Universidade Federal do Rio Grande do Sul

Instituto de Matemática

Programa de Pós-Graduação em Matemática

MODELOS DE REGRESSÃO LOGÍSTICA

por

CLEONIS VIATER FIGUEIRA

Porto Alegre, 31 de março de 2006.

Dissertação submetida por Cleonis Viater Figueira* como requisito parcial para obtenção do grau de Mestre em Matemática pelo Programa de Pós-Graduação em Matemática do Instituto de Matemática da Universidade Federal do Rio Grande do Sul.

Professora Orientadora:

Dra. Sílvia Regina Costa Lopes

Banca Examinadora:

Dr. Artur Oscar Lopes

Dra. Sara Ianda Correa Carmona

Dra. Sílvia Regina Costa Lopes

Dra. Hildete Prisco Pinheiro (IMECC/UNICAMP)

Data da Defesa: 31 de março de 2006.

*Bolsista do Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq

“Models are to be used, but not to be believed.”

Henri Theil

À minha família.

Agradecimentos

À Deus, pela existência das pessoas que nomeio a seguir.

À minha orientadora, Sílvia Regina Costa Lopes, por sua orientação acadêmica, pelos conselhos de mãe e principalmente pelo exemplo de profissional e de mulher.

Aos meus pais Antonio Viater e Iria Lucia Viater por terem me ensinado a ler e a escrever e pelo apoio incondicional que fez de mim mais forte do que pensei ser.

Ao meu cônjuge Jalves Sampaio Figueira por querer estar ao meu lado com sua presença e seu carinho. Ao meu filho Vicente Viater Figueira por fazer meus olhos brilharem.

Aos colegas da Pós-Graduação: Rosane, Edite, Eliane, Lisiane, Mauricio, Cícero, Marcio, Cleber, Edson, Gabriela, Marcus, e a todos os outros que freqüentam ou freqüentaram a "salinha" da pós. De modo especial refiro-me ao Guilherme Pumi e a Joice Hunch pela amizade construída e por todos os momentos de estudo e, também, à Virginia Silva Rodrigues por se tornar a irmã que a vida até então não havia me dado.

À Rosane, nossa querida secretária do Programa de Pós-Graduação, pelo carinho e dedicação para com todos os alunos.

Aos professores: Alexandre Baraviera, Alveri Alves S'Antana, Jaime Bruck Ripoll, Leonardo Prange Bonorino e Sara Ianda Correa Carmona pelas aulas ministradas, pela cordialidade no atendimento extra-classe e pelos bons exemplos que irei levar por toda esta vida. À professora Hildete Prisco Pinheiro, IMECC (Unicamp), por gentilmente fornecer os exemplos que foram utilizados neste trabalho.

Ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) pelo apoio financeiro que foi de grande valia para a conclusão de meu trabalho.

Resumo

O presente trabalho discute formalmente a Análise de Regressão Linear em suas formas simples, múltipla e multivariada e a Análise de Regressão Logística Nominal em suas formas binária, múltipla e multinomial. São apresentados estimadores aos parâmetros envolvidos em cada modelo, suas propriedades estatísticas e também critérios para se julgar a adequabilidade dos modelos. Exemplos de aplicação da teoria desenvolvida são comentados.

Abstract

The present study formally discusses the Linear Regression Analysis in its simple, multiple and multivariate forms, and the Nominal Logistic Regression Analysis in its binary, multiple, and multinomial forms. Estimators concerning the parameters involved in each model are presented as well as their statistical proprieties and criteria to judge the suitability of the models. Examples of the theory application are commented.

Índice

Lista de Figuras

Lista de Tabelas

Introdução

Ao se estudar um fenômeno, deve-se coletar informações e dados sobre ele. A partir destes dados, analisá-los e obter conclusões e propriedades. Entretanto, em muitas situações, é impossível efetuar a coleta de informações de toda uma população porque se tornaria muito despendiosa ou muito tempo seria gasto.

Este é o contexto em que se insere a Análise de Regressão, uma ferramenta extremamente poderosa.

Através dos recursos matemáticos e estatísticos oferecidos pela Análise de Regressão pode-se encontrar alguma função que estime o comportamento do conjunto de dados que não se dispõe, a partir dos dados coletados.

O termo regressão tem uma origem interessante. Apareceu pela primeira vez na literatura em Galton (1885). Trabalho este realizado pelo antropologista Sir Francis Galton que investigava a relação entre a altura dos pais e a de seus filhos. Concluiu, de forma não surpreendente, que pais altos tendem a ter filhos altos e pais baixos tendem a ter filhos baixos. Entretanto, Galton também percebeu que muitos dos pais de grande estatura tem filhos menores e muitos pais de pequena estatura tem filhos mais altos do que eles próprios. Galton chamou este fenômeno de *regression toward the mean*, ou seja, de regressão para a média (aqui a palavra regressão tem a conotação de retorno).

O objetivo desta dissertação é apresentar uma revisão bibliográfica sobre as idéias básicas dos modelos de regressão linear e logística, tais como suposições envolvidas, aspectos de inferência e exemplos de aplicações.

O presente trabalho é dividido em quatro capítulos, descritos resumidamente a seguir.

O Capítulo 1 apresenta a discussão do modelo de regressão linear geral, que se baseia no uso de uma função linear. Inicia com o caso simples onde figura apenas uma variável aleatória e uma variável independente, partindo-se, posteriormente, para a regressão linear múltipla e multivariada. Na seqüência são apresentadas suas respectivas tabelas ANOVA (do inglês, *Analysis of Variance*) e a Análise de Resíduos, com a finalidade de apresentar subsídios para a avaliação do modelo proposto. No final desde capítulo é apresentado exemplo de aplicação.

O Capítulo 2 é estruturado de forma análoga ao Capítulo 1 e descreve o modelo de regressão logística em sua forma binária, múltipla e multinomial. Também se faz exemplo de aplicação.

A análise de dados reais, com o uso do *software* SPSS 10.0 for *Windows* (*Statistical Package for the Social Sciences*), é apresentada no Capítulo 3. Os dados analisados são oriundos do *Institute of Public Health - Faculty of Health Sciencies*, na Dinamarca, através do trabalho de pesquisa do professor Morten Frydenberg com pacientes portadores de câncer de próstata e apresentam resultados de exames médicos básicos buscando diagnosticar a presença ou não de tumor na cápsula prostática.

As conclusões finais referentes a este trabalho acadêmico são apresentadas no Capítulo 4.

Capítulo 1

Regressão Linear

Neste capítulo propõe-se a discutir o modelo de regressão linear em sua forma mais simples (caso bivariado) e posteriormente os modelos de regressão múltipla e multivariada. Inicialmente, apresentam-se dois exemplos que ilustram como os modelos de regressão linear aparecem em determinadas situações. Outros exemplos podem ser encontrados em Bickel e Doksum (1976), Bisquerra, Sarriera e Martínez (2004), Bussab e Morettin (2004) e Mood, Graybill e Boes (1986).

Exemplo 1.1. A distância s em que uma partícula se desloca no intervalo de tempo x é dada pela expressão $s = \beta_0 + \beta_1 x$, onde β_0 é a posição no instante $t = 0$ e β_1 é a velocidade média. Se β_0 e β_1 não são conhecidos, então o valor de s pode ser observado para dois valores distintos de x e, desta forma, tem-se condições para determinar valores para β_0 e β_1. Suponha-se, agora, que por alguma razão a distância não possa ser medida de forma exata, mas que exista um erro de medida que possui natureza aleatória. Então, o valor s não pode ser observado. Suponha que possa ser observado o valor y, dado por $y = s + \varepsilon$, onde ε é um erro aleatório com média zero. Substituindo nesta última igualdade a expressão correspondente a s, tem-se $y = \beta_0 + \beta_1 x + \varepsilon$, onde y é uma variável aleatória observada, x é uma variável não-aleatória observada, ε é uma variável aleatória não observada e β_0 e β_1 são constantes (ou parâmetros) não conhecidas. Neste novo contexto, não se pode mais obter os valores para β_0 e β_1 a partir de duas observações de y e x, como apresentado anteriormente, pois aqui não temos relação ou expressão matemática que conecte tais variáveis. Então, utilizam-se métodos estatísticos e observação de vários dados de y e x para se obter estimadores de β_0, β_1, pois $s = \beta_0 + \beta_1$.

Exemplo 1.2. Considere a relação entre a altura y e o peso x de habitantes de uma certa cidade. A princípio não existe uma função que descreva a relação entre y e x. No entanto, parece haver algum tipo de relação. Considera-se, então, para estudo, a altura e o peso como variáveis aleatórias, respectivamente Y e X, e define-se que o vetor aleatório (X, Y) tenha distribuição normal bivariada. Então a esperança matemática de Y dado um valor x de X é dada por $\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x$, onde β_0 e β_1 são funções dos parâmetros da função densidade de uma normal bivariada. Embora não haja uma função entre Y e X decorre, da suposição de que possuem distribuição conjunta normal, que existe uma relação linear entre os pesos e o valor médio das alturas. Dessa forma, pode-se apresentar o seguinte: Y e X tem distribuição normal conjunta e, $\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x$ ou que $y = \beta_0 + \beta_1 x + \varepsilon$, onde ε é uma variável aleatória, com distribuição normal, denominada erro.

1.1 Regressão Linear Simples

Considera-se o caso de duas variáveis (análise bivariada). O objetivo é desenvolver um modelo estatístico que possa ser usado para prever valores de uma variável dependente Y em função de uma variável independente X. Ambas as variáveis aleatórias (v.a.'s) estão definidas sobre uma mesma população $\mathcal{P}$, em um mesmo espaço de probabilidade.

Supõe-se dispor de uma amostra de n unidades, e para cada observação, tem-se um par de valores das variáveis X e Y, denotada por (x_i, y_i), $i \in \{1, \cdots, n\}$.

Seja $\mu(\cdot)$ uma função afim definida em um domínio D, onde D pode ser o conjunto dos números reais ($\mathbb{R}$) ou mesmo um sub-intervalo de $\mathbb{R}$. Define-se $\mu(x) = \beta_0 + \beta_1 x$, onde x pertence ao domínio D. Para modelar as situações apresentadas nos Exemplos 1.1 e 1.2, assume-se que existe uma família de funções de distribuição acumulada (f.d.a.), uma para cada x do domínio, tal que a média da f.d.a. correspondente a um dado x_0 pertencente ao domínio, é $\beta_0 + \beta_1 x_0$. Então, as médias das f.d.a., estão na reta definida por $\mu(x) = \beta_0 + \beta_1 x$. Assim, procura-se tomar uma amostra para alguma f.d.a. e com base na amostra fazer inferências sobre os parâmetros β_0, β_1 e $\mu(x)$, com base num modelo linear, definido a seguir.

Definição 1.1. *Modelo Linear*: Seja $\mu(x) = \beta_0 + \beta_1 x$, para todo $x \in D$. Para cada x pertencente a D, seja $F_{Y_x}(\cdot)$ uma função de distribuição acumulada (f.d.a.), com média $\mu(x)$ e variância σ^2_ε. Para cada x_i, seja Y_i uma amostra aleatória de tamanho

um da f.d.a. $F_{Y_i}(\cdot)$, para $i \in \{1, \cdots, n\}$. Então, os pares (x_i, Y_i), $i \in \{1, \cdots, n\}$, formam um conjunto de n observações de forma que

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i \quad \text{e} \quad Var(Y_i) = \sigma_\varepsilon^2. \tag{1.1}$$

É interessante notar que decorre da Definição 1.1 que a equação (1.1) pode ser interpretada, para cada $i \in \{1, \cdots, n\}$, como

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{onde}$$

$$\mathbb{E}(\varepsilon_i | x_i) = 0 \quad \text{e} \quad Var(\varepsilon_i | x_i) = \sigma_\varepsilon^2. \tag{1.2}$$

A expressão (1.2) motiva as suposições para as v.a.'s envolvidas:

(i) A variável X é por hipótese controlada e não está sujeita a variações aleatórias. Diz-se que X é uma variável fixa ou determinística.

(ii) Para dado valor x de X, os erros distribuem-se ao redor da média $\beta_0 + \beta_1 x$ com média zero, isto é, $\mathbb{E}(\varepsilon_i | x_i) = 0$, para todo $i \in \{1, \cdots, n\}$.

(iii) Os erros devem ter variabilidade constante em torno de X, ou melhor, $Var(\varepsilon_i | x_i) = \sigma_\varepsilon^2$, para todo $i \in \{1, \cdots, n\}$.

(iv) Os erros são não-correlacionados.

Em algumas situações se farão necessárias suposições, para quaisquer pares (i, j) tais que $1 \leq i, j \leq n$, tais como:

- Caso A: assume-se que a f.d.a. $F_{Y_i}(\cdot)$ é normal e que as v.a. Y_i e Y_j são independentes, para todo $i, j \in \{1, \cdots, n\}$, com $i \neq j$.

- Caso B: assume-se que as v.a. Y_i e Y_j são não-correlacionadas, para todo $i, j \in \{1, \cdots, n\}$, com $i \neq j$.

Os procedimentos de inferência para estas duas situações serão apresentados a seguir. Primeiramente aborda-se o Caso A.

Caso A:

Considere que as v.a.'s $Y_1, Y_2, \cdots, Y_n$ são independentes e igualmente distribuídas com distribuição normal satisfazendo a equação (1.1).

Inicialmente definem-se os elementos a seguir

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}_{n\times 2}, \quad \mathcal{B} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}_{2\times 1}$$

$$e \quad \mathcal{E} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}.$$

Dessa forma, o modelo linear apresentado na Definição (1.1) pode ser escrito matricialmente como

$$\mathbf{Y} = \mathbf{X}\mathcal{B} + \mathcal{E}.$$

O teorema, a seguir, apresenta os estimadores de máxima verossimilhança para $\mathcal{B}$ e σ^2_ε.

Teorema 1.1. *Considere o modelo linear dado na Definição 1.1. Seja* $\mathbf{Y} = \mathbf{X}\mathcal{B}+\mathcal{E}$, *onde* $\mathbf{X}$ *possui posto* $n-2$. *Então o estimador de* $\mathcal{B}$ *é dado por*

$$\hat{\mathcal{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \;\; \textit{e possui distribuição } N_2(\mathcal{B}, \sigma^2_\varepsilon(\mathbf{X}'\mathbf{X})^{-1})$$

e é distribuído independentemente dos resíduos $\hat{\mathcal{E}} = \mathbf{Y} - \mathbf{X}\hat{\mathcal{B}}$. *E, ainda,*

$$n\hat{\sigma}^2_\varepsilon = \hat{\mathcal{E}}'\hat{\mathcal{E}} \;\; \textit{e possui distribuição } \sigma^2_\varepsilon\chi^2_{n-2},$$

onde $\hat{\sigma}^2_\varepsilon$ *é o estimador de máxima verossimilhança de* σ^2_ε.

Demonstração: A função de verossimilhança conjunta das variáveis aleatórias $Y_1, Y_2, \cdots, Y_n$ é dada por

$$\begin{aligned} \mathcal{L}(\mathcal{B}, \sigma^2_\varepsilon) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} \exp\left(\frac{-\mathcal{E}'\mathcal{E}}{2\sigma^2_\varepsilon}\right) \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n_\varepsilon} \exp\left[\frac{-(\mathbf{Y}-\mathbf{X}\mathcal{B})'(\mathbf{Y}-\mathbf{X}\mathcal{B})}{2\sigma^2_\varepsilon}\right]. \end{aligned} \tag{1.3}$$

Para σ^2_ε fixo, a função de verossimilhança é maximizada quando se minimiza a expressão $(\mathbf{Y}-\mathbf{X}\mathcal{B})'(\mathbf{Y}-\mathbf{X}\mathcal{B})$, ou seja, quando $\hat{\mathcal{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, o qual não depende do σ^2_ε adotado.

Agora, maximizando $\mathcal{L}(\hat{\mathcal{B}}, \sigma_\varepsilon^2)$ com respeito a σ_ε^2 tem-se que

$$\mathcal{L}(\hat{\mathcal{B}}, \hat{\sigma}_\varepsilon^2) = \frac{1}{(2\pi)^{\frac{n}{2}}(\hat{\sigma}_\varepsilon^2)^{\frac{n}{2}}} \exp\left(\frac{-n}{2}\right),$$

onde $\hat{\sigma}_\varepsilon^2 = \dfrac{(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})}{n}$.

Tem-se que $\hat{\mathcal{B}}$ e $\hat{\mathcal{E}}$ podem ser representados como combinações lineares das variáveis $\mathcal{E}$, que por suposição, apresentam distribuição normal. De forma mais específica, tem-se

$$\begin{bmatrix} \hat{\mathcal{B}} \\ \hat{\mathcal{E}} \end{bmatrix} = \begin{bmatrix} \mathcal{B} \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \end{bmatrix} \mathcal{E}.$$

Isto decorre do fato de que $\mathbf{Y} = \mathbf{X}\mathcal{B} + \mathcal{E}$ e, dessa forma, $\hat{\mathcal{B}}$ pode ser representado por

$$\begin{aligned} \hat{\mathcal{B}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\mathcal{B} + \mathcal{E}) \\ &= \mathcal{B} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathcal{E}, \end{aligned}$$

e, $\hat{\mathcal{E}}$, pode ser apresentado por

$$\begin{aligned} \hat{\mathcal{E}} &= [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} = [\mathbf{I} - \mathbf{X}('\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][\mathbf{X}\mathcal{B} + \mathcal{E}] \\ &= [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathcal{E}, \end{aligned}$$

visto que $[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{X} = \mathbf{X} - \mathbf{X} = \mathbf{0}$.

Com base nas propriedades de esperança matemática e de variância, segue-se que

$$\begin{aligned} \mathbb{E}(\hat{\mathcal{B}}) &= \mathcal{B} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}(\mathcal{E}) = \mathcal{B} \\ Var(\hat{\mathcal{B}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Var(\mathcal{E})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}, \end{aligned}$$

e ainda,

$$\begin{aligned} \mathbb{E}(\hat{\mathcal{E}}) &= [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbb{E}(\mathcal{E}) = \mathbf{0} \\ Var(\hat{\mathcal{E}}) &= [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']Var(\mathcal{E})[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\ &= \sigma_\varepsilon^2[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']. \end{aligned}$$

Agora, procura-se avaliar a $\mathbb{E}(\hat{\mathcal{E}}'\hat{\mathcal{E}})$, para tal é interessante lembrar que $\hat{\mathcal{E}}'\hat{\mathcal{E}}$ tem dimensão 1×1. Com base na definição de $\hat{\mathcal{E}}$, tem-se que

$$\begin{aligned}
\hat{\mathcal{E}}'\hat{\mathcal{E}} &= \mathcal{E}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathcal{E} \\
&= \mathcal{E}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathcal{E} \\
&= tr\{\mathcal{E}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathcal{E}\} \\
&= tr\{[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathcal{E}\mathcal{E}'\}. \qquad (1.4)
\end{aligned}$$

Aplicando-se esperança na igualdade obtida em (1.4), conclui-se que

$$\begin{aligned}
\mathbb{E}(\hat{\mathcal{E}}'\hat{\mathcal{E}}) &= tr\{[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbb{E}(\mathcal{E}\mathcal{E}')\} \\
&= \sigma_\varepsilon^2 tr[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \\
&= \sigma_\varepsilon^2 tr(\mathbf{I}) - \sigma_\varepsilon^2 tr[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \\
&= \sigma_\varepsilon^2 n - \sigma_\varepsilon^2 tr[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} \\
&= \sigma_\varepsilon^2 n - \sigma_\varepsilon^2 tr\,[\mathbf{I}_{2\times 2}] \\
&= \sigma_\varepsilon^2 (n-2). \qquad (1.5)
\end{aligned}$$

Da espressão obtida em (1.5) segue o resultado para $\hat{\sigma}_\varepsilon^2$. □

Como $\mathbf{X}$ fica fixo, tem-se que $\hat{\mathcal{B}}$ e $\hat{\mathcal{E}}$ apresentam distribuição normal conjunta e são independentes.

Agora, que o valor de $Cov(\hat{\mathcal{B}}, \hat{\mathcal{E}}) = \mathbf{0}$ é verificado da seguinte forma

$$\begin{aligned}
Cov(\hat{\mathcal{B}}, \hat{\mathcal{E}}) &= \mathbb{E}[(\hat{\mathcal{B}} - \mathcal{B})\hat{\mathcal{E}}'] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}(\mathcal{E}\mathcal{E}')[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \\
&= \sigma_\varepsilon^2 (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \mathbf{0}.
\end{aligned}$$

Agora, assume-se que $(\lambda, \mathbf{e})$ sejam o par de autovalor e autovetor de matriz $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, esta matriz é idempotente e assim

$$\begin{aligned}
\lambda\mathbf{e} &= [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{e} \\
&= [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']^2\mathbf{e} \\
&= \lambda[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{e} = \lambda^2\mathbf{e},
\end{aligned}$$

o que implica que $\lambda = 0$ ou $\lambda = 1$.

Por propriedades de matrizes sabe-se que $tr[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = n - 2$ e, ainda, $tr[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \lambda_1 + \cdots + \lambda_n$, onde $\lambda_1 \geq \cdots \geq \lambda_n$ são os autovalores de $[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']$.

Por conseqüência, existem exatamente $n-2$ autovalores com valor um e os demais são nulos. Assim, com base no Teorema da Decomposição Espectral (ver página 105 de Johnson e Wichern (1998)) tem-se que

$$[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \mathbf{e}_1\mathbf{e}_1' + \cdots + \mathbf{e}_{n-2}\mathbf{e}_{n-2}',$$

onde $\mathbf{e}_1, \cdots, \mathbf{e}_{n-2}$ são os autovetores normalizados associados respectivamente aos autovalores $\lambda_1 = \cdots = \lambda_{n-2} = 1$.

Seja V uma matriz definida como $\mathrm{V} = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_{n-2} \end{bmatrix} = \begin{bmatrix} \mathbf{e}_1'\mathcal{E} \\ \mathbf{e}_2'\mathcal{E} \\ \vdots \\ \mathbf{e}_{n-2}'\mathcal{E} \end{bmatrix}$.

Então, V possui distribuição normal com média $\mathbf{0}$ e

$$Cov(V_j, V_k) = \begin{cases} \sigma_\varepsilon^2 \mathbf{e}_j'\mathbf{e}_k, & j = k \\ 0, & \text{caso contrário.} \end{cases}$$

Pode-se concluir que os V_j são independentes e possuem distribuição $N(0, \sigma_\varepsilon^2)$ e ainda que

$$\begin{aligned} n\hat{\sigma}_\varepsilon^2 &= \hat{\mathcal{E}}'\hat{\mathcal{E}} = \mathcal{E}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathcal{E} \\ &= V_1^2 + V_2^2 + \cdots + V_{n-2}^2, \end{aligned}$$

possui distribuição $\sigma_\varepsilon^2 \chi_{n-2}^2$. □

Corolário 1.2. *Para os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ tem-se*

$$\mathbb{E}(\hat{\beta}_0) = \beta_0, \;\; \mathbb{E}(\hat{\beta}_1) = \beta_1, \;\; Var(\hat{\beta}_0) = \frac{\sigma_\varepsilon^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2},$$

$$Var(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \;\; e \;\; Cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma_\varepsilon^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})}. \tag{1.6}$$

Demonstração: Uma outra forma de se analisar o resultado obtido Teorema 1.1 acima é considerar $(\hat{\beta}_0, \hat{\beta}_1)$ como uma v.a. com distribuição normal bivariada. Dessa forma segue-se o resultado. □

Observação 1.1. Em algumas situações a serem apresentadas mais adiante mais adiante, serão necessários os estimadores de $Var(\hat{\beta}_0)$ e de $Var(\hat{\beta}_1)$, denotados por $\widehat{Var(\hat{\beta}_0)}$ e $\widehat{Var(\hat{\beta}_1)}$, respectivamente. Tais estimadores são obtidos substituindo-se σ_ε^2 por seu estimador nos resultados do corolário acima.

Corolário 1.3. *Para o estimador* $\hat{\mu}(x)$ *de* $\mu(x) = \beta_0 + \beta_1 x$ *tem-se*

$$\begin{aligned} \mathbb{E}[\hat{\mu}(x)] &= \mathbb{E}[\hat{\beta}_0 + \hat{\beta}_1 x] = \mu(x) \\ Var[\hat{\mu}(x)] &= Var(\hat{\beta}_0 + \hat{\beta}_1 x) = \sigma_\varepsilon^2 \left[\frac{1}{n} + \frac{(\bar{x} - x)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]. \end{aligned}$$

Demonstração: A primeira igualdade é obtida facilmente através das propriedades da esperança matemática e do resultado do Corolário 1.2. Para a segunda igualdade tem-se

$$\begin{aligned} Var[\hat{\mu}(x)] &= Var(\hat{\beta}_0) + 2x\, Cov(\hat{\beta}_0, \hat{\beta}_1) + x^2 Var(\hat{\beta}_1) \\ &= \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \left(\frac{\sum_{i=1}^n x_i^2}{n} - 2x\bar{x} + x^2 \right) \\ &= \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \left[(\bar{x} - x)^2 + \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right] \\ &= \sigma_\varepsilon^2 \left[\frac{1}{n} + \frac{(\bar{x} - x)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]. \end{aligned}$$

□

Observação 1.2. Em geral, estimadores de máxima verossimilhança não são estimadores não-viciados de uniforme mínima variância (UMVUE). A seguir, apresenta-se um resultado que garante esta propriedade aos estimadores $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\sigma}_\varepsilon^2$ dos parâmetros β_0, β_1 e σ_ε^2, respectivamente.

Teorema 1.4. *Seja o modelo linear dado pela Definição 1.1. Considere* $h(\beta_0, \beta_1, \sigma_\varepsilon^2)$ *uma função conhecida dos parâmetros* β_0, β_1 *e* σ_ε^2. *Então, existe um estimador de* $h(\beta_0, \beta_1, \sigma_\varepsilon^2)$ *que é uma função de* $\hat{\beta}_0$, $\hat{\beta}_1$ *e* $\hat{\sigma}_\varepsilon^2$, *denotado por* $\hat{h}(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_\varepsilon^2)$. *Este é o UMVUE de* $h(\beta_0, \beta_1, \sigma_\varepsilon^2)$.

Demonstração: Os estimadores não viciados de β_0, β_1 e σ_ε^2 são funções das estatísticas suficientes e completas $\sum_{i=1}^n Y_i$, $\sum_{i=1}^n Y_i^2$ e $\sum_{i=1}^n x_i Y_i$. Dessa forma, este teorema é um caso particular do Teorema de Lehmann-Scheffé. □

Do Teorema 1.4 decorrem os seguintes corolários.

Corolário 1.5. *O estimador UMVUE de cada um dos parâmetros* β_0, β_1 *e* σ_ε^2 *é dado, respectivamente, por* $\hat{\beta}_0$, $\hat{\beta}_1$ *e* $\hat{\sigma}_\varepsilon^2$.

Corolário 1.6. *O estimador UMVUE de* $\mu(x) = \beta_0 + \beta_1 x$, *para todo* $x \in D$, *é dado por* $\hat{\mu}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$.

Corolário 1.7. *Para quaisquer $c_0, c_1 \in \mathbb{R}$, o estimador UMVUE de $c_0\beta_0 + c_1\beta_1$ é dado por $c_0\hat{\beta}_0 + c_1\hat{\beta}_1$.*

As proposições a seguir fornecem meios de se obter intervalos a $100(1-\alpha)\%$ de confiança para os estimadores de β_0, β_1 e σ_ε^2.

Proposição 1.8. *O intervalo a $100(1-\alpha)\%$ de confiança para β_0 é dado por*

$$\left[\hat{\beta}_0 - t_{n-2,\frac{\alpha}{2}}\sqrt{\widehat{Var(\hat{\beta}_0)}}, \hat{\beta}_0 - t_{n-2,\frac{\alpha}{2}}\sqrt{\widehat{Var(\hat{\beta}_0)}}\right],$$

onde $t_{n-2,\frac{\alpha}{2}}$ é o quantil t-Student com $n-2$ graus de liberdade dado por $\mathbb{P}(t_{n-2} > t_{n-2,\frac{\alpha}{2}}) = \frac{\alpha}{2}$ e $\widehat{Var(\hat{\beta}_0)} = \dfrac{\hat{\sigma}_\varepsilon^2 \sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \bar{x})^2}$.

Demonstração: Decorrem do Teorema 1.1 e do Corolário 1.2 as seguintes afirmações.

(i) $Z = \dfrac{\hat{\beta}_0 - \beta_0}{\sqrt{Var(\hat{\beta}_0)}} = \dfrac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{\sigma_\varepsilon^2 \sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i-\bar{x})^2}}}$ é uma v.a. com função distribuição normal padrão.

(ii) $U = \dfrac{(n-2)\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2}$ é uma v.a. com distribuição qui-quadrado com $n-2$ graus de liberdade.

(iii) Z e U são v.a.'s independentes.

Desta forma a v.a. $T = \dfrac{Z}{\sqrt{\frac{U}{n-2}}}$ é uma v.a. com distribuição t-Student com $n-2$ graus de liberdade e pode ser utilizada como pivô.

Observe que

$$\begin{aligned}
T &= Z\sqrt{\frac{n-2}{U}} \\
&= \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{\sigma_\varepsilon^2 \sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i-\bar{x})^2}}} \times \sqrt{\frac{n-2}{\frac{(n-2)\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2}}} \\
&= \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{\hat{\sigma}_\varepsilon^2 \sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i-\bar{x})^2}}} \\
&= \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_\varepsilon}\sqrt{\frac{n\sum_{i=1}^n (x_i-\bar{x})^2}{\sum_{i=1}^n x_i^2}}. \qquad (1.7)
\end{aligned}$$

Segue-se da Observação 1.1 e do resultado obtido na expressão (1.7) que

$$T = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\widehat{Var(\hat{\beta}_0)}}}.$$

Assim, tem-se o que segue

$$\begin{aligned}
\alpha &= \mathbb{P}\left[-t_{n-2,\frac{\alpha}{2}} \leq T \leq t_{n-2,\frac{\alpha}{2}}\right] \\
&= \mathbb{P}\left[-t_{n-2,\frac{\alpha}{2}} \leq \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\widehat{Var(\hat{\beta}_0)}}} \leq t_{n-2,\frac{\alpha}{2}}\right] \\
&= \mathbb{P}\left[\hat{\beta}_0 - t_{n-2,\frac{\alpha}{2}}\sqrt{\widehat{Var(\hat{\beta}_0)}} \leq \beta_0 \leq \hat{\beta}_0 + t_{n-2,\frac{\alpha}{2}}\sqrt{\widehat{Var(\hat{\beta}_0)}}\right],
\end{aligned}$$

onde $\widehat{Var(\hat{\beta}_0)} = \dfrac{\hat{\sigma}^2_\varepsilon \sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \bar{x})^2}$. □

Proposição 1.9. *O intervalo a* $100(1-\alpha)\%$ *de confiança para* β_1 *é dado por*

$$\left[\hat{\beta}_1 - t_{n-2,\frac{\alpha}{2}}\sqrt{\widehat{Var(\hat{\beta}_1)}}, \hat{\beta}_1 + t_{n-2,\frac{\alpha}{2}}\sqrt{\widehat{Var(\hat{\beta}_1)}}\right],$$

onde $t_{n-2,\frac{\alpha}{2}}$ *é o quantil t-Student com* $n-2$ *graus de liberdade dado por* $\mathbb{P}(t_{n-2} > t_{n-2,\frac{\alpha}{2}}) = \frac{\alpha}{2}$ *e* $\widehat{Var(\hat{\beta}_1)} = \dfrac{\hat{\sigma}^2_\varepsilon}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

Demonstração: Decorrem do Teorema 1.1 e do Corolário 1.2 as seguintes afirmações.

(i) $Z = \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{Var(\hat{\beta}_1)}} = \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2_\varepsilon}{\sum_{i=1}^n (x_i - \bar{x})^2}}}$ é uma v.a. com distribuição normal padrão.

(ii) $U = \dfrac{(n-2)\hat{\sigma}^2_\varepsilon}{\sigma^2_\varepsilon}$ é uma v.a. qui-quadrado com $n-2$ graus de liberdade.

(iii) Z e U são v.a.'s independentes.

Desta forma a v.a. $T = \dfrac{Z}{\sqrt{\frac{U}{n-2}}}$ é uma v.a. com distribuição t-Student com $n-2$ graus de liberdade e pode ser utilizada como pivô.

Seguindo-se raciocínio análogo ao realizado na demonstração da Proposição 1.8, verifica-se que

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var(\hat{\beta}_1)}}}.$$

Assim, tem-se o que segue

$$\begin{aligned}
\alpha &= \mathbb{P}\left[-t_{n-2,\frac{\alpha}{2}} \leq T \leq t_{n-2,\frac{\alpha}{2}}\right] \\
&= \mathbb{P}\left[-t_{n-2,\frac{\alpha}{2}} \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var(\hat{\beta}_1)}}} \leq t_{n-2,\frac{\alpha}{2}}\right] \\
&= \mathbb{P}\left[\hat{\beta}_1 - t_{n-2,\frac{\alpha}{2}}\sqrt{\widehat{Var(\hat{\beta}_1)}} \leq \beta_1 \leq \hat{\beta}_1 + t_{n-2,\frac{\alpha}{2}}\sqrt{\widehat{Var(\hat{\beta}_1)}}\right],
\end{aligned}$$

onde $\widehat{Var(\hat{\beta}_1)} = \dfrac{\hat{\sigma}^2_\varepsilon}{\sum_{i=1}^n (x_i - \bar{x})^2}$. □

Proposição 1.10. *O intervalo a* $100(1-\alpha)\%$ *de confiança para* σ^2_ε *é dado por*

$$\left[\frac{(n-2)\hat{\sigma}^2_\varepsilon}{\chi^2_{n-2,1-\alpha}}, \frac{(n-2)\hat{\sigma}^2_\varepsilon}{\chi^2_{n-2,\alpha}}\right],$$

onde $\chi^2_{n-2,\alpha}$ *é o quantil qui-quadrado com* $n-2$ *graus de liberdade dado por* $\mathbb{P}(\chi^2_{n-2} > \chi^2_{n-2,\alpha}) = \frac{\alpha}{2}$.

Demonstração: Tem-se que $U = \frac{(n-2)\hat{\sigma}^2_\varepsilon}{\sigma^2_\varepsilon}$ é v.a. com distribuição qui-quadrado com $n-2$ graus de liberdade. A variável aleatória U então é o pivô utilizado. Segue que

$$\begin{aligned}
\alpha &= \mathbb{P}[\chi^2_{n-2,1-\alpha} \leq U \leq \chi^2_{n-2,\alpha}] \\
&= \mathbb{P}\left[\frac{(n-2)\hat{\sigma}^2_\varepsilon}{\chi^2_{n-2,1-\alpha}} \leq \sigma^2_\varepsilon \leq \frac{(n-2)\hat{\sigma}^2_\varepsilon}{\chi^2_{n-2,\alpha}}\right].
\end{aligned}$$

□

Proposição 1.11. *O intervalo a* $100(1-\alpha)\%$ *de confiança para* $\mu(x) = \beta_0 + \beta_1 x$ *é dado por*

$$\left[\hat{\mu}(x) - t_{n-2,\frac{\alpha}{2}}\sqrt{\widehat{Var[\hat{\mu}(x)]}}, \hat{\mu}(x) + t_{n-2,\frac{\alpha}{2}}\sqrt{\widehat{Var[\hat{\mu}(x)]}}\right],$$

onde $t_{n-2,\frac{\alpha}{2}}$ *é o quantil t-Student com* $n-2$ *graus de liberdade dado por* $\mathbb{P}(t_{n-2} > t_{n-2,\frac{\alpha}{2}}) = \frac{\alpha}{2}$ *e* $\widehat{Var[\hat{\mu}(x)]} = \widehat{Var(\hat{\beta}_0 + \hat{\beta}_1 x)}$.

Demonstração: Decorre da definição de $\mu(x)$ e do Corolário 1.3 os itens a seguir

(i) $Z = \dfrac{\hat{\mu}(x) - \mu(x)}{\sqrt{Var[\hat{\mu}(x)]}} = \dfrac{\hat{\mu}(x) - \mu(x)}{\sqrt{\sigma^2_\varepsilon\left[\frac{1}{n} + \frac{(\bar{x}-x)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]}}$ tem distribuição normal padrão.

(ii) $U = \dfrac{(n-2)\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2}$ tem distribuição qui-quadrado com $n-2$ graus de liberdade.

(iii) U e Z são independentes.

(iv) $T = \dfrac{Z}{\sqrt{\frac{U}{n-2}}}$ tem distribuição t-Student com $n-2$ graus de liberdade.

A variável aleatória T pode ser utilizada como pivô e é apresentada como

$$\begin{aligned} T &= Z\sqrt{\frac{n-2}{U}} \\ &= \frac{\hat{\mu}(x) - \mu(x)}{\sqrt{\sigma_\varepsilon^2\left[\frac{1}{n} + \frac{(\bar{x}-x)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]}} \times \sqrt{\frac{n-2}{\frac{(n-2)\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2}}} \\ &= \frac{\hat{\mu}(x) - \mu(x)}{\sqrt{\hat{\sigma}_\varepsilon^2\left[\frac{1}{n} + \frac{(\bar{x}-x)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]}}. \end{aligned} \tag{1.8}$$

Segue-se da Observação 1.1 e do resultado obtido na expressão (1.8) que

$$T = \frac{\hat{\mu}(x) - \mu(x)}{\sqrt{\widehat{Var[\hat{\mu}(x)]}}}.$$

Assim, tem-se o que segue

$$\begin{aligned} \alpha &= \mathbb{P}\left[-t_{n-2,\frac{\alpha}{2}} \leq T \leq +t_{n-2,\frac{\alpha}{2}}\right] = \mathbb{P}\left[-t_{n-2,\frac{\alpha}{2}} \leq \frac{\hat{\mu}(x) - \mu(x)}{\sqrt{\widehat{Var[\hat{\mu}(x)]}}} \leq t_{n-2,\frac{\alpha}{2}}\right] \\ &= \mathbb{P}\left[\hat{\mu}(x) - t_{n-2,\frac{\alpha}{2}}\sqrt{\widehat{Var[\hat{\mu}(x)]}} \leq \mu(x) \leq \hat{\mu}(x) + t_{n-2,\frac{\alpha}{2}}\sqrt{\widehat{Var[\hat{\mu}(x)]}}\right]. \end{aligned}$$

□

Outro procedimento utilizado em um modelo linear do Caso A é o Teste de Hipótese. Primeiramente, definem-se duas situações de teste,

$$\begin{aligned} &(i) \quad \mathcal{H}_0^0 : \beta_0 = b_0 \text{ vs } \mathcal{H}_1^0 : \beta_0 \neq b_0, \text{ para } b_0 \in \mathbb{R}. \\ &(ii) \quad \mathcal{H}_0^1 : \beta_1 = b_1 \text{ vs } \mathcal{H}_1^1 : \beta_1 \neq b_1, \text{ para } b_1 \in \mathbb{R}. \end{aligned} \tag{1.9}$$

Teorema 1.12. *No modelo linear dado pela Definição 1.1, o teste da razão de verossimilhança de tamanho α dado pela expressão (ii) de (1.9) consiste em rejeitar $\mathcal{H}_0^1$ se e somente se o intervalo de confiança dado por*

$$\left[\hat{\beta}_1 - t_{n-2,\frac{\alpha}{2}}\sqrt{\widehat{Var(\hat{\beta}_1)}}, \hat{\beta}_1 + t_{n-2,\frac{\alpha}{2}}\sqrt{\widehat{Var(\hat{\beta}_1)}}\right]$$

não contém b_1*, onde* $\widehat{Var(\hat{\beta}_1)} = \dfrac{\hat{\sigma}^2_\varepsilon}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

Demonstração: Assume-se que se deseja testar $\mathcal{H}_0^1 : \beta_1 = b_1$ vs $\mathcal{H}_1^1 : \beta_1 \neq b_1$. Uma estatística de teste que pode ser escolhida é

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var(\hat{\beta}_1)}}} = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}^2_\varepsilon} \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n x_i^2}}.$$

Segue da demonstração da Proposição 1.8 que, sob $\mathcal{H}_0^1$, a v.a. T tem distribuição t-Student com $n-2$ graus de liberdade. Um teste de tamanho α é dado por: rejeita-se $\mathcal{H}_0^1$ se e somente se $|T| > t_{n-2,\frac{\alpha}{2}}$. Comparando-se esta última expressão com o resultado obtido na Proposição 1.8, tem-se que ao se tomar o conjunto delimitado pelo intervalo de confiança de nível α para o parâmetro β_1, rejeita-se $\mathcal{H}_0^1$ se e somente se o intervalo de confiança não contém b_1. Agora, tem-se que mostrar que este é um teste de razão de verossimilhança. Para tal, primeiramente, definem-se os espaços de parâmetros Θ, Θ_0 e Θ_1, onde $\mathcal{B} = (\beta_0, \beta_1, \sigma^2_\varepsilon)$, dados por

$$\Theta = \Theta_0 \cup \Theta_1,$$

onde

$$\Theta = \{(\beta_0, \beta_1, \sigma^2_\varepsilon) \mid \beta_0, \beta_1 \in \mathbb{R}, \sigma^2_\varepsilon > 0\} \text{ e}$$
$$\Theta_0 = \{(\beta_0, \beta_1, \sigma^2_\varepsilon) \mid \beta_0 \in \mathbb{R}, \beta_1 = b_1, \sigma^2_\varepsilon > 0\}.$$

Denota-se λ como

$$\lambda = \frac{\sup_{\mathcal{B}\in\Theta_0} L(\mathcal{B}; y_1, \cdots, y_n)}{\sup_{\mathcal{B}\in\Theta} L(\mathcal{B}; y_1, \cdots, y_n)},$$

onde $L(\mathcal{B}; y_1, \cdots, y_n)$ é dada pela equação (1.3) e os valores de β_0, β_1 e σ^2_ε que a tornam máxima para $\mathcal{B} \in \Theta$ são os estimadores de máxima verossimilhança apresentados no Teorema 1.1. Assume-se, sem perda de generalidade, que $\beta_1 = 0$, dessa forma, tem-se que

$$\sup_{\mathcal{B}\in\Theta_0} L(\mathcal{B}; y_1, \cdots, y_n) = \left(\frac{1}{2\pi\tilde{\sigma}^2_\varepsilon}\right)^{\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{2\tilde{\sigma}^2_\varepsilon}\right], \tag{1.10}$$

onde $\tilde{\sigma}^2_\varepsilon = \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i^2$ e $\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. Substitui-se $\beta_1 = 0$ na expressão (1.10), obtendo-se

$$L(\beta_0, \sigma_\varepsilon^2) = \frac{1}{(2\pi\sigma_\varepsilon^2)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i - \beta_0)^2\right].$$

Os valores de β_0 e σ_ε^2 que maximizam a função de verossimilhança são os estimadores de máxima verossimilhança dados por

$$\begin{aligned} \beta_0^* &= \bar{y} \\ \sigma_\varepsilon^{*2} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2. \end{aligned} \tag{1.11}$$

Então, tem-se

$$\sup_{\mathcal{B}\in\Theta_0} L(\mathcal{B}; y_1, \cdots, y_n) = \frac{1}{(2\pi\sigma_\varepsilon^{*2})^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma_\varepsilon^{*2}}\sum_{i=1}^{n}(y_i - \beta_0^*)^2\right]. \tag{1.12}$$

Segue-se da equação (1.12) que

$$\lambda = \left(\frac{\tilde{\sigma}_\varepsilon^2}{\sigma_\varepsilon^{*2}}\right)^{\frac{n}{2}}.$$

A partir da expressão para λ obtida acima, examina-se, com base na expressão (1.11), a quantidade $\lambda^{-\frac{2}{n}} - 1$, a qual é uma função monótona em λ e é dada por

$$\begin{aligned} \lambda^{-\frac{2}{n}} - 1 &= \frac{\sigma_\varepsilon^{*2} - \tilde{\sigma}_\varepsilon^2}{\tilde{\sigma}_\varepsilon^2} \\ &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}. \end{aligned} \tag{1.13}$$

A expressão (1.13) assume a seguinte forma

$$\lambda^{-\frac{2}{n}} - 1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}[(y_i - \bar{y}) - \hat{\beta}_1(x_i - \bar{x})]^2}{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}. \tag{1.14}$$

Desenvolvendo-se, na equação (1.14), o quadrado que se encontra entre colchetes, resulta em

$$\lambda^{-\frac{2}{n}} - 1 = \frac{2\hat{\beta}_1\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) - \hat{\beta}_1\sum_{i=1}^{n}\hat{\beta}_1(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2},$$

e, assim segue-se

$$\lambda^{-\frac{2}{n}} - 1 = \frac{\hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}. \tag{1.15}$$

Multiplicando-se ambos os lados da igualdade (1.15) por $n-2$, pode-se obter

$$\begin{aligned}(n-2)[\lambda^{-\frac{2}{n}} - 1] &= \frac{\hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})^2}{\hat{\sigma}_\varepsilon^2} \\ &= \frac{\hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})^2 / \sigma_\varepsilon^2}{\hat{\sigma}_\varepsilon^2 / \sigma_\varepsilon^2} \equiv T^2. \end{aligned} \tag{1.16}$$

A equação (1.16), sob a hipótese $\mathcal{H}_0^1 : \beta_1 = 0$, é a razão entre os valores de duas v.a.'s independentes com distribuição qui-quadrado divididas, respectivamente, pelos seus graus de liberdade; 1 para o numerador e $n-2$ para o denominador. Portanto, a penúltina igualdade da equação (1.16) é uma v.a. com distribuição F-Snedecor com 1 e $n-2$ graus de liberdade.

