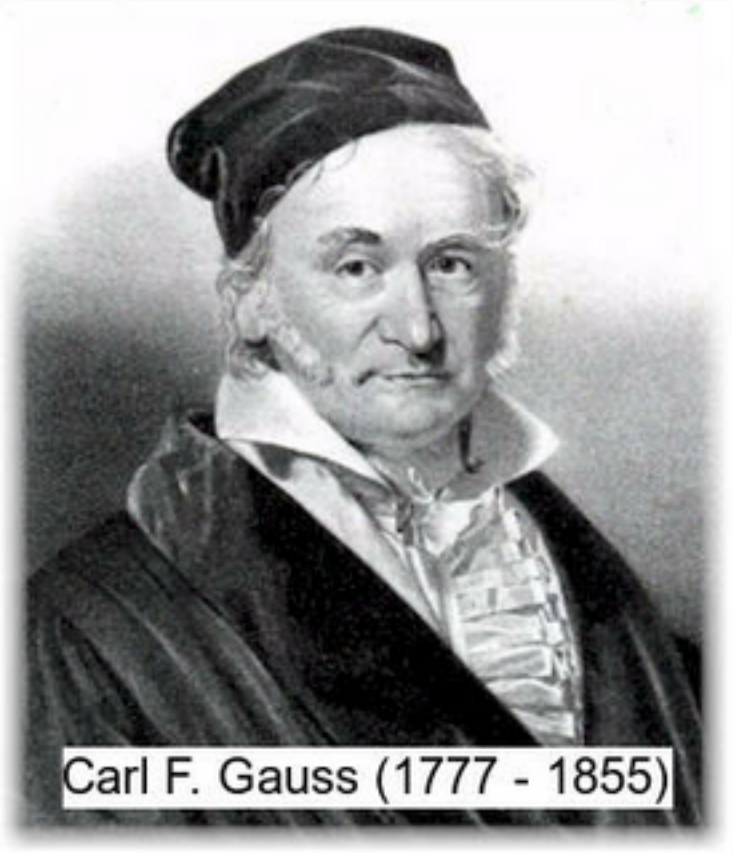
Inferência Estatística

Correlação e Regressão Linear Simples

Fabio Cop (fabiocopf@gmail.com)
Instituto do Mar - UNIFESP
Última atualização em 14 de março de 2022

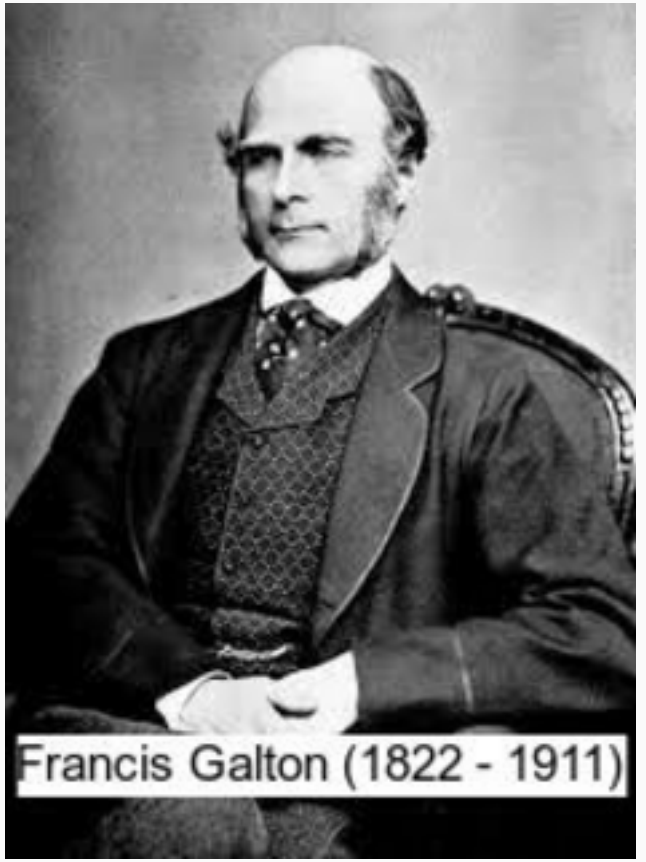
Método dos Mínimos Quadrados (MMQ)

A primeira solução para o problema da regressão (relacionar uma variável resposta Y a uma variável preditora X) foi o **Método dos Mínimos Quadrados (MMQ)**, publicado por por Gauss (1777 – 1855) em 1809, embora haja relatos históricos de que Gauss pensou e resolveu o problema quando tinha apenas 11 anos. Gauss aplicou o método para obter predições sobre as órbitas dos corpos ao redor do Sol a partir de observações astronômicas.



O termo Regressão

O termo **regressão** foi empregado por Francis Galton em 1866, um dos pais da Biometria e primo de *Charles Darwin*, no séc. XIX, para descrever o fenômeno biológico em que pais muito altos tenderiam a ter descendentes mais baixos que eles próprios e vice versa. A altura dos descendentes tenderia portanto a *regressar* à média da população.

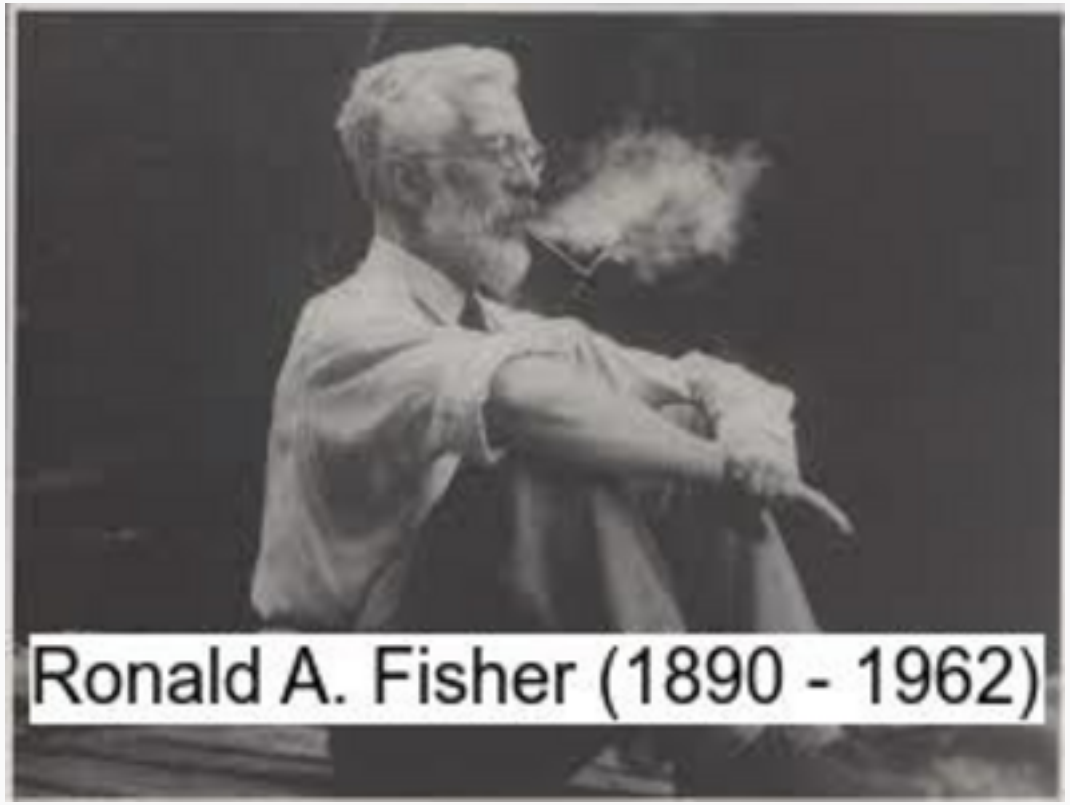


O coeficiente de correlação de Pearson

Galton propôs o **coeficiente de correlação** para medir a associação linear entre duas variáveis quantitativas. Suas idéias foram estendidas por Udny Yule e Karl Pearson para um contexto estatístico mais geral. No modelo de Udny e Pearson assume-se que a distribuição conjunta entre a variável resposta e a variável preditora $f(Y, X)$ é Gaussiana (Normal). Pearson foi um dos diversos autores que começaram a utilizar o termo **distribuição Normal** no início do século XX.

A formulação de Fisher

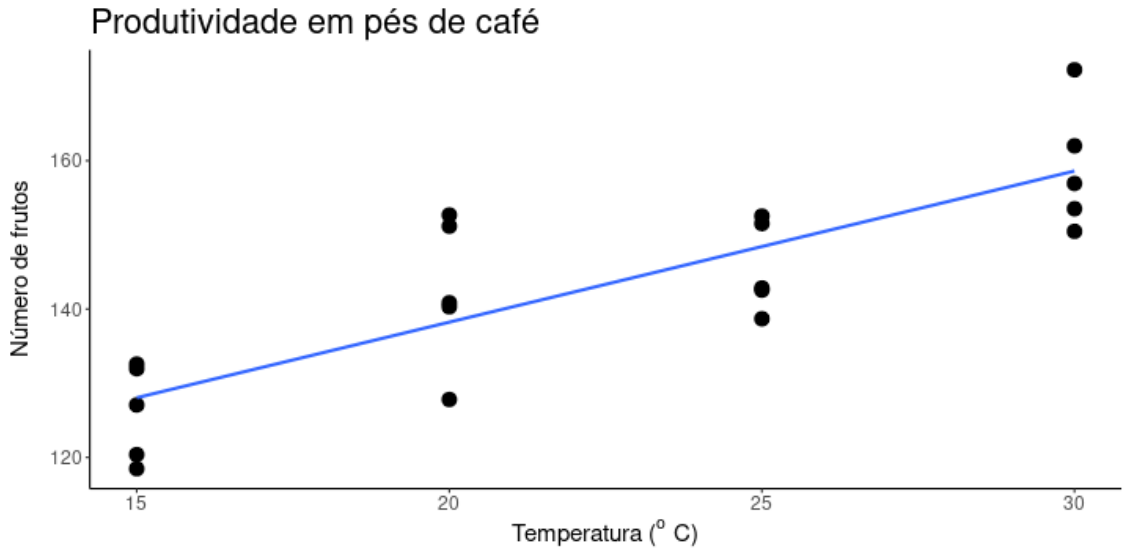
A suposição de Pearson confunde os conceitos de regressão e correlação. Esta suposição foi modificada por R. A. Fisher em 1922 e 1925. Fisher assumiu que a distribuição **condicional** da variável resposta $f(Y|X)$ seja Gaussiana - a conjunta não precisa ser. Esta solução é mais próxima daquela formulada por Gauss. Fisher desenvolveu também o método da **Máxima Verossimilhança (MV)**. Para uma variável em que $f(Y|X)$ é Gaussiana, a solução pelo **MMQ** e pela **MV** convergem. Fisher se dedicou também ao problema de encontrar uma distribuição estatística para o coeficiente de correlação de Pearson.

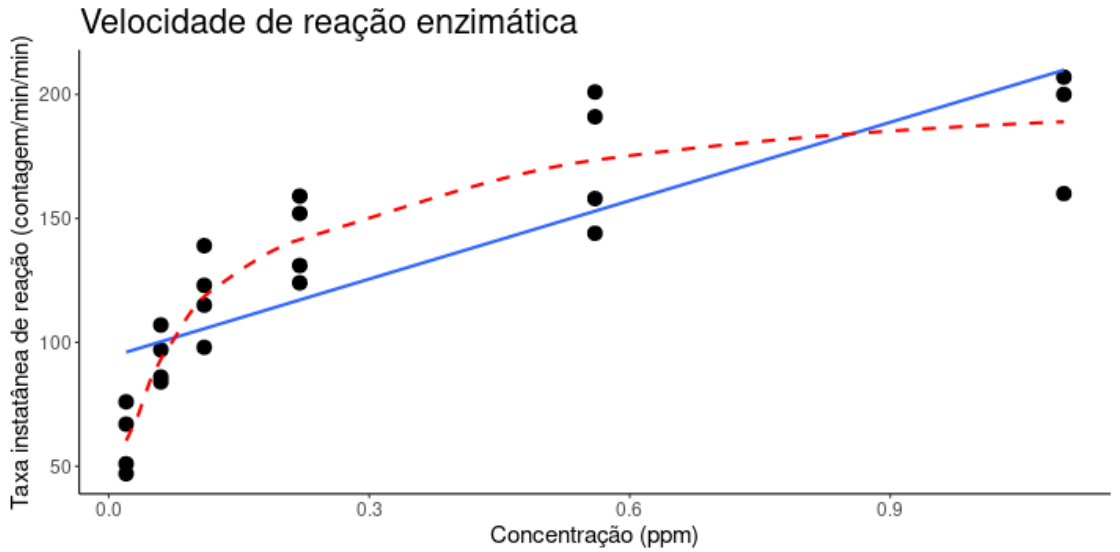


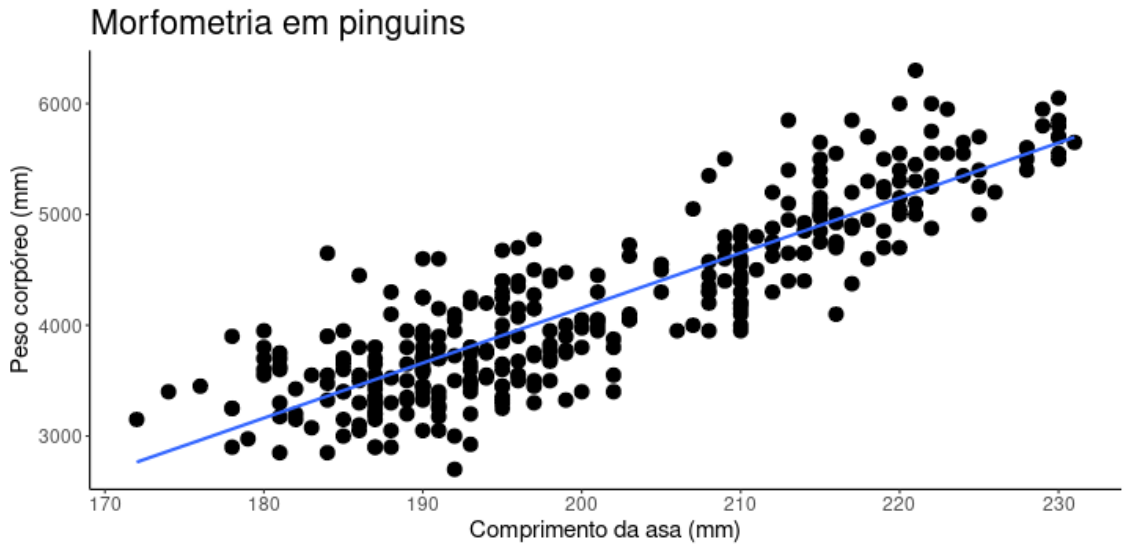
Conteúdo da aula

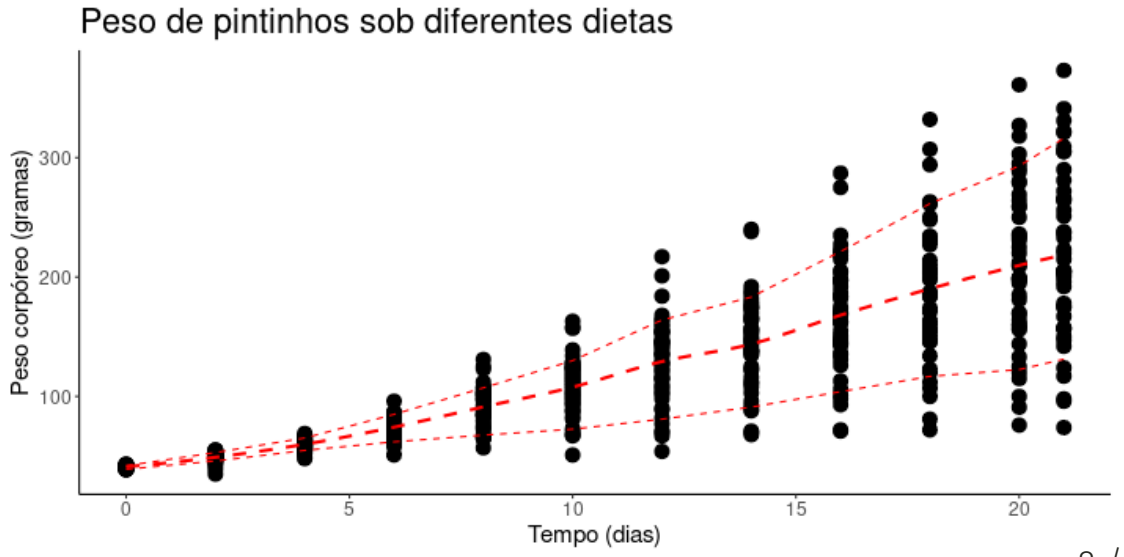
1. Medindo a intensidade de relações lineares
2. Variâncias e covariâncias
3. O coeficiente de correlaçao linear de Pearson
4. Teste de hipóteses sobre o r de Pearson
5. Regressão Linear Simples
6. Teste de hipóteses
7. Intervalos de confiança e de predição
8. Partição da Soma dos Quadrados e variação explicada
9. Os comandos em R
10. Pressupostos do modelo
11. Transformações lineares

