Lista 2 - MAE 0399 - Análise de Dados e Simulação

Guilherme Ventura (11340293), Milton Leal (8973974), Richard Sousa (11810898)

07/07/2021

Questão 4

Item a) R: A média dos valores da tabela é

$$\frac{(22,2+61,1+13,7+27,8+22,8+7,4+8,7+6,3+20,4+25,6+23,2+11,1+13+7,2+14,8)}{15}$$

$$= 19,01 km/h.$$

Para encontrarmos a mediana dos valores, ordenamos os valores:

$$6,3; 7,2; 7,4; 8,7; 11,1; 13; 13,7; 14,8; 20,4; 22,2; 22,7; 23,2; 25,6; 27,8; 61,1$$

agora, basta contar o total de números, se o total for par deve-se fazer a média aritmética dos valores centrais, caso seja ímpar (nosso caso) deve-se tomar o valor central. Ou seja, a mediana é $14,8$ km/h.

Para encontrarmos o desvio padrão, primeiro devemos calcular a variância:

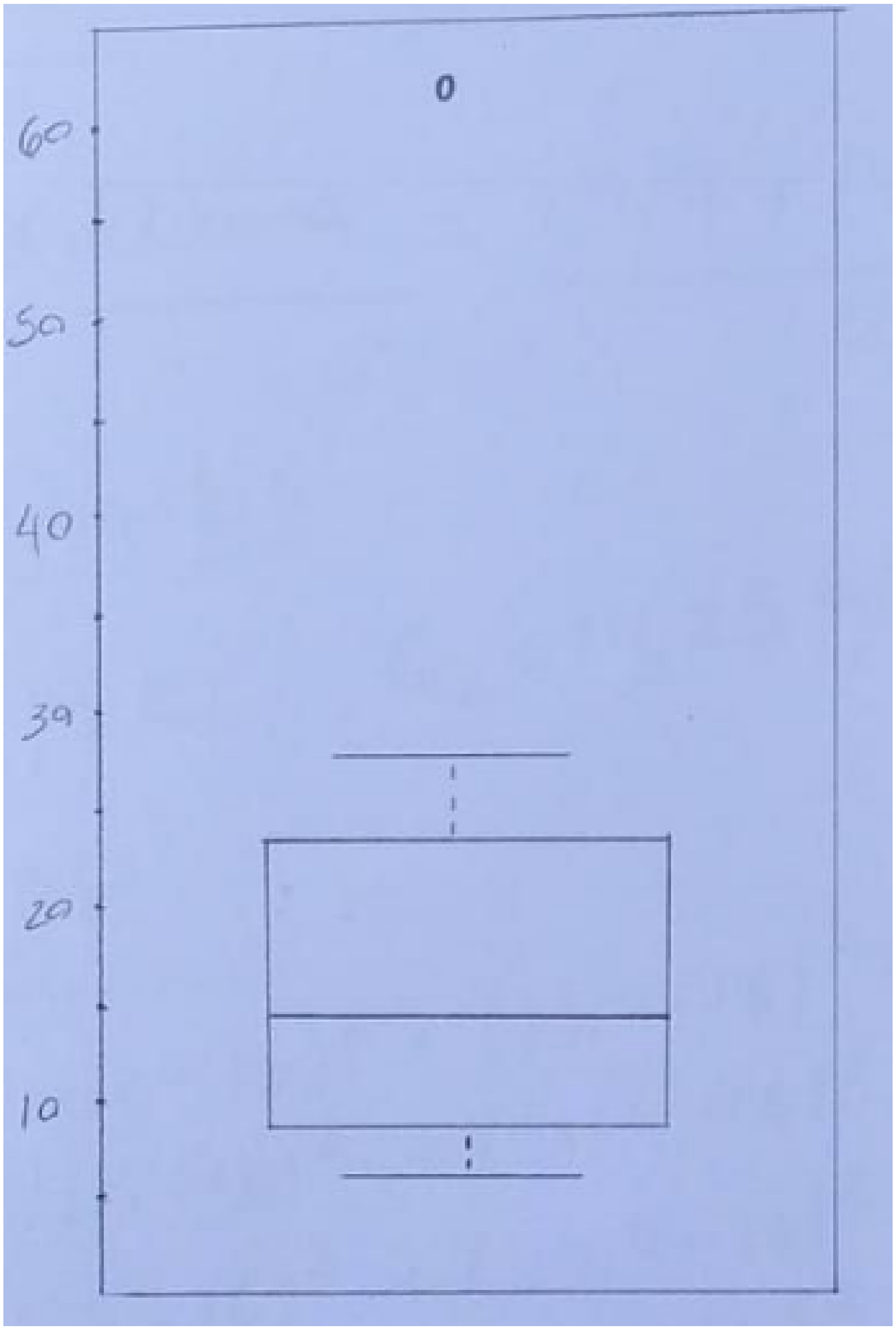
$((22,2-19,01)^2 + (61,1-19,01)^2 + (13,7-19,01)^2 + (27,8-19,01)^2 + (22,7-19,01)^2 + (7,4-19,01)^2 + (8,7-19,01)^2 + (6,3-19,01)^2 + (20,4-19,01)^2 + (25,6-19,01)^2 + (23,2-19,01)^2 + (11,1-19,01)^2 + (13-19,01)^2 + (7,2-19,01)^2 + (14,8-19,01)^2)/(15-1) = (10,17+1771,56+28,19+77,26+13,61+134,79+106,29+161,54+1,93+43,42+17,55+62,56+36,12+139,47+17,72)/14 = 187,3$

