3º Estágio / Trabalho Avaliativo

Ítallo Silva - 118110718 | Thiago Nascimento - 118110804 | João Marcelo Junior - 117110448

Questão 3

Carregando os dados

A seguir, fazemos a leitura dos dados, neles encontramos três colunas:

- faturamento, nossa variável *target*;
- pesquisa
- propaganda

```
dados <- read_xls("../datasets/faturamento-hamburguer-fast-food.xls")
```

Separação em treino e teste

Iremos separar nosso conjunto de dados em treino e teste, com as proporções de 80% e 20%, respectivamente. Com isso, poderemos avaliar o desempenho do modelo em dados que este não viu previamente descartando a possibilidade de *overfitting*.

```
set.seed(1223)
treino <- sample_frac(dados, .8)
teste <- setdiff(dados, treino)
```

Criando o modelo

Criaremos o modelo linear utilizando a função `lm`.

```
mod <- lm(faturamento ~ ., dados)
```

Testando os pressupostos

Normalidade dos resíduos

Para avaliar a normalidade dos resíduos, aplicaremos o teste de Shapiro-Wilk, nesse teste temos:

- H_0 : distribuição dos dados = normal
- H_1 : distribuição dos dados $\neq$ normal

```
shapiro.test(mod$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod$residuals
## W = 0,96, p-value = 0,06
```

Temos um p-value de aproximadamente 0.06, assim considerando um nível de significância de 5% podemos assumir a normalidade dos resíduos.

Outliers nos resíduos

```
summary(rstandard(mod))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1,9890 -0,7274 -0,0683 -0,0012  0,9150  1,6935
```

Vemos que a distribuição dos resíduos padronizados está entre -2 e 2, assim podemos concluir que não temos nenhum outlier.

Independência dos resíduos

Iremos verificar a independência dos resíduos a partir do teste Durbin-Watson, nele temos as seguintes hipóteses:

- $H_0 : \rho = 0$, não há autocorrelação dos resíduos
- $H_1 : \rho \neq 0$, há autocorrelação dos resíduos

```
car::durbinWatsonTest(mod)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1        -0,04637         2,041   0,886
##  Alternative hypothesis: rho != 0
```

Considerando um nível de significância de 5%, como o p-value é maior que ele, assumimos a hipótese nula, assim confirmando que os resíduos são independentes.

Homocedasticidade

Verificaremos a seguir a presença de homocedasticidade, ou seja, variância constante dos resíduos, temos como hipóteses:

- H_0 : há homocedasticidade
- H_1 : não há homocedasticidade

```r
lmtest::bptest(mod)
```

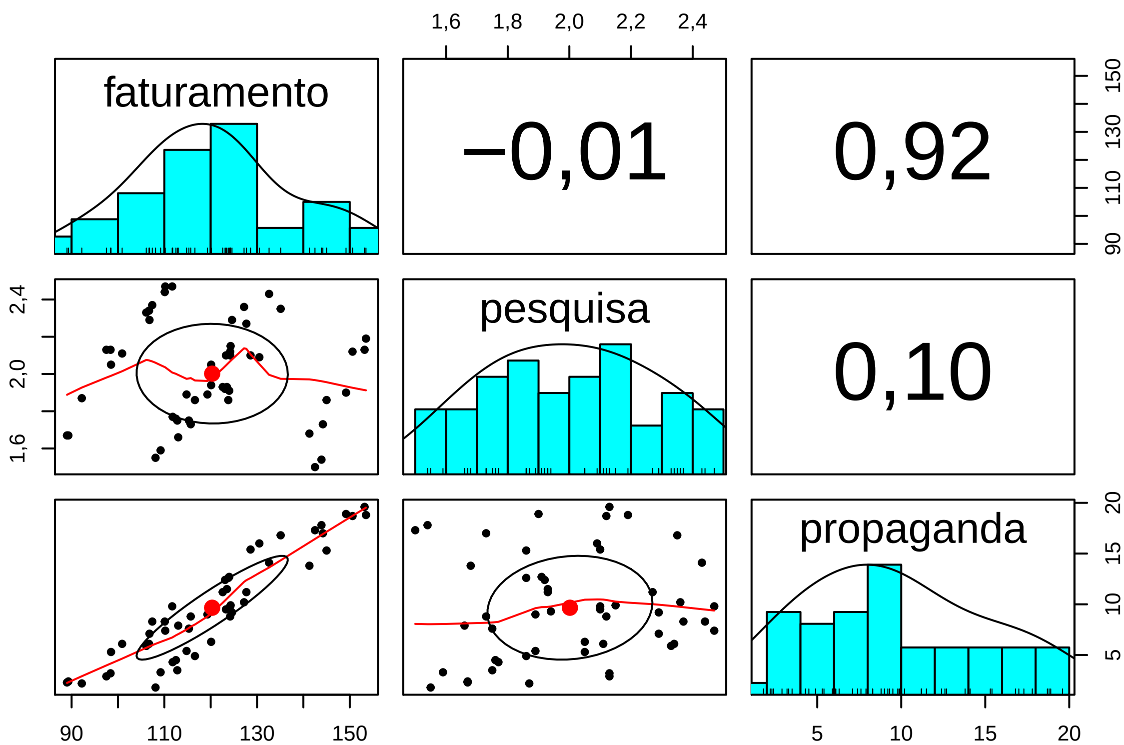
```
##
##  studentized Breusch-Pagan test
##
## data:  mod
## BP = 7,6, df = 2, p-value = 0,02
```

