**SELEÇÃO DE VARIÁVEIS
EM
REGRESSÃO L_1**

Rodrigo Andrade Tavares

DISSERTAÇÃO APRESENTADA AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA UNIVERSIDADE DE SÃO PAULO
PARA OBTENÇÃO DO GRAU DE
MESTRE EM ESTATÍSTICA

Área de Concentração: Estatística

Orientadora: Profa. Dra. Carmen Diva Saldiva de André

Durante a elaboração deste trabalho, o autor recebeu apoio financeiro do CNPq

-São Paulo, Agosto de 1998-

Este exemplar corresponde à redação final
da dissertação devidamente corrigida e
defendida por Rodrigo Andrade Tavares
e aprovada pela comissão julgadora.

São Paulo, 21 de Agosto de 1998.

Banca Examinadora:

- Profa. Dra. Carmen Diva Saldiva de André (Orientadora) - IME - USP
- Prof. Dr. Rinaldo Artes - IME - USP
- Prof. Dr. Subhash C. Narula - Virginia Commonwealth University

*“Assim que, se alguém está em Cristo, nova criatura é:
as coisas velhas já passaram; eis que tudo se fez novo”*

II aos Coríntios 5:17

Agradecimentos

Primeiramente devo agradecer Deus, pois sem Ele jamais teria chegado até aqui e nada disso valeria a pena. A cada instante pude sentir Sua presença fiel ao meu lado me dando força e me sustentando. Agradeço também por todas as pessoas que irei citar aqui, pois sei que foram, sem sombra de dúvida, instrumentos dEle na minha vida.

Aos meus pais por tudo que me deram: amor, amizade, educação, apoio, incentivo e compreensão. Muito obrigado e que Deus os abençoe.

À Profa. Carmen pela orientação segura e competente e, mais importante, pela amizade, dedicação, respeito e confiança no meu trabalho.

À Profa. Lisbeth pela amizade e por tudo que fez por mim. Aprendi muito com a sua postura e seriedade.

À Profa. Sílvia Elian e ao Prof. Subhash C. Narula pela valiosa ajuda e contribuição na elaboração deste trabalho.

Aos amigos de longa data Alexandre, Paulo, Daniel e Luciana pela amizade, companheirismo e apoio.

Um obrigado muito especial para a moçada do Cubo: Edson, Fernando, Gabriel, Gilberto, Jishu, Lílian, Lucas, Neto, Ribamar, Rafael, Ramon, Ratner, Raul, Ricardo e Ronaldo. Espero não ter esquecido ninguém. Sem a inestimável amizade e companhia de vocês, tudo teria sido muito mais difícil e chato. Com vocês foi tudo muito divertido e especial.

À turma que começou o Mestrado comigo: Alessandra, Carine, Esteban, Luís e Maria Paula. Começamos essa luta juntos e não foi fácil. Vou sempre me lembrar com muito carinho de vocês.

Aos demais amigos e colegas da Estatística. Foram anos e anos de agradável convivência que pudemos compartilhar.

Aos amigos da Computação, Licenciatura, Matemática e Matemática Aplicada que conheci nesta caminhada.

Aos professores e ao IME pela oportunidade de aprender e estudar numa das melhores instituições de ensino superior.

Resumo

O método mais utilizado no ajuste de modelos de regressão múltipla é o de mínimos quadrados, devido a suas propriedades estatísticas serem amplamente estudadas e facilidades computacionais. Contudo, este método é sensível a valores aberrantes, que são muito frequentes no caso da distribuição dos erros possuir caudas pesadas. O objetivo desta dissertação é apresentar o método de estimação L_1 , que é resistente a valores aberrantes na variável resposta. Será explorado, em particular, o problema de seleção de variáveis, sendo apresentados e desenvolvidos os critérios quando são analisadas todas as possíveis regressões, e procedimentos automáticos de seleção. Um estudo preliminar sobre os efeitos da multicolinearidade nas estimativas L_1 é também executado. São apresentados também, programas que tornam viável a utilização do método L_1 em problemas de regressão.

Abstract

A widely used method to fit a multiple regression model is the least squares method, due to its statistical properties and computational facilities. However, this approach can be affected by outliers, which are very frequent when the errors distribution has heavy tails. The aim of this dissertation is to present the L_1 method for regression models, exploring the variables selection problem. Criteria for variables selection are presented and developed when all possible regressions are considered and automatic selection procedures. A preliminary study about the multicollinearity effects over L_1 estimates is also performed. Programs which make possible the use of the L_1 method in regression problems are presented.

Índice

1. Introdução	01
2. Alguns Resultados em Regressão L_1	07
2.1. Algoritmos Computacionais.....	07
2.1.1. Regressão Linear Simples	07
2.1.2. Regressão Linear Múltipla	08
2.2. Programas Computacionais.....	09
2.3. Inferência Estatística	09
2.4. Robustez	12
3. Métodos de Seleção de Variáveis	14
3.1. O Problema de Seleção de Variáveis	14
3.2. Todas as Possíveis Regressões	19
3.2.1. Critério da Soma dos Erros Absolutos - Todos os Possíveis Modelos	20
3.2.2. Critério da Soma dos Erros Absolutos - Algoritmo de Narula e Wellington.....	21
3.2.3. Critério do Coeficiente de Determinação	27
3.2.4. Critério do Coeficiente de Determinação Ajustado	33
3.2.5. Critério da Soma dos Erros Absolutos Preditivos.....	36
3.3. Procedimentos Automáticos de Seleção de Variáveis	39
3.3.1. Seleção <i>Forward</i>	40
3.3.2. Eliminação <i>Backward</i>	41
3.3.3. <i>Stepwise</i>	42
3.4. Conclusões	43
4. Um Estudo Preliminar Sobre os Efeitos da Multicolinearidade.....	45
4.1. Introdução.....	45
4.2. Descrição do Estudo.....	46
4.3. Resultados Obtidos.....	48
4.4. Conclusões	50
5. Aplicação	51

6. Considerações Finais.....	57
Apêndice A - Programa de Estimação L_1	59
Apêndice B - Programas de Procedimentos Automáticos de Seleção de Variáveis.....	65
Apêndice C - Programa de Simulação	73
Bibliografia	77

Capítulo 1

Introdução

Considere o seguinte modelo de regressão linear múltipla:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (1.1)$$

onde $\mathbf{y}$ é um vetor $n \times 1$ de valores de uma variável resposta de interesse, $\mathbf{X}$ uma matriz $n \times p$ de valores das variáveis explicativas, que pode conter uma coluna de 1s referente ao termo do intercepto, $\boldsymbol{\beta} = (\beta_0 \dots \beta_{(p-1)})^t$ é um vetor $p \times 1$ dos parâmetros desconhecidos e $\boldsymbol{\varepsilon}$ é um vetor $n \times 1$ de erros aleatórios independentes não-observáveis.

O método de mínimos quadrados tem sido o mais utilizado na estimação dos parâmetros deste modelo devido à sua viabilidade computacional e propriedades estatísticas bastante estudadas dentro da classe dos modelos Gaussianos (Neter et al. (1996)). Contudo, sabe-se que a ocorrência de valores aberrantes na variável resposta, muito frequente quando a distribuição dos erros possui cauda pesada, tem grande influência nas estimativas obtidas por este método. Na verdade, um único valor aberrante pode ter um grande efeito nas estimativas dos parâmetros do modelo, conforme pode ser visto no exemplo a seguir, retirado de Birkes e Dodge (1993).

A Tabela 1.1 apresenta a taxa de nascimento (número de nascimentos por ano por 1000 pessoas) no período de 1980 a 1985, e a porcentagem urbana (porcentagem da população vivendo em cidades com mais de 100000 habitantes) em 1980, observadas em 14 países das Américas Central e do Norte com população acima de um milhão de pessoas em 1985.

Tabela 1.1. Dados de taxa de nascimento e porcentagem urbana.

País	Taxa de Nascimento	Porcentagem Urbana
Canadá	16,2	55,0
Costa Rica	30,5	27,3
Cuba	16,9	33,3
República Dominicana	33,1	37,1
El Salvador	40,2	11,5
Guatemala	38,4	14,2
Haiti	41,3	13,9
Honduras	43,9	19,0
Jamaica	28,3	33,1
México	33,9	43,2
Nicarágua	44,2	28,5
Panamá	28,0	37,7
Trinidad e Tobago	24,6	6,8
Estados Unidos	16,0	56,5

Fonte: Birkes e Dodge (1993)

No Gráfico 1.1 é apresentado o diagrama de dispersão das variáveis taxa de nascimento e porcentagem urbana. O gráfico sugere que países com porcentagem urbana alta têm taxa de nascimento menores e que a relação entre as duas variáveis pode ser explicada por uma reta. Adotando o método de mínimos quadrados, obtém-se a seguinte reta estimada:

$$\hat{y} = 43 - 0,40x. \quad (1.2)$$

No Gráfico 1.1, nota-se que o ponto Trinidad e Tobago se distancia da reta ajustada, sugerindo a ocorrência de um valor aberrante. Uma possibilidade é eliminar este ponto e ajustar novamente a reta por mínimos quadrados, obtendo-se:

$$\hat{y} = 49 - 0,55x. \quad (1.3)$$

Observa-se que a eliminação de Trinidad e Tobago acarretou uma modificação considerável nas estimativas dos parâmetros do modelo, indicando que este é um ponto de influência no conjunto de dados analisado.

Gráfico 1.1. Diagrama de dispersão da taxa de nascimento pela porcentagem urbana e as retas de regressão estimadas pelo método de mínimos quadrados com (linha contínua) e sem (linha tracejada) o ponto Trinidad e Tobago.

Entretanto, eliminar observações implica em perda de informação. Uma alternativa que permitiria ajustar um modelo de regressão sem eliminar observações seria adotar um método robusto de estimação.

Um dos métodos robustos mais conhecidos é o método L_1 , ou de mínima soma dos erros absolutos, ou de mínimos desvios absolutos. Os estimadores L_1 dos parâmetros do modelo (1.1) são os valores de $\beta_0, \dots, \beta_{(p-1)}$ que minimizam:

$$\sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_{(p-1)} x_{(p-1)i}|, \quad (1.4)$$

onde y_i é o valor da variável resposta para a i -ésima observação, $x_{1i}, \dots, x_{(p-1)i}$ são os respectivos valores das variáveis explicativas, e $\beta_0, \dots, \beta_{(p-1)}$ são os parâmetros do modelo.

Historicamente, o método de estimação L_1 é o mais velho de todos os métodos robustos, tendo sido introduzido quase 50 anos antes do método de mínimos quadrados, em 1757, por Roger Joseph Boscovich (Boscovich (1757)).

Boscovich propôs que a reta ajustada com base em três ou mais pontos não colineares em um plano, deveria satisfazer às seguintes condições:

- (i) A soma dos resíduos positivos e negativos deveriam ser iguais em magnitude, e
- (ii) A soma dos erros absolutos deveria ser mínima.

A regressão L_1 reapareceu em 1880 devido ao trabalho de Edgeworth. Suas contribuições incluem:

- (a) A condição (i) de Boscovich deveria ser descartada, de forma que a soma dos erros absolutos pudesse atingir seu mínimo (Edgeworth (1887));
- (b) A regressão L_1 pode não ser única;
- (c) Os estimadores L_1 são de máxima verossimilhança quando os erros seguem a distribuição de Laplace (Edgeworth (1888)).

Ele também desenvolveu um algoritmo para resolver o problema de regressão linear simples.

Até os anos 50, os problemas computacionais e de inferência estatística associados com a regressão L_1 impediram o seu uso. Desde então, vários algoritmos eficientes têm sido propostos, eliminando assim o problema computacional. Basset e Koenker (1978) deduziram a distribuição assintótica dos estimadores L_1 . Com base nestes resultados, procedimentos de inferência estatística foram propostos, removendo então as dificuldades inferenciais na utilização do método L_1 . Mais detalhes históricos podem ser encontrados em Farebrother (1987).

A reta L_1 ajustada para os dados da Tabela 1.1 é:

$$\hat{y} = 46 - 0,54x . \quad (1.5)$$

Comparando esta reta com a obtida por mínimos quadrados quando o ponto Trinidad e Tobago é excluído (1.3), é possível perceber que as estimativas estão bem próximas; em particular as do parâmetro de inclinação da reta.

O Gráfico 1.2 apresenta novamente o diagrama de dispersão das variáveis taxa de nascimento e porcentagem urbana e as retas ajustadas pelos métodos L_1 e de mínimos quadrados quando são consideradas todas as observações. Nota-se que a reta L_1 é menos influenciada pelo ponto aberrante.

Gráfico 1.2. Diagrama de dispersão da taxa de nascimento pela porcentagem urbana e as retas de regressão estimadas pelo método de mínimos quadrados (linha contínua) e pelo método L_1 (linha tracejada).

Ao contrário do método de mínimos quadrados, para o qual métodos de seleção de variáveis foram exhaustivamente estudados e comparados (Hocking (1976) e Neter et al. (1996), por exemplo), não é de nosso conhecimento um trabalho que descreva conjuntamente os procedimentos de seleção de variáveis existentes quando o método L_1 é adotado para o ajuste do modelo.

O objetivo deste trabalho é estudar diferentes métodos de seleção de variáveis na regressão L_1 . Serão apresentados também, programas computacionais desenvolvidos no aplicativo S-PLUS que auxiliam na parte inferencial e na seleção de variáveis.

O trabalho está organizado da seguinte forma: no Capítulo 2, são apresentados alguns resultados e comentários gerais sobre a regressão L_1 ; no Capítulo 3, são apresentados métodos de seleção de variáveis ilustrados através de um exemplo; no Capítulo 4, são feitas algumas considerações sobre o efeito da multicolinearidade na regressão L_1 ; no Capítulo 5, é feita uma aplicação da metodologia abordada, sendo apresentadas no Capítulo 6, conclusões e sugestões para pesquisas futuras.

Capítulo 2

Alguns Resultados em Regressão L_1

Neste capítulo, é feito um breve apanhado sobre alguns resultados inferenciais e métodos computacionais relacionados à regressão L_1 .

2.1. Algoritmos Computacionais

A obtenção das estimativas L_1 envolve um certo trabalho computacional, uma vez que os estimadores não possuem uma fórmula explícita. A seguir, são citados alguns algoritmos para a obtenção de tais estimativas.

2.1.1. Regressão Linear Simples

Seja y_i o valor da variável resposta associado a x_{1i} , o valor da variável explicativa para a i -ésima observação, $i=1, 2, \dots, n$, onde n é o número de observações. O modelo de regressão linear simples pode ser escrito como:

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i, \quad (2.1)$$

onde β_0 e β_1 são os parâmetros (intercepto e inclinação) desconhecidos do modelo e ε_i representa o erro aleatório não-observável.

O objetivo é determinar os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ de β_0 e β_1 , tais que $\sum_{i=1}^n |y_i - \hat{y}_i|$, a soma dos erros absolutos, seja mínima, onde $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$ é o valor ajustado para a i -ésima observação. Edgeworth (1888) propôs um algoritmo para encontrar as estimativas L_1 do modelo (2.1); contudo, seu método não parece ter sido largamente usado. Setenta anos depois, Karst (1958) propôs um algoritmo iterativo para o problema. Desde então, algoritmos foram sucessivamente propostos, por exemplo, Barrodale e Roberts (1973), Armstrong e Kung (1978), Abdelmalek (1980), Wesolowsky (1981), Klingman e Mote (1982) e Josavenger e Sposito (1983).

2.1.2. Regressão Linear Múltipla

Charnes et al. (1955) notaram que o problema da regressão L_1 é essencialmente um problema de programação linear. No trabalho de Wagner (1959) o problema foi formulado da seguinte forma:

$$\begin{aligned} &\text{Minimizar } \mathbf{1}'(\mathbf{e}^+ + \mathbf{e}^-) \\ &\text{sujeito a } \mathbf{X}\boldsymbol{\beta} + \mathbf{e}^+ - \mathbf{e}^- = \mathbf{y}, \\ &\mathbf{e}^+, \mathbf{e}^- \geq 0, \\ &\boldsymbol{\beta} \text{ irrestrito em sinal,} \end{aligned} \tag{2.2}$$

onde $\mathbf{1}$ é um vetor $n \times 1$ de 1s e $\mathbf{e}^+$ e $\mathbf{e}^-$ são vetores $n \times 1$ de resíduos positivos e negativos, respectivamente. Barrodale e Roberts (1973) propuseram um algoritmo muito eficiente para solucionar (2.2). Atualmente, vários algoritmos: Bartels et al. (1978), Armstrong et al. (1979), Bloomfield e Steiger (1983), Wesolowsky (1981), Ruzinsky e Olsen (1989), Coleman e Li (1992), Madsen e Nielsen (1993) e Zhang (1993), entre outros, estão disponíveis para resolver os modelos de regressão L_1 linear múltipla.

2.2. Programas Computacionais

Existe um grande número de programas computacionais para calcular as estimativas L_1 dos parâmetros dos modelos de regressão linear simples e múltipla. Por exemplo, programas são encontrados em Barrodale e Roberts (1974), Bartels et al. (1976), Armstrong e Kung (1978) e Armstrong et al. (1979).

Para o problema de regressão linear simples, o programa de Josavenger e Sposito (1983), que é baseado na modificação do algoritmo de Wesolowsky (1981) mostrou-se bom num estudo comparativo de Gentle et al. (1988). Num estudo para comparar a *performance* relativa dos programas para regressão linear múltipla, Gentle et al. (1987) relataram que, dentro das limitações do estudo, o programa de Armstrong et al. (1979) apresentou melhor desempenho que os demais.

O modelo L_1 pode também ser ajustado através de rotinas dos aplicativos SAS (procedimento *glm*) e S-PLUS (rotina *lfit*). Entretanto, essas rotinas não fornecem resultados inferenciais além das estimativas pontuais dos parâmetros. Um programa computacional desenvolvido no aplicativo S-PLUS, que fornece estimativas por intervalo e testes de hipóteses sobre os parâmetros do modelo, é apresentado no Apêndice A.

