ELAYNE PENHA VEIGA

A MEDIDA L COMO CRITÉRIO DE COMPARAÇÃO DE MODELOS: UMA REVISÃO DA LITERATURA

LAVRAS – MG

2012

ELAYNE PENHA VEIGA

A MEDIDA L COMO CRITÉRIO DE COMPARAÇÃO DE MODELOS:
UMA REVISÃO DA LITERATURA

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária para a obtenção do título de Mestre.

Orientador

Dr. Mario Javier Ferrua Vivanco

LAVRAS – MG

2012

Ficha Catalográfica Elaborada pela Divisão de Processos Técnicos da Biblioteca da UFLA

Veiga, Elayne Penha.
A medida L como critério de comparação de modelos : uma revisão da literatura / Elayne Penha Veiga. – Lavras : UFLA, 2012.
70 p. : il.

Dissertação (mestrado) – Universidade Federal de Lavras, 2012.
Orientador: Mario Javier Ferrua Vivanco.
Bibliografia.

1. Função perda quadrática. 2. Inferência preditiva bayesiana. 3. Seleção preditiva de modelos. I. Universidade Federal de Lavras. II. Título.

CDD – 519.542

ELAYNE PENHA VEIGA

A MEDIDA L COMO CRITÉRIO DE COMPARAÇÃO DE MODELOS:
UMA REVISÃO DA LITERATURA

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária para a obtenção do título de Mestre.

APROVADA em 28 de fevereiro de 2012.

Dr. Júlio Sílvio de Sousa Bueno Filho	UFLA
Dr. Washington Santos Silva	IFMG
Dr. Telde Natel Custódio	UFSJ

Dr. Mario Javier Ferrua Vivanco
Orientador

LAVRAS – MG

2012

Ao Vicente, meu pai; à Angela, minha mãe e à Renata, minha irmã; que fazem de mim uma pessoa melhor a cada dia.

DEDICO

AGRADECIMENTOS

Antes de qualquer coisa, agradeço a Deus, meu guia, e à Nossa Senhora Aparecida, minha mãe e alento nas horas de angústia.

À Universidade Federal de Lavras (UFLA).

Ao Departamento de Ciências Exatas (DEX), em especial aos professores que me acompanharam e me ajudaram tanto, e às funcionárias pela atenção e carinho.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa de estudos.

Ao professor Mário Javier Ferrua Vivanco, meu orientador, pelos ensinamentos, conselhos, e palavras certas nas horas certas.

Às minhas professoras do primário, "tia" Cássia, "tia" Vívian, "tia" Mônica e "tia" Mônica pelo carinho e atenção na hora de guiar meus primeiros passos na educação escolar e aos meus professores do Ensino Médio, todos sem exceção, pela orientação e ensinamentos.

Ao Departamento de Administração e Economia (DAE) e à UFLA Júnior Consultoria Administrativa.

Ao Centro Universitário de Lavras, em especial aos meus professores do curso de Matemática.

Aos meus amigos das faculdades e meus amigos da Empresa Júnior que se tornaram parte da minha família.

À Comissão de Formatura, que me fez crescer profissionalmente e pessoalmente e que são meus irmãos do coração.

Aos meus familiares, por torcerem por mim sempre.

Ao meu Anjo, pelo amor, carinho e apoio, mesmo longe.

À Angela, minha mãe, por ser a mulher em que me inspiro todos os dias da minha vida, e por ter me ajudado com os “deveres de casa”.

Ao Vicente, meu pai, por me mostrar que amar o que se faz é o segredo das coisas, e por ser o meu Super Herói.

À Renata, minha irmã, por me mostrar como é bom ter paciência e ser mais resiliente.

Às minhas amigas-irmãs, Juliana Goursand, Alessandra Casali, Suzana Duarte e Thaís Barros, pela amizade, carinho e força.

E a todas as pessoas que mesmo indiretamente, me ajudaram nesta caminhada.

“Só há um princípio motor: a faculdade de desejar.”

Autor desconhecido

RESUMO

Este estudo refere-se à Medida L e foi feito através de uma revisão de literatura com o objetivo didático de explicitar seu conceito e justificar o seu uso. A Medida L é um critério que se utiliza de conceitos bayesianos e é construída a partir da distribuição preditiva *a posteriori* dos dados. Pode ser escrita como a soma de dois componentes: um envolve a média desta distribuição e outro envolve as variâncias, e mede o desempenho de um modelo pela combinação de quão próximas as predições estão dos dados observados e qual a variabilidade das predições. Pela Teoria da Decisão, a Medida L é a função perda quadrática. Neste sentido, quando da tomada de decisão, o objetivo é diminuir esta perda ao se escolher um modelo em detrimento de outro. Ainda, o desenvolvimento algébrico da função perda quadrática, resulta no cálculo do Erro Quadrático Médio. Bons modelos terão pequenos valores para a medida L_m^2. Para exemplificar o cálculo da medida, estudos de comparação da Medida L com outros critérios, foram feitos em dois exemplos didáticos de dados de regressão linear múltipla com o intuito de ilustrar e analisar o critério e suas comparações. Os resultados dos dois exemplos diferem; enquanto AIC e BIC selecionaram o mesmo modelo, a Medida L selecionou outro modelo.

Palavras-chave: Medida L. Função Perda Quadrática. Inferência Preditiva Bayesiana. Seleção Preditiva de Modelos.

ABSTRACT

The study refers to L-Measure, and was done through a literature review with the aim of clarifying the concept, justify its use. The L-Measure is a criterion that uses Bayesian concepts and is constructed from the posterior predictive distribution of the data. It can be written as the sum of two components: one involves the mean of this distribution and the other involves the variances. It measures the performance of a model by the combination of how close the predictions are from the observed data and the variability of predictions. By Decision Theory, L-Measure is the quadratic loss function. In this sense, when the decision is taken, the goal is to reduce this loss when choosing one model over another. The algebraic development of the quadratic loss function will result in the Mean Squared Error. Good models will have small values of L_m^2. Comparison studies with other L-Measure criteria were made in two didactics examples with linear regression data with the aim to illustrate and analyze the criterion and their comparisons. The criteria AIC and BIC selected the same model, but L-Measure selected a model different to explain the dependent variable.

Keywords: L-Measure. Quadratic Loss Function. Bayesian Predictive Inference. Predictive Selection of Models.

LISTA DE SÍMBOLOS

a_i	Ação ou decisão.
θ_i	Estado da natureza; parâmetro.
$L(\theta_i, a_i)$	Função Perda. Perda ocorrida se tomada determinada ação a_i quando θ_i é o verdadeiro estado da natureza.
$U(r)$	Função Utilidade. Cada par (θ_i, a_i) determina uma recompensa r, que tem Utilidade $U(r)$.
$\rho[\pi(\theta\|x), a]$	Perda Esperada Bayesiana. Perda esperada dada a distribuição de probabilidade de θ e a ação tomada a.
$\pi(\theta\|x)$	Distribuição de θ observados os dados; Distribuição *a posteriori*.
$L(\theta, a) = (\theta - a)^2$	Função Perda Erro Quadrático.
$R[\theta, \delta(X)]$	Função Risco de uma regra de decisão $\delta(X)$.
$L(\boldsymbol{\theta}, \boldsymbol{a}) = (\boldsymbol{\theta} - \boldsymbol{a})'\boldsymbol{Q}(\boldsymbol{\theta} - \boldsymbol{a})$	Função Perda Quadrática.
$L(\theta; \boldsymbol{x})$	Função de Verossimilhança do parâmetro θ.
$\log L(\theta; \boldsymbol{x}) = l(\theta; \boldsymbol{x})$	Função Suporte.
$E_{\hat{G}}[\log f(\boldsymbol{x}\|\hat{\theta})] = \frac{1}{n}\sum_{i=1}^{n} \log f(x_i\|\hat{\theta})$	Estimador da Função Suporte. O subíndice $\hat{G}$ da Esperança significa que a esperança é calculada com respeito à função empírica $\hat{G}$.
H_0	Hipótese Nula
H_1	Hipótese Alternativa
$\boldsymbol{A_0} = \{(x_1, \dots, x_n) \epsilon \mathcal{X} \| d(x_1, \dots, x_n) = a_0\}$	Região de Aceitação (teste de hipótese), ou seja, pontos amostrais que levam à aceitação de H_0.
$\boldsymbol{A_1} = \{(x_1, \dots, x_n) \epsilon \mathcal{X} \| d(x_1, \dots, x_n) = a_1\}$	Região de Rejeição ou Região Crítica, ou seja, pontos amostrais que levam à rejeição de H_0.
$R(\theta_0, a) = \alpha$	Função Risco que define a probabilidade de ocorrência do erro tipo I, α.

$R(\theta_1, a) = \beta$	Função Risco que define a probabilidade de ocorrência do erro tipo II, β.
$\pi(\theta_1) = 1 - \beta$	Poder do Teste, ou seja, a probabilidade de rejeitar a hipótese nula, H_0, sendo falsa.
$L_0(\boldsymbol{x})$	Função de Verossimilhança de θ_0.
$L_1(\boldsymbol{x})$	Função de Verossimilhança de θ_1.
$I(g;f) = \int_{-\infty}^{+\infty} g(x)\, log\left(\frac{g(x)}{f(x)}\right) dx$	Informação de Kullback-Leibler.
$H_B = -\log\left[\frac{g(x)}{f(x)}\right]$	Variação da Entropia de Boltzmann.
AIC	Critério de Informação de Akaike.
$\log L(\hat{\theta})$	Função Suporte Maximizada.
BIC	Critério de Informação de Schawrz.
$p(\theta\|x)$	Distribuição de θ depois de observados os dados.
$p(x\|\theta)$	Distribuição dos dados, x.
$p(x,\theta)$	Distribuição conjunta de θ e x.
$p(\theta)$	Distribuição de θ.
$\pi(\theta)$	Distribuição *a priori* de θ.
$f(x\|\theta)$	Função de densidade de probabilidade de X.
$p(z\|x)$	Densidade preditiva da variável aleatória Z.
$g(z\|\theta)$	Densidade da variável aleatória Z.
$p_m(y\|\theta)$	Densidade dos dados y, dado θ e o modelo m.
$\pi_m(\theta)$	Distribuição *a priori* de θ dado o modelo m.
$\pi_m(\theta\|y)$	Distribuição *a posteriori* de θ observados os dados e o modelo m.
$p_m(z\|y)$	Densidade preditiva de z dado y e o modelo m ; ou, Densidade Preditiva do Experimento Replicado.
$p_m(z\|\beta^{(m)},\tau)$	Densidade de z dados os parâmetros $\beta^{(m)}$ e τ, e o modelo m.
$\pi_m(\beta^{(m)},\tau\|y)$	Distribuição *a posteriori* dos parâmetros $\beta^{(m)}$ e τ observados os dados y.

L_m^2	Medida L
$L(y, b, k)$	Forma Geral da Medida L

SUMÁRIO

1 INTRODUÇÃO

Os processos de tomada de decisão envolvem avaliações e decisões que são escolhas feitas com base em propósitos, são ações orientadas para determinado objetivo e o alcance deste objetivo determina a eficiência do processo. A decisão pode ser tomada a partir de probabilidades, possibilidades e/ou, alternativas.

Geralmente, tomadores de decisão baseiam-se em argumentos matemáticos e/ou estatísticos para conferir credibilidade às escolhas.

A teoria da decisão estatística é um conjunto de métodos para a tomada de decisão que permitem resultados confiáveis. Preocupa-se com decisões que envolvem incerteza.

Os modelos matemáticos ou determinísticos tentam explicar fenômenos quando todas as variáveis envolvidas são conhecidas, podendo então ser representadas. Os modelos estatísticos incluem variáveis envolvidas no processo que não são conhecidas, e, portanto não podem ser representadas matematicamente e que então compõem o erro da modelagem.

Na natureza, fenômenos e experimentos estudados pelos cientistas e pesquisadores não podem ser completamente conhecidos, já que são muito mais complexos e, geralmente no processo de coleta de dados e análise sempre há erros associados.

A modelagem estatística é uma das principais ferramentas do estudo estatístico de experimentos que auxilia no melhoramento de processos e produtos. De forma que a observação de um evento, qual seja, de forma controlada ou não, possa gerar padrões que modelados auxiliam em estudos futuros. Assim, é importante que o modelo seja válido, no sentido de ser mais próximo à realidade observada, já que este, sendo ótimo será usado para fazer previsões/inferências.

Em resumo, o que se faz em estatística é ajustar modelos a conjuntos de dados a partir de experimentos ou fenômenos aleatórios. A questão é: qual será o modelo mais apropriado para representar esse fenômeno aleatório? ou, entre diversos modelos, qual será o mais adequado?

Gelfand e Gosh (1998) comentam da importância da escolha entre modelos candidatos como atividade fundamental na análise de conjunto de dados. Ainda, citam a estatística razão de verossimilhanças como critério primário para a seleção de modelos, e alguns autores como Akaike (1974), Schwarz (1978), entre outros, que propuseram penalizações para esta estatística.

