Densidades Preditivas no Modelo de Regressão Linear

Magen Danielle Infante Rojas

Dissertação apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do grau de Mestre
em

Estatística

Área de Concentração: **Estatística**

Orientadora: **Profª Drª Silvia Nagib Elian**

Durante a elaboração deste trabalho o autor recebeu u apoio do CNPq

São Paulo - Março de 2000

Densidades Preditivas no Modelo de Regressão Linear

Este exemplar corresponde à redação
final da dissertação devidamente corrigida
e defendida por Magen Danielle Infante Rojas
e aprovada pela comissão julgadora.

São Paulo, 27 de Março de 2000

Banca examinadora:

Profª Drª Silvia Nagib Elian (orientadora) (IME-USP)
Profª Drª Elisete da Conceição Quintaneiro Aubin (IME-USP)
Prof. Dr. Josemar Rodrigues (UFSCar)

Ao Rei dos séculos,
ao imortal, ao invisível,
ao único e sábio Deus,
a quem pertencem a glória e a honra
pelos séculos dos séculos.

Agradecimentos

Aos meus pais **Raúl Infante** e **Haydeé Rojas**, que sempre serão os mais felizes por tudo que eu fizer ou puder conseguir e cujo amor é incondicional.

A meus irmãos, que sempre apostaram em mim.

À minha orientadora, professora **Silvia Nagib Elian**, que além da orientação séria e acertada na elaboração da dissertação, desde o início do programa constituiu-se um apoio integral.

Aos professores e funcionários com os quais tive contato de alguma forma, a gentileza e disposição de ajudar de todos eles foi sempre alentadora; em particular,

Ao professor **Vanderlei da Costa Bueno**, que em um momento importante deu-me o apoio necessário que permitiu-me continuar.

Ao professor **Luiz Augusto Fernandes de Oliveira**, pela ajuda e orientação em algums conceitos de Matemática que foram necessários neste trabalho.

Aos meus amigos todos, que de alguma forma estiveram me dando forças para não desanimar. Em especial à **Yna Rezza**, amiga em todas as circunstâncias.

Ao **CNPq** pelo apoio financeiro.

Resumo

Neste trabalho, dedicamo-nos ao estudo de funções de verossimilhança preditivas e densidades preditivas para um vetor de observações futuras com base num conjunto de dados observados, apresentando várias aplicações em modelos de regressão linear. Sob estes modelos, apresentamos quatro diferentes densidades preditivas, analisando propriedades relativas à consistência e otimalidade. Posteriormente, descrevemos um método para detectar pontos influentes na análise de regressão através do uso de densidades preditivas. Na última parte do trabalho, essas funções são utilizadas na seleção da melhor equação de regressão.

Abstract

In this dissertation we present predictive likelihood functions and predictive densities for an unobserved vector of random variables based on an observed sample, with many applications in the Linear Regression Model. Under this model, four predictive densities are described and some properties, like consistency and optimality, are analyzed. Further, it is shown a method of assessing the influence of specified subsets of the data in the regression analysis using predictive densities. The last part of the work is devoted to the use of predictive densities in the selection of the best linear regression model.

Sumário

Capítulo 1

Introdução

Na Inferência Estatística, modelos probabilísticos são adotados quando desejamos obter conclusões gerais a partir de um conjunto de dados observados.

Estas conclusões são, na maioria dos casos, referentes a uma ou mais quantidades de interesse, que seriam os parâmetros da distribuição de probabilidades especificada por tais modelos, ou ainda, dizem respeito a valores de variáveis aleatórias, cuja distribuição é definida, a menos de parâmetros desconhecidos, que podem ou não ser de interesse do pesquisador.

A primeira situação, correspondente à inferência sobre os parâmetros, consiste num problema de estimação, e a segunda, a da obtenção de informação sobre valores de particulares variáveis aleatórias, seria enquadrada na área de previsão.

Dentre os vários métodos de estimação existentes, o método da máxima verossimilhança tem sido intensivamente estudado e sua forma intuitiva, aliada a suas propriedades, o tornam bastante atraente.

Muito menos no entanto tem sido estudado sobre o uso de funções de verossimilhança com o objetivo de previsão.

Este trabalho dedica-se ao estudo de Funções de Verossimilhança Preditivas e Densidades Preditivas e sua utilização no Modelo de Regressão Linear Múltipla.

No início do Capítulo 2, fazemos uma revisão dos métodos de previsão comumente utilizados, apresentando as abordagens frequentista e bayesiana

para a previsão. Posteriormente, introduzimos o conceito de Funções de Verossimilhança Preditivas para abordar o problema da previsão de variáveis aleatórias não observadas. A maior parte do Capítulo 2 dedica-se ao estudo deste conceito, pouco explorado, porém muito importante. Deste modo, apresentamos os vários tipos de Funções de Verossimilhança Preditivas e analisamos as vantagens do uso de tais funções frente às outras abordagens.

Uma vez definidos os vários tipos de densidades preditivas presentes na literatura, incluindo as obtidas a partir de Funções de Verossimilhança Preditivas, nos dedicamos ao estudo dessas densidades no modelo de regressão linear em que a variável resposta tem distribuição normal, e tal estudo se constituiu na maior parte deste trabalho. Para este fim, o Capítulo 3 define os modelos lineares para as variáveis aleatórias observadas e não observadas ou futuras, apresentando quatro densidades preditivas para os dados não observados. Analisaremos ainda propriedades relativas à otimalidade e à consistência dessas funções quando utilizadas para estimar a densidade de probabilidades das observações futuras.

Para a determinação de pontos influentes no modelo de regressão através de densidades preditivas, tem-se desenvolvido na literatura a medida de divergência de Kullback-Leibler. Esta medida dá origem às chamadas *Funções de Influência Preditivas*, que serão estudadas no Capítulo 4. Demonstraremos a obtenção dessa medida utilizando uma particular densidade preditiva bayesiana e ilustramos com um exemplo numérico o uso dessas funções na caracterização de pontos ou conjunto de pontos influentes na previsão de um vetor de observações futuras.

O Capítulo 5 mostrará a utilização das densidades preditivas para selecionar variáveis explicativas no modelo de regressão linear. Esta seleção do modelo será feita através de uma *medida de verossimilhança* obtida do produto de todas as densidades preditivas correspondentes a cada uma das observações amostrais. Apresentamos neste capítulo a técnica conhecida como técnica de *reutilização da amostra*, que consiste em obter a densidade preditiva de uma observação dadas as demais, sendo que a densidade preditiva utilizada será também a bayesiana. Finalizamos esse capítulo apresentando um exemplo numérico de seleção de variáveis explicativas num modelo de regressão linear através dessa técnica.

Capítulo 2

Funções de Verossimilhança Preditivas

2.1 Introdução

O objetivo deste capítulo é apresentar uma ferramenta matemática unificada, baseada em funções de verossimilhança, para abordar o problema da previsão, seja qual for o enfoque, bayesiano ou frequentista. Estudaremos assim as contribuições desenvolvidas na literatura relativa à previsão, dando maior destaque a uma abordagem não bayesiana que tem algumas vantagens, como por exemplo, dispensar a especificação de uma densidade a priori para os parâmetros desconhecidos. É essencialmente o problema da previsão que trataremos neste capítulo, fornecendo o conceito de *verossimilhança preditiva* como uma base para a previsão de quantidades aleatórias.

Para isso, na próxima seção faremos uma revisão das duas abordagens preditivas presentes na literatura, a clássica e a bayesiana, destacando as dificuldades ao aplicá-las. Posteriormente, apresentaremos a definição de verossimilhança preditiva, a partir da qual iremos obter as chamadas *Funções de Verossimilhança Preditivas.*

Vamos nos concentrar na obtenção de três versões diferentes de *Funções de Verossimilhança Preditivas*, que surgem basicamente após a eliminação dos parâmetros *nuisance* através de maximização, integração ou condicionamento a estatísticas suficientes. Apresentaremos ainda alguns exemplos ilustrativos e finalizamos o capítulo destacando as vantagens do uso destas funções como a obtenção de densidades preditivas, previsores e regiões de previsão para a quantidade de interesse.

2.2 Previsão

O problema da previsão considera basicamente uma variável aleatória $\mathbf{X}$ observada e uma variável aleatória $\mathbf{Y}$ não observável ou futura ($\mathbf{X}$ e $\mathbf{Y}$ uni ou multidimensionais); sendo que suas distribuições de probabilidades dependem de um mesmo parâmetro θ desconhecido, cujo valor pode não ser de interesse.

Vamos supor que desejamos prever o valor de uma variável aleatória $\mathbf{Z} = h(\mathbf{Y})$, com distribuição de probabilidades que também depende de θ, sobre o qual o conjunto de dados prévio $\mathbf{X}$ forneceu alguma informação. Na maior parte dos casos, $\mathbf{X}$ será um vetor aletório, correspondendo a uma amostra de tamanho n. Observamos que o parâmetro θ, comum a $\mathbf{X}$ e $\mathbf{Z}$, será o elo de ligação entre essas duas quantidades.

Ao prever $\mathbf{Z}$, temos dois tipos de incerteza: incerteza devido à variação aleatória de $\mathbf{Y}$ e incerteza devido à falta de conhecimento sobre θ . Desejamos então fazer afirmações sobre $\mathbf{Z}$, e a este procedimento chamaremos de previsão.

Por exemplo, quando adotamos um modelo de regressão linear simples,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \cdots, n$$

com base numa amostra de n observações $(x_1, y_1), \cdots (x_n, y_n)$, geralmente desejamos prever o valor de uma observação futura y_{n+1} (dado x_{n+1}). O previsor de y_{n+1} é dado por

$$\hat{y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$$

onde $\hat{\beta}_0$ e $\hat{\beta}_1$ são estimadores de β_0 e β_1, obtidos a partir da amostra.

Quanto aos parâmetros β_0 e β_1, estes podem não ser de interesse, mas sua estimação foi necessária para a previsão de y_{n+1}.

Voltando ao problema original da previsão de $\mathbf{Z} = h(\mathbf{Y})$ com base nos dados observados $\mathbf{X}$, definimos então o previsor de $\mathbf{Z}$ como

$$\hat{T} = \hat{T}(\mathbf{X})$$

onde $\hat{T}$ é qualquer função de $\mathbf{X}$, não envolvendo quantidades desconhecidas.

O previsor $\hat{T}$ é dito não viciado para $\mathbf{Z}$ se

$$E(\hat{T} - \mathbf{Z}) = 0.$$

Finalizando, o Erro Quadrático Médio do previsor $\hat{T}$ é definido como

$$EQM(\hat{T}) = E(\hat{T} - \mathbf{Z})^2.$$

Existem dois enfoques para resolver o problema da previsão, o *enfoque clássico* (ou *frequentista*) e o *enfoque bayesiano.*

Enfoque Clássico (ou Frequentista)

Não existem regras muito claras de como fazer previsão de observações futuras na *inferência clássica.* Um dos procedimentos mais adotados é o de substituir na distribuição de probabilidades das observações futuras o valor do parâmetro θ desconhecido por alguma estimativa baseada em dados passados. Isto é, se $\mathbf{X}$ e $\mathbf{Z}$ tem funções densidade $f(\mathbf{x};\theta)$ e $f(\mathbf{z};\theta)$ respectivamente, onde $\mathbf{Z}$ é a variável aleatória a ser prevista com base nas observações passadas $\mathbf{X}$, então $\mathbf{Z}$ será prevista utilizando-se a distribuição $f(\mathbf{z};\hat{\theta})$, onde $\hat{\theta}$ é um estimador de θ baseado em $\mathbf{X}$. A falha desse procedimento é não levar em consideração a variabilidade associada à estimação de θ, pois a previsão é feita como se $\hat{\theta}$ fosse o verdadeiro valor deste parâmetro. Por isso, de acordo com GAMERMAN e MIGON (1997), [pg. 169], esse procedimento subestima a variabilidade do previsor obtido.

Uma abordagem que evita esses problemas consiste em obter uma *quantidade pivotal*, que dependeria da variável aleatória a ser prevista, mas cuja distribuição não depende de parâmetros desconhecidos.

Definição 2.2.1 *(GAMERMAN e MIGON (1997) [pg. 169])*
Uma variável aleatória $G = G(\mathbf{X}, \mathbf{Z})$ *é uma* **Quantidade Pivotal** *ou* **Pivot**, *se a distribuição de probabilidades de* G *não depende do parâmetro* θ, *isto é, se* $G(\mathbf{X}, \mathbf{Z})$ *tem a mesma distribuição de probabilidades para todo* θ.

A função G não depende de θ, e de acordo com GAMERMAN e MIGON (1997), deve depender da amostra $\mathbf{X}$ de uma forma ótima, por exemplo,

através de estatísticas suficientes minimais para θ.

Uma vez obtido G, conforme veremos no exemplo, podemos fazer afirmações probabilísticas sobre esta função, e como consequência, sobre a quantidade a ser prevista. O único problema desse procedimento é que nem sempre conseguimos encontrar a quantidade pivotal G.

Exemplo 2.1 Consideremos $\mathbf{X} = (X_1, \cdots, X_n)$ vetor aleatório observado e $\mathbf{Y} = (Y_1, \cdots, Y_m)$ vetor aleatório não observado ou de dados futuros, tais que $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, 2, \cdots n$ e $Y_j \sim \mathcal{N}(\mu, \sigma^2)$, $j = 1, 2, \cdots m$ são mutuamente independentes $\forall i, j$, com σ^2 conhecido e μ desconhecido. Nestas condições,

$$\bar{\mathbf{X}} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \qquad e \qquad \bar{\mathbf{Y}} \sim \mathcal{N}(\mu, \frac{\sigma^2}{m})$$

e são independentes e, tomando $\mathbf{Z} = \bar{\mathbf{Y}}$, a função

$$G(\mathbf{X}, \mathbf{Z}) = \frac{\bar{\mathbf{X}} - \mathbf{Z}}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)}} \quad \sim \quad \mathcal{N}(0,1)$$

é a chamada *quantidade pivotal*, porque depende da quantidade a ser prevista $\mathbf{Z}$, mas sua distribuição de probabilidades independe do parâmetro desconhecido μ .

Assim, fixado α, podemos obter os quantis $-z_\alpha$ e z_α da distribuição normal padrão, de modo que

$$P(-z_\alpha \leq G \leq z_\alpha) = P\left(-z_\alpha \leq \frac{\bar{\mathbf{X}} - \mathbf{Z}}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)}} \leq z_\alpha \right) \geq \alpha.$$

Isolando $\mathbf{Z}$, temos que

$$P\left(\bar{\mathbf{X}} - z_\alpha \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)} \leq \mathbf{Z} \leq \bar{\mathbf{X}} + z_\alpha \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)} \right) \geq \alpha,$$

e então, um intervalo de previsão para $\mathbf{Z}$ com coeficiente de confiança α fica

$$\left[\bar{\mathbf{X}} - z_\alpha\sqrt{\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)}\,;\,\bar{\mathbf{X}} + z_\alpha\sqrt{\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)}\right]. \qquad \square$$

Finalizando, conforme comentado no início desta seção, nos modelos de regressão linear múltipla da forma

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i, \qquad i = 1, \cdots, n\,,$$

com as suposições de que $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ e que ϵ_i, ϵ_j são variáveis aleatórias independentes, para $i \neq j$, o previsor de y dado $x_1, x_2, \cdots, x_k$ é dado por

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki},$$

onde $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\cdots$ $\hat{\beta}_k$ são usualmente os estimadores de mínimos quadrados de $\beta_0, \beta_1, \beta_2 \cdots \beta_k$. Neste caso, o previsor $\hat{y}$ nada mais é que o estimador de $E(y/\, x_1, x_2, \cdots, x_k)$, que, como consequência da hipótese de que $E(\epsilon_i) = 0$, é da forma $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$.

Enfoque Bayesiano

Através da abordagem bayesiana, o problema da previsão de observações futuras é facilmente resolvido, utilizando-se $f(\mathbf{z}/\mathbf{x})$, a densidade a posteriori da quantidade a ser prevista, $\mathbf{Z}$, dada a amostra observada, $\mathbf{x}$.

Tal densidade, denominada densidade preditiva a posteriori de $\mathbf{Z}$ dado $\mathbf{x}$ é definida como

$$f(\mathbf{z}/\mathbf{x}) = \int_{\theta\in\Theta} f(\mathbf{z}/\mathbf{x}, \theta) f(\theta/\mathbf{x}) d\theta \tag{2.1}$$

onde

$f(\mathbf{z}/\mathbf{x}, \theta)$ é a densidade condicional de $\mathbf{Z}$ dado $\mathbf{X} = \mathbf{x}$ e
$f(\theta/\mathbf{x})$ é a densidade a posteriori de θ dado $\mathbf{X} = \mathbf{x}$.

Observamos que $f(\mathbf{z}/\mathbf{x})$ corresponde a $E_\theta(f(\mathbf{z}/\mathbf{x}, \theta))$, ou seja, é a esperança em θ de $f(\mathbf{z}/\mathbf{x}, \theta)$, com relação à densidade a posteriori de θ dado $\mathbf{X} = \mathbf{x}$.

Verifica-se facilmente que, se $f(\theta)$ é a densidade a priori de θ, então

$$f(\mathbf{z}/\mathbf{x}) = \frac{\int_{\theta\in\Theta} f(\mathbf{x},\mathbf{z}/\theta)f(\theta)d\theta}{\int_{\theta\in\Theta} f(\mathbf{x}/\theta)f(\theta)d\theta}, \tag{2.2}$$

sendo que

$f(\mathbf{x}/\theta)$ é a densidade dos dados $\mathbf{X}$ e
$f(\mathbf{x},\mathbf{z}/\theta)$ é a densidade conjunta de $(\mathbf{X}, \mathbf{Z})$.

Com base na densidade preditiva $f(\mathbf{z}/\mathbf{x})$, uma região de confiança preditiva para $\mathbf{Z}$, com coeficiente de confiança $100\,\gamma$ % é dada por

$$R_\gamma(\mathbf{x}) = \{\, \mathbf{z} \mid f(\mathbf{z}/\mathbf{x}) \geq k \,\},$$

onde k é tal que

$$\int_{R_\gamma(\mathbf{x})} f(\mathbf{z}/\mathbf{x})\, d\mathbf{z} = \gamma.$$

Este procedimento apresenta uma única dificuldade que é a dependência de uma densidade a priori para o parâmetro θ.

2.3 Verossimilhança na previsão

Bayesianos e frequentistas baseiam-se no conceito fundamental da *verossimilhança* como uma ferramenta comum para a *inferência paramétrica*, embora, no problema da *previsão*, as abordagems bayesiana e frequentista sejam bastante diferentes.

Assim, a motivação principal para o estudo das *Funções de Verossimilhança Preditivas* é a quase inexistência de uma teoria de previsão não bayesiana e, por outro lado, o fato de que a teoria bayesiana exige a especificação de uma priori.

A teoria de previsão bayesiana já existente constituia-se somente numa parte da teoria de previsão. Especificamente, a previsão bayesiana se assemelha ao uso das *Funções de Verossimilhança Preditivas Marginais* que, conforme veremos posteriormente, são obtidas mediante a integração de uma

função de verossimilhança particular de $(\mathbf{z}, \theta)$ no parâmetro *nuisance* θ.

Segundo a literatura, tudo indica que a noção de verossimilhança preditiva foi sugerida por FISHER (1956), num estudo envolvendo a distribuição binomial. No entanto, formalmente não existiam *bases unificadas* para a previsão de eventos futuros dadas as observações. LAURITZEN (1974) publicou o primeiro artigo a respeito, no qual fez tentativas para desenvolver uma teoria *não bayesiana* que viabilizasse a previsão através de *Funções de Verossimilhança Preditivas.* Desta forma, o autor introduziu o conceito de *previsor de máxima verossimilhança*, considerando variáveis aleatórias discretas em Processos Estocásticos.

Conforme definimos anteriormente, o problema da previsão consiste em obter informação sobre $\mathbf{Z} = h(\mathbf{Y})$, uma função qualquer de $\mathbf{Y}$, uma vez observado o vetor de dados $\mathbf{X}$. LAURITZEN (1974) dedicou-se à previsão de $h(\mathbf{Y}) = \mathbf{Y}$, e, trabalhando com variáveis aleatórias discretas sugeriu como previsor para $\mathbf{Z} = h(\mathbf{Y})$, o valor $\hat{\mathbf{Z}}$ que maximiza

$$L_1(\mathbf{z}/\mathbf{x}) \;=\; f(\mathbf{x}/r(\mathbf{x},\mathbf{z})),$$

onde

$r(\mathbf{X},\mathbf{Z})$ é a Estatística Suficiente Minimal com base em $(\mathbf{X},\mathbf{Z})$ e

$f(\mathbf{x}/r(\mathbf{x},\mathbf{z}))$ é a densidade condicional de $\mathbf{X}$ dado que $r(\mathbf{X},\mathbf{Z}) = r(\mathbf{x},\mathbf{z})$.

Portanto, $\hat{\mathbf{Z}}$ é o previsor de $\mathbf{Z}$ com base em L_1, que foi utilizada para avaliar a verossimilhança dos valores de $\mathbf{Z}$, na presença dos dados observados.

O termo *verossimilhança preditiva* foi introduzido por HINKLEY (1979), expressando assim a necessidade e desejo de uma abordagem preditivista baseada na verossimilhança. O autor trabalhou com variáveis aleatórias discretas e contínuas, e para isso estendeu e modificou levemente L_1, de modo a cobrir ambos os casos, mas exigindo algumas condições, que formalmente descreveremos posteriormente. O autor também sugere o uso de verossimilhanças preditivas baseadas em *estatísticas ancilares.* Esta publicação é o primeiro grande trabalho a se destacar na construção de bases para a previsão via verossimilhança, e tem servido de inspiração para muitas das pesquisas na área.

MATHIASEN (1979) estudou várias funções de previsão, introduzindo inclusive a verossimilhança preditiva do tipo *profile*, que será definida na próxima seção.

Uma distribuição do tipo profile também foi proposta por LEJEUNE e FAULKENBERRY (1982) como alternativa à distribuição preditiva bayesiana. Os autores denominaram-na densidade preditiva de máxima verossimilhança e, sob determinadas condições, provaram a coincidência entre esta densidade preditiva e a densidade preditiva bayesiana com uma priori não informativa.

LEVY e PERNG (1984) utilizam a *Função de Verossimilhança Preditiva Profile* na previsão de dados não observados em modelos de regressão linear. Seu trabalho, que é de especial interesse nessa dissertação, será discutido com detalhes no Capítulo 3.

BUTLER (1986) define a *Função de Verossimilhaça Preditiva Condicional* de uma maneira similar às apresentadas por LAURITZEN (1974) e HINKLEY(1979). Neste trabalho, *Funções de Verossimilhança Preditivas* são usadas não somente para prever dados futuros, mas também na imputação de dados perdidos. Tal forma de imputação foi posteriormente comparada com a obtida através do algoritmo EM. São ainda apresentadas algumas aplicações desta metodologia em modelos de Análise de Variância. Numa posterior publicação de 1989, o autor apresenta uma *Função de Verossimilhança Preditiva Marginal* baseada em estatísticas ancilares.

HARRIS (1989) considera uma distribuição preditiva *bootstrap*, que é obtida integrando-se a densidade de $\mathbf{Z}$ para $\theta = \hat{\theta}$ com respeito à distribuição amostral do estimador $\hat{\theta}(\mathbf{X})$ e é dada por $\int f(\mathbf{z}; \hat{\theta}) dF_{\hat{\theta}(\mathbf{X})}(\hat{\theta})$.

Vários outros autores têm se dedicado ao estudo de *Funções de Verossimilhança Preditivas*. BJORNSTAD (1990) apresenta um excelente histórico sobre o assunto. O autor considera a *Função de Verossimilhança Preditiva Profile* e suas modificações, estuda as várias *Funções de Verossimilhança Preditivas Condicionais* e compara algumas das mais importantes *Funções de Verossimilhança Preditivas* presentes na literatura. Sugere unificações e simplificações das funções de verossimilhança baseadas na suficiência e, finalizando, destaca propriedades desejáveis para *Funções de Verossimilhança Preditivas* de modo geral.

Na próxima seção, definiremos formalmente o conceito de *Função de Ve-*

rossimilhança Preditiva, apresentaremos as principais funções e alguns exemplos.

2.4 Funções de Verossimilhança Preditivas

Da descrição feita na seção anterior, notamos que têm sido propostas diferentes versões de *Funções de Verossimilhança Preditivas*. Todas elas, no entanto, surgem da aplicação de alguma técnica de eliminação de parâmetros *nuisance* na verossimilhança conjunta. Esta verossimilhança conjunta será definida a seguir.

Definição 2.4.1 *(BERGER e WOLPER, (1988), [pg. 39])*
Sejam $\mathbf{X} = (X_1, \cdots, X_n)$ *e* $\mathbf{Y} = (Y_1, \cdots, Y_m)$ *vetores aleatórios observado e não observado respectivamente e* $\mathbf{Z} = h(\mathbf{Y})$ *a variável aleatória que desejamos prever.*
Se θ *é o parâmetro desconhecido,* Θ *é o espaço paramétrico e* $f(\mathbf{x}, \mathbf{z}; \theta)$ *é a densidade conjunta de* $(\mathbf{X}, \mathbf{Z})$, *a função*

$$l_{\mathbf{x}}(\mathbf{z}, \theta) = f(\mathbf{x}, \mathbf{z}; \theta)$$

é definida como a **Função de Verossimilhança de** $(\mathbf{Z}, \theta)$ *e coincide com a densidade* $f(\mathbf{x}, \mathbf{z}; \theta)$, *vista como função de* $(\mathbf{z}, \theta)$, *dado* $\mathbf{X} = \mathbf{x}$.

Conforme discutido anteriormente, nosso objetivo é prever $\mathbf{Z}$ com base nos dados observados $\mathbf{X} = (X_1, \cdots, X_n)$ e quando não há um interesse específico na estimação de θ, este é considerado um *parâmetro nuisance*. Assim, apresentamos a seguir uma definição geral de função de verossimilhaça para a previsão de $\mathbf{Z}$.

Definição 2.4.2 *(BJORNSTAD (1990))*
Nas condições da Definição 2.4.1, se $\mathbf{Z}$ *é a quantidade que deseja-se prever, então a* **Função de Verossimilhança Preditiva** *de* $\mathbf{Z}$, $L(\mathbf{z}/\mathbf{x})$, *é obtida após a eliminação do parâmetro nuisance* θ *da função de verossimilhança* $l_{\mathbf{x}}(\mathbf{z}, \theta)$.

Para a eliminação do parâmetro *nuisance*, existem vários métodos, e, dependendo do método usado, obtem-se diferentes *Funções de Verossimilhança Preditivas.*

BJORNSTAD (1990) apresenta vários tipos de *Funções de Verossimilhanças Preditivas* mas salienta que todas são provenientes basicamente da aplicação de uma das três operações sobre $l_{\mathbf{x}}(\mathbf{z}, \theta)$: Maximização, Integração com respeito a θ e Condicionamento a estatísticas suficientes.

A eliminação do parâmetro *nuisance* da função $l_{\mathbf{x}}(\mathbf{z}, \theta)$ através de maximização dá origem à *Função de Verossimilhança Preditiva Profile*, que denotaremos por $L_p(\mathbf{z}/\mathbf{x})$, e descreveremos a seguir.

Funções de Verossimilhança Preditivas Profile

Inicialmente, esta função preditiva foi sugerida por MATHIASEN (1979) para o caso especial em que o vetor aleatório observado $\mathbf{X}$ e o vetor de dados futuros $\mathbf{Y}$ são independentes.

Para o caso em que $\mathbf{X}$ e $\mathbf{Y}$ não são necessariamente independentes, LEVY e PERNG (1984) definem-na como uma *Função de Previsão Modificada de Fisher*, cujo conceito será discutido com detalhes no Capítulo 3. A partir daí, a *Função de Verossimilhança Preditiva Profile* está perfeitamente definida para ambos os casos.

A definição que apresentaremos a seguir foi a utilizada por todos os autores acima mencionados.

Definição 2.4.3 *Nas condições da Definição 2.4.1, a função*

$$L_p(\mathbf{z}/\mathbf{x}) = \sup_{\theta \in \Theta} l_{\mathbf{x}}(\mathbf{z}, \theta) \ = \sup_{\theta \in \Theta} f(\mathbf{x}, \mathbf{z}; \theta)$$

é definida como a **Função de Verossimilhança Preditiva Profile** *para* $\mathbf{Z}$.

É importante notarmos que, na Definição 2.4.3, como $\mathbf{X}$ é observado e $\mathbf{Z}$ não, $\mathbf{Z}$ será visto como variável aleatória e $\mathbf{X}$ como parâmetro, e $L_p(\mathbf{z}/\mathbf{x})$ pode ser encarada como uma função proporcional a uma distribuição condicional de $\mathbf{Z}$ dado $\mathbf{X}$.

Além disso, qualquer *Função de Verossimilhança Preditiva* pode ser padronizada de modo a se tornar uma densidade, conforme mostra a próxima definição.

Definição 2.4.4 *A* **Densidade Preditiva Profile** *é definida como*

$$\begin{aligned} f_p(\mathbf{z}/\mathbf{x}) &= k(\mathbf{x}) \sup_{\theta \in \Theta} f(\mathbf{x}, \mathbf{z}; \theta) \\ &= k(\mathbf{x})\, L_p(\mathbf{z}/\mathbf{x}), \end{aligned}$$

onde $k(\mathbf{x})$ *é uma constante de padronização que não envolve* $\mathbf{z}$ *e tal que*

$$\int_{-\infty}^{\infty} k(\mathbf{x}) L_p(\mathbf{z}/\mathbf{x}) d\mathbf{z} = 1.$$

Desta definição, é facil ver que duas *Funções de Verossimilhança Preditivas Profile* proporcionais forneceriam a mesma densidade preditiva profile. A padronização traz uma série de vantagens, que serão descritas posteriormente.

De acordo com LEJEUNE e FAULKENBERRY (1982), a *Função de Verossimilhança Preditiva Profile* L_p surge de uma idéia bastante intuitiva. Inicialmente, KALBFLEISH e SPROTT (1970), ao considerar a marginalização em problemas de verossimilhanças multiparamétricas em que θ_1 é o parâmetro de interesse, eliminaram o parâmetro *nuisance* θ_2 na função de verossimilhança $L(\theta_1, \theta_2/\mathbf{x})$, substituindo-o por seu estimador de máxima verossimilhança. Assim, logo após a padronização, os autores obtiveram o que chamaram *verossimilhança relativa máxima* de θ_1.

Tecnicamente, a *Função de Verossimilhança Preditiva Profile* é obtida da mesma forma, mas, considerando a variável aleatória $\mathbf{Z}$ ao invés do parâmetro θ_1. Desta forma, para o nosso estudo, $\mathbf{Z}$ corresponderia a θ_1, quantidade de interesse, e θ , similar a θ_2, é o parâmetro *nuisance*. Este método corresponde à *verossimilhança profile* na *inferência paramétrica* (KALBFLEISH e SPROTT (1970)) e por essa razão, $L_p(\mathbf{z}/\mathbf{x})$ é chamada *Função de Verossimilhaça Preditiva Profile.*

Os próximos exemplos ilustram o cálculo de $L_p(\mathbf{z}/\mathbf{x})$.

Exemplo 2.2 (LEJEUNE e FAULKENBERRY (1982))
Sejam $\mathbf{X} = (X_1, \cdots, X_n)$ o conjunto de variáveis aleatórias observáveis,

$\mathbf{Y} = (Y_1, \cdots, Y_m)$ o conjunto de variáveis aleatórias não observáveis ou futuras sendo que todas as variáveis aleatórias são independentes e têm distribuição exponencial com parâmetro θ desconhecido.

Vamos admitir que $\mathbf{Z} = \sum_{i=1}^{m} Y_i$ é a quantidade a ser prevista. Esta variável aleatória tem distribuição *gama* com parâmetros m e θ e assim, se $f(\mathbf{x};\theta)$ e $g(\mathbf{z};\theta)$ são as densidades de $\mathbf{X}$ e $\mathbf{Z}$ respectivamente, então

$$f(\mathbf{x};\theta) = \theta^n e^{-\theta \sum_{i=1}^{n} x_i}$$

e

$$g(\mathbf{z};\theta) = \frac{\theta^m}{\Gamma(m)} z^{m-1} e^{-\theta z}.$$

Logo,

$$\begin{aligned} l_{\mathbf{x}}(\mathbf{z},\theta) &= f(\mathbf{x};\theta)\, g(\mathbf{z};\theta) \\ &= \frac{\theta^{n+m}}{\Gamma(m)} z^{m-1} e^{-\theta(z+\sum_{i=1}^{n} x_i)} \\ &= \frac{\theta^{n+m}}{(m-1)!} z^{m-1} e^{-\theta(z+\sum_{i=1}^{n} x_i)}. \end{aligned} \tag{2.3}$$

Tomando logaritmo, obtemos

$$\ln\left(l_{\mathbf{x}}(\mathbf{z},\theta)\right) = (n+m)\ln\theta - \ln(m-1)! + (m-1)\ln z - \theta(z + \sum_{i=1}^{n} x_i)$$

e derivando com respeito a θ e igualando a zero segue que

$$\frac{d}{d\theta}\ln\left(l_{\mathbf{x}}(\mathbf{z},\theta)\right) = \frac{n+m}{\theta} - (z + \sum_{i=1}^{n} x_i) = 0$$

e

$$\hat{\theta} = \left(\frac{n+m}{\mathbf{z} + \sum_{i=1}^{n} x_i}\right)$$

é o valor de θ que maximiza $l_{\mathbf{x}}(\mathbf{z},\theta)$.

Substituindo em (2.3), obtemos a *Função de Verossimilhança Preditiva Profile*

$$L_p(\mathbf{z}/\mathbf{x}) = \sup_{\theta\in\Theta}\left(l_{\mathbf{x}}(\mathbf{z},\theta)\right)$$

$$
\begin{aligned}
&= \frac{(n+m)^{n+m}}{(z+\sum_{i=1}^{n} x_i)^{n+m}} \frac{z^{m-1}}{(m-1)!} \exp\{-(n+m)\} \\
&= \frac{(n+m)^{n+m}}{(m-1)!} \exp\{-(n+m)\} \frac{z^{m-1}}{(z+\sum_{i=1}^{n} x_i)^{n+m}} \\
&= c \frac{z^{m-1}}{(z+\sum_{i=1}^{n} x_i)^{n+m}} \\
&= c \frac{1}{(\sum_{i=1}^{n} x_i)^{n+m}} \frac{z^{m-1}}{\left(\frac{z}{\sum_{i=1}^{n} x_i}+1\right)^{n+m}} \\
&= c'(\mathbf{x}) \frac{1}{(\sum_{i=1}^{n} x_i)^{m-1}} \frac{z^{m-1}}{\left(\frac{z}{\sum_{i=1}^{n} x_i}+1\right)^{n+m}} \\
&= c'(\mathbf{x}) \left(\frac{z}{\sum_{i=1}^{n} x_i}\right)^{m-1} \left(1+\frac{z}{\sum_{i=1}^{n} x_i}\right)^{-n-m},
\end{aligned}
$$

onde $c'(\mathbf{x})$ é uma função que independe de $\mathbf{z}$.

Exemplo 2.3 Sejam $\mathbf{X}$ e $\mathbf{Y}$ como definidos no exemplo anterior mas tais que $X_i,\ Y_j \sim \mathcal{N}(\theta, \sigma^2)$ com σ^2 conhecido, para $i = 1, 2, \cdots, n$ e $j = 1, 2, \cdots, m$. Se a quantidade que desejamos prever é $\mathbf{Z} = \sum_{i=1}^{m} Y_i / m$, temos que $\mathbf{Z} \sim \mathcal{N}(\theta, \sigma^2 / m)$,

$$
f(\mathbf{x};\theta) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\{-(2\sigma^2)^{-1} \sum_{i=1}^{n} (x_i-\theta)^2\}
$$

e

$$
g(z;\theta) = m^{\frac{1}{2}} (2\pi)^{-\frac{1}{2}} \sigma^{-1} \exp\{-m(2\sigma^2)^{-1} (z-\theta)^2\}.
$$

Desta forma,

$$
\begin{aligned}
l_{\mathbf{x}}(z,\theta) &= f(\mathbf{x};\theta)\, g(z;\theta) \\
&= (2\pi\sigma^2)^{-\frac{(n+1)}{2}} m^{\frac{1}{2}} \exp\left\{-(2\sigma^2)^{-1} [m(z-\theta)^2 + \sum_{i=1}^{n} (x_i-\theta)^2]\right\} \\
&= c_1 \exp\left\{-(2\sigma^2)^{-1} [mz^2 - 2mz\theta + m\theta^2 + \sum_{i=1}^{n} x_i^2 - 2\theta \sum_{i=1}^{n} x_i + n\theta^2]\right\} \\
&= c_1 \exp\left\{-(2\sigma^2)^{-1} (\sum_{i=1}^{n} x_i^2 + mz^2) + \theta(\sigma^2)^{-1} (\sum_{i=1}^{n} x_i + mz) - (2\sigma^2)^{-1}(n+m)\theta^2\right\}.
\end{aligned}
$$

De forma similar ao exemplo anterior,

$$
\hat{\theta} = \left(\frac{\sum_{i=1}^{n} x_i + m\mathbf{z}}{n+m}\right)
$$

é o valor de θ que maximiza $l_{\mathbf{x}}(z,\theta)$.

Substituindo $\hat{\theta}$ em $l_{\mathbf{x}}(z,\theta)$, obtemos a *Função de Verossimilhança Preditiva Profile* de $\mathbf{Z}$,

$$
\begin{aligned}
L_p(\mathbf{z}/\mathbf{x}) &= l_{\mathbf{x}}(z,\hat{\theta}) \\
&= c_1\, exp\left\{-\frac{(\sum_{i=1}^n x_i^2 + mz^2)}{2\sigma^2} + \frac{(\sum_{i=1}^n x_i + mz)^2}{\sigma^2(n+m)} - \frac{(\sum_{i=1}^n x_i + mz)^2}{2\sigma^2(n+m)}\right\} \\
&= c_1\, exp\left\{-\frac{(\sum_{i=1}^n x_i^2 + mz^2)}{2\sigma^2} + \frac{(\sum_{i=1}^n x_i + mz)^2}{2\sigma^2\,(n+m)}\right\} \\
&= c_1\, exp\left\{-\frac{1}{2\sigma^2(n+m)}\left[(n+m)(\sum_{i=1}^n x_i^2 + mz^2) - (\sum_{i=1}^n x_i + mz)^2\right]\right\} \\
&= c_1\, exp\left\{-\frac{1}{2\sigma^2(n+m)}\left[(n+m)\sum_{i=1}^n x_i^2 + nmz^2 - (\sum_{i=1}^n x_i)^2 - 2mz\sum_{i=1}^n x_i\right]\right\} \\
&= c_1\, exp\left\{\frac{(\sum_{i=1}^n x_i)^2 - (n+m)\sum_{i=1}^n x_i^2}{2\sigma^2(n+m)}\right\}\, exp\left\{-\frac{(nmz^2 - 2mz\sum_{i=1}^n x_i)}{2\sigma^2(n+m)}\right\} \\
&= c_1\, c_2(\mathbf{x})\, exp\left\{-\frac{(nmz^2 - 2mz\sum_{i=1}^n x_i)}{2\sigma^2(n+m)}\right\} \\
&= c_1\, c_2(\mathbf{x})\, exp\left\{-\frac{(nmz^2 - 2mnz\bar{x})}{2\sigma^2(n+m)}\right\} \\
&= c_1\, c_2(\mathbf{x})\, exp\left\{-\frac{nm}{2\sigma^2(n+m)}(z^2 - 2\bar{x}z + \bar{x}^2 - \bar{x}^2)\right\} \\
&= c_1\, c_2(\mathbf{x})\, exp\left\{\frac{nm}{2\sigma^2(n+m)}(\bar{x}^2)\right\}\, exp\left\{-\frac{1}{2\sigma^2}\left(\frac{nm}{n+m}\right)(z-\bar{x})^2\right\} \\
&= c_1\, c_2(\mathbf{x})\, c_3(\mathbf{x})\, exp\left\{-\frac{1}{2\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)}(z-\bar{x})^2\right\} \\
&= c(\mathbf{x})\, exp\left\{-\frac{1}{2\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)}(z-\bar{x})^2\right\}.
\end{aligned}
$$

Após a padronização, obtemos a *Densidade Preditiva Profile* de $\mathbf{Z}$ dada por

$$
f(z\,/\,\mathbf{x}) = \frac{1}{\sqrt{2\pi}\;\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}\;\exp\left\{-\frac{1}{2\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)}(z-\bar{x})^2\right\},
$$

também definida por LEJEUNE e FAULKENBERRY (1982) e que coincide com a densidade de uma variável aleatória com distribuição normal com

média $\bar{x}$ e variância $\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)$. $\square$

A seguir, trataremos de *Funções de Verossimilhança Preditivas* baseadas no condicionamento em *estatísticas suficientes minimais.*

Funções de Verossimilhança Preditivas Condicionais

Segundo alguns autores, a abordagem mais simples e direta para a previsão seria substituir θ por um estimador $\hat{\theta}=\hat{\theta}(\mathbf{X})$ na função de verossimilhaça $l_{\mathbf{x}}(\mathbf{z},\theta)$ e após a padronização, utilizar

$$g_{\hat{\theta}}(\mathbf{z}/\mathbf{x}) = k(\mathbf{x})\, l_{\mathbf{x}}(\mathbf{z},\hat{\theta})$$

como a *Função de Verossimilhança Preditiva.* Neste caso, $g_{\hat{\theta}}(\mathbf{z}/\mathbf{x})$ é considerada como uma densidade condicional de $\mathbf{Z}$ dado $\mathbf{X}=\mathbf{x}$ para $\theta=\hat{\theta}$.

Este tipo de procedimento encontra justificativa pois, se θ é conhecido, $g_{\theta}(\mathbf{z}/\mathbf{x})$ seria a única verossimilhança preditiva (padronizada) e se θ é desconhecido, parece razoável substituí-lo por um estimador (BJORNSTAD (1990)).

Na procura de uma abordagem preditiva mais precisa através de condicionamento, tem-se sugerido vários conceitos de *Funções de Verossimilhança Preditivas Condicionais*, alguns dos quais passaremos a descrever. Antes, apresentaremos algumas definições, lembrando o conceito de *estatística suficiente minimal.*

Definição 2.4.5 *Seja* $\mathbf{X}=(X_1,\cdots,X_n)$ *uma amostra aleatória da densidade* $f(x/\theta)$. *A estatística* $T(\mathbf{X})$ *é uma* **Estatística Suficiente** *para* θ *se a distribuição condicional da amostra* $\mathbf{X}=(X_1,\cdots,X_n)$ *dado o valor de* $T(\mathbf{X})$, *não depende de* θ, *para todo valor de* $T(\mathbf{X})$.

