UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA UFSCar-USP

Gilberto Pereira de Alcântara Junior

Avaliação do lasso e métodos alternativos em modelos de regressão logística

Dissertação apresentada ao Departamento de Estatística - DEs - UFSCar e ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Estatística - Programa Interinstitucional de Pós-Graduação em Estatística. Área de Concentração: Estatística

Orientador: Prof. Dr. Gustavo Henrique Araújo Pereira

São Carlos
Março de 2021

Gilberto Pereira de Alcântara Junior

Lasso evaluation and alternative methods in logistic regression models

Master dissertation submitted to the Departamento de Estatística - DEs - UFSCar and to the Instituto de Ciências Matemáticas e de Computação - ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Joint Graduate Program in Statistics DEs-UFSCar/ICMC-USP.
Concentration Area: Statistics

Advisor: Prof. Dr. Gustavo Henrique Araújo Pereira

São Carlos
March 2021

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Gilberto Pereira de Alcântara Júnior, realizada em 11/03/2021.

Comissão Julgadora:

Prof. Dr. Gustavo Henrique de Araujo Pereira (UFSCar)

Profa. Dra. Mônica Carneiro Sandoval (IME-USP)

Prof. Dr. Izabela Regina Cardoso de Oliveira (UFLA)

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

"If one thing had been different, would everything be different today?"

Aaron Dessner e Taylor Swift.

À Ana Pereira de Alcântara.

Agradecimentos

Primeiramente agradeço sempre a Deus, por ter me concedido muita força, coragem e sempre me abençoar nessa caminhada que levou 2 anos, estando comigo nos momentos mais obscuros.

Aos meus pais, Gilberto Pereira de Alcântara e Edileuza Candida da Silva, por sempre me apoiarem, se preocuparem comigo e nunca me deixar desistir desse sonho que se torna realidade. Também a minha irmã Tainá e meu irmão Rikelme que me apoiaram emocionalmente sempre que foi preciso.

Aos meus queridos amigos Leonardo, Bruna, Ravena e Ana Julia, pelos quais tenho como minha família, pois sempre estiveram ao meu lado compartilhando momentos e risadas.

A minha querida avó que infelizmente não está mais presente fisicamente em nosso meio, porém sempre segue me guiando espiritualmente e por ter sido a primeira pessoa que me inspirou a sonhar. Apesar de não estar presente fisicamente, seu apoio foi herdado para minhas tias Lena e Silvia.

Ao meu orientador, Prof. Dr. Gustavo Henrique de Araujo Pereira, primeiramente por toda paciência e compreensão na difícil fase pela qual passei durante a pandemia em que vivemos, por sempre me orientar da melhor forma possível para que eu buscasse o aprendizado próprio, e por ter me agraciado com o prazer de ser seu orientando.

Aos amigos de graduação Érika, Cleidison, Sérgio e Matheus que sempre estiveram comigo desde o começo desse percurso. Ao meu amigo querido Giovanni que talvez tenha sido a maior companhia nessa jornada academica, sempre ao me lado.

À Universidade Federal de São Carlos, à Universidade de São Paulo e ao Programa Interinstitucional de Pós-Graduação em Estatística pela oportunidade de fazer o Mestrado em Estatística. Agradeço aos recursos disponibilizados do HPC pela Superintendencia de Tecnologia da Informacao da Universidade de São Paulo.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001"

Resumo

Alcântara Junior, G. P. **Avaliação do lasso e métodos alternativos em modelos de regressão logística**.

A regressão logística sempre foi uma importante ferramenta não só na área de estatística, mas também em diversas outras áreas como econômica, biológica e médica. Em muitas dessas áreas é comum se deparar com problemas de alta dimensionalidade, no qual o número de cováriaveis a serem testadas é maior do que o tamanho amostral. Métodos clássicos de estimação apresentam certos problemas em alta dimensionalidade. Uma das formas de solucionar esse problema é a estimação por métodos de penalização, como o lasso proposto por Tibshirani (1996). Apesar dos muitos trabalhos feitos da aplicação do lasso no modelo de regressão logística, nenhum apresenta um estudo completo de simulação do desempenho de predição do método utilizando alguma medida tradicional de avaliação de performance. Também não há na literatura trabalhos que comparam o desempenho de outras possíveis combinações feitas a partir do lasso, como por exemplo, o lasso para selecionar covariáveis e a estimação via máxima verossimilhança, ou a seleção via stepwise e a estimação via lasso. Neste trabalho é apresentado um extenso estudo de simulação sob diversos cenários criados com o objetivo de estudar e comparar o desempenho do lasso e outras 3 técnicas combinadas no modelo de regressão logística. Também foram estudados e analisados vários exemplos de aplicações em que o modelo logístico pode ser usado. Através dos resultados obtidos tanto pelas simulações quanto pelas aplicações, em relação ao poder preditivo, foi possível constatar que o lasso se sobressaía ou tinha desempenho similar aos outros métodos em todos os cenários apresentados. Em relação à comparação do modelo ajustado com o verdadeiro, nenhum dentre os métodos considerados se destaca em todos os cenários e em relação a todos os aspectos analisados.

Palavras-chave: Estimador de máxima verossimilhança, Lasso, Regressão Logística, Seleção de variáveis, Stepwise.

Abstract

Alcântara Junior, G. P. **Lasso evaluation and alternative methods in logistic regression models**.

Logistic regression has always been an important tool not only in the area of statistics, but also in several other areas such as economic, biological and medical. In many of these areas it is common to encounter problems of high dimensionality, in which the number of covariates to be tested is greater than the sample size. Classic estimation methods present certain problems in high dimensionality. One of the ways to solve this problem is the estimation by methods of penalty, as the lasso proposed by Tibshirani (1996). Despite the many works done on the application of lasso in the logistic regression model, none of them presents a complete study of simulation of the method's prediction performance using some traditional measure of performance evaluation. There are also no studies in the literature that compare the performance of other possible combinations made from lasso, such as lasso to select covariates and estimation via maximum likelihood, or selection via stepwise and estimation via lasso. In this work an extensive simulation study is presented under several scenarios created in order to study and compare the performance of the lasso and 3 other techniques combined in the logistic regression model. Several examples of applications in which the logistic model can be used were also studied and analyzed. Through the results obtained both by the simulations and by the applications, in relation to the predictive power, it was possible to verify that the lasso stood out or had similar performance to the other methods in all the presented scenarios. Regarding the comparison of the adjusted model with the real one, none of the methods considered stands out in all scenarios and in relation to all aspects analyzed.

Keywords:Lasso, Logistic Regression, Maximum Likelihood Estimator, Selection of variables, Stepwise.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Com o atual crescimento constante da massa de dados no mundo, se faz cada vez mais necessário saber trabalhar com grandes bases de dados. No cenário de regressão, isto pode significar uma base de dados em alta dimensionalidade, o que ocorre quando o número de covariáveis p é maior do que o tamanho amostral n. Neste ponto técnicas clássicas de estimação de parâmetros como a de mínimos quadrados se tornam problemáticas, pois quando $n < p$ a solução que minimiza a soma do quadrado dos erros não é única (Hastie et al., 2015). Quando o número de covariáveis é muito grande, também é difícil o uso de técnicas clássicas de seleção de covariáveis como stepwise (James et al., 2013), pois demandam tempo e custo computacionais muitas vezes inviáveis.

Para solucionar os problemas encontrados em alta dimensionalidade, nas últimas décadas surgiram vários métodos de regularização, isto é, métodos que adicionam uma penalização na equação de mínimos quadrados. Dentre esses métodos, um dos mais utilizados é o lasso (*last absolute shrinkage and selection operator*) proposto por Tibshirani (1996). O lasso minimiza a soma de quadrados adicionada a um peso, ou penalização não negativa de forma a criar esparsidade dentro do modelo, isto é, fazendo muitos coeficientes convergirem para zero. Dessa forma, o lasso, simultaneamente, estima os parâmetros e seleciona covariáveis para o modelo.

Sabendo que o lasso é um técnica de seleção e estimação, entra-se em questionamento se de fato ela se sobressai de maneira geral em contrapartida à combinação clássica da estimação via função de verossimilhança e seleção via stepwise. Kumar et al. (2019) fizeram um estudo de comparação do poder preditivo de cada um dos métodos no modelo de regressão linear utilizando um conjunto de dados reais, no qual constatou-se um desempenho melhor do lasso. Hastie et al. (2020) demonstraram em seu estudo de simulação também com modelo de regressão linear normal, que em diferentes cenários e métricas estatísticas cada método possui conjuntos de cenários específicos onde se sobressai em relação ao outro.

A técnica do lasso pode ser facilmente estendida para os demais modelos de regressão. A abordagem utilizada é semelhante a da estimação no modelo de regressão linear mas,

nesses casos, um termo de penalização não negativo é adicionado ao oposto do logaritmo da função de verossimilhança. Na literatura podemos encontrar diversos trabalhos sobre a aplicação do lasso em diferentes modelos de regressão como: Belloni et al. (2011) que propõe o uso do lasso para previsões de primeiro estágio e estimação em modelos de variáveis instrumentais lineares no caso canônico gaussiano; Das e Sobel (2015) que propõe o Dirichlet lasso para seleção de covariáveis em modelos de regressão com enfoque bayesiano; Ahmed et al. (2012) que propõe a estimação via lasso no modelo de regressão Weibull com censura e Tibshirani (1997) que apresenta o lasso aplicado ao modelo de cox.

A regressão logística é um dos principais métodos utilizados para realizar predição ou classificação quando a variável resposta é binária, isto é possui apenas duas classes, normalmente 0 ou 1. A regressão logística é muito popular na área de estudos biológicos e medicinais, nos quais a presença de alta dimensionalidade dos dados é ainda mais frequente. Um exemplo é o estudo de genética, no qual estudam-se milhares de genes e espera-se que poucos deles sejam de fato relacionados ao objeto em estudo. Kim et al. (2018) fizeram um estudo da performance de predição do modelo logístico com aplicação na predição de câncer de pulmão, usando os métodos do lasso e stepwise com estimação por máxima verossimilhança, no qual foi possível constatar que o lasso teve um maior poder preditivo.

Há na literatura diversos trabalhos que usam o lasso para selecionar as covariáveis e estimar os parâmetros na regressão logística (Steyerberg et al., 2000; Uh et al., 2007; Wang et al., 2004). Além disso, diversas extensões do lasso já foram propostas como Group lasso proposto por Meier et al. (2008), Adaptive lasso proposto por Zou (2006) e o Fused lasso proposto por Tibshirani et al. (2005). Porém, há apenas um trabalho que realiza um amplo estudo de simulação na regressão logística comparando o lasso com outros métodos. Van Calster et al. (2020) realizaram um estudo de simulação comparando o lasso com outros métodos de estimação quanto ao risco de subestimar ou superestimar eventos raros no modelo de regressão logística, isto é, quando uma das classes da variável resposta possui proporção muito baixa, o que é comum em estudos de doenças raras. Neste estudo, ele mostra que se o número de covariáveis é pequeno, com correlação alta e a proporção de uma das classes é muito baixa, o lasso tende a subestimar esses eventos raros. Porém, no estudo não há nenhuma comparação da performance de predição do lasso com nenhum outro método usando alguma medida usualmente utilizada para isso como a área sob a curva ROC. Além disso, o trabalho não considera o método que combina a seleção de variáveis pelo método stepwise e estimação por máxima verossimilhança. Além do trabalho mencionado, Machado (2018) fez um estudo de simulação comparando o lasso com o método Seleção de Variáveis via Busca Estocástica (SSVS). Porém, essa comparação foi feita sob um enfoque bayesiano e não considerou os demais métodos considerados neste trabalho.

Sabendo que o lasso é um método de seleção e estimação, também surge o questionamento de sua performance quando combinado a outros métodos para selecionar

ou estimar. Por exemplo, usar o lasso para estimar os parâmetros do modelo após um processo de seleção de variáveis via stepwise, ou uma seleção via lasso e estimação por máxima verossimilhança.

O objetivo deste trabalho é fazer um estudo aprofundado de comparação entre os seguintes métodos: lasso para estimação dos parâmetros do modelo combinado à seleção de covariáveis via stepwise, estimação via máxima verossimilhança combinado à seleção via stepwise, estimação via função de máxima verossimilhança e seleção via lasso e o lasso como método único para seleção e estimação. Todos esses métodos foram aplicados no modelo de regressão logística, em diversas aplicações e diferentes cenários de simulação para avaliar o seu desempenho.

A notação usada neste trabalho segue o mesmo padrão da sua principal referência (Hastie et al., 2015). Todos os vetores são definidos como vetores colunas. Somente os vetores de ordem n serão escritos em negrito, os demais como o vetor β serão escritos normalmente. Todas as matrizes serão escritas em negrito, independente de sua ordem. Para a matriz $\boldsymbol{X}$ de ordem $n \times p$, seus vetores coluna serão representados por $\boldsymbol{x}_k$ enquanto suas linhas são representados por x_i. No entanto, como todos os vetores são definidos como vetores colunas, x_i representa um vetor coluna cujos elementos são aqueles da i-ésima linha da matriz $\boldsymbol{X}$.

Este trabalho está organizado da seguinte forma. O Capítulo 2 descreve as principais informações sobre o lasso: sua implementação, suas vantagens e diversos aspectos teóricos por trás do método. No Capítulo 3 abordamos o modelo de regressão logística, falamos sobre suas principais características e descrevemos a estimação de seus parâmetros pelo método lasso. Em seguida, no Capítulo 4 apresentamos e detalhamos todo o estudo de simulação que foi desenvolvido neste trabalho, bem como os resultados e discussões sobre o mesmo. No Capítulo 5 são apresentadas, analisadas e discutidas 9 bases de dados que foram utilizadas como aplicação para comprovar os resultados vistos da simulação na prática. Por fim no último capítulo, Capítulo 6, são apresentadas as conclusões feitas sobre todo o estudo teórico, simulações e aplicações que foram desenvolvidas neste trabalho, bem como são abordados possíveis trabalhos futuros.

Capítulo 2

Lasso

O modelo linear geral é definido como

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \tag{2.1}$$

em que y_i representa o valor da variável resposta ou dependente da observação i ($i = 1,2,\ldots,n$) , x_{ij} é o valor da covariável ou variável independente j ($j = 1,2,\ldots,p$) do indivíduo i, β_0 e $\beta = (\beta_1,...,\beta_p)^T$ são os parâmetros desconhecidos do modelo e $\boldsymbol{\epsilon}$ é o vetor que contém os erros associados a cada observação. O erro é a variável aleatória que representa a variabilidade não explicada pelas covariáveis presentes no modelo.

Ao definir o modelo da Equação (2.1) os seguintes pressupostos são assumidos: todos os erros são independentes e identicamente distribuídos, com média 0 e variância constante. Além disso, na Equação (2.1) assumimos que a relação entre a média da variável resposta e cada uma das covariáveis é linear nos parâmetros. Usualmente, também é assumido que os erros tem distribuição normal.

Para encontrar as estimativas dos parâmetros do modelo linear geral frequentemente é utilizado o método de mínimos quadrados, onde devemos minimizar a soma de quadrados dos erros. Para isso basta encontrar a solução da seguinte função objetivo:

$$\underset{\beta_0,\beta}{\text{minimize}} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2. \tag{2.2}$$

Na literatura usualmente o método de minímos quadrados não é definido com a constante $\frac{1}{2n}$. Porém, neste trabalho será definido com essa constante, pois ela é conveniente para definição do lasso que é feita na Equação (2.3). Frisa-se que a presença da constante nada altera no estimador obtido por este método.

Em cenários de alta dimensionalidade, quando $p > n$, a solução encontrada pelo método de mínimos quadrados se torna um problema, pois ela não é única, ou seja, existem infinitas soluções para a função objetivo. Outro problema é a falta de esparsidade

do método que pode criar modelos com muitos parâmetros e de difícil interpretação. Para lidar com esses problemas, o lasso foi introduzido, sendo que sua função objetivo é dada por

$$\underset{\beta_0,\beta}{\text{minimize}} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \; sujeito\ a\ \|\beta\|_1 \leq t, \tag{2.3}$$

em que $\|\beta\|_1 = \left[\sum_{j=1}^{p} |\beta_j|\right]^{\frac{1}{1}}$ é o que chamamos de norma $L1$ (na notação adotada $\|\beta\|_k = (\sum_{j=1}^{p} |\beta_j|^k)^{\frac{1}{k}}$) e t é conhecido como parâmetro de penalização, sendo que é através dele que o número de covariáveis do modelo é controlado. O parâmetro de penalização na equação é assumido como conhecido e positivo, e um critério para sua estimação será discutido na Seção 2.3.

A função objetivo também pode ser reescrita no que chamamos de forma Lagrangiana (Hastie et al., 2009)

$$\underset{\beta_0,\beta}{\text{minimize}} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_1, \tag{2.4}$$

em que $\boldsymbol{y}$ é o vetor contendo as respostas das n observações, $\boldsymbol{X}$ é a matriz de ordem $n \times p$ na qual cada linha representa os valores das p covariáveis para cada observação, $\|\boldsymbol{x}\|_2^2 = \left[\left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}}\right]^2 = \sum_{i=1}^{n} x_i^2$. Na Equação (2.4) temos $\lambda \geq 0$. Observe que na forma Lagrangiana não aparece o intercepto. Isto é uma conveniência da forma Lagrangiana e será discutido melhor na Seção 2.4.

Há uma correspondência de um para um entre a restrição imposta na Equação (2.3) e na Equação (2.4). Isto é, para cada valor assumido de t em um intervalo de valores de t que não leva aos estimadores do lasso e de mínimos quadrados coincidirem, vai existir um valor correspondente de λ que garante a mesma solução pela forma Lagrangiana (Hastie et al., 2015). Outro ponto a ser notado é que quanto maior o valor que t assume na Equação (2.3), mais próxima a função objetivo está do método de mínimos quadrados. Os dois estimadores vão coincidir se t for maior que a soma dos coeficientes estimados por mínimos quadrados. Já na Equação (2.4) quanto menor o valor de λ, mais próxima a estimação se torna do método de mínimos quadrados, e ambos irão coincidir se $\lambda = 0$. Isto se deve ao fato de que se $\lambda = 0$ e $t \to \infty$, não estaremos impondo nenhuma restrição na função objetivo.

A forma Lagrangiana possui certa conveniência numérica, pois permite o uso do procedimento conhecido como coordenada descendente, que facilita a solução da equação objetivo e consequentemente a estimação dos parâmetros. Entraremos em mais detalhes sobre ele na Seção 2.4.

É comum encontrarmos na literatura a equação objetiva do lasso definida sem a constante $\frac{1}{2n}$, ou ainda substituída simplesmente por $\frac{1}{n}$ ou $\frac{1}{2}$. Essas são apenas reparametrizações da equação e que, na prática, não interferem na estimação do vetor β. Porém, o uso da constante $\frac{1}{2n}$ ou $\frac{1}{n}$ é interessante, pois torna os valores de λ comparáveis entre diferentes tamanhos amostrais, o que é útil, por exemplo, para a escolha de λ usando validação cruzada (Hastie et al., 2015).

Para estimar os parâmetros do modelo através do lasso, primeiro devemos padronizar as covariáveis, de maneira que elas sejam centradas em zero e possuam variância unitária. Devemos fazer isso para que a estimação não seja afetada pela unidade de medida nem pela ordem de grandeza das mesmas. Uma covariável, por exemplo, pode estar em metros, outra em quilômetros e uma terceira pode ser adimensional com ordem de grandeza muito diferente das demais.

2.1 Norma L1

Um dos importantes questionamentos a se fazer é o porque do lasso utilizar a norma $L1$ e não a norma $L2$ ($\sum_{j=1}^{p} \beta_j^2$), por exemplo. Isto se deve ao fato da norma $L1$ carregar uma propriedade que faz com que muitos parâmetros fiquem com estimativas iguais a zero, reduzindo assim o número de covariáveis do modelo.

Usando a norma ou aplicando a penalização $L2$ na função objetivo, ela assume a seguinte forma:

$$\underset{\beta_0,\boldsymbol{\beta}}{\text{minimize}} \frac{1}{2n} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \; sujeito\ a \; \sum_{j=1}^{p} \beta_j^2 \leq t^2. \tag{2.5}$$

Quando estimamos os parâmetros usando a norma $L2$, denominamos o modelo de regressão ridge. Maiores detalhes dessa técnica podem ser encontrados em Hoerl et al. (1975).

Para mostrar melhor como a norma $L1$ consegue fazer com que muitos parâmetros sejam nulos vamos comparar graficamente o que acontece com os coeficientes usando a norma $L1$ e a norma $L2$ (Hastie et al., 2015). Na Figura 2.1 temos dois gráficos representando o conjunto de solução para uma regressão com duas covariáveis, aplicando o lasso e a regressão ridge. Como podemos ver, a região da norma $L1$ assume uma forma quadrada, enquanto a norma $L2$ redonda. As elipses apresentadas na figura correspondem aos contornos que apresentam mesmo valor para a soma de quadrados dos resíduos ordinários. A solução para ambos os métodos é dada quando o contorno da elipse da soma de quadrados dos resíduos atinge o primeiro ponto do conjunto de solução. A norma $L1$ possui arestas em sua solução, de forma que, quando essa solução é atingida, um dos parâmetros é estimado como zero. Já a norma $L2$, por não possuir essas arestas,

dificilmente seu conjunto solução atingirá uma solução onde um dos parâmetros é nulo. Logo, o lasso, ao contrário da regressão ridge, possui essa propriedade de criar esparsidade dentro do modelo, encontrando, em geral, um conjunto de solução onde vários parâmetros são estimados como zero.

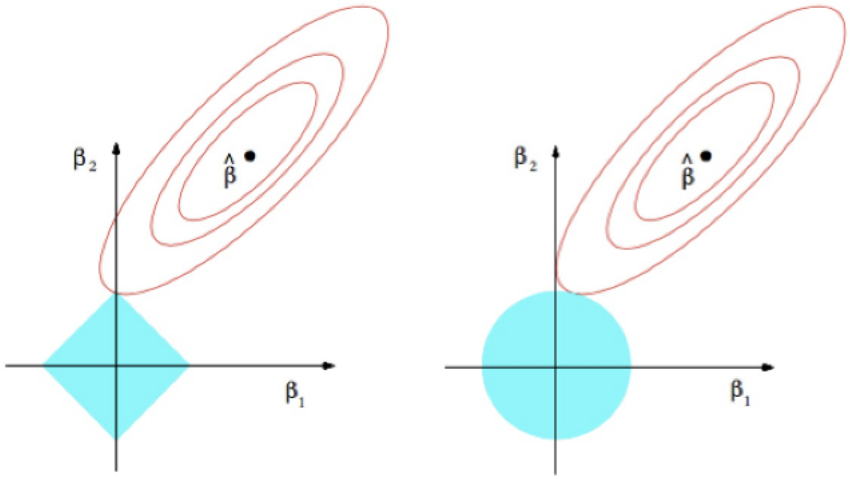

Figura 2.1: *Gráfico dos contornos da soma de quadrados dos resíduos para a regressão ridge e o lasso, e o conjunto solução para λ com 2 covariáveis.*

À esquerda temos a regressão lasso e à direita a regressão ridge. As áreas sólidas em azul representam o conjunto solução de λ para cada regresão $|\beta_1|+|\beta_2| \leq \lambda$ e $\beta_1^2+\beta_2^2 \leq \lambda^2$ respectivamente. Já as elipses são os contornos que apresentam mesmo valor para a soma de quadrado dos resíduos. Fonte: Hastie et al. (2015).

2.2 Balanço viés-variância

Uma questão importante no ajuste de modelos de regressão é o controle do viés e da variância dos estimadores do modelo. O viés do estimador da esperança da variável resposta, chamado aqui simplesmente de viés do modelo, mede o quão próximas, em média, as estimativas do modelo ajustado estão das verdadeiras respostas. Quanto menor o viés no modelo, mais próximo, em média, os valores estimados $\hat{\boldsymbol{y}}$ estão da média condicional de $\boldsymbol{y}|\boldsymbol{X}$. Apesar de parecer que o melhor modelo é sempre aquele cujo viés é zero, isto não é necessariamente verdade. Um modelo com viés zero pode ter superajuste, isto é, um modelo cuja as estimativas $\hat{\boldsymbol{y}}$ são próximas aos valores observados de $\boldsymbol{y}$, dando a falsa sensação de ótimo ajuste do modelo, quando na verdade isto pode ter ocorrido apenas naquela amostra de dados. As previsões da variável resposta baseada nesse modelo em uma outra amostra de dados estão, em geral, distantes dos valores observados. Isto se deve ao fato de que um modelo com viés baixo pode ter uma variância alta em seus estimadores e portanto baixa capacidade de generalização.

Para exemplificar, vamos supor o seguinte exemplo, baseado em Izbicki e dos Santos (2020). Foram geradas 10 observações seguindo o seguinte modelo:

$$y_i = \beta_0 + \sum_{j=1}^{4} x_i^j \beta_j + \epsilon_i, \;\; \epsilon \sim N(0,1), \;\; \boldsymbol{x} \sim U(0;1). \tag{2.6}$$

Em seguida foram ajustadas 2 modelos polinomiais com $p = 2$ e 4 covariáveis. Na Figura 2.2, temos o gráfico das curvas ajustadas e do modelo verdadeiro. Como podemos ver, apesar da regressão de grau 4 ser a verdadeira e portanto apresentar viés zero, a regressão ajustada de grau 2 está mais próxima do verdadeiro modelo. Isto se deve ao fato da regressão ajustada com 4 covariáveis apresentar super-ajuste naquela amostra.

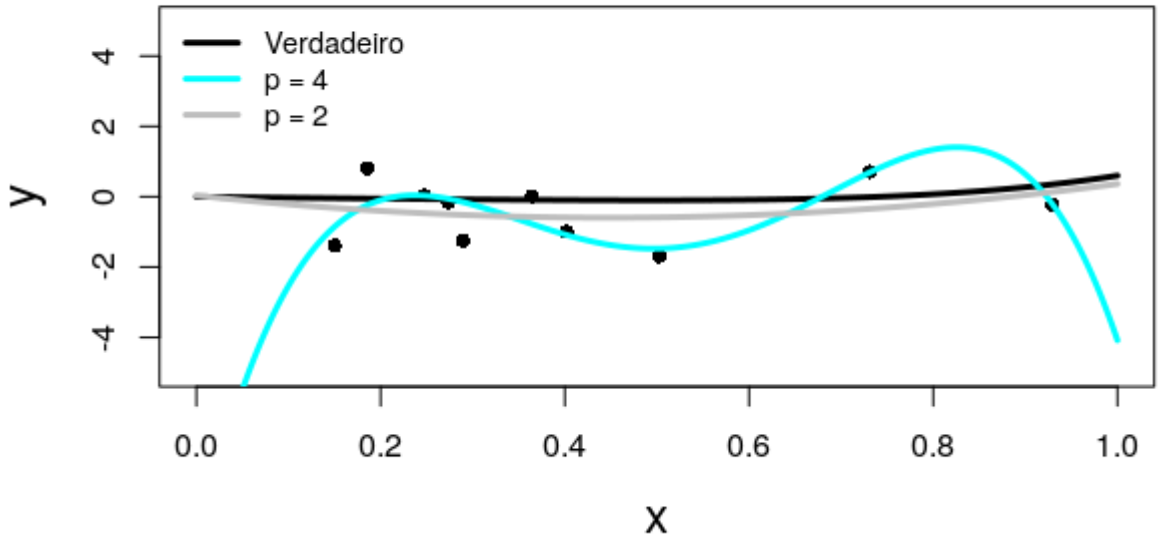

Figura 2.2: *Gráfico dos modelos polinomiais ajustados com 2 e 4 covariáveis.*

A variância do modelo (variância do estimador da esperança da variável resposta) nos mostra o quanto o modelo consegue se adaptar às mudanças de amostra. Se um modelo possui baixa variância ele pouco varia suas estimativas a cada mudança na amostra. Um modelo com variância muito pequena é ruim por ser pouco flexível. Por outro lado um modelo com variância muito grande também é ruim, porque muda suas estimativas consideravelmente se pequenas alterações na amostra são feitas. A variância presente no modelo aumenta se o número de covariáveis presentes também aumentam, conforme mostramos no exemplo a seguir (Izbicki e dos Santos, 2019). Foram geradas 20 amostras com 50 observações, seguindo o seguinte modelo:

$$Y|(X = x) \sim N(45tanh(x/1,9 - 7) + 57; 4^4), \;\; x \sim U(8;18). \tag{2.7}$$

Em seguida, para cada amostra, foi ajustado o seguinte modelo considerando-se $p = 2$, 10 e 30:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_i^j + \epsilon_i. \tag{2.8}$$

Nas Figuras 2.3 à 2.5 temos os gráficos das 20 retas ajustadas para cada amostra, sendo que a reta em vermelho corresponde ao ajuste da amostra cujas observações são apresentadas no gráfico. Podemos ver claramente o quanto o modelo de regressão varia em função da amostra gerada. Quanto maior o valor de p (mais covariáveis presentes), maior é a variância do modelo ajustado.

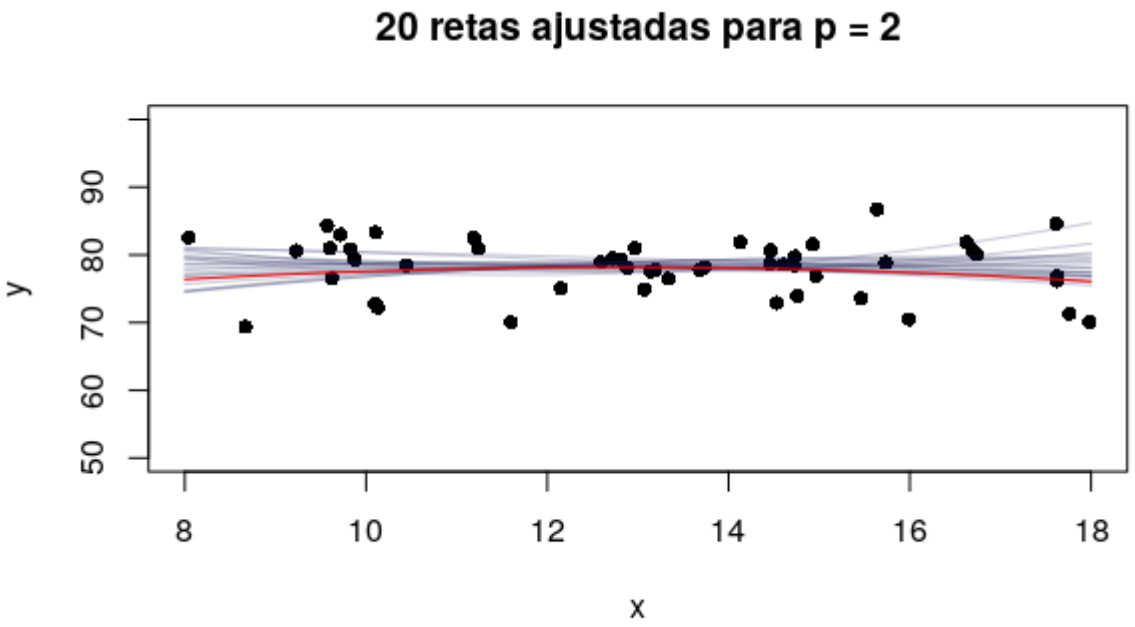

Figura 2.3: *20 retas ajustadas para $p = 2$.*

Uma das garantias téoricas da técnica de mínimos quadrados é ter os estimadores não viesados. Porém, os modelos podem ter grande variância. O lasso surge então como uma opção para diminuir a variância do modelo introduzindo um pouco de viés no mesmo. Essa troca entre viés e variância é controlada através do parâmetro de penalização, como vimos na Equação (2.4). Quanto menor o valor de λ, mais covariáveis entram no modelo, diminuindo assim o viés, mas aumentando a variância. No sentindo contrário, quanto maior o valor de λ, menos covariáveis no modelo e portanto menor variância e maior viés. O ideal é encontrar um valor para λ que consiga balancear o viés e a variância presente no modelo, possibilitando assim a obtenção de um modelo com bom poder preditivo.

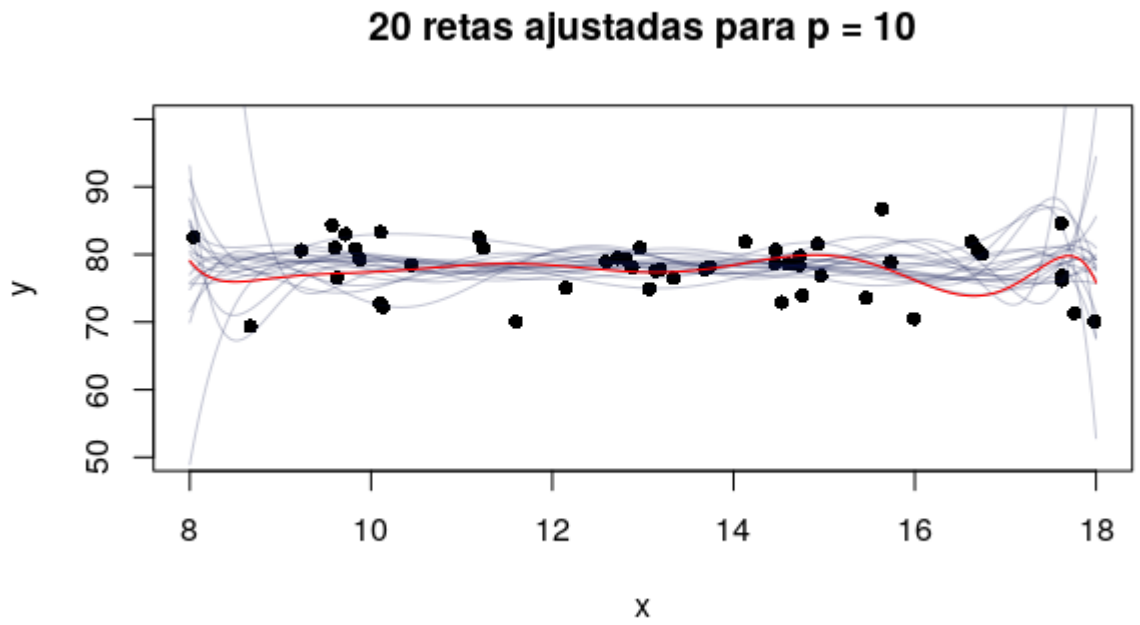

Figura 2.4: *20 retas ajustadas para p = 10.*

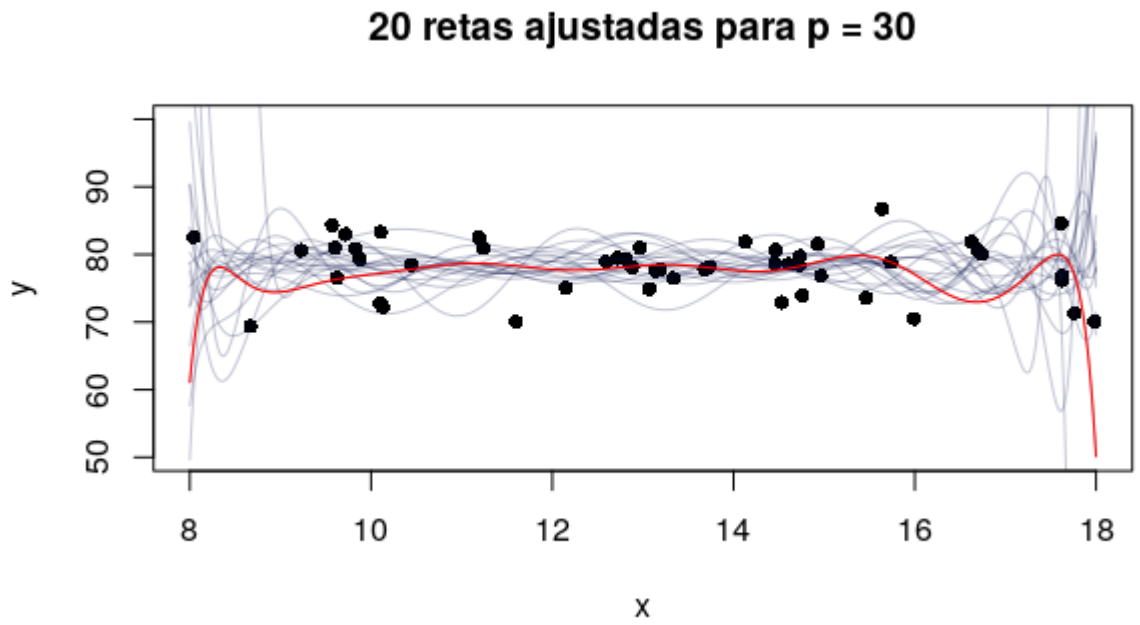

Figura 2.5: *20 retas ajustadas para p = 30.*

2.3 Validação cruzada

Como visto, o parâmetro de penalização λ tem um papel muito importante na regressão via lasso. Logo estimá-lo corretamente é extremamente importante. O método mais utilizado para estimar λ é denominado validação cruzada (Hastie et al., 2015).

