Ponderação de Modelos com Aplicação em Regressão Logística Binária

Juliane Bertini Brocco

Orientador: Prof^a. Dr^a. Cecília Candolo

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

São Carlos

Abril de 2006

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

B863pm

Brocco, Juliane Bertini.
Ponderação de modelos com aplicação em regressão logística binária / Juliane Bertini Brocco. -- São Carlos : UFSCar, 2006.
78 p.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2006.

1. Análise de regressão. 2. Ponderação de modelos. 3. Regressão logística. I. Título.

CDD: 519.536 (20ª)

Agradeço,

aos meus pais e ao meu irmão pelo apoio, força e incentivo que sempre me deram na vida e, principalmente, para a realização desta tese. Muito obrigada do fundo do meu coração, tenho certeza que sem vocês não teria conseguido.

ao meu marido pela compreensão e paciência.

à CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) pela assistência financeira.

à professora Dra. Cecília Candolo pela orientação e amizade.

e a todos os proferssores do departamento que contribuíram para minha formação e realização desta tese.

Dedico,

ao meu filho João Vitor.

Resumo

Esta dissertação considera o problema de incorporação da incerteza devido à escolha do modelo na inferência estatística, segundo a abordagem de ponderação de modelos, com aplicação em regressão logística. Será utilizada a abordagem de Buckland *et. al.* (1997), que propuseram um estimador ponderado para um parâmetro comum a todos os modelos em estudo, sendo que, os pesos desta ponderação são obtidos a partir do uso de critérios de informação ou do método *bootstrap*. Também será aplicada a ponderação bayesiana de modelos como apresentada por Hoeting *et. al.* (1999), onde a distribuição a posteriori do parâmetro de interesse é uma média da distribuição a posteriori do parâmetro sob cada modelo em consideração ponderado por suas respectivas probabilidades a posteriori.

O objetivo deste trabalho é estudar o comportamento do estimador ponderado, tanto na abordagem clássica como na bayesiana, em situações que consideram o uso de regressão logística binária, com enfoque na estimação da predição. O método de seleção de modelos *Stepwise* será considerado como forma de comparação da capacidade preditiva em relação ao método de ponderação de modelos.

Palavras-chave:Regressão Logística, Ponderação de Modelos.

Abstract

This work consider the problem of how to incorporate model selection uncertainty into statistical inference, through model averaging, applied to logistic regression. It will be used the approach of Buckland *et. al.* (1997), that proposed an weighed estimator to a parameter common to all models in study, where the weights are obtained by information criteria or bootstrap method. Also will be applied bayesian model averaging as shown by Hoeting *et. al.* (1999), where posterior probability is an average of the posterior distributions under each of the models considered, weighted by their posterior model probability.

The aim of this work is to study the behavior of the weighed estimator, both, in the classic approach and in the bayesian, in situations that consider the use of binary logistic regression, with foccus in prediction. The known model-choice selection method *Stepwise* will be considered as form of comparison of the predictive performance in relation to model averaging.

keywords: Logistic Regression, Model Averaging.

Sumário

Capítulo 1

Introdução

Uma abordagem típica de análise estatística consiste em vários estágios: exploração descritiva do conjunto de dados, definição da classe de modelos a ser considerada, seleção do melhor modelo dentro desta classe de acordo com algum critério pré-estabelecido e obtenção de inferências baseadas no modelo selecionado. Este ciclo é geralmente iterativo e envolve, além da aplicação dos conceitos e técnicas estatísticas, considerações subjetivas. A conclusão obtida no final deste processo depende do(s) modelo(s) escolhido(s). Quando a inferência é feita sem levar em consideração a incerteza devido à escolha do(s) modelo(s), pode acontecer uma subestimação da variabilidade de quantidades de interesse e/ou inferências super-otimistas ou viciadas (Buckland *et. al.*, 1997)

Nesta dissertação será considerado o problema de incorporação da incerteza devido à escolha do modelo na inferência estatística, segundo a abordagem de ponderação de modelos, com aplicação em regressão logística. Será utilizada a abordagem de Buckland *et. al.* (1997), que propuseram um estimador ponderado para um parâmetro comum a todos os modelos em estudo, sendo que, os pesos desta ponderação são obtidos a partir do uso de critérios de informação ou do método *bootstrap*. Também será aplicada a ponderação bayesiana de modelos como apresentada por Hoeting *et. al.* (1999), onde a distribuição a posteriori do parâmetro de interesse é uma média da distribuição a posteriori do parâmetro sob cada modelo em consideração ponderado por suas respectivas probabilidades a posteriori. A incorporação desta incerteza na inferência tem despertado o interesse de alguns pesquisadores e começou a ser tratada de forma sistemática recentemente. Candolo (2001) e Candolo *et. al.* (2003) desenvolveram o estimador ponderado

proposto por Buckland *et. al.* (1997), aprofundando seu estudo para modelos de regressão linear.

O objetivo deste trabalho é estudar o comportamento do estimador ponderado, tanto na abordagem clássica como na bayesiana, em situações que consideram o uso de regressão logística binária, com enfoque na estimação da predição. O método de seleção de modelos *Stepwise* será considerado como forma de comparação da capacidade preditiva em relação ao método de ponderação de modelos. O motivo da escolha do modelo de regressão logística binária é devido à sua vasta aplicação, como por exemplo, no ramo financeiro para determinação de concessão de crédito e no ramo biológico. O desenvolvimento da abordagem bayesiana foi obtido a partir do desenvolvimento de trabalhos similares na área de análise de sobrevivência e pesquisa social, ver Volinsky *et. al.* (1997) e Raftery (1995). A maior dificuldade encontrada na aplicação dessas abordagens é o esforço computacional requerido. Conjuntos de dados com muitas covariáveis fazem com que o número total de modelos a serem ajustados seja muito grande implicando, muitas vezes, na inviabilização do método apesar do ganho na capacidade preditiva.

Esta dissertação esta estruturada da seguinte maneira: no Capítulo 2 será apresentada a ponderação de modelos, sua aplicação em regressão linear e o uso do método *bootstrap* aplicado à metodologia de ponderação de modelos; no Capítulo 3 será apresentada a ponderação bayesiana de modelos e sua aplicação em regressão linear; no Capítulo 4 será apresentada toda a metodologia envolvendo regressão logística, incluindo a forma de estimação, qualidade do ajuste e formas de predição; no Capítulo 5 está a metodologia de ponderação de modelos aplicada à regressão logística e no Capítulo 6 serão apresentadas aplicações da metodologia de ponderação de modelos com o objetivo de estudar suas propriedades e de comparar a capacidade preditiva desta metodologia em relação ao método de seleção de modelos *Stepwise*. Essas aplicações englobam exemplos da literatura, estudos de simulação e um conjunto de dados reais.

Capítulo 2

Ponderação de Modelos

No contexto frequentista, Buckland *et. al.* (1997) desenvolveram uma metodologia de fácil aplicação, indicando o uso de critérios de informação e do método *bootstrap* na construção de pesos para ponderar modelos.

Esta abordagem de ponderação de modelos assume uma situação onde são considerados K modelos, $M_1, ...M_K$, com o objetivo de estimar um parâmetro de interesse θ. Cada modelo ajustado fornece um estimador deste parâmetro $\theta, \widehat{\theta}_k$, e um peso w_k, construídos de forma que $\sum_{k=1}^{K} w_k = 1$.

Desta forma, o estimador para o parâmetro θ, ponderado pelos pesos, será dado por

$$\widehat{\theta} = \sum_{k=1}^{K} w_k \widehat{\theta}_k. \tag{2.1}$$

Os pesos w_k para cada um dos K modelos, são obtidos via critérios de informação que têm uma forma geral dada por

$$I_k = -2\log(L_k) + q_k, \tag{2.2}$$

onde L_k é a função de verossimilhança maximizada para o modelo k e q_k é uma penalidade, função do número de parâmetros do modelo k ou do número de observações. Podem ser citadas duas opções para este critério, uma proposta por Akaike (1973), conhecida como AIC, onde $q = 2p$, sendo p o número de parâmetros do modelo em estudo. A outra, devida a Schwarz (1978), conhecida como BIC, considera $q = p.log(n)$, onde $n =$ número

de observações. Os pesos podem ainda ser obtidos através do uso do método *bootstrap*, no qual w_k é estimado pela proporção de amostras *bootstrap* nas quais M_k é identificado como o melhor modelo.

Usando critério de informação, os pesos são calculados da seguinte forma

$$w_k = \frac{\exp(-I_k/2)}{\sum_{l=1}^{K} \exp(-I_l/2)}, \ k = 1, ..., K. \tag{2.3}$$

Isso se deve ao fato de que quando dois modelos, k e l, são comparados usando critério de informação obtêm-se

$$\frac{L_k \exp(-q_k/2)}{L_l \exp(-q_l/2)} = \frac{\exp(-I_k/2)}{\exp(-I_l/2)}. \tag{2.4}$$

Sendo assim, a equação (2.3) é uma escolha plausível para determinação dos pesos w_k, pois, desta forma, garante-se que dois modelos com os mesmos valores de critério de informação receberão o mesmo peso, independentemente da penalidade definida para cada um deles.

Para o cálculo da variância do estimador ponderado $\widehat{\theta}$, Buckland *et. al.* (1997) consideram inicialmente um caso irreal onde os $\widehat{\theta}_k$ são identicamente distribuídos com média θ e os pesos w_k são constantes conhecidas. Sob estas condições obtêm-se

$$Var(\widehat{\theta}) = \sum_k w_k^2 var(\widehat{\theta}_k) + \sum_k \sum_{l \neq k} w_k w_l cov(\widehat{\theta}_k, \widehat{\theta}_l). \tag{2.5}$$

O problema encontrado no cálculo desta variância está em estimar a covariância entre $\widehat{\theta}_k$ e $\widehat{\theta}_l$. Sabe-se que a covariância será alta devido ao fato de cada modelo ser ajustado ao mesmo conjunto de dados. Sendo assim, fixa-se a covariância como sendo o maior valor possível, isto é, a média geométrica das variâncias estimadas considerando os modelos k e l. Desta forma, obtêm-se um limite superior para a $var(\widehat{\theta})$ dado por

$$var(\widehat{\theta}) \leq \left\{ \sum_k w_k \sqrt{var(\widehat{\theta}_k)} \right\}^2. \tag{2.6}$$

Porém, esta variância não incorpora o vício de má especificação do modelo. Suponha então que define-se $\theta_k = \theta + \beta_k$, onde β_k é o vício de má especificação que surge na

estimativa de θ sob o modelo k. Suponha também que $E(\beta_k) = 0$, quando todos os possíveis modelos estão sendo considerados. Desta forma obtêm-se

$$E(\widehat{\theta}_k/\beta_k) = \theta + \beta_k = \theta_k. \tag{2.7}$$

Se a média for calculada considerando-se todos os possíveis modelos, obtêm-se $E(\widehat{\theta}_k) = \theta$. Assumindo

$$Var(\widehat{\theta}_k/\beta_k) = E\left[(\widehat{\theta}_k - \theta_k)^2\right] \tag{2.8}$$

e

$$Var(\widehat{\theta}_k) = E\left[(\widehat{\theta}_k - \theta)^2\right], \tag{2.9}$$

então

$$Var(\widehat{\theta}_k) = Var(\widehat{\theta}_k/\beta_k) + \beta_k^2. \tag{2.10}$$

Desta forma, obtêm-se

$$Var(\widehat{\theta}) = \sum_k w_k^2 var(\widehat{\theta}_k) + \sum_k \sum_{l \neq k} w_k w_l cov(\widehat{\theta}_k, \widehat{\theta}_l), \tag{2.11}$$

para a qual, assumindo correlação perfeita, obtêm-se

$$Var(\widehat{\theta}) = \left\{ \sum_k w_k \sqrt{var(\widehat{\theta}_k/\beta_k) + \beta_k^2} \right\}^2. \tag{2.12}$$

Esta variância pode ser estimada substituindo-se $\widehat{\beta}_k = \widehat{\theta}_k - \widehat{\theta}$ e $\widehat{Var}(\widehat{\theta}_k/\beta_k)$. As estimativas $\widehat{\theta}_k$ e $\widehat{Var}(\widehat{\theta}_k/\beta_k)$ são obtidas através de métodos inferenciais usuais, assumindo-se que o modelo k é o modelo verdadeiro e $\widehat{\theta}$ é dado pela equação (2.1).

Outra abordagem para se obter a $Var(\widehat{\theta})$ é através do uso do método *bootstrap,* onde o estimador $\widehat{\theta}$ pode ser calculado em cada reamostra para estimar a variância. Este assunto será abordado com mais detalhes na Seção 2.2.

É importante ressaltar que a ponderação de modelos apenas faz sentido se as quantidades que estão sendo ponderadas têm a mesma interpretação para todos os modelos em consideração. Assim, ponderar valores de parâmetros ou estimativas relativas a um modelo particular não será pertinente no contexto de ponderação de modelos.

2.1 Ponderação de Modelos em Regressão Linear

Considere o modelo de regressão linear

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon = \mathbf{X}\beta + \varepsilon, \tag{2.13}$$

onde Y é um vetor de dimensão n x 1 de observações da variável resposta, X é uma matriz de dimensão n x $(p+1)$ de covariáveis observadas, $X_1, ..., X_p$ são vetores de dimensão 1 x $(p+1)$ desta matriz, β é o vetor de dimensão $(p+1)$ x 1 de parâmetros desconhecidos e ε é o vetor de erros de dimensão n x 1. Assume-se que os erros são independentes com distribuição $N(0, \sigma^2)$ e β e σ^2 são desconhecidos.

Muitas vezes o número de covariáveis presentes num modelo é muito grande e além disso, muitas delas podem não ser estatisticamente significantes. Uma alternativa a esta situação é utilizar um método de seleção de modelos na busca de um modelo "ótimo". Os métodos de seleção de modelos mais conhecidos são o *Stepwise*, o *Backward* e o *Forward*. Neste trabalho, será considerado o método *Stepwise*.

O método de seleção de modelos consiste em identificar as covariáveis mais significativas na previsão da variável resposta. Esta escolha é feita adicionando e removendo-se variáveis com base em um teste F, de forma a se obter um modelo "ótimo", isto é, o modelo que melhor prediz a variável resposta dentre todos os possíveis modelos. Mais detalhes sobre os métodos de seleção de modelos podem ser encontrados em Neter, Kutner, Nachtsheim & Wasserman (1996).

Considerando o modelo (2.13), com p potenciais variáveis explicativas, o número de possíveis modelos a serem considerados, K, é $K = 2^p$. Vale observar, porém, que muitos desses modelos tem pouco suporte dos dados. Neste caso, a ponderação poderia ser feita considerando apenas os melhores modelos como uma aproximação da ponderação sob todos os 2^p possíveis modelos.

Suponha uma situação onde têm-se apenas duas potenciais variáveis explicativas.

Nesta situação, os $K = (2^2) = 4$ possíveis modelos a serem considerados são

$$\begin{aligned} Modelo(1) &: y_i = \beta_0 + \varepsilon_i \\ Modelo(2) &: y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i \\ Modelo(3) &: y_i = \beta_0 + \beta_2 x_{2i} + \varepsilon_i \\ Modelo(4) &: y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i. \end{aligned} \tag{2.14}$$

O objetivo é predizer a média $\theta = \beta_0 + \beta_1 x_{1+} + \beta_2 x_{2+}$ de uma observação futura Y_+ para valores x_{1+} e x_{2+}. Os possíveis estimadores desta média são

$$\begin{aligned} Modelo(1) &: \widehat{\theta}_1 = \widehat{\beta}_0 \\ Modelo(2) &: \widehat{\theta}_2 = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1+} \\ Modelo(3) &: \widehat{\theta}_3 = \widehat{\beta}_0 + \widehat{\beta}_2 x_{2+} \\ Modelo(4) &: \widehat{\theta}_4 = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1+} + \widehat{\beta}_2 x_{2+}. \end{aligned}$$

Desta forma, o estimador para o parâmetro θ, ponderado pelos pesos, será dado por

$$\widehat{\theta} = w_1\widehat{\theta}_1 + w_2\widehat{\theta}_2 + w_3\widehat{\theta}_3 + w_4\widehat{\theta}_4. \tag{2.15}$$

Considerando o uso do AIC, os pesos w_k para cada um dos K modelos são calculados da seguinte forma

$$w_k = \frac{\exp(-AIC_k/2)}{\sum_{l=1}^{k} \exp(-AIC_l/2)},\ k = 1, ..., 4, \tag{2.16}$$

onde AIC_k é o critério de informação de Akaike para o modelo k. Quando a variância é conhecida, o logaritmo da função de verossimilhança maximizada para o modelo k é

$$\log L_k = const - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2} SQR_k, \tag{2.17}$$

onde SQR_k é a soma de quadrados dos resíduos da regressão para o modelo k. Já quando

a variância é desconhecida, o logaritmo da função de verossimilhança maximizada para o modelo k é

$$\log L_k = const - \frac{n}{2}\log(SQR_k/n). \tag{2.18}$$

A variância de $\widehat{\theta}$ pode ser obtida como em (2.12), e no caso de regressão linear, a $\widehat{Var}(\widehat{\theta}_k/\beta_k)$ é dada por

$$\widehat{Var}(\widehat{\theta}_k/\beta_k) = \widehat{\sigma}^2\left(1 + \underline{x}_+ \left(X'X\right)^{-1}\underline{x}'_+\right),$$

onde $\underline{x}_+ = (1, x_{1+}, x_{2+})$.

2.2 O Método *Bootstrap* e seu uso em Regressão Linear

O método *bootstrap* introduzido por Efron (1979), tem a vantagem de evitar desenvolvimentos analíticos e tem sido uma das técnicas mais utilizadas e expandidas, com aplicações nas mais diversas áreas. Referências básicas são os livros de Efron e Tibshirani (1993) e Davison e Hinkley (1997) que abordam amplamente a metodologia *bootstrap* e aplicações em diversas áreas. O texto desta Seção é baseado em Candolo (2001) visando a abordagem do método *bootstrap* de forma específica para a ponderação de modelos.

O método *bootstrap* é definido como segue: Seja $y_1, ..., y_n$ uma amostra aleatória de observações identicamente distribuídas. Os valores amostrais são os resultados obtidos das variáveis aleatórias independentes e identicamente distribuídas $Y_1, ..., Y_n$ com distribuição de probabilidade desconhecida, F, que depende de um parâmetro desconhecido θ, dado por $\theta = s(F)$, sendo $s(\cdot)$ a função que define θ. O parâmetro θ é estimado por $\widehat{\theta} = s(\widehat{F})$, onde $\widehat{F}$ é um estimador de F obtido a partir da amostra $y_1, ..., y_n$. Seja $\widehat{F}$ a distribuição empírica de $y_1, ..., y_n$, que atribui probabilidades iguais a $1/n$ para cada valor da amostra, isto é,

$$\widehat{F}(y) = \frac{\#\{y_j \leq y\}}{n}. \tag{2.19}$$

onde $\#\{y_j \leq y\}$ indica o número de vezes que $y_j \leq y$ ocorre. O método *bootstrap* tem como princípio substituir a distribuição desconhecida F por $\widehat{F}$ para estimar θ, aproximando a

distribuição de $\theta = s(F)$ pela de $\widehat{\theta^*} = s(\mathbf{y}^*, \widehat{F})$, onde $\mathbf{y}^*$ é uma amostra aleatória de tamanho n, retirada de $\widehat{F}$. A amostra $\mathbf{y}^* = (y_1^*, ..., y_n^*)$ é chamada de amostra *bootstrap* e sua definição é análoga a obter uma amostra aleatória de tamanho n retirada com reposição da população de tamanho n, $(y_1, ..., y_n)$. A partir da amostra *bootstrap* é calculada uma repetição *bootstrap* de $\widehat{\theta}$, $\widehat{\theta^*} = s(\mathbf{y}^*)$. Repetindo este procedimento um número suficientemente grande de vezes, calcula-se uma distribuição empírica de $\widehat{\theta^*}$ e, a partir desta, obtêm-se média, erro padrão, intervalo de confiança, etc. Observa-se que a única suposição feita é a de que a observações $(y_1, ..., y_n)$ são independentes e identicamente distribuídas.

