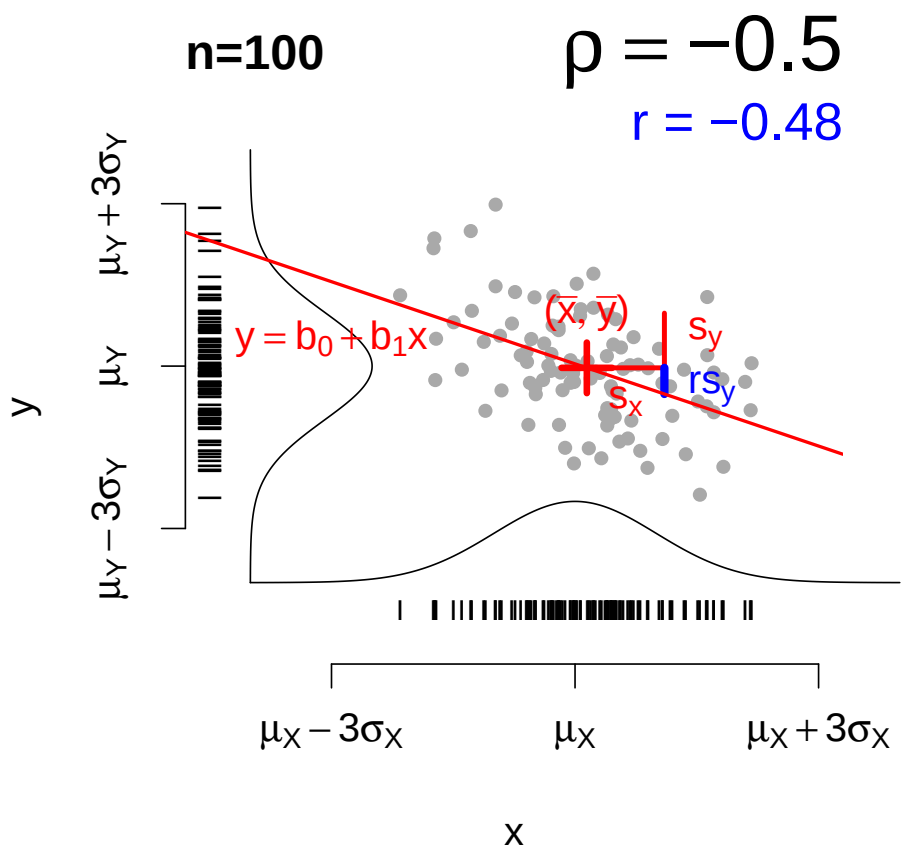
Linear regression

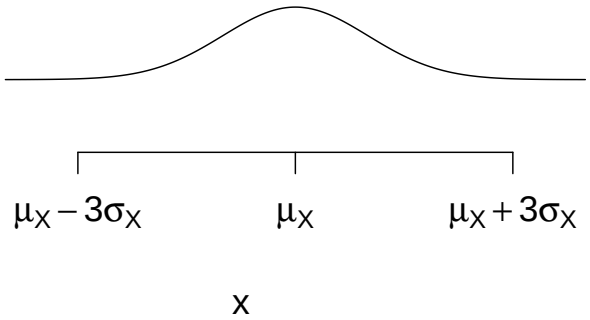


All examples are fictitious. All data are simulated and the graphics were created with the statistical program package R.

Sämtliche Beispiele sind frei erfunden. Alle Daten sind simuliert und die Grafiken wurden mit statistischen Programmpaket R erstellt.

Reminder: Correlation

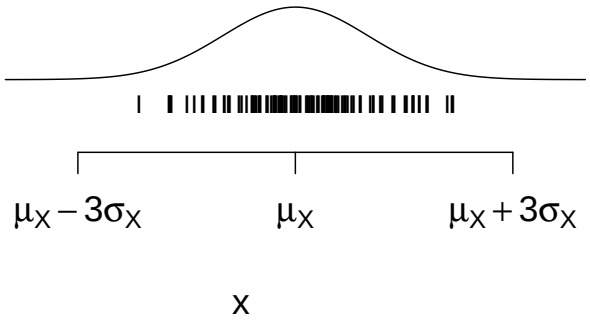
- $X_1, \ldots, X_n$ i.i.d. RVs, $X_1 \sim N(\mu_x, \sigma_x^2)$.



Reminder: Correlation

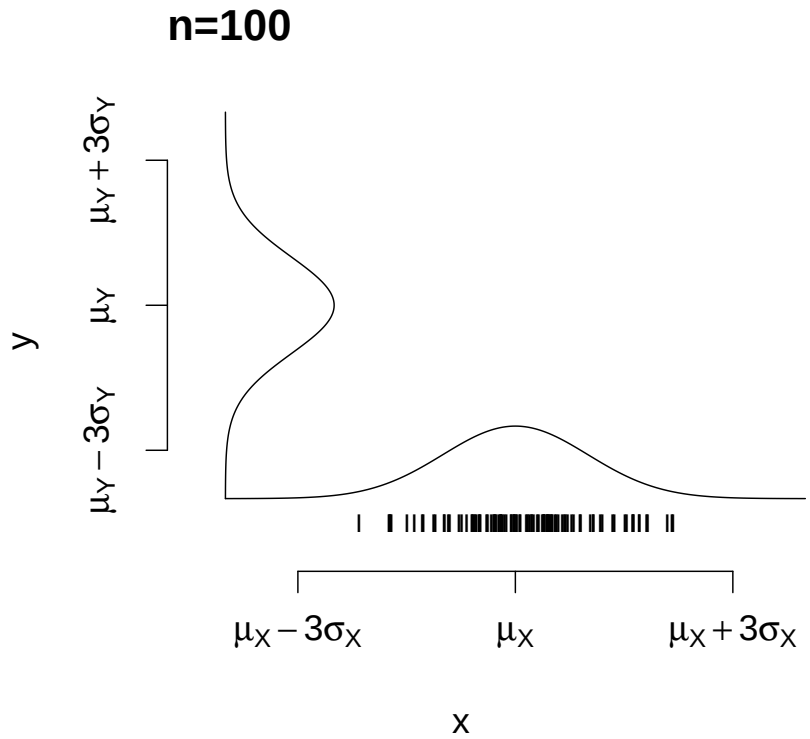
- $X_1, \ldots, X_n$ i.i.d. RVs, $X_1 \sim N(\mu_x, \sigma_x^2)$. Here $n = 100$

n=100



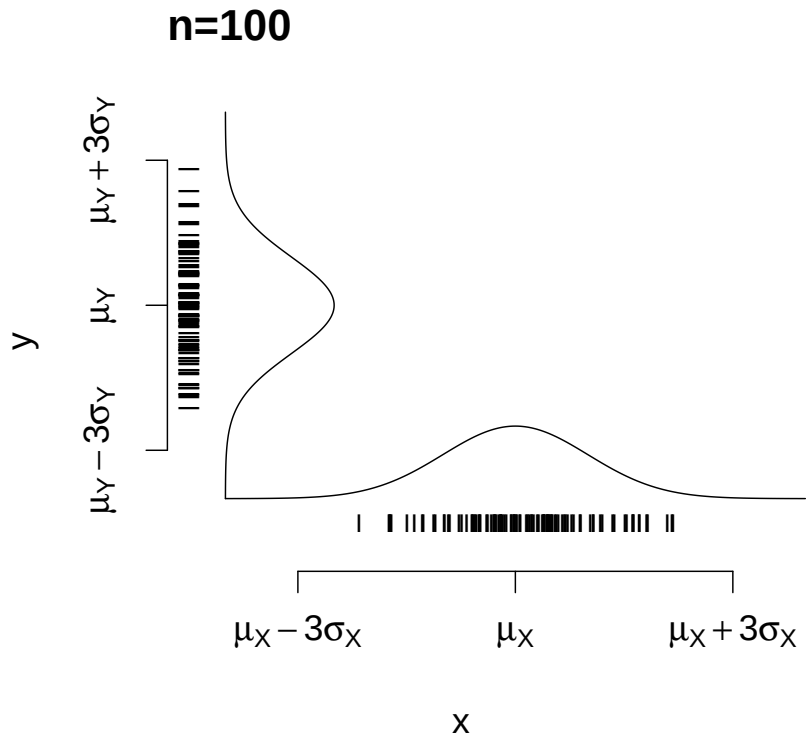
Reminder: Correlation

- $X_1, \ldots, X_n$ i.i.d. RVs, $X_1 \sim N(\mu_x, \sigma_x^2)$. Here $n = 100$
- $Y_1, \ldots, Y_n$ i.i.d. RVs, $Y_1 \sim N(\mu_y, \sigma_y^2)$



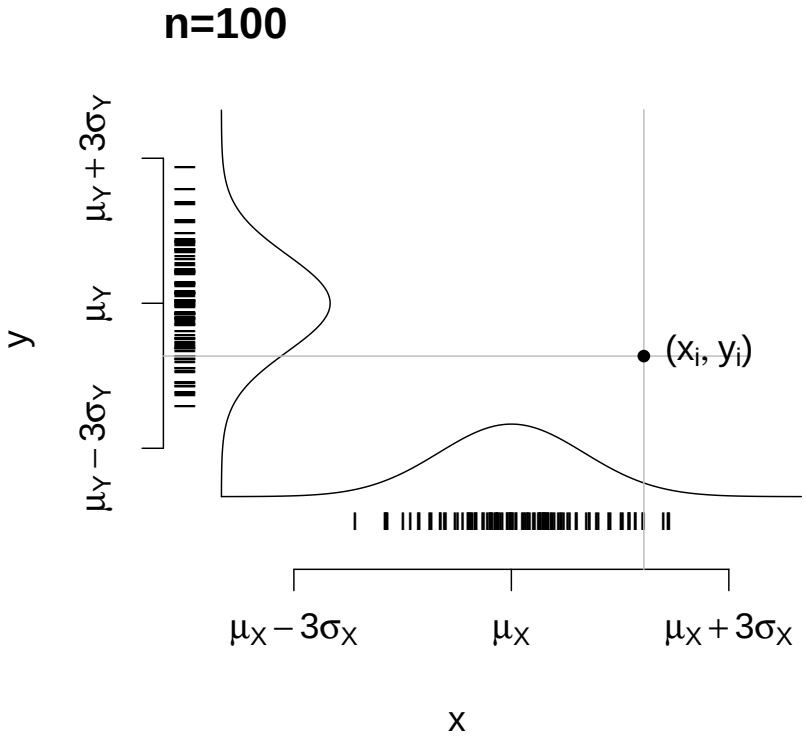
Reminder: Correlation

- $X_1, \ldots, X_n$ i.i.d. RVs, $X_1 \sim N(\mu_x, \sigma_x^2)$. Here $n = 100$
- $Y_1, \ldots, Y_n$ i.i.d. RVs, $Y_1 \sim N(\mu_y, \sigma_y^2)$



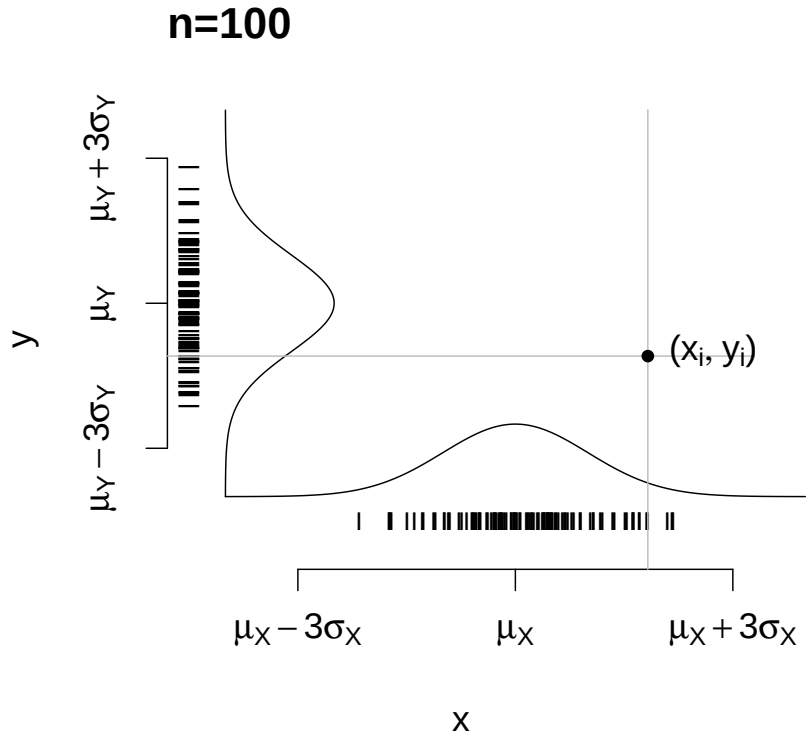
Reminder: Correlation

- $X_1, \ldots, X_n$ i.i.d. RVs, $X_1 \sim N(\mu_x, \sigma_x^2)$. Here $n = 100$
- $Y_1, \ldots, Y_n$ i.i.d. RVs, $Y_1 \sim N(\mu_y, \sigma_y^2)$
- also let the pairs $(X_i, Y_i)_{i=1,\ldots,n}$ be independent over $i = 1, 2, \ldots$



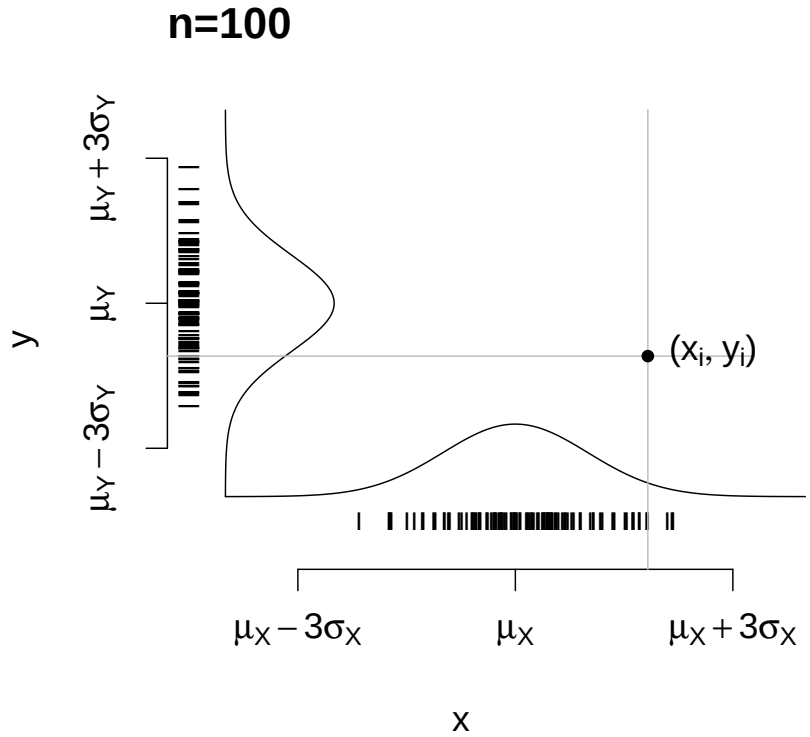
Reminder: Correlation

- $X_1, \ldots, X_n$ i.i.d. RVs, $X_1 \sim N(\mu_x, \sigma_x^2)$. Here $n = 100$
- $Y_1, \ldots, Y_n$ i.i.d. RVs, $Y_1 \sim N(\mu_y, \sigma_y^2)$
- also let the pairs $(X_i, Y_i)_{i=1,\ldots,n}$ be independent over $i = 1, 2, \ldots$
- So long, nothing said about the *relation* between X_i and Y_i



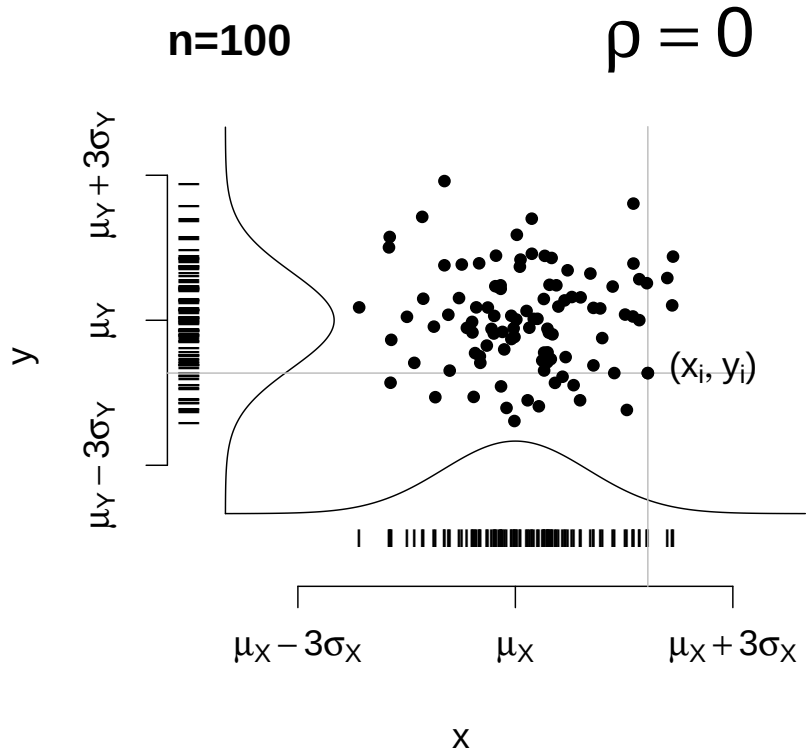
Reminder: Correlation

- $X_1, \ldots, X_n$ i.i.d. RVs, $X_1 \sim N(\mu_x, \sigma_x^2)$. Here $n = 100$
- $Y_1, \ldots, Y_n$ i.i.d. RVs, $Y_1 \sim N(\mu_y, \sigma_y^2)$
- also let the pairs $(X_i, Y_i)_{i=1,\ldots,n}$ be independent over $i = 1, 2, \ldots$
- So long, nothing said about the *relation* between X_i and Y_i
- This is accomplished (e.g.,) through the notion of *correlation*



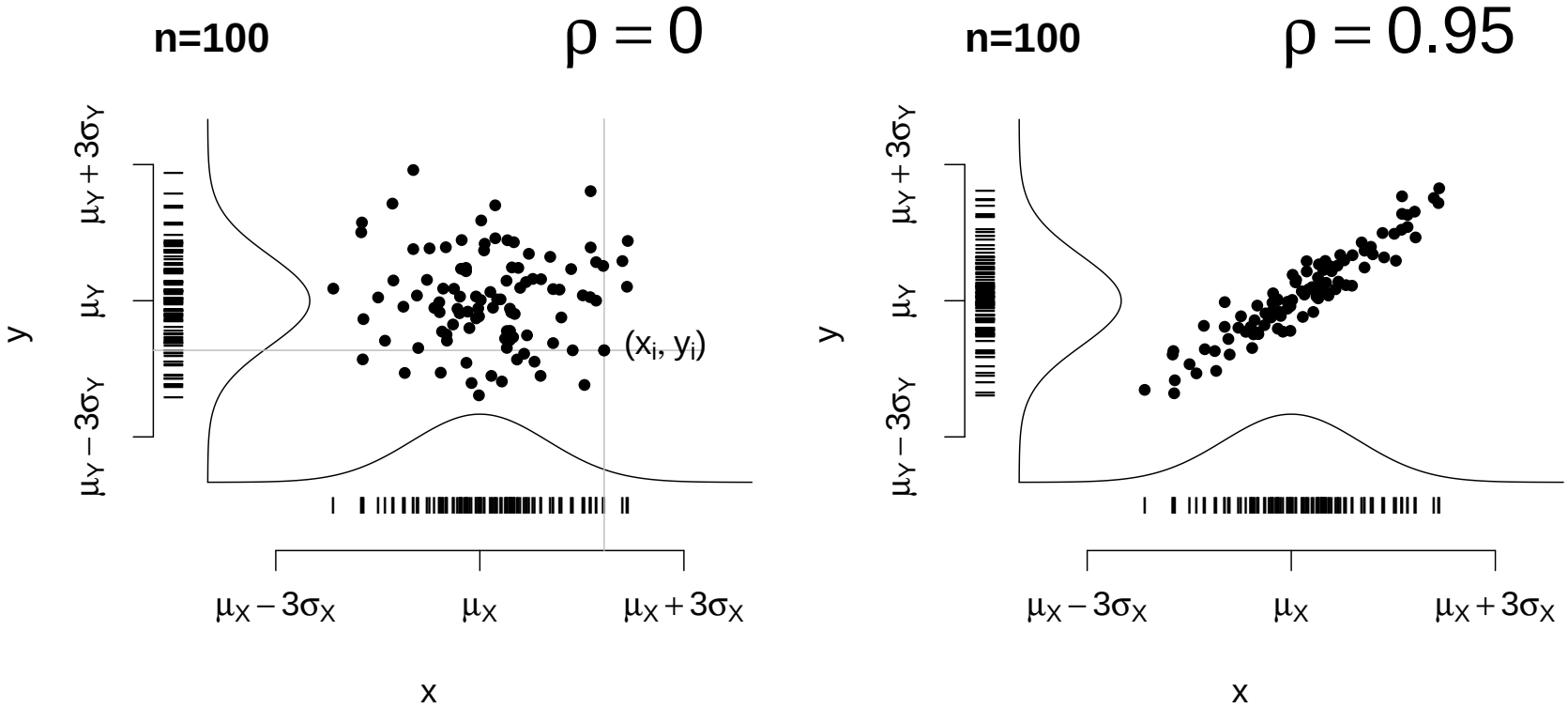
Reminder: Correlation

- $X_1, \ldots, X_n$ i.i.d. RVs, $X_1 \sim N(\mu_x, \sigma_x^2)$. Here $n = 100$
- $Y_1, \ldots, Y_n$ i.i.d. RVs, $Y_1 \sim N(\mu_y, \sigma_y^2)$
- also let the pairs $(X_i, Y_i)_{i=1,\ldots,n}$ be independent over $i = 1, 2, \ldots$
- So long, nothing said about the *relation* between X_i and Y_i
- This is accomplished (e.g.,) through the notion of *correlation*



Reminder: Correlation

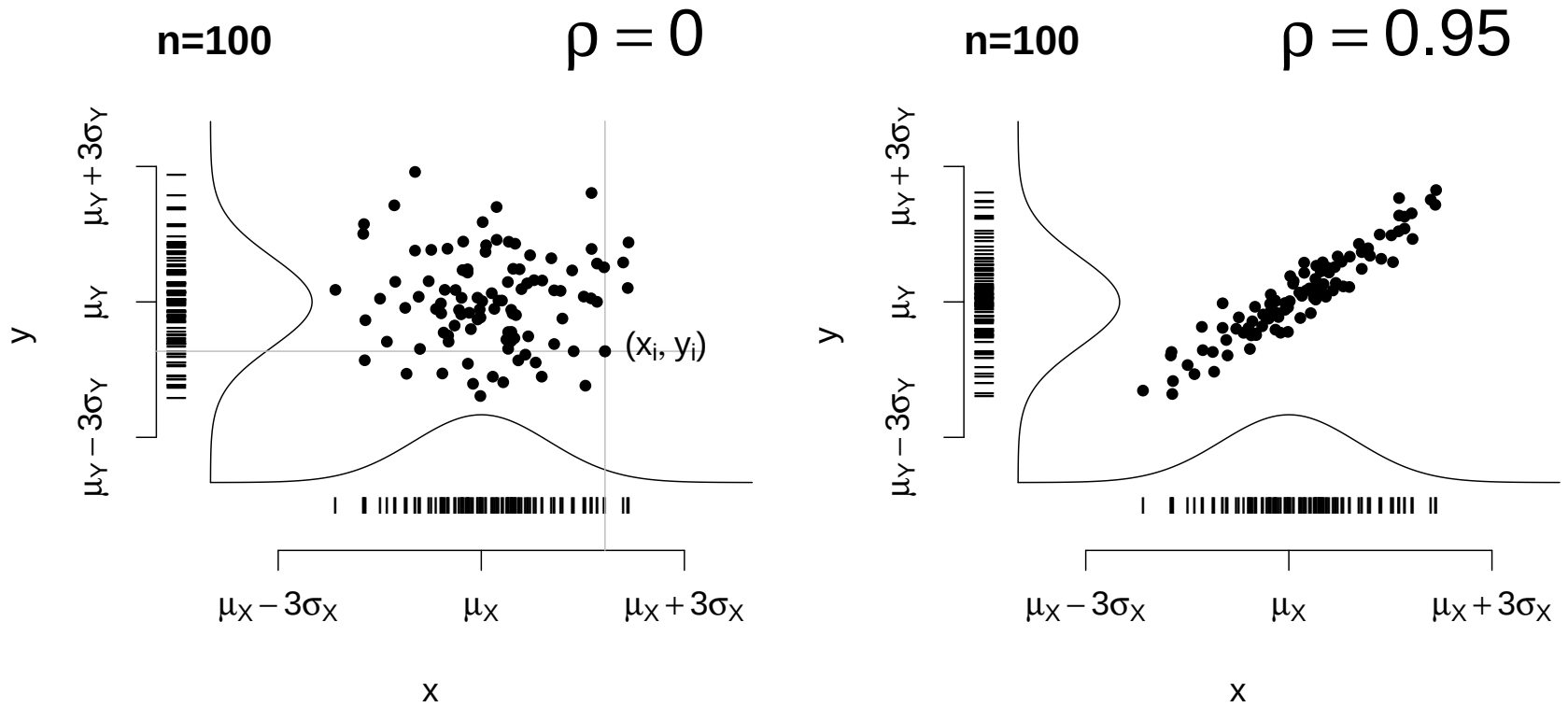
- $X_1, \ldots, X_n$ i.i.d. RVs, $X_1 \sim N(\mu_x, \sigma_x^2)$. Here $n = 100$
- $Y_1, \ldots, Y_n$ i.i.d. RVs, $Y_1 \sim N(\mu_y, \sigma_y^2)$
- also let the pairs $(X_i, Y_i)_{i=1,\ldots,n}$ be independent over $i = 1, 2, \ldots$
- So long, nothing said about the *relation* between X_i and Y_i
- This is accomplished (e.g.,) through the notion of *correlation*



Reminder: Correlation

- $X_1, \ldots, X_n$ i.i.d. RVs, $X_1 \sim N(\mu_x, \sigma_x^2)$. Here $n = 100$
- $Y_1, \ldots, Y_n$ i.i.d. RVs, $Y_1 \sim N(\mu_y, \sigma_y^2)$
- also let the pairs $(X_i, Y_i)_{i=1,\ldots,n}$ be independent over $i = 1, 2, \ldots$
- So long, nothing said about the *relation* between X_i and Y_i
- This is accomplished (e.g.,) through the notion of *correlation*
- Definition: For the RVs X and Y (with $\mathbb{V}ar(X), \mathbb{V}ar(Y) \in (0, \infty)$) their correlation ρ is given as

$$\rho := \mathbb{C}orr(X, Y)$$

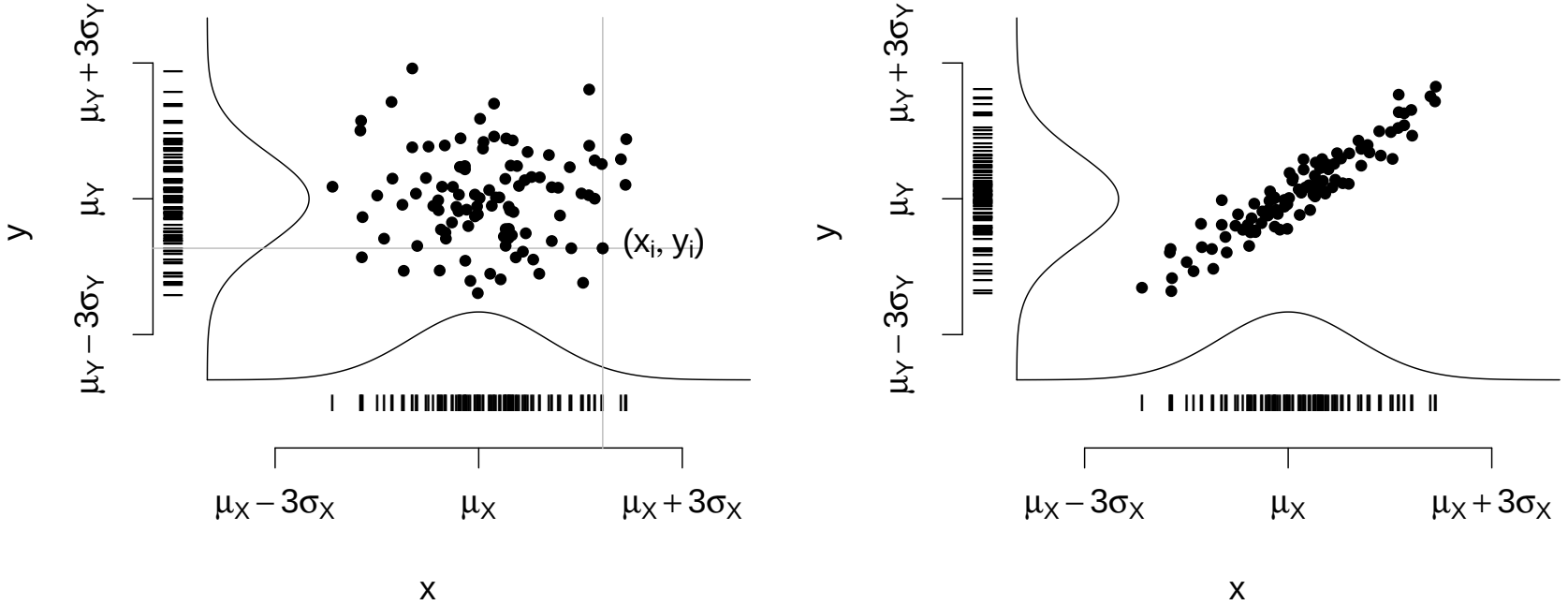


Reminder: Correlation

- $X_1, \ldots, X_n$ i.i.d. RVs, $X_1 \sim N(\mu_x, \sigma_x^2)$. Here $n = 100$
- $Y_1, \ldots, Y_n$ i.i.d. RVs, $Y_1 \sim N(\mu_y, \sigma_y^2)$
- also let the pairs $(X_i, Y_i)_{i=1,\ldots,n}$ be independent over $i = 1, 2, \ldots$
- So long, nothing said about the *relation* between X_i and Y_i
- This is accomplished (e.g.,) through the notion of *correlation*
- Definition: For the RVs X and Y (with $\mathbb{V}ar(X), \mathbb{V}ar(Y) \in (0, \infty)$) their correlation ρ is given as

$$\rho := \mathbb{C}orr(X, Y) := \frac{\mathbb{C}ov(X, Y)}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}}$$

n=100 $\rho = 0$ **n=100** $\rho = 0.95$

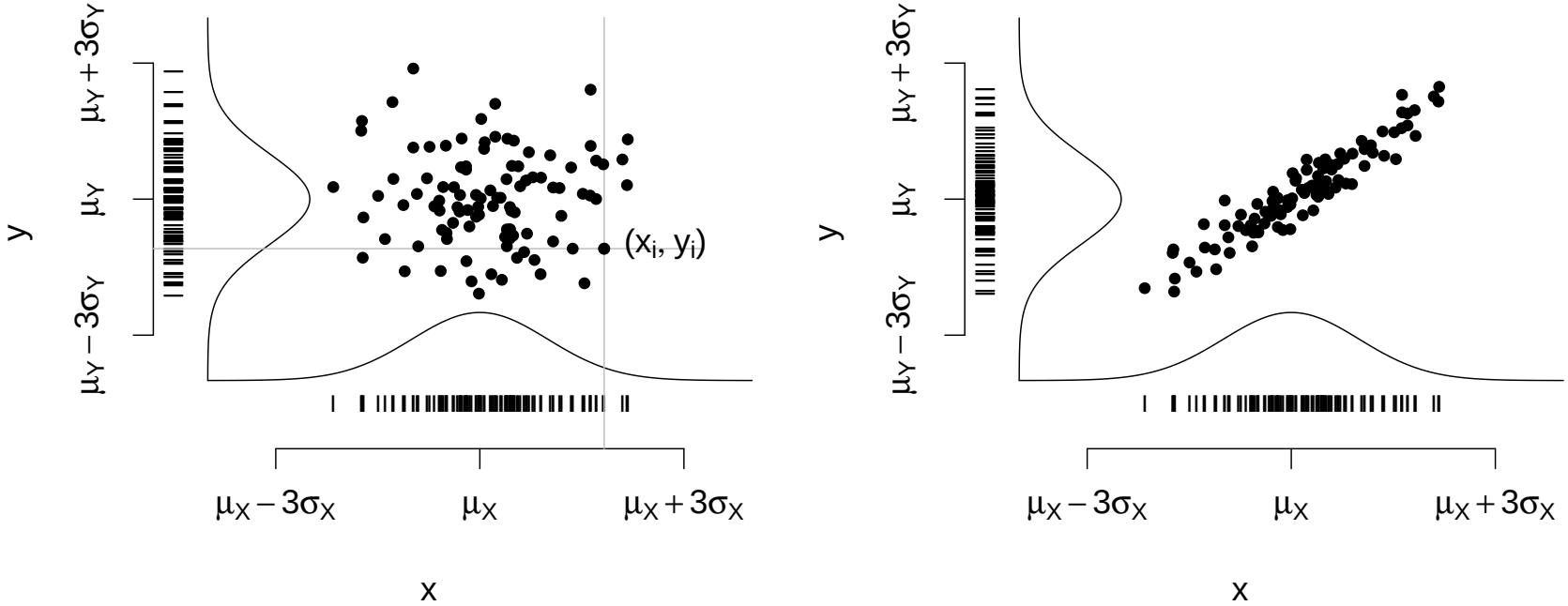


Reminder: Correlation

- $X_1, \ldots, X_n$ i.i.d. RVs, $X_1 \sim N(\mu_x, \sigma_x^2)$. Here $n = 100$
- $Y_1, \ldots, Y_n$ i.i.d. RVs, $Y_1 \sim N(\mu_y, \sigma_y^2)$
- also let the pairs $(X_i, Y_i)_{i=1,\ldots,n}$ be independent over $i = 1, 2, \ldots$
- So long, nothing said about the *relation* between X_i and Y_i
- This is accomplished (e.g.,) through the notion of *correlation*
- Definition: For the RVs X and Y (with $\mathbb{V}ar(X), \mathbb{V}ar(Y) \in (0, \infty)$) their correlation ρ is given as

$$\rho := \mathbb{C}orr(X, Y) := \frac{\mathbb{C}ov(X, Y)}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}} := \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}}$$

n=100 $\rho = 0$ **n=100** $\rho = 0.95$



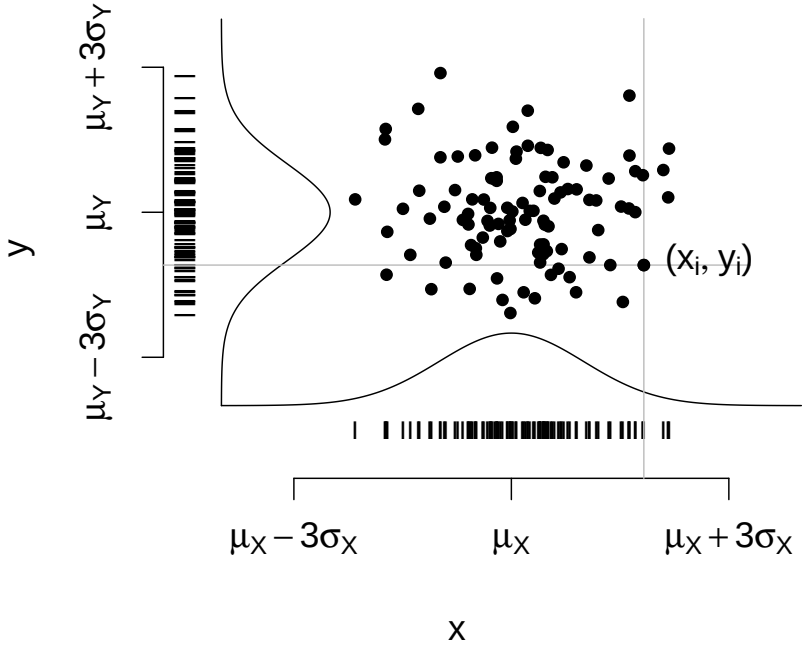
Reminder: Correlation

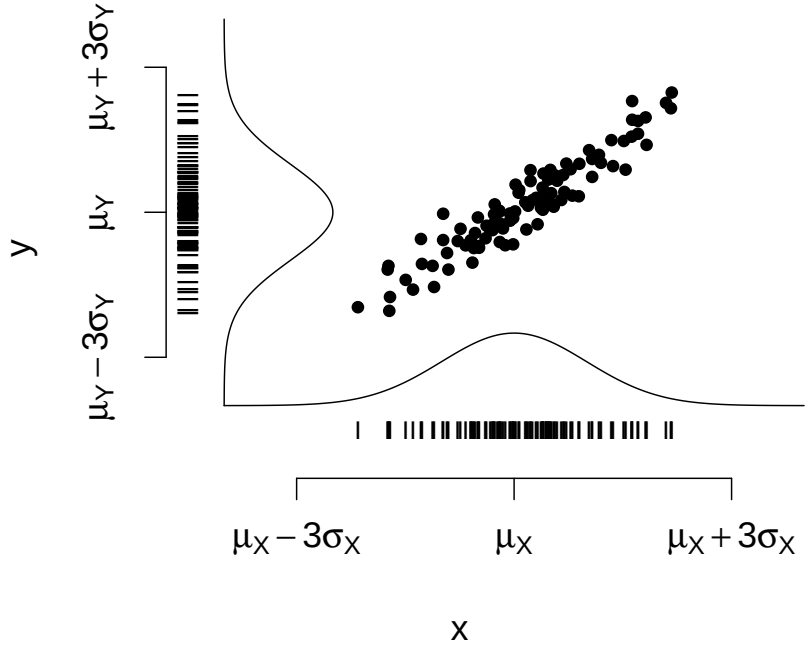
- $X_1, \ldots, X_n$ i.i.d. RVs, $X_1 \sim N(\mu_x, \sigma_x^2)$. Here $n = 100$
- $Y_1, \ldots, Y_n$ i.i.d. RVs, $Y_1 \sim N(\mu_y, \sigma_y^2)$
- also let the pairs $(X_i, Y_i)_{i=1,\ldots,n}$ be independent over $i = 1, 2, \ldots$
- So long, nothing said about the *relation* between X_i and Y_i
- This is accomplished (e.g.,) through the notion of *correlation*
- Definition: For the RVs X and Y (with $\mathbb{V}ar(X), \mathbb{V}ar(Y) \in (0, \infty)$) their correlation ρ is given as