Temos um p-value menor que 5%, sendo assim devemos assumir a hipótese alternativa e portanto não temos homocedasticidade.

Ausência de multicolinearidade

Vamos então avaliar se existe multicolinearidade, ou seja, alguma correlação muito alta entre as variáveis independentes. Consideraremos uma correlação alta se ela for maior que 0.8, em módulo.

```r
psych::pairs.panels(dados)
```



Pelo gráfico acima, vemos que a correlação entre as variáveis independentes é bastante baixa (0.10), assim indicando que não há multicolinearidade. Podemos ver ainda que a variável pesquisa tem uma baixa correlação com a variável alvo faturamente, sugerindo que talvez ela não seja necessária ao modelo.

Outra forma de verificar a ausência de multicolinearidade é através da função VIF (inflação da variância):

```
car::vif(mod)
```

```
##   pesquisa propaganda
##       1,01       1,01
```

Consideramos que existe multicolinearidade caso o valor de VIF da variável seja maior que 10, como vemos não é nosso caso. Assim, confirmamos que não temos multicolinearidade.

Analisando o modelo

```
summary(mod)
```

```
##
## Call:
## lm(formula = faturamento ~ ., data = dados)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11,557  -4,244  -0,403   5,330  10,129
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  104,786      6,483   16,16   <2e-16 ***
## pesquisa      -6,642      3,191   -2,08    0,043 *
## propaganda     2,984      0,167   17,88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## Residual standard error: 6,07 on 49 degrees of freedom
## Multiple R-squared:  0,867,  Adjusted R-squared:  0,862
## F-statistic:  160 on 2 and 49 DF,  p-value: <2e-16
```

Pelo exposto acima, vemos que a variável pesquisa apresenta uma baixa significância no modelo (o coeficiente da variável é próximo a zero), conforme suposto pela sua baixa correlação.

Podemos ver ainda que obtivemos um R^2 ajustado de 0.86, e vemos que a estatística F tem um p-value muito baixo, assim com muita significância podemos afirmar que o modelo é melhor que o modelo nulo (ou seja, simplesmente prever utilizando a média).

Este resultado pode indicar que temos um bom modelo ou que está havendo *overfitting*, precisamos então realizar um teste para avaliar o R^2 do modelo com dados desconhecidos.

Realizando predição com o modelo

Iremos realizar uma predição usando o nosso modelo treino e o conjunto de dados que separamos como conjunto de teste e avaliar o R^2 obtido.

```
predicoes <- predict.lm(mod, teste)
yardstick::rsq_vec(teste$faturamento, predicoes)
```

```
## [1] 0,8136
```

Obtivemos um R^2 de aproximadamente 0.81, indicando que temos de fato um bom modelo.

