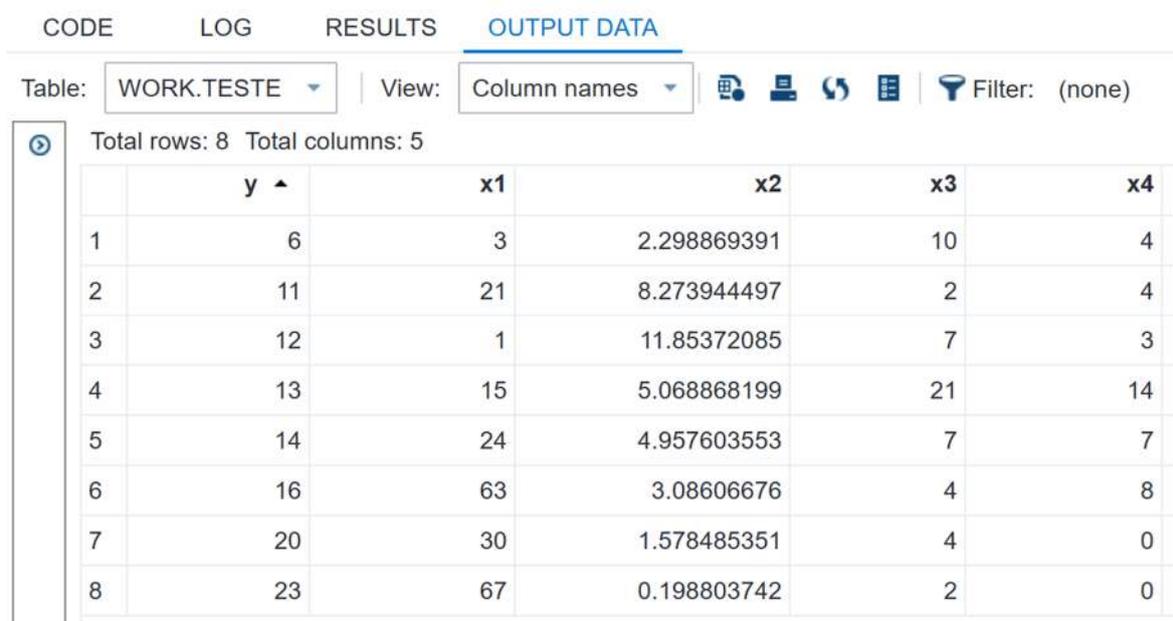
Ícone	Ação
	Ferramenta para buscar qualquer aspecto no Painel de Navegação.
	Para abrir um arquivo armazenado na sua conta da SAS On Demand for Academics.
	Caso desejar abrir uma nova aba de código, importar dados ou mais opções basta clicar nessa ferramenta.
	Muda a perspectiva de SAS Programmer para Visual Programmer.
	Ferramenta que contém opções de configuração do SAS Studio.
	Acessa os manuais online do SAS Studio bem como outras opções de autoajuda.

1.2. Conjunto de dados SAS (SAS Data Set)

Para que o SAS realize o processamento dos dados é necessário que esses dados sejam adequados a um formato padrão de conjuntos de dados estruturados em linhas e colunas (Tabela) de forma que o software entenda.

Portanto, uma tabela SAS é um arquivo estruturado similar a uma tabela Excel contendo linhas e colunas. Cada coluna é denominada de “Variável” e cada linha é denominada de “Observação”.

A Figura 3 mostra um print da tela do SAS Studio mostrando uma tabela SAS, com o nome “Teste” armazenada na livreria Work, que possui cinco variáveis com oito observações para cada variável.



The screenshot shows the SAS Studio interface with the 'OUTPUT DATA' tab selected. The table name is 'WORK.TESTE' and the view is 'Column names'. The table contains 8 rows and 5 columns. The columns are labeled 'y', 'x1', 'x2', 'x3', and 'x4'. The data is as follows:

	y	x1	x2	x3	x4
1	6	3	2.298869391	10	4
2	11	21	8.273944497	2	4
3	12	1	11.85372085	7	3
4	13	15	5.068868199	21	14
5	14	24	4.957603553	7	7
6	16	63	3.08606676	4	8
7	20	30	1.578485351	4	0
8	23	67	0.198803742	2	0

Figura 3. Data Set SAS mostrando as variáveis e observações.

Uma tabela SAS é um arquivo que contém duas partes: uma descritiva e uma parte de dados. A parte descritiva de um conjunto de dados SAS reserva as informações importantes para as estatísticas: i) Nome do conjunto de dados, ii) Data e hora em que o conjunto de dados foi criado, iii) número de observações e o número de variáveis. A parte de dados representa a coleção de valores dos dados organizados em uma tabela retangular contendo as variáveis e as observações.

Uma Tabela SAS é originada após um processamento de dados imputados no SAS que são armazenadas na Livreria do SAS localizada no painel de navegação do SAS Studio

e tem uma extensão “.sas7bdata”. Um conjunto de dados SAS pode armazenar milhares de variáveis.

1.2.1. Atributos de variáveis SAS

Além da escala em que uma variável mensurada é representada, cada coluna (Variável) de uma Tabela SAS é armazenada de acordo a um padrão de atributos SAS. Os principais atributos existentes para qualquer variável são: Nome, Tipo e Comprimento, conforme descrito no Quadro 4. Outros atributos também são armazenados em uma Tabela SAS, mas não são estabelecidos para todas variáveis (Format, Informat e Label).

Quadro 4. Características que uma variável deve assumir em uma Tabela SAS de acordo a SAS convention.

Nome da variável	<p>Deve começar com letra ou traço baixo (Letras modificadas por sinais gráficos (é, ã, ç, ...) não são consideradas letras pelo SAS); Continua com letra, número ou traço baixo (ou underscore); Pode ser letra maiúscula, minúscula ou misturado; Não pode conter espaço; Não pode conter caractere especial exceto traço baixo; Exemplo certo: Volume, VOlume, Volume1, Volume_1; Exemplo errado: 1Volume, Volume-1, Volume#1, Volume 1. Deve compreender entre 1 a 32 caracteres. Exemplo de um nome de variável com 28 caracteres: Crown_Social_Position_Xapuri.</p>
Tipo da informação na variável	<p>Pode ser numérica ou categórica; Numérica deve conter somente números entre 0 e 9 com separador decimal ponto. Pode assumir valor positivo, negativo (-33.6) ou na forma de notação científica (33E5); A representação de informação faltante em variável numérica é ponto (.); Categórica, representada por letra, número ou espaços em branco (CEP, 224-2018, 24321, #Tecnona grandis). Pode conter qualquer caractere que se possa incluir com o teclado do computador (letras, números, !@#%^&=+-, ..., etc). A representação de informação faltante em variável categórica é espaço ().</p>
Comprimento	<p>Informa quantos bytes estão sendo utilizados pelo SAS para armazenar uma variável na memória do computador; Toda variável numérica tem predefinido um comprimento de 8 bytes indiferente de quantos dígitos tiver.</p>

A seguir, um exemplo de uma Tabela SAS (Output 1) com 6 variáveis e as 5 primeiras observações de um conjunto de dados de ocorrência de furacão. A variável categórica “Summary” apresenta informações de texto conforme explicado no Quadro 4. A

representação de informações faltantes na variável é considerada pelo SAS como um espaço em branco para variável categórica e um ponto (.) para variável numérica.

Output 1. Tabela SAS com dados estruturados de ocorrência de furacão. A tabela mostra dados faltantes para a variável categórica “Summary” na observação 3 e para variável numérica “Cost” na observação 5.

Obs	Event	Date	Summary	Cost	Deaths
1	Hurricane Katrina	25AUG2005	Category 3 hurricane initially impacts the U.S. as a Category 1 near Miami, FL, then as a strong Category 3 along the eastern LA-western MS coastlines, resulting in severe storm surge damage (maximum surge probably exceeded 30 feet) along the LA-MS-AL coasts, wind damage, and the failure of parts of the levee system in New Orleans. Inland effects included high winds and some flooding in the states of AL, MS, FL, TN, KY, IN, OH, and GA.	161300000000	1833
2	Hurricane Harvey	25AUG2017	Category 4 hurricane made landfall near Rockport, Texas causing widespread damage. Harvey's devastation was most pronounced due to the large region of extreme rainfall producing historic flooding across Houston and surrounding areas. More than 30 inches of rainfall fell on 6.9 million people, while 1.25 million experienced over 45 inches and 11,000 had over 50 inches, based on 7-day rainfall totals ending August 31. This historic U.S. rainfall caused massive flooding that displaced over 30,000 people and damaged or destroyed over 200,000 homes and businesses.	125000000000	89
3	Hurricane Maria	19SEP2017		90000000000	65
	Hurricane Sandy	30OCT2012	Extensive damage across several northeastern states (MD, DE, NJ, NY, CT, MA, RI) due to high wind and coastal storm surge, particularly NY and NJ. Damage from wind, rain and heavy snow also extended more broadly to other states (NC, VA, WV, OH, PA, NH), as Sandy merged with a developing Nor'easter. Sandy's impact on major population centers caused widespread interruption to critical water / electrical services and also caused 159 deaths (72 direct, 87 indirect). Sandy also caused the New York Stock Exchange to close for two consecutive business days, which last happened in 1888 due to a major winter storm.	70900000000	159
5	Hurricane Irma	06SEP2017	Category 4 hurricane made landfall at Cudjoe Key, Florida after devastating the U.S. Virgin Islands - St John and St Thomas - as a category 5 storm. The Florida Keys were heavily impacted, as 25% of buildings were destroyed while 65% were significantly damaged. Severe wind and storm surge damage also occurred along the coasts of Florida and South Carolina. Jacksonville, FL and Charleston, SC received near-historic levels of storm surge causing significant coastal flooding. Irma maintained a maximum sustained wind of 185 mph for 37 hours, the longest in the satellite era. Irma also was a category 5 storm for longer than all other Atlantic hurricanes except Ivan in 2004.	.	97

1.3. Upload de dados no SAS Studio

Existem várias formas de introduzir dados no software SAS Studio, sendo duas as principais:

- i. Inserir dados diretamente dentro do editor (CODE) do SAS;
- ii. Importar os dados a partir de um arquivo de dados externo salvo em algum lugar (disco rígido, rede etc).

1.3.1. Inserindo os dados diretamente no editor do SAS

O método consiste em digitar os dados brutos diretamente no programa SAS (Editor) quando a base de dados é pequena e estruturada.

Para esse método utiliza-se as declarações INPUT e DATALINES (ou CARDS) do SAS considerando a seguinte sintaxe básica:

```
data sitio1;
  input Narv Capoeira$ D H IPAg Hegyi;
  datalines;
1 A 18.1 14.9 108.802 2.285
2 A 6.6 7 16.930 5.473
3 A 6.1 8 20.281 7.502
4 A 1.7 2.8 1.863 26.917
1 B 38.3 25 387.365 .
;
proc print data=sitio1;
run;
```

A declaração INPUT indica ao SAS o nome das variáveis do conjunto de dados. Para indicar ao SAS uma variável categórica utiliza-se o sinal de cifrão (\$) que deve ser adicionado logo após o nome da respectiva variável. Este é o caso da variável “Capoeira”.

A declaração DATALINES indica que os dados serão fornecidos diretamente dentro do editor do SAS. Desta forma, os dados devem ser inseridos logo a seguir da declaração

e se estende até o ponto e vírgula. O ponto e vírgula no final da lista de dados informa ao input o fim da leitura dos dados.

É importante salientar que o SAS considera como separador decimal o ponto (.) e não a vírgula (,).

Para informar ao SAS a existência de um valor faltante (missing value) no conjunto de dados, basta utilizar um ponto (.) para variável numérica e espaço vazio para variável categórica.

Após o processamento uma tabela SAS é criada e armazenada temporariamente na livreria Work. Para verificar os dados, basta abrir o arquivo ou solicitar ao SAS a impressão em tela dos dados utilizando o procedimento PROC PRINT conforme resultado do Output 2.

Output 2. Resultado do processamento dos dados com o procedimento PROC PRINT.

Obs	Narv	Capoeira	D	H	IPAg	Hegyí
1	1	A	18.1	14.9	108.802	2.285
2	2	A	6.6	7.0	16.930	5.473
3	3	A	6.1	8.0	20.281	7.502
4	4	A	1.7	2.8	1.863	26.917
5	1	B	38.3	25.0	387.365	.

Outra forma de fazer o upload de dados no SAS é copiar os dados diretamente da planilha eletrônica (p.e. MS Excel) e colar dentro do SAS quando o conjunto de dados é pequeno (Provavelmente até cerca de 5 variáveis e 100 observações).

Para isso, é necessário informar ao SAS a forma com que as variáveis estão separadas configurando SAS para ler os dados colados diretamente no editor. Para isso, no SAS Studio basta clicar em *preferences>Code and Log* e marcar a opção *Substitute spaces for tabs* e clicar em *Salvar*. A seguir, mostra-se o print da tela do SAS Studio com a janela de preferências aberta com a caixa marcada para a referida opção.

General
Start Up
Code and Log
Results
Tables
Tasks
Task Repositories
Background Jobs
Git Profiles
Git Repositories

Editor options

Enable autocomplete (Ctrl+spacebar or Command+spacebar)
 Enable hint
Tab width: spaces
 Substitute spaces for tabs
 Enable color coding
 Show line numbers
Font size:
 Enable autosave
Autosave interval: seconds

Log options

Show generated code in the SAS log
 Stream log updates while a procedure is running

With each submission

Automatically clear log
 Append log

[Reset to Defaults](#)

Save Cancel

1.3.2. Criando conjunto de dados SAS a partir de arquivo externo

Esse método permite importar dados externamente armazenados em arquivos de outros softwares para dentro do SAS como o MS Excel ou arquivos de texto. Para dados organizados em planilhas MS Excel existem várias opções para importar os dados para dentro do SAS.

Algumas formas de trabalhar com dados externos no SAS são:

- Usar procedimento de importação de dados para dentro do SAS com o PROC IMPORT;
- Ler arquivos externos em planilha eletrônica diretamente no SAS com a declaração LIBNAME.

1.3.2.1. Usando o procedimento PROC IMPORT para importar dados para o SAS

Neste caso, o procedimento PROC IMPORT irá realizar um escaneamento dos dados do MS Excel e automaticamente determinará os tipos de variáveis (Numérica ou Categórica), estabelecerá o comprimento da variável categórica e reconhecerá formatos de data.

```
proc import datafile=nome_arquivo;  
    dbms=extensão replace;  
    out=dataset;  
    getnames=yes;  
    sheet=nome_aba;
```

Onde *nome_arquivo* é o arquivo que se deseja importar, e *dataset* é o nome do conjunto de dados SAS a ser criado. A opção DBMS= indica o tipo de arquivo a ser importado (extensão), seja “xls” ou a mais recente “xlsx”. Caso o arquivo esteja separado por vírgula, deve-se indicar a extensão “csv”. A opção REPLACE indica ao SAS substituir um arquivo já existente nomeado na opção OUT=. O SAS considera o nome das variáveis indicado na primeira linha do Excel por padrão. Caso o pesquisador não queira importar o nome das variáveis, basta indicar GETNAMES=NO. Neste caso, o SAS nomeia as variáveis com letra do alfabeto (A, B, C..). Se o Excel tem mais do que uma aba de dados, você pode especificar o nome da aba que deseja importar na opção SHEET=.

1.3.2.2. Usando o comando LIBNAME para ler arquivos MS Excel

A declaração LIBNAME permite criar atalhos, através de bibliotecas SAS, para acessar diferentes tipos e formatos de dados estruturados. Um dos tipos de dados é a leitura de arquivo Excel como se fosse uma tabela SAS sem precisar importar esses dados para o SAS. A sintaxe padrão é a seguinte:

```
libname libref engine “path”;
```

A *libref* se refere ao nome da livraria que se deseja criar e que vai armazenar os dados externos. O nome da livraria deve começar com letra ou traço baixo e continuar com letra, traço baixo ou número considerando no máximo oito caracteres.

A *engine* refere-se o tipo de arquivo a ser lido pelo SAS. Existem dezenas de *engine* no SAS que permitem acessar diferentes dados com extensão: xlx, xlsx, csv, etc.

Em seguida o SAS faz a leitura do arquivo salvo no diretório informado no endereço (path) entre aspas. Logo, salva o arquivo em formato de tabela SAS na livreria SAS recém-criada.

A sintaxe a seguir acessa um arquivo Excel nomeado “*arvores*” com a extensão xlsx e salva em formato de tabela SAS na livreria a ser criada nomeada por *correl*:

```
libname correl xlsx "c:\Users\documents\arvores.xlsx";
```

Na página do autor no Youtube existem alguns vídeos de passo a passo para realizar a importação de dados do computador local para o SAS ODA. O endereço é: https://youtu.be/73Ch8tuNLBg?si=yhQ_4p3wmSZ8Xkr-.

1.4. Programação SAS

O SAS usa linguagem de programação específica e padronizada (palavras-chave) para realizar diversas análises estatísticas. Portanto, a programação SAS não é do tipo C+, R, Python ou Java por exemplo.

Basicamente é constituída por passos (Step's) reunidos por dois comandos: DATA Step e PROC Step. A Figura 4 apresenta um resumo de algumas funcionalidades do DATA STEP e PROC STEP para fins de entendimento.

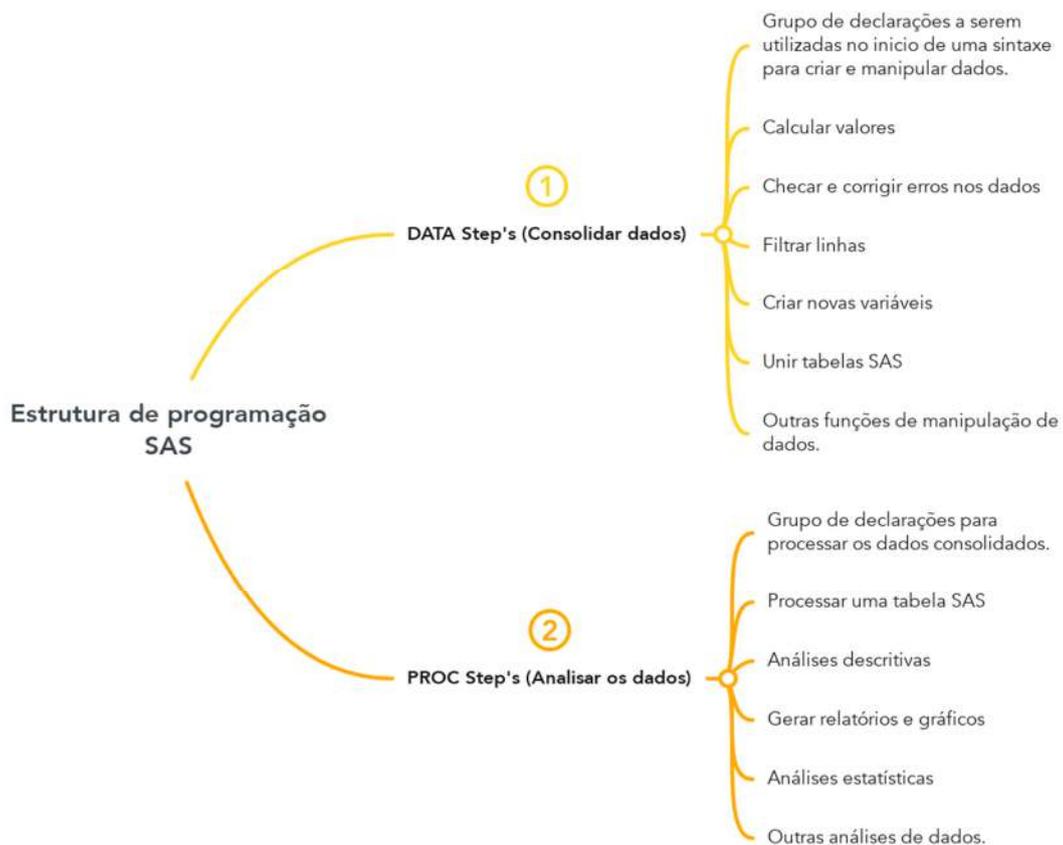


Figura 4. Estrutura de programação SAS para execução de análise de dados.

Um DATA STEP é utilizado para ler, modificar e criar arquivos de dados e sempre inicia com declaração DATA. São escritos de forma personalizada de acordo à análise de dados necessária (manipulação, tratamento etc.).

Por outro lado, PROC STEP's são procedimentos SAS padronizados, pré-escritos (built-in) e direcionados para realizar análises de dados de um conjunto de variáveis e observações criados ou manipulados em um DATA STEP.

Para cada tipo de análise estatística existe um procedimento específico. Para calcular o coeficiente de correlação de Pearson por exemplo, basta solicitar o procedimento PROC CORR. Neste caso, o nome PROC significa a abreviação de Procedure (Procedimento) e CORR de Correlation (Correlação).

No módulo SAS/STAT, por exemplo, existem vários procedimentos (PROC) dedicados para análises estatísticas que inclui análise de variância (PROC GLM, PROC ANOVA, PROC MIXED entre outros a depender do objetivo e tipo de dados), análise de

regressão (PROC REG, PROC NLIN, PROC GLM), regressão logística (PROC LOGISTIC, PROC GENMOD) e outros.

1.4.1. Erros comuns em programação SAS

Durante o aprendizado de programação SAS é comum consumir bastante tempo conferindo erros comuns que comumente ocorrem. Um dos erros mais comuns de acontecer durante o ensino de análise de dados no SAS Studio é omitir o sinal de ponto e vírgula em uma declaração SAS, por exemplo.

Um erro em um programa de computador causa um resultado indesejado e geralmente inesperado. Geralmente são divididos em três tipos: Sintaxe, Dado e Lógico (DELWICHE; SLAUGHTER, 2019).

Erro de **sintaxe** resulta de um processamento de programa em que não foram consideradas as regras para as palavras-chaves SAS (built-in) no programa. O Quadro 5 apresenta um resumo de alguns erros de sintaxe mais comuns durante o ensino de análises de dados no SAS.

Quadro 5. Alguns erros de sintaxe que comumente acontecem durante o ensino no SAS Studio. O Círculo vermelho indica a posição do erro na sintaxe.

Descrição do erro	Exemplo	Mensagem de erro no LOG
Omitir o sinal de ponto e vírgula necessário.	PROC MEANS DATA=POSCOPA	Syntax error, expecting one of the following....
Omitir sinal de igualdade quando necessário para identificar um conjunto de dados em um PROC STEP.	PROC MEANS DATA POSCOPA;	Expecting a =.
Escrever o nome do arquivo de dados faltando letra diferente do existente.	DATA POSCOPA; . . . PROC PRINT DATA=POCOPA;	File WORK.POCOPA.DATA does not exist.
Omitir o sinal de aspas quando necessário	TITLE "Efeito da copa";	The TITLE statement is ambiguous due to invalid options or unquoted text.
Omitir a declaração RUN ao final de um PROC STEP	PROC PRINT DATA=POSCOPA; ○	No error message, but no output appears in the output window until the next proc step is submitted.

A linguagem SAS é sensível para as palavras-chaves no quesito cores de texto. Isso é uma forma prática de ajudar a não cometer esse erro como mostra a Figura 5.

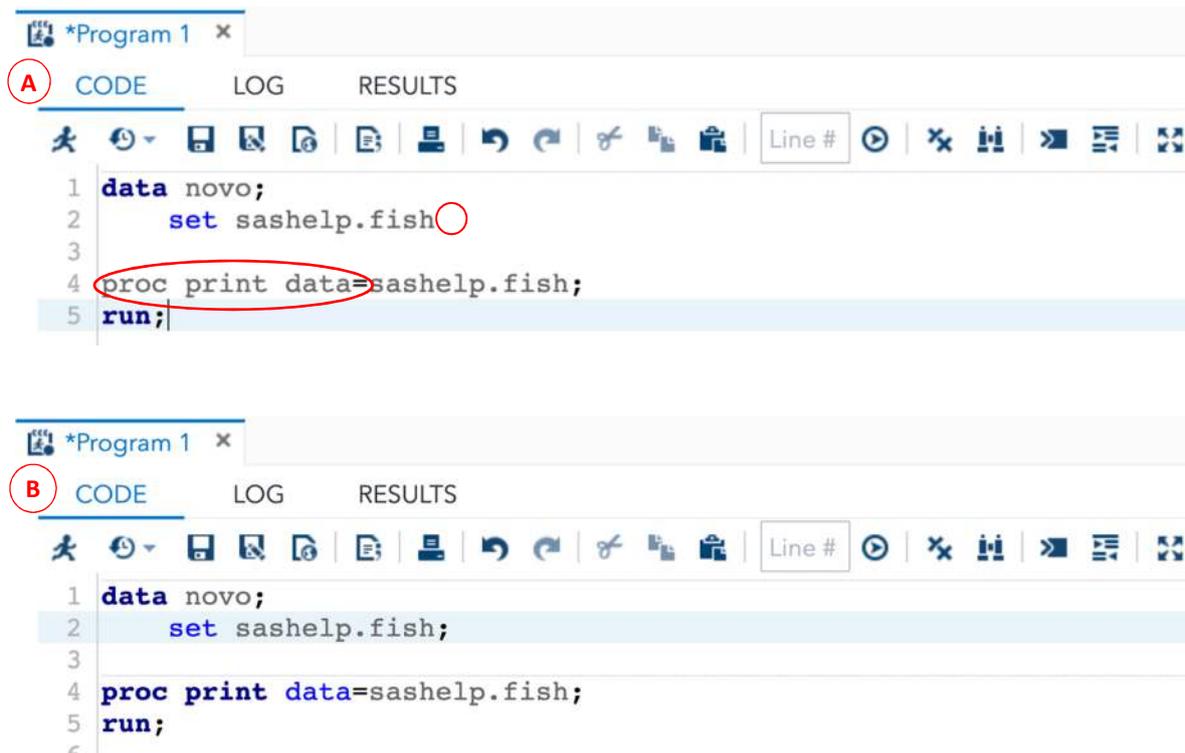


Figura 5. A= a omissão do sinal de ponto e vírgula afeta a correta cor da declaração do PROC PRINT e DATA. B= A cor azul padrão quando o erro é corrigido com a adição do sinal do ponto e vírgula.

Toda palavra-chave SAS (DATA Step ou PROC Step) aparecem em cor azul escuro, mas quando o sinal de ponto e vírgula não é utilizado no final da declaração “SET sashelp.fish” o procedimento PROC PRINT não fica azul como mostra a parte superior da figura. A forma correta do procedimento aparece na parte debaixo da figura quando se corrige o erro.

2. Processo para obtenção de dados para pesquisa

A obtenção de dados para pesquisa depende de vários passos importantes que devem ser conhecidos a priori, a fim de elaborar um planejamento robusto da pesquisa (projeto de pesquisa).

- 1) População e amostra;
- 2) Uso, tipos de variáveis e escala de medição;
- 3) Desenho de estudo estatístico a ser realizado; e
- 4) Obtenção de amostras.

A seguir descrevemos cada um dos tópicos abordados de forma a esclarecer detalhes sobre esse processo tão importante para uma pesquisa científica.

2.1. População e amostra

Uma população pode ser considerada como toda a informação de dados que interessa ser coletada para o estudo. Em todo o início de uma pesquisa científica é necessário definir qual a população (pessoas, animais, árvores, etc.) será abordada para estudá-la e desvendar algo sobre a mesma. Na área de pesquisa florestal, a população pode ser uma floresta nativa ou plantada, na qual pode ser de grande extensão (Reserva Extrativista Chico Mendes, que possui cerca de um milhão de hectares) até uma porção de floresta urbana como a área de floresta do Parque Zoobotânico da UFAC (Cerca de 100 hectares de floresta). Mas, a população pode ser também um conjunto de árvores ou até partes das árvores como podemos ter em pesquisas de laboratório.

Portanto, determinar de forma precisa qual será a população alvo da pesquisa é muito importante e depende do objetivo e seu alcance.

Se o objetivo da pesquisa é avaliar a variação da produção de frutos de castanheiras (*Bertholletia excelsa* Bonpl.) em floresta nativa de uma determinada localidade, a população alvo será o total de árvores com diâmetro a altura do peito maior ou igual a 50 cm que existem na localidade, visto que árvores silvestres de castanheiras iniciam a produção de frutos quando alcançam um diâmetro de 50 cm, em média.

Portanto, considerando toda a área florestal da Reserva Chico Mendes (1 milhão de ha), a depender do objetivo, a população pode ser considerada como:

- Número potencial de pontos sistematicamente distribuídos;
- 2.500.000 parcelas retangulares de 200 m · 20 m (caso possível); e
- Número total de árvores da espécie de interesse.

Neste sentido, uma população é toda a coleção de sujeitos/membros (árvores, animais, espécie, etc.) que possuem certas características em comum de interesse da pesquisa. Essas características são denominadas de parâmetro e usualmente são denotados por letras Gregas (KOZAK et al., 2008).

O parâmetro de uma população (descrição mais precisa) é obtido quando são realizadas medições de forma precisa em cada um dos membros da população. Esse método de abordagem é conhecido na área florestal como Censo ou Inventário 100%. Realizar um Censo representa custo alto devido ao tempo necessário para realizar a medição de todas as unidades. Entretanto, o censo florestal é necessário em casos para a obtenção do estoque de madeira de algumas espécies de árvores de valor comercial para fins de manejo florestal sustentável.

Por outro lado, uma amostra é uma parte menor da população, mas que possua características semelhantes à população alvo de pesquisa e que possa representá-la apropriadamente (amostra representativa).

Uma amostragem representativa da população fornece informações com níveis de precisão suficientes para realizar inferências em pesquisa científica. Cochran (1965) relatou que uma boa amostragem se baseia na retirada de material o mais homogêneo possível e exemplifica o fato de se obter diagnósticos laboratoriais precisos sobre a saúde humana utilizando apenas alguns mililitros de sangue. A presunção é que o sangue em circulação está sempre bem misturado e que uma gota conta a mesma história que qualquer outra.

A utilização de processos de amostragem probabilística juntamente com o cálculo do tamanho da amostra (n) assegura a obtenção de uma amostra representativa da população alvo da pesquisa.

2.1.1. Amostragem ou censo?

A decisão de realizar uma amostragem ou censo para a obtenção de dados para a pesquisa florestal leva em consideração alguns fatores, a saber:

- Realizar o censo implica em um grande investimento financeiro comparado a realizar uma amostragem visto que nesta, somente uma pequena porção da população é observada;
- O tempo disponível para a pesquisa é importante e deve ser considerado para a coleta de dados;
- Se a população de interesse para a pesquisa é pequena o suficiente pode-se conduzir um censo de modo a não acarretar em custo alto tampouco muito tempo a ser dedicado na coleta dos dados. Por outro lado, enumerar toda uma população muito grande pode não ser viável por questões de tempo e alto custo;
- No caso em que o alvo da pesquisa seja uma característica em particular como o diâmetro, a altura do peito de árvores de Seringueira, e esta variável apresentar uma pequena variação, então é possível observar apenas algumas unidades de observação para ter uma boa informação sobre a variável. Se a variação é alta, então uma amostragem pode falhar em capturar a alta dispersão na população. Neste caso, um censo pode ser mais apropriado;
- A obtenção de dados pode envolver a destruição da unidade de observação (exemplo a coleta de discos de árvores para estudo do crescimento ou testar a vida útil de um pneu do Skidder). Claramente, nestas situações um censo não seria adequado pois não sobriam árvores ou pneus;
- Em algumas situações a pesquisa necessita de um estudo aprofundado no detalhamento da unidade de observação. Neste caso, o tempo e orçamento destinados para a pesquisa favorecem a realização de uma amostragem.

Estas condições são resumidas no Quadro 6. Obviamente, na prática alguns dos fatores favorecem a realização da amostragem enquanto em outros favorecem a realização do censo na população.

Quadro 6. Condição em que o uso de amostragem ou censo deve ser considerado para obtenção de variáveis.

Fatores	Condições que favorecem o uso de:	
	Amostra	Censo
Orçamento disponível	Pequeno	Grande
Tempo disponível	Curto	Longo
Tamanho da população	Grande	Pequena
Variância da variável de resposta	Pequena	Grande
Custos relacionados a erros de amostragem	Baixo	-
Natureza da obtenção dos dados	Destrutiva	Não-destrutiva
Detalhamento profundo na medição da unidade de observação	Sim	Não
Obtenção do estoque de madeira comercial em uma floresta para manejo florestal madeireiro	Não	Sim

A fim de inferir os resultados obtidos a partir da amostra para a população (estimativas), a base matemática da estatística descritiva e inferencial possui vários termos técnicos que devem ser conhecidos a profundo para entender melhor a teoria na prática. A Figura 6 reúne alguns termos relacionados.

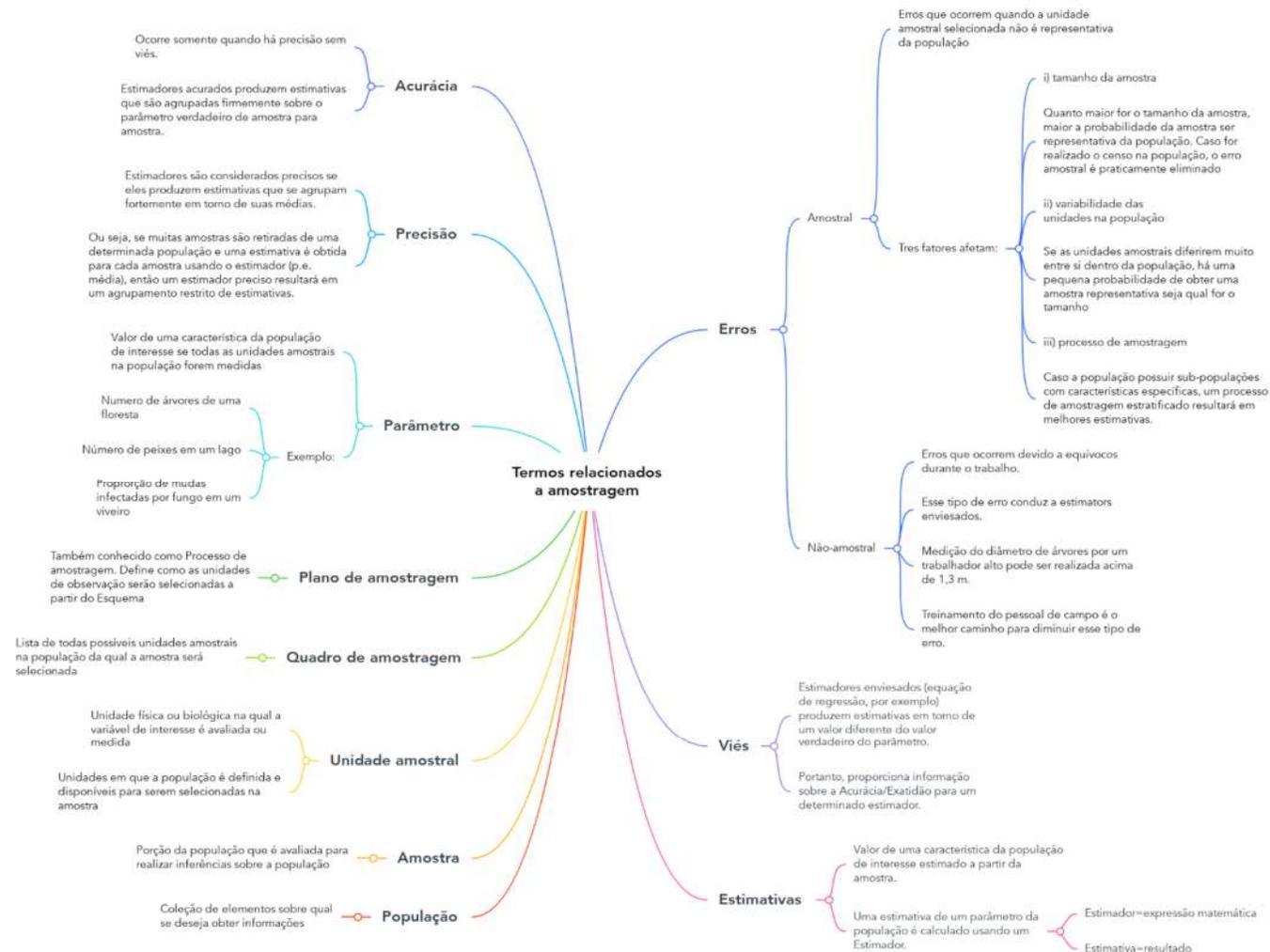


Figura 6. Termos relacionados à amostragem estatística de acordo a SHIVER e BORDES (1996).

Link para acesso ao mapa mental: <https://www.mindmeister.com/2768916757/termos-relacionados-a-amostragem>.

2.2. Tipos de variáveis e escala de medição

2.2.1. Variável dependente e variável independente

Variáveis são informações (características ou propriedades de dados) que são registradas dentro da unidade amostral podendo assumir diferentes valores ou quantidades.

Dados são informações obtidas a partir de uma unidade amostral física ou biológica que são organizados em planilhas ou textos. Uma unidade amostral física ou biológica pode ser uma árvore, animal, pessoa ou um objeto que está sendo avaliado para a pesquisa. Portanto, para cada árvore haverá uma gama de informações que representa uma característica particular denominada de Variável.

Em análise de dados para fins específicos (Análise de regressão, Análise de correlação, Comparação de médias, etc.) as variáveis coletadas são classificadas de acordo à sua função ou a forma como serão utilizadas na pesquisa em **variável dependente** e **variável independente**.

Variável dependente também conhecida como variável de resposta (Outcome em inglês) é considerada como aquela que seus valores são afetados ou dependem sob alguma suposição ou hipótese de valores de outras variáveis conhecidas, por sua vez como **variáveis independentes** (Predictor, em inglês).

O modelo de regressão linear a seguir mostra a dependência da variável v_i , que representa o volume de uma determinada árvore, que depende dos valores de diâmetro a altura do peito (D_i) e altura do fuste (hf_i) ambas **variáveis independentes** além das constantes consideradas:

$$v_i = \beta_0 + \beta_1 D_i + \beta_2 hf_i + \varepsilon_i$$

Em um caso específico da área florestal, a resposta da adubação nitrogenada na taxa de crescimento inicial de mudas exemplifica a dependência da variável crescimento (em altura ou diâmetro do colo) em função dos diferentes níveis do fator adubação nitrogenada (Ureia).

A **variável dependente** é representada, geralmente, pela letra “y” e a **variável independente** pela letra “x” (caso mais de uma, “x’s”). Isso facilita a representação dos resultados de forma gráfica, onde os valores da **variável dependente** sempre são plotados no eixo da Ordenada e a **variável independente** (ou os níveis do fator) no eixo da Abscissa.

2.2.2. Organização dos dados

Geralmente dados de pesquisa são armazenados em diferentes tipos de meio digital ou até em planilhas de papel personalizadas para pesquisa. É comum termos dados em planilhas eletrônicas (No Excel em .xls, .xlsx, .csv, etc.), em sistemas de gerenciamento de banco de dados (Access, Oracle, etc.) e até mesmo em arquivo de texto (No Word em .rtf, .dat, .txt).

A forma com que os dados são organizados em um arquivo, por exemplo no meio digital, irá influenciar como estes serão tratados antes de fazer qualquer análise estatística.

2.2.2.1. Dados estruturados

Arquivos de dados estruturados implica que o conjunto de dados está organizado em linhas e colunas. Portanto, dados estruturados são aqueles que se ajustam perfeitamente em uma planilha com linhas e colunas. As colunas representam cada uma das variáveis e as linhas representam as características particulares de cada unidade amostral.

O Quadro 7 mostra um exemplo de dados organizados em colunas e linhas. Neste caso, a variável NumArv (Número da árvore) representa a unidade amostral na qual foram mensuradas e observadas um total de sete características em cada uma das 17 unidades.

Quadro 7. Conjunto de dados para 17 árvores.

NumArv	Codigo	D	H	FormaCopa	PSocial	CFruto	NVizinhos
1	APULEI	62.5	28.9	Semi	2	0	4
2	DIPODO	53.5	27.5	Semi	2	0	6
3	APULEI	54.2	27.2	Irreg	2	1	2
4	AMBCEA	73.4	37.9	Circ	2	0	9
5	APULEI	65.5	36.8	Circ	2	0	6
6	DIPODO	76.1	28.1	Semi	2	1	9
7	DIPODO	88.1	37.9	Irreg	3	1	8
8	DIPODO	96.2	34.6	Irreg	3	0	3
9	APULEI	56.9	25.1	Semi	1	1	11
13	APULEI	86.5	27.8	Semi	2	1	4
11	AMBCEA	96.2	31.4	Semi	2	0	8
12	AMBCEA	77.1	29.5	Semi	2	1	5
13	AMBCEA	38.5	25.2	Semi	2	1	6
14	APULEI	73.5	31.6	Irreg	2	0	3
15	DIPODO	73.7	27.2	Semi	3	1	7
16	DIPODO	71.8	25.8	Irreg	3	0	8
17	DIPODO	42.5	21.9	Irreg	3	0	7

NumArv=número da árvore; Codigo=três primeiros nomes do gênero e epíteto da espécie avaliada (AMBCEA=*Amburana cearenses*, APULEI=*Apuleia leiocarpa*, DIPODO=*Dypterix odorata*); D=Diâmetro a altura do peito (cm); H=Altura total (m); FormaCopa=Forma da copa (Circ=circular, Semi=Semi-circular, Irreg=Irregular); PSocial=Posição social da árvore (1=dominante, 2= Codominante, 3=suprimida); CFruto=Presença de frutos na copa (0=não, 1=sim); NVizinhos=número de árvores ao redor da árvore objeto dentro de um raio estabelecido.

Cada coluna (variável) representa uma informação sobre uma característica particular de todas as árvores. Em inventário florestal para diagnóstico de produtividade, a unidade amostral é uma parcela utilizada para o levantamento de informações.

No MS Excel esse tipo de arquivo pode ser copiado e colado diretamente dentro do editor do SAS Studio como forma de introduzir os dados no software (Entretanto se o conjunto de dados possuir mais do que 100 observações e muitas variáveis esse processo é tedioso).

2.2.2.2. Dados não-estruturados

Em um conjunto de dados não estruturados, a organização dos mesmos não está separada por linhas e colunas, são informações que não estão organizadas de acordo com um modelo ou esquema de dados predefinidos.

Esse tipo de organização de dados é o mais abundante devido que a sua conformação pode ser qualquer coisa coletada por humanos ou até mesmo por uma máquina. Portanto, geralmente é composto por uma grande coleção de arquivos de fotos, mídia, e-mails, áudio, dados de um sensor e muito mais.

Exemplos de dados não estruturados:

- Dados de uma estação meteorológica;
- Registro de e-mails;
- Dados geoespaciais; e
- Dados de preferência de consumidores (Netflix, Amazon, etc.).

2.2.3. Tipos de variáveis e escala de medição

Em uma pesquisa de campo é comum realizar a coleta de diversas variáveis de acordo ao planejamento realizado ou considerando a oportunidade de estar em campo para mensurar variáveis potenciais.

Realizar a categorização dessas variáveis é de suma importância para, em seguida, designar qual método estatístico bem como a representação visual (gráfico) serão consideradas para responder o objetivo da pesquisa (STEVENS, 1946).

A Figura 7 mostra a divisão dos tipos de variáveis bem como a escala de medição consideradas.

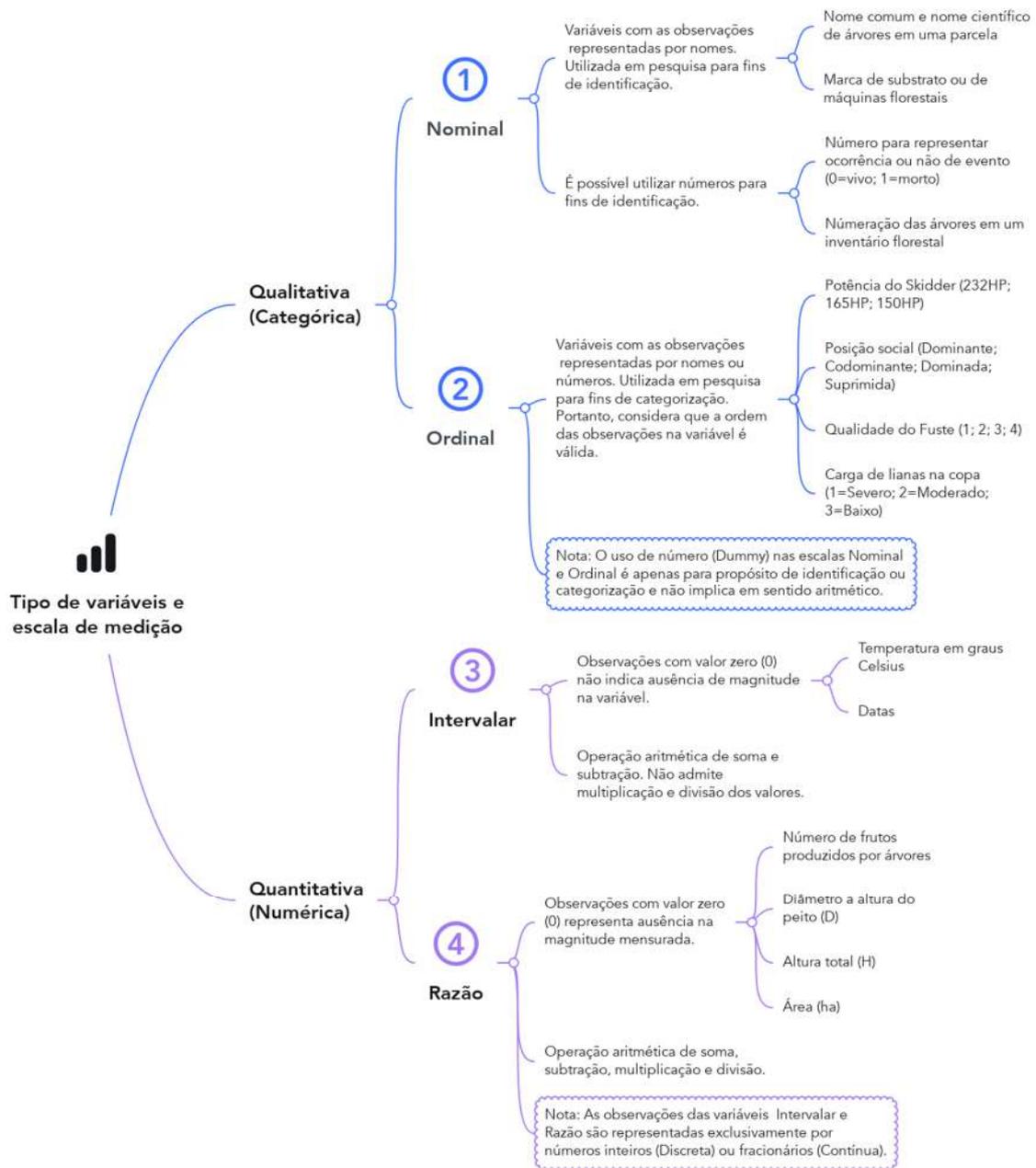


Figura 7. Tipos de variáveis e escala de medição em pesquisa científica.

Link de acesso ao mapa mental: <https://www.mindmeister.com/3332720604/tipo-de-vari-veis-e-escala-de-medi-o>.

Uma variável **qualitativa** possui níveis naturalmente diferenciáveis entre duas observações, portanto, apresentam distintas categorias. As observações são registradas em texto, número ou uma combinação dos dois (Alfanumérico) para representar a observação.

As variáveis *NumArv*, *Codigo*, *FormaCopa*, *PosSocial* e *CFruto* do Quadro 7 são exemplos de variáveis **qualitativas** que por sua vez se divide em escala **nominal** para fins apenas de identificação (*NumArv*, *Codigo*, *CFruto*) ou escala **ordinal** quando a ordem ou o rank são válidos (*FormaCopa*, *PosSocial*).

Neste caso, a forma da copa (*FormaCopa*) e a posição social (*PSocial*) apresentam observações em que a ordem importa (Uma árvore com posição social 1 está em melhores condições de iluminação do que uma árvore em posição 3. O mesmo vale para a forma da copa sendo que uma copa com forma circular (Circ), que a depender do objetivo da pesquisa, apresenta melhores condições de crescimento comparado a uma árvore com copa de categoria inferior.

Os números utilizados para representar os valores das variáveis *Codigo* e *PSocial* são apenas códigos (conhecidos como códigos Dummy) sem significado aritmético.

Para uma variável do tipo **quantitativa** as observações são representadas exclusivamente por números que representam o “quão grande” ou “quantos” em escala numérica. Esse tipo de variável se divide em **discreta** e **contínua**.

Variáveis **discretas** representam valores de observações restritas a assumir apenas um número específico de valores inteiros. Portanto, essas variáveis representam geralmente valores de contagem como por exemplo, a variável número de vizinhos (*NVizinhos*) do Quadro 7. Esse tipo de variável pode ter valores atribuídos para as quatro escalas de medição (Nominal, Ordinal, Intervalar e Razão).

Por outro lado, uma variável **contínua** pode assumir qualquer valor em um determinado intervalo e, portanto, gera valores com registros “decimais” como registrado nos valores do Quadro 7 para o diâmetro a altura do peito (*d*) e a altura total das árvores (*h*).

Os valores fracionários de uma variável **contínua** podem ser utilizados apenas para os níveis de mensuração, **intervalar** e **razão** sendo que a primeira classifica variáveis em que observações com registro zero (0) não indicam ausência de magnitude. Este é o caso da variável temperatura em que o valor zero graus (Centígrados) não indica ausência de temperatura, mas sim a temperatura em que a água congela.

Neste caso, o valor zero em observações de variável **intervalar** é arbitrário e, conseqüentemente, considera-se válido apenas as operações aritméticas de diferença e soma de valores.

Para uma variável numérica em que observações com valores zero (0) representam ausência de magnitude, ou seja, existe o conceito de zero absoluto, sua escala é do tipo **razão**. As variáveis *d* e *h* do Quadro 7 e a maioria das variáveis dendrométricas numéricas são do tipo **razão**. A temperatura em graus Kelvin é considerada uma variável do tipo razão visto que o zero absoluto especificado para esta escala é a temperatura mínima possível para a matéria (KERSHAW et al., 2017).

As operações aritméticas do tipo diferença, soma, multiplicação ou divisão são válidas para essa escala de medição.

São vastas as análises estatísticas para variável do tipo **quantitativa** com nível de mensuração **razão** visto que englobam as distribuições de probabilidade contínua (p.e. Normal, Gama, Weibull, exponencial, outras) e discreta (p.e. Poisson, outras) existentes. Para variáveis do tipo **qualitativa** não tem significado aritmético calcular soma, subtração, multiplicação ou divisão e, portanto, alguns cálculos estatísticos específicos são utilizados.

O Quadro 8 resume alguns métodos estatísticos designados de acordo ao nível de mensuração da variável. Para a análise gráfica, existem vários tipos de gráficos específicos e adequados para cada tipo de variável e nível de mensuração.

Quadro 8. Alguns métodos de análise de dados de acordo à escala de medição da **variável dependente (y)**. Adaptado de HUSCH et al. (1972).

Tipo de análise	Escala de medição da variável		
	Nominal	Ordinal	Razão
1. Medidas de tendência central	Moda ¹ .	Mediana, Percentis.	Média aritmética, Média Geométrica, Média Harmônica Mediana, Moda
2. Medidas de dispersão	-	Amplitude Interquartil.	Variância, Desvio Padrão, Amplitude, Variância da Média, Coeficiente de Variação.
3. Medidas de associação	Coeficiente de Contingência, Coeficiente Phi ² , V de Cramer ³ , Regressão Logística Binária ou Multinomial.	Correlação de Postos de Spearman, Tau-b ou Tau-c de Kendall, Regressão Logística Ordinal.	Correlação de Pearson, Análise de Regressão Linear e Não-linear.
4. Testes para inferência	Qui-quadrado.	Testes não-paramétricos (Mann-Whitney ⁴ U, teste Kruskal-Wallis ⁵ H, outros).	Testes paramétricos (Teste t, Teste F, outros).

¹= Identifica qual a espécie é mais comum entre todas da variável; ²= Adequado para tabelas de contingência 2x2; ³= Adequado para tabelas de contingência maiores que 2x2; ⁴= Comparar se a mediana é diferente ou não entre dois grupos; ⁵= Comparar se a mediana é diferente ou não entre dois ou mais grupos.

Importante destacar que para pesquisadores que irão trabalhar com dados secundários (já coletados por terceiros) é de suma importância realizar uma classificação minuciosa das variáveis de acordo ao objetivo da pesquisa, visto que a fase de planejamento da coleta de dados não foi contemplada.

2.3. Desenhos de estudos estatísticos

Refere-se à forma com que os dados coletados ou variáveis observadas estão relacionados à aplicação ou não de alguma intervenção no ambiente (campo, laboratório, animal, árvore, etc.). Na ciência florestal comumente emprega-se um dos dois tipos de estudos, a saber: Estudo Experimental e Estudo Observacional.

2.3.1. Estudo experimental (Manipulativo)

Neste tipo de estudo, o pesquisador primeiro altera os níveis de uma variável (variável independente ou Fator) e então mede o efeito dessa alteração (variável de resposta). Neste caso, os resultados são utilizados para testar a hipótese de causa-efeito.

Exemplo: Considere que uma investigação foi realizada com o objetivo de determinar se a redução da área basal proporciona maiores taxas de crescimento de árvores remanescentes em floresta natural.

Neste caso, o pesquisador deve definir os níveis de redução (ex.: 10; 25; 35 e 50 m² ha⁻¹) e aplicar em diferentes parcelas dentro da floresta. Logo, após um determinado tempo avaliar a variável de resposta (crescimento) nas árvores remanescentes.

A representação gráfica desses dados considerando o eixo x para a variável independente (níveis de redução da área basal) e o eixo y para a variável de resposta (crescimento das árvores) é apresentada na Figura 8.

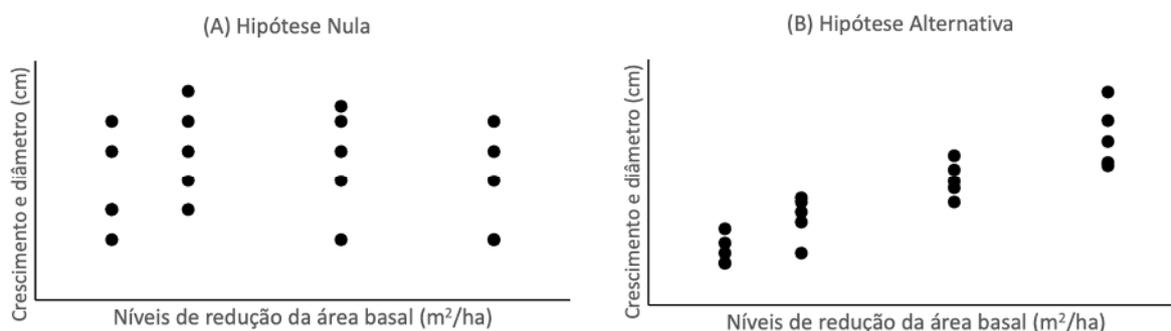


Figura 8. Relação hipotética entre a redução da área basal e o crescimento de árvores remanescentes em floresta natural.

Cada ponto representa uma parcela na qual foi aplicado o tratamento e medida a variável de resposta. A hipótese nula é de que a redução da área basal não tem efeito sobre o crescimento das árvores. A hipótese alternativa é de que a redução da área basal tem efeito sobre o crescimento das árvores proporcionando uma relação positiva entre as duas variáveis.

2.3.2. Estudo observacional

Neste tipo de estudo o pesquisador considera a variação natural presente na variável de interesse. Portanto, o ambiente ou condição não são alterados para avaliar seus efeitos na variável de resposta, ou seja, nenhum tratamento é aplicado.

A Figura 9 mostra um resumo dos tipos de desenhos de estudo aplicados na ciência florestal com alguns exemplos de aplicação.

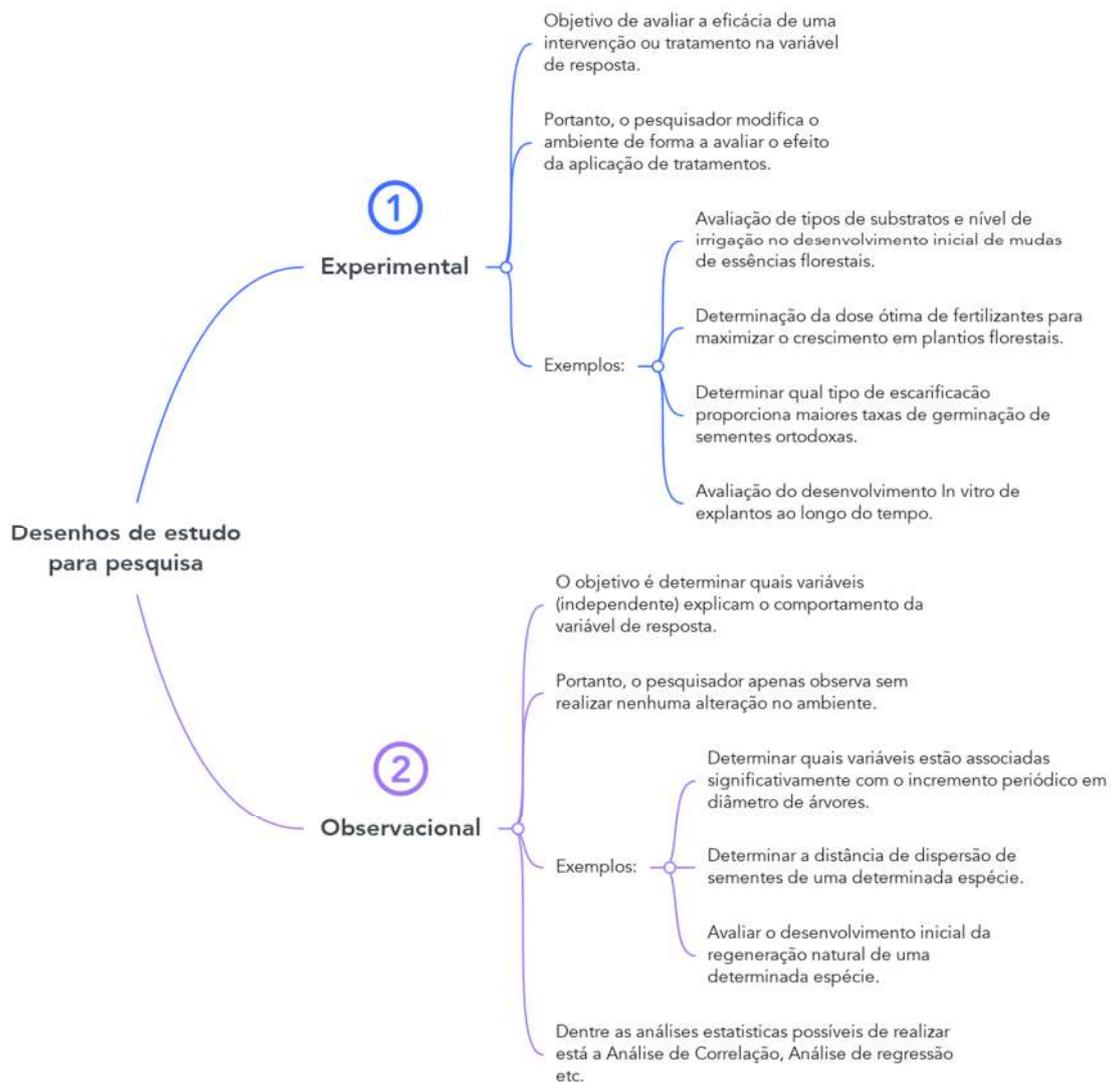


Figura 9. Tipos de desenhos de estudos para pesquisa. Em ambos os tipos é possível realizar de forma que a variável de resposta é monitorada ao longo de um período.

Fonte para acesso ao mapa mental: <https://www.mindmeister.com/2807063095/desenhos-de-estudo-para-pesquisa>.

2.4. Obtenção de amostras

Definida a população de interesse e se um censo não é aplicável, o pesquisador deve selecionar a amostra. De acordo com a população alvo, o primeiro passo a seguir é dividir a população em unidades amostrais a fim de construir o Desenho Amostral (Sampling Frame). A partir desse passo, o pesquisador conhece da primeira até a última unidade amostral, população finita. Quando possível, esse procedimento deve ser realizado de forma a abranger toda a *população* sem superposição e pode ser operacionalizado com o auxílio de imagens de satélite.

A título de exemplo apresentamos a Figura 10 que consiste no Plano Amostral de uma área de floresta, na qual apresenta uma listagem das 400 unidades amostrais delimitadas em parcelas de 0,1 ha desenvolvido por Loetsch e Haller (1964) e utilizado em Prodan (1997).

O esquema mostra também as coordenadas de localização de cada unidade amostral bem como outras subdivisões que serão discutidas em exemplos posteriores.

	I					II					III					IV					
1	130	153	153	112	200	106	100	147	118	165	0	0	12	0	35	0	18	0	0	24	A
2	124	106	136	130	165	141	194	212	136	88	100	0	12	65	88	0	100	30	12	47	
3	177	165	136	124	171	106	82	177	147	165	118	82	47	6	88	12	30	0	0	24	
4	165	112	124	118	153	118	224	136	118	159	141	65	35	24	0	30	30	53	53	30	
5	100	82	118	153	147	130	130	112	88	118	147	153	88	53	71	0	0	94	47	30	
6	224	247	217	230	130	259	277	100	147	171	200	171	118	141	82	59	71	6	0	0	B
7	253	200	135	271	277	271	230	206	242	177	141	200	135	153	106	153	124	71	30	6	
8	212	277	265	212	206	171	289	259	183	247	194	277	183	165	88	106	118	136	53	71	
9	224	283	247	300	100	318	277	306	177	200	177	271	141	71	124	71	188	171	159	94	
10	100	141	265	277	306	165	253	265	271	159	236	188	300	165	147	241	118	159	82	124	
11	277	330	253	218	177	353	330	253	171	194	241	177	177	118	88	106	118	188	77	165	C
12	224	212	159	224	141	183	283	188	147	183	206	183	130	88	59	130	141	112	106	94	
13	271	318	200	271	218	253	260	200	147	259	253	77	165	242	153	194	106	224	59	141	
14	277	277	206	236	230	230	294	165	294	212	259	159	94	124	212	100	159	124	218	200	
15	130	218	65	171	165	194	171	206	312	94	153	118	171	71	136	147	88	100	153	124	
16	218	130	118	130	82	171	147	124	177	183	159	94	124	212	100	159	124	100	82	71	D
17	106	147	153	118	159	153	153	130	112	177	88	12	41	18	24	88	53	41	0	18	
18	130	200	194	100	141	165	153	147	177	194	106	35	0	18	0	0	35	30	41	35	
19	77	165	159	159	183	118	124	124	94	159	71	0	100	18	6	6	0	0	0	30	
20	188	183	177	130	94	153	47	188	112	118	18	18	0	0	0	12	0	30	59	12	
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	

Figura 10. Exemplo de uma população finita com as unidades amostrais definidas (Quadro de Amostragem) como parcelas dentro de um povoamento florestal.

Fonte: LOETSCH e HALLER (1964).

Definido o Esquema, o pesquisador deve decidir o processo de como selecionar as unidades de dentro da população para conformar a amostra, a partir do total disponível.

A forma de como as unidades amostrais são obtidas a partir da população divide a amostragem em não-probabilística e probabilística.

2.4.1. Amostragem não-probabilística

A obtenção de unidades de amostra dentro de uma determinada população pode ser realizada considerando algumas situações em que o pesquisador pode aproveitar de forma a facilitar o acesso à população, a saber:

- i) Parcelas próximas a estradas de acesso dentro da floresta;
- ii) Parcelas localizadas em sítio com características específicas de interesse do estudo; e
- iii) Parcelas localizadas às margens de rios e igarapés.

Nestas situações a amostra obtida pode representar muito bem as condições nas quais foram instaladas as parcelas, pois o pesquisador as selecionou de forma intencional. Entretanto, pode não representar toda a população de interesse da pesquisa dada a situação específica do sítio em que a parcela foi instalada e mensurada.

Portanto, as informações obtidas a partir de unidades de amostra selecionadas dessa forma devem ser utilizadas somente para representar o sítio em questão e não podem ser utilizadas para representar toda população de uma determinada área de floresta, visto que a amostragem realizada intencionalmente próximo às vias de acesso frequentemente representam apenas características específicas do local (p.e. padrão do relevo) e, portanto, características da floresta como um todo podem não ser captadas.

Assim, caso o pesquisador desconsidere todas as outras potenciais parcelas da população de interesse pelo fato de a proximidade das vias de acesso facilitar a instalação das mesmas, por exemplo, a amostragem é considerada como amostragem não-probabilística. Desta forma, alguns cálculos não são permitidos de serem realizados para obter alguma conclusão sobre a população. Mais detalhes são apresentados na Figura 11.

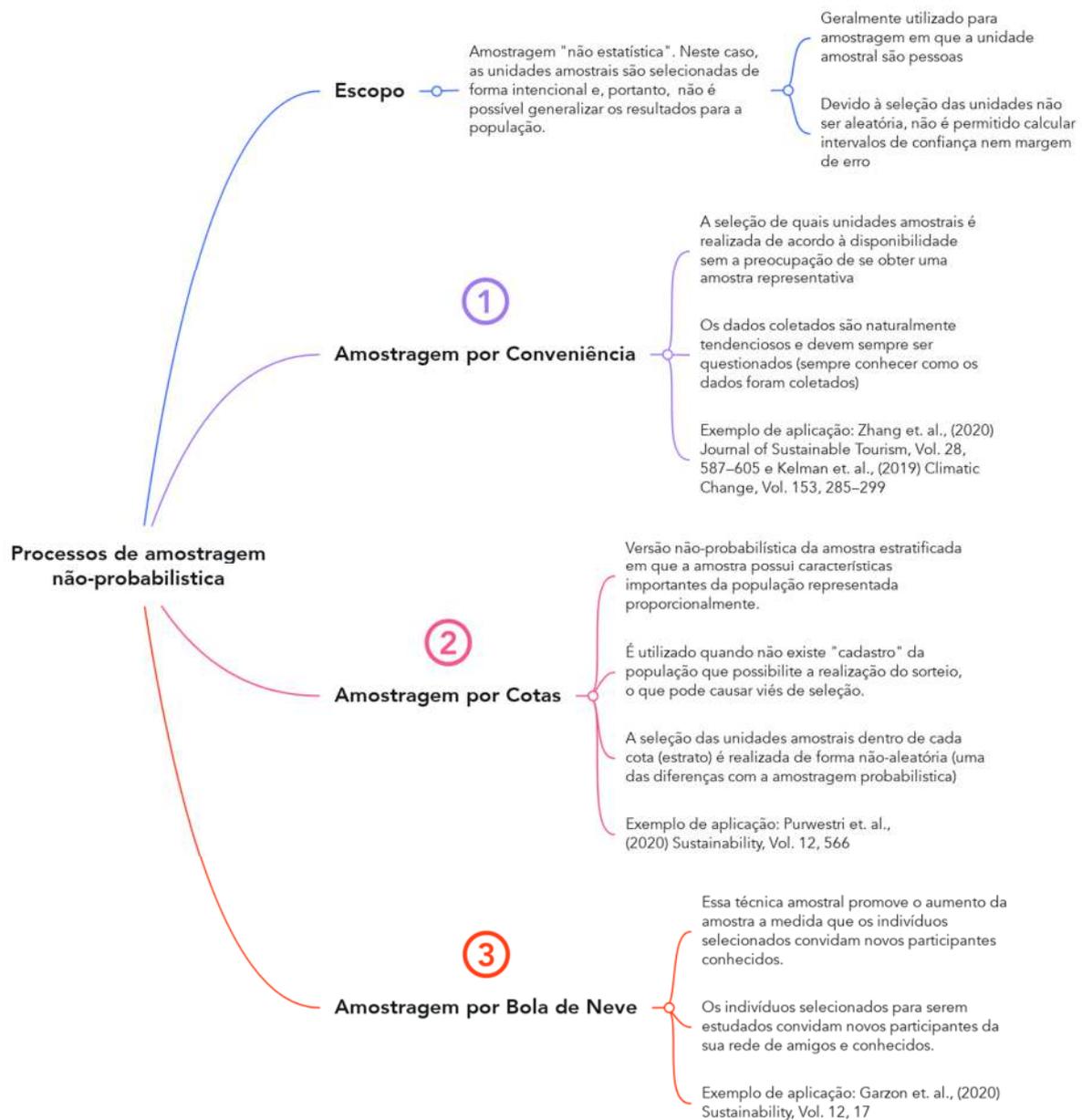


Figura 11. Alguns processos de amostragem não-probabilísticos com exemplos referenciados de aplicação na área florestal.

Fonte para acesso ao mapa mental: <https://www.mindmeister.com/2768925190/processos-de-amostragem-n-o-probabilistica>.

Os processos de amostragem não-probabilísticos comumente são aplicados quando a unidade amostral é uma pessoa, como por exemplo em pesquisas para:

- Intenção de votos para um determinado cargo;
- Determinar como produtos florestais madeireiros e não-madeireiros são utilizados por uma determinada cidade, região ou país;

- Percepção de pessoas com respeito à delimitação de uma unidade de conservação no entorno de uma cidade;
- Entrevista para caracterizar a prática silvicultural em uma determinada espécie madeireira que uma comunidade extrativista realiza.

A amostragem não-probabilística resulta em uma base de dados naturalmente tendenciosa devido à seleção intencional de unidades de amostra o que ocasiona um viés de seleção (amostra tem uma tendência embutida para excluir um determinado grupo ou característica na população) bem como o estudo pode obter resposta somente a partir de voluntários que quiseram participar da pesquisa (respondentes) e, nesta situação, caso a taxa de resposta seja pequena (percentagem de respondentes na amostra total) pode haver um viés de não-resposta. Portanto, essas situações impostas podem influenciar a representatividade e a amostra pode ser muito diferente da população alvo.

A previsão de vitória presidencial dos Estados Unidos para o candidato Thomas E. Dewey feita pelo jornal “The Chicago Daily Tribune”, em 1948 é o exemplo (Figura 12) mais famoso de métodos de pesquisa que utilizam dados obtidos de forma tendenciosa.



Figura 12. Presidente Harry S. Truman exibe alegremente uma edição inicial do Chicago Daily Tribune depois de derrotar Thomas E. Dewey na eleição presidencial de 1948 com a mensagem “Dewey derrota Truman”.

Foto: FRANK CANCELLARE.

2.4.2. Amostragem probabilística

Se cada uma das unidades amostrais do esquema tiverem a mesma probabilidade de serem selecionadas de forma independente, ou seja, a situação em que qualquer possível combinação de unidades amostrais da população tenha igual e independente chance de vir a ser a amostra (seleção aleatória das unidades de amostra), a amostragem é considerada como probabilística.

Assim, considerando que seja realizado uma amostragem aleatória simples com o intuito de estimar o volume de madeira total da área de floresta da Figura 10 (40 ha) e sejam selecionadas 20 parcelas da população, cada parcela terá uma probabilidade de seleção de $1/400$ sendo que uma parcela medida representaria outras 399 parcelas não medidas. Neste caso, cada unidade amostral tem a mesma probabilidade de ser selecionada ($1/400$).

A amostragem aleatória simples pode ser realizada com e sem reposição. Sem reposição implica na condição que após a seleção de uma unidade, esta não pode ser selecionada novamente. Para uma população hipotética com um total de 93 unidades e seja solicitada a seleção aleatória simples sem reposição de 6 unidades, um total de 762.245.484 possíveis réplicas de amostras podem ser obtidas.

Na Ciência Florestal o processo de amostragem aleatória simples é apropriado de usar quando:

- a) Existe pouca ou nenhuma informação auxiliar disponível para usar um processo de amostragem mais eficiente; e
- b) Quando se desejar utilizar métodos mais simples de estimativa dos atributos da população.

De outra forma, caso o pesquisador desejar selecionar as unidades de amostra considerando uma distância fixa entre as mesmas, o processo de amostragem será sistemático. Neste caso, as unidades são selecionadas seguindo um esquema que considera igual distância entre linhas e parcelas. Para que se cumpra o critério de probabilidade, é necessário que pelo menos a primeira unidade amostral (ponto de partida) seja selecionada de forma aleatória.

Em áreas com topografia ondulada de forma regular, deve-se considerar cautela para a adoção do processo de amostragem sistemática, pois o planejamento da seleção das unidades amostrais pode falhar causando tendenciosidade da variável de resposta. Isso pode acontecer caso as unidades selecionadas estejam localizadas, em grande parte,

no platô ou no baixio. Nesta situação, recomenda-se utilizar parcelas retangulares compridas (p.e. 20m · 200m) como forma de abranger a variação que possa existir devido às mudanças na topografia.

É possível combinar dois processos de amostragem o que pode resultar no aumento de precisão das estimativas. Neste caso, ambos os processos Aleatório Simples e Sistemático podem ser aplicados a diferentes estratos definidos dentro da população de interesse resultando assim em um processo Aleatório Estratificado e Sistemático Estratificado, respectivamente.

Esses processos de amostragem são considerados os mais usados para inventários florestais. Entretanto, existem outros processos que também resultam em uma boa precisão conforme descrito na Figura 13.

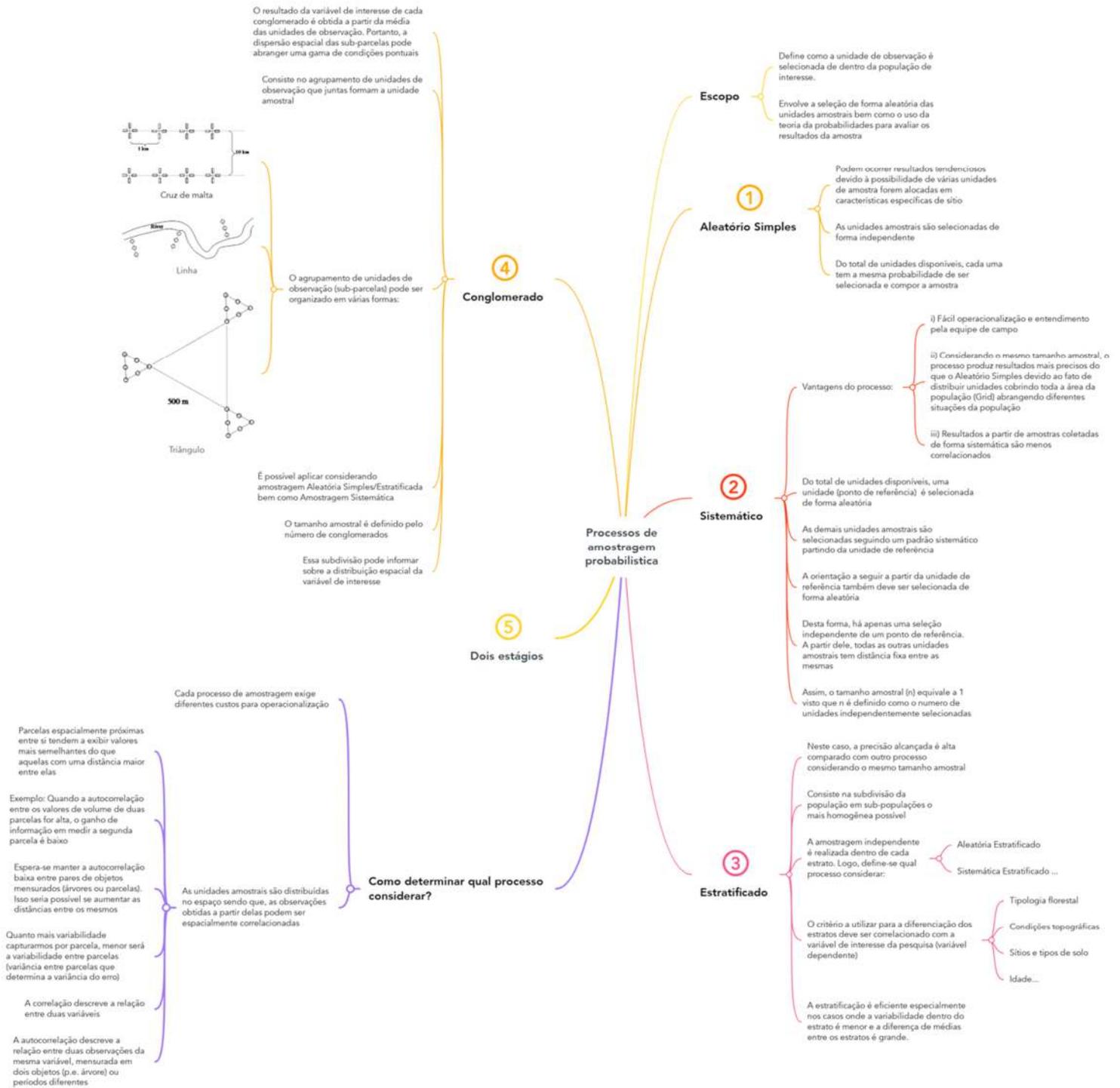


Figura 13. Alguns processos de amostragem probabilísticos utilizados na coleta de dados para pesquisa científica.

Fonte para acesso ao mapa mental: <https://www.mindmeister.com/2768905004/processos-de-amostragem-probabilistica>.

2.4.3. Métodos de amostragem

Definido a forma como será selecionada a unidade amostral dentro da população, o próximo passo é definir como será realizada a abordagem da população que se dá pela definição do formato da unidade amostral, ou seja, utilizar parcelas com uma área conhecida (área fixa) e forma definida (parcela quadrada, retangular, circular ou faixa) ou um método no qual não é possível determinar a área em que foi realizada a amostragem (área variável).

Entre os métodos de área variável o emprego do consagrado método de Bitterlich é muito eficiente para estimar variáveis que são correlacionadas com o tamanho da árvore. Em estudos para avaliar o efeito da competição, o método pode ser empregado para selecionar árvores vizinhas de acordo ao fator de área basal desejado. Para utilizar em floresta nativa em que comumente a visibilidade das árvores é limitada a poucos metros, o método pode ser empregado por meio do Vertex modelos III ou IV. Basta configurar o Fator de Área Basal desejado que o aparelho vai informar qual o diâmetro de seleção, a partir da distância que estiver entre o aparelho e o transponder. Caso a árvore em questão tiver um diâmetro igual ou maior do que o informado, esta entra na contabilização.

A Figura 14 reúne os principais métodos para amostragem.

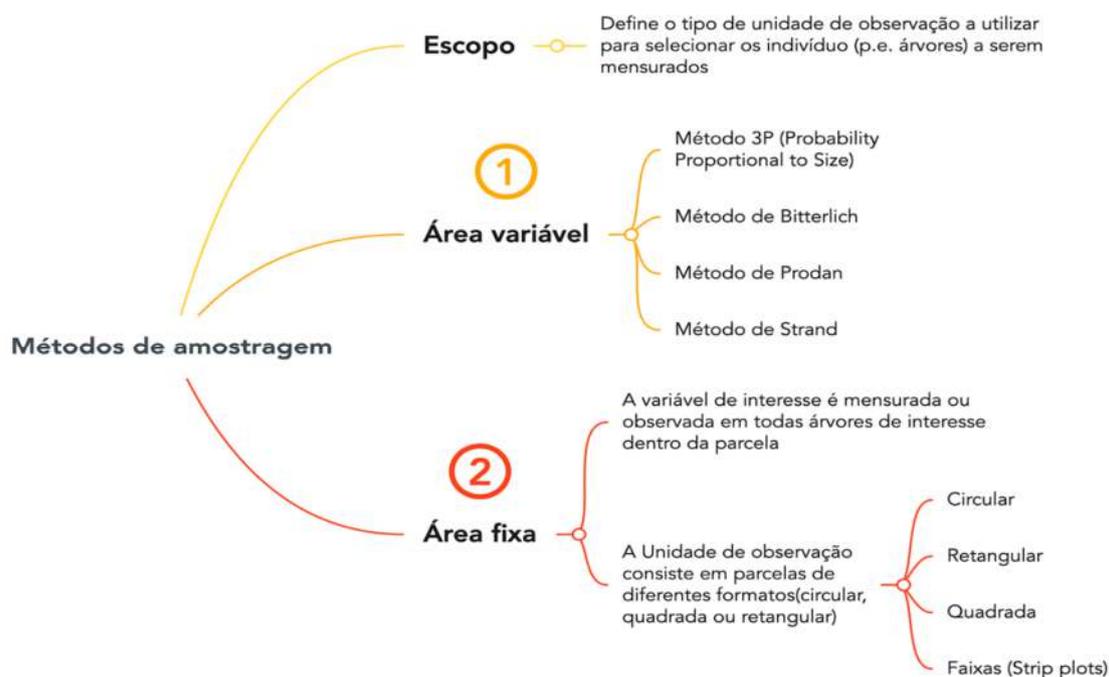


Figura 14. Alguns métodos de amostragem para obtenção de dados em pesquisa.

Fonte para acesso ao mapa mental: <https://www.mindmeister.com/2768894861/m-todos-de-amostragem>.

O objetivo dessa parte do manual é apenas de informar os processos e métodos existentes que podem ser utilizados para a obtenção de dados para pesquisa. As diversas fórmulas matemáticas e seu emprego com exemplos são explicadas em detalhes em alto nível de aprofundamento em Cochran (1965); Shiver e Borders (1996) e Péllico Netto e Brena (1997).

2.4.4. Seleção de amostras aleatórias no SAS Studio

A partir de uma população finita (Quadro de amostragem) é possível selecionar unidades amostrais de forma rápida e simples utilizando o procedimento PROC SURVEYSELECT do SAS.

Esse procedimento foi desenvolvido para realizar a seleção de unidades amostrais considerando processos de amostragem probabilística (Equal Probability Sampling) ou não-probabilística (Proportional Probability Sampling).

Os processos de amostragem probabilística suportados pelo PROC SURVEYSELECT são:

- i) Aleatória Simples sem reposição;
- ii) Aleatória Simples com reposição;
- iii) Aleatória Sistemática;
- iv) Aleatória Simples Sequencial;
- v) Amostragem Bootstrap; e
- vi) Amostragem Bernoulli.

