Técnicas Estatísticas de Predição

Exercício 2

João Pedro Gentil da Silveira

Questões

A) Retire da análise as observações drivewheel=4wd.

R:

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

# Retirada dos valores 4wd de drivewheel
base_sem_4wd = base[base$drivewheel != "4wd",]
```

B) Retire uma amostra de 120 carros utilizando a sintaxe com seed.

R:

```
# Tira a amostra de tamanho 120 da base de dados
set.seed(29072003)
amostra = base_sem_4wd[sample(nrow(base_sem_4wd), 120),]
```

C) Ajuste um modelo com interação considerando “price” como variável dependente e “carwidth” e “drivewheel” como variáveis independentes.

R:

```
# Cria e mostra um modelo
modelo = lm(price ~ carwidth * drivewheel, data = amostra)
summary(modelo)
```

```
## Call:
## lm(formula = price ~ carwidth * drivewheel, data = amostra)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3269  -2.4590  -0.4659   1.3129  23.1797
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -98.0942    27.6820  -3.544 0.000570 ***
## carwidth                   1.6512     0.4254   3.881 0.000173 ***
## drivewheelrwd            -58.5608    33.7337  -1.736 0.085224 .
## carwidth:drivewheelrwd     0.9719     0.5129   1.895 0.060562 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 4.933 on 116 degrees of freedom
## Multiple R-squared:  0.6681, Adjusted R-squared:  0.6595
## F-statistic: 77.83 on 3 and 116 DF,  p-value: < 2.2e-16

# Mostra o modelo depois de retirar a interação
modelo_sem_interacao = lm(price ~ carwidth + drivewheel, data = amostra)
summary(modelo_sem_interacao)

## Call:
## lm(formula = price ~ carwidth + drivewheel, data = amostra)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9857 -2.7395 -0.3203  1.3916 22.4904
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -141.6032    15.6369  -9.056 3.63e-15 ***
## carwidth         2.3200     0.2402   9.659  < 2e-16 ***
## drivewheelrwd    5.3386     1.0596   5.038 1.73e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.987 on 117 degrees of freedom
## Multiple R-squared:  0.6578, Adjusted R-squared:  0.652
## F-statistic: 112.5 on 2 and 117 DF,  p-value: < 2.2e-16
```

D) Analise se existe a necessidade de interação no modelo (caso não exista necessidade retire a interação do modelo) e interprete os coeficientes

R: Uma vez que o coeficiente da interação entre "carwidth" e "drivewheel" tem um p-valor de cerca de 0.061, é possível designar a interação como não significativa a um nível de significância de 0.05, pois o p-valor ultrapassa tal valor. Desse modo, a interação pode ser descartada - como feito na sintaxe da questão anterior.

Contudo, antes de analisar as variáveis presentes no modelo, é possível construir o seguinte modelo de regressão linear múltipla:

$$Preço = \ -141.6032 + 2.3200 \times carwidth + 5.3386 \times drivewheel$$

Quanto ao coeficiente de cada variável:

- **Cardwith:** Com um coeficiente estimado de 2.3200, sugere que, somente se as outras variáveis se mantiverem constantes, para cada aumento de uma unidade (polegada) na largura do carro, seu valor tem um aumento médio de 2.3200 x 1000 reais.
- **Drivewheel:** Com um coeficiente estimado de 5.3386, sugere que, no caso em que a outra variável se mantenha constante, em comparação com carros

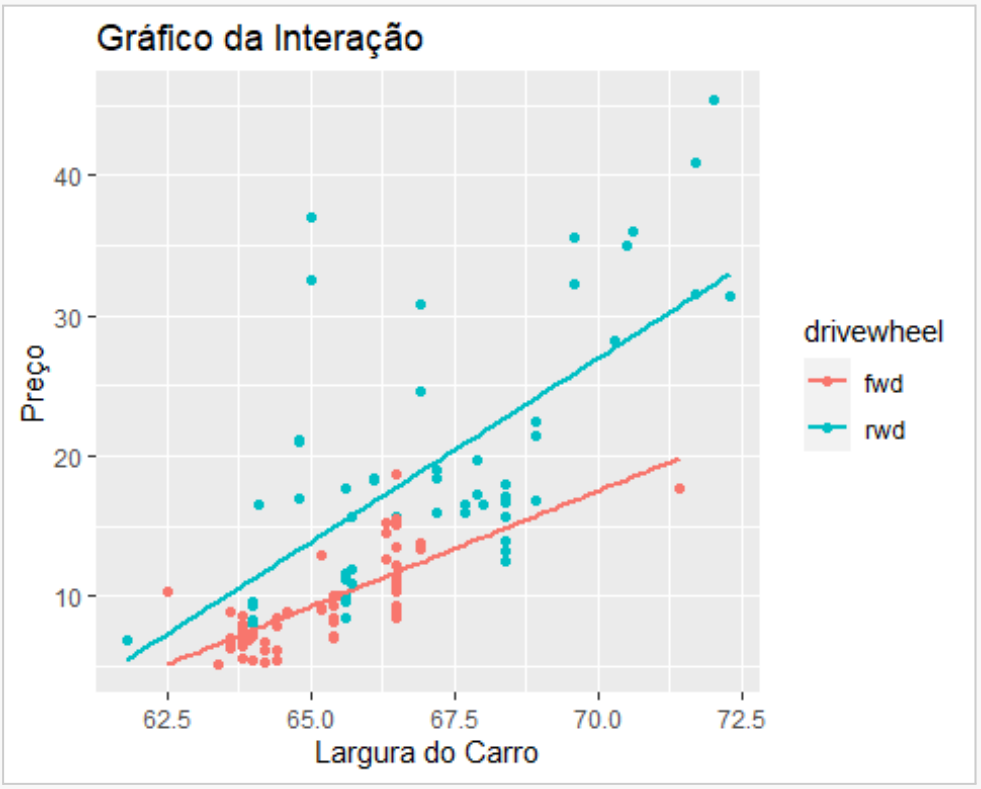
com a tração dianteira - valor de drivewheel igual a fwd -, carros com a tração traseira - valor de drivewheel igual rwd - têm um preço médio de 5.3386 x 1000 reais maior.

E) Apresente um gráfico para representar a interação no modelo.

R:

