Técnicas Estatísticas de Predição

Exercício 1

João Pedro Gentil da Silveira

Questões

A) Utilizando o software R, retire uma amostra aleatória de tamanho 80 da base de dados.

R:

```r
# Leitura da base de dados
base = read.csv2("apartamento.csv", dec=".")

# Converte os valores de Local para:
#   -> 0 = Região menos valorizada;
#   -> 1 = Região mais valorizada;
base$Local = ifelse(base$Local == 1, 1, 0)

# Tira a amostra de tamanho 80 da base de dados
amostra = base[sample(nrow(base), 80),]
```

B) Com a sua amostra retirada no item (a), ajuste um modelo de regressão linear múltipla utilizando somente as variáveis no modelo que apresentarem p-valor menor do que 0,05.

R: Como mostrado pelo output do `summary(modelo_inicial)`, a única variável com p-valor maior do que 0,05 foi Energia - retirada, logo em seguida, na construção de um novo modelo para visualização dos dados, apenas com variáveis consideradas estatisticamente significativas.

```r
# Cria e mostra um modelo inicial com todas as variáveis
modelo_inicial = lm(Valor ~ Area + Idade + Energia + Local,
data=amostra)
summary(modelo_inicial)
```

```
##
## Call:
## lm(formula = Valor ~ Area + Idade + Energia + Local, data = amostra)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41.76 -16.39  -1.06  14.90  44.24
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -35.99255   22.25230  -1.617   0.1100
## Area          1.01255    0.08602  11.771  < 2e-16 ***
```

```
## Idade        -2.15912    0.42722  -5.054 2.97e-06 ***
## Energia       0.26202    0.13996   1.872   0.0651 .
## Local        14.83759    5.71691   2.595   0.0114 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.29 on 75 degrees of freedom
## Multiple R-squared:   0.7504,  Adjusted R-squared:  0.7371
## F-statistic: 56.37 on 4 and 75 DF,  p-value: < 2.2e-16
```

```
# Mostra o modelo depois de retirar a variável energia
modelo_ajustado = lm(Valor ~ Area + Idade + Local, data=amostra)
summary(modelo_ajustado)
```

```
##
## Call:
## lm(formula = Valor ~ Area + Idade + Local, data = amostra)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.828 -18.734  -2.194  12.415  51.705
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.38942    8.79192   0.272  0.78653
## Area         1.04136    0.08602  12.106  < 2e-16 ***
## Idade       -2.23549    0.43222  -5.172 1.82e-06 ***
## Local       16.67035    5.72453   2.912  0.00471 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.67 on 76 degrees of freedom
## Multiple R-squared:  0.7388, Adjusted R-squared:  0.7284
## F-statistic: 71.64 on 3 and 76 DF,  p-value: < 2.2e-16
```

C) Interprete os coeficientes do modelo que permaneceram na análise.

R: Antes de analisar propriamente os coeficientes das variáveis permanecidas no modelo, é possível construir o seguinte modelo de regressão linear múltipla:

$$Valor = 2.38942 + 1.04136 \times Area - 2.23549 \times Idade + 16.67035 \times Local$$

Quanto ao coeficiente de cada variável:

- **Área:** Com um coeficiente estimado de 1.04136, sugere que, contanto que as outras variáveis se mantenham constantes, para cada m² aumentado na área do apartamento, seu valor tem um aumento médio de 1.04136 x 1000 reais.
- **Idade:** Com um coeficiente estimado de -2.23549, sugere que, no cenário em que as outras variáveis se mantenham constantes, para cada ano aumentado na idade do apartamento, seu valor tem uma diminuição média de 2.23549 x 1000 reais.

- **Local:** Com um coeficiente estimado de 16.67035, sugere que, mantendo as outras variáveis constantes, estar localizado em uma região mais valorizada (Local igual a 1), faz com que o apartamento tenha um aumento médio de 16.67035 x 1000 reais no seu valor quando em comparação com estar em uma região menos valorizada (Local igual a 0).