Portanto, o desvio padrão é $\sqrt{187,3} = 13,68 km/h$.

Para os quartis, temos: $Q_1 = 8,7$ km/h, $Q_2 = 14,8$ km/h, $Q_3 = 23,2$ km/h.

Para desenhar o boxplot, precisamos encontrar a distância interquartil: $dq = q_3 - q_1 = 23,2 - 8,7 = 15,1$ km/h. E, agora, calculamos os limites superior e inferior: $L_S = q_3 + (1,5)dq = 45,85$ km/h e $L_I = q_1 - (1,5)dq = -13,95$ km/h.

Segue abaixo o gráfico Boxplot:



Item b) R: Sim, existe um valor atípico, que é o $61,1$. Vamos removê-lo e refazer as contas: Para a nova média, devemos retirar o valor atípico, então o novo conjunto de dados é

$$6,3; 7,2; 7,4; 8,7; 11,1; 13; 13,7; 14,8; 20,4; 22,2; 22,7; 23,2; 25,6; 27,8.$$

Portanto,

$$\frac{(22,2+13,7+27,8+22,8+7,4+8,7+6,3+20,4+25,6+23,2+11,1+13+7,2+14,8)}{14} = 16,01 km/h.$$

Para a nova mediana, como o novo conjunto, descrito acima, tem um número par de valores então a mediana é dada pela soma dos valores centrais: $(13,7 + 14,8)/2 = 14,25$ km/h.

Para o novo desvio padrão, calculamos primeiro a variância sob o novo conjunto:

$$((22,2-16)^2 + (13,7-16)^2 + (27,8-16)^2 + (22,7-16)^2 + (7,4-16)^2 + (8,7-16)^2 + (6,3-16)^2 + (20,4-16)^2 + (25,6-16)^2 + (23,2-16)^2 + (11,1-16)^2 + (13-16)^2 + (7,2-16)^2 + (14,8-16)^2)/(14-1)$$

$$= \frac{(38,44 + 5,29 + 139,24 + 44,89 + 53,29 + 94,09 + 19,36 + 92,16 + 51,89 + 24,01 + 9 + 77,44 + 1,44)}{13}$$