O método consiste em criar novas amostras artificiais a partir da base de dados original para, através de divisões aleatórias, fazer a estimação da performance do modelo em cada uma das amostras sob diferentes valores de λ e encontrar aquele valor que maximiza a performance do modelo.

Mais precisamente a base de dados é dividida aleatoriamente em $k > 1$ grupos. Um destes k grupos é fixado como teste, para validar a performance do modelo, e os demais $k-1$ grupos são usados para estimação do modelo. Primeiro devemos construir o modelo usando o conjunto de treinamento, os $k-1$ grupos que restaram, para diferentes valores de λ. O grupo que ficou para teste é usado para medir o desempenho do modelo, por exemplo, estimar o Erro Quadrático Médio de Predição (EQMP). O EQMP é uma das medidas mais utilizadas para avaliar a performance de predição de um modelo. Ele mostra em média o quão próximo as estimativas do modelo estão dos verdadeiros valores, sendo calculado como

$$EQMP = \sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}. \tag{2.9}$$

Em seguida o procedimento é repetido k vezes, de maneira que todos os k grupos tenham a chance de fazer parte do grupo de teste. No final, para cada valor de λ dentro do intervalo considerado teremos k estimativas do EQMP. Calculamos então a média do EQMP para cada valor de λ e construímos um gráfico (curva de erro da validação cruzada, CEVC) em que colocamos λ (no pacote GLMNET é plotado o logaritmo de λ) no eixo x e a média do EQMP no eixo y. Através da CEVC conseguimos observar qual valor do parâmetro de penalização que maximiza a performance do modelo, ou seja para qual valor de λ o EQMP é menor.

Na Figura 2.6 temos a CEVC, feita através do pacote GLMNET, descrito na Seção 2.5. Este exemplo de regressão linear foi construído a partir da base de dados chamada *marketing*, retirado do pacote Datarium (Kassambara, 2019) no R (R Core Team, 2020). Nela podemos perceber que para valores bem pequenos de λ, a média da estimativa do EQMP é maior, e conforme λ cresce ela vai diminuindo até atingir o ponto de melhor λ, depois a mesma voltar a crescer. Esse padrão é observado na maioria dos bancos de dados. No gráfico existem dois pontos destacados como melhor λ. O primeiro representa aquele do qual provém o menor EQMP e o outro fornece o modelo mais regularizado (com menos covariáveis), que apresenta EQMP menor que o do modelo com menor EQMP mais um erro padrão.

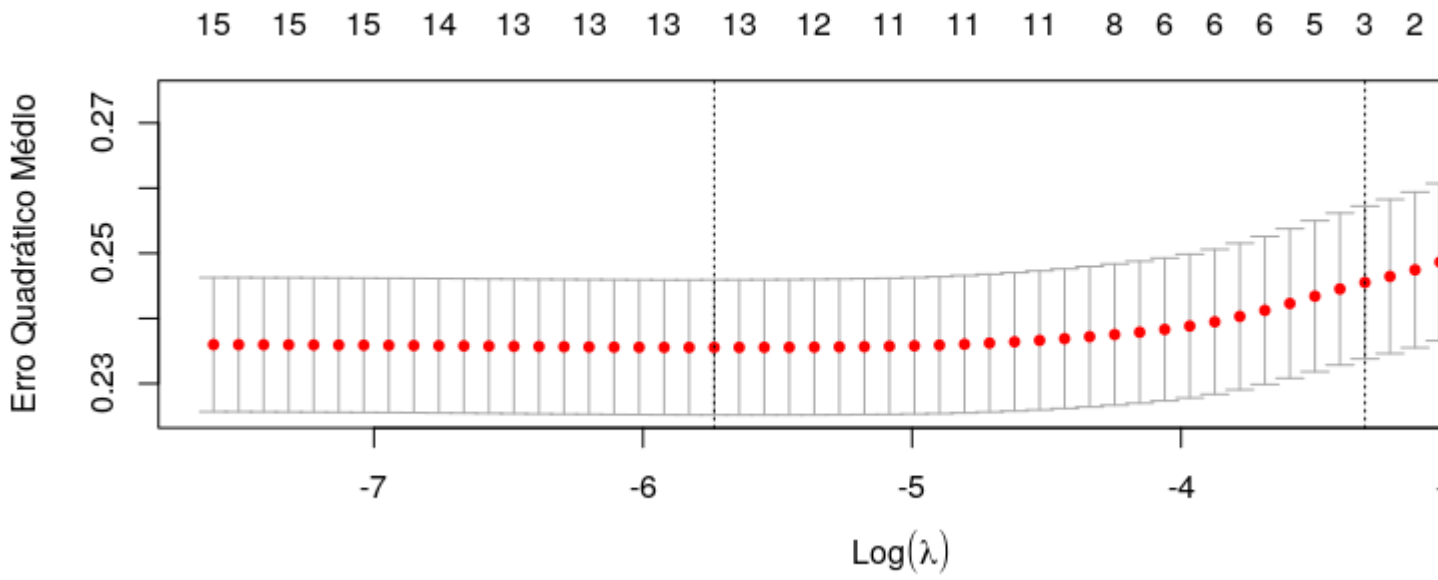

Figura 2.6: *Gráfico da curva de erro da validação cruzada feita através do pacote GLMNET.*

2.4 Coordenada descendente

A Equação (2.4) é um problema que pode ser resolvido através de uma programação quadrática com restrição convexa. E entre os diversos programas e métodos capazes de solucionarem nossa equação objetivo, um em específico é muito utilizado para construir o algoritmo de solução do lasso, por ser simples e bastante efetivo: o coordenada descendente (Hastie et al., 2015).

O primeiro passo na estimação via lasso, por coordenada descendente, é padronizar as covariáveis presentes no modelo, isto é, cada coluna da matrix $\boldsymbol{X}$ deve ter média zero e variância igual a um. Como já discutido anteriormente, padronizando as covariáveis conseguimos garantir que a estimação de seus coeficientes associados não sejam dependentes de sua unidade de medida. Também assumimos que o vetor de respostas $\boldsymbol{y}$ esteja centralizado, isto é, com média igual a zero. Centralizando $\boldsymbol{y}$ e padronizando as covariáveis, podemos convenientemente omitir a constante β_0 de nossa equação, pois necessariamente ela é estimada com o valor zero.

Consideremos agora, a forma Lagrangiana do lasso definida na Equação (2.4) para apenas uma covariável e com suas devidas padronizações. Isto é

$$\underset{\beta}{\text{minimize}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - z_i\beta)^2 + \lambda||\beta|| \tag{2.10}$$

onde z_i é o elemento de posição i do vetor $\boldsymbol{z}$ que corresponde a única covariável já padronizada, com média zero e variância um.

Como estamos enfrentando um problema comum de minimização de uma função, o procedimento padrão seria encontrar a primeira derivada da função com relação a β e igualar a zero. Entretanto como estamos trabalhando com módulo, pela definição de derivada de função módulo, temos que

$$f(x) = |x| \rightarrow f'(x) = \begin{cases} -1 & \text{se } x < 0, \\ 1 & \text{se } x \geq 0. \end{cases}$$

Por definição se as derivadas laterais, isto é, as derivadas considerando os dois casos $\beta < 0$ e $\beta \geq 0$ não coincidirem no ponto $\beta = 0$, então a derivada não existirá nesse ponto (Anton et al., 2014). Desenvolvendo a derivada da função objetivo definida na Equação (2.10) considerando esses dois casos, $\beta < 0$ e $\beta \geq 0$, temos que

$$\frac{\partial \frac{1}{2n}\sum_{i=1}^{n}(y_i - z_i\beta)^2 + \lambda\|\beta\|}{\partial\beta} = \begin{cases} \frac{1}{2n}\sum_{i=1}^{n} 2(\beta z_i^2 - y_i z_i) + \lambda & \text{se } \beta \geq 0, \\ \frac{1}{2n}\sum_{i=1}^{n} 2(\beta z_i^2 - y_i z_i) - \lambda & \text{se } \beta < 0. \end{cases}$$

Como dito anteriormente z_i representa a covariável padronizada, então $\frac{1}{n}\sum_{i=1}^{n} z_i^2 = 1$, dessa forma podemos reescrever as derivadas como

$$\frac{\partial \frac{1}{2n}\sum_{i=1}^{n}(y_i - z_i\beta)^2 + \lambda\|\beta\|}{\partial\beta} = \begin{cases} \beta - \frac{1}{n}\sum_{i=1}^{n}(y_i z_i) + \lambda & \text{se } \beta \geq 0, \\ \beta - \frac{1}{n}\sum_{i=1}^{n}(y_i z_i) - \lambda & \text{se } \beta < 0. \end{cases}$$

As derivadas laterais não coincidiram. Logo, nossa função objetivo não possui derivada no ponto $\beta = 0$, o que torna sua solução um pouco mais complexa. Entretanto, pela inspeção de sua função objetivo e observação de sua derivada, podemos perceber que

$$\hat{\beta} = \begin{cases} \frac{1}{n}\sum_{i=1}^{n}(y_i z_i) - \lambda & \text{se } \lambda < \frac{1}{n}\sum_{i=1}^{n}(y_i z_i), \\ 0 & \text{se } \lambda \geq \frac{1}{n}|\sum_{i=1}^{n}(y_i z_i)|, \\ \frac{1}{n}\sum_{i=1}^{n}(y_i z_i) + \lambda & \text{se } \lambda < -\frac{1}{n}\sum_{i=1}^{n}(y_i z_i). \end{cases}$$

Usando a notação usual de produto interno de vetores para denotar $\sum_{i=1}^{n}(y_i z_i)$ como $\langle \boldsymbol{z}, \boldsymbol{y} \rangle$, podemos reescrever o estimador β como

$$\hat{\beta} = \begin{cases} \frac{1}{n}\langle \boldsymbol{z}, \boldsymbol{y} \rangle - \lambda, & \text{se } \frac{1}{n}\langle \boldsymbol{z}, \boldsymbol{y} \rangle > \lambda, \\ 0, & \text{se } \frac{1}{n}|\langle \boldsymbol{z}, \boldsymbol{y} \rangle| \leq \lambda, \\ \frac{1}{n}\langle \boldsymbol{z}, \boldsymbol{y} \rangle + \lambda, & \text{se } \frac{1}{n}\langle \boldsymbol{z}, \boldsymbol{y} \rangle < -\lambda, \end{cases}$$

note que se $\lambda = 0$, o estimador coincide com o estimador de mínimos quadrados que nesse caso é dado por $\hat{\beta}_{MQ} = \frac{1}{n}\langle \boldsymbol{z}, \boldsymbol{y} \rangle$.

Nas Figuras 2.7 e 2.8 apresentamos dois exemplos com dados simulados para ilustrar

a relação entre $\hat{\beta}$ e λ. No primeiro exemplo temos que $\frac{1}{n}\sum_{i=1}^{n}(y_i z_i) > 0$ e no segundo $\frac{1}{n}\sum_{i=1}^{n}(y_i z_i) < 0$. Dessa forma podemos ver os dois possíveis comportamentos da função $\hat{\beta}$. O ponto de mudança de comportamento da função é justamente quando $\lambda = \frac{1}{n}\sum_{i=1}^{n}(y_i z_i)$.

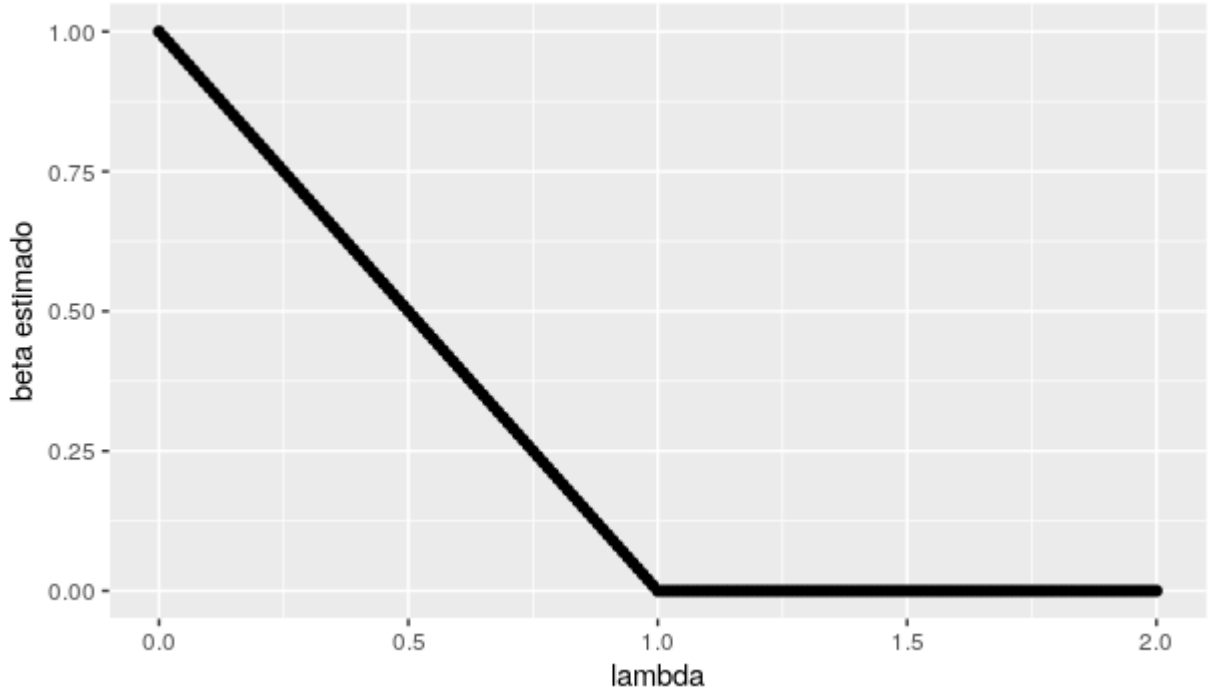

Figura 2.7: *Relação entre $\hat{\beta}$ e λ quando $\frac{1}{n}\sum_{i=1}^{n}(y_i z_i) > 0$.*

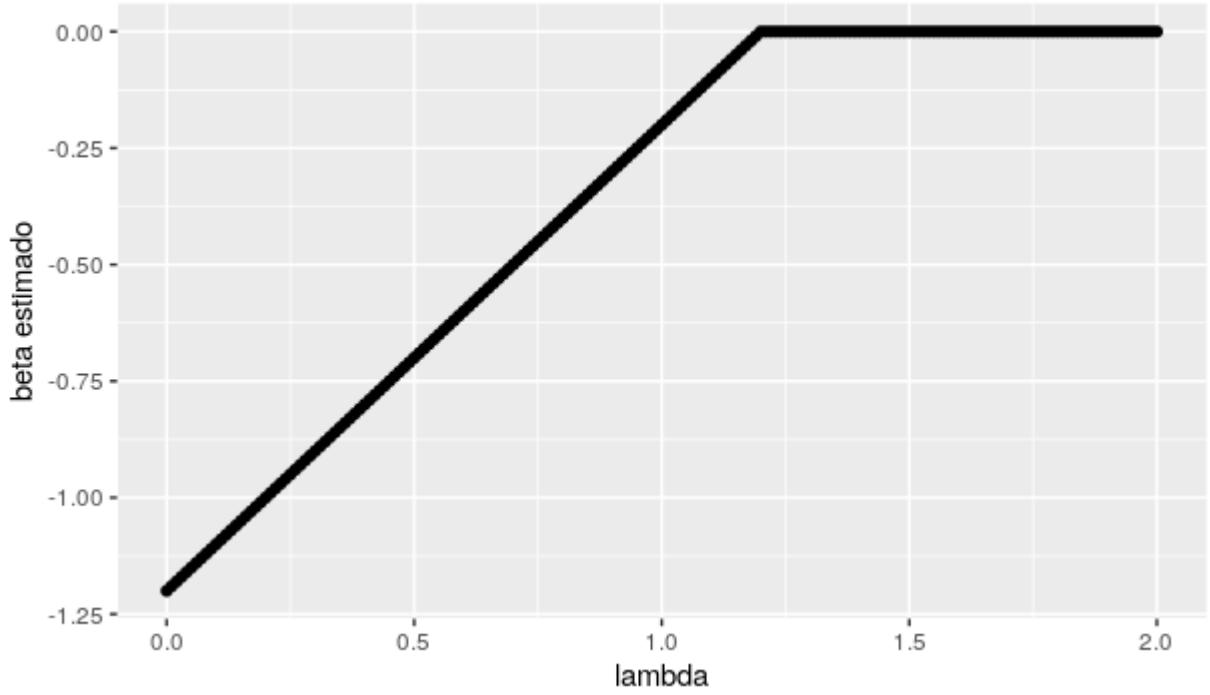

Figura 2.8: *Relação entre $\hat{\beta}$ e λ quando $\frac{1}{n}\sum_{i=1}^{n}(y_i z_i) < 0$.*

Analisando graficamente o comportamento da função, podemos ver que na estimação via lasso se $\frac{1}{n}\sum_{i=1}^{n}(y_i z_i)$, parte que representa a estimação via mínimos quadrados, é po-

sitiva e maior do λ, então a estimativa do coeficiente via lasso se resume a estimação via minimos quadrados reduzida pelo parâmetro de penalização. E se $\frac{1}{n}\sum_{i=1}^{n}(y_i z_i)$ é negativo e menor do que λ, a estimativa via lasso é a estimativa de mínimos quadrados reduzida em módulo pelo parâmetro de penalização. Quando a estimativa de mínimos quadrados em módulo é menor que λ, então, a estimativa de β via lasso é zero.

Podemos reescrever $\hat{\beta}$ de forma mais sucinta, através do seguinte operador (Hastie et al., 2015):

$$S_\lambda(x) = sinal(x)(|x| - \lambda)_+, \tag{2.11}$$

em que a função t_+ é o próprio valor de t se $t > 0$ e 0 caso contrário. Essa função transforma x em zero se |x|$\leq \lambda$ e leva x em direção a zero caso contrário. Apenas reescrevendo a equação chegamos a

$$S_\lambda(x) = \begin{cases} x - \lambda & \text{se } x > \lambda, \\ 0 & \text{se } |x| \leq \lambda, \\ x + \lambda & \text{se } x < -\lambda. \end{cases}$$

Dessa forma, usando esse operador para definir o estimador de β pelo método lasso temos que

$$\hat{\beta} = S_\lambda\left(\frac{1}{n}\langle \boldsymbol{z}, \boldsymbol{y}\rangle\right). \tag{2.12}$$

Essa é a solução para o caso com apenas uma covariável. Intuitivamente, um esquema ciclíco pode ser criado para o caso do lasso com mais covariáveis, em que uma coordenada ou parâmetro é atualizado por vez. Primeiro devemos criar uma ordem fixa para atualizar os parâmetros. Apesar de ser fixa essa ordem deve ser construída aleatoriamente. Suponha que estamos no j-ésimo passo, onde o parâmetro β_j será atualizado, enquanto todos os outros continuam fixados. Reescrevendo nossa função objetivo dada na Equação (2.4) obtemos

$$\underset{\boldsymbol{\beta}}{\text{minimize}}\ \frac{1}{2n}\sum_{i=1}^{n}(y_i - \sum_{k\neq j} x_{ik}\beta_k - x_{ij}\beta_j)^2 + \lambda|\beta_j| + \lambda\sum_{k\neq j}|\beta_k|. \tag{2.13}$$

Podemos ainda reescrever essa função em termos dos resíduos parciais, $r_i^{(j)} = y_i - \sum_{k\neq j} x_{ik}\hat{\beta}_k$. Logo em termos do resíduo parcial o j-ésimo coeficiente pode ser atualizado como

$$\hat{\beta}_j = S_\lambda\left(\frac{1}{n}\langle \boldsymbol{x_j}, \boldsymbol{r^{(j)}}\rangle\right), \tag{2.14}$$

em que $\boldsymbol{r^{(j)}}=(r_1^{(j)}, r_2^{(j)}, \ldots, r_n^{(j)})$. Equivalentemente, a atualização pode ser escrita como

$$\hat{\beta}_j \leftarrow S_\lambda\left(\hat{\beta}_j + \frac{1}{n}\langle \boldsymbol{x_j}, \boldsymbol{r}\rangle\right), \tag{2.15}$$

em que $\boldsymbol{r} = (r_1, r_2, \ldots, r_n)^\top$ e $r_i = y_i$ - $\sum_{j=1}^{p} x_{ij}\hat{\beta}_j$. O algoritmo é repetido de forma ciclíca através da ordem escolhida anteriormente, até que não haja mudanças na atualização dos parâmetros considerando-se um limite estabelecido.

O algoritmo coordenada descendente para o lasso é bastante rápido devido as coordenadas serem bastante explícitas, como podemos ver na Equação (2.14). Outro ponto positivo é que por criar uma solução convexa, para valores grandes de λ, muitos coeficientes serão nulos o que facilita ainda mais o método.

Na prática, encontrar a melhor solução para o lasso está submetido a escolha de um valor fixo para λ que nos leva ao melhor resultado sendo um critério pré-definido. Para isto devemos executar o algoritmo de estimação não sob apenas um valor de λ, mas sim para um intervalo de possíveis valores. Nosso interesse é escolher um intervalo de possíveis valores para λ de maneira que possamos encontrar sua melhor estimativa de forma rápida. Uma das maneiras de se fazer isso é através do método conhecido como: *pathwise coordinate descent*. Este método consiste em partir de um valor inicial para λ grande o suficiente de forma que a sua primeira solução ótima para os coeficientes seja um vetor nulo. Este valor seria: $\lambda_{max} = max_j|\frac{1}{n}\langle \boldsymbol{x_j}, \boldsymbol{y}\rangle|$. Em seguida ir diminuindo o valor de λ e aplicando o coordenada descendente até a sua perfeita convergência (Hastie et al. , 2015). Dessa forma percorremos um intervalo grande de λ e de uma forma que seja computacionalmente eficiente.

2.5 GLMNET

Toda a parte computacional desenvolvida neste trabalho foi feita através do software de domínio publico R (R Core Team, 2020). A estimação via lasso foi feita através de um dos principais pacotes no R chamado GLMNET desenvolvido por Friedman et al. (2010).

O pacote já vem configurado para ajustar a regressão via lasso na maioria das distribuições da familia exponencial (Nelder e Wedderburn, 1972) como o modelo binomial, poisson, gaussiano e muitos outros. Recentemente ele recebeu uma nova atualização, sendo possível agora ajustar novos modelos como os que consideram funções de ligação probito e complemento log-log no caso de variáveis respostas binárias.

O termo de penalização do pacote foi construído não apenas para a penalização via

lasso, mas sim na forma de penalização via elastic net (Zou e Hastie, 2005), que é uma generalização do lasso. Para um modelo qualquer, a equação objetivo via elastic net é dada por

$$\underset{\beta_0,\boldsymbol{\beta}}{\text{minimize}} \frac{-1}{N} l(\boldsymbol{y};\beta_0,\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \gamma_j \Big[(1-\alpha)\beta_j^2 + \alpha|\beta_j|\Big]. \tag{2.16}$$

É fácil ver que, na utilização do pacote, se definirmos os parâmetros α e γ_j iguais a 1 para todo j então temos a penalização via lasso.

Sobre o parâmetro de penalização, o pacote também faz o uso da técnica de validação cruzada para encontrar o melhor valor de λ para o modelo. Ele possui diferentes critérios, em sua base, para definir qual o melhor modelo como: área sobre a curva roc (Myerson et al., 2001), erro quadrático médio, entre outros. O padrão do pacote utiliza um intervalo de 100 valores de λ (na escala logarítmica) na validação cruzada, mas o mesmo também permite ao usuário a liberdade de definir essa quantidade no intervalo. Outro ponto que o pacote permite a escolha do usuário é em relação ao número de grupos na validação cruzada, sendo que o padrão é a criação de 10 grupos.

A estimação dos parâmetros do modelo é feita através do algoritmo coordenada descendente discutido na Seção 2.4. Por padrão, o pacote, ao fazer a estimação, centraliza a variável resposta e padroniza as covariáveis do modelo. Entretanto, as estimativas dos parâmetros fornecidas pelo pacote são na escala original da variável resposta e das covariáveis. Se fossemos realizar manualmente os procedimentos para retornar as estimativas para a escala original de $\boldsymbol{y}$ e $\boldsymbol{X}$, seriam necessário dois passos. Primeiro devemos ajustar as estimativas dos parâmetros de forma que os valores ajustados fiquem na escala original de $\boldsymbol{y}$ mesmo com as covariáveis padronizadas. Como $\boldsymbol{y}$ foi apenas centralizada, precisamos apenas somar a média amostral de $\boldsymbol{y}$ ($\overline{y}$) em cada valor ajustado. Para isso, basta somar $\overline{y}$ à estimativa do intercepto. Porém, conforme discutido na Seção 2.4, ao centralizarmos $\boldsymbol{y}$ e padronizarmos as covariáveis, β_0 pôde ser excluído do processo de estimação. Sendo assim, sua estimativa é necessariamente zero. Dessa forma, para que os valores ajustados fiquem na escala original de $\boldsymbol{y}$ mesmo com as covariáveis padronizadas, a estimativa do intercepto é igual a $\overline{y}$ e as estimativas dos demais parâmetros não mudam.

No segundo passo estamos interessados em encontrar os valores das estimativas dos coeficientes que associados às covariáveis na escala original ($\hat{\gamma_0}$ e $\hat{\gamma_j}$) levem aos mesmos valores ajustados que os valores das estimativas dos coeficientes associados às covariáveis na escala padronizada ($\hat{\beta_0}$ e $\hat{\beta_j}$). Dessa forma queremos encontrar $\hat{\gamma_0}$ e $\hat{\gamma_j}$ que satisfaçam para todo i

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j z_{ij} = \hat{\gamma}_0 + \sum_{j=1}^{p} \hat{\gamma}_j x_{ij}, \tag{2.17}$$

em que z_{ij} representa o valor da j-ésima covariável padronizada do i-ésimo indivíduo.

Seja $\overline{x}_j$ e s_j, respectivamente a média e o desvio padrão amostral da j-ésima covariável, então

$$\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j \left(\frac{x_{ij} - \overline{x}_j}{s_j} \right) = \left(\hat{\beta}_0 - \sum_{j=1}^{p} \frac{\hat{\beta}_j \overline{x}_j}{s_j} \right) + \sum_{j=1}^{p} \left(\frac{\hat{\beta}_j}{s_j} \right) x_j = \hat{\gamma}_0 + \sum_{j=1}^{p} \hat{\gamma}_j x_{ij}. \tag{2.18}$$

Logo, é evidente a seguinte igualdade:

$$\hat{\gamma}_0 = \hat{\beta}_0 - \sum_{j=1}^{p} \frac{\hat{\beta}_j \overline{x}_j}{s_j} \tag{2.19}$$

e

$$\hat{\gamma}_j = \frac{\hat{\beta}_j}{s_j}. \tag{2.20}$$

Considerando-se que, conforme discutido no primeiro passo, $\hat{\beta}_0 = \overline{y}$ e o resultado na Equação (2.20), podemos reescrever a Equação (2.19) como

$$\hat{\gamma}_0 = \overline{y} - \sum_{j=1}^{p} \overline{x}_j \hat{\gamma}_j. \tag{2.21}$$

O pacote em si possui diversas outras finalidades, como a possibilidade de trabalhar com a regressão ridge, outras extensões do lasso como o relaxed lasso, criação de gráficos da CEVC e muitas outras opções que são descritas em Friedman et al. (2010).

Capítulo 3

Estimação e seleção de variáveis na regressão logística

Neste capítulo será apresentado o modelo de regressão logística, que é um caso particular dos modelos lineares generalizados (Fox, 2015). Os modelos de regressão logística são muito usados atualmente em praticamente todas as áreas da ciência, pois é muito comum o interesse em estudar a relação entre uma variável resposta binária e suas covariáveis.

3.1 Modelos Lineares Generalizados

No modelo linear geral, nosso objetivo é modelar o comportamento de uma variável resposta de nosso interesse através da combinação linear de covariáveis explicativas. Para a construção do modelo assumimos certas suposições conforme apresentamos no Capítulo 2.

Propostos por Nelder e Wedderburn (1972), os Modelos Lineares Generalizados (MLG) estendem o modelo linear geral definido na Equação (2.1) em dois aspectos. O primeiro é que nos MLG uma função da média da variável resposta e não a própria média que deve ser uma função linear nos parâmetros das covariáveis. O outro aspecto é que a distribuição da variável resposta não precisar ser normal, mas pode pertencer a uma classe mais ampla de distribuições, a família exponencial (Annette, 2001). Assim, através dos MLG, diferentes tipos de variáveis respostas podem ser modelados como: contagem, binária e contínua assimétrica.

Um MLG é divido em 3 componentes; a componente aleatória, composta pelo vetor de observações $\boldsymbol{y}$ de tamanho n, em que as observações são independentes e identicamente distribuídas, provenintes de uma variável aleatória Y que possui um vetor de médias $\boldsymbol{\mu}$. O preditor linear, que é a parte sistemática do modelo e comprende as covariáveis e os seus respectivos parâmetros desconhecidos. E por último a função de ligação g que faz a ligação ou associação entre a média da variável resposta e a parte sistemática. Essa função deve ser monótona e pelo menos duplamente diferenciável.

Uma suposição dos MLG é a independência entre as observações, o que impossibilita a modelagem de bancos de dados com estruturas longitudinais, por exemplo, nos quais as unidades amostrais são medidas mais de uma vez ao longo do tempo. Porém esse problema pode ser resolvido, por exemplo, com o uso dos MLG mistos (McCulloch e Neuhaus, 2014).

3.1.1 Família exponencial

Para que a formalização dos MLG seja feita na Seção 3.1.2, precisamos definir a família exponencial de distribuições. Uma distribuição de probabilidade é dita pertencer à classe da família exponencial se sua função (densidade) de probabilidade pode ser escrita da seguinte forma:

$$f_Y(y|\theta,\phi) = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right], \tag{3.1}$$

em que $y \in 0,1$ é variável de interesse, θ é o parâmetro de localização, ϕ o parâmetro de escala, associado à variância do modelo. As funções a, b e c são funções específicas que determinam o modelo. Alguns exemplos de modelos que pertencem a família exponencial: Poisson, Bernoulli, Normal, Gama, Pascal e muitos outros.

Entre as propriedades dos modelos da família exponencial, temos que, a média da variável de interesse Y pode ser obtida através da primeira derivada da função $b(\theta)$

$$E(Y) = \mu = b'(\theta) = \frac{db(\theta)}{d\theta}. \tag{3.2}$$

Outra propriedade importante é que a variância pode ser obtida da seguinte forma:

$$V(Y) = \sigma^2 = b''(\theta)a(\phi) = \frac{d^2b(\theta)}{d\theta^2}a(\phi). \tag{3.3}$$

3.1.2 Modelos Lineares Generalizados

Os modelos lineares generalizados são definidos assumindo que a variável resposta pertença à família exponencial definida na Equação (3.1) juntamente com o seguinte componente sistemático (Nelder e Wedderburn, 1972):

$$g(\mu_i) = \eta_i = \beta_0 + x_i^\top \beta. \tag{3.4}$$

Toda distribuição pertencente a classe da família exponencial possui uma função de

ligação singular chamada de função de ligação canônica. Uma função de ligação é dita canônica quando $g(\boldsymbol{\mu}) = \theta$. O uso das ligações canônicas tem algumas vantagens, sendo uma delas que o algoritmo de estimação dos parâmetros é simplificado quando ela é usada. Entretanto outras funções de ligação, que não sejam as conônicas, também podem ser utilizadas para definir um MLG.

A estimação dos parâmetros dos MLG é feita numericamente, usualmente pelo método de máxima verossimilhança através de algum algoritmo iterativo, como por exemplo, o algoritmo do método de scoring de Fisher (Annette, 2001), que é derivado da expansão da série de Taylor. Um algoritmo de estimação muito utilizado é uma variação do método scoring de Fisher que é conhecido como mínimos quadrados reponderados iterativos (Green, 1984).

3.2 Regressão logística

O modelo de regressão logística é um caso particular dos MLG, no qual a variável resposta é binária, isto é, 0 ou 1. Por exemplo, na modelagem de crédito, onde estamos interessados em classificar uma série de indíviduos em bons (1) ou maus pagadores (0). Para isso assumimos que a variável resposta Y tem distribuição Bernoulli.

Seja μ_i a probabilidade da observação i assumir o valor 1. Podemos assim definir a função de probabilidade de Y da seguinte forma:

$$f_Y(y_i|\mu_i) = \mu_i^{y_i}(1-\mu_i)^{1-y_i} \quad 0 < \mu_i < 1. \tag{3.5}$$

Reescrevendo a Equação (3.5) no formato da família exponencial, obtemos

$$f_Y(y_i|\mu_i) = exp\left[y_i log\left(\frac{\mu_i}{1-\mu_i}\right) - (-log(1-\mu_i))\right]. \tag{3.6}$$

Após identificar a variável resposta devemos definir qual função de ligação utilizar. Para variáveis respostas binárias, a função de ligação logito é a mais utilizada. Quando essa função de ligação é utilizada, temos o conhecido modelo de regressão logística que pode ser escrito como

$$log\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i = \beta_0 + x_i^\top \beta. \tag{3.7}$$

Alternativamente, o modelo de regressão logística pode ainda ser expresso da seguinte forma:

$$P(Y_i = 1|x_i^\top) = \frac{exp\left(\beta_0 + x_i^\top \beta\right)}{1 + exp\left(\beta_0 + x_i^\top \beta\right)}. \tag{3.8}$$

Uma das características da regressão logística é ter um comportamento probabilístico em formato de letra S (Figura 3.1), do qual podemos observar que conforme $g(\boldsymbol{x}) = x_i^\top \beta \longrightarrow +\infty$ então $P((Y|x_i^\top) = 1) \longrightarrow 1$, e evidentemente $g(\boldsymbol{x}) = x_i^\top \beta \longrightarrow -\infty$ então $P((Y|x_i^\top) = 1) \longrightarrow 0$.