Os critérios de informação de Akaike (1974), AIC, e o critério de informação Bayesiano de Schwarz (1978), BIC, são medidas para escolha de modelos bastante utilizadas na literatura. Emiliano et al. (2009) observam que apesar do amplo uso do Critério de Informação de Akaike, a validação do critério precisa de grandes amostras o que, às vezes, leva a abusos na sua utilização. Para selecionar modelos através do BIC, assim como para o AIC, deve-se calculá-lo e escolher aquele que tem o menor valor da medida. Ambos os critérios são assintóticos.

Ibrahim e Laud (1994) propuseram um critério de comparação de modelos, chamado Medida L que é bem menos utilizada que o AIC e o BIC, entre outras razões, pelo fato de não estar implementada em pacotes estatísticos.

A Medida L baseia-se no preditivismo, cuja preocupação é a predição de valores advindos de um mesmo experimento ou de experimentos semelhantes. O cálculo desta medida usa do conceito de Densidade Preditiva em sua definição, portanto, compara o que é predito ao que é observado para fazer a escolha entre modelos.

Neste trabalho, objetiva-se:

1º Explicar conceitualmente o que é a Medida L,

2º Explicar os conceitos pertinentes à Medida L de maneira compreendê-la melhor como critério para seleção de modelos candidatos.

3º Apresentar dois exemplos didáticos para observar o comportamento da Medida L e de dois critérios bastante utilizados, a saber: AIC e BIC .

Neste trabalho, a medida é introduzida para modelos de regressão linear, ou seja, quando, a matriz de incidência é de posto completo e com erros homocedásticos. E será aplicada a tais modelos para satisfazer o terceiro objetivo.

2 REFERENCIAL TEÓRICO

Nesta seção serão apresentados conceitos importantes para tornar claro o conceito da Medida L e atingir o objetivo proposto neste trabalho.

A Medida L pode ser interpretada de formas diferentes, dependendo da abordagem, a saber: função Perda Quadrática; distância Euclidiana e Erro Quadrático Médio. Para cada uma destas interpretações, conceitos devem ser elucidados.

Como simples ilustração, nos exemplos didáticos, serão calculados os valores dos critérios de Informação de Akaike e Bayesiano. Estes conceitos serão apresentados abordando a Modelagem e conceitos pertinentes à fundamentação destes critérios de uma forma sintética antes dos conceitos pertinentes à Medida L.

2.1 Modelagem e critérios para comparação de modelos

A partir da observação de dados advindos de fenômenos ou de resultados de experimentos planejados, modelos são formulados na tentativa de resumir a informação disponível e fazer inferências. A modelagem é o desenvolvimento de expressões matemáticas que, de alguma forma, tentam descrever o comportamento de determinada variável de interesse.

Existem modelos determinísticos e modelos estatísticos. Os determinísticos são aqueles em que tudo que é observado é conhecido e é então possível de ser traduzido através de uma função ou simbologia específica no modelo. Já os estatísticos são aqueles em que existe uma parte sistemática e outra aleatória, ou seja, uma parte que é explicada pelo modelo de forma

determinística e outra que não é possível de ser traduzida sistematicamente e é atribuída ao acaso na parte aleatória.

Geralmente, na natureza, a complexidade dos fenômenos não permite que o comportamento de variáveis seja traduzido completamente em modelos determinísticos. Então, faz-se a aproximação do fenômeno por um modelo estatístico. Neste, haverá perda de informação, devida à parte aleatória, que deve ser minimizada.

Ainda, um ou mais modelos podem ser formulados a partir de um mesmo conjunto de dados. Assim, é interessante que haja alguma forma de selecionar aquele que melhor explica o comportamento dos dados, levando em consideração a qualidade do ajuste e complexidade, geralmente devida ao número de parâmetro: quanto mais parâmetros, mais complexo o modelo.

Para selecionar modelos, podemos fazer a seleção dos parâmetros fazendo testes de hipóteses ou ainda, usando de muitos critérios existentes na literatura estatística, entre eles, pode-se citar o Critério de Informação de Akaike, o Critério de Informação de Schwarz ou Bayesiano que são calculados com relação à verossimilhança dos dados e o conceito de Informação e Entropia.

Nas seções seguintes serão apresentados os conceitos de Teste de Hipóteses e os erros que se pode cometer ao aceitar ou rejeitar uma hipótese qualquer, conceitos de verossimilhança e Função Suporte, e dos Critérios de Informação de Akaike e Bayesiano.

2.1.1 Teste de hipóteses e erro tipo I e tipo II

Podemos selecionar modelos fazendo testes de hipóteses, selecionando os parâmetros que farão parte do modelo em questão.

O Teste de Hipóteses será apresentado, nesta revisão de literatura, com o objetivo de introduzir e esclarecer os tipos de erros incorridos ao aceitar ou rejeitar uma hipótese num teste de hipóteses.

Quando se dispõe de evidências, pode-se usá-las para tomar a decisão de aceitar ou rejeitar determinada afirmação. Essas afirmações ou hipóteses, como são chamadas na Teoria Estatística, devem ser testadas a fim de aceitá-las ou rejeitá-las.

De acordo com Mood, Graybill e Boes (1974), tem-se a definição de *hipótese estatística:*

Definição 1: *Uma hipótese estatística é uma asserção ou conjectura sobre a distribuição de uma ou mais variáveis aleatórias.*

Comumente, a hipótese de interesse é chamada *hipótese de nulidade*, representada por H_0, que é testada contra a *hipótese alternativa*, representada por H_1 ou H_a, que podem ser simples ou compostas. Definidos os espaços paramétricos Θ_0 e Θ_1, quando $\Theta_0 = \{\theta_0\}$ H_0 é simples, caso contrário, é composta – e da mesma forma para H_1. As notações comumente utilizadas são:

$$\begin{cases} H_0: \theta \in \Theta_0 \\ H_1: \theta \in \Theta_1 \end{cases}$$

Definição 2: *Chama-se "teste de uma hipótese estatística", a função de decisão* $d: \mathcal{X} \to \{a_0, a_1\}$*, em que* a_0 *corresponde à ação de considerar a hipótese* H_0 *como verdadeira,* a_1 *corresponde à ação de considerar a hipótese* H_1 *como verdadeira e* $\mathcal{X}$ *é o espaço amostral associado à amostra aleatória* $X_1, \ldots X_n$ *que é dividida nos dois conjuntos:*

$$\boldsymbol{A_0} = \{(x_1, \ldots, x_n) \in \mathcal{X} | d(x_1, \ldots, x_n) = a_0\}$$

$$\boldsymbol{A_1} = \{(x_1, \dots, x_n)\, \epsilon\, \mathcal{X} | d(x_1, \dots, x_n) = a_1\}$$

Sendo, $\boldsymbol{A_0} \cup \boldsymbol{A_1} = \mathcal{X}$ e $\boldsymbol{A_0} \cap \boldsymbol{A_1} = \emptyset$.

Como em $\boldsymbol{A_0}$ temos os pontos amostrais $\boldsymbol{x} = (x_1, \dots, x_n)$ que levam à aceitação de H_0, chama-se $\boldsymbol{A_0}$ de *região de aceitação*, e por analogia, $\boldsymbol{A_1}$ de *região de rejeição* de H_0, também chamada de *região crítica* (BOLFARINE; SANDOVAL, 2000)

No caso de testar duas hipóteses simples, como $H_0: \theta = \theta_0$ versus $H_1: \theta = \theta_1$, e considerando a função perda $l(\theta, a) = 0$ se a decisão for correta e $l(\theta, a) = 1$ se a decisão for incorreta, a função risco é calculada:

$$R(\theta_0, a) = E[l(\theta_0, a)] = 0P[\boldsymbol{X} \in \boldsymbol{A_0} | \theta_0] + 1P[\boldsymbol{X} \in \boldsymbol{A_1} | \theta_0]$$
$$\Rightarrow R(\theta_0, a) = P[\boldsymbol{X} \in \boldsymbol{A_1} | \theta_0] = \alpha$$

Ou seja, essa função risco define a probabilidade de ocorrência do erro tipo I, α, que é aquele que se comete ao rejeitar a hipótese H_0 sendo verdadeira. E,

$$R(\theta_1, a) = E[l(\theta_1, a)] = 0P[\boldsymbol{X} \in \boldsymbol{A_1} | \theta_1] + 1P[\boldsymbol{X} \in \boldsymbol{A_0} | \theta_1]$$
$$\Rightarrow R(\theta_1, a) = P[\boldsymbol{X} \in \boldsymbol{A_0} | \theta_1] = \beta$$

Ou seja, essa função risco define a probabilidade de ocorrência do erro tipo II, β, que é aquele que se comete ao aceitar a hipótese H_0 sendo falsa.

O que geralmente se faz é fixar a probabilidade do erro tipo I, α, e procurar a região crítica que forneça a menor probabilidade do erro tipo II, β, ou seja, o maior poder entre os testes com igual ou menor nível α.

O poder do teste $\pi(\theta_1)$ é definido pela probabilidade de rejeitar a hipótese nula, sendo falsa:

$$\pi(\theta_1) = P[\boldsymbol{X} \in \boldsymbol{A_1} | \theta_1]$$
$$\Rightarrow \pi(\theta_1) = 1 - \beta$$

2.1.2 Verossimilhança e função suporte

O conceito de Função de Verossimilhança é fundamental em toda a teoria estatística, e está envolvido em muitos processos de inferência e, inclusive, para a definição dos critérios de informação de Akaike e de Schwarz. Da mesma forma, a Função Suporte.

Bolfarine e Sandoval (2000) definem a função de verossimilhança como:

Definição 3: *Sejam $X_1, \ldots, X_n$ uma amostra aleatória de tamanho n da variável aleatória X com função de densidade (ou de probabilidade) $f(x|\theta)$, com $\theta \epsilon\ \Theta$, em que Θ é o espaço paramétrico. A função de verossimilhança de θ correspondente à amostra aleatória observada, calculada como o produtório das funções de densidade, é dada por*

$$L(\theta; \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\theta)$$

Definida a função acima, pode-se definir o Método de Máxima Verossimilhança, que é um método para a obtenção de um estimador para o parâmetro, θ, através do conceito acima.

Definição 4: *O estimador de máxima verossimilhança de* θ *é o valor de* $\theta \in \Theta$ *que maximiza a função de verossimilhança.*

Ou seja, o estimador dado pelo Método de Máxima Verossimilhança é o valor que maximiza a função de verossimilhança. Ou ainda, é o valor de θ que maximiza a probabilidade de se obter a amostra observada.

Diante disso, para maximizar a função, deve-se calcular a primeira derivada de $L(\theta; \boldsymbol{x})$ com respeito ao parâmetro θ, igualar a zero e resolver para θ. Obtêm-se, portanto, os pontos críticos. Se existir, aquele ponto que maximiza a função é o estimador de máxima verossimilhança de θ. Ou seja,

$$\frac{dL(\theta; \boldsymbol{x})}{d\theta} = 0$$

Para o caso em que se tem mais de um parâmetro, as derivadas tomadas são as parciais com relação a cada um dos parâmetros envolvidos. E procede-se de maneira análoga.

"Não é difícil verificar que o valor de θ que maximiza a função de verossimilhança $L(\theta; \boldsymbol{x})$, também maximiza $l(\theta; \boldsymbol{x})$, dada por $l(\theta; \boldsymbol{x}) = \log L(\theta; \boldsymbol{x})$." (BOLFARINE; SANDOVAL, 2000, p. 35). Esta função é chamada Função Suporte, muitas vezes é mais tratável numericamente, sendo também mais fácil de encontrar os pontos críticos.

A Função Suporte pode ser estimada. Substituindo em $E_g[\log f(\boldsymbol{x}|\theta)] = \int \log f(\boldsymbol{x}|\theta) g(\boldsymbol{x}) d\boldsymbol{x}$, o parâmetro ou vetor de parâmetros que foi estimado pelo método da máxima verossimilhança tem-se:

$$E_g[\log f(\boldsymbol{x}|\hat{\theta})] = \int \log f(\boldsymbol{x}|\hat{\theta}) g(\boldsymbol{x}) d\boldsymbol{x} = \int \log f(\boldsymbol{x}|\hat{\theta}) dG(\boldsymbol{x})$$

em que $g(\boldsymbol{x})$ é a verdadeira distribuição dos dados, $f(\boldsymbol{x}|\ddot{\theta})$ é o modelo que aproxima $g(\boldsymbol{x})$ após estimados os parâmetros e $G(\boldsymbol{x})$ é função de distribuição acumulada de $g(\boldsymbol{x})$.