Seja Λ uma variável aleatória que assume todos os possíveis valores de λ. Desta forma, a v.a. definida por $(n-2)[\Lambda^{-\frac{2}{n}} - 1]$ possui, por definição, distribuição F-Snedecor com 1 e $n-2$ graus de liberdade, sob $\mathcal{H}_0^1$. O teste de razão de verossimilhança rejeita $\mathcal{H}_0^1$ se e somente se $\lambda \leq \lambda_0$ ou então, se e somente se $(n-2)[\lambda^{-\frac{2}{n}} - 1] \geq (n-2)[{\lambda_0}^{-\frac{2}{n}} - 1] = {\lambda_0}^*$; ou ainda, se e somente se,

$$\left[\frac{\hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\hat{\sigma}_\varepsilon^2}\right] \geq {\lambda_0}^*,$$

onde ${\lambda_0}^*$ é escolhido de forma a atender o tamanho esperado do Erro Tipo I. □

Teorema 1.13. *No modelo linear dado pela Definição 1.1, o teste da razão de verossimilhança de tamanho α dado pela expressão (i) de (1.9) consiste em rejeitar $\mathcal{H}_0^0$ se e somente se o intervalo de confiança dado por*

$$\left[\hat{\beta}_0 - t_{n-2,\frac{\alpha}{2}} \sqrt{\widehat{Var(\hat{\beta}_0)}}, \hat{\beta}_0 - t_{n-2,\frac{\alpha}{2}} \sqrt{\widehat{Var(\hat{\beta}_0)}}\right]$$

não contém b_0, onde $\widehat{Var(\hat{\beta}_0)} = \frac{\hat{\sigma}_\varepsilon^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}$.

Demonstração: A demonstração segue raciocínio análogo ao realizado no Teorema 1.12, utilizando-se o intervalo de confiança obtido na Proposição 1.8. □

Caso B:

Na seqüência, apresentam-se os procedimentos de inferência para o Caso B. Assume-se que as v.a.s Y_i e Y_j sejam não-correlacionadas para quaisquer pares (i, j) tais que $1 \leq i, j \leq n$. Para esta situação, supõe-se que $Y_1, Y_2, \cdots, Y_n$ são v.a.s com média $\beta_0 + \beta_1 x_i$ e variância σ^2_ε, para $1 \leq i \leq n$. Se a função densidade de probabilidade (f.d.p.) conjunta das v.a.s Y_i para $i \in \{1, 2, \cdots, n\}$ não é especificada então os estimadores de máxima verossimilhança não podem ser obtidos para os parâmetros β_0 , β_1 e σ^2_ε, como feito no Caso A. Em situações em que a f.d.p. não é dada ou conhecida, o método de estimação adequado é o dos Mínimos Quadrados.

Definição 1.2. Sejam (y_i, x_i), com $i \in \{1, 2, \cdots, n\}$, n pares de observações que satisfazem o modelo linear apresentado na Definição 1.1. Os valores de β_0 e β_1 que minimizam a soma dos quadrados $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ são definidos como os *estimadores dos mínimos quadrados* de β_0 e β_1. No contexto da notação matricial, tem-se que o Método dos Mínimos Quadrados seleciona $\hat{\mathcal{B}}$ que minimiza a soma dos quadrados das diferenças $(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})$.

O procedimento para encontrar tais estimadores é apresentado pelo teorema a seguir.

Teorema 1.14. *No Caso B e no contexto da Definição 1.2 o estimador dos mínimos quadrados de $\mathcal{B}$, denotado por $\hat{\mathcal{B}}$, é dado por $\hat{\mathcal{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.*

Demonstração: Para se obter o estimador de $\mathcal{B}$, deve-se encontrar os valores que minimizam a expressão $(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})$.

Para tal, usa-se do fato que

$$\begin{aligned} \mathbf{Y} - \mathbf{X}\mathcal{B} &= \mathbf{Y} - \mathbf{X}\hat{\mathcal{B}} + \mathbf{X}\hat{\mathcal{B}} - \mathbf{X}\mathcal{B} \\ &= \mathbf{Y} - \mathbf{X}\hat{\mathcal{B}} + \mathbf{X}(\hat{\mathcal{B}} - \mathcal{B}). \end{aligned}$$

Dessa forma,

$$\begin{aligned} (\mathbf{Y} - \mathbf{X}\mathcal{B})'(\mathbf{Y} - \mathbf{X}\mathcal{B}) &= (\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}}) + (\hat{\mathcal{B}} - \mathcal{B})'\mathbf{X}'\mathbf{X}(\hat{\mathcal{B}} - \mathcal{B}) + \\ &\quad 2(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})'\mathbf{X}(\hat{\mathcal{B}} - \mathcal{B}) \\ &= (\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}}) + (\hat{\mathcal{B}} - \mathcal{B})\mathbf{X}'\mathbf{X}(\hat{\mathcal{B}} - \mathcal{B}), \quad (1.17) \end{aligned}$$

A primeira parcela de (1.17) não depende de $\mathcal{B}$ e a segunda é o quadrado de $\mathbf{X}(\hat{\mathcal{B}} - \mathcal{B})$. Como se assume que $\mathbf{X}$ tenha posto completo, ocorre que $\mathbf{X}(\hat{\mathcal{B}} - \mathcal{B}) \neq \mathbf{0}$

caso $\hat{\mathcal{B}} \neq \mathcal{B}$. Dessa forma, o mínimo da soma dos quadrados existe e é unico para $\hat{\mathcal{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. □

Proposição 1.15. *Para o estimador dos mínimos quadrados $\hat{\mathcal{B}}$, tem-se que suas respectivas esperança e variância são dadas pela expressão (1.6).*

Demonstração: Este resultado é conseqüência dos resultados do Teorema 1.1.

Para o Caso A, os estimadores de máxima verossimilhança encontrados, além de serem não viciados, apresentam a propriedade de uniforme mínima variância. Esta característica não permanece válida para os estimadores do Caso B, obtidos pelo método dos mínimos quadrados. Tal fato ocorre porque no Caso A as hipóteses referentes às v.a.'s são mais fortes que no Caso B, onde as v.a.'s Y_i, para $i \in \{1, \cdots, n\}$, tem f.d.p. desconhecida. Para o contexto dos estimadores obtidos pelo método dos mínimos quadrados uma propriedade que representa uma espécie de mínima variância será definida a seguir.

Definição 1.3. Sejam $Y_1, Y_2, \cdots, Y_n$ v.a.'s observadas tais que $\mathbb{E}(Y_i) = h_i(\theta)$, onde $h_i(\cdot)$ são funções conhecidas definidas para o parâmetro desconhecido $\theta = (\theta_1, \theta_2, \cdots, \theta_s)$. Para estimar qualquer θ_j com $j \in \{1, \cdots, s\}$, considera-se apenas a classe dos estimadores que são funções lineares das v.a.'s Y_i. Nesta classe, apenas alguns estimadores são não viciados para θ_j e existe um estimador de θ_j que apresenta a menor variância em comparação com qualquer outro estimador de θ_j nesta mesma classe. Este estimador é definido como o *melhor estimador linear não viciado* (BLUE) para θ_j.

Nesta definição o termo *melhor* se refere a propriedade de mínima variância. É interessante notar que existem dois fatos a serem observados antes da propriedade de mínima variância. Primeiramente, as funções $h_i(\cdot)$ devem ser lineares em θ, e, posteriormente, selecionam-se apenas aquelas que fornecem estimadores não viciados. Finalmente, então, investiga-se qual estimador apresenta a propriedade de mínima variância.

Na seqüência, apresenta-se o Teorema de Gauss-Markov que assegura que os estimadores obtidos pelo método dos mínimos quadrados são os melhores estimadores lineares não viciados.

Teorema 1.16. *Considere o modelo linear apresentado na Definição 1.1 e as hipóteses referentes ao Caso B. Então, o estimador dos mínimos quadrados $\hat{\mathcal{B}}$ é o melhor estimador linear não-viciado para $\mathcal{B}$.*

Demonstração: Para qualquer $\mathbf{c} \in \mathbb{R}^2$ fixo, seja $a'\mathbf{Y}$ um estimador não viciado de $\mathbf{c}'\mathcal{B}$. Então, $\mathbb{E}(\mathbf{a}'\mathbf{Y}) = \mathbf{c}'\mathcal{B}$, para qualquer que seja $\mathcal{B}$. Tem-se ainda, por hipótese, que $\mathbb{E}(\mathbf{a}'\mathbf{Y}) = \mathbb{E}(\mathbf{a}'\mathbf{X}\mathcal{B}+\mathbf{a}'\mathcal{E}) = \mathbf{a}'\mathbf{X}\mathcal{B}$. Igualando-se estas duas expressões obtidas para esperança, tem-se que $(\mathbf{c}' - \mathbf{a}'\mathbf{X})\mathcal{B} = \mathbf{0}$ para qualquer $\mathcal{B}$, inclusive para $\mathcal{B} = (\mathbf{c}' - \mathbf{a}'\mathbf{X})'$. O que acarreta que $\mathbf{c}' = \mathbf{a}'\mathbf{X}$.

Define-se $\mathbf{a}^* = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}$, dessa forma, $\mathbf{c}'\hat{\mathcal{B}} = \mathbf{a}^{*\prime}\mathbf{Y}$ é estimador não viciado. Então para qualquer $\mathbf{a}$ satisfazendo $\mathbf{c}' = \mathbf{a}'\mathbf{X}$, tem-se que

$$\begin{aligned} Var(\mathbf{a}'\mathbf{Y}) &= Var(\mathbf{a}'\mathbf{X}\mathcal{B} + \mathbf{a}'\mathcal{E}) \\ &= Var(\mathbf{a}'\mathcal{E}) \\ &= \sigma_\varepsilon^2 \mathbf{a}'\mathbf{a} \\ &= \sigma_\varepsilon^2 (\mathbf{a} - \mathbf{a}^* + \mathbf{a}^*)'(\mathbf{a} - \mathbf{a}^* + \mathbf{a}^*) \\ &= \sigma_\varepsilon^2 [(\mathbf{a} - \mathbf{a}^*)'(\mathbf{a} - \mathbf{a}^*) + \mathbf{a}^{*\prime}\mathbf{a}^*]. \end{aligned}$$

Decorre da definição de $\mathbf{a}^*$ que $\mathbf{a} - \mathbf{a}^{*\prime}\mathbf{a}^* = 0$. Como $\mathbf{a}^*$ é fixo e o termo $(\mathbf{a} - \mathbf{a}^*)'(\mathbf{a} - \mathbf{a}^*)$ é positivo para $\mathbf{a} \neq \mathbf{a}^*$, segue que $Var(\mathbf{a}'\mathbf{Y})$ é minimizada pela escolha $\mathbf{a}'\mathbf{Y} = \mathbf{c}'\mathcal{B}$. □

A seguir, inicia-se a uma discussão análoga à apresentada nesta seção para o contexto da regressão linear múltipla.

1.2 Regressão Linear Múltipla

O modelo de regressão linear simples apresenta suas limitações. Em algumas situações é necessário analisar a influência de várias váriaveis independentes sobre uma variável dependente. Este é o modelo de regressão linear múltipla. Para se fazer a extensão do modelo apresentado na Definição 1.1 ao modelo de regressão linear múltipla, primeiramente apresenta-se um exemplo.

Exemplo 1.3. A porcentagem de uma substância química absorvida por uma planta depende, não apenas da quantidade desta substância que é administrada ao solo, mas também da quantidade de água que a planta recebe. Um pesquisador pode selecionar uma variedade de combinações entre quantidade de determinada substância fornecida ao solo e quantidade de água que a planta recebe e observar as concentrações apresentadas pela planta em cada caso. Assumem-se que as médias

das concentrações da substância química na planta variam linearmente como função da quantidade de água e da quantidade de substância química disponíveis na planta. Assim, para uma amostra de n plantas, tem-se observações y_i, na forma, $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, com $i = 1, \cdots, n$. Onde y_i é a concentração da substância química na planta i, x_{i1} é a quantidade de substância química oferecida e x_{i2} é a quantidade de água administrada à planta i. Os valores β_i são os parâmetros desconhecidos da relação linear. Supõe-se que as v.a.'s ε_i sejam independentes e identicamente distribuídas com média zero.

O Exemplo 1.3 motiva o seguinte comentário. De forma mais específica, o modelo de regressão linear geral com r variáveis independentes e uma v.a. Y dependente toma a seguinte forma

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + \varepsilon. \tag{1.18}$$

Assumindo n independentes observações de Y associadas aos valores de x_j, para $j \in \{1, \cdots, r\}$, o modelo (1.18) apresenta-se como

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_r x_{1r} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \cdots + \beta_r x_{2r} + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \cdots + \beta_r x_{nr} + \varepsilon_n, \end{aligned} \tag{1.19}$$

onde os termos dos erros, ε_i, seguem as seguintes suposições, para $i \in \{1, \cdots, n\}$

$$\begin{aligned} &(i) \quad \mathbb{E}(\varepsilon_i) = 0. \\ &(ii) \quad Var(\varepsilon_i) = \sigma_\varepsilon^2. \\ &(iii) \quad Cov(\varepsilon_i, \varepsilon_l) = 0, \text{ se } i \neq l. \end{aligned} \tag{1.20}$$

Em notação matricial, o sistema (1.19) apresenta-se como

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1r} \\ 1 & x_{21} & x_{22} & \cdots & x_{2r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nr} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

ou ainda,

$$\mathbf{Y} = \mathbf{X}\mathcal{B} + \mathcal{E}, \tag{1.21}$$

onde, $\mathbf{Y}$ é uma matriz $n \times 1$, $\mathbf{X}$ é uma matriz $n \times (r+1)$, $\mathcal{B}$ é uma matriz $(r+1) \times 1$ e $\mathcal{E}$ é uma matriz $n \times 1$. Note que a primeira coluna da matriz $\mathbf{X}$ é composta por um fator constante. Pode-se introduzir uma variável artificial (na realidade é uma constante), $x_{i0} = 1$, para qualquer $i \in \{1, \cdots, n\}$, de modo que

$$y = \beta_0 x_0 + \beta_1 x_1 + \cdots + \beta_r x_r + \varepsilon,$$

sendo que então, cada coluna da matriz $\mathbf{X}$ é composta de n valores da variável independente correspondente enquanto que a i-ésima linha de $\mathbf{X}$ corresponde aos valores de todas as variáveis independentes na amostra i.

As suposições apresentadas em (1.20), em sua forma matricial são

$$(i) \quad \mathbb{E}(\mathcal{B}) = \mathbf{0}.$$
$$(ii) \quad Cov(\mathcal{B}) = \mathbb{E}(\mathcal{B}\mathcal{B}') = \sigma_\varepsilon^2 \mathbf{I}.$$

Tais comentários motivam a seguinte definição.

Definição 1.4. As v.a.'s $Y_1, \cdots, Y_n$ satisfazem um *modelo linear geral* se uma amostra de tamanho um de cada Y_i pode ser expressa como

$$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \cdots + \beta_r x_{ir} + \varepsilon_i, \tag{1.22}$$

onde x_{ij} são constantes conhecidas, β_j são parâmetros desconhecidos do modelo e ε_i são v.a.'s independentes, igualmente distribuídas com média zero e variância σ_ε^2.

Observação 1.3. É interessante comentar que da mesma forma que exposto na seção anterior, é necessário se fazer, nesta seção, distinção entre Caso A e Caso B, no que se refere ao conhecimento ou não da função de distribuição das v.a.'s $Y_1, \cdots, Y_n$.

Vamos inicialmente abordar a situação múltipla análoga ao Caso B.

Caso B:

Como feito na seção precedente, procuram-se estimadores para os parâmetros desconhecidos na equação matricial (1.21), representados por $\mathcal{B}$ através de uma generalização da Definição 1.2 que é apresentada a seguir.

Definição 1.5. Os valores $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_r$ que fornecem o valor mínimo para a expressão, $SQRes(\mathcal{B}) \equiv SQRes(\beta_0, \beta_1, \cdots, \beta_r)$ que é dada por

$$SQRes(\mathcal{B}) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 x_{i0} - \beta_1 x_{i1} - \cdots - \beta_r x_{ir})^2,$$

são chamados de *estimadores dos mínimos quadrados* dos parâmetros de regressão $\beta_0, \beta_1, \cdots, \beta_r$.

Observação 1.4. A Definição 1.5 poderia ser apresentada em termos matriciais, ficando, de forma análoga, definido o estimador $\hat{\mathcal{B}}$ para o parâmetro $\mathcal{B}$. Ou melhor,

$$SQRes(\mathcal{B}) = (\mathbf{Y} - \mathbf{X}\mathcal{B})'(\mathbf{Y} - \mathbf{X}\mathcal{B}),$$

e dessa maneira, o vetor de resíduos é representado por

$$\hat{\mathcal{E}} = \mathbf{Y} - \mathbf{X}\hat{\mathcal{B}}.$$

O procedimento para encontrar tais estimadores é descrito a seguir. Usa-se a forma matricial por simplicidade de notação.

Teorema 1.17. *Seja* $\mathbf{X}$ *uma matriz de posto completo* $r+1 \leq n$. *O estimador dos mínimos quadrados de* $\mathcal{B}$ *é dado por*

$$\hat{\mathcal{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{1.23}$$

Seja $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathcal{B}} = \mathbf{H}\mathbf{Y}$ *a representação dos valores ajustados de* $\mathbf{Y}$, *onde* $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ *é dita matriz chapéu, então os resíduos*

$$\hat{\mathcal{E}} = \mathbf{Y} - \hat{\mathbf{Y}} = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

satisfazem $\mathbf{X}'\hat{\mathcal{B}} = \mathbf{0}$ *e* $\hat{\mathbf{Y}}'\hat{\mathcal{E}} = \mathbf{0}$. *E ainda, a soma dos quadrados dos resíduos é dada por*

$$\begin{aligned}\hat{\mathcal{E}}'\hat{\mathcal{E}} &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 x_{i0} - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_r x_{ir})^2 \\ &= \mathbf{Y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\mathcal{B}}.\end{aligned} \tag{1.24}$$

Demonstração: Seja $\hat{\mathcal{B}}$ como definido acima. Então

$$\hat{\mathcal{E}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\mathcal{B}} = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}.$$

A matriz $[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']$ é simétrica, idempotente e $\mathbf{X}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \mathbf{X}' - \mathbf{X}' = \mathbf{0}$. Conseqüentemente, $\mathbf{X}'\hat{\mathcal{E}} = \mathbf{X}'(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{0}$, então $\hat{\mathbf{Y}}'\hat{\mathcal{E}} = \hat{\mathcal{B}}'\mathbf{X}'\hat{\mathcal{E}} = \mathbf{0}$.

Pela idempotência de $[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']$, segue-se a equação (1.24).

Para se verificar a expressão (1.23), escreve-se

$$\mathbf{Y} - \mathbf{X}\mathcal{B} = \mathbf{Y} - \mathbf{X}\hat{\mathcal{B}} + \mathbf{X}\hat{\mathcal{B}} - \mathbf{X}\mathcal{B} = \mathbf{Y} - \mathbf{X}\hat{\mathcal{B}} + \mathbf{X}(\hat{\mathcal{B}} - \mathcal{B}),$$

o que acarreta, pelos comentários da Observação 1.4,

$$\begin{aligned} SQRes(\mathcal{B}) &= (\mathbf{Y}-\mathbf{X}\mathcal{B})'(\mathbf{Y}-\mathbf{X}\mathcal{B}) \\ &= (\mathbf{Y}-\mathbf{X}\hat{\mathcal{B}})'(\mathbf{Y}-\mathbf{X}\hat{\mathcal{B}}) + (\hat{\mathcal{B}}-\mathcal{B})'\mathbf{X}'\mathbf{X}(\hat{\mathcal{B}}-\mathcal{B}) + \\ &\quad 2(\mathbf{Y}-\mathbf{X}\hat{\mathcal{B}})'\mathbf{X}(\hat{\mathcal{B}}-\mathcal{B}) \\ &= (\mathbf{Y}-\mathbf{X}\hat{\mathcal{B}})'(\mathbf{Y}-\mathbf{X}\hat{\mathcal{B}}) + (\hat{\mathcal{B}}-\mathcal{B})'\mathbf{X}'\mathbf{X}(\hat{\mathcal{B}}-\mathcal{B}), \qquad (1.25) \end{aligned}$$

desde que $(\mathbf{Y}-\mathbf{X}\hat{\mathcal{B}})'\mathbf{X} = \hat{\mathcal{E}}'\mathbf{X} = \mathbf{0}'$. O primeiro termo da expressão (1.25) não depende de $\mathcal{B}$, e o segundo termo é $\mathbf{X}(\hat{\mathcal{B}}-\mathcal{B})$ ao quadrado. Como, por hipótese, $\mathbf{X}$ é de posto completo, então $\mathbf{X}(\hat{\mathcal{B}}-\mathcal{B}) \neq \mathbf{0}$ caso $\mathcal{B} \neq \hat{\mathcal{B}}$. Dessa forma o mínimo da soma dos quadrados é único e ocorre para $\hat{\mathcal{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. □

Observação 1.5. Note que a inversa $(\mathbf{X}'\mathbf{X})^{-1}$ existe desde que a matriz $\mathbf{X}'\mathbf{X}$ tenha posto $r+1 \leq n$.

O estimador $\hat{\mathcal{B}}$ e o resíduo $\hat{\mathcal{E}}$ têm algumas propriedades que são apresentadas na proposição a seguir.

Proposição 1.18. *Seja o modelo linear dado pela Definição 1.4. O estimador dos mínimos quadrados de $\mathcal{B}$, denotado por $\hat{\mathcal{B}}$ e dado na equação (1.23), apresenta*

$$\mathbb{E}(\hat{\mathcal{B}}) = \mathcal{B} \quad e \quad Var(\hat{\mathcal{B}}) = \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}.$$

O resíduo $\hat{\mathcal{E}}$ apresenta

$$\mathbb{E}(\hat{\mathcal{E}}) = \mathbf{0}, \qquad Var(\hat{\mathcal{E}}) = \sigma_\varepsilon^2[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \sigma_\varepsilon^2(\mathbf{I}-\mathbf{H})$$
$$e \quad \mathbb{E}(\hat{\mathcal{E}}'\hat{\mathcal{E}}) = (n-r-1)\sigma_\varepsilon^2.$$

E ainda, definindo-se,

$$\hat{\sigma}_\varepsilon^2 \equiv \frac{\hat{\mathcal{E}}'\hat{\mathcal{E}}}{n-(r+1)} = \frac{\mathbf{X}'(\mathbf{I}-\mathbf{H})\mathbf{X}}{n-r-1},$$

tem-se que $\mathbb{E}(\hat{\sigma}_\varepsilon^2) = \sigma_\varepsilon^2$.

Demonstração: Sabe-se que $\mathbf{Y} = \mathbf{X}\mathcal{B}+\mathcal{E}$ e, dessa forma, $\hat{\mathcal{B}}$ pode ser representado por

$$\begin{aligned} \hat{\mathcal{B}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\mathcal{B}+\mathcal{E}) \\ &= \mathcal{B} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathcal{E}, \end{aligned}$$

e, $\hat{\mathcal{E}}$, pode ser apresentado por

$$\begin{aligned}\hat{\mathcal{E}} &= [\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}]'\mathbf{Y} = [\mathbf{I}-\mathbf{X}('\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][\mathbf{X}\mathcal{B}+\mathcal{E}] \\ &= [\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathcal{E}, \end{aligned} \tag{1.26}$$

visto que $[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{X} = \mathbf{X}-\mathbf{X} = \mathbf{0}$.

Com base nas propriedades de esperança matemática e de variância, segue-se que

$$\begin{aligned}\mathbb{E}(\hat{\mathcal{B}}) &= \mathcal{B}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}(\mathcal{E}) = \mathcal{B} \\ Var(\hat{\mathcal{B}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Var(\mathcal{E})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2_{\varepsilon}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2_{\varepsilon}(\mathbf{X}'\mathbf{X})^{-1},\end{aligned}$$

e ainda,

$$\begin{aligned}\mathbb{E}(\hat{\mathcal{E}}) &= [\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbb{E}(\mathcal{E}) = \mathbf{0} \\ Var(\hat{\mathcal{E}}) &= [\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']Var(\mathcal{E})[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\ &= \sigma^2_{\varepsilon}[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'].\end{aligned}$$

Agora, procura-se avaliar a $\mathbb{E}(\hat{\mathcal{E}}'\hat{\mathcal{E}})$. Com base na definição de $\hat{\mathcal{E}}$, tem-se que

$$\begin{aligned}\hat{\mathcal{E}}'\hat{\mathcal{E}} &= \mathcal{E}'[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}]'[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}]'\mathcal{E} \\ &= \mathcal{E}'[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathcal{E} \\ &= tr\{\mathcal{E}'(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathcal{E}\} \\ &= tr\{[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathcal{E}\mathcal{E}'\}. \end{aligned} \tag{1.27}$$

Aplicando-se esperança na igualdade obtida em (1.27), conclui-se que

$$\begin{aligned}\mathbb{E}(\hat{\mathcal{E}}'\hat{\mathcal{E}}) &= tr\{[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbb{E}(\mathcal{E}\mathcal{E}')\} \\ &= \sigma^2_{\varepsilon}tr[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \\ &= \sigma^2_{\varepsilon}tr(\mathbf{I}) - \sigma^2_{\varepsilon}tr[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \\ &= \sigma^2_{\varepsilon}n - \sigma^2_{\varepsilon}tr[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} \\ &= \sigma^2_{\varepsilon}n - \sigma^2_{\varepsilon}tr\left[\mathbf{I}_{(r+1)\times(r+1)}\right] \\ &= \sigma^2_{\varepsilon}(n-r-1). \end{aligned} \tag{1.28}$$

Da expressão obtida em (1.28) segue o resultado para $\hat{\sigma}^2_{\varepsilon}$. □

Assim, como visto na seção anterior, o estimador $\hat{\mathcal{B}}$ possui a propriedade de mínima variância estabelecida por Gauss e Markov. Agora apresentamos seu análogo para o contexto da regressão múltipla.

Teorema 1.19. *Seja* $\mathbf{Y} = \mathbf{X}\mathcal{B} + \mathcal{E}$, *onde* $\mathbb{E}(\mathcal{E}) = \mathbf{0}$ *e* $Var(\mathcal{E}) = \sigma_\varepsilon^2\mathbf{I}$, $\mathbf{X}$ *possui posto completo* $r+1$. *Para qualquer* $\mathbf{c} \in \mathbb{R}^r$, $\mathbf{c}'\hat{\mathcal{B}} = c_0\hat{\beta}_0 + \cdots + c_r\hat{\beta}_r$ *é o estimador de* $\mathbf{c}'\mathcal{B}$ *que tem a menor variância possível dentre todos os estimadores lineares na forma* $a'\mathbf{Y} = a_0y_0 + \cdots + a_ny_n$ *que são não viciados para* $\mathbf{c}'\mathcal{B}$.

Demonstração: Para qualquer $\mathbf{c} \in \mathbb{R}^r$ fixo, seja $a'\mathbf{Y}$ um estimador não viciado de $\mathbf{c}'\mathcal{B}$. Então, $\mathbb{E}(\mathbf{a}'\mathbf{Y}) = \mathbf{c}'\mathcal{B}$, para qualquer que seja $\mathcal{B}$. Tem-se ainda, por hipótese, que $\mathbb{E}(\mathbf{a}'\mathbf{Y}) = \mathbb{E}(\mathbf{a}'\mathbf{X}\mathcal{B}+\mathbf{a}'\mathcal{E}) = \mathbf{a}'\mathbf{X}\mathcal{B}$. Igualando-se estas duas expressões obtidas para esperança, tem-se que $(\mathbf{c}' - \mathbf{a}'\mathbf{X})\mathcal{B} = \mathbf{0}$ para qualquer $\mathcal{B}$, inclusive para $\mathcal{B} = (\mathbf{c}' - \mathbf{a}'\mathbf{X})'$. O que acarreta que $\mathbf{c}' = \mathbf{a}'\mathbf{X}$.

Define-se $\mathbf{a}^* = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}$, dessa forma, $\mathbf{c}'\hat{\mathcal{B}} = \mathbf{a}^{*\prime}\mathbf{Y}$ é estimador não viciado, pelo resultado da Proposição 1.18. Então para qualquer $\mathbf{a}$ satisfazendo $\mathbf{c}' = \mathbf{a}'\mathbf{X}$, tem-se que

$$\begin{aligned} Var(\mathbf{a}'\mathbf{Y}) &= Var(\mathbf{a}'\mathbf{X}\mathcal{B} + \mathbf{a}'\mathcal{E}) \\ &= Var(\mathbf{a}'\mathcal{E}) \\ &= \mathbf{a}'\mathbf{I}\sigma_\varepsilon^2\mathbf{a} \\ &= \sigma_\varepsilon^2(\mathbf{a} - \mathbf{a}^* + \mathbf{a}^*)'(\mathbf{a} - \mathbf{a}^* + \mathbf{a}^*) \\ &= \sigma_\varepsilon^2[(\mathbf{a} - \mathbf{a}^*)'(\mathbf{a} - \mathbf{a}^*) + \mathbf{a}^{*\prime}\mathbf{a}^*]. \end{aligned}$$

Decorre da definição de $\mathbf{a}^*$ que $\mathbf{a} - \mathbf{a}^{*\prime}\mathbf{a}^* = 0$. Como $\mathbf{a}^*$ é fixo e o termo $(\mathbf{a} - \mathbf{a}^*)'(\mathbf{a} - \mathbf{a}^*)$ é positivo para $\mathbf{a} \neq \mathbf{a}^*$, segue que $Var(\mathbf{a}'\mathbf{Y})$ é minimizada pela escolha $\mathbf{a}'\mathbf{Y} = \mathbf{c}'\mathcal{B}$. □

Observação 1.6. O resultado acima é muito poderoso, pois afirma que a substituição de $\mathcal{B}$ por $\hat{\mathcal{B}}$ na expressão $\mathbf{c}'\mathcal{B}$, para qualquer $\mathbf{c}$ que se tenha interesse, fornece o melhor estimador possível que é chamado o melhor estimador linear não-viciado (BLUE).

Para se descrever procedimentos de inferência, baseados no modelo apresentado na Definição 1.4, é necessário adicionar a hipótese de que a v.a. erro, $\mathcal{E}$, possui distribuição normal n-variada com $\mathbb{E}(\mathcal{E}) = \mathbf{0}$ e $Var(\mathcal{E}) = \sigma_\varepsilon^2\mathbf{I}$, denotada por $N_n(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I})$. Assim, é conseqüência da equação (1.21) que $\mathbb{E}(\mathbf{Y}|\mathbf{x}) = \beta_0x_0 + \cdots + \beta_rx_r$. Nesta situação, tem-se o análogo ao Caso A, visto na seção anterior.

Teorema 1.20. *Seja* $\mathbf{Y} = \mathbf{X}\mathcal{B} + \mathcal{E}$*, onde* $\mathbf{X}$ *tem posto completo* $r+1$ *e a v.a.* $\mathcal{E}$ *é distribuída como* $N_n(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I})$*. Então, o estimador de máxima verossimilhança de* $\hat{\mathcal{B}}$ *é o mesmo que o obtido pelo método dos mínimos quadrados. E,* $\hat{\mathcal{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ *tem distribuição* $N_{r+1}[\mathcal{B}, \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}]$ *e é independente dos resíduos* $\hat{\mathcal{E}} = \mathbf{Y} - \mathbf{X}\hat{\mathcal{B}}$*. E, ainda,* $n\hat{\sigma}_\varepsilon^2 = \hat{\mathcal{E}}'\hat{\mathcal{E}}$ *possui distribuição qui-quadrado,* $\hat{\sigma}_\varepsilon^2\chi^2_{n-r-1}$*, onde* $\hat{\sigma}_\varepsilon^2$ *é o estimador de* σ_ε^2*.*

Demonstração: Partindo-se da suposição de que a variável aleatória $\mathbf{Y}$ e a variável aleatória $\mathcal{E}$ apresentam distribuição normal, a função de verossimilhança para $\mathcal{B}$ e σ_ε^2 é dada por

$$\begin{aligned}\mathcal{L}(\mathcal{B}, \sigma_\varepsilon^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} \exp\left(\frac{-\mathcal{E}'\mathcal{E}}{2\sigma_\varepsilon^2}\right) \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma_\varepsilon^n} \exp\left[\frac{-(\mathbf{Y} - \mathbf{X}\mathcal{B})'(\mathbf{Y} - \mathbf{X}\mathcal{B})}{2\sigma_\varepsilon^2}\right].\end{aligned}$$

Para σ_ε^2 fixo, a função de verossimilhança é maximizada quando se minimiza a expressão $(\mathbf{Y} - \mathbf{X}\mathcal{B})'(\mathbf{Y} - \mathbf{X}\mathcal{B})$. Ao efetuar esta minimização, obtém-se o mesmo estimador obtido pelo método dos Mínimos Quadrados, ou seja, $\hat{\mathcal{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, o qual não depende do σ_ε^2 adotado.

Agora, maximizando $\mathcal{L}(\hat{\mathcal{B}}, \sigma_\varepsilon^2)$ com respeito a σ_ε^2 tem-se que

$$\mathcal{L}(\hat{\mathcal{B}}, \hat{\sigma}_\varepsilon^2) = \frac{1}{(2\pi)^{\frac{n}{2}}(\hat{\sigma}_\varepsilon^2)^{\frac{n}{2}}} \exp\left(\frac{-n}{2}\right),$$

onde $\hat{\sigma}_\varepsilon^2 = \dfrac{(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})}{n}$.

Segue-se da expressão (1.26) que $\hat{\mathcal{B}}$ e $\hat{\mathcal{E}}$ podem ser representados como combinações lineares das variáveis $\mathcal{E}$, que por suposição, apresentam distribuição normal. De forma mais específica, tem-se

$$\begin{bmatrix} \hat{\mathcal{B}} \\ \hat{\mathcal{E}} \end{bmatrix} = \begin{bmatrix} \mathcal{B} \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \end{bmatrix} \mathcal{E}.$$

Como $\mathbf{X}$ fica fixo, tem-se que $\hat{\mathcal{B}}$ e $\hat{\mathcal{E}}$ apresentam distribuição normal conjunta e são independentes, visto que se pode provar que $Cov(\hat{\mathcal{B}}, \hat{\mathcal{E}}) = \mathbf{0}$. Suas respectivas variâncias e esperanças são dadas no Teorema 1.20. Agora, assume-se que $(\lambda, \mathbf{e})$ sejam o par de autovalor e autovetor de matriz $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, esta matriz é idempotente e assim

$$\begin{aligned}\lambda\mathbf{e} &= [\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{e} \\ &= [\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']^2\mathbf{e} \\ &= \lambda[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{e} = \lambda^2\mathbf{e},\end{aligned}$$

o que implica que $\lambda = 0$ ou $\lambda = 1$.

Por propriedades de matrizes sabe-se que $tr[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = n-r-1$ e, ainda, $tr[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \lambda_1 + \cdots + \lambda_n$, onde $\lambda_1 \geq \cdots \geq \lambda_n$ são os autovalores de $[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']$.

Por conseqüência, existem exatamente $n-r-1$ autovalores com valor um e os demais são nulos. Assim, com base no Teorema da Decomposição Espectral tem-se que

$$[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})\mathbf{X}'] = \mathbf{e}_1\mathbf{e}_1' + \cdots + \mathbf{e}_{n-r-1}\mathbf{e}_{n-r-1}',$$

onde $\mathbf{e}_1, \cdots, \mathbf{e}_{n-r-1}$ são os autovetores normalizados associados respectivamente aos autovalores $\lambda_1 = \cdots = \lambda_{n-r-1} = 1$.

Seja V uma matriz definida como $\mathrm{V} = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_{n-r-1} \end{bmatrix} = \begin{bmatrix} \mathbf{e}_1'\mathcal{E} \\ \mathbf{e}_2'\mathcal{E} \\ \vdots \\ \mathbf{e}_{n-r-1}'\mathcal{E} \end{bmatrix}$.

Então V é possui distribuição normal com média $\mathbf{0}$ e

$$Cov(V_j, V_k) = \begin{cases} \sigma_\varepsilon^2\mathbf{e}_j'\mathbf{e}_k, & j = k \\ 0, & \text{caso contrário.} \end{cases}$$

Pode-se concluir, então, que os V_j são independentes e possuem distribuição $N(0, \sigma_\varepsilon^2)$ e ainda que

$$\begin{aligned}n\hat{\sigma}_\varepsilon^2 &= \hat{\mathcal{E}}'\hat{\mathcal{E}} = \mathcal{E}'[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathcal{E} \\ &= V_1^2 + \mathcal{V}_2^2 + \cdots + V_{n-r-1}^2,\end{aligned}$$

possui distribuição $\sigma_\varepsilon^2\chi^2_{n-r-1}$. □

Com base no estimador de máxima verossimilhança obtido para $\mathcal{B}$, pode-se construir um elipsóide de confiança para este parâmetro. Tal elipsóide é expresso em termos do estimador da matriz de variâncias-covariâncias $\hat{\sigma}_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}$, onde $\hat{\sigma}_\varepsilon^2 = \dfrac{\hat{\mathcal{E}}'\hat{\mathcal{E}}}{n-(r+1)}$.

Proposição 1.21. *Seja* $\mathbf{Y} = \mathbf{X}\mathcal{B}+\mathcal{E}$*, onde* $\mathbf{X}$ *tem posto completo* $r+1$ *e a v.a.* $\mathcal{E}$ *é distribuída como* $N_n(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I})$*. Então uma região de confiança nível* $100(1-\alpha)\%$ *para* $\mathcal{B}$ *é dada por* $(\mathcal{B}-\hat{\mathcal{B}})'\mathbf{X}'\mathbf{X}(\mathcal{B}-\hat{\mathcal{B}}) \leq (r+1)\sigma_\varepsilon^2 F_{r+1,n-r-1}(\alpha)$*, onde* $F_{r+1,n-r-1}(\alpha)$ *é o* (100α)*-ésimo percentil de uma distribuição* F*-Snedecor com* $r+1$ *e* $n-r-1$ *graus de liberdade e* $\hat{o}_\varepsilon^2 = \frac{\hat{\mathcal{E}}'\hat{\mathcal{E}}}{n-(r+1)}$*.*

Demonstração: Considere a matriz $(\mathbf{X}'\mathbf{X})^{\frac{1}{2}}$, pode-se verificar que esta matriz é simétrica. Define-se $V = (\mathbf{X}'\mathbf{X})^{\frac{1}{2}}(\hat{\mathcal{B}}-\mathcal{B})$, de forma que $\mathbb{E}(V) = \mathbf{0}$,

$$\begin{aligned} Var(V) &= (\mathbf{X}'\mathbf{X})^{\frac{1}{2}} Var(\hat{\mathcal{B}})(\mathbf{X}'\mathbf{X})^{\frac{1}{2}} \\ &= \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{\frac{1}{2}}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})^{\frac{1}{2}} \\ &= \sigma_\varepsilon^2\mathbf{I} \end{aligned}$$

e V seja normalmente distribuída (visto que ela consiste em combinações lineares de $\hat{\beta}_i$'s).

Dessa forma,

$$\begin{aligned} V'V &= (\hat{\mathcal{B}}-\mathcal{B})'(\mathbf{X}'\mathbf{X})^{\frac{1}{2}}(\mathbf{X}'\mathbf{X})^{\frac{1}{2}}(\hat{\mathcal{B}}-\mathcal{B}) \\ &= (\hat{\mathcal{B}}-\mathcal{B})'(\mathbf{X}'\mathbf{X})(\hat{\mathcal{B}}-\mathcal{B}), \end{aligned}$$

que apresenta distribuição $\sigma_\varepsilon^2\chi_{r+1}^2$. Pelo resultado apresentado, anteriormente, no Teorema 1.20, tem-se que $(n-r-1)\sigma_\varepsilon^2 = \mathcal{E}'\mathcal{E}$ possui distribuição $\sigma_\varepsilon^2\chi_{n-r-1}^2$, independentemente de $\hat{\mathcal{B}}$ e, desta forma, independentemente de V.

Uma conseqüência deste fato é que

$$\frac{\left[\dfrac{\chi_{r+1}^2}{r+1}\right]}{\left[\dfrac{\chi_{n-r-1}^2}{n-r-1}\right]} = \frac{\left[\dfrac{V'V}{r+1}\right]}{\sigma_\varepsilon^2}$$

possui distribuição F-Snedecor com $r+1$ e $n-r-1$ graus de liberdade e, com base nesta estatística, consegue-se o intervalo procurado para $\mathcal{B}$. □

Observação 1.7. O elipsóide de confiança é centrado no estimador de máxima verossimilhança $\hat{\mathcal{B}}$ e sua orientação e dimensão são determinadas pelos autovalores e autovetores de $\mathbf{X}'\mathbf{X}$. Se um autovalor é próximo de zero, o elipsóide de confiança será muito longo na direção do autovetor correspondente. Várias outras informações sobre intervalos de confiança n-dimensionais podem ser encontradas na página 392 e no apêndice do Capítulo V de Johnson e Wichern (1998).

Parte da análise de regressão objetiva descrever os efeitos de uma variável independente em particular sobre a variável dependente. Uma hipótese nula de interesse afirma que alguns x_i's não possuem influência sobre Y. Estas variáveis preditoras serão renomeadas como $x_{q+1}, \cdots, x_r$. A afirmação de que $x_{q+1}, \cdots, x_r$ não exercem influência sobre Y transcrita em termos estatísticos fica

$$\mathcal{H}_0 : \beta_{q+1} = \cdots = \beta_r = 0 \text{ ou } \mathcal{B}_2 = \mathbf{0}, \tag{1.29}$$

onde $\mathcal{B}_2 = [\beta_{q+1}, \cdots, \beta_r]'$.