Desta forma, uma *estatística suficiente* para o parâmetro θ capta toda a informação sobre θ contida na amostra. Com isto, dado o valor desta estatística, os elementos amostrais não fornecem nenhuma informação adicional sobre θ. Essas considerações conduzem à técnica de redução de dados conhecida como Principio da Suficiência.

Principio da Suficiência

Se $T(\mathbf{X})$ é uma estatística suficiente para θ então qualquer inferência sobre θ dependerá da amostra somente através de $T(\mathbf{X})$. Assim, se $\mathbf{x}$ e $\mathbf{y}$ são dois pontos amostrais tais que $T(\mathbf{x}) = T(\mathbf{y})$, então a inferência sobre θ será a mesma se $\mathbf{X} = \mathbf{x}$ ou $\mathbf{X} = \mathbf{y}$.

Definição 2.4.6 *Uma estatística suficiente* $T(\mathbf{X})$ *é chamada* **Estatística Suficiente Minimal**, *se para qualquer outra estatística suficiente* $T'(\mathbf{X})$, $T(\mathbf{X})$ *é uma função de* $T'(\mathbf{X})$.

Considerando variáveis aleatórias discretas e Z unidimensional, LAURITZEN (1974) sugere como previsor para Z o valor $\hat{z}$ que maximiza a *Função de Verossimilhança Preditiva Condicional* $L_1(z/\mathbf{x})$, cuja definição daremos a seguir.

Definição 2.4.7 *(LAURITZEN (1974))*
Sejam $\mathbf{X} = (X_1, \cdots, X_n)$ *e* $\mathbf{Y} = (Y_1, \cdots, Y_m)$ *vetores aleatórios discretos observáveis e não observáveis respectivamente, sendo que* $X_1, \cdots, X_n$, $Y_1, \cdots, Y_m$ *são variáveis aleatórias independentes e identicamente distribuídas. Adicionalmente, sejam*

$\mathbf{Z} = h(\mathbf{Y})$ *a variável aleatória que desejamos prever e*

$R = R(\mathbf{X}, \mathbf{Z})$ *a Estatística Suficiente Minimal com base em* $(\mathbf{X}, \mathbf{Z})$, *com valor* $r = r(\mathbf{x}, \mathbf{z})$ *para* $\mathbf{X} = \mathbf{x}$ *e* $\mathbf{Z} = \mathbf{z}$.

Nestas condições, define-se uma **Função de Verossimilhança Preditiva Condicional** *de* $\mathbf{Z}$ *como*

$$L_1(\mathbf{z}/\mathbf{x}) = f(\mathbf{x}/r(\mathbf{x}, \mathbf{z})).$$

onde $f(\mathbf{x}/r(\mathbf{x}, \mathbf{z}))$ *é a densidade condicional de* $\mathbf{X}$ *dado* $r(\mathbf{x}, \mathbf{z})$.

Assim, a função L_1 analisa a verossimilhança do valor de $\mathbf{z}$ tendo conhecimento de que $\mathbf{X} = \mathbf{x}$ e pode ser vista como a *frequência relativa* da observação $\mathbf{X} = \mathbf{x}$ dado o valor da redução suficiente minimal de $(\mathbf{X}, \mathbf{Z})$. De acordo com esta definição, poderíamos verificar sob que valor de $\mathbf{z}$ o valor observado $\mathbf{X} = \mathbf{x}$ é mais provável, e tal procedimento é semelhante à

utilização usual da função de verossimilhança para parâmetros. Conforme veremos na Seção 2.6, este procedimento gera um previsor de máxima verossimilhança de $\mathbf{Z}$, que é o valor de $\mathbf{z}$ que maximiza $L_1(\mathbf{z}/\mathbf{x})$.

Daremos a seguir uma definição que pode ser considerada extensão da definição anterior, porque não exige independência nem idêntica distribuição para $X_1, X_2, \cdots, X_n$ e $Y_1, Y_2, \cdots, Y_m$, embora exija independência entre $\mathbf{X}$ e $\mathbf{Y}$. Além disso, aplica-se também para variáveis aleatórias contínuas.

Definição 2.4.8 *(HINKLEY (1979))*
Sejam $\mathbf{X} = (X_1, \cdots, X_n)$ *e* $\mathbf{Y} = (Y_1, \cdots, Y_m)$ *vetores aleatórios independentes contendo os dados observados e futuros e* $\mathbf{Z} = h(\mathbf{Y})$ *a variável aleatória que desejamos prever. Além disso, sejam* S *e* T *estatísticas suficientes minimais com base respectivamente em* $\mathbf{X}$ *e* $\mathbf{Z}$, *e* R *a estatística suficiente minimal com base em* (S, T), *com valores* $r = r(S(\mathbf{x}), T(\mathbf{z}))$. *Admitindo que* $t = t(r, s)$ *é unicamente definida por* (r, s), *define-se a* **Verossimilhança Preditiva** *de* $T = t$ *dado* $S = s$ *como*

$$L_2(t/s) = f(s/r(s,t)),$$

e a **Verossimilhança Preditiva** *de* $\mathbf{Z} = \mathbf{z}$ *dado* $S = s$ *como*

$$L_2(\mathbf{z}/s) = f(\mathbf{z}/t)L_2(t/s) \qquad para \quad t = t(\mathbf{z}).$$

A **Função de Verossimilhança Preditiva Condicional** *de* $\mathbf{Z} = \mathbf{z}$ *dado* $\mathbf{X} = \mathbf{x}$ *é dada por*

$$L_2(\mathbf{z}/\mathbf{x}) = f(\mathbf{x}/s)L_2(\mathbf{z}/s) = f(\mathbf{x}/s)f(\mathbf{z}/t)f(s/r(s,t)),$$

para $t = T(\mathbf{z})$ *e* $s = S(\mathbf{x})$. *Quando* $\mathbf{X}$ *e* $\mathbf{Y}$ *são discretos, se* $t = t(r, s)$ *for única,*

$$L_2(\mathbf{z}/\mathbf{x}) = f(\mathbf{x}, \mathbf{z}/r(\mathbf{x}, \mathbf{z})).$$

Notamos que, devido à suficiência, a previsão de $\mathbf{Z}$ dado $\mathbf{X} = \mathbf{x}$ é estatisticamente equivalente a prever T dado $S = s$.

Segue também da suficiência que todas as funções presentes no cálculo das duas formas de $L_2(\mathbf{z}/\mathbf{x})$ independem de parâmetros desconhecidos e portanto, podem ser completamente determinadas.

Os exemplos a seguir, entre outras coisas, irão mostrar que as funções de verossimilhança L_1 e L_2 não são em geral iguais.

Exemplo 2.4 Consideremos $\mathbf{X} = (X_1, \cdots, X_n)$ e $\mathbf{Y} = (Y_1, \cdots, Y_m)$ os vetores de dados observados e futuros sendo que $X_1, \cdots, X_n$, $Y_1, \cdots, Y_m$ são variáveis aleatórias independentes com distribuição de *Bernoulli* com parâmetro θ. Vamos admitir que $\mathbf{Z} = \mathbf{Y}$ é a quantidade aleatória a ser prevista.

Verifica-se que, neste caso, as estatísticas suficientes minimais com base em $\mathbf{X}$ e $\mathbf{Y}$ são respectivamente $S = \sum_{i=1}^{n} X_i$, $T = \sum_{j=1}^{m} Y_j$ e que a estatística suficiente minimal com base em (S, T) é $R = S + T$.

Da Definição 2.4.7 temos que

$$\begin{aligned} L_1(\mathbf{z}/\mathbf{x}) &= f(\mathbf{x}/r(\mathbf{x}, \mathbf{z})) \\ &= P(\mathbf{X} = \mathbf{x}/R = s + t), \end{aligned}$$

para $s = \sum_{i=1}^{n} x_i$ e $t = \sum_{i=1}^{m} y_i$.

Portanto,

$$\begin{aligned} L_1(\mathbf{z}/\mathbf{x}) &= \frac{P(\mathbf{X} = \mathbf{x},\, S + T = s + t)}{P(S + T = s + t)} \\ &= \frac{P(\mathbf{X} = \mathbf{x},\, T = t)}{P(R = s + t)} \\ &= \frac{P(\mathbf{X} = \mathbf{x})\, P(T = t)}{P(R = s + t)} \\ &= \frac{\theta^s (1-\theta)^{n-s} \binom{m}{t} \theta^t (1-\theta)^{m-t}}{\binom{n+m}{s+t} \theta^{s+t} (1-\theta)^{n+m-s-t}} \\ &= \frac{\binom{m}{t}}{\binom{n+m}{s+t}}, \end{aligned}$$

onde $\mathbf{z} = \mathbf{y}$, $t = \sum_{i=1}^{m} y_i$ e $s = \sum_{i=1}^{n} x_i$.

Por outro lado, da Definição 2.4.8, para $\mathbf{X}$ e $\mathbf{Y}$ discretos

$$L_2(\mathbf{z}/\mathbf{x}) = f(\mathbf{x}, \mathbf{z}/R = s + t)$$

$$
\begin{aligned}
&= \frac{P(\mathbf{X}=\mathbf{x},\ \mathbf{Z}=\mathbf{z})}{P(R=s+t)} \\
&= \frac{\theta^s(1-\theta)^{n-s}\,\theta^t(1-\theta)^{m-t}}{\binom{n+m}{s+t}\theta^{s+t}(1-\theta)^{n+m-s-t}} \\
&= \frac{1}{\binom{n+m}{s+t}},
\end{aligned}
$$

para $s=\sum_{i=1}^{n} x_i$ e $t=\sum_{i=1}^{m} y_i$.

Assim, vemos que L_1 e L_2 podem resultar em funções diferentes. □

As duas definições anteriores exigem que $\mathbf{X}$ e $\mathbf{Z}$ sejam independentes, embora as mesmas possam ser espandidas para cobrir o caso de sequências de variáveis aleatórias dependentes, sendo necessário para isto modificar-se algumas das condições. Esta extensão considera que S, estatística suficiente para $\mathbf{X}$, pode não ser minimal, e que a estatística suficiente minimal de $\mathbf{Z}$ pode ser determinada por T e S, isto é, não necessariamente só por T (HINKLEY (1979)).

Finalizaremos o estudo de *Funções de Verossimilhanças Preditivas Condicionais* definindo esta extensão.

Definição 2.4.9 *(HINKLEY (1979))*
Sejam $\mathbf{X}=(X_1,\cdots,X_n)$ *e* $\mathbf{Y}=(Y_1,\cdots,Y_m)$ *os conjuntos de variáveis aleatórias observáveis e não observáveis respectivamente e* $\mathbf{Z}=h(\mathbf{Y})$ *a variável aleatória que desejamos prever. Além disso, consideremos* $R=R(\mathbf{X},\mathbf{Z})$ *a estatística suficiente minimal para* $(\mathbf{X},\mathbf{Z})$ *e* S *a estatística suficiente para* $\mathbf{X}$. *Admitindo que existe uma função* T *de* $(\mathbf{Z},S)$ *tal que*

(i) R *é determinada por* (S,T)

(ii) a Estatística Suficiente Minimal de $\mathbf{Z}$ *é função de* (S,T)

(iii) $t=t(r,s)$ *está unicamente determinada para cada* s, *onde* $t(r,s)$ *é inversa de* $r(s,t)$,

então, a **Verossimilhança Preditiva** *de* $T=t$ *dado* $S=s$ *é*

$$L_3(t/s)=f(s/r(s,t)),$$

e a **Verossimilhança Preditiva** *de* $\mathbf{Z} = \mathbf{z}$ *dado* $S = s$ *é*

$$L_3(\mathbf{z}/s) = f(\mathbf{z}/s,t)L_3(t/s) = f(\mathbf{z}/s,t)\, f(s/r(s,t)).$$

A última classe de *Funções de Verossimilhança Preditivas* que apresentaremos será aquela em que o parâmetro θ da função de verossimilhança $l_{\mathbf{x}}(\mathbf{z},\theta)$ é eliminado via integração.

Funções de Verossimilhança Preditivas Marginais

Apresentaremos agora o conceito de *Função de Verossimilhança Preditiva Marginal.* Este tipo de função apresenta algumas semelhanças com as funções preditivas bayesianas, semelhanças estas que serão discutidas logo após a seguinte definição.

Definição 2.4.10 *(BJORNSTAD (1990))*
Sejam $\mathbf{X} = (X_1, \cdots, X_n)$ *e* $\mathbf{Y} = (Y_1, \cdots, Y_m)$ *os conjuntos de variáveis aleatórias observáveis e não observáveis respectivamente e* $\mathbf{Z} = h(\mathbf{Y})$ *a variável aleatória que desejamos prever. Se* $f(\mathbf{x},\mathbf{z};\theta)$ *é a distribuição de probabilidades de* $(\mathbf{X},\mathbf{Z})$*, de modo que* $l_{\mathbf{x}}(\mathbf{z},\theta) = f(\mathbf{x},\mathbf{z};\theta)$ *é a Função de Verossimilhança de* $(\mathbf{Z},\theta)$*, então*

$$\begin{aligned} L_m(\mathbf{z}/\mathbf{x}) &= \int_{\theta\in\Theta} l_{\mathbf{x}}(\mathbf{z},\theta)d\theta \\ &= \int_{\theta\in\Theta} f(\mathbf{x},\mathbf{z};\theta)d\theta \end{aligned}$$

é denominada **Função de Verossimilhança Preditiva Marginal de Z.**

Para verificarmos a semelhança entre a densidade preditiva bayesiana e a *Função de Verossimilhança Preditiva Marginal*, lembramos que, no contexto bayesiano, a densidade preditiva de $\mathbf{Z}$ dado $\mathbf{X} = \mathbf{x}$, quando $f(\theta)$ é uma densidade a priori para θ, é dada por

$$\begin{aligned} f(\mathbf{z}/\mathbf{x}) &= \frac{\int_{\theta\in\Theta} f(\mathbf{x},\mathbf{z}/\theta)f(\theta)\,d\theta}{\int_{\theta\in\Theta} f(\mathbf{x}/\theta)f(\theta)\,d\theta} \\ &= k(\mathbf{x})\int_{\theta\in\Theta} f(\mathbf{x},\mathbf{z}/\theta)f(\theta)\,d\theta. \end{aligned}$$

Como

$$\begin{aligned} L_m(\mathbf{z}/\mathbf{x}) &= \int_{\theta\in\Theta} l_{\mathbf{x}}(\mathbf{z},\theta)\,d\theta \\ &= \int_{\theta\in\Theta} f(\mathbf{x},\mathbf{z};\theta)\,d\theta \\ &= \int_{\theta\in\Theta} f(\mathbf{x},\mathbf{z};\theta)(1)\,d\theta, \end{aligned}$$

observamos que $L_m(\mathbf{z}/\mathbf{x})$ é proporcional à correspondente densidade preditiva bayesiana com $f(\theta) = 1$, isto é, quando θ tem uma distribuição a priori constante, pois $f(\mathbf{x},\mathbf{z};\theta)$ na formulação clássica é igual a $f(\mathbf{x},\mathbf{z}/\theta)$ na abordagem bayesiana.

Exemplo 2.5 Sejam X_i, Y_j variáveis aleatórias independentes para $i = 1,2,\cdots,n$ e $j = 1,2,\cdots,m$, com distribuição de *Bernoulli* com parâmetro θ, $S = \sum_{i=1}^{n} X_i \sim \mathcal{B}inomial(n,\theta)$ e $Z = \sum_{j=1}^{m} Y_i \sim \mathcal{B}inomial(m,\theta)$ a quantidade que desejamos prever. Nestas condições, a densidade conjunta de $(\mathbf{X}, Z)$

$$\begin{aligned} l_{\mathbf{x}}(\mathbf{z},\theta) &= f(\mathbf{x};\theta)\,g(z;\theta) \\ &= \theta^s(1-\theta)^{n-s}\binom{m}{z}\theta^z(1-\theta)^{m-z} \\ &= \binom{m}{z}\theta^{s+z}(1-\theta)^{n+m-s-z} \end{aligned}$$

é a função de verossimilhança de $(Z;\theta)$. Logo, de acordo com a Definição 2.4.10, obtemos a *Função de Verossimilhança Preditiva Marginal* de Z, $L_m(\mathbf{z}/\mathbf{x})$, integrando $l_{\mathbf{x}}(\mathbf{z},\theta)$ em θ, ou seja,

$$\begin{aligned} L_m(\mathbf{z}/\mathbf{x}) &= \int_0^1 \binom{m}{z}\theta^{s+z}(1-\theta)^{n+m-s-z}\,d\theta \\ &= \binom{m}{z}\int_0^1 \theta^{s+z}(1-\theta)^{n+m-s-z}\,d\theta. \end{aligned}$$

A integral acima pode ser calculada através da função *beta*, pois

$$\int_0^1 \theta^{s+z}(1-\theta)^{n+m-s-z}\,d\theta = \mathcal{B}eta(s+z+1, n+m-s-z+1).$$

Como para $a > 0$ e $b > 0$, tem-se que $\mathcal{B}eta(a,b) = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}$, então,

$$\int_0^1 \theta^{s+z}\,(1-\theta)^{n+m-s-z}\,d\theta = \frac{\Gamma(s+z+1)\,\Gamma(n+m-s-z+1)}{\Gamma(n+m+2)}.$$

Além disso, $s+z+1$ e $n+m-s-z+1$ são inteiros, e lembrando que, para n inteiro, $\Gamma(n+1) = n!$, temos

$$\begin{aligned}
L_m(\mathbf{z}/\mathbf{x}) &= \binom{m}{z} \int_0^1 \theta^{s+z}\,(1-\theta)^{n+m-s-z}\,d\theta \\
&= \binom{m}{z} \frac{\Gamma(s+z+1)\,\Gamma(n+m-s-z+1)}{\Gamma(n+m+2)} \\
&= \binom{m}{z} \frac{(s+z)!(n+m-s-z)!}{(n+m+1)!} \\
&= \binom{m}{z} \frac{1}{(n+m+1)} \frac{1}{\binom{n+m}{s+z}} \\
&= \frac{1}{(n+m+1)} \frac{\binom{m}{z}}{\binom{n+m}{s+z}}, \qquad 0 \le z \le m.
\end{aligned}$$

Padronizando $L_m(\mathbf{z}/\mathbf{x})$, a função obtida pode ser identificada como uma distribuição *hipergeométrica negativa.* □

2.5 Funções de Verossimilhança Preditivas Padronizadas

Conforme comentado na Seção 2.4 no caso particular das *Funções de Verossimilhança Preditivas* do tipo profile, *Funções de Verossimilhança Preditivas* de um modo geral podem ser padronizadas de modo a se tornarem densidades.

Assim, a padronização da *Função de Verossimilhança Preditiva* $L(\mathbf{z}/\mathbf{x})$ é uma transformação desta numa função $g(\mathbf{z}/\mathbf{x})$ tal que

$$g(\mathbf{z}/\mathbf{x}) = k(\mathbf{x})\,L(\mathbf{z}/\mathbf{x}),$$

onde $k(\mathbf{x})$ é uma constante de padronização que não depende de $\mathbf{z}$, mas pode depender de $\mathbf{x}$ e tal que

$$\int_{-\infty}^{\infty} k(\mathbf{x})\,L(\mathbf{z}/\mathbf{x})\,d\mathbf{z} = 1.$$

O procedimento de padronização tem inúmeras vantagens. Além de permitir expressar e operar *Funções de Verossimilhança Preditivas* como densidades de probabilidade, facilita a comparação entre *Funções de Verossimilhanças Preditivas* diferentes.

Adicionalmente, alguns autores utilizam *Funções de Verossimilhança Preditivas* como estimadores da densidade condicional $f(\mathbf{z}/\mathbf{x})$ e, com esse objetivo, a padronização se torna obrigatória.

Outra vantagem do uso de *Funções de Verossimilhança Preditivas* padronizadas é a possibilidade de construção de intervalos de previsão, assunto que será discutido a seguir.

2.6 Utilização das Funções de Verossimilhança Preditivas

Uma das principais aplicações de *Funções de Verossimilhança Preditivas* é na determinação de previsores para a quantidade de interesse não observada, que estamos indicando por $\mathbf{Z}$. A próxima definição introduz o conceito de *previsor de máxima verossimilhança* para $\mathbf{Z}$.

Definição 2.6.1 *(BJORNSTAD (1990))*
Se $L(\mathbf{z}/\mathbf{x})$ *é a Função de Verossimilhança Preditiva de* $\mathbf{Z}$, *o* **Previsor de Máxima Verossimilhança** *de* $\mathbf{Z}$, *que indicaremos por* $\hat{\mathbf{z}}_{mv}$, *é o valor de* $\mathbf{z}$ *que maximiza* $L(\mathbf{z}/\mathbf{x})$. *Além disso, se a correspondente Função de Verossimilhança Preditiva padronizada* $g(\mathbf{z}/\mathbf{x})$ *for uma distribuição simétrica, então o previsor de máxima verossimilhança coincidirá com a média dessa distribuição, ou seja*

$$\hat{\mathbf{z}}_{mv} = E_g(\mathbf{Z}) = \int \mathbf{z}\, g(\mathbf{z}/\mathbf{x})\, d\mathbf{z},$$

denominada Esperança Preditiva de $\mathbf{Z}$.

O próximo exemplo ilustra o cálculo do *previsor de máxima verossimilhança*, que acabamos de definir.

Exemplo 2.6 Sejam X_i, Y_j variáveis aleatórias independentes com distribuição $\mathcal{N}(\theta, \sigma^2)$, $i = 1, 2, \cdots, n$ e $j = 1, 2, \cdots, m$, com σ^2 conhecido. Se $Z = \frac{\sum_{j=1}^{m} Y_j}{m}$ é a quantidade que desejamos prever, a densidade conjunta de $(\mathbf{X}, Z)$ será

$$l_{\mathbf{x}}(z, \theta) = f(\mathbf{x}; \theta)\, g(z;\, \theta),$$

onde

$$f(\mathbf{x}; \theta) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2\}$$

e

$$g(z;\, \theta) = (2\pi\sigma^2)^{-\frac{1}{2}}\, m^{\frac{1}{2}} \exp\{-\frac{m}{2\sigma^2}(z - \theta)^2\}.$$

Então, a *Função de Verossimilhança Preditiva Marginal* de $\mathbf{Z}$ é dada por

$$\begin{aligned}
L_m(z/\mathbf{x}) &= \int_{-\infty}^{\infty} l_{\mathbf{x}}(z, \theta)\, d\theta \\
&= \int_{-\infty}^{\infty} (2\pi\sigma^2)^{-\frac{n+1}{2}}\, m^{\frac{1}{2}}\, e^{\left\{-\frac{1}{2\sigma^2}\left[m(z-\theta)^2 + \sum_{i=1}^{n}(x_i-\theta)^2\right]\right\}}\, d\theta \\
&= (2\pi\sigma^2)^{-\frac{n+1}{2}}\, m^{\frac{1}{2}} \int_{-\infty}^{\infty} e^{\left\{-\frac{1}{2\sigma^2}\left[m(z-\theta)^2 + \sum_{i=1}^{n}(x_i-\theta)^2\right]\right\}}\, d\theta \\
&= k \int_{-\infty}^{\infty} e^{\left\{-\frac{1}{2\sigma^2}\left[mz^2 - 2mz\theta + m\theta^2 + \sum_{i=1}^{n} x_i^2 - 2\theta\sum_{i=1}^{n} x_i + n\theta^2\right]\right\}}\, d\theta \\
&= k\, e^{\left\{-\frac{1}{2\sigma^2}\left(mz^2 + \sum_{i=1}^{n} x_i^2\right)\right\}} \int_{-\infty}^{\infty} e^{\left\{-\frac{1}{2\sigma^2}\left[-2(mz + \sum_{i=1}^{n} x_i)\theta + (n+m)\theta^2\right]\right\}}\, d\theta \\
&= k_1(\mathbf{x}) \int_{-\infty}^{\infty} e^{\left\{-\frac{(n+m)}{2\sigma^2}\left[\theta^2 - 2\theta\left(\frac{mz + \sum_{i=1}^{n} x_i}{n+m}\right) + \left(\frac{mz + \sum_{i=1}^{n} x_i}{n+m}\right)^2 - \left(\frac{mz + \sum_{i=1}^{n} x_i}{n+m}\right)^2\right]\right\}}\, d\theta \\
&= k_1(\mathbf{x})\, e^{\left\{\frac{(n+m)}{2\sigma^2}\left(\frac{mz + \sum_{i=1}^{n} x_i}{n+m}\right)^2\right\}} \int_{-\infty}^{\infty} e^{\left\{-\frac{1}{2\frac{\sigma^2}{(n+m)}}\left[\theta - \left(\frac{mz + \sum_{i=1}^{n} x_i}{n+m}\right)\right]^2\right\}}\, d\theta \\
&= k_2(\mathbf{x}) \left(\frac{2\pi\sigma^2}{n+m}\right)^{\frac{1}{2}} \int_{-\infty}^{\infty} \left(\frac{n+m}{2\pi\sigma^2}\right)^{\frac{1}{2}} e^{\left\{-\frac{1}{2\frac{\sigma^2}{(n+m)}}\left[\theta - \left(\frac{mz + \sum_{i=1}^{n} x_i}{n+m}\right)\right]^2\right\}}\, d\theta \\
&= k_3(\mathbf{x}) \int_{-\infty}^{\infty} \left(\frac{n+m}{2\pi\sigma^2}\right)^{\frac{1}{2}} e^{\left\{-\frac{1}{2\frac{\sigma^2}{(n+m)}}\left[\theta - \left(\frac{mz + \sum_{i=1}^{n} x_i}{n+m}\right)\right]^2\right\}}\, d\theta.
\end{aligned}$$

A expressão dentro da integral é a densidade de uma variável aleatória com distribuição normal com média $\frac{mz + \sum_{i=1}^{n} x_i}{n+m}$ e variância $\frac{\sigma^2}{(n+m)}$. Portanto, a

Função de Verossimilhança Preditiva Marginal de $\mathbf{Z}$ fica

$$
\begin{aligned}
L_m(z/\mathbf{x}) &= k_3(\mathbf{x}) \\
&= k_2(\mathbf{x}) \left(\frac{2\pi\sigma^2}{n+m}\right)^{\frac{1}{2}} \\
&= k_1(\mathbf{x})\, e^{\left\{\frac{(n+m)}{2\sigma^2}\left(\frac{mz+\sum_{i=1}^n x_i}{n+m}\right)^2\right\}} \left(\frac{2\pi\sigma^2}{n+m}\right)^{\frac{1}{2}} \\
&= k\, e^{\left\{-\frac{1}{2\sigma^2}(mz^2+\sum_{i=1}^n x_i^2)\right\}}\, e^{\left\{\frac{(n+m)}{2\sigma^2}\left(\frac{mz+\sum_{i=1}^n x_i}{n+m}\right)^2\right\}} \left(\frac{2\pi\sigma^2}{n+m}\right)^{\frac{1}{2}} \\
&= k'\, e^{\left\{-\frac{1}{2\sigma^2}(mz^2+\sum_{i=1}^n x_i^2)\right\}}\, e^{\left\{\frac{(n+m)}{2\sigma^2}\left(\frac{mz+\sum_{i=1}^n x_i}{n+m}\right)^2\right\}} \\
&= k'\, e^{\left\{-\frac{1}{2\sigma^2}\left(mz^2+\sum_{i=1}^n x_i^2-\left(\frac{m^2z^2+(\sum_{i=1}^n x_i)^2+2mz\sum_{i=1}^n x_i}{n+m}\right)\right)\right\}} \\
&= k'\, e^{\left\{-\frac{1}{2\sigma^2(n+m)}\left(nmz^2+m^2z^2+n\sum_{i=1}^n x_i^2+m\sum_{i=1}^n x_i^2-m^2z^2-(\sum_{i=1}^n x_i)^2-2mz\sum_{i=1}^n x_i\right)\right\}} \\
&= k'\, e^{\left\{-\frac{nm}{2\sigma^2(n+m)}\left(z^2+\frac{\sum_{i=1}^n x_i^2}{m}+\frac{\sum_{i=1}^n x_i^2}{n}-\frac{(\sum_{i=1}^n x_i)^2}{nm}-2\frac{z}{n}\sum_{i=1}^n x_i\right)\right\}} \\
&= k'\, e^{\left\{-\frac{nm}{2\sigma^2(n+m)}\left(z^2+\left(\frac{1}{n}+\frac{1}{m}\right)\sum_{i=1}^n x_i^2-\frac{(\sum_{i=1}^n x_i)^2}{nm}-2z\bar{x}\right)\right\}} \\
&= k'\, e^{\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2-\frac{(\sum_{i=1}^n x_i)^2}{n+m}+\frac{nm}{n+m}\left((z-\bar{x})^2-\bar{x}^2\right)\right)\right\}} \\
&= k'\, e^{\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2-\frac{n^2}{n+m}\bar{x}^2-\frac{nm}{n+m}\bar{x}^2\right)\right\}}\, e^{\left\{-\frac{1}{2\sigma^2}\frac{(z-\bar{x})^2}{\left(\frac{1}{n}+\frac{1}{m}\right)}\right\}} \\
&= k_4(\mathbf{x})\, e^{\left\{-\frac{1}{2\sigma^2}\frac{(z-\bar{x})^2}{\left(\frac{1}{n}+\frac{1}{m}\right)}\right\}}.
\end{aligned}
$$

Observamos então que a função $L_m(\mathbf{z}/\mathbf{x})$ é proporcional à densidade de uma variável aleatória com distribuição normal com média $\bar{x}$ e variância $\left(\frac{1}{n}+\frac{1}{m}\right)\sigma^2$. Padronizando, obtemos a densidade preditiva de $\mathbf{Z}$ dada por

$$
g_m(z/\mathbf{x}) = \left(2\pi\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)\right)^{-\frac{1}{2}} exp\left\{-\frac{1}{2\sigma^2}\,\frac{(z-\bar{x})^2}{\left(\frac{1}{n}+\frac{1}{m}\right)}\right\}. \qquad (2.4)
$$

O previsor de máxima verossimilhança de $\mathbf{Z}$, obtido a partir da função de verossimilhança preditiva marginal, será o valor de $\mathbf{z}$ que maximiza $L_m(\mathbf{z}/\mathbf{x})$ ou, equivalentemente, $g_m(z/\mathbf{x})$. Este previsor é

$$
\hat{\mathbf{z}}_{mv} = E_{g_m}(\mathbf{Z}) = \int \mathbf{z}\, g_m(\mathbf{z}/\mathbf{x})\, d\mathbf{z} = \bar{x}. \qquad \square
$$

Uma vez obtido o previsor para a quantidade $\mathbf{Z}$, pode ser de interesse a construção de um *intervalo* ou *região de previsão* para esta variável. *Funções de Verossimilhança Preditivas Padronizadas* podem também ser utilizadas na determinação de tais intervalos, conforme veremos a seguir.

O problema da construção de um intervalo de previsão pode ser definido da seguinte maneira. Observado o vetor de dados $\mathbf{X} = (X_1, \cdots, X_n)$, desejamos fazer afirmações preditivas sobre $\mathbf{Z} = h(\mathbf{Y})$, onde $\mathbf{Y} = (Y_1, \cdots, Y_m)$ são observações futuras. Tal afirmação poderá ser na forma de um intervalo, possivelmente centrado num previsor pontual de $\mathbf{Z}$.

Na literatura, grande parte dos métodos para obter intervalos de previsão fazem uso de Testes de Hipóteses (FRASER e GUTTMAN (1956)), Aproximação Normal (NELSON (1968)), Quantidades Pivotais (HAHN (1969)), Condicionamento numa Estatística Suficiente (FAULKENBERRY (1973)), Verossimilhanças Preditivas (HINKLEY (1979), BJORNSTAD (1990)) e de Densidades Bayesianas a posteriori.

HINKLEY (1979) mostra como *Funções de Verossimilhança Preditivas* podem ser utilizadas em conjunto com métodos clássicos na construção de intervalos de previsão. O autor propõe o uso destas funções na escolha de uma região, dentre várias de mesmo tamanho, obtidas através do método da correspondência com testes de hipóteses. Sugere ainda idêntica utilização quando o intervalo é obtido através do método da quantidade pivotal.

BJORNSTAD (1990) define de forma geral e simples uma região de previsão com coeficiente de confiança $(1-\alpha)$ baseada nas *Funções de Verossimilhaça Preditivas Padronizadas*. Sua abordagem é aparentemente mais viável do ponto de vista prático.

Definição 2.6.2 *(BJORNSTAD (1990))*
Se $L(\mathbf{z}/\mathbf{x})$ *é uma função de verossimilhaça preditiva padronizada de* $\mathbf{Z}$, *observado* $\mathbf{X} = \mathbf{x}$, *uma* **Região de Previsão** *com coeficiente de confiança* $(1-\alpha)$ *para* $\mathbf{Z}$ *será o conjunto*

$$\mathcal{P}_\alpha(\mathbf{x}) = \{\mathbf{z} : L(\mathbf{z}/\mathbf{x}) \geq k_\alpha\},$$

onde k_α *é tal que*

$$\int_{\mathcal{P}_\alpha(\mathbf{x})} L(\mathbf{z}/\mathbf{x})d\mathbf{z} = 1-\alpha \qquad \textit{se } L \textit{ é contínua e}$$

$$\sum_{\mathbf{z}\in\mathcal{P}_\alpha(\mathbf{x})} L(\mathbf{z}/\mathbf{x}) = 1-\alpha \qquad \textit{se } L \textit{ é discreta.}$$

Assim, a **Região de Previsão** conterá os valores $\mathbf{z}$ com verossimilhanças preditivas mais altas.

Exemplo 2.7 Com os dados do Exemplo 2.6, construiremos a região de previsão com coeficiente de confiança $(1-\alpha)$ para $\mathbf{Z}$. Esta região é da forma

$$\mathbf{I_x} = \{z \,:\, L(z/\mathbf{x}) \geq k_\alpha\}\,.$$

Se utilizarmos a *Função de Verossimilhança Preditiva* padronizada dada em (2.4), temos

$$L(\mathbf{z}/\mathbf{x}) = f_m(z/\mathbf{x}) = \left(2\pi\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)\right)^{-\frac{1}{2}} e^{\left\{-\frac{1}{2\sigma^2}\frac{(z-\bar{x})^2}{\left(\frac{1}{n}+\frac{1}{m}\right)}\right\}}.$$

Como esta densidade é simétrica e coincide com a da distribuição normal com média $\bar{x}$ e variância $\left(\frac{1}{n}+\frac{1}{m}\right)\sigma^2$, temos que $\mathbf{I_x}$ é o intervalo $\{z \,:\, \hat{z}_{\frac{\alpha}{2}} \leq z \leq \hat{z}_{1-\frac{\alpha}{2}}\}$, de modo que $\int_{\hat{z}_{\frac{\alpha}{2}}}^{\hat{z}_{1-\frac{\alpha}{2}}} L(\mathbf{z}/\mathbf{x})\,d\mathbf{z} = 1-\alpha$.

Portanto, $\hat{z}_{\frac{\alpha}{2}}$ e $\hat{z}_{1-\frac{\alpha}{2}}$ são tais que

$$P\left(\hat{z}_{\frac{\alpha}{2}} \leq \mathbf{W} \leq \hat{z}_{1-\frac{\alpha}{2}}\right) = 1-\alpha,$$

onde $\mathbf{W} \sim \mathcal{N}\left(\bar{x},\, \sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)\right)$.

Como consequência,

$$\hat{z}_{\frac{\alpha}{2}} = \bar{x} - w\,\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}$$

e

$$\hat{z}_{1-\frac{\alpha}{2}} = \bar{x} + w\,\sigma\sqrt{\frac{1}{n}+\frac{1}{m}},$$

onde w é tal que $P(-w \leq \mathbf{W}^* \leq w) = 1-\alpha$, $\mathbf{W}^* \sim \mathcal{N}(0,1)$.

Portanto, a região de previsão com coeficiente de confiança $1-\alpha$ para $\mathbf{Z}$ é dada por

$$\mathbf{I_x} = \left\{z \,:\, \bar{x} - w\sigma\sqrt{\frac{1}{n}+\frac{1}{m}} \leq z \leq \bar{x} + w\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}\right\},$$

onde $\bar{x}$ é a média amostral observada. □.

Funções de Verossimilhança Preditivas podem ainda ser utilizadas na estimação de funções densidade.

De acordo com a literatura, é possível considerar uma *Função de Verossimilhança Preditiva* padronizada como uma estimativa da distribuição condicional de $\mathbf{Z}$, dado $\mathbf{x}$, $f(\mathbf{z}/\ \mathbf{x})$. Em particular, se $\mathbf{Z}$ e $\mathbf{X}$ são independentes, $f(\mathbf{z}/\ \mathbf{x})$ corresponderá à densidade de $\mathbf{Z}$. Sob determinadas condições, LEJEUNE e FAULKENBERRY (1982), LEVY e PERNG (1986) e HARRIS (1989) utilizaram densidades preditivas como estimadores de $f(\mathbf{z}/\ \mathbf{x})$ e de $f(\mathbf{z})$.

Neste contexto, certas propriedades de consistência assintótica são desejáveis. HINKLEY (1979) e MATHIASEN (1979) foram os primeiros a discutí-las.

Quando $\mathbf{X}$ e $\mathbf{Z}$ são independentes, para L padronizada, as propriedades em discussão podem ser formuladas do seguinte modo. Dispomos de uma variável aleatória $\mathbf{Z}$, correspondente à quantidade a ser prevista, com função densidade $f_\theta(\mathbf{z})$ desconhecida, pois depende do parâmetro desconhecido θ. Um possível estimador para $f_\theta(\mathbf{z})$ é a *Função de Verossimilhança Preditiva* $L(\mathbf{z}/\mathbf{x})$, que depende do valor observado de $\mathbf{X}$, vetor aleatório n-dimensional observado.

Diremos que $L(\mathbf{z}/\mathbf{X})$ é um estimador consistente de $f_\theta(\mathbf{z})$ se $L(\mathbf{z}/\mathbf{X})$ converge em probabilidade para $f_\theta(\mathbf{z})$, para $n \longrightarrow \infty$, ou seja, se

$$L(\mathbf{z}/\mathbf{X}) \xrightarrow{P} f_\theta(\mathbf{z}) \quad \text{quando} \quad n \longrightarrow \infty.$$

Esta definição nada mais é que a da consistência de estimadores, aplicada à estimação de funções densidade. Tal conceito será visto com mais detalhes no próximo capítulo, quando nos dedicaremos ao estudo de *Densidades Preditivas* no modelo de regressão linear.

Capítulo 3

Densidades Preditivas e Funções de Verossimilhança Preditivas no modelo de regressão linear

3.1 Introdução

Na literatura, têm sido propostas várias Funções de Verossimilhança Preditivas e Densidades Preditivas, para as mais variadas situações. Neste capítulo, iremos estudar algumas destas funções para a previsão de dados não observados, no modelo de regressão linear em que a variável resposta tem distribuição normal. Apresentaremos quatro densidades preditivas que podem também ser utilizadas como estimativas da densidade normal multivariada correspondente à variável resposta para os dados não observados. Estudaremos a *Densidade Pseudo Preditiva*, a *Densidade Preditiva Modificada de Fisher*, proposta por LEVY e PERNG (1984), a *Densidade Preditiva Bayesiana* e a *Densidade Preditiva Não Viciada de Mínima Variância.*

Na Seção 3.2, deduzimos cada uma dessas densidades. Na Seção 3.3, iremos abordar aspectos relativos à consistência de algumas dessas densidades preditivas, quando utilizadas como estimadores das densidades dos dados não observados. Finalizando, a Seção 3.4 demonstra que, sob certas condições, a *Densidade Preditiva Bayesiana* é a densidade preditiva ótima numa classe específica de densidades preditivas, segundo o critério proposto por LEVY e PERNG (1986).

Os modelos lineares, com os quais trabalharemos neste capítulo, são apresentados a seguir. Assim sejam

$$\mathbf{Y} = X\beta + \mathbf{u}, \tag{3.1}$$

$$\mathbf{Z} = W\beta + \mathbf{u}^*, \tag{3.2}$$

onde

$\mathbf{Y} = (Y_1, \cdots, Y_n)'$ e $\mathbf{Z} = (Z_1, \cdots, Z_m)'$ são respectivamente os vetores de variáveis aleatórias observadas e não observadas, de dimensões n e m,

X e W são matrizes de constantes conhecidas $n \times p$ e $m \times p$, sendo que X é de posto completo,

$\beta = (\beta_1, \cdots, \beta_p)'$ é o vetor de parâmetros de regressão desconhecidos, $\beta \in \Omega_\beta$, onde Ω_β é o espaço paramétrico para β e

$\mathbf{u}$ e $\mathbf{u}^*$ são vetores aleatórios independentes não observados, de dimensões n e m, respectivamente e tais que $\mathbf{u} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I)$ e $\mathbf{u}^* \sim \mathcal{N}_m(\mathbf{0}, \sigma^2 I)$, onde σ^2 é um parâmetro positivo desconhecido.

Nestas condições, o espaço paramétrico é $\Theta = \{\theta = (\beta, \sigma^2) : \beta \in \Omega_\beta;\ \sigma^2 > 0\}$.