O que se deseja é encontrar um bom estimador para a função suporte, já que, depois de estimarmos o parâmetro, passa-se a trabalhar com $f(\boldsymbol{x}|\ddot{\theta})$. Um estimador pode ser obtido substituindo $G(\boldsymbol{x})$ por uma distribuição empírica $\hat{G}(\boldsymbol{x})$. Emiliano ET al (2009) apresentaram o seguinte resultado como estimador da função suporte:

$$E_{\hat{G}}\left[\log f\left(\boldsymbol{x}|\hat{\theta}\right)\right] = \frac{1}{n}\sum_{i=1}^{n} \log f\left(x_i|\hat{\theta}\right)$$

Portanto, o estimador da função suporte esperada é $n^{-1}L(\ddot{\theta})$, e a função suporte, $L(\ddot{\theta})$, é um estimador para $nE_G\left[\log f\left(\boldsymbol{x}|\hat{\theta}\right)\right]$.

2.1.3 Critério de informação de Akaike (AIC)

Este critério é fundamentado nos conceitos de informação e entropia, e tem o objetivo de comparar modelos candidatos através de uma penalização da função suporte maximizada.

A partir dos conceitos de informação e entropia, foi estabelecido o conceito de Informação de Kullback-Leibler (K-L). Tal conceito de Informação encontra-se detalhado em Emiliano et al. (2009).

Akaike (1974) propôs utilizar a informação de K-L para seleção de modelos, estabelecendo uma relação entre esta e a Razão de Verossimilhança.

Essa relação é chamada Critério de Informação de Akaike (AIC). Para melhor entendimento da definição deste critério, a informação de K-L será apresentada.

Definição 5: *A informação de Kullback-Leibler é definida por:*

$$I(g;f) = E_g[-H_B] = E_g\left[log\left(\frac{g(x)}{f(x)}\right)\right] = \int_{-\infty}^{+\infty} g(x)\, log\left(\frac{g(x)}{f(x)}\right)dx \qquad (1.1)$$

em que H_B *é a entropia de Boltzmann;* g *é a distribuição da qual são gerados os dados;* f *é a distribuição utilizada para aproximar* g*; e,* E_g *é a esperança calculada com respeito a distribuição* g*.*

A partir de (1.1), pode-se observar que:

$$I(g;f) = E_g[\log(x)] - E_g[logf(x)]$$

$$I(g;f) = \int_{-\infty}^{+\infty} g(x)\log[g(x)]\,dx - \int_{-\infty}^{+\infty} g(x)\log[f(x)]\,dx$$

(1.2)

Isto é, de (1.2) pode-se deduzir que a informação de K-L quantifica a perda de informação quando avaliamos um modelo arbitrário especificado, por exemplo, $f(x)$, em comparação ao modelo verdadeiro, $g(x)$, de um conjunto $X = \{x_1, x_2, \dots, x_n\}$ de n observações independentes.

Emiliano et al. (2009) citam 3 propriedades da informação de K-L apresentadas por Konishi e Kitagawa (2008):

(P1) Para quaisquer funções de densidade de probabilidade f e g, $I(g;f) \geq 0$;

(P2) Se f e g são funções de densidade de probabilidade e $I(g;f) = 0$, então $f(x) = g(x)$, $\forall x \in \mathbb{R}$; e,

(P3) Se f e g são duas funções de densidade de probabilidade e $f \xrightarrow{d} g$, então $I(g; f) \xrightarrow{d} 0$.

Porém, observa-se que a informação de Kullback-Leibler pode ser de complicada aplicação para comparar modelos, já que na grande maioria das vezes não se conhece o modelo verdadeiro, $g(x)$.

Com essa motivação, Akaike (1974) propôs um critério para comparação de modelos em que não é necessário o conhecimento de $g(x)$.

Definição 6: *O Critério de Informação de Akaike, AIC, é definido por:*

$$AIC = -2(Função\ Suporte\ Maximizada) + 2(número\ de\ parâmetros)$$

ou seja,

$$AIC = -2 \log L(\hat{\boldsymbol{\theta}}) + 2(k)$$

em que k é o número de parâmetros no modelo. Esta expressão é obtida baseando-se na ideia que o viés tende ao número de parâmetros a serem estimados pelo modelo.

Em resumo, o AIC é uma ferramenta para comparação de modelos. Dado um conjunto de dados e os modelos candidatos gerados a partir destes dados, seleciona-se o melhor, - ou seja, aquele modelo entre os candidatos que melhor explica aquele conjunto de dados – selecionando aquele que apresentar o menor valor do AIC.

2.1.4 Critério de informação de Schwarz (BIC)

Este critério, proposto por Schwarz (1978), é um critério para a comparação de modelos candidatos e define-se a partir da distribuição *a posteriori* que será definida na seção 2.4.3.

Emiliano et al. (2009) apresentam a definição do BIC:

Definição 7: *Seja* $F(x_n|\ddot{\theta})$ *um modelo estatístico estimado através do método da máxima verossimilhança. Então o Critério de Informação Bayesiano, BIC, é dado por:*

$$BIC = -2\log f\left(x_n|\widehat{\boldsymbol{\theta}}\right) + p\,\log(n)$$

em que $f(x_n|\ddot{\boldsymbol{\theta}})$ é o modelo selecionado para o cálculo, p é o número de parâmetros a serem estimados e n é o número de observações da amostra.

Na comparação de modelos candidatos, calculam-se os valores BIC para cada um deles e seleciona-se aquele que apresentar menor valor.

2.2 A Medida L como função perda quadrática

Para introduzir o conceito de Função Perda Quadrática, é importante fazer uma breve discussão sobre o processo de tomada de decisão e Teoria da Decisão.

2.3 Tomada de decisão

A tomada de decisão, em estatística, envolve essencialmente tomar decisões em um ambiente de incerteza, em que estas incertezas podem ser trabalhadas probabilisticamente.

Alguns problemas de decisão amplamente discutidos são, por exemplo, o lançamento ou não de um novo produto farmacêutico, ou de um produto e tipo do tratamento de uma doença.

Decidir qual, entre muitos modelos candidatos, não é uma tarefa fácil, já que estes são uma representação da realidade e estão sujeitos a erros de ajuste, devidos à complexidade da natureza. Quando se testa a hipótese se um parâmetro qualquer está ou não incluído no modelo, se está sujeito a cometer dois tipos de erros, cujas probabilidades serão definidas na seção 2.3.1, quais sejam:

a) Rejeitar uma hipótese que na verdade deveria ser aceita, conhecido como Erro Tipo I; e,
b) Aceitar uma hipótese, que, na verdade, deveria ser rejeitada, conhecido como Erro Tipo II.

A comparação de modelos tem sido objeto de estudo na literatura estatística.

De acordo com Gelfand e Gosh (1998), na literatura clássica, ao tomar decisões sobre escolha de modelos, o critério estatístico primário é a razão de verossimilhanças. Como resultado, autores como Akaike (1974), tem proposto penalizar a verossimilhança usando funções que influenciam no número de parâmetros quando comparados modelos diferentes, em relação a tal número.

A área da estatística que se preocupa com a estrutura do processo de tomada de decisão é a Teoria da Decisão.

2.4 Teoria da decisão

Os dados observacionais ou advindos de um experimento planejado, quando organizados e passíveis de análise, transformam-se em informações com significado; os dados passam a ter relevância e propósito. Esses dados podem ser trabalhados estatisticamente na forma de modelos. Os modelos são formulações matemáticas que aproximam e que sintetizam informações importantes sobre o

comportamento dos dados observados. E é nessas informações que a Teoria da Decisão é fundamentada.

Berger (1985) comenta que a Teoria da Decisão é um conjunto de métodos para a tomada de decisão que permitem resultados confiáveis. E, ainda, preocupa-se com decisões que envolvem incerteza.

No processo de decisão Bayesiano, a quantidade desconhecida – o parâmetro ou vetor de parâmetros – afeta a decisão e é conhecida como o estado da natureza. O estado da natureza e a ação vão definir a *função perda*, importante elemento para esta teoria.

O estado da natureza, comumente representado por uma quantidade θ, por exemplo, afeta o processo de decisão. Pode-se representar por Θ, todos os possíveis estados da natureza. De acordo com Berger (1985), quando experimentos são planejados para obter informação sobre θ - estado da natureza -, geralmente as observações são distribuídas de acordo com alguma distribuição de probabilidade com θ como parâmetro desconhecido. Assim, Θ é chamado de espaço paramétrico.

Decisões são ações, e podem ser representadas por a, e o conjunto de todas as ações que podem ser tomadas podem ser chamadas de A.

Se uma particular ação a_1 é tomada e θ_1 é o estado da natureza, então a perda $L(\theta_1, a_1)$ ocorrerá (BERGER, 1985, p. 3).

Em resumo, num problema de decisão fica especificado:

i. O Estado da Natureza, θ, e o espaço dos estados da natureza ou espaço paramétrico, Θ;
ii. A ação tomada, a, e o espaço de todas as ações que podem ser tomada, $\mathcal{A}$; e,
iii. A perda incorrida definida pela Função Perda, $L(\theta, a)$.

2.4.1 Função utilidade e função perda

Como mencionado na seção anterior, na teoria da decisão, alguns conceitos são de extrema importância para a Teoria da Decisão, como o parâmetro desconhecido θ, as ações a, e a função perda.

As *ações* são tomadas pelo pesquisador com base nas informações sobre o verdadeiro estado da natureza. Essas ações incorrerão consequências, e estas podem ser avaliadas numericamente. Na teoria da decisão, os números que quantificam as consequências são chamados de *utilidades*.

Pode-se denominar o conjunto de todas as consequências de determinada ação tomada por $\mathcal{R}$. E essas consequências têm incertezas envolvidas na sua ocorrência, portanto, os resultados de ações são frequentemente distribuições de probabilidades em $\mathcal{R}$.

> Seja $\mathcal{P}$ o conjunto de todas as distribuições de probabilidade. É geralmente necessário trabalhar com valores e preferências sobre distribuições de probabilidade neste conjunto. Seria fácil fazer se o valor real da função $U(r)$ pudesse ser construída tal que o "valor" da distribuição de probabilidade $P \in \mathcal{P}$ fosse dada pela utilidade esperada $E^P[U(r)]$. Se tal função existe, é chamada de *função utilidade* (BERGER, 1985, p.47).

Sendo $r \in \mathcal{R}$.

Ainda de acordo com o autor, o objetivo é encontrar uma função utilidade, $U(r)$ que representa o verdadeiro padrão de preferência do tomador de decisão em $\mathcal{P}$. A função é tomada tal que, se P_1 e P_2 estão em $\mathcal{P}$, então P_2 é preferida ao invés de P_1 se e somente se

$$E^{P_2}[U(r)] > E^{P_1}[U(r)]$$

Concluindo, um problema que envolve decisão pode ser resolvido utilizando a função utilidade. De acordo com Berger (1985), cada par (θ, a) determina uma "recompensa" r - consequência de determinada ação tomada -, que tem utilidade $U(r)$. Esta função pode ainda ser representada por $U(\theta, a)$.

Definindo θ como o parâmetro desconhecido – estado da natureza – e a como a ação a ser tomada, e tendo em mente que $\theta \in \Theta$, espaço paramétrico, e $a \in \mathcal{A}$, então a função perda é definida por $L(\theta, a)$.

Uma vez que $U(\theta, a)$ foi obtida, a função perda pode ser simplesmente definida como

$$L(\theta, a) = -U(\theta, a)$$

Em decisões na presença de incerteza, a função perda não será conhecida com certeza. Diante disso, uma alternativa é considerar a perda esperada ao tomar uma decisão e então, como mencionado anteriormente, escolher a melhor opção.

A definição de Perda Esperada Bayesiana é dada por:

Definição 8: *se $\pi(\theta|x)$ é a distribuição de probabilidade de θ no momento da decisão, a Perda Esperada Bayesiana de uma ação a é*

$$\rho[\pi(\theta|x), a] = E^{\pi(\theta|x)} L(\theta, a)$$
$$= \int L(\theta, a) dF^{\pi(\theta|x)}(\theta)$$
$$= \int L(\theta, a)\pi(\theta|x) d\theta$$

em que a integral é calculada no espaço paramétrico Θ. E, $\pi(\theta|x)$ representa a distribuição a posteriori do parâmetro.

A Perda Esperada Bayesiana é também definida como Função Risco.

Na tomada de decisão ou na avaliação de uma regra de decisão, importantes funções perda padrões são utilizadas, como por exemplo, a função perda erro quadrático, definida por

$$L(\theta, a) = (\theta - a)^2$$

(1.3)

De acordo com Berger (1985) existem inúmeras razões para o uso de (1.3) para avaliar regras de decisão. As mais importantes, citadas pelo autor, são três:

1) A função perda erro quadrático será utilizada em problemas de inferência quando um estimador não viciado para o parâmetro fosse utilizado, já que a função risco seria a variância desse estimador. Esta variância será a esperança com relação ao parâmetro da função perda considerada, ou seja,

$$R[\theta, \delta(X)] = E_\theta[L(\theta, \delta(X)] = E_\theta[\theta - \delta(X)]^2$$

em que $R[\theta, \delta(X)]$ é a função risco de uma regra de decisão $\delta(X)$ e a esperança é tomada com relação ao parâmetro θ.

