Avaliação 1

Benilton Carvalho

Entrega: 6 de maio de 2020 (até ao meio-dia: via Moodle)

Instruções

- Esta avaliação deve ser realizada individualmente, sem troca de informações com colegas.
- Medidas cabíveis serão tomadas caso sejam identificados casos de compartilhamento de soluções entre colegas.
- O aluno poderá consultar todo o material disponibilizado, incluindo os livros apontados nas referências.
- A data/hora da entrega deverá ser seguida estritamente e materiais entregues depois do prazo não serão aceitos.
- Caso o aluno opte por uma solução manuscrita, esta solução deve estar completamente legível (tanto pela caligrafia do aluno, quanto pelo método de captura escolhido).
- Todo código que venha a ser utilizado para a resolução deve também ser submetido. Todos os códigos serão analisados pelo professor em sua acurácia na resolução e, também, em sua reprodutibilidade.
- Códigos com alta similaridade entre si (ao comparar soluções de diferentes alunos) serão analisados sob a possibilidade de serem classificados como plágio.

Questão 1 - Para o modelo $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$:

a. Derive manualmente os estimadores de mínimos quadrados (MQ) para os parâmetros β_0 e β_1 para o modelo abaixo:

b. Para o modelo acima, apresente os estimadores de máxima verossimilhança (MV). Que condições são necessárias para que os estimadores de MV sejam idênticos aos de MQ?

c. Apresente analiticamente os estimadores para a variância residual e para as variâncias dos parâmetros da reta.

Questão 2 - Responda Verdadeiro ou Falso, apresentando justificativas e contra-exemplos para todos os itens.

a. Um coeficiente de determinação, R^2, igual a zero indica que as variáveis preditora e resposta não são associadas.
b. Um coeficiente de determinação, R^2, muito alto indica que o modelo é adequado.
c. Em um modelo de regressão linear simples, como o da Questão 1, a reta de regressão sempre passa pelo ponto $(\bar{X}, \bar{Y})$. Apresente uma prova analítica para a sua resposta.
d. Em um modelo de regressão linear simples, como o da Questão 1, centralizar a variáveis resposta e preditora é suficiente para que o estimador de β_1 seja o coeficiente de correlação de Pearson. Apresente uma prova analítica para a sua resposta.
e. Em um modelo de regressão linear simples, como o da Questão 1, se o p-valor associado ao parâmetro β_1 for p, então o p-valor do teste F da mesma regressão será exatamente o mesmo valor. Apresente uma prova analítica para a sua resposta.

f. Conjuntos de dados diferentes, tendo sempre o mesmo número de observações (por exemplo, $n = 100$), podem apresentar diferentes graus de liberdade para a variância residual.
g. Mudar a escala dos dados (por exemplo, de milhas para quilômetros), a significância da regressão (medida pelo teste F) também muda.