Quando $\widehat{F}$ é definida como em (2.19), o método *bootstrap* é denominado de *bootstrap* não paramétrico, e quando $\widehat{F}$ é definida como uma distribuição de probabilidade específica, com os parâmetros substituídos pelas estimativas obtidas na amostra $(y_1, ..., y_n)$, o método *bootstrap* é denominado *bootstrap* paramétrico.

A aplicação do método *bootstrap* em modelos de regressão tem por objetivo a obtenção de propriedades dos estimadores dos parâmetros da regressão e de predições. O principal ponto a ser abordado neste tipo de aplicação diz respeito à forma de reamostragem em problemas de regressão, pois se a simulação é feita de maneira consistente com o modelo adotado, o resultado assintótico obtido pelo método *bootstrap* será o mesmo que aquele obtido pelos métodos analíticos.

Seja o modelo de regressão como definido em (2.13). O plano de reamostragem para que se obtenha o mesmo delineamento que têm os dados, isto é, $x_i^* \equiv x_i$, especifica que y_i^* tenha distribuição condicional a x_i^*, e é obtida por

$$y^* = X\widehat{\beta} + \varepsilon^*,$$

onde $\widehat{\beta}$ é a estimativa de β. O vetor de erros ε^* é aleatoriamente amostrado de $\widehat{F}_e$, a distribuição empírica dos resíduos $e_i = r_i - \overline{r}$, onde $r_i = y_i - \widehat{y}_i$ e $\overline{r}$ é sua média, para $i = 1, ..., n$. Davison e Hinkley (1997), Cap. 6, indicam que para melhores resultados práticos é melhor usar os resíduos estudentizados

$$r_i = \frac{y_i - \widehat{y}_i}{(1 - h_i)^{1/2}},$$

onde h_i é o i-ésimo elemento da diagonal da matriz $X(X^TX)^{-1}X^T$, pois a variância deste resíduo concorda com a de ε. A partir dos valores de $y_1^*, ..., y_n^*$ obtém-se $\widehat{\beta}^* = X(X^TX)^{-1}X^TY^*$ e a distribuição de $\sqrt{n}(\widehat{\beta}^* - \widehat{\beta})$ se aproxima da distribuição de $\sqrt{n}(\widehat{\beta} - \beta)$.

Uma abordagem diferente de reamostragem ocorre quando os dados são considerados como amostras de uma distribuição bivariada de (X, Y). Neste caso, $\widehat{F}$ é a distribuição empírica dos vetores de observações $(\mathbf{x}_i, \mathbf{y}_i)$, atribuindo probabilidade $1/n$ a cada um deles, para $i = 1, ..., n$, obtendo-se, então, uma amostra *bootstrap* de n pares. Há diferenças importantes entre estes dois métodos de reamostragem: o segundo método não faz suposição quanto a homogeneidade de variância, tendo por um lado mais robustez à heterocedasticidade, mas por outro lado pode ser ineficiente se o modelo de variância constante é correto. Outra diferença é que as amostras são obtidas com diferentes delineamentos, uma vez que os valores de X são obtidos aleatoriamente.

2.2.1 O Método *Bootstrap* em Ponderação de Modelos

O método *bootstrap*, como já dito anteriormente, é uma outra abordagem para se obter a $Var(\widehat{\theta})$, onde o estimador $\widehat{\theta}$ pode ser calculado em cada reamostra. No caso de situações de regressão, a reamostragem é, usualmente, feita a partir dos resíduos da regressão, pois, desta forma, a análise continua condicional aos valores das covariáveis. No caso de ponderação de modelos, o uso da abordagem de reamostragem dos resíduos provoca uma influência muito grande do modelo a partir do qual os resíduos foram calculados. Pode parecer mais adequado, neste caso, amostrar os vetores de observações, mas algumas alternativas têm sido consideradas. Buckland *et al.* (1997) sugerem três alternativas, além da reamostragem dos pares:

a. gerar todas as amostra *bootstrap* a partir do modelo selecionado na análise preliminar dos dados originais;

b. gerar as amostras a partir do modelo completo e

c. selecionar os modelos a partir dos pesos como calculados na equação (2.3) e, então, gerar a próxima amostra a partir do modelo selecionado.

Candolo (2001) apresenta uma outra alternativa considerando a reamostragem dos resíduos. Esta alternativa consiste em gerar K amostras a partir dos resíduos de cada um dos K modelos, $M_1, ..., M_K$, calcular os pesos w_k e $\widehat{\theta}_k$,como definido em (2.3), nas respecti-

vas amostras e, então, calcular $\widehat{\theta^*}$. Desta forma, os K modelos estarão sendo ajustados em amostras diferentes o que caracteriza um estimador diferente daquele definido em (2.1). O algoritmo para esta abordagem pode ser encontrado em Candolo (2001, pag.32).

Abaixo estão colocados os planos de reamostragem de pares e de resíduos.

Planos de Reamostragem

O algoritmo 1 descreve os passos para a reamostragem dos vetores de observações e o algoritmo 2 descreve os passos para a reamostragem dos resíduos, ambos aplicados a modelos de regressão linear como apresentado na Seção 2.1.

Algoritmo 1: Reamostragem dos Pares

(i) amostrar $i_1^*, ..., i_n^*$ aleatoriamente, com reposição, de $\{1, 2, ..., n\}$;

(ii) fazer $(y_j^*, \mathbf{x}_j^*) = (y_{i_j^*}, \mathbf{x}_{i_j^*})$, para $j = 1, ..., n$, obtendo a amostra *bootstrap* $(y_1^*, \mathbf{x}_1^*)$,...,$(y_n^*, \mathbf{x}_n^*)$;

(iii) repetir os passos (i) e (ii), B vezes.

Algoritmo 2: Reamostragem dos Resíduos

(i) amostrar ε_j^* de $r_j - \overline{r}$, para $j = 1, ..., n$, com reposição

(ii) fazer $y_j^* = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1j} + ... + \widehat{\beta}_p X_{pj} + \varepsilon_j^*,\ j = 1, ..., n$

(iii) repetir os passos (i) e (ii), B vezes.

O modelo a ser considerado no item (ii) do algoritmo 2 depende da especificação do modelo em consideração.

Capítulo 3

Ponderação Bayesiana de Modelos

A Ponderação Bayesiana de Modelos, ou BMA (do inglês *Bayesian Model Averaging*) é uma técnica Bayesiana utilizada para a incorporação da incerteza devido à escolha do modelo na inferência estatística.

Seja uma situação onde K modelos, $M_1, ..., M_K$, são considerados, tendo como objetivo estimar uma quantidade de interesse θ e seja D o conjunto de dados para análise. A distribuição a posteriori para este parâmetro de interesse é expressa por

$$P(\theta/D) = \sum_{k=1}^{K} P(\theta/M_k, D)P(M_k/D), \tag{3.1}$$

que é uma média da distribuição a posteriori sob cada um dos K modelos em consideração, ponderada por suas probabilidades a posteriori do modelo respectivo.

A probabilidade a posteriori para o modelo M_k é dada por

$$P(M_k/D) = \frac{P(D/M_k)P(M_k)}{\sum_{l=1}^{K} P(D/M_l)P(M_l)}, \tag{3.2}$$

onde

$$P(D/M_k) = \int P(D/\eta_k, M_k)P(\eta_k/M_k)d\eta_k \tag{3.3}$$

é a integral da verossimilhança do modelo M_k, η_k é o vetor de parâmetros do modelo M_k, $P(\eta_k/M_k)$ é a densidade a priori de η_k sob o modelo M_k, $P(D/\eta_k, M_k)$ é a verossimilhança e $P(M_k)$ é a probabilidade a priori do modelo M_k ser o verdadeiro modelo. Todas as probabilidades são implicitamente condicionais a $\mathcal{M} = \{M_1, ..., M_K\}$, conjunto de todos

os modelos que estão sendo considerados.

Assim, a média e a variância a posteriori de θ, podem ser escritas respectivamente por

$$E[\theta/D] = \sum_{k=1}^{K} E[\theta/D, M_k]\ P(M_k/D) \tag{3.4}$$

$$Var[\theta/D] = \sum_{k=1}^{K} (Var[\theta/D, M_k] + E[\theta/D, M_k]^2)\ P(M_k/D) - E[\theta/D]^2. \tag{3.5}$$

Segundo Hoeting *et. al.* (1999) os problemas encontrados para a implementação da Ponderação Bayesiana de Modelos são, basicamente:

- o número de termos na equação (3.1) pode ser muito grande, causando uma soma exaustiva;
- as integrais implícitas em (3.1) podem ser difíceis de calcular. Esse problema pode ser resolvido utilizando o método de Monte Carlo em Cadeia de Markov (MCMC);
- a especificação de $P(M_k)$, a distribuição a priori sobre os modelos, é importante e tem sido alvo de estudos.

Uma alternativa para resolver o primeiro problema é obter a média para um subconjunto dos modelos mais indicados pelos dados. Madigan & Raftery (1994) propuseram um método chamado *Occam 's Window*, com o qual se obtém a média de um conjunto de modelos parsimoniosos e indicados pelos dados, selecionados a partir da aplicação de técnicas padrões da pesquisa científica. Este assunto será abordado com mais detalhes na Seção 3.1. Outra alternativa é aproximar a soma em (3.1) usando a abordagem de Monte Carlo via Cadeias de Markov, também conhecida como MCMC. Madigan & York (1995) apresentam uma metodologia chamada composição de modelos via MCMC, que foi denominada de MC^3, a qual gera um processo estocástico que se move através do espaço de modelos. Este assunto será abordado com mais detalhes na Seção 3.2.

3.1 O Método *Occam's Window*

O método *occam's window* é baseado em um processo iterativo, no qual os modelos são comparados em relação à sua capacidade preditiva. Sendo assim, modelos com menor

capacidade preditiva serão descartados do conjunto de modelos em consideração.

Serão excluídos da equação (3.1) os modelos que não pertencerem ao conjunto $\mathcal{A}'$ definido a seguir, ou seja, os modelos que têm capacidade preditiva bem inferior ao modelo que tem a maior capacidade preditiva da classe

$$\mathcal{A}' = \left\{ M_k : \frac{\max_l \{P(M_l/D)\}}{P(M_k/D)} \leq C \right\}, \tag{3.6}$$

para alguma constante C. Segundo Madigan & Raftery (1994), o valor de C a ser usado dependerá do contexto. Em seus exemplos, o valor de C utilizado foi 20, em analogia ao usual ponto de corte de $0,05$ dos *p-valores*. Já Jeffreys (1961, app.B) sugere que se use um número entre 10 e 100.

Note que, a medida da capacidade preditiva de um modelo é feita através da $P(M_k/D)$ ao invés da $P(D/M_k)$. Nesse contexto, a verossimilhança é ponderada pela probabilidade a priori do modelo $P(M_k)$, de modo que esta reflita dados passados, resultando em uma probabilidade preditiva composta por dados presentes e passados.

O próximo passo para seleção dos modelos, considera a razão de *Occam's*. Seja E a evidência e $P(H/E)$ a probabilidade de uma hipótese específica H dado a evidência E. A razão de *Occam's* estabelece que se

$$P(H_1/E) = P(H_2/E) = ... = P(H_K/E)$$

para as hipóteses $H_1, ..., H_K$, então a hipótese a ser escolhida deverá ser a mais simples entre $H_1, ..., H_K$. Sendo assim, também serão excluídos da equação (3.1) os modelos pertencentes ao conjunto

$$\mathcal{B} = \left\{ M_k : \exists M_l \in \mathcal{A}', M_l \subset M_k, \frac{P(M_l/D)}{P(M_k/D)} > 1 \right\}, \tag{3.7}$$

onde M_l é um submodelo de M_k.

Assim, a equação (3.1) será substituída por

$$P(\theta/D) = \sum_{M_k \in A} P(\theta/M_k, D) P(M_k/D), \tag{3.8}$$

onde $\mathcal{A}=\mathcal{A}'\backslash\mathcal{B}$.

Esse procedimento reduz consideravelmente o número de modelos na soma em (3.1) e consequentemente simplifica o problema da incorporação da incerteza devido à escolha do modelo.

O questão se reduz então em definir o conjunto $\mathcal{A}$. A técnica proposta por Madigan & Raftery (1994) é uma variação do algoritmo de busca de *greedy*. A probabilidade a posteriori do modelo é usada como medida da busca. A estratégia trabalha dentro do espaço de modelos, comparando os modelos através da razão das probabilidades a posteriori em uma sequência de comparações aninhadas.

A estratégia é baseada em duas idéias principais: na primeira o algoritmo compara dois modelos encaixados e quando o modelo mais simples é rejeitado todos seus submodelos também serão. Na segunda, chamada *occam's window*, o aspecto crucial é a interpretação da razão das probabilidades a posteriori dos modelos $P(M_0/D)/P(M_1/D)$, onde M_0 é o modelo com uma variável preditora a menos do que o modelo M_1. A idéia principal deste princípio é mostrada na Figura 3-1 e pode ser interpretada como:

- se o logaritmo da razão das probabilidades a posteriori é positivo (ou seja, os dados dão mais evidência para o modelo M_0), rejeita-se M_1 e aceita-se M_0. Isso pode ser generalizado assumindo que essa razão seja maior do que uma contante positiva O_R antes de rejeitar o modelo M_1;

- se o logaritmo da razão das probabilidades a posteriori é pequeno e negativo, indicando que a evidência contra M_0 não é forte, considera-se os dois modelos;

- e, se o logaritmo da razão das probabilidades a posteriori é grande, em valor absoluto, e negativo (ou seja, menor do que $O_L = \log(C)$, onde C é definido pela equação (3.6)), rejeita-se M_0 e considera-se M_1.

O algoritmo completo pode ser encontrado em Madigan & Raftery (1994).

3.2 O Método MC^3

O método MC^3, proposto por Madigan & York (1995), usa o método de Monte Carlo via Cadeia de Markov (MCMC) para fazer uma aproximação direta de (3.1) gerando um processo estocástico que se move através do espaço de modelos.

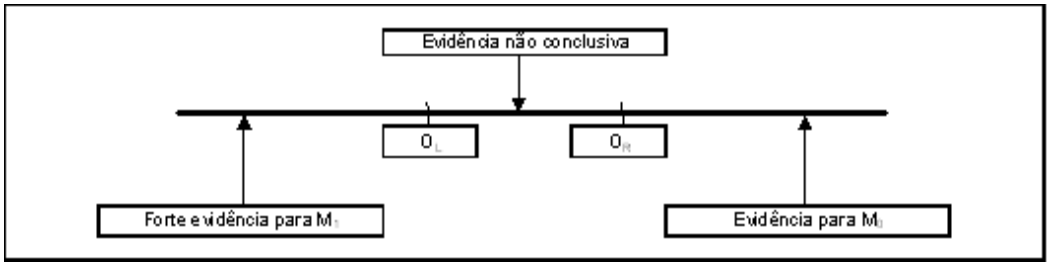

Figura 3-1: *Occam's Window*: interpretação do logaritmo da razão a posteriori.

Especifica-se $\mathcal{M}$ como sendo o espaço de estados dos modelos em consideração. Constrõe-se uma cadeia de Markov $\{M(T), t = 1, 2, ...\}$ com espaço de estados $\mathcal{M}$ e distribuição de equilíbrio $P(M_i/D)$. Simula-se esta cadeia obtendo-se as observações $M(1), ..., M(N)$. Assim, sob certas condições de regularidade, para qualquer função $g(M_i)$ definida em $\mathcal{M}$, a média

$$\widehat{G} = \frac{1}{N}\sum_{t=1}^{N} g(M(t)) \tag{3.9}$$

é a estimativa da $E(g(M))$. Para calcular $P(\theta/D)$ desta forma faz-se $g(M) = P(\theta/M, D)$.

Quanto à dificuldade no cálculo das integrais em $P(D/M_k)$, Hoeting *et al.* (1999) fornecem detalhes gerais de implementação do BMA para algumas classes de modelos, incluindo, além de regressão linear, modelos lineares generalizados, análise de sobrevivência e modelos gráficos. O uso do método de Laplace pode fornecer boas aproximações para $P(D/M_k)$, como pode ser visto em Tierney & Kadane (1986).

3.3 Ponderação Bayesiana de Modelos para Regressão Linear

O desenvolvimento do BMA para regressão linear é apresentado por Hoeting (1994) e por Raftery *et al.* (1997), que fornecem o cálculo apropriado de $P(D \ / \ M_k)$, usando a classe de distribuições a priori conjugadas, normais padrões, e a distribuição a posteriori preditiva, a partir do qual, são obtidas $\mathrm{E}(\theta \ / \ M_k, D)$ e $\mathrm{Var}(\theta \ / \ M_k, D)$.

Cada um dos K modelos em consideração tem a mesma forma descrita em (2.13).

Para regressão linear, utilizou-se uma distribuição a priori que abrangesse o valor do

parâmetro. Foi adotada uma classe de priori gamma-normal conjugada da forma

$$\begin{aligned} \beta &\sim N(\mu, \sigma^2 V), \\ \frac{v\lambda}{\sigma^2} &\sim \chi^2_v, \end{aligned}$$

onde v, λ, a matriz V, $(p+1)$ x $(p+1)$, e o vetor μ, $(p+1)$ x 1, são hiperparâmetros a serem escolhidos.

Para variáveis não categorizadas, assume-se que os $\beta's$ são independentes a priori, centraliza-se β em zero e escolhe-se $\mu = (\widehat{\beta}_0, 0, 0, ..., 0)$, onde $\widehat{\beta}_0$ é o estimador de mínimos quadrados de β_0. A matriz de covariância V é igual a σ^2 multiplicado pela matriz diagonal com os elementos da diagonal dados por $(S_Y^2, \phi^2 S_1^{-2}, \phi^2 S_2^{-2}, ..., \phi^2 S_p^{-2})$, onde S_Y^2 denota a variância amostral de Y, S_i^{-2} denota a variância amostral de X_i para $i = 1, ..., p$ e ϕ é um hiperparâmetro a ser escolhido. Hoeting (1994) fornece completa argumentação para a escolha dos valores destes hiperparâmetros e conclui que estes valores são: $v = 2,58$, $\lambda = 0,28$ e $\phi = 2,85$.

A verossimilhança marginal para Y sobre o modelo M_k baseada nas prioris determinadas acima é dada por

$$\begin{aligned} P(Y/\mu_k, V_k, X_k, M_k) &= \frac{\Gamma(\frac{v+n}{2})(v\lambda)^{v/2}}{\pi^{n/2}\Gamma(\frac{v}{2})|I + X_k V_k X_k^t|^{1/2}} x \\ \text{x} & \quad [\lambda v + (Y - X_k\mu_k)^t (I + X_k V_k X_k^t)^{-1} \\ \text{x} & \quad (Y - X_k\mu_k)]^{-(v+n)/2}, \end{aligned} \tag{3.10}$$

onde X_k é a matriz de delineamento, μ_k o vetor de médias de β e V_k é a matriz de covariância de β correspondente ao modelo M_k. Essa distribuição é uma t-Student não-central de dimensão n com ν graus de liberdade, média Xμ e variância $[\nu(\nu-2)]\lambda(I + XVX^T)$.

Assumindo, a priori, que todos os modelos são equiprováveis, então

$$P(M_k/D) = \frac{P(Y/\mu_k, V_k, X_k, M_k)}{\sum_{l=1}^{K} P(Y/\mu_k, V_l, X_l, M_l)}. \tag{3.11}$$

Seja $\theta = Z\beta + \varepsilon$, com $\varepsilon \sim N(0, \sigma^2 I)$, um valor de predição, onde $\underset{1x(p+1)}{Z}$ é a matriz dos preditores conhecidos e β é o vetor de parâmetros. As distribuições a priori dos parâmetros são

$$\begin{aligned} &\beta \sim N_{p+1}(\mu', \sigma^2 V'), \\ &\frac{\nu'\lambda'}{\sigma^2} \sim \chi^2_{\nu'}, \end{aligned}$$

onde μ', V', ν' e λ' são os parâmetros a posteriori dados por

$$\begin{aligned} \mu' &= (X^T X + V^{-1})^{-1}(X^T Y + V^{-1}\mu) \\ V' &= (X^T X + V^{-1})^{-1} \\ v' &= n + \nu \\ \lambda' &= \frac{1}{n+\nu}\left[\begin{array}{c} \nu\lambda + \mu^T V^{-1}\mu + Y^T Y - \\ \left(X^T Y + V^{-1}\mu\right)^T \left(X^T X + V^{-1}\right)^{-1} \left(X^T Y + V^{-1}\mu\right) \end{array}\right]. \end{aligned}$$

A distribuição a posteriori preditiva é dada por

$$\begin{aligned} f(\theta/Y) = &\frac{\Gamma\left(\frac{\nu+n+1}{2}\right)(\nu\lambda)^{\nu/2}}{\Gamma\left(\frac{1}{2}\right)\Gamma(v+n/2)}(v+n)^{(v+n)/2}\left|\lambda'\left(ZV'Z^T+1\right)\right|^{-1/2} \\ &\left\{(v+n) + \frac{(\theta - Z\mu')^2}{\lambda'(ZV'Z^T+1)}\right\}^{-(v+n+1)/2}, \end{aligned} \tag{3.12}$$

que é uma distribuição t-Student com $(n+v)$ graus de liberdade, média $Z\mu'$ e variância $[v + n/(v+n-2)]\lambda'(ZV'Z^T)$, onde n é o número de dados observados.