```
# Cria gráfico de interação com o modelo antigo
library(ggplot2)
ggplot(data = modelo, aes(x = carwidth, y = price, color = drivewheel))
+
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Gráfico da Interação",
       x = "Largura do Carro",
       y = "Preço")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



F) Construa um intervalo de confiança e intervalo de predição, considerando que o carro possui tração dianteira e tem largura de 70 polegadas. Explique a diferença do intervalo de confiança e do intervalo de predição no contexto do problema.

R: O intervalo de confiança é como uma estimativa da média do preço dos carros, considerando a variabilidade nos dados que temos e nos coeficientes do modelo, estimando onde a média do preço esteja para todos os carros. Por outro lado, o

intervalo de predição leva em conta não apenas a variabilidade nos dados e nos coeficientes do modelo, mas também a incerteza em torno de uma nova observação específica, fornecendo uma faixa dentro da qual espera-se que o preço de um carro individual possa variar.

Sendo assim, a diferença fundamental é que o intervalo de confiança é sobre a média esperada de todos os carros com características semelhantes, enquanto o intervalo de predição é sobre o preço esperado para um carro específico.

```
# Define os valores das variáveis independentes para o carro
carro_especifico = data.frame(carwidth = 70, drivewheel = "fwd")

# Constrói e mostra os intervalos de confiança e de predição
conf_interval <- predict(modelo_sem_interacao, newdata =
carro_especifico, interval = "confidence", level = 0.95)

conf_interval
##        fit      lwr      upr
## 1 20.79915 18.14912 23.44918

pred_interval <- predict(modelo_sem_interacao, newdata =
carro_especifico, interval = "prediction", level = 0.95)

pred_interval
##        fit      lwr      upr
## 1 20.79915 10.57323 31.02507
```

Sintaxe Completa

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

# Retirada dos valores 4wd de drivewheel
base_sem_4wd = base[base$drivewheel != "4wd",]

# Tira a amostra de tamanho 120 da base de dados
set.seed(29072003)
amostra = base_sem_4wd[sample(nrow(base_sem_4wd), 120),]