ρ is also known as *Pearson's coefficient of correlation*

$$\rho := \mathbb{C}orr(X, Y) := \frac{\mathbb{C}ov(X, Y)}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}} := \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}}$$

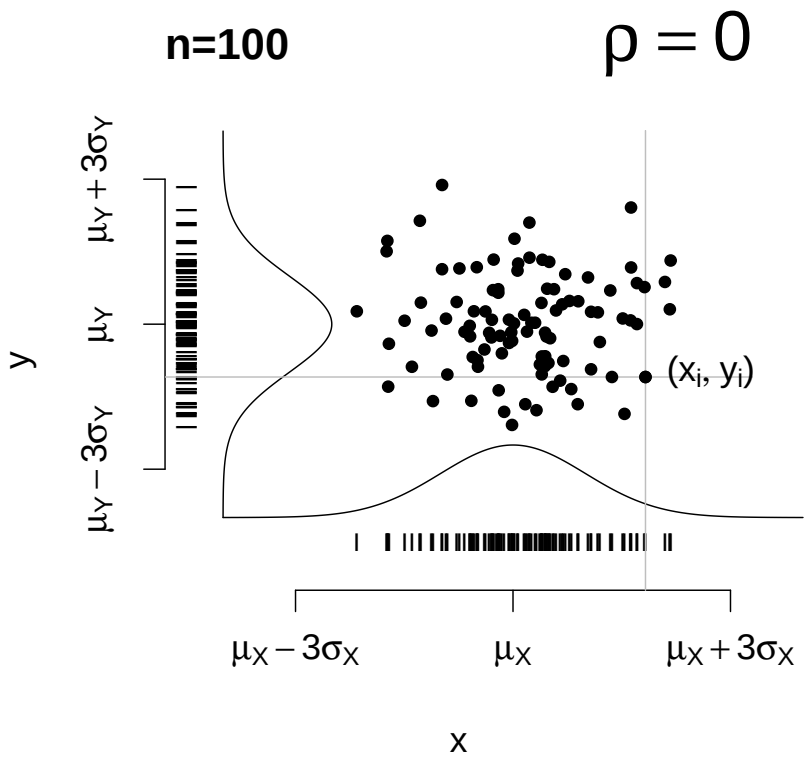
n=100 $\rho = 0$ **n=100** $\rho = 0.95$





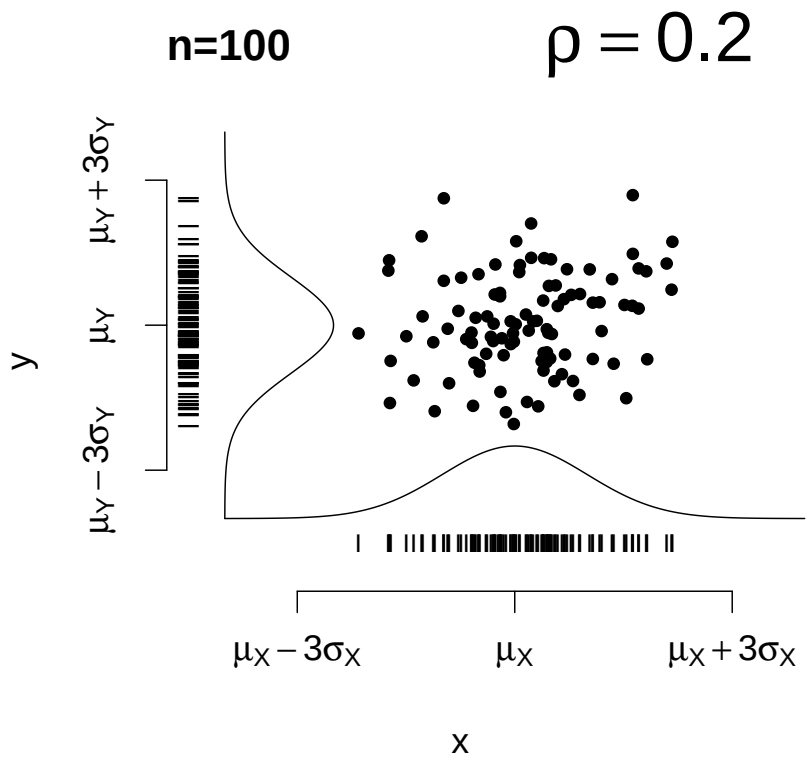
Examples: Correlation

$$\rho := \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}}$$



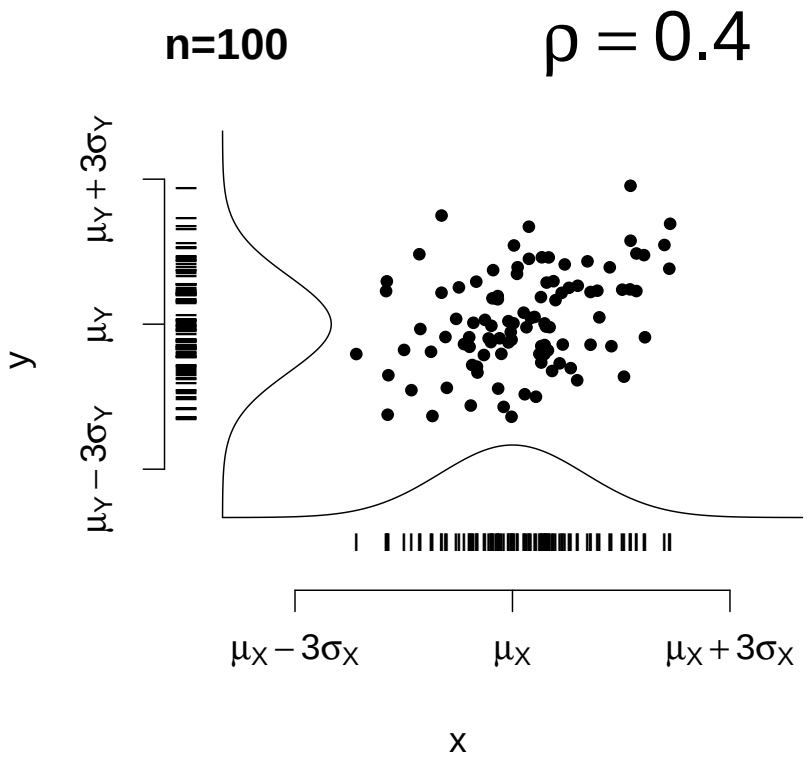
Examples: Correlation

$$\rho := \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}}$$



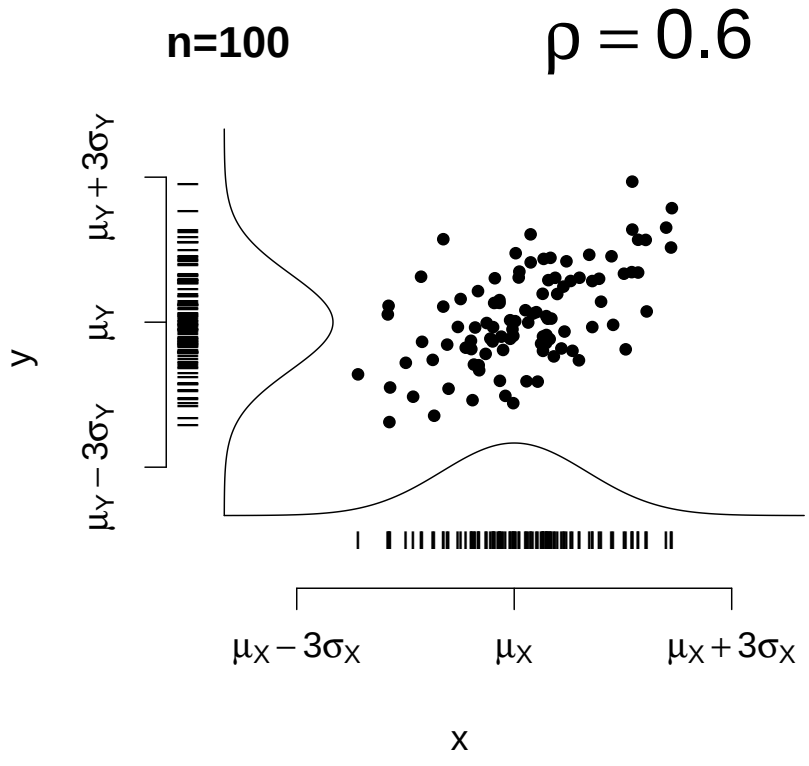
Examples: Correlation

$$\rho := \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}}$$



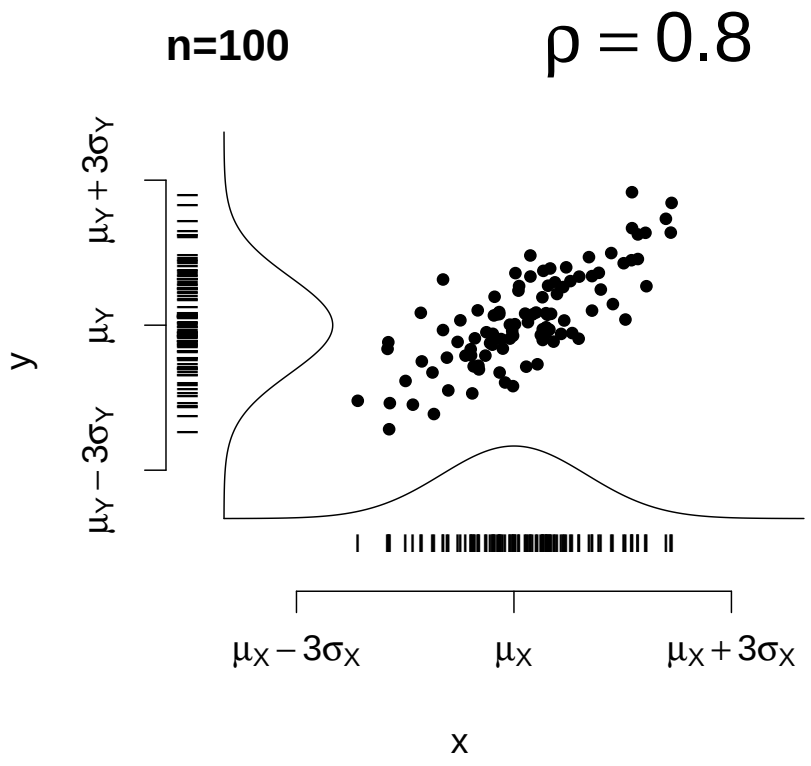
Examples: Correlation

$$\rho := \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}}$$



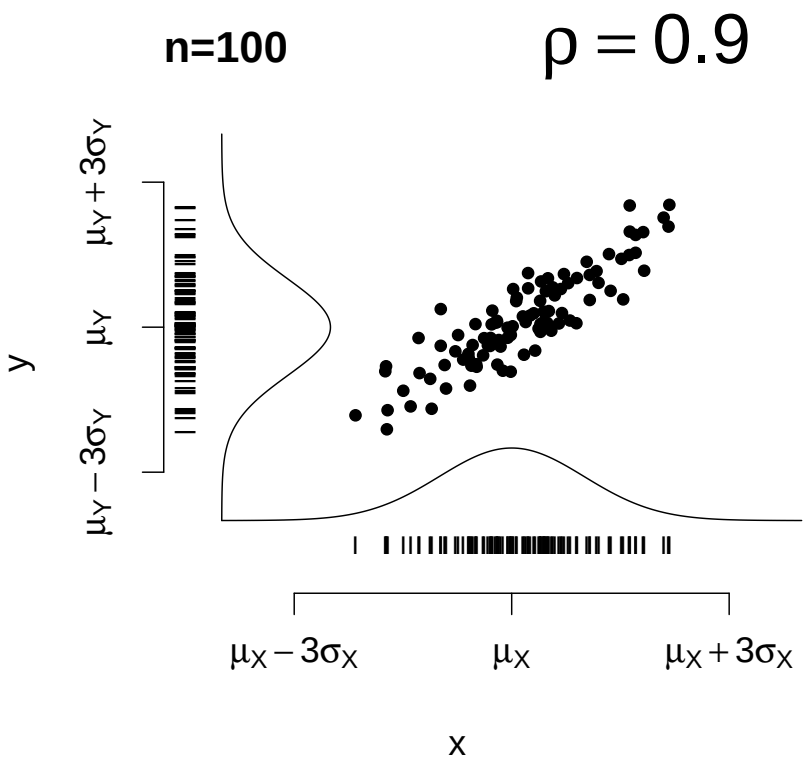
Examples: Correlation

$$\rho := \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}}$$



Examples: Correlation

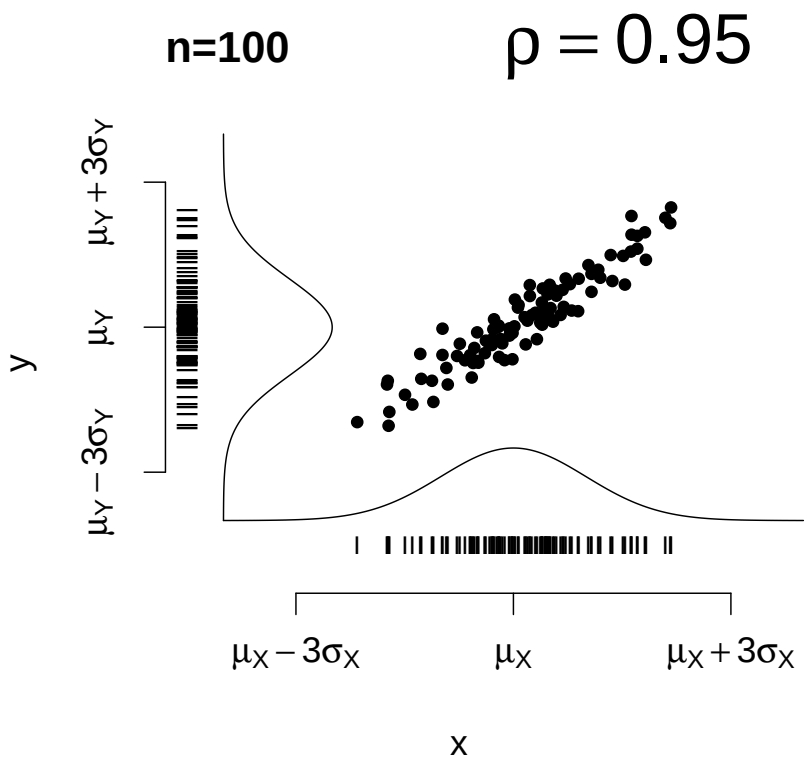
$$\rho := \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}}$$



- ρ positive: the product $(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])$ is positive in expectation. Naively: if X larger than its expectation, then tendentially also Y larger than its expectation, or the other way around, if X smaller than its expectation, then also Y tendentially smaller than its expectation

Examples: Correlation

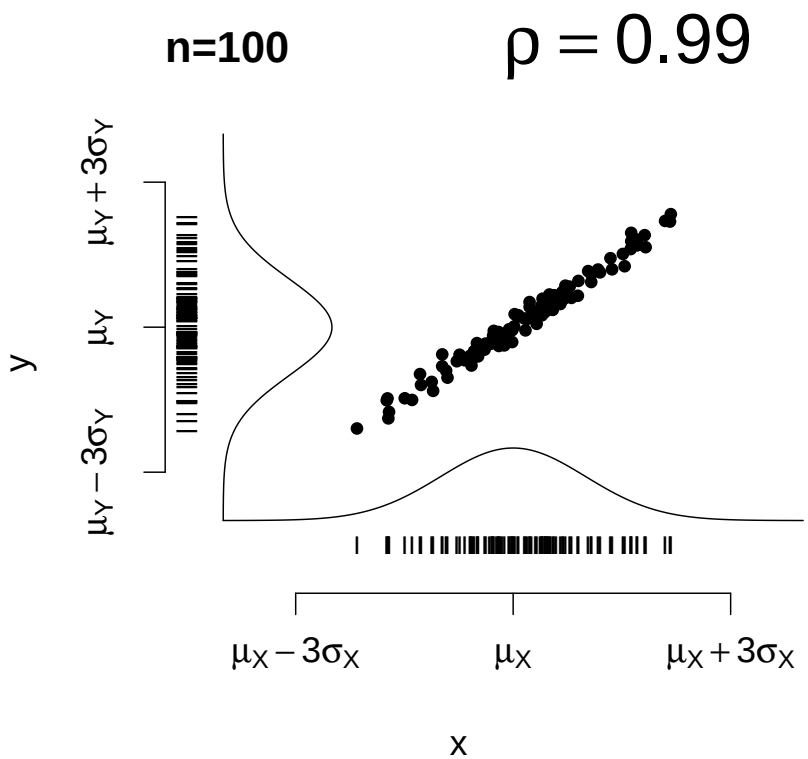
$$\rho := \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}}$$



- ρ positive: the product $(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])$ is positive in expectation. Naively: if X larger than its expectation, then tendentially also Y larger than its expectation, or the other way around, if X smaller than its expectation, then also Y tendentially smaller than its expectation

Examples: Correlation

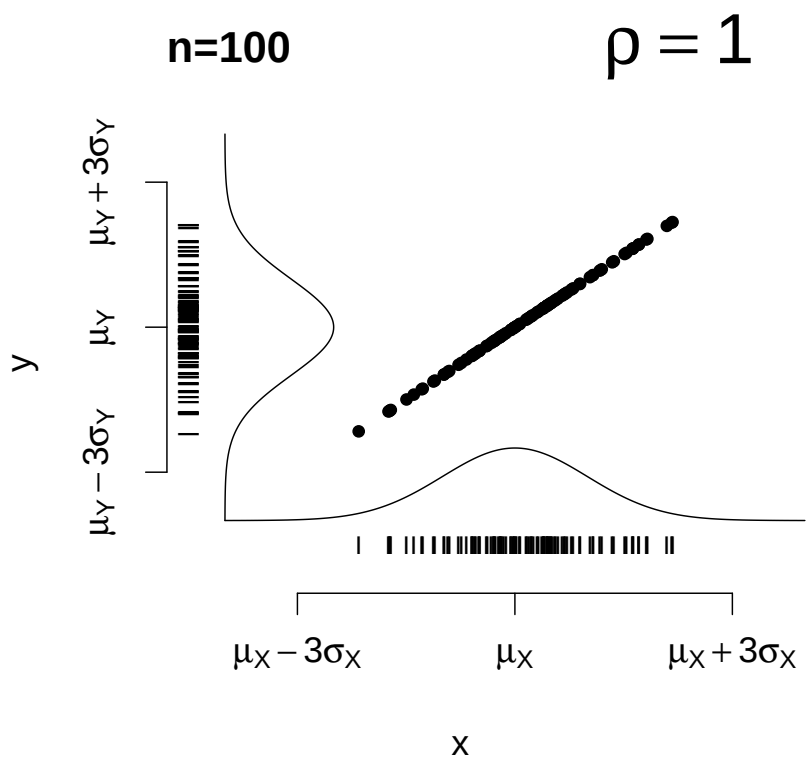
$$\rho := \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}}$$



- ρ positive: the product $(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])$ is positive in expectation. Naively: if X larger than its expectation, then tendentially also Y larger than its expectation, or the other way around, if X smaller than its expectation, then also Y tendentially smaller than its expectation

Examples: Correlation

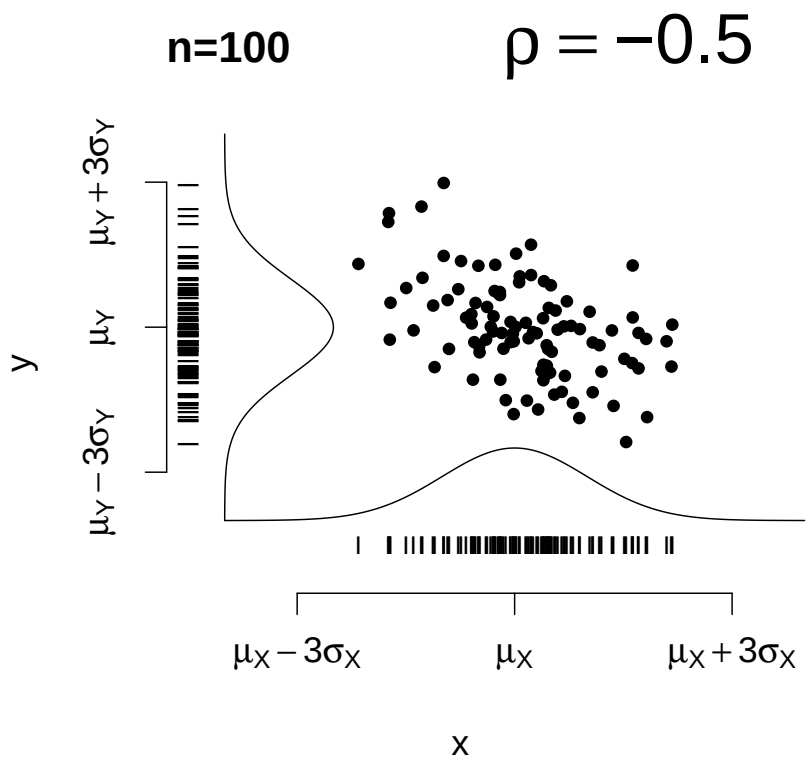
$$\rho := \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}}$$



- ρ positive: the product $(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])$ is positive in expectation. Naively: if X larger than its expectation, then tendentially also Y larger than its expectation, or the other way around, if X smaller than its expectation, then also Y tendentially smaller than its expectation

Examples: Correlation

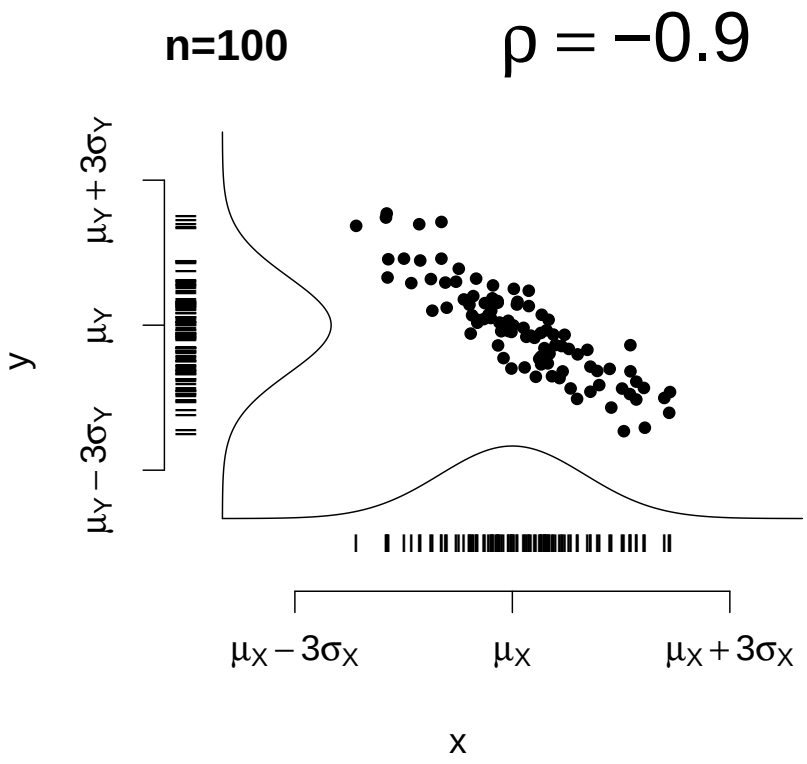
$$\rho := \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}}$$



- ρ negative: the product $(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])$ is negative in expectation. Naively: if X larger than its expectation, then tendentially Y smaller than its expectation, or vice versa

Examples: Correlation

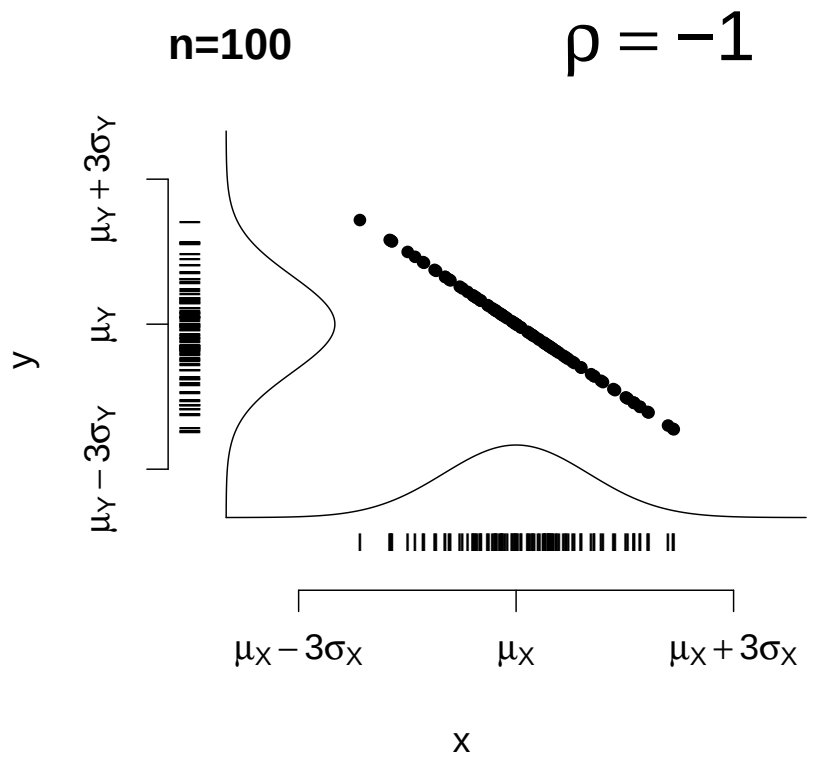
$$\rho := \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}}$$



- ρ negative: the product $(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])$ is negative in expectation. Naively: if X larger than its expectation, then tendentially Y smaller than its expectation, or vice versa

Examples: Correlation

$$\rho := \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}}$$

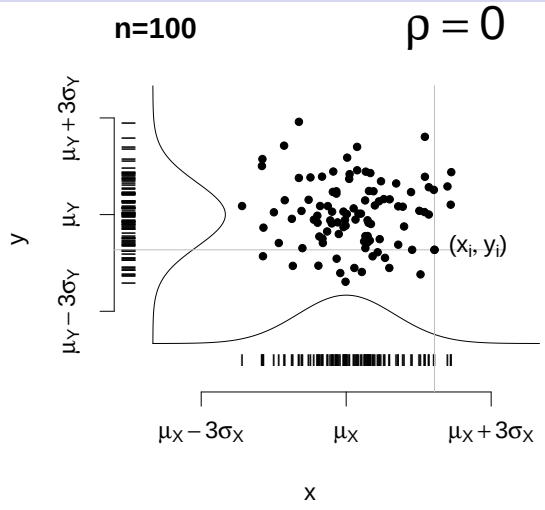


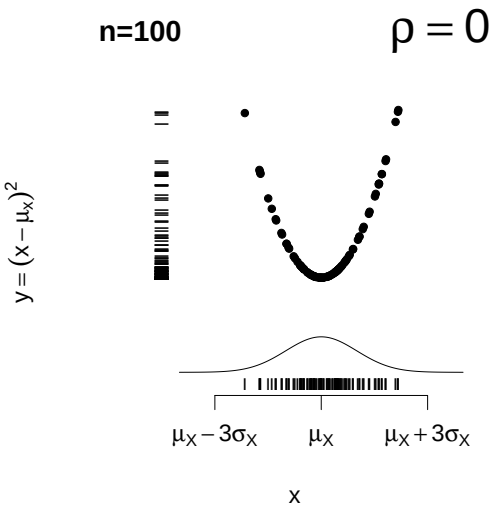
- ρ negative: the product $(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])$ is negative in expectation. Naively: if X larger than its expectation, then tendentially Y smaller than its expectation, or vice versa

Properties of the correlation

- The correlation ρ is a measure for the degree of the *linear* relation

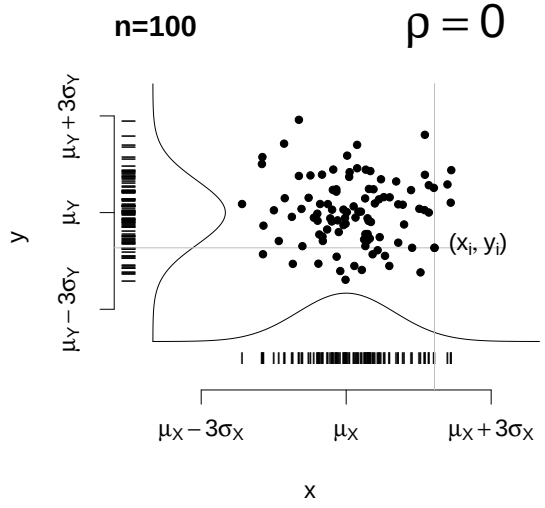
Properties of the correlation

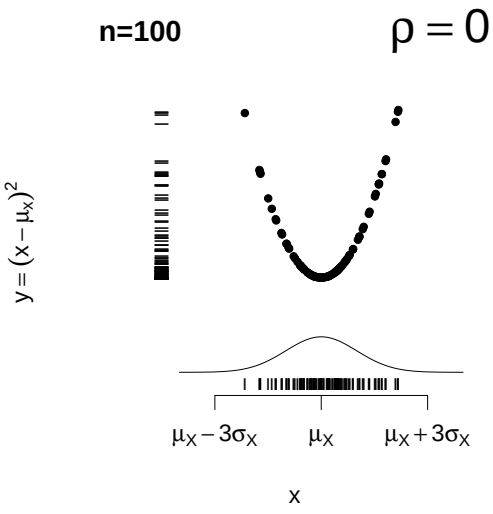


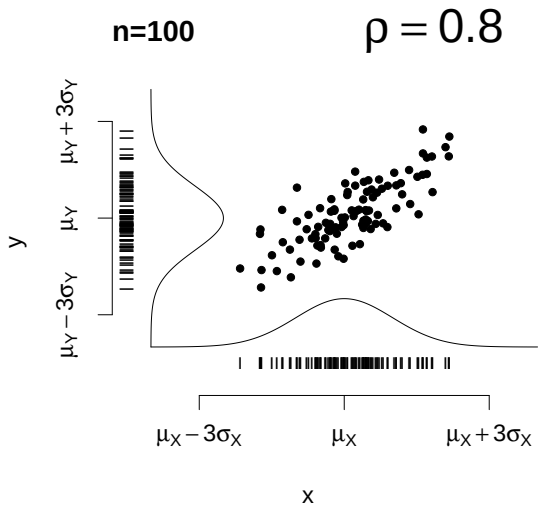


- The correlation ρ is a measure for the degree of the *linear* relation
- $\rho = 0 \Leftrightarrow$ no linear relation (say: X and Y are *uncorrelated*)