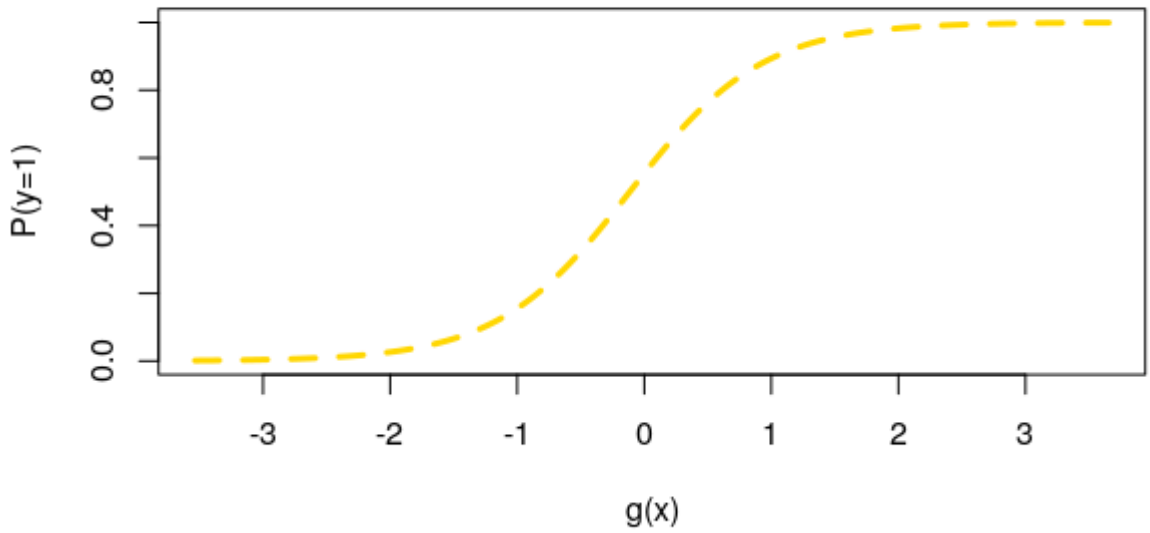

Figura 3.1: *Curva do modelo de regressão logística.*

Uma outra característica importante que faz o modelo logístico ser tão utilizado é a fácil interpretação dos parâmetros do modelo. Um valor positivo para o parâmetro associado a uma determina covariável indica que quanto maior o valor desta, maior é a probabilidade de ocorrência de sucesso na variável resposta. Já um valor negativo indica que essa probabilidade decresce com o aumento da covariável.

De forma mais precisa, os parâmetros do modelo podem ser interpretados em função da razão de chances (Agresti, 2003). No modelo logístico a razão das chances é obtida através da exponencial do coeficiente de uma covariável e as interpretações variam de acordo com a classificação da covariável. Suponha, por exemplo, uma covariável discreta binária, 0 e 1, cujo exponencial do coeficiente associado a essa covariável foi de 1,23. Nesse caso, estima-se que a chance de se obter o evento de sucesso na variável resposta ($P((Y|x_i^\top) = 1)$) é 23% maior para o grupo 1 do que para o grupo 0, mantidas as demais covariáveis constantes. Se a covariável é contínua e o exponencial de seu coeficiente associado for maior que 1, por exemplo, 1,56, estima-se que o aumento de uma unidade

daquela covariável resulta em um aumento de 56% das chances de sucesso, mantidas as demais covariáveis constantes. Quando o exponencial do coeficiente for menor que 1 isto indica decréscimo nas chances de sucesso da variável resposta à medida que aumenta o valor da covariável em estudo e mantém-se as demais constantes.

3.3 Estimação e seleção de variáveis na regressão logística

Como caso particular dos modelos lineares generalizados, os parâmetros do modelo de regressão logística são geralmente estimados por máxima verossimilhança. Para estimar os parâmetros do modelo por máxima verossimilhança, precisamos definir a função de verossimilhança do modelo de regressão logística. Por definição sabemos que a função de verossimilhança de uma variável aleatória discreta, independente e identicamente distribuída, é dada pelo produtório de sua função de probabilidade, isto é

$$L(\beta_0, \beta) = \prod_{i=1}^{n} f_Y(y_i|\mu_i) = \prod_{i=1}^{n} \mu_i^{y_i}(1-\mu_i)^{1-y_i} = \prod_{i=1}^{n} \frac{exp\left(y_i(\beta_0 + x_i^{\top t}\beta)\right)}{1 + exp\left(\beta_0 + x_i^{\top t}\beta\right)}. \tag{3.9}$$

Aplicando-se o logaritmo na função de verossimilhança, obtemos

$$l(\beta_0, \beta) = log(L(\beta_0, \beta)) = \sum_{i=1}^{n} \left[y_i(\beta_0 + x_i^{\top}\beta) - log\left(1 + exp\left(\beta_0 + x_i^{\top}\beta\right)\right)\right]. \tag{3.10}$$

3.3.1 Lasso na regressão logística

Como já discutido anteriormente, quando aplicamos o lasso em um modelo de regressão, adicionamos a penalização ao termo negativo do logaritmo da função de verossimilhança. Assim a função objetivo para a estimação dos parâmetros pelo método lasso é dada por

$$\underset{\beta_0,\beta}{\text{minimize}} \frac{-1}{n} \sum_{i=1}^{n} \left[y_i(\beta_0 + x_i^{\top}\beta) - log\left(1 + exp\left(\beta_0 + x_i^{\top}\beta\right)\right)\right] + \lambda||\beta||_1 \tag{3.11}$$

No Capítulo 2 foi discutido o algoritmo de coordenada descendente para estimação dos parâmetros do modelo de regressão linear geral via lasso. O mesmo algoritmo pode ser aplicado para estimar os parâmetros de outros MLG como a regressão logística pelo método lasso através da Equação (3.11). A desvantagem de usar este método na prática é que ao aplicá-lo usando a função objetivo definida na Equação (3.11), as coordenadas não são tão explícitas e de fácil análise como era no modelo linear geral. Segundo Hastie et al. (2015), em sua experiência na prática, a melhor maneira de estimar os parâmetros

dos demais MLG é aplicando a coordenada descendente na aproximação quadrática do logaritmo da função de máxima verossimilhança dada pela expansão da série de Taylor. Logo se fizermos a aproximação quadrática do logaritmo da função de máxima verossimilhança da regressão logística, por expansão da série de Taylor temos

$$l_Q(\beta_0, \beta) = \frac{1}{2n} \sum_{i=1}^{n} w_i \left(z_i - \beta_0 - x_i^\top \beta \right)^2 + C(\tilde{\beta_0}, \tilde{\beta}), \tag{3.12}$$

onde $w_i = \tilde{p}(x_i^\top)\left(1 - \tilde{p}(x_i^\top)\right)$, $z_i = \tilde{\beta_0} + x_i^\top \tilde{\beta} + \frac{y - \tilde{p}(x_i^\top)}{w_i}$, $\tilde{\beta_0}$ e $\tilde{\beta}$ são os valores atuais estimados dos parâmetros, $\tilde{p}(x_i^\top)$ é o atual valor estimado da probabilidade de ocorrência de sucesso na variável resposta e C denota uma função que não depende dos parâmetros (β_0, β). Com isso minimizar a Equação (3.12) se torna um simples problema de mínimos quadrados ponderados.

Após definir a aproximação quadrática do logaritmo da função de máxima verossimilhança da regressão logística, podemos então definir a função objetivo para estimação via lasso, que é dada pelo negativo da aproximação quadrática adicionado ao termo de penalização, isto é

$$\underset{\beta_0, \beta}{\text{minimize}} \, -l_Q(\beta_0, \beta) + \lambda \|\beta\|_1. \tag{3.13}$$

A solução para a Equação (3.13) pode ser encontrada a partir de um algoritmo conhecido como mapa aproximado de Newton. Maiores detalhes dessa técnica podem ser encontradas em Lee et al. (2014). De maneira geral a técnica consiste em fazer a estimação em 3 loops. No primeiro loop o valor de λ é diminuído. No segundo a aproximação da função quadrática é atualizada pelos valores correntes dos parâmetros. No terceiro loop o algoritmo de coordenada descendente é realizado para encontrar a solução da Equação (3.13). O pacote do R GLMNET que nós utilizamos para fazer a estimação do modelo de regressão logística faz o uso desse algoritmo de Newton.

Conforme discutido na Seção 2.3, ao fazer validação cruzada para encontrar o termo de penalização, devemos sempre definir um critério para avaliar a performance do modelo. Especificamente no modelo logístico, para realização da validação cruzada para estimar o parâmetro λ usando o pacote GLMNET, podemos escolher entre 4 diferentes critérios: a deviance (Paula, 2013), a área sob a curva roc, o erro médio absoluto e também a taxa de erro de classificação. Para o desenvolvimento deste trabalho foi adotado a área sob a curva roc.

Uma característica importante da regressão logística no GLMNET é que, para fazer futuras previsões, pode ser definido pelo usuário o ponto de classificação da variável resposta, isto é, para qual valor de probabilidade devemos classificar a variável resposta

como sucesso ou fracasso. O padrão do pacote é classificar como sucesso se a probabilidade estimada for maior que 0,5.

3.3.2 Stepwise

Stepwise é o método mais popular quando o assunto é seleção de covariáveis em regressão. O stepwise é composto pela junção dos métodos forward e backward (Neter et al. , 2008). O método *forward* consiste em, partindo de um modelo nulo, adicionar covariáveis ao modelo de regressão caso, adicionando essa covariável, o desempenho do modelo melhore. Por exemplo, suponhamos que o desempenho do modelo seja analisado pelo critério BIC (Akaike, 1978; Schwarz, 1978). Se a adição de uma determinada covariável melhora consideravelmente o desempenho do modelo (reduz o BIC), então ela é adicionada. A adição de covariáveis é feita até que o acréscimo de uma nova covariável não traga melhora ao desempenho do modelo (não reduza o BIC). O *backward* faz a seleção em sentido contrário. Ele parte de um modelo com todas as covariáveis disponíveis, e as vai removendo se a sua ausência no modelo não gerar perda de desempenho do mesmo (não gerar aumento do BIC). Enquanto um só remove e outro só acrescenta covariáveis, o stepwise atua nos dois sentidos. Após cada adição de covariável, ele verifica se uma das covariáveis já presentes pode ser excluída. A seleção de variáveis é finalizada quando nem a adição e nem a exclusão de uma covariável melhora o desempenho do modelo.

Ao utilizar o stepwise devemos definir qual medida será utilizada para avaliar o desempenho do modelo, isto é, qual métrica utilizar para adicionar ou remover covariáveis. As métricas mais frequentes a serem utilizadas são: AIC (Akaike, 1974), BIC, entre outras. Neste trabalho foi utilizado a técnica de stepwise presente no pacote "MASS" (Venables e Ripley, 2002) do R. As configurações padrões foram mantidas, no qual o critério de seleção é o AIC, critério de informação de Akaike em tradução livre.

3.3.3 Lasso e máxima verossimilhança

O método de máxima verossimilhança é um dos métodos mais utilizados para estimação de parâmetros em modelos de regressão. Além de ser um método bastante intuitivo, ele apresenta interessantes propriedades teóricas. Os estimadores obtidos por esse método são consistentes. Os estimadores de máxima verossimilhança também carregam a propriedade de serem invariantes sob transformações monotônicas (Casella e Berger, 2002). Os estimadores obtidos por esse método são assintoticamente não viesados e apresentam mínima variância assintótica entre os estimadores assintoticamente não viesados. Por fim, os estimadores de máxima verossimilhança possuem distribuição assintoticamente normal, o que é bastante conveniente para a construção de intervalos de confiança e a realização de testes de hipóteses.

Como a estimação por máxima verossimilhança é muito utilizada, por suas diversas

vantagens e propriedades, é natural a ideia de uni-la a um método de seleção de covariáveis para assim poder analisar o seu poder preditivo. Entre os muitos métodos de seleção na literatura, neste trabalho, ele foi combinado com o método de stepwise.

Diversos trabalhos já constataram o bom desempenho do lasso para predição, e por ele ser um método também capaz de selecionar covariáveis, também é interessante combinar este método com a estimação por máxima verossimilhança e analisar o desempenho dessa combinação.

Na combinação de lasso e máxima verossimilhança o ajuste do modelo é feito em dois passos. No primeiro passo temos a seleção de covariáveis, onde o lasso é aplicado no modelo. Em seguida o modelo, com as covariáveis escolhidas via lasso, tem seus parâmetros estimados por máxima verossimilhança.

Um método proposto na literatura e que possui uma certa similaridade com a combinação de lasso com máxima verossimilhança é o relaxed lasso (Meinshausen, 2007). O relaxed lasso é uma variação do lasso que estamos apresentando nesse trabalho. Ele consiste em "aplicar" o lasso duas vezes em seu modelo. Primeiro apenas para selecionar as melhores covariáveis e uma segunda vez para estimar melhor os parâmetros do que o lasso em sua versão original costuma fazer. Um fato muito interessante acontece quando um dos parâmetros do relaxed lasso, ϕ, assume o valor 0. Nesse caso, ele se torna uma combinação do lasso para seleção de covariáveis e da estimação via mínimos quadrados. Isso é semelhante a um do métodos considerados neste trabalho. A diferença em relação a esse caso é que no nosso trabalho a estimação é feita por máxima verossimilhança e não por mínimos quadrados.

3.3.4 Stepwise e lasso

No método discutido na Seção 3.3.3 usamos o lasso como método de seleção de covariáveis. Entretanto, sabemos que o lasso também é um método de estimação. Logo, podemos pensar em avaliar o desempenho do poder preditivo do lasso para estimação combinado a um método de seleção como o stepwise.

A combinação dos métodos stepwise e lasso tem uma particularidade interessante. No primeiro passo o stepwise vai ser responsável por selecionar as covariáveis que são consideradas importantes e devem ser mantidas no modelo. No segundo passo o lasso vai ser responsável por estimar os coeficientes das covariáveis escolhidas pelo stepwise. Entretanto o lasso é um método que ao mesmo tempo estima e seleciona covariáveis. Portanto, nessa combinação fazemos duas vezes a seleção de covariáveis, sendo esta a particularidade mencionada. Por combinarmos dois métodos que fazem seleção de variáveis o esperado é que eles criem modelos, em média, com menos covariáveis do que os demais métodos em estudo.

Capítulo 4

Estudos de simulação

Neste capítulo são apresentados os estudos de simulação de Monte Carlo desenvolvidos neste trabalho. Toda a parte de programação foi desenvolvida no software R, utilizando o pacote "GLMNET" para todo o procedimento envolvendo o método do lasso.

O objetivo desses estudos de simulação é avaliar o desempenho dos 4 métodos (Lasso, Lasso+MV, Stepwise+Lasso e Stepwise+MV) na regressão logística e sob diversos cenários, isto é, com diferentes tamanhos amostrais, correlação entre as covariáveis, número de covariáveis e percentual de covariáveis associados com a resposta. Nosso objetivo é analisar qual método se sobressai, em quais métricas e em quais cenários específicos. Conforme discutido no Capítulo 1, Hastie et al. (2020) em seu estudo de simulação comprovou que o lasso, na regressão linear, pode se sobressair em cenários específicos, como também pode não se destacar em outros. Portanto, é fundamental buscar criar cenários diversificados.

Os métodos foram avaliados neste trabalho em dois aspectos: quanto ao poder preditivo e em relação à proximidade entre o modelo ajustado e o verdadeiro modelo. Embora nosso principal interesse seja no estudo do poder preditivo do modelo, na prática algumas vezes é de interesse tanto a avaliação do poder preditivo quanto o ajuste de um modelo que consiga explicar bem a relação entre a variável resposta e as covariáveis. Dessa forma, é interessante também a avaliação dos métodos considerados neste trabalho em relação ao segundo aspecto mencionado.

4.1 Avaliação do poder preditivo

Para o primeiro grupo de estudos de simulação de Monte Carlo foi utilizada a técnica *datasplitting* que consiste em dividir a base de dados aleatoriamente em 2 partes, uma para treinar o modelo e outra para testá-lo (Butcher e Smith, 2020). Essa divisão é importante para evitar problemas como super ou sobre ajuste do modelo. Foram consideradas neste trabalho 500 réplicas de Monte Carlo e para cada uma delas utilizou-se *data splitting*.

No primeiro grupo foi avaliado o desempenho de predição e o número médio de

covariáveis selecionadas por cada método. No total foram avaliados 144 cenários de simulação, no qual variamos os tamanhos amostrais ($n = 200, 500, 1000$ e 5000), a correlação entre as covariáveis (ρ = 0, 0,2, 0,5 e 0,8), o número de covariáveis presentes (p = 10, 30 e 50) e também a porcentagem de covariáveis com parâmetro associado não nulo (20%, 40% e 60%). A divisão em base de treinamento e teste foi feita considerando-se 70% para treinamento e 30% para teste de forma aleatória. As covariáveis do modelo foram simuladas de uma distribuição normal multivariada com vetor de médias igual a 0 e matriz de covariâncias de tal forma que a variância de todas as covariáveis fosse igual a 1 e a correlação entre qualquer par de variáveis fosse igual a ρ. Elas foram mantidas fixas nas 500 réplicas de Monte Carlo. Os valores dos coeficientes não nulos foram fixados em 50% iguais a -3 e 50% iguais a 2, da mesma forma que foi considerado em Ijaz et al. (2019), sendo que os coeficientes nulos ocupavam as primeiras posições do vetor.

Quanto ao balanceamento de classes, isto é, a proporção de zeros e uns presentes em cada réplica de Monte Carlo, a maneira que as amostras foram construídas (valores do parâmetros e a distribuição utilizada para gerar os dados) garantem uma proporção de zeros e uns, em média, em 50% para todos os cenários.

A métrica utilizada para medir o poder preditivo dos 4 métodos foi o coeficiente de Gini (Thomas et al., 2002). O coeficiente de Gini (CG) é um medida que varia entre 0 e 1, em que 1 representa perfeita predição do modelo e 0 representa predição equivalente a uma escolha aleatória do valor de uma variável resposta binária. O coeficiente de Gini é uma função da área sob a curva roc (auc), isto é,

$$CG = 2 \times (auc - 0,5) \tag{4.1}$$

Embora a área sob a curva roc seja mais conhecida, o coeficiente de Gini é uma melhor forma de observar o poder preditivo do modelo, pois ele transforma o valor da área sob a curva roc em um coeficiente no intervalo $[0, 1]$.

Outra métrica utilizada foi o número médio de covariáveis selecionadas. O objetivo nesse caso é avaliar qual método ajusta modelos mais esparsos, isto é, modelos com menos covariáveis e portanto mais fáceis de serem interpretados. O esperado é que a combinação lasso com stepwise seja responsável pelas menores médias, por serem dois métodos combinados que selecionam covariáveis.

Quanto às escolhas do pacote GLMNET dentro do R, para estimar o parâmetro λ foram utilizadas as configurações padrões de validação cruzada, exceto pela métrica utilizada para a escolha de λ. Nesse trabalho, a métrica utilizada para a escolha de λ foi a área sob a curva roc, e assim escolhemos λ que maximiza essa medida. Um dos programas utilizados para a realização deste estudo de simulação pode ser encontrado no repositório deste autor no Github (Alcântara Junior, 2020).

4.1.1 Resultados da avaliação do poder preditivo

Nessa seção são apresentados os resultados do estudo de simulação relacionados com a avaliação de performance dos métodos. Para melhor organização do texto, apresentamos nesta seção apenas as tabelas que trazem os resultados para os cenários com correlação zero entre as covariáveis e que apresentam 20% de covariáveis importantes. As demais tabelas se encontram nos Apêndices A e B.

Na Tabela 4.1 temos os resultados da média do coeficiente de Gini para os cenários mencionados no parágrafo anterior. Podemos observar que conforme a dimensionalidade do modelo aumenta, ou seja, o número de covariáveis aumenta em relação ao tamanho amostral, a performance de todos os 4 métodos sofre um perda. O lasso é o método com menor perda de perfomance e o stepwise combinado com a máxima verossimilhança apresenta a maior. Observe que quando p=10 e n=200 a média estimada do coeficiente de Gini do stepwise com máxima verossimilhança era de 0,86918. Mantido fixo o tamanho amostral e aumentando-se o número de covariáveis para p=50, a média do coeficiente de Gini decaiu para 0,48779, uma perda acima de 50% no poder preditivo. Isto reflete o fato, já mencionado, de como a presença de alta dimensionalidade na base de dados pode afetar a estimação via máxima verossimilhança e a seleção via stepwise. Se o tamanho amostral é suficientemente grande para o número de covariáveis, isto é, para n=5000 a performance dos 4 métodos é semelhante na média estimada do coeficiente de Gini, independente do número de covariáveis.

Quanto ao poder preditivo dos 4 métodos pela média estimada do coeficiente de Gini, o lasso é indiscutivelmente o melhor método em relação aos cenários considerados na Tabela 4.1. Ele se sobressai em relação aos demais métodos quando a relação entre o número de covariáveis e o tamanho amostral é relativamente grande. Quando o tamanho amostral é muito grande ele se iguala aos demais métodos sem perda de performance para nenhum outro. Além de maior poder preditivo médio, apresenta também um menor desvio padrão do coeficiente de Gini. Isso é interessante porque indica que a performance do lasso varia menos quando mudamos de uma amostra para outra em uma mesma população.

Observando as demais tabelas do Apêndice A, é possível constatar que as mudanças de correlação e porcentagem de covariáveis com parâmetros associados não nulos não implicam em perda ou ganho considerável de desempenho preditivo estimado dos métodos. Aumentando o tamanho amostral e fixado o número de covariáveis, o desempenho de todos os 4 métodos aumentam, até que, para um tamanho amostral suficientemente grande, o poder preditivo dos 4 métodos se tornam semelhantes. O aumento de covariáveis nos demais cenários, fixado o tamanho amostral, também foi um fator que implicou em mudanças no desempenho preditivo estimado dos métodos estudados. Conforme o número de covariáveis aumenta, para um tamanho amostral pequeno, a média do coeficiente de Gini estimada de todos os métodos cai, sendo a combinação stepwise com

máxima verossimilhança o método com maior perda e o lasso com a menor. É possível também observarmos que parece existir uma relação entre o número de covariáveis e o tamanho amostral que garantem um desempenho superior ao lasso. Para aqueles cenários onde a relação p/n é maior ou igual que 0,1, o lasso obeteve um desempenho de predição superior a todos os outros métodos em todos os cenários, e quanto maior o valor dessa relação maior é a superioridade do lasso.

Tabela 4.1: *Resultados da média do coeficiente de Gini e seu desvio padrão para os cenários com correlação nula e 20% de covariáveis importantes.*

$\rho = 0$ 20% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
n=200	Lasso	0,87505 (0,05530)	0,92269 (0,04656)	0,89306 (0,05558)
	Lasso+MV	0,87143 (0,05872)	0,86944 (0,10529)	0,78978 (0,10309)
	Lasso+Step	0,87422 (0,05617)	0,89124 (0,07603)	0,68262 (0,13659)
	Step+MV	0,86918 (0,05899)	0,69919 (0,12172)	0,48779 (0,13701)
n=500	Lasso	0,86275 (0,03815)	0,94781 (0,02165)	0,95498 (0,01844)
	Lasso+MV	0,86123 (0,03848)	0,94321 (0,02444)	0,93351 (0,05651)
	Lasso+Step	0,86152 (0,03830)	0,94164 (0,02423)	0,94116 (0,03094)
	Step+MV	0,85943 (0,03898)	0,93617 (0,03206)	0,78171 (0,07950)
n=1000	Lasso	0,88057 (0,02552)	0,94695 (0,01415)	0,96634 (0,00980)
	Lasso+MV	0,87947 (0,02607)	0,94572 (0,01447)	0,96373 (0,01117)
	Lasso+Step	0,87987 (0,02583)	0,94445 (0,01456)	0,96088 (0,01156)
	Step+MV	0,87896 (0,02615)	0,94349 (0,01474)	0,95902 (0,01231)
n=5000	Lasso	0,87406 (0,01101)	0,94969 (0,00598)	0,970127 (0,00424)
	Lasso+MV	0,87381 (0,01100)	0,94930 (0,00603)	0,96951 (0,00440)
	Lasso+Step	0,87385 (0,01098)	0,94913 (0,00604)	0,96929 (0,00438)
	Step+MV	0,87367 (0,01099)	0,94894 (0,00606)	0,96908 (0,00440)

No início deste capítulo foi mencionado que o esperado é que a combinação entre lasso e stepwise seja responsável pelos modelos com maior esparsidade, isto é, modelos com menores média de covariáveis selecionadas. De fato isto ocorreu para os cenários com 10 covariáveis considerados na Tabela 4.2. Nos modelos com 30 covariáveis, sem correlação e apenas 20% de covariáveis significativas, o lasso sozinho foi capaz de construir modelos, em média, com menos covariáveis independente do tamanho amostral. Isto também ocorreu para $p = 50$ e $n \geq 500$.

Era esperado também que o lasso criasse modelos mais esparsos em relação ao stepwise combinado com a máxima verossimilhança. Nos cenários considerados na Tabela 4.2, isso ocorreu para aqueles em que $p \geq 30$, mas não para os que apresentam $p = 10$. Caso em determinado problema em estudo haja interesse em obter modelos esparsos, pode-se considerar outras formas de escolher o parâmetro λ no lasso (Hastie e Qian, 2014).

Nos demais cenários, que podem ser encontrados no Apêndice B, podemos observar que o aumento do tamanho amostral, fixado o número de covariáveis, também implica em um aumento da esparsidade do modelo estimado pelos 4 métodos. Se o tamanho amostral é muito grande, a média de covariáveis selecionadas pela combinação lasso com stepwise e stepwise com máxima verossimilhança fica bastante similar entre si. Conforme a porcentagem de covariáveis importantes aumenta, também aumenta o número médio de covariáveis selecionadas. Porém isso já era esperado, pois são mais covariáveis com coeficientes associados não nulos para selecionar. A correlação não implicou em mudanças consideráveis no número médio de covariáveis selecionadas por nenhum método. Dentre os 4 métodos, os modelos selecionados pelo stepwise combinado com o lasso são na maioria dos cenários, em média, mais esparsos do que os demais métodos. Comparando especificamente lasso e stepwise com máxima verossimilhança, o lasso cria modelos, na maior parte dos cenários, em média, com mais covariáveis do que o stepwise com máxima verossimilhança.

Tabela 4.2: *Resultados da média do número covariáveis selecionados para os cenários com correlação 0 e 20% de covariáveis importantes.*

$\rho = 0$ 20% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
n=200	Lasso	3,952	11,816	23,562
	Lasso+MV	3,952	11,816	23,562
	Lasso+Step	3,268	12,036	22,826
	Step+MV	3,536	15,252	24,472
n=500	Lasso	3,514	11,012	19,092
	Lasso+MV	3,514	11,012	19,092
	Lasso+Step	3,124	11,546	22,160
	Step+MV	3,378	11,794	28,966
n=1000	Lasso	3,474	9,470	16,724
	Lasso+MV	3,474	9,470	16,724
	Lasso+Step	3,146	10,230	19,338
	Step+MV	3,356	10,366	19,528
n=5000	Lasso	3,258	8,856	14,924
	Lasso+MV	3,258	8,856	14,924
	Lasso+Step	3,138	9,918	16,570
	Step+MV	3,270	9,962	16,596

4.2 Avaliação da proximidade entre o modelo ajustado e o verdadeiro

No segundo grupo, não houve *data splitting*, pois tratava-se de avaliar o viés e a raiz do erro quadrático médio (EQM) dos estimadores dos parâmetros do modelo, sendo desnecessária portanto a aplicação desse procedimento. Buscando tornar os resultados mais equiparáveis e próximos do primeiro grupo, houve mudanças nos tamanhos amostrais para corresponder ao que seria o tamanho do conjunto treinamento do primeiro grupo, isto é, foram utilizados n=140, 350, 700 e 3500. Quanto a correlação, número de covariáveis presentes, porcentagem de covariáveis não nulas, o número de repetições e a forma como as covariáveis foram simuladas seguiram o mesmo padrão do primeiro grupo.

O viés e a raiz do EQM são muito utilizados para a avaliação de estimadores. Entretanto, como o número de covariáveis é muito grande, analisar cada estimador separadamente seria inviável. Portanto, trabalhamos com estimativas para a média do módulo do viés e para a média da raiz do EQM em dois grupos distintos. O primeiro contém os estimadores de parâmetros não nulos e o segundo contém os demais.

Também obtivemos estimativas para o viés e raiz do EQM do modelo conforme definido na Seção 2.2. Novamente trabalhamos com a média do viés e raiz do EQM, já que neste caso temos uma estimativa dessas medidas para cada uma das n observações. As últimas métricas consideradas foram as estimativas de proporções de vezes que os métodos selecionavam o número correto de covariáveis, isto é, a proporção de vezes que o método selecionou a mesma quantidade de covariáveis que o modelo verdadeiro, mas não necessariamente as covariáveis certas. Também foi obtida estimativa para a proporção de vezes que o modelo selecionou exatamente as mesmas covariáveis associadas a parâmetros não nulos do modelo original, isto é, selecionou o verdadeiro modelo.

Dentro do software R, as configurações dos pacotes utilizados foram definidas da mesma forma que no primeiro grupo.

4.2.1 Resultados da avaliação da proximidade entre o modelo ajustado e o verdadeiro

Nessa seção são apresentados os resultados do estudo de simulação relacionados com a avaliação de estimação dos modelos ajustados pelos 4 métodos estudados. Novamente para melhor organização do texto, apresentamos nesta seção apenas as tabelas que trazem os resultados para os cenários com correlação zero entre as covariáveis e que apresentam 20% de covariáveis importantes. As demais tabelas se encontram nos Apêndices C, D, E e F.

Como discutido no Capítulo 2, o lasso tende, na maioria das vezes, a construir modelos com maior poder preditivo, o que de fato aconteceu nas simulações. Entretanto para aumentar esse poder preditivo, ele também, na maioria das vezes, tende a aumentar o viés do modelo ajustado. Isto pode ser visto em nossos estudos de simulação.

Na Tabela 4.3 temos os resultados da média do módulo do viés e a média da raiz do EQM dos modelos estimados para os cenários definidos no ínicio da seção. Como podemos ver, o lasso apresentou um viés médio estimado maior do que os outros métodos combinados, o que já era esperado. O aumento do tamanho amostral, fixado o número de covariáveis, diminui as médias do viés e raiz do EQM estimados para todos os métodos. Um fato que chama atenção é que os métodos onde o lasso é responsável pela estimação, são aqueles que apresentaram maiores médias de viés estimado e de raiz do EQM. Os métodos onde a estimação é feita por máxima verossimilhança apresentaram os menores valores e também, em geral, apresentaram resultados próximos. É interessante ressaltar ainda que na maioria dos cenários apresentados na Tabela 4.3, o stepwise combinado

Tabela 4.3: *Resultados da média do módulo do viés e a média da raiz quadrada do EQM do modelo para os cenários com correlação 0 e 20% de covariáveis importantes.*

$\rho = 0$ 20% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
n=140	Lasso	Viés	0,09721	0,10240	0,07667
		$\sqrt{EQM}$	0,13224	0,14815	0,16291
	Lasso+MV	Viés	0,00419	0,00910	0,00637
		$\sqrt{EQM}$	0,06878	0,13052	0,17921
	Step+Lasso	Viés	0,06036	0,03351	0,05552
		$\sqrt{EQM}$	0,11270	0,14643	0,20407
	Step+MV	Viés	0,00425	0,00589	0,00581
		$\sqrt{EQM}$	0,07518	0,18187	0,18226
n=350	Lasso	Viés	0,07625	0,09515	0,09207
		$\sqrt{EQM}$	0,09848	0,11412	0,11541
	Lasso+MV	Viés	0,00310	0,00794	0,00805
		$\sqrt{EQM}$	0,04140	0,06506	0,08565
	Step+Lasso	Viés	0,04502	0,01908	0,03132
		$\sqrt{EQM}$	0,08184	0,07385	0,10928
	Step+MV	Viés	0,00239	0,00397	0,00480
		$\sqrt{EQM}$	0,04508	0,07222	0,13242
n=700	Lasso	Viés	0,04566	0,08246	0,08832
		$\sqrt{EQM}$	0,05682	0,09407	0,09870
	Lasso+MV	Viés	0,00124	0,00322	0,00601
		$\sqrt{EQM}$	0,02760	0,03992	0,04794
	Step+Lasso	Viés	0,02248	0,01526	0,01092
		$\sqrt{EQM}$	0,04360	0,05212	0,05687
	Step+MV	Viés	0,00104	0,00244	0,00309
		$\sqrt{EQM}$	0,02988	0,04596	0,05758
n=3500	Lasso	Viés	0,02286	0,04546	0,05616
		$\sqrt{EQM}$	0,02787	0,05005	0,06017
	Lasso+MV	Viés	0,000595	0,00087	0,00161
		$\sqrt{EQM}$	0,01248	0,01637	0,01923
	Step+Lasso	Viés	0,00964	0,00819	0,00665
		$\sqrt{EQM}$	0,01954	0,02416	0,02358
	Step+MV	Viés	0,00054	0,00075	0,00091
		$\sqrt{EQM}$	0,01359	0,01924	0,02232

com máxima verossimilhança apresentou menor média do viés e o lasso combinado com máxima verossimilhança apresentou menor média da raiz do EQM.

Nos demais cenários (Apêndice C) é possível ver que o aumento ou decréscimo da correlação presente nas covariáveis não afetam o viés médio estimado ou a média estimada da raiz do EQM. O lasso foi o único método em que o viés médio estimado diminui se a porcentagem de covariáveis importantes aumentam. E novamente em todos os cenários o aumento de cováriaveis, fixado o tamanho amostral, implica em um aumento das médias estimadas do viés e da raiz do EQM. O aumento do tamanho amostral, fixado o número de covariáveis, implica em decréscimo das médias estimadas do viés e da raiz do EQM. A estimação por máxima verossimilhança associada ao método de seleção via stepwise apresentou as menores médias estimadas para o viés em quase todos os cenários. Para a maioria dos cenários, o lasso combinado com máxima verossimilhança apresentou a segunda menor média do viés. Já em relação à raiz do EQM, a ordenação entre os métodos varia mais entre os cenários. Porém, na maior partes dos cenários, o lasso ou o stepwise combinado com o lasso apresentaram maior média da raiz do EQM.

Nos cenários presentes na Tabela 4.4, onde temos as médias estimadas de viés e raiz de EQM dos estimadores para as covariáveis com parâmetros associados nulos, vemos que o lasso apresentou as menores médias estimadas de maneira geral. O aumento do tamanho

amostral (fixado o número de covariáveis) diminui as médias do viés e da raiz do EQM estimadas. Já o aumento do número de covariáveis (fixado o tamanho amostral) aumenta as médias do viés e da raiz do EQM estimadas. Se o tamanho amostral é relativamente pequeno para o número de covariáveis, as combinações onde o método de máxima verossimilhança é responsável pela estimação tem elevadas médias estimadas tanto para o viés quanto para raiz quadrada do EQM. Isso ocorre porque, para algumas réplicas de Monte Carlo, as estimativas dos parâmetros ficam muito longe dos verdadeiros valores dos parâmetros. Logo, é sugerido nesses cenários não estimar os parâmetros do modelo por máxima verossimilhança se o objetivo for avaliar a relação entre a média da variável resposta e cada uma das covariáveis. Nesses cenários, o lasso é a melhor forma de fazer a estimação do modelo.

Analisando a média do módulo do viés e a média da raiz do EQM dos estimadores para as variáveis com parâmetro associado igual a zero (Apêndice D) conseguimos verificar que mantendo o tamanho amostral fixo e aumentando o número de covariáveis, tanto a média do viés estimado, quanto a média estimada da raiz do EQM dos estimadores aumentam para todos os métodos. Para as combinações em que o método de máxima verossimilhança é responsável pela estimação, os valores estimados de viés e raiz de EQM apresentam valores muito altos onde a razão entre o número de covariáveis e tamanho amostral é relativamente grande. O aumento de porcentagem de covariáveis importantes no modelo, em alguns cenários, causa um grande aumento nas médias estimadas de viés e raiz do EQM nos métodos associados à estimação por máxima verossimilhança. Já o aumento da correlação amostral não indicou mudanças consideráveis nas médias estimadas. O aumento do tamanho amostral, fixado o número de covariáveis, diminui as médias estimadas tanto de viés, quanto raiz do EQM. Para todos os cenários considerados, o lasso é o método com menor viés e menor raiz do EQM para a estimação das covariáveis com coeficientes associados nulos.

