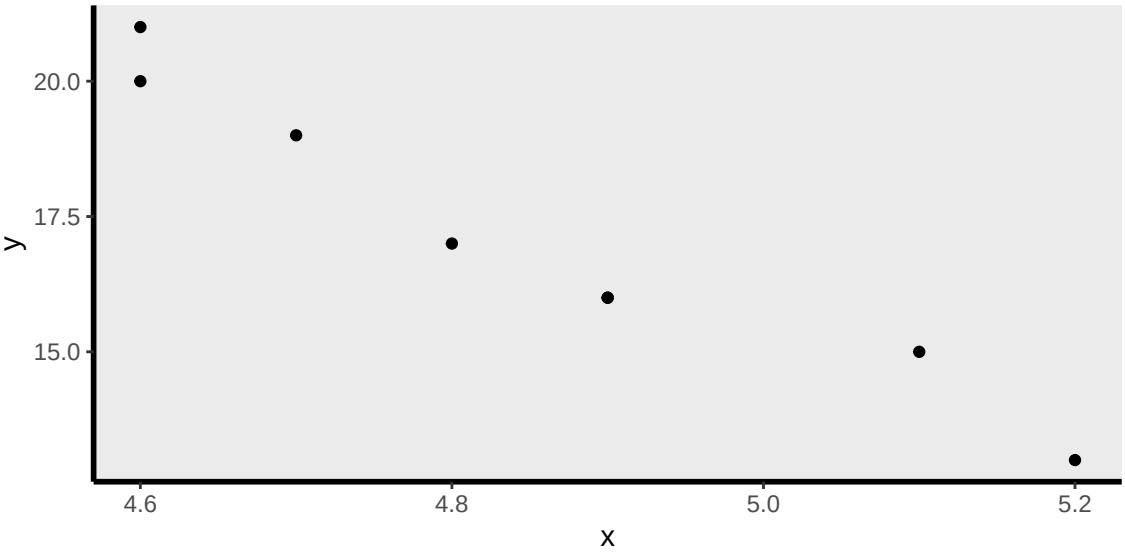
Regressão Linear Simples

EXERCÍCIO 1

a) Gráfico de dispersão dos dados



b) Modelo de regressão linear simples

A equação estimada é $\hat{y}_i = 74,897 - 11,912 \cdot x_i$.

c) Análise do modelo de regressão

```
##
## Call:
## lm(formula = y ~ x, data = dados_1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72059 -0.52941 -0.02941  0.27941  0.89706
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   74.897      5.514   13.58 9.88e-06 ***
## x            -11.912      1.136  -10.49 4.42e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6624 on 6 degrees of freedom
## Multiple R-squared:  0.9483, Adjusted R-squared:  0.9396
## F-statistic:   110 on 1 and 6 DF,  p-value: 4.416e-05
```

c.1) Resíduos

Os resultados apontam que os resíduos se encontram no intervalo $[-0,72059; 0.89706]$, com uma mediana próxima de zero, o que corrobora a hipótese do resíduo ter média zero.

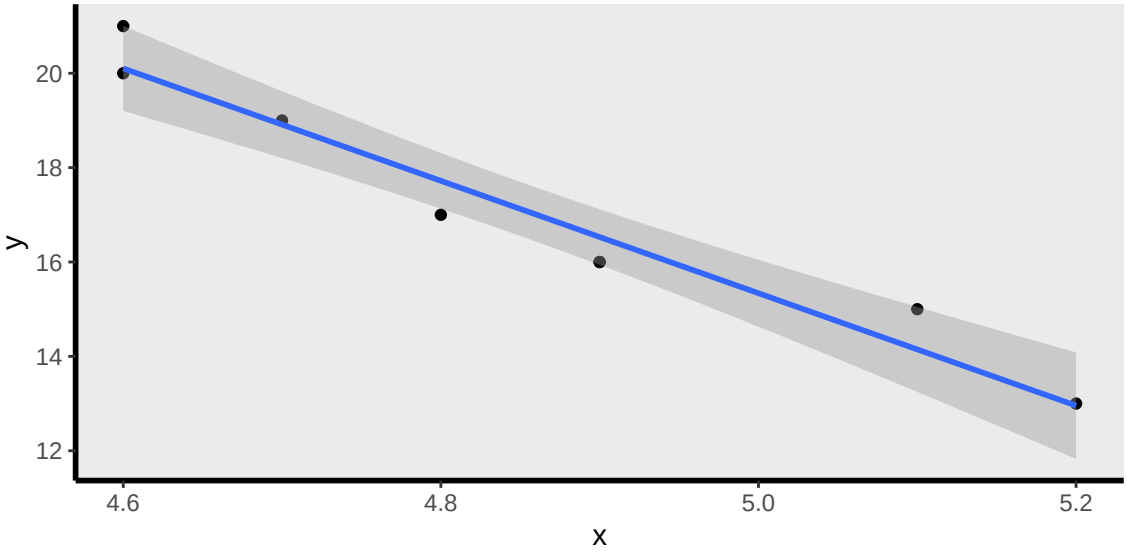
c.2) Significância estatística dos coeficientes