A sintaxe básica do PROC SURVEYSELECT para a seleção aleatória de unidades de amostra dentro de uma população finita é descrita a seguir.

```
proc surveysselect
  data=quadro_de_amostragem
  out=arquivo_com_a_amostra
  method= processo_de_amostragem
  seed=número_inteiro
  sampsize= número_inteiro;
  strata=variável_estrato / opções_de_alocação;
run;
```

DATA= Identifica o conjunto de dados da população (Quadro de amostragem) de onde será extraída a amostra.

OUT= Nome do conjunto de dados que contem a amostra selecionada. É possível solicitar que o arquivo da população seja dividido (subset) de forma que indique qual unidade amostral foi selecionada. Neste caso, basta incluir a opção OUTALL e a amostra é diferenciada da população pela nova variável criada denominada de selected.

METHOD= Especifica o processo de amostragem aleatória a ser utilizado para a seleção de amostras:

Abreviação do processo no SAS	Significado
METHOD=SRS	Realiza a seleção aleatória simples sem reposição das unidades de amostra dentro da população (SRS=Simple Random Sampling).
METHOD=URS	Realiza a seleção aleatória simples com reposição das unidades de amostra dentro da população (URS=Unrestricted Random Sampling).
METHOD=SYS	Realiza a seleção aleatória simples sistemática das unidades de amostra (SYS=Systematic Sampling).
METHOD=SRS adicionando a declaração STRATA	Realiza a seleção aleatória simples das unidades de amostra dentro de cada estrato da população.

SEED= Especifica o valor somente inicial para o SAS gerar os números aleatórios para a seleção da amostra. É opcional e deve ser número inteiro. Caso essa opção não seja utilizada, o SAS considera o valor da hora do sistema como valor semente. Com o valor semente indicado, cada vez que o procedimento é executado, uma amostra idêntica é criada.

SAMPSIZE= Indica o número de unidades de amostra (por exemplo 550) a ser selecionada a partir da população. O tamanho da amostra deve ser calculado de acordo ao objetivo da pesquisa e o tipo de análise estatística a ser realizada. Deve considerar critérios como o poder do teste, erro tipo 1 entre outros.

STRATA= Especifica a variável de estratificação da população. Com essa declaração a seleção aleatória simples das unidades de amostra é realizada dentro de cada estrato da população. Essa declaração permite opções de seleção de unidades amostrais considerando algumas características dos estratos como a proporção do tamanho do estrato, a variância dentro dos estratos e o custo do levantamento de uma unidade amostral dentro dos estratos.

2.4.4.1. Preparação dos dados da população

Para realizar a seleção de unidades a partir de uma população finita, será considerado como exemplo a população do quadro de amostragem da Figura 10. De acordo a Loetsch e Haller (1964), o quadro de amostragem possui uma subdivisão macro que resulta em agrupamento de parcelas em Zonas (variando de I a IV), em Sítios (variando de A a D) e Estratos (variando de E1 a E3).

Em seguida, a separação entre cada uma das parcelas se dá pela divisão da população em linhas (variando de 1 a 20) e em colunas (variando de “a” até “t”), sendo que todas parcelas são contínuas dentro da área e sem sobreposição. Cada parcela possui uma área de 0,1 ha totalizando 40 ha para a população ($400 \cdot 0,1=40$).

Os valores dentro de cada parcela representam o volume em m^3 $0,1 \text{ ha}^{-1}$ obtido a partir do censo florestal dentro de cada parcela. Portanto, parcelas com valores zero (0) significa ausência de árvores. O volume total das 400 parcelas equivale a 5457 m^3 .

Como resultado, o conjunto de dados de volume por parcela possui sete variáveis conforme mostra a Figura 15.

	A	B	C	D	E	F	G
1	Zone	Site	Y	X	Coordinates	Stratum	Volume
2	I	A	1	a	1a	E2	130
3	I	A	1	b	1b	E2	153
4	I	A	1	c	1c	E2	153
5	I	A	1	d	1d	E2	112
6	I	A	1	e	1e	E2	200
7	II	A	1	f	1f	E2	106
8	II	A	1	g	1g	E2	100
9	II	A	1	h	1h	E2	147
10	II	A	1	i	1i	E2	118
11	II	A	1	j	1j	E2	165
12	III	A	1	k	1k	E1	0
13	III	A	1	l	1l	E1	0
14	III	A	1	m	1m	E1	12
15	III	A	1	n	1n	E1	0
16	III	A	1	o	1o	E1	35
17	IV	A	1	p	1p	E1	0
18	IV	A	1	q	1q	E1	18
19	IV	A	1	r	1r	E1	0
20	IV	A	1	s	1s	E1	0
21	IV	A	1	t	1t	E1	24
22	I	A	2	a	2a	E2	124
23	I	A	2	b	2b	E2	106
24	I	A	2	c	2c	E2	136
25	I	A	2	d	2d	E2	130
26	I	A	2	e	2e	E2	165

Figura 15. Conjunto de dados estruturados criado a partir do quadro de amostragem da população de LOETSCH e HALLER (1964) contendo sete variáveis e 400 observações (O restante das observações foi omitida). A variável “Coordinates” foi criada a partir da junção da variável y e da variável x.

2.4.5. Amostragem aleatória simples

2.4.5.1. Amostragem simples sem reposição

Para fins de aplicação prática, vamos considerar o Quadro de Amostragem da Figura 10, para selecionar de forma aleatória simples sem reposição 30 parcelas. A decisão por n=30 foi arbitrária, mas pode ser calculada.

O primeiro passo foi carregar o SAS com o conjunto de dados a partir de um Data Step. Logo, para aplicar a seleção das 30 parcelas, utilizou-se o PROC SURVEYSELECT. A sintaxe para carregar os dados e selecionar as parcelas é a seguinte:

```

Data Loetsch;
  input zone$ site$ x y$ coordenates$ stratum$ volume;
  datalines;
I A 1 a 1a E2 130
I A 1 b 1b E2 153
I A 1 c 1c E2 153
I A 1 d 1d E2 112
I A 1 e 1e E2200
I A 1 f 1f E2 106
.
.
.
IV D 20 t 20t E1 12
;

```

```

Title "Amostra aleatória simples sem reposição";
proc surveyselect data=Loetsch
  method=srs
  seed=1500
  sampsize=30
  out=asimples_n30;
run;

```

Após o processamento, o SAS gera a amostra solicitada de acordo ao Output 3.

Output 3. Resumo do procedimento SURVEYSELECT com informações solicitadas para obter a amostra.

Selection Method	Simple Random Sampling
Input Data Set	Loetsch
Random Number Seed	1500
Sample Size	30
Selection Probability	0.075
Sampling Weight	13.3333
Output Data Set	asimples_n30

Note que o tipo de seleção da amostra é indicado na primeira tabela seguido das informações consideradas para a seleção da amostra na tabela seguinte. Neste caso, a probabilidade de seleção (Selection Probability) considerada na amostragem foi de 0,075 que equivale ao quociente entre o tamanho da amostra (30) e o tamanho da população (400).

O peso amostral (Sampling Weight) de cada parcela selecionada equivale a 13,3333. Esse valor indica o número de unidades (parcelas) do quadro de amostragem (população) que a unidade amostrada representa, ou seja, a unidade selecionada representa a si mesma e a 12 outras unidades comparáveis que não foram amostradas na população de estudo. O peso de amostragem é calculado pelo inverso da probabilidade de seleção de acordo à seguinte expressão:

$$\omega_i = \frac{1}{Pr(i \in S)}$$

Em que:

ω_i =Peso de amostragem (valor maior ou igual a 1);

$Pr(i \in S)$ =Probabilidade de seleção da *i-ésima* unidade amostral pertencente a população S (valor restrito entre 0 e 1).

Assegurar que a probabilidade de seleção seja igual entre todas as unidades amostrais possível da população é um requisito para obtenção de estimativas não-enviesadas em população finita.

O arquivo contendo a amostra selecionada (asimples_n30) encontra-se na Livraria WORK do SAS System. Para visualizar a amostra basta abrir a tabela ou solicitar ao SAS a impressão em tela com a seguinte sintaxe:

```
proc print data= asimples_n30;  
run;
```

Output 4. Amostra de 30 unidades selecionadas de forma aleatória simples.

Obs	Zone	Site	y	x	Coordenates	Stratum	Volume
1	I	A	2	b	2b	E2	106
2	II	A	2	j	2j	E2	88
3	I	A	3	b	3b	E2	165
4	III	A	4	l	4l	E1	65
5	II	B	6	i	6i	E2	147
6	III	B	6	n	6n	E2	141
7	IV	B	6	q	6q	E1	71
8	III	B	8	o	8o	E2	88
9	IV	B	8	r	8r	E2	136
10	II	B	9	g	9g	E3	277
11	III	B	9	m	9m	E2	141
12	I	B	10	e	10e	E3	306
13	IV	B	10	r	10r	E2	159
14	I	C	12	d	12d	E3	224
15	II	C	12	g	12g	E3	283
16	IV	C	14	q	14q	E2	159
17	I	C	15	a	15a	E2	130
18	II	C	15	j	15j	E2	94
19	IV	C	15	t	15t	E2	124
20	I	D	17	d	17d	E2	118
21	IV	D	17	q	17q	E1	53
22	IV	D	17	s	17s	E1	0
23	II	D	18	f	18f	E2	165
24	III	D	18	l	18l	E1	35
25	II	D	19	j	19j	E2	159
26	III	D	19	l	19l	E1	0
27	I	D	20	e	20e	E2	94
28	III	D	20	n	20n	E1	0
29	III	D	20	o	20o	E1	0
30	IV	D	20	r	20r	E1	30

A partir da variável “coordenates” foi possível localizar no quadro de amostragem as parcelas selecionadas aleatoriamente conforme mostra a Figura 16.

	I					II					III					IV					
1	130	153	153	112	200	106	100	147	118	165	0	0	12	0	35	0	18	0	0	24	A
2	124	106	136	130	165	141	194	212	136	88	100	0	12	65	88	0	100	30	12	47	
3	177	165	136	124	171	106	82	177	147	165	118	82	47	6	88	12	30	0	0	24	
4	165	112	124	118	153	118	224	136	118	159	141	65	35	24	0	30	30	53	53	30	
5	100	82	118	153	147	130	130	112	88	118	147	153	88	53	71	0	0	94	47	30	
6	224	247	217	230	130	259	277	100	147	171	200	171	118	141	82	59	71	6	0	0	B
7	253	200	135	271	277	271	230	206	242	177	141	200	135	153	106	153	124	71	30	6	
8	212	277	265	212	206	171	289	259	183	247	194	277	183	165	88	106	118	136	53	71	
9	224	283	247	300	100	318	277	306	177	200	177	271	141	71	124	71	188	171	159	94	
10	100	141	265	277	306	165	253	265	271	159	236	188	300	165	147	241	118	159	82	124	
11	277	330	253	218	177	353	330	253	171	194	241	177	177	118	88	106	118	188	77	165	C
12	224	212	159	224	141	183	283	188	147	183	206	183	130	88	59	130	141	112	106	94	
13	271	318	200	271	218	253	260	200	147	259	253	77	165	242	153	194	106	224	59	141	
14	277	277	206	236	230	230	294	165	294	212	259	159	94	124	212	100	159	124	218	200	
15	130	218	65	171	165	194	171	206	312	94	153	118	171	71	136	147	88	100	153	124	
16	218	130	118	130	82	171	147	124	177	183	159	94	124	212	100	159	124	100	82	71	D
17	106	147	153	118	159	153	153	130	112	177	88	12	41	18	24	88	53	41	0	18	
18	130	200	194	100	141	165	153	147	177	194	106	35	0	18	0	0	35	30	41	35	
19	77	165	159	159	183	118	124	124	94	159	71	0	100	18	6	6	0	0	0	30	
20	188	183	177	130	94	153	47	188	112	118	18	18	0	0	0	12	0	30	59	12	
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	

Figura 16. Quadro de amostragem com demarcação das unidades amostrais selecionadas (círculo vermelho) considerando o processo de amostragem aleatória simples.

Fonte: Adaptado de LOETSCH e HALLER (1964).

O PROC SURVEYSELECT também possui a opção de selecionar a amostra considerando uma porcentagem da população utilizando a opção SAMPRATE ao invés de um número fixo de unidades a serem selecionadas. A seguinte sintaxe pode ser utilizada para selecionar um total de 5% da população.

```
proc surveyselect data=loetsch
  method=srs
  seed=1500
  samprate=0.05
  out=asimples_n5percnt;

run;
```

A partir da amostra selecionada, as medidas descritivas podem ser calculadas considerando as expressões matemáticas descritas no Quadro 9.

Quadro 9. Fórmulas de medidas descritivas univariada para volume (v) considerando a população finita com total de 400 unidades amostrais (N) e uma amostra de 45 unidades (n).

Medidas descritivas	Expressão matemática
Média populacional	$\bar{v} = \frac{\sum_{i=1}^{400} v_i}{N}$
Variância populacional	$S^2 = \frac{\sum_{i=1}^{45} (v_i - \bar{v})^2}{N-1}$
Média amostral	$\hat{v} = \frac{\sum_{i=1}^{45} v_i}{n}$
Variância amostral	$s^2 = \frac{\sum_{i=1}^{45} (v_i - \hat{v})^2}{n-1}$
Variância da média amostral	$var(\hat{v}) = \frac{s^2}{n}$
Desvio padrão da média amostral	$desvpad(\hat{v}) = \sqrt{s^2}$
Erro padrão da média amostral	$erropad(\hat{v}) = \sqrt{\frac{s^2}{n}}$
Intervalo de confiança	$\hat{v} \pm t_{\alpha/2; n-1} \sqrt{\frac{s^2}{n}}$

Entretanto, considerando que a população é finita e possui $N=400$ é necessário incluir um fator de correção $1-(n/N)$ para o cálculo das medidas de variabilidade. Caso contrário, a variância será superestimada e, conseqüentemente, a estimativa do volume terá menor precisão para fins de inferência para a população principalmente para os intervalos de confiança.

O fator de correção é multiplicador e, portanto, sua influência no cálculo da variância é verificada pela seguinte expressão:

$$var(\hat{v}) = \frac{s^2}{n} \left(1 - \frac{n}{N}\right)$$

Nota-se, portanto que à medida que n se aproxima do valor de N o fator de correção tende a zero ao ponto de anular as medidas de variabilidade na situação $n=N$. O Efeito do fator de correção é maior à medida que o tamanho da amostra aumenta. O Quadro 10 mostra a variação do fator para diferentes tamanhos de amostra.

Quadro 10. Variação do Fator de Correção para população finita considerando uma população de $N=400$.

Tamanho da amostra (n)	n/N	Fator de Correção 1- (n/N)
1	0,003	0,998
5	0,013	0,988
10	0,025	0,975
30	0,075	0,925
45	0,113	0,888
100	0,250	0,750
150	0,375	0,625
200	0,500	0,500
250	0,625	0,375
300	0,750	0,250
400	1,000	0,000

Cochran (1965) informa que o uso do Fator de Correção para calcular as medidas de variabilidade pode ser ignorado quando a fração de amostragem (n/N) não exceder a 0,05. Esse é o caso de uma amostra de 10 unidades em uma população $N=400$ ($n/N=0,025$; Quadro 10).

Para demonstrar o efeito do fator de correção na estimativa de variabilidade, considerou-se a sintaxe PROC MEANS do SAS que calcula a estatística descritiva para amostras sem considerar o uso do fator de correção da população finita. Os mesmos resultados podem ser obtidos pelo uso do PROC SURVEYMEANS. Esse procedimento pertence à família Survey de procedimentos SAS sendo específico para o cálculo de medidas descritivas de amostras obtidas em populações.

As sintaxes do PROC MEANS e do PROC SURVEYMEANS são as seguintes:

```

title1 "Calculando descritivas para a amostra";
title2 "sem considerar fator de correção";

proc means data= asimples_n30 maxdec=3 mean var std stderr lclm uclm;
    var volume;
run;

proc surveymeans data=asimples_n30 mean var;
    var volume;
run;

```

Após o processamento o Output 5 é gerado contendo as estatísticas solicitadas pelo PROC MEANS.

Output 5. Medidas descritivas para a amostra de 30 unidades selecionadas de forma aleatória simples. Neste caso não se considerou o fator de correção da população finita.

Resultado do PROC MEANS

Analysis Variable: Volume					
Mean	Variance	Std Dev	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean
118.60 00	6676.1103	81.7075	14.9177	88.0899	149.1101

Resultado do PROC SURVEYMEANS. O restante dos resultados foi omitido intencionalmente.

Statistics				
Variable	Label	Mean	Std Error of Mean	Var of Mean
Volume	Volu	118.600000	14.917674	222.53 7011

Os intervalos de confiança para a média no PROC MEANS são calculados considerando 95% de probabilidade de confiança e valor t de Student de 2,0452296421 para 29 graus de liberdade. O valor de t pode ser solicitado no SAS pela função TINV (0,975, 29).

Uma forma de considerar o fator de correção para a população finita é utilizar o PROC SURVEYMEANS do SAS. A sintaxe necessária para o cálculo das estatísticas considera o uso da opção TOTAL= na qual se indica o valor do total de unidades da população (N). Logo, o SAS calcula o fator de correção considerando o valor de n como o número de observações da amostra. A sintaxe é a seguinte:

```
title1 "Calculando descritivas para a amostra";
title2 "considerando o fator de correção";
proc surveymeans data=asimples_n30 total=400 mean var;
  var volume;
run;
```

Output 6. Medidas descritivas para a amostra de 30 unidades selecionadas de forma aleatória simples. Neste caso, o fator de correção da população finita é considerado.

Statistics						
Variable	Label	Mean	Std Error of Mean	Var of Mean	95% CL for Mean	
Volume	Volume	118.6000 00	14.347360	205.846736	89.25635 43	147.943646

Com exceção da média aritmética, o valor das medidas de variabilidade foi afetado pelo fator de correção da população. Neste caso, a variância foi um pouco menor, mas quando calculado os limites de confiança observa-se uma mudança maior:

$$\hat{v} - t_{\alpha/2;n-1} \sqrt{\frac{s^2}{n}} = 118,6 - 2,045229 \cdot 14,347360 = 89,2563m^3$$

$$\hat{v} + t_{\alpha/2;n-1} \sqrt{\frac{s^2}{n}} = 118,6 + 2,045229 \cdot 14,347360 = 147,9437m^3$$

2.4.5.2. Amostragem simples com reposição

A amostragem com reposição utilizando o PROC SURVEYSELECT é realizada quando o METHOD=URS é utilizado. Neste caso, a seleção das unidades amostrais será irrestrita (unrestricted random sampling) e algumas unidades amostrais podem ser duplicadas. Para visualizar as unidades duplicadas basta utilizar a opção OUTHITS.

A sintaxe do PROC SURVEYSELECT para amostragem aleatória simples com reposição das unidades é a seguinte:

```
proc surveyselect data=loetsch
  method=urs
  seed=1500
  sampsiz=30
  out=asimples_n30CR
  outhits;
run;
```

2.4.6. Amostragem aleatória estratificada

A estratificação é a divisão do quadro de amostragem (população) em subambientes denominados de estratos. Os estratos são naturalmente diferenciáveis entre si e exaustivos.

A justificativa para a realização da estratificação é pelo aumento da precisão das estimativas para uma população com certa variação da variável de interesse. Portanto, o emprego desse processo depende da escolha apropriada da variável que diferencia os estratos. Ademais, a estratificação garante a representação de subgrupos menos prevalentes na população (LEWIS, 2017).

A estratificação de um conjunto de dados pode ser realizada considerando a variação de variáveis numéricas ou considerando variável categórica. No caso de uma população florestal em que se tenha a variável tipologia florestal é possível considerá-la como estrato. Para uma variável numérica, é possível considerar a variação do volume por parcela para criar grupos homogêneos considerando uma amplitude de variação determinada.

A Figura 17 apresenta a população extraída de Loetsch e Haller (1964) modificada por Prodan et al. (1997) e dividida em três estratos delimitados de acordo a classes de volume sendo Estrato 1 com volume entre 0 e 100 m³; Estrato 2 com volume entre 101 a 200 m³ e Estrato 3 com parcelas apresentando volume maior do que 201 m³.

	I					II					III					IV						
1	130	153	153	112	200	106	100	147	118	165	0	0	12	0	35	0	18	0	0	24	A	E1
2	124	106	136	130	165	141	194	212	136	88	100	0	12	65	88	0	100	30	12	47		
3	177	165	136	124	171	106	82	177	147	165	118	82	47	6	88	12	30	0	0	24		
4	165	112	124	118	153	118	224	136	118	159	141	65	35	24	0	30	30	53	53	30		
5	100	82	118	153	147	130	130	112	88	118	147	153	88	53	71	0	0	94	47	30		
6	224	247	217	230	130	259	277	100	147	171	200	171	118	141	82	59	71	6	0	0	B	E3
7	253	200	135	271	277	271	230	206	242	177	141	200	135	153	106	153	124	71	30	6		
8	212	277	265	212	206	171	289	259	183	247	194	277	183	165	88	106	118	136	53	71		
9	224	283	247	300	100	318	277	306	177	200	177	271	141	71	124	71	188	171	159	94		
10	100	141	265	277	306	165	253	265	271	159	236	188	300	165	147	241	118	159	82	124		
11	277	330	253	218	177	353	330	253	171	194	241	177	177	118	88	106	118	188	77	165	C	E2
12	224	212	159	224	141	183	283	188	147	183	206	183	130	88	59	130	141	112	106	94		
13	271	318	200	271	218	253	260	200	147	259	253	77	165	242	153	194	106	224	59	141		
14	277	277	206	236	230	230	294	165	294	212	259	159	94	124	212	100	159	124	218	200		
15	130	218	65	171	165	194	171	206	312	94	153	118	171	71	136	147	88	100	153	124		
16	218	130	118	130	82	171	147	124	177	183	159	94	124	212	100	159	124	100	82	71	D	E1
17	106	147	153	118	159	153	153	130	112	177	88	12	41	18	24	88	53	41	0	18		
18	130	200	194	100	141	165	153	147	177	194	106	35	0	18	0	0	35	30	41	35		
19	77	165	159	159	183	118	124	124	94	159	71	0	100	18	6	6	0	0	0	30		
20	188	183	177	130	94	153	47	188	112	118	18	18	0	0	0	12	0	30	59	12		
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t		

Figura 17. População estratificada definida em LOETSCH e HALLER (1964).

Para aplicação do processo de amostragem estratificado o procedimento PROC SURVEYSELECT requer que as unidades amostrais estejam organizadas de forma ascendente ou descendente pelo variável estrato. O procedimento PROC SORT organiza os dados de acordo à sintaxe a seguir:

```
title "Organizando os dados por estrato";
proc sort data=loetsch;

by estrato;
run;
```

Esse procedimento organiza os dados de forma ascendente (por padrão) a partir da variável declarada em BY.

Logo, o procedimento PROC FREQ foi utilizado para verificar a representação de cada estrato com a seguinte sintaxe:

```
title "Calculando frequência de unidades por estrato";  
proc freq data=loetsch;  
    tables estrato;  
run;
```

O resultado é uma tabela de frequência absoluta e relativa de cada estrato conforme mostrado no Output 7.

Output 7. Frequência absoluta e relativa de cada estrato.

Estrato				
Estrato	Frequency	Percent	Cumulative Frequency	Cumulative Percent
E1	100	25.00	100	25.00
E2	198	49.50	298	74.50
E3	102	25.50	400	100.00

Neste caso, o resultado da tabela de frequência revela que cada estrato possui diferente número de unidades amostrais sendo que o estrato 2 é o que possui cerca de 50% do total das parcelas da população. Deve-se, portanto, ter cuidado quando selecionar uma quantidade específica de unidades por estrato de forma a assegurar a representatividade de cada um.

```

title "Selecionando 15 unidades amostrais por estrato";
proc surveyselect data=loetsch2
  method=srs
  n=15
  seed=1500
  out=samplestrata3;
  strata estrato;

run;

```

O procedimento inclui a declaração STRATA que identifica ao programa a variável índice utilizada para classificar os estratos. O processo de amostragem utilizado é o Aleatório Simples (Simple Random Sampling, SRS) no qual extrai, dentro de cada estrato, 15 parcelas de forma aleatória. O resultado do PROC SURVEYSELECT é apresentado a seguir com um resumo da seleção das amostras.

Output 8. Resumo com informações solicitadas para obter a amostra.

Selection Method	Simple Random Sampling
Strata Variable	Estrato

Input Data Set	LOETSCH2
Random Number Seed	1500
Stratum Sample Size	15
Number of Strata	3
Total Sample Size	45
Output Data Set	SAMPLESTRATA3

O resultado SAS da seleção aleatória das 15 unidades amostrais (parcelas) para cada estrato é armazenado em uma segunda tabela SAS localizado na Livraria WORK. Para visualizar a amostra basta abrir a tabela ou solicitar ao SAS a impressão em tela com a seguinte sintaxe:

```

title "Verificando a amostra";
proc print data= samplestrata3;
run;

```

O Output 9 mostra o resultado do arquivo "SampleStrata3" com as 45 unidades amostrais. A variável SelectionProb contém a probabilidade de seleção para cada parcela na amostra. Considerando que cada parcela foi selecionada com igual probabilidade, o valor da probabilidade de seleção é igual ao número de amostra do estrato dividido pelo total de unidades de amostra do estrato. Portanto, a diferença de probabilidade de seleção entre os estratos é devido a diferença no número de unidades de amostra total de cada estrato.

Output 9. Amostra de 45 unidades (15 para cada estrato).

Obs	Estrato	Zone	Site	y	x	Coordenates	Volume	SelectionProb	SamplingWeight
1	E1	IV	A	1	p	1p	0	0.15000	6.6667
2	E1	IV	A	1	r	1r	0	0.15000	6.6667
3	E1	IV	A	2	q	2q	100	0.15000	6.6667
4	E1	IV	A	3	r	3r	0	0.15000	6.6667
5	E1	IV	A	3	t	3t	24	0.15000	6.6667
6	E1	III	A	5	o	5o	71	0.15000	6.6667
7	E1	IV	A	5	r	5r	94	0.15000	6.6667
8	E1	IV	B	6	s	6s	0	0.15000	6.6667
9	E1	IV	D	17	t	17t	18	0.15000	6.6667
10	E1	III	D	18	k	18k	106	0.15000	6.6667
11	E1	IV	D	18	r	18r	30	0.15000	6.6667
12	E1	IV	D	18	s	18s	41	0.15000	6.6667
13	E1	III	D	19	n	19n	18	0.15000	6.6667
14	E1	III	D	20	k	20k	18	0.15000	6.6667
15	E1	IV	D	20	t	20t	12	0.15000	6.6667
16	E2	I	A	2	d	2d	130	0.07576	13.2000
17	E2	II	A	3	h	3h	177	0.07576	13.2000
18	E2	II	A	4	f	4f	118	0.07576	13.2000
19	E2	II	A	4	j	4j	159	0.07576	13.2000
20	E2	III	A	4	k	4k	141	0.07576	13.2000
21	E2	II	A	5	f	5f	130	0.07576	13.2000
22	E2	II	A	5	i	5i	88	0.07576	13.2000
23	E2	III	B	7	m	7m	135	0.07576	13.2000

Obs	Estrato	Zone	Site	y	x	Coordinates	Volume	SelectionProb	SamplingWeight
24	E2	IV	B	8	r	8r	136	0.07576	13.2000
25	E2	IV	B	9	s	9s	159	0.07576	13.2000
26	E2	III	C	14	l	14l	159	0.07576	13.2000
27	E2	II	D	16	g	16g	147	0.07576	13.2000
28	E2	I	D	17	e	17e	159	0.07576	13.2000
29	E2	I	D	18	c	18c	194	0.07576	13.2000
30	E2	II	D	18	f	18f	165	0.07576	13.2000
31	E3	I	B	7	a	7a	253	0.14706	6.8000
32	E3	II	B	7	g	7g	230	0.14706	6.8000
33	E3	II	B	7	h	7h	206	0.14706	6.8000
34	E3	II	B	7	j	7j	177	0.14706	6.8000
35	E3	I	B	8	c	8c	265	0.14706	6.8000
36	E3	I	B	9	c	9c	247	0.14706	6.8000
37	E3	I	B	10	c	10c	265	0.14706	6.8000
38	E3	I	C	11	d	11d	218	0.14706	6.8000
39	E3	I	C	11	e	11e	177	0.14706	6.8000
40	E3	II	C	13	f	13f	253	0.14706	6.8000
41	E3	II	C	13	h	13h	200	0.14706	6.8000
42	E3	III	C	13	k	13k	253	0.14706	6.8000
43	E3	I	C	14	a	14a	277	0.14706	6.8000
44	E3	I	C	14	c	14c	206	0.14706	6.8000
45	E3	II	C	14	f	14f	230	0.14706	6.8000

Como observado na tabela de frequência, cada estrato possui um número diferente de parcelas. A representatividade dos estratos variou, sendo que uma parcela da amostra representa outras 7 parcelas para os estratos E1 e E3, enquanto para o estrato E2, uma parcela selecionada representa outras 13.

Uma das justificativas de realizar a estratificação de uma população é o aumento substancial da precisão nas estimativas. Portanto, na obtenção da estatística descritiva a partir da amostra estratificada, deve-se considerar o efeito de estrato no cálculo.

Uma forma de obtenção da estatística descritiva para amostras obtidas a partir de população estratificada é pelo uso do PROC SURVEYMEANS do SAS. Para fins de demonstração de precisão nas estimativas, considerou-se a sintaxe sem estratificação (como se fosse realizado a amostragem aleatória simples em toda a população) e o efeito da estratificação conforme a seguir:

```

title1 "Calculando descritivas para a amostra estratificada";
title2 "sem considerar o efeito de estrato";
proc surveymeans data= samplestrata3 total=400 mean var clm;
    var volume;
run;

title1 "Calculando descritivas para a amostra estratificada";
title2 "considerando o efeito de estrato";
proc surveymeans data= samplestrata3 total=400 mean var clm;
    strata stratum;
    var volume;
run;

```

Observa-se que a única diferença entre os dois procedimentos é o uso da declaração STRATA na qual se informa a variável que estratifica a população no conjunto de dados (stratum). Após o processamento, os resultados são apresentados no Output 10.

Output 10. Medidas de variabilidade a partir de uma amostra de 45 unidades amostrais.

Resultados sem considerar o efeito de estratificação.

Data Summary	
Number of Observations	45

Statistics						
Variable	Label	Mean	Std Error of Mean	Var of Mean	95% CL for Mean	
Volume	Volume	137.466667	12.181970	148.400399	112.915519	162.017814

Resultados considerando o efeito de estratificação.

Data Summary	
Number of Observations	45

Statistics						
Variable	Label	Mean	Std Erro of Mean	Var of Mean	95% CL for Mean	
Volume	Volume	137.466667	4.737367	22.442648	127.906273	147.027061

Os resultados para o erro padrão e variância da média amostral foram drasticamente reduzidos quando o efeito de estratificação foi considerado no cálculo dessas estatísticas. Neste caso, a precisão é refletida diretamente nos limites de confiança com valores mínimo e máximo mais próximos da média.

Isto se deve ao processo de cálculo conforme denota a expressão a seguir para o cálculo da variância da média amostral:

$$var(\hat{v}) = \sum_{h=1}^{H=3} \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right)$$

Em que:

$var(\hat{v})$ =variância da média amostral para população estratificada;

N_h =número de unidades dentro do estrato específico;

N =número total de unidades da população;

n_h =número de unidades amostradas dentro do estrato específico;

s_h^2 =variância amostral dentro do estrato específico.

2.4.7. Amostragem aleatória estratificada com alocação otimizada

Partindo de um número de unidades definido (Tamanho amostral), é possível distribuir um número de unidades em cada estrato considerando situações específicas como tamanho do estrato, a variância dentro do estrato e o custo do levantamento de uma unidade amostral dentro do estrato.

O PROC SURVEYSELECT distribui as unidades amostrais dentro de cada estrato por meio da utilização de alguns métodos de alocação de unidades amostrais utilizando a opção `ALLOC=método_de_alocação` na declaração STRATA.

Os três métodos de alocação disponíveis no PROC SURVEYSELECT são descritos a seguir com exemplo de aplicação.

2.4.7.1. Método de alocação Proporcional (Proportional)

Esse método aloca o tamanho total da amostra proporcional ao tamanho dos estratos. O tamanho do estrato equivale ao número total de unidades amostrais dentro do estrato (parcelas, pessoas, animais, área, etc.).

O cálculo da proporção de alocação para cada estrato h é obtido a partir da seguinte fórmula:

$$f_h^* = \frac{N_h}{N}$$

Onde: f_h^* =proporção de alocação das unidades amostrais; N_h =número de unidades amostrais no estrato h ; N =número total de unidades amostrais para todos estratos.

Desta forma, caso seja solicitada o número total de unidades (n) a serem selecionadas pela opção `SAMPsize=`, o procedimento divide o número de unidades amostrais entre os estratos h de acordo com a fórmula:

$$n_h^* = f_h^* \cdot n$$

Onde: n_h^* =número de unidades amostrais destinadas para o estrato h . Esse valor deve ser inteiro de acordo a restrições de arredondamento consideradas pelo procedimento; n =número de unidades amostrais solicitadas na opção `SAMPsize=`.

A seguinte sintaxe seleciona um total de 45 unidades de amostra entre os estratos considerando a alocação proporcional. O resultado é apresentado no Output 11.

```
title1 "Selecionando amostras por estrato";
title2 "Proporcional ao tamanho do estrato";
proc surveysselect data=loetsch
  method=srs
  n=45
  seed=1500
  out=samplestrata4;
  strata estrato / alloc=proportional;
run;
```

Output 11. Resumo com informações solicitadas para obter a amostra.

Selection Method	Simple Random Sampling
Strata Variable	Stratum
Allocation	Proportional

Input Data Set	LOETSCH2
Random Number Seed	1500
Number of Strata	3
Total Sample Size	45
Output Data Set	SAMPLESTRATA4

Após solicitado a impressão em tela, o resultado da seleção das 45 unidades de amostra é apresentado no Output 12. considerando a proporção do tamanho de cada estrato.

Output 12. Resultado da amostra selecionada.

Obs	Estrato	Zone	Site	y	x	Coordenates	Volume	Total	AllocProportion	SampleSize
1	E1	IV	A	1	p	1p	0	100	0.250	11
2	E1	IV	A	1	r	1r	0	100	0.250	11
3	E1	IV	A	2	r	2r	30	100	0.250	11
4	E1	IV	A	3	t	3t	24	100	0.250	11
5	E1	IV	A	5	p	5p	0	100	0.250	11
6	E1	IV	A	5	t	5t	30	100	0.250	11
7	E1	III	D	18	m	18m	0	100	0.250	11
8	E1	III	D	19	l	19l	0	100	0.250	11
9	E1	III	D	19	m	19m	100	100	0.250	11
10	E1	IV	D	19	r	19r	0	100	0.250	11
11	E1	III	D	20	o	20o	0	100	0.250	11
12	E2	I	A	1	e	1e	200	198	0.495	22
13	E2	I	A	2	c	2c	136	198	0.495	22
14	E2	II	A	3	g	3g	82	198	0.495	22
15	E2	I	A	4	e	4e	153	198	0.495	22
16	E2	II	A	4	h	4h	136	198	0.495	22
17	E2	II	A	4	i	4i	118	198	0.495	22
18	E2	I	A	5	d	5d	153	198	0.495	22
19	E2	II	A	5	g	5g	130	198	0.495	22
20	E2	III	B	6	m	6m	118	198	0.495	22

Obs	Estrato	Zone	Site	y	x	Coordenates	Volume	Total	AllocProportion	SampleSize
21	E2	IV	B	7	q	7q	124	198	0.495	22
22	E2	IV	B	8	p	8p	106	198	0.495	22
23	E2	IV	B	9	p	9p	71	198	0.495	22
24	E2	IV	B	9	q	9q	188	198	0.495	22
25	E2	IV	C	13	q	13q	106	198	0.495	22
26	E2	II	C	15	g	15g	171	198	0.495	22
27	E2	III	C	15	o	15o	136	198	0.495	22
28	E2	IV	C	15	s	15s	153	198	0.495	22
29	E2	I	D	16	b	16b	130	198	0.495	22
30	E2	IV	D	16	q	16q	124	198	0.495	22
31	E2	II	D	17	g	17g	153	198	0.495	22
32	E2	II	D	17	j	17j	177	198	0.495	22
33	E2	II	D	20	h	20h	188	198	0.495	22
34	E3	II	B	7	h	7h	206	102	0.255	12
35	E3	I	B	8	d	8d	212	102	0.255	12
36	E3	I	B	10	d	10d	277	102	0.255	12
37	E3	II	C	11	f	11f	353	102	0.255	12
38	E3	II	C	11	g	11g	330	102	0.255	12
39	E3	II	C	13	h	13h	200	102	0.255	12
40	E3	II	C	13	j	13j	259	102	0.255	12
41	E3	I	C	14	c	14c	206	102	0.255	12
42	E3	I	C	14	d	14d	236	102	0.255	12
43	E3	II	C	14	f	14f	230	102	0.255	12
44	E3	II	C	14	i	14i	294	102	0.255	12
45	E3	II	C	14	j	14j	212	102	0.255	12

Neste caso, o número de unidades amostrais variou de acordo com o tamanho do estrato, sendo que o estrato E2 teve maior número de parcelas (22 parcelas) e os demais estratos com 11 e 12 parcelas.

2.4.7.2. Método de alocação de Neyman (Neyman)

Esse método aloca o tamanho total da amostra entre os estratos em proporção ao tamanho dos estratos e à sua variância (s^2). A obtenção do número de unidades amostrais para o estrato h se dá por meio do emprego da seguinte fórmula:

$$f_h^* = N_h \cdot S_h / \sum_{i=1}^H N_i \cdot S_i$$

Onde: f_h^* =proporção de alocação das unidades amostrais; N_h =número de unidades amostrais no estrato h ; S_h =desvio padrão do estrato h ; H =número total de estratos.

Para utilizar essa opção deve-se fornecer o valor da variância da variável de interesse para cada estrato na opção VAR=() separados por vírgula ou espaço em branco na mesma ordem do conjunto de dados de input.

A sintaxe a seguir solicita a seleção de 45 parcelas da população estratificada considerando a variância de cada estrato. Antes, a variância de cada estrato foi obtida pelo procedimento PROC MEANS (Neste caso, considerou-se a variância populacional).

```

title 'population volume means by stratum';
proc means data=loetsch mean var std min max;
  var volume;
  by estrato;
run;

title 'stratified sampling with neyman allocation';
proc surveysselect data=loetsch
  method=srs
  seed=1500
  sampsize=45
  out=sample_neyman;
  strata estrato / alloc=neyman var=(1025.79 1452.31 2844.03);
run;

```

2.4.7.3. Método de alocação Ótima (Optimal)

Esse método aloca o tamanho total da amostra entre os estratos em proporção ao tamanho, variação e custos dos estratos. Para utilizar essa opção deve-se fornecer a variação do estrato na opção VAR= e opção COST=(). Os valores a serem informados em COST devem ser positivos e representam o custo de estratos separados por espaço em

branco ou vírgula, ou seja, para cada estrato deve ser informado o custo de levantamento de uma única unidade no estrato.

O PROC SURVEYSELECT realiza a alocação ótima de unidades amostrais para cada estrato h de acordo aos resultados da seguinte fórmula:

$$f_h^* = \frac{N_h \cdot S_h}{\sqrt{C_h}} / \sum_{i=1}^H \frac{N_i \cdot S_i}{\sqrt{C_i}}$$

Onde: f_h^* =proporção de alocação das unidades amostrais em cada h estrato; N_h =número de unidades amostrais no estrato h ; S_h =desvio padrão do estrato h ; C_h =custo de levantamento de uma unidade amostral no estrato h ; H =número total de estratos.

Neste caso, haverá um balanceamento no quantitativo de amostras para cada estrato considerando uma minimização da variância total para um determinado custo ou uma minimização do custo total para uma determinada variância. Neste caso, o procedimento considera uma redução de unidades amostrais para estratos com maior custo e, conseqüentemente, aumento de unidades a serem amostradas para estratos de menor custo.

A sintaxe a seguir solicita a seleção de 45 parcelas da população estratificada considerando a variância e o custo de levantamento de uma unidade de amostra em cada estrato. O custo informado foi arbitrário par fins de exemplificação.

```
title 'stratified sampling with optimal allocation';
proc surveyselect data=loetsch
  method=srs
  seed=1500
  sampsize=45
  out=sample_optimal;
  strata estrato / alloc=optimal
    var=(1025.79 1452.31 2844.03)
    cost=(900 280 250);
run;
```

Considerando as informações de variância e custos na sintaxe indicadas e a população Loetsch da Figura 17, procedeu-se ao cálculo da proporção ótima para cada um dos três estratos da seguinte maneira:

Crítérios	Estrato 1	Estrato 2	Estrato 3
N_h	100	198	102
S_h	32,0280	38,1092	53,3294
C_h	900	280	250
$\sum_{i=1}^H \frac{N_i \cdot S_i}{\sqrt{C_i}}$	901,7276		

Para o estrato E1 a proporção de alocação de unidades é:

$$f_1^* = \frac{100 \cdot 32,0280}{\sqrt{900}} / 901,7276 = 0,1184$$

Para o estrato E2 a proporção de alocação de unidades é:

$$f_2^* = \frac{198 \cdot 38,1092}{\sqrt{280}} / 901,7276 = 0,5001$$

Para o estrato E3 a proporção de alocação de unidades é:

$$f_3^* = \frac{102 \cdot 53,3294}{\sqrt{250}} / 901,7276 = 0,3815$$

Portanto, a partir de uma amostra total de 45 unidades, a alocação de unidades a serem amostradas em cada estrato é a seguinte:

Para o estrato E1 o número de unidades é:

$$n_1^* = 0,1184 \cdot 45 = 5,3280 \approx 5 \text{ parcelas}$$

Para o estrato E2 o número de unidades é:

$$n_2^* = 0,5001 \cdot 45 = 22,5045 \approx 23 \text{ parcelas}$$

Para o estrato E3 o número de unidades é:

$$n_3^* = 0,3815 \cdot 45 = 17,1675 \approx 17 \text{ parcelas}$$

O Quadro 11 mostra a distribuição do quantitativo de unidades amostrais otimizadas pelos métodos de alocação de unidades amostrais pelo processo de amostragem aleatório simples dentro de cada estrato.

Partindo de um total de 45 unidades amostrais, o estrato E2 foi o mais representado referente ao número de unidades independentemente do método de otimização. Isso se deve ao maior tamanho do estrato entre os demais (198 parcelas).

A otimização de NEYMAN considerando a variância aumentou de 12 para 15 o número de unidades para o estrato E3 devido a que este apresenta a maior variância para o volume.

Considerando o custo do levantamento de uma parcela em campo, a otimização OPTIMAL aumentou duas unidades amostrais para o estrato E3.

Vale destacar que a otimização considerando custos (ALLOC=OPTIMAL) reduziu o número de parcelas do estrato E1 de 11 para 5 parcelas considerando que esse estrato foi o que apresentou o maior custo.

Quadro 11. Impacto na distribuição do número de unidades amostrais por estrato de acordo ao método de otimização considerado no PROC SURVEYSELECT.

Estratos	N	S ²	Custo* (R\$)	Método de otimização da seleção de unidades			
				Sem otimização	PROPORTIONAL	NEYMAN	OPTIMAL
E1	100	1025,79	900	15	11	9	5
E2	198	1452,31	280	15	22	21	23
E3	102	2844,03	250	15	12	15	17
Total de unidades				45	45	45	45

*Valores de custo indicados pelos autores para fins de exemplificação.

Cada método de otimização proporcionou diferentes níveis de acurácia para a média amostral e os intervalos de confiança de acordo a Figura 18.

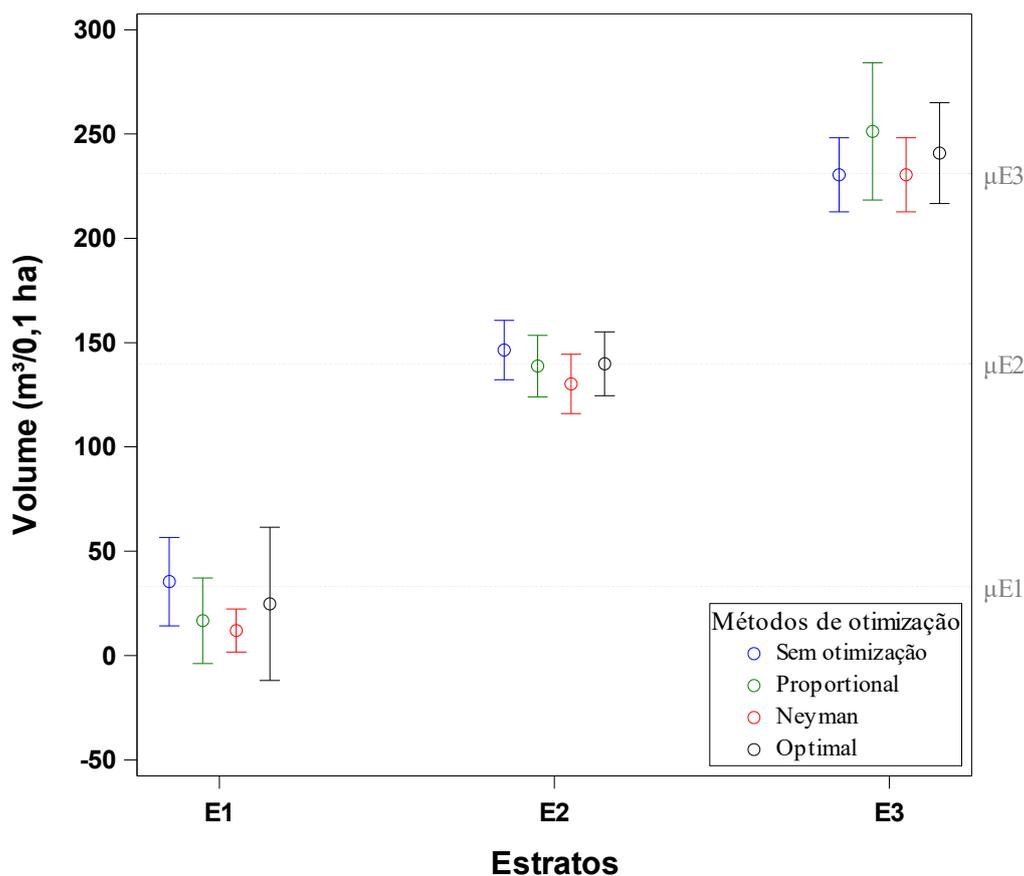


Figura 18. Comportamento da média aritmética e limites de confiança por estrato, estimados a partir de amostras obtidas pelos métodos de otimização. A linha de referência no eixo y (descontínua) representa a média populacional de cada estrato ($\mu E1=33 \text{ m}^3$, $\mu E2=140 \text{ m}^3$, $\mu E3=231 \text{ m}^3$).

Entre os estratos avaliados, a estimativa da média amostral mais próxima da média populacional do estrato foi obtida com a distribuição equitativa de 15 unidades entre os três estratos (cor azul).

Houve variação entre os métodos de otimização na estimativa da média e dos intervalos de confiança com destaque para o método de Neyman que apresentou os intervalos de confiança mais próximos da média (estreitos) no estrato 1 (E1).

2.4.8. Teorema Central do Limite e o tamanho da amostra

De acordo com o Teorema Central do Limite (TCL), a distribuição de uma sequência de médias calculada a partir de amostras independentes e aleatórias de tamanho “ n ” sempre terá distribuição Normal desde que a amostra (n) seja suficientemente grande.

Neste sentido, a população (N) não necessariamente precisaria ter uma distribuição Normal para que o TCL seja válido.

Assim, se uma população em que a variável dependente tem uma distribuição muito diferente da Normal e nesta forem extraídas várias amostras com reposição (réplicas), e para cada amostra calcular a média aritmética, a distribuição dessas médias terá uma distribuição Normal.

Em parte, a apropriação do TCL é importante pois permite a utilização de cálculos para realizar inferências estatísticas para população com distribuição desconhecida desde que a variável sob análise seja numérica.

A aproximação da distribuição Normal para as médias é maior à medida que se aumenta o tamanho da amostra. Entretanto, em estudo observacional existem questionamentos sobre o quão grande deveria ser uma amostra para considerar a afirmação do TCL. Uma amostra seria grande o suficiente com um $n=10$, $n=30$ ou $n=100$?

A fim de demonstrar na prática a teoria do TCL vamos considerar o caso florestal 1 a seguir.

Caso florestal 1: Comprovação do Teorema Central do Limite.

Considere a população finita em que a variável volume não tenha distribuição Normal. Para tal, utilize a população Loetsch da Figura 10. Logo, pede-se:

- a) Verifique a aderência à distribuição Normal para a variável volume da população. Construa um histograma de frequência incluindo a curva Normal hipotética no gráfico;
- b) Solicite ao SAS Studio a seleção aleatória simples de unidades de amostra com tamanho (n) igual a $n=10$, $n=30$ e $n=100$. Para cada amostra solicite um total de 500 réplicas;
- c) Construa um histograma de frequência para a população e para cada amostra incluindo a curva Normal hipotética no histograma.

Para avaliar a aderência à distribuição Normal e construir os histogramas de frequência será utilizado o procedimento PROC UNIVARIATE do SAS. Esse procedimento será abordado em capítulos posteriores. A sintaxe do procedimento é apresentada a seguir:

```

Data Loetsch;
  input zone$ site$ x y$ coordenates$ stratum$ volume;
  datalines;
I A 1 a 1a E2 130
I A 1 b 1b E2 153
I A 1 c 1c E2 153
I A 1 d 1d E2 112
I A 1 e 1e E2200
I A 1 f 1f E2 106
.
.
.
IV D 20 t 20t E1 12
;

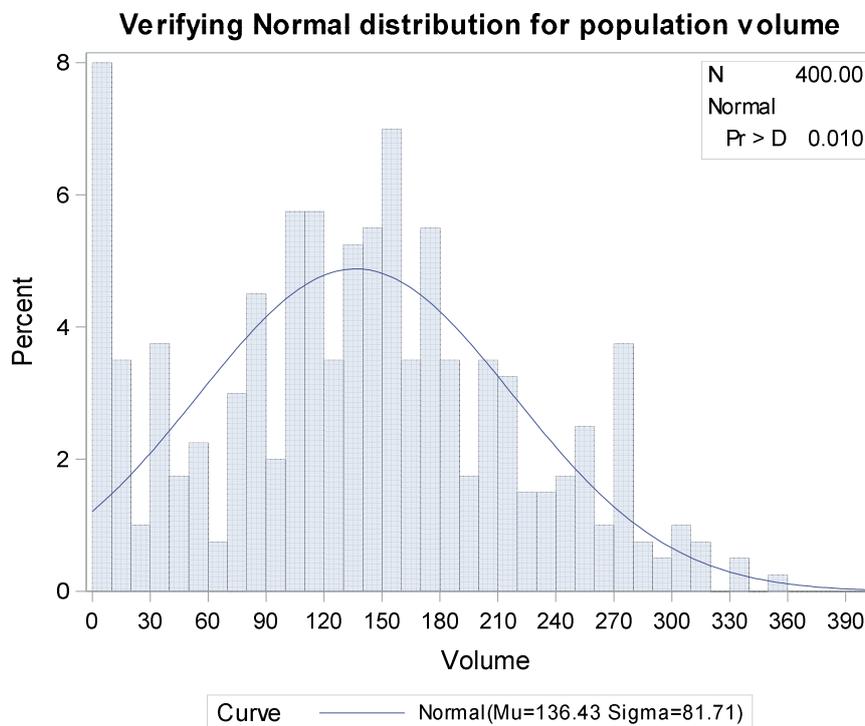
/*****item "a" Caso Florestal 1*****/
title "Verifying Normal distribution for population volume";
ods select Histogram ParameterEstimates GoodnessOfFit;

proc univariate data=Loetsch normaltest;
  var volume;
  histogram volume / normal odstitle = title;
  inset n normal(ksdpval) / pos = ne format = 6.3;
run;

```

O resultado no PROC UNIVARIATE é apresentado no Output 13 contendo o Histograma de frequência com a distribuição Normal hipotética. Verifica-se que a variável “volume” não se ajusta a uma distribuição Normal de acordo ao teste de Kolmogorov-Smirnov ($P > D = 0,010$).

Output 13. Distribuição de frequência para o volume da população Loetsch. Shapiro-Wilk: $p < 0,0001$. As tabelas com demais resultados foram omitidas intencionalmente.



É válido notar que existem situações em que a não normalidade de uma variável tende a ocorrer como:

- i) Restrições sobre os valores da variável sob análise: Caso em que a variável é restrita a valores positivos ou negativos e variável numérica discreta.
- ii) Quando a variável sob análise é resultante de outras duas variáveis: No caso de uma variável dependente (ou resposta) for resultante de um quociente. Isso acontece, por exemplo para uma variável de produtividade de frutos de castanha do Brasil que resulta da divisão do número de frutos pela área de floresta avaliada.

Para resolver o item “b” do caso florestal 1 vamos considerar o PROC SURVEYSELECT com a opção REPS= que faz réplicas do tamanho amostral informando em SAMPSIZE=. A sintaxe para extrair 500 réplicas é a seguinte:

```
title "500 samples of n=10 each";
proc surveysselect data=loetsch
    method=srs
    seed=1234
    sampsize=10
    reps=500
    out=sample_n10_500_reps;
run;

title "500 samples of n=30 each";
proc surveysselect data=loetsch
    method=srs
    seed=1234
    sampsize=30
    reps=500
    out=sample_n30_500_reps;
run;

title "500 samples of n=100 each";
proc surveysselect data=loetsch
    method=srs
    seed=1234
    sampsize=100
    reps=500
    out=sample_n100_500_reps;
run;

title "Printing the 15 first observations";
proc print data= sample_n10_500_reps (obs=15);
run;
```

A sintaxe indicada gera um arquivo com tabela SAS nomeado como “sample_n10_500_reps” contendo 500 amostras (REPS=500) de tamanho 10 cada (SAMPSIZE=10) extraídas da população “loetsch” de forma aleatória simples (SRS). Portanto, o tamanho total da amostra será de $10 \cdot 500=5000$.

A tabela SAS gerada com réplicas de tamanho $n=10$ é apresentada no Output 14 com as 15 primeiras observações (Total de 5000 observações) impressa pelo PROC PRINT. Para identificar cada uma das 500 réplicas (neste caso, de tamanho $n=10$), o SAS cria uma nova variável denominada “Replicate”.

Output 14. Primeiras 15 observações da amostragem aleatória simples com 500 réplicas de tamanho $n=10$.

Obs	Replicate	Zone	Site	y	x	Coordenates	Stratum	Volume
1	1	I	A	5	d	5d	E2	153
2	1	II	A	5	g	5g	E2	130
3	1	IV	B	6	q	6q	E1	71
4	1	IV	B	6	r	6r	E1	6
5	1	I	C	12	a	12a	E3	224
6	1	II	C	12	f	12f	E3	183
7	1	III	C	15	k	15k	E2	153
8	1	IV	C	15	p	15p	E2	147
9	1	III	D	20	n	20n	E1	0
10	1	III	D	20	o	20o	E1	0
11	2	IV	A	2	t	2t	E1	47
12	2	III	A	3	n	3n	E1	6
13	2	III	B	6	l	6l	E2	171
14	2	I	B	7	b	7b	E3	200
15	2	II	B	9	h	9h	E3	306

Para responder o item “c” do caso florestal 1 será necessário calcular a média aritmética de cada amostra replicada. No SAS esse cálculo é realizado utilizando o procedimento PROC MEANS. Esse procedimento calculará a média de cada uma das réplicas gerando um arquivo de 500 médias (observações) calculadas a partir de $n=10$; $n=30$ e $n=100$.

Juntamente com o PROC MEANS, será solicitado o teste de aderência para a distribuição Normal e o histograma utilizando o PROC UNIVARIATE. A sintaxe com os dois

procedimentos utilizados para o tamanho $n=10$ é apresentada a seguir (Os demais não são apresentados a fim de diminuir espaço):

```
/******item "c" Caso Florestal 1******/

title "500 samples of n=10 each";
proc surveysselect data=loetsch2
    method=srs
    seed=1234
    sampsize=10
    reps=500
    out=sample_n10_500_reps;
run;

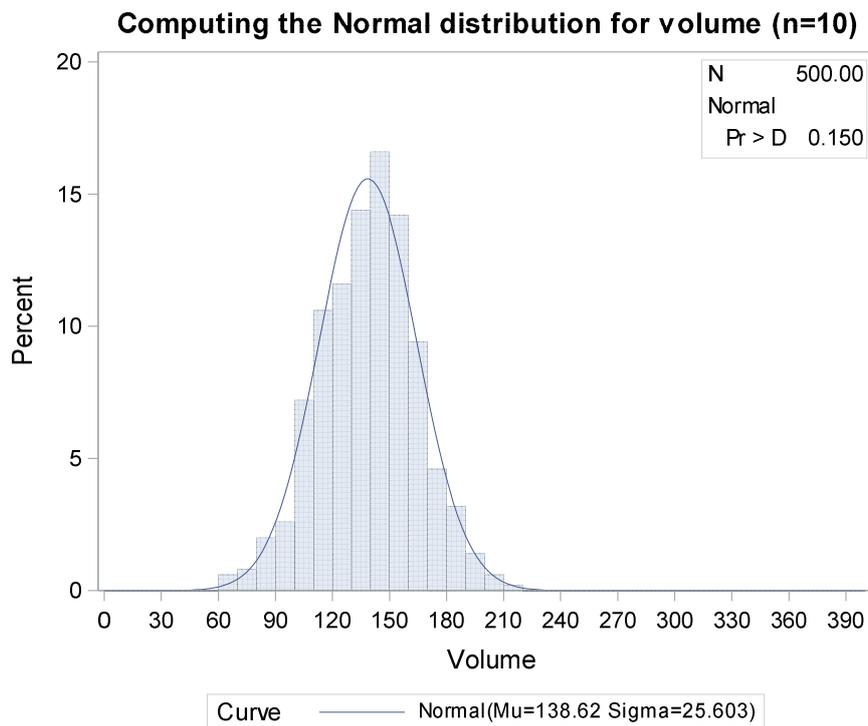
title 'Computing the mean of each 500 reps of n=10';
proc means data=sample_n10_500_reps noprint;
    var volume;
    by replicate;
    output out=medias_n10_500_reps mean=mean_vol_n10_500_reps;
run;

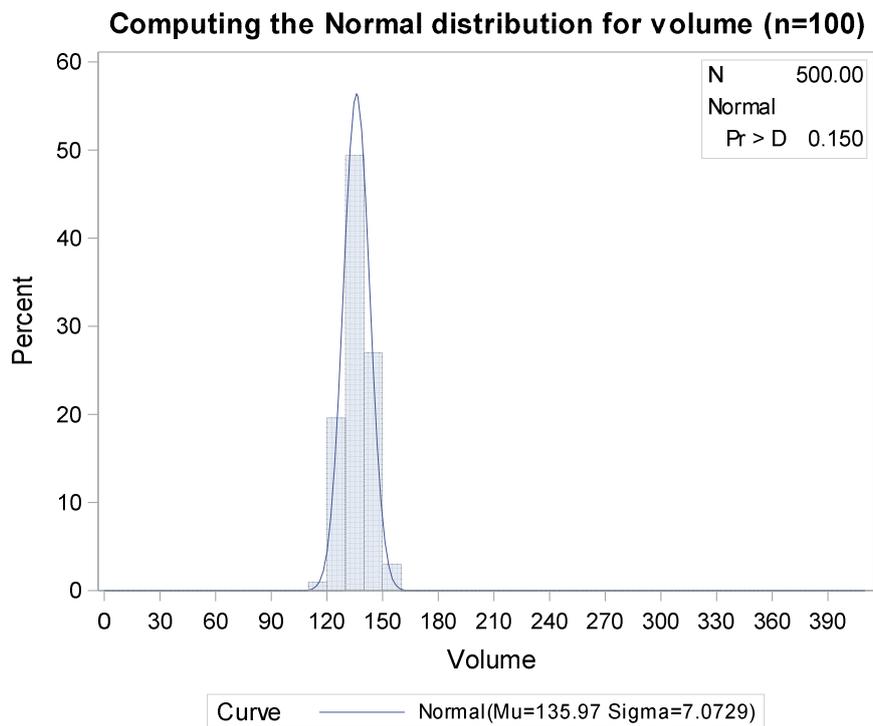
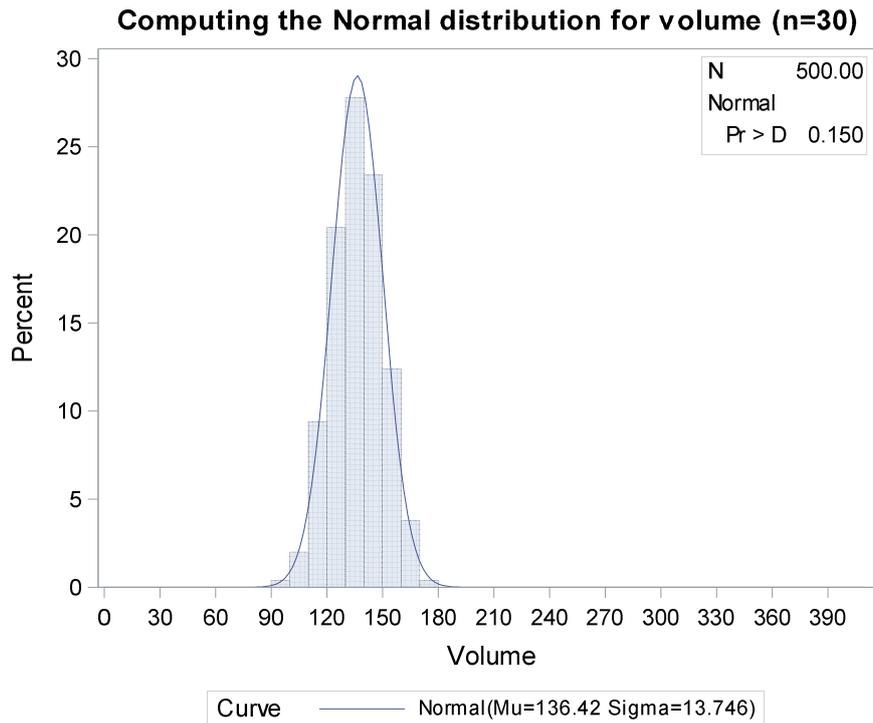
title1 'Computing the Normal distribution for volume (n=10)';
title2 '500 replication of n=10';
proc univariate data=medias_n10_500_reps normaltest;
    var mean_vol_n10_500_reps;
    histogram mean_vol_n10_500_reps / normal endpoints = 0 to 400 by 10
odstitle=title;
    inset n normal(ksdpval) / pos=ne format=6.3;
run;
```

A sintaxe do PROC MEANS inclui a opção OUTPUT OUT= para gerar uma tabela SAS (medias_n10_500_reps) contendo uma nova variável denominada “mean_vol_n10_500_reps” que representa a média aritmética da variável “volume” para cada réplica (REPLICATE).

Os resultados são apresentados no Output 15 contendo apenas o gráfico de histograma de frequência da distribuição Normal hipotética juntamente com os dados observados para as médias.

Output 15. Distribuição de 500 médias aritméticas (N) calculadas a partir de amostras cada uma de tamanho $n=10$, $n=30$ e $n=100$ extraídas da população Loetsch. Kolmogorov-Smirnov: $p>D=0.150$.





Os gráficos do Output 15 mostram claramente que todas as amostras extraídas da população Loetsch apresentam uma distribuição Normal, independentemente do tamanho amostral considerado para as réplicas.

Ademais, observa-se que o aumento do tamanho da amostra (n) levou à diminuição do desvio padrão de $S=25,6 \text{ m}^3$ com $n=10$ para $S=7,0 \text{ m}^3$ para $n=100$, refletindo em maior precisão (Para amostras de tamanho $n=100$ a frequência de dados tornou a distribuição mais estreita).

2.4.9. Cálculo do tamanho da amostra

Independente da análise estatística a ser realizada, determinar o tamanho adequado da amostra é etapa importante de se realizar no planejamento da pesquisa visto que, ter a variabilidade da população presente nas amostras (Representatividade) consiste no fundamento básico para realizar inferências sobre a população.

Por razões econômicas, o tamanho da amostra é importante considerando que uma amostra superdimensionada (muitas repetições, por exemplo) irá utilizar mais recursos econômicos do que realmente é necessário e expor a unidade experimental (árvores, pessoas, animais) a riscos enquanto um número reduzido de amostras pode desperdiçar recursos por não ter a capacidade de detectar o efeito verdadeiro dos tratamentos.

Basicamente, para determinar o tamanho da amostra, primeiramente deve-se decidir qual o tipo de desenho de estudo que será conduzido na pesquisa bem como a quantidade de médias que serão avaliadas.

2.4.9.1. Tamanho da amostra: Uma média

Durante o planejamento da coleta de dados para pesquisa, um passo importante para determinar o tamanho da amostra (n) a fim de estimar a média de uma população é estabelecer a probabilidade de confiança ($1 - \alpha$) e a margem de erro (ME).

Além da ME, é necessário conhecer a priori uma medida de variabilidade como a variância, erro padrão ou o coeficiente de variação. Essas informações podem ser obtidas a partir de valores de referência, quando disponível, ou a partir de uma amostra piloto.

O número de amostra piloto a considerar deve seguir uma condição estatística a ser satisfeita, ou seja, o número de unidades de amostra deverá ser igual a no mínimo 5% do tamanho da população $N: n \geq 0,05 \cdot N$ (Para população finita).

A ME ou diferença máxima entre a média amostral e a média populacional, deve ser estabelecida de acordo ao nível de exigência para o estudo. O Cálculo da ME é realizado com a seguinte expressão matemática:

$$ME = t_{\alpha/2;n-1} \frac{s}{\sqrt{n}}$$

Em que:

t =valor “ t ” obtido na tabela com $n-1$ graus de liberdade e nível de significância $\alpha/2$;

$\frac{s}{\sqrt{n}}$ = erro padrão (s = desvio padrão, n =tamanho da amostra).

Observa-se que a partir da ME é possível determinar o tamanho da amostra (n):

$$\sqrt{n} = \frac{t_{\alpha/2;n-1} s}{ME}$$

Logo, ambos os termos ao quadrado resultam em:

$$n = \frac{(t_{\alpha/2;n-1})^2 s^2}{ME^2}$$

O tamanho da amostra (n), neste caso, é calculado de acordo com a margem de erro estabelecida e determinado nível de significância. Esse cálculo pode ser realizado considerando o coeficiente de variação (CV) e a porcentagem de erro (PE):

$$n = \frac{(t_{\alpha/2;n-1})^2 cv^2}{PE^2}$$

O processo para determinar o tamanho da amostra é iterativo, isto é, a expressão matemática para determinar “ n ” é aplicada repetidas vezes variando o valor de “ t ” correspondente ao número de amostra calculado na interação 1. Logo, esse processo é realizado até que o valor de “ n ” se estabilize assintoticamente para uma solução de tamanho da amostra (Processo de interação convergir).

Para calcular o tamanho da amostra considerando uma probabilidade de confiança de 95%, Stauffer (1982) recomenda utilizar um valor inicial para $t=2$ no caso em que os valores para o Coeficiente de Variação (CV) e Porcentagem de Erro (PE) sejam arbitrários ou imprecisos. Desta forma, o tamanho da amostra (n) necessária para estimar a média dentro de uma margem de erro estabelecida é de, aproximadamente:

$$n = \frac{4 cv^2}{PE^2}$$

Vale destacar que, o cálculo do tamanho da amostra pode resultar em uma manipulação estatística ineficiente em casos que a estimativa da variância seja deficiente (FREESE, 1967).

Para fins de aplicação do cálculo do tamanho da amostra, vamos considerar a amostragem aleatória simples realizada na população Loetsch da secção 2.4.5.1. Neste caso, vamos considerar a amostra de 30 parcelas (unidades amostrais) como piloto para estimar a variância e verificar se é necessário amostrar mais unidades do que as 30 consideradas no exemplo.

As informações disponíveis para a população Loetsch a partir das 30 parcelas amostradas inicialmente são as seguintes: média estimada do volume=118,6 m³/0,1ha; desvio padrão=78,5837 m³/0,1ha. Portanto, considerando que se deseja estimar a média com uma margem de erro de 2 m³/ha, o tamanho da amostra a ser extraída da população pelo processo de amostragem aleatório simples é calculado com 95% de probabilidade de confiança da seguinte maneira:

$$n_0 = \frac{(t_{\alpha/2;n-1})^2 s^2}{ME^2} = \frac{2,045^2 \cdot 7,85837^2}{2^2} = \frac{258,2567}{4} = 64,5642 \approx 65$$

$$n_1 = \frac{(t_{\alpha/2;n-1})^2 s^2}{ME^2} = \frac{1,998^2 \cdot 7,85837^2}{2^2} = \frac{246,5221}{4} = 61,6305 \approx 62$$

$$n_2 = \frac{(t_{\alpha/2;n-1})^2 s^2}{ME^2} = \frac{2,00^2 \cdot 7,85837^2}{2^2} = \frac{247,0159}{4} = 61,7540 \approx 62$$

Neste caso, o tamanho da amostra calculado na interação zero (n_0) foi igual a 65 parcelas. Substituindo o valor de “t” na equação considerando 65-1 graus de liberdade, a interação 1 resultou em $n_1=62$ parcelas. Observe que na interação 2 o valor de n variou dentro do valor de 62 parcelas (arredondamento para cima) e, portanto, convergindo.

Portanto, para estimar a média seria necessário realizar um levantamento de 32 parcelas a mais do que as 30 parcelas já consideradas na amostragem inicial.

Um detalhe importante de se notar é a necessidade de manter a mesma unidade e dimensão para a variância e o erro admitido na fórmula do cálculo do tamanho da amostra. Observe que o desvio padrão obtido na amostragem inicial equivale a 78,5837 m³/0,1ha e foi ajustado para a mesma dimensão do erro admitido (hectare).

Outra aplicação do cálculo de “ n ” é solicitada no Caso Florestal 2.

Caso florestal 2: Determinação do número de toras para estudo do Coeficiente de Rendimento Volumétrico (CRV) para madeiras tropicais em serrarias.

Considere que uma pesquisa será conduzida para determinar o rendimento de madeira serrada a partir do desdobro de toras de *Dipteryx odorata* (Cumaru-ferro) em uma Serraria. A pesquisa deu-se devido atualmente o limite estabelecido do CRV para espécies nativas da Amazônia é de 35% (Estabelecido por órgãos competentes) e que este pode variar de acordo com a espécie, maquinário, bitola de peças, etc.

O Coeficiente de Variação (CV) determinado em estudos anteriores para a mesma espécie, maquinário semelhante, peças padronizadas e diâmetro menor da tora variando entre 70 a 150 cm foi de 39%.

Neste sentido, deve-se determinar quantas toras devem ser utilizadas para o desdobro se o pesquisador deseja ter 95% de confiança que a média amostral esteja a menos/mais de 10% da média do CRV.

Aplicando a fórmula temos:

$$n_0 = \frac{4 cv^2}{PE^2} = \frac{4 * 39^2}{10^2} = \frac{6084,0000}{100} = 60,8400 \approx 61 \text{ toras}$$

Neste caso, observe que para a próxima interação (n_1), o valor de “ t ” para 61-1 graus de liberdade ($t=2$) resultará no mesmo valor para n_0 .

O tamanho da amostra para duas ou mais médias pode ser calculado para algumas análises estatísticas específicas. Essa técnica é denominada de Poder da Análise (Power Analysis) e será abordada no próximo tópico.

2.4.10. Poder da Análise (Power Analysis)

O processo de cálculo do tamanho da amostra aleatória simples (capítulo anterior) é comumente empregado em inventário florestal diagnóstico para determinar o número de unidades amostrais considerando valores de entrada, a margem de erro, variância esperada e valor t de student (o Z em algumas condições).

Por outro lado, quando se deseja determinar o tamanho da amostra na condução de um delineamento experimental, por exemplo, e aplicar teste específico para comparar as

médias (Teste F, Teste de Mann-Whitney, Wilcoxon, outros) é necessário considerar outra técnica denominada de Poder da Análise (Ou Power Analysis, em inglês).

Portanto, trata-se de uma solução analítica muito eficiente que permite obter uma estimativa do número de amostras necessárias para algumas análises estatísticas como uma simples comparação de médias entre duas populações (Teste *t* pareado ou não pareado) ou o tamanho da amostra necessário para obter um determinado valor de coeficiente de determinação (R^2) para uma análise de Regressão Linear, análise de Correlação (*r*), número de repetições em uma ANOVA etc. Neste caso, o processo de determinação do tamanho da amostra (*n*) é calculado considerando que seja possível detectar um efeito significativo se ele realmente existir.

A expressão a seguir representa, de forma geral, a proporcionalidade (\propto) do Poder da Análise com relação a outros componentes:

$$Poder \propto \frac{TE \cdot \alpha \cdot \sqrt{n}}{s}$$

Em que:

Poder=Poder do teste;

TE=Tamanho do efeito a detectar (Tamanho efetivo);

α =Nível de significância;

n =Tamanho da amostra;

S=Desvio padrão.

2.4.10.1. Nível de significância

O nível de significância (α) é a probabilidade de rejeitar a hipótese nula quando esta for verdadeira, ou seja, concluir que existe diferença entre as duas médias quando na verdade são iguais (Falso positivo). Portanto, α é referenciado como o erro Tipo I em estatística.

Geralmente esse valor é estabelecido em 5% que resulta em uma probabilidade de confiança de $1-\alpha=95\%$ (Decisão correta).

2.4.10.2. Poder do teste e Tamanho da amostra

O poder do teste é a probabilidade (valor entre 0 e 1) de rejeitar a hipótese nula quando a hipótese alternativa é verdadeira (Decisão correta). Matematicamente, pode ser

escrito como $1 - \beta$, em que β representa o erro do Tipo II em teste de hipótese (Falso negativo).

Por outro lado, o tamanho da amostra é o número total de unidades amostrais necessárias para conduzir um estudo seja observacional ou experimental. Para estudo experimental, o tamanho da amostra será a soma do número de repetições por tratamento para um determinado delineamento experimental.

Valores para $1 - \beta$ de 0,8 a 0,9 (80 a 90%) são amplamente considerados nas pesquisas científicas o que resulta em 80% ou 90% de certeza que pelo menos um tratamento será diferente dos demais analisados.

A relação entre o número de amostras (n) e o poder do teste é positiva apresentando uma assíntota quando o poder do teste alcança valor de 1 (100%). Neste ponto, o aumento de " n " não proporciona aumento no poder do teste conforme observado na Figura 19 que representa o número de repetições necessárias para diferentes níveis de poder do teste (0,8; 0,9 e 0,99).

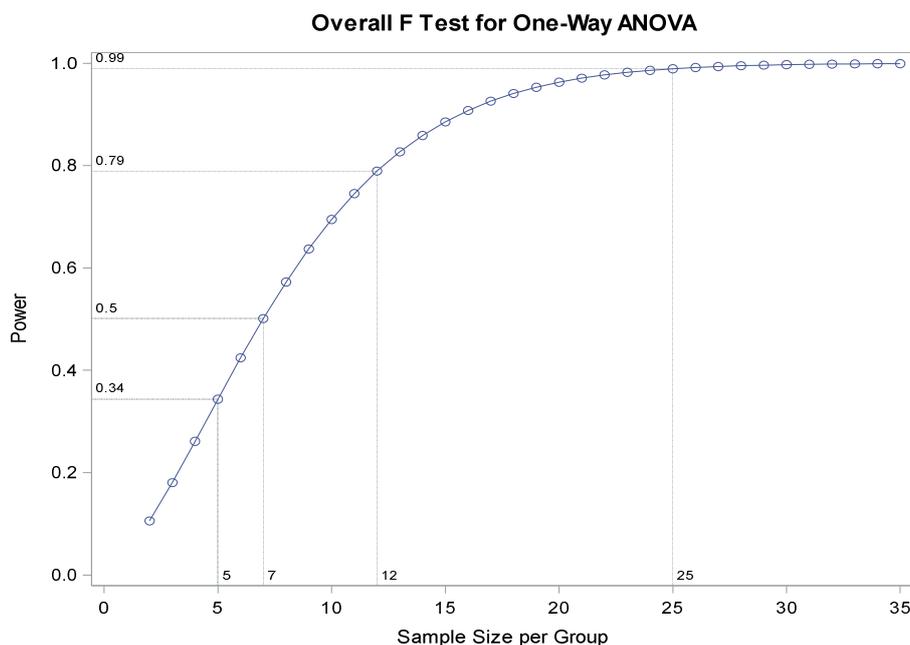


Figura 19. Evolução do número de repetições por grupo (Sample Size per Group) necessárias em função do poder do teste (Power) estabelecido em uma simulação de Análise de Variância para um delineamento com quatro tratamentos ($\mu_1=41$; $\mu_2=22$; $\mu_3=29$; $\mu_4=28$ com $\alpha=5\%$ e $s=14$).

A partir da Figura 19 observa-se que, caso o delineamento seja conduzido com 5 repetições a pesquisa estará sujeita a não captar um efeito significativo em 66% das vezes ($1-\beta=$). Com 7 repetições ambas probabilidades para cometer o erro Tipo II (Falso negativo) e de tomar a decisão correta são iguais (50%).

Por outro lado, com 25 repetições (tamanho da amostra = 100 unidades experimentais) há 99% de probabilidade de detectar um efeito verdadeiro quando realmente ele existir (Um tratamento com efeito significativo).

Portanto, considerar apenas 5 repetições no delineamento simulado não são suficientes, mas por outro lado, 25 repetições podem causar uma inflação excessiva no orçamento da pesquisa.

Para as condições estabelecidas no experimento (Simulação), alcançar um poder do teste de 80% (Um dos valores mais recomendados) é necessário considerar 12 repetições.

Geralmente o valor de $\beta=20\%$ é o mais utilizado em pesquisas das Ciências Agrárias implicando que se existir um efeito verdadeiro, este não será observado em 20% das vezes. Diante do exposto surge a seguinte pergunta: Um poder do teste de 80% é suficiente considerando que sobra 20% de chance de cometer o erro Tipo II na pesquisa?

Para responder essa pergunta é necessário aprofundar mais sobre os tipos de erros que estamos sujeitos a cometer em análise estatística (Erro Tipo I e erro Tipo II) em teste de hipótese.

Neste sentido, considere o Quadro 12 onde se resume os erros associados a teste de hipótese em estatística.

Quadro 12. Tipos de erros estatísticos associados a teste de hipótese. Observa-se que o poder do teste ($1-\beta$) é inversamente proporcional ao erro Tipo II (β).

Decisão baseada na amostra	Realidade	
	Grupos são iguais (h_0 verdadeira)	Grupos são diferentes (h_1 verdadeira)
Grupos são iguais → Não rejeitar h_0 .	Decisão correta ($1-\alpha$)	Erro Tipo II (β) Falso negativo
Grupos são diferentes → Rejeitar h_0 .	Erro Tipo I (α) Falso positivo	Decisão correta ($1-\beta$)

O Quadro 12 mostra que quando a tomada de decisão baseada na amostra diverge da realidade (Desconhecida) ocorre os tipos de erro durante o teste de hipótese. Neste caso, não rejeitar h_0 quando na verdade deveria ser rejeitada, estamos cometendo o erro Tipo II (β), ou seja, h_0 é falsa, mas o teste de hipótese na pesquisa considerando a amostra leva a sua não-rejeição.

Neste caso, a conclusão da pesquisa é de que não existe diferença entre as duas médias, por exemplo, quando na verdade estas são diferentes (conhecido como falso negativo).

O Quadro 13 mostra as probabilidades calculadas para cada tipo de erro bem como a probabilidade para a tomada de decisão correta considerando um cenário de realidade em que a hipótese nula e a hipótese alternativa sejam verdadeiras em uma probabilidade prévia de 50%.

Quadro 13. Erros estatísticos associados a teste de hipótese e resultados de probabilidades para tomada de decisão. Para a simulação, considerou-se um nível de significância $\alpha=5\%$ e um poder do teste de $1-\beta=80\%$.

Decisão baseada na amostra	Realidade	
	Grupos são iguais (h_0 verdadeira) 1/2	Grupos são diferentes (h_1 verdadeira) 1/2
Grupos são iguais → Não rejeitar h_0 . (Efeito não significativo) $1-\alpha=95\%$; $\beta=20\%$	Decisão correta ($1-\alpha$) $95\%/2=47,5\%$	Erro Tipo II (β) $20\%/2=10\%$ Falso negativo
Grupos são diferentes → Rejeitar h_0 . (Efeito significativo) $\alpha=5\%$; $1-\beta=80\%$	Erro Tipo I (α) $5\%/2=2,5\%$ Falso positivo	Decisão correta ($1-\beta$) $80\%/2=40\%$

De acordo ao estabelecido no Quadro 13, o resultado mostra que as probabilidades de tomar uma decisão correta ou de cometer um dos tipos de erro são as seguintes:

- 2,5% de probabilidade de cometer o Erro tipo I (Falso positivo);
- 10% de probabilidade de cometer o Erro tipo II (Falso negativo);
- 40% de probabilidade de encontrar um efeito significativo (Existe diferença entre as médias, por exemplo); e
- 47,5% de probabilidade de encontrar um efeito não-significativo (Não existe diferença entre as médias, por exemplo).

Observa-se que mesmo estabelecendo um poder do teste de 80%, a situação mais provável resulta em 47,5% de probabilidade de não encontrar nenhum efeito significativo o que é frustrante em um resultado de pesquisa. Se aumentar o poder do teste para $1 - \beta = 99\%$ mantendo a probabilidade de confiança em 95% teríamos os seguintes resultados:

- 2,5% de probabilidade de cometer o Erro tipo I (Falso positivo);
- 0,5% de probabilidade de cometer o Erro tipo II (Falso negativo);
- 49,5% de probabilidade de encontrar um efeito significativo (existe diferença entre as médias); e
- 47,5% de probabilidade de encontrar um efeito não-significativo (não existe diferença entre as médias).

