REGRESSÃO LINEAR NO R

Para que serve a Regressão Linear Simples?

Utilizamos a regressão linear simples para descrever a relação linear entre duas variáveis. Com isso, ela é útil em algumas circunstâncias:

- Quando queremos prever o valor de uma variável pelo valor da outra
- Para entender se uma variável está relacionada com a outra
- Criar um modelo base antes de criar modelos de Regressão Linear Múltipla

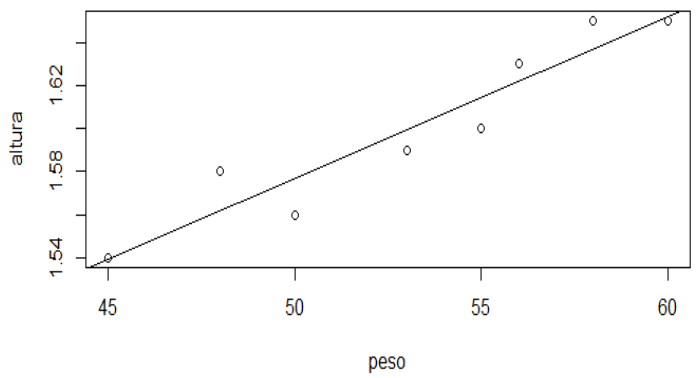
Como exemplos práticos, podemos:

- Avaliar o coeficiente de inteligência de acordo com a idade
- Entender se "insônia" é um preditor de "depressão"

CRIANDO AS VARIÁVEIS E PLOTANDO O GRÁFICO

No R, dados em tabelas são objetos do tipo *data frame*, nos quais cada coluna corresponde a uma variável e cada linha corresponde a uma observação. Neste exemplo, utilizaremos um conjunto de dados em que a variável resposta (Y) é o PESO, e a variável explicativa (X) é a ALTURA.

```
peso <- c(45,50,60,55,58,56,48,53)
altura <- c(1.54,1.56,1.65,1.60,1.65,1.63,1.58,1.59)

plot(peso, altura)
```



É possível observar um crescimento nos valores da variável ALTURA de acordo com o aumento dos valores da PESO. Portanto, esperamos que o efeito da ALTURA sobre o PESO seja positivo: $\beta > 0$.

Toda equação de linha reta tem uma estrutura padrão que é resumida pela seguinte fórmula: $Y = a + bx$.

Repare que, na fórmula, o valor "a" será sempre constante, sem a influência a outro coeficiente. É chamado, portanto, de coeficiente linear. Já o "b" é sempre multiplicado pelo ponto X, sendo alterado de acordo com este ponto. Desta forma, é considerado o coeficiente angular.

A função que realiza o ajuste da reta ou modelo de regressão linear no R é a *lm()*.

```
lm(altura ~ peso)
```

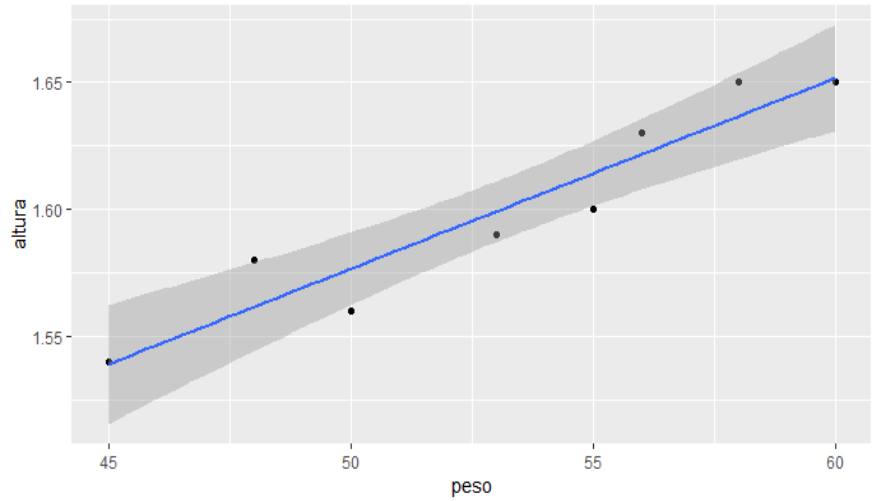
```
Call:
lm(formula = altura ~ peso)

Coefficients:
(Intercept)         peso
   1.200575     0.007519
```