2.3. Inferência Estatística

Suponha que os componentes do vetor de erros são variáveis aleatórias tais que:

(S1) $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ são independentes, identicamente distribuídos com função de distribuição F .

(S2) F possui mediana única que, sem perda de generalidade, é zero e F tem uma derivada contínua f numa vizinhança de zero tal que $f(0) > 0$.

(S3) $(1/n)X^tX \rightarrow C$ à medida que $n \rightarrow \infty$, onde C é uma matriz positiva definida.

(S4) Para todo $j=1, \dots, p$, $\max_{1 \leq i \leq n} \frac{(x_{ij} - \bar{x}_j)^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \rightarrow 0$, à medida que $n \rightarrow \infty$.

Defina um parâmetro de escala τ como $\tau=(2(f(0)))^{-1}$, isto é, τ^2/n é a variância da mediana em amostra de tamanho n da distribuição dos erros (Cox e Hinkley (1991)).

A partir destas suposições, Basset e Koenker (1978) mostraram que o estimador L_1 de β no modelo de regressão é assintoticamente não-viciado, consistente e segue uma distribuição Normal Multivariada com matriz de variância-covariância $\tau^2(X'X)^{-1}$. Uma implicação importante deste resultado é que o estimador L_1 produz um elipsóide de confiança menor do que o do estimador de mínimos quadrados para β , para qualquer distribuição em que a mediana amostral é um estimador mais eficiente do que a média.

É sabido que os estimadores L_1 são também os de máxima verossimilhança e, por isso, são assintoticamente não viciados e eficientes, quando os erros têm distribuição de Laplace.

Com base nos resultados assintóticos, fórmulas para construção de intervalos de confiança e testes de hipóteses sobre os parâmetros do modelo têm sido desenvolvidas (Dielmann e Pfaffenberger (1982) e Narula (1987)). A seguir, encontram-se algumas fórmulas para intervalos de confiança e testes de hipóteses sobre combinações lineares $r'\beta$ dos parâmetros, onde r é um vetor $p \times 1$ de constantes conhecidas.

Um intervalo com $(1-\alpha)\%$ de confiança para $r'\beta$ pode ser escrito como:

$$r'\hat{\beta} \pm z_{\alpha/2} \hat{\tau} \sqrt{r'(X'X)^{-1}r}, \quad (2.3)$$

onde, $z_{\alpha/2}$ é o percentil de ordem $(1-\alpha/2)$ da distribuição Normal(0,1), $\hat{\beta}$ é o vetor das estimativas dos parâmetros e $\hat{\tau}$ é um estimador consistente de τ .

Vários estimadores de τ têm sido propostos. Um estimador consistente recomendado por Birkes e Dodge (1993) e McKean e Schrader (1987) é:

$$\hat{\tau} = \sqrt{n^*} (e_{(n^*-m+1)} - e_{(m)}) / 4, \quad (2.4)$$

onde, $m=(n^*+1)/2 - n^{*1/2}$, n^* é o número de resíduos diferentes de zero, e $e_{(1)}, e_{(2)}, \dots, e_{(n^*)}$ são os resíduos não nulos em ordem crescente.

Um intervalo com $(1-\alpha)\%$ de confiança para um único componente de β é dado por:

$$\hat{\beta}_i \pm z_{\alpha/2} \hat{\tau} \sqrt{(X^t X)^{-1}_{ii}}, \quad (2.5)$$

onde, $(X^t X)^{-1}_{ii}$ é o i -ésimo elemento da diagonal de $(X^t X)^{-1}$ e $\hat{\beta}_i$ é o i -ésimo componente de $\hat{\beta}$.

Um intervalo com $(1-\alpha)\%$ de confiança para a resposta média $y_i = X_i^t \beta$ pode ser obtido da seguinte forma:

$$X_i^t \hat{\beta} \pm z_{\alpha/2} \hat{\tau} \sqrt{X_i^t (X^t X)^{-1} X_i}, \quad (2.6)$$

onde, X_i^t é a i -ésima linha da matriz X .

Um intervalo de previsão com $(1-\alpha)\%$ de confiança para uma nova observação y_{nova} , sendo X_{novo} o vetor de valores das variáveis explicativas correspondente à nova observação, tem a seguinte forma:

$$X_{\text{novo}}^t \hat{\beta} \pm z_{\alpha/2} \hat{\tau} \sqrt{(1 + X_{\text{novo}}^t (X^t X)^{-1} X_{\text{novo}})}. \quad (2.7)$$

Para testar a hipótese nula $H_0: r^t \beta = \rho$ versus $H_1: r^t \beta \neq \rho$ ao nível de significância α , a regra de decisão é rejeitar H_0 quando:

$$z^* = \left| \frac{r^t \hat{\beta} - \rho}{\hat{\tau} \sqrt{r^t (X^t X)^{-1} r}} \right| > z_{\alpha/2}. \quad (2.8)$$

Para testar a hipótese nula $H_0: \beta_i = 0$ versus $H_1: \beta_i \neq 0$, a regra de decisão é rejeitar H_0 quando:

$$z^* = \left| \frac{\hat{\beta}_i}{\hat{\tau} \sqrt{(X^t X)^{-1}_{ii}}} \right| > z_{\alpha/2}. \quad (2.9)$$

Para testar a contribuição conjunta de q variáveis regressoras na explicação da variável resposta, ou seja, a hipótese $H_0: \beta_{(p-q)} = \dots = \beta_{(p-1)} = 0$, McKean e Schrader (1987) propuseram a estatística:

$$D_q = \frac{SEA_{\text{reduzido}} - SEA_{\text{completo}}}{\hat{\tau} / 2}, \quad (2.10)$$

onde $SEA_{\text{completo}} = \sum_{i=1}^n |y_i - \hat{y}_i|$ representa a soma dos erros absolutos no modelo com $p-1$ variáveis explicativas, e SEA_{reduzido} é definida de forma análoga no modelo com $p-q-1$ variáveis explicativas. Sob H_0 , D_q , para valores grandes de n , possui uma distribuição χ^2_q .

Procedimentos de inferência para pequenas amostras têm sido investigados usando estudos de Monte Carlo (por exemplo, Stangenhuis e Narula (1991)). Os resultados mostraram que a distribuição amostral do estimador L_1 apresentou-se Normal para tamanhos de amostra de 20 e 30 quando os erros seguem distribuição Normal ou Normal contaminada, respectivamente. No caso da distribuição dos erros ser Cauchy ou Laplace, os tamanhos de amostra necessários para a distribuição amostral ser Normal foram, naquele estudo, respectivamente 100 e 150.

Além disso, para as distribuições dos erros mencionadas no parágrafo anterior, a distribuição Normal pode ser usada para construção de intervalos de confiança e em testes de hipóteses dos parâmetros do modelo, mesmo que a distribuição amostral não convirja para Normal com tamanhos de amostra a partir de 10.

Stangenhuis et al. (1993) também propuseram procedimentos *bootstrap* para inferência em pequenas amostras.

2.4. Robustez

Appa e Smith (1973) mostraram que:

- (i) Pelo menos um hiperplano que minimiza a soma dos erros absolutos passa por p das n observações;

- (ii) Sob a hipótese de que nenhum conjunto de $p+1$ observações fica em um hiperplano em p dimensões, o hiperplano L_1 satisfaz à condição: $|n^+ - n^-| < p$, onde n^+ e n^- denotam o número de observações com resíduos positivos e negativos, respectivamente.

Como pelo menos um hiperplano de regressão L_1 passa por p das n observações, então os resíduos observados $e_i = y_i - \hat{y}_i$ são iguais a zero para p observações. As observações com resíduo zero são chamadas definidoras ou básicas e determinam completamente as estimativas L_1 dos parâmetros do modelo.

É interessante notar que a regressão L_1 está para a de mínimos quadrados, assim como a mediana amostral está para a média amostral. Por exemplo, tanto a média amostral como os estimadores de mínimos quadrados são determinados e influenciados por todas as observações, ao passo que a mediana amostral e os estimadores L_1 são determinados somente por um subconjunto das observações.

Assim como o valor da mediana amostral não é afetado se uma observação é alterada, desde que continue do mesmo lado (acima ou abaixo) da mediana amostral, um resultado análogo é verdadeiro para a variável dependente na regressão L_1 : o modelo ajustado não se modifica se um valor observado da variável resposta for alterado de forma a permanecer no mesmo lado do hiperplano de regressão L_1 . Narula e Wellington (1985) apresentam um estudo de sensibilidade, que fornece limites inferiores e superiores para cada observação da variável resposta de tal forma que, se ela for uma observação definidora, continue sendo, mesmo que o valor de $\hat{\beta}$ seja alterado, e se ela não for definidora e assumir valores dentro desses limites, o hiperplano de regressão L_1 não muda.

Além disso, a regressão L_1 permanece a mesma se o valor das variáveis independentes, para as observações com resíduos não nulos, fica dentro de certos intervalos. O processo para calcular tais intervalos foi proposto em Narula e Wellington (1998). Estes intervalos proporcionam ao analista informação útil sobre a precisão das observações que deixa a regressão L_1 inalterada.

Capítulo 3

Métodos de Seleção de Variáveis

3.1. O Problema de Seleção de Variáveis

No ajuste de um modelo de regressão, assume-se, em geral, que as $p-1$ variáveis explicativas sejam todas relevantes para explicar a variabilidade da variável resposta. Contudo, algumas variáveis explicativas podem não ter uma contribuição significativa por serem pouco correlacionadas com a resposta, ou por serem muito correlacionadas entre si. Frequentemente é possível selecionar um modelo com $k < p-1$ variáveis sem perder, essencialmente, a informação sobre a variável resposta contida nas $p-1$ variáveis. Além disso, um modelo com muitas variáveis pode ser difícil de ser mantido, ou ser desvantajoso do ponto de vista econômico. Neste capítulo serão discutidos diferentes métodos de seleção de variáveis em modelos de regressão múltipla, quando é adotado o método de estimação L_1 .

As condições de coleta de dados para a construção de um modelo de regressão variam de acordo com a natureza do estudo a ser conduzido. Neter et al. (1996) considera que um experimento pode ser classificado em um dos seguintes tipos:

- *Experimentos Controlados*: os níveis das variáveis explicativas são controlados e, a cada unidade experimental, é associado um tratamento que consiste na combinação dos níveis das variáveis explicativas, sendo então observado o valor da variável resposta. Em experimentos controlados, as variáveis explicativas são frequentemente chamadas de fatores. Por exemplo, um horticultor conduziu um experimento para estudar os efeitos de variedade de flor (fator A: variedades LP e WB) e nível de mistura de adubo (fator B: baixo e alto) na produção de flores.

Os tratamentos consistem nas combinações dos diferentes níveis das variáveis variedade de flor e nível de mistura.

- *Experimentos Controlados com Variáveis Suplementares*: além dos fatores, é considerada informação adicional sobre características das unidades experimentais, com o objetivo de reduzir a variância dos termos de erro no modelo de regressão. Às vezes, contudo, não é possível incorporar esta informação adicional na etapa de planejamento do experimento. Ao invés disso, pode ser possível incorporar esta informação através da introdução de variáveis suplementares (covariáveis) no modelo de regressão. No exemplo anterior sobre a produção de flores, o pesquisador suspeita que o tamanho do canteiro pode afetar a produção de maneira significativa. Então, o pesquisador inclui o tamanho dos canteiros como covariável, o que torna a análise dos efeitos das variáveis explicativas na variável resposta mais precisa.
- *Estudos Observacionais Confirmatórios*: este tipo de estudo é baseado em dados observacionais, isto é, dados que foram obtidos através de estudos não-experimentais, nos quais não é possível controlar as variáveis explicativas de interesse. O objetivo deste estudo é confirmar ou não hipóteses derivadas de estudos prévios ou intuição. Para estes estudos, os dados são coletados para variáveis explicativas que, sabidamente, mostraram afetar a variável resposta em outros estudos, bem como para nova(s) variável(eis) envolvida(s) na hipótese a ser averiguada. Neste contexto, as variáveis explicativas relacionadas com hipótese de interesse são chamadas variáveis principais e, as variáveis incluídas por refletir o conhecimento existente do problema em questão são chamadas variáveis controle (também denominadas fatores de risco em epidemiologia). Aqui, as variáveis controle são utilizadas porque sabe-se que explicam parte da variabilidade da variável resposta. Por exemplo, em um estudo observacional do efeito da vitamina E na incidência de um certo tipo de câncer, fatores de risco conhecidos, como idade, sexo, e raça, poderiam ser incluídos como variáveis controle e a quantidade de vitamina E ingerida diariamente seria a variável explicativa principal. A variável resposta seria a ocorrência ou não do tipo de câncer em estudo no período de interesse.

- *Estudos Observacionais Exploratórios*: em vários campos de pesquisa, frequentemente não é possível conduzir experimentos controlados. Mais ainda, não se tem o conhecimento suficiente para conduzir um estudo observacional confirmatório. O pesquisador procura por variáveis explicativas que possam estar relacionadas com a variável resposta. O conjunto de potenciais variáveis explicativas pode ser muito grande. Por exemplo, as vendas de uma companhia em uma cidade podem ser afetadas pelo tamanho da população, renda *per capita*, porcentagem da população em áreas urbanas, porcentagem da população com menos de 50 anos, porcentagem de famílias com crianças, etc... Algumas destas variáveis podem ser eliminadas, pois uma variável explicativa pode não ser fundamental para o problema, ser sujeita a erros de medida, duplicar a informação dada por uma outra variável ou nem mesmo ser medida.

Após as etapas de escolha do tipo de estudo e ser conduzido e de coleta de dados, começa o problema da escolha do modelo que melhor explica as questões de interesse, segundo os objetivos do pesquisador. Isto nada mais é que um problema de seleção de variáveis e, para cada um dos quatro tipos de estudo comentados anteriormente, esta questão é abordada de maneira apropriada:

- *Experimentos Controlados*: a redução de variáveis explicativas na fase de construção do modelo não é uma questão relevante neste tipo de experimento. O pesquisador escolhe as variáveis explicativas para investigação, e um modelo de regressão é desenvolvido de tal maneira a capacitá-lo para estudar os efeitos destas variáveis na resposta.
- *Experimentos Controlados com Variáveis Suplementares*: alguma redução nas variáveis suplementares pode ser razoável, pois o pesquisador, com frequência, não tem certeza de antemão quais variáveis suplementares escolhidas serão úteis na explicação da variabilidade da variável resposta. O número de variáveis suplementares consideradas num experimento controlado costuma ser pequeno, e portanto, nenhum problema em especial é encontrado em determinar se algumas ou todas as variáveis suplementares devem ser descartadas do modelo.

- *Estudos Observacionais Confirmatórios*: geralmente nenhuma redução de variáveis explicativas é usada neste tipo de estudo. As variáveis controle são escolhidas com base em resultados de estudos anteriores, e as variáveis explicativas principais são aquelas cuja influência na variável resposta se deseja examinar e, por isso, devem estar no modelo.
- *Estudos Observacionais Exploratórios*: conforme foi comentado anteriormente, aqui o número de variáveis explicativas é tipicamente grande. Mais ainda, muitas destas variáveis são correlacionadas. Portanto, a redução do número de variáveis explicativas é pertinente e desejável. Na regressão de mínimos quadrados, a presença de variáveis muito correlacionadas no modelo pode provocar um aumento na variância dos estimadores dos coeficientes e prejudicar a capacidade preditiva do modelo ajustado; não é de nosso conhecimento um estudo sobre o efeito da multicolinearidade na regressão L_1 e este assunto será abordado no Capítulo 4. Uma das etapas dos estudos observacionais exploratórios consiste então na identificação de alguns “bons subconjuntos” das variáveis em estudo. No entanto, este processo é um dos problemas mais difíceis na análise de regressão, uma vez que pode não haver um único melhor modelo, e a escolha entre esses modelos “igualmente bons” varia conforme os critérios e os objetivos do pesquisador. A seleção das variáveis mais importantes para o problema em estudo deve ser feita com muito cuidado pois, as variáveis mais importantes não devem ser eliminadas para não prejudicar o poder de explicação do modelo. Além disso, selecionadas as variáveis, o modelo final pode ser obtido verificando-se a existência de termos de maior grau e interações.

Neste capítulo, serão apresentados dois conjuntos de procedimentos de seleção de variáveis quando o método de estimação L_1 é adotado. O primeiro, que é recomendado para um número pequeno ou moderado de variáveis explicativas, considera todos os subconjuntos de variáveis e identifica aqueles que são “bons” segundo critérios especificados pelo investigador. O segundo faz uso de procedimentos automáticos de busca para chegar a um único subconjunto, e pode ser utilizado em estudos envolvendo um número qualquer de variáveis. É claro que o uso destes procedimentos automáticos deve ser acompanhado de julgamento subjetivo, ou seja, a

escolha do modelo final deve se basear também no conhecimento do pesquisador sobre o problema.

Para ilustrar os diferentes métodos de seleção de variáveis a serem abordados, serão usados os dados relativos a imóveis, retirados de Narula e Wellington (1977), contendo as variáveis imposto, em centenas de dólares (x_1), área do terreno, em milhares de pés quadrados (x_2), área construída, em milhares de pés quadrados (x_3), idade da residência, em anos (x_4), e a variável resposta, preço de venda da casa (y), em milhares de dólares. A Tabela 3.1 apresenta este conjunto de dados.

