Universidade Federal de Minas Gerais
Departamento de Estatística
Programa de Pós-Graduação

O MÉTODO LASSO PARA O MODELO DE COX E SUA COMPARAÇÃO COM PROPOSTAS TRADICIONAIS DE SELEÇÃO DE VARIÁVEIS

Carolina Gontijo Miranda
carolinagontijo@ufmg.br

Enrico Antônio Colosimo
enricoc@est.ufmg.br

Rosângela Helena Loschi
loschi@est.ufmg.br

Belo Horizonte, junho de 2006

Agradecimentos

Agradecimento maior aos meus orientadores Enrico Colosimo e Rosângela Loschi pelo projeto proposto, a confiança depositada, a atenção dispensada, o acompanhamento contínuo, a amizade e as horas gastas analisando cada detalhe desse trabalho.

Agradecimentos ao professor Marcelo Azevedo pela disponibilidade, presteza e pelo auxílio inestimável na programação do método LASSO. Aos professores Suely Giolo, Sueli Mingoti e Marcelo Azevedo pelas ótimas sugestões e correções apresentadas e pela atenção dispensada na avaliação dessa dissertação.

À Telemar por disponibilizar os dados utilizados na aplicação real.

À minha família, pelo amor e apoio incondicionais durante esses anos de estudo. À mamãe, amiga e iniciadora nas artes da Estatística. Ao vovô pelas preces, pelos colos, pelo carinho e doçura inspiradores.

Agradecimentos a todos os amigos que acreditaram nessa conquista e que souberam entender os momentos de ausência e principalmente a Deus, por proporcionar-me cumprir mais esta jornada.

Resumo

Neste trabalho utilizou-se o método LASSO *(Least Absolute Shrinkage and Selection Operator)* para seleção de variáveis no modelo de Cox assim como alguns outros métodos sugeridos na literatura, a saber, o método passo-a-passo e os critérios AIC *(Akaike Information Criterion)* e BIC *(Bayesian Information Criterion)*. O método LASSO é comparado com tais métodos através de simulações Monte Carlo e seu uso é ilustrado em problemas reais envolvendo o modelo de Cox. Um algorítmo alternativo foi utilizado para implementar o LASSO.

O método LASSO é um método de seleção de variáveis inicialmente formulado para os modelos lineares. No contexto do modelo de Cox, o método LASSO procura maximizar o logaritmo da verossimilhança parcial, sujeito à restrição de que a soma dos valores absolutos dos coeficientes sejam menores do que uma constante pré-especificada. Devido à natureza da restrição, o método pode gerar valores exatamente nulos para alguns coeficientes.

Simulações demonstram que, em geral, o método LASSO produz resultados melhores que o método passo-a-passo e critérios de seleção de modelos como o AIC e BIC.

Palavras-chave: AIC, BIC, Método Passo-a-Passo, Verossimilhança Parcial.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Uma análise cuidadosa de dados deve sempre considerar o problema de determinação do modelo, isto é, o problema da avaliação e escolha do modelo que melhor represente a situação em estudo. *"A escolha do melhor modelo não pode ser desvinculada do objetivo do estudo, mais especificamente, do uso que se pretende dar ao modelo, e da classe de modelos que estão em competição"* (Paulino *et al.*, 2003, pag. 348)

Muitos autores têm estudado a questão de seleção de modelos e novos métodos de selecionar o modelo mais parcimonioso têm sido sugeridos na literatura. Kadane e Lazar (2004), por exemplo, sugerem um método bayesiano baseado na comparação das probabilidades *a posteriori* de cada modelo. Faraggi e Simon (1998) apresentam um método bayesiano de seleção de variáveis para dados de sobrevivência censurados no qual a escolha do modelo mais parcimonioso é baseada na comparação da função de perda e percentual do erro explicado por cada um dos modelos em competição. Tibshirani (1994) propõe um método que procura maximizar o logaritmo da verossimilhança parcial, sujeito à restrição de que a soma dos valores absolutos dos coeficientes sejam menores do que uma constante pré-especificada e Breiman (1995) apresenta o método GARROTE que é baseado nos métodos *subset selection* e *ridge regression*.

Na área Médica, por exemplo, a maioria dos registros de fatores preditivos para determinadas doenças são constantemente atualizados. A ocorrência de um aumento repentino dos fatores de risco para alguma patologia pode ser um indício de um aumento no número de indivíduos doentes. Neste caso, o ideal seria um sistema de

seleção de variáveis capaz de identificar, com alta precisão, quais seriam estes grupos de covariáveis que deveriam ser monitoradas a fim de detectar este novo padrão o mais rápido possível, permitindo a adoção de ações preventivas adequadas. Esta rápida detecção seria benéfica tanto para os indivíduos quanto para a sociedade, no sentido de reduzir despesas com medicamentos ou evitar que a doença se agrave, devido ao diagnóstico tardio.

Já na área de Ciências Sociais, muitas vezes, pode-se estar interessado em identificar quais variáveis comportamentais podem vir a explicar determinado tempo de falha. Por exemplo, o setor de cobrança de uma empresa tem interesse nas variáveis preditoras para o tempo até a ocorrência da retirada por inadimplência, visando tomar medidas preventivas para evitar que um cliente se torne inadimplente e corretivas para tentar diminuir a quantidade de inadimplência na sua carteira de clientes.

Em análise de sobrevivência, o modelo de regressão de Cox é o mais utilizado devido a sua versatilidade. No modelo de Cox, assim como em outros modelos, métodos de seleção de variáveis são particularmente importantes, pois relacionam quais das muitas covariáveis mensuradas podem ser tomadas como preditoras. Assim, estes métodos, sob a perspectiva do modelo de Cox, objetivam identificar quais grupos de covariáveis explicam melhor o tempo até a ocorrência do evento de interesse, denominado tempo de falha. A principal característica de dados de sobrevivência é a presença de censura, que é a observação parcial da resposta. Isto se refere a situações em que, por algum motivo, o acompanhamento da observação foi interrompido. Isto significa que toda informação referente à resposta se resume ao conhecimento de que o tempo de falha é superior àquele observado.

O termo análise de sobrevivência refere-se basicamente a situações médicas envolvendo dados censurados. Entretanto, condições similares ocorrem em outras áreas em que se usam as mesmas técnicas de análise de dados. Em engenharia, são comuns os estudos em que produtos ou componentes são colocados sob teste para se estimar características relacionadas aos seus tempos de vida, tais como o tempo médio ou a probabilidade de um certo produto durar mais do que 5 anos. O mesmo acontece em ciências socias, em que várias situações de interesse têm como resposta o tempo entre eventos.

Este trabalho foi motivado por um problema real na área de Telefonia e outro

na área Médica. Sua principal contribuição é entender e implementar um algoritmo alternativo para o método LASSO *(Least Absolute Shrinkage and Selection Operator)* para seleção de variáveis no modelo de Cox, visto que o disponível na literatura (Tibshirani, 1996) se mostra ineficiente para tamanhos amostrais maiores que 20, devido a enorme quantidade de operações matriciais envolvidas, em particular, a inversão de matrizes. Neste estudo também são implementados o método passo-a-passo e os critérios de seleção de modelos AIC *(Akaike Information Criterion)* e BIC *(Bayesian Information Criterion)*. O método LASSO é comparado com estes procedimentos através de um estudo Monte Carlo. Também utilizou-se estes métodos e critérios para determinação do "melhor"modelo em duas situações reais, a saber, identificação dos fatores preditivos para a morte devido à cirrose e fatores determinantes para a retirada do telefone do cliente no caso em que o mesmo se encontrava inadimplente junto à operadora de telefonia.

Este trabalho está assim organizado. No Capítulo 2 são apresentados, brevemente, o modelo de regressão de Cox e o método de máxima verossimilhança parcial para a realização de inferência sobre os parâmetros do modelo. O Capítulo 3 apresenta detalhadamente o método LASSO para seleção de variáveis e alguns outros métodos e critérios de seleção de modelos comumente utilizados para seleção de variáveis que serão posteriormente comparados ao método LASSO. O Capítulo 4 apresenta um estudo de simulação comparando o método LASSO com tais métodos e critérios. No Capítulo 5, encontram-se dois estudos de casos reais. Finalmente, no Capítulo 6, são apresentadas as considerações finais e propostas de trabalhos futuros.

Capítulo 2

Modelo de Regressão de Cox

O modelo de regressão de Cox abriu uma nova fase na modelagem de dados de tempos de vida. Uma evidência quantitativa deste fato pode ser encontrada em Stigler (1994). O autor usa citações feitas a periódicos indexados de todas as áreas, entre os anos de 1987 e 1989, para quantificar a importância de algumas publicações na literatura estatística. O artigo de Cox (1972), em que o modelo é apresentado, foi neste período o segundo artigo mais citado.

O objetivo deste capítulo é apresentar este importante modelo para a análise de dados de sobrevivência. O modelo de regressão de Cox, assim como os modelos de tempos de vida acelerado, permitem a análise de dados provenientes de estudo de tempos de vida em que a resposta é o tempo até a ocorrência de um evento de interesse.

A principal razão da popularidade do modelo de regressão de Cox é a presença de um componente não paramétrico, que o torna bastante flexível. Alguns resultados importantes sobre o modelo de Cox podem ser encontrados, por exemplo, em Cox e Hinkley (1974). Aplicações recentes deste modelo podem ser vistas, por exemplo, em Colosimo e Giolo (2006).

Este capítulo está assim organizado: A Seção 2.1 apresenta alguns conceitos básicos em análise de sobrevivência. A Seção 2.2 apresenta o modelo de Cox e o método da máxima verossimilhança parcial.

2.1 Preliminares

Uma das principais funções probabilísticas usadas para descrever estudos de sobrevivência, é a função de sobrevivência que é definida como sendo a probabilidade de um elemento não falhar até o tempo t, ou equivalentemente, como sendo a probabilidade de um elemento sobreviver ao tempo t. Em termos probabilísticos, isto é escrito como:

$$S(t) = P(T \geq t), \tag{2.1}$$

em que T denota a variável aleatória resposta, tempo até a falha.

A probabilidade da falha ocorrer em um intervalo de tempo $[t_1, t_2)$ pode ser expressa em termos da função de sobrevivência como sendo:

$$S(t_1) - S(t_2). \tag{2.2}$$

No entanto, a função de taxa de falha é mais usada para a modelagem de dados de sobrevivência.