Definição 9: Uma regra de decisão δ é uma função definida em Ω que assume valores em $\mathcal{A}$, ou seja: $\delta: \Omega \rightarrow \mathcal{A}$.

2) Existe uma relação entre a função perda erro-quadrático e a teoria clássica de mínimos quadrados, já que se a função perda quadrática é empregada, a perda esperada reduz-se ao Erro Quadrático Médio.

3) Para a maioria dos problemas de análise de decisão, o uso da função perda erro quadrático torna os cálculos relativamente fáceis e simples.

Uma generalização da função perda erro-quadrático é a *função perda quadrática*, que é uma extensão natural para situações multivariadas, e é dada por:

$$L(\boldsymbol{\theta}, \boldsymbol{a}) = (\boldsymbol{\theta} - \boldsymbol{a})' \boldsymbol{Q} (\boldsymbol{\theta} - \boldsymbol{a})$$

em que $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ é o vetor de parâmetros a ser estimado por $\boldsymbol{a} = (a_1, \ldots, a_p)'$ e $\boldsymbol{Q}$ é uma matriz positiva definida $p \times p$. Se $\boldsymbol{Q}$ é diagonal, então:

$$L(\boldsymbol{\theta}, \boldsymbol{a}) = \sum_{i=1}^{p} q_i (\theta_i - a_i)^2$$

Posteriormente, ver-se-á que a Medida L, objetivo deste estudo, é uma função Perda Quadrática.

2.5 Alguns conceitos sobre inferência Bayesiana

Por Inferência estatística, entende-se fazer inferência sobre o estado da natureza em termos de probabilidade.

A estatística Bayesiana fundamenta-se na ideia da probabilidade condicional, traduzida pelo Teorema de Bayes, e também na ideia de probabilidade como grau de credibilidade.

O conhecimento sobre algum evento de interesse pode ser traduzido através de sua probabilidade de ocorrência. A crença em determinado evento de

interesse depende da familiaridade do pesquisador/analista com o mesmo, no sentido de saber estabelecer a probabilidade com maior ou menor propriedade; tendo ou não observado eventos semelhantes anteriormente.

Diante disso, os conceitos de Distribuição *a priori* e *a posteriori* são formulados. O primeiro, traduzindo o conhecimento prévio do pesquisador, e o segundo, a atualização do primeiro via Teorema de Bayes.

2.5.1 Teorema de Bayes

O Teorema de Bayes é simplesmente uma afirmação sobre probabilidades condicionais.

Para sua definição, suponha um conjunto de eventos mutuamente exclusivos, $A_1, \dots, A_k$, em que os eventos B e A_j são de interesse especial.

De acordo com Press (2003), o Teorema fornece um modo de encontrar a probabilidade condicional de um evento A_j dada a ocorrência de outro evento B nos termos da probabilidade condicional de B dado A_j. Dessa forma:

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)}; P(B) > 0$$

em que $P(A_j|B)$ é a probabilidade de ocorrer A_j dado que já ocorreu B; e $P(A_j)$ é a probabilidade de ocorrer A_j.

O Teorema pode ser entendido em função de funções de probabilidade e funções de densidade de probabilidade.

Quando exposto por Bayes (1763), os dados tinham distribuição Binomial e a distribuição *a priori* para o parâmetro era Uniforme. Porém, o teorema não é tão limitado e tem sido generalizado, incluindo uma grande variedade de distribuições para os dados e de distribuições *a priori*.

Em termos de variáveis aleatórias e parâmetros, suponha θ o parâmetro desconhecido e X uma variável aleatória. Vamos atualizar o conhecimento sobre o parâmetro, observados os dados. Sendo $p(\theta)$ o conhecimento prévio sobre θ, e $p(X|\theta)$ a função de distribuição dos dados. Após observar $X = \boldsymbol{x}$ tem-se:

$$p(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x},\theta)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{\int p(\boldsymbol{x}|\theta)p(\theta)d\theta}$$

em que $p(\theta|x)$ é a distribuição do parâmetro depois de observados os dados. Ou seja, é a atualização do conhecimento que se tem sobre θ depois de observada a realização da amostra aleatória.

Pode-se observar que $\frac{1}{p(x)}$ não depende de θ e funciona como uma constante normalizadora de $p(\theta|x)$, que pode ser facilmente calculada, já que $p(\theta|x) = kp(x|\theta)p(\theta)$, fazendo:

$$k^{-1} = \int p(x|\theta)p(\theta)d\theta$$

que também é chamada de densidade marginal ou preditiva de x.

Diante disso, pode-se reescrever o teorema da seguinte forma:

$$p(\theta|x) \propto l(\theta; x)p(\theta)$$

em que $l(\theta; x) = p(x|\theta)$ é a verossimilhança do parâmetro, e o símbolo $\propto$ indica proporcionalidade.

O conhecimento prévio sobre o(s) parâmetro(s) pode ser representado através da Função de Distribuição *a priori*.

2.5.2 Distribuição *a priori*

Probabilidades *a priori* são graus de crença que o pesquisador/analista tem antes de observar qualquer dado que pode resultar de um problema. Em casos em que não há dados disponíveis, a probabilidade *a priori* é de extrema importância. Em casos em que existem dados disponíveis, têm-se duas situações observáveis: se a amostra for grande, os dados "falarão por si mesmos"; mas caso a amostra seja pequena, probabilidades *a priori* podem pesar em contraste com a pequena quantidade de dados observados.

De acordo com Box e Tiao (1992), a distribuição *a priori* tem um papel importante na análise Bayesiana. Ela representa o conhecimento dos parâmetros desconhecidos antes dos dados estarem disponíveis. Ainda, pode ser usada para representar um conhecimento primeiro ou "ignorância relativa".

Distribuições *a priori* podem ser objetivas e subjetivas, e nessas categorias, ainda se subdividir em informativas, não informativas, próprias e impróprias. Existem muitas questões a se considerar na escolha de *prioris*. Algumas opções incluem *prioris* conjugadas, que têm a vantagem da conveniência matemática; *prioris* não-informativas, quando a crença prévia sobre determinado evento não é tão forte e/ou não desejamos influenciar a análise; e *prioris* informativas, quando a crença é forte o suficiente e queremos que esta influencie na análise.

Distribuições *a priori* objetivas são aquelas em que se tenta traduzir muito pouca informação disponível sobre o parâmetro antes que qualquer observação seja feita, e que ainda, seja a crença comum da maioria das pessoas. Neste sentido, temos as distribuições *a priori* para Políticas Públicas, usadas pelos tomadores de decisão, no sentido de, por exemplo, refletir a opinião de um grupo grande de pessoas.

Ainda, neste contexto, de prioris objetivas, tem-se o Princípio da Razão Insuficiente de Laplace que sugere que na ausência de qualquer razão ao contrário, todos os valores do parâmetro desconhecido deveriam ser igualmente prováveis *a priori*. E, Jeffreys (1961 citado por PRESS, 2003), que seguiu essencialmente o mesmo princípio, concluiu que, quando o parâmetro desconhecido encontra-se em um intervalo finito, a distribuição Uniforme atende a necessidade de traduzir pouca informação, como política pública, ou distribuição *a priori* objetiva. Quando, pelo menos, um ponto do domínio do parâmetro não é finito, a distribuição *a priori* objetiva se torna imprópria (ou seja, não integra 1).

Distribuições *a priori* subjetivas, como o próprio nome sugere, traduzem um conhecimento prévio subjetivo, incorporado através de observações de eventos semelhantes. A questão subjetiva traduz-se no sentido que diferentes pesquisadores têm diferentes ideias sobre distribuições *a prioris* do mesmo parâmetro numa mesma situação.

É importante comentar que geralmente não é fácil encontrar distribuições *a priori* subjetivas já que não é sempre fácil traduzir um conhecimento prévio subjetivo em uma distribuição de probabilidade com significado.

Supondo que a *priori* não seja vaga, no sentido de não refletir indiferença a todos os valores do parâmetro, mas tendo alguma informação a ser traduzida por uma distribuição de probabilidade, muitas vezes será suficiente que o grau de crença *a priori* seja representado por uma distribuição que é membro específico de uma família de funções de distribuições *a priori*, comumente chamada de *família conjugada natural*.

Distribuições *a priori* conjugadas refletem a ideia das distribuições *a priori* e *a posteriori* pertencerem à mesma classe de distribuições, assim,

atualiza-se o conhecimento sobre o parâmetro θ somente com a mudança nos hiperparâmetros.

Chen, Ibrahim e Yiannoutsos (1999) examinam o problema da elicitação de distribuições *a priori* informativas para parâmetros de regressão, assim como para seleção Bayesiana de variáveis na regressão logística. A construção da *priori* proposta é baseada em estudos anteriores que medem a mesma variável resposta e covariáveis do estudo em questão.

Chen e Ibrahim (2000) estabeleceram *prioris* importantes para parâmetros de modelos de séries temporais. Assim como no artigo de Chen, Ibrahim e Yiannoutsos (1999), citado anteriormente, a construção d*a priori* para o caso proposto é baseada na noção da existência de dados históricos.

Deste modo, é bastante útil comentar que, na estatística clássica o(s) parâmetro(s), θ, que pertence(m) ao espaço paramétrico, Θ, é um escalar ou vetor de escalares desconhecidos fixo; enquanto que no ponto de vista bayesiano, este mesmo é um escalar ou vetor aleatório. Neste sentido, Paulino et al. (2003) comentam que este parâmetro ou vetor aleatório é incerto e que esta incerteza deve ser quantificada em termos de probabilidade.

Em resumo, de acordo com Paulino et al. (2003), a distribuição *a priori* pode traduzir-se formalmente por uma distribuição de probabilidade, geralmente subjetiva, para θ, seja $\pi(\theta)$.

A atualização do conhecimento traduzido pela distribuição *a priori*, após observar os dados se dá através da Distribuição *a posteriori*.

2.5.3 Distribuição *a posteriori*

A Distribuição *a posteriori* é a atualização da informação inicial sobre o parâmetro, ou seja, da distribuição *a priori*, através do Teorema de Bayes.

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}$$

em que $\pi(x|\theta)$ é a distribuição *a posteriori* de X dado o parâmetro θ,

$f(x|\theta)$ é a função de densidade de probabilidade de X,

$\pi(\theta)$ é a distribuição *a priori* de θ, e,

o denominador desta expressão é distribuição marginal de X.

A distribuição *a posteriori* revela o conhecimento do parâmetro desconhecido, quando se possui um conhecimento prévio, descrito através da *priori*, e das informações contida nos dados, traduzidas pela função de densidade de probabilidade (no caso, $f(x|\theta)$).

Definidos os conceitos de Distribuição *a priori* e Distribuição *a posteriori*, introduz-se a inferência preditiva, que é uma abordagem centrada nas predições. Ou seja, as quantidades desconhecidas de interesse não são os parâmetros, mas sim as variáveis aleatórias futuras.

2.6 Abordagem preditiva em modelos

Tanto na abordagem Clássica quanto na Bayesiana, muitas vezes, a inferência sobre o parâmetro desconhecido não é necessariamente o objeto do estudo, mas somente uma ferramenta ou caminho para se fazer predições de dados futuros.

A filosofia preditiva envolve o uso do sistema de crenças sobre o que é observável e não observável na ciência e a filosofia da metodologia científica que implementa esse sistema de crenças, ou seja, o desenvolvimento de modelos e hipóteses estatísticas, baseados em dados observados, que podem ser então usados para predizer novas observações.

Geisser e Eddy (1979) comentam em sua introdução, duas questões que têm confrontado pesquisadores: “Quando existem m modelos candidatos $M_1, \ldots, M_m$ qual deles explica melhor um dado conjunto de dados?” e “Qual destes modelos $M_1, \ldots, M_m$ gera melhores predições para observações futuras do mesmo processo que gerou o dado conjunto de dados?”. A segunda questão, apesar de mais difícil de responder, justifica a razão de ter todos esses modelos.

De acordo com Press (2003), predizer novas observações advindas de experimentos científicos tem sido o principal objetivo da ciência experimental por séculos.

Problemas preditivos, portanto, são aqueles em que as variáveis desconhecidas de interesse são variáveis aleatórias futuras. Formular modelos estatísticos para descrever determinado acontecimento e/ou experimento, é a alternativa comumente usada na ciência para estudar determinado fenômeno. A partir desses modelos, como dito anteriormente e ratificado por vários autores, é interessante prever observações futuras deste fenômeno, a fim de entendê-lo e tomar decisões diante dessas previsões.

Martini e Spezzaferri (1984); Ibrahim e Laud (1994); Chen e Ibrahim (2000); Ibrahim, Chen e Sinha (2001); Chen e Ibrahim (2004); Ibrahim, Chen e Sinha (2004) usam da abordagem preditiva em seus trabalhos.