Tabela 3.1. Dados relativos a imóveis.

y	x_1	x_2	x_3	x_4
25,9	4,9176	3,4720	0,998	42
29,5	5,0208	3,5310	1,500	62
27,9	4,5429	2,2750	1,175	40
25,9	4,5573	4,0500	1,232	54
29,9	5,0597	4,4550	1,121	42
29,9	3,8910	4,4550	0,988	56
30,9	5,8980	5,8500	1,240	51
28,9	5,6039	9,5200	1,501	32
84,9	15,4202	9,8000	3,420	42
82,9	14,4598	12,8000	3,000	14
35,9	5,8282	6,4350	1,225	32
31,5	5,3003	4,9883	1,552	30
31,0	6,2712	5,5200	0,975	30
30,9	5,9592	6,6660	1,121	32
30,0	5,0500	5,0000	1,020	46
36,9	8,2464	5,1500	1,664	50
41,9	6,6969	6,9020	1,488	22
40,5	7,7841	7,1020	1,376	17
43,9	9,0384	7,8000	1,500	23
37,5	5,9894	5,5200	1,256	40
37,9	7,5422	4,0000	1,690	22
44,5	8,7951	9,8900	1,820	50
37,9	6,0931	6,7265	1,652	44
38,9	8,3607	9,1500	1,777	48
36,9	8,1400	8,0000	1,504	3
45,8	9,1416	7,3262	1,831	31
41,0	12,0000	5,0000	1,200	30

Fonte: Narula e Wellington (1977)

3.2. Todas as Possíveis Regressões

Este procedimento considera todos os possíveis subconjuntos de variáveis explicativas que podem ser construídos a partir do conjunto original de variáveis. Na Tabela 3.2 estão listados todos os 16 modelos de regressão com intercepto, que podem ser formados a partir do exemplo dos imóveis e seus respectivos número de parâmetros (p).

Tabela 3.2. Número de parâmetros (p) para todos os modelos de regressão possíveis, utilizando os dados de imóveis.

Variáveis no Modelo	p
Nenhuma	1
x_1	2
x_2	2
x_3	2
x_4	2
x_1, x_2	3
x_1, x_3	3
x_1, x_4	3
x_2, x_3	3
x_2, x_4	3
x_3, x_4	3
x_1, x_2, x_3	4
x_1, x_2, x_4	4
x_1, x_3, x_4	4
x_2, x_3, x_4	4
x_1, x_2, x_3, x_4	5

Em algumas situações, pode ser impossível examinar todos os possíveis subconjuntos de variáveis. Por exemplo, se fossem consideradas 10 variáveis em um estudo, haveria $2^{10} = 1024$ regressões, o que tornaria inviável o uso da estratégia de investigar todos os modelos possíveis.

Quando o método de mínimos quadrados é utilizado, diferentes critérios de comparação de modelos podem ser usados para seleção dentre todas as possíveis regressões: o coeficiente de determinação múltipla (R^2), a soma de quadrados de resíduos, o coeficiente de determinação

múltipla ajustado (R_a^2), a soma dos erros quadráticos preditivos (PRESS) e o C_p (ver Neter et al. (1996), por exemplo). Serão apresentados a seguir alguns critérios de comparação próprios à regressão L_1 .

3.2.1. Critério da Soma dos Erros Absolutos - Todos os Possíveis Modelos

A Soma dos Erros Absolutos (SEA) para um modelo de regressão com $p-1$ variáveis explicativas é dada por:

$$SEA = \sum_{i=1}^n |y_i - \hat{y}_i|. \quad (3.1)$$

O objetivo deste critério é determinar qual o modelo que produz a menor SEA, e tal que a inclusão de variáveis adicionais faz com que a SEA não diminua significativamente. No Gráfico 3.1 estão representados todos os valores das SEA para todos os possíveis modelos apresentados na Tabela 3.2, como uma função de p . Ao lado de cada ponto estão indicadas as variáveis consideradas. As SEA são também apresentadas na Tabela 3.5.

Gráfico 3.1. Valores da SEA em função do número de parâmetros no modelo.

O menor valor da SEA, excluindo o modelo completo ($p=5$), corresponde ao modelo com as variáveis x_1 , x_3 e x_4 (78,3789). Este valor é muito próximo da SEA do modelo completo (77,8051), ou seja, quase não há ganho com a inclusão de x_2 no modelo. Contudo, o modelo com x_1 e x_3 (84,0668) possui um valor SEA próximo ao valor do modelo com x_1 , x_3 e x_4 , sugerindo que há pouco ganho com a inclusão de x_4 no modelo. Portanto, segundo o critério da SEA, o modelo com as variáveis x_1 e x_3 é o escolhido.

3.2.2. Critério da Soma dos Erros Absolutos - Algoritmo de Narula e Wellington

Como foi comentado, a busca do “melhor” modelo considerando-se todas as possíveis regressões pode se tornar impraticável quando o número de variáveis independentes é muito grande. Um algoritmo eficiente para determinar este “melhor” modelo sob o critério da SEA, entre todos os modelos possíveis, foi proposto por Narula e Wellington (1979). Considerando

propriedades da SEA, esse procedimento possibilita não examinar todos os modelos, reduzindo assim os cálculos e o custo computacional.

A seleção do melhor modelo de tamanho k ($1 \leq k < p-1$), será apresentada utilizando os dados da Tabela 3.1.

Primeiramente deve ser observado que a soma dos erros absolutos em um modelo com m variáveis explicativas é sempre menor ou igual à soma dos erros absolutos de qualquer modelo com $k = 1, 2, \dots, m-1$ variáveis explicativas que pode ser obtido a partir dessas m variáveis, ou seja:

$$SEA_m \leq SEA_k, k = 1, 2, \dots, m-1. \quad (3.2)$$

Portanto, o seguinte teste de sondagem pode ser usado: se a soma dos erros absolutos de um modelo com determinadas m variáveis explicativas é maior ou igual à soma dos erros absolutos de um modelo com um número menor de variáveis explicativas, então os modelos com $k = 1, 2, \dots, m-1$ variáveis que podem ser obtidos a partir dessas m variáveis não necessitam ser investigados.

A Figura 3.1 apresenta uma *árvore*, que consiste num arranjo de todos os modelos possíveis para o exemplo dos imóveis, sendo cada modelo representado por um *nó*. A importância de se construir uma *árvore* e segui-la de maneira adequada, reside na meta de que se tenha a menor quantidade possível de cálculos.

Figura 3.1. *Árvore* completa com as somas dos erros absolutos de cada modelo para o exemplo dos imóveis.

Algumas relações entre modelos podem ser definidas, através de uma *árvore*. Se um modelo com $m-1$ variáveis é obtido retirando-se uma variável de um com m variáveis, o modelo com m variáveis é chamado *pai* e o com $m-1$ variáveis, *filho*. Todos os modelos com $m-1$ variáveis obtidos a partir das mesmas m variáveis são denominados *irmãos*. A *raiz* da *árvore* representa o modelo com todas as variáveis explicativas e os demais *nós* são os modelos obtidos a partir dos subconjuntos dessas variáveis. Os *nós* são obtidos retirando-se, a partir da esquerda (ou direita), uma variável explicativa de cada vez. No caso da Figura 3.1, começando pela *raiz*, os *nós* foram obtidos retirando-se as variáveis explicativas a partir da esquerda e descendo na *árvore* pela direita, ou seja, uma vez que foram obtidos os *nós* 234, 134, 124 e 123, foram obtidos os *filhos* de 234 (34, 24 e 23), em seguida os de 134 e assim por diante, tomando o cuidado de não considerar *nós* repetidos.

A forma de percorrer uma *árvore* pode ser resumida da seguinte forma:

- Inicia-se pela *raiz* da *árvore* e move-se de um *pai* para um *filho* sucessivamente.

- Quando se atingir um *nó* terminal, ou seja, que não possua *filhos*, move-se para o próximo *irmão* ou, se não há *irmãos*, para outro *tio*, ou, se não há *tios*, para outro *irmão* do avô etc...

Por exemplo, uma possível forma de percorrer a *árvore* da Figura 3.1 é: 1234, 234, 34, 4, 3, 24, 2, 23, 134, 14, 1, 13, 124, 12 e 123. Esta construção pode ser visualizada na Figura 3.2:

Figura 3.2. Exemplo de como percorrer uma *árvore* completa para quatro variáveis.

Para iniciar o processo de busca, precisam ser estabelecidos valores iniciais para as somas dos erros absolutos para os modelos com 1, 2, ..., $p-1$ variáveis. Uma forma de se obter esses limites seria, adotando um procedimento *stepwise*, selecionar modelos com 1, 2, ..., $p-1$ variáveis explicativas pelo método de mínimos quadrados. Selecionadas as variáveis para compor os modelos de cada tamanho, ajustá-los pelo método L_1 , obtendo a soma dos erros absolutos que funcionariam como valores iniciais.

No exemplo, utilizando o procedimento de seleção *forward* para o método de mínimos quadrados, foram selecionados os melhores modelos de cada tamanho, e em seguida foram ajustadas as regressões L_1 com as variáveis selecionadas, para se obter as somas dos erros

absolutos associadas a estes modelos. Estas somas, que são os limites superiores iniciais, se encontram na Tabela 3.3.

Tabela 3.3. Limites superiores iniciais da SEA para os modelos do exemplo dos imóveis.

Número de Variáveis	Variáveis	SEA
1	x_1	103,4292
2	x_1, x_2	103,2933
3	x_1, x_2, x_3	83,1100

Em resumo, o algoritmo pode ser descrito da seguinte forma:

- Primeiramente obtém-se um limite superior para a soma dos erros absolutos para um modelo de cada tamanho.
- Durante o algoritmo, esses limites são substituídos pela soma dos erros absolutos do melhor modelo de cada tamanho encontrado até aquele ponto.
- Um modelo de cada dimensão é investigado seguindo o percurso estabelecido na Figura 3.2. Se a sua soma dos erros absolutos é menor que o limite para modelos do mesmo tamanho, então esse limite deve ser alterado passando a ser a SEA do modelo que está sendo pesquisado. Se sua soma dos erros absolutos é maior que o limite para modelos de menor dimensão, então os modelos de menor dimensão que podem ser obtidos a partir desse modelo não serão analisados, devido à propriedade dada em (3.2).
- Um modelo entre aqueles ainda não investigados é então escolhido.
- A soma dos erros absolutos desse modelo é comparada com os limites superiores atuais para verificar se o limite superior para modelos do mesmo tamanho deve ser alterado e se outros modelos de menor tamanho podem ser descartados.
- O processo continua dessa forma até que seja obtido o melhor modelo de cada tamanho.

O caminho percorrido no exemplo é ilustrado na Figura 3.3 e descrito em seguida.

Figura 3.3. Caminho percorrido no exemplo dos imóveis.

No exemplo, foi inicialmente ajustado o modelo de regressão L_1 com as quatro variáveis, cuja soma dos erros absolutos é 77,8051. Em seguida foi investigado o modelo com as variáveis x_2 , x_3 e x_4 . A soma dos erros absolutos deste modelo (94,7217) é superior ao limite inicial para modelos com três variáveis. Portanto, o limite inicial para três variáveis não é alterado. Contudo, a SEA deste modelo é inferior ao limite superior para modelos com duas variáveis, o que significa que os modelos com duas variáveis obtidos a partir de x_2 , x_3 e x_4 precisam ser pesquisados. Foi então investigado o modelo com x_3 e x_4 . A soma dos erros absolutos do modelo com estas variáveis (101,5501) é inferior ao limite inicial para duas variáveis, tornando-se então o novo limite para modelos deste tamanho. Essa soma também é inferior ao limite para modelos com uma variável e, portanto, modelos desse tamanho necessitam ser pesquisados. Descendo um nível na árvore, nota-se que a soma dos erros absolutos do modelo com x_4 (206,8500) é superior ao limite para modelos com uma variável, sendo então descartado. O mesmo ocorre com o modelo com x_3 . Analisou-se então, o modelo com x_2 e x_4 , cuja soma dos erros absolutos (159,6650) é superior ao limite para modelos com duas variáveis; essa SEA é também superior ao limite para modelos com uma variável e, portanto, não é necessário pesquisar-se o modelo x_2 . Passa-se para o modelo com x_2 e x_3 , obtendo-se soma dos erros absolutos igual a 109,1555, que é superior ao limite para modelos com duas variáveis. Partiu-se então para o modelo com x_1 , x_3 e

x_4 , obtendo-se soma dos erros absolutos igual a 78,3789, que é inferior ao limite para modelo com três variáveis e por isso passou a ser o novo limite para modelos desse tamanho. Como a SEA deste modelo é também inferior ao limite superior para modelos com duas variáveis, continua-se neste ramo. Foi então examinado o modelo com x_1 e x_4 , sendo que sua soma dos erros absolutos (103,3377) é inferior ao limite para uma variável, portanto este ramo não foi descartado. Descendo mais um nível nesse mesmo ramo, foi avaliado o modelo com x_1 , pois o modelo com x_4 já foi descartado anteriormente. Por esta razão o melhor modelo com uma variável é com x_1 . Avaliando, em seguida, o modelo com x_1 e x_3 , a soma dos erros absolutos associada a este modelo foi 84,0668, que passou a ser o novo limite para modelos com duas variáveis.

Avaliando o modelo com x_1 , x_2 e x_4 , foi obtida uma soma dos erros absolutos igual a 103,2552 que é superior ao limite para duas variáveis (84,0668). Portanto, o modelo com x_1 e x_2 deste ramo não foi analisado. Finalmente, a soma dos erros absolutos do modelo com x_1 , x_2 e x_3 (83,1100) é superior ao limite para três variáveis (78,3789) e este modelo foi descartado.

Portanto, para o exemplo considerado, o melhor modelo com três variáveis é com x_1 , x_3 e x_4 , o melhor com duas variáveis é com x_1 e x_3 e, o melhor com uma variável é com x_1 .

Apesar de ser eficiente, este algoritmo pode apresentar alguns problemas se o número de variáveis explicativas for muito grande. Finalmente, ao término do processo haverá mais de um modelo à disposição do pesquisador, um de cada tamanho. Um programa em FORTRAN para este algoritmo pode ser encontrado em Wellington e Narula (1981); este programa pode resolver um problema com até 20 variáveis explicativas e 100 observações.

3.2.3. Critério do Coeficiente de Determinação

O coeficiente de determinação R^2 , definido como:

$$R^2 = (SQT - SQR) / SQT, \quad (3.3)$$

onde SQR é a Soma de Quadrados dos Resíduos e SQT a Soma de Quadrados Total, é frequentemente utilizado, na regressão de mínimos quadrados, como uma medida intuitiva do

ajuste de um modelo, uma vez que pode ser interpretado como a porcentagem da variabilidade total na variável resposta que é explicada pelo modelo adotado. Entretanto, este coeficiente não é robusto, e mesmo um único valor aberrante pode afetar drasticamente o seu valor. O método L_1 é uma boa alternativa ao de mínimos quadrados quando a distribuição dos erros tem caudas pesadas. Nas situações em que esse método é adotado, seria interessante a utilização de um coeficiente de determinação que também seja resistente a valores aberrantes. Buscando definir uma estatística com essa característica, McKean e Sievers (1987), inicialmente, consideraram que, no caso de variáveis independentes não-estocásticas, um coeficiente de determinação R deve possuir as seguintes propriedades:

- (P1) R deve estar diretamente ligado ao critério de ajuste, uma vez que ele é usado para medir a qualidade do ajuste de um modelo.
- (P2) R deve medir o quanto a inclusão das variáveis explicativas melhora o ajuste do modelo e estar relacionado ao teste da contribuição conjunta das $p-1$ variáveis explicativas.
- (P3) R deve ser adimensional, invariante perante mudanças de locação e escala na variável resposta ou nas explicativas.
- (P4) $0 \leq R \leq 1$, com 1 correspondendo ao ajuste perfeito e 0 à falta de ajuste.
- (P5) R deve aumentar quando se acrescentam novas variáveis preditoras ao modelo.
- (P6) R deve ser robusto.

Com o objetivo de obter um coeficiente com tais propriedades, os autores consideraram o seguinte coeficiente:

$$R_1 = \frac{RSEA}{\sum_{i=1}^n |y_i - \hat{\beta}_0|}, \quad (3.4)$$

onde $\hat{\beta}_0 = \text{mediana}(y_1, \dots, y_n)$ e $RSEA$ é a redução na soma dos erros absolutos quando se passa de um modelo ajustado sem nenhuma variável explicativa, ou seja, $\hat{y}_i = \text{mediana}(y_1, \dots, y_n) = \hat{\beta}_0$, para um modelo ajustado com $p-1$ variáveis explicativas, isto é:

$$RSEA = \sum_{i=1}^n |y_i - \hat{\beta}_0| - \sum_{i=1}^n |y_i - \hat{y}_i|. \quad (3.5)$$

Sua interpretação na regressão L_1 é a mesma que a da soma dos quadrados da regressão no método de mínimos quadrados. A soma dos desvios absolutos das observações y_i em relação à sua mediana pode ser denominada soma dos erros absolutos total. Portanto, o coeficiente de determinação proposto em (3.4) representa a proporção de redução na soma dos erros absolutos total devido ao ajuste de um modelo com $p-1$ variáveis explicativas, tendo então interpretação muito semelhante ao R^2 de mínimos quadrados.

McKean e Sievers mostraram que esse coeficiente não é robusto quando um valor aberrante é observado, contudo espera-se que R_1 seja mais robusto que R^2 .

Para se obter um coeficiente de determinação robusto baseado no método L_1 , foi considerado inicialmente que, quando é adotado o método de mínimos quadrados, a estatística para testar a contribuição conjunta das $p-1$ variáveis explicativas no modelo, isto é, para testar a hipótese:

$$H_0: \beta^* = \mathbf{0} \text{ versus } H_1: \beta^* \neq \mathbf{0}, \quad (3.6)$$

onde $\beta^* = (\beta_1, \dots, \beta_{(p-1)})^t$, está relacionada com o coeficiente de determinação R^2 da seguinte forma:

$$F = ((n-p) / (p-1))(R^2 / (1-R^2)). \quad (3.7)$$

Quando o método L_1 é adotado, RSEA é uma estatística natural para o teste da hipótese (3.6), sendo que valores grandes indicam a rejeição de H_0 . Contudo isto deve ser medido numa escala apropriada. A padronização adequada que proporciona a estatística do teste é:

$$D_{p-1} = RSEA / (\hat{\tau}/2), \quad (3.8)$$

onde $\hat{\tau}$ é um estimador consistente de τ .

