

RT-MAE 9831

*COEFICIENTES DE DETERMINAÇÃO NA
REGRESSÃO DE MÍNIMA SOMA DOS ERROS
ABSOLUTOS E SUA UTILIZAÇÃO EM
PROBLEMAS DE SELEÇÃO DE VARIÁVEIS*

by

*Carmon D. Saldiva de André, Sílvia Nagib Ellen
and
Rodrigo Tavares*

Palavras-Chave: coeficiente de determinação ajustado, regressão L_1 , seleção de variáveis.

Classificação AMS: 62J05.
(AMS Classification)

COEFICIENTES DE DETERMINAÇÃO NA REGRESSÃO DE MÍNIMA SOMA DOS ERROS ABSOLUTOS E SUA UTILIZAÇÃO EM PROBLEMAS DE SELEÇÃO DE VARIÁVEIS

Carmen D. Saldiva de André

Sílvia Nagib Elian

Rodrigo Tavares

IME-USP

Resumo

Nosso objetivo é modificar o coeficiente de determinação robusto para a regressão MSAE proposto por McKean e Sievers (1987) de tal forma que ele apresente boas propriedades. Propomos também um coeficiente de determinação ajustado que é apropriado para comparar modelos com diferentes números de variáveis. Mostramos que, se o valor desse coeficiente decresce quando novas variáveis são adicionadas ao modelo, então a contribuição dessas variáveis não é significativa.

Palavras chave : coeficiente de determinação ajustado, regressão L_1 , seleção de variáveis.

1. Introdução

Consideremos o modelo de regressão linear múltipla :

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{(p-1)} x_{i(p-1)} + e_i, \quad (1.1)$$

onde

y_i é o valor da variável resposta para a i -ésima observação,
 $x_{i1}, \dots, x_{i(p-1)}$ são os respectivos valores das variáveis explicativas,
 $\beta_1, \dots, \beta_{p-1}$ são os parâmetros do modelo, e
 e_i são erros casuais, $i = 1, \dots, n$.

Faremos as seguintes suposições sobre os componentes do modelo:

1.1) $e_1, \dots, e_n$ são variáveis aleatórias independentes e identicamente distribuídas com função distribuição F .

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

1.3) Seja $X = (x_{ij})_{n \times p-1}$ tal que $(1/n) X'X \rightarrow C$, quando $n \rightarrow \infty$, onde C é uma matriz positiva definida.

1.4) Para todo $j = 1, \dots, p-1$, $\max_{1 \leq i \leq n} (x_{ij} - \bar{x})^2 / \sum_{i=1}^n (x_{ij} - \bar{x})^2 \rightarrow 0$, quando $n \rightarrow \infty$.

Os parâmetros do modelo são usualmente estimados pelo método de mínimos quadrados. Quando este método é utilizado, o coeficiente de determinação R^2 , definido como:

$$R^2 = (SQT - SQR) / SQT, \quad (1.1)$$

onde SQR é a Soma de Quadrados dos Resíduos e SQT a Soma de Quadrados Total, é frequentemente utilizado como uma medida informal do ajuste do 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 da mínima soma dos erros absolutos (MSAE) é uma boa alternativa ao de mínimos quadrados quando a distribuição dos erros tem caudas pesadas. Nas situações em que esse método é adotado, é interessante a utilização de um coeficiente de determinação que também seja resistente a *valores aberrantes*. Neste trabalho apresentamos, na seção 2, os coeficientes de determinação mais utilizados quando o

método MSAE é adotado e propriedades desses coeficientes. Mostramos, através de um exemplo, que o coeficiente mais utilizado pode não ser adequado para comparar diferentes modelos, pois o valor desse coeficiente pode diminuir mesmo quando uma variável que contribui na explicação da variável resposta é adicionada ao modelo. Buscando solucionar esse problema, propomos, na seção 3, um novo coeficiente de determinação. Na seção 4 introduzimos, para a regressão MSAE, um coeficiente com propriedades semelhantes ao R^2 ajustado para graus de liberdade utilizado na regressão de mínimos quadrados.