O resultado mostra que a probabilidade de detectar um efeito verdadeiro se esse realmente existir de fato aumentou de 40% para 49,5%. Entretanto, observa-se que o aumento no poder do teste foi muito maior (15%) do que a resultante de probabilidade para a decisão correta (9,5%). Ademais, esse esforço resultaria em um aumento significativo do número de amostras como observado a relação positiva entre poder do teste e tamanho da amostra na Figura 19.

Se o grupo de pesquisa decidir diminuir a probabilidade de cometer o Erro tipo I de 5% para 1% resultaria em 99% de probabilidade de confiança. Os resultados mantendo as condições de poder do teste $1 - \beta = 80\%$ são:

- 0,5% de probabilidade de cometer o Erro tipo I (Falso positivo);
- 10% de probabilidade de cometer o Erro tipo II (Falso negativo);
- 40% de probabilidade de encontrar um efeito significativo (existe diferença entre as médias); e

- 49,5% de probabilidade de encontrar um efeito não-significativo (não existe diferença entre as médias).

Observa-se que a mudança não influenciou a probabilidade de detectar efeito significativo que permaneceu em 40% contra 49,5% de não existir diferença entre as médias.

Caso o grupo de pesquisa esteja ciente com 70% de probabilidade de que a hipótese alternativa (H_1) seja verdadeira no início da coleta de dados, ou seja, executou um bom planejamento, o resultado da simulação mostra que a probabilidade de encontrar um efeito significativo aumenta para 56% contra 28,5% para efeito não significativo.

É possível mudar os valores para os tipos de erros comumente considerados em pesquisas florestais, mas a decisão deve ser baseada, em parte, na importância de se cometer um dos erros para a inferência a depender do objetivo da pesquisa.

Dependendo da área de concentração como saúde, por exemplo, um desses dois tipos de erros bem como seus níveis pode ser mais grave.

Neste caso, Neyman e Pearson (1933) relataram a seguinte frase:

“É mais grave condenar um inocente ou absolver um culpado?”

Neste caso, se um inocente é condenado e essa pessoa não fez nada, mas estamos inferindo que essa pessoa fez alguma coisa, então isso é um erro do Tipo I. Por outro lado, se um culpado é absolvido, essa pessoa fez alguma coisa errada, mas estamos inferindo que nenhum crime foi cometido, então esse é um erro do Tipo II.

2.4.10.3. Tamanho efetivo

O Tamanho Efetivo ou Tamanho de Efeito (d) é calculado considerando a média de dois grupos dividido pelo desvio padrão. Portanto, informa o número mínimo de desvio padrão de diferença entre os grupos de interesse que podem ser detectados.

O desvio padrão pode ser estimado a partir de uma amostra piloto ou considerando uso de valores de referência a depender do objeto de estudo tendo em vista a literatura existente.

Jacob Cohen propôs valores de referência equivalentes a 0,2; 0,5 e 0,8 a serem utilizados para representar efeitos pequeno, médio e grande, respectivamente (Cohen, 1988). Neste caso, esses valores são utilizados quando não for possível o cálculo de d .

O tamanho efetivo (d) e o tamanho da amostra (n) são inversamente proporcionais para um mesmo valor de nível de confiança e poder do teste. A Figura 20 mostra a mudança do tamanho da amostra e do poder do teste para diferentes níveis de Tamanho Efetivo considerando valores de referência propostos por Jacob Cohen em uma simulação considerando um teste t para duas médias.

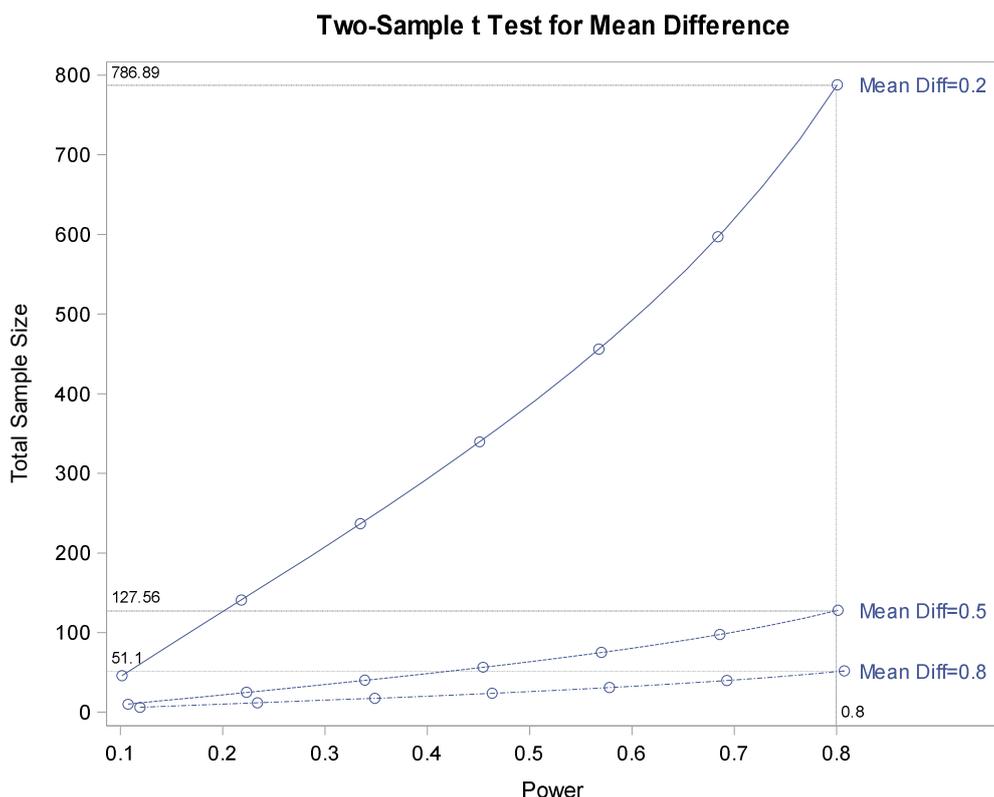


Figura 20. Evolução do número de amostras necessárias para um poder de teste de 80% para cada um dos valores de referência do Tamanho Efetivo (Mean Diff/1) para um teste t de comparação de duas médias. Para a simulação considerou-se um desvio padrão $S=1$.

A Figura 20 mostra um aumento significativo no tamanho da amostra à medida que o tamanho do efeito a ser detectado diminui. Para um tamanho de efeito pequeno (0,2) a grande (0,8) existe uma diferença de 736 unidades amostrais. Portanto, considerando um poder do teste de 80%, será necessário um total de 52 unidades amostrais (duas amostras de 26 unidades para cada grupo).

2.4.10.4. Aplicação da Análise de Poder no SAS Studio

Em vez de usar fórmulas matemáticas longas e complicadas combinando os componentes para calcular o tamanho da amostra é possível utilizar softwares estatísticos para tal. Neste caso, o SAS possui dois procedimentos específicos o PROC POWER e o PROC GLMPOWER.

A sintaxe padrão do PROC POWER depende de especificar o tipo de análise que se deseja calcular, o tamanho da amostra ou poder do teste (ou outro componente) conforme a seguir:

```
proc power;  
    <tipo_de_análise> <opções>;  
    plot <opções> / <opções_de_gráficos>;  
run;
```

De forma a simplificar o processo de determinação do tamanho da amostra, poder do teste ou tamanho efetivo no PROC POWER os seguintes passos devem ser considerados:

Passo 1: Identificar o tipo de análise estatística

Para calcular o tamanho da amostra, poder do teste ou tamanho efetivo no PROC POWER é necessário indicar o tipo de análise estatística a ser realizada na linha de programação <tipo_de_análise>.

Portanto, é de suma importância saber a priori qual a análise estatística será conduzida para responder os objetivos da pesquisa e em seguida considerar qual a sintaxe será utilizada. O Quadro 14 indica os tipos de análise estatística suportados pelo PROC POWER.

Quadro 14. Tipos de análises e o correspondente código a informar na sintaxe do PROC POWER.

Tipo de análise estatística	Código no PROC POWER
Testes <i>t</i>	ONESAMPLEMEANS, PAIREDMEANS, TWOSAMPLEMEANS
Intervalo de confiança para médias	ONESAMPLEMEANS, PAIREDMEANS, TWOSAMPLEMEANS
Testes de proporções	ONESAMPLEFREQ, PAIREDFREQ, TWOSAMPLEFREQ
ANOVA (Um fator)	ONEWAYANOVA
Teste Wilcoxon Mann-Whitney	TWOSAMPLEWILCOXON
Correlação	ONECORR, TWOCORR
Regressão logística	LOGISTIC
Regressão linear múltipla	MULTREG
Análise de sobrevivência	TWOSAMPLESURVIVAL

Passo 2: Formular as hipóteses nula e alternativa

O teste de hipótese é um procedimento estatístico que permite tomar uma decisão sobre a população utilizando os dados observados a partir de uma amostra. Também pode ser definido como uma regra que especifica se deve rejeitar ou não uma alegação sobre uma população de acordo com as provas fornecidas por uma amostra de dados.

A decisão ou alegação sobre a população é formulada mediante as hipóteses nula (H_0) e hipótese alternativa (H_1) que são avaliadas considerando um nível de significância estabelecido (α).

A formulação de uma hipótese nula e alternativa é baseada no objetivo da pesquisa e, conseqüentemente, determinará se um teste será unilateral ou bilateral. Esse conhecimento prévio é necessário de ser informado na sintaxe do PROC POWER. No Quadro 15 são apresentadas algumas formulações de acordo à análise e objetivo da pesquisa.

Quadro 15. Alguns exemplos de objetivos e suas respectivas hipóteses para algumas análises estatísticas.

Objetivo da pesquisa	Tipo de análise estatística	Hipóteses
Determinar se existe uma associação positiva entre o diâmetro do colo e altura de mudas de <i>Cedrela odorata</i> .	Análise de correlação.	$H_0: r=0$ $H_1: r>0$
Avaliar se a produção de frutos de castanha do Brasil para um grupo de árvores é maior do que a média regional de 230 frutos.	Teste t uma média.	$H_0: m=230$ $H_1: m>230$
Avaliar se o incremento em diâmetro de um grupo de árvores remanescentes de <i>Swietenia macrophylla</i> é diferente quando se realiza a liberação da competição de árvores vizinhas em floresta natural.	Teste t duas médias independentes.	$H_0: m_1 - m_2 = 0$ $H_1: m_1 - m_2 \neq 0$
Avaliar se o incremento em diâmetro de um grupo de árvores remanescentes de <i>Swietenia macrophylla</i> é diferente quando submetidas a três diferentes níveis de liberação da competição de árvores vizinhas em floresta natural.	Teste F três médias.	$H_0: m_1 - m_2 - m_3 = 0$ $H_1: \text{Pelo menos um par de médias são diferentes}$

Passo 3: Especificar o componente a ser determinado

Por meio do PROC POWER é possível informar três componentes a fim de calcular o quarto. Neste caso, para calcular o tamanho da amostra basta informar o valor para o nível de significância, poder do teste e o tamanho efetivo. Por outro lado, caso se deseje determinar o poder do teste basta informar o valor para o nível de significância, tamanho da amostra e o tamanho efetivo.

A seguir a sintaxe do PROC POWER para um teste F para diferença entre quatro médias:

```
proc power;
  onewayanova
  test=overall
  groupmeans=(41 22 29 28)
  stddev=.
  power=.
  npergroup= .
  alpha=0.05;
run;
```

O usuário pode resolver qualquer um dos componentes da sintaxe indicados anteriormente com valor missing “.” mantendo o restante com valores. Portanto, caso se deseje determinar o poder do teste (POWER) basta substituir o valor missing dos componentes STDDEV e NTOTAL por valores desejados na sintaxe do PROC POWER conforme demonstrado a seguir:

```
proc power;
  onewayanova
  test=overall
  groupmeans=(41 22 29 28)
  stddev=14
  power=.
  alpha=0.05
  npergroup= 6;
run;
```

A opção TEST=OVERAL indica que a análise será realizada considerando o teste de F para diferenças entre dois ou mais grupos. Na opção GROUPMEANS= é informado o valor das médias esperadas para cada um dos grupos (Tratamentos).

Logo, na opção STDDEV= indica-se o valor do desvio padrão. Esse valor deve ser obtido de uma amostra piloto ou de referências de estudos anteriores. A opção NPERGROUP= quando informado um valor missing, retorna o valor para o número de repetições.

Os resultados da análise de poder são apresentados no Output 16.

Output 16. Determinação do poder do teste para um delineamento com valores simulados.

Fixed Scenario Elements	
Method	Exact
Alpha	0.05
Group Means	41 22 29 28
Standard Deviation	14
Sample Size per Group	6

Computed Power	
Power	
	0.424

Neste caso, para as condições do delineamento o poder do teste foi de 0,424. Certamente o elevado valor para o desvio padrão ($s=14$) colaborou para o baixo poder do teste.

Para determinar o número de repetições (Sample Size per Group) para um poder do teste de 80% basta especificar o valor 0,8 em POWER e deixar NPERGROUP= com valor missing conforme a seguinte sintaxe adicionada de um gráfico de tamanho da amostra versus poder do teste:

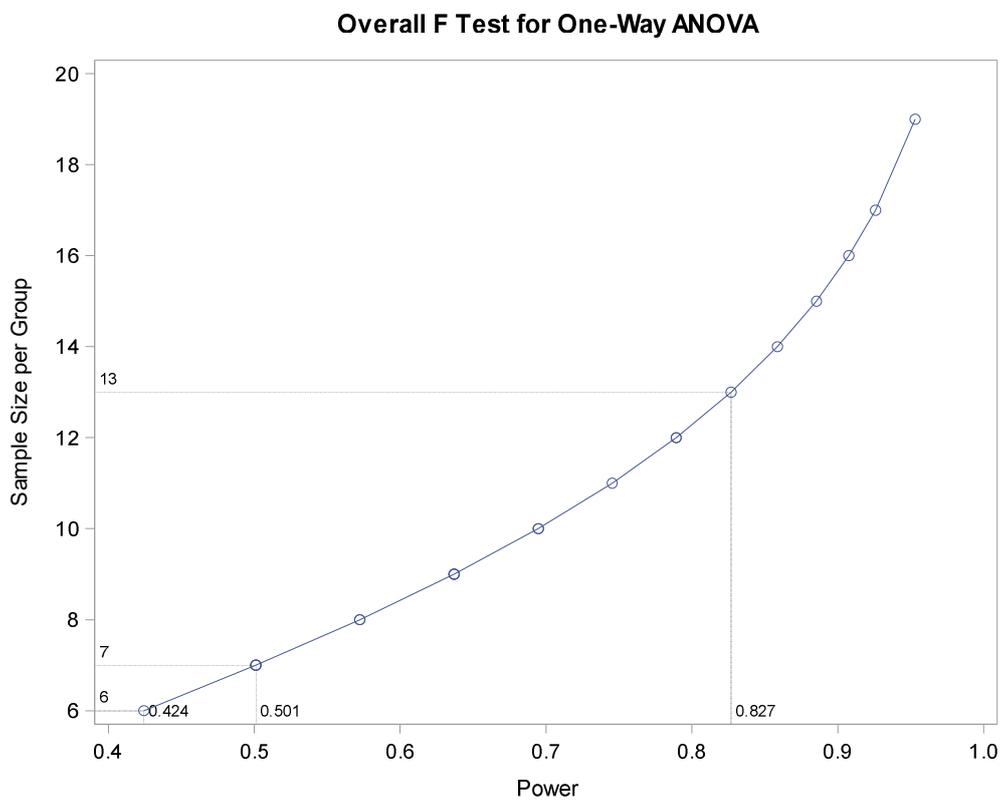
```
proc power;
  onewayanova
  test=overall
  groupmeans=(41 22 29 28)
  stddev=14
  power=0.8
  alpha=0.05
  npergroup= .;
  plot x=power min=0.4 max=0.95 yopts=(ref=6 7 13 crossref=yes);
run;
```

Neste caso, a sintaxe solicita um gráfico na opção PLOT contendo no eixo x os valores de poder do teste variando de 0,4 a 0,95. Ademais, solicita a determinação do poder do teste para tamanho da amostra variando de 6, 7 e 13 pela opção YOPTS. Os resultados da análise de poder são apresentados no Output 17.

Output 17. Determinação do número de repetições para um poder do teste de 80% em um delineamento com valores simulados.

Fixed Scenario Elements	
Method	Exact
Alpha	0.05
Group Means	41 22 29 28
Standard Deviation	14
Nominal Power	0.8

Computed N per Group	
Actual Power	N per Group
0.827	13



Observa-se que são necessárias 13 repetições para conduzir um delineamento experimental que tenha como expectativa os valores de entrada.

2.4.10.5. Utilizando o Task and Utilities do SAS Studio

É possível calcular qualquer um dos componentes utilizando a opção de point and click do SAS por meio da ferramenta Task and Utilities. Neste caso, vamos considerar o seguinte Caso Florestal.

Caso florestal 3: Determinação do tamanho da amostra utilizando a ferramenta Task and Utilities do SAS Studio.

Uma empresa madeireira deseja exportar peças de madeira para utilização em pisos do tipo Deck. Entretanto, a empresa deseja avaliar se a dureza Janka da madeira da nova espécie apresenta padrão internacional de dureza estabelecido para esse tipo de piso. Em média a dureza Janka para esta finalidade deve ser de 3500 libras. Portanto, para realizar a exportação, compradores do mercado internacional solicitaram que um teste estatístico seja realizado.

As informações técnicas repassadas aos pesquisadores é que a dureza média medida na nova espécie não deve apresentar uma diferença para mais ou para menos do que 250 libras comparado à média estabelecida sendo que o desvio padrão apresentado em estudo piloto foi de 1.035 libras.

Ademais, foi requerido que o teste seja realizado com um poder de pelo menos 0.8 e um nível de significância de 0.05.

Para responder o caso florestal 3 será necessário considerar o teste t para uma média amostral a fim de comparar com a média padrão estabelecida. Esse procedimento é facilmente realizado no SAS Studio utilizando o Painel de Navegação a aba **Task and Utilities** opção **Power and Sample Size** e logo, clicar em **t Tests** conforme descreve a Figura 21.

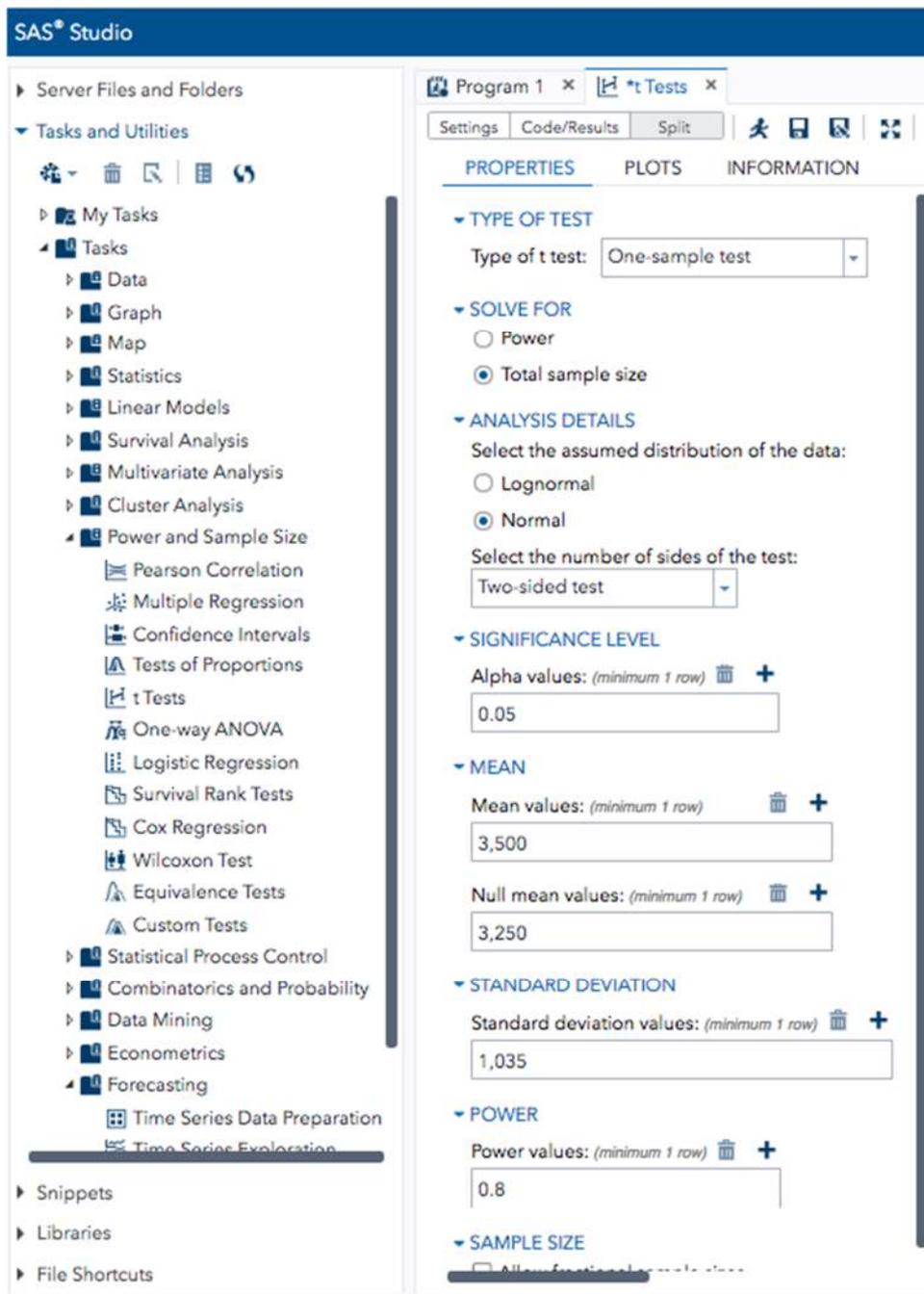


Figura 21. Task and Utilities para determinação do tamanho da amostra para o Teste t de uma média no SAS Studio.

- 1) A opção “**Type of test**” deve selecionar a opção **One-sample test**;
- 2) Em “**Solve for**” marque a opção **Total sample size**;
- 3) Selecione a opção **Normal** para a distribuição dos dados;
- 4) O nível de significância deve ser selecionado na opção **Significant level**.
Neste caso, utilizaremos 5%;

- 5) Nesta janela é necessário digitar a média populacional (mean value) que será comparada com a média hipotética (Null mean value). No caso do exemplo utilizaremos a média considerada como padrão para a dureza Janka de 3.500 libras;
- 6) Nesta janela, deve-se digitar o valor do desvio padrão determinado para o estudo;
- 7) Aqui deve-se indicar o poder do teste que no exemplo foi solicitado como 80%.

O poder do teste pode ser interpretado como a probabilidade de se obter um resultado estatisticamente significativo quando existe uma diferença verdadeira entre os tratamentos. Geralmente utiliza-se no mínimo 80% para assegurar uma alta probabilidade de se observar o efeito do tratamento.

Após preencher os valores basta solicitar o processamento e o SAS irá apresentar os resultados conforme o Output 18.

Output 18. Tamanho da amostra necessária para atender ao solicitado no caso florestal 3.

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Number of Sides	2
Null Mean	3250
Alpha	0.05
Mean	3500
Standard Deviation	1035
Nominal Power	0.8

Computed N Total	
Actual Power	N Total
0.802	137

O relatório produzido indica que, para as condições solicitadas, deve-se considerar 137 unidades amostrais (peças de madeira) para conduzir o experimento.

Vale ressaltar que a amostra calculada deve ser grande o suficiente para que a probabilidade de encontrar diferenças entre os grupos por mero acaso seja baixa e para que a probabilidade de se detectar diferenças verdadeiras e significantes seja alta.

A opção do Task and Utilities funciona considerando uma programação do PROC POWER “por trás” do point and click. Portanto, ao mesmo tempo que se preenche os valores, a seguinte sintaxe é criada de forma automática na aba CODE:

```
proc power;
  onesamplemeans test=t sides=2
  mean=3500
  nullmean=3250
  stddev=1035
  power=0.8
  alpha=0.05
  ntotal=.;
run;
```

3. Análise de regressão

Em análise de associação bivariada entre variáveis, existem diferentes estratégias de análise dos dados a considerar de acordo com a classificação da variável dependente (resposta) e da variável independente (preditora).

Neste capítulo, a análise de regressão será abordada, considerando-se a possibilidade de análise de variáveis qualitativas (Categóricas) e quantitativas de acordo com as possibilidades indicadas no Quadro 16.

Quadro 16. Tipos de análises estatísticas a realizar de acordo com a classificação da variável.

Variável dependente \ Variável independente	Qualitativa	Quantitativa	Quantitativa e Qualitativa
Quantitativa	Análise de Variância (ANOVA)	Regressão Linear ou Não-linear	Análise de Covariância (ANCOVA)
Qualitativa	Tabela de Contingência ou Regressão Logística	Regressão Logística	Regressão Logística

Em cada possibilidade de análise de regressão, um tópico específico será abordado neste capítulo.

3.1. Regressão Linear

Objetivos de aprendizagem desse capítulo:

- i) Descrever a diferença entre os tipos de regressão linear simples e múltipla;
- ii) Demonstrar os cálculos necessários para o ajuste de modelos de regressão linear;
- iii) Mostrar a utilidade da análise de regressão para análise exploratória e predição.

Uma regressão é dita linear quando os coeficientes de regressão (Betas) se apresentam na forma aditiva ou subtrativa e elevados ao expoente unitário (Schneider et al., 2009). O modelo de regressão a seguir é um exemplo de regressão linear, mesmo sendo um modelo de polinômio do segundo grau (Curvilnear):

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

Neste caso os coeficientes de regressão Betas ($\beta_0 + \beta_1 x_i + \beta_2 x_i^2$) estão na forma aditiva e elevados a expoente unitário.

Ademais, ao calcular a derivada do modelo com respeito aos coeficientes de regressão os resultados não envolvem coeficientes de regressão, ou seja, o modelo é linear. As derivadas do polinômio do segundo grau para cada coeficiente de regressão são:

- Derivada de y_i em relação a $\beta_0 \rightarrow \frac{\partial y}{\partial \beta_0} = 1$
- Derivada de y_i em relação a $\beta_1 \rightarrow \frac{\partial y}{\partial \beta_1} = x_i$
- Derivada de y_i em relação a $\beta_2 \rightarrow \frac{\partial y}{\partial \beta_2} = x_i^2$

Caso o resultado das derivadas de um modelo de regressão qualquer seja dependente de coeficientes de regressão, o modelo é não-linear.

Portanto, em estatística um modelo é considerado linear ou não-linear de acordo com a forma com que os coeficientes de regressão são inseridos na função do modelo. Alguns modelos em que sua representação gráfica aparece côncavo ou convexo são de fato modelos lineares como por exemplo o modelo do polinômio do segundo grau acima.

A análise de regressão linear tem por finalidade explicar a variação de uma variável dependente (y) utilizando um grupo de variáveis independentes (x 's) por meio de uma expressão matemática. A quantidade de variáveis independentes utilizadas no modelo estatístico indica se é uma regressão linear simples (apenas uma variável x) ou múltipla (mais de uma variável x no modelo).

Alguns exemplos de utilização da análise de regressão no campo florestal compreendem:

- Estimar o volume de árvores (variável dependente) a partir da medição do diâmetro a altura do peito e altura total (variáveis preditoras);
- Explorar o perfil do tronco de árvores por meio de modelos de afilamento (tapering) com o objetivo de estimar o sortimento de madeira;

- Estudar o comportamento do crescimento em diâmetro, altura e volume ao longo de toda a vida da árvore. Neste caso, em geral utilizam-se modelos não-lineares como o modelo biológico de Chapman-Richards.

3.1.1. Regressão Linear Simples

Considera no modelo estatístico apenas uma variável independente (preditora). O objetivo dessa análise é:

- Avaliar a significância da variável independente na explicação da variabilidade ou comportamento da variável dependente;
- Prevê valores da variável dependente dado os valores da variável independente.

Como exemplo, podemos estudar a associação entre a altura total e diâmetro a altura do peito (Relação hipsométrica) de uma amostra de árvores, o modelo de regressão linear a utilizar tem a seguinte equação:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Em que:

$i = 1, 2, 3, \dots$, unidades amostrais aonde se realizam as medições de “y” e “x” (árvores);

y_i = altura total (m) observada na i-ésima árvore;

β_0 = coeficiente intercepto, significa a altura total média da população de árvores;

x_i = variável independente observada na i-ésima árvore. No caso da pesquisa, representa o diâmetro a altura do peito (cm);

β_1 = coeficiente angular, que significa a mudança que ocorre na altura das árvores devido ao incremento em uma unidade de diâmetro (x_i);

ε_i = Resíduos, efeito aleatório associado com a i-ésima observação \sim NIID (0, σ^2).

Para cumprir com algumas exigências do modelo linear, os resíduos devem assumir uma distribuição normal (N), serem independentes (I) e identicamente distribuídos (ID=cada um dos resíduos se distribuem com o mesmo parâmetro de média zero e variância σ^2). Esses pressupostos serão discutidos em capítulos posteriores.

Na Figura 22 podemos observar a dispersão dos valores de diâmetro a altura do peito (d) e altura total (h) de árvores em uma determinada floresta nativa.

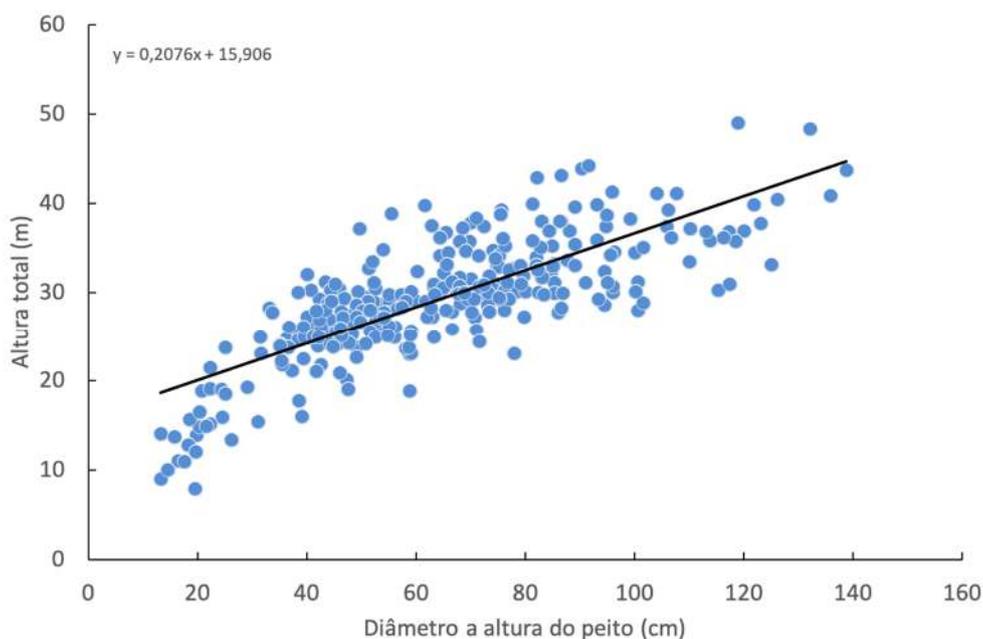


Figura 22. Valores observados e estimados pela regressão linear simples para altura total em função do diâmetro de árvores de uma determinada floresta nativa na Amazônia.

3.1.2. Regressão Linear Múltipla

Os resultados de uma análise de regressão linear simples muitas vezes são restritos para o uso na prática, pois o modelo considera que toda variação da variável dependente (y) pode ser explicada, em parte, pela única variável independente (x). No campo florestal muitas vezes isso não acontece!

Para o exemplo da Figura 22, o modelo linear simples considera que para um determinado valor de diâmetro (d), existe apenas um valor de altura total (h). Entretanto, não é o que acontece em florestas visto que a altura total pode variar entre indivíduos com o mesmo valor de diâmetro.

Neste caso, para aumentar a acurácia das estimativas de h e modelar uma relação que considere outras variáveis independentes, podemos utilizar mais do que uma variável no modelo de regressão. Em outras palavras, utilizaríamos uma regressão linear múltipla.

Logo, a regressão linear múltipla considera no modelo estatístico duas ou mais variáveis independentes e tem como principal objetivo:

- Predição: Desenvolver um modelo para prever valores presentes ou futuros de uma variável dependente (y) baseado na relação com outras variáveis independentes (x 's);
- Análise Exploratória: Desenvolver um entendimento da relação entre a variável dependente e as variáveis independentes, sendo muito utilizado para entender o “efeito biológico” de algumas variáveis dendrométricas sobre uma resposta de árvores ou povoamentos florestais.

Neste caso, o modelo de regressão linear tem a seguinte equação matemática:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i$$

Em que:

$i = 1, 2, 3, \dots$, unidades amostrais aonde se realizam as medições (p.e. árvores);

y_i = variável dependente observada na i -ésima unidade amostral;

β_0 = coeficiente intercepto, significa o valor médio da variável dependente da população;

x_{1i} = valor da variável x_1 observada na i -ésima unidade amostral;

β_1 = coeficiente angular, significa mudança que ocorre na variável dependente devido ao incremento em uma unidade da variável x_1 dado x_2 constante;

x_{2i} = valor da variável x_2 observada na i -ésima unidade amostral;

β_2 = coeficiente angular, significa mudança que ocorre na variável dependente devido ao incremento em uma unidade da variável x_2 dado x_1 constante;

ε_i = Resíduos, efeito aleatório associado com a i -ésima observação \sim NIID $(0, \sigma^2)$.

A regressão linear múltipla tem a grande vantagem de possibilitar a investigação da relação entre y e diversas variáveis independentes simultaneamente. Entretanto, por ser um modelo mais complexo, torna-se mais difícil estabelecer qual modelo é o melhor bem como interpretar os modelos.

Aplicação na Predição – Os termos no modelo, os valores dos coeficientes de regressão e a significância estatística são de importância secundária. Neste caso, o foco está em obter um modelo final com os melhores resultados de bondade de ajuste para prever valores futuros de y_i em função de valores de x 's. É importante salientar que a aplicação desses modelos se dá para conjunto de dados grande o suficiente para dividi-lo em dados de treino, validação e teste, ou seja, o modelo será ajustado com dados de treino, logo seu desempenho é avaliado em novos casos (dados de validação). Portanto, para este

objetivo, o modelo de regressão linear múltiplo concentra o foco nos valores estimados da variável dependente conforme indicado com traço baixo na seguinte equação:

$$\underline{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i$$

Aplicação na Análise Exploratória – O foco está no entendimento da relação entre a variável dependente y e as variáveis independentes x 's. Conseqüentemente, a significância estatística, a magnitude e os sinais dos coeficientes de regressão (+ ou -) são importantes. Para este objetivo, os coeficientes do modelo de regressão são o foco da pesquisa como indicado com um traço baixo na seguinte equação:

$$y_i = \underline{\beta}_0 + \underline{\beta}_1 x_{1i} + \underline{\beta}_2 x_{2i} + \dots + \underline{\beta}_p x_{pi} + \varepsilon_i$$

Na Tabela 1 são apresentados alguns modelos de regressão consolidados na área florestal para descrever o volume de árvores considerando a combinação de duas variáveis dendrométricas.

Tabela 1. Modelos de regressão linear simples e múltipla utilizados na ciência florestal para representar dados de volumetria de árvores a partir de cubagem rigorosa.

Autor	Modelo
1. Kopezky - Gehrhardt	$v_i = \beta_0 + \beta_1 D_i^2 + \varepsilon_i$
2. Dissescu - Meyer	$v_i = \beta_1 d_i + \beta_2 D_i^2 + \varepsilon_i$
3. Hohenadl - Krenn	$v_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2 + \varepsilon_i$
4. Brenac	$\ln v_i = \beta_0 + \beta_1 \ln D_i + \beta_2 \frac{1}{D_i} + \varepsilon_i$
5. Spurr. (var. Combinada)	$v_i = \beta_0 + \beta_1 D_i H_i + \varepsilon_i$
6. Stöate	$v_i = \beta_0 + \beta_1 D_i^2 + \beta_2 H_i + \beta_3 D_i^2 H_i + \varepsilon_i$
7. Näslund	$v_i = \beta_1 d_i^2 + \beta_2 D_i^2 H_i + \beta_3 D_i H_i^2 + \beta_4 H_i^2 + \varepsilon_i$
8. Spurr. logaritmo	$\ln v_i = \beta_0 + \beta_1 \ln D_i^2 H_i + \varepsilon_i$
9. Schumacher-Hall	$\ln v_i = \beta_0 + \beta_1 \ln D_i + \beta_2 \ln H_i + \varepsilon_i$

Em que: i =corresponde à i -ésima árvore; v_i = volume individual com casca (m^3); D_i = diâmetro a altura do peito (cm); H_i = altura total (m); $\beta_0; \beta_1, \dots, \beta_p$ = coeficientes de regressão; ε_i = variação não explicada pela regressão \sim NIID (0, σ^2); \ln = logaritmo natural de base e .

Fonte: LOETSCH et al. (1973).

Dentro de regressão linear múltipla existe um tipo específico de regressão que são os modelos polinomiais, nos quais potências e/ou interação da variável independente são incluídas no modelo como variáveis independentes. Alguns exemplos de modelos de regressão polinomial são apresentados a seguir.

- Modelo do polinômio do segundo grau:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

- Modelo do polinômio do terceiro grau:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$$

- Modelo do polinômio com interação:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i$$

Na Tabela 1 o modelo 3 (Hohenadl - Krenn) é um caso de regressão polinomial do segundo grau e o modelo 6 (Stöate) um caso de polinômio com interação.

3.2. Avaliação preliminar para análise de regressão

Neste tópico serão abordadas técnicas para examinar a dispersão dos dados da variável dependente em função de cada uma das variáveis independentes para análise de regressão como forma de avaliação preliminar. Portanto, os objetivos de aprendizagem desse tópico são:

- i) Demonstrar a importância da avaliação prévia do comportamento dos dados observados da variável dependente e variáveis independentes para análise de regressão;
- ii) Apresentar os procedimentos SAS para construção de gráficos de dispersão;
- iii) Analisar o comportamento da dispersão dos dados observados para tomada de decisão;
- iv) Verificar e consolidar os dados para decidir qual modelo de regressão pode ser apropriado para representar os dados.

Realizar a análise exploratória dos dados antes de realizar qualquer ajuste de modelos de regressão é de suma importância pois possibilita verificar potenciais problemas durante a construção de um modelo de regressão como, por exemplo, presença de correlação significativa entre variáveis independentes (multicolinearidade) e presença de observações “estranhas” no conjunto de dados.

A Figura 23 mostra um resumo da sequência recomendada para a construção de modelos de regressão considerando várias etapas.

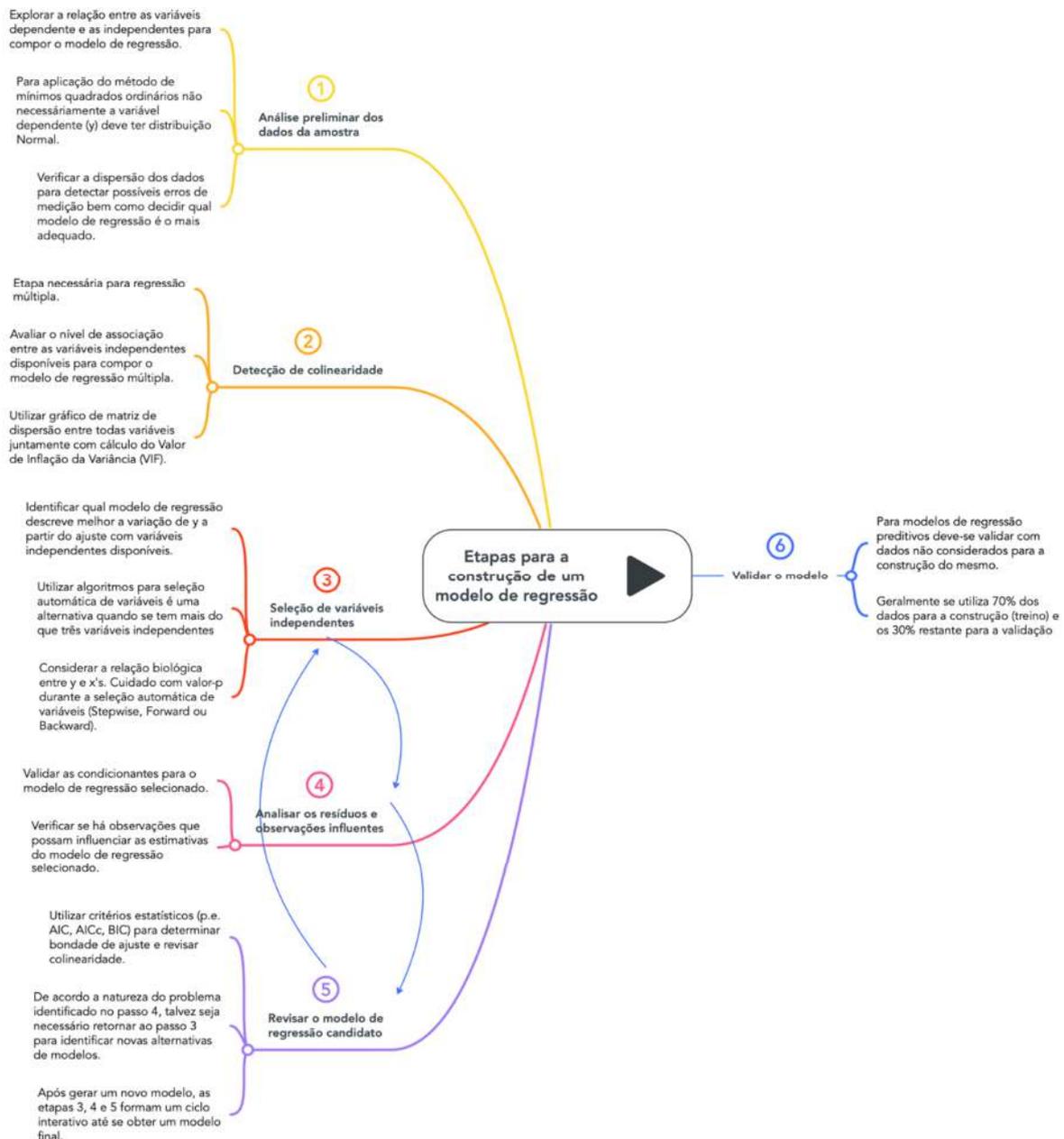


Figura 23. Passos a serem realizados para a construção de um modelo de regressão linear múltipla.

O uso de gráficos de dispersão é uma excelente opção para realizar a análise exploratória de variáveis para a análise de regressão. A construção de gráficos de dispersão pode ser facilmente operacionalizada no SAS System com o uso dos procedimentos PROC SGSCATTER e PROC SGPLOT além de outros procedimentos.

O procedimento PROC SGSCATTER cria painéis de gráficos de dispersão para várias combinações de variáveis de acordo com a necessidade. Esse procedimento possui

três declarações que podem ser utilizadas para obter gráficos de dispersão em painéis: COMPARE, MATRIX e PLOT. Exemplos da sintaxe simplificada do procedimento são:

- utilizando a declaração compare para criar um painel de gráficos de dispersão com eixos y's compartilhados para fins de comparação. É possível construir gráficos com mais de uma variável "x" e "y" bastando apenas declarar as variáveis dentro dos parênteses conforme o exemplo de aplicação resultando no gráfico da Figura 24:

```
proc sgscatter data=nome_do_dataset;
compare x=variável_x|(variável-1...Variável-n) y= variável_y|(variável-1...Variável-n)
/options;
run;

*/Sintaxe do procedimento com aplicação para dados de diâmetro a altura do peito (D) e
altura total (H) que serão plotados com a variável incremento*/

data sitio1;
input Narv Capoeira$ D H IPAg Hegyi;
datalines;
1 A 18.1 14.9 108.802 2.285
2 A 6.6 7 16.930 5.473
3 A 6.1 8 20.281 7.502
4 A 1.7 2.8 1.863 26.917
1 B 38.3 25 387.365 0.356
.
.
.
6 J 31.4 31 139.212 2.529
;
proc sgscatter data=sitio1;
compare x=(D H) y=ipag;
label D="Diâmetro a altura do peito (cm)"
H="Altura total (m)"
ipag="Incremento periódico anual em área transversal (cm^2)";
run;
```

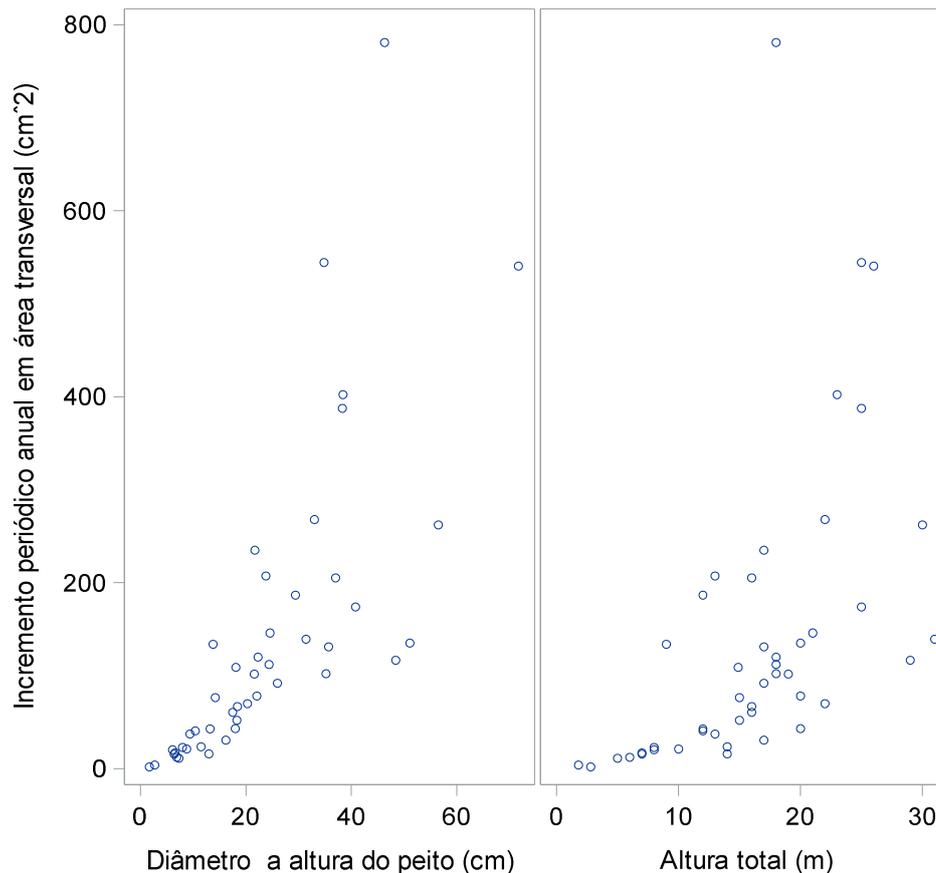


Figura 24. Painel de gráfico com eixo y compartilhado para duas variáveis. A opção LABEL foi adicionada para criar a legenda customizada para o nome das variáveis. Se esta opção não for utilizada, o SAS considera a etiqueta da variável como declarada na linha do INPUT.

- Utilizando a declaração MATRIX para criar uma matriz de gráficos de dispersão conforme o exemplo de aplicação resultando no gráfico da Figura 25. Também é possível construir matriz de dispersão com agrupamento dos dados bem como a inserção de elipse de predição (ELLIPSE) para cada par de variáveis para avaliar caso uma nova amostra seja realizada em uma população normal bivariada, se 95% dos dados estariam dentro da região da elipse no gráfico. O resultado é o gráfico da Figura 26:

```

proc sgscatter data=nome_do_dataset;
    matriz variável_numérica-1 variável_numérica-2 <...variável_numérica-n> /options;
run;

*/Sintaxe do procedimento com aplicação para dados de diâmetro (D) e altura (H) que serão
plotados com a variável incremento*/

data sitio1;
    input Narv Capoeira$ D H IPAg Hegyi;
    datalines;
1 A 18.1 14.9 108.802 2.285
2 A 6.6 7 16.930 5.473
3 A 6.1 8 20.281 7.502
4 A 1.7 2.8 1.863 26.917
1 B 38.3 25 387.365 0.356
.
.
.
6 J 31.4 31 139.212 2.529
;
proc sgscatter data=sitio1;
    matriz IPAg D H Hegyi;
run;

proc sgscatter data=sitio1;
    matriz IPAg D H Hegyi / group=capoeira start=topleft
    ellipse=(alpha=0.05 type=predicted) legend=(noborder title="capoeira"
    position=bottom);
run;

```

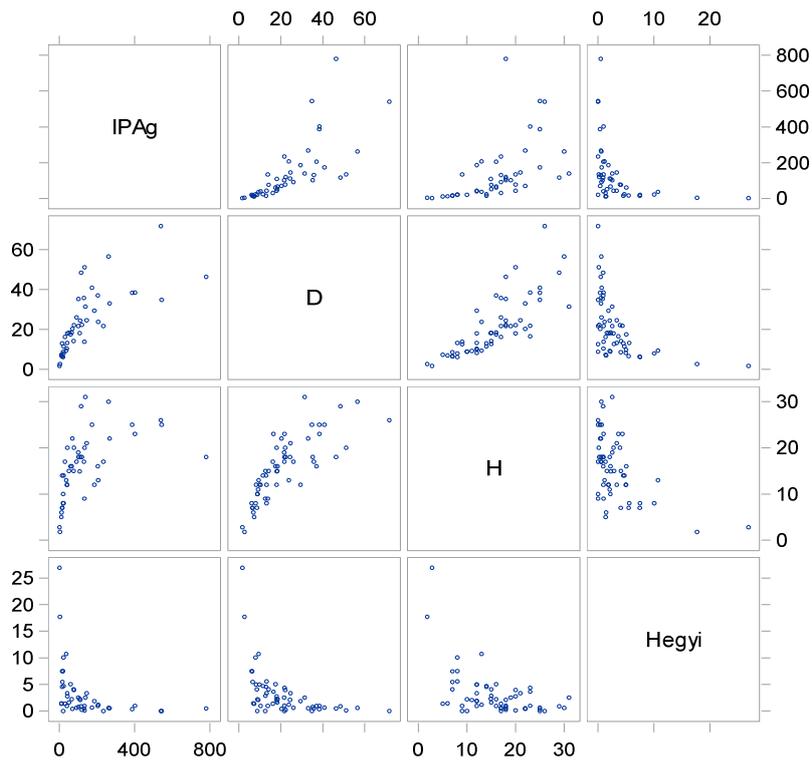


Figura 25. Matriz de dispersão para avaliar comportamento entre pares de variáveis.

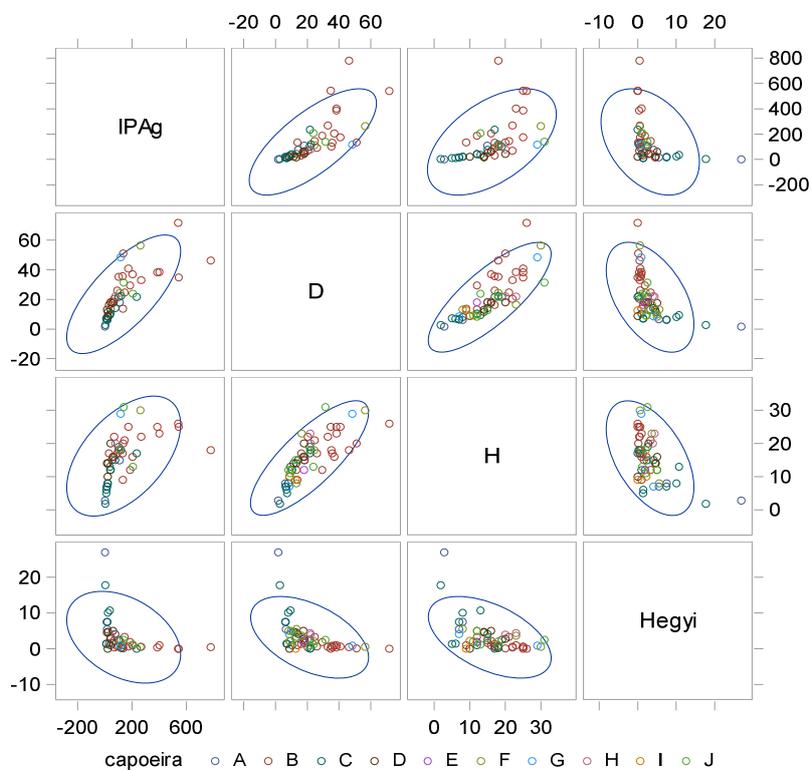


Figura 26. Matriz de dispersão para avaliar comportamento entre pares de variáveis agrupados (Capoeira) com inserção de elipse de predição.

- Sintaxe simplificada do SGSCATTER utilizando a declaração PLOT para criar um painel contendo gráficos de dispersão independentes conforme resultado do processamento na Figura 27:

```
proc sgscatter data=nome_do_dataset;
    plot (variável_numérica-1 variável_numérica-2 <...variável_numérica-n> )*
    (variável_numérica-3 variável_numérica-4 <...variável_numérica-n> ) /options;
run;
```

/Sintaxe do procedimento com aplicação para dados de diâmetro a altura do peito (D) e altura total (H) que serão plotados com a variável incremento/

```
data sitio1;
    input Narv Capoeira$ D H IPAg Hegyi;
    datalines;
1 A 18.1 14.9 108.802 2.285
2 A 6.6 7 16.930 5.473
3 A 6.1 8 20.281 7.502
4 A 1.7 2.8 1.863 26.917
1 B 38.3 25 387.365 0.356
.
.
.
6 J 31.4 31 139.212 2.529
;
proc sgscatter data=sitio1;
    plot (D H)*ipag;
    label D="Diâmetro a altura do peito (cm)"
        H="Altura total (m)"
        ipag="IPAg (cm^2)";
run;
```

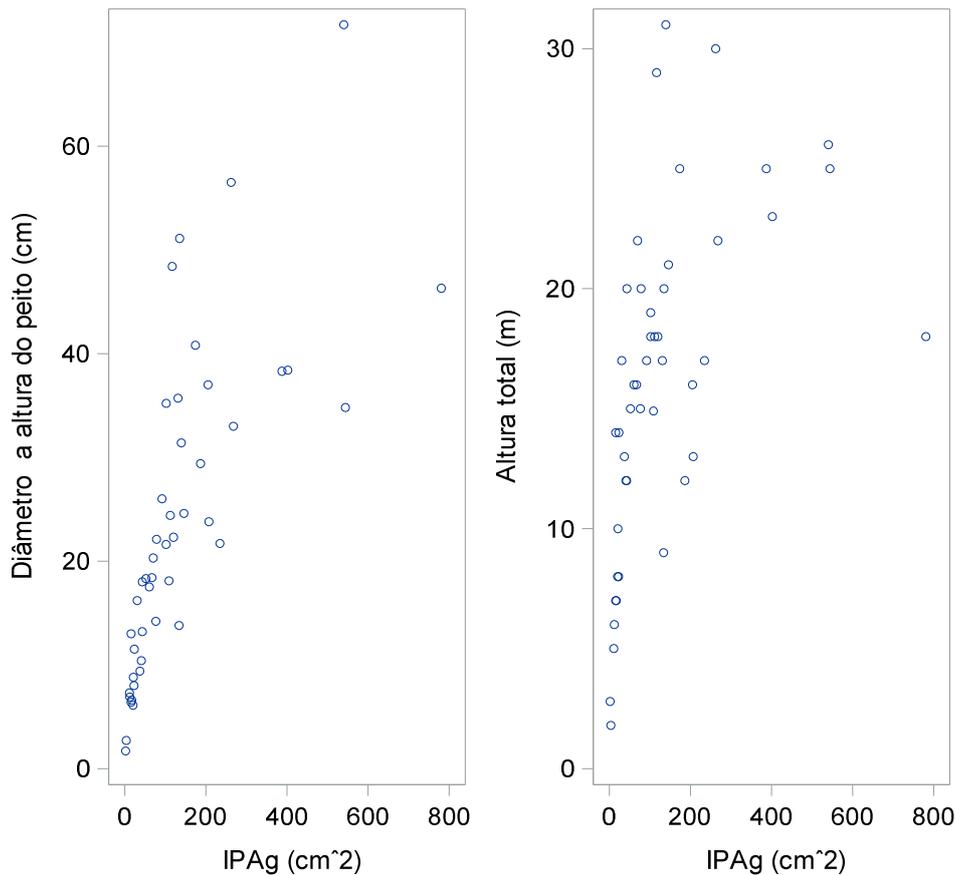


Figura 27. Painel de gráficos de dispersão.

Todas as três declarações possuem a possibilidade de edição gráfica com a inclusão de várias opções descritas a profundo e com exemplo nos manuais SAS.

O argumento `OPTIONS` é um exemplo que inclui vasta opções de gráfico no SAS. A exemplo a opção `GROUP=variável_categórica`, divide os valores da variável dependente (y) em grupos de acordo a variável categórica indicada.

A seguir um exemplo da aplicação da declaração `COMPARE` no `PROC SGSCATTER` considerando dados de Incremento Periódico Anual em área basal individual ($ipag$ no eixo y) comparando com o diâmetro a altura do peito (d), altura total (h) e o Índice de Competição de Hegyi ($hegyi$). O pesquisador ficou interessado em avaliar a dispersão dos dados agrupados para cada capoeira (floresta secundária) utilizando no argumento `GROUP` a variável categórica “Capoeira”. O resultado está na Figura 28.

```

data sitio1;
  input Narv Capoeira$ D H IPAg Hegyi;
  datalines;
1 A 18.1 14.9 108.802 2.285
2 A 6.6 7 16.930 5.473
3 A 6.1 8 20.281 7.502
4 A 1.7 2.8 1.863 26.917
1 B 38.3 25 387.365 0.356
.
.
.
6 J 31.4 31 139.212 2.529
;
proc sgscatter data=sitio1;
  compare x=(D H Hegyi) y= IPAg / group=capoeira;
  label ipag="IPAg (cm^2)";
run;

```

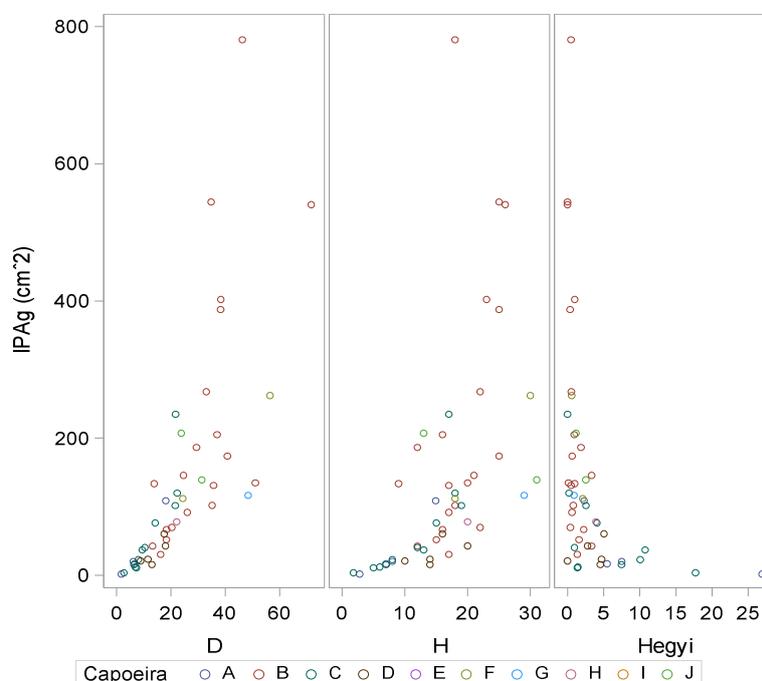


Figura 28. Painel de comparação entre a variável de Incremento Periódico Anual em área transversal (IPAg, cm^2), diâmetro a altura do peito (D , cm), altura total (H , m) e Índice de Competição de Hegyi agrupados por capoeiras (A , B , ..., J).

A análise das Figuras 3 e 4 também informam possível problema de colinearidade em um eventual modelo de regressão considerando as variáveis independentes em questão, visto que, alguns pares de variáveis apresentam associação entre si com as variáveis diâmetro com altura e altura com Hegyi.

Outro procedimento gráfico do SAS muito útil para análise exploratória de dados é o procedimento PROC SGPLOT. Por meio da declaração SCATTER o procedimento constrói gráficos de dispersão para avaliar o comportamento e a relação entre a variável dependente e a variável independente. Nesta fase, pode-se analisar visualmente se há a necessidade de se considerar uma regressão polinomial caso o comportamento dos dados seja curvilíneo. Para demonstração de uso do Procedimento vamos considerar o caso florestal 4.

Caso florestal 4: Uso do procedimento proc sgplot para análise exploratória

Considere que um pesquisador está interessado em estudar o efeito de um aditivo químico na resistência de um determinado papel utilizado para embalagens. Para tal, o pesquisador selecionou uma amostra representativa e mediu a quantidade de aditivo químico utilizado (variável independente) e a resistência ao rasgamento (variável dependente) representada pela quantidade de força necessária para rasgar o papel.

Para a análise exploratória desses dados utilizou-se o procedimento PROC SGPLOT considerando a declaração SCATTER para dispersão dos dados observados conforme sintaxe SAS a seguir:

```
data paper;
  input obs amount strength;
  datalines;
1 1 2.6
2 1 2.4
3 1 2.7
4 2 2.6
5 2 2.7
6 2 2.6
7 2 2.5
8 2 2.8
9 3 3
10 3 3
11 3 2.8
12 3 2.8
13 4 2.9
14 4 2.9
15 4 3
16 4 3.1
17 4 3
18 5 2.8
19 5 2.9
20 5 3
21 5 2.9
22 5 2.9
;
title "Verificando dispersão dos dados";
proc sgplot data=paper;
  scatter x=amount y=strength;
  label amount="Quantidade de aditivo químico" strength="Resistência";
run;
```

A Figura 29 indica que os dados de resistência do papel apresentam um comportamento curvilíneo à medida que aumenta a quantidade de produto químico e, portanto, o uso de uma regressão linear simples é inadequado para representar os dados.

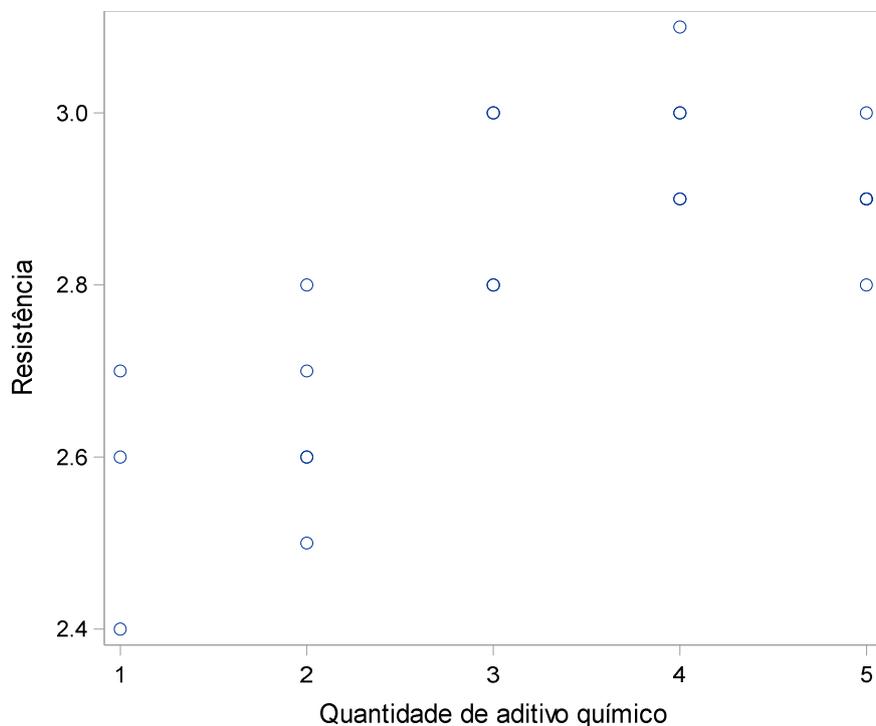


Figura 29. Tendência dos dados de resistência do papel em relação à quantidade de produto químico adicionado.

Ao solicitar a inclusão de uma linha de tendência linear aos dados no PROC SGPLOT, por meio da declaração REG, comprova-se a inadequação da regressão linear simples aos dados observados conforme indica a Figura 30.

```
title "Ajuste da reta aos dados";  
proc sgplot data=paper;  
  reg x=amount y=strength / lineattrs=(color=brown pattern=solid)  
  legendlabel="Linear";  
  label amount="Quantidade de aditivo químico" strength="Resistência";  
run;
```

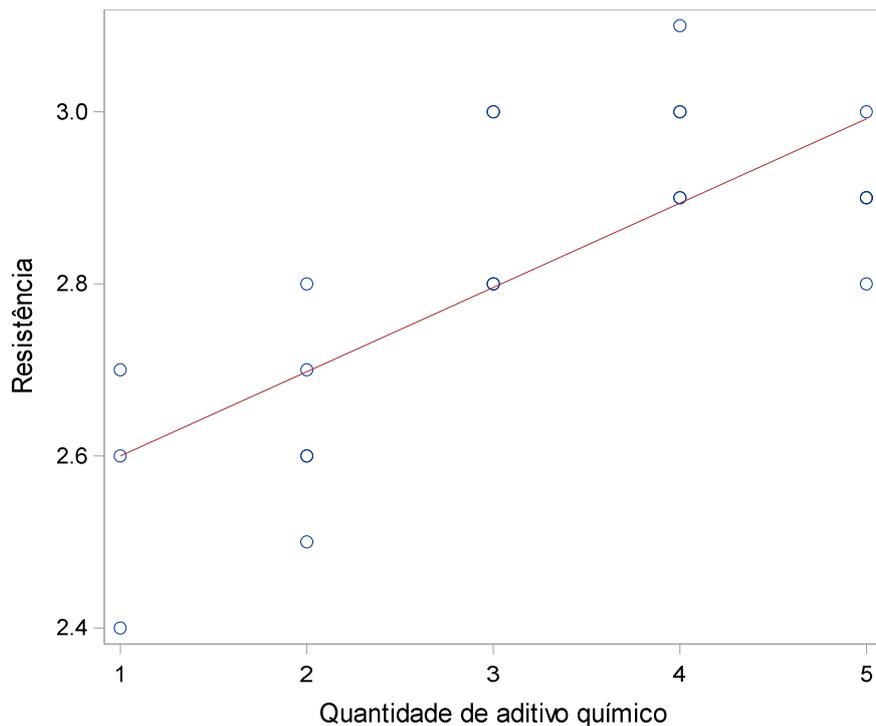


Figura 30. Tendência dos dados observados e reta de regressão para resistência do papel em relação à quantidade de produto químico adicionado.

Neste caso, a Figura 30 mostra a necessidade de inclusão de um termo quadrático no modelo de regressão. Desta forma, pode-se fazer uma avaliação prévia dos dados mediante o ajuste de um modelo de regressão polinomial do segundo grau no PROC SGPLOT com a seguinte sintaxe SAS:

```

title "Ajuste polinômio segundo grau";
proc sgplot data=paper;
  reg x=amount y=strength / degree=2 lineattrs=(color=green pattern=mediumdash)
  legendlabel="2 Grau";
  label amount="Quantidade de aditivo químico" strength="Resistência";
run;

```

O resultado é apresentado na Figura 31 onde de forma visual é possível avaliar que o modelo de regressão polinomial de segundo grau é mais adequado para os dados observados.

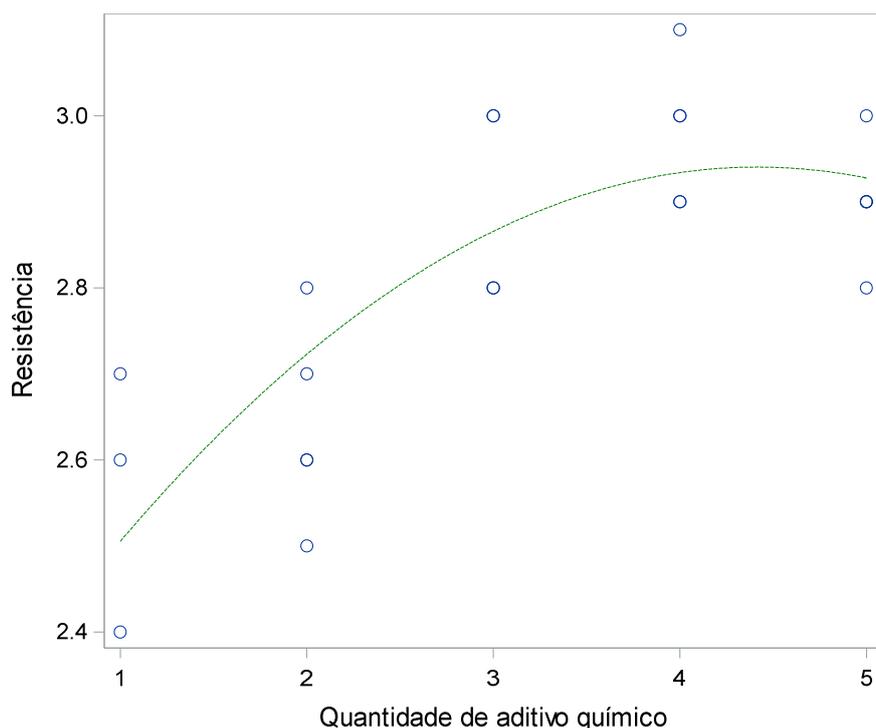


Figura 31. Tendência dos dados observados e polinômio do segundo grau para resistência do papel em relação à quantidade de produto químico adicionado.

Os procedimentos apresentados para a análise exploratória de dados via análise visual de gráficos, apenas fornecem subsídios para análise subjetiva do comportamento ou tendência dos dados. Neste caso, a adequação dos dados para determinado modelo de regressão deve ser realizada analiticamente considerando critérios de bondade de ajuste combinado com análise de resíduos mediante outros procedimentos SAS a serem descritos nos próximos capítulos.

3.3. Ajuste de modelos de regressão linear

Objetivos de aprendizagem desse capítulo:

- i) Demonstrar as fórmulas utilizadas para estimar os coeficientes de regressão linear simples;
- ii) Demonstrar os procedimentos do SAS System para o ajuste e realizar a interpretação dos resultados.

3.3.1. Estimativa dos coeficientes de regressão linear

Após a verificação dos dados mediante análise exploratória é hora de estimar os coeficientes de regressão (Betas) e realizar as análises de regressão.

De acordo com a especificidade dos dados, pode-se calcular os coeficientes de regressão mediante o emprego das seguintes técnicas:

- Mínimos Quadrados Ordinários;
- Mínimos Quadrados Generalizados;
- Máxima Verossimilhança.

Para n observações da variável dependente e variável independente, o modelo de regressão em notação matricial é:

$$y = X\beta + \varepsilon \quad (1)$$

Sendo que os vetores e matrizes são representados da seguinte forma:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n \cdot 1}, X = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix}_{n \cdot p}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{p \cdot 1}, \varepsilon = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{n \cdot 1}$$

Em que:

y = é um vetor $n \cdot 1$ que representa a variável dependente;

X = matriz das variáveis independentes (matriz design) com dimensão $n \cdot (p+1)$. A primeira coluna contém uma série de valores 1 que representa o intercepto (β_0) e as demais colunas contém os valores das variáveis independentes;

β = representa o vetor dos coeficientes de regressão $\beta_0, \beta_1, \dots, \beta_p$;

ε = representa o vetor dos resíduos.

O método de Mínimos Quadrados Ordinários busca minimizar a variância dos resíduos, ou seja, considera um critério capaz de encontrar um estimador $\hat{\beta}$ que minimiza a soma de quadrados dos resíduos.

Os resíduos são calculados pela diferença entre os valores observados e valores estimados pela regressão:

$$\varepsilon = y - X\hat{\beta} \quad (2)$$

Por sua vez, a soma de quadrados dos resíduos é representada por $\varepsilon'\varepsilon$ como:

$$[e_1 \ e_2 \ \dots \ \dots \ e_n]_{1..n} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ \vdots \\ e_n \end{bmatrix}_{n..1} = [e_1 \ e_1 + e_2 \ e_2 + \dots \ \dots \ e_n e_n]_{1..1} \quad (3)$$

Também podemos representar a soma de quadrados como:

$$\begin{aligned} \varepsilon'\varepsilon &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\ &= y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta} \end{aligned} \quad (4)$$

Para determinar $\hat{\beta}$ que minimiza a soma de quadrados dos resíduos, é necessário obter a derivada da equação 4 com respeito a $\hat{\beta}$:

$$\frac{\partial \varepsilon'\varepsilon}{\partial \beta} = -2X'y + 2X'X\hat{\beta} = 0 \quad (5)$$

A partir da equação 5 podemos obter o que se conhece por equações normais:

$$(X'X)\hat{\beta} = X'y \quad (6)$$

Em cálculo matricial a divisão de matrizes é realizada considerando o produto de uma matriz com sua inversa. Desta forma, para isolar $\hat{\beta}$ da equação 6, deve-se multiplicar ambos os termos pela inversa $(X'X)^{-1}$:

$$(X'X)^{-1}X'X\hat{\beta} = (X'X)^{-1}X'y \quad (7)$$

Por definição temos que $(X'X)^{-1}(X'X)$ é igual à matriz identidade (I). Portanto, o resultado é:

$$I\hat{\beta} = (X'X)^{-1}X'y; \hat{\beta} = (X'X)^{-1}X'y = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} \quad (8)$$

Para que a aplicação do método de mínimos quadrados ordinários seja válida, algumas condicionantes devem ser atendidas:

- Homocedasticidade. Na qual se estabelece que a variância de e_i é a mesma (σ^2) para todas observações i , ou seja $\text{Var}(\varepsilon) = \sigma^2 \forall i$ o que implica em variâncias constantes;
- Independência dos resíduos (auto correlação ausente). Para uma observação qualquer, saber algo sobre o resíduo não diz nada sobre o resíduo de outra observação $\text{Cov}(e_i, e_j) = 0 \forall i \neq j$.

Essas duas condicionantes podem ser representadas matricialmente como:

$$\text{Var}(\varepsilon) = \sigma^2 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = \sigma^2 I \quad (9)$$

A matriz de variância e covariância indica variâncias constantes (diagonal principal) e ausência de correlação entre as observações (valores zero nas demais linhas e colunas).

A estimação dos coeficientes de regressão de um modelo linear simples pode ser realizada usando as seguintes fórmulas:

Estimativa de $\hat{\beta}_0$:

$$\hat{\beta}_0 = \frac{(\sum y \cdot \sum x^2) - (\sum xy \cdot \sum x)}{(n \cdot \sum x^2) - (\sum x)^2} \quad (10)$$

Estimativa do $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{(n \cdot \sum xy) - (\sum x \cdot \sum y)}{(n \cdot \sum x^2) - (\sum x)^2} \quad (11)$$

Neste caso, a variável y representa a variável dependente e a variável x , a independente. Para aplicação dessas fórmulas considere o seguinte caso florestal:

Caso florestal 5: Ajuste manual de um modelo de regressão linear simples

Considere que um pesquisador deseja ajustar um modelo de regressão para estimar incremento periódico anual em área transversal dos últimos quatro anos (IPAg) a partir de valores do diâmetro a altura do peito (d) em árvores de *Cedrela odorata* L. Para tal, o pesquisador selecionou uma amostra representativa de 10 árvores ($n=10$) em uma floresta nativa da Amazônia e obteve as variáveis. Para o ajuste, considerou o seguinte modelo de regressão:

$$IPAg_i = \beta_0 + \beta_1 D_i + \varepsilon_i$$

Para determinar os coeficientes de regressão com as fórmulas 10 e 11 procede-se à criação das variáveis necessárias de acordo ao quadro 17.

Quadro 17. Dados observados e criação de variáveis e somatórios para a aplicação das fórmulas para estimar os coeficientes de regressão linear simples.

	x	y	x ²	x·y
Árvore	D	IPAg	D ²	D·IPAg
1	47.1	136.7	2218.4	6437.2
2	20.2	18.9	408.0	381.3
3	89.1	165.7	7938.8	14765.1
4	83.6	410.9	6989.0	34352.4
5	53.3	258.2	2840.9	13763.1
6	52.1	67.1	2714.4	3498.2
7	42.9	98.2	1840.4	4213.1
8	62.5	168.4	3906.3	10527.9
9	70.0	238.6	4900.0	16704.4
10	96.0	132.1	9216.0	12680.9
Soma	616.8000	1694.9198	42972.1800	117323.5585

Estimativa de $\hat{\beta}_0$ a partir dos valores do quadro 1:

$$\begin{aligned}\hat{\beta}_0 &= \frac{(\sum y \cdot \sum x^2) - (\sum xy \cdot \sum x)}{(n \cdot \sum x^2) - (\sum x)^2} \\ &= \frac{(1694,9198 \cdot 42972,1800) - (117323,5585 \cdot 616,8000)}{(10 \cdot 42972,1800) - (616,8000)^2} \\ &= \frac{469227,7242}{42279,5600} = \mathbf{9,5218}\end{aligned}$$

Estimativa de $\hat{\beta}_1$ a partir dos valores do quadro 1:

$$\begin{aligned}\hat{\beta}_1 &= \frac{(n \cdot \sum xy) - (\sum x \cdot \sum y)}{(n \cdot \sum x^2) - (\sum x)^2} = \frac{(10 \cdot 117323,5585) - (616,8000 \cdot 1694,9198)}{(10 \cdot 42972,1800) - (616,8000)^2} \\ &= \frac{127809,0542}{42279,5600} = \mathbf{2,5936}\end{aligned}$$

Portanto, a equação que estima os valores de ipag a partir do diâmetro a altura do peito é a seguinte:

$$\widehat{IPA}_g = 9,5218 + 2,5936 \cdot D$$

3.3.2. Procedimentos SAS para ajuste de regressão linear

Para ajustar modelos de regressão linear, o SAS System possui vários procedimentos, alguns deles como o PROC REG, PROC GLMSELECT e o PROC GLM serão abordados neste capítulo. O procedimento REG é amplamente utilizado para análise de regressão linear, pois possui várias opções de análise visual gráfica para a tomada de decisão.

O uso de um procedimento sobre o outro depende dos objetivos da pesquisa e complexidade dos resultados necessários para responder o problema de pesquisa. Diferenças existem na eficiência de cálculo, opções de diagnósticos e especificação de variáveis com interação no modelo.

Ademais, caso seja necessário a avaliação de variáveis independentes do tipo categórica no modelo de regressão, cada procedimento possui vantagens para uso.

3.3.2.1. Ajuste de modelo de regressão com variáveis preditoras contínuas no PROC IML

Os coeficientes do modelo de regressão do caso florestal 5 podem ser estimados pela resolução de sistemas lineares por operações de matrizes. O SAS possui um procedimento exclusivo para álgebra matricial, o PROC IML.

Para a estimativa dos coeficientes de regressão pelo sistema de equações normais (fórmula 8) é necessário criar a matriz das variáveis independentes e o vetor da variável dependente.

Desta forma, a aplicação da equação 8 para a estimativa dos coeficientes de regressão pode ser realizada no PROC IML pela seguinte sintaxe:

```
/*Alternativa 1 para aplicação do sistema de equações normais (fórmula 8)*/  
  
proc iml;  
X={1 47.1,  
1 20.2,  
1 89.1,  
1 83.6,  
1 53.3,  
1 52.1,  
1 42.9,  
1 62.5,  
1 70.0,  
1 96.0};  
  
y={136.7,  
18.9,  
165.7,  
410.9,  
258.2,  
67.1,  
98.2,  
168.4,  
238.6,  
132.1};
```

```

print "matriz de design X";
print x;

print "vetor y";
print y;

print "matriz X transposta";
xt = x';
print xt;

print "produto entre matrizes X' e X";
xtx = xt*x;
print xtx;

print "inversa de xtx";
inv_xtx = inv(xtx);
print inv_xtx;

print 'produto xty';
xty = xt*y;
print xty;

print "estimativa betas pela formula 8";
betas = inv_xtx*xty;
print betas;

/*Alternativa 2 para aplicação do sistema de equações normais (fórmula 8)*/
data caso2;
  input arvore d ipag;
  datalines;
1 47.1 136.7
2 20.2 18.9
3 89.1 165.7
4 83.6 410.9
5 53.3 258.2
6 52.1 67.1
7 42.9 98.2

```

```

8 62.5 168.4
9 70.0 238.6
10 96.0 132.1
;
proc iml;
use caso2;
read all;

n=nrow(d);
u=j(n,1,1);

*vetor y;
y=ipag;

*matriz de design;
x=u||d;

print "matriz X transposta";
xt = x`;
print xt;

print "produto entre matrizes X` e X";
xtx = xt*x;
print xtx;

print "inversa de xtx";
inv_xtx = inv(xtx);
print inv_xtx;

print 'produto xty';
xty = xt*y;
print xty;

print "estimativa betas pela formula 8";
betas = inv_xtx*xty;
print betas;

```

Após o processamento de qualquer uma das duas alternativas do PROC IML, os resultados do SAS são apresentados no output 19.

Output 19. Resultados das operações de matrizes utilizando o PROC IML do SAS System para ajuste do modelo de regressão do caso florestal 5.

vetor y

y
136.7
18.9
165.7
410.9
258.2
67.1
98.2
168.4
238.6
132.1

matriz de design x

x
1 47.1
1 20.2
1 89.1
1 83.6
1 53.3
1 52.1
1 42.9
1 62.5
1 70
1 96

matriz x transposta

xt									
1	1	1	1	1	1	1	1	1	1
47.1	20.2	89.1	83.6	53.3	52.1	42.9	62.5	70	96

produto entre matrizes x' e X

<i>xtx</i>	
10	616.8
616.8	42972.18

inversa de xtx

<i>inv_xtx</i>	
0.8720082	-0.012516
-0.012516	0.0002029

produto xty

<i>xyt</i>
1694.8
117314.81

estimativa betas pela formula 8

<i>betas</i>
9.5267867
2.5932752

3.3.2.2. Ajuste de modelo de regressão com variáveis preditoras contínua no PROC REG

O procedimento PROC REG possui a seguinte sintaxe padrão para ajustar modelos de regressão linear seja simples ou múltipla:

```
title "Regressão com o intercepto";  
proc reg data= nome_do_dataset;  
  model variável_dependente = variáveis_independentes;  
run;  
  
title "Regressão sem o intercepto";  
proc reg data= nome_do_dataset;  
  model variável_dependente = variáveis_independentes / noint;  
run;
```

Na declaração MODEL não é necessário incluir os coeficientes de regressão do modelo, apenas as variáveis independentes separadas por espaço. Caso se deseje ajustar uma regressão passando pela origem, basta indicar ao SAS a opção NOINT. Existem centenas de opções para análise de regressão no procedimento PROC REG e algumas delas serão aplicados nos capítulos específicos.

