AULA 4: MODELOS LINEARES GERAIS I

Análise Quantitativa de Dados Ambientais

Thiago S. F. Silva - tsfsilva@rc.unesp.br

31 de Agosto de 2015

Programa de Pós Graduação em Geografia - IGCE/UNESP

OUTLINE

MODELOS LINEARES GERAIS

Classe de modelos do tipo $\mathbf{Y} = \mathbf{BX} + \mathbf{U}$, onde $\mathbf{Y}$ é um vetor de respostas, $\mathbf{B}$ é a matriz desenho (*design matrix*), $\mathbf{X}$ é uma matriz de variáveis explicativas, e $\mathbf{U}$ é uma matriz contendo os erros.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ 1 & x_{31} & \dots & x_{3k} \\ . & . & . & . \\ . & . & . & . \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

CUIDADO!

A nomenclatura é bastante confusa:

- General Linear Models (GLM)
- Generalized Linear Models (GLM)
- Generalized Linear Mixed Models (GLMM)
- Generalized Least Squares (GLS)

Regressão Linear Simples?

Regressão Linear Simples?

Sir Francis Galton, no século IX, observou que a relação entre alturas de pais e filhos parecia ”reverter”, ou ”regredir”para a média do grupo. A partir dessa observação, Sir Galton desenvolveu uma primeira formulação matemática para a regressão.

Regressão **Linear** Simples?

Regressão **Linear** Simples?

Os modelos de regressão são formados a partir de combinações lineares de variáveis, através de parâmetros

$Y = \beta_0 + \beta_1 X$ é uma combinação linear de um parâmetro linear β_0 e um parâmetro linear β_1 que multiplica X

$Y = \beta_0 + \beta_1 X + \beta_2 X^2$ também é uma combinação linear, de um parâmetro linear β_0, um parâmetro linear β_1 que multiplica X, e um parâmetro linear β_2 que multiplica X^2

Regressão **Linear** Simples?

$Y = \beta_0 + e^{\beta_1 X}$ não é uma combinação linear

A linearidade se refere aos parâmetros, e não às variáveis

$Y = \beta_0 + e^{\beta_1 X}$ não é uma combinação linear

Mas alguns modelos não-lineares podem ser linearizados:

$$Ln(Y) = Ln(\beta_0) + \beta_1 X$$

$$Z = W + \beta_1 X$$

$$Z = Ln(Y)$$

$$W = Ln(\beta 0)$$

Regressão Linear **Simples?**

Regressão Linear **Simples?**

Apenas duas variáveis são usadas, uma dependente (Y)
e uma independente (X)

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- X é a variável **independente**, também chamada de **preditor**

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- X é a variável **independente**, também chamada de **preditor**
- Y é a variável **dependente**, também chamada de **variável resposta**

TERMINOLOGIA

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- X é a variável **independente**, também chamada de **preditor**
- Y é a variável **dependente**, também chamada de **variável resposta**
- β_0 e β_1 são os **parâmetros** ou **coeficientes** da regressão

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- X é a variável **independente**, também chamada de **preditor**
- Y é a variável **dependente**, também chamada de **variável resposta**
- β_0 e β_1 são os **parâmetros** ou **coeficientes** da regressão
- ε é o **termo de erro**

CARACTERÍSTICAS DE UM MODELO DE REGRESSÃO

Os modelos de regressão expressam essencialmente:

- Uma tendência de Y em variar sistematicamente de acordo com o preditor X

CARACTERÍSTICAS DE UM MODELO DE REGRESSÃO

Os modelos de regressão expressam essencialmente:

- Uma tendência de Y em variar sistematicamente de acordo com o preditor X
- Uma dispersão de pontos ao redor da reta que descreve uma relação estatística

CARACTERÍSTICAS DE UM MODELO DE REGRESSÃO

Os modelos de regressão expressam essencialmente:

- Uma tendência de Y em variar sistematicamente de acordo com o preditor X
- Uma dispersão de pontos ao redor da reta que descreve uma relação estatística

Essas características são expressas através das pressuposições:

- Existe uma distribuição de probabilidade de Y para cada nível (valor) de X

CARACTERÍSTICAS DE UM MODELO DE REGRESSÃO

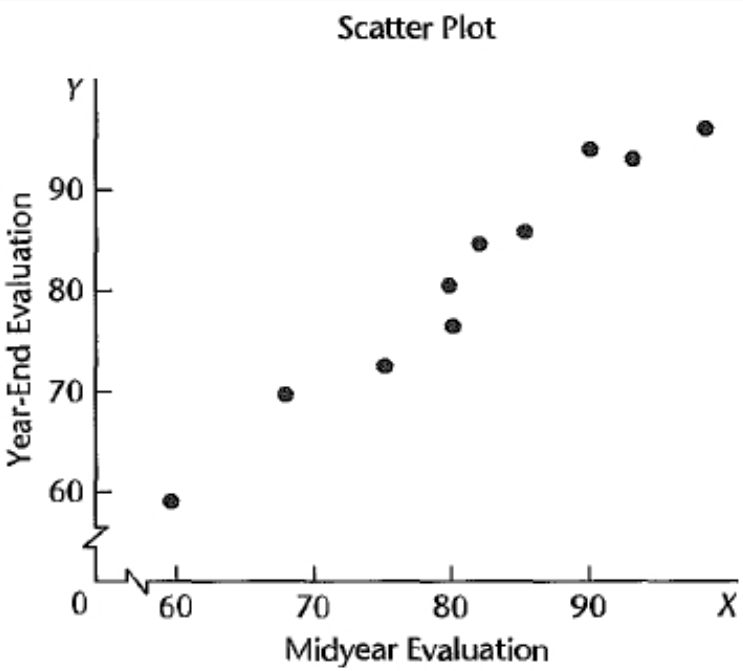
Os modelos de regressão expressam essencialmente:

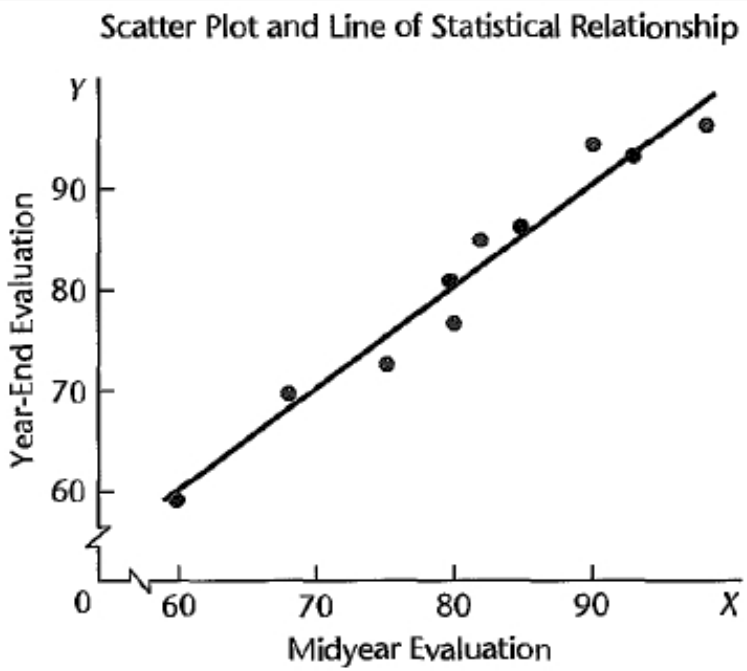
- Uma tendência de Y em variar sistematicamente de acordo com o preditor X
- Uma dispersão de pontos ao redor da reta que descreve uma relação estatística

Essas características são expressas através das pressuposições:

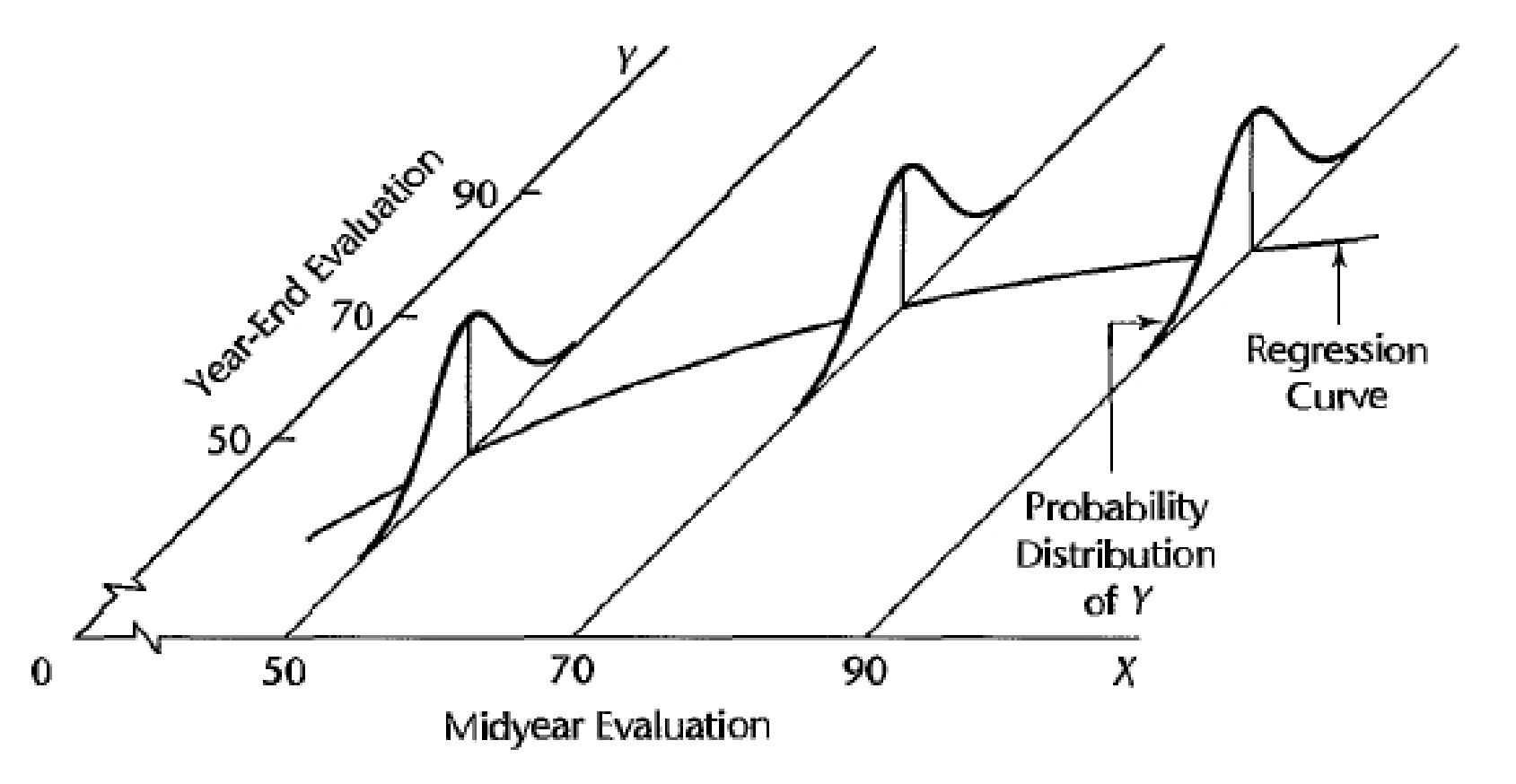
- Existe uma distribuição de probabilidade de Y para cada nível (valor) de X
- A média destas distribuições varia sistematicamente com X

EXEMPLO


Scatter Plot
Y
90
80
70
60
0
60
70
80
90
X
Year-End Evaluation
Midyear Evaluation

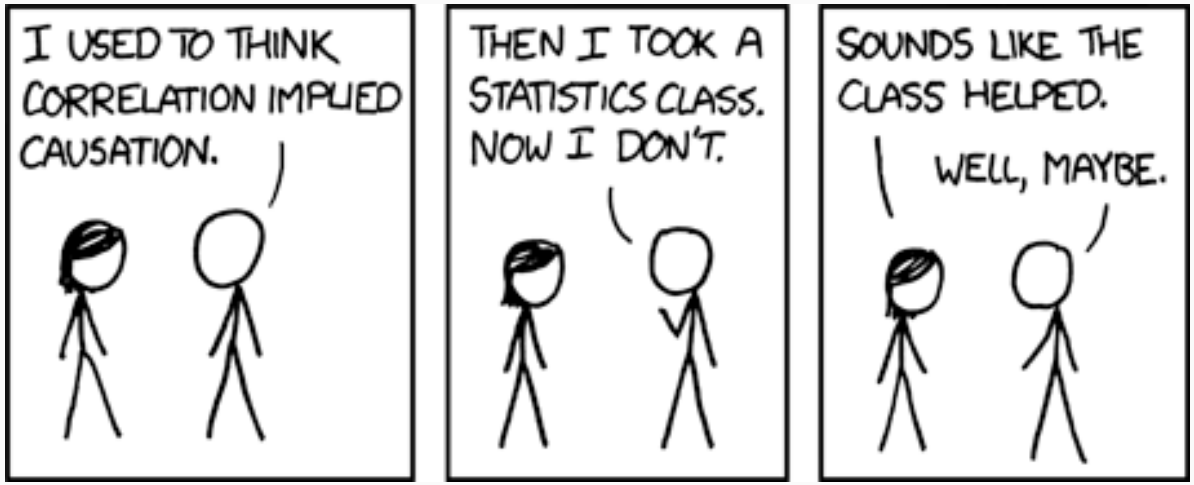

Scatter Plot and Line of Statistical Relationship
Y
90
80
70
60
0
60
70
80
90
X
Year-End Evaluation
Midyear Evaluation

EXEMPLO


Y
Year-End Evaluation
90
70
50
Regression
Curve
Probability
Distribution
of Y
0
50
70
90
X
Midyear Evaluation

A existência de uma (cor)relação estatística entre duas variáveis não implica em uma relação real de causalidade ou dependência.