1. Medindo a intensidade de relações lineares

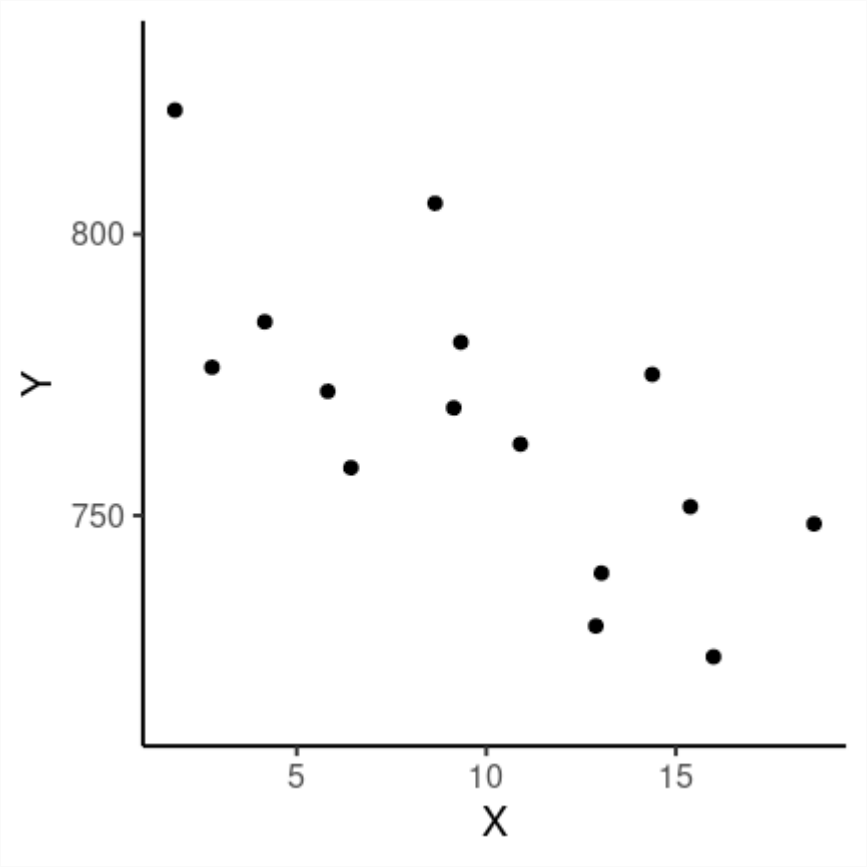




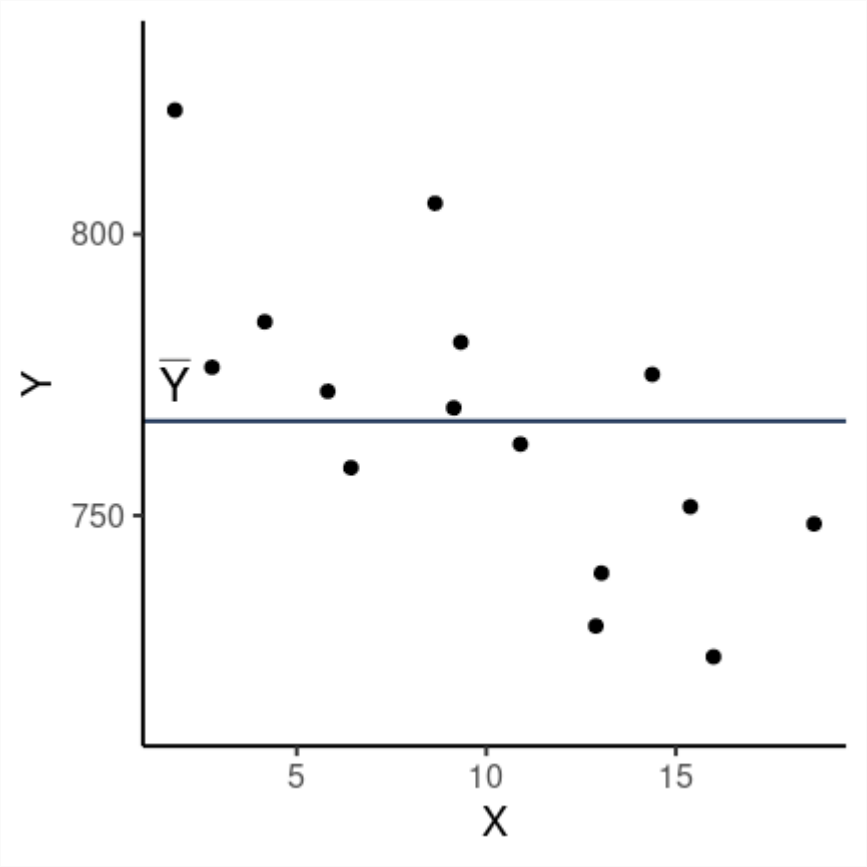




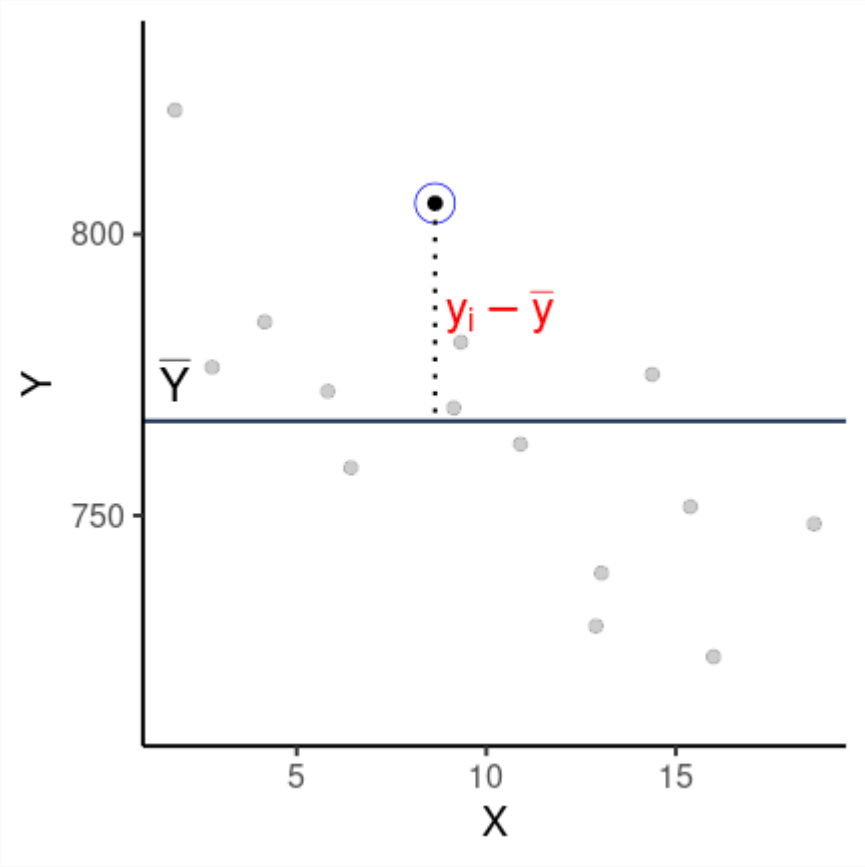
2. Variâncias e Covariâncias



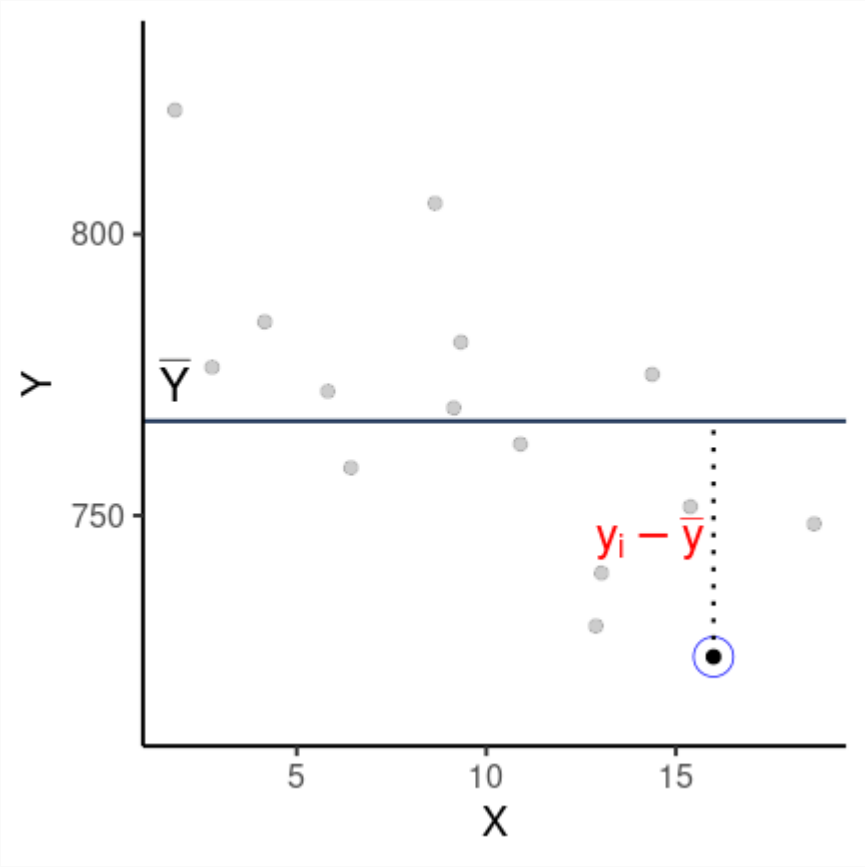
2. Variâncias e Covariâncias

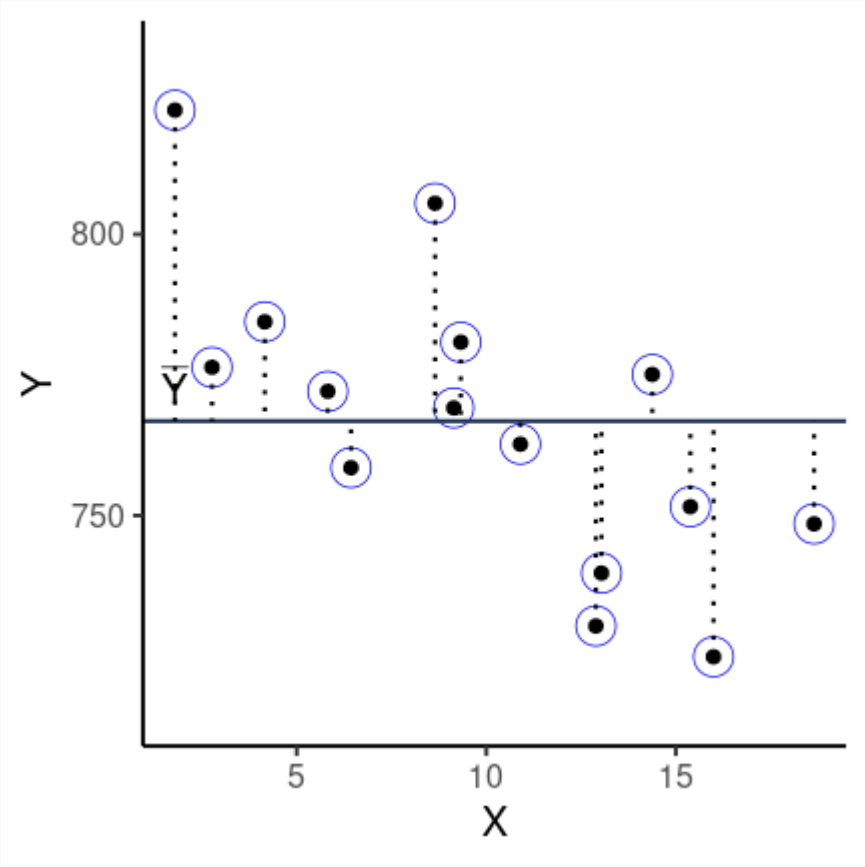


2. Variâncias e Covariâncias



2. Variâncias e Covariâncias



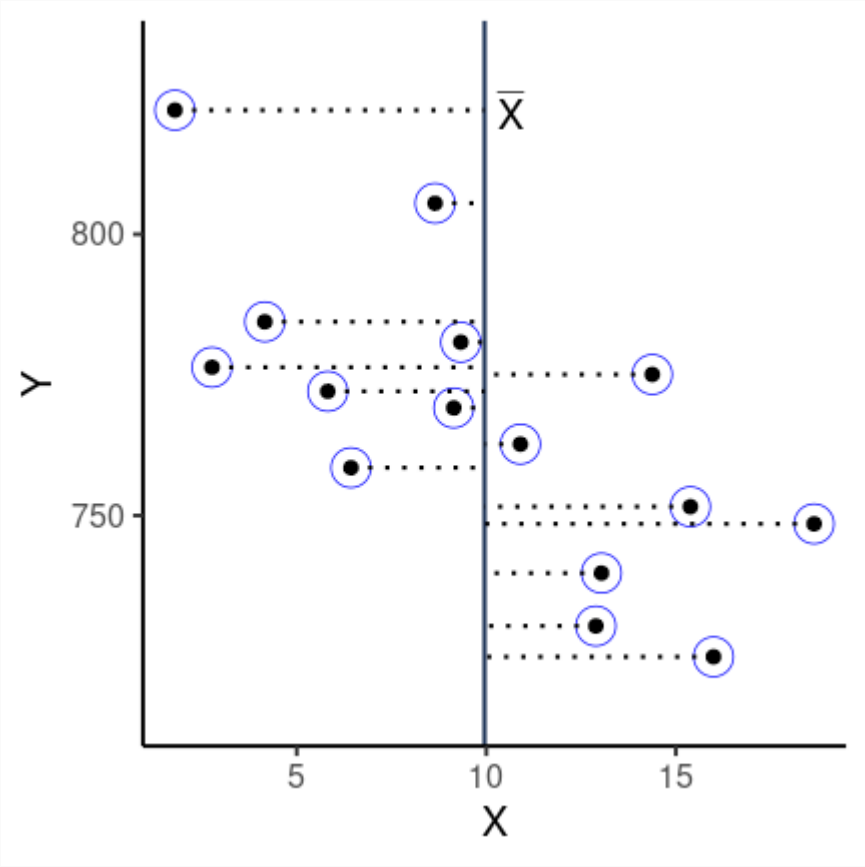


Soma dos Quadrados de Y

$$SQ_Y = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})$$

Variância amostral de Y

$$s_Y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

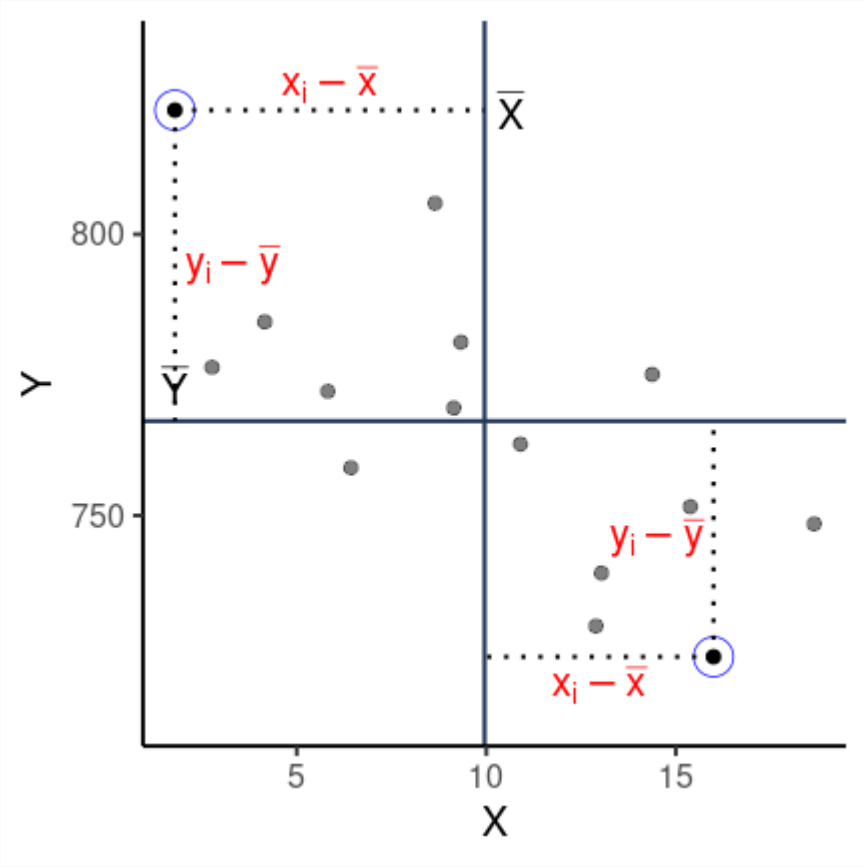


Soma dos Quadrados de X

$$SQ_X = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})$$

Variância amostral de X

$$s_X^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

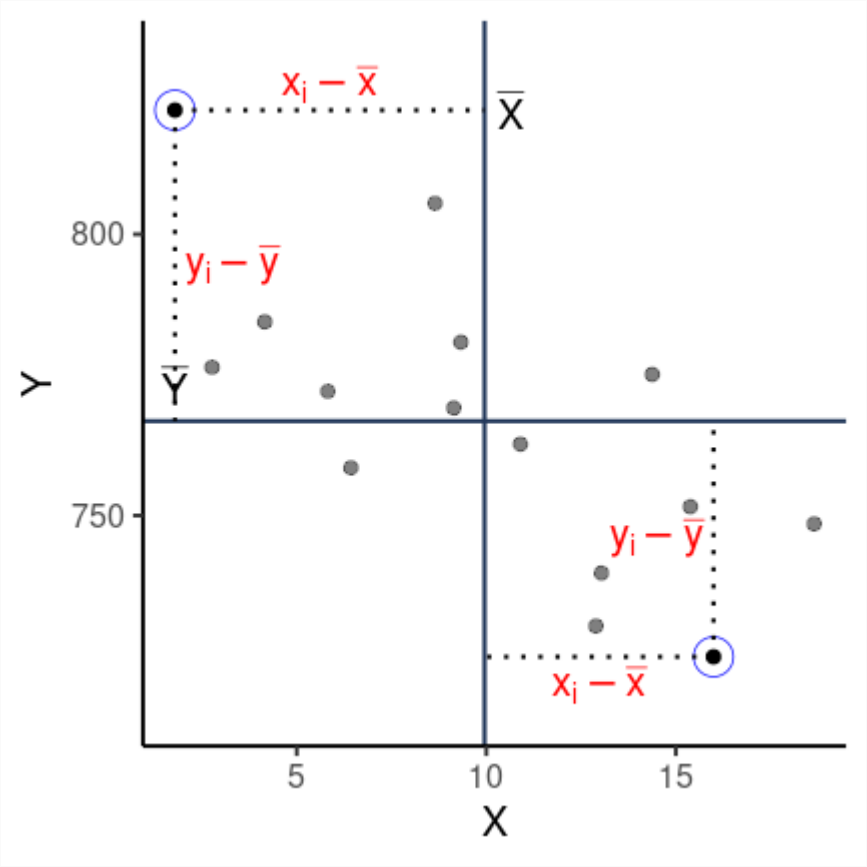


Soma dos produtos cruzados de Y e X

$$SQ_{YX} = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$$

Covariância amostral entre Y e X

$$s_{YX} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1}$$



Se

$$(y_i - \bar{y}) > 0; (x_i - \bar{x}) < 0$$

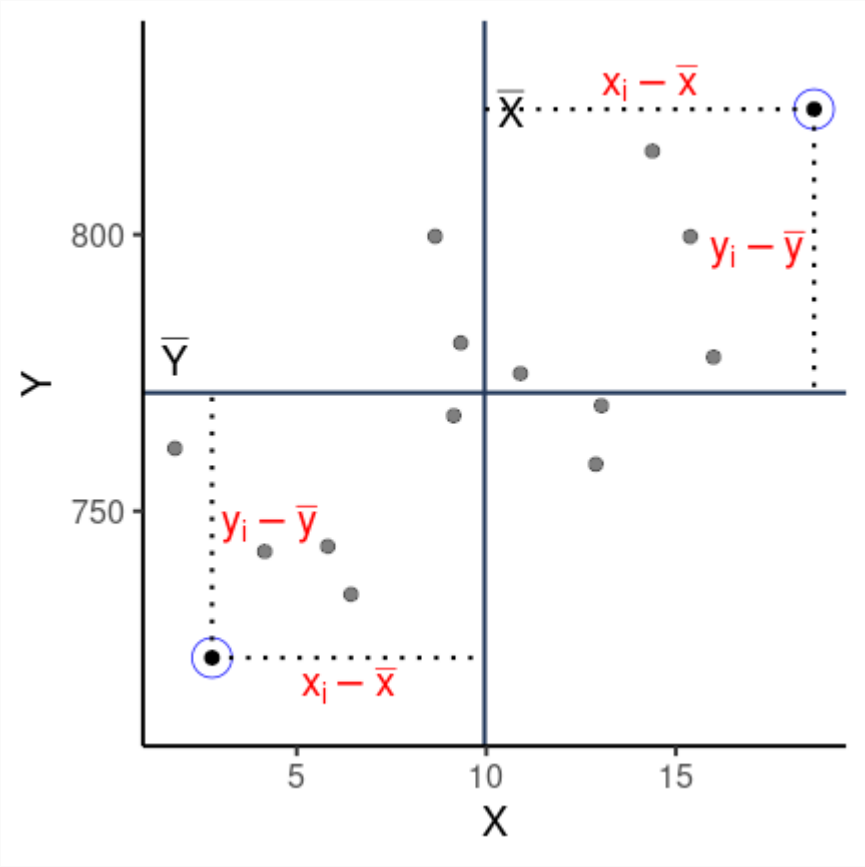
ou

$$(y_i - \bar{y}) < 0; (x_i - \bar{x}) > 0$$

temos

$$s_{YX} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1} < 0$$

A covariância pode ser **NEGATIVA**



Se

$$(y_i - \bar{y}) > 0; (x_i - \bar{x}) > 0$$

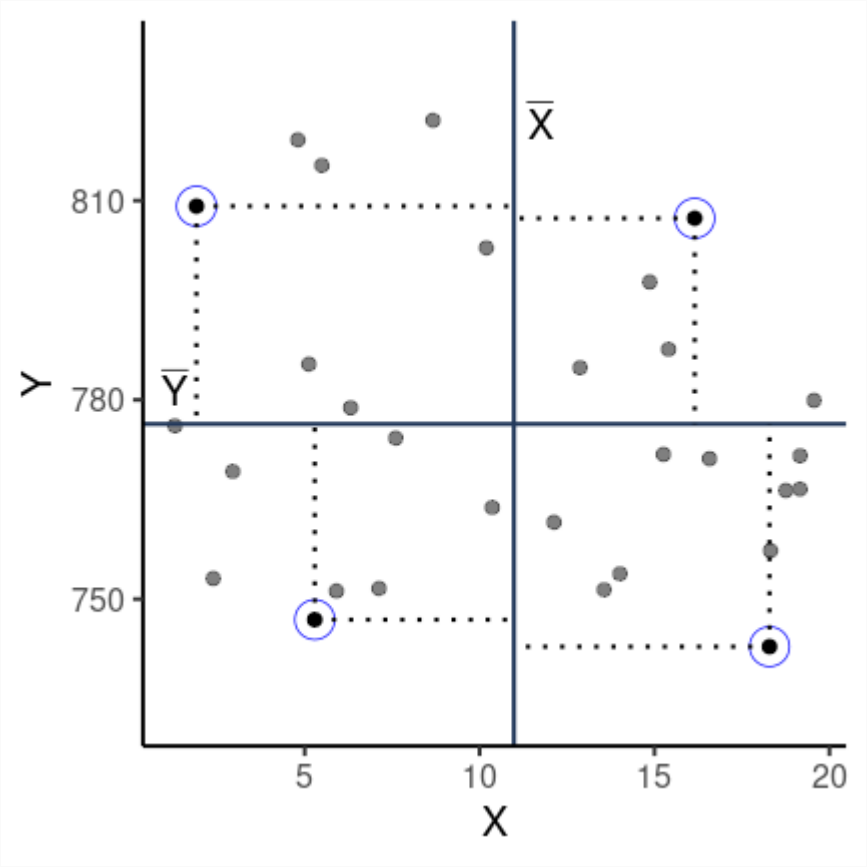
ou

$$(y_i - \bar{y}) < 0; (x_i - \bar{x}) < 0$$

temos

$$s_{YX} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1} > 0$$

A covariância pode ser **POSITIVA**



Se

$$(y_i - \bar{y}) \approx 0; (x_i - \bar{x}) \approx 0$$

ou

$$(y_i - \bar{y}) \approx 0; (x_i - \bar{x}) \approx 0$$

Temos

$$s_{YX} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1} \approx 0$$

A covariância pode ser **NULA**

3. O coeficiente de correlaçao linear de Pearson

Covariância amostral entre Y e X

$$s_{YX} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1}$$

Variância amostral de Y

$$s_Y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

Variância amostral de X

$$s_X^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

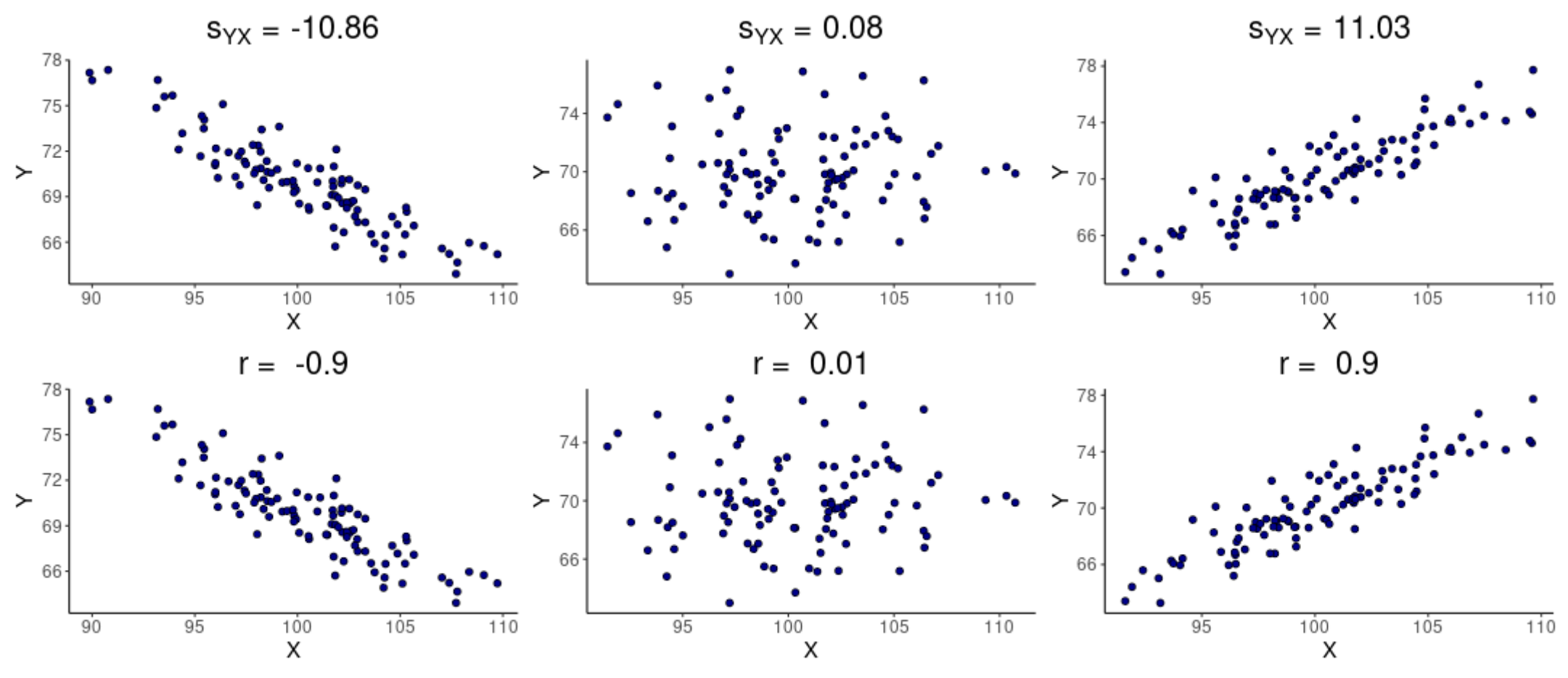
O coeficiente de correlaçao linear de Pearson r

$$r = \frac{s_{YX}}{\sqrt{s_Y^2} \times \sqrt{s_X^2}}$$

O r de Pearson é a covariância **padronizada** pelos desvios padrões de Y e X

3. O coeficiente de correlaçao linear de Pearson

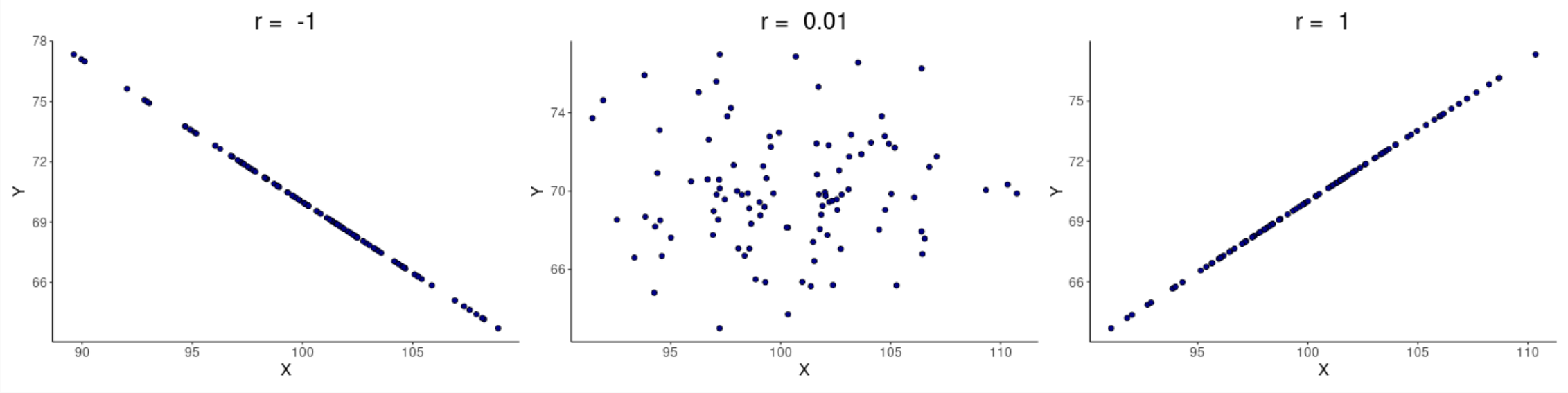
A covariância não tem limites negativos ou positivos. A escala depende das magnitudes de Y e de X.



O r de Pearson varia entre -1 e $+1$.

3. O coeficiente de correlaçao linear de Pearson

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$



- $r = -1$ (Associção linear perfeitamente **negativa**)
- $r = 0$ (Associção linear inexistente)
- $r = 1$ (Associção linear perfeitamente **positiva**)

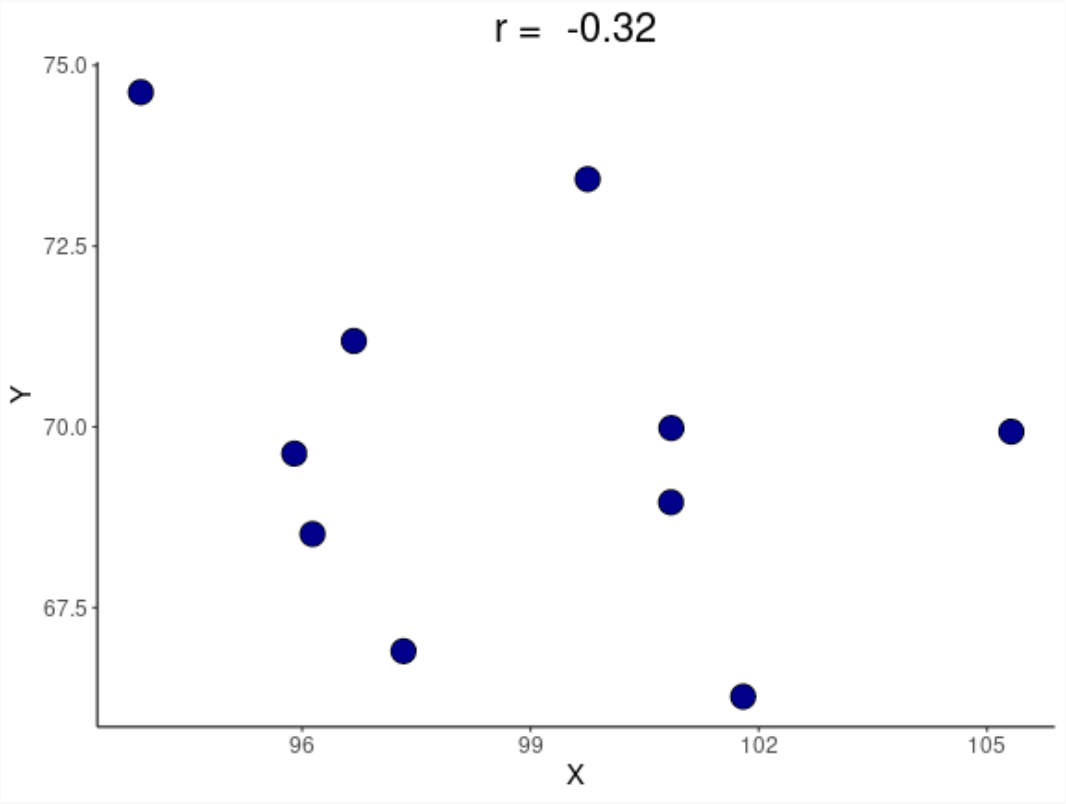
4. Teste de hipóteses sobre o r de Pearson

Dada uma **amostra** com n observações para os pares Y e X, a correlação entre Y e X na **população estatística** é diferente de zero?

$H_0 : \rho = 0$

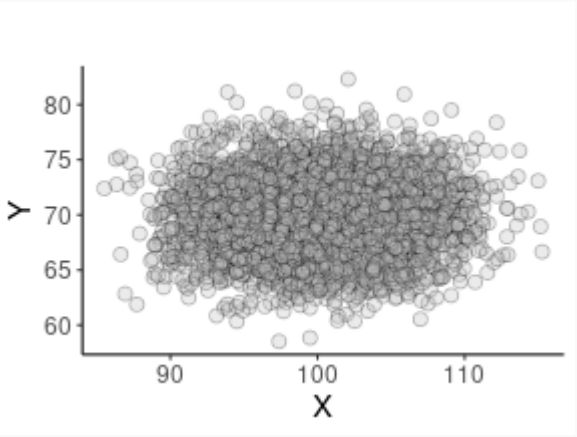
$H_a : \rho \neq 0$

$n = 10$

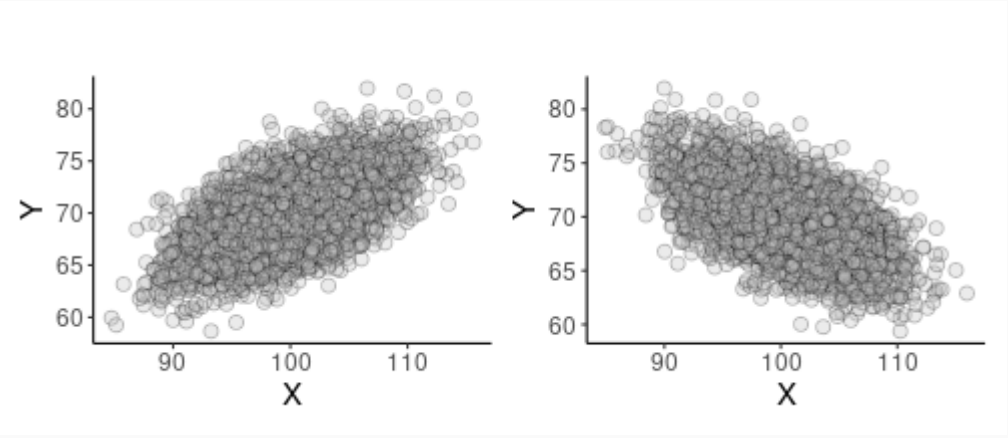


4. Teste de hipóteses sobre o r de Pearson

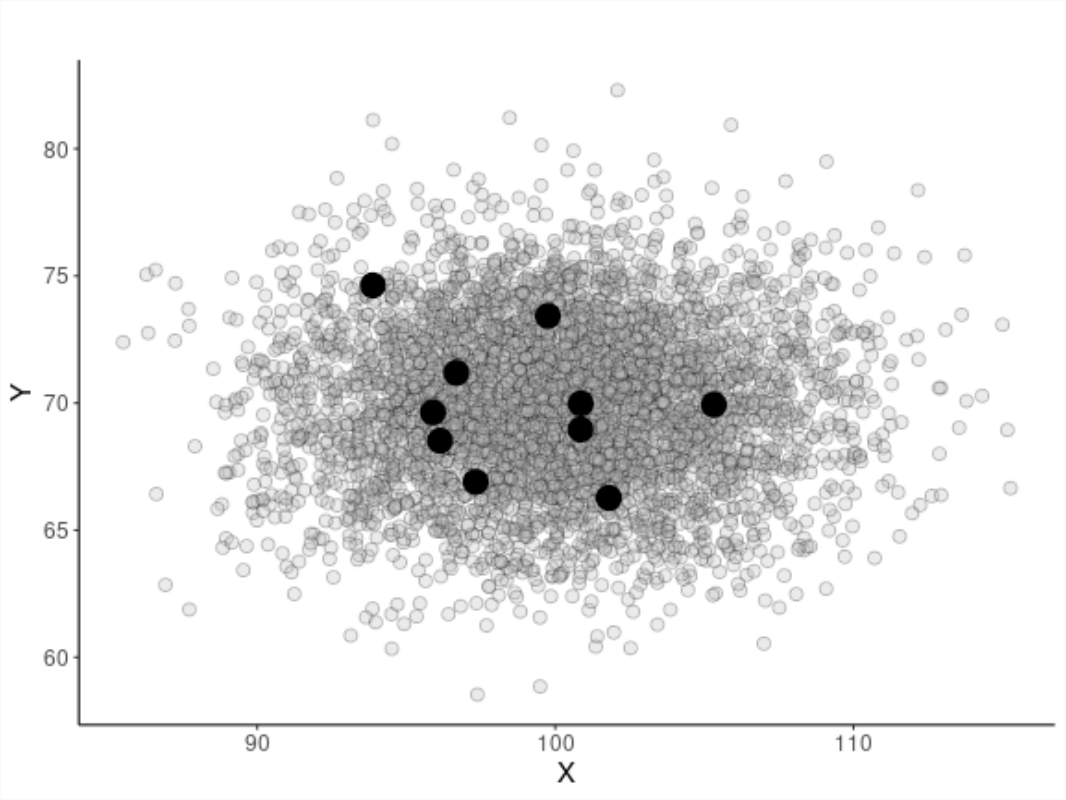
$H_0 : \rho = 0$



$H_a : \rho \neq 0$



Os dados segundo H_0



4. Teste de hipóteses sobre o r de Pearson

Assumimos que distribuição conjunta entre $f(Y, X)$ é Normal.

$H_0 : \rho = 0$

$H_a : \rho \neq 0$

$\alpha = 0.05$

$n = 10$

$r = -0.32$

Estatística do teste - t

$$t_{calculado} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Segundo H_0

4. Teste de hipóteses sobre o r de Pearson

Teste de hipótese sobre ρ

$\overline{Y} = 98.85$; $\overline{X} = 69.94$; $n = 10$

$r = -0.32$

$$t_{calculado} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{-0.32}{\sqrt{\frac{1-(-0.32)^2}{8}}} = -0.965$$

$p = 0.363$

Assumindo $\alpha = 0.05$, **Aceito** H_0:

> Não há evidências de correlação entre Y e X.

Cálculo do coeficiente de correlação

	Y	X	$\sum (y_i - \overline{y})^2$	$\sum (x_i - \overline{x})^2$	$(y_i - \overline{y})(x_i - \overline{x})$
1	95.9	69.63	8.72	0.10	0.92
2	101.8	66.27	8.68	13.50	-10.83
3	100.85	69.98	4.01	0.00	0.08
4	99.75	73.43	0.81	12.13	3.14
5	93.88	74.63	24.69	21.95	-23.28
6	97.33	66.9	2.30	9.27	4.62
7	96.68	71.19	4.71	1.55	-2.70
8	100.85	68.96	3.99	0.97	-1.97
9	96.14	68.52	7.36	2.03	3.87
10	105.32	69.93	41.87	0.00	-0.06
$\sum$			**107.15**	**61.52**	**-26.21**

4. Teste de hipóteses sobre o r de Pearson

Aumentando o tamanho amostral

$H_0 : \rho = 0$

$H_a : \rho \neq 0$

$\alpha = 0.05$

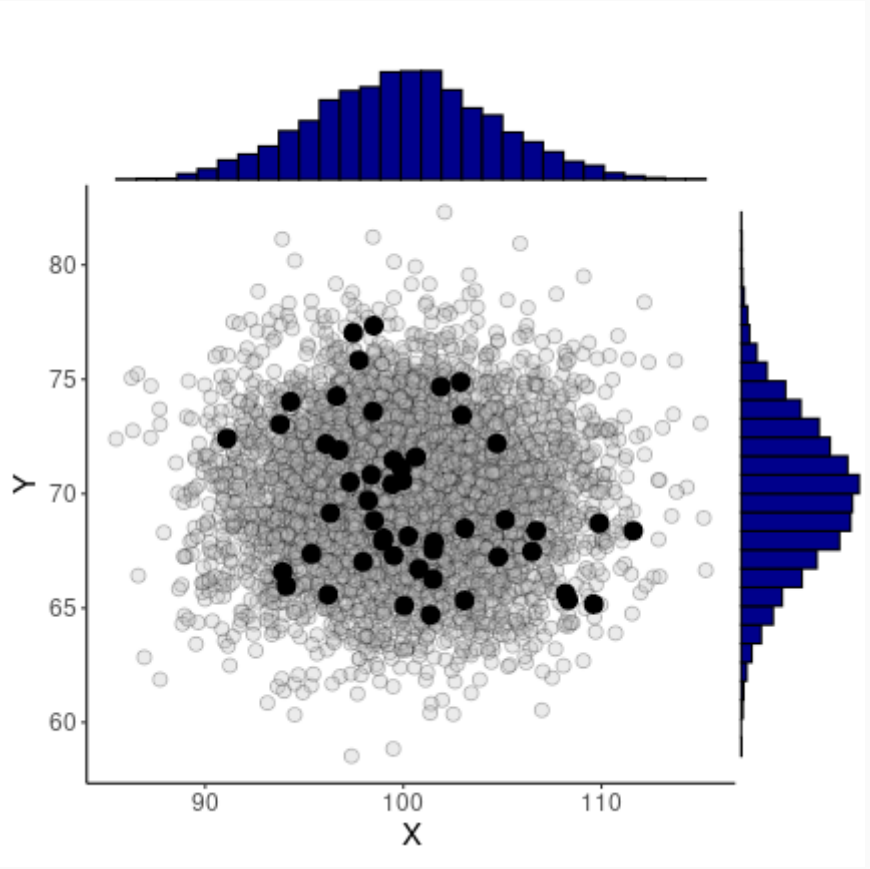
$n = 50$

$r = -0.32$

Estatística do teste - t

$$t_{calculado} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Segundo H_0



4. Teste de hipóteses sobre o r de Pearson

Teste de hipótese sobre ρ

$\overline{Y} = 100.41; \overline{X} = 69.64; n = 50$

$r = -0.32$

$t_{calculado} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{-0.32}{\sqrt{\frac{1-(-0.32)^2}{48}}} = -2.363$

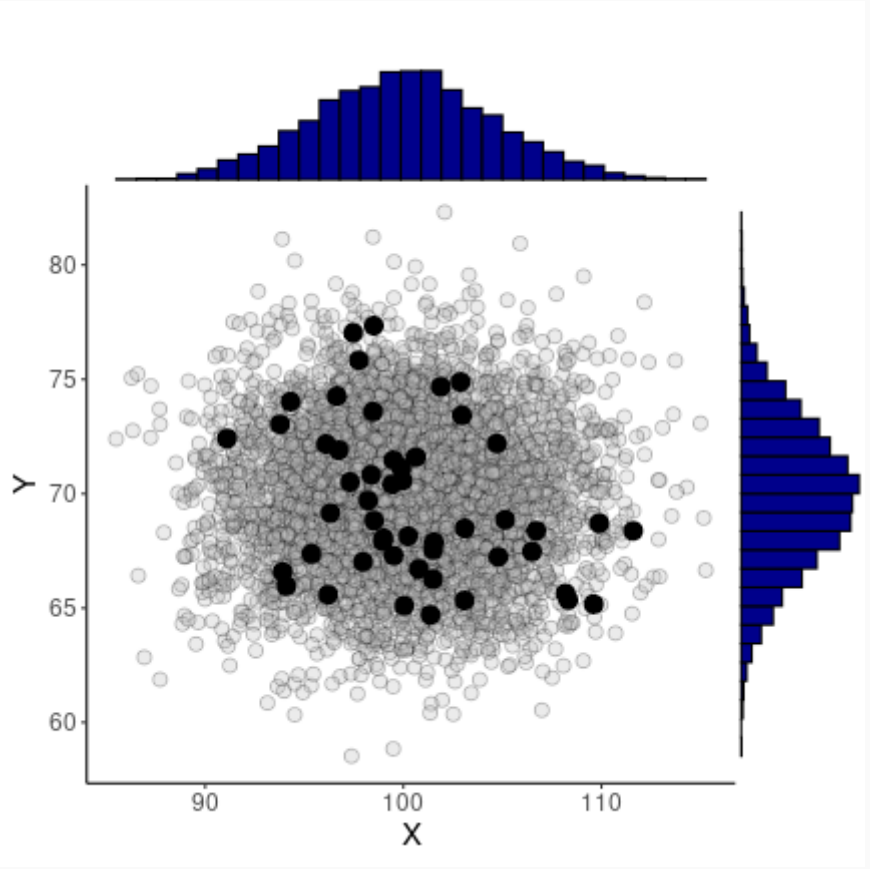
$p = 0.022$

Assumindo $\alpha = 0.05$, **Rejeito** H_0:

> Há evidências de correlação entre Y e X

.