Caso florestal 6: Ajuste de um modelo de regressão linear múltipla

Em uma pesquisa de campo, em cada uma de 123 árvores foram obtidas as seguintes variáveis: diâmetro a altura do peito (d), altura total (h), forma da copa (formcopa) e a carga de lianas na copa (lianas). Ademais, em cada árvore extraiu-se uma amostra do tronco com trado de Pressler para calcular a área do alburno (aalb). As variáveis “ d ”, “ h ” e “aalb” são quantitativas contínuas e as variáveis formcopa e lianas são categóricas do tipo ordinal.

A pergunta de pesquisa foi a seguinte: Um modelo de regressão pode ser desenvolvido para explicar os valores de área de alburno a partir das variáveis independentes?

Para ajustar o modelo de regressão do caso florestal 6 primeiramente será considerado apenas as variáveis independentes do tipo quantitativas utilizando o datajob SAS para ajustar o seguinte modelo de regressão:

$$aalb_i = \beta_0 + \beta_1 D_i + \beta_2 H_i + \varepsilon_i$$

```

data bnut;
  input arvore aalb d h formcopa$ lianas$;
  datalines;

30 183.3829 83.10 33.00 perfeita sem
32 227.5006 226.00 44.00 perfeita sem
34 225.1188 131.10 43.00 perfeita sem
36 107.7894 59.20 32.00 tolerável com
38 198.1581 105.10 40.00 perfeita sem
.
.
.
Mais dados...
.
;
proc reg data= bnut plots=none;
  model aalb=d h;
run;

```

A opção PLOTS=NONE no PROC REG foi solicitada para suprimir intencionalmente todos os gráficos a fim de simplificar os resultados. Esses gráficos serão abordados no capítulo sobre análise de resíduos.

O modelo de regressão foi informado na linha MODEL considerando apenas as variáveis dependentes e independentes sem os coeficientes de regressão e o termo do resíduo. Observe que neste caso, é ajustado no PROC REG com apenas três linhas de programação. O resultado é mostrado no output 20.

Output 20. Resultados do ajuste do modelo de regressão do caso florestal 6. Neste caso, apresentando apenas as tabelas sem os gráficos de análise residual.

. Number of Observations Read	123
Number of Observations Used	123

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	325897	162948	63.00	<.0001
Error	120	310356	2586.29795		
Corrected Total	122	636253			

Root MSE	50.85566	R-Square	0.5122
Dependent Mean	162.61640	Adj R-Sq	0.5041
Coeff Var	31.27339		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standar Error	t Value	Pr > t
Intercept	Intercept	1	1.76221	25.53638	0.07	0.9451
d	d	1	1.34241	0.13675	9.82	<.0001
h	h	1	0.64289	0.77890	0.83	0.4108

A primeira tabela apresenta dados da amostra com observações lidas e utilizadas para o ajuste. É importante que o número de observações lidas seja igual ao número de observações utilizadas. Caso contrário, algum problema com os valores de alguma variável tornará esses valores diferentes.

A tabela da Análise de Variância (Analysis of Variance) mostra uma análise da variabilidade observada nos dados e a variação explicada pela linha de regressão. Cada coluna da tabela ANOVA para a regressão ajustada está organizada nas seguintes colunas:

Source (Fonte de variação) nome da fonte de variação.

DF graus de Liberdade associados a cada fonte de variação.

Sum of Squares (Soma de quadrados) montante de variação associada com cada fonte de variação.

Mean Square (Quadrado médio) resulta do quociente entre a soma de quadrados e graus de liberdade de cada fonte de variação.

F Value (Valor de F) resulta do quociente entre o quadrado médio do modelo e o

quadrado médio do erro. Esta razão compara a quantidade de variação explicada pela regressão com a variação não explicada pela regressão.

Pr>F (Valor de p para F) para realizar teste de hipótese.

O valor de F testa se o coeficiente angular ($\hat{\beta}_1$) da variável independente é igual a zero (Hipótese de nulidade, h_0).

A coluna **Source** é aplicada para as seguintes fontes de variação:

Model é a variação explicada pelo modelo de regressão ajustado;

Error é a variação não explicada pelo modelo (Resíduos);

Corrected Total é a variação total nos dados.

Logo em seguida, a tabela de critérios estatísticos possui cinco critérios estatísticos descritos a seguir:

Root MSE (Raiz quadrada do Quadrado médio do erro) valor estimado para o Erro Padrão de estimativa.

Dependent Mean (Média aritmética de y) Valor médio da variável dependente observada.

Coeff Var Coeficiente de variação que representa o desvio padrão relativo à média.

R Square Coeficiente de determinação. Este valor varia entre 0 e 1. Informa a proporção de variação observada nos dados que é explicada pela equação de regressão ajustada.

Adj R Sq Coeficiente de determinação ajustado. Este valor varia entre 0 e 1. Considera no cálculo o número de coeficientes no modelo de regressão.

A tabela de parâmetros estimados (Parameter Estimates) mostra o modelo ajustado para os dados com as seguintes informações:

DF representa os Graus de Liberdade associados com cada termo no modelo.

Parameter Estimate são os valores estimados dos parâmetros de regressão do modelo.

Standard Error é o erro padrão de cada parâmetro estimado.

t Value valor de t , o qual é calculado mediante o quociente entre o valor do parâmetro estimado e seu respectivo erro padrão.

Pr > |t| valor de p (p-value) associado com a estatística t . Testa se o parâmetro associado com cada termo do modelo é diferente de zero (0). Para o exemplo, o valor do coeficiente angular associado a variável independente é estatisticamente diferente de zero (0).

Neste caso, a equação associada ao modelo por meio da amostra de 123 árvores que estima a área de alburno (aalb) a partir do diâmetro a altura do peito (d) e a altura total (h) é a seguinte:

$$\widehat{aalb}_i = 1,76221 + 1,34241D_i + 0,64289H_i$$

No caso em que se deseje avaliar o efeito de variáveis independentes do tipo qualitativa no modelo de regressão é necessário a criação de variáveis indicadoras (Dummy) pois o procedimento PROC REG não aceita a inclusão de variável independente qualitativa, diretamente na declaração model.

Para fins de demonstração do ajuste considerando variável independente do tipo qualitativa juntamente com o diâmetro e a altura, consideremos uma modificação no caso florestal 6 em que também foi quantificado a carga de lianas na copa “Liana” e registrado a forma da copa “fcopa” em cada árvore. Neste caso, as variáveis são do tipo qualitativa ordinais com 2 e 5 categorias, respectivamente.

O ajuste do modelo para responder o caso florestal 6 modificado será realizado utilizando os procedimentos PROC REG, PROC GLM e PROC GLMSELECT a seguir.

3.3.3. Ajuste de modelo incluindo variável qualitativa no PROC REG

Para incorporar as variáveis de carga de liana e da forma de copa dentro do modelo de regressão para área de alburno do caso florestal 6 deve-se primeiro criar variáveis indicadoras (Indicator Variables) para os níveis das variáveis categóricas e, em seguida, incluí-las na seleção do modelo.

Variáveis indicadores também são conhecidos como variáveis Dummy e utilizam escala binária (1 ou 0) sendo o valor 1 caso uma condição existe e 0 do contrário. Portanto, variáveis Dummy servem como um substituto da variável categórica assim como um manequim de teste de colisão é um substituto para uma vítima de colisão.

No caso da variável “lianas” que possui 2 categorias de nível ordinal (com e sem), quando uma árvore apresentar uma observação “com” para a variável “lianas”, o SAS atribuirá valor 1 e zero do contrário.

Para a variável categórica de forma da copa (formcopa), será necessário criar três variáveis Dummy para cada nível ordinal de forma da copa.

As variáveis Dummy são criadas a seguir da declaração input do SAS conforme o seguinte datajob SAS.

```

data bnut;
    input arvore aalb D H formcopa$ lianas$;

*criando variáveis Dummy para Lianas;
if lianas ="com" then pres_liana=1;
else if lianas = "sem" then pres_liana=0;

*criando variáveis Dummy para Forma da Copa (formcopa);
fc_perf=0;
fc_boa=0;
fc_tole=0;

if formcopa ="perfeita" then fc_perf=1;
if formcopa = "boa" then fc_boa=1;
if formcopa = "tolerável" then fc_tole=1;

    datalines;

30 183.3829 83.10 33.00 perfeita sem
32 227.5006 226.00 44.00 perfeita sem
34 225.1188 131.10 43.00 perfeita sem
36 107.7894 59.20 32.00 tolerável com
38 198.1581 105.10 40.00 perfeita sem
.
.
.
Mais dados...
.
;

proc print data=bnut (obs=5);
    run;

proc reg data= bnut plots=none;
    model aalb=d h pres_liana fc_perf fc_boa fc_tole;
    run;

```

Utilizou-se a opção PLOTS=NONE no procedimento de forma intencional para suprimir os gráficos para diagnóstico do modelo de regressão. O procedimento gera automaticamente um painel de gráficos para análise de resíduos e pontos influentes. Esse tema será avaliado em detalhe no capítulo sobre análise de resíduos. Caso o usuário desejar visualizar os gráficos basta apagar o comando por inteiro.

Observe que na declaração MODEL deve-se incluir o modelo de regressão considerando apenas as variáveis dependente e independentes sem os coeficientes de regressão.

Após a solicitação de análise de regressão, o SAS gera os resultados organizados em quatro tabelas que inclui a ANOVA e parâmetros estimados de acordo ao Output 21:

Output 21. Resultados para a análise de regressão com o PROC REG para o caso florestal 6. O painel de gráficos de diagnóstico foi suprimido da apresentação.

Tabela contendo as primeiras 5 observações solicitadas no PROC PRINT (OBS=5) para mostrar as variáveis indicadoras criadas para representar os níveis das variáveis categóricas.

Obs	Arvore	aalb	D	H	Formcopa	Lianas	pres_liana	fc_perf	fc_boa	fc_tole
1	30	183.3829	83.10	33.00	perfeita	sem	0	1	0	0
2	32	227.5006	226.00	44.00	perfeita	sem	0	1	0	0
3	34	225.1188	131.10	43.00	perfeita	sem	0	1	0	0
4	36	107.7894	59.20	32.00	tolerável	com	1	0	0	1
5	38	198.1581	105.10	40.00	perfeita	sem	0	1	0	0

Tabelas da análise de regressão. Os gráficos foram suprimidos intencionalmente pela opção PLOTS=NONE.

Number of Observations Read	123
Number of Observations Used	123

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	356300	71260	29.78	<.0001
Error	117	279953	2392.75697		
Corrected Total	122	636253			

Root MSE	48.91582	R-Square	0.5600
Dependent Mean	162.61640	Adj R-Sq	0.5412
Coeff Var	30.08049		

Note: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

$$fc_tole = \text{Intercept} - fc_perf - fc_boa$$

Parameter Estimates						
Variable	Label	DF	ParameterEstimate	Standard Error	t Value	Pr > t
Intercept	Intercept	B	-27.11528	26.18926	-1.04	0.3026
d	D	1	1.43680	0.13440	10.69	<.0001
h	H	1	0.16836	0.76541	0.22	0.8263
pres_liana		1	3.18550	13.08719	0.24	0.8081
fc_perf		B	46.71479	13.42797	3.48	0.0007
fc_boa		B	30.39320	14.52028	2.09	0.0385
fc_tole		0	0	.	.	.

Nota: Extrapolação do modelo para dados além da amplitude amostrada é inapropriado. Não podemos assumir que a relação observada se mantém para valores superiores ou inferiores não amostrados.

Para realizar estimativas, o modelo ajustado (equação) assume os seguintes valores para os coeficientes de regressão:

$$\widehat{aalb} = -27,11528 + (1,43680 \cdot D) + (0,16836 \cdot H) + (3,18550 \cdot pres_liana) + (46,71479 \cdot fc_perf) + (30,39320 \cdot fc_boa)$$

Portanto, o valor estimado da área de alburno para uma árvore com $D=83,10$ cm, $H=33$, sem presença de lianas na copa ($pres_liana=0$) e com forma de copa perfeita ($fc_perf=1$; $fc_boa=0$) é:

$$\widehat{aalb} = -27,11528 + (1,43680 \cdot 83,10) + (0,16836 \cdot 33) + (3,18550 \cdot 0) + (46,71479 \cdot 1) + (30,39320 \cdot 0) = 144,55 \text{ cm}^2$$

Observe que caso a árvore em questão tivesse uma forma de copa tolerável, ambos os valores dos coeficientes de regressão associados à forma perfeita e forma boa serão multiplicados por valores 0. Neste caso, o intercepto foi ajustado para considerar a forma de copa tolerável.

Um detalhe importante que deve ser avaliado é o valor do coeficiente de regressão associado à presença de liana na copa. O sinal positivo do coeficiente indica associação positiva entre a área de alburno e à presença de liana na copa. Esse comportamento diverge do que realmente acontece na floresta, pois árvores com presença de lianas tendem a apresentar menor área de alburno quando comparado a árvores sem lianas na copa.

Esse comportamento do coeficiente se deve a alguns fatores. Um deles pode estar associado a um problema de Colinearidade. Esse coeficiente não deve ser considerado na equação pois o valor-p indica que o mesmo é não significativo. Esse detalhe será descrito no capítulo a seguir sobre teste de hipótese.

É possível solicitar ao PROC REG a seleção automática de variáveis considerando métodos de seleção bem como critérios de bondade de ajuste. As sintaxes a seguir mostram algumas alternativas para o ajuste do modelo de regressão pelo PROC REG.

- a) Seleção automática do melhor modelo considerando alguns critérios de bondade. Neste caso, o procedimento lista uma série de modelos combinando as variáveis e finaliza com o melhor modelo de acordo aos critérios:

```
proc reg data= bnut plots=none;  
  model aalb= D H pres_liana fc_perf fc_boa fc_tole / selection=adjrsq bic ;  
run;
```

- b) Seleção automática do melhor modelo considerando alguns critérios de bondade. Ademais, solicita ao procedimento que se deseje um modelo final com o mínimo duas e no máximo 3 variáveis (START=2 STOP=3). Também é possível solicitar que se mostre apenas os 10 melhores modelos com a opção BEST=10:

```

proc reg data= bnut plots=none;
  model aalb= D H pres_liana fc_perf fc_boa fc_tole / selection=adjrsq bic
        start=2 stop=3 best=10;
run;

```

- c) Seleção automática do melhor modelo considerando métodos de seleção Forward, Backward e Stepwise utilizando valor-p como nível de significância para entrar (SLENTY) e permanecer (SLSTAY) no modelo final:

```

proc reg data= bnut plots=none;
  model aalb=D H pres_liana fc_perf fc_boa fc_tole / selection=forward slentry=0.05;
run;

proc reg data= bnut plots=none;
  model aalb=D H pres_liana fc_perf fc_boa fc_tole / selection=backward slstay=0.05;
run;

proc reg data= bnut plots=none;
  model aalb=D H pres_liana fc_perf fc_boa fc_tole / selection=stepwise slentry=0.05
        slstay=0.05;
run;

```

3.3.4. Ajuste de modelo incluindo variável qualitativa no PROC GLM

O procedimento PROC GLM também pode ser utilizado para ajustar modelos de regressão linear além de outros modelos lineares gerais. Esse procedimento possui a declaração CLASS que possibilita incluir diretamente as variáveis categóricas dentro do modelo de regressão sem a necessidade de criar previamente as variáveis indicadores (Dummy). Portanto, procedimentos SAS que possuam a declaração CLASS, a criação de variáveis Dummy é realizada de forma automática durante o processamento do modelo estatístico contendo variáveis categóricas.

A sintaxe do procedimento para ajustar o modelo do caso florestal 6 é a seguinte:

```

data bnut;
  input arvore aalb D H formcopa$ lianas$;
  datalines;

30 183.3829 83.10 33.00 perfeita sem
32 227.5006 226.00 44.00 perfeita sem
34 225.1188 131.10 43.00 perfeita sem
36 107.7894 59.20 32.00 tolerável com
38 198.1581 105.10 40.00 perfeita sem
.
.
.
Mais dados...
.
;

proc glm data= bnut plots=none;
  class lianas formcopa;
  model aalb= D H lianas formcopa / solution;
run;

```

Por padrão o PROC GLM não imprime os parâmetros de regressão do modelo ajustado. Neste caso, basta adicionar a opção SOLUTION na declaração MODEL que o SAS cria uma tabela com os coeficientes do modelo.

Após o processamento o resultado apresenta as mesmas tabelas do procedimento PROC REG com exceção de duas tabelas a mais com a soma de quadrados do tipo I e do tipo III para os efeitos do modelo. Os resultados são apresentados no Output 22:

Output 22. Resultados para a análise de regressão com o PROC GLM para o caso florestal 6. O painel de gráficos de diagnóstico foi suprimido da apresentação.

Tabela Anova

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	356300.0095	71260.0019	29.78	<.0001
Error	117	279952.5653	2392.7570		
Corrected Total	122	636252.5748			

Tabela com critérios estatísticos para avaliar a bondade de ajuste do modelo

R-Square	Coeff Var	Root MSE	Aalb Mean
0.559998	30.08049	48.91582	162.6164

Tabelas com a soma de quadrados do tipo I e tipo II para os efeitos do modelo de regressão

Source	DF	Type I SS	Mean Square	F Value	Pr > F
D	1	324134.9241	324134.9241	135.47	<.0001
H	1	1761.8963	1761.8963	0.74	0.3926
Lianas	1	503.5861	503.5861	0.21	0.6473
Formcopa	2	29899.6030	14949.8015	6.25	0.0026

Source	DF	Type III SS	Mean Square	F Value	Pr > F
d	1	273440.8168	273440.8168	114.28	<.0001
h	1	115.7686	115.7686	0.05	0.8263
Lianas	1	141.7626	141.7626	0.06	0.8081
Formcopa	2	29899.6030	14949.8015	6.25	0.0026

Tabela com os coeficientes de regressão

Parameter	Estimate		Standard Error	t Value	Pr > t
Intercept	-27.11528068	B	26.18926347	-1.04	0.3026
d	1.43679658		0.13440422	10.69	<.0001
h	0.16836125		0.76541347	0.22	0.8263
Lianas com	3.18550173	B	13.08718823	0.24	0.8081
Lianas sem	0.00000000	B	.	.	.
Formcopa boa	30.39319557	B	14.52028394	2.09	0.0385
Formcopa perfeita	46.71479458	B	13.42797265	3.48	0.0007
Formcopa tolerável	0.00000000	B	.	.	.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

3.3.5. Ajuste de modelo incluindo variável qualitativa no PROC GLMSELECT

A sintaxe para ajustar o modelo de regressão do caso florestal 6 utilizando o PROC GLMSELECT é a seguinte:

```
data bnut;
  input arvore aalb D H formcopa$ lianas$;
  datalines;

30 183.3829 83.10 33.00 perfeita sem
32 227.5006 226.00 44.00 perfeita sem
34 225.1188 131.10 43.00 perfeita sem
36 107.7894 59.20 32.00 tolerável com
38 198.1581 105.10 40.00 perfeita sem
.
.
.
Mais dados...
.
;

proc glmselect data= bnut plots=none;
  class lianas formcopa;
  model aalb= D H lianas formcopa / selection=none showpvalues;
run;
```

A opção `SELECTION=NONE` na declaração `MODEL` ajusta o modelo como especificado. Sem o uso dessa opção o procedimento irá aplicar o método de seleção automática Stepwise por padrão para escolher as variáveis no modelo. Essa opção foi utilizada apenas para fins de comparação entre os procedimentos. Maiores detalhes sobre os métodos de seleção serão abordados no capítulo 6 sobre seleção de variáveis para construção de modelos de regressão linear.

Após o processamento o SAS gera os resultados apresentados no Output 23 conforme descrito a seguir.

Output 23. Resultados para a análise de regressão para o caso florestal 6 utilizando o PROC GLMSELECT. O painel de gráficos de diagnóstico foi suprimido da apresentação.

Tabela ANOVA

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	356300	71260	29.78	<.0001
Error	117	279953	2392.75697		
Corrected Total	122	636253			

Tabela com critérios estatísticos para avaliar a bondade de ajuste do modelo

Root MSE	48.91582
Dependent Mean	162.61640
R-Square	0.5600
Adj R-Sq	0.5412
AIC	1087.81351
AICC	1088.78742
SBC	979.68661

Tabela com os coeficientes de regressão.

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	-27.115281	26.189263	-1.04	0.3026
d	1	1.436797	0.134404	10.69	<.0001
h	1	0.168361	0.765413	0.22	0.8263
Lianas com	1	3.185502	13.087188	0.24	0.8081
Lianas sem	0	0	.	.	.
Formcopa boa	1	30.393196	14.520284	2.09	0.0385
Formcopa perfeita	1	46.714795	13.427973	3.48	0.0007
Formcopa tolerável	0	0	.	.	.

3.3.6. Comparação de ajuste entre os procedimentos REG, GLM E GLMSELECT

Como observado, os três procedimentos podem ser utilizados para o ajuste de modelos de regressão linear e apresentaram valores idênticos para a tabela ANOVA, tabela de critérios de bondade e ajuste e tabela de parâmetros da regressão.

A grande diferença nos resultados foram alguns critérios de bondade de ajuste a mais fornecidos pelo PROC GLMSELECT (AIC, AICC E SBC).

Cada procedimento possui capacidades adicionais e a escolha por um deles depende muito do objetivo e da disponibilidade de variáveis da pesquisa.

O Quadro 18 mostra um resumo de aplicações dos procedimentos para ajuste de modelos de regressão linear.

Quadro 18. Comparação entre os procedimentos SAS para análise de regressão linear.

Procedimento	Algoritmo para seleção de variáveis	Inclusão variáveis categóricas na análise	Painel para análise de resíduos
PROC REG	Sim	Não*	Sim
PROC GLM	Não	Sim	Sim
PROC GLMSELECT	Sim	Sim	Sim

*=É possível por meio da criação de variáveis do tipo Dummy.

Os procedimentos PROC GLM e PROC GLMSELECT possuem a declaração CLASS que suporta a seleção e avaliação de variáveis independentes do tipo categóricas no modelo de regressão.

Para a avaliação de gráficos de resíduos, todos os procedimentos constroem por padrão painéis de gráficos de resíduos úteis para avaliar condicionantes de regressão e pontos influentes de forma visual.

O PROC GLMSELECT combina recursos dos procedimentos GLM e REG para análise de regressão. Foi disponibilizado na versão 9.1 do SAS e foi desenvolvido intencionalmente para ser um procedimento de seleção automática de variáveis independentes para modelos lineares geral. Portanto, possui várias opções de customização de uso de critérios para a seleção e pausa de variáveis independentes em um modelo de regressão.

3.3.6.1. O que acontece se aparece a letra “B” na tabela de parâmetros estimados?

Quando se inclui variáveis categóricas no modelo de regressão, o SAS imprime uma mensagem sobre a situação da matriz de variáveis independentes (Matriz design) na estimativa dos parâmetros do modelo ajustado.

Quando ajustado no PROC GLM a mensagem é a seguinte:

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

A mensagem que o SAS imprime pode causar dúvidas sobre a confiança dos resultados visto que os parâmetros de regressão são seguidos pela letra “B” são enviesados. Será que podemos confiar nos resultados do ajuste para esse modelo?

Para responder à pergunta é necessário entender a forma com que a matriz de design é composta quando variáveis categóricas são incluídas no modelo de regressão.

Ao considerarmos o caso florestal 5 com 123 observações (n) e sete variáveis independentes (considerando variáveis Dummy para os níveis das variáveis categóricas), a matriz design X do modelo de regressão em notação matricial tem a seguinte notação matricial:

$$y = X\beta + \varepsilon,$$
$$X = \begin{bmatrix} 1 & 83,1 & 33 & 0 & 1 & 0 & 1 & 0 \\ 1 & 226,0 & 44 & 0 & 1 & 0 & 1 & 0 \\ 1 & 131,1 & 43 & 0 & 1 & 0 & 1 & 0 \\ 1 & 59,2 & 32 & 1 & 0 & 0 & 0 & 1 \\ 1 & 105,1 & 40 & 0 & 1 & 0 & 1 & 0 \\ 1 & 67,0 & 33 & 0 & 1 & 0 & 1 & 0 \\ 1 & 139,9 & 38 & 0 & 1 & 0 & 1 & 0 \\ 1 & 164,6 & 57 & 0 & 1 & 1 & 0 & 0 \\ \vdots & \vdots \\ 1 & 105,1 & 49 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}_{123 \times 8}$$

Neste caso a dimensão da matriz design é obtida pelo produto $n \cdot (p+1)$ sendo, p =número de parâmetros e n =número de observações da amostra.

O produto da matriz design transporta pela matriz design $X'X$ gera uma matriz quadrada 8×8 com a seguinte estrutura:

$$X'X = \begin{bmatrix} 123 & 12656,2 & 4347,9 & 17 & 106 & 33 & 72 & 18 \\ 12656,2 & 1470093 & 459739,7 & 1791,8 & 10864,4 & 3542 & 7096,8 & 2017,4 \\ 4347,9 & 459739,7 & 158866 & 625,6 & 3722,3 & 1165,6 & 2571,6 & 610,7 \\ 17 & 1791,8 & 625,6 & 17 & 0 & 5 & 7 & 5 \\ 106 & 10864,4 & 3722,3 & 0 & 106 & 28 & 65 & 13 \\ 33 & 3542 & 1165,6 & 5 & 28 & 33 & 0 & 0 \\ 72 & 7096,8 & 2571,6 & 7 & 65 & 0 & 72 & 0 \\ 18 & 2017,4 & 610,7 & 5 & 13 & 0 & 0 & 18 \end{bmatrix}$$

Neste caso, a matriz $X'X$ é singular pois a mesma tem um determinante igual a zero (0). O fato de $X'X$ ser singular implica nas estimativas dos coeficientes de regressão que são calculados por:

$$\hat{\beta} = (X'X)^{-1}X'y = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$$

$X'X$ é singular, e, portanto, não é possível obter sua inversa. Mas o que significa uma matriz singular em termos práticos? Significa que existe colunas linearmente dependente entre si, ou seja, uma determinada coluna da matriz $X'X$ é gerada a partir da combinação aritmética de outras duas colunas.

Isso se deve à inclusão de variáveis Dummy para representar as variáveis categóricas no modelo de regressão. Esse detalhe é exposto pelo PROC REG no aviso emitido nos resultados:

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

`fc_tole =` Intercept - fc_perf - fc_boa

Por esta razão a tabela de parâmetros estimados apresenta valor zero para a variável “Lianas sem” e para “Formcopa tolerável” e, conseqüentemente, um valor faltante para o erro padrão, valor de t e valor-p do respectivo parâmetro.

Mas como o SAS ajusta o modelo de regressão se a matriz $X'X$ é singular? Neste caso, utiliza-se a inversa generalizada de $X'X$ para calcular os parâmetros de regressão.

Existem vários tipos de inversa generalizada para matriz singular com a inversa de Moore-Penrose. Entretanto, o SAS considera a matriz inversa de Sweep no PROC GLM para a estimativa de coeficientes de regressão (GOODNIGHT, 1979).

Outra informação importante é que o modelo ajustado está com parâmetros de regressão a mais do que deveria ter (over-parameterized).

Para reverter essa situação, basta indicar ao SAS o valor de referência de cada variável categórica considerando a opção PARAM=REF na declaração CLASS do PROC GLMSELECT conforme sintaxe a seguir.

```
proc glmselect data= bnut plots=none;
  class lianas formcopa / param=ref;
  model aalb= D H lianas formcopa / selection=none showpvalues;
run;
```

Esta opção posiciona as variáveis das colunas da matriz design que são linearmente dependentes como valores de referência, e, portanto, gera resultados de parâmetros estimados que elimina as variáveis anteriormente incluídas com valores zero para os parâmetros estimados. A tabela de parâmetros estimados é apresentada no output 24 (os demais resultados do output foram excluídos).

Output 24. Resultado do ajuste do modelo considerando valor de referência para as variáveis categóricas.

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	-27.115281	26.189263	-1.04	0.3026
D	1	1.436797	0.134404	10.69	<.0001
H	1	0.168361	0.765413	0.22	0.8263
Lianas com	1	3.185502	13.087188	0.24	0.8081
Formcopa boa	1	30.393196	14.520284	2.09	0.0385
Formcopa perfeita	1	46.714795	13.427973	3.48	0.0007

Note que os valores dos parâmetros de regressão estimados são exatamente iguais aos estimados pelo PROC REG e PROC GLM.

Observa-se que para utilizar o modelo, o intercepto foi ajustado para considerar os parâmetros com valores zero. Por exemplo para árvores sem lianas: $aalb(\text{Lianas sem}) = \text{intercepto} + \text{Lianas sem} = -27,11528068 + 0 = -27,11528068$.

Com o novo ajuste, os coeficientes de regressão que ficaram com valores zero desapareceram. Neste caso: $aalb(\text{Lianas sem}) = \text{intercepto} = -27,11528068$ que é exatamente igual ao modelo anterior.

Portanto, observa-se que não há nenhum problema na estimação dos coeficientes de regressão quando aparecer o aviso no results do SAS com o PROC REG ou PROC GLM.

3.3.7. Outros procedimentos SAS para ajuste de modelos geral de regressão linear

O SAS possui outros procedimentos que podem ser utilizados para ajustar modelos de regressão linear, entretanto foram desenvolvidos para outros tipos de análises.

Para modelos lineares geral (GLM) o SAS possui diferentes procedimentos para ajuste de regressão. Neste caso, alguns procedimentos considerando a versão 9 são:

Procedimento	Módulo SAS	Especificidade
PROC MODEL	SAS/ETS	Ajuste e análise de modelos para série temporal e econometria
PROC RSREG	SAS/Stat	
PROC MIXED	SAS/Stat	Ajuste de modelos lineares mistos
PROC NLIN	SAS/Stat	Ajuste de modelos de regressão não-linear
PROC HPREG	SAS/Stat	Ajuste de modelos de regressão linear com alta performance de processamento dos dados (H=High, P=Performance)
PROC GLIMMIX	SAS/Stat	Ajuste de modelos lineares generalizados, mas também ajusta modelos geral.
PROC NLMIXED	SAS/Stat	Ajuste de modelos não-lineares mistos.

3.4. Testes de hipóteses e ANOVA em análise de regressão

Objetivos de aprendizagem desse capítulo:

- i) Determinar a significância estatística para um modelo ou coeficiente de regressão;
- ii) Determinar como decompor a variância total de uma análise de regressão em termos de soma de quadrados.

A avaliação da adequação do modelo de regressão aos dados observados é realizada, analiticamente, pela estatística F calculada na tabela de Análise de Variância (ANOVA). Assim, para a tomada de decisão é necessário realizar um teste de hipótese considerando a significância dos coeficientes angulares do modelo de regressão utilizando ($\beta_1 = \beta_2 = \dots = \beta_p$). Portanto, não se considera, no teste de hipótese, o coeficiente intercepto.

Na Figura 32 mostra-se a dispersão de dados observados para a produção de látex de árvores de seringueira (*Hevea brasiliensis*) em função da altura do fuste (H_f) juntamente com uma reta de regressão linear (linha vermelha). Observa-se que a reta é paralela ao eixo do eixo x indicando que à medida que aumenta os valores para a altura do fuste não ocorre mudança significativa nos valores de látex produzido pela árvore na ocasião da medição.

Neste caso, o modelo de regressão não se ajusta aos dados melhor do que o modelo médio padrão (média aritmética de látex). Portanto, existem evidências suficientes para dizer que a variável independente altura do fuste (H_f) não explica uma quantidade significativa de variabilidade observada para a variável látex. Essa afirmação deve ser comprovada mediante teste de hipótese.

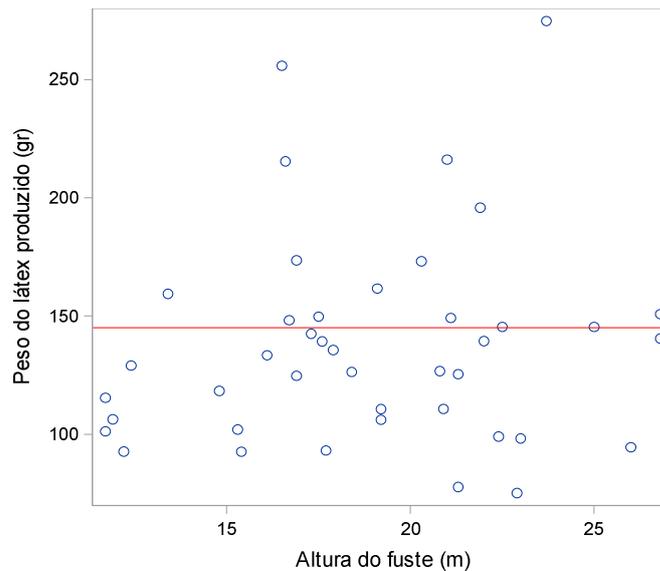


Figura 32. Modelo padrão médio no qual a variável independente altura do fuste (h_f) não explica a variação observada para a produção de látex, ou seja, não existe associação entre y e x .

A formulação da Hipótese Nula e Hipótese Alternativa para uma regressão é a seguinte:

- Hipótese Nula: $h_0 = \beta_1 = \beta_2 = \dots = \beta_p = 0$. O modelo de regressão não se ajusta aos dados observados da variável dependente melhor do que o modelo médio padrão (Todos os coeficientes de regressão são iguais a zero e, portanto, não há regressão entre y e x);
- Hipótese Alternativa: $h_1 =$ Nem todos os coeficientes de regressão β_i s são iguais a zero). O modelo de regressão se ajusta aos dados observados da variável dependente melhor do que o modelo médio padrão.

Para realizar o teste de hipótese geral, considera-se o valor F de Snedecor calculado seguido do valor-p a partir da tabela de Análise de Variância (ANOVA). Assim, caso o valor-p seja menor do que o nível de significância estabelecido (α , geralmente 5%), rejeita-se a hipótese nula.

A tabela ANOVA para análise de regressão organiza as fontes de variação devido à regressão e ao acaso bem como a soma das duas fontes de variação. A variação é representada em números pela soma de quadrados sendo que cada fonte de variação possui sua respectiva soma de quadrados de acordo com o Quadro 19.

Quadro 19. Análise de variância mostrando os cálculos necessários para a obtenção do valor F de uma regressão linear.

Fonte de variação	Graus de Liberdade	Soma de quadrados	Quadrado médio	Valor F	p-valor
Devido à regressão	$p-1$	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{p-1}$	$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / p-1}{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / n-p}$	Se $< \alpha$ (predefinido, p.e.0,05) então, modelo significativo
Resíduos	$n-p$	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-p}$		
Total corrigido	$n-1$	$\sum_{i=1}^n (y_i - \bar{y})^2$			

Onde: p =número de coeficientes de regressão; n =número de observações da unidade amostral.

A soma de quadrados da tabela ANOVA tem sua origem a partir do cálculo dos desvios obtidos a partir dos dados observados, estimados e a média aritmética da variável dependente conforme descrito graficamente pela Figura 33.

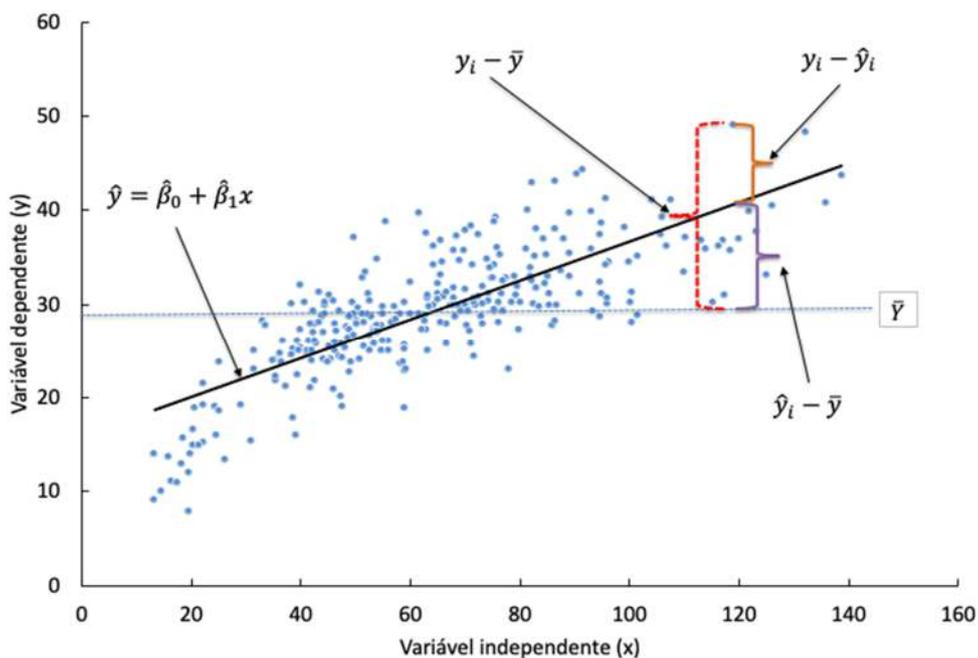


Figura 33. Representação da obtenção dos desvios para uma regressão linear simples.

A partir da Figura 33 é possível avaliar a eficiência da regressão por meio da avaliação dos diferentes desvios, como:

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + y_i - \hat{y} \quad (12)$$

Onde:

$y_i - \bar{y}$ = Desvio total;

$\hat{y}_i - \bar{y}$ = Desvio considerando o valor ajustado em relação à média;

$y_i - \hat{y}$ = Desvio considerando o valor observado em relação ao estimado.

Considerando que o somatório para as observações resultará em um valor zero, cada termo da equação 12 foi elevado ao quadrado incluindo o somatório o que resulta em:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (13)$$

Portanto, a soma de quadrados para cada fonte de variação da Tabela 19 assume a forma da equação 13.

Considerando os resultados do Output 25 para o caso florestal 6, a tabela ANOVA apresenta um F calculado de 63,00 com um valor-p de <0,0001. Nesta situação, há evidências suficientes para rejeitar h_0 . Isso significa que nem todos os coeficientes de regressão ajustados são iguais a zero e que o modelo de regressão proposto se ajusta aos dados observados de incremento periódico anual em área basal.

O teste de hipótese continua para cada um dos coeficientes de regressão. Assim, ao avaliar a tabela d do Output 25 (parâmetros estimados do caso florestal 6), observa-se que somente a variável independente diâmetro a altura do peito (d) foi significativamente diferente de zero (valor-p <0,0001) sendo a variável altura total (h) não significativa, ou seja, igual a zero (valor-p =0,4108).

Uma alternativa para o modelo é reajustá-lo sem a variável altura total e verificar a significância para o intercepto. Essa estratégia de construção de modelos de regressão de forma manual e automática será abordada no capítulo sobre modelagem.

3.5. Avaliação de modelos de regressão após o ajuste

Após o ajuste de um modelo de regressão, além dos critérios de bondade de ajuste algumas etapas devem ser consideradas para determinar se a equação obtida é a mais apropriada para representar os dados observados e assegurar previsões ou análise dos coeficientes de regressão.

Este tópico será dedicado a uma avaliação criteriosa do modelo de regressão ajustado considerando, portanto, além dos critérios de bondade de ajuste, os seguintes tópicos serão explorados a profundo:

- i) as condicionantes de regressão pela análise de resíduos;
- ii) os métodos para identificar pontos influentes e outliers; e
- iii) os efeitos da multicolinearidade de modelos de regressão linear múltipla.

3.5.1. Critérios de bondade de ajuste

Após determinar os coeficientes de regressão é desejável saber o quão bom o modelo de regressão se ajusta aos dados observados da variável dependente (y) com as variáveis independentes (x 's). Para tal propósito utilizam-se algumas estatísticas que expressam a bondade de ajuste do modelo de regressão ajustado.

3.5.1.1. Erro padrão da estimativa

O erro padrão da estimativa (S_{yx}) mede a dispersão entre os valores estimados pela regressão em torno da média aritmética considerando a unidade da variável dependente.

$$S_{yx} = \sqrt{\left[\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p} \right]}$$

Em que: n =número de observações; p =número de coeficientes de regressão; y_i =valor observado da i -ésima observação da variável dependente; \hat{y}_i =valor estimado da i -ésima observação da variável dependente.

O valor para a operação matemática entre colchetes é representado nos resultados da ANOVA do SAS por MSE=Mean Square Error que significa a variância do modelo de regressão.

Deve-se considerar a utilização desse critério para comparação entre modelos de regressão somente se os mesmos estiverem com a mesma unidade de medida da variável dependente.

Caso seja necessário comparar a bondade de ajuste entre modelos com variáveis dependentes em escalas diferentes (por exemplo escala logarítmica *versus* escala aritmética), o Erro Padrão da Estimativa deve ser recalculado para a variável de interesse.

Neste sentido, o Erro Padrão da Estimativa de um modelo em que a variável dependente (y) não sofreu transformação (escala original), deve ser comparado com o Índice de Furnival (IF) do modelo em que a variável dependente está transformada para logaritmo ($\ln(y)$).

Para calcular o IF, Silva e Bailey (1991) realizaram uma modificação no Índice conforme descrito a seguir.

$$IF = \text{Exp} \left[\frac{\sum_{i=1}^n \ln(y_i)}{n} \right] \cdot S_{yx}$$

Em que: Exp =exponencial; \ln =Logaritmo natural; n =número de observações; y_i =valor observado da i -ésima observação da variável dependente.

3.5.1.2. Coeficiente de Variação

Outro critério comparador para ajuste entre modelos de regressão é o coeficiente de variância (CV) também conhecido como Erro Padrão da Estimativa percentual.

$$CV = \frac{S_{yx}}{\bar{y}} \cdot 100$$

Em que: S_{yx} =erro padrão da estimativa; \bar{y} =média aritmética da variável dependente.

Para comparação de ajuste entre modelos ajustados com e sem transformação logarítmica, recomenda-se considerar o coeficiente de variação de modelo em escala original de y com o Índice de Furnival em porcentagem (IF%) do modelo logarítmico. O IF% é calculado conforme expressão a seguir:

$$IF\% = \frac{IF}{\bar{y}} \cdot 100$$

Em que: \bar{y} =média aritmética da variável dependente.

3.5.1.3. Coeficiente de determinação

O coeficiente de determinação (R^2) expressa a quantidade de variação total observada que é explicada pela regressão. Portanto, seu valor varia entre 0 a 1 sendo o modelo com melhor ajuste quanto mais próximo de 1.

$$R^2 = \frac{\left[\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \right]}{\left[\sum_{i=1}^n (y_i - \bar{y})^2 \right]}$$

Em que: n=número de observações; \bar{y} =média aritmética da variável dependente; y_i =valor observado da i-ésima observação da variável dependente; \hat{y}_i =valor estimado da i-ésima observação da variável dependente.

Esse coeficiente é amplamente utilizado para comparar o ajuste de modelos de regressão concorrentes. Entretanto é sensível à inclusão de novas variáveis independentes no modelo e, portanto, uma modificação em R^2 deve ser considerada. Uma modificação penalizando o número de coeficientes de regressão foi desenvolvida assegurando a comparação da bondade de ajuste de modelos de regressão com diferentes quantidades de coeficientes.

$$R_{aj}^2 = R^2 - \left[\left(\frac{p-1}{n-p} \right) \cdot (1 - R^2) \right]$$

Em que: R_{aj}^2 =coeficiente de determinação ajustado; n=número de observações; p=número de coeficientes de regressão.

Por outro lado, para a comparação do ajuste entre modelos com transformação da variável dependente em diferentes escalas Kvalseth (1985) seguido de Scott e Wild (1991) relataram que o uso do R^2 é inadequado.

Para o caso florestal 6 os critérios de bondade de ajuste foram organizados no Quadro 20 a seguir.

Quadro 20. Critérios para diagnóstico de modelos de regressão ajustado com os dados do Caso florestal 6.

Critério estatístico	Abreviação no SAS	Valor para o caso florestal 6
Erro padrão da estimativa	RootMSE*	50,86 cm ²
Coefficiente de variância	Coeff Var	31,27%
Coefficiente de determinação	R-Square	0,51
Coefficiente de determinação ajustado	Adj R-Square	0,50

*=Root square of the Mean Square Error que em tradução livre significa Raiz quadrada do Quadrado Médio dos Resíduos.

Heeringa et al. (2010) relataram que para o coeficiente de determinação, não existe um valor mínimo universalmente aceito com o qual o pesquisador deve buscar alcançar visto que, a qualidade do ajuste depende dos dados bem como da área objeto de estudo sendo que na área de física é comum se obter modelos de regressão com R^2 próximos a 0,99, na área de química $R^2 < 0,90$ e na área de Ciências Humanas modelos de regressão com R^2 de no máximo 0,20 a 0,40.

3.5.1.4. Critérios de informação para comparação entre modelos

O Quadro 21 descreve alguns critérios de informação amplamente utilizados para comparar modelos de regressão.

Quadro 21. Critérios de informação para seleção de modelos de regressão geral (GLM-Modelo Linear Geral).

Abreviação SAS	Critério	Fórmula padrão	Componente de penalidade	Fonte
AIC	Akaike	$n \text{Ln} \left(\frac{SQ_{res}}{n} \right)$	$2p + n + 2$	Akaike (1973)
AICC	Akaike corrigido		$\frac{n(n+p)}{n-p-2}$	Hurvich & Tsai (1989)
BIC	Bayesian		$2(p+2)q - 2q^2$	Sawa (1978)
SBC	Schwarz Bayesian		$p \text{Log}(n)$	Schwarz (1978)

Em que: n=número de observações; Ln= logaritmo natural; SQres=Soma de Quadrados dos Resíduos; p=número de coeficientes no modelo incluindo o intercepto (para modelos não-lineares, soma-se 1 ao valor de p. Caso um modelo possuir variável independente categórica, o valor de p é adicionado ao número de classes); $q = \frac{n\hat{\sigma}^2}{SQ_{res}}$; $\hat{\sigma}^2$ =estimativa da variância do erro puro para o modelo completo.

Fonte: RODRIGUES, 2023.

Destaca-se que os critérios não são utilizados para avaliar a qualidade do modelo, servem apenas para comparação entre modelos e não indicam se o modelo é bom ou ruim. Portanto, o modelo que apresente o menor valor é o mais adequado.

Cada critério de informação indica um modelo que minimiza a variabilidade não explicada com o menor número possível de efeitos no modelo.

O uso de um critério ou outro depende da situação e do objetivo da pesquisa. Neste caso, o Critério de Informação AICC deve ser usado em preferência de AIC quando se trabalha com tamanho de amostras pequenas, ou seja, quando o número de observações (n) é menor do que 12 vezes o número de coeficientes de regressão.

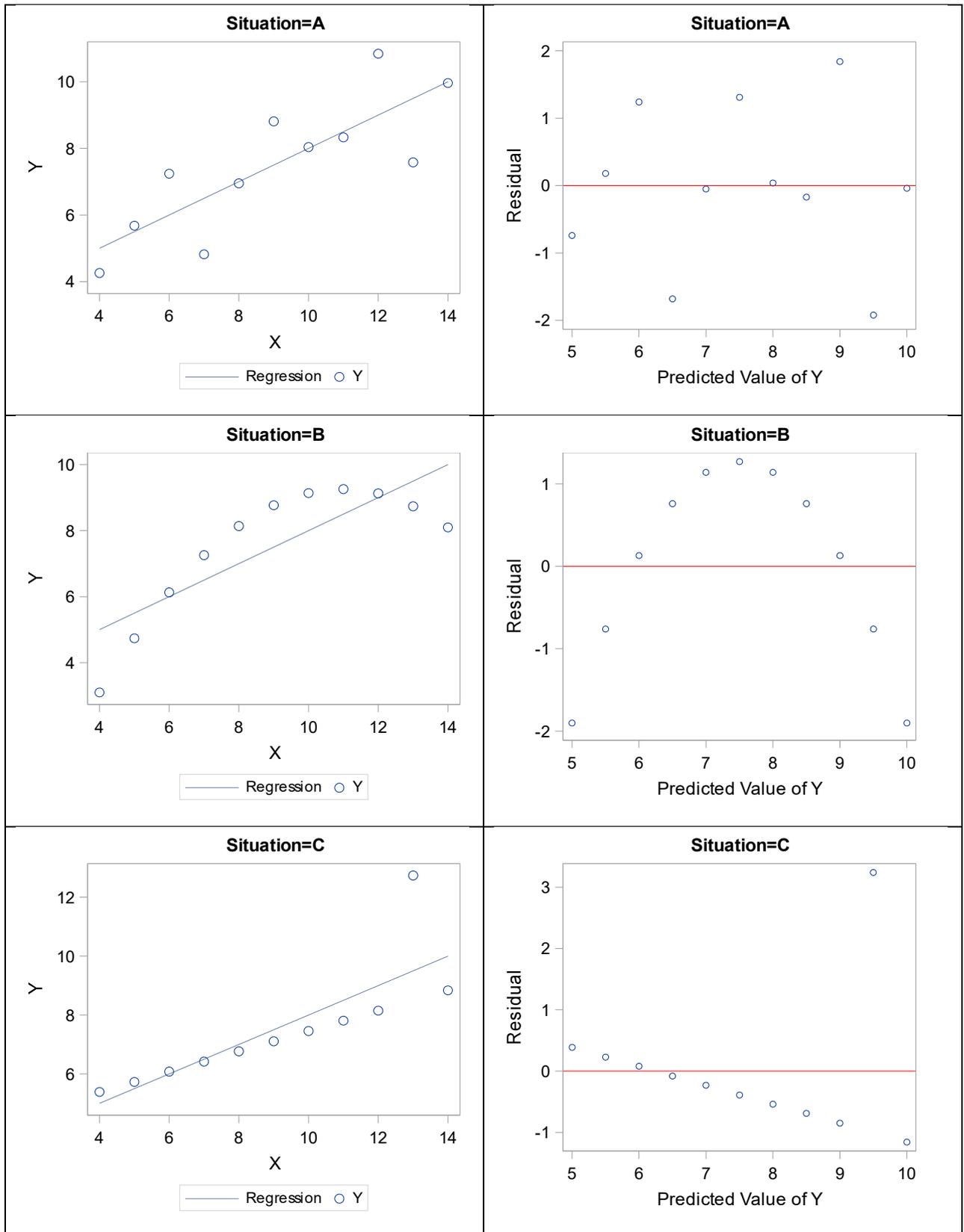
Um detalhe importante a considerar quando utilizar critérios de informação é que não se deve utilizar o critério de informação de Akaike para comparação entre o modelo com diferentes escalas na variável dependente (CORDEIRO; DEMÉTRIO, 2010), ou seja, um modelo de regressão para incremento em diâmetro não pode ser comparado com um modelo de logaritmo de incremento em diâmetro.

Para determinar o melhor modelo entre duas alternativas, considera-se o que apresentar o menor valor para os critérios de informação.

3.5.2. Análise de resíduos e observações influentes

A avaliação da qualidade de ajuste de modelos de regressão deve ser realizada não somente a partir de critérios de bondade de ajuste, mas pela avaliação visual da dispersão dos dados de resíduos obtidos a partir do modelo de regressão ajustado considerando os resíduos que são plotados no eixo das ordenadas em função dos valores de cada variável independente (x 's) ou variável dependente estimada.

Os gráficos de resíduos podem ser construídos considerando no eixo x os valores estimados ou os valores das variáveis independentes. Neste caso, ao avaliar a dispersão dos resíduos para cada situação de Anscombe (1973), facilmente observa-se problemas de ajuste em três situações (Situação B, C e D) conforme indicado na Figura 34.



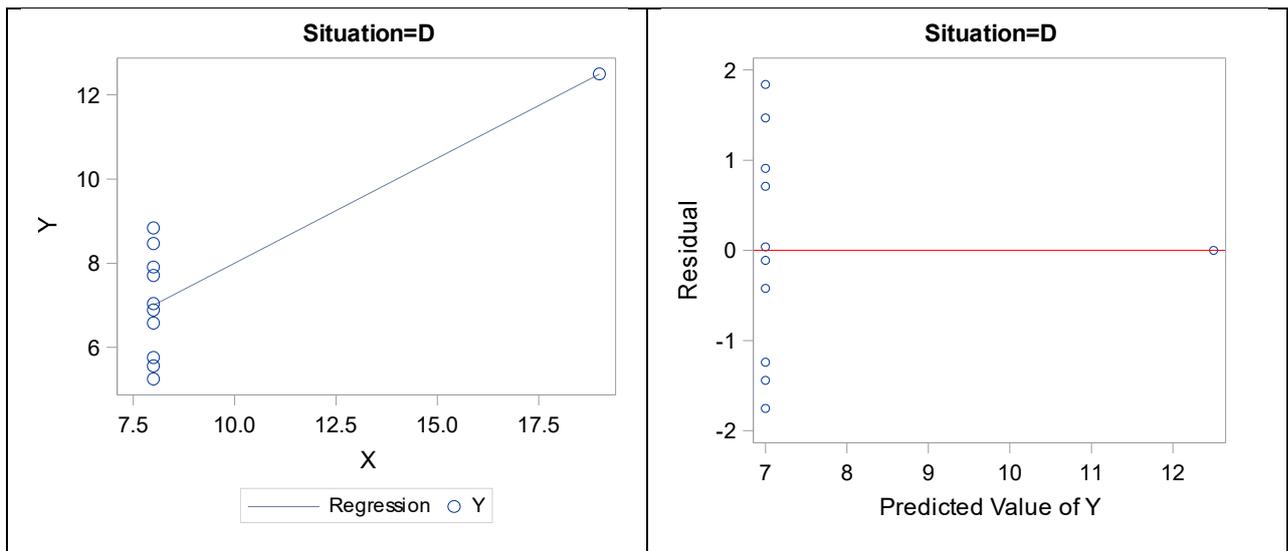


Figura 34. Quarteto de Anscombe (esquerda) mostrando a regressão linear simples ajustada para dados observados e distribuição residual (ANSCOMBE, 1973).

A avaliação visual das diferentes situações da Figura 34 mostra algumas situações não desejáveis para o comportamento dos resíduos. A situação B da Figura 34 mostra um modelo inadequado para os dados observados. Neste caso, é necessário um termo extra no modelo (um termo quadrático) ou a necessidade de transformação da variável dependente antes da análise (DRAPER; SMITH, 1981).

Para a situação C e situação D, da Figura 34 aparece um ponto influente, ou seja, caso esses pontos sejam excluídos da base de dados, a tendência da regressão pode ser diferente da atual e, conseqüentemente, os critérios de ajuste do modelo de regressão serão afetados.

3.5.2.1. Avaliação das condicionantes da regressão

A análise de resíduos além de possibilitar a avaliação da adequação do modelo de regressão utilizado para os dados, também possibilita a avaliação das condicionantes de regressão como forma de validar testes de hipótese, os intervalos de confiança entre outros quesitos necessários para o método de Mínimos Quadrados Ordinários conforme indicado no Quadro 22.

Os resíduos (ε_i) são valores obtidos pela diferença entre os valores observados e os estimados a partir do modelo matemático a seguir:

$$\varepsilon_i = y - X\hat{\beta}$$

Para que as estimativas do modelo de regressão sejam válidas, os resíduos devem atender três condicionantes representadas pela notação matemática a seguir:

$$\varepsilon_i \approx N I I D(0, \sigma^2)$$

Essa notação indica que os resíduos: $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$, devem ter a média aritmética zero (0) e a variância constante representada por sigma ao quadrado. De acordo com a notação, as letras representam as condicionantes da regressão:

- **(N)** Normalidade: Os resíduos possuem uma distribuição Normal para cada valor da variável independente;
- **(I)** Independentes: Os resíduos são independentes: $\sigma^2 I$;
- **(ID)** Identicamente Distribuídos: Os resíduos são **identicamente distribuídos**, ou seja, possuem a mesma variância para cada valor da variável independente (Homocedasticidade).

Quadro 22. Consequências para Mínimos Quadrados Ordinários diante do não atendimento às condicionantes de regressão (MODIFICADO DE HAMILTON, 1992).

Problema	Consequências não desejadas			
	Betas viesados	Erro padrão viesado	Teste t e F inválidos*	Inflação da variância de Betas
Não normalidade para distribuição de ε	Não	Não	Sim	Sim
Valores de ε correlacionados	Não	Sim	Sim	Sim
Heterocedasticidade	Não	Sim	Sim	Sim

*=para testes de hipótese.

A condicionante de “Normalidade” também deve ser atendida para validar os intervalos de confiança (Predição) para os coeficientes de regressão.

Caso o modelo de regressão atenda às condicionantes, o gráfico de resíduos versus valores estimados apresentará uma dispersão aleatória dos pontos em torno da linha de referência (zero) conforme demonstrado na Figura 11, situação A.

Analicamente as condicionantes podem ser verificadas juntamente com gráficos de resíduos por meio de testes estatísticos concebidos na literatura estatística.

No módulo SAS/STAT existem vários procedimentos que incluem testes estatísticos e possibilidades de elaboração de gráficos avançados para avaliação das condicionantes de regressão.

3.5.2.1.1. Avaliação da condicionante de “Normalidade” para os resíduos

A condicionante “Normalidade” para os resíduos pode ser verificada utilizando o procedimento PROC UNIVARIATE. Esse procedimento calcula quatro estatísticas de teste formal sob hipótese nula de que a frequência dos resíduos observada se ajusta a uma frequência normal hipotética. Os testes estatísticos para avaliar a hipótese nula são descritos no Quadro 23.

Quadro 23. Testes utilizados para avaliar aderência dos resíduos à distribuição Normal.

Nome do teste	Estatística	Hipótese Nula	Condição para uso
Shapiro-Wilk	W*	Os dados da variável sob análise (resíduos) assumem uma distribuição Normal.	Considerar para tamanho de amostra entre 7 e 2000 observações (SHAPIRO; WILK, 1965)**.
Kolmogorov-Smirnov	D		
Cramer-von Mises	W-sq		
Anderson-Darling	A-sq		

*= O valor de W é positivo e menor do que 1 sendo que, valores próximos de 1 indicam normalidade da variável sob análise. **=O SAS suprime o valor da estatística W quando o número de observações é maior do que 2000.

Além das estatísticas de teste, o PROC UNIVARIATE produz gráficos específicos para avaliar a condicionante de normalidade juntamente com os testes estatísticos. A sintaxe básica do procedimento é apresentada a seguir.

```
proc univariate data=nome_do_dataset options;  
  var variável_numérica;  
  histogram /options;  
  qqplot / options;  
  ppplot / options;  
run;
```

Neste caso, a declaração VAR analisa a distribuição e os testes para resíduos. A declaração HISTOGRAM constrói um gráfico mostrando os resíduos no eixo x em categorias (intervalos) proporcional para o total de observações e sobrepõe sob as barras a curva da distribuição Normal hipotética para comparação. A declaração QQPLOT e PPLOT gera como resultado um gráfico de resíduos em quantil-quantil e percentil-percentil.

A seguir, apresenta-se a sintaxe do procedimento utilizando dados de resíduos obtidos a partir do caso florestal 6 arquivos “saída3”. Os resultados são apresentados no Output 26:

- Parte 1 apresenta os testes formais do Quadro 5 para avaliar e outras tabelas;
- Parte 2 apresenta os gráficos solicitados.

```
/*ajuste do modelo de regressão para obtenção dos valores de resíduos no arquivo de
saída “saída3” */

data bnut;
    input arvore aalb D H formcopa$ lianas$;
    datalines;

30 183.3829 83.10 33.00 perfeita sem
32 227.5006 226.00 44.00 perfeita sem
34 225.1188 131.10 43.00 perfeita sem
36 107.7894 59.20 32.00 toleravel com
38 198.1581 105.10 40.00 perfeita sem
.
.
.
Mais dados...
.
;
proc reg data= bnut plots=none;
    model aalb=D H;
    output out=saída3 p=aald_est r=resíduos;
    run;

/*uso do procedimento para analisar a condicionante de normalidade para os resíduos*/

proc univariate data=saída3 normal;
    var resíduos;
    qqplot resíduos /normal (mu=est sigma=est color=red l=1);
    pplot resíduos /normal (mu=est sigma=est color=red l=1);
    histogram /normal (color=maroon w=4) cfill=blue cframe=ligr;
    inset kurtosis skewness mean std /cfill=blank format=5.2;
```

Output 26. Resultados do procedimento PROC UNIVARIATE para avaliar a condicionante de normalidade para os resíduos da regressão do caso florestal 6.

Output 26: Parte 1

Moments			
N	46	Sum Weights	46
Mean	0	Sum Observations	0
Std Deviation	111.608333	Variance	12456.42
Skewness	1.95854588	Kurtosis	7.36786675
Uncorrected SS	560538.901	Corrected SS	560538.901
Coeff Variation	.	Std Error Mean	16.455751

Basic Statistical Measures			
Location		Variability	
Mean	0.00000	Std Deviation	111.60833
Median	-4.96635	Variance	12456
Mode	.	Range	682.43961
		Interquartile Range	60.38179

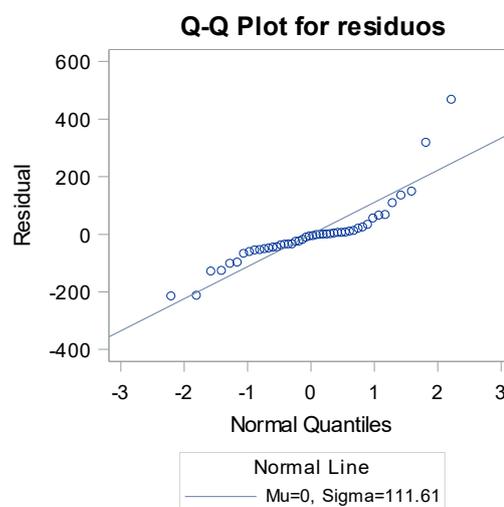
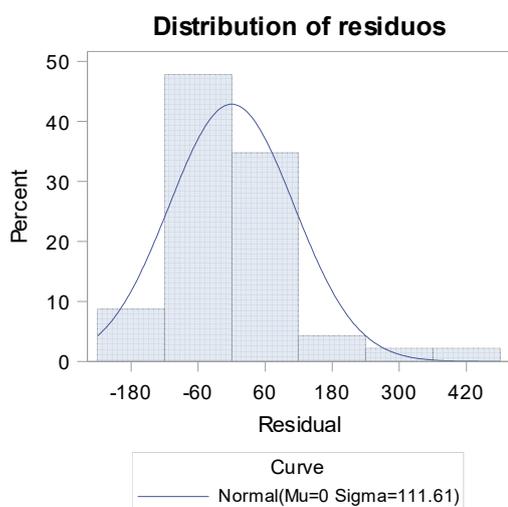
Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	0	Pr > t 	1.0000
Sign	M	-3	Pr >= M 	0.4614
Signed Rank	S	-95.5	Pr >= S 	0.3019

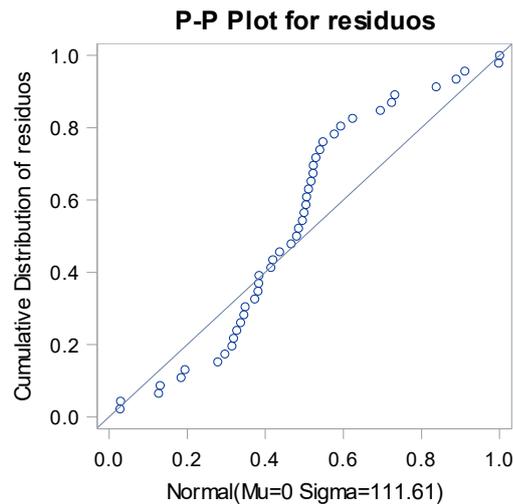
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.807127	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.2135	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.504779	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.731394	Pr > A-Sq	<0.0050

Quantiles (Definition 5)	
Level	Quantile
100% Max	468.74417
99%	468.74417
95%	150.12353
90%	109.89140
75% Q3	13.28351
50% Median	-4.96635
25% Q1	-47.09828
10%	-100.46467
5%	-127.54917
1%	-213.69543
0% Min	-213.69543

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-213.695	14	109.891	29
-211.517	43	136.158	5
-127.549	41	150.124	24
-125.640	19	319.718	10
-100.465	8	468.744	17

Output 26: Parte 2





De acordo ao resultado do Output 26, parte 1, observa-se que o procedimento produz várias tabelas no resultado da análise. A tabela nomeada “Tests for normality” apresenta os quatro testes formais e todos rejeitam a hipótese nula para normalidade dos resíduos.

Neste caso, quando o número de observações for menor do que 2000, recomenda-se considerar apenas a estatística W para o teste de hipótese de normalidade. Portanto, para amostras com o número de observações maior do que 2000, considerar o teste de Kolmogorov-Smirnov.

Na parte 2 do Output 26, observa-se que todos os gráficos indicam que os resíduos não seguem uma distribuição Normal. Neste caso, a distribuição de frequência dos resíduos (histograma) tende a se ajustar à distribuição Normal hipotética no primeiro gráfico.

O comportamento dos dados de resíduos distribuídos nos gráficos de Q-Q plot e P-P plot também indicam não aderência à distribuição Normal. Basicamente esse tipo de gráfico compara a frequência dos resíduos obtidos a partir do modelo de regressão ajustado com uma frequência esperada de resíduos com distribuição Normal. Portanto, se os pontos no gráfico são semelhantes entre si ou próximos, estes serão distribuídos bem próximos à linha diagonal de referência e, portanto, indica que os resíduos possuem distribuição Normal. Caso ocorra algum distanciamento das observações com respeito à linha diagonal de referência, como no resultado do Output 26, a suposição de resíduos Normais é violada.

Embora, visualmente atraentes, os gráficos de resíduos para normalidade não fornecem critérios objetivos para determinar a normalidade de uma variável. Portanto, devem ser analisados juntamente com os testes formais.

Para fins de comparação com os gráficos da parte 2 do Output 26, a Figura 35 mostra o comportamento dos resíduos em um histograma, Q-Q plot e P-P plot quando os resíduos possuem distribuição Normal ($n=123$; $W=0,9827$; $Pr<W= <0,1170$). Os dados considerados para a construção dos gráficos diferem do caso florestal 6.

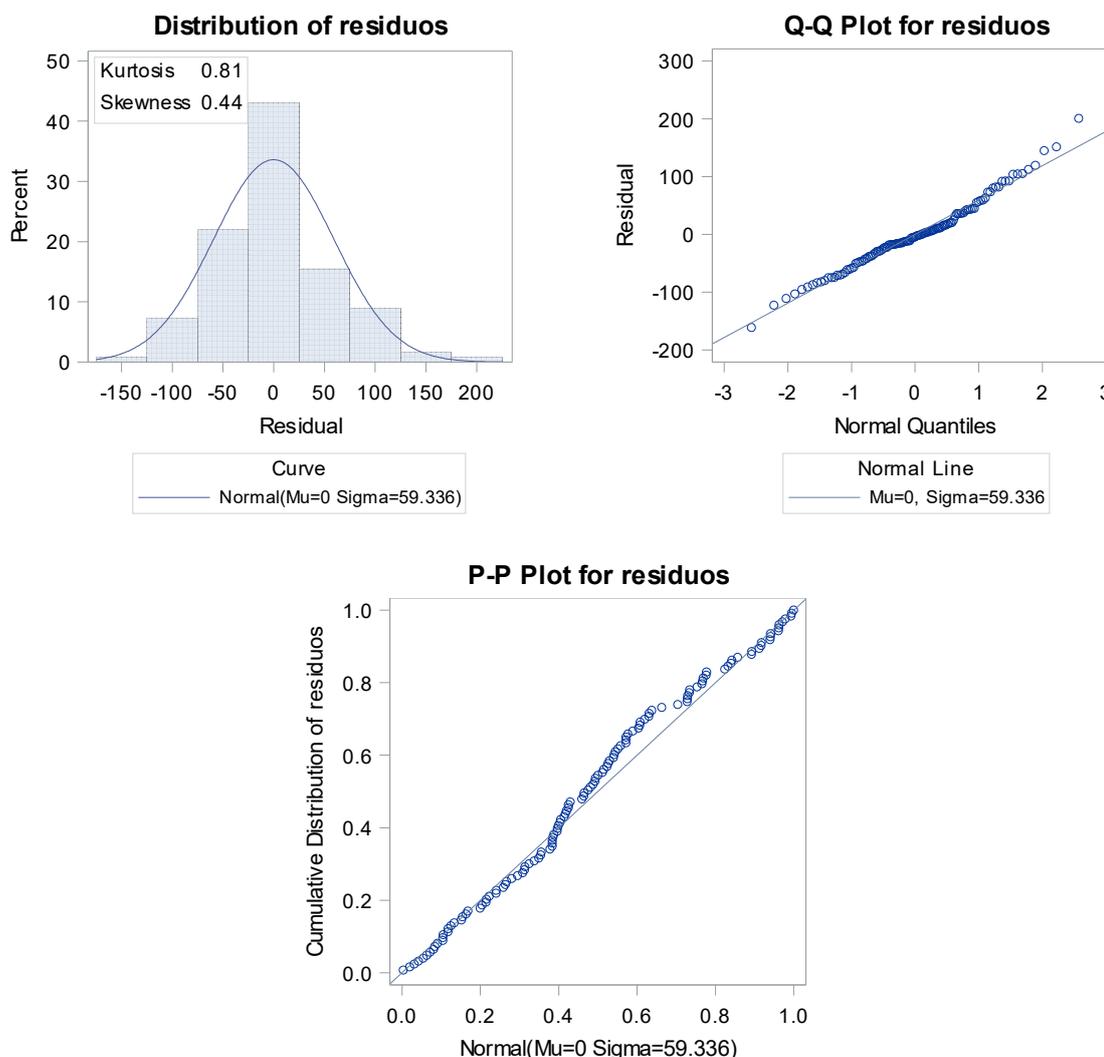


Figura 35. Comportamento da distribuição de resíduos nos gráficos de Histograma, Q-Q Plot e P-P Plot para uma situação em que a condicionante de normalidade é aceita. Observa-se que todos os pontos estão muito próximos e aderentes à linha diagonal de referência nos gráficos de Q-Q Plot e P-P Plot.

3.5.2.1.2. Avaliação das condicionantes de independência e homoscedasticidade para os resíduos

A avaliação das condicionantes de independência e homoscedasticidade para os resíduos pode ser realizada por método gráfico e por meio de testes estatísticos formais concebidos na literatura.

Para a avaliação gráfica, os resíduos são plotados no eixo das ordenadas contra os valores estimados da variável dependente no eixo da abscissa. Draper e Smith (1981) recomendaram plotar os resíduos em função da variável dependente estimada. Entretanto, gráficos de resíduos com as variáveis independentes são usuais para avaliar o efeito de cada variável na contribuição da variação.

Para avaliar se existe correlação entre qualquer observação dos resíduos (Independência), recomenda-se que se realize uma primeira avaliação sobre a fonte dos dados, ou seja, de que forma os dados da variável dependente (y) foram coletados em termos de periodicidade e espacialidade.

Dados coletados ao longo do tempo (horas, dias, meses anos etc.) possivelmente serão correlacionados e, conseqüentemente produzirão resíduos dependentes. Esse é o caso para estudos em que se coleta dados de crescimento de árvores, seja por parcelas permanentes, análise de tronco completa ou a partir de dendrômetros de crescimento.

Por outro lado, caso o comportamento da variável dependente apresente um aumento da variação em função da variável independente ocorre violação da condicionante de variâncias homogêneas. Caso se ajuste uma regressão linear simples a esse tipo de dados, os resíduos apresentarão um padrão em forma de cone com aumento de seus valores à medida que se aumenta o valor estimado. Neste caso, o modelo seria tendencioso pois apresenta menor precisão da estimativa com o aumento da variável do eixo x .

O Quadro 24 indica os meios para a avaliação das condicionantes de independência e homogeneidade de variâncias.

Quadro 24. Formas de avaliação das condicionantes de regressão.

Condicionante	Método gráfico	Método numérico
Resíduos não-correlacionados	Dispersão de resíduos versus variável dependente estimada ou variável independente de tempo. Caso a condicionante seja violada, a dispersão dos resíduos se apresenta em forma sequencial com um padrão.	Teste de Durbin-Watson $d = \frac{\sum_{i=1}^n (\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^n \varepsilon_i^2}$ A estatística d é limitada a testar apenas a autocorrelação de primeira ordem (ε_t e ε_{t-1} , em que t=tempo). A estatística d testa a hipótese nula (H_0) de que os resíduos são independentes.
Resíduos com variâncias homogêneas	Dispersão de resíduos versus variável dependente estimada e ou variáveis independentes. Caso a condicionante seja violada, o formato da dispersão dos resíduos apresenta-se com um padrão de um cone.	Teste de White, testa a hipótese nula (H_0) de que a variância dos resíduos é homogênea.

Para fins de avaliação das condicionantes de independência e homoscedasticidade para os resíduos, considere o gráfico da Figura 36 que mostra o comportamento do crescimento em diâmetro à altura do peito para 16 árvores em função do tempo bem como os valores estimados para diâmetro (linha vermelha descontínua) obtidos a partir do ajuste do modelo de crescimento de Chapman-Richards aos dados observados.

O comportamento das curvas de crescimento mostra que ao longo do tempo algumas árvores cresceram melhor do que outras causando um aumento da variação do diâmetro entre as árvores à medida que se aumenta a idade.

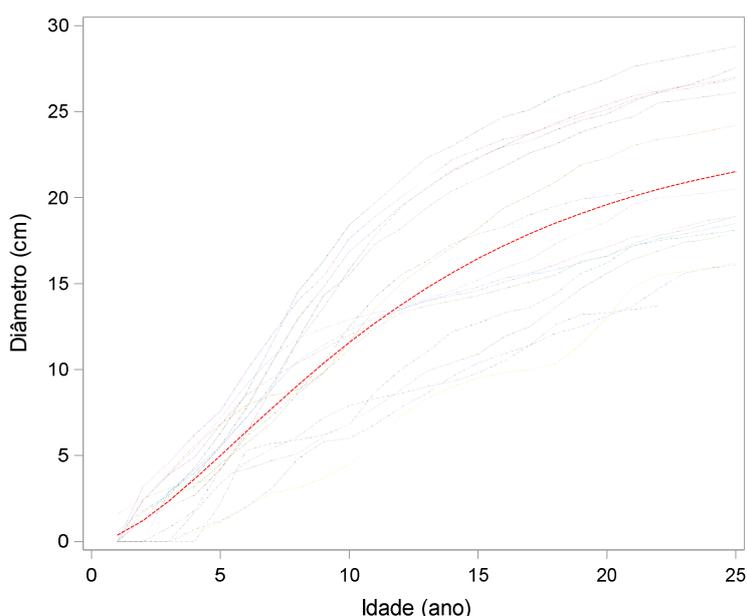


Figura 36. Comportamento observado do crescimento em diâmetro ao longo da idade de árvores. A linha vermelha descontínua representa os dados estimados para diâmetro.

Os gráficos de dispersão dos resíduos em função do diâmetro estimado e da variável independente idade são mostrados na Figura 37. A partir da análise visual observa-se claramente um ordenamento sequencial dos resíduos denominado como correlação serial (Autocorrelação), ou seja, correlação entre dados de resíduos consecutivos.

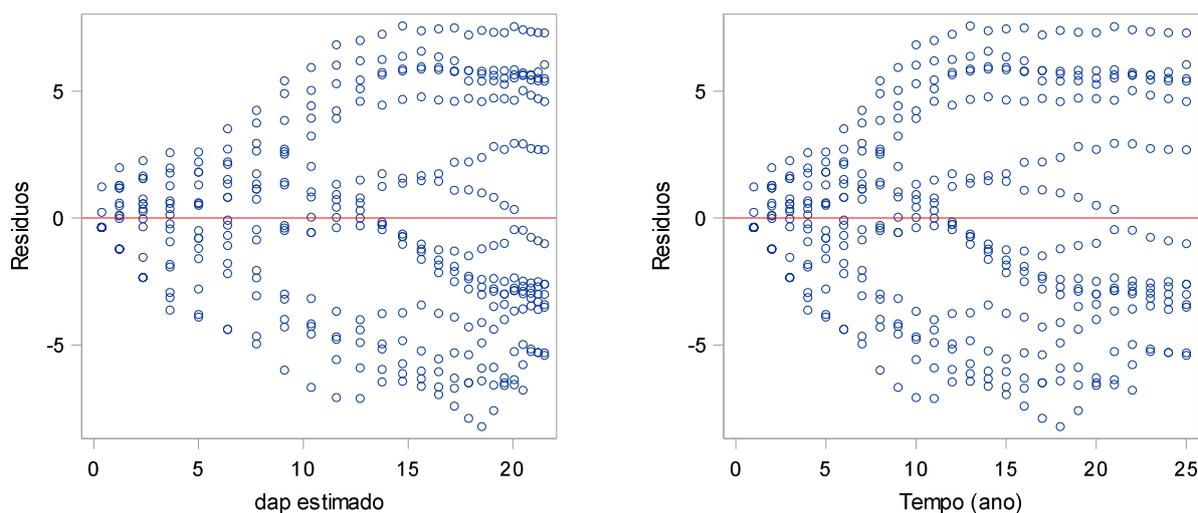


Figura 37. Padrão de dependência entre os resíduos consecutivos mostrando padrão não aleatório da dispersão. Um gráfico de resíduos versus a variável de tempo também pode informar sobre a falta de independência dos resíduos.

Para realizar teste de hipótese a fim de verificar a existência ou não de resíduos autocorrelacionados utiliza-se o teste de Durbin-Watson (d) que calcula a estatística de autocorrelação de primeira ordem. Neste caso, o teste é utilizado para testar a hipótese nula de que os resíduos possuem correlação nula ($\rho=0$) contra a hipótese alternativa de que os resíduos possuem correlação maior do que zero ($\rho>0$).

A estatística de Durbin-Watson (d) tem uma limitação de calcular apenas a autocorrelação dos resíduos em primeira ordem, ou seja, a correlação entre o valor atual e anterior de ε no tempo t . Portanto, a estatística d , não informa a existência de autocorrelação em outro nível de ordem para resíduos, como em ε_t e ε_{t-2} (segunda ordem) ou em ε_t e ε_{t-3} (terceira ordem).

A estatística d de Durbin-Watson é aproximadamente relacionada com o coeficiente de correlação r na seguinte proporção:

$$d \approx 2(1 - r)$$

Desta forma, o valor d varia de 0 a 4 sendo que na ausência de correlação ($r=0$) o valor d equivale aproximadamente a 2 indicando ausência de autocorrelação nos resíduos. Caso o valor d calculado for menor ou maior do que 2, indica autocorrelação positiva e negativa, respectivamente.

Para a conclusão do teste utilizando a estatística d de Durbin-Watson é necessário comparar o valor dessa estatística com valores críticos da tabela de Durbin-Watson d_L (Lower) e d_U (Upper), considerando-se as seguintes regras de decisão:

- Decisão 1: Se $d < d_L \rightarrow$ Rejeitar H_0
- Opção 2: $d > d_U \rightarrow$ Não rejeitar H_0
- Opção 3: $d_L < d < d_U \rightarrow$ Teste inconclusivo

Para avaliar a condicionante homogeneidade de variâncias para os resíduos (Homocedasticidade) pode-se realizar uma análise gráfica dos resíduos em que os valores são plotados em função da variável dependente estimada. A condição é que a dispersão dos valores de resíduos não deve apresentar um padrão definido com tendência, e sim, um padrão totalmente aleatório como visto na distribuição de resíduos da Figura 25.

Ocorre Heterocedasticidade quando a distribuição dos resíduos tende a aumentar ou diminuir à medida que se aumenta a variável estimada. Este é o caso da distribuição residual da Figura 37 que dá indícios suficientes de não cumprimento dessa condicionante e indica que a variância dos resíduos não é constante, ou seja, os resíduos são heterocedásticos.

A detecção da heterocedasticidade pode ser realizada por meio de gráficos produzidos automaticamente após o ajuste de um modelo de regressão no procedimento PROC REG do SAS. Esse procedimento oferece como resultado gráficos avançados para análise residual organizados em painéis.

É possível solicitar os testes formais para avaliar se os resíduos cumprem ou não com as condicionantes de independência e de homocedasticidade no procedimento PROC REG do SAS. Basta adicionar a opção DW DWPROB SPEC logo após a especificação do modelo de regressão.

A opção DW e DWPROB considera o teste de Durbin-Watson para a hipótese nula de que os resíduos são independentes.

Quando a opção SPEC é solicitada no procedimento PROC REG o SAS calcula uma versão do teste conhecido como “White’s test” que é frequentemente utilizado para testar heterocedasticidade em modelos de regressão descrito no teorema 2 do artigo original de

Halbert White (White, 1980). Desta forma a opção SPEC testa em conjunto a hipótese nula de que:

- i) os resíduos são homocedásticos;
- ii) os resíduos são independentes e
- iii) o modelo está corretamente especificado.

White (1980) enfatiza que se a hipótese nula for rejeitada existe heterocedasticidade. Entretanto, a rejeição da hipótese nula pode ser atribuída à falta de independência dos resíduos que por sua vez é, comumente, causada pela heterocedasticidade devido à dependência da variância dos resíduos nas variáveis preditoras.

Se a conclusão da rejeição da hipótese nula for devido a especificação do modelo, o teorema 2 de White (WHITE, 1980) indica apenas que algo está errado com o mesmo. Neste caso, se o pesquisador estiver confiante que o modelo está corretamente especificado, a causa da rejeição é a heterocedasticidade.