2.7 Inferência preditiva Bayesiana

Problemas preditivos, na estatística, são aqueles em que as quantidades desconhecidas de interesse são variáveis aleatórias futuras.

Frequentemente, as inferências sobre os parâmetros do modelo postulado não são um fim em si, mas antes, um meio visando predizer dados amostrais futuros. (PAULINO et al., 2003).

Press (2003) diz que cientistas testam uma teoria usando da formulação matemática, chamada "modelo", e então predizem valores de observações futuras baseadas neste modelo. Mas, sabe-se que os valores preditos não serão os mesmos dos observados anteriormente por duas razões apresentadas pelo autor, a saber:

1ª) a natureza é geralmente mais complexa, e esta complexidade não é facilmente traduzida por parâmetros isoladamente no modelo.

2ª) observações tem de ser medidas, e medições sempre têm erros associados.

Essa diferença entre o predito e o observado é o chamado "erro de predição", que em "bons experimentos" tem um valor pequeno. Daí, vêm as questões apresentadas por Press (2003): "o erro de predição é muito grande?", "quão grande é muito grande?". Ainda de acordo com o autor, a qualidade de qualquer teoria científica é medida por quão bem a teoria prediz observações futuras.

Suponha que se queira comparar duas teorias, seja *teoria A* e *teoria B*, para prever uma nova observação de uma realização de uma variável aleatória X. A probabilidade preditiva de uma observação futura Y, dada esta observação X, é a média ponderada dos valores preditos de Y assumindo que a *teoria A* é a correta, e os valores de Y dado que a *teoria B* é a correta. Em termos:

$$P[Y|X] = P[Y|teoria\ A]P[teoria\ A|X] + P[Y|teoria\ B]P[teoria\ B|X]$$

em que : X são as variáveis aleatórias observadas, Y são as variáveis aleatórias preditas e $P[teoria\ A|X]$ e $P[teoria\ B|X]$ são as probabildiades *a posteriori* obtidas pelo teorema de Bayes das duas teorias, dados os resultados X.

Exemplo 2: Suponha-se que já tenham sido calculadas as probabilidades das teorias dados os dados X . Sendo $P[teoria\ A|X] = 0{,}2$ e

$P[teoria\ B|X] = 0{,}8$. E, que se deseja estudar um experimento em particular em que só dois resultados são possíveis: ou o efeito é observado (sucesso) ou não é observado (falha) e que este experimento pode ser repetido muitas vezes.

Dados que a probabilidade p_A de sucesso da *teoria A* é $p_A = 0{,}1$, se esta é a correta; e que a probabilidade p_B da *teoria B* é $p_B = 0{,}5$, se esta é a correta. Assim, as probabilidades de Y dadas as teorias são:

$$P[Y|teoria\ A] = P[Y|p_A = 0{,}1] = \binom{2}{Y}(0{,}1)^Y(0{,}9)^{2-Y}$$

$$P[Y|teoria\ B] = P[Y|p_B = 0{,}5] = \binom{2}{Y}(0{,}5)^2$$

$$\Rightarrow P[Y|X] = (0{,}2)\binom{2}{Y}(0{,}1)^Y(0{,}9)^{2-Y} + (0{,}8)\binom{2}{Y}(0{,}5)^2$$

O exemplo acima mostra a distribuição preditiva para variáveis aleatórias discretas.

Na situação em que as variáveis aleatórias são contínuas, Berger (1985) comenta uma típica situação que envolve a predição de uma variável aleatória Z, com densidade $g(z|\theta)$, com θ desconhecido, onde existem dados disponíveis, x, com densidade $f(x|\theta)$. Por exemplo, x poderia ser dados de um estudo de regressão, e deseja-se predizer a futura variável resposta, Z.

De Berger (1985) teremos:

Assumindo X e Z independentes. A ideia da inferência preditiva Bayesiana é que, já que $\pi(\theta|x)$ é a distribuição *a posteriori* de θ, então $g(z|\theta)\pi(\theta|x)$ é a distribuição conjunta de z e θ dado x, e integrando com relação a θ teremos a distribuição de z dado x.

Definição 10: *a densidade preditiva de* Z *dado* x*, quando a priori para* θ *é* $\pi(\theta)$ *é definida por:*

$$p(z|x) = \int g(z|\theta)\pi(\theta|x)d\theta = \int g(z|\theta)\pi(\theta)p(x|\theta)d\theta$$

em que $p(z|x)$ é a densidade preditiva da variável aleatória Z,

$g(z|\theta)$ é a densidade de z dado θ, e,

$\pi(\theta|x)$ é a distribuição *a posteriori* de θ.

Importante salientar que Z tem a mesma densidade de X.

A partir do conceito de densidade preditiva, introduz-se a abordagem preditiva para a seleção de modelos.

2.8 Seleção preditiva de modelos

Laud e Ibrahim (1995) comentam três problemas na seleção de modelos, quais sejam: (a) selecionar um modelo adequado entre uma classe de possíveis modelos; (b) escolher adequadas transformações do preditor e/ou variáveis resposta em regressão linear, e, (c) selecionar funções de variância apropriadas em modelos lineares heterocedástico.

Dentre muitos critérios para seleção de modelos propostos na literatura, o Critério de Informação de Akaike, AIC, e o Critério de Informação Bayesiano, BIC, são amplamente aceitos (LAUD; IBRAHIM, 1995). Ainda de acordo com os autores citados, um problema inerente a esses critérios é que eles não permitem a inclusão de informações prévias, *prioris*, para a escolha do modelo mais adequado; e que, as definições e/ou calibrações baseiam-se fortemente em considerações assintóticas. Diante disso, propõem três critérios que podem ser usados para a seleção de modelos que dão ênfase aos fatores observáveis ao

invés dos parâmetros e são baseados na densidade preditiva Bayesiana. A medida L é objeto de estudo neste trabalho.

Para introduzir as medidas, os autores consideram o problema de seleção de variáveis em regressão linear.

Laud e Ibrahim (1995) começam a partir de uma matriz de incidência composta por uma coluna de 1´s para o intercepto seguida de k colunas, cada uma representando uma variável independente. O modelo completo é definido por:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{1.4}$$

em que Y é um vetor de respostas n-dimensional, β é um vetor de coeficientes de regressão de tamanho $k+1$ e ϵ é um vetor de erros aleatórios n-dimensional.

Isto é, (1.4) é:

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \dots & x_{1k} \\ 1 & x_{21} & x_{22} \dots & x_{2k} \\ & & \dots & \\ 1 & x_{n1} & x_{n2} \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_n \end{bmatrix}$$

Usualmente, ϵ tem distribuição normal multivariada com média 0 e matriz de precisão τI, onde τ é um escalar positivo e I é uma matriz identidade com dimensão $n \times n$. Em termos:

$$\epsilon|\tau \sim N_{o_n}(0, \tau^{-1} I)$$

A variância do erro enfatiza o caso trabalhado, qual seja, homocedástico com covariâncias iguais a zero.

O interesse é nos 2^k modelos obtidos a partir de (1.4) selecionando vários subconjuntos das últimas k colunas da matriz X e modificando o comprimento de β de acordo com a modificação feita em X.

Mitchell e Beauchamp (1988), em seu artigo sobre a seleção bayesiana de variáveis em modelos de regressão linear apresentam a justificativa para a restrição desse conjunto de variáveis na matriz X, qual seja:

> A procura por um sub-modelo melhor é chamada *seleção de variáveis* ou *seleção de subconjunto*. Algumas razões para esta procura são (a) expressar a relação entre y e os preditores tão simples quanto possível, (b) reduzir o custo futuro da predição, (c) identificar preditores importantes ou negligenciáveis, ou (d) aumentar a precisão dos estimadores estatísticos e predições. (MITCHELL; BEAUCHAMP, 1988, p. 2)

Seja m o subconjunto de inteiros contendo 0, e seja k_m o número de elementos de m. Este último identifica o modelo com intercepto e uma escolha específica de $k_m - 1$ variáveis preditoras. Assim, o modelo em (1.4) pode ser escrito:

$$Y = X_m \beta^{(m)} + \epsilon, \quad m \in \mathcal{M}$$

(1.5)

em que $\mathcal{M}$ é o conjunto de todos os 2^k modelos considerados, X_m é a matriz de incidência sob o modelo m, de posto completo e dimensão $n \times k_m$; e $\beta^{(m)}$ o vetor de coeficientes.

Escolher entre os modelos na equação acima é o objetivo dos métodos de seleção de variáveis, ou seja, os métodos de seleção de variáveis vão selecionar quais variáveis independentes explicarão a variável dependente.

Adota-se a abordagem preditiva Bayesiana que nos permite diminuir a importância dos parâmetros e focar nas observações, para seleção de modelos considerando os modelos de probabilidades para as observações Y condicionadas em cada modelo m e vetor de parâmetros $\theta^{(m)}$. Então a expressão (1.5) pode ser representada por:

$$p_m(y|\theta), \quad m \in \mathcal{M}, \quad \theta \in \Theta^{(m)}$$

(1.6)

em que $\Theta^{(m)}$ é o espaço paramétrico para o modelo m.

Em relação às *prioris* para $\theta|m$, serão construídas de alguma forma automatizada a partir de uma predição anterior para Y e esta distribuição *a priori* não será usada no espaço de modelos $\mathcal{M}$.

Suponha-se agora que *a priori*, $\pi_m(\theta)$, tenha sido especificada para cada θ com $m \in \mathcal{M}$. Então, *a posteriori* para cada parâmetro sob cada modelo m, dados $Y = y$ é dada por:

$$\pi_m(\theta|y) = \frac{\pi_m(\theta)p_m(y|\theta)}{\int \pi_m(\theta)p_m(y|\theta)d\theta^{(m)}}$$

(1.7)

em que $\pi_m(\theta)$ é a *priori* para o parâmetro θ no modelo m,

$\pi_m(\theta|y)$ é *a posteriori* do parâmetro θ observados os dados no modelo m,

$p_m(y|\theta)$ é a função de densidade dos dados no modelo m.

$\int \pi_m(\theta)p_m(y|\theta)d\theta^{(m)}$ é a densidade marginal ou preditiva dos dados no modelo m.

Da mesma forma que em Laud e Ibrahim (1995), usando o artifício de replicar o experimento e denotando por Z o vetor de resposta que pode resultar desta replicação. Sendo que os parâmetros θ do modelo, neste caso, são $(\beta^{(m)},\tau)$ e cada modelo m especifica a matriz de preditores X_m.

A ideia de usar um vetor de respostas futuras Z para desenvolver um critério para avaliação de um modelo ou comparação de vários modelos tem sido bastante motivada na literatura por Geisser (1993) e referências como Ibrahim e Laud (1994), Laud e Ibrahim (1995), e Gelfand e Ghosh (1998). O experimento replicado imaginado faz Y e Z diretamente comparáveis e permutáveis *a priori*. (IBRAHIM; CHEN; SINHA, 2001)

O experimento conceitual replicado tem o mesmo desenho da matriz X do experimento realmente realizado (com as observações y).

Sob o modelo $m \in \mathcal{M}$, tem-se a matriz X_m. A densidade preditiva de Z sob o modelo m é:

$$p_m(z|y) =_{\forall\theta} \int p_m(z|\theta)\pi_m(\theta|y)d\theta \tag{1.8}$$

em que $p_m(z|y)$ é chamada *Densidade Preditiva do Experimento Replicado*.

$p_m(z|\theta)$é a densidade de z dado θ e o modelo m, e,

$\pi_m(\theta|y)$é *a posteriori* de θ dado y e o modelo m.

Esta densidade preditiva será denominada por DPER.

Em Ibrahim e Laud (1994) os autores definem a mesma densidade fazendo os parâmetros, θ, do modelo m iguais a $(\beta^{(m)},\tau)$, que é a apresentada a seguir. A densidade preditiva do experimento replicado para o modelo (1.5) é:

$$p_m(z|y) = \int\int p_m(z|\beta^{(m)},\tau)\pi_m(\beta^{(m)},\tau|y)d\beta^{(m)}d\tau$$

Para facilitar a notação, os autores nomeiam esta densidade simplesmente por f_m.

O experimento replicado é uma "ferramenta" imaginária que coloca a densidade preditiva para uso inferencial, adaptando a filosofia adotada em Geisser (1971).

Como já foi citada, esta replicação imaginária faz Y e Z comparáveis e permutáveis *a priori*. Ainda, os autores comentam que os parâmetros no modelo têm papel mínimo na replicação. E, assim como é feito em Ibrahim e Laud (1994), a partir das considerações feitas acima sobre a replicação, densidades e predições, parece claro que bons modelos deveriam fazer predições próximas ao que foi observado no experimento idêntico.