Definem-se as matrizes $\mathbf{X}_1$ de tamanho $n \times (q+1)$, $\mathbf{X}_2$ de tamanho $n \times (r-q)$, $\mathcal{B}_1$ de tamanho $(q+1) \times 1$, $\mathcal{B}_2$ de tamanho $(r-q) \times 1$ de maneira que

$$\mathbf{X} = [\mathbf{X}_1 | \mathbf{X}_2] \text{ e } \mathcal{B} = \left[\frac{\mathcal{B}_1}{\mathcal{B}_2}\right].$$

Dessa forma, a equação (1.21) pode ser expressa através do produto de matrizes particionadas na forma

$$\mathbf{Y} = \mathbf{X}_1\mathcal{B}_1 + \mathbf{X}_2\mathcal{B}_2 + \mathcal{E}.$$

Sobre a hipóptese nula $\mathcal{H}_0 : \mathcal{B}_2 = \mathbf{0}$, tem-se que $\mathbf{Y} = \mathbf{X}_1\mathcal{B}_1 + \mathcal{E}$. O teste de razão de verossimilhança para $\mathcal{H}_0$ é baseado na soma de quadrados definida por

$$S(\mathbf{X}) = (\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}}).$$

Proposição 1.22. *Seja $\mathbf{Y} = \mathbf{X}\mathcal{B} + \mathcal{E}$, onde $\mathbf{X}$ tem posto completo $r+1$ e a v.a. $\mathcal{E}$ é distribuída como $N_n(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I})$. O teste de razão de verossimilhança de $\mathcal{H}_0 : \mathcal{B}_2 = \mathbf{0}$ consiste em rejeitar $\mathcal{H}_0$ se*

$$\frac{S(\mathbf{X}_1) - S(\mathbf{X})}{\hat{\sigma}_\varepsilon^2(r-q)} \geq F_{r-q,n-r-1}(\alpha),$$

onde $F_{r-q,n-r-1}(\alpha)$ é o (100α)-ésimo percentil de uma distribuição F-Snedecor com $r-q$ e $n-r-1$ graus de liberdade e $\hat{\sigma}_\varepsilon^2 = \frac{\hat{\mathcal{E}}'\hat{\mathcal{E}}}{n-(r+1)}$.

Demonstração: Assumindo-se que os dados apresentam distribuição normal, a função de verossimilhança associada aos parâmetros $\mathcal{B}$ e σ_ε^2 é dada por

$$\mathcal{L}(\mathcal{B}, \sigma_\varepsilon^2) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma_\varepsilon^n} \exp\left[\frac{-(\mathbf{Y} - \mathbf{X}\mathcal{B})'(\mathbf{Y} - \mathbf{X}\mathcal{B})}{2\sigma_\varepsilon^2}\right] \leq \frac{1}{(2\pi)^{\frac{n}{2}}\hat{\sigma}_\varepsilon^n} \exp\left(\frac{-n}{2}\right),$$

com o máximo ocorrendo em $\hat{\mathcal{B}} = (\mathbf{X}'\mathbf{X})\mathbf{X}'\mathbf{Y}$ e $\hat{\sigma}_\varepsilon^2 = \dfrac{(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})}{n}$.

Sobre a restrição da hipótese nula, tem-se que $\mathbf{Y} = \mathbf{X}_1\mathcal{B}_1 + \mathcal{E}$ e

$$\max_{\mathcal{B}_1,\sigma_\varepsilon^2} \mathcal{L}(\mathcal{B}_1, \sigma_\varepsilon^2) = \frac{1}{(2\pi)^{\frac{n}{2}}\hat{\sigma}_\varepsilon^n} \exp\left(\frac{-n}{2}\right),$$

onde o máximo ocorre em $\hat{\mathcal{B}}_1 = (\mathbf{X}_1'\mathbf{X}^1)^{-1}\mathbf{Y}$. E, desta forma,

$$\hat{\sigma}_\varepsilon^2 = \frac{(\mathbf{Y} - \mathbf{X}_1\hat{\mathcal{B}}_1)'(\mathbf{Y} - \mathbf{X}_1\hat{\mathcal{B}}_1)}{n}.$$

Rejeitar $\mathcal{H}_0 : \mathcal{B}_2 = \mathbf{0}$ para pequenos valores da razão de verossimilhança

$$\begin{aligned}\frac{\max_{\mathcal{B}_1,\sigma_\varepsilon^2} \mathcal{L}(\mathcal{B}_1, \sigma_\varepsilon^2)}{\max_{\mathcal{B},\sigma_\varepsilon^2} \mathcal{L}(\mathcal{B}, \sigma_\varepsilon^2)} &= \left(\frac{\sigma_{1\varepsilon}^2}{\hat{\sigma}_\varepsilon^2}\right)^{\frac{-n}{2}} = \left(\frac{\hat{\sigma}_\varepsilon^2 + \hat{\sigma}_{1\varepsilon}^2 - \sigma_\varepsilon^2}{\sigma_\varepsilon^2}\right)^{\frac{-n}{2}} \\ &= \left(1 + \frac{\hat{\sigma}_{1\varepsilon}^2 - \sigma_\varepsilon^2}{\sigma_\varepsilon^2}\right)^{\frac{-n}{2}},\end{aligned}$$

é equivalente a rejeitar $\mathcal{H}_0$ para valores grandes de $\frac{(\hat{\sigma}_{1\varepsilon}^2 - \hat{\sigma}_\varepsilon^2)}{\hat{\sigma}_\varepsilon^2}$, ou seja, na sua escala inversa,

$$\frac{\dfrac{n(\sigma_{1\varepsilon}^2 - \sigma_\varepsilon^2)}{r-q}}{\dfrac{n\sigma_\varepsilon^2}{n-r-1}} = \frac{S(\mathbf{X}_1) - S(\mathbf{X})}{\hat{\sigma}_\varepsilon^2(r-q)} \geq F_{r-q,n-r-1}(\alpha).$$

□

De forma mais geral, é possível formular hipótese nula com base em $r - q$ combinações lineares de $\mathcal{B}$, na forma $\mathcal{H}_0 : \mathbf{C}\mathcal{B} = \mathbf{0}$. Considere a matriz $\mathbf{C}$, de tamanho $(r-q) \times (r+1)$, de posto completo. Sobre estas circunstâncias, $\mathbf{C}\hat{\mathcal{B}}$ tem distribuição $N_{r-q}(\mathbf{C}\mathcal{B}, \sigma_\varepsilon^2\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')$. Rejeita-se $\mathcal{H}_0 : \mathbf{C}\mathcal{B} = \mathbf{0}$ com nível α se $\mathbf{0}$ não pertence ao elipsóide de confiança para $\mathbf{C}\mathcal{B}$ com $100(1-\alpha)\%$ de confiança.

Vamos utilizar os resultados obtidos nos parágrafos anteriores para resolver o problema de calcular a esperança da v.a. Y dado um valor da variável independente $\mathbf{X} = (x_0, x_1, \cdots, x_r)$. Os valores $\mathbf{X}$ e $\hat{\mathcal{B}}$ podem ser utilizados para estimar o valor de Y dado $\mathbf{X}$.

Seja Y_0 o valor obtido pela variável independente $\mathbf{X}_0$. De acordo com o modelo (1.22), o valor esperado de Y_0 dado $\mathbf{X} = \mathbf{X}_0$ é dado por

$$\mathbb{E}(Y_0|\mathbf{X}_0) = \beta_0 x_{00} + \beta_1 x_{01} + \cdots + \beta_r x_{0r} = \mathbf{X}_0'\mathcal{B}.$$

Proposição 1.23. *Para o modelo linear dado pela expressão (1.21),* $\mathbf{X}_0'\hat{\mathcal{B}}$ *é o estimador obtido pelo método dos mínimos quadrados para* $\mathbb{E}(Y_0|\mathbf{X}_0)$ *com mínima variância* $Var(\mathbf{X}_0'\hat{\mathcal{B}}) = \sigma_\varepsilon^2\,\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0$*. Se o erro* $\mathcal{E}$ *é normalmente distribuído, então um intervalo com nível* $100(1-\alpha)\%$ *de confiança para* $\mathbb{E}(Y_0|\mathbf{X}_0)$ *é obtido por*

$$\mathbf{X}_0'\hat{\mathcal{B}} \pm t_{n-r-1,\frac{\alpha}{2}}\sqrt{\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\hat{\sigma}_\varepsilon^2},$$

onde $t_{n-r-1,\frac{\alpha}{2}}$ *é o* $100(\frac{\alpha}{2})$*-ésimo percentil de uma distribuição t-Student com* $n-r-1$ *graus de liberdade e* $\hat{\sigma}_\varepsilon^2 = \frac{\hat{\mathcal{E}}'\hat{\mathcal{E}}}{n-(r+1)}$*.*

Demonstração: Para $\mathbf{X}_0$ fixo, tem-se que $\mathbf{X}_0'\mathcal{B}$ é uma combinação linear de β_i's. Dessa forma, o Teorema 1.19 pode ser aplicado. Pelo resultado da Proposição 1.18, segue-se que

$$Var(\mathbf{X}_0'\hat{\mathcal{B}}) = \mathbf{X}_0'Var(\hat{\mathcal{B}})\mathbf{X}_0 = \sigma_\varepsilon^2\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0.$$

Usando-se da hipótese de normalidade do erro, pelo Teorema 1.20, segue-se que $\hat{\mathcal{B}}$ possui distribuição $N_{r+1}(\mathcal{B}, \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1})$ e é independente de $\frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2}$ que possui distribuição $\frac{\chi^2_{n-r-1}}{n-r-1}$. Conseqüentemente, a combinação linear $\mathbf{X}_0'\hat{\mathcal{B}}$ é distribuída como $N_{r+1}(\mathbf{X}_0'\mathcal{B}, \sigma_\varepsilon^2\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0)$ e

$$\frac{\dfrac{\mathbf{X}_0'\mathcal{B} - \mathbf{X}_0'\hat{\mathcal{B}}}{\sqrt{\sigma_\varepsilon^2\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0}}}{\sqrt{\dfrac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2}}} = \frac{\mathbf{X}_0'\mathcal{B} - \mathbf{X}_0'\hat{\mathcal{B}}}{\sqrt{\hat{\sigma}_\varepsilon^2\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0}}$$

possui distribuição t-Student com $n-r-1$ graus de liberdade. □

1.3 Regressão Linear Multivariada

Nesta seção, considera-se o problema de modelar a relação entre m variáveis dependentes $\mathbf{Y}_1, \cdots, \mathbf{Y}_m$ e um conjunto de variáveis independentes $\mathbf{X}_1, \cdots, \mathbf{X}_r$. Cada $\mathbf{Y}_k$, com $k \in \{1, \cdots, m\}$, é assumida seguindo seu próprio modelo de regressão. Desta forma, se ao se assumir que as variáveis independentes tenham valores particulares $\mathbf{x}_1, \cdots, \mathbf{x}_r$, o modelo apresenta-se como

$$\begin{aligned}
\mathbf{Y}_1 &= \beta_{01} + \beta_{11}\mathbf{x}_1 + \cdots + \beta_{r1}\mathbf{x}_r + \mathcal{E}_1 \\
\mathbf{Y}_2 &= \beta_{02} + \beta_{12}\mathbf{x}_1 + \cdots + \beta_{r2}\mathbf{x}_r + \mathcal{E}_2 \\
&\vdots \\
\mathbf{Y}_m &= \beta_{0m} + \beta_{1m}\mathbf{x}_1 + \cdots + \beta_{rm}\mathbf{x}_r + \mathcal{E}_m,
\end{aligned}$$

O vetor aleatório $\mathcal{E} = [\varepsilon_1, \cdots, \varepsilon_m]'$ apresenta $\mathbb{E}(\mathcal{E}) = \mathbf{0}$ e $Var(\mathcal{E}) = \mathbf{\Sigma}$. Dessa forma, os erros associados a diferentes variáveis dependentes podem estar correlacionados. Para se estabelecer uma notação para o modelo linear de regressão neste caso, assume-se que $[x_{i0}, \cdots, x_{ir}]$ denota os valores das variáveis independentes para a i-ésima coleta de informações, com $i \in \{1, \cdots, n\}$, $\mathbf{Y}_i = [Y_{i1}, \cdots, Y_{im}]'$ seja o valor obtido na i-ésima coleta de informações, e ainda $\mathcal{E}_i = [\varepsilon_{i1}, \cdots, \varepsilon_{im}]'$ seja o vetor dos erros, os quais não se sabe qual distribuição possui, ou seja, tem-se um análogo ao Caso B visto nas seções precedentes.

Em notação matricial, tem-se que

$$\mathbf{X} = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1r} \\ x_{20} & x_{21} & \cdots & x_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{nr} \end{bmatrix},$$

onde $x_{i0} \equiv 1$ para todo $i \in \{1, \cdots, n\}$.

As outras matrizes, referentes ao modelo, podem ser apresentadas como matrizes particionadas a seguir

$$\mathbf{Y} = \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1m} \\ Y_{21} & Y_{22} & \cdots & Y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nm} \end{bmatrix} = [Y_1|Y_2|\cdots|Y_m],$$

$$\mathcal{B} = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{r1} & \beta_{r2} & \cdots & \beta_{rm} \end{bmatrix} = [\mathcal{B}_1|\mathcal{B}_2|\cdots|\mathcal{B}_m]$$

e

$$\mathcal{E} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1m} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nm} \end{bmatrix} = [\mathcal{E}_1|\mathcal{E}_2|\cdots|\mathcal{E}_m].$$

A vantagem de se utilizar matrizes particionadas consiste em se resumir os resultados encontrados na seção precedente como casos particulares dos contextos em estudo nesta seção. Desta forma, as demonstrações podem ser reduzidas e simplificadas.

Os comentários anteriores motivam a definição a seguir.

Definição 1.6. O *modelo de regressão linear multivariado* é dado por

$$\mathbf{Y} = \mathbf{X}\mathcal{B} + \mathcal{E}$$

com

$$\mathbb{E}(\mathcal{E}_k) = \mathbf{0} \quad \text{e} \quad Cov(\mathcal{E}_k, \mathcal{E}_l) = \sigma_{kl}\mathbf{I}, \quad k, l \in \{1, \cdots, m\}.$$

As m observações na i-ésima amostra possuem matriz de variâncias-covariâncias $\mathbf{\Sigma}_i = (\sigma_{kl})$, mas observações de diferentes amostras são não-correlacionadas, emq que $\mathcal{B}$ e σ_{kl} são parâmetros desconhecidos, e a i-ésima linha da matriz $\mathbf{X}$ é $[x_{i0}, \cdots, x_{ir}]$.

Observação 1.8. Segundo a Definição 1.6, a k-ésima variável de resposta, Y_k, segue o modelo de regressão apresentado na Definição 1.4, para $k \in \{1, \cdots, m\}$, com $Var(\mathcal{E}_k) = \sigma_{kk}\mathbf{I}$. Entretanto, os erros de diferentes variáveis resposta em uma mesma amostra podem estar correlacionados.

Teorema 1.24. *Dada a matriz* $\mathbf{Y}$ *e a matriz* $\mathbf{X}$ *com posto completo, tem-se que o estimador de* $\mathcal{B}$*, pelo método dos mínimos quadrados, é dado por*

$$\hat{\mathcal{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{1.30}$$

Seja $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathcal{B}}$ *a representação dos valores ajustados de* $\mathbf{Y}$*. Então, os resíduos*

$$\hat{\mathcal{E}} = \mathbf{Y} - \hat{\mathbf{Y}} = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}$$

satisfazem $\mathbf{X}'\hat{\mathcal{E}} = \mathbf{0}$ *e* $\hat{\mathbf{Y}}'\hat{\mathcal{E}} = \mathbf{0}$*. E ainda, a soma dos quadrados dos resíduos é dada por*

$$\begin{aligned} \hat{\mathcal{E}}'\hat{\mathcal{E}} &= \mathbf{Y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{Y} - \hat{\mathcal{B}}'\mathbf{X}'\mathbf{X}\hat{\mathcal{B}}. \end{aligned}$$

Demonstração: Dada a matriz de valores de respostas $\mathbf{Y}$ e a matriz de valores da variável independente $\mathbf{X}$, com base na Observação 1.8, pode-se determinar os estimadores $\hat{\mathcal{B}}_i$ de cada observação $\mathbf{Y}_i$ na i-ésima resposta. De acordo com o resultado

obtido pelo Teorema 1.17, tem-se que $\hat{\mathcal{B}}_i$ é dado pela expressão (1.23). Agrupando os m estimadores obtidos em uma matriz particionada, segue da expressão (1.30) que

$$\begin{aligned}\hat{\mathcal{B}} &= [\hat{\mathcal{B}}_1|\hat{\mathcal{B}}_2|\cdots|\hat{\mathcal{B}}_m] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[Y_1|Y_2|\cdots|Y_m] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.\end{aligned}$$

Para qualquer escolha dos parâmetros $\mathbf{B} = [\mathbf{b}_1|\mathbf{b}_2|\cdots|\mathbf{b}_m]$, a matriz de erros é $\mathbf{Y} - \mathbf{X}\mathcal{B}$. Dessa forma, tem-se que $(\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB})$

$$= \begin{bmatrix} (\mathbf{Y}_1 - \mathbf{X}\mathbf{b}_1)'(\mathbf{Y}_1 - \mathbf{X}\mathbf{b}_1) & \cdots & (\mathbf{Y}_1 - \mathbf{X}\mathbf{b}_1)'(\mathbf{Y}_m - \mathbf{X}\mathbf{b}_m) \\ \vdots & \ddots & \vdots \\ (\mathbf{Y}_m - \mathbf{X}\mathbf{b}_m)'(\mathbf{Y}_1 - \mathbf{X}\mathbf{b}_1) & \cdots & (\mathbf{Y}_m - \mathbf{X}\mathbf{b}_m)'(\mathbf{Y}_m - \mathbf{X}\mathbf{b}_m) \end{bmatrix}.$$

A seleção $\mathbf{b}_k = \mathcal{B}_k$ minimiza a i-ésima soma de quadrados dada por $(\mathbf{Y}_k - \mathbf{X}\mathbf{b}_k)'(\mathbf{Y}_k - \mathbf{X}\mathbf{b}_k)$. Conseqüentemente, ocorre que $tr[(\mathbf{Y} - \mathbf{X}\mathcal{B})'(\mathbf{Y} - \mathbf{X}\mathcal{B})]$ é minimizado pela escolha $\mathbf{B} = \hat{\mathcal{B}}$.

Usando o estimador dos mínimos quadrados $\hat{\mathcal{B}}$, pode-se formar as seguintes matrizes

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{X}\hat{\mathcal{B}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ \hat{\mathcal{E}} &= \mathbf{Y} - \hat{\mathbf{Y}} = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}.\end{aligned}$$

As condições de ortogonalidade entre os resíduos, valores preditos e as colunas de $\mathbf{X}$, que são válidas para o modelo de regressão linear múltipla, também são válidas para o modelo de regressão linear multivariada. Dessa forma, segue-se que $\mathbf{X}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \mathbf{X}' - \mathbf{X}' = \mathbf{0}$. Especificamente, tem-se

$$\mathbf{X}'\hat{\mathcal{E}} = \mathbf{X}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} = \mathbf{0},$$

assim os resíduos $\hat{\mathcal{E}}_k$ são perpendiculares às colunas de $\mathbf{X}$.

E, ainda,

$$\hat{\mathbf{Y}}'\hat{\mathcal{E}} = \hat{\mathcal{B}}'\mathbf{X}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} = \mathbf{0},$$

confirmando que os valores preditos $\hat{\mathbf{Y}}_k$ são perpendiculares aos resíduos $\hat{\mathcal{E}}_k$.

Devido ao fato de que $\mathbf{Y} = \hat{\mathbf{Y}} + \hat{\mathcal{E}}$, pode-se obter

$$\mathbf{Y}'\mathbf{Y} = (\hat{\mathbf{Y}} + \hat{\mathcal{E}})'(\hat{\mathbf{Y}} + \hat{\mathcal{E}}) = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \hat{\mathcal{E}}'\hat{\mathcal{E}}.$$

Segue-se, então, que a soma dos quadrados dos resíduos pode ser escrita como

$$\hat{\mathcal{E}}'\hat{\mathcal{E}} = \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \mathbf{Y}'\mathbf{Y} - \hat{\mathcal{B}}'\mathbf{X}'\mathbf{X}\hat{\mathcal{B}}.$$

□

De posse do estimador $\hat{\mathcal{B}}$, obtido pelo método dos mínimos quadrados, para $\mathcal{B}$, pode-se agora apresentar algumas propriedades do estimador e algumas inferências sobre o modelo proposto pela Definição 1.6.

Proposição 1.25. *Para o estimador dos mínimos quadrados $\hat{\mathcal{B}} = [\hat{\mathcal{B}}_1|\cdots|\hat{\mathcal{B}}_m]$ obtido no Teorema 1.24 determinado sob as condições da Definição 1.6, com a matriz $\mathbf{X}$ de posto completo $r+1 < n$, tem-se que*

$$\begin{aligned} \mathbb{E}(\hat{\mathcal{B}}_k) = \mathcal{B}_k, \qquad & \mathbb{E}(\hat{\mathcal{B}}) = \mathcal{B} \quad e \\ Cov(\hat{\mathcal{B}}_k, \hat{\mathcal{B}}_l) = \sigma_{kl}(\mathbf{X}'\mathbf{X})^{-1}, \qquad & k, l \in \{1, \cdots, m\}. \end{aligned}$$

Os resíduos $\hat{\mathcal{E}} = [\hat{\mathcal{E}}_1|\cdots|\hat{\mathcal{E}}_m] = \mathbf{Y} - \mathbf{X}\hat{\mathcal{B}}$ satisfazem $\mathbb{E}(\hat{\mathcal{E}}_k) = \mathbf{0}$ e $\mathbb{E}(\hat{\mathcal{E}}_k'\hat{\mathcal{E}}_l) = (n-r-1)\sigma_{kl}$. Então,

$$\mathbb{E}(\hat{\mathcal{E}}) = \mathbf{0} \quad e \quad \mathbb{E}\left(\frac{1}{n-r-1}\hat{\mathcal{E}}'\hat{\mathcal{E}}\right) = \mathbf{\Sigma}.$$

Além disso, $\hat{\mathcal{E}}$ e $\hat{\mathcal{B}}$ são não correlacionados.

Demonstração: Segue-se do modelo de regressão linear múltipla que a k-ésima variável resposta aparesenta

$$\mathbf{Y}_k = \mathbf{X}\mathcal{B}_k + \mathcal{E}_k, \quad \mathbb{E}(\mathcal{E}_k) = \mathbf{0}, \quad \text{e} \quad \mathbb{E}(\mathcal{E}_k\mathcal{E}_k') = \sigma_{\varepsilon}^2{}_{kk}\mathbf{I}.$$

E ainda,

$$\hat{\mathcal{B}}_k - \mathcal{B}_k = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_k - \mathcal{B}_k = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathcal{E}}_k \quad \text{e}$$

$$\begin{aligned} \hat{\mathcal{E}}_k &= \mathbf{Y}_k - \hat{\mathbf{Y}}_k = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}_k \\ &= [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathcal{E}_k, \end{aligned}$$

o que implica $\mathbb{E}(\hat{\mathcal{B}}_k) = \mathcal{B}_k$ e $\mathbb{E}(\hat{\mathcal{E}}_k) = \mathbf{0}$.

Na seqüência, tem-se

$$\begin{aligned} Cov(\hat{\mathcal{B}}_k, \hat{\mathcal{B}}_l) &= \mathbb{E}[(\hat{\mathcal{B}}_k - \mathcal{B}_k)(\hat{\mathcal{B}}_l - \mathcal{B}_l)'] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}(\mathcal{E}_k\mathcal{E}_l')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma_{kl}(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

Usando-se das propriedades do traço de uma matriz e dos resultados obtidos na Proposição 1.18, para $\mathbf{U}$ um vetor aleatório qualquer e $\mathbf{A}$ uma matriz fixada, tem-se que

$$\mathbb{E}[\mathbf{U}'\mathbf{A}\mathbf{U}] = \mathbb{E}[tr(\mathbf{A}\mathbf{U}\mathbf{U}')] = tr[\mathbf{A}\mathbb{E}(\mathbf{U}\mathbf{U}')].$$

Conseqüentemente,

$$\begin{aligned} \mathbb{E}(\hat{\mathcal{E}}_k'\hat{\mathcal{E}}_l) &= \mathbb{E}\{\mathcal{E}_k'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathcal{E}_l\} \\ &= tr\{[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\sigma_{kl}\mathbf{I}\} \\ &= \sigma_{kl}tr[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \\ &= \sigma_{kl}(n - r - 1). \end{aligned}$$

Ao se dividir cada elemento $\hat{\mathcal{E}}_k'\hat{\mathcal{E}}_l$ de $\hat{\mathcal{E}}'\hat{\mathcal{E}}$ por $n-r-1$ obtém-se o estimador não viciado para $\mathbf{\Sigma}$.

Finalmente,

$$\begin{aligned} Cov(\hat{\mathcal{B}}_k, \hat{\mathcal{E}}_l) &= \mathbb{E}\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathcal{E}_k\mathcal{E}_l'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}(\mathcal{E}_k\mathcal{E}_l')[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\sigma_{kl}'\mathbf{I}[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \\ &= \sigma_{kl}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \mathbf{0}, \end{aligned}$$

de onde se conclui que cada elemento de $\hat{\mathcal{B}}$ é não correlacionado com cada elemento de $\hat{\mathcal{E}}$. □

Estimadores pelo método de máxima verossimilhança para os parâmetros do modelo apresentado na Definição 1.6 podem ser obtidos quando os erros $\mathcal{E}$ possuem distribuição normal. E, assim, tem-se um contexto análogo ao Caso A abordado nas seções precedentes.

Proposição 1.26. *Seja o modelo apresentado na Definição 1.6, com a matriz* $\mathbf{X}$ *de posto completo* $r+1$*, de forma que* $n \geq r+1 = m$*, e os erros* $\mathcal{E}$ *tenham distribuição*

normal. Então,

$$\hat{\mathcal{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

é o estimador de máxima verossimilhança de $\mathcal{B}$ *e, ainda,* $\mathcal{B}$ *tem distribuição normal com* $\mathbb{E}(\hat{\mathcal{B}}) = \mathcal{B}$ *e* $Cov(\hat{\mathcal{B}}_k, \hat{\mathcal{B}}_l) = \sigma_{kl}(\mathbf{X}'\mathbf{X})^{-1}$*. Também,* $\hat{\mathcal{B}}$ *é independente do estimador de máxima verossimilhança da matriz positiva definida* $\mathbf{\Sigma}$*, que é dado por*

$$\hat{\mathbf{\Sigma}} = \frac{\hat{\mathcal{E}}'\hat{\mathcal{E}}}{n} = \frac{(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})}{n},$$

onde $n\hat{\mathbf{\Sigma}}$ *possui distribuição de Wishart com* r *e* $n - r - 1$ *graus de liberdade, denotada por* $W_{r,n-r-1}$*.*

Demonstração: De acordo com o modelo de regressão, a função de verossimilhança é determinada com base nos dados $\mathbf{Y} = [\mathbf{Y}_1|\mathbf{Y}_2|\cdots|\mathbf{Y}_n]'$ que possuem linhas independentes, onde $\mathbf{Y}_i$ apresenta distribuição $N_m(\mathcal{B}'\mathbf{X}_i, \mathbf{\Sigma})$, com $i \in \{1, \cdots, n\}$.

Primeiramente, note-se que $\mathbf{Y} - \mathbf{X}\mathcal{B} = [\mathbf{Y}_1 - \mathcal{B}'\mathbf{X}_1|\mathbf{Y}_2 - \mathcal{B}'\mathbf{X}_2|\cdots|\mathbf{Y}_n - \mathcal{B}'\mathbf{X}_n]'$, dessa forma

$$(\mathbf{Y} - \mathbf{X}\mathcal{B})'(\mathbf{Y} - \mathbf{X}\mathcal{B}) = \sum_{i=1}^{n}(\mathbf{Y}_i - \mathcal{B}'\mathbf{X}_i)(\mathbf{Y}_i - \mathcal{B}'\mathbf{X}_i)', \quad \text{e}$$

$$\begin{aligned}
\sum_{i=1}^{n}(\mathbf{Y}_i - \mathcal{B}'\mathbf{X}_i)'\mathbf{\Sigma}^{-1}(\mathbf{Y}_i - \mathcal{B}'\mathbf{X}_i) &= \sum_{i=1}^{n} tr[(\mathbf{Y}_i - \mathcal{B}'\mathbf{X}_i)'\mathbf{\Sigma}^{-1}(\mathbf{Y}_i - \mathcal{B}'\mathbf{X}_i)] \\
&= \sum_{i=1}^{n} tr[\mathbf{\Sigma}^{-1}(\mathbf{Y}_i - \mathcal{B}'\mathbf{X}_i)(\mathbf{Y}_i - \mathcal{B}'\mathbf{X}_i)'] \\
&= tr[\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\mathcal{B})'(\mathbf{Y} - \mathbf{X}\mathcal{B})']. \qquad (1.31)
\end{aligned}$$

Outro cálculo preliminar que pode ser desenvolvido, com base nos resultados apresentados no Teorema 1.24, é

$$\begin{aligned}
(\mathbf{Y} - \mathbf{X}\mathcal{B})'(\mathbf{Y} - \mathbf{X}\mathcal{B}) &= [\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}} + \mathbf{X}(\hat{\mathcal{B}} - \mathcal{B})]'[\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}} + \mathbf{X}(\hat{\mathcal{B}} - \mathcal{B})] \\
&= (\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}}) + (\hat{\mathcal{B}} - \mathcal{B})'\mathbf{X}'\mathbf{X}(\hat{\mathcal{B}} - \mathcal{B}) \\
&= \hat{\mathcal{E}}'\hat{\mathcal{E}} + (\hat{\mathcal{B}} - \mathcal{B})'\mathbf{X}'\mathbf{X}(\hat{\mathcal{B}} - \mathcal{B}). \qquad (1.32)
\end{aligned}$$

Usando-se as expressões obtidas em (1.31) e (1.32), obtém-se a função de verossimilhança como segue

$$
\begin{aligned}
\mathcal{L}(\mathcal{B},\mathbf{\Sigma}) &= \prod_{i=1}^{n}\frac{1}{(2\pi)^{\frac{m}{2}}\det(\mathbf{\Sigma})^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(\mathbf{Y}_i-\mathcal{B}'\mathbf{X}_i)'\mathbf{\Sigma}^{-1}(\mathbf{Y}_i-\mathcal{B}'\mathbf{X}_i)\right]\\
&= \frac{1}{(2\pi)^{\frac{nm}{2}}\det(\mathbf{\Sigma})^{\frac{n}{2}}}\exp\left\{-\frac{1}{2}tr[\mathbf{\Sigma}^{-1}(\hat{\mathcal{E}}'\hat{\mathcal{E}}+(\hat{\mathcal{B}}-\mathcal{B})'\mathbf{X}'\mathbf{X}(\hat{\mathcal{B}}-\mathcal{B}))]\right\}\\
&= \frac{1}{(2\pi)^{\frac{nm}{2}}\det(\mathbf{\Sigma})^{\frac{n}{2}}}\ \times\\
&\quad \exp\left\{-\frac{1}{2}tr(\mathbf{\Sigma}^{-1}\hat{\mathcal{E}}'\hat{\mathcal{E}})-\frac{1}{2}tr[\mathbf{X}(\hat{\mathcal{B}}-\mathcal{B})\mathbf{\Sigma}^{-1}(\hat{\mathcal{B}}-\mathcal{B})'\mathbf{X}']\right\}. \qquad (1.33)
\end{aligned}
$$

Pode-se provar que a matriz $\mathbf{X}(\hat{\mathcal{B}}-\mathcal{B})\mathbf{\Sigma}^{-1}(\hat{\mathcal{B}}-\mathcal{B})'\mathbf{X}'$ é não negativa definida. Assim, seus autovalores são todos não negativos e seu traço (que corresponde à soma dos autovalores) assumirá o valor mínimo, zero, se $\mathcal{B}=\hat{\mathcal{B}}$. Esta escolha é única pois $\mathbf{X}$ possui posto completo e, ainda, $\hat{\mathcal{B}}_i-\mathcal{B}_i\neq\mathbf{0}$ implica que $\mathbf{X}(\hat{\mathcal{B}}_i-\mathcal{B}_i)\neq\mathbf{0}$ e, portanto, $tr[\mathbf{X}(\hat{\mathcal{B}}-\mathcal{B})\mathbf{\Sigma}^{-1}(\hat{\mathcal{B}}-\mathcal{B})'\mathbf{X}']\geq\mathbf{a}'\mathbf{\Sigma}^{-1}\mathbf{a}>0$, onde $\mathbf{a}'$ é qualquer linha de $\mathbf{X}(\hat{\mathcal{B}}-\mathcal{B})$.

Para se estabelecer os resultados com referência às distribuições de probabilidade utiliza-se a Proposição 1.25 da qual se pode obter que os elemntos $\hat{\mathcal{B}}_i$ e $\hat{\mathcal{E}}_i$ são combinações lineares dos elementos de $\mathcal{E}$.

Pode-se provar que $\hat{\mathcal{B}}_1,,\hat{\mathcal{B}}_m,\hat{\mathcal{E}}_1,,\hat{\mathcal{E}}_m$ possuem distribuição conjunta normal e assim suas esperanças e covariâncias são obtidas pela Proposição 1.25. E, $\hat{\mathcal{E}}$ e $\hat{\mathcal{B}}$ são independentes visto que possuem matriz de variâncias-covariâncias nula.

Usando-se de um procedimento análogo ao realizado no Teorema 1.20, tem-se

$$
\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'=\sum_{l=1}^{n-r-1}\mathbf{e}_l\mathbf{e}_l',\ \text{ onde }\ \mathbf{e}_l'\mathbf{e}_l=1,
$$

para $l\neq k$, tem-se que $\mathbf{e}_l'\mathbf{e}_k=0$.

Define-se

$$
\begin{aligned}
V_l &= \mathcal{E}'\mathbf{e}_l=[\mathcal{E}_1'\mathbf{e}_l|\mathcal{E}_2'\mathbf{e}_l|\cdots|\mathcal{E}_m'\mathbf{e}_l]'\\
&= e_{l1}\mathcal{E}_1+e_{l2}\mathcal{E}_2+\cdots+e_{ln}\mathcal{E}_n,
\end{aligned}
$$

devido ao fato de que V_l, com $l\in\{1,\cdots,n-r-1\}$, são combinações lineares dos elementos de $\mathcal{E}$, e apresentam distribuição normal conjunta com $\mathbb{E}(V_l)=\mathbb{E}(\mathcal{B}')\mathbf{e}_l=\mathbf{0}$.

Para $l \neq k$, V_l e V_k possuem matriz de variâncias-covariâncias dada por $\mathbf{e}_l'\mathbf{e}_k\Sigma = (0)\Sigma = \mathbf{0}$. Conseqüentemente, os elementos V_l apresentam distribuição $N_m(\mathbf{0}, \Sigma)$ e são independentes.

Finalmente,

$$\begin{aligned}\hat{\mathcal{E}}'\hat{\mathcal{E}} &= \hat{e}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathcal{E} \\ &= \sum_{l=1}^{n-r-1} \mathcal{E}'\mathbf{e}_l\mathbf{e}_l'\mathcal{E} \\ &= \sum_{l=1}^{n-r-1} V_l V_l',\end{aligned}$$

que, por definição, possui distribuição Wishart com r e $n-r-1$ graus de liberdade. A distribuição de Wishart é a generalização multivariada da distribuição qui-quadrado. Detalhes e propriedades sobre a distribuição de Wishart podem ser encontrados no Capítulo IV Johnson e Wichern (1998) ou no Capítulo VI de Chatfield e Collins (1980). □

A Proposição 1.27, a seguir, fornece informação extra sobre o uso dos estimadores pelo método dos mínimos quadrados. Quando os erros são normalmente distribuídos, $\hat{\mathcal{B}}$ e $n^{-1}\hat{\mathcal{E}}'\hat{\mathcal{E}}$ são os estimadores de máxima verossimilhança para $\mathcal{B}$ e $\mathbf{\Sigma}$, respectivamente. Entretanto, para grandes amostras, eles tem aproximadamente a menor variância possível. Ou seja, tem-se aqui um análogo à propriedade apresentada na Proposição 1.16.

O teste de hipótese do caso multivariado, análogo à situação apresentada em (1.29), apresenta a hipótese de que as variáveis Y_k, com $k \in \{1, \cdots, m\}$, não dependem de $\mathbf{X}_{q+1}, \mathbf{X}_{q+2}, \cdots, \mathbf{X}_r$, e é representado por

$$\mathcal{H}_0 : \mathcal{B}_2 = \mathbf{0}, \tag{1.34}$$

onde $\mathcal{B}_2 = [\mathcal{B}_{q+1}, \cdots, \beta_r]'$,

Definem-se as matrizes $\mathbf{X}_1$ de tamanho $n \times (q+1)$, $\mathbf{X}_2$ de tamanho $n \times (r-q)$, $\mathcal{B}_1$ de tamanho $(q+1) \times 1$, $\mathcal{B}_2$ de tamanho $(r-q) \times 1$ de maneira que

$$\mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2] \quad \text{e} \quad \mathcal{B} = \left[\frac{\mathcal{B}_1}{\mathcal{B}_2}\right].$$

Dessa forma, decorre da Definição 1.6 que

$$\mathbb{E}(\mathbf{Y}) = \mathbf{X}\mathcal{B} = \mathbf{X}_1\mathcal{B}_1 + \mathbf{X}_2\mathcal{B}_2.$$

Sob a hipótese (1.34), tem-se que $\mathbf{Y} = \mathbf{X}_1\mathcal{B}_1 + \mathcal{E}$ e o teste de razão de verossimilhança de $\mathcal{H}_0$ é baseado em quantidades que envolvem a expressão

$$n(\hat{\mathbf{\Sigma}}_1 - \hat{\mathbf{\Sigma}}) = (\mathbf{Y} - \mathbf{X}_1\hat{\mathcal{B}}_1)'(\mathbf{Y} - \mathbf{X}_1\hat{\mathcal{B}}_1) - (\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}}),$$

onde $\hat{\mathcal{B}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}$ e $\hat{\mathbf{\Sigma}}_1 = n^{-1}(\mathbf{Y} - \mathbf{X}_1\hat{\mathcal{B}}_1)'(\mathbf{Y} - \mathbf{X}_1\hat{\mathcal{B}}_1)$.

Proposição 1.27. *Seja o modelo dado pela Definição 1.6. Assume-se que* $\mathbf{X}$ *possui posto completo* $r+1$ *e que* $r+1+m \leq n$. *Seja* $\mathcal{E}$ *v.a. com distribuição normal multivariada. Sob a hipótese (1.34), tem-se que* $n\hat{\mathbf{\Sigma}}$ *possui distribuição* $W_{r,n-r-1}$ *e é independente de* $n(\hat{\mathbf{\Sigma}}_1 - \hat{\mathbf{\Sigma}})$, *a qual possui distribuição* $W_{r,r-q}$. *O teste de razão de verossimilhança para* $\mathcal{H}_0$ *é equivalente a rejeitar a hipótese nula para valores altos de*

$$-2\ln\Lambda = -n\ln\left(\frac{|\hat{\mathbf{\Sigma}}|}{|\hat{\mathbf{\Sigma}}_1|}\right) = -n\ln\frac{|n\hat{\mathbf{\Sigma}}|}{|n\hat{\mathbf{\Sigma}} + n(\hat{\mathbf{\Sigma}}_1 - \hat{\mathbf{\Sigma}})|}.$$

Para n *suficientemente grande, a estatística*

$$-\left[n - r - 1 - \frac{1}{2}(m - r + q + 1)\right]\ln\left(\frac{|\hat{\mathbf{\Sigma}}|}{|\hat{\mathbf{\Sigma}}_1|}\right)$$

tem, aproximadamente, distribuição qui-quadrado com $m(r-q)$ *graus de liberdade.*

Demonstração: A demonstração desta afirmação encontra-se no Apêndice A. □

Observação 1.9. Tecnicamente, os valores $n - r$ e $n - m$ são suficientemente grandes para se obter uma boa aproximação da distribuição qui-quadrado com $m(r-q)$ graus de liberdade.

Observação 1.10. Se $\mathbf{X}$ não apresenta posto completo, mas possui posto $r_1 + 1$, então $\hat{\mathcal{B}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$, onde $(\mathbf{X}'\mathbf{X})^{-}$ é dita a inversa generalizada de $\mathbf{X}$. A inversa generalizada é definida como $(\mathbf{X}'\mathbf{X})^{-} = \sum_{j=1}^{r_1+1}\lambda_j^{-1}\mathbf{e}_j\mathbf{e}_j'$, onde $\lambda_1 \geq \cdots \geq \lambda_{r_1+1} > 0$ e $\lambda_{r_1+2} = \cdots = \lambda_{r+1} = 0$. Desta forma a expressão $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ possui posto r_1+1 e gera uma única projeção de $\mathbf{Y}$ no espaço expandido pelas colunas independentes de $\mathbf{X}$. Este é verdadeiro para qualquer escolha da inversa generalizada. Para maiores detalhes recomenda-se Rao (1973).

Supõe-se que o modelo proposto na Definição 1.6, admita $\mathcal{E}$ com distribuição normal, já esteja adequado. Assim, tal modelo pode ser utilizado com a finalidade de se fazer predições.

Um problema abordado é predizer o valor da média das variáveis dependentes correspondente a um valor fixo $\mathbf{X}_0$ das variáveis independentes. Inferências sobre a média das variáveis dependentes pode ser feita usando-se dos resultados obtidos na Proposição 1.26.

A partir deste resultado pode-se verificar que a v.a. $\hat{\mathcal{B}}'\mathbf{X}_0$ possui distribuição normal multivariada $N_m(\mathcal{B}'\mathbf{X}_0, \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\boldsymbol{\Sigma})$ que é independente da v.a. $n\hat{\boldsymbol{\Sigma}}$ que apresenta distribuição $W_{r,n-r-1}$.

Proposição 1.28. *A região a* $100(1-\alpha)\%$ *de confiança para* $\mathcal{B}'\mathbf{X}_0$ *é dada por*

$$(\mathcal{B}'\mathbf{X}_0 - \hat{\mathcal{B}}'\mathbf{X}_0)' \left(\frac{n}{n-r-1}\hat{\boldsymbol{\Sigma}}\right)^{-1} (\mathcal{B}'\mathbf{X}_0 - \hat{\mathcal{B}}'\mathbf{X}_0) \leq \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0 \times \left(\frac{m(n-r-1)}{n-r-m} F_{m,n-r-m}(\alpha)\right), \quad (1.35)$$

onde $F_{m,n-r-m}(\alpha)$ *é o* (100α)*-ésimo percentil de uma distribuição* F*-Snedecor com* m *e* $n-r-m$ *graus de liberdade.*

Demonstração: Com base nos resultados apresentados na Proposição 1.26, sabe-se que $\hat{\mathcal{B}}'\mathbf{X}_0$ possui distribuição $N_m[\mathcal{B}'\mathbf{X}_0, \mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\boldsymbol{\Sigma}]$ e é independente de $n\hat{\boldsymbol{\Sigma}}$ que possui distribuição $W_{r,n-r-1}$.