Tanto o intercepto quanto a variável explicativa x foram significativas a 1% de significância, com o resultado da regressão apontando que, tudo o mais constante, há uma redução de 11,912 unidades em y quando x aumenta em 1 unidade.

c.3) Percentual da variância explicada pelo modelo

Com um $R^2 = 0,9396$, obtemos que $93,96\%$ das variações em y são explicadas por variações em x.

d) Gráfico de dispersão com reta de regressão



EXERCÍCIO 2

a) Estrura e sumário estatístico dos dados

A estrutura dos dados é apresentada abaixo:

```
## 'data.frame':    1704 obs. of  8 variables:
##  $ pais      : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ ano       : int  1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 ...
##  $ pop       : num  8425333 9240934 10267083 11537966 13079460 ...
##  $ continente: chr  "Asia" "Asia" "Asia" "Asia" ...
##  $ expVida   : num  28.8 30.3 32 34 36.1 ...
##  $ pibPercap : num  779 821 853 836 740 ...
##  $ lpibPercap: num  6.66 6.71 6.75 6.73 6.61 ...
##  $ lexpVida  : num  3.36 3.41 3.47 3.53 3.59 ...
```

À princípio todas as variáveis parecem estar com formato correto. Verificando em seguida o sumário estatístico dos dados:

```
##      pais                ano            pop              continente
##  Length:1704        Min.   :1952   Min.   :6.001e+04   Length:1704
##  Class :character   1st Qu.:1966   1st Qu.:2.794e+06   Class :character
##  Mode  :character   Median :1980   Median :7.024e+06   Mode  :character
##                     Mean   :1980   Mean   :2.960e+07
##                     3rd Qu.:1993   3rd Qu.:1.959e+07
##                     Max.   :2007   Max.   :1.319e+09
##     expVida        pibPercap          lpibPercap        lexpVida
##  Min.   :23.60   Min.   :   241.2   Min.   : 5.485   Min.   :3.161
##  1st Qu.:48.20   1st Qu.:  1202.1   1st Qu.: 7.092   1st Qu.:3.875
##  Median :60.71   Median :  3531.8   Median : 8.170   Median :4.106
##  Mean   :59.47   Mean   :  7215.3   Mean   : 8.159   Mean   :4.060
##  3rd Qu.:70.85   3rd Qu.:  9325.5   3rd Qu.: 9.141   3rd Qu.:4.261
##  Max.   :82.60   Max.   :113523.1   Max.   :11.640   Max.   :4.414
```

Nenhuma variável apresentou valores discrepantes.