Na Tabela 4.5, temos os resultados das covariáveis cujo coeficiente associado é não nulo para o cenário mencionado no começo desta seção. Analisando os resultados podemos ver que, assim como no caso das covariáveis com parâmetros associados nulos, quando o método de estimação é o de máxima verossimilhança, tanto o viés quanto a raiz do EQM aumentam se o número de covariáveis aumenta e o tamanho amostral é fixado. Diferente do caso anterior, o lasso não apresentou as menores médias de viés e raiz do EQM médio estimadas em geral. Pelo contrário, apresentou as maiores médias estimadas quando o tamanho amostral é relativamente grande para o número de covariáveis. Quando isso acontece a combinação lasso com máxima verossimilhança se sobressai. Entretanto se o tamanho amostral é pequeno para o número de covariáveis, como por exemplo, $n = 140$ e $p = 50$ então, o lasso tem as menores médias estimadas de viés e raiz do EQM.

Nos outros cenários (Apêndice E) podemos notar que se aumentarmos o tamanho amostral, como esperado, a média estimada do viés e da raiz do EQM dos estimadores decrescem para todos os métodos. Mudanças na correlação não afetam consideravelmente

Tabela 4.4: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado igual a zero para os cenários com correlação 0 e 20% de covariáveis importantes.*

$\rho = 0$ 20% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
n=140	Lasso	Viés	0,00397	0,01997	0,07997
		$\sqrt{EQM}$	0,13443	0,32224	0,34704
	Lasso+MV	Viés	0,00857	5,12e+11	9,16e+10
		$\sqrt{EQM}$	0,26426	1,41e+13	2,05e+12
	Step+Lasso	Viés	0,00499	0,20931	1,30571
		$\sqrt{EQM}$	0,17393	3,70509	6,55817
	Step+MV	Viés	0,00588	2,53e+11	4988,162
		$\sqrt{EQM}$	0,29192	5,66e+12	105645,7
n=350	Lasso	Viés	0,00397	0,00824	0,01591
		$\sqrt{EQM}$	0,07409	0,07009	0,13825
	Lasso+MV	Viés	0,00821	0,02905	3,20e+11
		$\sqrt{EQM}$	0,13281	0,22959	6,67e+12
	Step+Lasso	Viés	0,00262	0,00856	0,06003
		$\sqrt{EQM}$	0,10607	0,20512	1,50454
	Step+MV	Viés	0,00389	0,01058	1,97e+11
		$\sqrt{EQM}$	0,14854	0,28086	4,40e+12
n=700	Lasso	Viés	0,00161	0,00332	0,00538
		$\sqrt{EQM}$	0,04250	0,03842	0,03122
	Lasso+MV	Viés	0,00384	0,01105	0,02635
		$\sqrt{EQM}$	0,07949	0,11102	0,15787
	Step+Lasso	Viés	0,00222	0,00407	0,00895
		$\sqrt{EQM}$	0,06521	0,11248	0,20451
	Step+MV	Viés	0,00255	0,00560	1,41908
		$\sqrt{EQM}$	0,08913	0,13951	31,76588
n=3500	Lasso	Viés	0,00045	0,00079	0,00132
		$\sqrt{EQM}$	0,02025	0,01424	0,01251
	Lasso+MV	Viés	0,00141	0,00229	0,00531
		$\sqrt{EQM}$	0,03558	0,04041	0,04787
	Step+Lasso	Viés	0,00094	0,00144	0,00197
		$\sqrt{EQM}$	0,03150	0,04432	0,05294
	Step+MV	Viés	0,00103	0,00166	0,00214
		$\sqrt{EQM}$	0,04027	0,05288	0,06134

Tabela 4.5: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado diferente de zero para os cenários com correlação 0 e 20% de covariáveis importantes.*

$\rho = 0$ 20% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
n=140	Lasso	Viés	1,08837	1,17046	0,97271
		$\sqrt{EQM}$	1,30880	1,93547	1,66446
	Lasso+MV	Viés	0,37117	1,04e+13	1,42e+12
		$\sqrt{EQM}$	0,87868	1,09e+14	3,17e+13
	Step+Lasso	Viés	0,63518	10,19228	9,59017
		$\sqrt{EQM}$	1,03655	20,50888	19,03262
	Step+MV	Viés	0,43340	2,56e+12	22474,39
		$\sqrt{EQM}$	0,89335	5,65e+13	386445,3
n=350	Lasso	Viés	0,93665	1,34956	1,42822
		$\sqrt{EQM}$	1,07010	1,41552	1,65008
	Lasso+MV	Viés	0,11869	0,53062	5,70e+12
		$\sqrt{EQM}$	0,35506	0,91448	6,62e+13
	Step+Lasso	Viés	0,52604	0,02105	5,53145
		$\sqrt{EQM}$	0,82613	0,79302	11,09787
	Step+MV	Viés	0,13971	0,78856	2,14e+12
		$\sqrt{EQM}$	0,36642	1,15895	4,78e+13
n=700	Lasso	Viés	0,68554	1,27892	1,47662
		$\sqrt{EQM}$	0,75643	1,32038	1,49334
	Lasso+MV	Viés	0,04410	0,17219	0,42811
		$\sqrt{EQM}$	0,22776	0,37164	0,68242
	Step+Lasso	Viés	0,33417	0,16891	0,24501
		$\sqrt{EQM}$	0,51176	0,50553	1,18075
	Step+MV	Viés	0,05085	0,25597	21,62975
		$\sqrt{EQM}$	0,22879	0,42528	465,0683
n=3500	Lasso	Viés	0,40276	0,91647	1,16528
		$\sqrt{EQM}$	0,44503	0,94194	1,17930
	Lasso+MV	Viés	0,00273	0,03543	0,05518
		$\sqrt{EQM}$	0,09853	0,13042	0,15171
	Step+Lasso	Viés	0,17225	0,17669	0,14692
		$\sqrt{EQM}$	0,26159	0,31440	0,26755
	Step+MV	Viés	0,00301	0,04768	0,08408
		$\sqrt{EQM}$	0,09838	0,13396	0,16570

as médias estimadas. Houve cenários onde o aumento na porcentagem de covariáveis importantes aumentou as médias estimadas para os métodos associados a estimação por máxima verossimilhança. Diferente do caso das covariáveis com parâmetros associados nulos, o lasso neste caso não apresenta a menor das médias estimadas do viés ou da raiz do EQM. Em geral, ele apresenta menor média estimada do viés e da raiz do EQM que os demais métodos apenas quando o tamanho amostral é pequeno em relação ao número de covariáveis. Nos cenários com tamanho amostral relativamente grande para o número de covariáveis, as combinações com a estimação feita por máxima verossimilhança obtiveram as menores médias e bem próximas entre si. Houve cenários onde a máxima verossimilhança com lasso foi melhor e outros onde a máxima verossimilhança com stepwise se destacou.

Na Tabela 4.6 temos os resultados para a proporção de vezes que o método de seleção usado acertou o número de covariáveis do modelo verdadeiro e a proporção de vezes que ele selecionou exatamente as mesmas covariáveis do modelo original, considerando-se o cenário mencionado no início da seção. O objetivo de se avaliar a segunda proporção é verificar com que frequência é selecionado um modelo semelhante ao correto. Podemos observar que quanto maior o tamanho amostral, fixado o número de covariáveis, melhor é o desempenho dos métodos em acertar o número de covariáveis corretas e o modelo

Tabela 4.6: *Resultados da proporção de vezes em que número correto de covariáveis foi identificado e o modelo correto foi selecionado para os cenários com correlação 0 e 20% de covariáveis importantes.*

$\rho = 0$ 20% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
n=140	Lasso	Variáveis corretas	0,506	0,050	0,002
		Modelo correto	0,506	0,050	0
	Lasso+MV	Variáveis corretas	0,506	0,050	0,002
		Modelo correto	0,506	0,050	0
	Step+Lasso	Variáveis corretas	0,340	0,032	0
		Modelo correto	0,340	0,026	0
	Step+MV	Variáveis corretas	0,190	0	0
		Modelo correto	0,190	0	0
n=350	Lasso	Variáveis corretas	0,542	0,142	0,010
		Modelo correto	0,542	0,142	0,010
	Lasso+MV	Variáveis corretas	0,542	0,142	0,010
		Modelo correto	0,542	0,142	0,010
	Step+Lasso	Variáveis corretas	0,376	0,018	0,004
		Modelo correto	0,376	0,018	0,004
	Step+MV	Variáveis corretas	0,234	0,012	0
		Modelo correto	0,234	0,012	0
n=700	Lasso	Variáveis corretas	0,546	0,332	0,074
		Modelo correto	0,546	0,332	0,074
	Lasso+MV	Variáveis corretas	0,546	0,332	0,074
		Modelo correto	0,546	0,332	0,074
	Step+Lasso	Variáveis corretas	0,390	0,026	0
		Modelo correto	0,390	0,026	0
	Step+MV	Variáveis corretas	0,266	0,006	0
		Modelo correto	0,266	0,006	0
n=3500	Lasso	Variáveis corretas	0,578	0,468	0,248
		Modelo correto	0,578	0,468	0,248
	Lasso+MV	Variáveis corretas	0,578	0,468	0,248
		Modelo correto	0,578	0,468	0,248
	Step+Lasso	Variáveis corretas	0,358	0,052	0,002
		Modelo correto	0,358	0,052	0,002
	Step+MV	Variáveis corretas	0,274	0,014	0
		Modelo correto	0,274	0,014	0

correto. Por outro lado se aumentarmos o número de covariáveis, fixando o tamanho amostral, pior é o desempenho. O lasso como método de seleção de covariáveis teve um desempenho melhor, tanto em encontrar o modelo correto, quanto em selecionar a mesma quantidade de covariáveis em todos os casos dos cenários apresentados na Tabela 4.6.

De acordo com as tabelas apresentadas no Apêndice F, conforme a correlação entre as covariáveis aumenta, o desempenho dos 3 métodos associados ao lasso em acertar o número de covariáveis e o modelo correto decai. Já a combinação stepwise com máxima verossimilhança não é afetada consideravelmente. Quanto menor a porcentagem de covariáveis importantes, modelo mais esparso, melhor o desempenho do lasso em encontrar o verdadeiro modelo e pior o desempenho do stepwise. Quanto maior a porcentagem de covariáveis importantes, pior o desempenho do lasso e melhor é o desempenho do stepwise.

De maneira geral para todos os métodos, se o tamanho amostral aumenta (fixado o número de covariáveis) também cresce a proporção de vezes que os métodos selecionam tanto o modelo correto, quanto o número de covariáveis corretas. Se fixarmos o tamanho amostral, conforme aumentamos o número de covariáveis, todos os métodos de seleção decaem em desempenho.

Dois fatos que chamam bastante atenção nos resultados apresentados é a similaridade

Tabela 4.7: *Frequência de seleção das covariáveis por cada método para os cenários com correlação 0 e 20% de covariáveis importantes, n = 200 e p = 10.*

$\rho = 0$ e 20% de covariáveis importantes	Lasso	Step	Step+Lasso
Variável 1	86	84	68
Variável 2	117	117	92
Variável 3	93	104	72
Variável 4	96	101	75
Variável 5	95	90	68
Variável 6	109	105	75
Variável 7	113	98	76
Variável 8	105	94	73
Variável 9	500	500	500
Variável 10	500	500	500

forte entre os valores das duas proporções analisadas em todos os cenários considerados, bem como os baixos valores para as proporções. Na maioria das vezes as proporções são iguais, isto é, os métodos quase nunca identificam modelo incorreto com mesmo número de variáveis do modelo correto. Isso ocorre porque consideramos valores, em módulo, grandes para os parâmetros do modelo. Quando consideramos cenários com valores menores para os parâmetros, como, por exemplo, 0,5, a divergência entre as proporções de número correto de covariáveis e modelo correto se torna mais vísivel. Já o fator das baixas porcentagens está associado ao número grande de covariáveis. Quando consideramos um modelo com apenas 4 covariáveis esses números ficam bem maiores.

Como foi visto na Tabela 4.2, em média os métodos tendem a construir modelos com mais covariáveis que o correto. E na Tabela 4.6 observamos que raramente o modelo correto era de fato selecionado. Logo podemos nos questionar se apesar de não selecionar o modelo correto os métodos ainda selecionavam as covariáveis corretas e outras incorretas (ruídos). Também é interessante avaliar com qual frequência as covariáveis corretas eram selecionadas e como se dava a seleção de covariáveis incorretas para o modelo. Para isso observamos quantas vezes cada covariável entrava no modelo em cada uma das 500 réplicas de Monte Carlo. Na Tabela 4.7 temos os resultados para o cenário com $n = 200$, $p = 10$, correlação 0 e 20% de covariáveis importantes. Como podemos ver, as últimas variáveis (9 e 10), que são as covariáveis importantes, isto é, com coeficientes associados não nulos, foram selecionadas em todas as réplicas. As demais covariáveis foram selecionadas de maneira semelhante e aleatória entre as réplicas em todos os métodos. Com isso, podemos observar que, mesmo quando o modelo correto não é selecionado, todos os métodos conseguem identificar de fato as covariáveis corretas. Entretanto, eles também selecionam para o modelo covariáveis com coeficiente associado nulo.

Capítulo 5

Aplicações

Após os estudos de simulação, nove bases de dados de diferentes áreas e aplicações foram utilizadas para comparar na prática o desempenho dos quatro métodos estudados. Novamente os estudos foram conduzidos no software R com o auxilio do pacote GLMNET. Nas aplicações, a técnica de *data spliting* foi usada em todas as bases de dados, sendo que a divisão foi feita em 70% para treinamento dos modelos e o restante para validação de maneira aleatória. Nas aplicações, as divisões foram refeitas 100 vezes e para avaliar o desempenho de cada método utilizamos novamente o coeficiente de Gini e o número médio de covariáveis selecionadas. Sobre as escolhas para a validação cruzada, stepwise e outras características do pacote GLMNET, o mesmo padrão das simulações foi adotado.

As nove bases de dados foram retiradas dos repositórios do "Kaggle" (Kaggle, 2010), da Universidade da Califórnia (UCI da Universidade da California, 1987) e do pacote "ahaz"(Gorst-Rasmussen e Scheike, 2012). As sete primeiras bases de dados possuem covariáveis quantitativas e as duas últimas são compostas apenas por covariáveis categóricas. Em cada grupo, as aplicações estão ordenadas em relação ao valor de p/n, iniciando-se pela aplicação com menor valor dessa razão. Outras informações sobre cada base de dados são apresentadas na Tabela 5.1.

Como podemos observar, temos duas aplicações em alta dimensionalidade e como já mencionado, enfrentamos alguns problemas quando fazemos a estimação por máxima verossimilhança destes modelos com $p > n$. No software R, quando usamos a função glm com $p > n$, ele considera apenas as primeiras n covariáveis para estimação. Porém, neste trabalho, não enfrentamos este problema porque estamos combinando a estimação por máxima verossimihança com um método de seleção de variáveis. Assim, no nosso caso, quando usamos o lasso, ele já seleciona um número de variáveis menor do que n para que os parâmetros sejam estimados por máxima verossimilhança. E quando usamos o stepwise, a cada etapa do processo de seleção de variáveis estão sendo ajustados modelos em que $p < n$.

A primeira base de dados que trata sobre crédito bancário é a nossa maior base em tamanho amostral e cerca de 22% dos clientes faziam o pagamento classificado como

Tabela 5.1: *Características resumo sobre as nove bases de dados utilizadas.*

Base de dados	n	p	Classes	Referências
Aplicação 1	30000	23	6636 Padrão/ 23364 Não padrão	Yeh e Lien (2009)
Aplicação 2	3656	15	557 Doentes/ 3099 Não doentes	Detrano et al. (1989)
Aplicação 3	392	8	130 Doentes/ 262 Não doentes	Ramana et al. (2011)
Aplicação 4	569	30	212 Malignos / 357 Benignos	Street et al. (1993)
Aplicação 5	351	34	126 Bons / 225 Maus	Sigillito et al. (1989)
Aplicação 6	195	22	147 Doentes / 48 Não doentes	Little et al. (2007)
Aplicação 7	115	550	38 Doentes / 77 Não doentes	Sørlie et al. (2003)
Aplicação 8	123	6	61 Positivos / 62 Negativos	Thrun et al. (1991)
Aplicação 9	70	205	29 Homens / 41 Mulheres	Zarchi et al. (2018)

padrão. A segunda base de dados possui cerca de 15% dos participantes com doença do coração, sendo a menor proporção dentre as bases de dados para uma classe, possuindo também um tamanho amostral relativamente grande para o seu número de covariáveis. A terceira aplicação trata de um estudo sobre doenças hepáticas, sendo a segunda base de dados que possui o menor número de covariáveis, apenas 8, e em suas classes cerca de 33% dos pacientes apresentam alguma doença hepática. A quarta aplicação se aproxima muito de um dos cenários de nossos estudos de simulação ($p = 30$ e $n = 500$), sendo um estudo de câncer de mama. A quinta base de dados, que é sobre radares na ionosfera, possui um tamanho amostral relativamente menor para o número de covariáveis presentes nela, sendo que cerca de 36% são classificados como bons. A base de dados sobre doença de Parkinson, sexta aplicação, também é outra base de dados com tamanho amostral relativamente pequeno para o número de covariáveis, na qual cerca de um quarto dos pacientes não possuem a doença. A sétima aplicação é uma base de dados sobre cancêr, apresentando covariáveis quantitativas e em alta dimensionalidade com 550 covariáveis, sendo a maior de todas em relação a esse aspecto. Já a oitava aplicação trata-se de uma base de dados artificial muito conhecida e utilizada na literatura, conhecida como "Monk's problems", possuindo apenas 6 covaráveis, menor quantidade entre as aplicações, sendo que todas as covariáveis são categóricas. Por fim, na última aplicação sobre crianças com deficiência física e motora, temos uma base de dados em alta dimensionalidade em que suas covariáveis são todas categóricas.

5.1 Resultado das aplicações

Na Tabela 5.2 temos os resultados do coeficiente médio estimado de Gini em cada uma das aplicações, bem como seu desvio padrão estimado e o número médio de covariáveis selecionadas por cada um dos 4 métodos estudados.

Na primeira base de dados, podemos ver que os métodos não obtiveram um bom desempenho e todos apresentaram resultados similares. Quanto ao número médio de covariáveis selecionadas, o stepwise conseguiu obter em média um número menor de

Tabela 5.2: *Performance preditiva e esparsidade dos modelos ajustados.*

Aplicação	Método	Coeficiente de Gini		Média de variáveis
		Média	Desvio Padrão	Selecionadas
1	Lasso	0,44531	0,01231	21,34000
	Lasso+MV	0,44590	0,01222	21,34000
	Step+Lasso	0,44494	0,01229	16,36000
	Step+MV	0,44548	0,01223	16,36000
2	Lasso	0,46197	0,03317	10,38000
	Lasso+MV	0,46094	0,03281	10,38000
	Step+Lasso	0,46014	0,03322	6,98000
	Step+MV	0,45972	0,03315	6,98000
3	Lasso	0,69330	0,06056	6,44000
	Lasso+MV	0,69414	0,06070	6,44000
	Step+Lasso	0,68995	0,06013	4,18000
	Step+MV	0,68964	0,06056	8,17000
4	Lasso	0,98401	0,01347	11,96000
	Lasso+MV	0,96653	0,03393	11,96000
	Step+Lasso	0,97877	0,01492	10,19000
	Step+MV	0,89201	0,03827	13,54000
5	Lasso	0,81297	0,06537	15,84000
	Lasso+MV	0,78427	0,07483	15,84000
	Step+Lasso	0,73896	0,11366	18,47000
	Step+MV	0,70958	0,0948	22,48000
6	Lasso	0,78383	0,07896	8,53000
	Lasso+MV	0,77562	0,08932	8,53000
	Step+Lasso	0,79045	0,09116	7,68000
	Step+MV	0,78135	0,10211	8,17000
7	Lasso	0,44405	0,11958	13,76000
	Lasso+MV	0,33511	0,16420	13,76000
	Step+Lasso	0,29434	0,17340	7,84000
	Step+MV	0,11530	0,14531	22,64000
8	Lasso	0,52804	0,13257	3,27000
	Lasso+MV	0,52995	0,18092	3,27000
	Step+Lasso	0,53243	0,12933	2,66000
	Step+MV	0,52946	0,13187	2,66000
9	Lasso	0,66153	0,18238	14,64000
	Lasso+MV	0,59118	0,19675	14,64000
	Step+Lasso	0,39861	0,31520	6,94000
	Step+MV	0,24290	0,23939	16,19000

covariáveis.

Novamente na segunda base de dados podemos ver que os 4 métodos também não obtiveram um bom desempenho. O coeficiente médio estimado de Gini dos todos os métodos estão bem próximos entre si, sendo que o lasso apresenta um desempenho médio ligeiramente melhor que os demais. Quanto à quantidade de covariáveis selecionadas por cada método, o stepwise foi capaz de criar maior esparsidade, isto é, obteve um número

médio menor de covariáveis selecionadas do que o lasso.

Na terceira aplicação, as estimativas da média do coeficiente de Gini novamente ficaram bem próximas entre os 4 métodos com um desempenho rázoavel. A combinação entre lasso e stepwise selecionou, em média, uma quantidade menor de covariáveis.

Já em nossa quarta base de dados, o lasso apresentou o melhor desempenho preditivo (0,98401) e com menor desvio padrão (0,01347) estimados. A combinação entre stepwise e máxima verossimilhança obteve o pior desempenho (0,89201). Sobre o número médio de covariáveis selecionadas a combinação entre o lasso e o stepwise obteve maior esparsidade (10,19). Essa base de dados é bem próxima de um dos cenários de simulação (p=30 e n=500) e as conclusões são parecidos com as da simulação, já que na maioria desses cenários o lasso é o método de melhor desempenho e a combinação entre stepwise e máxima verossimihança apresenta os piores resultados.

Novamente em nossa quinta aplicação, a estimativa média do desempenho de predição do lasso foi superior aos demais métodos (0,81297), também a combinação do stepwise com a máxima verossimilhança obteve o pior desempenho (0,70958). O lasso foi o método de seleção responsável por criar maior esparsidade. Essa base de dados é uma das bases que possui um alto número de covariáveis em relação ao tamanho amostral entre aquelas com $p < n$, e assim como nos estudos de simulação, se refletiu a superioridade do lasso nessas situações.

Na base de dados de Parkinson, sexta base de dados, o desempenho de todos os métodos ficaram bem próximos, quase empatados, exceto pela leve superioridade do coeficiente de Gini estimado pela combinação entre stepwise e lasso (0,79045). Também houve pouca diferença no número médio de covariáveis selecionadas, sendo que a combinação entre stepwise e lasso também foi capaz de criar um pouco mais de esparsidade.

Na sétima base de dados onde todas as covariáveis eram numéricas e com a presença de alta dimensionalidade, observamos novamente uma superioridade do lasso em relação aos demais métodos. Apesar de seu desempenho não ser satifatório (0,44405), ele ainda é bem superior aos demais métodos.

Na primeira aplicação onde todas as covariáveis eram categóricas (oitava aplicação) e sem presença de alta dimensionalidade, tivemos resultados similares entre todos os métodos quanto ao coeficiente estimado de Gini. Todos os métodos também apresentaram resultados próximos em relação à esparsidade.

Os resultados da última base de dados novamente comprovam a superioridade do lasso em presença de alta dimensionalidade também quando todas as covariáveis são categóricas. Enquanto os métodos combinados ao stepwise tiveram desempenho péssimo, o lasso, sozinho, obteve um desempenho razoável no coeficiente estimado de Gini (0,66153). Novamente a combinação de dois métodos de seleção, stepwise com lasso, criou, em média, modelos mais esparsos.

No Apêndice G podemos observar o número de vezes que cada covariável foi selecionada por cada um dos métodos para as aplicações sem a presença de alta dimensionali-

dade e em todas as 100 divisões feitas nas bases de dados. Podemos observar que apesar das diferenças no número médio de covariáveis selecionadas, assim como observamos nos estudos de simulação, todos os métodos são semelhantes na escolha das covariáveis mais importantes, isto é, as covariáveis que foram escolhidas mais vezes entre as 100 divisões da base entre treinamento e teste.

Nos estudos de simulação vimos que, se o número de covariáveis é relativamente grande para o tamanho amostral, então podemos esperar, em média, um desempenho de predição melhor para o lasso. Isto de fato ocorreu nas aplicações, já que em duas das três bases (aplicações 4 e 5) em que o tamanho amostral não era tão grande em relação o número de covariáveis, o lasso apresentou maior coeficiente médio estimado de Gini. Por outro lado, nesses casos em que o número de covariáveis é relativamente grande para o tamanho amostral, o lasso levou, em geral, a modelos menos esparsos nos estudos de simulação e mais esparsos nas aplicações 4 e 5. Também vimos que para um tamanho amostral relativamente grande, os quatro métodos obtinham resultados similares. Nas aplicações isso também ocorreu já que nas duas bases com maior tamanho amostral em relação ao número de covariáveis (aplicações 1 e 2), a média do coeficiente de Gini ficou próxima entre os 4 métodos. Nas aplicações não foi observado efeito da presença apenas de covariáveis qualitativas no desempenho relativo dos métodos, mas a ocorrência de alta dimensionalidade afeta bastante os resultados. Na base de dados apenas com variáveis categóricas e baixa dimensionalidade, o lasso não se destacou em relação aos demais métodos. Já nas duas aplicações com alta dimensionalidade (com covariáveis quantitativas e apenas com categóricas) o lasso apresentou desempenho bem superior aos demais métodos.

Capítulo 6

Conclusão

Neste trabalho são considerados 4 métodos para a seleção de variáveis e estimação dos parâmetros em modelos de regressão logística. Foram desenvolvidos estudos de simulação para avaliar o desempenho dos 4 métodos estudados. No desempenho dos métodos foi avaliado o poder preditivo de cada um deles para novas observações e também o quão próximo os modelos estimados pelos métodos estavam do verdadeiro modelo. Diferentes métricas foram utilizadas para avaliar esse desempenho: Coeficiente de Gini, numéro médio de covariáveis selecionadas, média do módulo do viés e raiz do EQM dos estimadores dos parâmetros, média do módulo do viés e raiz do EQM dos modelos em si e também as proporções de vezes em que os métodos estudados encontraram o número de variáveis corretas e o modelo correto.

Conforme discutido no Capítulo 4, quanto ao desempenho de predição dos métodos, o lasso é indiscutivelmente o melhor nos cenários considerados. Ele apresenta um coeficiente médio estimado de Gini superior ou similar aos demais métodos com um desvio padrão também menor nos cenários que apresenta maior média. Analisando o número médio de covariáveis com parâmetros associados não nulos, a combinação lasso e stepwise em geral conseguiu modelos mais esparsos e mais próximos do verdadeiro número de covariáveis com parâmetros associados não nulos do modelo original.

O lasso também é o método, nos cenários considerados, com menor viés médio e raiz do EQM para os estimadores associados a parâmetros nulos, principalmente em cenários onde o número de covariáveis era relativamente grande para o tamanho amostral. Nesses cenários ele também apresenta os melhores índices de média do viés e média da raiz do EQM para os estimadores de parâmetros não nulos. Porém quando o tamanho amostral é muito grande o mesmo perde em performance para os demais métodos. Na avaliação dos modelos pela média do viés e média da raiz do EQM, o lasso obteve em geral um desempenho muito inferior aos demais métodos. Principalmente em relação ao viés, isso é coerente com o que é discutido na literatura, já que o lasso introduz mais viés no modelo afim de conseguir melhores predições. Na seleção do número de covariáveis e modelo correto, o lasso se sai melhor quando o número de covariáveis importantes é pequeno,

enquanto o stepwise é melhor quando esse número é grande.

De uma maneira geral, podemos afirmar que o lasso pode se sobressair em quase todas as métricas aqui apresentadas se o cenário a ser estudado possui muitas covariáveis e um tamanho amostral não tão grande, o que reforça a performance superior do lasso em alta dimensionalidade.

Quanto às aplicações estudadas, elas refletiram aquilo que já se esperava a partir dos estudos de simulação. O lasso se sobressaindo ou tendo um desempenho muito parecido na média do coeficiente de Gini estimado em comparação aos demais métodos. E o lasso combinado com o stepwise sendo o método com maior esparsidade nos modelos estimados.

Diante das diversas análises consideradas neste trabalho, se na prática nosso objetivo for construir um modelo com maior poder preditivo, então, o lasso é o melhor método, em relação aos demais aqui apresentados, para a construção desse modelo. Se o objetivo é explicativo e a relação do número de covariáveis com o tamanho amostral (p/n) for grande, o lasso também é o método mais indicado. Entretanto, se o tamanho amostral for suficientemente grande, conforme vimos nos estudos de viés e raiz do EQM, os dois métodos em que a estimação é feita por máxima verossimilhança são a melhor indicação.

6.1 Trabalhos futuros

Os estudos de simulação aqui apresentados foram conduzidos apenas sob a perspectiva de todas as covariáveis provirem da mesma distribuição, a normal padrão. Entretanto sabemos que na prática isto raramente ocorre. É comum as bases ou bancos de dados serem diversificados quanto a distribuição que cada covariável segue. Então um dos pontos a serem trabalhados futuramente é considerar estudos de simulação para outros tipos de covariáveis como, por exemplo, as categóricas. Também pretendemos considerar cenários em que a correlação entre as covariáveis não é constante.

Em nossos estudos de simulação, pela maneira como geramos as covariáveis e definimos os coeficiente dos parâmetros, não há desbalanceamento nas proporções das classes da variável resposta. Futuramente é interessante estudar qual efeito causado por esse desbalancemanto nas proporções.

Recentemente o pacote utilizado para estimar os modelos de regressão no R através do lasso, o GLMNET, foi atualizado com novas funções de ligação para modelos cuja variável resposta é binária, como por exemplo: probito e complemento log-log. Posteriormente, seria interessante ainda estudar o desempenho relativo dos métodos considerados neste trabalho com essas funções de ligação e até mesmo compará-lo com o desempenho do modelo de regressão logística.

Apêndice A

Resultados completos dos estudos de simulação para o coeficiente de Gini

Tabela A.1: *Resultados da média do coeficiente de Gini e seu desvio padrão para os cenários com correlação nula e 40% de covariáveis importantes.*

$\rho = 0$ 40% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	0,90972 (0,04638)	0,92107 (0,04979)	0,87318 (0,07407)
	Lasso+MV	0,90936 (0,04731)	0,81341 (0,09671)	0,78105 (0,10170)
	Lasso+Step	0,91154 (0,04596)	0,86843 (0,08764)	0,71732 (0,12338)
	Step+MV	0,90940 (0,04703)	0,69120 (0,12579)	0,53521 (0,12275)
$n = 500$	Lasso	0,92539 (0,02529)	0,96630 (0,01573)	0,96133 (0,01613)
	Lasso+MV	0,92488 (0,02547)	0,95434 (0,04044)	0,86694 (0,06378)
	Lasso+Step	0,92561 (0,02510)	0,96395 (0,01750)	0,94203 (0,03380)
	Step+MV	0,92497 (0,02533)	0,92924 (0,06608)	0,79467 (0,06465)
$n = 1000$	Lasso	0,91249 (0,01943)	0,96467 (0,01058)	0,97517 (0,00890)
	Lasso+MV	0,91222 (0,01937)	0,96391 (0,01119)	0,96926 (0,02068)
	Lasso+Step	0,91273 (0,01949)	0,96462 (0,01078)	0,97348 (0,00964)
	Step+MV	0,91254 (0,01947)	0,96426 (0,01088)	0,95424 (0,04709)
$n = 5000$	Lasso	0,93444 (0,00737)	0,97466 (0,00365)	0,98327 (0,00287)
	Lasso+MV	0,93433 (0,00735)	0,97441 (0,00367)	0,98296 (0,00291)
	Lasso+Step	0,93439 (0,00737)	0,97453 (0,00365)	0,98305 (0,00286)
	Step+MV	0,93431 (0,00738)	0,97446 (0,00366)	0,98296 (0,00287)

Tabela A.2: *Resultados da média do coeficiente de Gini e seu desvio padrão para os cenários com correlação nula e 60% de covariáveis importantes*

$\rho = 0$ 60% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	0,92567 (0,04307)	0,91944 (0,07490)	0,84118 (0,07590)
	Lasso+MV	0,92817 (0,04350)	0,83531 (0,08376)	0,77886 (0,09602)
	Lasso+Step	0,92849 (0,04116)	0,83269 (0,08732)	0,67810 (0,12038)
	Step+MV	0,92887 (0,04465)	0,65001 (0,11183)	0,51029 (0,12342)
$n = 500$	Lasso	0,95663 (0,01914)	0,96176 (0,01695)	0,95399 (0,04794)
	Lasso+MV	0,95630 (0,01930)	0,93508 (0,05711)	0,86272 (0,05310)
	Lasso+Step	0,95709 (0,01904)	0,96323 (0,01683)	0,93275 (0,03950)
	Step+MV	0,95668 (0,01929)	0,92903 (0,06390)	0,78123 (0,07024)
$n = 1000$	Lasso	0,95202 (0,01360)	0,97600 (0,00850)	0,97821 (0,00812)
	Lasso+MV	0,95196 (0,01361)	0,97558 (0,00881)	0,93691 (0,06073)
	Lasso+Step	0,95211 (0,01360)	0,97679 (0,00837)	0,97860 (0,00800)
	Step+MV	0,95199 (0,01356)	0,97646 (0,00857)	0,92166 (0,06010)
$n = 5000$	Lasso	0,95148 (0,00581)	0,98217 (0,00285)	0,98747 (0,00226)
	Lasso+MV	0,95146 (0,00580)	0,98209 (0,00289)	0,98740 (0,00227)
	Lasso+Step	0,95150 (0,00578)	0,98222 (0,00285)	0,98762 (0,00223)
	Step+MV	0,95147 (0,00579)	0,98218 (0,00286)	0,98760 (0,00224)

Tabela A.3: *Resultados da média do coeficiente de Gini e seu desvio padrão para os cenários com correlação 0,2 e 20% de covariáveis importantes.*

$\rho = 0{,}2$ 20% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	0,80808 (0,07638)	0,92250 (0,04710)	0,91127 (0,04939)
	Lasso+MV	0,80488 (0,076635)	0,85611 (0,10943)	0,80050 (0,09263)
	Lasso+Step	0,80782 (0,07582)	0,88323 (0,08048)	0,72359 (0,12917)
	Step+MV	0,80341 (0,07638)	0,69550 (0,11887)	0,52032 (0,13380)
$n = 500$	Lasso	0,85114 (0,04135)	0,93792 (0,02184)	0,94702 (0,02202)
	Lasso+MV	0,84944 (0,04206)	0,93454 (0,02417)	0,93376 (0,04340)
	Lasso+Step	0,85013 (0,04169)	0,93115 (0,02411)	0,93448 (0,03007)
	Step+MV	0,84859 (0,04210)	0,92824 (0,02490)	0,80001 (0,09678)
$n = 1000$	Lasso	0,86025 (0,02590)	0,94149 (0,01631)	0,95532 (0,01322)
	Lasso+MV	0,85937 (0,02567)	0,93936 (0,01741)	0,95177 (0,01478)
	Lasso+Step	0,85975 (0,02572)	0,93904 (0,01748)	0,95016 (0,01490)
	Step+MV	0,85900 (0,02573)	0,93797 (0,01781)	0,94884 (0,01537)
$n = 5000$	Lasso	0,86025 (0,01177)	0,94424 (0,00644)	0,96618 (0,00407)
	Lasso+MV	0,86008 (0,01177)	0,94374 (0,00654)	0,96562 (0,00420)
	Lasso+Step	0,86008 (0,01179)	0,94366 (0,00650)	0,96550 (0,00417)
	Step+MV	0,86000 (0,01183)	0,94347 (0,00650)	0,96533 (0,00418)