Define-se a estatística T^2-Hotelling como segue

$$T^2 = \left(\frac{\hat{\mathcal{B}}'\mathbf{X}_0 - \mathcal{B}'\mathbf{X}_0}{\sqrt{\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0}}\right)' \left(\frac{n\hat{\boldsymbol{\Sigma}}}{n-r-1}\right)^{-1} \left(\frac{\hat{\mathcal{B}}'\mathbf{X}_0 - \mathcal{B}'\mathbf{X}_0}{\sqrt{\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0}}\right),$$

onde T^2 é produto entre um vetor aleatório normal multivariado e o inverso de uma matriz aleatória com distribuição de Wishart com suas entradas divididas pelo grau de liberdade.

Dessa forma, por definição, T^2 possui distribuição $\frac{m(n-r-1)}{n-r-m}F_{m,n-r-m}$ e, assim, a região a $100(1-\alpha)\%$ de confiança é dada por (1.35).

A estatística T^2-Hotelling é a generalização multivariada da distribuição t-Student e para maiores detalhes sobre suas propriedades recomenda-se o Capítulo V de Johnson e Wichern (1998) e o Capítulo VI de Chatfield e Collins (1980). □

1.4 Análise da Variância e Análise de Resíduos

Nesta seção apresentam-se ferramentas matemáticas e estatísticas capazes de fornecer uma análise dos estimadores e dos resultados obtidos em um modelo de regressão linear com base nos dados amostrais utilizados.

Inicialmente apresenta-se uma discussão a respeito do modelo linear simples e *a posteriori* são apresentados comentários análogos aos casos múltiplo e multivariado.

Análise da Variância e de Resíduos para o Modelo Linear Simples

Após estimar os parâmetros, tanto para o Caso A quanto para o Caso B, em um modelo linear, deve-se avaliar, então, a significância das variáveis no modelo. Uma forma de testar a significância do coeficiente de uma variável independente em um modelo é comparar os valores observados para a v.a. dependente com dois modelos auxiliares, um em que conste a variável independente em questão e outro que não a inclua.

A função matemática usada para comparar os valores preditos e observados depende das particularidades do problema em análise. Se os valores observados com o uso da variável no modelo são melhores ou, em algum sentido, mais acurados do que quando a variável não está no modelo, então percebe-se que a variável em questão é significante.

Na regressão linear, a avaliação do significado do coeficiente angular, β_1, por exemplo, é realizada através da tabela da análise da variância, chamada ANOVA (abreviação da expressão *Analysis of Variance*, ou seja, *Análise da Variância*). Esta tabela apresenta a soma total dos quadrados dos desvios em relação à média de forma parcelada. Uma parcela corresponde à soma dos quadrados dos desvios com respeito à linha de regressão, $SQRes$. A outra parcela, diz respeito à soma dos quadrados dos valores preditos, baseados no modelo de regressão, usando-se da média da v.a. dependente, $SQReg$.

No modelo de regressão linear, a comparação entre os valores observados e medidos é baseada no quadrado da distância entre os dois. Se y_i denota o valor observado e $\hat{y}_i$ denota o valor predito para a i-ésima observação (ou seja, $\hat{y}_i = \hat{\mu}(x_i)$), então a estatística utilizada para avaliar esta comparação é

$$SQRes = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = \hat{\mathcal{E}}'\hat{\mathcal{E}}, \tag{1.36}$$

que possui distribuição qui-quadrado com $n-2$ graus de liberdade.

No modelo que não inclui a variável independente em questão, o único parâmetro é β_0, e assim, $\hat{\beta}_0 = \bar{y}$, que é a média da variável resposta. Neste caso, $\hat{y}_i = \bar{y}$ e $SQRes$ equivale à variância total, $SQTotal$. Onde $\mathcal{B} = (\beta_0, \beta_1)$.

Onde $SQTotal$ fica definido como

$$SQTotal = (\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{Y} - \bar{\mathbf{Y}}),$$

que possui distribuição qui-quadrado com $n-1$ graus de liberdade.

Quando se inclui a variável independente no modelo, um pequeno decréscimo no valor de SQRes é esperado, devido ao fato de que $\beta_1 \neq 0$.

A mudança no valor de $SQRes$ é denotada por $SQReg$, e, desta forma tem-se a seguinte igualdade

$$SQReg = (\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{Y} - \bar{\mathbf{Y}}) - (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = SQTotal - SQRes, \quad (1.37)$$

que apresenta distribuição qui-quadrado com $n-2$ graus de liberdade.

Observação 1.11. Quanto maior for o módulo de $\hat{\beta}_1$, maior será a redução da soma dos quadrados dos resíduos. No modelo de regressão linear um grande valor de $SQReg$ sugere que a variável independente é importante enquanto que um valor pequeno indica que tal variável não é útil na formulação do modelo.

Pode-se resumir todas estas observações em uma tabela ANOVA, ilustrada, a seguir, pela Tabela 1.1. Onde F.V. significa fonte de variação; g.l., graus de liberdade; SQ, soma de quadrados; QM, quadrados médios e F é estatística de teste que segue distribuição F-Snedecor com 1 e $n-2$ graus de liberdade.

Tabela 1.1: Tabela ANOVA para o modelo de regressão linear simples.

F.V.	g.l.	SQ	QM	F
Regressão	1	$SQReg$	$SQReg$	$\frac{SQReg}{\hat{\sigma}^2_\varepsilon}$
Resíduo	$n-2$	$SQRes$	$\frac{SQRes}{n-2}$	
Total	$n-1$	$SQTotal$	$\frac{SQTotal}{n-1}$	

Proposição 1.29. *Um estimador não viciado para σ^2_ε é dado por $\frac{SQRes}{n-2}$.*

Demonstração: Esta prova segue dos resultados apresentados no Teorema 1.1 e da expressão (1.36) □

Na proposição a seguir, apresenta-se um resultado relacionado com a estatística F apresentada na tabela ANOVA.

Proposição 1.30. *A estatística* $F = \frac{SQReg}{\hat{\sigma}_\varepsilon^2}$ *possui distribuição* F*-Snedecor com* 1 *e* $n-2$ *graus de liberdade.*

Demonstração: Sabe-se que $(n-2)\hat{\sigma}_\varepsilon^2$ possui distribuição qui-quadrado com $(n-2)$ graus de liberdade e, também, que $SQReg$ possui distribuição qui-quadrado com 1 grau de liberdade. Dessa forma, efetuando-se o quociente

$$\frac{\frac{SQReg}{1}}{\frac{(n-2)\hat{\sigma}_\varepsilon^2}{n-2}} = \frac{SQReg}{\hat{\sigma}_\varepsilon^2},$$

segue, por definição, que $\frac{SQReg}{\hat{\sigma}_\varepsilon^2}$ possui distribuição F-Snedecor com 1 e $n-2$ graus de liberdade. □

Também pode-se medir o lucro relativo ao se introduzir a variável independente no modelo através da estatística R^2, que é definida como

$$R^2 = \frac{SQReg}{SQTotal}. \tag{1.38}$$

Dessa forma, R^2 mede a proporção da variação total da média $\bar{\mathbf{Y}}$ explicada pela regressão e, por isso, $0 \leq R^2 \leq 1$. É frequentemente expressa como uma porcentagem.

Sobre a estatística R^2 é interessante notar que:

(i) Se todas as observações pertencem à linha de regressão (situação de regressão perfeita) então $SQRes = 0$ e, desta forma, $R^2 = 1$.

(ii) Se a linha de regressão é horizontal (não existe contribuição de X na predição de Y) então $SQRes = SQTotal$ e $R^2 = 0$.

(iii) A estatística R^2 mede a relação linear entre X e Y.

Dois resultados interessantes a respeito da estatística R^2 são apresentados a seguir.

Proposição 1.31. *A estatística* $R^2 = \frac{SQReg}{SQTotal}$ *possui distribuição* $B(\frac{1}{2}, \frac{n-2}{2})$*, ou seja, distribuição beta com parâmetros* $\frac{1}{2}$ *e* $\frac{n-2}{2}$*.*

Demonstração: Para a verificação desta informação são necessários dois fatos importantes. Primeiro, sabe-se que uma v.a. com distribuição qui-quadrado com n graus de liberdade, χ^2_2, é equivalente a uma v.a. com distribuição gama com parâmetros $\frac{n}{2}$ e 2, $\Gamma(\frac{n}{2}, n)$. Segundo, por propriedade de operações entre variáveis aleatórias, tem-se que se A e B são variáveis aleatórias, respectivamente, com distribuições $\Gamma(a, c)$ e $\Gamma(b, c)$ então a v.a. $\frac{A}{A+B}$ possui distribuição beta com parâmetros a e b, $B(a, b)$.

Ainda tem-se que $SQTotal = SQRes + SQReg$. Assim, ao se definir $A = SQRes$ que possui distribuição χ^2_{n-2} e, conseqüentemente, distribuição $\Gamma(\frac{n-2}{2}, 2)$, e $B = SQReg$ que possui distribuição χ^2_1 e, conseqüentemente, distribuição $\Gamma(\frac{1}{2}, 2)$, segue o resultado. □

Proposição 1.32. *A estatística* R^2 *é equivalente ao quadrado do coeficiente de Pearson entre* $\hat{y}_i$ *e* y_i*, para* $i \in \{1, \cdots, n\}$*.*

Demonstração: O coeficiente de correlação de Pearson entre U e V é definido como

$$r_{UV} = \frac{\hat{\sigma}_{UV}}{\hat{\sigma}_U \hat{\sigma}_V}, \tag{1.39}$$

onde

$$\begin{aligned}
\hat{\sigma}_{UV} &= \frac{\sum_{i=1}^n (u_i - \bar{u})(v_i - \bar{v})}{n-1}, \\
\hat{\sigma}_U &= \sqrt{\frac{\sum_{i=1}^n (u_i - \bar{u})^2}{n-1}}, \\
\hat{\sigma}_V &= \sqrt{\frac{\sum_{i=1}^n (v_i - \bar{v})^2}{n-1}}.
\end{aligned}$$

Fazendo-se $U = \hat{Y}$ e $V = Y$ na expressão (1.39), após algumas simplificações, consegue-se

$$\begin{aligned}
r_{\hat{Y}Y} &= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \\
&= \frac{\sum_{i=1}^n (\hat{y}_i - \hat{y}_i + \hat{y}_i - \bar{y})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \\
&= \frac{\sum_{i=1}^n (y_i - \hat{y})(\hat{y}_i - \bar{y}) + \sum_{i=1}^n (\hat{y}_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}.
\end{aligned} \tag{1.40}$$

Pode-se verificar que a parcela $\sum_{i=1}^{n}(y_i-\hat{y})(\hat{y}_i-\bar{y})$ que aparece na expressão (1.40) é nula, resultando

$$\begin{aligned} r_{\hat{Y}Y} &= \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})(\hat{y}_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}. \\ &= \frac{\sqrt{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})(\hat{y}_i-\bar{y})}}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}, \end{aligned} \tag{1.41}$$

elevando-se ambos os lados da equação (1.41) ao quadrado tem-se o resultado. Este fato é verdadeiro para qualquer regressão linear com qualquer número de variáveis independentes envolvidas. □

Uma estatística relacionada com R^2, chamada de R^2-ajustado, denotada por R_a^2 é definida como

$$R_a^2 = 1-(1-R^2)\left(\frac{n-1}{n-2}\right). \tag{1.42}$$

Observação 1.12. O "ajuste" é feito para que haja correspondência entre os graus de liberdade de $SQTotal$ e $SQRes$. A estatística R_a^2 possui valor relativamente próximo ao de R^2.

É possível expressar as estatísticas R^2 e F uma em função da outra, como se apresenta na proposição a seguir.

Proposição 1.33. *As estatísticas F e R^2, podem ser expressas, respectivamente, como segue*

$$F = \frac{(n-2)R^2}{1-R^2} \tag{1.43}$$

$$R^2 = \frac{F}{F+n-2}. \tag{1.44}$$

Demonstração: Basta substituir as expressões (1.37) e (1.38) na definição de F e após algumas simplificações se obtém (1.43) e posteriormente se isola R^2 no resultado obtido para se verificar (1.44). □

Historicamente, as tabelas de valores da distribuição F-Snedecor eram mais difundidas do que as tabelas da distribuição beta. Desta forma, testes de hipóteses utilizando-se a estatística R^2 raramente eram realizados.

Para se verificar a adequabilidade de um modelo é necessário investigar se as suposições feitas para sua formulação estão sendo satisfeitas. Para tal, estuda-se o

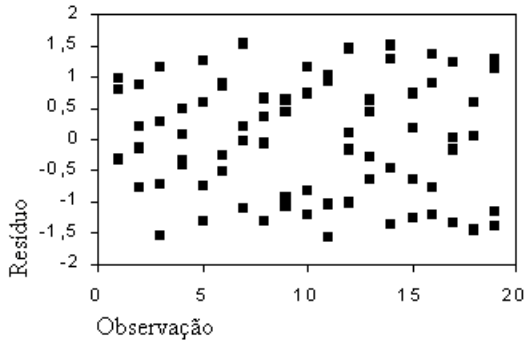

Figura 1.1: Situação ideal

comportamento do modelo com base no conjunto de dados observados, verificam-se discrepâncias entre os valores observados e os valores ajustados pelo modelo, ou melhor, faz-se uma *análise de resíduos*.

O i-ésimo resíduo é dado por $\hat{\varepsilon}_i = y_i - \hat{y}_i$, para $i \in \{1, \cdots, n\}$. O objetivo nesta etapa é estudar o comportamento individual e o conjunto dos resíduos, com base nas suposições feitas sobre os verdadeiros erros ε_i.

Uma representação gráfica bastante útil é obtida plotando-se os pares $(x_i, \hat{\varepsilon}_i)$, para $i \in \{1, \cdots, n\}$. Em outras situações, é de maior utilidade fazer a representação gráfica dos chamados *resíduos padronizados*,

$$\hat{z}_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}_\varepsilon} = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_\varepsilon},$$

plotando-se, então, os pares $(x_i, \hat{z}_i)$, para $i \in \{1, \cdots, n\}$. Estes dois gráficos terão formas semelhantes, havendo apenas mudança de escala das suas ordenadas.

Outro resíduo utilizado é o chamado *resíduo estudentizado*, definido por

$$\hat{r}_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_\varepsilon \sqrt{1 - v_i}}, \tag{1.45}$$

onde $v_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. O denominador na expressão (1.45) corresponde ao desvio-padrão de ε_i, para $i \in \{1, \cdots, n\}$.

De posse do gráfico dos resíduos, precisa-se identificar possíveis inadequações. As Figuras 1.1, 1.2, 1.3, 1.4 e 1.5 a seguir ilustram alguns tipos de gráficos de resíduos. A Figura 1.1 representa a situação ideal para os resíduos, ou seja, distribuídos aleatoriamente ao redor do zero, sem nenhuma observação discrepante.

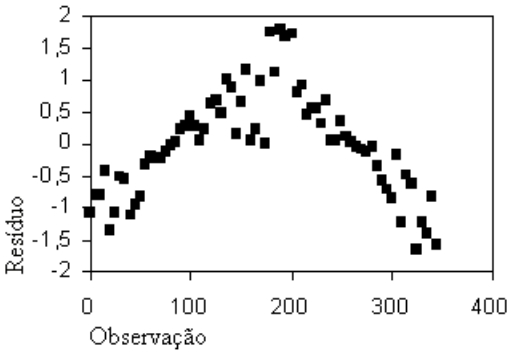

Figura 1.2: Modelo inadequado

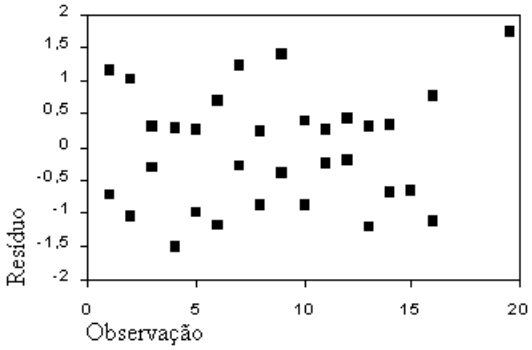

Figura 1.3: Elemento discrepante

Na situação apresentada pela Figura 1.2 existe inadequação do modelo ajustado e, ainda, a curva sugere que é necessário procurar outras funções matemáticas que representem melhor o fenômeno em análise.

A Figura 1.3 representa a situação em que existe um elemento discrepante, e é interessante investigar a razão deste fato. Pode ser um erro de medida ou ser efetivamente real. Em situações como essa, em que há observações muito diferentes das demais, métodos chamados robustos devem ser utilizados (para tal indica-se o Capítulo VI de Draper e Smith (1981)). A Figura 1.4 indica que a suposição de mesma variância, ou seja, homocedasticidade, não está sendo satisfeita. Uma das hipóteses, em modelos de regressão, é a de homocedasticidade pois as variâncias são constantes em relação ao tempo t, ou seja, vale o item (iii) associado à expressão (1.2). No entanto, muitas vezes não se pode garantir, *a priori* esta suposição. Pode-se, então, fazer uma inspeção visual ou um teste para se obter

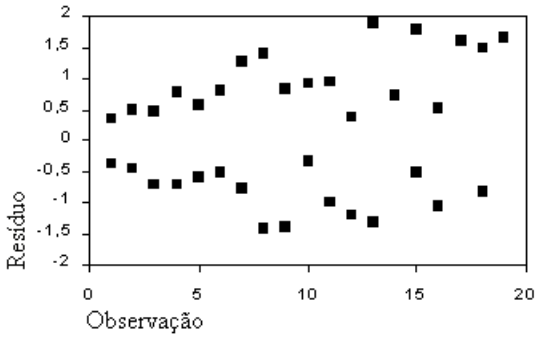

Figura 1.4: Heterocedasticidade

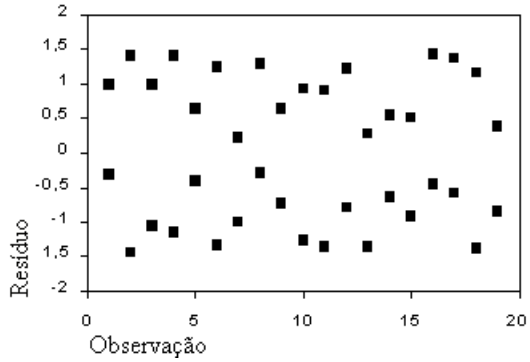

Figura 1.5: Não-normalidade

uma resposta referente a validade ou não desta hipótese. Um teste utilizado neste contexto é o teste de Bartlett, para detalhes sugere-se Dixon e Massey (1957).

Na Figura 1.5, parece haver maior incidência de observações nos extremos, mostrando que a suposição de normalidade não está sendo satisfeita. A verificação da hipótese de normalidade pode ser realizada fazendo-se um histograma dos resíduos ou um gráfico de quantis. Para detalhes a respeito da construção de histogramas e de gráficos de quantis recomenda-se o Capítulo III de Bussab e Morettin (2004).

O teste de Bartlett, pode ser resumido em cinco etapas apresentadas a seguir, onde n_l representa o número de unidades observadas para o nível l e $\hat{\sigma}^2_\varepsilon(l)$ a variância amostral do nível l, para $l \in \{1, \cdots, I\}$, com $n = n_1 + n_2 + \cdots + n_I$.

(i) calcule a variância amostral comum $\hat{\sigma}^2_\varepsilon$;

(ii) calcule $M = (n-I)\ln\hat{\sigma}^2_\varepsilon - \sum_{l=1}^{I}(n_l-1)\ln\hat{\sigma}^2_\varepsilon(l)$;

(ii) calcule $M = (n-I)\ln\hat{\sigma}^2_\varepsilon - \sum_{l=1}^{I}(n_l-1)\ln\hat{\sigma}^2_\varepsilon(l)$;

(iii) calcule $C = 1 + \frac{1}{3(I-1)}\left[\sum_{l=1}^{I}\frac{1}{n_l-1} - \frac{1}{n-I}\right]$;

(iv) construa a estatística M/C, que segue uma distribuição aproximada qui-quadrado, com $I-1$ graus de liberdade, para amostras grandes;

(v) compare o p valor da estatística M/C com o valor tabelado de uma distribuição qui-quadrado com $I-1$ graus de liberdade à $\alpha = 5\%$ de significância. (1.46)

Análise da Variância e de Resíduos para o Modelo Linear Múltiplo

Uma tabela ANOVA, semelhante à apresentada na Tabela 1.1, pode ser aplicada sem dificuldades ao modelo de regressão múltipla, que se encontra abaixo na Tabela 1.2.

As estatísticas $SQReg$, $SQRes$ e $SQTotal$ são definidas da mesma forma que na seção precedente, ou seja, são dadas pela expressões (1.37) e (1.36).

Tabela 1.2: Tabela ANOVA para o modelo de regressão linear múltipla.

F.V.	g.l.	SQ	QM	F
Regressão	$r-1$	$SQReg$	$\frac{SQReg}{r-1}$	$\frac{SQReg/(r-1)}{SQRes/(n-r)}$
Resíduo	$n-r$	$SQRes$	$\frac{SQRes}{n-r}$	
Total	$n-1$	$SQTotal$	$\frac{SQTotal}{n-1}$	

É possível apresentar um estimador não viaciado para σ^2_ε de forma semelhante à apresentada na regressão linear simples.

Proposição 1.34. *Um estimador não viciado para σ^2_ε é dado por $\frac{SQRes}{n-r-1}$.*

Demonstração: A prova desta afirmação segue raciocínio análogo ao realizado na Proposição 1.29. □

Também é possível utilizar a ANOVA para verificar se $\beta_j = 0$ para algum $j \in \{0, \cdots, r\}$ da mesma forma feita anteriormente. Entretanto, quando há muitas variáveis preditoras no modelo é usual apresentar sua contribuição pela média da chamada soma de quadrados extra, $SQExtra$. Uma $SQExtra$ mede a redução marginal na $SQRes$ quando uma ou mais variáveis preditoras são acrescentadas ao modelo.

Define-se $SQExtra(x_l|x_j)$, por exemplo, como sendo a medida do aperfeiçoamento do modelo que já possui a variável x_j pela adição da variável x_l, com $l \neq j$,

$$\begin{aligned} SQExtra(x_l|x_j) &= SQRes(x_j) - SQRes(x_l, x_j) \\ &= SQReg(x_l, x_j) - SQReg(x_j). \end{aligned}$$

A $SQExtra$ é especialmente utilizada em testes de *performance* dos coeficientes de regressão para se detectar situações de multicolinearidade. Para maiores detalhes recomenda-se o Capítulo VII de Marques de Sá (2003).

A seguir apresenta-se, respectivamente, um análogo à Proposição 1.30 e à Proposição 1.31 para o contexto da regressão linear múltipla.

Proposição 1.35. *A estatística* $F = \frac{SQReg/(r-1)}{SQRes/(n-r)}$ *possui distribuição F-Snedecor com* $r-1$ *e* $n-r$ *graus de liberdade.*

Demonstração: A prova desta afirmação segue procedimento análogo ao desenvolvido na Proposição 1.30. □

Proposição 1.36. *A estatística* R^2 *possui distribuição* $B(\frac{r-1}{2}, \frac{n-r}{2})$*, ou seja, distribuição beta com parâmetros* $\frac{r-1}{2}$ *e* $\frac{n-r}{2}$*.*

Demonstração: A verificação deste resultado com argumentos semelhantes aos utilizados na verificação da Proposição 1.31. □

Um análogo da Proposição 1.33 também é válido e é apresentado na seqüência.

Proposição 1.37. *As estatísticas* F *e* R^2 *podem ser expressas como funções uma da outra conforme segue*

$$\begin{aligned} F &= \frac{(n-r-1)R^2}{r(1-R^2)} \\ R^2 &= \frac{rF}{rF+n-r-1}, \end{aligned}$$

onde r *representa o número de variáveis independentes no modelo.*

Demonstração: A prova é semelhante a realizada na Proposição 1.33. □

A estatística R^2-ajustado é definida por

$$R_a^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-r}\right). \tag{1.47}$$

Há outra estatística relacionada com R^2, chamada *estatística C_p-Mallow*, definida por

$$C_p = \frac{SQRes_p}{\hat{\sigma}_\varepsilon^2} - (n - 2p),$$

onde $SQRes_p$ é a soma dos quadrados dos resíduos contendo p parâmetros (sempre incluindo β_0).

Uma descrição completa sobre a estatística C_p-Mallow é obtida no Capítulo VI de Draper e Smith (1981).

Se o modelo é válido, cada resíduo $\hat{\varepsilon}_i$ é um estimador do erro ε_i, para o qual se assume ter distribuição normal com média zero e variância σ_ε^2. Dessa forma, o resíduo $\hat{\mathcal{E}}$ possui valor esperado $\mathbf{0}$ e sua matriz de variâncias e covariâncias é dada como na Proposição 1.18, ou seja,

$$Var(\hat{\mathcal{E}}) = \sigma_\varepsilon^2[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \sigma_\varepsilon^2(\mathbf{I} - \mathbf{H}). \tag{1.48}$$

Pelo fato do vetor de resíduos $\hat{\mathcal{E}}$ apresentar matriz de variâncias-covariâncias dada por (1.48), as variâncias dos $\hat{\varepsilon}_i$ podem variar muito se os elementos da diagonal de $\mathbf{H}$, denotados por h_{ii}, forem substancialmente diferentes.

De maneira análoga à apresentada na seção precedente, definem-se os *resíduos padronizados* como

$$\hat{z}_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}_\varepsilon} = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_\varepsilon}$$

e os *resíduos estudentizados* como

$$\hat{r}_i = \frac{\hat{\varepsilon}_i}{\sqrt{\hat{\sigma}_\varepsilon^2(1 - h_{ii})}},$$

para $i \in \{1, \cdots, n\}$.

Os resíduos podem ser apresentados através de gráficos de diferentes maneiras para se detectar anomalias, conforme já discutido anteriormente.

1.5 Aplicação da Teoria

Nesta seção apresenta-se uma aplicação do modelo de regressão linear através de um exemplo. Os dados numéricos e informações que constituem o exemplo foram extraídos da página 413 de Bussab e Morettin (2004).

Exemplo: Um psicólogo está investigando a relação entre o tempo que um indivíduo leva para reagir a um estímulo visual e alguns fatores, como idade e acuidade visual. Definem-se as variáveis Y como sendo o tempo de resposta ao estímulo, X_1 como a idade e X_2 como a acuidade visual (medida em porcentagem). Na Tabela 1.3, têm-se os tempos de resposta ao estímulo visual com relação à idade, sexo e acuidade visual para vinte indivíduos.

Tabela 1.3: Tempos de reação a um estímulo (Y) e acuidade visual (X_2) de vinte indivíduos, segundo a idade (X_1).

Indivíduo	Y	X_1	X_2	Indivíduo	Y	X_1	X_2
1	96	20	90	11	109	30	90
2	92	20	100	12	100	30	80
3	106	20	80	13	112	35	90
4	100	20	90	14	105	35	80
5	98	25	100	15	118	35	70
6	104	25	90	16	108	35	90
7	110	25	80	17	113	40	90
8	101	25	90	18	112	40	90
9	116	30	70	19	127	40	60
10	106	30	90	20	117	40	80

Inicialmente, busca-se um modelo que relacione a v.a. dependente Y, tempo de reação, e a variável independente X_1, idade. Para tal, traça-se o gráfico de dispersão de idade e reação ao estímulo, representado aqui pela Figura 1.6. Neste gráfico, aparece também a reta ajustada.

A Figura 1.6 mostra que os dados observados apresentam uma tendência crescente que pode ser representada por uma reta, ou seja, o tempo médio de reação pode ser entendido como uma função linear da idade.

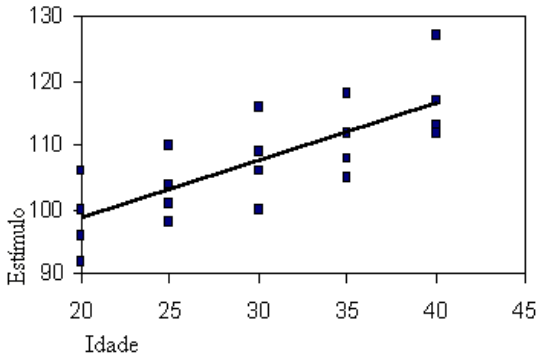

Figura 1.6: Gráfico de dispersão de idade e reação ao estímulo, com a reta ajustada.

Tanto X_1 (idade) como Y (tempo de resposta ao estímulo) são v.a.'s contínuas, assim, com base no contexto da Definição 1.1, um modelo razoável para descrever a $\mathbb{E}(Y|x_1)$ é dado pela expressão (1.1).

O modelo pode ser escrito como

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i, \quad i \in \{1, \cdots, 20\}, \tag{1.49}$$

onde

$$\mathbb{E}(\varepsilon_i|x_{i1}) = 0 \ \text{ e } \ Var(\varepsilon_i|x_{i1}) = \sigma^2_\varepsilon,$$

devendo-se encontrar os prováveis valores para β_0 e β_1, segundo critérios apresentados na Seção 1.1 deste capítulo. Ou seja, buscam-se os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$, que são os mesmos tanto para o Caso A quanto para o Caso B.

Em sua forma matricial, o modelo dado por (1.49), assume a seguinte configuração

$$\mathbf{Y} = \mathbf{X}_1\mathcal{B} + \mathcal{E}, \quad \mathbb{E}(\mathcal{E}|\mathbf{X}_1) = \mathbf{0} \ \text{ e } \ Var(\mathcal{E}|\mathbf{X}_1) = \sigma^2_\varepsilon, \tag{1.50}$$

onde tem-se que $\mathbf{X}_1$ é uma matriz 20×2, $\mathcal{B}$ é uma matriz 2×1 e $\mathcal{E}$ é uma matriz 20×1.

Para ajustar o modelo (1.50), com $\mathbf{Y}$ sendo o vetor de apresenta os tempos de reação dos 20 indivíduos, $\mathbf{X}_1$ sendo a matriz que fornece as idades dos indivíduo, $\mathcal{E}$ sendo o vetor erro das observações, precisa-se utilizar os resultados apresentados no Teorema 1.1.

Sendo que

$$\mathbf{X}_1 = \begin{bmatrix} 1 & x_{11} \\ \vdots & \dots \\ 1 & x_{1,20} \end{bmatrix}.$$

Da Tabela 1.3 obtém-se a informação

$$\mathbf{X}_1'\mathbf{X}_1 = \begin{bmatrix} 20 & 600 \\ 600 & 19000 \end{bmatrix},$$

que implica em

$$(\mathbf{X}_1'\mathbf{X}_1)^{-1} = \begin{bmatrix} 0,95 & -0,03 \\ -0,03 & 0,001 \end{bmatrix}. \tag{1.51}$$

Substituindo-se o resultado (1.51) no resultado fornecido pelo Teorema 1.1, resulta

$$\begin{aligned} \hat{\mathcal{B}} &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y} \\ &= \begin{bmatrix} 0,95 & -0,03 \\ -0,03 & 0,001 \end{bmatrix} \mathbf{X}_1\mathbf{Y} \\ &= \begin{bmatrix} 0,90 \\ 80,50 \end{bmatrix} \end{aligned} \tag{1.52}$$

Com o resultado encontrado para $\hat{\mathcal{B}}$ em (1.52), segue-se que o modelo dado por (1.50) assume a seguinte configuração

$$\hat{\mathbf{Y}} = \mathbf{X} \begin{bmatrix} 0,90 \\ 80,50 \end{bmatrix} + \mathcal{E}. \tag{1.53}$$

É interessante ressaltar que o modelo (1.50) parece adequado, não apresentando nenhum ponto discrepante. Entretanto, apresenta-se, na seqüência, a avaliação deste modelo através da sua respectiva tabela ANOVA.

Utilizam-se, novamente os dados correspondentes às colunas Y e X_1 da Tabela 1.3 e o exemplo da tabela ANOVA, apresentado na Tabela 1.1, onde

$$\begin{aligned} SQRes &= (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) \\ SQTotal &= (\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{Y} - \bar{\mathbf{Y}}) \\ SQReg &= SQtotal - SQRes \\ F &= \frac{SQReg}{\hat{\sigma}_\varepsilon^2}. \end{aligned} \tag{1.54}$$

Assim, obtém-se a Tabela 1.4.

Tabela 1.4: Tabela ANOVA para o modelo dado por (1.50).

F.V.	g.l.	SQ	QM	F
Regressão	1	810	810	25,89
Resíduo	18	563	31,28	
Total	19	1373	72,26	

Com base nos dados da Tabela 1.4, pode-se obter a estatística R^2 dada pela expressão (1.38), ou seja, $R^2 = \dfrac{810}{1373} = 0,5899$, ou ainda, $R^2 = 58,99\%$.

Desta informação pode-se concluir que o modelo proposto diminui a variância residual em mais da metade. Na coluna QM da Tabela 1.4, o valor de 72,26 em contraste com 31,28 explica aproximadamente 59% da variabilidade. A estatística R_a^2, dada pela expressão (1.42), assume valor $R_a^2 = 0,57$, que é um valor relativamente próximo do encontrado para R^2.

Conclui-se assim, que é vantajoso adotar o modelo de regressão linear simples para explicar o tempo médio de reação ao estímulo em função da idade. Onde o valor obtido para $\hat{\beta}_1 = 0,90$ indica que o acréscimo em uma unidade na idade do indivíduo observado, provoca aumento no tempo de resposta ao estímulo em 0,90 unidade.

Da Tabela 1.4, com base na Proposição 1.29, pode-se retirar a informação $\dfrac{SQRes}{n-2} = \hat{\sigma}_\varepsilon^2 = 31,28$, implicando em $\hat{\sigma}_\varepsilon = 5,59$.

De posse deste último resultado é possível apresentar intervalos a $100(1-\alpha)\%$ de confiança para os parâmetros β_0 e β_1. Utilizam-se $\alpha = 0,05$ como coeficiente de confiança e as Proposições 1.8 e 1.9. Denota-se $t_{n-2,\alpha}$ o valor obtido em uma tabela da distribuição t-Student com coeficiente de confiança α e $n-2$ $(20-2=18)$ graus de liberdade. Assim, $t_{18;0,05} = 2,101$.

O intervalo a 95% de confiança para β_0, denotado por $IC(\beta_0, 95\%)$, é dado por

$$\begin{aligned} IC(\beta_0, 95\%) &= 80,50 \pm 2,101 \times 5,59\sqrt{\frac{19000}{20 \times 1000}} \\ &= 80,50 \pm 11,45, \end{aligned}$$

ou seja,

$$IC(\beta_0, 95\%) = [69,05; 91,95]. \tag{1.55}$$

O intervalo a 95% de confiança para β_1, denotado por $IC(\beta_1, 95\%)$, é dado por

$$\begin{aligned} IC(\beta_1, 95\%) &= 0,90 \pm 2,101 \times 5,59\sqrt{\frac{1}{1000}} \\ &= 0,90 \pm 0,30, \end{aligned}$$

ou seja,

$$IC(\beta_0, 95\%) = [0,60; 1,20]. \tag{1.56}$$

É interessante notar que o intervalo (1.56) não contém o zero e esta é uma evidência de que $\beta_1 \neq 0$. Os intervalos de confiança obtidos em (1.55) e (1.56) podem ser utilizados para testar hipóteses.

Testa-se, separadamente, por exemplo, as hipóteses apresentadas em (1.9) para $b_1 = b_2 = 0$. Em ambos os casos, rejeita-se que os parâmetros sejam iguais à zero. Com base nos Teoremas 1.12 e 1.13 e na Proposição 1.30, pode-se utilizar a estatística F que aparece na tabela ANOVA para testar as hipóteses (1.9) para $b_1 = b_2 = 0$.

No contexto do exemplo em estudo, pode-se estar interessado em saber qual o tempo de reação para um indivíduo de 28 anos. Aqui é importante ficar claro qual é o objetivo: estimar o tempo médio para o grupo etário de 28 anos ou o tempo de reação provável para uma pessoa de 28 anos.

A estimativa pontual é a mesma nos dois casos, mas a estimativa por intervalo de confiança é distinta. Para uma discussão sobre tais diferenças recomenda-se o Capítulo XV de Bussab e Morettin (2004).

A estimativa pontual dada pelo modelo (1.49) para o tempo de reação de um indivíduo de 28 anos é dada por

$$\hat{y}(28) = 80,5 + 0,9 \times 28 = 105,7.$$

Com o resultado acima, pode-se, então, calcular o intervalo de confiança para a média através da Proposição 1.11, do Corolário 1.3 e da Observação 1.1. Para tal, utilizam-se os seguintes valores

$$\hat{y}(28) = 105,7, \quad t_{18;0,05} = 2,101, \quad \hat{\sigma}^2_\varepsilon = 5,59,$$
$$n = 20, \quad \bar{x}_1 = 30 \quad \text{e} \quad \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2 = 1000$$

e obtém-se que

$$\begin{aligned} IC(\mu(28), 5\%) &= 105,7 \pm 2,101 \times 5,59\sqrt{\frac{1}{20} + \frac{(28-30)^2}{1000}} \\ &= 105,7 \pm 2,7 \end{aligned}$$

ou ainda

$$IC(\mu(28), 5\%) = [103; 108,4]. \tag{1.57}$$

O intervalo de confiança apresentado em (1.57) pode ser interpretado como o intervalo em que 95% das amostras coletadas pode-se encontrar o verdadeiro valor do tempo de resposta ao estímulo visual para um indivíduo com 28 anos de idade.

Os resíduos $\hat{\varepsilon}_i$, com $i \in \{1, \cdots, 20\}$, para o modelo (1.50) são apresentados na Tabela 1.5, a seguir, bem como os resíduos padronizados, $\hat{z}_i$, e os resíduos estudentizados, $\hat{r}_i$.

Tabela 1.5: Resíduos para o modelo dado por (1.50).

Idade	$\hat{\varepsilon}_i$	$\hat{z}_i$	$\hat{r}_i$	Idade	$\hat{\varepsilon}_i$	$\hat{z}_i$	$\hat{r}_i$
20	-2,5	-0,45	-0,49	30	1,5	0,27	0,28
20	-6,5	-1,16	-1,26	30	-7,5	-1,34	-1,37
20	7,5	1,34	1,45	35	0,0	0,0	0,0
20	1,5	0,27	0,29	35	-7,0	-1,25	-1,30
25	-5,0	-0,89	-0,92	35	6,0	1,07	1,11
25	1,0	0,18	0,19	35	-4,0	-0,72	-0,75
25	7,0	1,25	1,30	40	-4,5	-0,80	-0,86
25	-2,0	-0,36	0,37	40	-5,5	-0,98	-1,06
30	8,5	1,52	1,56	40	9,5	1,70	1,84
30	-1,5	-0,27	-0,28	40	-0,5	-0,09	-0,10

Os resíduos e os resíduos padronizados são apresentados, respectivamente, nas Figuras 1.7 e 1.8.

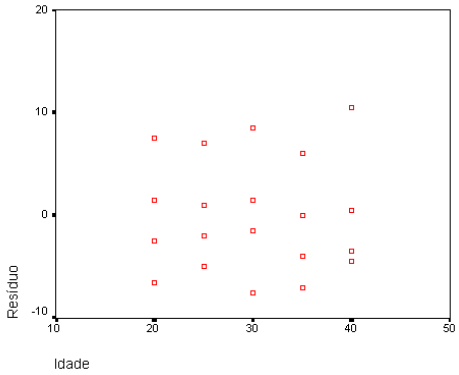

Figura 1.7: Gráfico de resíduo versus idade para o modelo dado por (1.50).

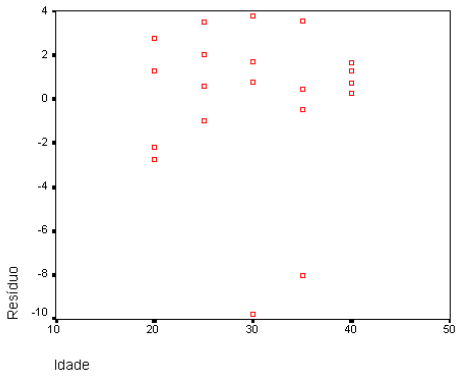

Figura 1.8: Gráfico de resíduo padronizado versus idade para o modelo dado por (1.50).

A análise dos resíduos permite concluir que as suposições de média zero e variância comum são satisfeitas.

A Figura 1.10 apresenta o histograma dos resíduos e a Figura 1.9 mostra um gráfico quantil-quantil. Em ambas percebe-se que os resíduos não são normalmente distribuídos.

Para se verificar a validade da suposição de mesma variância aplica-se o Teste de Bartlett. Para se efetuar este teste reorganizam-se as informações que se encontram na Tabela 1.3 de forma que os dados sejam fornecidos por faixa etária e de cada faixa etária se obtém sua respectiva variância, conforme sugere a Tabela 1.6.

Para efetuar o teste de Bartlett, seguem-se os passos $(i) - (v)$ dados em (1.46),

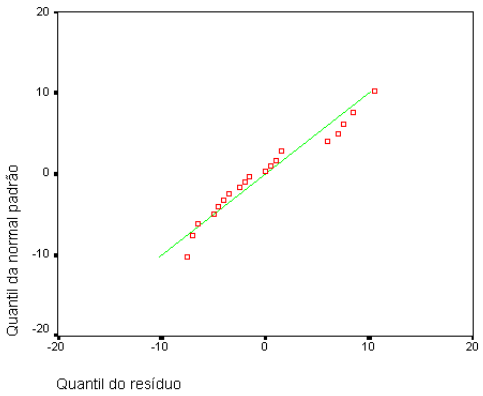

Figura 1.9: Gráfico quantil×quantil para o modelo dado por (1.50).

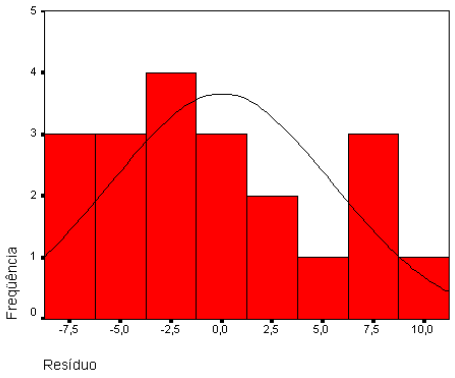

Figura 1.10: Histograma de resíduo para o modelo dado por (1.50).

Tabela 1.6: Tabela da variância em função do grupo etário.

Grupo etário	20	25	30	35	40
Tamanho da amostra	4	4	4	4	4
Variância	35,67	26,25	44,25	31,58	46,92

com base nos dados apresentados na Tabela 1.6. Dessa forma, obtêm-se os seguintes resultados

$$
\begin{aligned}
&(i) \quad \hat{\sigma}_{\varepsilon}^2 = 36,93 \\
&(ii) \quad M = (20-5)\ln(36,93) - [3\ln(35,67) + 3\ln(26,25) + \\
&\qquad 3\ln(44,25) + 3\ln(31,58) + 3\ln(46,92)] = 0,3369 \\
&(iii) \quad C = 1 + \frac{1}{3\times 4}\left[\frac{1}{3}+\frac{1}{3}+\frac{1}{3}+\frac{1}{3}+\frac{1}{3}-\frac{1}{15}\right] = 1,1333 \\
&(iv) \quad \frac{M}{C} = \frac{0,3369}{1,1333} = 0,2972 \\
&(v) \quad \frac{M}{C} < 9,488 = \chi^2_{4;0,05}.
\end{aligned}
\tag{1.58}
$$

Decorre, das etapas apresentadas em (1.58), que não se rejeita a hipótese de igualdade das variâncias.