Box (1980 citado por GELFAND; GOSH, 1998) citam dizendo que a abordagem Bayesiana emprega distribuições preditivas para a "crítica do modelo à luz dos dados atuais". E que, examinando uma coleção de modelos, distribuições preditivas serão comparáveis enquanto distribuições *a posteriori* não. Além disso, parece natural avaliar a performance de um modelo comparando o que é predito com o que foi observado.

3 A MEDIDA L

Em Ibrahim e Laud (1994), Laud e Ibrahim (1995) e Ibrahim, Chen e Sinha (2001), a Medida L é construída a partir da distribuição preditiva *a posteriori* dos dados e pode ser escrita como a soma de dois componentes, um envolve a média desta distribuição e outro envolve as variâncias. É um critério bayesiano de ajuste estatístico que mede o desempenho de um modelo pela combinação de quão próximo as predições estão dos dados observados e as variabilidades das predições.

Chen e Ibrahim (2000) definem a medida para dados de séries temporais, justificada por ser importante avaliar quão bem um modelo pode predizer futuras observações de uma série. A forma da medida proposta neste artigo, combina tanto a variabilidade preditiva do modelo quanto a performance deste quanto aos futuros pontos observados. Ainda comentam que uma vantagem desta medida é que é bem definida sob distribuições *a priori* impróprias.

Gelfand e Gosh (1998) apresentam a medida como uma função de perda quadrática.

Seja um experimento com as observações $y = (y_1, \dots, y_n)$ com densidade conjunta amostral dada por $p(y|\theta)$, em que θ é um vetor de parâmetros. As observações, $Y_{i's}$ podem ser totalmente observadas, censuradas a direita – tempos de falha ou censuras – ou ainda de censura intervalar $[a_{l_i}, a_{r_i}]$.

Considere, ainda, valores futuros de um experimento replicado imaginário, $z = (z_1, \dots z_2)$ com mesma densidade amostral de $Y_{i's}$. Replicar um experimento imaginário é uma ferramenta que coloca a densidade preditiva para uso inferencial e ainda, faz y e z diretamente comparáveis.

Diante disso, Laud e Ibrahim (1995) e Ibrahim e Laud (1994) citados por Ibrahim, Chen e Sinha (2001) definem a medida como a esperança do quadrado da distância Euclidiana entre Y e Z:

$$L_m^2 = E[(Z - Y)'(Z - Y)]$$

(1.9)

Em que a esperança é calculada com relação à distribuição preditiva *a posteriori* de $z|y$ dada por:

$$p(z|y) = \int p(z|\theta)p(\theta|y)d\theta$$

em que $p(z|\theta)$ é a distribuição amostral de z; $p(\theta|y)$ é a distribuição *a posteriori* de θ.

Ibrahim e Laud (1994), Laud e Ibrahim (1995) e Ibrahim, Chen e Sinha (2001) apresentam a decomposição da medida como a soma de dois componentes, um envolvendo as médias da distribuição preditiva e outro envolvendo as variâncias dos valores futuros, como apresentado em (2.0):

$$L_m^2 = \sum_{i=1}^{n}[\{E(Z_i) - y_i\}^2 + var(Z_i)]$$

(2.0)

A primeira componente de L_m^2 é interpretada como um viés quadrado que é compensado pela variância. Então, modelos que produzem predições viesadas podem ser adequados se este viés for compensado pela redução na variância. (IBRAHIM; LAUD, 1994).

Ibrahim, Chen e Sinha (2001) apresentam a forma geral da medida:

$$L(y, b, k) = E[(z - y)'(z - y)] + k(y - b)'(y - b)$$

em que a esperança também é tomada com respeito a distribuição preditiva de $z|y$. E, $b = (b_1, ..., b_n)$ é um vetor de posição arbitrário e k é um escalar não negativo que pondera a discrepância entre os valores futuros em relação aos observados. Ainda, se $b = y$, temos a medida L como apresentada na expressão (1.9). O caso em que $k = 0$ pode ser interpretado como a perda quadrática medida nos valores futuros.

Em notação escalar:

$$L(y, b, k) = \sum_{i=1}^{n} \{Var(z_i|y) + (\mu_i - b_i)^2 + k(y_i - b_i)^2\}$$

em que $\mu_i = E(z_i|y)$.

Selecionando b como o valor que minimiza $L(y, b, k)$:

$$\hat{b} = (1 - v)\mu_i + vy_i$$

em que $v = \frac{k}{k+1}$, então:

$$L(y, v) = \sum_{i=1}^{n} Var(z_i|y) + v\sum_{i=1}^{n}(\mu_i - y_i)^2$$

De Ibrahim, Chen e Sinha (2001), se $o < v < 1$, $v = 0$ se $k = 0$, e $v \to 1$ se $k \to \infty$. Em Ibrahim e Laud (1994) e Laud e Ibrahim (1995), $v = 1$, ou seja, o "peso" do viés e da variância é igual.

4 APLICAÇÃO DA MEDIDA L

A aplicação da Medida L será em um modelo de regressão linear múltipla, como apresentado na seção 2.8, em que se deseja selecionar variáveis independentes que melhor explicam o comportamento da variável dependente observada e que ainda faça boas previsões.

As *prioris* utilizadas são a não informativa de Jeffreys e a Conjugada Normal-Gama.

Os conjuntos de dados serão apresentados na seção 4.4

4.1 Modelo de regressão linear múltipla

Um modelo de Regressão é aquele em que uma variável, Y, chamada de variável dependente, é explicada em função da variação de outra ou outras variáveis, X, chamada(s) independente(s). O objetivo é estabelecer uma relação entre a variável dependente e a(s) variável(eis) independente(s).

Nos Modelos de Regressão Linear Múltipla descreve-se Y como a soma de uma parte determinística e uma aleatória. Na parte determinística, mais geral, pode-se expressar o valor esperado de Y como função de várias variáveis regressoras.

Matricialmente, um modelo de regressão linear múltipla é representado na forma:

$$\boldsymbol{Y} = \boldsymbol{X\theta} + \boldsymbol{\epsilon}, \text{ e } \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

em que $\boldsymbol{\theta}$ é vetor de coeficientes para as variáveis regressoras;

$\boldsymbol{X}$ é o vetor de variáveis regressoras;

$\boldsymbol{\epsilon}$ é a parte aleatória da regressão, representada pelo erro aleatório com distribuição Normal, com vetor de médias $\mathbf{0}$ e matriz de variância e covariância $\sigma^2\boldsymbol{I}$, em que $\boldsymbol{I}$ é matriz identidade $n \times n$.

Convém observar que o número de colunas de $\boldsymbol{X}$ é igual ao número de elementos em $\boldsymbol{\theta}$ e o número de linhas é o tamanho da amostra. A primeira coluna, dessa matriz $\boldsymbol{X}$, é um vetor de dimensão n, cujos elementos são todos iguais a 1, que corresponde ao intercepto; e as demais são vetores de dimensão também n formados pelos valores correspondentes às observações da amostra.

O vetor $\boldsymbol{Y}$, correspondente aos valores das variáveis regressoras dados em $\boldsymbol{X}$, tem distribuição de probabilidade Normal Multivariada, de ordem n, com vetor de médias e matriz de variâncias e covariâncias, dados, respectivamente por:

$$E[\boldsymbol{Y}] = E[\boldsymbol{X\theta} + \boldsymbol{\epsilon}]$$
$$= E[\boldsymbol{X\theta}] + E[\boldsymbol{\epsilon}]$$
$$= \boldsymbol{X\theta} + \mathbf{0}$$
$$= \boldsymbol{X\theta}$$

$$V[\boldsymbol{Y}] = V[\boldsymbol{X\theta} + \boldsymbol{\epsilon}]$$
$$= V[\boldsymbol{X\theta}] + V[\boldsymbol{\epsilon}]$$
$$= \sigma^2\boldsymbol{I}$$

$$\Rightarrow \boldsymbol{Y} \sim N_n(\boldsymbol{X\theta}, \sigma^2\boldsymbol{I})$$

Algumas pressuposições devem ser atentamente observadas:

a) A relação entre $\boldsymbol{X}$ e $\boldsymbol{Y}$ é linear;

b) Os valores de $\boldsymbol{X}$ são fixos;

c) $E[\boldsymbol{\epsilon}] = \boldsymbol{0}$;

d) V[$\boldsymbol{\epsilon}$] $= E[\boldsymbol{\epsilon^2}] = \boldsymbol{I}\sigma^2$, ou seja, os erros são homocedásticos;

e) Independência dos erros, ou seja, os erros são não-correlacionados; e

f) Os erros têm distribuição Normal.

A estimação dos coeficientes $\boldsymbol{\theta}$ pode ser feita através do Método de Mínimos quadrados, que consiste em tornar mínima a soma de quadrados dos erros, e tem como resultado:

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

Partindo do modelo linear Gauss-Markov, duas consequências advêm a partir do Sistema de Equações Normais, a saber: $\widehat{\boldsymbol{\theta}}$ é um estimador não viciado para $\boldsymbol{\theta}$; e, a matriz de variâncias e covariâncias de $\widehat{\boldsymbol{\theta}}$ é dada por $V(\widehat{\boldsymbol{\theta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$. Em resumo:

$$\widehat{\boldsymbol{\theta}} \sim N_n(\boldsymbol{\theta}; (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2)$$

Diante das especificações apresentadas acima sobre o Modelo de Regressão Linear Múltipla e seus parâmetros, pode-se propor *prioris*.

4.2 Distribuições *a priori*

Antes de estabelecer as *prioris*, é importante definir a distribuição de $\boldsymbol{Y}$. Apresentada por Box e Tiao (1992), é importante para o cálculo da Densidade Preditiva, DPER, que será calculada usando a *posteriori*. A *posteriori*, como apresentada na seção 2.4.3, por definição é proporcional ao produto da *priori* com a distribuição dos dados, $\boldsymbol{Y}$.

As distribuições *a priori* apresentadas estão caracterizadas em Press (2003), e também são apresentadas por Laud e Ibrahim (1995), bem como as distribuições *a posteriori* consequentes e os resultados para a densidade preditiva, DPER.

4.2.1 Distribuição de $\boldsymbol{Y}$

Considere a distribuição para $\boldsymbol{Y}$ Normal multivariada, com densidade:

$$p(\boldsymbol{Y}|\boldsymbol{\theta},\sigma^2) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \sigma^{-n} \exp[-\frac{1}{2\sigma^2}(\boldsymbol{Y}-\boldsymbol{X\theta})'(\boldsymbol{Y}-\boldsymbol{X\theta})]$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}[\nu s^2 + (\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})]\right\}$$

em que $\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$

$\nu = n - k$

$s^2 = \left(\frac{1}{\nu}\right)(\boldsymbol{Y}-\widehat{\boldsymbol{Y}})'(\boldsymbol{Y}-\widehat{\boldsymbol{Y}})$ e $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\theta}}$

E, algumas observações:

a) $\widehat{\boldsymbol{\theta}}$ é um vetor de estatísticas conjuntamente suficientes para o vetor de parâmetros $\boldsymbol{\theta}$ se σ^2 for conhecido;

b) $\widehat{\boldsymbol{\theta}}$ e σ^2 são estatísticas conjuntamente suficientes para $(\boldsymbol{\theta}, \sigma^2)$;

c) $\hat{\boldsymbol{\theta}}$ tem distribuição multivariada Normal: $\hat{\boldsymbol{\theta}} \sim N_n[\boldsymbol{\theta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]$; e é solução de Mínimos Quadrados do Sistema de Equações Normais para o modelo de regressão linear múltipla, como apresentado em 4.1; e,

d) $\frac{\nu s^2}{\sigma^2} \sim \chi^2_\nu$ e é distribuído independentemente de $\hat{\boldsymbol{\theta}}$.

Supondo o posto da matriz $\boldsymbol{X}$ igual a k, σ conhecido; e que $E[\boldsymbol{Y}] = \boldsymbol{X\theta}$, a função de verossimilhança é da forma:

$$L(\boldsymbol{\theta}, \sigma|\boldsymbol{Y}) \propto \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X\theta})'(\boldsymbol{Y} - \boldsymbol{X\theta})\right]$$

A forma quadrática $(\boldsymbol{Y} - \boldsymbol{X\theta})'(\boldsymbol{Y} - \boldsymbol{X\theta})$ pode ser escrita na forma:

$$(\boldsymbol{Y} - \boldsymbol{X\theta})'(\boldsymbol{Y} - \boldsymbol{X\theta}) = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})'(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$$

E já que $\hat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ e $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\theta}}$, $(\boldsymbol{Y} - \hat{\boldsymbol{Y}})'(\boldsymbol{Y} - \hat{\boldsymbol{Y}})$ é uma função dos dados que não envolvem os parâmetros $\boldsymbol{\theta}$. Desta forma, pode-se escrever a função de verossimilhança:

$$L(\boldsymbol{\theta}, \sigma|\boldsymbol{Y}) \propto \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right]$$

(2.1)

Agora, supondo σ e os parâmetros $\boldsymbol{\theta}$ desconhecidos, a função de verossimilhança pode ser escrita como:

$$L(\boldsymbol{\theta}, \sigma|\boldsymbol{Y}) \propto \left(\frac{s}{\sigma}\right)^{n} \exp\left[-\frac{(n-k)s^2}{2\sigma^2} - \frac{s^2}{2\sigma^2}\frac{(\boldsymbol{\theta}-\ddot{\boldsymbol{\theta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\theta}-\ddot{\boldsymbol{\theta}})}{s^2}\right]$$

(2.2)

em que,

$$s^2 = \frac{1}{(n-k)}(\boldsymbol{Y}-\widehat{\boldsymbol{Y}})'(\boldsymbol{Y}-\widehat{\boldsymbol{Y}})$$

Observe que, tanto quando os parâmetros são conhecidos ou desconhecidos, a verossimilhança tem a mesma forma, como apresentado nas expressões (2.1) e (2.2).

