3. Regressão linear múltipla

3.1. A *Vend* dedica-se à comercialização, exploração e manutenção de equipamentos de venda automática, denominada (*vending*). Um fator importante neste negócio é o tempo que os motoristas demoram a reabastecer as máquinas e a realizar pequenas tarefas de manutenção e limpeza. Um engenheiro industrial desta empresa, sugeriu que as variáveis mais relevantes para explicar o tempo na prestação do serviço às máquinas de venda automática de refrigerantes num aeroporto (y), em minutos, são o número de caixas de produto reabastecidas (x_1) e a distância percorrida a pé pelo motorista (x_2), em metros. Para estudar esta relação, foram recolhidas 25 observações disponíveis no ficheiro *vending.csv*.

a) Desenhe as nuvens de pontos para cada par das 3 variáveis observadas. Comente.
b) Calcule a matriz de correlações entre todos os pares de variáveis observadas. Comente.
c) Ajuste o modelo de regressão linear múltipla resultante de modelar o tempo com base nos dois preditores disponíveis. Interprete os coeficientes de regressão.
d) Teste, ao nível de significância de 5%, a hipótese de que os coeficientes de regressão das variáveis explicativas são conjuntamente nulos.
e) Para cada variável explicativa, teste, ao nível de significância de 5%, a hipótese de que o respetivo coeficiente de regressão é nulo.
f) Construa intervalos de confiança a 90% para cada um dos coeficientes de regressão.
g) Qual o valor estimado para o tempo necessário na prestação de serviço à máquina nas seguintes situações:

Número de caixas	3	10	20
Distância (em metros)	100	500	1000

h) Com base em cada um dos pares de observações anteriores, obtenha:
   i) Um intervalo de confiança (95%) para o valor esperado para o tempo necessário na prestação de serviço associado a esses valores das variáveis preditoras.
   ii) Um intervalo de predição (95%) para o tempo necessário na prestação de serviço individual.
i) Ajuste os modelos de regressão lineares simples para modelar o tempo com base no número de caixas, e o tempo com base na distância. Compare os coeficientes destes modelos com os obtidos com o modelo de regressão linear múltipla.
j) No modelo de regressão múltipla ajustado em c), qual a variável com maior contribuição relativa para o modelo?

3.2. Considere o conjunto de dados `marketing` disponível no package `datarium` do R, que contém informação sobre os valores gastos em anúncios de um certo produto em três tipos de medias e o impacto nas vendas desse produto:

- `youtube` – quantia gasta em anúncios no Youtube;
- `facebook` - quantia gasta em anúncios no Facebook;
- `newspaper` - quantia gasta em anúncios em jornais;
- `sales` – montante das vendas.

a) Ajuste o modelo de regressão linear considerando apenas os efeitos principais.
b) Teste a significância do modelo geral.
c) Teste a significância de cada um dos coeficientes.

d) Com base nos resultados anteriores, ajuste o modelo que lhe parecer mais adequado.
e) Verifique se obteria um modelo diferente, caso tivesse usado os métodos sequenciais: *forward*, *backward* e *stepwise*.
f) Considere o modelo ajustado em d).
   i) Obtenha intervalos de confiança a 95% para os coeficientes e interprete.
   ii) Qual a variável com maior contribuição relativa para o modelo?
   iii) Qual o valor previsto para as vendas se forem gastos 100 milhares de dólares em anúncios no youtube, 50 milhares de dólares no Facebook e 80 milhares de dólares em jornais?
   iv) Verifique se a interação entre os preditores facebook e youtube é significativa.
   v) Valide os pressupostos do modelo.
g) Caso o modelo ajustado em d) não satisfaça os pressupostos, tente encontrar um modelo alternativo. Experimente:
   i) Transformar as variáveis (resposta, preditoras ou ambas).
   ii) Incluir de termos polinomiais e/ou interações.
   iii) Remover pontos influentes, caso existam.
h) Caso em g) tenha optado por um modelo com variáveis transformadas e/ou interações, interprete o modelo obtido.