2. Coeficientes de determinação na regressão MSAE

Com o objetivo de definir um coeficiente de determinação robusto apropriado à regressão MSAE, McKean e Sievers (1987), inicialmente consideraram que, no caso de variáveis explicativas 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 de que os coeficientes de regressão dessas variáveis são nulos.

(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 total de ajuste.

(P5) R deve aumentar quando se acrescentam novas variáveis preditoras ao modelo.

(P6) R deve ser robusto.

Fazendo analogia com o R^2 dado em (1.1), os autores definiram o seguinte coeficiente :

$$R_1 = \frac{RSEA}{\sum_{i=1}^n |y_i - \hat{\beta}_0|}, \quad (2.1)$$

onde :

$$RSEA = \sum_{i=1}^n |y_i - \hat{\beta}_0| - \sum_{i=1}^n |y_i - \hat{y}_i|,$$

$$\hat{\beta}_0 = \text{mediana}(y_1, \dots, y_n) \ominus$$

$\hat{y}_i$ é o valor previsto da i -ésima observação através do modelo ajustado pelo método MSAE.

Observemos que $RSEA$ corresponde à diferença que ocorre na soma dos erros absolutos quando passamos do modelo ajustado sem nenhuma variável explicativa, ou seja, $\hat{y}_i = \text{mediana}(y_1, \dots, y_n) = \hat{\beta}_0$, para o modelo ajustado com as $(p-1)$ variáveis explicativas. Sua interpretação na regressão MSAE é a mesma que a da soma de 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 de soma dos erros absolutos total. Segundo os autores R_1 pode ser interpretado como a proporção na soma dos desvios absolutos total devida à regressão. Apesar de satisfazer às propriedades (P1), (P3), (P4) e (P5), foi demonstrado que R_1 não é robusto quando um valor aberrante extremo é observado, e portanto, não satisfaz à propriedade (P6), embora seja mais robusto que o coeficiente de determinação R^2 .

Para obter um coeficiente de determinação robusto para a regressão MSAE, os autores consideraram 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, testar a hipótese:

$$H_0: \beta^* = 0 \text{ versus } H_1: \beta^* \neq 0, \quad (2.2)$$

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

$$F = ((n-p) / (p-1))(R^2 / (1-R^2)). \quad (2.3)$$

Sob as suposições (1.1), (1.3) e (1.4), e supondo $\text{var}(e_i) < \infty$, Arnold (1981) mostrou que $(p-1)F$ tem distribuição assintótica $\chi^2_{(p-1)}$

Quando o método MSAE é utilizado, $RSEA$ é uma estatística natural para o teste da hipótese de interesse (2.2), sendo que valores grandes dessa estatística indicam a rejeição de H_0 . Contudo isto deve ser medido numa escala apropriada. O parâmetro de escala utilizado na regressão MSAE é dado por

$$\tau = (2 f(0))^{-1},$$

onde f é a função densidade de e_i , $i=1, \dots, n$. Um estimador consistente de τ , dado em Schrader e McKean (1982) é:

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

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

Uma estatística apropriada para o teste da hipótese (2.2) é então dada por:

$$D_{p-1} = RSEA / (\hat{\tau}/2). \quad (2.4)$$

Schrader e McKean (1987) mostraram ainda que, sob as suposições apresentadas no Apêndice A, D_{p-1} tem distribuição assintótica $\chi^2_{(p-1)}$ sob H_0 . Portanto a hipótese H_0 será rejeitada se D_{p-1} exceder um valor crítico da distribuição χ^2 com $(p-1)$ graus de liberdade.

A relação entre R^2 e a estatística F (2.3) sugere que um coeficiente de determinação R_2 , para regressão L_1 , pode ser definido a partir da estatística (2.4) considerando-se a relação $(n-p)R_2 / (1-R_2) = D_{p-1}$. Substituindo D_{p-1} por sua expressão dada em (2.4), temos :

$$R_2 = RSEA / (RSEA + (n-p)(\hat{\tau}/2)) \quad (2.5)$$

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, conforme pode ser observado no exemplo a seguir. Este fato pode gerar resultados confusos, no caso do coeficiente R_2 ser utilizado como um indicativo do modelo a ser selecionado entre todas as possíveis regressões.

Exemplo - Os dados da Tabela 2.1 são relativos a imóveis, e foram 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.