Para os modelos apresentados, as distribuições de $\mathbf{Y}$ e $\mathbf{Z}$, denotadas por $p_n(\mathbf{y}, X/\beta, \sigma^2)$ e $p_m(\mathbf{z}, W/\beta, \sigma^2)$ são

$$p_n(\mathbf{y}, X/\beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\{-(2\sigma^2)^{-1}(\mathbf{y} - X\beta)'(\mathbf{y} - X\beta)\} \tag{3.3}$$

e

$$p_m(\mathbf{z}, W/\beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{m}{2}} \exp\{-(2\sigma^2)^{-1}(\mathbf{z} - W\beta)'(\mathbf{z} - W\beta)\}. \tag{3.4}$$

Como $\mathbf{Z}$ é o vetor de dados não observados, nosso objetivo nesse capítulo é o da determinação das densidades preditivas de $\mathbf{Z}$. No modelo linear normal $\mathbf{Y} = X\beta + \mathbf{u}$, os estimadores de *máxima verossimilhança* de β e σ^2 são dados por

$$\hat{\beta} = (X'X)^{-1}X'\mathbf{Y} \tag{3.5}$$

e

$$\hat{\sigma}^2 = \frac{(\mathbf{Y} - X\hat{\beta})'(\mathbf{Y} - X\hat{\beta})}{n}. \tag{3.6}$$

Desta forma, se $\theta = (\beta, \sigma^2)$ é o parâmetro de interesse, o estimador de máxima verossimilhança de θ será

$$\hat{\theta} = \left(\hat{\beta}, \hat{\sigma}^2\right). \tag{3.7}$$

Os modelos (3.1) e (3.2) podem ser escritos numa forma condensada, que apresentaremos a seguir e que será útil para a dedução das densidades preditivas *Modificada de Fisher* e *Bayesiana.*

Assim, se $V = [\mathbf{Y}'\ \mathbf{Z}']'$, o modelo de regressão linear com vetor de variáveis resposta V fica

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{Z} \end{bmatrix}_{(n+m)\times 1} = \begin{bmatrix} X \\ W \end{bmatrix}_{(n+m)\times p} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}_{p\times 1} + \begin{bmatrix} u_1 \\ \vdots \\ u_n \\ u_1^* \\ \vdots \\ u_m^* \end{bmatrix}_{(n+m)\times 1}$$

ou equivalentemente

$$\mathbf{V} = H\beta + \mathbf{e}, \tag{3.8}$$

onde $\beta = (\beta_1, \cdots, \beta_p)'$,

$H = \begin{bmatrix} X \\ W \end{bmatrix}_{(n+m)\times p}$ é uma matriz de constantes de posto completo,

$\mathbf{e} = \begin{bmatrix} \mathbf{u} \\ \mathbf{u}^* \end{bmatrix}_{(n+m)\times 1}$ é o vetor de erros aleatórios tal que $\mathbf{e} \sim \mathcal{N}_{n+m}(\mathbf{0}, \sigma^2 I)$.

Se denotarmos por $\hat{\beta}_{\mathbf{z}}$ e $\hat{\sigma}^2_{\mathbf{z}}$ os estimadores de máxima verossimilhança de β e σ^2 respectivamente, baseados em $(\mathbf{Y}, \mathbf{Z})$, isto é, sob o modelo $\mathbf{V} = H\beta + \mathbf{e}$, então

$$\hat{\beta}_{\mathbf{z}} = (H'H)^{-1}H'\mathbf{V} \tag{3.9}$$

e

$$\hat{\sigma}^2_{\mathbf{z}} = \frac{(\mathbf{V} - H\hat{\beta}_{\mathbf{z}})'(\mathbf{V} - H\hat{\beta}_{\mathbf{z}})}{n+m}. \tag{3.10}$$

De forma similar ao modelo anterior, se $\theta = (\beta, \sigma^2)$ é o parâmetro de interesse, o estimador de máxima verossimilhança de θ baseado em $(\mathbf{Y}, \mathbf{Z})$ será

$$\hat{\theta}_{\mathbf{z}} = \left(\hat{\beta}_{\mathbf{z}}, \hat{\sigma}^2_{\mathbf{z}}\right). \tag{3.11}$$

Destacamos que, na verdade, o cálculo de $\hat{\theta}_{\mathbf{z}}$ é apenas uma construção auxiliar para determinarmos as densidades preditivas de interesse, pois $\hat{\theta}_{\mathbf{z}}$ envolve o vetor $\mathbf{z}$ que é desconhecido.

Tendo definido os modelos e os correspondentes estimadores de máxima verossimilhança, passamos a descrever o procedimento para se obter as densidades preditivas de $\mathbf{Z}$.

3.2 Densidades Preditivas

Nesta seção, sob os modelos (3.1) e (3.2), desenvolvemos inicialmente três densidades preditivas para $\mathbf{Z}$, que foram propostas por LEVY e PERNG (1984). Na obtenção de algumas destas funções, ficam ilustradas as técnicas de eliminação dos parâmetros *nuisance*, estudadas no Capítulo 2.

Densidade Pseudo Preditiva ou Clássica

Uma possível forma de construir funções de previsão para $\mathbf{Z}$ é substituir os parâmetros desconhecidos β e σ^2 na densidade (3.4) por estimadores consistentes calculados com base nos dados observados (LEVY E PERNG (1984)). Em particular, para o modelo linear considerado em (3.2), se substituimos em (3.4) os parâmetros β e σ^2 por seus respectivos estimadores de máxima verossimilhança baseados nas observações $\mathbf{Y}$, iremos obter a densidade *Clássica* ou *Pseudo Preditiva*, proposta por GEISSER (1970),

$$\begin{aligned} q(\mathbf{z}, W/\mathbf{y}, X) &= p_m(\mathbf{z}, W/\hat{\beta}, \hat{\sigma}^2) \\ &= (2\pi\hat{\sigma}^2)^{-\frac{m}{2}} \exp\{-(2\hat{\sigma}^2)^{-1}(\mathbf{z} - W\hat{\beta})'(\mathbf{z} - W\hat{\beta})\}, \end{aligned} \tag{3.12}$$

sendo $\hat{\beta}$ e $\hat{\sigma}^2$ definidos em (3.5) e (3.6).

Densidade Preditiva Modificada de Fisher

MATHIASEN (1979), admitindo que $\mathbf{Y}$ e $\mathbf{Z}$ são vetores aleatórios independentes, sugere a utilização das seguintes funções de verossimilhança preditivas

$$LL(\mathbf{z}/\mathbf{y}) = \underset{\theta\epsilon\Theta}{Sup} \{L_p(\theta;\mathbf{y})\, L_f(\theta;\mathbf{z})\}$$

e

$$FL(\mathbf{z}/\mathbf{y}) = \frac{Sup_{\theta\epsilon\Theta} \{L_p(\theta;\mathbf{y})\, L_f(\theta;\mathbf{z})\}}{Sup_{\theta\epsilon\Theta} L_p(\theta;\mathbf{y})\; Sup_{\theta\epsilon\Theta} L_f(\theta;\mathbf{z})},$$

onde L_p e L_f são as funções de verossimilhança de θ com base em $\mathbf{y}$ e $\mathbf{z}$ respectivamente.

Notamos que $LL(\mathbf{z}/\mathbf{y})$ é a *Função de Verossimilhança Preditiva Profile*, definida no Capítulo 2. Já, $FL(\mathbf{z}/\mathbf{y})$ pode ser encarada como a estatística do teste da razão de verossimilhança para $H_0 : \theta_1 = \theta_2$ contra $H_a : \theta_1 \neq \theta_2$, quando $\mathbf{Y}$ e $\mathbf{Z}$ são independentes com densidades $L_p(\theta_1, \mathbf{y})$ e $L_p(\theta_2, \mathbf{z})$ respectivamente. Assim, como estamos admitindo que $\theta_1 = \theta_2 = \theta$, FL seria uma medida da *plausibilidade* de cada possível valor de $\mathbf{z}$, observados os dados amostrais $\mathbf{y}$. Desta forma, uma vez observado $\mathbf{y}$, o valor de $\mathbf{z}$ que maximiza $FL(\mathbf{z}/\mathbf{y})$ seria o mais compatível com essa hipótese.

LEVY e PERNG (1984) propõe uma pequena modificação em $FL(\mathbf{z}/\mathbf{y})$, resultando na *Função de Verossimilhança Modificada de Fisher*, dada por

$$LL^*(\mathbf{z}/\mathbf{y}) = \frac{Sup_{\theta\epsilon\Theta} L_{pf}(\theta;\mathbf{y},\mathbf{z})}{Sup_{\theta\epsilon\Theta} L_p(\theta;\mathbf{y})},$$

onde

$L_p(\theta;\mathbf{y})$ é a função de verossimilhança de θ baseada nas observações $\mathbf{y}$,

$L_{pf}(\theta;\mathbf{y},\mathbf{z})$ é a função de verossimilhança de θ baseada nas observações $(\mathbf{y}, \mathbf{z})$ e

θ é o parâmetro desconhecido.

Da forma como está definida, esta função também se aplica para o caso em que $\mathbf{Y}$ e $\mathbf{Z}$ não são independentes.

Definindo

$\hat{\theta}_{mv}$ o estimador de máxima verossimilhança de θ baseado nas observações $\mathbf{Y}$ e

$\hat{\theta}^*_{mv}$ o estimador de máxima verossimilhança de θ baseado em $(\mathbf{Y}, \mathbf{Z})$,

então

$$LL^*(\mathbf{z}/\mathbf{y}) = \frac{L_{pf}(\hat{\theta}^*_{mv}; \mathbf{y}, \mathbf{z})}{L_p(\hat{\theta}_{mv}; \mathbf{y})}.$$

Para os modelos lineares já descritos, temos $\hat{\theta}_{mv} = \hat{\theta} = \left(\hat{\beta}, \hat{\sigma}^2\right)$ e $\hat{\theta}^*_{mv} = \hat{\theta}_{\mathbf{z}} = \left(\hat{\beta}_{\mathbf{z}}, \hat{\sigma}^2_{\mathbf{z}}\right)$ de (3.7) e (3.11) respectivamente. Desta forma,

$$\begin{aligned} LL^*(\mathbf{z}/\mathbf{y}) &= LL^*(\mathbf{z}, W/\mathbf{y}, X) \\ &= \frac{L_{pf}(\hat{\theta}_{\mathbf{z}}; \mathbf{y}, \mathbf{z})}{L_p(\hat{\theta}; \mathbf{y})} \\ &= \frac{L_{pf}(\hat{\beta}_{\mathbf{z}}, \hat{\sigma}^2_{\mathbf{z}}; \mathbf{y}, \mathbf{z})}{L_p(\hat{\beta}, \hat{\sigma}^2; \mathbf{y})}, \end{aligned}$$

que, com a notação introduzida em (3.3) e (3.4) fica

$$\begin{aligned} LL^*(\mathbf{z}, W/\mathbf{y}, X) &= \frac{p_{n+m}((\mathbf{y}, \mathbf{z}), (X, W)/\hat{\beta}_{\mathbf{z}}, \hat{\sigma}^2_{\mathbf{z}})}{p_n(\mathbf{y}, X/\hat{\beta}, \hat{\sigma}^2)} \\ &= \frac{(2\pi\hat{\sigma}^2_{\mathbf{z}})^{-\frac{n+m}{2}} \exp\{-(2\hat{\sigma}^2_{\mathbf{z}})^{-1}(\mathbf{v} - H\hat{\beta}_{\mathbf{z}})'(\mathbf{v} - H\hat{\beta}_{\mathbf{z}})\}}{(2\pi\hat{\sigma}^2)^{-\frac{n}{2}} \exp\{-(2\hat{\sigma}^2)^{-1}(\mathbf{y} - X\hat{\beta})'(\mathbf{y} - X\hat{\beta})\}}. \end{aligned}$$

Substituindo $\hat{\sigma}^2$ e $\hat{\sigma}^2_{\mathbf{z}}$ pelas expressões (3.6) e (3.10) respectivamente, segue que

$$\begin{aligned} LL^*(\mathbf{z}, W/\mathbf{y}, X) &= \frac{(2\pi\hat{\sigma}^2_{\mathbf{z}})^{-\frac{n+m}{2}} \exp\left\{-\frac{(n+m)}{2}\frac{(\mathbf{v}-H\hat{\beta}_{\mathbf{z}})'(\mathbf{v}-H\hat{\beta}_{\mathbf{z}})}{(\mathbf{v}-H\hat{\beta}_{\mathbf{z}})'(\mathbf{v}-H\hat{\beta}_{\mathbf{z}})}\right\}}{(2\pi\hat{\sigma}^2)^{-\frac{n}{2}} \exp\left\{-\frac{n}{2}\frac{(\mathbf{y}-X\hat{\beta})'(\mathbf{y}-X\hat{\beta})}{(\mathbf{y}-X\hat{\beta})'(\mathbf{y}-X\hat{\beta})}\right\}} \\ &= \frac{(2\pi\hat{\sigma}^2_{\mathbf{z}})^{-\frac{n+m}{2}} \exp\left\{-\frac{(n+m)}{2}\right\}}{(2\pi\hat{\sigma}^2)^{-\frac{n}{2}} \exp\left\{-\frac{n}{2}\right\}} \\ &= (2\pi)^{-\frac{m}{2}} \exp\{-\frac{m}{2}\}(\hat{\sigma}^2)^{\frac{n}{2}}(\hat{\sigma}^2_{\mathbf{z}})^{-\frac{n+m}{2}}. \end{aligned} \tag{3.13}$$

Substituindo o $\hat{\sigma}_{\mathbf{z}}^2$ restante pela forma dada em (A.3) do apêndice, a *Função de Verossimilhança Preditiva Modificada de Fisher* fica

$$LL^*(\mathbf{z}, W/\mathbf{y}, X) = \frac{(2\pi)^{-\frac{m}{2}} e^{-\frac{m}{2}} (\hat{\sigma}^2)^{\frac{n}{2}}}{\left[(n+m)^{-1}\{ n\hat{\sigma}^2 + (\mathbf{z} - W\hat{\beta})'(I + WM^{-1}W')^{-1}(\mathbf{z} - W\hat{\beta})\}\right]^{\frac{(n+m)}{2}}},$$

onde $M = X'X$.

Desenvolvendo esta última expressão,

$$\begin{aligned} LL^*(\mathbf{z}, W/\mathbf{y}, X) &= \frac{(2\pi)^{-\frac{m}{2}} e^{-\frac{m}{2}} (\hat{\sigma}^2)^{\frac{n}{2}}}{(n+m)^{-\frac{(n+m)}{2}} n^{\frac{(n+m)}{2}} (\hat{\sigma}^2)^{\frac{(n+m)}{2}}} \times \\ &\quad \left[1 + (\mathbf{z} - W\hat{\beta})'[n\hat{\sigma}^2(I + WM^{-1}W')]^{-1}(\mathbf{z} - W\hat{\beta})\right]^{-\frac{n+m}{2}} \\ &= \frac{e^{-\frac{m}{2}} \left(\frac{n+m}{n}\right)^{\frac{n}{2}} \left(\frac{n+m}{2}\right)^{\frac{m}{2}}}{(\pi n)^{\frac{m}{2}} (\hat{\sigma}^2)^{\frac{m}{2}}} \times \\ &\quad \left[1 + (1/n)(\mathbf{z} - W\hat{\beta})'[\hat{\sigma}^2(I + WM^{-1}W')]^{-1}(\mathbf{z} - W\hat{\beta})\right]^{-\frac{n+m}{2}}. \end{aligned} \tag{3.14}$$

Verifica-se que a função de previsão LL^* pode ser padronizada de modo a se tornar uma densidade preditiva, e que a correspondente constante de padronização é

$$\mathcal{K} = e^{\frac{m}{2}} (1 + \frac{m}{n})^{-\frac{n}{2}} \left(\frac{n+m}{2}\right)^{-\frac{m}{2}} \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})[\det(I + WM^{-1}W')]^{\frac{1}{2}}}.$$

Desta forma, a *Densidade Preditiva Modificada de Fisher* é obtida multiplicando (3.14) pela constante de padronização $\mathcal{K}$, resultando em

$$\begin{aligned} t(\mathbf{z}, W/\mathbf{y}, X) &= \frac{\Gamma(\frac{n+m}{2})}{(\pi n)^{\frac{m}{2}} \Gamma(\frac{n}{2})[\det \hat{\sigma}^2(I + WM^{-1}W')]^{\frac{1}{2}}} \times \\ &\quad \left[1 + (1/n)(\mathbf{z} - W\hat{\beta})'[\hat{\sigma}^2(I + WM^{-1}W')]^{-1}(\mathbf{z} - W\hat{\beta})\right]^{-\frac{n+m}{2}}, \end{aligned} \tag{3.15}$$

que coincide com a densidade de uma variável aleatória com distribuição *t-Student* multivariada com parâmetro de locação $W\hat{\beta}$, n graus de liberdade e matriz de correlação $R = \hat{\sigma}^2(I + WM^{-1}W')$ (ver Resultado 7, Seção A.2 do apêndice).

Densidade Preditiva Bayesiana

Do enfoque bayesiano, se $\mathbf{y}$ é o vetor de dados observados e $\mathbf{z}$ o de dados futuros, temos que a *Densidade Preditiva Bayesiana* é dada por

$$f(\mathbf{z}/\mathbf{y}) = \int_{\theta\in\Theta} f(\mathbf{z}/\mathbf{y},\theta) f(\theta/\mathbf{y})\, d\theta.$$

No caso dos modelos lineares (3.1) e (3.2), $f(\mathbf{z}/\mathbf{y},\theta) = f(\mathbf{z}, W/\mathbf{y}, X, \beta, \sigma^2)$ é a densidade condicional de $\mathbf{z}$ dado $\mathbf{y}$ e $f(\theta/\mathbf{y}) = f(\beta, \sigma^2/\mathbf{y}, X)$ é a densidade a posteriori de (β, σ^2) dado $\mathbf{y}$. Desta forma, a *Densidade Preditiva Bayesiana* fica

$$r(\mathbf{z}, W/\mathbf{y}, X) = \int_{\theta\in\Theta} f(\mathbf{z}, W/\mathbf{y}, X, \beta, \sigma^2)\ f(\beta, \sigma^2/\mathbf{y}, X)\ d\beta\, d\sigma^2. \quad (3.16)$$

Por outro lado, a densidade a posteriori de $\theta = (\beta, \sigma^2)$ dado $\mathbf{Y} = \mathbf{y}$ será

$$f(\beta, \sigma^2/\mathbf{y}, X) = \frac{f(\beta, \sigma^2)\, p_n(\mathbf{y}, X/\beta, \sigma^2)}{\int_{\theta\in\Theta} f(\beta, \sigma^2)\, p_n(\mathbf{y}, X/\beta, \sigma^2)\, d\beta\, d\sigma^2},$$

onde $f(\beta, \sigma^2)$ é a densidade a priori para (β, σ^2).

LEVY e PERNG (1984) admitem que $\theta = (\beta, \sigma^2)$ tem uma priori *não informativa*, dada por

$$f(\theta) = f(\beta, \sigma^2) \propto \frac{1}{\sigma^2}.$$

Como as variáveis aleatórias $\mathbf{Y}$ e $\mathbf{Z}$ são independentes dado θ, temos que $f(\mathbf{z}, W/\mathbf{y}, X; \beta, \sigma^2) = f(\mathbf{z}, W/\beta, \sigma^2) = p_m(\mathbf{z}, W/\beta, \sigma^2)$. A correspondente *Densidade Preditiva Bayesiana* para $\mathbf{Z}$ é portanto,

$$\begin{aligned} r(\mathbf{z}, W/\mathbf{y}, X) &= \int_{\theta\in\Theta} p_m(\mathbf{z}, W/\beta, \sigma^2)\, f(\beta, \sigma^2/\mathbf{y}, X)\, d\beta\, d\sigma^2 \\ &= \frac{\int_{\theta\in\Theta} (\sigma^2)^{-1}\ p_m(\mathbf{z}, W/\beta, \sigma^2)\ p_n(\mathbf{y}, X/\beta, \sigma^2)\ d\beta\, d\sigma^2}{\int_{\theta\in\Theta} (\sigma^2)^{-1}\, p_n(\mathbf{y}, X/\beta, \sigma^2)\, d\beta\, d\sigma^2}. \end{aligned}$$

Pela independência de $\mathbf{Z}$ e $\mathbf{Y}$ dado $\theta = (\beta, \sigma^2)$,

$$r(\mathbf{z}, W/\mathbf{y}, X) = \frac{\int_{\theta\in\Theta} (\sigma^2)^{-1}\, p_{n+m}((\mathbf{y}', \mathbf{z}')', (X', W')'/\beta, \sigma^2)\, d\beta\, d\sigma^2}{\int_{\theta\in\Theta} (\sigma^2)^{-1}\, p_n(\mathbf{y}, X/\beta, \sigma^2)\, d\beta\, d\sigma^2}$$

$$= \frac{\int_{\theta\in\Theta} (\sigma^2)^{-1}(2\pi\sigma^2)^{-\frac{n+m}{2}} e^{\left\{-(2\sigma^2)^{-1}[(\mathbf{y}-X\beta)'(\mathbf{y}-X\beta)+(\mathbf{z}-W\beta)'(\mathbf{z}-W\beta)]\right\}}\, d\beta\, d\sigma^2}{\int_{\theta\in\Theta} (\sigma^2)^{-1}(2\pi\sigma^2)^{-\frac{n}{2}} e^{\left\{-(2\sigma^2)^{-1}(\mathbf{y}-X\beta)'(\mathbf{y}-X\beta)\right\}}\, d\beta\, d\sigma^2}$$

$$= \frac{\int_{\theta\in\Theta} (\sigma^2)^{-1}(2\pi\sigma^2)^{-\frac{n+m}{2}} e^{\left\{-(2\sigma^2)^{-1}[(\mathbf{y}-X\beta)', (\mathbf{z}-W\beta)'] \begin{bmatrix} \mathbf{y}-X\beta \\ \mathbf{z}-W\beta \end{bmatrix}\right\}} d\beta d\sigma^2}{\int_{\theta\in\Theta} (\sigma^2)^{-1}(2\pi\sigma^2)^{-\frac{n}{2}} e^{\left\{-(2\sigma^2)^{-1}(\mathbf{y}-X\beta)'(\mathbf{y}-X\beta)\right\}} d\beta d\sigma^2}$$

$$= \frac{\int_{\theta\in\Theta} (\sigma^2)^{-1}(2\pi\sigma^2)^{-\frac{n+m}{2}} e^{\left\{-(2\sigma^2)^{-1} \begin{bmatrix} \mathbf{y}-X\beta \\ \mathbf{z}-W\beta \end{bmatrix}' \begin{bmatrix} \mathbf{y}-X\beta \\ \mathbf{z}-W\beta \end{bmatrix}\right\}} d\beta\, d\sigma^2}{\int_{\theta\in\Theta} (\sigma^2)^{-1}(2\pi\sigma^2)^{-\frac{n}{2}} e^{\left\{-(2\sigma^2)^{-1}(\mathbf{y}-X\beta)'(\mathbf{y}-X\beta)\right\}}\, d\beta\, d\sigma^2}$$

$$= \frac{\int_{\theta\in\Theta} (\sigma^2)^{-1}(2\pi\sigma^2)^{-\frac{n+m}{2}} e^{\left\{-(2\sigma^2)^{-1}(\mathbf{v}-H\beta)'(\mathbf{v}-H\beta)\right\}}\, d\beta\, d\sigma^2}{\int_{\theta\in\Theta} (\sigma^2)^{-1}(2\pi\sigma^2)^{-\frac{n}{2}} e^{\left\{-(2\sigma^2)^{-1}(\mathbf{y}-X\beta)'(\mathbf{y}-X\beta)\right\}}\, d\beta\, d\sigma^2}. \qquad (3.17)$$

Na Seção A.2 do apêndice, calculamos as integrais do numerador e do denominador de (3.17), designados respectivamente por $I(n)$ e $I(d)$. Assim, de (A.11) e (A.13),

$$I(n) = (\pi)^{\frac{p-(n+m)}{2}} \Gamma(\frac{n+m-p}{2})\, |M|^{-\frac{1}{2}} |I + WM^{-1}W'|^{-\frac{1}{2}} (n\hat{\sigma}^2)^{-\frac{n+m-p}{2}} \times$$
$$\left[1 + (\mathbf{z} - W\hat{\beta})'[n\hat{\sigma}^2(I + WM^{-1}W')]^{-1}(\mathbf{z} - W\hat{\beta})\right]^{-\frac{n+m-p}{2}}$$

e

$$I(d) = (\pi)^{\frac{p-n}{2}} \Gamma(\frac{n-p}{2})\, |M|^{-\frac{1}{2}} (n\hat{\sigma}^2)^{-\frac{n-p}{2}}.$$

Dividindo as duas últimas expressões e substituindo em (3.17) (ver (A.14), Seção A.2 do apêndice), a *Densidade Preditiva Bayesiana* para $\mathbf{Z}$ resulta em

$$r(\mathbf{z}, W/\mathbf{y}, X) = \frac{\Gamma(\frac{n+m-p}{2})}{[\pi(n-p)]^{\frac{m}{2}}\, \Gamma(\frac{n-p}{2})} \left| n\hat{\sigma}^2 \frac{(I + WM^{-1}W')}{(n-p)} \right|^{-\frac{1}{2}}$$

$$\times \left[1 + \frac{1}{(n-p)}(\mathbf{z} - W\hat{\beta})'[n\hat{\sigma}^2\frac{(I + WM^{-1}W')}{(n-p)}]^{-1}(\mathbf{z} - W\hat{\beta})\right]^{-\frac{n+m-p}{2}}. \tag{3.18}$$

Esta densidade coincide com a densidade de uma variável aleatória com distribuição *t-Student* multivariada com parâmetro de locação $W\hat{\beta}$, $(n-p)$ graus de liberdade e matriz de correlação $R = [n\hat{\sigma}^2(I + WM^{-1}W')]/(n-p)$ (Resultado 7, Seção A.2 do apêndice).

Densidade Preditiva Não Viciada de Mínima Variância

Para o modelo de regressão linear em que a variável resposta tem distribuição normal, O'REILLY (1976) fornece uma condição necessária e suficiente para a existência de um estimador não viciado uniformemente de variância mínima para a função densidade do vetor de observações futuras $\mathbf{Z}$, associado à matriz de variáveis independentes W.

No caso particular em que o modelo de regressão tem posto completo (X é de posto completo), que é o caso do modelo (3.1), esta condição exige que o espaço linear gerado pelas linhas de W seja um subespaço do espaço linear gerado pelas linhas de X e também que $(I - W(X'X)^{-1}W')$ seja positiva definida e de posto menor ou igual a $n-p$.

Nestas condições, o estimador não viciado de variância mínima de $p_m(\mathbf{z}, W/\beta, \sigma^2)$, que é uma possível densidade preditiva para $\mathbf{Z}$, é dado por

$$\hat{p}_m(\mathbf{z}, W/\beta, \sigma^2) =$$

$$= \begin{cases} \dfrac{\Gamma(\frac{n-p}{2})}{[\pi(n-p)]^{\frac{m}{2}}\Gamma(\frac{n-m-p}{2})} \dfrac{\left|\hat{\sigma}^2(I-W(X'X)^{-1}W')\right|^{-\frac{1}{2}}}{\left[1-(\mathbf{z}-W\hat{\beta})'[(n-p)\hat{\sigma}^2(I-W(X'X)^{-1}W')]^{-1}(\mathbf{z}-W\hat{\beta})\right]^{-\frac{n-p-m-2}{2}}}, \\ \qquad\qquad se \quad (\mathbf{z}-W\hat{\beta})'[(n-p)\hat{\sigma}^2(I-W(X'X)^{-1}W')]^{-1}(\mathbf{z}-W\hat{\beta}) < 1\ ; \\ 0 \qquad\qquad caso\ contrário. \end{cases} \tag{3.19}$$

Salientamos que essa densidade não foi obtida através dos processos descritos no Capítulo 2, mas que tem importância como estimador de $p_m(\mathbf{z}, W/\beta, \sigma^2)$, assunto que será discutido na próxima seção.

Verifica-se ainda que $\hat{p}_m(\mathbf{z}, W/\beta,\sigma^2)$ pertence à classe de densidades preditivas proposta por LEVY E PERNG (1986), que definiremos na Seção 3.4.

Na próxima seção, estudaremos aspectos relativos à consistência de algumas das densidades apresentadas na presente seção.

3.3 Aspectos relativos à consistência das Densidades Preditivas

Conforme comentado na seção anterior, uma possível utilização de densidades preditivas é na estimação da função densidade associada a observações futuras. Neste caso, uma importante propriedade a ser analisada é a da consistência.

A Definição 3.3.1 introduz o conceito de consistência de uma densidade preditiva quando utilizada com o objetivo de estimar a verdadeira densidade das observações futuras.

Algumas definições adicionais e resultados auxiliares são apresentados no apêndice, Seção A.3.

Definição 3.3.1 *Consideremos os modelos (3.1), (3.2) e* $p_m(\mathbf{z}, W/\beta, \sigma^2)$, *a densidade do vetor aleatório* $\mathbf{Z}$. *A densidade preditiva* $L(\mathbf{z}, W/\mathbf{y}, X)$ *é consistente para* $p_m(\mathbf{z}, W/\beta, \sigma^2)$ *se*

$$L(\mathbf{z}, W/\mathbf{Y}, X) \underset{n \to \infty}{\overset{P}{\longrightarrow}} p_m(\mathbf{z}, W/\beta, \sigma^2).$$

Dada esta definição, discutiremos a seguir a consistência de algumas densidades preditivas da seção anterior. O Lema 3.3.1, cuja prova se encontra no Lema A.3.1, Seção A.3 do apêndice, será útil nesse aspecto.

Lema 3.3.1 *(LEVY e PERNG, (1984))*
Sejam $\mathbf{Y}$, X, $\mathbf{Z}$ *e* W, *vetores e matrizes definidos anteriormente e vamos supor que* $(X'X)/n \to D$ *quando* $n \to \infty$, D *positiva definida. Se* $\hat{\beta}$, $\hat{\sigma}^2$, $\hat{\beta}_{\mathbf{z}}$ *e* $\hat{\sigma}^2_{\mathbf{z}}$ *são os estimadores definidos em (3.5), (3.6), (3.9) e (3.10)*

respectivamente, então,

(i) $\hat{\beta} - \beta = O_p(n^{-\frac{1}{2}})$

(ii) $\hat{\sigma}^2 - \sigma^2 = O_p(n^{-\frac{1}{2}})$

(iii) $\hat{\beta}_z - \hat{\beta} = O_p(n^{-1})$

(iv) $\hat{\sigma}_z^2 - \hat{\sigma}^2 = O_p(n^{-1})$.

Teorema 3.3.1 *(FULLER (1976), [pg. 192])*
Seja $\mathbf{X}_n = (X_{1n}, X_{2n}, \cdots, X_{kn})'$, *um vetor aleatório* k*-dimensional,* $\mathbf{a} = (a_1, a_2, \cdots, a_k)'$ *um vetor de constantes,* r_n *uma função de* n, $n \in \mathbb{N}$, *tal que* $r_n \to 0$ *para* $n \to \infty$ *e*

$$\mathbf{X}_n = \mathbf{a} + O_p(r_n).$$

Se $g(\mathbf{x})$ *é uma função real definida no espaço euclidiano* k*-dimensional, com derivadas parciais contínuas de ordem* 2 *em* $\mathbf{x} = \mathbf{a}$, *então*

$$\begin{aligned} g(\mathbf{X}_n) &= g(\mathbf{a}) + \sum_{j=1}^{k} \frac{\partial g(\mathbf{a})}{\partial x_j}(X_{jn} - a_j) \\ &\quad + \sum_{j=1}^{k}\sum_{i=1}^{k} \frac{1}{2!}\frac{\partial^2 g(\mathbf{a})}{\partial x_j \partial x_i}(X_{jn} - a_j)(X_{in} - a_i) \\ &\quad + O_p(r_n^3) \end{aligned}$$

onde

$\partial g(\mathbf{a})/\partial x_j$	*é a derivada parcial de* $g(\mathbf{x})$ *com respeito a* x_j *avaliada em* $\mathbf{x} = \mathbf{a}$ *e*
$\partial^2 g(\mathbf{a})/\partial x_j \partial x_i$	*é a derivada parcial de* $g(\mathbf{x})$ *com respeito a* x_j *e* x_i *avaliada em* $\mathbf{x} = \mathbf{a}$.

Utilizando este resultado, podemos provar a consistência das densidades *Pseudo Preditiva* e *Preditiva Modificada de Fisher.*

Consistência da Densidade Preditiva Clássica ou Pseudo Preditiva

A prova da consistência da *Densidade Preditiva Clássica* é direta. Do Lema 3.3.1, temos que $\hat{\beta} - \beta = O_p(n^{-\frac{1}{2}})$ e $\hat{\sigma}^2 - \sigma^2 = O_p(n^{-\frac{1}{2}})$ e com isso, pelo Teorema A.3.3

$$\begin{pmatrix} \hat{\beta} \\ \hat{\sigma}^2 \end{pmatrix} - \begin{pmatrix} \beta \\ \sigma^2 \end{pmatrix} = O_p(n^{-\frac{1}{2}}).$$

Aplicando o Teorema 3.3.1 para $\mathbf{X}_n = (\hat{\beta}', \hat{\sigma}^2)'$, $\mathbf{a} = (\beta', \sigma^2)'$ e $g(\mathbf{X}_n) = p_m(\mathbf{z}, W/\hat{\beta}, \hat{\sigma}^2)$, segue que

$$\begin{aligned} p_m(\mathbf{z}, W/\hat{\beta}, \hat{\sigma}^2) &= p_m(\mathbf{z}, W/\beta, \sigma^2) + \sum_{j=1}^{2} \frac{\partial p_m(\mathbf{z}, W/\beta, \sigma^2)}{\partial X_{jn}}(X_{jn} - a_j) \\ &\quad + \sum_{j=1}^{2}\sum_{i=1}^{2} \frac{1}{2!} \frac{\partial^2 p_m(\mathbf{z}, W/\beta, \sigma^2)}{\partial X_{jn} \partial X_{in}}(X_{jn} - a_j)(X_{in} - a_i) \\ &\quad + O_p((n^{-\frac{1}{2}})^3) \end{aligned}$$

onde X_{jn} é a j-ésima componente do vetor $\mathbf{X}_n$ e a_j a j-ésima componente do vetor $\mathbf{a}$.

Portanto

$$\begin{aligned} p_m(\mathbf{z}, W/\hat{\beta}, \hat{\sigma}^2) &= p_m(\mathbf{z}, W/\beta, \sigma^2) + O_p(n^{-\frac{1}{2}}) + O_p(n^{-1}) + O_p(n^{-\frac{3}{2}}) \\ &= p_m(\mathbf{z}, W/\beta, \sigma^2) + O_p(n^{-\frac{1}{2}}), \end{aligned} \tag{3.20}$$

o que implica que $q(\mathbf{z}, W/\mathbf{Y}, X) = p_m(\mathbf{z}, W/\hat{\beta}, \hat{\sigma}^2)$ é um estimador consistentes de $p_m(\mathbf{z}, W/\beta, \sigma^2)$.

Consistência da Densidade Preditiva Modificada de Fisher

O seguinte teorema, cuja prova é dada no Teorema A.3.6 do apêndice, permitirá a demonstração da consistência da *Densidade Preditiva Modificada de Fisher.*

Teorema 3.3.2 *(LEVY e PERNG (1984))*
Sejam $\mathbf{Y}$, X, $\mathbf{Z}$ *e* W, *vetores e matrizes definidos anteriormente e vamos supor que* $(X'X)/n \to D$ *quando* $n \to \infty$, D *positiva definida. Nestas condições,*

$$LL^*(\mathbf{z}, W/\mathbf{Y}, X) = q(\mathbf{z}, W/\mathbf{Y}, X)\,(1 + \mathsf{O}_p(n^{-1})).$$

O próximo teorema mostra a consistência da *Densidade Preditiva Modificada de Fisher.*

Teorema 3.3.3 *(LEVY e PERNG (1984))*
Sob as hipóteses do teorema anterior,

$$LL^*(\mathbf{z}, W/\mathbf{Y}, X) = p_m(\mathbf{z}, W/\beta, \sigma^2) + \mathsf{O}_p(n^{-\frac{1}{2}}).$$

Prova: Temos que

$$q(\mathbf{z}, W/\mathbf{Y}, X) = p_m(\mathbf{z}, W/\hat{\beta}, \hat{\sigma}^2),$$

e assim, pelo teorema anterior,

$$LL^*(\mathbf{z}, W/\mathbf{Y}, X) = p_m(\mathbf{z}, W/\hat{\beta}, \hat{\sigma}^2)(1 + \mathsf{O}_p(n^{-1})).$$

Substituindo $p_m(\mathbf{z}, W/\hat{\beta}, \hat{\sigma}^2)$ pela expressão dada em (3.20),

$$\begin{aligned} LL^*(\mathbf{z}, W/\mathbf{Y}, X) &= \{p_m(\mathbf{z}, W/\beta, \sigma^2) + \mathsf{O}_p(n^{-\frac{1}{2}})\}\,(1 + \mathsf{O}_p(n^{-1})) \\ &= p_m(\mathbf{z}, W/\beta, \sigma^2) + \mathsf{O}_p(n^{-1})\,p_m(\mathbf{z}, W/\beta, \sigma^2) \\ &\quad + \mathsf{O}_p(n^{-\frac{1}{2}}) + \mathsf{O}_p(n^{-\frac{1}{2}})\,\mathsf{O}_p(n^{-1}) \\ &= p_m(\mathbf{z}, W/\beta, \sigma^2) + \mathsf{O}_p(n^{-\frac{1}{2}}). \qquad \square \end{aligned}$$

Como $p_m(\mathbf{z}, W/\beta, \sigma^2)$ é a densidade de $\mathbf{Z}$, desconhecida, foi provado que $LL^*(\mathbf{z}, W/\mathbf{Y}, X) \underset{n\to\infty}{\xrightarrow{P}} p_m(\mathbf{z}, W/\beta, \sigma^2)$, sendo portanto um estimador consistente de $p_m(\mathbf{z}, W/\beta, \sigma^2)$.

A *Densidade Preditiva Modificada de Fisher* foi proposta e analisada com detalhes por LEVY e PERNG (1984). Neste artigo, é provada ainda sua consistência forte, ou seja, a convergência quase certa de LL^* para $p_m(\mathbf{z}, W/\beta, \sigma^2)$.

Num artigo posterior, os autores dão maior destaque à *Densidade Preditiva Bayesiana*, provando sua otimalidade numa particular classe de densidades preditivas. Este estudo será abordado na próxima seção.

3.4 Densidade Preditiva Ótima

Sob o modelo (3.1), como o objetivo de previsão dos dados futuros $\mathbf{Z}$, LEVY e PERNG (1986) definiram uma classe de densidades preditivas Ψ, que contem as quatro densidades estudadas nas seções anteriores.

Lembrando que uma possível utilização de densidades preditivas é na estimação da densidade de $\mathbf{Z}$, denotada por $p_m(\mathbf{z}, W/\beta, \sigma^2)$, os autores obtêm a densidade preditiva ótima dentro de Ψ, definida como aquela que é a mais próxima de $p_m(\mathbf{z}, W/\beta, \sigma^2)$, no sentido de minimizar a medida de divergência de Kullback-Leibler.

Apresentamos a seguir as definições formais dos conceitos que acabamos de apresentar e posteriormente a densidade preditiva ótima de $\mathbf{Z}$.

Construção da classe de Densidades Preditivas Ψ

Ao invés de considerar todas as possíveis densidades preditivas de $\mathbf{Z}$, LEVY e PERNG (1986) restringem a atenção à classe de densidades preditivas Ψ.

Se W, $\hat{\beta}$ e $\hat{\sigma}^2$ são os elementos definidos respectivamente em (3.2), (3.5) e (3.6), consideremos a estatística

$$\mathbf{t} = \mathbf{t}(\mathbf{y}, \mathbf{z}) = \frac{(\mathbf{z} - W\hat{\beta})}{(n\hat{\sigma}^2)^{\frac{1}{2}}}. \tag{3.21}$$

A classe Ψ é definida como a coleção de todas as densidades preditivas de $\mathbf{Z}$ que são funções de $\mathbf{t}$, ou seja,

$$\Psi = \{s(\mathbf{z}, W/\mathbf{y}, X) : s(\mathbf{z}, W/\mathbf{y}, X) = g(\mathbf{t}) = g(\mathbf{t}(\mathbf{y}, \mathbf{z}))\}.$$

Verifica-se facilmente que a classe Ψ contem as densidades preditivas definidas nas seções anteriores. Além disso, sua construção permite obter de forma relativamente rápida a densidade preditiva ótima. O critério de otimalidade utilizado é definido a seguir.

Critério de Kullback-Leibler

AITCHISON (1975) e LARIMORE (1983) propuseram o uso da medida de informação de Kullback-Leibler (KULLBACK e LEIBLER (1951)) como

uma forma natural de quantificar a proximidade de duas densidades. Tal medida foi utilizada para verificar a qualidade de uma particular densidade preditiva como estimador de $p_m(\mathbf{z}, W/\beta, \sigma^2)$.

Neste contexto, se $s(\mathbf{z}, W/\mathbf{y}, X)$ é a densidade preditiva de $\mathbf{Z}$, LEVY e PERNG (1986) definem a divergência de s com relação a p_m por

$$\begin{aligned} D_{\beta,\sigma^2}(p_m, s) &= E_{\mathbf{Y},\mathbf{Z}}\left[\log \frac{p_m(\mathbf{Z}, W/\beta, \sigma^2)}{s(\mathbf{Z}, W/\mathbf{Y}, X)}\right] \\ &= \int_{\mathbb{R}^n} p_n(\mathbf{y}, X/\beta, \sigma^2) \int_{\mathbb{R}^m} p_m(\mathbf{z}, W/\beta, \sigma^2)\, log\Big\{\frac{p_m(\mathbf{z}, W/\beta, \sigma^2)}{s(\mathbf{z}, W/\mathbf{y}, X)}\Big\}\, d\mathbf{z}\, d\mathbf{y}. \end{aligned}$$

Observamos que $D_{\beta,\sigma^2}(p_m, s)$ será tão menor quanto mais próxima estiver s de p_m.

Dada essa definição, passaremos à determinação da densidade preditiva ótima.

Cálculo da Densidade Preditiva Ótima

O problema de obter a densidade preditiva ótima dentre as pertencentes à classe de densidades Ψ consiste em minimizar D_{β,σ^2} da definição anterior. Podemos desenvolver essa medida, obtendo

$$E_{\mathbf{Y},\mathbf{Z}}\left[\log\{\frac{p_m(\mathbf{Z}, W/\beta, \sigma^2)}{s(\mathbf{Z}, W/\mathbf{Y}, X)}\}\right] =$$

$$\begin{aligned} &= E_{\mathbf{Y},\mathbf{Z}}\left[\log\{\frac{(2\pi\sigma^2)^{-\frac{m}{2}} \exp\{-(2\sigma^2)^{-1}(\mathbf{Z}-W\beta)'(\mathbf{Z}-W\beta)\}}{s(\mathbf{Z}, W/\mathbf{Y}, X)}\}\right] \\ &= E_{\mathbf{Y},\mathbf{Z}}\left[\frac{m}{2}\log(\frac{1}{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(\mathbf{Z}-W\beta)'(\mathbf{Z}-W\beta)\right] - E_{\mathbf{Y},\mathbf{Z}}\left[\log\, s(\mathbf{Z}, W/\mathbf{Y}, X)\right] \\ &= \frac{m}{2}\log(\frac{1}{2\pi\sigma^2}) - \frac{1}{2}E_{\mathbf{Y},\mathbf{Z}}\Big\{\frac{(\mathbf{Z}-W\beta)'(\mathbf{Z}-W\beta)}{\sigma^2}\Big\} - E_{\mathbf{Y},\mathbf{Z}}\left[\log\, s(\mathbf{Z}, W/\mathbf{Y}, X)\right] \\ &= \frac{m}{2}\log(\frac{1}{2\pi\sigma^2}) - \frac{m}{2} - E_{\mathbf{Y},\mathbf{Z}}\left[\log\, s(\mathbf{Z}, W/\mathbf{Y}, X)\right]. \end{aligned} \tag{3.22}$$

Como para todo $s \in \Psi$, $s(\mathbf{z}, W/\mathbf{y}, X) = g(\mathbf{t})$, o problema equivale a maximizar $E_{\mathbf{t}}[\log g(\mathbf{t})]$.