Segundo H_0



4. Teste de hipóteses sobre o r de Pearson

$r = -0.32; n = 10$

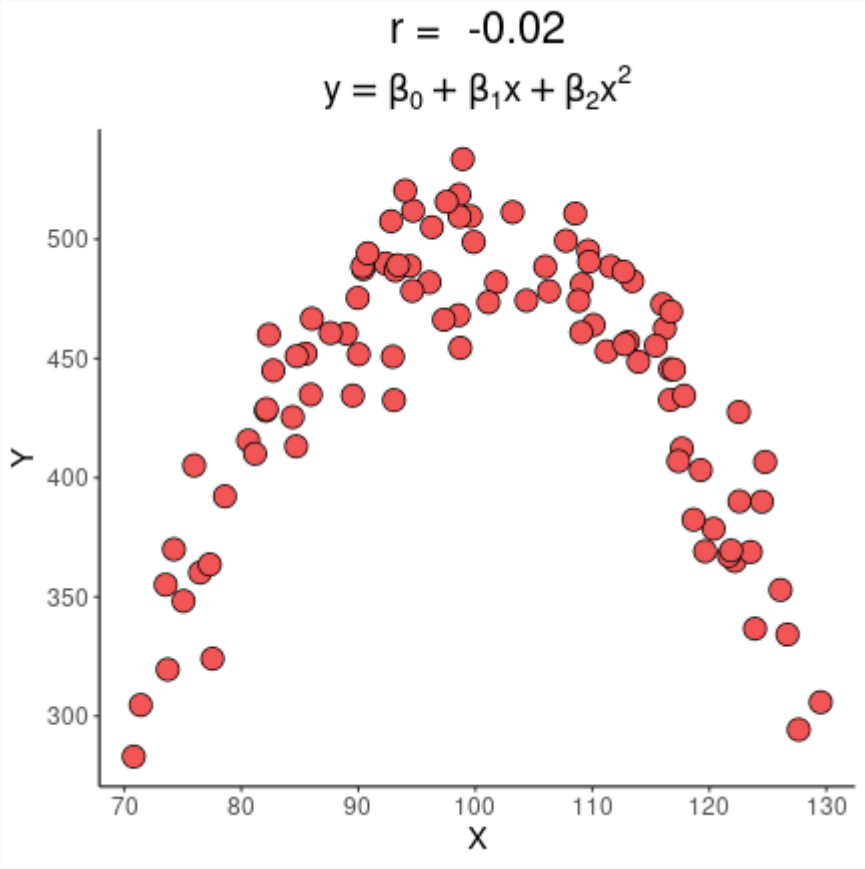
$t_{calculado} = -0.965; p = 0.363$

$r = -0.32; n = 50$

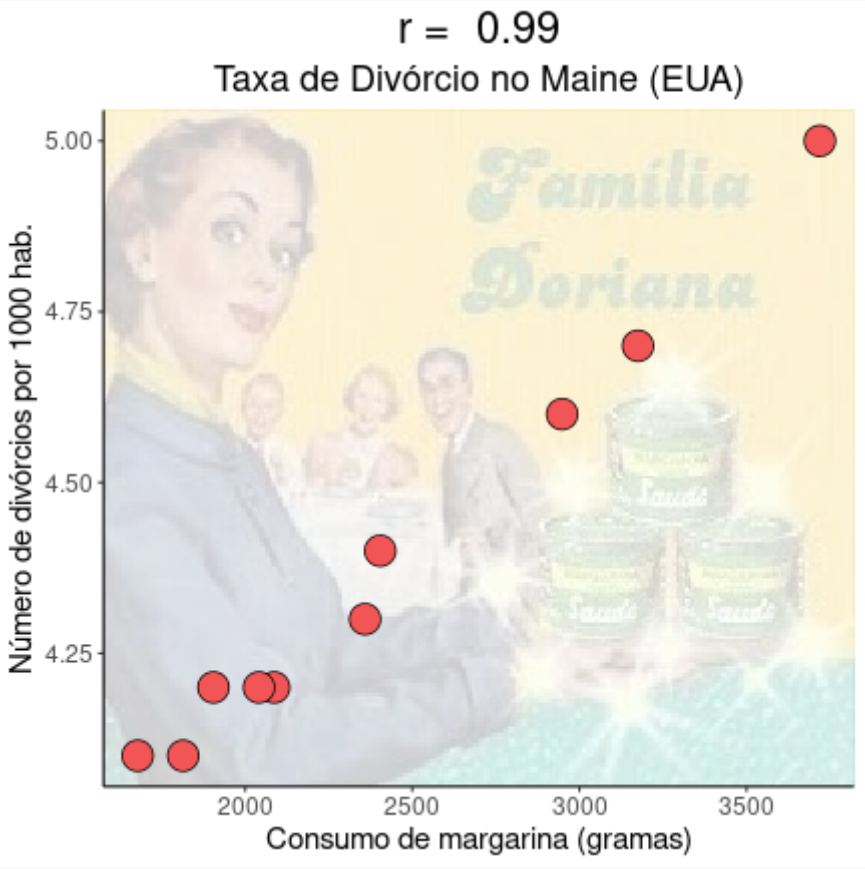
$t_{calculado} = -2.363; p = 0.022$

4. Teste de hipóteses sobre o r de Pearson

O r mede associações **lineares**



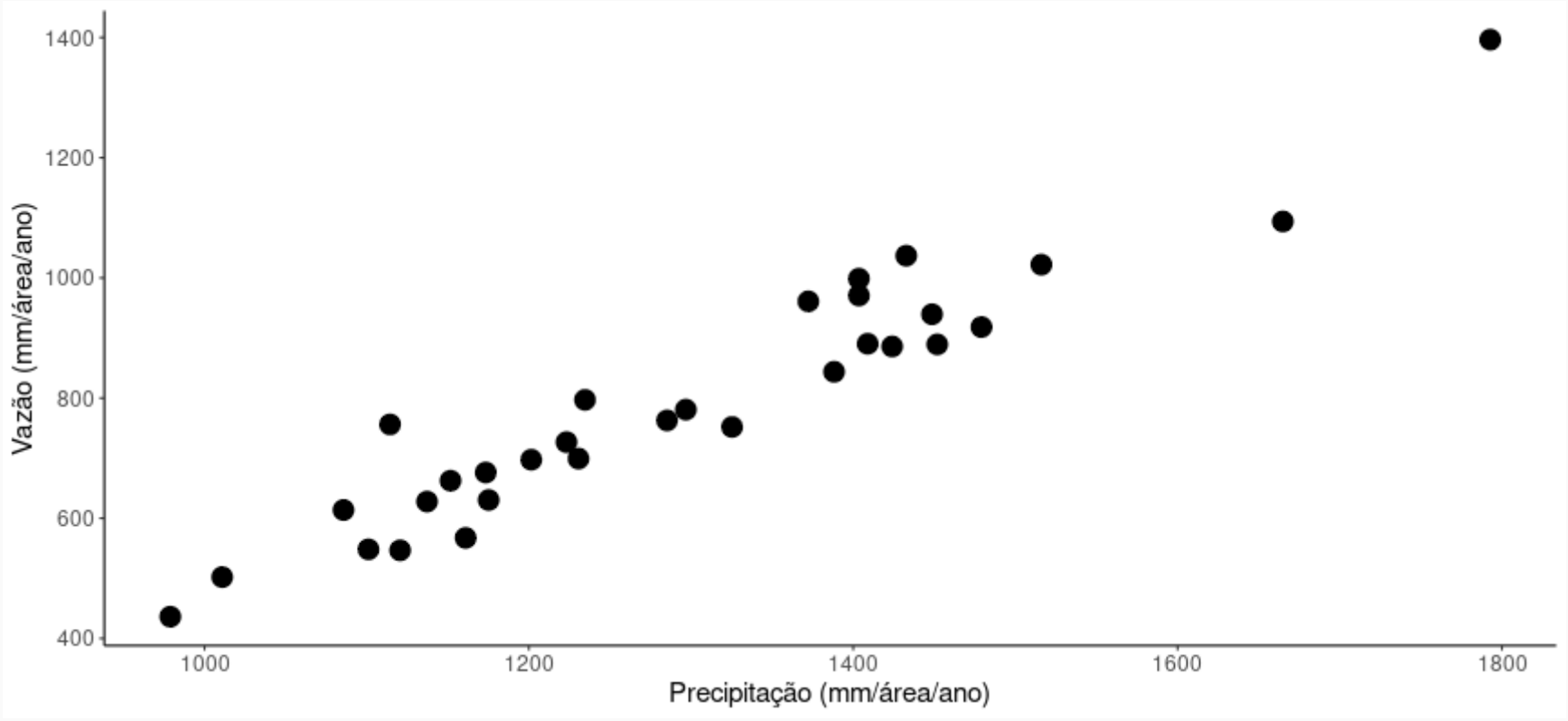
Correlação **não implica** causalidade



5. Regressão linear simples: descrevendo relações funcionais

Um serviço ecossistêmico essencial de bacias hidrográficas é o fornecimento hídrico. Em 1955, o Serviço Florestal americano estabeleceu a Floresta Experimental de **Hubbard Brook (HBEF)** como um centro de pesquisa hidrológica.

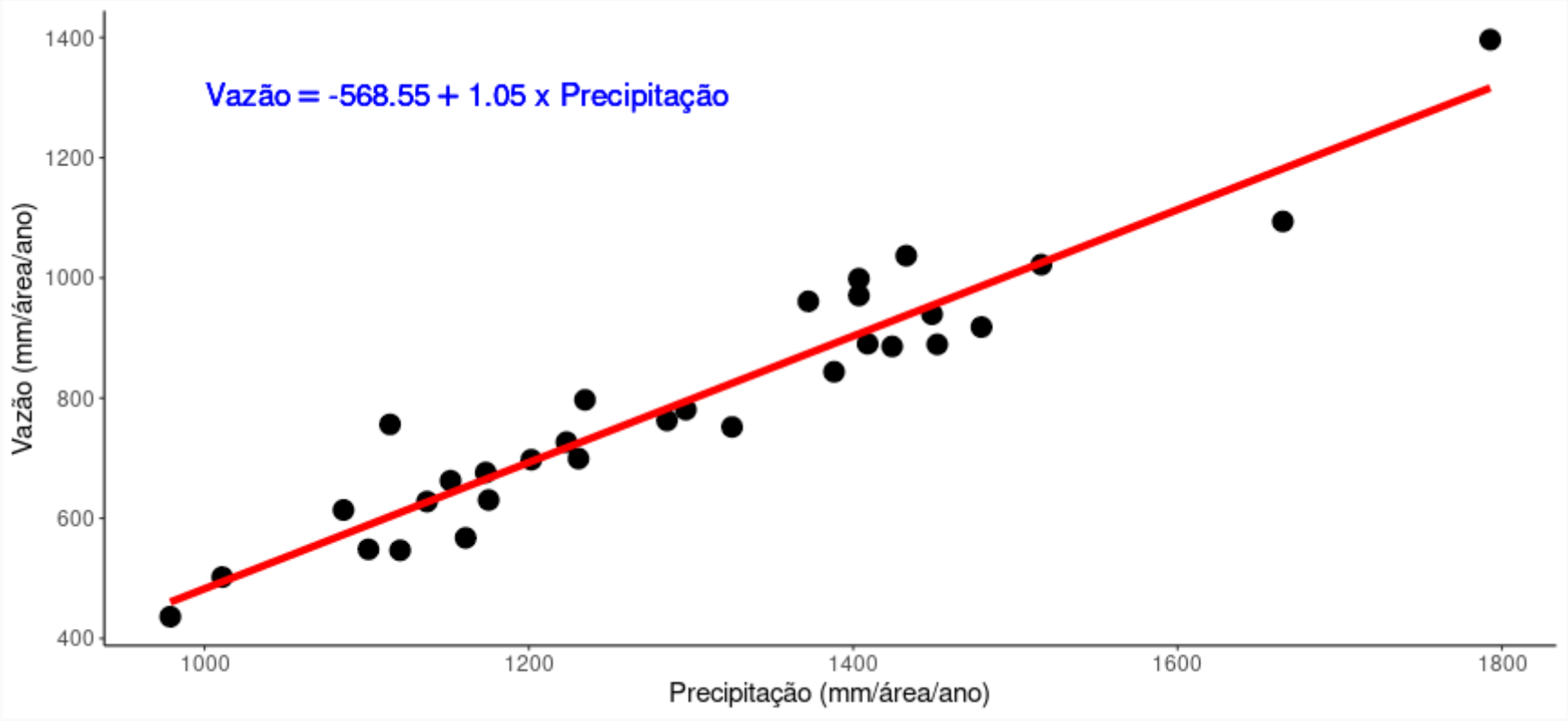
Podemos supor que o volume de água anual que uma bacia pode fornecer tem relação com o volume de chuva na região.



5. Regressão linear simples: descrevendo relações funcionais

Um serviço ecossistêmico essencial de bacias hidrográficas é o fornecimento hídrico. Em 1955, o Serviço Florestal americano estabeleceu a Floresta Experimental de **Hubbard Brook (HBEF)** como um centro de pesquisa hidrológica.

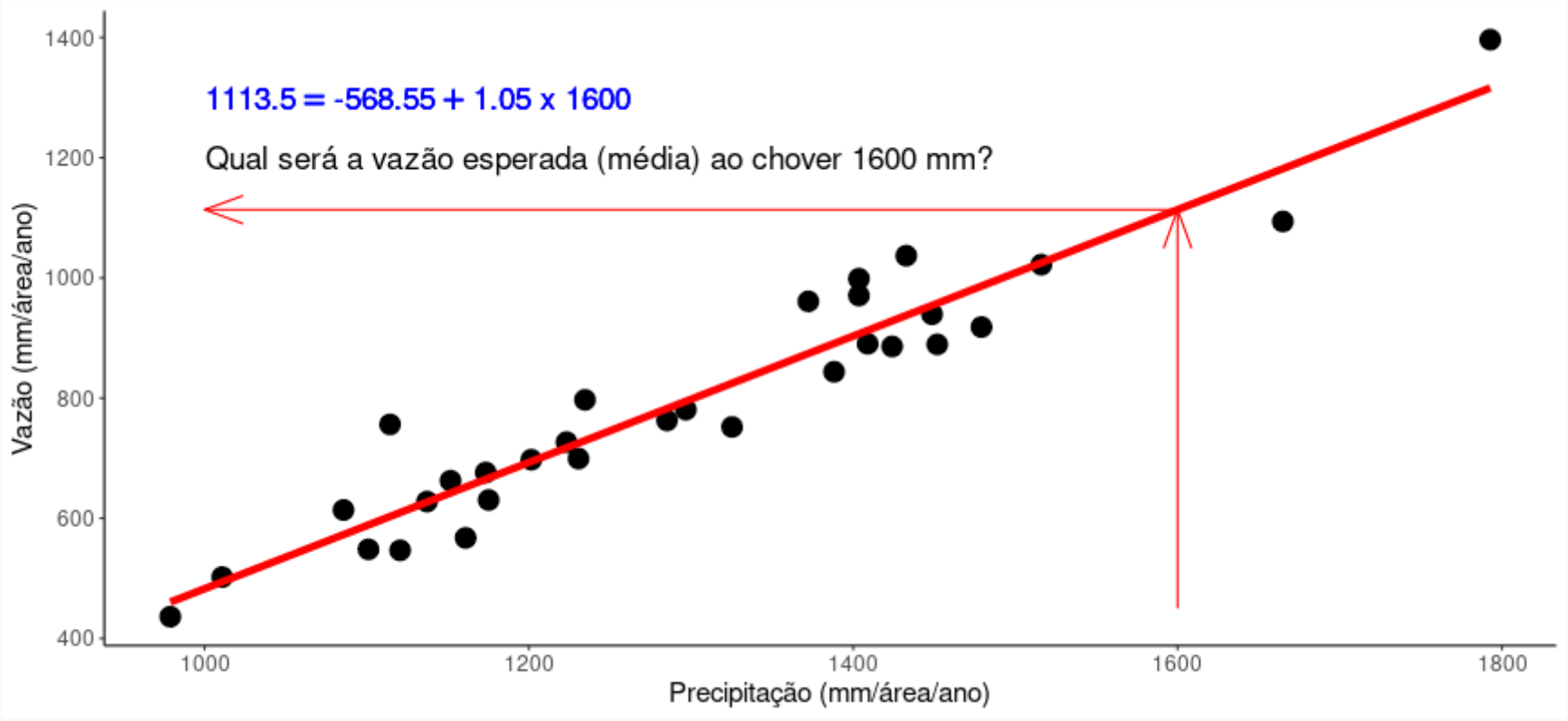
Podemos supor que o volume de água anual que uma bacia pode fornecer tem relação com o volume de chuva na região.



5. Regressão linear simples: descrevendo relações funcionais

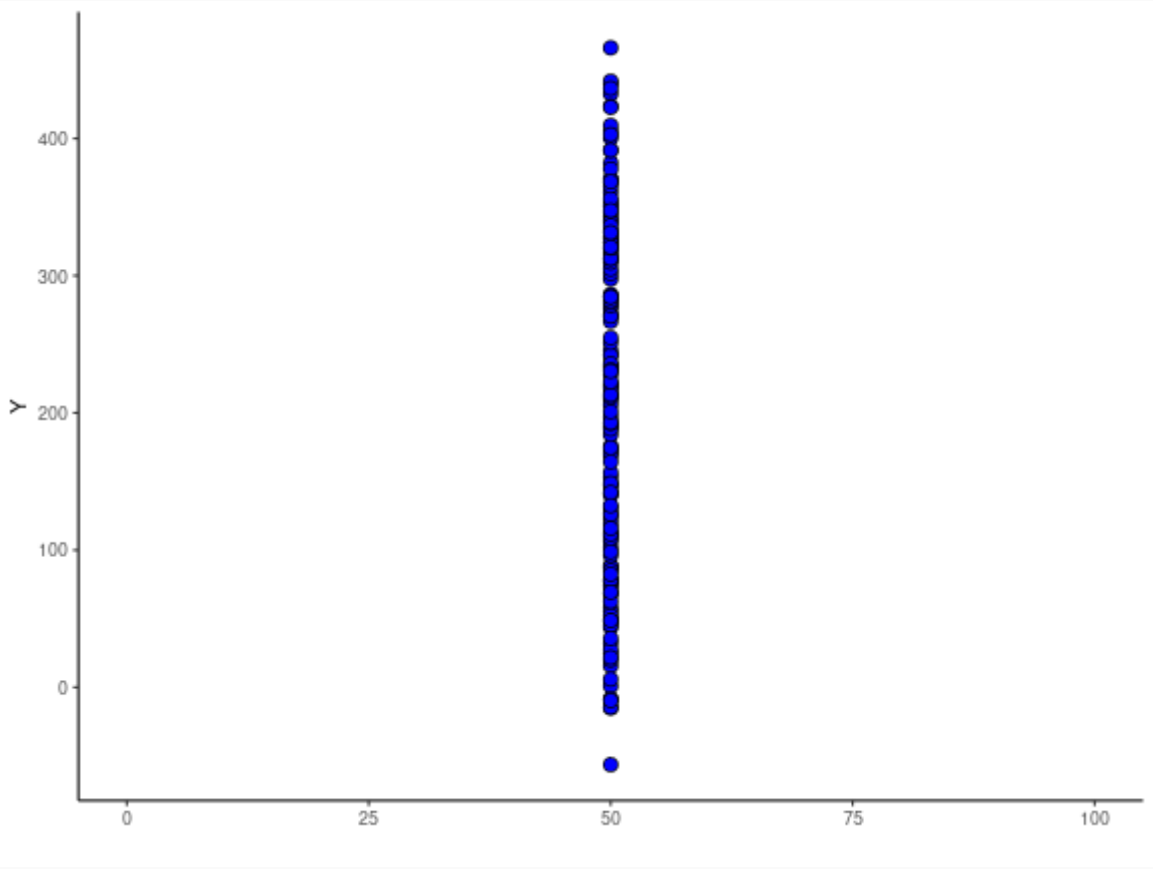
Um serviço ecossistêmico essencial de bacias hidrográficas é o fornecimento hídrico. Em 1955, o Serviço Florestal americano estabeleceu a Floresta Experimental de **Hubbard Brook (HBEF)** como um centro de pesquisa hidrológica.

Podemos supor que o volume de água anual que uma bacia pode fornecer tem relação com o volume de chuva na região.



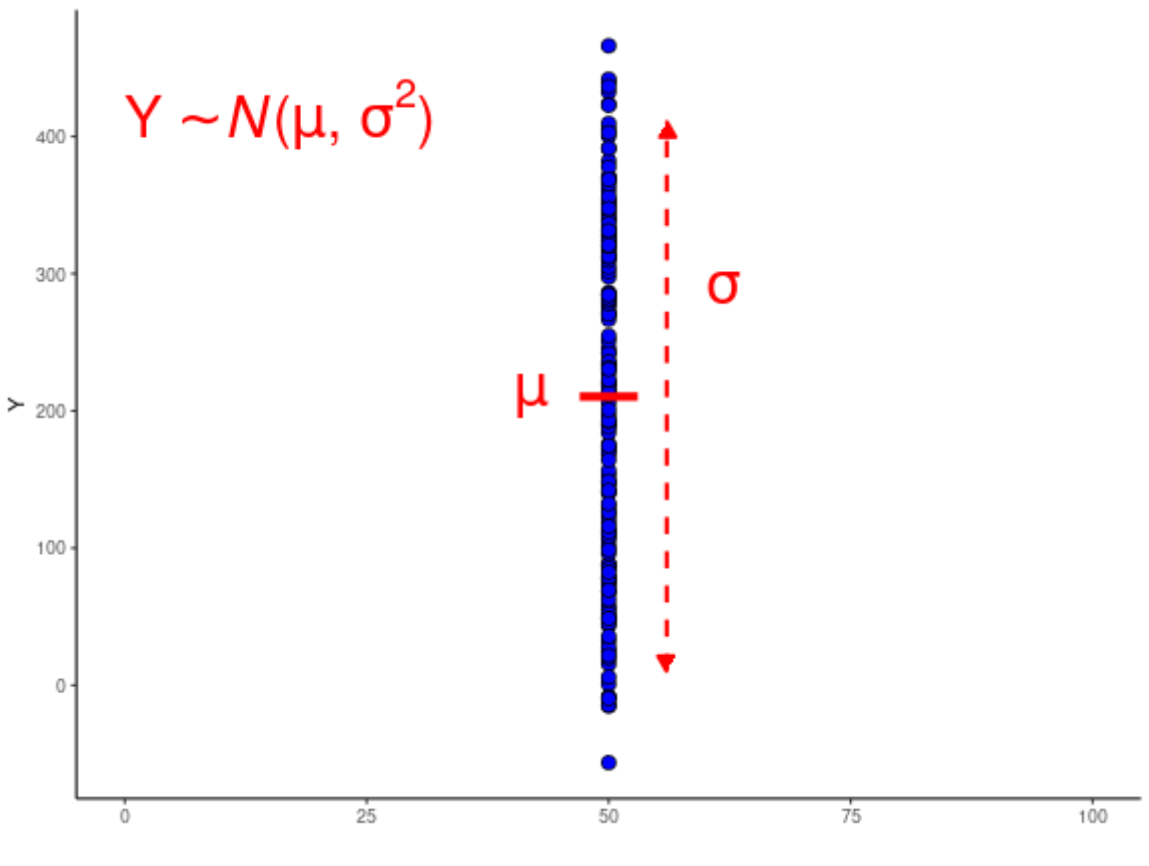
5. Regressão linear simples: estrutura geral do modelo

Seja uma variável aleatória Y com distribuição normal proveniente de um *experimento aleatório*.



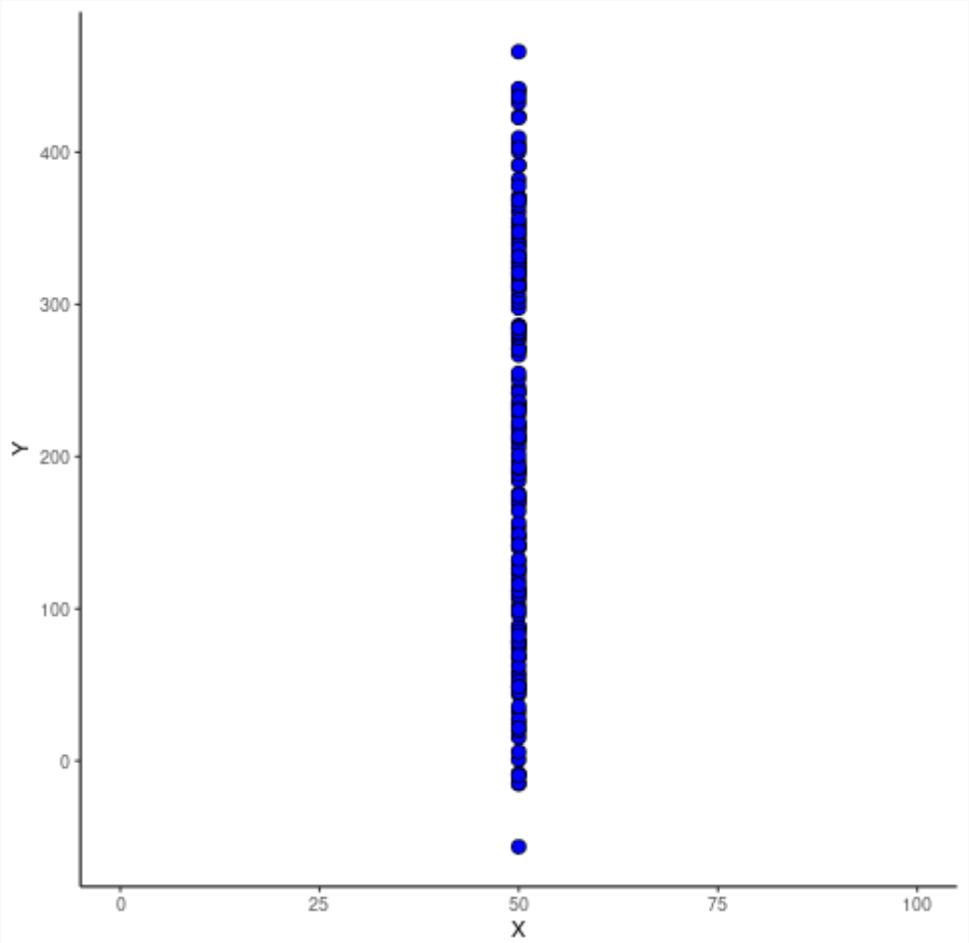
5. Regressão linear simples: estrutura geral do modelo

Seja uma variável aleatória Y com distribuição normal proveniente de um *experimento aleatório*.



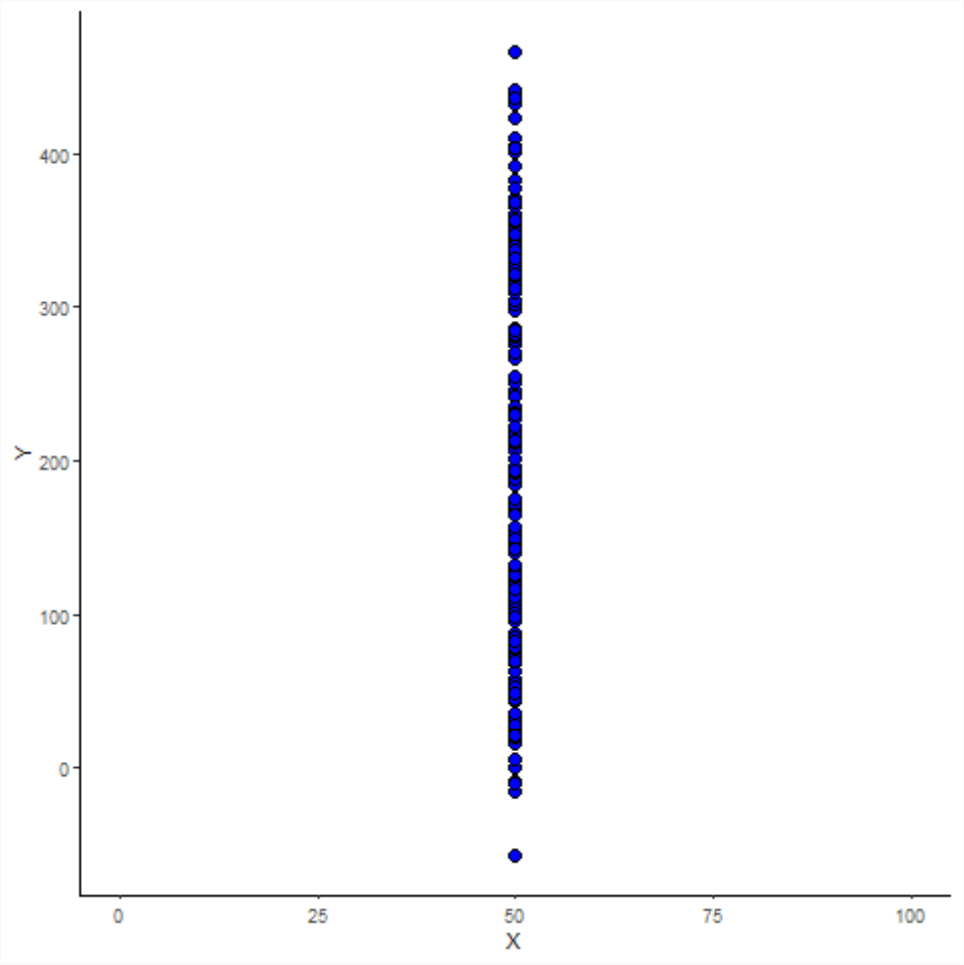
5. Regressão linear simples: estrutura geral do modelo

Para cada observação y_i é conhecida também uma informação sobre x_i.



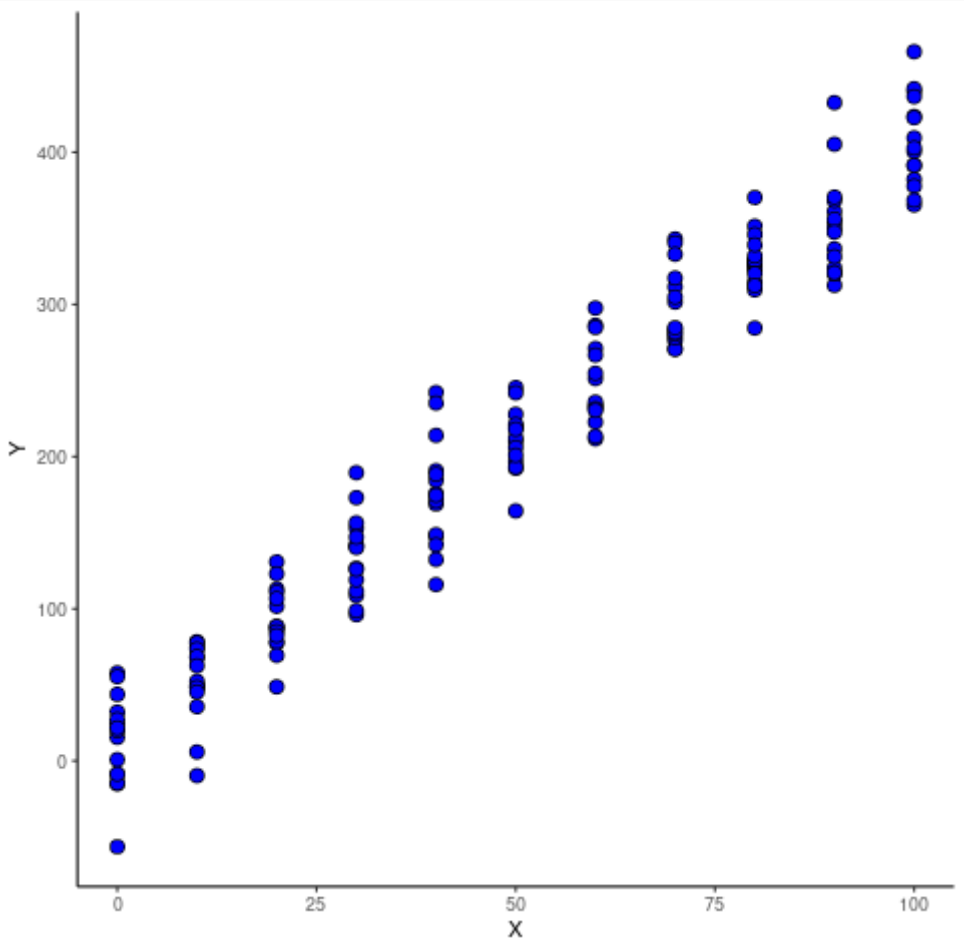
5. Regressão linear simples: estrutura geral do modelo

Para cada observação y_i é conhecida também uma informação sobre x_i.