b) Classificação das variáveis

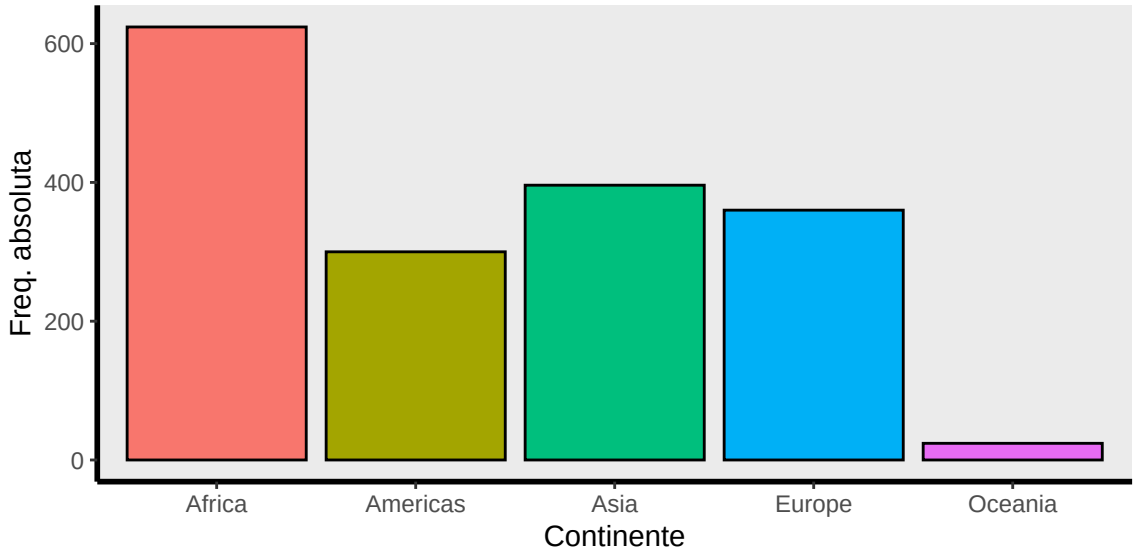
1. País - Qualitativa nominal
2. Ano - Quantitativa discreta
3. Pop - Quantitativa discreta
4. Continente - Qualitativa nominal
5. Exp. Vida - Contínua
6. pibPercap - Contínua

c) Tabelas de frequência do número de observações por continente (absoluta e relativa, respectivamente)

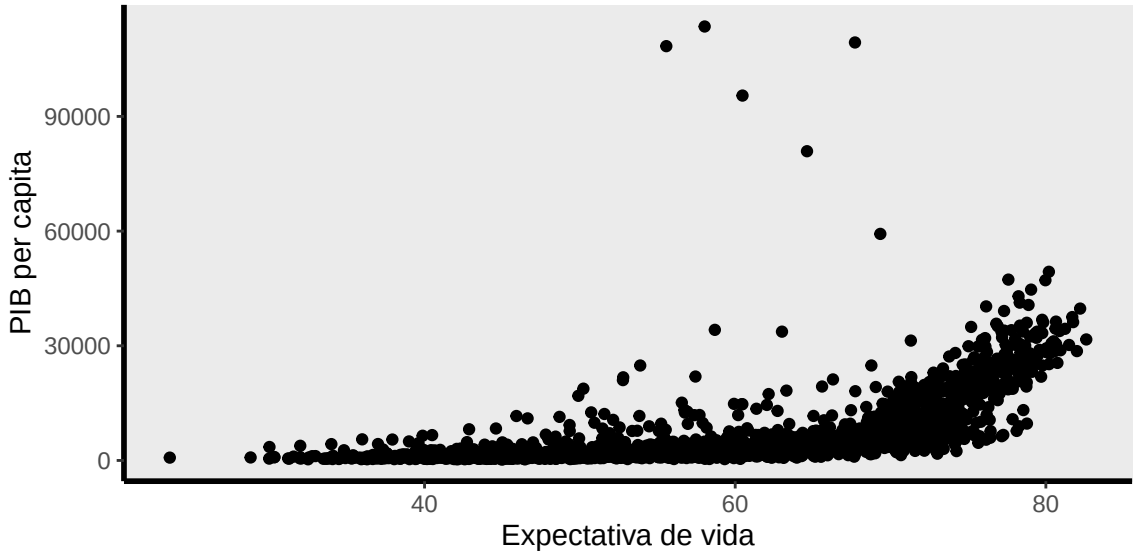
Continente	Observações
Africa	624
Americas	300
Asia	396
Europe	360
Oceania	24

Continente	Frequência
Africa	0.37
Americas	0.18
Asia	0.23
Europe	0.21
Oceania	0.01