Properties of the correlation

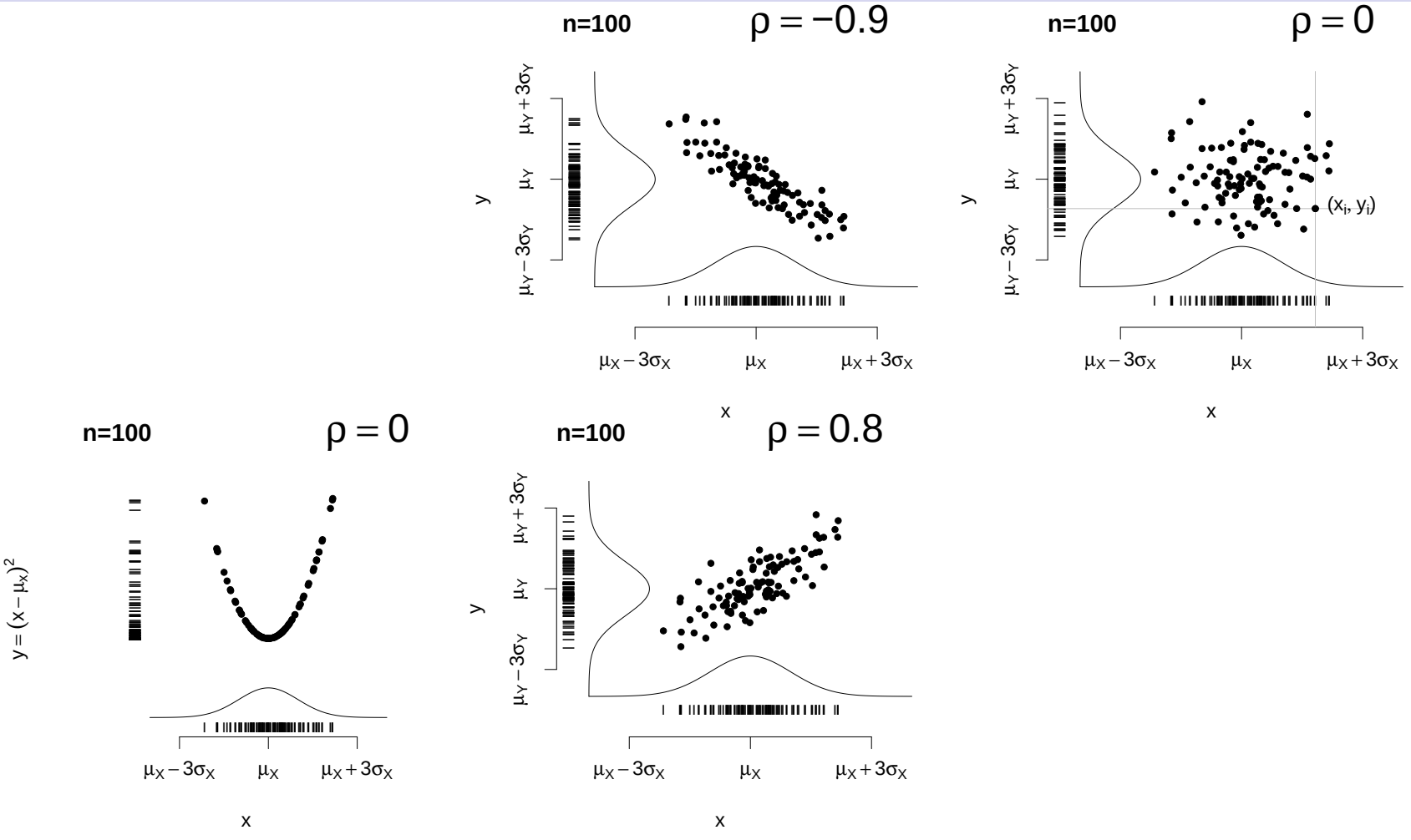






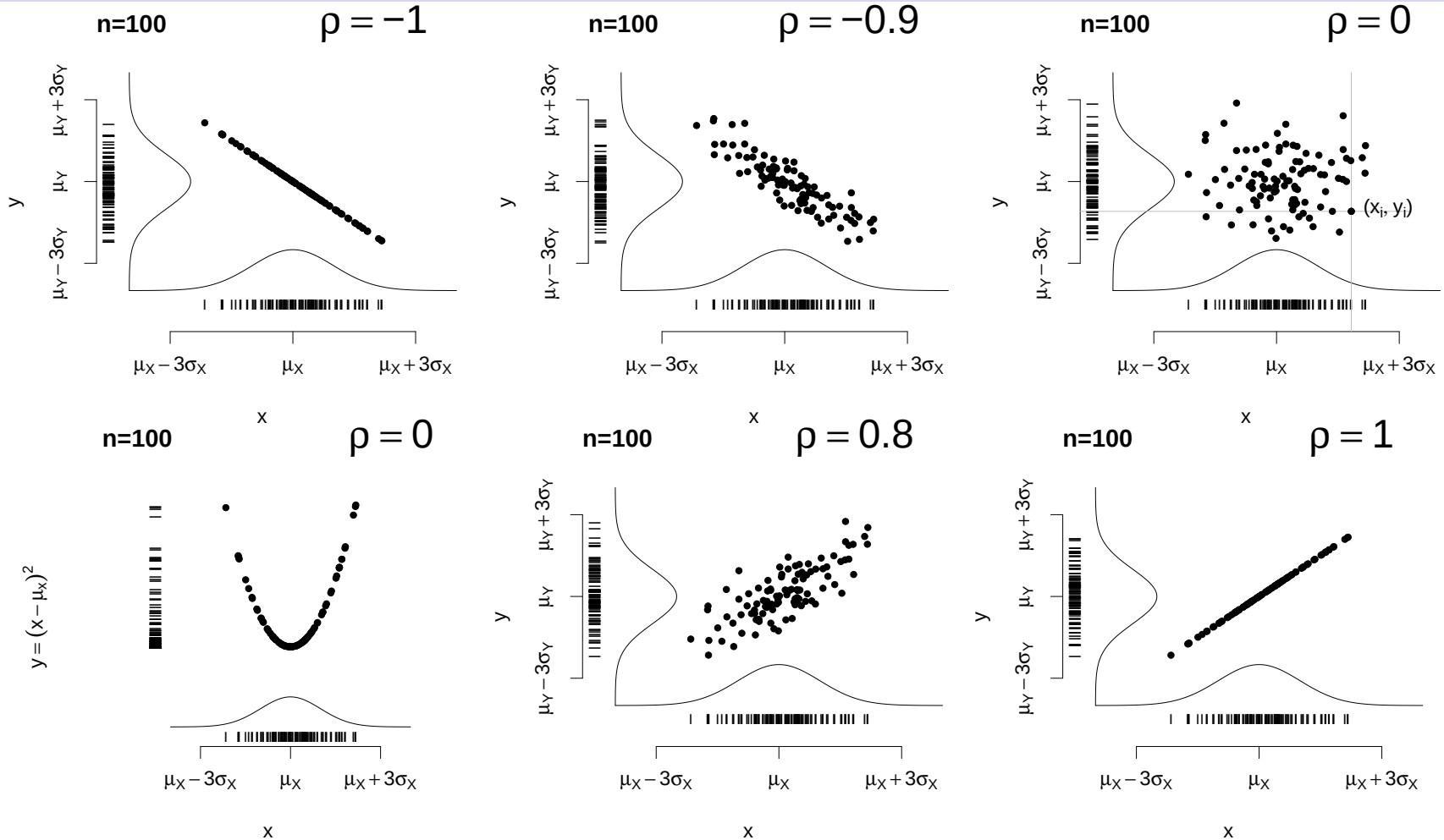
- The correlation ρ is a measure for the degree of the *linear* relation
- $\rho = 0 \Leftrightarrow$ no linear relation (say: X and Y are *uncorrelated*)
- $\rho > 0 \Leftrightarrow$ positive linear relation

Properties of the correlation



- The correlation ρ is a measure for the degree of the *linear* relation
- $\rho = 0 \Leftrightarrow$ no linear relation (say: X and Y are *uncorrelated*)
- $\rho > 0 \Leftrightarrow$ positive linear relation
- $\rho < 0 \Leftrightarrow$ negative linear relation

Properties of the correlation

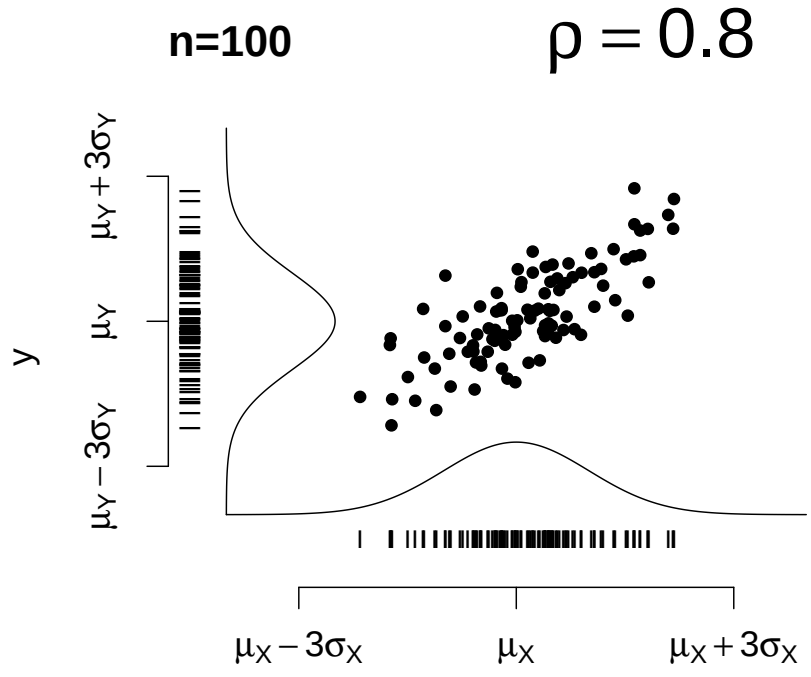


- The correlation ρ is a measure for the degree of the *linear* relation
- $\rho = 0 \Leftrightarrow$ no linear relation (say: X and Y are *uncorrelated*)
- $\rho > 0 \Leftrightarrow$ positive linear relation
- $\rho < 0 \Leftrightarrow$ negative linear relation
- $|\rho| = 1 \Leftrightarrow$ perfect linear relation

It holds $\rho \in [-1, 1]$

Empirical correlation

$$\rho := \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}}$$

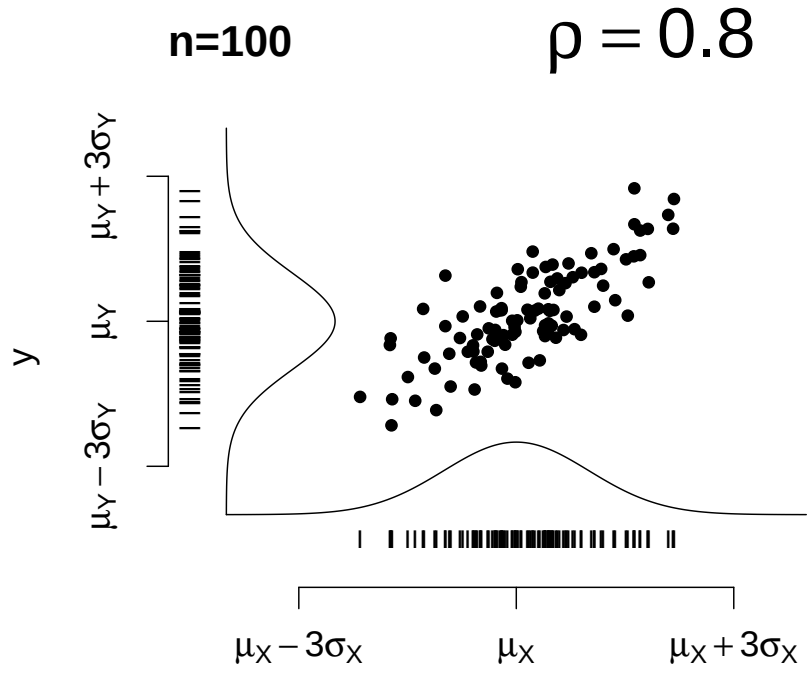


- For realizations $(x_i, y_i)_{i=1,2,\ldots,n}$ estimate ρ through the *empirical correlation*

$$r := \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y}$$

Empirical correlation

$$\rho := \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}}$$



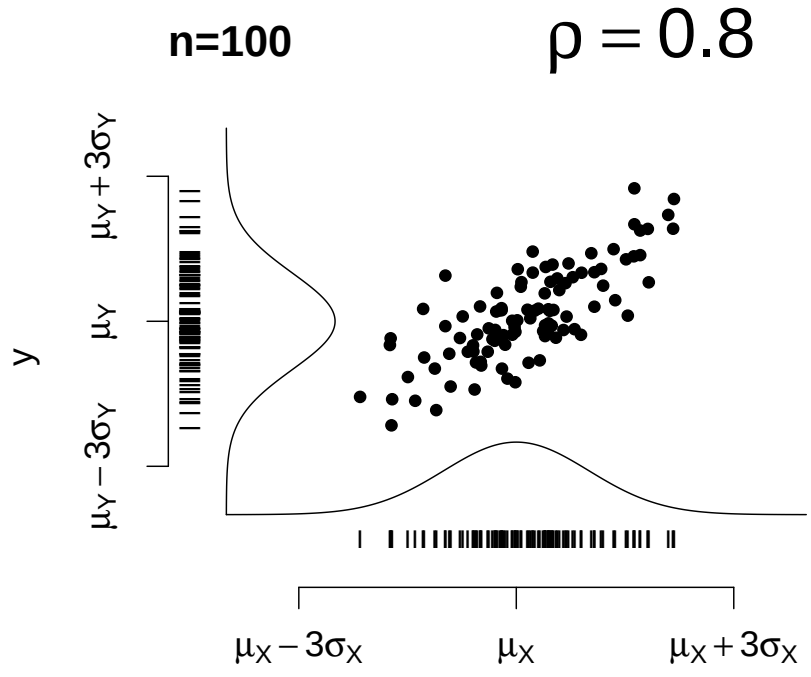
- For realizations $(x_i, y_i)_{i=1,2,\ldots,n}$ estimate ρ through the *empirical correlation*

$$r := \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y}$$

- in R via `cor()`

Empirical correlation

$$\rho := \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}}$$



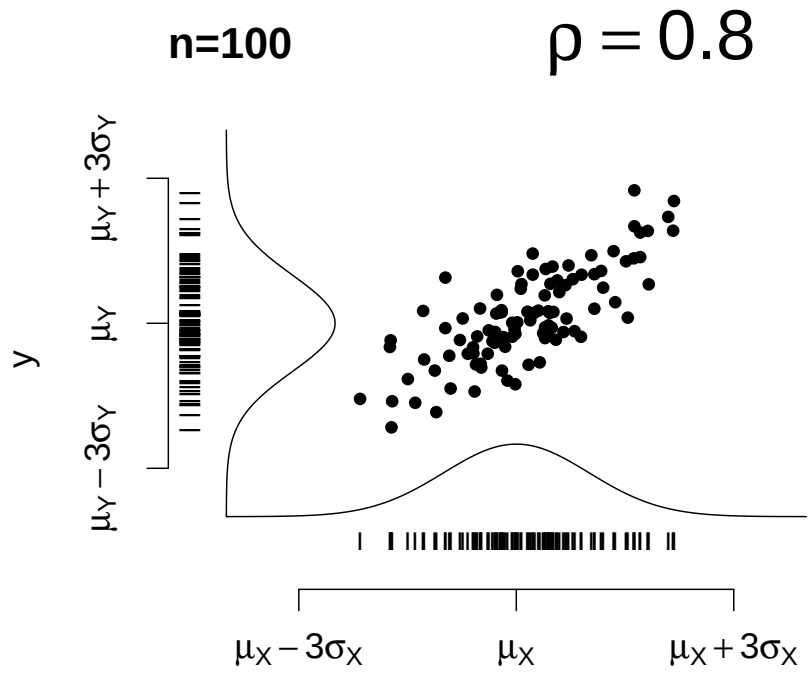
- For realizations $(x_i, y_i)_{i=1,2,\ldots,n}$ estimate ρ through the *empirical correlation*

$$r := \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y}$$

- in R via `cor()` here: $r \approx 0.78$

Empirical correlation

$$\rho := \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}}$$

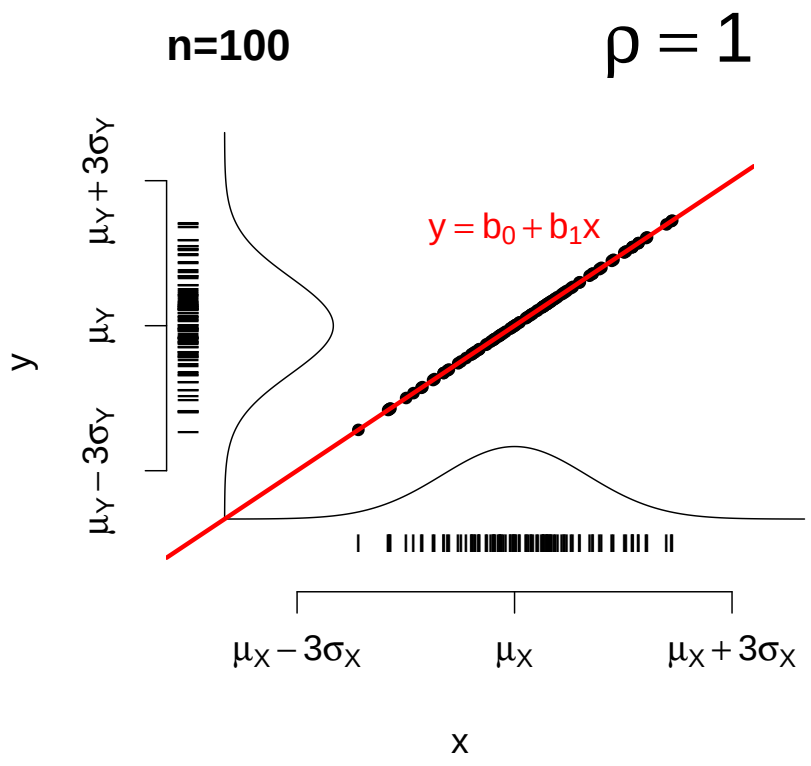


- For realizations $(x_i, y_i)_{i=1,2,\ldots,n}$ estimate ρ through the *empirical correlation*

$$r := \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y}$$

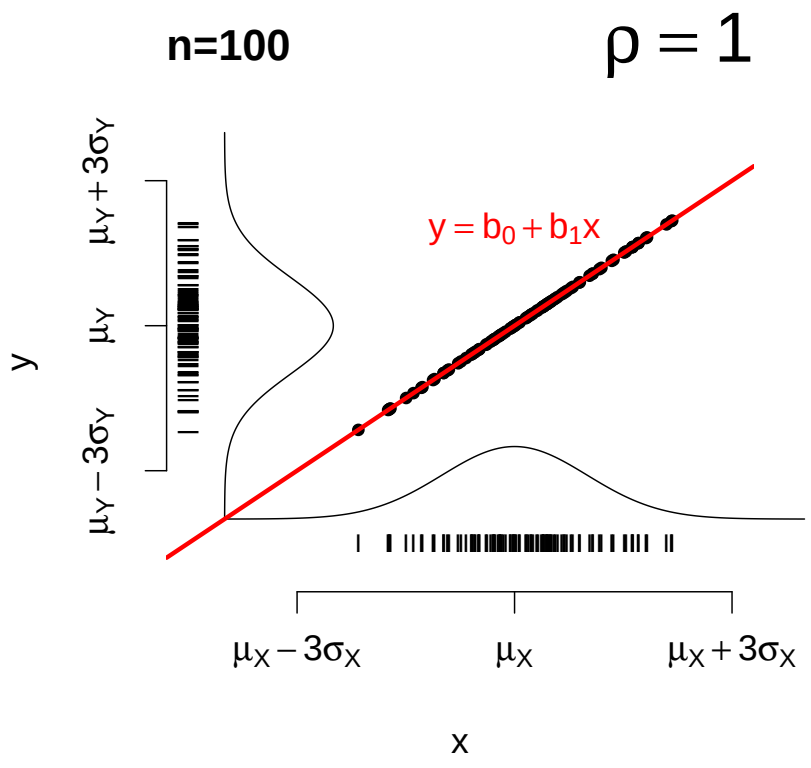
- in R via `cor()` here: $r \approx 0.78$ It is $r \in [-1, 1]$

Perfect linear relation for $\rho = 1$



- $\rho = 1$: the points lie on a line $y = b_0 + b_1 x$.

Perfect linear relation for $\rho = 1$

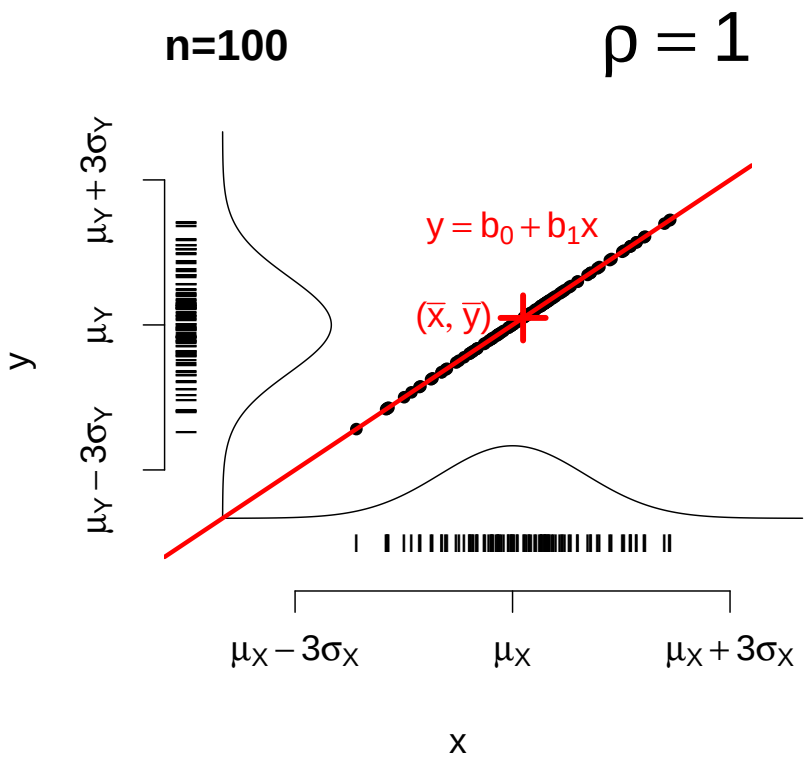


- $\rho = 1$: the points lie on a line $y = b_0 + b_1 x$.
- For the slope b_1 and the intercept b_0 it holds

$$b_1 = \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$

Perfect linear relation for $\rho = 1$

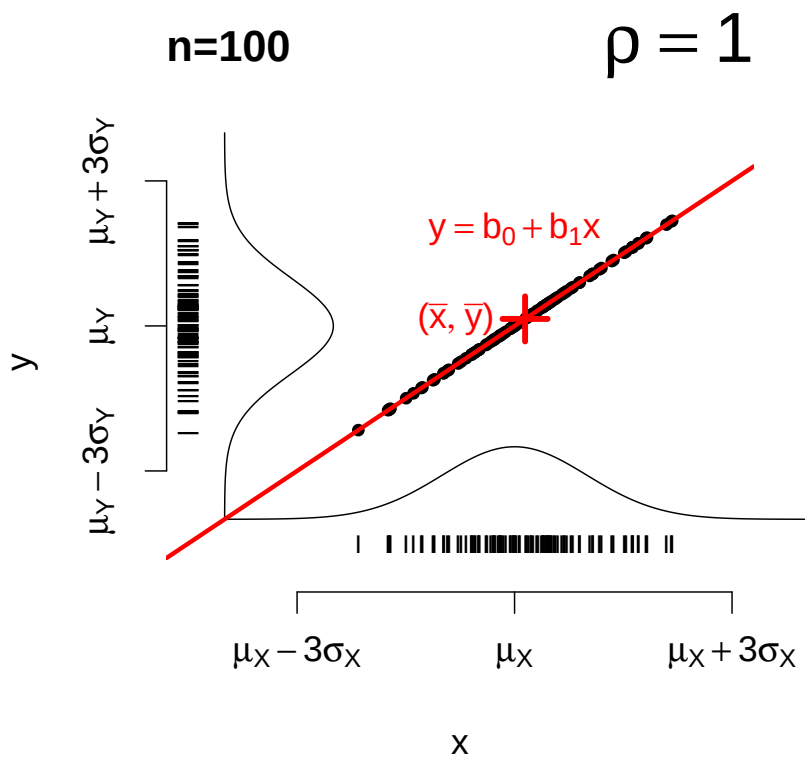
$$b_1 = \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$



- ad intercept b_0:

Perfect linear relation for $\rho = 1$

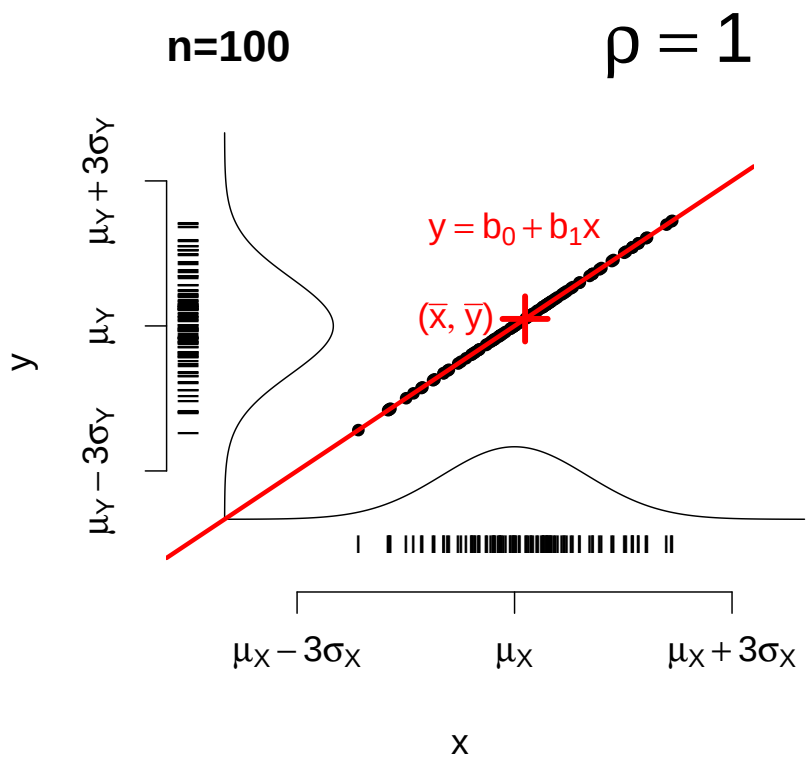
$$b_1 = \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$



- ad intercept b_0:
  - it is $y_i = b_0 + b_1 x_i$ for all $i = 1, 2, \ldots, n$

Perfect linear relation for $\rho = 1$

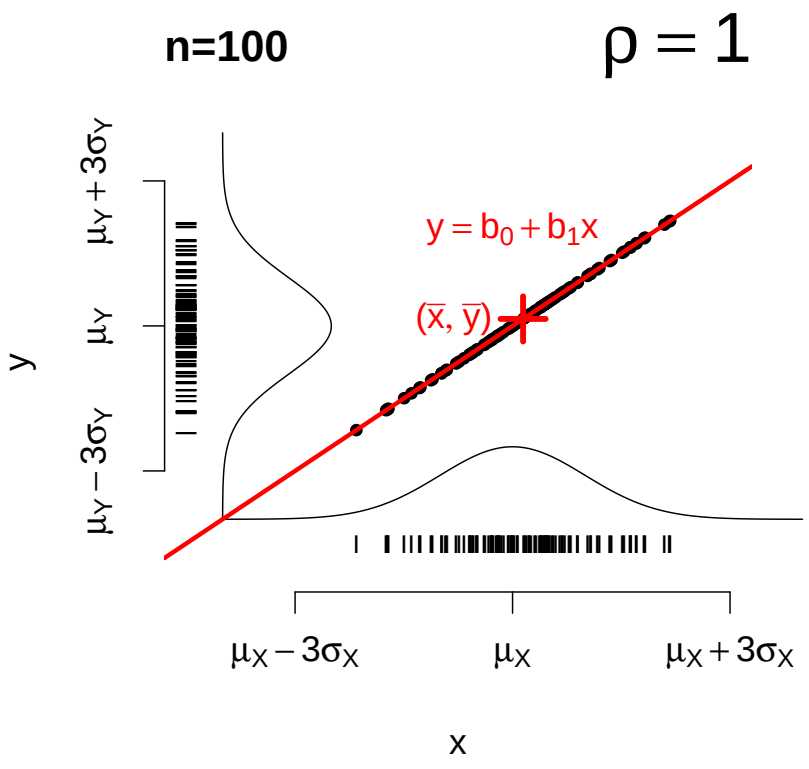
$$b_1 = \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$



- ad intercept b_0:
  - it is $y_i = b_0 + b_1 x_i$ for all $i = 1, 2, \ldots, n$
  - Summation: $\sum_i y_i = \sum_i (b_0 + b_1 x_i) = n b_0 + b_1 \sum_i x_i$

Perfect linear relation for $\rho = 1$

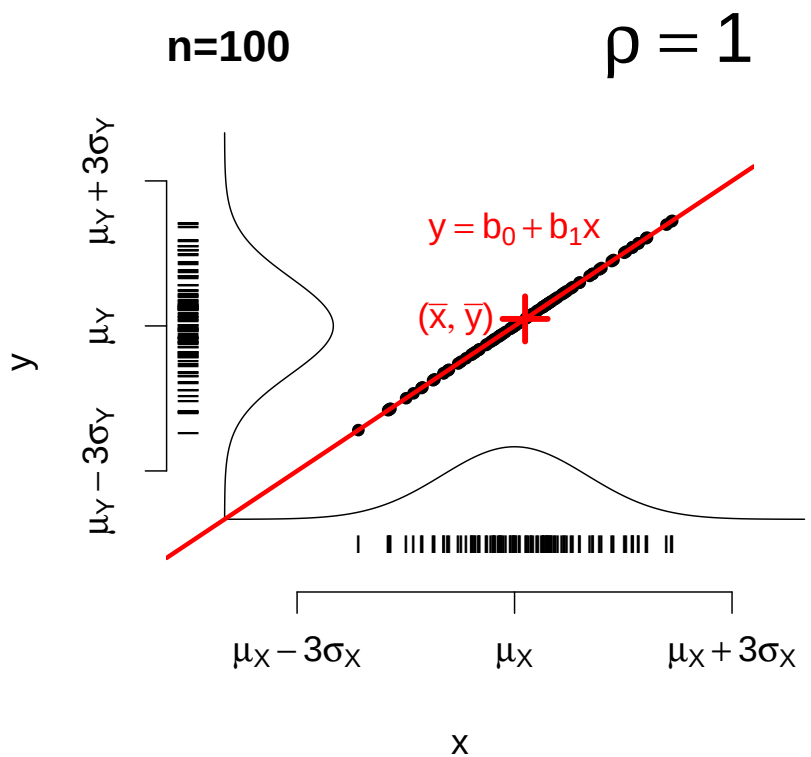
$$b_1 = \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$



- ad intercept b_0:
  - it is $y_i = b_0 + b_1 x_i$ for all $i = 1, 2, \ldots, n$
  - Summation: $\sum_i y_i = \sum_i (b_0 + b_1 x_i) = n b_0 + b_1 \sum_i x_i$
  - division through n yields: $\bar{y} = b_0 + b_1 \bar{x}$

Perfect linear relation for $\rho = 1$

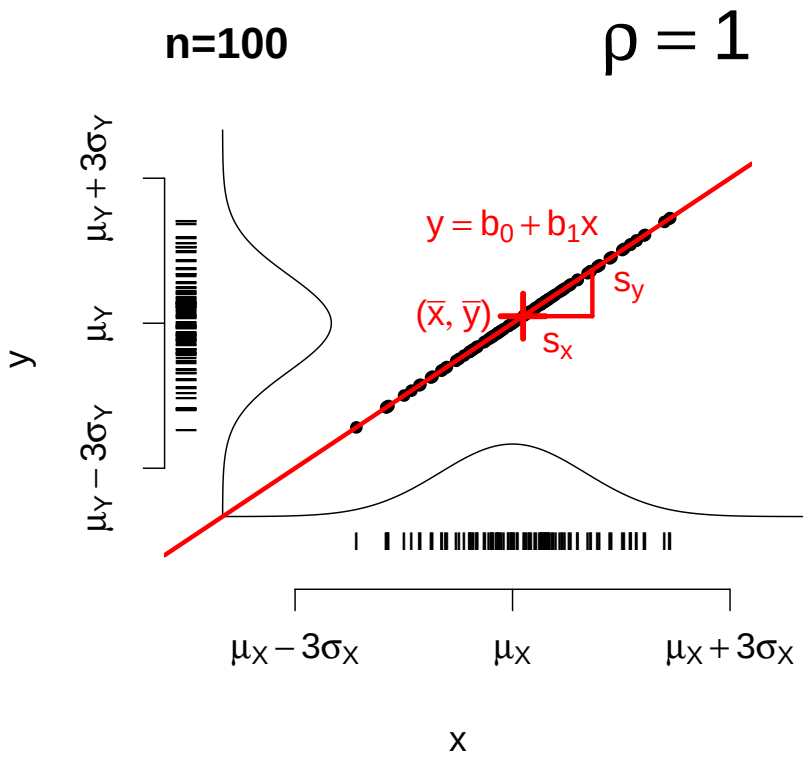
$$b_1 = \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$



- ad intercept b_0:
  - it is $y_i = b_0 + b_1 x_i$ for all $i = 1, 2, \ldots, n$
  - Summation: $\sum_i y_i = \sum_i (b_0 + b_1 x_i) = n b_0 + b_1 \sum_i x_i$
  - division through n yields: $\bar{y} = b_0 + b_1 \bar{x}$
- Thus: $b_0 = \bar{y} - b_1 \bar{x}$. Graphically: the line passes the center of mass $(\bar{x}, \bar{y})$

Perfect linear relation for $\rho = 1$

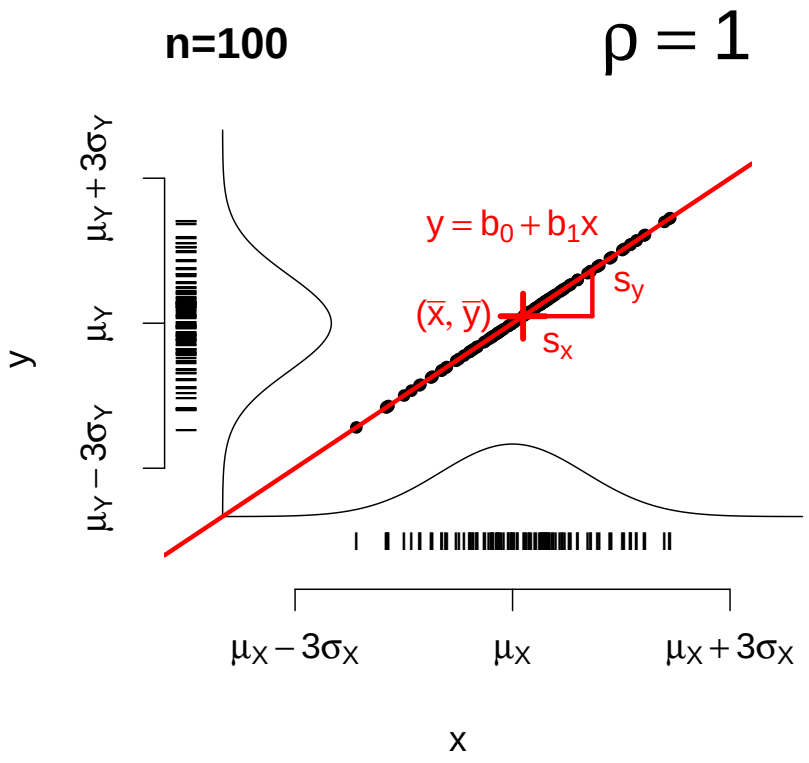
$$b_1 = \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$



- ad slope b_1:

Perfect linear relation for $\rho = 1$

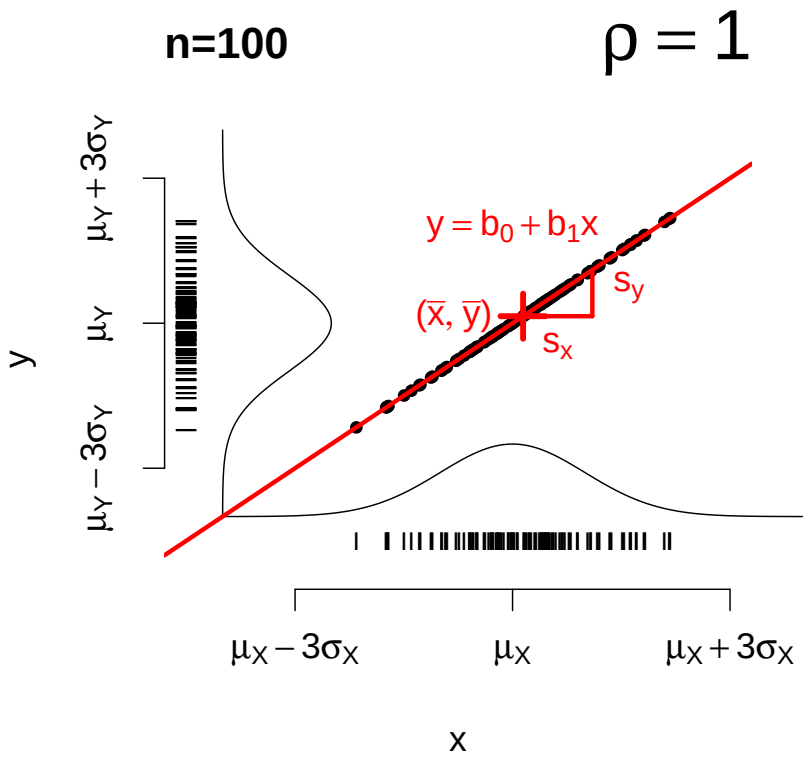
$$b_1 = \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$