Agora, busca-se um modelo que relacione a v.a. dependente Y, tempo de reação, e as variáveis independentes X_1, idade, e X_2, acuidade visual. Ou seja, assume-se o contexto da regressão linear múltipla.

Tanto a idade X_1 quanto a acuidade visual X_2 são v.a.'s contínuas. Assim, com base no contexto da Definição 1.4, um modelo razoável para descrever a relação entre tais variáveis é

$$
y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \quad i \in \{1, \cdots, 20\}, \tag{1.59}
$$

onde

$$
\mathbb{E}(\varepsilon_i|\mathbf{x}_i) = 0, \quad Var(\varepsilon_i|\mathbf{x}_i) = \sigma_{\varepsilon}^2 \quad \text{e} \quad \mathbf{x}_i = (1, x_{i1}, x_{i2}).
$$

Devem-se encontrar os prováveis valores para β_0, β_1 e β_2, ou então para $\mathcal{B} = (\beta_0, \beta_1, \beta_2)$, segundo os critérios apresentados na seção 1.2 deste capítulo. Ou seja, buscam-se os estimadores $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$, que são os mesmos tanto para o Caso A quanto para o Caso B.

Para ajustar o modelo dado por (1.59), com y_i sendo o tempo de reação, x_{i1} a idade e x_{i2} a acuidade visual do i-ésimo indivíduo, precisa-se utilizar os resultados teóricos apresentados no Teorema 1.17.

Nesta etapa, então, evidencia-se o uso de um pacote computacional (SPSS 10.0) para a obtenção de informações e estatísticas que auxiliem a análise.

O modelo dado por (1.59) parece estar adequado. No entanto, apresenta-se sua avaliação com base em sua respectiva tabela ANOVA.

Utilizam-se os dados correspondentes às colunas Y, X_1 e X_2 da Tabela 1.3 e o exemplo de tabela ANOVA para o caso da regressão linear múltipla apresentado na Tabela 1.2. A Tabela 1.7 apresenta a análise da variância para este caso.

Tabela 1.7: Tabela ANOVA para o modelo dado por (1.59).

F.V.	g.l.	SQ	QM	F
Regressão	2	1139,03	569,51	41,38
Resíduo	17	233,97	13,76	
Total	19	1373	72,26	

Com base na Tabela 1.7 e pela Proposição 1.37 pode-se obter a estatística R^2, como segue

$$\begin{aligned} R^2 &= \frac{rF}{rF+n-r-1} \\ &= \frac{2\times 41,38}{2\times 41,38+20-2-1} \\ &= 0,8295, \end{aligned}$$

ou ainda, $R^2 = 82,95\%$.

Desta informação pode-se concluir que o modelo proposto diminui a variância residual em "quase" sua totalidade. Na coluna QM da Tabela 1.7, o valor 72,26 em contraste com 13,76, explica aproximadamente 83% da variabilidade. A estatística R_a^2, para este caso, tem valor 0,81, ou seja, possui valor muito próximo de R^2.

Conclui-se, assim, a princípio, que é vantajoso adotar o modelo de regressão linear múltiplo para explicar o tempo médio de reação ao estímulo em função da

idade e da acuidade visual.

Ao se fazer o contraste das tabelas ANOVA dadas pelas Tabelas 1.4 e 1.7, percebe-se que o modelo linear múltiplo é vantajoso também em relação ao modelo de regressão linear simples, pois como comentado no início deste exemplo o modelo (1.49) explica apenas 59% da variabilidade. Entretanto, vale comentar que nem sempre a inserção de novas variáveis contribui para a sua significativa melhora.

Os valores estimados para os parâmetros β_0, β_1 e β_2 e seus respectivos intervalos a 95% de confiança são apresentados na tabela abaixo.

Tabela 1.8: Intervalos a 95% de confiança para $\mathcal{B}$ para o modelo dado por (1.59).

Estimador	Valor Estimado	Intervalo de Confiança
$\hat{\beta}_0$	126,564	[105,27;147,85]
$\hat{\beta}_1$	0,650	[0,38;0,92]
$\hat{\beta}_2$	-0,454	[-0,65;-0,25]

Todos os intervalos apresentados na Tabela 1.8 não contém o número zero, esta é uma evidência de que $\mathcal{B} \neq \mathbf{0}$.

Os resíduos $\hat{\varepsilon}_i$, com $i \in \{1, \cdots, 20\}$, para o modelo dado por (1.59) são apresentados na Tabela 1.9, a seguir, bem como os resíduos padronizados, $\hat{z}_i$, e os resíduos estudentizados, $\hat{r}_i$. Os resíduos e os resíduos padronizados são apresentados graficamente nas Figuras 1.11 e 1.12, abaixo.

A análise dos resíduos permite concluir que as suposições de média zero e variância comum são satisfeitas.

A Figura 1.13 representa o histograma dos resíduos e a Figura 1.14 mostra um gráfico quantil-quantil. Em ambas as situações percebe-se que os resíduos não são normalmente distribuídos. Tal fato pode ser justificado devido ao pequeno número de dados observados, apenas vinte; e, pela possível falta de outras variáveis independentes que poderiam descrever melhor o problema, tais como, por exemplo, sexo, pressão ocular e índice de álcool no organismo.

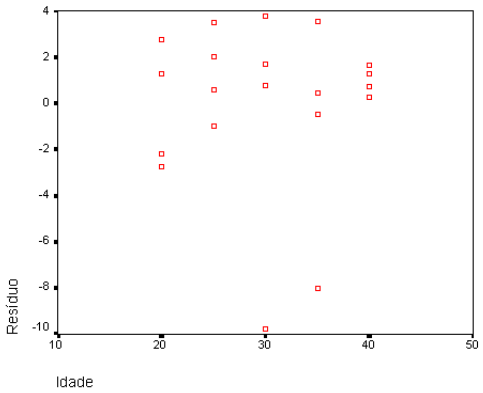

Figura 1.11: Gráfico de resíduo versus idade para o modelo dado por (1.59).

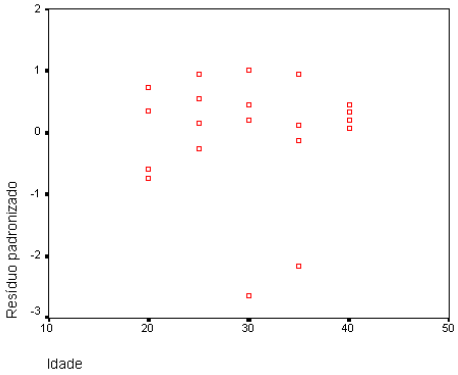

Figura 1.12: Gráfico de resíduo padronizado versus idade para o modelo dado por (1.59).

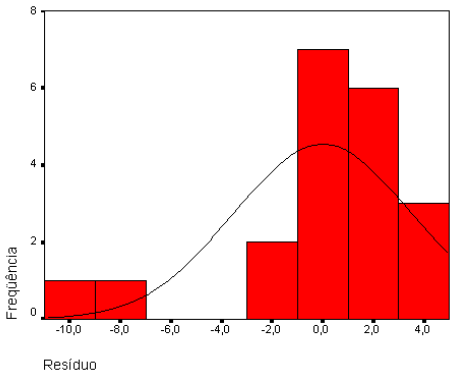

Figura 1.13: Histograma de resíduo para o modelo dado por (1.59).

Tabela 1.9: Resíduos para o modelo dado por (1.59).

Idade	$\hat{\varepsilon}_i$	$\hat{z}_i$	$\hat{r}_i$	Idade	$\hat{\varepsilon}_i$	$\hat{z}_i$	$\hat{r}_i$
20	-2,72	-0,73	-0,79	30	3,76	1,01	1,05
20	-2,18	-0,58	-0,66	30	-9,76	-2,63	-2,72
20	2,73	0,73	0,83	35	3,51	0,94	1,01
20	1,27	0,34	0,37	35	-8,02	-2,16	-2,25
25	0,55	0,15	0,16	35	0,44	0,11	0,13
25	2,02	0,54	0,56	35	-0,48	-0,13	-0,13
25	3,48	0,93	0,99	40	1,26	0,34	0,38
25	-0,97	-0,26	-0,27	40	0,26	0,07	0,08
30	1,69	0,45	0,50	40	1,65	0,44	0,56
30	0,76	0,20	0,21	40	0,72	0,19	0,21

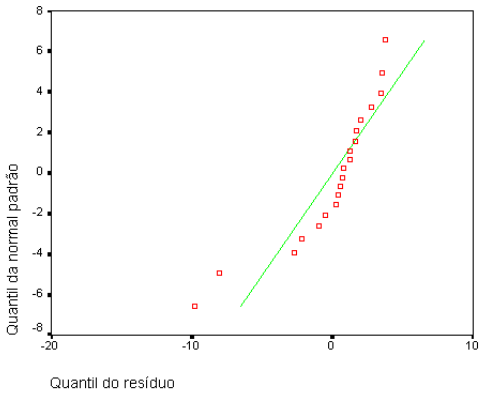

Figura 1.14: Gráfico quantil×quantil para o modelo dado por (1.59).

Capítulo 2

Regressão Logística

Nos modelos de regressão linear simples ou múltipla, a variável dependente Y é uma variável aleatória de natureza contínua. No entanto, em algumas situações, a variável dependente é qualitativa e expressa por duas ou mais categorias, ou seja, admite dois ou mais valores. Neste caso, o método dos mínimos quadrados não oferece estimadores plausíveis. Uma boa aproximação é obtida pela regressão logística que permite o uso de um modelo de regressão para se calcular ou prever a probabilidade de um evento específico.

As categorias (ou valores) que a variável dependente assume podem possuir natureza nominal ou ordinal. Em caso de natureza ordinal, há uma ordem natural entre as possíveis categorias e, então tem-se o contexto da Regressão Logística Ordinal. Quando esta ordem não existe entre as categorias da variável independente assume-se o contexto da Regressão Logística Nominal.

Neste trabalho aborda-se apenas o caso da Regressão Logística Nominal. Detalhes e informações sobre a Regressão Logística Ordinal podem ser encontrados em Kleinbaum e Klein (2002), Agresti (1984) e Agresti (1990).

Inicialmente apresenta-se exemplo que ilustra uma situação em que a variável dependente possui natureza nominal. Outros exemplos podem ser obtidos em Hosmer e Lemeshow (2000), Draper e Smith (1981), Pampel (2000) e Menard (2002).

Exemplo 2.1. Suponha que se deseje estudar a toxicidade de uma certa droga. Neste contexto, dosagens $x_1 < x_2 < \cdots < x_p$ são fixadas. A dosagem x_i geralmente é medida como o logaritmo na base dez da concentração da droga em uma solução

e é administrada em uma quantidade c_i de animais. Após este procedimento, ocorre um número p_i de mortes para cada i, com $1 \leq i \leq n$. Assume-se que $\pi(x)$ é a probabilidade que um animal escolhido aleatoriamente sucumba com a dosagem x. Dessa forma, p_i, $1 \leq i$, são v.a.'s independentes com distribuição binomial $B(c_i, \pi(x_i))$, com $i \in \{1, \cdots, n\}$. O objetivo aqui é encontrar um modelo no qual, para cada valor da variável independente x_i, é possível predizer a variável dependente $p(x_i)$, a qual é binomial com probabilidade de sucesso $\pi(x_i)$.

2.1 Regressão Logística Binária

Nesta seção apresenta-se o contexto em que a variável resposta possui apenas duas categorias, ou seja, natureza binária ou dicotômica, e apenas uma variável independente envolvida.

Antes de se iniciar a discussão sobre a Regressão Logística, é interessante fazer um breve comentário sobre Modelos Lineares Generalizados. Maiores detalhes podem ser obtidos em Agresti (1984) e Agresti (1990).

Um modelo linear generalizado (m.l.g.) é especificado por três componentes: uma componente aleatória, a qual identifica à distribuição de probabilidade da variável dependente, uma componente sistemática, que especifica uma função linear entre as variáveis independentes e uma função de ligação que descreve a relação matemática entre a componente sistemática e o valor esperado da componente aleatória.

Em outras palavras, a componente aleatória de um m.l.g. consiste nas observações da variável aleatória Y, ou seja com o vetor $\mathbf{y} = (y_1, \cdots, y_n)$.

A componente sistemática do m.l.g. é definida através de um vetor $\eta = (\eta_1, \cdots, \eta_n)$ que está associado ao conjunto das variáveis independentes através de um modelo linear $\eta = \mathbf{X}\mathcal{B}$, onde $\mathbf{X}$ é uma matriz que consiste nas variáveis independentes das n observações e $\mathcal{B}$ é um vetor de parâmetros do modelo.

A terceira componente do m.l.g. é a função de ligação entre as componentes aleatória e sistemática. Seja $\mu_i = \mathbb{E}(Y_i|x_i)$, com $i \in \{1, \cdots, n\}$ então η_i é definida por $\eta_i = g(\mu_i)$, onde g é uma função monotônica e diferenciável.

Dessa forma, a função de ligação conecta os valores esperados das observações

às variáveis explanatórias, para $i \in \{1, \cdots, n\}$, através da fórmula

$$g(\mu_i) = \sum_{j=1}^{n} \beta_j x_{ij}. \tag{2.1}$$

É interessante comentar que se a função g, dada por (2.1), for a função identidade tem-se então o modelo de regressão linear.

Há duas classes importantes de modelos lineares generalizados. Uma delas é constituída pelos modelos *logit*, nos quais a variável dependente pode ser associada a uma variável aleatória Bernoulli (aqui se enquadra a regressão logística), e pelos modelos *loglinear* no qual a variável dependente é associada a uma variável aleatória Poisson. Para maiores detalhes sobre os modelos *loglinear* recomenda-se Ardersen (1996), Agresti (1984) e Agresti (1990).

Na seqüência apresenta-se o modelo de regressão logística binária, que é um caso particular dos modelos lineares generalizados, mas especificamente dos modelos *logit*.

Para se analisar $\pi(x)$, tomam-se as observações independentes $x_1, \cdots, x_n$. Neste contexto, é razoável assumir, como suposição inicial, que $\pi(x)$ é uma função monotônica com valores entre zero e um, quando x varia na reta real, ou seja $\pi(x)$ é uma função de distribuição de probabilidade.

Como $\pi(\cdot)$ varia entre zero e um, uma representação linear simples para π sobre todos os valores possíveis de x é não é adequada (pois, caso contrário, tem-se o modelo de regressão linear), então considera-se a transformação logística de $\pi(\cdot)$ sob a forma linear

$$\ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = g(x), \tag{2.2}$$

onde

$$g(x) = \beta_0 + \beta_1 x \tag{2.3}$$

ou equivalentemente,

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}, \tag{2.4}$$

onde é necessário que $\beta_1 < 0$ para que π seja crescente e que $\beta_1 > 0$ para que π seja decrescente. Quando x tende ao infinito, $\pi(x)$ tende a zero quando $\beta_1 < 0$ e tende a um quando $\beta_1 > 0$. Assim, dessa forma, define-se a função de ligação necessária ao modelo. Caso $\beta_1 = 0$, a variável de resposta Y é independente da variável X.

Se na equação (2.4), tem-se $\beta_0 = 0$ e $\beta_1 = -1$ então $\pi(x)$ é chamada de *função de distribuição logística*. Sua respectiva *função de distribuição acumulada* (f.d.a.) é dada por

$$F(x) = \frac{\exp(-x)}{1 + \exp(-x)}. \tag{2.5}$$

E, conseqüentemente, sua *função de densidade de probabilidade* (f.d.p.), obtida por derivação de (2.5), é dada por

$$f(x) = \frac{\exp(-x)}{[1 + \exp(-x)]^2}.$$

Proposição 2.1. *Caso haja reparametrização, em (2.5), dada por*

$$\mu = \frac{-\beta_0}{\beta_1} \ \ e \ \ \sigma = \frac{-1}{\beta_1},$$

tem-se que $\pi(\cdot)$ dada pela expressão (2.4) assume a forma

$$\pi(x) = F\left(\frac{x - \mu}{\sigma}\right).$$

Demonstração: Primeiramente analisa-se a reparamentrização

$$\frac{x - \mu}{\sigma} = \frac{x + \frac{\beta_0}{\beta_1}}{\frac{-1}{\beta_1}} = -\beta_0 - \beta_1 x.$$

Dessa forma, pelas expressões (2.3), (2.4) e (2.5), tem-se que

$$F\left(\frac{x - \mu}{\sigma}\right) = F[-g(x)] = \pi(x).$$

□

Observação 2.1. Diz-se que μ é o *centro de simetria* e que σ é o *parâmetro de escala* da reparametrização.

O exposto acima motiva a seguinte definição.

Definição 2.1. *Modelo Logístico:* Seja $g(x) = \beta_0 + \beta_1 x$, para qualquer $x \in \mathbb{R}$. Seja $F(\cdot)$ a f.d.a. definida em (2.5) correspondente a $\beta_0 = 0$ e $\beta_1 = -1$. Ao reparametrizar-se $\mu = -\beta_0(\beta_1)^{-1}$ e $\sigma = -(\beta_1)^{-1}$, para cada x_i, para $i \in \{1, \cdots, n\}$, segue-se que

$$\begin{aligned} F\left(\frac{x_i - \mu}{\sigma}\right) &= \pi(x_i) \\ \mathbb{E}(Y_i|x_i) &= \pi(x_i), \end{aligned} \tag{2.6}$$

onde y_i representa uma amostra aleatória de tamanho um de $F(\cdot)$.

Na seqüência, apresenta-se algumas diferenças entre o modelo de regressão logístico e o modelo de regressão linear.

O que distingue um modelo de regressão logístico simples e binário de um modelo de regressão linear simples é a variável aleatória dependente. No modelo logístico ela se apresenta na escala nominal. Esta diferença se reflete na escolha do modelo paramétrico e nas hipóteses envolvidas.

Em qualquer problema de regressão, a quantidade ou grandeza chave é a esperança da v.a. dependente dado o valor da variável independente, ou melhor, $\mathbb{E}(Y|x)$. Em virtude do exposto nos parágrafos anteriores, ao se trabalhar com dados de natureza binária, a esperança condicional deve ser menor ou igual a um e maior ou igual a zero, $0 \leq \mathbb{E}(Y|x) \leq 1$.

Uma transformação de $\pi(x)$ que é central para o estudo da regressão logística é a transformação *logit*. Esta transformação é definida em termos de $\pi(x)$ em (2.2).

A função $g(x)$, definida em (2.3), é denominada de função *logit* e tem muitas das propriedades da função $\mu(\cdot)$ do modelo linear pois é linear em seus parâmetros e contínua.

Os parâmetros β_0 e β_1 tem significados similares aos seus análogos na regressão linear. Se $x = 0$ na expressão (2.4) segue-se que a expressão (2.2) assume valor β_0.

Agora, efetuando-se o mesmo procedimento para x e $x+1$, para qualquer $x \in \mathbb{R}$, é possível obter, pela expressão (2.2),

$$\begin{aligned} \ln\left(\frac{\pi(x+1)}{1-\pi(x+1)}\right) - \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) &= \beta_0 + \beta_1(x+1) - (\beta_0 + \beta_1 x) \\ &= \beta_1. \end{aligned} \tag{2.7}$$

Então, β_1 é o incremento no valor da expressão (2.2) (*log odds*) devido ao aumento de uma unidade em x. E, β_0 corresponde a (*log odds*) de "sucesso" contra "fracasso" no caso em que $x = 0$.

Calculando a exponencial de (2.7), tem-se

$$\exp(\beta_1) = \frac{\dfrac{\pi(x+1)}{1-\pi(x+1)}}{\dfrac{\pi(x)}{1-\pi(x)}}. \tag{2.8}$$

A expressão do lado direito em (2.8) é conhecida como *razão de chances* (*odds ratio*) e compara a probabilidade de sucesso para $x+1$ com a probabilidade de sucesso para x.

Na regressão logística esta razão é uma função constante quando analisada em relação à variável x. De (2.8), finalmente, pode-se obter que

$$\frac{\pi(x+1)}{1-\pi(x+1)} = \exp(\beta_1)\frac{\pi(x)}{1-\pi(x)},$$

ou seja, $\exp(\beta_1)$ é a mudança multiplicativa nas probabilidades de sucesso, correspondente ao aumento de uma unidade em x.

Outra diferença importante entre o modelo de regressão linear e o modelo de regressão logístico é o que diz respeito à distribuição condicional da v.a. Y. No modelo linear assume-se que uma observação possa ser expressa como $y = \mathbb{E}(Y|x) + \varepsilon$. A hipótese mais comum é a de que ε possui distribuição normal com média zero e variância que é constante para todas as possibilidades da variável Y. Deste fato, segue que a distribuição condicional de Y dado x possui distribuição normal com média $\mathbb{E}(Y|x)$ e variância constante. Este não é o caso quando a v.a. Y possui natureza binária. Neste contexto, pode-se expressar uma observação como $y = \pi(x) + \varepsilon$. A v.a. ε pode assumir uma de duas possibilidades. Se $y = 1$ então $\varepsilon = 1-\pi(x)$ com probabilidade $\pi(x)$, e se $y = 0$ então $\varepsilon = -\pi(x)$ com probabilidade $1-\pi(x)$.

Proposição 2.2. *A v.a. ε tem distribuição Bernoulli com média zero e variância $\pi(x)[1-\pi(x)]$.*

Demonstração: Primeiramente avalia-se a esperança de ε dado que $X = x$

$$\mathbb{E}(\varepsilon|x) = \sum_{k=1}^{2} \varepsilon_k \mathbb{P}(\varepsilon = \varepsilon_k) = -\pi(x)(1-\pi(x)) + (1-\pi(x))\pi(x) = 0.$$

Agora, procura-se avaliar a variância condicional de ε,

$$\begin{aligned} Var(\varepsilon|x) &= \sum_{k=1}^{2} \varepsilon_k^2 \mathbb{P}(\varepsilon = \varepsilon_k) = [-\pi(x)]^2(1-\pi(x)) + [1-\pi(x)]^2\pi(x) \\ &= \pi(x) - 2\pi^2(x) + \pi^3(x) + \pi^2(x) - \pi^3(x) = \pi(x) - \pi^2(x). \end{aligned}$$

□

Observação 2.2. Esta proposição indica que independentemente dos erros serem grandes ou pequenos, pode-se esperar que sua média seja nula.

Decorre da Definição 2.1 que as afirmações na expressão (2.6) podem ser interpretadas, para cada $i \in \{1, \cdots, n\}$, como

$$y_i = \frac{\exp[g(x_i)]}{1 + \exp[g(x_i)]} + \varepsilon_i,$$

onde os erros, ε_i, seguem as seguintes suposições, para todo $i, l \in \{1, \cdots, n\}$

$$\begin{aligned} &(i) \quad \mathbb{E}(\varepsilon_i | x_i) = 0. \\ &(ii) \quad Var(\varepsilon_i | x_i) = \pi(x_i)[1 - \pi(x_i)]. \\ &(iii) \quad Cov(\varepsilon_i, \varepsilon_l) = 0, \text{ se } i \neq l. \end{aligned} \tag{2.9}$$

A expressão (2.9) motiva as seguintes suposições para a as v.a.'s envolvidas:

(i) A variável X é por hipótese controlada e não está sujeita a variações aleatórias.

(ii) Para um dado valor x de X, os erros ε têm distribuição Bernoulli com média zero e variância $\pi(x)[1 - \pi(x)]$.

Supõe-se uma amostra de n independentes observações de pares (x_i, y_i), para $i \in \{1, \cdots, n\}$, onde y_i denota o valor da variável aleatória binária e x_i, o valor da variável independente para a i-ésima observação. Assume-se também que a v.a. Y pode ser decodificada pelos valores 0 ou 1, representando, respectivamente, a ausência ou a presença da característica em estudo. Para se ajustar o modelo de regressão logística apresentado na Definição 2.1 a um conjunto de dados deve-se estimar os valores de β_0 e β_1, parâmetros não conhecidos.

No capítulo anterior utilizou-se o método dos mínimos quadrados para apresentar estimadores aos parâmetros envolvidos no modelo. Tais estimadores apresentam propriedades interessantes e importantes. Infelizmente, quando este método é aplicado a um modelo cuja variável dependente é de natureza nominal, os estimadores obtidos não conservam as mesmas propriedades. Entretanto para n suficientemente grande pode-se utilizar o método dos mínimos quadrados ponderados, para detalhes recomenda-se o Capítulo XIII de Agresti (1990).

Outro método de estimação utilizado conjuntamente ao modelo de regressão linear é o de máxima verossimilhança. Este método embasará a procura dos estimadores para o modelo de regressão logística.

Proposição 2.3. *Assume-se o contexto da Definição 2.1. Seja* $\mathcal{B} = (\beta_0, \beta_1)$ *o vetor de parâmetros relacionado com a probabilidade condicional* $\mathbb{P}(Y_i = 1|x_i) = \pi(x_i)$. *Então, o estimador, pelo método de máxima verossimilhanca, de* $\mathcal{B}$, *denotado por* $\hat{\mathcal{B}}$, *é a solução das equações de verossimilhança*

$$\sum_{i=1}^{n}[y_i - \pi(x_i)] = 0 \tag{2.10}$$

$$\sum_{i=1}^{n} x_i[y_i - \pi(x_i)] = 0. \tag{2.11}$$

Demonstração: Seguem-se da Definição 2.1 as probabilidades $\mathbb{P}(y_i = 1|x_i) = \pi(x_i)$ e $\mathbb{P}(y_i = 0|x_i) = 1 - \pi(x_i)$. Então, para os pares (x_i, y_i) tais que $y_i = 1$, a contribuição para a função de verossimilhança é $\pi(x_i)$, e para os pares tais que $y_i = 0$, a contribuição para a função de verossimilhança é $1 - \pi(x_i)$, onde a quantidade $\pi(x_i)$ denota o valor de $\pi(x)$ avaliado em x_i. Uma maneira conveniente de expressar estas contribuições à função de verossimilhança é

$$\pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i},$$

para valores $y_i = 0$ ou $y_i = 1$, para todo $i \in \{1, \cdots, n\}$.

Como assume-se que as observações são independentes, a função de verossimilhança obtida é dada por

$$L(\mathcal{B}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}. \tag{2.12}$$

O princípio da máxima verossimilhança atesta que o estimador $\hat{\mathcal{B}}$ é o valor que maximiza a expressão (2.12). Para encontrar este valor, diferencia-se a equação (2.12) com respeito a β_0 e β_1 e igualam-se as expressões resultantes a zero, obtendo-se, respectivamente, os resultados (2.10) e (2.11). □

Aplicando-se logaritmo em ambos os lados da equação (2.12), tem-se a expressão

$$\mathcal{L}(\mathcal{B}) = \sum_{i=1}^{n} \left[y_i \ln \pi(x_i) + (1 - y_i) \ln(1 - \pi(x_i)) \right]. \tag{2.13}$$

Observação 2.3. No modelo de regressão linear as equações de verossimilhança são facilmente resolvidas. Para o modelo de regressão logística, tais equações são não-lineares nos parâmetros e desta forma, requer-se o uso de um procedimento iterativo conhecido como o método de Newton-Raphson. De forma resumida, este

método pode ser descrito como segue. O primeiro passo requer o uso de uma solução inicial (candidato) para os valores que maximizam a função de verossimilhança. A função é aproximada, em uma vizinhança da solução inicial por um polinômio de segundo grau. A segunda solução obtida, no processo iterativo, é o ponto de máximo valor do polinômio, e assim por diante. Dessa forma o método de Newton-Raphson gera uma seqüência de soluções que convergem para o ponto de máximo da função de verosimilhança.

Observação 2.4. Pode-se provar que a expressão (2.13) é uma função estritamente convexa. Desta forma, as equações de verossimilhança admitem solução e ainda, esta solução é única. Ou seja, têm-se os estimadores de máxima verossimilhança.

Para maiores detalhes sobre as Observações 2.3 e 2.4 indica-se a leitura da seção 3 do Capítulo XII de Casella e Berger (2002).

Observação 2.5. Uma conseqüência que pode ser obtida da equação (2.10) é que $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i)$, significando que a soma dos valores observados de y é igual à soma dos valores preditos de y.

Um método geral para avaliar a significância das variáveis foi ilustrado no modelo linear, e vamos utilizá-lo para motivar a forma usada na regressão logística.

Na regressão linear, a avaliação do significado do coeficiente angular β_1, por exemplo, é realizada através da tabela da análise da variância.

No modelo de regressão logística o princípio é o mesmo. Entretanto, a comparação entre os valores observados e os valores preditos é baseada no logaritmo da função de máxima verossimilhança.

Para entender melhor esta comparação, é útil, de forma conceitual, pensar o valor observado da variável dependente como sendo uma variável preditora resultante de um modelo saturado (um modelo é dito saturado se contém tantos parâmetros quantos dados observados).

Um exemplo simples de modelo saturado é ajustar um conjunto de dados com apenas dois pontos amostrais ao modelo dado pela Definição 2.1.

Proposição 2.4. *Para o modelo saturado associado ao modelo proposto pela expressão (2.6) tem-se que* $\hat{\pi}(x_i) = y_i$, *com* $i \in \{1, 2\}$.

Demonstração: Assume-se o conjunto de dados amostrais dados pelos pontos (x_1, y_1) e (x_2, y_2). Procura-se estimadores para $\hat{\pi}(x_i)$, com $i \in \{1, 2\}$. Com base nas equações (2.10) e (2.11), tem-se o seguinte sistema

$$\begin{aligned} &y_1 - \hat{\pi}(x_1) + y_2 - \hat{\pi}(x_2) = 0 \\ &x_1 y_1 - x_1\hat{\pi}(x_1) + x_2 y_2 - x_2\hat{\pi}(x_2) = 0. \end{aligned} \tag{2.14}$$

Da primeira equação do sistema (2.14) obtém-se que

$$y_2 - \hat{\pi}(x_2) = -y_1 + \hat{\pi}(x_1). \tag{2.15}$$

Ao se substituir a expressão (2.15) na segunda equação do sistema (2.14), consegue-se

$$[y_1 - \hat{\pi}(x_1)](x_1 - x_2) = 0. \tag{2.16}$$

Como tem-se que x_1 e x_2 são quaisquer, a equação (2.16) será identicamente nula se $y_1 = \hat{\pi}(x_1)$.

Um desenvolvimento semelhante pode ser realizado para se verificar que $y_2 = \hat{\pi}(x_2)$. Ficando, assim, demonstrada a proposição. □

Proposição 2.5. *A função de verossimilhança do modelo saturado vale um.*

Demonstração: Segue-se do resultado da Proposição 2.4 e de um raciocínio semelhante ao efetuado na Proposição 2.3. Para os pares (x_i, y_i) tais que $y_i = 1$, a contribuição para a função de verossimilhança é $\hat{\pi}(x_i) = y_i = 1$ e para os pares tais que $y_i = 0$, a contribuição para função de verossimilhança é $1 - \hat{\pi}(x_i) = 1 - y_i = 1$. O que implica que a função de verossimilhança dada pela expressão (2.12) assume valor um. □

A comparação entre valores observados e preditos usando-se a função de verossimilhança é baseada na seguinte proposição.

Proposição 2.6. *A estatística D, definida por*

$$D = -2\ln\left[\frac{\textit{verossimilhança do modelo ajustado}}{\textit{verossimilhança do modelo saturado}}\right], \tag{2.17}$$

pode ser representada por

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1 - y_i)\ln\left(\frac{1 - \hat{\pi}_i}{1 - y_i}\right)\right], \tag{2.18}$$

onde $\hat{\pi}_i = \hat{\pi}(x_i)$.

Demonstração: A função de verossimilhança do modelo ajustado é obtida a partir de (2.12) utilizando-se $\hat{\pi}(x_i)$ no lugar de $\pi(x_i)$, e para o modelo saturado utiliza-se $\hat{\pi}(x_i) = y_i$.

Assim

$$D = -2\ln\left[\frac{\prod_{i=1}^{n}\hat{\pi}(x_i)^{y_i}(1-\hat{\pi}(x_i))^{1-y_i}}{\prod_{i=1}^{n}y_i^{y_i}(1-y_i)^{1-y_i}}\right],$$

que, por propriedades dos logaritmos e do produtório, pode ser reescrita como

$$\begin{aligned} D &= -2\sum_{i=1}^{n}\ln\left[\left(\frac{\hat{\pi}(x_i)}{y_i}\right)^{y_i}\left(\frac{1-\hat{\pi}(x_i)}{1-y_i}\right)^{1-y_i}\right] \\ &= -2\sum_{i=1}^{n}\left[y_i\ln\left(\frac{\hat{\pi}(x_i)}{y_i}\right)+(1-y_i)\ln\left(\frac{1-\hat{\pi}(x_i)}{1-y_i}\right)\right], \end{aligned} \tag{2.19}$$

denotando-se, por simplicidade, $\hat{\pi}(x_i) = \hat{\pi}_i$, segue o resultado. □

Observação 2.6. A estatística D é chamada de *deviance* e faz o mesmo papel que a $SQRes$ na regressão linear, ou seja, é uma estatística que auxilia na comparação entre os valores observados e preditos.

Da Proposição 2.6 pode-se extrair o seguinte corolário.

Corolário 2.7. *A expressão (2.18) pode ser representada pela forma*

$$D = -2\ln(\textit{verossimilhança do modelo ajustado}).$$

Demonstração: Decorre da Proposição 2.5 que a função de verossimilhança do modelo saturado quando a variável Y é dicotômica, ou seja, $y = 0$ ou $y = 1$, é igual a um. Por simples substituição em (2.17), segue o resultado. □

Designa-se D_0 para indicar a função de verossimilhança sem a variável independente.

Como Y é uma variável aleatória dicotômica, se $n_1 = \sum_{i=1}^{n} y_i$ e $n_0 = \sum_{i=1}^{n}(1 - y_i)$ e $\mathbb{P}(Y = 1) = \frac{n_1}{n}$, segue-se da expressão (2.19) que

$$D_0 = -2\left\{n_1\ln[\mathbb{P}(Y=1)] + n_0\ln[\mathbb{P}(Y=0)]\right\}.$$

É interessante notar que n_1, da forma com está definido, representa a contagem do número de vezes em que a v.a. Y assumiu valor um e, analogamente, n_0 representa o número de vezes em que a v.a. Y assumiu valor zero.

Assim, com o propósito de verificar a significância de uma variável independente, compara-se o valor de D com e sem tal variável na equação, representados respectivamente por D_M e D_0. A alteração no valor de D esperada pela inclusão da variável independente no modelo é obtida através de

$$G = D_0 - D_M. \tag{2.20}$$

Devido à definição da função de verossimilhança do modelo saturado (ela vale um) é comum expressar-se a estatística G por

$$G = -2\ln\left[\frac{\text{verossimilhança sem a variável}}{\text{verossimilhança com a variável}}\right]. \tag{2.21}$$

Para o específico caso de haver uma única variável independente, é fácil verificar que se a variável não está no modelo, o estimador de máxima verossimilhança de β_0 é $\hat{\beta}_0 = \ln\left(\frac{n_1}{n_0}\right)$, onde

$$n_1 = \sum_{i=1}^{n} y_i \quad \text{e} \quad n_0 = \sum_{i=1}^{n}(1 - y_i) \tag{2.22}$$

e o valor predito é constante, $\frac{n_1}{n}$. Neste caso, o valor da estatística G em (2.21) é dada por

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{1-y_i}}\right],$$

ou ainda,

$$G = 2\left\{\sum_{i=1}^{n}[y_i\ln(\hat{\pi}_i) + (1-y_i)\ln(1-\hat{\pi}_i)] - [n_1\ln(n_1) + n_0\ln(n_0) - n\ln(n)]\right\}.$$

Sob a hipótese de que β_1 é igual a zero, a estatística G apresenta distribuição assintótica qui-quadrado com 1 grau de liberdade, e este fato motiva o seguinte teorema.

Teorema 2.8. *No modelo dado pela Definição 2.1, o teste de razão de verossimilhança de tamanho α dado pela expressão*

$$\mathcal{H}_0 : \beta_1 = b_1 \ \ vs \ \ \mathcal{H}_1 : \beta_1 \neq b_1, \ \ para \ \ b_1 \in \mathbb{R},$$

consiste em rejeitar $\mathcal{H}_0$ se $\mathbb{P}[\chi_1^2 > G] < \alpha$.

Demonstração: Indicações sobre a validade desta afirmação podem ser obtidas nas páginas 14 e 15 de Hosmer e Lemeshow (2000). A prova de que a estatística G apresenta distribuição assintótica χ^2 será apresentada na Seção 2.4 deste capítulo. □

Existem outros testes, estatisticamente equivalentes ao apresentado no Teorema 2.8, bastante usados na literatura: Teste de Wald e Teste de Score. As hipóteses necessárias para estes dois testes são as mesmas utilizadas no teste da razão de verossimilhança. Apresentam-se, a seguir, alguns comentários sobre estes testes:

(i) O Teste de Wald é baseado na estatística

$$W = \frac{\hat{\beta}_1 - b_1}{\widehat{SE}(\hat{\beta}_1)},$$

ou seja, é obtido comparando-se o estimador de β_1, do método da máxima verossimilhança, com o seu erro padrão (isto é, seu desvio padrão estimado), denotado por $\widehat{SE}(\hat{\beta}_1)$. Sabe-se que sob $\mathcal{H}_0 : \beta_1 = 0$, a estatística W possui distribuição normal padrão. Hauck e Donner (1977) examinaram o Teste de Wald e observaram que ele tem um comportamento aberrante, ou melhor, freqüentemente falha na rejeição da hipótese nula quando o coeficiente de β_1 é significante. Jennings (1986) também verificou esta mesma propriedade. Por estas razões, estes autores recomendam o uso do teste da razão de verossimilhança.

(ii) O Teste de Score não exige o cálculo do estimador de máxima verossimilhança de β_1. No caso univariado, este teste é baseado na distribuição condicional da derivada da equação (2.11) dada a derivada da equação (2.10). Neste caso, pode-se apresentar uma expressão para o Teste de Score. Este teste usa o valor da equação (2.11) calculado para $\beta_0 = \ln\left(\frac{n_1}{n_0}\right)$ e $\beta_1 = 0$, obtendo-se assim, $\hat{\pi} = \bar{y} = \frac{n_1}{n}$. Então o lado esquerdo da equação (2.11) torna-se $\sum_{i=1}^{n} x_i(y_i - \bar{y})$. Pode-se mostrar que o estimador para a variância é $\bar{y}(1-\bar{y})\sum_{i=1}^{n}(x_i - \bar{x})^2$. Desta forma, a estatística para o Teste de Score é

$$S = \frac{\sum_i^n x_i(y_i - \bar{y})}{\sqrt{\bar{y}(1-\bar{y})\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Uma discussão completa sobre estes testes pode ser obtida em Rao (1973).

Um importante adicional à discussão feita anteriormente é obter e interpretar intervalos de confiança para os parâmetros de interesse. Como no caso da regresssão linear, pode-se obter os intervalos de confiança para β_0, β_1, $g(x)$ e $\pi(x)$.

A base teórica para a construção dos intervalos de estimação é a mesma usada para formular o teste de significância do modelo. Os intervalos de confiança para os estimadores de β_o e β_1 são baseados em seus respectivos Teste de Wald, como afirmam as proposições a seguir.

Proposição 2.9. *O intervalo a* $100(1-\alpha)\%$ *de confiança para* β_0 *é dado por*

$$[\hat{\beta}_0 - z_{\frac{\alpha}{2}}\widehat{SE}(\hat{\beta}_0), \hat{\beta}_0 + z_{\frac{\alpha}{2}}\widehat{SE}(\hat{\beta}_0)],$$

onde $z_{\frac{\alpha}{2}}$ *é o quantil de uma normal padrão dado por* $\mathbb{P}(z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$ *e* $\widehat{SE}(\hat{\beta}_0)$ *denota o estimador baseado no desvio padrão de* $\hat{\beta}_0$.

Proposição 2.10. *O intervalo a* $100(1-\alpha)\%$ *para* β_1 *é dado por*

$$[\hat{\beta}_1 - z_{\frac{\alpha}{2}}\widehat{SE}(\hat{\beta}_1), \hat{\beta}_1 + z_{\frac{\alpha}{2}}\widehat{SE}(\hat{\beta}_1)],$$

onde $z_{\frac{\alpha}{2}}$ *é o quantil de uma normal padrão dado por* $\mathbb{P}(z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$ *e* $\widehat{SE}(\hat{\beta}_1)$ *denota o estimador baseado no desvio padrão de* $\hat{\beta}_1$.

Como o *logit* é a parte linear do modelo de regressão logístico, ele pode ser estimado, semelhantemente ao feito no modelo de regressão linear, por $\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$.

O estimador da variância de $\hat{g}(x)$, representado por $\widehat{Var}[\hat{g}(x)]$, requer a obtenção da variância da soma. Neste caso

$$\widehat{Var}[\hat{g}(x)] = \widehat{Var}(\hat{\beta}_0) + x^2\widehat{Var}(\hat{\beta}_1) + 2x\,\widehat{Cov}(\hat{\beta}_0, \hat{\beta}_1), \tag{2.23}$$

pois $Var(X+Y) = Var(X) + Var(Y) + 2\,Cov(X,Y)$, para quaisquer v.a.'s X e Y. Os valores de $Cov(\hat{\beta}_0, \hat{\beta}_1)$ são obtidos, no método de máxima verossimilhança, através da matriz de variâncias-covariâncias que corresponde ao inverso da Informação de Fisher. Para maiores detalhes recomenda-se o Capítulo VII de Kleinbaum e Klein (2002) e Agresti (1990). Dessa forma, segue-se a seguinte proposição.

Proposição 2.11. *O intervalo a* $100(1-\alpha)\%$ *de confiança para* $g(x)$ *é dado por*

$$\left[\hat{g}(x) - z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{g}(x)]}, \hat{g}(x) + z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{g}(x)]}\right],$$

onde $z_{\frac{\alpha}{2}}$ *é o quantil de uma normal padrão dado por* $\mathbb{P}(z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$ *e* $\widehat{Var}[\hat{g}(x)]$ *denota o estimador de* $Var[\hat{g}(x)]$ *dado pela expressão (2.23).*

Demonstração: Decorre das Proposições 2.9 e 2.10. □

Proposição 2.12. *O intervalo a* $100(1-\alpha)\%$ *de confiança para* $\pi(x)$ *é dado por*

$$\frac{\exp\left[\hat{g}(x) - z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{g}(x)]}\right]}{1+\exp\left[\hat{g}(x) - z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{g}(x)]}\right]}, \frac{\exp\left[\hat{g}(x) + z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{g}(x))}\right]}{1+\exp\left[\hat{g}(x) + z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{g}(x)]}\right]},$$

onde $z_{\frac{\alpha}{2}}$ *é o quantil de uma normal padrão dado por* $\mathbb{P}(z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$ *e* $\widehat{Var}[\hat{g}(x)]$ *denota o estimador de* $Var[\hat{g}(x)]$ *dado pela expressão (2.23).*

Demonstração: A partir do resultado da Proposição 2.11, pode-se obter as seguintes expressões

$$\exp\left[\hat{g}(x) + z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{g}(x)]}\right] \geq \exp[g(x)] \quad \geq \exp\left[\hat{g}(x) - z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{g}(x)]}\right] \tag{2.24}$$

e

$$1+\exp\left[\hat{g}(x) + z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{g}(x)]}\right] \geq 1+\exp[g(x)] \quad \geq 1+\exp\left[\hat{g}(x) - z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{g}(x)]}\right]. \tag{2.25}$$

Como a função $f(x) = \exp(x)$ é sempre positiva, para qualquer $x \in \mathbb{R}$, então ao se dividir a expressão (2.24) por (2.25), obtém-se

$$\frac{\exp\left[\hat{g}(x) + z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{g}(x)]}\right]}{1+\exp\left[\hat{g}(x) + z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{g}(x)]}\right]} \geq \frac{\exp[g(x)]}{1+\exp[-g(x)]} \quad \geq \frac{\exp\left[\hat{g}(x) - z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{g}(x)]}\right]}{1+\exp\left[\hat{g}(x) - z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{g}(x)]}\right]}. \tag{2.26}$$

Usando-se das equações (2.3) e (2.4) na expressão (2.26), verifica-se o resultado.