Schrader e McKean (1982) mostraram, sob as suposições apresentadas no Capítulo 2, que D_{p-1} tem distribuição assintótica χ^2_{p-1} sob H_0 . Portanto a hipótese H_0 será rejeitada se D_{p-1} exceder um valor crítico da distribuição qui-quadrado com $p-1$ graus de liberdade.

A relação entre R^2 e a estatística F (3.7) sugere que um novo coeficiente de determinação R_2 , para regressão L_1 , pode ser definido a partir da estatística D_{p-1} considerando-se a relação $(n-p)R_2 / (1-R_2) = D_{p-1}$. Utilizando a igualdade (3.8) nesta relação vem:

$$R_2 = \text{RSEA} / (\text{RSEA} + (n-p)(\hat{\tau} / 2)) \quad (3.9)$$

Com relação às seis propriedades mencionadas anteriormente, McKean e Sievers mostraram que R_2 satisfaz às propriedades (P1), (P2), (P3), (P4) e (P6). A propriedade (P5) só é satisfeita se a estimativa de τ diminui à medida que novas variáveis são acrescentadas no modelo. Contudo isto não ocorre necessariamente, quando o estimador de τ dado em (2.4) é utilizado, conforme pode ser observado na Tabela 3.4 que apresenta os valores de R_2 de todos os modelos possíveis para os dados de imóveis.

Tabela 3.4. Valores de R_2 e do Erro Absoluto Médio (EAM) para o exemplo dos imóveis.

Variáveis no Modelo	$\hat{\tau}$	R_2	EAM
Nenhuma	10,9629	0,0000	8,3852
x_1	4,0079	0,7105	3,8307
x_2	8,0285	0,3830	6,0779
x_3	9,8973	0,4681	4,3519
x_4	11,8750	0,1164	7,6611
x_1, x_2	3,9590	0,7215	3,8257
x_1, x_3	5,4301	0,6860	3,1136
x_1, x_4	3,9636	0,7212	3,8273
x_2, x_3	8,0913	0,5470	4,0428
x_2, x_4	6,5918	0,4576	5,9135
x_3, x_4	6,6577	0,6098	3,7611
x_1, x_2, x_3	6,1331	0,6701	3,0781
x_1, x_2, x_4	7,3275	0,5937	3,8243
x_1, x_3, x_4	5,4875	0,7011	2,9029
x_2, x_3, x_4	4,0022	0,7410	3,5082
x_1, x_2, x_3, x_4	4,0246	0,7704	2,8817

Por exemplo, $\hat{\tau}$ no modelo com a variável x_1 (4,0079) é menor que $\hat{\tau}$ do modelo com x_1 e x_3 (5,4301), fazendo com que o R_2 do modelo com x_1 (0,7105) seja maior que o R_2 com x_1 e x_3 (0,6860). Entretanto a redução na SEA é grande quando se passa do modelo com x_1 (103,4292) para o modelo com x_1 e x_3 (84,0668), sugerindo que x_3 é uma variável que deve ser selecionada.

Portanto, devido à instabilidade das estimativas de τ , pode ocorrer que, mesmo sendo introduzida uma variável que traga contribuição ao modelo, a estimativa de τ aumente quando se passa do modelo reduzido para o completo, diminuindo o valor de R_2 , ocasionando então que o modelo com essa variável não seja selecionado.

Sabe-se (Elian et al. (1997)) que, quando os erros seguem uma distribuição de Laplace, o estimador de máxima verossimilhança de τ é o Erro Absoluto Médio (EAM), dado por:

$$EAM = SEA/n. \quad (3.10)$$

Apesar das condições de regularidade usuais não estarem satisfeitas (Cox e Hinkley (1991)), o EAM é um estimador consistente de τ (Bain e Engelhardt (1973)). Os valores de EAM para o exemplo dos imóveis se encontram na Tabela 3.4.

É possível, então, definir um coeficiente de determinação R_3 a partir de (3.9) para o caso da distribuição dos erros ser Laplace, substituindo $\hat{\tau}$ por EAM, obtendo-se:

$$R_3 = RSEA / (RSEA + (n-p)(EAM / 2)). \quad (3.11)$$

Os valores de R_3 obtidos são apresentados na Tabela 3.5.

Gráfico 3.2. Valores de R_3 em função do número de parâmetros do modelo.

O Gráfico 3.2 apresenta os valores de R_3 para os possíveis modelos no exemplo em questão, como função de p . A intenção ao usar o R_3 como critério é detectar qual modelo de cada tamanho produz o maior valor de R_3 , e o ponto a partir do qual a adição de mais variáveis faz com que o R_3 aumente muito pouco. A partir do gráfico nota-se que, quando se passa do modelo

com x_1 e x_3 para o modelo com x_1 , x_3 e x_4 , o R_3 aumenta de 0,7921 para 0,8160, ou seja, a inclusão de x_4 no modelo contribuiu pouco para a explicação da variabilidade da resposta. Além disso, o R_3 do modelo com x_1 , x_3 e x_4 está muito próximo do R_3 do modelo completo (0,8242). Portanto, usando o R_3 como critério de escolha, o melhor modelo selecionado é o com as variáveis x_1 e x_3 .

3.2.4. Critério do Coeficiente de Determinação Ajustado

A utilização do coeficiente de determinação R^2 para selecionar o melhor modelo na regressão de mínimos quadrados é feita buscando-se estabelecer, para um dado valor de p , qual o modelo com maior R^2 e, a partir de qual valor de p , a introdução de novas variáveis no modelo proporciona pouca contribuição no aumento do valor desse coeficiente. Uma forma alternativa de seleção de modelos seria a utilização do coeficiente de determinação ajustado $\bar{R}_p^2$ (ver, por exemplo, Montgomery e Peck (1992)), dado por:

$$\bar{R}_p^2 = 1 - (1 - R_p^2) \left(\frac{n}{n - p} \right), \quad (3.12)$$

onde R_p^2 é o coeficiente dado em (3.3), em um modelo com $p-1$ variáveis explicativas. Deve ser notado que, na expressão (3.12) é considerado o número de variáveis independentes no modelo através do número de graus de liberdade do resíduo $n-p$. Além disso, a estatística $\bar{R}_p^2$ não aumenta necessariamente quando novas variáveis regressoras são adicionadas ao modelo.

Na verdade, sabe-se que, se q novas variáveis são consideradas, se $\bar{R}_{p+q}^2$ é menor que $\bar{R}_p^2$, então a estatística F parcial para testar a contribuição adicional dessas variáveis é menor que 1 (Seber (1977)). Sendo assim, pode ser feita uma comparação direta e o modelo com o maior valor de $\bar{R}_p^2$ é escolhido.

André et al. (1998) propuseram um coeficiente de determinação que tem na regressão L_1 papel análogo ao $\bar{R}_p^2$ na regressão de mínimos quadrados. O coeficiente proposto é dado por:

$$\overline{R}_3^p = 1 - (1 - R_3^p) \left(\frac{n}{n - p} \right), \quad (3.13)$$

onde R_3^p é o coeficiente de determinação apresentado em (3.11).

Os autores mostraram que:

$$D_q > q \Rightarrow \overline{R}_3^{p+q} > \overline{R}_3^p, \quad (3.14)$$

ou seja,

$$\overline{R}_3^{p+q} \leq \overline{R}_3^p \Rightarrow D_q \leq q, \quad (3.15)$$

onde D_q é a estatística para testar a contribuição conjunta de q variáveis regressoras na explicação da variável resposta, dada em (2.10). Portanto, se com a introdução de q variáveis explicativas no modelo for observado um decréscimo no valor do coeficiente de determinação ajustado, então essas variáveis não trazem contribuição adicional na explicação da variável resposta pois, aos níveis de significância usuais, o valor crítico da distribuição χ^2_q é maior do que q .

Na Tabela 3.5 são apresentados os valores de $\overline{R}_3^p$ para os dados de imóveis. No Gráfico 3.3 estão representados os valores de $\overline{R}_3^p$ para cada modelo.

Gráfico 3.3. Valores de $\bar{R}_3^p$ em função do número de parâmetros do modelo.

A partir do Gráfico 3.3 é possível observar que o valor de $\bar{R}_3^p$ do modelo com x_1 , x_3 e x_4 (0,7840) está muito próximo do valor de $\bar{R}_3^p$ do modelo completo (0,7842), além de ser o maior valor entre todos os modelos, excetuando-se o completo. Além disso, deve ser notado que os valores de $\bar{R}_3^p$ dos modelos com x_1 e x_2 (0,6944) e com x_1 , x_2 e x_4 (0,6911) são menores que o valor de $\bar{R}_3^p$ do modelo com x_1 (0,6973), ou seja, de acordo com a expressão (3.15), as variáveis x_2 e x_4 não trazem contribuição adicional na explicação da variável resposta.

Portanto, segundo o critério do coeficiente de determinação ajustado, o melhor modelo é com as variáveis x_1 , x_3 e x_4 . Vale ressaltar que, os valores de $\bar{R}_3^p$ dos modelos com x_1 e x_3 (0,7661) e com x_1 , x_2 e x_3 (0,7664) estão muito próximos do valor obtido para o modelo com x_1 , x_3 e x_4 .

3.2.5. Critério da Soma dos Erros Absolutos Preditivos

O critério da Soma dos Erros Absolutos Preditivos (SEAP) avalia a qualidade preditiva do modelo, através do seguinte procedimento:

- Retira-se a i -ésima observação e se ajusta o modelo com as $n-1$ restantes.
- Ajusta-se o valor da variável resposta para a i -ésima observação ($\hat{y}_{(i)}$) usando este modelo.
- A diferença entre os valores observado (y_i) e ajustado ($\hat{y}_{(i)}$) da variável resposta para a i -ésima observação é calculada.
- Repete-se esta operação para $i=1, \dots, n$ e se obtém a SEAP calculando-se:

$$SEAP = \sum_{i=1}^n |y_i - \hat{y}_{(i)}|. \quad (3.16)$$

Modelos com SEAP pequena são candidatos a bons e, além disso, a SEAP não diminui necessariamente com a introdução de novas variáveis no modelo, permitindo assim uma comparação direta entre modelos com diferentes números de variáveis. Os valores da SEAP para os dados dos imóveis estão na Tabela 3.5, e são representados graficamente, como uma função do número de variáveis preditoras no modelo no Gráfico 3.4.

Gráfico 3.4. Valores da SEAP em função do número de parâmetros no modelo.

A partir do Gráfico 3.4, nota-se que o menor valor da SEAP foi produzido pelo modelo com as variáveis x_2 , x_3 e x_4 (105,4403), configurando-se então este como o melhor modelo. O modelo com as variáveis x_1 , x_3 e x_4 produziu uma SEAP bem próxima do mínimo (107,0446), sendo também um candidato a bom modelo.

Um critério análogo na regressão de mínimos quadrados consiste no cálculo da expressão:

$$\text{PRESS} = \sum_{i=1}^n (y_i - \hat{y}_{(i)})^2. \quad (3.17)$$

Quando este método de estimação é utilizado, o trabalho computacional para a escolha do modelo com menor PRESS é reduzido pois sabe-se que:

$$\text{PRESS} = \sum_{i=1}^n \left(\frac{e_i}{1 - h_{ii}} \right)^2, \quad (3.18)$$

onde e_i representa o i -ésimo resíduo e h_{ii} o i -ésimo elemento da diagonal da matriz $\hat{H} = X(X'X)^{-1}X'y$. Portanto, não é necessário ajustar-se o modelo cada vez que uma das n observações é retirada.

Não é conhecido um resultado semelhante para a regressão L_1 , talvez devido à particular forma da matriz $\hat{H}$ quando este método é usado (ver Narula e Wellington (1985)).

Portanto, quando o método L_1 é utilizado, o custo computacional exigido para o cálculo da SEAP de um dado modelo é alto, pois devem ser ajustadas n regressões. Se o número de variáveis e/ou o número de observações é muito grande, a utilização deste critério pode se tornar impraticável.

A sugestão é que este critério de escolha seja usado como um desempate entre modelos selecionados por outros procedimentos.

Tabela 3.5. Critérios de seleção para todas as possíveis regressões.

Variáveis no Modelo	SEA	R_3	$\bar{R}_3^p$	SEAP
Nenhuma	226,4000	0,0000	0,0000	233,4000
x_1	103,4292	0,7197	0,6973	114,5606
x_2	164,1021	0,4505	0,4066	193,5543
x_3	117,5025	0,6669	0,6402	134,9930
x_4	206,8500	0,1695	0,1031	217,8521
x_1, x_2	103,2933	0,7284	0,6944	128,7110
x_1, x_3	84,0668	0,7921	0,7661	121,7719
x_1, x_4	103,3377	0,7282	0,6942	132,2737
x_2, x_3	109,1555	0,7073	0,6707	161,0147
x_2, x_4	159,6650	0,4846	0,4202	198,7830
x_3, x_4	101,5501	0,7345	0,7013	133,3945
x_1, x_2, x_3	83,1100	0,8019	0,7674	133,6073
x_1, x_2, x_4	103,2552	0,7368	0,6911	152,3043
x_1, x_3, x_4	78,3789	0,8160	0,7840	107,0446
x_2, x_3, x_4	94,7217	0,7655	0,7247	105,4403
x_1, x_2, x_3, x_4	77,8051	0,8242	0,7842	118,9192

Vale lembrar que os valores obtidos nas Tabelas 3.4 e 3.5 foram calculados por intermédio do programa desenvolvido no S-PLUS apresentado no Apêndice A.

3.3. Procedimentos Automáticos de Seleção de Variáveis

Quando o número de variáveis explicativas envolvidas é moderado ou grande, torna-se impraticável o estudo de todas as possíveis regressões. Para atacar esse problema foram desenvolvidos procedimentos automáticos de seleção de variáveis que, embora possam ser utilizados em modelos com poucas variáveis, são mais recomendados para problemas nos quais o número de possíveis modelos a serem analisados é grande.

Na regressão de mínimos quadrados, o procedimento automático de busca *stepwise* é o mais usado (ver por exemplo Neter et al. (1996)). Essencialmente, este método desenvolve uma sequência de modelos de regressão, adicionando ou removendo a cada passo uma variável explicativa, com base na contribuição adicional dessa variável na explicação da variável resposta, dadas as outras variáveis que compõem o modelo.

Procedimentos tipo *stepwise* para a regressão L_1 foram formalmente propostos por André et al. (1996). Assim como para mínimos quadrados, uma variável de cada vez é incluída ou excluída do modelo de acordo com a sua contribuição para a explicação da variável resposta, dado que outras variáveis foram incluídas no modelo. A contribuição de uma variável é avaliada através da redução que ocorre na soma dos erros absolutos quando ela é introduzida no modelo.

Em outras palavras, para determinar a contribuição da variável x_j , dado que as variáveis $x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_p$ já estão no modelo, as somas dos erros absolutos correspondentes aos seguintes modelos são calculadas:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_{(p-1)} x_{(p-1)i} + \varepsilon_i \quad (3.19)$$

e

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_{(j-1)} x_{(j-1)i} + \beta_{j+1} x_{(j+1)i} + \dots + \beta_{(p-1)} x_{(p-1)i} + \varepsilon_i \quad (3.20)$$

Sejam SEA e $SEA_{(j)}$ as somas dos erros absolutos para o modelo (3.19) com $p-1$ variáveis explicativas, e para o (3.20), sem a variável x_j , respectivamente. A redução na soma dos erros

absolutos é dada por: $SEA_{(j)} - SEA$. O teste da hipótese $H_0: \beta_j = 0$, é feito considerando-se a hipótese proposta por McKean e Schrader (1987):

$$D_1 = (SEA_{(j)} - SEA) / (\hat{\tau}/2), \quad (3.21)$$

onde $\hat{\tau}$ é o estimador consistente de τ dado em (2.4). Para n grande, sob H_0 , D_1 tem distribuição χ^2_1 .

Os procedimentos tipo *stepwise* são classificados em três categorias: seleção *forward*, eliminação *backward* e *stepwise*, que combina os dois primeiros.

3.3.1. Seleção *Forward*

O procedimento começa com o modelo $y_i = \beta_0 + \varepsilon_i$, ou seja, o modelo ajustado é $\hat{y}_i = \text{mediana}(y)$, $i=1, \dots, n$. A primeira variável a entrar no modelo é aquela para a qual for observado o valor máximo de D_1 em (3.21), se sua contribuição for estatisticamente significativa, a um nível de significância α . Considerando o modelo com uma variável, a próxima variável a entrar no modelo é aquela que apresentar o maior valor da estatística do teste, se essa estatística for significativa a um nível α . O processo é repetido até nenhuma variável ser incluída no modelo. Pode-se resumir o procedimento todo da seguinte forma:

- Passo 1F: Calcular a soma dos erros absolutos para o modelo ajustado: $\hat{y}_i = \text{mediana}(y_1, \dots, y_n)$.
- Passo 2F: Calcular a estatística D_1 em (3.21) para cada variável. Selecionar aquela com o maior valor de D_1 e testar se é significativa a um nível α . Se não for significativa, parar; caso contrário, incluir esta variável no modelo e ir para o Passo 3F.
- Passo 3F: Repetir o Passo 2F até nenhuma nova variável entrar no modelo.

Os resultados da aplicação da seleção *forward* nos dados de imóveis, com auxílio de um programa desenvolvido no aplicativo S-PLUS (ver Apêndice B), se encontram na Tabela 3.6.

Tabela 3.6. Resultados da seleção *forward*.

α	Modelo Selecionado
0,01	x_1, x_3
0,05	x_1, x_3
0,10	x_1, x_3

A partir da Tabela 3.6, nota-se que para todos os níveis de significância considerados, o modelo indicado como melhor pela seleção *forward* é aquele com as variáveis x_1 e x_3 .