Neste sentido, o teste de White é considerado generalista (THURSBY, 1982) sendo que o resultado pode ser significativo quando os resíduos são homocedásticos, mas por outro lado, o modelo pode não estar corretamente especificado.

A sintaxe do procedimento PROC REG a seguir solicita esses testes formais para o caso florestal 6:

```
/*Avaliação das condicionantes de regressão de Independência e Homocedasticidade com o  
proc reg*/  
  
proc reg data= bnut plots=none;  
    model aalb=d h / dw dwprob spec;  
run;
```

Após o processamento o SAS apresenta a tabela ANOVA e duas novas tabelas contendo os valores dos testes de Durbin-Watson e o teste de White conforme indicado no Output 27. As tabelas da ANOVA, bondade de ajuste e os gráficos de resíduos foram omitidas para apresentar apenas os testes de condicionantes solicitados.

Output 27. Resultados obtidos a partir do processamento dos dados do caso florestal 6 para o cálculo de Durbin-Watson e do teste de White.

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
5	4.65	0.4607

Durbin-Watson D	1.914
Pr < DW	0.3572
Pr > DW	0.6428
Number of Observations	46
1st Order Autocorrelation	0.040

O SAS calcula o valor de Qui-quadrado sob hipótese nula de que a variância é a mesma, porém desconhecida, para cada uma das classes de resíduos formada. De acordo ao valor de probabilidade ($Pr > ChiSq = 0,4607$) há evidências suficientes para não rejeitar a hipótese nula, ou seja, o teste indica que os resíduos cumprem com a condicionante de Homocedasticidade.

A estatística d de Durbin-Watson calculada foi de 1,914 e está próximo ao valor 2 que indica resíduos independentes. A tabela de valores críticos de Durbin-Watson para um nível de significância de 0,05, três coeficientes de regressão (K) e um $n=45$ indica valores críticos de $d_L=1,38$ e de $d_U=1,67$. Assim, de acordo a regra de decisão o valor d é maior do que o valor crítico d_U o que leva a não rejeição de h_0 .

O teste de hipótese para Durbin-Watson também pode ser realizado pelo valor-p obtido a partir da declaração `dwprob`. Neste caso, tanto para resíduos positivos ($Pr < DW$) como negativos ($Pr > DW$) o valor-p é maior do que o nível de significância indicando a não rejeição de h_0 .

3.5.2.2. Avaliação de observações influentes

Observações influentes são valores contidos na base de dados que podem proporcionar mudança substancial em alguma parte do processo de modelagem em análise de regressão mesmo após a avaliação da bondade de ajuste e das condicionantes de regressão.

Desta forma, caso uma observação com valor incomum seja excluída dos dados e cause mudança drástica no resultado dos coeficientes de regressão, medidas de bondade

de ajuste e, conseqüentemente, valores estimados da variável dependente, essa observação é considerada como um ponto influente nos dados.

Portanto, caso um modelo de regressão linear simples, ou múltipla, seja ajustado a um conjunto de dados que contenha uma observação (ou mais de uma) substancialmente diferente das demais observações, o resultado gráfico dos dados observados e estimados seria semelhante aos gráficos da Situação C e D da Figura 38.

O conjunto de gráficos da Figura 38 faz parte de um trabalho publicado por Anscombe (1973) para demonstrar a importância de se avaliar os efeitos de observações influentes nas propriedades estatísticas em uma análise de regressão.

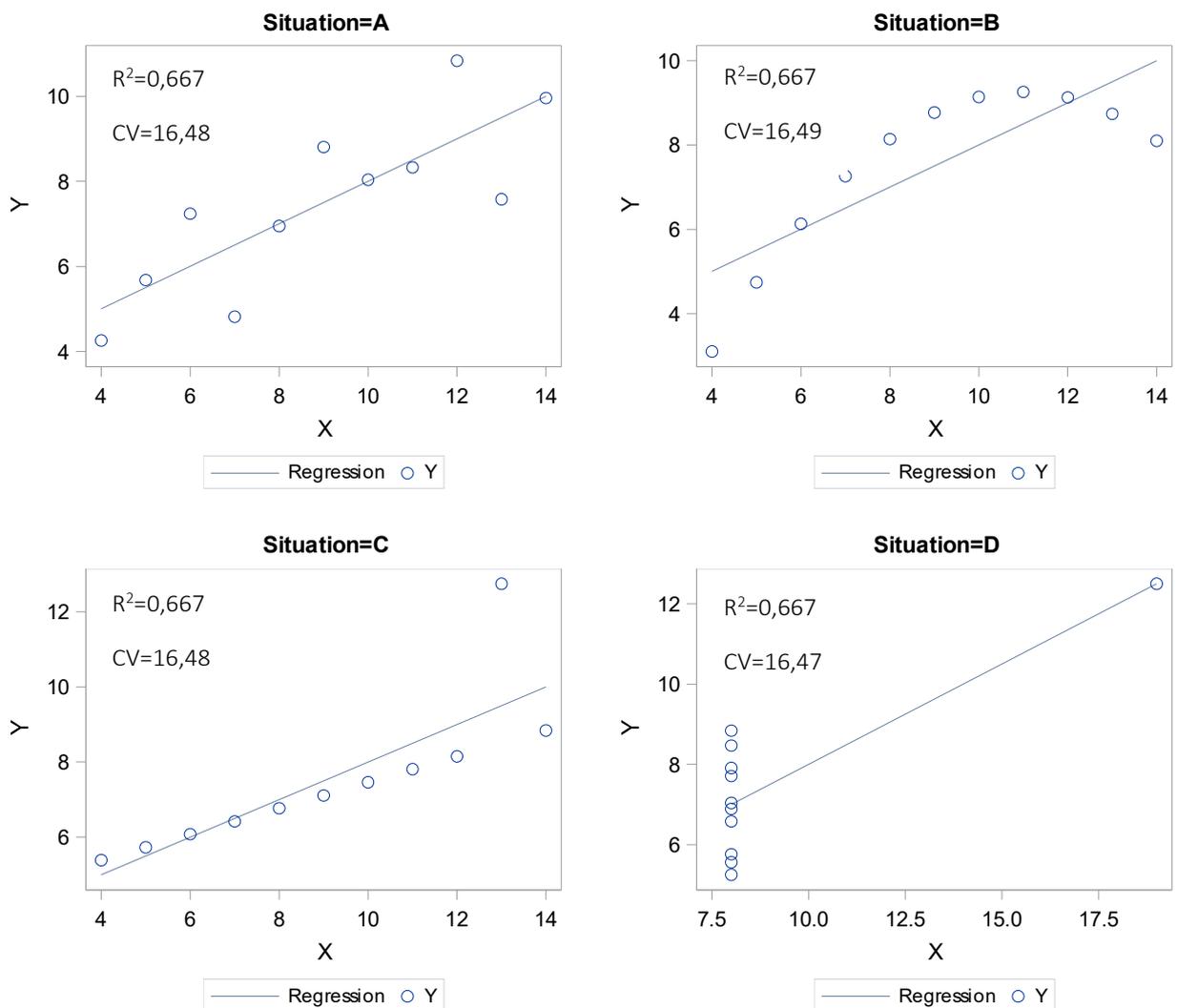


Figura 38. Quarteto de Anscombe mostrando a regressão linear simples ajustada para dados observados com dispersão diferente, mas com mesmos valores de ajuste de acordo a ANSCOMBE (1973).

Para a situação “A” a linha de regressão aparentemente representa bem os dados observados. Entretanto, para as demais situações a análise de regressão deve ser avaliada com cautela. Para a situação “B” a regressão linear simples não é apropriada para o conjunto de dados em que y tem relação curvilínea para valores de x.

A observação influente contida na situação C da Figura 38 pode ser classificada como Outlier visto que o valor não segue a tendência geral dos dados. Por sua vez, o quão esse valor pode influenciar na análise é indicado pela Leverage (Ponto de alavancagem). Este é o caso da situação D da Figura 38.

Para a situação “D”, provavelmente não haveria regressão significativa caso seja excluído o valor atípico dos dados.

A Figura 39 mostra duas retas de regressão, uma ajustada com os dados considerando o ponto influente da situação C de Anscombe (linha azul) e outra reta ajustada após a eliminação do ponto influente (laranja).

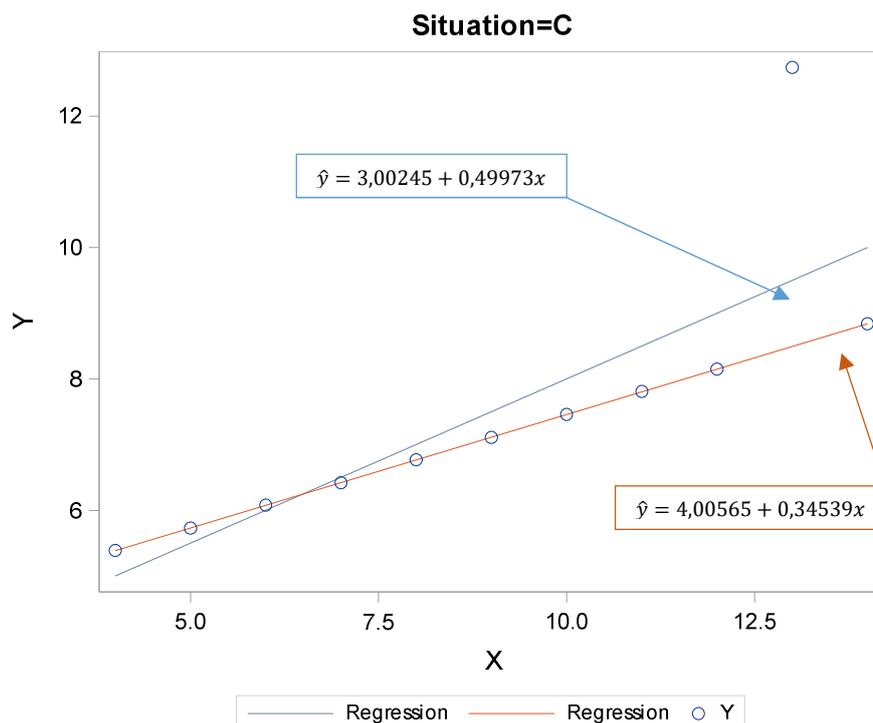


Figura 39. Efeito de um ponto influente no ajuste de duas retas de regressão. A linha azul representa valores estimados da regressão considerando o Outlier nos dados, enquanto a linha laranja representa os valores estimados sem a presença do Outlier no ajuste da regressão.

Decidir pela exclusão de pontos incomum da base de dados não é tarefa simples, visto que, essas observações podem ser altamente informativas e, conseqüentemente, não podem ser excluídas automaticamente sem uma análise prévia justificada.

Se a informação é um dado errado, esta poderia ser corrigida facilmente. Para tal, é necessário conferir as planilhas de campo. Mas se for realmente uma situação que ocorre em campo, esta deve ser determinada o quão influencia no processo de modelagem e deve-se considerar critérios estatísticos para a tomada de decisão quanto a sua exclusão.

A detecção de pontos incomuns em análise de regressão linear simples pode ser realizada, em parte, por um gráfico de dispersão como na análise da Figura 39. Entretanto, um valor detectado como outlier em um gráfico de dispersão pode não influenciar os coeficientes de regressão ou vice-versa.

O uso de estatísticas para detecção de observações influentes é outra opção em que se calculam valores que apoiam a tomada de decisão no processo de construção de modelos de regressão simples e regressão múltipla. As principais estatísticas de diagnóstico utilizadas em análise de regressão são descritas no Quadro 25.

Quadro 25. Estatísticas para detectar observações influentes na análise de regressão linear simples ou múltipla. Adaptado de BELSLEY et al. (1980).

Estatística	Descrição	Diagnóstico
$RStudent_i = \frac{\varepsilon_i}{s_{(i)}\sqrt{1-h_i}}$	Resíduos estudentizados considerando a variância sem a <i>i</i> -ésima observação influente ($s_{(i)}^2$) como forma de avaliar influência da observação excluída dos dados. Considera-se as propriedades da distribuição <i>t</i> de Student com $n - k - 1$ grau de liberdade.	Valores de resíduos estudentizados maiores do que $ 3 $ merecem atenção e podem ser excluídos da base de dados, mas essa decisão deve ser realizada com cautela.
$DFBetas_{ik} = \frac{\beta_k - \beta_{k(i)}}{s_{(i)}/\sqrt{SQres_k}}$	Utilizado quando existe observações influentes no conjunto de dados e se deseja determinar em qual <i>k</i> coeficiente de regressão a observação influente está impactando. O impacto é medido na mudança do erro padrão do <i>k</i> coeficiente de regressão caso seja excluída uma <i>i</i> -ésima observação influente da base de dados.	Valor entre $\pm 2/\sqrt{n}$ indica que existem observações que são influentes na estimativa do respectivo coeficiente de regressão.
$C_i = \left(\frac{\varepsilon_i^2}{p+1} \right) \cdot \left(\frac{h_i}{1-h_i} \right)$	O valor da distância de Cook (C_i) representa a medida do grau para o qual os valores estimados mudam se a observação estranha é eliminada da base de dados. Portanto, <i>C</i> mede a influência no modelo de regressão como um todo.	Neter et al. (1996) sugeriram investigar qualquer ponto com valor de $C_i > 1$ ou $> 4/n$.
$DFFits_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{s_{(i)}\sqrt{h_i}}$	Medida de mudança no valor predito para a <i>i</i> -ésima observação calculada a partir da eliminação de uma <i>i</i> -ésima observação influente.	Valores entre $ 2 \cdot \sqrt{\left(\frac{k}{n}\right)} $ indicam que a observação é relativamente influenciável.

Sendo: ε_i = *i*-ésimo resíduo; *S* = desvio padrão; $s_{(i)}$ = desvio padrão estimado sem a *i*-ésima observação influente; h_i = valor de Leverage; $SQres_k$ = Soma de Quadrado dos Resíduos obtida a partir da regressão da variável dependente associada ao *k* coeficiente de regressão com as variáveis independentes resultantes (sem excluir a observação).

O procedimento PROC REG calcula as estatísticas propostas por Besley et al. (1980) e descritas no Quadro 25, para medir a influência de cada observação nas estimativas. Para tal, basta adicionar a opção INFLUENCE na linha de programação MODEL conforme indicado no programa SAS com dados de crescimento periódico anual em área basal (ipag) de árvores de cedro (*Cedrela odorata* L.). Incluiu-se a opção **id** seguido da variável código para o SAS associar cada critério estatístico à variável que, neste caso, representa a junção no número da árvore e da capoeira a qual foi medida a respectiva árvore de cedro.

```
data cedrela;
  input arvore capoeira$ código$ D hegyi ipag;
datalines;
1 A 1A 18.1 2.285 108.802
2 A 2A 6.6 5.473 16.930
3 A 3A 6.1 7.502 20.281
.
.
.
56 J 56J 12.7 2.866 .
57 J 57J 23.8 1.174 207.133
58 J 58J 31.4 2.529 139.212
;
proc reg data= cedrela plots=none;
  id codigo;
  model ipag=D hegyi / influence;
run;
```

O resultado do processamento do PROC REG é apresentado no Output 28 contendo apenas a tabela dos critérios para medir a influência de cada observação nas estimativas somente das primeiras 20 observações do conjunto de dados do caso florestal 6 (para resumo da apresentação).

Output 28. Critérios estatísticos para análise de observações influentes.

Output Statistics									
Obs	codigo	Residual	RStudent	Hat DiagH	Cov Ratio	DFFITS	DFBETAS		
							Intercept	D	Hegyí
1	1A	11.0789	0.0973	0.0287	1.1041	0.0167	0.0133	-0.0077	-0.0064
2	2A	6.3544	0.0564	0.0487	1.1278	0.0128	0.0100	-0.0086	-0.0012
3	3A	13.2835	0.1181	0.0523	1.1312	0.0277	0.0155	-0.0143	0.0059
4	4A	26.1860	0.3463	0.5704	2.4765	0.3990	-0.1575	0.1018	0.3752
5	1B	136.1576	1.2257	0.0423	1.0084	0.2575	-0.0239	0.1431	-0.0173
6	2B	-96.4034	-0.8636	0.0497	1.0711	-0.1974	0.0452	-0.1311	-0.0101
7	3B	-65.7969	-0.5801	0.0283	1.0783	-0.0990	-0.0570	0.0118	0.0455
8	4B	-100.4647	-0.8942	0.0362	1.0522	-0.1732	-0.0033	-0.0795	0.0224
9	5B	2.9159	0.0256	0.0251	1.1008	0.0041	0.0009	0.0011	-0.0003
10	6B	319.7181	3.1299	0.0358	0.5934	0.6032	0.0775	0.2170	-0.1502
11	7B	-1.4903	-0.0130	0.0220	1.0972	-0.0020	-0.0006	-0.0002	-0.0001
12	8B	-44.4428	-0.3927	0.0369	1.1020	-0.0768	-0.0636	0.0355	0.0476
13	9B	-36.2335	-0.3203	0.0387	1.1082	-0.0642	0.0066	-0.0360	0.0003
14	10B	-213.6954	-2.0389	0.0954	0.8938	-0.6622	0.3023	-0.5466	-0.1160
15	11B	35.0551	0.3549	0.2669	1.4508	0.2142	-0.1420	0.2013	0.0710
16	12B	56.8000	0.5011	0.0317	1.0886	0.0906	0.0150	0.0293	-0.0200
17	13B	468.7442	5.5393	0.0715	0.2230	1.5367	-0.5850	1.1909	0.2190
18	14B	-52.6696	-0.4662	0.0387	1.0992	-0.0935	-0.0840	0.0558	0.0520
19	15B	-125.6399	-1.1234	0.0347	1.0172	-0.2130	-0.0011	-0.1000	0.0187
20	16B	68.8799	0.6142	0.0492	1.0988	0.1397	0.1319	-0.0953	-0.0861
...

Entre as observações apresentadas na tabela do Output 28, a árvore 13 localizada na capoeira B (código 13B) apresenta um DFBetas de 1,1909 para o coeficiente associado à variável independente d . Neste caso, considera-se que a árvore 13B no ajuste da regressão acarreta em um aumento de 1,19 no erro padrão do coeficiente β_1 a mais do que se a observação fosse excluída para um novo ajuste.

A demonstração do cálculo do DFBetas para a observação 17 (código 13B) é a seguinte:

β_D =coeficiente de regressão para a variável d considerando a observação 17 no ajuste da regressão=7,60886;

$\beta_{D(17)}$ = coeficiente de regressão para a variável d excluindo a observação 17 no ajuste da regressão = 6,42580;

$s_{(i)}$ = desvio padrão estimado excluindo a observação 17 no ajuste da regressão = 87,81836;

SQR_k = Soma de Quadrado dos Resíduos obtida a partir da regressão considerando a variável dependente como o d e a variável independente resultante Hegyi (sem excluir a observação 17) = 7814, 25063.

$$DFBetas_{17,DAP} = \frac{\beta_D - \beta_{D(17)}}{\frac{s_{(17)}}{\sqrt{SQR_D}}} = \frac{7,60886 - 6,42580}{\frac{87,81836}{\sqrt{7814,25063}}} = \frac{1,18306}{\frac{87,81836}{88,39826}} = \frac{1,18306}{0,99344} = 1,1909$$

A decisão de excluir a observação 17 do conjunto de dados pode ser de acordo ao ponto de corte estabelecido para o DFBetas como valor limite de $\pm \frac{2}{\sqrt{n}} = 0,29488$. Neste caso, o valor de 1,19 excede o limite e a observação é considerada influente visto que, houve uma influência de 1,19 unidades no coeficiente de regressão para a variável d .

O gráfico de resíduos estudentizados e a distância de Cook indicam claramente que a observação 17 (árvore 13B) é um ponto influente. Neste caso, a influência da observação 17 é avaliada para o modelo de regressão como um todo o que difere do critério DFBetas que mede o impacto de uma observação influente em cada coeficiente de regressão.

Portanto, caso o objetivo da pesquisa com análise de regressão linear seja o entendimento da relação entre a variável dependente y e as variáveis independentes x 's, a magnitude e os sinais dos coeficientes de regressão serão considerados para as conclusões. Neste caso, uma análise de pontos influentes com impacto nos coeficientes de regressão deve ser considerada (DFBetas).

Outra opção para análise de observações influentes é a avaliação de gráficos personalizados construídos por demanda a partir da opção PLOTS= do procedimento reg. A sintaxe a seguir inclui a opção PLOTS=(RSTUDENTBYLEVERAGE(LABEL) COOKSD(LABEL) DFBETAS(LABEL)).

```

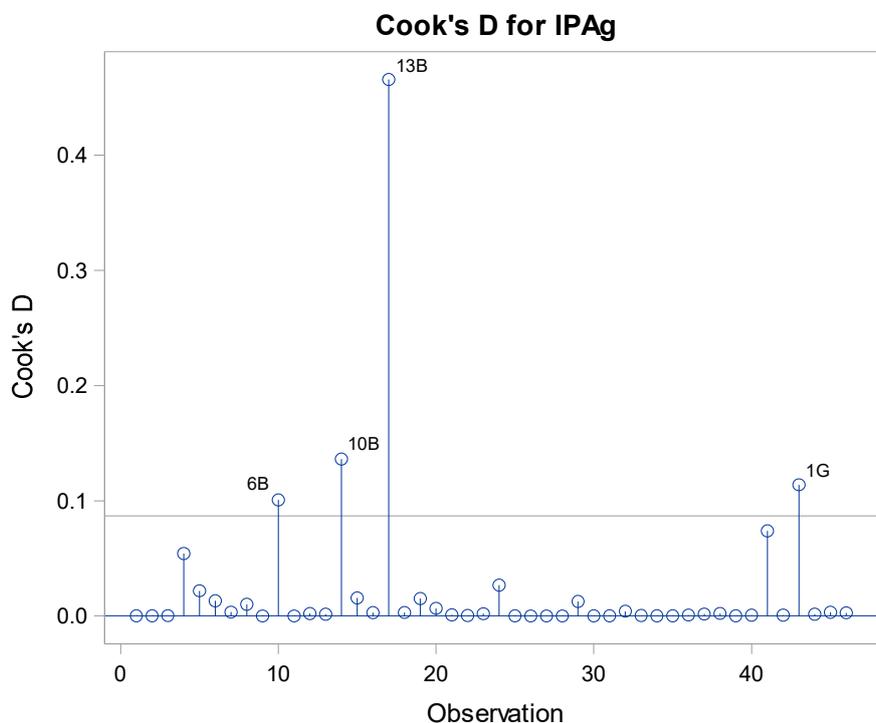
proc reg data= cedrela plots=(RStudentByLeverage(label) CooksD(label) DFBetas(label));
  id codigo;
  model ipag=D hegyi;
run;

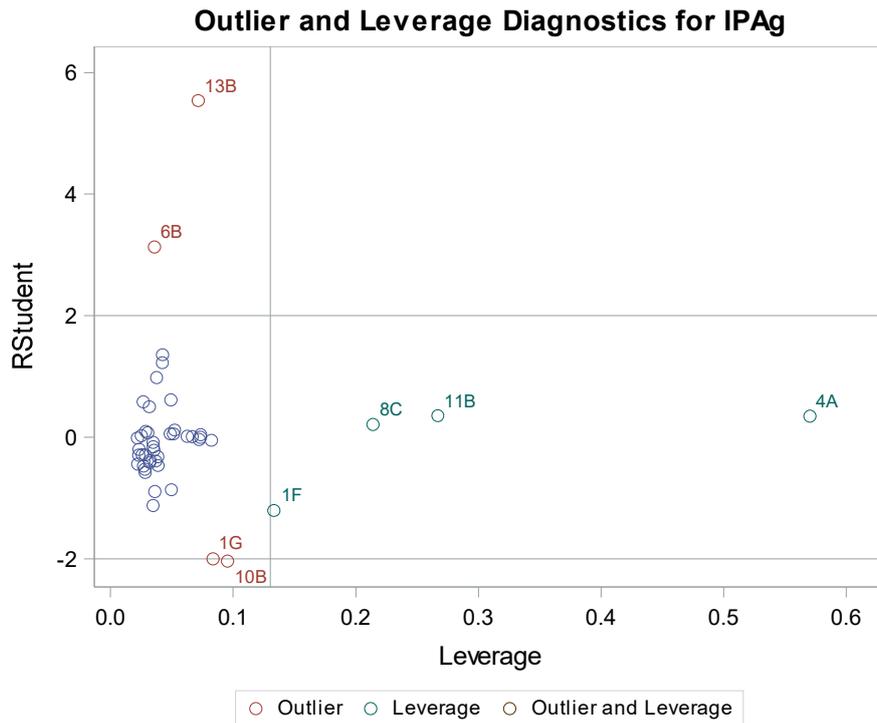
```

Os gráficos solicitados ao SAS são apresentados no Output 29 e revelam que existem quatro outliers de acordo com a distância de Cook e do gráfico de resíduos estudentizados. Os outliers identificados em ambos os gráficos são as árvores 13B, 10B, 1G e 6B.

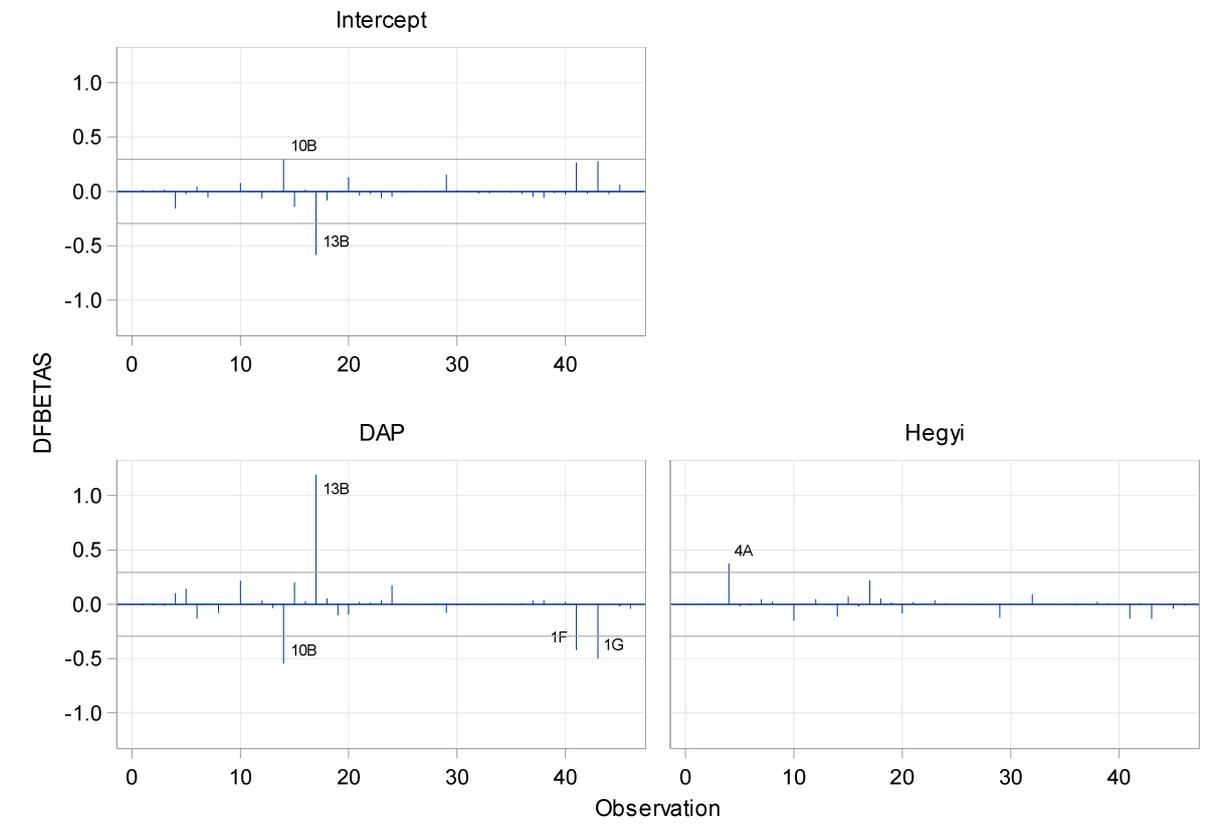
Ademais, algumas observações influentes foram identificadas em cada coeficiente de regressão (gráficos de DFBetas) que indicam além de outras, a árvore 1F para a variável *d* e 4G para a variável *Hegy*.

Output 29. Gráficos para avaliação e detecção de observações influentes.



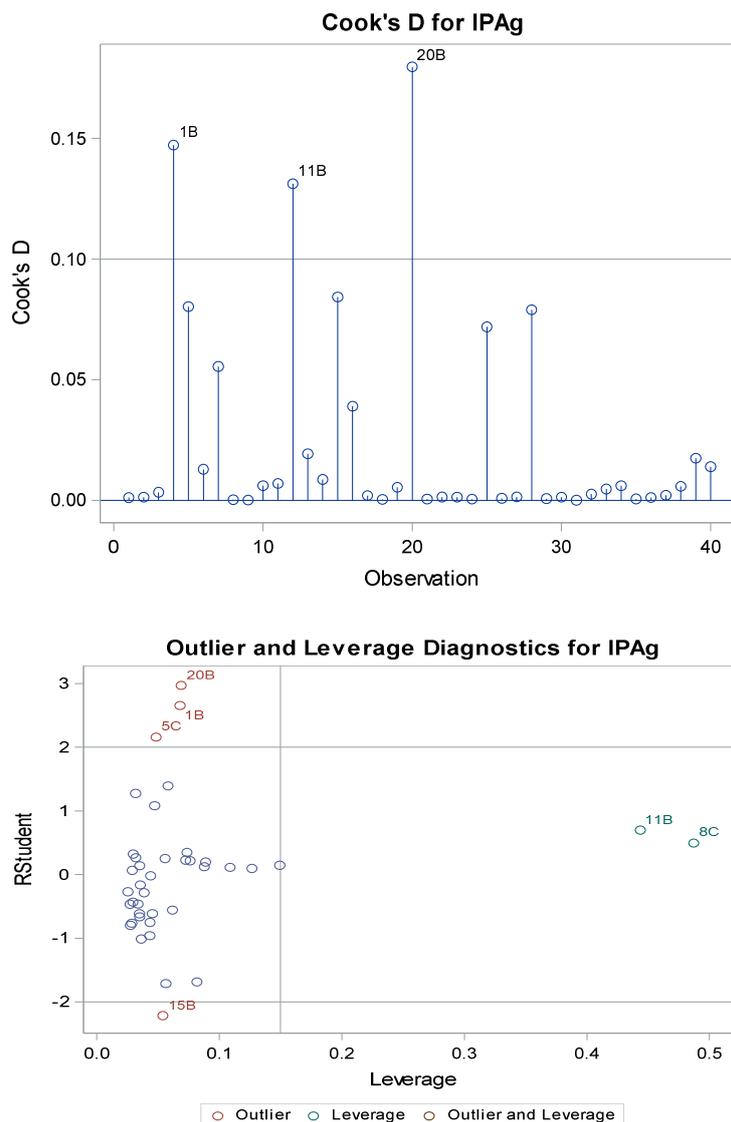


Influence Diagnostics for IPAg



Após o reajuste do modelo de regressão excluindo as observações influentes e os outliers identificados, os resultados mostram novos pontos influentes e outliers identificados conforme o Output 30.

Output 30. Gráficos personalizados de distância de Cook e resíduos estudentizados após reajuste do modelo de regressão sem as observações influentes 13B, 10B, 1G, 6B, 1F e 4A.



O fato de permanecer vários pontos influentes no conjunto de dados pode estar associado a uma inadequação do modelo para descrever a variação observada na variável dependente (incremento em área basal).

Portanto, é recomendado que se realize, inicialmente, uma análise da distribuição dos resíduos para apoiar a construção de modelos de regressão. A Figura 40 mostra um painel de gráficos de resíduos para o modelo de regressão em questão e revela um padrão na distribuição dos resíduos bem como outros problemas com o modelo considerado.

Conseqüentemente, para prosseguir com a análise de pontos influentes e outliers primeiramente recomenda-se uma avaliação e remediação das condicionantes de regressão.

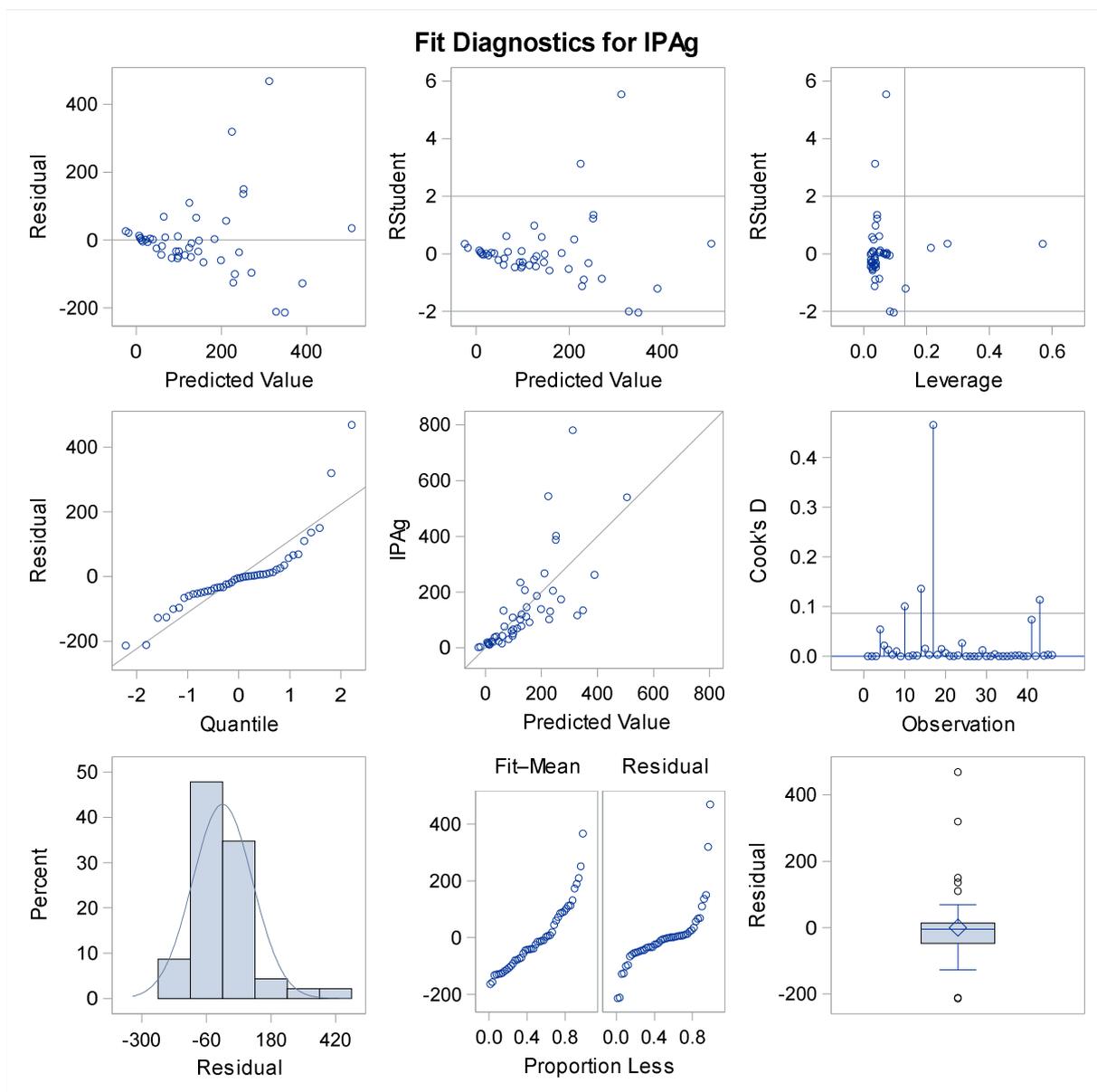


Figura 40. Padrão na distribuição dos resíduos em função das variáveis independentes revelando problemas de condicionantes de regressão.

3.5.2.3. Como remediar não atendimento as condicionantes de regressão?

O não atendimento a condicionante de Normalidade não é um problema caso a amostra seja razoavelmente grande visto que o Teorema do Limite Central garante que os coeficientes de regressão estimados terão uma distribuição aproximadamente Normal mesmo se ε_i não tenha distribuição Normal. Portanto, assegura-se o uso de valores-p e intervalo de confiança calculados a partir da tabela Normal.

Por outro lado, ao remediar a condicionante de Normalidade a condicionante de Homocedasticidade é influenciada. Por esta razão, ambas são remediadas em conjunto, visto que, geralmente a alternativa utilizada para normalizar os resíduos também os torna com variâncias homogêneas.

Neste sentido, Neter et al. (1996) recomendaram que a transformação seja realizada primeiramente para estabilizar a variância dos resíduos e em seguida, verificar se a transformação também normalizou os resíduos.

A transformação da variável dependente para a escala de logaritmo é uma das alternativas mais comuns em análise de regressão. A outra opção é a determinação de uma potência como a raiz quadrada.

Outra forma de transformar a variável dependente é determinar uma potência de forma automatizada e que corrige a assimetria e a variância heterogênea da distribuição dos resíduos.

O procedimento de transformação Box-Cox (BOX; COX, 1964) identifica a transformação adequada para a variável dependente (y') a partir de uma família de potências lambda (λ) de acordo a forma $y' = y^\lambda$. É importante destacar que a transformação da variável dependente é realizada para normalizar os resíduos obtidos a partir do modelo de regressão. A forma simples de transformação é apresentada no Quadro 26.

Quadro 26. Valores de lambda para potenciação da variável dependente e sua respectiva indicação de uso.

Parâmetro Lambda (λ)	Potência para y	Indicação de uso a partir do comportamento dos resíduos
$\lambda = 2$	$y' = y^2$	Transformação quadrada de y , recomendada quando o aumento/incremento da variância dos resíduos não ocorre rapidamente com o aumento de x .
$\lambda = 1$	$y' = y^1$	Nenhuma transformação recomendada para os dados de y .
$\lambda = 0,5$	$y' = \sqrt{y}$	Transformação raiz quadrada de y , recomendada quando a variância dos resíduos incrementa repentinamente com aumento de x .
$\lambda = 0$	$y' = \text{Log}_e y$	Transformação logaritmo natural ¹ de y (por definição). Recomendada quando a variância dos resíduos incrementa repentinamente com aumento de x . Neste caso, para compensar a discrepância logarítmica para equações em que y foi transformada para logaritmo natural (Ln) é necessário utilizar o fator de correção (FC) proposto por Sprugel, (1983): $FC = \text{Exp} \left[\frac{(S_{yx})^2}{2} \right]$ Exp=exponencial; S_{yx} = erro padrão da estimativa do modelo logaritmo. Para obter o valor correto do fator de correção, Sprugel, (1983) recomenda que S_{yx} seja calculado a partir do modelo em que y seja transformado com logaritmo natural. Caso y seja transformado com logaritmo base 10, S_{yx} deve ser convertido multiplicando-o por 2,303.
$\lambda = -0,5$	$y' = \frac{1}{\sqrt{y}}$	Transformação recíproca da raiz quadrada de y , recomendada quando a variância dos resíduos decresce com o aumento de x .
$\lambda = -1$	$y' = \frac{1}{y}$	Transformação recíproca de y , recomendada quando a variância dos resíduos decresce com o aumento de x .

¹= Também pode ser utilizado a transformação para logaritmo base 10. O uso do logaritmo base 10 em vez de logaritmo natural não afeta a transformação logarítmica. Ademais, a interpretação da variável transformada para logaritmo base 10 é melhor do que para base natural.

Para ilustrar o uso da transformação Box-Cox vamos considerar o resultado da análise de resíduos para o caso florestal 6. Os gráficos do Output 31 do capítulo anterior mostram a falta de aderência dos resíduos à distribuição Normal conforme demonstrado no tópico sobre avaliação das condicionantes de regressão.

O procedimento PROC TRANSREG do SAS realiza a transformação da variável ipag mediante a seguinte sintaxe.

```
data cedrela;
  input arvore capoeira$ código$ D hegyi ipag;
datalines;
1 A 1A 18.1 2.285 108.802
2 A 2A 6.6 5.473 16.930
3 A 3A 6.1 7.502 20.281
.
.
.
56 J 56J 12.7 2.866.
57 J 57J 23.8 1.174 207.133
58 J 58J 31.4 2.529 139.212
;

%let Xvars = D hegyi;
%let Yvar = ipag;
proc transreg data= cedrela ss2 details plots=(boxcox);
  model BoxCox(&Yvar / convenient lambda=-2 to 2 by 0.1) = identity(&Xvars);
  output out=TransOut residual;
run;
```

Na linha MODEL do procedimento PROC TRANSREG é declarado o modelo de regressão desejado com a variável dependente (y) as variáveis independentes (Ambas em modo macro do SAS para facilitar a modelagem). Também inclui opções para determinar quais os valores do parâmetro lambda serão consideradas para a transformação.

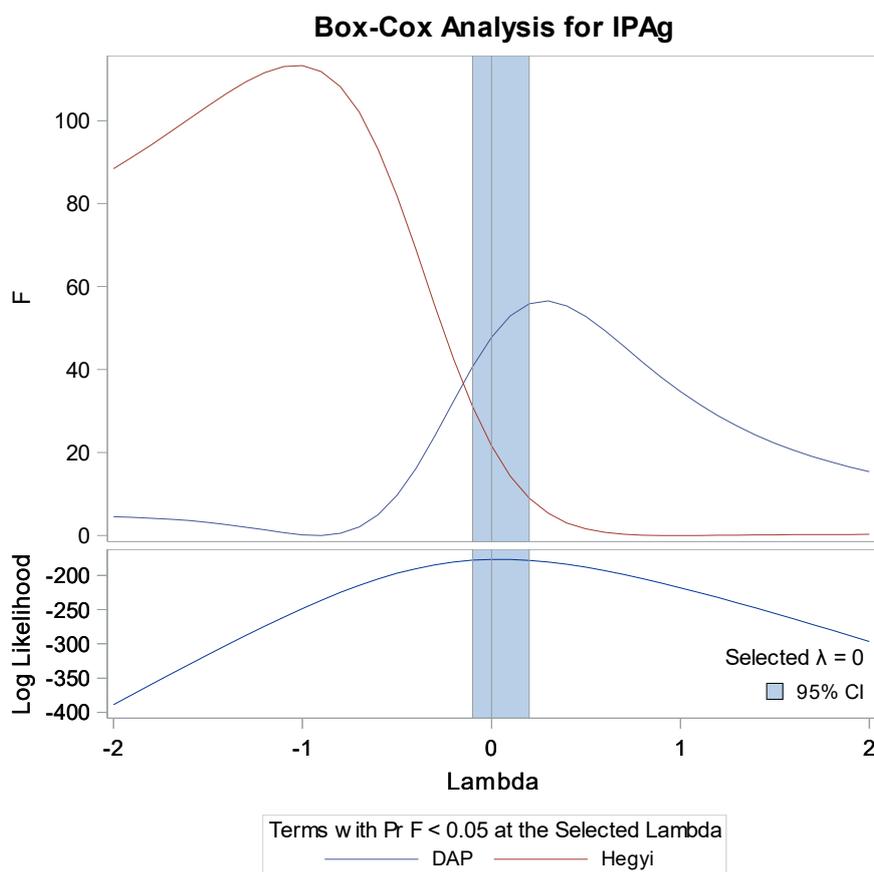
A opção CONVENIENT LAMBDA solicita ao SAS a indicação de um valor próximo de lambda para o qual a transformação da variável dependente (y) seja de fácil entendimento e interpretação. Desta forma, o SAS pode indicar um $\lambda = -0,5$ em vez de um

$\lambda = -0,63$, ou um $\lambda=0$ em vez de um $\lambda=0,18$. Para isso, o SAS considera o intervalo de confiança para λ .

O intervalo considerado para a interação foi de -2 a 2 em intervalos de 0,1 unidades resultando em 41 modelos de regressão ajustados sendo um para cada lambda estabelecido. Quando a opção lambda= não é utilizada no programa, o SAS considera o intervalo entre -3 e 3 com intervalo de 0,25 unidades por padrão.

Para resumir os resultados de ajuste dos 41 modelos de regressão, o SAS apresenta um painel com dois gráficos apresentados no Output 31.

Output 31. Gráfico Box-Cox com a indicação da potência adequada para a transformação da variável ipag.



O gráfico inferior do painel mostra a função de verossimilhança para cada valor de lambda examinado (para a normalidade dos resíduos). A função é maximizada para um valor de lambda zero (0). A banda azul em ambos os gráficos mostra o intervalo de confiança.

Por sua vez o gráfico superior do painel mostra o valor da estatística F para cada uma das variáveis independentes dos 41 modelos de regressão solicitados. Quanto maior o valor de F , melhor será o lambda utilizado. Neste caso, o processo de transformação Box-Cox determinou que a soma de quadrados dos resíduos (SQR) obteve um menor valor quando o modelo de regressão ajustado considerou um lambda zero. Neste caso, o modelo logaritmo é o recomendado.

Para fins de verificação ao atendimento à condicionante de normalidade, solicitou-se ao SAS a construção de gráficos de histograma, Q-Q Plot e P-P Plot conforme mostra a Figura 41.

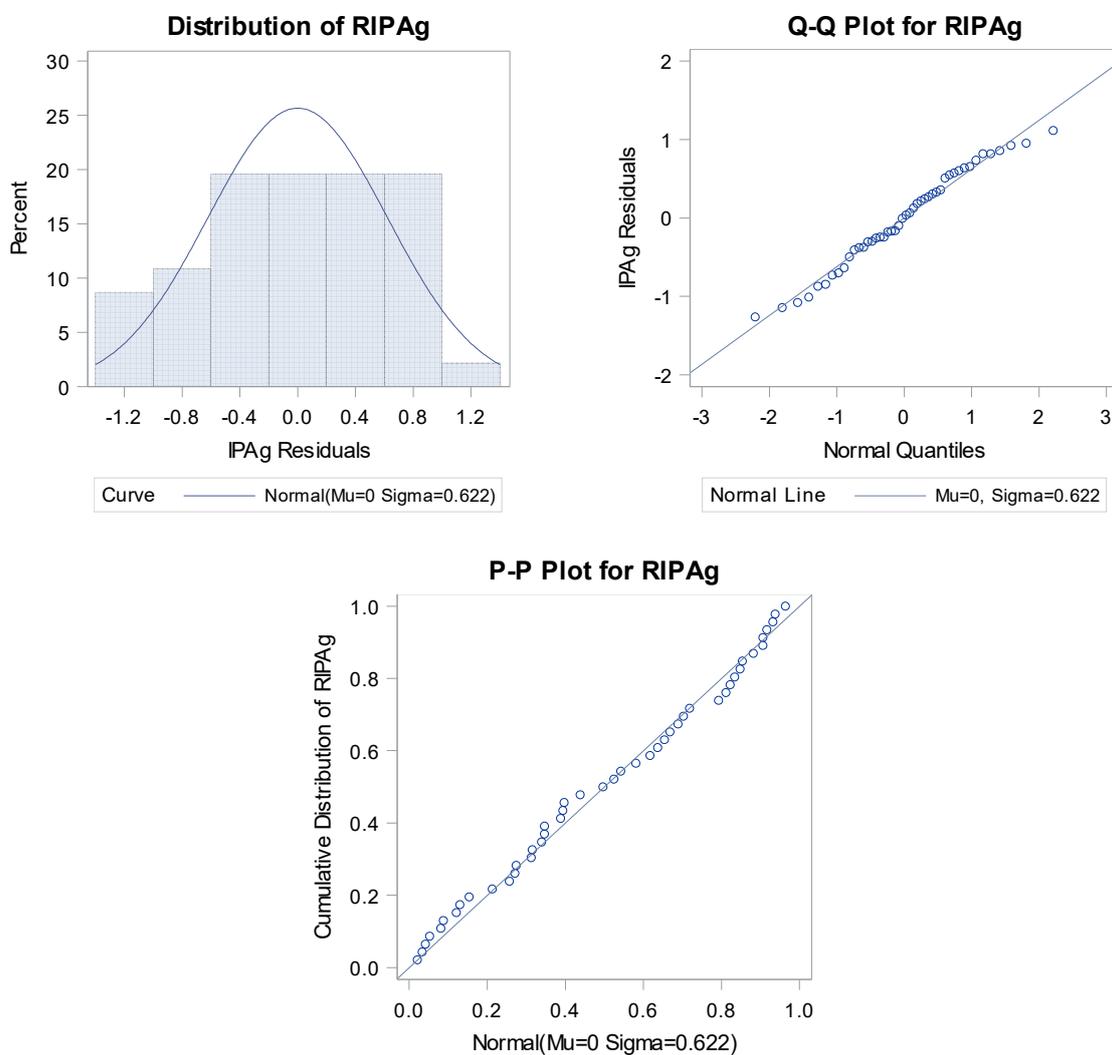


Figura 41. Comportamento normalizado dos resíduos obtidos a partir do modelo de regressão logaritmo ($\lambda=0$).

Outras formas de corrigir problemas com o não atendimento às condicionantes de regressão, como variâncias heterogêneas podem ser utilizadas na modelagem. Uma opção é utilizar mínimos quadrados ponderados. No SAS é possível utilizar alguns procedimentos que consideram dados com variâncias heterogêneas e resíduos correlacionados a partir de observações repetidas, na modelagem como:

- i) PROC GENMOD ou PROC GLIMMIX utilizando a opção DIST= para indicar a distribuição apropriada;
- ii) PROC MIXED com a opção GROUP= que possibilita definir os efeitos de heterogeneidade na estrutura da matriz de covariância.

Para dados coletados a partir de séries temporais, medidas repetidas na mesma unidade amostral com correlação entre as observações e, conseqüentemente, nos resíduos, é possível utilizar outros procedimentos do SAS que consideram observações correlacionadas. Para dados obtidos a partir de séries temporais pode-se utilizar o procedimento PROC AUTOREG ou PROC ARIMA do SAS.

O não atendimento à condicionante de homoscedasticidade produz estimativas não consistentes para o erro padrão dos coeficientes de regressão influenciando diretamente nos testes estatísticos. Uma alternativa é ajustar o modelo com resíduos heterogêneos considerando estimador de covariância consistente de heterocedasticidade (Heterocedasticity Consistent Covariance). Esse método produz estimativas consistentes para o erro padrão e pode ser solicitado no PROC REG do SAS por meio da opção HCC conforme sintaxe a seguir:

```
proc reg data=dados;  
  model y=x / hcc;  
run;
```

3.5.3. Colinearidade

A colinearidade se refere à associação linear entre duas variáveis independentes (x 's) presentes no modelo de regressão. Também ocorre colinearidade quando uma variável independente é resultante da combinação linear de outras variáveis no modelo.

Portanto, em modelos de regressão em que novas variáveis são criadas a partir da combinação ou transformação de variáveis existentes, pode ocorrer colinearidade. O termo também é conhecido como multicolinearidade no caso em que duas ou mais variáveis independentes sejam correlacionadas em um modelo de regressão.

A Figura 42 que mostra instabilidade de um modelo de regressão quando existe a presença de variáveis independentes altamente correlacionadas entre si (gráfico da esquerda). Neste caso, a instabilidade refere-se ao fato de que qualquer plano não pode ser sustentado quando apoiado sob os pontos situado ao longo dos eixos de dispersão como visto no gráfico esquerdo da Figura 42.

Por outro lado, um plano poderia ser bem apoiado sob os pontos situados ao longo dos eixos de dispersão do gráfico a direita da Figura 42 quando as variáveis x_1 e x_2 não são correlacionadas (Ortogonais).

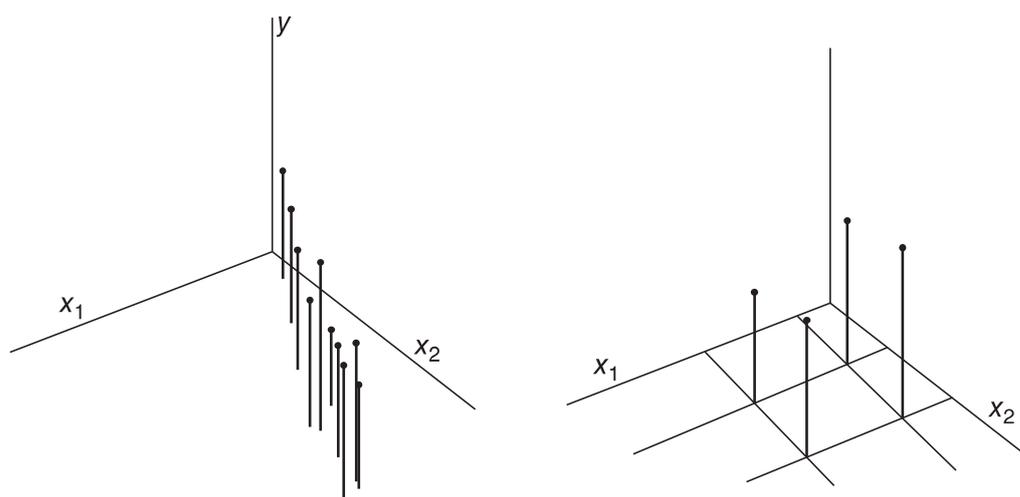


Figura 42. Ilustração indicando a presença (esquerda) e ausência (direita) de multicolinearidade entre as variáveis independentes x_1 e x_2 em um modelo de regressão de acordo a MONTGOMERY et al. (2012).

A colinearidade não é considerada como uma condicionante de regressão. Ademais, as estimativas dos coeficientes de regressão por mínimos quadrados ordinários não são afetadas quando existe colinearidade ou multicolinearidade.

Os efeitos da colinearidade podem causar sérios problemas em procedimentos de construção de modelos a partir de seleção automática de variáveis independentes (por exemplo, Stepwise) considerando o nível de significância como fator para inclusão e

exclusão de variáveis. O modelo final pode não conter variáveis importantes devido a que sua significância pode ser mascarada pelo efeito da colinearidade.

Esta é uma justificativa plausível para primeiramente analisar a colinearidade antes de realizar a seleção automática de variáveis por algoritmos.

3.5.3.1. Efeitos da multicolinearidade

Quando duas ou mais variáveis independentes são correlacionadas (não há ortogonalidade) a variância dos coeficientes de regressão das variáveis correlacionadas é afetada o que pode causar os seguintes problemas na modelagem:

- 1) Aumento do erro padrão dos coeficientes de regressão e, conseqüentemente, limites de predição com maior abrangência do que deveria ser;
- 2) Presença de colinearidade pode ocultar a relação verdadeira entre a variável dependente (y) e independente (x). Quando duas variáveis independentes são correlacionadas e utilizadas no modelo de regressão, ambas podem ser estatisticamente não significativas, mas quando usadas em separado podem ter significância na modelagem;
- 3) A colinearidade pode causar mudança no sinal algébrico dos coeficientes de regressão causando conflito na interpretação biológica dos coeficientes.

3.5.3.2. Diagnóstico da multicolinearidade

No SAS a detecção de colinearidade é facilmente acessada no procedimento PROC REG por meio do uso das opções:

- i) VIF → Mede a magnitude da colinearidade considerando o fator de inflação da variância (VIF) calculado para cada variável independente do modelo;
- ii) COLLIN → Inclui o intercepto na avaliação do VIF. Ademais, fornece informações necessárias para identificar o conjunto de variáveis independentes fonte do problema;
- iii) COLLINOINT → Exclui o intercepto na avaliação do VIF.

O VIF mede o incremento da variância devido à colinearidade. É calculado para cada variável do modelo considerando o coeficiente de determinação (R_i^2) obtido quando a

i -ésima variável independente é modelada em função das demais: A fórmula do VIF é a seguinte:

$$VIF_i = \frac{1}{1 - R_i^2}$$

Desta forma, para um modelo de regressão linear múltiplo com três variáveis independentes x_1 , x_2 e x_3 , o VIF calculado para x_2 , por exemplo, será obtido da seguinte forma:

- Ajustar o modelo de regressão e obter o coeficiente de determinação R^2 considerando a variável x_2 como dependente: $x_2 = \beta_0 + \beta_1 x_1 + \beta_2 x_3$.
- Substituir o coeficiente de determinação (R_i^2) com o resultado obtido após o ajuste do modelo $x_2 = \beta_0 + \beta_1 x_1 + \beta_2 x_3$ e calcular o VIF para a variável x_2 .

Portanto, caso uma variável independente possua um valor de VIF maior do que 10 é um indicativo de que a colinearidade está influenciando na modelagem por mínimos quadrados ordinários (NETER et al., 1996). Na prática, a fórmula do VIF indica que para uma variável independente causar problemas de colinearidade, esta deve ter um coeficiente de determinação acima de 0,9.

Podemos visualizar na prática os três problemas causados pela colinearidade mencionados no subcapítulo 5.3.1 quando o modelo de volume de Meyer (LOETSCH et al., 1973) é ajustado para dados de cubagem rigorosa de árvores. O modelo de Meyer tem a seguinte expressão matemática:

$$v_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2 + \beta_3 D_i h_{fi} + \beta_4 D_i^2 h_{fi} + \beta_5 h_{fi} + \varepsilon_i$$

Em que:

v_i =Volume (m^3) da i -ésima árvore;

D_i =Diâmetro a 1,3 m do solo (cm) da i -ésima árvore;

h_{fi} =Altura do fuste (m) da i -ésima árvore.

A sintaxe do PROC REG para ajustar o modelo de Meyer e calcular o VIF é a seguinte:

```
proc reg data= volume;
  model v = D D2 Dh_m D2h_m h_m / vif;
run;
```

O valor do VIF é mostrado na tabela de coeficientes de regressão juntamente com os valores dos coeficientes conforme Output 32.

Output 32. Ajuste do modelo de Meyer com valores de VIF (Variance Inflation Factor) para cada variável independente solicitado pela opção VIF na declaração MODEL. As demais tabelas e gráficos foram omitidos intencionalmente.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	0.03694	0.00739	68.51	<.0001
Error	34	0.00367	0.00010782		
Corrected Total	39	0.04060			

Root MSE	0.01038	R-Square	0.9097
Dependent Mean	0.09348	Adj R-Sq	0.8964
Coeff Var	11.10769		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	1	0.10362	0.51667	0.20	0.8422	0
D	1	-0.01641	0.07790	-0.21	0.8344	9611.43754
D2	1	0.00101	0.00290	0.35	0.7308	9714.13163
Dh_m	1	0.00095116	0.00796	0.12	0.9056	21456
D2h_m	1	-0.00002028	0.00029845	-0.07	0.9462	12222
h_m	1	-0.00593	0.05249	-0.11	0.9107	3626.94042

A tabela Anova mostra que o modelo é significativo (p-valor <0,0001). Em seguida a tabela de critérios de bondade de ajuste indica que o modelo explica 91% da variação observada no volume rigoroso da amostra de árvores ($R^2=0,9097$) com coeficiente de variação considerado baixo ($CV=11,1$).

A última tabela mostra os valores estimados para os coeficientes de regressão bem como o valor do fator de inflação da variância (*VIF*) para cada variável independente do modelo. Todas variáveis apresentam valor de *VIF* maior do que o valor limite de 10. A variável com maior valor de *VIF* é *Dh_m* seguido da variável *D2h_m*.

Um detalhe importante na tabela dos coeficientes de regressão é o valor-p indicando que todos os coeficientes do modelo de Meyer são não-significativos. Isso equivale a que cada um dos coeficientes tem valor zero (0). Entretanto, a tabela Anova indica o contrário!

Isso é ocasionado pela inflação dos valores do erro padrão (Standard Error) devido a presença da colinearidade afetando diretamente o cálculo do valor *t* (*t Value*) que, conseqüentemente, influencia o valor-p.

Outro detalhe se nota no sinal algébrico negativo associado aos coeficientes de regressão para o diâmetro e a altura indicando relação contrária com o volume das árvores.

3.5.3.3. Alternativas para reparar o efeito da multicolinearidade

Quando a colinearidade é detectada é comum eliminar as variáveis independentes do modelo e, neste caso, o modelo é reajustado novamente. Entretanto, esse procedimento pode desconsiderar uma ou mais variáveis relevantes para o processo de modelagem.

No caso do Output 33 o modelo deveria ser reajustado após a eliminação das variáveis com maiores valores de *VIF*. Entretanto, deve-se considerar que algumas variáveis possuem importância biológica na estimativa do volume de árvores individuais como no caso de diâmetro e altura comercial.

Desta forma, procedeu-se a eliminar todas variáveis criadas a partir do diâmetro e altura comercial reajustando o modelo novamente de acordo à sintaxe do PROC REG.

```
proc reg data= volume;  
  model v = D h_m / vif;  
run;
```

O Output 33 mostra que o modelo continua significativo ($Pr>F = <0,0001$) com praticamente os mesmos valores para a bondade de ajuste com redução do valor *VIF* para abaixo do nível aceitável de colinearidade.

Outro detalhe importante é a significativa redução do erro padrão para os coeficientes de regressão para d e h_m tornando-os significativos bem como mudando o sinal para positivo para ambos.

Output 33. Resultado do ajuste do modelo sem as variáveis criadas a partir do diâmetro e altura comercial.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.03649	0.01825	164.29	<.0001
Error	37	0.00411	0.00011106		
Corrected Total	39	0.04060			

Root MSE	0.01054	R-Square	0.8988
Dependent Mean	0.09348	Adj R-Sq	0.8933
Coeff Var	11.27312		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	1	-0.12665	0.01472	-8.60	<.0001	0
D	1	0.01470	0.00081238	18.09	<.0001	1.01486
h_m	1	0.00293	0.00089107	3.28	0.0022	1.01486

Outra alternativa é utilizar análise de regressão de cumeeira (Ridge) que é uma técnica enviesada, mas permite reduzir a variância dos coeficientes. Também é possível utilizar a regressão por componentes principais.

3.6. Seleção de variáveis para construção de modelos de regressão

Objetivos de aprendizagem desse tópico:

- i) Descrever as ferramentas utilizadas para seleção de variáveis para compor modelos de regressão;

- ii) Descrever as técnicas de seleção de modelos: Forward, Backward e Stepwise;
- iii) Obter um modelo de regressão candidato utilizando o procedimento GLMSELECT;
- iv) Descrever os critérios de seleção utilizados para selecionar variáveis para o modelo e avaliar os modelos construídos;
- v) Construir modelo de regressão preditiva.

Estudos observacionais na área de crescimento e produção florestal geralmente produzem grande quantidade de variáveis preditoras (independentes) a fim de explicar a grande variação de dados de crescimento de árvores.

Por exemplo, em estudos de avaliação do crescimento de árvores individuais em florestas naturais, uma grande quantidade de variáveis é necessária para explicar o crescimento em diâmetro.

Determinar qual modelo de regressão seria o mais adequado diante de inúmeras variáveis independentes é um desafio, visto que, à medida que aumenta o número de variáveis, a quantidade de modelos candidatos aumenta na proporção de 2^k , sendo que k = número de variáveis independentes do estudo.

Desta forma, caso em uma pesquisa exista seis variáveis independentes, teríamos um total de 2^6 possíveis modelos de regressão. Caso esse número seja de 25 variáveis teríamos um total de 2^{25} possíveis modelos. Isso significa que com seis ou 25 preditores, há mais de 60 ou 1.000.000 possíveis modelos de regressão, respectivamente.

Portanto, encontrar a melhor combinação para um modelo de regressão para responder ao objetivo de sua pesquisa é desafiador.

O SAS System possui procedimentos de análise de regressão que incluem potentes ferramentas para limitar o número de modelos candidatos de forma que possibilite a escolha de um modelo apropriado de acordo a experiência do usuário/pesquisador e prioridades da pesquisa.

É possível construir um modelo de regressão manualmente por meio do ajuste do modelo com todas as variáveis independentes e posterior exclusão de variáveis não-significativas passo a passo considerando o valor-p.

O procedimento seria eliminar a primeira variável não-significativa e ajustar o modelo remanescente e proceder à eliminação de variáveis menos significativas do modelo. Esse procedimento é repetido até obter um modelo com todas variáveis significativas. Esse

procedimento manual de construção de um modelo pode ser realizado na situação em que se tem poucas variáveis independentes.

Considerando uma situação de pesquisa em que se tenha sete variáveis independentes para construir um modelo de regressão que explique a variação de y , tem-se o seguinte arranjo de todos possíveis modelos de regressão indicado no quadro 27.

Quadro 27. Número de modelos de regressão candidatos a partir do número de variáveis independentes disponíveis.

Número de modelos	Modelos de regressão candidatos
1	Modelo somente com o intercepto sem variável independente
7	Modelos somente com uma variável independente
21	Modelos com duas variáveis independentes
35	Modelos com três variáveis independentes
35	Modelos com quatro variáveis independentes
21	Modelos com cinco variáveis independentes
7	Modelos com seis variáveis independentes
1	Modelo com sete variáveis independentes
Total= 128	

Observa-se que a quantidade de modelos candidatos para um total de sete variáveis independentes equivale a: $2^7=128$ modelos. Ressalta-se que essa quantidade não considera variáveis criadas a partir das atuais como, por exemplo, efeitos de polinômio e interação.

O SAS System possui três algoritmos que seleciona variáveis independentes de forma automatizada e constrói saídas com equações que inserem as variáveis independentes mais importantes para explicar o y .

Os procedimentos PROC REG e PROC GLMSELECT suportam a aplicação dos métodos tradicionais de seleção de variáveis Forward, Backward e Stepwise descritos no quadro 28.

Quadro 28. Descrição dos métodos de seleção de variáveis independentes automatizada para a construção de modelos de regressão.

Método de seleção para modelagem	Funcionamento básico do método
Forward	Esse método inicia o processo de seleção de variáveis a partir de um modelo sem nenhuma variável independente. Logo, calcula a estatística F calculada para cada variável independente e examina o maior valor de F. Caso essa variável correspondente seja significativa, de acordo ao nível especificado, a mesma é adicionada no modelo. Essa mesma variável permanece no modelo mesmo se tornar não significativa após a inclusão de novas variáveis. O método adiciona novas variáveis ao modelo até o momento em que nenhuma das candidatas sejam significativas de acordo ao nível de significância pré-estabelecido no SAS. O nível de significância pré-estabelecido no SAS para a inclusão de variáveis é de p-valor=0,50.
Backward	Conhecido como método de eliminação, inicia com um modelo de regressão contendo todas as variáveis independentes disponíveis. Calcula a estatística F para cada modelo com uma variável e, caso seja significativa, permanece no modelo. O método repete o processo para a próxima variável até que uma ou mais variáveis restantes sejam não significativas e são eliminadas. A variável excluída não é reavaliada novamente. O nível de significância pré-estabelecido no SAS para a exclusão de variáveis é de p-valor=0,10.
Stepwise	Esse método combina aspectos do método Forward e Backward. Inicia com um modelo de regressão sem nenhuma variável independente e seleciona cada variável por vez semelhante ao método Forward. Entretanto, caso uma variável se torne não significativa, à medida que se inclui novas variáveis, o método exclui a mesma semelhante ao método Backward. O processo finaliza se nenhuma variável pode ser adicionada ou eliminada do modelo ou quando a variável a ser adicionada no modelo seja uma que já foi excluída em um passo anterior do processo. O valor de nível de significância pré-estabelecido no SAS é de p-valor=0,15 tanto para entrada como eliminação de variáveis. Recomendado quando se tem grande volume de variáveis independentes candidatas.

Os resultados da seleção de variáveis são diferentes para cada método. Portanto, a escolha do método a utilizar pode ser baseada na análise crítica do modelo final considerando o critério da importância biológica e a facilidade de obtenção das variáveis selecionadas.

A solicitação de uso de um dos métodos tradicionais de seleção de variáveis independentes automatizada pode ser realizada utilizando o procedimento PROC GLMSELECT ou PROC REG do SAS. Neste caso, basta indicar qual o método desejado após o sinal de igualdade na declaração SELECTION= conforme exemplo da sintaxe a seguir que utiliza o método stepwise.

```
proc reg data=teste;  
    model y= x x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 / selection=stepwise;  
run;  
  
proc glmselect data=teste;  
    model y= x x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 / selection=stepwise;  
run;
```

O grande diferencial do procedimento PROC GLMSELECT é a possibilidade de selecionar variáveis independentes utilizando os métodos mais recentes de seleção como o LASSO (TIBSHIRANI, 1996) e o LAR (EFRON et al., 2004).

Ademais, o procedimento possui outras características específicas que os procedimentos PROC REG e PROC GLM não possuem, tais como:

- Inclui efeitos de classificação no processo de modelagem;
- Suporta a inclusão de qualquer nível de interação (efeito cruzado) ou efeitos aninhados e
- Possibilita a partição interna do dataset em dados de treinamento, validação e teste.

Especificamente, o PROC GLMSELECT possui uma opção que possibilita a customização do processo de seleção de variáveis como uma forma alternativa para medir a melhoria no ajuste do modelo a ser construído além da estatística F dos métodos tradicionais de seleção.

Para especificar uma estatística alternativa para aferir o ganho no ajuste do modelo com a inclusão ou exclusão de variáveis utiliza-se a declaração SELECT= sendo possível utilizar o nível de significância desejado ou critérios de bondade de ajuste conforme descrito no Quadro 29.

Quadro 29. Critérios estatísticos para a aplicação nos métodos de seleção de variáveis para a construção de modelos de regressão linear.

Opção SELECT=	Descrição
SELECT=SL	<p>Especifica ao SAS que as variáveis independentes candidatas serão controladas considerando o nível de significância (significance level). Adicionalmente é possível solicitar que a entrada, permanência ou saída das variáveis com valores-p estabelecidos, incluindo-se o nível de significância desejado na opção para entrar (SLENTY=) e permanecer (SLSTAY=) no modelo de regressão. Entretanto, os valores p não podem ser interpretados de forma confiável como probabilidades. A sintaxe a seguir mostra um exemplo considerando o método de seleção stepwise com o nível de significância de p-valor=0,05 para a entrada e permanência de variáveis candidatas tanto para o PROC GLMSELECT como para o PROC REG:</p> <pre data-bbox="432 887 1442 1375"> proc glmselect data=teste; model y= x x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 / selection=stepwise select=sl slstay=0.05 slentry=0.05; run; proc reg data=teste; model y= x x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 / selection=stepwise select=sl slstay=0.05 slentry=0.05; run; </pre>
SELECT=AIC	<p>Considera o critério de informação de Akaike (AIC) para a seleção de variáveis. É possível especificar outros critérios de informação como o Akaike Corrigido (AICC) e outros como o BIC e SBC. A seleção das variáveis se dá de acordo à contribuição na redução da variância do modelo. Além dos critérios de informação é possível indicar o Coeficiente de Determinação Ajustado (ADJRSQ) o Cp de Mallows (CP) e outras estatísticas. A sintaxe a seguir mostra um exemplo utilizando o critério de Akaike:</p> <pre data-bbox="432 1720 1442 1964"> proc glmselect data=teste; model ipag= x x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 / selection=stepwise select=aic; run; </pre>

Uma questão a ser considerada para um modelo de regressão construído a partir da aplicação do método de seleção automática de variáveis independentes em regressão linear é descrita por Harrell (2001):

- i) O valor para o coeficiente de determinação é altamente enviesado;
- ii) A estatística de F e Qui-quadrado não tem a distribuição que deveria seguir;
- iii) O valor do erro padrão dos coeficientes de regressão é menor do que deveria ser, influenciando diretamente nos testes de hipóteses e amplitude do intervalo de confiança.

Portanto, especificar qualquer um dos métodos de seleção automática de variáveis, acarretará em uma seleção arbitrária de variáveis desconsiderando a relação entre as variáveis independentes e a dependente. Alguns autores como Flom e Cassell (2007) criticam a aplicação do método de seleção por Forward, Backward e Stepwise baseado no nível de significância para modelos de regressão linear ordinária.

3.6.1. Aplicação dos métodos de controle de seleção de variáveis

Para fins de aplicação dos métodos de seleção e controle de variáveis independentes, será considerado o caso florestal 7 para produzir os modelos candidatos com vistas a explicar a variação do incremento periódico anual em área transversal (IPAg).

Caso florestal 7: Construção de modelos pelo método de seleção “Stepwise”

Considere que uma investigação foi realizada com o objetivo de construir um modelo de regressão para estimar o incremento periódico anual em área transversal (IPAg) da regeneração natural de árvores de *Bertholletia excelsa* (Castanha do Brasil) considerando um grupo de 13 variáveis independentes obtidas em cada árvore individual.

3.6.1.1. Seleção de variáveis considerando nível de significância

Para resolver o caso florestal 7 será utilizado a seguinte sintaxe do procedimento PROC GLMSELECT:

```

data castanha;
  set book.bnregen;

%let interval = balmod compcopa diamcopa forcopa formalcopa glovholl gralesbel h hc
indabrag indsali propcopa possocial;

ods graphics on;
proc glmselect data=castanha plots=all;
  model ipag= &interval / selection=stepwise details=steps select=sl slstay=0.05
slentry=0.05 showpvalues;
run;

```

A opção PLOTS=ALL solicita a impressão de todos os gráficos disponíveis para fins de avaliação das etapas da modelagem.

A variável dependente ipag é especificada em MODEL juntamente com todas variáveis independentes da pesquisa incluídas na opção macro &INTERVAL do SAS.

A opção SELECTION=STEPWISE solicita ao PROC GLMSELECT que considere essa técnica de seleção de modelos candidatos. Essa técnica é a considerada padrão quando não especificada no SAS. A opção DETAILS=STEPS exibe uma tabela e gráficos das variáveis de entrada selecionadas em cada etapa do processo. Em seguida, SELECT= especifica o critério que determina a ordem em que as variáveis entram ou saem em cada etapa do método stepwise. O valor padrão é SELECT=SBC, mas, neste exemplo foi considerado SL para nível de significância como critério para selecionar variáveis para o modelo. Neste caso, o nível de significância foi alterado para 0,05. Para o SAS exibir os valores p, utilizou-se SHOWPVALUES.

Após o processamento, o SAS gera o resultado para os modelos candidatos selecionados conforme mostra Output 34:

Output 34. Resultado do PROC GLMSELECT com informações solicitadas para obter o modelo de regressão.

Essa primeira parte dos resultados fornece informações básicas sobre o processo de seleção do modelo. Os níveis de significância considerados também são indicados.

Data Set	WORK.CASTANHA
Dependent Variable	IPAg
Selection Method	Stepwise
Select Criterion	Significance Level
Stop Criterion	Significance Level
Entry Significance Level (SLE)	0.05
Stay Significance Level (SLS)	0.05
Effect Hierarchy Enforced	None

Number of Observations Read	251
Number of Observations Used	251

Dimensions	
Number of Effects	13
Number of Parameters	13

Effect Entered: Intercept

A próxima parte do resultado são as informações de cada passo realizado pelo método Stepwise. A etapa zero mostra a análise para um modelo somente com intercepto.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	0	0	.	.	.
Error	250	39839	159.35792		
Corrected Total	250	39839			

Root MSE	12.62370
Dependent Mean	6.78596
R-Square	0.0000
Adj R-Sq	0.0000
AIC	1526.85735
AICC	1526.90573
SBC	1277.38280

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	6.785960	0.796801	8.52	<.0001

Effect Entered: h

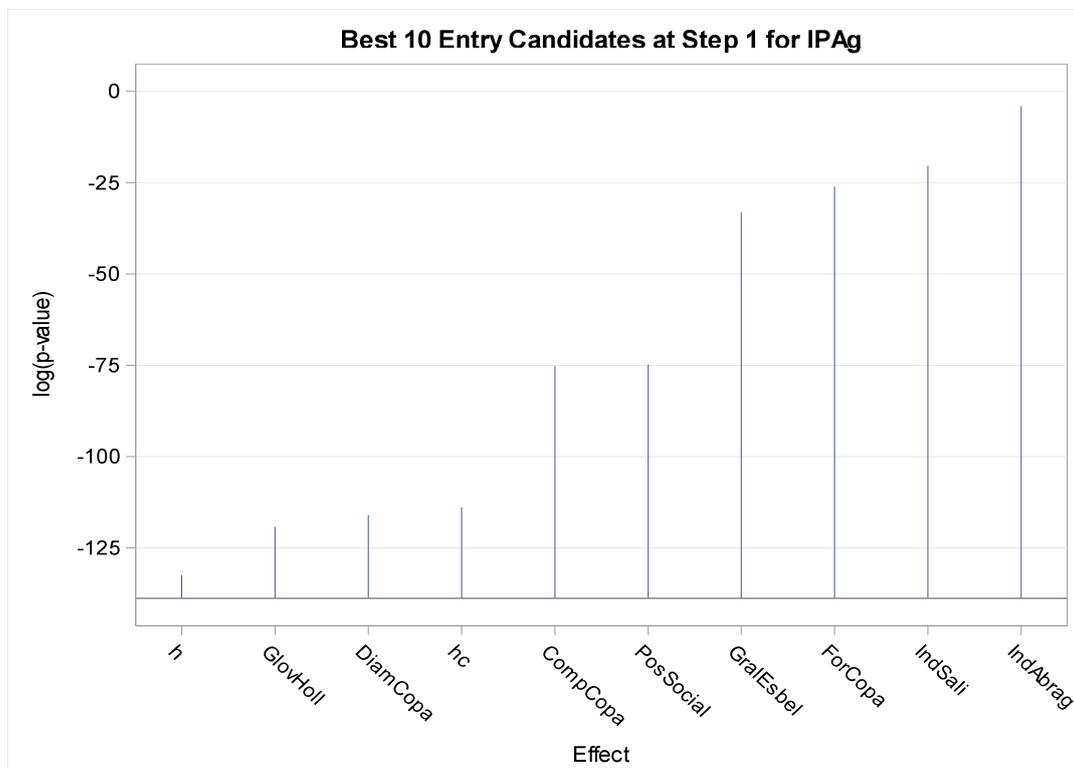
Na etapa 1 a variável *h* entrou no modelo primeiro, o que indica que de todos os modelos com uma variável independente, altura total (*h*) foi a variável com o nível de significância mais alto. O SAS mostra uma tabela e um gráfico com os 10 melhores modelos de uma variável independente com maiores níveis de significância. As variáveis são classificadas de acordo com o valor-*p* mais significativo dentro dos limites especificados na opção *SLEntry* e *SLStay*.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	25776	25776	456.37	<.0001
Error	249	14064	56.48005		
Corrected Total	250	39839			

Root MSE	7.51532
Dependent Mean	6.78596
R-Square	0.6470
Adj R-Sq	0.6456
AIC	1267.49773
AICC	1267.59490
SBC	1021.54864

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	-9.756543	0.908102	-10.74	<.0001
<i>h</i>	1	1.741659	0.081527	21.36	<.0001

Best 10 Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	h	-132.4085	<.0001
2	GlovHoll	-119.2428	<.0001
3	DiamCopa	-116.0432	<.0001
4	hc	-113.9597	<.0001
5	CompCopa	-75.2618	<.0001
6	PosSocial	-74.8116	<.0001
7	GralEsbel	-33.1357	<.0001
8	ForCopa	-26.0990	<.0001
9	IndSali	-20.4064	<.0001
10	IndAbrig	-4.0028	0.0183



Effect Entered: GlovHoll

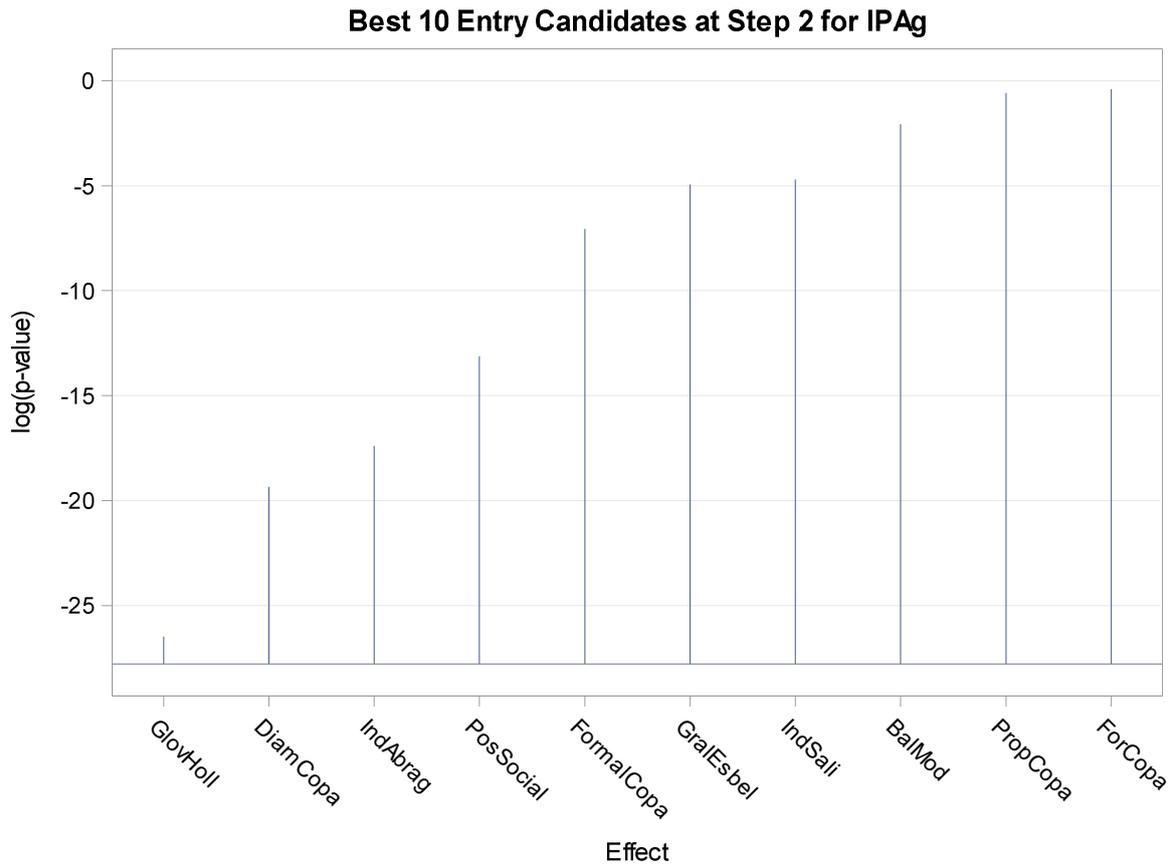
No passo 2 a variável *GlovHoll* entrou no modelo. Neste caso, o método de seleção (Stepwise) verifica se a variável *h* (anteriormente selecionada) se torna não significativa com a inclusão da segunda variável, em caso afirmativo o método a remove do modelo candidato. A variável *h* permaneceu significativa como mostra a tabela de valores dos coeficientes de regressão (Parameters Estimates). Observa-se na Tabela e Gráfico dos 10 melhores candidatos que a variável *h* não aparece pois já está incluída no modelo.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	28282	14141	303.45	<.0001
Error	248	11557	46.60130		
Corrected Total	250	39839			

Root MSE	6.82651
Dependent Mean	6.78596
R-Square	0.7099
Adj R-Sq	0.7076
AIC	1220.23068
AICC	1220.39328
SBC	977.80704

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	-9.009018	0.831145	-10.84	<.0001
GlovHoll	1	14.544275	1.983194	7.33	<.0001
h	1	1.085533	0.116140	9.35	<.0001

Best 10 Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	GlovHoll	-26.4779	<.0001
2	DiamCopa	-19.3392	<.0001
3	IndAbrig	-17.3987	<.0001
4	PosSocial	-13.1220	<.0001
5	FormalCopa	-7.0597	0.0009
6	GralEsbel	-4.9471	0.0071
7	IndSali	-4.6943	0.0091
8	BalMod	-2.0669	0.1266
9	PropCopa	-0.5751	0.5626
10	ForCopa	-0.3923	0.6755



Os passos 3 e 4 foram omitidos como forma de resumir os resultados!