Tabela A.4: *Resultados da média do coeficiente de Gini e seu desvio padrão para os cenários com correlação 0,2 e 40% de covariáveis importantes.*

$\rho = 0{,}2$ 40% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	0,92872 (0,04070)	0,91275 (0,06011)	0,88736 (0,06289)
	Lasso+MV	0,92662 (0,04376)	0,80342 (0,09625)	0,81523 (0,09009)
	Lasso+Step	0,92921 (0,03983)	0,86180 (0,09639)	0,73058 (0,11369)
	Step+MV	0,92459 (0,04477)	0,69093 (0,11964)	0,55561 (0,12849)
$n = 500$	Lasso	0,91824 (0,02717)	0,95811 (0,01830)	0,95922 (0,01794)
	Lasso+MV	0,91731 (0,02770)	0,95171 (0,02789)	0,86183 (0,06109)
	Lasso+Step	0,91835 (0,02739)	0,95703 (0,01900)	0,94652 (0,03242)
	Step+MV	0,91763 (0,02782)	0,93981 (0,05317)	0,79959 (0,06612)
$n = 1000$	Lasso	0,91288 (0,02088)	0,96210 (0,01123)	0,97405 (0,00897)
	Lasso+MV	0,91249 (0,02074)	0,96165 (0,01130)	0,96611 (0,02812)
	Lasso+Step	0,91286 (0,02076)	0,96320 (0,01115)	0,97293 (0,00947)
	Step+MV	0,91250 (0,02075)	0,96310 (0,01117)	0,95459 (0,04659)
$n = 5000$	Lasso	0,92115 (0,00822)	0,97256 (0,00391)	0,98336 (0,00273)
	Lasso+MV	0,92104 (0,00826)	0,97232 (0,00392)	0,98292 (0,00279)
	Lasso+Step	0,92114 (0,00823)	0,97256 (0,00389)	0,98323 (0,00275)
	Step+MV	0,92105 (0,00823)	0,97249 (0,00390)	0,98312 (0,00276)

Tabela A.5: *Resultados da média do coeficiente de Gini e seu desvio padrão para os cenários com correlação 0,2 e 60% de covariáveis importantes.*

$\rho = 0{,}2$ 60% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	0,92193 (0,04573)	0,93293 (0,07147)	0,84451 (0,09176)
	Lasso+MV	0,92123 (0,04993)	0,84949 (0,08540)	0,77157 (0,10069)
	Lasso+Step	0,92416 (0,04608)	0,84563 (0,09062)	0,70928 (0,11481)
	Step+MV	0,92216 (0,05083)	0,66779 (0,11023)	0,54186 (0,12082)
$n = 500$	Lasso	0,94128 (0,02093)	0,96151 (0,01696)	0,95060 (0,04716)
	Lasso+MV	0,94153 (0,02112)	0,92655 (0,06327)	0,85681 (0,04783)
	Lasso+Step	0,94247 (0,02067)	0,96281 (0,01678)	0,92569 (0,03743)
	Step+MV	0,94244 (0,02072)	0,91481 (0,07237)	0,76869 (0,06491)
$n = 1000$	Lasso	0,94101 (0,01511)	0,98019 (0,00747)	0,97701 (0,00804)
	Lasso+MV	0,94095 (0,01515)	0,97877 (0,00938)	0,94119 (0,05370)
	Lasso+Step	0,94134 (0,01508)	0,98055 (0,00759)	0,97773 (0,00792)
	Step+MV	0,94129 (0,01513)	0,97958 (0,00946)	0,92974 (0,05984)
$n = 5000$	Lasso	0,94328 (0,00677)	0,98087 (0,00302)	0,98807 (0,00224)
	Lasso+MV	0,94329 (0,00680)	0,98075 (0,00302)	0,98795 (0,00226)
	Lasso+Step	0,94336 (0,00679)	0,98095 (0,00301)	0,98823 (0,00223)
	Step+MV	0,94335 (0,00679)	0,98091 (0,00301)	0,98820 (0,00224)

Tabela A.6: *Resultados da média do coeficiente de Gini e seu desvio padrão para os cenários com correlação 0,5 e 20% de covariáveis importantes.*

$\rho = 0{,}2$ 60% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	0,92193 (0,04573)	0,93293 (0,07147)	0,84451 (0,09176)
	Lasso+MV	0,92123 (0,04993)	0,84949 (0,08540)	0,77157 (0,10069)
	Lasso+Step	0,92416 (0,04608)	0,84563 (0,09062)	0,70928 (0,11481)
	Step+MV	0,92216 (0,05083)	0,66779 (0,11023)	0,54186 (0,12082)
$n = 500$	Lasso	0,94128 (0,02093)	0,96151 (0,01696)	0,95060 (0,04716)
	Lasso+MV	0,94153 (0,02112)	0,92655 (0,06327)	0,85681 (0,04783)
	Lasso+Step	0,94247 (0,02067)	0,96281 (0,01678)	0,92569 (0,03743)
	Step+MV	0,94244 (0,02072)	0,91481 (0,07237)	0,76869 (0,06491)
$n = 1000$	Lasso	0,94101 (0,01511)	0,98019 (0,00747)	0,97701 (0,00804)
	Lasso+MV	0,94095 (0,01515)	0,97877 (0,00938)	0,94119 (0,05370)
	Lasso+Step	0,94134 (0,01508)	0,98055 (0,00759)	0,97773 (0,00792)
	Step+MV	0,94129 (0,01513)	0,97958 (0,00946)	0,92974 (0,05984)
$n = 5000$	Lasso	0,94328 (0,00677)	0,98087 (0,00302)	0,98807 (0,00224)
	Lasso+MV	0,94329 (0,00680)	0,98075 (0,00302)	0,98795 (0,00226)
	Lasso+Step	0,94336 (0,00679)	0,98095 (0,00301)	0,98823 (0,00223)
	Step+MV	0,94335 (0,00679)	0,98091 (0,00301)	0,98820 (0,00224)

Tabela A.7: *Resultados da média do coeficiente de Gini e seu desvio padrão para os cenários com correlação 0,5 e 40% de covariáveis importantes.*

$\rho = 0{,}5$ 40% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	0,87176 (0,06373)	0,90787 (0,05301)	0,85971 (0,06797)
	Lasso+MV	0,87270 (0,06120)	0,78732 (0,09915)	0,75611 (0,09808)
	Lasso+Step	0,87471 (0,06067)	0,85566 (0,08745)	0,75227 (0,09481)
	Step+MV	0,87459 (0,06022)	0,67381 (0,12001)	0,56387 (0,11615)
$n = 500$	Lasso	0,87821 (0,03558)	0,95565 (0,01816)	0,96430 (0,01702)
	Lasso+MV	0,87713 (0,03587)	0,94332 (0,04192)	0,86556 (0,05737)
	Lasso+Step	0,87922 (0,03541)	0,95500 (0,01861)	0,94018 (0,03181)
	Step+MV	0,87830 (0,03590)	0,93739 (0,05293)	0,78886 (0,06163)
$n = 1000$	Lasso	0,89927 (0,02145)	0,95704 (0,01200)	0,97586 (0,00809)
	Lasso+MV	0,89886 (0,02158)	0,95564 (0,01248)	0,96121 (0,03505)
	Lasso+Step	0,89921 (0,02152)	0,95762 (0,01197)	0,97457 (0,00868)
	Step+MV	0,89910 (0,02157)	0,95720 (0,01206)	0,94475 (0,05595)
$n = 5000$	Lasso	0,88986 (0,01047)	0,96594 (0,00481)	0,98111 (0,00316)
	Lasso+MV	0,88971 (0,01049)	0,96561 (0,00482)	0,98071 (0,00322)
	Lasso+Step	0,88988 (0,01049)	0,96593 (0,00482)	0,98117 (0,00317)
	Step+MV	0,88982 (0,01050)	0,96584 (0,00483)	0,98107 (0,00319)

Tabela A.8: *Resultados da média do coeficiente de Gini e seu desvio padrão para os cenários com correlação 0,5 e 60% de covariáveis importantes.*

$\rho = 0{,}5$ 60% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	0,92551 (0,04366)	0,92099 (0,05133)	0,84987 (0,07553)
	Lasso+MV	0,92434 (0,04476)	0,83449 (0,07973)	0,76653 (0,10385)
	Lasso+Step	0,92761 (0,04206)	0,84891 (0,07849)	0,75204 (0,10231)
	Step+MV	0,92582 (0,04388)	0,67220 (0,10897)	0,56625 (0,12620)
$n = 500$	Lasso	0,91022 (0,02952)	0,95853 (0,01731)	0,95516 (0,01959)
	Lasso+MV	0,91039 (0,02931)	0,93642 (0,05317)	0,86017 (0,04819)
	Lasso+Step	0,91142 (0,02923)	0,95931 (0,01696)	0,92596 (0,03386)
	Step+MV	0,91158 (0,02925)	0,92892 (0,06238)	0,76839 (0,06373)
$n = 1000$	Lasso	0,91917 (0,01903)	0,97491 (0,00808)	0,97743 (0,00835)
	Lasso+MV	0,91897 (0,01903)	0,97407 (0,00842)	0,93379 (0,05492)
	Lasso+Step	0,91953 (0,01894)	0,97571 (0,00796)	0,97756 (0,00874)
	Step+MV	0,91950 (0,01892)	0,97523 (0,00819)	0,92040 (0,06070)
$n = 5000$	Lasso	0,92437 (0,00799)	0,97963 (0,00323)	0,98913 (0,00211)
	Lasso+MV	0,92436 (0,00800)	0,97957 (0,00323)	0,98903 (0,00214)
	Lasso+Step	0,92449 (0,00799)	0,97978 (0,00321)	0,98931 (0,00208)
	Step+MV	0,92448 (0,00799)	0,97977 (0,00321)	0,98928 (0,00209)

Tabela A.9: *Resultados da média do coeficiente de Gini e seu desvio padrão para os cenários com correlação 0,8 e 20% de covariáveis importantes.*

$\rho = 0{,}8$ 20% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	0,65496 (0,11252)	0,82292 (0,07597)	0,84924 (0,07237)
	Lasso+MV	0,64955 (0,11341)	0,77911 (0,11446)	0,73166 (0,11539)
	Lasso+Step	0,65667 (0,11002)	0,80194 (0,08823)	0,72368 (0,11374)
	Step+MV	0,65437 (0,10882)	0,70288 (0,15109)	0,51935 (0,11908)
$n = 500$	Lasso	0,69107 (0,05880)	0,86692 (0,03656)	0,91997 (0,02629)
	Lasso+MV	0,68735 (0,05918)	0,85918 (0,03873)	0,90353 (0,04024)
	Lasso+Step	0,68911 (0,05863)	0,86409 (0,03775)	0,90614 (0,03232)
	Step+MV	0,68697 (0,05884)	0,86161 (0,03856)	0,85682 (0,09173)
$n = 1000$	Lasso	0,69553 (0,04140)	0,88978 (0,02290)	0,93592 (0,01573)
	Lasso+MV	0,69399 (0,04179)	0,88530 (0,02382)	0,93018 (0,01754)
	Lasso+Step	0,69551 (0,04154)	0,88822 (0,02334)	0,93285 (0,01653)
	Step+MV	0,69456 (0,04161)	0,88709 (0,02352)	0,93134 (0,01697)
$n = 5000$	Lasso	0,69532 (0,01873)	0,89109 (0,01037)	0,94092 (0,00723)
	Lasso+MV	0,69500 (0,01881)	0,89011 (0,01049)	0,93974 (0,00739)
	Lasso+Step	0,69531 (0,01882)	0,89070 (0,01040)	0,94055 (0,00726)
	Step+MV	0,69510 (0,01887)	0,89046 (0,01040)	0,94031 (0,00728)

Tabela A.10: *Resultados da média do coeficiente de Gini e seu desvio padrão para os cenários com correlação 0,8 e 40% de covariáveis importantes.*

$\rho = 0{,}8$ 40% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	0,79817 (0,08057)	0,90480 (0,05231)	0,88875 (0,05694)
	Lasso+MV	0,80029 (0,08079)	0,80012 (0,09763)	0,79379 (0,09298)
	Lasso+Step	0,80411 (0,08039)	0,85997 (0,07395)	0,81426 (0,08232)
	Step+MV	0,80458 (0,08035)	0,67674 (0,10442)	0,62005 (0,11307)
$n = 500$	Lasso	0,81703 (0,04672)	0,93663 (0,02392)	0,95346 (0,02061)
	Lasso+MV	0,81597 (0,04727)	0,93277 (0,02589)	0,84796 (0,06908)
	Lasso+Step	0,81872 (0,04699)	0,93729 (0,02416)	0,92901 (0,03322)
	Step+MV	0,81818 (0,04699)	0,93508 (0,02770)	0,77004 (0,06329)
$n = 1000$	Lasso	0,81829 (0,03336)	0,95240 (0,01378)	0,97250 (0,00962)
	Lasso+MV	0,81816 (0,03321)	0,95032 (0,01439)	0,96497 (0,02482)
	Lasso+Step	0,81937 (0,03324)	0,95264 (0,01393)	0,97093 (0,01083)
	Step+MV	0,81935 (0,03326)	0,95198 (0,01418)	0,95629 (0,04209)
$n = 5000$	Lasso	0,82851 (0,01320)	0,95683 (0,00572)	0,97797 (0,00333)
	Lasso+MV	0,82837 (0,01321)	0,95646 (0,00581)	0,97756 (0,00336)
	Lasso+Step	0,82860 (0,01320)	0,95695 (0,00571)	0,97810 (0,00335)
	Step+MV	0,82856 (0,01320)	0,95687 (0,00574)	0,97801 (0,00337)

Tabela A.11: *Resultados da média do coeficiente de Gini e seu desvio padrão para os cenários com correlação 0,8 e 60% de covariáveis importantes.*

$\rho = 0{,}8$ 60% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	0,86683 (0,06202)	0,90131 (0,05774)	0,92797 (0,03996)
	Lasso+MV	0,86853 (0,06056)	0,79730 (0,09047)	0,87108 (0,08209)
	Lasso+Step	0,86860 (0,06175)	0,85284 (0,07356)	0,88523 (0,06468)
	Step+MV	0,86982 (0,06115)	0,66713 (0,10930)	0,69779 (0,09802)
$n = 500$	Lasso	0,88089 (0,03347)	0,95604 (0,01917)	0,96178 (0,01769)
	Lasso+MV	0,88212 (0,03345)	0,94133 (0,04508)	0,88012 (0,04746)
	Lasso+Step	0,88336 (0,03267)	0,95582 (0,02090)	0,93317 (0,03007)
	Step+MV	0,88386 (0,03284)	0,93555 (0,05814)	0,78259 (0,05642)
$n = 1000$	Lasso	0,89127 (0,02287)	0,97117 (0,00958)	0,97870 (0,00783)
	Lasso+MV	0,89119 (0,02297)	0,97078 (0,00976)	0,93080 (0,05562)
	Lasso+Step	0,89193 (0,02276)	0,97210 (0,00937)	0,97709 (0,00878)
	Step+MV	0,89200 (0,02288)	0,97189 (0,00950)	0,91520 (0,06153)
$n = 5000$	Lasso	0,88946 (0,01039)	0,97851 (0,00345)	0,98897 (0,00205)
	Lasso+MV	0,88938 (0,01040)	0,97841 (0,00346)	0,98886 (0,00210)
	Lasso+Step	0,88959 (0,01037)	0,97862 (0,00342)	0,98915 (0,00205)
	Step+MV	0,88954 (0,01038)	0,97859 (0,00342)	0,98913 (0,00206)

Apêndice B

Resultados completos dos estudos de simulação para o número médio de covariáveis selecionadas

Tabela B.1: *Resultados da média do número covariáveis selecionados para os cenários com correlação 0 e 40% de covariáveis importantes.*

$\rho = 0$ 40% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	6,528	22,428	33,484
	Lasso+MV	6,528	22,428	33,484
	Lasso+Step	5,086	15,730	22,920
	Step+MV	5,208	16,150	23,428
$n = 500$	Lasso	5,702	18,696	33,688
	Lasso+MV	5,702	18,696	33,688
	Lasso+Step	4,882	17,062	26,576
	Step+MV	5,022	18,168	27,366
$n = 1000$	Lasso	5,928	18,866	33,840
	Lasso+MV	5,928	18,866	33,840
	Lasso+Step	4,912	15,458	29,474
	Step+MV	4,942	15,476	30,206
$n = 5000$	Lasso	5,096	16,416	27,762
	Lasso+MV	5,096	16,416	27,762
	Lasso+Step	4,860	14,942	25,216
	Step+MV	4,954	14,944	25,218

Tabela B.2: *Resultados da média do número covariáveis selecionados para os cenários com correlação 0 e 60% de covariáveis importantes.*

$\rho = 0$ 60% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	8,344	25,826	41,438
	Lasso+MV	8,344	25,826	41,438
	Lasso+Step	6,858	17,724	25,190
	Step+MV	6,920	17,870	25,344
$n = 500$	Lasso	7,596	27,046	45,290
	Lasso+MV	7,596	27,046	45,290
	Lasso+Step	6,668	21,860	32,080
	Step+MV	6,732	22,076	32,162
$n = 1000$	Lasso	7,380	26,266	44,446
	Lasso+MV	7,380	26,266	44,446
	Lasso+Step	6,564	20,640	37,410
	Step+MV	6,610	20,644	38,118
$n = 5000$	Lasso	7,080	23,020	41,614
	Lasso+MV	7,080	23,020	41,614
	Lasso+Step	6,630	19,970	33,570
	Step+MV	6,668	19,980	33,570

Tabela B.3: *Resultados da média do número covariáveis selecionados para os cenários com correlação 0,2 e 20% de covariáveis importantes.*

$\rho = 0{,}2$ 20% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	4,156	14,226	24,992
	Lasso+MV	4,156	14,226	24,992
	Lasso+Step	3,270	13,026	20,336
	Step+MV	3,524	15,160	22,166
$n = 500$	Lasso	3,928	10,116	19,174
	Lasso+MV	3,928	10,116	19,174
	Lasso+Step	3,254	11,146	23,034
	Step+MV	3,398	11,386	29,792
$n = 1000$	Lasso	3,478	10,534	18,444
	Lasso+MV	3,478	10,534	18,444
	Lasso+Step	3,168	10,400	18,608
	Step+MV	3,288	10,466	18,652
$n = 5000$	Lasso	3,152	8,996	16,158
	Lasso+MV	3,152	8,996	16,158
	Lasso+Step	3,122	9,882	16,684
	Step+MV	3,278	9,928	16,688

Tabela B.4: *Resultados da média do número covariáveis selecionados para os cenários com correlação 0,2 e 40% de covariáveis importantes.*

$\rho = 0{,}2$ 40% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	5,998	22,242	35,568
	Lasso+MV	5,998	22,242	35,568
	Lasso+Step	5,048	16,042	21,590
	Step+MV	5,284	16,498	21,806
$n = 500$	Lasso	5,862	20,614	37,044
	Lasso+MV	5,862	20,614	37,044
	Lasso+Step	4,932	17,218	26,802
	Step+MV	5,010	17,604	27,588
$n = 1000$	Lasso	5,674	21,700	35,794
	Lasso+MV	5,674	21,700	35,794
	Lasso+Step	4,948	15,420	29,340
	Step+MV	5,034	15,422	29,870
$n = 5000$	Lasso	5,494	18,240	30,660
	Lasso+MV	5,494	18,240	30,660
	Lasso+Step	4,886	15,040	25,262
	Step+MV	4,938	15,040	25,262

Tabela B.5: *Resultados da média do número covariáveis selecionados para os cenários com correlação 0,2 e 60% de covariáveis importantes.*

$\rho = 0{,}2$ 60% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	8,264	25,468	39,438
	Lasso+MV	8,264	25,468	39,438
	Lasso+Step	6,844	17,572	23,144
	Step+MV	6,886	17,664	23,242
$n = 500$	Lasso	8,042	27,046	45,310
	Lasso+MV	8,042	27,046	45,310
	Lasso+Step	6,626	21,990	31,900
	Step+MV	6,668	22,258	31,992
$n = 1000$	Lasso	7,772	25,872	45,158
	Lasso+MV	7,772	25,872	45,158
	Lasso+Step	6,678	20,632	37,430
	Step+MV	6,704	20,668	37,888
$n = 5000$	Lasso	7,618	24,642	43,338
	Lasso+MV	7,618	24,642	43,338
	Lasso+Step	6,648	20,054	33,604
	Step+MV	6,648	20,054	33,604

Tabela B.6: *Resultados da média do número covariáveis selecionados para os cenários com correlação 0,5 e 20% de covariáveis importantes.*

$\rho = 0{,}5$ 20% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	3,876	15,348	23,900
	Lasso+MV	3,876	15,348	23,900
	Lasso+Step	3,290	14,870	21,588
	Step+MV	3,514	16,768	22,896
$n = 500$	Lasso	3,640	13,280	20,504
	Lasso+MV	3,640	13,280	20,504
	Lasso+Step	3,126	11,026	22,328
	Step+MV	3,320	11,078	28,892
$n = 1000$	Lasso	3,742	11,080	21,044
	Lasso+MV	3,742	11,080	21,044
	Lasso+Step	3,288	10,416	18,502
	Step+MV	3,390	10,492	18,510
$n = 5000$	Lasso	3,484	10,802	18,534
	Lasso+MV	3,484	10,802	18,534
	Lasso+Step	3,232	9,926	16,928
	Step+MV	3,324	9,926	16,928

Tabela B.7: *Resultados da média do número covariáveis selecionados para os cenários com correlação 0,5 e 40% de covariáveis importantes.*

$\rho = 0{,}5$ 40% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	6,660	22,156	32,484
	Lasso+MV	6,660	22,156	32,484
	Lasso+Step	5,134	15,802	20,728
	Step+MV	5,196	16,166	21,372
$n = 500$	Lasso	6,410	21,502	36,364
	Lasso+MV	6,410	21,502	36,364
	Lasso+Step	5,052	16,982	26,676
	Step+MV	5,092	17,310	27,284
$n = 1000$	Lasso	5,960	21,588	37,070
	Lasso+MV	5,960	21,588	37,070
	Lasso+Step	4,922	15,510	30,302
	Step+MV	4,966	15,512	31,050
$n = 5000$	Lasso	5,874	19,060	35,652
	Lasso+MV	5,874	19,060	35,652
	Lasso+Step	4,984	14,958	25,248
	Step+MV	5,004	14,958	25,248

Tabela B.8: *Resultados da média do número covariáveis selecionados para os cenários com correlação 0,5 e 60% de covariáveis importantes.*

$\rho = 0{,}5$ 60% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	8,012	25,508	35,120
	Lasso+MV	8,012	25,508	35,120
	Lasso+Step	6,810	16,748	20,426
	Step+MV	6,876	16,894	20,644
$n = 500$	Lasso	8,644	26,778	45,198
	Lasso+MV	8,644	26,778	45,198
	Lasso+Step	6,700	21,826	31,322
	Step+MV	6,706	21,986	31,408
$n = 1000$	Lasso	8,272	26,242	45,376
	Lasso+MV	8,272	26,242	45,376
	Lasso+Step	6,686	20,600	37,474
	Step+MV	6,690	20,602	38,034
$n = 5000$	Lasso	7,960	25,544	44,938
	Lasso+MV	7,960	25,544	44,938
	Lasso+Step	6,598	19,946	33,490
	Step+MV	6,604	19,946	33,490

Tabela B.9: *Resultados da média do número covariáveis selecionados para os cenários com correlação 0,8 e 20% de covariáveis importantes.*

$\rho = 0{,}8$ 20% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	5,120	15,658	23,902
	Lasso+MV	5,120	15,658	23,902
	Lasso+Step	3,516	14,456	19,826
	Step+MV	3,576	15,760	21,914
$n = 500$	Lasso	4,196	14,538	24,792
	Lasso+MV	4,196	14,538	24,792
	Lasso+Step	3,398	10,788	23,632
	Step+MV	3,474	10,806	25,438
$n = 1000$	Lasso	4,196	14,016	23,870
	Lasso+MV	4,196	14,016	23,870
	Lasso+Step	3,368	10,194	18,304
	Step+MV	3,404	10,194	18,304
$n = 5000$	Lasso	4,068	13,044	23,598
	Lasso+MV	4,068	13,044	23,598
	Lasso+Step	3,390	10,020	16,658
	Step+MV	3,422	10,020	16,658

Tabela B.10: *Resultados da média do número covariáveis selecionados para os cenários com correlação 0,8 e 40% de covariáveis importantes.*

$\rho = 0{,}5$ 40% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	7,010	19,560	26,302
	Lasso+MV	7,010	19,560	26,302
	Lasso+Step	5,170	14,606	16,350
	Step+MV	5,196	15,454	17,110
$n = 500$	Lasso	6,900	21,686	36,746
	Lasso+MV	6,900	21,686	36,746
	Lasso+Step	5,016	16,192	27,562
	Step+MV	5,024	16,202	28,548
$n = 1000$	Lasso	6,788	21,484	36,608
	Lasso+MV	6,788	21,484	36,608
	Lasso+Step	4,936	15,406	29,366
	Step+MV	4,946	15,408	29,812
$n = 5000$	Lasso	6,552	20,912	36,554
	Lasso+MV	6,552	20,912	36,554
	Lasso+Step	4,954	14,914	25,142
	Step+MV	4,954	14,914	25,142

Tabela B.11: *Resultados da média do número covariáveis selecionados para os cenários com correlação 0,8 e 60% de covariáveis importantes.*

$\rho = 0{,}8$ 60% de covariáveis importantes		$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	8,438	22,064	15,544
	Lasso+MV	8,438	22,064	15,544
	Lasso+Step	6,652	15,124	11,358
	Step+MV	6,664	15,680	12,768
$n = 500$	Lasso	8,550	26,542	41,064
	Lasso+MV	8,550	26,542	41,064
	Lasso+Step	6,630	21,212	26,812
	Step+MV	6,630	21,288	26,988
$n = 1000$	Lasso	8,626	26,882	44,798
	Lasso+MV	8,626	26,882	44,798
	Lasso+Step	6,692	20,438	37,512
	Step+MV	6,692	20,438	38,132
$n = 5000$	Lasso	8,484	26,086	44,802
	Lasso+MV	8,484	26,086	44,802
	Lasso+Step	6,654	20,100	33,580
	Step+MV	6,654	20,100	33,580

Apêndice C

Resultados completos dos estudos de simulação para o módulo do viés e a raiz quadrado do EQM do modelo

Tabela C.1: *Resultados da média do módulo do viés e a média da raiz quadrada do EQM do modelo para os cenários com correlação 0 e 40% de covariáveis importantes.*

$\rho = 0$ 40% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	Viés	0,08505	0,07340	0,06131
		$\sqrt{EQM}$	0,11925	0,13126	0,13706
	Lasso+MV	Viés	0,00734	0,00494	0,00445
		$\sqrt{EQM}$	0,06933	0,13262	0,10997
	Step+Lasso	Viés	0,06367	0,04003	0,06514
		$\sqrt{EQM}$	0,10581	0,13939	0,17464
	Step+MV	Viés	0,00508	0,00469	0,00445
		$\sqrt{EQM}$	0,06972	0,13797	0,10997
$n = 500$	Lasso	Viés	0,06534	0,06785	0,07001
		$\sqrt{EQM}$	0,08872	0,10072	0,09989
	Lasso+MV	Viés	0,00263	0,00680	0,00357
		$\sqrt{EQM}$	0,04113	0,07135	0,08884
	Step+Lasso	Viés	0,03561	0,01736	0,03482
		$\sqrt{EQM}$	0,06611	0,07220	0,10342
	Step+MV	Viés	0,00241	0,00432	0,00341
		$\sqrt{EQM}$	0,04051	0,07260	0,09289
$n = 1000$	Lasso	Viés	0,05140	0,07685	0,06354
		$\sqrt{EQM}$	0,06442	0,09325	0,08048
	Lasso+MV	Viés	0,00128	0,00386	0,00519
		$\sqrt{EQM}$	0,02716	0,04182	0,05387
	Step+Lasso	Viés	0,02452	0,01650	0,01863
		$\sqrt{EQM}$	0,04673	0,04915	0,05811
	Step+MV	Viés	0,00105	0,00225	0,00311
		$\sqrt{EQM}$	0,02700	0,04252	0,06391
$n = 5000$	Lasso	Viés	0,02823	0,04960	0,05065
		$\sqrt{EQM}$	0,03361	0,05479	0,05588
	Lasso+MV	Viés	0,00050	0,00134	0,00207
		$\sqrt{EQM}$	0,01145	0,01729	0,02015
	Step+Lasso	Viés	0,01238	0,00936	0,00919
		$\sqrt{EQM}$	0,02301	0,02297	0,02288
	Step+MV	Viés	0,00046	0,00076	0,00086
		$\sqrt{EQM}$	0,01168	0,01752	0,01963

Tabela C.2: *Resultados da média do módulo do viés e a média da raiz quadrada do EQM do modelo para os cenários com correlação 0 e 60% de covariáveis importantes*

$\rho = 0$ 60% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,06598	0,05037	0,07718
		$\sqrt{EQM}$	0,10807	0,11360	0,12934
	Lasso+MV	Viés	0,00645	0,00410	0,00237
		$\sqrt{EQM}$	0,07328	0,09643	0,07715
	Step+Lasso	Viés	0,05100	0,05282	0,06585
		$\sqrt{EQM}$	0,10079	0,13782	0,16126
	Step+MV	Viés	0,00480	0,00406	0,00235
		$\sqrt{EQM}$	0,07231	0,09623	0,07592
$n = 350$	Lasso	Viés	0,04775	0,03062	0,04209
		$\sqrt{EQM}$	0,07131	0,08302	0,08117
	Lasso+MV	Viés	0,00265	0,00433	0,00310
		$\sqrt{EQM}$	0,04064	0,07858	0,07953
	Step+Lasso	Viés	0,03091	0,01253	0,03732
		$\sqrt{EQM}$	0,05791	0,07219	0,10020
	Step+MV	Viés	0,00253	0,00379	0,00317
		$\sqrt{EQM}$	0,03859	0,07652	0,07988
$n = 700$	Lasso	Viés	0,04258	0,04882	0,04670
		$\sqrt{EQM}$	0,06266	0,06983	0,07167
	Lasso+MV	Viés	0,00203	0,00310	0,00348
		$\sqrt{EQM}$	0,02864	0,04198	0,07493
	Step+Lasso	Viés	0,02317	0,01729	0,01973
		$\sqrt{EQM}$	0,04615	0,04641	0,05625
	Step+MV	Viés	0,00133	0,00250	0,00246
		$\sqrt{EQM}$	0,02728	0,04021	0,06804
$n = 3500$	Lasso	Viés	0,02701	0,03499	0,03162
		$\sqrt{EQM}$	0,03390	0,04324	0,04241
	Lasso+MV	Viés	0,00057	0,00125	0,00153
		$\sqrt{EQM}$	0,01238	0,01711	0,02089
	Step+Lasso	Viés	0,01133	0,00908	0,00762
		$\sqrt{EQM}$	0,02150	0,02120	0,02187
	Step+MV	Viés	0,00045	0,00077	0,00100
		$\sqrt{EQM}$	0,01215	0,01629	0,01947

Tabela C.3: *Resultados da média do módulo do viés e a média da raiz quadrada do EQM do modelo para os cenários com correlação 0,2 e 20% de covariáveis importantes.*

$\rho = 0{,}2$ 20% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,08730	0,10574	0,11444
		$\sqrt{EQM}$	0,11623	0,14225	0,16070
	Lasso+MV	Viés	0,00575	0,00904	0,00531
		$\sqrt{EQM}$	0,07124	0,12507	0,14298
	Step+Lasso	Viés	0,05426	0,04881	0,06478
		$\sqrt{EQM}$	0,09886	0,13909	0,19209
	Step+MV	Viés	0,00462	0,00678	0,00535
		$\sqrt{EQM}$	0,07487	0,17984	0,15236
$n = 350$	Lasso	Viés	0,06046	0,09124	0,08057
		$\sqrt{EQM}$	0,07875	0,10784	0,10599
	Lasso+MV	Viés	0,00262	0,00552	0,00912
		$\sqrt{EQM}$	0,04325	0,06233	0,09663
	Step+Lasso	Viés	0,02653	0,02227	0,02505
		$\sqrt{EQM}$	0,06115	0,07519	0,10244
	Step+MV	Viés	0,00259	0,00416	0,00556
		$\sqrt{EQM}$	0,04565	0,07171	0,13270
$n = 700$	Lasso	Viés	0,04980	0,06517	0,07297
		$\sqrt{EQM}$	0,06199	0,07684	0,08440
	Lasso+MV	Viés	0,00154	0,00414	0,00575
		$\sqrt{EQM}$	0,02889	0,04401	0,04945
	Step+Lasso	Viés	0,02415	0,01286	0,01033
		$\sqrt{EQM}$	0,04795	0,04779	0,05445
	Step+MV	Viés	0,00136	0,00249	0,00322
		$\sqrt{EQM}$	0,03141	0,04665	0,05592
$n = 3500$	Lasso	Viés	0,02341	0,03619	0,04716
		$\sqrt{EQM}$	0,02862	0,04073	0,05137
	Lasso+MV	Viés	0,00038	0,00089	0,00177
		$\sqrt{EQM}$	0,01242	0,01870	0,02119
	Step+Lasso	Viés	0,00959	0,00558	0,00571
		$\sqrt{EQM}$	0,01968	0,02053	0,02300
	Step+MV	Viés	0,00043	0,00078	0,00089
		$\sqrt{EQM}$	0,01371	0,01998	0,02275

Tabela C.4: *Resultados da média do módulo do viés e a média da raiz quadrada do EQM do modelo para os cenários com correlação 0,2 e 40% de covariáveis importantes.*

$\rho = 0{,}2$ 40% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,09782	0,06960	0,05133
		$\sqrt{EQM}$	0,13407	0,13281	0,13382
	Lasso+MV	Viés	0,00738	0,00467	0,00380
		$\sqrt{EQM}$	0,07462	0,13577	0,10676
	Step+Lasso	Viés	0,06254	0,05213	0,05991
		$\sqrt{EQM}$	0,11158	0,15515	0,17097
	Step+MV	Viés	0,00496	0,00472	0,00380
		$\sqrt{EQM}$	0,07429	0,14158	0,10591
$n = 350$	Lasso	Viés	0,05777	0,06243	0,06740
		$\sqrt{EQM}$	0,07951	0,09691	0,09620
	Lasso+MV	Viés	0,00403	0,00455	0,00639
		$\sqrt{EQM}$	0,04392	0,10051	0,07109
	Step+Lasso	Viés	0,02724	0,02530	0,02005
		$\sqrt{EQM}$	0,05899	0,10031	0,07279
	Step+MV	Viés	0,00255	0,00411	0,00422
		$\sqrt{EQM}$	0,04248	0,10573	0,07354
$n = 700$	Lasso	Viés	0,04750	0,05757	0,04933
		$\sqrt{EQM}$	0,06215	0,07833	0,06761
	Lasso+MV	Viés	0,00135	0,00569	0,00448
		$\sqrt{EQM}$	0,03081	0,05753	0,04588
	Step+Lasso	Viés	0,02072	0,01625	0,01295
		$\sqrt{EQM}$	0,04346	0,05681	0,04537
	Step+MV	Viés	0,00109	0,00300	0,00245
		$\sqrt{EQM}$	0,02999	0,05958	0,04274
$n = 3500$	Lasso	Viés	0,02521	0,03915	0,03924
		$\sqrt{EQM}$	0,03090	0,04518	0,04479
	Lasso+MV	Viés	0,00060	0,00167	0,00172
		$\sqrt{EQM}$	0,01242	0,02131	0,01825
	Step+Lasso	Viés	0,00954	0,00749	0,00767
		$\sqrt{EQM}$	0,01873	0,02191	0,02084
	Step+MV	Viés	0,00059	0,00104	0,00085
		$\sqrt{EQM}$	0,01244	0,01985	0,01777