Desta forma obtem-se:

$$E(\theta/M_k, D) = Z\mu'_k. \tag{3.13}$$

$$Var(\theta/M_k, D) = \frac{n+\nu}{n+\nu-2}\lambda'_k\left(ZV'_kZ^T\right). \tag{3.14}$$

Finalmente, a média e variância a posteriori de θ são dadas por

$$E(\theta/D) = \sum_{l=1}^{K} E(\theta/M_l, D)\, P(M_l/D). \tag{3.15}$$

$$Var\left(\theta/D\right) = \sum_{l=1}^{K} Var\left(\theta/M_l, D\right) P\left(M_l/D\right) + \sum_{l=1}^{K} \left[E\left(\theta/M_l, D\right) - E\left(\theta/D\right)^2\right] P\left(M_l/D\right). \tag{3.16}$$

Hoeting *et. al.* (1999) fornecem indicação de como obter programas em S-PLUS para o cálculo das probabilidades a posteriori dos modelos e atualmente os mesmos já encontram-se implementados no software R.

Capítulo 4

Regressão Logística

Em muitas aplicações de regressão, a variável resposta é do tipo binária, onde a resposta medida em cada unidade é um "sucesso" ou um "fracasso". Para esse tipo de aplicação o modelo de regressão logística é geralmente o mais utilizado.

Considere que a variável resposta do tipo binária Y esteja sendo modelada como função de uma covariável x. A variável resposta Y é representada por ensaios de Bernoulli com probabilidades de sucesso π e fracasso $1-\pi$ e com $E(Y)=\pi$ e $Var(Y)=\pi(1-\pi)$. O modelo de regressão linear correspondente seria $y=\beta_0+\beta_1 x+\varepsilon$ onde $\varepsilon \sim N(0,\sigma^2)$ e a função que representa a relação entre a variável resposta y e a covariável x é dada por

$$E(Y)=\pi=\beta_0+\beta_1 x \tag{4.1}$$

Este modelo necessita de algumas suposições para que seja válido, entretanto no caso de respostas binárias estas suposições não são satisfeitas. São elas:

- os erros não tem distribuição normal e variância constante;
- nada garante que o campo de variação de $\beta_0+\beta_1 x$ esteja entre 0 e 1 e
- a relaçao entre π e $\beta_0+\beta_1 x$ não é linear.

No caso de regressão logística, a relação entre a variável resposta y e a covariável x, no modelo de regressão logística, é descrita por uma curva sigmoidal, que tem uma forma curvilínea lembrando um S. A linearização desta relação é feita através das funções de ligação. As mais utilizadas são:

- **Transformação Logística**: a transformação logística para a probabilidade de

sucesso π, denotada por $logito(\pi)$, é dada pela transformação $log\{\pi/(1-\pi)\}$, que é o logaritmo da razão de sucesso. Assim, os valores de π no intervalo $(0,1)$ correspondem aos valores do $logito(\pi)$ no intervalo (-∞, ∞).

- Transformação Probito: o probito de uma probabilidade π é definido como sendo os valores de ξ para os quais

$$\frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^{\xi} \exp\left(-\frac{1}{2}u^2\right) du = \pi.$$

Essa integral é a função distribuição de uma variável aleatória normal padrão, U, e então $\pi = P(U \leqslant \xi)$. A função distribuição normal padrão é denotada por $\Phi(\xi)$, e ξ é tal que $\Phi(\xi) = \pi$. Rearranjando, $\xi = \Phi^{-1}(\pi)$, onde a função $\Phi^{-1}(\pi)$ inversa é a transformação probito de π, denotada por $probito(\pi)$.

- Transformação Completmento Log-Log: a transformação complemento log-log da probabilidade π é $log[-log(1-\pi)]$, que também transforma os valores no intervalo $(0,1)$ para valores no intervalo (-∞, ∞).

A transformação complemento log-log é limitada a situações onde a probabilidade de sucesso é assimétrica. Já as transformações logística e probito são bem parecidas, porém a transformação logística é mais conveniente do ponto de vista computacional. Neste trabalho será considerado o uso da transformação logística devido ao fato desta ser mais utilizada e adequada nos casos aqui considerados.

A formulação do modelo de regressão logística é dada da seguinte forma: considere uma amostra aleatória dos pares de observações (y_1,$\underline{x}_1$),...,(y_n,$\underline{x}_n$), onde cada observação y_i corresponde ao resultado de um ensaio de Bernoulli com probabilidade de sucesso π_i e de fracasso $1-\pi_i$ e $\underline{x}_1, ..., \underline{x}_n$ correspondem aos vetores 1 x p de covariáveis. Os momentos são $E(Y) = \pi_i$ e $Var(Y) = \pi_i(1-\pi_i)$.

A função distribuição de probabilidade de y_i é dada por

$$f_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, \tag{4.2}$$

para $y_i = 0, 1$ e $i = 1, ..., n$.

Aplicando-se a transformação logística à equação (4.2) obtêm-se

$$logit(\pi_i) = \log(\pi_i/(1-\pi_i)) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}, \tag{4.3}$$

para $i = 1, ..., n$. O modelo logístico é obtido pela transformação inversa e é dado por

$$E(Y) = \pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}. \tag{4.4}$$

A forma de estimação dos parâmetros deste modelo será apresentada na Seção seguinte.

4.1 Estimação em Regressão Logística

Os procedimentos de estimação e inferência a serem utilizados em regressão logística são um caso particular da metodologia de modelos lineares generalizados apresentado em detalhes no Apêndice A.

A função de verossimilhança para o modelo logístico é dada por

$$L(\beta_0, \beta_1, ..., \beta_p) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}. \tag{4.5}$$

Como a verossimilhança depende da probabilidade de sucesso desconhecida π_i, que por sua vez depende dos β's, a função de verossimilhança pode ser vista como função de β. O problema agora é obter os valores de $\widehat{\beta}_0$, $\widehat{\beta}_1, ...,$ e $\widehat{\beta}_p$ que maximizam $L(\beta_0, \beta_1, ..., \beta_p)$, ou equivalemtemente os valores que maximizam o $log(L(\beta_0, \beta_1, ..., \beta_p))$.

O logarítmo da função de verossimilhança é dado por

$$\begin{aligned} log(L(\beta_0, \beta_1, ..., \beta_p)) &= \sum_i \{y_i \log \pi_i + (1-y_i)\log(1-\pi_i)\} && (4.6) \\ &= \sum_i \left\{ y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i) \right\} && (4.7) \\ &= \sum_i yi(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}) - \log[1 + \exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})] && (4.8) \end{aligned}$$

As derivadas do logaritmo da função de verossimilhança com relação aos parâmetros

desconhecidos β são

$$
\begin{aligned}
U_1 &= \frac{\partial log(L(\beta_0,\beta_1,...,\beta_p))}{\partial\beta_0} = \sum_i \left\{ yi - \left[\frac{\exp(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip})}{1+\exp(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip})} \right] \right\} = \sum_i (yi-\pi_i) \\
U_2 &= \frac{\partial log(L(\beta_0,\beta_1,...,\beta_p))}{\partial\beta_1} = \sum_i \left\{ yix_i - \left[\frac{\exp(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip})}{1+\exp(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip})} \right] \right\} = \\
&= \sum_i x_i(yi-\pi_i) \\
&\vdots \\
U_p &= \frac{\partial log(L(\beta_0,\beta_1,...,\beta_p))}{\partial\beta_p} = \sum_i \left\{ yix_i - \left[\frac{\exp(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip})}{1+\exp(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip})} \right] \right\} = \\
&= \sum_i x_i(yi-\pi_i)
\end{aligned}
$$

Para a solução do sistema de equações acima, utiliza-se o método Score, descrito no Apêndice A, e chamado de método iterativo de mínimos quadrados ponderados.

4.2 Qualidade do Ajuste

Depois de se ajustar um modelo a um conjunto de dados, é natural questionar qual a diferença entre os valores ajustados da variável resposta sob o modelo e os valores observados. Se a diferença entre as observações e os correspondentes valores ajustados é pequena, então o modelo é aceito. Caso contrário, a forma corrente do modelo não será aceita e este precisará ser revisado. Esse aspecto de adequabilidade do modelo será referenciado como qualidade do ajuste. Este texto foi escrito baseado em Collett (1991).

Uma maneira de se medir a discrepância entre a probabilidade de sucesso observada, π_i, e as probabilidade ajustadas, $\widehat{\pi}_i$, pelo modelo assumido é através da função de verossimilhança, pois esta resume a informação que os dados dão sobre um parâmetro desconhecido em um dado modelo. A estatística mais utilizada para verificar esta discrepância, considerando a função de verossimilhança, é a *deviance*, definida como

$$D = -2\log(\widehat{L}_c/\widehat{L}_s) = -2[\log\widehat{L}_c - \log\widehat{L}_s], \tag{4.9}$$

onde $\widehat{L}_c$ é o máximo da verossimilhança sob o modelo corrente e $\widehat{L}_s$ o máximo da verossimilhança do modelo saturado (neste modelo, os valores ajustados coincidem com as observações, ou seja, o modelo ajusta os dados perfeitamente).

Grandes valores de D são encontrados quando $\widehat{L}_c$ é relativamente menor que $\widehat{L}_s$, indicando que o modelo atual é ruim. Por outro lado, pequenos valores de D são obtidos quando $\widehat{L}_c$ é próximo de $\widehat{L}_s$, indicando que o modelo atual é bom.

A estatística *deviance* tem distribuição assintoticamente χ^2 com $(n-p)$ graus de liberdade, onde n representa o número de observações e p o número de parâmetros do modelo corrente.

No caso especial de dados binários onde $n_i = 1,\ i = 1, ..., n$, a *deviance* depende apenas das probabilidades de sucesso ajustadas π_i, e então é não informativa sobre a qualidade do ajuste do modelo.

A verossimilhança para n observações binárias, como função dos parâmetros β, é

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}, \tag{4.10}$$

onde $\pi_i = \frac{\exp(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip})}{1+\exp(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip})}$.

Assim, o logaritmo da função de verossimilhança maximizado considerando o modelo corrente é dado por

$$\log\ \widehat{L}_c = \sum_i \left\{ y_i \log \widehat{\pi}_i + (1-y_i) \log(1-\widehat{\pi}_i) \right\}. \tag{4.11}$$

Para o modelo completo, $\widehat{\pi}_i = y_i$, e como $[y_i \log y_i]$ e $[(1-y_i)\log(1-y_i)]$ valem zero para os únicos dois possíveis valores de $y_i,\ 0$ e $1,\ \log \widehat{L}_S = 0$. Assim a *deviance* para dados binários fica

$$\begin{aligned} D &= -2 \sum \left\{ y_i \log \widehat{\pi}_i + (1-y_i) \log(1-\widehat{\pi}_i) \right\} \\ &= -2 \sum \left\{ y_i \log(\widehat{\pi}_i/(1-\widehat{\pi}_i)) + \log(1-\widehat{\pi}_i) \right\}. \end{aligned} \tag{4.12}$$

Podendo ser reescrita como

$$D = -2 \sum \{ \widehat{\pi}_i logit(\widehat{\pi}_i) + logit(1-\widehat{\pi}_i) \}. \tag{4.13}$$

Assim, a *deviance* depende das observações binárias y_i apenas através das probabilidades ajustadas $\widehat{\pi}_i$, e, portanto, não informando a respeito da discrepância entre as proba-

bilidades observadas e suas correpondentes probabilidades ajustadas. Consequentemente, a *deviance* para modelos ajustados para respostas binárias não deve ser usada como uma medida de qualidade de ajuste do modelo. Então, no caso de respostas binárias, usa-se apenas a diferença de *deviances* para comparar modelos. Esta diferença, é utilizada, por exemplo, pelo método de seleção de modelos na escolha do melhor modelo.

4.3 Predição em Regressão Logística

Um objetivo na modelagem logística é predizer o valor de uma variável resposta binária. A probabilidade da resposta predita pode consequentemente formar a base para se classificar um indivíduo de acordo com um dos dois grupos (0 ou 1). Para se fazer esta classificação, o conjunto de dados deve ser dividido em dois subconjuntos: um conjunto de construção (D^C) e um conjunto de teste (D^T). O primeiro conjunto (D^C) é usado para o ajuste dos modelos e o segundo conjunto (D^T) é usado para se predizer a probabilidade da resposta para um novo indivíduo, a qual será depois comparada com o valor observado.

Para se classificar um indivíduo em dois grupos, com base na probabilidade da resposta predita, um valor limiar π_c tem que ser identificado. Este valor é tal que o indivíduo será classificado no grupo 1 se $\pi_0 < \pi_c$ e no grupo 2 se $\pi_0 \geqslant \pi_c$, onde π_0 é um valor de predição obtido pelo ajuste do modelo. Geralmente $\pi_c = 0,5$ é um valor razoável, entretanto, se os dois grupos não podem ser classificados como simétricos, um valor diferente de 0,5 deve ser considerado. Uma maneira de se determinar este valor limiar π_c, também conhecido como ponto de corte, é através da curva **ROC** (Receiver Operating Characteristics), a qual permite avaliar a capacidade preditiva de um modelo usando o ponto de corte escolhido. Este texto foi escrito baseado em Abreu (2004).

Duas medidas bastante utilizadas para se avaliar a capacidade preditiva de um modelo após a classificação das observações em um dos dois grupos, 0 ou 1, são a sensibilidade e a especificidade. A sensibilidade é definida como a probabilidade de um indivíduo ser classificado como zero, dado que realmente é zero e a especificidade é a probabilidade de um indivíduo ser classificado como um, dado que realmente é um.

A curva ROC (Zweig & Campbell, 1993) é construída variando os pontos de corte ao longo das probabilidades preditas pelos modelos, a fim de se obter as diferentes classifi-

cações dos indivíduos e obtendo consequentemente os respectivos valores para as medidas de sensibilidade e especificidade para cada ponto de corte estabelecido. Assim, a curva ROC é obtida tendo no seu eixo horizontal os valores de 1-Especificidade, ou seja, a proporção de uns que são classificados como zero pelo modelo, e, no eixo vertical a sensibilidade, que é a proporção de zeros que são realmente classificados como zeros. Uma curva ROC obtida ao longo da diagonal principal corresponde a uma classificação obtida sem a utilização de qualquer ferramenta preditiva, ou seja, sem a presença de modelos. Consequentemente, a curva ROC deve ser interpretada de forma que quanto mais a curva estiver distante da diagonal principal melhor o desempenho do modelo associado a ela. Esse fato sugere que quanto maior for a área entre a curva ROC produzida e a diagonal principal, melhor o desempenho global do modelo.

A curva ROC apresenta sempre um contrabalanço entre a sensibilidade e a especificidade ao se variar os pontos de corte ao longo das probabilidade preditas, e, pode ser usada para auxiliar na decisão de onde se localiza no melhor ponto de corte. Em geral, o melhor ponto de corte produz valores para sensibilidade e especificidade que se localiza no "ombro" da curva, ou próximo dele, ou seja, no ponto mais à esquerda e superior possível.

Quando se tem interesse em avaliar o modelo em um único ponto de corte, constrõe-se uma tabela 2 x 2 para o ponto de corte escolhido, denominada de matriz de confusão, representada na Figura 4-1. A partir deste matriz a sensibilidade e especificidade são obtidas. Neste trabalho, estas medidas são utilizadas com uma nomenclatura diferente, a sensibilidade será denominada de capacidade de acerto dos zeros e a especificidade como capacidade de acerto dos uns. Estas medidas são definidas como:

Capacidade de acerto total $(CAT) = \frac{b_1+m_0}{n}$

Capacidade de acerto dos zeros $(CAZ) = \frac{m_0}{A}$ (Sensibilidade)

Capacidade de acerto dos uns $(CAU) = \frac{b_1}{B}$ (Especificidade)

onde

n = número total de observações na amostra;

b_1 = número de uns que foram classificados como um (acerto);

m_0 = número de zeros que foram classificados como zero (acerto);

m_1 = número de uns que foram classificados como zero (erro);

b_0 = número de zeros que foram classificados como um (erro);

	Situação Real		
Previsão do Modelo	0	1	Total
0	m_0	m_1	a
1	b_0	b_1	b
Total	A	B	n

Figura 4-1: Matriz de Confusão

A = número de zeros na amostra

B = número de uns na amostra

a = número total de observações classificadas como zero na amostra

b = número total de observações classificadas como um na amostra.

Como geralmente, nas amostras de validação, onde os modelos são avaliados, se conhece a verdadeira resposta, torna-se possível comparar essa classificação obtida com a verdadeira resposta. A forma mais utilizada para estabelecer a matriz de confusão é determinar um ponto de corte na probabilidade preditiva e classificar os indivíduos com base nesse ponto. Essa matriz descreve portanto uma tabulação cruzada entre a classificação predita através de um único ponto de corte e a condição real e conhecida de cada indivíduo, onde a diagonal principal representa as classificações corretas e valores fora dessa diagonal correspondem a erros de classificação.

Uma outra medida que pode ser utilizada para avaliar a capacidade preditiva de um modelo é o logaritmo do score preditivo proposto por Good (1952). Esta medida vem mostrando ser um índice robusto e sensível. Para obtê-la o conjunto de dados deve ser dividido em dois subconjuntos: um conjunto de construção (D^C) e um conjunto de teste (D^T). O primeiro conjunto (D^C) é usado para o ajuste dos modelos e o segundo conjunto (D^T) é usado para se predizer a probabilidade da resposta para um novo indivíduo.

Desta forma, o logaritmo do score preditivo para um dado modelo M_k considerando a abordagem clássica é dado por

$$\sum_{d \in D^T} \log(\widehat{\theta}_k^d w_k) \tag{4.14}$$

onde d corresponde as observações individuais do D^T e $\widehat{\theta}_k^d$ é o valor da predição obtido pelo modelo k para a observação d do conjunto de teste.

Para a abordagem bayesiana.

$$\sum_{d \in D^T} \log \left\{ P(\theta / M_k, D^C) P(M_k / D^C) \right\} \tag{4.15}$$

De forma similar o logaritmo do score preditivo para a ponderação de modelos considerando a abordagem clássica é dada por

$$\sum_{d \in D^T} \log \left\{ \sum_{k=1}^{K} (\widehat{\theta}_k^d w_k) \right\}, \tag{4.16}$$

e por

$$\sum_{d \in D^T} \log \left\{ \sum_{k=1}^{K} P(\theta / M_k, D^C) P(M_k / D^C) \right\} \tag{4.17}$$

para a abordagem bayesiana.

Quanto maior o valor do logaritmo do score preditivo melhor a capacidade preditiva do modelo, permitindo então, medir o desempenho preditivo da metodologia em estudo.

Capítulo 5

Ponderação de Modelos em Regressão Logística

5.1 Abordagem Clássica

A abordagem clássica da ponderação de modelos em regressão logística segue a mesma metodologia da ponderação de modelos em regressão linear apresentada na seção 2.1. Cada um dos K modelos em consideração tem a mesma forma de (4.3).

Suponha uma situação onde têm-se apenas duas potenciais variáveis explicativas. Nesta situação existem 4 (2^2) possíveis modelos a serem considerados

$$\begin{aligned}
Modelo(1) &: \quad logit(\pi_i) = \beta_0 \\
Modelo(2) &: \quad logit(\pi_i) = \beta_0 + \beta_1 x_1 \\
Modelo(3) &: \quad logit(\pi_i) = \beta_0 + \beta_2 x_2 \\
Modelo(4) &: \quad logit(\pi_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.
\end{aligned}$$

O objetivo é predizer a média $\theta = \pi = \frac{\exp(\beta_0+\beta_1 x_{1+}+\beta_2 x_{2+})}{1+\exp(\beta_0+\beta_1 x_{1+}+\beta_2 x_{2+})}$ de uma variável futura Y_+ para valores x_{1+} e x_{2+}.