Tabela 3.1. Dados relativos a imóveis.

y	x₁	x₂	x₃	x₄
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	58
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	6,8860	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

Na Tabela 3.2 apresentamos todos os possíveis modelos que podem ser obtidos a partir das variáveis explicativas consideradas, e as correspondentes estimativas $\hat{\tau}$ obtidas após o ajuste desses modelos pelo método MSAE.

Tabela 3.2. Número de parâmetros (p) e $\hat{\tau}$ para todos os modelos de regressão possíveis, utilizando os dados de imóveis.

Variáveis no Modelo	p	$\hat{\tau}$	R_2
Nenhuma	1	10,9629	0,0000
x_1	2	4,0079	0,7105
x_2	2	8,0285	0,3830
x_3	2	9,8973	0,4681
x_4	2	11,8750	0,1164
x_1, x_2	3	3,9590	0,7215
x_1, x_3	3	5,4301	0,6860
x_1, x_4	3	3,9636	0,7212
x_2, x_3	3	8,0913	0,5470
x_2, x_4	3	6,5918	0,4576
x_3, x_4	3	6,6577	0,6098
x_1, x_2, x_3	4	6,1331	0,6701
x_1, x_2, x_4	4	7,3275	0,5937
x_1, x_3, x_4	4	5,4875	0,7011
x_2, x_3, x_4	4	4,0022	0,7410
x_1, x_2, x_3, x_4	5	4,0246	0,7704

Observamos que $\hat{\tau}$ do 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 contribuição adicional da variável x_3 , dado que x_1 está no modelo, é significativa (nível descritivo menor que 0.01).

Portanto, devido à instabilidade das estimativas de τ , pode ocorrer que, mesmo sendo introduzida uma variável que traga contribuição na explicação da variável resposta, a estimativa de τ aumente quando se passa do modelo reduzido para o completo. Este fato diminuiria o valor de R_2 , fazendo com que esse modelo não seja selecionado, se o coeficiente de determinação for o critério utilizado para escolha do modelo final entre todas as possíveis regressões.

3. Coeficiente de determinação proposto

Nesta seção vamos propor um coeficiente de determinação a ser utilizado quando os parâmetros do modelo são estimados pelo método MSAE que satisfaça às propriedades (P1) a (P6) citadas na seção 2.

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

$$EAM = SEA/n.$$

Apesar das condições de regularidade usuais não estarem satisfeitas, esse estimador é consistente (Engelhardt e Bain (1973)). (2.6)

É possível, então, definir um coeficiente de determinação a partir de (2.5) para o caso da distribuição dos erros ser Laplace, substituindo $\hat{\tau}$ por EAM, obtendo-se:

$$R_3 = RSEA / (RSEA + (n-p)(EAM / 2)). \quad (2.7)$$

Este coeficiente difere de R_2 apenas pelo estimador do parâmetro de escala τ , e portanto possui as mesmas propriedades que R_2 , satisfazendo ainda a propriedade (P5), uma vez que o erro médio absoluto diminui à medida que novas variáveis são introduzidas no modelo.

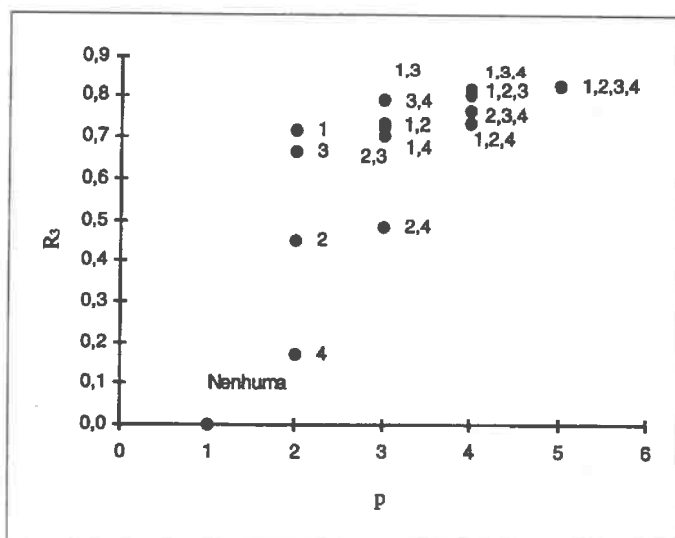
Assumindo que os erros seguem distribuição de Laplace no exemplo dado na seção 2, foram calculados os valores de $\hat{\tau}$ (EMA) e R_3 para cada possível modelo. Os valores obtidos são apresentados na Tabela 3.3.