4.2.2 Distribuição *a priori* não informativa

Para os modelos da seção 4.2.1, tanto quando os parâmetros são conhecidos ou quando alguns são desconhecidos, *prioris* são propostas com o objetivo de usar os conceitos apresentados na seção 2.8.

A *priori* não informativa traduz a "ignorância" a *priori* sobre os parâmetros e, como comentado na seção 2.5.2, não interfere na análise, já que reflete indiferença sobre os valores dos parâmetros em estudo. Neste sentido, uma *priori* não informativa pode ser utilizada para o cálculo da Medida L, no sentido de não influenciar nas predições.

Portanto, antes de estabelecer *prioris* informativas, considere a distribuição *a priori* não informativa de Jeffreys modificada definida por:

Definição 11: Considere a amostra aleatória $\boldsymbol{Y} = (y_1, y_2, \ldots, y_3)$ com função de densidade $p(\boldsymbol{Y}|\boldsymbol{\theta})$. A distribuição *a priori* de Jeffreys para o caso multiparamétrico é dada por:

$$h(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{\frac{1}{2}}$$

em que $I(\boldsymbol{\theta})$ é a matriz de Informação de Fisher de $\boldsymbol{\theta}$ que é dada por $I(\boldsymbol{\theta}) = E\left[\left(\frac{\partial \log f(Y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^2\right]$.

Esta *priori* é adotada para os parâmetros justificada pela regra de Jeffreys para problemas multiparamétricos, apresentada por Box e Tiao (1992), a saber: a distribuição *a priori* não informativa para um conjunto de parâmetros é tomada ser proporcional a raiz quadrada do determinante da matriz informação de Fisher.

Mais especificamente, na seleção de *prioris* em problemas envolvendo parâmetros de posição e escala, uma *priori* não informativa para ambos os parâmetros é aquela para qual $\boldsymbol{\theta}$ e $\log\sigma$ são aproximadamente uniformes locais, da forma:

$$p(\theta, \log\sigma) \propto c$$

ou equivalente,

$$p(\theta, \sigma) \propto \sigma^{-1}$$

Esta *priori* é considerada em ambos os casos, quando a média tem distribuição Normal com média $\boldsymbol{\theta}$ e desvio padrão σ; e nos casos em que o modelo é linear Normal e com desvio padrão, σ, desconhecido.

Ainda, geralmente é apropriado considerar os parâmetros de posição, θ, e escala, σ, são distribuídos independentemente. Qualquer ideia *a priori* que se pode ter sobre a distribuição do parâmetro de posição não deveria ser influenciada sobre a ideia do valor do parâmetro de escala. Então se pode

considerar $p(\boldsymbol{\theta}|\sigma) = p(\boldsymbol{\theta})$. E considerando o caso da distribuição Normal, quando σ é conhecido, a *priori* não informativa para $\boldsymbol{\theta}$ é obtida, fazendo $p(\boldsymbol{\theta}|\sigma)$ localmente uniforme; e já que os parâmetros são independentes, implica que $p(\boldsymbol{\theta})$ deveria ser uniforme, logo,

$$p(\boldsymbol{\theta}, \sigma) = p(\boldsymbol{\theta})p(\sigma) \propto 1/\sigma$$

Multiplicando esta *priori* com a distribuição Normal dos dados, a distribuição *a posteriori* correspondente é então:

$$p(\boldsymbol{\theta}, \sigma|\boldsymbol{y}) \propto \sigma^{-(n+1)} \exp\left[-\frac{(n-k)s^2}{2\sigma^2} - \frac{(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})}{2\sigma^2}\right]$$

com $-\infty < \theta_i < +\infty$, $i = 1, \ldots, k$; $\sigma > 0$; e, $s^2 = \frac{1}{(n-k)}(\boldsymbol{y}-\hat{\boldsymbol{y}})'(\boldsymbol{y}-\hat{\boldsymbol{y}})$. E que pode ser escrita na forma:

$$p(\boldsymbol{\theta}, \sigma|\boldsymbol{y}) \propto \sigma^{-(n+1)} \exp\left[-\frac{(\boldsymbol{y}-\hat{\boldsymbol{y}})'(\boldsymbol{y}-\hat{\boldsymbol{y}}) + (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})}{2\sigma^2}\right]$$

(2.3)

A distribuição *a posteriori* (2.3), dada acima, é importante no cálculo da Distribuição Preditiva do Experimento Replicado apresentada na seção 2.7 na equação (1.8), já que a *posteriori* para os parâmetros é multiplicada pela verossimilhança dos dados $\boldsymbol{z}$ do experimento replicado.

A distribuição *a posteriori* de $\boldsymbol{z}|\boldsymbol{y}$, DPER, será uma distribuição t multivariada:

$$\boldsymbol{z} \sim t_n(\nu, \boldsymbol{\mu}, \Sigma)$$

em que $\nu = n - k$, tal que n é o tamanho da amostra e k é o número de coeficientes da regressão;

$\boldsymbol{\mu} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ é o parâmetro de posição;

$\Sigma = s^2(\boldsymbol{I} - P)$ é a matriz de dispersão, tal que $s^2 = (n-k)^{-1}\boldsymbol{y}'(\boldsymbol{I} - P)\boldsymbol{y}$, e $P = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ é o projetor ortogonal no espaço coluna de $\boldsymbol{X}$.

A função de densidade $\boldsymbol{z}|\boldsymbol{y}$ é dada por:

$$p(\boldsymbol{z}|\boldsymbol{y}) \propto \left[(n-k) + (\boldsymbol{z} - P\boldsymbol{y})'\left(s^2(\boldsymbol{I} + P)\right)(\boldsymbol{z} - P\boldsymbol{y})\right]^{-\frac{(2n-k)}{2}} \quad (2.4)$$

4.2.3 Distribuição *a priori* conjugada

Para *prioris* informativas para os parâmetros do Modelo Linear Normal, adota-se uma *priori* natural conjugada para $\boldsymbol{\theta}$ e σ^2 da forma:

$$p(\boldsymbol{\theta}, \sigma^2) = p(\boldsymbol{\theta}|\sigma^2)p(\sigma^2)$$

em que $\boldsymbol{\theta}|\sigma^2 \sim N_n(\tilde{\boldsymbol{\theta}}, t\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$, sendo $\tilde{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$; e,

$\sigma^2 \sim GI(\frac{\delta}{2}, \frac{\rho}{2})$

E, fazendo $\frac{1}{\sigma^2} = \tau$, então $\tau \sim G(\frac{\delta}{2}, \frac{\rho}{2})$

$$(\boldsymbol{\theta}, \tau) \sim N - G\left(\tilde{\boldsymbol{\theta}}, \sigma^2, \frac{\delta}{2}, \frac{\rho}{2}\right)$$

$$\Rightarrow p(\boldsymbol{\theta},\tau) \propto \tau^{\frac{\delta+1}{2-1}} \exp\left[-\frac{\tau}{2}\left(\rho + c\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\right)\right)^2\right]$$

em que $c = (\boldsymbol{X}'\boldsymbol{X})^{-1}$

Multiplicando esta *priori* com a distribuição Normal dos dados, a distribuição *a posteriori* dos parâmetros correspondente é então:

$$p(\boldsymbol{\theta},\sigma^2|\boldsymbol{y}) \propto \sigma^{-2\delta+2-n} \exp\left[-\frac{1}{2\sigma^2}\left(\nu s^2 + \left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\right)'\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\theta}-\right.\right.\right.$$
$$\left.\left.\left.\widehat{\boldsymbol{\theta}}\right)\right) + \left(\rho + + \boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\right)\right)^2\right] \qquad (2.5)$$

Da mesma forma que no caso da *priori* não informativa, esta distribuição *a posteriori* dada em (2.5) é importante no cálculo da Distribuição Preditiva do Experimento Replicado apresentada na seção 2.8 na equação (1.8), já que a *posteriori* para os parâmetros é multiplicada pela verossimilhança dos dados z do experimento replicado.

E, também, a distribuição *a posteriori* de $\boldsymbol{z}|\boldsymbol{y}$, DPER, será uma distribuição t multivariada:

$$\boldsymbol{z} \sim t_n(\nu_1, \boldsymbol{\mu_1}, \Sigma_1)$$

em que $\nu_1 = n - \delta$

$\boldsymbol{\mu_1} = P[\gamma\eta + (1-\gamma)\boldsymbol{y}]$, com η é um vetor fixo independente do modelo sob consideração e $\gamma = \frac{t}{1+t}$;

$\Sigma_1 = s_1^2[\mathbf{I} + (1-\gamma)\mathrm{P}]$, com $s_1^2 = (n+\delta)^{-1}(q + \gamma p + n\sigma^2)$, sendo $q = \boldsymbol{y}'(\boldsymbol{I}-P)\boldsymbol{y}$ e $p = (\boldsymbol{y}-\eta)'P(\boldsymbol{y}-\eta)$.

$$P = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

A função de densidade $\boldsymbol{z}|\boldsymbol{y}$ é dada por:

$$p(\boldsymbol{z}|\boldsymbol{y}) \propto \left[(n+\delta) + (\boldsymbol{z}-\boldsymbol{\mu_1})'\left(s_1^2(\boldsymbol{I}+(1-\gamma)P)\right)(\boldsymbol{z}-\boldsymbol{\mu_1})\right]^{-\frac{(2n+\delta)}{2}} \quad (2.6)$$

A DPER para *priori* não informativa pode ser obtida de (2.6) fazendo $\gamma = 0, \delta = -k$.

4.3 A Medida L para modelos de regressão linear múltipla

A Medida L para Modelos de Regressão Linear Múltipla é calculada, como apresentada na seção 3.0, de acordo com a expressão (2.0), a saber:

$$L_m^2 = \sum_{i=1}^{n}[\{E(Z_i) - y_i\}^2 + var(Z_i)]$$

em que a esperança é calculada a respeito da distribuição preditiva de $\boldsymbol{z}|\boldsymbol{y}$.

A partir das DPER´s apresentadas nas seções 4.2.2 e 4.2.3, nas expressões (2.4) e (2.5) respectivamente, a Medida L é dada por:

$$L = [(1+\lambda)q + \gamma(\gamma+\lambda)p + \lambda\rho]^{\frac{1}{2}}$$

em que $\lambda = \frac{n+(1-\gamma)k}{n+\delta-2}$;

$p = (\boldsymbol{y} - \eta)'P(\boldsymbol{y} - \eta)$ representa a penalidade por uma suposição ruim *a priori* para $\boldsymbol{Y}$;

$q = \boldsymbol{y}'(\boldsymbol{I} - P)\boldsymbol{y}$ representa a soma de quadrado dos erros sob o modelo em consideração.

k é o número de coeficientes da regressão e $\gamma = \frac{t}{1+t}$ é o "peso" dado à suposta predição, e são definidos como nas seções 4.2.2 e 4.2.3 respectivamente.

4.4 Aplicação aos dados e comparação com AIC e BIC

Todos os cálculos das medidas foram feitos usando o software R, ressaltando que a Medida L não está implementada no mesmo.

4.4.1 Aplicação em dados da produção em um processo químico

Para realização desta primeira análise, o conjunto de dados avaliado foi o apresentado como exercício por Charnet et al. (2008) no capítulo 8, somente com o intuito de ilustrar a Medida L e compará-la com AIC e BIC.

Trata-se de um experimento que foi realizado para se verificar o efeito da temperatura, x_1, e da concentração, x_2, na produção (em litros), y, de um processo químico. Em resumo, selecionar o melhor modelo é responder qual das variáveis independentes ou ambas, explica melhor a produção do processo químico em questão.

Na tabela 1A dos dados, nos anexos, para o cálculo do AIC e BIC foram usadas todas as 20 observações, já para o cálculo da Medida L, foram usadas as mesmas como dados observados e, para o vetor η, as mesmas variáveis foram consideradas como o vetor Z de observações, como se o experimento tivesse sido replicado e estas seriam então as respostas obtidas.

Na Tabela 1, abaixo, observa-se que o modelo selecionado pelos Critérios de Informação de Akaike, de Schwarz e Medida L foi o modelo 1.