# Cria e mostra um modelo
modelo = lm(price ~ carwidth * drivewheel, data = amostra)
summary(modelo)
```

```
## Call:
## lm(formula = price ~ carwidth * drivewheel, data = amostra)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3269  -2.4590  -0.4659   1.3129  23.1797
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -98.0942    27.6820  -3.544 0.000570 ***
## carwidth                  1.6512     0.4254   3.881 0.000173 ***
## drivewheelrwd           -58.5608    33.7337  -1.736 0.085224 .
## carwidth:drivewheelrwd    0.9719     0.5129   1.895 0.060562 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.933 on 116 degrees of freedom
## Multiple R-squared:  0.6681, Adjusted R-squared:  0.6595
## F-statistic: 77.83 on 3 and 116 DF,  p-value: < 2.2e-16
```

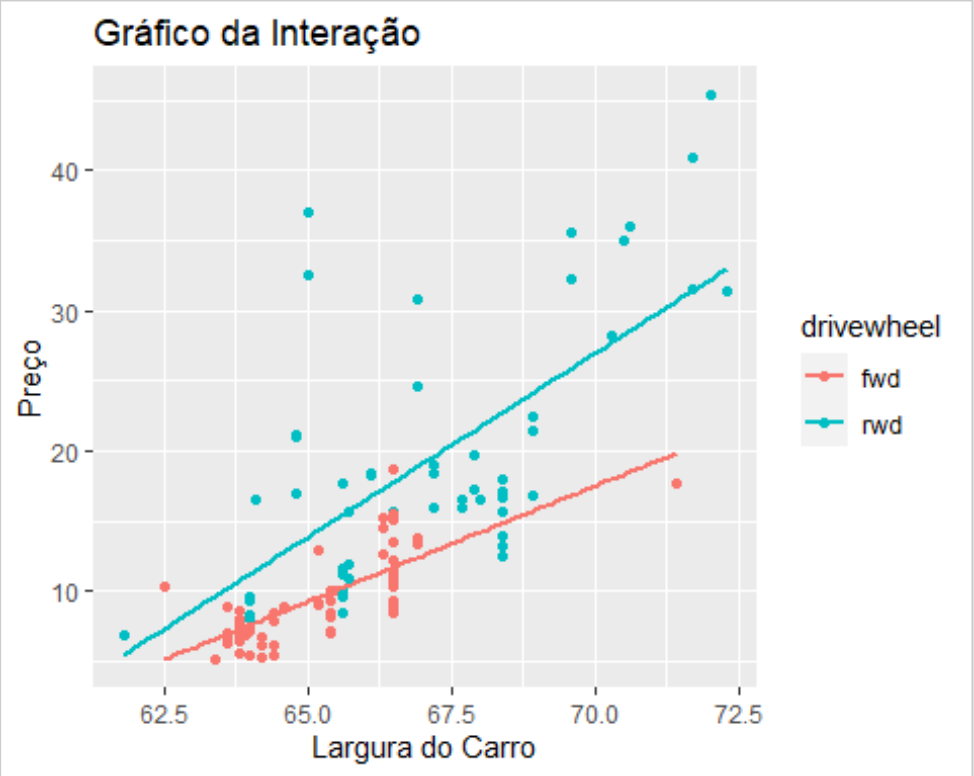
```r
# Mostra o modelo depois de retirar a interação
modelo_sem_interacao = lm(price ~ carwidth + drivewheel, data = amostra)
summary(modelo_sem_interacao)
```

```
## Call:
## lm(formula = price ~ carwidth + drivewheel, data = amostra)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9857 -2.7395 -0.3203  1.3916 22.4904
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -141.6032    15.6369  -9.056 3.63e-15 ***
## carwidth         2.3200     0.2402   9.659  < 2e-16 ***
## drivewheelrwd    5.3386     1.0596   5.038 1.73e-06 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.987 on 117 degrees of freedom
## Multiple R-squared:  0.6578, Adjusted R-squared:  0.652
## F-statistic: 112.5 on 2 and 117 DF,  p-value: < 2.2e-16
```

```
# Cria gráfico de interação com o modelo antigo
library(ggplot2)
ggplot(data = modelo, aes(x = carwidth, y = price, color = drivewheel))
+
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Gráfico da Interação",
       x = "Largura do Carro",
       y = "Preço")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Define os valores das variáveis independentes para o carro
carro_especifico = data.frame(carwidth = 70, drivewheel = "fwd")

# Constrói e mostra os intervalos de confiança e de predição
conf_interval <- predict(modelo_sem_interacao, newdata =
carro_especifico, interval = "confidence", level = 0.95)

conf_interval
##        fit      lwr      upr
## 1 20.79915 18.14912 23.44918
```

```r
pred_interval <- predict(modelo_sem_interacao, newdata =
carro_especifico, interval = "prediction", level = 0.95)

pred_interval
##        fit      lwr      upr
## 1 20.79915 10.57323 31.02507
```