

# ME613 - Análise de Regressão

Parte 1

Samara F. Kihl - IMECC - UNICAMP

## Introdução

### Introdução

Análise de regressão é uma metodologia estatística que utiliza a relação entre duas ou mais variáveis de maneira que uma variável pode ser predita pela(s) outra(s).

É uma metodologia amplamente utilizada em diversas áreas, pois é de relativamente fácil interpretação.

### Francis Galton e os dados sobre altura



Galton inventou os conceitos de correlação e regressão.

Em um de seus estudos, investigou a idéia de que filhos de pais altos eram altos, mas não tanto quanto seus pais. Filhos de pais mais baixos eram baixos, mas nem tanto quanto seus pais.

Ele se referiu a isso como "regressão para a média". Seus resultados foram publicados no artigo [Regression toward Mediocrity in Hereditary Stature](#).

## Francis Galton e os dados sobre altura

O trabalho de Galton, publicado seu trabalho em 1886, ainda é relevante.

Em 2009 pesquisadores compararam o modelo de Galton com modelos utilizando dados genômicos e o modelo de Galton se mostrou superior. Os resultados foram publicados no artigo [Predicting human height by Victorian and genomic methods](#).

Considere que tenhamos interesse nas seguintes questões:

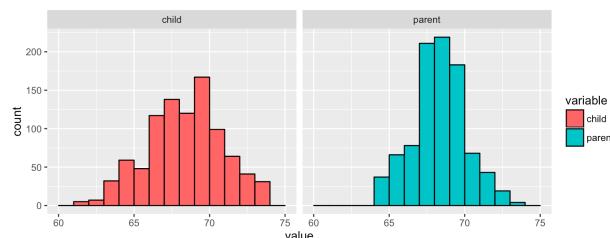
- Usar a altura dos pais para predizer a dos filhos.
- Encontrar uma maneira simples de explicar a relação entre a altura média dos pais e a altura média dos filhos.
- Investigar a variação na altura dos filhos que parece não estar relacionada com a altura dos pais (variação residual).
- Descobrir quais as suposições são necessárias para generalizar as descobertas realizadas com os dados em mãos para outros dados.
- Por que filhos de pais muito altos tendem a ser altos, mas nem tanto quanto seus pais.
- Por que filhos de pais muito baixos tendem a ser baixos, mas nem tanto quanto seus pais.

Modelos de regressão podem nos ajudar a responder essas perguntas.

5/49

6/49

## Francis Galton e os dados sobre altura



- Dados de 1885, usados por Galton.
- Observamos primeiramente as distribuições marginais: apenas dos pais, apenas dos filhos.
- Alturas das mulheres foram multiplicadas por 1.08

## Encontrando o centro via mínimos quadrados

Considerando apenas as alturas dos filhos:

- Se você fosse tentar adivinhar um valor para a altura de um dos filhos, utilizando apenas o histograma/dados dos filhos, como você faria?
- Como descrever o "centro/meio"?
- Um definição: seja  $Y_i$  a altura do filho  $i$  para  $i = 1, \dots, n = 928$ , então definimos o centro como o valor de  $\mu$  que minimiza

$$\sum_{i=1}^n (Y_i - \mu)^2$$

- É o centro de massa do histograma.
- Temos que  $\mu = \bar{Y}$  (demonstrar!)

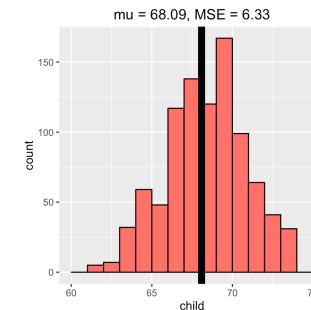
7/49

8/49

## Encontrando o centro via mínimos quadrados usando RStudio

```
library(UsingR); data(galton); library(reshape); long <- melt(galton)
library(manipulate)
library(ggplot2)
myHist <- function(mu){
  mse <- mean((galton$child - mu)^2)
  g <- ggplot(galton, aes(x = child)) + geom_histogram(fill = "salmon",
                                                       colour = "black", binwidth=1)
  g <- g + geom_vline(xintercept = mu, size = 3)
  g <- g + ggtitle(paste("mu = ", mu, ", MSE = ", round(mse, 2), sep = ""))
  g
}
manipulate(myHist(mu), mu = slider(62, 74, step = 0.5))
```