Os possíveis estimadores desta média são

$$\begin{aligned}
Modelo(1) &: \widehat{\theta}_1 = \widehat{\pi}_1 = [1+\exp(-\widehat{\beta}_0)]^{-1} \\
Modelo(2) &: \widehat{\theta}_2 = \widehat{\pi}_2 = [1+\exp(-\widehat{\beta}_0 - \widehat{\beta}_1 x_{1+})]^{-1} \\
Modelo(3) &: \widehat{\theta}_3 = \widehat{\pi}_3 = [1+\exp(-\widehat{\beta}_0 - \widehat{\beta}_2 x_{2+})]^{-1} \\
Modelo(4) &: \widehat{\theta}_4 = \widehat{\pi}_4 = [1+\exp(-\widehat{\beta}_0 - \widehat{\beta}_1 x_{1+} - \widehat{\beta}_2 x_{2+})]^{-1}.
\end{aligned}$$

Desta forma, o estimador para o parâmetro θ ponderado pelos pesos será dado por

$$\widehat{\theta} = w_1\widehat{\theta}_1 + w_2\widehat{\theta}_2 + w_3\widehat{\theta}_3 + w_4\widehat{\theta}_4. \tag{5.1}$$

Considerando o uso do AIC, os pesos w_k para cada um dos K modelos são calculados da seguinte forma

$$w_k = \frac{\exp(-AIC_k/2)}{\sum_{l=1}^{k}\exp(-AIC_l/2)},\ k = 1, ..., 4, \tag{5.2}$$

onde AIC_k é o critério de informação de Akaike para o modelo k. O logaritmo da função de verossimilhança maximizada para o modelo k no caso Bernouilli é

$$\log L_k = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_j + \beta_2 x_j) - \sum_{i=1}^{n}\log[1+\exp(\beta_0 + \beta_1 x_j + \beta_2 x_j)]. \tag{5.3}$$

A variância de $\widehat{\theta}$ pode ser obtida como em (2.12), e, no caso de regressão logística a $\widehat{Var}(\widehat{\theta}_k/\beta_k)$ é dada por $\widehat{Var}(\widehat{\theta}_k/\beta_k) = \sum_{j=1}^{p} x_{j+}^2 Var(\widehat{\beta}_j) + 2\sum_{j=1}^{p}\sum_{h=1}^{j} x_{h+}x_{j+}Cov(\widehat{\beta}_h, \widehat{\beta}_j)$. Outra abordagem para se obter a variância de $\widehat{\theta}$, como visto no caso de regressão linear, é através do uso do método *bootstrap* discutido na Seção seguinte.

5.2 O Método *Bootstrap* e seu uso em Regressão Logística

Davison & Hinkley (1997) consideram quatro planos de reamostragem para modelos lineares generalizados: reamostragem dos resíduos de Pearson padronizados, dos resíduos

padronizados na escala do preditor linear, dos resíduos deviance e reamostragem dos vetores de observações. Visto que, o modelo de regressão logística é um caso particular de modelos lineares generalizados, estes planos podem ser utilizados. Davison & Hinkley (1997) desenvolveram um estudo de simulação para comparar essas quatro abordagens e chegaram a conclusão que os resultados obtidos em cada uma delas são bastante similares. Sendo assim, neste trabalho, serão considerados apenas dois destes planos: o de reamostragem dos resíduos de Pearson padronizados e o de reamostragem dos vetores de observações.

Para a aplicação do plano de reamostragem dos resíduos de Pearson padronizados, os resíduos serão definidos como

$$r_{Pi} = \frac{y_i - \widehat{\mu}_i}{\{c_i \widehat{k} V(\widehat{\mu}_i)(1-h_i)\}^{1/2}},\ i = 1, ..., n, \tag{5.4}$$

onde h_i é o i-ésimo elemento da diagonal da matriz $X(X^TX)^{-1}X^T$, c_i são os pesos conhecidos e k é desconhecido. No caso de dados binários, $k = 1$ e $c_i = 1$. A partir dos valores de $y_1^*, ..., y_n^*$ obtém-se $\widehat{\beta}^* = X(X^TX)^{-1}X^TY^*$. Em grandes amostras espera-se que r_{Pj} tenha média próximo de zero e variância próxima de um, assim como em modelos de regressão linear.

Os métodos de reamostragem aqui considerados, seguem basicamente os mesmos procedimentos utilizados em modelos de regressão linear, como apresentado na Seção 2.2.1.

Abaixo seguem os algoritmos a serem utilizados em cada uma das abordagens consideradas. O algoritmo 3 descreve os passos para a reamostragem dos pares dos vetores de observações estratificada e o algoritmo 4 descreve os passos para a reamostragem dos resíduos de Pearson padronizados.

Algoritmo 3: Reamostragem dos Pares Estratificada

(i) separar a amostra em zeros e uns formando dois estratos, um de tamanho n_0 e outro de tamanho n_1, respectivamente;

(i) amostrar $i_1^*, ..., i_{n0}^*$ e $i_1^*, ..., i_{n1}^*$ aleatoriamente, com reposição, dentro do seu respectivo estratos, de $\{1, 2, ..., n_0\}$ e de $\{1, 2, ..., n_1\}$;

(ii) fazer $(y_j^*, \mathbf{x}_j^*) = (y_{i_j^*}, \mathbf{x}_{i_j^*})$, para $j = 1, ..., n$, obtendo a amostra *bootstrap*
$(y_1^*, \mathbf{x}_1^*)$,...,$(y_n^*, \mathbf{x}_n^*)$, tal que $n = n_0 + n_1$;

(iii) repetir os passos (i) e (ii), B vezes.

Algoritmo 4: Reamostragem dos Resíduos

(i) amostrar $\varepsilon_1^*, ..., \varepsilon_n^*$ aleatoriamente, com reposição, de $\varepsilon_1, ..., \varepsilon_n$, onde $\varepsilon_i = r_{Pi} - \overline{r}_P$, para $i = 1, ..., n$ e $\overline{r}_P$ é a média dos r_{Pi};

(ii) fazer $y_j^* = \widehat{\mu}_j + [c_i V(\widehat{\mu}_i)]^{1/2} \varepsilon_j^*$, para $j = 1, ..., n$, obtendo a amostra bootstrap $y_1^*, ..., y_n^*$;

(iii) repetir os passos (i) e (ii), B vezes.

O modelo a ser considerado no item (ii) do algoritmo 4 depende da especificação do modelo em consideração.

5.3 Abordagem Bayesiana

A abordagem bayesiana de ponderação de modelos em regressão logística segue a formulação geral de ponderação bayesiana de modelos apresentada no Capítulo 3. O desenvolvimento do BMA para modelos lineares generalizados é apresentado brevemente por Hoeting *et. al.* (1999), mas não há uma apresentação mais detalhada para o modelo de regressão logística, como pode ser encontrado, por exemplo, para análise de sobrevivência (Volinsky *et. al.*, 1997). O desenvolvimento aqui apresentado foi baseado em Volinsky *et. al.* (1997) e Raftery (1995).

Um ponto importante a ser considerado na implementação do BMA é a especificação da probabilidade a priori dos modelos, $P(M_k)$. Quando se tem pouca informação a priori sobre a plausibilidade dos modelos que estão sendo considerados, uma escolha razoável é assumir que todos os modelos são equiprováveis a priori. Raftery *et. al.* (1997), Madigan & Raftery (1994), e Madigan *et. al.* (1996) verificaram que quando o espaço de modelos é muito grande (mais de 10^{12} modelos) não há efeito perceptível em se atribuir uma distribuição uniforme a priori para os modelos. Já quando se tem informação a priori sobre a importância de uma variável, a probabilidade a priori do modelo M_k pode ser especificada como

$$P(M_k) = \prod_{j=1}^{p} \gamma_j^{\delta_{ij}} (1 - \gamma_j)^{1-\delta_{ij}}, \tag{5.5}$$

onde $\gamma_j \in [0, 1]$ é a probabilidade a priori de que $\beta_j \neq 0,\ j = 1, ..., p$, e δ_{kj} é uma variável indicadora de quando a variável j é ou não incluída no modelo M_k. Atribuir $\gamma_j = 0,5$,

para todo j, é correspondente a atribuir uma priori uniforme no espaço de modelos. Fazer $\gamma_j < 0,5$, para todo j, impõe uma penalidade para modelos com muitas covariáveis e usar $\gamma_j = 1$ faz com que a variável j seja incluída em todos os modelos. Usando essa metodologia, a definição da probabilidade a priori para os modelos é simples e dispensa a necessidade da definição das prioris para um grande número de modelos.

Segundo Raftery (1995), quando todos os modelos são considerados iguais a priori, ou seja, usando $\gamma_j = 0,5$, a probabilidade a posteriori para o modelo M_k pode ser aproximada por

$$P(M_k/D) \approx \exp(-\frac{1}{2}BIC_k)/\sum_{l=1}^{K}\exp(-\frac{1}{2}BIC_l). \tag{5.6}$$

O critério de informação bayesiana (BIC), desenvolvido por Schwarz (1978), pode ser obtido aproximando-se a integral presente na equação (3.3) via método de Laplace. Raftery (1996) apresenta todo o desenvolvimento para se obter o BIC e mosta que este pode ser calculado como

$$BIC_k = L_k^2 - df_k \log n, \tag{5.7}$$

onde L_k^2 é a deviance do modelo k considerando-se a distribuição de Bernouilli, df_k são os graus de liberdade correspondente e n o número de observações.

Na equação (3.1), a distribuição preditiva de θ, dado um modelo particular M_k, é encontrada integrando-se em relação ao parâmetro do modelo, β_k :

$$P(\theta/M_k, D) = \int P(\theta/\beta_k, M_k, D)P(\beta_k/M_k, D)d\beta_k. \tag{5.8}$$

Como esta integral não tem uma forma fechada, utiliza-se a aproximação

$$P(\theta/M_k, D) \approx P(\theta/M_k, \widehat{\beta}_k, D), \tag{5.9}$$

onde $\widehat{\beta}_k$ é o estimador de máxima verossimilhança de β_k obtido via (A.28).

No contexto de incorporação da incerteza devido a escolha do modelo, esta aproximação foi utilizada por Taplin (1993) que encontrou uma excelente aproximação para problema de regressão de séries temporais e posteriormente utilizada por Taplin e Raftery (1994) e Draper (1995).

A média e a variância a posteriori são dadas por

$$E[\theta/D] = \sum_{k=0}^{K} E[\widehat{\theta}/D, M_k]\ P(M_k/D) \tag{5.10}$$

$$Var[\theta/D] = \sum_{k=0}^{K} (Var[\widehat{\theta}/D, M_k] + E[\widehat{\theta}/D, M_k]^2)\ P(M_k/D) - E[\widehat{\theta}/D]^2. \tag{5.11}$$

É importante ressaltar que a priori utilizada para a obtenção destes resultados considera $\gamma_j = 0,5$, ou seja, todos os modelos são igualmente prováveis a priori.

Capítulo 6

Aplicação

No capítulo 5 foram apresentadas as metodologias utilizadas para as abordagens de ponderação de modelos em regressão logística.

Neste capítulo serão apresentados exemplos com a aplicação da metodologia de ponderação de modelos e do método de seleção de modelos *Stepwise*, como forma de comparar o desempenho preditivo do método de ponderação. Na Seção 6.1 será apresentado um exemplo de regressão logística com apenas duas covariáveis. Neste exemplo serão aplicadas a ponderação clássica (incluindo o uso do método *bootstrap*), a abordagem bayesiana e o método *Stepwise* e verificado o resultado. Com base neste exemplo será feito um estudo de simulação com o objetivo de avaliar as propriedades das abordagens clássica e bayesiana (através do vício e variância do estimador) e avaliar também o desempenho do uso *bootstrap* para a obtenção da variância do estimador ponderado. Na Seção 6.2 será apresentado uma aplicação em uma situação de regressão logística com 13 covariáveis e alto grau de incerteza na escolha do modelo, com o objetivo de comparar a capacidade preditiva da ponderação com a do *Stepwise*. Um estudo de simulação será realizado para tentar obter um resultado mais conclusivo. E na Seção 6.3 a metodologia de ponderação de modelos será aplicada a um conjunto de dados reais.

Em todos os exemplos será fornecida a fonte dos dados. Os dados do exemplo 3 foram gentilmente cedidos por um pesquisador e não estão disponíveis para utilização sem prévia autorização.

Os cálculos e gráficos foram feitos usando-se o software R e os programas estão no Apêndice B. Os cálculos envolvendo o BMA foram baseados na função BIC_GLM, de-

senvolvido por Chris Volinsky usando-se o software S-PLUS e obtida como indicado em Hoeting *et. al.* (1999) e atualmente essas funções encontram-se diponíveis no software R.

6.1 Exemplo 1

Este exemplo foi obtido de Neter, Kutner, Nachtsheim & Wasserman (1996), página 619. Uma clínica de saúde enviou avisos à seus clientes para encorajá-los, principalmente os mais idosos que tem maiores riscos de complicações, a tomarem injeção contra a gripe, visando proteção em uma esperada epidemia. Em um estudo piloto, 50 clientes foram selecionados aleatoriamente e questionados se eles tinham ou não recebido uma injeção contra a gripe. Foram coletados também a idade deste pacientes (X_1) e seu conhecimento sobre a doença. Estes dados foram combinados em um índice de conhecimento sobre a doença (X_2), para os quais, valores altos indicam grande conhecimento. O cliente que recebeu a injeção contra gripe foi codificado como $Y = 1$, e o cliente que não recebeu a injeção contra gripe como $Y = 0$.

Como o exemplo considera duas covariáveis, os 2^2 possíveis modelos a serem ajustados são:

$$\begin{aligned}
Modelo(1) &: logit(\pi_i) = \beta_0 \\
Modelo(2) &: logit(\pi_i) = \beta_0 + \beta_1 x_1 \\
Modelo(3) &: logit(\pi_i) = \beta_0 + \beta_2 x_2 \\
Modelo(4) &: logit(\pi_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.
\end{aligned}$$

Os resultados dos ajustes obtidos foram

$$\begin{aligned}
Modelo(1) &: logit(\widehat{\pi}_i) = -0,3228 \\
Modelo(2) &: logit(\widehat{\pi}_i) = -6,5763 + 0,1331 x_1 \\
Modelo(3) &: logit(\widehat{\pi}_i) = -7,3902 + 0,1349 x_2 \\
Modelo(4) &: logit(\widehat{\pi}_i) = -21,5846 + 0,2218 x_1 + 0,2035 x_2.
\end{aligned}$$

Considerando o método de seleção de modelos, *Stepwise*, o modelo selecionado foi o

modelo completo, ou seja,

$$Modelo(4) : logit(\widehat{\pi}_i) = -21,5846 + 0,2218x_1 + 0,2035x_2.$$

Na Tabela 6-1 estão apresentados, para os 4 possíveis modelos, os pesos e as probabilidades a posteriori.

Tabela 6-1: Pesos dos 4 possíveis modelos

Modelo	**W**	**PostProb**
1	$< 0,0001$	$< 0,0001$
2	$< 0,0001$	$< 0,0001$
3	$0,0006$	$0,0015$
4	$0,9994$	$0,9984$

Pela análise da Tabela 6-1, nota-se que não há incerteza devido a escolha do modelo, sendo o modelo 4 claramente o favorito.

O estimador ponderado $\widehat{\theta}$ foi obtido para a predição em todas as 50 observações do conjunto de dados. As variâncias destas estimativas para cada uma das abordagens consideradas estão representadas na Figura 6-1. Observa-se, pelo gráfico, que há uma completa concordância entre as variâncias das abordagens clássica, bayesiana e do método *Stepwise*, o que era esperado devido a não haver incerteza quanto à escolha do modelo. A variância da abordagem *bootstrap* - reamostragem dos resíduos acompanha o comportamento das anteriores embora sejam maiores nos pontos mais extremos e a variância *bootstrap* - reamostragem dos pares estratificado tem uma variância maior, em média para todos os pontos do que a apresentada pelos outros métodos, além de um comportamento diferente.

Para aplicação do método *bootstrap*, tanto no caso de reamostragem dos resíduos como para a reamostragem dos pares estratificada, foram realizadas $B = 10000$ reamostras.

Na Figura 6-2 estão apresentados os gráficos de convergência dos métodos *bootstrap* para um exemplo de um valor de predição. Pode-se notar que para o método *bootstrap* - reamostragem dos resíduos a variância se estabilizou num valor próximo de $B = 4000$ e para o método *bootstrap* - reamostragem dos pares estratificados a variância se estabilizou num valor próximo de $B = 7000$. Estes valores de B foram utilizados no estudo de

do exemplo.pdf

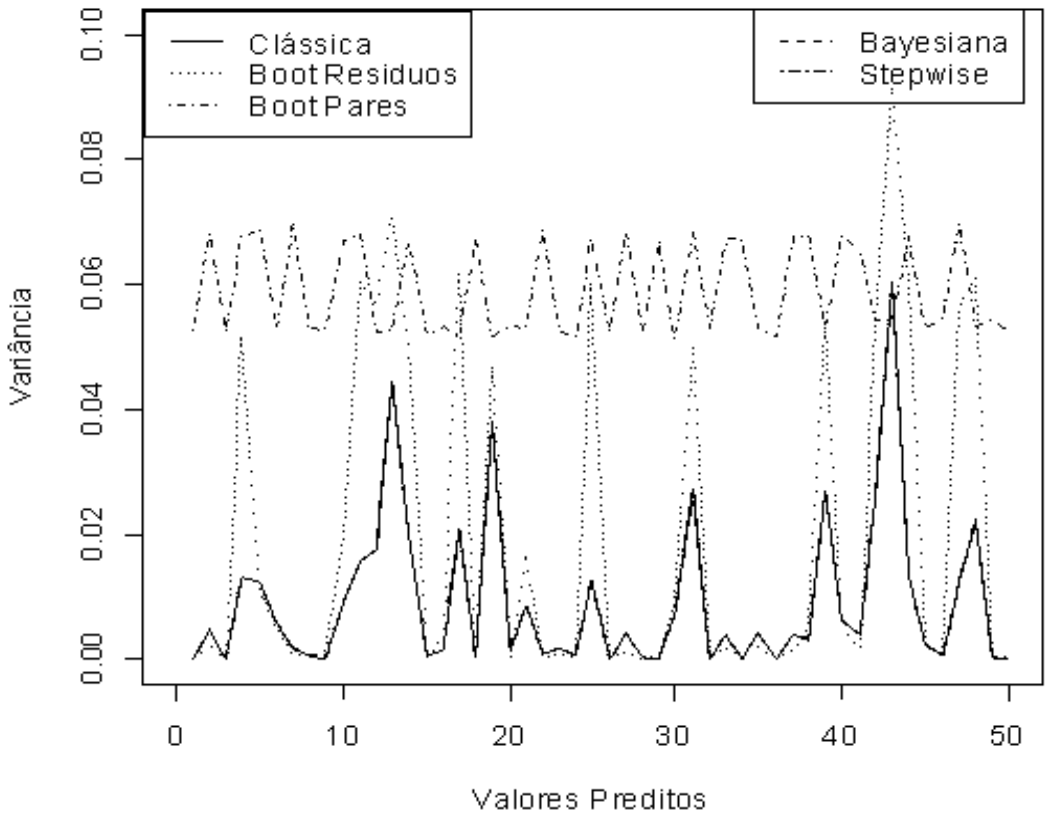

Figura 6-1: Variância do estimador ponderado $\widehat{\theta}$ nas abordagens clássica, bayesiana, no método Stepwise e nas abordagens bootstrap - reamostragem dos resíduos e reamostragem dos pares estratificada. As linhas correspondentes as variâncias das abordagens clássica, bayesiana e Stepwise estão sobrepostas.

simulação apresentado na Seção seguinte.

