Regressão Linear Múltipla

Larysa Mendes, Vitória Maria, Welly Remígio

2024-10-02

# Setup para o relatório Quarto  
  
knitr::opts\_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)  
  
# Definindo o espelho do CRAN  
options(repos = c(CRAN = "https://cloud.r-project.org/"))

# Introdução

Este relatório tem por objetivo ajustar um modelo de regressão linear múltiplo com o intuito de investigar a influência de determinadas características associadas a vendas de cadeirinhas de carro para crianças e diversos fatores que podem influenciar essas vendas.

Neste contexto a regressão será realizada sobre base de dados Carseats, que trata das vendas de cadeirinhas de carro para crianças (Child Car Seats) em diferentes locais. Essa base está na biblioteca ISLR (An Introduction to Statistical Learning with Applications in R) que é um pacote em R que acompanha o famoso livro “An Introduction to Statistical Learning” (ISLR). Ela contém vários conjuntos de dados usados para demonstrar técnicas de aprendizado estatístico e machine learning. Essa base de dados contém onze variáveis, sendo três destas quanlitativas, as demais são medidas quantitativas.

# Os dados

É possível baixar os dados do livro [An Introduction to Statistical Learning with applications in R](https://www.statlearning.com/), mas o pacote ISLR pode ser baixado diretamente do *R*, para acessar a base de dados, Carseats.

install.packages("ISLR", quiet=TRUE) #instalando pacote

package 'ISLR' successfully unpacked and MD5 sums checked

library(ISLR) #chamando pacote  
  
setwd("C:\\Users\\Notebook\\Documents\\Relatorio\_Estatistica")  
write.table(Carseats , file="nome\_do\_arquivo.csv", sep=";", dec=",") #criando arquivo .csv no local de   
#help("Carseats") #descrição das variáveis  
#glimpse(Carseats)

Descrição da Base de Dados: Número de observações: 400 Variável Resposta: - Sales: Vendas de cadeirinhas de carro em diferentes locais. (Milhares de unidades)

Variáveis Explicativas: - CompPrice: Preço da cadeirinha na loja concorrente. (em dólares).  
- Income: Renda média dos consumidores naquela região. (em milhares de dólares).  
- Advertising: Valor gasto em publicidade para aquela região (em milhares de dólares).  
- Population: População da região. (em milhares).  
- Price: Preço da cadeirinha de carro na loja.(em dólares).  
- ShelveLoc: Qualidade da localização da prateleira na loja. (“Good”, “Medium”, “Bad”).  
- Age: Idade média da população na região. (em anos).  
- Education: Nível médio de educação da população na região.(em anos).  
- Urban: Um fator que indica se a região é urbana. (“Yes” ou “No”).  
- US: Um fator que indica se a loja está nos EUA. (“Yes” ou “No”).

## Análise exploratória dos dados

library(skimr)  
  
dados <- Carseats  
  
skim(dados)

Data summary

|  |  |
| --- | --- |
| Name | dados |
| Number of rows | 400 |
| Number of columns | 11 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 3 |
| numeric | 8 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| ShelveLoc | 0 | 1 | FALSE | 3 | Med: 219, Bad: 96, Goo: 85 |
| Urban | 0 | 1 | FALSE | 2 | Yes: 282, No: 118 |
| US | 0 | 1 | FALSE | 2 | Yes: 258, No: 142 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sales | 0 | 1 | 7.50 | 2.82 | 0 | 5.39 | 7.49 | 9.32 | 16.27 | ▁▆▇▃▁ |
| CompPrice | 0 | 1 | 124.97 | 15.33 | 77 | 115.00 | 125.00 | 135.00 | 175.00 | ▁▅▇▃▁ |
| Income | 0 | 1 | 68.66 | 27.99 | 21 | 42.75 | 69.00 | 91.00 | 120.00 | ▇▆▇▆▅ |
| Advertising | 0 | 1 | 6.64 | 6.65 | 0 | 0.00 | 5.00 | 12.00 | 29.00 | ▇▃▃▁▁ |
| Population | 0 | 1 | 264.84 | 147.38 | 10 | 139.00 | 272.00 | 398.50 | 509.00 | ▇▇▇▇▇ |
| Price | 0 | 1 | 115.80 | 23.68 | 24 | 100.00 | 117.00 | 131.00 | 191.00 | ▁▂▇▆▁ |
| Age | 0 | 1 | 53.32 | 16.20 | 25 | 39.75 | 54.50 | 66.00 | 80.00 | ▇▆▇▇▇ |
| Education | 0 | 1 | 13.90 | 2.62 | 10 | 12.00 | 14.00 | 16.00 | 18.00 | ▇▇▃▇▇ |

### Comentários:

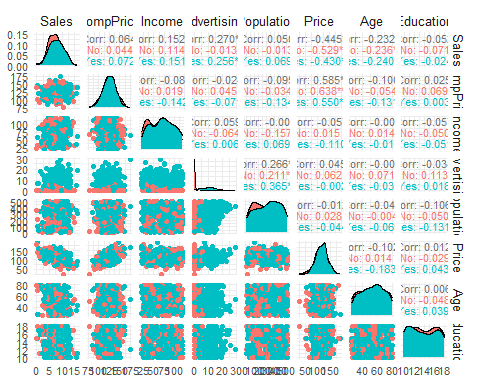
De acordo com a tabela é possível observar que não há ausência de dados. # INCOMPLETO