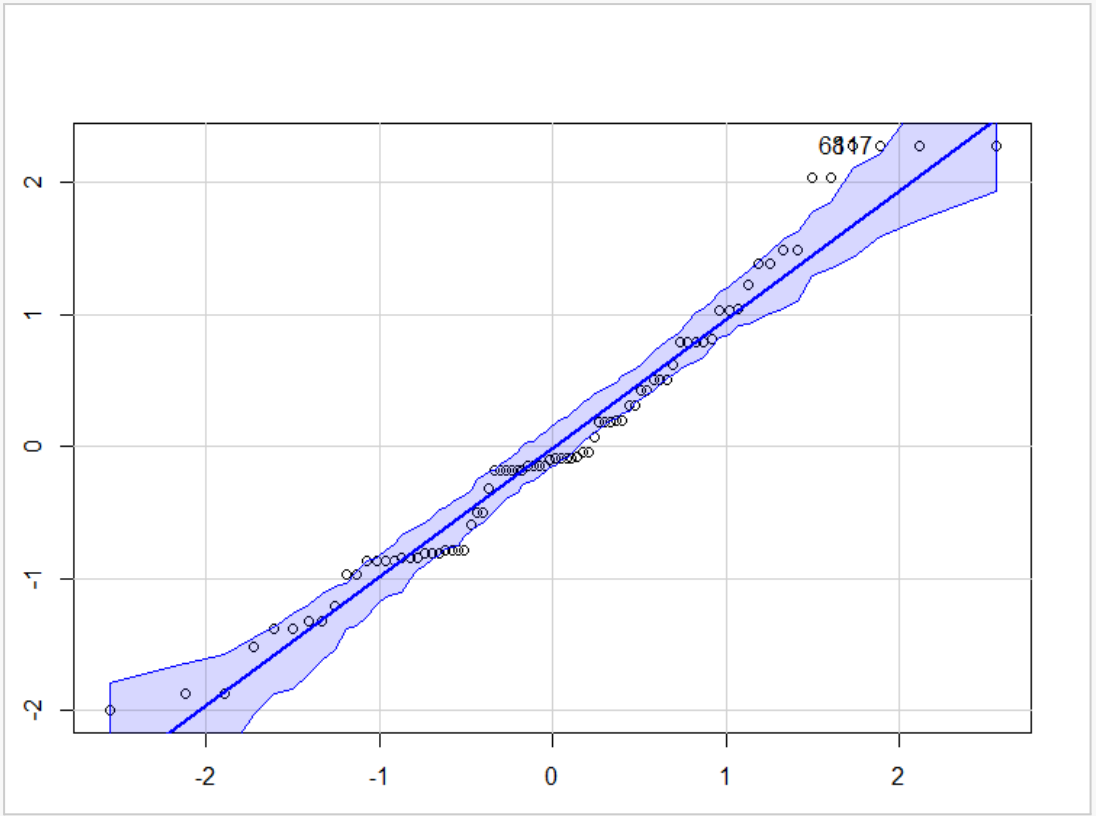
D) Interprete o coeficiente de determinação ajustado (R2 ajustado).

R: Um R^2 ajustado com coeficiente de 0.7284 permite inferir que próximo de 72,84% da variabilidade possível na variável dependente (Valor) é explicada pelas variáveis independentes (Área, Idade e Local) inclusas no modelo.

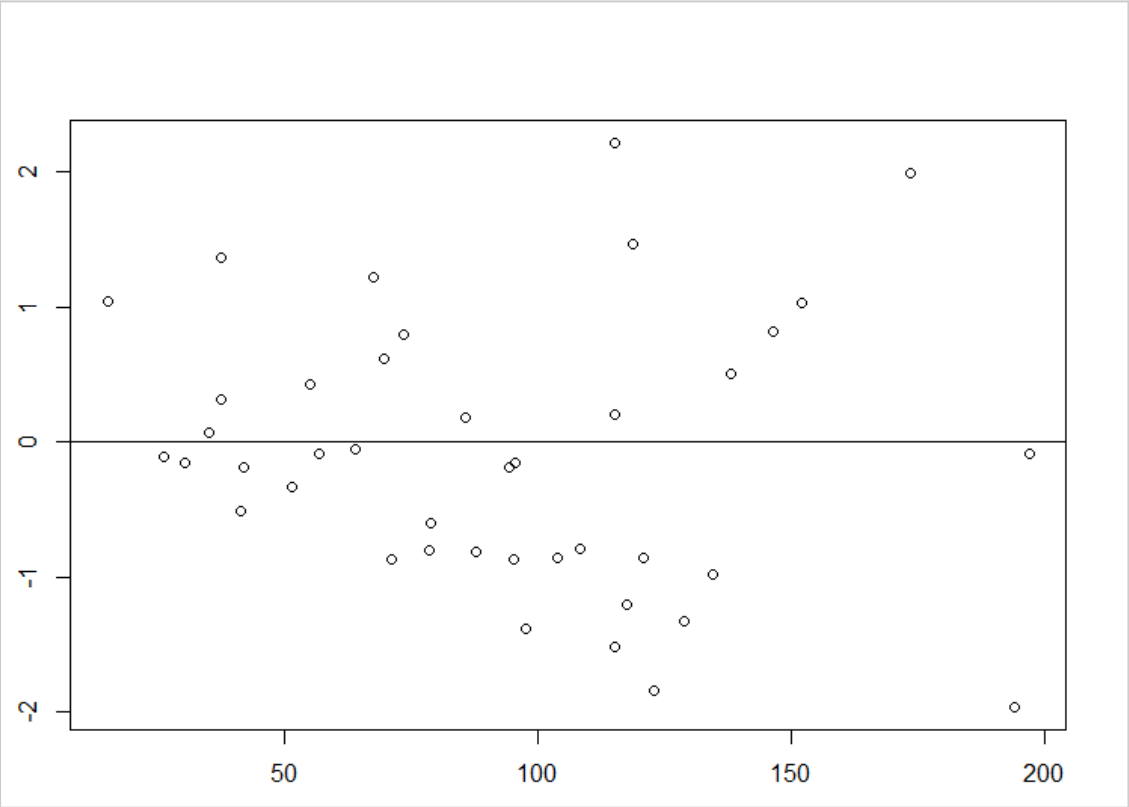
E) Faça a análise da Qualidade do Ajuste do modelo e indique se o modelo está bem ajustado indicando como você está chegando a essa conclusão.