- ad slope b_1:
  - it is $y_i = b_0 + b_1 x_i$ for all $i = 1, 2, \ldots, n$ as well as $\bar{y} = b_0 + b_1 \bar{x}$

Perfect linear relation for $\rho = 1$

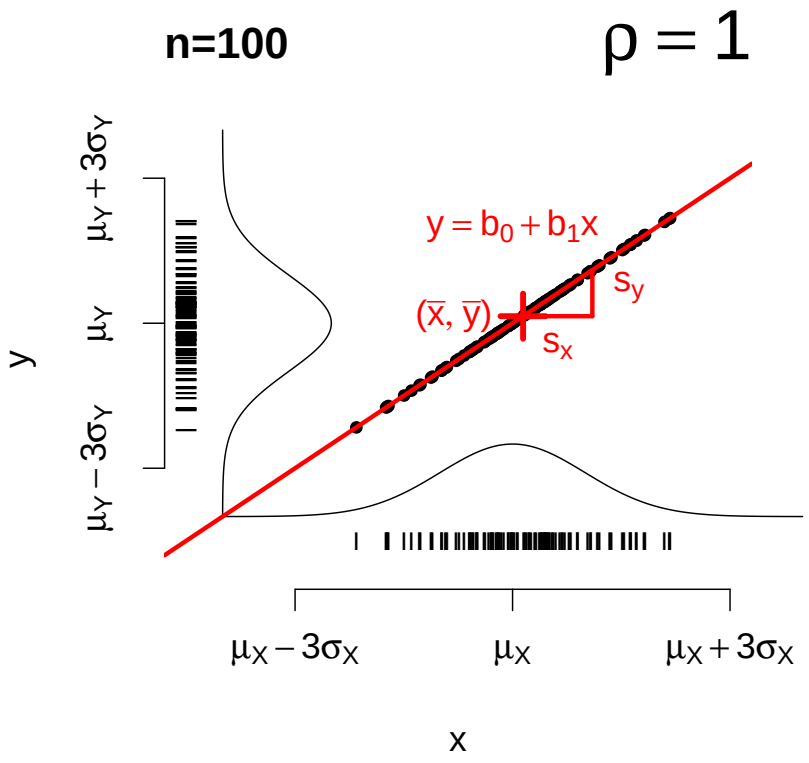
$$b_1 = \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$



- ad slope b_1:
  - it is $y_i = b_0 + b_1 x_i$ for all $i = 1, 2, \ldots, n$ as well as $\bar{y} = b_0 + b_1 \bar{x}$
  - difference and squaring: $(y_i - \bar{y})^2 = b_1^2 (x_i - \bar{x})^2$ for all $i = 1, 2, \ldots, n$

Perfect linear relation for $\rho = 1$

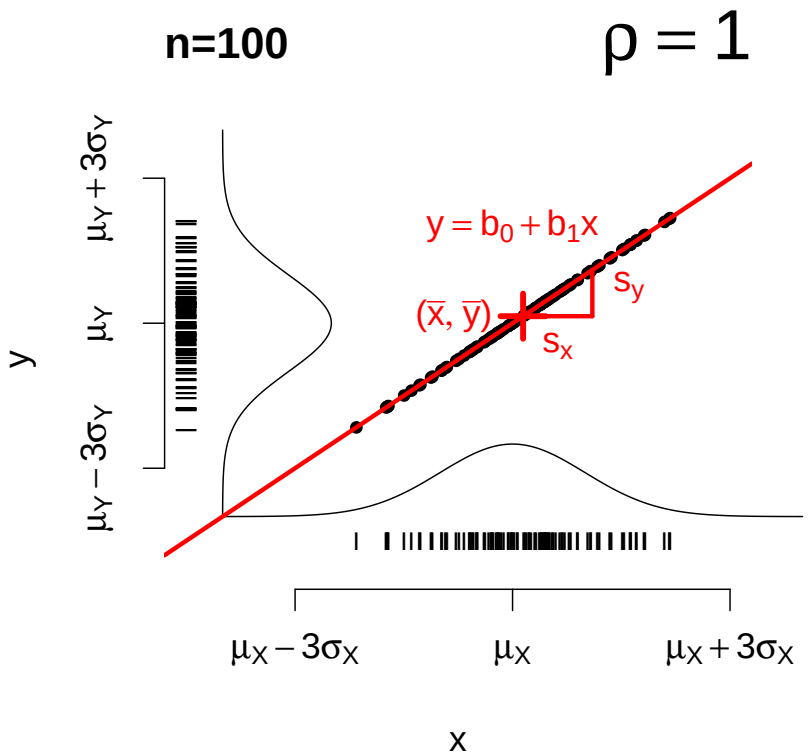
$$b_1 = \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$



- ad slope b_1:
  - it is $y_i = b_0 + b_1 x_i$ for all $i = 1, 2, \ldots, n$ as well as $\bar{y} = b_0 + b_1 \bar{x}$
  - difference and squaring: $(y_i - \bar{y})^2 = b_1^2 (x_i - \bar{x})^2$ for all $i = 1, 2, \ldots, n$
  - summation and division through $n - 1$: $\frac{1}{n-1} \sum_i (y_i - \bar{y})^2 = b_1^2 \cdot \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$

Perfect linear relation for $\rho = 1$

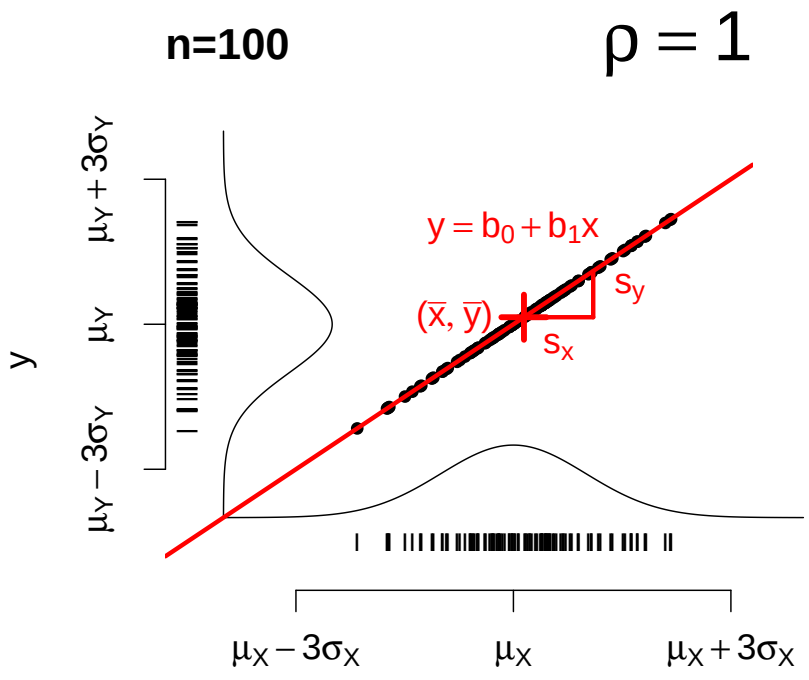
$$b_1 = \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$



- ad slope b_1:
  - it is $y_i = b_0 + b_1 x_i$ for all $i = 1, 2, \ldots, n$ as well as $\bar{y} = b_0 + b_1 \bar{x}$
  - difference and squaring: $(y_i - \bar{y})^2 = b_1^2 (x_i - \bar{x})^2$ for all $i = 1, 2, \ldots, n$
  - summation and division through $n - 1$: $\frac{1}{n-1} \sum_i (y_i - \bar{y})^2 = b_1^2 \cdot \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$
  - square-root: $b_1 = s_y / s_x$

Perfect linear relation for $\rho = 1$

$$b_1 = \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$

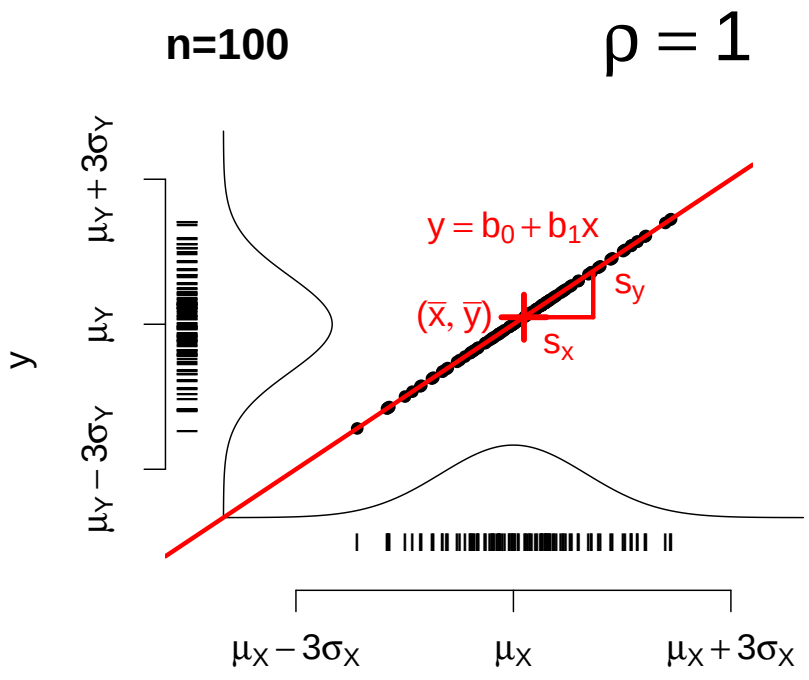


In summary:

- y_i is 'explained' by x_i

Perfect linear relation for $\rho = 1$

$$b_1 = \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$

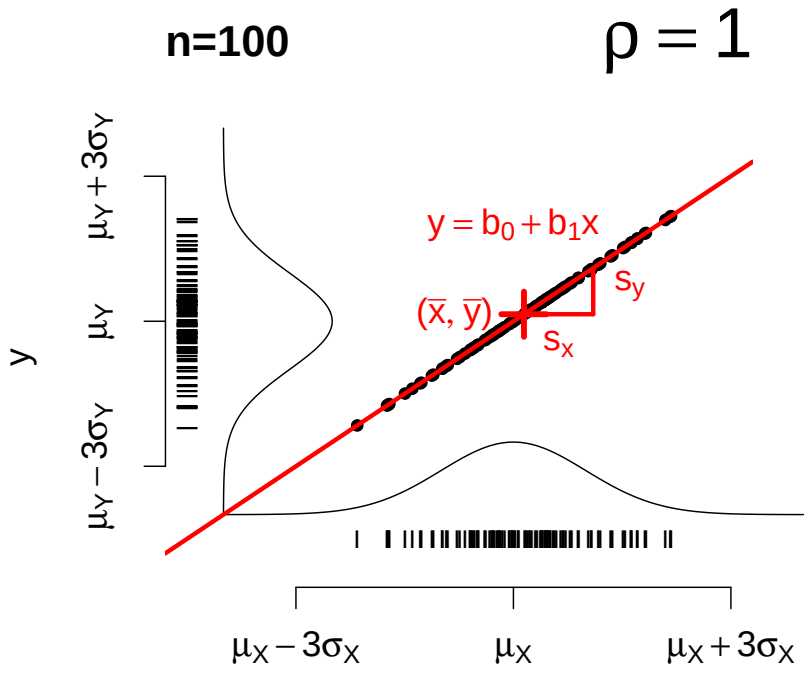


In summary:

- y_i is 'explained' by x_i
- the line passes the center of mass $(\bar{x}, \bar{y})$

Perfect linear relation for $\rho = 1$

$$b_1 = \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$

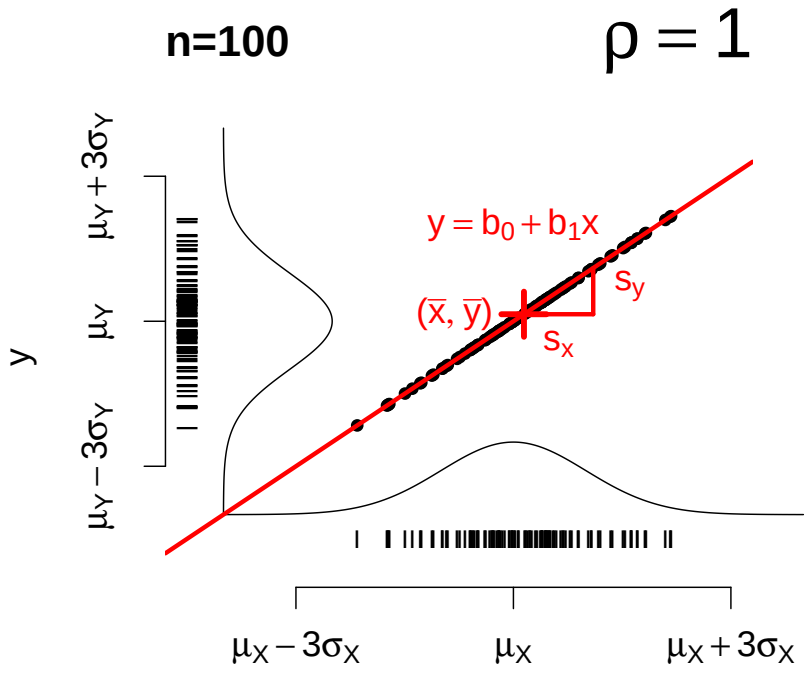


In summary:

- y_i is 'explained' by x_i
- the line passes the center of mass $(\bar{x}, \bar{y})$
- regarding the slope think in the standard deviations

Perfect linear relation for $\rho = 1$

$$b_1 = \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$

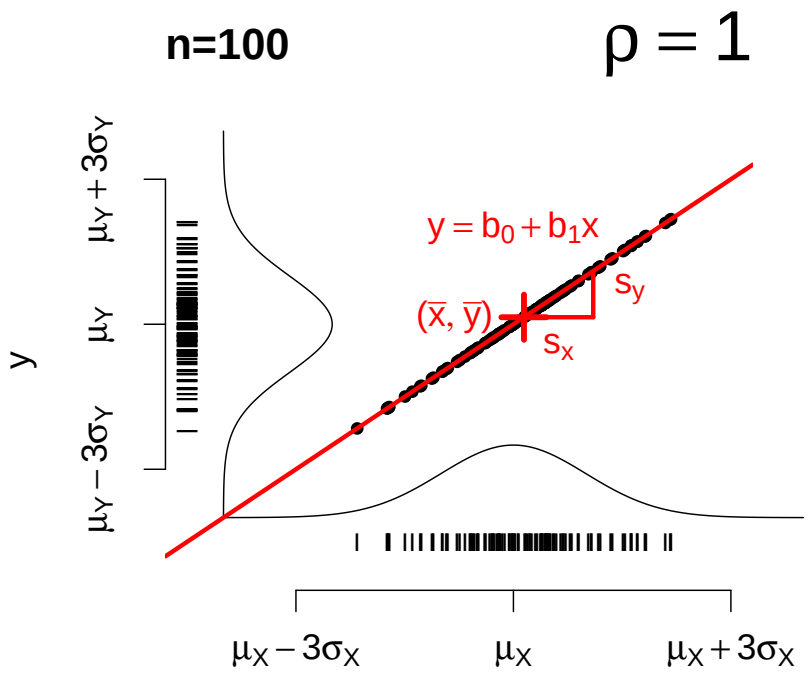


In summary:

- y_i is 'explained' by x_i
- the line passes the center of mass $(\bar{x}, \bar{y})$
- regarding the slope think in the standard deviations
  - one step to the right of size s_x results in an increase of size s_y

Perfect linear relation for $\rho = 1$

$$b_1 = \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$

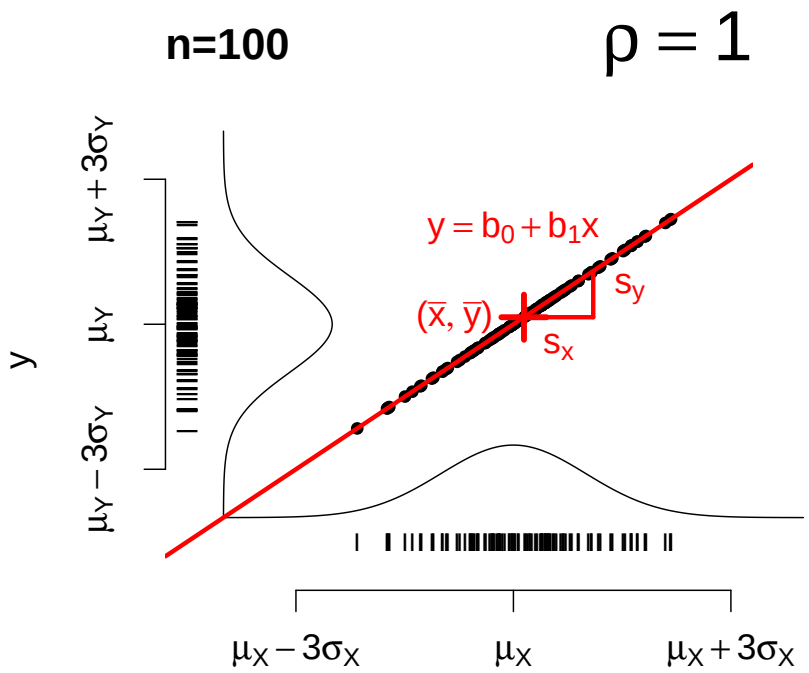


In summary:

- y_i is 'explained' by x_i
- the line passes the center of mass $(\bar{x}, \bar{y})$
- regarding the slope think in the standard deviations
  - one step to the right of size s_x results in an increase of size s_y
  - but this particular slope is a consequence of the special case $\rho = 1$

Perfect linear relation for $\rho = 1$

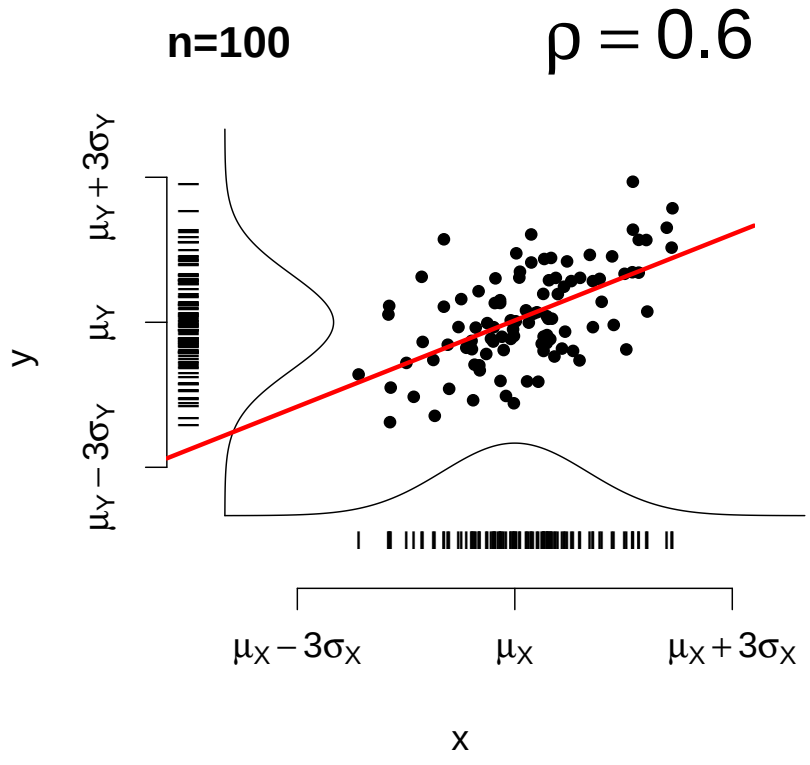
$$b_1 = \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$



In summary:

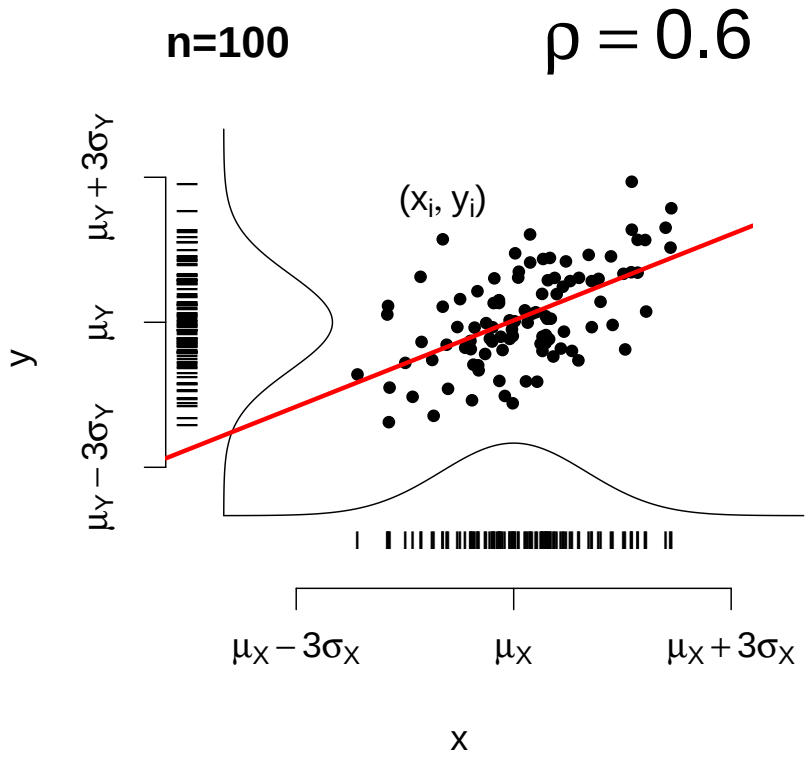
- y_i is 'explained' by x_i
- the line passes the center of mass $(\bar{x}, \bar{y})$
- regarding the slope think in the standard deviations
  - one step to the right of size s_x results in an increase of size s_y
  - but this particular slope is a consequence of the special case $\rho = 1$
  - general ρ induces the factor r (empirical correlation)...

General: linear relation plus error



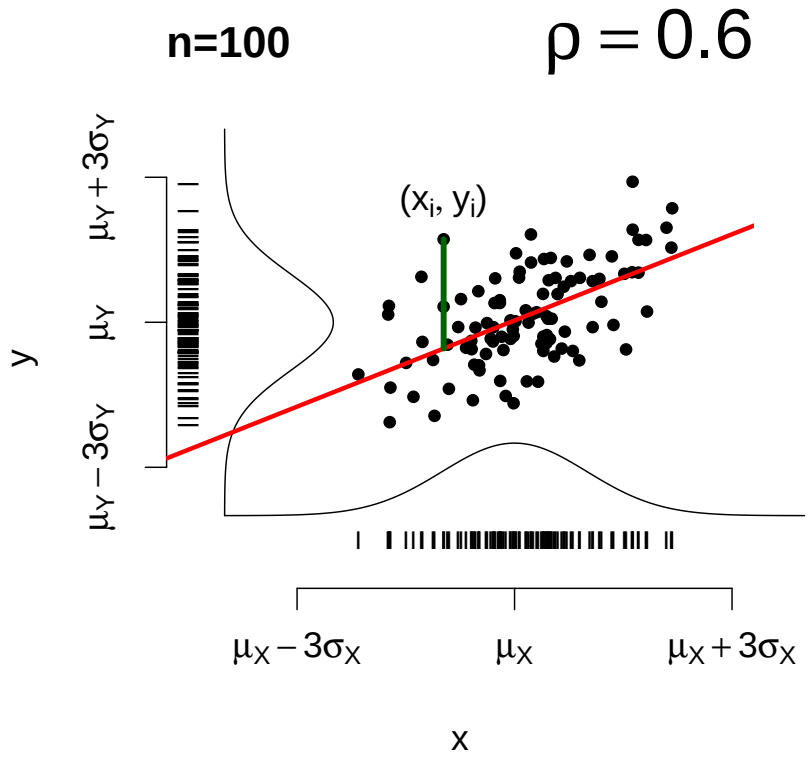
- For $|\rho| \neq 1$ the relation is 'only' *approximately* linear

General: linear relation plus error



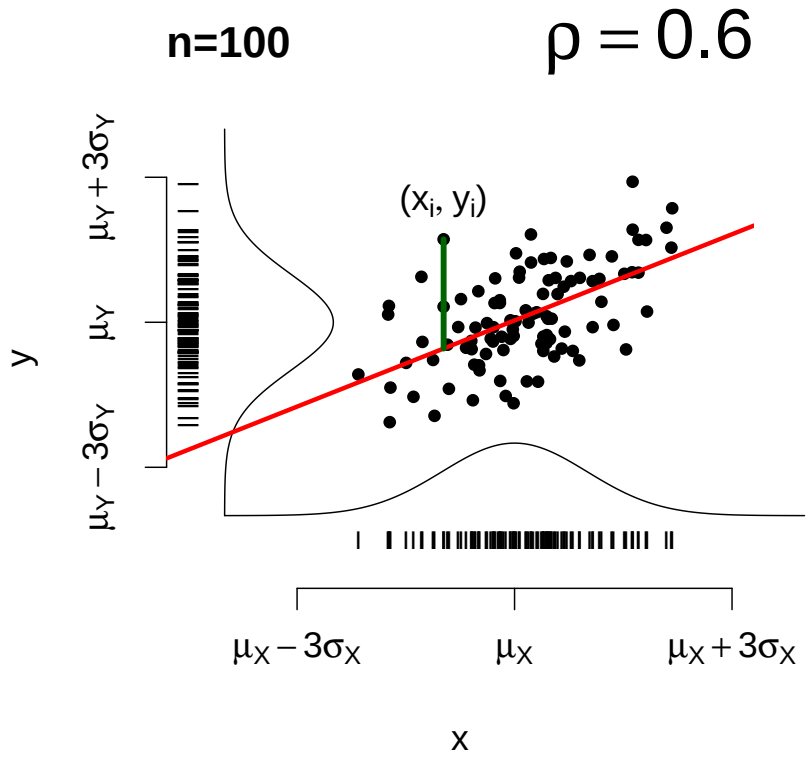
- For $|\rho| \neq 1$ the relation is 'only' *approximately* linear
- $y_i = \beta_0 + \beta_1 x_i$

General: linear relation plus error



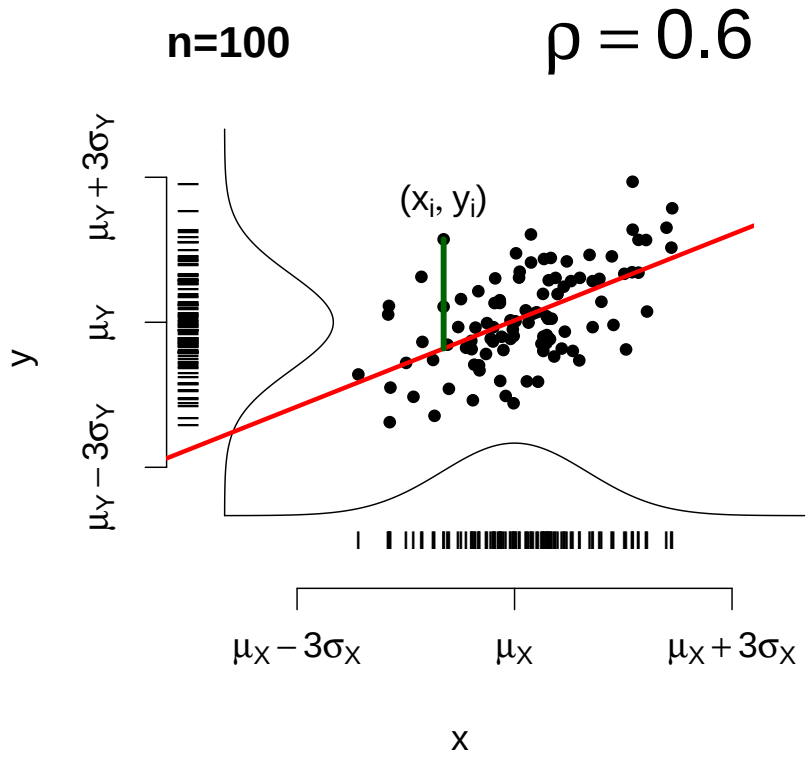
- For $|\rho| \neq 1$ the relation is 'only' *approximately* linear
- $y_i = \beta_0 + \beta_1 x_i + e_i$

General: linear relation plus error



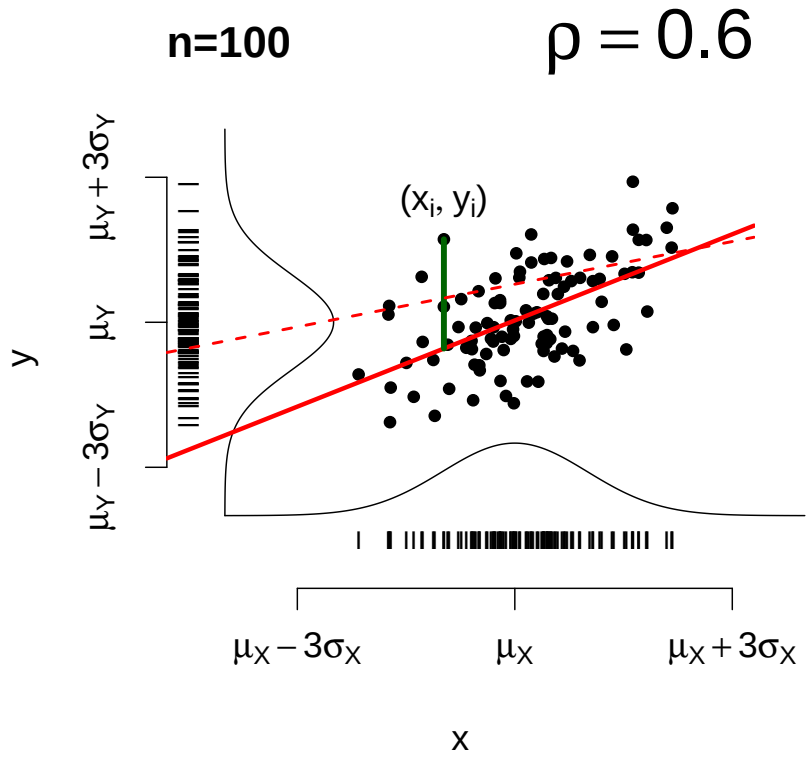
- For $|\rho| \neq 1$ the relation is 'only' *approximately* linear
- $y_i = \beta_0 + \beta_1 x_i + e_i$
  - while e_i is denoted the (i-th) *error*, respectively, the (i-th) *residual*

General: linear relation plus error



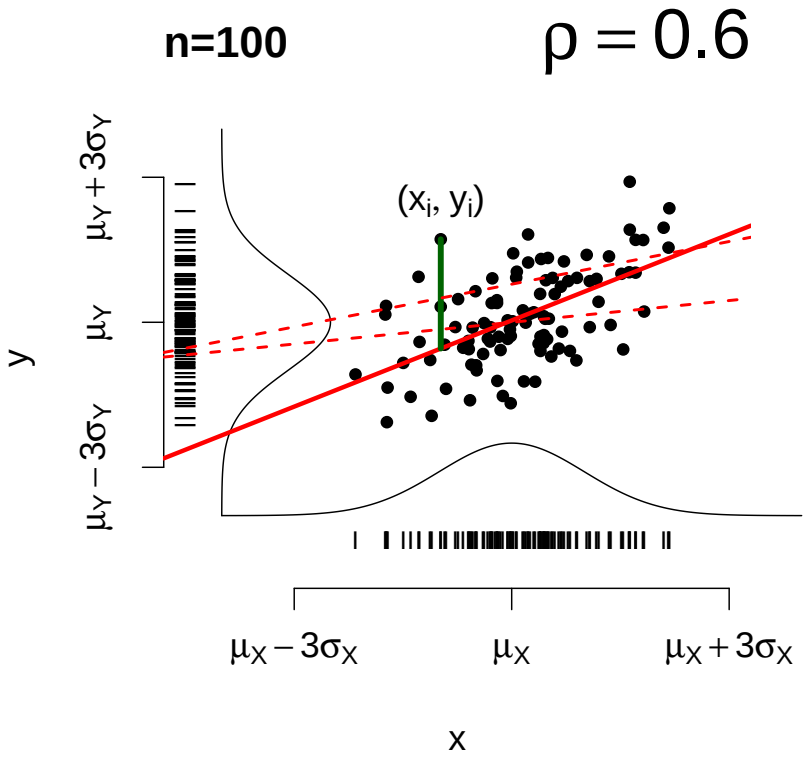
- For $|\rho| \neq 1$ the relation is 'only' *approximately* linear
- $y_i = \beta_0 + \beta_1 x_i + e_i$
  - while e_i is denoted the (i-th) *error*, respectively, the (i-th) *residual*
  - thus the assumed relation is: linear proportion plus error

Regression line google: C. F. Gauß



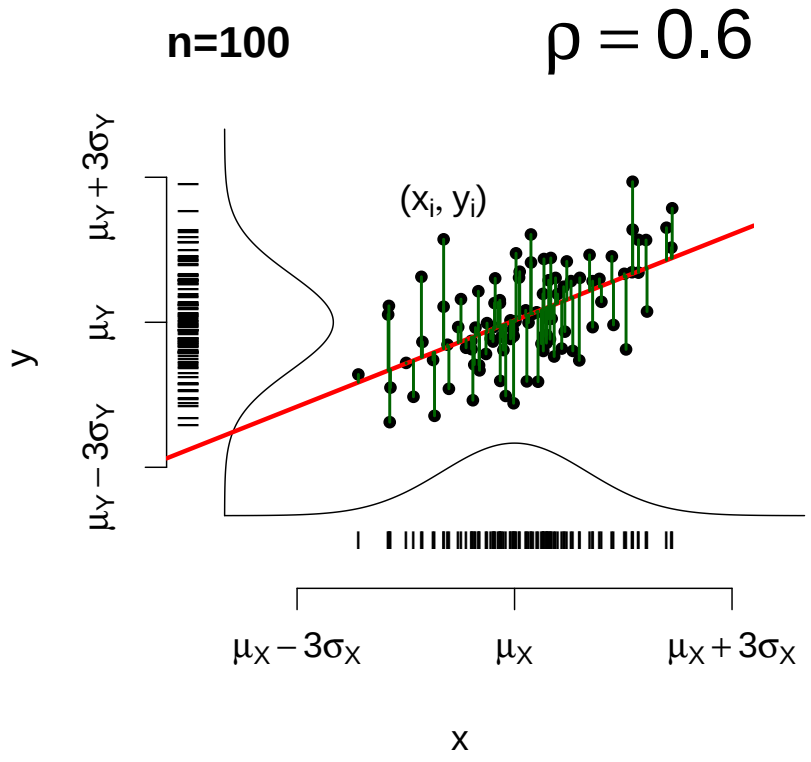
- $y_i = \beta_0 + \beta_1 x_i + e_i$
- in principle there are many possible lines

Regression line google: C. F. Gauß



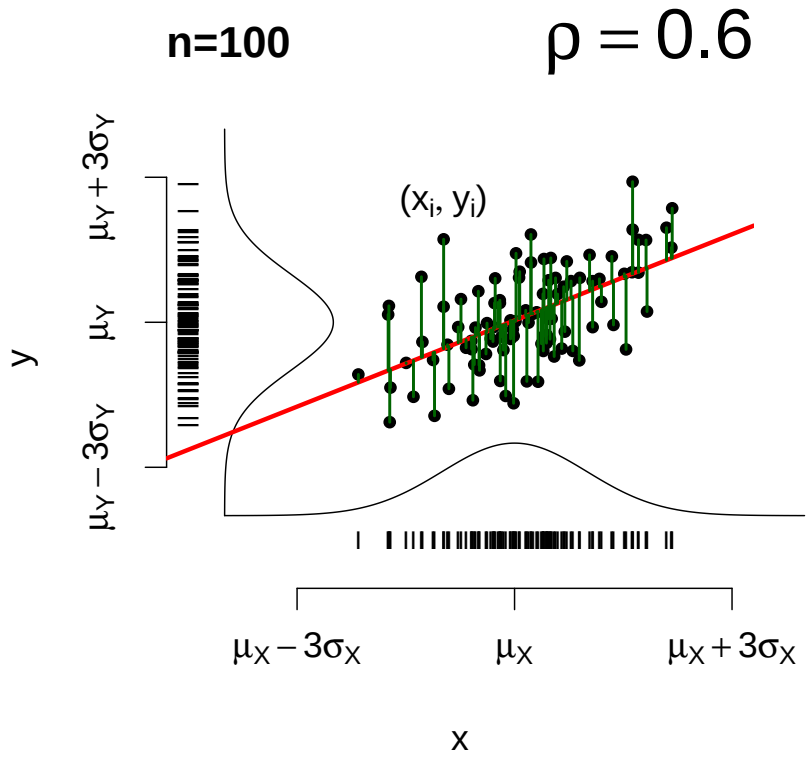
- $y_i = \beta_0 + \beta_1 x_i + e_i$
- in principle there are many possible lines