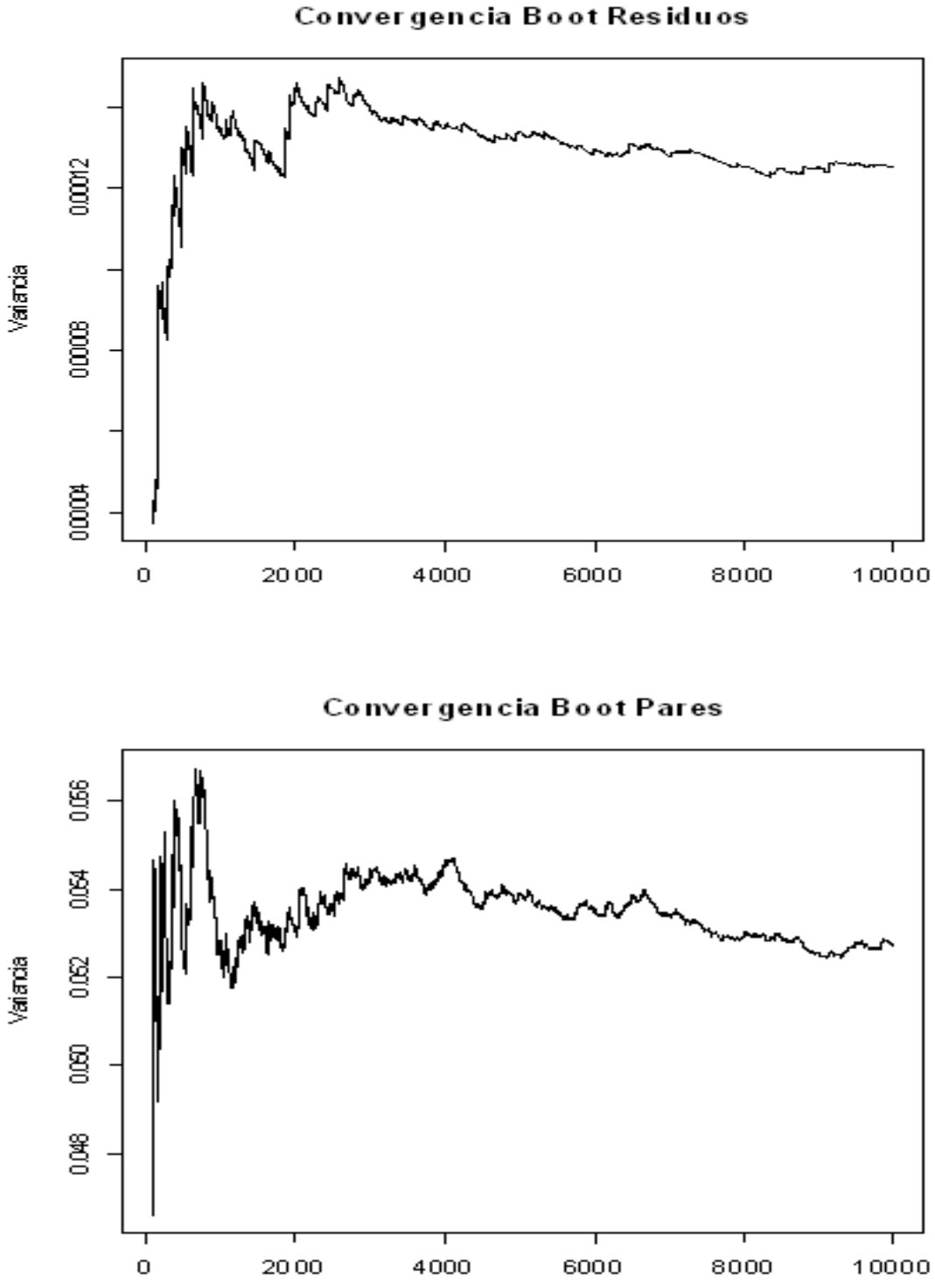

Figura 6-2: Estudo da estabilização da variância de $\widehat{\theta}$ no método bootstrap - reamostragem dos resíduos e reamostragem dos pares estratificada.

A classificação do estimador ponderado $\widehat{\theta}$, para cada uma das abordagens consideradas, em um dos dois grupos, 0 ou 1, será feita, considerando o uso da curva ROC, apresentada na Seção 4.3. O valor do ponto de corte, para cada um dos métodos, foi determinado como sendo o ponto máximo da soma da sensibilidade e da especificidade. Os gráficos das curvas ROC correspondentes a cada uma das abordagens estão apresentados na Figura 6-3, 6-4 e 6-5, respectivamente. Na Tabela 6-2 estão apresentados os pontos de corte, as medidas de capacidade de acerto (CAT, CAZ, CAU) e a área sob a curva ROC (AUC) para cada abordagem.

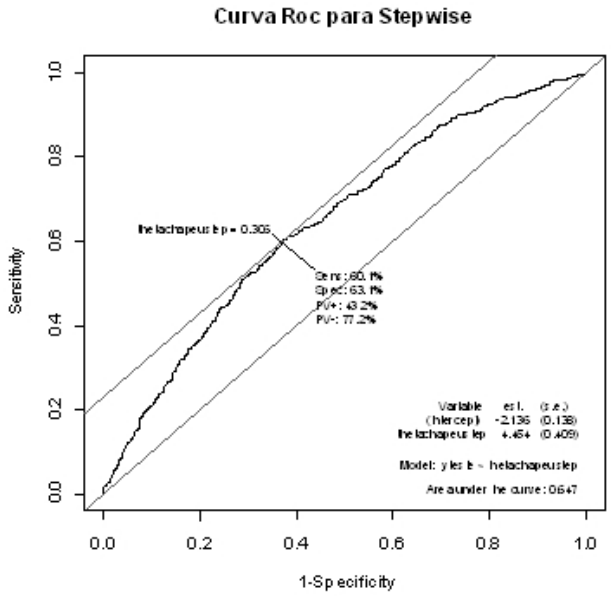

Figura 6-3: Curva Roc do método de seleção de modelos *Stepwise*.

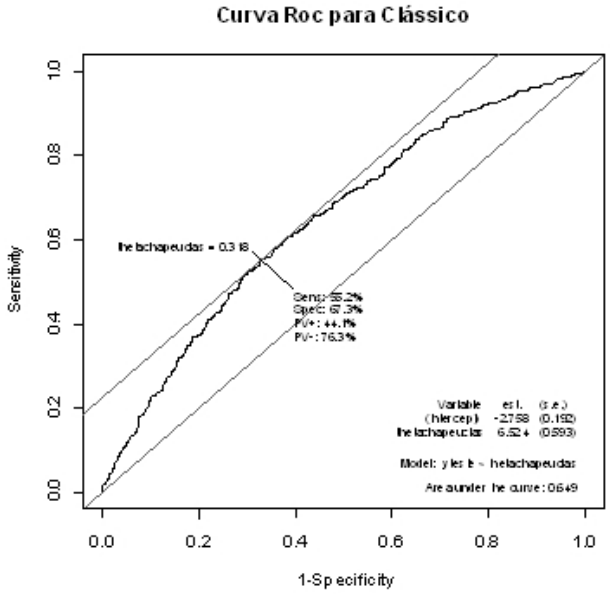

Figura 6-4: Curva ROC do método de ponderação de modelos abordagem clássica.

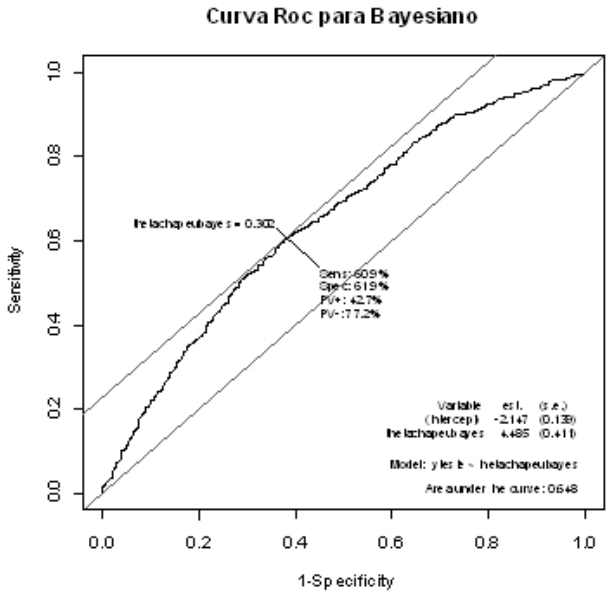

Figura 6-5: Curva ROC do método de ponderação de modelos abordagem bayesiana.

Tabela 6-2: Pontos de corte, medidas de capacidade de acerto e área sob a curva ROC.

	Ponto de Corte	**CAT**	**CAU**	**CAZ**	**AUC**
Stepwise	0,651	86%	96,6%	71,4%	0,923
Clássico	0,651	86%	96,6%	71,4%	0,923
Bayesiano	0,651	86%	96,6%	71,4%	0,923

Pela Tabela 6-2, nota-se que o ponto de corte, as medidas de capacidade preditiva e as áreas sob as curvas ROC para as abordagens clássica, bayesiana e o método *stepwise* são as mesmas. Não havendo, desta forma, para este exemplo, uma abordagem que se destacasse como sendo a melhor. Este resultado deve-se, provavelmente, ao fato de não haver a presença de incerteza na escolha do modelo.

6.1.1 Estudo de Simulação

O estudo de simulação aqui apresentado foi realizado com base no exemplo da Seção anterior. Este estudo tem por objetivo verificar as propriedades dos dois métodos de ponderação, clássico e bayesiano, através do vício e variância correspondentes e o desempenho das estimativas das variâncias *bootstrap*. Para a realização do estudo foram considerados dois tamanhos de amostra $n = 20$, retirado aleatoriamente do conjunto de dados original, e $n = 50$.

Foram considerados dois conjuntos de valores dos parâmetros do modelo, $\beta_0, \beta_1, \beta_2$ com o objetivo de intensificar a incerteza. Estes dois conjuntos considerados foram: **conjunto 1 -** valores dos parâmetros estimados pelo ajuste dos modelos completos considerando os dois tamanhos de amostra em estudo ($n = 20$ e $n = 50$). **Conjunto 2** - os valores anteriores foram modificados de forma a aumentar a incerteza na escolha do modelo. Os dois conjuntos de valores de β's utilizados para fazer o estudo de simulação estão apresentados na Tabela 6-3.

Tabela 6-3: Conjunto de valores de β's utilizados para fazer o estudo de simulação.

	n	β_0	β_1	β_2
Conjunto 1	50	-21,5846	0,2218	0,2035
	20	-23,5284	0,1925	0,2885
Conjunto 2	50	-21,5846	0,1757	0,1578
	20	-23,5284	0,1500	0,2000

O procedimento de simulação constituiu em gerar $R = 10000$ novos vetores de respostas y a partir de uma distribuição binomial com probabilidade de sucesso

$$\widehat{\pi}_i = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}. \tag{6.1}$$

Os vetores x_1 e x_2 foram mantidos iguais ao do exemplo da Seção 6.1. Para cada novo conjunto de dados, formado pelo novo vetor da variável resposta y e pelas covariáveis x_1 e x_2, as metodologias apresentadas foram aplicadas obtêndo-se o estimador ponderado $\widehat{\theta}$, a variância Buckland (2.12) e a variância BMA (5.10). O vício foi calculado como sendo a diferença entre o valor verdadeiro, obtido com os $\beta_0, \beta_1, \beta_2$ especificados acima, e as médias das estimativas obtidas na simulação. As variâncias das estimativas simuladas serão chamadas de variâncias simuladas.

Para o método *bootstrap* foram feitas $R = 100$ replicações e $B = 4000$ e $B = 7000$ reamostras para a reamostragem dos resíduos e para a reamostragem dos pares, respectivamente. O valor de R, neste caso foi reduzido devido ao fato do algoritmo computacional requerer muito tempo de execução.

Para avaliar as propriedades do estimador $\widehat{\theta}$, foram comparados os valores dos vícios nas abordagens clássica e bayesiana, e das variâncias simuladas nas abordagens clássica e bayesiana, a média das variâncias obtidas por Buckland (2.12), as médias das variâncias segundo a abordagem bayesiana (5.10) e as médias das variâncias nas duas abordagens *bootstrap*.

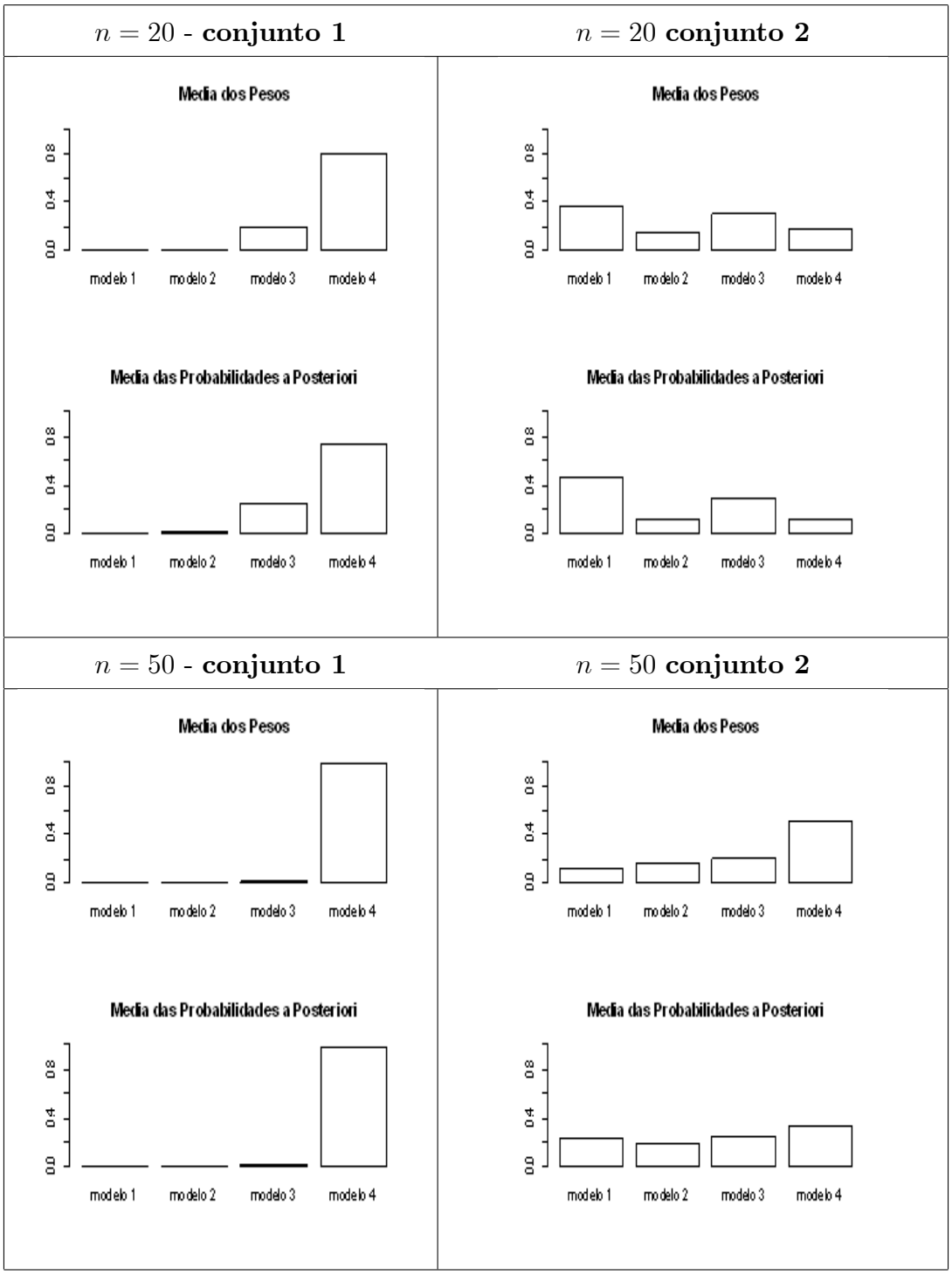

Figura 6-6: Histograma da média dos pesos e das probabilidades a posteriori dos modelos dos 10000 valores simulados.

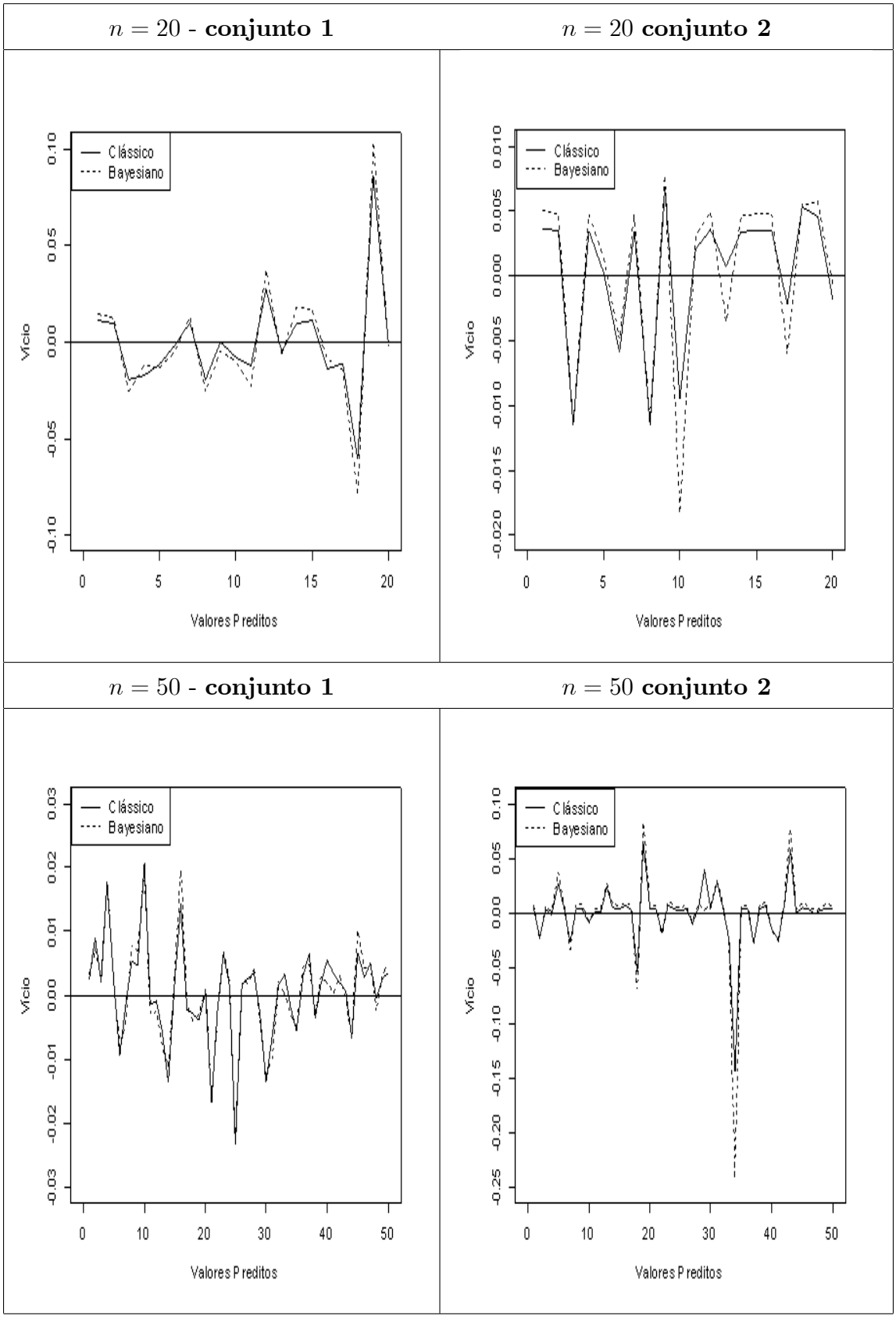

Figura 6-7: Vício do estimador ponderado $\widehat{\theta}$ nas abordagens clássica e bayesiana.

