Universidade de Brasília
Departamento de Estatística
EST0077 - 2023/1

SOLUÇÃO DOS EXERCÍCIOS PARA ENTREGA 4

Igor de Oliveira Barros Faluhelyi
Prof. George

Brasília
2023

Sumário

1 Exercício 29 da Lista 5

Enunciado:

Discuta sobre possíveis vantagens e desvantagens das medidas MSE (Erro Quadrático Médio) e MAE (Erro Absoluto Médio) nas seguintes condições:

(a) Regressão Linear em que $n >>> p$ (Ex. n = 5000 e p = 5)

(b) Regressão Linear em que $p \approx n$ (Ex. n = 150 e p = 100)

Resposta:

Em (a) temos n = 5000 e k = 5, em (b) temos n = 150 e k = 100:

$$QM_{reg} = SQ_{reg}/k$$

$$QM_{res} = SQ_{res}/(n-k-1)$$

$$SQ_{res} = n\dot{M}SE$$

Sob H_0 no teste F:

$$F_0 = QM_{reg}/QM_{res}$$

$$F_0 \sim F_{k,n-k-1}$$

$$F_0 \sim F_{k,n-k-1}$$

$$\mathbb{P}(F_{k,n-k-1} > F_c)$$

E, rejeita-se H_0 à α se $F_0 > F_c$.

$$H_0 : \beta_1 = \ldots = \beta_k = 0$$

Ou seja, Estatísticas F grandes querem dizer que ao menos uma variável explanatória está relacionada com a variável resposta. No contexto com n grande, a Estatística F que é pouco maior que 1 já indica evidência contra H_0.

Essa abordagem usando a Estatística F para fazer o teste de associação entre preditores e a resposta funciona quando k é relativamente pequeno comparado à n.

2 Exercício 30 da Lista 5

Enunciado:

Defina as seguintes medidas para seleção de modelos e discuta sobre vantagens e desvantagens.

(a) AIC - Critério de Informação de Akaike

(b) BIC - Critério de Informação Bayesiano

(c) Medida Cp de Mallows

Resposta:

O critério AIC é definido para uma grande classe de modelos ajustados por máxima verossimilhança. No caso do modelo com erros gaussianos, máxima verossimilhança e mínimos quadrados são a mesma coisa. Neste caso, o AIC é dado por:

$$MSE = RSS/n$$

$$AIC = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

Para um modelo de mínimos quadrados ajustado contendo d preditores, a estimativa de C_p é calculado usando a equação:

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

Observa-se que, nesse caso, Cp e AIC são proporcionais entre si. Em que $\hat{\sigma}^2$ é a estimativa da variância do erro ε. Essencialmente, a estatística Cp acrescenta uma penalidade de $2d\hat{\sigma}^2$ para o RSS calculado nos dados de treino, a fim de ajustar para o fato de que o treinamento tende a subestimar o erro do teste. A penalidade aumenta à medida que o número de preditores no modelo aumenta. Cp é não viesado para estimar MSE nos dados de teste, como consequência, a estatística Cp tende a assumir um valor pequeno para modelos com baixo erro de teste, por isso, ao determinar qual dos modelos é o melhor, escolhemos o modelo com o menor valor de Cp.

Para o modelo de mínimos quadrados com d preditores, o BIC é dado por:

$$BIC = \frac{1}{n}(RSS + log(n)d\hat{\sigma}^2)$$

Como Cp, o BIC tenderá a assumir um pequeno valor para um modelo com um

baixo erro nos dados de teste, e por isso geralmente selecionamos o modelo que tem o menor valor para seu BIC.

Observe que o BIC substitui o $2d\hat{\sigma}^2$ usado por Cp por $log(n)d\hat{\sigma}^2$, em que n é o número de observações. Desde que $logn > 2$ para qualquer $n > 7$ (log na base e), a estatística BIC geralmente coloca uma penalidade mais pesada em modelos com muitas variáveis e, portanto, resulta na seleção de modelos mais parcimoniosos que o Cp.

3 Exercício 31 da Lista 5

Enunciado:

I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$.

(a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \varepsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(b) Answer (a) using test rather than training RSS.

(c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer

(d) Answer (c) using test rather than training RSS

Resposta:

(a) Os termos extras do polinômio permitem um ajuste mais próximo (mais graus de liberdade) aos dados de treinamento, então é esperado que o RSS nos dados de treinamento para regressão cúbica fosse menor do que para regressão linear simples.