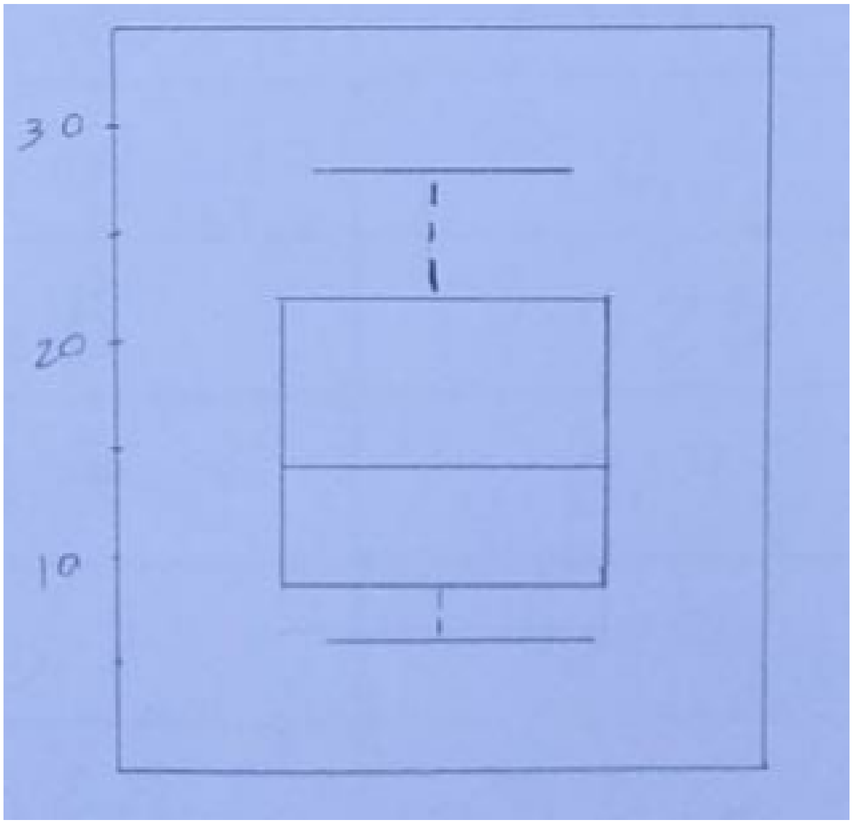
$$= 55,72.$$

Daqui, o desvio padrão é $\sqrt{55,72} = 7,46$ km/h.

Os novos quartis são dados por: $Q_1 = 8,7$ km/h, $Q_2 = 14,25$ km/h e $Q_3 = 22,7$ km/h.

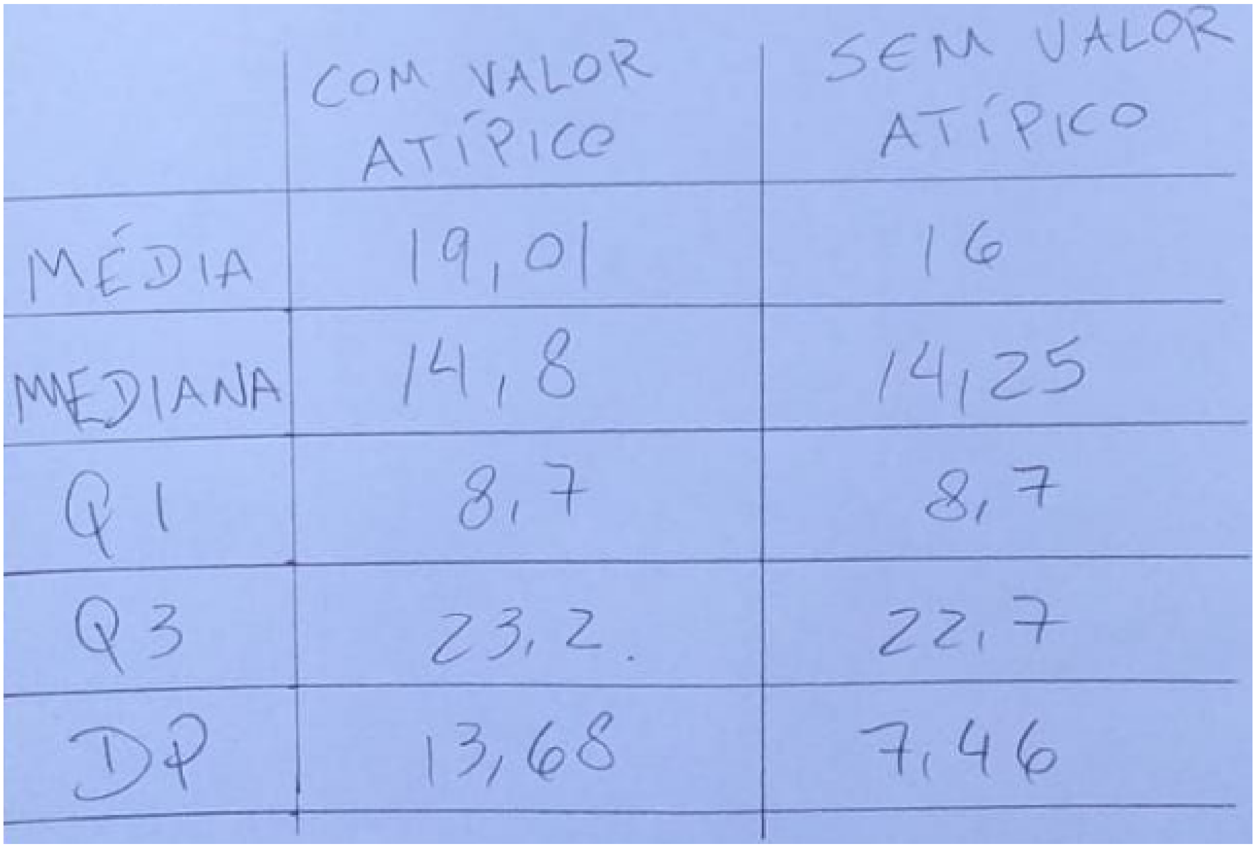
E para o novo boxplot, calculamos a distância interquartil: $dq = q_3 - q_1 = 22,7 - 8,7 = 14$ km/h. E os limites superior e inferior: $L_S = q_3 + (1,5)dq = 43,7$ km/h e $L_I = q_1 - (1,5)dq = -12,3$ km/h.