## A estimativa por mínimos quadrados é a média



$$\bar{y} = 68.09$$

9/49

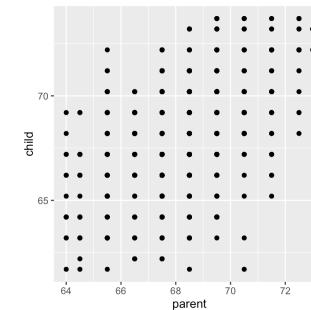
10/49

## Matematicamente

$$\begin{aligned} \sum_{i=1}^n (Y_i - \mu)^2 &= \sum_{i=1}^n (Y_i - \bar{Y} + \bar{Y} - \mu)^2 \\ &= \sum_{i=1}^n (Y_i - \bar{Y})^2 + 2 \sum_{i=1}^n (Y_i - \bar{Y})(\bar{Y} - \mu) + \sum_{i=1}^n (\bar{Y} - \mu)^2 \\ &= \sum_{i=1}^n (Y_i - \bar{Y})^2 + 2(\bar{Y} - \mu) \sum_{i=1}^n (Y_i - \bar{Y}) + \sum_{i=1}^n (\bar{Y} - \mu)^2 \\ &= \sum_{i=1}^n (Y_i - \bar{Y})^2 + 2(\bar{Y} - \mu)(\sum_{i=1}^n Y_i - n\bar{Y}) + \sum_{i=1}^n (\bar{Y} - \mu)^2 \\ &= \sum_{i=1}^n (Y_i - \bar{Y})^2 + \sum_{i=1}^n (\bar{Y} - \mu)^2 \\ &\geq \sum_{i=1}^n (Y_i - \bar{Y})^2 \end{aligned}$$

## Utilizando os dados dos pais

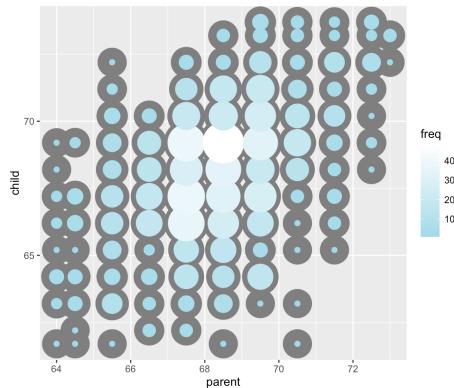
- Será que utilizando a informação dos pais conseguimos uma predição melhor para a altura do filho?
- Será que há alguma relação entre essas alturas? Como descobrir?



11/49

12/49

## Utilizando os dados dos pais



## Regressão pela origem

- Encontrar uma reta que melhor se ajusta aos pontos do gráfico.
- $X_i$  é a altura média dos pais da criança  $i$  e  $Y_i$  é a altura da criança  $i$ .
- $Y_i = \alpha + X_i\beta + \varepsilon_i$ , isto é  $\alpha + X_i\beta$  é a reta e  $\varepsilon$  um termo de erro.
- Precisamos encontrar o intercepto ( $\alpha$ ) e o coeficiente angular ( $\beta$ ) para definir a reta. Inicialmente, para simplificar,  $\alpha = 0$  (reta passa pela origem).
- Procuramos  $\beta$  que minimiza

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

- Estamos procurando uma reta, a partir da origem, que minimiza o quadrado das distâncias verticais dos pontos do gráfico até ela.