A taxa de falha no intervalo $[t_1, t_2)$ é definida como a probabilidade de que a falha ocorra neste intervalo dado que não ocorreu antes de t_1, dividida pelo comprimento do intervalo. De forma geral, redefinindo o intervalo como $[t, t + \triangle t)$ e assumindo $\triangle t$ bem pequeno, tem-se que, a função taxa de falha $\lambda(t)$ é definida como:

$$\lambda(t) = \lim_{\triangle t \to 0} \frac{P(t \leq T < t + \triangle t \mid T \geq t)}{\triangle t}. \tag{2.3}$$

Observe que as taxas de falha são números positivos sem limite superior. A função de taxa de falha $\lambda(t)$ é bastante útil para descrever a distribuição do tempo de vida de pacientes, de equipamentos eletrônicos, dentre outros.

Como pode ser observado na Figura 2.1, a função de taxa de falha pode apresentar várias formas. Por exemplo, a função crescente indica que a taxa de falha do paciente ou do equipamento aumenta com o transcorrer do tempo. Este comportamento mostra um efeito gradual do envelhecimento ou desgaste. A funçao constante indica que a taxa de falha não se altera com o passar do tempo e, a função decrescente, mostra que a taxa de falha diminui à medida que o tempo passa. Ver detalhes

em Colosimo e Giolo (2006).

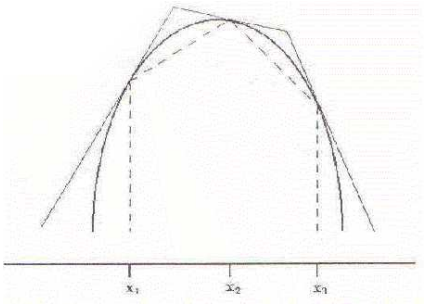

Figura 2.1: Função de Taxa de Falha.

2.2 O Modelo de Cox

Considere dados usuais de sobrevivência, isto é, assuma $(t_1, \mathbf{x}_1, \delta_1), \ldots, (t_n, \mathbf{x}_n, \delta_n)$, em que t_j denota o tempo de sobrevivência do *j-ésimo* elemento, n é o número total de elementos na amostra, δ_j é a função indicadora de falha para o *j-ésimo* elemento, em que $\delta_j = 1$ indica que observou-se uma falha para o elemento j e $\delta_j = 0$ indica a ocorrência de uma censura à direita (o tempo de ocorrência do evento de interesse está à direita do tempo registrado) e $\mathbf{x}$ denota a matriz usual de covariáveis $(x_{j1}, x_{j2}, \ldots, x_{jp})'$. Note que se $\delta_j = 1$, t_j fornece uma informação completa do tempo de ocorrência do evento de interesse e se $\delta_j = 0$, t_j fornece uma informação incompleta, ou seja, toda a informação obtida sobre este elemento é que o seu tempo até a ocorrência do evento de interesse é superior ao tempo registrado até o último acompanhamento.

Considere d_j o número de falhas no tempo t_j.

O modelo de riscos proporcionais para dados censurados, também conhecido como modelo de Cox, modela o comportamento da função de risco no tempo t, dado os valores das covariáveis $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$ da seguinte forma

$$\lambda(t|\mathbf{x}) = \lambda_0(t) \exp\left(\sum_{l=1}^{p} x_l \beta_l\right), \tag{2.4}$$

em que $\lambda_0(t)$ é uma função de risco de base arbitrária e $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p), \beta_i \in \mathbb{R}, i = 1, ..., p$, é o vetor de coeficientes do modelo.

Note, que o modelo de Cox é composto pelo produto de dois componentes, um não paramétrico e outro paramétrico. O componente não paramétrico, $\lambda_0(t)$, não é especificado e é uma função não-negativa do tempo. Ele é usualmente chamado de função de risco de base, pois $\lambda(t|\mathbf{x}) = \lambda_0(t)$ quando $\mathbf{x} = 0$. O componente paramétrico é, frequentemente, usado na forma exponencial, o que garante que $\lambda(t|\mathbf{x})$ será sempre positiva. Este modelo é chamado de modelo de riscos proporcionais, pois a razão de taxas de falha de dois elementos diferentes é constante ao longo do tempo.

Assumindo que o mecanismo de censura é não informativo, ou seja, que a distribuição dos dados censurados não agregam informações ao modelo, uma estimativa usual dos parâmetros $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$, sem a especificação de $\lambda_0(t)$, é obtida maximizando-se a função de verossimilhança parcial, através do método iterativo de Newton-Raphson (Cox, 1972) que, no caso em que não há empate nos tempos de falha, é dada por:

$$L(\boldsymbol{\beta} \mid \mathbf{x}) = \prod_{j=1}^{n} \left[\frac{\exp(x_j' \boldsymbol{\beta})}{\sum_{l \in C_j} \exp(x_l' \boldsymbol{\beta})} \right]^{\delta_j}, \tag{2.5}$$

em que C_j é o conjunto dos índices dos elementos sob risco no tempo t_j. No caso em que há empates nos tempos de falha, a expressão (2.5) sofre algumas modificações, ver Hosmer e Lemeshow (1989).

Nos métodos clássicos passo-a-passo de seleção de modelos, que serão descritos no próximo capítulo, usa-se, muito frequentemente, o teste da razão de verossimilhança parcial. A meta é, através da comparação de modelos encaixados, via teste de hipóteses, decidir qual deles "melhor"explica o comportamento dos dados amostrais. Na construção deste teste, utiliza-se a estatística da razão de verossimilhanças para modelos encaixados (Cox e Hinkley, 1974). Isto significa que deve ser identificado um modelo completo (com todas as covariáveis e sem interações) tal que os modelos restritos sejam casos particulares deste. O teste é realizado a partir de dois ajustes: (1) modelo completo e obtenção do valor do logaritmo de sua função de verossimilhança parcial ($\log L(\hat{\beta}_C)$); (2) modelo restrito e obtenção do valor do logaritmo de sua função de verossimilhança parcial ($\log L(\hat{\beta}_R)$), em que $\hat{\beta}$ são os estimadores de máxima verossimilhança sob cada modelo. A partir destes valores é possível calcular

a estatística para o teste da razão de verossimilhanças (TRV), isto é:

$$TRV = -2\log\left[\frac{L(\hat{\beta}_R)}{L(\hat{\beta}_C)}\right] = 2[\log L(\hat{\beta}_C) - \log L(\hat{\beta}_R)], \tag{2.6}$$

que, sob H_0, tem uma distribuição aproximadamente qui-quadrado com os graus de liberdade igual a diferença do número de parâmetros dos modelos que estão sendo comparados Hosmer e Lemeshow (1989).

No Capítulo 3, são apresentados alguns métodos de seleção de variáveis e alguns critérios de comparação de modelos que serão, posteriormente, comparados (Capítulo 4) na seleção de variáveis para o modelo de Cox.

Capítulo 3

Métodos de Seleção de Variáveis no Modelo de Cox

3.1 Conceitos Gerais

A escolha do "melhor modelo"é um tópico extremamente importante na análise de dados de tempo de vida. Busca-se o modelo mais parcimonioso, isto é, o modelo que envolva o mínimo de parâmetros possíveis a serem estimados e que explique bem o comportamento da variável resposta. Existem vários métodos de seleção de variáveis na literatura. Kadane e Lazar (2004), por exemplo, sugerem um método bayesiano baseado na comparação das probabilidades *a posteriori* de cada modelo. Faraggi e Simon (1998) apresentam um método bayesiano de seleção de variáveis para dados de sobrevivência censurados no qual a escolha do modelo mais parcimonioso é baseada na comparação da função de perda e percentual do erro explicado por cada um dos modelos em competição. Breiman (1995) apresenta o método GARROTE que é baseado nos métodos *subset selection* e *ridge regression*.

Como visto no Capítulo 2, o modelo de regressão de Cox é indexado pelos coeficientes $\boldsymbol{\beta}$ que medem os efeitos das covariáveis na função de taxa de falha. Neste capítulo serão apresentados alguns métodos de seleção de variáveis e critérios de com-

paração de modelos, aplicados ao modelo de Cox, dando atenção especial ao método LASSO *(Least Absolute Shrinkage and Selection Operator)*. O objetivo do método LASSO para seleção de variáveis no Modelo de Cox, como visto em Tibshirani (1997), é selecionar da forma mais eficiente possível, o sub-conjunto de covariáveis que melhor explica o tempo até a ocorrência do evento de interesse. Alguns outros métodos de seleção de variáveis e critérios de seleção de modelos serão apresentados. Organiza-se este capítulo da seguinte forma. Na Seção 3.2 apresentam-se os métodos clássicos do tipo passo-a-passo. Na Seção 3.3 descreve-se detalhadamente o método LASSO. Finalmente, na Seção 3.4 são apresentados alguns critérios de seleção do modelo.

3.2 Métodos Clássicos do Tipo Passo-a-Passo

Nesta seção serão apresentados dois métodos clássicos do tipo passo-a-passo para seleção de variáveis no modelo de Cox. Estes métodos fornecem, em cada passo, novos modelos por adição ou eliminação de uma covariável do modelo obtido no passo anterior e são conhecidos como métodos passo-a-passo (*stepwise*).

3.2.1 Métodos Passo-a-Passo

Os métodos passo-a-passo englobam toda a classe dos métodos *forward, backward e stepwise*. Covariáveis podem ser selecionadas para inclusão no modelo de riscos proporcionais usando os métodos de seleção passo-a-passo, da mesma forma como são usados em outros modelos de regressão, tais como o modelo de regressão linear ou o modelo logístico. A estatística de teste mais utilizada para comparação de modelos encaixados é a da razão de verossimilhanças parcial, dada em (2.6). Entretanto, alguns softwares utilizam o teste de Wald e o teste Escore (Cox e Hinkley, 1974). O processo completo de seleção via *stepwise* consiste de inclusão em *forward* seguido por eliminação em *backward*. O processo de seleção *forward* parte do modelo nulo e adiciona passo-a-passo, ao modelo, as covariáveis que são estatisticamente significativas. O processo de eliminação *backward*, parte do modelo completo e verifica, a cada passo, se cada covariável deve ser mantida no modelo. Estes dois métodos podem ser utilizados separadamente ou conjuntamente.