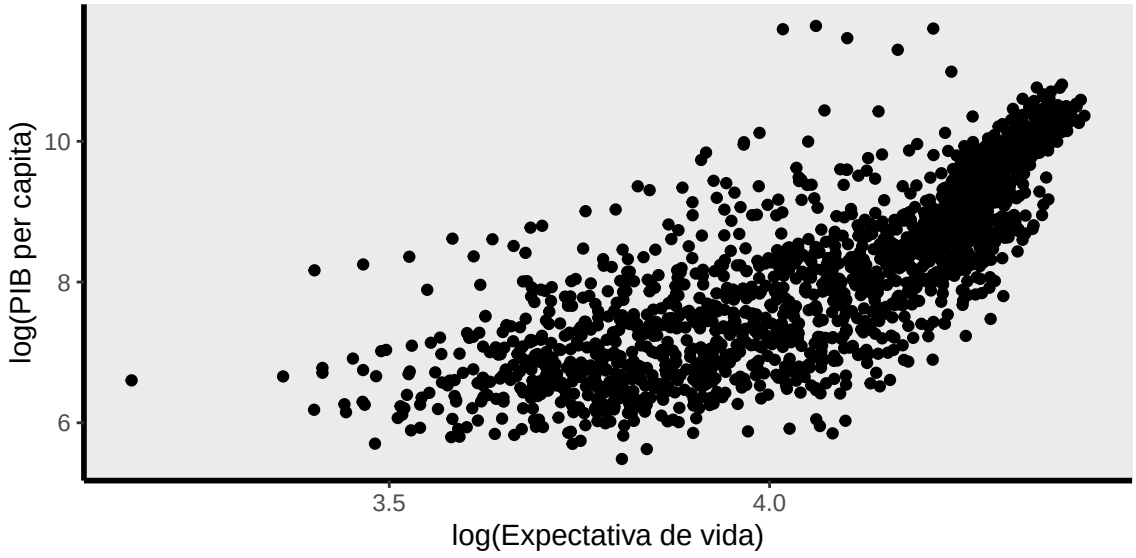
d) Gráfico de barras - frequência absoluta dos continentes



e) Gráfico de dispersão: PIB per capita x expectativa de vida



f) Gráfico de dispersão: log(PIB per capita) x log(expectativa de vida)



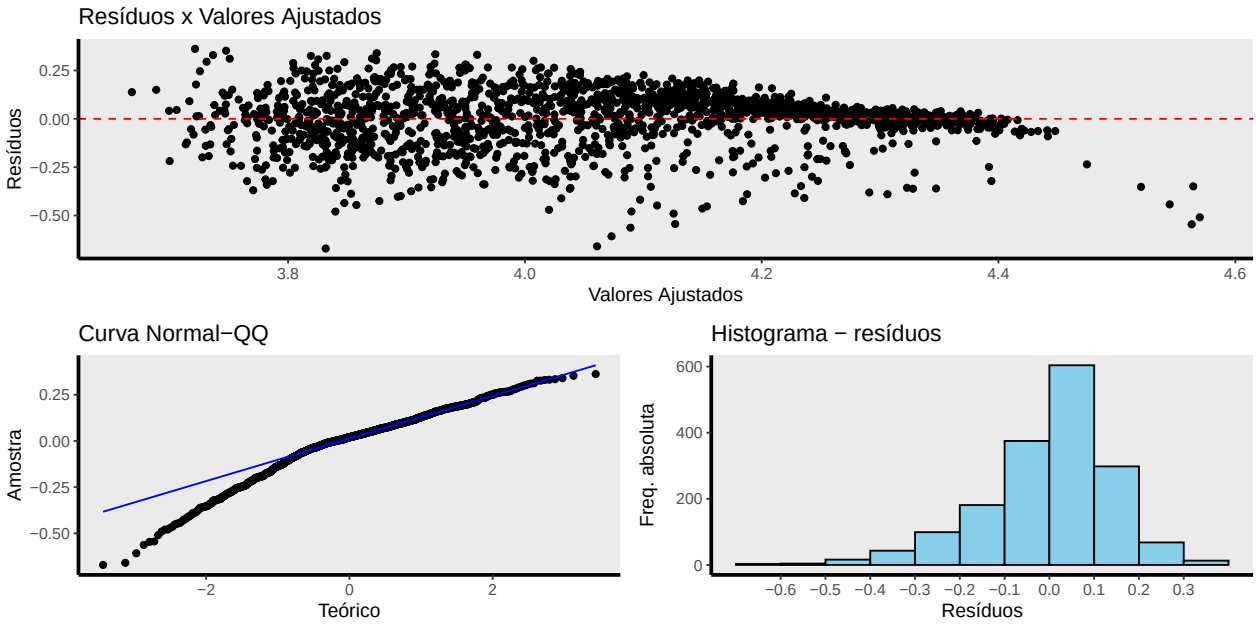
g) Modelo de regressão linear: log(expectativa de vida) em função do log(PIB per capita)