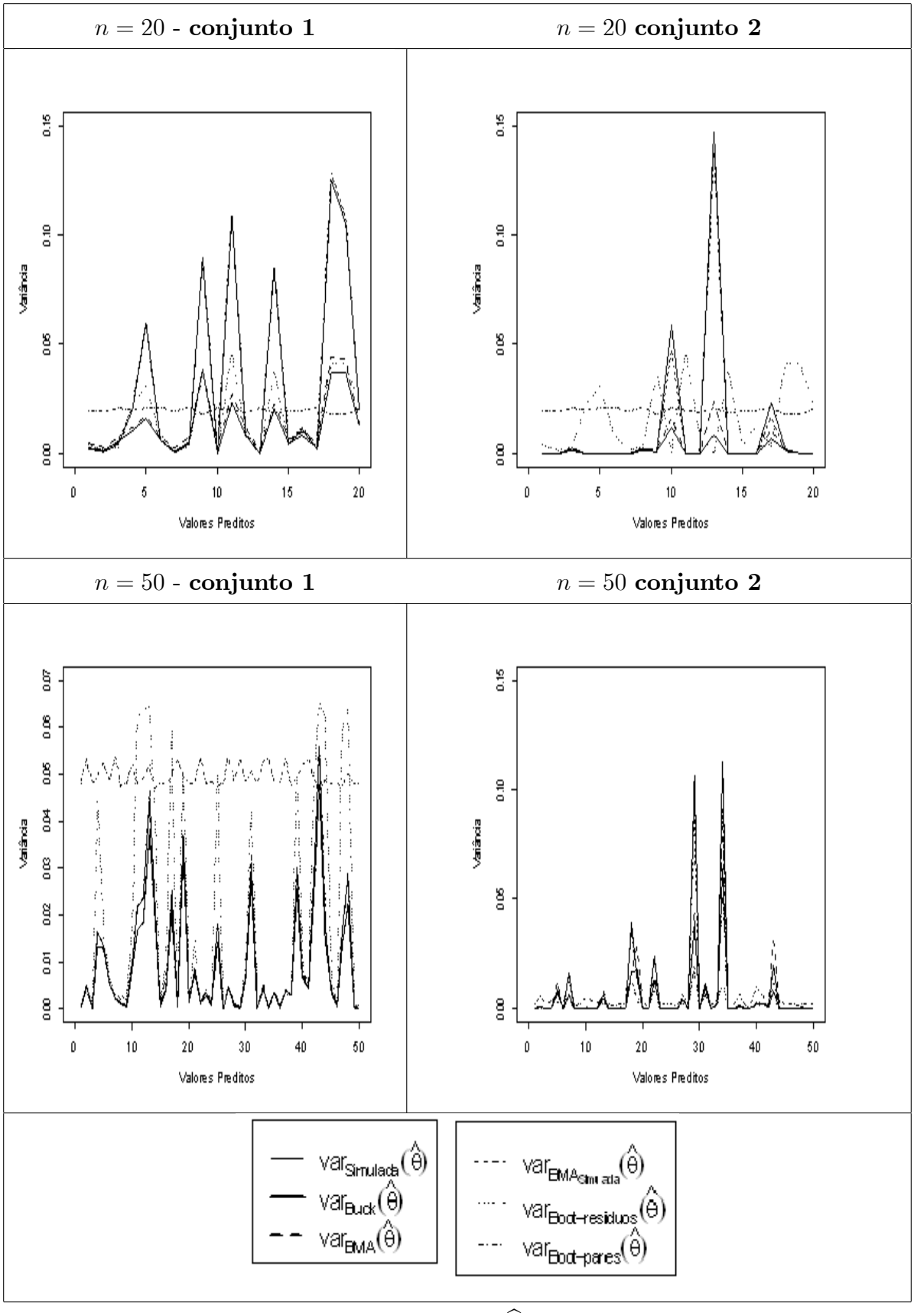

Figura 6-8: Variância do estimador ponderado $\widehat{\theta}$ nas abordagens clássica,bayesiana e nos dois métodos de reamostragem *bootstrap*.

A Figura 6-6 mostra os histogramas da média dos pesos e das probabilidades a posteriori dos modelos dos 10000 valores simulados. Observa-se que no **conjunto 1** os pesos e as probabilidades a posteriori dos modelos estão mais concentrados em apenas um modelo e no **conjunto 2**, estes valores encontram-se mais dispersos entre os possíveis modelos mostrando o aumento da incerteza devido à escolha do modelo.

A Figura 6-7 mostra os vícios de estimação nas abordagens clássica e bayesiana. Há praticamente uma concordância entre os vícios destas abordagens em todas as situações consideradas. Vale observar que aumentando o tamanho da amostra há um diminuição nos valores dos vícios.

A Figura 6-8, mostra as variâncias estimadas nas abordagens clássica, bayesiana e nas duas abordagens *bootstrap* consideradas. Observa-se que as estimativas obtidas pela proposta de Buckland *et. al.* (1997) e do estimador BMA acompanham o comportamento das variâncias obtidas por simulação, e, por serem médias dos valores obtidos em cada simulação, apresentam um comportamento suavizado nos picos. O método *bootstrap* - reamostragem dos resíduos apresenta um comportamento próximo aos anteriores. Já as estimativas obtidas para o método *boostrap* - reamostragem dos pares estratificado apresentam um comportamento diferente das outras abordagens. Apenas no **conjunto 2**, para $n = 50$, o comportamento obtido é o mesmo das outras abordagens consideradas.

Pode-se concluir de forma geral, para as configurações consideradas nesta simulação, que com o aumento do tamanho da amostra observa-se maior concordância entre as variâncias, ou seja, as abordagens produzem resultados mais próximos entre si.

6.2 Exemplo 2

Nesta seção será apresentado um exemplo de aplicação do método de ponderação de modelos em um exemplo de regressão logística com 13 covariáveis e alto grau de incerteza na escolha do modelo. Será realizado um estudo de simulação com o objetivo de verificar se há, em média, um aumento das medidas de capacidade preditiva do métodos de ponderação em relação ao método de seleção de modelos *Stepwise*.

6.2.1 Exemplo

A porcentagem de gordura corporal é comumente utilizada como indicador de saúde. Esta medida pode ser obtida por vários métodos, como medidas feitas em baixo d 'água e medidas feita por impulsos elétricos. O problema é que esses métodos requerem equipamentos e profissionais especializados. Uma alternativa para resolver esse problema, é obter medidas simples de características corporais, como peso e altura, na tentativa de se predizer se uma pessoa está ou não acima da faixa do percentual de gordura ideal.

O conjunto de dados aqui utilizado foi obtido em Johnson (1996). Este conjunto é formado de 252 observações feitas em pessoas do sexo masculino. Para cada indivíduo, a porcentagem de gordura corporal, o peso, a idade, a altura e 10 medidas circunferênciais foram obtidas. A descrição desta medidas esta apresentada na Tabela 6-4.

Tabela 6-4: Descrição do conjunto de dados.

Variável	**Descrição**
X1	Idade (anos)
X2	Peso (libras)
X3	Altura (polegada)
X4	Circunferência do pescoço (cm)
X5	Circunferência toráxica (cm)
X6	Circunferência abdominal (cm)
X7	Circunferência do quadril (cm)
X8	Circunferência da coxa (cm)
X9	Circunferência do joelho (cm)
X10	Circunferência do tornozelo (cm)
X11	Circunferência do extensor do bíceps (cm)
X12	Circunferência do antebraço (cm)
X13	Circunferência do punho (cm)

A observação 42 foi omitida devido ao fato da existência de uma medida aparentemente errônea. Cada indivíduo pertencente ao conjunto de dados foi classificado de acordo com a faixa do percentual de gordura ideal de acordo com os índices determinados na Tabela 6-

5, obtidos no site http://www.saudeemmovimento.com.br/saude/tabelas. Desta forma, o indivíduo foi classificado como 1 quando seu percentual de gordura corporal se encontrava abaixo do índice determinado na Tabela 6-5 e como 0 caso contrário.

Tabela 6-5: Faixa de Percentual de Gordura Ideal de acordo com Sexo e Idade

Faixa Etária	**Homens**	**Mulheres**
de 18 a 29 anos	14%	19%
de 30 a 39 anos	16%	21%
de 40 a 49 anos	17%	22%
de 50 a 59 anos	18%	23%
acima de 60 anos	21%	26%

Como o exemplo considera 13 covariáveis, existem então $2^{13} = 8192$ modelos a serem ajustados.

Para o método de seleção de modelos *Stepwise*, o modelo selecionado foi:

$$y = \beta_0 + \beta_1 x1 + \beta_2 x2 + \beta_4 x4 + \beta_6 x6 + \beta_7 x7 + \beta_8 x8 + \beta_{12} x12 + \beta_{13} x13$$

Para a aplicação, tanto da metodologia de ponderação de modelos como do método de seleção de modelos *Stepwise*, o conjunto de dados foi dividido, de forma aleatória, em dois subconjuntos: um conjunto de construção (D^C) composto por 142 observações e um conjunto de teste (D^T) formado pelas 109 observações restantes. Utilizando os dados de construção (D^C), os 8192 possíveis modelos foram ajustados e destes foram selecionados os modelos cujos pesos ou probabilidades a posteriori somavam 90% de incerteza, resultando em 881 modelos selecionados pela abordagem clássica e 733 modelos selecionados pela abordagem bayesiana. Na Tabela 6-6 estão apresentados, em ordem decrescente, os AIC's e pesos dos 10 modelos com maiores pesos e na Tabela 6-7 os BIC's e probabilidades a posteriori dos 10 modelos com maiores probabilidades a posteriori.

Tabela 6-6: Os 10 modelos com maiores pesos

Modelo	**AIC**	**W**
2219	170,2938	0,0181
3921	170,9800	0,0128
3011	171,3707	0,0106
931	171,3761	0,0105
1426	171,4178	0,0103
3976	171,8140	0,0085
3986	171,9792	0,0078
5064	172,0616	0,0075
3990	172,2129	0,0069
3725	172,2556	0,0068

Tabela 6-7: Os 10 modelos com maiores probabilidades a posteriori

Modelo	**BIC**	**PostProb**
2219	-470,2534	0,0291
931	-469,8102	0,0233
1426	-469,195	0,0166
3921	-468,9281	0,0150
3011	-468,5374	0,0123
2198	-468,1456	0,0102
932	-468,1115	0,0100
3976	-468,0941	0,0099
2148	-467,9340	0,0091
3986	-467,9290	0,0091

O que pode ser observado pela análise das Tabelas 6-6 e 6-7 é que, tanto na ponderação clássica como na bayesiana, há uma considerável incerteza devido à escolha do modelo, pois o modelo com maior peso $(0,0181)$ representa apenas $1,81\%$ do peso total e o modelo com maior probabilidade a posteriori $(0,0291)$ representa apenas $2,91\%$ da probabilidade a posteriori total, o que indica que não há nenhum modelo que poderia ser selecionado e

tido como o melhor modelo para se fazer a predição que se deseja.

Utilizando agora os dados de teste (D^T) e apenas os modelos selecionados, as predições para cada uma das abordagens consideradas foram obtidas. As Figuras 6-9, 6-10 e 6-11 mostram as curvas ROC para cada uma das abordagens consideradas. Na Tabela 6-8, estão apresentados os pontos de corte, as medidas de capacidade preditiva, a área sob a curva ROC e o logaritmo do score preditivo, descritos na Seção 4.3, obtidos para cada uma das abordagens consideradas.

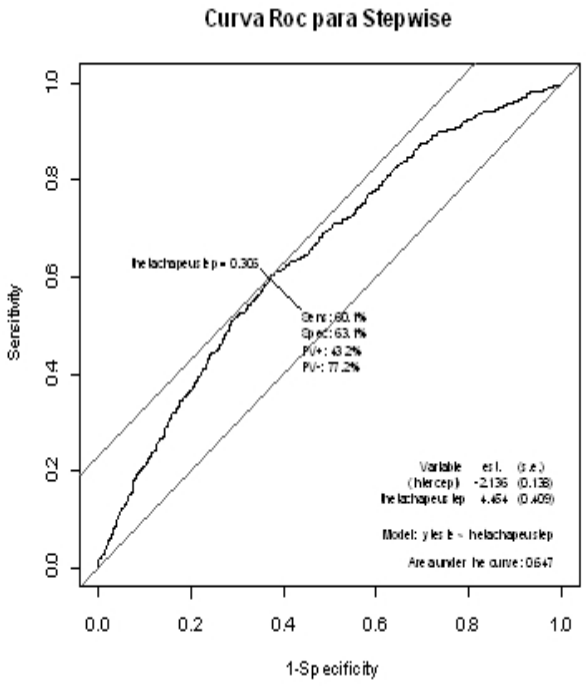

Figura 6-9: Curva ROC do método de seleção de modelos *Stepwise*.

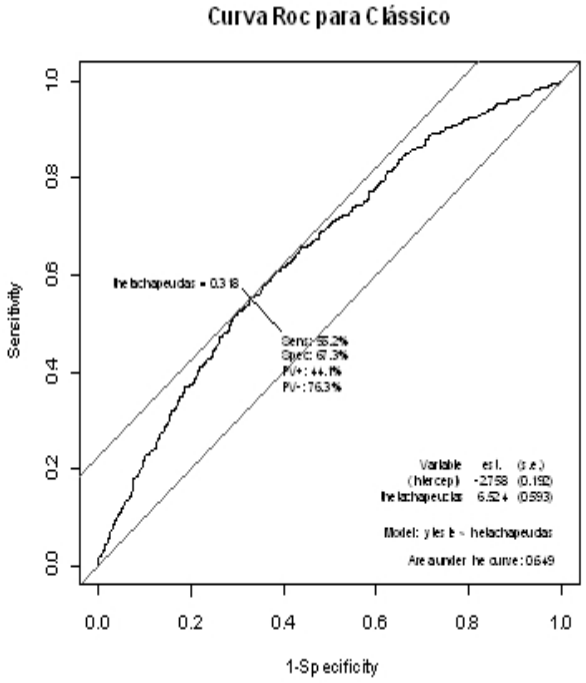

Figura 6-10: Curva ROC do método de ponderação de modelos abordagem clássica.

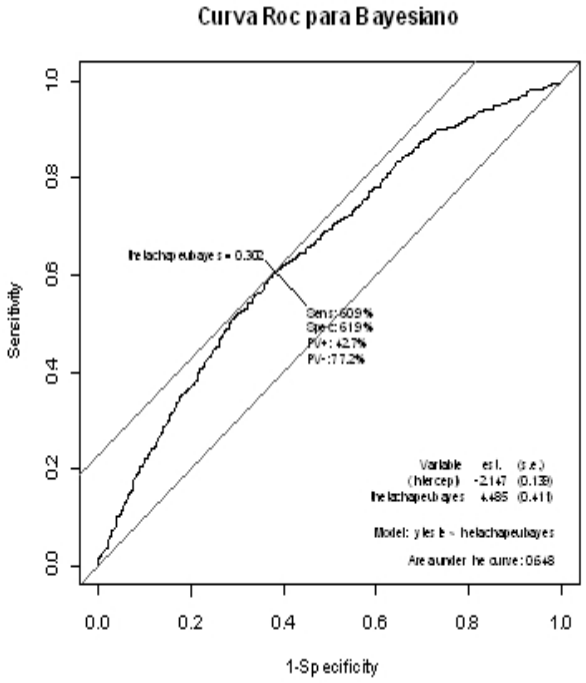

Figura 6-11: Curva ROC do método de ponderação de modelos abordagem bayesiana.

Tabela 6-8: Medidas de Capacidade Preditiva

	Pto de Corte	**CAT**	**CAU**	**CAZ**	**AUC**	**Log Score**
Stepwise	0,451	87,2%	86,9%	87,5%	0,926	-295,7205
Clássico	0,552	87,2%	88,5%	85,4%	0,917	-281,2239
Bayesiano	0,594	87,5%	88,5%	85,4%	0,915	-275,4816

Pela análise da capacidade preditiva total (CAT), nota-se que não houve difirença entre as abordagens. Nas outras medidas obteve-se um pouco de variação, no caso da capacidade preditiva dos uns (CAU) esta é maior para o método de ponderação de modelos do que para o método de seleção de modelos *Stepwise*. Já no caso da capacidade preditivados zeros (CAZ) essas medidas se inverteram. Já pela análise da do logaritmo do score preditivo observa-se que o método de ponderação de modelos é melhor do que o método de seleção de modelos *Stepwise*. A melhora no score preditivo do método de ponderação de modelos em relação ao método de seleção de modelos *Stepwise* foi de $\delta_1 = 20,239$ na abordagem bayesiana e de $\delta_2 = 14,4966$ na abordagem clássica. O conjunto de teste era composto de $n_{teste} = 109$ observações, então, o resultado significa que, em média, a probablidade preditiva do método de ponderação de modelos, tanto na abordagem clássica como na bayesiana, foi maior do que o método de seleção de modelos *Stepwise* por um fator de $\exp(\delta_1/n_{teste}) = 1,204$, ou seja, $20,40\%$ na abordagem bayesiana e $\exp(\delta_2/n_{teste}) = 1,142$, ou seja, $14,20\%$ na abordagem clássica. Em outras palavras, o método de ponderação de modelos prediz se um indivíduo está abaixo da faixa de percentual de gordura ideal $20,40\%$ melhor na abordagem bayesiana e $14,2\%$ na abordagem clássica de ponderação de modelos do que o método de seleção de modelos *Stepwise*. A análise obtida pelas medidas de capacidade preditiva aqui utilizadas (CAT, CAU, CAZ, AUC) não estão em concordância com a análise obtida via logaritmo do score preditivo. Não é possível chegar a um resultado conclusivo com base em apenas um exemplo. Para verificar se ocorre, em média, alguma diferença entre a capacidade preditiva para estas abordagens foi realizado um estudo de simulação que está descrito a seguir.

6.2.2 Estudo de Simulação

O estudo desenvolvido tem por objetivo verificar se há, em média, um aumento na capacidade preditiva. Esta verificação será feita considerando-se as médias das medidas de capacidade preditiva, as médias das áreas sob a curva ROC e as médias do logaritmo do score preditivo, como descrito na Seção 4.3.

O procedimento de simulação constituiu em gerar $R = 1000$ novos vetores da variável resposta y, a partir de uma distribuição binomial com probabilidade de sucesso

$$\widehat{\pi}_i = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{12} x_{12} + \beta_{13} x_{13})}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{12} x_{12} + \beta_{13} x_{13})},$$

onde os valores de $\beta_0, ..., \beta_{13}$ foram os estimados no exemplo anterior. Os vetores das covariáveis foram mantidos igual ao do exemplo anterior. Neste estudo o número de replicações foi fixado em apenas 1.000 devido ao grande número de modelos em estudo, o que tornou o procedimento lento. Para cada novo conjunto de dados, formado pelo novo vetor da variável resposta y e pelas 13 covariáveis, o mesmo procedimento realizado no exemplo anterior foi aplicado, apenas para os modelos que somavam 90% da incerteza. Em cada replicação foram obtidas as medidas de capacidade preditiva, área sob a curva ROC e o logaritmo do score preditivo, além das estimativas dos estimadores ponderados. Na Tabela 6-9 estão descritos os valores médios obtidos para cada uma das medidas de capacidade preditiva.

Tabela 6-9: Medidas de Capacidade Preditiva

	CAT	**CAU**	**CAZ**	**AUC**	**Log Score**
Stepwise	85,3	83,9	87,5	0,918	-338,3515
Clássico	85,9	84,4	88,0	0,925	-308,7339
Bayesiano	85,9	84,5	88,1	0,953	-300,7323

Pela análise da Tabela 6-9 verifica-se que, em média, o método de ponderação de modelos é melhor do que o método de seleção de modelos *Stepwise*. Todas as medidas de capacidade, CAT, CAU, CAZ e AUC apresentaram-se maior no método de ponderação de modelos do que no método de seleção de modelos *Stepwise*. A melhora no score preditivo para a ponderação de modelos - abordagem bayesiana - em relação ao método de seleção de

modelos *Stepwise* foi de $\delta_1 = 37,6192$ enquanto que a melhora do método de ponderação de modelos - abordagem clássica - em relação ao método de seleção de modelos *Stepwise* foi de $\delta_2 = 29,6176$. O conjunto de teste era composto de $n_{teste} = 109$ observações, então, o resultado significa que, em média, a probablidade preditiva do método de ponderação de modelos, tanto na abordagem clássica como na bayesiana, foi maior do que o método de seleção de modelos *Stepwise* por um fator de $\exp(\delta_1/n_{teste}) = 1,412174$, ou seja, $41,22\%$ na abordagem bayesiana e $\exp(\delta_2/n_{teste}) = 1,312221$, ou seja, $31,22\%$ na abordagem clássica. Em outras palavras, o método de ponderação de modelos prediz se um indivíduo está abaixo da faixa de percentual de gordura ideal $41,22\%$ melhor na abordagem bayesiana e $31,22\%$ na abordagem clássica de ponderação de modelos do que o método de seleção de modelos *Stepwise*.