Estas rotinas automáticas para seleção de covariáveis estão implementadas e, portanto, disponíveis em softwares estatísticos. Entretanto, estas rotinas possuem algumas desvantagens. Tipicamente, elas tendem a identificar um conjunto particular de covariáveis, ao invés de possíveis conjuntos igualmente bons para explicar a resposta. Este fato impossibilita que dois ou mais conjuntos de covariáveis igualmente bons sejam apresentados para o pesquisador, para a escolha do mais relevante na sua área de aplicação. Isto significa que estes métodos são automáticos ao determinar o "melhor"modelo. Na realidade, o que se defende é que o estatístico, juntamente com o pesquisador, tenham uma postura pró-ativa neste processo. Isto implica, por exemplo, que covariáveis importantes em termos clínicos, ou de negócio, devem ser incluídas independente da significância estatística, assim como a importância clínica, ou de negócio, deve ser considerada em cada passo de inclusão ou exclusão no processo de seleção de variáveis.

Frente à estas limitações das rotinas automáticas, optou-se por utilizar um método, que será apresentado na próxima seção, que permite a interferência mais de perto do analista. Mais detalhes sobre os métodos *stepwise* podem ser encontrados em Drapper e Smith (1998).

3.2.2 Um Método Passo-a-Passo Alternativo

Este método alternativo é uma estratégia de seleção de modelos derivada da proposta de Collet (1994). Os passos utilizados no processo de seleção são apresentados a seguir:

1. Ajustar todos os modelos contendo uma única covariável. Incluir todas as covariáveis que forem significativas ao nível de 0,10. Neste passo, é aconselhável utilizar o teste da razão de verossilhanças, apresentado na Seção 2.2.

2. As covariáveis significativas no passo 1 são então ajustadas conjuntamente. Na presença de certas covariáveis, outras podem deixar de ser significativas. Consequentemente, ajustam-se modelos reduzidos, excluindo-se uma única covariável de cada vez. Verifica-se as covariáveis que provocam um aumento estatisticamente significativo na estatística da razão de verossimilhanças, dada em (2.6). Somente aquelas que atingirem a significância permanecem no modelo.

3. Ajusta-se um novo modelo com as covariáveis que atingiram a significância no passo 2. Neste passo, as covariáveis excluídas no passo 2 retornam, uma a uma, ao modelo para confirmar que elas não são estatisticamente significativas.

4. As eventuais covariáveis significativas no passo 3 são incluídas no modelo juntamente com aquelas do passo 2. Neste passo, retorna-se com as covariáveis excluídas no passo 1, uma a uma, para confirmar que elas não são estatisticamente significativas.

5. Ajusta-se um modelo incluindo as covariáveis significativas no passo 4. Neste passo é testado se alguma delas pode ser retirada do modelo.

6. Utilizando as covariáveis que foram significativas no passo 5 ajusta-se o modelo final para os efeitos principais.

Ao ser utilizado este procedimento de seleção, deve-se incluir as informações clínicas, ou de negócio, no processo de decisão e evitar ser muito rigoroso ao testar cada nível individual de significância. Para decidir se um termo deve ser incluído, o nível de significância não deve ser muito pequeno, sendo recomendado um valor próximo de 0,10. Variações deste método de seleção de variáveis podem ser encontrados na literatura. Hosmer e Lemeshow (1989), por exemplo, discutem estes métodos com mais profundidade.

Este método foi programado por sintaxe no SPSS 12.0 e será utilizado no estudo de simulação e nas aplicações reais em que é comparado ao método LASSO e aos critérios de seleção de variáveis apresentados na Seção 3.4.

3.3 Método LASSO (*Least Absolute Shrinkage and Selection Operation*)

O método LASSO *(Least Absolute Shrinkage and Selection Operator)* é um método de seleção de variáveis inicialmente formulado para os modelos lineares. Nesta trabalho serão abordados os principais algorítmos utilizados para o método LASSO bem como a ilustração de seu desempenho em simulações e em problemas envolvendo o modelo de Cox.

A proposta deste método é estimar os coeficientes $\boldsymbol{\beta}$ do modelo completo (todas as covariáveis e nenhuma interação), maximizando o logaritmo da função de

verossimilhança parcial, dada em (2.5), sujeito à restrição de que a soma dos valores absolutos dos parâmetros é limitada por uma constante. O modelo escolhido será aquele formado pelo sub-conjunto das covariáveis cujos coeficientes estimados pelo método LASSO forem não nulos.

Mais formalmente, pode-se apresentar o método LASSO como segue. Denote o logaritmo da função de verossimilhança parcial dada em (2.5) por $\mathrm{l}(\boldsymbol{\beta} \mid \mathbf{x})$, isto é $\mathrm{l}(\boldsymbol{\beta} \mid \mathbf{x}) = \log L(\boldsymbol{\beta} \mid \mathbf{x})$, e assuma que as covariáveis estejam padronizadas. A proposta do método LASSO é estimar $\boldsymbol{\beta}$ através do seguinte critério:

$$\widehat{\boldsymbol{\beta}} = \max \mathrm{l}(\boldsymbol{\beta} \mid \mathbf{x}), \tag{3.1}$$

sujeito à restrição que $\sum |\boldsymbol{\beta}| \leq s$, em que $s > 0$ é um valor inicialmente especificado pelo usuário. Caso não haja restrição, o valor de s seria definido pela soma dos valores usuais das estimativas de máxima verossimilhança parcial.

Tibshirani (1996) propõe dois algorítmos para a obtenção do ajuste do método LASSO. Os algorítmos são iterativos e, geralmente, são necessárias entre p e $2p$ iterações, em que p é o número de covariáveis do modelo completo.

No primeiro algorítmo proposto por Tibshirani (1994), a estratégia para resolver a expressão dada em (3.1) é utilizar o algorítmo de Newton-Raprson. A Figura 3.1 apresenta uma descrição visual desse algorítmo.

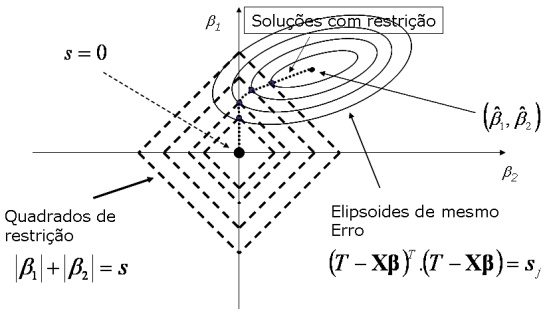

Figura 3.1: Avaliação visual do Método LASSO.

Denote por $\mathbf{x}$ a matriz de covariáveis e por $\boldsymbol{\eta} = \mathbf{x}\boldsymbol{\beta}$. Defina $\mathbf{u} = \partial l/\partial \boldsymbol{\eta}$ e $\mathbf{A} = -\partial^2 l/\partial \boldsymbol{\eta}\boldsymbol{\eta}'$ e considere $\mathbf{z} = \boldsymbol{\eta} + A^{-1}\mathbf{u}$. Então, o primeiro termo da expansão em série de Taylor para $\mathrm{l}(\boldsymbol{\beta})$ tem a seguinte forma:

$$(\mathbf{z} - \boldsymbol{\eta})' \mathbf{A} (\mathbf{z} - \boldsymbol{\eta}). \tag{3.2}$$

Utilizando estes resultados, o seguinte procedimento é utilizado para obter $\hat{\boldsymbol{\beta}}$:

1. Fixe s e inicialize $\hat{\boldsymbol{\beta}} = 0$.
2. Calcule $\boldsymbol{\eta}$,$\mathbf{u}$, $\mathbf{A}$, $\mathbf{z}$, baseados no valor atual de $\hat{\boldsymbol{\beta}}$.
3. Minimize $(\mathbf{z} - \boldsymbol{\eta})' \mathbf{A} (\mathbf{z} - \boldsymbol{\eta})$ sujeito à restrição $\sum |\boldsymbol{\beta}| \leq s$.
4. Repita os passos 2 e 3 até que $\hat{\boldsymbol{\beta}}$ se estabilize.

Segundo Tibshirani (1994), possíveis escolhas para s são: o valor que minimiza a função de validação cruzada ou o valor que maximiza o BIC *(Bayesian Information Criterion)*, a ser apresentado na Seção 3.4.

Uma característica atrativa da restrição $\sum |\boldsymbol{\beta}| \leq s$ é que, frequentemente, alguns dos coeficientes são exatamente zero, o que justificaria excluir do modelo as covariáveis correspondentes aos coeficientes nulos. Em outras palavras, a forma da restrição fornece um modelo final mais estável que os produzidos pelos métodos passo-a-passo ou outros métodos de seleção de covariáveis. O modelo mais parcimonioso será obtido pelo sub-conjunto das covariáveis cujas estimativas dos seus coeficientes forem não nulas.

O algorítmo descrito foi implementado no software R 1.7.1, mas para amostras de tamanho maior que 20, o mesmo já não se mostrava eficiente devido a enorme quantidade de operações matriciais envolvidas, em particular, a inversão de matrizes. Na tentativa de contornar este problema, considerou-se o segundo algorítmo proposto por Tibshirani.

A segunda proposta apresentada por Tibshirani é o algorítmo *forward selection*, ver Tibshirani, Hatie e Friedman (2003). Segue os passos deste algoritmo:

1. Inicialize $\hat{\alpha_k} = 0$, $k = 1$ até K.
2. Seja ξ o tamanho do passo e M a quantidade máxima de passos.

Defina $\xi > 0$ bem pequeno e M o maior possível.

3. **Para** $m = 1$ **até** M calcule:

$(\boldsymbol{\beta}^*, k^*)$=argmin$_{\boldsymbol{\beta},k} \sum_{i=1}^{N}[t_i - \sum_{l=1}^{k}\alpha_l \ \mathrm{T}_l(x_i) - \boldsymbol{\beta} T_k(x_i)]^{\ 2}$

$\alpha_k* \leftarrow \alpha_k * + \xi \ \text{sign} \ (\boldsymbol{\beta}^*)$

4. Saída: $f(x) = \sum_{k=1}^{K} \alpha_k \mathrm{T}_k(x)$

A partir do algorítmo descrito, Azevedo (2005) apresentou uma proposta alternativa que foi programada no MATLAB 7.0 e por se mostrar muito mais eficiente que a anterior, será utilizada nos estudos de simulação e nos exemplos reais (Capítulos 4 e 5, respectivamente). A Figura 3.2 apresenta uma descrição visual desse algorítmo. Esta proposta baseia-se no algorítmo *forward selection*, ver Tibshirani, Hatie e Friedman (2003), e pode ser definida como um procedimento de busca por varredura que se mostra mais eficiente quanto menor for o tamanho do passo dentro dessa varredura e maior a quantidade de passos.