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 este 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 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 , 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 R_3 como critério de escolha, o modelo selecionado é o com as variáveis x_1 e x_3 .

Tabela 3.3. Valores de R_3 e do Erro Absoluto Médio (EAM) para o exemplo dos imóveis.

Variáveis no Modelo	R_3	EAM
Nenhuma	0,0000	8,3852
x_1	0,7197	3,8307
x_2	0,4505	6,0779
x_3	0,6669	4,3519
x_4	0,1695	7,6611
x_1, x_2	0,7284	3,8257
x_1, x_3	0,7921	3,1136
x_1, x_4	0,7282	3,8273
x_2, x_3	0,7073	4,0428
x_2, x_4	0,4846	5,9135
x_3, x_4	0,7345	3,7611
x_1, x_2, x_3	0,8019	3,0781
x_1, x_2, x_4	0,7368	3,8243
x_1, x_3, x_4	0,8160	2,9029
x_2, x_3, x_4	0,7655	3,5082
x_1, x_2, x_3, x_4	0,8242	2,8817

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



4. Coeficiente de determinação ajustado

Na regressão de mínimos quadrados, escolher como melhor modelo aquele que possui o maior valor de R^2 não é uma conduta adequada, uma vez que desta forma, seria sempre escolhido o modelo com todas variáveis explicativas. A utilização do coeficiente R^2 na escolha do melhor modelo 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 contribui pouco para o aumento do valor desse coeficiente. Uma forma alternativa de seleção de modelos, quando são consideradas todas as possíveis regressões, seria a utilização do coeficiente de determinação ajustado $\bar{R}_p^2$ (ver, por exemplo, Seber (1977)), dado por:

$$\bar{R}_p^2 = 1 - (1 - R_p^2) \left(\frac{n}{n-p} \right), \quad (4.1)$$

onde R_p^2 é o coeficiente dado em (1.1). Observemos que na expressão do coeficiente ajustado, é considerado o número de variáveis independentes no modelo, através do número de graus de liberdade do resíduo $(n-p)$. A estatística $\bar{R}_p^2$ não aumenta necessariamente quando novas variáveis preditoras são adicionadas ao modelo. Na verdade, sabe-se que, se q novas variáveis são consideradas, $\bar{R}_{p+q}^2$ é maior que $\bar{R}_p^2$ se e somente se a estatística F parcial para testar a contribuição adicional dessas variáveis for maior que 1 (Seber (1977)). É escolhido o modelo com maior $\bar{R}_p^2$.

Nosso objetivo nesta seção é propor um coeficiente de determinação que tenha na regressão MSAE papel análogo ao de $\bar{R}_p^2$ na regressão de mínimos quadrados. Sugerimos, então, a seguinte estatística:

$$\bar{R}_p^* = 1 - (1 - R_p^*) \left(\frac{n}{n-p} \right) \quad (4.2)$$

onde R_p^* é o coeficiente de determinação dado em (3.2).

Com o objetivo de estudar as propriedades desse coeficiente, lembremos inicialmente que a estatística para testar a contribuição adicional de q novas variáveis explicativas em relação a um modelo com $p-1$ variáveis preditoras na regressão MSAE é dada por (Schrader e McKean (1982)):

$$D_q = \frac{RSEA_{p+q} - RSEA_p}{\hat{\sigma}_{p+q}^2 / 2}, \quad (4.3)$$

$$\text{onde, } RSE_{p+q} = \sum_{i=1}^n |y_i - \hat{\beta}_0| - \sum_{i=1}^n |y_i - \hat{y}_{i,p+q}|,$$

$$\hat{\beta}_0 = \text{mediana } (y_1, \dots, y_n),$$

$\hat{y}_{i,p+q}$ é o valor previsto pelo modelo ajustado com $p+q-1$ variáveis explicativas,

RSE_p é definido de forma análoga à RSE_{p+q} para o modelo com $p-1$ variáveis explicativas e

$\hat{\tau}_{p+q}$ é um estimador consistente de τ em um modelo com $p+q-1$ variáveis explicativas.