3.3.2. Eliminação *Backward*

Este procedimento começa com todas as variáveis no modelo. A estatística D_1 para cada variável no modelo é calculada, como se ela fosse a última a entrar no modelo, isto é, a SEA com todas as variáveis no modelo é calculada, bem como a SEA do modelo sem a variável que está sendo investigada. O menor valor de D_1 é testado a um nível α de significância. Se não é significativo, a variável é retirada. Começando com este modelo, o processo é repetido até nenhuma variável ser retirada. O processo pode ser sintetizado como segue:

- Passo 1B: Ajustar o modelo com todas as variáveis e calcular sua soma de erros absolutos.
- Passo 2B: Calcular o valor de D_1 em (3.21) para cada variável, como se ela fosse a última a entrar no modelo. Selecionar a variável correspondente ao menor valor de D_1 e testar sua significância a um nível α . Se for significativo, parar; caso contrário, eliminar a variável e ir para o Passo 3B.
- Passo 3B: Repetir o Passo 2B até nenhuma variável ser eliminada.

Os resultados da aplicação da eliminação *backward* nos dados de imóveis, com auxílio de um programa desenvolvido no aplicativo S-PLUS (ver Apêndice B), se encontram na Tabela 3.7.

Tabela 3.7. Resultados da eliminação *backward*.

α	Modelo Selecionado
0,01	x_1, x_3
0,05	x_1, x_3
0,10	x_1, x_3

A partir da Tabela 3.7, nota-se que para todos os níveis de significância considerados, o modelo indicado como melhor pela eliminação *backward*, assim como na seleção *forward* é aquele com as variáveis x_1 e x_3 .

3.3.3. Stepwise

Este método consiste em combinar a seleção *forward* e a eliminação *backward*. O processo começa da mesma maneira que o *forward*. Se três variáveis entram no modelo, a contribuição de cada variável que já está no modelo é testada na presença da última a ter entrado. Isto é feito calculando a estatística D_1 para cada variável como se ela fosse a última a ter entrado e testando sua significância a um nível α ; se o valor de D_1 não é significativo, a variável é eliminada. Esta eliminação pode acontecer porque a inclusão de uma variável no modelo pode tornar a contribuição de uma outra variável já selecionada não-significante. O processo é repetido até nenhuma variável ser incluída ou excluída do modelo. No *stepwise* podem ser usados dois níveis de significância: um de entrada (α) e outro de saída (β). Entretanto, é recomendado que seja adotado o mesmo valor para α e β , pois dependendo dos valores escolhidos o processo pode entrar em *looping*. Pode-se resumir o processo da seguinte maneira:

- Passo 1S (*Backward*): Supondo que existam ($q \geq 3$) variáveis no modelo e x_q foi a última selecionada. Calcular D_1 em (3.21) para cada variável $x_1, \dots, x_{q-1}$, como se tivesse sido a última a ter entrado no modelo. Selecionar a variável correspondente ao menor valor de D_1 e testar sua significância a um nível α . Se for significativa, ir para o Passo 2S; senão, excluir a variável e repetir este passo com o modelo reduzido até uma variável não poder ser excluída, e ir para o Passo 2S.

- Passo 2S (*Forward*): Começando com o modelo no Passo 1S, calcular o valor de D_1 em (3.21) para cada variável que não está no modelo. Selecionar a variável correspondente ao maior valor de D_1 e testar sua significância a um nível α . Se não for significativa, parar; caso contrário, incluir a variável e ir para o Passo 1S.

Os resultados da aplicação do *stepwise* nos dados dos imóveis foram os mesmos da seleção *forward* pois, como só entraram duas variáveis no modelo (x_1 e x_3), não foi possível executar o Passo 1S do *stepwise*.

Deve ser notado que, como na regressão de mínimos quadrados, cada um dos procedimentos descritos nesta seção pode conduzir a um modelo diferente.

Como pode ser observado, os procedimentos automáticos, ao contrário do de todas as possíveis regressões, identificam um único modelo como sendo o “melhor”. Esta característica dos procedimentos automáticos consiste no seu maior defeito, pois às vezes estes procedimentos podem indicar um modelo de regressão “pobre” como sendo o “melhor”. Além disso, a identificação de um único modelo pode esconder a existência de outros bons modelos.

Mediante este problema, uma boa estratégia é adotar o modelo escolhido pelo procedimento automático como um ponto de partida na busca de outros subconjuntos de variáveis. Por exemplo, pode-se considerar o número de variáveis explicativas do modelo indicado pelo procedimento automático como sendo um “tamanho ideal” e utilizar o procedimento de todas as possíveis regressões para subconjuntos deste tamanho ou de tamanhos próximos ao selecionado.

3.4. Conclusões

O procedimento de todas as possíveis regressões identifica alguns subconjuntos que são “bons” segundo algum critério específico. No exemplo considerado, o critério da SEA indicou o modelo com as variáveis x_1 e x_3 ; o critério de Narula e Wellington selecionou os seguintes subconjuntos: x_1 para os modelos com uma variável, x_1 e x_3 para os com duas variáveis, e x_1 , x_3 e x_4 para os com três variáveis. O uso do Coeficiente de Determinação resultou na escolha do

modelo com as variáveis x_1 e x_3 , enquanto que o Coeficiente de Determinação Ajustado selecionou o modelo com x_1 , x_3 e x_4 .

Finalmente, o critério da SEAP apontou o modelo com x_2 , x_3 e x_4 . Os procedimentos automáticos seleção *forward* e eliminação *backward* selecionaram o modelo com x_1 e x_3 .

Portanto, os modelos selecionados segundo os diferentes critérios não foram, em geral, os mesmos. Na verdade, pode-se garantir que apenas os critérios do R_3 e da SEA apontam para o mesmo modelo, devido à equivalência entre os métodos. Na escolha entre eles deve-se considerar o conhecimento do pesquisador na área de aplicação do problema em estudo. Deve-se também lembrar que cada um dos critérios apresentados, quando são consideradas todas as possíveis regressões, indicam um melhor modelo, mas modelos “quase igualmente bons” também podem ser selecionados.

Capítulo 4

Um Estudo Preliminar Sobre os Efeitos da Multicolinearidade

4.1. Introdução

A presença de multicolinearidade é um fato bastante frequente nos modelos de regressão linear múltipla. Quando o método de estimação utilizado é o de mínimos quadrados, o assunto está bem desenvolvido (ver por exemplo Montgomery e Peck (1992) ou Neter et al. (1996)), tendo sido apresentadas técnicas para detectar a multicolinearidade, estudadas suas consequências e propostos procedimentos para solucionar esse problema.

Não é de nosso conhecimento a existência de estudos sobre o efeito da multicolinearidade quando a estimação é feita através do método L_1 . Como uma das possíveis correções para este problema é a redução no número de variáveis independentes que irão compor o modelo final, o problema da seleção de variáveis pode estar relacionado com o da multicolinearidade. Por esse motivo esse problema será abordado neste capítulo um estudo preliminar sobre os efeitos da multicolinearidade nas estimativas L_1 .

Este estudo foi motivado pelo fato de que, devido à forma como são obtidas, as estimativas L_1 não são alteradas, se os valores observados das variáveis preditoras ou da variável resposta para as observações não-definidoras forem modificados de forma a permanecerem dentro de determinados intervalos (Narula e Wellington (1998)). Portanto, ao contrário do que acontece na regressão de mínimos quadrados na presença de multicolinearidade, não são observadas mudanças nas estimativas quando uma observação é modificada. Serão investigados,

através de simulação, a existência de alguns efeitos da multicolinearidade nas estimativas L_1 já conhecidos na regressão de mínimos quadrados; a saber: o aumento no erro padrão dos estimadores, a inversão do sinal algébrico das estimativas dos coeficientes e o efeito da eliminação de uma observação nas estimativas dos parâmetros.

4.2. Descrição do Estudo

No estudo a ser apresentado, foi considerado o modelo de regressão linear múltipla sem intercepto com duas variáveis independentes, x_1 e x_2 , dado por:

$$\begin{aligned} y_i &= \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \\ i &= 1, \dots, 15. \end{aligned} \quad (4.1)$$

Foram inicialmente gerados os valores x_{1i} a partir de uma distribuição $N(0,1)$. Geraram-se, então, distúrbios v_i a partir de uma distribuição Uniforme. Os valores de x_{1i} e v_i foram transformados e rotacionados, via Componentes Principais, de forma a satisfazer às seguintes condições:

$$\begin{aligned} \sum_{i=1}^{15} x_{1i} &= \sum_{i=1}^{15} v_i = 0, \\ \sum_{i=1}^{15} x_{1i}^2 &= 1, \\ \sum_{i=1}^{15} x_{1i} v_i &= 0. \end{aligned} \quad (4.2)$$

Os valores x_{2i} foram, então, calculados a partir da relação:

$$x_{2i} = \alpha x_{1i} + v_i, \quad (4.3)$$

onde α é uma constante fixada. A transformação feita nos x_{1i} e v_i garantiu também que

$$\sum_{i=1}^{15} x_{2i} = 0.$$

Desta forma tem-se que o coeficiente de correlação entre x_1 e x_2 é dado por:

$$r(x_1, x_2) = \frac{\sum_{i=1}^{15} x_{1i} x_{2i}}{\sqrt{x_{1i}^2} \sqrt{x_{2i}^2}}. \quad (4.4)$$

Mas, $\sum_{i=1}^{15} x_{1i} x_{2i} = \sum_{i=1}^{15} x_{1i} (\alpha x_{1i} + v_i) = \alpha$ e, assim:

$$r(x_1, x_2) = \frac{\alpha}{\sqrt{x_{2i}^2}} = \frac{\alpha}{\sqrt{\alpha^2 + \sum_{i=1}^{15} v_i^2}}. \quad (4.5)$$

Portanto, para gerarem-se valores de x_1 e x_2 com correlação próxima de 1 (ou -1), os distúrbios v_i podem ser gerados a partir de uma distribuição Uniforme com variância $\sum_{i=1}^{15} v_i^2$ próxima de zero.

Obtidos os valores de x_1 e x_2 , as correspondentes observações da variável resposta podem ser calculadas fixando-se os valores para β_1 e β_2 , e gerando-se os erros e_i a partir de uma mesma função densidade de probabilidade com média zero.

Fixando $\beta_1 = 1$ e $\beta_2 = 1$, com auxílio do MINITAB e de um programa desenvolvido no S-PLUS (ver Apêndice C), foram obtidos 4 conjuntos de observações nas seguintes condições:

- Conjunto 1: Fixou-se $\alpha = 1,5$, v_i foram gerados a partir de uma $U(-0,25;0,25)$, e e_i a partir de uma $N(0,1)$. Observou-se neste conjunto de dados $r(x_1, x_2) = 0,94$.
- Conjunto 2: Fixou-se $\alpha = 0,5$, v_i foram gerados a partir de uma $U(-100,100)$, e e_i a partir de uma $N(0,1)$. Observou-se neste conjunto de dados $r(x_1, x_2) = 0,14$.

- Conjunto 3: foi gerado de forma análoga ao Conjunto 1, a menos que os erros e_i foram obtidos a partir de uma distribuição de Laplace com média 0 e parâmetro de escala 1. Observou-se $r(x_1, x_2) = 0,93$.

- Conjunto 4: foi gerado de forma análoga ao Conjunto 2, a menos que os erros e_i foram obtidos a partir de uma distribuição de Laplace com média 0 e parâmetro de escala 1. Observou-se $r(x_1, x_2) = 0,13$.

4.3. Resultados Obtidos

A cada conjunto de dados o modelo foi ajustado pelos métodos L_1 e de mínimos quadrados. Os resultados obtidos, através do programa apresentado no Apêndice C, encontram-se resumidos nas Tabelas 4.1 e 4.2.

Tabela 4.1. Resultados obtidos no ajuste do modelo pelo método L_1 .

	Distribuição dos Erros	Parâmetros	Estimativas	Erros Padrão
Conjunto 1 $r(x_1, x_2) = 0,94$	Normal(0,1)	β_1	6,4149	3,1486
		β_2	-0,9182	1,9800
Conjunto 2 $r(x_1, x_2) = 0,14$	Normal(0,1)	β_1	-0,2809	1,3709
		β_2	0,4056	0,3801
Conjunto 3 $r(x_1, x_2) = 0,93$	Laplace(0,1)	β_1	0,9079	4,1818
		β_2	1,7749	2,5910
Conjunto 4 $r(x_1, x_2) = 0,13$	Laplace(0,1)	β_1	2,4599	1,5181
		β_2	0,6845	0,4037

Tabela 4.2. Resultados obtidos no ajuste do modelo pelo método de mínimos quadrados.

	Distribuição dos Erros	Parâmetros	Estimativas	Erros Padrão
Conjunto 1 $r(x_1, x_2) = 0,94$	Normal(0,1)	β_1	5,1498	3,0580
		β_2	-0,6769	1,9231
Conjunto 2 $r(x_1, x_2) = 0,14$	Normal(0,1)	β_1	0,6581	1,0537
		β_2	0,5679	0,2921
Conjunto 3 $r(x_1, x_2) = 0,93$	Laplace(0,1)	β_1	6,6687	4,1480
		β_2	-2,3589	2,5701
Conjunto 4 $r(x_1, x_2) = 0,13$	Laplace(0,1)	β_1	0,6684	1,6420
		β_2	0,8580	0,4367

Para erros com distribuição Normal e Laplace, pode ser observado que os erros padrão das estimativas L_1 aumentam quando cresce a correlação entre x_1 e x_2 . Além disso, a inflação observada é da mesma ordem que a obtida quando o modelo é ajustado por mínimos quadrados. Isto indica que a presença de multicolinearidade inflaciona os erros padrão dos estimadores L_1 , da mesma forma que no método de mínimos quadrados.

Um outro possível efeito da multicolinearidade nas estimativas de mínimos quadrados é observar-se sinal do coeficiente ajustado oposto ao esperado teoricamente. Quando os erros têm distribuição Normal, os dois métodos sofreram esse efeito. Entretanto, quando a distribuição dos erros é Laplace, tal efeito foi notado apenas no método de mínimos quadrados.

Com o objetivo de verificar se as estimativas dos parâmetros se alteram quando uma observação é retirada, cada uma das 15 observações foi excluída e o modelo foi reajustado pelo método L_1 através do programa que se encontra no Apêndice C, para os 4 conjuntos de dados que foram gerados. Os resultados são apresentados na Tabela 4.3.

Tabela 4.3. Resultados das estimativas pelo método L_1 com a retirada de observações.

observação retirada	Conjunto 1		Conjunto 2		Conjunto 3		Conjunto 4	
	obs. definidoras: 8 e 14		obs. definidoras: 4 e 14		obs. Definidoras: 9 e 12		obs. definidoras: 10 e 11	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
1	5,7205	-1,1677	-0,2809	0,4056	0,9079	1,7749	2,4599	0,6845
2	5,7205	-1,1677	0,4526	0,7393	0,9079	1,7749	2,4599	0,6845
3	6,4149	-0,9182	-0,5731	0,4271	0,9079	1,7749	2,4599	0,6845
4	6,4149	-0,9182	-0,5731	0,4271	0,6596	2,1727	2,4599	0,6845
5	7,8677	-2,2278	-0,5731	0,4271	0,6596	2,1727	2,4599	0,6845
6	5,7205	-1,1677	-0,2809	0,4056	0,6596	2,1727	0,4857	0,5704
7	6,4149	-0,9182	-0,2809	0,4056	0,9079	1,7749	2,4599	0,6845
8	7,8677	-2,2278	-0,5731	0,4271	0,9079	1,7749	2,4599	0,6845
9	5,7205	-1,1677	-0,2809	0,4056	0,6596	2,1727	2,4599	0,6845
10	6,4149	-0,9182	-0,5731	0,4271	1,8141	0,3237	1,5157	0,9420
11	6,4149	-0,9182	-0,2809	0,4056	0,6596	2,1727	0,4857	0,5704
12	6,4149	-0,9182	-0,2809	0,4056	12,0303	-4,3809	2,4599	0,6845
13	5,7205	-1,1677	-0,2809	0,4056	0,9079	1,7749	2,4599	0,6845
14	5,7205	-1,1677	-0,4663	0,3213	0,9079	1,7749	2,4599	0,6845
15	6,4149	-0,9182	-0,2809	0,4056	1,8141	0,3237	2,4599	0,6845

Observa-se, como já era esperado devido à forma peculiar como as estimativas L_1 são obtidas, que os coeficientes podem permanecer inalterados após a retirada de uma observação.

4.4. Conclusões

Os resultados obtidos na seção anterior sugerem que diagnósticos informais de multicolinearidade, utilizados na regressão de mínimos quadrados, tais como: inversão de sinal dos coeficientes ajustados em relação ao que é esperado teoricamente, e mudança nas estimativas dos coeficientes quando uma observação é eliminada ou alterada não podem ser adotados, quando o modelo é ajustado pelo método L_1 . Além disso, neste pequeno estudo foi observado que a magnitude da inflação dos erros padrão dos estimadores, devido à presença de multicolinearidade, foi aproximadamente da mesma ordem para os dois métodos de estimação.

Capítulo 5

Aplicação

Neste capítulo é feita uma aplicação dos procedimentos de seleção de variáveis discutidos ao longo deste trabalho. O conjunto de dados que será explorado foi retirado de Grácio (1991) e consiste em algumas características demográficas de 49 países. Os dados apresentados na Tabela 5.1 representam as seguintes características: mortalidade infantil por 100 crianças nascidas (x_1), número de habitantes para cada médico (x_2), população por km^2 (x_3), população por 1000 hectares de agricultura no país (x_4), porcentagem de alfabetizados maiores de 15 anos (x_5), estudantes matriculados no nível superior para cada 100000 habitantes (x_6), e a variável resposta, produto interno bruto de 1957 calculado em dólares (y).