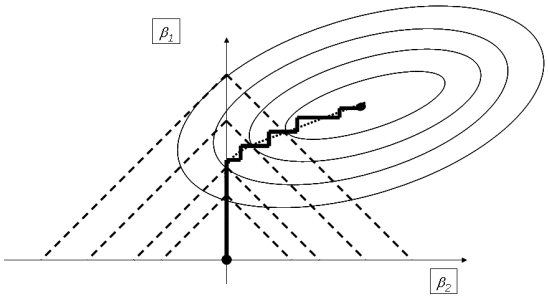

Figura 3.2: Avaliação visual do Método LASSO.

Segue os passos do algorítmo:

Adaptação do algorítimo Forward Selection para o método LASSO

1. Inicialize $\hat{\boldsymbol{\beta}} = 0$.
2. Seja ξ o tamanho do passo e M a quantidade máxima de passos.

Defina $\xi > 0$ bem pequeno e M o maior possível.

Para $m = 1$ **até** M

Armazene a verossimilhança parcial atual, dada em (2.5)

Para $l = 1$ **até** p

Realize a varredura na direção positiva e negativa,

armazenando a verossimilhança parcial dos dois casos.

Compare as verossimilhanças calculadas e guarde a maior.

Fim Para

Identifique o $\hat{\boldsymbol{\beta}}_l$ que resulta na maior verossimilhança.

Atualize o vetor $\hat{\boldsymbol{\beta}}$ com o $\hat{\boldsymbol{\beta}}_l$ ótimo.

Fim Para

3. Calcular BIC para cada um dos M passos.
4. Encontrar a posição do BIC máximo.
5. Produza um gráfico com os $\hat{\boldsymbol{\beta}}$ e trace uma linha vertical no valor encontrado no passo 5.

A interpretação do algorítmo descrito é visual. Para saber quais váriaveis ficaram no modelo final, deve-se analisar o gráfico produzido no passo 5 e identificar quais variáveis apresentam coeficientes diferentes de zero na linha vertical.

3.4 Critérios de Seleção de Modelos

Alguns critérios comuns na literatura também podem ser utilizados para seleção de variáveis. Estes critérios levam em consideração a complexidade do modelo no critério de seleção. São critérios que "penalizam"a verossimilhança, essencialmente, utilizando o número de covariáveis do modelo e, eventualmente, o tamanho da amostra. Critérios de seleção de modelos como o *Akaike Information Criterion* (AIC) e *Bayesian Information Criterion* (BIC) são frequentemente utilizados para selecionar modelos em problemas de sobrevivência para dados censurados. Segundo estes critérios, o modelo mais parcimonioso será aquele que apresentar maior valor de AIC e BIC. Estes critérios foram programados no MATLAB 7.0 e serão comparados com o método LASSO no estudo de simulação (Capítulo 4) e nas aplicações reais (Capítulo 5).

3.4.1 AIC

O *Akaike Information Criterion* (AIC) para o modelo M_i pode ser calculado por:

$$AIC_i = [2lnL(\widehat{\boldsymbol{\beta}}^i|\mathbf{x}, M_i)] - 2p_i, \tag{3.3}$$

em que $\widehat{\boldsymbol{\beta}}^i$ é o estimador de máxima verossimilhança para $\boldsymbol{\beta}^i$ sob o modelo M_i e p_i é a dimensão de $\boldsymbol{\beta}^i$.

Este critério é baseado em considerações frequentistas de eficiência assintótica.

3.4.2 BIC

Considerando a mesma notação utilizada em (3.3), define-se o *Bayesian Information Criterion* (BIC) para o modelo M_i em competição como sendo:

$$BIC_i = 2lnL(\widehat{\boldsymbol{\beta}}^i|\mathbf{x}, M_i) - p_i \; log \; n. \tag{3.4}$$

Capítulo 4

Estudo Comparativo do Método LASSO e Outros Métodos de Seleção de Variáveis

Neste capítulo, comparam-se via simulação Monte Carlo, os métodos LASSO *(Least Absolute Shrinkage and Selection Operator)* e passo-a-passo alternativo e os critérios de seleção de modelos AIC *(Akaike Information Criterion)* e BIC *(Bayesian Information Criterion)*, apresentados no Capítulo 3, para seleção de variáveis no modelo de Cox. O objetivo é avaliar o desempenho do método LASSO se comparado a procedimentos de seleção de modelos comumente utilizados na literatura, para diferentes proporções de censuras, diferentes tamanhos amostrais, diferentes números de covariáveis e diferentes parâmetros para a distribuição Weibull assumida, para a variável tempo de falha.

Este capítulo está organizado como segue: na Seção 4.1, descreve-se o estudo de simulação realizado. Os resultados das simulações são apresentados e analisados na Seção 4.2.

4.1 Descrição dos Cenários

Para comparar os métodos, foram considerados quatro cenários diferentes, que serão descritos a seguir. Considerou-se dois tamanhos de amostra (n = 50 e 100), percentuais de censuras diferentes (0% e 30% de censura), assumiu-se distintas quantidades de covariáveis (4 ou 5) e diferentes parâmetros para a distribuição dos tempos de falha. Em cada cenário, foram consideradas 1000 réplicas Monte Carlo.

Para todos os cenários a variável resposta t foi gerada da seguinte forma:

$$t \sim Weibull(\alpha(x), \rho); \tag{4.1}$$

em que $\alpha(x) = \exp\left(\sum_{l=1}^{p} x_l\beta_l\right)$ e considerou-se $\rho = \frac{1}{2}, 1$ e 2. Observe que para $\rho = 1$ tem-se a distribuição exponencial cuja funçaõ de taxa de falha é constante, e para $\rho = \frac{1}{2}$ e 2 tem-se a Weibull com funções taxa de falha decrescente e crescente, respectivamente. A variável resposta foi gerada a partir de uma distribuição Weibull, pois esta pertence a classe de modelos de riscos proporcionais.

Assumindo, nas simulações, que o vetor $\boldsymbol{\beta}$ seja igual a (1,0,0,1), no caso de 4 covariáveis, e igual a (1,0,0,1,1) para 5 covariáveis, segue a descrição detalhada dos quatro cenários.

Cenário 1: Amostra de tamanho n=50, sem censura e com 4 covariáveis, sendo 2 são geradas da Normal (0,1) e 2 vindas da Bernoulli ($\frac{1}{2}$), nesta ordem;

Cenário 2: Amostra de tamanho n=100, sem censura e com 4 covariáveis, em que, 2 são geradas da Normal (0,1) e 2 vindas da Bernoulli ($\frac{1}{2}$), nesta ordem;

Cenário 3: Amostra de tamanho n=100, com 30% de censura e com 4 covariáveis, sendo que, 2 são geradas da Normal (0,1) e 2 vindas da Bernoulli ($\frac{1}{2}$), nesta ordem;

Cenário 4: Amostra de tamanho n=50, sem censura e com 5 covariáveis, em que 2 são geradas da Normal (0,1), 2 vindas da Bernoulli ($\frac{1}{2}$) e 1 é gerada da Exponencial (1), nesta ordem.

Para os dois primeiros cenários (amostra de tamanhos n=50 e 100, sem censura e quatro covariáveis), foram considerados três valores para o parâmetro de forma da Weibull: $\frac{1}{2}$, 1 e 2. Para os Cenários 3 e 4, o parâmetro de forma utilizado para a Weibull foi 1.

Na implementação do método LASSO testou-se vários valores para ξ e M. A

escolha dos valores 0,001 e 20000 para ξ e M respectivamente, foi baseada na análise de convergência do método.

Como os métodos passo-a-passo alternativo e LASSO e os critérios de seleção de modelos AIC e BIC não foram programados no mesmo software utilizou-se o seguinte procedimento para garantir que seriam usadas as mesmas amostras no estudo de simulação:

1. Gerar amostra no MATLAB e salvar em um arquivo txt.

2. Cálculo do método LASSO e critérios AIC e BIC no MATLAB, ulitizando a amostra obtida no passo 1.

3. Leitura do arquivo obtido no passo 1, por sintaxe no SPSS, e cálculo do método passo-a-passo alternativo.

4.2 Análise dos Resultados

Para avaliar o desempenho do método do LASSO, se comparado aos outros métodos, foram calculados o percentual de acerto do método em cada cenário e o percentual de acerto de cada método para cada covariável individualmente, em cada um dos cenários. Estes resultados estão sumarizados nas Tabelas 4.1, 4.2, 4.3 e 4.4 e nas Figuras 4.1, 4.2, 4.3 e 4.4, a seguir.

4.2.1 O Efeito do Tamanho da Amostra e de Diferentes Parâmentros para a Distribuição do Tempo de Falha

A Tabela 4.1 mostra que os melhores resultados são obtidos quando o parâmetro de forma da Weibull é 1, isto é, quando temos uma função de risco constante. Percebe-se ainda que o método LASSO sempre apresenta o melhor desempenho (selecionando o modelo correto em 87% das vezes ou mais), seguido pelos métodos passo-a-passo alternativo, AIC e BIC, nessa ordem. Observa-se também que os resultados para a Weibull com taxa de falha decrescente são ligeiramente superiores àqueles obtidos para taxa de falha crescente, em todos os métodos.

Similar ao que foi observado para o Cenário 1, nota-se da Tabela 4.2 que os melho-

res resultados para o Cenário 2 são obtidos quando o parâmetro de forma da Weibull é 1. Percebe-se ainda que o método LASSO apresenta o melhor desempenho, seguido pelos métodos passo-a-passo alternativo, BIC e AIC, nessa ordem.