□

2.2 Regressão Logística Múltipla

Nesta seção generaliza-se o modelo logístico apresentado na Definição 2.1 para o caso em que se apresentam mais de uma variável independente, ou seja, este é o caso múltiplo. Dessa forma, ainda se discute o contexto de uma v.a. Y dependente com natureza dicotômica, o que acarreta semelhança no modelo paramétrico adotado e nas hipóteses envolvidas na seção precedente.

Antes de se iniciar a discussão teórica apresenta-se um exemplo para ilustrar o novo contexto.

Exemplo 2.2. Um estudo sobre o peso de nascimento de bebês é realizado. A variável em análise Y pode assumir valor "baixo peso" se o bebê ao nascer apresenta peso inferior a 2500g ou então assumir o valor "peso adequado" se ao nascer o bebê apresenta peso igual ou superior a 2500g. Ou seja Y é uma v.a. de natureza nominal e pode ser quantizada atribuindo-se o valor zero para a situação de baixo peso e valor um para a situação de peso adequado. Algumas variáveis independentes que podem ajudar a modelar este problema são a idade da mãe, o peso da mãe, número de visitas pré-natal e condição sócio econômica.

Considera-se uma coleção de r variáveis independentes denotadas por $\mathbf{X} = (X_1, \cdots, X_r)'$, onde $\mathbf{x} = (x_1, \cdots, x_r)'$ é um valor particular e uma v.a. dependente Y de natureza binária. A proporção do número de indivíduos com a característica Y, denotada por $\mathbb{E}(Y|\mathbf{x})$, sendo uma proporção, é tal que $0 \leq \mathbb{E}(Y|\mathbf{x}) \leq 1$.

Seja $g(\cdot)$, a chamada função *logit*, uma representação da combinação linear das r variáveis preditoras

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r. \tag{2.27}$$

O principal objetivo nesta etapa é apresentar um modelo para predizer $\mathbb{P}(Y = 1|\mathbf{x}) = \pi(\mathbf{x})$, onde

$$\pi(\mathbf{x}) = \frac{\exp[g(\mathbf{x})]}{1 + \exp[g(\mathbf{x})]}. \tag{2.28}$$

Assumindo-se n independentes observações de Y, denotadas por $y_1, \cdots, y_n$, associadas aos valores de $\mathbf{x}_i = (x_{i1}, \cdots, x_{ir})$, para $i \in \{1, \cdots, n\}$, o *logit*, dado pela

expressão (2.27), apresenta-se como

$$\begin{aligned} g_1 &= g_1(\mathbf{x}_1) = \beta_0 + \beta_1 x_{11} + \cdots + \beta_r x_{1r} + \varepsilon_1 \\ g_2 &= g_2(\mathbf{x}_2) = \beta_0 + \beta_1 x_{21} + \cdots + \beta_r x_{2r} + \varepsilon_2 \\ &\vdots \\ g_n &= g_n(\mathbf{x}_n) = \beta_0 + \beta_1 x_{n1} + \cdots + \beta_r x_{nr} + \varepsilon_n, \end{aligned} \tag{2.29}$$

onde os erros, ε_i, seguem as seguintes suposições, para todo $i, l \in \{1, \cdots, n\}$

$$\begin{aligned} &(i) \quad \mathbb{E}(\varepsilon_i|\mathbf{x}_i) = 0. \\ &(ii) \quad Var(\varepsilon_i|\mathbf{x}_i) = \pi(\mathbf{x}_i)[1 - \pi(\mathbf{x}_i)]. \\ &(iii) \quad Cov(\varepsilon_i, \varepsilon_l) = 0, \text{ se } i \neq l. \end{aligned} \tag{2.30}$$

Os comentários acima motivam a seguinte definição.

Definição 2.2. As v.a.'s $Y_1, \cdots, Y_n$ satisfazem um *modelo logístico múltiplo* se uma amostra de tamanho um de cada Y_i pode ser expressa como

$$y_i = \frac{\exp(g_i)}{1 + \exp(g_i)} + \varepsilon_i,$$

onde g_i é obtida pela expressão (2.29), para a qual x_{ij} é constante conhecida, β_j é parâmetro desconhecido do modelo e os erros ε_i possuem as suposições dadas em (2.30).

Observação 2.7. No modelo apresentado pela Definição 2.2 pode-se ter várias variáveis discretas, do tipo escala nominal, cujos diversos números usados para representar os diversos níveis dessa escala são meramente identificadores e não possuem significado numérico. Estas são as variáveis *dummies*. A expressão apresentada em (2.27) pode ser reescrita como

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \sum_{l=1}^{k_j - 1} \beta_{jl} x_{jl} + \cdots + \beta_r x_r,$$

quando temos uma variável na escala nominal com k possíveis valores. Foram introduzidas $k - 1$ variáveis *dummies* no modelo onde a j-ésima variável está na escala nominal com k_j níveis; cada uma das $k_j - 1$ variáveis *dummies* é denotada por x_{jl} e seu coeficiente por β_{jl}, com $l \in \{1, \cdots, k_j - 1\}$.

Como nesta seção ainda tem-se o contexto de que Y é uma v.a. de natureza dicotômica, pode-se expressar uma observação como $y = \pi(\mathbf{x}) + \varepsilon$, e assim, a v.a.

ε pode assumir uma de duas possibilidades. Se $y = 1$ então $\varepsilon = 1 - \pi(\mathbf{x})$ com probabilidade $\pi(\mathbf{x})$, e se $y = 0$ então $\varepsilon = -\pi(\mathbf{x})$ com probabilidade $1 - \pi(\mathbf{x})$.

Assim, um análogo à Proposição 2.2 pode ser apresentado.

Proposição 2.13. *Assume-se o contexto da Definição 2.2. A v.a. ε tem distribuição de Bernoulli com média zero e variância $\pi(\mathbf{x})[1 - \pi(\mathbf{x})]$.*

Demonstração: A demonstração é semelhante a realizada na Proposição 2.2. □

De posse do modelo, precisa-se determinar seus coeficientes $\beta_0, \cdots, \beta_r$. Os coeficientes são parâmetros que devem ser estimados pelo método de máxima verossimilhança.

Antes de iniciar-se esta discussão, definem-se as seguintes matrizes,

$$\begin{aligned}
\mathbf{Y} &= (y_1, \cdots, y_n)'_{1\times n} \\
\mathbf{\Pi} &= (\pi_1, \cdots, \pi_n)'_{1\times n} \\
\mathcal{B} &= (\beta_0, \cdots, \beta_r)'_{1\times (r+1)} \\
\mathbf{X} &= \begin{bmatrix} 1 & x_{11} & \cdots & x_{1r} \\ 1 & x_{21} & \cdots & x_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nr} \end{bmatrix}_{n\times(r+1)} \\
\mathbf{\Sigma} &= \begin{bmatrix} \pi_1(1-\pi_1) & 0 & \cdots & 0 \\ 0 & \pi_2(1-\pi_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_n(1-\pi_n) \end{bmatrix}_{n\times n},
\end{aligned} \tag{2.31}$$

onde $\mathbf{Y}$, $\mathbf{\Pi}$ e $\mathcal{B}$ são matrizes $n \times 1$, $\mathbf{X}$ é uma matriz $n \times (r + 1)$ e $\mathbf{\Sigma}$ é uma matriz $n \times n$.

A função de verossimilhança, neste caso, é idêntica à expressão (2.12), com a única alteração de que agora $\pi(\cdot)$ é dada pela expressão (2.28), de forma que é definida como a função massa de probabilidade conjunta de n v.a.'s. Especificamente, para uma amostra de tamanho n, tem-se que

$$L(\mathcal{B}) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}, \text{ com } y_i \in \{0, 1\}. \tag{2.32}$$

Assim, pode-se apresentar uma proposição análoga à Proposição 2.3.

Proposição 2.14. *Assume-se o contexto da Definição 2.2. Seja $\mathcal{B}$ o vetor de parâmetros relacionado com a probabilidade condicional $\mathbb{P}(Y_i = 1|\mathbf{x}_i) = \pi(x_i)$ para $i \in \{1, \cdots, n\}$. Então, o estimador de $\mathcal{B}$, pelo método da máxima verrossimilhança, denotado por $\hat{\mathcal{B}}$, é a solução das equações de verossimilhança*

$$\begin{aligned} &\sum_{i=1}^{n}(y_i - \pi_i) = 0 \\ &\sum_{i=1}^{n} x_{ij}(y_i - \pi_i) = 0, \ \ \textit{para} \ \ j \in \{1, \cdots, r\}. \end{aligned} \tag{2.33}$$

Demonstração: A prova desta afirmação segue raciocínio análogo ao desenvolvido na Proposição 2.3. Entretando, agora, tem-se $r + 1$ equações de verossimilhança que são obtidas ao se diferenciar a função logaritmo de verossimilhança dada por

$$\mathcal{L}(\mathcal{B}) = \sum_{i=1}^{n}[y_i \ln \pi_i + (1 - y_i) \ln(1 - \pi_i)], \tag{2.34}$$

com respeito a cada um dos $r + 1$ coeficientes. A expressão (2.34) é obtida a partir do logaritmo da função (2.32) e do uso das propriedades de somatório e de logaritmos.

As expressões das equações normais são apresentadas abaixo,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \beta_0} &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\pi}_i = 0 \\ \frac{\partial \mathcal{L}}{\partial \beta_j} &= \sum_{i=1}^{n} x_{ij} y_i - \sum_{i=1}^{n} x_{ij} \hat{\pi}_i = 0, \ \ \text{para} \ \ j \in \{1, \cdots, r\}. \end{aligned} \tag{2.35}$$

onde $\hat{\pi}_i$ indica o estimador pelo método da máxima verossimilhança de π_i.

Com base nas propriedades de somatório, a partir de (2.35) segue-se o resultado. □

Aqui são válidos comentários semelhantes aos apresentados nas Observações 2.3 e 2.4, ou seja, como o sistema apresentado em (2.33) não é linear, as equações de verossimilhança não são facilmente resolvidas e, então, se faz necessário o uso de algum método de aproximação. O método utilizado é o método iterativo de Newton-Raphson; e, ainda, pode-se provar que a expressão (2.34) é uma função estritamente convexa, logo o sistema (2.33) admite um único ponto de máximo, ou seja, uma única solução.

De forma mais compacta pode-se representar todas as $r+1$ equações de verossimilhança, em notação matricial, através de

$$\frac{\partial \mathcal{L}(\mathcal{B})}{\partial \mathcal{B}} \mathbf{X}'(\mathbf{Y} - \mathbf{\Pi}) = \mathbf{0}.$$

Um método para avaliar a significância das variáveis foi apresentado na seção precedente através do uso da estatística G, definida em (2.21).

Para acessar a significância de todos os $r+1$ coeficientes no modelo apresentado na Definição 2.2, o teste de razão de verossimilhança é baseado na mesma estatística G, com a diferença de que os valores ajustados $\hat{\pi}_i$, sob o modelo, estão baseados em um vetor contendo $r+1$ parâmetros, e desta forma a estatística G, dada por (2.21) apresenta distribuição qui-quadrado com r graus de liberdade.

Teorema 2.15. *Assume-se o contexto da Definição 2.2. O teste de razão de verossimilhança de tamanho α é dado por*

$$\mathcal{H}_0 : \mathcal{B} = \mathbf{B} \ \ vs \ \ \mathcal{H}_1 : \mathcal{B} \neq \mathbf{B}, \ \ para \ \ \mathbf{B} \in \mathbb{R}^{r+1},$$

e rejeita-se $\mathcal{H}_0$ se $\mathbb{P}(\chi^2_r > G) < \alpha$.

Demonstração: A prova de que a estatística G possui distribuição qui-quadrado com r graus de liberdade é conseqüência da Proposição 2.27. □

Na seção precedente, foram discutidos dois testes equivalentes ao teste de razão de verossimilhança, o Teste de Wald e o Teste de Score. Faz-se, então, uma breve apresentação de suas versões na regressão logística múltipla:

(i) O análogo multivariado do Teste de Wald é obtido pela expressão

$$W = \hat{\mathcal{B}}'[\widehat{Var}(\hat{\mathcal{B}})]^{-1}\hat{\mathcal{B}} = \hat{\mathcal{B}}'(\mathbf{X}'\mathbf{\Sigma}\mathbf{X})\hat{\mathcal{B}}, \tag{2.36}$$

onde $\mathbf{\Sigma}$ é dada por (2.31). Sabe-se que sob $\mathcal{H}_0 : \mathcal{B} = \mathbf{0}$, a estatística W, dada por (2.36), possui distribuição qui-quadrado com $r+1$ graus de liberdade. Como este teste exige a execução de operações entre matrizes e a obtenção de $\hat{\mathcal{B}}$, então não há vantagens computacionais sobre o teste de razão de verossimilhança para se testar a significância do modelo.

(ii) O análogo multivariado do Teste de Score para a avaliação da significância do modelo é baseado na distribuição das r derivadas de (2.34) com respeito a $\mathcal{B}$. As dificuldades computacionais para este teste são as mesmas do Teste de Wald.

Para o leitor com interesse em obter mais informações sobre estes testes recomenda-se Cox e Hinkley (1974) e Dobson (1990).

Os intervalos de confiança para os coeficientes β_j, com $j \in \{0, \cdots, r\}$, são feitos de forma análoga ao desenvolvido no modelo de regressão logística binária, conforme afirma a proposição a seguir.

Proposição 2.16. *O intervalo a* $100(1-\alpha)\%$ *de confiança para* β_j, *com* $j \in \{0, \cdots, r\}$ *é dado por*

$$[\hat{\beta}_j - z_{\frac{\alpha}{2}} \widehat{SE}(\hat{\beta}_j), \hat{\beta}_j + z_{\frac{\alpha}{2}} \widehat{SE}(\hat{\beta}_j)],$$

onde $z_{\frac{\alpha}{2}}$ *é o quantil de uma normal padrão dado por* $\mathbb{P}(z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$ *e* $\widehat{SE}(\hat{\beta}_j)$ *denota o estimador do desvio padrão de* $\hat{\beta}_j$.

Para a obtenção do intervalo de confiança para o estimador do *logit* utiliza-se a mesma idéia apresentada na Proposição 2.11. Entretanto, aqui expressa-se o *logit* em sua notação vetorial $\hat{g}(\mathbf{x}) = \mathbf{x}'\hat{\mathcal{B}}$, onde o vetor $\hat{\mathcal{B}}' = (\hat{\beta}_0, \cdots, \hat{\beta}_r)$ denota o estimador dos $r+1$ coeficientes e o vetor $\mathbf{x}'$ representa a constante e o conjunto de valores das r variáveis independentes do modelo, onde $x_0 = 1$.

O estimador da variância de $\hat{g}(\mathbf{x})$, representado por $\widehat{Var}[\hat{g}(\mathbf{x})]$, requer a obtenção da variância da soma. Neste caso

$$\widehat{Var}[\hat{g}(\mathbf{x})] = \sum_{j=0}^{r} x_j^2 \widehat{Var}(\hat{\beta}_j) + 2 \sum_{j=0}^{r} \sum_{k=j+1}^{r} x_j x_k \widehat{Cov}(\hat{\beta}_j, \hat{\beta}_k). \tag{2.37}$$

O resultado da expressão (2.37) pode ser apresentado de forma mais compacta através de sua notação matricial.

Define-se como *matriz da informação* a expressão

$$\widehat{Var}(\mathcal{B}) = (\mathbf{X}'\mathbf{\Sigma}\mathbf{X})^{-1}.$$

Dessa forma,

$$\widehat{Var}[\hat{g}(\mathbf{x})] = \mathbf{x}'\widehat{Var}(\mathcal{B})\mathbf{x} = \mathbf{x}'(\mathbf{X}'\mathbf{\Sigma}\mathbf{X})^{-1}\mathbf{x},$$

onde $\mathbf{X}$ e $\mathbf{\Sigma}$, são dadas em (2.31).

No modelo de regressão logística a matriz de informação observada coincide com a matriz de informação esperada porque as segundas derivadas na função log de verossimilhança não dependem da variável aleatória dependente Y.

Observação 2.8. A matriz da informação foi utilizada na expressão (2.36) para a construção da estatística de Wald.

Proposição 2.17. *O intervalo a* $100(1-\alpha)\%$ *de confiança para* $g(\mathbf{x})$ *é dado por*

$$\left[\hat{g}(\mathbf{x}) - z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{g}(\mathbf{x})]}, \hat{g}(\mathbf{x}) + z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{g}(\mathbf{x})]}\right],$$

onde $\widehat{Var}[\hat{g}(\mathbf{x})]$ *é dada pela expressão (2.37) e* $z_{\frac{\alpha}{2}}$ *é o quantil de uma normal padrão dado por* $\mathbb{P}(z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$.

Demonstração: Decorre da Proposição 2.16. □

Proposição 2.18. *O intervalo a* $100(1-\alpha)\%$ *de confiança para* $\pi(\mathbf{x})$ *é dado por*

$$\left\{\frac{\exp\left[\hat{g}(\mathbf{x}) - z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{g}(\mathbf{x}))}\right]}{1+\exp\left[\hat{g}(\mathbf{x}) - z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{g}(\mathbf{x}))}\right]}, \frac{\exp\left[\hat{g}(\mathbf{x}) + z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{g}(\mathbf{x}))}\right]}{1+\exp\left[\hat{g}(\mathbf{x}) + z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{g}(\mathbf{x}))}\right]}\right\},$$

onde $\widehat{Var}[\hat{g}(\mathbf{x})]$ *é dada pela expressão (2.37) e* $z_{\frac{\alpha}{2}}$ *é o quantil de uma normal padrão dado por* $\mathbb{P}(z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$.

Demonstração: A prova desta afirmação segue raciocínio análogo ao realizado na Proposição 2.12. □

2.3 Regressão Logística Multinomial

O modelo apresentado na Definição 2.2 pode ser modificado de maneira que a v.a. dependente, que possui natureza nominal, apresente mais de dois níveis de codificação. Esta situação é ilustrada pelo exemplo abaixo.

Exemplo 2.3. Uma grande corporação realiza um estudo para escolher um plano de saúde para os seus funcionários a partir de três opções oferecidas pela empresa prestadora de serviços. Dessa forma, a variável em análise é o tipo do plano de saúde escolhido, que possui natureza nominal e seus três níveis são denotados por A, B e C. As variáveis independentes utilizadas para a escolha do plano de saúde são: a idade do funcionário, o tamanho de sua família e o rendimento mensal. O objetivo deste estudo é modelar as escolhas de plano de saúde como uma função das variáveis envolvidas e apresentar os resultados em termos das proporções de escolha dos diferentes planos.

Nas seções anteriores os modelos de regressão dados pelas Definições 2.1 e 2.2 utilizavam-se de uma v.a. binária, ou seja, que poderia assumir, por exemplo, apenas os valores zero e um. Assim o modelo era parametrizado em termos do *logit* de $Y = 1$ versus $Y = 0$.

Considera-se uma coleção de $r + 1$ variáveis independentes denotadas por $\mathbf{X} = (X_0, X_1, \cdots, X_r)$, onde $\mathbf{x} = (x_0, x_1, \cdots, x_r)$ com $x_0 = 1$ e uma v.a. Y de natureza nominal que pode assumir os níveis $0, 1, \cdots, q$.

Uma abordagem análoga à realizada nas seções anteriores é descrever o *logit* comparando-se $Y = k$ com $Y = 0$ para $k \in \{1, \cdots, q\}$. O valor *zero* então é denominado *categoria de referência*.

Denota-se as funções *logit* como sendo

$$\begin{aligned} g_k &\equiv g_k(\mathbf{x}) = \ln\left[\frac{\mathbb{P}(Y = k|\mathbf{x})}{\mathbb{P}(Y = 0|\mathbf{x})}\right] \\ &= \beta_{k0}x_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kr}x_r \\ &= \mathbf{x}'\mathcal{B}_k, \text{ para } k \in \{0, \cdots, q\}, \end{aligned} \tag{2.38}$$

onde $\mathcal{B}_k = (\beta_{k0}, \cdots, \beta_{kr})'$ e $x_{k0} = 1$.

Assumindo-se n independentes observações de Y, denotadas por $y_1, \cdots, y_n$, associadas aos valores de $\mathbf{x}_i = (x_{i0}, \cdots, x_{ir})$, para $i \in \{1, \cdots, n\}$, o *logit*, dado em (2.38), apresenta-se como

$$\begin{aligned} g_{k1} &\equiv g_{k1}(\mathbf{x}_1) = \beta_{k0}x_{10} + \beta_{k1}x_{11} + \cdots + \beta_{kr}x_{1r} + \varepsilon_1 \\ g_{k2} &\equiv g_{k2}(\mathbf{x}_2) = \beta_{k0}x_{20} + \beta_{k1}x_{21} + \cdots + \beta_{kr}x_{2r} + \varepsilon_2 \\ &\vdots \\ g_{kn} &\equiv g_{kn}(\mathbf{x}_n) = \beta_{k0}x_{n0} + \beta_{k1}x_{n1} + \cdots + \beta_{kr}x_{nr} + \varepsilon_n, \end{aligned}$$

onde $x_{i0} = 1$, para $i \in \{1, \cdots, n\}$ e os erros, ε_i, seguem as seguintes suposições, para todo $i, l \in \{1, \cdots, n\}$

$$\begin{aligned} &(i) \quad \mathbb{E}(\varepsilon_i|\mathbf{x}_i) = 0. \\ &(ii) \quad Var(\varepsilon_i|\mathbf{x}_i) = Var(Y_i|\mathbf{x}_i). \\ &(iii) \quad Cov(\varepsilon_i, \varepsilon_l) = 0, \text{ se } i \neq l. \end{aligned} \tag{2.39}$$

O exposto acima motiva a seguinte definição.

Definição 2.3. As v.a.'s $Y_1, \cdots, Y_n$ satisfazem um *modelo logístico multinomial* se uma amostra de tamanho um de cada Y_i pode ser expressa como

$$\pi_{ki} \equiv \pi_{ki}(\mathbf{x}) = \frac{\exp(g_{ki})}{1 + \exp(g_{ki})}, \tag{2.40}$$

onde g_{ki} é obtida pela expressão (2.38), para a qual x_{ij} é constante conhecida e β_{kj} é parâmetro desconhecido, os erros ε_i possuem as suposições dadas em (2.68) e $\pi_{ki}(\mathbf{x})$ representa $\mathbb{P}(Y_i = k|\mathbf{x})$, com $i \in \{1, \cdots, n\}$, $j \in \{0, \cdots, r\}$ e $k \in \{0, \cdots, q\}$.

Observação 2.9. Decorre da primeira igualdade apresentada na expressão (2.38) que $\exp[g_{0i}(\mathbf{x})] = 1$, e desta forma $\beta_{0j} = 0$, para qualquer $j \in \{0, \cdots, r\}$. E, ainda, que para cada nível que a v.a. Y pode assumir tem-se $r + 1$ coeficientes. Ou seja, o modelo apresenta um total de $q(r + 1)$ coeficientes.

Seguindo-se a convenção apresentada no modelo dado por (2.1), tem-se que

$$\pi_k(\mathbf{x}) = \mathbb{P}(Y = k|\mathbf{x}), \text{ para } k \in \{0, \cdots, q\}. \tag{2.41}$$

Proposição 2.19. *Uma expressão geral para as probabilidades condicionais em um modelo com $q + 1$ categorias é dada por*

$$\mathbb{P}(Y = k|\mathbf{x}) = \frac{\exp[g_k(\mathbf{x})]}{\sum_{k=0}^{q} \exp[g_k(\mathbf{x})]},$$

onde $g_k(\mathbf{x})$ é dado pela expressão (2.38), para $k \in \{1, \cdots, q\}$ e $g_0(\mathbf{x}) = 0$.

Demonstração: Da expressão (2.38) pode-se obter, pelo uso das propriedades de logaritmos,

$$\exp[g_k(\mathbf{x})] = \frac{\mathbb{P}(Y = k|\mathbf{x})}{\mathbb{P}(Y = 0|\mathbf{x})}, \text{ para } k \in \{0, \cdots, q\},$$

ou ainda

$$\mathbb{P}(Y = k|\mathbf{x}) = \exp[g_k(\mathbf{x})]\mathbb{P}(Y = 0|\mathbf{x}). \tag{2.42}$$

Pela propriedade da probabilidade total tem-se que a soma das $q + 1$ equações apresentadas em (2.41) totalizam valor um, ou seja,

$$\sum_{k=1}^{q} \mathbb{P}(Y = k|\mathbf{x}) = 1. \tag{2.43}$$

Substituindo a expressão (2.42) em (2.43), obtém-se

$$\sum_{k=1}^{q} \exp[g_k(\mathbf{x})]\mathbb{P}(Y = 0|\mathbf{x}) = 1,$$

o que, por propriedade de somatório, acarreta

$$\mathbb{P}(Y=0|\mathbf{x}) = \frac{1}{\sum_{k=1}^{q}\exp[g_k(\mathbf{x})]},$$

e, substituindo-se este último resultado em (2.42), segue a afirmação. □

Como nesta seção tem-se que T é uma v.a. de natureza politômica, com $q+1$ valores, pode-se expressar uma observação como $y = \pi(\mathbf{x}) + \varepsilon$, e assim, a v.a. ε pode assumir $q+1$ valores. Se $y_i = k$ então $\varepsilon_i = k - \pi(\mathbf{x}_i)$ com probabilidade $\mathbb{P}(Y=k|\mathbf{x}_i)$, para qualquer $k \in \{0,\cdots,q\}$ e $i \in \{1,\cdots,n\}$.

Proposição 2.20. *A v.a. ε tem distribuição Multinomial com média zero e variância igual a da v.a. Y.*

Demonstração: Primeiramente avalia-se a esperança de ε dado x_i,

$$\begin{aligned}
\mathbb{E}(\varepsilon|\mathbf{x}_i) &= \sum_{k=0}^{q}\varepsilon_k\mathbb{P}(\varepsilon=\varepsilon_k|\mathbf{x}_i)\\
&= -\pi(\mathbf{x}_i)\mathbb{P}(Y=0|\mathbf{x}_i) + \cdots + (q-\pi(\mathbf{x}_i))\mathbb{P}(Y=q|\mathbf{x}_i)\\
&= -\pi(\mathbf{x}_i)\sum_{k=0}^{q}\mathbb{P}(Y=k|\mathbf{x}_i) + \sum_{k=0}^{q}k\,\mathbb{P}(Y=k|\mathbf{x}_i)\\
&= -\pi(\mathbf{x}_i) + \mathbb{E}(Y|\mathbf{x}_i) = -\pi(\mathbf{x}_i) + \pi(\mathbf{x}_i) = 0.
\end{aligned}$$

Agora, avalia-se a variância condicional de ε,

$$\begin{aligned}
Var(\varepsilon|\mathbf{x}_i) &= \sum_{k=0}^{q}\varepsilon_k^2\mathbb{P}(\varepsilon=\varepsilon_k|\mathbf{x}_i) = \sum_{k=0}^{q}(k-\pi(\mathbf{x}_i))^2\mathbb{P}(Y=k|\mathbf{x}_i)\\
&= \sum_{k=0}^{q}(k^2 - 2k\pi(\mathbf{x}_i) + \pi(\mathbf{x}_i)^2)\mathbb{P}(Y=k|\mathbf{x}_i)\\
&= \sum_{k=0}^{q}k^2\mathbb{P}(Y=k|\mathbf{x}_i) - 2\pi(\mathbf{x}_i)\sum_{k=0}^{q}k\mathbb{P}(Y=k|\mathbf{x}_i) +\\
&\quad \pi(\mathbf{x}_i)^2\sum_{k=0}^{q}\mathbb{P}(Y=k|\mathbf{x}_i)\\
&= \mathbb{E}(Y^2|\mathbf{x}_i) - 2\pi(\mathbf{x}_i)\mathbb{E}(Y|\mathbf{x}_i) + \pi(\mathbf{x}_i)^2\\
&= \mathbb{E}(Y^2|\mathbf{x}_i) - \pi(\mathbf{x}_i)^2 = Var(Y|\mathbf{x}_i).
\end{aligned}$$

□

Observação 2.10. Para se construir a função de verossimilhança é necessário introduzir $q+1$ variáveis auxiliares com o objetivo de simplicar a notação utilizada,

mas que não são empregadas em nenhuma análise posterior. As variáveis auxiliares são apresentadas da seguinte forma, se $Y = 0$ então $Y_0 = 1, Y_1 = 0, \cdots, Y_q = 0$, se $Y = 1$ então $Y_0 = 0, Y_1 = 1, Y_2 = 0, \cdots, Y_q = 0$, e assim, de forma geral, se $Y = k$ então $Y_k = 1$ e $Y_l = 0$, para $l \neq k$ e $l \in \{0, 1, \cdots, q\}$.

Com base na Observação **??**, pode-se definir uma matriz auxiliar Q como segue

$$Q = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots \\ 0 & 0 & & 1 \end{bmatrix}_{(q+1)\times(q+1)}, \tag{2.44}$$

onde os elementos q_{rs} da linha r correspondem, respectivamente, aos valores assumidos pelas variáveis auxiliares Y_k, com $r \in \{1, \cdots, q\}$ e $k = r - 1$, quando $Y = k$.

É facil perceber que não importa qual seja o valor assumido por Y, o somatório $\sum_{k=1}^{q} Y_k$ sempre totaliza um.

Proposição 2.21. *A função de verossimilhança $L(\mathcal{B})$ para uma amostra de n observações independentes é dada por*

$$L(\mathcal{B}) = \prod_{i=1}^{n} [\pi_0(\mathbf{x}_i)^{Y_{0i}} \pi_1(\mathbf{x}_i)^{Y_{1i}} \cdots \pi_q(\mathbf{x}_i)^{Y_{qi}}], \tag{2.45}$$

onde $\pi_i = \pi(\mathbf{x}_i)$, $\mathbf{x}_i = (x_{i0}, \cdots, x_{ir})$ e $i \in \{1, \cdots, n\}$.

Demonstração: Seguindo-se raciocínio desenvolvido na Proposição 2.3 para se obter a função de verossimilhança, se $y_i = k$, a contribuição para a função de verossimilhança é $\pi_k(\mathbf{x}_i)$ e usando-se a linha $i = r - 1$ da matriz dada por (2.44), então tem-se

$$\pi_0(\mathbf{x}_i)^{Y_{0i}} \pi_1(\mathbf{x}_i)^{Y_{1i}} \cdots \pi_q(\mathbf{x}_i)^{Y_{qi}}.$$

Como sabemos as n observações são independentes, segue o resultado. □

É comum se utilizar a função log de verossimilhança, obtida após aplicação de logaritmo natural em ambos os lados da expressão (2.45), assumindo a forma

$$\mathcal{L}(\mathcal{B}) = \ln \left\{ \prod_{i=1}^{n} \left[\pi_0(\mathbf{x}_i)^{Y_{0i}} \pi_1(\mathbf{x}_i)^{Y_{1i}} \cdots \pi_q(\mathbf{x}_i)^{Y_{qi}} \right] \right\}. \tag{2.46}$$

Com base no resultado apresentado pela Proposição 2.21, pode-se apresentar uma forma de se obter o estimador de $\mathcal{B}$ pelo método de máxima verossimilhança.

Teorema 2.22. *Assume-se o contexto da Definição 2.3. Seja $\mathcal{B}$ o vetor de parâmetros relacionados com a probabilidade $\mathbb{P}(Y_i = k|\mathbf{x}_i)$, para $i \in \{1, \cdots, n\}$ e $k \in \{0, \cdots, q\}$. Então o estimador de $\mathcal{B}$, pelo método de máxima verossimilhança, denotado por $\hat{\mathcal{B}}$, é a solução das equações*

$$\frac{\partial \mathcal{L}(\mathcal{B})}{\partial \beta_{kj}} = \sum_{i=1}^{n} x_{ij}(y_{ki} - \pi_{ki}), \tag{2.47}$$

para $k \in \{1, \cdots, q\}$, $j \in \{0, \cdots, r\}$ e $\pi_{ki} = \pi_k(\mathbf{x}_i)$, com $x_{0i} = 1$ para qualquer i.

Demonstração: Ao se aplicar as propriedades de somatório e de logaritmo na função apresentada em (2.46), pode-se conseguir

$$\mathcal{L}(\mathcal{B}) = \sum_{i=1}^{n} \left\{ \sum_{k=1}^{q} y_{ki} g_k(\mathbf{x}_i) - \ln \left[\sum_{k=1}^{q} \exp[g_k(\mathbf{x}_i)] \right] \right\}. \tag{2.48}$$

As equações de verossimilhança (2.47) são obtidas através das primeiras derivadas parciais de (2.48) com respeito a cada um dos $q(r+1)$ parâmetros desconhecidos.

Para simplificar a notação, define-se $\pi_{ki} = \pi_k(\mathbf{x}_i)$. Assim sendo, a forma geral das equações apresentadas em (2.48) é

$$\frac{\partial \mathcal{L}(\mathcal{B})}{\partial \beta_{kj}} = \sum_{i=1}^{n} x_{ij}(y_{ki} - \pi_{ki}),$$

para $k \in \{1, \cdots, q\}$, $j \in \{1, \cdots, r\}$ e $x_{0i} = 1$, para qualquer $i \in \{1, \cdots, n\}$. □

Observação 2.11. O estimador de máxima verossimilhança, $\hat{\mathcal{B}}$, é obtido igualando-se cada equação a zero e resolvendo o sistema para $\mathcal{B}$. A solução requer alguma técnica de cálculo iterativo da mesma forma que se fez necessário para o cálculo do estimador nos modelos com variável dependente binária.

Um método para avaliar a significância das variáveis foi apresentado na Seção 2.1 através do uso da estatística G, definida em (2.21).

Para acessar a significância dos $q(r+1)$ coeficientes no modelo apresentado pela Definição 2.3, o teste de razão de verossimilhança é baseado nesta mesma estatística, G, com a diferença de que a mesma apresenta, neste contexto, distribuição qui-quadrado com $q(r+1) - r$ graus de liberdade.

Teorema 2.23. *Assume-se o contexto da Definição 2.3. O teste de razão de verossimilhança de tamanho α é dado por*

$$\mathcal{H}_0 : \mathcal{B} = \mathbf{B} \;\; vs \;\; \mathcal{H}_1 : \mathcal{B} \neq \mathbf{B}, \;\; para \;\; \mathbf{B} \in \mathbb{M}_{(q+1)\times(r+1)}, \tag{2.49}$$

e rejeita-se $\mathcal{H}_0$ *se* $\mathbb{P}(\chi^2_{q(r+1)-r} > G) < \alpha$*, onde* $\mathbb{M}_{(q+1)\times(r+1)}$ *representa o conjunto de todas as matrizes de dimensão* $(q+1)\times(r+1)$*.*

Demonstração: A prova de que a estatística G segue distribuição qui-quadrado com com $q(r+1)-r$ graus de liberdade é conseqüência da demonstração da Proposição 2.27 e da Observação 2.15, que se encontram na Seção 2.4 deste capítulo.

A matriz das segundas derivadas parciais é necessária para se obter a matriz informação, $\mathbf{I}(\hat{\mathcal{B}})$, e o estimador da matriz de variâncias-covariâncias para $\hat{\mathcal{B}}$. A expressão geral dos elementos na matriz das segundas derivadas parciais é

$$\frac{\partial \mathcal{L}(\mathcal{B})}{\partial \beta_{kj}\partial \beta_{kl}} = -\sum_{i=1}^{n} x_{il}x_{ij}\pi_{ki}(1-\pi_{ki}) \quad \text{e}$$
$$\frac{\partial \mathcal{L}(\mathcal{B})}{\partial \beta_{kj}\partial \beta_{kl}} = \sum_{i=1}^{n} x_{il}x_{ij}\pi_{ki}\pi_{mi}, \tag{2.50}$$

para $k, m \in \{1, \cdots, q\}$ e $j, l \in \{0, 1, \cdots, r\}$.

A matrix $\mathbf{I}(\hat{\mathcal{B}})$, de ordem $2(r+1)$, possui elementos que são simétricos aos valores encontrados nas expressões (2.50) quando avaliados em $\hat{\mathcal{B}}$.

O estimador da matriz de variâncias-covariâncias de $\hat{\mathcal{B}}$ é a inversa da matriz da informação, ou seja,

$$\widehat{Var}(\hat{\mathcal{B}}) = \mathbf{I}(\hat{\mathcal{B}})^{-1}. \tag{2.51}$$

Com base no resultado (2.51), pode-se apresentar o análogo multinomial ao Teste de Wald. A estatística deste teste é dada pela expressão

$$W = \hat{\mathcal{B}}'[\mathbf{I}(\hat{\mathcal{B}})]\hat{\mathcal{B}}'.$$

Sabe-se que sob $\mathcal{H}_0 : \mathcal{B} = \mathbf{0}$, a estatística W possui distribuição qui-quadrado com $q(r+1)$ graus de liberdade. E, da mesma forma que seus análogos, este teste não apresenta vantagens computacionais sobre o teste da razão de verossimilhança.

Apresentam-se os intervalos de confiança para os coeficientes β_{kj}, com $k \in \{0, \cdots, q\}$ e $j \in \{1, \cdots, r\}$, de forma análoga ao caso binário, conforme afirma a proposição abaixo.

Proposição 2.24. *O intervalo a* $100(1-\alpha)\%$ *de confiança para* β_{kj} *é dado por*

$$\left[\hat{\beta}_{kj} - z_{\frac{\alpha}{2}}\widehat{SE}(\hat{\beta}_{kj}), \hat{\beta}_{kj} + z_{\frac{\alpha}{2}}\widehat{SE}(\hat{\beta}_{kj})\right],$$

onde $z_{\frac{\alpha}{2}}$ *é o quantil de uma normal padrão dado por* $\mathbb{P}(z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$ *e* $\widehat{SE}(\hat{\beta}_{kj})$ *representa o estimador do desvio padrão de* $\hat{\beta}_{kj}$.

Observação 2.12. O estimador de $\widehat{SE}(\hat{\beta}_{kj})$ é a raiz quadrada do elemento da k-ésima linha e j-ésima coluna da matriz $\mathbf{I}(\hat{\mathcal{B}})^{-1}$.

Para o obtenção dos intervalos de confiança para os estimadores das funções *logits*, $g_k(\mathbf{x})$, com $k \in \{0, \cdots, q\}$, utiliza-se a mesma idéia apresentada na Proposição 2.17. Entretanto, aqui expressa-se o *logit* em sua notação dada por (2.38).

O estimador da variância de $\hat{g}_k(\mathbf{x})$, representado por $\widehat{Var}[\hat{g}_k(\mathbf{x})]$, requer a obtenção da variância da soma, como feito em (2.37), resultando em

$$\widehat{Var}[\hat{g}_k(\mathbf{x})] = \mathbf{x}' \widehat{Var}(\hat{\mathcal{B}}_k)\mathbf{x}. \tag{2.52}$$

O estimador $\widehat{Var}[\hat{g}_k(\mathbf{x})]$ pode ser obtido atraés do método Delta. Este método fornece valores dos desvios-padrão de estatísticas que podem ser representadas como função de outras estatísticas que possuem distribuição assintótica conjunta normal. A formalização do método Delta pode ser obtida em Agresti (1984) e Agresti (1990).

Proposição 2.25. *O intervalo a* $100(1-\alpha)\%$ *de confiança para* $g_k(\mathbf{x})$, *com* $k \in \{0, \cdots, q\}$ *é dado por*

$$\left[\hat{g}_k(\mathbf{x}) - z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{g}_k(\mathbf{x})]}, \hat{g}_k(\mathbf{x}) + z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{g}_k(\mathbf{x})]}\right],$$

onde $\widehat{Var}[\hat{g}_k(\mathbf{x})]$ *é dada pela expressão (2.52) e* $z_{\frac{\alpha}{2}}$ *é o quantil de uma normal padrão dado por* $\mathbb{P}(z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$.

Demonstração: Decorre da Proposição 2.24, da expressão (2.23). □

Proposição 2.26. *O intervalo a* $100(1-\alpha)\%$ *de confiança para* $\pi_k(\mathbf{x})$ *é dado por*

$$\left\{\frac{\exp\left[\hat{g}_k(\mathbf{x}) - z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{g}_k(\mathbf{x})]}\right]}{1+\exp\left[\hat{g}_k(\mathbf{x}) - z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{g}_k(\mathbf{x})]}\right]}, \frac{\exp\left[\hat{g}_k(\mathbf{x}) + z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{g}_k(\mathbf{x})]}\right]}{1+\exp\left[\hat{g}_k(\mathbf{x}) + z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{g}_k(\mathbf{x})]}\right]}\right\},$$

onde $\widehat{Var}[\hat{g}_k(\mathbf{x})]$ *é dada pela expressão (2.52) e* $z_{\frac{\alpha}{2}}$ *é o quantil de uma normal padrão dado por* $\mathbb{P}(z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$.

Demonstração: A prova desta afirmação segue raciocínio análogo ao realizado na Proposição 2.18. Sua validade é justificada através do Método Delta. Detalhes sobre este método podem ser obtidos em Agresti (1984) e Agresti (1990). □

2.4 Testes de Significância, Medidas de Dimensão e Interpretação do Logit.

Nesta seção apresentam-se ferramentas matemáticas e estatísticas capazes de fornecer uma análise dos resultados obtidos no modelo de regressão logística com base nos dados amostrais utilizados.

Na regressão linear a avaliação do modelo é baseada em soma de quadrados. A partir da soma de quadrados definem-se as estatísticas F e R^2 para a avaliação do modelo.

Estatísticas semelhantes a R^2 e F existem no modelo de regressão logística. Da mesma maneira que a soma dos quadrados é usada como critério para a seleção dos parâmetros no modelo de regressão linear, a função logaritmo de verossimilhança, dada por (2.13), é o critério para se selecionar parâmetros na regressão logística. A diferença definida em (2.20) representa este critério, e é denominada de modelo qui-quadrado, representando o papel da estatística F do modelo linear.

Observação 2.13. Na regressão logística, a diferença de duas funções de logaritmo de verossimilhança, quando multiplicada por -2, pode ser interpretada como uma estatística χ^2 se as funções são originárias de dois diferentes modelos ou quando as variáveis independentes do primeiro modelo são um subconjunto próprio das variáveis independentes do segundo modelo. Para uma discussão completa sobre esta observação, recomenda-se McCullagh e Nelder (1989).

Muitas estatísticas análogas à estatística R^2 da regressão linear são apresentadas no contexto da regressão logística, e são conhecidas como medidas de dimensão de efeito (*measures of effect size*). Para detalhes veja Veall e Zimmerman (1996) e Menard (2000).

Se for mantida a analogia entre a função logaritmo de verossimilhança na regressão logística e a soma de quadrados para a regressão linear, então, uma escolha análoga à estatística R^2, dada por (1.72), é definida por

$$R^2_{\mathcal{L}} = \frac{G}{D_0} = \frac{G}{G + D_M},$$

onde D_0 e D_M são dadas por

$$
\begin{aligned}
D_0 &= -2\left\{n_1 \ln[\mathbb{P}(Y=1)] + n_0 \ln[\mathbb{P}(Y=0)]\right\}, \\
D_M &= -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}(x_i)}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}(x_i)}{1-y_i}\right)\right],
\end{aligned}
$$

para $n_1 = \sum_{i=1}^{n} y_i$ e $n_0 = \sum_{i=1}^{n}(1-y_i)$ e $\mathbb{P}(Y=1) = \frac{n_1}{n}$.