Mesmo quando há causalidade, cuidado com a direção da relação: X causa Y, ou Y causa X?



http://xkcd.com/552/

FORMULAÇÃO DO MODELO DE REGRESSÃO

O modelo de regressão completo pode ser escrito como:

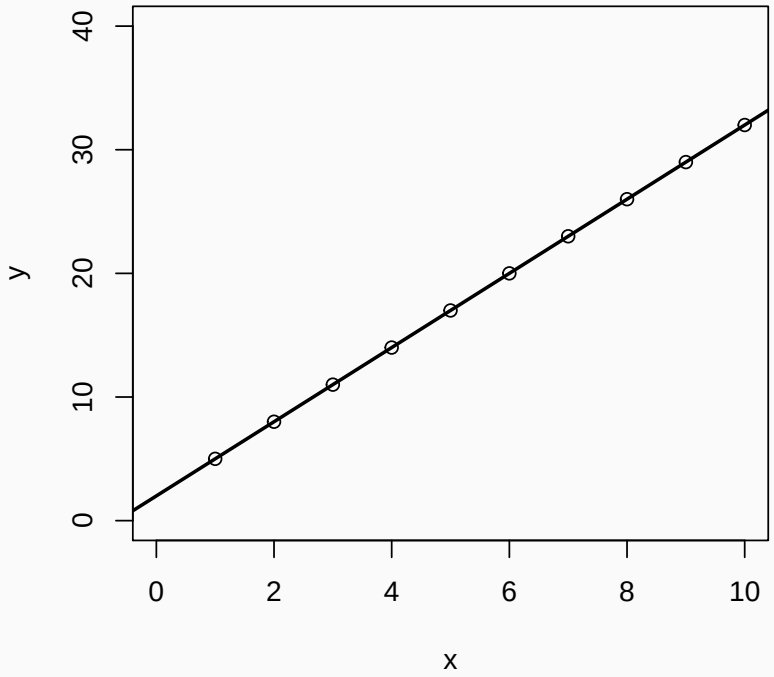
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

"O i-ésimo valor de Y é função de um parâmetro constante β_0, somado ao i-ésimo valor de X multiplicado por um parâmetro constante β_1, somado a um i-ésimo valor específico de erro".

Se Y_i puder ser predito exatamente por X_i, então

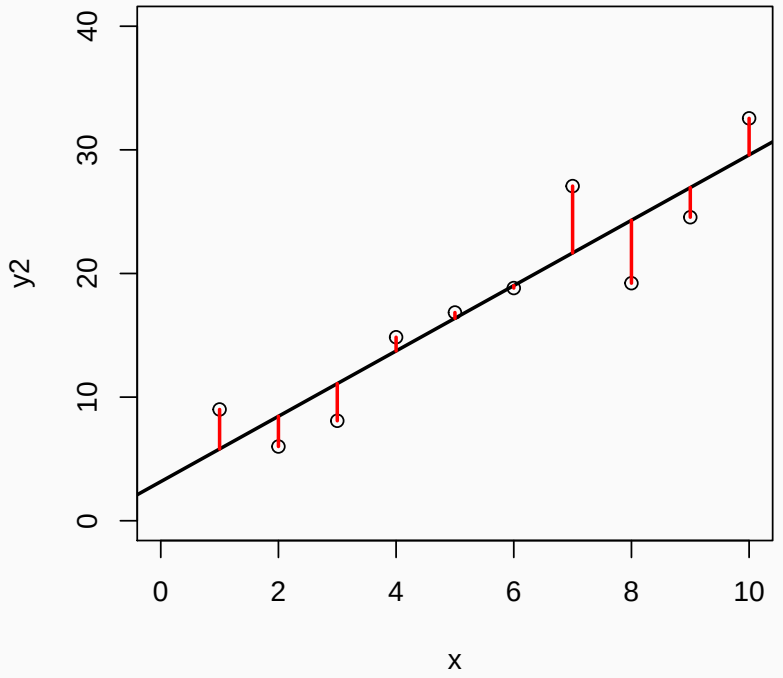
$\varepsilon_i \sim N(0, 0)$

$Y_i = 2 + 3X_i$



Se Y_i pode ser aproximado por X_i, então $\varepsilon_i \sim N(0, \sigma)$

$Y_i = 2 + 3X_i + \varepsilon \sim N(0, 3)$

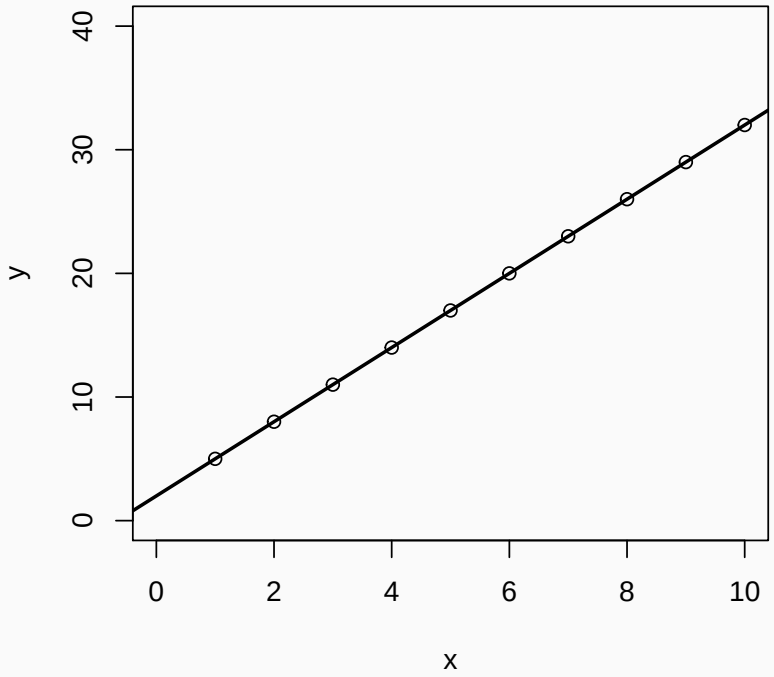


FORMULAÇÃO DO MODELO DE REGRESSÃO

Se Y_i puder ser predito exatamente por X_i, então

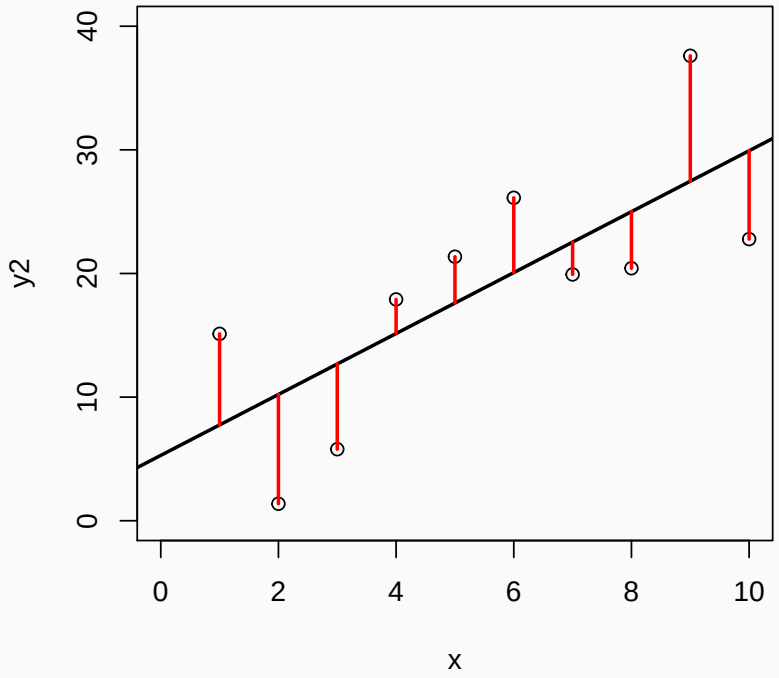
$\varepsilon_i \sim N(0, 0)$

$Y_i = 2 + 3X_i$



Se Y_i pode ser aproximado por X_i, então $\varepsilon_i \sim N(0, \sigma)$

$Y_i = 2 + 3X_i + \varepsilon \sim N(0, 6)$

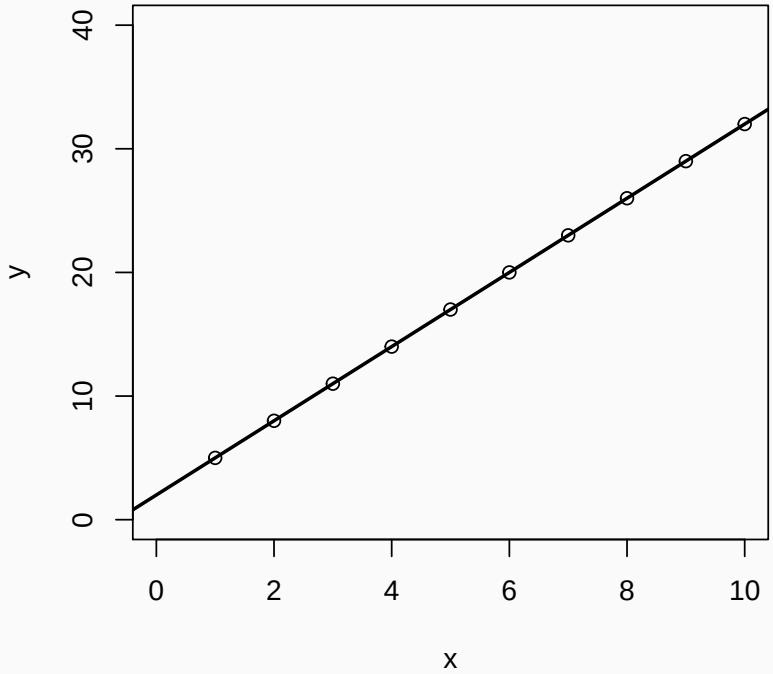


FORMULAÇÃO DO MODELO DE REGRESSÃO

Se Y_i puder ser predito exatamente por X_i, então

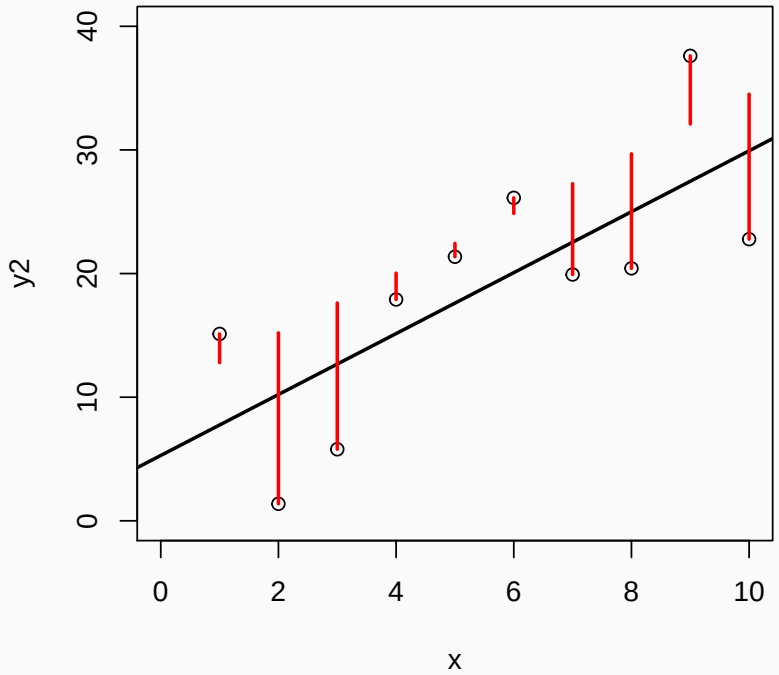
$\varepsilon_i \sim N(0, 0)$

$Y_i = 2 + 3X_i$



Se ε não possuir média zero, os erros não "regridem"para a linha de tendência central