### Análise de Correlação

install.packages("GGally", quite=TRUE)

package 'GGally' successfully unpacked and MD5 sums checked  
  
The downloaded binary packages are in  
 C:\Users\Notebook\AppData\Local\Temp\RtmpeuJ7zI\downloaded\_packages

library(GGally)# Pacote para função ggpairs  
library(ggplot2)  
  
  
### Multicolinearidade: r > 0.9 (ou 0.8)  
  
graf1 <- ggpairs(dados,   
 columns = c("Sales", "CompPrice", "Income", "Advertising", "Population", "Price", "Age", "Education"),   
 ggplot2::aes(colour=US),   
 upper = list(continuous = wrap("cor", size = 3)),   
 lower = list(continuous = "points")) +   
 theme\_minimal() +   
 theme(axis.text = element\_text(size = 8),   
 legend.position = "bottom",   
 strip.text = element\_text(size = 10))   
  
graf1



# Salvando o gráfico em .jpeg  
ggsave("Grafico\_dispersao\_carseats.jpeg")

#### Comentários

Com relação à análise de correlação é algo desejável observar altas correlações das variáveis independentes com relação à variável dependente que no presente caso é Sales.

Por outro lado, altas correlações entre as demais variáveis a serem utilizadas como variáveis independentes nos dá indícios de que haverá **problemas de multicolinearidade** ao ajustar o MRLM. **Como regra geral** isto ocorre quando há **correlações** ou **entre** as **variáveis preditoras**.

Dito isto, é possível observar que:

1. A variável dependente Sales:
2. não apresenta correlação linear significante com a variável CompPrice (r= 0,064, p > 0.10);
3. apresenta correlação linear significante com a variável Income (r= 0.152, p < 0.001);

iii.apresenta correlação linear significante com a variável Advertising (r= 0.270, p < 0.001);

1. não apresenta correlação linear significante com a variável Population (r= 0.050, p > 0.10);
2. apresenta correlação linear significante com a variável Price (r= -0.445, p < 0.001);
3. apresenta correlação linear significante com a variável Age (r= -0.232, p < 0.001);
4. não apresenta correlação linear significante com a variável Education (r= -0.052, p > 0.10);

Uma maneira de identificar a multicolinearidade é através do **Fator de Inflação da Variância (VIF)**.

O **VIF é calculado** para cada preditor fazendo uma regressão linear desse preditor em todos os outros preditores e, então, obtendo o dessa regressão. O VIF é apenas .

**Como interpretar o VIF?**

Um **valor maior que 5** indica correlação **potencialmente grave** entre uma determinada variável preditora e outras variáveis preditoras no modelo. Nesse caso, as estimativas de coeficiente e os valores de p na saída da regressão provavelmente não são confiáveis.

**VIF > 10** indica **problema de multicolinearidade**.

Veja abaixo **como obter o vif** usando o *R*, desconsiderando a variável Sales no modelo inicial.

O problema da multicolinearidade não pode ser maior que 0,8/0,9

modelo1 <- lm(Sales ~ . - ShelveLoc - US - Urban, data = dados)  
  
library(car)   
  
vif(modelo1)

CompPrice Income Advertising Population Price Age   
 1.549720 1.014598 1.084521 1.101597 1.534323 1.015975   
 Education   
 1.014991

Como podemos ver, são as variáveis x e y que apresentam vif maior que 10 e são exatamente elas que também apresentaram forte correlação positiva. Fiquemos atentos ao resultados da regressão associados a estas duas variáveis preditoras que dão indício para o posterior problema de multicolinearidade.

# Ajuste do modelo 1 (Sem as vaiáveis CompPrice, US e Urban)

summary(modelo1)

Call:  
lm(formula = Sales ~ . - ShelveLoc - US - Urban, data = dados)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-5.0598 -1.3515 -0.1739 1.1331 4.8304   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 7.7076934 1.1176260 6.896 2.15e-11 \*\*\*  
CompPrice 0.0939149 0.0078395 11.980 < 2e-16 \*\*\*  
Income 0.0128717 0.0034757 3.703 0.000243 \*\*\*  
Advertising 0.1308637 0.0151219 8.654 < 2e-16 \*\*\*  
Population -0.0001239 0.0006877 -0.180 0.857092   
Price -0.0925226 0.0050521 -18.314 < 2e-16 \*\*\*  
Age -0.0449743 0.0060083 -7.485 4.75e-13 \*\*\*  
Education -0.0399844 0.0371257 -1.077 0.282142   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 1.929 on 392 degrees of freedom  
Multiple R-squared: 0.5417, Adjusted R-squared: 0.5335   
F-statistic: 66.18 on 7 and 392 DF, p-value: < 2.2e-16

Ao observar o ajuste do primeiro modelo com o intuito de prever a variável Sales, tem-se que nem todas as variáveis explicam de forma estatísticamente significativa, exceto Population e Education e com um ajuste indicado pelo Coeficiente de determinação ajustado (Adjusted R-squared:0.5335).