Assumindo que os erros seguem distribuição de Laplace, vamos considerar nesta seção o estimador de máxima verossimilhança dado em (2.6).

Da definição do coeficiente de determinação R_3 dada em (2.5), temos que, em um modelo com $p+q-1$ variáveis explicativas, RSE_{p+q} pode ser escrito em função de R_3 como:

$$RSE_{p+q} = \frac{R_3^{p+q}}{1 - R_3^{p+q}} (n - p - q) \frac{\hat{\tau}_{p+q}}{2}. \quad (4.4)$$

De forma análoga podemos obter a expressão de RSE_p em função de R_3^p . Substituindo essas expressões em (4.3), temos que a estatística para o teste da contribuição adicional das q variáveis explicativas pode ser escrita em função dos coeficientes de determinação e dos estimadores do parâmetro de escala τ da seguinte forma:

$$D_q = \frac{\frac{R_3^{p+q}}{1 - R_3^{p+q}} (n - p - q) \frac{\hat{\tau}_{p+q}}{2} - \frac{R_3^p}{1 - R_3^p} (n - p) \frac{\hat{\tau}_p}{2}}{\frac{\hat{\tau}_{p+q}}{2}} \quad (4.5)$$

Desta relação segue-se que:

$$R_3^{p+q} = \frac{(1 - R_3^p) D_q \hat{\tau}_{p+q} + R_3^p (n - p) \hat{\tau}_p}{(1 - R_3^p) D_q \hat{\tau}_{p+q} + R_3^p (n - p) \hat{\tau}_p + (n - p - q) \hat{\tau}_{p+q} (1 - R_3^p)} \quad (4.6)$$

Considerando essa expressão e a definição do coeficiente de determinação ajustado dada em (4.2), obtemos que:

$$\begin{aligned} \bar{R}_3^{p+q} &= 1 - \frac{[(1 - R_3^p) \hat{\tau}_{p+q}] n}{(1 - R_3^p) \hat{\tau}_{p+q} (D_q + (n - p - q)) + (n - p) \hat{\tau}_p R_3^p} \\ &= 1 - (1 - R_3^p) \frac{n}{n - p} \cdot \frac{\hat{\tau}_{p+q} (n - p)}{(1 - R_3^p) \hat{\tau}_{p+q} (D_q - q + n - p) + (n - p) \hat{\tau}_p R_3^p} \end{aligned}$$

$$\begin{aligned}
&= 1 - (1 - \bar{R}_3^p) \frac{n}{n-p} \cdot \frac{\hat{\tau}_{p+q}(n-p)}{(1 - R_3^p) \hat{\tau}_{p+q}(D_q - q + n - p) + (n-p) \hat{\tau}_{p+q} R_3^p \left(\frac{D_q + 2n}{2n} \right)} \\
&= 1 - (1 - \bar{R}_3^p) \frac{n}{n-p} \cdot \frac{(n-p)}{(1 - R_3^p)(D_q - q + n - p) + (n-p) R_3^p \left(\frac{D_q}{2n} + 1 \right)} \quad (4.7)
\end{aligned}$$

Fazendo a diferença entre o denominador e o numerador (DIF) do termo que multiplica $[(1 - \bar{R}_3^p) n / (n-p)]$ em (4.7), obtemos que

$$DIF = D_q(1 - R_3^p) + \frac{D_q(n-p)}{2n} R_3^p + (1 - R_3^p)(-q).$$

Se $D_q > q$, então

$$\begin{aligned}
DIF &> (1 - R_3^p)q + \frac{(n-p)}{2n} q R_3^p + (1 - R_3^p)(-q) \\
&= \frac{n-p}{2n} q R_3^p > 0.
\end{aligned}$$

Portanto,

$$D_q > q \Rightarrow \bar{R}_3^{p+q} > \bar{R}_3^p,$$

ou seja,

$$\bar{R}_3^{p+q} \leq \bar{R}_3^p \Rightarrow D_q \leq q.$$

Portanto, se com a introdução de q variáveis explicativas no modelo observarmos um decréscimo no valor do coeficiente de determinação ajustado, então o valor da estatística para o teste da contribuição adicional dessas variáveis na explicação da variável resposta é não significativa. Isto ocorre pois, aos níveis de significância usuais, o valor crítico da distribuição χ^2 com q graus de liberdade é maior que q .