Segue abaixo o gráfico Boxplot do novo conjunto de dados:



Tendo em vista os novos valores acima, é possível notar que a média e o desvio padrão foram as medidas mais afetadas pela retirada do valor atípido. Porém, quando abordamos a mediana, o 1º quartil e o 3º quartil, notamos que essas medidas sofreram alterações pequenas quando comparadas à alteração sofrida pela medidas anteriores. Isso mostra que a mediana e os quartis são medidas mais robustas.

Segue abaixo a tabela de comparação das medidas com e sem o valor atípico:



	COM VALOR ATÍPICO	SEM VALOR ATÍPICO
MÉDIA	19,01	16
MEDIANA	14,8	14,25
Q1	8,7	8,7
Q3	23,2.	22,7
DP	13,68	7,46

Questão 5

R: Sejam P_i os pontos observados e P_c o ponto a ser classificado. Vamos calcular a distância euclidiana entre P_i e P_c:

$$d(P_1, P_c) = ((0-0)^2 + (3-0)^2 + (0-0)^2)^{\frac{1}{2}} = 3$$
$$d(P_2, P_c) = ((2-0)^2 + (0-0)^2 + (0-0)^2)^{\frac{1}{2}} = 2$$
$$d(P_3, P_c) = ((0-0)^2 + (1-0)^2 + (3-0)^2)^{\frac{1}{2}} = \sqrt{10}$$
$$d(P_4, P_c) = ((0-0)^2 + (1-0)^2 + (2-0)^2)^{\frac{1}{2}} = \sqrt{5}$$
$$d(P_5, P_c) = ((-1-0)^2 + (0-0)^2 + (1-0)^2)^{\frac{1}{2}} = \sqrt{2}$$
$$d(P_6, P_c) = ((1-0)^2 + (1-0)^2 + (1-0)^2)^{\frac{1}{2}} = \sqrt{3}$$

Para $k = 1$, classificamos P_c como **Verde**, pois P_5 é o ponto mais próximo de P_c.

Para $k = 3$, classificamos P_c como **Vermelho**, pois os pontos mais próximos P_2, P_5 e P_6 são vermelho, verde e vermelho e temos uma probabilidade de ser vermelho igual a $\frac{2}{3}$.