Tabela C.5: *Resultados da média do módulo do viés e a média da raiz quadrada do EQM do modelo para os cenários com correlação 0,2 e 60% de covariáveis importantes.*

$\rho = 0{,}2$ 60% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,05114	0,05089	0,04944
		$\sqrt{EQM}$	0,10137	0,12250	0,13063
	Lasso+MV	Viés	0,00573	0,00351	0,00372
		$\sqrt{EQM}$	0,07777	0,11014	0,08728
	Step+Lasso	Viés	0,03610	0,06275	0,08028
		$\sqrt{EQM}$	0,09167	0,16551	0,22436
	Step+MV	Viés	0,00515	0,00352	0,00357
		$\sqrt{EQM}$	0,07428	0,11025	0,08361
$n = 350$	Lasso	Viés	0,04348	0,04087	0,03732
		$\sqrt{EQM}$	0,07077	0,07731	0,07522
	Lasso+MV	Viés	0,00302	0,00450	0,00273
		$\sqrt{EQM}$	0,04361	0,07732	0,07372
	Step+Lasso	Viés	0,02346	0,02133	0,03956
		$\sqrt{EQM}$	0,05492	0,06930	0,10420
	Step+MV	Viés	0,00263	0,00377	0,00273
		$\sqrt{EQM}$	0,04101	0,07876	0,07372
$n = 700$	Lasso	Viés	0,04685	0,03250	0,03446
		$\sqrt{EQM}$	0,06308	0,05894	0,05822
	Lasso+MV	Viés	0,00174	0,00362	0,00260
		$\sqrt{EQM}$	0,02864	0,04452	0,06634
	Step+Lasso	Viés	0,02675	0,01377	0,02010
		$\sqrt{EQM}$	0,04917	0,04547	0,05214
	Step+MV	Viés	0,00141	0,00267	0,00249
		$\sqrt{EQM}$	0,02755	0,04168	0,06902
$n = 3500$	Lasso	Viés	0,02279	0,02247	0,02361
		$\sqrt{EQM}$	0,02948	0,03169	0,03423
	Lasso+MV	Viés	0,00057	0,00141	0,00131
		$\sqrt{EQM}$	0,01258	0,01825	0,02103
	Step+Lasso	Viés	0,00939	0,00701	0,00741
		$\sqrt{EQM}$	0,01847	0,01939	0,02131
	Step+MV	Viés	0,00056	0,00087	0,00091
		$\sqrt{EQM}$	0,01211	0,01692	0,01949

Tabela C.6: *Resultados da média do módulo do viés e a média da raiz quadrada do EQM do modelo para os cenários com correlação 0,5 e 20% de covariáveis importantes.*

$\rho = 0{,}5$ 20% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,07364	0,07902	0,08795
		$\sqrt{EQM}$	0,11090	0,13009	0,16224
	Lasso+MV	Viés	0,00867	0,01162	0,00939
		$\sqrt{EQM}$	0,08453	0,14880	0,17613
	Step+Lasso	Viés	0,03878	0,03092	0,06575
		$\sqrt{EQM}$	0,09391	0,14872	0,24204
	Step+MV	Viés	0,00498	0,00676	0,00613
		$\sqrt{EQM}$	0,08488	0,19370	0,18699
$n = 350$	Lasso	Viés	0,05018	0,07322	0,07450
		$\sqrt{EQM}$	0,06883	0,09296	0,09872
	Lasso+MV	Viés	0,00280	0,00755	0,00959
		$\sqrt{EQM}$	0,04633	0,07005	0,08889
	Step+Lasso	Viés	0,02529	0,01861	0,03038
		$\sqrt{EQM}$	0,05899	0,07308	0,10224
	Step+MV	Viés	0,00225	0,00415	0,00490
		$\sqrt{EQM}$	0,04795	0,07419	0,13767
$n = 700$	Lasso	Viés	0,03772	0,06004	0,05807
		$\sqrt{EQM}$	0,05115	0,07350	0,07389
	Lasso+MV	Viés	0,00150	0,00430	0,00679
		$\sqrt{EQM}$	0,03343	0,05054	0,05937
	Step+Lasso	Viés	0,01549	0,01192	0,00822
		$\sqrt{EQM}$	0,03853	0,05049	0,05755
	Step+MV	Viés	0,00141	0,00258	0,00321
		$\sqrt{EQM}$	0,03424	0,05147	0,06011
$n = 3500$	Lasso	Viés	0,01880	0,02860	0,03420
		$\sqrt{EQM}$	0,02408	0,03442	0,03940
	Lasso+MV	Viés	0,00040	0,00126	0,00182
		$\sqrt{EQM}$	0,01447	0,02153	0,02456
	Step+Lasso	Viés	0,00700	0,00490	0,00479
		$\sqrt{EQM}$	0,01763	0,02114	0,02334
	Step+MV	Viés	0,00045	0,00105	0,00095
		$\sqrt{EQM}$	0,01515	0,02163	0,02389

Tabela C.7: *Resultados da média do módulo do viés e a média da raiz quadrada do EQM do modelo para os cenários com correlação 0,5 e 40% de covariáveis importantes.*

$\rho = 0{,}5$ 40% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,06377	0,05779	0,05377
		$\sqrt{EQM}$	0,10689	0,12801	0,13011
	Lasso+MV	Viés	0,00746	0,00682	0,00416
		$\sqrt{EQM}$	0,08300	0,15440	0,12207
	Step+Lasso	Viés	0,03507	0,03778	0,07392
		$\sqrt{EQM}$	0,08881	0,16273	0,20334
	Step+MV	Viés	0,00528	0,00597	0,00405
		$\sqrt{EQM}$	0,07731	0,16341	0,11995
$n = 350$	Lasso	Viés	0,03524	0,05355	0,04685
		$\sqrt{EQM}$	0,06121	0,08915	0,08273
	Lasso+MV	Viés	0,00362	0,00626	0,00364
		$\sqrt{EQM}$	0,04749	0,07714	0,10043
	Step+Lasso	Viés	0,01791	0,01698	0,02581
		$\sqrt{EQM}$	0,04957	0,07191	0,09497
	Step+MV	Viés	0,00236	0,00443	0,00367
		$\sqrt{EQM}$	0,04354	0,07286	0,10413
$n = 700$	Lasso	Viés	0,03492	0,04338	0,04198
		$\sqrt{EQM}$	0,05148	0,06095	0,06602
	Lasso+MV	Viés	0,00156	0,00441	0,00463
		$\sqrt{EQM}$	0,03219	0,04629	0,05997
	Step+Lasso	Viés	0,01416	0,01281	0,00979
		$\sqrt{EQM}$	0,03664	0,04408	0,05415
	Step+MV	Viés	0,00118	0,00250	0,00322
		$\sqrt{EQM}$	0,03048	0,04273	0,05869
$n = 3500$	Lasso	Viés	0,01864	0,02636	0,02638
		$\sqrt{EQM}$	0,02524	0,03327	0,03438
	Lasso+MV	Viés	0,00073	0,00116	0,00187
		$\sqrt{EQM}$	0,01477	0,02008	0,02310
	Step+Lasso	Viés	0,00624	0,00611	0,00603
		$\sqrt{EQM}$	0,01628	0,01955	0,02158
	Step+MV	Viés	0,00070	0,00070	0,00109
		$\sqrt{EQM}$	0,01417	0,01850	0,02078

Tabela C.8: *Resultados da média do módulo do viés e a média da raiz quadrada do EQM do modelo para os cenários com correlação 0,5 e 60% de covariáveis importantes.*

$\rho = 0{,}5$ 60% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,04168	0,04088	0,05936
		$\sqrt{EQM}$	0,09655	0,11642	0,12993
	Lasso+MV	Viés	0,00586	0,00411	0,00443
		$\sqrt{EQM}$	0,07928	0,12639	0,10049
	Step+Lasso	Viés	0,02793	0,05471	0,09288
		$\sqrt{EQM}$	0,08290	0,16302	0,21078
	Step+MV	Viés	0,00609	0,00411	0,00296
		$\sqrt{EQM}$	0,07364	0,12372	0,09024
$n = 350$	Lasso	Viés	0,03909	0,03481	0,02830
		$\sqrt{EQM}$	0,06769	0,07506	0,06970
	Lasso+MV	Viés	0,00330	0,00438	0,00270
		$\sqrt{EQM}$	0,04804	0,07618	0,07617
	Step+Lasso	Viés	0,02080	0,01529	0,03435
		$\sqrt{EQM}$	0,05422	0,06753	0,10286
	Step+MV	Viés	0,00304	0,00379	0,00270
		$\sqrt{EQM}$	0,04504	0,07493	0,07617
$n = 700$	Lasso	Viés	0,03125	0,02459	0,02801
		$\sqrt{EQM}$	0,04821	0,05347	0,05519
	Lasso+MV	Viés	0,00149	0,00293	0,00339
		$\sqrt{EQM}$	0,03109	0,04604	0,07658
	Step+Lasso	Viés	0,01754	0,01042	0,01554
		$\sqrt{EQM}$	0,03819	0,04358	0,05186
	Step+MV	Viés	0,00129	0,00237	0,00269
		$\sqrt{EQM}$	0,02931	0,04276	0,06870
$n = 3500$	Lasso	Viés	0,01520	0,02053	0,01737
		$\sqrt{EQM}$	0,02307	0,02899	0,02797
	Lasso+MV	Viés	0,00058	0,00116	0,00114
		$\sqrt{EQM}$	0,01407	0,01836	0,02047
	Step+Lasso	Viés	0,00680	0,00681	0,00591
		$\sqrt{EQM}$	0,01642	0,01921	0,02007
	Step+MV	Viés	0,00057	0,00077	0,00091
		$\sqrt{EQM}$	0,01318	0,01705	0,01886

Tabela C.9: *Resultados da média do módulo do viés e a média da raiz quadrada do EQM do modelo para os cenários com correlação 0,8 e 20% de covariáveis importantes.*

$\rho = 0{,}8$ 20% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,05181	0,06906	0,07266
		$\sqrt{EQM}$	0,09886	0,12461	0,15888
	Lasso+MV	Viés	0,00618	0,01196	0,01456
		$\sqrt{EQM}$	0,09652	0,15331	0,19948
	Step+Lasso	Viés	0,02619	0,02565	0,06097
		$\sqrt{EQM}$	0,08674	0,14944	0,23641
	Step+MV	Viés	0,00472	0,00774	0,00815
		$\sqrt{EQM}$	0,08962	0,19149	0,21538
$n = 350$	Lasso	Viés	0,03852	0,04148	0,05435
		$\sqrt{EQM}$	0,06141	0,08093	0,09159
	Lasso+MV	Viés	0,00269	0,00831	0,01011
		$\sqrt{EQM}$	0,05527	0,09221	0,10734
	Step+Lasso	Viés	0,01438	0,00905	0,01094
		$\sqrt{EQM}$	0,05278	0,07728	0,09851
	Step+MV	Viés	0,00218	0,00480	0,00554
		$\sqrt{EQM}$	0,05324	0,08323	0,11434
$n = 700$	Lasso	Viés	0,02694	0,03827	0,03883
		$\sqrt{EQM}$	0,04527	0,05840	0,06052
	Lasso+MV	Viés	0,00122	0,00374	0,00541
		$\sqrt{EQM}$	0,04190	0,05928	0,06633
	Step+Lasso	Viés	0,00956	0,00719	0,00690
		$\sqrt{EQM}$	0,03932	0,05163	0,05826
	Step+MV	Viés	0,00134	0,00300	0,00348
		$\sqrt{EQM}$	0,04016	0,05504	0,06198
$n = 3500$	Lasso	Viés	0,01389	0,01916	0,02145
		$\sqrt{EQM}$	0,02036	0,02728	0,02988
	Lasso+MV	Viés	0,00064	0,00129	0,00191
		$\sqrt{EQM}$	0,01701	0,02571	0,02824
	Step+Lasso	Viés	0,00500	0,00342	0,00388
		$\sqrt{EQM}$	0,01648	0,02232	0,02424
	Step+MV	Viés	0,00044	0,00094	0,00107
		$\sqrt{EQM}$	0,01690	0,02388	0,02553

Tabela C.10: *Resultados da média do módulo do viés e a média da raiz quadrada do EQM do modelo para os cenários com correlação 0,8 e 40% de covariáveis importantes.*

$\rho = 0{,}8$ 40% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,04718	0,05028	0,07281
		$\sqrt{EQM}$	0,09501	0,11932	0,13155
	Lasso+MV	Viés	0,00668	0,00821	0,00964
		$\sqrt{EQM}$	0,08789	0,15229	0,11198
	Step+Lasso	Viés	0,03131	0,04397	0,06424
		$\sqrt{EQM}$	0,08787	0,15945	0,16730
	Step+MV	Viés	0,00611	0,00610	0,00300
		$\sqrt{EQM}$	0,08186	0,15942	0,10215
$n = 350$	Lasso	Viés	0,02949	0,03578	0,03338
		$\sqrt{EQM}$	0,06013	0,07147	0,07288
	Lasso+MV	Viés	0,00390	0,00609	0,00367
		$\sqrt{EQM}$	0,05531	0,07593	0,09422
	Step+Lasso	Viés	0,01339	0,01226	0,02506
		$\sqrt{EQM}$	0,05093	0,06650	0,09734
	Step+MV	Viés	0,00275	0,00462	0,00368
		$\sqrt{EQM}$	0,04948	0,07250	0,09697
$n = 700$	Lasso	Viés	0,02224	0,02555	0,03163
		$\sqrt{EQM}$	0,04410	0,05149	0,05489
	Lasso+MV	Viés	0,00220	0,00338	0,00438
		$\sqrt{EQM}$	0,03960	0,05067	0,05830
	Step+Lasso	Viés	0,00962	0,00821	0,00928
		$\sqrt{EQM}$	0,03670	0,04404	0,05135
	Step+MV	Viés	0,00165	0,00255	0,00325
		$\sqrt{EQM}$	0,03600	0,04544	0,05852
$n = 3500$	Lasso	Viés	0,01204	0,01496	0,01652
		$\sqrt{EQM}$	0,02042	0,02495	0,02629
	Lasso+MV	Viés	0,00091	0,00109	0,00148
		$\sqrt{EQM}$	0,01693	0,02215	0,02300
	Step+Lasso	Viés	0,00436	0,00404	0,00414
		$\sqrt{EQM}$	0,01576	0,01945	0,02019
	Step+MV	Viés	0,00083	0,00086	0,00102
		$\sqrt{EQM}$	0,01553	0,01980	0,02047

Tabela C.11: *Resultados da média do módulo do viés e a média da raiz quadrada do EQM do modelo para os cenários com correlação 0,8 e 60% de covariáveis importantes.*

$\rho = 0{,}8$ 60% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,03921	0,05509	0,10580
		$\sqrt{EQM}$	0,09073	0,14354	0,16877
	Lasso+MV	Viés	0,00631	0,01046	0,03860
		$\sqrt{EQM}$	0,08420	0,13327	0,09881
	Step+Lasso	Viés	0,02814	0,04729	0,07567
		$\sqrt{EQM}$	0,08392	0,15681	0,15696
	Step+MV	Viés	0,00576	0,00485	0,00280
		$\sqrt{EQM}$	0,08055	0,12547	0,06454
$n = 350$	Lasso	Viés	0,02407	0,02411	0,03070
		$\sqrt{EQM}$	0,05748	0,06749	0,07670
	Lasso+MV	Viés	0,00267	0,00461	0,00269
		$\sqrt{EQM}$	0,05112	0,07991	0,08665
	Step+Lasso	Viés	0,01557	0,01440	0,03228
		$\sqrt{EQM}$	0,05101	0,06773	0,10584
	Step+MV	Viés	0,00246	0,00407	0,00273
		$\sqrt{EQM}$	0,04712	0,08097	0,08700
$n = 700$	Lasso	Viés	0,01759	0,01986	0,02092
		$\sqrt{EQM}$	0,04068	0,04906	0,05131
	Lasso+MV	Viés	0,00171	0,00282	0,00303
		$\sqrt{EQM}$	0,03593	0,04693	0,06745
	Step+Lasso	Viés	0,00890	0,00785	0,01204
		$\sqrt{EQM}$	0,03542	0,04328	0,05326
	Step+MV	Viés	0,00173	0,00237	0,00264
		$\sqrt{EQM}$	0,03333	0,04355	0,06862
$n = 3500$	Lasso	Viés	0,00853	0,01077	0,01130
		$\sqrt{EQM}$	0,01813	0,02171	0,02286
	Lasso+MV	Viés	0,00068	0,00089	0,00114
		$\sqrt{EQM}$	0,01530	0,01898	0,02033
	Step+Lasso	Viés	0,00417	0,00428	0,00433
		$\sqrt{EQM}$	0,01485	0,01782	0,01895
	Step+MV	Viés	0,00066	0,00081	0,00098
		$\sqrt{EQM}$	0,01400	0,01750	0,01871

Apêndice D

Resultados completos dos estudos de simulação para a média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado igual a zero

Tabela D.1: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado igual a zero para os cenários com correlação 0 e 40% de covariáveis importantes.*

$\rho = 0$ 40% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,02282	0,09059	0,16671
		$\sqrt{EQM}$	0,23942	0,43403	0,43179
	Lasso+MV	Viés	0,28818	5,63e+11	2,89001
		$\sqrt{EQM}$	6,18501	9,73e+12	7,08707
	Step+Lasso	Viés	0,00853	0,53533	1,10474
		$\sqrt{EQM}$	0,22133	4,10820	5,40389
	Step+MV	Viés	2,09821	255,6397	759,1928
		$\sqrt{EQM}$	47,88681	3443,221	12158,97
$n = 350$	Lasso	Viés	0,00314	0,02441	0,04196
		$\sqrt{EQM}$	0,10751	0,19254	0,20630
	Lasso+MV	Viés	0,00873	7,49e+10	4,38e+11
		$\sqrt{EQM}$	0,18483	1,68e+12	8,22e+12
	Step+Lasso	Viés	0,00472	0,01834	0,14646
		$\sqrt{EQM}$	0,12052	0,46876	1,43520
	Step+MV	Viés	0,00628	23,30299	1228,962
		$\sqrt{EQM}$	0,17824	733,4966	27472,73
$n = 700$	Lasso	Viés	0,00204	0,00789	0,01845
		$\sqrt{EQM}$	0,05919	0,07075	0,10092
	Lasso+MV	Viés	0,00527	0,02440	2,45e+11
		$\sqrt{EQM}$	0,11382	0,18533	6,83e+12
	Step+Lasso	Viés	0,00228	0,00668	0,01534
		$\sqrt{EQM}$	0,08317	0,14148	0,48648
	Step+MV	Viés	0,00319	0,00770	1761,313
		$\sqrt{EQM}$	0,11231	0,19165	44168,83
$n = 3500$	Lasso	Viés	0,00086	0,00231	0,00384
		$\sqrt{EQM}$	0,02503	0,02138	0,02361
	Lasso+MV	Viés	0,00177	0,00689	0,01198
		$\sqrt{EQM}$	0,04606	0,06071	0,08196
	Step+Lasso	Viés	0,00103	0,00201	0,00215
		$\sqrt{EQM}$	0,03760	0,05142	0,06142
	Step+MV	Viés	0,00122	0,00242	0,00243
		$\sqrt{EQM}$	0,04802	0,06254	0,07775

Tabela D.2: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado igual a zero para os cenários com correlação 0 e 60% de covariáveis importantes.*

$\rho = 0$ 60% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,04071	0,15850	0,11355
		$\sqrt{EQM}$	0,34092	0,58274	0,29791
	Lasso+MV	Viés	0,46810	1,08e+11	2,16795
		$\sqrt{EQM}$	12,667	2,43e+12	5,39200
	Step+Lasso	Viés	0,03064	0,51653	0,84643
		$\sqrt{EQM}$	0,33709	3,69442	4,86995
	Step+MV	Viés	2,60039	338,2444	2,10e+7
		$\sqrt{EQM}$	62,48737	5562,869	3,32e+8
$n = 350$	Lasso	Viés	0,00495	0,04600	0,07845
		$\sqrt{EQM}$	0,13165	0,68922	0,31993
	Lasso+MV	Viés	0,01075	5,97e+11	2,24321
		$\sqrt{EQM}$	0,22734	1,32e+13	10,23536
	Step+Lasso	Viés	0,00787	0,03469	0,13586
		$\sqrt{EQM}$	0,13300	0,64005	1,12852
	Step+MV	Viés	0,01270	31,31595	111,4992
		$\sqrt{EQM}$	0,19696	795,532	1702,38
$n = 700$	Lasso	Viés	0,00983	0,00928	0,029190
		$\sqrt{EQM}$	0,09599	0,11818	0,24197
	Lasso+MV	Viés	0,01086	0,01891	1,02e+12
		$\sqrt{EQM}$	0,14105	0,30902	1,851307e+13
	Step+Lasso	Viés	0,00470	0,00666	0,01836
		$\sqrt{EQM}$	0,08695	0,17861	0,35883
	Step+MV	Viés	0,00418	0,00814	1299,746
		$\sqrt{EQM}$	0,12151	0,27606	34114,94
$n = 3500$	Lasso	Viés	0,00198	0,00529	0,00775
		$\sqrt{EQM}$	0,03685	0,04196	0,05494
	Lasso+MV	Viés	0,00295	0,00954	0,01113
		$\sqrt{EQM}$	0,05848	0,08687	0,10736
	Step+Lasso	Viés	0,00151	0,00320	0,00268
		$\sqrt{EQM}$	0,04407	0,05943	0,06966
	Step+MV	Viés	0,00158	0,00378	0,00311
		$\sqrt{EQM}$	0,05510	0,07419	0,08636

Tabela D.3: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado igual a zero para os cenários com correlação 0,2 e 20% de covariáveis importantes.*

$\rho = 0{,}2$ 20% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,00865	0,01927	0,04395
		$\sqrt{EQM}$	0,13760	0,25825	0,22892
	Lasso+MV	Viés	0,01774	1,50e+10	1,97e+11
		$\sqrt{EQM}$	0,27965	3,36e+11	4,33e+12
	Step+Lasso	Viés	0,00709	0,15726	0,91535
		$\sqrt{EQM}$	0,18244	3,59613	6,63696
	Step+MV	Viés	0,01342	1,47e+7	389,1407
		$\sqrt{EQM}$	0,29484	3,28e+8	4426,303
$n = 350$	Lasso	Viés	0,00272	0,00534	0,01988
		$\sqrt{EQM}$	0,06815	0,06626	0,14022
	Lasso+MV	Viés	0,00542	0,01862	1,87e+11
		$\sqrt{EQM}$	0,12287	0,22556	4,09e+12
	Step+Lasso	Viés	0,00449	0,00777	0,05349
		$\sqrt{EQM}$	0,10158	0,20906	1,33142
	Step+MV	Viés	0,00559	0,01359	3,70e+11
		$\sqrt{EQM}$	0,13269	0,29559	9,30e+12
$n = 700$	Lasso	Viés	0,00218	0,00474	0,00700
		$\sqrt{EQM}$	0,04580	0,04201	0,04054
	Lasso+MV	Viés	0,00358	0,01621	0,03089
		$\sqrt{EQM}$	0,08207	0,13166	0,19447
	Step+Lasso	Viés	0,00251	0,00505	0,00880
		$\sqrt{EQM}$	0,07131	0,11726	0,20475
	Step+MV	Viés	0,00343	0,00573	0,01318
		$\sqrt{EQM}$	0,09333	0,14644	0,31386
$n = 3500$	Lasso	Viés	0,00047	0,00086	0,001581
		$\sqrt{EQM}$	0,02084	0,01833	0,01649
	Lasso+MV	Viés	0,00098	0,00224	0,00576
		$\sqrt{EQM}$	0,03577	0,05004	0,05695
	Step+Lasso	Viés	0,00115	0,00160	0,00156
		$\sqrt{EQM}$	0,03299	0,04940	0,05640
	Step+MV	Viés	0,00126	0,00193	0,00174
		$\sqrt{EQM}$	0,04174	0,05641	0,06451

Tabela D.4: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado igual a zero para os cenários com correlação 0,2 e 40% de covariáveis importantes.*

$\rho = 0{,}2$ 40% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,01257	0,10270	0,16660
		$\sqrt{EQM}$	0,18722	0,58870	0,43990
	Lasso+MV	Viés	0,02147	1,68e+11	1,57e+11
		$\sqrt{EQM}$	0,38308	3,56e+12	3,50e+12
	Step+Lasso	Viés	0,00577	0,27674	1,25151
		$\sqrt{EQM}$	0,21270	3,32834	5,50501
	Step+MV	Viés	0,01097	5197,159	1,62e+11
		$\sqrt{EQM}$	0,38421	115671,8	3,63e+12
$n = 350$	Lasso	Viés	0,01012	0,01828	0,04756
		$\sqrt{EQM}$	0,10094	0,22847	0,24022
	Lasso+MV	Viés	0,02089	4,01e+11	7,70e+11
		$\sqrt{EQM}$	0,17799	1,09e+13	1,44e+13
	Step+Lasso	Viés	0,00564	0,01237	0,15150
		$\sqrt{EQM}$	0,12037	0,456033	1,42094
	Step+MV	Viés	0,00722	355,2289	673,3846
		$\sqrt{EQM}$	0,16525	8277,622	14568,78
$n = 700$	Lasso	Viés	0,00213	0,01494	0,02584
		$\sqrt{EQM}$	0,06423	0,09670	0,12037
	Lasso+MV	Viés	0,00500	0,02924	1,74e+11
		$\sqrt{EQM}$	0,11352	0,23440	4,85e+12
	Step+Lasso	Viés	0,00252	0,00636	0,01200
		$\sqrt{EQM}$	0,08331	0,15604	0,31401
	Step+MV	Viés	0,00243	0,00773	102,4604
		$\sqrt{EQM}$	0,10732	0,20295	2287,875
$n = 3500$	Lasso	Viés	0,00140	0,00387	0,00401
		$\sqrt{EQM}$	0,02831	0,02843	0,03229
	Lasso+MV	Viés	0,00278	0,00945	0,00891
		$\sqrt{EQM}$	0,04772	0,07391	0,09644
	Step+Lasso	Viés	0,00192	0,00244	0,00260
		$\sqrt{EQM}$	0,03868	0,05898	0,06841
	Step+MV	Viés	0,00259	0,00255	0,00336
		$\sqrt{EQM}$	0,04814	0,06993	0,08376

Tabela D.5: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado igual a zero para os cenários com correlação 0,2 e 60% de covariáveis importantes.*

ρ = 0,2 60% de covariáveis importantes			p = 10	p = 30	p = 50
n = 140	Lasso	Viés	0,00713	0,14763	0,25422
		$\sqrt{EQM}$	0,45629	0,61713	0,52045
	Lasso+MV	Viés	0,86792	3,01347	3,29475
		$\sqrt{EQM}$	19,19554	17,03503	6,42099
	Step+Lasso	Viés	0,01368	0,63882	1,11170
		$\sqrt{EQM}$	0,28162	3,75207	5,35900
	Step+MV	Viés	1,81029	4,96e+11	9111822
		$\sqrt{EQM}$	46,3961	1,11e+13	2,04e+8
n = 350	Lasso	Viés	0,00920	0,04236	0,08613
		$\sqrt{EQM}$	0,14424	0,41329	0,35083
	Lasso+MV	Viés	0,01574	9,72e+11	2,17653
		$\sqrt{EQM}$	0,22802	1,68e+13	9,50488
	Step+Lasso	Viés	0,00657	0,02958	0,21552
		$\sqrt{EQM}$	0,14153	0,71044	1,44277
	Step+MV	Viés	0,00879	78,24471	2208,198
		$\sqrt{EQM}$	0,19013	2256,712	48051,83
n = 700	Lasso	Viés	0,00514	0,02780	0,04121
		$\sqrt{EQM}$	0,08833	0,16396	0,21831
	Lasso+MV	Viés	0,01043	0,03180	5,43e+11
		$\sqrt{EQM}$	0,14734	0,29466	1,58e+13
	Step+Lasso	Viés	0,00339	0,00981	0,01806
		$\sqrt{EQM}$	0,09555	0,17397	0,30062
	Step+MV	Viés	0,00292	0,01124	613,7037
		$\sqrt{EQM}$	0,13092	0,24070	15987,64
n = 3500	Lasso	Viés	0,00103	0,00835	0,01028
		$\sqrt{EQM}$	0,03776	0,05682	0,06775
	Lasso+MV	Viés	0,00165	0,00852	0,01032
		$\sqrt{EQM}$	0,05917	0,09445	0,12269
	Step+Lasso	Viés	0,00129	0,00304	0,00252
		$\sqrt{EQM}$	0,04243	0,06379	0,07708
	Step+MV	Viés	0,00180	0,00331	0,00313
		$\sqrt{EQM}$	0,05229	0,07478	0,09724

Tabela D.6: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado igual a zero para os cenários com correlação 0,5 e 20% de covariáveis importantes.*

$\rho = 0{,}5$ 20% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,01448	0,03517	0,08103
		$\sqrt{EQM}$	0,17843	0,32533	0,44001
	Lasso+MV	Viés	0,02743	9,78e+11	5,19e+11
		$\sqrt{EQM}$	0,30009	2,17e+13	1,01e+13
	Step+Lasso	Viés	0,00788	0,13147	1,37134
		$\sqrt{EQM}$	0,21641	3,71103	7,00009
	Step+MV	Viés	0,01026	1537,049	4,70e+11
		$\sqrt{EQM}$	0,30598	35305,5	1,04e+13
$n = 350$	Lasso	Viés	0,00445	0,01079	0,02045
		$\sqrt{EQM}$	0,08963	0,08586	0,10123
	Lasso+MV	Viés	0,00987	0,03239	2,20e+11
		$\sqrt{EQM}$	0,15880	0,27346	5,29e+12
	Step+Lasso	Viés	0,00484	0,00969	0,06697
		$\sqrt{EQM}$	0,12836	0,23715	1,49296
	Step+MV	Viés	0,00518	0,01214	2,86e+11
		$\sqrt{EQM}$	0,16983	0,31675	6,40e+12
$n = 700$	Lasso	Viés	0,00244	0,00588	0,01169
		$\sqrt{EQM}$	0,06347	0,05767	0,06854
	Lasso+MV	Viés	0,00479	0,01652	0,03556
		$\sqrt{EQM}$	0,10765	0,15550	0,24280
	Step+Lasso	Viés	0,00332	0,00509	0,00958
		$\sqrt{EQM}$	0,08721	0,13547	0,21273
	Step+MV	Viés	0,00421	0,00597	0,01037
		$\sqrt{EQM}$	0,11261	0,16365	0,26079
$n = 3500$	Lasso	Viés	0,00075	0,00164	0,00246
		$\sqrt{EQM}$	0,02511	0,02676	0,02571
	Lasso+MV	Viés	0,00099	0,00427	0,00699
		$\sqrt{EQM}$	0,04422	0,06489	0,08057
	Step+Lasso	Viés	0,00144	0,00275	0,00259
		$\sqrt{EQM}$	0,03888	0,05779	0,06909
	Step+MV	Viés	0,00179	0,00318	0,00291
		$\sqrt{EQM}$	0,04811	0,06644	0,07863

Tabela D.7: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado igual a zero para os cenários com correlação 0,5 e 40% de covariáveis importantes.*

$\rho = 0{,}5$ 40% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,02007	0,10440	0,17502
		$\sqrt{EQM}$	0,23476	0,79103	0,54369
	Lasso+MV	Viés	0,03296	9,72e+11	9,12e+10
		$\sqrt{EQM}$	0,43290	2,37e+13	2,04e+12
	Step+Lasso	Viés	0,01058	0,51305	0,92722
		$\sqrt{EQM}$	0,26115	4,74147	4,4784
	Step+MV	Viés	0,01154	340,4201	3,18e+8
		$\sqrt{EQM}$	0,37732	8273,358	7,12e+9
$n = 350$	Lasso	Viés	0,01520	0,02305	0,05525
		$\sqrt{EQM}$	0,16039	0,22542	0,32495
	Lasso+MV	Viés	0,02276	1,95e+11	4,80e+11
		$\sqrt{EQM}$	0,25170	4,38e+12	1,23e+13
	Step+Lasso	Viés	0,00539	0,015809	0,11524
		$\sqrt{EQM}$	0,16855	0,33855	1,34001
	Step+MV	Viés	0,00698	1,61417	7,95e+9
		$\sqrt{EQM}$	0,21465	50,10747	1,77e+11
$n = 700$	Lasso	Viés	0,00433	0,01739	0,02406
		$\sqrt{EQM}$	0,09115	0,11054	0,17676
	Lasso+MV	Viés	0,00661	0,03675	2,63e+11
		$\sqrt{EQM}$	0,14626	0,27991	6,28e+12
	Step+Lasso	Viés	0,00319	0,00823	0,014530
		$\sqrt{EQM}$	0,10369	0,18244	0,40645
	Step+MV	Viés	0,00336	0,01015	79,60571
		$\sqrt{EQM}$	0,13090	0,23930	1908,541
$n = 3500$	Lasso	Viés	0,00095	0,00321	0,00855
		$\sqrt{EQM}$	0,03727	0,04478	0,05539
	Lasso+MV	Viés	0,00240	0,00585	0,01201
		$\sqrt{EQM}$	0,06212	0,09673	0,12558
	Step+Lasso	Viés	0,00159	0,00245	0,00332
		$\sqrt{EQM}$	0,04743	0,07006	0,08655
	Step+MV	Viés	0,00220	0,00272	0,00388
		$\sqrt{EQM}$	0,05748	0,08214	0,10205

Tabela D.8: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado igual a zero para os cenários com correlação 0,5 e 60% de covariáveis importantes.*

$\rho = 0{,}5$ 60% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,01121	0,23367	0,28659
		$\sqrt{EQM}$	0,34173	0,88528	0,62958
	Lasso+MV	Viés	0,02789	12,34975	4,77819
		$\sqrt{EQM}$	0,56137	233,4623	15,64939
	Step+Lasso	Viés	0,02084	0,79159	1,10321
		$\sqrt{EQM}$	0,30133	4,19887	4,57954
	Step+MV	Viés	0,03526	406,487	1,71e+6
		$\sqrt{EQM}$	0,47593	6310,137	3,82e+7
$n = 350$	Lasso	Viés	0,00939	0,02921	0,15537
		$\sqrt{EQM}$	0,17092	0,47176	0,48127
	Lasso+MV	Viés	0,01476	8,03e+11	3,71726
		$\sqrt{EQM}$	0,26179	1,55e+13	12,83968
	Step+Lasso	Viés	0,00629	0,02453	0,30508
		$\sqrt{EQM}$	0,16637	0,63061	1,61827
	Step+MV	Viés	0,00894	26,71853	387,0412
		$\sqrt{EQM}$	0,21555	511,3851	8027,334
$n = 700$	Lasso	Viés	0,00590	0,02032	0,04140
		$\sqrt{EQM}$	0,11885	0,21776	0,36292
	Lasso+MV	Viés	0,00861	0,01877	1,42e+12
		$\sqrt{EQM}$	0,18353	0,35942	2,66e+13
	Step+Lasso	Viés	0,00394	0,00692	0,02235
		$\sqrt{EQM}$	0,12049	0,22025	0,50719
	Step+MV	Viés	0,00649	0,01133	2584,078
		$\sqrt{EQM}$	0,15424	0,28835	76110,23
$n = 3500$	Lasso	Viés	0,00186	0,00683	0,01042
		$\sqrt{EQM}$	0,05215	0,06586	0,09057
	Lasso+MV	Viés	0,00179	0,01038	0,00896
		$\sqrt{EQM}$	0,07589	0,12184	0,15976
	Step+Lasso	Viés	0,00195	0,00404	0,00280
		$\sqrt{EQM}$	0,05249	0,07938	0,10249
	Step+MV	Viés	0,00224	0,00487	0,00400
		$\sqrt{EQM}$	0,06194	0,09641	0,12463