Na tabela 4.1 apresentamos os valores da estatística proposta observados nos possíveis modelos de regressão. Observamos, por exemplo, que os valores de $\bar{R}_3^p$ nos modelos com as variáveis $(x_1 \text{ e } x_4)$ e $(x_1, x_2 \text{ e } x_4)$ são menores que no modelo com a variável x_1 . Este fato implica que a contribuição adicional conjunta das variáveis x_2 e x_4 em relação à explicação obtida em um modelo que inclui apenas a variável x_1 não é significativa. Notamos também que os modelos com as variáveis $x_1 \text{ e } x_3$, $x_1, x_2 \text{ e } x_3$, $x_1, x_3 \text{ e } x_4$ apresentaram valores de $\bar{R}_3^p$ bastante próximos ao valor observado no modelo com todas as variáveis explicativas, e portanto, são possíveis candidatos a melhor modelo segundo este critério.

Tabela 4.1. Valores de $\bar{R}_3$ no exemplo dos imóveis.

Variáveis no Modelo	$\bar{R}_3$
Nenhuma	0,0000
x_1	0,6973
x_2	0,4066
x_3	0,6402
x_4	0,1031
x_1, x_2	0,6944
x_1, x_3	0,7661
x_1, x_4	0,6942
x_2, x_3	0,6707
x_2, x_4	0,4202
x_3, x_4	0,7013
x_1, x_2, x_3	0,7674
x_1, x_2, x_4	0,6911
x_1, x_3, x_4	0,7840
x_2, x_3, x_4	0,7247
x_1, x_2, x_3, x_4	0,7842

Bibliografia

Arnold, S.F. (1981). *The Theory of Linear Models and Multivariate Analysis*, John Wiley & Sons, New York.

Engelhardt, M. and Bain L.J. (1973). Interval Estimation for the Two - parameter Double Exponential Distribution. *Technometrics*, 15, 4, 875 - 887.

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

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

Schrader, R.M. and McKean , J. W. (1987). Least 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.

Seber , G.A.F.(1977). *Linear Regression Analysis*, John Wiley & Sons New York.

ÚLTIMOS RELATÓRIOS TÉCNICOS PUBLICADOS

- 9801 - GUIOL, H. Some properties of K-step exclusion processes. 1998. 21p. (RT-MAE-9801)
- 9802 - ARTES, R., JØRSENSEN, B. Longitudinal data estimating equations for dispersion models. 1998. 18p. (RT-MAE-9802)
- 9803 - BORGES, W.S.; HO, L.L. On capability index for non-normal processes. 1998. 13p. (RT-MAE-9803)
- 9804 - BRANCO, M.D. , BOLFARINE, H., IGLESIAS, P., ARELLANO-VALLE, R. Bayesian analysis of the calibration problem under elliptical distributions. 1998. 15p. (RT-MAE-9804)
- 9805 - IGLESIAS, P., MATÓS, F., PEREIRA, C.A.B., TANAKA, N.I. On finite sequences conditionally uniform given minima and maxima. 1998. 13p. (RT-MAE-9805)
- 9806 - IGLESIAS, P., PEREIRA, C.A.B., TANAKA, N.I. Characterizations of multivariate spherical distribution in L_2 norm. 1998. 19p. (RT-MAE-9806)
- 9807 - GIAMPAOLI, V., SINGER, J.M. A note on the comparison of two normal populations with restricted means. 1998. 12p. (RT-MAE-9807)
- 9808 - BUENO, V.C. A note on conditional reliability importance of components. 1998. 07p. (RT-MAE-9808)
- 9809 - GIMENEZ, P., BOLFARINE, H. Consistent estimation in comparative calibration models. 1998. 19p. (RT-MAE-9809)
- 9810 - ARTES, R., PAULA, G.A., RANVAUD, R. Estudo da direção tomada por pombos através de equações de estimação para dados circulares longitudinais. 1998. 16p. (RT-MAE-9810)
- 9811 - LIMA, A.C.P., SEM, P.K. Time-Dependent Coefficients in a Multi-Event Model for Survival Analysis. 1998. 21p. (RT-MAE-9811)