R: De acordo com os seguintes gráficos para análise de resíduos do modelo, é bem possível afirmar que o modelo está bem ajustado, uma vez que os gráficos revelam uma variância de resíduos relativamente constante, distanciando-se do problema de heterocedasticidade. Fora isso, a plotagem dos resíduos não mostra uma tendência positiva ou negativa e a plotagem de probabilidade normal é até que linear.

```r
# Análise de Resíduos do modelo
library(car)
qqPlot(modelo_ajustado)
```



```r
plot(fitted(modelo_ajustado), rstandard(modelo_ajustado))
abline(0,0)
```

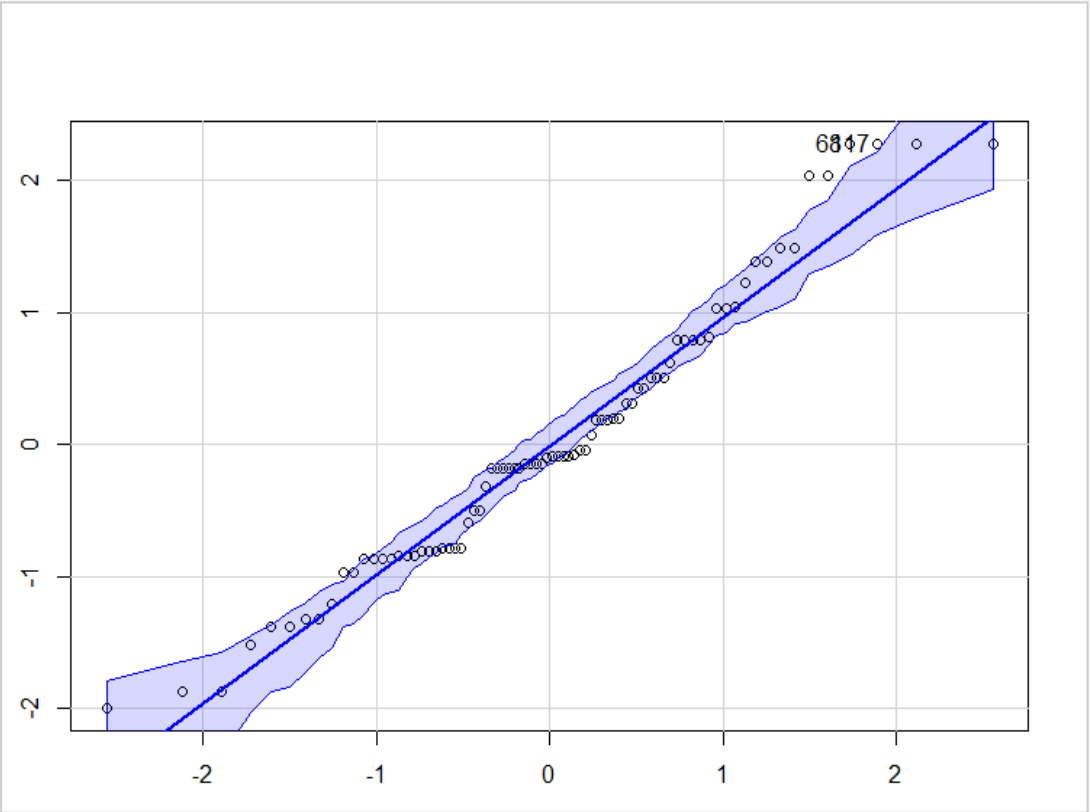


Sintaxe Completa

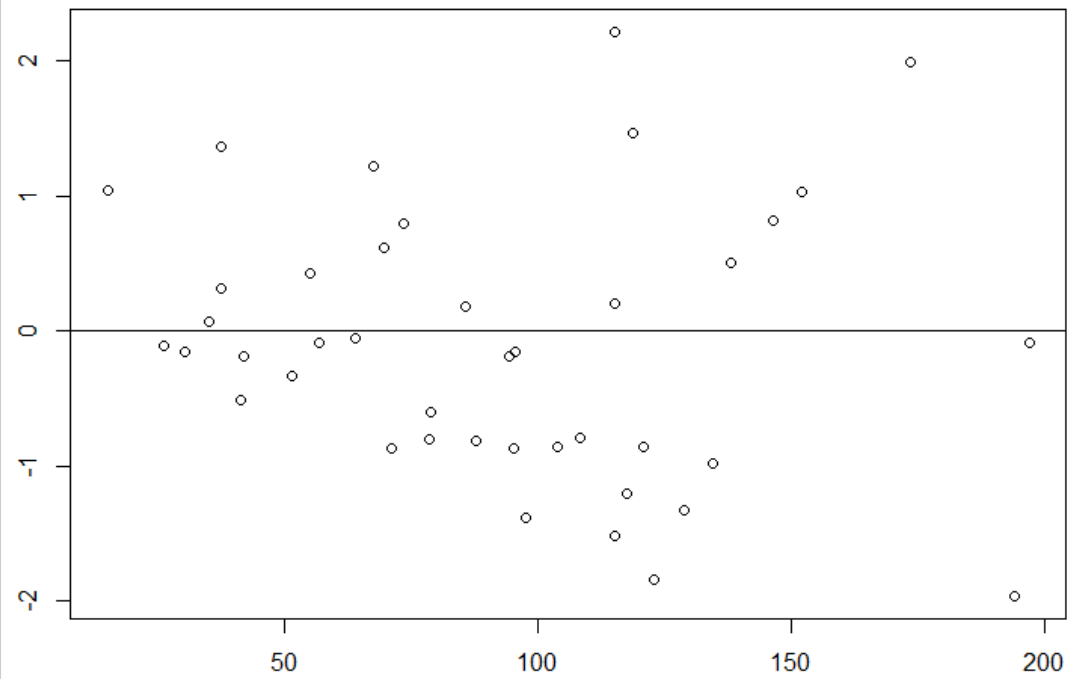
```r
# Leitura da base de dados
base = read.csv2("apartamento.csv", dec=".")

# Converte os valores de Local para:
#   -> 0 = Região menos valorizada;
#   -> 1 = Região mais valorizada;
base$Local = ifelse(base$Local == 1, 1, 0)

# Tira a amostra de tamanho 80 da base de dados
amostra = base[sample(nrow(base), 80),]

# Cria e mostra um modelo inicial com todas as variáveis
modelo_inicial = lm(Valor ~ Area + Idade + Energia + Local,
data=amostra)
summary(modelo_inicial)

##
## Call:
## lm(formula = Valor ~ Area + Idade + Energia + Local, data = amostra)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41.76 -16.39  -1.06  14.90  44.24
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -35.99255   22.25230  -1.617   0.1100
