Análise de Regressão Linear

Neste exemplo, analisaremos um conjunto de dados conhecido como **mtcars**. A base de dados contém características de diferentes modelos de carros e descreve como tais características se relacionam com o consumo de combustível medido em galões por milha (mpg).

Em nossa análise, utilizaremos o peso do automóvel (wt) como variável independente e a relacionaremos com a variável dependente (mpg) através de um modelo de regressão linear simples.

```
library(ggplot2)
data("mtcars"); head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Ajuste da regressão linear

Em nossa análise, utilizaremos o *software* estatístico R. Para modelos de regressão linear (simples ou múltipla), a aplica-se a função **lm**.

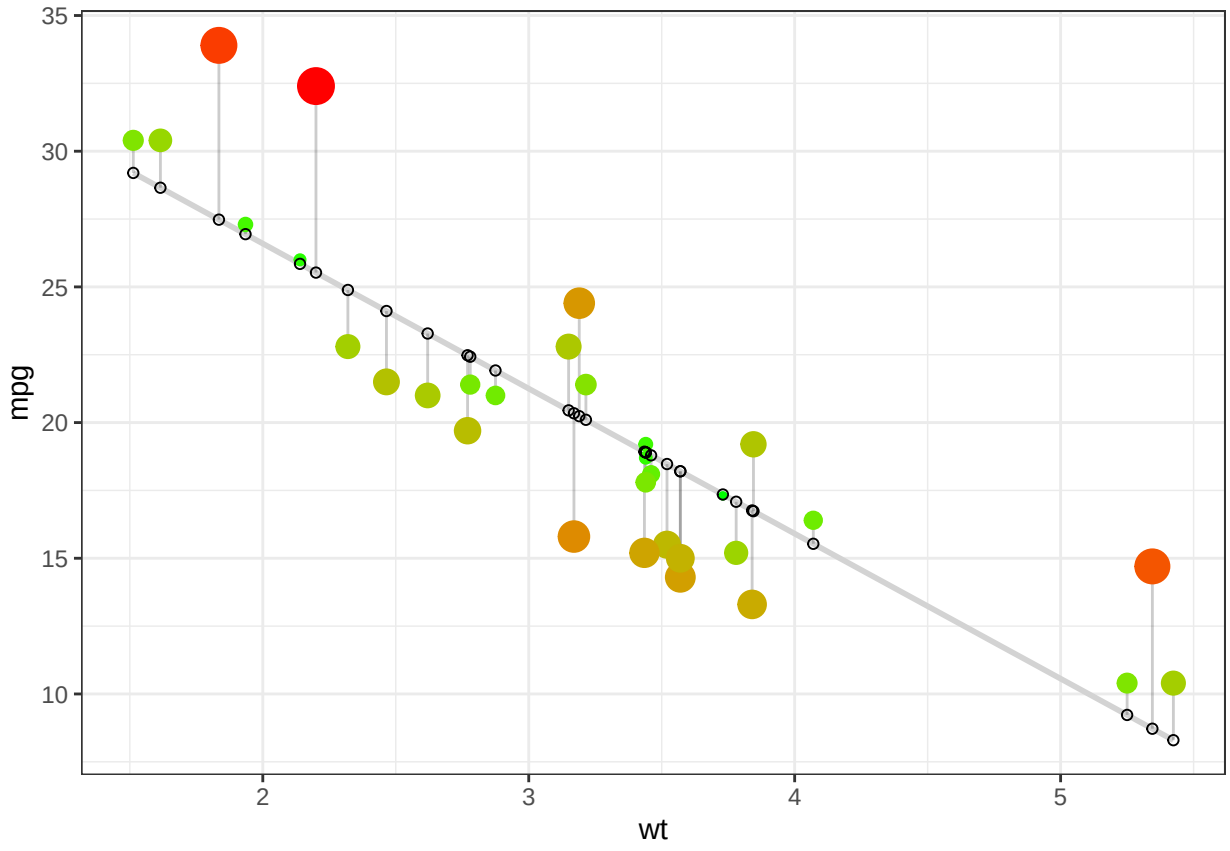
```
fit <- lm(mpg ~ wt, data = mtcars) # fit the model
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

Uma vez ajustado o modelo, podemos visualizar a reta resultante e os dados observados. A figura a seguir apresenta graficamente o tamanho do valor do resíduo atrevés da cor e do tamanho do ponto.

```r
mtcars$predicted <- predict(fit)   # Save the predicted values
mtcars$residuals <- residuals(fit) # Save the residual values
ggplot(mtcars, aes(x = wt, y = mpg)) +
  # regression line
  geom_smooth(method = "lm", se = FALSE, color = "lightgrey") +
  # draw line from point to line
  geom_segment(aes(xend = wt, yend = predicted), alpha = .2) +
  # size of the points
  geom_point(aes(color = abs(residuals), size = abs(residuals))) +
  # colour of the points mapped to residual size - green smaller, red larger
  scale_color_continuous(low = "green", high = "red") +
  # Size legend removed
  guides(color = "none", size = "none") +
  geom_point(aes(y = predicted), shape = 1) +
  theme_bw()
```