Inicialmente, temos a sumarização da regressão realizada:

```
##
## Call:
## lm(formula = lexpVida ~ lpibPercap, data = dados_2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67059 -0.06453  0.01978  0.09086  0.36156
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.864177   0.023283  123.02   <2e-16 ***
## lpibPercap  0.146549   0.002821   51.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1445 on 1702 degrees of freedom
## Multiple R-squared:  0.6132, Adjusted R-squared:  0.613
## F-statistic:  2698 on 1 and 1702 DF,  p-value: < 2.2e-16
```

A análise dos resíduos corrobora a hipótese de que eles têm média zero. Ao analisarmos a significância estatística das variáveis, verificamos que tanto o intercepto quanto o logaritmo do pib per capita são estatisticamente significativos a 1%. Além disso, a interpretação do modelo é de que um aumento de 1% no pib per capita está relacionado com um acréscimo de 0,14% na expectativa de vida. Por fim, $61,30\%$ das variações no variável dependente são explicadas por variações na variável independente.

Procedendo agora a análise dos gráficos de diagnósticos, temos o seguinte:



Podemos verificar que embora inicialmente o gráfico dos resíduos x valores ajustados pareça apontar que a hipóteses relativas aos resíduos estão sendo satisfeitas, conforme o valor ajustado aumenta, começamos a perceber a presença de *outliers* e uma tendência de redução da magnitude do resíduo, o que pode ser um indicativo de heterocedasticidade. De fato, se realizarmos o teste de Breusch-Pagan para heterocedasticidade, iremos obter que

```
##
##  studentized Breusch-Pagan test
##
## data:  reg_2
## BP = 38.206, df = 1, p-value = 6.366e-10
```

ou seja, como o p-valor é muito baixo, rejeitamos a hipótese nula de homocedasticidade.

Analisando agora o gráfico da curva QQ-Normal, podemos observar que inicialmente os resíduos não se ajustam bem à distribuição teórica. Apesar disso, vemos que o ajuste melhor depois, com o histograma confirmando que a ditribuição aparenta ser assimétrica à direita. Dessa forma, não podemos dizer que a hipótese de normalidade dos resíduos foi satisfeita.

EXERCÍCIO 3

a) Dicionário e carregamento dos dados

Analisando o dicionário, disponível no arquivo "dicionario.csv", podemos verificar que algumas variáveis foram importados no formato incorreto. Como não iremos usá-las na análise em questão, podemos seguir em frente.