Verifica-se que $\mathbf{t} = (\mathbf{z} - W\hat{\beta})/[n\hat{\sigma}^2]^{\frac{1}{2}}$ tem distribuição *t-Student* multivariada, ou seja,

$$\frac{(\mathbf{z} - W\hat{\beta})}{(n\hat{\sigma}^2)^{\frac{1}{2}}} \sim St_m(\mathbf{0},\, n-p,\, A/(n-p)),$$

sendo $A = I + W(X'X)^{-1}W'$.

Devemos então maximizar

$$E_{\mathbf{t}}[\log g(\mathbf{t})] = k \int_{\mathbb{R}^m} (1 + \mathbf{t}'A^{-1}\mathbf{t})^{-\frac{(m+n-p)}{2}} \log g(\mathbf{t})\, d\mathbf{t},$$

onde

$$k = \frac{\Gamma\left(\frac{m+n-p}{2}\right)}{((n-p)\pi)^{\frac{m}{2}}\Gamma\left(\frac{n-p}{2}\right)\left|\frac{A}{(n-p)}\right|^{\frac{1}{2}}}.$$

Este procedimento é equivalente a maximizar o funcional

$$J[g] = \int_{\mathbb{R}^m} (1 + \mathbf{t}'A^{-1}\mathbf{t})^{-\frac{(m+n-p)}{2}} \log g(\mathbf{t}) d\mathbf{t},$$

sujeito à restrição

$$K[g] = \int_{\mathbb{R}^m} g(\mathbf{t}) d\mathbf{z} = 1. \tag{3.23}$$

Temos que

$$\mathbf{t} = \frac{(\mathbf{z} - W\hat{\beta})}{(n\hat{\sigma}^2)^{\frac{1}{2}}} = (n\hat{\sigma}^2)^{-\frac{1}{2}}\mathbf{z} - (n\hat{\sigma}^2)^{-\frac{1}{2}}W\hat{\beta}$$

e com isso,

$$d\mathbf{t} = d\left((n\hat{\sigma}^2)^{-\frac{1}{2}}\mathbf{z}\right) = (n\hat{\sigma}^2)^{-\frac{m}{2}} d\mathbf{z},$$

o que implica em

$$d\mathbf{z} = (n\hat{\sigma}^2)^{\frac{m}{2}} d\mathbf{t}. \tag{3.24}$$

Substituindo (3.24) em (3.23), a restrição fica

$$K[g] = \int_{\mathbb{R}^m} (n\hat{\sigma}^2)^{\frac{m}{2}} g(\mathbf{t}) d\mathbf{t} = 1. \tag{3.25}$$

Apresentamos a seguir o teorema dado em LEVY e PERNG (1986), que exibe a densidade preditiva ótima de $\mathbf{Z}$, detalhando um pouco mais sua demonstração.

Teorema 3.4.1 *(LEVY e PERNG (1986))*
Sejam Ψ, $\mathbf{t}$ *e* $D_{\beta,\sigma^2}(p_m, s)$ *como definidas previamente. Nestas condições, a densidade preditiva*

$$g^*(\mathbf{z}, W/\mathbf{y}, X) = g^*(\mathbf{t}(\mathbf{y}, \mathbf{z})) = st_m(W\hat{\beta},\, n-p,\, n\hat{\sigma}^2 A/(n-p))$$

é o único mínimo de $D_{\beta,\sigma^2}(p_m, s)$ *dentre todas as densidades* s *em* Ψ, *para todo* $\hat{\beta}$ *e* $\hat{\sigma}^2$.

Prova: Conforme visto anteriormente, devemos maximizar

$$J[g] = \int_{\mathbb{R}^m} F(\mathbf{t}, g(\mathbf{t}))d\mathbf{t}$$

sujeito à condição

$$K[g] = \int_{\mathbb{R}^m} G(\mathbf{t}, g)d\mathbf{t} = 1,$$

onde

$$F(\mathbf{t}, g) = (1 + \mathbf{t}'A^{-1}\mathbf{t})^{-\frac{(m+n-p)}{2}} \log g(\mathbf{t}) \tag{3.26}$$

e

$$G(\mathbf{t}, g) = (n\hat{\sigma}^2)^{\frac{m}{2}} g(\mathbf{t}). \tag{3.27}$$

Essa maximização de $J[g]$, sujeita à condição $K[g] = 1$, é um problema isoperimétrico de cálculo variacional (GELFAND (1963)).

Verifica-se que se $\bar{g}$ maximiza J sujeito à $K[g] = 1$, então existe $\lambda \in \mathbb{R}^m$, multiplicador de Lagrange, tal que $\bar{g}$ é o ponto crítico do funcional

$$L[g] = \int_{\mathbb{R}^m} [F(\mathbf{t}, g(\mathbf{t})) - \lambda G(\mathbf{t}, g(\mathbf{t}))]d\mathbf{t}.$$

Verifica-se também que o ponto crítico desse funcional é obtido quando a derivada direcional (ver (A.18), Seção A.4 do apêndice) de L na direção $h \in \mathcal{C}^1(\mathbb{R}^m)$ é igual a zero ($\bar{g}$ e $h(\mathbf{t})$ são unidimensionais e $\mathcal{C}^1(\mathbb{R}^m)$ é a classe das funções contínuas com primeira derivada contínua, com domínio em $\mathbb{R}^m$). Desta forma, devemos determinar $\bar{g}$ tal que

$$\frac{\partial L(\bar{g})}{\partial h} = \lim_{t\to 0}\frac{L(\bar{g}+th(\mathrm{t}))-L(\bar{g})}{t} = 0 \qquad \forall\, h\in\mathcal{C}^1(\mathbb{R}^m). \tag{3.28}$$

De (A.18) do apêndice,

$$\frac{\partial L(\bar{g})}{\partial h} =$$

$$= \lim_{t\to 0}\frac{\int_{\mathbb{R}^m}[F(\mathrm{t},\bar{g}(\mathrm{t})+th(\mathrm{t}))-\lambda G(\mathrm{t},\bar{g}(\mathrm{t})+th(\mathrm{t}))]d\mathrm{t} - \int_{\mathbb{R}^p}[F(\mathrm{t},\bar{g}(\mathrm{t}))-\lambda G(\mathrm{t},\bar{g}(\mathrm{t}))]d\mathrm{t}}{t}$$

$$= \lim_{t\to 0}\int_{\mathbb{R}^m}\frac{F(\mathrm{t},\bar{g}(\mathrm{t})+th(\mathrm{t}))-\lambda G(\mathrm{t},\bar{g}(\mathrm{t})+th(\mathrm{t})) - F(\mathrm{t},\bar{g}(\mathrm{t}))+\lambda G(\mathrm{t},\bar{g}(\mathrm{t}))}{t}d\mathrm{t}$$

$$= \lim_{t\to 0}\int_{\mathbb{R}^m}\left[\frac{F(\mathrm{t},\bar{g}(\mathrm{t})+th(\mathrm{t}))-F(\mathrm{t},\bar{g}(\mathrm{t}))}{t} - \lambda\frac{G(\mathrm{t},\bar{g}(\mathrm{t})+th(\mathrm{t}))-G(\mathrm{t},\bar{g}(\mathrm{t}))}{t}\right]d\mathrm{t}.$$

Admitindo-se que $\left[\frac{F(\mathrm{t},\bar{g}(\mathrm{t})+th(\mathrm{t}))-F(\mathrm{t},\bar{g}(\mathrm{t}))}{t} - \lambda\frac{G(\mathrm{t},\bar{g}(\mathrm{t})+th(\mathrm{t}))-G(\mathrm{t},\bar{g}(\mathrm{t}))}{t}\right]$ é integrável, o que deve ser averiguado posteriormente, segue que

$$\frac{\partial L(\bar{g})}{\partial h} =$$

$$\int_{\mathbb{R}^m}\lim_{t\to 0}\left[\frac{F(\mathrm{t},\bar{g}(\mathrm{t})+th(\mathrm{t}))-F(\mathrm{t},\bar{g}(\mathrm{t}))}{t} - \lambda\frac{G(\mathrm{t},\bar{g}(\mathrm{t})+th(\mathrm{t}))-G(\mathrm{t},\bar{g}(\mathrm{t}))}{t}\right]d\mathrm{t}.$$

Devido à derivação feita em (A.21), esta expressão se reduz a

$$\frac{\partial L(\bar{g})}{\partial h} = \int_{\mathbb{R}^m}\frac{\partial F}{\partial g}(\mathrm{t},\bar{g}(\mathrm{t})).h(\mathrm{t})\,d\mathrm{t} - \lambda\int_{\mathbb{R}^m}\frac{\partial G}{\partial g}(\mathrm{t},\bar{g}(\mathrm{t})).h(\mathrm{t})\,d\mathrm{t}.$$

Lembrando que $\bar{g}$ é tal que $\partial L(\bar{g})/\partial h = 0$, então

$$\frac{\partial L(\bar{g})}{\partial h} = \int_{\mathbb{R}^m}\left[\frac{\partial F}{\partial g}(\mathrm{t},\bar{g}(\mathrm{t})) - \lambda\frac{\partial G}{\partial g}(\mathrm{t},\bar{g}(\mathrm{t}))\right].h(\mathrm{t})\,d\mathrm{t} = 0 \qquad \forall\, h(\mathrm{t})\in\mathcal{C}^1(\mathbb{R}^m),$$

onde $\mathcal{C}$ é a classe das funções contínuas em $\mathbb{R}^m$, do que segue que

$$\frac{\partial F}{\partial g}(\mathrm{t},\bar{g}(\mathrm{t})) - \lambda\frac{\partial G}{\partial g}(\mathrm{t},\bar{g}(\mathrm{t})) = 0 \qquad \forall\, \mathrm{t}\in\mathbb{R}^m.$$

As derivadas parciais das funções em (3.26) e (3.27) com respeito a g, avaliadas em $g = \bar{g}$ são

$$\frac{\partial}{\partial g} F(\mathbf{t}, \bar{g}) = \frac{(1 + \mathbf{t}' A^{-1} \mathbf{t})^{-\frac{(m+n-p)}{2}}}{\bar{g}} \tag{3.29}$$

e

$$\frac{\partial}{\partial g} G(\mathbf{t}, \bar{g}) = (n\hat{\sigma}^2)^{\frac{m}{2}}. \tag{3.30}$$

Logo,

$$\frac{(1 + \mathbf{t}' A^{-1} \mathbf{t})^{-\frac{(m+n-p)}{2}}}{\bar{g}} - \lambda (n\hat{\sigma}^2)^{\frac{m}{2}} = 0$$

e finalmente

$$\bar{g} = \frac{(1 + \mathbf{t}' A^{-1} \mathbf{t})^{-\frac{(m+n-p)}{2}}}{\lambda (n\hat{\sigma}^2)^{\frac{m}{2}}},$$

onde o valor de λ pode ser calculado da restrição $\int_{\mathbb{R}^m} (n\hat{\sigma}^2)^{\frac{m}{2}} \bar{g}(\mathbf{t})\, d\mathbf{t} = 1$. Após cálculos temos que

$$\lambda = \frac{\pi^{\frac{m}{2}} \Gamma\left(\frac{(n-p)}{2}\right) |A|^{\frac{1}{2}}}{\Gamma\left(\frac{(m+n-p)}{2}\right)}$$

e

$$\bar{g}(\mathbf{t}) = \frac{\Gamma\left(\frac{(m+n-p)}{2}\right)}{\Gamma\left(\frac{(n-p)}{2}\right) \pi^{\frac{m}{2}} |A|^{\frac{1}{2}} (n\hat{\sigma}^2)^{\frac{m}{2}}} (1 + \mathbf{t}' A^{-1} \mathbf{t})^{-\frac{(m+n-p)}{2}}.$$

Verifica-se que a função $\left[\frac{F(\mathbf{t}, \bar{g}(\mathbf{t}) + th(\mathbf{t})) - F(\mathbf{t}, \bar{g}(\mathbf{t}))}{t} - \lambda \frac{G(\mathbf{t}, \bar{g}(\mathbf{t}) + th(\mathbf{t})) - G(\mathbf{t}, \bar{g}(\mathbf{t}))}{t}\right]$, para $\bar{g}(\mathbf{t})$ obtido, é majorada por uma função integrável, sendo portanto integrável.

Verifica-se ainda que $\bar{g}$ fornece um único máximo para $J[g]$ sob a restrição em (3.25).

Como consequência, a densidade preditiva ótima de $\mathbf{Z}$ é

$$\begin{aligned}\bar{g}(\mathbf{t}(\mathbf{z})) &= g^*(\mathbf{z}, W/\mathbf{y}, X) \\ &= \frac{\Gamma\left(\frac{(m+n-p)}{2}\right)}{\Gamma\left(\frac{(n-p)}{2}\right)\pi^{\frac{m}{2}}|A|^{\frac{1}{2}}(n\hat{\sigma}^2)^{\frac{m}{2}}}(1+\frac{1}{n-p}(\mathbf{z}-W\hat{\beta})'[\frac{n\hat{\sigma}^2 A}{n-p}]^{-1}(\mathbf{z}-W\hat{\beta}))^{-\frac{(m+n-p)}{2}}\end{aligned}$$

e resulta em $St_m(W\hat{\beta},\ n-p,\ n\hat{\sigma}^2 A/(n-p))$.

Esta densidade, que minimiza a medida de Kullback-Leibler, é a *Densidade Preditiva Bayesiana* (3.18). Portanto concluimos que tal densidade é ótima na classe Ψ, que contem todas as densidade de interesse já descritas.

Capítulo 4

Densidades Preditivas na determinação de pontos influentes no modelo de regressão linear

4.1 Introdução

Diagnóstico em regressão é um tópico bastante estudado. Na literatura, existem muitas medidas de diagnóstico que basicamente são usadas para comparar o ajuste do modelo considerado na ausência e presença de um ou vários elementos amostrais. É de nosso interesse em especial descrever o uso de densidades preditivas para detectar uma ou mais observações influentes num modelo linear normal. Serão avaliadas as densidades preditivas na presença e ausência de observações e a discrepância entre elas será quantificada através da medida de divergência de Kullback-Leibler (KULLBACK e LEIBLER (1951)).

Iniciamos este capítulo apresentando os modelos lineares com os quais iremos trabalhar e fornecendo as notações necessárias para denotar as observações retiradas e o modelo sem estas observações. Logo após, introduzimos a densidade preditiva que será utilizada e as definições de *Funções de Influência Preditivas* baseadas nas medidas de divergência de Kullback-Leibler. Para o caso da previsão de um vetor de observações, as *Funções de Influência Preditivas* são desenvolvidas de forma detalhada de modo a se conseguir uma clara interpretação da influência das observações.

Em particular, para a previsão de um vetor aleatório no modelo linear normal em que $\mathbf{Y} \sim \mathcal{N}(X\beta, \theta I)$, a densidade utilizada será a densidade preditiva bayesiana para a priori $f(\beta, \theta) = \theta^{-1}$. Quando o parâmetro θ é conhecido, a densidade preditiva torna-se uma normal multivariada, e, com base nela, calcularemos as *Funções de Influência Preditivas* que descreveremos na Seção 4.3. Se o parâmetro *nuisance* θ é desconhecido, a densidade preditiva bayesiana resultará na *t-Student* multivariada, já estudada no Capítulo 3. Neste caso, não será possível calcular as *Funções de Influência Preditivas* exatas, pela complexidade na resolução das integrais envolvidas e consequentemente, calcularemos *Funções de Influência Preditivas* aproximadas.

Neste estudo, estamos apresentando a abordagem proposta por JOHNSON e GEISSER (1983), que utiliza somente a densidade preditiva bayesiana na deteção de pontos influentes. Uma análise similar pode, evidentemente, ser feita adotando-se qualquer outra densidade preditiva.

Finalizamos o capítulo com um exemplo ilustrativo no qual estaremos caracterizando as observações ou grupo de observações influentes de maior a menor influência, com base nas *Funções de Influência Preditivas* aproximadas.

4.2 Previsão para um vetor de observações

Consideremos o modelo linear

$$\mathbf{Y} = X\beta + \mathbf{u} \tag{4.1}$$

em que $\mathbf{Y} = (Y_1, \cdots, Y_n)'$ é o vetor aleatório $n \times 1$ de dados observados, onde

X é uma matriz de constantes $n \times p$ de posto completo,

$\beta = (\beta_1, \cdots, \beta_p)'$ é o vetor $p \times 1$ de coeficientes de regressão desconhecidos onde $\beta \in \Omega_\beta$ e Ω_β é o espaço paramétrico para β e

$\mathbf{u}$ é o vetor de erros aleatórios não observável $n \times 1$ tal que $\mathbf{u} \sim \mathcal{N}_n(\mathbf{0}, \theta I)$, I é a matriz identidade de ordem n e θ é um parâmetro positivo.

Desejamos prever o vetor de observações futuras $\mathbf{Z}$ no modelo

$$\mathbf{Z} = W\beta + \mathbf{u}^* \tag{4.2}$$

onde $\mathbf{Z} = (Z_1, \cdots, Z_m)'$ é o vetor aleatório $m \times 1$ de variáveis aleatórias não observadas,

W é uma matriz de constantes $m \times p$ e

$\mathbf{u}^*$ é um vetor aleatório não observável $m \times 1$ tal que $\mathbf{u}^* \sim \mathcal{N}_m(\mathbf{0}, \theta I)$ e $\mathbf{u}$ e $\mathbf{u}^*$ são independentes.

Com esse objetivo, JOHNSON e GEISSER (1983) admitem a densidade a priori não informativa para (β, θ), dada por

$$f(\beta, \theta) \quad \propto \quad \theta^{-1}.$$

Quando $\mathbf{Y} = \mathbf{y}$ é observado, a densidade preditiva bayesiana de $\mathbf{Z}$ é

$$f(\mathbf{z}/W, X, \mathbf{y}) \propto \int f(\mathbf{z}/W, \beta, \theta)\, f(\beta, \theta/X, \mathbf{y})\, d\beta\, d\theta$$

sendo $f(\beta, \theta/X, \mathbf{y})$ a densidade a posteriori de (β, θ) e $f(\mathbf{z}/W, \beta, \theta)$ a densidade de $\mathbf{Z}$ dado β e θ, normal multivariada com média $W\beta$ e matriz de covariância θI. Como a priori $f(\beta, \theta)$ não depende do parâmetro β, temos dois casos a considerar:

(i) θ conhecido: A densidade preditiva $f(\mathbf{z}/W, X, \mathbf{y})$ é normal multivariada com média $W\hat{\beta}$ e matriz de covariância $\theta(I + W(X'X)^{-1}W')$, que indicaremos por

$$\mathcal{N}_m(W\hat{\beta},\, \theta(I + W(X'X)^{-1}W')) \tag{4.3}$$

sendo que $\hat{\beta} = (X'X)^{-1}X'\mathbf{Y}$ é o estimador de mínimos quadrados de β. O cálculo desta densidade encontra-se na Seção A.5, apêndice.

(ii) θ desconhecido: Neste caso, a densidade preditiva de $\mathbf{Z}$ dado $\mathbf{Y}$ foi obtida no Capítulo 3 e é uma *t-Student* multivariada com $(n - p)$ graus de liberdade, vetor de locação $W\hat{\beta}$ e matriz de correlação $s^2(I + W(X'X)^{-1}W')$, onde $s^2 = (\mathbf{y} - X\hat{\beta})'(\mathbf{y} - X\hat{\beta})/(n - p)$ e será indicada por

$$St_m(W\hat{\beta}, n - p, s^2(I + W(X'X)^{-1}W')). \tag{4.4}$$

Previsão quando observações são retiradas

Se, no modelo (4.1), retira-se um subconjunto de dados de tamanho k, indicaremos por "(i)" os dados restantes e por "i" os dados retirados. Desta forma, reescrevemos a matriz X e o vetor $\mathbf{Y}$ como

$$X = \begin{pmatrix} X_i \\ X_{(i)} \end{pmatrix} \qquad e \qquad \mathbf{Y} = \begin{pmatrix} \mathbf{Y}_i \\ \mathbf{Y}_{(i)} \end{pmatrix}.$$

Assim, X_i é a matriz $k \times p$ das variáveis auxiliares associadas às observações retiradas, $X_{(i)}$ é a matriz correspondente às observações restantes e vale uma interpretação análoga para $\mathbf{Y}_i$ e $\mathbf{Y}_{(i)}$. O modelo linear (4.1) poderá então ser escrito na forma

$$\mathbf{Y}' = (\mathbf{Y}_i', \mathbf{Y}_{(i)}') = \beta'(X_i', X_{(i)}') + (\mathbf{u}_i', \mathbf{u}_{(i)}').$$

Nestas condições, denotaremos por $f_{(i)} = f_{(i)}(\mathbf{z}/W, X_{(i)}, \mathbf{y}_{(i)})$ à densidade preditiva de $\mathbf{Z}$ calculada após a retirada do grupo de observações "i". Esta densidade será deduzida a seguir para os casos em que θ é conhecido e desconhecido.

(i) θ conhecido: De maneira análoga ao cálculo de (4.3), obtem-se a densidade preditiva de $\mathbf{Z}$ quando o modelo linear nas observações é $\mathbf{Y}_{(i)} = X_{(i)}\beta + \mathbf{u}_{(i)}$, resultando na densidade

$$\mathcal{N}_m(W\hat{\beta}_{(i)},\ \theta(I + W(X_{(i)}'X_{(i)})^{-1}W')), \tag{4.5}$$

onde $\hat{\beta}_{(i)} = (X_{(i)}'X_{(i)})^{-1}X_{(i)}'\mathbf{Y}_{(i)}$ é o estimador de mínimos quadrados de β com base nos dados restantes.

(ii) θ desconhecido: De maneira similar a (4.4), quando o conjunto "i" é retirado, a densidade preditiva de $\mathbf{Z}$ é

$$St_m(W\hat{\beta}_{(i)},\ n-k-p,\ s_{(i)}^2(I + W(X_{(i)}'X_{(i)})^{-1}W')) \tag{4.6}$$

onde $s_{(i)}^2 = (\mathbf{y}_{(i)} - \hat{\mathbf{y}}_{(i)})'(\mathbf{y}_{(i)} - \hat{\mathbf{y}}_{(i)})/(n-k-p)$ e $\hat{\mathbf{y}}_{(i)} = X_{(i)}\hat{\beta}_{(i)}$, com $\hat{\beta}_{(i)} = (X_{(i)}'X_{(i)})^{-1}X_{(i)}'\mathbf{Y}_{(i)}$.

Adicionalmente, introduziremos a seguinte notação para as estatísticas a serem usadas:

$$
\begin{aligned}
S &= X'X & S_{(i)} &= X'_{(i)}X_{(i)} \\
H &= XS^{-1}X' & H^{(i)} &= XS^{-1}_{(i)}X' \\
\hat{\mathbf{y}} &= X\hat{\beta} & \hat{\mathbf{y}}^{(i)} &= X\hat{\beta}_{(i)} \\
\mathbf{r} &= \mathbf{y} - \hat{\mathbf{y}} & \mathbf{r}_{(i)} &= \mathbf{y} - \hat{\mathbf{y}}^{(i)} \\
a^2 &= \mathbf{r}'\mathbf{r} & a^2_{(i)} &= \mathbf{r}'_{(i)}\mathbf{r}_{(i)} \\
s^2 &= a^2/(n-p) & s^2_{(i)} &= a^2_{(i)}/(n-k-p). \qquad (4.7)
\end{aligned}
$$

Utilizaremos ainda as estatísticas

$$
\begin{aligned}
V_i &= X_i S^{-1} X_i' \\
U_i &= X_i S^{-1}_{(i)} X_i' \\
\hat{\mathbf{y}}_i &= X_i\hat{\beta} \\
\mathbf{r}_i &= \mathbf{y}_i - \hat{\mathbf{y}}_i \\
t_i^2 &= a^{-2}\mathbf{r}_i'(I - V_i)^{-1}\mathbf{r}_i \\
T_i^2 &= a^2 a^{-2}_{(i)} t_i^2. \qquad (4.8)
\end{aligned}
$$

No modelo de regressão linear $\mathbf{Y} = X\beta + \mathbf{u}$ em que β é estimado por mínimos quadrados, define-se *resíduo internamente studentizado* para um grupo de observações (COOK e WEISBERG (1982), [pg. 30]) como

$$
e_{(i)} = \frac{\mathbf{r}_i'(I - V_i)^{-1}\mathbf{r}_i}{\hat{\sigma}^2},
$$

onde $\hat{\sigma}^2 = s^2$.

Observamos então que t_i^2 é proporcional à essa generalização do *resíduo internamente studentizado.* Observamos ainda que V_i é uma submatriz da *matriz hat* H, pois

$$
\begin{aligned}
H &= X(X'X)^{-1}X' \\
&= \begin{bmatrix} X_i \\ X_{(i)} \end{bmatrix} (X'X)^{-1} \begin{bmatrix} X_i' & X_{(i)}' \end{bmatrix} \\
&= \begin{bmatrix} X_i(X'X)^{-1}X_i' & X_i(X'X)^{-1}X_{(i)}' \\ X_{(i)}(X'X)^{-1}X_i' & X_{(i)}(X'X)^{-1}X_{(i)}' \end{bmatrix} \\
&= \begin{bmatrix} X_i S^{-1} X_i' & X_i S^{-1} X_{(i)}' \\ X_{(i)} S^{-1} X_i' & X_{(i)} S^{-1} X_{(i)}' \end{bmatrix}.
\end{aligned}
$$

Em ambos os casos, θ conhecido ou não, as densidades preditivas para os dados completos e os dados com observações retiradas serão comparadas para medir a influência dos dados retirados na previsão. Esta comparação será através da medida de divergência de Kullback-Leibler que, para o caso específico da previsão, foi definida como *Função de Influência Preditiva*, inicialmente por JOHNSON e GEISSER (1982) considerando a previsão de uma única observação. Os mesmos autores, em 1983, redefiniram as *Funções de Influência Preditivas* quando o interesse é a previsão de um vetor de observações. A seguir, descreveremos estas funções.

4.3 Funções de Influência Preditivas

Nesta seção, introduziremos o principal conceito do capítulo, que é a definição de *Funções de Influência Preditivas.* Com este objetivo, apresentaremos algumas definições preliminares.

Medida de falta de ajuste $D_i^2(\cdot)$

No modelo de regressão (4.1), uma classe de estatísticas que medem a influência de um subconjunto de dados quando o parâmetro β é estimado via mínimos quadrados é dada por

$$D_i^2(Q) = (\hat{\beta} - \hat{\beta}_{(i)})' Q (\hat{\beta} - \hat{\beta}_{(i)}), \tag{4.9}$$

onde $\hat{\beta}$ é o estimador de mínimos quadrados de β, $\hat{\beta}_{(i)}$ é o estimador de mínimos quadrados de β quando um subconjunto de observações "i" é retirado e Q é uma matriz positiva semi-definida.

Para $k = 1$ e $Q = S$ como definida em (4.7), a estatística $D_i^2(S)$ é proporcional à medida *D de Cook*, proposta por COOK (1977) para avaliar a influência da i-ésima observação quando se estima β via mínimos quadrados. Quando $k > 1$, vale uma interpretação similar.

As identidades

$$\begin{aligned} \hat{\beta}_{(i)} &= \hat{\beta} - \mathbf{r}_i'(I - V_i)^{-1} X_i S^{-1} \\ S_{(i)}^{-1} &= S^{-1} + S^{-1} X_i'(I - V_i)^{-1} X_i S^{-1} \\ a^2 a_{(i)}^{-2} &= 1 + T_i^2 \end{aligned}$$

$$\begin{aligned} T_i^2 &= t_i^2(1-t_i^2)^{-1} \\ U_i &= V_i(I-V_i)^{-1} \end{aligned} \tag{4.10}$$

demonstradas por BINGHAM (1977), serão úteis no cálculo e interpretação das *Funções de Influência Preditivas.*

Usando a forma de $\hat{\beta}_{(i)}$ dada em (4.10), obtemos a identidade

$$D_i^2(Q) = \mathbf{r}_i'(I-V_i)^{-1}X_iS^{-1}QS^{-1}X_i'(I-V_i)^{-1}\mathbf{r}_i.$$

Para $Q = S$, a medida $D_i^2(S)$ resulta em

$$D_i^2(S) = \mathbf{r}_i'(I-V_i)^{-1}V_i(I-V_i)^{-1}\mathbf{r}_i.$$

Da definição em (4.9) temos que

$$\begin{aligned} D_i^2(S) &= (\hat{\beta}-\hat{\beta}_{(i)})'S(\hat{\beta}-\hat{\beta}_{(i)}) \\ &= (\hat{\beta}-\hat{\beta}_{(i)})'X'X(\hat{\beta}-\hat{\beta}_{(i)}) \\ &= [X(\hat{\beta}-\hat{\beta}_{(i)})]'[X(\hat{\beta}-\hat{\beta}_{(i)})] \\ &= (X\hat{\beta}-X\hat{\beta}_{(i)})'(X\hat{\beta}-X\hat{\beta}_{(i)}) \\ &= (\hat{\mathbf{y}}-\hat{\mathbf{y}}^{(i)})'(\hat{\mathbf{y}}-\hat{\mathbf{y}}^{(i)}), \end{aligned} \tag{4.11}$$

resultado obtido ao utilizar as definições de $\hat{\mathbf{y}}$ e $\hat{\mathbf{y}}^{(i)}$ dadas em (4.7).

Assim, de (4.11), observamos que $D_i^2(S)$ é o quadrado da *distância euclidiana* entre o previsor de $\mathbf{Y}$ baseado no conjunto de dados completos e o correspondente previsor baseado no conjunto de dados com observações retiradas.

Divergências de Kullback-Leibler

A medida de *distância* ou *divergência* entre duas populações foi definida por KULLBACK e LEIBLER (1951). Assim, se f_1 e f_2 são funções densidade e E_{f_i} é o operador que toma esperança com respeito às densidades f_i, $i = 1,2$, os autores introduzem as *divergências direcionadas*

$$I(f_1,f_2) = E_{f_1}\, ln(f_1/f_2),$$

$$I(f_2,f_1) = E_{f_2}\, ln(f_2/f_1),$$

e a *divergência* entre f_1 e f_2

$$J(f_1, f_2) = I(f_1, f_2) + I(f_2, f_1).$$

Verifica-se que essas medidas são bem definidas e não negativas, sendo nulas apenas se $f_1 = f_2$. Além disso, se f_1 e f_2 são densidades de vetores aleatórios com distribuição $\mathcal{N}_n(\mathbf{u}_i, \Sigma_i)$, com Σ_i positiva definida, $i = 1, 2$ respectivamente, verifica-se (Resultado 11, Seção A.5 do apêndice) que

$$2I(f_1, f_2) = (\mathbf{u}_2 - \mathbf{u}_1)'\Sigma_2^{-1}(\mathbf{u}_2 - \mathbf{u}_1) + tr\Sigma_1\Sigma_2^{-1} - ln|\Sigma_1\Sigma_2^{-1}| - n\,, \quad (4.12)$$

onde $|\Sigma_1\Sigma_2^{-1}|$ representa o determinante da matriz $\Sigma_1\Sigma_2^{-1}$.

Funções de Influência Preditivas

Como a inferência sobre valores futuros baseada em distribuições preditivas depende da informação contida na amostra, é de interesse determinar quais subconjuntos de tal amostra são mais influentes para o propósito da previsão.

Conforme definido anteriormente, as funções $f = f(\mathbf{z}/W, X, \mathbf{y})$ e $f_{(i)} = f(\mathbf{z}/W, X_{(i)}, \mathbf{y}_{(i)})$, correspondem respectivamente à densidade preditiva de $\mathbf{Z}$ baseada em todas as observações e à densidade preditiva obtida após uma observação ou um grupo de observações ter sido retirado.

Para medir a influência na previsão de uma observação futura quando um particular elemento ou grupos de elementos amostrais é retirado, JOHNSON e GEISSER (1982) definiram as *Funções de Influência Preditivas* como as medidas de divergência entre f e $f_{(i)}$. Posteriormente, os autores redefiniram as *Funções de Influência Preditivas* para o caso mais geral da previsão de vetores (JOHNSON e GEISSER (1983)). Em ambos os casos, define-se as *Funções de Influência Preditivas* como as *divergências direcionadas*

$$I(f, f_{(i)}) = E_f\, ln(f/f_{(i)}),$$

$$I(f_{(i)}, f) = E_{f_{(i)}}\, ln(f_{(i)}/f),$$

ou como a divergência entre f e $f_{(i)}$,

$$J(f, f_{(i)}) = I(f, f_{(i)}) + I(f_{(i)}, f).$$

Apresentaremos a seguir as *Funções de Influência Preditivas* no modelo linear definido em (4.1) e (4.2). Conforme comentado anteriormente, quando θ é conhecido, calcularemos as *Funções de Influência Preditivas* exatas através de (4.3) e (4.5), e, para θ desconhecido, obteremos as *Funções de Influência Preditivas* aproximadas usando (4.4) e (4.6).

I. Funções de Influência Preditivas exatas, θ conhecido

Quando o parâmetro θ é conhecido nos modelos (4.1) e (4.2), *Funções de Influência Preditivas* exatas podem ser calculadas utilizando-se as densidades preditivas da seção anterior. Na prática, dificilmente o parâmetro θ é conhecido, mas este estudo é importante para ilustrar o cálculo das *Funções de Influência Preditivas* e facilitar a compreensão do caso em que θ é desconhecido.

Os resultados a seguir foram obtidos por JOHNSON e GEISSER (1983) e permitirão uma avaliação da influência das observações na previsão do próprio vetor de dados $\mathbf{Y}$, caso em que $W = X$.

Resultado 4.1
Nos modelos (4.1) e (4.2), com θ conhecido e $W = X$, a *Função de Influência Preditiva* $I(f, f_{(i)})$ é tal que

$$\begin{aligned} 2I(f, f_{(i)}) &= \frac{1}{2}D_i^2(\theta^{-1}\{S - \frac{1}{2}X_i'(I - \frac{1}{2}V_i)^{-1}X_i\}) \\ &\quad + \{ln|I + \frac{1}{2}V_i(I - V_i)^{-1}| - \frac{1}{2}tr[V_i(I - \frac{1}{2}V_i)^{-1}]\}. \end{aligned}$$

Prova : Como as densidades preditivas de f e $f_{(i)}$ são $\mathcal{N}_n(\hat{\mathbf{y}}, \theta(I + H))$ e $\mathcal{N}_n(\hat{\mathbf{y}}^{(i)}, \theta(I + H^{(i)}))$ respectivamente, então devido a (4.12)

$$\begin{aligned} 2I(f, f_{(i)}) &= (\hat{\mathbf{y}} - \hat{\mathbf{y}}^{(i)})'[\theta(I + H^{(i)})]^{-1}(\hat{\mathbf{y}} - \hat{\mathbf{y}}^{(i)}) + tr[\theta(I + H)\theta^{-1}(I + H^{(i)})^{-1}] \\ &\quad - ln|\theta(I + H)\theta^{-1}(I + H^{(i)})^{-1}| - n. \end{aligned}$$

Lembrando que $\hat{\mathbf{y}}^{(i)} = X\hat{\beta}_{(i)}$,

$$\begin{aligned} 2I(f, f_{(i)}) &= (X\hat{\beta} - X\hat{\beta}_{(i)})'\theta^{-1}(I + H^{(i)})^{-1}(X\hat{\beta} - X\hat{\beta}_{(i)}) \\ &\quad + \{tr[(I + H)(I + H^{(i)})^{-1}] - ln|(I + H)(I + H^{(i)})^{-1}| - n\} \\ &= (\hat{\beta} - \hat{\beta}_{(i)})'X'\theta^{-1}(I + H^{(i)})^{-1}X(\hat{\beta} - \hat{\beta}_{(i)}) \\ &\quad + \{tr[(I + H)(I + H^{(i)})^{-1}] - ln|(I + H)(I + H^{(i)})^{-1}| - n\}. \end{aligned}$$

Como $D_i^2(Q) = (\hat{\beta} - \hat{\beta}_{(i)})'Q(\hat{\beta} - \hat{\beta}_{(i)})$, esta expressão fica

$$\begin{aligned} 2I(f, f_{(i)}) &= D_i^2(X'\theta^{-1}(I + H^{(i)})^{-1}X) \\ &\quad + \{tr[(I + H)(I + H^{(i)})^{-1}] - ln|(I + H)(I + H^{(i)})^{-1}| - n\}. \end{aligned}$$

As matrizes $I+H$ e $(I+H^{(i)})^{-1}$ são simétricas e assim, $tr[(I+H)(I+H^{(i)})^{-1}] = tr[(I + H^{(i)})^{-1}(I + H)]$. Portanto, do Resultado 12 (vi) e (vii) (Seção A.5 do apêndice),

$$\begin{aligned} 2I(f, f_{(i)}) &= D_i^2[\theta^{-1}X'(I - \frac{1}{2}H - \frac{1}{2}XS^{-1}X_i'(I - \frac{1}{2}V_i)^{-1}X_iS^{-1}X')X] \\ &\quad + \{n - \frac{1}{2}tr[V_i(I - \frac{1}{2}V_i)^{-1}] - ln|(I + H)(I + H^{(i)})^{-1}| - n\} \\ &= D_i^2[\theta^{-1}(X'X - \frac{1}{2}X'HX - \frac{1}{2}X'XS^{-1}X_i'(I - \frac{1}{2}V_i)^{-1}X_iS^{-1}X'X)] \\ &\quad - \frac{1}{2}tr[V_i(I - \frac{1}{2}V_i)^{-1}] - ln|(I + H)(I + H^{(i)})^{-1}| \\ &= D_i^2[\theta^{-1}(X'X - \frac{1}{2}X'XS^{-1}X'X - \frac{1}{2}X'XS^{-1}X_i'(I - \frac{1}{2}V_i)^{-1}X_iS^{-1}X'X)] \\ &\quad - \frac{1}{2}tr[V_i(I - \frac{1}{2}V_i)^{-1}] - ln|(I + H)(I + H^{(i)})^{-1}| \\ &= D_i^2[\theta^{-1}(\frac{1}{2}S - \frac{1}{2}X_i'(I - \frac{1}{2}V_i)^{-1}X_i)] \\ &\quad - \frac{1}{2}tr[V_i(I - \frac{1}{2}V_i)^{-1}] - ln|(I + H)(I + H^{(i)})^{-1}|. \end{aligned}$$

Do Resultado 12 (i) e (iii) (Seção A.5 do apêndice) e sendo que para duas matrizes A e B, quadradas e de mesma ordem, $|AB| = |A||B|$, então

$$\begin{aligned} 2I(f, f_{(i)}) &= D_i^2[\theta^{-1}(\frac{1}{2}S - \frac{1}{2}X_i'(I - \frac{1}{2}V_i)^{-1}X_i)] \\ &\quad - \frac{1}{2}tr[V_i(I - \frac{1}{2}V_i)^{-1}] - ln[2^p(2^p)^{-1}|I + \frac{1}{2}V_i(I - V_i)^{-1}|^{-1}] \\ &= \frac{1}{2}D_i^2[\theta^{-1}(S - X_i'(I - \frac{1}{2}V_i)^{-1}X_i)] \\ &\quad + ln|I + \frac{1}{2}V_i(I - V_i)^{-1}| - \frac{1}{2}tr[V_i(I - \frac{1}{2}V_i)^{-1}]. \qquad \square \end{aligned}$$

Resultado 4.2
Nos modelos (4.1) e (4.2), com θ conhecido e $W = X$, a função de influência preditiva $I(f_{(i)}, f)$ é tal que,

$$2I(f_{(i)},f) = \frac{1}{2}D_i^2(\theta^{-1}S) + \{\frac{1}{2}tr[V_i(I-V_i)^{-1}] - ln|I+\frac{1}{2}V_i(I-V_i)^{-1}|\}.$$

Prova : De forma similar à prova anterior, usamos as densidades preditivas de f e $f_{(i)}$ e a relação (4.12) para obter

$$\begin{aligned} 2I(f_{(i)},f) &= (\hat{\mathbf{y}}^{(i)}-\hat{\mathbf{y}})'[\theta(I+H)]^{-1}(\hat{\mathbf{y}}^{(i)}-\hat{\mathbf{y}}) \\ &+ \{tr[\theta(I+H^{(i)})\theta^{-1}(I+H)^{-1}] - ln|\theta(I+H^{(i)})\theta^{-1}(I+H)^{-1}| - n\}. \end{aligned}$$

Como $\hat{\mathbf{y}}^{(i)} = X\hat{\beta}_{(i)}$,

$$\begin{aligned} 2I(f_{(i)},f) &= (X\hat{\beta}_{(i)}-X\hat{\beta})'[\theta(I+H)]^{-1}(X\hat{\beta}_{(i)}-X\hat{\beta}) \\ &\quad + \{tr[(I+H^{(i)})(I+H)^{-1}] - ln|(I+H^{(i)})(I+H)^{-1}| - n\} \\ &= (\hat{\beta}_{(i)}-\hat{\beta})'[\theta^{-1}X'(I+H)^{-1}X](\hat{\beta}_{(i)}-\hat{\beta}) \\ &\quad + \{tr[(I+H^{(i)})(I+H)^{-1}] - ln|(I+H^{(i)})(I+H)^{-1}| - n\}. \end{aligned}$$

Da definição da estatística $D_i^2(Q)$ em (4.9) e do Resultado 12 (i), (ii), (iv) e (iii), (Seção A.5 do apêndice) respectivamente

$$\begin{aligned} 2I(f_{(i)},f) &= D_i^2[\theta^{-1}X'(I-\frac{1}{2}H)X] \\ &+ \{n+\frac{1}{2}tr[V_i(I-V_i)^{-1}] - ln[2^p(2^p)^{-1}|I+\frac{1}{2}V_i(I-V_i)^{-1}|] - n\} \\ &= D_i^2(\theta^{-1}\frac{1}{2}S) + \{\frac{1}{2}tr[V_i(I-V_i)^{-1}] - ln|I+\frac{1}{2}V_i(I-V_i)^{-1}|\} \\ &= \frac{1}{2}D_i^2(\theta^{-1}S) + \{\frac{1}{2}tr[V_i(I-V_i)^{-1}] - ln|I+\frac{1}{2}V_i(I-V_i)^{-1}|\}. \end{aligned}$$ □

Devido à (4.11), observamos que o primeiro componente de cada uma das *Funções de Influência Preditivas* mede a falta de ajuste do conjunto de dados "(i)" relativo à uma métrica que depende de V_i. Já o segundo componente só depende de V_i, que sendo uma submatriz da matriz *Hat*, está relacionada à influencia das observações. Assim, o efeito do subconjunto de dados "i" é medido por uma estatística que mostra simultaneamente falta de ajuste e influência.

