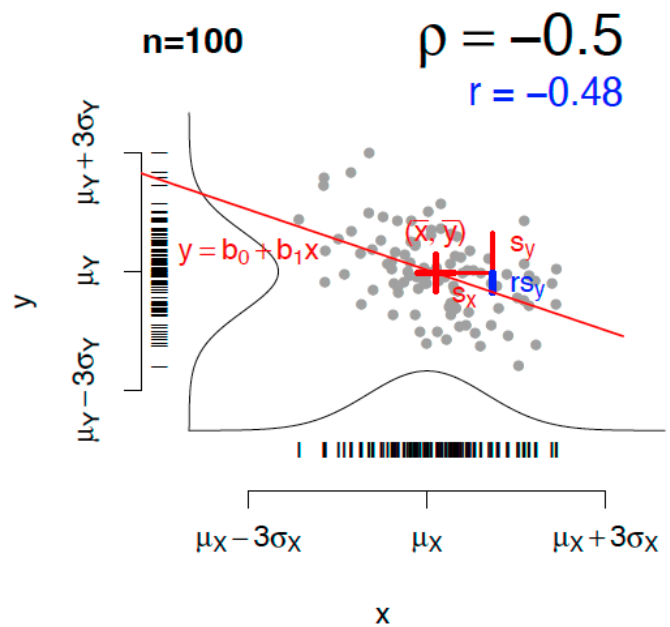
TECHNISCHE UNIVERSITÄT WIEN

Linear regression



Slides were developed together with Dr. Michael Messer, former ASTAT member
Please send your comments to: T. Levajković, tijana.levajkovic@tuwien.ac.at