Assim, pode ser verificado um aumento na capacidade preditiva em relação as medidas obtidas no exemplo da Seção 6-3-1.

6.3 Exemplo 3

O conjunto de dados aqui utilizado é advindo de uma instituição financeira de grande porte e foi gentilmente cedido pelo Prof. Dr. Francisco Louzada Neto.

Várias são as aplicações de regressão logística a dados financeiros e são geralmente, vinculadas a classificação de clientes, como por exemplo *Credit Score*. A técnica é utilizada para determinar risco de crédito. Levando em consideração um modelo de regressão logística já ajustado, a probabilidade de perda, isto é, a probabilidade de um cliente não pagar o empréstimo tomado, é calculada considerando-se fatores de riscos, tais como, idade, condição sócio-econômica, histórico de inadimplência, setor de atividades, etc. e/ou fatores de riscos característicos da operação, valor total do empréstimo, prazo de pagamento, tipos de garantia (Abreu, 2004). O desenvolvimento do modelo de *Credit Score* consiste de uma forma geral, em buscar características dos clientes que estão relacionadas significativamente como seu risco de crédito. Normalmente esses modelos são desenvolvidos a partir de bases históricas de desempenho de crédito dos clientes e também de informações pertinentes ao produto.

O conjunto de dados é constituído de 7321 clientes, sendo 2227 clientes inadimplentes.

As variáveis consideradas foram: tipo de cliente, tempo de emprego, sexo, idade, estado civil, limite de crédito, tempo de residência, região e profissão.

O procedimento foi o mesmo que o realizado anteriormente. O conjunto de dados foi dividido, de forma aleatória, em dois subconjuntos: um conjunto de construção (D^C) representando 70% dos dados originais e um conjunto de teste (D^T) constituído pelos 30% restantes dos dados. Como estão sendo consideradas 10 covariáveis, existem $2^{10} = 1024$ possíveis modelos a serem ajustados.

Para o método de seleção de modelos *Stepwise*, o modelo selecionado foi:

$$y = \beta_0 + \beta_1 x1 + \beta_2 x2 + \beta_3 x3 + \beta_4 x4 + \beta_5 x5 + \beta_7 x7 + \beta_8 x8$$

Utilizando os dados de construção (D^C), os 1024 possíveis modelos foram ajustados e destes foram selecionados os modelos cujos pesos, ou probabilidades a posteriori, somavam 90% de incerteza, resultando em 909 modelos selecionados considerando abordagem clássica e 17 modelos considerando a abordagem bayesiana. Nas Tabelas 6-10 e 6-11 estão apresentados, em ordem decrescente, os AIC 's e pesos dos 10 modelos com maiores pesos os BIC's e probabilidades a posteriori dos 10 modelos com maiores probabilidades a posteriori.

Tabela 6-10: Os 10 modelos com maiores pesos

Modelo	**AIC**	**W**
1010	8,5429	0,0011
961	8,5440	0,0011
1023	8,5444	0,0011
1020	8,5449	0,0011
1007	8,5455	0,0011
1004	8,5460	0,0011
966	8,5461	0,0011
1024	8,5464	0,0011
967	8,5466	0,0011
1019	8,5475	0,0011

Tabela 6-11: Os 10 modelos com maiores probabilidades a posteriori

Modelo	**BIC**	**PostProb**
1010	-8992,209	0,2212
961	-8991,507	0,1557
1023	-8990,307	0,0855
1020	-8989,854	0,0681
1007	-8989,610	0,0603
966	-8989,474	0,0564
1004	-8989,162	0,0482
967	-8988,960	0,0436
846	-8988,150	0,0291
1024	-8987,948	0,0263

Observando-se as Tabelas 6-10 e 6-11 verifica-se que, tanto na ponderação clássica como na bayesiana, há uma considerável incerteza devido à escolha do modelo, pois o modelo com maior peso $(0,0011)$ representa apenas $0,11\%$ do peso total e o modelo com maior probabilidade a posteriori $(0,2212)$ representa $22,12\%$ da probabilidade a posteriori total, o que indica que não há nenhum modelo que poderia ser selecionado e tido como o melhor modelo para se fazer a predição que se deseja.

Utilizando apenas os modelos selecionados e o conjunto de teste (D^T) as predições foram obtidas para as duas abordagens de ponderação de modelos e para o método de seleção de modelos *Stepwise*.

As Figuras 6-12, 6-13 e 6-14 mostram as curvas ROC para cada uma das abordagens consideradas. Na Tabela 6-12 estão apresentados os pontos de corte, as medidas de capacidade preditiva, a área sob a curva e o logaritmo do score preditivo, descritos na Seção 4.3, obtidos para cada uma das abordagens consideradas.

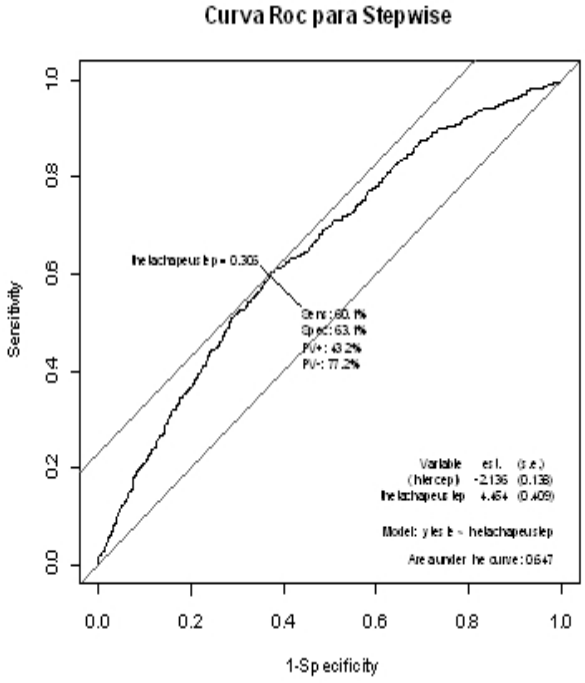

Figura 6-12: Curva ROC do método de seleção de modelos *Stepwise*.

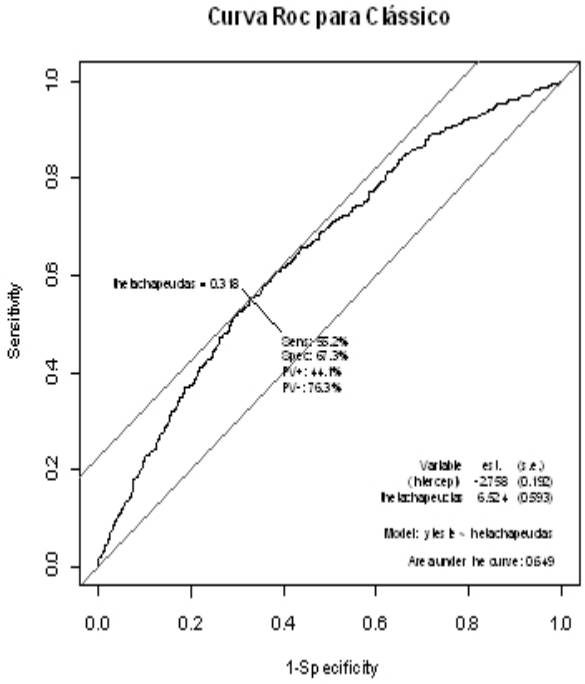

Figura 6-13: Curva ROC do método de ponderação de modelos abordagem clássica.

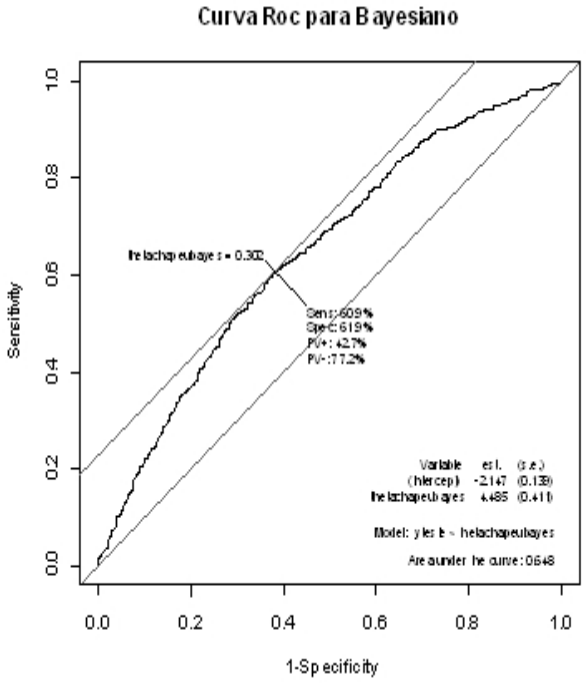

Figura 6-14: Curva ROC do método de ponderação de modelos abordagem bayesiana.

Tabela 6-18: Medidas de Capacidade Preditiva

	Pto de Corte	**CAT**	**CAU**	**CAZ**	**AUC**	**Log Score**
Stepwise	0.306	62.1	63.1	60.1	0.647	-2841.38
Clássico	0.318	63.4	67.2	55.2	0.649	-2743.37
Bayesiano	0.302	61.6	61.9	60.9	0.648	-2836.96

Pela análise das medidas de capacidade CAT, CAU, CAZ e AUC apenas para o método de ponderação de modelos abordagem clássica houve um aumento em relação ao método de seleção de modelos *Stepwise*. Já para o método de ponderação de modelos abordagem bayesiana verifica-se que estas medidas são praticamente iguais as do método de seleção de modelos *Stepwise*.

Pela análise do logaritmo do score preditivo observa-se que o método de ponderação de modelos - abordagem clássica - tem maior capacidade preditiva do que o método de seleção de modelos *Stepwise*. A melhora no score preditivo do método de ponderação de modelos

em relação ao método de seleção de modelos *Stepwise* foi de $\delta_1 = 4,42$ na bayesiana e de $\delta_2 = 98,01$ na clássica. O conjunto de teste era composto de $n_{teste} = 2196$ observações, então, o resultado significa que, em média, a probablidade preditiva do método de ponderação de modelos, tanto na abordagem clássica como na bayesiana, foi maior do que o método de seleção de modelos *Stepwise* por um fator de $\exp(\delta_1/n_{teste}) = 1,002015$, ou seja, $0,202\%$ na abordagem bayesiana e $\exp(\delta_2/n_{teste}) = 1,045642$, ou seja, $4,56\%$ na abordagem clássica. Em outras palavras, o método de ponderação de modelos prediz se um indivíduo será inadimplente $0,202\%$ melhor na abordagem bayesiana e $4,56\%$ na abordagem clássica de ponderação de modelos do que o método de seleção de modelos *Stepwise*.

Capítulo 7

Conclusão

Com o estudo de simulação apresentado na Seção 6.1.1 pode-se concluir que as abordagens de ponderação apresentadas produzem propriedades similares. Esta similaridade era esperada uma vez que as prioris e aproximações utilizadas na abordagem bayesiana a formulação final da distribuição a posteriori preditiva fica muito próxima da versão clássica. Os vícios das abordagens clássica e bayesiana são muito próximos e diminuem com o aumento do tamanho da amostra. Com relação a comparação das variâncias (Figura 6-8) verifica-se que as estimativas obtidas pela proposta de Buckland *et. al.* (1997) e do estimador BMA acompanham o comportamento das variâncias obtidas por simulação, e, por serem médias dos valores obtidos em cada simulação apresentam um comportamento suavizado nos picos. O método *bootstrap* - reamostragem dos resíduos apresenta um comportamento próximo aos anteriores. Já as estimativas obtidas para o método *boostrap* - reamostragem dos pares estratificado apresentam um comportamento diferente das outras abordagens. Apenas no **conjunto 2**, para $n = 50$, o comportamento obtido é o mesmo das outras abordagens consideradas. Pode-se concluir de forma geral, para as configurações consideradas nesta simulação, que, com o aumento do tamanho da amostra observa-se maior concordância entre as variâncias, ou seja, as abordagens produzem resultados mais próximos.

A aplicação feita na Seção 6.2 mostrou, através do logaritmo do score preditivo, que o método de ponderação de modelos aumentou em mais de 14% a capacidade preditiva em relação ao método de seleção de modelos. O desempenho observado nas outras medidas de capacidade preditiva não foi conclusivo. Os resultados do estudo de simulação mostraram

que, em média, há um aumento da capacidade preditiva do método de ponderação em relação ao método de seleção de modelos *Stepwise*. Este aumento é medido de forma mais significativa ao utilizar o logaritmo do score preditivo e foi de 31% para a abordagem clássica e de 41% para a abordagem bayesiana.

A aplicação da metodologia de ponderação de modelos em um conjunto de dados real, apresentou aumento na capacidade preditiva, através do logaritmo do score preditivo, apenas na abordagem clássica $(4,56\%)$. A abordagem bayesiana teve o mesmo desempenho obtido pelo método de seleção de modelos *Stepwise*. As outras medidas de capacidade preditiva, novamente, foram inconclusivas.

Nos exemplos de aplicação, os resultados não foram conclusivos havendo discordância entre o logaritmo do score preditivo e as medidas de capacidade preditiva CAT, CAU, CAZ e AUC. Foi interessante o resultado do estudo de simulação do exemplo da Seção 6.3 que apresentou , em média, melhoria da capacidade preditiva quando se usa o método de ponderação de modelos tanto na abordagem clássica como na bayesiana.

Um ponto observado neste trabalho foi a dificuldade de se captar o ganho na capacidade preditiva devido ao fato da necessidade de classificação da variável resposta nas categorias 0 e 1. Se for considerado o caso de regressão linear, a medida de capacidade preditiva compara a estimativa diretamente com o valor observado. Assim, toda pequena diferença fornecida pelo estimador pode ser um ganho. No caso logístico, categorizar o valor predito para 0 ou 1, faz com que se perca estes pequenos ganhos.

Apesar de não se poder generalizar os resultados aqui obtidos, pode-se dizer que esse aumento na capacidade preditiva obtida pelo método de ponderação de modelos, embora pequeno, em geral, é de grande interesse pois, sempre quando se deseja predizer um evento espera-se que a predição obtida seja a mais precisa possível. Desta forma, considerando o fato da facilidade da aplicação da técnica, salvo a conjuntos de dados com muitas covariáveis, e consequentemente muitos modelos, que tornam o procedimento lento e muitas vezes inviável, esta deve ser utilizada para se garantir uma melhora na capacidade preditiva, sempre que houver incerteza quanto à escolha de um melhor modelo.

Referências Bibliográficas

[1] ABREU, H.J. Aplicação de Análise de Sobrevicência em um problema de Credit Score e comparação com a Regressão Logística. São Carlos, 2004. 116p. Tese de Mestrado - UFSCar.

[2] AKAIKE, H. Information theory and an extension of the maximum likelihood priniple. In **Breakthroughs in Statistics**, V.1, eds. S. Kotz & N. L. Johnson, p.610-624. New York:Springer. 1973.

[3] BUCKLAND, S.T., BURNHAN, K.P. and AUGUSTIN, N.H. Model selection: An integral part of inference. **Biometrics**, V.53, p.603-618, 1997

[4] COLLETT, D. **Modelling Binary Data**, London:Chapman & Hall, 1991, 289p.

[5] CANDOLO, C. A incorporação da incerteza devido a escolha de modelos na inferência estatística com aplicação em modelos de regressão linear. Piracicaba, 2001. 80p. Tese (Doutorado) - ESALQ-USP.

[6] CANDOLO, C., SILVEIRA, R.M., Um Estudo da Incorporação da Incerteza na Seleção de Modelos em Regressão Logística. Iniciação científica, FAPESP, 2003.

[7] DAVISON, A. C., HINKLEY, D.V. **Bootstrap Methods and their Application**. Cambridge University Press, 1997. 582p.

[8] DEMÉTRIO, C.G.B. **Modelos Lineares Generalizados na Experimentação Agronômica, V SEAGRO e XXXVIII RBRAS**. Porto Alegre: DE/UFRGS, 1993, 125p.

[9] DOBSON, A.J. **An Introduction to Generalized Linear Models**, London: Chapman & Hall, 1990, 174p.

[10] DRAPER, D. Assessment and propagation of model uncertainty (with Discussion). **Journal of Royal Statistical Society**, Série B, V.57, p.45-97, 1995.

[11] EFRON, B. Bootstrap methods: another look at the jacknife. **Annals of statistics**, V.7, p.1-26, 1979.

[12] EFRON, B. & TIBSHIRANI, R. J. **An Introduction to the Bootstrap**. New York:Chapman & Hall, 1993, 436p.

[13] GOOD, I.J. Rational decisions. **Journal of Royal Statistical Society**, Série B, V.14, n.1, p.107-114, 1952.

[14] HOETING, J.A. Accounting for Model Uncertainty in Linear Regression. Seattle, 1994. 167p. Thesis (Ph.D.) - University of Washington, 1994.

[15] HOETING, J.A., MADIGAN, D., RAFTERY, A,E, & VOLINSKY, C.T. Bayesian model averaging: a tutorial (with Dicussion). **Statistical Science**, V.14, p.382-417, 1999.

[16] JEFFREYS, H. Theory of probability (3^{rd} ed.), Oxford, U.K, **Oxford University Press**, 1961.

[17] JOHNSON, R.W. Fitting percentage of body fat to simple body measurements. **Journal od Statistics Education** V.4, 1996.

[18] MADIGAN, D., ANDERSSON, S.A., PERLMAN, M. AND VOLINSKY, C.T. Bayesian model averaging and model selection for markov equivalence classes of acyclic digraphs. **Communs Statistics Theory Methodology**, 25, 2493-2520, 1996.

[19] MADIGAN, D. & RAFTERY, A.E. Model selection and accounting for model uncertainty in graphical models using Occam's window. **Journal of the American Statistical Association**, V.89, p.1535-1546, 1994.

[20] MADIGAN, D. & YORK, J. Baeysian graphical models for discrete data. **International Statistical Review**, V. 63, p.215-232, 1995.

[21] NETER, J., KUTNER, M.H., NACHTSHEIN, C.J. e WASSERMAN, W. **Applied Linear Statistical Models**. 3ed. Chicago:Irwin, 1996, 1408p.

[22] RAFTERY, A.E. Bayesian model selection in social research (with Discussion). In **Sociological Methodology**, p.111-196, 1995.

[23] RAFTERY, A.E. Approximated Bayes factor and accounting for model uncertainty in generalised linear models. **Biometrika**, V.83, p.251-266, 1996.

[24] RAFTERY, A.E., MADIGAN, D. & HOETING, J.A. Bayesian model averaging for linear regression models. **Journal of the American Statistical Association**, V.92, p.179-191, 1997.

[25] SCHWARZ, G. Estimating the dimensions of a model. **Annals os Statistics**, V.6, p.461-463, 1978.

[26] TAPLIN, R.H. Robust likehood calculation for time series. **Journal of Royal Statistical Society,** Série B, V.55, p.829-836, 1993.

[27] TAPLIN, R.H. & RAFTERY, A.E. Analisys of agricultural field trials in the presence of outliers an fertility jumps. **Biometrics**, V.50, p.764-781, 1994.

[28] TIERNEY, L. & KADANE, J.B. Accurate approximations for posterior moments and marginal densities. **Journal of the American Statistical Association**, V.81, p.82-86, 1986.

[29] VOLINSKY, C.T., MADIGAN, D., RAFTERY, A.E. and KRONMAL, R.A. Bayesian model averaging in proportional hazard models: assessing the risk os a stroke. **Applied Statistics**, V.46, n.4, p.433-448. 1997.

[30] ZWEIG, M. H., Receiver-operating characteristic (ROC) plots. Campbell, G., **Clin. Chem.**, 29, 561-577, 1993.

Apêndice A

Estimação em Modelos Lineares Generalizados

Para a construção deste texto, que apresentará a metodologia de modelos lineares generalizados, podem ser citados como referências Dobson(1990) e Demétrio (1993).