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

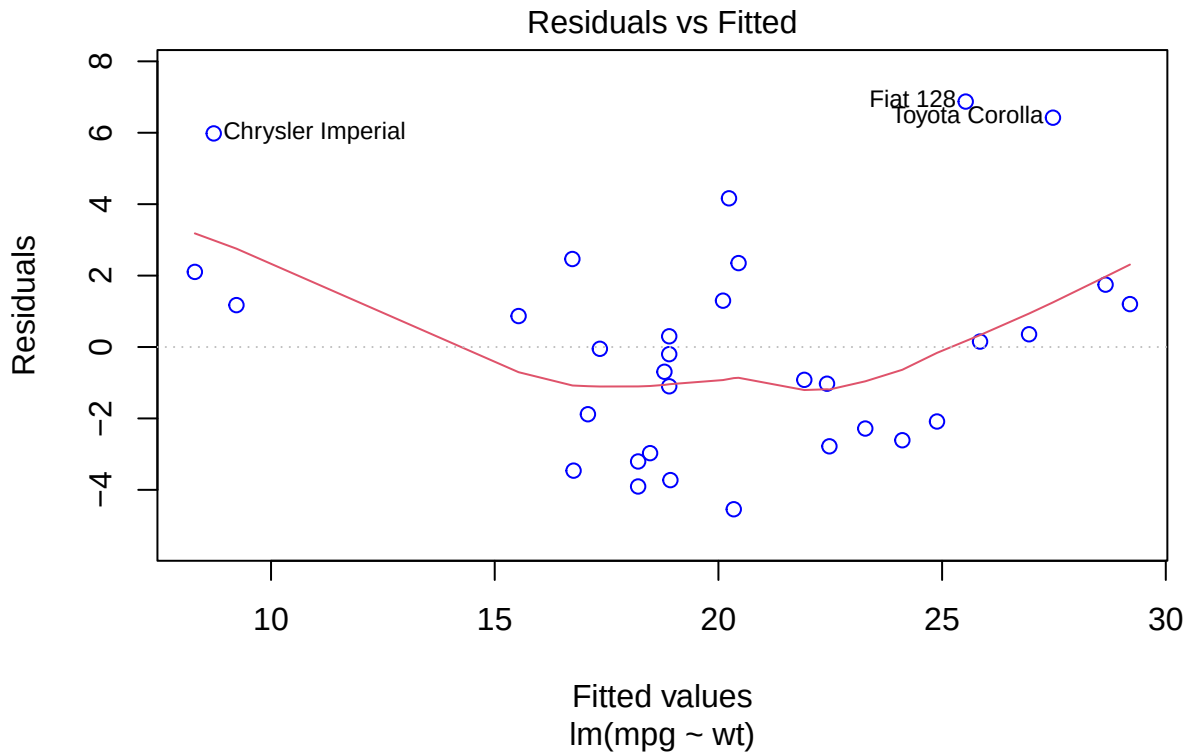


Análise de resíduos

Em um modelo de regressão linear, os resíduos representam a variação que não foi explicada.

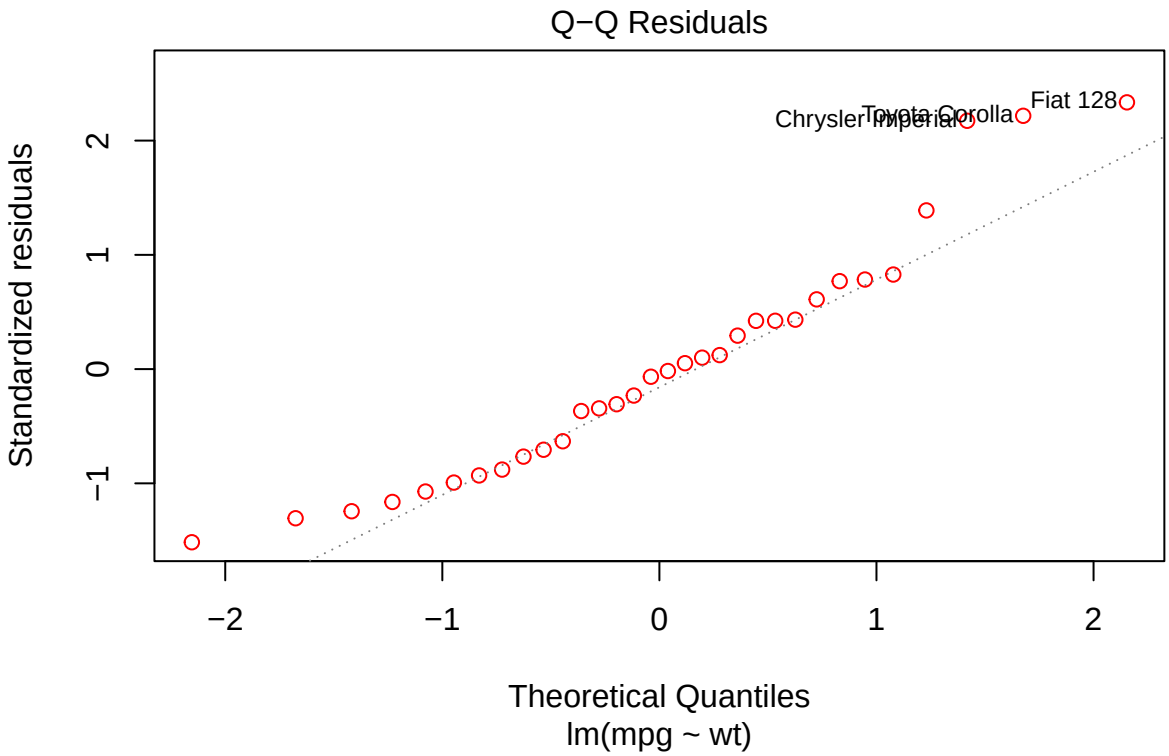
Usualmente, regresentações gráficas dos resíduos são utilizadas com o intuito de identificar padrões que indiquem, possivelmente, que o modelo possa não estar corretamente especificado.

```r
plot(fit, which=1, col=c("blue")) # Residuals vs Fitted Plot
```