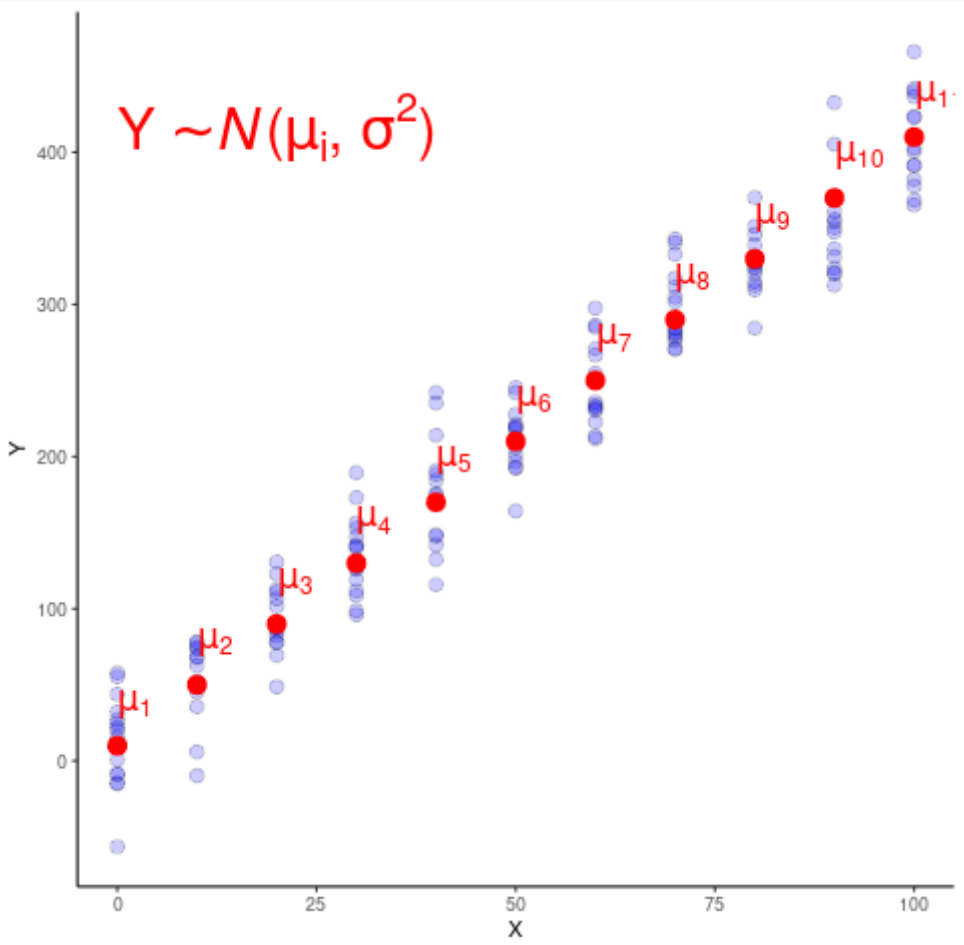
5. Regressão linear simples: estrutura geral do modelo

Para cada observação y_i é conhecida também uma informação sobre x_i.



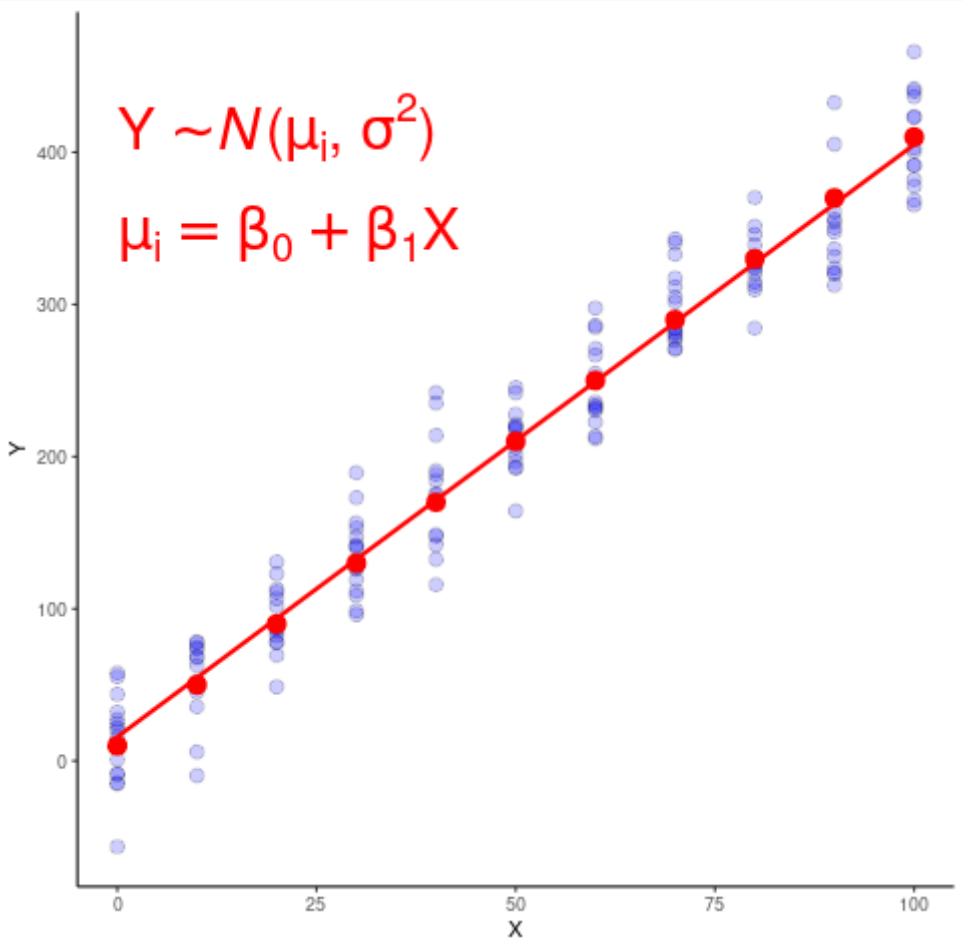
5. Regressão linear simples: estrutura geral do modelo

Para cada observação y_i é conhecida também uma informação sobre x_i.



5. Regressão linear simples: estrutura geral do modelo

Para cada observação y_i é conhecida também uma informação sobre x_i.



5. Regressão linear simples: estrutura geral do modelo

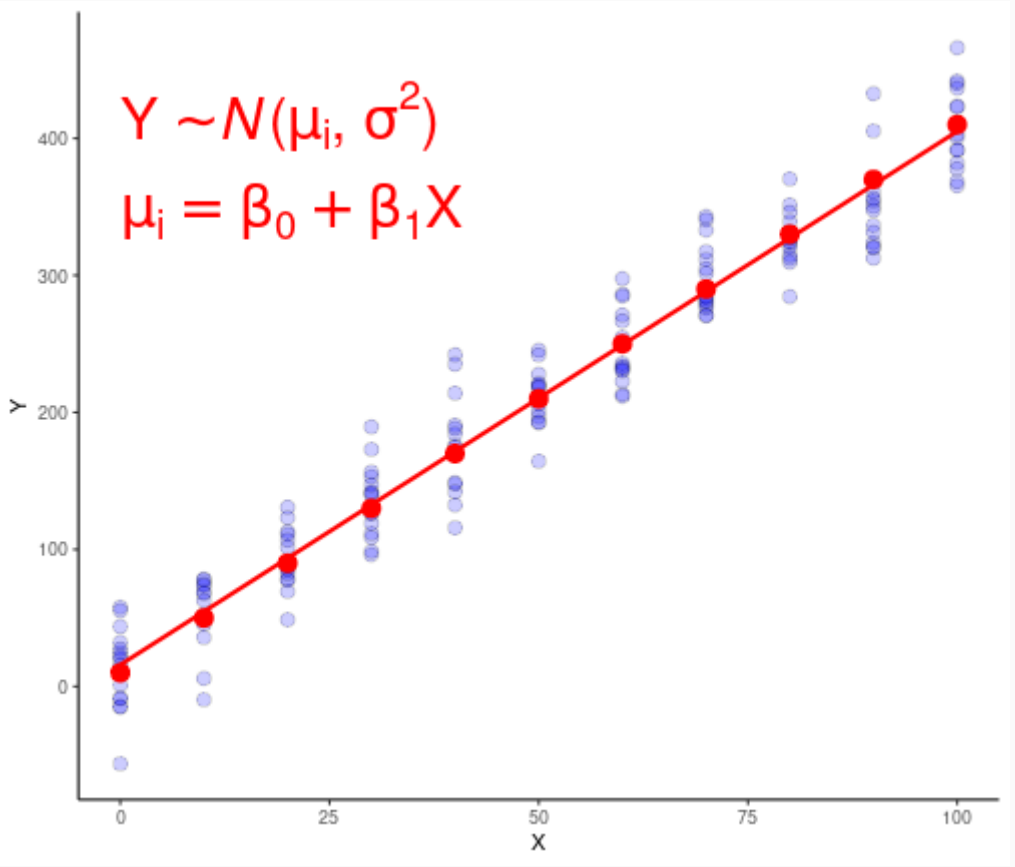
1 - As observações em Y e X compõem um par (y_i, x_i) de modo que:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}, X = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}$$

2 - X é determinada **experimentalmente** e **sem erros**.

3 - Y é uma variável aleatória normalmente distribuída, com μ_i variância σ^2.

$$Y \sim \mathcal{N}(\mu_i, \sigma^2)$$



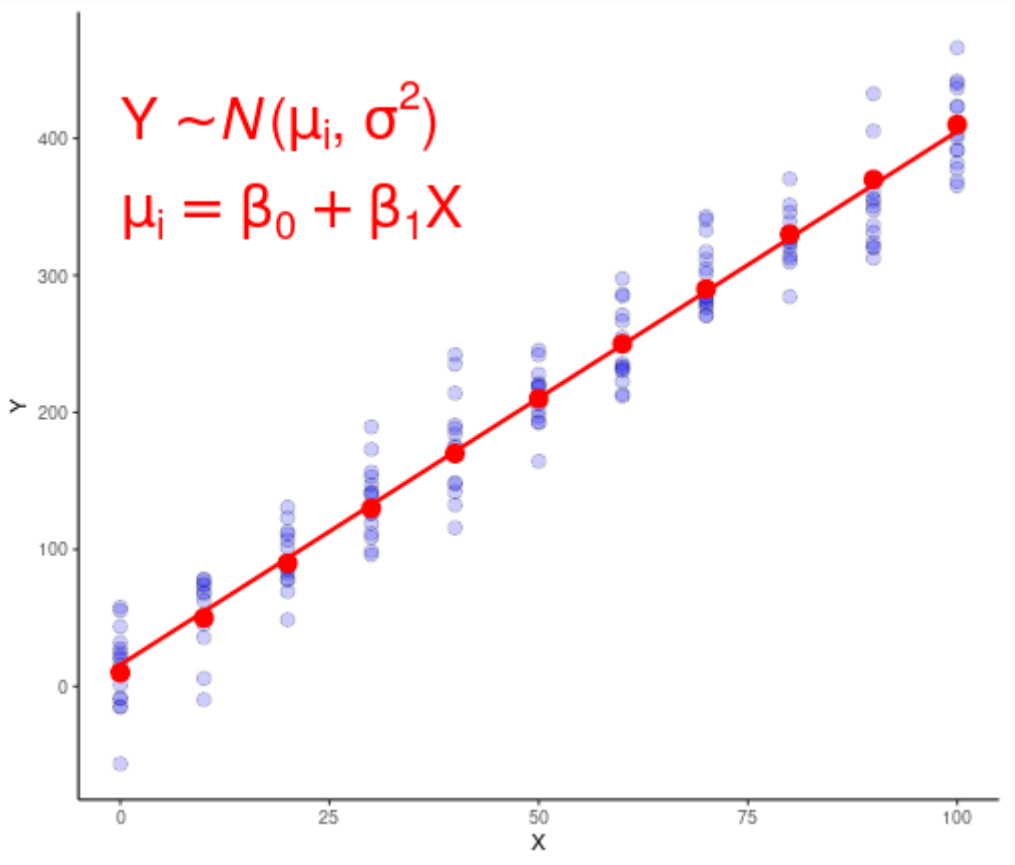
5. Regressão linear simples: estrutura geral do modelo

4 - μ_i é representado por um **modelo linear** que expressa o valor esperado de y_i para um dado valor de x_i. Compõe a **parcela determinística** do modelo.

$$E(Y|x_i) = \mu_i = \beta_0 + \beta_1 x_i$$

5 - β_0 e β_1 são as contantes a serem estimadas, representando o **intercepto** e o **coeficience de inclinação da reta**, repectivamente.

6 - σ^2 é a **variância** de Y e ser estimada. σ^2 é **constante** para todos os valores em X.



5. Regressão linear simples: o modelo matemático

Variáveis envolvidas

Y: variável resposta (dependente);

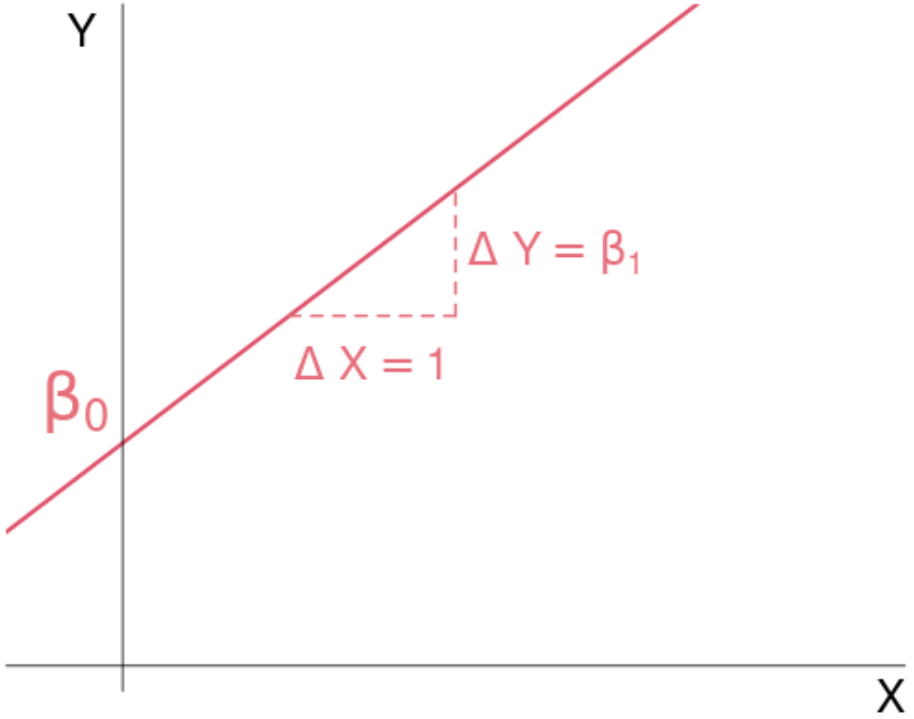
X: variável preditora (**in**dependente);

$$E(Y|x_i) = \beta_0 + \beta_1 x_i$$

Parâmetros do modelo

β_0: Intercepto;

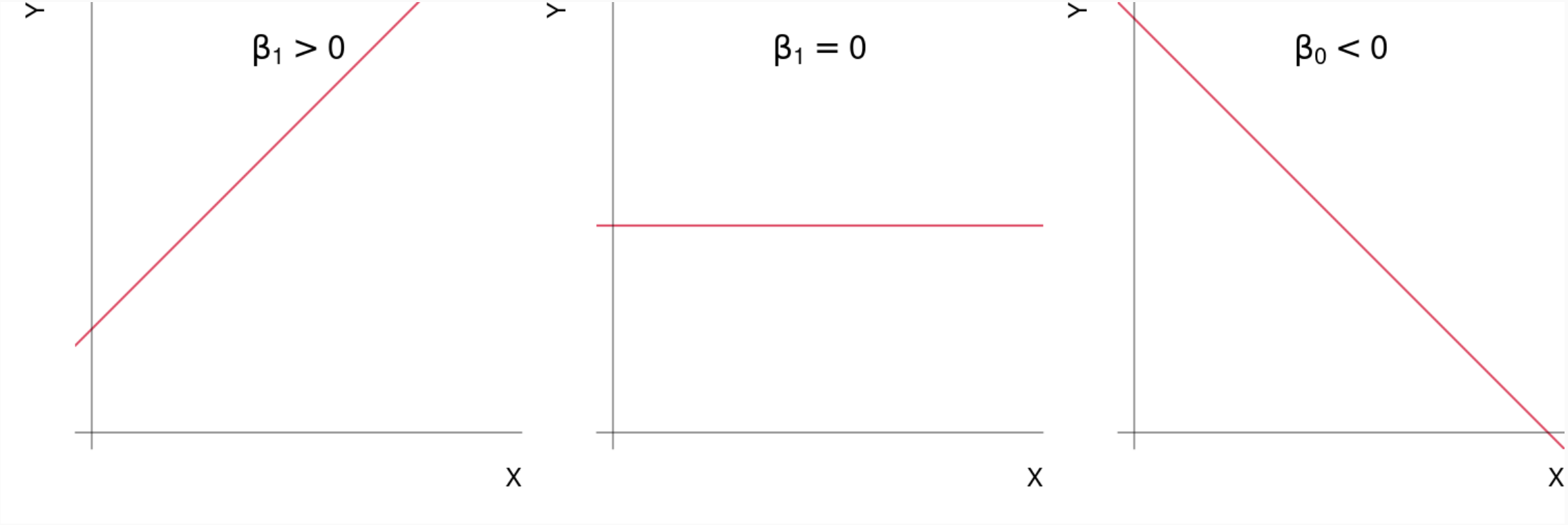
β_1: coeficiente de inclinação da reta (**coeficiente de regressão**);



5. Regressão linear simples: o modelo matemático

Se o intercepto β_0 e a inclinação β_1 são conhecidos, podemos **PREDIZER** qualquer valor y_i para um dado valor em x_i.

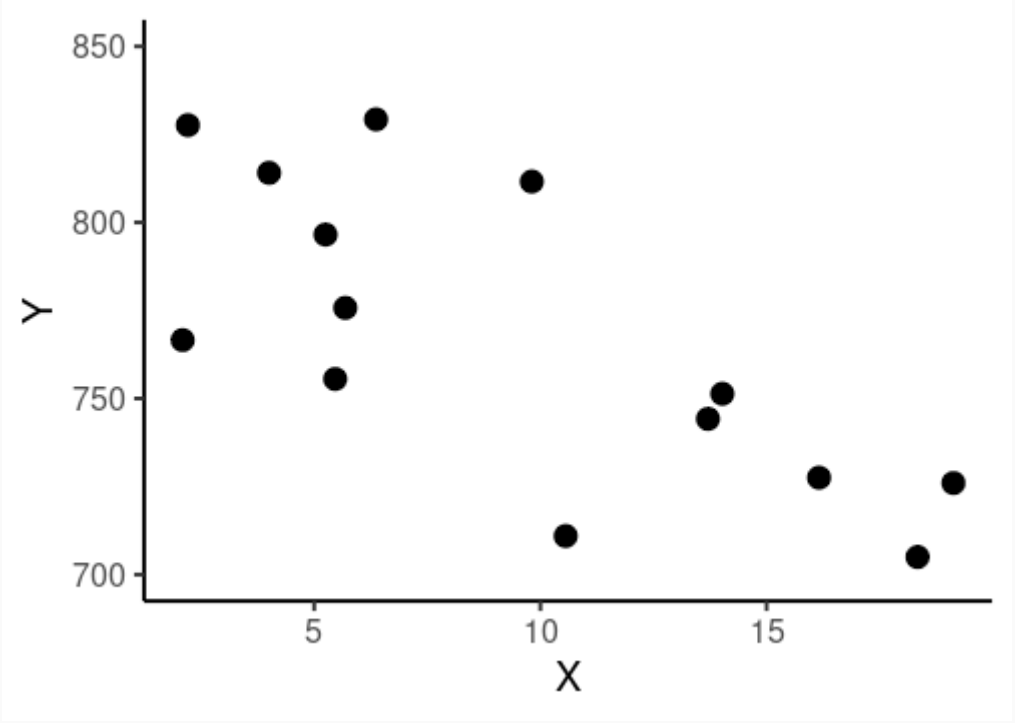
$$E(Y|x_i) = \beta_0 + \beta_1 x_i$$



5. Regressão linear simples: a tabela e o gráfico de dispersão

$$y_i|x_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

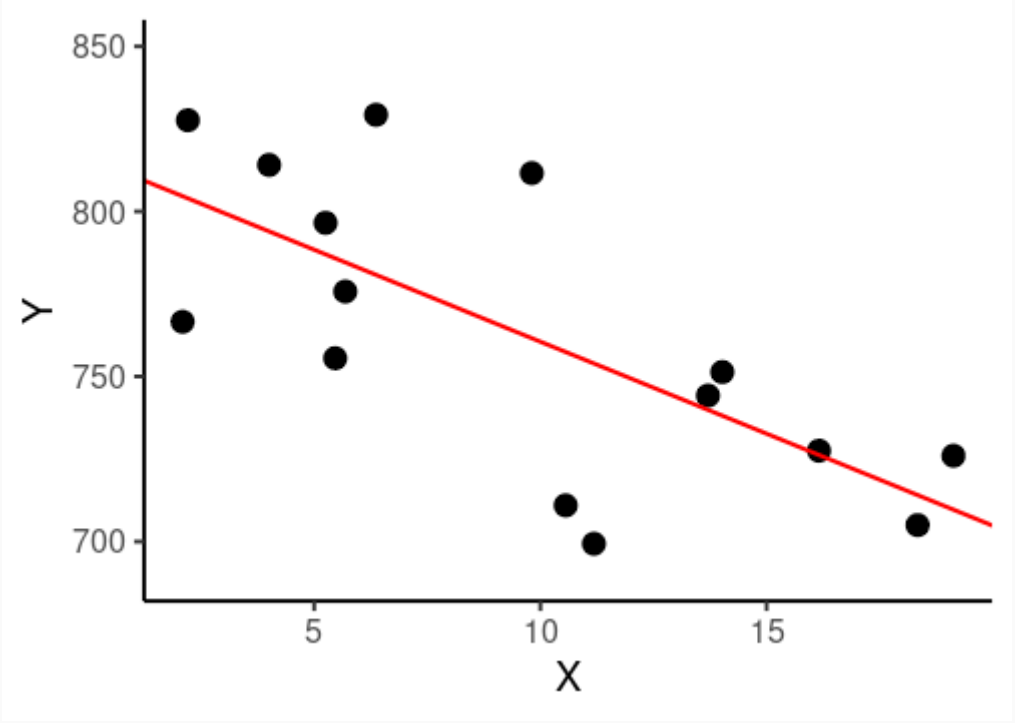
x_i	y_i
2.09	766.58
2.21	827.62
4.00	814.09
5.25	796.54
5.47	755.55
5.69	775.74
6.37	829.26
9.81	811.61
10.56	710.96
11.18	699.34
13.70	744.22
14.02	751.35
16.16	727.52
18.34	704.99
19.13	726.00



5. Regressão linear simples: o modelo estatístico

$$y_i|x_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

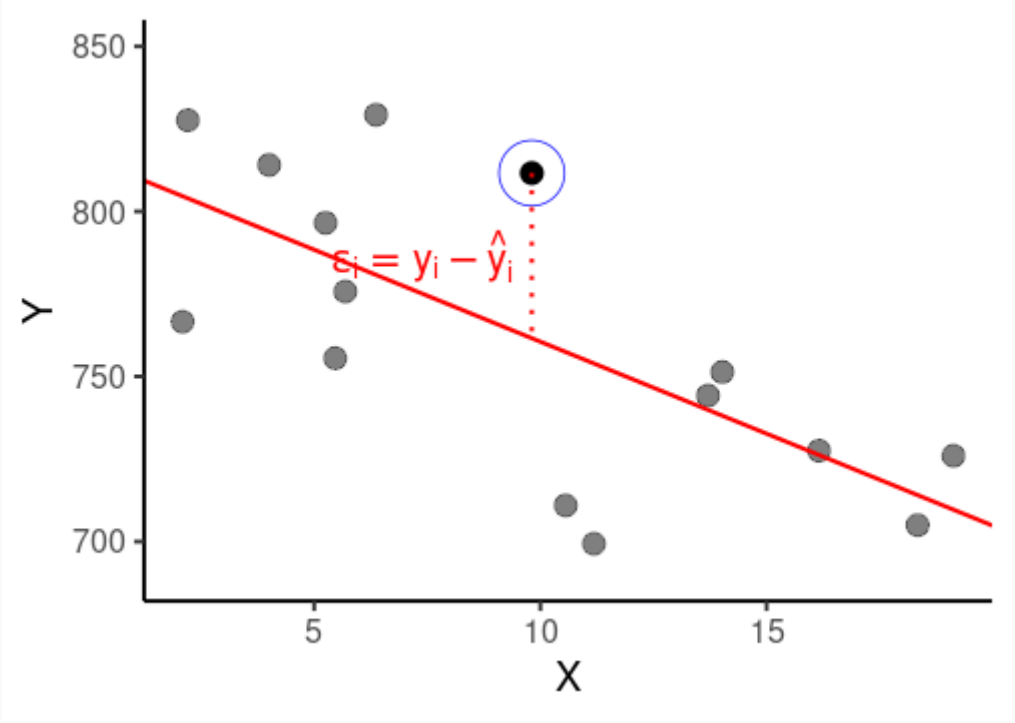
x_i	y_i	$\hat{y}_i$
2.09	766.58	804.59
2.21	827.62	803.94
4.00	814.09	793.94
5.25	796.54	786.98
5.47	755.55	785.79
5.69	775.74	784.55
6.37	829.26	780.76
9.81	811.61	761.58
10.56	710.96	757.41
11.18	699.34	753.93
13.70	744.22	739.88
14.02	751.35	738.11
16.16	727.52	726.20
18.34	704.99	714.06
19.13	726.00	709.64



5. Regressão linear simples: o modelo estatístico

$$y_i|x_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

x_i	y_i	$\hat{y}_i$	ϵ_i
2.09	766.58	804.59	-38.01
2.21	827.62	803.94	23.68
4.00	814.09	793.94	20.15
5.25	796.54	786.98	9.56
5.47	755.55	785.79	-30.25
5.69	775.74	784.55	-8.82
6.37	829.26	780.76	48.50
9.81	**811.61**	**761.58**	**50.03**
10.56	710.96	757.41	-46.45
11.18	699.34	753.93	-54.59
13.70	744.22	739.88	4.34
14.02	751.35	738.11	13.24
16.16	727.52	726.20	1.32
18.34	704.99	714.06	-9.06
19.13	726.00	709.64	16.36

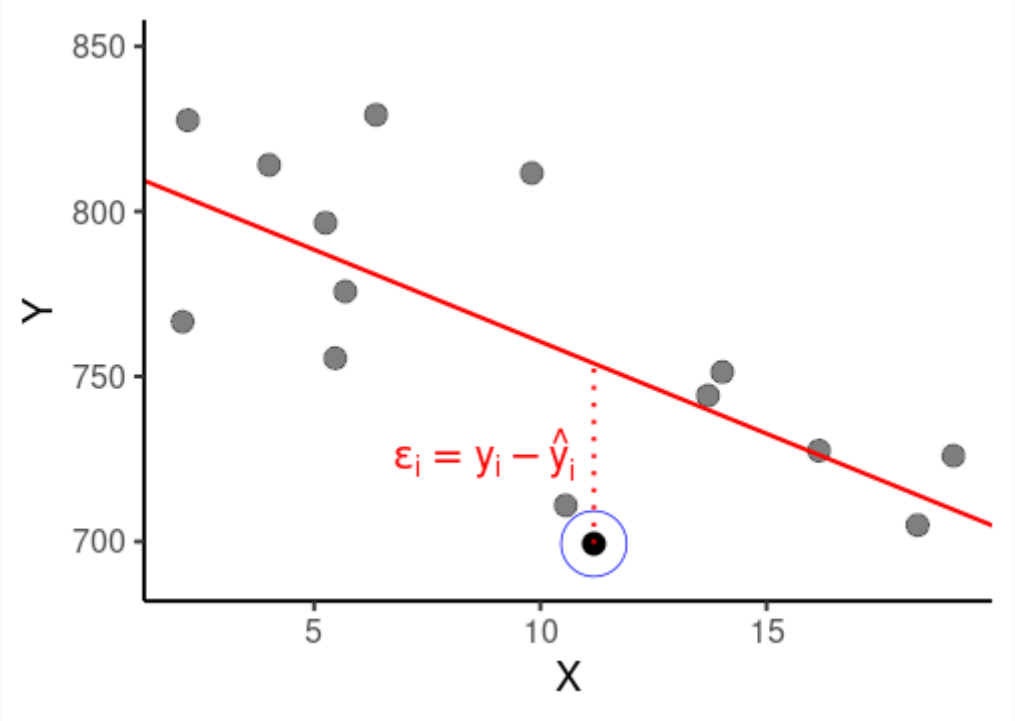


ε_i: resíduo - responsável pela variação de y_i em torno do valor **predito** $(\hat{y}_i)$ pela reta de regressão.