Regression line google: C. F. Gauß



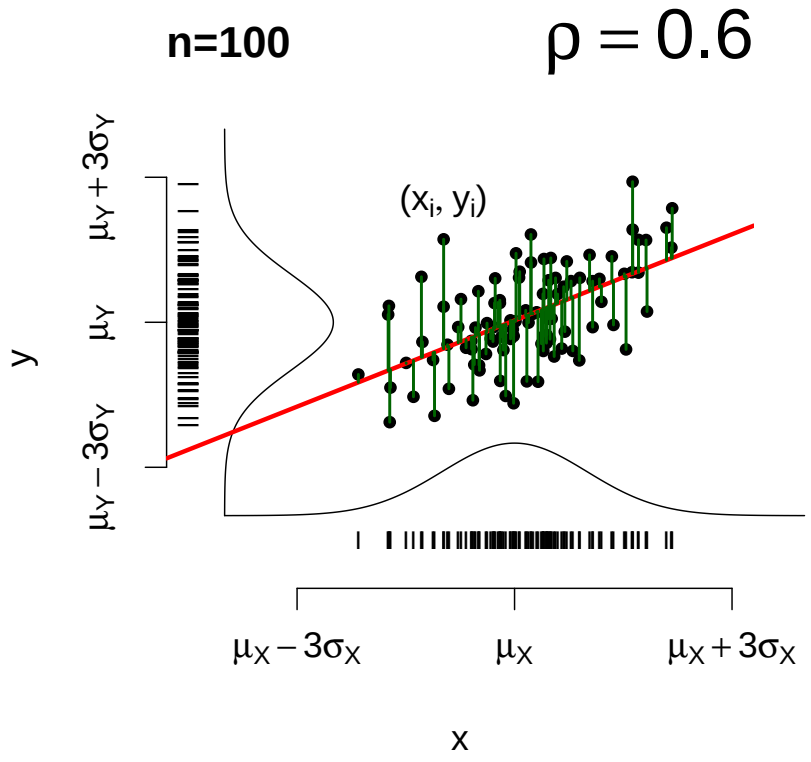
- $y_i = \beta_0 + \beta_1 x_i + e_i$
- in principle there are many possible lines
- Definition: the line, which minimizes the sum of squares of the residuals, is called the *regression line*

Regression line google: C. F. Gauß



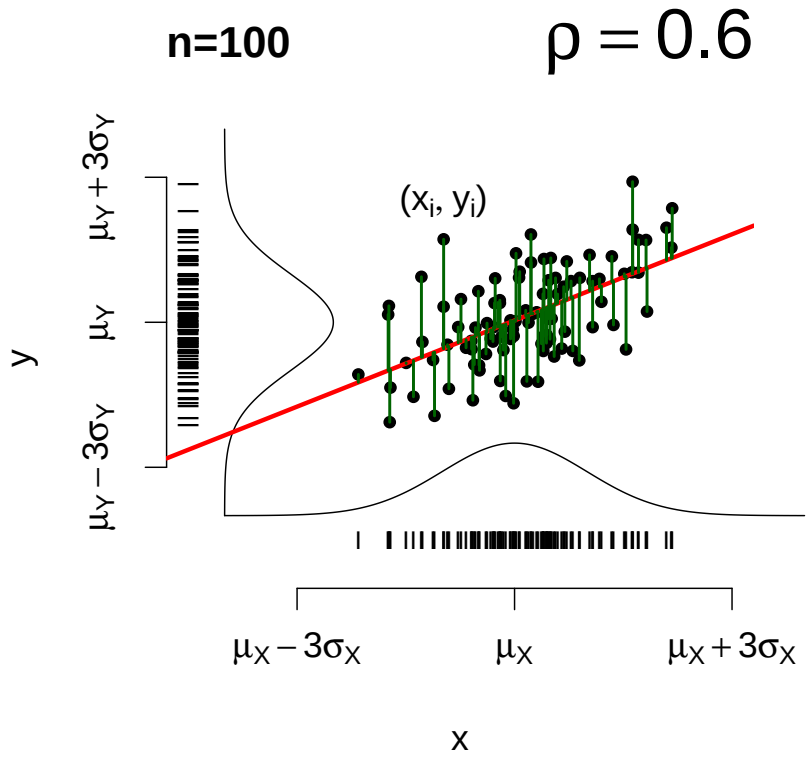
- $y_i = \beta_0 + \beta_1 x_i + e_i$
- in principle there are many possible lines
- Definition: the line, which minimizes the sum of squares of the residuals, is called the *regression line*
- i.e., search β_0 and β_1 such that $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_i])^2$ minimal

Regression line google: C. F. Gauß



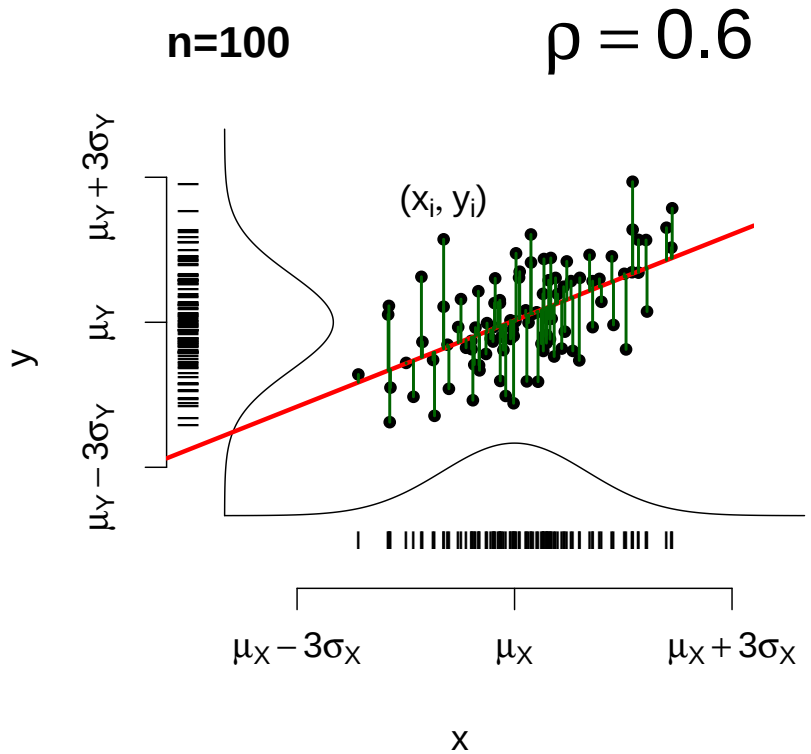
- $y_i = \beta_0 + \beta_1 x_i + e_i$
- in principle there are many possible lines
- Definition: the line, which minimizes the sum of squares of the residuals, is called the *regression line*
- i.e., search β_0 and β_1 such that $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_i])^2$ minimal
- the minimizers b_0 and b_1 yield the regression line $y = b_0 + b_1 x$

Regression line google: C. F. Gauß



- $y_i = \beta_0 + \beta_1 x_i + e_i$
- in principle there are many possible lines
- Definition: the line, which minimizes the sum of squares of the residuals, is called the *regression line*
- i.e., search β_0 and β_1 such that $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_i])^2$ minimal
- the minimizers b_0 and b_1 yield the regression line $y = b_0 + b_1 x$
- procedure called *'method of least squares'*

Regression line google: C. F. Gauß



- $y_i = \beta_0 + \beta_1 x_i + e_i$
- in principle there are many possible lines
- Definition: the line, which minimizes the sum of squares of the residuals, is called the *regression line*
- i.e., search β_0 and β_1 such that $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_i])^2$ minimal
- the minimizers b_0 and b_1 yield the regression line $y = b_0 + b_1 x$
- procedure called *'method of least squares'*
- the estimators b_0 and b_1 are the *least-squares estimators* for β_0 and β_1

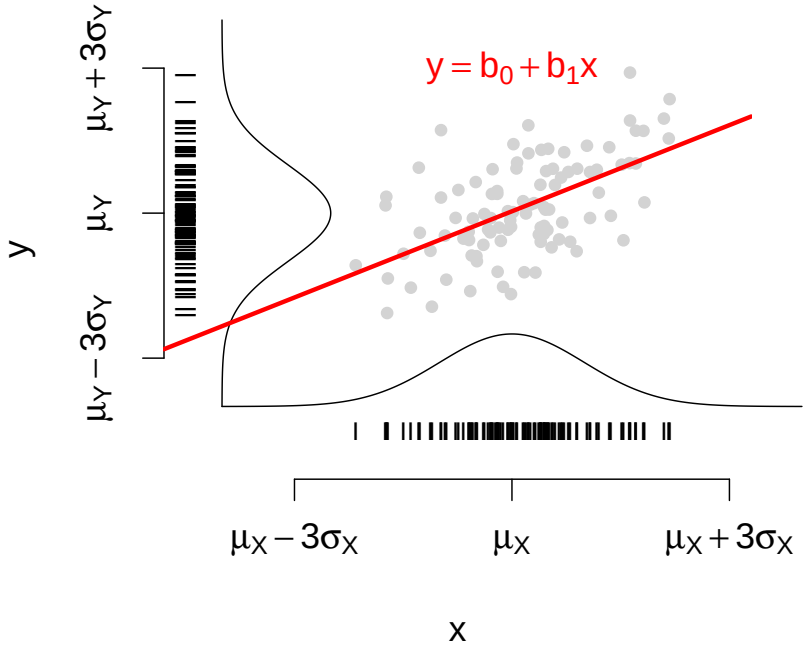
  greek $\beta_j \leftrightarrow$ parameters ('unknown'), latin $b_j \leftrightarrow$ statistics / estimators ('known', functions of the $(x_i, y_i)_i$)

Regression line: b_0 and b_1

For the slope and the intercept of the regression line it holds

$$b_1 = r \cdot \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$

n=100 $\rho = 0.6$



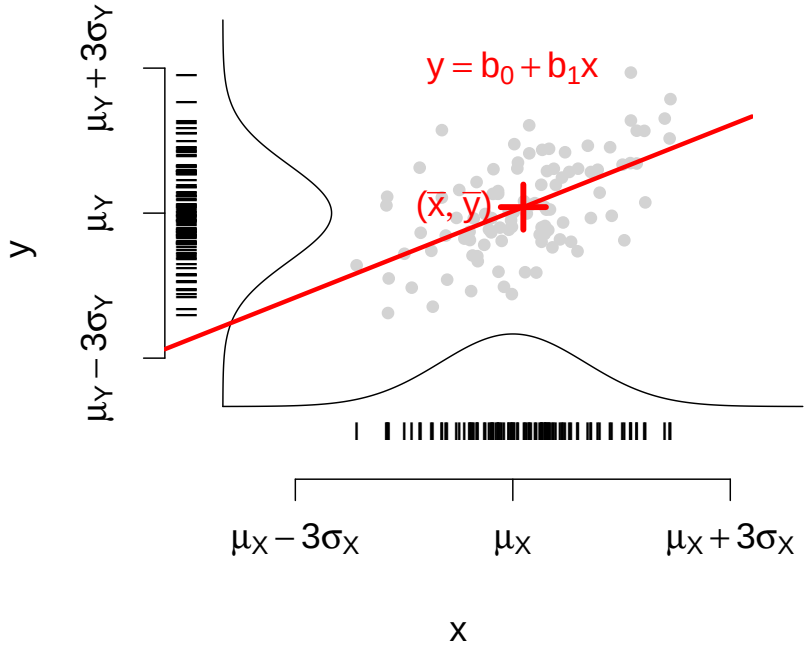
Meaning:

Regression line: b_0 and b_1

For the slope and the intercept of the regression line it holds

$$b_1 = r \cdot \frac{s_y}{s_x} \quad \text{and} \quad b_0 = \bar{y} - b_1 \cdot \bar{x}$$

n=100 $\rho = 0.6$



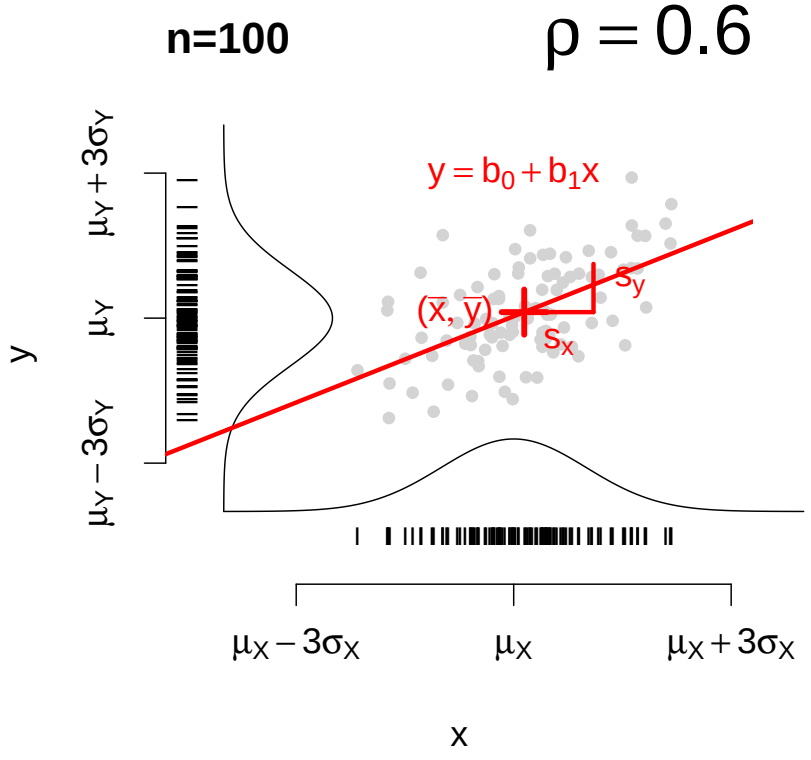
Meaning:

- the regression line passes the center of mass $(\bar{x}, \bar{y})$

Regression line: b_0 and b_1

For the slope and the intercept of the regression line it holds

$$b_1 = r \cdot \frac{s_y}{s_x} \quad \text{and} \quad b_0 = \bar{y} - b_1 \cdot \bar{x}$$



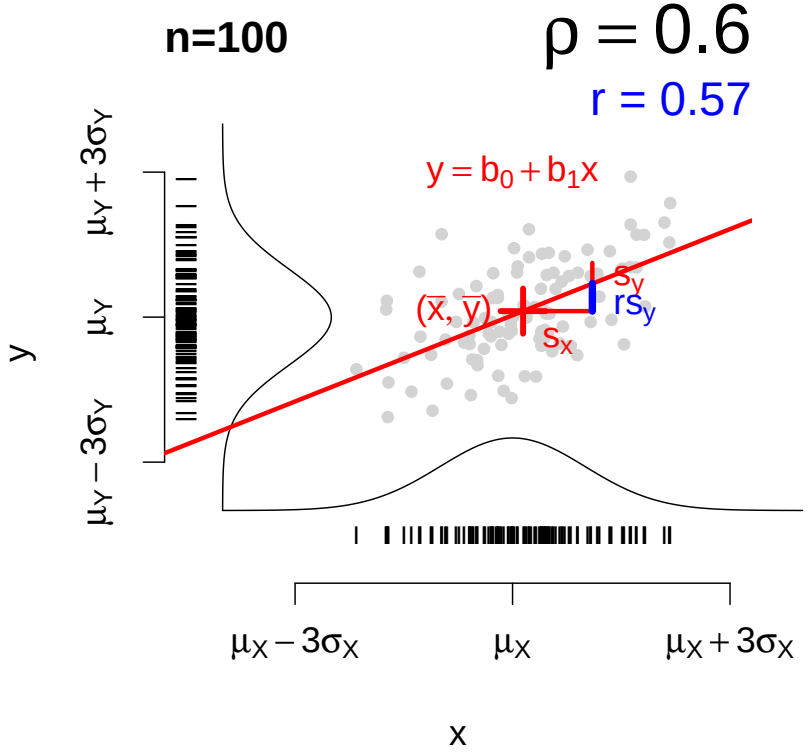
Meaning:

- the regression line passes the center of mass $(\bar{x}, \bar{y})$
- one step to the right of size s_x yields an increase of size $r \cdot s_y$

Regression line: b_0 and b_1

For the slope and the intercept of the regression line it holds

$$b_1 = r \cdot \frac{s_y}{s_x} \quad \text{and} \quad b_0 = \bar{y} - b_1 \cdot \bar{x}$$



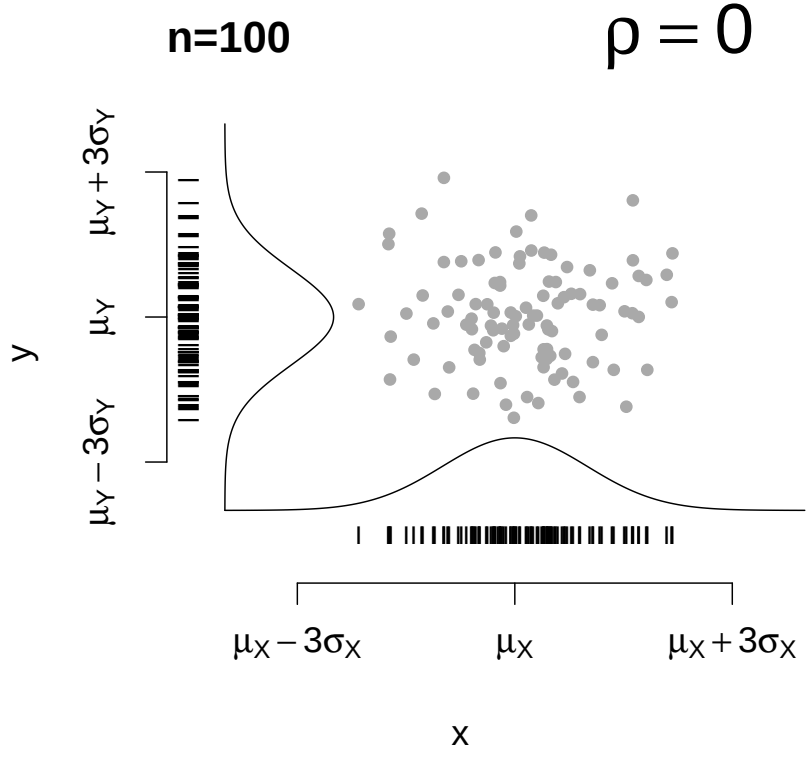
Meaning:

- the regression line passes the center of mass $(\bar{x}, \bar{y})$
- one step to the right of size s_x yields an increase of size $r \cdot s_y$

For the derivation of b_1 and b_0 see e.g., Messer, M. and Schneider, G. *Statistik: Theorie und Praxis im Dialog*, Springer Berlin

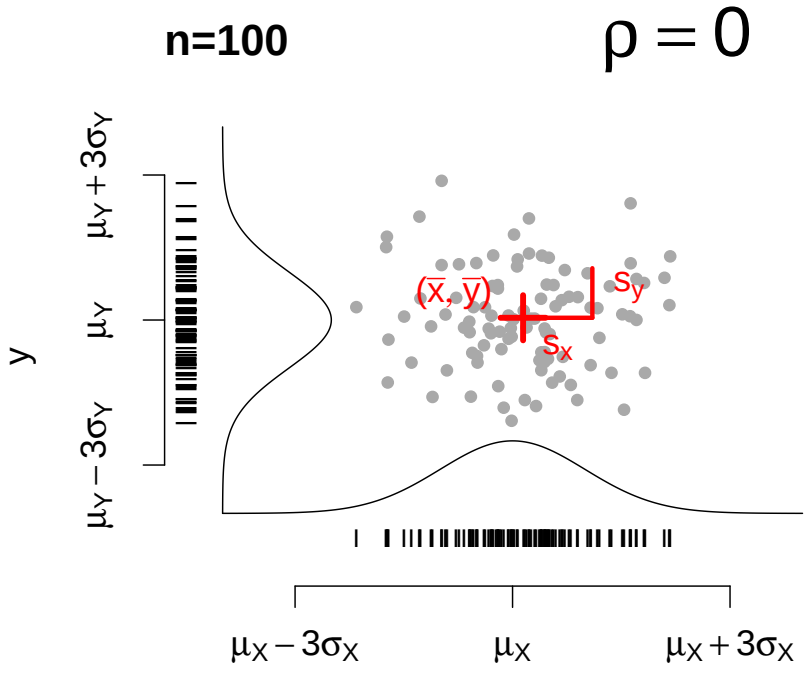
Regression line: examples

$$b_1 = r \cdot \frac{s_y}{s_x} \quad \text{and} \quad b_0 = \bar{y} - b_1 \cdot \bar{x}$$



Regression line: examples

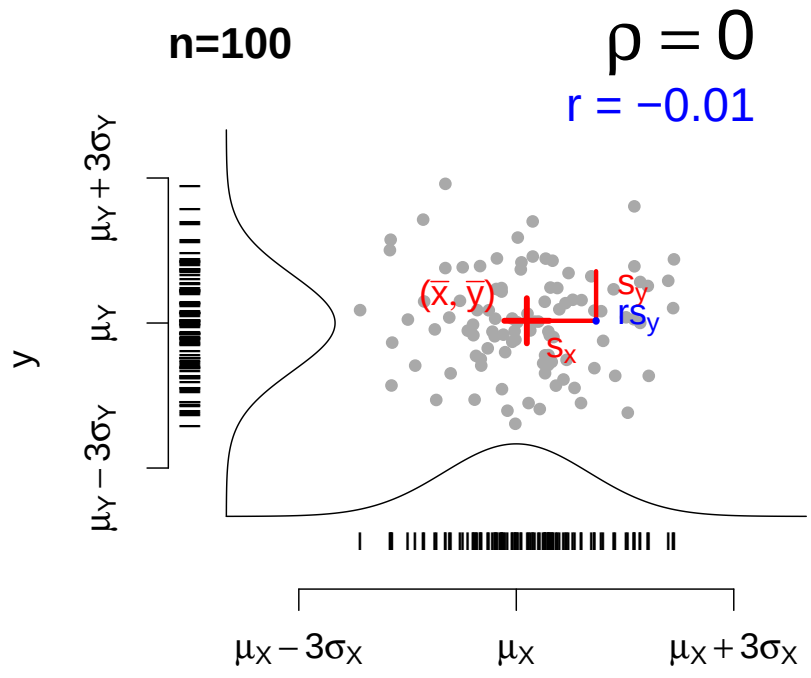
$$\boxed{b_1 = r \cdot \frac{s_y}{s_x}} \quad \text{and} \quad \boxed{b_0 = \bar{y} - b_1 \cdot \bar{x}}$$



- in fact, the data have their own standard deviations s_x and s_y

Regression line: examples

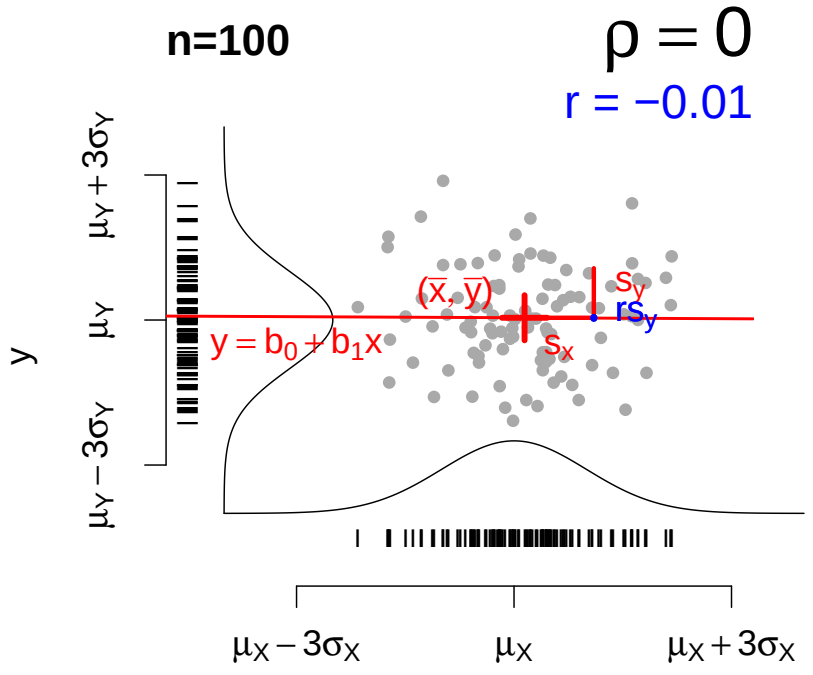
$$\boxed{b_1 = r \cdot \frac{s_y}{s_x}} \quad \text{and} \quad \boxed{b_0 = \bar{y} - b_1 \cdot \bar{x}}$$



- in fact, the data have their own standard deviations s_x and s_y
- however, the relation is negligible $r \approx -0.01$

Regression line: examples

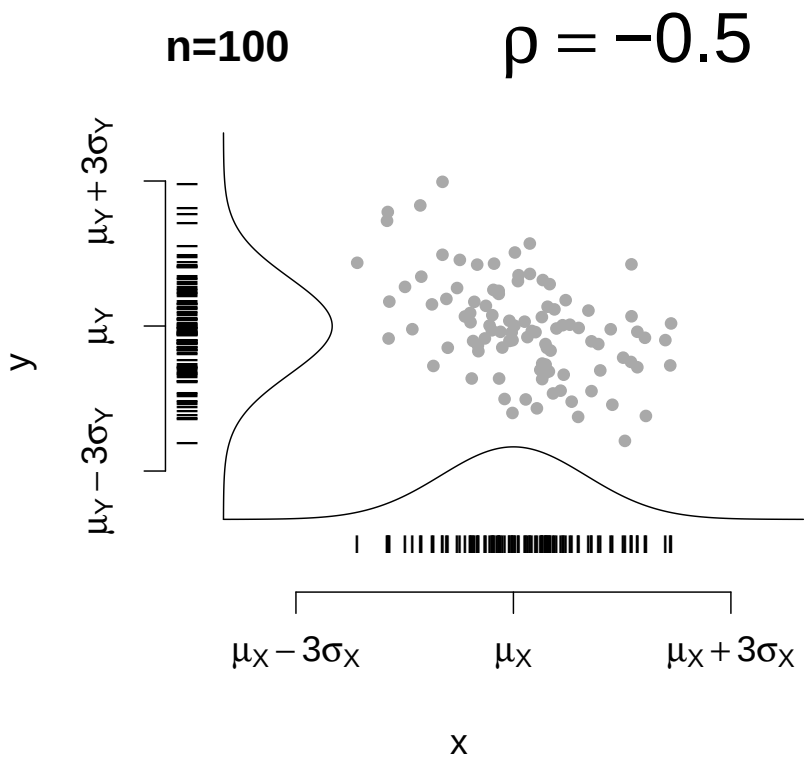
$$b_1 = r \cdot \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$



- in fact, the data have their own standard deviations s_x and s_y
- however, the relation is negligible $r \approx -0.01$
- and thus, the regression line is found flat

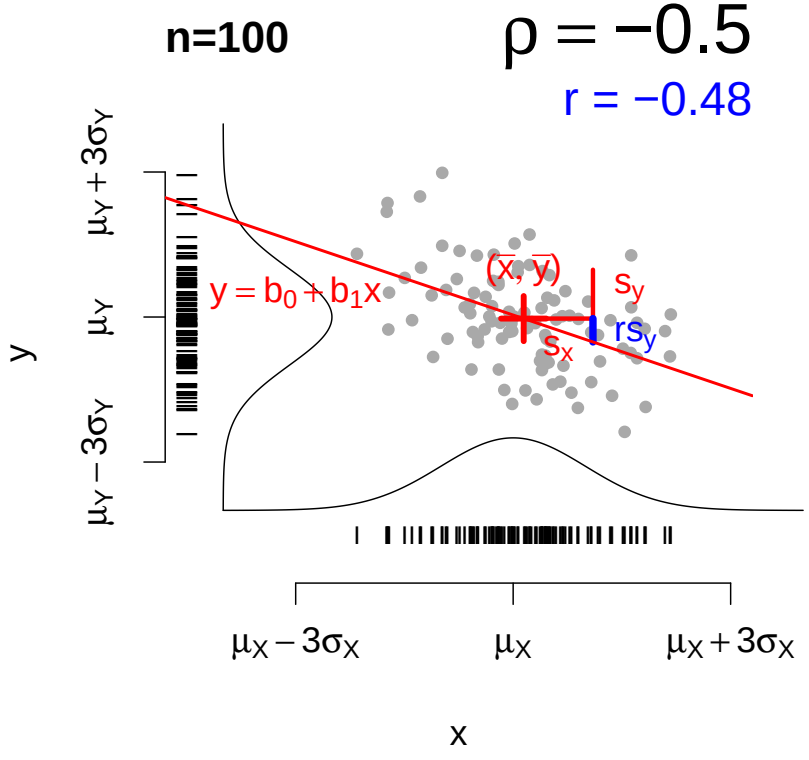
Regressionsgerade: examples

$$b_1 = r \cdot \frac{s_y}{s_x} \quad \text{and} \quad b_0 = \bar{y} - b_1 \cdot \bar{x}$$



Regressionsgerade: examples

$$\boxed{b_1 = r \cdot \frac{s_y}{s_x}} \quad \text{and} \quad \boxed{b_0 = \bar{y} - b_1 \cdot \bar{x}}$$

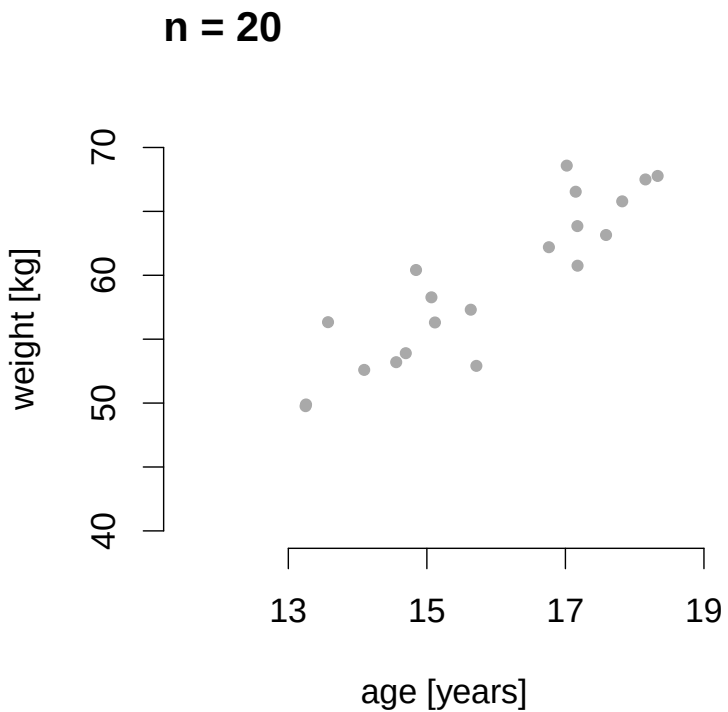


Data analysis: regression line

- Is there a relation between age and weight in teenage years?