Com a finalidade de garantir que a distribuição da variável resposta seja Laplace, Grácio (1991) sugeriu a seguinte transformação de Box e Cox na variável resposta: $y^* = y^{(-0,5)}$. O resultado desta transformação foi considerado satisfatório, ou seja, a variável transformada y^* possui distribuição de Laplace.

O Gráfico 5.1 ilustra o comportamento de cada uma das variáveis explicativas em relação à variável resposta, bem como as possíveis associações entre as variáveis explicativas. A partir deste gráfico, é possível notar que os países que apresentam valores altos nas variáveis x_1 e x_2 possuem valores altos na variável resposta transformada y^* e que países que apresentam valores baixos nas variáveis x_5 e x_6 apresentam valores altos na variável resposta transformada y^* . Além disso, valores altos de x_2 estão associados a valores altos de x_1 e a valores baixos de x_5 .

Portanto, através de uma análise gráfica, é possível concluir que as variáveis x_1 , x_2 , x_5 e x_6 são as que possuem maior potencial para explicar a variável y^* .

Tabela 5.1. Dados relativos à características demográficas de 49 países.

País	y	x ₁	x ₂	x ₃	x ₄	x ₅	x ₆
Alemanha Ocidental	927	33,8	798	217	3631	98,5	528
Austrália	1316	19,5	860	1	21	98,5	856
Áustria	670	37,5	695	84	1720	98,5	546
Barbados	200	60,4	3000	548	7121	91,1	24
Bélgica	1196	35,4	819	301	5257	96,7	536
Bulgária	365	45,1	740	72	1380	85,0	456
Canadá	1947	27,3	900	2	257	97,5	645
Chile	379	127,9	1700	11	1164	80,1	257
Chipre	467	29,9	1400	62	1042	60,5	78
Costa Rica	357	78,9	2600	24	948	79,4	326
Dinamarca	1057	23,7	830	107	1434	98,5	570
El Salvador	219	76,3	5400	127	1497	39,4	89
Espanha	293	43,5	1000	61	1347	87,0	258
Estados Unidos	2577	26,4	780	20	399	98,0	1983
Finlândia	794	21,0	1600	13	1512	98,5	529
França	943	27,4	1014	83	1288	96,4	667
Guatemala	189	91,9	6400	36	1365	29,4	135
Guiana Britânica	235	67,1	3900	3	192	74,0	27
Holanda	836	16,5	900	346	4855	98,5	923
Hong Kong	272	41,5	3300	3092	98143	57,5	176
Hungria	490	47,6	650	108	1370	97,5	258
Índia	73	225,0	5200	138	2279	19,3	220
Irlanda	550	30,5	1000	40	598	98,5	362
Islândia	572	22,4	840	2	79	98,5	445
Itália	516	48,7	746	164	2323	87,5	362
Iugoslávia	265	100,0	1637	73	1215	77,0	524
Jamaica	316	58,7	4300	143	3410	77,0	42
Japão	306	37,7	930	254	7563	98,0	750
Luxemburgo	1388	31,5	910	123	2286	96,5	36
Malásia	356	68,9	6400	54	2980	38,4	475
Malta	377	38,3	980	1041	8050	57,6	142
Maurícios	225	69,5	4500	352	4711	51,8	14
México	262	77,7	1700	18	296	50,0	258
Nicarágua	160	71,7	2800	10	824	38,4	110
Noruega	1130	20,2	946	11	3420	98,5	258
Nova Zelândia	1310	22,8	700	9	170	98,5	839
Panamá	329	54,8	3200	15	838	67,5	371
Polônia	475	74,7	1100	96	1411	95,0	351
Porto Rico	563	52,4	2200	271	4030	81,0	1192
Portugal	224	77,5	1394	100	1087	55,9	272
Reino Unido	1189	24,1	935	217	2677	98,5	460
Romênia	360	75,7	788	78	1248	89,0	226
Singapura	400	32,3	2400	2904	108214	50,0	437
Suécia	1380	16,6	1089	17	1705	98,5	401
Suíça	1428	21,1	765	133	2320	98,5	398
Tailândia	161	30,5	1500	305	10446	54,0	329
Tchecoslováquia	680	31,0	620	108	1821	97,5	398
Trindade	423	45,4	2300	168	4383	73,8	61
URSS	600	35,0	578	10	339	95,0	539

Fonte: Grácio (1991)

Gráfico 5.1. Matriz de diagramas de dispersão.

Nesta aplicação existem seis variáveis explicativas, o que implica em $2^6 = 64$ modelos de regressão possíveis. Numa situação como essa, a utilização do procedimento de todas as possíveis regressões como primeira estratégia na seleção de algum modelo torna-se impraticável.

O mais indicado é começar pelos procedimentos automáticos de seleção de variáveis que podem indicar alguns modelos, ou apenas sugerir o número ideal de variáveis que o modelo final deve ter. Os resultados indicados na Tabela 5.2 mostram que a seleção *Forward* e a eliminação *Backward* apontam para um modelo com as variáveis explicativas x_1 , x_5 e x_6 . Portanto, os procedimentos automáticos sugeriram o mesmo modelo de três variáveis.

Tabela 5.2. Resultados dos procedimentos automáticos.

Método	Nível de Significância	Variáveis Seleccionadas
<i>Forward</i>	0,01	x_1, x_5, x_6
	0,05	x_1, x_5, x_6
	0,10	x_1, x_5, x_6
<i>Backward</i>	0,01	x_1, x_5, x_6
	0,05	x_1, x_5, x_6
	0,10	x_1, x_5, x_6

O algoritmo proposto por Narula e Wellington, que faz uso da Soma dos Erros Absolutos (SEA), também pode ser utilizado pois identifica o melhor modelo de cada tamanho dentre todas as possíveis regressões, sem a necessidade de pesquisar todos os modelos possíveis. A Tabela 5.3 apresenta os modelos seleccionados por este algoritmo, bem como os valores observados das estatísticas relacionadas aos outros critérios de seleção apresentados no Capítulo 3. Nota-se que o modelo com as variáveis x_1, x_5 e x_6 foi o melhor com três variáveis, coincidindo com os procedimentos automáticos. Mais ainda, a redução na SEA e o aumento de R_2, R_3 e $\bar{R}_3^p$ são muito pequenos quando se passa do modelo com três variáveis para o modelo com quatro variáveis. Além disso, o modelo com x_1, x_5 e x_6 é o que produz a menor Soma Absoluta dos Erros Preditivos (SEAP).

Tabela 5.3. Modelos seleccionados pelo algoritmo proposto por Narula e Wellington.

Variáveis no Modelo	$\hat{\tau}$	SEA	R_3	$\bar{R}_3^p$	SEAP
x_1	0,0193	0,4325	0,5603	0,5416	0,4940
x_1, x_5	0,0097	0,3146	0,7213	0,7031	0,3572
x_1, x_5, x_6	0,0081	0,2852	0,7586	0,7372	0,2933
x_1, x_3, x_5, x_6	0,0076	0,2822	0,7658	0,7392	0,2999
x_1, x_3, x_4, x_5, x_6	0,0075	0,2785	0,7739	0,7423	0,3181

Como os procedimentos automáticos e o critério da mínima SEA seleccionaram um modelo com três variáveis, será feita uma comparação de todos os modelos possíveis desse tamanho, considerando-se agora os critérios $R_3, \bar{R}_3^p$ e SEAP. Com esta finalidade foi construída a

Tabela 5.4, que apresenta os valores observados dessas estatísticas para os 20 modelos possíveis com três variáveis. Apenas como ilustração, são também apresentados nessa tabela os valores de $\hat{\tau}$ e SEA. Percebe-se que o modelo com as variáveis x_1 , x_5 e x_6 produz os maiores valores de R_3 e $\bar{R}_3^p$ e os menores de SEA e SEAP, configurando-se mais uma vez como o melhor. O subconjunto formado por x_1 , x_3 e x_5 pode ser considerado como o segundo melhor, dentre os relacionados pela Tabela 5.4.

Tabela 5.4. Critérios de seleção calculados para todos os modelos com três variáveis.

Variáveis no Modelo	$\hat{\tau}$	SEA	R_3	$\bar{R}_3^p$	SEAP
x_1, x_2, x_3	0,0124	0,3562	0,6755	0,6467	0,3797
x_1, x_2, x_4	0,0143	0,3641	0,6655	0,6358	0,4011
x_1, x_2, x_5	0,0105	0,3139	0,7264	0,7021	0,3868
x_1, x_2, x_6	0,0099	0,3323	0,7049	0,6787	0,3620
x_1, x_3, x_4	0,0158	0,3909	0,6302	0,5973	0,4446
x_1, x_3, x_5	0,0098	0,3125	0,7281	0,7039	0,3519
x_1, x_3, x_6	0,0126	0,3379	0,6981	0,6713	0,3729
x_1, x_4, x_5	0,0108	0,3137	0,7268	0,7025	0,3620
x_1, x_4, x_6	0,0144	0,3444	0,6902	0,6626	0,3743
x_1, x_5, x_6	0,0081	0,2852	0,7586	0,7372	0,2933
x_2, x_3, x_4	0,0195	0,5006	0,4605	0,4125	0,5566
x_2, x_3, x_5	0,0174	0,4117	0,6012	0,5658	0,4986
x_2, x_3, x_6	0,0142	0,4422	0,5562	0,5168	0,4909
x_2, x_4, x_5	0,0169	0,4120	0,6008	0,5653	0,4775
x_2, x_4, x_6	0,0140	0,4436	0,5542	0,5146	0,4968
x_2, x_5, x_6	0,0132	0,3807	0,6438	0,6122	0,4815
x_3, x_4, x_5	0,0163	0,4128	0,5996	0,5640	0,5319
x_3, x_4, x_6	0,0208	0,5608	0,3456	0,2874	0,6422
x_3, x_5, x_6	0,0155	0,3773	0,6483	0,6171	0,4736
x_4, x_5, x_6	0,0157	0,3773	0,6484	0,6171	0,4702

Após a aplicação dos critérios de seleção apresentados neste trabalho aos dados da Tabela 5.1, o melhor modelo reduzido para explicar a variável transformada y^* é com as variáveis x_1 , x_5 e x_6 . Este resultado é coerente com a análise do Gráfico 5.1, que indicou as variáveis x_1 , x_2 , x_5 e x_6 como candidatas em potencial para comporem um bom modelo.

Determinadas quais são as variáveis mais importantes na explicação do problema, a eventual inclusão de interações e termos de maior grau deve ser feita em conjunto com o

pesquisador na área do problema em estudo. Vale lembrar que a adição de um termo de interação de qualquer ordem, não deve ser feita apenas com base em resultados de testes estatísticos, uma vez que nem sempre o pesquisador pode fornecer uma interpretação adequada para a interação de duas ou mais variáveis.

Capítulo 6

Considerações Finais

Foram apresentados neste trabalho alguns procedimentos para seleção de variáveis em modelos de regressão linear múltipla, quando o ajuste é feito pelo método L_1 . Quando o número de variáveis no modelo inicial é pequeno, todas as possíveis regressões podem ser analisadas e comparadas segundo os critérios da soma dos resíduos absolutos, coeficiente de determinação, coeficiente de determinação ajustado e soma dos erros absolutos preditivos. Esses critérios não selecionam necessariamente o mesmo modelo, a menos dos critérios da SEA e do coeficiente de determinação proposto R_3 . Modelos “igualmente bons” podem ser escolhidos. Na indicação do modelo final devem ser considerados o julgamento profissional na área de interesse, e aspectos práticos e econômicos relacionados à observação das variáveis preditoras que vão compor o modelo final.

Os métodos de seleção automáticos apresentados podem ser aplicados quando o número inicial de variáveis preditoras for tal que seja impossível analisar-se todas as possíveis regressões. Os modelos selecionados por estes métodos podem servir de indicação sobre o número ideal de variáveis independentes que irão compor o modelo final. Todos os modelos desse tamanho podem, então, ser analisados e comparados segundo os critérios mencionados no parágrafo anterior.

Um estudo preliminar sobre o efeito da multicolinearidade indicou que formas de diagnósticos informais para detectar esse problema na regressão de mínimos quadrados não devem ser utilizadas quando o ajuste é feito pelo método L_1 . Entretanto, um estudo mais detalhado deste problema deve ser realizado.

A utilização na regressão L_1 de outros métodos de seleção de variáveis adotados em mínimos quadrados deve ser estudada. Em particular, o método *lasso*, proposto por Tibshirani (1996) deve ser considerado. O *lasso* minimiza a soma dos quadrados do resíduo sujeito à restrição de que a soma dos valores absolutos dos coeficientes seja menor que uma constante. Essa restrição tende a produzir algumas estimativas dos coeficientes iguais a zero. A obtenção das estimativas é feita através de programação não-linear.

No caso da regressão L_1 , esse método minimizaria a soma dos erros absolutos do resíduo, sujeito às mesmas restrições. As estimativas L_1 poderiam, então, ser obtidas por técnicas de programação linear. As propriedades desse método de seleção poderiam ser estudadas e os modelos selecionados por essa técnica comparados com os obtidos por outros critérios de seleção.

Ainda como sugestão para a continuidade do trabalho, seria interessante a comparação entre as estimativas do parâmetro de escala τ dadas em (2.4) e (3.10). Através de simulação, poderia ser determinado qual o tamanho de amostra a partir do qual as duas estimativas são próximas, considerando-se diferentes distribuições dos erros e diferentes números de variáveis explicativas no modelo.

Apêndice A

Programa de Estimação L_1

Listagem do programa L1.TXT, desenvolvido no pacote S-PLUS:

```
# PROGRAMA L1.TXT - AGOSTO/1998
#
# Este programa ajusta modelos de regressao L1, com ou sem intercepto,
# imprimindo as observacoes definidoras, valores observados, ajustados e
# residuos, estimativas dos parametros, estimativas dos erros padrao
# dos estimadores, intervalos de confianca para os parametros, testes de
# hipoteses individuais de cada parametro, teste do efeito de regressao e
# estatisticas para selecao de variaveis.
#
# Variaveis a serem definidas pelo usuario:
# intercepto: recebendo o valor 0, ajusta o modelo sem intercepto e recebendo
#             o valor 1, ajusta o modelo com intercepto.
# imprime: recebendo o valor 1, imprime os valores observados, ajustados e os
#           residuos do modelo e recebendo 0, nao imprime tais valores.
# laplace: recebendo o valor 1, imprime as estatisticas de selecao de variaveis
#           que podem ser usadas no caso da distribuicao dos dados ser de La-
#           place e recebendo 0, nao imprime tais estatisticas.
#
# Verifica se o usuario pediu modelo com ou sem intercepto
if (intercepto==0) {
  fit<-llfit(X, y, intercept=F)
  Xmat<-X
  SEAm<-sum(abs(y))
  fator<-0
}
if (intercepto==1) {
  fit<-llfit(X, y, intercept=T)
  Xmat<-cbind(1,X)
  SEAm<-sum(abs(y-median(y)))
  fator<-1
}
```

```

# Calculo da Estimativa de tao
r<-cbind(residuals(fit))
coef1<-cbind(coef(fit))
aju<-(y-r)
n<-(nrow(X)-ncol(X)-fator)
m<-((n+1)/2)-sqrt(n)
res<-cbind(sort(r))
res1<-matrix(0,(nrow(res)-nrow(coef1)),1)
def<-matrix(0,nrow(coef1),1)
j<-1
for(i in 1:nrow(r)) if (r[i]==0) ((def[j]<-i) & (j<-j+1))
j<-1
for(i in 1:nrow(res)) if (res[i]!=0) ((res1[j]<-res[i]) & (j<-j+1))
pos1<-(n-m+1)
pos2<-m
e1<-round(pos1)
e2<-round(pos2)
if (e1>n) (e1<-n)
if (e2==0) (e2<-1)
tao<-sqrt(n)*((res1[e1]-res1[e2])/4)

# Calculos dos Intervalos de Confianca, Estatisticas de Testes e
# Niveis Descritivos
SEA<-sum(abs(res))
XX<-solve(t(Xmat)%*%Xmat)
DX<-diag(XX)
IC<-matrix(0,nrow(coef1),2)
TH<-matrix(0,nrow(coef1),2)
DP<-matrix(0,nrow(coef1),1)
aux<-cbind(sqrt(DX))
DP<-tao*aux
IC[,1]<-(coef1-(1.96*tao*aux))
IC[,2]<-(coef1+(1.96*tao*aux))
E<-matrix(0,nrow(coef1),1)
for(i in 1:ncol(Xmat)) E[i]<-((coef1[i])/(tao*aux[i]))
E<-abs(E)
TH[,1]<-E
TH[,2]<-2*(1-pnorm(E,0,1))
EReg<-matrix(0,1,2)
reg<-(SEAm-SEA)/(tao/2)
EReg[1,1]<-reg
EReg[1,2]<-(1-pchisq(reg,ncol(X)))

# Calculos das Estatisticas para Selecao de Variaveis
EAM<-SEA/(nrow(y))
RSEA<-SEAm-SEA
R2<-RSEA/(RSEA+(nrow(X)-ncol(X)-fator)*(tao/2))
R3<-RSEA/(RSEA+(nrow(X)-ncol(X)-fator)*(EAM/2))
Raju<-1-(1-R3)*(nrow(y)/(nrow(y)-ncol(X)-fator))