13/49

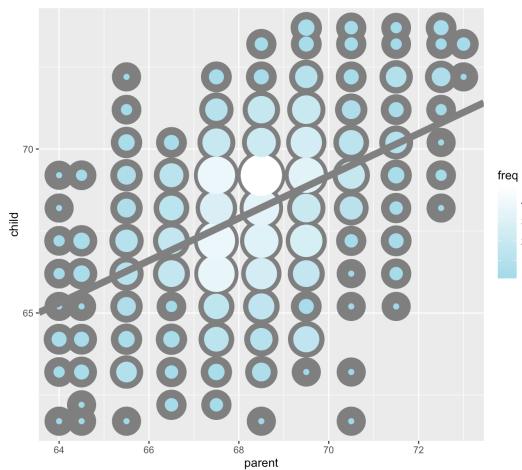
14/49

## Solução direta pelo R

```
myPlot <- function(beta){  
  y <- galton$child - mean(galton$child)  
  x <- galton$parent - mean(galton$parent)  
  freqData <- as.data.frame(table(x, y))  
  names(freqData) <- c("child", "parent", "freq")  
  plot(  
    as.numeric(as.vector(freqData$parent)),  
    as.numeric(as.vector(freqData$child)),  
    pch = 21, col = "black", bg = "lightblue",  
    cex = .15 * freqData$freq,  
    xlab = "parent",  
    ylab = "child"  
  )  
  abline(0, beta, lwd = 3)  
  points(0, 0, cex = 2, pch = 19)  
  mse <- mean((y - beta * x)^2)  
  title(paste("beta = ", beta, "mse = ", round(mse, 3)))  
}  
manipulate(myPlot(beta), beta = slider(0.6, 1.2, step = 0.02))
```

15/49

16/49



## Regressão Linear Simples

17/49

### Modelo de Regressão Linear Simples

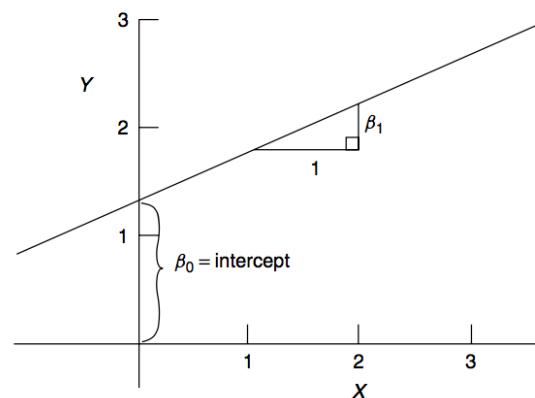
Em muitas situações a relação entre duas variáveis pode ser descrita por uma reta.

Podemos assumir que a reta de regressão da variável  $Y$  na variável  $X$  tem a forma  $\beta_0 + \beta_1 X$ .

Desta maneira, consideramos o modelo:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

ou seja, para um dado  $X_i$ , o valor correspondente de  $Y_i$  consiste no valor  $\beta_0 + \beta_1 X_i$  mais um termo  $\varepsilon_i$ , que é o incremento indicando a distância de  $Y_i$  da reta  $\beta_0 + \beta_1 X_i$ .



19/49

20/49

## Estimação por Mínimos Quadrados

$\beta_0, \beta_1$  e  $\varepsilon_i$  são desconhecidos.

$\varepsilon_i$  depende da observação  $i$ , mas  $\beta_0$  e  $\beta_1$  são constantes para todo  $i$ , portanto mais simples de serem estimados.

Suponha que tenhamos  $n$  pares de observações:  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ .

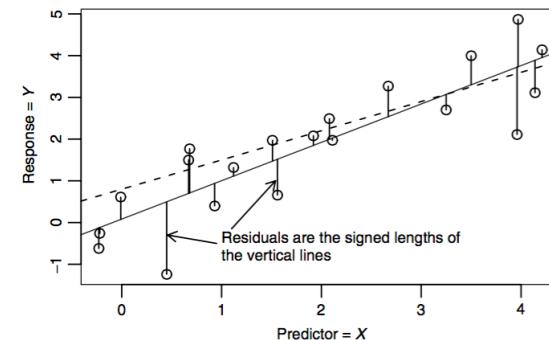
Equação do modelo de regressão linear:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

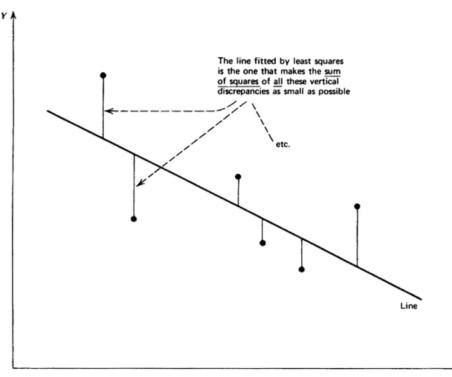
A soma dos desvios ao quadrado de cada observação  $Y_i$  até a reta  $\beta_0 + \beta_1 X_i$  é dada por:

$$S = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$

## Estimação por Mínimos Quadrados



## Estimação por Mínimos Quadrados



21/49

## Estimação por Mínimos Quadrados

Queremos encontrar  $\hat{\beta}_0$  e  $\hat{\beta}_1$  de tal forma que a reta resultante ao substituirmos esses valores no lugar de  $\beta_0$  e  $\beta_1$  minimiza a soma dos desvios ao quadrado.