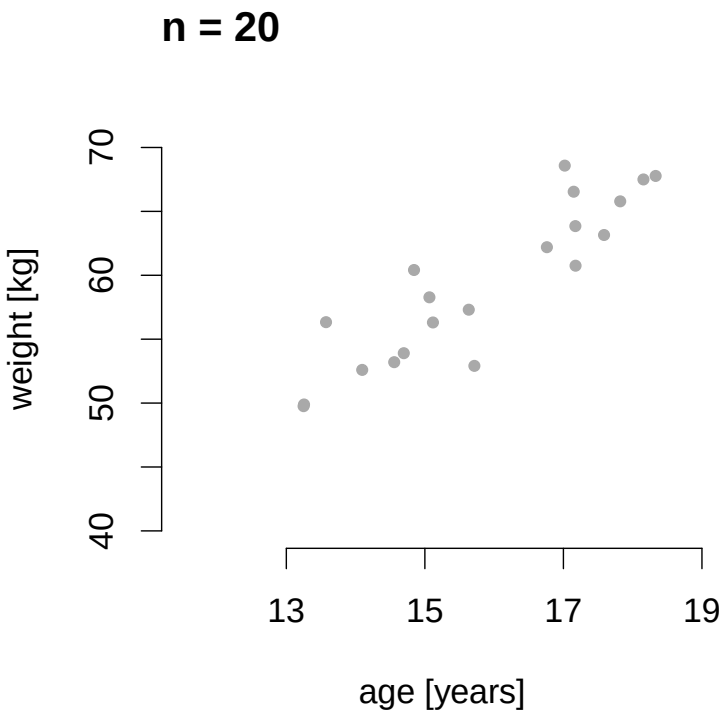
Data analysis: regression line

- Is there a relation between age and weight in teenage years?
- $n = 20$ teenagers of age between 13 and 19 years interviewed $\rightarrow$ data $(x_i, y_i)_{i=1,\ldots,n}$



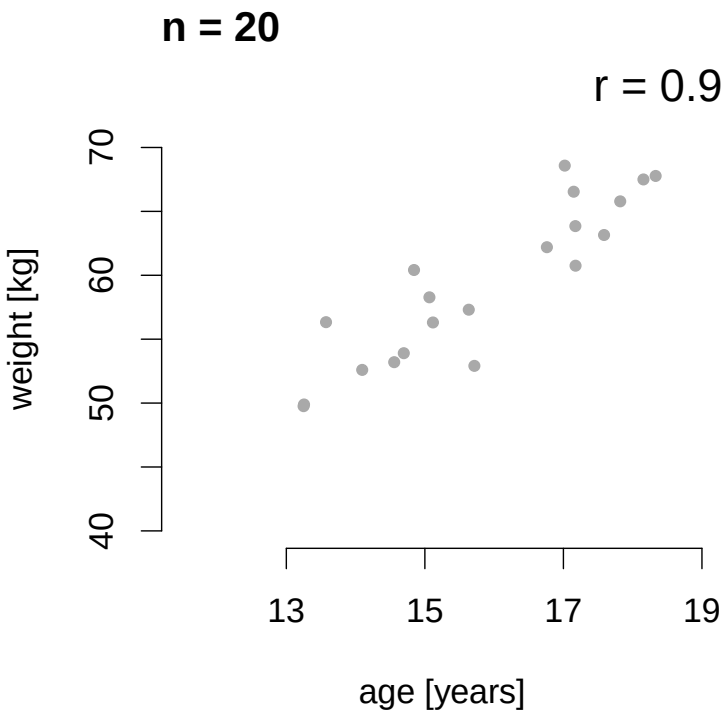
Data analysis: regression line

- Is there a relation between age and weight in teenage years?
- $n = 20$ teenagers of age between 13 and 19 years interviewed $\rightarrow$ data $(x_i, y_i)_{i=1,\ldots,n}$
- the relation is approximately linear, i.e., $y_i = \beta_0 + \beta_1 x_i + e_i$



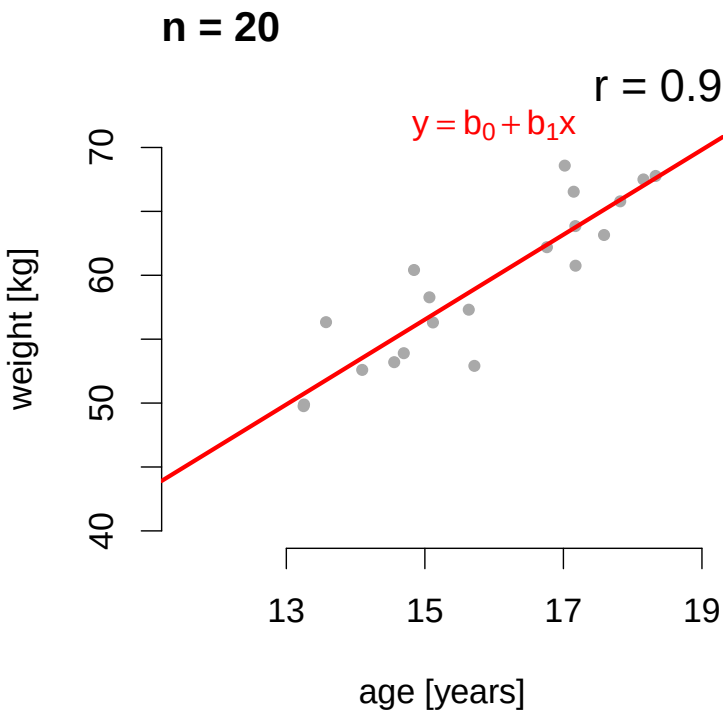
Data analysis: regression line

- Is there a relation between age and weight in teenage years?
- $n = 20$ teenagers of age between 13 and 19 years interviewed $\rightarrow$ data $(x_i, y_i)_{i=1,\ldots,n}$
- the relation is approximately linear, i.e., $y_i = \beta_0 + \beta_1 x_i + e_i$
- the data show a strong positive correlation, $r \approx 0.9$



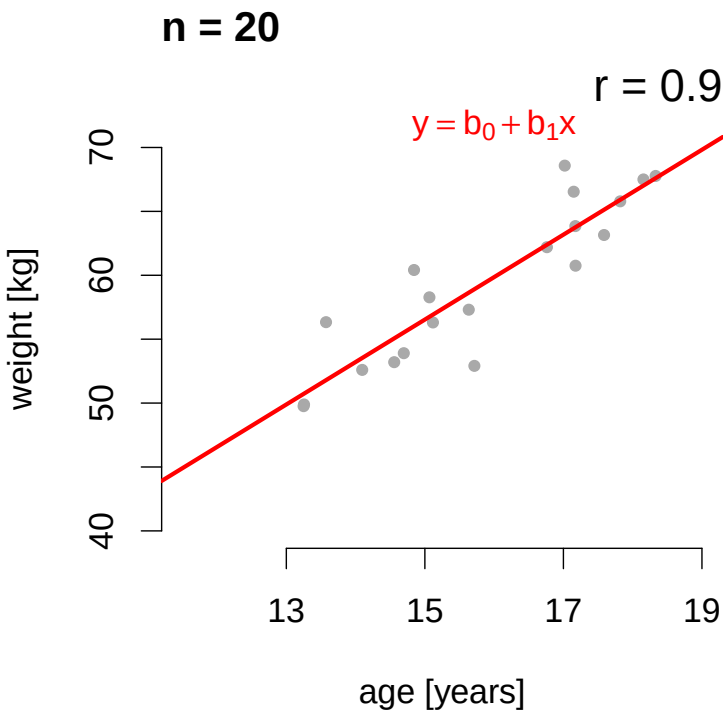
Data analysis: regression line

- Is there a relation between age and weight in teenage years?
- $n = 20$ teenagers of age between 13 and 19 years interviewed $\rightarrow$ data $(x_i, y_i)_{i=1,\ldots,n}$
- the relation is approximately linear, i.e., $y_i = \beta_0 + \beta_1 x_i + e_i$
- the data show a strong positive correlation, $r \approx 0.9$
- For the regression line we estimate $b_0 \approx 6.7$[kg] and $b_1 \approx 3.3$[kg/year]



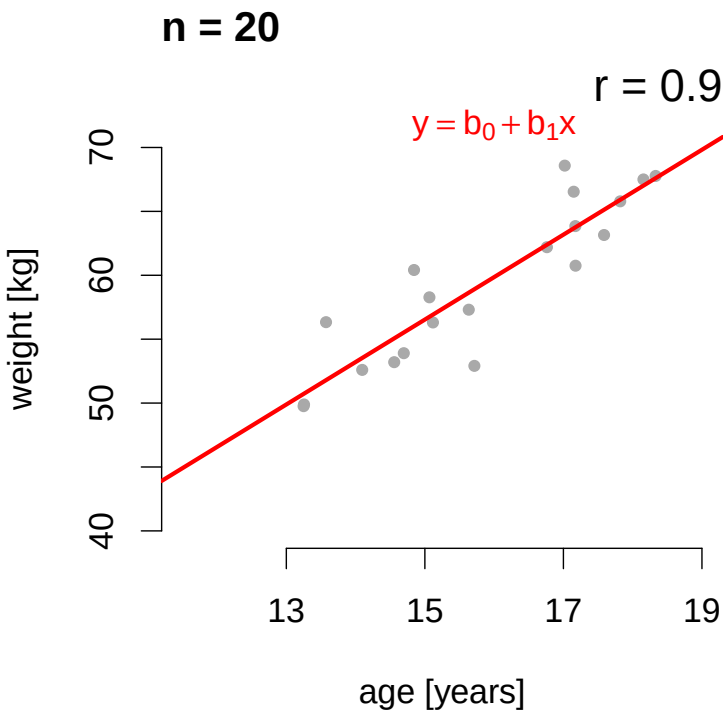
Data analysis: regression line

- Is there a relation between age and weight in teenage years?
- $n = 20$ teenagers of age between 13 and 19 years interviewed $\rightarrow$ data $(x_i, y_i)_{i=1,\ldots,n}$
- the relation is approximately linear, i.e., $y_i = \beta_0 + \beta_1 x_i + e_i$
- the data show a strong positive correlation, $r \approx 0.9$
- For the regression line we estimate $b_0 \approx 6.7$[kg] and $b_1 \approx 3.3$[kg/year]
- interpretation: per year the weight of a teenager increases about 3.3kg in the mean



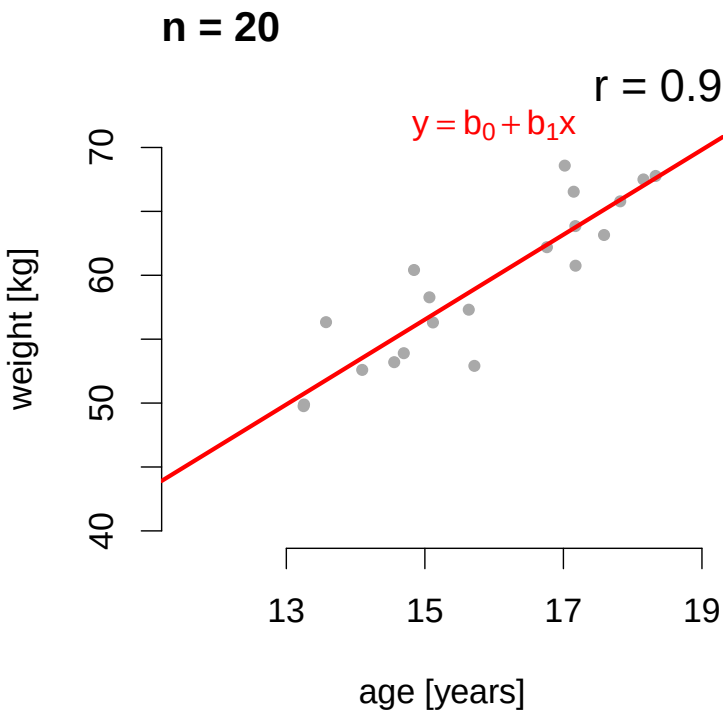
Data analysis: regression line

- Is there a relation between age and weight in teenage years?
- $n = 20$ teenagers of age between 13 and 19 years interviewed $\rightarrow$ data $(x_i, y_i)_{i=1,\ldots,n}$
- the relation is approximately linear, i.e., $y_i = \beta_0 + \beta_1 x_i + e_i$
- the data show a strong positive correlation, $r \approx 0.9$
- For the regression line we estimate $b_0 \approx 6.7$[kg] and $b_1 \approx 3.3$[kg/year]
- interpretation: per year the weight of a teenager increases about 3.3kg in the mean
- prediction: a 16-year old weighs in the mean $6.7 + 3.3 \cdot 16 = 59.5$kg



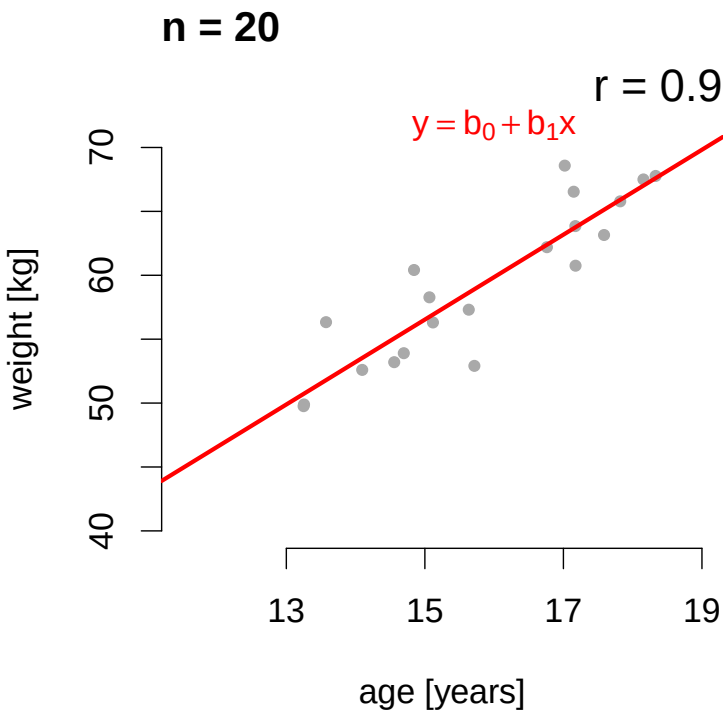
Data analysis: regression line

- Is there a relation between age and weight in teenage years?
- $n = 20$ teenagers of age between 13 and 19 years interviewed $\rightarrow$ data $(x_i, y_i)_{i=1,\ldots,n}$
- the relation is approximately linear, i.e., $y_i = \beta_0 + \beta_1 x_i + e_i$
- the data show a strong positive correlation, $r \approx 0.9$
- For the regression line we estimate $b_0 \approx 6.7$[kg] and $b_1 \approx 3.3$[kg/year]
- interpretation: per year the weight of a teenager increases about 3.3kg in the mean
- prediction: a 16-year old weighs in the mean $6.7 + 3.3 \cdot 16 = 59.5$kg
- Attention: predictions meaningful only in the observed range $[13, 19]$.



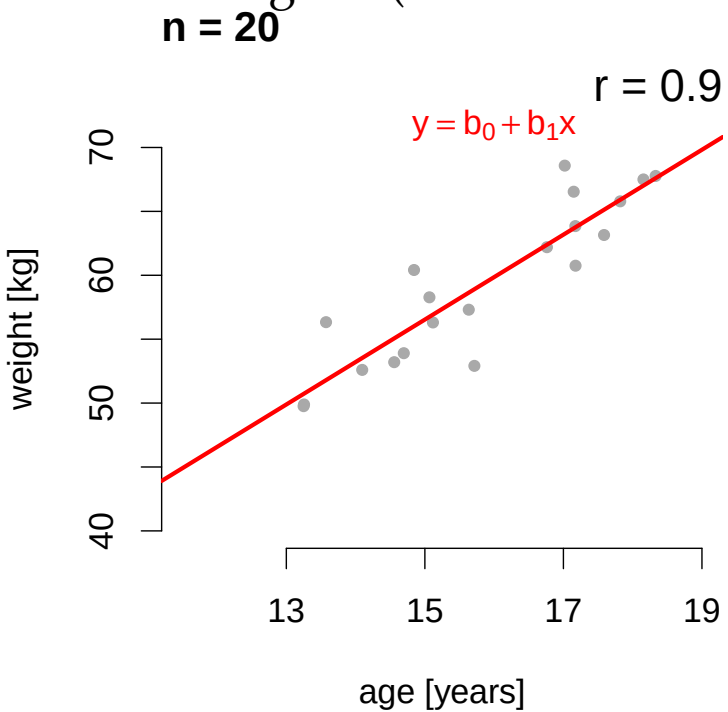
Data analysis: regression line

- Is there a relation between age and weight in teenage years?
- $n = 20$ teenagers of age between 13 and 19 years interviewed $\rightarrow$ data $(x_i, y_i)_{i=1,\ldots,n}$
- the relation is approximately linear, i.e., $y_i = \beta_0 + \beta_1 x_i + e_i$
- the data show a strong positive correlation, $r \approx 0.9$
- For the regression line we estimate $b_0 \approx 6.7$[kg] and $b_1 \approx 3.3$[kg/year]
- interpretation: per year the weight of a teenager increases about 3.3kg in the mean
- prediction: a 16-year old weighs in the mean $6.7 + 3.3 \cdot 16 = 59.5$kg
- Attention: predictions meaningful only in the observed range $[13, 19]$. 80-year old people do not weigh about 270kg.



Data analysis: regression line

- Is there a relation between age and weight in teenage years?
- $n = 20$ teenagers of age between 13 and 19 years interviewed $\rightarrow$ data $(x_i, y_i)_{i=1,\ldots,n}$
- the relation is approximately linear, i.e., $y_i = \beta_0 + \beta_1 x_i + e_i$
- the data show a strong positive correlation, $r \approx 0.9$
- For the regression line we estimate $b_0 \approx 6.7$[kg] and $b_1 \approx 3.3$[kg/year]
- interpretation: per year the weight of a teenager increases about 3.3kg in the mean
- prediction: a 16-year old weighs in the mean $6.7 + 3.3 \cdot 16 = 59.5$kg
- Attention: predictions meaningful only in the observed range $[13, 19]$. 80-year old people do not weigh about 270kg. Similarly, the intercept $b_0 = 6.7$ is biologically meaningless (newborns don't weigh about 6.7kg in the mean)



Regression line in R

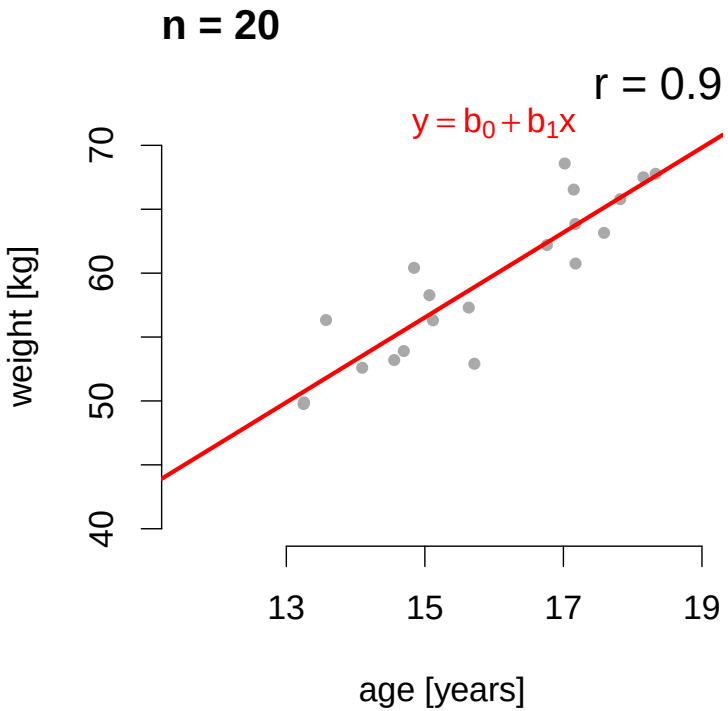
```
# Enter data, x- and y- values as vectors
x <- c(...)
y <- c(...)
# Calculate regression line
lm(y~x)
# Output
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
      6.701        3.322
```

- `lm(y~x)` means: describe the y_i as a linear function of the x_i plus error, thus $y_i = \beta_0 + \beta_1 \cdot x_i + e_i$, and estimate the intercept β_0 and the slope β_1 through least-squares (`lm()` for 'linear model').
- the estimated intercept is $b_0 \approx 6.7$ and the estimated slope is $b_1 \approx 3.3$.
- the regression line can be added to a plot via `abline(lm(y~x))`.
- Alternatively 'by hand' as before: $b_1 = r \cdot s_y/s_x$ and $b_0 = \bar{y} - b_1 \cdot \bar{x}$.

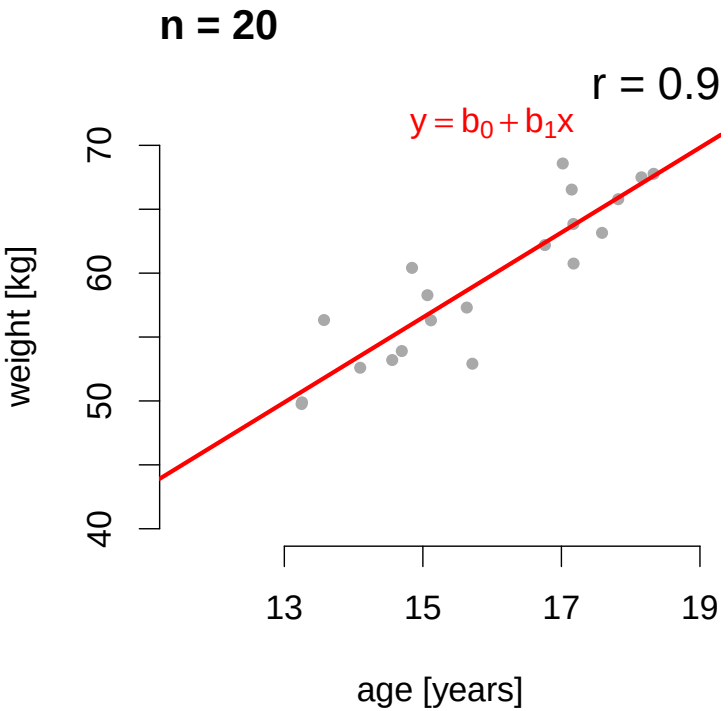
Significance test for the slope

- $n = 20$ teenagers asked for their age and weight $\rightarrow$ data $(x_i, y_i)_{i=1,\ldots,n}$
- $b_0 \approx 6.7$ and $b_1 \approx 3.3$



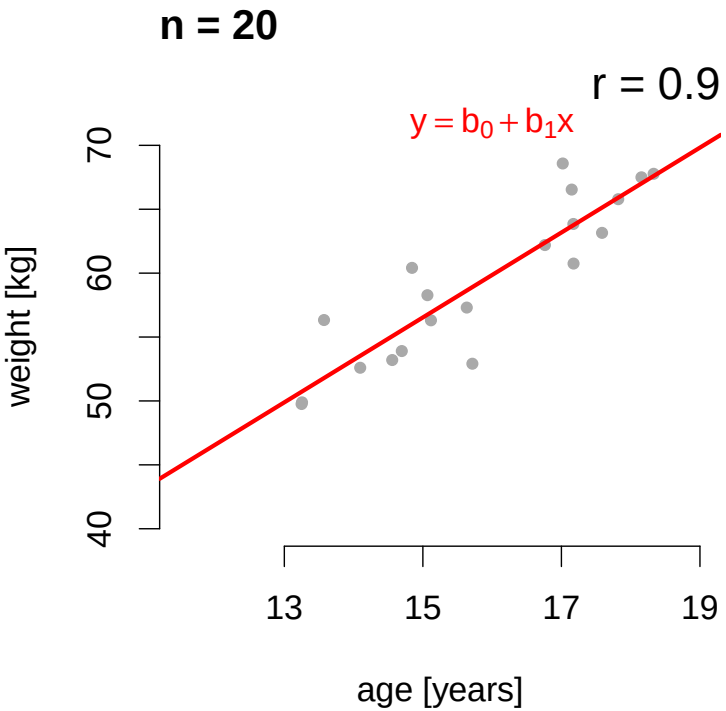
Significance test for the slope

- $n = 20$ teenagers asked for their age and weight $\rightarrow$ data $(x_i, y_i)_{i=1,\ldots,n}$
- $b_0 \approx 6.7$ and $b_1 \approx 3.3$
- question: can the positive relation observed in the data have easily happened by chance, if there was actually no difference in the mean weights in population of all teenagers between 13 and 19 years?



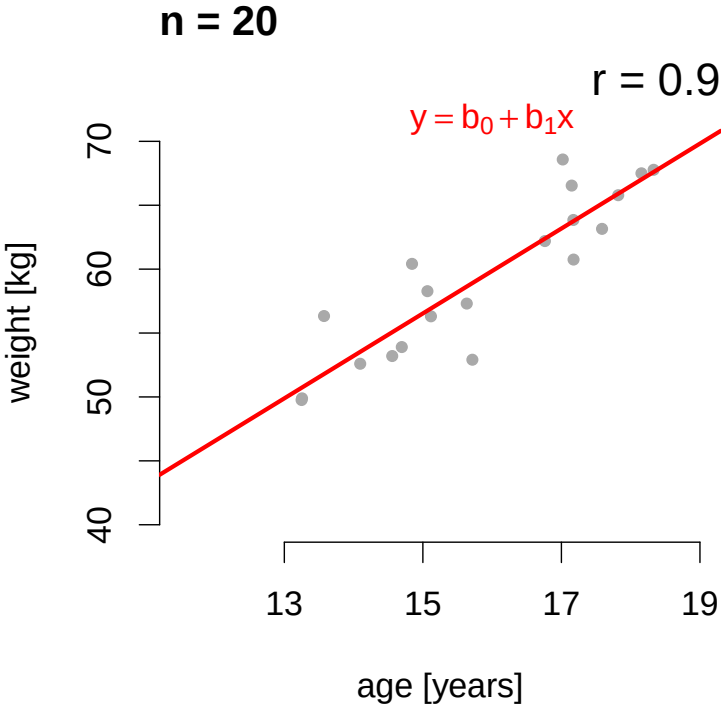
Significance test for the slope

- $n = 20$ teenagers asked for their age and weight $\rightarrow$ data $(x_i, y_i)_{i=1,\ldots,n}$
- $b_0 \approx 6.7$ and $b_1 \approx 3.3$
- question: can the positive relation observed in the data have easily happened by chance, if there was actually no difference in the mean weights in population of all teenagers between 13 and 19 years?
- Answers: this depends on the variability of the estimated slope b_1



Significance test for the slope

- $n = 20$ teenagers asked for their age and weight $\rightarrow$ data $(x_i, y_i)_{i=1,\ldots,n}$
- $b_0 \approx 6.7$ and $b_1 \approx 3.3$
- question: can the positive relation observed in the data have easily happened by chance, if there was actually no difference in the mean weights in population of all teenagers between 13 and 19 years?
- Answers: this depends on the variability of the estimated slope b_1
- More precisely, need a statistical model in which we can speak about the variability of the estimated slope



Linear regression model

- Model: For $i = 1, \ldots, n$ let

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \sigma Z_i,$$

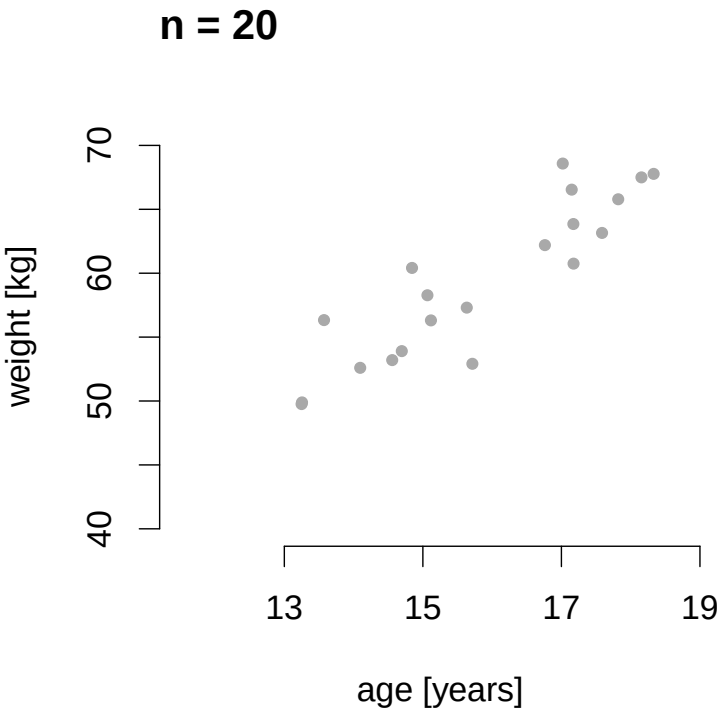
with $Z_1, \ldots, Z_n$ i.i.d. RVs and $Z_1 \sim N(0,1)$, and $(\beta_0, \beta_1, \sigma) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$

Linear regression model

- Model: For $i = 1, \ldots, n$ let

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \sigma Z_i,$$

with $Z_1, \ldots, Z_n$ i.i.d. RVs and $Z_1 \sim N(0,1)$, and $(\beta_0, \beta_1, \sigma) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$



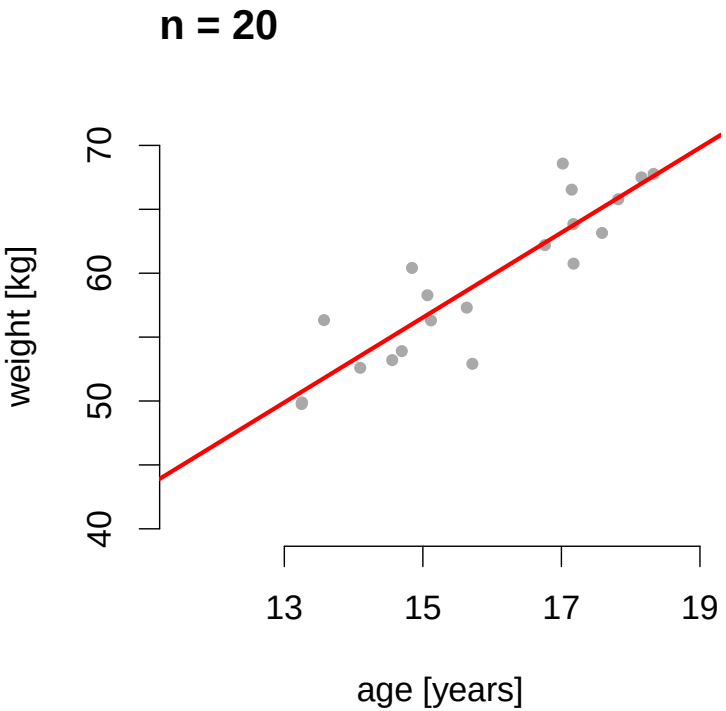
- The pairs $(x_i, y_i)_i$ are given data and y_i is interpreted as a realization of Y_i

Linear regression model

- Model: For $i = 1, \dots, n$ let

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \sigma Z_i,$$

with $Z_1, \dots, Z_n$ i.i.d. RVs and $Z_1 \sim N(0,1)$, and $(\beta_0, \beta_1, \sigma) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$



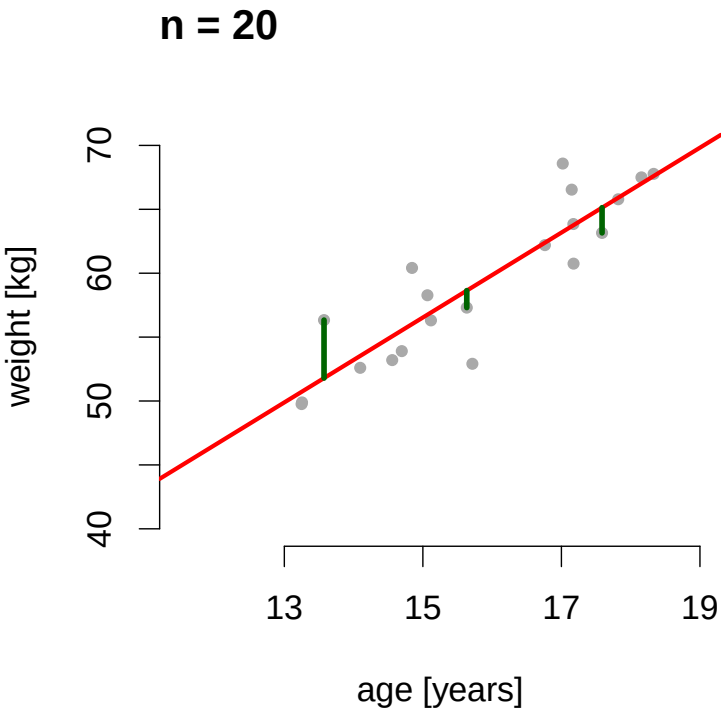
- The pairs $(x_i, y_i)_i$ are given data and y_i is interpreted as a realization of Y_i
- Y_i depends on linear proportion $\beta_0 + \beta_1 \cdot x_i$ as well as a *random* error σZ_i

Linear regression model

- Model: For $i = 1, \dots, n$ let

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \sigma Z_i,$$

with $Z_1, \dots, Z_n$ i.i.d. RVs and $Z_1 \sim N(0,1)$, and $(\beta_0, \beta_1, \sigma) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$



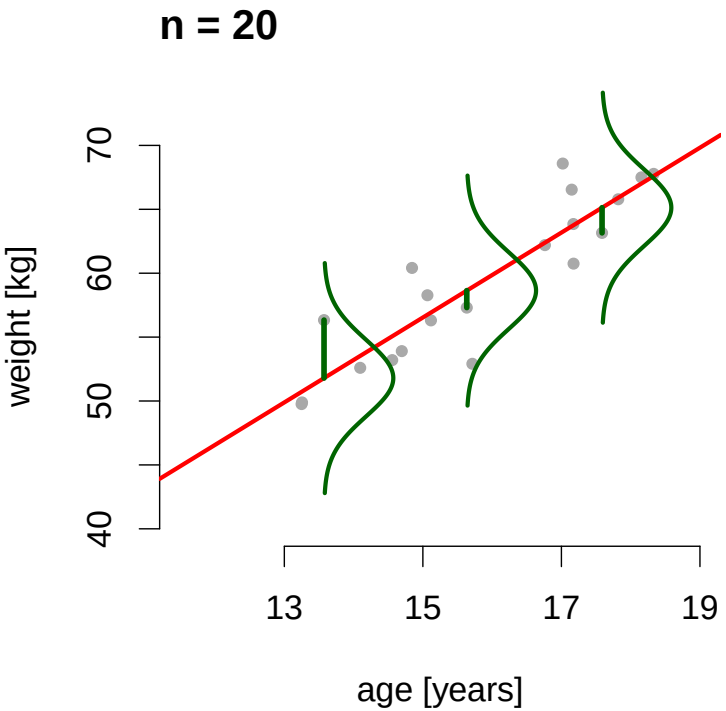
- The pairs $(x_i, y_i)_i$ are given data and y_i is interpreted as a realization of Y_i
- Y_i depends on linear proportion $\beta_0 + \beta_1 \cdot x_i$ as well as a *random* error σZ_i

Linear regression model

- Model: For $i = 1, \ldots, n$ let

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \sigma Z_i,$$

with $Z_1, \ldots, Z_n$ i.i.d. RVs and $Z_1 \sim N(0,1)$, and $(\beta_0, \beta_1, \sigma) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$



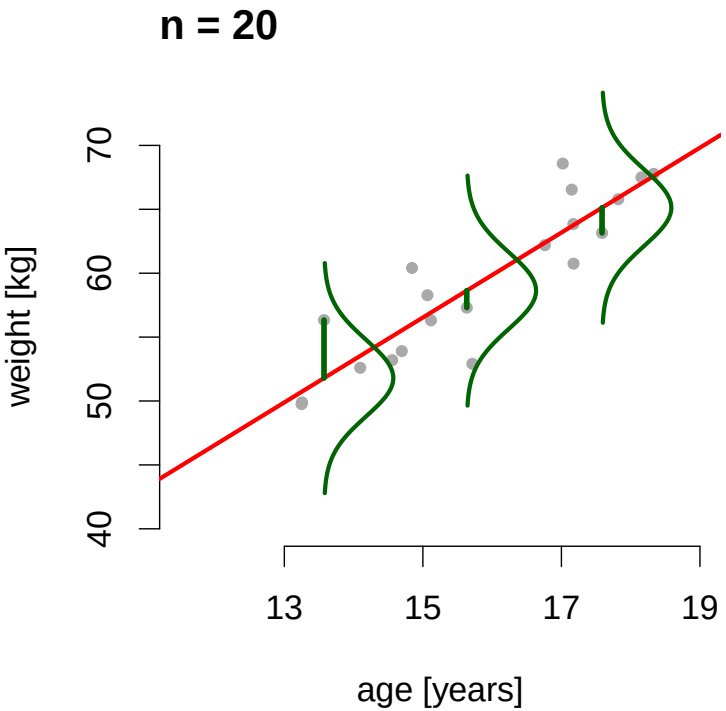
- The pairs $(x_i, y_i)_i$ are given data and y_i is interpreted as a realization of Y_i
- Y_i depends on linear proportion $\beta_0 + \beta_1 \cdot x_i$ as well as a *random* error σZ_i

Linear regression model

- Model: For $i = 1, \ldots, n$ let

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \sigma Z_i,$$

with $Z_1, \ldots, Z_n$ i.i.d. RVs and $Z_1 \sim N(0,1)$, and $(\beta_0, \beta_1, \sigma) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$



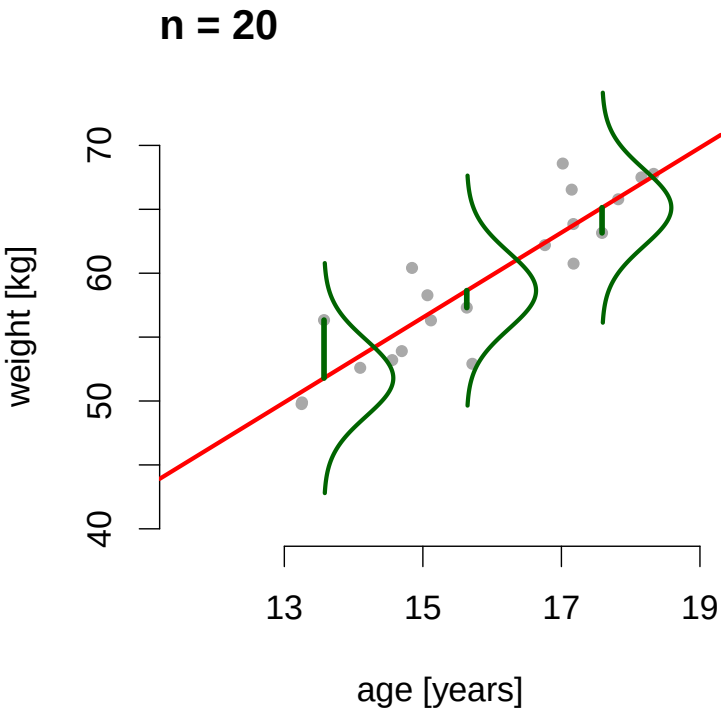
- The pairs $(x_i, y_i)_i$ are given data and y_i is interpreted as a realization of Y_i
- Y_i depends on linear proportion $\beta_0 + \beta_1 \cdot x_i$ as well as a *random* error σZ_i
- consequence: in the context of the model, the least-squares estimators B_0 and B_1 for β_0 and β_1 become random statistics...