```

```

SEAP<-0
for(i in 1:nrow(X)) {
  X1<-matrix(0,(nrow(X)-1),ncol(X))
  y1<-matrix(0,(nrow(X)-1),1)
  k<-1
  for(j in 1:nrow(X)) if (j!=i) ((X1[k,]<-X[j,]) & (y1[k]<-y[j]) & (k<-k+1))
    if (intercepto==0) (fit1<-llfit(X1,y1,intercept=F))
    if (intercepto==1) (fit1<-llfit(X1,y1,intercept=T))
  coeiaux<-cbind(coef(fit1))
  yaju<-(Xmat[i,]%*%coeiaux)
  parc<-abs(y[i]-yaju)
  SEAP<-SEAP+parc
}

# Impressao dos Resultados
cat(" ",fill=T)
cat(" ",fill=T)
cat("*****",fill=T)
cat("*****",fill=T)
cat("***** SAIDA DO PROGRAMA L1.TXT *****",fill=T)
cat("*****",fill=T)
cat("*****",fill=T)
cat(" ",fill=T)
cat(" ",fill=T)
cat("PARTE INFERENCIAL",fill=T)
cat(" ",fill=T)
cat("Observacoes Definidoras:",fill=T)
print(def)
cat(" ",fill=T)

# Verifica se o usuario pediu a impressao dos Valores Observados,
# Ajustados e Residuos
if (imprime==1) {
  OAR<-cbind(y,aju,r)
  cat("Valores Observados, Ajustados e Residuos:",fill=T)
  cat(" ")
  print(OAR)
  cat(" ",fill=T)
}
tao<-round(tao,dig=4)
cat("Estimativa de tao:",tao,fill=T)
cat(" ",fill=T)
cat("Estimativas dos Parametros do Modelo:",fill=T)
coef1<-round(coef1,dig=4)
print(coef1)
cat(" ",fill=T)
cat("Estimativas dos Erros Padrao dos Estimadores do Modelo:",fill=T)
DP<-round(DP,dig=4)
print(DP)
cat(" ",fill=T)

```



```

cat("Intervalos de Confianca (95%):", fill=T)
IC<-round(IC,dig=4)
print(IC)
cat(" ", fill=T)
cat("Estatística do Teste Individual dos Parametros e Nivel Descritivo:", fill=T)
TH<-round(TH,dig=4)
print(TH)
cat(" ", fill=T)
cat("Estatística do Teste do Efeito de Regressão e Nivel Descritivo:", fill=T)
EReg<-round(EReg,dig=4)
print(EReg)
cat(" ", fill=T)
cat(" ", fill=T)
cat("ESTATISTICAS PARA SELECAO DE VARIABEIS", fill=T)
cat(" ", fill=T)
SEA<-round(SEA,dig=4)
cat("Soma dos Erros Absolutos (SEA):", SEA, fill=T)
cat(" ", fill=T)
EAM<-round(EAM,dig=4)
cat("Erro Absoluto Medio (EAM):", EAM, fill=T)
cat(" ", fill=T)
R2<-round(R2,dig=4)
cat("Coeficiente de Determinacao Multipla - R2:", R2, fill=T)
cat(" ", fill=T)

# Verifica se o usuario pediu a impressao das Estatisticas usadas para
# a Distribuicao de Laplace
if (laplace==1) {
  R3<-round(R3,dig=4)
  cat("Coeficiente de Determinacao Multipla - R3:", R3, fill=T)
  cat(" ", fill=T)
  Raju<-round(Raju,dig=4)
  cat("Coeficiente de Determinacao Multipla Ajustado:", Raju, fill=T)
  cat(" ", fill=T)
}
SEAP<-round(SEAP,dig=4)
cat("Soma dos Erros Absolutos Preditivos (SEAP):", SEAP, fill=T)
cat(" ", fill=T)
cat("*****", fill=T)
cat(" ", fill=T)

```

Exemplo de sequência de comandos no S-PLUS para rodar o programa L1.TXT:

```
> # Exemplo de Narula e Wellington (1977)
> dados<-scan("c:\\rodrigo\\mestrado\\tese\\imoveis.txt",what=list(preco=0, imposto=0, area=0,
areac=0, idade=0))
> attach(dados)
> # Modelo com as variaveis x1 e x3
> X<-cbind(imposto,areac)
> y<-cbind(preco)
> intercepto<-1
> imprime<-0
> laplace<-1
> source("c:\\rodrigo\\mestrado\\tese\\L1.txt")
```

Leitura dos dados.

Matriz com as variáveis explicativas.

Vetor com a variável resposta.

Modelo com intercepto.

Não vai imprimir os valores observados, ajustados e resíduos.

Vai imprimir as estatísticas para o caso da distribuição de Laplace.

Chamada do programa L1.TXT

Exemplo de saída gerada pelo programa L1.TXT:

```
*****
*****
***** SAIDA DO PROGRAMA L1.TXT *****
*****
```

PARTE INFERENCIAL

Observacoes Definidoras:

```
[,1]
[1,] 9
[2,] 14
[3,] 23
```

Observações com resíduo nulo.

Estimativa de tao: 5.4301 Obtido a partir de (2.4).

Estimativas dos Parametros do Modelo:

```
[,1]
Intercept 0.9787
imposto 2.6678
areac 12.5100
```

Estimativas dos Erros Padrao dos Estimadores do Modelo:

```
[,1]
[1,] 3.1080
[2,] 0.6667
[3,] 3.4420
```

Intervalos de Confianca (95%):

```
[,1] [,2]
[1,] -5.1130 7.0704
[2,] 1.3610 3.9745
[3,] 5.7636 19.2563
```

Intervalos para os parâmetros do modelo, obtidos a partir de (2.5).

Estatistica do Teste Individual dos Parametros e Nivel Descritivo:

```
[,1] [,2]
[1,] 0.3149 0.7528
[2,] 4.0014 0.0001
[3,] 3.6345 0.0003
```

Em cada linha é impresso o valor da estatística apresentada em (2.9) e o respectivo nível descritivo.

Estatística do Teste do Efeito de Regresso e Nível Descritivo:
 [,1] [,2]

[1,] 52.4238 0

Valor da estatística para testar a contribuição conjunta
 de todas as variáveis, dada em (2.10), e o nível descritivo.

ESTATISTICAS PARA SELECAO DE VARIABEIS

Soma dos Erros Absolutos (SEA): 84.0668

Obtido a partir de (3.1).

Erro Absoluto Medio (EAM): 3.1136

Obtido a partir de (3.10).

Coeficiente de Determinacao Multipla - R2: 0.686

Obtido a partir de (3.9).

Coeficiente de Determinacao Multipla - R3: 0.7921

Obtido a partir de (3.11).

Coeficiente de Determinacao Multipla Ajustado: 0.7661

Obtido a partir de (3.13).

Soma dos Erros Absolutos Preditivos (SEAP): 121.7719

Obtido a partir de (3.16).

Apêndice B

Programas de Procedimentos Automáticos de Seleção de Variáveis

Listagem do programa FORWARD.TXT, desenvolvido no pacote S-PLUS:

```
# PROGRAMA FORWARD.TXT - AGOSTO/1998
#
# Este programa realiza seleção de variaveis em regressao Ll atraves do
# procedimento automatico Selecao Forward, para modelos com ou sem
# intercepto.
#
# Variaveis a serem definidas pelo usuario:
# intercepto: recebendo o valor 0, trabalha com o modelo sem intercepto e
#             recebendo o valor 1, tarbalha com o modelo com intercepto.
# laplace: recebendo o valor 1, 'usa o Erro Absoluto Medio como estimativa de
#           tao e recebendo 0, usa o estimador dado em (2.4).
# alfa: nivel de significancia de entrada; pode receber qualquer valor entre
#       0 e 1.

cat(" ", fill=T)
cat(" ", fill=T)
cat("*****", fill=T)
cat("*****", fill=T)
cat("***** SELECAO FORWARD *****", fill=T)
cat("*****", fill=T)
cat("*****", fill=T)
cat(" ", fill=T)
cat(" ", fill=T)
cat("Nivel de Significancia de Entrada:", alfa, fill=T)
cat(" ", fill=T)
mod<-1
sim<-1
tam<-ncol(X)
```

```

# Verifica se o usuario pediu modelo com ou sem intercepto
if (intercepto==0) (SEAi<-sum(abs(y)))
if (intercepto==1) (SEAi<-sum(abs(y-median(y))))
modelo<-matrix(0,ncol(X),1)

# Enquanto houver variaveis a serem incluídas e estas forem significantes
while (sim==1) {
  SEAi<-matrix(0,tam,1)
  VF<-matrix(0,tam,1)
  for (j in 1:tam) {
    if (tam==ncol(X)) (Xmod<-cbind(X[,j]))
    if (tam!=ncol(X)) (Xmod<-cbind(Xmod1,Xmat[,j]))

    # Verifica se o usuario pediu modelo com ou sem intercepto
    if (intercepto==0) {
      fit<-llfit(Xmod, y, intercept=F)
      fator<-0
    }
    if (intercepto==1) {
      fit<-llfit(Xmod, y, intercept=T)
      fator<-1
    }

    # Estima tao e calcula a estatística do teste
    r<-cbind(residuals(fit))
    coef1<-cbind(coef(fit))
    SEAj<-sum(abs(r))

    # Verifica qual estimador de tao foi pedido pelo usuario
    if (laplace==0) {
      res<-cbind(sort(r))
      res1<-matrix(0,(nrow(res)-nrow(coef1)),1)
      n<-(nrow(Xmod)-ncol(Xmod)-fator)
      m<-((n+1)/2)-sqrt(n)
      k<-1
      for (l in 1:nrow(res)) if (res[l]!=0) ((res1[k]<-res[l]) & (k<-k+1))
      pos1<-(n-m+1)
      pos2<-m
      e1<-round(pos1)
      e2<-round(pos2)
      tao<-sqrt(n)*((res1[e1]-res1[e2])/4)
    }
    if (laplace==1) (tao<-SEAj/nrow(X))
    VF[j]<-(SEAi-SEAj)/(tao/2)
  }
  posmax<-0
  variavel<-0
  MF<-max(VF)
  for (j in 1:tam) if (VF[j]==MF) (posmax<-j)
}

```

```

# Identifica qual variavel apresentou o maior valor da estatistica
if (tam==ncol(X)) (variavel<-posmax)
if (tam!=ncol(X)) for (j in 1:ncol(X)) {
    igual<-0
    for (k in 1:nrow(X)) if (Xmat[k,posmax]==X[k,j]) (igual<-igual+1)
    if (igual==nrow(X)) (variavel<-j)
}
PMF<-(1-pchisq(MF,1))

# Se a variavel for siginificante
if (PMF<alfa) {
    tam<-tam-1
    if (tam!=0) {
        Xmat1<-matrix(0,nrow(y),tam)
        SEAi<-SEA[posmax]
        k<-1
        if (tam==ncol(X)-1) {
            for (j in 1:ncol(X)) if (j != posmax) ((Xmat1[,k]<-X[,j]) & (k<-k+1))
            Xmod1<-cbind(X[,posmax])
        }
        if (tam!=ncol(X)-1) {
            for (j in 1:ncol(Xmat)) if (j!=posmax) ((Xmat1[,k]<-Xmat[,j]) & (k<-k+1))
            Xmod1<-cbind(Xmod1,Xmat[,posmax])
        }
        Xmat<-Xmat1
    }
    cat("Variavel Incluida:",variavel,fill=T)
    MF<-round(MF,dig=4)
    cat("Valor da Estatistica:",MF,fill=T)
    cat(" ",fill=T)
    modelo[mod]<-variavel
    mod<-mod+1
}
if ((PMF>=alfa) || (tam==0)) (sim<-0)
}

# Impressao dos resultados
modelo<-cbind(sort(modelo))
if (sum(modelo)!=0) {
    cat("Variaveis no Modelo Final:")
    for (i in 1:nrow(modelo)) {
        if ((modelo[i]!=0) && (i!=nrow(modelo))) (cat(modelo[i],","))
        if ((modelo[i]!=0) && (i==nrow(modelo))) (cat(modelo[i],))
    }
    cat(" ",fill=T)
    cat(" ",fill=T)
}
if (sum(modelo)==0) {
    cat("Nenhuma Variavel Incluida",fill=T)
    cat(" ",fill=T)
}

```

```
cat("*****", fill=T)
cat(" ", fill=T)
```

Exemplo de sequência de comandos no S-PLUS para rodar o programa FORWARD.TXT:

```
> # Exemplo de Narula e Wellington (1977)
> dados<-scan("c:\\rodrigo\\mestrado\\tese\\imoveis.txt", what=list(preco=0, imposto=0, area=0,
areac=0, idade=0))
> attach(dados)
> X<-cbind(imposto, area, areac, idade)
> y<-cbind(preco)
> intercepto<-1
> laplace<-0
> alfa<-0.05
> source("c:\\rodrigo\\mestrado\\tese\\forward.txt")
```

Leitura dos dados.

Matriz com as variáveis explicativas.

Vetor com a variável resposta.

Modelo com intercepto.

Estimador de τ dado em (2.4).

Nível de significância de entrada.

Chamada do programa FORWARD.TXT

Exemplo de saída gerada pelo programa FORWARD.TXT:

```
*****
*****
***** SELECAO FORWARD *****
*****
*****
```

Nivel de Significancia de Entrada: 0.05

Variavel Incluida: 1

Valor da Estatistica: 61.3645

Obtida a partir de (3.21).

Variavel Incluida: 3

Valor da Estatistica: 7.1315

Obtida a partir de (3.21).

Variaveis no Modelo Final:1 ,3

```
*****
```

Listagem do programa BACKWARD.TXT, desenvolvido no pacote S-PLUS:

```
# PROGRAMA BACKWARD.TXT - AGOSTO/1998
#
# Este programa realiza seleção de variaveis em regressão L1 através do
# procedimento automatico Eliminacao Backward, para modelos com ou sem
# intercepto.
#
# Variaveis a serem definidas pelo usuario:
# intercepto: recebendo o valor 0, trabalha com o modelo sem intercepto e
#             recebendo o valor 1, trabalha com o modelo com intercepto.
# laplace: recebendo o valor 1, usa o Erro Absoluto Medio como estimativa de
#          tao e recebendo 0, usa o estimador dado em (2.4).
# alfa: nivel de significancia de saida; pode receber qualquer valor entre
#       0 e 1.

cat(" ", fill=T)
cat(" ", fill=T)
cat("*****", fill=T)
cat("*****", fill=T)
cat("***** ELIMINACAO BACKWARD *****", fill=T)
cat("*****", fill=T)
cat("*****", fill=T)
cat(" ", fill=T)
cat(" ", fill=T)
cat("Nivel de Significancia de Saidas:", alfa, fill=T)
cat(" ", fill=T)
sim<-1
tam<-ncol(X)
modelo<-matrix(0,ncol(X),1)
for (i in 1:ncol(X)) modelo[i]<-i

# Enquanto houver variaveis a serem retiradas e estas forem nao-significantes
while (sim==1) {
  SEA<-matrix(0,tam,1)
  VF<-matrix(0,tam,1)
  if (tam==ncol(X)) (Xmat<-X)

  # Verifica se o usuario pediu modelo com ou sem intercepto
  if (intercepto==0) {
    fit<-llfit(Xmat, y, intercept=F)
    fator<-0
  }
  if (intercepto==1) {
    fit<-llfit(Xmat, y, intercept=T)
    fator<-1
  }
}
```



```

# Estima tao e calcula a estatistica do teste
r<-cbind(residuals(fit))
coefl<-cbind(coef(fit))
SEAi<-sum(abs(r))

# Verifica qual estimador de tao foi pedido pelo usuario
if (laplace==0) {
  res<-cbind(sort(r))
  resl<-matrix(0,(nrow(res)-nrow(coefl)),1)
  n<-(nrow(Xmat)-ncol(Xmat)-fator)
  m<-((n+1)/2)-sqrt(n)
  k<-1
  for (l in 1:nrow(res)) if (res[l]!=0) ((resl[k]<-res[l]) & (k<-k+1))
  pos1<-(n-m+1)
  pos2<-m
  e1<-round(pos1)
  e2<-round(pos2)
  tao<-sqrt(n)*((resl[e1]-resl[e2])/4)
}
if (laplace==1) (tao<-(SEAi/nrow(X)))
for (j in 1:tam) {
  if (tam!=1) {
    Xaux<-matrix(0,nrow(Xmat),tam-1)
    k<-1
    for (i in 1:ncol(Xmat)) if (i!=j) ((Xaux[,k]<-Xmat[,i]) & (k<-k+1))
    if (intercepto==0) (fit1<-llfit(Xaux, y, intercept=F))
    if (intercepto==1) (fit1<-llfit(Xaux, y, intercept=T))
    r<-cbind(residuals(fit1))
    SEA[j]<-sum(abs(r))
    VF[j]<-(SEA[j]-SEAi)/(tao/2)
  }
  if (tam==1) {

    # Verifica se o usuario pediu modelo com ou sem intercepto
    if (intercepto==0) (SEA[j]<-sum(abs(y)))
    if (intercepto==1) (SEA[j]<-sum(abs(y-median(y))))
    VF[j]<-(SEA[j]-SEAi)/(tao/2)
  }
}
posmin<-0
variavel<-0
MF<-min(VF)
for (j in 1:tam) if (VF[j]==MF) (posmin<-j)

```

```

# Identifica qual variavel apresentou o menor valor da estatística
if (tam==ncol(X)) (variavel<-posmin)
if (tam!=ncol(X)) for (j in 1:ncol(X)) {
    igual<-0
    for (k in 1:nrow(X)) if (Xmat[k,posmin]==X[k,j]) (igual<-igual+1)
    if (igual==nrow(X)) (variavel<-j)
}
PMF<-(1-pchisq(MF,1))

# Se a variavel for nao-siginificante
if (PMF>=alfa) {
    tam<-tam-1
    if (tam!=0) {
        Xmat1<-matrix(0,nrow(y),tam)
        k<-1
        if (tam==ncol(X)-1) {
            for (j in 1:ncol(X)) if (j!=posmin) ((Xmat1[,k]<-X[,j]) & (k<-k+1))
            Xmat<-Xmat1
        }
        if (tam!=ncol(X)-1) {
            for (j in 1:ncol(Xmat)) if (j!=posmin) ((Xmat1[,k]<-Xmat[,j]) & (k<-k+1))
            Xmat<-Xmat1
        }
    }
    cat("Variavel Eliminada:",variavel,fill=T)
    MF<-round(MF,dig=4)
    cat("Valor da Estatística:",MF,fill=T)
    cat(" ",fill=T)
    modelo[variavel]<-0
}
if ((PMF<alfa) || (tam==0)) (sim<-0)
}

# Impressao dos resultados
modelo<-cbind(sort(modelo))
if (sum(modelo)!=0) {
    cat("Variaveis no Modelo Final:")
    for (i in 1:nrow(modelo)) {
        if ((modelo[i]!=0) && (i!=nrow(modelo))) (cat(modelo[i],", "))
        if ((modelo[i]!=0) && (i==nrow(modelo))) (cat(modelo[i],))
    }
    cat(" ",fill=T)
    cat(" ",fill=T)
}
if (sum(modelo)==0) {
    cat("Todas as Variaveis Eliminadas",fill=T)
    cat(" ",fill=T)
}
cat("*****",fill=T)
cat(" ",fill=T)

```

Exemplo de sequência de comandos no S-PLUS para rodar o programa BACKWARD.TXT:

```
> # Exemplo de Narula e Wellington (1977)
> dados<-scan("c:\\rodrigo\\mestrado\\tese\\imoveis.txt",what=list(preco=0, imposto=0, area=0,
areac=0, idade=0))
> attach(dados)
> X<-cbind(imposto,area,areac,idade)
> y<-cbind(preco)
> intercepto<-1
> laplace<-0
> alfa<-0.05
> source("c:\\rodrigo\\mestrado\\tese\\backward.txt")
```

Leitura dos dados.