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^n X_i(Y_i - \beta_0 - \beta_1 X_i)$$

Igualamos cada equação acima a zero para encontrar  $\hat{\beta}_0$  e  $\hat{\beta}_1$  que minimizam  $S$ .

23/49

22/49

24/49

## Estimação por Mínimos Quadrados

Equações normais:

$$\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$\sum_{i=1}^n X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

ou

$$\sum_{i=1}^n Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^n X_i = 0$$

$$\sum_{i=1}^n X_i Y_i - \hat{\beta}_0 \sum_{i=1}^n X_i - \hat{\beta}_1 \sum_{i=1}^n X_i^2 = 0$$

## Estimação por Mínimos Quadrados

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

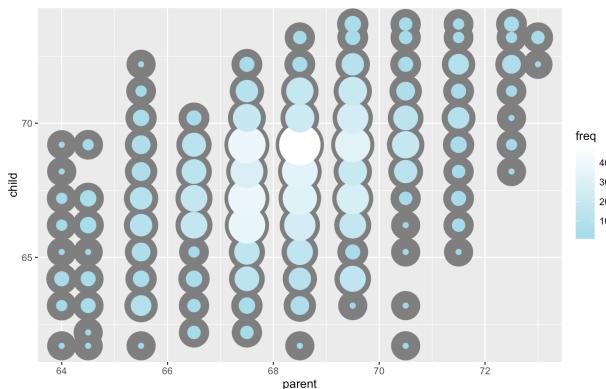
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Obtemos então a **equação de regressão estimada**:

$$\hat{Y}_i = \hat{E}(Y | X = X_i) = \hat{\beta}_0 + \hat{\beta}_1 X_i = \bar{Y} + \hat{\beta}_1(X_i - \bar{X})$$

[\(vídeo\)](#)

## Dados de Galton - estimação por mínimos quadrados



25/49

## Dados de Galton - estimação por mínimos quadrados

- Seja  $Y_i$  a altura da criança  $i$  e  $X_i$  a altura média dos pais da criança  $i$ .
- Encontrar a "melhor reta" (que minimiza  $S$ )
- $Y_i = \beta_0 + \beta_1 X_i$
- Mínimos quadrados: minimizar

$$S = \sum_{i=1}^n \{Y_i - (\beta_0 + \beta_1 X_i)\}^2$$

$$\text{Resultado: } \hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = Cor(X, Y) \frac{DP(X)}{DP(Y)} \text{ e } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

27/49

26/49

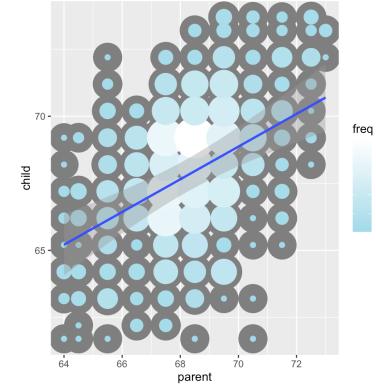
28/49

## Usando o R

O R tem a função `lm` para ajuste de regressão linear.

```
y <- galton$child
x <- galton$parent
beta1 <- cor(y, x) * sd(y) / sd(x)
beta0 <- mean(y) - beta1 * mean(x)
rbind(c(beta0, beta1), coef(lm(y ~ x)))

##      (Intercept)      x
## [1,]    23.94153 0.6462906
## [2,]    23.94153 0.6462906
```



$$\hat{\beta}_0 = 23.9415302 \text{ e } \hat{\beta}_1 = 0.6462906.$$

29/49

30/49

## Suposições do modelo de regressão linear simples

Nenhuma suposição sobre distribuição de probabilidade foi feita até o momento.

A partir de agora iremos assumir:

1.  $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, 2, \dots, n$
2.  $\varepsilon_i$  uma v.a. em que  $E(\varepsilon_i) = 0$  e  $Var(\varepsilon_i) = \sigma^2$  desconhecida, para  $i = 1, 2, \dots, n$ .
3.  $\varepsilon_i$  e  $\varepsilon_j$  são não-correlacionados para  $i \neq j$ , portanto  $Cov(\varepsilon_i, \varepsilon_j) = 0$  para  $i \neq j$ .