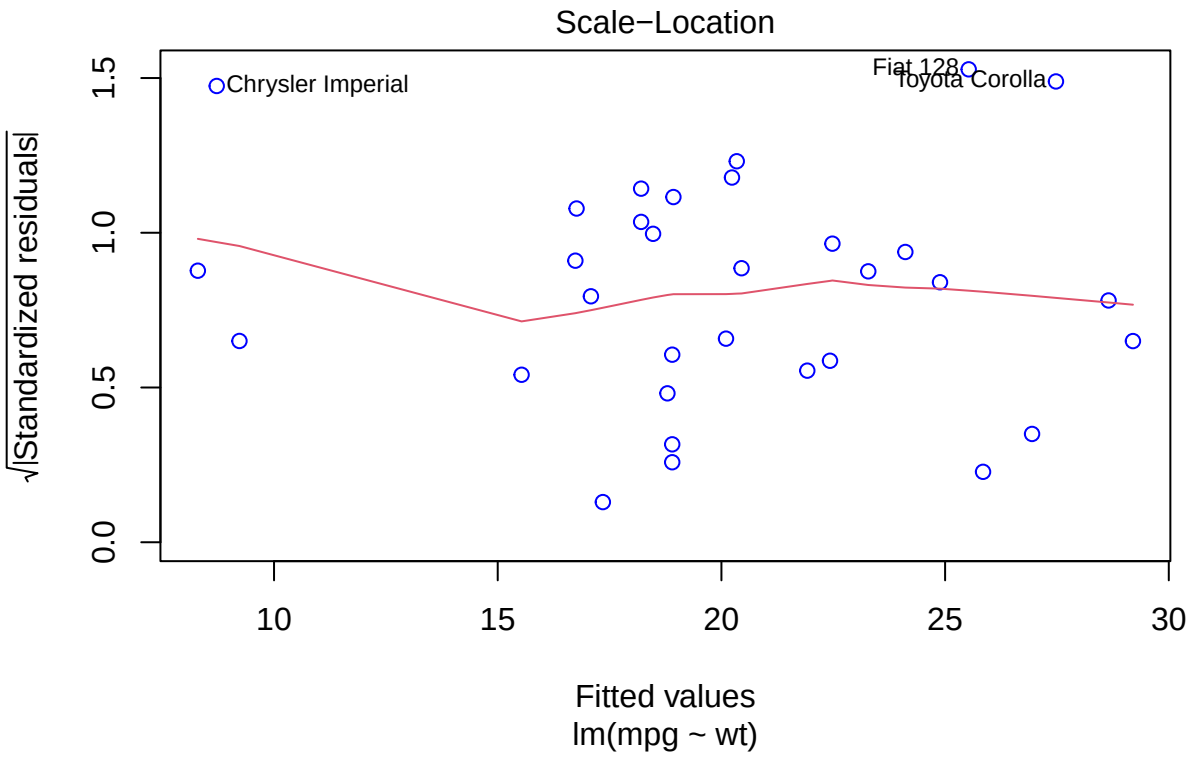
De acordo com as suposições dos modelos de regressão linear, assumimos que os erros possuem distribuição normal. A forma mais comum de verificar se tal suposição é satisfeita se dá ao representarmos os resíduos através de um *Q-Q plot* (*quantile-quantile plot*).

```r
plot(fit, which=2, col=c("red"))  # Q-Q Plot
```



Outra suposição do modelo assume variância constante nos modelos de regressão linear. Tal suposição pode ser verificada observando se os resíduos possuem ou não variância constante.

```
plot(fit, which=3, col=c("blue"))  # Scale-Location Plot
```


Scale−Location
Chrysler Imperial
Fiat 128
Toyota Corolla
√|Standardized residuals|
0.0
0.5
1.0
1.5
10
15
20
25
30
Fitted values
lm(mpg ~ wt)