II. Funções de Influência Preditivas aproximadas, θ desconhecido

As *Funções de Influência Preditivas* são construídas com base nas densidades preditivas, que, quando o parâmetro θ é desconhecido, são densidades *t-Student* multivariadas. Infelizmente, não é possível obter expressões das *Funções de Influência Preditivas* a partir desta distribuição, devido à complexidade na solução das integrais envolvidas no cálculo. Pode-se, no entanto, obter *Funções de Influência Preditivas* aproximadas, substituindo apropriadamente densidades *t-Student* multivariadas por normais multivariadas. JOHNSON e GEISSER (1983) propõe o uso das densidades preditivas

$$\tilde{f}: \quad \mathcal{N}_n(\hat{\mathbf{y}}, (n-p)(n-p-2)^{-1}s^2(I+H)) \qquad e$$

$$\tilde{f}_{(i)}: \quad \mathcal{N}_n(\hat{\mathbf{y}}^{(i)}, (n-k-p)(n-k-p-2)^{-1}s^2_{(i)}(I+H^{(i)}))$$

como aproximação para as densidades *t-Student* (4.4) e (4.6), respectivamente para os dados completos e para os dados com observações retiradas.

A aproximação foi feita de modo que (ver Resultado 7, Seção A.2 do apêndice)

$$E_f(\mathbf{Z}) = E_{\tilde{f}}(\mathbf{Z}),$$

$$E_{f_{(i)}}(\mathbf{Z}) = E_{\tilde{f}_{(i)}}(\mathbf{Z}),$$

$$Var_f(\mathbf{Z}) = Var_{\tilde{f}}(\mathbf{Z}),$$

$$Var_{f_{(i)}}(\mathbf{Z}) = Var_{\tilde{f}_{(i)}}(\mathbf{Z}),$$

ou seja, se $\mathbf{Z}$ tinha distribuição *t-Student* multivariada com vetor de médias μ e matriz de covariâncias Σ, sua densidade foi aproximada pela densidade da normal multivariada com vetor de médias μ e matriz de covariâncias Σ.

Nestas condições, definem-se as *Funções de Influência Preditivas* aproximadas

$$\hat{I}(f, f_{(i)}) = I(\tilde{f}, \tilde{f}_{(i)})$$

e

$$\hat{I}(f_{(i)}, f) = I(\tilde{f}_{(i)}, \tilde{f}).$$

A seguir, apresentamos os resultados que servirão para a detecção de pontos influentes no caso do parâmetro θ ser desconhecido, no problema de previsão já descrito.

Resultado 4.3
Nos modelos (4.1) e (4.2), com θ desconhecido e $W = X$, a função de influência preditiva aproximada $\hat{I}(f, f_{(i)})$ é tal que

$$\begin{aligned}
2\,\hat{I}(f, f_{(i)}) &= 2\,I(\tilde{f}, \tilde{f}_{(i)}) \\
&= \frac{1}{2}\,c_1\,D_i^2[s_{(i)}^{-2}(S - X_i'(I - \frac{1}{2}V_i)^{-1}X_i)] \\
&\quad + \{ln|I + \frac{1}{2}V_i(I - V_i)^{-1}| - \frac{1}{2}\,b\,tr[V_i(I - \frac{1}{2}V_i)^{-1}]\} \\
&\quad + n\,\{b(1 - t_i^2)^{-1} - ln\,[b(1 - t_i^2)^{-1}] - 1\} \\
&\quad + \{-\frac{1}{2}\,b\,t_i^2(1 - t_i^2)^{-1}\,tr[V_i(I - \frac{1}{2}V_i)^{-1}]\},
\end{aligned}$$

onde $b = (n - p - 2)^{-1}(n - k - p - 2)$ e $c_1 = (n - k - p)^{-1}(n - k - p - 2)$.

Prova : De (4.12)

$$\begin{aligned}
2I(\tilde{f}, \tilde{f}_{(i)}) &= (\hat{\mathbf{y}} - \hat{\mathbf{y}}^{(i)})'(n - k - p)^{-1}(n - k - p - 2)s_{(i)}^{-2}(I + H^{(i)})^{-1}(\hat{\mathbf{y}} - \hat{\mathbf{y}}^{(i)}) \\
&+ tr\left[\frac{(n-p)}{(n-p-2)}s^2(I + H)\frac{(n-k-p-2)}{(n-k-p)}s_{(i)}^{-2}(I + H^{(i)})^{-1}\right] \\
&- ln\left|\frac{(n-p)}{(n-p-2)}s^2(I + H)\frac{(n-k-p-2)}{(n-k-p)}s_{(i)}^{-2}(I + H^{(i)})^{-1}\right| - n \\
&= \frac{(n-k-p-2)}{(n-k-p)}(\hat{\mathbf{y}} - \hat{\mathbf{y}}^{(i)})'s_{(i)}^{-2}(I + H^{(i)})^{-1}(\hat{\mathbf{y}} - \hat{\mathbf{y}}^{(i)}) \\
&+ tr\left[\frac{(n-k-p-2)}{(n-p-2)}(n-p)s^2(n-k-p)^{-1}s_{(i)}^{-2}(I + H)(I + H^{(i)})^{-1}\right] \\
&- ln\left|\frac{(n-k-p-2)}{(n-p-2)}(n-p)s^2(n-k-p)^{-1}s_{(i)}^{-2}(I + H)(I + H^{(i)})^{-1}\right| - n.
\end{aligned}$$

Tomando $c_1 = (n-k-p)^{-1}(n-k-p-2)$ e $b = (n-p-2)^{-1}(n-k-p-2)$ e substituindo $\hat{\mathbf{y}}$ por $X\hat{\beta}$ e $\hat{\mathbf{y}}^{(i)}$ por $X\hat{\beta}_{(i)}$,

$$\begin{aligned}
2I(\tilde{f}, \tilde{f}_{(i)}) &= c_1\left[(X\hat{\beta} - X\hat{\beta}_{(i)})'s_{(i)}^{-2}(I + H^{(i)})^{-1}(X\hat{\beta} - X\hat{\beta}_{(i)})\right] \\
&\quad + tr\left[b(n-p)s^2(n-k-p)^{-1}s_{(i)}^{-2}(I + H)(I + H^{(i)})^{-1}\right] \\
&\quad - ln\left|b(n-p)s^2(n-k-p)^{-1}s_{(i)}^{-2}(I + H)(I + H^{(i)})^{-1}\right| - n \\
&= c_1\left[(\hat{\beta} - \hat{\beta}_{(i)})'X's_{(i)}^{-2}(I + H^{(i)})^{-1}X(\hat{\beta} - \hat{\beta}_{(i)})\right] \\
&\quad + tr\left[ba^2a_{(i)}^{-2}(I + H)(I + H^{(i)})^{-1}\right] \\
&\quad - ln\left|ba^2a_{(i)}^{-2}(I + H)(I + H^{(i)})^{-1}\right| - n,
\end{aligned}$$

devido às relações em (4.7).

De (4.10), temos ainda que

$$\begin{aligned} a^2 a_{(i)}^{-2} &= 1 + T_i^2 \\ &= 1 + t_i^2 (1 - t_i^2)^{-1} \\ &= (1 - t_i^2)^{-1} \end{aligned}$$

e, utilizando o Resultado 12 (vi) (Seção A.5 do apêndice),

$$\begin{aligned} 2I(\tilde{f}, \tilde{f}_{(i)}) &= c_1 D_i^2 [s_{(i)}^{-2} X'(I + H^{(i)})^{-1} X] \\ &\quad + tr \left[b a^2 a_{(i)}^{-2} (I + H)(I + H^{(i)})^{-1} \right] \\ &\quad - ln \left| b a^2 a_{(i)}^{-2} (I + H)(I + H^{(i)})^{-1} \right| - n \\ &= c_1 D_i^2 \left[s_{(i)}^{-2} X'(I - \frac{1}{2} H - \frac{1}{2} X S^{-1} X_i' (I - \frac{1}{2} V_i)^{-1} X_i S^{-1} X') X \right] \\ &\quad + tr \left[b (1 - t_i^2)^{-1} (I + H)(I + H^{(i)})^{-1} \right] \\ &\quad - ln \left| b (1 - t_i^2)^{-1} (I + H)(I + H^{(i)})^{-1} \right| - n \\ &= c_1 D_i^2 \left[s_{(i)}^{-2} (X'X - \frac{1}{2} X'HX - \frac{1}{2} X'XS^{-1} X_i' (I - \frac{1}{2} V_i)^{-1} X_i S^{-1} X'X) \right] \\ &\quad + b (1 - t_i^2)^{-1} \, tr \left[(I + H)(I + H^{(i)})^{-1} \right] \\ &\quad - ln \left[b (1 - t_i^2)^{-1} \right]^n - ln |(I + H)(I + H^{(i)})^{-1}| - n \\ &= c_1 D_i^2 \left[s_{(i)}^{-2} (S - \frac{1}{2} S - \frac{1}{2} X_i' (I - \frac{1}{2} V_i)^{-1} X_i) \right] \\ &\quad + b (1 - t_i^2)^{-1} (n - \frac{1}{2} tr[V_i (I - \frac{1}{2} V_i)^{-1}]) \\ &\quad - n \, ln[b (1 - t_i^2)^{-1}] - ln |(I + H)(I + H^{(i)})^{-1}| - n \\ &= \frac{1}{2} c_1 D_i^2 \left[s_{(i)}^{-2} (S - X_i' (I - \frac{1}{2} V_i)^{-1} X_i) \right] \\ &\quad + n \, b (1 - t_i^2)^{-1} - \frac{1}{2} b (1 - t_i^2)^{-1} \, tr[V_i (I - \frac{1}{2} V_i)^{-1}] \\ &\quad - n \, ln[b (1 - t_i^2)^{-1})] - ln |(I + H)(I + H^{(i)})^{-1}| - n \\ &= \frac{1}{2} c_1 D_i^2 \left[s_{(i)}^{-2} (S - X_i' (I - \frac{1}{2} V_i)^{-1} X_i) \right] \\ &\quad + n \, b (1 - t_i^2)^{-1} - \frac{1}{2} b \, tr[V_i (I - \frac{1}{2} V_i)^{-1}] \\ &\quad - \frac{1}{2} b \, t_i^2 (1 - t_i^2)^{-1} \, tr[V_i (I - \frac{1}{2} V_i)^{-1}] - n \, ln[b (1 - t_i^2)^{-1})] \end{aligned}$$

$$- ln|(I+H)(I+H^{(i)})^{-1}| - n.$$

Do Resultado 12 (i) e (iii) (Seção A.5 do apêndice),

$$\begin{aligned}
2I(\tilde{f},\tilde{f}_{(i)}) &= \frac{1}{2}c_1 D_i^2\left[s_{(i)}^{-2}(S - X_i'(I-\frac{1}{2}V_i)^{-1}X_i)\right] \\
&\quad + n\,b(1-t_i^2)^{-1} - \frac{1}{2}b\,tr[V_i(I-\frac{1}{2}V_i)^{-1}] \\
&\quad - \frac{1}{2}b\,t_i^2(1-t_i^2)^{-1}\,tr[V_i(I-\frac{1}{2}V_i)^{-1}] - n\,ln[b(1-t_i^2)^{-1})] \\
&\quad - ln[2^p 2^{-p}|I+\frac{1}{2}V_i(I-\frac{1}{2}V_i)^{-1}|^{-1}] - n \\
&= \frac{1}{2}c_1 D_i^2\left[s_{(i)}^{-2}(S - X_i'(I-\frac{1}{2}V_i)^{-1}X_i)\right] \\
&\quad + \{\,ln|I+\frac{1}{2}V_i(I-\frac{1}{2}V_i)^{-1}| - \frac{1}{2}b\,tr[V_i(I-\frac{1}{2}V_i)^{-1}]\,\} \\
&\quad + n\{b(1-t_i^2)^{-1} - ln[b(1-t_i^2)^{-1}] - 1\} \\
&\quad + \{-\frac{1}{2}b\,t_i^2(1-t_i^2)^{-1}\,tr[V_i(I-\frac{1}{2}V_i)^{-1}]\,\}. \qquad \square
\end{aligned}$$

Resultado 4.4
Nos modelos (4.1) e (4.2), com θ desconhecido e $W = X$, a função de influência preditiva aproximada $I(\tilde{f}_{(i)},\tilde{f})$ é tal que

$$\begin{aligned}
2\,\hat{I}(f_{(i)},f) &= 2\,I(\tilde{f}_{(i)},\tilde{f}) \\
&= \frac{1}{2}c_2 D_i^2(s^{-2}S) \\
&\quad + \{\frac{1}{2}b^{-1}\,tr[V_i(I-V_i)^{-1}] - ln|I+\frac{1}{2}V_i(I-V_i)^{-1}|\} \\
&\quad + n\,\{b^{-1}(1-t_i^2) - ln[b^{-1}(1-t_i^2)] - 1\} \\
&\quad + \{-\frac{1}{2}b^{-1}\,t_i^2\,tr\left[V_i(I-V_i)^{-1}\right]\},
\end{aligned}$$

onde $b = (n-k-p-2)(n-p-2)^{-1}$ e $c_2 = (n-p-2)(n-p)^{-1}$.

Prova : De (4.12),

$$\begin{aligned}
2I(\tilde{f}_{(i)},\tilde{f}) &= (\hat{\mathbf{y}}^{(i)} - \hat{\mathbf{y}})'[(n-p)^{-1}(n-p-2)s^{-2}(I+H)^{-1}](\hat{\mathbf{y}}^{(i)} - \hat{\mathbf{y}}) \\
&+ tr\left[\frac{(n-k-p)}{(n-k-p-2)}s_{(i)}^2(I+H^{(i)})\frac{(n-p-2)}{(n-p)}s^{-2}(I+H)^{-1}\right]
\end{aligned}$$

$$
\begin{aligned}
&- \left\{ ln \left| \frac{(n-k-p)}{(n-k-p-2)} s^2_{(i)} (I + H^{(i)}) \frac{(n-p-2)}{(n-p)} s^{-2} (I+H)^{-1} \right| - n \right\} \\
&= \frac{(n-p-2)}{(n-p)} (\hat{\mathbf{y}} - \hat{\mathbf{y}}^{(i)})' [s^{-2}(I+H)^{-1}] (\hat{\mathbf{y}} - \hat{\mathbf{y}}^{(i)}) \\
&+ tr \left[\frac{(n-p-2)}{(n-k-p-2)} (n-k-p) s^2_{(i)} (n-p)^{-1} s^{-2} (I + H^{(i)})(I+H)^{-1} \right] \\
&- \left\{ ln \left| \frac{(n-p-2)}{(n-k-p-2)} (n-k-p) s^2_{(i)} (n-p)^{-1} s^{-2} (I + H^{(i)})(I+H)^{-1} \right| - n \right\}.
\end{aligned}
$$

Definindo $c_2 = (n-p-2)(n-p)^{-1}$ e $b = (n-k-p-2)(n-p-2)^{-1}$ e, substituindo $\hat{\mathbf{y}}$ por $X\hat{\beta}$ e $\hat{\mathbf{y}}^{(i)}$ por $X\hat{\beta}_{(i)}$,

$$
\begin{aligned}
2I(\tilde{f}_{(i)}, \tilde{f}) &= c_2 (\hat{\beta} - \hat{\beta}_{(i)})' [X' s^{-2} (I+H)^{-1} X] (\hat{\beta} - \hat{\beta}_{(i)}) \\
&\quad + tr[b^{-1}(n-k-p) s^2_{(i)} (n-p)^{-1} s^{-2} (I + H^{(i)})(I+H)^{-1}] \\
&\quad - ln|b^{-1}(n-k-p) s^2_{(i)} (n-p)^{-1} s^{-2} (I + H^{(i)})(I+H)^{-1}| - n \\
&= c_2 D_i^2 (s^{-2} X' (I+H)^{-1} X) \\
&\quad + \{ tr[b^{-1} a^2_{(i)} a^{-2} (I + H^{(i)})(I+H)^{-1}] \\
&\quad - ln|b^{-1} a^2_{(i)} a^{-2} (I + H^{(i)})(I+H)^{-1}| - n \},
\end{aligned}
$$

da definição de $D_i^2(\cdot)$ e usando as relações de (4.7).

Como consequência de (4.10) e do Resultado 12 (ii) (Seção A.5 do apêndice),

$$
\begin{aligned}
2I(\tilde{f}_{(i)}, \tilde{f}) &= c_2 D_i^2 (s^{-2} X' (I - \frac{1}{2} H) X) \\
&\quad + tr[b^{-1}((1-t_i^2)^{-1})^{-1} (I + H^{(i)})(I+H)^{-1}] \\
&\quad - ln|b^{-1}((1-t_i^2)^{-1})^{-1} (I + H^{(i)})(I+H)^{-1}| - n \\
&= c_2 D_i^2 (s^{-2} (X'X - \frac{1}{2} X'HX)) \\
&\quad + tr[b^{-1}(1-t_i^2)(I + H^{(i)})(I+H)^{-1}] \\
&\quad - ln|b^{-1}(1-t_i^2)(I + H^{(i)})(I+H)^{-1}| - n \\
&= c_2 D_i^2 (s^{-2} (S - \frac{1}{2} S)) \\
&\quad + tr[b^{-1}(1-t_i^2)(I + H^{(i)})(I+H)^{-1}] \\
&\quad - ln|b^{-1}(1-t_i^2)(I + H^{(i)})(I+H)^{-1}| - n \\
&= \frac{1}{2} c_2 D_i^2 (s^{-2} S) \\
&\quad + b^{-1}(1-t_i^2)\, tr[(I + H^{(i)})(I+H)^{-1})] \\
&\quad - ln[b^{-1}(1-t_i^2)]^n - ln|(I + H^{(i)})(I+H)^{-1}| - n.
\end{aligned}
$$

Usando o Resultado 12 (iv), (i) e (iii) (Seção A.5 do apêndice) respectivamente,

$$
\begin{aligned}
2I(\tilde{f}_{(i)}, \tilde{f}) &= \frac{1}{2} c_2 D_i^2(s^{-2}S) \\
&\quad + b^{-1}(1-t_i^2)\,(n + \frac{1}{2} tr\,[V_i(I-V_i)^{-1}]) \\
&\quad - n\,ln[b^{-1}(1-t_i^2)] - ln|I + \frac{1}{2}V_i(I-V_i)^{-1}| - n \\
&= \frac{1}{2} c_2 D_i^2(s^{-2}S) \\
&\quad + n\,b^{-1}(1-t_i^2) + \frac{1}{2} b^{-1}(1-t_i^2)\,tr\,[V_i(I-V_i)^{-1}] \\
&\quad - n\,ln[b^{-1}(1-t_i^2)] - ln|I + \frac{1}{2}V_i(I-V_i)^{-1}| - n \\
&= \frac{1}{2} c_2 D_i^2(s^{-2}S) \\
&\quad + \{n\,b^{-1}(1-t_i^2) + \frac{1}{2} b^{-1} tr[V_i(I-V_i)^{-1}] - \frac{1}{2} b^{-1} t_i^2\,tr\,[V_i(I-V_i)^{-1}] \\
&\quad - n\,ln[b^{-1}(1-t_i^2)] - ln|I + \frac{1}{2}V_i(I-V_i)^{-1}| - n\} \\
&= \frac{1}{2} c_2 D_i^2(s^{-2}S) \\
&\quad + \frac{1}{2} b^{-1} tr\,[V_i(I-V_i)^{-1}] \\
&\quad - ln|I + \frac{1}{2}V_i(I-V_i)^{-1}| \\
&\quad + n\{b^{-1}(1-t_i^2) - ln[b^{-1}(1-t_i^2)] - 1\} \\
&\quad - \frac{1}{2} b^{-1} t_i^2\,tr\,[V_i(I-V_i)^{-1}]. \qquad \square
\end{aligned}
$$

Observamos que o primeiro termo em $\hat{I}(f, f_{(i)})$ é proporcional à função distância de Cook para o caso k arbitrário, sendo que os dois primeiros componentes são muito semelhantes aos dos Resultados 4.1 e 4.2 e podem ser interpretados de maneira similar.

O último termo, que é sempre negativo, vai produzir um decréscimo em $\hat{I}(f, f_{(i)})$. Tal decréscimo será tanto maior quanto mais influente for o conjunto de dados "i" e vai compensar um possível aumento nas demais parcelas ocorrido devido à falta de ajuste.

Encerrado o estudo sobre *Funções de Influência Preditivas* nos modelos

lineares (4.1) e (4.2), apresentaremos a seguir um exemplo que ilustrará algumas das idéias desenvolvidas nesta seção.

4.4 Exemplo

Nesta seção, apresentamos a análise feita por JOHNSON e GEISSER (1983), com o objetivo de ilustrar o uso de $\hat{I}(f_{(i)}, f)$ (Resultado 4.4) na detecção de subconjuntos de observações influentes.

Os autores utilizam os dados de COOK e WEISBERG (1980), relativos a um experimento conduzido com o objetivo de avaliar o sucesso do tratamento de nuvens na tentativa de incrementar chuvas.

O experimento FACE (Florida Area Cumulus Experiment) foi realizado para verificar a existência de efeito do uso de "iodeto de prata" para aumentar a quantidade de chuva. A análise foi feita numa área de aproximadamente 3000 milhas quadradas ao norte e leste de Coral Gables, Florida.

Neste experimento, 24 dias no verão de 1975 foram considerados apropriados para aplicação do tratamento, com base num particular critério. Um dia seria analisado se

$$S - N_e \geq 1.5, \tag{4.13}$$

onde

S é a previsão da diferença entre as alturas máximas de uma nuvem se tratada e não tratada e

N_e é um fator que será tanto maior quanto melhores forem as chances de se obter chuva naturalmente.

Em geral, os dias apropriados foram aqueles nos quais $S - N_e$ era grande e a primeira chuva natural nesse dia foi pequena. Em cada dia considerado como apropriado, a decisão de aplicar o tratamento baseou-se numa aleatorização. Assim, aplicou-se o tratamento a 12 dias e analisou-se mais 12 dias sem o tratamento.

Na seleção dos dias através do critério da inequação (4.13), os investigadores consideraram somente dias com $C \leq 13\,\%$, onde

C é a porcentagem de nuvens na área experimental, medida usando radar.

Esta variável é tal que um dia com distúrbios ou não apropriado teria $C > 13\,\%$.

Em cada dia, foram medidas a variável resposta "quantidade de chuva", uma variável indicadora expressando se as nuvens foram ou não tratadas naquele dia e mais 9 variáveis, incluindo interações, consideradas importantes na explicação da variável resposta.

Foram consideradas a variável resposta

Y : *quantidade de chuva na região num período de* 6 *horas, medida em metros cúbicos* $\times\ 10^7$ *em cada dia analisado,*

e as seguintes variáveis explicativas:

$S - N_e$ *e* C *definidas anteriormente*
P : *quantidade de chuva na região destino, medida em metros cúbicos* $\times\ 10^7$, *uma hora antes da aplicação do tratamento*
T : *número de dias decorridos após o início do experimento* ($T = 0$ *para* 16 *de junho*, 1975).

A variável T é importante, pois possibilita a análise de uma tendência natural de chuva com o passar dos dias de experimento ou modificações nas técnicas experimentais. Foram ainda utilizadas as variáveis explicativas indicadoras:

$$A = \begin{cases} 1 & \textit{se houve aplicação do tratamento nas nuvens nesse dia,} \\ 0 & \textit{caso contrário} \end{cases}$$

e

$$E = \begin{cases} 1 & \textit{quando existe mensagem de movimento de radar,} \\ 2 & \textit{para mensagem de radar estacionário.} \end{cases}$$

A Tabela 4.1 fornece os valores dessas variáveis para a amostra de 24 dias.

Tabela 4.1 - Valores das variáveis para a amostra de 24 dias

Caso	A	T	S	C	P	E	Y
1	0	0	1,75	13,4	0,274	2	12,85
2	1	1	2,70	37,9	1,267	1	5,52
3	1	3	4,10	3,9	0,198	2	6,29
4	0	4	2,35	5,3	0,526	1	6,11
5	1	6	4,25	7,1	0,250	1	2,45
6	0	9	1,60	6,9	0,018	2	3,61
7	0	18	1,30	4,6	0,307	1	0,47
8	0	25	3,35	4,9	0,194	1	4,56
9	0	27	2,85	12,1	0,751	1	6,35
10	1	28	2,20	5,2	0,084	1	5,06
11	1	29	4,40	4,1	0,236	1	2,76
12	1	32	3,10	2,8	0,214	1	4,05
13	0	33	3,95	6,8	0,796	1	5,74
14	1	35	2,90	3,0	0,124	1	4,84
15	1	38	2,05	7,0	0,144	1	11,86
16	0	39	4,00	11,3	0,398	1	4,45
17	0	53	3,35	4,2	0,237	2	3,66
18	1	55	3,70	3,3	0,960	1	4,22
19	0	56	3,80	2,2	0,230	1	1,16
20	1	59	3,40	6,5	0,142	2	5,45
21	1	65	3,15	3,1	0,073	1	2,02
22	0	68	3,15	2,6	0,136	1	0,82
23	1	82	4,01	8,3	0,123	1	1,09
24	0	83	4,65	7,4	0,168	1	0,28

O modelo adotado foi

$$\begin{aligned} LY &= \beta_0 + \beta_1 A + \beta_2 T + \beta_3 (S - N_e) + \beta_4 C + \beta_5 LP + \beta_6 E \\ &\quad + \beta_{13}(A \times (S - N_e)) + \beta_{14}(A \times C) + \beta_{15}(A \times LP) \\ &\quad + \beta_{16}(A \times E) + \epsilon, \end{aligned} \tag{4.14}$$

onde $LY = log_{10}Y$, $LP = log_{10}P$ e os termos de produto cruzado correspondem às interações entre as variáveis.

O objetivo da análise é descrever a diferença entre as quantidades médias de chuva para dias com nuvens tratadas e dias com nuvens sem tratamento, fixadas as demais variáveis, dada por

$$\Delta LY = E(LY \mid A = 1) - E(LY \mid A = 0)$$

$$= \beta_1 + \beta_{13}(S - N_e) + \beta_{14}C + \beta_{15}LP + \beta_{16}E.$$

Desta forma, o subconjunto de parâmetros $\beta_1, \beta_{13}, \beta_{14}, \beta_{15}$ e β_{16} é de primordial interesse.

A Tabela 4.2 fornece o valor da *componente de locação* (função de D_i^2), a soma dos três membros restantes de $I(\tilde{f}_{(i)}, \tilde{f})$, que será chamada de *componente covariacional* e o valor da *soma* $I(\tilde{f}_{(i)}, \tilde{f})$ dado no Resultado 4.4.

Partindo dos dados completos, isto é, n=24 observações, a Tabela 4.2 mostra os 5 subconjuntos de dados mais influentes quando retiradas uma, duas e três observações respectivamente, em ordem decrescente com relação às magnitudes de $I(\tilde{f}_{(i)}, \tilde{f})$.

Tabela 4.2 - Grupos de observações mais influentes

	Observação	Componente de Locação	Componente Covariacional	Soma $\hat{I}(f_{(i)}, f)$
	2	7,02	16,32	23,24
	7	4,01	5,26	9,27
k=1	24	2,39	3,92	6,31
	6	2,53	0,79	3,32
	18	1,43	1,60	3,03
	2,5	40,10	22,85	62,95
	2,21	18,08	29,09	47,17
k=2	2,15	1,35	30,30	31,65
	2,14	8,21	21,24	29,45
	2,11	9,13	19,12	28,25
	2,5,21	125,22	41,87	167,09
	2,5,23	49,33	56,88	106,21
k=3	2,14,21	32,91	47,29	80,20
	2,5,15	32,56	44,56	77,12
	2,5,14	45,77	29,31	75,08

Quando $k = 1$, as observações mais influentes em ordem decrescente com respeito à *componente de locação* foram 2, 7, 6, 24 e 18, no entanto, com respeito à *componente covariacional* foram 2, 7, 24, 18 e 6. Da tabela, vemos que quando k=1, as observações 2, 7 e 24 são as mais influentes com respeito à *soma*.

Quando $k = 2$ ou $k = 3$, os cinco conjuntos mais influentes contêm a observação 2. A partir dos resultados apresentados na Tabela 4.2, os autores consideraram a observação 2 como "discrepante" . Verificou-se que essa observação possuia um alto valor em uma das variáveis explicativas, estando fora das condições planejadas para o experimento. Suspeitando-se que essa observação poderia ocultar características importantes dos dados restantes, ela foi retirada.

Após a retirada da observação 2, foram refeitos os cálculos, desta vez com base na amostra de n=23 observações e os resultados encontram-se na Tabela 4.3.

Tabela 4.3 - Subconjuntos de observações mais influentes, eliminada a observação 2

	Observação	Componente de Locação	Componente Covaracional	Soma $\hat{I}(f_{(i)}, f)$
	7	3,71	5,13	8,84
	24	2,21	3,79	6,00
k=1	6	2,25	0,81	3,06
	18	1,18	1,68	2,86
	17	0,56	0,93	1,49
	7,24	2,09	14,81	16,19
	4,24	4,22	8,75	12,97
k=2	1,17	8,17	4,18	23,35
	7,16	5,23	6,22	11,45
	4,7	6,35	4,86	11,21
	5,11,23	83,43	49,02	132,45
	1,13,17	23,39	13,77	36,16
k=3	1, 6,13	18,17	12,78	30,95
	4,16,24	9,61	18,16	27,77
	1, 8,17	19,28	3,59	22,87

Esta tabela mostra que as observações 7 e 24 são as mais influentes individualmente, e também aos pares, devido a seu efeito na estrutura de covariância, principalmente na influência por pares. Sendo assim, era de se esperar que o par (7,24) fosse o mais influente em grupos de 3 observações, mas isso não aconteceu. Nos grupos de 3 observações mais influentes, a ob-

servação 24 apareceu na quarta tripla mais influente, e o par (7,24) apareceu no sétima tripla mais influente.

Desta análise, os autores concluiram que o par (7,24) não se ajusta ao modelo assumido e que a tripla (5,11,23) é influente principalmente por apresentar valores das variáveis distantes da massa de dados. Além disso, nenhuma outra tripla mostrou-se tão inconsistente com o modelo.

Os autores da análise também concluiram que os primeiros sete pares mais influentes correspondem a dias sem tratamento, já, as primeiras dez triplas mais influentes envolvem sem distinção dias com nuvens tratadas e não tratadas.

Em particular, a tripla (5,11,23) é muito influente com respeito a ambos componentes de locação e covariacional, especialmente com respeito ao primeiro. Essas observações incluiram somente dias com nuvens tratadas mas respostas relativamente baixas e não se mostraram influentes individualmente nem aos pares. O par mais influente de (5,11,23) foi (5,23), que era o 85º em ordem de magnitude de $I(\tilde{f}_{(i)}, \tilde{f})$.

Um posterior teste de outlier mostrou que não existia inconsistência da tripla (5,11,23) com o modelo, talvez devido ao fato de que todos os subconjuntos de (5,11,23) são pouco influentes.

Finalizando, destacamos que, com base na medida *D de Cook*, COOK e WEISBERG (1980) concluiram que o par (7,24) deveria ser removido. A análise feita utilizando *Funções de Influência Preditivas* levou à mesma conclusão.

Capítulo 5

Densidades Preditivas na seleção de modelos de regressão

5.1 Introdução

Neste capítulo, apresentaremos o uso de densidades preditivas para a seleção de variáveis independentes no modelo de regressão linear múltipla, isto é, na escolha da melhor equação de regressão.

Esta escolha será feita com base no uso de uma *medida de verossimilhança preditiva*, obtida através da técnica de *Reutilização da Amostra* (GEISSER, (1975)).

De acordo com essa técnica, os dados amostrais são particionados em observações retidas e retiradas e para cada possível modelo, calcula-se uma *medida de verossimilhança preditiva* baseada em uma densidade preditiva qualquer das observações retiradas dado as retidas.

Em particular, GEISSER e EDDY (1979) utilizam duas densidades preditivas. Em ambos os casos, os autores obtêm a densidade preditiva de cada observação amostral dado as $n-1$ restantes e utilizam o produto dessas n funções como medida de verossimilhança preditiva para o modelo considerado. O modelo mais apropriado ou provável será o que apresentar o maior valor dessa medida.

A próxima seção descreve inicialmente o critério para selecionar o modelo quando a distribuição das observações está completamente especificada. Logo após, apresentamos o critério de *Reutilização da Amostra*, que permite

selecionar o melhor modelo quando um ou mais parâmetros da distribuição das observações são desconhecidos, construindo a medida de verossimilhança preditiva a partir do produto das densidades preditivas de cada observação dadas as demais.

Neste capítulo, para o cálculo das medidas de verossimilhança preditivas, utilizaremos a densidade preditiva bayesiana. Salientamos no entanto que, qualquer densidade preditiva, incluindo as obtidas através das Funções de Verossimilhança Preditivas, pode ser utilizada para obter as medidas de verossimilhança preditivas que vamos precisar. Também faremos menção do uso da densidade preditiva de quase verossimilhança considerada por GEISSER e EDDY (1979).

Na Seção 5.2, assumindo que as observações provêm de populações normais, ilustramos a utilização do critério para selecionar um modelo dentre três modelos possíveis, usando a densidade preditiva bayesiana.

Finalmente, na Seção 5.3, descrevemos o cálculo da medida de verossimilhança com o uso da densidade preditiva bayesiana desenvolvida por GEISSER (1965), quando as observações obedecem a um modelo de regressão linear múltipla. Desta forma, mostramos como o critério proposto é útil na seleção de variáveis explicativas no modelo de regressão. Através de um exemplo numérico, ilustraremos o procedimento, comparando-o com o método conhecido como *Todas as Possíveis Regressões* (DRAPER e SMITH (1981)), usando como critério de seleção o *Quadrado Médio do Resíduo* e a estatística C_p (MALLOWS (1973)).

5.2 Seleção do melhor modelo

Na literatura, uma questão de importância é escolher o modelo que melhor represente um conjunto de observações dadas. Com esse objetivo, vários procedimentos têm sido propostos. A seguir, descrevemos o critério apresentado por GEISSER e EDDY (1979) para a seleção do melhor modelo.

Vamos supor que $M_1, \cdots, M_m$ são todos os possíveis modelos que poderiam ter gerado um conjunto de observações dadas, ou, pelo menos, oferecem uma descrição satisfatória dos mesmos. Sejam $X_1, \cdots, X_n$, variáveis aleatórias independentes consistindo na amostra ou dados, e $\mathbf{x} = (x_1, \cdots, x_n)$ o conjunto dos correspondentes valores dessas observações. Em geral, considera-

se que os modelos M_k, $k = 1, \cdots, m$, especificam parcialmente ou completamente as distribuições das variáveis aleatórias X_j.

Admitiremos também que, para cada valor de X_j, $j = 1, 2, \cdots, n$, existe um conjunto de valores de l variáveis independentes ou explicativas $\mathbf{s}_j = (s_{j1}, \cdots, s_{jl})$. Em particular, quando dispomos de duas amostras de duas populações, um elemento de $\mathbf{s}_j$ poderia ser uma variável indicadora que descreveria a qual das populações a observação pertence.

Supondo que, sob cada modelo possível M_k, a distribuição das variáveis aleatórias $X_1, \cdots, X_n$ é completamente especificada, se $f(\mathbf{x}/\mathbf{s}, M_k)$ é a densidade conjunta do vetor aleatório $\mathbf{X} = (X_1, \cdots, X_n)$ sob M_k, e se $f_j(x_j/\mathbf{s}_j, M_k)$ é a densidade da variável aleatória X_j, para $j = 1, 2, \cdots, n$, então, a função de verossimilhança sob o modelo M_k é definida como

$$L_k = f(\mathbf{x}/\mathbf{s}, M_k) = \Pi_{j=1}^{n} f_j(x_j/\mathbf{s}_j, M_k),$$

onde $\mathbf{s} = (\mathbf{s}_1, \cdots, \mathbf{s}_n)$. Neste caso, observada a amostra, pode-se obter o valor de L_k, já que $f_j(x_j/\mathbf{s}_j, M_k)$ não depende de parâmetros desconhecidos.

Na maior parte dos casos, os modelos M_k só especificam a distribuição da amostra aleatória, de maneira que um subconjunto de parâmetros fica desconhecido. Assim, uma alternativa para obter o valor de L_k, sob um particular modelo M_k, é considerar densidades preditivas, de modo que os parâmetros desconhecidos possam ser eliminados por algum dos métodos explicados no Capítulo 2.

A seguir, apresentamos o método proposto por GEISSER (1975), que permite obter o valor da medida de verossimilhança utilizando as densidades preditivas de cada variável aleatória X_j, dado os valores das demais variáveis X_k, para todo $k \neq j$, $k = 1, 2, \cdots, n$.

Critério de Reutilização da Amostra

GEISSER (1975) propôs particionar os dados da amostra em observações retidas e observações retiradas, e posteriormente, utilizar uma densidade preditiva qualquer para a previsão das observações retiradas com base nas retidas.

Para nosso estudo, seja $f(z/\mathbf{x}, \mathbf{z}, M_k)$ uma densidade preditiva qualquer para uma observação futura z quando M_k é o verdadeiro modelo, sendo

que $\mathbf{x} = (x_1, \cdots, , x_n)$ corresponde aos dados observados e $\mathbf{z}$ é o vetor de variáveis explicativas associado a z. Se denotamos $\mathbf{x}_{(j)} = (x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_n)$ como o conjunto de observações $\mathbf{x}$ com a j-ésima observação retirada, e se a função $f_j(x_j/\mathbf{x}_{(j)}, \mathbf{z}_j, M_k)$ é a densidade preditiva para X_j dado $\mathbf{X}_{(j)} = \mathbf{x}_{(j)}$, a medida de verossimilhança preditiva sob o modelo M_k será

$$L_k = \Pi_{j=1}^{n} f_j(x_j/\mathbf{x}_{(j)}, \mathbf{z}_j, M_k), \tag{5.1}$$

$k = 1, 2, \cdots, m$, onde m é o número de possíveis modelos para uma particular situação.

Como L_k é obtida através do produto de densidades preditivas dos X_j, $j = 1, 2, \cdots, n$, estamos admitindo uma "independência preditiva" entre essas variáveis.

Calculadas todas as medidas de verossimilhança L_k, $k = 1, \cdots, m$, consideramos M_{k^*} como o modelo mais provável se L_{k^*} é o máximo de $L_1, \cdots, L_m$.

O critério de *Reutilização da Amostra* apresentado é definido para qualquer densidade preditiva $f(z/\mathbf{x}, \mathbf{z}, M_k)$, sendo z o valor que desejamos prever e $\mathbf{x}$ o conjunto de variáveis aleatórias observadas. Na seleção de modelos, GEISSER e EDDY (1979) consideraram duas particulares densidades preditivas:

(i) Densidade Preditiva de Quase-Verossimilhança

Sob o modelo M_k, seja $g(z/\mathbf{z}, \theta_k, M_k)$ a densidade de uma observação futura z, associada ao vetor de variáveis independentes ou explicativas $\mathbf{z}$, com parâmetro θ_k.

A densidade preditiva de quase verossimilhança de z é obtida substituindo-se θ_k em $g(z/\mathbf{z}, \theta_k, M_k)$ por seu estimador de máxima verossimilhança $\hat{\theta}_k$, calculado com base nos dados observados.

Assim, no cálculo de L_k em (5.1), utilizamos como densidade preditiva

$$f_j(x_j/\mathbf{x}_{(j)}, \mathbf{z}_j, M_k) = g(x_j/\mathbf{z}_j, \hat{\theta}_{k(j)}, M_k)$$

para cada $j = 1, 2, \cdots, n$, onde $\hat{\theta}_{k(j)}$ é o estimador de máxima verossimilhança de θ_k calculado em função de $\mathbf{x}_{(j)} = (x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_n)$.

(ii) Densidade Preditiva Bayesiana

Conforme definido no Capítulo 2, a densidade preditiva bayesiana de uma observação futura z baseada nos dados $\mathbf{x}$, sob o modelo M_k, é calculada como

$$f(z/\mathbf{x},\mathbf{z},M_k) = \int f(z/\mathbf{x},\theta_k,\mathbf{z},M_k) f(\theta_k/\mathbf{x},\mathbf{z},M_k) d\,\theta_k,$$

onde $f(\theta_k/\mathbf{x},\mathbf{z},M_k)$ é a densidade a posteriori de θ_k e $f(z/\mathbf{x},\theta_k,\mathbf{z},M_k)$ é a densidade condicional de Z dado $\mathbf{X}=\mathbf{x}$.

Neste trabalho estamos admitindo Z e $\mathbf{X}$ independentes dado θ_k, e com isso,

$$f(z/\mathbf{x},\mathbf{z},M_k) = \int f(z/\theta_k,\mathbf{z},M_k) f(\theta_k/\mathbf{x},\mathbf{z},M_k) d\,\theta_k. \tag{5.2}$$

A densidade a posteriori $f(\theta_k/\mathbf{x},\mathbf{z},M_k)$ é obtida como

$$f(\theta_k/\mathbf{x},\mathbf{z},M_k) = \frac{f(\mathbf{x},\theta_k/\mathbf{z},M_k)}{f(\mathbf{x}/\mathbf{z},M_k)} = \frac{f(\mathbf{x}/\theta_k,\mathbf{z},M_k) f(\theta_k/M_k)}{f(\mathbf{x}/\mathbf{z},M_k)}.$$

Substituindo essa expressão em (5.2), segue que

$$f(z/\mathbf{x},\mathbf{z},M_k) \;=\; \int f(z/\theta_k,\mathbf{z},M_k) \frac{f(\mathbf{x}/\theta_k,\mathbf{z},M_k) f(\theta_k/M_k)}{f(\mathbf{x}/\mathbf{z},M_k)} d\,\theta_k$$

onde $f(\mathbf{x}/\theta_k,\mathbf{z},M_k)$ é a densidade de $\mathbf{X}$, $f(\theta_k/M_k)$ é a densidade a priori para θ_k e $f(\mathbf{x}/\mathbf{z},M_k) = \int f(\mathbf{x},\theta_k/\mathbf{z},M_k)\, d\theta_k$.