E temos a equação da reta ajustada:

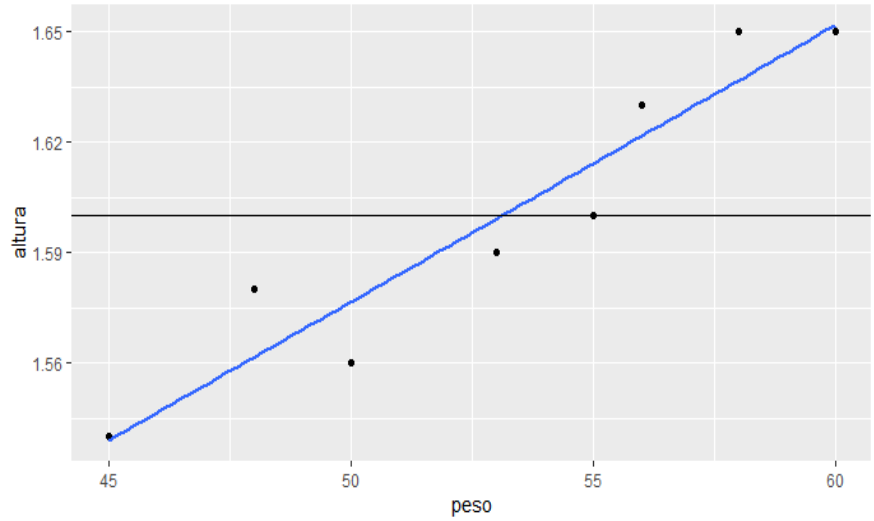
$$E(Y) = 1.200575 + 0.007519 * peso$$

```
abline(1.200575, 0.007519)
```

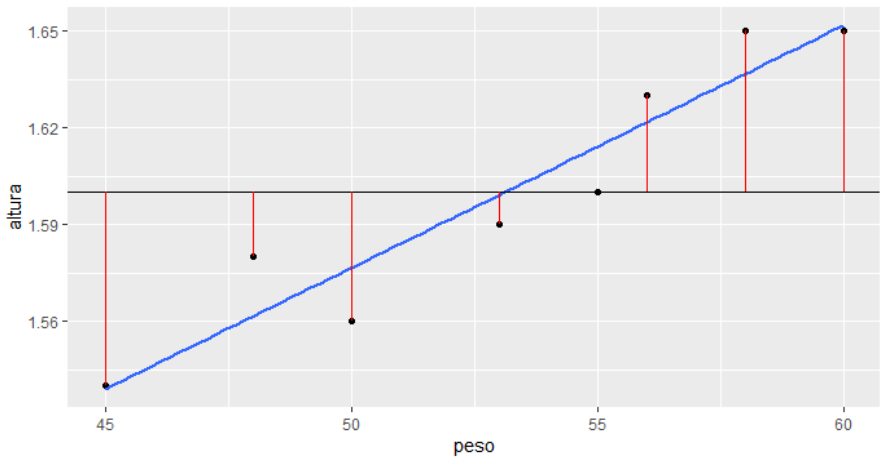
```
library(ggplot2)

ggplot(mapping = aes(peso, altura)) +
  geom_point() +
  geom_smooth(method = "lm")
```