Questão 6

R: Queremos que β torne mínima a soma dos quadrados dos erros

$$\Sigma_{i=1}^{n} \varepsilon_i^2 = \Sigma_{i=1}^{n} (Y_i - \beta X_i)^2$$

Portanto, o estimador para β é aquele que minimiza a função acima. Vamos derivar a função:

$$\frac{\mathrm{d}}{\mathrm{d}\beta} = \Sigma_{i=1}^{n}(Y_i - \beta X_i)^2 = -2\Sigma_{i=1}^{n} X_i Y_i + 2\beta\Sigma_{i=1}^{n} X_i^2$$

Igualando a zero, temos:

$$-2\Sigma X_i Y_i + 2\beta\Sigma_{i=1}^{n} X_i^2 = 0 \Rightarrow \hat{\beta} = \frac{\Sigma_{i=1}^{n} X_i Y_i}{\Sigma_{i=1}^{n} X_i^2}$$

$\hat{\beta}$ é o estimador que minimiza a função, pois a segunda derivada é positiva.

Questão 10

Item a) R:

```
library(ISLR)
attach(Auto)
#ajuste do modelo linear
ajuste <- lm(mpg~horsepower)
summary(ajuste)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

i) Sim, há uma relação entre as variáveis horsepower e mpg, conforme determinado pelo teste da hipótese nula de todos os coeficientes de regressão serem iguais a zero. Como a estatística F é muito maior do que 1 e o p-valor é próximo de zero, podemos rejeitar a hipótese nula e afirmar que há uma relação estatisticamente significativa entre horsepower e mpg.

ii) Para calcular o erro residual relativo à variável resposta, usamos a média da resposta e o RSE. A média de mpg é $23,4459$. O RSE do ajuste foi de $4,906$, o que indica um erro percentual de $20,9248\%$. O R^2 do ajuste foi de cerca de $0,6059$, o que significa que $60,6\%$ da variação em mpg é explicada pelo horsepower.

iii) A relação entre as variáveis é negativa, pois o coefiente encontrado para "horsepower" foi de $-0,1578$. Ou seja, quanto mais horsepower tem um automóvel, a regressão linear indica que menor será o mpg.

iv)

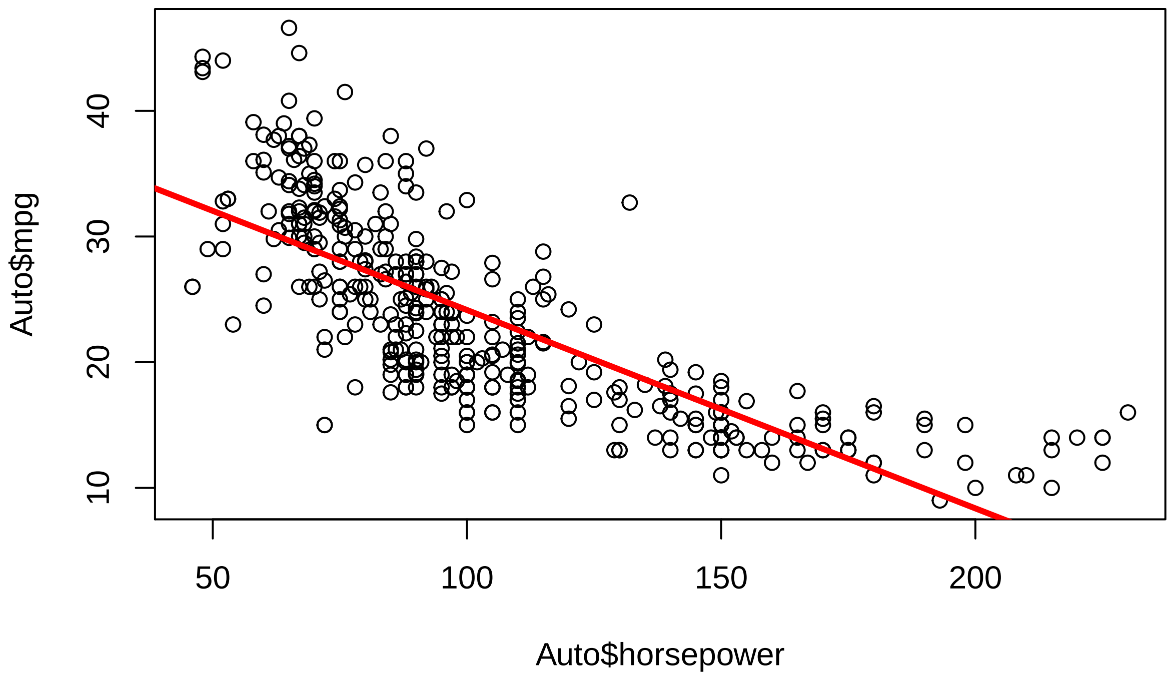
```
predict(ajuste, data.frame(horsepower=98),interval="confidence")
```