Após o último passo, o SAS apresenta um resumo de todo o processo de seleção de modelos. A Tabela de Resumo (*Stepwise Selection Summary*) apresenta as variáveis que entraram e saíram do modelo. Neste caso, todas que entraram permaneceram no modelo final. Os valores dos coeficientes de regressão para o modelo final podem ser encontrados no passo anterior. Logo a seguir o SAS informa que o processo de seleção parou devido a variável para entrada ter um *SLEntry* maior do que 0,05 e a variável para remoção ter um *SLStay* menor do que 0,05.

Em seguida o SAS apresenta uma tabela (*Stop Details*) mostrando que a variável candidata à entrada no modelo *BalMod* não cumpriu o critério de entrada bem como a variável candidata à remoção do modelo *PosSocial* não atendeu ao critério de remoção do modelo.

Stepwise Selection Summary					
Step	Effect Entered	Effect Removed	Number Effects In	F Value	Pr > F
0	Intercept		1	0.00	1.0000
1	h		2	456.37	<.0001
2	GlovHoll		3	53.78	<.0001
3	IndAbrag		4	31.66	<.0001
4	PosSocial		5	5.96	0.0153

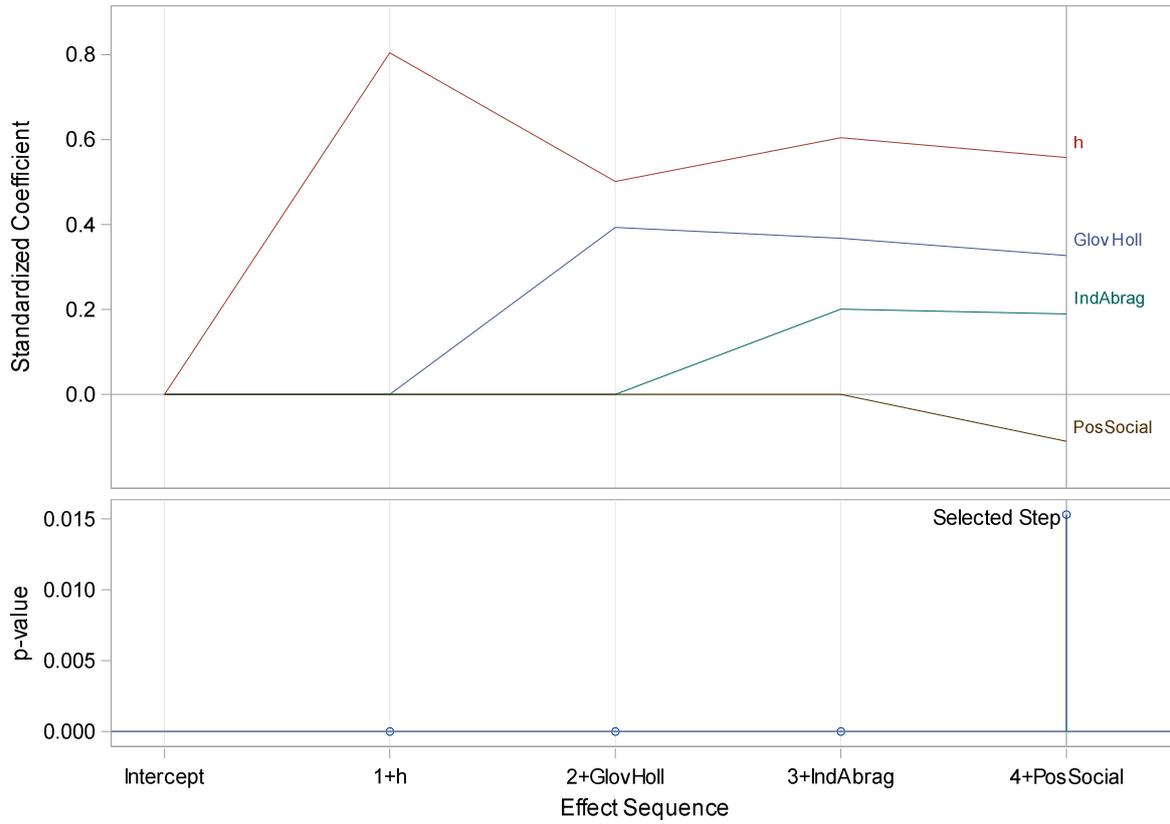
Selection stopped because the candidate for entry has SLE > 0.05 and the candidate for removal has SLS < 0.05.

Stop Details				
Candidate For	Effect	Candidate Significance	Compare Significance	
Entry	BalMod	0.1434	> 0.0500	(SLE)
Removal	PosSocial	0.0153	< 0.0500	(SLS)

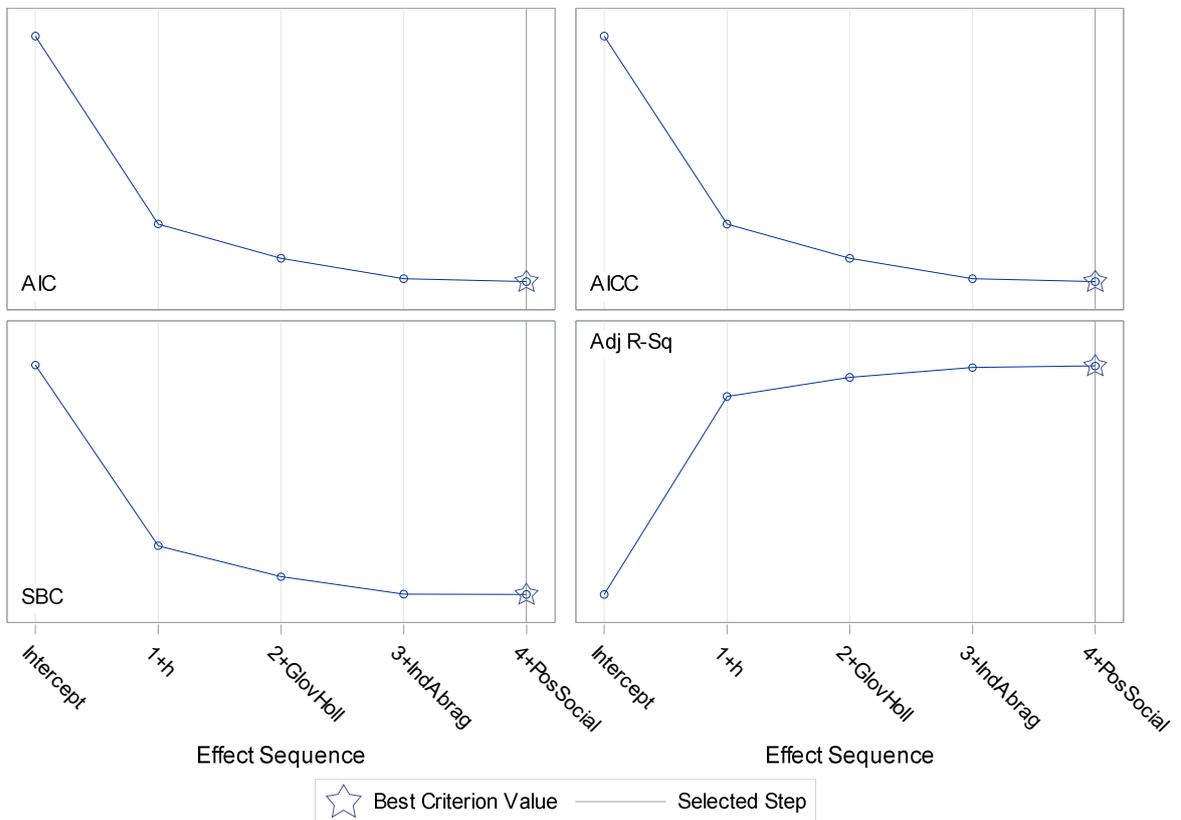
O SAS apresenta em seguida, dois painéis de gráficos mostrando o comportamento dos valores dos coeficientes padronizados. Quando a variável *h* entrou no modelo o valor padronizado foi de aproximadamente 0,8. Em seguida quando a variável *GlovHoll* entrou no modelo o valor do coeficiente padronizado da variável *h* diminuiu para aproximadamente 0,5. Portanto, esse gráfico acompanha a mudança dos coeficientes padronizados sendo possível verificar quando estabiliza.

O segundo painel mostra a mudança nos valores dos critérios de bondade de ajuste (AIC, SBC, AICC e R^2). O símbolo “Estrela” indica o melhor modelo dos cinco avaliados.

Coefficient Progression for IPAg

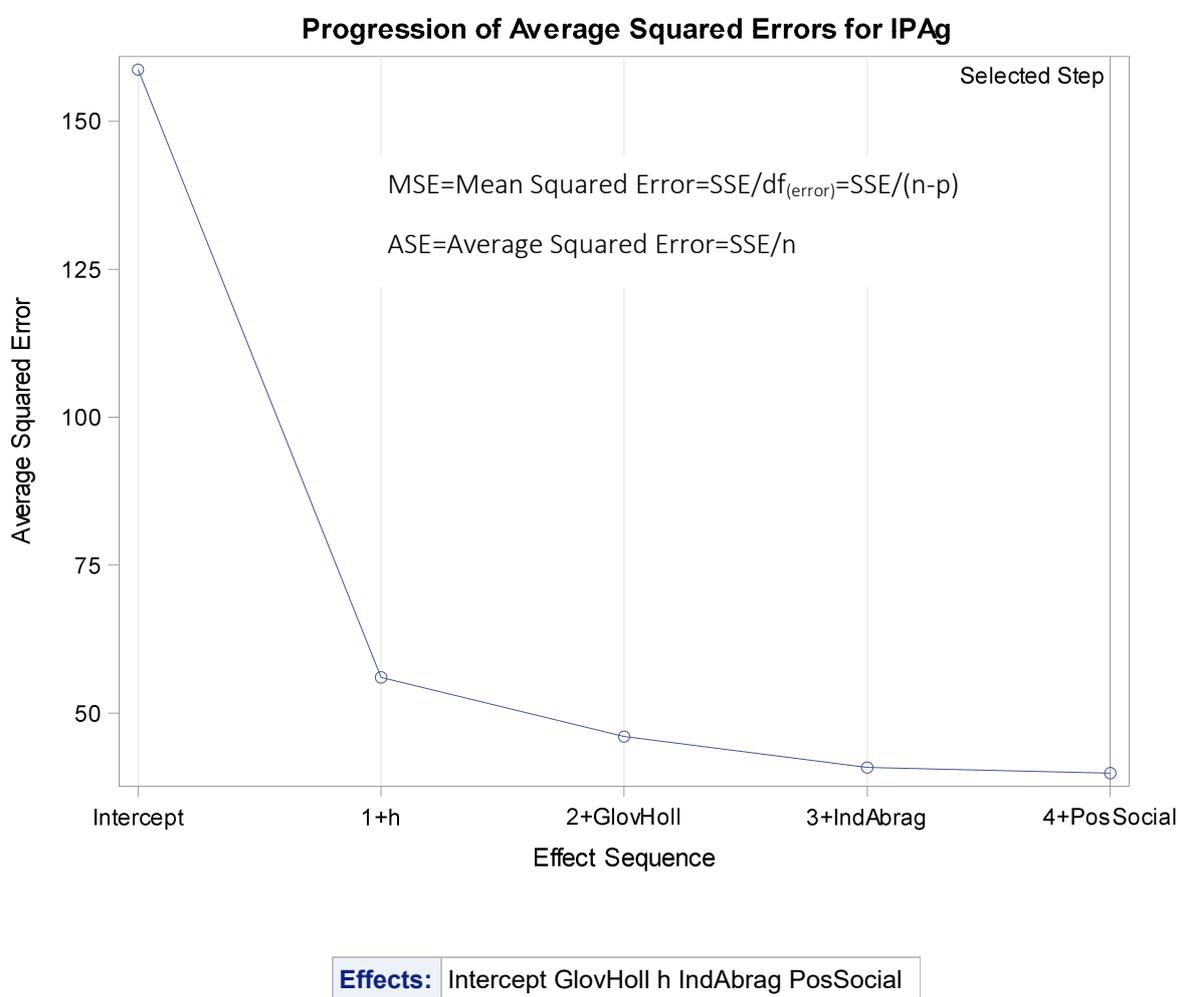


Fit Criteria for IPAg



Aqui o SAS mostra um gráfico de progressão do erro padrão médio (ASE). O processo de cálculo desse critério é similar ao erro padrão da estimativa considerando que um denominador n no cálculo em vez de $n-p$ (Dentro do gráfico foi inserido a fórmula de cálculo). Observa-se que a inclusão de novas variáveis proporcionou diminuição significativa do erro.

Esse gráfico mostra uma pequena diferença na variação não explicada dos modelos nos passos 3 e 4. Desta forma, mesmo o método Stepwise indicando o melhor modelo no passo 4, é possível decidir por usar o modelo candidato do passo 3. O painel de critérios de bondade de ajuste também colabora nesta decisão pelo fato da pequena contribuição da variável PosSocial no aumento do coeficiente de determinação.



Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	29838	7459.45593	183.47	<.0001
Error	246	10002	40.65714		
Corrected Total	250	39839			

Root MSE	6.37630
Dependent Mean	6.78596
R-Square	0.7490
Adj R-Sq	0.7449
AIC	1187.94834
AICC	1188.29260
SBC	952.57560

Parameter Estimates					
Parameter	DF	Estimate	Standar Error	t Value	Pr > t
Intercept	1	-10.122841	3.220044	-3.14	0.0019
GlovHoll	1	12.093362	1.958508	6.17	<.0001
h	1	1.207978	0.122506	9.86	<.0001
IndAbrag	1	22.749173	4.265951	5.33	<.0001
PosSocial	1	-1.549967	0.634647	-2.44	0.0153

Para solicitar ao SAS os demais métodos de seleção de modelos candidatos Forward e Backward, basta utilizar as seguintes sintaxes:

```
%let interval = balmod compcopa diamcopa forcopa formalcopa glovholl gralesbel h hc
indabrag indsali propcopa possocial;
```

```
ods graphics on;
```

```
proc glmselect data=castanha plots=all;
```

```
forward: model ipag= &interval / selection=forward details=steps select=sl slentry=0.05
```

```
showpvalues;
```

```
run;
```

```
ods graphics on;
```

```
proc glmselect data=castanha plots=all;
```

```
backward: model ipag= &interval / selection=backward details=steps select=sl slentry=0.05  
showpvalues;  
run;
```

O pesquisador deve ter cautela quando utilizar os métodos de seleção automática de modelos candidatos visto que, uma série de fatores não são considerados durante o processo de seleção de modelos o que gera problemas como:

- i) Caso o pesquisador considere diferentes níveis de significância para a entrada e permanência no modelo, o procedimento PROC GLMSELECT irá produzir modelos totalmente diferentes dos apresentados no caso florestal do exemplo;
- ii) A seleção de modelos automática não considera efeitos de colinearidade das variáveis selecionadas. Caso as variáveis independentes selecionadas apresentem alta correlação entre si, deve-se primeiramente reduzir a colinearidade e então rodar o procedimento de seleção novamente;
- iii) Modelos produzidos a partir dessa técnica resultam em vieses nas estimativas de parâmetros, previsões, erro padrão, cálculo incorreto dos graus de liberdade e os valores p tendem a errar no lado da significância superestimada. Alguns desses problemas são gerados quando considerado valores p para seleção de variáveis;
- iv) Valores-p são destinados a testar uma hipótese, não dezenas de hipóteses em muitos modelos com conjunto sobreposto de modelos. Portanto, considere utilizar a mesma base de dados (amostra) tanto para escolher o modelo quanto para avaliá-lo também é outra causa dos vieses nas estimativas dos parâmetros e valores-p.

Para aliviar alguns dos problemas de estimativas enviesadas, deve-se utilizar outros critérios para a entrada e saída de variáveis como os critérios de ajuste de bondade. Também se recomenda que a base de dados seja particionada, uma destinada para a modelagem e outra para validação do modelo.

Neste sentido, recomenda-se utilizar a metade dos dados disponíveis para a seleção do modelo. Em seguida, o modelo poderia ser aplicado em uma amostra de dados diferentes daquela que foi utilizada para desenvolver o modelo (outra metade).

Entretanto, muitas vezes o pesquisador não tem dados suficientes para o particionamento em conjunto de dados de treinamento e retenção para validação. Neste caso, com conjunto de dados pequenos ou moderados, a divisão de dados é ineficiente. Em tal situação, técnicas de reamostragem de Bootstrap podem auxiliar.

3.6.1.2. Seleção de variáveis considerando critérios de informação

Outra forma de controlar a seleção de variáveis de forma automática para modelos candidatos é o uso de critérios de informação em vez de valores-p como mostrado no tópico anterior. Alguns dos critérios de informação são considerados no procedimento PROC GLMSELECT do SAS bastando apenas informar qual critério na opção SELECT=, conforme sintaxe básica a seguir:

```
%let interval = balmod compcopa diamcopa forcopa formalcopa glovholl gralesbel h hc
indabrag indsali propcopa possocial;

ods graphics on;
Title "Seleção pelo critério de informação de Akaike corrigido";
proc glmselect data=castanha plots=all;
  stepwiseaic: model ipag= &interval / selection=stepwise details=steps select=aicc;
run;
```

Os resultados do processamento para métodos de seleção considerando apenas o critério de informação AICC são apresentados no Output 35:

Output 35. Resultado do GLMSELECT com informações solicitadas para obter o modelo de regressão solicitado pela técnica Stepwise considerando o critério de informação de Akaike Corrigido (AICC).

Data Set	WORK.CASTANHA
Dependent Variable	IPAg
Selection Method	Stepwise
Select Criterion	AICC
Stop Criterion	AICC
Effect Hierarchy Enforced	None

Number of Observations Read	251
Number of Observations Used	251

Dimensions	
Number of Effects	13
Number of Parameters	13

Effect Entered: Intercept

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	0	0	.	.	.
Error	250	39839	159.35792		
Corrected Total	250	39839			

Root MSE	12.62370
Dependent Mean	6.78596
R-Square	0.0000
Adj R-Sq	0.0000
AIC	1526.85735
AICC	1526.90573
SBC	1277.38280

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	6.785960	0.796801	8.52	<.0001

Effect Entered: h

No modelo somente com o intercepto no passo anterior, o valor de AICC é de 1526 e com a inclusão da variável h no modelo o critério reduziu para 1267. A variável h foi selecionada para entrar no modelo pois apresentou o menor valor de AICC entre as 10 melhores avaliadas conforme a tabela a seguir.

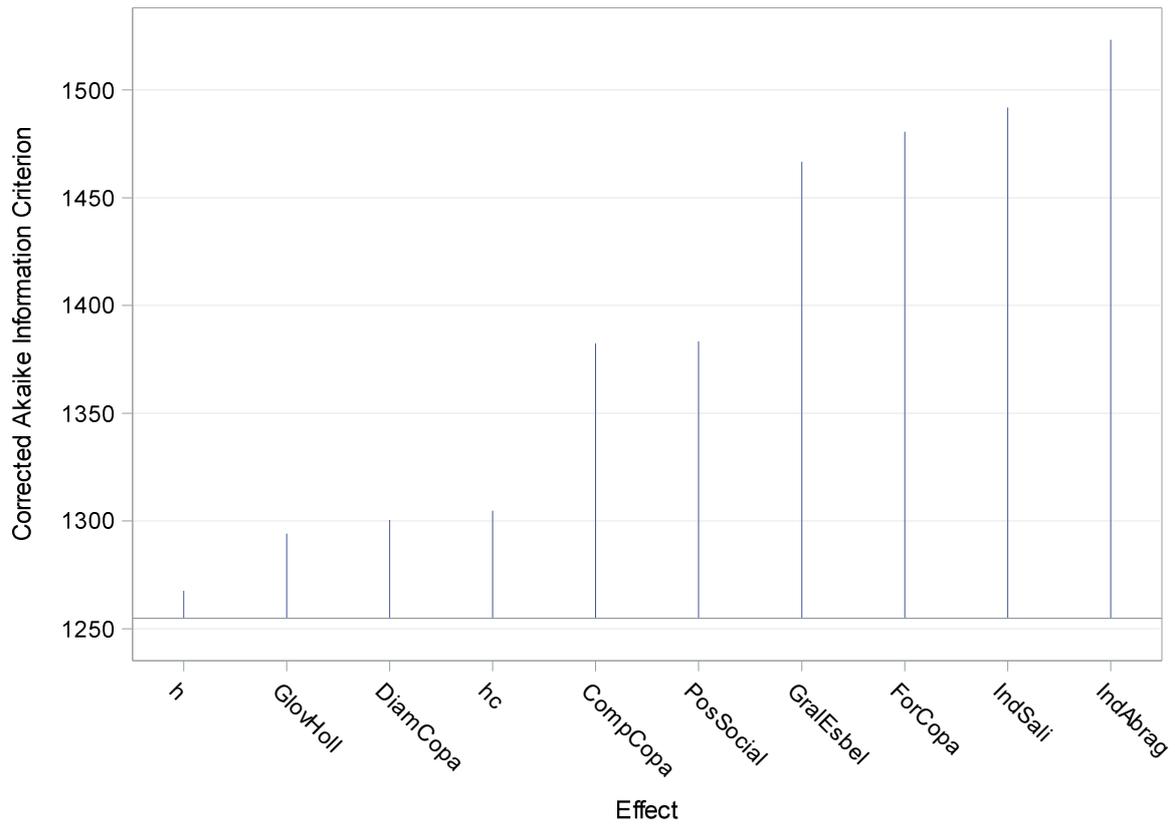
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	25776	25776	456.37	<.0001
Error	249	14064	56.48005		
Corrected Total	250	39839			

Root MSE	7.51532
Dependent Mean	6.78596
R-Square	0.6470
Adj R-Sq	0.6456
AIC	1267.49773
AICC	1267.59490
SBC	1021.54864

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	-9.756543	0.908102	-10.74	<.0001
h	1	1.741659	0.081527	21.36	<.0001

Best 10 Entry Candidates		
Rank	Effect	AICC
1	h	1267.5949
2	GlovHoll	1294.0755
3	DiamCopa	1300.5092
4	hc	1304.6985
5	CompCopa	1382.4284
6	PosSocial	1383.3315
7	GralEsbel	1466.6699
8	ForCopa	1480.6361
9	IndSali	1491.8793
10	IndAbrag	1523.3279

Best 10 Entry Candidates at Step 1 for IPAg

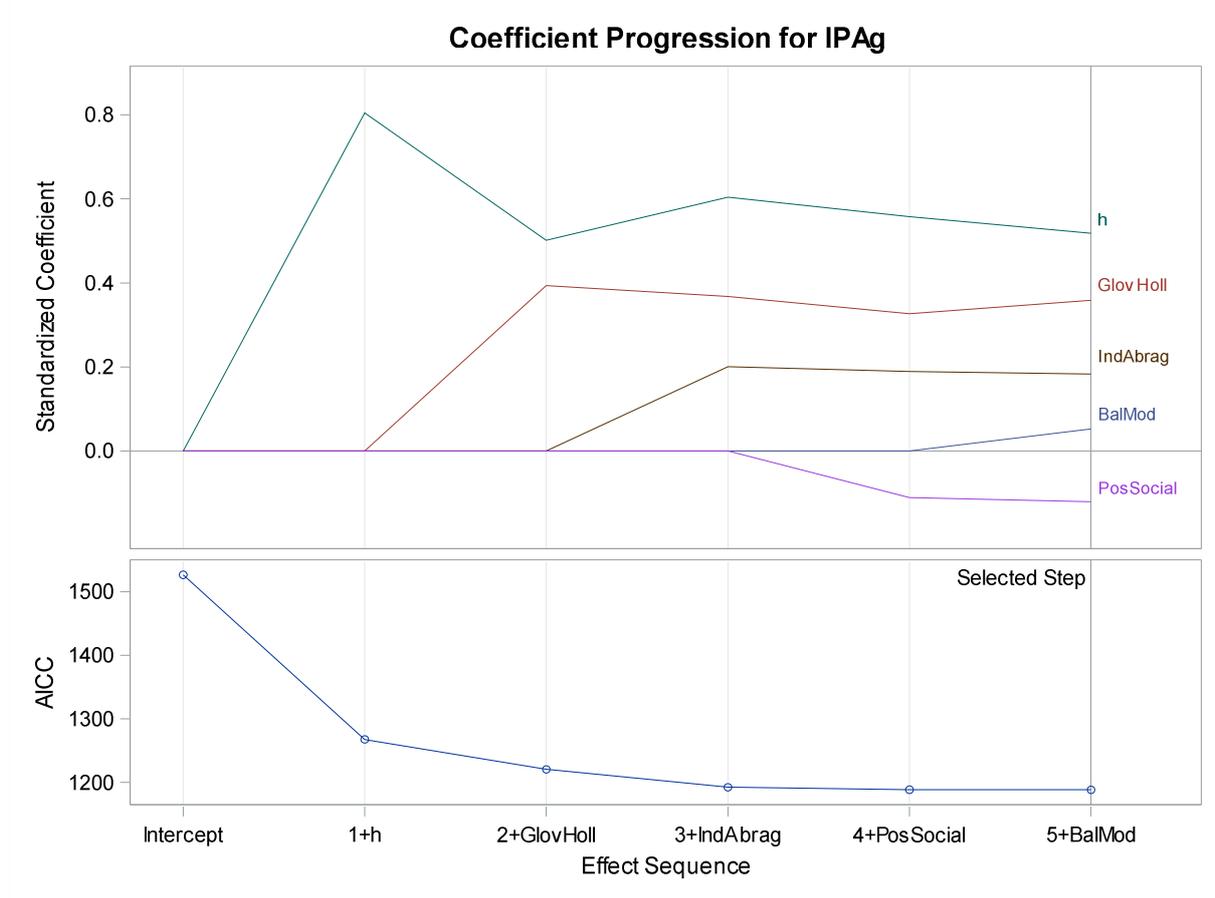


Os demais passos foram omitidos para resumir os resultados!

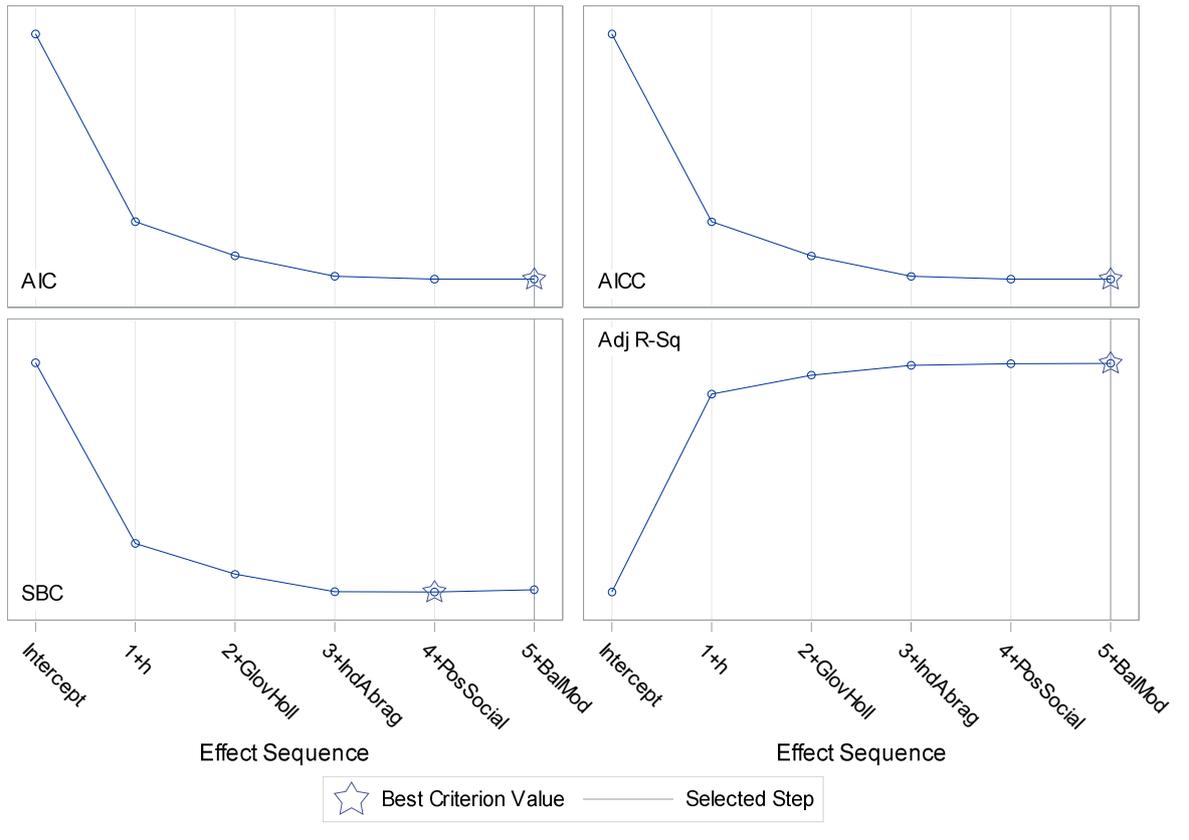
Stepwise Selection Summary				
Step	Effect Entered	Effect Removed	Number Effects In	AICC
0	Intercept		1	1526.9057
1	h		2	1267.5949
2	GlovHoll		3	1220.3933
3	IndAbrag		4	1192.2065
4	PosSocial		5	1188.2926
5	BalMod		6	1188.2111*
* Optimal Value of Criterion				

Selection stopped at a local minimum of the AICC criterion.

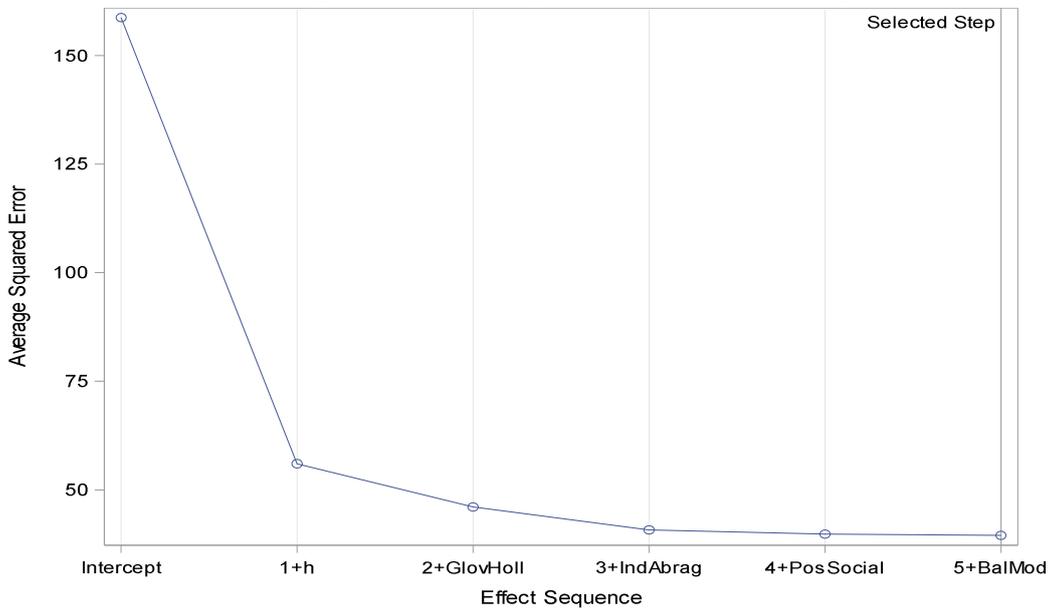
Stop Details			
Candidate For	Effect	Candidate AICC	Compare AICC
Entry	ForCopa	1188.2379	> 1188.2111
Removal	BalMod	1188.2926	> 1188.2111



Fit Criteria for IPAg



Progression of Average Squared Errors for IPAg



Effects: Intercept BalMod GlovHoll h IndAbrag PosSocial

Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	29925	5985.00651	147.90	<.0001
Error	245	9914.44844	40.46714		
Corrected Total	250	39839			

Root MSE	6.36138
Dependent Mean	6.78596
R-Square	0.7511
Adj R-Sq	0.7461
AIC	1187.75016
AICC	1188.21107
SBC	955.90288

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	-9.856425	3.217633	-3.06	0.0024
BalMod	1	2.462459	1.677413	1.47	0.1434
GlovHoll	1	13.257312	2.108668	6.29	<.0001
h	1	1.121935	0.135547	8.28	<.0001
IndAbrag	1	22.022365	4.284671	5.14	<.0001
PosSocial	1	-1.686837	0.639990	-2.64	0.0089

De acordo aos resultados do Output 35 o gráfico de painel mostra que o modelo adequado para explicar a variação de IPA_g possui cinco variáveis independentes conforme a última tabela de resumo dos parâmetros de regressão.

Entretanto, o painel indica que houve baixa contribuição na redução dos valores dos critérios de informação a partir do passo 3 quando a variável IndAbrag é adicionada. Esse comportamento também é indicado no gráfico de redução do erro.

Desta forma, dependendo dos custos e tempo de medição das variáveis indicadas no passo 4 e passo 5, essas variáveis poderiam ser excluídas do modelo.

Vale destacar que o grande propósito dos critérios de informação é evitar o overfitting dos dados, ou seja, o modelo possui um desempenho excelente para os dados de treino, porém, quando utilizado para o restante dos dados (Validação) o ajuste é ruim, resultando em um modelo ineficaz quando se realiza modelagem com particionamento de dados.

Portanto, o que se deseja em modelagem é apenas variável de entrada úteis no modelo que o ajudem a prever bem, e não variáveis irrelevantes que podem levar ao overfitting.

Neste caso, ao adicionar uma variável independente (x_1) no modelo somente com intercepto, deve-se verificar sua contribuição na diminuição do critério SBC como critério para mantê-la ou não no modelo.

Caso o valor de SBC diminuir 10 para 8, justifica manter x_1 no modelo, caso contrário, se a variação do valor de SBC for mínima, esta variável deve ser eliminada do modelo.

3.6.2. Validação do modelo de regressão

Na análise de regressão exploratória, o objetivo principal é o entendimento da relação entre a variável dependente y e as variáveis independentes x 's de forma a utilizar os resultados da análise da amostra observada e obter conclusões sobre toda a população alvo da pesquisa.

Por outro lado, na análise de regressão preditiva, o principal objetivo é generalizar os resultados para estimar/predizer os valores da variável dependente a partir de novas observações de variáveis independentes. Portanto, para que a aplicação do modelo seja aceitável, a validação é necessária.

Portanto, para generalizar um modelo de regressão torna-se necessário o uso de grande quantidade de dados observados para que seja possível dividi-lo em subconjuntos de dados para Treinamento, Validação e, algumas vezes, Teste.

Para a criação de subconjunto de dados, comumente considera-se 70% para dados de treinamento e 30% para a validação. Vale ressaltar que, caso se tenha um conjunto de dados pequeno, essa partição pode não ser eficiente e, quando possível, existe a possibilidade de realizar a validação cruzada para auxiliar em situação de conjunto de dados pequeno ou médio.

Desta forma, o subconjunto de dados de Treinamento será utilizado para o ajuste dos modelos. Logo, o subconjunto Validação será utilizado para selecionar o melhor modelo construído a partir dos dados de treinamento.

Hastie et al. (2016) informaram que é difícil ter uma regra definida para o número de observações necessárias para cada subconjunto de dados e recomendaram que a divisão seja de 50% para treinamento e 25% para validação e 25% para teste.

3.6.2.1. Avaliação do desempenho de um modelo

O processo de selecionar um modelo de regressão para uso inclui etapas para a avaliação do desempenho de cada um com vistas a escolher o melhor. Avaliar a performance de um modelo implica uma análise detalhada de erros de estimativas bem como a capacidade de generalização e simplicidade do mesmo.

Machado e Figueiredo Filho (2003) relataram que erro é um desvio do valor real e está associado à ideia de inexato e, portanto, não é um valor errado. A partir da análise de estimativas com relação ao valor verdadeiro de uma variável, podemos ter as seguintes magnitudes relacionadas a erro:

- Acurácia= é a proximidade do valor de uma medição com o valor verdadeiro que se deseja obter (HUSCH et al., 2003);
- Precisão= expressa o grau de concordância entre valores obtidos a partir de várias medições (HUSCH et al., 2003);
- Viés= refere-se a erros sistemáticos causados por estimativas tendenciosas de um modelo de regressão, por exemplo. Esse erro possui uma característica de ser acumulativo e proporciona informação sobre a Acurácia/Exatidão de um modelo de regressão. O termo viés-variância é muito discutido no âmbito do aprendizado de máquina para fins de predição de valores da variável dependente (target);
- Variância= proporciona informação para aferir a Precisão das estimativas de um modelo de regressão, ou seja, informa sobre o grau de concordância entre uma série de estimativas, obtidas a cada novo ajuste do modelo, e o valor verdadeiro (parâmetro) que se deseja estimar.

A Figura 43 representa, na prática, o comportamento de estimativas (pontos azuis) em relação ao parâmetro populacional (localizado no centro do alvo) com o qual se deseja estimar seu valor pelos modelos de regressão sob avaliação.

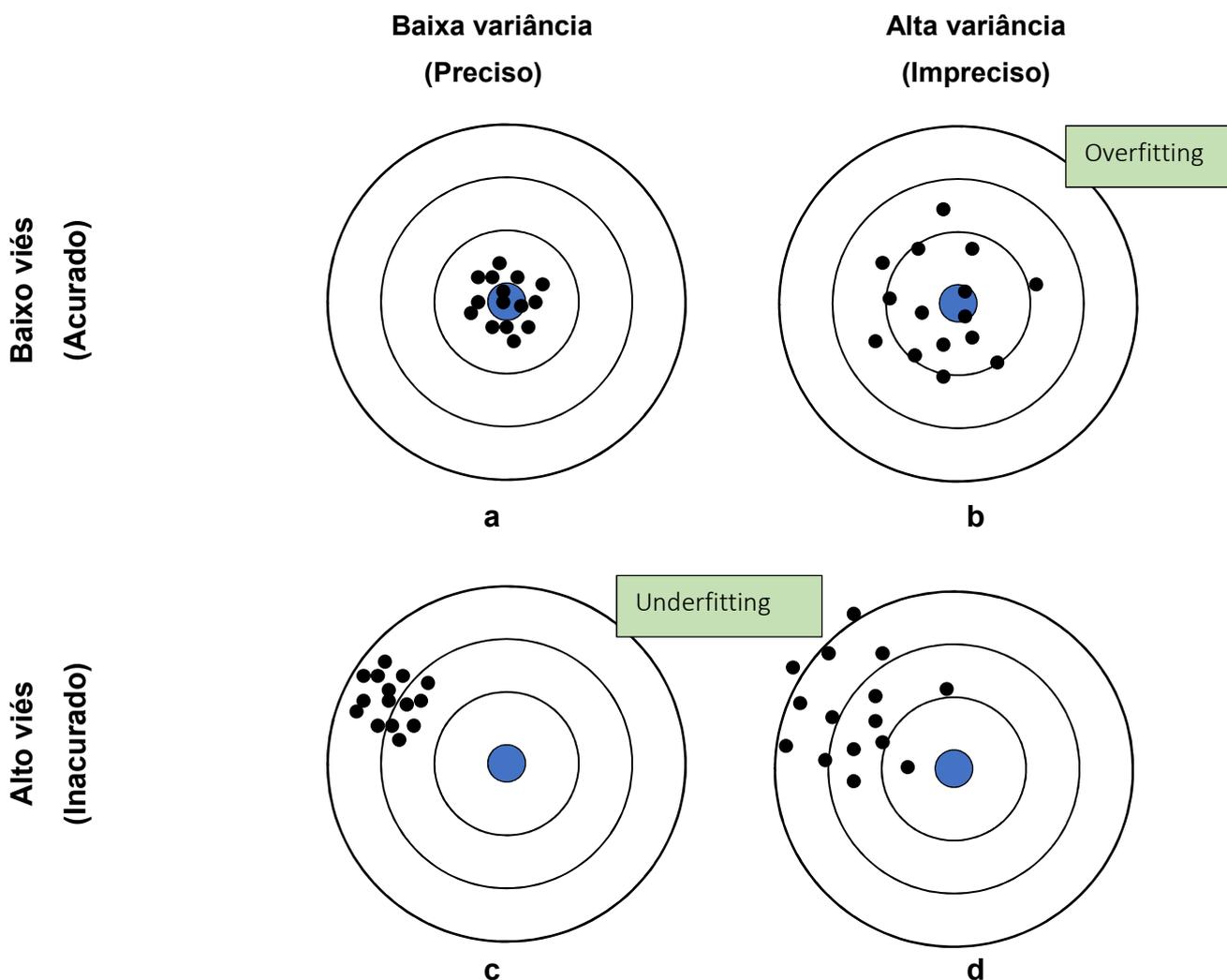


Figura 43. Representação da precisão e acurácia utilizando a dispersão de valores estimados por um modelo de regressão a partir de novas amostras de variáveis independentes (15 amostras=15 pontos) em relação ao valor do parâmetro que se deseja estimar (ponto central azul).

Fonte: Adaptado de HUSCH et al. (2003).

A partir da Figura 43 é possível obter quatro situações de acordo ao significado de Viés e Variância:

Situação 1: Modelo com baixo Viés (Acurado) e baixa Variância (Preciso). Esta situação é representada pelo alvo “a” que se apresenta como uma situação ideal em que os valores das estimativas apresentam pequenos desvios entre si e são próximos ao valor do parâmetro que se deseja estimar. No entanto, é muito difícil na prática ter um modelo com essas características;

Situação 2: Modelo com baixo Viés (Acurado) e alta Variância (Impreciso). Representada pelo alvo “b”. Neste caso, o modelo produz estimativas próximas ao valor do

parâmetro que se deseja estimar (representado pelo alvo azul no centro), mas os valores estimados apresentam elevada variação entre si, ou seja, o modelo está super-ajustando (**Overfitting**) e não generaliza para novos dados;

Situação 3: Modelo com alto Viés (Inacurado) e baixa Variância (Preciso). Representada pelo alvo “c”. Significa que o modelo produz estimativas com pequenos desvios (baixa Variância), mas seus valores são muito diferentes do valor do parâmetro devido a que se apresentam com alto Viés (Tendência). Neste caso, o modelo está subajustando (**Underfitting**);

Situação 4: Modelo com alto Viés (Inacurado) e alta Variância (Impreciso). Representada pelo alvo “d”. Nesta situação, o modelo é considerado inconsistente visto que é tendencioso e com grandes desvios entre as estimativas. Geralmente esse tipo de modelo é obtido na fase inicial de construção.

Diante das situações apresentadas na Figura 43, fica claro que o grande objetivo é obter um modelo de regressão considerando o ponto de partida da modelagem a situação “d” em direção da situação “a”.

Entretanto, garantir um modelo Acurado e Preciso simultaneamente é muito difícil pois à medida que se aumenta a complexidade de um modelo (adicionar mais e mais variáveis independentes) em busca de melhores estimativas, o viés diminui, mas como consequência, a variância aumenta. Por outro lado, reduzir a complexidade de um modelo proporciona redução da variância, mas provoca aumento do viés. Esse comportamento é conhecido como Dilema Viés-Variância (em inglês Bias-Variance Tradeoff).

Isso se deve ao fato de que o erro para qualquer algoritmo depende de três componentes:

- Erro irreduzível
- Erro de Viés
- Erro de Variância

A seguinte expressão demonstra o desdobramento do erro da estimativa ($Erro(x)$) em três componentes de erro (HASTIE et al., 2016):

$$Erro(x) = \sigma_{\varepsilon}^2 + \underbrace{[E\hat{f}(x) - f(x)]^2}_{\text{Viés}^2} + \underbrace{E[\hat{f}(x) - E\hat{f}(x)]^2}_{\text{Variância}}$$

Em que:

σ_{ε}^2 = representa a variância da variável dependente com relação à média verdadeira (Erro irreduzível);

$E\hat{f}(x)$ = representa o valor esperado (média) das estimativas do modelo de regressão ajustado;

$f(x)$ = representa a média verdadeira que se deseja estimar;

$\hat{f}(x)$ = representa o valor estimado pelo modelo de regressão ajustado.

O segundo termo da expressão, lado direito da igualdade, representa o viés (Bias) quadrático, ou seja, o valor pelo qual a média das estimativas do modelo ($E\hat{f}(x)$) difere da média verdadeira $f(x)$. O último termo representa a Variância, ou seja, o desvio quadrático entre os valores estimados pelo modelo de regressão ajustado ($\hat{f}(x)$) e o valor médio das estimativas.

Em termos de regressão linear temos que $\hat{f}(x)$ é:

$$\hat{f}(x) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p$$

Um modelo se torna mais complexo à medida que se aumenta o número de variáveis independentes. Portanto, um modelo de polinômio do segundo grau será menos complexo do que um modelo de polinômio do quinto grau. Portanto, um modelo “ideal” deve ter um balanceamento entre complexidade e capacidade de generalização.

A Figura 44 mostra o comportamento do dilema Viés-Variância com o aumento da complexidade do modelo.

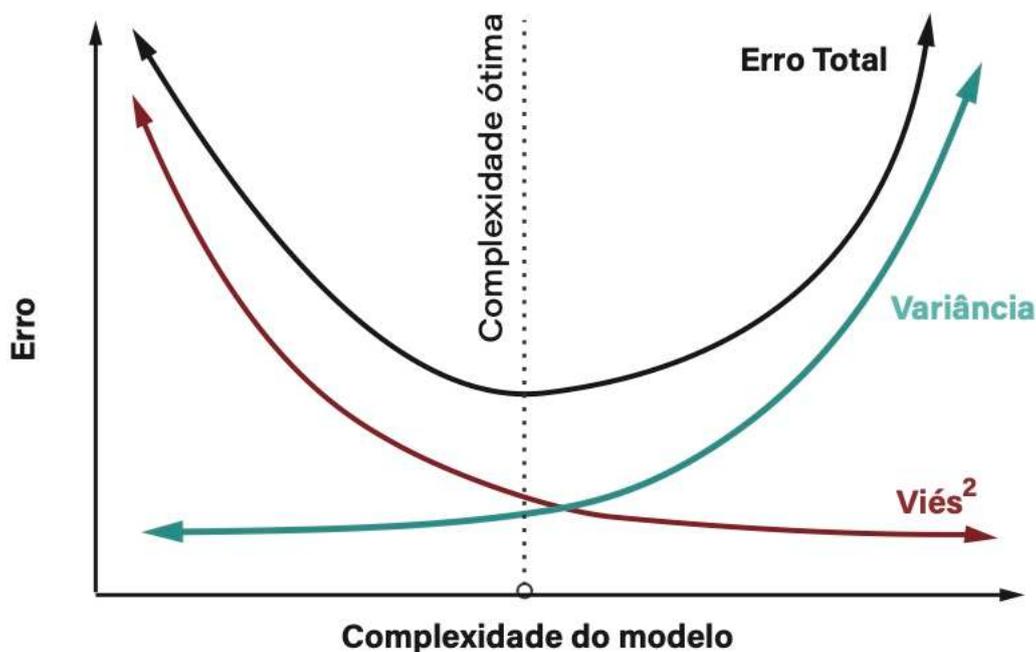


Figura 44. Comportamento dos componentes de erro à medida que o modelo se torna mais complexo.

Fonte: Adaptado de RODRIGUES (2023).

Para acessar o desempenho de modelos de regressão o procedimento PROC GLMSELECT imprime um gráfico da trajetória do erro padrão médio (ASE=average square error) para cada modelo construído. A expressão para calcular o ASE é a seguinte:

$$ASE = \frac{\sum_{i=1}^n \{ [Y_i - (\hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_j X_j)]^2 \}}{n}$$

3.6.2.2. Particionamento de dados

O procedimento PROC GLMSELECT possibilita a construção de modelos de regressão em duas situações: i) quando os dados observados para o ajuste estão particionados em conjuntos de dados para treinamento e dados para validação ou ii) o conjunto de dados observados único sem o particionamento.

Em ambas situações o PROC GLMSELECT pode construir modelos de regressão iniciando pelo treinamento e depois com a validação.

O procedimento particiona os dados automaticamente em dois (treinamento e validação) ou três (treinamento, validação e teste) subconjuntos de dados (amostras) a partir do conjunto geral. A seguinte sintaxe particiona os dados.

```

proc glmselect data= dataset_de_treinamento seed= número;
  partition fraction(test=fração validate=fração);
  model y= x x1 x2 x3 x4 x5 x6 x7 x8 x9 x10;
run;

```

A opção SEED= especifica um número inteiro maior do que 1 para iniciar o processo de seleção aleatória. Caso não seja especificado, o SAS considera a hora registrada no computador como número semente. Entretanto, recomenda-se utilizar um valor quando se deseja replicar os mesmos resultados de seleção.

A opção FRACTION da declaração PARTITION especifica a fração (proporção a partir do conjunto de dados geral) de observações com a qual se deseja particionar os dados em amostras para a validação e teste do modelo. Essas observações são selecionadas de forma aleatória. Neste caso, a soma da partição deve ser menor do que 1 sendo que, o restante dos dados será utilizado no treinamento.

Para demonstrar o particionamento de dados na construção de modelos de regressão preditiva, consideraremos o mesmo conjunto de dados de castanha do Brasil utilizado na modelagem. A sintaxe do PROC GLMSELECT para particionar os dados de castanha é o seguinte.

```

options validvarname=v7;
libname BNRegen xlsx "x:/Projects/Manual SAS/Apoio/Data/BNRegeneration.xlsx";

data castanha;
  set bnregen.base;

%let interval = balmod compcopa diamcopa forcopa formalcopa glovholl gralesbel h hc
indabrag indsali propcopa possocial;

ods graphics on;
proc glmselect data=castanha seed=1234;
  partition fraction(validate=0.3 test=0.15)
  model ipag= &interval;
run;

```

Neste caso, considerou-se que 30% do conjunto de observações “Castanha” serão utilizados para a validação do modelo e 15% para teste. Portanto, o restante de 55% será utilizado para treino do modelo. Os resultados da partição dos dados são apresentados no Output 36:

Output 36. Resultado da partição do conjunto de dados “castanha”. O restante da análise foi suprimido para fins didáticos.

Data Set	WORK.CASTANHA
Dependent Variable	IPAg
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	SBC
Effect Hierarchy Enforced	None
Random Number Seed	1234

Number of Observations Read	251
Number of Observations Used	251
Number of Observations Used for Training	134
Number of Observations Used for Validation	72
Number of Observations Used for Testing	45

Dimensions	
Number of Effects	14
Number of Parameters	14

A partir do total de 251 observações o SAS reservou 72 e 45 observações de forma aleatória para a validação e treino, respectivamente.

3.6.2.3. Validação de um modelo de regressão linear

Para fins de demonstração, vamos considerar o caso florestal 7 utilizando o procedimento PROC GLMSELECT do SAS considerando a sintaxe para a construção de modelos de regressão linear com partição de dados para treinamento e validação de acordo a seguinte sintaxe.

```

options validvarname=v7;
libname BNRegen xlsx "x:/Projects/Manual SAS/Apoio/Data/BNRegeneration.xlsx";

data castanha;
  set bnregen.base;

%let interval = balmod compcopa diamcopa forcopa formalcopa glovholl gralesbel h hc
indabrag indsali propcopa;
%let categorical =PosSocial;

ods graphics on;

proc glmselect data=castanha seed=1234 plots=(ASEPlot Coefficient);
  partition fraction(validate=0.3);
  class &categorical / param=ref;

  model IPAg=&interval &categorical / selection=forward (select=sbc choose=validate)
                                details=steps(candidates ParameterEstimates);

run;

```

A declaração DATA indica o conjunto de dados (castanha) de onde será extraído a amostra para validação de 30% do total de observações, sendo que o restante será destinado para treinar o modelo. Essa partição de dados foi realizada pela opção PARTITION.

A opção PLOTS solicita ao SAS que imprima o gráfico de ASE (Average Squared Error) que é simplesmente o valor médio da diferença ao quadrado do valor observado menos o valor estimado, sendo n o número de observações. É possível solicitar mais gráficos com a opção PLOTS=ALL.

Como existe uma variável categórica utilizada como input, utilizou-se a declaração CLASS para indicar a posição social (possocial) como variável candidata para o modelo. Neste caso, considerou-se a opção PARAM=FIRST para especificar o método de parametrização para a variável categórica, ou seja, indicar ao SAS o valor de referência de cada nível da variável categórica PosSocial.

Na declaração MODEL considerou-se o incremento periódico anual em área basal como variável de resposta (target). O método de seleção utilizado foi Forward com seleção pelo critério SBC para determinar quais variáveis (inputs) entram no modelo.

A opção CHOOSE=VALIDATE especifica que o melhor modelo será selecionado baseado no menor valor de ASE (Average Squared Error) para os dados de validação.

O resultado da análise é apresentado no Output 37 que resume em algumas tabelas e gráfico de ASE gerado pelo PROC GLMSELECT para o modelo de regressão linear preditiva com particionamento de dados.

Output 37. Tabela do modelo final selecionado por Forward e gráfico de ASE.

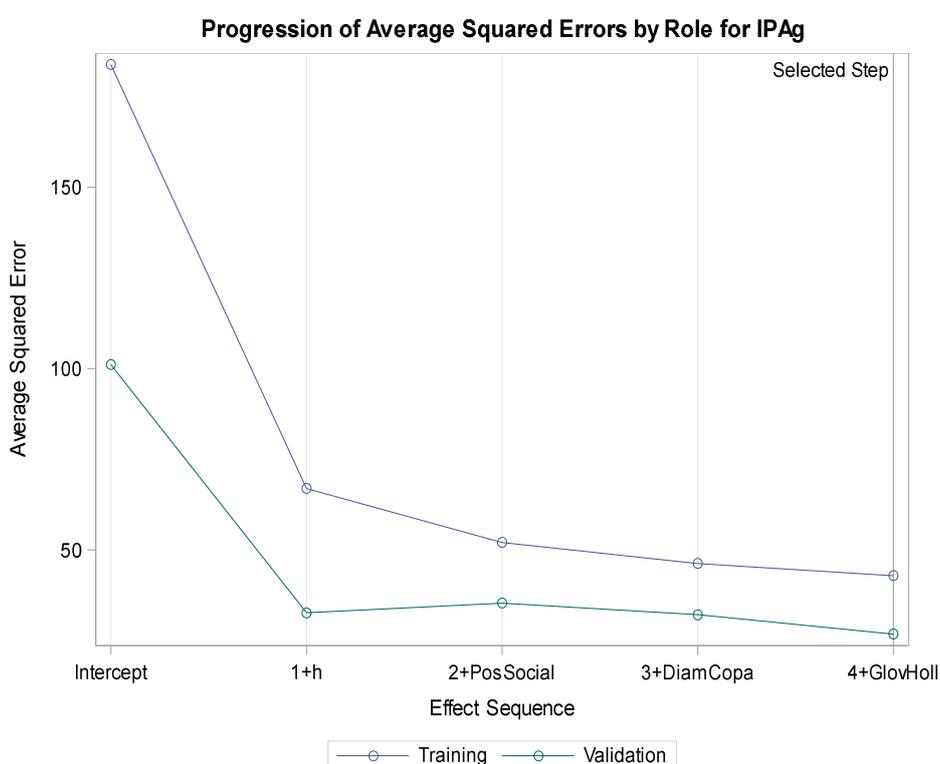
Forward Selection Summary						
Step	Effect Entered	Number Effects In	Number Parns In	SBC	ASE	Validation ASE
0	Intercept	1	1	917.6017	183.8141	101.1099
1	h	2	2	745.9830	66.9358	32.6620
2	PosSocial	3	5	717.4993	52.0618	35.3583
3	DiamCopa	4	6	702.0179	46.2681	32.1350
4	GlovHoll	5	7	694.0641*	42.9265	26.8084*
* Optimal Value of Criterion						

Effects: Intercept DiamCopa GlovHoll h PosSocial

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	6	24655	4109.22324	91.90
Error	168	7512.13419	44.71508	
Corrected Total	174	32167		

Root MSE	6.68693
Dependent Mean	7.00776
R-Square	0.7665
Adj R-Sq	0.7581
AIC	848.91056
AICC	849.77803
SBC	694.06406
ASE (Train)	42.92648
ASE (Validate)	26.80838

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	-9.134784	1.145685	-7.97
DiamCopa	1	2.577742	0.664242	3.88
GlovHoll	1	8.436071	2.332765	3.62
h	1	0.684598	0.167842	4.08
PosSocial 1	1	9.829625	2.675333	3.67
PosSocial 2	1	2.465117	2.597871	0.95
PosSocial 3	1	-1.607902	1.213608	-1.32



Os resultados mostram que o PROC GLMSELECT construiu cinco modelos de regressão conforme a tabela de sumário do método Forward. Logo, utilizou os dados de validação para decidir se o modelo com as quatro variáveis selecionadas (DiamCopa, GlovHoll, h e PosSocial) foi o modelo que melhor equilibrava a parcimônia e a precisão da predição.

O gráfico da Average Squared Error (ASE) mostra a trajetória do erro padrão médio para cada um dos cinco modelos construídos pelo PROC GLMSELECT iniciando pelo modelo somente com o intercepto (intercept). A aplicação do modelo para os dados de

validação mostrou pouco incremento na acurácia entre os modelos com somente a variável altura (1+h) e o modelo final (4+GlovHoll).

3.7. Regressão não-linear

Objetivos de aprendizagem:

- i) Descrever os principais modelos não-lineares considerados para o crescimento de árvores;
- ii) Demonstrar o uso de dois procedimentos SAS para o ajuste de modelos não-lineares;
- iii) Mostrar como realizar a comparação dos coeficientes de regressão para uma covariável de sítio;
- iv) Indicar os critérios estatísticos considerados para avaliar o ajuste de modelos não-lineares.

Um modelo de regressão é considerado não-linear se as derivadas com relação aos coeficientes de regressão dependem de um ou mais coeficientes. Um modelo de regressão utilizado na Ciência Florestal para descrever o comportamento assintótico de y_i em função de x_i sem ponto de inflexão é o de Mitscherlich (MITSCHERLICH, 1919) que tem a seguinte expressão matemática básica:

$$y_i = A(1 - e^{(-kx_i)}) + \varepsilon_i$$

Em que:

y_i = variável dependente na i -ésima unidade amostral (Taxa de germinação, diâmetro, altura, volume, peso, etc.);

x_i = variável independente (Tempo em dias, meses anos ou em outra escala; quantidade de irrigação, fertilizante e etc.);

e =base do logaritmo natural=2,718282;

A, k =coeficientes de regressão a serem estimados;

ε_i =resíduos independentes distribuídos $n(0, \sigma^2)$;

As derivadas do modelo de Mitscherlich de y_i com relação aos coeficientes resultam em:

- Derivada de y_i em relação a $A \rightarrow \frac{\partial y_i}{\partial A} = 1 - \exp(-k \cdot x_i)$
- Derivada de y_i em relação a $k \rightarrow \frac{\partial y_i}{\partial k} = A \cdot x_i \cdot \exp(-k \cdot x_i)$

Observa-se que as derivadas resultantes para A e k dependem de coeficientes de regressão.

Antes do advento computacional, alguns modelos não-lineares eram ajustados por mínimos quadrados ordinários devido à facilidade do processo de cálculo comparado ao ajuste de um modelo não-linear. Esse processo é possível desde que um modelo não-linear seja passível de linearização, ou seja, tornar um modelo não-linear para linear. O processo de tornar-se um modelo não-linear para linear é realizado considerando as seguintes propriedades operatórias dos logaritmos:

- $\text{Log}_a(xy) = \text{Log}_a(x) + \text{Log}_a(y)$
- $\text{Log}_a\left(\frac{x}{y}\right) = \text{Log}_a(x) - \text{Log}_a(y)$
- $\text{Log}_a(x)^p = p \text{Log}_a(x)$

Neste caso, para o modelo não-linear $y_i = \beta_0 \cdot x_i^{\beta_1} + \varepsilon_i$ a aplicação das propriedades operatórias dos logaritmos o tornará linear da seguinte forma:

- Passo 1 - Logaritmo em ambos os termos do modelo:
- $\text{Log}(y_i) = \text{Log}(\beta_0 \cdot x_i^{\beta_1})$
- Passo 2 - Transformar o produto em soma para tornar os coeficientes na forma aditiva: $\text{Log}(y_i) = \text{Log}(\beta_0) + \text{Log}(x_i^{\beta_1})$
- Passo 3 – Retirar a potência β_1 para a forma multiplicativa por x :
- $\text{Log}(y_i) = \beta_0 + \beta_1 \text{Log}(x_i)$

Neste caso, o modelo linearizado $\text{Log}(y_i) = \beta_0 + \beta_1 \text{Log}(x_i) + \phi_i$ pode ser ajustado por mínimos quadrados ordinários, entretanto, o termo de resíduo ϕ_i não é equivalente ao termo de resíduo ε_i . Portanto, a linearização resulta apenas em uma aproximação da forma não-linear desejada.

O processo de linearização originou os modelos de regressão logaritmos existentes hoje como, por exemplo, o modelo de volume de Schumacher e Hall amplamente utilizado na Ciência Florestal. Entretanto, Neter et al. (1996) ressaltaram que a linearização de um modelo afeta o termo de resíduos (ε_i) e, portanto, indica que a versão não-linear de um modelo é preferível do que sua versão linearizada.

Vale ressaltar que nem todos os modelos de regressão não-linear podem ser linearizados como o exemplo a seguir:

$$y_i = \text{sen}(\beta_1 x_i) + \varepsilon_i$$

Outra diferença importante de modelos de regressão não-lineares é que o número de coeficientes de regressão não necessariamente é diretamente relacionado ao número de variáveis independentes no modelo (NETER et al., 1996).

Exemplo clássico de um modelo não-linear é o exponencial simples com a seguinte forma matemática descrita na Figura 45:

$$y_i = \beta e^{kx_i} + \varepsilon_i$$

Em que:

y_i =variável dependente observado na i -ésima unidade amostral;

x_i =variável independente observada;

β =coeficiente de regressão que representa o valor inicial de y_i ;

e =base do logaritmo natural=2,718282;

K =coeficiente de regressão que controla a taxa de mudança da curva. Neste caso, se $k>0$ os valores de y_i na curva aumentam em função de x_i e diminuem se $k<0$;

ε_i =resíduos independentes distribuídos $n(0, \sigma^2)$.

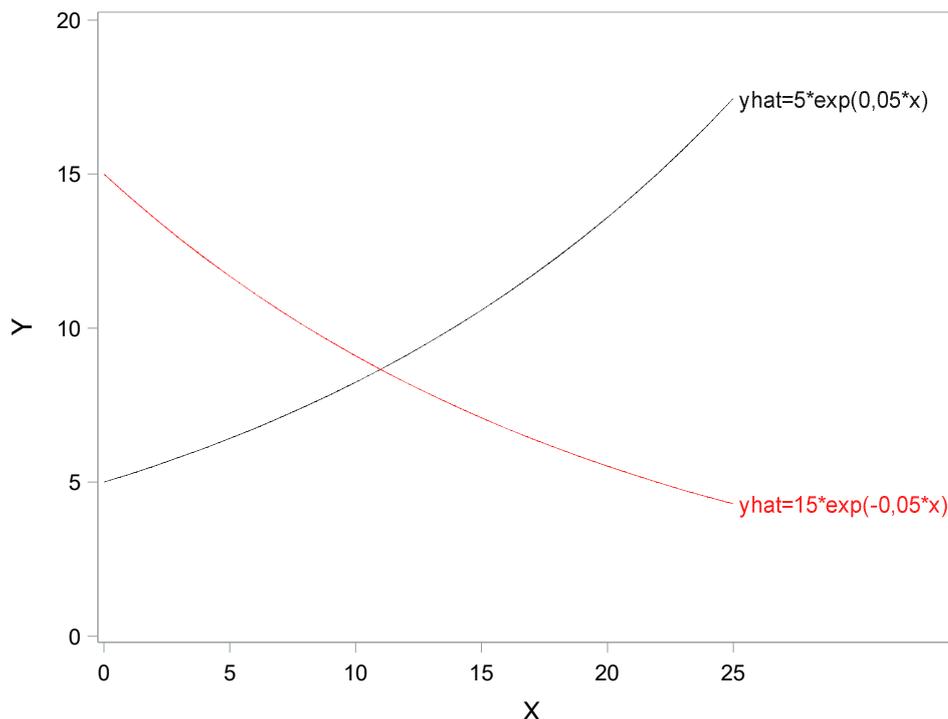


Figura 45. Comportamento de aumento exponencial e decaimento de acordo ao valor do coeficiente k no modelo exponencial simples.

3.7.1. Modelos de curva de tamanho-idade

Na ciência florestal é muito comum coletar dados para o estudo do crescimento de árvores (como por exemplo, crescimento em diâmetro, altura ou volume quando associado à idade dos indivíduos), bem como a construção de curvas de índice de sítio. Neste caso o comportamento bivariado tem uma consideração biológica na qual a resposta apresenta um valor assintótico à medida que a variável independente cresce e tende ao infinito.

A forma da curva resultado dessa relação é, geralmente, complexa (Sigmoidal, por exemplo) ou mais simples para indivíduos que apresentam taxa de crescimento muito alta ao longo da vida como por exemplo árvores de Eucalipto que cresceram em sítio bom.

A Figura 46 mostra o comportamento do crescimento em diâmetro a altura do peito (d) em função da idade observado em árvores de uma determinada espécie. Os dados foram obtidos a partir de análise de tronco completa em um povoamento de 25 anos.

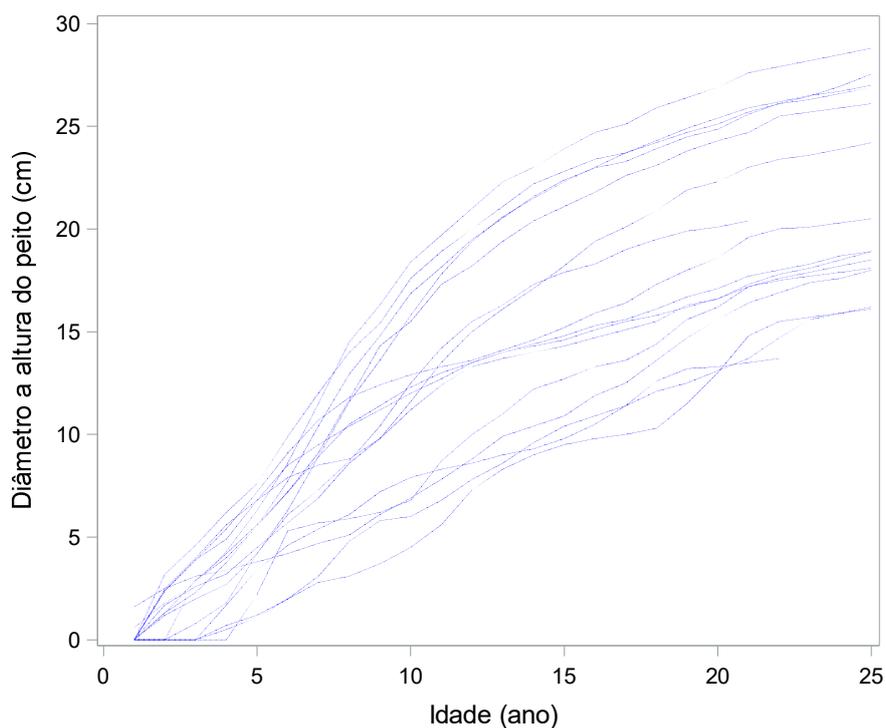


Figura 46. Comportamento do diâmetro a altura do peito (d) em função da idade para 16 árvores.

No caso dos dados de crescimento observado na Figura 46 seria possível ajustar modelos lineares diversos da literatura ou considerar modelos não-lineares para representar o crescimento em diâmetro ao longo do tempo.

Modelos lineares empíricos são estabelecidos a partir da relação entre a variável dependente e variáveis independentes (modelo linear simples, modelo polinomial, etc.).

Por outro lado, modelos de crescimento são funções não-lineares analíticas ou funcionais derivadas a partir de proposições lógicas sobre as relações entre as variáveis.

Portanto, a utilização de modelos não-lineares para representar dados de crescimento como os da Figura 46 é justificado em base às seguintes características (Adaptado de RICHARDS, 1959):

- i) Modelos não-lineares são construídos com base em considerações biológicas com conceituação quantitativa do processo em estudo;
- ii) Os coeficientes de regressão de um modelo não-linear de crescimento geralmente têm interpretação direta do processo em estudo e permite avançar no conhecimento das relações funcionais entre as variáveis;
- iii) Modelos não-lineares são flexíveis e podem ter restrições nos coeficientes de regressão.

Alguns modelos de regressão não-linear utilizados na área florestal para descrever matematicamente o comportamento do crescimento de árvores em função do tempo ou idade são apresentados no Quadro 30.

Quadro 30. Alguns modelos de regressão não-linear para descrever a curva tamanho-idade utilizados na Ciência Florestal.

Autor ou designação	Expressão matemática da curva tamanho-idade	Características		
		Número de coeficientes	Simetria com relação ao ponto de inflexão	Ponto de inflexão (Localização)
1. Chapman-Richards	$y_i = A(1 - e^{(-kx_i)})^c + \varepsilon_i$	3	Simétrico	$y_{inf} = [1 - (1/c)]^c$
2. Logística	$y_i = \frac{A}{1 + Be^{(-kx_i)}} + \varepsilon_i$	3	Simétrico	$y_{inf} = \frac{A}{2}$ $x_{inf} = \frac{1}{k} \cdot \ln(B)$
3. Gompertz	$y_i = Ae^{(-Be^{(-kx_i)})} + \varepsilon_i$	3	Assimétrico	$y_{inf} = \frac{A}{e}$ $x_{inf} = \frac{B}{k}$
4. Mitscherlich	$y_i = A(1 - Be^{(-kx_i)}) + \varepsilon_i$	3	-	Não tem!
5. Richards	$y_i = A(1 - Be^{(-kx_i)})^{\frac{1}{1-m}} + \varepsilon_i$	4	Simétrico	$y_{inf} = Am^{1/1-m}$ $x_{inf} = \frac{\ln(\frac{B}{1-m})}{k}$
6. Schnute	$y_i = \left[y_1^b + (y_2^b - y_1^b) \frac{1 - e^{(-a(x_i - \tau_1))}}{1 - e^{(-a(\tau_2 - \tau_1))}} \right]^{1/b} + \varepsilon_i$	4	-	

Em que: y_i = Tamanho (Diâmetro, altura, volume, peso, etc.) mensurado na idade x_i ; e =base do logaritmo natural=2,718282; $A, a, B, b, c, k, m, y_1, y_2$ =coeficientes de regressão a serem estimados em cada modelo; τ_1, τ_2 = idade no tempo 1 (início) e 2 (final) dos dados observados, respectivamente; ε_i =resíduos independentes distribuídos $N(0, \sigma^2)$; y_{inf} =valor da variável dependente no ponto de inflexão; x_{inf} =idade no ponto de inflexão.

Fonte: Adaptado de LOETSCH et al. (1973); MITSCHERLICH e SONNTAG (1982); RICHARDS (1959); WINSOR (1932); SWEDA e KOUKETSU (1984); LEI e ZANG (2006).

Em todos os modelos de regressão do Quadro 30 o coeficiente de regressão “A” representa a assíntota, ou seja, a variável dependente (diâmetro, altura, volume, biomassa, etc.) continuará a aumentar até chegar ao valor da Assíntota e a partir daí estabilizará. Em geral, o coeficiente da Assíntota está presente na maioria das funções de crescimento e indica, o valor máximo que a variável dependente pode atingir.

Por outro lado, o coeficiente “k” representa a taxa de crescimento, ou seja, indica quão rapidamente os valores de y_i subirá uma vez iniciada a curva ao longo da idade. Portanto, influencia o declive/active da curva de crescimento na qual valores maiores indicam uma subida rápida de zero em direção à assíntota. Em modelos de crescimento k sempre será positivo e possui a mesma unidade de y .

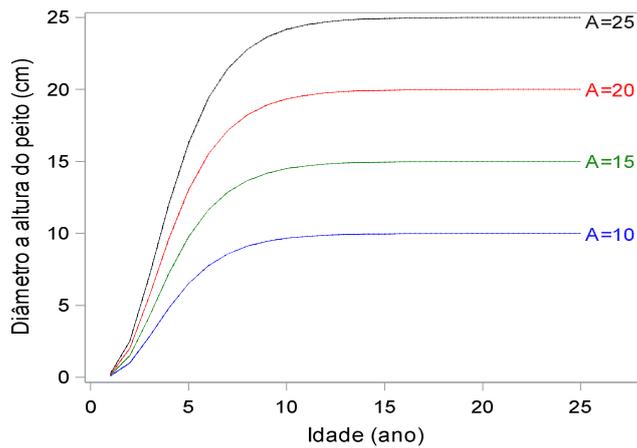
3.7.2. Características e formas das curvas de tamanho-idade

O modelo de Chapman-Richards é utilizado amplamente para descrever o crescimento de árvores e povoamentos florestais sendo derivado do modelo de von Bertalanffy, onde a constante de valor igual a $2/3$ é substituída pelo coeficiente de regressão “c” (ZEIDE, 1993). De outra forma, se o valor $c=1$ ou $c=-1$, o modelo de Chapman-Richards assume a função Monomolecular ou a função Logística, respectivamente.

Esse modelo de regressão é uma versão com três coeficientes do modelo de Richards e possui o ponto de partida da curva na origem dos eixos na qual, sua utilização é adequada para dados de crescimento iniciando em zero (TJORV; TJORVE, 2010).

O coeficiente B presente nos modelos 2 ao 4 do Quadro 30 é utilizado para ajustar o valor inicial do período de crescimento quando a suposição $x=0, y=0$ não ocorrer, ou seja, dados observados em que $y_0 \neq 0$ ou $x_0 \neq 0$. Por exemplo, Ware et al. (1982) utilizaram esse coeficiente para considerar o efeito da concentração de nutrientes (x) no rendimento de tecidos vegetais (y) sob a suposição inicial de que em $x=0$ existe algum rendimento (y_0). Portanto, esse coeficiente tem pouco significado biológico visto que reflete a escolha do valor zero no tempo de estudo (RICHARDS, 1959) sendo conhecido como coeficiente de escala.

O modelo de Chapman-Richards produz várias formas de curvas de acordo aos valores de cada um dos três coeficientes de regressão. O impacto da mudança nos valores dos coeficientes na forma da curva de crescimento é observado na Figura 47 a partir de dados simulados de diâmetro a 1,3 m (D_i) em função da idade (I_i).

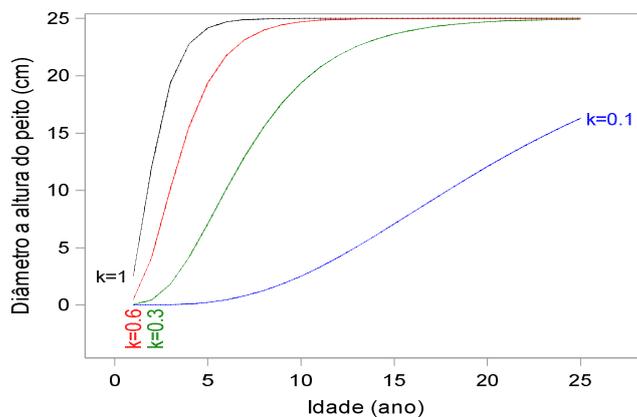


$$D = A. [1 - e^{(-k.I_i)}]^c + \varepsilon_i$$

$A = \text{variando}$

$k = 0,5$

$c = 5,0$

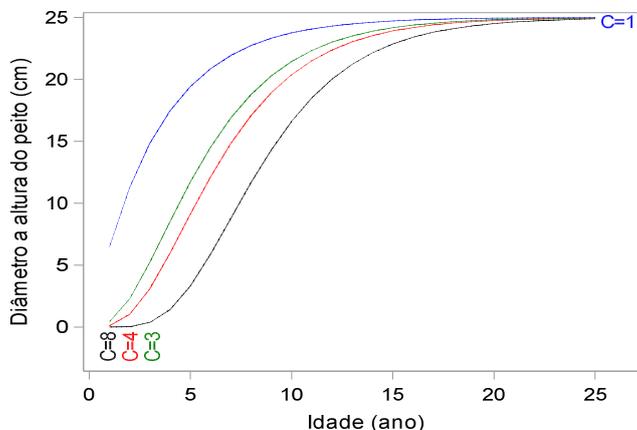


$$D = A. [1 - e^{(-k.I_i)}]^c + \varepsilon_i$$

$A = 25$

$k = \text{variando}$

$c = 5,0$



$$D = A. [1 - e^{(-k.I_i)}]^c + \varepsilon_i$$

$A = 25$

$k = 0,3$

$c = \text{variando}$

Figura 47. Flexibilidade do modelo de Chapman-Richards (Modelo 1 do Quadro 30) diante da variação dos coeficientes de regressão para curvas simuladas.

A partir da Figura 47 é possível observar que a forma sigmoide das curvas permanece invariável diante da variação dos valores da assíntota (A).

Por outro lado, as curvas com a variação do coeficiente relacionado à taxa de crescimento (k) impacta diretamente a posição do ponto de inflexão. À medida que o valor do coeficiente k aumenta, mais cedo em idade, o ponto de inflexão estará localizado sob a

curva. Adicionalmente, a mudança desse coeficiente impacta a idade em que o organismo alcança a assíntota. Quando $k=0,1$ a curva indica claramente que a assíntota ocorrerá em idade maior do que 25 anos e quando $k=0,6$ a assíntota é alcançada próximo a idade de 10 anos.

Quando ocorre variação no coeficiente c de Chapman-Richards, as curvas mantem a mesma localização da assíntota com forma sigmoide exceto para $c=1$ quando a curva demonstra uma representação do modelo de Mitscherlich (Modelo 4 do Quadro 30), também conhecido como modelo de crescimento natural, ou seja, uma curva sem ponto de inflexão.

3.7.3. Características e variações dos modelos de crescimento

Em relação à localização e número de coeficientes de regressão, existem várias modificações a partir dos modelos de crescimento do Quadro 30 considerados por diversos autores.

Uma adequação ao modelo de Chapman-Richards pode ser realizada com o intuito de adicionar um coeficiente de regressão no modelo original para representar o valor de ponto de partida da curva (T_0) modificando o modelo para a expressão matemática:

$$y_i = A(1 - e^{(-k(x_i - T_0))})^c + \varepsilon_i$$

ou

$$y_i = \frac{A}{(1 - e^{(-k(x_i - T_0))})^{-c}} + \varepsilon_i$$

O modelo Logístico do Quadro 30 também pode ser modificado. Uma variação desse modelo foi utilizada por Price et al. (2017) para descrever a relação de resposta frente a dose de herbicidas para o controle de plantas daninhas. A expressão matemática considerada pelos autores incluiu um coeficiente de regressão T_i :

$$y_i = \frac{A}{1 + e^{(k(x_i - T_i))}} + \varepsilon_i$$

Neste caso, a modificação do modelo Logístico possibilita obter a localização do ponto de inflexão diretamente pelo coeficiente de regressão estimado sendo a coordenada no plano cartesiano do mesmo dada por: T_i ; $A/2$.

Outra modificação do modelo Logístico inclui um coeficiente de regressão que representa o valor inicial da variável dependente quando o valor de x for zero ($x=0$) representada pelo coeficiente y_0 (TJORVE; TJORVE, 2010):

$$y_i = \frac{A}{1 + \left(\frac{y_0}{A-1}\right) e^{(-kx_i)}} + \varepsilon_i$$

É importante notar que o fato do modelo Logístico ser simétrico com relação ao ponto de inflexão, torna-o menos flexível que outras funções sigmóides não simétricas e, portanto, essa propriedade pode limitar sua capacidade de se ajustar bem aos dados de crescimento observado.

O modelo de Gompertz foi originalmente derivado por Benjamin Gompertz para estimar a mortalidade humana e publicado em 1825 no periódico Phil. Trans. Roy. Soc. London. Uma característica importante do modelo de Gompertz é a assimetria da curva de crescimento com respeito ao ponto de inflexão. Desta forma, recomenda-se utilizar a curva de crescimento de Gompertz para modelar o crescimento sigmoide nos quais o ponto de inflexão está localizado em aproximadamente 1/3 do tamanho máximo possível a ser atingido no tempo.

Originalmente o modelo de Mitscherlich foi proposto para avaliar o efeito do crescimento de culturas agrícolas tratadas com diferentes doses de adubação química em experimentos agrícolas (MITSCHERLICH, 1919). A forma mais comum desse modelo é:

$$y_i = A(1 - e^{(-kx_i)}) + \varepsilon_i$$

Em que:

y_i =rendimento da i -ésima cultura;

x_i = i -ésima dose de adubação no experimento;

A =máximo rendimento alcançado no experimento;

k =constante de proporcionalidade ou fator de efeito de x_i no rendimento.