$Y_i = 2 + 3X_i + \varepsilon \sim N(6, 6)$



Na prática, estamos modelando a **esperança** de Y para cada nível de X:

$$E[\,Y_i] = [\beta_0 + \beta_1 X_i + \varepsilon_i]$$

Na prática, estamos modelando a **esperança** de Y para cada nível de X:

$$E[Y_i] = [\beta_0 + \beta_1 X_i + \varepsilon_i]$$

$$E[Y_i] = \beta_0 + \beta_1 X_i + E[\varepsilon_i]$$

Na prática, estamos modelando a **esperança** de Y para cada nível de X:

$$E[Y_i] = [\beta_0 + \beta_1 X_i + \varepsilon_i]$$

$$E[Y_i] = \beta_0 + \beta_1 X_i + E[\varepsilon_i]$$

Mas nós sabemos que $E[\varepsilon_i] = 0$, então:

Na prática, estamos modelando a **esperança** de Y para cada nível de X:

$$E[Y_i] = [\beta_0 + \beta_1 X_i + \varepsilon_i]$$

$$E[Y_i] = \beta_0 + \beta_1 X_i + E[\varepsilon_i]$$

Mas nós sabemos que $E[\varepsilon_i] = 0$, então:

$$E[Y] = \beta_0 + \beta_1 X$$

Estas relações implicam nas seguintes propriedades:

1) Y_i é a soma de um termo constante ($E[Y]$) e um termo aleatório (ε), então Y_i é uma variável aleatória

Estas relações implicam nas seguintes propriedades:

1) Y_i é a soma de um termo constante ($E[Y]$) e um termo aleatório (ε), então Y_i é uma variável aleatória

2) A função de regressão para o modelo é $E[Y] = \beta_0 + \beta_1 X$

Estas relações implicam nas seguintes propriedades:

1) Y_i é a soma de um termo constante ($E[Y]$) e um termo aleatório (ε), então Y_i é uma variável aleatória

2) A função de regressão para o modelo é $E[Y] = \beta_0 + \beta_1 X$

3) Y_i desvia do valor determinado pela função de regressão por um erro ε_i

4) Pressupõe-se que os erros ε_i tem uma variância constante σ^2. Se isso é verdade, Y_i tem a mesma variância:

$$Var[\beta_0 + \beta_1 X_i + \varepsilon_i] = Var[\varepsilon_i] = \sigma^2$$

4) Pressupõe-se que os erros ε_i tem uma variância constante σ^2. Se isso é verdade, Y_i tem a mesma variância:

$$Var[\beta_0 + \beta_1 X_i + \varepsilon_i] = Var[\varepsilon_i] = \sigma^2$$

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

4) Pressupõe-se que os erros ε_i tem uma variância constante σ^2. Se isso é verdade, Y_i tem a mesma variância:

$$Var[\beta_0 + \beta_1 X_i + \varepsilon_i] = Var[\varepsilon_i] = \sigma^2$$

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

$$Var[Y_i] = Var[\beta_0 + \beta_1 X_i + \varepsilon_i]$$

4) Pressupõe-se que os erros ε_i tem uma variância constante σ^2. Se isso é verdade, Y_i tem a mesma variância:

$$Var[\beta_0 + \beta_1 X_i + \varepsilon_i] = Var[\varepsilon_i] = \sigma^2$$

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

$$Var[Y_i] = Var[\beta_0 + \beta_1 X_i + \varepsilon_i]$$

$$Var[Y_i] = \sigma^2$$

5) Pressupõe-se que os erros ε_i são independentes (não-correlacionados). Se quaisquer ε_i e ε_j são independentes, então Y_i e Y_j também são independentes:

Em resumo: Um modelo de regressão linear simples pressupõe que a resposta Y_i vem de uma distribuição de probabilidade com média $E[Y_i]$ e variância σ^2 constante para todos os níveis de X, e que quaisquer Y_i e Y_j são independentes.

OS PARÂMETROS DA REGRESSÃO

PARÂMETROS DO MODELO DE REGRESSÃO

Os parâmetros ou coeficientes da regressão $E[Y] = \beta_0 + \beta_1 X$ possuem nomes e significados específicos:

PARÂMETROS DO MODELO DE REGRESSÃO

Os parâmetros ou coeficientes da regressão $E[Y] = \beta_0 + \beta_1 X$ possuem nomes e significados específicos:

- O parâmetro β_0 é chamado de ?

PARÂMETROS DO MODELO DE REGRESSÃO

Os parâmetros ou coeficientes da regressão $E[Y] = \beta_0 + \beta_1 X$ possuem nomes e significados específicos:

- O parâmetro β_0 é chamado de **intercepto**

PARÂMETROS DO MODELO DE REGRESSÃO

Os parâmetros ou coeficientes da regressão $E[Y] = \beta_0 + \beta_1 X$ possuem nomes e significados específicos:

- O parâmetro β_0 é chamado de **intercepto**
  - β_0 representa ?

PARÂMETROS DO MODELO DE REGRESSÃO

Os parâmetros ou coeficientes da regressão $E[\,Y] = \beta_0 + \beta_1 X$ possuem nomes e significados específicos:

- O parâmetro β_0 é chamado de **intercepto**
  - β_0 representa $E[\,Y_i]$ quando $X = 0$

PARÂMETROS DO MODELO DE REGRESSÃO

Os parâmetros ou coeficientes da regressão $E[Y] = \beta_0 + \beta_1 X$ possuem nomes e significados específicos:

- O parâmetro β_0 é chamado de **intercepto**
  - β_0 representa $E[Y_i]$ quando $X = 0$

- O parâmetro β_1 é chamado de ?

PARÂMETROS DO MODELO DE REGRESSÃO

Os parâmetros ou coeficientes da regressão $E[Y] = \beta_0 + \beta_1 X$ possuem nomes e significados específicos:

- O parâmetro β_0 é chamado de **intercepto**
  - β_0 representa $E[Y_i]$ quando $X = 0$

- O parâmetro β_1 é chamado de **inclinação (*slope*)** da reta

PARÂMETROS DO MODELO DE REGRESSÃO

Os parâmetros ou coeficientes da regressão $E[Y] = \beta_0 + \beta_1 X$ possuem nomes e significados específicos:

- O parâmetro β_0 é chamado de **intercepto**
  - β_0 representa $E[Y_i]$ quando $X = 0$

- O parâmetro β_1 é chamado de **inclinação (*slope*)** da reta
  - β_1 representa ?

PARÂMETROS DO MODELO DE REGRESSÃO

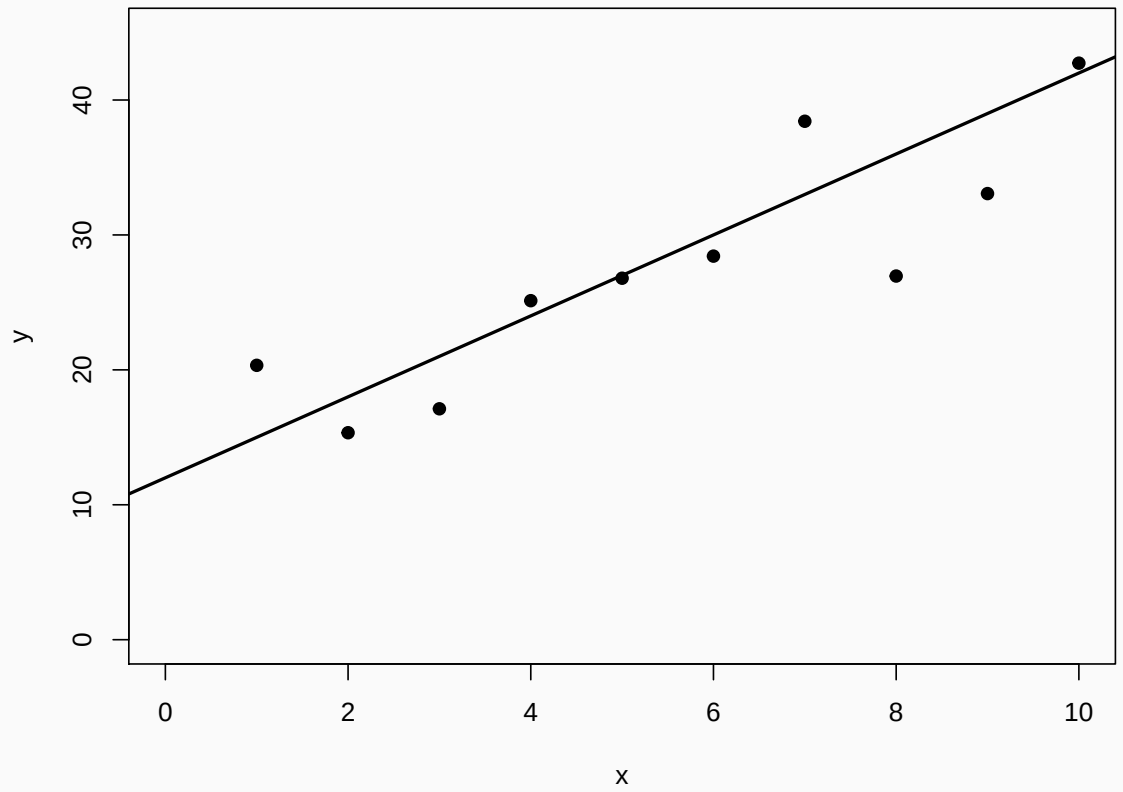
Os parâmetros ou coeficientes da regressão $E[Y] = \beta_0 + \beta_1 X$ possuem nomes e significados específicos:

- O parâmetro β_0 é chamado de **intercepto**
  - β_0 representa $E[Y_i]$ quando $X = 0$

- O parâmetro β_1 é chamado de **inclinação (*slope*)** da reta
  - β_1 representa o aumento em $E[Y_i]$ para um aumento unitário em X