Linear regression model

- Model: For $i = 1, \ldots, n$ let

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \sigma Z_i,$$

with $Z_1, \ldots, Z_n$ i.i.d. RVs and $Z_1 \sim N(0,1)$, and $(\beta_0, \beta_1, \sigma) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$



- The pairs $(x_i, y_i)_i$ are given data and y_i is interpreted as a realization of Y_i
- Y_i depends on linear proportion $\beta_0 + \beta_1 \cdot x_i$ as well as a *random* error σZ_i
- consequence: in the context of the model, the least-squares estimators B_0 and B_1 for β_0 and β_1 become random statistics...and have a standard error

Linear regression model

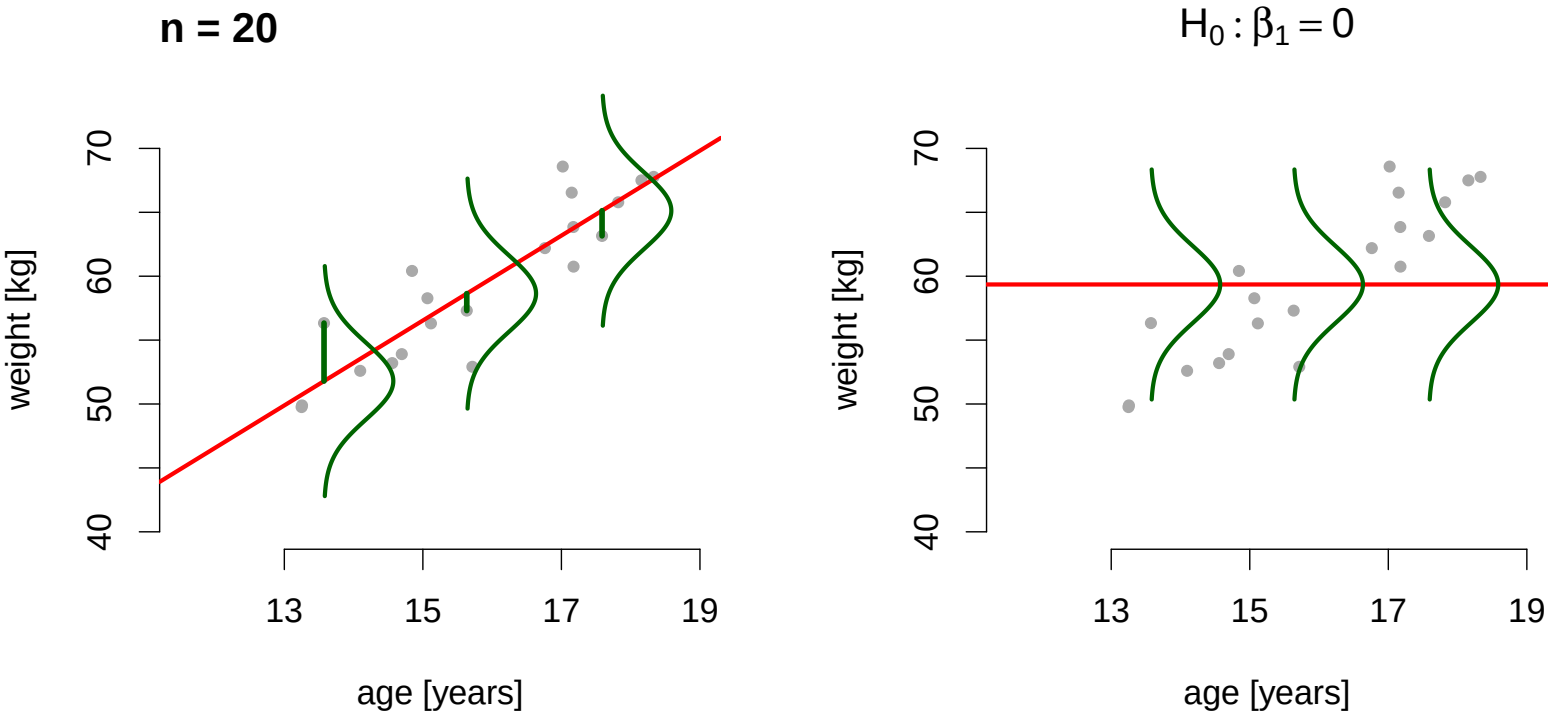
- Model: For $i = 1, \ldots, n$ let

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \sigma Z_i,$$

with $Z_1, \ldots, Z_n$ i.i.d. RVs and $Z_1 \sim N(0,1)$, and $(\beta_0, \beta_1, \sigma) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$

- Null hypothesis: $H_0 : \beta_1 = 0$ (no relation of age and weight)

Intuitively implausible, as around $x_i \approx 14$ all y_i in the lower tail, while for $x_i \approx 18$ all y_i in the upper tail



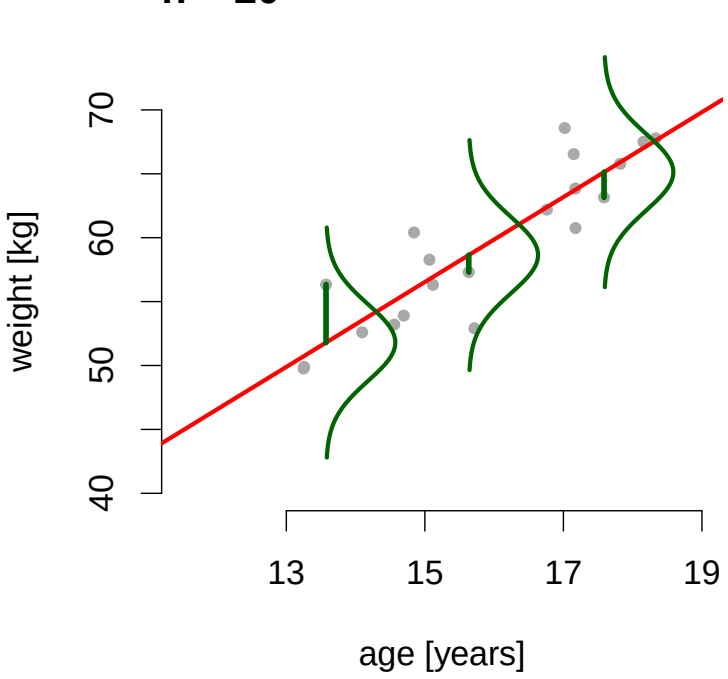
- The pairs $(x_i, y_i)_i$ are given data and y_i is interpreted as a realization of Y_i
- Y_i depends on linear proportion $\beta_0 + \beta_1 \cdot x_i$ as well as a *random* error σZ_i
- consequence: in the context of the model, the least-squares estimators B_0 and B_1 for β_0 and β_1 become random statistics...and have a standard error

Linear regression model

- Model: For $i = 1, \ldots, n$ let

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \sigma Z_i,$$

with $Z_1, \ldots, Z_n$ i.i.d. RVs and $Z_1 \sim N(0,1)$, and $(\beta_0, \beta_1, \sigma) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$



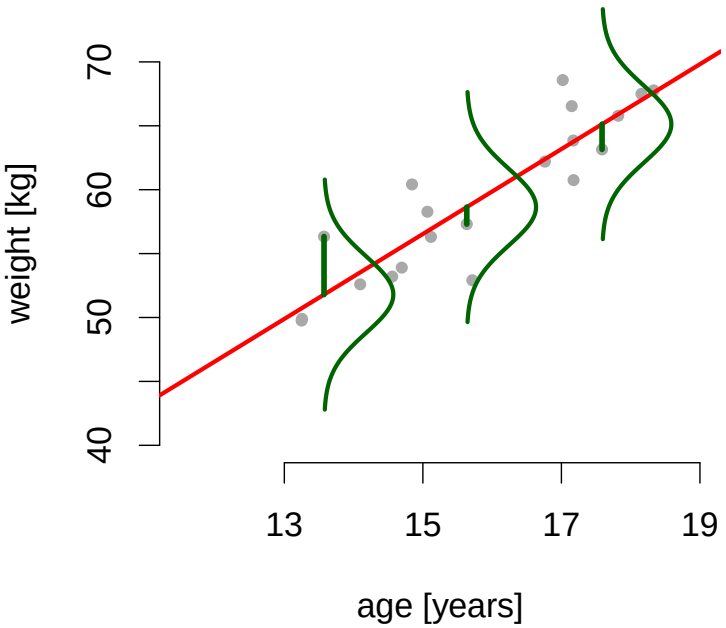
- B_0 and B_1 least-squares estimators for β_0 and β_1

Linear regression model

- Model: For $i = 1, \ldots, n$ let

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \sigma Z_i,$$

with $Z_1, \ldots, Z_n$ i.i.d. RVs and $Z_1 \sim N(0,1)$, and $(\beta_0, \beta_1, \sigma) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$



- B_0 and B_1 least-squares estimators for β_0 and β_1
- the variance σ^2 is estimated from the residuals via

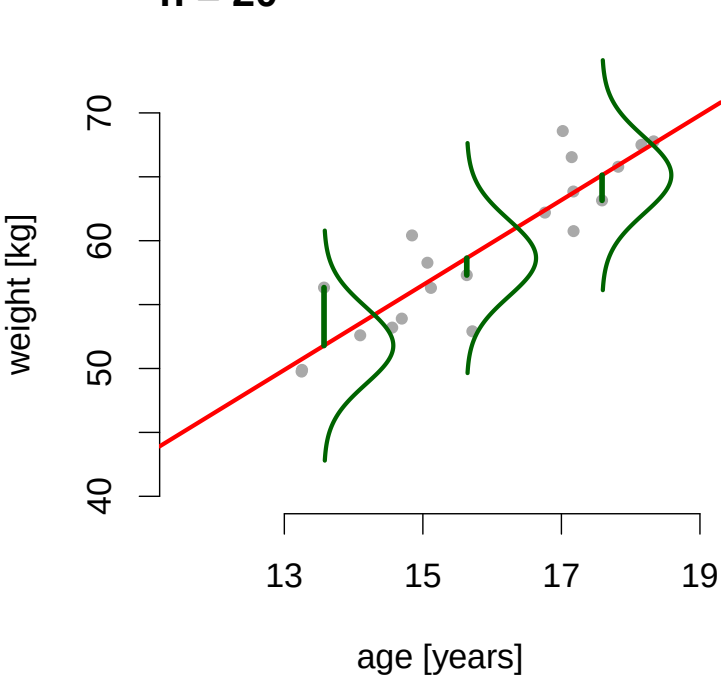
$$S_r^2 := \frac{1}{n-2} \sum_{i=1}^{n} [Y_i - (B_0 + B_1 \cdot x_i)]^2$$

Linear regression model

- Model: For $i = 1, \ldots, n$ let

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \sigma Z_i,$$

with $Z_1, \ldots, Z_n$ i.i.d. RVs and $Z_1 \sim N(0,1)$, and $(\beta_0, \beta_1, \sigma) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$



- B_0 and B_1 least-squares estimators for β_0 and β_1
- the variance σ^2 is estimated from the residuals via

$$S_r^2 := \frac{1}{n-2} \sum_{i=1}^{n} [Y_i - (B_0 + B_1 \cdot x_i)]^2$$

- Definition: S_r is called the *standard error of the regression*

Standard error of the slope B_1 – Definition

- B_0 and B_1 least-squares estimators for β_0 and β_1
- Definition: The *standard error of the slope* B_1 is

$$\boxed{SE_{B_1} := \frac{S_r}{s_x \cdot \sqrt{n-1}}}$$

while

$$S_r = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} [Y_i - (B_0 + B_1 \cdot x_i)]^2}$$

denotes the standard error of the regression and

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

is the (non-random) empirical variance of the data $(x_i)_i$

Standard error of the slope B_1 – Definition

- B_0 and B_1 least-squares estimators for β_0 and β_1
- Definition: The *standard error of the slope* B_1 is

$$\boxed{SE_{B_1} := \frac{S_r}{s_x \cdot \sqrt{n-1}}}$$

while

$$S_r = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}[Y_i - (B_0 + B_1 \cdot x_i)]^2}$$

denotes the standard error of the regression and

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

is the (non-random) empirical variance of the data $(x_i)_i$

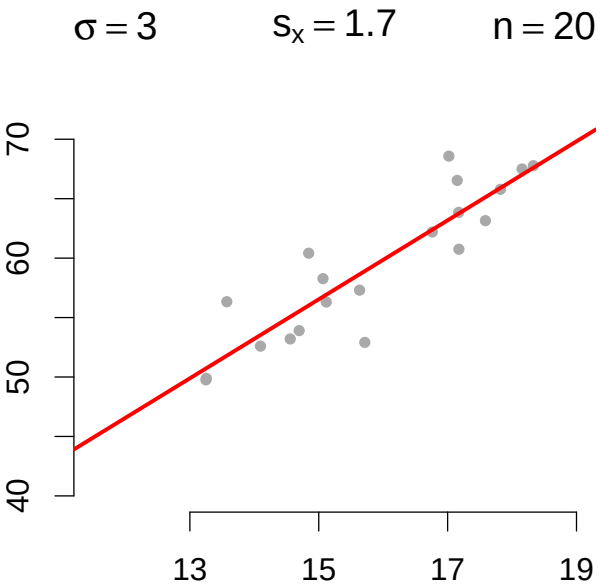
Note: In general a *standard error* denotes an estimator for the standard deviation of a statistic, for example

- the standard error of the regression S_r estimates $\mathbb{V}ar(Y_1)^{1/2} = \sigma$
- the standard error of the slope SE_{B_1} estimates $\mathbb{V}ar(B_1)^{1/2} = \sigma/(s_x\sqrt{n-1})$ (latter equality not shown)
- the standard error of the mean SEM estimates $\mathbb{V}ar(\bar{Z})^{1/2} = \sigma/\sqrt{n}$ (not used in this context)

Standard error of the slope B_1 – Intuition

- The *standard error of the slope* B_1 is

$$SE_{B_1} := \frac{S_r}{s_x \cdot \sqrt{n-1}}$$

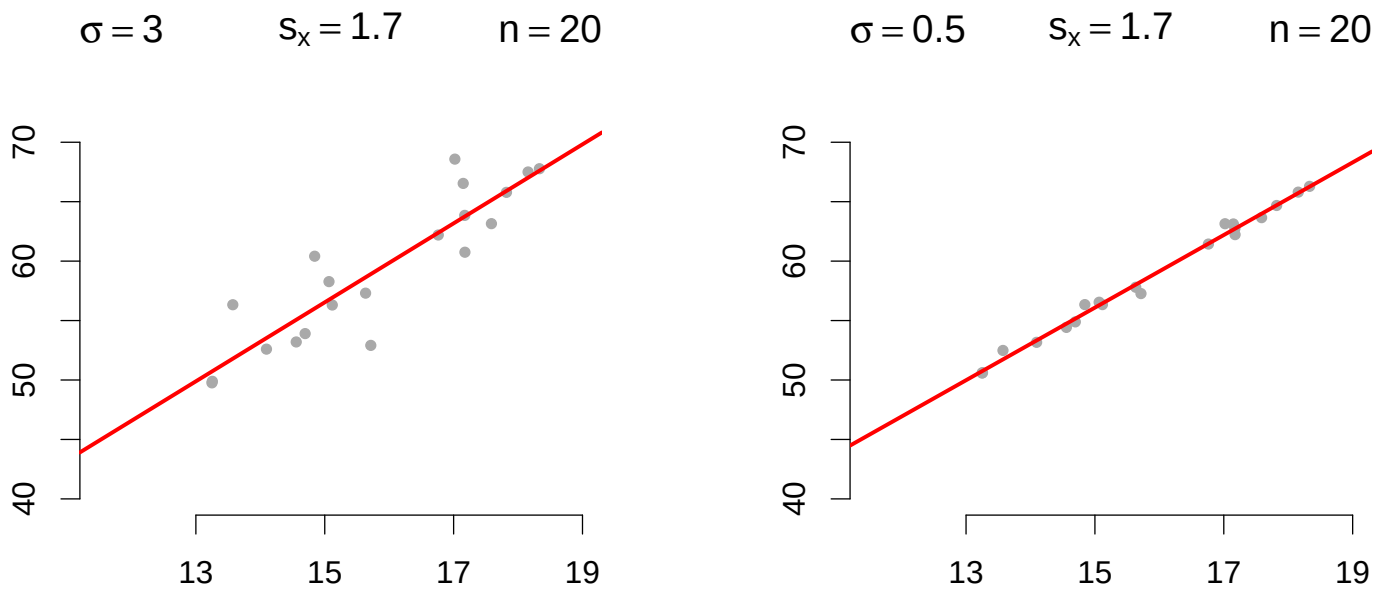


Why is SE_{B_1} plausible? In other words: How 'variable' is the regression line?

Standard error of the slope B_1 – Intuition

- The *standard error of the slope* B_1 is

$$SE_{B_1} := \frac{S_r}{s_x \cdot \sqrt{n-1}}$$



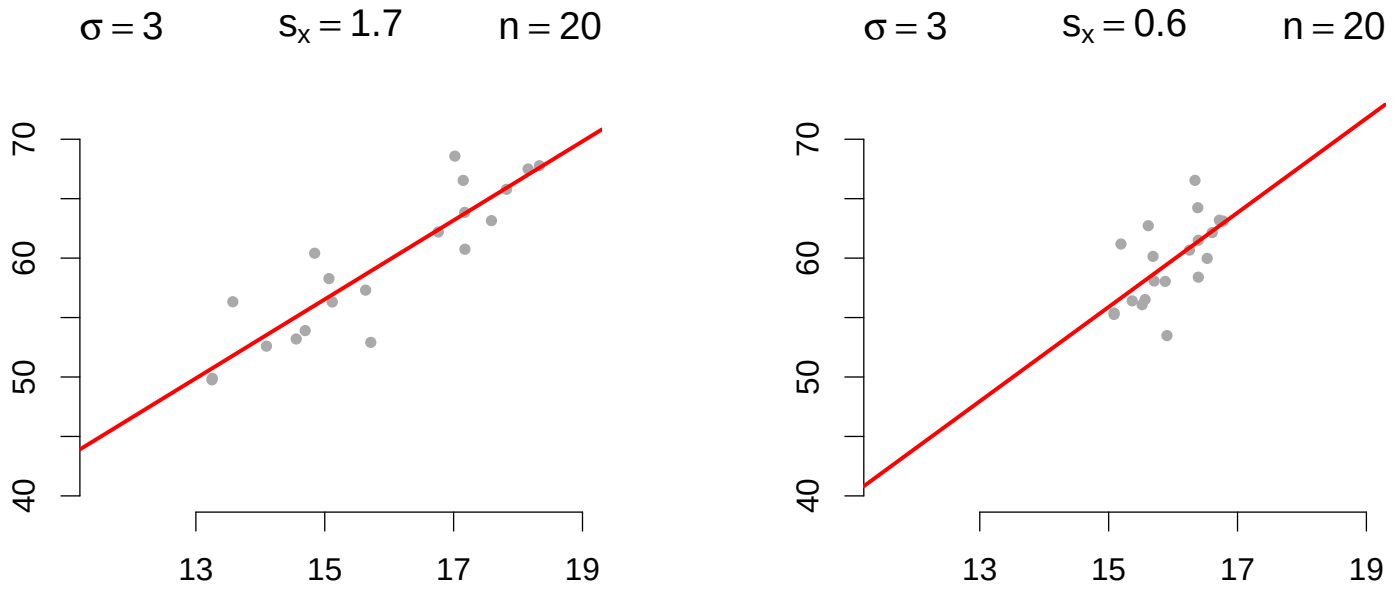
Why is SE_{B_1} plausible? In other words: How 'variable' is the regression line? It is 'intuitively stable' if

- S_r (the estimator for σ) is small here on right side

Standard error of the slope B_1 – Intuition

- The *standard error of the slope* B_1 is

$$SE_{B_1} := \frac{S_r}{s_x \cdot \sqrt{n-1}}$$



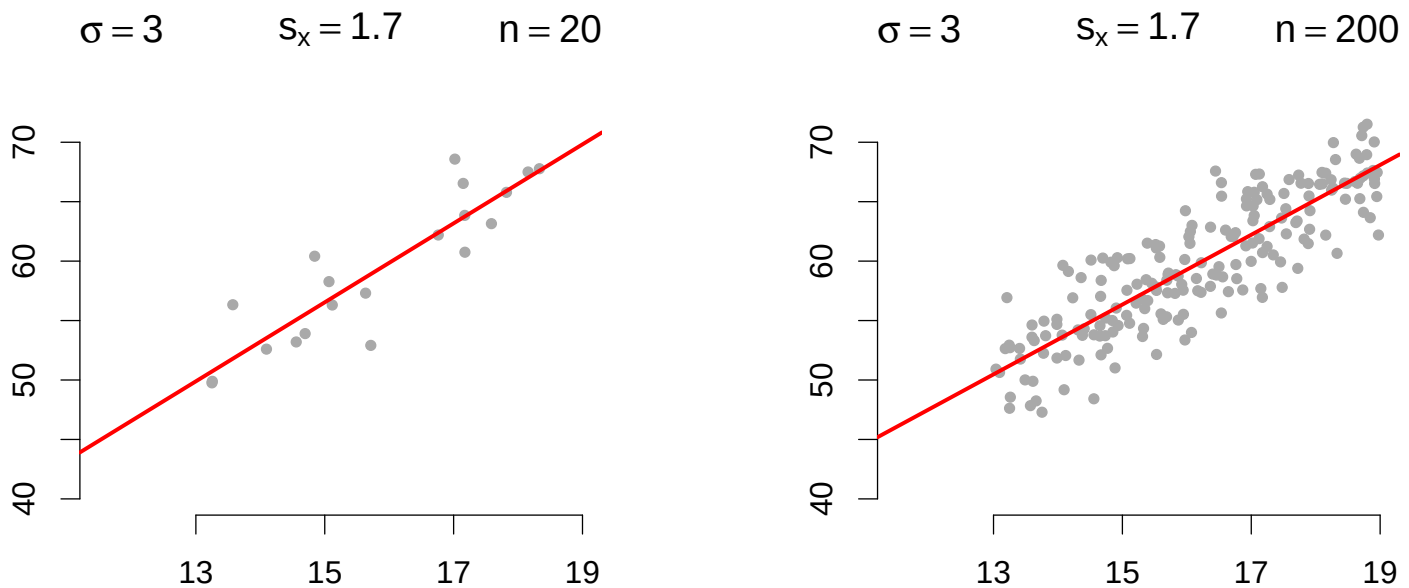
Why is SE_{B_1} plausible? In other words: How 'variable' is the regression line? It is 'intuitively stable' if

- S_r (the estimator for σ) is small
- s_x (the variability of the $(x_i)_i$ data) is large here on left side

Standard error of the slope B_1 – Intuition

- The *standard error of the slope* B_1 is

$$SE_{B_1} := \frac{S_r}{s_x \cdot \sqrt{n-1}}$$



Why is SE_{B_1} plausible? In other words: How 'variable' is the regression line? It is 'intuitively stable' if

- S_r (the estimator for σ) is small
- s_x (the variability of the $(x_i)_i$ data) is large
- n (the number of observations) is large

here on right side

Significance test for the slope B_1

- For $i = 1, \ldots, n$ let
$$Y_i = \beta_0 + \beta_1 \cdot x_i + \sigma Z_i,$$
with $Z_1, \ldots, Z_n$ i.i.d. RVs and $Z_1 \sim N(0,1)$, and $(\beta_0, \beta_1, \sigma) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$
- $q_{1-\alpha/2}$ the $(1-\alpha/2)$-quantile of the $t(n-2)$-distribution

Under $H_0 : \beta_1 = \beta_1^{(0)}$ it holds

$$T := \frac{B_1 - \beta_1^{(0)}}{SE_{B_1}} \sim t(n-2)$$

Significance test for the slope B_1

- For $i = 1, \ldots, n$ let
$$Y_i = \beta_0 + \beta_1 \cdot x_i + \sigma Z_i,$$
with $Z_1, \ldots, Z_n$ i.i.d. RVs and $Z_1 \sim N(0,1)$, and $(\beta_0, \beta_1, \sigma) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$
- $q_{1-\alpha/2}$ the $(1-\alpha/2)$-quantile of the $t(n-2)$-distribution

Under $H_0 : \beta_1 = \beta_1^{(0)}$ it holds

$$T := \frac{B_1 - \beta_1^{(0)}}{SE_{B_1}} \sim t(n-2)$$

and equivalently: the confidence interval

$$I := \left(B_1 - q_{1-\alpha/2} \cdot SE_{B_1},\ B_1 + q_{1-\alpha/2} \cdot SE_{B_1}\right)$$

overlaps $\beta_1^{(0)}$ with probability $1-\alpha$

- known structure: $T = (\spadesuit - \spadesuit)/\spadesuit$ and $I = (\spadesuit - q \cdot \spadesuit,\ \spadesuit + q \cdot \spadesuit)$

Significance test for the slope B_1

- For $i = 1, \ldots, n$ let
$$Y_i = \beta_0 + \beta_1 \cdot x_i + \sigma Z_i,$$
with $Z_1, \ldots, Z_n$ i.i.d. RVs and $Z_1 \sim N(0,1)$, and $(\beta_0, \beta_1, \sigma) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$
- $q_{1-\alpha/2}$ the $(1-\alpha/2)$-quantile of the $t(n-2)$-distribution

Under $H_0 : \beta_1 = \beta_1^{(0)}$ it holds

$$T := \frac{B_1 - \beta_1^{(0)}}{SE_{B_1}} \sim t(n-2)$$

and equivalently: the confidence interval

$$I := \left(B_1 - q_{1-\alpha/2} \cdot SE_{B_1},\ B_1 + q_{1-\alpha/2} \cdot SE_{B_1}\right)$$

overlaps $\beta_1^{(0)}$ with probability $1-\alpha$

- known structure: $T = (\spadesuit - \spadesuit)/\spadesuit$ and $I = (\spadesuit - q \cdot \spadesuit,\ \spadesuit + q \cdot \spadesuit)$
- typically: null hypothesis $\beta_1^{(0)} = 0 \leftrightarrow$ no relation

Significance test for the slope B_1

- For $i = 1, \ldots, n$ let
$$Y_i = \beta_0 + \beta_1 \cdot x_i + \sigma Z_i,$$
with $Z_1, \ldots, Z_n$ i.i.d. RVs and $Z_1 \sim N(0,1)$, and $(\beta_0, \beta_1, \sigma) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$
- $q_{1-\alpha/2}$ the $(1-\alpha/2)$-quantile of the $t(n-2)$-distribution

Under $H_0 : \beta_1 = \beta_1^{(0)}$ it holds

$$T := \frac{B_1 - \beta_1^{(0)}}{SE_{B_1}} \sim t(n-2)$$

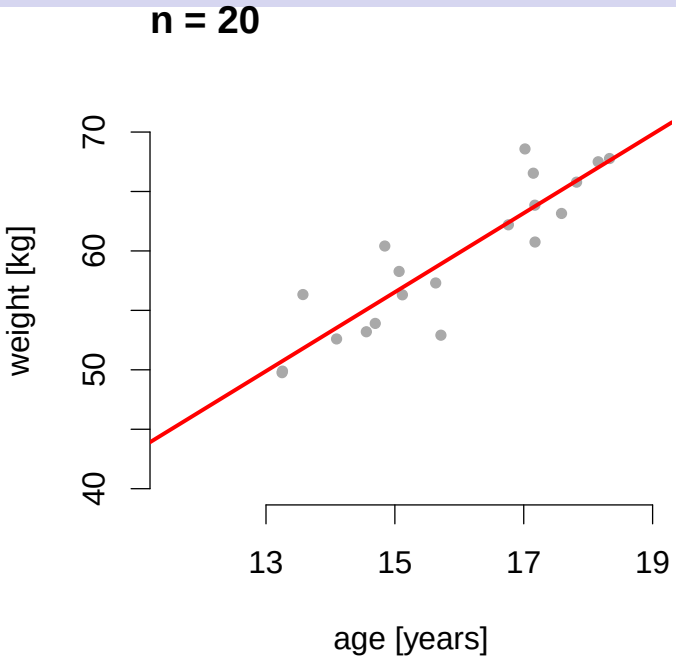
and equivalently: the confidence interval

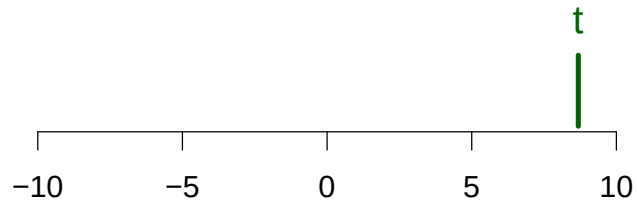
$$I := \left(B_1 - q_{1-\alpha/2} \cdot SE_{B_1},\ B_1 + q_{1-\alpha/2} \cdot SE_{B_1}\right)$$

overlaps $\beta_1^{(0)}$ with probability $1-\alpha$

- known structure: $T = (\spadesuit - \spadesuit)/\spadesuit$ and $I = (\spadesuit - q \cdot \spadesuit,\ \spadesuit + q \cdot \spadesuit)$
- typically: null hypothesis $\beta_1^{(0)} = 0 \leftrightarrow$ no relation
- In the model both β_0 and β_1 are estimated, thus $df = n - 2$

Data analysis: significance test for the slope

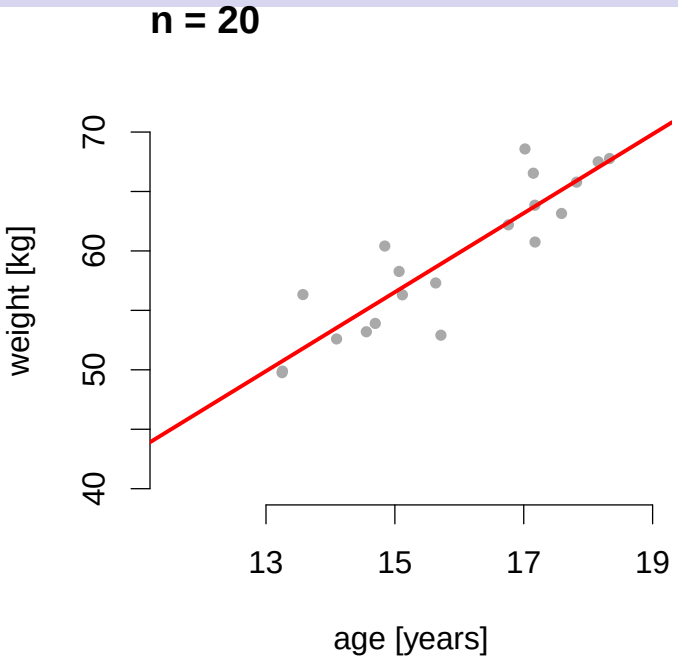


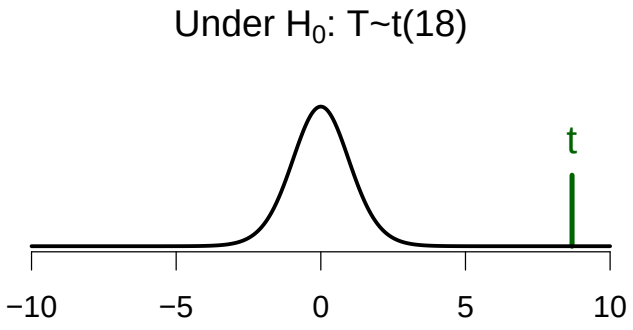


- evaluation of the data

$$t = \frac{b_1 - 0}{se_{b_1}} \approx 8.7$$

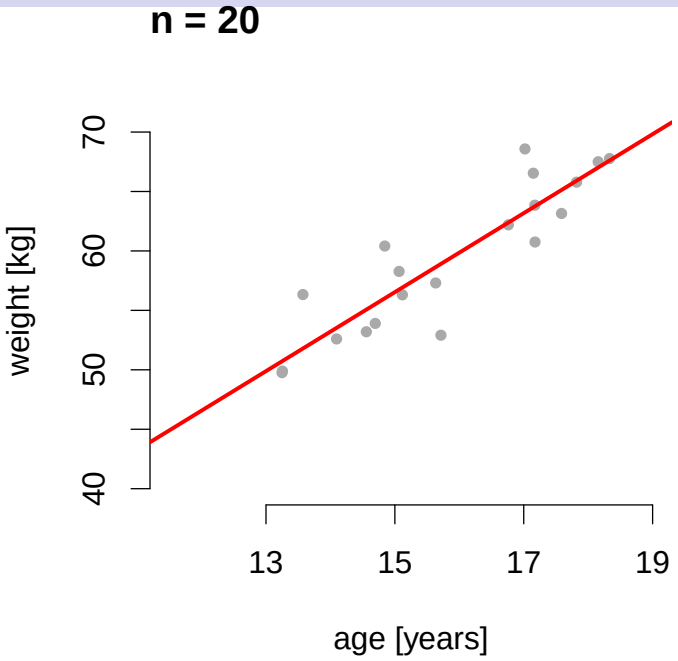
Data analysis: significance test for the slope

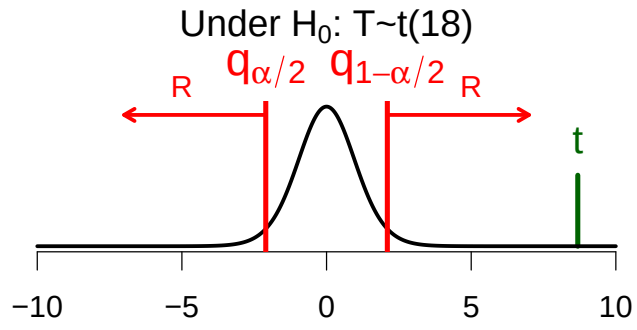




- evaluation of the data $t = \dfrac{b_1 - 0}{se_{b_1}} \approx 8.7$
- for the significance level $\alpha = 5\%$ it is $q_{1-\alpha/2} \approx 2.1$

Data analysis: significance test for the slope



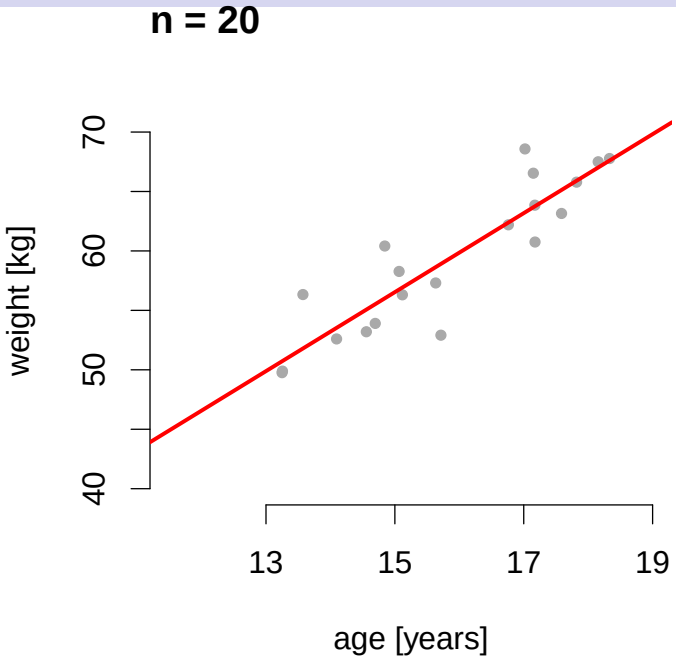


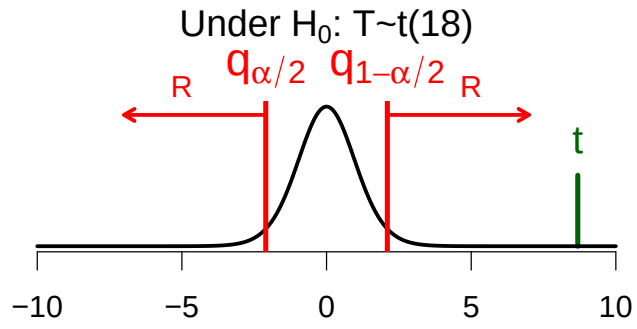
- evaluation of the data

$$t = \frac{b_1 - 0}{se_{b_1}} \approx 8.7$$

- for the significance level $\alpha = 5\%$ it is $q_{1-\alpha/2} \approx 2.1$
- $\rightarrow$ rejection area $R \approx (-\infty, -2.1] \cup [2.1, \infty)$ (two-sided)