A suposição desse modelo criada por Mitscherlich é que a relação entre a quantidade de fertilizante e o rendimento da colheita é descrita pela Lei dos incrementos decrescentes. Ademais, a constante de proporcionalidade (k) inicialmente foi considerada como constante e igual para todas culturas agrícolas, época do ano ou sítio.

O modelo de Mitscherlich não possui ponto de inflexão e pode ser utilizado para expressar o desenvolvimento de árvores na fase inicial ou final do crescimento ou para espécies que apresentam comportamento do crescimento ao longo do tempo sem um ponto de inflexão definido.

É possível generalizar o modelo de Mitscherlich a fim de incluir mais variáveis considerando a hipótese de que à medida que se aumenta a quantidade de fertilizante (ou outra variável) pode haver efeito negativo na cultura, denominado por Mitscherlich de “Fator de Injúria”. Neste caso, a expressão matemática com mais de uma variável independente é:

$$y_i = A(1 - e^{(-k_1x_1)})(1 - e^{(-k_2x_2)}) \dots (1 - e^{(-k_nx_n)})$$

Essa modificação torna o modelo de Mitscherlich mais flexível e tem pouca influência no valor da assíntota. Entretanto, pode não ser útil e sem sentido biológico para descrever dados de relação altura do fuste – diâmetro a 1,3 m, por exemplo, visto que à medida que o diâmetro aumenta a altura não diminuirá.

Comparado aos demais modelos do Quadro 30, o de Richards (1959) possui um quarto coeficiente de regressão “*m*” permitindo uma maior flexibilidade para se ajustar a uma gama de curvas sigmóides pois a depender do valor estimado para o coeficiente “*m*”, o modelo de Richards representa uma transição para as seguintes curvas (RICHARDS, 1959):

- $m=0 \rightarrow$ modelo Monomolecular;
- $m=2/3 \rightarrow$ modelo von Bertalanffy;
- $m \approx 1 \rightarrow$ modelo se limita à de Gompertz;
- $m \approx 1$ a $2 \rightarrow$ transição entre o modelo de Gompertz para a Logística.

O ponto de inflexão sob a curva estará localizado em menor idade (mais cedo) à medida que for menor o valor de “*m*” e maior “*k*”. Portanto, a correlação entre “*m*” e a assíntota “*A*” é negativa.

Esse modelo foi desenvolvido a partir de uma generalização do modelo de crescimento de von Bertalanffy no qual originalmente o valor de “*m*” equivale 2/3 denominado por Bertalanffy de constante alométrica.

Pineaar e Turbull (1973) relataram que em seu estudo, Richards (1959) juntamente com Chapman (1960) generalizaram o modelo de von Bertalanffy como forma de tornar seu

uso mais amplo para outras formas de crescimento de organismos vivos com a mudança da constante 2/3 para o coeficiente de regressão “*m*” que denominaram de constante de alometria da relação entre *y* e *x*.

Várias modificações desse modelo foram realizadas podendo ter as seguintes expressões matemáticas para representar T_0 e T_i (TJORVE; TJORVE, 2010):

$$y_i = A(1 - (1 - m)e^{(-k(x_i - T_i))})^{\frac{1}{1-m}} + \varepsilon_i$$

Ou

$$y_i = A \left(1 - \left(\frac{1}{m} \right) e^{(-k(x_i - T_i))} \right)^m + \varepsilon_i$$

O modelo de Gompertz foi alterado para estudar a relação entre a altura total de árvores (*h*) e o diâmetro a altura do peito (*d*) pela adição da constante 1,3 no modelo sendo a expressão matemática:

$$y_i = 1,3 + Ae^{-Be^{-kx_i}} + \varepsilon_i$$

Em que: y_i =Altura total da *i*-ésima árvore; x_i =Diâmetro a 1,3 m.

O modelo de crescimento de Schnute foi desenvolvido por Schnute (1981) inicialmente para aplicação em pesquisa pesqueira, mas com o passar do tempo, teve sua aplicação na ciência florestal iniciando por Bredenkamp e Gregoire (1988) e depois outras pesquisas como a de Lei e Zang (2006) para descrever a relação hipsométrica de árvores.

Esse modelo possui quatro coeficientes de regressão que impõem grande flexibilidade na curva de crescimento-idade de acordo aos valores dos coeficientes “*a*” (relacionado à taxa de crescimento) e “*b*” (relacionado à forma da curva) tornando a curva sigmoidal quando os valores dos coeficientes $a > 0$ e $0 < b < 1$ ou $a > 0$ e $b < 0$ (SCHNUTE, 1981).

Os coeficientes y_1 e y_2 representam o valor de *y* (tamanho da variável dependente) na idade de início do crescimento (τ_1) e idade máxima dos dados observados (τ_2), respectivamente. A assíntota no modelo de Schnute é calculada considerando a combinação de todos os quatro coeficientes de acordo à seguinte expressão:

$$\text{Assíntota} = \left[\frac{e^{a\tau_2} y_2^b - e^{a\tau_1} y_1^b}{e^{a\tau_2} - e^{a\tau_1}} \right]^{\frac{1}{b}}$$

A curva de Schnute pode assumir a forma côncava (sem ponto de inflexão) quando os valores dos coeficientes $a > 0$ e $b > 1$, neste caso, utilizado por Coble e Lee (2006) para a construção de curvas de índice de sítio. Portanto, o modelo de Schnute pode se ajustar automaticamente a um conjunto de dados com comportamento sigmoide ou não caracterizando esse modelo como mais flexível do que o modelo de Chapman-Richards.

Quando o modelo de Schnute assume valor zero para o coeficiente “ b ” a forma do modelo assume um caso especial do modelo de Gompertz com a seguinte expressão matemática (SCHNUTE, 1981):

$$y_i = y_1 \left[\exp \left(\text{Log} \left(\frac{y_2}{y_1} \right) \frac{1 - e^{-a(x_i - \tau_1)}}{1 - e^{-a(\tau_2 - \tau_1)}} \right) \right] + \varepsilon_i; \text{ para } a > 0 \text{ e } b = 0$$

3.7.4. Ajuste de regressão não-linear no SAS System

O ajuste de modelos não-lineares ocorre pelo método iterativo em que os valores dos coeficientes de regressão são modificados até que a Soma de Quadrados dos Resíduos (SQR) seja a mínima possível a partir de diferentes fases iterativas pelo método de mínimos quadrados ordinários não-lineares (BATES; WATTS, 2007).

Portanto, modelos não lineares são mais difíceis de ajustar do que modelos lineares. A dificuldade aumenta à medida que um modelo é mais complexo em sua expressão matemática e na quantidade de coeficientes de regressão.

No SAS os procedimentos PROC NLIN e PROC NLMIXED ajustam modelos de regressão não-linear para variáveis dependentes (y_i) do tipo contínua com efeito fixo no contexto de modelo geral. Entretanto, o PROC NLMIXED foi desenvolvido especificamente para modelos não-lineares considerando efeito fixo e aleatório (modelo misto) no contexto de modelos generalizados.

Para utilizar o procedimento PROC NLIN é necessário especificar:

- O nome dos coeficientes de regressão (geralmente letras com números utilizadas nos modelos de regressão, como: $\beta_0, \beta_1, \theta_1, A, k$, etc.);
- Os valores “sementes” para cada coeficiente;
- A expressão matemática do modelo de regressão; e

- As derivadas parciais do modelo com respeito a cada coeficiente de regressão $\partial y_i / \partial \beta_0, \partial y_i / \partial \beta_1, \dots$ (Opcional).

As derivadas parciais do modelo a ser ajustado passou a ser opcional a partir da versão 6.12 do SAS mediante a inclusão de uma solução que calcula as derivadas (Differentiator) automaticamente durante o processo interativo de ajuste do modelo.

Toda essa especificação é utilizada no processo interativo por meio de algoritmos de otimização que o PROC NLIN considera para estimar os coeficientes de regressão a partir dos valores sementes. Os algoritmos de otimização disponíveis no PROC NLIN são:

- Método de Gauss-Newton modificado;
- Método de Marquardt;
- Método de gradient;
- Método de Newton.

Todos os quatro métodos utilizam derivadas ou aproximações às derivadas da SQR em relação aos coeficientes de regressão para orientar o software à busca pelos valores estimados dos coeficientes de regressão. A escolha de cada um dos algoritmos vai depender da complexidade do modelo de regressão. Geralmente, se o método de Gauss-Newton falhar em determinar os coeficientes de regressão (não convergir), recomenda-se utilizar o método de Marquardt.

A sintaxe padrão do PROC NLIN suficiente para ajustar um modelo de regressão não-linear em que a variável dependente é contínua é a seguinte:

```
proc nlin data= nome_do_dataset method=algoritmo maxiter=número_de_interações;
  parameters coeficiente1=valor_semente coeficiente2=valor_semente...;
  model variável_dependente = expressão_do_modelo;
  der.coeficiente1=expressão_da_derivada;
  der.coeficiente2= expressão_da_derivada;
  output out=nome_arquivo_saída;
run;
```

Na declaração METHOD= deve-se indicar qual algoritmo de otimização será utilizado para calcular os coeficientes de regressão por meio do processo interativo. Caso essa declaração não seja utilizada, o SAS considera o método de Gauss-Newton por padrão.

Por padrão, o SAS considera um máximo de 100 interações para minimizar a SQR. É possível que 100 interações não sejam suficientes para determinar os valores dos coeficientes de regressão a partir dos valores sementes indicados. Neste caso, utilizando a opção MAXITER= é possível aumentar o número de interações do algoritmo para verificar se ocorre a convergência.

Uma estratégia para indicação de valores sementes é utilizar os prováveis valores dos coeficientes otimizados na última interação que falhou em convergir. Logo, repetir o processamento dos dados.

A declaração PARAMETERS (Ou PARMS) define os valores sementes de cada coeficiente do modelo de regressão indicados na declaração. É possível estabelecer um grid de opções para os valores sementes de cada coeficiente de regressão a fim de tornar mais eficiente o processo interativo dos algoritmos. Neste caso, basta indicar a amplitude desejada que o PROC NLIN irá limitar o processo interativo com os valores indicados conforme exemplo a seguir extraído do manual SAS para um modelo não-linear com cinco coeficientes de regressão (Betas):

parameters B0=0

B1=4 to 8

B2=0 to 0.6 by 0.2

B3=1, 10, 100

B4=0, 0.5, 1 to 4;

Com essa conformação o PROC NLIN calcula a SQR (Soma de Quadrados dos Resíduos) para cada uma das 360 combinações possíveis de valores sementes ($1 \cdot 5 \cdot 4 \cdot 3 \cdot 6=360$). Entretanto, para determinar essas amplitudes de valores é necessário conhecimento prévio do comportamento da curva a ser ajustada bem como uma interpretação biológica do modelo de regressão a ser ajustado.

A expressão matemática do modelo de regressão é declarada em MODEL podendo ser escrita na íntegra ou dividida em partes por meio da criação de expressões dentro do procedimento.

A declaração $DER \cdot B0=$ especifica a derivada parcial com respeito ao coeficiente de regressão $B0$, por exemplo. Também é possível utilizar a segunda derivada parcial por meio da seguinte forma da declaração $DER \cdot B0 \cdot B1=$.

3.7.4.1. Aplicação com o modelo de Mitscherlich

Para fins de aplicação do ajuste de um modelo não-linear, será considerado o caso florestal 8 para modelar a relação hipsométrica de 15 espécies florestais nativas da Amazônia.

Caso florestal 8: Ajuste do modelo de Mitscherlich para dados de Altura - Diâmetro

Considere que uma investigação foi realizada com o objetivo de descrever a relação hipsométrica de árvores da floresta Amazônica. Para tal, se dispõe de um conjunto de dados de diâmetro a altura do peito (d) e altura do fuste (hf) para 15 espécies totalizando 131252 árvores oriundas de áreas de Plano de Manejo Florestal Sustentável. Para o estudo considerou-se: i) a altura do fuste foi medida e não a altura total, ii) o desenvolvimento final da altura do fuste ocorre antes do que da altura total, mas o diâmetro a altura do peito continua seu desenvolvimento, iii) a relação hipsométrica de árvores não apresenta ponto de inflexão e iv) considerou-se árvores com $D \geq 30$ cm a exceção da espécie *Swietenia macrophylla* (SWIMAC). Portanto, considerou-se o modelo não-linear de Mitscherlich como adequado sem o coeficiente de regressão original b . A variável "spcode" representa o nome científico de cada espécie resumido pelas três primeiras letras do gênero seguido das três primeiras letras do epíteto.

A expressão matemática do modelo é a seguinte:

$$hf_i = A(1 - e^{(-kD_i)}) + \varepsilon_i$$

Em que:

hf_i =Altura do fuste (m) da i -ésima árvore;

D_i =Diâmetro a 1,3 m (cm) da i -ésima árvore;

A =representa o valor da altura do fuste quando diâmetro a 1,3 m tende ao infinito (Assíntota);

k =coeficiente relacionado ao desenvolvimento da altura do fuste;

ε_i =efeito dos resíduos

Para resolver o caso florestal 8 será utilizado a seguinte sintaxe do procedimento PROC NLIN:

```

data hipso;
  input obs umf upa ut arvore spcode d hf;
  datalines;
1 1 2r 1 20 AMBACR 90 16
2 1 2r 1 25 DIPODO 120 16
3 1 2r 1 1 BEREXC 76 15
.
.
.
131252 1 4 3 88 SWIMAC 115 17
;
proc sort data=hipso;
  by spcode;
run;

title "Modelo de Mitscherlich por espécie";
ods output ParameterEstimates=Parms;
proc nlin data=hipso method=marquardt;
  parameters A=25
             k=0.03 ;
  model hf = A*(1-exp(-k*d));
  by spcode;
  output out=Mitscherlich p=hf_Mitscherlich r=res_Mitscherlich;
run;

proc print data=Parms noobs;
run;

```

obs=observação; umf=número da unidade de manejo florestal; upa=número da unidade de produção anual; ut=número da unidade de trabalho; arvore=número da árvore; spcode=código da espécie; d=diâmetro a 1,3 m (cm); hf=altura do fuste (m).

Para ajustar o mesmo modelo para cada uma das 15 espécies utilizou-se a opção BY spcode. Neste caso, o SAS ajusta o modelo indicado na declaração MODEL para cada uma das 15 espécies desde que a variável “spcode” em questão esteja ordenada de forma ascendente ou descendente. Para isto, utilizou o PROC SORT antes do PROC NLIN. Os resultados são apresentados no Output 38 de forma resumida para a primeira espécie.

Output 38. Resultado do ajuste do modelo não-linear de Mitscherlich para a primeira espécie (Amburana acreana=AMBACR) ordenada pela variável spcode.

Iterative Phase			
Iter	A	k	Sum of Squares
0	25.0000	0.0300	51280.4
1	18.2990	0.0369	19119.9
2	18.6310	0.0400	18501.8
3	18.6689	0.0398	18500.9
4	18.6679	0.0398	18500.9
5	18.6679	0.0398	18500.9

NOTE: Convergence criterion met.

Estimation Summary	
Method	Marquardt
Iterations	5
R	8.523E-7
PPC(k)	1.903E-6
RPC(k)	0.000025
Object	1.13E-10
Objective	18500.91
Observations Read	1679
Observations Used	1679
Observations Missing	0

Note: An intercept was not specified for this model.

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Model	2	504221	252111	22852.4	<.0001
Error	1677	18500.9	11.0321		
Uncorrected Total	1679	522722			

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
A	18.6679	0.2082	18.2595	19.0764
k	0.0398	0.00217	0.0356	0.0441

	A	k
A	1.0000000	-0.9078247
k	-0.9078247	1.0000000

socode	Parameter	Estimate	StdErr	Alpha	LowerCL	UpperCL	tValue	Probt
AMBACR	A	18.6679	0.2082	0.05	18.2595	19.0764	89.65	<.0001
AMBACR	k	0.0398	0.00217	0.05	0.0356	0.0441	18.34	<.0001
APULEI	A	18.6566	0.0738	0.05	18.5119	18.8013	252.81	<.0001
APULEI	k	0.0587	0.00288	0.05	0.0531	0.0644	20.43	<.0001
ASTLEC	A	26.2065	0.0916	0.05	26.0268	26.3861	285.98	<.0001
ASTLEC	k	0.0444	0.000752	0.05	0.0429	0.0459	59.04	<.0001
BEREXC	A	26.7274	0.0383	0.05	26.6523	26.8025	697.81	<.0001
BEREXC	k	0.0439	0.000595	0.05	0.0427	0.0450	73.75	<.0001
CARMIC	A	24.3735	0.0567	0.05	24.2623	24.4847	429.74	<.0001
CARMIC	k	0.0312	0.000339	0.05	0.0305	0.0318	92.02	<.0001
CEDODO	A	17.0605	0.1858	0.05	16.6961	17.4248	91.81	<.0001
CEDODO	k	0.0354	0.00138	0.05	0.0327	0.0381	25.75	<.0001
CEIPEN	A	19.6712	0.0638	0.05	19.5461	19.7963	308.26	<.0001
CEIPEN	k	0.0533	0.00147	0.05	0.0504	0.0562	36.36	<.0001
CLARAC	A	17.8237	0.0713	0.05	17.6839	17.9636	249.86	<.0001
CLARAC	k	0.0438	0.000771	0.05	0.0423	0.0453	56.84	<.0001
DINEXC	A	19.4094	0.0793	0.05	19.2540	19.5648	244.81	<.0001
DINEXC	k	0.0274	0.000530	0.05	0.0263	0.0284	51.66	<.0001
DIPODO	A	18.4417	0.0651	0.05	18.3141	18.5693	283.32	<.0001
DIPODO	k	0.0429	0.000755	0.05	0.0415	0.0444	56.88	<.0001
HANIMP	A	25.7673	0.4633	0.05	24.8566	26.6779	55.61	<.0001
HANIMP	k	0.0340	0.00228	0.05	0.0295	0.0385	14.91	<.0001
HANSER	A	26.3209	0.2452	0.05	25.8402	26.8016	107.35	<.0001
HANSER	k	0.0267	0.000688	0.05	0.0253	0.0280	38.79	<.0001
HYMEXC	A	21.2864	0.1180	0.05	21.0551	21.5176	180.42	<.0001
HYMEXC	k	0.0382	0.000901	0.05	0.0364	0.0400	42.42	<.0001
MEZITA	A	17.5365	0.1414	0.05	17.2593	17.8137	124.02	<.0001
MEZITA	k	0.0461	0.00185	0.05	0.0425	0.0497	24.86	<.0001
SWIMAC	A	18.1318	0.1991	0.05	17.7411	18.5225	91.09	<.0001
SWIMAC	k	0.0323	0.00156	0.05	0.0293	0.0354	20.76	<.0001

A primeira tabela do Output 38 mostra as mudanças ocorridas nos valores sementes (mostrados na interação zero) ao longo das cinco interações necessárias para haver a convergência, ou seja, a otimização por Marquardt da Soma de Quadrados dos Resíduos até estabilizar no valor de 18500,9. O número de interações necessárias para a convergência depende da proximidade dos valores sementes aos valores da curva.

A terceira tabela mostra a análise da variância (Anova) indicando que o modelo de Mitscherlich foi significativo (p -valor $<0,0001$) e, portanto, adequado para descrever a relação altura diâmetro das árvores.

A tabela de parâmetros estimados está logo a seguir da Anova, e mostra os valores dos coeficientes de regressão com seus respectivos intervalos de confiança. Neste caso, o valor estimado para a assíntota (A) da espécie analisada equivale a 18,7 m, ou seja, a partir dos dados observados, estima-se que a altura do fuste alcançará no máximo esse valor independentemente do aumento do diâmetro a 1,3 m (diâmetro $\rightarrow \infty$).

Essa tabela também é utilizada para realizar o teste de hipótese para cada coeficiente de regressão. Neste caso, se o intervalo de confiança de um determinado coeficiente de regressão apresentar valores negativos e positivos, indica não significância, pois o valor estimado compreende o zero dentro do intervalo. Este não é o caso para todos os dois coeficientes do modelo sob análise.

A última tabela do PROC NLIN mostra a correlação entre os coeficientes de regressão. Neste caso o coeficiente k mostra uma correlação forte (-0,90) com a assíntota (A). Essa correlação se justifica pelo fato de a taxa de crescimento impactar diretamente a assíntota, ou seja, maior taxa de crescimento a assíntota será alcançada em menores valores de diâmetro.

A linha de programação “ODS OUTPUT PARAMETERESTIMATES=PARMS;” salva em um arquivo separado denominado “Parms” contendo todos os coeficientes de regressão para cada espécie considerada no estudo bem como outras informações importantes como a probabilidade t para cada coeficiente. Esse arquivo é salvo na livreria WORK do SAS e pode ser acessado com um duplo clique sob o arquivo ou solicitando um PROC PRINT do arquivo. Essa tabela é útil para realizar teste de hipótese para os coeficientes de regressão (Probt) e para realizar uma análise específica dos coeficientes por espécie.

A linha de programação “OUTPUT OUT=Mitscherlich P=hf_Mitscherlich R=res_Mitscherlich;” do PROC NLIN foi considerada para salvar os valores estimados da relação hipsométrica para cada espécie. Esses valores são utilizados para gerar diversos

gráficos de análise como as curvas produzidas pelas estimativas de Mitscherlich por meio da seguinte sintaxe do PROC SGPLOT:

```
proc sort data=Mitscherlich out=sorted;
  by spcode d;
run;

ods graphics / reset width=18cm height=18cm imagemap noborder;
proc sgplot data=sorted;
  series x=d y=hf_Mitscherlich / group=spcode transparency=0.1;

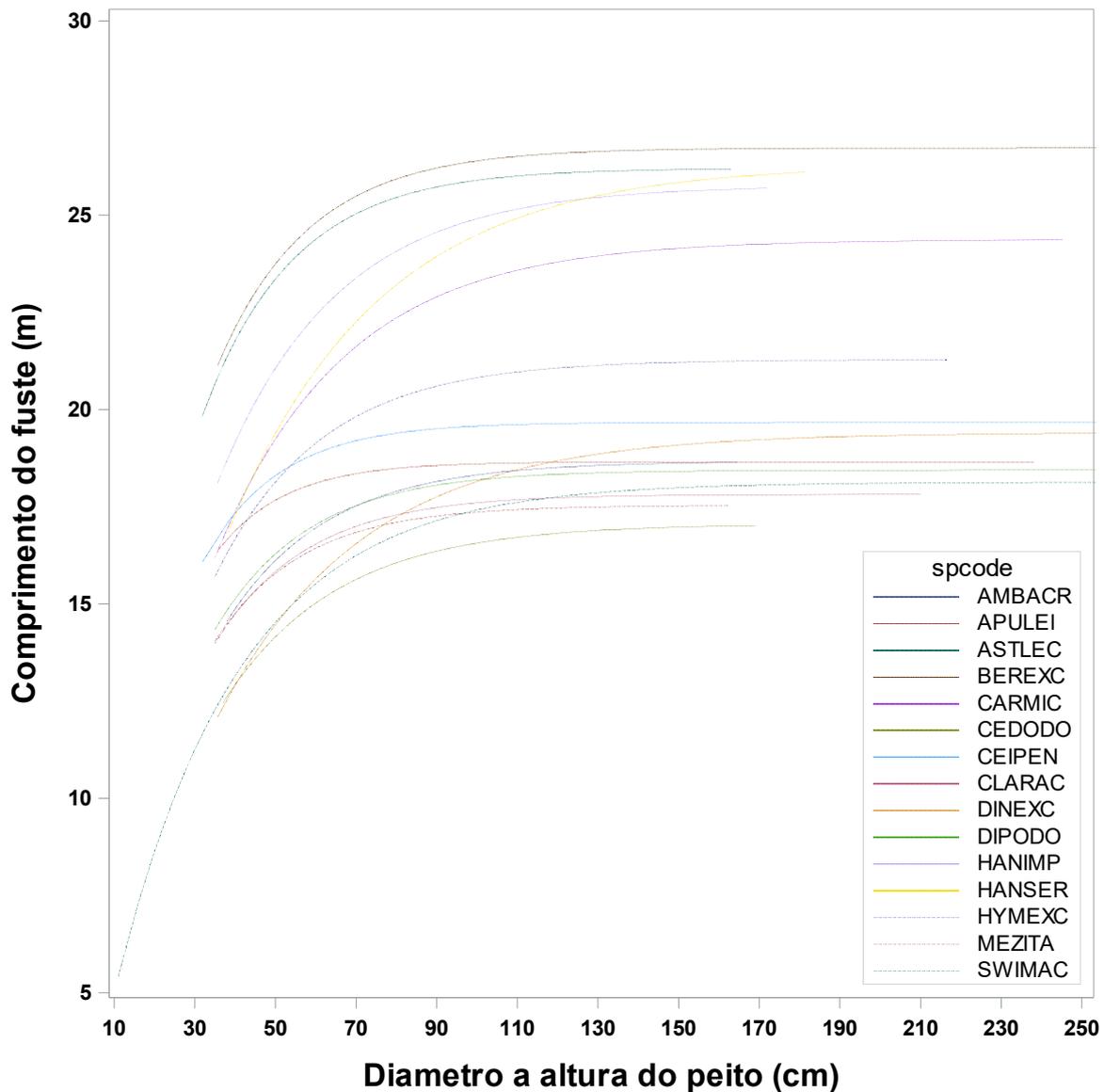
  xaxis integer values=(10 to 250 by 20) label="Diâmetro a altura do peito (cm)"
    LABELATTRS=(Color=black Family=Arial Size=13 Weight=Bold)
    VALUEATTRS=(Color=black Family=Arial Weight=Bold);

  yaxis integer values=(5 to 30 by 5) label="Comprimento do fuste (m)"
    LABELATTRS=(Color=black Family=Arial Size=13 Weight=Bold)
    VALUEATTRS=(Color=black Family=Arial Weight=Bold);

  keylegend / location=inside position=bottomright across=1;
run;
```

A linha de programação ODS GRAPHICS estabelece a largura e comprimento de 18 cm para o gráfico sem borda. Em seguida a declaração SERIES produz as curvas considerando no eixo x o diâmetro a 1,3 m e no eixo y os valores estimados da altura do fuste a partir do modelo de Mitscherlich. A opção GROUP=spcode plota todas as curvas para cada espécie e inclui uma legenda. É realizada uma edição para os eixos do gráfico dentro das declarações XAXIS e YAXIS para o tipo e tamanho de fonte bem como para negrito para a etiqueta e para os valores nos eixos. O resultado é um gráfico mostrado no Output 39:

Output 39. Curvas da relação hipsométrica para 15 espécies (spcode) nativas da Amazônia obtidas a partir de estimativas do modelo de regressão de Mitscherlich. A espécie *Swietenia macrophylla* (SWIMAC) apresenta valores de diâmetro a 1,3 a partir de 10 cm, as demais a partir de 30 cm.



A partir deste gráfico é possível observar um comportamento assintótico variável entre as espécies avaliadas com maior altura de fuste (26,7 m) para a espécie *Bertholletia excelsa* (BEXEXC) e menor altura de fuste (17,1 m) para a espécie *Cedrela odorata* (CEDODO).

O gráfico também mostra uma curva sem uma definição clara para a Assíntota na espécie *Handroanthus serratifolius* (HANSER) apesar do valor do coeficiente estimado ser de 26,3 m.

3.7.4.2. Aplicação com o modelo de Chapman-Richards para dados de crescimento

Para fins de aplicação o modelo 1 Chapman-Richards (Quadro 30) será considerado como exemplo de programação SAS no PROC NLIN para o ajuste a dados de crescimento em diâmetro em função da idade apresentados na Figura 48. A expressão matemática do modelo é a seguinte:

$$D_i = A(1 - e^{(-kI_i)})^c + \varepsilon_i$$

Em que:

D_i =Diâmetro a 1,3 m (cm);

I_i =idade da árvore em um determinado tempo;

A =representa o valor do diâmetro quando a idade tende ao infinito (Assíntota);

k =coeficiente relacionado à velocidade do crescimento;

c =coeficiente relacionado ao ponto de inflexão na curva;

ε_i =efeito dos resíduos.

A sintaxe a seguir constrói o gráfico de valores observados e estimados bem como o gráfico de resíduos. Os dados de saída do PROC NLIN foram suprimidos com a opção NOPRINT visto que tem o mesmo formato do ajuste descrito na secção anterior.

```
title "modelo de Chapman-Richards";
proc nlin data=uva_japao method=Marquardt noprint;
  parameters A=20
             k=0.1
             c=2;
  model Di = A*(1-exp(-k*I))**c;
  output out=C_Richards p=Di_estimado r=res_C_Richards;
run;
```

```

proc sort data=C_Richards out=nlin_out;
  by I;
run;

ods graphics / reset noborder width=12cm height=10cm imagemap;
proc sgplot data=nlin_out noautolegend;
  series x=I y=Di / group=árvore lineattrs=(color=blue thickness=1 pattern=solid)
transparency=0.5;
  series x=I y=Di_estimado / lineattrs=(color=red thickness=3 pattern=solid) transparency=0.1;
  label I="idade (ano)" Di="diâmetro a altura do peito (cm)";
run;

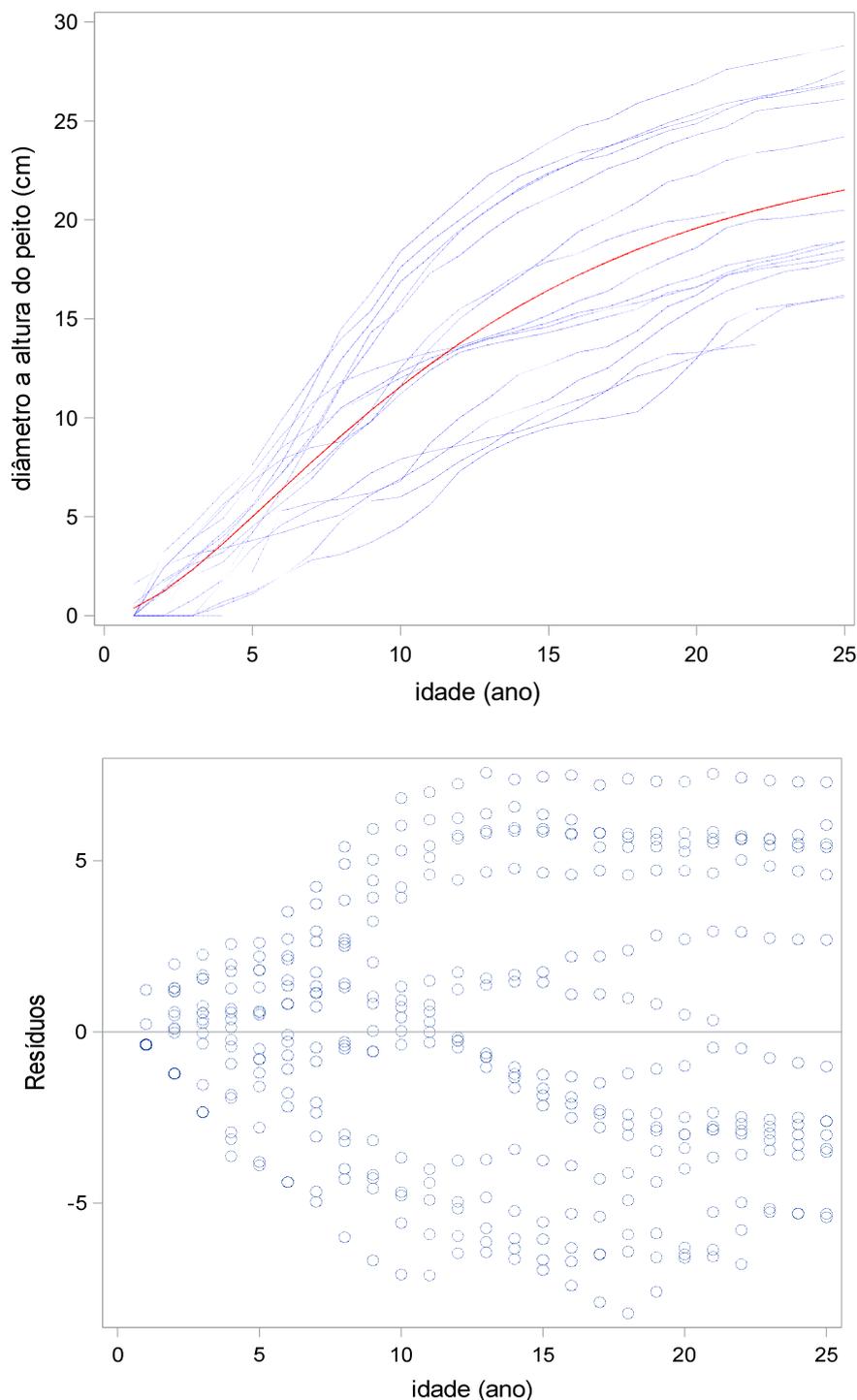
proc sgplot data=nlin_out noautolegend;
  scatter x=I y=res_C_Richards / transparency=0.1;
  refline 0 / axis=y;

  label I="idade (ano)" res_C_Richards="Resíduos";
run;

```

O resultado dos dois procedimentos PROC SGPLOT são os gráficos apresentados no Output 40. As curvas de diâmetro (D_i) observado (linhas azuis) e estimado pelo modelo de Chapman-Richards ($D_{i_estimado}$ =linha vermelha) são apresentadas no primeiro gráfico. Os dados observados apresentam aumento da variação do crescimento entre árvores à medida que aumenta a idade. Essa variação é refletida no gráfico dos resíduos e é causada, em parte, à qualidade do sítio em que as árvores cresceram (Neste caso, os dados são oriundos de três sítios), mas também devido à competição diferente que as árvores têm na floresta que afetam o crescimento em diâmetro.

Output 40. Curvas de crescimento para diâmetro a 1,3 m e gráfico de resíduos.



Portanto, para modelar o crescimento em diâmetro dentro de cada sítio, basta incluir no procedimento PROC NLIN a opção BY para ajustar o modelo dentro de cada um dos três sítios (sítio=sítio natural). Ademais, solicitou-se ao SAS a construção do gráfico de curva estimada por sítio pelo procedimento PROC SGPLOT. Os resultados são apresentados no Output 40 resultado da sintaxe a seguir.

```

title 'modelo de Chapman-Richards'; *y=A*[1-exp(-k*t)]**c;
proc nlin data=uva_japao method=marquardt;
  parameters A=20
             k=0.1
             c=2;
  model Di = A*(1-exp(-k*I))**c;
  by sitio;
  output out=C_Richards_SN p=Di_estimado_SN r=res_C_Richards_SN;
run;

proc sort data=C_Richards_SN;
  by I;
run;

ods graphics / reset noborder width=12cm height=10cm imagemap;
proc sgplot data=C_Richards_SN;
  series x=I y=Di / group=arv lineattrs=(color=blue thickness=1 pattern=solid)
  transparency=0.8;
  series x=I y=Di_estimado_SN /group=sitio lineattrs=(thickness=3 pattern=solid)
  transparency=0.1;

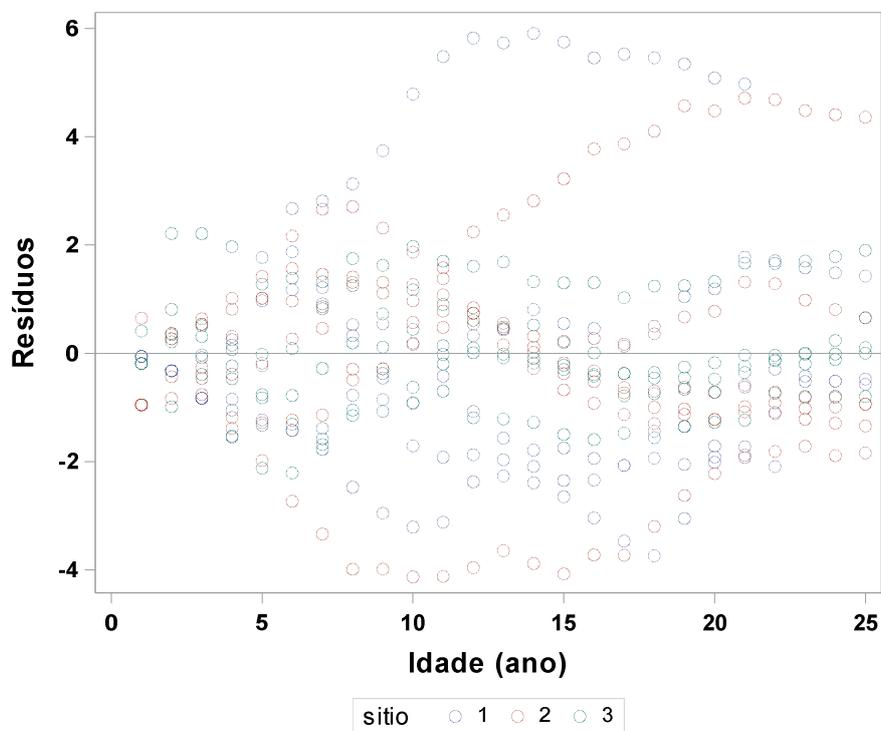
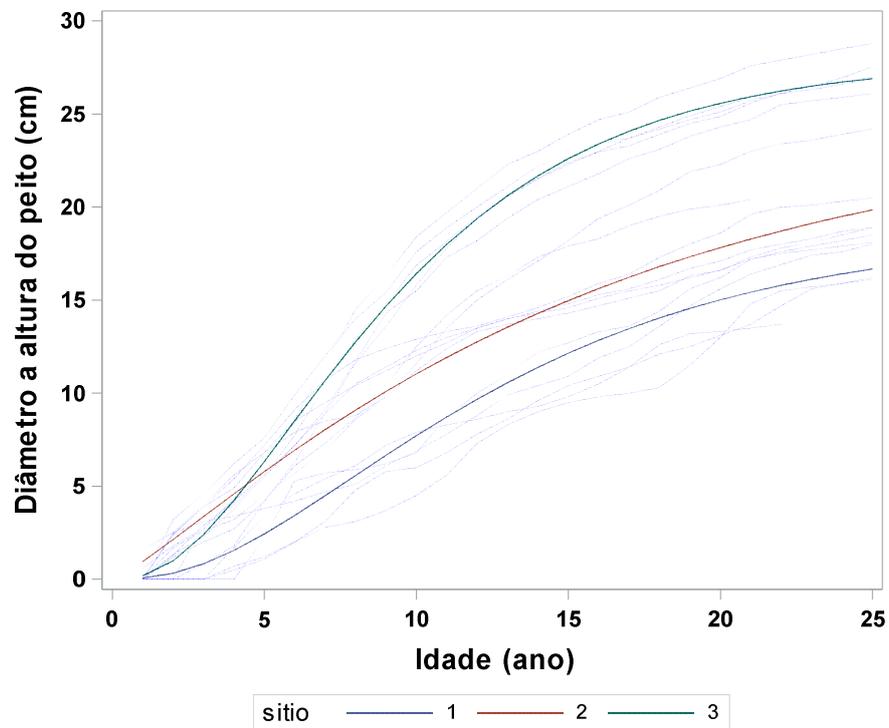
  xaxis label="Idade (ano)";
  yaxis label="Diâmetro a altura do peito (cm)";
run;

proc sort data=C_Richards_SN out= nlinoutbysn;
  by sitio;
run;

proc sgplot data=nlinoutbysn;
  scatter x=I y=resíduos / group=sitio transparency=0.1;
  refline 0 / axis=y;
  label i="idade (ano)" resíduos="Resíduos";
run;

```

Output 41. Curvas de diâmetro a 1,3 m observadas e estimadas por sítio natural (sítio) e resíduos estudentizados.



3.7.4.2.1. Inclusão da covariável sítio no modelo de Chapman-Richards

Em estudos de crescimento e produção bem como outros relacionados à Ciências Agrárias é comum ter dados estratificados seja por sítio ou outro tipo de classificação como posição social, gradiente topográfico, tratamentos ou considerando outra fonte de variação.

Neste caso, é possível comparar as curvas entre si que estão agrupadas por alguma fonte de variação denominada aqui de covariável. Essa comparação permite determinar se existe diferença significativa entre os coeficientes de regressão (assíntotas, a taxa de crescimento e/ou ponto de inflexão) das curvas ajustadas para os dados observados entre os valores da covariável (Sítio, por exemplo) sob a hipótese nula de igualdade.

Considerando o exemplo das curvas de crescimento do Output 41 e o objetivo da pesquisa em determinar se a assíntota (A) do crescimento em diâmetro das árvores (D_i) em função da idade (I) é diferente, estatisticamente, entre os sítios do estudo, o modelo de Chapman-Richards é modificado de forma a incluir variáveis indicadoras do tipo *Dummy* para discriminar os três sítios conforme descrito no Quadro 31.

Quadro 31. Modelo de Chapman-Richards com variáveis *Dummy* para testar a hipótese nula de assíntotas (A) iguais entre os sítios ($H_0: A_{\text{sítio1}}=A_{\text{sítio2}}=0$).

Covariável sítio natural	Modelo a ser ajustado para testar a Assíntota
Sítio 1	$D_i = (A + A_1d_1(1) + A_2d_2(0))(1 - e^{(-kI_i)})^c + \varepsilon_i$
Sítio 2	$D_i = (A + A_1d_1(0) + A_2d_2(1))(1 - e^{(-kI_i)})^c + \varepsilon_i$
Sítio 3	$D_i = (A + A_1d_1(0) + A_2d_2(0))(1 - e^{(-kI_i)})^c + \varepsilon_i$
	ou $D_i = A(1 - e^{(-kI_i)})^c + \varepsilon_i$

É possível determinar o efeito da covariável utilizando o princípio da soma de quadrados extra (DRAPER; SMITH, 1981) em que se compara a diferença da soma de quadrados dos resíduos a partir de um modelo simples (sem a covariável) com um modelo complexo (com covariável) mediante a seguinte expressão matemática:

$$SQres_{extra} = \frac{\frac{SQres_s - SQres_c}{P_c - P_s}}{\frac{SQres_c}{n - P_c}} \sim F_{P_c - P_s, n - P_c}$$

Em que:

$SQres_s$ = Soma de Quadrados dos Resíduos do modelo simples;

$SQres_c$ = Soma de Quadrados dos Resíduos do modelo complexo;

P_s, P_c = número de coeficientes de regressão do modelo simples e complexo, respectivamente;

n = número de observações.

Nota-se que é possível testar a hipótese nula de igualdade entre os modelos simples e complexo comparando o resultado da Soma de Quadrados dos Resíduos Extra ($SQres_{extra}$) com o valor tabelado da distribuição F desde que, a condição de resíduos Normais seja atendida.

Para exemplificar o método, a seguinte sintaxe SAS incluindo a criação das variáveis *Dummy* e o ajuste dos modelos no PROC NLIN é apresentada:

```
data uva_japão;
input sitio árvore | Di;

S1=0;
S2=0;
if sitio eq 1 then S1=1;
if sitio eq 2 then S2=1;

datalines;
1 1 25 23.4
1 1 24 23.2
1 1 23 23.1
.
.
.
3 18 1 0.0
;
title "modelo de Chapman-Richards sem covariável";
proc nlin data=uva_japao method=Marquardt;
parameters A=20
           k=0.1
```

```

c=2;
model Di = A*(1-exp(-k*I))**c;
run;

title 'modelo de Chapman-Richards com covariavel';
proc nlin data=uva_japao method=marquardt;
parameters A=25
           A1=20
           A2=15
           k=0.1
           c=2;
model d = (A+(A1*S1)+(A2*S2))*((1-exp(-k*I))**c);
run;

```

sítio=sítio natural em que as árvores foram coletadas (S1=sítio 1); arvore=número da árvore; I=idade em anos; Di=diâmetro a 1,3 m do solo.

A sintaxe SAS mostra a criação das variáveis Dummy para representar os sítios (S1=sítio 1) seguido do PROC NLIN para ajustar o modelo de Chapman-Richards único para todos os três sítios. Logo, o modelo inclui as covariáveis de sítio no coeficiente assíntota. Os resultados de forma resumida são apresentados no Output 42:

Output 42. Resultado do ajuste do modelo de Chapman-Richards sem e com covariável para sítio natural.

Modelo sem covariável para sítio:

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Model	3	83757.7	27919.2	1758.36	<.0001
Error	390	6192.4	15.8780		
Uncorrected Total	393	89950.1			

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
A	24.2030	1.6223	21.0134	27.3925
K	0.1115	0.0215	0.0691	0.1538
C	1.8559	0.3146	1.2373	2.4744

Modelo com covariável para sítio:

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Model	5	88459.1	17691.8	4604.01	<.0001
Error	388	1491.0	3.8427		
Uncorrected Total	393	89950.1			

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
A	29.9710	0.7517	28.4931	31.4490
A1	-13.3116	0.4922	-14.2793	-12.3438
A2	-9.0744	0.4220	-9.9041	-8.2447
K	0.1330	0.0107	0.1119	0.1542
c	2.0810	0.1748	1.7374	2.4247

Na tabela anova é possível observar uma redução do Quadrado Médio do Resíduo (Variância) para o modelo com covariável comparado com o modelo sem covariável. Os valores estimados da assíntota para os sítios 1 e 2 revelam valores negativos, ou seja, a partir da assíntota do sítio 3 (A=29,97 cm) é possível obter a assíntota do sítio 1 pela diferença 29,97-13,31=16,66 cm.

A aplicação da Soma de Quadrados do Resíduo extra é a seguinte:

$$SQ_{res_{extra}} = \frac{\frac{SQ_{res_s} - SQ_{res_c}}{P_c - P_s}}{\frac{SQ_{res_c}}{n - P_c}} = \frac{\frac{6192,4 - 1491,0}{5 - 3}}{\frac{1491,0}{393 - 5}} = \frac{\frac{4701,4}{2}}{\frac{1491,9}{388}} = \frac{2350,7}{3,8451} = \mathbf{611,3}$$

O valor de F tabelado com 2 e 388 graus de liberdade no numerador e denominador, respectivamente equivale a 3. Portanto, com 95% de probabilidade de confiança, rejeita-se a hipótese nula de igualdade entre as assíntotas por sítio.

Os cálculos apresentados até o momento indicam haver diferença entre pelo menos duas assíntotas de diferentes sítios.

É possível realizar uma comparação entre pares da covariável indicada, neste caso, sítio. Essa comparação (contraste) entre os coeficientes do modelo de regressão por sítio é realizada em outro procedimento SAS para ajuste de modelos não lineares, o PROC NLMIXED.

Esse procedimento foi desenvolvido para o ajuste de modelos não-lineares mistos em que a variável dependente tenha distribuição Normal, Poisson, Binomial e outras no âmbito de modelos generalizados com mais de um componente de variância.

Possui grande capacidade de ajuste de modelos por meio de várias declarações disponíveis como a declaração CONTRAST que realiza a comparação por contraste de médias. Neste caso, o PROC NLIN não aceita o uso da declaração CONTRAST dentro da sintaxe do procedimento.

Para ajustar um modelo não linear fixo (sem efeito aleatório nos coeficientes) em que se deseja considerar a distribuição Normal de y , a sintaxe básica do PROC NLMIXED é a seguinte:

```
proc nlmixed data= nome_do_dataset;  
parameters coeficiente1=valor_semente...sigma2= valor_estimado da variância;  
  
mu = expressão_do_modelo;  
model y~normal(mu, 2igma2);  
run;
```

Neste caso, a sintaxe é semelhante ao do PROC NLIN diferindo na declaração model em que é necessário indicar a distribuição da variável dependente (y) seguido da média (μ) e variância (σ^2). Para ajustar um modelo misto, basta adicionar a declaração RANDOM para especificar os efeitos aleatórios incluídos na declaração MODEL bem como outros detalhes que não será discutido aqui.

Uma breve comparação entre os procedimentos PROC NLIN e PROC NLMIXED é apresentado no Quadro 32:

Quadro 32. Comparação de recursos disponíveis nos procedimentos SAS para ajuste de modelos não-lineares.

Recurso	proc nlin	proc nlmixed
Método de ajuste iterativo	•	•
Valores sementes indicados em um grid de busca	•	•
Derivadas do modelo calculadas automaticamente	•	•
Estimação dos coeficientes de regressão	Mínimos Quadrados Não lineares	Máxima Verossimilhança
Habilitado para efeitos aleatório		•
Variância residual indicada em um coeficiente		•
Permite modelar a variável dependente (y) com distribuição além da Normal		•

Caso seja de interesse da pesquisa, além da assíntota também é possível comparar os demais coeficientes de regressão entre os sítios avaliados ou outra variável de classificação. A sintaxe do PROC NL MIXED contendo o modelo de Chapman-Richards e comparando por contraste todos os coeficientes de regressão para cada um dos três sítios é a seguinte:

```
proc sort data=uva_japão;
  by sitio árvore;
run;

proc nlmixed data=uva_japao;
  parameters A1=15 k1=0.1 c1=1
             A2=20 k2=0.2 c2=2
             A3=25 k3=0.3 c3=3
             Sigma2=15.8;

  if sitio=1 then do;
    A = A1;
    k=k1;
    c=c1;
  end;
endrun;
```

```

end;
else if sitio=2 then do;
  A = A2;
  k=k2;
  c=c2;
end;
else if sitio=3 then do;
  A = A3;
  k=k3;
  c=c3;
end;

mu = A*(1-exp(-k*t))**c;
model Di~normal(mu, Sigma2);

contrast 'Assintota A1-A2' A1-A2;
contrast 'Assintota A1-A3' A1-A3;
contrast 'Assintota A2-A3' A2-A3;

contrast 'Taxa k1-k2' k1-k2;
contrast 'Taxa k1-k3' k1-k3;
contrast 'Taxa k2-k3' k2-k3;

contrast 'Inflexão c1-c2' c1-c2;
contrast 'Inflexão c1-c3' c1-c3;
contrast 'Inflexão c2-c3' c2-c3;

predict mu out = pred;

run;

```

A criação das variáveis do tipo Dummy é realizada diretamente dentro do procedimento PROC NLMIXED.

Na declaração PARAMETERS além da especificação dos coeficientes de regressão com seus respectivos valores sementes, é necessário especificar o valor semente para a variância estimada representada no exemplo como sigma2. O valor da variância e dos

valores sementes para os coeficientes de cada sítio podem ser estimados previamente pelo PROC NLIN em um modelo por sítio utilizando a opção BY sitio.

Logo em seguida, após a declaração parameters, foi criado variáveis do tipo *Dummy* incluídas aqui para especificar cada um dos sítios e possibilitar o ajuste do modelo único contendo todos os três sítios naturais. Neste caso, cada coeficiente do modelo recebeu um número para especificar o sítio, sendo A1=assíntota do sítio 1 e assim, sucessivamente.

A variável “*mu*” representa a expressão matemática para o modelo de Chapman-Richards (Modelo 1 do quadro 30). Em seguida a declaração model especifica o modelo para estimar o diâmetro a 1,3 metros (*Di*) considerando uma distribuição Normal com média igual a “*mu*” e variância “sigma2”.

É possível ajustar modelos não-lineares generalizados no PROC NLMIXED mudando apenas a distribuição considerada para a variável dependente na declaração model.

Em seguida a declaração CONTRAST é solicitada para comparar cada um dos coeficientes de regressão do modelo entre os três sítios avaliados. Neste caso, a linha de programação “contrast 'Assintota A (A1-A2)' A1-A2;” irá comparar o valor estimado da Assíntota entre os sítios 1 e 2 e assim sucessivamente para os demais coeficientes do modelo.

Toda a informação entre aspas (aqui simples, mas pode ser duplas também) não será processada pelo SAS e serve apenas como etiqueta de texto (Label).

Finaliza a programação uma solicitação de valores estimados que sejam informados no arquivo de saída nomeado como “pred”. Os resultados da análise são apresentados no Output 43:

Output 43. Resultado da comparação de curvas pelo procedimento PROC NLMIXED.

Specifications	
Data Set	WORK.UVA_JAPAO
Dependent Variable	D
Distribution for Dependent Variable	Normal
Optimization Technique	Dual Quasi-Newton
Integration Method	None

Dimensions	
Observations Used	393
Observations Not Used	0
Total Observations	393
Parameters	10

Initial Parameters										
A1	k1	C1	A2	k2	C2	A3	k3	C3	sigma2	Negative Log Likelihood
18	0.12	2.7	25	0.07	1.2	28	0.17	2.7	15.8	947.826448

Iteration History					
Iteration	Calls	Negative Log Likelihood	Difference	Maximum Gradient	Slope
1	6	946.2745	1.55193	222.667	-2242.04
2	10	946.2313	0.04317	232.156	-157.332
3	13	946.2003	0.031062	233.075	-3.54722
4	17	943.4400	2.760267	184.936	-19.9607
5	23	923.1800	20.25997	1015.71	-1.86034
6	30	864.2389	58.94111	2077.67	-40.5856
7	37	821.1868	43.05209	1851.43	-119.432
8	40	801.7429	19.44398	1879.62	-526.931
9	43	796.8135	4.929416	653.162	-25.7847
10	46	796.1547	0.658731	304.949	-5.70793
11	48	795.4199	0.734831	106.031	-1.31240
12	51	795.3226	0.097317	111.809	-0.29079
13	55	795.0400	0.282594	102.648	-0.48308
14	57	794.7998	0.24022	92.0723	-0.36947
15	60	794.6365	0.16327	15.2020	-0.30496
16	63	794.5391	0.097427	85.3172	-0.05208
17	66	794.5248	0.014296	10.2397	-0.01857
18	68	794.5024	0.022375	77.7794	-0.00751
19	72	794.3776	0.124756	183.954	-0.04710
20	74	794.2638	0.113825	95.4444	-0.10461
21	77	794.2491	0.014708	20.1488	-0.03812
22	80	794.2463	0.002773	5.06667	-0.00610
23	83	794.2459	0.00041	2.43498	-0.00055
24	87	794.2448	0.001083	5.80932	-0.00028
25	91	794.2423	0.0025	9.86943	-0.00112
26	94	794.2417	0.000647	1.47707	-0.00128
27	97	794.2415	0.000156	5.42749	-0.00006

Iteration History					
Iteration	Calls	Negative Log Likelihood	Difference	Maximum Gradient	Slope
28	103	794.2345	0.007036	8.81670	-0.00023
29	106	794.2338	0.000702	0.43994	-0.00130
30	109	794.2338	4.852E-6	0.32105	-9.22E-6
31	112	794.2338	9.92E-8	0.007899	-2.07E-7

NOTE: GCONV convergence criterion satisfied.

Fit Statistics	
-2 Log Likelihood	1588.5
AIC (smaller is better)	1608.5
AICC (smaller is better)	1609.0
BIC (smaller is better)	1648.2

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr > t	95% Confidence Limits		Gradient
A1	18.6414	1.3476	393	13.83	<.0001	15.9920	21.2908	-0.00008
k1	0.1288	0.02370	393	5.43	<.0001	0.08219	0.1754	-0.00715
C1	2.7347	0.5428	393	5.04	<.0001	1.6677	3.8018	0.000166
A2	24.4780	2.2117	393	11.07	<.0001	20.1298	28.8263	0.000040
k2	0.07408	0.01744	393	4.25	<.0001	0.03980	0.1084	0.006550
C2	1.2307	0.1573	393	7.83	<.0001	0.9215	1.5399	-0.00047
A3	27.8825	0.6150	393	45.34	<.0001	26.6735	29.0916	0.000021
k3	0.1733	0.01539	393	11.26	<.0001	0.1430	0.2035	0.007899
C3	2.7204	0.3257	393	8.35	<.0001	2.0802	3.3607	-0.00038
sigma2	3.3335	0.2378	393	14.02	<.0001	2.8660	3.8010	-9.6E-6

Contrasts				
Label	Num DF	Den DF	F Value	Pr > F
Assintota A (A1-A2)	1	393	5.08	0.0248
Assintota A (A1-A3)	1	393	38.92	<.0001
Assintota A (A2-A3)	1	393	2.20	0.1389
Taxa cresc k (k1-k2)	1	393	3.46	0.0638
Taxa cresc k (k1-k3)	1	393	2.48	0.1158
Taxa cresc k (k2-k3)	1	393	18.20	<.0001
Inflexão C (C1-C2)	1	393	7.08	0.0081
Inflexão C (C1-C3)	1	393	0.00	0.9819
Inflexão C (C2-C3)	1	393	16.97	<.0001

O layout dos resultados do PROC NLMIXED comparado ao PROC NLIN mostra algumas tabelas adicionais com destaque para a tabela de critérios para avaliar a bondade de ajuste entre modelos por meio dos critérios de Akaike e demais estatísticas (Fit Statistics).

A tabela de contraste do modelo mostra que houve diferença significativa dos valores estimados para a assíntota entre os sítios 1 e sítio 2 bem como sítio 1 com sítio 3. O sítio 2 e sítio 3 compartilham a mesma assíntota ($Pr>F=0,1389$).

Os demais coeficientes do modelo de Chapman-Richards mostraram diferença significativa. Por exemplo, a taxa de crescimento do sítio 2 é diferente do sítio 3 ($k_2-k_3 \rightarrow Pr>F=<0,0001$) bem como o coeficiente relacionado à inflexão da curva para a comparação dos sítios 1 com 2 e 2 com 3 ($C_1-C_2 \rightarrow Pr>F=0,0081$; $C_2-C_3 \rightarrow Pr>F=<0,0001$).

3.7.4.3. Modelo para a descrição da forma do tronco de árvores

Sabe-se da literatura que, em geral, porções do perfil dos fustes de árvores poderiam ser representados por figuras geométricas. Deste modo, a base do tronco se assemelharia a um Neilóide, a porção mediana a um parabolóide e o topo por um conoide.

Esse conhecimento teórico foi a base para o surgimento dos modelos segmentados de forma que foi apresentado pela primeira vez por Max e Burkhart (1976) que dividiram o tronco em três segmentos sendo a base modelada como um sólido neiloidal; porção mediana como um sólido paraboloidal e o topo como um conoide ilustrado na Figura 48 que exemplifica o funcionamento do sistema proposto por esses autores.

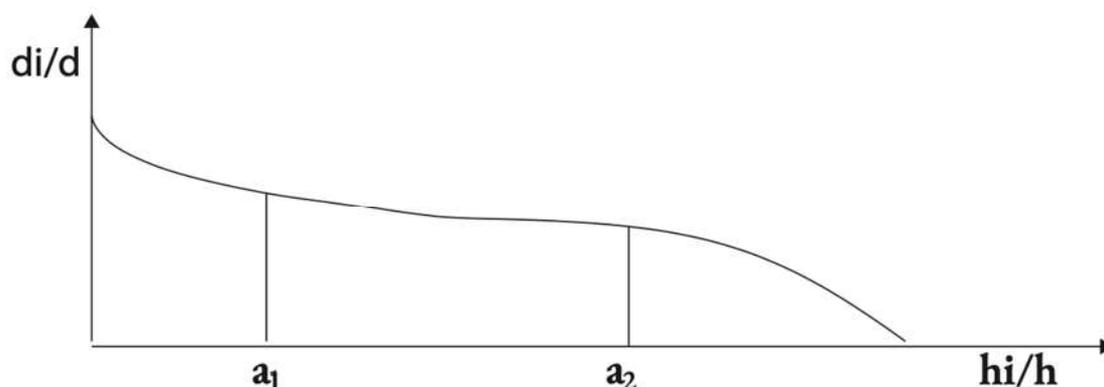


Figura 48. Segmentação do perfil do tronco de uma árvore proposto por MAX e BURKHART (1976). d_i = diâmetro comercial do tronco (cm), d =Diâmetro a 1,3 m de altura do solo, h =Altura total da árvore (m), h_i =Altura em uma determinada altura no tronco da árvore (cm).

O modelo não-linear que Max e Burkhart usaram para representar o tronco tem a seguinte expressão matemática:

$$d_i = d[\beta_1(X - 1) + \beta_2(X^2 - 1) + \beta_3(a_1 - X)^2 I_1 + \beta_4(a_2 - X)^2 I_2]^{0,5} + \varepsilon_i$$

Em que:

d_i = diâmetro comercial do tronco (cm);

d =Diâmetro a 1,3 m de altura do solo;

$x=hi/h$ quociente entre uma determinada altura e a altura total da árvore.

$\beta_1, \beta_2, \beta_3, \beta_4$ =Coeficientes de regressão a serem estimados;

a_1, a_2 =coeficientes livres que interligam os polinômios.

Os coeficientes a_1 e a_2 separam os três segmentos e teoricamente, o primeiro deveria se posicionar em um ponto entre 5 e 20% da altura total da árvore, enquanto o segundo deveria estar entre 70 e 90% da altura total. É possível limitar os valores para esses coeficientes livres dentro da sintaxe do PROC NLIN do SAS utilizando a declaração BOUND.

A sintaxe do PROC NLIN para ajustar o modelo de Max e Burkhart considerando valores sementes para os coeficientes de regressão para um povoamento jovem de *Pinus caribaea* é a seguinte.

```
proc nlin data=taper method=marquardt;
  parameters a1=.05 a2=.8 b1=-3 b2=2 b3=22 b4=-1;
  bounds .05<a1<.2;
  bounds .6<a2<.9;
  l1=0;
  l2=0;

  if x<=a1 then
    l1=1;

  if x<=a2 then
    l2=1;

  w1=b3*(a1-x)**2;
```

```
w2=b4*(a2-x)**2;
```

```
Model di=d*((b1*(x-1)+b2*(x2-1)+w1*I1+w2*I2)**0.5);
```

```
output out=stats p=pred r=res;
```

```
run;
```

3.7.4.4. Avaliação de modelos não-lineares

A melhor forma de selecionar um modelo de regressão não-linear apropriado para descrever dados de crescimento é considerar o comportamento da curva aos dados observados combinado com a interpretação biológica dos coeficientes de regressão juntamente com a avaliação da bondade de ajuste por meio de critérios estatísticos.

É importante notar que o melhor modelo será aquele que, além dos melhores valores para bondade de ajuste, também seja parcimonioso de acordo ao descrito na Figura 44.

Entre os critérios estatísticos considerados para selecionar o melhor modelo de regressão o coeficiente de determinação (R^2) é consagrado e amplamente utilizado para modelos lineares.

Entretanto, modelos não-lineares não possuem um coeficiente que representa a média populacional como no caso do intercepto (β_0) em modelos lineares. Ademais, em modelos não-lineares, a soma de quadrados total não resulta da junção entre a soma de quadrados da regressão mais a soma de quadrados dos resíduos. Esses detalhes influenciam no processo de cálculo do R^2 tornando esse critério não apropriado como indicador de bondade de ajuste para modelos não-lineares.

Em um estudo com mais de mil simulações de ajuste de modelos não-lineares, Spiess e Neumeyer (2010) reportaram que quando se utiliza o coeficiente de determinação para modelos não-lineares, alguns problemas acontecem, como:

- R^2 é consistentemente alto para modelos excelentes como para modelos ruins;
- R^2 não aumenta sempre para melhores modelos;
- Apenas em 28 a 43% das vezes, o R^2 conduz à seleção do verdadeiro melhor modelo.

Portanto, os critérios de informação de Akaike (AIC), Akaike corrigido para amostra pequenas (AICc) e variações do critério Bayesian (BIC) descritos no Quadro 30 são utilizados para a tomada de decisão de qual modelo é o mais próximo de ser o correto. O

critério BIC apresenta alta penalidade para o número de coeficientes de regressão do modelo.

Os critérios de informação são fornecidos por padrão pelo procedimento PROC NLMIXED como demonstrado na tabela Fit Statistics no Output 43.

Os seis modelos de regressão não-lineares do Quadro 30 foram ajustados para a seleção do mais adequado aos dados de crescimento em diâmetro a 1,3 cm. Os valores dos coeficientes com os limites de confiança e os critérios de informação de AIC, AICc e BIC são apresentados na Tabela 2.

Tabela 2. Valores estimados para os coeficientes de regressão dos modelos do Quadro 30 e critérios de informação para seleção do modelo mais adequado. Os critérios de informação foram obtidos pelo PROC NLMIXED.

Modelo de crescimento	Coeficiente	Coeficiente estimado	Limites de confiança 95%		Critérios de informação		
			Menor	Maior	AIC	AICc	BIC
1. Chapman-Richards	<i>A</i>	24,2030	21,0134	27,3925	2206,9	2207,0	2222,8
	<i>k</i>	0,1115	0,0691	0,1538			
	<i>c</i>	1,8559	1,2373	2,4744			
2. Logístico	<i>A</i>	20,9119	19,8185	22,0052	2220,9	2221,0	2236,8
	<i>b</i>	12,6688	8,1809	17,1567			
3. Gompertz	<i>k</i>	0,2675	0,2238	0,3111	2210,5	2210,6	2226,4
	<i>A</i>	22,2213	20,6248	23,8177			
	<i>b</i>	3,5700	2,8413	4,2987			
4. Mitscherlich	<i>k</i>	0,1686	0,1368	0,2004	2209,0	2209,1	2224,9
	<i>A</i>	29,5768	24,1503	35,0034			
	<i>b</i>	1,0881	1,0235	1,1526			
5. Richards	<i>k</i>	0,0581	0,0388	0,0773	2241,8	2242,0	2261,7
	<i>A</i>	24,9748	19,8951	30,0544			
	<i>b</i>	1,0842	0,8368	1,3316			
	<i>k</i>	0,0977	0,0300	0,1654			
6. Schnute	<i>m</i>	0,3150	-0,2877	0,9177	2208,6	2208,8	2228,5
	<i>a</i>	0,09766	0,03133	0,1640			
	<i>b</i>	0,6850	0,1043	1,2657			
	<i>y</i> ₁	0,06321	-1,0968	1,2232			
	<i>y</i> ₂	21,6101	20,4936	22,7267			

A Tabela 2 mostra os limites de confiança para cada coeficiente de regressão dos modelos. Os valores são utilizados para testar a hipótese sobre um coeficiente de regressão específico. Portanto, se os valores dos limites de confiança apresentarem valores positivos e negativos, a hipótese nula não é rejeitada.

Os critérios de informação de Akaike e Akaike corrigido são praticamente semelhantes para cada modelo sob avaliação da Tabela 2. O critério AICc tem uma melhor performance para amostras pequenas, mas à medida que a amostra aumenta o valor do critério AICc tende a se aproximar do valor de AIC, por esta razão recomenda-se que se utilize sempre AICc como padrão (BURNHAM; ANDERSON, 2002).

Outro detalhe importante de se notar é o valor do critério de BIC. Considerando a penalidade para o número de coeficientes de regressão dos modelos sob comparação, o valor de BIC conduz à seleção de modelos com menor número de coeficientes como relatado por Burnham e Anderson, (2002) que ressaltaram a tendência do BIC em selecionar modelos demasiado simples.

Este é o caso do modelo de Mitscherlich que teve o segundo menor valor para BIC. Entretanto, esse modelo não é adequado para representar o crescimento em diâmetro a 1,3 m visto que o mesmo não apresenta ponto de inflexão na expressão matemática, influenciando diretamente a estimativa da assíntota maior entre os demais modelos de regressão como demonstrado na Figura 49 (Cor rosa da curva). Ademais, a curva de Mitscherlich também mostra tendência para valores negativos para idade inicial indicando um subajuste (Underfitting, Figura 49).

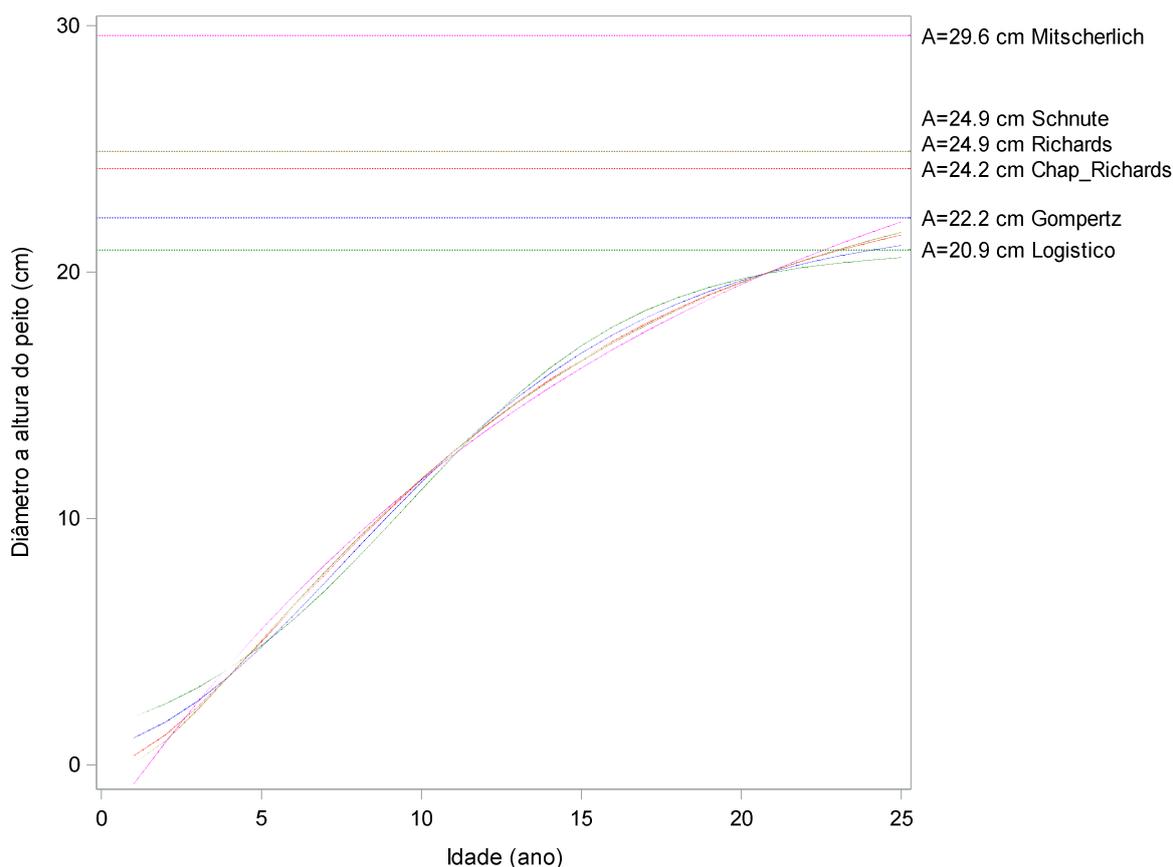


Figura 49. Desenvolvimento das curvas de crescimento estimadas para cada modelo de regressão do Quadro 30 para o mesmo conjunto de dados. A assíntota (A) para cada um dos modelos é projetada pela linha de referência descontínua.

Os modelos de crescimento apresentaram a curva na forma sigmoide, exceto a curva estimada por Mitscherlich, com pequena variação entre as curvas para os valores estimados na idade inicial e Assíntota variando de 29,6 cm para o modelo de Mitscherlich e 20,9 cm para o modelo Logístico. Entre os demais modelos, a Assíntota estimada variou de 24,9 cm para os modelos de Richards e modelo de Schnute e de 20,9 cm para o modelo Logístico.

Uma característica do modelo de crescimento Logístico é a simetria da localização para o ponto de inflexão sob a curva localizado exatamente em $A/2$ no eixo y. Esta necessidade matemática de simetria “moldou” a curva alavancando as estimativas para menor na Assíntota e maior para valores de diâmetro na idade inicial como indicado no gráfico da Figura 49 (linha verde).

Por outro lado, o modelo de Gompertz possui ponto de inflexão localizado a 1/3 da assíntota e, portanto, produz uma curva assimétrica com respeito ao ponto de inflexão. A estimativa da Assíntota foi maior do que a estimada pela Logística (linha azul).

Os modelos de Schunte e Richards apresentaram mesmo valor para a Assíntota ($A=24,9$ cm) com exatamente a mesma forma da curva que estão sobrepostas (linha cinza). Em seguida da estimativa por Chapman-Richards (linha vermelha) estima a assíntota em 24,2 cm com os melhores resultados para os critérios de informação.

Esse comportamento variado na estimativa da Assíntota e variação das estimativas na idade inicial, demonstra a importância de uma análise prévia das curvas e, portanto, não considerar apenas os critérios de bondade de ajuste para selecionar o modelo mais adequado para descrever dados de crescimento visto que, a simples comparação numérica pode indicar um modelo que não é o correto no sentido biológico.

3.8. Regressão Logística

Objetivos de aprendizagem:

- i) Mostrar a deficiência do uso de regressão ordinária para ajuste de modelo de regressão com variável dependente binária;
- ii) Demonstrar o uso de dois procedimentos SAS para ajuste de regressão logística binária para estimativa da probabilidade de ocorrência de evento;
- iii) Indicar os critérios estatísticos considerados para avaliar o ajuste de modelos logísticos;
- iv) Demonstrar a utilização de métodos de seleção automática de variáveis independentes em regressão logística.

A regressão logística é uma ferramenta analítica cada vez mais popular e importante para análise de dados em que o objetivo é prever a probabilidade de que o evento de interesse ocorra como uma função linear de uma ou mais variáveis independentes contínua e/ou categórica.

É uma solução analítica amplamente utilizada na área das Humanas em estudos de doenças como prevenção do Câncer, por exemplo. Entretanto, na ciência florestal é cada vez mais comum a obtenção de dados categóricos como variável dependente durante uma pesquisa de campo, inventário florestal ou outro levantamento de dados.

A base matemática da regressão logística (diferente de modelo de regressão logístico do Quadro 30) considera que a variável dependente seja categórica do tipo binária (dois níveis) ou ordinal (mais de dois níveis). Portanto, a regressão linear ordinária, em que a variável y é do tipo numérica, não é adequada para modelar variável dependente do tipo categórica por Mínimos Quadrados Ordinários (MQO).

Portanto, a diferença importante entre o que está sendo estimado por um modelo de regressão logística e o que é estimado por um modelo linear é que o primeiro estima a probabilidade de uma unidade em análise adquirir o evento de interesse por meio de uma função linear de uma ou mais variáveis independentes, e o segundo estima o valor da variável dependente (y) por meio de uma função linear de uma ou mais variáveis independentes.

O exemplo a seguir mostra na prática o ajuste de modelo de regressão linear por MQO para mortalidade de árvores em função do diâmetro a 1,3 m. Neste caso, se a árvore estava morta na ocasião do inventário, registrou-se valor 1, caso contrário, valor 0 para viva. Assim, os dados da variável dependente correspondem a valores binários (0 ou 1). Os dados do exemplo estão descritos no caso florestal 9.

Caso florestal 9: Calcular a regressão linear para estimar a mortalidade de árvores em uma floresta nativa.

Considere que uma investigação foi realizada com o objetivo de avaliar a mortalidade de árvores nativas de diferentes espécies em uma floresta nativa da Amazônia após 12 anos de monitoramento em parcelas permanentes. O interesse da pesquisa é determinar se o porte das árvores, representado pelo diâmetro a 1,3 m do solo (d_{2004}), registrado no início do monitoramento (2004) tem influência na mortalidade das árvores (situation).

A expressão matemática do modelo é a seguinte:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Em que:

y_i = mortalidade registrada para a i -ésima árvore (1 para árvore morta e 0 para árvore viva);

x_i = diâmetro a 1,3 m (cm) da i -ésima árvore medido no ano de 2004 (d_{2004});

β_0 = valor da mortalidade quando o diâmetro 1,3 m for zero (Intercepto);

β_1 = mudança de mortalidade com o aumento de um centímetro em diâmetro das árvores (Coeficiente angular);

ε_i = efeito dos resíduos \sim NIID(0, σ^2).

Para ajustar o modelo de mortalidade, primeiramente criou-se a variável dependente “situation” dentro do SAS utilizando o comando IF e THEN. Neste caso, solicitou-se ao SAS comparar os valores de diâmetro a 1,3 m do solo do ano de 2016 com o diâmetro registrado em 2004. Se menor, o SAS atribuiu valor 1 para a variável “situation” e zero do contrário. Portanto, árvores com valor 1 para variável “situation” foi considerada como morta e zero do contrário.

A sintaxe SAS utilizada foi a seguinte:

```
data mortalidade;
  input num GrupoEc d_2004 d_2016;

  if d_2016 lt d_2004 then situation =1;
  else situation=0;

  datalines;
1 Pioneira 26.4 29.3
2 ni 44.2 0
3 SecTar 66.9 70.1
.
.
.
5286 ni 101.8 0
;

proc print data=mortalidade (firstobs=501 obs=508) noobs;
  run;

Title "OLS wrong model for binary data";
proc reg data=mortalidade;
  model situation=d_2004;
  output out=saída1 predicted=situation_est;
  run;
  title;
```

```

proc sort data=saída1 out=sorted;
  by d_2004;
run;

ods graphics / reset width=12cm height=12cm imagemap noborder;
proc sgplot data=sorted;
  scatter x=d_2004 y=situation / transparency=0.7;
  series x=d_2004 y=situation_est / lineattrs=(color=red thickness=2 pattern=solid)
transparency=0.1;

  xaxis integer values=(0 to 200 by 25) label="Diâmetro a altura do peito (cm)"
  LABELATTRS=(Color=black Family=Arial Size=13 Weight=Bold)
  VALUEATTRS=(Color=black Family=Arial Weight=Bold);

  yaxis integer values=(-0.5 to 1.5 by 0.25) label="Situação"
  LABELATTRS=(Color=black Family=Arial Size=13 Weight=Bold)
  VALUEATTRS=(Color=black Family=Arial Weight=Bold);

  label situation="Vivo=0; Morto=1"
  situation_est="Estimativa";

  keylegend / location=inside position=bottomleft across=1;

run;

```

num=número da árvore; grupo=grupo ecológico (pioner=pioneira; ni=não definido; shatol=tolerante a sombra); d2004=diâmetro a 1,3 m (cm) medido no ano de 2004; d2016=diâmetro a 1,3 m (cm) medido no ano de 2016 na mesma árvore.

Solicitou-se a impressão de algumas observações selecionadas ao acaso (501 a 508) para fins de verificação dos dados pelo PROC PRINT e logo ajustou-se o modelo de regressão no PROC REG. Por último, solicitou-se um arquivo de saída (saída_1) para registrar os valores estimados para a situação a fim de construir um gráfico de valores estimados e observados em função do diâmetro a 1,3 m com o PROC SGPLOT. O resultado da análise é mostrado no Output 44.

Output 44. Resultado da análise para a mortalidade de árvores por mínimos quadrados ordinários (MQO) em que a variável dependente é dicotômica (0=Vivo; 1=Morto).

The PRINT Procedure

Num	GrupoEc	d_2004	d_2016	Situation
501	SecTar	31.6	0.0	1
502	Pioneira	25.4	37.7	0
503	ni	54.9	0.0	1
504	SecTar	22.5	30.8	0
505	SecTar	29.5	30.4	0
506	SecTar	28.7	30.7	0
507	ni	20.3	0.0	1
508	SecTar	45.6	0.0	1

The REG Procedure

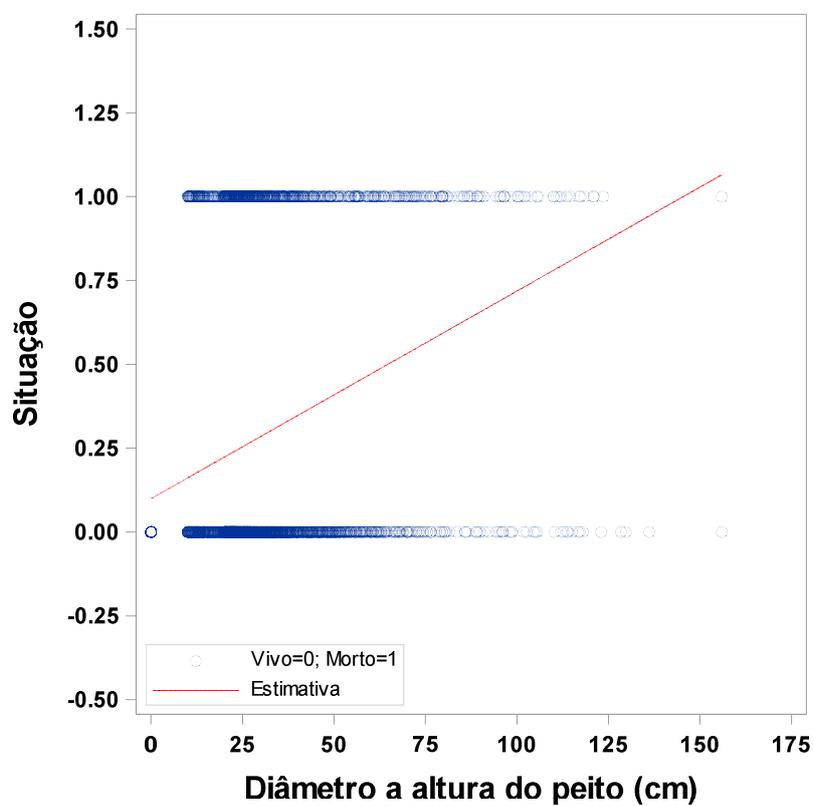
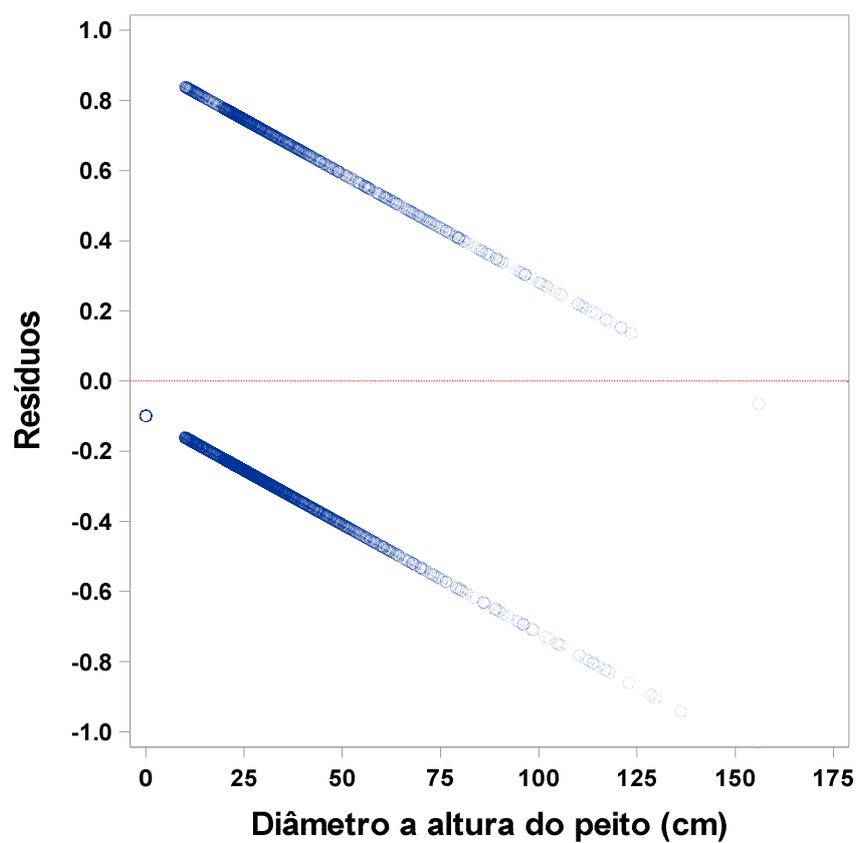
Number of Observations Read	5287
Number of Observations Used	5286
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	86.44139	86.44139	502.80	<.0001
Error	5284	908.42297	0.17192		
Corrected Total	5285	994.86436			

Root MSE	0.41463	R-Square	0.0869
Dependent Mean	0.25142	Adj R-Sq	0.0867
Coeff Var	164.91677		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	0.09948	0.00886	11.23	<.0001
d_2004	d_2004	1	0.00619	0.00027600	22.42	<.0001

The SGPLOT Procedure



A partir da impressão dos dados com o PROC PRINT é possível verificar que as árvores 501, 503, 507 e 508 morreram e receberam valor 1 para a variável dependente “situation”.