Sejam as variáveis aleatórias independentes $Y_1, ...Y_n$ com médias $\mu_1, ..., \mu_n$, isto é,

$$E(Y_i) = \mu_i \ , \ i = 1, 2, ..., n,$$

tais que Y_i tem distribuição pertencente a família exponencial com as seguintes propriedades:

1. A distribuição de cada Y_i pertence à família exponencial na forma canônica e dependem de um único parâmetro θ_i, isto é,

$$f(y_i; \theta_i) = \exp[y_i b_i(\theta_i) + c_i(\theta_i) + d_i(y_i)], \tag{A.1}$$

onde $b(.)$ e $c(.)$ são funções conhecidas e o parâmetro θ_i é chamado de parâmetro natural da família exponencial.

2. A distribuição de todos os $Y_i's$ são da mesma forma.

Assim, a função densidade de probabilidade conjunta dos $Y_i's$ pode ser expressa por

$$f(y_1, ..., y_n; \theta_1, ..., \theta_n) = \exp[\sum_{i=1}^{n} y_i b_i(\theta_i) + \sum_{i=1}^{n} c_i(\theta_i) + \sum_{i=1}^{n} d_i(y_i)]. \tag{A.2}$$

Para a especificação do modelo, os parâmetros θ_i não são de interesse direto (desde que há um para cada observação) e sim um conjunto menor de parâmetros $\beta_1, ..., \beta_p$ $(p < n)$ de tal forma que a combinação linear dos $\beta's$ seja igual a alguma função do valor esperado de Y_i, isto é,

$$g(\mu_i) = x_i^T \beta, \tag{A.3}$$

onde g é uma função monótona e diferenciável chamada de função de ligação, x_i é o vetor de variáveis explicativas de dimensão px1 e $\beta = [\beta_1, ..., \beta_p]$ o vetor de parâmetros de dimensão px1.

Assim, o modelo linear generalizado é definido por três componentes:

1. um componente aleatório representado pelas variáveis respostas $Y_i,\ i = 1, 2, ..., n$, vindas de uma mesma distribuição que faz parte da família exponencial;

2. um componente sistemático que especifica as variáveis explicativas usadas como preditoras no modelo, ou seja, um conjunto de parâmetros β e as variáveis explicativas

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix};$$

3. uma função de ligação, $g(.)$,que descreve a relação funcional entre o componente sistemático e o valor esperado (média) do componente aleatório, ou seja,

$$g(\mu_i) = x_i^T \beta, \tag{A.4}$$

onde

$$\mu_i = E(Y_i). \tag{A.5}$$

Além disso, a variância é uma função explicita da média μ,

$$Var(Y_i) = \phi V(\mu), \tag{A.6}$$

onde $V(\cdot)$ é uma função de variância conhecida e ϕ o parâmetro de dispersão, que geralmente é desconhecido.

A partir da definição de um modelo linear generalizado, obtêm-se o estimador de

máxima verossimilhança dos parâmetros β no ajuste do modelo. O logaritmo da função de verossimilhança (A.2) é dado por

$$l(\theta, y) = \sum_{i=1}^{n} y_i b_i(\theta_i) + \sum_{i=1}^{n} c_i(\theta_i) + \sum_{i=1}^{n} d_i(y_i), \tag{A.7}$$

e, a média e a variância dos $Y_i's$ são dadas, respectivamente, por

$$E(Y_i) = \mu_i = -c'(\theta_i)/b'(\theta_i), \tag{A.8}$$

e

$$Var(Y_i) = [b'(\theta_i)c'(\theta_i) - c'(\theta_i)b'(\theta_i)]/[b'(\theta_i)]^3. \tag{A.9}$$

A função de ligação pode ser reescrita como

$$g(\mu_i) = x_i^T \beta = \eta_i. \tag{A.10}$$

Uma propriedade da família exponencial é que ela satisfaz as condições de regularidade para se encontrar um máximo global do logaritmo da função de verossimilhança, e que é obtido unicamente pela solução do sistema de equações $\frac{\partial l}{\partial \beta} = 0$.

Dobson (1990) mostra que

$$U_j = \frac{\partial l(\theta; y)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j}, \tag{A.11}$$

onde

$$l_i = y_i b_i(\theta_i) + c_i(\theta_i) + d_i(y_i). \tag{A.12}$$

Para obter U_j utilizamos a relação

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} \tag{A.13}$$

Diferenciando (A.12) e substituindo em (A.8) obtemos

$$\frac{\partial l_i}{\partial \theta_i} = y_i b'(\theta_i) + c'(\theta_i) = b'(\theta_i)(y_i - \mu_i). \tag{A.14}$$

Diferenciando (A.8) e substituindo em (A.9) tem-se

$$\frac{\partial \mu_i}{\partial \theta_i} = -\frac{c'(\theta_i)}{b'(\theta_i)} + \frac{c'(\theta_i)b''(\theta_i)}{[b'(\theta_i)]^2} = b''(\theta_i)var(Y_i), \tag{A.15}$$

e diferenciando (A.10)

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}\frac{\partial \mu_i}{\partial \eta_i}. \tag{A.16}$$

Então,

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \mu_i}{\partial \beta_j}/\frac{\partial \mu_i}{\partial \theta_i} = \frac{(y_i - \mu_i)x_{ij}}{var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right). \tag{A.17}$$

Assim,

$$\frac{\partial l}{\partial \beta_j} = U_j = \sum_{i=1}^{N}\frac{(y_i - \mu_i)x_{ij}}{var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right). \tag{A.18}$$

As equações $U_j = 0$ são não lineares e devem ser resolvidas por iteração numérica. Utilizando o método de Newton-Raphson a m-ésima aproximação é dada por

$$b^{(m)} = b^{(m-1)} - \left[\frac{\partial^2 l(\theta; y)}{\partial \beta_j \partial \beta_k}\right]^{-1}_{\beta = b^{(m-1)}} U^{(m-1)}, \tag{A.19}$$

onde $\left[\frac{\partial^2 l(\theta;y)}{\partial \beta_j \partial \beta_k}\right]_{\beta = b^{(m-1)}}$ é a matriz da segunda derivada de l, calculada em $\beta = b^{(m-1)}$ e $U^{(m-1)}$ é o vetor das primeiras derivadas $U_j = \frac{\partial l(\theta;y)}{\partial \beta_j}$, calculada em $\beta = b^{(m-1)}$.

Um procedimento alternativo ao método de Newton-Raphson é o método Score. Este consiste em substituir a matriz das segundas derivadas pela matriz dos valores esperados

$$E\left[\frac{\partial^2 l(\theta; y)}{\partial \beta_j \partial \beta_k}\right], \tag{A.20}$$

que é igual à matriz negativa de variância-covariância dos U_j's. A matriz de informação $\Im = E[UU^T]$ é formada pelos elementos

$$\Im_{jk} = E[U_j U_k] = E\left[\frac{\partial l}{\partial \beta_j}\frac{\partial l}{\partial \beta_k}\right] = -E\left[\frac{\partial^2 l(\theta; y)}{\partial \beta_j \partial \beta_k}\right]. \tag{A.21}$$

Assim, a equação (A.19) pode ser substituída por

$$b^{(m)} = b^{(m-1)} + [\Im^{(m-1)}]^{-1}U^{(m-1)}, \tag{A.22}$$

onde $\Im^{(m-1)}$ é a matriz de informação calculada em $b^{(m-1)}$. Multiplicando-se ambos os

lados da equação (A.22) por $\Im^{(m-1)}$ obtêm-se

$$\Im^{(m-1)}b^{(m)} = \Im^{(m-1)}b^{(m-1)} + U^{(m-1)}. \tag{A.23}$$

Dado que os elementos da matriz de informação são definidos por $\Im_{jk} = E[U_j U_k]$, então,

$$E\left[\frac{\partial l_i}{\partial \beta_j}\frac{\partial l_i}{\partial \beta_k}\right] = E\left[\frac{(y_i-\mu_i)^2 x_{ij}}{\{var(Y_i)\}^2}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right] = \frac{x_{ij}x_{ik}}{var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2. \tag{A.24}$$

Portanto o $(j,k)-$ésimo elemento de $\Im$ é

$$\Im_{jk} = \sum_{i=1}^{N}\frac{x_{ij}x_{ik}}{var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2. \tag{A.25}$$

Sendo assim, $\Im$ pode ser escrito como $\Im = X^T W X$,onde W é uma matriz diagonal nxn formada pelos elementos

$$w_{ii} = \frac{1}{var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2. \tag{A.26}$$

O lado direito da expressão (A.23) é um vetor com elementos

$$\sum_k\sum_i \frac{x_{ij}x_{ik}}{var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 b_k^{(m-1)} + \sum_i \frac{(y_i-\mu_i)x_{ij}}{var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right) \tag{A.27}$$

calculados em $b^{(m-1)}$. Então, o lado direito da expressão (A.23) pode ser reescrito como $X^T W z$,onde z tem os elementos $z_i = \sum_k x_{ij} b_k^{(m-1)} + (y_i - \mu_i)\left(\frac{\partial \mu_i}{\partial \eta_i}\right)$,com μ_i e $\frac{\partial \mu_i}{\partial \eta_i}$ avaliados em $b^{(m-1)}$.

Por fim, a equação iterativa para o método Score, (A.23), pode ser escrita como

$$X^T W X b^{(m)} = X^T W z. \tag{A.28}$$

A equação iterativa para o método de Score tem a mesma forma das equações normais para modelos lineares obtidos por mínimos quadrados ponderados. Então, para modelos lineares generalizados os estimadores de máxima verossimilhança são obtidos por um procedimento iterativo de mínimos quadrados ponderados, chamado de mínimos quadrados ponderados iterativamente.

Apêndice B

Programas Desenvolvidos para as Aplicações

1. Função que faz os cálculos da ponderação clássica de modelos

```
# Seleção dos melhores modelos
#matriz com os resultados dos 8192 ajustes: indice do modelo, aic
resultclas<-matrix(0,ncol=2,nrow=8192)
dimnames(resultclas)<-list(NULL,c("MODEL","AIC"))
## ajuste do modelo so com bo
ajuste1<-glm(y~1, data=dados, family=binomial(link=logit))
resultclas[1,1]<-1
resultclas[1,2]<-ajuste1$aic
#ajuste do 8191 modelos restantes
for (i in 2:8192){
xsel<-x[,XM[i,]]
yx<-data.frame(y=y,xsel)
ajuste1<-glm(y~.,data=yx,family=binomial(link=logit))
resultclas[i,1]<-i
resultclas[i,2]<-ajuste1$aic
}
#Calculo dos W
w<-exp(-resultclas[,2]/2)/sum(exp(-resultclas[,2]/2))
```

```
resultclas<-cbind(resultclas,w)
dimnames(resultclas)<-list(NULL,c("MODEL","AIC","W"))
resultclas[1:5,]
#ordenaçao pelo W
resultclasordenado<-resultclas[order(resultclas[,3],resultclas[,1]),1:3]
round(resultclasordenado[8192:8172,],digits=4)
XM[resultclasordenado[8192:8172,1],]
#calculo de w acumulados
aux<-cumsum(resultclasordenado[,3])
#escolha dos modelos que acumulam os maiores 90% dos w's
numclas<-length(aux[aux>0.1])
resultclasordenado2<-cbind(resultclasordenado,aux)
dimnames(resultclasordenado2)<-list(NULL,c("MODEL","AIC","W","WCUM"))
index<-resultclasordenado2[resultclasordenado2[,4]>0.1,1]
#matrix de modelos com ostop models
XMtopclas<-XM[index,]
# Cálculo das predições usando os modelos selecionados
#matriz com os resultados dos top model
resultclas2<-matrix(0,ncol=(length(yteste)+2),nrow=numclas)
for(i in 1:numclas){
x<-xconst[,XMtopclas[i,]]
yx<-data.frame(y=yconst,x)
x<-xteste[,XMtopclas[i,]]
xfinal<-data.frame(x)
ajuste1<-glm(y~.,data=yx,family=binomial(link=logit))
pred<-predict.glm(ajuste1, newdata=xfinal,se.fit=T,type="response")
resultclas2[i,1]<-i
resultclas2[i,2]<-ajuste1$aic
resultclas2[i,3:(length(yteste)+2)]<-pred$fit
}
#Calculo dos W
```

```
w2<-exp(-resultclas2[,2]/2)/sum(exp(-resultclas2[,2]/2))
#Calculo do Thetachapeu
thetachapeuclas<-t(as.matrix(w2))%*%resultclas2[,3:(length(yteste)+2)]
```

2. Função que faz o cálculo da ponderação de modelos abordagem bayesiana

```
# Seleção dos melhores modelos
prior.weight.denom<-0.5^13
#matriz com os resultados dos 8192 ajustes: indice do modelo, aic
resultbayes<-matrix(0,ncol=4,nrow=8192)
dimnames(resultbayes)<-list(NULL,c("MODEL","DEVIANCE","DF","BIC"))
## ajuste do modelo so com bo
ajuste1<-glm(y~1, data=dados, family=binomial(link=logit))
resultbayes[1,1]<-1
resultbayes[1,2]<-ajuste1$deviance
resultbayes[1,3]<-ajuste1$df.residual
resultbayes[1,4]<-resultbayes[1,2]-resultbayes[1,3]*log(length(dados))-
2*log(prior.weight.denom)
#ajuste do 8191 modelos restantes
for (i in 2:8192){
xsel<-x[,XM[i,]]
yx<-data.frame(y=y,xsel)
ajuste1<-glm(y~.,data=yx, family=binomial(link=logit))
resultbayes[i,1]<-i
resultbayes[i,2]<-ajuste1$deviance
resultbayes[i,3]<-ajuste1$df.residual
resultbayes[i,4]<-resultbayes[i,2]-resultbayes[i,3]*log(length(dados))-
2*log(prior.weight.denom)
}
#Calculo das probabilidades a posteriori
postprob<-exp(-0.5*(resultbayes[,4]-min(resultbayes[,4]))) /
sum(exp(-0.5*(resultbayes[,4]-min(resultbayes[,4]))))
```

```
resultbayes<-cbind(resultbayes,postprob)
dimnames(resultbayes)<-list(NULL,c("MODEL","DEVIANCE","DF",
"BIC","POSTPROB"))
resultbayes[1:5,]
#ordenaçao pela postprob
resultbayesordenado<-resultbayes[order(resultbayes[,5],resultbayes[,1]),1:5]
round(resultbayesordenado[8192:8172,],digits=4)
XM[resultbayesordenado[8192:8172,1],]
#calculo das postprob acumulados
aux<-cumsum(resultbayesordenado[,5])
#escolha dos modelos que acumulam os maiores 90% das post prob
numbayes<-length(aux[aux>0.1])
numbayes
resultbayesordenado2<-cbind(resultbayesordenado,aux)
dimnames(resultbayesordenado2)<-list(NULL,c("MODEL","DEVIANCE","DF",
"BIC","PPOST","PPOSTCUM"))
index<-resultbayesordenado2[resultbayesordenado2[,6]>0.1,1]
#matrix de modelos com top models
XMtopbayes<-XM[index,]
# Cálculo das predições usando os modelos selecionados
#matriz com os resultados dos top model
resultbayes2<-matrix(0,ncol=(length(yteste)+4),nrow=numbayes)
for(i in 1:numbayes){
x<-xconst[,XMtopbayes[i,]]
yx<-data.frame(y=yconst,x)
x<-xteste[,XMtopbayes[i,]]
xfinal<-data.frame(x)
ajuste1<-glm(y~.,data=yx,family=binomial(link=logit))
pred<-predict.glm(ajuste1, newdata=xfinal,se.fit=T,type="response")
resultbayes2[i,1]<-i
resultbayes2[i,2]<-ajuste1$deviance
```

```
resultbayes2[i,3]<-ajuste1$df.residual
resultbayes2[i,4]<-resultbayes2[i,2]-resultbayes2[i,3]*log(length(dadosconst))-
2*log(prior.weight.denom)
resultbayes2[i,5:(length(yteste)+4)]<-pred$fit
}
#Calculo das probabilidades a posteriori
postprob2<-exp(-0.5*(resultbayes2[,4]-min(resultbayes2[,4]))) /
sum(exp(-0.5*(resultbayes2[,4]-min(resultbayes2[,4]))))
#Calculo do Thetachapeu
thetachapeubayes<-t(as.matrix(postprob2))
%*%resultbayes2[,5:(length(yteste)+4)]
```

3. Cálculo da predição utilizando o método de seleção de modelos Stepwise

```
ajuste<-glm(y~.,data=dados, family=binomial(link=logit))
step(ajuste)
ajustestep<-glm(y~x4 + x6 + x11 + x12 + x13 ,data=dadosconst,
family=binomial(link=logit))
pred<-predict.glm(ajustestep, newdata=dadosteste,se.fit=T,type="response")
thetachapeustep<-pred$fit
```

4. Função bootstrap utilizada para o estudo de simulação

```
# função a ser utilizada no bootstrap
calculos<-function(nomedf){
#intercepto
ajuste1<-glm(y~1, data=nomedf, family=binomial(link=logit))
aic1<-ajuste1$aic
p1<-predict.glm(ajuste1,type="response")
#x1
ajuste2<-glm(y~x1, data=nomedf, family=binomial(link=logit))
aic2<-ajuste2$aic
```

```
p2<-predict.glm(ajuste2,type="response")
#x2
ajuste3<-glm(y~x2, data=nomedf, family=binomial(link=logit))
aic3<-ajuste3$aic
p3<-predict.glm(ajuste3,type="response")
#x1 e x2
comp<-glm(y~.,data=nomedf, family=binomial(link=logit))
aic4<-comp$aic
p4<-predict.glm(comp,type="response")
w1<-exp(-aic1/2)/(exp(-aic1/2)+exp(-aic2/2)+exp(-aic3/2)+exp(-aic4/2))
w2<-exp(-aic2/2)/(exp(-aic1/2)+exp(-aic2/2)+exp(-aic3/2)+exp(-aic4/2))
w3<-exp(-aic3/2)/(exp(-aic1/2)+exp(-aic2/2)+exp(-aic3/2)+exp(-aic4/2))
w4<-exp(-aic4/2)/(exp(-aic1/2)+exp(-aic2/2)+exp(-aic3/2)+exp(-aic4/2))
th<-w1*p1 + w2*p2 + w3*p3 + w4*p4
th
}
#funcao do bootstrap residuos (mod completo)
bootfun1e2<-function(data,i) {
y<-data$fit + sqrt(data$fit*(1-data$fit))*data$pearson[i]
y[y < 0]<-0
y[y > 1]<-1
y[y > 0.5]<-1
y[y < 0.5]<-0
data$y<-y
calculos(data) }
#funcao bootstrap pares
bootfun4<-function(data,i) {
calculos(data[i,]) }
```

5. Gráficos das Curvas ROC

```
m4<-list(thetachapeustep,yteste)
names(m4)<-list("Preditos","Original")
```

```
pred1 <- prediction(m4$Preditos, m4$Original)
perf1<- performance(pred1,"tpr","fpr")
ROC(thetachapeustep,yteste,plot="ROC",PV=TRUE,AUC=TRUE,MX=TRUE,
main="Curva ROC Clássico")
m4<-list(thetachapeuclas,yteste)
names(m4)<-list("Preditos","Original")
pred2 <- prediction(m4$Preditos, m4$Original)
perf2<- performance(pred2,"tpr","fpr")
ROC(thetachapeuclas,yteste,plot="ROC",PV=TRUE,AUC=TRUE,MX=TRUE,
main="Curva ROC Stepwise")
names(m4)<-list("Preditos","Original")
pred3 <- prediction(m4$Preditos, m4$Original)
perf3<- performance(pred3,"tpr","fpr")
ROC(thetachapeubayes,yteste,plot="ROC",PV=TRUE,AUC=TRUE,MX=TRUE,
main="Curva ROC Bayesiano")
```

6. Classificação das predições e cálculos das medidas de capacidade preditiva

```
vetor<-rep(0,length(yteste))
vetor[thetachapeustep<0.451]<-0
vetor[thetachapeustep>0.451]<-1
tabstep<-table(vetor,yteste)
tabstep
A<-sum(tabstep[,1])
B<-sum(tabstep[,2])
a<-sum(tabstep[1,])
b<-sum(tabstep[2,])
n<-sum(tabstep)
cat<-(tabstep[2,2]+tabstep[1,1])/n
round(cat,digits=3)
caz<-tabstep[1,1]/A
round(caz,digits=3)
```

```
cau<-tabstep[2,2]/B
round(cau,digits=3)
logclas<-sum(log(thetachapeuclas))
logstep<-sum(log(thetachapeustep))
logbayes<-sum(log(thetachapeubayes))
```