5. Regressão linear simples: o modelo estatístico

$$y_i|x_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

x_i	y_i	$\hat{y}_i$	ϵ_i
2.09	766.58	804.59	-38.01
2.21	827.62	803.94	23.68
4.00	814.09	793.94	20.15
5.25	796.54	786.98	9.56
5.47	755.55	785.79	-30.25
5.69	775.74	784.55	-8.82
6.37	829.26	780.76	48.50
9.81	811.61	761.58	50.03
10.56	710.96	757.41	-46.45
11.18	**699.34**	**753.93**	**-54.59**
13.70	744.22	739.88	4.34
14.02	751.35	738.11	13.24
16.16	727.52	726.20	1.32
18.34	704.99	714.06	-9.06
19.13	726.00	709.64	16.36



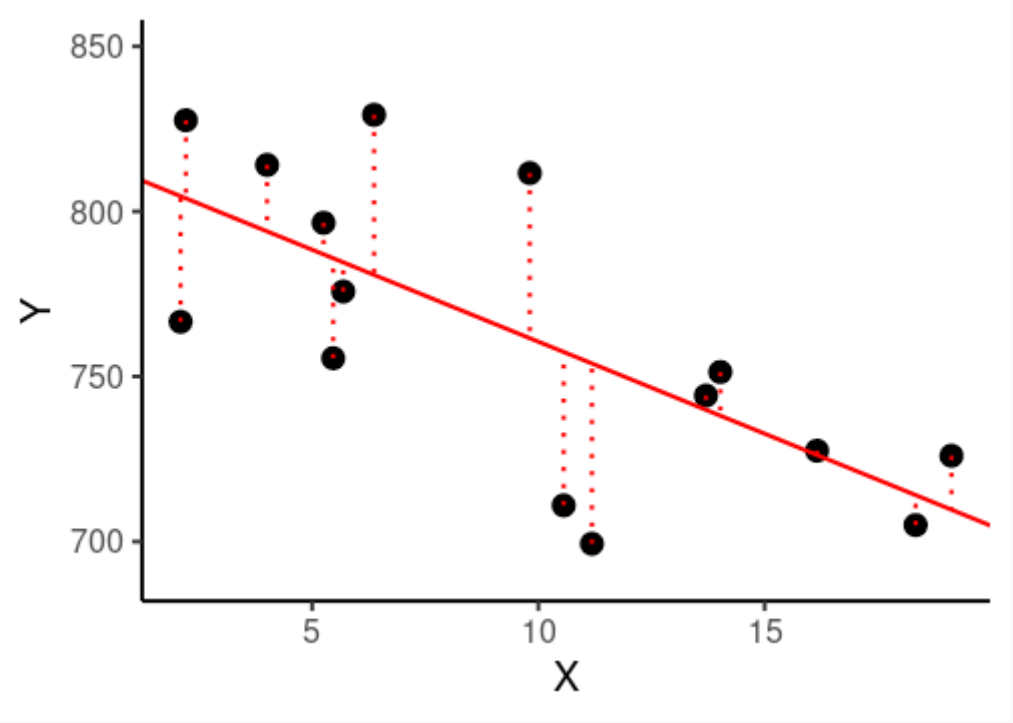
ε_i: resíduo - responsável pela variação de y_i em torno do valor **predito** $(\hat{y}_i)$ pela reta de regressão.

$$y_i|x_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

O resíduo associado a cada observação diminui ou aumenta à medida que o ponto está mais próximo ou distante da reta de regressão.

Assume-se que os resíduos têm distribuição Normal de probabilidades com média zero e variância σ^2.

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$



σ^2 elevada

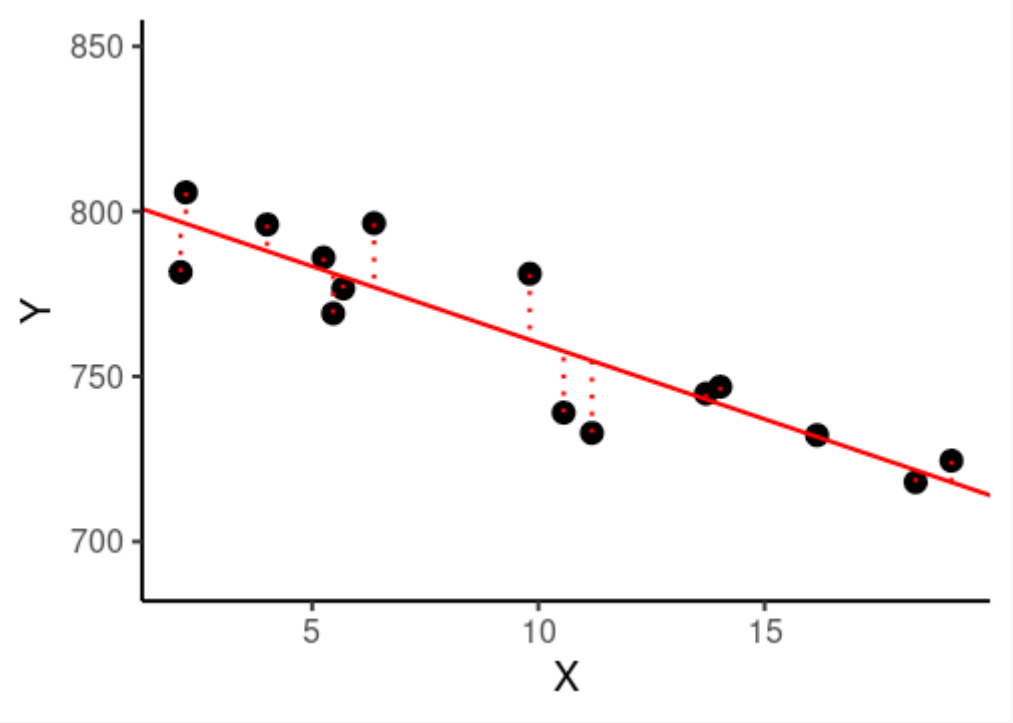
- pontos distantes da reta

$$y_i|x_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

O resíduo associado a cada observação diminui ou aumenta à medida que o ponto está mais próximo ou distante da reta de regressão.

Assume-se que os resíduos têm distribuição Normal de probabilidades com média zero e variância σ^2.

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$



σ^2 reduzida

- pontos próximos da reta

5. Regressão linear simples: o modelo estatístico

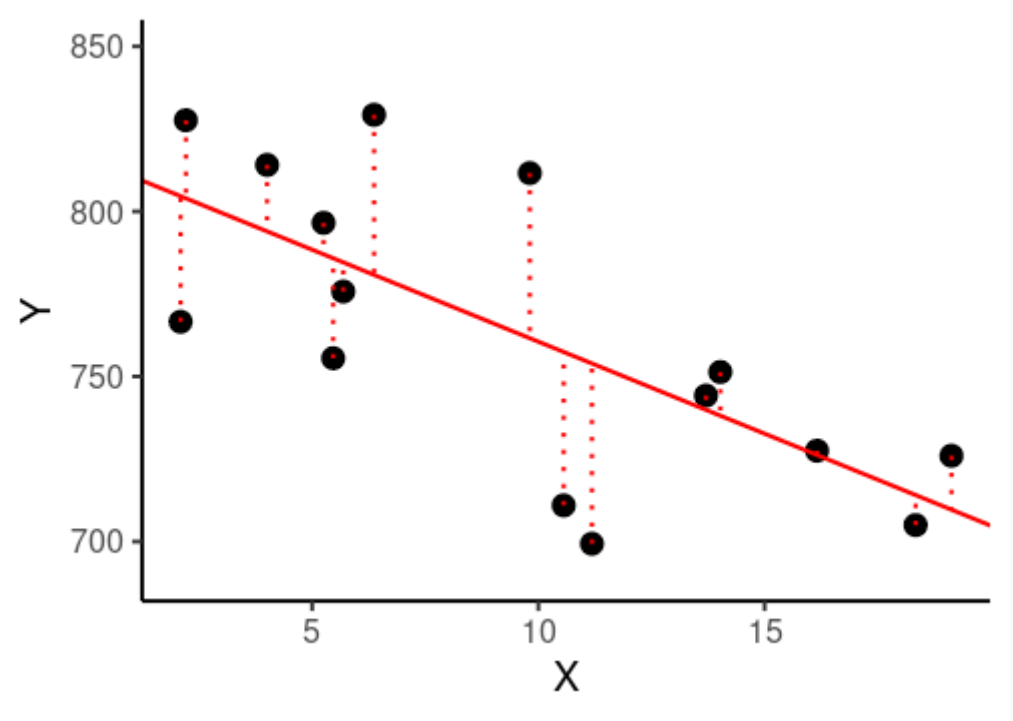
$$y_i|x_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Variáveis e quantias envolvidas

- y_i: variável resposta - i: $1 \cdots n$;
- x_i: variável preditora - i: $1 \cdots n$;
- n: tamanho da amostra;

Parâmetros do modelo

- β_0: intercepto;
- β_1: coeficiente inclinação da reta;
- σ^2: variância do resíduo.



5. Regressão linear simples: o modelo estatístico

$$y_i|x_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Parte determinística: β_0 e β_1

$$\beta_0 + \beta_1 x_i$$

Parte estocástica: σ^2

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

5. Regressão linear simples: estimativa dos parâmetros

$$y_i|x_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

5. Regressão linear simples: estimativa dos parâmetros

$$y_i|x_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

5. Regressão linear simples: estimativa dos parâmetros

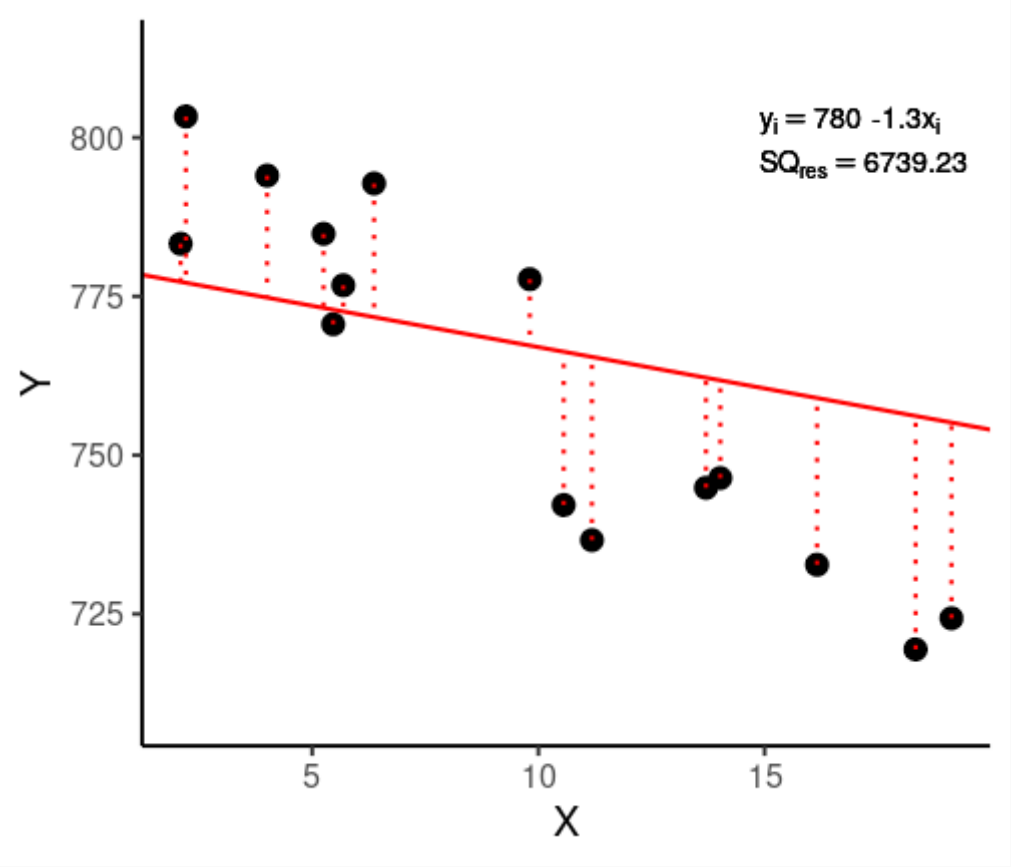
$$y_i|x_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

5. Regressão linear simples: estimativa dos parâmetros

$$y_i|x_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

5. Regressão linear simples: estimativa dos parâmetros

O Método dos Mínimos Quadrados



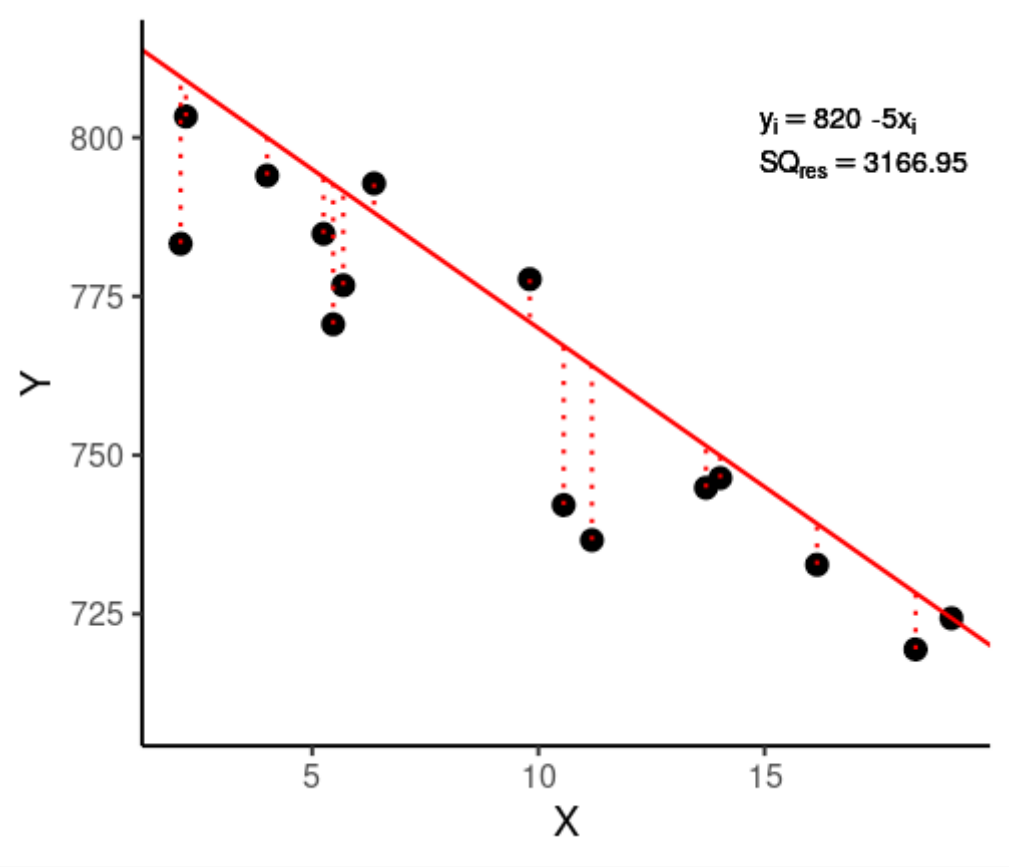
Soma dos quadrados dos resíduos (SQ_{Res})

$$SQ_{Res} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

O método dos mínimos quadrados consiste em encontrar a reta que **MINIMIZA** o somatório dos quadrados dos resíduos.

5. Regressão linear simples: estimativa dos parâmetros

O Método dos Mínimos Quadrados



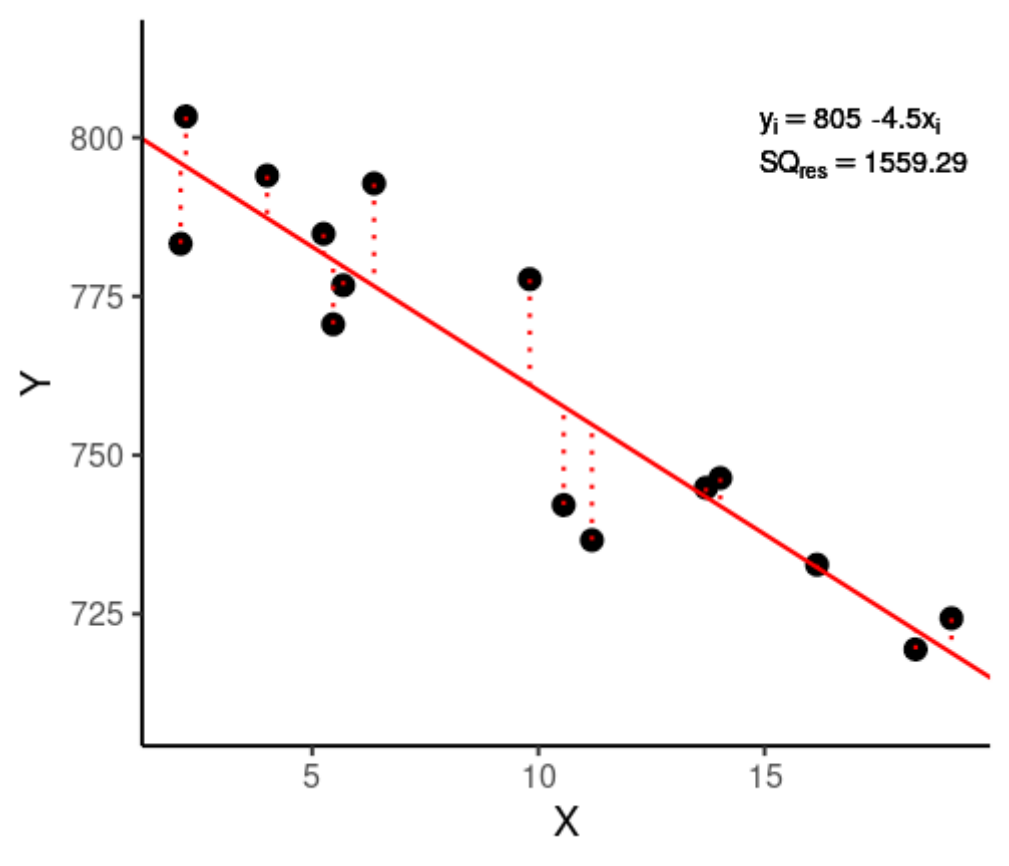
Soma dos quadrados dos resíduos (SQ_{Res})

$$SQ_{Res} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

O método dos mínimos quadrados consiste em encontrar a reta que **MINIMIZA** o somatório dos quadrados dos resíduos.

5. Regressão linear simples: estimativa dos parâmetros

O Método dos Mínimos Quadrados



Soma dos quadrados dos resíduos (SQ_{Res})

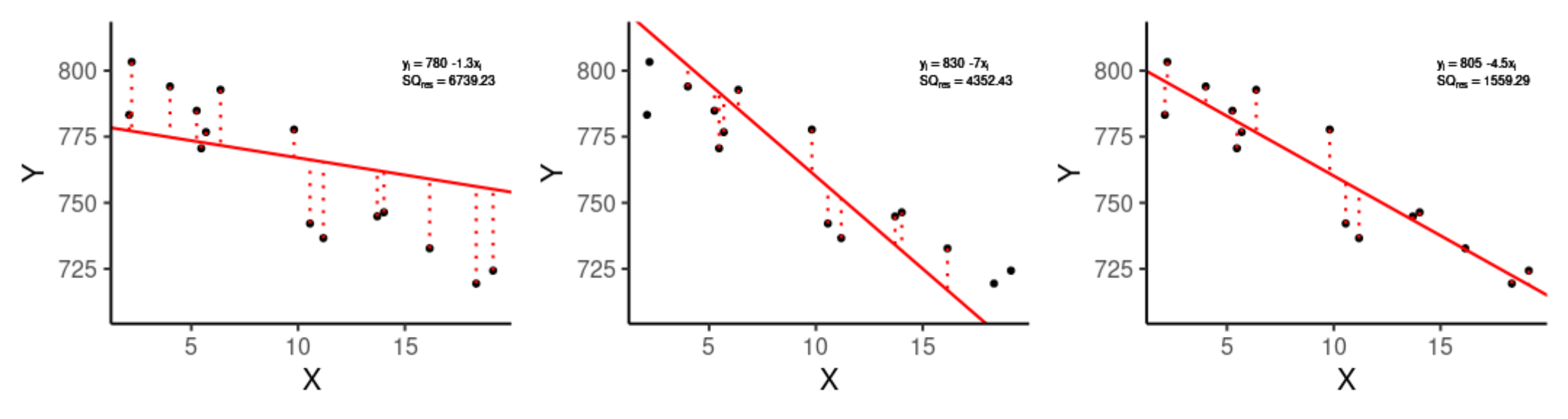
$$SQ_{Res} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

O método dos mínimos quadrados consiste em encontrar a reta que **MINIMIZA** o somatório dos quadrados dos resíduos.

5. Regressão linear simples: estimativa dos parâmetros

O Método dos Mínimos Quadrados

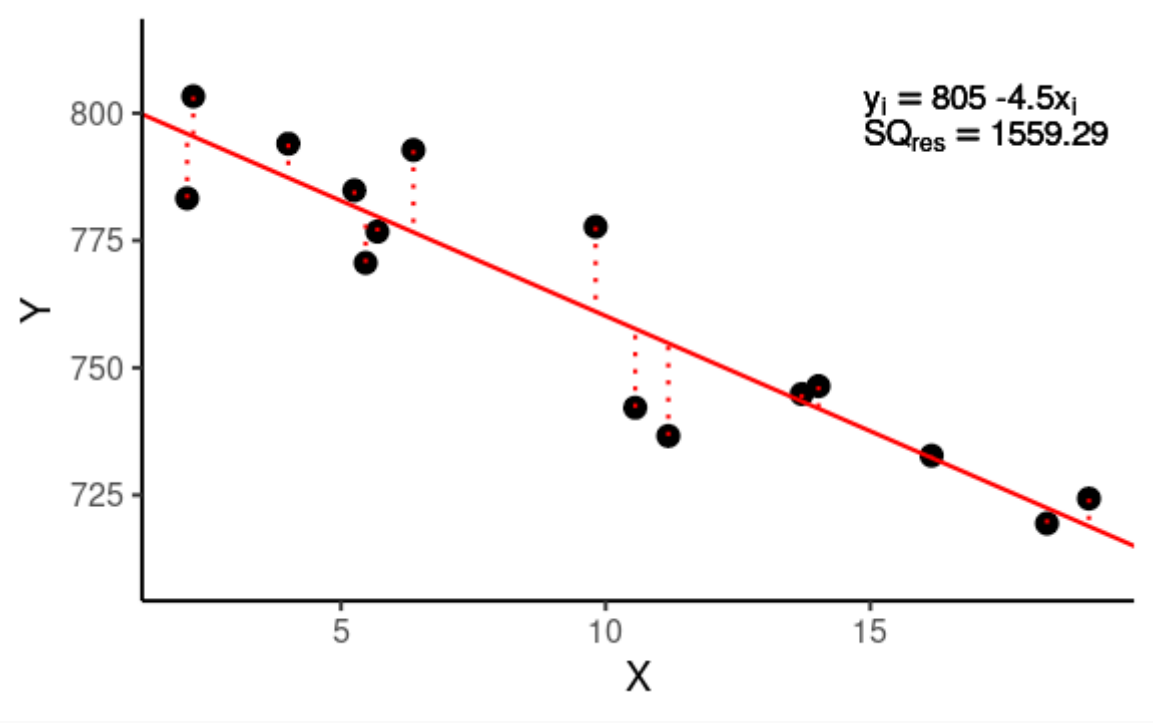
$$SQ_{Res} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



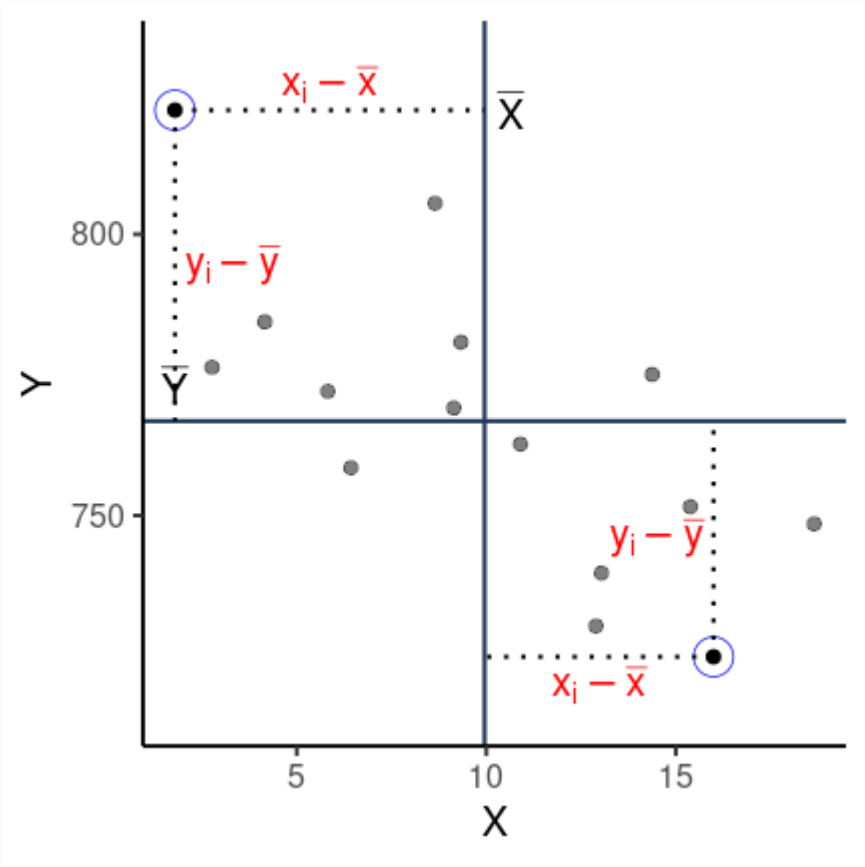
5. Regressão linear simples: estimativa dos parâmetros

---> *Estime* $\hat{\beta}_0$ e $\hat{\beta}_1$ que minimize a quantia:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$



5. Relembrando da Covariância entre Y e X



Soma dos produtos cruzados de Y e X

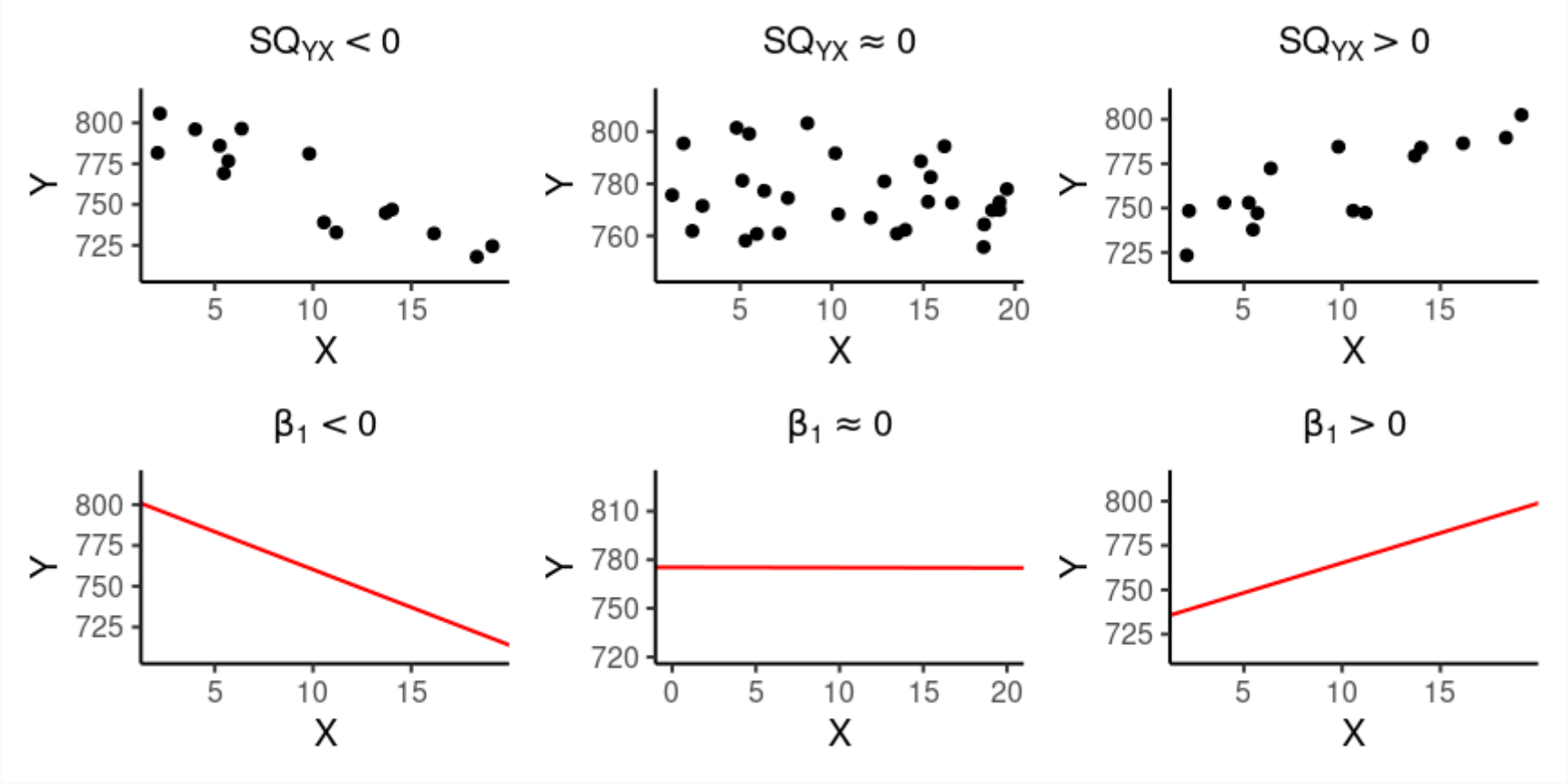
$$SQ_{YX} = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$$

Covariância amostral entre Y e X

$$s_{YX} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1}$$

5. Regressão linear simples: estimando β_1

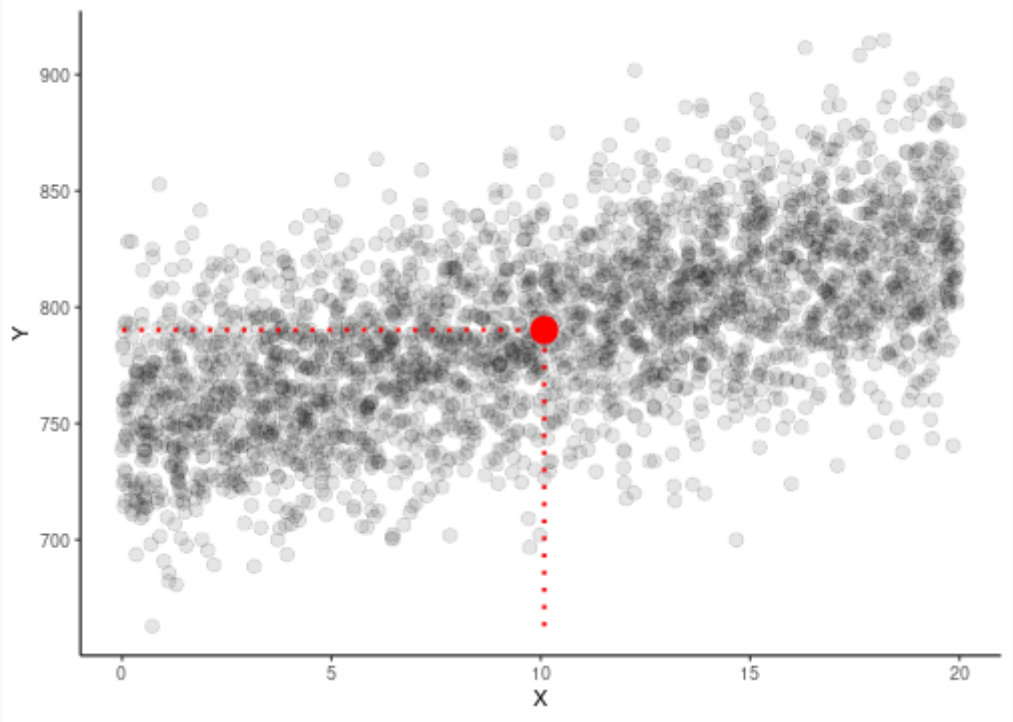
$$\hat{\beta}_1 = \frac{SQ_{YX}}{SQ_X} = \frac{s_{XY}}{s_X^2}$$