Perceba das Tabelas 4.1 e 4.2 que quando aumentamos o tamanho da amostra, o BIC se mostra melhor que o AIC, provavelmente pela própria natureza do método que leva em conta o tamanho amostral. Também nota-se que, com o aumento do tamanho da amostra, os resultados do LASSO e do passo-a-passo alternativo são mais próximos, tendo um percentual de acerto do modelo em torno de 97%.

Analisando as Tabelas 4.1 e 4.2, observa-se que quando aumentamos o tamanho da amostra todos os métodos apresentam melhor desempenho, como é esperado.

Tabela 4.1: Percentual de Acerto de Cada Método - Cenário 1 - Amostra de tamanho n=50, sem censura e com 4 covariáveis

Modelo	**Método**	**Percentual de Acerto**
$T \sim W(\alpha; 1)$	**LASSO**	0,943
	Passo-a-Passo Alternativo	0,907
	AIC	0,798
	BIC	0,735
$T \sim W(\alpha; 0,5)$	**LASSO**	0,901
	Passo-a-Passo Alternativo	0,813
	AIC	0,754
	BIC	0,716
$T \sim W(\alpha; 2)$	**LASSO**	0,879
	Passo-a-Passo Alternativo	0,781
	AIC	0,722
	BIC	0,659

Tabela 4.2: Percentual de Acerto de Cada Método - Cenário 2 - Amostra de tamanho n=100, sem censura e com 4 covariáveis

Modelo	Método	Percentual de Acerto
$T \sim W(\alpha; 1)$	**LASSO**	0,974
	Passo-a-Passo Alternativo	0,972
	AIC	0,808
	BIC	0,854
$T \sim W(\alpha; 0,5)$	**LASSO**	0,945
	Passo-a-Passo Alternativo	0,928
	AIC	0,789
	BIC	0,847
$T \sim W(\alpha; 2)$	**LASSO**	0,936
	Passo-a-Passo Alternativo	0,917
	AIC	0,763
	BIC	0,832

4.2.2 O Efeito da Proporção de Censura

Como obtido nos Cenários 1 e 2, nota-se da Tabela 4.3 que o LASSO se mostrou melhor que os outros métodos. E, diferentemente do que foi observado no Cenário 2, o BIC tem pior desempenho que o AIC. As Tabelas 4.2 e 4.3 mostram que, na presença de censura, o percentual de acerto diminui no método LASSO, passo-a-passo alternativo e BIC. Ressalta-se que esta diminuição é menos acentuada para o LASSO e mais acentuada para o BIC.

Tabela 4.3: Percentual de Acerto de Cada Método - Cenário 3 - Amostra de tamanho n=100, com 30% de censura e com 4 covariáveis

Modelo	Método	Percentual de Acerto
$T \sim W(\alpha; 1)$	**LASSO**	0,951
	Passo-a-Passo Alternativo	0,910
	AIC	0,809
	BIC	0,754

4.2.3 O Efeito do Número de Covariáveis

Da mesma forma que nos Cenários 1 e 3, da Tabela 4.4, também observa-se melhor desempenho do método LASSO e pior desempenho do BIC. Comparando as Tabelas 4.1 e 4.4 nota-se que os métodos AIC e BIC sofreram as maiores pioras e que o método passo-a-passo foi o que sofreu menos influência do acréscimo de covariáveis.

Comparando os resultados mostrados nas Tabelas 4.1 e 4.4 observa-se que o aumento do número de covariáveis, reduz o desempenho de todos os métodos na seleção de variáveis.

Tabela 4.4: Percentual de Acerto de Cada Método - Cenário 4 - Amostra de tamanho n=50, sem censura e com 5 covariáveis

Modelo	Método	Percentual de Acerto
$T \sim W(\alpha; 1)$	**LASSO**	0,874
	Passo-a-Passo Alternativo	0,859
	AIC	0,680
	BIC	0,623

Em resumo, pode-se concluir que, no contexto do modelo de Cox, em todos os cenários, o desempenho do método do LASSO é superior aos demais. Percebe-se ainda que a eficiência dos métodos diminuem à medida que incluímos censura ou aumentamos o número de covariáveis. Ressalta-se que o AIC não parece ter sido influenciado com o aumento do percentual de censura. Pode-se concluir que quando o parâmetro

de forma da distribuição do tempo de falha é 1, os métodos apresentam uma melhor performace. Nota-se também que para amostras de tamanho 100, não encontramos grandes vantagens em usarmos um algorítmo tão complexo quanto o método LASSO, visto que os resultados apresentados estão muito próximos aos obtidos para o método passo-a-passo alternativo.

4.2.4 Desempenho dos Métodos Covariável-a-Covariável

A seguir estão apresentados o percentual de acerto por covariável para cada um dos cenários, em cada um dos métodos.

Observando as Figuras 4.1, 4.2, 4.3 e 4.4 não se percebe grandes diferenças entre o percentual de acerto para cada covariável, em cada método, nos cenários analisados. No entanto, pequenas diferenças podem ser pontuadas.

Por exemplo, para amostras de tamanho n=100 na presença de censura (Cenario 3), nota-se que o percentual de acertos de todos os métodos para covariáveis geradas de distribuições Bernoulli é menor que o percentual observado para covariáveis geradas de distribuições Normal. Isto pode indicar que os métodos são menos eficientes na seleção de variáveis discretas.

Nota-se também, no Cenário 4, que para todos os métodos o percentual de acertos para covariáveis geradas da distribuição Exponencial é menor que o percentual observado para covariáveis geradas das distribuições Normal e Bernoulli ($\frac{1}{2}$), o que pode indicar uma eficiência menor dos métodos de seleção de covariáveis vindas de distribuições assimétricas.

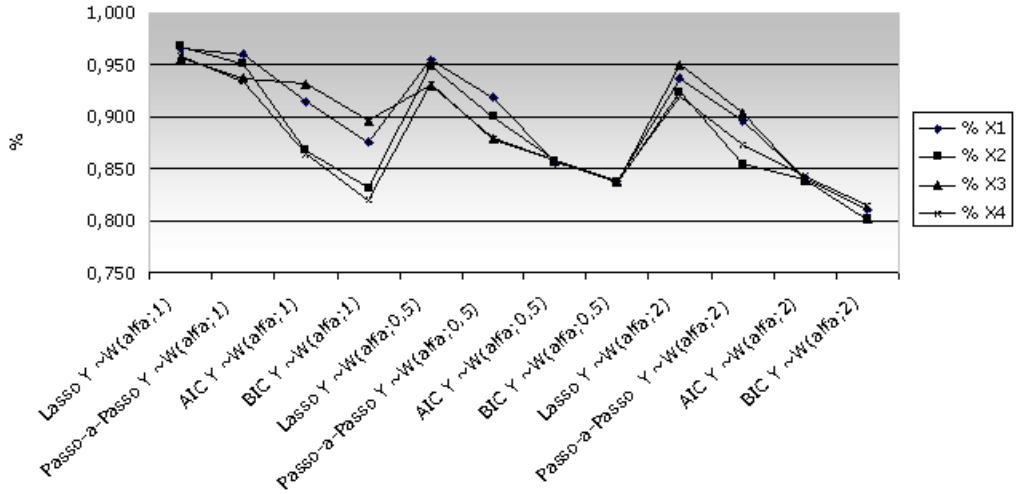

Figura 4.1: Percentual de Acerto de Cada Método - Cenário 1 - Amostra de tamanho n=50, sem censura e com 4 covariáveis

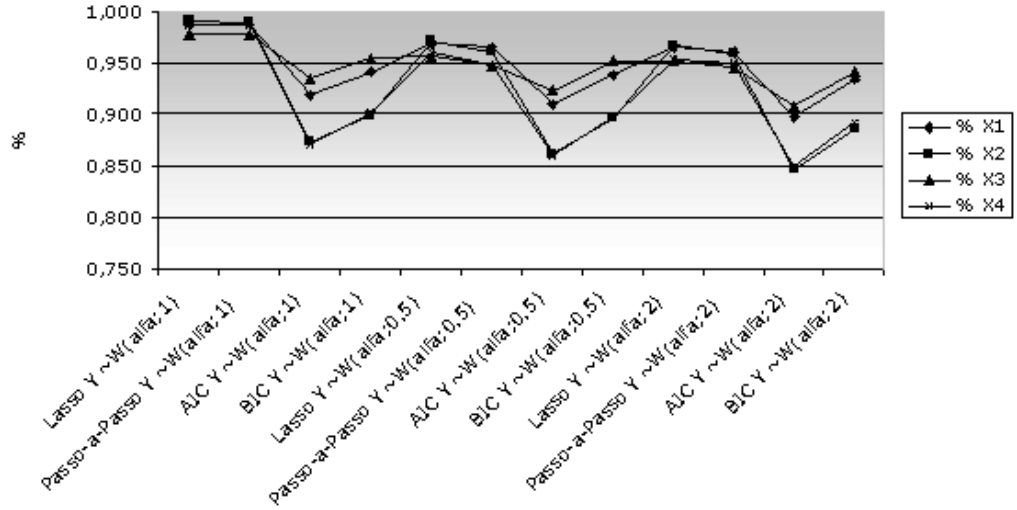

Figura 4.2: Percentual de Acerto de Cada Método - Cenário 2 - Amostra de tamanho n=100, sem censura e com 4 covariáveis

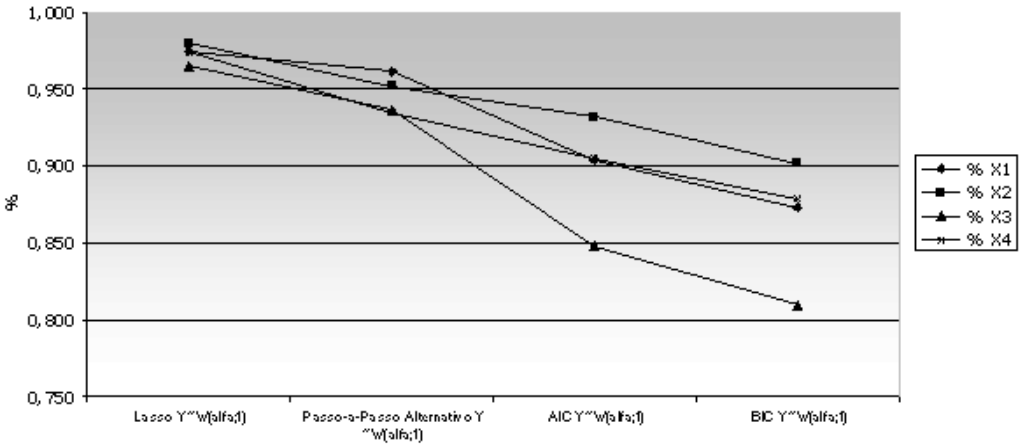

Figura 4.3: Percentual de Acerto de Cada Método - Cenário 3 - Amostra de tamanho n=100, com 30% de censura e com 4 covariáveis