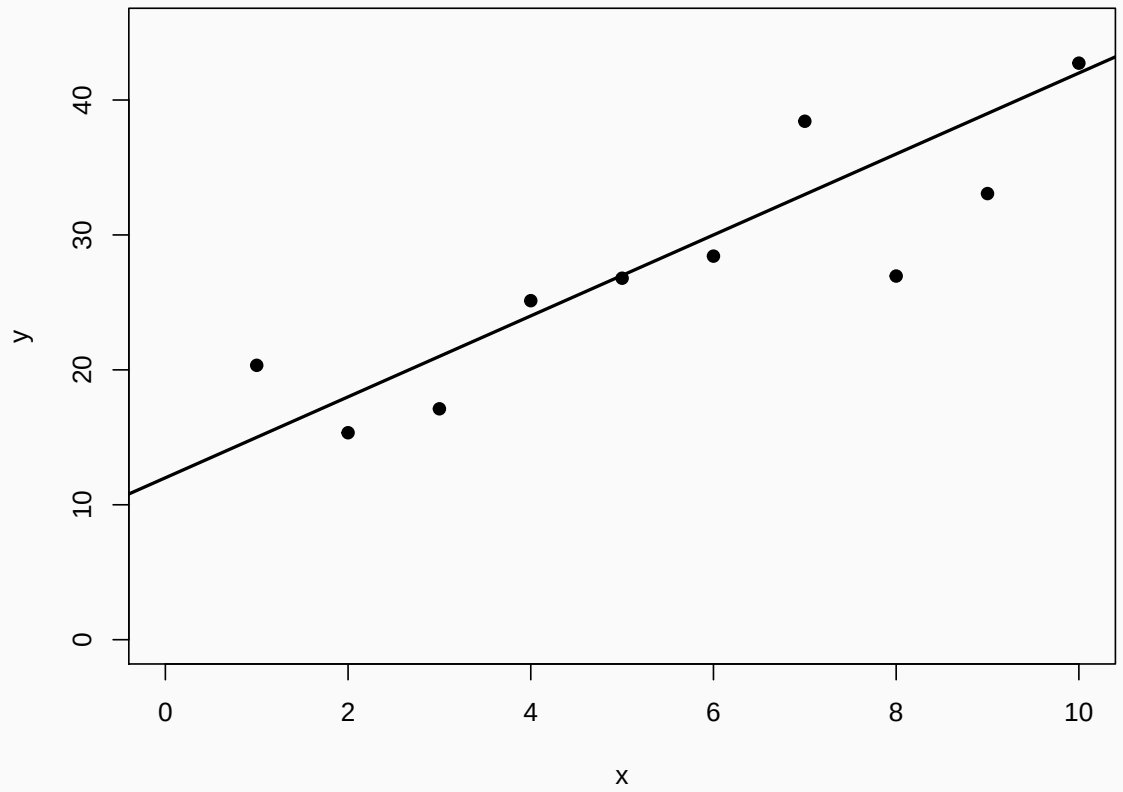
PARÂMETROS DO MODELO DE REGRESSÃO

$$E[Y] = 12 + 3X$$



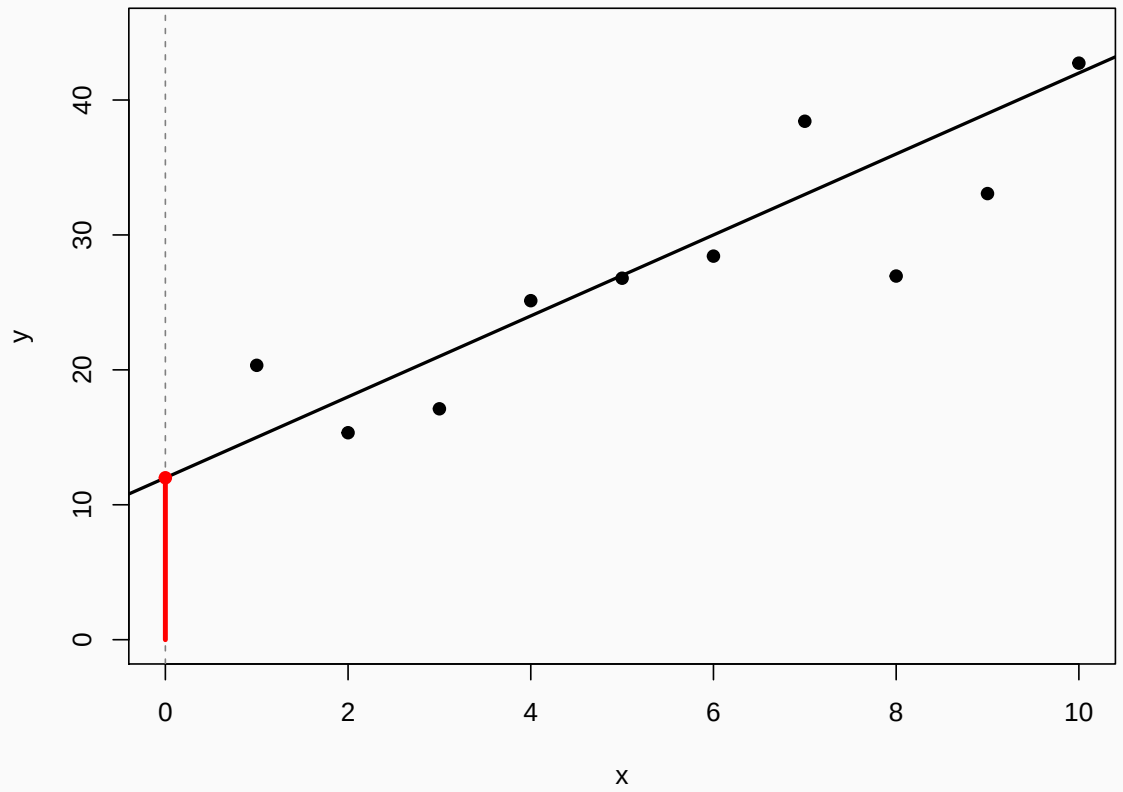
PARÂMETROS DO MODELO DE REGRESSÃO

$$E[Y] = 12 + 3X, \beta_0 = ?$$



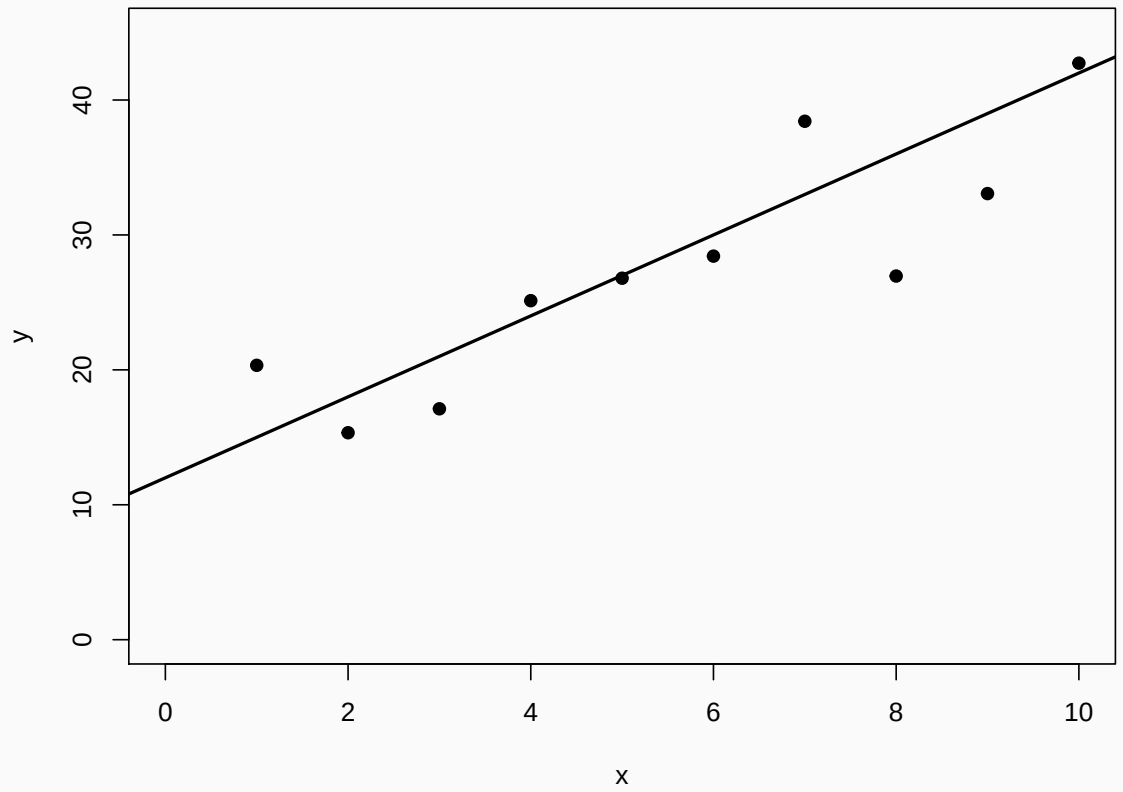
PARÂMETROS DO MODELO DE REGRESSÃO

$$E[Y] = 12 + 3X, \beta_0 = 12$$



PARÂMETROS DO MODELO DE REGRESSÃO

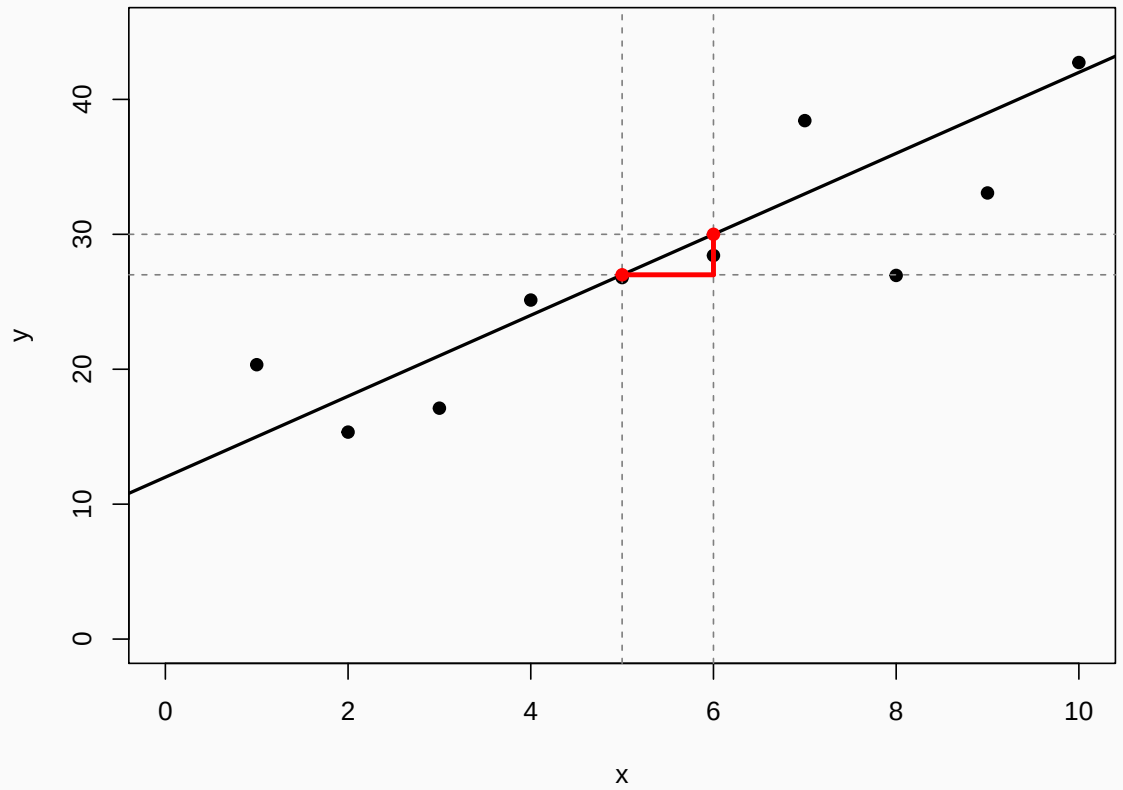
$$E[Y] = 12 + 3X, \beta_1 = ?$$



PARÂMETROS DO MODELO DE REGRESSÃO

$$E[Y] = 12 + 3X, \beta_1 = 3$$

$$X = 5,\ Y = 27;\ X = 6,\ Y = 30;\ 30 - 27 = 3$$



ESTIMANDO O MODELO

Assim como outras estatísticas, assume-se que o modelo $E[Y] = \beta_0 + \beta_1 X + \varepsilon$ corresponde a uma população.

Assim como outras estatísticas, assume-se que o modelo $E[Y] = \beta_0 + \beta_1 X + \varepsilon$ corresponde a uma população.