Tabela D.9: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado igual a zero para os cenários com correlação 0,8 e 20% de covariáveis importantes.*

$\rho = 0{,}8$ 20% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,01767	0,05899	0,10851
		$\sqrt{EQM}$	0,28111	0,61639	0,64944
	Lasso+MV	Viés	0,02634	9,42e+11	1,40e+12
		$\sqrt{EQM}$	0,45318	2,02e+13	2,70e+13
	Step+Lasso	Viés	0,00991	0,14897	1,29530
		$\sqrt{EQM}$	0,31839	3,796561	7,74063
	Step+MV	Viés	0,01324	593,0504	955,6392
		$\sqrt{EQM}$	0,41812	17542,64	15608,45
$n = 350$	Lasso	Viés	0,00559	0,02623	0,03271
		$\sqrt{EQM}$	0,13293	0,20666	0,20797
	Lasso+MV	Viés	0,01039	0,04575	0,54225
		$\sqrt{EQM}$	0,22854	0,43250	14,04299
	Step+Lasso	Viés	0,00412	0,01549	0,04324
		$\sqrt{EQM}$	0,17986	0,31959	0,88362
	Step+MV	Viés	0,00743	0,01701	7651,454
		$\sqrt{EQM}$	0,21964	0,37310	171137,3
$n = 700$	Lasso	Viés	0,00230	0,00749	0,01506
		$\sqrt{EQM}$	0,09746	0,11185	0,12866
	Lasso+MV	Viés	0,00351	0,01576	0,03239
		$\sqrt{EQM}$	0,15914	0,25328	0,37747
	Step+Lasso	Viés	0,00318	0,00811	0,01037
		$\sqrt{EQM}$	0,12856	0,20129	0,29724
	Step+MV	Viés	0,00491	0,00955	0,01219
		$\sqrt{EQM}$	0,15360	0,23127	0,35276
$n = 3500$	Lasso	Viés	0,00181	0,00301	0,00492
		$\sqrt{EQM}$	0,04041	0,04586	0,05514
	Lasso+MV	Viés	0,00273	0,00616	0,01050
		$\sqrt{EQM}$	0,06803	0,10199	0,13244
	Step+Lasso	Viés	0,00149	0,00332	0,00421
		$\sqrt{EQM}$	0,05652	0,08262	0,10231
	Step+MV	Viés	0,00176	0,00383	0,00484
		$\sqrt{EQM}$	0,06923	0,09292	0,11575

Tabela D.10: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado igual a zero para os cenários com correlação 0,8 e 40% de covariáveis importantes.*

$\rho = 0{,}8$ 40% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,00736	0,21939	0,11844
		$\sqrt{EQM}$	0,38728	0,96199	0,48247
	Lasso+MV	Viés	0,01470	1,36e+12	3,09e+11
		$\sqrt{EQM}$	0,64265	2,96e+13	6,91e+12
	Step+Lasso	Viés	0,01632	0,78820	0,71582
		$\sqrt{EQM}$	0,38882	5,11464	5,08611
	Step+MV	Viés	0,02196	1,31e+12	361,7816
		$\sqrt{EQM}$	0,56444	3,35e+13	5864,736
$n = 350$	Lasso	Viés	0,01544	0,04285	0,12658
		$\sqrt{EQM}$	0,21793	0,34667	0,62639
	Lasso+MV	Viés	0,02304	1,66316	1,30e+12
		$\sqrt{EQM}$	0,33277	35,32324	2,32e+13
	Step+Lasso	Viés	0,01251	0,02306	0,34100
		$\sqrt{EQM}$	0,22248	0,58205	2,74020
	Step+MV	Viés	0,01496	61,24704	1563,356
		$\sqrt{EQM}$	0,27167	1558,667	31786,76
$n = 700$	Lasso	Viés	0,00948	0,02096	0,03791
		$\sqrt{EQM}$	0,15487	0,22357	0,23530
	Lasso+MV	Viés	0,01352	0,02814	7,59e+11
		$\sqrt{EQM}$	0,22981	0,41636	1,62e+13
	Step+Lasso	Viés	0,00529	0,01010	0,02263
		$\sqrt{EQM}$	0,16154	0,27926	0,63413
	Step+MV	Viés	0,00437	0,01218	678,4188
		$\sqrt{EQM}$	0,19230	0,33670	14646,81
$n = 3500$	Lasso	Viés	0,00329	0,00627	0,01159
		$\sqrt{EQM}$	0,06021	0,08799	0,10161
	Lasso+MV	Viés	0,00561	0,00736	0,01355
		$\sqrt{EQM}$	0,09316	0,15519	0,20412
	Step+Lasso	Viés	0,00413	0,00292	0,00580
		$\sqrt{EQM}$	0,06691	0,10942	0,13935
	Step+MV	Viés	0,00502	0,00374	0,00628
		$\sqrt{EQM}$	0,07901	0,12399	0,16258

Tabela D.11: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado igual a zero para os cenários com correlação 0,8 e 60% de covariáveis importantes.*

$\rho = 0{,}8$ 60% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,02416	0,26384	0,17134
		$\sqrt{EQM}$	0,47345	1,22606	0,53064
	Lasso+MV	Viés	0,04206	3,17e+11	3,8509
		$\sqrt{EQM}$	0,76766	8,23e+12	32,33462
	Step+Lasso	Viés	0,01353	0,89702	0,91742
		$\sqrt{EQM}$	0,46171	6,19420	4,83553
	Step+MV	Viés	0,02065	2,10e+10	1,53e+9
		$\sqrt{EQM}$	0,65150	4,71e+11	3,43e+10
$n = 350$	Lasso	Viés	0,00690	0,08470	0,18617
		$\sqrt{EQM}$	0,28946	0,84052	0,81096
	Lasso+MV	Viés	0,01252	1,12e+12	2,84e+11
		$\sqrt{EQM}$	0,40954	2,82e+13	6,35e+12
	Step+Lasso	Viés	0,00740	0,07157	0,46429
		$\sqrt{EQM}$	0,26220	1,11776	2,83073
	Step+MV	Viés	0,01100	172,3099	5,39e+11
		$\sqrt{EQM}$	0,32321	4570,694	1,20e+13
$n = 700$	Lasso	Viés	0,00624	0,03302	0,07654
		$\sqrt{EQM}$	0,18932	0,33165	0,59607
	Lasso+MV	Viés	0,01069	0,03055	2,50e+12
		$\sqrt{EQM}$	0,25972	0,55764	3,72e+13
	Step+Lasso	Viés	0,01321	0,00993	0,06923
		$\sqrt{EQM}$	0,17910	0,35619	0,99316
	Step+MV	Viés	0,01457	0,01234	3537,532
		$\sqrt{EQM}$	0,20719	0,44522	87387,24
$n = 3500$	Lasso	Viés	0,00268	0,00898	0,01832
		$\sqrt{EQM}$	0,08039	0,12895	0,15982
	Lasso+MV	Viés	0,00170	0,00764	0,01353
		$\sqrt{EQM}$	0,10927	0,19348	0,26124
	Step+Lasso	Viés	0,00156	0,00515	0,00660
		$\sqrt{EQM}$	0,07304	0,13095	0,17197
	Step+MV	Viés	0,00183	0,00636	0,00758
		$\sqrt{EQM}$	0,08384	0,15101	0,20479

Apêndice E

Resultados completos dos estudos de simulação para a média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado diferente de zero

Tabela E.1: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado diferente de zero para os cenários com correlação 0 e 40% de covariáveis importantes.*

$\rho = 0$ 40% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,95693	1,04264	1,28404
		$\sqrt{EQM}$	1,35768	1,76575	1,64536
	Lasso+MV	Viés	3,53480	6,19e+12	16,97045
		$\sqrt{EQM}$	41,78399	7,22e+13	20,91735
	Step+Lasso	Viés	0,67714	5,78991	4,45821
		$\sqrt{EQM}$	1,14885	13,30789	12,33599
	Step+MV	Viés	33,29064	2520,098	3245,042
		$\sqrt{EQM}$	381,0948	11381,22	30709,15
$n = 350$	Lasso	Viés	0,93026	1,04118	1,35519
		$\sqrt{EQM}$	1,08483	1,43672	1,64540
	Lasso+MV	Viés	0,20217	1,17e+12	1,00e+13
		$\sqrt{EQM}$	0,47017	2,61e+13	7,02e+13
	Step+Lasso	Viés	0,52819	0,63415	3,62989
		$\sqrt{EQM}$	0,76549	2,72265	8,15380
	Step+MV	Viés	0,20215	736,4087	11109,25
		$\sqrt{EQM}$	0,46886	5393,225	145960,2
$n = 700$	Lasso	Viés	0,84624	1,33853	1,40383
		$\sqrt{EQM}$	0,92859	1,42197	1,53484
	Lasso+MV	Viés	0,08022	0,40857	8,68e+12
		$\sqrt{EQM}$	0,27116	0,65407	7,17e+13
	Step+Lasso	Viés	0,41101	0,23340	0,93109
		$\sqrt{EQM}$	0,60953	0,62664	3,52642
	Step+MV	Viés	0,08123	0,46137	34286,34
		$\sqrt{EQM}$	0,27042	0,68601	520727,6
$n = 3500$	Lasso	Viés	0,56372	1,11966	1,28340
		$\sqrt{EQM}$	0,61168	1,14443	1,30013
	Lasso+MV	Viés	0,01582	0,05684	0,14865
		$\sqrt{EQM}$	0,10893	0,16064	0,23437
	Step+Lasso	Viés	0,25325	0,25802	0,28851
		$\sqrt{EQM}$	0,37380	0,38702	0,41814
	Step+MV	Viés	0,01643	0,06245	0,15347
		$\sqrt{EQM}$	0,10895	0,16331	0,23685

Tabela E.2: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado diferente de zero para os cenários com correlação 0 e 60% de covariáveis importantes.*

$\rho = 0$ 60% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,82456	0,86235	1,73085
		$\sqrt{EQM}$	1,46133	1,64912	1,86968
	Lasso+MV	Viés	7,45839	7,92e+11	12,49846
		$\sqrt{EQM}$	68,43208	1,77e+13	15,33172
	Step+Lasso	Viés	0,58380	3,58266	3,37613
		$\sqrt{EQM}$	1,24236	10,80387	10,88943
	Step+MV	Viés	59,21736	2357,416	8,37e+7
		$\sqrt{EQM}$	706,3044	19076,31	1,32e+9
$n = 350$	Lasso	Viés	0,80522	0,46126	0,83701
		$\sqrt{EQM}$	0,97823	4,35930	1,47685
	Lasso+MV	Viés	0,25880	1,45e+13	48,24265
		$\sqrt{EQM}$	0,56044	9,84e+13	55,06846
	Step+Lasso	Viés	0,53028	1,50378	1,96879
		$\sqrt{EQM}$	0,76252	4,12888	5,96230
	Step+MV	Viés	0,23900	1007,633	2002,573
		$\sqrt{EQM}$	0,54615	7287,887	9407,335
$n = 700$	Lasso	Viés	0,76540	1,06897	1,05535
		$\sqrt{EQM}$	0,92449	1,23122	1,88723
	Lasso+MV	Viés	0,10737	0,90235	3,83e+13
		$\sqrt{EQM}$	0,30890	1,26941	1,30e+14
	Step+Lasso	Viés	0,44408	0,22802	0,66564
		$\sqrt{EQM}$	0,64335	0,81574	2,80555
	Step+MV	Viés	0,10342	0,84216	28235,7
		$\sqrt{EQM}$	0,30603	1,18046	302159,3
$n = 3500$	Lasso	Viés	0,59327	0,98685	1,00880
		$\sqrt{EQM}$	0,65831	1,04980	1,10203
	Lasso+MV	Viés	0,02081	0,10399	0,21812
		$\sqrt{EQM}$	0,11995	0,19955	0,30529
	Step+Lasso	Viés	0,26457	0,30295	0,26115
		$\sqrt{EQM}$	0,37712	0,43400	0,39039
	Step+MV	Viés	0,02050	0,09823	0,19507
		$\sqrt{EQM}$	0,11992	0,19595	0,28535

Tabela E.3: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado diferente de zero para os cenários com correlação 0,2 e 20% de covariáveis importantes.*

ρ = 0,2 20% de covariáveis importantes			p = 10	p = 30	p = 50
n = 140	Lasso	Viés	0,96083	1,16453	1,41318
		$\sqrt{EQM}$	1,1516	1,74475	1,69319
	Lasso+MV	Viés	0,34762	1,87e+12	4,91e+12
		$\sqrt{EQM}$	0,79111	4,19e+13	6,75e+13
	Step+Lasso	Viés	0,55411	7,27700	9,53700
		$\sqrt{EQM}$	0,90147	18,91078	20,44676
	Step+MV	Viés	0,38495	1,60e+8	3773,81
		$\sqrt{EQM}$	0,79252	3,58e+9	13760,64
n = 350	Lasso	Viés	0,81921	1,29162	1,27549
		$\sqrt{EQM}$	0,92638	1,34932	1,52862
	Lasso+MV	Viés	0,11424	0,49510	2,79e+12
		$\sqrt{EQM}$	0,34089	0,85240	4,44e+13
	Step+Lasso	Viés	0,33957	0,05338	3,70837
		$\sqrt{EQM}$	0,59715	0,80128	9,23597
	Step+MV	Viés	0,12431	0,75912	4,49e+12
		$\sqrt{EQM}$	0,34233	1,09077	7,16e+13
n = 700	Lasso	Viés	0,72732	1,07631	1,34509
		$\sqrt{EQM}$	0,80251	1,11222	1,36694
	Lasso+MV	Viés	0,02830	0,20626	0,51081
		$\sqrt{EQM}$	0,21860	0,39040	0,77941
	Step+Lasso	Viés	0,34814	0,14126	0,20252
		$\sqrt{EQM}$	0,53590	0,42776	0,76711
	Step+MV	Viés	0,03652	0,25453	0,91158
		$\sqrt{EQM}$	0,22102	0,41695	1,80161
n = 3500	Lasso	Viés	0,40143	0,74617	1,03157
		$\sqrt{EQM}$	0,44559	0,76750	1,04534
	Lasso+MV	Viés	0,00907	0,02980	0,05898
		$\sqrt{EQM}$	0,09722	0,12103	0,15029
	Step+Lasso	Viés	0,16708	0,11328	0,11892
		$\sqrt{EQM}$	0,25798	0,20258	0,22230
	Step+MV	Viés	0,01046	0,03611	0,07590
		$\sqrt{EQM}$	0,09751	0,12297	0,15698

Tabela E.4: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado diferente de zero para os cenários com correlação 0,2 e 40% de covariáveis importantes.*

ρ = 0,2 40% de covariáveis importantes			p = 10	p = 30	p = 50
n = 140	Lasso	Viés	1,09323	0,57827	1,12654
		$\sqrt{EQM}$	1,35840	2,04920	1,52132
	Lasso+MV	Viés	0,48582	2,51e+12	1,23e+12
		$\sqrt{EQM}$	1,21950	4,04e+13	2,74e+13
	Step+Lasso	Viés	0,68363	5,84157	5,00572
		$\sqrt{EQM}$	1,09503	13,83680	12,91324
	Step+MV	Viés	0,50850	26435,5	7,80e+11
		$\sqrt{EQM}$	1,23510	518648,8	1,74e+13
n = 350	Lasso	Viés	0,90536	1,05894	1,28212
		$\sqrt{EQM}$	1,03541	1,67348	1,61415
	Lasso+MV	Viés	0,16749	6,70e+12	1,62e+13
		$\sqrt{EQM}$	0,46851	7,78e+13	1,03e+14
	Step+Lasso	Viés	0,43295	0,39652	3,45751
		$\sqrt{EQM}$	0,67222	2,26251	7,21326
	Step+MV	Viés	0,16683	5309,892	12689,8
		$\sqrt{EQM}$	0,45939	95709,04	96654,71
n = 700	Lasso	Viés	0,78835	1,03932	1,22237
		$\sqrt{EQM}$	0,87184	1,13753	1,37138
	Lasso+MV	Viés	0,05479	0,47991	3,76e+12
		$\sqrt{EQM}$	0,26735	0,69420	5,01e+13
	Step+Lasso	Viés	0,35484	0,13702	0,25925
		$\sqrt{EQM}$	0,53495	0,56621	1,79104
	Step+MV	Viés	0,05260	0,44982	1970,445
		$\sqrt{EQM}$	0,26597	0,65601	28628,25
n = 3500	Lasso	Viés	0,49169	0,94893	1,11708
		$\sqrt{EQM}$	0,54079	0,97660	1,13897
	Lasso+MV	Viés	0,01603	0,06808	0,15366
		$\sqrt{EQM}$	0,10731	0,16800	0,24211
	Step+Lasso	Viés	0,19595	0,20439	0,22509
		$\sqrt{EQM}$	0,28111	0,32435	0,35086
	Step+MV	Viés	0,01612	0,07059	0,14112
		$\sqrt{EQM}$	0,10728	0,16920	0,23156

Tabela E.5: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado diferente de zero para os cenários com correlação 0,2 e 60% de covariáveis importantes.*

ρ = 0,2 60% de covariáveis importantes			p = 10	p = 30	p = 50
n = 140	Lasso	Viés	0,62222	0,60832	1,40954
		$\sqrt{EQM}$	3,11602	1,69616	1,68953
	Lasso+MV	Viés	7,6047	42,03383	10,77489
		$\sqrt{EQM}$	120,9817	55,84776	13,55478
	Step+Lasso	Viés	0,45205	3,75502	3,58253
		$\sqrt{EQM}$	1,19367	10,62498	11,16002
	Step+MV	Viés	25,17507	1,45e+12	60430650
		$\sqrt{EQM}$	329,9107	3,24e+13	1351233773
n = 350	Lasso	Viés	0,71442	0,53140	0,89741
		$\sqrt{EQM}$	0,90745	2,34100	1,45672
	Lasso+MV	Viés	0,23996	2,20e+13	40,10643
		$\sqrt{EQM}$	0,50750	1,12e+14	44,61603
	Step+Lasso	Viés	0,37903	1,09800	1,81748
		$\sqrt{EQM}$	0,64178	4,43960	5,63417
	Step+MV	Viés	0,22112	2570,79	9760,292
		$\sqrt{EQM}$	0,49254	16347,9	165370,1
n = 700	Lasso	Viés	0,80896	0,71716	0,94583
		$\sqrt{EQM}$	0,92273	0,99741	1,31717
	Lasso+MV	Viés	0,10772	0,71247	3,04e+13
		$\sqrt{EQM}$	0,30813	0,99112	1,13e+14
	Step+Lasso	Viés	0,46996	0,16278	0,11710
		$\sqrt{EQM}$	0,67091	0,63972	1,96624
	Step+MV	Viés	0,10345	0,64117	13947,7
		$\sqrt{EQM}$	0,30586	0,90343	115803,7
n = 3500	Lasso	Viés	0,50020	0,70068	0,84438
		$\sqrt{EQM}$	0,56003	0,76953	0,93049
	Lasso+MV	Viés	0,01871	0,08506	0,23408
		$\sqrt{EQM}$	0,11538	0,18091	0,32320
	Step+Lasso	Viés	0,21850	0,22666	0,26020
		$\sqrt{EQM}$	0,30682	0,32461	0,36528
	Step+MV	Viés	0,01777	0,07633	0,20620
		$\sqrt{EQM}$	0,11477	0,17548	0,29889

Tabela E.6: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado diferente de zero para os cenários com correlação 0,5 e 20% de covariáveis importantes.*

$\rho = 0{,}5$ 20% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,85094	0,95872	1,03566
		$\sqrt{EQM}$	1,03987	1,54959	1,80695
	Lasso+MV	Viés	0,24059	6,19e+12	6,57e+12
		$\sqrt{EQM}$	0,62843	8,20e+13	7,91e+13
	Step+Lasso	Viés	0,39545	6,50815	8,36422
		$\sqrt{EQM}$	0,76016	16,40893	16,65842
	Step+MV	Viés	0,27228	23996,56	1,87e+12
		$\sqrt{EQM}$	0,65050	269722,3	4,03e+13
$n = 350$	Lasso	Viés	0,66310	1,0888	1,28924
		$\sqrt{EQM}$	0,77325	1,1528	1,35088
	Lasso+MV	Viés	0,10569	0,42861	2,97e+12
		$\sqrt{EQM}$	0,35584	0,78371	4,76e+13
	Step+Lasso	Viés	0,30543	0,02297	3,25070
		$\sqrt{EQM}$	0,55643	0,73259	7,99215
	Step+MV	Viés	0,11411	0,59673	2,19e+12
		$\sqrt{EQM}$	0,35436	0,95027	4,88e+13
$n = 700$	Lasso	Viés	0,54261	0,96150	1,15069
		$\sqrt{EQM}$	0,61976	1,00553	1,18328
	Lasso+MV	Viés	0,05193	0,17198	0,51360
		$\sqrt{EQM}$	0,22292	0,36531	0,73903
	Step+Lasso	Viés	0,21576	0,11110	0,18691
		$\sqrt{EQM}$	0,36789	0,38905	0,59930
	Step+MV	Viés	0,05387	0,20161	0,64313
		$\sqrt{EQM}$	0,22039	0,37977	0,86662
$n = 3500$	Lasso	Viés	0,30736	0,57935	0,79064
		$\sqrt{EQM}$	0,34283	0,60546	0,80443
	Lasso+MV	Viés	0,00775	0,03426	0,07597
		$\sqrt{EQM}$	0,09513	0,13245	0,16661
	Step+Lasso	Viés	0,11353	0,08893	0,08131
		$\sqrt{EQM}$	0,18720	0,17366	0,18828
	Step+MV	Viés	0,00854	0,03703	0,07773
		$\sqrt{EQM}$	0,09452	0,13033	0,16576

Tabela E.7: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado diferente de zero para os cenários com correlação 0,5 e 40% de covariáveis importantes.*

$\rho = 0{,}5$ 40% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,75205	0,44616	1,15435
		$\sqrt{EQM}$	1,05086	2,49503	1,68438
	Lasso+MV	Viés	0,47533	1,24e+13	9,80e+11
		$\sqrt{EQM}$	0,99797	1,00e+14	2,19e+13
	Step+Lasso	Viés	0,34188	7,12079	2,53758
		$\sqrt{EQM}$	0,82207	15,11409	8,99590
	Step+MV	Viés	0,43910	4502,55	8,56e+8
		$\sqrt{EQM}$	0,93608	40009,19	1,91e+10
$n = 350$	Lasso	Viés	0,54982	0,87197	0,98770
		$\sqrt{EQM}$	0,72167	1,23683	1,49615
	Lasso+MV	Viés	0,15449	1,52e+12	8,48e+12
		$\sqrt{EQM}$	0,42856	3,41e+13	7,01e+13
	Step+Lasso	Viés	0,25120	0,144079	2,58655
		$\sqrt{EQM}$	0,51272	1,06294	5,43165
	Step+MV	Viés	0,14492	23,70527	8,79e+10
		$\sqrt{EQM}$	0,41660	359,007	1,97e+12
$n = 700$	Lasso	Viés	0,57391	0,93571	0,98071
		$\sqrt{EQM}$	0,67931	1,00581	1,21462
	Lasso+MV	Viés	0,07700	0,42345	3,77e+12
		$\sqrt{EQM}$	0,27467	0,64856	5,00e+13
	Step+Lasso	Viés	0,22872	0,15295	0,62002
		$\sqrt{EQM}$	0,40079	0,51974	1,61288
	Step+MV	Viés	0,07368	0,38665	1024,591
		$\sqrt{EQM}$	0,27090	0,61443	12885,51
$n = 3500$	Lasso	Viés	0,35030	0,68927	0,84568
		$\sqrt{EQM}$	0,39460	0,72268	0,87128
	Lasso+MV	Viés	0,02277	0,06979	0,14936
		$\sqrt{EQM}$	0,11469	0,17251	0,25201
	Step+Lasso	Viés	0,11871	0,15714	0,16323
		$\sqrt{EQM}$	0,18872	0,25199	0,28109
	Step+MV	Viés	0,02213	0,06207	0,12852
		$\sqrt{EQM}$	0,11387	0,16751	0,23647

Tabela E.8: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado diferente de zero para os cenários com correlação 0,5 e 60% de covariáveis importantes.*

$\rho = 0{,}5$ 60% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,43676	0,34118	1,56747
		$\sqrt{EQM}$	1,07793	1,91010	1,8941
	Lasso+MV	Viés	0,87251	92,57681	13,76097
		$\sqrt{EQM}$	1,74437	842,644	43,47315
	Step+Lasso	Viés	0,22208	3,40430	1,79694
		$\sqrt{EQM}$	0,97276	9,42423	7,74742
	Step+MV	Viés	0,79399	2622,525	3,57e+7
		$\sqrt{EQM}$	1,68462	21974,11	7,95e+8
$n = 350$	Lasso	Viés	0,59028	0,40735	0,80704
		$\sqrt{EQM}$	0,78598	1,97906	1,39310
	Lasso+MV	Viés	0,20448	1,11e+13	38,30982
		$\sqrt{EQM}$	0,49015	9,37e+13	43,26784
	Step+Lasso	Viés	0,28739	0,79078	1,49766
		$\sqrt{EQM}$	0,56824	3,00573	4,52243
	Step+MV	Viés	0,19059	691,9445	3852,283
		$\sqrt{EQM}$	0,47781	3057,259	30755,3
$n = 700$	Lasso	Viés	0,55165	0,54686	0,65647
		$\sqrt{EQM}$	0,67255	0,85776	1,48199
	Lasso+MV	Viés	0,10134	0,63118	3,49e+13
		$\sqrt{EQM}$	0,31547	0,95180	1,26e+14
	Step+Lasso	Viés	0,30625	0,07200	0,56538
		$\sqrt{EQM}$	0,48305	0,59227	2,12970
	Step+MV	Viés	0,09317	0,54165	52990,98
		$\sqrt{EQM}$	0,31014	0,86721	490566,4
$n = 3500$	Lasso	Viés	0,00186	0,66625	0,73071
		$\sqrt{EQM}$	0,38679	0,71866	0,80888
	Lasso+MV	Viés	0,02029	0,11424	0,23641
		$\sqrt{EQM}$	0,11693	0,21935	0,33637
	Step+Lasso	Viés	0,14897	0,21728	0,21250
		$\sqrt{EQM}$	0,22199	0,32749	0,34330
	Step+MV	Viés	0,01866	0,10321	0,20478
		$\sqrt{EQM}$	0,11619	0,21245	0,31116

Tabela E.9: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado diferente de zero para os cenários com correlação 0,8 e 20% de covariáveis importantes.*

$\rho = 0{,}8$ 20% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,68839	0,76112	0,99836
		$\sqrt{EQM}$	0,95080	1,96001	1,84420
	Lasso+MV	Viés	0,21293	6,00e+12	1,23e+13
		$\sqrt{EQM}$	0,73492	9,65e+13	1,25e+14
	Step+Lasso	Viés	0,28907	3,82210	6,22635
		$\sqrt{EQM}$	0,70138	11,10891	14,23642
	Step+MV	Viés	0,20027	9143,492	7199,07
		$\sqrt{EQM}$	0,68530	61234,5	45199,33
$n = 350$	Lasso	Viés	0,56638	0,70254	0,92347
		$\sqrt{EQM}$	0,68616	0,86819	1,06043
	Lasso+MV	Viés	0,06722	0,41249	6,52317
		$\sqrt{EQM}$	0,38716	0,75943	64,08405
	Step+Lasso	Viés	0,18530	0,08037	1,02459
		$\sqrt{EQM}$	0,43175	0,59218	2,677224
	Step+MV	Viés	0,06476	0,38021	45484,95
		$\sqrt{EQM}$	0,37040	0,71184	950373,3
$n = 700$	Lasso	Viés	0,42327	0,64981	0,83382
		$\sqrt{EQM}$	0,51676	0,72908	0,90410
	Lasso+MV	Viés	0,04259	0,19217	0,52206
		$\sqrt{EQM}$	0,28663	0,43688	0,79896
	Step+Lasso	Viés	0,13295	0,01880	0,17848
		$\sqrt{EQM}$	0,32135	0,38393	0,59407
	Step+MV	Viés	0,04128	0,18144	0,50988
		$\sqrt{EQM}$	0,27429	0,42298	0,78335
$n = 3500$	Lasso	Viés	0,22805	0,38844	0,52835
		$\sqrt{EQM}$	0,26360	0,41847	0,55445
	Lasso+MV	Viés	0,00456	0,03128	0,07104
		$\sqrt{EQM}$	0,11424	0,15653	0,19587
	Step+Lasso	Viés	0,08019	0,05142	0,05612
		$\sqrt{EQM}$	0,14464	0,15960	0,18969
	Step+MV	Viés	0,00419	0,02885	0,06319
		$\sqrt{EQM}$	0,10997	0,15118	0,18814

Tabela E.10: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado diferente de zero para os cenários com correlação 0,8 e 40% de covariáveis importantes.*

$\rho = 0{,}8$ 40% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,55520	0,62984	1,62344
		$\sqrt{EQM}$	0,91334	2,16079	2,00363
	Lasso+MV	Viés	0,45796	8,10e+12	1,02e+12
		$\sqrt{EQM}$	0,97196	8,85e+13	2,28e+13
	Step+Lasso	Viés	0,29397	4,07020	1,82200
		$\sqrt{EQM}$	0,82991	10,93662	9,25054
	Step+MV	Viés	0,41612	4,63e+12	1829,461
		$\sqrt{EQM}$	0,94441	8,06e+13	14165,84
$n = 350$	Lasso	Viés	0,43591	0,63237	0,83822
		$\sqrt{EQM}$	0,62782	0,97645	1,62594
	Lasso+MV	Viés	0,13922	9,24233	9,72e+12
		$\sqrt{EQM}$	0,45451	135,4496	7,20e+13
	Step+Lasso	Viés	0,16585	0,31540	2,92674
		$\sqrt{EQM}$	0,46494	1,24880	6,52336
	Step+MV	Viés	0,11859	563,6146	10528,09
		$\sqrt{EQM}$	0,43628	7386,627	85283,52
$n = 700$	Lasso	Viés	0,35835	0,59703	0,85561
		$\sqrt{EQM}$	0,49042	0,77393	1,01239
	Lasso+MV	Viés	0,07200	0,39151	6,57e+12
		$\sqrt{EQM}$	0,31173	0,69658	6,70e+13
	Step+Lasso	Viés	0,13782	0,04481	0,60793
		$\sqrt{EQM}$	0,33655	0,51570	1,69314
	Step+MV	Viés	0,06412	0,32478	5514,882
		$\sqrt{EQM}$	0,30312	0,64109	92023,02
$n = 3500$	Lasso	Viés	0,21290	0,41285	0,60801
		$\sqrt{EQM}$	0,26551	0,46515	0,65508
	Lasso+MV	Viés	0,02072	0,05828	0,17149
		$\sqrt{EQM}$	0,13433	0,20819	0,30935
	Step+Lasso	Viés	0,07429	0,08651	0,09001
		$\sqrt{EQM}$	0,15459	0,21859	0,26783
	Step+MV	Viés	0,01919	0,04842	0,14468
		$\sqrt{EQM}$	0,13160	0,20277	0,28956

Tabela E.11: *Resultados da média do módulo do viés e a média da raiz do EQM para as covariáveis com parâmetro associado diferente de zero para os cenários com correlação 0,8 e 60% de covariáveis importantes.*

$\rho = 0{,}8$ 60% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Viés	0,43284	0,80080	1,81659
		$\sqrt{EQM}$	1,02535	2,24991	2,14572
	Lasso+MV	Viés	0,63162	1,20e+12	12,12804
		$\sqrt{EQM}$	1,34479	3,49e+13	76,07272
	Step+Lasso	Viés	0,23914	3,70752	1,14717
		$\sqrt{EQM}$	0,95982	11,07878	6,36349
	Step+MV	Viés	0,55944	3,33e+10	3,56e+9
		$\sqrt{EQM}$	1,29707	7,45e+11	3,43e+10
$n = 350$	Lasso	Viés	0,36774	0,16343	0,65468
		$\sqrt{EQM}$	0,63609	1,94366	1,70909
	Lasso+MV	Viés	0,14595	1,21e+13	1,23e+12
		$\sqrt{EQM}$	0,53202	8,71e+13	2,75e+13
	Step+Lasso	Viés	0,21252	0,82185	2,45330
		$\sqrt{EQM}$	0,54125	2,65697	5,80077
	Step+MV	Viés	0,12394	2433,16	1,02e+12
		$\sqrt{EQM}$	0,51503	13591,27	2,29e+13
$n = 700$	Lasso	Viés	0,29987	0,44278	0,42339
		$\sqrt{EQM}$	0,46892	0,78401	1,55975
	Lasso+MV	Viés	0,09069	0,65834	4,21e+13
		$\sqrt{EQM}$	0,34505	1,02742	1,43e+14
	Step+Lasso	Viés	0,12831	0,06741	1,00132
		$\sqrt{EQM}$	0,36975	0,68526	2,56936
	Step+MV	Viés	0,08262	0,56950	22908,7
		$\sqrt{EQM}$	0,33924	0,94333	265149,9
$n = 3500$	Lasso	Viés	0,17567	0,37291	0,51067
		$\sqrt{EQM}$	0,24680	0,46077	0,61445
	Lasso+MV	Viés	0,01934	0,09717	0,26042
		$\sqrt{EQM}$	0,13997	0,25624	0,42121
	Step+Lasso	Viés	0,08321	0,11830	0,11065
		$\sqrt{EQM}$	0,16869	0,26971	0,33501
	Step+MV	Viés	0,01751	0,08478	0,22484
		$\sqrt{EQM}$	0,13826	0,24914	0,39334

Apêndice F

Resultados completos dos estudos de simulação para a proporção de vezes em que o número correto de covariáveis foi identificado e o modelo correto foi selecionado

Tabela F.1: *Resultados da proporção de vezes em que o número correto de covariáveis foi identificado e o modelo correto foi selecionado para os cenários com correlação 0 e 40% de covariáveis importantes.*

$\rho = 0$ 40% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Variáveis corretas	0,294	0	0
		Modelo correto	0,294	0	0
	Lasso+MV	Variáveis corretas	0,294	0	0
		Modelo correto	0,294	0	0
	Step+Lasso	Variáveis corretas	0,412	0,028	0,070
		Modelo correto	0,412	0,006	0
	Step+MV	Variáveis corretas	0,290	0,020	0,070
		Modelo correto	0,290	0,004	0
$n = 350$	Lasso	Variáveis corretas	0,396	0,006	0,002
		Modelo correto	0,396	0,006	0,002
	Lasso+MV	Variáveis corretas	0,396	0,006	0,002
		Modelo correto	0,396	0,006	0,002
	Step+Lasso	Variáveis corretas	0,410	0,016	0,012
		Modelo correto	0,410	0,016	0,008
	Step+MV	Variáveis corretas	0,346	0,016	0,006
		Modelo correto	0,346	0,016	0,006
$n = 700$	Lasso	Variáveis corretas	0,404	0,114	0
		Modelo correto	0,404	0,114	0
	Lasso+MV	Variáveis corretas	0,404	0,114	0
		Modelo correto	0,404	0,114	0
	Step+Lasso	Variáveis corretas	0,398	0,024	0,002
		Modelo correto	0,398	0,024	0,002
	Step+MV	Variáveis corretas	0,338	0,022	0,002
		Modelo correto	0,338	0,022	0,002
$n = 3500$	Lasso	Variáveis corretas	0,534	0,210	0,024
		Modelo correto	0,534	0,210	0,024
	Lasso+MV	Variáveis corretas	0,534	0,210	0,024
		Modelo correto	0,534	0,210	0,024
	Step+Lasso	Variáveis corretas	0,424	0,044	0
		Modelo correto	0,424	0,044	0
	Step+MV	Variáveis corretas	0,360	0,040	0
		Modelo correto	0,360	0,040	0