(b) A relação verdadeira é linear e, portanto, a regressão linear simples generalizaria melhor para dados não vistos, como tal é esperado que tivesse um RSS menor nos dados de teste. O modelo cúbico provavelmente se ajustou demais aos dados de treinamento e, portanto, é esperado que ele tivesse um RSS mais alto nos dados de teste.

(c) A regressão cúbica terá um melhor ajuste aos dados não-lineares e, portanto, seu RSS nos dados de treinamento será menor.

(d) O RSS nos dados de teste depende de quão longe de linear é a relação verdadeira dada por $f(x)$. Se $f(x)$ é mais linear do que cúbica, então a regressão cúbica pode sofrer *overfitting*, então o RSS cúbico será maior e o RSS linear será menor. Se $f(x)$ é mais cúbico do que linear, então a regressão linear pode não ser bem ajustada, então o RSS linear será maior e o RSS cúbico será menor.

4 Exercício 32 da Lista 5

Resposta:

(a), (b) e (c):

Código .R em anexo.

$Y = 2 + 2x_1 + 0.3x_2 + \epsilon$, em que ϵ é N(0,1). $\beta_0 = 2$, $\beta_1 = 2$, $\beta_2 = 0.3$.

Coeficientes da regressão: $\hat{\beta}_0 = 2.04$, $\hat{\beta}_1 = 2.85$, $\hat{\beta}_2 = -1.04$. O Std. Error é alto e aumenta desde x1 para o x2, ainda, seus respectivos coeficientes estimados $\hat{\beta}_1$ and $\hat{\beta}_2$ são imprecisos quando comparados ao $\beta_1 = 2$ e $\beta_2 = 0.3$ .O p-value para $\hat{\beta}_1$ está abaixo de 0.05 e por isso podemos rejeitar $H_0 : \hat{\beta}_1 = 0$. O p-value para $\hat{\beta}_2$ é muito superior a 0,05 e, portanto, não podemos rejeitar $H_0 : \hat{\beta}_2 = 0$.

(d)

RSE e R2 são similares como quando foi utilizado x1+x2, F-statistic ficou um pouco maior.

P-value perto de zero para x1, então $H_0 : \hat{\beta}_1 = 0$ pode ser rejeitada.

(e)

P-value perto de zero para x2, então $H_0 : \hat{\beta}_2 = 0$ pode ser rejeitada.

(f)

Não, eles não se contradizem, pois a diferença entre (c) e (e) pode ser explicada pelo problema da colinearidade. Ao usar duas variáveis que são altamente colineares, o efeito sobre a resposta de uma variável pode ser mascarado por outra. A colinearidade também faz com que o erro padrão aumente - como pode ser visto o std. erro de x1+x2 é maior que x1 ou x2.

(g)

Pela Figura 2, considerando x1+x2 no modelo, O ponto 101 é de alavanca, mas não é um outlier.

Pela Figura 3, considerando apenas x1 no modelo, o ponto 101 é um outlier.

Pela Figura 4, considerando apenas x2 no modelo, o ponto 101 é de alavanca.

Figura 1: Observa-se que o novo ponto (101) aparentemente é um outlier. Vamos investigar.