Ao tomarmos uma amostra de valores de X e Y, queremos estimar o modelo $\hat{Y} = b_0 + b_1 X + e$

Assim como outras estatísticas, assume-se que o modelo $E[Y] = \beta_0 + \beta_1 X + \varepsilon$ corresponde a uma população.

Ao tomarmos uma amostra de valores de X e Y, queremos estimar o modelo $\hat{Y} = b_0 + b_1 X + e$

Idealmente, gostaríamos de usar um método onde $\hat{Y}$, b_0, b_1 e e sejam bons estimadores (não-tendenciosos) de Y, β_0, β_1 e ε.

Para isso, podemos usar o método dos **Mínimos Quadrados Comuns (*Ordinary Least Squares, OLS*)**. Este método considera as diferenças entre cada valor Y_i e seu valor esperado $E[Y_i]$:

$$Y_i - E[Y_i] = Y_i - (\beta_0 + \beta_1 X_i)$$

Para isso, podemos usar o método dos **Mínimos Quadrados Comuns (*Ordinary Least Squares, OLS*)**. Este método considera as diferenças entre cada valor Y_i e seu valor esperado $E[Y_i]$:

$$Y_i - E[Y_i] = Y_i - (\beta_0 + \beta_1 X_i)$$

Como no caso das variâncias, estamos interessados no quadrado destas diferenças, para que elas não se cancelem:

Para isso, podemos usar o método dos **Mínimos Quadrados Comuns (*Ordinary Least Squares, OLS*)**. Este método considera as diferenças entre cada valor Y_i e seu valor esperado $E[Y_i]$:

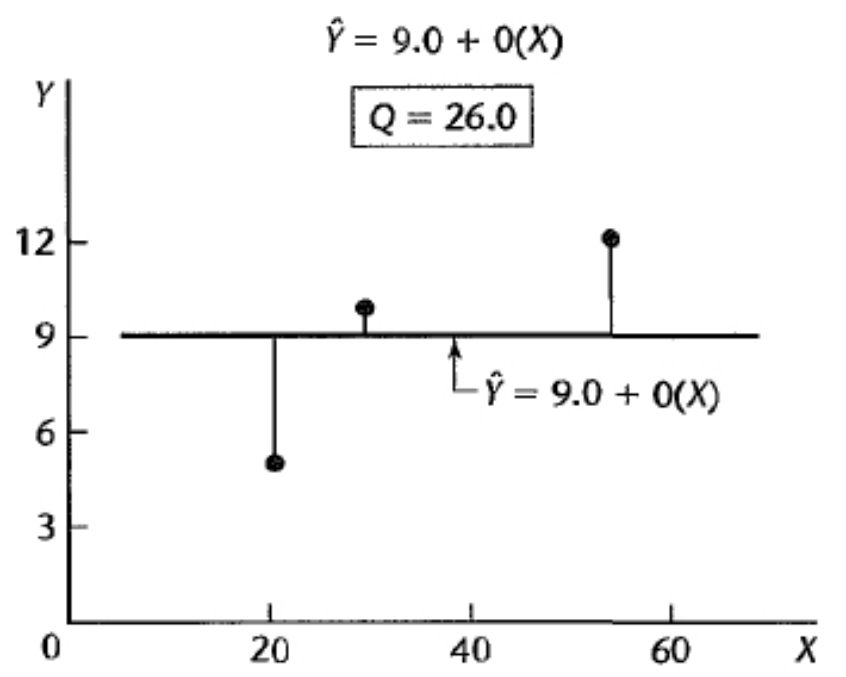
$$Y_i - E[Y_i] = Y_i - (\beta_0 + \beta_1 X_i)$$

Como no caso das variâncias, estamos interessados no quadrado destas diferenças, para que elas não se cancelem:

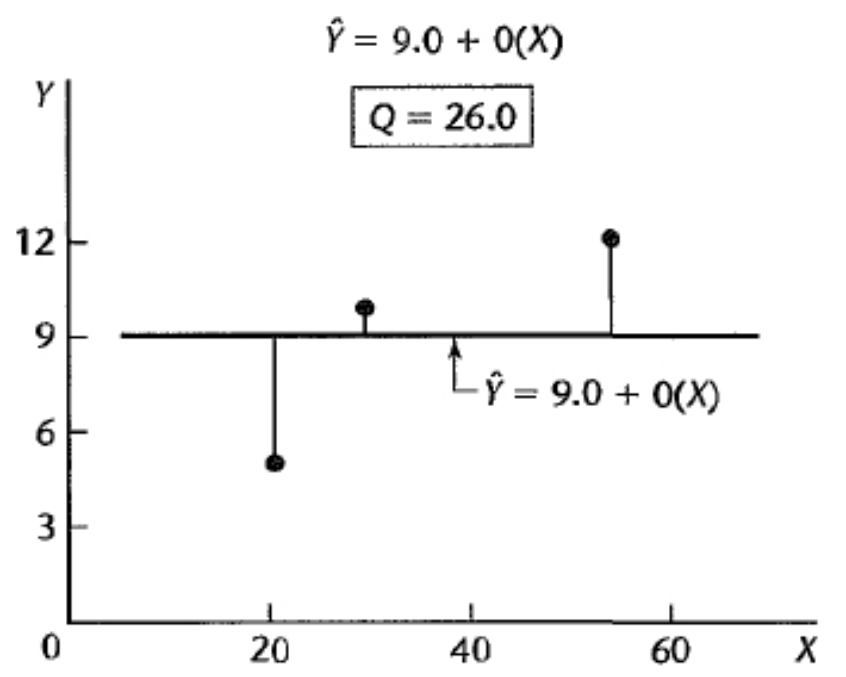
$$Q = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

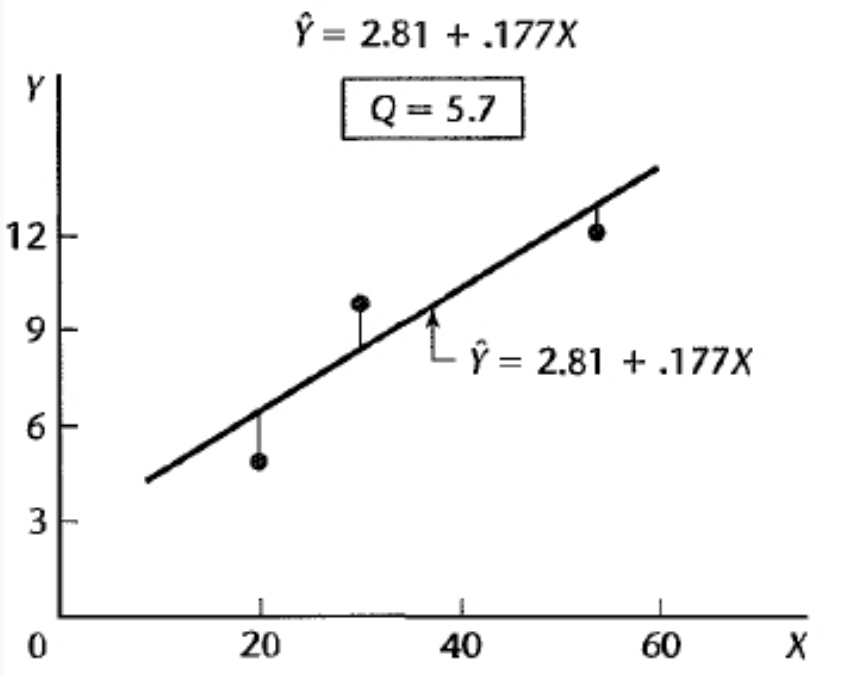
De acordo com a formulação do método OLS, os melhores estimadores de β_0 e β_1 são os valores b_0 e b_1 que minimizam o critério Q para as amostras obtidas.

De acordo com a formulação do método OLS, os melhores estimadores de β_0 e β_1 são os valores b_0 e b_1 que minimizam o critério Q para as amostras obtidas.



De acordo com a formulação do método OLS, os melhores estimadores de β_0 e β_1 são os valores b_0 e b_1 que minimizam o critério Q para as amostras obtidas.





Os estimadores b_0 e b_1 que satisfazem o critério de mínimos quadrados podem ser determinados de duas maneiras:

Os estimadores b_0 e b_1 que satisfazem o critério de mínimos quadrados podem ser determinados de duas maneiras:

- Numericamente, através de procedimentos de busca computacional.
- Analiticamente. Este método só funciona para Modelos Lineares Gerais.

Os estimadores b_0 e b_1 que satisfazem o critério de mínimos quadrados podem ser determinados de duas maneiras:

- Numericamente, através de procedimentos de busca computacional.
- Analiticamente. Este método só funciona para Modelos Lineares Gerais.

Para a regressão simples, o método analítico nos dá:

$$b_0 = \bar{Y} - b_1 \bar{X} \qquad b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

Ver Kutner et al. (2005) Applied Linear Statistical Models. 5th ed. McGraw Hill. para a dedução analítica. I know you want to.

EXEMPLO

X	Y	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$
9.09	12.10	14.97	24.14
4.82	7.28	-1.14	0.42
2.53	7.25	2.98	2.71
7.28	12.76	11.51	9.66
1.66	8.92	0.35	6.35
2.64	8.30	1.16	2.35
6.11	11.22	4.17	3.74
3.88	12.96	-1.16	0.09
4.79	9.71	0.41	0.38
3.66	10.34	-0.66	0.26
3.47	9.44	-0.27	0.49
0.70	1.97	24.65	12.12
9.12	15.51	31.94	24.49
2.76	5.56	4.93	1.99
0.12	2.52	26.52	16.46
$\bar{X}$	$\bar{Y}$	$\sum(X-\bar{X})(Y-\bar{Y})$	$\sum(X-\bar{X})^2$
4.18	9.06	120.35	105.66

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

EXEMPLO

X	Y	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$
9.09	12.10	14.97	24.14
4.82	7.28	-1.14	0.42
2.53	7.25	2.98	2.71
7.28	12.76	11.51	9.66
1.66	8.92	0.35	6.35
2.64	8.30	1.16	2.35
6.11	11.22	4.17	3.74
3.88	12.96	-1.16	0.09
4.79	9.71	0.41	0.38
3.66	10.34	-0.66	0.26
3.47	9.44	-0.27	0.49
0.70	1.97	24.65	12.12
9.12	15.51	31.94	24.49
2.76	5.56	4.93	1.99
0.12	2.52	26.52	16.46
$\bar{X}$	$\bar{Y}$	$\sum(X-\bar{X})(Y-\bar{Y})$	$\sum(X-\bar{X})^2$
4.18	9.06	120.35	105.66

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$b_1 = \frac{120.35}{105.66}$$

EXEMPLO

X	Y	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$
9.09	12.10	14.97	24.14
4.82	7.28	-1.14	0.42
2.53	7.25	2.98	2.71
7.28	12.76	11.51	9.66
1.66	8.92	0.35	6.35
2.64	8.30	1.16	2.35
6.11	11.22	4.17	3.74
3.88	12.96	-1.16	0.09
4.79	9.71	0.41	0.38
3.66	10.34	-0.66	0.26
3.47	9.44	-0.27	0.49
0.70	1.97	24.65	12.12
9.12	15.51	31.94	24.49
2.76	5.56	4.93	1.99
0.12	2.52	26.52	16.46
$\bar{X}$	$\bar{Y}$	$\sum(X-\bar{X})(Y-\bar{Y})$	$\sum(X-\bar{X})^2$
4.18	9.06	120.35	105.66