A relação entre as variáveis G e D_M é apresentada por $G = D_0 - D_M$.

A estatística $R^2_{\mathcal{L}}$ é uma medida da redução proporcional, no valor absoluto, da função logaritmo de verossimilhança e indica o quanto a inclusão das variáveis independentes no modelo reduz a variação. Esta estatística pode assumir valores entre 0 e 1. O valor $R^2_{\mathcal{L}} = 0$ representa um modelo no qual $G = 0$ e $D_M = D_0$ e as variáveis independentes não são importantes na predição da variável dependente; e, $R^2_{\mathcal{L}} = 1$ representa um modelo no qual $G = D_0$ e $D_M = 0$ e as variáveis independentes descrevem perfeitamente o modelo em análise.

Uma alternativa ao cálculo de $R^2_{\mathcal{L}}$ pode ser a obtenção do coeficiente de contingência de Alderich e Nelson ou também chamado de Pseudo-R^2, denotado por R^2_C e definido por

$$R^2_C = \frac{G}{G+n},$$

onde n representa o número de dados observados. Esta estatística pode ser utilizada tanto em casos dicotômicos como em politômicos e mais informações podem ser encontradas em Hagle e Mitchell (1992).

Nos resultados de uma análise através do *software* SPSS 10.0 figuram duas estatísticas R^2. Uma delas é a estatística R^2 de Cox e Snell que é baseada na função de verossimilhança, mas seu valor máximo pode ser, e geralmente é, inferior a um, ocasionando assim dificuldades na análise. A outra estatística é R^2 de Nagelkerke, que nada mais é do que uma variação da estatística proposta por Cox e Snell buscando assegurar sua variação entre zero e um. Para saber mais sobre tais estatísticas recomendam-se, respectivamente, Cox e Snell (1989) e Nagelkerke (1991).

Antes de se apresentar a estatística qui-quadrado de Pearson, faz-se a introdução de duas variáveis auxiliares.

Se no modelo de regressão logística binária tem-se uma variável independente X, e denota-se por J o número de diferentes valores observados para x. Se algumas

observações x assumem o mesmo valor, tem-se então que $n > J$. Denota-se o número de observações em que $x = x_i$ por m_i, para $i \in \{1, \cdots, J\}$. Segue-se que $\sum_{i=1}^{J} m_i = n$.

Seja y_{m_i} o número de vezes em que a v.a. Y assumiu valor um dentre todas as m_i observações em que $x = x_i$. Segue-se que $\sum_{i=1}^{J} y_{m_i} = n_1$, onde n_1 é definido em (2.22).

Pode-se mostrar que o número esperado de respostas positivas, ou seja, em que $y = 1$, é

$$\hat{y}_{m_i} = m_i\hat{\pi}_i = m_i\frac{\exp[\hat{g}(x_i)]}{1+\exp[\hat{g}(x_i)]},$$

onde $\hat{g}(x_i)$ representa o estimador do *logit* de x_i.

O resíduo de Pearson é definido como

$$r_i = \frac{y_{m_i} - m_i\hat{\pi}_i}{\sqrt{m_i\hat{\pi}_i(1-\hat{\pi}_i)}}, \tag{2.54}$$

e, a estatística baseada no resíduo (2.54) é chamada de estatística qui-quadrado de Pearson e é representada por

$$\chi^2 = \sum_{i=1}^{J} r_i^2.$$

O resíduo *deviance* é definido como

$$d_i = \pm\left\{2\left[y_{m_i}\ln\left(\frac{y_{m_i}}{m_i\hat{\pi}_i}\right) + (m_i - y_{m_i})\ln\left(\frac{m_i - y_{m_i}}{m_i(1-\hat{\pi}_i)}\right)\right]\right\}^{\frac{1}{2}}, \tag{2.55}$$

onde, o sinal da expressão (2.55) é o mesmo sinal de $(y_{m_i} - m_i\hat{\pi}_i)$.

É interessante comentar que na expressão (2.55), y_{m_i} pode ser interpretado como a quantidade de valores *um* que são observados e $m_i y_i$ como o número esperado de valores *um*. O valor $(m_i - m_i\hat{\pi}_i)$ corresponde à diferença entre as quantidades observadas e esperadas, assim como na estatística clássica χ^2 de Person em tabelas de contingência para testes de homogeneidade e de independência.

A estatística baseada no resíduo (2.55) é chamada de *deviance* e é dada por

$$D = \sum_{i=1}^{J} d_i^2. \tag{2.56}$$

Em um conjunto de dados em que $J = n$ a expressão dada por (2.56) é equivalente à (2.18).

A distribuição das estatísticas χ^2 e G sob a hipótese de que o modelo ajustado está correto é aproximadamente qui-quadrado com $J-2$ graus de liberdade. A proposição e o teorema apresentados a seguir confirmam esta afirmação.

Proposição 2.27. *A estatística χ^2, para amostras grandes, é uma aproximação da estatística G.*

Demonstração: Suponha que o número J está fixado e que $\pi_i > 0$, para $i \in \{1, \cdots, J\}$. Para verificar a afirmação, expressa-se G como

$$\begin{aligned} G &= 2\sum_{i=1}^{J} n_i \ln\left(\frac{n_i}{m_i}\right) \\ &= 2n\sum_{i=1}^{J} p_i \ln\left(1+\frac{p_i-\hat{\pi}_i}{\hat{\pi}_i}\right), \end{aligned} \tag{2.57}$$

onde $p_i = \dfrac{\sum_{i=1}^{J} y_{m_i}}{n} = \dfrac{n_i}{n}$.

Procura-se expressar G segundo a forma dada por (2.57) para se facilitar a aplicação da expansão em série de Taylor da função $y = \ln(1+x)$, que é dada por

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots, \tag{2.58}$$

onde $|x| < 1$.

Se aplicarmos a expressão (2.58), para $x = \dfrac{p_i-\hat{\pi}_i}{\hat{\pi}_i}$, na equação dada por (2.57) então tem-se, para n suficientemente grande, que

$$\begin{aligned} G &= 2n\sum_{i=1}^{J}[\hat{\pi}_i + (p_i-\hat{\pi}_i)]\left[\frac{p_i-\hat{\pi}_i}{\hat{\pi}_i} - \left(\frac{1}{2}\right)\frac{(p_i-\hat{\pi}_i)^2}{\hat{\pi}_i^2} + \cdots\right] \\ &= 2n\sum_{i=1}^{J}\left[(p_i-\hat{pi}_i) - \left(\frac{1}{2}\right)\frac{(p_i-\hat{\pi}_i)^2}{\hat{\pi}_i^2} + \frac{(p_i-\hat{\pi}_i)^2}{\hat{\pi}_i^2} + \mathcal{O}(p_i-\hat{\pi}_i)^3\right] \\ &= n\sum_{i=1}^{J}\frac{(p_i-\hat{\pi}_i)^2}{\hat{\pi}_i^2} + 2n\,\mathcal{O}(p_i-\hat{\pi}_i)^3 \\ &= \chi^2 + 2n\,\mathcal{O}(p_i-\hat{\pi}_i)^3, \end{aligned}$$

onde pode-se provar que $\dfrac{p_i-\hat{\pi}_i}{\hat{\pi}_i}$ converge, em probabilidade para zero.

Desta forma, tem-se que $2n\,\mathcal{O}(p_i-\hat{\pi}_i)^3$ converge para zero em probabilidade e, como conseqüência, G possui distribuição assintótica qui-quadrado com $J-2$ graus de liberdade. □

Observação 2.14. A notação $\mathcal{O}(z_n)$ representa uma v.a. que para todo $\epsilon > 0$, existe uma constante K e um inteiro n_0, de forma que $\mathbb{P}\left[\frac{|\mathcal{O}(z_n)|}{z_n} < K\right] > 1-\epsilon$ para todo $n > n_0$.

Teorema 2.28. *(***Aproximação geral da razão de verossimilhança***). Sob a hipótese de que o modelo ajustado é correto e n é suficientemente grande, então a estatística G tem aproximadamente distribuição qui-quadrado com $J-2$ graus de liberdade.*

Demonstração: Este teorema passa a ser um corolário do resultado apresentado na Proposição 2.27. □

Observação 2.15. O resultado apresentado pela Proposição 2.27 pode ser generalizado para o caso multinomial, sendo que sua demonstração fica praticamente a mesma, apenas havendo distinção nos valores p_i e n_i, que devem estar adequados ao contexto.

Como G tem aproximadamente uma distribuição qui-quadrado, é usada para avaliar a significância na regressão logística de forma análoga ao uso da soma dos quadrados dos erros na regressão linear. O teste de razão de verossimilhança, também conhecido como teste modelo qui-quadrado, foi apresentado nos contextos binário e múltiplo, onde a variável de resposta é dicotômica, e multinomial, em que a variável de resposta é politômica, respectivamente pelos Teoremas 2.8, 2.15 e 2.23.

Neste mesmo contexto, há o teste qui-quadrado, ou mais conhecido como, *Hosmer and Lemeshow's goodness of fit test* (não confundir com o teste C-chápeu que possui a denominação de *Hosmer e Lemeshow's goodness of fit index* que está obsoleto). Sobre tais informações recomenda-se o endereço eletrônico http://www2.-chass.ncsu.edu. Se a estatística do teste for maior que o nível de significância α adotado, rejeita-se a hipótese de que não há diferença entre os valores observados e preditos implicando, assim, que o modelo descreve bem os dados no nível adotado.

A estatística W, definida no Teste de Wald, é uma alternativa comumente utilizada para testar a significância individualmente dos coeficientes de cada variável independente e é análogo ao teste de significância dos coeficientes na regressão linear.

A função *logit* é analisada através dos estimadores de seus coeficientes e da

razão de chances que foi definida em (2.8).

Tabelas de classificação são tabelas de ordem dois, para o caso da regressão logística dicotômica, e de ordem $2 \times (q+1)$ para o caso da regressão logística politômica. Estas tabelas registram os estimadores corretos e incorretos. As colunas apresentam os valores preditos da variável dependente e as linhas fornecem os valores observados para a variável independente. Em um modelo perfeito, todos os casos estariam na diagonal principal e a porcentagem de acertos seria de 100%. Na seqüência apresentam-se dois exemplos que ilustram tais tabelas.

A Tabela 2.1 apresenta esta classificação para um determinado modelo em análise. Esta tabela informa que 100% das previsões em que a v.a. Y assume valor zero estão corretas e que nenhuma das previsões em que a v.a. Y assume valor um esteve correta, e, ainda, que no geral, houve um acerto de 78,06%, valor este que pode ser encarado como moderadamente bom. É obtido pela porcentagem de acerto, ou seja, divide-se o número 370 pelo total de dados observados, 474.

Tabela 2.1: Exemplo de Tabela de Classificação de ordem dois.

Obs\Pred	0	1	Correta (%)	Geral (%)
0	370	0	100	
1	104	0	0	78,06

Tabela 2.2, que é uma tabela de classificação de ordem três, é apresentada nas eqüência.

Tabela 2.2: Exemplo de Tabela de Classificação de ordem três.

Obs\Pred	1	2	3	Correta (%)	Geral (%)
0	0	0	5	100	
1	4	1	0	80	
2	1	4	0	80	86,7

Na Tabela 2.2 tem-se que, nas previsões de valor 1 para a v.a. Y, houve apenas 1 erro e 4 acertos, o mesmo ocorrendo para as previsões quando Y assume valor 2. Totalizando em ambos os casos 80% de acerto. As previsões para a v.a. Y assumir valor 0 tiveram 100% de acerto, ocasionando, assim uma porcentagem de

acerto geral 86,7%. O valor de 86,7% é obtido somando-se o total de acertos, 13, e dividindo-se pelo total de observações, 15.

Observação 2.16. Se o modelo de regressão logística admitisse homocedasticidade (observe que homocedasticidade não é uma suposição do modelo logístico), a porcentagem de acertos seria aproximadamente a mesma em todas as linhas das tabelas de classificação.

O histograma das probabilidades preditas, chamado de *classplot* é uma alternativa para o acesso às corretas e incorretas predições na regressão logística. O eixo das abcissas representa a probabilidade predita, com valores entre zero e um, para a variável dependente. O eixo das ordenadas representa a freqüência, ou seja, o número de casos classificados. No corpo do gráfico, as colunas são constituídas de símbolos 0 e 1 (ou equivalentemente, por exemplo, como N e Y).

Na Figura 2.1 apresenta-se um exemplo de *classplot*. Neste histograma, cada símbolo N e Y representa os casos preditos e o eixo das ordenadas indica a freqüência, ou seja, o número de casos classificados. No corpo de gráfico, as colunas representam os valores um e zero assumidos pela v.a. Y, denotados, respectivamente, por Y e N. Examinando-se este gráfico, pode-se dizer que o modelo tem dificuldade de classificar casos em que a probabilidade está próxima de 0,5 e, ainda, que muitos casos que deveriam ser classificados com valor um tiveram valor de classificação zero. E ainda, que prócima da probabilidade de 0,25 há uma coluna com cinco N's e apenas um Y, isto indica que seis casos foram preditos com valor Y com probabilidade próxima de 0,25 e foram classificados como N.

Na regressão linear, a análise de resíduos possui um papel fundamental para a avaliação de um determinado modelo. Os resíduos são estatísticas que também estão definidos para o modelo de regressão logística. Para maiores informações sobre este assunto recomenda-se o material disponibilizado no endereço eletrônico http://www.ime.usp.br/ giapaula/livro.pdf, de autoria do professor Gilberto Alvarenga Paula.

Alguns autores, Leinbaum e Klein (2002) e Smith (2001), por exemplo, não recomendam nem utilizam resíduos em suas análises. A verificação da validade de um modelo de regressão logística acaba sendo, muitas vezes, baseada na comparação com outros modelos matemáticos.

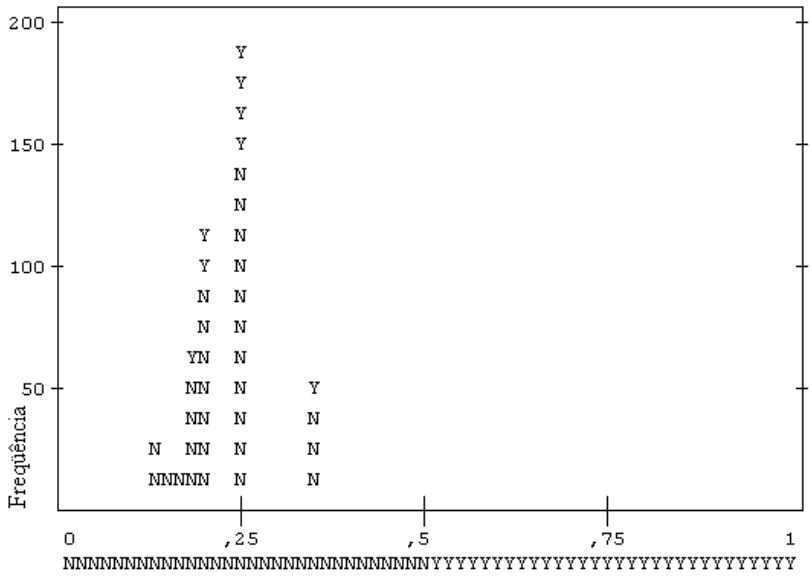

Figura 2.1: Exemplo de histograma de probabilidades preditas.

2.5 Aplicação da Teoria

Nesta seção apresentam-se exemplos de aplicação do modelo de regressão logística. O primeiro exemplo aborda a regressão logística binária e trata da análise da sobrevivência de insetos após terem recebido dosagens diversificadas de uma determinada substância. O segundo exemplo apresenta um caso de aplicação do modelo de regressão logística multinomial e mostra a relação entre pacientes que receberam medicamento e placebo para tratamento do sintoma de uma doença e tiveram diferentes níveis de reação.

Os dados utilizados para o primeiro e segundo exemplo, foram gentilmente cedidos, pela professora Hildete Prisco Pinheiro.

Os procedimentos utilizados para a análise dos exemplos abaixo seguem indicações apresentadas no Apêndice A de Kleinbaum e Klein (2002) e Smith (2001).

Exemplo 2.4. Um entomologista deseja fazer um estudo sobre a resistência de uma espécie de besouros com relação a uma determinada substância nociva. Para tal administra diferentes dosagens da substância a alguns besouros desta espécie e verifica se há óbito ou não do inseto. A dosagem da substância fornecida ao

besouro é representada pela variável X e a resistência à substância é representada pela variável dicotômica Y que assume valor *um* em caso de morte do inseto e valor *zero* em caso de sobrevivência do inseto. As 482 informações obtidas pelo pesquisador são apresentadas na Tabela 2.3.

Tabela 2.3: Dados referentes ao Exemplo 1.

Dosagem - x_i	Sobrevivência - y_i	Número de insetos
1,619	0	53
1,619	1	6
1,724	0	47
1,724	1	13
1,755	0	44
1,755	1	18
1,784	0	28
1,784	1	28
1,811	0	11
1,811	1	52
1,837	0	6
1,837	1	53
1,861	0	1
1,861	1	61
1,884	0	0
1,884	1	6

Analisa-se o modelo que relaciona a v.a. dependente Y e a variável independente X. Como Y é uma v.a. binária e X é variável numérica, um modelo razoável para descrever $\mathbb{E}(Y|x)$ é fornecido pela equação (2.6).

O modelo pode ser escrito como

$$y_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} + \varepsilon_i, \tag{2.59}$$

para $i \in \{1, \cdots, 482\}$, onde

$$\mathbb{E}(\varepsilon_i|x_i) = 0 \text{ e } Var(\varepsilon_i|x_i) = \pi(x_i)(1 - \pi(x_i)),$$

devendo-se encontrar os prováveis valores para β_0 e β_1, segundo critérios apresentados na Seção 2.1 deste capítulo. Ou seja, buscam-se os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$.

Aqui também utiliza-se a definição da função *logit* como

$$g(x_i) = \beta_0 + \beta_1 x_i, \tag{2.60}$$

para $i \in \{1, \cdots, 482\}$.

Inicialmente, a Tabela 2.4, apresenta-se resumidamente um teste para o modelo dado por (2.59). No Passo 0 verifica-se a validade da hipótese $\mathcal{H}_0 : \beta_1 = 0$, ou seja, se o coeficiente da variável X_1 é nulo. Esta hipótese é rejeitada, com qualquer nível de significância, porque o respectivo p-valor é nulo, ou seja, zero é o menor nível de significância para o qual se aceita $\mathcal{H}_0$ com base na amostra observada.

Tabela 2.4: Variáveis na equação para o modelo dado por (2.59).

Passo 0	$\hat{\mathcal{B}}$	$\widehat{SE}(\hat{\mathcal{B}})$	W	g.l.	p-valor	$\widehat{\exp}(\hat{\mathcal{B}})$
β_0	0,44	0,09	22,55	1	0,00	1,56
Passo 1	$\hat{\mathcal{B}}$	$\widehat{SE}(\hat{\mathcal{B}})$	W	g.l.	p-valor	$\widehat{\exp}(\hat{\mathcal{B}})$
β_0	-49,83	4,71	111,88	1	0,00	0,00
β_1	28,21	2,64	113,98	1	0,00	$1{,}79\times10^{12}$

Analisando-se os resultados apresentados no Passo 1, percebe-se que a variável independente X é importante ao modelo, visto que seu p-valor é menor que 0,05, ou seja, $0,00 \leq 0,05$. Através de um mesmo argumento, tem-se que β_0, nesta etapa, desempenha um papel importante, pois seu p-valor também é menor do que o nível de significância adotado.

Ainda com base na Tabela 2.4, no Passo 1, na coluna $\hat{\mathcal{B}}$, obtém-se que $\hat{\beta}_0 = -49,83$ e $\hat{\beta}_1 = 28,21$. Com uso das Proposições 2.9 e 2.10, com $\alpha = 0,05$, e das colunas $\hat{\mathcal{B}}$ e $\widehat{SE}(\hat{\mathcal{B}})$ ($\widehat{SE}(\hat{\mathcal{B}})$ indica o desvio padrão estimado) podem ser obtidos os respectivos intervalos de confiança.

O intervalo a 95% de confiança para β_0, denotado por $IC(\beta_0, 95\%)$ é dado por

$$\begin{aligned} IC(\beta_0, 95\%) &= -49,83 \pm 1,96 \times 4,71 \\ &= -49,83 \pm 9,23, \end{aligned}$$

ou seja,

$$IC(\beta_0, 95\%) = [-59,06; -40,60]. \tag{2.61}$$

O intervalo (2.61) pode ser interpretado como o intervalo em que se tem 95% de confiança de se obter o verdadeiro valor para β_0. É interessante notar que o intervalo (2.61) não contém o valor zero. Esta é uma evidência estatística de que a constante não é nula, sendo importante ao modelo dado por (2.59).

O intervalo a 95% de confiança para β_1, denotado por $IC(\beta_1, 95\%)$ é dado por

$$\begin{aligned} IC(\beta_1, 95\%) &= 28,21 \pm 1,96 \times 2,64 \\ &= 28,21 \pm 5,17. \end{aligned}$$

ou seja,

$$IC(\beta_1, 95\%) = [23,04; 33,38]. \tag{2.62}$$

Pode-se dizer, então, que tem-se 95% de confiança de que a variação na função *logit* devido ao acréscimo de uma unidade na variável X está compreendido no intervalo (2.62), ou seja, a variação em uma unidade na dosagem da substância nociva acarreta, para a função *logit*, variação entre 23,04 e 33,38 unidades.

Na Tabela 2.4, a coluna $\widehat{\exp}(\hat{\mathcal{B}})$ indica as razões de chances nas variáveis independente e dependente. Como já foi comentado em seção precedente, ela prediz a alteração por unidade de incremento na correspondente variável independente. Assim, ao se observar novamente o Passo 1, percebe-se que o incremento de uma unidade na variável X, ou seja, o aumento de uma unidade na concentração da substância oferecida ao besouro, acarreta acréscimo na respectiva função *logit* dada por (2.60), visto que seu valor é muito superior a um.

As estatísticas R^2 de Cox e Snell, R^2 de Nagelkerke e *deviance* podem ser obtidas na Tabela 2.5.

Tabela 2.5: Sumário do modelo dado por (2.59).

D	R^2 (Cox e Snell)	R^2 (Nagelkerke)
390,52	0,41	0,55

As estatísticas R^2 apresentadas acima indicam que é pequena a relação linear existente entre as variáveis X e Y visto que seus valores são inferiores a 60%.

A estatística D pode ser utilizada de forma análoga ao $SQRes$ na regressão linear e seu valor será importante em comparação com outros modelos.

O teste qui-quadrado, também denominado Teste de Hosmer e Lemeshow, apresenta a estatística de teste $\chi^2 = 40,37$, com seis graus de liberdade e p-valor=0,00. O valor de 0,00 (maior que 0,05) para o nível de significância indica que se aceita a hipótese de que não há diferença entre os valores preditos e os observados, ou seja, aqui tem-se evidência estatística de que o modelo dado por (2.59) se adecua aos dados em análise.

Procedimentos para a obtenção dos intervalos de confiança para a função *logit* e para $\pi(x)$ são apresentados a seguir.

Inicialmente constrói-se o intervalo a 95% de confiança para a função *logit*, dada por (2.60), quando avaliada, por exemplo, para $x = 1,8$, ou seja, o intervalo de confiança para $\hat{g}(3) = 0,95$.

Utiliza-se a Proposição 2.11 e a expressão (2.26), com o adicional de que $\widehat{Cov}(\hat{\beta}_0, \hat{\beta}_1) = 0$. Os valores de $\widehat{Var}(\hat{\beta}_j)$, com $j \in \{0,1\}$, são obtidos através da Tabela 2.4, por

$$\widehat{Var}(\hat{\beta}_j) = \left[\widehat{SE}(\hat{\beta}_j)\right]^2.$$

Assim, $\widehat{Var}(\hat{\beta}_0) = 22,18$, $\widehat{Var}(\hat{\beta}_1) = 6,96$ e

$$\begin{aligned} \widehat{Var}(\hat{g}(3)) &= \widehat{Var}(\hat{\beta}_0) + (1,8)^2\widehat{Var}(\hat{\beta}_1) \\ &= 22,18 + 3,24 \times 6,96 = 44,73. \end{aligned}$$

Dessa forma, o intervalo a 95% de confiança é dado por

$$\begin{aligned} IC(g(3), 95\%) &= 0,95 \pm 1,96 \times \sqrt{44,73} \\ &= 0,95 \pm 6,61, \end{aligned}$$

ou seja,

$$IC(g(3), 95\%) = [-5,66; 7,56]. \tag{2.63}$$

Assim, tem-se, com confiança de 95%, que o valor do *logit* de uma dosagem equivalente a 1,8 pertence ao intervalo dado por (2.63).

De posse dos valores para $\widehat{Var}(\hat{g}(1,8)) = 44,73$ e $\hat{g}(3) = 0,95$, pode-se apresentar o intervalo a 95% de confiança para $\hat{\pi}(1,8)$, ou seja, a probabilidade logística estimada de que a dosagem de 1,8 cause a morte do inseto.

O valor para $\hat{\pi}(1,8)$ é obtido através da expressão (2.59), ou seja,

$$\begin{aligned}\hat{\pi}(3) &= \frac{\exp[\hat{g}(1,8)]}{1+\exp[\hat{g}(1,8)]} \\ &= \frac{\exp(0,95)}{1+\exp(0,95)} = \frac{2,59}{3,59} \\ &= 0,72. \end{aligned} \tag{2.64}$$

Com base na Proposição 2.12 e nos resultados anteriores, tem-se que

$$\begin{aligned} IC(\pi(1,8),95\%) &= \frac{\exp(0,95 \pm 1,96 \times \sqrt{44,73})}{1+\exp(0,95 \pm 1,96 \times \sqrt{44,73})} \\ &= \left[\frac{0,003}{1+0,003}, \frac{2059,05}{1+2059,05}\right] = [0;1]. \end{aligned} \tag{2.65}$$

O valor encontrado em (2.64) é um estimador da média, ou melhor, da proporção de besouros que receberam a dosagem 1,8 e morreram, na população amostrada. Este valor encontrado informa que existe uma proporção de 72% de besouros que viram a morte após receberem a dosagem de 1,8.

Cada besouro que recebeu a dosagem de 1,8 poderá ou não morrer , entretanto o intervalo apresentado em (2.65) sugere que esta média poderá variar, com nível de confiança de 95%.

Na Figura 2.2, apresenta-se o histograma das probabilidades preditas, o *classplot*. Neste histograma, cada símbolo 0 e 1 representa 5 casos preditos, o eixo das coordenadas indica a freqüência, ou seja, o número de casos classificados. No corpo do gráfico, as colunas apresentam os valores um e zero assumidos pela v.a. Y, denotados, respectivamente, por 0 e 1.

Examinando-se este histograma, pode-se dizer que o modelo classifica relativamente bem os casos de morte e sobrevivência, visto que todas as colunas são longas. Fato este que confere com as informações apresentadas pela Tabela 2.6.

Tabela 2.6: Tabela de classificação para o modelo dado por (2.59).

Obs\Pred	0	1	Correta (%)	Geral (%)
0	143	45	76,1	
1	38	255	87	82,7

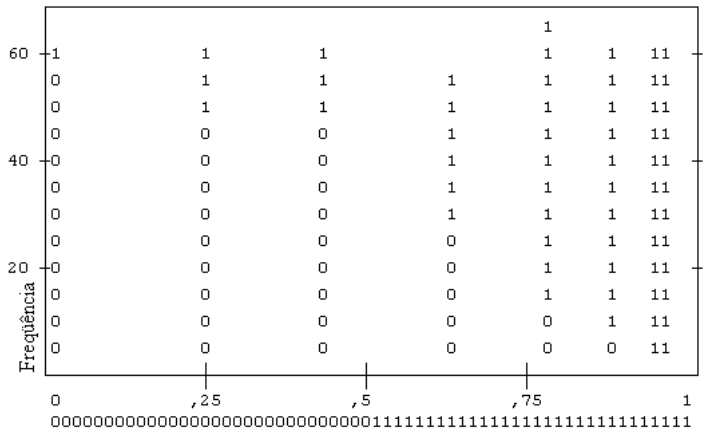

Figura 2.2: Histograma de probabilidades preditas para o modelo dado por (2.59).

A Tabela 2.6 informa que das 188 situações em que o inseto sobreviveu a dosagem fornecida pelo pesquisador apenas 45 delas tiveram sua previsão errada. E que, do total de 293 casos em que houve a morte do besouro, apenas 38 casos tiveram sua previsão errada pelo modelo. Ocasionando dessa forma um acerto total de 82,7%.

Agora, apresenta-se um exemplo no contexto da regressão logística multinomial.

Exemplo 2.5. Um médico realiza uma pesquisa buscando avaliar a eficiência de um medicamento. Para tal realizou um estudo com 83 voluntários dos quais se constatou a idade, o sexo e se o paciente havia recebido o medicamento em questão ou alguma espécie de placebo sem efeito farmacológico. Após o período de atuação da droga verificou-se o grau de melhora para cada paciente.

Os dados referentes ao Exemplo 2.5 encontram-se no Apêndice B.

Definem-se as seguintes variáveis para o exemplo. O grau de melhora do paciente é dado por Y, que é uma v.a. politômica e assume os valores 0, 1 e 2, respectivamente, para as situações de muita, alguma ou nenhuma melhora.

A variável númerica idade será denotada por X_1 e as variáveis categorizadas sexo e medicação serão denotadas, respectivamente, por X_2 e X_3 de forma que

$$x_2 = \begin{cases} 1, & \text{se feminino} \\ 0, & \text{se masculino,} \end{cases}$$

e

$$x_3 = \begin{cases} 1, & \text{se o paciente recebeu medicamento} \\ 0, & \text{se o paciente recebeu placebo.} \end{cases}$$

Um modelo razoável para descrever $\mathbb{P}(Y = k|\mathbf{x})$, onde $k \in \{0, 1, 2\}$ e $\mathbf{x} = (1, x_1, x_2, x_3)$, é fornecido pela expressão (2.40).

Como assume-se o contexto da regressão logística multinomial, uma das categorias da variável resposta deve ser designada como categoria de referência e as demais serão comparadas com esta referência. A escolha é arbitrária e, portanto, assume-se o valor *zero* como categoria de referência.

Primeiramente, definem-se as funções *logit* associadas aos valores $k \in \{0, 1, 2\}$ e $i \in \{1, \cdots, 83\}$,

$$g_{ki} \equiv g_2(\mathbf{x}_i) = \beta_{k0} + \beta_{k1}x_{i1} + \beta_{k2}x_{i2} + \beta_{k3}x_{i3}, \tag{2.66}$$

lembrando que $g_{0i} = g_0(\mathbf{x}_i) = 0$, para qualquer i.

Dessa forma, pode-se apresentar o modelo

$$y_k(\mathbf{x}_i) = \frac{\exp(g_{ki})}{\sum_{k=0}^{2} \exp(g_{ki})} + \varepsilon_i, \tag{2.67}$$

com $i \in \{1, \cdots, 84\}$, $k \in \{0, 1, 2\}$ e

$$\begin{aligned} &(i) \quad \mathbb{E}(\varepsilon_i|\mathbf{x}_i) = 0. \\ &(ii) \quad Var(\varepsilon_i|\mathbf{x}_i) = Var(Y|\mathbf{x}_i). \\ &(iii) \quad Cov(\varepsilon_i, \varepsilon_l) = 0, \text{ se } i \neq l. \end{aligned} \tag{2.68}$$

As estatísticas R^2 obtidas para o modelo dado por (2.67) são apresentadas na tabela a seguir. Em ambas as estatísticas tem-se assegurado de que a relação linear entre a variável dependente e as variáveis independentes é pequena, visto que os valores destas estatística é próximo de 30%.

Tabela 2.7: Sumário do modelo dado por (2.67).

R^2 (Cox e Snell)	R^2 (Nagelkerke)
0,27	0,31

Tabela 2.8: Testes de verossimilhança.

$\mathcal{B}$	D	χ^2	g.l.	p-valor
β_0	146,14	19,69	2	0,000
β_1	132,44	5,99	2	0,050
β_2	133,42	6,97	2	0,031
β_3	140,41	13,96	2	0,001

Algumas estatísticas relacionadas a testes de máxima verossimilhança são apresentadas na Tabela 2.8.

Da Tabela 2.8, pode-se extrair que não há evidência estatística de que os coeficientes que descrevem o modelo são nulos, pois os níveis de significância apresentados nos testes são menores ou iguais ao nível de significância adotado de 5%.

Na Tabela 2.9 são apresentados os estimadores dos parâmetros envolvidos no modelo em análise e algumas estatísticas relacionadas.

Tabela 2.9: Variáveis na equação para o modelo dado por (2.67).

Grupo		$\hat{\mathcal{B}}$	$\widehat{SE}(\hat{\mathcal{B}})$	W	g.l.	p-valor	$\widehat{\exp}(\hat{\mathcal{B}})$
1	β_{10}	-0,68	2,01	0,12	1	0,730	
	β_{11}	0,01	0,03	0,04	1	0,830	1,01
	β_{12}	0,35	0,91	0,15	1	0,700	1,42
	β_{13}	-1,01	0,70	2,09	1	0,150	0,36
2	β_{20}	4,90	1,47	11,06	1	0,001	
	β_{21}	-0,04	0,02	3,90	1	0,001	0,95
	β_{22}	-1,35	0,65	4,26	1	0,039	0,26
	β_{23}	-2,10	0,61	11,94	1	0,001	0,12

Na Tabela 2.9, verifica-se o fato teórico da existência de $q(r+1)$ coeficientes, visto que neste exemplo $q = 2$, pois tem-se três valores para a v.a. Y que fazem analogia com as quantidades 0, 1 e 2. E ainda, as variáveis independentes são em número de três, logo $r = 3$. Totalizando desta forma $2(3+1) = 8$ coeficientes a serem estimados. E, seguindo a analogia com a teoria, os coeficientes na forma β_{0j},

com $j \in \{0, 1, 2\}$, são nulos.

Os valores estimados para tais coeficientes são apresentados na coluna $\hat{\mathcal{B}}$ e seus respectivos desvios-padrão estimados na coluna $\widehat{SE}(\hat{\mathcal{B}})$. Assim, com uso do resultado apresentado na Proposição 2.24, apresentam-se os intervalos de confiança aos parâmetros.

Tabela 2.10: Intervalos a 95% de confiança para os parâmetros do modelo dado por (2.67).

$\mathcal{B}$	$\hat{\mathcal{B}}$	Intervalo de Confiança
β_{10}	-0,68	[-4,61;3,25]
β_{11}	0,01	[-0,05;0,07]
β_{12}	0,35	[-1,43;2,13]
β_{13}	-1,01	[-2,38;0,36]
β_{20}	4,90	[2,02;7,78]
β_{21}	-0,04	[-0,08;0,00]
β_{22}	-1,35	[-2,62;-0,08]
β_{23}	-2,10	[-3,29;-0,91]

Novamente os intervalos de confiança para as funções g_{ki} e π_{ki}, dadas respectivamente por (2.66) e (2.67) são apresentados a seguir. Para tal, supõe-se que um paciente do sexo masculino com 65 anos de idade recebeu tratamento com placebo. Assim, este indivíduo terá seu perfil representado por $\mathbf{x}_0 = (1, 65, 0, 0)$.

Com base nas expressões dadas em (2.66) e nos estimadores dos coeficientes apresentados na Tabela 2.9, tem-se $\hat{g}_1(\mathbf{x}_0) = -0,03$ e $\hat{g}_2(\mathbf{x}_0) = 2,3$.

Para se apresentar um intervalo de confiança para $\hat{g}_1(\mathbf{x}_0)$ e $\hat{g}_2(\mathbf{x}_0)$, precisa-se, inicialmente, avaliar a $\widehat{Var}[\hat{g}_k(\mathbf{x}_0)]$, para $k \in \{1, 2\}$, que é dada pela expressão

$$\widehat{Var}[\hat{g}_k(\mathbf{x}_0)] = \mathbf{x}_0 \widehat{Var}(\hat{\mathcal{B}}_k)\mathbf{x}_0. \tag{2.69}$$

A expressão (2.69) depende da $\widehat{Var}(\hat{\mathcal{B}}_k)$, que é obtida através da matriz de informação e da matriz de variâncias e covariâncias $\mathcal{V} = \widehat{Var}(\hat{\mathcal{B}})$, dadas por (2.57). Na expressão (2.70) apresenta-se a matriz de variâncias e covariâncias para o modelo

dado por (2.67),

$$\mathcal{V} = \begin{pmatrix} 4,04 & -0,05 & -0,79 & -0,31 & 1,36 & -0,02 & -0,22 & -0,20 \\ -0,05 & 0,00 & 0,00 & 0,00 & -0,18 & 0,00 & 0,00 & 0,00 \\ -0,79 & 0,00 & 0,83 & 0,09 & -0,23 & 0,00 & 0,25 & 0,05 \\ -0,31 & 0,00 & 0,09 & 0,49 & -0,14 & 0,00 & 0,02 & 0,19 \\ 1,36 & -0,02 & -0,23 & -0,14 & 2,17 & -0,03 & -0,42 & -0,33 \\ -0,18 & 0,00 & 0,00 & 0,00 & -0,03 & 0,00 & 0,00 & 0,00 \\ -0,22 & 0,00 & 0,25 & 0,02 & -0,42 & 0,00 & 0,43 & 0,11 \\ -0,20 & 0,00 & 0,50 & 0,19 & -0,33 & 0,00 & 0,11 & 0,37 \end{pmatrix}. \tag{2.70}$$

A matriz correspondente a $\widehat{Var}(\hat{\mathcal{B}}_1)$ é obtida pela interseção das quatro primeiras linhas com as quatro primeiras colunas e $\widehat{Var}(\hat{\mathcal{B}}_2)$ é obtida pela interseção das quatro últimas linhas com as quatro últimas colunas da matriz (2.70).

Assim, com base no exposto acima e na expressão (2.69), tem-se que $\widehat{Var}(\hat{\mathcal{B}}_1) = 0,74$ e $\widehat{Var}(\hat{\mathcal{B}}_2) = 1,73$.

Dessa forma, tem-se o intervalo para $g_1(\mathbf{x}_0)$

$$\begin{aligned} IC(g_1(\mathbf{x}_0), 95\%) &= -0,03 \pm 1,96 \times \sqrt{0,74} \\ &= -0,03 \pm 1,68, \end{aligned}$$

ou seja,

$$IC(g_1(\mathbf{x}_0), 95\%) = [-1,71; 1,65], \tag{2.71}$$

e para $g_2(\mathbf{x}_0)$

$$\begin{aligned} IC(g_2(\mathbf{x}_0), 95\%) &= 2,30 \pm 1,96 \times \sqrt{1,73} \\ &= 2,30 \pm 2,57, \end{aligned}$$

ou seja,

$$IC(g_1(\mathbf{x}_0), 95\%) = [-0,27; 4,87]. \tag{2.72}$$

O intervalo de confiança apresentado em (2.71) indica que a função *logit* associada à situação de alguma melhora, para o paciente de sexo masculino, com 65 anos de idade e que foi submetido ao placebo apresenta valores entre -1,71 e 1,65, com 95% de confiança. Um raciocínio semelhante pode ser feito para o intervalo dado em (2.72), ou seja, a função *logit* associada à situação de nenhuma melhora,

para o indivíduo com perfil dado por $\mathbf{x}_0$, possui valor entre -0,27 e 4,87, com 95% de confiança.

Agora, tem-se condições de apresentar intervalos de confiança para as probabilidades $\pi_1(\mathbf{x}_0)$ e $\pi_2(\mathbf{x}_0)$. Utilizando-se a expressão (2.67) e os valores obtidos para $\hat{g}_{10}$ e $\hat{g}_{20}$, tem-se que $\hat{\pi}_1(\mathbf{x}_0) = 0,20$ e $\hat{\pi}_2(\mathbf{x}_0) = 0,37$.

Deste resultado pode-se concluir que um indivíduo que apresente 65 anos de idade, do sexo masculino e que recebeu placebo tem probabilidade predita de 20% de apresentar alguma melhora, e probabilidade predita de 37% de não apresentar melhora e 42% de chances de apresentar muita melhora.

Os intervalos de confiança para as probabilidades preditas são obtidos através da Proposição 2.26, por procedimento semelhante ao realizado anteriormente e correspondem a

$$\begin{aligned} IC(\pi_1(\mathbf{x}_0), 95\%) &= [0,15; 0,83] \\ IC(\pi_2(\mathbf{x}_0), 95\%) &= [0,43; 0,99]. \end{aligned} \tag{2.73}$$

Ou seja, cada paciente de 65 anos, do sexo masculino e que recebeu placebo pode apresentar diferentes níveis de melhora. Entretanto, a probabilidade de apresentar alguma melhora varia entre 15% e 83% e a probabilidade de apresentar muita melhora varia entre 43% e 99%. Assim, em 95% dos casos, os intervalos dados por (2.73) contém, respectivamente, a probabilidade de apresentar alguma e muita melhora.

Outra informação importante é apresentada pela Tabela 2.11, que é uma tabela de classificação de ordem três.

Tabela 2.11: Tabela de classificação para o modelo dado por (2.67).

Obs\Pred	0	1	2	Correta (%)	Geral (%)
0	16	0	11	59,3	
1	6	0	8	0	
2	9	0	33	78,6	59

Na Tabela 2.11, tem-se que, nas previsões de alguma melhora, não houve acertos, ocorrendo para as previsões mehuma melhora um total de 33 acertos que

equivalem a um acerto parcial de 78,6%, e para as revisões em que tem-se muita melhora a porcentagem correta foi de 59,3%. Totalizando, em geral, um acerto de 59%.

Capítulo 3

Análise de Dados Reais

Neste capítulo apresenta-se análise de dados reais, obtidos no endereço eletrônico http://www.biostat.au.dk/teaching/postregCont.htm, através do arquivo, em forma de planilha, prossub.sav, fruto dos estudos do pesquisador e professor Morten Frydenberg na *Faculty of Health Sciencies - Institute of Public Health*, na Dinamarca. Os dados são referentes a exames básicos em pacientes com câncer de próstata em que se verifica ou não a penetração do tumor na cápsula prostática.

Os dados constituem um total de 380 observações e estão organizados como descrito na seqüência. A variável aleatória dependente que indica se o tumor penetrou ou não a cápsula prostática é representada por Y e é denotada por 0 para o caso negativo e 1 para o caso positivo. As demais variáveis independentes verificadas no exame foram X_1, idade em anos, X_2, quantidade de nódulos presentes no exame de toque, X_3, detecção da existência de envoltória capsular prostática ($x_3 = 1$, caso negativo e $x_3 = 2$, caso positivo) e X_4, a concentração do antígeno específico prostático, conhecido popularmente como psa, medido em mg/ml.

Inicialmente apresenta-se a tentativa de descrever o conjuntos de dados seguindo o modelo de regressão linear múltipla. Como já se sabe, de antemão, a natureza dicotômica da variável dependente em estudo, esta etapa deveria ser suprimida em uma situação prática. Entretanto, para fins didáticos faz-se a tentiva de ajuste linear. Os resultados encontrados para esta abordagem são descritos, resumidamente, a seguir.