Como $f(\mathbf{x}/\mathbf{z},M_k)$ não depende de z e nem de θ_k e pela independência dos X_j para todo $j=1,2,\cdots,n$,

$$f(z/\mathbf{x},\mathbf{z},M_k) \propto \int f(z/\theta_k,\mathbf{z},M_k)\, \Pi_{j=1}^{n} f(x_j/\theta_k,\mathbf{z}_j,M_k)\, f(\theta_k/M_k)\, d\theta_k.$$

Desta forma, a densidade preditiva bayesiana para uma observação amostral x_j dadas as demais, sob o modelo M_k, será

$$f_j(x_j/\mathbf{x}_{(j)},\mathbf{z}_j,M_k) \propto \int f(x_j/\theta_k,\mathbf{z}_j,M_k)\, \Pi_{i=1,\, i\neq j}^{n} f(x_i/\theta_k,\mathbf{z}_i,M_k)\, f(\theta_k/M_k)\, d\theta_k,$$

sendo x_j o j-ésimo elemento de $\mathbf{x}$, $\mathbf{z}_i$ o vetor de variáveis independentes associado ao i-ésimo elemento amostral e $\mathbf{x}_{(j)}$ como definido previamente.

Ilustraremos a seguir o método descrito utilizando densidades preditivas bayesianas para a seleção de um modelo, quando as observações provêm de populações normais.

Consideremos então que os dados $\mathbf{x}_i = (x_{i1}, \cdots, x_{in_i})$ correspondem a duas amostras de tamanhos n_i, $i = 1, 2$ de duas populações normais, cujas densidades são:

$$f(x_i/\mu_i, \sigma_i^2) = (2\pi\sigma_i^2)^{-\frac{1}{2}} exp\left(-\frac{1}{2\sigma_i^2}(x_i - \mu_i)^2\right), \qquad i = 1, 2$$

e que existem três possíveis relações entre os parâmetros μ_i , σ_i, dando origem aos modelos M_1, M_2 e M_3 tais que

$$\begin{aligned} M_1: &\quad \mu_1 \neq \mu_2\ , \quad \sigma_1^2 \neq \sigma_2^2, \\ M_2: &\quad \mu_1 \neq \mu_2\ , \quad \sigma_1^2 = \sigma_2^2 \ \ e \\ M_3: &\quad \mu_1 = \mu_2\ , \quad \sigma_1^2 = \sigma_2^2. \end{aligned}$$

GEISSER (1964), utilizando a densidade a priori

$$g(\mu_i, \sigma_i) \ \propto \ \sigma_i^{-1}, \quad i = 1, 2, \tag{5.3}$$

obtem a densidade preditiva bayesiana de uma observação futura z, para os modelos M_1 e M_2.

(i) Densidade Preditiva Bayesiana sob M_1 : $\mu_1 \neq \mu_2$, $\sigma_1^2 \neq \sigma_2^2$

Sob M_1, a densidade preditiva bayesiana de uma observação futura z com base em $\mathbf{x}_i = (x_{i1}, \cdots, x_{in_i})$, proveniente da i-ésima população, $i = 1, 2$, para uma priori como em (5.3), é dada por

$$f(z/\bar{x}_i, \mathbf{z}_i, M_1) = \left(\frac{n_i}{\pi(n_i^2 - 1)}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{1}{2}n_i)}{\Gamma(\frac{1}{2}(n_i - 1))s_i} \left(1 + \frac{n_i(\bar{x}_i - z)^2}{(n_i^2 - 1)s_i^2}\right)^{-\frac{1}{2}n_i} \tag{5.4}$$

onde

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} \qquad \text{e} \qquad s_i^2 = \frac{\sum_j (x_{ij} - \bar{x}_i)^2}{n_i - 1}$$

são os estimadores de μ_i e σ_i^2 respectivamente e $\mathbf{z}_i = i$ é a variável independente que, neste caso, só indica de qual população foi obtida a amostra.

Assim, se x_{ij} é a j-ésima observação amostral da população i, reescrevendo (5.4), a densidade preditiva da observação x_{ij} dadas as demais será

$$f(x_{ij}/\bar{x}_{i(j)}, i, M_1) = \left(\frac{n_i - 1}{\pi(n_i - 2)n_i}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{1}{2}(n_i - 1))}{\Gamma(\frac{1}{2}(n_i - 2))s_{i(j)}} \left(1 + \frac{(n_i - 1)(\bar{x}_{i(j)} - x_{ij})^2}{n_i(n_i - 2)s_{i(j)}^2}\right)^{-\frac{1}{2}(n_i - 1)}$$

onde

$$\bar{x}_{i(j)} = \frac{\sum_{\substack{t=1 \\ t \neq j}}^{n_i} x_{it}}{n_i - 1} \qquad \text{e} \qquad s_{i(j)}^2 = \frac{\sum_{\substack{t=1 \\ t \neq j}}^{n_i} (x_{it} - \bar{x}_{i(j)})^2}{n_i - 2}.$$

Logo, a medida de verossimilhança preditiva usando o critério descrito será

$$L_1 = \Pi_{i=1}^{2} \Pi_{j=1}^{n_i} \left[\frac{n_i - 1}{\pi(n_i - 2)n_i}\right]^{\frac{1}{2}} \frac{\Gamma(\frac{1}{2}(n_i - 1))}{\Gamma(\frac{1}{2}(n_i - 2))s_{i(j)}} \left[1 + \frac{(n_i - 1)(\bar{x}_{i(j)} - x_{ij})^2}{n_i(n_i - 2)s_{i(j)}^2}\right]^{-\frac{1}{2}(n_i - 1)}.$$

(ii) Densidade Preditiva Bayesiana sob M_2 : $\mu_1 \neq \mu_2$, $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Neste modelo, utilizando a priori dada em (5.3), que neste caso se reduz a $1/\sigma$, GEISSER (1964) obteve a densidade preditiva bayesiana para uma observação futura z dado $\bar{\mathbf{x}}_i$ observado,

$$f(z/\bar{x}_i, \mathbf{z}_i, M_2) = \left(\frac{n_i}{\pi(n - 2)(n_i + 1)}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{1}{2}(n - 1))}{\Gamma(\frac{1}{2}(n - 2))s} \left[1 + \frac{n_i(\bar{x}_i - z)^2}{(n_i + 1)(n - 2)s^2}\right]^{-\frac{1}{2}(n-1)}$$

onde

$$s^2 = \frac{\sum_i (n_i - 1)s_i^2}{n - 2},$$

$\bar{x}_i$ e s_i^2 foram definidos no item **(i)** e $n = n_1 + n_2$.

Com isso, a densidade preditiva bayesiana para uma observação x_{ij} dadas as demais é

$$f(x_{ij}/\bar{x}_{i(j)}, i, M_2) = \left(\frac{n_i - 1}{\pi(n-3)n_i}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{1}{2}(n-2))}{\Gamma(\frac{1}{2}(n-3))s_{(ij)}} \left[1 + \frac{(n_i-1)(\bar{x}_{i(j)} - x_{ij})^2}{n_i(n-3)s^2_{(ij)}}\right]^{-\frac{1}{2}(n-2)}$$

onde

$$s^2_{(ij)} = (n-3)^{-1}\left[(n_i - 2)s^2_{i(j)} + (n_{3-i} - 1)s^2_{3-i}\right],$$

e $s^2_{i(j)}$ foi definido anteriormente, para $i = 1,2$ e $j = 1, \cdots, n_i$.

Finalizando, a medida de verossimilhança preditiva é dada por

$$L_2 = \Pi^2_{i=1}\Pi^{n_i}_{j=1}\left[\frac{n_i - 1}{\pi(n-3)n_i}\right]^{\frac{1}{2}} \frac{\Gamma(\frac{1}{2}(n-2))}{\Gamma(\frac{1}{2}(n-3))s_{(ij)}} \left[1 + \frac{(n_i-1)(\bar{x}_{i(j)} - x_{ij})^2}{n_i(n-3)s^2_{(ij)}}\right]^{-\frac{1}{2}(n-2)}.$$

(iii) Densidade Preditiva Bayesiana sob M_3 : $\mu_1 = \mu_2$, $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Sob M_3, e usando a priori $1/\sigma$, deduzimos a densidade preditiva bayesiana de uma observação futura z com base em $\bar{x}_i = \bar{x}$, $i = 1,2$, a partir da densidade preditiva dada em GEISSER e EDDY (1979),

$$f(z/\bar{x}_i, \mathbf{z}_i, M_3) = \left(\frac{n}{\pi(n^2-1)}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{1}{2}n)}{\Gamma(\frac{1}{2}(n-1))t} \left(1 + \frac{n(\bar{x} - z)^2}{(n^2-1)t^2}\right)^{-\frac{1}{2}(n-1)}$$

onde

$$\bar{x} = \frac{\sum_{j=1}^{n_1} x_{1j} + \sum_{j=1}^{n_2} x_{2j}}{n_1 + n_2},$$

e

$$t^2 = (n-1)^{-1}\sum_i \sum_j (x_{ij} - \bar{x})^2,$$

e $\mathbf{z}_i = i$ é a variável indicadora para identificar de que população foram obtidos os dados.

Assim, se x_{ij} é a j-ésima observação da amostra da população i, sob o modelo M_3, a densidade preditiva da observação x_{ij} dadas as demais é

$$f(x_{ij}/\bar{x}_{i(j)}, i, M_3) = \left(\frac{(n-1)}{\pi n(n-2)}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{1}{2}(n-1))}{\Gamma(\frac{1}{2}(n-2))t_{(ij)}} \left(1 + \frac{(n-1)(\bar{x}_{(ij)} - x_{ij})^2}{n(n-2)t^2_{(ij)}}\right)^{-\frac{1}{2}(n-2)}$$

onde

$$\bar{x}_{(ij)} = (n-1)^{-1} \sum_{k,l}^{(ij)} x_{kl},$$

$$t^2_{(ij)} = (n-2)^{-1} \sum_{k,l}^{(ij)} (x_{kl} - \bar{x}_{(ij)})^2$$

e $\sum_{k,l}^{(ij)}$ denota a soma sobre todos os valores de k, l exceto o par $k = i$ e $l = j$, para $i = 1, 2$ e $j = 1, \cdots, n_i$.

Desta forma, a medida de verossimilhança preditiva para M_3 fica

$$L_3 = \Pi_{i=1}^{2} \Pi_{j=1}^{n_i} \left[\frac{n-1}{\pi(n-2)n} \right]^{\frac{1}{2}} \frac{\Gamma(\frac{1}{2}(n-1))}{\Gamma(\frac{1}{2}(n-2)) t_{(ij)}} \left(1 + \frac{(n-1)(\bar{x}_{(ij)} - x_{ij})^2}{n(n-2)t^2_{(ij)}} \right)^{-\frac{1}{2}(n-2)}.$$

Conforme comentamos anteriormente, para a seleção do melhor modelo, escolhemos o máximo entre L_1, L_2 e L_3.

Ilustramos este critério com os dados obtidos de LINDLEY (1965) [pg. 124], analizados por GEISSER e EDDY (1979). Estes dados correspondem ao índice cefálico para duas amostras de bois de diferentes raças e estão dispostos na Tabela 5.1.

Tabela 5.1 - Indices cefálicos para as duas amostras

Amostra I	74,1	77,7	74,4	74,0	73,8	79,3	75,8
	82,8	72,2	75,2	78,2	77,1	78,4	76,3
	76,8						
Amostra II	70,8	74,9	74,2	70,4	69,2	72,2	76,8
	72,4	77,4	78,1	72,8	74,3	74,7	

Para os dados dessa tabela temos $log\, L_1 = -70,95$, $log\, L_2 = -69,70$ e $log\, L_3 = -72,05$, e assim, através do método proposto, selecionamos o modelo M_2. Admitimos portanto igualdade entre as variâncias das distribuições dos índices cefálicos, mas desigualdade de médias.

No exemplo apresentamos a utilização das densidades preditivas na seleção da distribuição de probabilidades associada a um modelo de duas distribuições

normais independentes.

Quando as observações seguem um modelo de regressão linear múltipla, um problema muito comum é o da seleção das variáveis independentes presentes na equação de regressão.

Neste caso, dadas p variáveis independentes, os possíveis modelos serão constituídos por todas as equações de regressão contendo cada subconjunto dessas variáveis. A próxima seção mostrará como selecionar um modelo de regressão linear múltipla através do uso de densidades preditivas.

5.3 Seleção de variáveis no modelo de regressão linear

O problema da seleção do melhor modelo de regressão consiste em determinar o subconjunto de variáveis independentes que melhor explique a variável resposta, segundo algum critério.

Inicialmente vamos considerar os modelos do Capítulo 3 para a previsão, no caso particular em que $\mathbf{Z}$ é uma variável aleatória unidimensional, isto é, na previsão de uma única observação.

Desta forma, sejam

$$\begin{aligned} \mathbf{Y} &= X\beta + \mathbf{u}, \\ \mathbf{Z} &= W\beta + \mathbf{u}^* \end{aligned} \tag{5.5}$$

sendo que

$\mathbf{Y} = (Y_1, \cdots, Y_n)'$ é o vetor de variáveis aleatórias observadas,

$\mathbf{Z} = Z$ é a variável aleatória não observada

X é uma matriz de constantes $n \times p$ de posto completo,

$W = \mathbf{w}$ um vetor linha p-dimensional,

$\beta = (\beta_1, \cdots, \beta_p)'$ é o vetor de parâmetros de regressão desconhecidos, $\beta \in \Omega_\beta$, onde Ω_β é o espaço paramétrico para β e

$\mathbf{u}$, $\mathbf{u}^*$ são vetores aleatórios independentes $n \times 1$ e 1×1 respectivamente, tais que $\mathbf{u} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I)$ e $\mathbf{u}^* \sim \mathcal{N}(0, \sigma^2)$.

Para estes modelos, GEISSER (1965) obteve a densidade preditiva de $c(\mathbf{Z} - \mathbf{w}\hat{\beta})^2/a^2$, onde

$$c = 1 - \mathbf{w}(X'X)^{-1}\mathbf{w}',$$

$$a^2 = (\mathbf{y} - X\hat{\beta})'(\mathbf{y} - X\hat{\beta}),$$

$\mathbf{Z}$ é m-dimensional e $\hat{\beta} = (X'X)^{-1}X'\mathbf{Y}$, admitindo densidade a priori para (β, σ^2) da forma $g(\beta, \sigma^2) \propto \sigma^{-1}$. A partir daí, verificamos que $f(z/\mathbf{y}, \mathbf{w}, X)$, a densidade preditiva de Z, unidimensional, é dada por

$$f(z/\mathbf{y}, \mathbf{w}, X) = \left(\frac{c}{\pi a^2}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{1}{2}(n-p))}{\Gamma(\frac{1}{2}(n-p-1))} \left[1 + \frac{c(z - \mathbf{w}\hat{\beta})^2}{a^2}\right]^{-\frac{1}{2}(n-p)}.$$

Portanto, utilizando esta relação, considerando p variáveis preditoras, a densidade preditiva para uma observação retirada y_j dadas as observações retidas $\mathbf{y}_{(j)} = (y_1, \cdots, y_{j-1}, y_{j+1}, \cdots, y_n)'$ será

$$f(y_j/\mathbf{y}_{(j)}, X_j, X_{(j)}) = \left(\frac{c_j}{\pi a^2_{(j)}}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{1}{2}(n-p-1))}{\Gamma(\frac{1}{2}(n-p-2))} \left[1 + \frac{c_j(y_j - X_j\hat{\beta}_{(j)})^2}{a^2_{(j)}}\right]^{-\frac{1}{2}(n-p-1)}$$

para $j = 1, 2, \cdots, n$, onde

$$\begin{aligned}
c_j &= 1 - X_j(X'X)^{-1}X_j', \\
a^2_{(j)} &= (\mathbf{y}_{(j)} - X_{(j)}\hat{\beta}_{(j)})'(\mathbf{y}_{(j)} - X_{(j)}\hat{\beta}_{(j)}), \\
\hat{\beta}_{(j)} &= (X'_{(j)}X_{(j)})^{-1}X'_{(j)}\mathbf{y}_{(j)}, \\
X &= (X_1', \cdots, X_n')', \\
X_{(j)} &= (X_1', \cdots, X'_{j-1}, X'_{j+1}, \cdots, X_n')' \text{ e} \\
X_j &= (x_{j1}, \cdots, x_{jp}).
\end{aligned}$$

Desta forma, a medida de verossimilhança preditiva considerando as p variáveis preditoras será

$$\begin{aligned}
L &= \Pi_{j=1}^n f(y_j/\mathbf{y}_{(j)}, X_j, X_{(j)}) \\
&= \Pi_{j=1}^n \left(\frac{c_j}{\pi a^2_{(j)}}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{1}{2}(n-p-1))}{\Gamma(\frac{1}{2}(n-p-2))} \left[1 + \frac{c_j(y_j - X_j\hat{\beta}_{(j)})^2}{a^2_{(j)}}\right]^{-\frac{1}{2}(n-p-1)}.
\end{aligned}$$

Para cada subconjunto de tamanho l arbitrário das p variáveis independentes, temos $\binom{p}{l}$ possíveis modelos com l variáveis independentes, para $l = 1, \cdots, p$. Considerando todos os possíveis modelos de regressão linear, teremos $N = \binom{p}{1} + \cdots + \binom{p}{p} = 2^p - 1$ modelos. A medida de verossimilhança preditiva associada ao k-ésimo modelo será

$$L_k = \Pi_{j=1}^{n} \left(\frac{c_j}{\pi a_{(j)}^2} \right)^{\frac{1}{2}} \frac{\Gamma(\frac{1}{2}(n-l-1))}{\Gamma(\frac{1}{2}(n-l-2))} \left[1 + \frac{c_j (y_j - X_j^k \hat{\beta}_{(j)})^2}{a_{(j)}^2} \right]^{-\frac{1}{2}(n-l-1)}$$

para $k = 1, 2, \cdots, N$, $l = 1, 2, \cdots, p$ onde l é o número de variáveis independentes e sendo que X_j^k é o vetor de valores das variáveis independentes presentes no modelo considerado, associado a y_j. Portanto, neste caso, a matriz $W = \mathrm{w}$ do modelo (5.5) corresponde a X_j^k.

Desta forma, com base no critério apresentado na seção anterior, escolhemos o modelo M_{k^*} tal que L_{k^*} é o maior de todos os L_k, para $k = 1, \cdots, N$.

Apresentamos a seguir um exemplo numérico da utilização do critério descrito. A análise, desenvolvida por GEISSER e EDDY (1979, 1980) consiste na seleção de variáveis independentes no modelo de regressão para o conjunto de dados de Hald (DRAPER e SMITH (1981)) através do cálculo da medida L_k.

Para os dados da Tabela 5.2, foram avaliadas quatro variáveis independentes X_1, X_2, X_3 e X_4, e a variável resposta Y.

Tabela 5.2 - Conjunto de Dados de Hald

Y	X_1	X_2	X_3	X_4
78,5	7	26	6	60
74,3	1	29	15	52
104,3	11	56	8	20
87,6	11	31	8	47
95,9	7	52	6	33
109,2	11	55	9	22
102,7	3	71	17	6
72,5	1	31	22	44
93,1	2	54	18	22
115,9	21	47	4	26
83,8	1	40	23	34
113,3	11	66	9	12
109,4	10	68	8	12

A Tabela 5.3 apresenta os valores de log L_k calculados por GEISSER e EDDY (1980) para os quinze possíveis modelos de regressão com intercepto, obtidos com base em cada subconjunto das variáveis independentes. A Tabela 5.3 fornece também o *quadrado médio do resíduo* e o valor da estatística C_p de Mallows após o ajuste de cada modelo.

Tabela 5.3 - Critério de Seleção

Variáveis Independentes	Quadrado Médio do Resíduo	C_p	Log L_k
1	115,06	202,55	-51,15
2	82,39	142,49	-49,80
3	176,31	315,15	-54,04
4	80,35	138,73	-49,06
12	5,79	2,68	-32,46
13	122,71	198,07	-52,50
14	7,48	5,50	-34,36
23	41,54	62,40	-46,13
24	86,89	138,23	-50,90
34	17,57	22,37	-40,22
123	5,35	3,04	-32,84
124	5,33	3,02	-32,56
134	5,65	3,50	-32,98
234	8,20	7,34	-35,60
1234	5,98	5,00	-34,25

O modelo com o maior valor de L_k é o que contem as variáveis independentes X_1 e X_2 e assim, seria o modelo escolhido sob o critério proposto.

Observamos também que a ordenação dos modelos segundo a medida $log\, L_k$ é similar à obtida utilizando-se o valor de C_p. É importante lembrar que modelos com baixos valores de C_p são usualmente selecionados. De acordo com GEISSER e EDDY (1980), pode ser verificado que os critérios C_p e $Log\, L_k$ são assintoticamente equivalentes quando a variável resposta Y tem distribuição normal.

Salientamos ainda que o modelo selecionado pelo critério proposto coincide com o obtido por DRAPER e SMITH (1981) utilizando o procedimento de *stepwise*.

5.4 Considerações finais

Existem muitas referências que estudam técnicas bayesianas e não bayesianas para abordar o problema de seleção de variáveis em um modelo de regressão linear através de densidades preditivas. Por exemplo, SAN MARTINI e SPEZZAFERRI (1984) apresentam um procedimento para selecionar um modelo que irá produzir a melhor previsão de observações futuras. As distribuições preditivas associadas a cada modelo são comparadas por intermédio de uma função de utilidade logarítmica. Desta forma, o modelo h é preferido com relação ao modelo k se $\int log\left[\frac{p_h(z)}{p_k(z)}\right] p^*(z)\ dz > 0$, onde $p_h(z)$ e $p_k(z)$ são densidades preditivas da observação futura z sob os modelos h e k respectivamente e $p^*(z) = \sum_{j=1}^{N} \Pi_j p_j(z)$, onde Π_j representa a probabilidade do modelo M_j ser o verdadeiro, $j = 1, \cdots, n$.

GUTIERREZ-PEÑA (1998) dedica-se ao mesmo problema considerando uma classe mais ampla de possíveis modelos e funções de utilidade a posteriori. O procedimento proposto pelo autor teria ainda a vantagem de dispensar a especificação das probabilidades Π_j associadas a cada modelo, especificação essa que pode ser inviável do ponto de vista prático.

LAUD e IBRAHIM (1995) consideram o problema de selecionar um particular modelo dentre vários plausíveis utilizando uma medida que envolve a média e a variância de densidades preditivas. A adequação de cada modelo é avaliada verificando-se quão próximas estão as previsões associadas a um conjunto de dados observados e também a variabilidade dessas previsões. No cálculo da medida associada a cada modelo, é utilizada uma particular densidade preditiva e diferentes prioris para os parâmetros. Esta metodologia é utilizada em três importantes problemas associados a modelos de regressão linear: seleção de variáveis, seleção de transformações e estimação da variância desconhecida. Relações entre o critério proposto e outros conhecidos na literatura são também analisadas.

Apêndice A

Apêndice

A.1 Resultados Matriciais

Apresentaremos nesta seção resultados matriciais necessários para a derivação das densidades preditivas da Seção 3.2. O resultado básico é dado em (A.3) e sua demonstração depende dos fatos apresentados preliminarmente.

Resultado 1
Se A e D são matrices não singulares $m \times m$ e $k \times k$ respectivamente e B é uma matriz $m \times k$, então

$$(A + BDB')^{-1} = A^{-1} - A^{-1}B(B'A^{-1}B + D^{-1})^{-1}B'A^{-1}.$$

Lema A.1.1 *Sejam* $H = [X', W']'$ *e* $X'X = M$. *Então,*

$$(i) \quad (H'H)^{-1} = (M + W'W)^{-1} = M^{-1} - M^{-1}W'(I + WM^{-1}W')^{-1}WM^{-1}.$$

$$(ii) \quad I - (I + WM^{-1}W')^{-1}WM^{-1}W' = (I + WM^{-1}W')^{-1}.$$

$$(iii) \quad W(M + W'W)^{-1}W' = WM^{-1}W'(I + WM^{-1}W')^{-1}.$$

Prova:
(i) Temos que

$$H'H = [X'\ W'] \begin{bmatrix} X' \\ W' \end{bmatrix} = X'X + W'W = M + W'W,$$

e assim, $(H'H)^{-1} = (M + W'W)^{-1}$.

Tomando $A = M$, $B = W'$ e $D = I$ no Resultado 1,

$$\begin{aligned}(H'H)^{-1} &= (M + W'W)^{-1} \\ &= (M + W'IW)^{-1} \\ &= M^{-1} - M^{-1}W'(I + WM^{-1}W')^{-1}WM^{-1}. \qquad \square\end{aligned}$$

(ii) Aplicando o Resultado 1 para $A = I$, $B = I$ e $D = WM^{-1}W'$, temos

$$\begin{aligned}(I + WM^{-1}W')^{-1} &= I - I[I + (WM^{-1}W')^{-1}]^{-1} \\ &= I - \{(WM^{-1}W')^{-1}[WM^{-1}W' + I]\}^{-1} \\ &= I - (WM^{-1}W' + I)^{-1}WM^{-1}W'. \qquad \square\end{aligned}$$

(iii) Aplicando os resultados *(i)* e *(ii)* do Lema,

$$\begin{aligned}W(M + W'W)^{-1}W' &= W[M^{-1} - M^{-1}W'(I + WM^{-1}W')^{-1}WM^{-1}]W' \\ &= WM^{-1}W' - WM^{-1}W'(I + WM^{-1}W')^{-1}WM^{-1}W' \\ &= WM^{-1}W'[I - (I + WM^{-1}W')^{-1}WM^{-1}W'] \\ &= WM^{-1}W'(I + WM^{-1}W')^{-1}. \quad \square\end{aligned}$$

Resultado 2
Sejam $\hat{\beta}$, $\hat{\beta}_z$, $\hat{\sigma}^2$ e $\hat{\sigma}^2_z$ como foram definidos em (3.5), (3.6), (3.9) e (3.10). Nestas condições,

$$\hat{\sigma}^2_z = (n+m)^{-1}\left[n\hat{\sigma}^2 + (\hat{\beta} - \hat{\beta}_z)'M(\hat{\beta} - \hat{\beta}_z) + (\mathbf{Z} - W\hat{\beta}_z)'(\mathbf{Z} - W\hat{\beta}_z)\right].$$

Prova: De (3.10) temos que

$$\begin{aligned}\hat{\sigma}^2_z &= \frac{(\mathbf{V} - H\hat{\beta}_z)'(\mathbf{V} - H\hat{\beta}_z)}{(n+m)} \\ &= \frac{1}{(n+m)}\left[\begin{pmatrix}\mathbf{Y}\\ \mathbf{Z}\end{pmatrix} - \begin{pmatrix}X\\ W\end{pmatrix}\hat{\beta}_z\right]'\left[\begin{pmatrix}\mathbf{Y}\\ \mathbf{Z}\end{pmatrix} - \begin{pmatrix}X\\ W\end{pmatrix}\hat{\beta}_z\right]\end{aligned}$$

$$= \frac{1}{(n+m)} \begin{pmatrix} \mathbf{Y} - X\hat{\beta}_z \\ \mathbf{Z} - W\hat{\beta}_z \end{pmatrix}' \begin{pmatrix} \mathbf{Y} - X\hat{\beta}_z \\ \mathbf{Z} - W\hat{\beta}_z \end{pmatrix}$$

$$\begin{aligned}
&= (n+m)^{-1}\{(\mathbf{Y} - X\hat{\beta}_z)'(\mathbf{Y} - X\hat{\beta}_z) + (\mathbf{Z} - W\hat{\beta}_z)'(\mathbf{Z} - W\hat{\beta}_z)\} \\
&= (n+m)^{-1}\{(\mathbf{Y} - X\hat{\beta} + X\hat{\beta} - X\hat{\beta}_z)'(\mathbf{Y} - X\hat{\beta} + X\hat{\beta} - X\hat{\beta}_z) \\
&\quad + (\mathbf{Z} - W\hat{\beta}_z)'(\mathbf{Z} - W\hat{\beta}_z)\} \\
&= (n+m)^{-1}\{[(\mathbf{Y} - X\hat{\beta})' + (X\hat{\beta} - X\hat{\beta}_z)'][(\mathbf{Y} - X\hat{\beta}) + (X\hat{\beta} - X\hat{\beta}_z)] \\
&\quad + (\mathbf{Z} - W\hat{\beta}_z)'(\mathbf{Z} - W\hat{\beta}_z)\} \\
&= (n+m)^{-1}\{(\mathbf{Y} - X\hat{\beta})'(\mathbf{Y} - X\hat{\beta}) + (\mathbf{Y} - X\hat{\beta})'(X\hat{\beta} - X\hat{\beta}_z) \\
&\quad + (X\hat{\beta} - X\hat{\beta}_z)'(\mathbf{Y} - X\hat{\beta}) + (X\hat{\beta} - X\hat{\beta}_z)'(X\hat{\beta} - X\hat{\beta}_z) \\
&\quad + (\mathbf{Z} - W\hat{\beta}_z)'(\mathbf{Z} - W\hat{\beta}_z)\} \\
&= (n+m)^{-1}\{(\mathbf{Y} - X\hat{\beta})'(\mathbf{Y} - X\hat{\beta}) + (X'\mathbf{Y} - X'X\hat{\beta})'(\hat{\beta} - \hat{\beta}_z) \\
&\quad + (\hat{\beta} - \hat{\beta}_z)'(X'\mathbf{Y} - X'X\hat{\beta}) + (X\hat{\beta} - X\hat{\beta}_z)'(X\hat{\beta} - X\hat{\beta}_z) \\
&\quad + (\mathbf{Z} - W\hat{\beta}_z)'(\mathbf{Z} - W\hat{\beta}_z)\}.
\end{aligned}$$

Como $X'X\hat{\beta} = X'\mathbf{Y}$, esta expressão fica

$$\begin{aligned}
\hat{\sigma}_z^2 &= (n+m)^{-1}\{(\mathbf{Y} - X\hat{\beta})'(\mathbf{Y} - X\hat{\beta}) + \mathbf{0}'(\hat{\beta} - \hat{\beta}_z) + (\hat{\beta} - \hat{\beta}_z)'\mathbf{0} \\
&\quad + (X\hat{\beta} - X\hat{\beta}_z)'(X\hat{\beta} - X\hat{\beta}_z) + (\mathbf{Z} - W\hat{\beta}_z)'(\mathbf{Z} - W\hat{\beta}_z)\} \\
&= (n+m)^{-1}\{(\mathbf{Y} - X\hat{\beta})'(\mathbf{Y} - X\hat{\beta}) + (X\hat{\beta} - X\hat{\beta}_z)'(X\hat{\beta} - X\hat{\beta}_z) \\
&\quad + (\mathbf{Z} - W\hat{\beta}_z)'(\mathbf{Z} - W\hat{\beta}_z)\} \\
&= (n+m)^{-1}\{n\frac{(\mathbf{Y} - X\hat{\beta})'(\mathbf{Y} - X\hat{\beta})}{n} + (\hat{\beta} - \hat{\beta}_z)'X'X(\hat{\beta} - \hat{\beta}_z) \\
&\quad + (\mathbf{Z} - W\hat{\beta}_z)'(\mathbf{Z} - W\hat{\beta}_z)\} \\
&= (n+m)^{-1}\{n\hat{\sigma}^2 + (\hat{\beta} - \hat{\beta}_z)'M(\hat{\beta} - \hat{\beta}_z) + (\mathbf{Z} - W\hat{\beta}_z)'(\mathbf{Z} - W\hat{\beta}_z)\}. \quad \square
\end{aligned}$$

Resultado 3
Nas condições dos resultados anteriores

$$(i) \quad \hat{\beta}_z = (H'H)^{-1}M\hat{\beta} + (H'H)^{-1}W'\mathbf{Z}$$

$$(ii) \quad (\hat{\beta} - \hat{\beta}_z) = (H'H)^{-1}W'(W\hat{\beta} - \mathbf{Z})$$

Prova:
(i) De (3.8) e (3.9) do Capítulo 3,

$$\hat{\beta}_z = (H'H)^{-1}H'\mathbf{V}$$

$$
\begin{aligned}
&= (H'H)^{-1}[X'\,W'] \begin{bmatrix} \mathbf{Y} \\ \mathbf{Z} \end{bmatrix} \\
&= (H'H)^{-1}[X'\mathbf{Y} + W'\mathbf{Z}] \\
&= (H'H)^{-1}X'\mathbf{Y} + (H'H)^{-1}W'\mathbf{Z} \\
&= (H'H)^{-1}(X'X)(X'X)^{-1}X'\mathbf{Y} + (H'H)^{-1}W'\mathbf{Z} \\
&= (H'H)^{-1}(X'X)\hat{\beta} + (H'H)^{-1}W'\mathbf{Z} \\
&= (H'H)^{-1}M\hat{\beta} + (H'H)^{-1}W'\mathbf{Z}. \qquad \square
\end{aligned}
$$

(ii) Usando (i) e o Lema A.1.1-*(i)*

$$
\begin{aligned}
(\hat{\beta} - \hat{\beta}_{\mathbf{z}}) &= \hat{\beta} - (H'H)^{-1}M\hat{\beta} - (H'H)^{-1}W'\mathbf{Z} \\
&= (I - (H'H)^{-1}M)\hat{\beta} - (H'H)^{-1}W'\mathbf{Z} \\
&= (H'H)^{-1}((H'H) - M)\hat{\beta} - (H'H)^{-1}W'\mathbf{Z} \\
&= (H'H)^{-1}(M + W'W - M)\hat{\beta} - (H'H)^{-1}W'\mathbf{Z} \\
&= (H'H)^{-1}W'W\hat{\beta} - (H'H)^{-1}W'\mathbf{Z} \\
&= (H'H)^{-1}W'(W\hat{\beta} - \mathbf{Z}). \qquad \square
\end{aligned}
$$

Resultado 4
Para as matrizes H e M definidas no Lema A.1.1, temos

$$W(H'H)^{-1}M(H'H)^{-1}W' = WM^{-1}W'(I + WM^{-1}W')^{-2}.$$

Prova: Dos resultados *(i)* e *(iii)* do Lema A.1.1, e pela simetria das matrizes,

$$
\begin{aligned}
W(H'H)^{-1}M(H'H)^{-1}W' &= W[M^{-1} - M^{-1}W'(I + WM^{-1}W')^{-1}WM^{-1}]^{-1} \\
&\quad M[M^{-1} - M^{-1}W'(I + WM^{-1}W')^{-1}WM^{-1}]^{-1}W' \\
&= [I - WM^{-1}W'(I + WM^{-1}W')^{-1}]^{-1} \\
&\quad WM^{-1}MM^{-1}W'[I - (I + WM^{-1}W')^{-1}WM^{-1}W']^{-1} \\
&= [I - (I + WM^{-1}W')^{-1}WM^{-1}W']^{-1} \\
&\quad WM^{-1}W'[I - (I + WM^{-1}W')^{-1}WM^{-1}W']^{-1} \\
&= (I + WM^{-1}W')^{-1}WM^{-1}W'(I + WM^{-1}W')^{-1} \\
&= WM^{-1}W'(I + WM^{-1}W')^{-2}. \qquad \square
\end{aligned}
$$

Resultado 5
Para $\hat{\beta}_z$ como definido no Resultado 2,

$$(\mathbf{Z} - W\hat{\beta}_z) = (I + WM^{-1}W')^{-1}(\mathbf{Z} - W\hat{\beta}).$$

Prova: Do Resultado 3, e utilizando *(i)*, *(iii)*, *(ii)*, *(i)* e *(ii)* do Lema A.1.1 nessa ordem, temos

$$\begin{aligned}
(\mathbf{Z} - W\hat{\beta}_z) &= \{\mathbf{Z} - W[(H'H)^{-1}M\hat{\beta} + (H'H)^{-1}W'\mathbf{Z}]\} \\
&= (I - W(H'H)^{-1}W')\mathbf{Z} - W(H'H)^{-1}M\hat{\beta} \\
&= (I - W(M + W'W)^{-1}W')\mathbf{Z} - W(M + W'W)^{-1}M\hat{\beta} \\
&= (I - WM^{-1}W'(I + WM^{-1}W')^{-1})\mathbf{Z} - W(M + W'W)^{-1}M\hat{\beta} \\
&= (I - (I + WM^{-1}W')^{-1}WM^{-1}W')\mathbf{Z} - W(M + W'W)^{-1}M\hat{\beta} \\
&= (I + WM^{-1}W')^{-1}\mathbf{Z} - W(M + W'W)^{-1}M\hat{\beta} \\
&= (I + WM^{-1}W')^{-1}\mathbf{Z} \\
&\quad - W[M^{-1} - M^{-1}W'(I + WM^{-1}W')^{-1}WM^{-1}]M\hat{\beta} \\
&= (I + WM^{-1}W')^{-1}\mathbf{Z} - W(I - M^{-1}W'(I + WM^{-1}W')^{-1}W)\hat{\beta} \\
&= (I + WM^{-1}W')^{-1}\mathbf{Z} - [W - WM^{-1}W'(I + WM^{-1}W')^{-1}W]\hat{\beta} \\
&= (I + WM^{-1}W')^{-1}\mathbf{Z} - [I - WM^{-1}W'(I + WM^{-1}W')^{-1}]W\hat{\beta} \\
&= (I + WM^{-1}W')^{-1}\mathbf{Z} - (I - (I + WM^{-1}W')^{-1}WM^{-1}W')W\hat{\beta} \\
&= (I + WM^{-1}W')^{-1}\mathbf{Z} - (I + WM^{-1}W')^{-1}W\hat{\beta} \\
&= (I + WM^{-1}W')^{-1}(\mathbf{Z} - W\hat{\beta}). \qquad \square
\end{aligned}$$

Como consequência do Resultado 3, obtemos ainda

$$(\hat{\beta} - \hat{\beta}_z)'M(\hat{\beta} - \hat{\beta}_z) = (W\hat{\beta} - \mathbf{Z})'WM^{-1}W'(I + WM^{-1}W')^{-2}(W\hat{\beta} - \mathbf{Z}). \qquad \square \tag{A.1}$$

Devido ao Resultado 5, segue que

$$(\mathbf{Z} - W\hat{\beta}_z)'(\mathbf{Z} - W\hat{\beta}_z) = (\mathbf{Z} - W\hat{\beta})'(I + WM^{-1}W')^{-2}(\mathbf{Z} - W\hat{\beta}). \qquad \square \tag{A.2}$$

Resultado 6
Para $\hat{\beta}$ e $\hat{\sigma}^2$ definidos em (3.5) e (3.6) do Capítulo 3 e $\hat{\sigma}_{\mathbf{z}}^2$ como definido no Resultado 2,

$$\hat{\sigma}_{\mathbf{z}}^2 = (n+m)^{-1}\left[n\hat{\sigma}^2 + (\mathbf{Z}-W\hat{\beta})'(I+WM^{-1}W')^{-1}(\mathbf{Z}-W\hat{\beta})\right]. \tag{A.3}$$

Prova: Do Resultado 2 e utilizando (A.1) e (A.2),

$$\begin{aligned}
\hat{\sigma}_{\mathbf{z}}^2 &= (n+m)^{-1}\left[n\hat{\sigma}^2 + (\hat{\beta}-\hat{\beta}_{\mathbf{z}})'M(\hat{\beta}-\hat{\beta}_{\mathbf{z}}) + (\mathbf{Z}-W\hat{\beta}_{\mathbf{z}})'(\mathbf{Z}-W\hat{\beta}_{\mathbf{z}})\right] \\
&= (n+m)^{-1}[n\hat{\sigma}^2 + (W\hat{\beta}-\mathbf{Z})'WM^{-1}W'(I+WM^{-1}W')^{-2}(W\hat{\beta}-\mathbf{Z}) \\
&\quad + (\mathbf{Z}-W\hat{\beta})'(I+WM^{-1}W')^{-2}(\mathbf{Z}-W\hat{\beta})] \\
&= (n+m)^{-1}[n\hat{\sigma}^2 + (\mathbf{Z}-W\hat{\beta})'WM^{-1}W'(I+WM^{-1}W')^{-2}(\mathbf{Z}-W\hat{\beta}) \\
&\quad + (\mathbf{Z}-W\hat{\beta})'(I+WM^{-1}W')^{-2}(\mathbf{Z}-W\hat{\beta})] \\
&= (n+m)^{-1}\{n\hat{\sigma}^2 + [(\mathbf{Z}-W\hat{\beta})'WM^{-1}W' + (\mathbf{Z}-W\hat{\beta})'] \\
&\quad \times (I+WM^{-1}W')^{-2}(\mathbf{Z}-W\hat{\beta})\} \\
&= (n+m)^{-1}[n\hat{\sigma}^2 + (\mathbf{Z}-W\hat{\beta})'(WM^{-1}W'+I)(I+WM^{-1}W')^{-2}(\mathbf{Z}-W\hat{\beta})] \\
&= (n+m)^{-1}\left[n\hat{\sigma}^2 + (\mathbf{Z}-W\hat{\beta})'(I+WM^{-1}W')^{-1}(\mathbf{Z}-W\hat{\beta})\right]. \quad \square
\end{aligned}$$

A.2 Cálculo de Integrais

Para o cálculo de $r(\mathbf{z}, W/\mathbf{y}, X)$ na Seção 3.2, é necessário desenvolver a expressão (3.17), dada por

$$r(\mathbf{z}, W/\mathbf{y}, X) = \frac{\int_{\theta\in\Theta}\left(\frac{1}{\sigma^2}\right)\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n+m}{2}}\exp\left\{-(2\sigma^2)^{-1}(\mathbf{v}-H\beta)'(\mathbf{v}-H\beta)\right\}d\beta\,d\sigma^2}{\int_{\theta\in\Theta}\left(\frac{1}{\sigma^2}\right)\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}\exp\{-(2\sigma^2)^{-1}(\mathbf{y}-X\beta)'(\mathbf{y}-X\beta)\}\,d\beta\,d\sigma^2}. \tag{A.4}$$

Neste cálculo, utilizaremos os resultados mostrados a seguir.