5. Regressão linear simples: estimando β_0

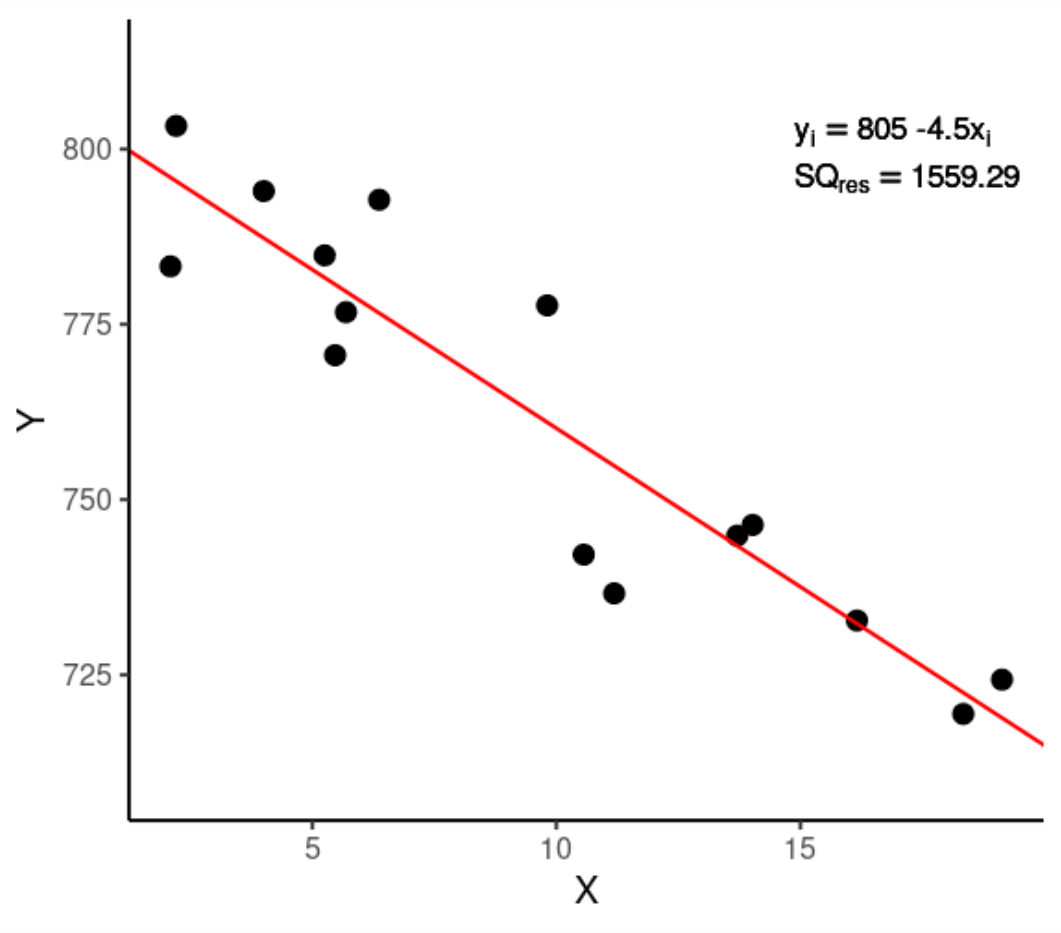
$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



5. Regressão linear simples: estimando σ^2

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i$$



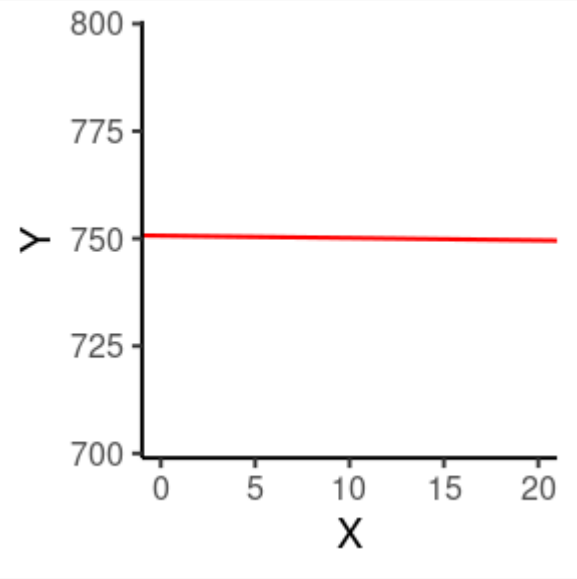
O Quadrado Médio do Resíduo (QM_{Res})

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

7. Teste de hipóteses

Hipótese nula

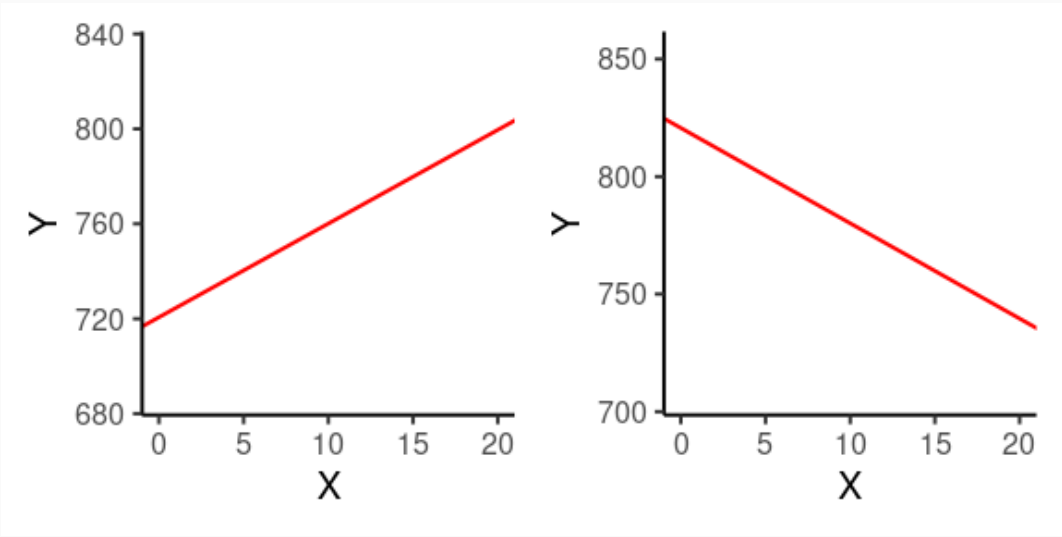
$$H_0 : \beta_1 = 0$$



$$y_i = \beta_0 + \varepsilon_i$$

Hipótese alternativa

$$H_1 : \beta_1 \neq 0$$

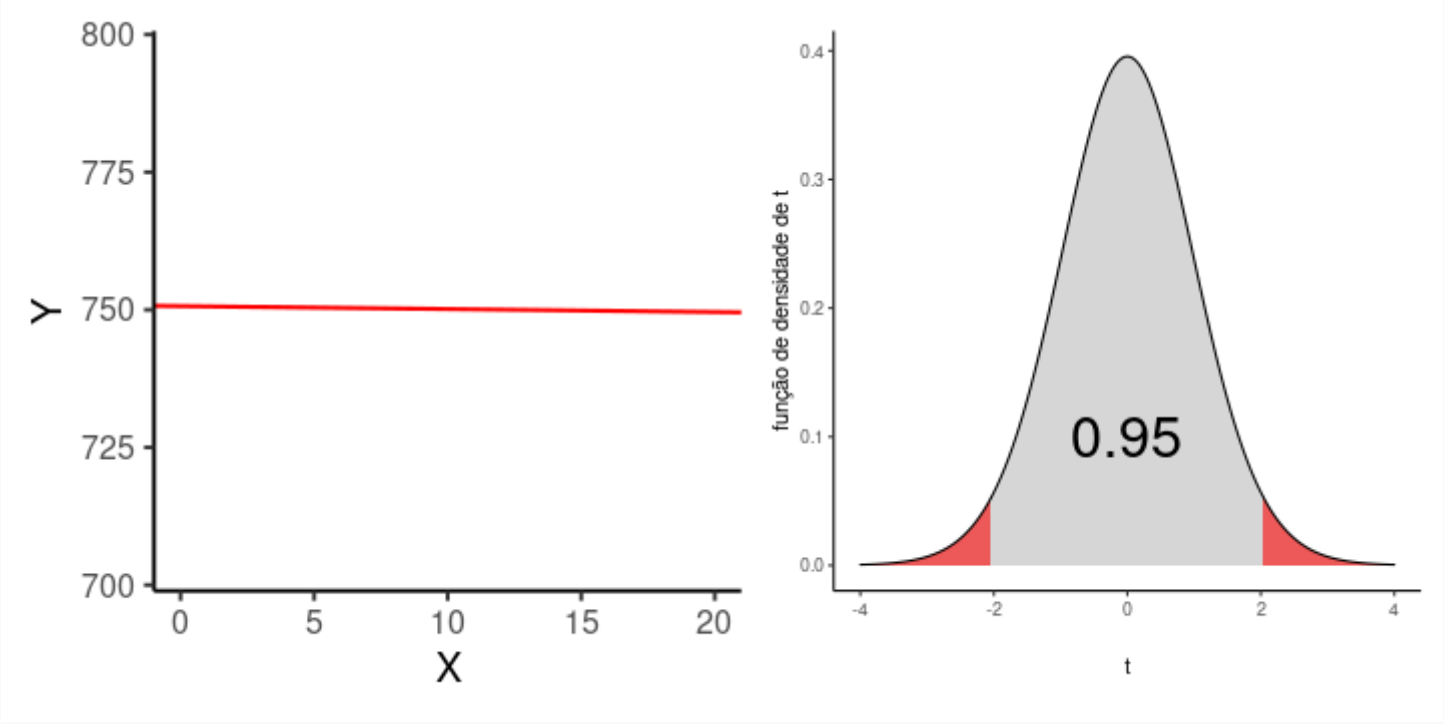


$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

7. Teste de hipóteses

H_0 pode ser testada por meio do teste t para o estimador $\hat{\beta}_1$

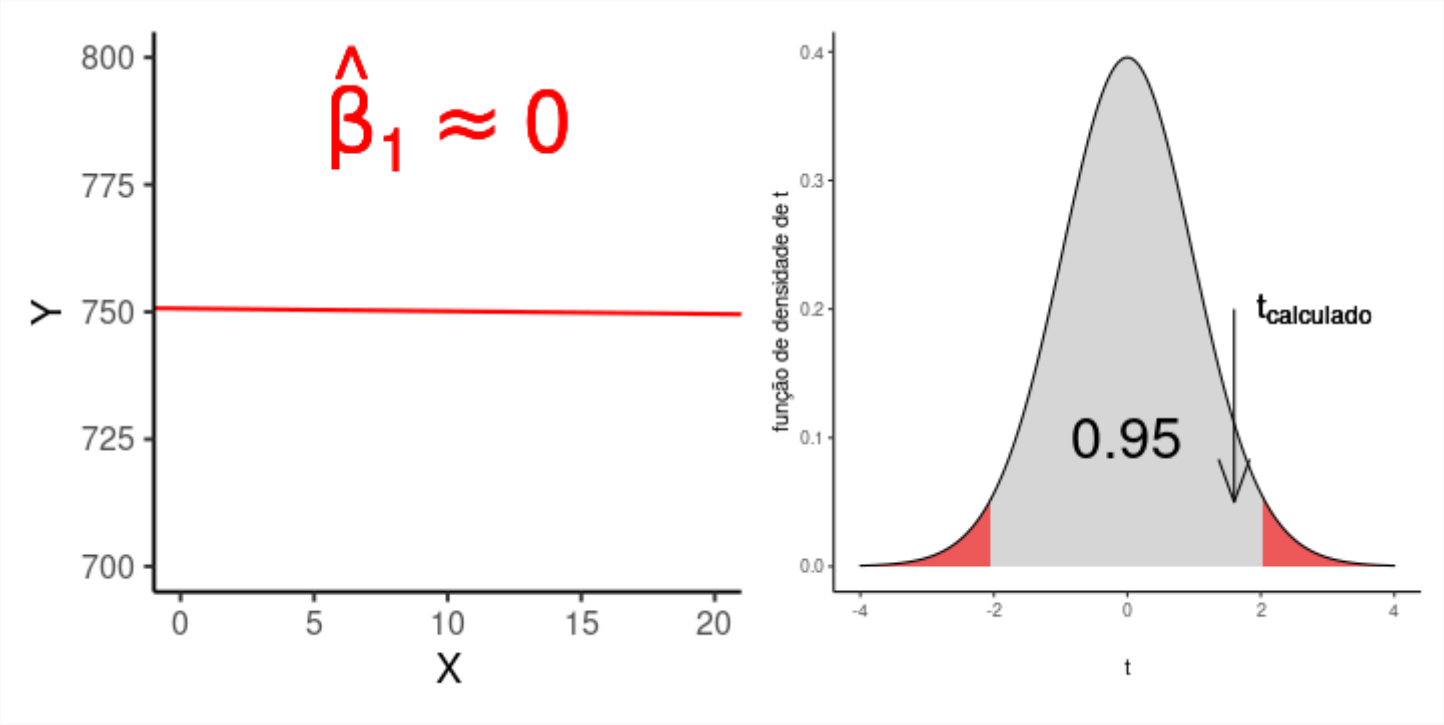
$$t_{calculado} = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}; \; s_{\hat{\beta}_1} = \sqrt{\frac{s^2}{SQ_X}}$$



7. Teste de hipóteses

$t_{calculado}$ depende da **magnitude de** $\hat{\beta}_1$

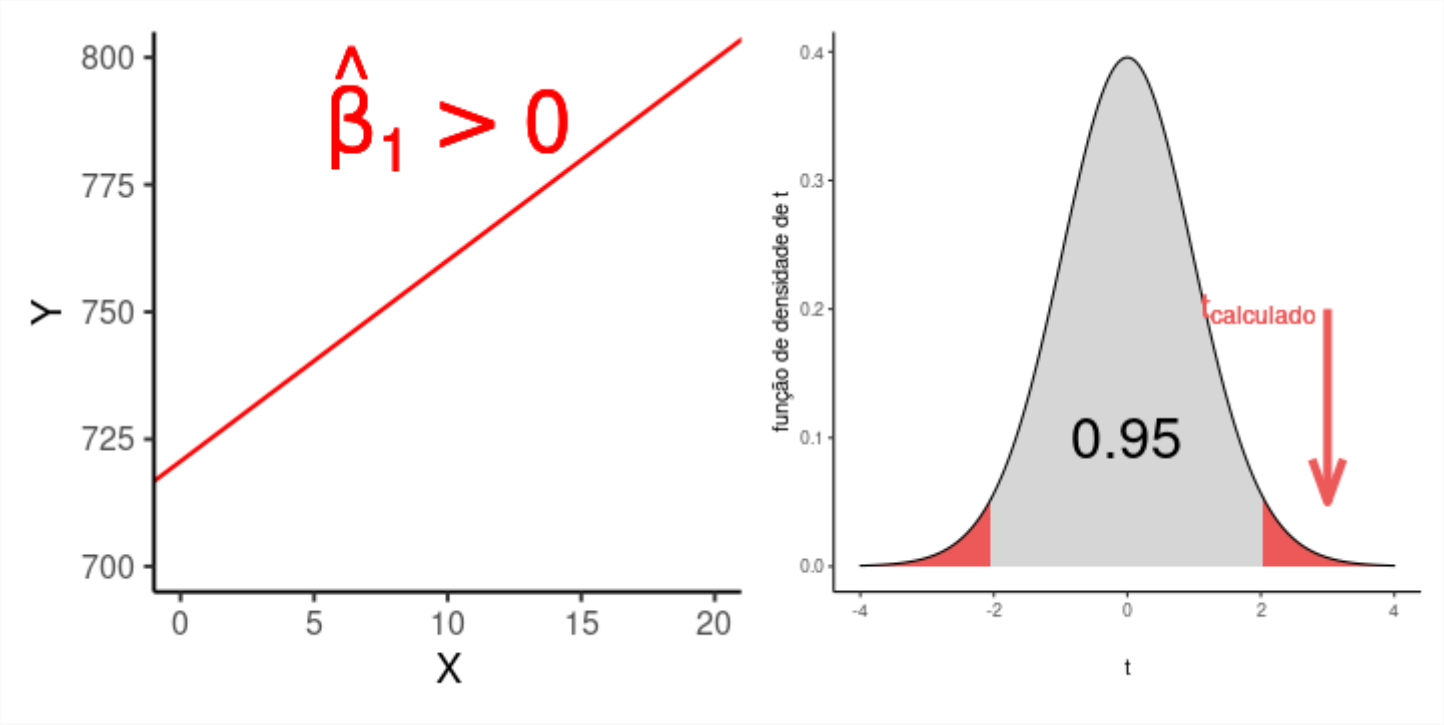
$$t_{calculado} = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$$



7. Teste de hipóteses

$t_{calculado}$ depende da **magnitude de** $\hat{\beta}_1$

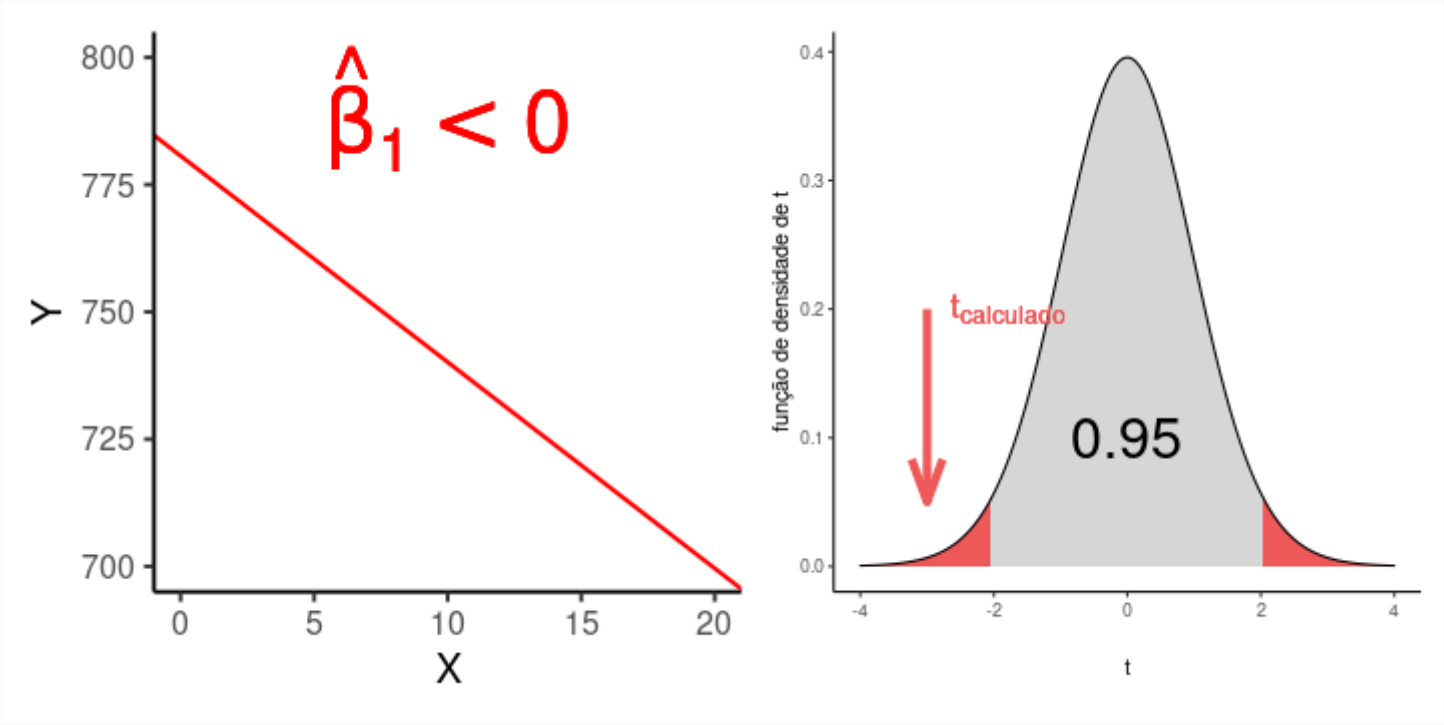
$$t_{calculado} = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$$



7. Teste de hipóteses

$t_{calculado}$ depende da **magnitude de** $\hat{\beta}_1$

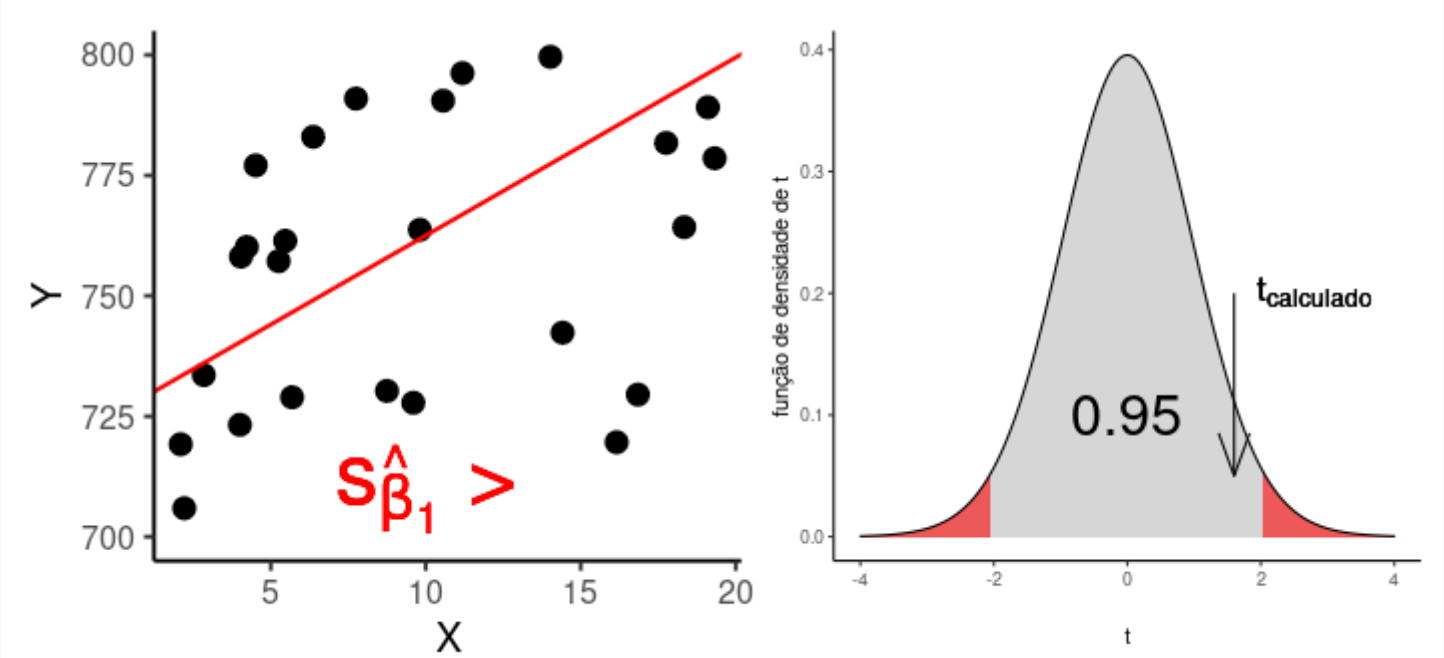
$$t_{calculado} = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$$



7. Teste de hipóteses

$t_{calculado}$ depende da **variância residual** - $s_{\hat{\beta}_1} = \sqrt{\frac{s^2}{SQ_X}}$

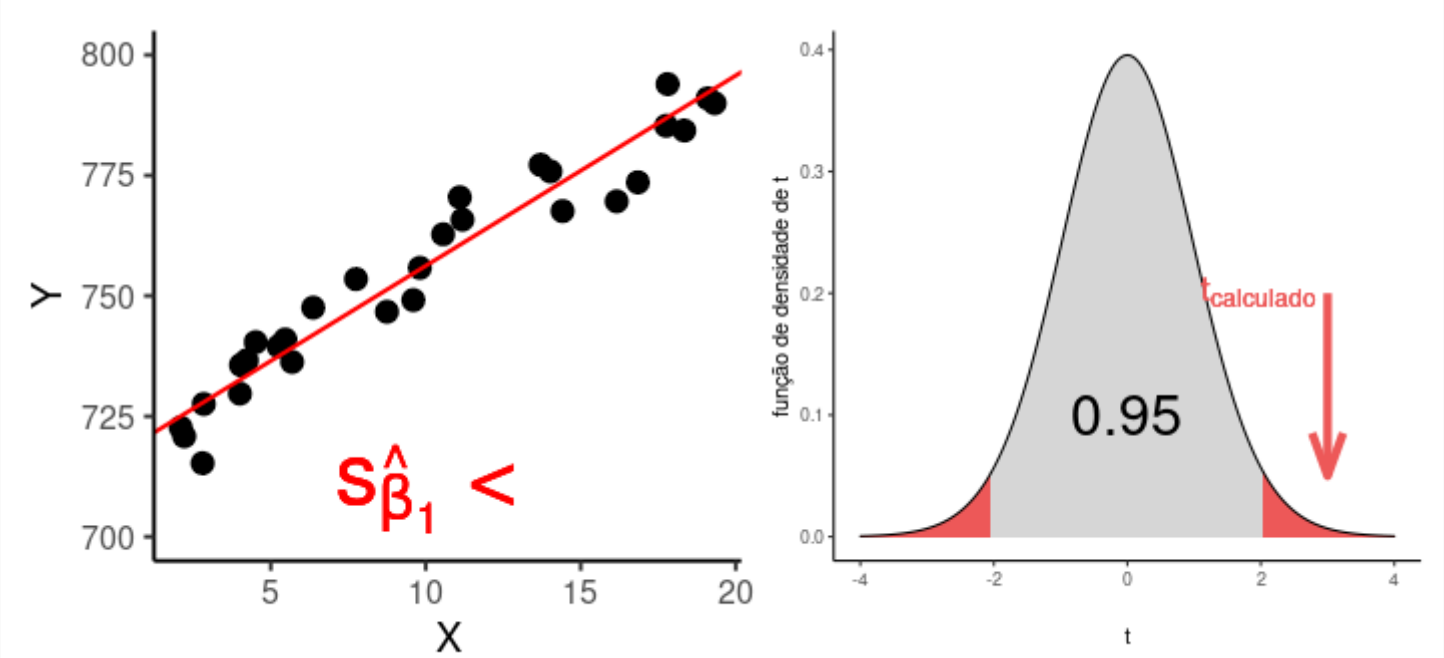
$$t_{calculado} = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$$



7. Teste de hipóteses

$t_{calculado}$ depende da **variância residual** - $s_{\hat{\beta}_1} = \sqrt{\frac{s^2}{SQ_X}}$

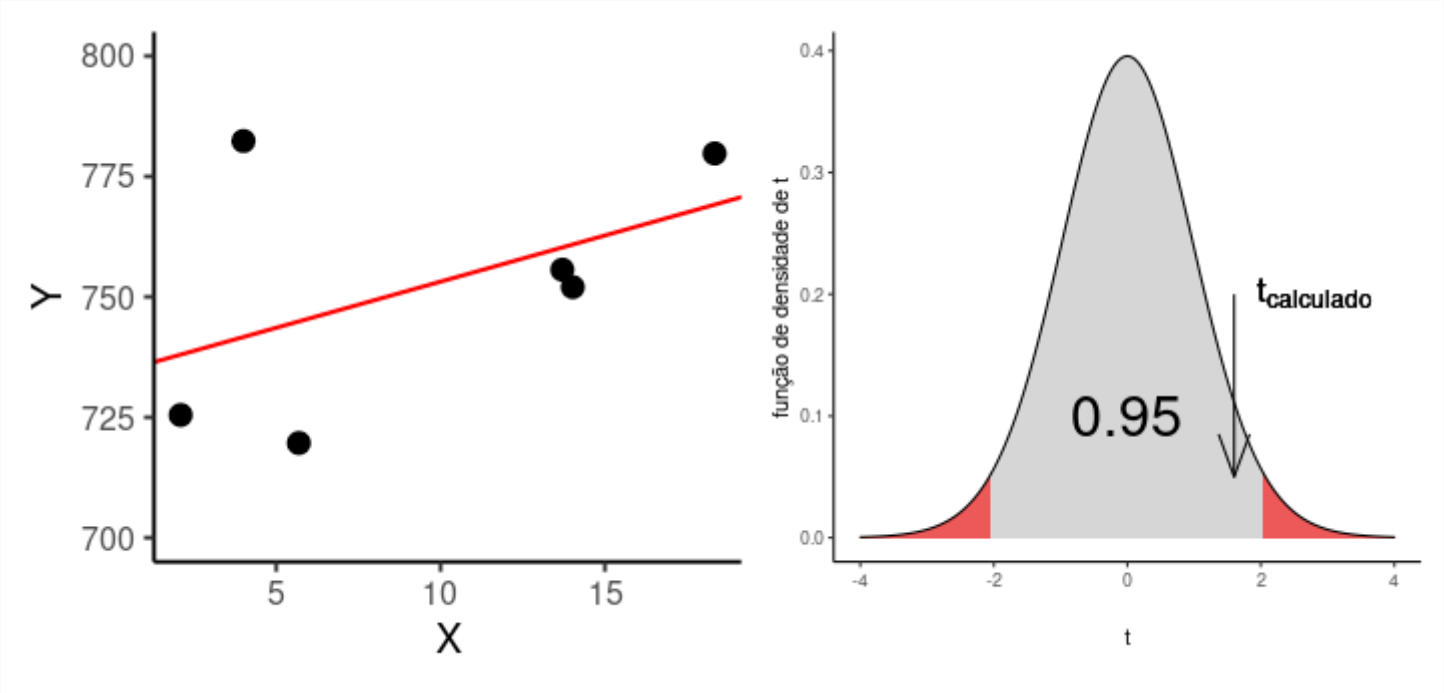
$$t_{calculado} = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$$



7. Teste de hipóteses

$t_{calculado}$ depende do **tamanho da amostra** - n

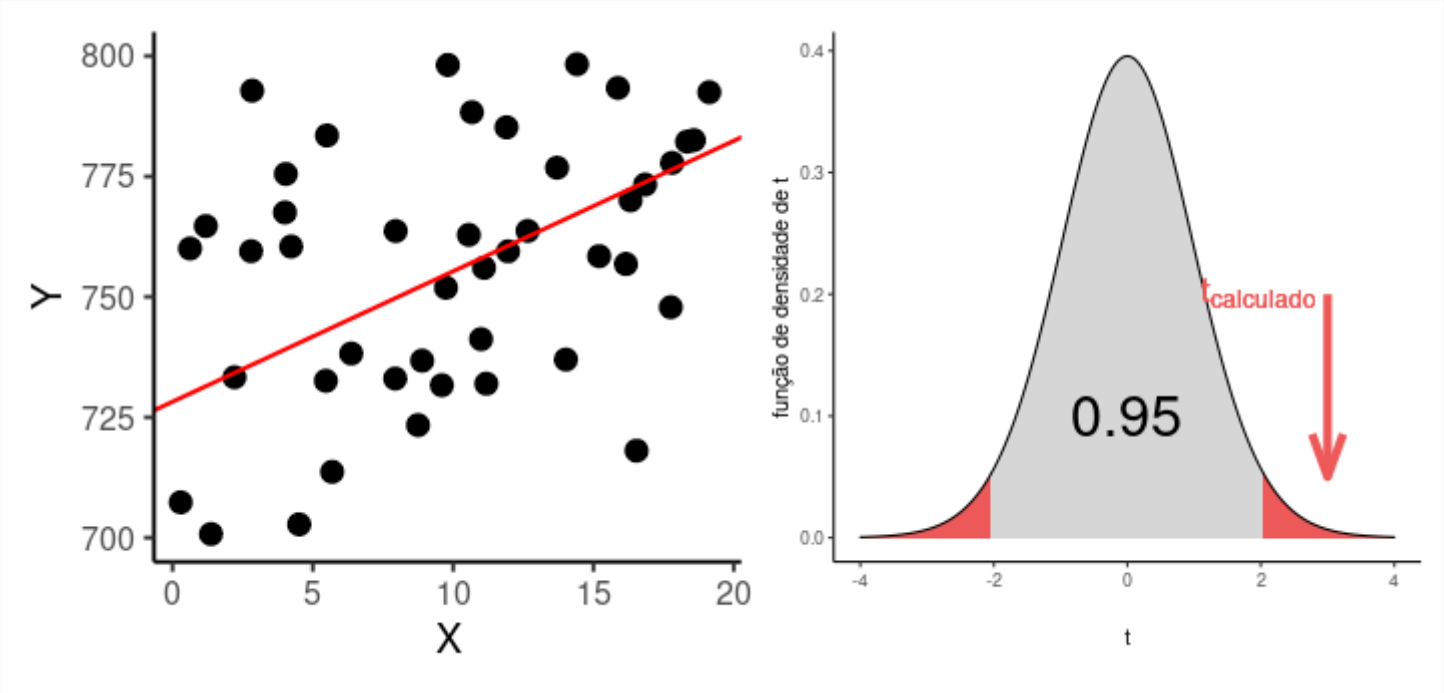
$$t_{calculado} = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$$



7. Teste de hipóteses

$t_{calculado}$ depende do **tamanho da amostra** - n

$$t_{calculado} = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$$



7. Teste de hipóteses

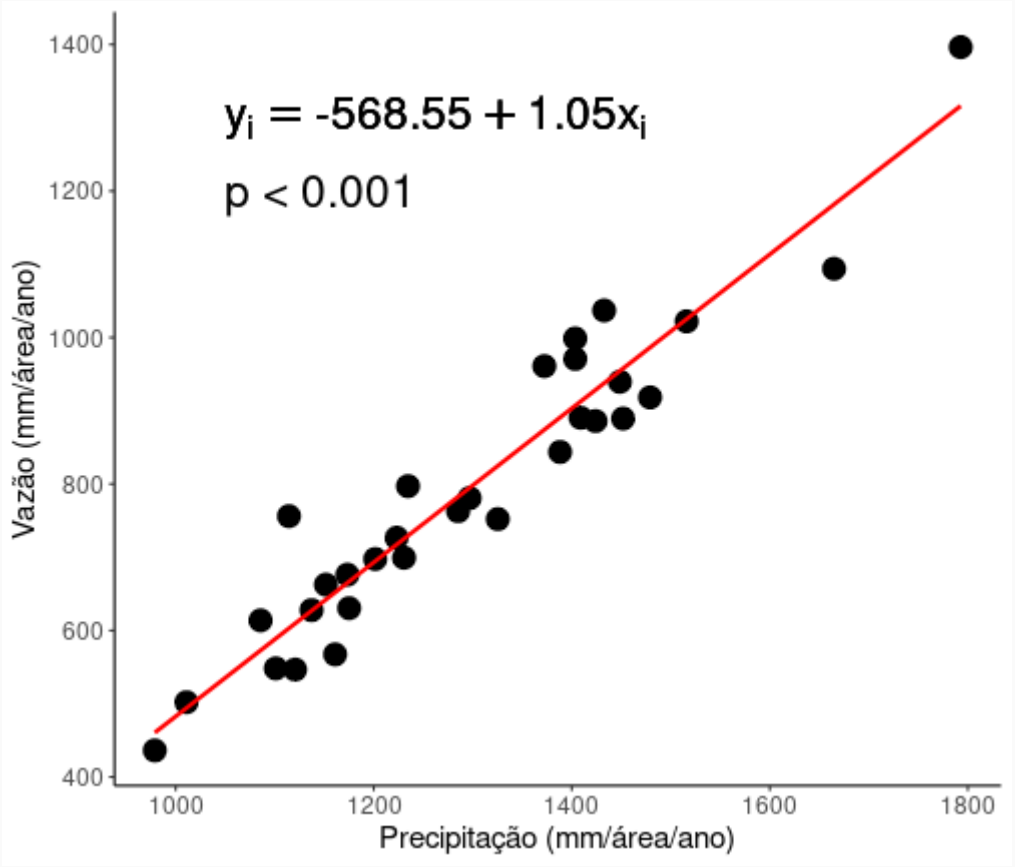
Na figura ao lado, os coeficientes de regressão foram estimados pelo MMQ em $\hat{\beta}_0 = -568.55$ e $\hat{\beta}_1 = 1.05$.