O modelo de regressão linear, no contexto em análise, pode ser descrito como

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i, \tag{3.1}$$

com $i \in \{1, \cdots, 380\}$ e $\mathbf{x}_i = (1, x_{i1}, x_{i2}, x_{i3}, x_{i4})$, onde

$$\mathbb{E}(\varepsilon_i|\mathbf{x}_i) = 0 \text{ e } Var(\varepsilon_i|\mathbf{x}_i) = \sigma_\varepsilon^2.$$

A variável X_3 é uma variável categorizada e assume-se que

$$x_3 = \begin{cases} 1, & \text{se não há envoltória capsular prostática} \\ 0, & \text{em caso contrário.} \end{cases}$$

As principais estatísticas utilizadas no modelo linear, para avaliar a adequabilidade do modelo, são a estatística F, obtida na tabela ANOVA, e a estatística R^2. É interessante lembrar que uma pode ser obtida em função da outra com o uso da Proposição 1.33.

A Tabela 3.1 apresenta a análise da variância para o modelo dado por (3.1).

Tabela 3.1: Tabela ANOVA para o modelo dado por (3.1).

F.V.	g.l.	SQ	QM	F
Regressão	4	17,25	4,31	21,81
Resíduo	375	74,15	0,20	
Total	379	91,40	4,51	

Com base na Tabela 3.1 e na Proposição 1.33, pode-se obter a estatística $R^2 = 0,19$. Desta informação conclui-se que o modelo proposto diminui a variância residual em quase 20%. Esta informação é confirmada com o uso da equação (1.45) que fornece o valor da estatística $R_a^2 = 0,20$.

Também apresenta-se, para a avaliação do modelo proposto por (3.1), recursos gráficos como o histograma de resíduos, o gráfico quantil-quantil e o gráfico da penetração na cápsula prostática versus idade, respectivamente, nas Figuras 3.1, 3.2 e 3.3.

O histograma de resíduos, apresentado na Figura 3.1, mostra que a distribuição deste é, aproximadamente, binomial e tal fato é confirmado pela aparência do

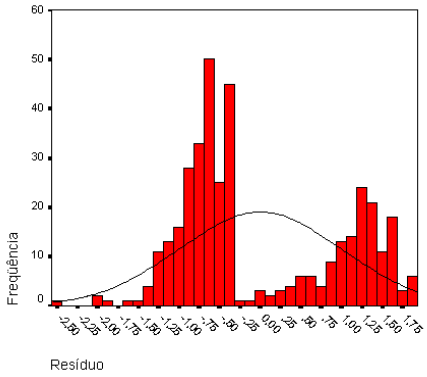

Figura 3.1: Histograma de resíduos para o modelo dado por (3.1).

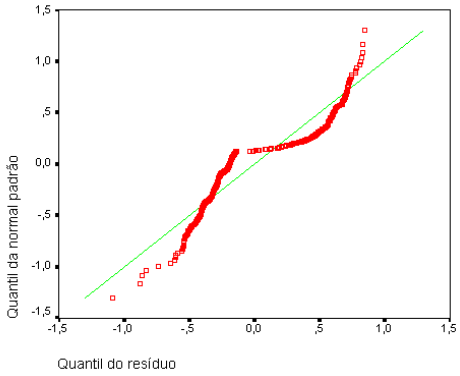

Figura 3.2: Gráfico quantil-quantil para o modelo dado por (3.1).

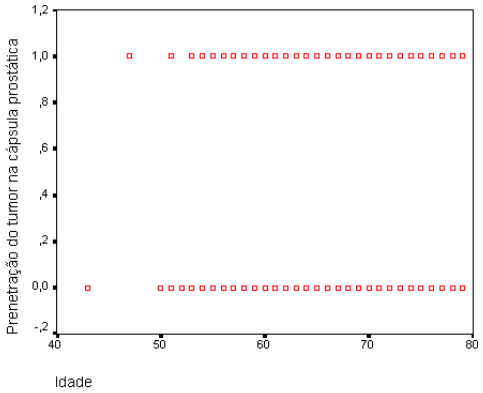

Figura 3.3: Gráfico da penetração na cápsula prostática versus idade.

gráfico quantil-quantil. A Figura 3.3 apresenta forte indício de que a variável Y possui natureza dicotômica.

Uma alternativa adequada é abordar o problema com o uso do modelo de regressão logística. Desta forma, utilizam-se as mesmas variáveis já descritas, para o seguinte modelo

$$y_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{i3} x_{i3} + \beta_4 x_{i4})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{i3} x_{i3} + \beta_4 x_{i4})} + \varepsilon_i, \tag{3.2}$$

para $i \in \{1, \cdots, 380\}$ e $\mathbf{x}_i = (1, x_{i1}, x_{i2}, x_{i3}, x_{i4})$, onde

$$\mathbb{E}(\varepsilon_i | \mathbf{x}_i) = 0 \text{ e } Var(\varepsilon_i | \mathbf{x}_i) = \pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i)).$$

Deseja-se encontrar os prováveis valores para β_i, com $i \in \{0, 1, 2, 3, 4\}$, segundo critérios do Capítulo 2. A variável X_3 é uma variável categorizada e assume-se, novamente, que se

$$x_3 = \begin{cases} 1, & \text{se não há envoltória capsular prostática} \\ 0, & \text{em caso contrário.} \end{cases}$$

Define-se a função *logit* como

$$g(\mathbf{x}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{i3} x_{i3} + \beta_4 x_{i4}, \tag{3.3}$$

com $i \in \{1, \cdots, 380\}$ e $\mathbf{x}_i = (1, x_{i1}, x_{i2}, x_{i3}, x_{i4})$.

Uma primeira avaliação do modelo dado por (3.2) pode ser obtida através dos estimadores para os parâmetros envolvidos e suas respectivas estatísticas estão representados na Tabela 3.2.

No Passo 0, apresenta-se um teste inicial para o modelo dado por (3.2), no qual verifica-se a hipótese $\mathcal{H}_0 : \mathcal{B} = \mathbf{0}$. Esta hipótese é rejeitada, com qualquer nível de significância, porque o respectivo p-valor é nulo, ou seja, zero é o menor nível de significância para o qual se aceita $\mathcal{H}_0$, com base na amostra observada.

Agora, ao se analisar o Passo 1 da Tabela 3.2, pode-se afirmar que as variáveis independentes X_2, X_3 e X_4 são importantes ao modelo, pois seus respectivos p-valores são menores do que 0,05. Entretanto, a variável X_1 e a constante, representadas, respectivamente, por β_1 e β_0, não são importantes ao modelo porque os respectivos p-valores são maiores de que o nível de significância adotado (que é de 0,05). Ou seja, a idade (em anos) não está sendo uma variável importante ao

Tabela 3.2: Variáveis na equação para o modelo dado por (3.2) - Parte I.

Passo 0	$\hat{\mathcal{B}}$	$\widehat{SE}(\hat{\mathcal{B}})$	W	g.l.	p-valor	$\widehat{\exp}(\hat{\mathcal{B}})$
β_0	-0,39	0,10	14,22	1	0,00	0,67
Passo 1	$\hat{\mathcal{B}}$	$\widehat{SE}(\hat{\mathcal{B}})$	W	g.l.	p-valor	$\widehat{\exp}(\hat{\mathcal{B}})$
β_0	-0,95	1,29	0,54	1	0,46	0,39
β_1	-0,01	0,02	0,22	1	0,64	0,99
β_2	0,59	0,13	22,39	1	0,00	1,81
β_3	-0,99	0,43	5,43	1	0,02	0,37
β_4	0,04	0,01	21,48	1	0,03	1,04

modelo, ao passo que as informações dadas pelos exames básicos de detecção da quantidade de nódulos e da existência de envoltória capsular prostática e a taxa de psa estão sendo importantes ao modelo.

Como o p-valor da variável X_1 é superior ao p-valor da constante, retira-se a variável X_1 do modelo proposto em (3.2) e, na seqüência, realiza-se novamente a mesma análise, que é apresentada na Tabela 3.3 abaixo.

Tabela 3.3: Variáveis na equação para o modelo dado por (3.2) - Parte II.

Passo 0	$\hat{\mathcal{B}}$	$\widehat{SE}(\hat{\mathcal{B}})$	W	g.l.	p-valor	$\widehat{\exp}(\hat{\mathcal{B}})$
β_0	-0,40	0,11	14,23	1	0,00	0,67
Passo 1	$\hat{\mathcal{B}}$	$\widehat{SE}(\hat{\mathcal{B}})$	W	g.l.	p-valor	$\widehat{\exp}(\hat{\mathcal{B}})$
β_0	-1.50	0,55	7,42	1	0,01	0,22
β_2	0,60	0,13	22,56	1	0,00	1,81
β_3	-0,99	0,43	5,40	1	0,02	0,37
β_4	0,04	0,01	21,44	1	0,03	1,04

Na Tabela 3.3, novamente, no Passo 0, apresenta-se um teste inicial para o modelo dado por (3.2). Entretanto, o teste é realizado sem a presença da idade do paciente, ou seja, sem a variável X_1. Verifica-se a hipótese de que todos os coefientes são nulos. Esta hipótese é rejeitada, com qualquer nível de significância, porque o respectivo p-valor é nulo, ou seja, zero é o menor nível de significância

para o qual se aceita a possibilidade de todos os coeficientes serem zero, com base na amostra observada.

Ao se analisar o Passo 1 da Tabela 3.3, pode-se concluir que as variáveis independentes X_2, X_3 e X_4 e a constante são importantes ao modelo, pois seus respectivos p-valores são menores do que 0,05. Dessa forma, os exames de detecção do número de nódulos no exame de toque e da presença de envoltória capsular prostática e o exame de psa são importantes ao modelo, ou melhor, o modelo dado por (3.2) fica melhor descrito sem a presença da variável idade.

Assim, os estimadores para os parâmetros envolvidos e suas respectivas estatísticas ficam representados pela Tabela 3.3.

Ainda com base na Tabela 3.3, a coluna $\widehat{\exp}(\hat{\mathcal{B}})$ indica as razões de chance nas variáveis independentes e dependente. O valor de $\widehat{\exp}(\hat{\mathcal{B}})$ para X_2 é 1,81 e isto significa que a variação em uma unidade no valor da variável X_2 provoca aumento no valor da função *logit* dada por (3.3). Ou melhor, a detecção de um nódulo a mais no exame de toque provoca um aumento multiplicativo de 1,81 nas probabilidades de haver a penetração do câncer na cápsula prostática. O valor de $\widehat{\exp}(\hat{\mathcal{B}})$ para X_3 é de 0,37 e isto acarreta que a variação em uma unidade no valor da variável X_3 provoca decréscimo na função *logit*. Ou seja, a detecção de envoltória capsular prostática propicia uma diminuição na probabilidade de haver penetração do câncer na cápsula prostática com fator multiplicativo de 0,37. E, finalmente, o valor de $\widehat{\exp}(\hat{\mathcal{B}})$ para X_4 muito próximo da unidade informa que a variação de uma unidade na variável X_4 não traz mudanças significativas na função *logit*, ou ainda, que a variação de apenas uma unidade no resultado do exame de psa não produz mudanças significativas nas probabilidades de se ter ou não a penetração do tumor na cápsula prostática.

Para se obter os intervalos a 95% de confiança para os parâmetros utiliza-se a Proposição 2.16 e as informações apresentadas nas colunas $\hat{\mathcal{B}}$ e $\widehat{SE}(\hat{\mathcal{B}})$ da Tabela 3.3. Os resultados obtidos estão apresentados na Tabela 3.4.

A interpretação dos intervalos de confiança, obtidos pela Tabela 3.4, é apresentada na seqüência.

Pode-se afirmar, por exemplo, com 95% de confiança de que o verdadeiro valor para o parâmetro β_0 (a constante do modelo) se encontra entre os valores -1,50 e 0,42. E, também com 95% de confiança que a variação na função *logit* devido

Tabela 3.4: Intervalos a 95% de confiança para os parâmetros do modelo dado por (3.2).

$\mathcal{B}$	$\hat{\mathcal{B}}$	Intervalo de Confiança
β_0	-1,50	[-2,58;0,42]
β_2	0,60	[0,35;0,85]
β_3	-0,99	[-1,83;-0,15]
β_4	0,04	[0,02;0,06]

ao acréscimo de uma unidade na variável X_2 está compreendido no intervalo de 0,60 a 0,85 unidades, ou seja, o aumento em uma unidade no número de nódulos percebido no exame de toque propicia um aumento compreendido entre 0,60 e 0,85 unidades na função *logit*.

Seguindo este mesmo raciocínio, pode-se afirmar que a variação em uma unidade no resultado do exame de psa não afeta significativamente a função *logit*, pois o coeficiente da variável X_4, com 95% de confiança, tem valores compreendidos entre 0,02 e 0,06. E, ainda que a detecção de envoltória capsular prostática provoca queda no valor da função *logit* porque o coeficiente da variável X_3 está conpreendido entre os valores -1,83 e -0,15, com 95% de confiança.

As estatísticas R^2 são apresentadas na Tabela 3.5 e indicam que pouco é explicado pelo modelo dado por (3.2) porque seus valores são menores do que 0,30, ou seja, 30%.

Tabela 3.5: Sumário do modelo dado por (3.2).

R^2 (Cox e Snell)	R^2 (Nagelkerke)
0,20	0,27

O teste de Hosmer e Lemeshow para o modelo em análise apresenta a estatística $\chi^2 = 16,04$ com oito graus de liberdade e p-valor de 0,04. O valor de 0,04 para o nível de significância indica que se rejeita a hipótese de que há diferença entre os valores preditos e observados.

Na seqüência apresenta-se exemplo de construção de intervalo a 95% de con-

fiança para a função $g(\mathbf{x}_0)$, dada por (3.3), e para $\pi(\mathbf{x}_0)$, dada por

$$\pi(\mathbf{x}) = \frac{\exp[g(\mathbf{x})]}{1 + \exp g(\mathbf{x})}, \tag{3.4}$$

para a situação em que $\mathbf{x}_0 = (1, 70, 1, 0, 40)$. O valor de $\mathbf{x}_0$ representa um paciente com 70 anos de idade que no exame de toque apresentou presença de um nódulo, não apresentou envoltória capsular prostática e sua taxa de psa é de 40mg/ml.

O valor $\hat{g}(\mathbf{x}_0) = 0,70$ é obtido através da expressão (3.3) com base nas informações da Tabela 3.3. É interessante lembrar que a idade, ou melhor, a variável X_1, não está integrando o modelo em análise e não fará parte dos cálculos efetuados.

Para se obter o intervalo de confiança para $\hat{g}(\mathbf{x}_0)$ é necessário avaliar sua variância $\widehat{Var}[\hat{g}(\mathbf{x}_0)]$. Utiliza-se a expressão (2.42) e o fato de que $Cov(\beta_j, \beta_l) = 0$, para $j, l \in \{0, 2, 3, 4\}$. Para se obter os valores de $\widehat{Var}(\hat{\beta}_j)$, para $j \in \{0, 2, 3, 4\}$, utiliza-se a Tabela 3.3 e a expressão

$$\widehat{Var}(\hat{\beta}_j) = \left[\widehat{SE}(\hat{\beta}_j)\right]^2.$$

Assim, $\widehat{Var}[\hat{g}(x_0)] = 0,32$.

Usando-se a Proposição 2.17 e $\widehat{Var}[\hat{g}(\mathbf{x}_0)] = 0,32$, tem-se que

$$IC(\hat{g}(\mathbf{x}_0), 95\%) = [-0,39; 1,79].$$

Agora, tem-se condições de se apresentar o intervalo a 95% de confiança para $\pi(\mathbf{x}_0)$, ou seja, a probabilidade logística estimada para o paciente que apresenta o perfil dado por $\mathbf{x}_0$.

O valor de $\hat{\pi}(\mathbf{x}_0) = 0,67$ é obtido através da expressão (3.2) usando-se dos valores $\widehat{Var}[\hat{g}(\mathbf{x}_0)] = 0,32$ e $\hat{g}(\mathbf{x}_0) = 0,70$.

Deste resultado pode-se concluir que a probabilidade estimada de um paciente, com o perfil dado por $\mathbf{x}_0$, apresentar penetração do tumor na cápsula prostática é de 67%.

Com base na Proposição 2.18 e nos resultados acima, tem-se que

$$IC(\pi(\mathbf{x}_0), 95\%) = [0,40; 0,86].$$

Cada paciente com perfil dado por $\mathbf{x}_0$, ou seja, cada paciente com 70 anos de idade e que no exame de toque apresentou presença de um nódulo, não apresentou envoltória capsular prostática e sua taxa de psa é de 40mg/ml, poderá ou

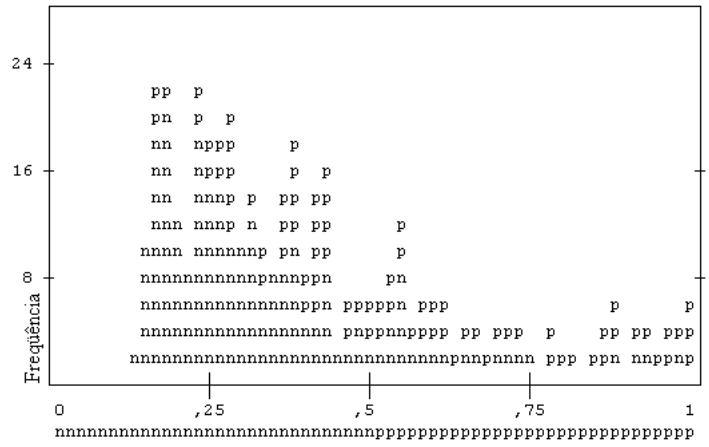

Figura 3.4: Histograma de probabilidades preditas para o modelo dado por (3.2).

não apresentar penetração do tumor na cápsula prostática, entretanto, o intervalo $IC(\pi(\mathbf{x}_0), 95\%)$ sugere que o modelo dado por (3.2) prevê que as chances de apresentar a penetração do tumor na cápsula prostática podem variar entre 40% e 86%.

Na Figura 3.3, apresenta-se o histograma das probabilidades preditas. Neste histograma, cada símbolo n e p representam 2 casos preditos. Onde n representa a situação em que a v.a. Y assume valor zero e p representa a situação em que a v.a. Y assume valor um.

No *classplot* pode-se verificar que as probabilidades com valores inferiores a 0,5 estão sendo preditas de forma satifatória e que as probabilidades com valores acima de 0,5 não estão sendo preditas corretamente. Ou seja, vários pacientes tiveram a previsão de que o tumor penetrou na cápsula prostática quando de fato isto não ocorreu e os pacientes em que a previsão indicou a não penetração do tumor na cápsula prostática tiveram sua previsão acertada.

Na seqüência, apresenta-se a Tabela 3.6 que traz uma comparação entre os valores preditos e observados em relação ao modelo dado por (3.2), e pode ser encarada como um resumo do modelo em estudo.

Da Tabela 3.6 pode-se concluir que das 227 observações de que não haveria penetração do tumor na cápsula prostática houve um acerto de 100%. As 153 observações em que houve penetração do tumor na cápsula prostática foram prevista erradas. Esta tabela ratifica as conclusões apresentadas sobre o *classplot*.

Tabela 3.6: Tabela de classificação para o modelo dado por (3.2).

Obs\Pred	0	1	Correta (%)	Geral (%)
0	227	0	100	
1	153	0	0	59,7

De forma geral, o modelo dado por (3.2) apresentou algumas limitações para o tratamento dos dados analisados. Tais limitações poderiam ser melhoradas com o aumento do número de observações coletadas e com a utilização de variáveis adicionais ao modelo, como por exemplo, número de casos da doença na família na primeira geração anterior ao paciente, peso corporal e propenção à infecção urinária. No entanto, apresentou vantagens em relação ao modelo dado por (3.4) pois incorpora a hipótese de que os erros seguem distribuição binomial e de que a variável aleatória dependente possui natureza nominal. Portanto, vê-se claramente a necessidade do uso da regressão logística dicotômica.

Capítulo 4

Conclusões Finais

Historicamente as técnicas de regressão tem se firmado como forma de abordagem para tratamento e análise de informações coletadas em experimentos e pesquisas científicas.

A mais conhecida e difundida modalidade é a regressão linear que possui vasto campo de aplicações. Entretanto, situações em que a variável dependente possui natureza nominal não são descritas de forma satisfatória pelo modelo linear.

Uma abordagem alternativa neste contexto é o modelo de regressão logística. Neste trabalho apresentou-se os modelos de regressão logística em sua forma mais simples (caso bivariado em que figuram apenas uma variável dependente dicotômica e uma variável independente) e nas formas múltipla (onde tem-se uma variável dependente dicotômica e mais de uma variável independente) e multivariada (em que há mais de duas categorias na descrição da variável dependente).

O modelo de regressão linear simples considera o caso em que se deseja desenvolver um modelo estatístico, baseado numa função afim, para prever valores de uma variável aleatória dependente Y, de natureza contínua, em função de uma variável independente X também contínua.

Para tal se dispõe de n observações de pares de valores das variáveis X e Y onde, por definição, a esperança de Y dado que $X = x$ é dada pela função afim quando avaliada em x. Entretanto, os valores preditos pelo modelo nem sempre equivalem aos valores observados na amostra de tamanho n. Esta diferença entre tais valores é a variável aleatória chamada erro.

Devido a isto se faz necessário assumir algumas suposições com respeito às variáveis aleatórias envolvidas:

(i) A variável X é por hipótese controlada e não está sujeita a variações aleatórias. Diz-se que X é uma variável fixa ou determinística.

(ii) Para dado valor x de X, os erros, ε distribuem-se ao redor da média $\beta_0 + \beta_1 x$ com média zero, isto é, $\mathbb{E}(\varepsilon_i|x_i) = 0$, para todo $i \in \{1, \cdots, n\}$.

(iii) Os erros devem ter variabilidade constante em torno de X, ou melhor, $Var(\varepsilon_i|x_i) = \sigma_\varepsilon^2$, para todo $i \in \{1, \cdots, n\}$.

(iv) Os erros são não-correlacionados.

Em algumas situações se faz necessário suposições, para quaisquer pares (i, j) tais que $1 \leq i, j \leq n$, tais como:

• Caso A: assume-se que a função de distribuição da variável Y é normal e que as v.a. Y_i e Y_j são independentes, para todo $i, j \in \{1, \cdots, n\}$, com $i \neq j$.

• Caso B: assume-se que as v.a. Y_i e Y_j são não-correlacionadas, para todo $i, j \in \{1, \cdots, n\}$, com $i \neq j$.

Para se obter um modelo de regressão linear é necessário apresentar os estimadores para os coeficientes da função afim (parâmetros).

Ao se assumir o Caso A, os estimadores são obtidos através do método da Máxima Verossimilhança, visto que se dispõe de uma função de distribuição.

Os estimadores obtidos possuem propriedades estatísticas importantes. São não viciados e possuem uniforme mínima variância. Ou seja, de todos os estimadores não-viciados possíveis para os parâmetros do modelo, os estimadores encontrados pelo método de Máxima Verossimilhança apresentam a menor variância possível, para o modelo de regressão linear.

No Caso B, a função de distribuição não é conhecida, dessa forma, não há como se definir os estimadores de máxima verossimilhança. Em situações como esta, um método adequado é o dos Mínimos Quadrados.

Este método consiste em minimizar a soma dos quadrados dos erros, ou seja, minimizar a soma dos quadrados das diferenças entre os valores preditos e observados.

Coincidentemente, ao se assumir que os erros são gaussianos, os estimadores obtidos pelo método dos Mínimos Quadrados são iguais (numericamente) aos obtidos pelo método de Máxima Verossimilhança.

Os estimadores obtidos pelo método dos Mínimos Quadrados são não viciados mas, não mantém a propriedade de uniforme mínima variância (propriedade esta que tem natureza global). O método dos Mínimos Quadrados assegura aos estimadores do modelo de regressão linear a propriedade de melhor estimador linear não-viciado, ou seja, a propriedade de mínima variância só é assegurada dentre todos os estimadores lineares. Este resultado é conhecido como Teorema de Gauss-Markov.

O modelo de regressão linear múltipla possui a mesma estrutura teórica que o modelo de regressão linear simples, com a diferença de que neste novo contexto existe mais de uma variável independente mas, ainda se mantém apenas uma variável dependente.

Para o modelo de regressão linear multivariada, considera-se o problema de se modelar a relação entre m variáveis independentes com base em um conjunto de r variáveis dependentes. A abordagem teórica é assumir cada variável dependente em conjunto com as variáveis independentes, como um caso de regressão linear múltipla e, devido a este artifício, os resultados são análogos ao modelo de regressão linear múltipla.

De posse dos estimadores para o modelo linear, é, então necessário julgar se o modelo obtido é bom ou adequado.

Uma abordagem que pode ser utilizada é a Análise da Variância através da Tabela ANOVA. Esta tabela fornece estatísticas e valores para se testar hipóteses a respeito dos coeficientes do modelo e fornece subsídios para se avaliar os resíduos do modelo.

Os resíduos também podem ser empregados em diferentes tipos de gráficos com o objetivo de detectar anomalias no modelo.

Entretanto, mesmo uma ferramenta tão poderosa como a regressão linear apre-

senta suas limitações, não sendo adequada, por exemplo, nas situações em que a variável em estudo possui natureza nominal e, por este motivo, não assume valores no conjunto dos números reais ou em intervalos da reta.

Adequada ao contexto em que a variável aleatória dependente assume um número finito de valores, a regressão logística se constitui em uma ferramenta bastante eficaz.

Devido a este fato, existem diferenças entre os dois modelos, que vão das suposições envolvidas às funções que os descrevem.

Neste trabalho foram abordados os modelos de regressão logística binária, múltipla e multinomial, nos quais a variável dependente envolvida possui natureza nominal.

No modelo de regressão logística binária, de forma análoga ao modelo linear simples, busca-se apresentar um modelo estatístico baseado na função definida em (2.3).

Esta função é exponencial para os coeficientes e a variável independente. Dessa forma, define-se uma função auxiliar chamada *logit* que possui propriedades semelhantes à função afim do modelo de regressão linear. Ou seja, o *logit* é uma função linear nos parâmetros e na variável independente, dessa forma, tem-se um caso particular de modelo linear generalizado.

Para se estimar os coeficientes da função *logit* dispõe-se de uma amostra de n observações de pares de variáveis X e Y. A variável Y, no caso binário, assume apenas dois valores que podem ser representados por zero e um. Onde *zero* denota a ausência de uma determinada característica e *um* denota a ocorrência desta característica.

Assim como comentado para o caso da regressão linear, os erros correspondem, na regressão logística, à diferença entre os valores preditos e observados, sendo então uma variável aleatória.

As suposições para as variáveis envolvidas no modelo de regressão logística são de que X é por hipótese controlada e não sujeita a variações e que para um dado valor x de X, os erros seguem distribuição binomial com média zero e variância igual a $Var(Y|x)$.

Na regressão logística os estimadores são obtidos pelo método da Máxima Verossimilhança e, com base na função de verossimilhança, podem ser apresentadas diversas estatísticas que são utilizadas na avaliação do modelo e em testes de hipóteses.

O modelo de regressão logística múltipla possui analogias teóricas ao modelo de regressão logística binária. Aqui o modelo emprega uma variável dependente de natureza binária e mais de uma variável independente. Alguns dos resultados são obtidos através de álgebra matricial.

Para o modelo de regressão logística multinomial considera-se o problema de se modelar uma variável aleatória dependente Y que pode assumir $q + 1$ valores com base em sua relação com as r variáveis independentes. Neste caso os erros apresentam disbribuição multinomial com média zero e variância $Var(Y|x)$.

A função *logit* é definida com base na comparação entre duas probabilidades, conforme indica a expressão (2.43).

De posse dos coeficientes, o modelo logístico pode ser avaliado através de testes de máxima verossimilhança, testes qui-quadrado e das diferentes estatísticas R^2.

Outros recursos que estão disponíveis são as tabelas de classificação e os histogramas de probabilidades preditas. Tais meios constituem uma maneira eficaz de se avaliar quando as probabilidades preditas estão corretamente determinadas ou não.

Os exemplos desenvolvidos nos capítulos 1 e 2, nas seções de aplicação da teoria, serviram como base para se perceber como se processam os modelos de regressão e foram estímulo para o manuseio do *software* SPSS 10.0.

Na análise de dados reais procurou-se repetir de forma resumida os procedimentos realizados nos exemplos. Percebeu-se que os dados utilizados não se adequavam ao modelo de regressão linear, pelo fato da variável dependente ter natureza nominal. O modelo de regressão logística utilizado pareceu ser mais adequado em comparação com o linear. Entretanto, o uso de outros procedimentos de regressão associados ao modelo utilizado e a inserção de outras variáveis envolvidas, por exemplo, podem ajudar a melhorar a análise obtida.

Referências

[1] Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. New York: John Wiley.

[2] Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley.

[3] Andersen, E.B. (1996). *Introduction to the Statistical Analysis of Categorical Data*. New York: Springer.

[4] Bickel, P.J. e Doksum, K.A. (1976). *Mathematical Statistics - Basic Ideas and Selected Topics*. Sydney: Holden-Day.

[5] Bisquerra, R., Sarriera, J.C. e Martínez, F. (2004). *Introdução à Estatística - Enfoque Informático com o Pacote Estatístico SPSS*. Porto Alegre: Editora Artmed.

[6] Box, G.E.P. (1949). "A General Distribution Theory for a Class of Likelihood Criteira". *Biometrika*, Vol. 36, 317-346.

[7] Bussab, W.O. e Morettin, P.A. (2004). *Estatística Básica*. São Paulo: Editora Saraiva, 5^a Edição.

[8] Casella, G. e Berger, R.L. (2002). *Statistical Inference*. Pacifc Grove: Duxbury, 2^a Edição.

[9] Chatfield, C. e Collins, A.J. (1980). *Introduction to Multivariate Analysis*. Science Paperbacks. Cambridge: University Press.

[10] Cox, D.R. e Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman & Hall.

[11] Cox, D.R. e Snell, E.J. (1989). *Analysis of Binary Data*. London: Chapman & Hall, 2^a Edição.

[12] Dixon, W.J. e Massey, F.J. (1957). *Introduction to Statistical Analysis*. New York: Mc-Graw Hill, 2^a Edição.

[13] Dobson, A. (1990). *An Introduction to Generalized Models*. London: Chapman & Hall.

[14] Draper, N.R. e Smith, H. (1981). *Applied Regression Analysis*. New York: John Wiley, 2^a Edição.

[15] Galton, F. (1885). "Regression Toward Mediocrity in Heredity Stature". *Journal of the Anthropological Institute*, Vol. 15, 246-263.

[16] Hagle, T.M. e Mitchell, G.E. (1992). "Goodness-of-fit Measures for Probit e Logit." *American Journal of Political Science*, Vol. 36, 762-784.

[17] Hauck, W.W. e Donner, A. (1977). "Wald's Test as Applied to Hypotheses in Logit Analysis". *Journal of the American Statistical Association*, Vol. 72, 851-853.

[18] Hosmer, D.W. e Lemeshow, S. (2000). *Applied Logistic Regression*. New York: John Wiley, 2^a Edição.

[19] Jennings, D.E. (1986a). "Judging Inference Adequacy in Logistic Regression". *Journal of the American Statistical Association*, Vol. 81, 471-476.

[20] Jennings, D.E. (1986b). "Outliers and Residual Distributions in Logistic Regression". *Journal of the American Statistical Association*, Vol. 81, 987-990.

[21] Johnson, R.A. e Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis*. New Jersey: Prentice Hall, 4^a Edição.

[22] Kleinbaum, D.G. e Klein, M. (2002). *Logistic Regression - A Self-learning Text*. New York: Spring, 2^a Edição.

[23] Marquês de Sá, J.P. (2003). *Applied Statistics Using SPSS, STATISTICA, and MATLAB*. New York: Springer-Verlag.

[24] Martins, G.A. (2002). *Estatística Geral e Aplicada*. São Paulo: Editora Atlas, 2^a Edição.

[25] McCullagh, P. e Nelder, J.A. (1989). *Generalized Linear Models*. London: Chapman & Hall, 2^a Edição.

[26] Menard, S. (2002). *Applied Logistic Regression Analysis*. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-106, Thousand Oaks: Sage.

[27] Menard, S. (2000). "Coefficients of Determination for Multiple Logistic Regression Analysis". *American Statistician*, Vol. 54, 17-24.

[28] Mood, A.M., Graybill, F.A. e Boes, D.C. (1986). *Introduction to the Theory of Statistics - International Student Edition*. Singapore: Mc-Graw Hill, 3^a Edição.

[29] Nagelkerke, N.J.D. (1991). "A Note on a General Definition of the Coefficient of Determination". *Biometrika*, Vol. 78, n^o 3, 691-692.

[30] Pampel, F.C. (2000). *Logistic Regression - A primer*. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-132. Thousand Oaks: Sage.

[31] Rao, C.R. (1973). *Linear Statistical Inference and Its Application*. New York: John Wiley, 2^a Edição.

[32] Smith, J.Z. (2001). *Multiple Logistic Regression: Predicting Coronary Heart Disease*. Monografia de Conclusão de Bacharelado em Matemática, Southern Oregon University, Ashland. Disponível em: http://martini.nu/justin/SOUthesis.htm. Acesso em 26/ago/2005.

[33] Veall, M.R. e Zimmerman, K.F. (1996). "Pseudo-R^2 Measures for Come Common Limited Dependent Variable Models". *Journal of Economic Surveys*, Vol. 10, 241-260.

Apêndice A

O desenvolvimento deste apêndice estabelece o resultado apresentado na Proposição 1.27.

Sabe-se que $n\hat{\mathbf{\Sigma}} = \mathbf{Y}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}$ e que sobre $\mathcal{H}_0$, $n\hat{\mathbf{\Sigma}}_1 = \mathbf{Y}(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{Y}$ com $\mathbf{Y} = \mathbf{X}_1\mathcal{B}_1 + \mathcal{E}$.

Seja $\mathbf{A} = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']$.

Como se sabe que

$$\begin{aligned} \mathbf{0} &= [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{X} \\ &= [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][\mathbf{X}_1|\mathbf{X}_2] \\ &= [\mathbf{A}\mathbf{X}_1|\mathbf{A}\mathbf{X}_1], \end{aligned}$$

então, tem-se que as colunas de $\mathbf{X}$ são perpendiculares a $\mathbf{A}$.

Dessa forma, pode-se escrever que

$$\begin{aligned} n\hat{\mathbf{\Sigma}} &= (\mathbf{X}\mathcal{B} + \mathcal{E})'\mathbf{A}(\mathbf{X}\mathcal{B} + \mathcal{E}) = \mathcal{E}'\mathbf{A}\mathcal{E} \\ n\hat{\mathbf{\Sigma}}_1 &= (\mathbf{X}_1\mathcal{B}_1 + \mathcal{E})'\mathbf{A}_1(\mathbf{X}_1\mathcal{B}_1 + \mathcal{E}) = \mathcal{E}'\mathbf{A}_1\mathcal{E}, \end{aligned}$$

onde $\mathbf{A}_1 = \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$.

Utiliza-se o processo de ortogonalização de Gram-Schmidt para se construir um conjunto de vetores ortonormais, representados por $\mathbf{G} = [\mathbf{g}_1|\cdots|\mathbf{g}_{q+1}]$, com base nas colunas de $\mathbf{X}_1$. Na seqüência, obtém-se um conjunto ortonormal de vetores baseado em $[\mathbf{G}|\mathbf{X}_2]$, e finalmente completa-se este conjunto obtido com $n - r - 1$ vetores ortonormais arbitrários que são ortogonais aos vetores previamente obtidos.

Ou seja, consegue-se $\mathbf{g}_1, , \mathbf{g}_{q+1}$ vetores ortonormais com base nas colunas de $\mathbf{X}_1$, $\mathbf{g}_{q+2}, \cdots, \mathbf{g}_{r+1}$ vetores ortonormais com base nas colunas de $\mathbf{X}_2$ e perpendiculares às colunas de $\mathbf{X}_1$ e, $\mathbf{g}_{r+2}, \cdots, \mathbf{g}_n$ vetores ortonormais arbitrários e ortogonais às colunas de $\mathbf{X}$.

Seja $(\lambda, \mathbf{e})$ um par de autovalor e autovetor correspondente à matriz idenpotente

$\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$. Segue que

$$\begin{aligned}\lambda\mathbf{e} &= \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{e} \\ &= [\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1']^2\mathbf{e} \\ &= \lambda[\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1']\mathbf{e} \\ &= \lambda^2\mathbf{e},\end{aligned}$$

e, por este motivo, os autovalores de $\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$ assumem apenas valores zero ou um.

Entretanto,

$$\begin{aligned}tr[\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'] &= tr[(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_1] \\ &= tr\mathbf{I}_{(q+1)\times(q+1)} = q+1 \\ &= \lambda_1+\lambda_2+\cdots+\lambda_{q+1},\end{aligned}$$

onde $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{q+1} > 0$ são os autovalores de $\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$, ou seja, esta matriz possui $q+1$ autovalores iguais a 1.

Como $[\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1']\mathbf{X}_1 = \mathbf{X}_1$ então qualquer combinação linear de $\mathbf{X}_1\mathbf{a}_l$ de comprimento unitário é um autovetor correspondente a um autovalor 1.

Os vetores ortonormais $\mathbf{g}_l$, com $l \in \{1,\cdots,q+1\}$, são autovetores de $\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$ visto que são formados por combinações lineares particulares de $\mathbf{X}_1$. Pelo Teorema da Decomposição Espectral, decorre que

$$\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1' = \sum_{l=1}^{q+1}\mathbf{g}_l\mathbf{g}_l'.$$

Similarmente, representando $\mathbf{X} = [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{X}$, pode-se afirmar que a combinação linear $\mathbf{X}\mathbf{a}_l - \mathbf{g}_l$, por exemplo, é um autovetor de $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ com autovalor 1 e, dessa forma, conclui-se que $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \sum_{l=1}^{r+1}\mathbf{g}_l\mathbf{g}_l'$.

Continuando, tem-se que $\mathbf{A}\mathbf{X} = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{X} = \mathbf{X} - \mathbf{X} = \mathbf{0}$, assim $\mathbf{g}_l = \mathbf{X}\mathbf{a}_l$, com $l \leq r+1$, são autovetores de $\mathbf{A}$ com autovalores nulos. E, por construção $\mathbf{g}_l$, com $l > r+1$, implica $\mathbf{X}'\mathbf{g}_l = \mathbf{0}$. O que resulta $\mathbf{A}\mathbf{g}_l = \mathbf{g}_l$.

Conseqüentemente, os vetores $\mathbf{g}_l$ são autovetores da matriz $\mathbf{A}$ correspondentes a $n-r-1$ autovalores 1.

Decorre do Teorema da Decomposição Espectral que $\mathbf{A} = \sum_{l=r+2}^{n} \mathbf{g}_l \mathbf{g}_l'$ e

$$n\hat{\mathbf{\Sigma}} = \mathcal{E}'\mathbf{A}\mathcal{E} = \sum_{l=r+2}^{n} (\mathcal{E}'\mathbf{g}_l)(\mathcal{E}'\mathbf{g}_l)' = \sum_{l=r+2}^{n} V_l V_l',$$

onde, devido ao fato de que $Cov(V_{li}, V_{jk}) = \mathbb{E}(\mathbf{g}_l' \mathcal{E}_i \mathcal{E}_k' \mathbf{g}_j) = \sigma_{ik} \mathbf{g}_l' \mathbf{g}_j = 0$, para $l \neq j$, tem-se que $\mathcal{E}'\mathbf{g}_l = V_l$ são independentemente e possuem distribuição $N_m(\mathbf{0}, \mathbf{\Sigma})$. Indicando, por definição, que $n\hat{\mathbf{\Sigma}}$ possui distribuição de Wishart com r e $n-r-1$ graus de liberdade.

Seguindo procedimento semelhante, se

$$\mathbf{A}_1 \mathbf{g}_l = \begin{cases} \mathbf{g}_l, & l > q+1 \\ \mathbf{0}, & l \leq q+1 \end{cases},$$

então $\mathbf{A}_1 = \sum_{l=q+2}^{n} \mathbf{g}_l \mathbf{g}_l'$.

Pode-se escrever a expressão $n(\hat{\mathbf{\Sigma}}_1 - \hat{\mathbf{\Sigma}})$ como

$$\begin{aligned} n(\hat{\mathbf{\Sigma}}_1 - \hat{\mathbf{\Sigma}}) &= \mathcal{E}'(\mathbf{A}_1 - \mathbf{A})\mathcal{E} \\ &= \sum_{l=q+2}^{r+1} (\mathcal{E}'\mathbf{g}_l)(\mathcal{E}'\mathbf{g}_l)' \\ &= \sum_{l=q+2}^{r+1} V_l V_l', \end{aligned}$$

onde os elementos V_l são independentes e com distribuição $N_m(\mathbf{0}, \mathbf{\Sigma})$. O que acarreta, por definição, que $n(\hat{\mathbf{\Sigma}}_1 - \hat{\mathbf{\Sigma}})$ possui distribuição de Wishart com r e $r-q$ graus de liberdade independentemente de $n\hat{\mathbf{\Sigma}}$ (desde que $n(\hat{\mathbf{\Sigma}}_1 - \hat{\mathbf{\Sigma}}$ envolva um conjunto deferente de V_l's independentes).

Para uma amostra suficientemente grande, a estatística

$$-\left[n - r - 1 - \frac{1}{2}(m - r + q + 1)\right] \ln\left(\frac{|\hat{\mathbf{\Sigma}}|}{|\hat{\mathbf{\Sigma}}_1|}\right)$$

tem, aproximadamente, distribuição qui-quadrado com $m(r-q)$ graus de liberdade. Esta afirmação pode ser encontrada em Box (1949). □

Apêndice B

Tabela 4.1: Dados referentes ao Exemplo 2.5.

Indivíduo	X_1	X_2	X_3	Indivíduo	X_1	X_2	X_3
1	27	0	1	43	46	0	1
2	29	0	1	44	58	0	1
3	30	0	1	45	59	0	1
4	32	0	1	46	59	0	1
5	63	0	1	47	63	0	1
6	64	0	1	48	64	0	1
7	69	0	1	49	70	0	1
8	23	1	1	50	32	1	1
9	37	1	1	51	41	1	1
10	41	1	1	52	48	1	1
11	48	1	1	53	55	1	1
12	55	1	1	54	56	1	1
13	57	1	1	55	57	1	1
14	57	1	1	56	58	1	1
15	59	1	1	57	59	1	1
16	60	1	1	58	61	1	1
17	62	1	1	59	62	1	1
18	66	1	1	60	67	1	1
19	68	1	1	61	69	1	1
20	69	1	1	62	70	1	1
21	37	0	0	63	44	0	0
22	50	0	0	64	51	0	0
23	52	0	0	65	53	0	0
24	59	0	0	66	59	0	0
25	62	0	0	67	62	0	0

Tabela 4.2: Dados referentes ao Exemplo 2.5. - Continuação

Indivíduo	X_1	X_2	X_3	Indivíduo	X_1	X_2	X_3
26	63	0	0	68	23	1	0
27	30	1	0	69	30	1	0
30	44	1	0	72	45	1	0
28	31	1	0	70	32	1	0
29	33	1	0	71	37	1	0
31	46	1	0	73	48	1	0
32	49	1	0	74	51	1	0
33	53	1	0	75	54	1	0
34	54	1	0	75	54	1	0
35	55	1	0	76	57	1	0
36	57	1	0	77	58	1	0
37	59	1	0	78	59	1	0
38	61	1	0	79	63	1	0
39	64	1	0	80	65	1	0
40	66	1	0	81	66	1	0
41	66	1	0	82	68	1	0
42	74	1	0	83	68	1	0