Resultado 7 *(TONG (1990), [pg. 204])*
O vetor aleatório m-dimensional $\mathbf{t}$ tem distribuição *t-Student* multivariada com média $\mathbf{t}_0$, matriz de correlação R e n graus de liberdade se sua função densidade de probabilidades é

$$h(\mathbf{t};\mathbf{t}_0,R,n) = \frac{\Gamma(\frac{m+n}{2})}{(n\pi)^{\frac{m}{2}}\,\Gamma(\frac{n}{2})\,|R|^{\frac{1}{2}}}\left(1+\frac{1}{n}(\mathbf{t}-\mathbf{t}_0)'R^{-1}(\mathbf{t}-\mathbf{t}_0)\right)^{-\frac{m+n}{2}},$$

com $\mathbf{t}\in\mathbb{R}^m$ e R positiva definida. Esta densidade será indicada por $St_m(\mathbf{t}_0,n,R)$. Verifica-se ainda que se $n\geq 3$, a matriz de covariância de $\mathbf{t}$ é $\frac{n}{n-2}R$.

Resultado 8 *(BOX e TIAO (1973), [pg. 144])*
Se $\Gamma(\alpha)$ é a função gama real, então, para $a>0$ e $c>0$, verifica-se que

$$\int_0^\infty x^{-d-1}e^{-\frac{c}{x}}dx = c^{-d}\,\Gamma(d).$$

Resultado 9 *(GRAYBILL (1976) [pg. 48])*
Sejam a_0 e b_0 constantes escalares, $\mathbf{a}$ e $\mathbf{b}$ vetores $p\times 1$ de constantes e A e B matrizes de constantes, A simétrica e B positiva definida, ambas de dimensão p. Se $\mathbf{x}$ é um vetor $p\times 1$, com componentes $x_1,x_2,\ldots x_p$, o valor da integral múltipla

$$I = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(\mathbf{x}'A\mathbf{x}+\mathbf{x}'\mathbf{a}+a_0)\exp\{-[\mathbf{x}'B\mathbf{x}+\mathbf{x}'\mathbf{b}+b_0]\}\,dx_1\,dx_2\,\ldots\,dx_p \tag{A.5}$$

é dado por

$$I = \frac{1}{2}\pi^{\frac{p}{2}}|B|^{-\frac{1}{2}}e^{\frac{1}{4}\mathbf{b}'B^{-1}\mathbf{b}-b_0}\{tr(AB^{-1})-\mathbf{b}'B^{-1}\mathbf{a}+\frac{1}{2}\mathbf{b}'B^{-1}AB^{-1}\mathbf{b}+2a_0\}. \tag{A.6}$$

No cálculo da expressão (A.4), trabalharemos em separado com o numerador e o denominador, $I(n)$ e $I(d)$ respectivamente.

(i) Cálculo do numerador $I(n)$:
Pode-se escrever

$$(\mathbf{v}-H\beta)'(\mathbf{v}-H\beta)=\mathbf{v}'\mathbf{v}-2\mathbf{v}'H\beta+\beta'H'H\beta$$

e assim o numerador de (A.4) fica

$$I(n) = \int_{\theta\in\Theta} \left(\frac{1}{\sigma^2}\right)\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n+m}{2}} \exp\{-(2\sigma^2)^{-1}[\mathbf{v}'\mathbf{v} - 2\mathbf{v}'H\beta + \beta'H'H\beta]\}\, d\beta\, d\sigma^2.$$

Vamos integrar inicialmente com respeito a β, usando (A.6). Para isso, identificamos os elementos na equação (A.5) como $A = \mathbf{O}$, $\mathbf{a} = \mathbf{0}$, $a_0 = (\frac{1}{\sigma^2})(\frac{1}{2\pi\sigma^2})^{\frac{n+m}{2}}$, $B = \frac{H'H}{2\sigma^2}$, $\mathbf{b} = \frac{-H'\mathbf{v}}{\sigma^2}$ e $b_0 = \frac{\mathbf{v}'\mathbf{v}}{2\sigma^2}$. Nestas condições,

$$(\mathbf{x}'A\mathbf{x} + \mathbf{x}'\mathbf{a} + a_0) = \left(\frac{1}{\sigma^2}\right)\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n+m}{2}}$$

e

$$\exp\{-[\mathbf{x}'B\mathbf{x} + \mathbf{x}'\mathbf{b} + b_0]\} = \exp\left\{-\left[\frac{\beta'H'H\beta}{2\sigma^2} - \frac{2\beta H'\mathbf{v}}{2\sigma^2} + \frac{\mathbf{v}'\mathbf{v}}{2\sigma^2}\right]\right\}.$$

Desta forma, depois de integrar com respeito a β, tem-se

$$I(n) = \int_{\sigma^2>0} \frac{\pi^{\frac{p}{2}}}{2}\left|\frac{H'H}{2\sigma^2}\right|^{-\frac{1}{2}} e^{\left\{\frac{1}{4}\left(\frac{-\mathbf{v}'H}{\sigma^2}\right)\left(\frac{H'H}{2\sigma^2}\right)^{-1}\left(\frac{-H'\mathbf{v}}{\sigma^2}\right)-\frac{\mathbf{v}'\mathbf{v}}{2\sigma^2}\right\}} \left[\frac{2}{\sigma^2}\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n+m}{2}}\right] d\sigma^2$$

$$= \int_{\sigma^2>0} \pi^{\frac{p}{2}}\left(\frac{1}{\sigma^2}\right)\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n+m}{2}}\left|\frac{H'H}{2\sigma^2}\right|^{-\frac{1}{2}} e^{\left\{-\frac{1}{2\sigma^2}[\mathbf{v}'\mathbf{v}-\mathbf{v}'H(H'H)^{-1}H'\mathbf{v}]\right\}} d\sigma^2$$

$$= \left(\frac{1}{2\pi}\right)^{\frac{n+m-p}{2}} |H'H|^{-\frac{1}{2}} \int_{\sigma^2>0}\left(\frac{1}{\sigma^2}\right)^{\frac{n+m-p}{2}+1} e^{\left\{-\frac{1}{2}[\mathbf{v}'\mathbf{v}-\mathbf{v}'H(H'H)^{-1}H'\mathbf{v}](\sigma^2)^{-1}\right\}} d\sigma^2$$

$$= (2\pi)^{\frac{p-(n+m)}{2}}|H'H|^{-\frac{1}{2}} \int_{\sigma^2>0} (\sigma^2)^{-[\frac{n+m-p}{2}+1]} e^{\left\{-\{\frac{1}{2}[\mathbf{v}'\mathbf{v}-\mathbf{v}'H(H'H)^{-1}H'\mathbf{v}]\}(\sigma^2)^{-1}\right\}} d\sigma^2. \tag{A.7}$$

Para o cálculo da integral com respeito a σ^2, usamos o Resultado 8 com $x = \sigma^2$, $\mathbf{c} = \{\frac{1}{2}[\mathbf{v}'\mathbf{v} - \mathbf{v}'H(H'H)^{-1}H'\mathbf{v}]\}$ e $d = \frac{n+m-p}{2}$. Logo, (A.7) fica

$$I(n) = (2\pi)^{\frac{p-(n+m)}{2}}|H'H|^{-\frac{1}{2}}\{\tfrac{1}{2}[\mathbf{v}'\mathbf{v}-\mathbf{v}'H(H'H)^{-1}H'\mathbf{v}]\}^{-\frac{n+m-p}{2}}\Gamma(\tfrac{n+m-p}{2})$$

$$= (2\pi)^{\frac{p-(n+m)}{2}}|H'H|^{-\frac{1}{2}}(\frac{1}{2})^{-\frac{n+m-p}{2}}[\mathbf{v}'\mathbf{v} - \mathbf{v}'H(H'H)^{-1}H'\mathbf{v}]^{-\frac{n+m-p}{2}}\Gamma(\frac{n+m-p}{2})$$

$$= (\pi)^{\frac{p-(n+m)}{2}}|H'H|^{-\frac{1}{2}}[\mathbf{v}'\mathbf{v} - \mathbf{v}'H(H'H)^{-1}H'\mathbf{v}]^{-\frac{n+m-p}{2}}\Gamma(\frac{n+m-p}{2})$$

$$= (\pi)^{\frac{p-(n+m)}{2}}\Gamma(\frac{n+m-p}{2})|M + W'W|^{-\frac{1}{2}}[\mathbf{v}'\mathbf{v} - \mathbf{v}'H(H'H)^{-1}H'\mathbf{v}]^{-\frac{n+m-p}{2}}. \tag{A.8}$$

Vamos desenvolver a expressão (A.8) para obter o resultado requerido. Temos que

$$\begin{aligned}
\mathbf{v}'\mathbf{v} - \mathbf{v}'H(H'H)^{-1}H'\mathbf{v} &= \mathbf{v}'\mathbf{v} - \mathbf{v}'H(H'H)^{-1}H'\mathbf{v} - \mathbf{v}'H(H'H)^{-1}H'\mathbf{v} \\
&\quad + \mathbf{v}'H(H'H)^{-1}H'\mathbf{v} \\
&= \mathbf{v}'\mathbf{v} - \mathbf{v}'H\hat{\beta}_{\mathbf{z}} - \hat{\beta}_{\mathbf{z}}'H'\mathbf{v} + \mathbf{v}'H(H'H)^{-1}H'\mathbf{v} \\
&= \mathbf{v}'\mathbf{v} - \mathbf{v}'H\hat{\beta}_{\mathbf{z}} - \hat{\beta}_{\mathbf{z}}'H'\mathbf{v} + \mathbf{v}'H(H'H)^{-1}(H'H)(H'H)^{-1}H'\mathbf{v} \\
&= \mathbf{v}'\mathbf{v} - \mathbf{v}'H\hat{\beta}_{\mathbf{z}} - \hat{\beta}_{\mathbf{z}}'H'\mathbf{v} + \hat{\beta}_{\mathbf{z}}'(H'H)\hat{\beta}_{\mathbf{z}} \\
&= \mathbf{v}'(\mathbf{v} - H\hat{\beta}_{\mathbf{z}}) - \hat{\beta}_{\mathbf{z}}'H'(\mathbf{v} - H\hat{\beta}_{\mathbf{z}}) \\
&= (\mathbf{v} - H\hat{\beta}_{\mathbf{z}})'(\mathbf{v} - H\hat{\beta}_{\mathbf{z}}) \\
&= (n+m)\hat{\sigma}_{\mathbf{z}}^2. \qquad \square
\end{aligned}$$

Substituindo $\hat{\sigma}_{\mathbf{z}}^2$ pela expressão dada na equação (A.3),

$$\mathbf{v}'\mathbf{v} - \mathbf{v}'H(H'H)^{-1}H'\mathbf{v} =$$

$$\begin{aligned}
&= (n+m)\left\{(n+m)^{-1}[n\hat{\sigma}^2 + (\mathbf{z} - W\hat{\beta})'(I + WM^{-1}W')^{-1}(\mathbf{z} - W\hat{\beta})]\right\} \\
&= n\hat{\sigma}^2 + (\mathbf{z} - W\hat{\beta})'(I + WM^{-1}W')^{-1}(\mathbf{z} - W\hat{\beta}). \qquad \text{(A.9)}
\end{aligned}$$

Também, devido a resultados para matrizes particionadas (GRAYBILL (1976) [pg. 20])

$$\begin{aligned}
|H'H| = |M + W'W| &= |M(I + M^{-1}W'W)| \\
&= |M||I + M^{-1}W'W| \\
&= |M|\left|\begin{pmatrix} I & -M^{-1}W' \\ W & I \end{pmatrix}\right| \\
&= |M|\left|\begin{pmatrix} I & -M^{-1}W' \\ W & I \end{pmatrix}'\right| \\
&= |M|\left|\begin{pmatrix} I & W' \\ -WM^{-1} & I \end{pmatrix}\right| \\
&= |M||I||I + WM^{-1}W'| \\
&= |M||I + WM^{-1}W'|. \qquad \square \qquad \text{(A.10)}
\end{aligned}$$

Finalmente, de (A.9) e (A.10), a integral do numerador de (A.4) será

$$\begin{aligned}
I(n) &= (\pi)^{\frac{p-(n+m)}{2}}\,\Gamma\left(\frac{n+m-p}{2}\right)|M|^{-\frac{1}{2}}|I + WM^{-1}W'|^{-\frac{1}{2}} \\
&\quad \times \left[n\hat{\sigma}^2 + (\mathbf{z} - W\hat{\beta})'(I + WM^{-1}W')^{-1}(\mathbf{z} - W\hat{\beta})\right]^{-\frac{n+m-p}{2}}
\end{aligned}$$

$$= (\pi)^{\frac{p-(n+m)}{2}} \Gamma(\frac{n+m-p}{2}) |M|^{-\frac{1}{2}} |I + WM^{-1}W'|^{-\frac{1}{2}} (n\hat{\sigma}^2)^{-\frac{n+m-p}{2}}$$
$$\times \left[1 + (\mathbf{z} - W\hat{\beta})'[n\hat{\sigma}^2(I + WM^{-1}W')]^{-1}(\mathbf{z} - W\hat{\beta})\right]^{-\frac{n+m-p}{2}}. \quad \square$$

(A.11)

(ii) Cálculo do denominador $I(d)$:

Como

$$(\mathbf{y} - X\beta)'(\mathbf{y} - X\beta) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'X\beta + \beta'X'X\beta,$$

a expressão do denominador de (A.4) fica

$$I(d) = \int_{\theta\in\Theta} (\frac{1}{\sigma^2})(\frac{1}{2\pi\sigma^2})^{\frac{n}{2}} \exp\{-(2\sigma^2)^{-1}[\mathbf{y}'\mathbf{y} - 2\mathbf{y}'X\beta + \beta'X'X\beta]\}\, d\beta\, d\sigma^2.$$

De forma análoga ao numerador, integramos primeiro com respeito a β. Identificando as variáveis na expressão (A.5), temos que $A = \mathbf{O}$, $\mathbf{a} = \mathbf{0}$, $a_0 = (\frac{1}{\sigma^2})(\frac{1}{2\pi\sigma^2})^{\frac{n}{2}}$, $B = \frac{X'X}{2\sigma^2}$, $\mathbf{b} = \frac{-X'\mathbf{y}}{\sigma^2}$ e $b_0 = \frac{\mathbf{y}'\mathbf{y}}{2\sigma^2}$, então

$$(\mathbf{x}'A\mathbf{x} + \mathbf{x}'\mathbf{a} + a_0) = \left(\frac{1}{\sigma^2}\right)\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}$$

e

$$\exp\{-[\mathbf{x}'B\mathbf{x} + \mathbf{x}'\mathbf{b} + b_0]\} = \exp\left\{-\left[\frac{\beta'X'X\beta}{2\sigma^2} - \frac{2\beta X'\mathbf{y}}{2\sigma^2} + \frac{\mathbf{y}'\mathbf{y}}{2\sigma^2}\right]\right\}.$$

Desta forma, usando (A.6), integramos em β, obtendo

$$I(d) = (2\pi)^{\frac{p-n}{2}} |X'X|^{-\frac{1}{2}} \int_{\sigma^2>0} (\sigma^2)^{-[\frac{n-p}{2}+1]} e^{\left\{-\{\frac{1}{2}[\mathbf{y}'\mathbf{y} - \mathbf{y}'X(X'X)^{-1}X'\mathbf{y}]\}(\sigma^2)^{-1}\right\}} d\sigma^2. \tag{A.12}$$

Voltando a usar o Resultado 8 para integrar (A.12), temos que $x = \sigma^2$, $c = \{\frac{1}{2}[\mathbf{y}'\mathbf{y} - \mathbf{y}'X(X'X)^{-1}X'\mathbf{y}]\}$ e $d = \frac{n-p}{2}$, logo

$$I(d) = (\pi)^{\frac{p-n}{2}} \Gamma(\frac{n-p}{2}) |X'X|^{-\frac{1}{2}} [\mathbf{y}'\mathbf{y} - \mathbf{y}'X(X'X)^{-1}X'\mathbf{y}]^{-\frac{n-p}{2}}$$
$$= (\pi)^{\frac{p-n}{2}} \Gamma(\frac{n-p}{2}) |M|^{-\frac{1}{2}} [\mathbf{y}'\mathbf{y} - \mathbf{y}'X(X'X)^{-1}X'\mathbf{y}]^{-\frac{n-p}{2}}.$$

Através de um cálculo similar ao feito em (A.9), temos que

$$\mathbf{y}'\mathbf{y} - \mathbf{y}'X(X'X)^{-1}X'\mathbf{y} = n\hat{\sigma}^2$$

e com isso

$$I(d) \;=\; (\pi)^{\frac{p-n}{2}}\,\Gamma(\frac{n-p}{2})\,|M|^{-\frac{1}{2}}(n\hat{\sigma}^2)^{-\frac{n-p}{2}}. \qquad \square \qquad \text{(A.13)}$$

Efetuando a divisão de (A.11) por (A.13), obtemos

$$\begin{aligned}
\frac{I(n)}{I(d)} &= \frac{\Gamma(\frac{n+m-p}{2})}{\pi^{\frac{m}{2}}\Gamma(\frac{n-p}{2})}\left|(I+WM^{-1}W')\right|^{-\frac{1}{2}}(n\hat{\sigma}^2)^{-\frac{m}{2}} \\
&\quad\times\left[1+(\mathbf{z}-W\hat{\beta})'[n\hat{\sigma}^2(I+WM^{-1}W')]^{-1}(\mathbf{z}-W\hat{\beta})\right]^{-\frac{n+m-p}{2}} \\
&= \frac{\Gamma(\frac{n+m-p}{2})}{\pi^{\frac{m}{2}}\Gamma(\frac{n-p}{2})}\left|n\hat{\sigma}^2(I+WM^{-1}W')\right|^{-\frac{1}{2}} \\
&\quad\times\left[1+(\mathbf{z}-W\hat{\beta})'[n\hat{\sigma}^2(I+WM^{-1}W')]^{-1}(\mathbf{z}-W\hat{\beta})\right]^{-\frac{n+m-p}{2}} \\
&= \frac{\Gamma(\frac{n+m-p}{2})}{[(n-p)\pi]^{\frac{m}{2}}\Gamma(\frac{n-p}{2})}\left|n\hat{\sigma}^2\frac{(I+WM^{-1}W')}{(n-p)}\right|^{-\frac{1}{2}} \\
&\quad\times\left[1+\frac{1}{(n-p)}(\mathbf{z}-W\hat{\beta})'[n\hat{\sigma}^2\frac{(I+WM^{-1}W')}{(n-p)}]^{-1}(\mathbf{z}-W\hat{\beta})\right]^{-\frac{n+m-p}{2}}. \qquad \square
\end{aligned} \qquad \text{(A.14)}$$

A.3 Resultados para provar a Consistência das Densidades Preditivas

Na Seção 3.3, provando a consistência de algumas densidades preditivas, são utilizadas várias definições e resultados, que vamos expor a seguir.

Definição A.3.1 *Sejam* $\{a_n\}_{n\geq 1}$ *e* $\{b_n\}_{n\geq 1}$ *sequências de números reais; então diremos que*
(i) $a_n = O(b_n)$ *se existirem um número real* $k > 0$ *e um número inteiro possitivo* $n_o = n_o(k)$ *tal que*

$$\left|\frac{a_n}{b_n}\right| \leq k\,, \qquad\qquad \forall\, n \geq n_o$$

isto é, $a_n = O(b_n)$ *se a razão* $\left|\frac{a_n}{b_n}\right|$ *for limitada para todo n suficientemente grande.*

(ii) $a_n = o(b_n)$ *se para todo* $\varepsilon > 0$ *existir um número inteiro positivo* $n_o = n_o(\varepsilon)$ *tal que*

$$\left|\frac{a_n}{b_n}\right| < \varepsilon, \qquad \forall n \geq n_o$$

isto é, $a_n = o(b_n)$ *se* $\frac{a_n}{b_n} \to 0$ *quando* $n \to \infty$.

Definição A.3.2 *Sejam* $\{X_n\}_{n\geq 1}$ *uma sequência de variáveis aleatórias e* $\{b_n\}_{n\geq 1}$ *uma sequência de números reais (ou de variáveis aleatórias). Diremos que*
(i) $X_n = O_p(b_n)$ *se para todo número real* $\eta > 0$ *existirem um número real positivo* $k = k(\eta)$ *e um número inteiro* $n_o = n_o(\eta)$ *tais que*

$$P\left(\left|\frac{X_n}{b_n}\right| \geq k\right) \leq \eta, \qquad \forall n \geq n_o,$$

isto é, $X_n = O_p(b_n)$ *se a sequência* $\{X_n/b_n\}_{n\geq 1}$ *for limitada em probabilidade para todo n suficientemente grande.*
(ii) $X_n = o_p(b_n)$ *se para todo número real* $\varepsilon > 0$ *e para todo número real* $\eta > 0$ *existir um número inteiro positivo* $n_o = n_o(\varepsilon, \eta)$ *, tal que*

$$P\left(\left|\frac{X_n}{b_n}\right| \geq \varepsilon\right) \leq \eta, \qquad \forall n \geq n_o,$$

isto é, $X_n = o_p(b_n)$ *se para todo número real* $\varepsilon > 0$, $P\left(\left|\frac{X_n}{b_n}\right| \geq \varepsilon\right) \to 0$, *quando* $n \to \infty$.

Definição A.3.3 *(FULLER (1976), [pg. 182])*
(i) *Seja* $\mathbf{B}_n$ *uma matriz* $k \times r$ *de variáveis aleatórias, cujos elementos são* b_{ijn}. *Dizemos que* $\mathbf{B}_n$ *é de ordem no máximo* g_n *e indicamos por*

$$\mathbf{B}_n = O_p(g_n),$$

se, para todo $\epsilon > 0$, *existe um numero real positivo* M_ϵ *tal que*

$$P\left(|b_{ijn}| \geq M_\epsilon g_n\right) \leq \epsilon,$$

para $i = 1, 2, \cdots, k, \quad j = 1, 2, \cdots, r$ *e para todo* n.

(ii) *Dizemos que* $\mathbf{B}_n$ *é de menor ordem em probabilidade que* g_n *e indicamos por*

$$\mathbf{B}_n = o_p(g_n)$$

se, para todo $\epsilon > 0$ *e* $\delta > 0$, *existe um* N *tal que para todo* $n > N$,

$$P\left(|b_{ijn}| \geq \epsilon g_n\right) \leq \delta,$$

para $i = 1, 2, \cdots, k, \quad j = 1, 2, \cdots, r.$

Definição A.3.4 *(LEITE e SINGER (1990), [pg. 12])*
Seja $\{\mathbf{a}_n\}_{n\geq 1}$ *uma sequência de vetores* $p \times 1$ *e* $\{b_n\}_{n\geq 1}$ *uma sequência de números reais, onde*

$$\mathbf{a}_n = [a_{ni}] = \begin{pmatrix} a_{n1} \\ \vdots \\ a_{np} \end{pmatrix}.$$

Nestas condições,

$$\mathbf{a}_n = O(b_n) \qquad se \qquad \|\mathbf{a}_n\| = O(b_n);$$

$$\mathbf{a}_n = o(b_n) \qquad se \qquad \|\mathbf{a}_n\| = o(b_n);$$

onde $\|\cdot\|$ *é a norma do vetor.*

Teorema A.3.1 *(LEITE e SINGER (1990), [pg. 12])*
Seja $\{\mathbf{a}_n\}_{n\geq 1}$ *uma sequência de vetores* $p \times 1$ *e* $\{b_n\}_{n\geq 1}$ *uma sequência de números reais, então*

$$\mathbf{a}_n = O(b_n) \qquad se \qquad a_{ni} = O(b_n) \qquad \forall i = 1, \ldots, p.$$

$$\mathbf{a}_n = o(b_n) \qquad se \qquad a_{ni} = o(b_n) \qquad \forall i = 1, \ldots, p.$$

Definição A.3.5 *Seja* $\{\mathbf{A}_n\}_{n\geq 1}$ *uma sequência de vetores* $p \times 1$ *onde cada elemento* $\mathbf{a}_{ni}$ *de* $\mathbf{A}_n$ *é um vetor* $1 \times p$ *i.e.,* $\mathbf{a}_{ni} = (a_{ni1}, \cdots, a_{nip})$ *e* $\{b_n\}_{n\geq 1}$ *uma sequência de números reais tal que*

$$\mathbf{A}_n = [\mathbf{a}_{ni}]_{p\times 1} = \begin{pmatrix} \mathbf{a}_{n1} \\ \vdots \\ \mathbf{a}_{np} \end{pmatrix},$$

então

$$\mathbf{A}_n = \mathsf{O}(b_n) \qquad se \qquad \|\mathbf{A}_n\| = \mathsf{O}(b_n)$$

$$\mathbf{A}_n = \mathsf{o}(b_n) \qquad se \qquad \|\mathbf{A}_n\| = \mathsf{o}(b_n)$$

onde $\|\cdot\|$ *é a norma do vetor.*

Teorema A.3.2 *Seja* $\{\mathbf{A}_n\}_{n\geq 1}$ *uma sequência de vetores* $p\times 1$ *e* $\{b_n\}_{n\geq 1}$ *uma sequência de números reais. Nestas condições,*

$$\mathbf{A}_n = \mathsf{O}(b_n) \qquad se \qquad \mathbf{a}_{ni} = \mathsf{O}(b_n) \qquad \forall i = 1, \ldots, p$$

$$\mathbf{A}_n = \mathsf{o}(b_n) \qquad se \qquad \mathbf{a}_{ni} = \mathsf{o}(b_n) \qquad \forall i = 1, \ldots, p.$$

Definição A.3.6 *(LEITE e SINGER (1990), [pg. 42])*
Seja $\{\mathbf{X}_n\}_{n\geq 1} = \{(X_{n1}, X_{n2}, \cdots, X_{np})\}_{n\geq 1}$ *uma sequência de vetores aleatórios* $p \times 1$, *com* $p \geq 2$ *e* $\{b_n\}_{n\geq 1}$ *uma sequência de números reais. Diremos que*

$$\mathbf{X}_n = \mathsf{O}_p(b_n) \qquad se \qquad \|\mathbf{X}_n\| = \mathsf{O}_p(b_n)$$

e

$$\mathbf{X}_n = \mathsf{o}_p(b_n) \qquad se \qquad \|\mathbf{X}_n\| = \mathsf{o}_p(b_n),$$

onde $\|\cdot\|$ *é a norma do vetor.*

A seguir apresentaremos o resultado que pode ser utilizado para reduzir o caso vetorial ao caso unidimensional.

Teorema A.3.3 *(LEITE e SINGER (1990), [pg. 42])*
Seja $\{\mathbf{X}_n\}_{n\geq 1} = \{(X_{n1}, X_{n2}, \cdots, X_{np})\}_{n\geq 1}$ *uma sequência de vetores aleatórios* $p \times 1$, *com* $p \geq 2$ *e* $\{b_n\}_{n\geq 1}$ *uma sequência de números reais. Nestas condições,*

$\mathbf{X}_n = \mathrm{O}_p(b_n)$ *se e somente se* $X_{nj} = \mathrm{O}_p(b_n)$ *para* $j = 1, \cdots, p$

e

$\mathbf{X}_n = \mathrm{o}_p(b_n)$ *se e somente se* $X_{nj} = \mathrm{o}_p(b_n)$ *para* $j = 1, \cdots, p.$

Teorema A.3.4 *(FULLER (1976), [pg. 185])*
Seja $\{X_n\}_{n\geq 1}$ *uma sequência de variáveis aleatórias e* $\{a_n\}_{n\geq 1}$ *uma sequência de números reais positivos. Se* $E\{X_n^2\} = \mathrm{O}(a_n^2)$ *então*

$$X_n = \mathrm{O}_p(a_n).$$

Teorema A.3.5 *(FULLER (1976), [pg. 186])*
Nas condições do Teorema A.3.4, se $E\{(X_n - E\{X_n\})^2\} = \mathrm{O}(a_n^2)$ *e* $E\{X_n\} = \mathrm{O}(a_n)$ *então*

$$X_n = \mathrm{O}_p(a_n).$$

Apresentamos a seguir o lema dado em LEVY e PERNG (1984), correspondente ao Lema 3.3.1 do Capítulo 3, que é fundamental na prova dos resultados relativos à consistência da Seção 3.3.

Lema A.3.1 *(LEVY e PERNG (1984))*
Sejam $\mathbf{Y}$, X, $\mathbf{Z}$ *e* W, *vetores e matrizes definidos no Capítulo 3 e vamos supor que* $(X'X)/n \to D$ *quando* $n \to \infty$, D *positiva definida. Se* $\hat{\beta}$, $\hat{\sigma}^2$, $\hat{\beta}_z$ *e* $\hat{\sigma}_z^2$ *são os estimadores definidos em (3.5), (3.6), (3.9) e (3.10) respectivamente, então,*
(i) $\hat{\beta} - \beta = \mathrm{O}_p(n^{-\frac{1}{2}})$

(ii) $\hat{\sigma}^2 - \sigma^2 = \mathrm{O}_p(n^{-\frac{1}{2}})$

(iii) $\hat{\beta}_z - \hat{\beta} = \mathrm{O}_p(n^{-1})$

(iv) $\hat{\sigma}_z^2 - \hat{\sigma}^2 = \mathrm{O}_p(n^{-1})$.

Prova:
(i) Como $\mathbf{Y} \sim \mathcal{N}(X\beta, \sigma^2 I)$, verifica-se facilmente que $\hat{\beta} \sim \mathcal{N}(\beta, (X'X)^{-1}\sigma^2)$. Além disso, se $n^{-1}(X'X) \to D$ quando $n \to \infty$, prova-se que $n(X'X)^{-1} \to D^{-1}$ quando $n \to \infty$ (ELIAN (1991), [pg. 209]).

Como consequência, se $Var(\hat{\beta}) = (X'X)^{-1}\sigma^2$, então

$$\frac{Var(\hat{\beta})}{(1/n)} = n\, Var(\hat{\beta}) \to D^{-1}\sigma^2\,, \qquad n \to \infty.$$

Logo, se $b_n = n^{-1}$, pela Definição A.3.1, para cada elemento v_{ij} da matriz $Var(\hat{\beta})$, $\exists\ k_{ij} > 0$ e um número inteiro $n_{ij} = n_{ij}(k_{ij})$ tal que $|\frac{v_{ij}}{b_n}| < k_{ij}$, $\forall n \geq n_{ij}$. Em particular, se escrevermos $Var(\hat{\beta}) = (\mathbf{v}_1, \cdots, \mathbf{v}_p)$, onde $\mathbf{v}_j$ é um vetor p-dimensional, existe $k_i = max\{k_{ij}\,,\ j = 1, \cdots, p\} > 0$ e um número inteiro $n_i = max\{n_{ij}\,,\ j = 1, \cdots, p\}$ tal que

$$\left\|\frac{\mathbf{v}_i}{b_n}\right\| = \left\|\frac{\mathbf{v}_i}{(1/n)}\right\| < \frac{1}{b_n}\sqrt{\sum_{j=1}^{p} b_n^2\, k_{ij}^2} = \sqrt{\sum_{j=1}^{p} k_{ij}^2} \leq \sqrt{\sum_{j=1}^{p} k_i^2}\,, \qquad \forall n \geq n_i,$$

para cada $i = 1, \cdots, p$.

Então $\mathbf{v}_i = \mathsf{O}(b_n) = \mathsf{O}(n^{-1})$, $i = 1, \cdots, p$ e pelo Teorema A.3.2,

$$Var(\hat{\beta}) = E\left[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))'\right] = \mathsf{O}(n^{-1}) = \mathsf{O}(n^{-\frac{1}{2}})^2.$$

Da última expressão, vemos que cada um dos elementos da diagonal de $Var(\hat{\beta})$ é tal que

$$E(\hat{\beta}_j - E(\hat{\beta}_j))^2 = \mathsf{O}(n^{-1})$$

para todo $j = 1, \cdots, p$ e pelo Teorema A.3.4,

$$\hat{\beta}_j - E(\hat{\beta}_j) = \hat{\beta}_j - \beta_j = \mathsf{O}_p(n^{-\frac{1}{2}})$$

para todo $j = 1, \cdots, p$.

Logo, usando o Teorema A.3.3

$$\hat{\beta} - \beta = \mathsf{O}_p(n^{-\frac{1}{2}}). \qquad \square$$

(ii) Da teoria de Modelos Lineares para $\mathbf{Y} \sim \mathcal{N}(X\beta, \sigma^2 I)$,

$$\frac{1}{\sigma^2}(\mathbf{Y} - X\hat{\beta})'(\mathbf{Y} - X\hat{\beta}) \sim \chi^2_{(n-p)}$$

e portanto,

$$\frac{n}{\sigma^2}\frac{(\mathbf{Y} - X\hat{\beta})'(\mathbf{Y} - X\hat{\beta})}{n} = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-p)}.$$

Como consequência,

$$E\left(\frac{n\hat{\sigma}^2}{\sigma^2}\right) = \frac{n}{\sigma^2}E\left(\hat{\sigma}^2\right) = n - p\,,$$

$$Var\left(\frac{n\hat{\sigma}^2}{\sigma^2}\right) = \frac{n^2}{\sigma^4}Var\left(\hat{\sigma}^2\right) = 2(n-p)\,,$$

e com isto,

$$Var\left(\hat{\sigma}^2\right) \;=\; E(\hat{\sigma}^2 - E(\hat{\sigma}^2))^2 = \frac{2\sigma^4(n-p)}{n^2} = \mathsf{O}(n^{-1}) = \mathsf{O}(n^{-\frac{1}{2}})^2.$$

Assim, pelo Teorema A.3.4

$$\hat{\sigma}^2 - E(\hat{\sigma}^2) = \hat{\sigma}^2 - \frac{(n-p)}{n}\sigma^2 = \mathsf{O}_p(n^{-\frac{1}{2}}),$$

o que implica em

$$\hat{\sigma}^2 - \sigma^2 = \mathsf{O}_p(n^{-\frac{1}{2}}). \qquad \square$$

(iii) Do modelo descrito na Seção 3.1, temos que

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'\mathbf{Y} \\
&= (X'X)^{-1}X'[X\beta + \mathbf{u}] \\
&= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\mathbf{u} \\
&= \beta + (X'X)^{-1}X'\mathbf{u}.
\end{aligned}$$

Além disso, devido a (3.9), temos ainda que

$$\begin{aligned}
\hat{\beta}_{\mathbf{z}} &= (H'H)^{-1}H'\mathbf{V} \\
&= (H'H)^{-1}H'\begin{bmatrix}\mathbf{Y}\\ \mathbf{Z}\end{bmatrix}
\end{aligned}$$

$$
\begin{aligned}
&= (H'H)^{-1}H' \begin{bmatrix} X\beta + \mathbf{u} \\ W\beta + \mathbf{u}^* \end{bmatrix} \\
&= (H'H)^{-1}H' \begin{bmatrix} X\beta \\ W\beta \end{bmatrix} + (H'H)^{-1}H' \begin{bmatrix} \mathbf{u} \\ \mathbf{u}^* \end{bmatrix} \\
&= (H'H)^{-1}H' \begin{bmatrix} X \\ W \end{bmatrix} \beta + (H'H)^{-1}H' \begin{bmatrix} \mathbf{u} \\ \mathbf{u}^* \end{bmatrix} \\
&= (H'H)^{-1}H'[X', W']'\beta + (H'H)^{-1}H' \begin{bmatrix} \mathbf{u} \\ \mathbf{u}^* \end{bmatrix} \\
&= (H'H)^{-1}H'H\beta + (H'H)^{-1}[X'\,W'] \begin{bmatrix} \mathbf{u} \\ \mathbf{u}^* \end{bmatrix} \\
&= \beta + (H'H)^{-1}[X'\mathbf{u} + W'\mathbf{u}^*] \\
&= \beta + (H'H)^{-1}X'\mathbf{u} + (H'H)^{-1}W'\mathbf{u}^*.
\end{aligned}
$$

Subtraindo $\hat{\beta}_{\mathbf{z}}$ de $\hat{\beta}$,

$$
\begin{aligned}
\hat{\beta} - \hat{\beta}_{\mathbf{z}} &= \beta + (X'X)^{-1}X'\mathbf{u} - \beta - (H'H)^{-1}X'\mathbf{u} - (H'H)^{-1}W'\mathbf{u}^* \\
&= (X'X)^{-1}X'\mathbf{u} - (H'H)^{-1}X'\mathbf{u} - (H'H)^{-1}W'\mathbf{u}^* \\
&= [(X'X)^{-1} - (H'H)^{-1}]X'\mathbf{u} - (H'H)^{-1}W'\mathbf{u}^* \\
&= [M^{-1} - (H'H)^{-1}]X'\mathbf{u} - (H'H)^{-1}W'\mathbf{u}^*,
\end{aligned}
$$

e com isto,

$$
\begin{aligned}
(\hat{\beta} - \hat{\beta}_{\mathbf{z}})(\hat{\beta} - \hat{\beta}_{\mathbf{z}})' = {} & M^{-1}X'\mathbf{u}\mathbf{u}'XM^{-1} - (H'H)^{-1}X'\mathbf{u}\mathbf{u}'XM^{-1} \\
& -(H'H)^{-1}W'\mathbf{u}^*\mathbf{u}'XM^{-1} - M^{-1}X'\mathbf{u}\mathbf{u}'X(H'H)^{-1} \\
& +(H'H)^{-1}X'\mathbf{u}\mathbf{u}'X(H'H)^{-1} + (H'H)^{-1}W'\mathbf{u}^*\mathbf{u}'X(H'H)^{-1} \\
& -M^{-1}X'\mathbf{u}\mathbf{u}^{*'}W(H'H)^{-1} + (H'H)^{-1}X'\mathbf{u}\mathbf{u}^{*'}W(H'H)^{-1} \\
& +(H'H)^{-1}W'\mathbf{u}^*\mathbf{u}^{*'}W(H'H)^{-1}.
\end{aligned}
$$

Calculando a esperança dessa expressão, todos os termos que apresentam $\mathbf{u}$ e $\mathbf{u}^*$ juntos se anulam, pois segundo os modelos na Seção 3.1, $\mathbf{u}$ e $\mathbf{u}^*$ são independentes e com esperança nula. Assim

$$
\begin{aligned}
E(\hat{\beta} - \hat{\beta}_{\mathbf{z}})(\hat{\beta} - \hat{\beta}_{\mathbf{z}})' = {} & \sigma^2[M^{-1} - (H'H)^{-1} - (H'H)^{-1} + (H'H)^{-1}M(H'H)^{-1}] \\
& + \sigma^2(H'H)^{-1}W'W(H'H)^{-1}.
\end{aligned}
$$

Substituindo o segundo termo dentro do colchete pela expressão (i) do Lema A.1.1, temos

$$
E(\hat{\beta} - \hat{\beta}_{\mathbf{z}})(\hat{\beta} - \hat{\beta}_{\mathbf{z}})' = \sigma^2\,[M^{-1} - M^{-1} + M^{-1}W'(I + WM^{-1}W')^{-1}WM^{-1}]
$$

$$
\begin{aligned}
& - \sigma^2 (H'H)^{-1} + \sigma^2 (H'H)^{-1} M (H'H)^{-1} \\
& + \sigma^2 (H'H)^{-1} W'W (H'H)^{-1} \\
= \; & \sigma^2 M^{-1} W' (I + WM^{-1}W')^{-1} WM^{-1} \\
& - \sigma^2 (H'H)^{-1} MM^{-1} + \sigma^2 (H'H)^{-1} M (H'H)^{-1} \\
& + \sigma^2 (H'H)^{-1} W'W (H'H)^{-1} \\
= \; & \sigma^2 M^{-1} W' (I + WM^{-1}W')^{-1} WM^{-1} \\
& - \sigma^2 (H'H)^{-1} M \left[M^{-1} - (H'H)^{-1} \right] \\
& + \sigma^2 (H'H)^{-1} W'W (H'H)^{-1}.
\end{aligned}
$$

Aplicando novamente o resultado (i) do Lema A.1.1 no segundo termo da última expressão, segue que

$$
\begin{aligned}
E(\hat{\beta} - \hat{\beta}_z)(\hat{\beta} - \hat{\beta}_z)' \; = \; & \sigma^2 M^{-1} W' (I + WM^{-1}W')^{-1} WM^{-1} \\
& - \sigma^2 (H'H)^{-1} M \left[M^{-1} W' (I + WM^{-1}W')^{-1} WM^{-1} \right] \\
& + \sigma^2 (H'H)^{-1} W'W (H'H)^{-1} \\
= \; & \sigma^2 \left[I - (H'H)^{-1} M \right] M^{-1} W' (I + WM^{-1}W')^{-1} WM^{-1} \\
& + \sigma^2 (H'H)^{-1} W'W (H'H)^{-1} \\
= \; & \sigma^2 \left[M^{-1} - (H'H)^{-1} \right] MM^{-1} W' (I + WM^{-1}W')^{-1} WM^{-1} \\
& + \sigma^2 (H'H)^{-1} W'W (H'H)^{-1} \\
= \; & \sigma^2 \left[M^{-1} - (H'H)^{-1} \right] W' (I + WM^{-1}W')^{-1} WM^{-1} \\
& + \sigma^2 (H'H)^{-1} W'W (H'H)^{-1}.
\end{aligned}
$$

Finalizando, devido ao Lema A.1.1, *(i)* e *(iii)* e considerando que a matriz $M^{-1}W'W(H'H)^{-1}W'W(H'H)^{-1}$ é simétrica,

$$
\begin{aligned}
& E(\hat{\beta} - \hat{\beta}_z)(\hat{\beta} - \hat{\beta}_z)' = \\
= \; & \sigma^2 M^{-1} W' (I + WM^{-1}W')^{-1} WM^{-1} MM^{-1} W' (I + WM^{-1}W')^{-1} WM^{-1} \\
& + \sigma^2 (H'H)^{-1} W'W (H'H)^{-1} \\
= \; & \sigma^2 M^{-1} W' (I + WM^{-1}W')^{-1} WM^{-1} W' (I + WM^{-1}W')^{-1} WM^{-1} \\
& + \sigma^2 (H'H)^{-1} W'W (H'H)^{-1} \\
= \; & \sigma^2 M^{-1} W'W (H'H)^{-1} W' (I + WM^{-1}W')^{-1} WM^{-1} (W'W)(W'W)^{-1} \\
& + \sigma^2 (H'H)^{-1} W'W (H'H)^{-1} \\
= \; & \sigma^2 M^{-1} W'W (H'H)^{-1} W'W (H'H)^{-1} W'W (W'W)^{-1} \\
& + \sigma^2 (H'H)^{-1} W'W (H'H)^{-1} \\
= \; & \sigma^2 M^{-1} W'W (H'H)^{-1} W'W (H'H)^{-1} \\
& + \sigma^2 (H'H)^{-1} W'W (H'H)^{-1}
\end{aligned}
$$

$$= \sigma^2 (H'H)^{-1}(W'W)M^{-1}(W'W)(H'H)^{-1} + \sigma^2 (H'H)^{-1}W'W(H'H)^{-1}.$$

Verificamos que essa última expressão é $O(n^{-2})$ pois $M^{-1} = O(n^{-1})$ por hipótese, $W'W = O(1)$ e $(H'H)^{-1} = (M + W'W)^{-1} = O(n^{-1})$.