- 9812 - KOLEV, N., MINKOVA, L. Over-and underdispersed models for ruin probabilities. 1998. 34p. (RT-MAE-9812)
- 9813 - KOLEV, N. Negative Binomial Cross-Classifications. 1998. 16p. (RT-MAE-9813)
- 9814- BAKEVA, V., GEORGIEVA, M., KOLEV, N. Two Characterizations of the Geometric Distribution Related to an Unreliable Geo/ $G_0/1$ System. 1998. 11p. (RT-MAE-9814)
- 9815 - KOLEV, N., MINKOVA, L. Some basic Inflated-Parameter Discrete Distributions. 1998. 19p. (RT-MAE-9815)
- 9816 - ARTES, R., PAULA, G.A., RANVAUD, R. Strategy for the analysis of circular longitudinal data based on generalized estimating equations. 1998. 15p. (RT-MAE-9816)
- 9817 - PAULA, G.A., ARTES, R. A Score-Type Test to Assess Overdispersion in Linear Logistic Models. 1998. 17p. (RT-MAE-9817)
- 9818 - LOSCHI, R.H., IGLESIAS, P.L., WECHSLER, S. Unpredictability and Probability Updating. 1998. 11p. (RT-MAE-9818)
- 9819 - KOLEV, N., MINKOVA, L. New Over-/Under- Dispersed Class of Inflated-Parameter Discrete Probability Distributions. 1998. 10p. (RT-MAE-9819)
- 9820 - KOLEV, N., MINKOVA, L. Inflated-Parameter Family of Generalized Power Series Distributions. 1998. 11p. (RT-MAE-9820)
- 9821 - AUBIN, E.C.Q., STANGENHAUS, G. A Robust Estimation of Growth Curves. 1998. 7p. (RT-MAE-9821)
- 9822 - BORGES, W., HO, L. L., A Fraction Non-Conforming Based Capability Index. 1998. 13p. (RT-MAE-9822)
- 9823 - GALEA-ROJAS, M., BOLFARINE, H. Hypothesis Testing in Functional Comparative Calibration Models. 1998. 13p. (RT-MAE-9823)
- 9824 - FONTES, L. R. G., ISOPI, M., NEWMAN, C. M. Chaotic Time Dependence in a Disordered Spin System. 1998. 23p. (RT-MAE-9824)

- 9825 - FERRARI, P. A., FONTES, L. R. G., VARES, M. E. The Asymmetric Simple Exclusion Model With Multiple Shocks. 1998. 24p. (RT-MAE-9825)
- 9826 - CABRAL, C. R. B., PAULA, G. A., BOLFARINE, H. One-Sided Tests for the Analysis of Pretest-Posttest Data. 1998. 12p. (RT-MAE-9826)
- 9827 - GASCO, L., BOLFARINE, H., ARELLANO-VALLE, R. B. Asymptotic properties of the sample mean vector and sample covariance matrix with application in elliptical multiplicative measurement error models. 1998. 17p. (RT-MAE-9827)
- 9828 - FERRARI, S. L. P., CRIBARI-NETO, F. On the robustness of analytical and bootstrap corrections to score tests in regression models. 1998. 11p. (RT-MAE-9828)
- 9829 - PERES, C. A., ANDRE, C. D. S., NARULA, S. C. Whyn the observations within a treatment of a random-effects model are correlated? 1998. 8p. (RT-MAE-9829)
- 9830 - FONTES, L.R., MATHIEU, P., PICCO, P. On the averaged dynamics of the random field curie-weiss model. 1998. 27p. (RT-MAE-9830)

The complete list of "Relatórios do Departamento de Estatística", IME-USP, will be sent upon request.

Departamento de Estatística
IME-USP
Caixa Postal 66.281
05315-970 - São Paulo, Brasil