## Propriedades dos estimadores

32/49

## Propriedades de $Y_i$

O valor esperado para a resposta  $Y_i$  é:

$$\begin{aligned} E(Y_i) &= E(\beta_0 + \beta_1 X_i + \varepsilon_i) \\ &= \beta_0 + \beta_1 X_i + E(\varepsilon_i) \\ &= \beta_0 + \beta_1 X_i \end{aligned}$$

A variância para a resposta  $Y_i$  é:

$$\begin{aligned} Var(Y_i) &= Var(\beta_0 + \beta_1 X_i + \varepsilon_i) \\ &= Var(\varepsilon_i) = \sigma^2 \end{aligned}$$

A resposta  $Y_i$  vem de uma distribuição de probabilidade com  $E(Y_i) = \beta_0 + \beta_1 X_i$  (**função de regressão**) e  $Var(Y_i) = \sigma^2$ . A resposta,  $Y_i$ , está acima ou abaixo da função de regressão por um termo de erro  $\varepsilon_i$ .

## Propriedades de $\hat{\beta}_1$

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} \left( \sum_{i=1}^n (X_i - \bar{X})Y_i - \underbrace{\sum_{i=1}^n (X_i - \bar{X})\bar{Y}}_0 \right) \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = \sum_{i=1}^n a_i Y_i \end{aligned}$$

33/49

34/49

## Propriedades de $\hat{\beta}_1$

$$\begin{aligned} E(\hat{\beta}_1) &= E\left(\sum_{i=1}^n a_i Y_i\right) = \sum_{i=1}^n a_i E(Y_i) \\ &= \sum_{i=1}^n a_i (\beta_0 + \beta_1 X_i) = \beta_0 \sum_{i=1}^n a_i + \beta_1 \sum_{i=1}^n a_i X_i \\ &= \beta_0 \underbrace{\sum_{i=1}^n a_i}_0 + \beta_1 \underbrace{\sum_{i=1}^n a_i X_i}_1 \\ &= \beta_1 \end{aligned}$$

## Propriedades de $\hat{\beta}_1$

$$\begin{aligned} Var(\hat{\beta}_1) &= Var\left(\sum_{i=1}^n a_i Y_i\right) \\ &= \sum_{i=1}^n a_i^2 Var(Y_i) = \sum_{i=1}^n a_i^2 \sigma^2 \\ &= \sigma^2 \sum_{i=1}^n a_i^2 \\ &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \end{aligned}$$

35/49

36/49

## Precisão de $\hat{\beta}_0$

Mostre<sup>1</sup> que

$$Var(\hat{\beta}_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)$$

Mostre<sup>2</sup> que

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Mostre<sup>3</sup> que  $E(\hat{\beta}_0) = \beta_0$ .

## Propriedades dos estimadores

Não importa a distribuição de probabilidade de  $e_i$  (anteriormente definimos apenas os momentos nas suposições do modelo), o método de mínimos quadrados fornece estimadores não-viesados para  $\beta_0$  e  $\beta_1$ .

37/49

38/49

## Análise de Variância

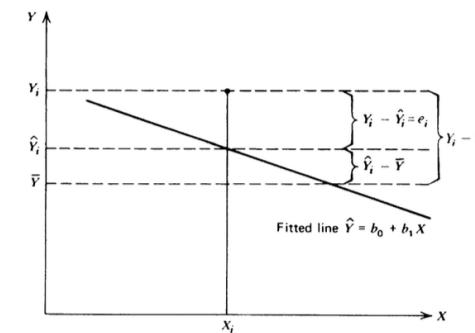
Quanto da variabilidade dos dados foi capturada pela reta ajustada?

$$Y_i - \hat{Y}_i = Y_i - \bar{Y} - (\hat{Y}_i - \bar{Y})$$

O resíduo  $e_i = Y_i - \hat{Y}_i$  é a diferença entre:

1. o valor observado  $Y_i$  e a média  $\bar{Y}$
2. o valor ajustado  $\hat{Y}_i$  e a média  $\bar{Y}$ .