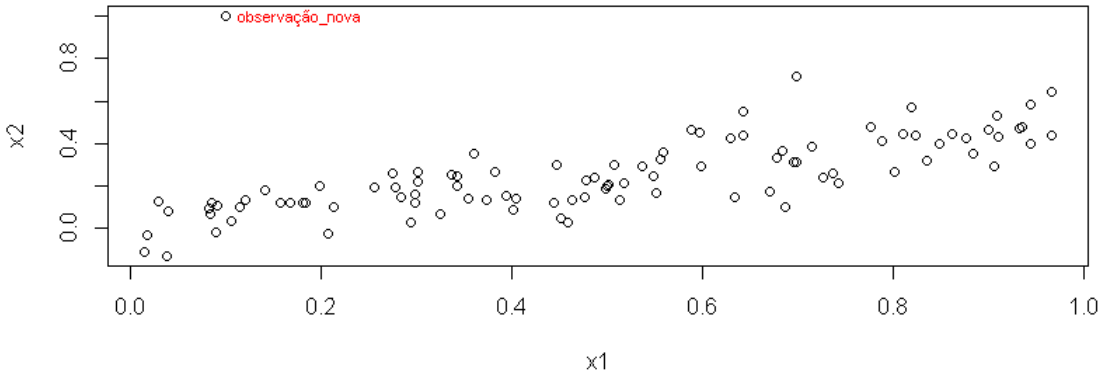

Figura 2: x1+x2

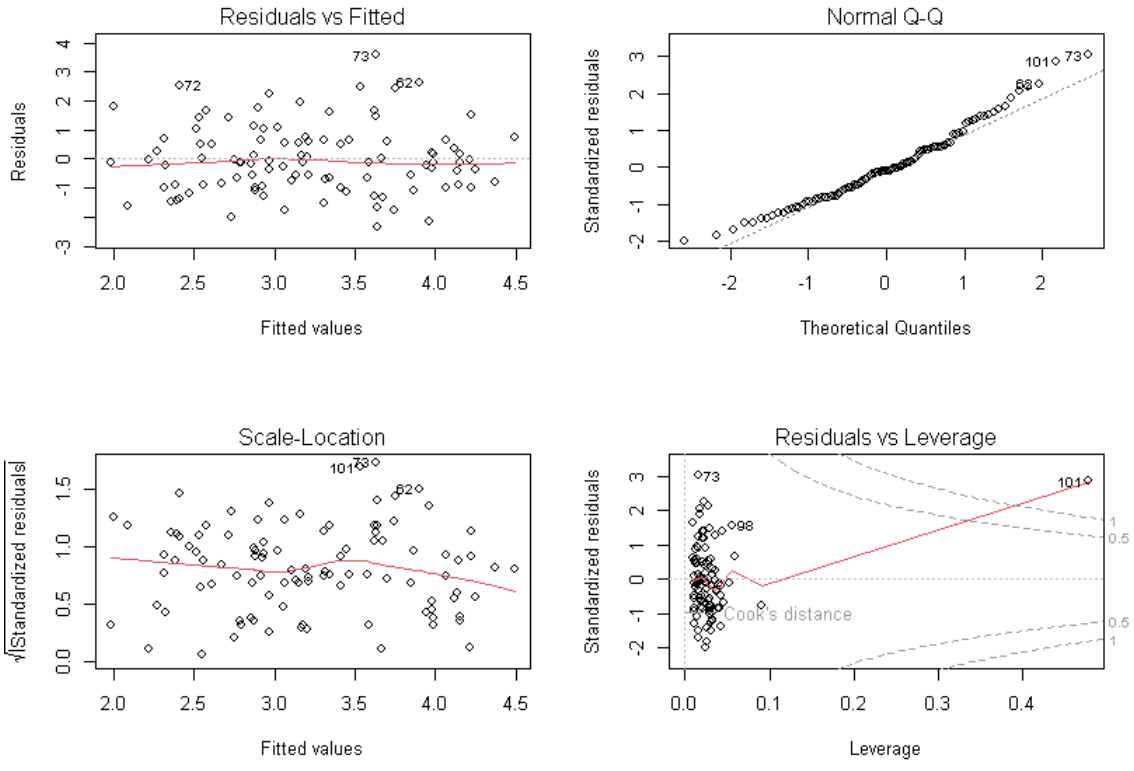

Figura 3: x1