A tabela ANOVA mostra que o modelo é significativo mesmo com valor para o coeficiente de determinação próximo de zero. A tabela de valores estimados para os coeficientes de regressão mostra que a variável independente d_2004 é altamente significativa ($Pr>|t|<0.0001$) e os dados da situação de mortalidade estimada são obtidos a partir da equação:

$$\hat{y}_i = 0,09948 + 0,00619x_i$$

Como observado, foi possível ajustar um modelo de regressão por mínimos quadrados ordinários mesmo para uma variável dependente binária. Entretanto, ao analisar o gráfico de resíduos observa-se que o ajuste por MQO produz resíduos que viola algumas condicionantes de regressão como descrito no Quadro 33:

Quadro 33. Violação das condicionantes de regressão quando utilizado ajuste por MQO para variável dependente dicotômica.

Motivo	Situação	Violação
Para qualquer valor de x_i existe somente duas possibilidades de valores para os resíduos	Se $y_i=1$, então: $\varepsilon_i=1-(0,09948+0,00619x_i)$ Se $y_i=0$, então: $\varepsilon_i=-(0,09948+0,00619x_i)$	Condicionante de normalidade dos resíduos.
A variância de ε_i é diferente para diferentes observações e varia em função de x_i	$\text{Var}(\varepsilon_i)=(\beta_0 + \beta_1x_i)(1-\beta_0 - \beta_1x_i)$	Condicionante de Homocedasticidade dos resíduos

A análise do gráfico de valores estimados também mostra limitação do uso de regressão por MQO, visto que produz estimativas de situação de mortalidade entre 0 e 1 sem interpretação. Uma árvore de 50 cm de diâmetro possui uma estimativa de situação de 0,41. Por outro lado, para árvores com diâmetro a partir de 170 cm o valor estimado de situação é maior do que 1.

Desta forma, faz-se necessário o uso correto de uma fundamentação matemática que considere a modelagem de dados categóricos na variável dependente com estimativas de probabilidade entre 0 e 1.

3.8.1. Ajuste da regressão logística no SAS System

O método de ajuste de uma regressão logística ocorre pela maximização da função de Verossimilhança. Desta forma, considerando a distribuição Binomial, a probabilidade de se ter um determinado valor para a variável dependente y dado os valores de tamanho da amostra (n) e a probabilidade de evento específico (π), temos:

$$P(y|n, \pi) = \binom{n}{y} \pi^y (1 - \pi)^{(n-y)}$$

A probabilidade de se observar um valor y da amostra caso o verdadeiro valor do parâmetro seja o valor testado, ou seja, a verossimilhança para uma observação y correspondente para a distribuição Binomial é obtida por:

$$L(\pi|n, y) = \binom{n}{y} \pi^y (1 - \pi)^{(n-y)}$$

A maximização do valor da função de verossimilhança (a fim de tornar a probabilidade de a amostra observada ocorrer seja a maior possível) é realizada considerando o logaritmo da função de verossimilhança, por fins práticos e facilidade de cálculo.

O procedimento PROC LOGISTIC considera a maximização da função de verossimilhança para ajustar modelos de regressão logística. Entretanto, outros procedimentos SAS também podem ser utilizados para o ajuste de regressão logística com resultados semelhantes para a estimativa dos coeficientes de regressão, a saber:

- PROC GENMOD: ajusta modelos lineares generalizados (GLiMs) por meio da especificação da distribuição dos dados na declaração DIST= juntamente com a função de ligação específica na declaração LINK=. Ademais, possibilita análise Bayesiana utilizando a declaração BAYES.
- PROC GLIMMIX: ajusta GLiMs com possibilidade de inclusão de efeito aleatório no modelo (modelo misto);

- PROC CATMOD: procedimento dedicado a análise de dados categóricos. Ajusta modelos considerando maximização da função de verossimilhança ou mínimos quadrados ponderados. Também utilizado para análise de tabela de contingência;
- PROC SURVEYLOGISTIC: ajusta modelo de regressão logística considerando especificidades da amostra como estratificação, ponderação, cluster e outros;
- PROC HPLOGISTIC: procedimento de alta performance (HP=High Performance) concebidos para big data podendo ser utilizado para amostras pequenas. O grande diferencial desse procedimento é a possibilidade de divisão dos dados observados da amostra em subconjuntos de dados (Treino=para ajuste do modelo, Validação=para a seleção do modelo e Teste=para validar o modelo selecionado) durante o processo de ajuste do modelo de regressão logística com resultados gráficos de análise. O PROC LOGISTIC utiliza todos os dados observados para o ajuste do modelo com gráficos dedicados para a amostra inteira.

Como visto, cada procedimento possui recursos especiais que os tornam úteis para determinadas aplicações e especificidades dos dados da amostra. Por exemplo, caso a amostragem dos dados seja realizada por estratificada (ou em Cluster) e cada estrato possua diferente número de unidades de amostra/unidade experimental (algumas unidades super amostradas) e o processo de ajuste do modelo de regressão logística desconsidere essa situação, as estimativas dos coeficientes de regressão não serão afetadas, mas o erro padrão pode ser subestimado impactando no teste de hipótese para a variável independente.

Neste caso, o uso do procedimento PROC SURVEYLOGISTIC é o mais adequado visto que possui a capacidade de ajustar o modelo de regressão logística considerando o peso da amostragem em cada estrato.

O procedimento PROC LOGISTIC é o mais específico e dedicado para análise de regressão logística em que os dados da amostra não possuam grupos com superamostragem ou subamostragem. Esse procedimento possui a opção de seleção de variáveis candidatas por meio de algoritmos como: Forward, Backward e Stepwise bem como métodos de seleção de efeito. Ademais, possui uma variedade de análises pós-ajuste seja para modelar variável dependente binária ou do tipo ordinal.

A sintaxe padrão do PROC LOGISTIC para ajustar um modelo de regressão logística é bem simples, entretanto, existem inúmeras opções (options) e recursos que faz com que esse procedimento seja o mais popular para realizar ajuste de modelos binários:

```
proc logistic data= nome_do_dataset options;  
  class variável_de_classe1 variável_de_classe2...;  
  model variável_dependente = variável_independente / options;  
run;
```

A declaração CLASS é utilizada quando se deseja analisar variáveis independentes do tipo categórica no modelo de regressão. Quando utilizado, o procedimento cria automaticamente variáveis indicadoras (Dummy) para representar cada nível da variável independente declarada no modelo. Portanto, se uma variável independente do tipo categórica possuir dois ou mais níveis representados por caracteres essa deve ser declarada em CLASS. A variável posição social com os níveis “Dominante”, “Codominante”, “Suprimida” deve ser declarada como categórica.

Por outro lado, se uma variável independente possuir dois níveis representados por valores indicadores do tipo Dummy (0 ou 1), não é necessário declará-la em CLASS para considerá-la no ajuste do modelo.

Na declaração MODEL deve-se inserir o modelo de regressão logística da mesma forma que no procedimento PROC REG.

3.8.1.1. Aplicação para variável dependente binária

Neste caso, a variável dependente (y) somente pode assumir um valor de duas alternativas (dicotômico) em que, usualmente, recebe um valor zero (0) para quando o evento não ocorrer e código um (1) para quando o evento ocorrer.

Alguns exemplos de variável dependente dicotômica na Ciência Florestal são indicados no Quadro 34:

Quadro 34. Definição da variável dependente dicotômica de acordo ao objetivo da pesquisa.

Objetivo	Variável dependente (y)	Variável independente (x's)
Avaliar o efeito da mortalidade de insetos de acordo à dosagem de um determinado inseticida orgânico.	Mortalidade: 1=morre; 0=não morre	Doses do inseticida orgânico (por exemplo: D1; D2; D3; Controle).
Estudar o sucesso da brotação de gemas de acordo ao tipo de meio de cultura utilizado.	Brotação: 1=brota; 0=não brota	Diferentes tipos de meio de culturas (por exemplo: MC1; MC2; MC3; MC4).
Avaliar a sobrevivência/mortalidade de árvores em função do diâmetro a 1,3 m do solo e o nível de competição.	Mortalidade: 1=morre; 0=não morre	Diâmetro a 1,3 m do solo e o nível de competição calculado pelo índice de Hegyi, por exemplo.
Determinar a ocorrência de incêndio para em função do valor de umidade relativa do ar em uma determinada área ou região.	Incêndio: 1=ocorre; 0=não ocorre	Umidade relativa do ar.

Neste caso, a probabilidade de um dos eventos do Quadro 34 ocorrer é representado por π_i e a não-ocorrência em uma probabilidade $1 - \pi_i$. Ao calcular a proporção $\frac{\pi_i}{1-\pi_i}$ e transformar com Logaritmo obtém-se uma nova variável dependente (η_i). Essa transformação é usualmente referida como Logit ou Log-odds e utilizada na regressão logística para estimar os valores dos coeficientes de regressão por meio do seguinte modelo em função das variáveis independentes (x's):

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}$$

Em que:

$\eta_i = \text{Logit}(\pi_i) = \text{Log}\left(\frac{\pi}{1-\pi}\right)$ = Variável dependente;

x_i = Variável independente;

$\beta_0, \beta_1, \beta_k$ = coeficientes de regressão a serem estimados.

Neste caso, logaritmo natural ou logaritmo base 10 podem ser utilizados. A diferença entre os dois impactará no valor do coeficiente intercepto (β_0). É importante notar que a função Logit transforma os dados da variável dependente resultando em valores positivos e negativos. Entretanto, os valores de probabilidade permanecem fixos entre zero e 1 conforme demonstra o Quadro 35:

Quadro 35. Correspondência entre os valores de probabilidade (π_i) e Logit (η_i).

π_i	$1 - \pi_i$	$Logit(\pi_i) = \eta_i$
0,01	0,99	-4,60
0,02	0,98	-3,89
0,05	0,95	-2,94
0,10	0,90	-2,20
0,50	0,50	0
0,95	0,05	2,94
0,99	0,01	4,60

Após o ajuste do modelo de regressão logística é necessário realizar a transformação inversa utilizando exponencial (neste caso, $\exp = e^x$) por meio da seguinte expressão:

$$\pi_i = \left[\frac{\exp(\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki})} \right]$$

Esse procedimento é essencial para realizar interpretação dos resultados em escala de probabilidade, ou seja, valores variando de 0 a 1.

Para um modelo simples com valor para o intercepto $\beta_0=0$ e valor para $\beta_1=1$, a probabilidade (π_i) é representada pela curva estimada com forma sigmoideal de acordo à Figura 50. Observa-se que independente do valor de x_i , os valores para a probabilidade estimada estará sempre entre 0 e 1.

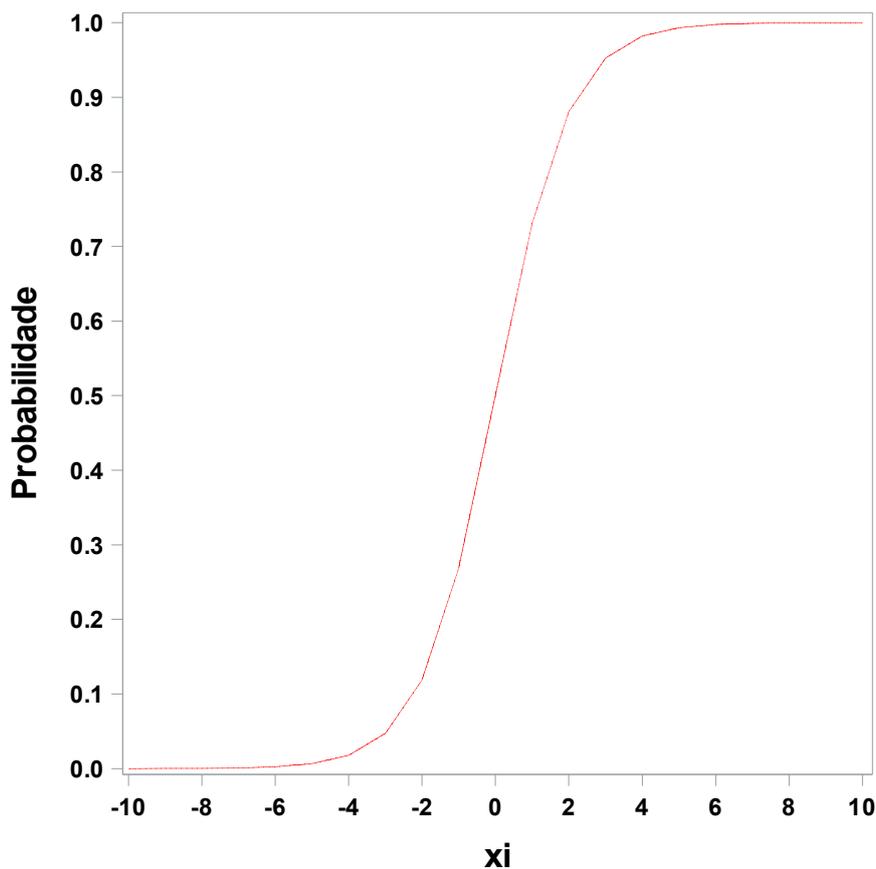


Figura 50. Comportamento da probabilidade em função da variável independente xi.

Ademais, é possível calcular a razão de chances (Odds ratio) para cada variável independente considerando a exponenciação do coeficiente de regressão (Betas) associado à variável pela seguinte fórmula:

$$Odds\ ratio = e^{\hat{\beta}} = 2,71828^{\hat{\beta}}$$

Para fins de aplicação, será considerado o caso florestal 9, mas com o ajuste de um modelo de regressão logística com a seguinte expressão matemática:

$$\eta_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

A sintaxe do procedimento PROC LOGISTIC é solicitada no SAS da seguinte maneira:

```

proc logistic data=mortalidade;
  model situation (event="1")=d2001;
  effectplot fit(x=d_2004);
run;

```

Por padrão, o PROC LOGISTIC modela o menor valor para a variável dependente. Neste sentido, se os dados estão organizados em esquema zero/um para a variável dependente, os resultados para o ajuste do modelo será a probabilidade estimada de o evento zero (0) ocorrer (menor valor para y). Portanto, se o interesse da pesquisa é avaliar o efeito do diâmetro a 1,3 m do solo na mortalidade das árvores (morrer=1), o resultado do processamento será um modelo para estimar o oposto ao objetivo da pesquisa (probabilidade de a árvore não morrer em função do diâmetro).

Para informar ao SAS que o modelo deve ser ajustado para estimar a probabilidade de o evento ocorrer basta adicionar a opção (EVENT="1") logo após a variável dependente na declaração MODEL.

A opção EFFECTPLOT FIT(x=) solicita ao SAS o gráfico de probabilidade estimada em função da variável independente diâmetro. Esse gráfico é útil para avaliar o comportamento da probabilidade. Os resultados do ajuste do modelo logístico são apresentados no Output 45:

Output 45. Resultado da análise para a mortalidade de árvores por maximização da função de verossimilhança pelo PROC LOGISTIC em que a variável dependente é dicotômica (0=Vivo; 1=Morto).

Model Information	
Data Set	WORK.MORTALIDADE
Response Variable	Situation
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	4069
Number of Observations Used	4068

Response Profile		
Ordered Value	Situation	Total Frequency
1	0	2834
2	1	1234

Probability modeled is Situation=1.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

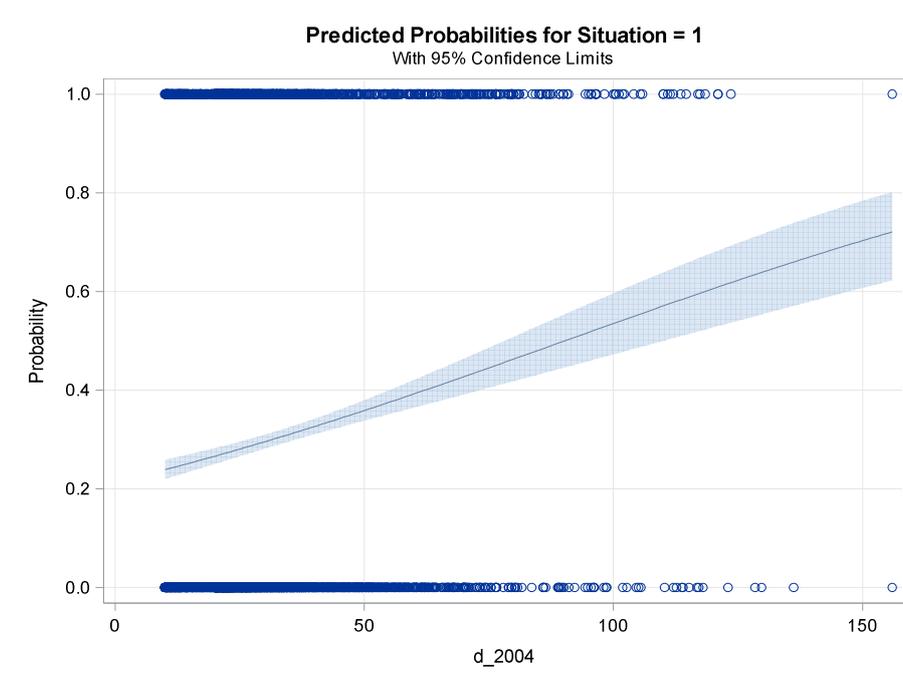
Model Fit Statistics		
Criterion	InterceptbOnly	Intercept and Covariates
AIC	4994.822	4935.393
SC	5001.133	4948.014
-2 Log L	4992.822	4931.393

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	61.4299	1	<.0001
Score	64.3292	1	<.0001
Wald	61.4201	1	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.3027	0.0702	344.4610	<.0001
d_2004	1	0.0144	0.00184	61.4201	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
d_2004	1.015	1.011	1.018

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	55.6	Somers' D	0.114
Percent Discordant	44.2	Gamma	0.115
Percent Tied	0.2	Tau-a	0.048
Pairs	3497156	c	0.557



A Tabela “Model Information” mostra que a variável dependente “situation” é binária (0 ou 1) e indica que o método numérico para maximizar a função de verossimilhança foi o de Fisher. Em seguida a Tabela “Response Profile” mostra que o número de observações para árvores vivas é bem maior do que a frequência de árvores mortas na amostra utilizada. Logo a seguir mostra a mensagem que a probabilidade a ser estimada pelo modelo será para o evento 1 conforme solicitado na opção EVENT=“1”.

A tabela “Model Fit Statistics” possui os valores para os critérios de informação de Akaike (AIC), Schwarz (SC) e o valor maximizado para o logaritmo da verossimilhança (-2LogL) para o modelo somente com o intercepto e o modelo contendo a variável independente diâmetro. Os valores para os critérios de Akaike e Schwarz são calculados a partir das fórmulas do Quadro 30 e são penalizados de acordo ao número de coeficientes de regressão sendo o mais rigoroso o de Schwarz para comparação entre dois modelos com diferentes números de coeficientes de regressão ajustados para o mesmo conjunto de

dados sendo o menor valor, o melhor. Neste caso, não existe um valor de referência para determinar o quão bom é um modelo como o R^2 .

A tabela “Testing Global Null Hypothesis: BETA=0” é utilizada para testar a hipótese nula de que o coeficiente de regressão associado à variável independente é igual a zero. Essa tabela é exclusiva para avaliar o efeito das variáveis independentes no modelo. Todos os três testes são significativos e, portanto, rejeita-se H_0 e conclui-se que ao menos um dos coeficientes associado a variável independente é significativo no modelo. Neste caso, só existe um!

O valor do χ^2 (Chi-Square) para o teste Likelihood ratio é obtido pela diferença entre o valor -2LogL do modelo somente com intercepto e o modelo contendo a variável independente. Portanto, basta realizar a diferença desses valores a partir da tabela “Model Fit Statistics”.

A próxima tabela “Analysis of Maximum Likelihood Estimates” contempla os valores estimados para cada um dos coeficientes de regressão do modelo juntamente com seus valores-p para testar a hipótese nula de que cada coeficiente é igual a zero. Neste caso, todos os dois coeficientes são estatisticamente significativos.

Outra tabela importante é a “Odds Ratio Estimates” que apresenta o valor exponencial do coeficiente associado à variável independente. Neste caso, $\exp(0,0283)=1,029$.

Assim, a Logit da probabilidade da equação de regressão logística é:

$$\text{Logit}(\pi_i) = -1,3027 + 0,0144D_i$$

Portanto, a probabilidade estimada de uma árvore morrer para um valor qualquer (dentro dos valores da amostra) do diâmetro a 1,3 metros do solo é calculada por:

$$\hat{\pi}_i = \frac{e^{-1,3027+0,0144D_i}}{1 + e^{-1,3027+0,0144D_i}}$$

Por outro lado, a probabilidade de uma árvore não morrer é calculada por:

$$\hat{\pi}_i = \frac{1}{1 + e^{-1,3027+0,0144D_i}}$$

Como observado na tabela “Analysis of Maximum Likelihood Estimates” o diâmetro a 1,3 m do solo tem influência significativa na mortalidade das árvores, mas em que intensidade ocorre essa influência à medida que aumenta o diâmetro das árvores? Para

responder essa pergunta vamos considerar o valor da exponencial do coeficiente relacionado ao diâmetro apresentado na tabela “Odds Ratio Estimates” e adicionar uma simples operação matemática como visto a seguir:

$$\{[e^{(0,0144)} \cdot 100] - 100\} = 1,44\%$$

A partir desse valor obtém-se a seguinte interpretação: Quando o diâmetro a 1,3 m do solo é aumentado em uma unidade (1 cm), a probabilidade de uma árvore morrer aumenta em 1,44%.

A curva de probabilidade estimada para a mortalidade das árvores em função do diâmetro é apresentada no gráfico do Output 45. Observa-se que à medida que aumenta o porte das árvores a probabilidade de uma determinada árvore morrer aumenta. Portanto, para duas árvores com um diâmetro a 1,3 m do solo de 20 cm e outra de 100 cm, a probabilidade de mortalidade estimada é dada por:

$$\hat{\pi}_i = \frac{e^{-1,3027+0,0144 \cdot 20}}{1 + e^{-1,3027+0,0144 \cdot 20}} = 0,2661$$
$$\hat{\pi}_i = \frac{e^{-1,3027+0,0144 \cdot 100}}{1 + e^{-1,3027+0,0144 \cdot 100}} = 0,5343$$

Uma análise mais detalhada da mortalidade de árvores pode ser realizada com a inclusão da variável categórica grupo ecológico no modelo de regressão logística. Essa opção é possível pela utilização da declaração **class** considerando a seguinte sintaxe:

```
proc logistic data=mortalidade;
  class GrupoEc (ref="Pioneira") / param=ref
  model situation (event="1")=d2001;
  effectplot slicefit(x=d_2004 sliceby=GrupoEc);
run;
```

A declaração CLASS inclui a opção REF="Pioneira" para que o grupo de árvores Pioneiras seja a referência de comparação com os demais níveis do grupo ecológico. Caso essa opção não for utilizada, o proc logistic considera o nível de comparação pela ordem alfabética que neste caso seria o nível Sem_Id. A opção PARAM=REF posiciona as variáveis das colunas da matriz design, linearmente dependentes, como valores de

referência sendo uma para cada nível da variável exceto para Pioneira. Esse passo é importante pois gera resultados de coeficientes estimados considerando um grupo de referência evitando over-parameterization.

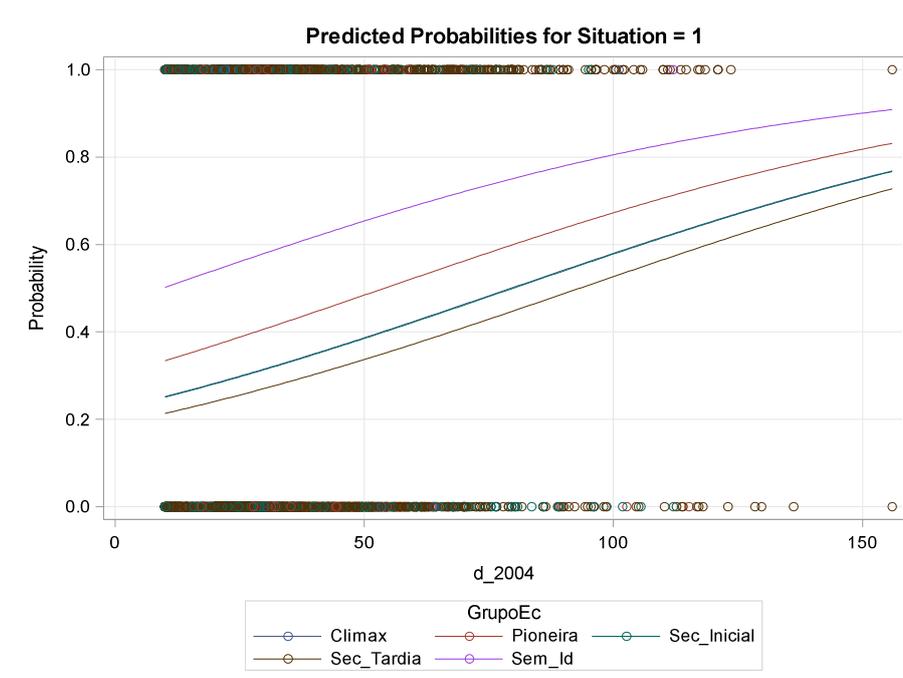
O gráfico de probabilidade de mortalidade estimada com curvas separadas para cada grupo ecológico é solicitado pela declaração EFFECTPLOT SLICEFIT (x=) SLICEBY=. Os resultados do ajuste estão no Output 46:

Output 46. Ajuste do modelo de regressão logística com variáveis independente contínua e categórica para o caso florestal 9. O restante dos resultados foi omitido.

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
d_2004	1	69.9882	<.0001
GrupoEc	4	37.1768	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.8478	0.1263	45.0520	<.0001
d_2004		1	0.0157	0.00187	69.9882	<.0001
GrupoEc	Climax	1	-0.3981	0.5369	0.5498	0.4584
GrupoEc	Sec_Inicial	1	-0.4049	0.1324	9.3578	0.0022
GrupoEc	Sec_Tardia	1	-0.6142	0.1227	25.0754	<.0001
GrupoEc	Sem_Id	1	0.6995	0.4023	3.0235	0.0821

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
d_2004	1.016	1.012	1.020
GrupoEc Climax vs Pioneira	0.672	0.234	1.924
GrupoEc Sec_Inicial vs Pioneira	0.667	0.515	0.865
GrupoEc Sec_Tardia vs Pioneira	0.541	0.425	0.688
GrupoEc Sem_Id vs Pioneira	2.013	0.915	4.428



Os resultados mostram que o grupo ecológico das árvores influencia na mortalidade ($Pr > \text{Chisq} = <0,0001$). A tabela “Analysis of Maximum Likelihood Estimates” mostra a comparação entre os níveis da variável grupo ecológico considerando o nível de referência (Pioneira). Comparado ao grupo de árvores Pioneiras, as árvores classificadas como Secundária Inicial (Sec_Inicial) e Secundária Tardia (Sec_Tardia) apresentaram efeito significativo na probabilidade de mortalidade. Neste caso, o valor positivo para ambos os coeficientes de regressão (Sec_Inicial=-0,4049 e Sec_Tardia=-0,6142) indica que as árvores desses grupo são menos propensas a morrer do que árvores Pioneiras.

Esse comportamento é visualizado no gráfico de curvas de probabilidade estimada para a mortalidade entre os grupos ecológicos do Output 46. Esse gráfico também mostra os valores observados da variável dependente “situation” (círculos).

Outra opção para essa análise é avaliar os valores estimados para a razão de chances de cada par de comparação da tabela “Odds Ratio Estimates”.

É possível personalizar a comparação entre os grupos ecológicos por meio da declaração CONTRAST conforme sintaxe exemplo a seguir que realiza a comparação de árvores Climax com árvores Secundária inicial.

```
proc logistic data=mortalidade;
  class GrupoEc (ref="Pioneira") / param=ref
  model situation (event="1")=d2001;
  contrast "Climax vs. SecTardia" GrupoEc 1 0 0 -1;
  run;
```

3.8.2. Medidas de ajuste da regressão logística

Para avaliar a qualidade do modelo de regressão logística Allisson (2012) sugere dois grupos de medidas de ajuste:

- 1) Critérios que informam o quão bom a variável dependente é estimada a partir das variáveis independentes:
 - a. Coeficiente de determinação generalizado;
 - b. Área abaixo da curva ROC.
- 2) Critérios para avaliar a bondade de ajuste:
 - a. Deviance;
 - b. Pearson Qui-quadrado;
 - c. Teste de Hosmer-Lemeshow.

O coeficiente de determinação generalizado (R^2 generalizado) é calculado pelo PROC LOGISTIC com a opção RSQ na declaração MODEL. Entretanto, esse coeficiente tem um limite máximo bem abaixo do valor 1 diferindo, portanto, do coeficiente de determinação para regressão ordinária. Para corrigir esse detalhe o PROC LOGISTIC imprime nos resultados outro valor de R^2 generalizado denominado "Max-rescaled Rsquare" que possui valor máximo de 1.

Para os critérios de bondade de ajuste, os valores da Deviance e Pearson Qui-quadrado são solicitados no PROC LOGISTIC com a opção AGGREGATE SCALE=NONE. Valor-p não significativo de Deviance e Pearson indicam que o modelo ajusta bem aos dados. Isso devido a que o critério de Deviance é calculado pela diferença do Log da Verossimilhança entre o modelo saturado (Ajuste perfeito aos dados observados) e o modelo sob análise, sendo o resultado da diferença positiva e multiplicado por dois.

Portanto, quanto mais próximo o modelo sob análise estiver do modelo ideal (saturado), menor será a diferença entre ambos para a Log da Verossimilhança implicando em não significância para $Pr > ChiSq$.

Entretanto, os critérios Deviance e Pearson somente devem ser considerados na avaliação de um modelo quando este possuir poucas variáveis independentes e do tipo categórica. Isso se deve a que durante o processo de cálculo dos critérios, considera-se uma frequência mínima de observações dentro de grupos (profiles) criado pelo teste. Caso um modelo possuir muitas variáveis e/ou algumas do tipo contínua, a frequência das observações dentro dos grupos será pequena ao ponto de estes testes não serem úteis para avaliar o modelo devido à não aderência à distribuição Qui-quadrado (ALLISSON, 2012).

Neste caso, deve-se considerar outra estatística utilizada para avaliar a qualidade do modelo denominado falta de ajuste (Lack of fit). Para solicitar esse teste, basta utilizar a opção LACKFIT na declaração model do PROC LOGISTIC. Esse teste também considera um valor-p não significativo como indicativo de modelo adequado aos dados.

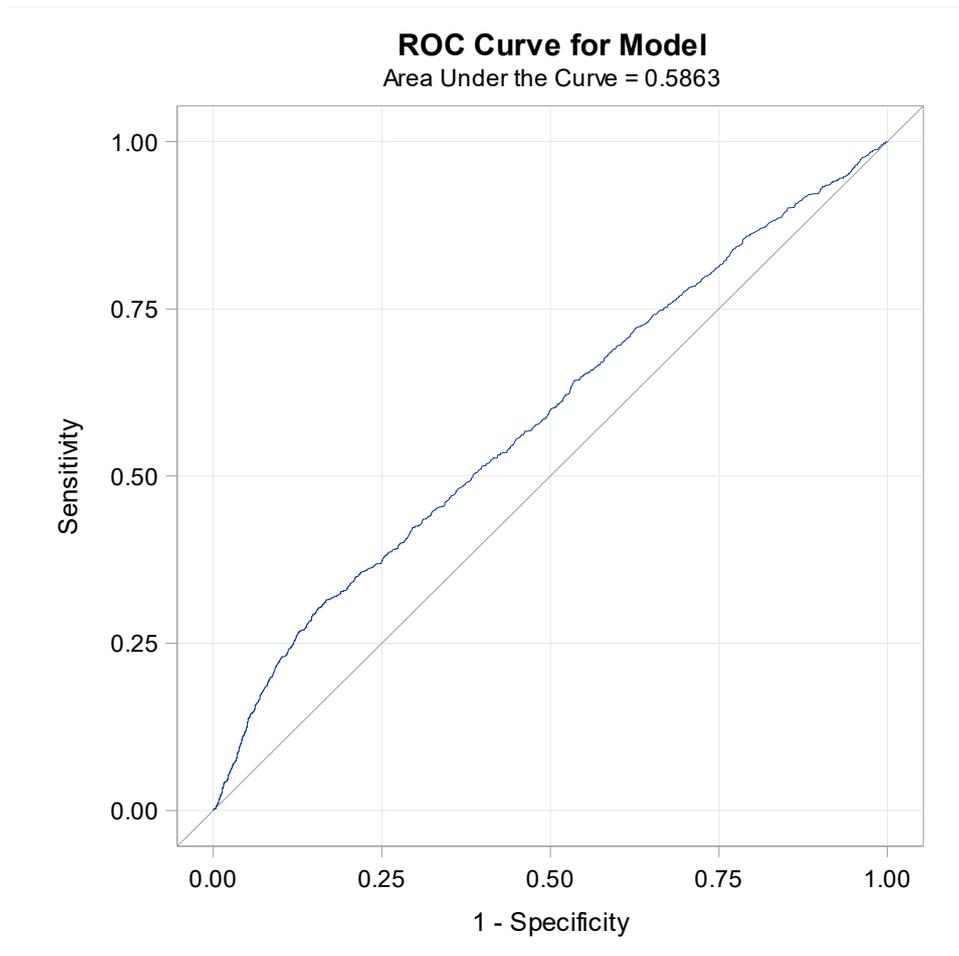
A seguir a sintaxe do PROC LOGISTIC contendo as opções para calcular as medidas de ajuste para o modelo de regressão logística do caso florestal 9. Os resultados são apresentados no Output 47:

```
/*1) Coeficiente de determinação e curva roc*/  
proc logistic data=mortalidade;  
  class GrupoEc (ref="Pioneira") / param=ref  
  model situation (event="1")=d2001 / aggregate scale=none;  
  run;  
  
/*2) Deviance, Pearson e Lack off it*/  
proc logistic data=mortalidade;  
  class GrupoEc (ref="Pioneira") / param=ref  
  model situation (event="1")=d2001 / aggregate scale=none lackfit;  
  run;
```

Output 47. Medidas de ajuste para avaliar o modelo de regressão logística do caso florestal 9.

Coefficiente de determinação e curva roc.

R-Square	0.0238	Max-rescaled R-Square	0.0337
-----------------	--------	------------------------------	--------



Critérios de bondade de ajuste Deviance, Pearson e Lackfit.

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	2107.2377	1659	1.2702	<.0001
Pearson	1670.5856	1659	1.0070	0.4160

Number of unique profiles: 1665

Partition for the Hosmer and Lemeshow Test					
Group	Total	Situation = 1		Situation = 0	
		Observed	Expected	Observed	Expected
1	403	96	90.66	307	312.34
2	410	96	100.68	314	309.32
3	408	108	103.69	300	304.31
4	407	112	107.29	295	299.71
5	406	121	112.52	285	293.48
6	404	106	116.63	298	287.37
7	407	118	123.13	289	283.87
8	407	103	134.51	304	272.49
9	407	162	152.69	245	254.31
10	409	212	192.20	197	216.80

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
19.5530	8	0.0122

O resultado para o coeficiente de determinação generalizado mostra um baixo poder preditivo do modelo sob análise como menos de 1% neste caso. O gráfico da curva ROC mostra uma linha de 45 graus que representa o modelo somente com o intercepto e a curva para o modelo sob análise. Portanto, quando mais afastada estiver essa curva da linha de 45 graus melhor será o poder preditivo do modelo. Neste caso, o afastamento entre ambas linhas é representado pela área sob a curva do modelo que equivale a 0,5863.

Neste caso, o valor da estatística Deviance foi significativo ($Pr > ChiSq < 0,0001$) enquanto Pearson foi não significativo ($Pr > ChiSq = 0,4160$). Ambos critérios não são adequados de utilizar para o modelo sob análise visto que a variável independente d_2004 é contínua o que levou a criação de 1665 grupos (profiles) que, conseqüentemente,

possuem poucas observações em cada grupo fazendo com que haja falta de ajuste para a distribuição Qui-quadrado.

Por outro lado, o teste de falta de ajuste de Hosmer e Lemeshow mostra um valor de Qui-quadrado 19,5 significativo ($Pr > ChiSq = 0,0122$) indicando que o modelo sob análise não é adequado para os dados. Essa conclusão é corroborada pelo baixo valor preditivo do modelo sob análise de acordo ao coeficiente de determinação generalizado.

3.8.3. Modelagem em regressão logística

A depender do objetivo da pesquisa, a construção de um modelo de regressão logística (ou outra) pode ser realizada mediante a seleção automática das variáveis independentes considerando as seguintes situações:

- i) o objetivo da pesquisa é a predição de valores presentes ou futuros para a variável dependente e, portanto, sem foco no entendimento da relação biológica entre y e as variáveis independentes;
- ii) dispõe-se de grande número de variáveis independentes na base de dados com pouca orientação teórica para sua escolha.
- iii) grande quantidade de modelos candidatos a partir da combinação de variáveis no modelo (número de possíveis modelos aumenta em uma proporção 2^k , sendo que k = número de variáveis independentes do estudo).

Para um total de 10 variáveis independentes (considerado número pequeno em modelagem) um total de $2^{10} = 1024$ possíveis modelos de regressão para comparar. Para 20 variáveis independentes (número moderado), um total de 1.048.576 modelos candidatos podem ser ajustados.

3.8.3.1. Modelagem com o PROC LOGISTIC

O procedimento PROC LOGISTIC disponibiliza aos usuários algoritmos para a seleção automática de variáveis independentes na modelagem de regressão logística. Os algoritmos são os mesmos descritos anteriormente no Quadro 35.

Um dos objetivos da utilização do processo de seleção automática de variáveis independentes é reduzir o número de modelos candidatos por meio da criação de um subconjunto de variáveis independentes a partir da minimização do valor do critério de

informação de Akaike (AIC), por exemplo. Entretanto, comparar o ajuste de cada um dos 1024 modelos possíveis (para 10 variáveis independentes) seria impraticável.

No procedimento PROC LOGÍSTIC os métodos de seleção de variáveis são limitados para uso combinado com o nível de significância estabelecido para a entrada, permanência ou saída do modelo em processo de construção.

Portanto, uma alternativa é a utilização do método Stepwise combinado com o critério de seleção utilizando nível de significância com valor próximo a 1 (0,99) para que uma variável independente entre no modelo (SLENTY) e nele permaneça (SLSTAY).

Essa alternativa força que o SAS inclua cada variável independente no modelo a cada passo. Portanto, o número de passos será igual ao número de variáveis independentes disponíveis desde o modelo somente com intercepto (null model) até o último passo que finaliza com o modelo contendo todas as variáveis independentes (full model).

Neste caso, o SAS considera a maximização do incremento na Verossimilhança para a inclusão de cada variável independente em cada passo. Para fins de demonstração da seleção de variáveis independentes (subconjunto) de forma a minimizar o critério de informação de Akaike, considere o caso florestal 10 a seguir:

Caso florestal 10: Modelagem de danos em árvores causados por tempestades em florestas urbanas.

Após a passagem do furacão Irma na Flórida uma equipe de pesquisadores realizou o levantamento dos danos em árvores urbanas dentro de parcelas previamente estabelecidas de 0,04 ha cada distribuídas aleatoriamente em três localidades: Tampa, Gainesville e Orange County. Portanto, em cada parcela as variáveis registradas foram:

- Location: Gainesville, Orange County, Tampa
- plot_ID: numero da parcela (0.1 acre plot)
- pct_impervious: % de superfícies impermeáveis na parcela
- ba_plot_sqft: Área basal
- live_trees_plot: número de árvores
- num_bldgs_plot: número de prédios ao redor da parcela
- num_spp: número de espécies de árvores
- prop_native: proporção de árvores nativas
- avg_dbh: média do diâmetro a 1,3 m do solo
- avg_total_ht: média da altura total
- avg_pct_missing: média de perda de copa
- avg_cle: media de exposição da copa à luz variando de 5=totalmente aberto, 0=sem luz direta na copa
- std_ht: desvio padrão da altura das árvores
- damage: Dano (1 =dano presente, 0 = sem presença de dado)

A variável dependente binária “damage” foi modelada em função das 12 variáveis a nível de parcela com o objetivo de determinar se os danos têm relação com alguma das variáveis independentes. O seguinte modelo de regressão foi considerado como ponto de partida para a modelagem utilizando a opção **stepwise** do **proc logistic**.

$$\text{Logit}(\pi_i) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

Essa base de dados tem origem da pesquisa publicada no seguinte endereço: <https://www.sciencedirect.com/science/article/pii/S0169204622001165>.

Como visto, existem um total de 12 variáveis independentes que totalizam $2^{12} = 4096$ possíveis modelos candidatos de regressão logística para estimar a probabilidade de danos nas árvores devido ao efeito do furacão Irma.

Neste sentido, para realizar a criação do subconjunto de variáveis independentes a partir das 12 variáveis bem como a construção de um gráfico para a avaliação do critério de Akaike, será utilizado a seguinte sintaxe do PROC LOGISTIC para a seleção das variáveis independentes e o PROC SGPLOT para a construção do gráfico de comportamento do critério de informação de Akaike para o caso florestal 10.

```

ods output ModelBuildingSummary=Resumo;
ods output FitStatistics=Ajuste;

Title "Subset of predictor variables";
proc logistic data=hurricane;
  class location / param=ref;
  model damage (event="1")=location live_trees_plot num_spp prop_native avg_dbh
    avg_total_ht avg_pct_missing avg_cle ba_plot_sqft pct_impervious
    num_bldgs_plot std_ht / selection=stepwise slentry=0.99 slstay=0.99;
run;
title;

proc print data=Ajuste;
  run;

proc print data=Resumo;
  run;

ods graphics / reset noborder width=14cm height=10cm imagemap;
proc sgplot data=Ajuste;
  where (criterion="aic");
  series x=step y=interceptandcovariates / lineattrs=(color=blue thickness=2
    pattern=solid) transparency=0.4;
  xaxis label="número de variáveis independentes" integer values=(0 to 12 by 1)
    labelattrs=(family=calibri size=13 weight=bold)
    valueattrs=(family=calibri size=12 weight=normal);

  yaxis label="critério de informação de akaike"
    labelattrs=(family=calibri size=13 weight=bold)
    valueattrs=(family=calibri size=12 weight=normal);
run;

```

Além do PROC LOGISTIC a sintaxe solicita que um conjunto de dados nomeado “Ajuste” seja criado a partir da linha de programação ODS OUTPUT FITSTATISTICS. Esse arquivo contém uma tabela de valores de -2LogL , AIC e SBC para cada um dos 12 modelos ajustados.

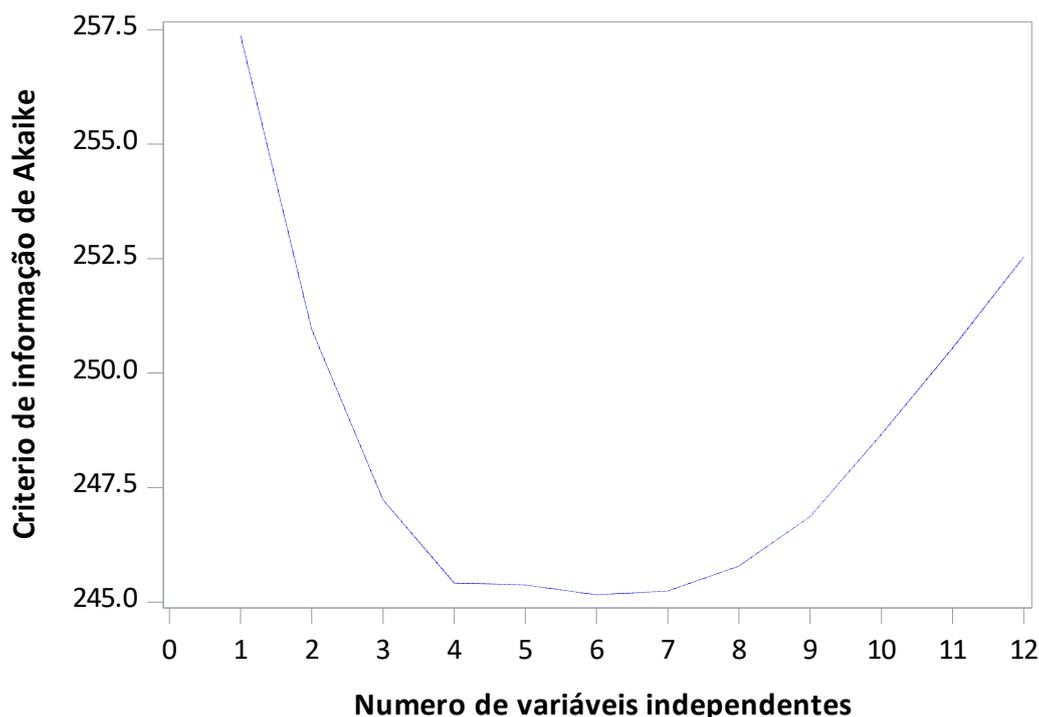
Outro arquivo de dados nomeado “Resumo” solicitado na linha de programação ODS OUTPUT MODELBUILDINGSUMMARY também é criado contendo o valor-p para cada modelo ajustado e outras estatísticas em uma tabela.

Para acessar os valores basta clicar no arquivo localizado na pasta Libraries subpasta Work do painel de navegação do SAS. As tabelas foram impressas em tela pelo PROC PRINT e os resultados são apresentadas do Output 48:

Output 48. Tabela “Resumo” contendo as variáveis independentes que entraram no modelo em cada um dos passos e arquivo “Ajuste” com as estatísticas de ajuste (Criterion) para cada um dos 12 modelos ajustados (Step). O gráfico de comportamento do critério de informação de Akaike também é mostrado.

Obs	Step	EffectEntered	EffectRemoved	DF	NumberInModel	ScoreChiSq	WaldChiSq	ProbChiSq
1	1	ba_plot_sqft		1	1	22.7507	—	<.0001
2	2	location		2	2	9.8897	—	0.0071
3	3	avg_dbh		1	3	5.5946	—	0.0180
4	4	num_spp		1	4	3.9505	—	0.0469
5	5	prop_native		1	5	2.0048	—	0.1568
6	6	live_trees_plot		1	6	2.0911	—	0.1482
7	7	std_ht		1	7	1.9915	—	0.1582
8	8	avg_cle		1	8	1.4683	—	0.2256
9	9	pct_impervious		1	9	0.9137	—	0.3391
10	10	num_bldgs_plot		1	10	0.2033	—	0.6521
11	11	avg_total_ht		1	11	0.1177	—	0.7315
12	12	avg_pct_missing		1	12	0.0069	—	0.9337

Obs	Step	RowId	Criterion	Equals	InterceptOnly	InterceptAndCovariates
1	0	M2LOGL0	-2 Log L	=	277.164000	277.164
2	1	AIC	AIC		279.164000	257.376
3	1	SBC	SC		282.506334	264.060
4	1	M2LOGL	-2 Log L		277.164000	253.376
5	2	AIC	AIC		279.164000	250.954
6	2	SBC	SC		282.506334	264.323
7	2	M2LOGL	-2 Log L		277.164000	242.954
8	3	AIC	AIC		279.164000	247.235
9	3	SBC	SC		282.506334	263.947
10	3	M2LOGL	-2 Log L		277.164000	237.235
11	4	AIC	AIC		279.164000	245.417
12	4	SBC	SC		282.506334	265.471
13	4	M2LOGL	-2 Log L		277.164000	233.417
14	5	AIC	AIC		279.164000	245.373
15	5	SBC	SC		282.506334	268.769
16	5	M2LOGL	-2 Log L		277.164000	231.373
17	6	AIC	AIC		279.164000	245.160
18	6	SBC	SC		282.506334	271.899
19	6	M2LOGL	-2 Log L		277.164000	229.160
20	7	AIC	AIC		279.164000	245.242
21	7	SBC	SC		282.506334	275.323
22	7	M2LOGL	-2 Log L		277.164000	227.242
23	8	AIC	AIC		279.164000	245.784
24	8	SBC	SC		282.506334	279.208
25	8	M2LOGL	-2 Log L		277.164000	225.784
26	9	AIC	AIC		279.164000	246.869
27	9	SBC	SC		282.506334	283.635
28	9	M2LOGL	-2 Log L		277.164000	224.869
29	10	AIC	AIC		279.164000	248.664
30	10	SBC	SC		282.506334	288.772
31	10	M2LOGL	-2 Log L		277.164000	224.664
32	11	AIC	AIC		279.164000	250.547
33	11	SBC	SC		282.506334	293.997
34	11	M2LOGL	-2 Log L		277.164000	224.547
35	12	AIC	AIC		279.164000	252.540
36	12	SBC	SC		282.506334	299.333
37	12	M2LOGL	-2 Log L		277.164000	224.540



A tabela “Resumo” mostra cada um dos modelos de regressão logística ajustado em cada um dos 12 passos. Cada variável entrou no modelo por ordem de importância aferido pelo índice “ScoreChisq”. Neste caso, é possível visualizar que o modelo com três variáveis independentes (Step 4) estima a probabilidade de danos às árvores urbanas causados por tormentas a partir dos dados de área basal da parcela (ba_plot_sqft), localização das parcelas (location), média do diâmetro a 1,3 m do solo dentro da parcela (avg_dbh) e número de espécies dentro da parcela (num_spp).

O gráfico mostra o decaimento do valor para o critério de informação de Akaike à medida que se inclui as variáveis independentes em um modelo único até o passo 4 (Step 4) com um valor de 245,417. Em seguida o critério apresenta pouca variação mesmo incluindo mais variáveis e logo aumenta.

Sobre os resultados alcançados até aqui é importante salientar que o fato de ter minimizado o critério de informação de Akaike não implica no melhor modelo obtido no passo 4 do processo (Três variáveis independentes adicionado do intercepto). Isso porque os resultados exibiram apenas uma sequência passo a passo e não todos os modelos possíveis.

Outra opção de modelagem é considerar a seleção do melhor modelo a partir das variáveis independentes candidatas pelo método de pontuação (score) do SAS que

seleciona o modelo de acordo do valor score Schisq (Qui-quadrado) por meio da declaração METHOD=SCORE combinado com as opções START=num1, STOP=num2 e BEST=b.

As opções START=s1, STOP=s2 restringe os modelos a possuírem apenas um número de variáveis independentes entre os valores declarados em num1 e num2. Logo, para cada modelo a opção BEST=b irá construir o “b” melhor modelo com a maior pontuação (score) de Qui-quadrado (Chisq). Por padrão, o PROC LOGISTIC imprime o valor score Schisq (pontuação de Qui-quadrado) na tabela “Testing Global Null Hypothesis: BETA=0”.

A sintaxe exemplo para a modelagem pelo método score no PROC LOGISTIC é a seguinte:

```
proc logistic data=exemplo;  
  model y = x1 x2 x3 x4 / selection=score start=1 stop=4 best=3;  
run;
```

Neste caso, o conjunto de dados do exemplo possui quatro variáveis independentes e de acordo com as configurações na declaração SELECTION= o resultado será os seguintes 10 modelos candidatos:

- Quatro modelos com uma variável independente cada: [x1], [x2], [x3], [x4];
- Seis modelos com duas variáveis independentes cada: [x1 x2], [x1 x3], [x1 x4], [x3 x3], [x2 x4], [x3 x4];
- Quatro modelos com três variáveis independentes cada: [x1 x2 x3], [x1 x2 x4], [x1 x3 x4], [x2 x3 x4];
- Um modelo com quatro variáveis independentes cada: [x1 x2 x3 x4].

Para cada opção de modelo, o PROC LOGISTIC seleciona os três modelos com maiores valores de score Chisq.

Importante notar que o método de seleção score não suporta o uso de variável independente do tipo categórica declarada em CLASS na modelagem.

3.8.3.2. Modelagem com o PROC HPLOGISTIC

Outro procedimento capaz de realizar a seleção automática de variáveis para regressão logística é o PROC HPLOGISTIC. Esse procedimento também incorpora os algoritmos de seleção automática de variáveis para construir modelos preditivos igual que o PROC LOGISTIC. Entretanto, o PROC HPLOGISTIC adiciona outros métodos de seleção de variáveis (AIC, SBC e outros) além do nível de significância como utilizado pelo PROC LOGISTIC.

Na sintaxe a seguir o PROC HPLOGISTIC utiliza o método FORWARD considerando a entrada de novas variáveis pelo critério de informação de Schwarz (SBC) para o caso florestal 10.

```
proc hplogistic data=hurricane;
  class location / param=ref;
  model damage (event="1")=location live_trees_plot num_spp prop_native avg_dbh
    avg_total_ht avg_pct_missing avg_cle ba_plot_sqft pct_impervious
    num_bldgs_plot std_ht;

  selection method=forward (select=sbc choose=sbc stop=none);
run;
```

O processo de cálculo para o critério de informação SBC considera o número de graus de liberdade do modelo de regressão e o tamanho da amostra utilizada para o ajuste do modelo. Conseqüentemente, o uso desse critério na modelagem favorece a seleção de um modelo parcimonioso apoiando na tomada de decisão para evitar problemas de modelos Overfitting.

Vale ressaltar que se um modelo de regressão possuir, por exemplo, uma variável independente categórica, o número de graus de liberdade do modelo é maior do que a quantidade de coeficientes de regressão. Neste caso, o PROC HPLOGISTIC considera esse detalhe no cálculo do critério SBC.

Com a opção SELECT=SBC a entrada de uma nova variável no modelo ocorre para aquela que causar uma diminuição no critério de Schwarz entre todas aquelas disponíveis para entrar. Em seguida, a opção CHOOSE=SBC seleciona o modelo, entre todos ajustados pelo forward, que resulta no menor valor para o sbc. A opção STOP=NONE

solicita que o PROC HPLOGISTIC continue realizando o processo de entrada de variáveis até que todas tenham sido avaliadas. O resultado com a seleção de variáveis pelo PROC HPLOGISTIC é apresentado no Output 49 com algumas tabelas suprimidas.

Output 49. Seleção de variáveis independentes considerando o modelo com menor valor para o critério de informação de Schwarz.

Selection Summary			
Step	Effect Entered	Number Effects In	SBC
0	Intercept	1	282.51
1	ba_plot_sqft	2	265.10
2	avg_dbh	3	262.28*
3	location	4	264.23
4	num_spp	5	265.34
5	prop_native	6	268.81
6	live_trees_plot	7	272.02
7	std_ht	8	275.25
8	avg_cle	9	279.20
9	pct_impervious	10	283.64
10	num_bldgs_plot	11	288.77
11	avg_total_ht	12	294.00
12	avg_pct_missing	13	299.33

*** Optimal Value of Criterion**

Selection stopped because all effects are in the model.

The model at step 2 is selected where SBC is 262,2826.

Selected Effects: Intercept avg_dbh ba_plot_sqft

Fit Statistics	
-2 Log Likelihood	246.15
AIC (smaller is better)	252.15
AICC (smaller is better)	252.27
BIC (smaller is better)	262.18

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	31.0116	2	<.0001

Parameter Estimates					
Parameter	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	-1.9263	0.3502	Infty	-5.50	<.0001
avg_dbh	0.06007	0.02310	Infty	2.60	0.0093
ba_plot_sqft	0.1055	0.04050	Infty	2.60	0.0092

A primeira tabela mostra que o valor de SBC minimizou quando o modelo de regressão considera as variáveis independentes avg_dbh e ba_plot_sqft com valor de SBC=262,28. A última tabela mostra que ambas variáveis são significativas.

A Figura 51 mostra as curvas de comportamento para os valores dos critérios de informação de Akaike (AIC) e Schwarz obtidos no Output 49. Observa-se que o critério de Akaike é minimizado a partir de três variáveis independentes e o critério de Schwarz começa a aumentar quando o modelo possui duas variáveis independentes. Essa diferença mostra o rigor do critério de SBC à medida que mais variáveis ingressam no modelo demonstrando.

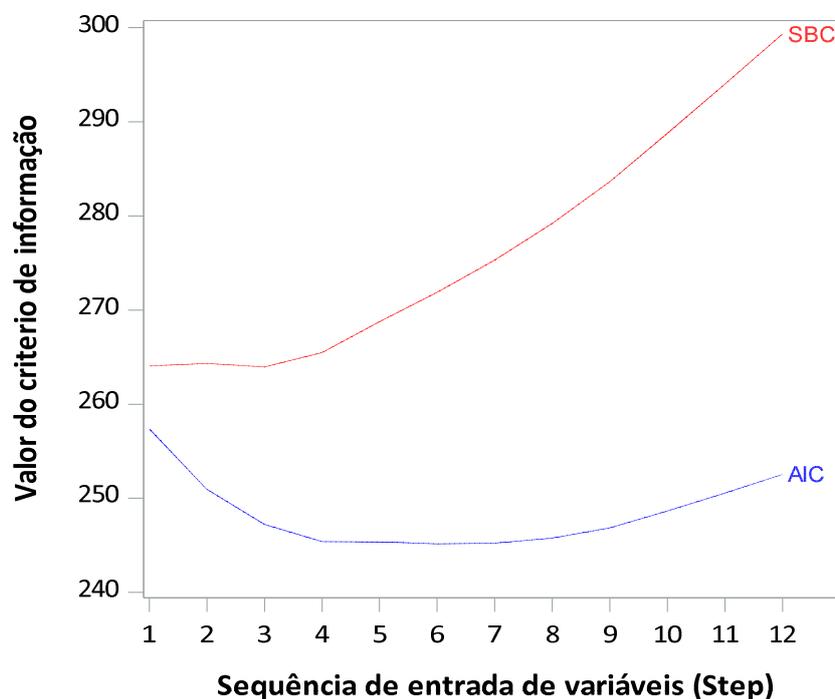


Figura 51. Comportamento dos critérios de informação de AKAIKE (AIC) e SCHWARZ (SBC) à medida que novas variáveis entram no modelo de regressão logística para estimar a probabilidade de danos em árvores do caso florestal 10. O passo (Step) 1 refere-se ao modelo somente com o intercepto.

REFERÊNCIAS

- Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. In: Petrov, B.N. and Csaki, F., Eds., **International Symposium on Information Theory**, p. 267-281, 1973.
- Allison, P.D. **Logistic regression using SAS: Theory and application**. SAS Press. 2012.
- Anscombe, F.J. Graphs in Statistical Analysis. **The American Statistician**, Vol. 27, p.17-21, 1973.
- Bates, D.M.; Watts, D.G. **Nonlinear Regression Analysis and Its Applications**. John Wiley & Sons, Inc. 1988.
- Bates, D.M.; Watts, D.G., **Nonlinear regression analysis and its applications**. New York John Wiley, 2007. 329p.
- Besley, D.A.; Kuh, E.; Welsch, R.E. **Regression Diagnostics: Identifying Influential Data and Sources of Collinearity**. John Wiley & Sons, Inc. 1980. 310 p.
- Box, G.E.P.; Cox, D.R. An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 26, N.2, p.211-252, 1964.
- Bredenkamp, B.V.; Gregoire. T.G. A Forestry Application of Schnute's Generalized Growth Function. **Forest Science**, Vol.34, Issue. 3, p.790–797, 1988.
- Burnham, K.P. and Anderson, D.R. **Model Selection and Inference: A Practical Information-Theoretic Approach**. 2nd Edition, Springer-Verlag, New York, 2002.
- Chapman, D.G. **Statistical problems in dynamics of exploited fisheries populations**. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. 1960.
- Coble, D.W.; Lee, Y. Use of a generalized sigmoid growth function to predict site index for unmanaged loblolly and slash pine plantations in East Texas. **Gen. Tech. Rep. SRS-92**. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. pp. 291-295, 2006.
- Cochran, W.G. **Técnicas de Amostragem**. John Wiley & Sons, INC. 1965, 555 p.
- Cordeiro, M.G.; Demétrio, C.G.B. **Apostila Modelos lineares generalizados e extensões**. 2010, 255p.
- Delwiche, L.D.; Slaughter, S.J. **The Little SAS Book: A Primer**. Cary, NC: SAS Institute, 2019. 346p.
- Draper, N.R.; Smith, H. **Applied Regression Analysis**. Wiley-Blackwell. 1981. 736 p.

Efron, B.; Hastie, T.; Tibshirani, I.J.R. Least angle regression. **Ann. Statist**, v.32, n.2, p.407-499, 2004.

Flom, P.L.; Cassell, D.L. **Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use**. NorthEast SAS Users Group, 2007.

Freese, F. **Elementary statistical methods for foresters**. Agriculture handbook US Department of Agriculture. 1967. 87 p.

Goodnight, J.H. et. al. **SAS Users Guide**. SAS Institute, Inc, 1979.

Hamilton, L.C. **Regression with graphics: a second course in applied statistics**. Pacific Grove. 1992. 363 p.

Harrell, F.E. **Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis**. Springer-Verlag, New York, 2001.

Hartie, T.; Tibshirani, R.; Friedman, F. **The elements of Statistical Learning: Data Mining, Inference and Prediction**. Springer second edition. 2016, 767p.

Heeringa, S.G.; West, B.T.; Berglund, P.A. **Applied Survey Data Analysis**. CRC Press. 2010. 442 p.

Hosmer, D.W.; Lemeshow, S. **Applied Logistic Regression**. John Wiley & Sons, Inc. 2000.

Hurvich, C.M.; Tsai, C.L. Bias of the corrected AIC criterion for underfitted regression and time series models. **Biometrika**, Vol. 78, Issue. 3, p.499–509, September 1991.

Husch, B.; Miller, C.I.; Kershaw, J. **Forest mensuration**. 2. ed. New Jersey: John Willey e Sons, Inc, 1972. 410 p.

Kershal, J.A.; Ducey, M.J.; Beers, T.W.; Husch, B. **Forest Mensuration**. 5. ed. John Willey e Sons, Inc, 2017. 633 p.

Kleinn, C. **Lecture Notes for the Teaching Module Forest Inventory. Department of Forest Inventory and Remote Sensing**. Faculty of Forest Science and Forest Ecology, Georg-August-Universität Göttingen. 2007, 164p.

Kozak, A.; Kozak, R.A.; Staudhammer, C.L.; Watts, S.B. **Introductory probability and statistics: applications for forestry and natural sciences**. CABI International, 2008, 448 p.

Kvalseth, T.O. Cautionary note about r. **The American Statistician**, vol.39, p.279–285, 1985.

Lei, Y.; Zhang, S.Y. Comparison and selection of growth models using the Schnute model **Journal of Forest Science**, vol. 52, n.4, p.188-196, 2006.

Lewis, T.H. **Complex Survey Data Analysis with SAS**. CRC Press. 2017. 326 p.

Loetsch, F.; Haller, K.E. **Forest inventory**. BLV-Munche: Basel, Wien. 1964, 436p.

- Loetsch, F.; Zoehrer, F.; Haller, K. E. **Forest inventory**. v. 2. Munchen: BVL, 1973, 469p.
- Machado, S.B.; Figueiredo Filho, A. **Dendrometria**. Curitiba, PR: Editado pelos autores, 2003. 309p.
- Mitscherlich, E.A. **Das Gesetz des Pflanzenwachstums**. Landw. Jahrb v.53 p 167-182. 1919.
- Mitscherlich, G.; Sonntag, G. Paperversuche: Modell für eine regenerata und Neupotz-papel- ertragstafel im Oberheingebiet. Allg. **Forst und Jg**, n.153, p.213-219, 1982.
- Montgomery, D.C.; Peck, E.A.; Vining, G.G. **Introduction to Linear Regression Analysis**. Wiley, 2012, 545p.
- Neter, J.; Wasserman, W.; Kutner, M.H. **Applied Linear Regression Models**. Richard D Irwin Inc. 1983. 547p.
- Neyman, J.; Pearson, E. S. On the problems of the most efficient tests of statistical hypotheses. **Philosophical Transactions of the Royal Society of London**, v.231A, p.289-338, 1933.
- Péllico N.S.; Brena, D. A. **Inventário Florestal**. Curitiba, Edição Autores, 1997. 316 p.
- Pienaar, L.V.; Turnbull, K.J. The Chapman-Richards Generalization of Von Bertalanffy's Growth Model for Basal Area Growth and Yield in Even-Aged Stands. **Forest Science**, v.19, p.2-22, 1973.
- Price, W.J.; Shafii, B.; Seefeldt, S.S. Estimation of Dose–Response Models for Discrete and Continuous Data in Weed Science. **Weed Technology**, v.26, p.587-601, 2017.
- Prodan, M.; Peters, R.; Cox, F.; Real, P. **Mensura Forestal**. San José: GTZ. 1997. 561 p.
- Rawlings, J.O.; Pantula, S.G.; Dickey, D.A. **Applied Regression Analysis: A Research Tool**. 2nd Edition, Springer, Berlin. 1998, 671p.
- Richards, F.J. A Flexible Growth Function for Empirical Use. **Journal of Experimental Botany**, v.10, p.290-300, 1959.
- Robert T. Regression Shrinkage and Selection via the Lasso. **Journal of the Royal Statistical Society**, Vol. 58, N.1, p.267-288, 1996.
- Rodrigues, R.N., **Building Regression models with SAS a guide for data scientists**. SAS Institute. 2023, 464p.
- Rodriguez, R.N. **Building Regression Models with SAS: A Guide for Data Scientists**. Cary, NC: SAS Institute Inc. 2023, 452p.
- Shapiro, S.S.; Wilk, M.B. An Analysis of Variance Test for Normality (Complete Samples). **Biometrika**, Vol. 52, N. 3/4, p. 591-611, 1965.

Sawa, T. Information Criteria for Discriminating among Alternative Regression Models. **Econometrica**, vol.46, p.1273–1282, 1978.

Schneider, P.R. et al. **Análise de regressão aplicada à engenharia florestal**. 2. ed. Santa Maria: Editora FACOS, 2009.

Schnute, J. A Versatile Growth Model with Statistically Stable Parameters. **Canadian Journal of Fisheries and Aquatic Sciences**, vol.38, p.1128-1140, 1981.

Schwarz, G. Estimating the Dimension of a Model. **Ann. Statist**, vol.6, n.2, p.461-464, 1978.

Scott, A.; Wild, W. **Transformations and R²**. American Statistical Association, vol.45, 1991.

Shiver, B.D.; Borders, B.E. **Sampling techniques for forest resource inventory**. John Wiley & Sons, INC. 1996. 356p.

Silva, J.A.A.; Bailey, R.L. Considerações teóricas sobre o uso correto do índice de Furnival na seleção de equações volumétricas. **Revista árvore**, vol.15, p. 323-327, 1991.

Sprugel, D. G. Correcting for Bias in Log-Transformed Allometric Equations. **Ecology**, vol.64, p.209-210, 1983.

Stauffer, H.B. A Sample Size Table for Forest Sampling. **Forest Science**, vol.28, p.777-784, 1982.

Stevens, S.S. On The Theory of Scales of Measurements. **Science**, vol.103, p.677-680, 1946.

Sweda, T.; Kouketsu, S. Applicability of Growth Equations to Growth of Trees in Stem Radius (II) Application to Jack Pine. **Journal of Japanese Forest Science**, v.66, p.402-411, 1984.

Thursby, J.G. Misspecification, Heteroscedasticity, and the Chow and Goldfeld-Quandt Tests. **The Review of Economics and Statistics**, vol.64, p.314-321, 1982.

Tjørve, E; Tjørve, K.M.C. A unified approach to the Richards-model family for use in growth analyses: Why we need only two model forms. **Journal of Theoretical Biology**, vol.267. p.417-425, 2010.

Ware, G.O.; Ohki, K.; Moon, L.C. The Mitscherlich Plant Growth Model for Determining Critical Nutrient Deficiency Levels. **Agronomy Journal**, vol.74, n.1, 88-91, 1982.

White, A. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. **Econometrica**, vol.48, p.817-838, 1980.

Winsor, C.P. The Gompertz Curve as a Growth Curve. **Proceedings of the National Academy of Sciences of the United States of America**, v.18, p.1-8, 1932.

Zeide, B. Analysis of Growth Equations. **Forest Science**, v.39, p.594-616, 1993.

Zhang, L. Cross-validation of Non-linear Growth Functions for Modelling Tree Height–Diameter Relationships. **Annals of Botany**, v.79, p.251-257, 1997.

AUTORES

Thiago Augusto da Cunha



Engenheiro Florestal pela Universidad Mayor de San Simón (2007) com mestrado e doutorado (2009 e 2013) na área de Manejo Florestal e linha de pesquisa Crescimento e Produção Florestal pela Universidade Federal de Santa Maria (UFSM). Fez estágio de doutorado sanduiche na Universität für Bodenkultur (BOKU) em Viena - Áustria. Atualmente é professor Associado I da Universidade Federal do Acre desde 2013 atuando no curso de Engenharia Floresta campus Rio Branco e professor permanente do Programa de Pós-graduação em Ciência Florestal (PPG Ciflor) desde 2016. Na graduação é professor regente das disciplinas i) Experimentação aplicada à Ciências Agrárias e ii) Exploração e Transporte Florestal e na pós-graduação pelas disciplinas i) Métodos Estatísticos Aplicados à Ciência Florestal e ii) Modelagem do Crescimento e da Produção Florestal. Coordenou o PPG Ciflor desde a primeira turma entre 2016 a 2018 e no período de 2020 a 2022.

Afonso Figueiredo Filho



Engenheiro Florestal pela Universidade Federal do Paraná (1976) onde também fez seu Mestrado e Doutorado (1983 e 1991) na área de Manejo Florestal. Pós-Doutorado na Universidade da Geórgia, EUA (1994-95). Professor na Universidade Federal do Paraná-UFPR (1978-1997), onde ainda atua como Professor Sênior na Pós-Graduação em Engenharia Florestal. É Professor Associado da Universidade Estadual do Centro-Oeste-UNICENTRO desde 1999. Coordenou convênio de intercâmbio acadêmico e coordena projeto de pesquisa e extensão internacionais. Atualmente é membro do Comitê de Assessores da Fundação Araucária e pesquisador 1B do CNPq. Também é Membro Titular da Academia Nacional de Engenharia (ANE) desde 2021 e Consultor de várias revistas científicas nacionais e internacionais. Atua na área de Recursos Florestais e Engenharia Florestal, com pesquisas em Florestas Naturais e Plantadas, com ênfase em estudos de dinâmica, modelagem do crescimento e produção, mensuração, inventário e manejo florestal.

ÍNDICE REMISSIVO

A

Acurácia: 86, 120, 237, 238 e 247.

Amostragem

Probabilística: 38, 53, 56 e 60.

Não-probabilística: 53, 55 e 60.

Aleatória: 56, 61, 63, 67, 72, 77, 79, 92 e 98.

Sistemática: 56.

ANOVA (Análise de Variância): 34, 100, 108, 117, 149, 153, 158, 160, 161, 166, 167, 168, 169, 170, 188, 209, 210, 266, 277 e 296.

Assíntota: 101, 253, 254, 255, 257, 258, 266, 269, 274, 276, 277, 279, 281, 284, 288, 289 e 290.

B

Backward (método de seleção): 156, 212, 213, 214, 217, 227, 228 e 298.

Bias: 239 e 240.

Bias-Variance: 239.

Box-Cox (Transformação): 200, 202, 203 e 204.

C

Coefficiente de determinação (R²): 100, 150, 172, 173, 207, 208, 216, 217, 286, 296, 310, 311, 312, 313 e 314.

Coefficiente de variação (CV): 49, 96, 97, 99, 150, 171 e 209.

Crescimento: 39, 42, 47, 50, 119, 184, 185, 193, 212, 247, 250, 251, 253, 254, 255, 256, 257, 258, 266, 269, 270, 271, 274, 284, 286, 287, 288, 289 e 290.

Censo: 38, 39, 40, 52 e 62.

Crítérios de informação: 173, 174, 216, 229, 235, 286, 287, 288, 290, 305 e 324.

AIC: 160, 161, 173, 174, 216, 219, 220, 222, 224, 227, 229, 230, 231, 232, 233, 235, 245, 283, 286, 287, 288, 304, 305, 315, 317, 319, 322, 323 e 324.

AICc: 160, 161, 173, 174, 216, 219, 220, 222, 224, 227, 229, 230, 231, 232, 233, 235, 245, 283, 286, 287, 288 e 323.

BIC: 155, 156, 173, 216, 283, 286, 287, 288 e 323.

Covariância: 117, 139 e 205.

D

Datalines: 29, 64, 89, 125, 127, 129, 131, 133, 143, 148, 152, 157, 159, 179, 193, 202, 263, 275 e 292.

Derivada: 118, 138, 247, 248, 251, 259, 260, 261 e 279.

Dummy: 47, 151, 152, 156, 161, 162, 163, 274, 275, 276, 280, 281 e 299.

Distância de Cook: 192, 195, 196 e 198.

E

Estimador: 137 e 205.

Escala de medição: 37, 42, 45, 46, 48 e 49.

Nominal: 47 e 49.

Ordinal: 47 e 49.

Intervalar: 47.

Razão: 47, 48 e 49.

Estudo Experimental: 49, 50 e 101.

Observacional: 49, 51 e 88.

Exatidão: 237.

Estratificação: 61, 72, 77, 78, 79, 298.

Erro padrão médio: 226, 241 e 246.

F

Fator de correção: 68, 69, 70, 71 e 201.

Forward (Método de seleção): 156, 212, 213, 214, 217, 227, 244, 245, 246, 298 e 322.

G

Glm (Procedimento SAS): 34, 35, 141, 151, 156, 157, 158, 161, 162, 164, 165, 173 e 215.

Glmselect (Procedimento SAS): 141, 151, 159, 160, 161, 164, 212, 213, 215, 216, 217, 218, 227, 228, 229, 241, 242, 243, 244, 245 e 246.

H

Heterocedasticidade: 177, 187, 188 e 205.

Hipótese: 42, 50, 100, 101, 102, 103, 105, 108, 109, 150, 155, 166, 167, 169, 176, 177, 178, 182, 185, 186, 187, 188, 189, 217, 228, 257, 266, 274, 275, 277, 288, 298 e 306.

Nula: 50, 100, 103, 108, 167, 178, 182, 185, 186, 187, 188, 189, 274, 275, 277, 288 e 306.

Alternativa: 50, 100, 103, 105, 108, 167 e 186.

I

Interativo: 14, 97, 259, 260, 261 e 279.

Input: 29, 30, 64, 75, 81, 83, 89, 125, 126, 127, 129, 131, 133, 143, 148, 151, 152, 157, 159, 179, 193, 202, 244, 245, 263, 275 e 292.

Import (Procedimento SAS): 23, 31 e 32.

Iml (Procedimento SAS): 142 e 145.

K

Kolmogorov-Smirnov: 89, 94, 178, 180 e 182.

L

Logística: 35, 49, 108, 117, 252, 253, 257, 290, 291, 297, 298, 299, 300, 301, 302, 306, 307, 308, 310, 311, 312, 314, 316, 320, 322 e 324.

Libname (Declaração SAS): 31, 32, 33, 242 e 244.

M

Média aritmética: 49, 71, 87, 88, 92, 93, 150, 166, 168, 170, 171, 172 e 177.

Means (Procedimento SAS): 35, 69, 70, 71, 77, 78, 83, 92, 93, 109, 110, 112 e 154.

Missing values: 294.

Mínimos Quadrados Ordinários: 137, 139, 176, 177, 206, 208, 248, 259, 291, 294 e 296.

Matriz: 126, 127, 128, 137, 138, 139, 142, 143, 144, 145, 162, 163, 164, 205 e 307.

Identidade: 138.

Inversa: 164.

Inversa de Moore-Penrose: 164.

N

Nível de significância: 97, 100, 103, 108, 109, 114, 156, 167, 189, 206, 214, 215, 216, 217, 218, 220, 315 e 322.

Nlin (Procedimento SAS): 35, 165, 259, 260, 261, 262, 263, 266, 269, 270, 271, 272, 275, 276, 278, 279, 281, 284 e 285.

Nlmixed (Procedimento SAS): 165, 259, 277, 278, 279, 280, 281, 284 e 287.

Normalidade: 90, 177, 178, 179, 180, 182, 183, 200, 203, 204 e 296.

O

Otimização: 86, 87, 260 e 266.

P

Ponderado: 205 e 298.

Print (Procedimento SAS): 24, 26, 29, 30, 35, 36, 65, 76, 91, 92, 143, 144, 152, 153, 263, 266, 269, 292, 293, 294, 296, 317 e 318.

População: 37, 38, 39, 40, 52, 53, 55, 56, 57, 59, 60, 61, 62, 63, 65, 67, 68, 69, 70, 71, 72, 73, 74, 77, 78, 79, 83, 84, 88, 90, 92, 94, 95, 96, 98, 108, 119, 121, 126 e 236.

Polinômio: 118, 123, 135, 136, 213, 240 e 285.

Precisão: 38, 57, 68, 72, 77, 79, 96, 184, 237, 238 e 246.

Ponto de inflexão: 247, 252, 254, 255, 256, 257, 258, 262, 269, 274, 288, 289 e 290.

R

Regressão linear: 42, 49, 100, 108, 117, 118, 119, 120, 121, 122, 123, 124, 133, 134, 136, 137, 140, 141, 146, 156, 159, 161, 165, 166, 168, 170, 176, 184, 190, 191, 192, 195, 208, 216, 217, 240, 243, 245 e 291.

não-linear: 165, 247, 248, 249, 251, 252, 259, 260, 286 e 287.

logística: 35, 49, 108, 117, 290, 291, 297, 298, 299, 300, 301, 302, 306, 307, 308, 310, 311, 312, 314, 316, 320, 322 e 324.

Richards (Modelo de crescimento): 119, 185, 252, 253, 254, 255, 257, 259, 269, 270, 272, 274, 275, 276, 279, 281, 284, 287, 289 e 290.

Reg (Procedimento SAS): 134, 135, 141, 146, 147, 148, 151, 152, 153, 155, 156, 157, 161, 163, 164, 165, 179, 187, 188, 193, 196, 205, 207, 208, 209, 210, 213, 215, 216, 292 e 294.

S

Sampling frame: 52.

Sample size: 60, 61, 64, 72, 80, 83, 84, 90, 91, 92 e 93.

Surveyselect: 60, 63, 64, 67, 72, 73, 75, 79, 80, 83, 84, 86, 90, 91 e 93.

Shapiro-Wilk: 90, 178 e 180.

SAS/Stat: 13, 14, 34, 165 e 177.

SAS/Graph: 13 e 14.

Step (Data Step e Proc Step): 33, 34, 35, 36 e 63.

Scatter (Procedimento SAS): 133, 134, 271, 273 e 294.

Sgplot (Procedimento SAS): 132, 133, 270, 272 e 293.

T

Tamanho da amostra: 38, 61, 65, 68, 69, 87, 88, 96, 97, 98, 99, 100, 101, 102, 104, 106, 107, 109, 111, 114, 115, 297 e 322.

Tamanho Efetivo: 100, 105, 106, 107 e 109.

Tradeoff: 239.

Transreg (Procedimento SAS): 202.

U

Univariate: 88, 89, 92, 178, 179 e 180.

Univariada: 68.

V

Variável categórica: 27, 28, 29, 30, 31, 72, 130, 151, 164, 244 e 307.

Numérica: 27, 28, 30, 31, 47, 48, 72, 88, 90, 127, 129, 178, 290 e 291.

Dependente: 42, 43, 49, 88, 90, 117, 118, 119, 120, 121, 122, 123, 130, 132, 137, 140, 142, 148, 150, 153, 164, 167, 168, 170, 171, 172, 174, 176, 184, 185, 187, 190, 192, 195, 198, 200, 201, 202, 207, 208, 217, 236, 237, 240, 247, 249, 253, 256, 258, 259, 260, 278, 279, 281, 290, 291, 292, 294, 296, 297, 298, 299, 300, 301, 303, 305, 307, 309, 314 e 316.

Independente: 42, 43, 50, 56, 86, 87, 95, 96, 109, 118, 119, 120, 121, 122, 123, 129, 132, 137, 140, 141, 142, 146, 147, 148, 150, 151, 153, 161, 162, 166, 167, 169, 170, 172, 173, 174, 175, 184, 185, 186, 187, 188, 189, 192, 194, 195, 199, 202, 204, 205, 206, 207,

208, 209, 210, 212, 213, 214, 215, 216, 217, 218, 220, 235, 236, 238, 239, 240, 250, 251, 252, 257, 290, 291, 296, 298, 299, 300, 301, 302, 303, 305, 306, 308, 310, 311, 313, 314, 315, 316, 317, 318, 320, 321, 322, 323 e 324.

ISBN: 978-65-80261-37-6

BR



9 786580 261376

DOI: 10.35170/ss.ed.9786580261376