## Análise de Variância



39/49

40/49

## Análise de Variância

$$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

Mostre<sup>4</sup> que elevando ambos os lados da equação ao quadrado e somando, temos:

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

## Análise de Variância

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \underbrace{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}_{SQT} + \underbrace{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}_{SQReg}$$

- SQT é a soma dos quadrados total (ajustada), representa a variação total de  $Y$  em torno de sua média.
- SQReg é a soma de quadrados da regressão, representa a variação total de  $\hat{Y}$  em torno de sua média.
- SQE é a soma dos quadrados do erro, representa a variação de  $Y$  em torno da reta ajustada.

41/49

42/49

## Análise de Variância

Queremos avaliar quão bom é nosso modelo de regressão.

Caso ideal: todas as observações se ajustam perfeitamente à reta. Neste caso:

$$SQE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = 0.$$

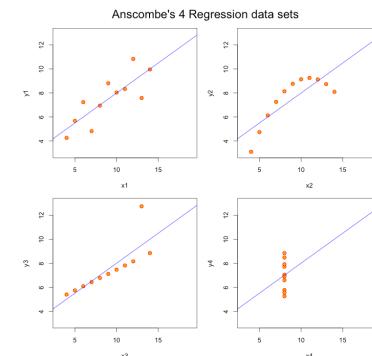
Para avaliar o modelo: observar quanto da SQT está contida em SQReg e quanto está na SQE.

Podemos utilizar para avaliar o modelo:

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

conhecido como **coeficiente de determinação**, que é a proporção da variabilidade total explicada pelo modelo de regressão ajustado.

Cuidado!



$R^2 = 0.66$  em todos os exemplos acima.

43/49

44/49

## Análise de Variância

| Fonte de Variação | gl      | SQ   |
|-------------------|---------|--|
| Regressão         | 1       | $SQReg = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ |
| Erro              | $n - 2$ | $SQE = \sum_{i=1}^n (Y_i - \hat{Y})^2$         |
| Total (ajustada)  | $n - 1$ | $SQT = \sum_{i=1}^n (Y_i - \bar{Y})^2$         |

## Análise de Variância

Utilizamos a tabela de análise de variância para comparar dois modelos:

$$1. Y_i = \beta_0 + \varepsilon_i \quad i = 1, 2, \dots, n$$

$$2. Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, 2, \dots, n$$

Para o modelo 1, vimos que  $\hat{\beta}_0 = \bar{Y}$ , portanto a soma dos quadrados dos resíduos deste modelo é dada por:

$$\sum_{i=1}^n (Y_i - \hat{\beta}_0)^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

Para o modelo 2, a soma dos quadrados dos resíduos é dada por:

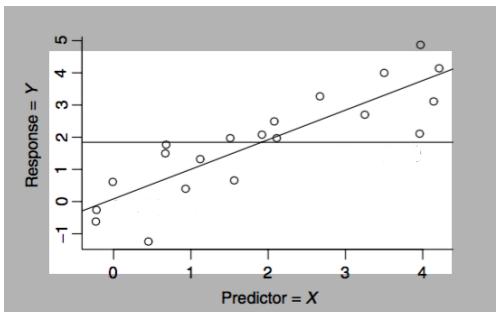
$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

\$

45/49

46/49

## Análise de Variância



Comparando modelo 1 (reta horizontal,  $X$  e  $Y$  não apresentam relação linear) com o modelo 2.

## $\sigma^2$ é desconhecido

Como  $\sigma^2$  desconhecido,  $Var(\hat{\beta}_0)$ ,  $Var(\hat{\beta}_1)$  e  $Cov(\hat{\beta}_0, \hat{\beta}_1)$  também são desconhecidos.

Um estimador não viésado para  $\sigma^2$  é:

$$s^2 = QME = \frac{SQE}{n - 2} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - 2}$$

47/49

48/49

## Leitura

- Applied Linear Statistical Models: Seções 1.1 a 1.7, 2.1, 2.2.
- Caffo - [Regression Models for Data Science in R](#): Introduction, Notation, Ordinary least squares, Regression to the mean, Statistical linear regression models, Residuals.
- Draper & Smith - [Applied Regression Analysis](#): Capítulo 0, 1.1 a 1.3.
- Weisberg - [Applied Linear Regression](#): Capítulo 1, 2.1 a 2.5, 2.7.