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$b_1 = \frac{120.35}{105.66}$$

$$b_1 = 1.14$$

EXEMPLO

X	Y	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$
9.09	12.10	14.97	24.14
4.82	7.28	-1.14	0.42
2.53	7.25	2.98	2.71
7.28	12.76	11.51	9.66
1.66	8.92	0.35	6.35
2.64	8.30	1.16	2.35
6.11	11.22	4.17	3.74
3.88	12.96	-1.16	0.09
4.79	9.71	0.41	0.38
3.66	10.34	-0.66	0.26
3.47	9.44	-0.27	0.49
0.70	1.97	24.65	12.12
9.12	15.51	31.94	24.49
2.76	5.56	4.93	1.99
0.12	2.52	26.52	16.46
$\bar{X}$	$\bar{Y}$	$\sum(X-\bar{X})(Y-\bar{Y})$	$\sum(X-\bar{X})^2$
4.18	9.06	120.35	105.66

$$b_0 = \bar{Y} - b_1\bar{X}$$

EXEMPLO

X	Y	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$
9.09	12.10	14.97	24.14
4.82	7.28	-1.14	0.42
2.53	7.25	2.98	2.71
7.28	12.76	11.51	9.66
1.66	8.92	0.35	6.35
2.64	8.30	1.16	2.35
6.11	11.22	4.17	3.74
3.88	12.96	-1.16	0.09
4.79	9.71	0.41	0.38
3.66	10.34	-0.66	0.26
3.47	9.44	-0.27	0.49
0.70	1.97	24.65	12.12
9.12	15.51	31.94	24.49
2.76	5.56	4.93	1.99
0.12	2.52	26.52	16.46
$\bar{X}$	$\bar{Y}$	$\sum(X-\bar{X})(Y-\bar{Y})$	$\sum(X-\bar{X})^2$
4.18	9.06	120.35	105.66

$$b_0 = \bar{Y} - b_1\bar{X}$$

$$b_0 = 9.06 - 1.14 \times 4.18$$

EXEMPLO

X	Y	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$
9.09	12.10	14.97	24.14
4.82	7.28	-1.14	0.42
2.53	7.25	2.98	2.71
7.28	12.76	11.51	9.66
1.66	8.92	0.35	6.35
2.64	8.30	1.16	2.35
6.11	11.22	4.17	3.74
3.88	12.96	-1.16	0.09
4.79	9.71	0.41	0.38
3.66	10.34	-0.66	0.26
3.47	9.44	-0.27	0.49
0.70	1.97	24.65	12.12
9.12	15.51	31.94	24.49
2.76	5.56	4.93	1.99
0.12	2.52	26.52	16.46
$\bar{X}$	$\bar{Y}$	$\sum(X-\bar{X})(Y-\bar{Y})$	$\sum(X-\bar{X})^2$
4.18	9.06	120.35	105.66

$$b_0 = \bar{Y} - b_1\bar{X}$$

$$b_0 = 9.06 - 1.14 \times 4.18$$

$$b_0 = 9.06 - 4.77$$

EXEMPLO

X	Y	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$
9.09	12.10	14.97	24.14
4.82	7.28	-1.14	0.42
2.53	7.25	2.98	2.71
7.28	12.76	11.51	9.66
1.66	8.92	0.35	6.35
2.64	8.30	1.16	2.35
6.11	11.22	4.17	3.74
3.88	12.96	-1.16	0.09
4.79	9.71	0.41	0.38
3.66	10.34	-0.66	0.26
3.47	9.44	-0.27	0.49
0.70	1.97	24.65	12.12
9.12	15.51	31.94	24.49
2.76	5.56	4.93	1.99
0.12	2.52	26.52	16.46
$\bar{X}$	$\bar{Y}$	$\sum(X-\bar{X})(Y-\bar{Y})$	$\sum(X-\bar{X})^2$
4.18	9.06	120.35	105.66

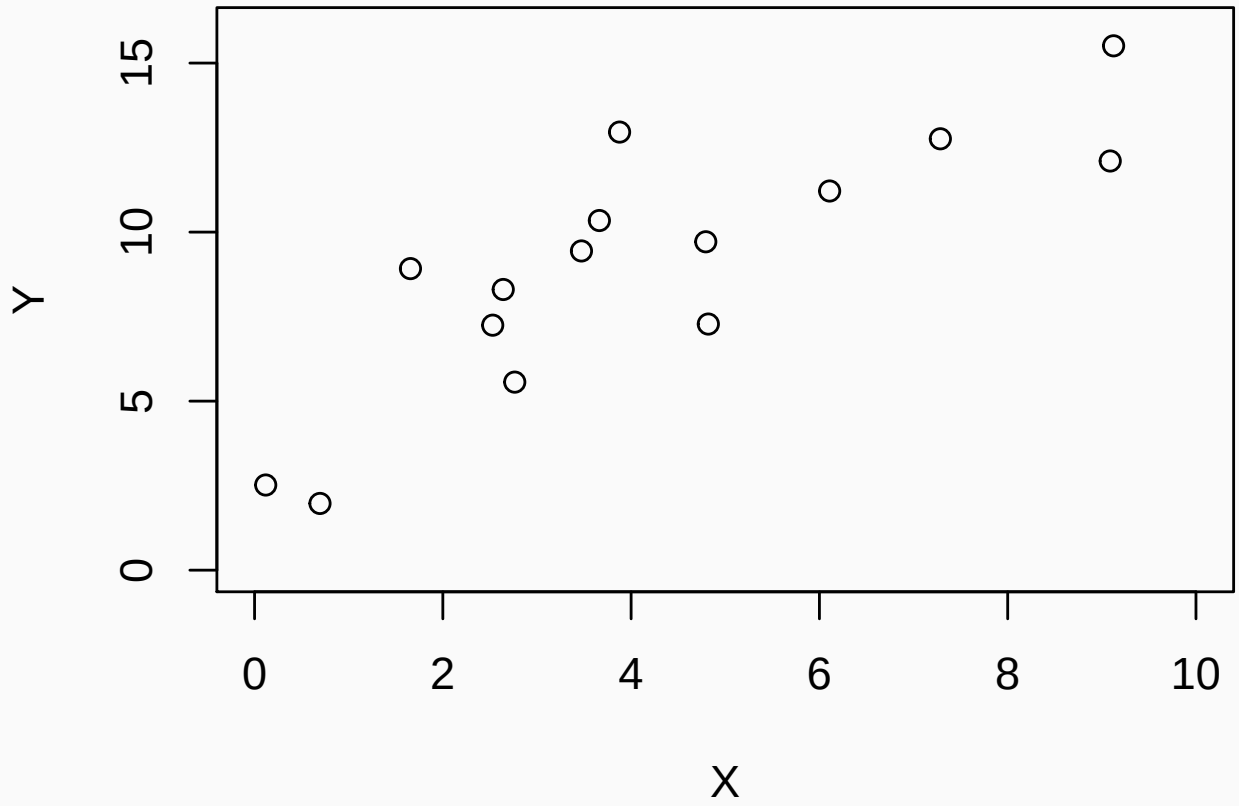
$$b_0 = \bar{Y} - b_1\bar{X}$$

$$b_0 = 9.06 - 1.14 \times 4.18$$

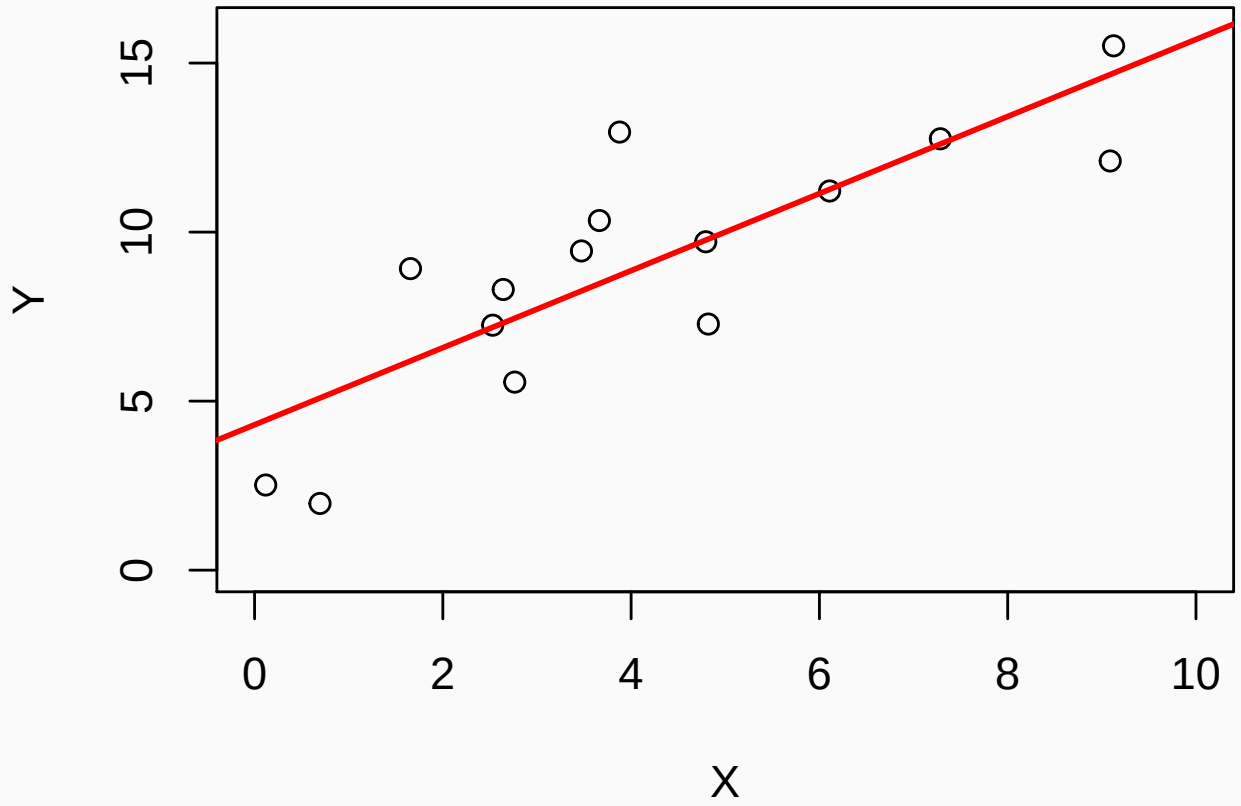
$$b_0 = 9.06 - 4.77$$

$$b_0 = 4.3$$

$$\hat{Y} = 4.3 + 1.14X$$



$$\hat{Y} = 4.3 + 1.14X$$



Olhando a nossa equação estimada, será que está faltando alguma coisa?

$$\hat{Y}_i = b_0 + b_1 X_i$$

Olhando a nossa equação estimada, será que está faltando alguma coisa?

$$\hat{Y}_i = b_0 + b_1 X_i$$

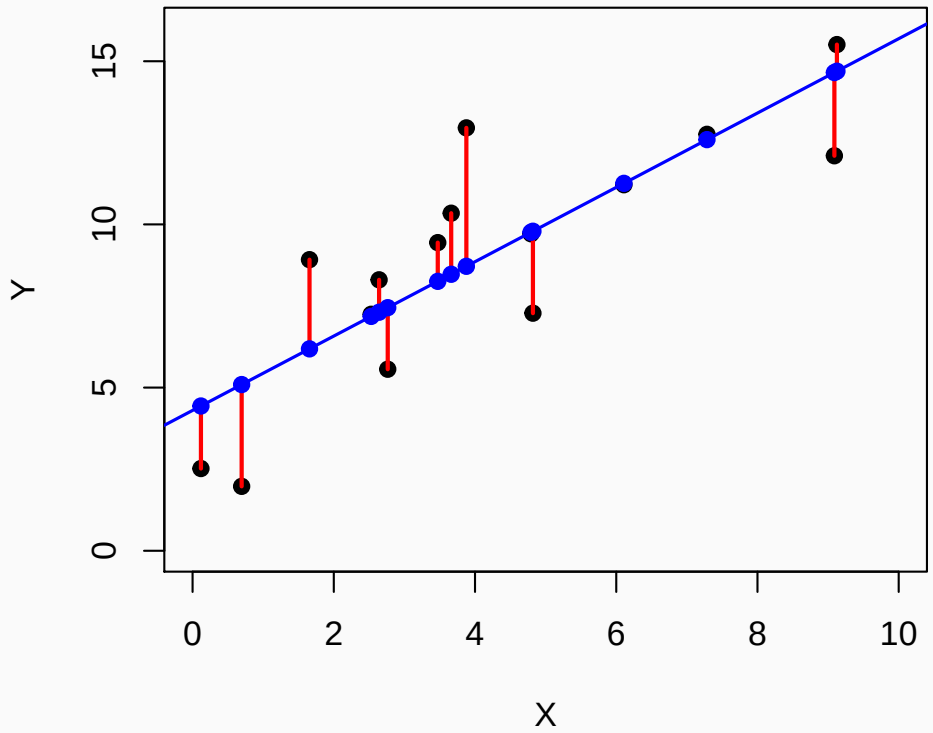
Onde está o termo de erro estimado, e?