O valor de t foi:

$$t_{calculado} = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} = \frac{1.05 - 0}{0.06} = 17.451$$

O valor de $p < 0.001$ associado a este resultado, se interpretado ao nível de significância $\alpha = 0,05$, é dito estatísticamente significativo, o que nos leva a **rejeitar** H_0 .

A conclusão é de que **existe** uma relação crescente entre a Precipitação e a Vazão na bacia hidrográfica.



8. Intervalo de confiança para $\hat{Y}$

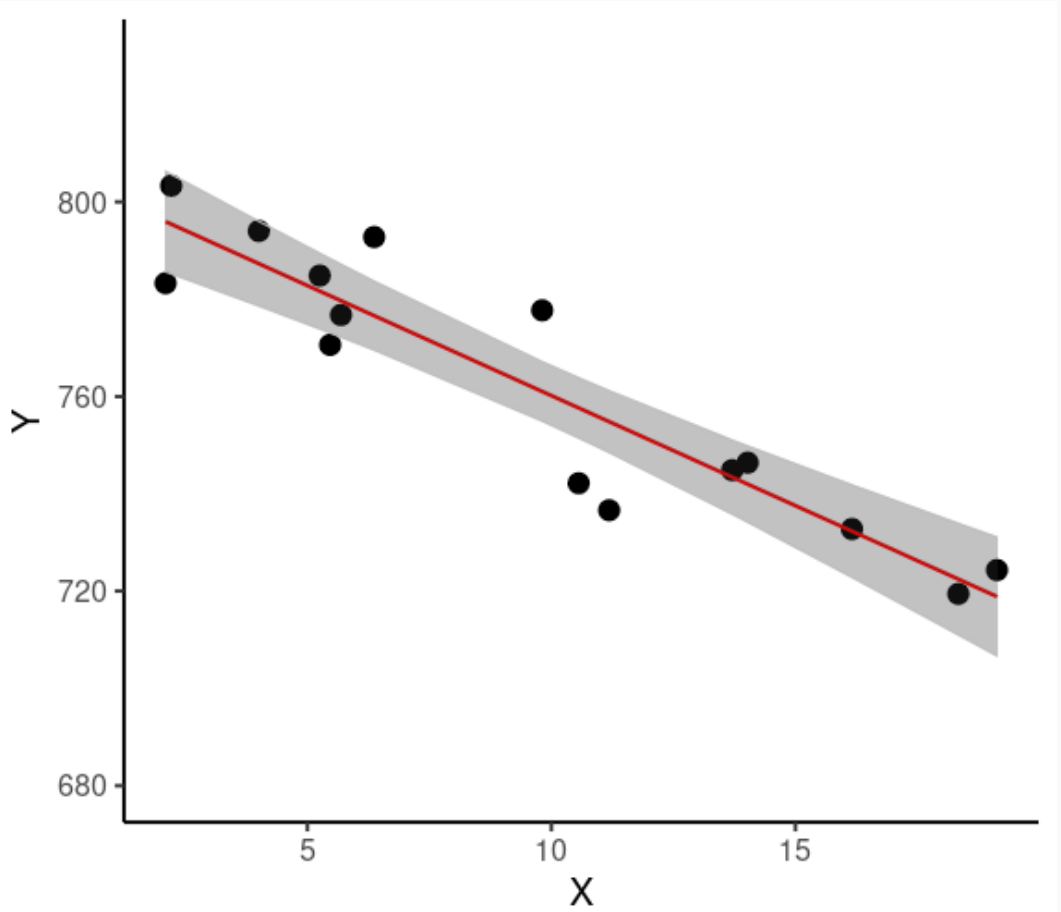
Cada repetição do experimento com amostra de tamanho n irá resultar em diferentes valores de $\hat{Y}$ e consequentemente diferentes **retas de regressão**. O erro padrão de $\hat{Y}$ é dado por:

$$s_{\hat{Y}|X} = \sqrt{s^2 \left(\frac{1}{n} + \frac{(X_i - \overline{X})^2}{SQ_X} \right)}$$

O intervalo de confiância de $\hat{Y}$ é dado por:

$$\hat{Y} \pm t_{(\alpha, n-2)} s_{\hat{Y}|X}$$

A confiança para $\hat{Y}$ aumenta ao redor de $\overline{X}$ e diminui nos extremos da distribuição de X_i



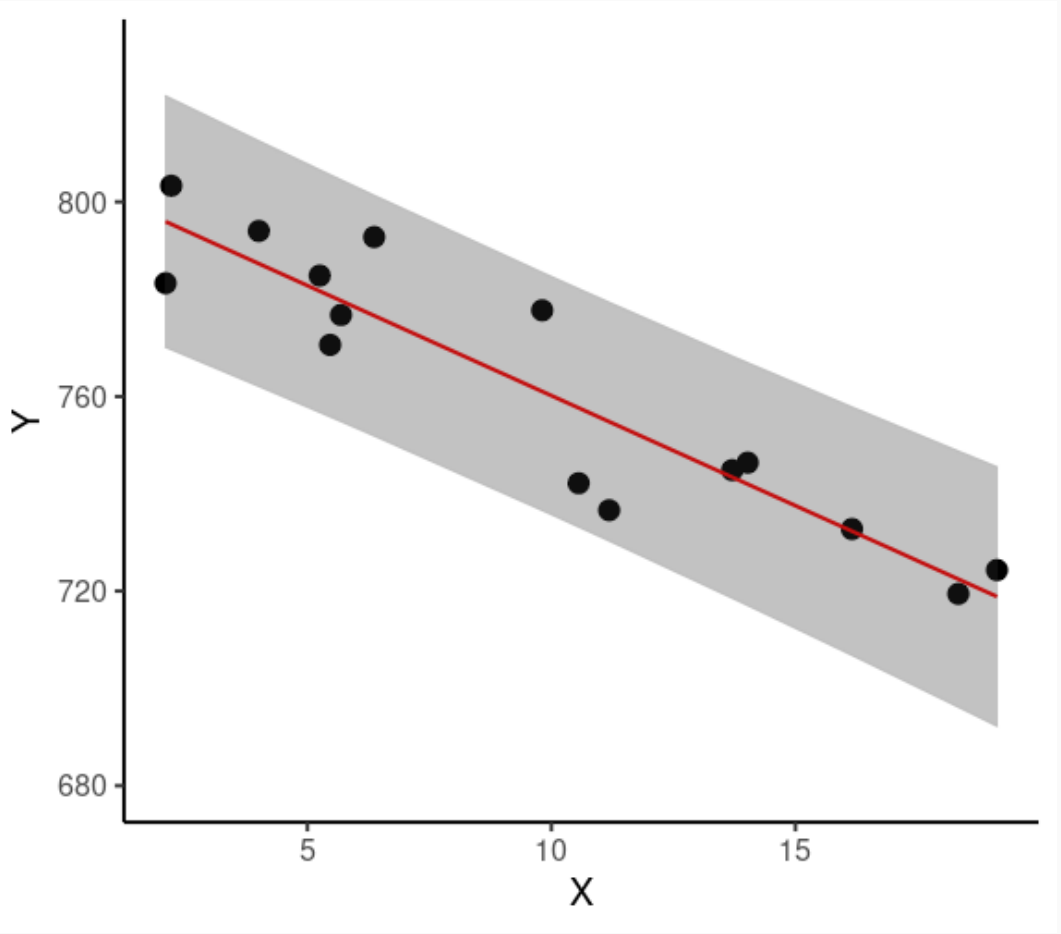
8. Intervalo de predição para Y^*

Tendo um modelo de regressão ajustado, o que esperar para Y se obtivermos **um novo dado** em X^*? O erro padrão de Y^* é dado por:

$$s_{Y^*|X^*} = \sqrt{s^2 \left(1 + \frac{1}{n} + \frac{(X^* - \overline{X})^2}{SQ_X}\right)}$$

O intervalo de **predição** de Y^* é dado por:

$$Y^* \pm t_{(\alpha, n-2)} s_{Y^*|X^*}$$

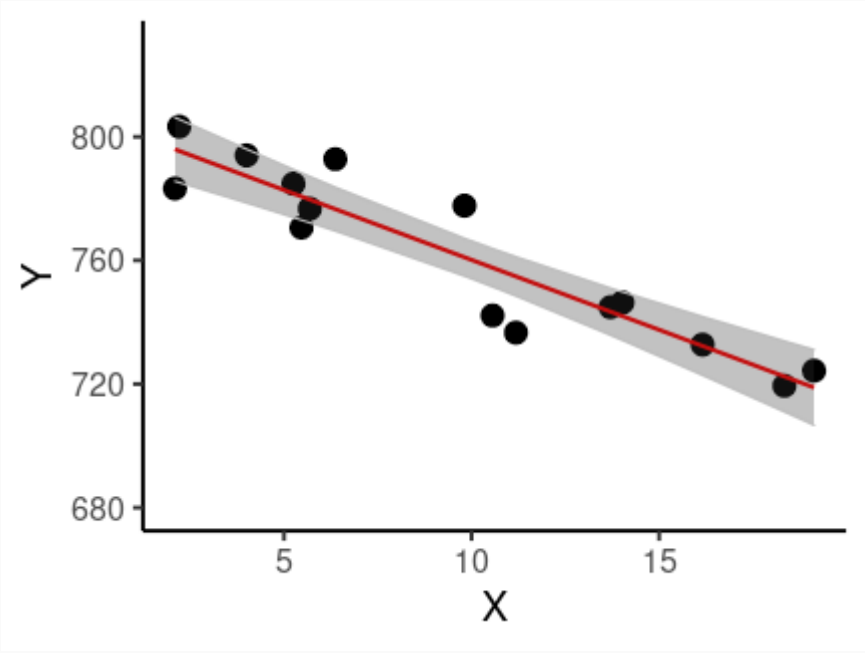


8. Intervalo de confiança *vs* intervalo de predição

Intervalo de confiança de $\hat{Y}$

$$s_{\hat{Y}|X} = \sqrt{s^2\left(\frac{1}{n} + \frac{(X_i - \overline{X})^2}{SQ_X}\right)}$$

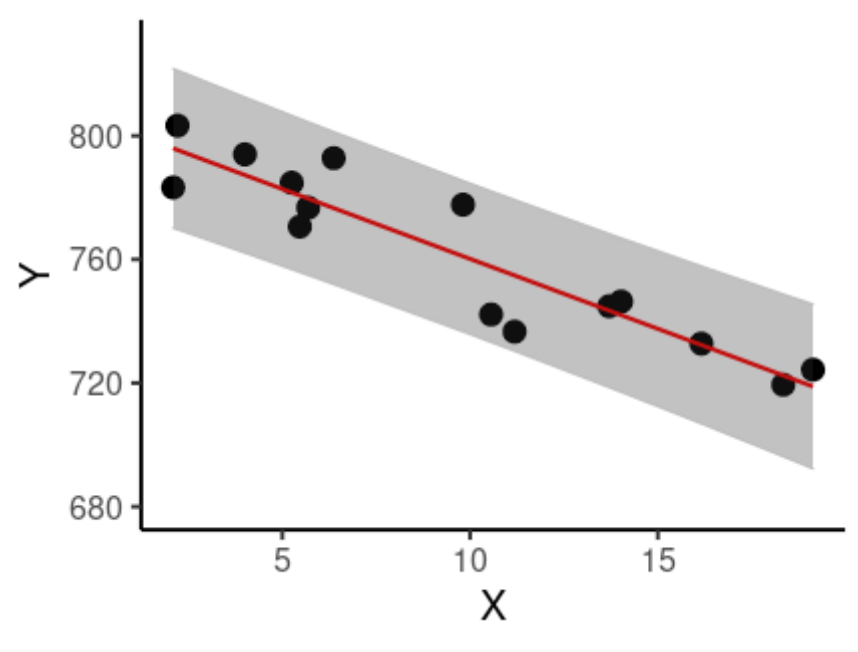
$$\hat{Y} \pm t_{(\alpha, n-2)} s_{\hat{Y}|X}$$



Intervalo de predição de Y_i^*

$$s_{Y^*|X^*} = \sqrt{s^2\left(1 + \frac{1}{n} + \frac{(X^* - \overline{X})^2}{SQ_X}\right)}$$

$$Y^* \pm t_{(\alpha, n-2)} s_{Y^*|X^*}$$



8. Intervalos de confiança para $\hat{\beta}_0$ e $\hat{\beta}_1$

Para $X_i = 0$, $\hat{Y} = \hat{\beta}_0$ de modo que:

$$s_{\hat{\beta}_0} = \sqrt{s^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{SQ_X} \right)}$$

O intervalo de confiância de $\hat{\beta}_0$ é dado por:

$$\hat{\beta}_0 \pm t_{(\alpha, n-2)} s_{\hat{\beta}_0}$$

Para $\hat{\beta}_1$ temos:

$$s_{\hat{\beta}_1} = \sqrt{\left(\frac{s^2}{SQ_X} \right)}$$

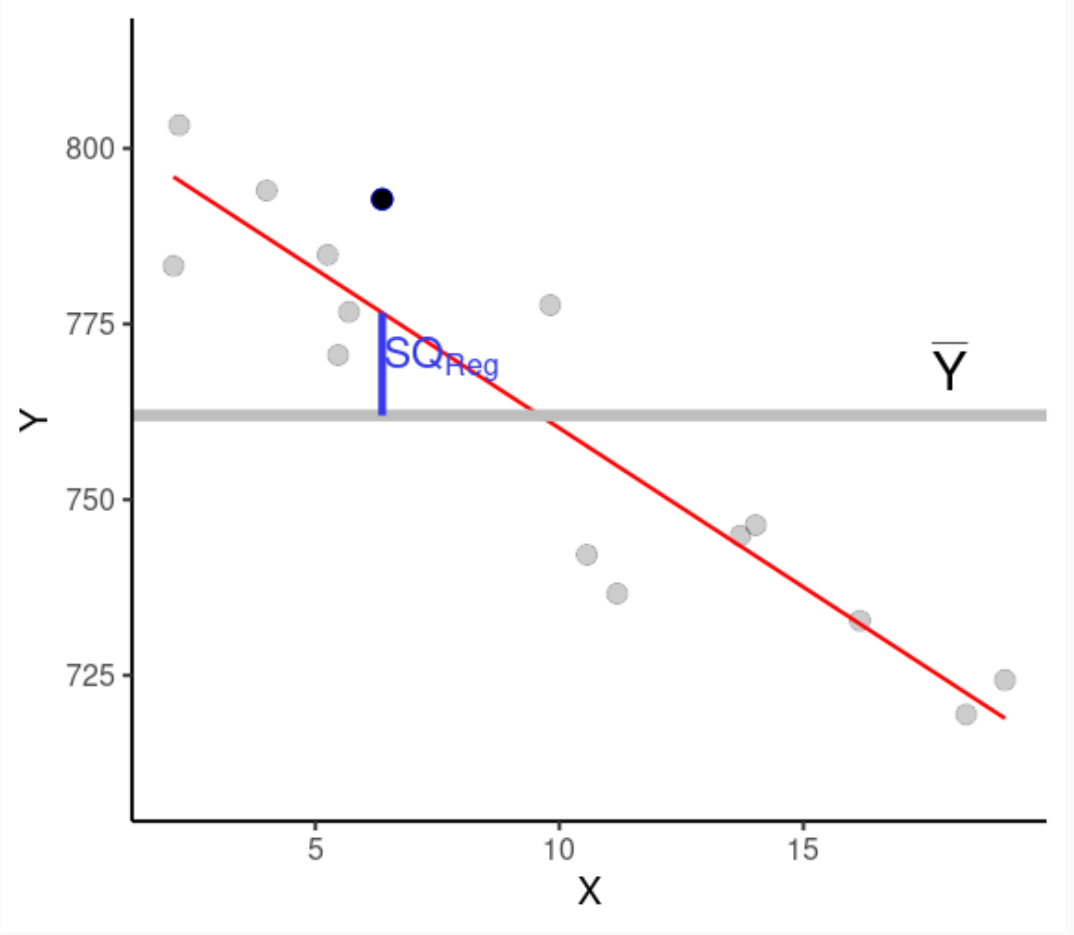
O intervalo de confiância de $\hat{\beta}_1$ é dado por:

$$\hat{\beta}_1 \pm t_{(\alpha, n-2)} s_{\hat{\beta}_1}$$

9. Partição da Soma dos Quadrados e variação explicada

Somatório dos Quadrados da Regressão - SQ_{Reg}

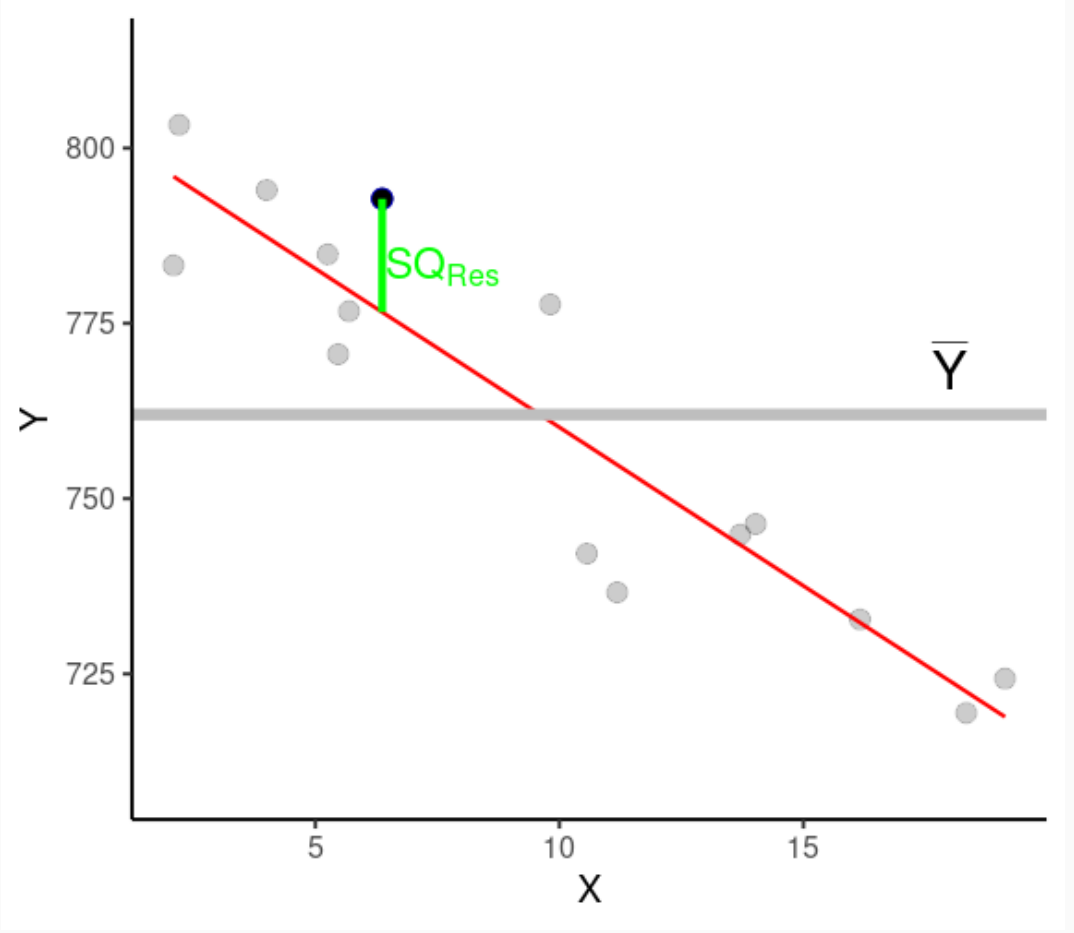
$$SQ_{Reg} = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$



9. Partição da Soma dos Quadrados e variação explicada

Somatório dos Quadrados do Resíduo - SQ_{Res}

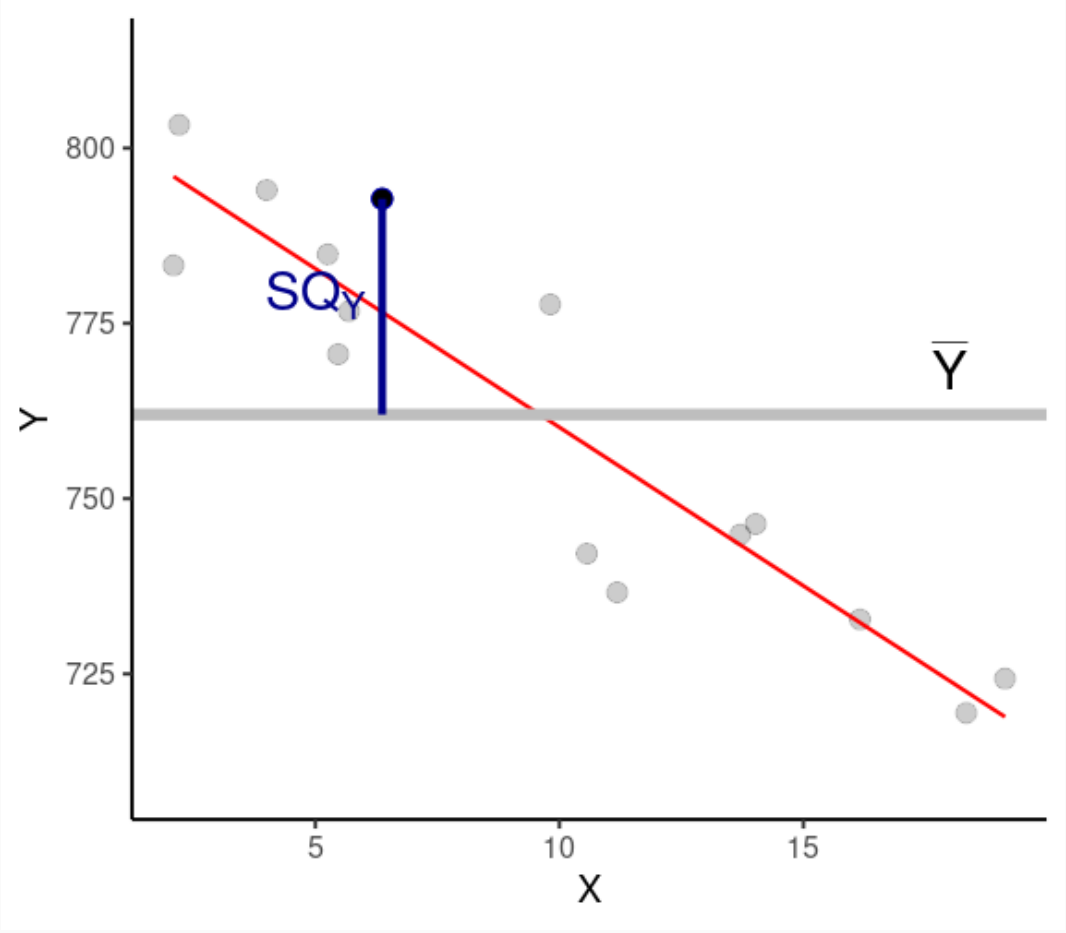
$$SQ_{Res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$



9. Partição da Soma dos Quadrados e variação explicada

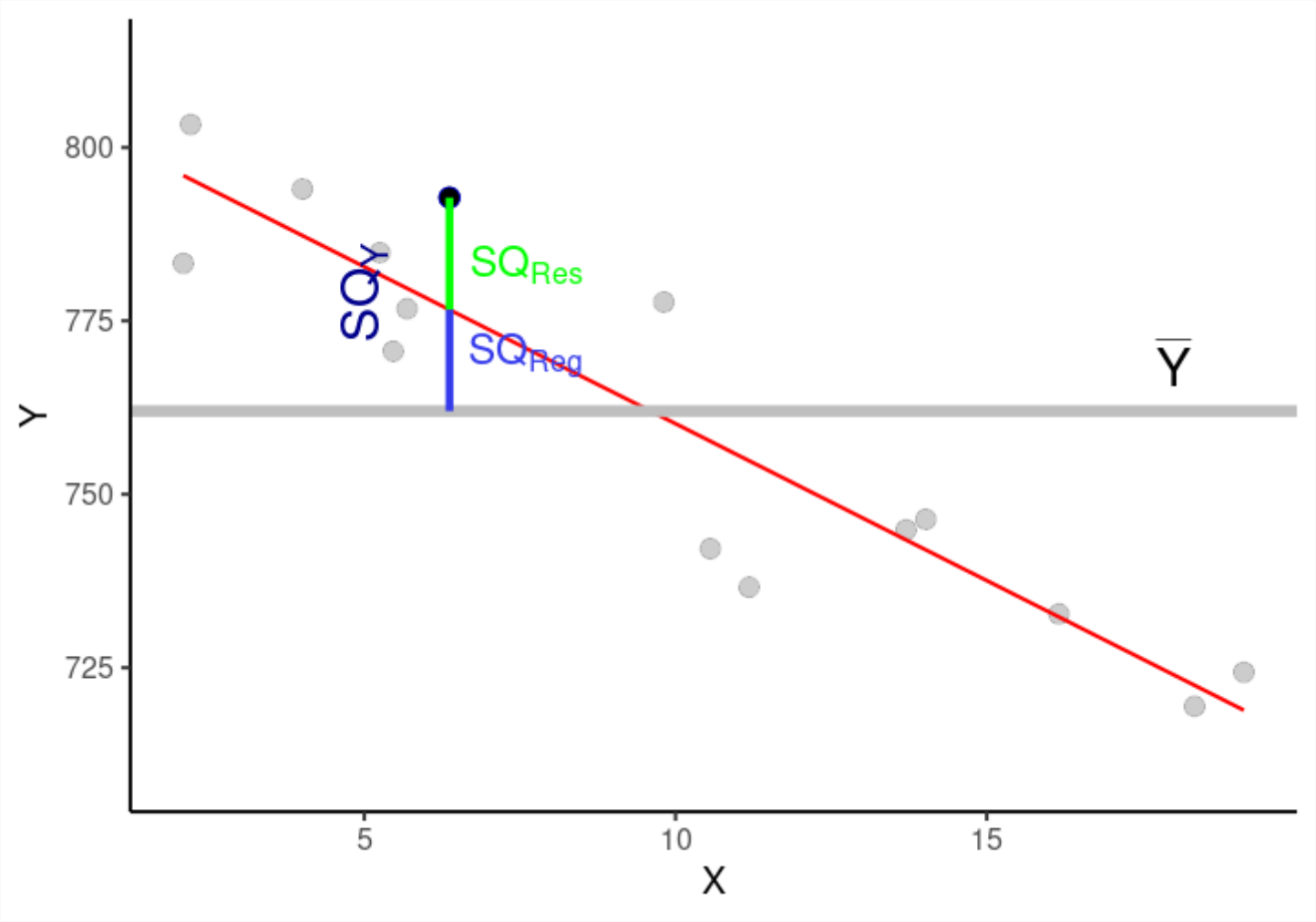
Somatório dos Quadrados de Y - SQ_Y

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



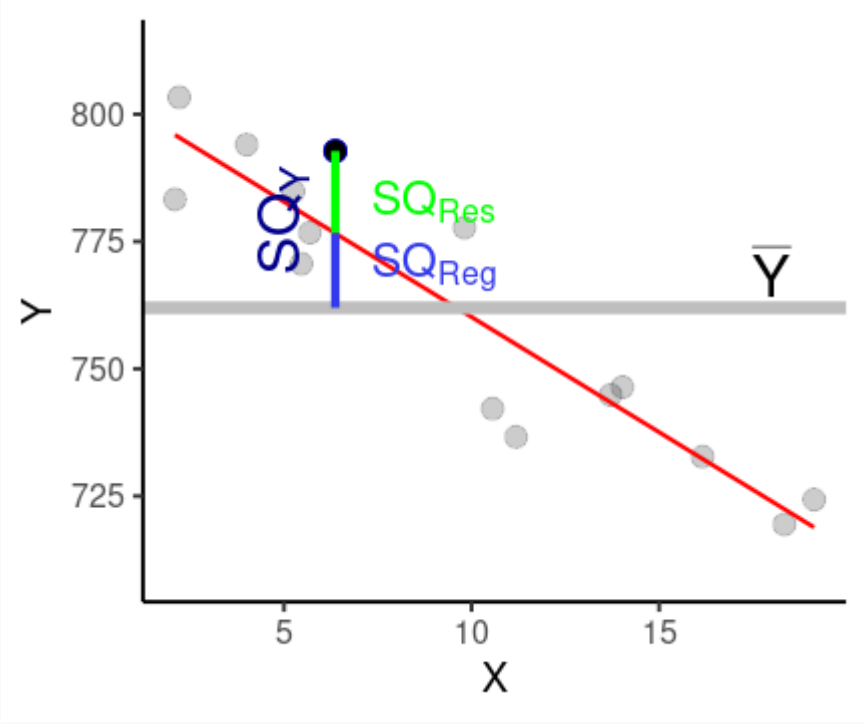
9. Partição da Soma dos Quadrados e variação explicada

$$SQ_Y = SQ_{Reg} + SQ_{Res}$$



O Coeficiente de Determinação

$$R^2 = \frac{SQ_{Reg}}{SQ_Y} = \frac{SQ_{Reg}}{SQ_{Reg} + SQ_{Res}}$$



R^2 estima a proporção da variação em Y que pode ser atribuída ao modelo de regressão linear.

$$0 \leq R^2 \leq 1$$

> Numericamente, o R^2 é igual ao coeficiente de correlação linear de Pearson elevado ao quadrado.