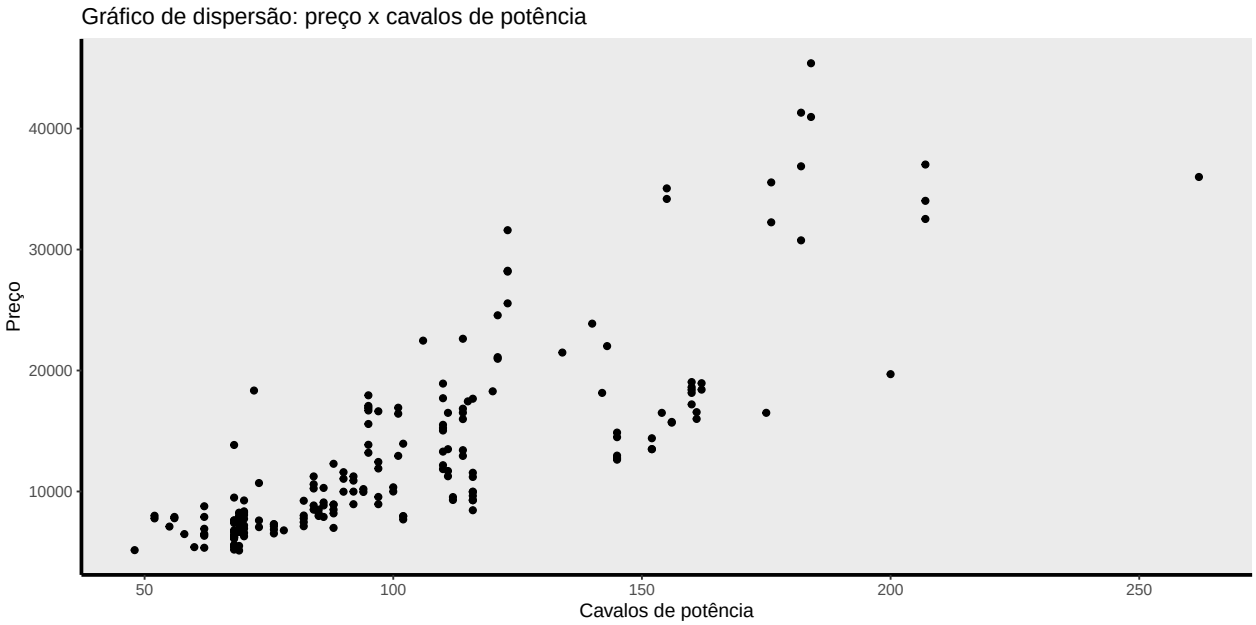
b) Exploração da base dados

Abaixo, temos os sumários estatísticos dos dados

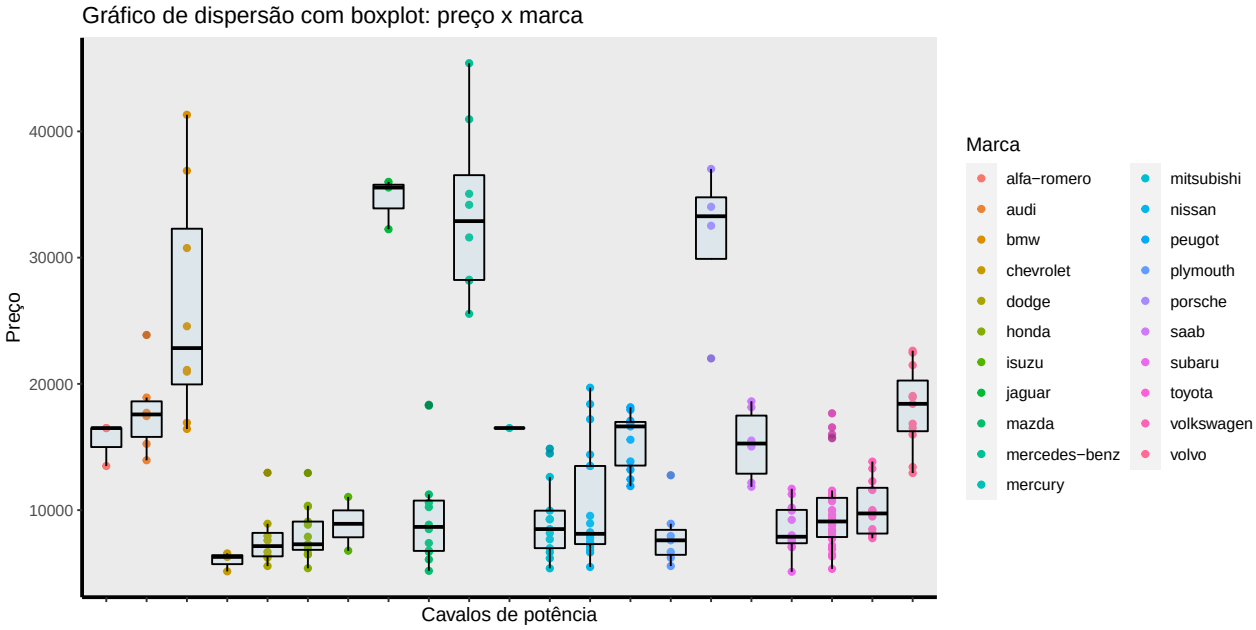
```
##      make            fuel.type          aspiration          num.doors
##  Length:193         Length:193         Length:193         Min.   :2.000
##  Class :character   Class :character   Class :character   1st Qu.:2.000
##  Mode  :character   Mode  :character   Mode  :character   Median :4.000
##                                                           Mean   :3.161
##                                                           3rd Qu.:4.000
```

```
##                                                         Max.   :4.000
##   body.style        drive.wheels       engine.location     wheel.base
##  Length:193         Length:193         Length:193         Length:193
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     length            width              height           curb.weight
##  Length:193         Length:193         Length:193         Min.   :1488
##  Class :character   Class :character   Class :character   1st Qu.:2145
##  Mode  :character   Mode  :character   Mode  :character   Median :2414
##                                                          Mean   :2562
##                                                          3rd Qu.:2952
##                                                          Max.   :4066
##  engine.type        num.cylinders    engine.size    fuel.system
##  Length:193         Min.   : 3.00   Min.   : 61.0   Length:193
##  Class :character   1st Qu.: 4.00   1st Qu.: 98.0   Class :character
##  Mode  :character   Median : 4.00   Median :120.0   Mode  :character
##                     Mean   : 4.42   Mean   :128.1
##                     3rd Qu.: 4.00   3rd Qu.:146.0
##                     Max.   :12.00   Max.   :326.0
##      bore              stroke          compression.ratio    horsepower
##  Length:193         Length:193         Length:193         Min.   : 48.0
##  Class :character   Class :character   Class :character   1st Qu.: 70.0
##  Mode  :character   Mode  :character   Mode  :character   Median : 95.0
##                                                          Mean   :103.5
##                                                          3rd Qu.:116.0
##                                                          Max.   :262.0
##     peak.rpm       city.mpg      highway.mpg        price
##  Min.   :4150   Min.   :13.00   Min.   :16.00   Min.   : 5118
##  1st Qu.:4800   1st Qu.:19.00   1st Qu.:25.00   1st Qu.: 7738
##  Median :5100   Median :25.00   Median :30.00   Median :10245
##  Mean   :5100   Mean   :25.33   Mean   :30.79   Mean   :13285
##  3rd Qu.:5500   3rd Qu.:30.00   3rd Qu.:34.00   3rd Qu.:16515
##  Max.   :6600   Max.   :49.00   Max.   :54.00   Max.   :45400
```