$$\hat{Y}_i = b_0 + b_1 X_i + e$$

OS RESÍDUOS

Os erros estimados e_i são chamados de **resíduos** da regressão:

$$e_i = Y_i - \hat{Y}_i$$



1. A soma dos resíduos é zero : $\sum e_i = 0$

PROPRIEDADES DO MODELO DE REGRESSÃO

1. A soma dos resíduos é zero : $\sum e_i = 0$
2. A soma dos quadrados dos resíduos, $\sum e_i^2$ é um mínimo

PROPRIEDADES DO MODELO DE REGRESSÃO

1. A soma dos resíduos é zero : $\sum e_i = 0$
2. A soma dos quadrados dos resíduos, $\sum e_i^2$ é um mínimo
3. A soma dos valores observados Y_i é igual a soma dos valores ajustados $\hat{Y}_i$

PROPRIEDADES DO MODELO DE REGRESSÃO

1. A soma dos resíduos é zero : $\sum e_i = 0$
2. A soma dos quadrados dos resíduos, $\sum e_i^2$ é um mínimo
3. A soma dos valores observados Y_i é igual a soma dos valores ajustados $\hat{Y}_i$
4. A soma dos resíduos ponderada pelos valores de X_i é zero: $\sum X_i e_i = 0$

PROPRIEDADES DO MODELO DE REGRESSÃO

1. A soma dos resíduos é zero : $\sum e_i = 0$
2. A soma dos quadrados dos resíduos, $\sum e_i^2$ é um mínimo
3. A soma dos valores observados Y_i é igual a soma dos valores ajustados $\hat{Y}_i$
4. A soma dos resíduos ponderada pelos valores de X_i é zero: $\sum X_i e_i = 0$
5. Devido a 1) e 4), a soma dos resíduos ponderada pelos valores de $\hat{Y}_i$ também é zero: $\sum \hat{Y}_i e_i = 0$

1. A soma dos resíduos é zero : $\sum e_i = 0$
2. A soma dos quadrados dos resíduos, $\sum e_i^2$ é um mínimo
3. A soma dos valores observados Y_i é igual a soma dos valores ajustados $\hat{Y}_i$
4. A soma dos resíduos ponderada pelos valores de X_i é zero: $\sum X_i e_i = 0$
5. Devido a 1) e 4), a soma dos resíduos ponderada pelos valores de $\hat{Y}_i$ também é zero: $\sum \hat{Y}_i e_i = 0$
6. **A reta de regressão sempre passa pelo ponto $(\bar{X}, \bar{Y})$**

ESTIMANDO A VARIÂNCIA DO MODELO

ESTIMANDO A VARIÂNCIA DE UMA POPULAÇÃO

A variância de ε também precisa ser estimada, a fim de caracterizar a distribuição de probabilidade de Y para cada nível de X, e permitir inferências sobre o modelo.

ESTIMANDO A VARIÂNCIA DE UMA POPULAÇÃO

A variância de ε também precisa ser estimada, a fim de caracterizar a distribuição de probabilidade de Y para cada nível de X, e permitir inferências sobre o modelo.

Relembrando: A variância de uma população $Y(\sigma^2)$ pode ser estimada pela variância de uma amostra (s^2), através da **soma dos quadrados** dos desvios de Yi a partir de $\bar{Y}$:

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

ESTIMANDO A VARIÂNCIA DE UMA POPULAÇÃO

A variância de ε também precisa ser estimada, a fim de caracterizar a distribuição de probabilidade de Y para cada nível de X, e permitir inferências sobre o modelo.

Relembrando: A variância de uma população $Y(\sigma^2)$ pode ser estimada pela variância de uma amostra (s^2), através da **soma dos quadrados** dos desvios de Yi a partir de $\bar{Y}$:

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

Para obter s^2, nós dividimos a soma dos quadrados pelos graus de liberdade associados com essa soma:

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

ESTIMANDO A VARIÂNCIA DO MODELO

Uma das pressuposições do modelo de regressão linear é que s^2 é constante. A única diferença para a variância comum é que, no modelo, a distribuição de Y_i varia de acordo com o nível de X, então os desvios são calculados em relação a $\hat{Y}$, e não $\bar{Y}$:

[1] Essa nomenclatura varia, vejam exatamente quem é quem ao ler um livro/artigo

ESTIMANDO A VARIÂNCIA DO MODELO

Uma das pressuposições do modelo de regressão linear é que s^2 é constante. A única diferença para a variância comum é que, no modelo, a distribuição de Y_i varia de acordo com o nível de X, então os desvios são calculados em relação a $\hat{Y}$, e não $\bar{Y}$:

$$Y_i - \hat{Y} = e_i$$

[1] Essa nomenclatura varia, vejam exatamente quem é quem ao ler um livro/artigo

ESTIMANDO A VARIÂNCIA DO MODELO

Uma das pressuposições do modelo de regressão linear é que s^2 é constante. A única diferença para a variância comum é que, no modelo, a distribuição de Y_i varia de acordo com o nível de X, então os desvios são calculados em relação a $\hat{Y}$, e não $\bar{Y}$:

$$Y_i - \hat{Y} = e_i$$

E a soma do quadrado destes valores é chamada de **Soma dos Quadrados dos Erros/Resíduos** (SQ_{res}[1]):

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

[1] Essa nomenclatura varia, vejam exatamente quem é quem ao ler um livro/artigo

A SQ_{res} tem $n-2$ graus de liberdade, por que dois graus são perdidos estimando-se β_0 e β_1. Assim, temos que s^2 é:

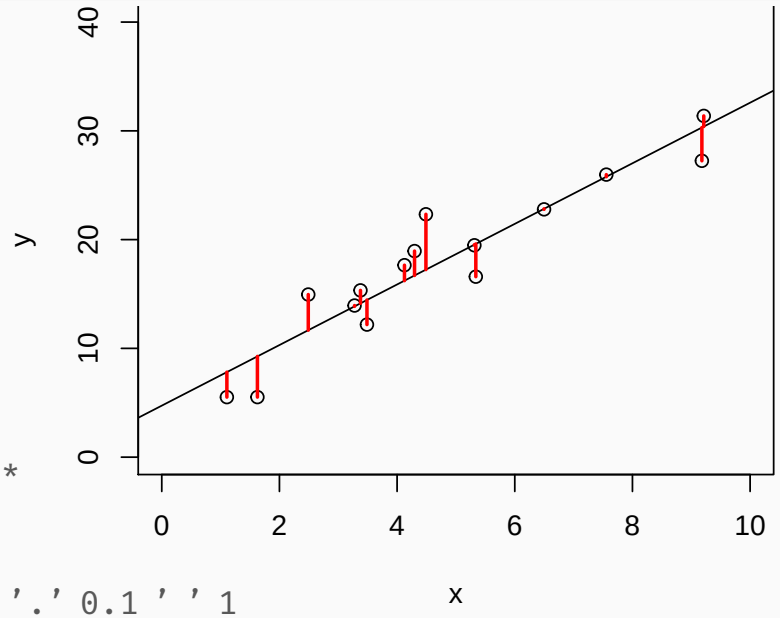
$$s^2 = MQ_{res} = \frac{SQ_{res}}{n-2} = \frac{\sum(Y_i - \hat{Y})^2}{n-2} = \frac{\sum e_i^2}{n-2}$$

A divisão por $n-2$ é apenas uma normalização para proporção. Mas não é necessária para entender a quantidade de variação.

```r
set.seed(1979)
x <- runif(15,1,10)
y <- 2 + 3*x + rnorm(15,0,3)
m <- lm(y ~ x)
summary(m)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.742 -2.281  0.078  1.308  5.090
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.7349    1.5059   3.144  0.00776 **
## x            2.7854    0.2828   9.848 2.15e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.617 on 13 degrees of freedom
## Multiple R-squared:  0.8818,^^IAdjusted R-squared:  0.8727
## F-statistic: 96.98 on 1 and 13 DF,  p-value: 2.149e-07
```

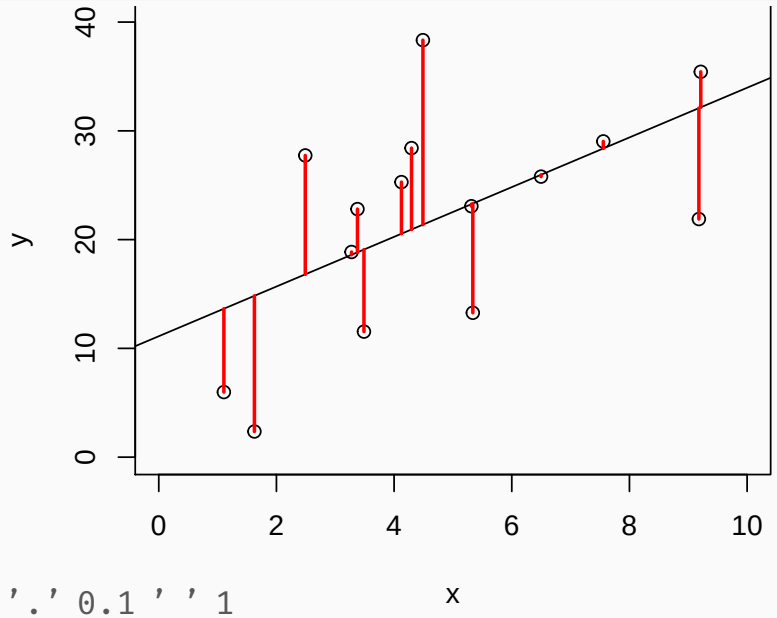
```r
ypred <- predict(m)
plot(x,y,xlim=c(0,10),ylim=c(0,40))
abline(m)
segments(x0=x, x1=x,y0=ypred,y1=y,
         col='red',lwd=2)
```



```r
set.seed(1979)
x <- runif(15,1,10)
y <- 2 + 3*x + rnorm(15,0,10)
m <- lm(y ~ x)
summary(m)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.4732  -7.6035   0.2601   4.3598  16.9658
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.1165     5.0196   2.215   0.0453 *
## x             2.2847     0.9428   2.423   0.0307 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.722 on 13 degrees of freedom
## Multiple R-squared:  0.3112,^^IAdjusted R-squared:  0.2582
## F-statistic: 5.872 on 1 and 13 DF,  p-value: 0.03072
```

```r
ypred <- predict(m)
plot(x,y,xlim=c(0,10),ylim=c(0,40))
abline(m)
segments(x0=x, x1=x,y0=ypred,y1=y,
         col='red',lwd=2)
```



Agora, já sabemos como estimar todos os componentes do modelo: $\hat{Y}$, b_0, b_1, e s^2. O que mais precisamos?

PARTIÇÃO DA VARIÂNCIA E O COEFICIENTE DE DETERMINAÇÃO

Agora, já sabemos como estimar todos os componentes do modelo: $\hat{Y}$, b_0, b_1, e s^2. O que mais precisamos?

- De um método para avaliar a qualidade da estimação

PARTIÇÃO DA VARIÂNCIA E O COEFICIENTE DE DETERMINAÇÃO

Agora, já sabemos como estimar todos os componentes do modelo: $\hat{Y}$, b_0, b_1, e s^2. O que mais precisamos?

- De um método para avaliar a qualidade da estimação
- De um método para avaliar o ajuste do modelo

PARTIÇÃO DA VARIÂNCIA E O COEFICIENTE DE DETERMINAÇÃO

- Como vimos antes, SQ_{res} nos dá a variancia dos resíduos.