Questão 3 - Interpretação de Modelo de Regressão

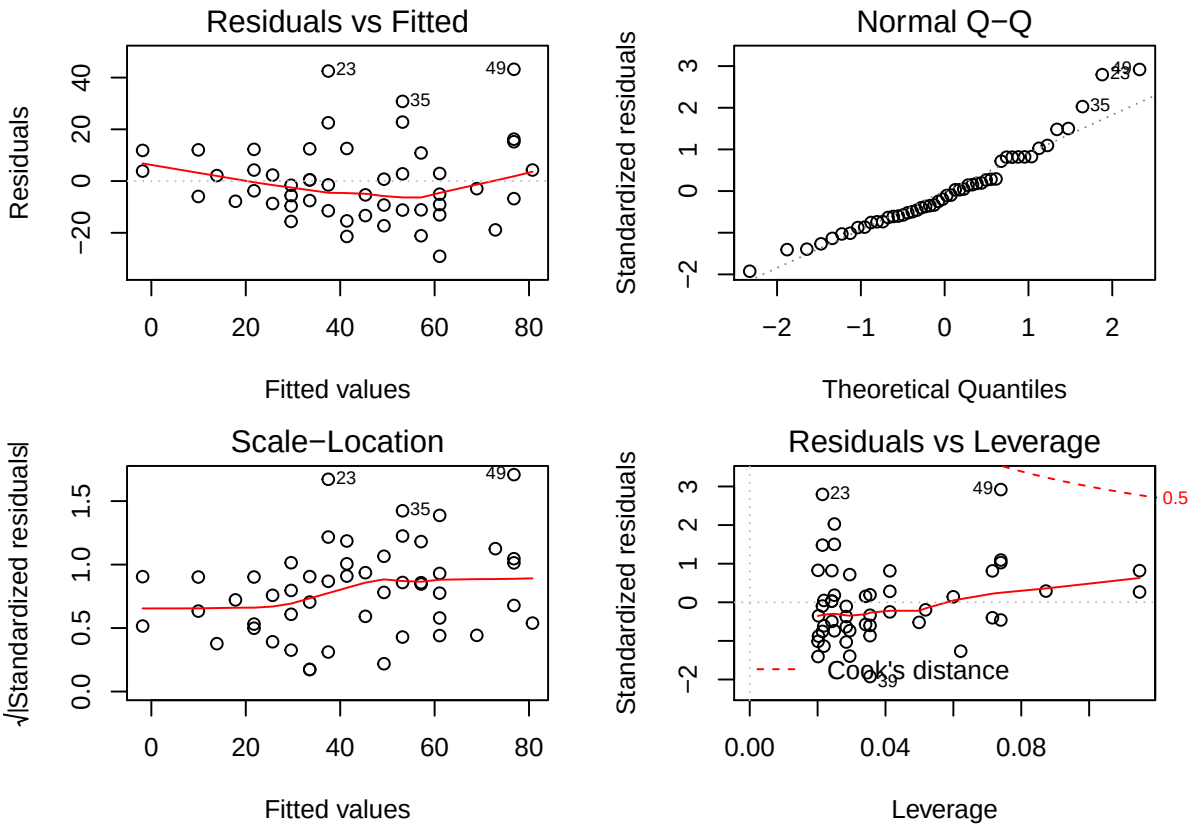
Um modelo de regressão linear simples foi ajustado com o R, conforme apresentado abaixo:

```
data(cars)
fit = lm(dist ~ speed, data=cars)
summary(fit)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Responda às seguintes perguntas, apresentando justificativas em todos os casos:

a. Qual é o valor do teste-t para o coeficiente da variável `speed`?
b. Utilizando um tamanho de teste de 1%, podemos dizer este coeficiente é igual a zero?
c. Demonstre, analiticamente, como calcular um estimador para a variância residual. Qual é esta estimativa? Coincide com aquela apresentada pelo sumário do modelo?
d. Demonstre, analiticamente, como calcular o intervalo de confiança para o intercepto. Apresente o IC de 95% de confiança para este parâmetro.
e. Demonstra, analiticamente, como calcular o intervalo de confiança para o coeficiente angular. Apresente o IC de 95% de confiança para este parâmetro.
f. Interprete cada um dos parâmetros ajustados.
g. Analisando criticamente os gráficos de diagnóstico apresentados abaixo, qual é o seu parecer acerca da qualidade do modelo proposto? Quais são as hipóteses acerca dos resíduos que devem ser analisadas? Que pontos devem ser tratados imediatamente com o objetivo de melhorá-lo? (ignore o gráfico de resíduos vs. *leverage*)



Questão 4 - Análise de Dados

O conjunto de dados `cancer.xlsx` possui três colunas:

- TYPE: tipo de câncer;
- LSCD: número total de divisões de células-tronco ao longo da vida;
- RISK: risco, em toda a vida, de desenvolver câncer.

a. Suspeita-se que o número de divisões celulares possa se associar ao risco de desenvolvimento de câncer. Realize uma análise de dados, incluindo descritiva, culminando na proposição de um modelo de regressão que mostre a existência (ou não) de associação entre estas duas variáveis (risco de câncer deve ser a variável resposta). Apresente gráficos, intervalos de confiança, testes de hipótese e qualquer outro recurso estatístico para justificar suas decisões.

b. Leia a reportagem da BBC e escreva um parecer técnico a respeito da reportagem.