Matriz com as variáveis explicativas.

Vetor com a variável resposta.

Modelo com intercepto.

Estimador de τ dado em (2.4).

Nível de significância de saída.

Chamada do programa BACKWARD.TXT

Exemplo de saída gerada pelo programa BACKWARD.TXT:

```
*****
*****
***** ELIMINACAO BACKWARD *****
*****
*****
```

Nível de Significancia de Saida: 0.05

Variavel Eliminada: 2

Valor da Estatistica: 0.2851

Obtida a partir de (3.21).

Variavel Eliminada: 4

Valor da Estatistica: 2.0731

Obtida a partir de (3.21).

Variaveis no Modelo Final:1 ,3

```
*****
```

Apêndice C

Programa de Simulação

Listagem do programa MULTICOL.TXT, desenvolvido no pacote S-PLUS:

```
# PROGRAMA MULTICOL.TXT - AGOSTO/1998
#
# Este programa faz a simulacao descrita no Capitulo 4, imprimindo
# todos os dados que estao apresentados nas tabelas deste capitulo.
#
# Variavel a ser definida pelo usuario:
# alfa: constante fixada.

x<-cbind(x)
v<-cbind(v)
e<-cbind(e)
xbarra<-mean(x)
vbarra<-mean(v)
xc<-matrix(0,nrow(x),1)
vc<-matrix(0,nrow(v),1)
xc<-(x-xbarra)
vc<-(v-vbarra)
XVc<-cbind(xc,vc)

# Uso de Componentes Principais
comp<-princomp(XVc,covlist=NULL)
cargas<-cbind(loadings(comp))

# Obtencao dos valores de x1 e x2
XVt<-(XVc%*%cargas)
dpxt<-sqrt(var(XVt[,1]))
dpvt<-sqrt(var(XVt[,2]))
numero<-sqrt(nrow(x)-1)
xt<-cbind(XVt[,1]/dpxt)
xt<-xt/numero
vt<-cbind(XVt[,2])
x1<-cbind(xt)
```

```

x2<-cbind((alfa*x1)+vt)
y15<-x1+x2+e
y15<-cbind(y15)
Xmat<-cbind(x1,x2)
CL1<-matrix(0,nrow(Xmat),ncol(Xmat))
X<-Xmat
y<-y15
fitl1<-l1fit(X, y, intercept=F)
r<-cbind(residuals(fitl1))
coefl<-cbind(coef(fitl1))
aju<-(y-r)
n<-(nrow(X)-ncol(X))
m<-((n+1)/2)-sqrt(n)
res<-cbind(sort(r))
res1<-matrix(0,(nrow(res)-nrow(coefl)),1)
def<-matrix(0,nrow(coefl),1)
j<-1
for(i in 1:nrow(r)) if (r[i]==0) ((def[j]<-i) & (j<-j+1))
j<-1
for(i in 1:nrow(res)) if (res[i]!=0) ((res1[j]<-res[i]) & (j<-j+1))
pos1<-(n-m+1)
pos2<-m
e1<-round(pos1)
e2<-round(pos2)
tao<-sqrt(n)*((res1[e1]-res1[e2])/4)
OAR<-cbind(y,aju,r)
XX<-solve(t(X)%*%X)
DX<-diag(XX)
DP<-matrix(0,nrow(coefl),1)
aux<-cbind(sqrt(DX))
DP<-tao*aux

# Impressao de Resultados do Metodo L1
cat(" ",fill=T)
cat("RESULTADOS PELO METODO L1",fill=T)
cat(" ",fill=T)
cat("Observacoes Definidoras:",fill=T)
print(def)
cat(" ",fill=T)
tao<-round(tao,dig=4)
cat("Estimativa de tao:",tao,fill=T)
cat(" ",fill=T)
cat("Coeficientes Estimados e Erros Padrao:",fill=T)
PDP<-cbind(coefl,DP)
PDP<-round(PDP,dig=4)
print(PDP)
cat(" ",fill=T)

```

```

# Retira cada observacao e reajusta pelo Metodo L1
for(i in 1:nrow(Xmat)) {
  Xmat1<-matrix(0, (nrow(Xmat)-1), ncol(Xmat))
  yaux<-matrix(0, (nrow(Xmat)-1), 1)
  k<-1
  for(j in 1:nrow(Xmat)) if (j!=i) ((Xmat1[k,]<-Xmat[j,]) & (yaux[k]<-y15[j]) & (k<-k+1))
  fit1<-llfit(Xmat1, yaux, intercept=F)
  coef1<-cbind(coef(fit1))
  CL1[i,]<-t(coef1)
}

# Ajuste de Minimos Quadrados e Impressao de Resultados
fitmq<-lm(y15~x1+x2-1)
cat("RESULTADOS PELO METODO DE MINIMOS QUADRADOS", fill=T)
cat(" ", fill=T)
print(summary(fitmq))
cat(" ", fill=T)
cat("Estimativas dos Parametros com a Retirada das Observacoes:", fill=T)
CL1<-round(CL1, dig=4)
print(CL1)
cat(" ", fill=T)
corre<-cor(cbind(x1, x2))
cat("Matriz de Correlacao entre x1 e x2:", fill=T)
corre<-round(corre, dig=2)
print(corre)
cat(" ", fill=T)

```

Exemplo de sequência de comandos no S-PLUS para rodar o programa MULTICOL.TXT:

```

# Gera o Conjunto de Dados 2 do Capitulo 4
> dadosn<-scan("c:\\rodrigo\\mestrado\\tese\\normalb.txt", what=list(x=0, v=0, e=0))
> attach(dadosn)
> alfa<-0.5
> source("c:\\rodrigo\\mestrado\\tese\\multicol.txt")

```

Constante fixada pelo usuário.
Chamada do programa MULTICOL.TXT

Exemplo de saída gerada pelo programa MULTICOL.TXT:

RESULTADOS PELO METODO L1

Observacoes Definidoras:

[,1]

[1,] 4
[2,] 14

Observações com resíduo nulo.

Estimativa de tao: 1.3576

Obtido a partir de (2.4).

Coefficientes Estimados e Erros Padrao:

[,1] [,2]
X1 -0.2809 1.3709
X2 0.4057 0.3801

RESULTADOS PELO METODO DE MINIMOS QUADRADOS

```
Call: lm(formula = y15 ~ x1 + x2 - 1)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-2.019 -0.9724 -0.5692  0.1272  1.197
```

```
Coefficients:
```

```
      Value Std. Error t value Pr(>|t|)
x1 0.6581  1.0537      0.6245  0.5431
x2 0.5679  0.2921      1.9437  0.0739
```

Resultados do ajuste de Mínimos Quadrados.

```
Residual standard error: 1.044 on 13 degrees of freedom
```

```
Multiple R-Squared: 0.2611
```

```
F-statistic: 2.296 on 2 and 13 degrees of freedom, the p-value is 0.1399
```

```
Correlation of Coefficients:
```

```
      x1
x2 -0.1386
```

```
Estimativas dos Parametros com a Retirada das Observacoes:
```

```
      [,1] [,2]
[1,] -0.2809 0.4057
[2,]  0.4526 0.7393
[3,] -0.5731 0.4271
[4,] -0.5731 0.4271
[5,] -0.5731 0.4271
[6,] -0.2809 0.4057
[7,] -0.2809 0.4057
[8,] -0.5731 0.4271
[9,] -0.2809 0.4057
[10,] -0.5731 0.4271
[11,] -0.2809 0.4057
[12,] -0.2809 0.4057
[13,] -0.2809 0.4057
[14,] -0.4663 0.3213
[15,] -0.2809 0.4057
```

```
Matriz de Correlacao entre x1 e x2:
```

```
      [,1] [,2]
[1,] 1.00 0.14
[2,] 0.14 1.00
```

Bibliografia

- Abdelmalek, N. N. (1980). L_1 solution of over determined system of linear equations. **ACM Transactions on Mathematical Software**, 6, 220-227.
- André, C. D. S., Elia, S. N., Narula, S. C. e Aubin, E. C. Q. (1996). **Stepwise procedure for selecting variables in the minimum sum of absolute errors regression**. Relatório Técnico RT-MAE 9619, IME-USP. 9p.
- André, C. D. S., Elia, S. N. e Tavares, R. A. (1998). **Coeficientes de determinação na regressão de mínima soma dos erros absolutos e sua utilização em problemas de seleção de variáveis**. Relatório Técnico RT-MAE, IME-USP. 13p.
- Appa, G. e Smith, C. (1973). On L_1 and Chebychev estimation. **Mathematical Programming**, 5, 73-87.
- Armstrong, R. D., Frome, E. e Kung, D. S. (1979). A revised simplex algorithm for the absolute deviation curve fitting problem. **Communications in Statistics**, B8, 175-190.
- Armstrong, R. D. e Kung, D. S. (1978). AS132: Least absolute value estimates for a simple linear regression problem. **Applied Statistics**, 27, 363-366.
- Bain, L. J. e Engelhardt, M. (1973). Interval estimation for the two-parameter double exponential distribution. **Technometrics**, 15, 875-887.
- Barrodale, I. e Roberts, F. D. K. (1973). An improved algorithm for Discrete L_1 linear approximation. **SIAM Journal of Numerical Analysis**, 10, 839-848.

Barrodale, I. e Roberts, F. D. K. (1974). Solution of an overdetermined system of equations in the L_1 norm. **Communications of the ACM**, 17, 319-320.

Bartels, R. H., Conn, A. R. e Sinclair, J. W. (1976). **A FORTRAN program for solving overdetermined system of linear equations in the L_1 sense**. Technical Report No. 236, Mathematical Sciences Department, John Hopkins University, Baltimore, MD, USA.

Bartels, R. H., Conn, A. R. e Sinclair, J. W. (1978). Minimization techniques for piecewise differentiable functions: The L_1 solution to an overdetermined linear system. **SIAM Journal of Numerical Analysis**, 15, 224-241.

Boscovich, R. J. (1757). Delitteria expeditone per pontificiam ditionem, et synopsis amplioris operis, ac habentur plura ejus ex exemplaria etiam sensorum impressa. **Bononiesi Sientiarum st Artum Institute Atque Academia Commentarii**, 4, 353-396.

Basset, G. e Koenker, R. (1978). Asymptotic theory of least absolute error regression. **Journal of the American Statistical Association**, 73, 618-622.

Birkes, D. e Dodge, Y. (1993). **Alternative Methods of Regression**. John Wiley and Sons, Inc., New York. 228p.

Bloomfield, P. e Steiger, W. (1983). **Least Absolute Deviations: Theory, Applications and Algorithms**. Birkhauser, Boston, MA.

Charnes, A., Cooper, W. W. e Ferguson, R. D. (1955). Optimal estimation of executive compensation by linear programming. **Management Science**, 1, 138-151.

Coleman, T. F. e Li, Y. (1992). A globally and quadratically convergent affine scaling method for linear L_1 problems. **Mathematical Programming**, 56, 189-222.

Cox, D. H. e Hinkley D. V. (1991). **Statistical theory and modelling**. Chapman Hall, London. 349p.

Dielmann, T. e Pfaffenberger, R. (1982). LAV (Least Absolute Value) estimation in the regression model: A review. **TIMS Studies in the Management Sciences**, 19, 31-52.

Edgeworth, F. Y. (1887). On observations relating to several quantities. **Hermathena**, 6, 279-285.

Edgeworth, F. Y. (1888). On a new method of reducing observations relating to several quantities. **Philosophical Magazine**, 25, 184-191.

Elian, S. N., André, C. D. S. e Narula, S. C. (1997). **Influence measure for the minimum sum of absolute errors regression**. Relatório Técnico RT-MAE 9733, IME-USP. 9p.

Farebrother, R. W. (1987). The historical development of the L_1 and L_∞ estimation procedures. **Statistical Data Analysis Based on the L_1 -Norm and Related Methods**. (Y. Dodge, editor). North-Holland, Amsterdam, Holland, 37-64.

Gentle, J. E., Sposito, V. A. e Narula, S. C. (1987). Algorithms for unconstrained L_1 linear regression. **Statistical Data Analysis Based on the L_1 -Norm and Related Methods**. (Y. Dodge, editor). North-Holland, Amsterdam, Holland, 83-94.

Gentle, J. E., Sposito, V. A. e Narula, S. C. (1988). Algorithms for unconstrained L_1 simple linear regression. **Computational Statistics and Data Analysis**, 6, 335-339.

Grácio, M. C. C. (1991). **Transformações e Diagnósticos em Regressão**. Dissertação de Mestrado, IMECC-UNICAMP. 129p.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. **Biometrics**, **32**, 1-49.

Josavenger, L. A. e Sposito, V. A. (1983). L_1 -norm estimates for the simple linear regression problem. **Communications in Statistics: Simulation and Computation**, **B12**, 215-221.

Karst, O. J. (1958). Linear curve fitting using least deviations. **Journal of the American Statistical Association**, **53**, 118-132.

Klingman, D. e Mote, J. (1982). Generalized network approaches for solving least absolute value and Tchebycheff regression problem. **TIMS Studies in the Management Sciences**, **19**, 53-66.

Madsen, K. e Nielsen, H. B. (1993). A finite smoothing algorithm for linear L_1 estimation. **SIAM Journal on Optimization**, **3**, 223-235.

McKean, J. W. e Schrader, R. M. (1987). Lest absolute errors analysis of variance. **Statistical Data Analysis Based on the L_1 -Norm and Related Methods** (Y. Dodge, editor), Elsevier Science Publishers B. V., 297-305.

McKean, J. W. e Sievers, G. L. (1987). Coefficient of Determination for least absolute deviation analysis. **Statistics and Probability Letters**, **5**, 49-54.

Montgomery, D. C. e Peck, E. A. (1992). **Introduction to linear regression analysis**. 2 ed. John Wiley and Sons, Inc., New York. 527p.

Narula, S. C. (1987). The minimum sum of absolute errors regression. **Journal of Quality Technology**, **19**, 37-45.

Narula, S. C. e Wellington, J. F. (1977). Prediction, linear regression and minimum sum of absolute errors. **Technometrics**, **19**, 185-190.

- Narula, S. C. e Wellington, J. F. (1979). Selection of variables in linear regression using the minimum sum of weighted absolute errors criterion. **Technometrics**, **21**, 299-306.
- Narula, S. C. e Wellington, J. F. (1985). Interior analysis for the minimum sum of absolute errors regression. **Technometrics**, **27**, 181-188.
- Narula, S. C. e Wellington, J. F. (1998). Intervals that leave the fitted MSAE multiple linear regression model unchanged. To appear.
- Neter, J., Kutner, M. H., Nachtsheim, C. J. e Wasserman, W. (1996). **Applied linear statistical models**. 4 ed. Homewood: Richard D. Irwin. 1410p.
- Ruzinsky, S. e Olsen, E. (1989). L_1 and L_∞ Minimization via a variant of Karmarkar's algorithm. **IEEE Transactions on Acoustics, Speech and Signal Processing**, **37**, 245-253.
- Schrader, R. M. e McKean, J. W. (1982). **Least absolute errors analysis of variance**. Technical Report. Dept of Math. and Stat., University of New Mexico, Albuquerque, NM.
- Seber, G. A. F. (1977). **Linear regression analysis**. John Wiley and Sons, Inc., New York. 456p.
- Stangenhuis, G. e Narula, S. C. (1991). Inference procedures for the L_1 regression. **Computational Statistics and Data Analysis**, **12**, 1, 79-85.
- Stangenhuis, G., Narula, S. C. e Ferreira, P. F. (1993). Bootstrap confidence intervals for the minimum sum of absolute errors regression. **Journal of Statistical Computation and Simulation**, **48**, 127-133.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. **Journal of The Royal Statistical Society**, **B58**, 1, 267-288.

Wagner, H. M. (1959). Linear programming techniques for regression analysis. **Journal of the American Statistical Association**, **54**, 206-212.

Wellington, J. W. e Narula, S. C. (1981). Variable selection in multiple linear regression using the minimum sum of weighted absolute errors criterion. **Communications in Statistics: Simulation and Computation**, **B10**, 641-648.

Wesolowsky, G. O. (1981). A new descent algorithm for the least absolute value regression problem. **Communications in Statistics: Simulation and Computation**, **B10**, 479-491.

Zhang, Y. (1993). Primal-dual interior point approach for computing L_1 solutions and L_∞ solutions of overdetermined systems. **Journal of Optimization Theory and Applications**, **77**, 323-341.