PARTIÇÃO DA VARIÂNCIA E O COEFICIENTE DE DETERMINAÇÃO

- Como vimos antes, SQ_{res} nos dá a variancia dos resíduos.
- A variância contida em SQ_{res} é a quantidade de variação que é aleatória, e não pôde ser capturada pelo modelo

PARTIÇÃO DA VARIÂNCIA E O COEFICIENTE DE DETERMINAÇÃO

- Como vimos antes, SQ_{res} nos dá a variancia dos resíduos.
- A variância contida em SQ_{res} é a quantidade de variação que é aleatória, e não pôde ser capturada pelo modelo
- Essa variação é uma parte da variância total de Y (Soma dos Quadrados Totais, SQ_{tot})

PARTIÇÃO DA VARIÂNCIA E O COEFICIENTE DE DETERMINAÇÃO

- Como vimos antes, SQ_{res} nos dá a variancia dos resíduos.
- A variância contida em SQ_{res} é a quantidade de variação que é aleatória, e não pôde ser capturada pelo modelo
- Essa variação é uma parte da variância total de Y (Soma dos Quadrados Totais, SQ_{tot})
- Podemos então definir a "variância explicada pelo modelo", como sendo a diferença entre a variância total de Y (SQ_{tot}), e a variância dos resíduos (SQ_{res}):

$$SQ_{reg} = SQ_{tot} - SQ_{res}$$

PARTIÇÃO DA VARIÂNCIA E O COEFICIENTE DE DETERMINAÇÃO

Essa partição pode ser melhor entendida ao se considerarem duas situações extremas:

Essa partição pode ser melhor entendida ao se considerarem duas situações extremas:

- Se todos os valores de Y caíssem exatamente em cima da reta, $SQ_{res} = 0$, e $SQ_{reg} = SQ_{tot}$

PARTIÇÃO DA VARIÂNCIA E O COEFICIENTE DE DETERMINAÇÃO

Essa partição pode ser melhor entendida ao se considerarem duas situações extremas:

- Se todos os valores de Y caíssem exatamente em cima da reta, $SQ_{res} = 0$, e $SQ_{reg} = SQ_{tot}$

- Se não há relação entre X e Y, $\beta_1 = 0$, e $Y = \beta_0 + \varepsilon$. Nesse caso, $Y \sim N(\beta_0, \sigma)$, e $SQ_{tot} = SQ_{res}$

PARTIÇÃO DA VARIÂNCIA E O COEFICIENTE DE DETERMINAÇÃO

Essa partição pode ser melhor entendida ao se considerarem duas situações extremas:

- Se todos os valores de Y caíssem exatamente em cima da reta, $SQ_{res} = 0$, e $SQ_{reg} = SQ_{tot}$

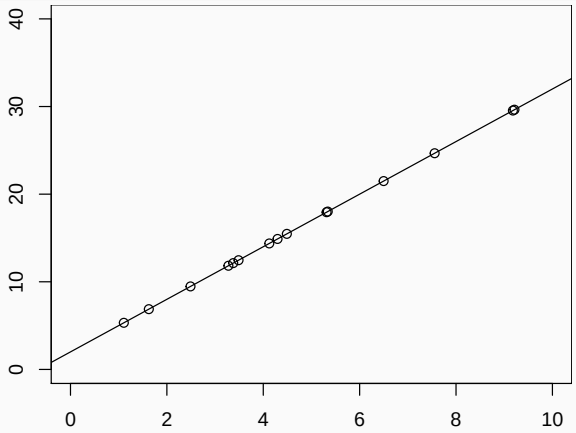
- Se não há relação entre X e Y, $\beta_1 = 0$, e $Y = \beta_0 + \varepsilon$. Nesse caso, $Y \sim N(\beta_0, \sigma)$, e $SQ_{tot} = SQ_{res}$

```r
set.seed(1979)
x <- runif(15,1,10)
y <- 2 + 3*x
m <- lm(y ~ x)
anova(m)["Sum Sq"]
```

```
##           Sum Sq
## x         770.26
## Residuals   0.00
```

```r
var(y)*(15-1)
```

```
## [1] 770.2641
```

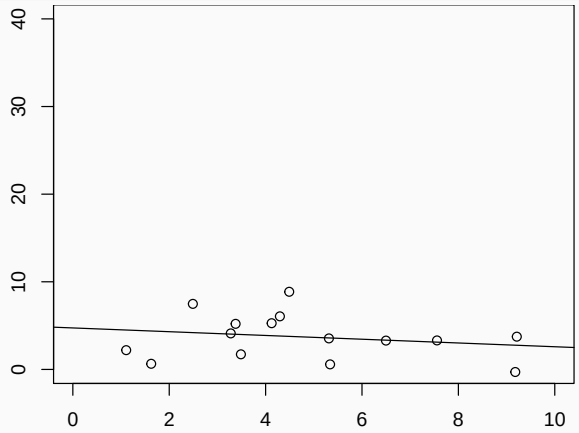



```r
set.seed(1979)
x <- runif(15,1,10)
y <- 2 + rnorm(15,0,3)
m <- lm(y ~ x)
anova(m)["Sum Sq"]
```

```
##           Sum Sq
## x          3.941
## Residuals 89.006
```

```r
var(y)*(15-1)
```

```
## [1] 92.94772
```



COEFICIENTE DE DETERMINAÇÃO

A partir desta formulação, chegamos ao **Coeficiente de Determinação (R^2)**:

$$R^2 = \frac{SQ_{reg}}{SQ_{tot}} = \frac{SQ_{reg}}{SQ_{reg} + SQ_{res}}$$

COEFICIENTE DE DETERMINAÇÃO

A partir desta formulação, chegamos ao **Coeficiente de Determinação (R^2)**:

$$R^2 = \frac{SQ_{reg}}{SQ_{tot}} = \frac{SQ_{reg}}{SQ_{reg} + SQ_{res}}$$

Como podemos interpretar o valor de R^2?

COEFICIENTE DE DETERMINAÇÃO

A partir desta formulação, chegamos ao **Coeficiente de Determinação (R^2)**:

$$R^2 = \frac{SQ_{reg}}{SQ_{tot}} = \frac{SQ_{reg}}{SQ_{reg} + SQ_{res}}$$

Como podemos interpretar o valor de R^2?
R^2 nos dá a proporção da variância total de Y explicada pelo modelo de regressão

* r^2 se refere a um modelo simples, e R^2 a um modelo multivariado.

EM RESUMO

- Um modelo de **regressão linear** estima uma relação estatística entre X e Y, através de coeficientes lineares β
- Esta relação é caracterizada por uma **co-variação** entre os níveis de X e $E[Y]$
- A existência de co-variação não implica em **causalidade**
- A variância de Y não capturada pelo modelo constitui o **erro da regressão** (ε)
- A **variância de** Y_i a cada nível de X é dada pela variância de ε
- A **variância total** de Y é dada pela variância de ε + a relação $\beta_0 + \beta_1 X$

EM RESUMO

- Um modelo de **regressão linear** estima uma relação estatística entre X e Y, através de coeficientes lineares β
- Esta relação é caracterizada por uma **co-variação** entre os níveis de X e $E[Y]$
- A existência de co-variação não implica em **causalidade**
- A variância de Y não capturada pelo modelo constitui o **erro da regressão** (ε)
- A **variância de** Y_i a cada nível de X é dada pela variância de ε
- A **variância total** de Y é dada pela variância de ε + a relação $\beta_0 + \beta_1 X$

EM RESUMO

- Um modelo de **regressão linear** estima uma relação estatística entre X e Y, através de coeficientes lineares β
- Esta relação é caracterizada por uma **co-variação** entre os níveis de X e $E[Y]$
- A existência de co-variação não implica em **causalidade**
- A variância de Y não capturada pelo modelo constitui o **erro da regressão** (ε)
- A **variância de** Y_i a cada nível de X é dada pela variância de ε
- A **variância total** de Y é dada pela variância de ε + a relação $\beta_0 + \beta_1 X$

EM RESUMO

- Um modelo de **regressão linear** estima uma relação estatística entre X e Y, através de coeficientes lineares β
- Esta relação é caracterizada por uma **co-variação** entre os níveis de X e $E[Y]$
- A existência de co-variação não implica em **causalidade**
- A variância de Y não capturada pelo modelo constitui o **erro da regressão** (ε)
- A **variância de** Y_i a cada nível de X é dada pela variância de ε
- A **variância total** de Y é dada pela variância de ε + a relação $\beta_0 + \beta_1 X$

EM RESUMO

- Um modelo de **regressão linear** estima uma relação estatística entre X e Y, através de coeficientes lineares β
- Esta relação é caracterizada por uma **co-variação** entre os níveis de X e $E[Y]$
- A existência de co-variação não implica em **causalidade**
- A variância de Y não capturada pelo modelo constitui o **erro da regressão** (ε)
- A **variância de** Y_i a cada nível de X é dada pela variância de ε
- A **variância total** de Y é dada pela variância de ε + a relação $\beta_0 + \beta_1 X$

EM RESUMO

- Um modelo de **regressão linear** estima uma relação estatística entre X e Y, através de coeficientes lineares β
- Esta relação é caracterizada por uma **co-variação** entre os níveis de X e $E[Y]$
- A existência de co-variação não implica em **causalidade**
- A variância de Y não capturada pelo modelo constitui o **erro da regressão** (ε)
- A **variância de** Y_i a cada nível de X é dada pela variância de ε
- A **variância total** de Y é dada pela variância de ε + a relação $\beta_0 + \beta_1 X$

EM RESUMO

- Os coeficientes β_0 e β_1 determinam o **intercepto** e a **inclinação da reta**
- A reta de regressão ($\hat{Y} = b_0 + b_1 X + e$) é estimada pelo método de **mínimos quadrados**, que busca minizar Soma dos Quadrados dos Erros (SQ_{res})
- A partir da diferença entre SQ_{res} e a Soma dos Quadrados Totais ($SQ_{tot} = Var[Y]$), podemos estimar a Soma dos Quadrados da Regressão (SQ_{reg})
- A relação $\frac{SQ_{reg}}{SQ_{tot}}$ é chamada de R^2, e nos diz a proporção da variância total de Y explicada pelo modelo

EM RESUMO

- Os coeficientes β_0 e β_1 determinam o **intercepto** e a **inclinação da reta**
- A reta de regressão ($\hat{Y} = b_0 + b_1 X + e$) é estimada pelo método de **mínimos quadrados**, que busca minizar Soma dos Quadrados dos Erros (SQ_{res})
- A partir da diferença entre SQ_{res} e a Soma dos Quadrados Totais ($SQ_{tot} = Var[Y]$), podemos estimar a Soma dos Quadrados da Regressão (SQ_{reg})
- A relação $\frac{SQ_{reg}}{SQ_{tot}}$ é chamada de R^2, e nos diz a proporção da variância total de Y explicada pelo modelo

EM RESUMO

- Os coeficientes β_0 e β_1 determinam o **intercepto** e a **inclinação da reta**
- A reta de regressão ($\hat{Y} = b_0 + b_1 X + e$) é estimada pelo método de **mínimos quadrados**, que busca minizar Soma dos Quadrados dos Erros (SQ_{res})
- A partir da diferença entre SQ_{res} e a Soma dos Quadrados Totais ($SQ_{tot} = Var[Y]$), podemos estimar a Soma dos Quadrados da Regressão (SQ_{reg})
- A relação $\frac{SQ_{reg}}{SQ_{tot}}$ é chamada de R^2, e nos diz a proporção da variância total de Y explicada pelo modelo

EM RESUMO

- Os coeficientes β_0 e β_1 determinam o **intercepto** e a **inclinação da reta**
- A reta de regressão ($\hat{Y} = b_0 + b_1 X + e$) é estimada pelo método de **mínimos quadrados**, que busca minizar Soma dos Quadrados dos Erros (SQ_{res})
- A partir da diferença entre SQ_{res} e a Soma dos Quadrados Totais ($SQ_{tot} = Var[Y]$), podemos estimar a Soma dos Quadrados da Regressão (SQ_{reg})
- A relação $\frac{SQ_{reg}}{SQ_{tot}}$ é chamada de R^2, e nos diz a proporção da variância total de Y explicada pelo modelo

Interpretando o output do modelo de regressão no R ...até agora.

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7244 -3.9253  0.2196  3.5752  7.5188
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.3632     2.0016   0.181    0.858
## x             3.3127     0.2058  16.099 3.93e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.81 on 18 degrees of freedom
## Multiple R-squared:  0.9351,^^IAdjusted R-squared:  0.9315
## F-statistic: 259.2 on 1 and 18 DF,  p-value: 3.925e-12
```