Tabela F.2: *Resultados da proporção de vezes em que o número correto de covariáveis foi identificado e o modelo correto foi selecionado para os cenários com correlação 0 e 60% de covariáveis importantes.*

$\rho = 0$ 60% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Variáveis corretas	0,174	0,002	0,014
		Modelo correto	0,174	0	0
	Lasso+MV	Variáveis corretas	0,174	0,002	0,014
		Modelo correto	0,174	0	0
	Step+Lasso	Variáveis corretas	0,440	0,218	0,014
		Modelo correto	0,440	0	0
	Step+MV	Variáveis corretas	0,384	0,214	0,014
		Modelo correto	0,384	0	0
$n = 350$	Lasso	Variáveis corretas	0,218	0	0
		Modelo correto	0,218	0	0
	Lasso+MV	Variáveis corretas	0,218	0	0
		Modelo correto	0,218	0	0
	Step+Lasso	Variáveis corretas	0,430	0,022	0,212
		Modelo correto	0,430	0,022	0,014
	Step+MV	Variáveis corretas	0,418	0,022	0,210
		Modelo correto	0,418	0,022	0,014
$n = 700$	Lasso	Variáveis corretas	0,272	0,022	0
		Modelo correto	0,272	0,022	0
	Lasso+MV	Variáveis corretas	0,272	0,022	0
		Modelo correto	0,272	0,022	0
	Step+Lasso	Variáveis corretas	0,504	0,068	0,014
		Modelo correto	0,504	0,068	0,014
	Step+MV	Variáveis corretas	0,482	0,068	0,012
		Modelo correto	0,482	0,068	0,012
$n = 3500$	Lasso	Variáveis corretas	0,432	0,068	0
		Modelo correto	0,432	0,068	0
	Lasso+MV	Variáveis corretas	0,432	0,068	0
		Modelo correto	0,432	0,068	
	Step+Lasso	Variáveis corretas	0,492	0,118	0,028
		Modelo correto	0,492	0,118	0,028
	Step+MV	Variáveis corretas	0,468	0,116	0,028
		Modelo correto	0,468	0,116	0,028

Tabela F.3: *Resultados da proporção de vezes em que o número correto de covariáveis foi identificado e o modelo correto foi selecionado para os cenários com correlação 0,2 e 20% de covariáveis importantes.*

$\rho = 0{,}2$ 20% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Variáveis Corretas	0,426	0,030	0,004
		Modelo Correto	0,426	0,030	0,004
	Lasso+MV	Variáveis Corretas	0,426	0,030	0,004
		Modelo Correto	0,426	0,030	0,004
	Step+Lasso	Variáveis Corretas	0,338	0,040	0,004
		Modelo Correto	0,338	0,034	0
	Step+MV	Variáveis Corretas	0,214	0	0
		Modelo Correto	0,214	0	0
$n = 350$	Lasso	Variáveis Corretas	0,496	0,172	0
		Modelo Correto	0,496	0,172	0
	Lasso+MV	Variáveis Corretas	0,496	0,172	0
		Modelo Correto	0,496	0,172	0
	Step+Lasso	Variáveis Corretas	0,326	0,030	0
		Modelo Correto	0,326	0,030	0
	Step+MV	Variáveis Corretas	0,254	0,014	0
		Modelo Correto	0,254	0,014	0
$n = 700$	Lasso	Variáveis Corretas	0,558	0,142	0,024
		Modelo Correto	0,558	0,142	0,024
	Lasso+MV	Variáveis Corretas	0,558	0,142	0,024
		Modelo Correto	0,558	0,142	0,024
	Step+Lasso	Variáveis Corretas	0,392	0,022	0
		Modelo Correto	0,392	0,022	0
	Step+MV	Variáveis Corretas	0,268	0,016	0
		Modelo Correto	0,268	0,016	0
$n = 3500$	Lasso	Variáveis Corretas	0,600	0,240	0,076
		Modelo Correto	0,600	0,240	0,076
	Lasso+MV	Variáveis Corretas	0,600	0,240	0,076
		Modelo Correto	0,600	0,240	0,076
	Step+Lasso	Variáveis Corretas	0,352	0,024	0
		Modelo Correto	0,352	0,024	0
	Step+MV	Variáveis Corretas	0,270	0,022	0
		Modelo Correto	0,270	0,022	0

Tabela F.4: *Resultados da proporção de vezes em que o número correto de covariáveis foi identificado e o modelo correto foi selecionado para os cenários com correlação 0,2 e 40% de covariáveis importantes.*

$\rho = 0{,}2$ 40% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Variáveis Corretas	0,364	0	0
		Modelo Correto	0,364	0	0
	Lasso+MV	Variáveis Corretas	0,364	0	0
		Modelo Correto	0,364	0	0
	Step+Lasso	Variáveis Corretas	0,378	0,056	0,074
		Modelo Correto	0,378	0,022	0
	Step+MV	Variáveis Corretas	0,312	0,038	0,076
		Modelo Correto	0,312	0,018	0
$n = 350$	Lasso	Variáveis Corretas	0,274	0,008	0
		Modelo Correto	0,274	0,008	0
	Lasso+MV	Variáveis Corretas	0,274	0,008	0
		Modelo Correto	0,274	0,008	0
	Step+Lasso	Variáveis Corretas	0,350	0,016	0,004
		Modelo Correto	0,350	0,016	0,004
	Step+MV	Variáveis Corretas	0,322	0,012	0,004
		Modelo Correto	0,322	0,012	0,004
$n = 700$	Lasso	Variáveis Corretas	0,352	0,004	0
		Modelo Correto	0,352	0,004	0
	Lasso+MV	Variáveis Corretas	0,352	0,004	0
		Modelo Correto	0,352	0,004	0
	Step+Lasso	Variáveis Corretas	0,378	0,030	0
		Modelo Correto	0,378	0,030	0
	Step+MV	Variáveis Corretas	0,348	0,030	0
		Modelo Correto	0,348	0,030	0
$n = 3500$	Lasso	Variáveis Corretas	0,494	0,054	0
		Modelo Correto	0,494	0,054	0
	Lasso+MV	Variáveis Corretas	0,494	0,054	0
		Modelo Correto	0,494	0,054	0
	Step+Lasso	Variáveis Corretas	0,406	0,044	0,004
		Modelo Correto	0,406	0,044	0,004
	Step+MV	Variáveis Corretas	0,378	0,042	0,004
		Modelo Correto	0,378	0,042	0,004

Tabela F.5: *Resultados da proporção de vezes em que o número correto de covariáveis foi identificado e o modelo correto foi selecionado para os cenários com correlação 0,2 e 60% de covariáveis importantes.*

$\rho = 0{,}2$ 60% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Variáveis Corretas	0,186	0	0,008
		Modelo Correto	0,186	0	0
	Lasso+MV	Variáveis Corretas	0,186	0	0,008
		Modelo Correto	0,186	0	0
	Step+Lasso	Variáveis Corretas	0,452	0,264	0,002
		Modelo Correto	0,452	0	0
	Step+MV	Variáveis Corretas	0,430	0,264	0
		Modelo Correto	0,430	0	0
$n = 350$	Lasso	Variáveis Corretas	0,200	0,002	0
		Modelo Correto	0,200	0,002	0
	Lasso+MV	Variáveis Corretas	0,200	0,002	0
		Modelo Correto	0,200	0,002	0
	Step+Lasso	Variáveis Corretas	0,494	0,044	0,194
		Modelo Correto	0,494	0,044	0
	Step+MV	Variáveis Corretas	0,492	0,044	0,198
		Modelo Correto	0,492	0,044	0
$n = 700$	Lasso	Variáveis Corretas	0,372	0	0
		Modelo Correto	0,372	0	0
	Lasso+MV	Variáveis Corretas	0,372	0	0
		Modelo Correto	0,372	0	0
	Step+Lasso	Variáveis Corretas	0,530	0,074	0,016
		Modelo Correto	0,530	0,074	0,014
	Step+MV	Variáveis Corretas	0,498	0,074	0,016
		Modelo Correto	0,498	0,074	0,014
$n = 3500$	Lasso	Variáveis Corretas	0,364	0	0
		Modelo Correto	0,364	0	0
	Lasso+MV	Variáveis Corretas	0,364	0	0
		Modelo Correto	0,364	0	0
	Step+Lasso	Variáveis Corretas	0,502	0,136	0,022
		Modelo Correto	0,502	0,136	0,022
	Step+MV	Variáveis Corretas	0,498	0,136	0,022
		Modelo Correto	0,498	0,136	0,022

Tabela F.6: *Resultados da proporção de vezes em que o número correto de covariáveis foi identificado e o modelo correto foi selecionado para os cenários com correlação 0,5 e 20% de covariáveis importantes.*

$\rho = 0{,}5$ 20% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Variáveis corretas	0,276	0,006	0,006
		Modelo correto	0,276	0,006	0
	Lasso+MV	Variáveis corretas	0,276	0,006	0,006
		Modelo correto	0,276	0,006	0
	Step+Lasso	Variáveis corretas	0,236	0,008	0,004
		Modelo correto	0,236	0,008	0
	Step+MV	Variáveis corretas	0,192	0,002	0
		Modelo correto	0,192	0,002	0
$n = 350$	Lasso	Variáveis corretas	0,404	0,042	0,002
		Modelo correto	0,404	0,042	0,002
	Lasso+MV	Variáveis corretas	0,404	0,042	0,002
		Modelo correto	0,404	0,042	0,002
	Step+Lasso	Variáveis corretas	0,326	0,018	0
		Modelo correto	0,326	0,018	0
	Step+MV	Variáveis corretas	0,230	0,014	0
		Modelo correto	0,230	0,014	0
$n = 700$	Lasso	Variáveis corretas	0,410	0,070	0,002
		Modelo correto	0,410	0,070	0,002
	Lasso+MV	Variáveis corretas	0,410	0,070	0,002
		Modelo correto	0,410	0,070	0,002
	Step+Lasso	Variáveis corretas	0,272	0,018	0,002
		Modelo correto	0,272	0,018	0,002
	Step+MV	Variáveis corretas	0,218	0,016	0
		Modelo correto	0,218	0,016	0
$n = 3500$	Lasso	Variáveis corretas	0,462	0,110	0,010
		Modelo correto	0,462	0,110	0,010
	Lasso+MV	Variáveis corretas	0,462	0,110	0,010
		Modelo correto	0,462	0,110	0,010
	Step+Lasso	Variáveis corretas	0,266	0,012	0,002
		Modelo correto	0,266	0,012	0,002
	Step+MV	Variáveis corretas	0,218	0,010	0,002
		Modelo correto	0,218	0,010	0,002

Tabela F.7: *Resultados da proporção de vezes em que o número correto de covariáveis foi identificado e o modelo correto foi selecionado para os cenários com correlação 0,5 e 40% de covariáveis importantes.*

$\rho = 0{,}5$ 40% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Variáveis corretas	0,144	0,002	0,004
		Modelo correto	0,144	0	0
	Lasso+MV	Variáveis corretas	0,144	0,002	0,004
		Modelo correto	0,144	0	0
	Step+Lasso	Variáveis corretas	0,328	0,012	0,172
		Modelo correto	0,328	0,004	0
	Step+MV	Variáveis corretas	0,316	0,008	0,16
		Modelo correto	0,316	0,002	0
$n = 350$	Lasso	Variáveis corretas	0,124	0,008	0
		Modelo correto	0,124	0,008	0
	Lasso+MV	Variáveis corretas	0,124	0,008	0
		Modelo correto	0,124	0,008	0
	Step+Lasso	Variáveis corretas	0,316	0,022	0,006
		Modelo correto	0,316	0,022	0,002
	Step+MV	Variáveis corretas	0,310	0,022	0,004
		Modelo correto	0,310	0,022	0,002
$n = 700$	Lasso	Variáveis corretas	0,286	0,006	0
		Modelo correto	0,286	0,006	0
	Lasso+MV	Variáveis corretas	0,286	0,006	0
		Modelo correto	0,286	0,006	0
	Step+Lasso	Variáveis corretas	0,358	0,046	0
		Modelo correto	0,358	0,046	0
	Step+MV	Variáveis corretas	0,344	0,046	0
		Modelo correto	0,344	0,046	0
$n = 3500$	Lasso	Variáveis corretas	0,290	0,014	0
		Modelo correto	0,290	0,014	0
	Lasso+MV	Variáveis corretas	0,290	0,014	0
		Modelo correto	0,290	0,014	0
	Step+Lasso	Variáveis corretas	0,318	0,050	0,008
		Modelo correto	0,318	0,050	0,008
	Step+MV	Variáveis corretas	0,314	0,050	0,008
		Modelo correto	0,314	0,050	0,008

Tabela F.8: *Resultados da proporção de vezes em que o número correto de covariáveis foi identificado e o modelo correto foi selecionado para os cenários com correlação 0,5 e 60% de covariáveis importantes.*

$\rho = 0{,}5$ 60% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 140$	Lasso	Variáveis corretas	0,072	0,004	0,072
		Modelo correto	0,072	0	0
	Lasso+MV	Variáveis corretas	0,072	0,004	0,072
		Modelo correto	0,072	0	0
	Step+Lasso	Variáveis corretas	0,408	0,198	0
		Modelo correto	0,408	0,002	0
	Step+MV	Variáveis corretas	0,408	0,190	0
		Modelo correto	0,408	0,002	0
$n = 350$	Lasso	Variáveis corretas	0,136	0,002	0
		Modelo correto	0,136	0,002	0
	Lasso+MV	Variáveis corretas	0,136	0,002	0
		Modelo correto	0,136	0,002	0
	Step+Lasso	Variáveis corretas	0,492	0,024	0,148
		Modelo correto	0,492	0,022	0
	Step+MV	Variáveis corretas	0,490	0,024	0,148
		Modelo correto	0,490	0,022	0
$n = 700$	Lasso	Variáveis corretas	0,212	0	0
		Modelo correto	0,212	0	0
	Lasso+MV	Variáveis corretas	0,212	0	0
		Modelo correto	0,212	0	0
	Step+Lasso	Variáveis corretas	0,498	0,080	0,008
		Modelo correto	0,498	0,080	0,006
	Step+MV	Variáveis corretas	0,490	0,080	0,008
		Modelo correto	0,490	0,080	0,006
$n = 3500$	Lasso	Variáveis corretas	0,202	0	0
		Modelo correto	0,202	0	0
	Lasso+MV	Variáveis corretas	0,202	0	0
		Modelo correto	0,202	0	0
	Step+Lasso	Variáveis corretas	0,504	0,146	0,014
		Modelo correto	0,504	0,146	0,014
	Step+MV	Variáveis corretas	0,502	0,146	0,014
		Modelo correto	0,502	0,146	0,014

Tabela F.9: *Resultados da proporção de vezes em que o número correto de covariáveis foi identificado e o modelo correto foi selecionado para os cenários com correlação 0,8 e 20% de covariáveis importantes.*

$\rho = 0{,}8$ 20% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	Variáveis corretas	0,154	0,008	0,006
		Modelo correto	0,154	0,006	0
	Lasso+MV	Variáveis corretas	0,154	0,008	0,006
		Modelo correto	0,154	0,006	0
	Step+Lasso	Variáveis corretas	0,226	0,006	0,004
		Modelo correto	0,226	0,004	0
	Step+MV	Variáveis corretas	0,204	0,004	0
		Modelo correto	0,204	0,004	0
$n = 500$	Lasso	Variáveis corretas	0,254	0,006	0
		Modelo correto	0,254	0,006	0
	Lasso+MV	Variáveis corretas	0,254	0,006	0
		Modelo correto	0,254	0,006	0
	Step+Lasso	Variáveis corretas	0,274	0,006	0
		Modelo correto	0,274	0,006	0
	Step+MV	Variáveis corretas	0,260	0,006	0
		Modelo correto	0,260	0,006	0
$n = 1000$	Lasso	Variáveis corretas	0,224	0,014	0
		Modelo correto	0,224	0,014	0
	Lasso+MV	Variáveis corretas	0,224	0,014	0
		Modelo correto	0,224	0,014	0
	Step+Lasso	Variáveis corretas	0,256	0,014	0
		Modelo correto	0,256	0,014	0
	Step+MV	Variáveis corretas	0,244	0,014	0
		Modelo correto	0,244	0,014	0
$n = 5000$	Lasso	Variáveis corretas	0,340	0,02	0
		Modelo correto	0,340	0,02	0
	Lasso+MV	Variáveis corretas	0,340	0,02	0
		Modelo correto	0,340	0,02	0
	Step+Lasso	Variáveis corretas	0,302	0,016	0,002
		Modelo correto	0,302	0,016	0,002
	Step+MV	Variáveis corretas	0,282	0,016	0,002
		Modelo correto	0,282	0,016	0,002

Tabela F.10: *Resultados da proporção de vezes em que o número correto de covariáveis foi identificado e o modelo correto foi selecionado para os cenários com correlação 0,8 e 40% de covariáveis importantes.*

$\rho = 0{,}8$ 40% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	Variáveis corretas	0,160	0,008	0,032
		Modelo correto	0,156	0	0
	Lasso+MV	Variáveis corretas	0,160	0,008	0,032
		Modelo correto	0,156	0	0
	Step+Lasso	Variáveis corretas	0,324	0,076	0,020
		Modelo correto	0,312	0	0
	Step+MV	Variáveis corretas	0,314	0,058	0,028
		Modelo correto	0,302	0	0
$n = 500$	Lasso	Variáveis corretas	0,076	0	0,002
		Modelo correto	0,076	0	0
	Lasso+MV	Variáveis corretas	0,076	0	0,002
		Modelo correto	0,076	0	0
	Step+Lasso	Variáveis corretas	0,332	0,018	0,028
		Modelo correto	0,332	0,016	0
	Step+MV	Variáveis corretas	0,332	0,018	0,016
		Modelo correto	0,332	0,016	0
$n = 1000$	Lasso	Variáveis corretas	0,104	0,002	0
		Modelo correto	0,104	0,002	0
	Lasso+MV	Variáveis corretas	0,104	0,002	0
		Modelo correto	0,104	0,002	0
	Step+Lasso	Variáveis corretas	0,302	0,032	0
		Modelo correto	0,302	0,032	0
	Step+MV	Variáveis corretas	0,300	0,032	0
		Modelo correto	0,300	0,032	0
$n = 5000$	Lasso	Variáveis corretas	0,154	0	0
		Modelo correto	0,154	0	0
	Lasso+MV	Variáveis corretas	0,154	0	0
		Modelo correto	0,154	0	0
	Step+Lasso	Variáveis corretas	0,348	0,030	0,004
		Modelo correto	0,348	0,030	0,004
	Step+MV	Variáveis corretas	0,348	0,030	0,004
		Modelo correto	0,348	0,030	0,004

Tabela F.11: *Resultados da proporção de vezes em que o número correto de covariáveis foi identificado e o modelo correto foi selecionado para os cenários com correlação 0,8 e 60% de covariáveis importantes.*

$\rho = 0{,}8$ 60% de covariáveis importantes			$p = 10$	$p = 30$	$p = 50$
$n = 200$	Lasso	Variáveis corretas	0,082	0,040	0,042
		Modelo correto	0,076	0	0
	Lasso+MV	Variáveis corretas	0,082	0,040	0,042
		Modelo correto	0,076	0	0
	Step+Lasso	Variáveis corretas	0,444	0,042	0
		Modelo correto	0,402	0	0
	Step+MV	Variáveis corretas	0,442	0,054	0
		Modelo correto	0,400	0	0
$n = 500$	Lasso	Variáveis corretas	0,070	0,002	0,008
		Modelo correto	0,070	0	0
	Lasso+MV	Variáveis corretas	0,070	0,002	0,008
		Modelo correto	0,070	0	0
	Step+Lasso	Variáveis corretas	0,482	0,040	0,074
		Modelo correto	0,482	0,020	0
	Step+MV	Variáveis corretas	0,482	0,038	0,072
		Modelo correto	0,482	0,020	0
$n = 1000$	Lasso	Variáveis corretas	0,086	0	0
		Modelo correto	0,086	0	0
	Lasso+MV	Variáveis corretas	0,086	0	0
		Modelo correto	0,086	0	0
	Step+Lasso	Variáveis corretas	0,488	0,082	0,014
		Modelo correto	0,488	0,082	0,004
	Step+MV	Variáveis corretas	0,488	0,082	0,014
		Modelo correto	0,488	0,082	0,004
$n = 5000$	Lasso	Variáveis corretas	0,078	0	0
		Modelo correto	0,078	0	0
	Lasso+MV	Variáveis corretas	0,078	0	0
		Modelo correto	0,078	0	0
	Step+Lasso	Variáveis corretas	0,518	0,116	0,020
		Modelo correto	0,518	0,116	0,020
	Step+MV	Variáveis corretas	0,518	0,116	0,020
		Modelo correto	0,518	0,116	0,020

Apêndice G

Resultados da seleção de covariáveis feita por cada método nas aplicações sem a presença de alta dimensionalidade

Tabela G.1: *Frequência de seleção das covariáveis por cada método na aplicação 1*

	Lasso	Step	Step+Lasso
Variável 1	100	100	100
Variável 2	100	100	100
Variável 3	100	100	100
Variável 4	100	100	100
Variável 5	100	100	100
Variável 6	100	100	100
Variável 7	100	99	99
Variável 8	100	100	100
Variável 9	95	34	34
Variável 10	99	60	60
Variável 11	90	12	12
Variável 12	100	100	100
Variável 13	68	52	51
Variável 14	88	58	58
Variável 15	53	14	14
Variável 16	66	21	21
Variável 17	75	22	22
Variável 18	100	100	100
Variável 19	100	100	100
Variável 20	100	62	62
Variável 21	100	81	81
Variável 22	100	81	81
Variável 23	100	57	57

Tabela G.2: *Frequência de seleção das covariáveis por cada método na aplicação 2*

	Lasso	Step	Step+Lasso
Variável 1	100	100	100
Variável 2	100	100	100
Variável 3	49	10	10
Variável 4	41	3	3
Variável 5	100	98	98
Variável 6	64	15	15
Variável 7	74	40	40
Variável 8	98	54	54
Variável 9	43	0	0
Variável 10	79	62	62
Variável 11	100	100	100
Variável 12	21	8	8
Variável 13	32	2	2
Variável 14	31	6	6
Variável 15	100	100	100

Tabela G.3: *Frequência de seleção das covariáveis por cada método na aplicação 3*

	Lasso	Step	Step+Lasso
Variável 1	83	37	37
Variável 2	100	100	100
Variável 3	40	6	6
Variável 4	77	13	13
Variável 5	33	7	4
Variável 6	100	92	92
Variável 7	97	90	90
Variável 8	97	74	74

Tabela G.4: *Frequência de seleção das covariáveis por cada método na aplicação 4*

	Lasso	Step	Step+Lasso
Variável 1	2	0	0
Variável 2	42	0	0
Variável 3	1	44	16
Variável 4	0	40	32
Variável 5	3	41	10
Variável 6	9	28	6
Variável 7	39	36	28
Variável 8	89	70	44
Variável 9	19	53	50
Variável 10	28	39	39
Variável 11	79	31	19
Variável 12	31	33	16
Variável 13	4	65	59
Variável 14	14	43	36
Variável 15	41	44	26
Variável 16	63	36	24
Variável 17	4	34	24
Variável 18	9	36	29
Variável 19	24	59	34
Variável 20	65	68	46
Variável 21	90	42	33
Variável 22	100	63	57
Variável 23	29	43	43
Variável 24	26	69	69
Variável 25	91	38	32
Variável 26	1	36	30
Variável 27	74	36	35
Variável 28	94	35	19
Variável 29	94	42	40
Variável 30	13	44	41

Tabela G.5: *Frequência de seleção das covariáveis por cada método na aplicação 5*

	Lasso	Step	Step+Lasso
Variável 1	100	100	100
Variável 2	0	0	0
Variável 3	100	70	71
Variável 4	40	40	58
Variável 5	100	88	88
Variável 6	79	84	92
Variável 7	97	42	42
Variável 8	100	92	92
Variável 9	48	39	46
Variável 10	83	49	5
Variável 11	19	61	3
Variável 12	20	29	84
Variável 13	2	20	50
Variável 14	35	39	45
Variável 15	24	64	53
Variável 16	27	57	81
Variável 17	7	29	45
Variável 18	68	59	66
Variável 19	7	53	84
Variável 20	16	16	29
Variável 21	20	39	50
Variável 22	95	81	86
Variável 23	40	46	61
Variável 24	35	51	61
Variável 25	73	74	81
Variável 26	23	26	37
Variável 27	94	96	100
Variável 28	14	33	49
Variável 29	41	68	82
Variável 30	69	87	9
Variável 31	68	64	71
Variável 32	17	30	52
Variável 33	7	29	41
Variável 34	80	82	91

Tabela G.6: *Frequência de seleção das covariáveis por cada método na aplicação 6*

	Lasso	Step	Step+Lasso
Variável 1	45	33	31
Variável 2	50	14	13
Variável 3	39	9	9
Variável 4	34	63	56
Variável 5	31	21	18
Variável 6	33	55	46
Variável 7	25	57	53
Variável 8	27	52	44
Variável 9	8	45	39
Variável 10	13	32	32
Variável 11	14	29	24
Variável 12	11	31	24
Variável 13	44	35	33
Variável 14	28	43	35
Variável 15	21	10	7
Variável 16	24	13	13
Variável 17	43	45	40
Variável 18	44	31	28
Variável 19	90	25	25
Variável 20	61	79	75
Variável 21	53	27	27
Variável 22	36	76	76

Tabela G.7: *Frequência de seleção das covariáveis por cada método na aplicação 8*

	Lasso	Step	Step+Lasso
Variável 1	99	100	100
Variável 2	59	43	43
Variável 3	31	13	13
Variável 4	30	9	9
Variável 5	100	100	100
Variável 6	8	1	1

Referências Bibliográficas

Agresti, A. (2003), *Categorical data analysis*, Vol. 482, John Wiley & Sons. 22

Ahmed, S. E., Hossain, S. e Doksum, K. A. (2012), 'Lasso and shrinkage estimation in weibull censored regression models', *Journal of Statistical Planning and Inference* **142**(6), 1273–1284. 2

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **19**(6). 25

Akaike, H. (1978), 'A bayesian analysis of the minimum aic procedure', *Annals of the Institute of Statistical Mathematics A* **30**, 9–14. 25

Alcântara Junior, G. (2020), 'Construção de modelos de regressão logística no r', https://github.com/GilAlcantara/logistic_regression. 28

Annette, J. D. (2001), *An Introduction to Generalized Linear Models, Second Edition*, 2 edn, Chapman Hall. 19, 21

Anton, H., Bivens, I. e Davis, S. (2014), *Cálculo-Volume I*, 10 edn, Bookman Editora. 13

Belloni, A., Chernozhukov, V. e Hansen, C. (2011), 'Lasso methods for gaussian instrumental variables models', *IT Department of Economics Working Paper* pp. 11–14. 2

Butcher, B. e Smith, B. J. (2020), 'Feature engineering and selection: A practical approach for predictive models', *The American Statistician* **74**(3), 308–309. 27

Casella, G. e Berger, R. L. (2002), *Statistical inference*, Vol. 2, Duxbury Pacific Grove, CA. 25

Das, K. e Sobel, M. (2015), 'Dirichlet lasso: A bayesian approach to variable selection', *Statistical Modelling* **15**(3), 215–232. 2

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K. H., Lee, S. e Froelicher, V. (1989), 'International application of a new probability algorithm for the diagnosis of coronary artery disease', *The American journal of cardiology* **64**(5), 304–310. 40

Fox, J. (2015), *Applied regression analysis and generalized linear models*, Sage Publications. 19

Friedman, J., Hastie, T. e Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', *Journal of Statistical Software* **33**(1), 1–22. 16, 18

Gorst-Rasmussen, A. e Scheike, T. H. (2012), 'Coordinate descent methods for the penalized semiparametric additive hazards model', *Journal of Statistical Software* **47**(9), 1–17. 39

Green, P. J. (1984), 'Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives', *Journal of the Royal Statistical Society: Series B (Methodological)* **46**(2), 149–170. 21

Hastie, T. e Qian, J. (2014), 'Glmnet vignette'.
URL: *https://web.stanford.edu/ hastie/glmnet/glmnet$_a$lpha.html* 30

Hastie, T., Tibshirani, R. e Friedman, J. (2009), *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media. 5

Hastie, T., Tibshirani, R., Tibshirani, R. et al. (2020), 'Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons', *Statistical Science* **35**(4), 579–592. 1, 27

Hastie, T., Tibshirani, R. e Wainwright, M. (2015), *Statistical Learning with Sparsity: The Lasso and Generalizations*, Chapman Hall/CRC. 1, 3, 5, 6, 7, 10, 12, 15, 16, 23

Hoerl, A. E., Kannard, R. W. e Baldwin, K. F. (1975), 'Ridge regression: some simulations', *Communications in Statistics-Theory and Methods* **4**(2), 105–123. 6

Ijaz, M., Asghar, Z. e Gul, A. (2019), 'Ensemble of penalized logistic models for classification of high-dimensional data', *Communications in Statistics-Simulation and Computation* pp. 7,8. 28

Izbicki, R. e dos Santos (2019), 'Lista do curso de machine learning (programa interinstitucional de pós-graduação em estatística)'. Accessed: 2020-06-20.
URL: *https://d1b10bmlvqabco.cloudfront.net/attach/jyj4xf9mkro44u/iiloo0nfn0wia/jzvh87klv7k9/hw1.pdf* 8

Izbicki, R. e dos Santos, T. M. (2020), *Aprendizado de máquina: uma abordagem estatística*. 7

James, G., Witten, D., Hastie, T. e Tibshirani, R. (2013), *An introduction to statistical learning*, Vol. 112, Springer. 1

Kaggle (2010), 'Kaggle, your home for data science'.
URL: *https://www.kaggle.com/* 39

Kassambara, A. (2019), 'Data bank for statistical analysis and visualization'.
URL: *https://cran.r-project.org/web/packages/datarium/datarium.pdf* 11

Kim, S. M., Kim, Y., Jeong, K., Jeong, H. e Kim, J. (2018), 'Logistic lasso regression for the diagnosis of breast cancer using clinical demographic data and the bi-rads lexicon for ultrasonography', *Ultrasonography* **37**(1), 36. 2

Kumar, S., Attri, S. e Singh, K. (2019), 'Comparison of lasso and stepwise regression technique for wheat yield prediction', *Journal of Agrometeorology* **21**(2), 188–192. 1

Lee, J. D., Sun, Y. e Saunders, M. A. (2014), 'Proximal newton-type methods for minimizing composite functions', *SIAM Journal on Optimization* **24**(3), 1420–1443. 24

Little, M. A., McSharry, P. E., Roberts, S. J., Costello, D. A. e Moroz, I. M. (2007), 'Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection', *Biomedical engineering online* **6**(1), 23. 40

Machado, A. A. (2018), 'Seleçao de variáveis em modelos de regressao logıstica'. Monografia (Bacharel em Estatística), UFF (Universidade Federal Fluminense), Rio de Janeiro, Brasil. 2

McCulloch, C. E. e Neuhaus, J. M. (2014), 'Generalized linear mixed models', *Wiley StatsRef: Statistics Reference Online* . 20

Meier, L., Van De Geer, S. e Bühlmann, P. (2008), 'The group lasso for logistic regression', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(1), 53–71. 2

Meinshausen, N. (2007), 'Relaxed lasso', *Computational Statistics & Data Analysis* **52**(1), 374–393. 26

Myerson, J., Green, L. e Warusawitharana, M. (2001), 'Area under the curve as a measure of discounting', *Journal of the experimental analysis of behavior* **76**(2), 235–243. 17

Nelder, J. A. e Wedderburn, R. W. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society: Series A (General)* **135**(3), 370–384. 16, 19, 20

Neter, J., Kutner, M. H., Nachtsheim, C. J. e Wasserman, W. (2008), 'Applied linear statistical models', *Irwin Chicago* . 25

Paula, G. A. (2013), 'Modelos de regressão com apoio computacional'.
URL: *https://www.ime.usp.br/ giapaula/texto*$_2$013.*pdf* 24

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/* 11, 16

Ramana, B. V., Babu, M. S. P., Venkateswarlu, N. et al. (2011), 'A critical study of selected classification algorithms for liver disease diagnosis', *International Journal of Database Management Systems* **3**(2), 101–114. 40

Schwarz, G. (1978), 'Estimating the dimension of a model', *The annals of statistics* **6**(2), 461–464. 25

Sigillito, V. G., Wing, S. P., Hutton, L. V. e Baker, K. B. (1989), 'Classification of radar returns from the ionosphere using neural networks', *Johns Hopkins APL Technical Digest* **10**(3), 262–266. 40

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S. et al. (2003), 'Repeated observation of breast tumor subtypes in independent gene expression data sets', *Proceedings of the national academy of sciences* **100**(14), 8418–8423. 40

Steyerberg, E. W., Eijkemans, M. J., Harrell Jr, F. E. e Habbema, J. D. F. (2000), 'Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets', *Statistics in medicine* **19**(8), 1059–1079. 2

Street, W. N., Wolberg, W. H. e Mangasarian, O. L. (1993), Nuclear feature extraction for breast tumor diagnosis, *in* 'Biomedical image processing and biomedical visualization', Vol. 1905, International Society for Optics and Photonics, pp. 861–870. 40

Thomas, L. C., Edelman, D. B. e Crook, J. N. (2002), *Credit scoring and its applications*, SIAM. 28

Thrun, S., Bala, J., Bloedorn, E., Bratko, I., Cestnik, B., Cheng, J., Jong, K. D., Dzeroski, S., Hamann, R., Kaufman, K., Keller, S., Kononenko, I., Kreuziger, J., Michalski, R., Mitchell, T., Pachowicz, P., Roger, B., Vafaie, H., de Velde, W. V., Wenzel, W., Wnek, J. e Zhang, J. (1991), The monkâs problems: A performance comparison of different learning algorithms, Technical Report CMU-CS-91-197, Carnegie Mellon University, Pittsburgh, PA. 40

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288. 1

Tibshirani, R. (1997), 'The lasso method for variable selection in the cox model', *Statistics in medicine* **16**(4), 385–395. 2

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. e Knight, K. (2005), 'Sparsity and smoothness via the fused lasso', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(1), 91–108. 2

UCI da Universidade da California (1987), 'Uci machine learning repository'.
URL: *https://archive.ics.uci.edu/ml/index.php* 39

Uh, H.-W., Mertens, B. J., van der Wijk, H. J., Putter, H., van Houwelingen, H. C. e Houwing-Duistermaat, J. J. (2007), Model selection based on logistic regression in a

highly correlated candidate gene region, *in* 'BMC proceedings', Vol. 1, Springer, p. S114. 2

Van Calster, B., van Smeden, M., De Cock, B. e Steyerberg, E. W. (2020), 'Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study', *Statistical methods in medical research* **29**(11), 3166–3178. 2

Venables, W. N. e Ripley, B. D. (2002), *Modern Applied Statistics with S*, fourth edn, Springer, New York. 25

Wang, D., Zhang, W. e Bakhai, A. (2004), 'Comparison of bayesian model averaging and stepwise methods for model selection in logistic regression', *Statistics in medicine* **23**(22), 3451–3467. 2

Yeh, I.-C. e Lien, C.-h. (2009), 'The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients', *Expert Systems with Applications* **36**(2), 2473–2480. 40

Zarchi, M. S., Bushehri, S. F. e Dehghanizadeh, M. (2018), 'Scadi: A standard dataset for self-care problems classification of children with physical and motor disability', *International journal of medical informatics* **114**, 81–87. 40

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *Journal of the American statistical association* **101**(476), 1418–1429. 2

Zou, H. e Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the royal statistical society: series B (statistical methodology)* **67**(2), 301–320. 17