Questão 4

Carregando os dados

A seguir, fazemos a leitura dos dados, neles encontramos três colunas:

- faturamento, nossa variável *target*;
- pesquisa
- propaganda

```
dados <- read_csv("../datasets/House_price.csv", col_types = cols_only(
  `Avg. Area Income` = col_double(),
  `House Age` = col_double(),
  `Number of Rooms` = col_double(),
  `Number of Bedrooms` = col_double(),
  `Area Population` = col_double(),
  Price = col_double()
))
```

Separação em treino e teste

Iremos separar nosso conjunto de dados em treino e teste, com as proporções de 80% e 20%, respectivamente. Com isso, poderemos avaliar o desempenho do modelo em dados que este não viu previamente descartando a possibilidade de *overfitting*.

```
set.seed(1223)
treino <- sample_frac(dados, .8)
teste <- setdiff(dados, treino)
```

Criando o modelo

Criaremos o modelo linear utilizando a função `lm`.

```
mod <- lm(Price ~ ., dados)
```

Testando os pressupostos

Normalidade dos resíduos

Para avaliar a normalidade dos resíduos, aplicaremos o teste de Shapiro-Wilk, nesse teste temos:

- H_0 : distribuição dos dados = normal
- H_1 : distribuição dos dados $\neq$ normal

```
shapiro.test(mod$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod$residuals
## W = 1, p-value = 0,3
```

Temos um p-value de aproximadamente 0.06, assim considerando um nível de significância de 5% podemos assumir a normalidade dos resíduos.

Outliers nos resíduos

```
summary(rstandard(mod))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -3,346  -0,695  -0,001   0,000   0,684   3,435
```

Vemos que a distribuição dos resíduos padronizados está entre -2 e 2, assim podemos concluir que não temos nenhum outlier.

Independência dos resíduos

Iremos verificar a independência dos resíduos a partir do teste Durbin-Watson, nele temos as seguintes hipóteses:

- $H_0 : \rho = 0$, não há autocorrelação dos resíduos
- $H_1 : \rho \neq 0$, há autocorrelação dos resíduos

```
car::durbinWatsonTest(mod)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0,0001884             2   0,984
##  Alternative hypothesis: rho != 0
```

Considerando um nível de significância de 5%, como o p-value é maior que ele, assumimos a hipótese nula, assim confirmando que os resíduos são independentes.

Homocedasticidade

Verificaremos a seguir a presença de homocedasticidade, ou seja, variância constante dos resíduos, temos como hipóteses:

- H_0 : há homocedasticidade
- H_1 : não há homocedasticidade

```
lmtest::bptest(mod)
```

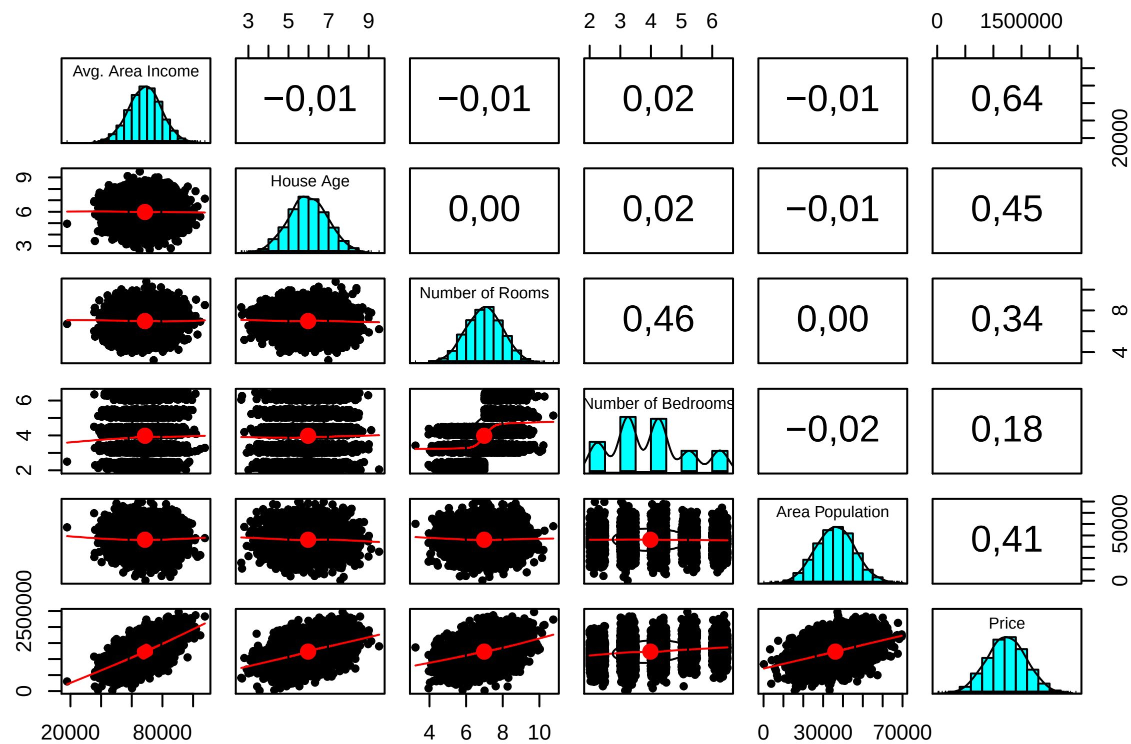
```
##
##  studentized Breusch-Pagan test
##
## data:  mod
## BP = 5,9, df = 5, p-value = 0,3
```