Ambas as variáveis explicam a produção do produto químico. O modelo selecionado foi:

$$\hat{Y} = 93{,}18 + 0{,}87X_1 + 2{,}42X_2$$

Para o cálculo da Medida L, a predição da produção do processo químico η em litros, utilizada foi:

$$\eta = (189{,}203{,}222{,}234{,}261{,}204{,}212{,}223{,}246{,}273,$$
$$220{,}228{,}252{,}263{,}291{,}226{,}232{,}259{,}268{,}294)$$

Tabela 1 Resultados do estudo de produção de um processo químico.

Nº	**Modelo**	**Medida L**	**AIC**	**BIC**
1	$\boldsymbol{y = \mu + x_1 + x_2 + \epsilon}$	31,86	80,83	138,74
2	$\boldsymbol{y = \mu + x_1 + \epsilon}$	133,48	170,14	191,88
3	$\boldsymbol{y = \mu + x_2 + \epsilon}$	137,46	173,13	194,88

4.4.2 Aplicação em dados das horas trabalhadas no departamento de contabilidade de uma empresa

Para realização desta análise, o conjunto de dados avaliado foi o apresentado como exercício por Charnet et al. (2008) no capítulo 11.

Deste conjunto de dados, com trinta observações de sete variáveis; uma dependente, y, e seis independentes, foram selecionadas para a aplicação da medida, quatro variáveis independentes e 20 observações com o objetivo de

diminuir o número de modelos possíveis e facilitar a análise e visualização dos resultados. E, da mesma forma que no exemplo anterior, na tabela 2A dos dados, nos anexos, para o cálculo do AIC e BIC foram usadas todas as 20 observações, já para o cálculo da Medida L, foram usadas as mesmas como dados observados e, para o vetor η, as mesmas variáveis foram consideradas como o vetor Z de observações, como se o experimento tivesse sido replicado e estas seriam então as respostas obtidas.

Trata-se de um estudo para se determinar as atividades mais importantes dos funcionários do departamento de contabilidade de uma empresa em que foram observadas, durante 30 dias, as seguintes variáveis: número de horas trabalhadas por dia, y; número de cheques descontados (pagos e cobrados), x_1; número de pagamentos recebidos pelos funcionários da empresa, x_2; número de documentos processados e enviados ao banco para compensação, x_3; e, número de ordens de pagamento, certificados e recibos de vendas emitidos pelos funcionários, x_4.

Na Tabela 2, observa-se que o mesmo modelo, o modelo 1, foi selecionado pelos Critérios de Informação de Akaike e de Schwarz, ou seja, obtiveram os menores valores para ambos os critérios. Porém, a Medida L selecionou outras variáveis para explicar as horas trabalhadas por dia pelos funcionários, selecionando o modelo 6.

A variáveis selecionada pelos Critérios de Akaike e Schwarz for x_1,ou seja, os funcionários do departamento gastam mais horas por dia nas atividades descontando cheques. Já a Medida L selecionou somente x_1 e x_3 para explicar y, ou seja, as horas trabalhadas são descontando cheques, e, dando ordens de pagamento e emitindo certificados e recibos de vendas.

O modelo selecionado pelo AIC e pelo BIC foi:

$$\hat{Y} = 84{,}88 + 0{,}06X_1$$

Para o cálculo da Medida L, a predição do número de horas trabalhadas por dia η, utilizada foi:

$$\eta = (130.7{,}113{,}125.4{,}131.1{,}133.2{,}178.2{,}121.1{,}135.5{,}109.8{,}119{,}$$
$$103.8{,}114.2{,}118.4{,}104.6{,}134{,}140.2{,}110.9{,}101.2{,}122.9{,}97.5)$$

O modelo selecionado pela Medida L foi:

$$\ddot{Y} = 86{,}50 + 0{,}07X_1 - 0{,}04X_3$$

Tabela 2 Resultados do estudo das atividades mais importantes do departamento de contabilidade.

N^0	**Modelo**	**Medida L**	**AIC**	**BIC**
1	$y = \mu + x_1 + \epsilon$	81,31	170,38	173,36
2	$y = \mu + x_2 + \epsilon$	88,66	173,84	176,86
3	$y = \mu + x_3 + \epsilon$	96,60	177,33	180,32
4	$y = \mu + x_4 + \epsilon$	96,65	177,35	180,33
5	$y = \mu + x_1 + x_2 + \epsilon$	81,88	172,38	176,36
6	$y = \mu + x_1 + x_3 + \epsilon$	79,52	171,53	175,52
7	$y = \mu + x_1 + x_4 + \epsilon$	81,03	171,96	175,94
8	$y = \mu + x_2 + x_3 + \epsilon$	86,18	174,44	178,43
9	$y = \mu + x_2 + x_4 + \epsilon$	88,13	175,35	179,33
10	$y = \mu + x_3 + x_4 + \epsilon$	96,71	179,09	183,08
11	$y = \mu + x_1 + x_2 + x_3 + \epsilon$	80,00	173,16	178,14
12	$y = \mu + x_1 + x_2 + x_4 + \epsilon$	81,59	173,95	178,93
13	$y = \mu + x_1 + x_3 + x_4 + \epsilon$	80,07	173,19	178,17
14	$y = \mu + x_2 + x_3 + x_4 + \epsilon$	85,67	175,92	180,90
15	$y = \mu + x_1 + x_2 + x_3 + x_4 + \epsilon$	79,70	174,73	180,71

5 CONSIDERAÇÕES FINAIS

Fundamentado em conceitos importantes como a Densidade Preditiva, e ainda permitindo que conhecimentos ou crenças, a *priori*, sejam incorporados na análise através da distribuição *a priori*, a Medida L é uma alternativa aos outros critérios já utilizados na literatura, em que a abordagem da predição é preferida às outras abordagens de ajuste de modelos.

Os resultados obtidos, nas duas análises, feitas neste trabalho, diferem. A explicação para esta diferença será apresentada em trabalhos posteriores, assim como suas aplicações em tipos de modelos diferentes, sua calibração e discutir critérios para estabelecer distribuições *a prioris*.

Os critérios de Informação de Akaike e Bayesiano diferem da Medida L já que esta última é calculada através da distribuição preditiva enquanto AIC e BIC são calculados através da verossimilhança ou de razão de posterioris somente.

A performance do modelo é dada pela combinação de quão próximos as predições estão dos dados observados e da variabilidade das predições.

Pela Teoria da Decisão, a medida L é a função perda quadrática. Neste sentido, quando da tomada de decisão, o objetivo é diminuir esta perda ao se escolher um modelo em detrimento de outro.

O desenvolvimento algébrico da função perda quadrática resultará no cálculo do Erro Quadrático Médio.

Bons modelos terão pequenos valores para a medida L_m^2, dada em (1.9). O cálculo deste erro quadrático médio está condicionado a cada ponto observado y. Em outras palavras, observados os valores iniciais Y e predizendo Z, a

variabilidade entre estes dois valores deve ser bem pequena para que o modelo seja adequado.

5.1 Estudos futuros

- Estudar o comportamento da Medida L em diferentes tipos de modelos e situações.
- Discutir o questionamento: será que *prioris* diferentes das sugeridas por Laud e Ibrahim (1995) fornecerão o mesmo comportamento da Medida L quanto à comparação de modelos?
- Dada a ausência de um "comando" ou algoritmo para o cálculo da medida L nos pacotes do software livre R, outro objetivo é desenvolve-lo para permitir a aplicação, quando conveniente.
- Estudar o comportamento assintótico da Medida L, quando comparado com outros critérios de comparação de modelos como AIC, BIC etc.

REFERÊNCIAS

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, Boston, v.9, n.6, p. 716-723, Dec. 1974.

Bayes, Thomas. 1763. ‘‘An essay towards solving a problem in the doctrine of chances.’’ **Philosphical Transactions of the Royal Society** 53:370--418.

BERGER, J. O. **Statistical Decision Theory and Bayesian Analysis.**2ª ed. New York: Springer-Verlag, 1985. 617p.

BOLFARINE, H.; SANDOVAL, M. C. **Introdução à inferência estatística.** São Paulo: Sociedade Brasileira de Matemática, 2000. 125 p.

BOX, G. E. P.; TIAO, G. C. **Bayesian Inference in statistical analysis.** Wiley Classics Library, 1992. 588 p.

CHARNET, R. et al. **Análise de modelos de regressão linear:** com aplicações.2. ed. Campinas, SP: Editora da UNICAMP, 2008. 356 p.

CHEN, M-H.; DEY, D. K.; IBRAHIM, J. G. Bayesian criterion based model assessment for categorical data. **Biometrika.** v. 91, n. 1, p. 45-63, 2004.

CHEN, M-H.; IBRAHIM, J. G. Bayesian Predictive Inference for Time Series Count Data. **Biometrics**, v. 56, p. 678-685, 2000.

CHEN, M-H.; IBRAHIM, J. G.; YIANNOUTSOS, C. Prior Elicitation, Variable Selection and Bayesian Computation for Logistic Regression Models. **Journal of the Royal Statistical Society Series B**, v. 61, n.1, p. 223-242, 1999.

EMILIANO, P. C.; VIVANCO, M. J. F; MENEZES, F. S. M.; AVELAR, F. G. Fundamentos e comparação de critérios de informação: Akaike and Bayesian. **Rev.Bras. Biom**, São Paulo, v.27, n.3, p.394-411, 2009.

GEISSER, S.; EDDY, W. F. A predictive approach to model selection. **Jounal of the American Statistical Association**, v. 74, n. 365, p. 153-160, 1979.

GELFAND, A. E.; GOSH, S. K. Model Choice: A minimum posterior predictive loss approach. **Biometrika**, v.85, n.1, p.1-11, 1998.

IBRAHIM, J. G.; CHEN, M-H.; SINHA, D. Criterion-based for Bayesian Model Assessment. **Statistica Sinica**, v.11, p.419-443, 2001.

IBRAHIM, J. G.; CHEN, M-H.; SINHA, D. Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials. **Statistica Sinica,** v. 14, p. 863-883, 2004.

IBRAHIM, J. G.; LAUD, P. W. A Predictive Approach to the Analysis of Designed Experiments. **Journal of the American Statistical Association**, v.89, n.425, p.309-319, 1994.

KONISHI, S.; KITAGAWA, G. **Information criteria and statistical modeling.** New York: Springer, 2008. 321p.

LAUD, P. W.; IBRAHIM, J. G. Predictive Model Selection. **Journal of the Royal Statistical Society Series B**, v.57, n.1, p. 247-262, 1995.

MARTINI, A. S.. SPEZZAFERRI, F. A predictive model selection criterion. **Journal of the Royal Statistical Society Series B**, v. 46, n. 2, p. 296-303, 1984.

MITCHELL, T. J.; BEAUCHAMP, J. J. Bayesian variable selection in linear regression.**Journal of the American Statistical Association**, v. 83, n. 404, p. 1023-1032, 1988.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the theory of statistics.** 3. ed. New York: J. Wiley, 1974. 564 p.

PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B. **Estatística Bayesiana.** Lisboa: Fundação Calouste Gulbenkian, 2003. 446 p.

PRESS, S. J. **Subjective and objective Bayesian statistics:** principals, models and applications. 2. ed. New Jersey: Wiley-Interscience, 2003. 591 p.

R DEVELOPMENT CORE TEAM. **R:** a language and environment for statistical computing. Vienna: R Foundation fos Statistical Computing, 2009.

SCHWARZ, G. Estimating the dimensional of a model. **Annals of Statistics**, Hayward, v.6, n.2, p.461-464, Mar. 1978.

ANEXOS

Tabela 1A Dados do experimento realizado para avaliar a produção de um processo químico em função de sua concentração e temperatura.

Produção (*l*)	**Temperatura (^{o}C)**	**Concentração (%)**
189	80	10
203	100	10
222	120	10
234	140	10
261	160	10
204	80	15
212	100	15
223	120	15
246	140	15
273	160	15
220	80	20
228	100	20
252	120	20
263	140	20
291	160	20
226	80	25
232	100	25
259	120	25
268	140	25
294	160	25

Tabela 2A Dados das horas trabalhadas no departamento de contabilidade de uma empresa.

Nº de horas trabalhadas	**Nº de cheques descontados**	**Nº de pagamentos recebidos**	**Nº de documentos processados**	**Nº de ordens de pagamento**
130.7	654	683	183	123
113.0	457	479	89	49
125.4	429	823	196	115
131.1	483	735	157	82
133.2	915	1018	211	116
178.2	813	857	218	169
121.1	616	924	312	105
135.5	936	1247	428	82
109.8	550	965	461	94
119.0	448	688	244	101
103.8	505	561	261	121
114.2	501	735	154	103
118.4	712	943	162	83
104.6	642	758	252	64
134.0	491	809	149	82
140.2	590	638	198	99
110.9	517	671	116	48
101.2	455	516	139	112
122.9	723	835	300	89
97.5	416	578	112	238