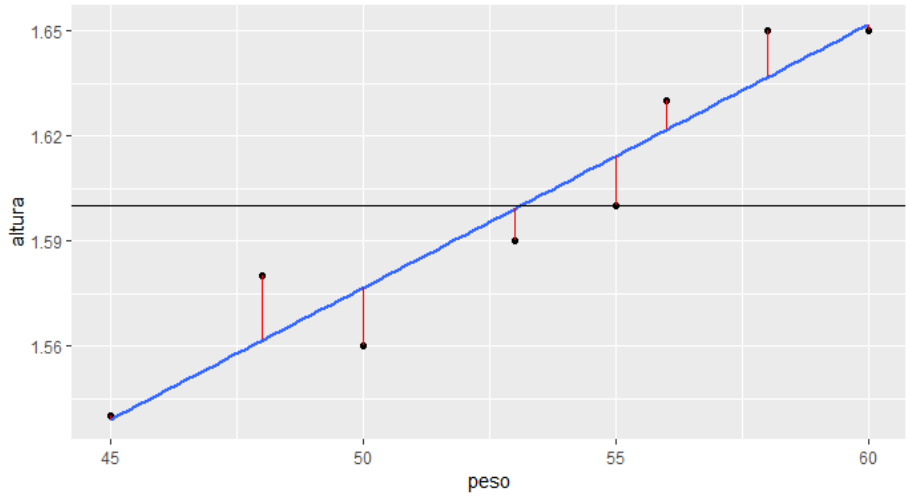

```
retas <- ggplot(mapping = aes(peso, altura)) +
  geom_point() +
  geom_smooth(se = FALSE, method = "lm") +
  geom_hline(yintercept = mean(altura))
retas
```




```
retas +
  geom_segment(aes(x = peso, y = altura,
  xend = peso, yend = mean(altura)), color="red")
```



```
retas +
  geom_segment(aes(x = peso, y = altura,
                   xend = peso, yend = predict(lm(altura ~ peso))),
color="red")
```



O comando *summary()* poderá indicar se os seus parâmetros estimados são significativos ou não, ou seja, se é possível assumir que são diferentes de zero.

```
summary(lm(altura ~ peso))
```

```
Call:
lm(formula = altura ~ peso)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0165044 -0.0103195 -0.0003009  0.0096247  0.0185328

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.200575   0.054293  22.113 5.59e-07 ***
peso        0.007519   0.001018   7.387 0.000316 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01384 on 6 degrees of freedom
Multiple R-squared:  0.9009,     Adjusted R-squared:  0.8844
F-statistic: 54.57 on 1 and 6 DF,  p-value: 0.0003158
```

Quanto mais asteriscos presentes ao lado do efeito estimado, maior o nível de confiança com que podemos afirmar que o efeito não é nulo.

Por mais que nosso R^2 tenha um efeito grande, também precisamos olhar o valor de significância da estatística F. Ele nos indica se o nosso modelo é significativamente diferente um modelo nulo.

```
SQt = sum((mean(altura) - altura)**2)
SQres = sum((predict(lm(altura ~ peso)) - altura)**2)
R2 = (SQt - SQres) / SQt
R2
```

```
0.9009349
```

O R^2 é uma medida que nos diz em quantos porcentos o a regressão linear é capaz de explicar a variação dos dados observados. No exemplo fica mais fácil: nosso R^2 é de 0,90. Com isso, podemos dizer que a regressão é capaz de determinar 90% da variação dos dados.

Não existe uma maneira padronizada de quanto deve ser o valor de R^2, devemos sempre interpretá-lo com base na teoria sobre nossas variáveis. Para fins de exemplo, podemos considerar que 90% da variação Da altura pode ser explicada pelo peso é um efeito grande.

Quanto ao R^2, ao utilizar apenas uma variável é normal que o valor não seja extremamente alto. De qualquer maneira, na prática, 0.56 é um valor bastante razoável.

A SEGUIR IREMOS FAZER UMA SIMULAÇÃO DE UMA PREDIÇÃO PARA ISSO CHAMAMOS A FUNÇÃO PREDICT.

```
predict(lm(altura ~ peso))
```

```
   1        2        3        4        5        6        7        8
1.538911 1.576504 1.651690 1.614097 1.636653 1.621616 1.561467 1.599060
```

Após predizer os valores iremos consultar valores de possivéis para pesos especificos nesse caso vamos prever a altura de 3 pessoas que tenham pesos 48,51 e 62.

```
pesos <- data.frame(peso = c(48, 51, 62))
predict(lm(altura ~ peso), pesos)
```

```
1        2        3
1.561467 1.584023 1.666728
```

Temos o resultado da previsão que nós retorna que a pessoa com 48 kg tenha uma altura de 1,56m, a pessoa com 51 kg tenha uma altura de 1,58m e a pessoa com 62kg tenha 1.66m.