Logo, cada elemento de $E(\hat{\beta} - \hat{\beta}_z)(\hat{\beta} - \hat{\beta}_z)'$ é tal que

$$E(\hat{\beta}_i - \hat{\beta}_{zi})(\hat{\beta}_j - \hat{\beta}_{zj})' = O(n^{-2})$$

para todo $i,\ j = 1, \cdots, p$.

Em particular, para todo $i = j$, temos

$$E(\hat{\beta}_i - \hat{\beta}_{zi})^2 = O(n^{-2}) = O(n^{-1})^2,$$

e assim, pelo Teorema A.3.4,

$$(\hat{\beta}_i - \hat{\beta}_{zi}) = O_p(n^{-1}),$$

para todo $i = 1, \cdots, p$.

Portanto, pelo Teorema A.3.3,

$$\hat{\beta} - \hat{\beta}_z = O_p(n^{-1}). \qquad \square$$

(iv) Inicialmente, vamos mostrar que

$$(\mathbf{Z} - W\hat{\beta}) \ = \ O_p(1).$$

Do Resultado 3-*(ii)* temos

$$(\hat{\beta} - \hat{\beta}_z) = (H'H)^{-1}W'(W\hat{\beta} - \mathbf{Z}).$$

Também verificamos que $W = O(1)$, $(H'H)^{-1} = O(n^{-1})$ e do item **(iii)** temos que $\hat{\beta} - \hat{\beta}_z = O_p(n^{-1})$. Como consequência, temos que

$$O(n^{-1})(W\hat{\beta} - \mathbf{Z}) = O_p(n^{-1}),$$

do que concluimos que

$$(\mathbf{Z} - W\hat{\beta}) = O_p(1).$$

Usando (A.3),

$$\begin{aligned}\hat{\sigma}_{\mathbf{z}}^2 &= (n+m)^{-1}\left[n\hat{\sigma}^2 + (\mathbf{Z}-W\hat{\beta})'(I+WM^{-1}W')^{-1}(\mathbf{Z}-W\hat{\beta})\right] \\ &= (n+m)^{-1}n\hat{\sigma}^2 + (n+m)^{-1}(\mathbf{Z}-W\hat{\beta})'(I+WM^{-1}W')^{-1}(\mathbf{Z}-W\hat{\beta})\end{aligned}$$

e com isso,

$$\hat{\sigma}_{\mathbf{z}}^2 - (n+m)^{-1}n\hat{\sigma}^2 = (n+m)^{-1}(\mathbf{Z}-W\hat{\beta})'(I+WM^{-1}W')^{-1}(\mathbf{Z}-W\hat{\beta}).$$

Do fato que $(I+WM^{-1}W')^{-1} = \mathsf{O}(1)$, temos que

$$\begin{aligned}\hat{\sigma}_{\mathbf{z}}^2 - (n+m)^{-1}n\hat{\sigma}^2 &= \mathsf{O}(n^{-1})\,\mathsf{O}_p(1)\,\mathsf{O}(1)\,\mathsf{O}_p(1) \\ &= \mathsf{O}_p(n^{-1}),\end{aligned}$$

e como consequência

$$\hat{\sigma}_{\mathbf{z}}^2 - \hat{\sigma}^2 = \mathsf{O}_p(n^{-1}). \qquad \square$$

O próximo teorema corresponde ao Teorema 3.3.2 do Capítulo 3.

Teorema A.3.6 *(LEVY e PERNG (1984))*
Sejam $\mathbf{Y}$, X, $\mathbf{Z}$ *e* W, *vetores e matrizes como definidos anteriormente. Se* $(X'X)/n \to D$ *quando* $n \to \infty$, D *positiva definida, então*

$$LL^*(\mathbf{z}, W/\mathbf{Y}, X) = q(\mathbf{z}, W/\mathbf{Y}, X)(1+\mathsf{O}_p(n^{-1})).$$

Prova: Seja

$$U_n = \frac{(\mathbf{Z}-W\hat{\beta})'(\mathbf{Z}-W\hat{\beta})}{\hat{\sigma}^2}. \tag{A.15}$$

Verifica-se (*ELIAN (1991) [pág. 210]*) que se

$$\hat{\sigma}^2 - \sigma^2 = \mathsf{O}_p(n^{-\frac{1}{2}})$$

então

$$\frac{1}{\hat{\sigma}^2} - \frac{1}{\sigma^2} = \mathsf{O}_p(n^{-\frac{1}{2}}),$$

o que implica em

$$\begin{aligned}\frac{1}{\hat{\sigma}^2} &= \mathsf{O}_p(n^{-\frac{1}{2}}) + \frac{1}{\sigma^2} \\ &= \mathsf{O}_p(n^{-\frac{1}{2}}) + \mathsf{O}_p(1) \\ &= \mathsf{O}_p(1). \end{aligned} \tag{A.16}$$

Assim, substituindo em (A.15) e do fato que $(\mathbf{Z} - W\hat{\beta}) = \mathsf{O}_p(1)$ temos que

$$U_n = \mathsf{O}_p(1)\,\mathsf{O}_p(1)\,\mathsf{O}_p(1) = \mathsf{O}_p(1).$$

Portanto, de (3.12) e (3.13),

$$\begin{aligned}\frac{LL^*(\mathbf{z}, W/\mathbf{Y}, X)}{q(\mathbf{z}, W/\mathbf{Y}, X)} &= \frac{(\hat{\sigma}^2)^{\frac{n}{2}}(2\pi\ e)^{-\frac{m}{2}}(\hat{\sigma}_{\mathbf{z}}^2)^{-\frac{n+m}{2}}}{(2\pi\hat{\sigma}^2)^{-\frac{m}{2}}\exp\{-(2\hat{\sigma}^2)^{-1}(\mathbf{z} - W\hat{\beta})'(\mathbf{z} - W\hat{\beta})\}} \\ &= (\hat{\sigma}^2)^{\frac{n+m}{2}}(\hat{\sigma}_{\mathbf{z}}^2)^{-\frac{n+m}{2}}\exp\{-\frac{1}{2}[m - \frac{(\mathbf{z} - W\hat{\beta})'(\mathbf{z} - W\hat{\beta})}{\hat{\sigma}^2}]\} \\ &= (\hat{\sigma}^2)^{\frac{n+m}{2}}(\hat{\sigma}_{\mathbf{z}}^2)^{-\frac{n+m}{2}}\exp\{-\frac{1}{2}[m - U_n]\} \\ &= \left(\frac{\hat{\sigma}_{\mathbf{z}}^2}{\hat{\sigma}^2}\right)^{-\frac{n+m}{2}}\exp\{-\frac{1}{2}[m - U_n]\}. \end{aligned}$$

Resta mostrar que esta última expressão é $(1 + \mathsf{O}_p(n^{-1}))$. Da expressão (A.3) e do Lema A.1.1-*(ii)*,

$$\begin{aligned}\hat{\sigma}_{\mathbf{z}}^2 &= (n+m)^{-1}\left[n\hat{\sigma}^2 + (\mathbf{Z} - W\hat{\beta})'(I + WM^{-1}W')^{-1}(\mathbf{Z} - W\hat{\beta})\right] \\ &= (n+m)^{-1}\left[n\hat{\sigma}^2 + (\mathbf{Z} - W\hat{\beta})'(\mathbf{Z} - W\hat{\beta})\right] - \\ &\quad (n+m)^{-1}\left[(\mathbf{Z} - W\hat{\beta})'(I + WM^{-1}W')^{-1}WM^{-1}W'(\mathbf{Z} - W\hat{\beta})\right]. \end{aligned}$$

Do Lema A.1.1-*(iii)*

$$\begin{aligned}\hat{\sigma}_{\mathbf{z}}^2 &= (n+m)^{-1}\left[n\hat{\sigma}^2 + m\hat{\sigma}^2 - m\hat{\sigma}^2 + \hat{\sigma}^2 U_n - \hat{\sigma}^2\frac{(\mathbf{Z} - W\hat{\beta})'W(H'H)^{-1}W'(\mathbf{Z} - W\hat{\beta})}{\hat{\sigma}^2}\right] \\ &= (n+m)^{-1}\left[\hat{\sigma}^2(n+m) - \hat{\sigma}^2[m - U_n] - \hat{\sigma}^2\frac{(\mathbf{Z} - W\hat{\beta})'W(H'H)^{-1}W'(\mathbf{Z} - W\hat{\beta})}{\hat{\sigma}^2}\right] \\ &= \hat{\sigma}^2(n+m)^{-1}\left[(n+m) - [m - U_n] - \frac{(\mathbf{Z} - W\hat{\beta})'W(H'H)^{-1}W'(\mathbf{Z} - W\hat{\beta})}{\hat{\sigma}^2}\right] \end{aligned}$$

$$= \hat{\sigma}^2 \left[1 - (n+m)^{-1}[m - U_n] - (n+m)^{-1} \frac{(\mathbf{Z} - W\hat{\beta})' W (H'H)^{-1} W' (\mathbf{Z} - W\hat{\beta})}{\hat{\sigma}^2} \right]$$

$$= \hat{\sigma}^2 \left[1 - (n+m)^{-1}[m - U_n] - (n+m)^{-1} \frac{1}{\hat{\sigma}^2} (\mathbf{Z} - W\hat{\beta})' W (H'H)^{-1} W' (\mathbf{Z} - W\hat{\beta}) \right].$$

Usando (A.16) e o fato que $W(H'H)^{-1}W' = \mathsf{O}(n^{-1})$, obtemos

$$\begin{aligned} \hat{\sigma}_{\mathbf{z}}^2 &= \hat{\sigma}^2 \left[1 - (n+m)^{-1}[m - U_n] + \mathsf{O}(n^{-1})\, \mathsf{O}_p(1)\, \mathsf{O}(n^{-1})\, \mathsf{O}_p(1) \right] \\ &= \hat{\sigma}^2 \left[1 - (n+m)^{-1}[m - U_n] + \mathsf{O}_p(n^{-2}) \right]. \end{aligned}$$

Assim,

$$\left(\frac{\hat{\sigma}_{\mathbf{z}}^2}{\hat{\sigma}^2} \right) = \left[1 - (n+m)^{-1}[m - U_n] \right] + \mathsf{O}_p(n^{-2}).$$

Se definirmos $X_n = (n+m)\, \mathsf{O}_p(n^{-2}) = \mathsf{O}_p(n^{-1})$, temos

$$\left(\frac{\hat{\sigma}_{\mathbf{z}}^2}{\hat{\sigma}^2} \right) = 1 - (n+m)^{-1}[m - U_n + X_n]$$

e

$$\left(\frac{\hat{\sigma}_{\mathbf{z}}^2}{\hat{\sigma}^2} \right)^{-\frac{n+m}{2}} = \left[1 - (n+m)^{-1}(m - U_n + X_n) \right]^{-\frac{n+m}{2}}.$$

Multiplicando essa expressão por $e^{-\frac{1}{2}[m-U_n]}$, segue que

$$\left(\frac{\hat{\sigma}_{\mathbf{z}}^2}{\hat{\sigma}^2} \right)^{-\frac{n+m}{2}} e^{-\frac{1}{2}(m-U_n)} = \left[1 - (n+m)^{-1}[m - U_n + X_n] \right]^{-\frac{n+m}{2}} e^{-\frac{1}{2}[m-U_n]}$$

e tomando logaritmo,

$$\begin{aligned} \ln \left[\left(\frac{\hat{\sigma}_{\mathbf{z}}^2}{\hat{\sigma}^2} \right)^{-\frac{n+m}{2}} e^{-\frac{1}{2}[m-U_n]} \right] &= -\frac{(n+m)}{2} \ln \left[1 - (n+m)^{-1}[m - U_n + X_n] \right] \\ &\quad - \frac{1}{2}[m - U_n]. \end{aligned} \tag{A.17}$$

Da expansão de $f(x) = \ln\left[1-(n+m)^{-1}[m-U_n+X_n]\right]$ em série de Taylor ao redor de $x_o = 1$, obtemos

$$\begin{aligned}
\ln(x) &= \frac{\ln^{(0)}(x_o)}{0!}(x-x_o)^0 + \frac{\ln^{(1)}(x_o)}{1!}(x-x_o)^1 + \frac{\ln^{(2)}(x_o)}{2!}(x-x_o)^2 + o_p(x-x_o)^2 \\
&= \frac{\ln(x)|_{(x=1)}}{0!}(x-1)^0 + \frac{\frac{1}{x}|_{(x=1)}}{1!}(x-1)^1 + \frac{-\frac{1}{x^2}|_{(x=1)}}{2!}(x-x_o)^2 + o_p(x-1)^2 \\
&= (x-1) - \frac{1}{2}(x-1)^2 + o_p(x-1)^2.
\end{aligned}$$

Tomando $x = \left[1-(n+m)^{-1}[m-U_n+X_n]\right]$, esta expansão fica

$$\begin{aligned}
&\ln\left[1-(n+m)^{-1}[m-U_n+X_n]\right] = \\
&= -(n+m)^{-1}(m-U_n+X_n) - \frac{1}{2}(n+m)^{-2}(m-U_n+X_n)^2 \\
&\quad + o_p((n+m)^{-1}(m-U_n+X_n))^2 \\
&= -\left(\frac{m-U_n+X_n}{n+m}\right) - \frac{1}{2}\left(\frac{m-U_n+X_n}{n+m}\right)^2 + o_p\left(\frac{m-U_n+X_n}{n+m}\right)^2.
\end{aligned}$$

Substituindo esta última expressão em (A.17) e como $U_n = O_p(1)$ e $X_n = O_p(n^{-1})$ implica em $(m-U_n+X_n) = O_p(1)$, segue que

$$\begin{aligned}
&\ln\left[\left(\frac{\hat{\sigma}_z^2}{\hat{\sigma}^2}\right)^{-\frac{n+m}{2}} \exp\left[-\tfrac{1}{2}(m-U_n)\right]\right] = \\
&= -\frac{(n+m)}{2}\left[-\left(\frac{m-U_n+X_n}{n+m}\right) - \frac{1}{2}\left(\frac{m-U_n+X_n}{n+m}\right)^2 + o_p\left(\frac{m-U_n+X_n}{n+m}\right)^2\right] \\
&\quad - \frac{1}{2}(m-U_n) \\
&= \left[\frac{1}{2}(m-U_n+X_n) + \frac{1}{4(n+m)}(m-U_n+X_n)^2 - \frac{n+m}{2}o_p\left(\frac{m-U_n+X_n}{n+m}\right)^2\right] \\
&\quad - \frac{1}{2}(m-U_n) \\
&= \frac{1}{2}X_n + \frac{1}{4(n+m)}(m-U_n+X_n)^2 - \frac{n+m}{2}o_p\left(\frac{m-U_n+X_n}{n+m}\right)^2 \\
&= O_p(n^{-1}) + O_p(n^{-1}) + o_p(n^{-1}) \\
&= O_p(n^{-1}).
\end{aligned}$$

Como consequência,

$$\left(\frac{\hat{\sigma}_z^2}{\hat{\sigma}^2}\right)^{-\frac{n+m}{2}} e^{-\frac{1}{2}(m-U_n)} = \exp[O_p(n^{-1})].$$

Expandindo e^x em série de Taylor ao redor de $x_o = 0$ e calculando em $x = O_p(n^{-1})$ tem-se

$$\begin{aligned}\left(\frac{\hat{\sigma}_z^2}{\hat{\sigma}^2}\right)^{-\frac{n+m}{2}} e^{-\frac{1}{2}(m-U_n)} &= \frac{\exp^{(0)}\{0\}\ (O_p(n^{-1})-0)^0}{0!} + O_p(O_p(n^{-1})-0)^1 \\ &= 1 + O_p(n^{-1}),\end{aligned}$$

o que prova o teorema. □

A.4 Resultados Auxiliares para a Obtenção da Densidade Preditiva Ótima

Derivada direcional: Sejam $F : U \to \mathbb{R}$ definida no aberto U $\subset \mathbb{R}^p$, $a \in U$ e $\mathbf{v} \in \mathbb{R}^p$. A **derivada direcional** de F no ponto **a**, segundo o vetor **v**, é por definição o limite

$$\frac{\partial F(\mathbf{a})}{\partial \mathbf{v}} = \lim_{t\to 0} \frac{F(\mathbf{a}+t\mathbf{v}) - F(\mathbf{a})}{t}, \tag{A.18}$$

quando tal limite existe.

Diferencial de uma função:
(i) Seja $F : U \to \mathbb{R}$ definida no aberto $U \subset \mathbb{R}^p$, diferenciável em todo U em particular no ponto $\mathbf{a} \in U$. A *diferencial de F no ponto* **a** é o *funcional linear*

$$dF(\mathbf{a}) : \mathbb{R}^p \to \mathbb{R}$$

cujo valor aplicado no vetor $\mathbf{v} = (v_1, v_2, \ldots, v_p)$ é dado por

$$dF(\mathbf{a}).\mathbf{v} = \left(\frac{\partial F(\mathbf{a})}{\partial x_1}, \ldots, \frac{\partial F(\mathbf{a})}{\partial x_p}\right).\mathbf{v}$$

$$= \left(\frac{\partial F(\mathbf{a})}{\partial x_1}, \ldots, \frac{\partial F(\mathbf{a})}{\partial x_p}\right) \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_p \end{pmatrix}$$

$$= \sum_{i=1}^{p} \frac{\partial F(\mathbf{a})}{\partial x_i} . v_i$$

$$= \frac{\partial F(\mathbf{a})}{\partial \mathbf{v}}. \qquad \text{(A.19)}$$

(ii) A aplicação dF é contínua se e somente se, cada uma das suas funções cordenadas $\partial F/\partial x_i : U \to \mathbb{R}$ é contínua, isto é, se e somente se, F é de classe $\mathcal{C}^1$.

Teorema do Valor Médio: Seja $F : U \to \mathbb{R}$ diferenciável em todos os pontos de segmento de reta aberto $]\mathbf{a}, \mathbf{a}+\mathbf{v}[$ tal que sua restrição ao segmento fechado $[\mathbf{a}, \mathbf{a}+\mathbf{v}] \subset U \subset \mathbb{R}^p$ seja contínua. Nessas condições, $\theta \in]0,1[$ tal que

$$\begin{aligned} F(\mathbf{a}+\mathbf{v}) - F(\mathbf{a}) &= dF(\mathbf{a}+\theta\mathbf{v}).\mathbf{v} \\ &= \frac{\partial F}{\partial \mathbf{v}}(\mathbf{a}+\theta\mathbf{v}) \\ &= \sum_{i=1}^{p} \frac{\partial F}{\partial x_i}(\mathbf{a}+\theta\mathbf{v}) v_i. \end{aligned}$$

Para $p = 2$, $(x_1, x_2) = (x, y)$, $\mathbf{a} = (x, \bar{y})$ onde $\bar{y}$ é uma função de x e $\mathbf{v} = (0, th)$, esta expressão fica

$$\begin{aligned} F(\mathbf{a}+\mathbf{v}) - F(\mathbf{a}) &= dF(\mathbf{a}+\theta\mathbf{v}).\mathbf{v} \\ &= \frac{\partial F}{\partial \mathbf{v}}(\mathbf{a}+\theta\mathbf{v}) \\ &= \sum_{i=1}^{2} \frac{\partial F}{\partial x_i}(\mathbf{a}+\theta\mathbf{v}).v_i \\ &= \frac{\partial F}{\partial x_1}(\mathbf{a}+\theta\mathbf{v})v_1 + \frac{\partial F}{\partial x_2}(\mathbf{a}+\theta\mathbf{v})v_2 \\ &= \frac{\partial F}{\partial x}((x,\bar{y})+\theta(0,th)).0 + \frac{\partial F}{\partial y}((x,\bar{y})+\theta(0,th)).th \\ &= \frac{\partial F}{\partial y}((x,\bar{y})+\theta(0,th)).th \\ &= \frac{\partial F}{\partial y}(x, \bar{y}+\theta th).th\,. \qquad \text{(A.20)} \end{aligned}$$

Então

$$\begin{aligned}\lim_{t\to 0}\frac{F(\mathbf{a}+\mathbf{v})-F(\mathbf{a})}{t} &= \lim_{t\to 0}\frac{\frac{\partial F}{\partial y}(x,\bar{y}+\theta th).th}{t}\\ &= \lim_{t\to 0}\frac{\partial F}{\partial y}(x,\bar{y}+\theta th).h\\ &= \frac{\partial F}{\partial y}(x,\bar{y}).h\,. \qquad \text{(A.21)}\end{aligned}$$

A.5 Resultados para a Obtenção das Funções de Influência Preditivas

Para os modelos lineares em (4.1) e (4.2), quando θ é conhecido, com distribuição a priori não informativa

$$f(\beta,\theta) \quad \propto \quad \theta^{-1}$$

a densidade a posteriori de (β,θ) dado $\mathbf{Y}=\mathbf{y}$ será

$$f(\beta,\theta/X,\mathbf{y}) = \frac{f(\beta,\theta)p_n(\mathbf{y},X/\beta,\theta)}{\int_{\Omega_\beta} f(\beta,\theta)\,p_n(\mathbf{y},X/\beta,\theta)\,d\beta},$$

onde $p_n(\mathbf{y},X/\beta,\theta)$ é a densidade da normal multivariada definida em (3.3). Logo, a densidade preditiva de $\mathbf{Z}$ será

$$f(\mathbf{z}/W,X,\mathbf{y}) = \frac{\int_{\Omega_\beta} f(\mathbf{z}/W,\beta,\theta)\,f(\beta,\theta)\,p_n(\mathbf{y},X/\beta,\theta)\,d\beta}{\int_{\Omega_\beta} f(\beta,\theta)\,p_n(\mathbf{y},X/\beta,\theta)\,d\beta}.$$

Como $f(\beta,\theta)\propto\theta^{-1}$ é independente de β, então esta expressão fica

$$\begin{aligned}f(\mathbf{z}/W,X,\mathbf{y}) &= \frac{\int_{\Omega_\beta} f(\mathbf{z}/W,\beta,\theta)\,p_n(\mathbf{y},X/\beta,\theta)\,d\beta}{\int_{\Omega_\beta} p_n(\mathbf{y},X/\beta,\theta)\,d\beta}\\ &= \frac{\int_{\Omega_\beta} p_m(\mathbf{z},W/\beta,\theta)\,p_n(\mathbf{y},X/\beta,\theta)\,d\beta}{\int_{\Omega_\beta} p_n(\mathbf{y},X/\beta,\theta)\,d\beta}\\ &= \frac{\int_{\Omega_\beta} p_{n+m}((\mathbf{y}'\mathbf{z}')',(X'W')'/\beta,\theta)\,d\beta}{\int_{\Omega_\beta} p_n(\mathbf{y},X/\beta,\theta)\,d\beta},\end{aligned}$$

sendo $p_n(\cdot/\beta,\theta)$, $p_m(\cdot/\beta,\theta)$ e $p_{n+m}(\cdot/\beta,\theta)$ as densidades normais multivariadas de $\mathbf{Y}$, $\mathbf{Z}$ e $(\mathbf{Y},\mathbf{Z})$ respectivamente. Lembrando que $\mathbf{v} = (\mathbf{y}', \mathbf{z}')'$ e $H = (X'\ W')'$,

$$\begin{aligned} f(\mathbf{z}/W,X,\mathbf{y}) &= \frac{\int_{\Omega_\beta}(2\pi\theta)^{-\frac{n+m}{2}} exp\{-(2\theta)^{-1}(\mathbf{v}-H\beta)'(\mathbf{v}-H\beta)\}\,d\beta}{\int_{\Omega_\beta}(2\pi\theta)^{-\frac{n}{2}} exp\{-(2\theta)^{-1}(\mathbf{y}-X\beta)'(\mathbf{y}-X\beta)\}\,d\beta} \\ &= \frac{\int_{\Omega_\beta}(2\pi\theta)^{-\frac{n+m}{2}} exp\{-(2\theta)^{-1}[\mathbf{v}'\mathbf{v}-2\mathbf{v}'H\beta+\beta'H'H\beta]\}\,d\beta}{\int_{\Omega_\beta}(2\pi\theta)^{-\frac{n}{2}} exp\{-(2\theta)^{-1}[\mathbf{y}'\mathbf{y}-2\mathbf{y}'X\beta+\beta'X'X\beta]\}\,d\beta} \\ &= \frac{\int_{\Omega_\beta}(2\pi\theta)^{-\frac{n+m}{2}} exp\{-[\frac{\beta'H'H\beta}{2\theta}-\frac{2\mathbf{v}'H\beta}{2\theta}+\frac{\mathbf{v}'\mathbf{v}}{2\theta}]\}\,d\beta}{\int_{\Omega_\beta}(2\pi\theta)^{-\frac{n}{2}} exp\{-[\frac{\beta'X'X\beta}{2\theta}-\frac{2\mathbf{y}'X\beta}{2\theta}+\frac{\mathbf{y}'\mathbf{y}}{2\theta}]\}d\beta}. \end{aligned}$$

Para calcular esta última integral, utilizamos o Resultado 9 e analogamente aos cálculos feitos na Seção A.2, obtemos

$$f(\mathbf{z}/W,X,\mathbf{y}) = \frac{(2\pi\theta)^{\frac{p-(n+m)}{2}}|H'H|^{-\frac{1}{2}} exp\{-[(2\theta)^{-1}(\mathbf{v}'\mathbf{v}-\mathbf{v}'H(H'H)^{-1}H'\mathbf{v})]\}}{(2\pi\theta)^{\frac{p-n}{2}}|X'X|^{-\frac{1}{2}} exp\{-[(2\theta)^{-1}(\mathbf{y}'\mathbf{y}-\mathbf{y}'X(X'X)^{-1}X'\mathbf{y})]\}}.$$

Como $M = X'X$ e $H'H = M + W'W$, utilizando (A.10), esta expressão fica

$$f(\mathbf{z}/W,X,\mathbf{y}) = \frac{|M|^{-\frac{1}{2}}|I+WM^{-1}W'|^{-\frac{1}{2}} exp\{-[(2\theta)^{-1}(\mathbf{v}'\mathbf{v}-\mathbf{v}'H(H'H)^{-1}H'\mathbf{v})]\}}{(2\pi)^{\frac{m}{2}}\theta^{\frac{m}{2}}|M|^{-\frac{1}{2}} exp\{-[(2\theta)^{-1}(\mathbf{y}'\mathbf{y}-\mathbf{y}'X(X'X)^{-1}X'\mathbf{y})]\}}.$$

Para as expressões nas exponencias, utilizaremos ainda as identidades obtidas na Seção A.2:

$$\mathbf{v}'\mathbf{v}-\mathbf{v}'H(H'H)^{-1}H'\mathbf{v} = n\hat{\sigma}^2 + (\mathbf{z}-W\hat{\beta})'(I+WM^{-1}W')^{-1}(\mathbf{z}-W\hat{\beta})$$

e

$$\mathbf{y}'\mathbf{y}-\mathbf{y}'X(X'X)^{-1}X'\mathbf{y} = n\hat{\sigma}^2,$$

no numerador e denominador respectivamente, resultando em

$$\begin{aligned} f(\mathbf{z}/W,X,\mathbf{y}) &= \frac{|M|^{-\frac{1}{2}}|I+WM^{-1}W'|^{-\frac{1}{2}}}{(2\pi)^{\frac{m}{2}}\theta^{\frac{m}{2}}|M|^{-\frac{1}{2}} exp\{-[(2\theta)^{-1}n\hat{\sigma}^2]\}} \\ &\times\ exp\{-[(2\theta)^{-1}(n\hat{\sigma}^2+(\mathbf{z}-W\hat{\beta})'(I+WM^{-1}W')^{-1}(\mathbf{z}-W\hat{\beta}))]\} \\ &= \frac{1}{(2\pi)^{\frac{m}{2}}\theta^{\frac{m}{2}}|I+WM^{-1}W'|^{\frac{1}{2}}} \\ &\times\ exp\{-[(2)^{-1}(\mathbf{z}-W\hat{\beta})'[(I+WM^{-1}W')\theta]^{-1}(\mathbf{z}-W\hat{\beta})]\}, \end{aligned} \tag{A.22}$$

que é a densidade normal multivariada dada na expressão (4.3).

Os próximos resultados são fundamentais na derivação das funções de influência preditivas do Capítulo 4.

Resultado 10 *(GRAYBILL (1976))*
Se $\mathbf{X}$ é um vetor aleatório com $E(\mathbf{X}) = \mathbf{u}$, $V(\mathbf{X}) = \Sigma$ e A uma matriz simétrica, então

$$E(\mathbf{X}'A\mathbf{X}) = tr(A\Sigma) + \mathbf{u}'A\mathbf{u}.$$

Resultado 11
Sejam f_1 e f_2 densidades de vetores aleatórios com distribuição normal multivariada com média $\mathbf{u}_i$ e matriz de covariância Σ_i, positiva definida $i = 1, 2$. Nessas condições, a divergência de Kullback-Leibler entre f_1 e f_2, $I(f_1, f_2)$, é tal que

$$2I(f_1, f_2) = (\mathbf{u}_1 - \mathbf{u}_2)'\Sigma_2^{-1}(\mathbf{u}_1 - \mathbf{u}_2) + \{tr\Sigma_1\Sigma_2^{-1} - ln|\Sigma_1\Sigma_2^{-1}| - n\}$$

Prova : Sejam f_1 a densidade $N_n(\mathbf{u}_1, \Sigma_1)$ e f_2 $N_n(\mathbf{u}_2, \Sigma_2)$, de modo que

$$f_1(\mathbf{y}) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} |\Sigma_1|^{-\frac{1}{2}} exp\{-\frac{1}{2}(\mathbf{y} - \mathbf{u}_1)'\Sigma_1^{-1}(\mathbf{y} - \mathbf{u}_1)\},$$

$$f_2(\mathbf{y}) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} |\Sigma_2|^{-\frac{1}{2}} exp\{-\frac{1}{2}(\mathbf{y} - \mathbf{u}_2)'\Sigma_2^{-1}(\mathbf{y} - \mathbf{u}_2)\}$$

e

$$\begin{aligned} ln\frac{f_1(\mathbf{y})}{f_2(\mathbf{y})} &= -\frac{n}{2}ln(2\pi) - \frac{1}{2}ln|\Sigma_1| - \frac{1}{2}(\mathbf{y} - \mathbf{u}_1)'\Sigma_1^{-1}(\mathbf{y} - \mathbf{u}_1) \\ &\quad + \frac{n}{2}ln(2\pi) + \frac{1}{2}ln|\Sigma_2^{-1}| + \frac{1}{2}(\mathbf{y} - \mathbf{u}_2)'\Sigma_2(\mathbf{y} - \mathbf{u}_2) \\ &= \frac{1}{2}ln\frac{|\Sigma_2|}{|\Sigma_1|} + \frac{1}{2}(\mathbf{y} - \mathbf{u}_2)'\Sigma_2^{-1}(\mathbf{y} - \mathbf{u}_2) \\ &\quad - \frac{1}{2}(\mathbf{y} - \mathbf{u}_1)'\Sigma_1^{-1}(\mathbf{y} - \mathbf{u}_1). \end{aligned}$$

De acordo com a definição, $I(f_1, f_2) = E_{f_1}[ln\frac{f_1}{f_2}]$ e portanto, tomando esperança com respeito à f_1,

$$E_{f_1}\left(ln\frac{f_1}{f_2}\right) = \int ln\frac{f_1(\mathbf{y})}{f_2(\mathbf{y})}f_1(\mathbf{y})d\mathbf{y}$$
$$= \frac{1}{2}ln\frac{|\Sigma_2|}{|\Sigma_1|} + \frac{1}{2}E_{f_1}\left[(\mathbf{Y}-\mathbf{u}_2)'\Sigma_2^{-1}(\mathbf{Y}-\mathbf{u}_2) - (\mathbf{Y}-\mathbf{u}_1)'\Sigma_1^{-1}(\mathbf{Y}-\mathbf{u}_1)\right].$$

Como, sob f_1, $\mathbf{Y} \sim \mathcal{N}(\mathbf{u}_1, \Sigma_1)$, então $(\mathbf{Y}-\mathbf{u}_1) \sim \mathcal{N}(\mathbf{0}, \Sigma_1)$ e $(\mathbf{Y}-\mathbf{u}_2) \sim \mathcal{N}(\mathbf{u}_1-\mathbf{u}_2, \Sigma_1)$ e assim, devido ao Resultado 10,

$$E_{f_1}\left[(\mathbf{Y}-\mathbf{u}_2)'\Sigma_2^{-1}(\mathbf{Y}-\mathbf{u}_2)\right] = tr(\Sigma_2^{-1}\Sigma_1) + (\mathbf{u}_1-\mathbf{u}_2)'\Sigma_2^{-1}(\mathbf{u}_1-\mathbf{u}_2)$$

e

$$E_{f_1}\left[(\mathbf{Y}-\mathbf{u}_1)'\Sigma_1^{-1}(\mathbf{Y}-\mathbf{u}_1)\right] = tr(\Sigma_1^{-1}\Sigma_1) = n.$$

Portanto,

$$E_{f_1}\left(ln\frac{f_1}{f_2}\right) = \frac{1}{2}ln|\Sigma_2\,\Sigma_1|^{-1} + \frac{1}{2}\left[(\mathbf{u}_1-\mathbf{u}_2)'\Sigma_2^{-1}(\mathbf{u}_1-\mathbf{u}_2) + tr(\Sigma_2^{-1}\Sigma_1) - n\right]$$
$$= \frac{1}{2}(\mathbf{u}_1-\mathbf{u}_2)'\Sigma_2^{-1}(\mathbf{u}_1-\mathbf{u}_2) + \frac{1}{2}\left\{tr(\Sigma_2^{-1}\Sigma_1) - ln|\Sigma_1\,\Sigma_2^{-1}| - n\right\}.$$

Como Σ_1 e Σ_2 são simétricas de mesma ordem, $tr(\Sigma_2^{-1}\Sigma_1) = tr(\Sigma_1\Sigma_2^{-1})$, o que encerra a demonstração. □

Resultado 12 (JOHNSON e GEISSER (1983))
Seja $H = X(X'X)^{-1}X'$, a matriz *Hat*. Nestas condições, valem os seguintes resultados:

(i) $|I+H| = 2^p$.

(ii) $(I+H)^{-1} = I - \frac{1}{2}H$.

(iii) $|I+H^{(i)}| = 2^p|I+\frac{1}{2}V_i(I-V_i)^{-1}|$.

(iv) $tr[(I+H)^{-1}(I+H^{(i)})] = n + \frac{1}{2}trV_i(I-V_i)^{-1}$.

(v) $I - X(S+S_{(i)})^{-1}X' = (I+H^{(i)})^{-1}$.

(vi) $(I+H^{(i)})^{-1} = I - \frac{1}{2}H - \frac{1}{2}XS^{-1}X_i'(I-\frac{1}{2}V_i)^{-1}X_iS^{-1}X'$.

(vii) $tr[(I+H^{(i)})^{-1}(I+H)] = n - \frac{1}{2}trV_i(I-\frac{1}{2}V_i)^{-1}$.

Referências Bibliográficas

Aitchison, J. (1975). *Goodness of Prediction Fit.* Biometrika, 62: 547-554.

Berger J. O. e Wolper, R. L. (1988). *The Likelihood Principle.* 2. ed. Hayward. 208 p.

Bingham, C. O. (1977). *Some Identities Useful in the Analysis of Residuals From Linear Regression.* Technical Report Nro. 300, School of Statistics, University of Minnesota. 16 p.

Bjornstad, J. F. (1990). *Predictive Likelihood: a Review.* Statistical Science, 5(1): 242-265.

Box, G. E. P. e Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Reading, MA: Addison-Wesley. 588 p.

Butler, R. W. (1986). *Predictive Likelihood Inference with Applications.* Journal of the Royal Statistical Society, Series B, 48(1): 1-38.

Butler, R. W. (1989). *Approximate Predictive Pivots and Densities.* Biometrika, 76: 489-501.

Cook, R. D. (1977). *Detection of Influential Observations in Linear Regression.* Technometrics, 19: 15-18.

Cook R. D. e Weisberg S. (1980). *Characterizations of an Empirical Influence Function for Detecting Influencial Cases in Regression.* Technometrics, 22: 495-508.

Cook R. D. e Weisberg S. (1982). *Residuals and Influence in Regression.* New York: Chapman and Hall. 230 p.

Draper, N. e Smith, H. (1981). *Applied Regression Analysis.* New York: Wiley. 709 p.

Elian, S. N. (1991). *Análise de Regressão em Populações Finitas.* Sâo Paulo. 229 p. Tese - Doutorado, IME-USP.

Faulkenberry, G. D. (1973). *A Method of Obtaining Prediction Intervals.* Journal of the American Statistical Association, 68: 433-435.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference.* Edinburgh, Oliver and Boyd, London. 175 p.

Fraser, D.A.S. e Guttman, I. (1956). *Tolerance Regions.* Annals of Mathematical Statistics, 27: 162-179.

Fuller, W. A. (1976). *Introduction to Statistical Time Series.* New York: John Wiley. 470 p.

Gamerman, D. e Migon, H. S. (1997). *Inferência Estatística: Uma Abordagem Integrada.* Rio de Janeiro, IM-UFRJ. 206 p.

Geisser, S. (1964). *Posterior Odds for Multivariate Normal Classifications.* Journal of the Royal Statistical Society, Series B, 26: 69-76.

Geisser, S. (1965). *Bayesian Estimation in Multivariate Analysis.* Annals of Mathematical Statistics, 36: 150-159.

Geisser, S. (1970). *The Inferencial Use of Predictive Distributions.* Foundations of Statistical Inference, Eds. V.P. Godambe and D.A. Sprott, pp. 459-469. Toronto: Halt, Rinehart and Winston.

Geisser, S. (1975). *The Predictive Sample Reuse Method with Applications.* Journal of the American Statistical Association, 70: 320-328.

Geisser, S. e Eddy, W. F. (1979). *A Predictive Approach to Model Selection.* Journal of the American Statistical Association, 74(365): 150-153.

Geisser, S. e Eddy, W. F. (1980). *Corrigenda.* Journal of the American Statistical Association, 75: 765.

Gelfand, (1963). *Calculus of Variations.* Englewood Cliffs, Prentice-Hall. 232 p.

Graybill, F. A. (1976). *Theory and Application of the Linear Model.* North Scituate: Duxbury Press. 704 p.

Gutierrez-Peña, E. (1998). *A Bayesian Predictive Approach to Variable Selection and Model Comparison in Regression.* Departamento de Probabilidad y Estadística, IIMAS-UNAM, México.

Hahn, G. J. (1969). *Factors for Calculating Two-sided Prediction Intervals for Samples from a Normal Distribution.* Journal of the American Statistical Association, 64: 878-888.

Harris, I. R. (1989). *Predictive Fit for Natural Exponential Families.* Biometrika, 76: 675-684.

Hinkley, D. V. (1979). *Predictive Likelihood.* The Annals of Statistics, 7: 718-728. Corrigendum 8: 694.

Johnson, W. e Geisser, S. (1982). *Assessing the Predictive Influence of Observations.* Essays in Honor of C.R. Rao, Ed. Kalianpur, Krishnaiah and Ghosh, Amsterdam: North Holland, 343-358.

Johnson, W. e Geisser, S. (1983). *A Predictive View of the Detection and Characterization of Influential Observations in Regression Analysis.* Journal of the American Statistical Association, 74(381): 137-144.

Kalbfleish, S. C. e Sprott, D. A. (1970). *Applications of Likelihood Methods to Models Involving Large Numbers of Parameters.* Journal of the Royal Statistical Society, Series B, 32: 175-208.

Kullback, S. e Leibler, R. A. (1951). *On Information and Sufficiency.* Annals of Mathematical Statistics, 22: 79-86.

Larimore, W. E. (1983). *Predictive Inference, Sufficiency, Entropy and Asymptotic Likelihood Principle.* Biometrika, 70, 1: 175-182.

Laud, P. e Ibrahim, J. (1995). *Predictive Model Selection.* Journal of the Royal Statistical Society, B 57: 247-262.

Lauritzen, S. L. (1974). *Sufficiency, Prediction and Extreme Models.* Scandinavian Journal of Statistics, 1: 128-134.

Leite, J. G. e Singer, J. M. (1990). *Métodos Assintóticos em Estatística: Fundamentos e Aplicações.* São Paulo, IME-USP. 130 p.

Lejeune, M. e Faulkenberry, G. D. (1982). *A Simple Predictive Density Function.* Journal of the American Statistical Association, 77: 654-657.

Levy, M. e Perng, S. K. (1984). *A Maximum Likelihood Prediction Function for de Linear Model with Consistency Results.* Communications in Statistics, A-Theory and Methods, 13: 1257-1273.

Levy, M. e Perng, S. K. (1986). *An Optimal Prediction Function for the Normal Linear Model.* Journal of the American Statistical Association, 81(393): 196-198.

Lindley, D. V. (1965). *Introduction to Probability and Statistics: from a Bayesian Viewpoint.* Cambridge, University Press. vol. 2, 292 p.

Mallows, C. L. (1973). *Some Comments on C_p.* Technometrics, 15: 661-675.

Mathiasen, P. E. (1979). *Prediction Functions.* Scandinavian Journal of Statistics, 6: 1-21.

Nelson, W.B. (1968). *Two-Sample Prediction.* General Electric Company TIS Report 68-C-404. (Available from Distribution Unit, Researsh and Development Center, General Electric Company, Schenectady, N. Y.).

O'Reilly, F. J. (1976). *On a Criterion for Simultaneous Extrapolation in Nonfull Rank Normal Regression.* Annals of Statistics, 4: 625-628.

San Martini, A. e Spezzaferri, F. (1984). *A Predictive Model Selection Criterion.* Journal of the Royal Statistical Society, B 46, 2: 296-303.

Tong, Y. L. (1990). *The Multivariate Normal Distribution.* Springer-Verlag, Berlin, Heldelberg and New York. 271 p.