Temos um p-value menor que 5%, sendo assim devemos assumir a hipótese alternativa e portanto não temos homocedasticidade.

Ausência de multicolinearidade

Vamos então avaliar se existe multicolinearidade, ou seja, alguma correlação muito alta entre as variáveis independentes. Consideraremos uma correlação alta se ela for maior que 0.8, em módulo.

```
psych::pairs.panels(dados)
```



Pelo gráfico acima, vemos que a correlação entre as variáveis independentes é bastante baixa (0.10), assim indicando que não há multicolinearidade. Podemos ver ainda que a variável pesquisa tem uma baixa correlação com a variável alvo faturamente, sugerindo que talvez ela não seja necessária ao modelo.

Outra forma de verificar a ausência de multicolinearidade é através da função VIF (inflação da variância):

```
car::vif(mod)
```

```
##   `Avg. Area Income`          `House Age`    `Number of Rooms`
##                1,001                1,001                1,276
## `Number of Bedrooms`    `Area Population`
##                1,277                1,001
```

Consideramos que existe multicolinearidade caso o valor de VIF da variável seja maior que 10, como vemos não é nosso caso. Assim, confirmamos que não temos multicolinearidade.

Analisando o modelo

```r
summary(mod)
```

```
##
## Call:
## lm(formula = Price ~ ., data = dados)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -338141  -70192    -101   69104  347218
##
## Coefficients:
##                          Estimate   Std. Error t value Pr(>|t|)
## (Intercept)          -2639277,424    17905,578 -147,40   <2e-16 ***
## `Avg. Area Income`         21,610        0,140  153,93   <2e-16 ***
## `House Age`            165341,729     1513,725  109,23   <2e-16 ***
## `Number of Rooms`      121144,906     1682,489   72,00   <2e-16 ***
## `Number of Bedrooms`     1458,480     1376,650    1,06     0,29
## `Area Population`          15,188        0,151  100,36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## Residual standard error: 101000 on 4542 degrees of freedom
## Multiple R-squared:  0,919,  Adjusted R-squared:  0,919
## F-statistic: 1,03e+04 on 5 and 4542 DF,  p-value: <2e-16
```

Pelo exposto acima, vemos que a variável pesquisa apresenta uma baixa significância no modelo (o coeficiente da variável é próximo a zero), conforme suposto pela sua baixa correlação.

Podemos ver ainda que obtivemos um R^2 ajustado de 0.86, e vemos que a estatística F tem um p-value muito baixo, assim com muita significância podemos afirmar que o modelo é melhor que o modelo nulo (ou seja, simplesmente prever utilizando a média).

Este resultado pode indicar que temos um bom modelo ou que está havendo *overfitting*, precisamos então realizar um teste para avaliar o R^2 do modelo com dados desconhecidos.

Realizando predição com o modelo

Iremos realizar uma predição usando o nosso modelo treino e o conjunto de dados que separamos como conjunto de teste e avaliar o R^2 obtido.

```r
predicoes <- predict.lm(mod, teste)
yardstick::rsq_vec(teste$Price, predicoes)
```

```
## [1] 0,9232
```

Obtivemos um R^2 de aproximadamente 0.73, indicando que temos de fato um bom modelo.