Em seguida, alguns gráficos exploratórios interessantes são



que mostra que, de forma geral, quanto maior a quantidade de cavalos de potência, maior o preço e



que mostra como se dá a distribuição dos preços de acordo com a marca do veículo.

c) Regressão do preço dos carros em função dos cavalos de potência

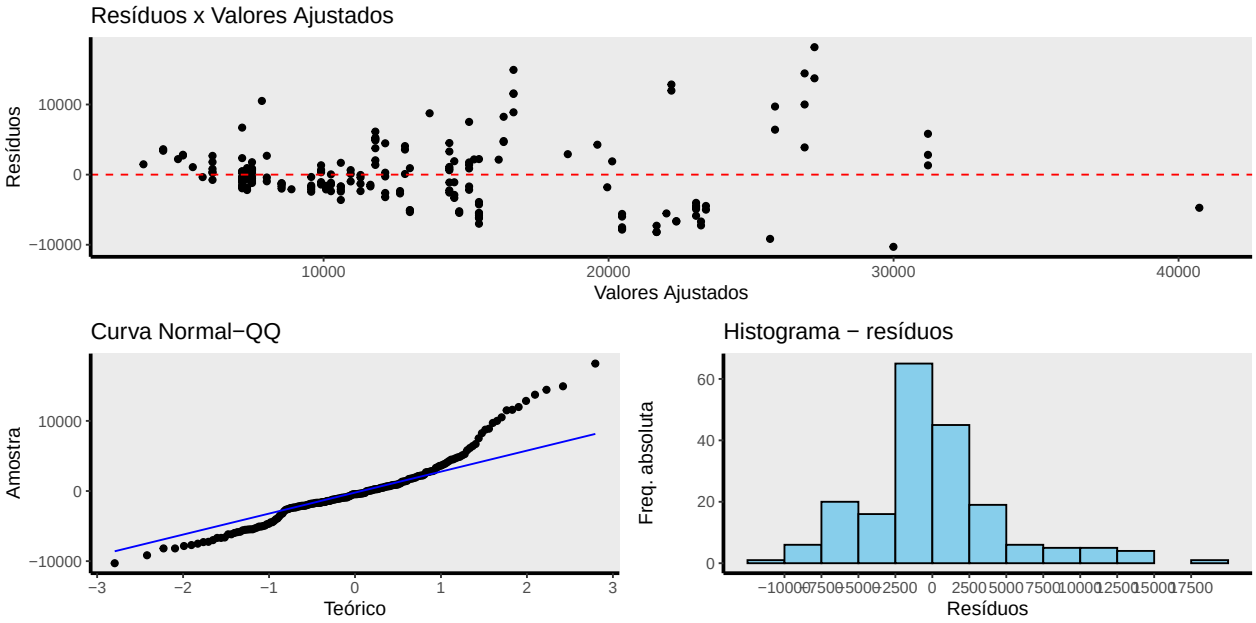
Intuitivamente, faz sentido que quanto maior a potência do carro, mais caro ele seja, uma vez que os componentes dos veículos mais potentes provavelmente são mais caros que os dos demais. Isso é confirmado ao observamos o resultado da regressão abaixo:

```
##
## Call:
## lm(formula = price ~ horsepower, data = dados_3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -10296.1  -2243.5   -450.1   1794.7  18174.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4630.70     990.58  -4.675 5.55e-06 ***
## horsepower    173.13       8.99  19.259  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4728 on 191 degrees of freedom
## Multiple R-squared:  0.6601, Adjusted R-squared:  0.6583
## F-statistic: 370.9 on 1 and 191 DF,  p-value: < 2.2e-16
```

Os resíduos obtidos são bem grandes, indicando que o ajuste do modelo não é tão bom. Apesar disso, tanto o intercepto quanto a variável de potência são estatisticamente significativos a 1% de significância. Temos que um aumento de uma unidade na potência eleva o preço do veículo em 173,13 unidades. Por fim, 66,01% das variações no preço são explicadas por variações na potência.

Analisando agora os graficos de diagnóstico, temos:



O gráfico dos resíduos x valores ajustados não parece bom. Os resíduos não aparentam se distribuir aleatoriamente ao redor do valor 0 e a variância não parece constante. De fato, ao realizarmos o teste de Breusch-Pagan,

```
##
##  studentized Breusch-Pagan test
##
## data:  reg_3
## BP = 51.873, df = 1, p-value = 5.92e-13
```

rejeitamos a hipótese nula de homocedasticidade.

Quanto a distribuição dos resíduos, observando a Curva Normal-QQ e o histograma, podemos perceber que claramente eles não seguem uma distribuição normal.

c.1) Interpretação do ajuste

Veja que o valor do intercepto é $-4630,70$, o que não faz sentido. Como a variável dependente mede preços, ela não pode ser negativa. Para corrigir isso, podemos estimar uma regressão que passa pela origem, assegurando que se a potência é 0, o preço também será. Os resultados dessa estimação são apresentados abaixo:

```
##
## Call:
## lm(formula = price ~ horsepower + 0, data = dados_3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7055.9 -3303.4 -2154.4   747.4 20805.9
##
## Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
## horsepower  133.663      3.252    41.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4979 on 192 degrees of freedom
## Multiple R-squared:  0.8979, Adjusted R-squared:  0.8974
## F-statistic:  1689 on 1 and 192 DF,  p-value: < 2.2e-16
```

c.2) Análise do modelo

Repare que o valor do R^2 obtido com o modelo com intercepto, $66,01\%$, é um valor relativamente alto. Apesar disso, devido ao problema de heterocedasticidade visto no modelo, podemos concluir que a potência do carro não é suficiente para uma boa previsão do preço do carro (claro que isso depende do que se considera como uma "boa previsão"). O modelo provavelmente pode ser melhorado com a inclusão de mais variáveis.