```
##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108
```

o valor de predição para $horsepower = 98$ é $mpg = 24,46708$ com o intervalo de confiança de 95% $[23,97308; 24,96108]$

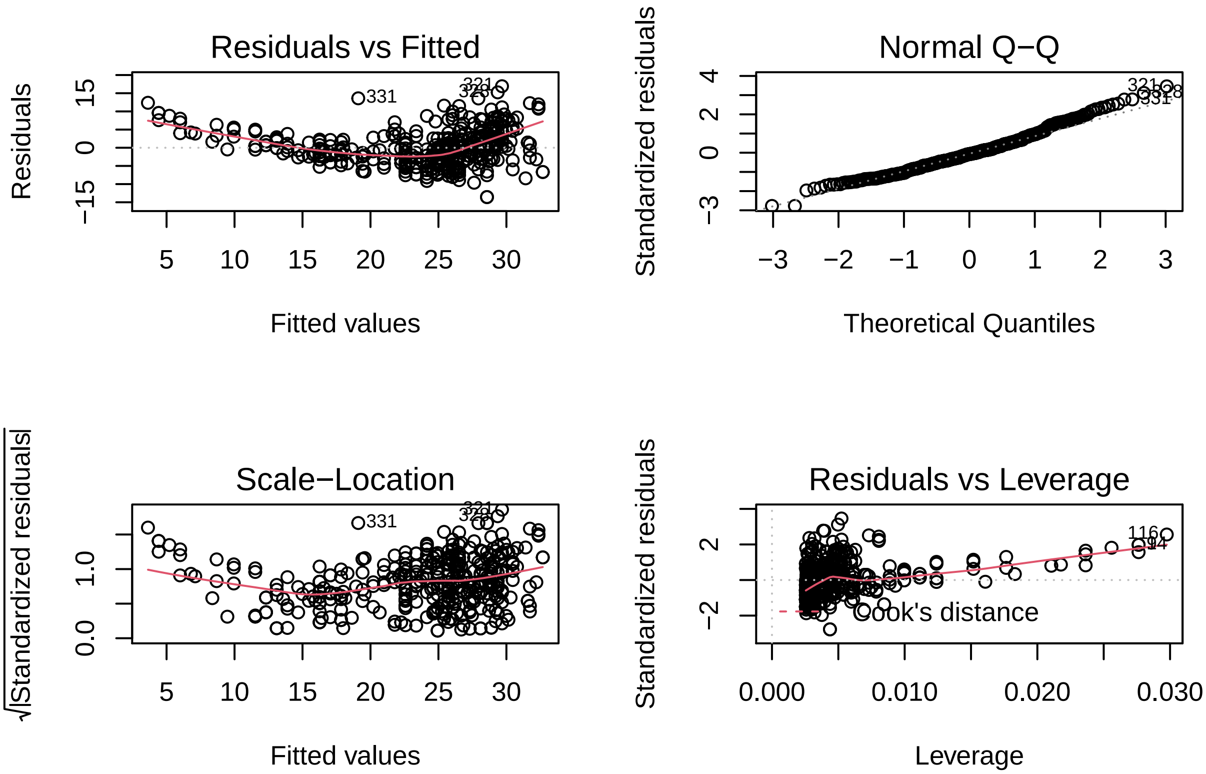
Item b) R:

```
plot(Auto$horsepower, Auto$mpg)
abline(ajuste, lwd=3, col="red")
```



Item c) R:

```
par(mfrow=c(2,2))
plot(ajuste)
```



O gráfico "Residuals x Fitted" mostra que os resíduos do ajustes apresentam tendência não linear, indicada pelo gráfico em formato de U.

Com o gráfico"Normal Q-Q", é possível visualizar que os resíduos estão distribuídos de acordo com uma Normal.

O gráfico "Scale-Location" mostra que as variâncias dos resíduos não são constantes, ou seja, é um caso de heterocedasticidade.

O gráfico "Residuals vs Leverage" mostra que não existem pontos que podem distorcer o ajuste da regressão linear.