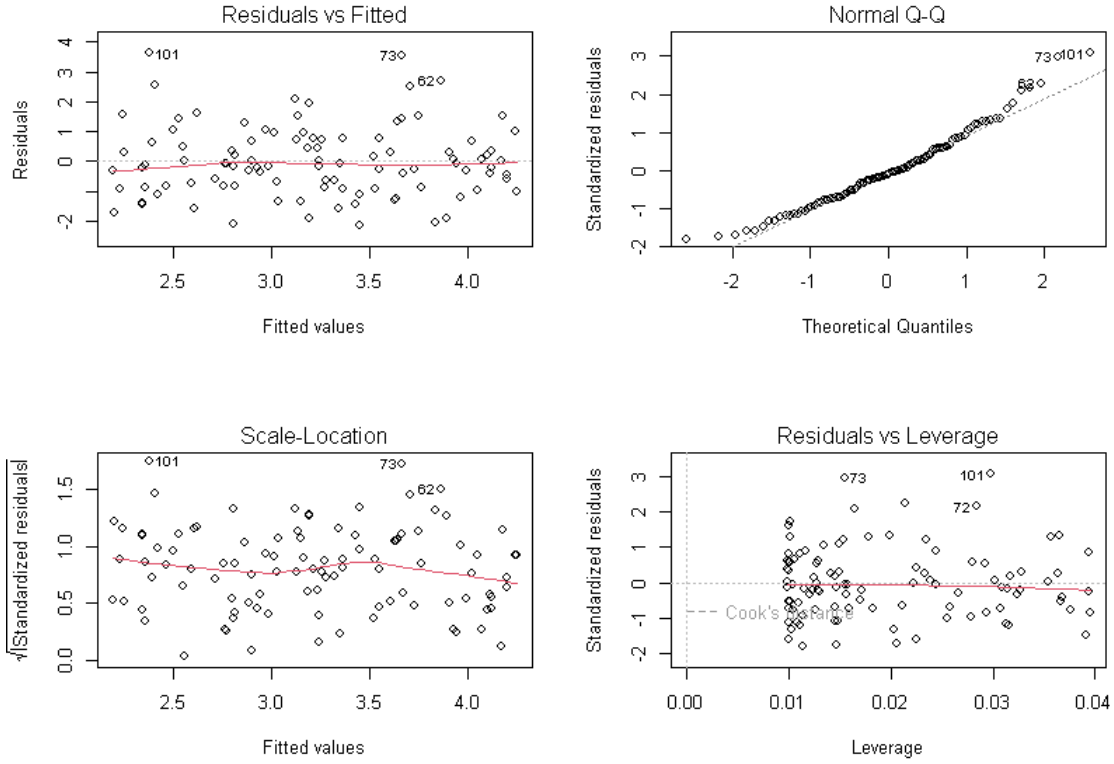

Figura 4: x2

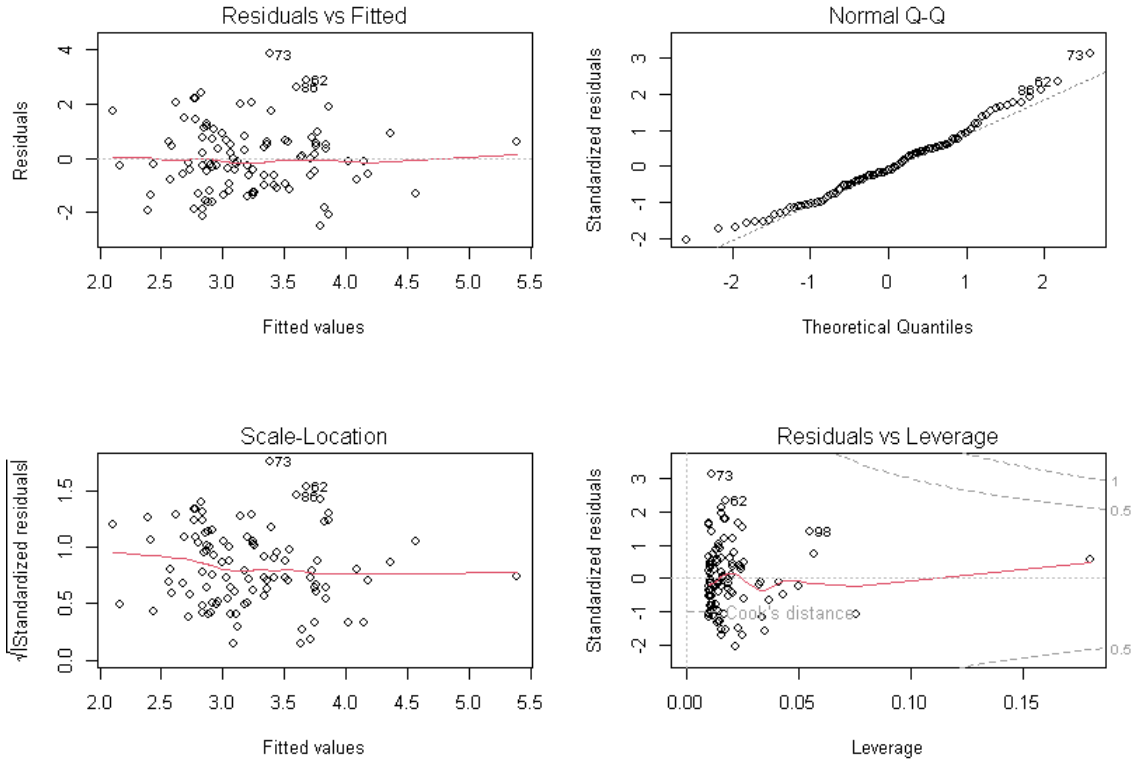

Figura 5: Cor(x1,x2) = 0.79