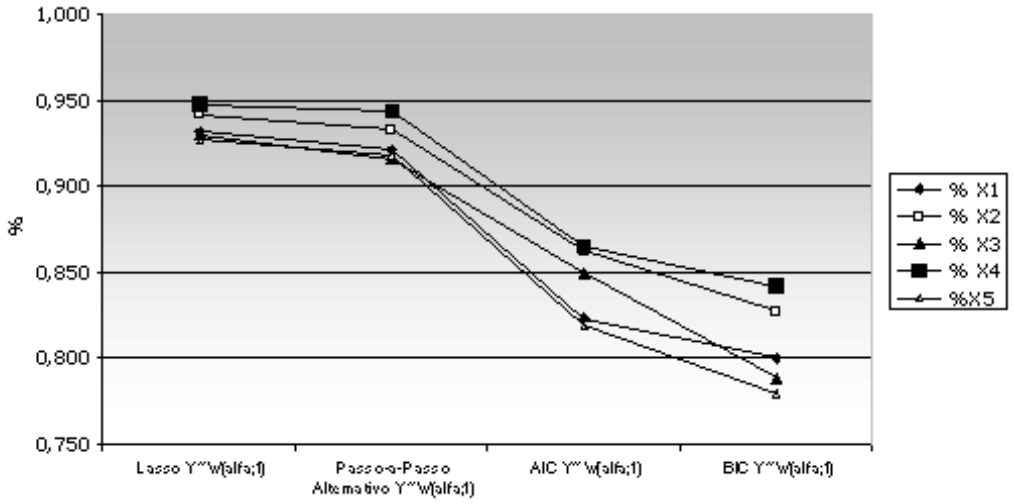

Figura 4.4: Percentual de Acerto de Cada Método - Cenário 4 - Amostra de tamanho n=50, sem censura e com 5 covariáveis

Capítulo 5

Aplicações

Neste capítulo, utilizam-se os métodos LASSO *(Least Absolute Shrinkage and Selection Operator)* e passo-a-passo alternativo e os critérios AIC *(Akaike Information Criterion)* e BIC *(Bayesian Information Criterion)* para seleção de variáveis em dois conjuntos de dados reais. O primeiro envolve a identificação dos fatores preditivos para a morte devido à cirrose e, o segundo, fatores determinantes para a retirada do telefone do cliente no caso em que o mesmo se encontrava inadimplente junto à operadora de telefonia.

Este capítulo está assim organizado. Na Seção 5.1, apresenta-se o estudo de tempo de vida de pacientes cirróticos e, na Seção 5.2 o tempo de planta até a retirada por inadimplência.

5.1 Tempo de Vida de Pacientes Cirróticos

A cirrose é uma doença rara, fatal e de causas desconhecidas que tem uma prevalência de cerca de 50 casos por milhão. A série de dados usadas neste exemplo se encontra no apêndice do livro de Fleming e Harrington (1991).

A descrição clínica da experimentação usada e as covariáveis medidas estão descritas a seguir. Uma discussão mais detalhada pode ser encontrada em Dickson *et al.* (1989). Este mesmo exemplo foi utilizado em Tibshirani (1996).

Esta base de dados foi coletada na Clínica Mayo e inclui todos os pacientes

cirróticos atendidos entre os anos de 1974 e 1984. Pacientes cirróticos, atendidos na Clínica Mayo durante esse intervalo de 10 anos e que atendiam aos critérios de elegibilidade, foram submetidos a um tratamento duplo cego para se testar a eficiência de uma droga. A variável resposta de interesse era o tempo até a morte devido à cirrose. Até o final do estudo, alguns pacientes ainda não tinham morrido, ou tinham sido submetidos ao transplante de fígado, ou tinham morrido de causa diferente da estudada, o que caracterizam censuras. Os pacientes que apresentaram dados faltantes foram excluídos do estudo, totalizando uma amostra de 276 pacientes. Neste exemplo, o percentual de censura foi cerca de 38%. As variáveis consideradas no estudos são apresentadas a seguir.

T: Tempo, em dias, entre o registro do paciente e a morte devido à cirrose;
δ: 1 se Falha e 0 se Censura;
X1: Tratamento (1 se Droga e 2 se Placebo);
X2: Idade, em dias;
X3: Sexo (0 se Masculino e 1 se Feminino);
X4: Presença de *ascites*;
X5: Presença de *hepatomegaly*;
X6: Presença de *spiders*;
X7: Presença de edema;
X8: Medida de bilirubina, em mg/dl;
X9: Medida do colesterol, em mg/dl;
X10: Medida da albumina, em g/dl;
X11: Medida de cobre na urina, em mg/dias;
X12: Medida de alcalinidade da fosfatose, em u/l;
X13: Medida de SGOT, em u/l;
X14: Medida do triglicerides, em mg/dl;
X15: Contagem das plaquetas;
X16: Tempo de *prothombine*, em segundos;
X17: Estágio histológico da doença, classificado em 1, 2, 3 ou 4.

O objetivo do estudo é encontrar o modelo mais parcimonioso que melhor explique o comportamento do tempo em dias, entre o registro do paciente e a morte devido à cirrose. Para isto, serão utilizados os métodos e critérios de seleção de variáveis apresentados no Capítulo 3. Os resultados para o método LASSO estão apresentados nas Figuras 5.1 e 5.2.

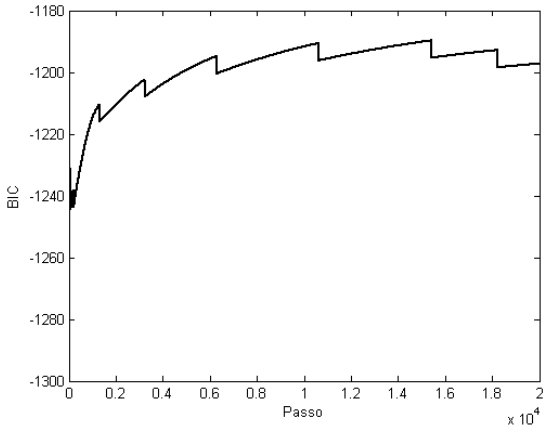

Figura 5.1: Método LASSO - BIC

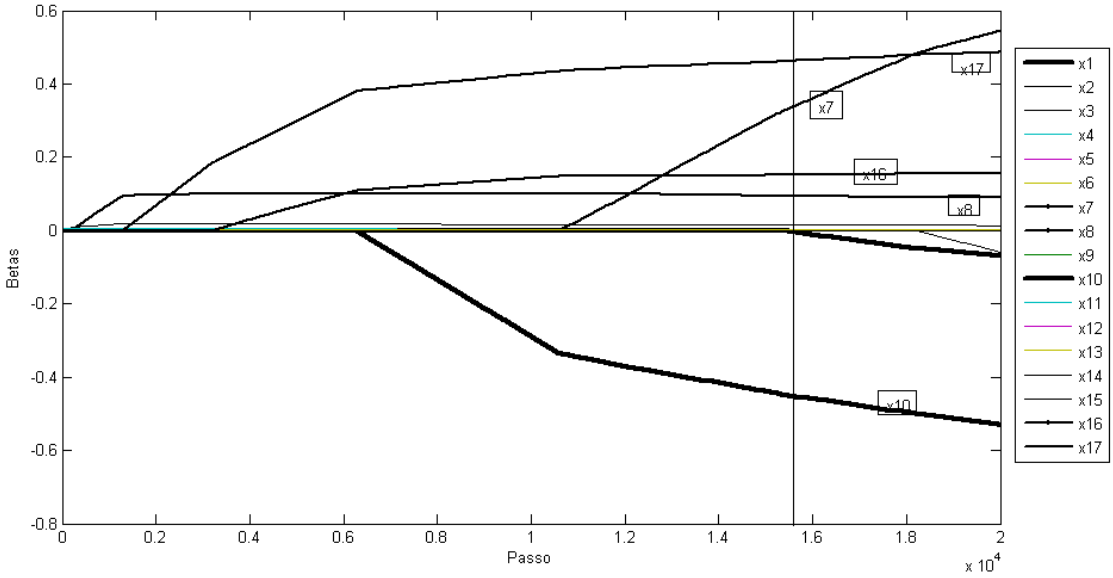

Figura 5.2: Método LASSO - Seleção de Variáveis

A Figura 5.1 apresenta os valores de BIC, calculados para cada um dos passos. O valor máximo está perto do passo 16.000.

Para identificar quais variáveis foram selecionadas pelo método LASSO deve-se traçar uma linha vertical na posição do BIC máximo, conforme feito na Figura 5.2, e observar quais variáveis apresentam coeficientes que são diferentes de zero nesse ponto. Dessa forma, pode-se concluir que o método LASSO selecionou as variáveis X7, X8, X10, X16 e X17.

A Tabela 5.1 apresenta os resultados do método passo-a-passo alternativo. Neste método usou-se o nível de 0,10.