Data analysis: significance test for the slope



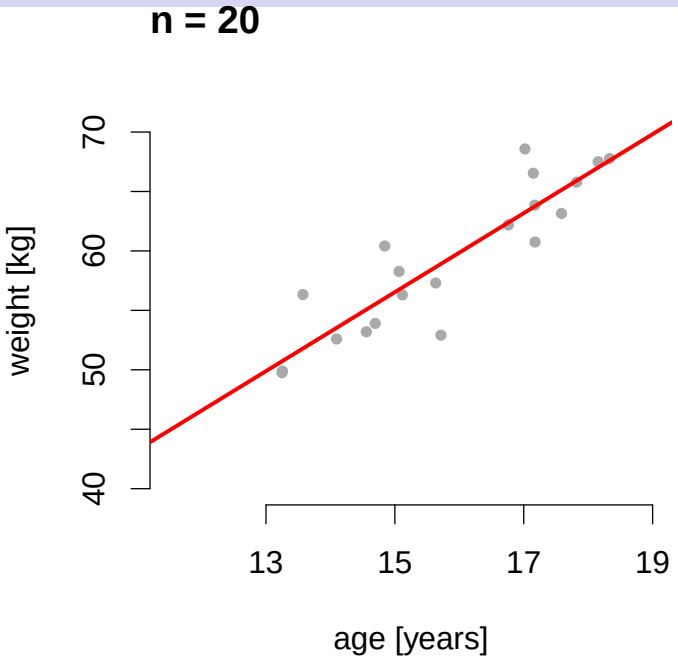


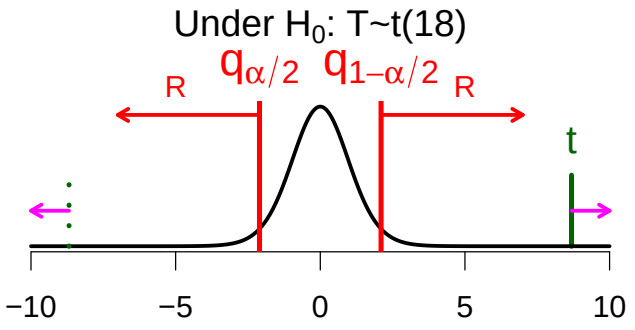
- evaluation of the data

$$t = \frac{b_1 - 0}{se_{b_1}} \approx 8.7$$

- for the significance level $\alpha = 5\%$ it is $q_{1-\alpha/2} \approx 2.1$
- $\rightarrow$ rejection area $R \approx (-\infty, -2.1] \cup [2.1, \infty)$ (two-sided)
- As $t \in R$ we reject $H_0 : \beta_1 = 0$ on the 5%-level

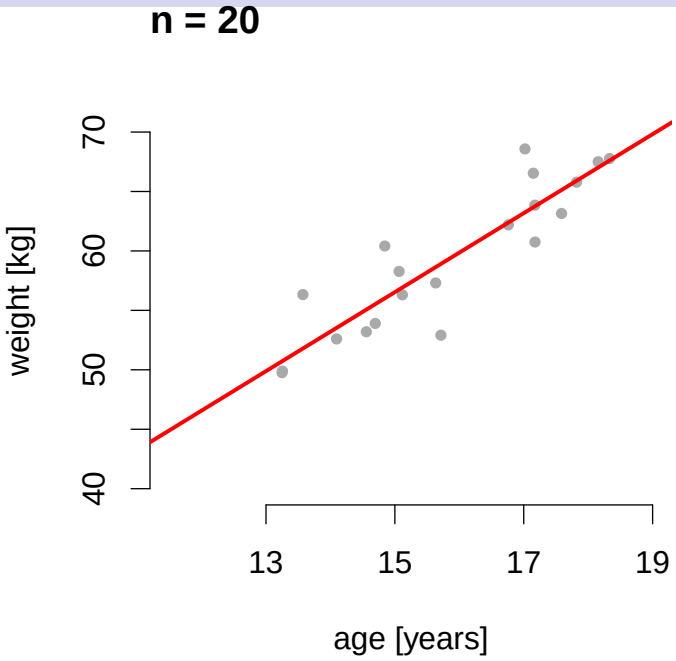
Data analysis: significance test for the slope

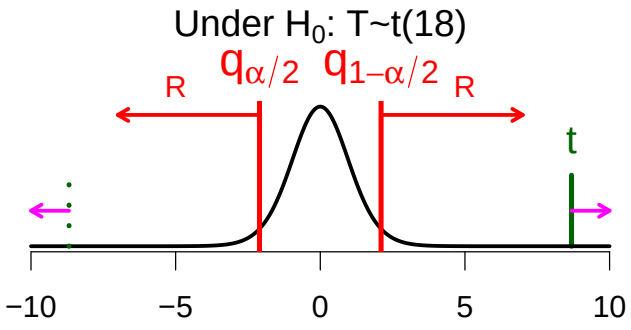




- evaluation of the data $t = \dfrac{b_1 - 0}{se_{b_1}} \approx 8.7$
- for the significance level $\alpha = 5\%$ it is $q_{1-\alpha/2} \approx 2.1$
- $\rightarrow$ rejection area $R \approx (-\infty, -2.1] \cup [2.1, \infty)$ (two-sided)
- As $t \in R$ we reject $H_0 : \beta_1 = 0$ on the 5%-level
- the p-value is $p = \mathbb{P}_{H_0}(|T| \geqslant |t|) \approx 7.5 \cdot 10^{-8}$ (tiny)

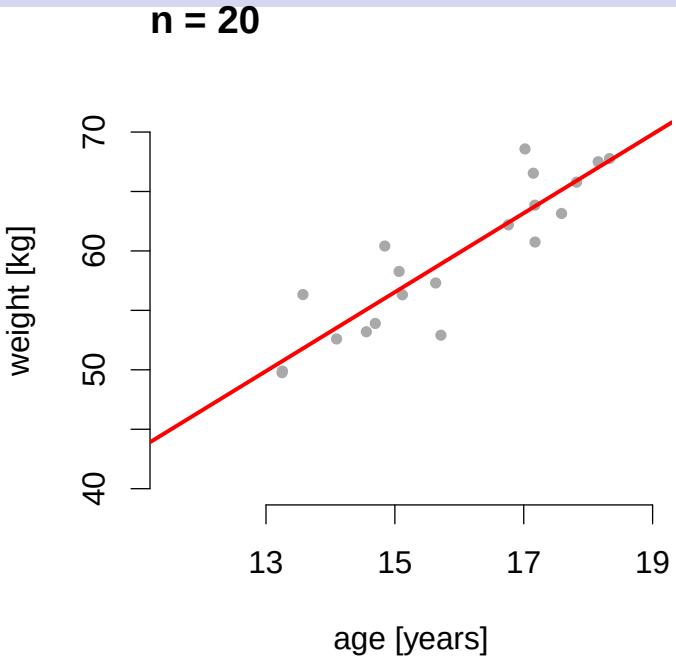
Data analysis: significance test for the slope

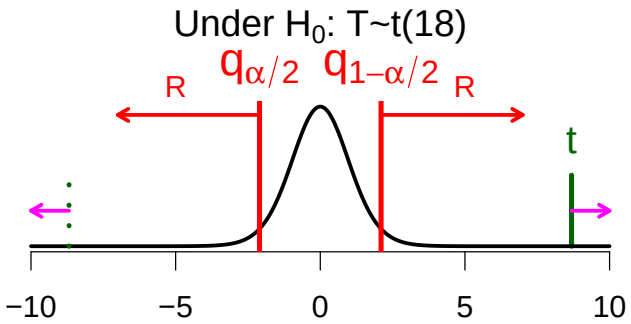




- evaluation of the data $t = \dfrac{b_1 - 0}{se_{b_1}} \approx 8.7$
- for the significance level $\alpha = 5\%$ it is $q_{1-\alpha/2} \approx 2.1$
- $\rightarrow$ rejection area $R \approx (-\infty, -2.1] \cup [2.1, \infty)$ (two-sided)
- As $t \in R$ we reject $H_0 : \beta_1 = 0$ on the 5%-level
- the p-value is $p = \mathbb{P}_{H_0}(|T| \geqslant |t|) \approx 7.5 \cdot 10^{-8}$ (tiny)
- Interpretation: The positive relation observed in the data is barely compatible with the assertion that there is no relation.

Data analysis: significance test for the slope





- evaluation of the data $t = \dfrac{b_1 - 0}{se_{b_1}} \approx 8.7$
- for the significance level $\alpha = 5\%$ it is $q_{1-\alpha/2} \approx 2.1$
- $\rightarrow$ rejection area $R \approx (-\infty, -2.1] \cup [2.1, \infty)$ (two-sided)
- As $t \in R$ we reject $H_0 : \beta_1 = 0$ on the 5%-level
- the p-value is $p = \mathbb{P}_{H_0}(|T| \geqslant |t|) \approx 7.5 \cdot 10^{-8}$ (tiny)
- Interpretation: The positive relation observed in the data is barely compatible with the assertion that there is no relation. If H_0 holds true, then we observe in less than one of 10^7 cases a relation that is at least as extreme as the relation observed ($p < 10^{-7}$)

Significance test for slope B_1 using R

```
# Enter data, x- and y- values as vectors
x <- c(...)
y <- c(...)
# compute regression line
rg <- lm(y~x)
# perform test
summary(rg)
# Output
...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.7005     6.0949   1.099    0.286
x              3.3222     0.3826   8.684 7.47e-08 ***
...
```

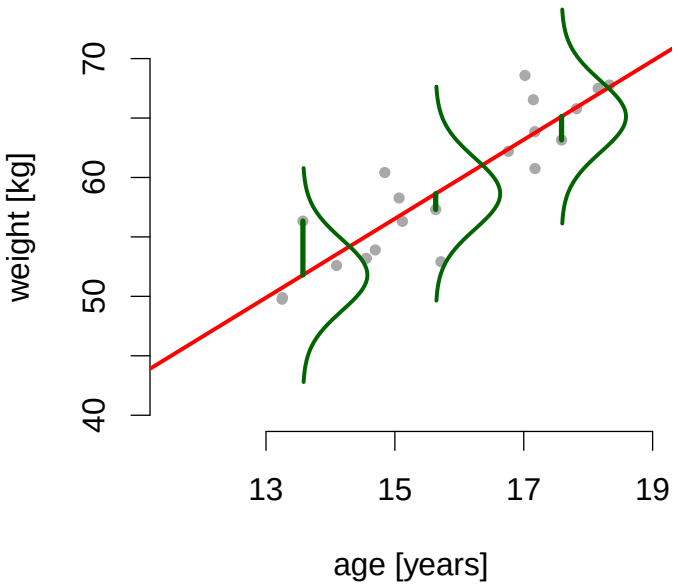
- 2nd row: slope, 1st row: intercept (usually not interesting)
- $b_1 \approx 3.3$, $se_{b_1} \approx 0.4$, $t \approx 8.7$, $p \approx 7.5 \cdot 10^{-8}$
- there are also other statistics returned (not shown here), for example a summary of the residuals or the standard error of the regression s_r, etc.

Check model assumptions

- Model: For $i = 1, \ldots, n$ let

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \sigma Z_i,$$

with $Z_1, \ldots, Z_n$ i.i.d. RVs and $Z_1 \sim N(0,1)$, and $(\beta_0, \beta_1, \sigma) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$



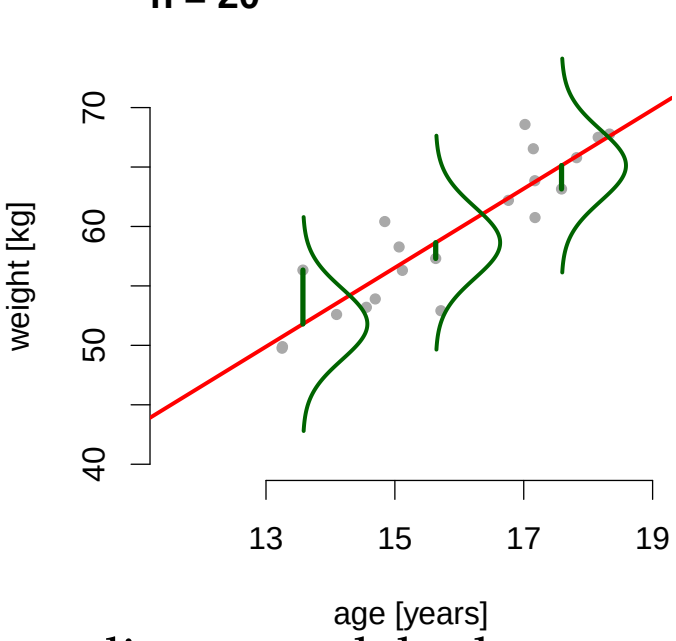
- linear model: observe a linear proportion plus error

Check model assumptions

- Model: For $i = 1, \ldots, n$ let

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \sigma Z_i,$$

with $Z_1, \ldots, Z_n$ i.i.d. RVs and $Z_1 \sim N(0,1)$, and $(\beta_0, \beta_1, \sigma) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$



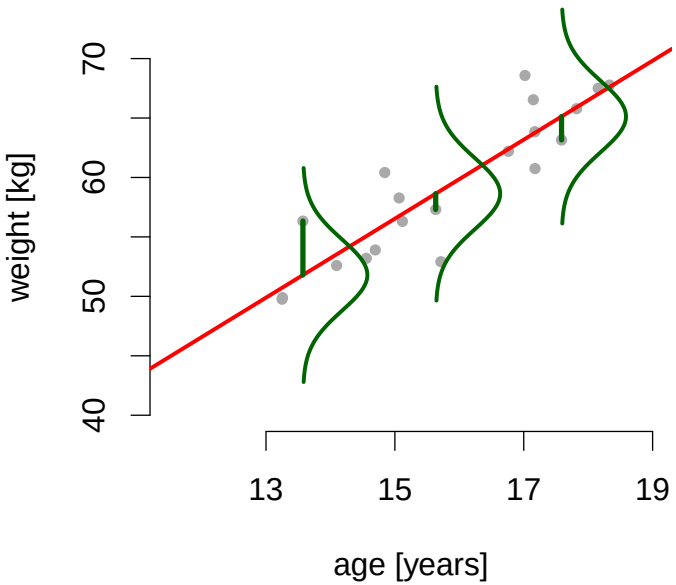
- linear model: observe a linear proportion plus error
- normal distributed errors: the data are distribute bell-shaped around the line

Check model assumptions

- Model: For $i = 1, \ldots, n$ let

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \sigma Z_i,$$

with $Z_1, \ldots, Z_n$ i.i.d. RVs and $Z_1 \sim N(0,1)$, and $(\beta_0, \beta_1, \sigma) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$



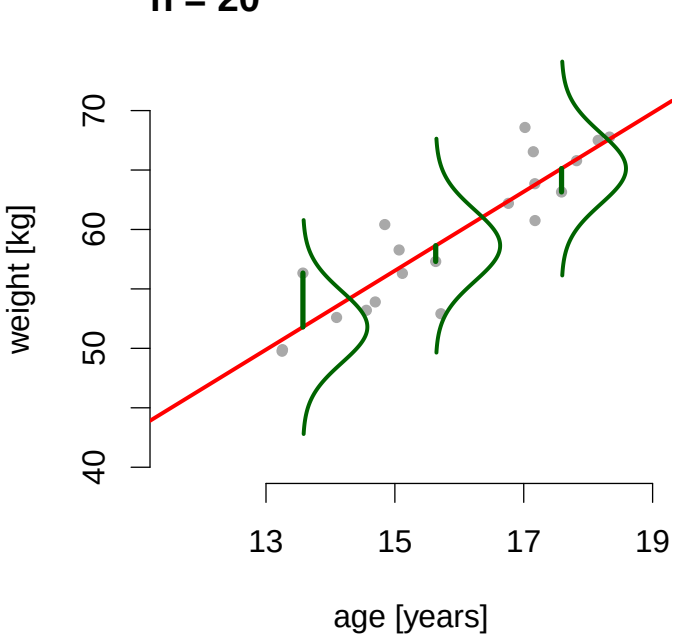
- linear model: observe a linear proportion plus error
- normal distributed errors: the data are distribute bell-shaped around the line
- constant variance σ^2: the spread of the errors does not change with age

Check model assumptions

- Model: For $i = 1, \ldots, n$ let

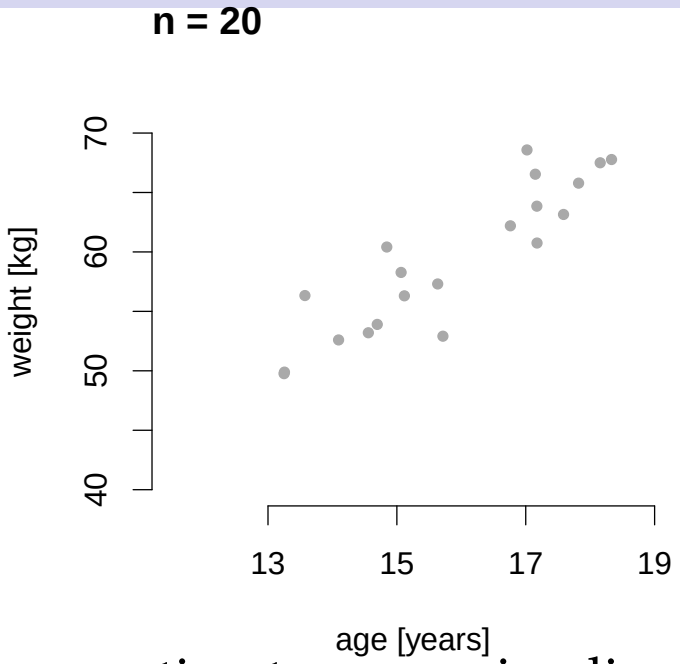
$$Y_i = \beta_0 + \beta_1 \cdot x_i + \sigma Z_i,$$

with $Z_1, \ldots, Z_n$ i.i.d. RVs and $Z_1 \sim N(0,1)$, and $(\beta_0, \beta_1, \sigma) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$



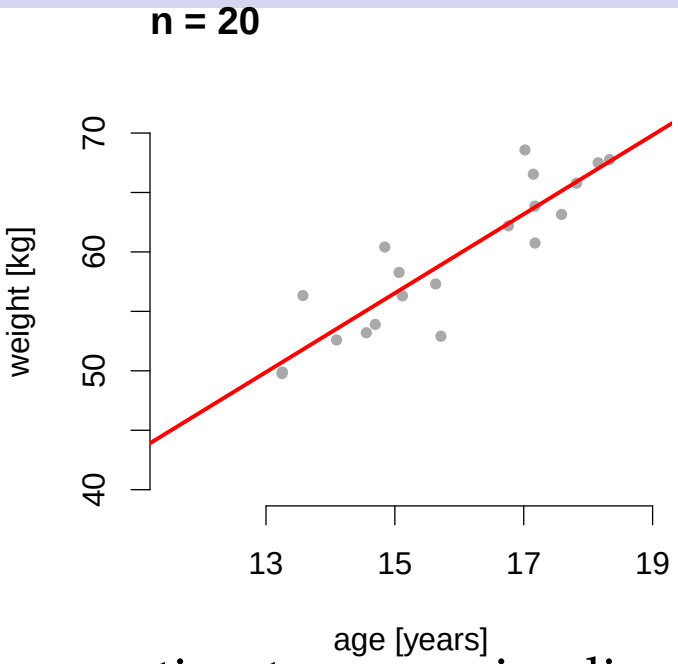
- linear model: observe a linear proportion plus error
- normal distributed errors: the data are distribute bell-shaped around the line
- constant variance σ^2: the spread of the errors does not change with age
- independence: it is plausible to assume the errors as independent, as the teenagers were 'randomly' chosen for the survey. Besides linearity we observe no further 'structure' in the data

Linear regression naively



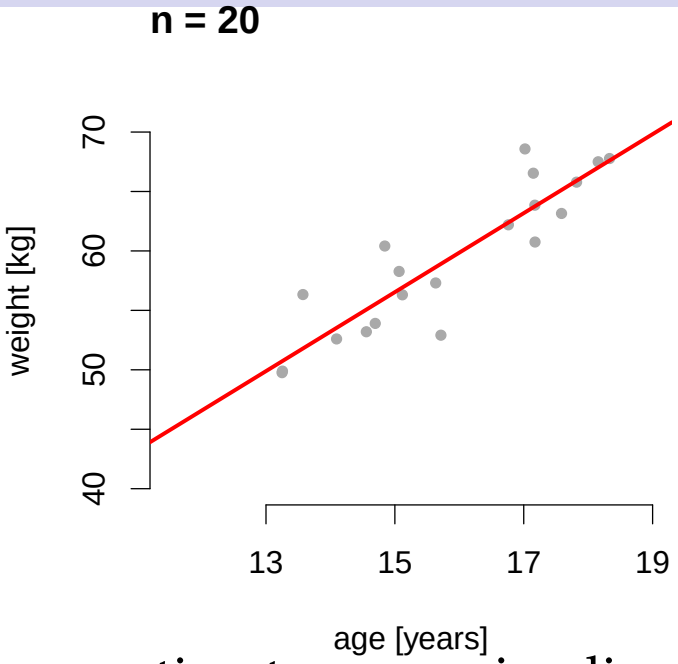
- estimate regression line via eye

Linear regression naively



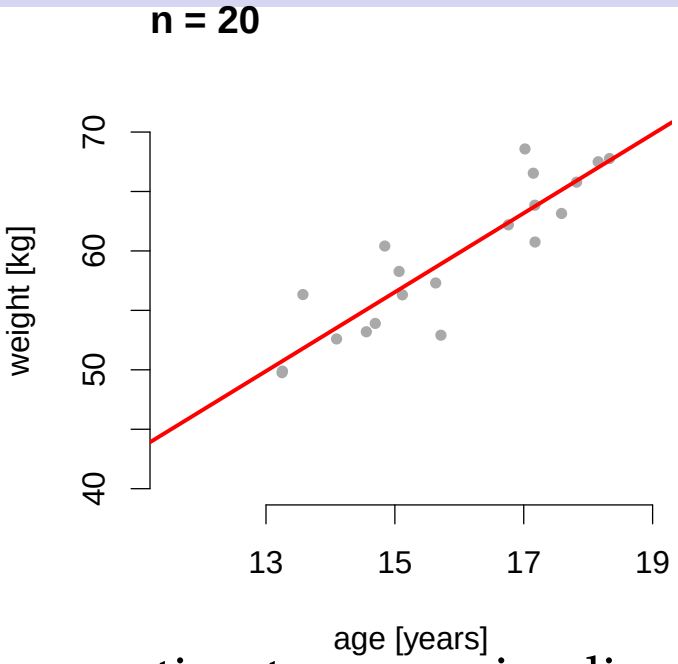
- estimate regression line via eye

Linear regression naively



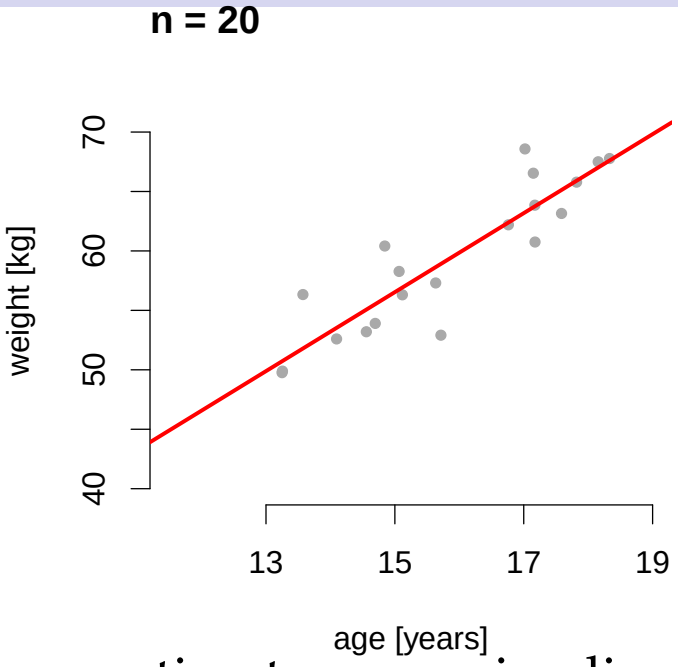
- estimate regression line via eye
- slope: $b_1 \approx (70 - 50)/(19 - 13) = 20/6 \approx 3$

Linear regression naively



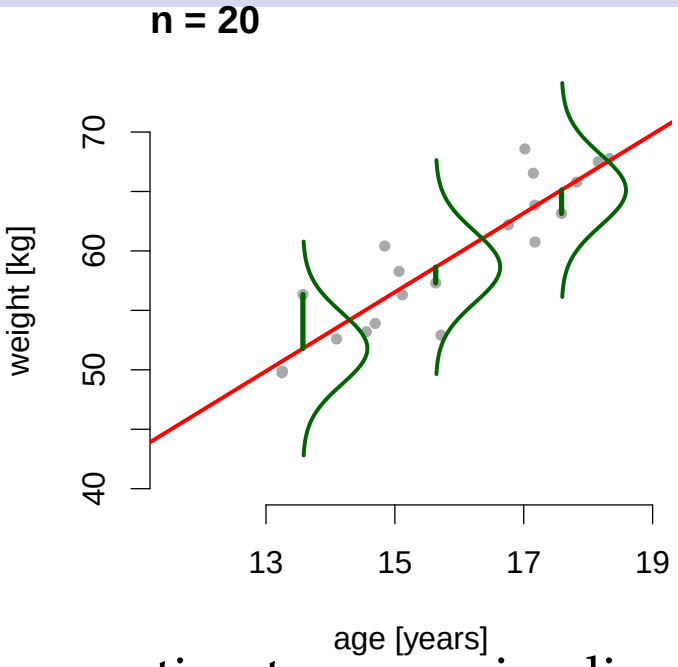
- estimate regression line via eye
- slope: $b_1 \approx (70 - 50)/(19 - 13) = 20/6 \approx 3$
- standard error of the slope: $se_{b_1} = s_r/(s_x \cdot \sqrt{n-1})$
  - $\sqrt{n-1} = \sqrt{19} \approx 4$

Linear regression naively



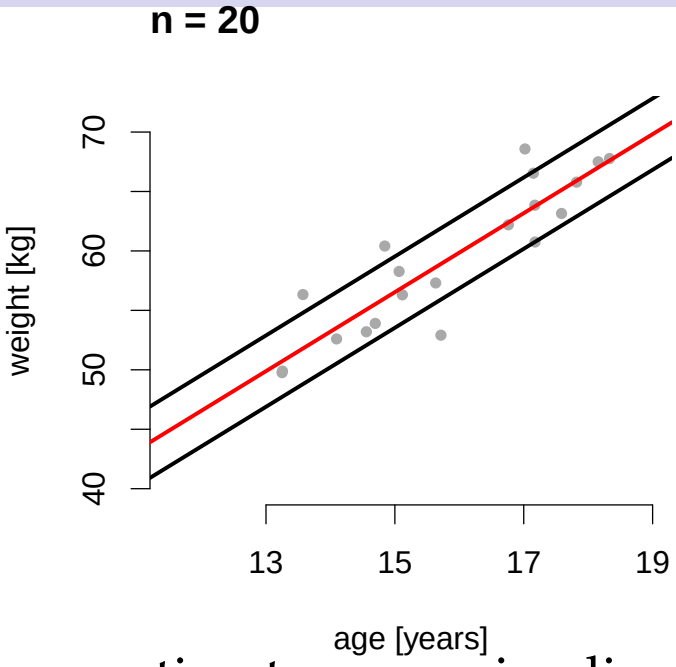
- estimate regression line via eye
- slope: $b_1 \approx (70 - 50)/(19 - 13) = 20/6 \approx 3$
- standard error of the slope: $se_{b_1} = s_r/(s_x \cdot \sqrt{n-1})$
  - $\sqrt{n-1} = \sqrt{19} \approx 4$
  - standard deviation of the data $(x_i)_i$: $s_x \approx 2$, as $\bar{x} \approx 16$ and about 2/3 of the data in $[14, 18]$

Linear regression naively



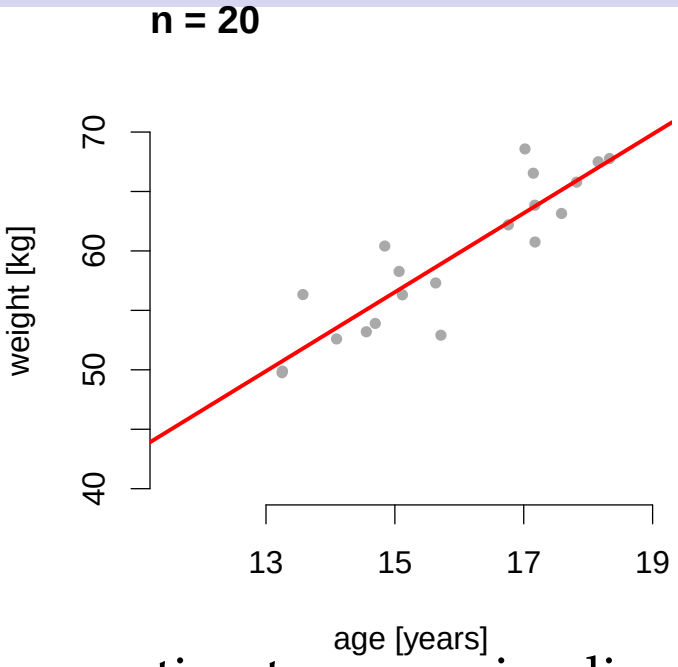
- estimate regression line via eye
- slope: $b_1 \approx (70-50)/(19-13) = 20/6 \approx 3$
- standard error of the slope: $se_{b_1} = s_r/(s_x \cdot \sqrt{n-1})$
  - $\sqrt{n-1} = \sqrt{19} \approx 4$
  - standard deviation of the data $(x_i)_i$: $s_x \approx 2$, as $\bar{x} \approx 16$ and about 2/3 of the data in $[14, 18]$
  - standard error of the regression: $s_r \approx 3$, captures about 2/3 of the data 'around the regression line'

Linear regression naively



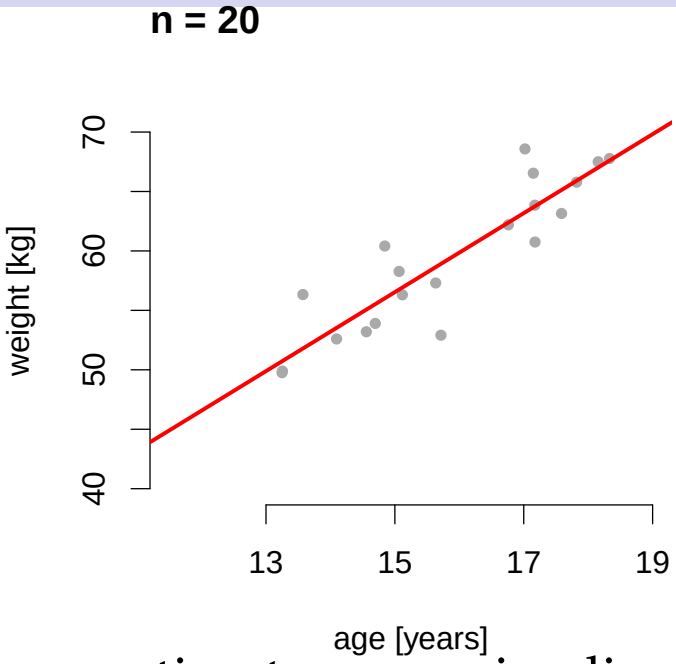
- estimate regression line via eye
- slope: $b_1 \approx (70-50)/(19-13) = 20/6 \approx 3$
- standard error of the slope: $se_{b_1} = s_r/(s_x \cdot \sqrt{n-1}) \approx 3/(2 \cdot 4) = 3/8$
  - $\sqrt{n-1} = \sqrt{19} \approx 4$
  - standard deviation of the data $(x_i)_i$: $s_x \approx 2$, as $\bar{x} \approx 16$ and about 2/3 of the data in $[14, 18]$
  - standard error of the regression: $s_r \approx 3$, captures about 2/3 of the data 'around the regression line'

Linear regression naively



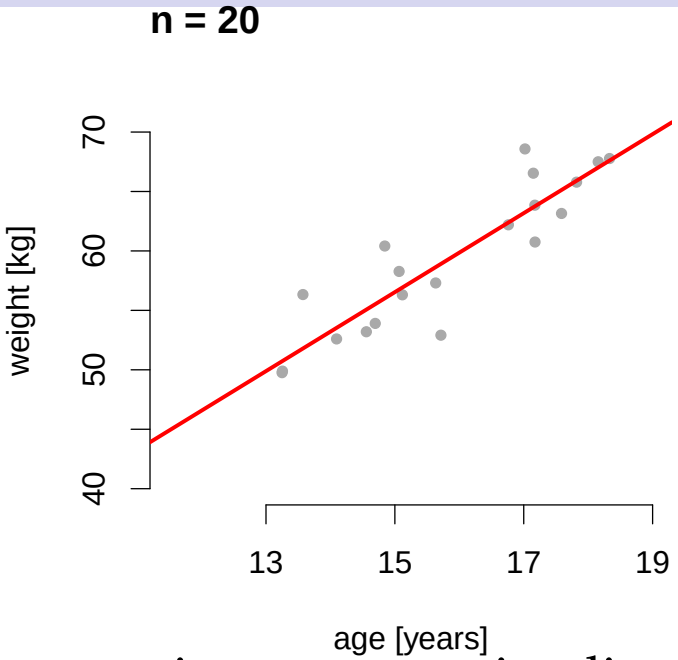
- estimate regression line via eye
- slope: $b_1 \approx (70-50)/(19-13) = 20/6 \approx 3$
- standard error of the slope: $se_{b_1} = s_r/(s_x \cdot \sqrt{n-1}) \approx 3/(2 \cdot 4) = 3/8$
  - $\sqrt{n-1} = \sqrt{19} \approx 4$
  - standard deviation of the data $(x_i)_i$: $s_x \approx 2$, as $\bar{x} \approx 16$ and about 2/3 of the data in $[14, 18]$
  - standard error of the regression: $s_r \approx 3$, captures about 2/3 of the data 'around the regression line'
- $t = b_1/se_{b_1} \approx 8$ (huge!)

Linear regression naively



- estimate regression line via eye
- slope: $b_1 \approx (70-50)/(19-13) = 20/6 \approx 3$
- standard error of the slope: $se_{b_1} = s_r/(s_x \cdot \sqrt{n-1}) \approx 3/(2 \cdot 4) = 3/8$
  - $\sqrt{n-1} = \sqrt{19} \approx 4$
  - standard deviation of the data $(x_i)_i$: $s_x \approx 2$, as $\bar{x} \approx 16$ and about 2/3 of the data in $[14, 18]$
  - standard error of the regression: $s_r \approx 3$, captures about 2/3 of the data 'around the regression line'
- $t = b_1/se_{b_1} \approx 8$ (huge!)
- for $\alpha = 5\%$ the rejection area is $R \approx (-\infty, -2] \cup [2, \infty)$

Linear regression naively



- estimate regression line via eye
- slope: $b_1 \approx (70-50)/(19-13) = 20/6 \approx 3$
- standard error of the slope: $se_{b_1} = s_r/(s_x \cdot \sqrt{n-1}) \approx 3/(2 \cdot 4) = 3/8$
  - $\sqrt{n-1} = \sqrt{19} \approx 4$
  - standard deviation of the data $(x_i)_i$: $s_x \approx 2$, as $\bar{x} \approx 16$ and about 2/3 of the data in $[14, 18]$
  - standard error of the regression: $s_r \approx 3$, captures about 2/3 of the data 'around the regression line'
- $t = b_1/se_{b_1} \approx 8$ (huge!)
- for $\alpha = 5\%$ the rejection area is $R \approx (-\infty, -2] \cup [2, \infty)$
- reject H_0

Thank you!