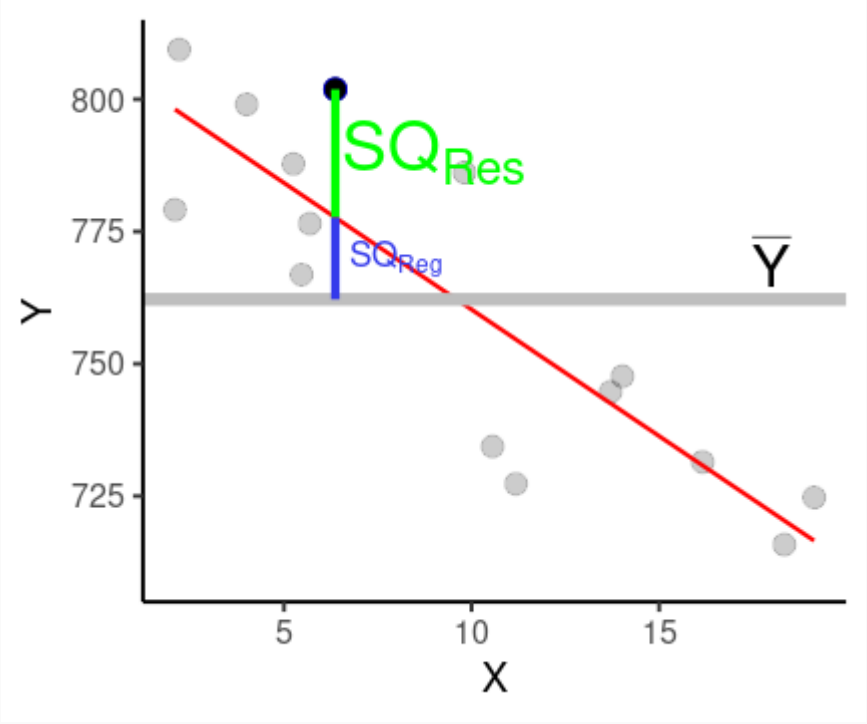
A Análise de Variância da Regressão

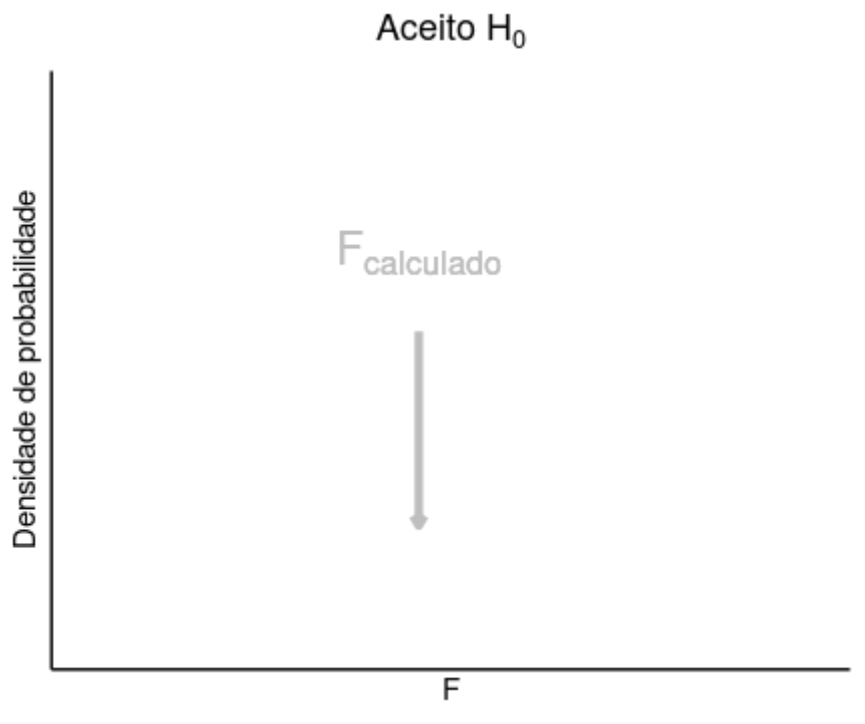
	gl	SQ	QM	F	p
X	1	9294.914	9294.914	77.49312	8e-07
Resíduo	13	1559.285	119.945	NA	NA

- gl: graus de liberdade
- SQ: soma dos quadrados
- QM: quadrado médio
- F: estatística F
- p: valor de probabilidade na distribuição F

A Análise de Variância da Regressão

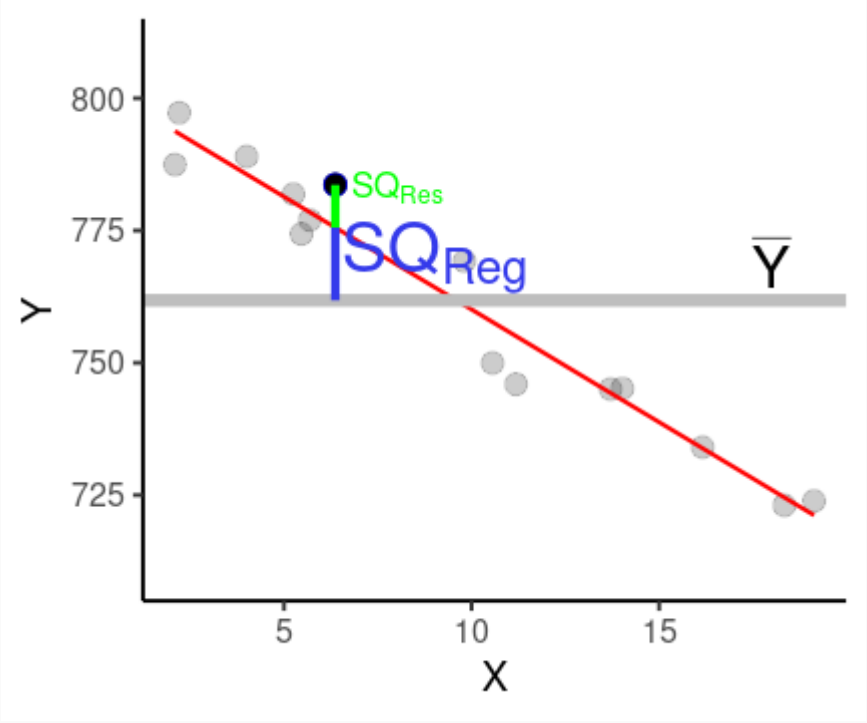
$$F_{calculado} = \frac{QM_{Reg}}{QM_{Res}}$$

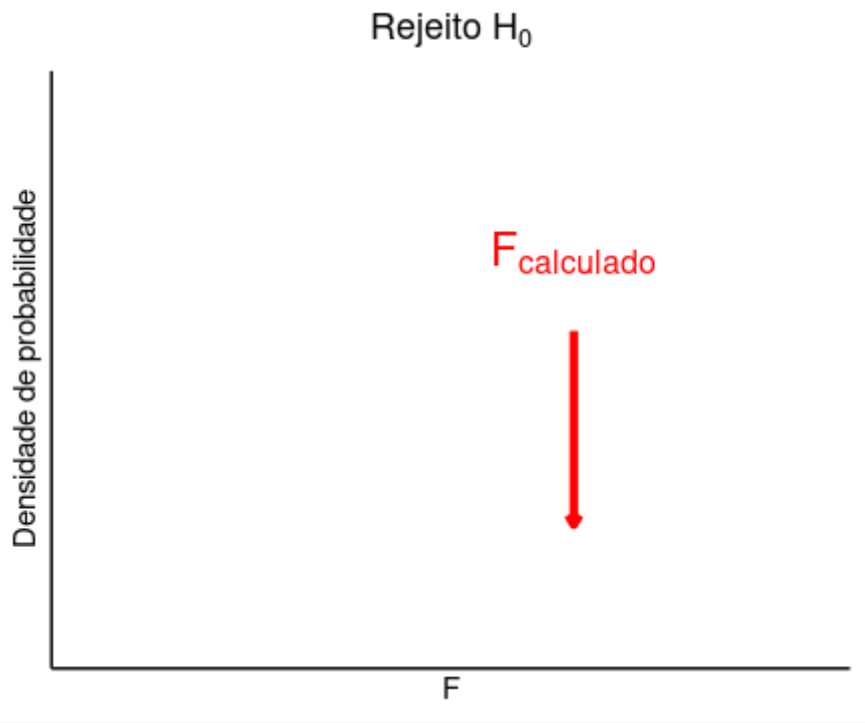




A Análise de Variância da Regressão

$$F_{calculado} = \frac{QM_{Reg}}{QM_{Res}}$$

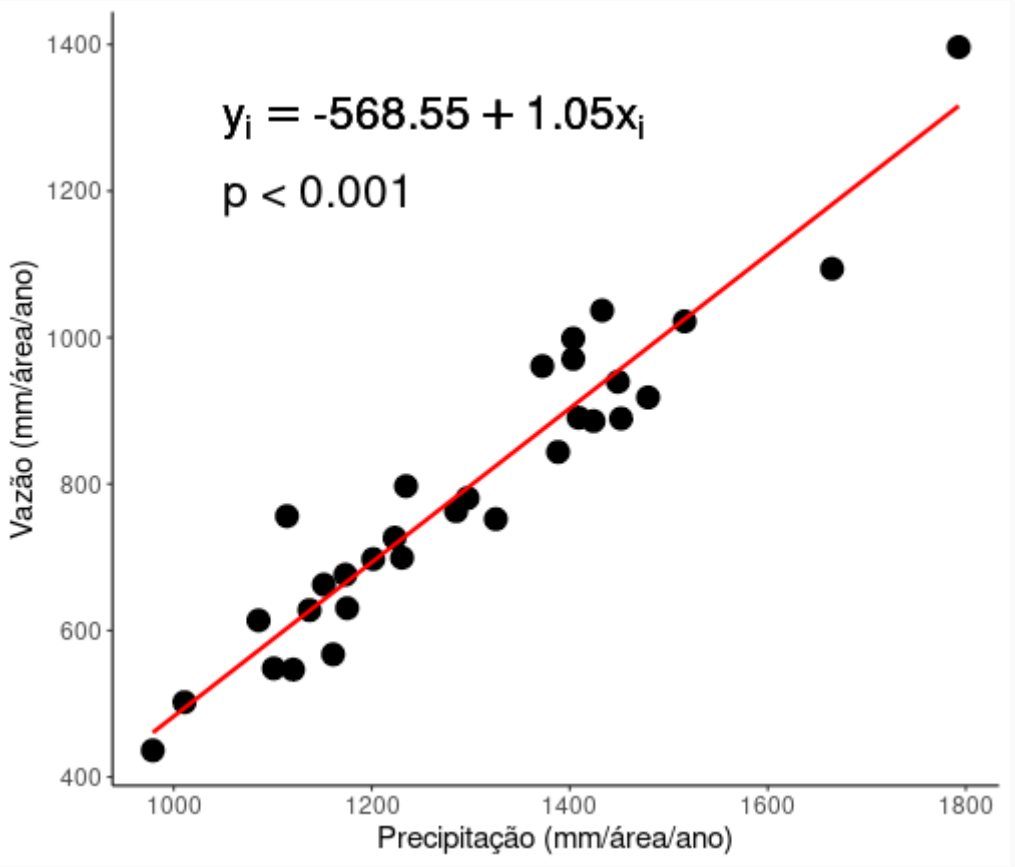




10. Os comandos em R

```
m_regressao = lm(Flow ~ Precipitation , data = st_ref)
summary(m_regressao)
```

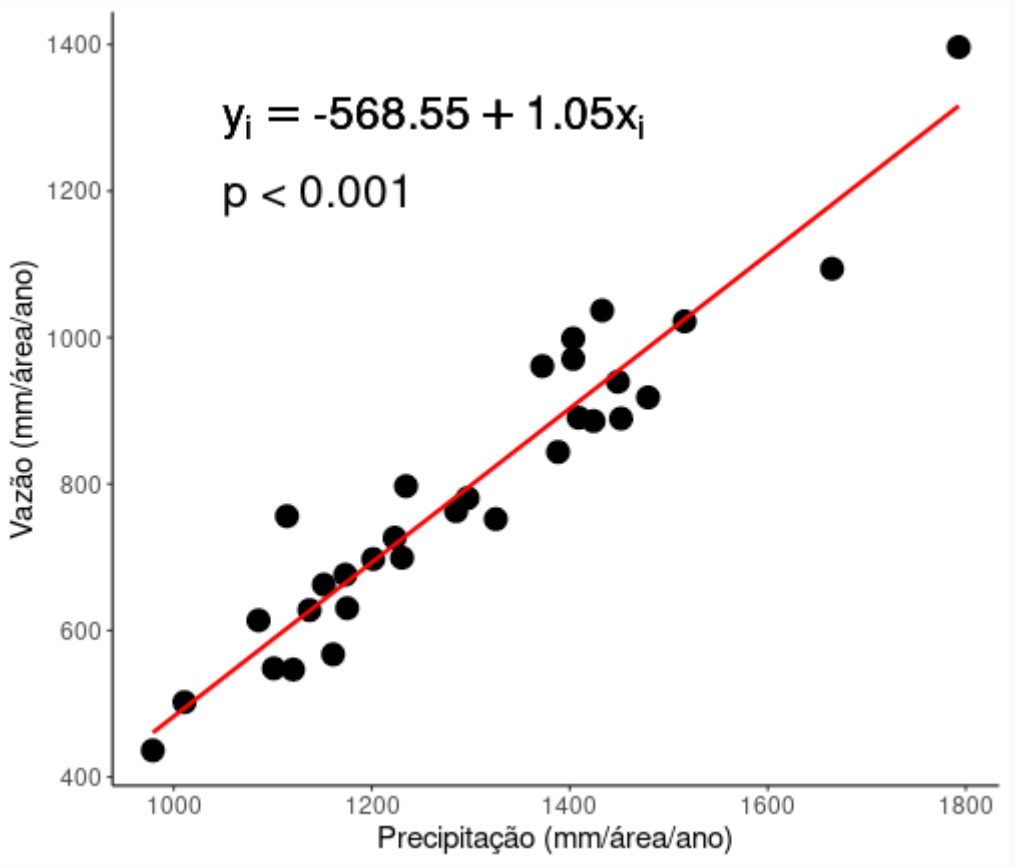
```
##
## Call:
## lm(formula = Flow ~ Precipitation, data = st_ref)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -87.86 -41.68 -14.03  30.64 153.09
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -568.54529   78.88794  -7.207  6.2e-08 ***
## Precipitation    1.05130    0.06024  17.451  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61.51 on 29 degrees of freedom
## Multiple R-squared:  0.9131,    Adjusted R-squared:  0.9101
## F-statistic: 304.5 on 1 and 29 DF,  p-value: < 2.2e-16
```



10. Os comandos em R

```
anova(m_regressao)
```

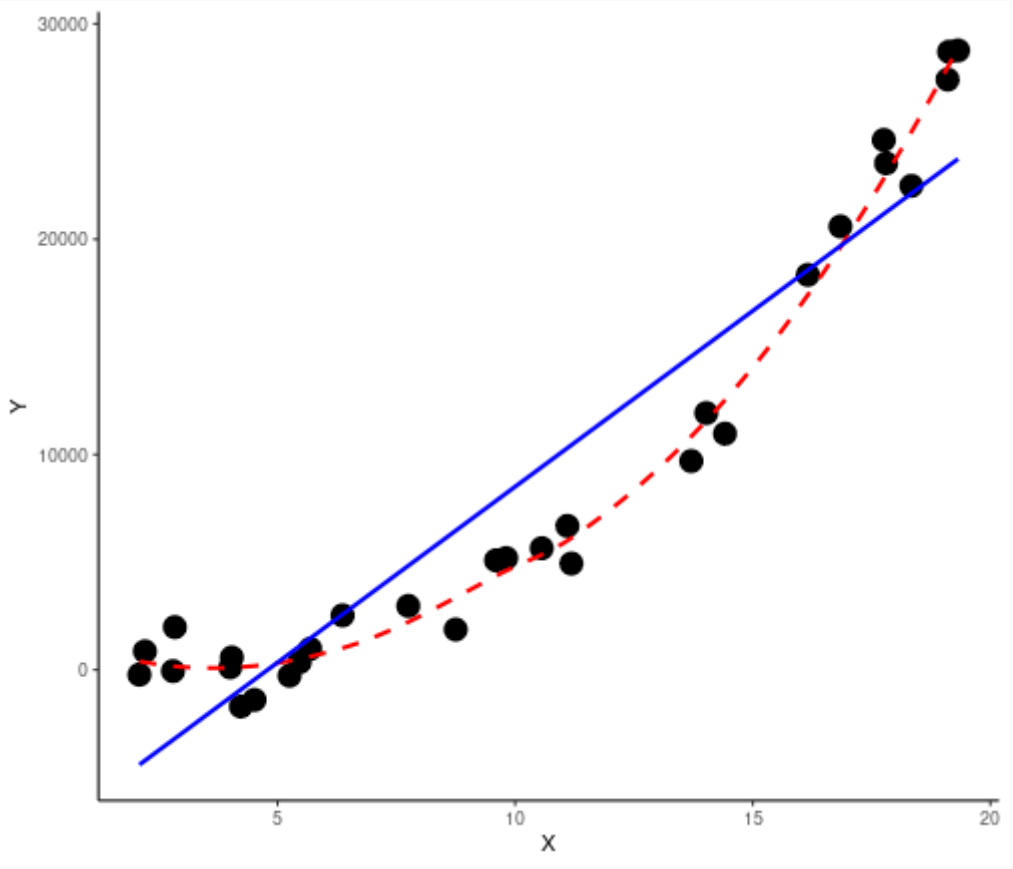
```
## Analysis of Variance Table
##
## Response: Flow
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## Precipitation  1 1152275 1152275  304.53 < 2.2e-16 ***
## Residuals     29  109730    3784
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



11. Pressupostos do modelo

Ao realizar uma regressão linear simples, devemos assumir/testar alguns pressupostos.

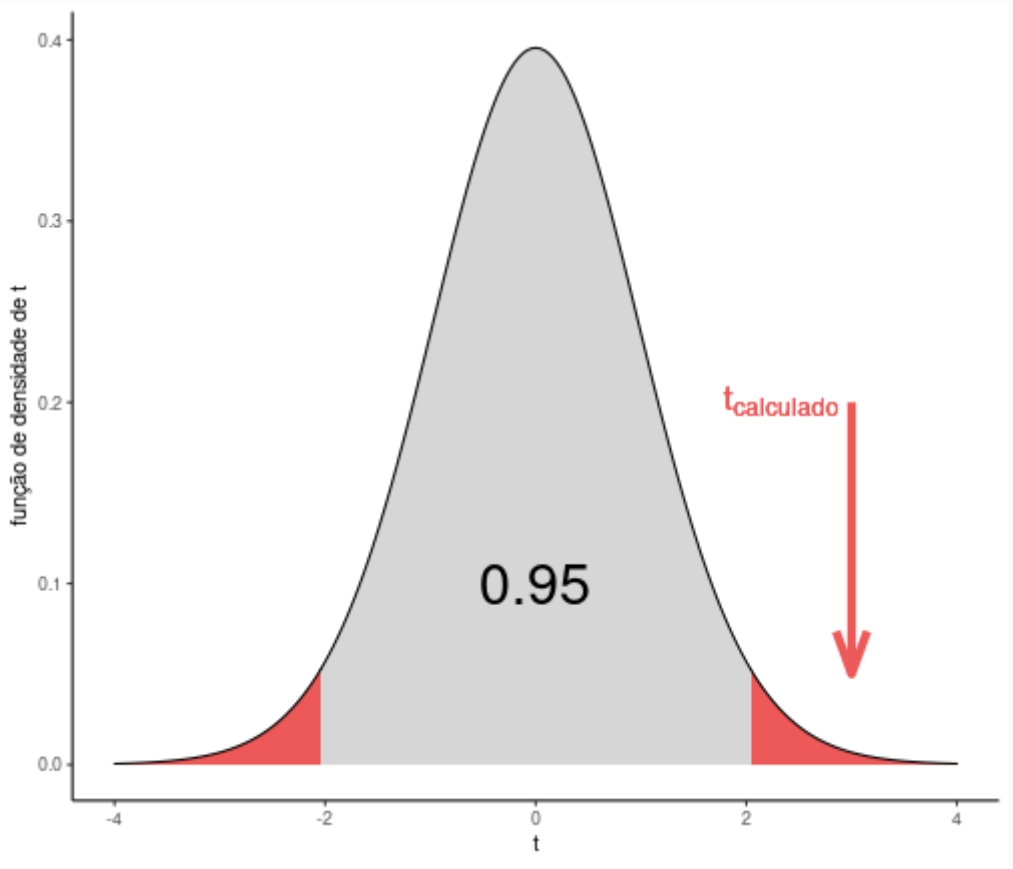
1. O modelo linear descreve adequadamente a relação funcional entre Y e X;
2. Cada par de observação (y_i, x_i) é independente dos demais;
3. A variável X é medida sem erros;
4. Os resíduos têm distribuição normal;
5. A variância residual σ^2 é constante ao longo de X.



11. Pressupostos do modelo

Ao realizar uma regressão linear simples, devemos assumir/testar alguns pressupostos.

1. O modelo linear descreve adequadamente a relação funcional entre Y e X;
2. Cada par de observação (y_i, x_i) é independente dos demais;
3. A variável X é medida sem erros;
4. Os resíduos têm distribuição normal;
5. A variância residual σ^2 é constante ao longo de X.



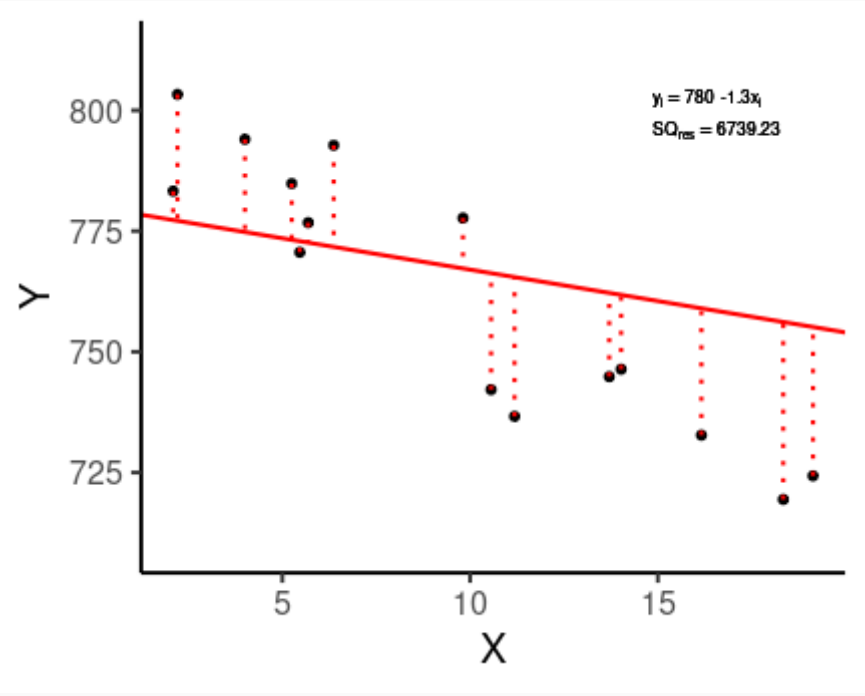
11. Pressupostos do modelo

Ao realizar uma regressão linear simples, devemos assumir/testar alguns pressupostos.

1. O modelo linear descreve adequadamente a relação funcional entre Y e X;
2. Cada par de observação (y_i, x_i) é independente dos demais;
3. A variável X é medida sem erros;
4. Os resíduos têm distribuição normal;
5. A variância residual σ^2 é constante ao longo de X.

$$Y \sim \mathcal{N}(\mu_i, \sigma^2)$$

$$E(Y|x_i) = \mu_i = \beta_0 + \beta_1 x_i$$

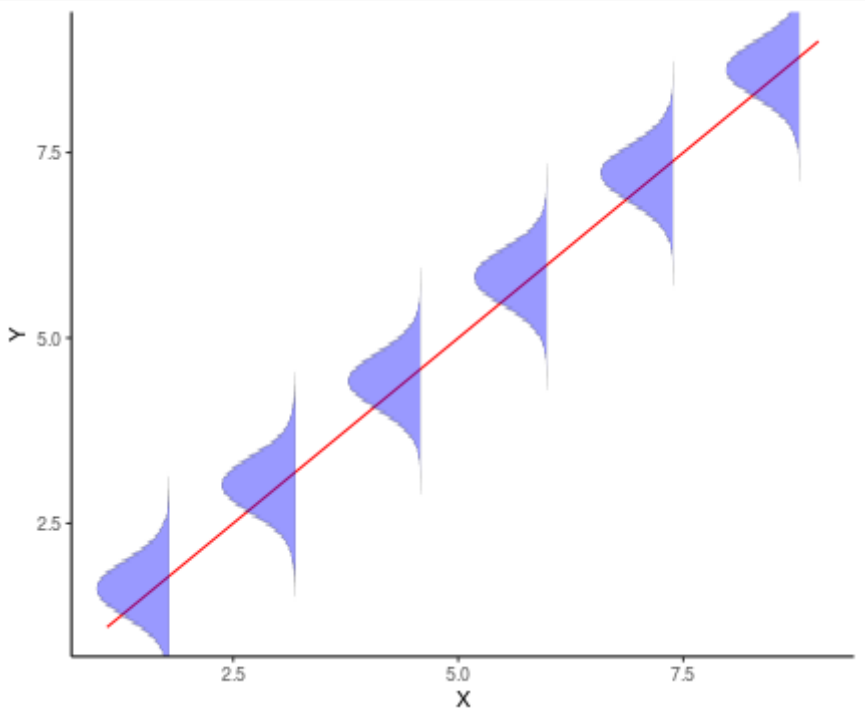


11. Pressupostos do modelo

Ao realizar uma regressão linear simples, devemos assumir/testar alguns pressupostos.

1. O modelo linear descreve adequadamente a relação funcional entre Y e X;
2. Cada par de observação (y_i, x_i) é independente dos demais;
3. A variável X é medida sem erros;
4. Os resíduos têm distribuição normal;
5. A variância residual σ^2 é constante ao longo de X.

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

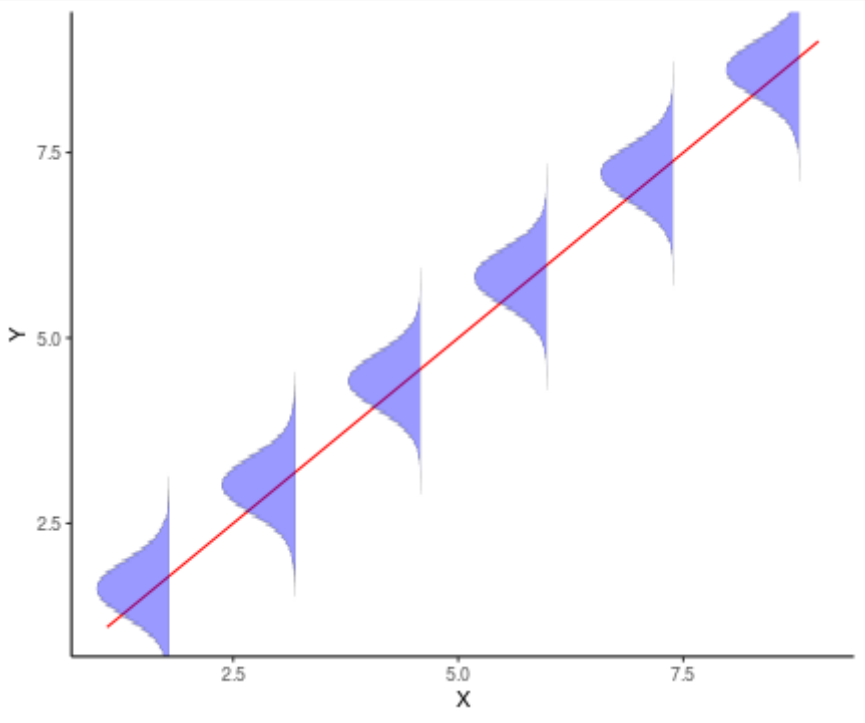


11. Pressupostos do modelo

Ao realizar uma regressão linear simples, devemos assumir/testar alguns pressupostos.

1. O modelo linear descreve adequadamente a relação funcional entre Y e X;
2. Cada par de observação (y_i, x_i) é independente dos demais;
3. A variável X é medida sem erros;
4. Os resíduos têm distribuição normal;
5. A variância residual σ^2 é constante ao longo de X.

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$



Um modelo de regressão linear **NÃO** precisa ser uma linha reta

> O que define um modelo estatístico como linear é a posição dos seus parâmetros com relação a(s) variável(is) preditora(s). Os parametros a serem estimados devem estar na **MESMA LINHA** da variável dependente.
>
> Os métodos discutidos para regressão linear simples também também se aplicam aos modelos de regressão múltipla e aos modelos polinomiais.

$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$: Regressão Linear Simples

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} \epsilon_i$: Regressão Linear Múltipla

$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 \epsilon_i$: Regressão Polinomial

---

12. Transformações lineares

Outros modelos podem ser *linearizados* por meio de uma **função de ligação** do tipo:

$$\eta = g(\beta_i X_i)$$

Modelos Lineares Generalizados

$\eta = \beta_0 + \beta_1 X_i$

$Y \sim \mathcal{N}(\mu = \beta_0 + \beta_1 X_i, \sigma^2)$: Modelo Normal (ex. Regressão Linear Simples)

$\eta = log(\beta_0 X_i^{\beta_1}) = log(\beta_0) + \beta_1 log(X_i)$

$Y \sim \mathcal{P}ois(\lambda = e^{\eta})$: Modelo de Poisson (ex. variáveis de contagem)

$\eta = logit(\eta) = log\left(\frac{\eta}{1-\eta}\right)$

$Y \sim \mathcal{B}inon(n = 1, p = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}})$: Modelo Binomial ou Regressão Logística (ex. variáveis categóricas binárias)