Tabela 5.1: Seleção de covariáveis considerando o método passo-a-passo alternativo

Passos	**Modelo**	**Valor p**
Passo 1	X1	0,230
	X2	0,592
	X3	0,080
	X4	0,911
	X5	0,699
	X6	0,982
	X7	0,032
	X8	0,001
	X9	0,372
	X10	0,011
	X11	0,004
	X12	0,423
	X13	0,148
	X14	0,461
	X15	0,286
	X16	0,091
	X17	0,023

Passos	Modelo	Valor p
Passo 2	X3+X7+X8+X10+X11+X16+X17	-
	X7+X8+X10+X11+X16+X17	0,230
	X3+X8+X10+X11+X16+X17	0,000
	X3+X7+X10+X11+X16+X17	0,000
	X3+X7+X8+X11+X16+X17	0,000
	X3+X7+X8+X10+X16+X17	0,000
	X3+X7+X8+X10+X11+X17	0,000
	X3+X7+X8+X10+X11+X16	0,000
Passo 3	X7+X8+X10+X11+X16+X17	-
	X3+X7+X8+X10+X11+X16+X17	0,230
Passo 4	X7+X8+X10+X11+X16+X17	-
	X1+X7+X8+X10+X11+X16+X17	0,268
	X2+X7+X8+X10+X11+X16+X17	0,438
	X4+X7+X8+X10+X11+X16+X17	0,949
	X5+X7+X8+X10+X11+X16+X17	0,612
	X6+X7+X8+X10+X11+X16+X17	0,945
	X9+X7+X8+X10+X11+X16+X17	0,193
	X12+X7+X8+X10+X11+X16+X17	0,455
	X13+X7+X8+X10+X11+X16+X17	0,169
	X14+X7+X8+X10+X11+X16+X17	0,608
	X15+X7+X8+X10+X11+X16+X17	0,882

O modelo final selecionado pelo método passo-a-passo alternativo inclui as variáveis X7, X8, X10, X11, X16 e X17. Os passos 5 e 6 não foram necessários.

Da forma como os critérios AIC e BIC estão definidos no Capítulo 3, seria necessário rodar os critérios 2^p vezes para a escolha do modelo final. Neste caso seriam necessárias 2^{17} rodadas para chegar ao modelo final. Para simplificar, foi feita uma pré seleção de variáveis, através de uma análise univariada, considerando uma significância de 0,20. Dessa forma, reduzimos a quantidade de modelos para 2^8. O modelo escolhido pelo método AIC inclui as variáveis X3, X8, X10 e X16 e o BIC forneceu um modelo

final com as seguintes covariáveis X7, X10, X11, X16 e X17.

A Tabela 5.2 compara os modelos finais selecionados segundo cada método.

Tabela 5.2: Apresentação dos modelos

Método	Modelo
Lasso	X7, X8, X10, X16 e X17
Passo-a-Passo Alternativo	X7, X8, X10, X11, X16 e X17
AIC	X3, X8, X10 e X16
BIC	X7, X10, X11, X16 e X17

As variáveis X10 e X16 estão presentes em todos os modelos. O método que inclui mais covariáveis é o passo-a-passo alternativo, 7 covariáveis, seguido pelo BIC e LASSO com 5 covariáveis.

Os métodos LASSO e passo-a-passo alternativo retornaram modelos finais bem parecidos. De acordo com os estudos de simulação apresentados no Capítulo 4, o melhor método é o LASSO, seguido pelo passo-a-passo alternativo. Portanto, nesse estudo real, o modelo final que poderia ser indicado como melhor é aquele que inclui as covariáveis X7, X8, X10, X16 e X17.

Ressalta-se que os resultados obtidos foram os mesmos de Tibshirani (1996).

5.2 Tempo de Planta até a Retirada por Inadimplência

No mercado de telecomunicações, a TELEMAR tem sido *benchmarking* em seu processo de cobrança. Tendo o objetivo de superar as melhorias já alcançadas, a partir de técnicas e ferramentas mais atualizadas, a TELEMAR pretende alcançar um patamar acima de suas conquistas, tornando-se um *benchmarking* para o processo de cobrança no mercado em geral.

O processo de cobrança da TELEMAR tinha um histórico de baixa inadimplência. Essa situação alterou-se muito rapidamente tão logo começaram as expansões, para atender um público maior, mais precisamente após a sua privatização. Para fazer face a inadimplência que surge com o atendimento a classes menos favorecidas, os procedimentos para a recuperação de crédito foram todos padronizados e enrijecidos.

Após um domínio maior de seu novo mercado, a TELEMAR volta a buscar uma maior flexibilização e, para isto, resolve investir em ferramentas estatísticas para tentar descobrir quais variáveis influenciam no tempo até a retirada do cliente da base de dados da TELEMAR por inadimplência. O modelo de COX se aplica a esta situação. Para efeito de análise, considera-se como variável resposta o tempo até a retirada do cliente por inadimplência, sendo que o tempo zero é a data de entrada do cliente na base. Pode-se exemplificar como casos de censuras, os clientes que saíram espontaneamente da base ou, ainda, os clientes que permanecem na base por serem adimplentes.

A análise foi feita com uma amostra representativa de 500 clientes da TELEMAR varejo com mais de seis meses de planta (tempo que está na base). O percentual de censura nesta amostra foi de 28%. Abaixo segue uma descrição das variáveis do banco.

T: Tempo, em dias, entre a entrada do cliente na base da TELEMAR e a retirada do telefone do cliente devido à falta de pagamento;
δ: 1 se Falha 0 se Censura;
X1: Classificação do cliente na TELEMAR (1 se diamante; 2 se ouro; 3 se prata; 4 se bronze; e 5 se outros);
X2: Cliente tem Fale local? - Produto de restrição que só permite ligações locais (1-Sim O-Não);
X3: Cliente tem Minha linha? - Produto de restrição que impede o telefone de realizar chamadas após um número fixo de minutos gastos (1 se Sim O se Não);
X4: Cliente tem Bloqueio temporário? - Produto de restrição que impede o telefone de realizar qualquer tipo de ligação (1 se Sim O se Não);
X5: Número de parcelamentos já feitos na TELEMAR;
X6: Cliente é BPI? - Cliente reincidente na inadimplência ao qual foi dado mais uma chance de permanecer na base (1 se Sim O se Não);
X7: Cliente tem DACC? - Débito automático em conta corrente (1 se Sim O se Não);
X8: Cliente é FPD? - Cliente que não pagou a primeira conta (1 se Sim O se Não);
X9: Cliente é Convergente? - Cliente que possui mais de um produto TELEMAR, por exemplo, celular da Oi ou Velox (1 se Sim O se Não).

As Figuras 5.3, 5.4 e 5.5, mostram o perfil da TELEMAR, segundo as variáveis previamente descritas.

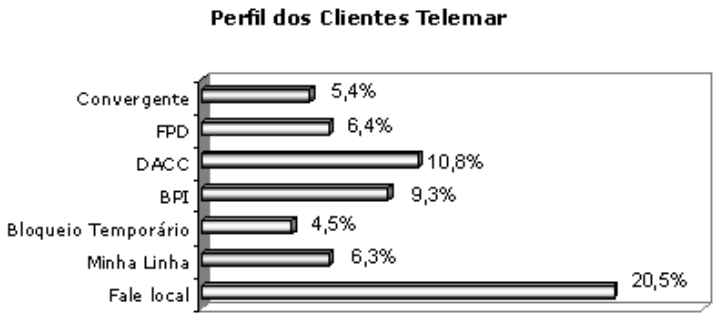

Figura 5.3: Perfil descritivo dos clientes TELEMAR.

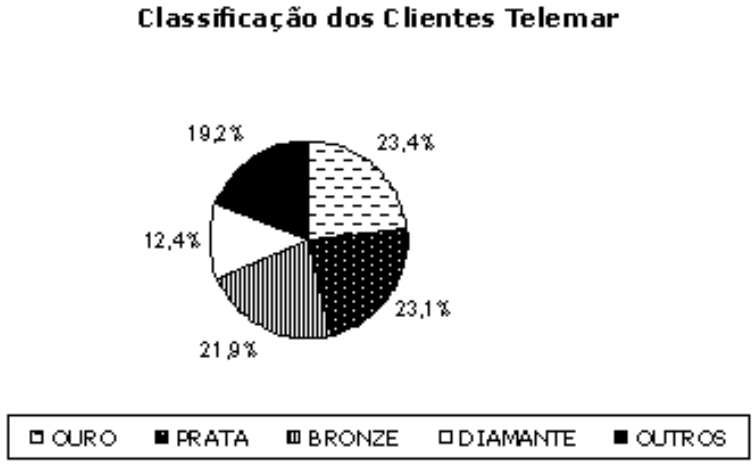

Figura 5.4: Classificação dos clientes TELEMAR.

As Figuras 5.3, 5.4 e 5.5, mostram que 20,5% dos clientes TELEMAR têm o produto de restrição fale local, 6,3% minha linha e 4,5% bloqueio temporário. Observa-se ainda que apenas 10,8% dos clientes têm débito automático em conta corrente, 9,3% já passaram por BPI, 6,4% não pagaram a primeira conta e, apenas 5,4% dos clientes, têm outro produto da companhia. Nota-se também que a grande maioria dos clientes nunca parcelaram suas contas.

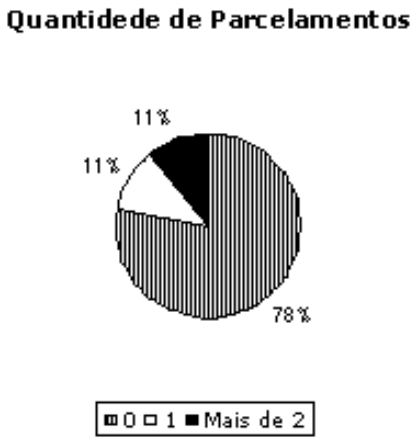

Figura 5.5: Quantidade de parcelametos por cliente.

Como o objetivo é encontrar o modelo mais parcimonioso e que melhor explique o comportamento do número de dias entre o registro do cliente na TELEMAR e a retirada do seu telefone por motivo de falta de pagamento serão utilizados os métodos e critérios de seleção de variáveis apresentados no Capítulo 3. Os resultados são apresentados a seguir. Vale ressaltar que as variáveis X1, X7 e X9 não entraram na análise por não serem relevantes do ponto de vista de negócio.

A Figura 5.6 representa os valores de BIC, calculados para cada um dos passos. O valor máximo está perto do passo 19000.

Para identificar quais variáveis foram selecionadas pelo método LASSO deve-se traçar uma linha vertical na posição do BIC máximo, conforme feito na Figura 5.7, e observar quais variáveis são diferentes de zero nesse ponto. Dessa forma, pode-se concluir que o método do LASSO selecionou as variáveis X2, X3, X4, X5, X6 e X8.