3.3. Uma empresa de tecnologia pretende prever a taxa de retenção de utilizadores de uma aplicação móvel (*app*). Foram analisados dados de 27 utilizadores, medindo-se as seguintes variáveis relativas à interação dos utilizadores com a *app*:

- Taxa de retenção (Ret), em %: proporção de dias em que o utilizador acedeu à app durante um período de 30 dias (y),
- Tempo médio de sessão (TMS), em minutos,
- Número médio de notificações recebidas por dia (Notif),
- Número de funcionalidades utilizadas regularmente (Func),
- Taxa de interação com conteúdo personalizado (Interac), em %.

	Ret	TMS	Notif	Fuc	Interac
Ret	1.00000	-0.08618	-0.05741	0.17778	0.55346
TMS	-0.08618	1.00000	0.33811	0.20362	0.30437
Notif	-0.05741	0.33811	1.00000	0.08943	0.38161
Fuc	0.17778	0.20362	0.08943	1.00000	0.64198
Interac	0.55346	0.30437	0.38161	0.64198	1.00000

Foi ajustado um modelo de regressão linear múltipla utilizando todas as concentrações de catiões como preditores. Obtiveram-se os seguintes resultados:

```
Call: lm(formula = Ret ~ TMS + Notif + Fuc + Interact, data = app)
Coefficients:
            Estimate  Std. Error   t value  Pr(>|t|)
(Intercept)  -815.76      446.45    -1.827   0.08126 .
TMS          -102.21       84.52    -1.209   0.23935
Notif        -405.78      207.13    -1.959   0.06291 .
Fu            -57.68       29.61    -1.948   0.06432 .
Interact      668.17      145.34     4.597   0.00014 ***
---
Residual standard error: 68.07 on 22 degrees of freedom
Multiple R-squared: 0.5147, Adjusted R-squared: 0.4264
F-statistic: 5.832 on 4 and 22 DF, p-value: 0.002344
```

a) Teste o ajustamento global do modelo. Discuta a qualidade desse ajustamento, tendo também em conta os valores dos coeficientes de determinação usual e modificado.
b) Interprete, no contexto do problema sob estudo, o significado do valor −57.68 na primeira coluna da tabela, indicando as unidades de medida do referido valor.
c) Um analista afirma que quando aumenta o número médio de funcionalidades utilizadas regularmente melhora a retenção dos utilizadores. Admitindo a validade do modelo, e dando o ónus da prova à hipótese do investigador, qual a conclusão a que se pode chegar com base nos dados ($\alpha = 0.05$)?
d) Foi afirmado que com base nos resultados do modelo acima ajustado, a única de entre as quatro variáveis preditoras que é importante na modelação da retenção dos utilizadores é a variável Interact. Sem fazer quaisquer contas, diga se considera legítima esta afirmação.

3.4. Considere o conjunto de dados `mtcars` disponível no package R, que contém informação sobre o consumo de combustível (`mpg`) e dez características do design e desempenho de 32 automóveis.:

- `mpg` - consumo medido em milhas/(EUA) galão,
- `cil` - número de cilindros,
- `disp` - deslocamento (cu.in.),
- `cv` - potência bruta,
- `drat` - relação do eixo traseiro,
- `wt` - peso (1000 libras),
- `qsec` - tempo em 1/4 milha,
- `vs` - motor (0 = em forma de V, 1 = recto),
- `am` - transmissão (0 = automática, 1 = manual),
- `gear` - número de velocidades para a frente,
- `carb` - número de carburadores.

a) Defina as variáveis `vs` e `am` como fatores.
b) Utilizando uma estratégia sequencial ajuste um modelo de regressão linear múltiplo considerando apenas os.
c) Teste a significância do modelo geral.
d) Teste a significância de cada um dos coeficientes.
e) Avalie a necessidade de incluir termos de interação de 2ª ordem entre as variáveis presentes no modelo.
f) Valide os pressupostos do modelo.
g) Assumindo que o modelo ajustado é adequado,
   i) Obtenha intervalos de confiança a 95% para os coeficientes e interprete.
   ii) Interprete o modelo obtido.