## Area          1.01255    0.08602  11.771  < 2e-16 ***
## Idade        -2.15912    0.42722  -5.054 2.97e-06 ***
## Energia       0.26202    0.13996   1.872   0.0651 .
## Local        14.83759    5.71691   2.595   0.0114 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.29 on 75 degrees of freedom
## Multiple R-squared:   0.7504,  Adjusted R-squared:  0.7371
## F-statistic: 56.37 on 4 and 75 DF,  p-value: < 2.2e-16

# Mostra o modelo depois de retirar a variável energia
modelo_ajustado = lm(Valor ~ Area + Idade + Local, data=amostra)
summary(modelo_ajustado)

##
## Call:
## lm(formula = Valor ~ Area + Idade + Local, data = amostra)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.828 -18.734  -2.194  12.415  51.705
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.38942    8.79192   0.272  0.78653
## Area         1.04136    0.08602  12.106  < 2e-16 ***
## Idade       -2.23549    0.43222  -5.172 1.82e-06 ***
## Local       16.67035    5.72453   2.912  0.00471 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.67 on 76 degrees of freedom
## Multiple R-squared:  0.7388, Adjusted R-squared:  0.7284
## F-statistic: 71.64 on 3 and 76 DF,  p-value: < 2.2e-16
```

```
# Análise de Resíduos do modelo
library(car)
qqPlot(modelo_ajustado)
```




```
plot(fitted(modelo_ajustado), rstandard(modelo_ajustado))
abline(0,0)
```


2
1
0
-1
-2
50
100
150
200