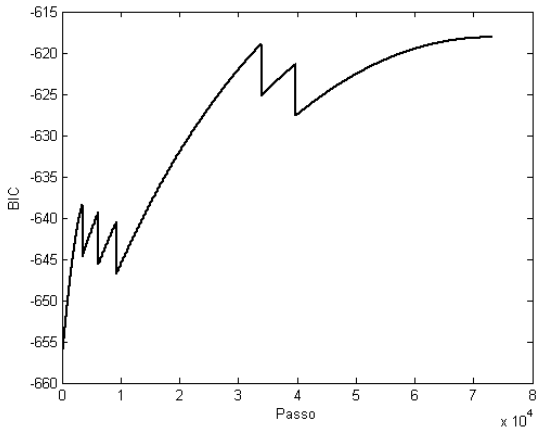

Figura 5.6: Método LASSO - BIC

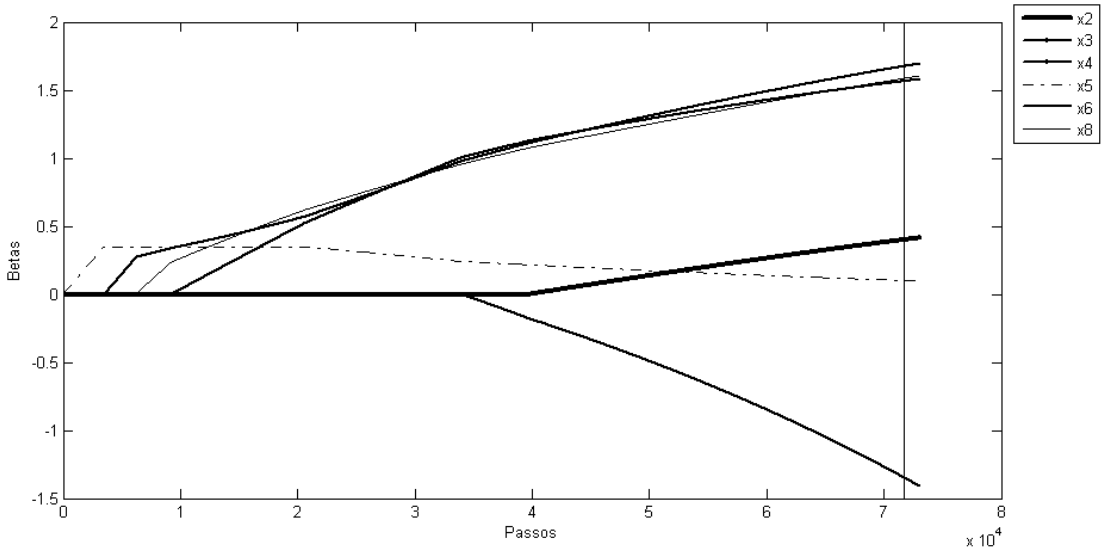

Figura 5.7: Método LASSO - Seleção de Variáveis

A Tabela 5.3 apresenta os resultados do método passo-a-passo alternativo.

Tabela 5.3: Seleção de covariáveis considerando o método-passo-a passo alternativo

Passos	**Modelo**	**Valor p**
Passo 1	X2	0,195
	X3	0,000
	X4	0,067
	X5	0,373
	X6	0,000
	X8	0,000
Passo 2	X3+X4+X6+X8	-
	X4+X6+X8	0,253
	X3+X6+X8	0,000
	X3+X4+X8	0,000
	X3+X4+X6	0,000
Passo 3	X4+X6+X8	-
	X3+X4+X6+X8	0,253
Passo 4	X4+X6+X8	-
	X2+X4+X6+X8	0,003
	X5+X4+X6+X8	0,385
Passo 5	X2+X4+X6+X8	-
	X4+X6+X8	0,000
	X2+X6+X8	0,000
	X2+X4+X8	0,000
	X2+X4+X6	0,000

O modelo final selecionado pelo método passo-a-passo alternativo inclui as variáveis X2, X4, X6 e X8. O passo 6 não foi necessário.

Neste exemplo, como a quantidade de covariáveis é reduzida, todas as 2^6 rodadas para chegar ao modelo final segundo os critérios AIC e BIC foram feitas. O modelo escolhido pelo método AIC inclui as variáveis X2, X3, X4 e X8 e o BIC forneceu um modelo com as seguintes covariáveis, X2, X3, X4, X5, X6 e X8.

A Tabela 5.4 compara os modelos finais selecionados segundo cada método.

Tabela 5.4: Apresentação dos modelos

Método	**Modelo**
Lasso	X2, X3, X4, X5, X6 e X8
Passo-a-Passo Alternativo	X2, X4, X6 e X8
AIC	X2, X3, X4 e X8
BIC	X2, X3, X4, X5, X6 e X8

As variáveis X2 e X8 estão presentes em todos os modelos. O BIC e o LASSO não excluíram nenhuma covariável. O método AIC e passo-a-passo alternativo incluíram 4. De acordo com o estudo de simulação apresentado no Capítulo 4, o melhor modelo é obtido através do método LASSO. Então a indicação do modelo mais parcimonioso para a TELEMAR deveria incluir as covariáveis X2, X3, X4, X5, X6 e X8.

Capítulo 6

Considerações Finais e Trabalhos Futuros

Neste trabalho utilizou-se e comparou-se o método LASSO *(Least Absolute Shrinkage and Selection Operator)* com o método passo-a-passo alternativo e com os critérios de seleção de modelos AIC *(Akaike Information Criterion)* e BIC *(Bayesian Information Criterion)*.

Além disso, foi implementado um algoritmo alternativo para o método LASSO para seleção de variáveis no modelo de Cox, visto que o algoritmo sugerido na literatura por Tibshirani (1996) se mostra ineficiente para tamanhos amostrais maiores que 20. Também foram implementados o método passo-a-passo e os critérios de seleção de modelos AIC e BIC que tiveram seus desempenhos comparados com o do LASSO por meio de um estudo Monte Carlo. Também utilizou-se estes métodos e critérios para determinação do modelo em duas situações reais, a saber, identificação dos fatores preditivos para a morte devido à cirrose e fatores determinantes para a retirada do telefone do cliente no caso em que o mesmo se encontrava inadimplente junto à operadora de telefonia.

Como visto nos estudos de simulação, em geral, o método LASSO se mostrou eficiente na seleção de variáveis no contexto do modelo de Cox. No entanto, para base de dados de tamanho 100 o seu desempenho não foi tão superior a outros métodos mais simples, como por exemplo os métodos passo-a-passo. A presença de censuras e

a inclusão de covariáveis também tiveram um impacto relevante na eficiência de todos os métodos de seleção de variáveis aqui apresentados. Observa-se também que, para todos os modelos, os melhores resultados são obtidos quando o parâmetro de forma da Weibull é 1. Nota-se ainda que os resultados para a Weibull com taxa de falha decrescente são ligeiramente superiores àqueles com taxa de falha crescente.

Neste trabalho, a eficiência do método LASSO foi comparada a outros métodos, como o passo-a-passo alternativo, AIC e BIC. Porém, comparações com outros métodos de seleção de variáveis como, por exemplo, os métodos bayesianos apresentados em Kadane e Lazar (2004), Faraggi e Simon (1998) e Dellaportas e Smith (1993) e os métodos clássicos, como o GARROTE (Breiman,1995) são motivações para trabalhos futuros.

Um outro tópico interessante para pesquisas futuras é uma análise mais profunda do efeito de covariáveis assimétricas e discretas na eficiência dos métodos de seleção de variáveis.

Referências Bibliográficas

[1] Azevedo, M. (2005) - Seleção de Variáveis em Modelos Lineares Generalizados via Método LASSO - Seminário apresentado no Departamento de Estatística da UFMG, Brasil.

[2] Breiman, L. (1995). "Better Subbset selection Using the Nonnegative Garrote", *Technometrics.* **37**, 373-384.

[3] Colosimo, E. A., Giolo, S.R. (2006). *Análise de Sobrevivência Aplicada*, ABE-Projeto Fisher. São Paulo: Blucher.

[4] Collet, D. (1994). *Modelling Survival Data in Medical Research*, Chapman and Hall, Londron.

[5] Cox, D.R. (1972). "Regression Models and Life Tables (with discussion)", *Journal of the Royal Statistical Society.* **34**, 187-220.

[6] Cox, D.R., Hinkley, D.V. (1974). *Theorical Statistics*, Chapman and Hall, Londron.

[7] Dellaportas, P., Smith, A.F.M. (1993). "Bayesian Inference for generalized Linear and Proportional Hazards via Gibbs Sampling ", *Applied Statistics.* **42**, 443-459.

[8] Draper, D., Smith, H. (1998). *Applied Regression Analysis*, New York: Wiley.

[9] Faraggi, D., Simon, R. (1998). "Bayesian Variable Selection Method for Censored Survival Data", *Biometrics.* **54**, 1475-1485.

[10] Fleming, T. R, Harrington, D.P. (1991). *Counting Processes and Survival Analysis*, New York: Wiley.

[11] George, E.I. (2000). "The variable Selection Procedure ," *Journal of the American Statistical Association.* **95**, 1304-1308.

[12] Hosmer, D.W., Lemershow, S. (1989). *Applied Logistic Regression*, Wiley Series in Probability and Statistics.

[13] Kadane, J.C., Lazar, N. A. (2004). "Methods and Criteria for Model Selection", *Journal of the American Statistical Association.* **99**, 279-290.

[14] Paulino, C.D., Turkman, M.A.A., Murteira, B. (2003). *Estatística Bayesiana*, Fundação Calouste Gulbenkian, Lisboa.

[15] Stigler, S.M. (1994). "Citation Patterns in the Journals of Statistics and Probability", *Statistical Science.* **9**, 94-108.

[16] Tibshirani, R. (1994). *A Proposal for Variable Selection in the Cox Model*, *Technical Report*. Toronto, Toronto.

[17] Tibshirani, R. (1996). "The Lasso method for Variable Selection in the Cox Model", *Journal of the Royal Statistical Society.* **58**, 267-288.

[18] Tibshirani, R. (1997). "Regression Shrinkage and Selection via the Lasso ", *Statistics in Medicine.* **16**, 385-395.

[19] Tibshirani, R., Hatie, T., Friedman, J. (2003). *The Elements of Statistical Learning Data Mining, Inference and Prediction*, New York: Springer.

[20] Wiesner R.H., Porayko M.K., Dickson E.R., Gores G.J. (1989). "Liver Transplantation: The Hepatology Perspective ", *Hepatology.* **10**, 1-7.