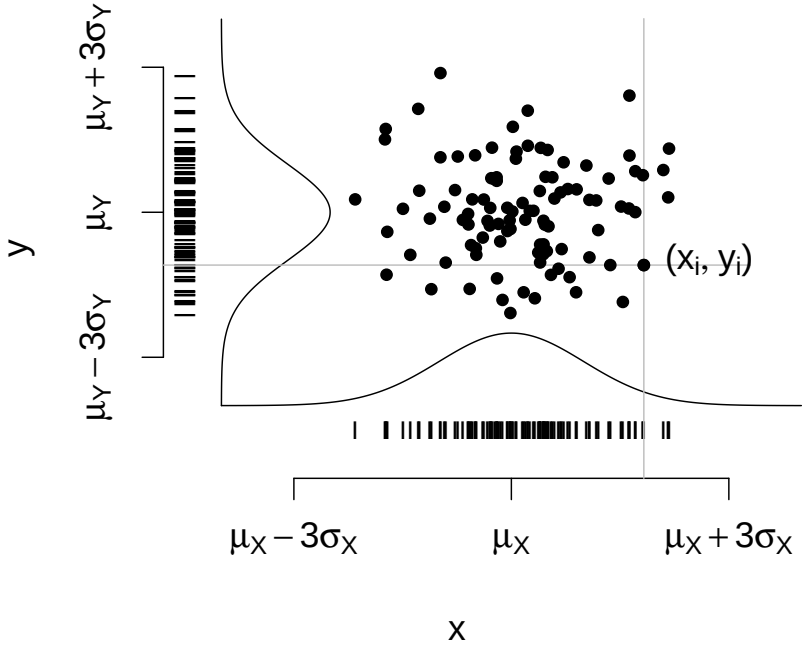
Reminder: Correlation

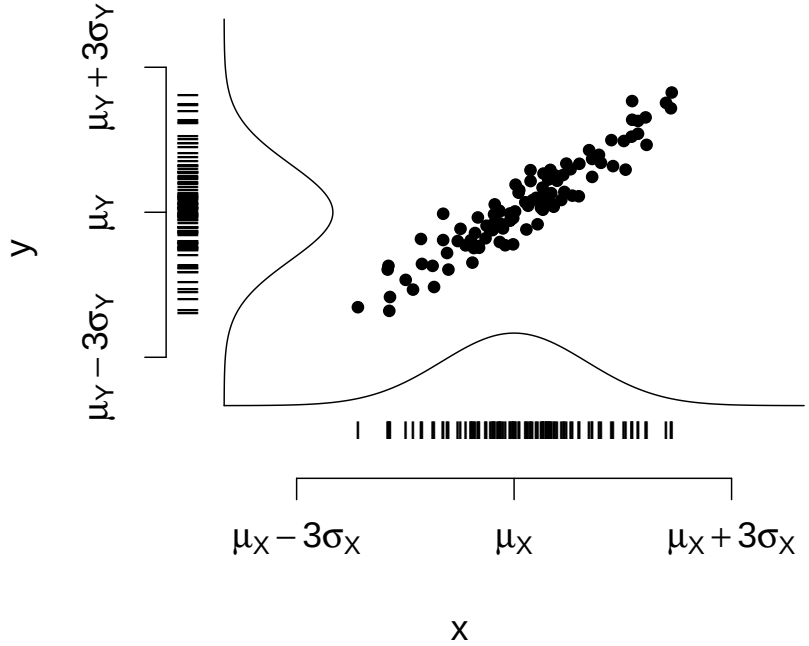
- $X_1, \ldots, X_n$ i.i.d. RVs, $X_1 \sim N(\mu_x, \sigma_x^2)$. Here $n = 100$
- $Y_1, \ldots, Y_n$ i.i.d. RVs, $Y_1 \sim N(\mu_y, \sigma_y^2)$
- also let the pairs $(X_i, Y_i)_{i=1,\ldots,n}$ be independent over $i = 1, 2, \ldots$
- So long, nothing said about the *relation* between X_i and Y_i
-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

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

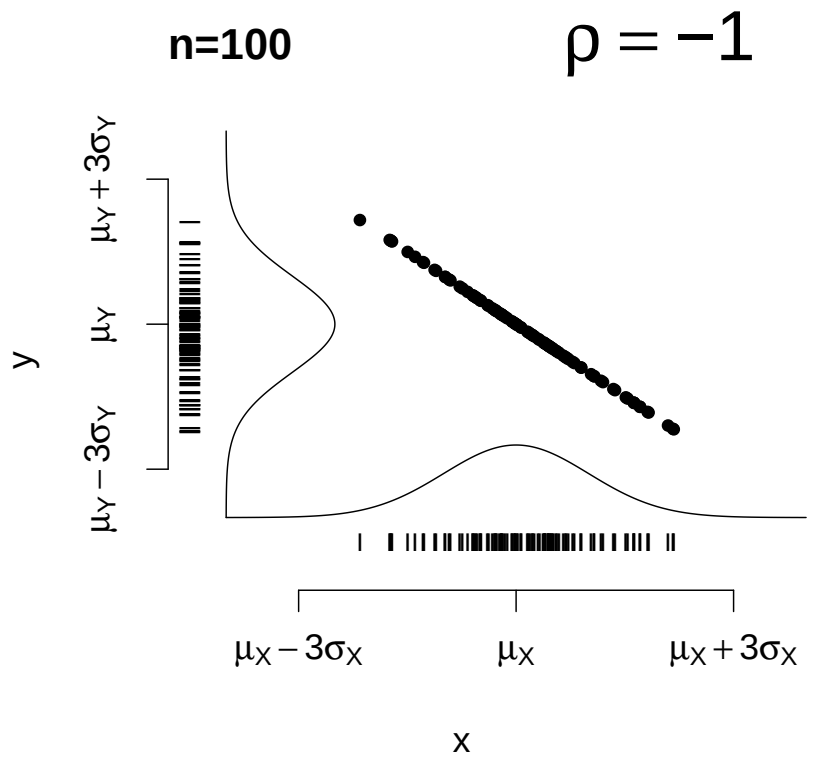
n=100 $\rho = 0$ **n=100** $\rho = 0.95$





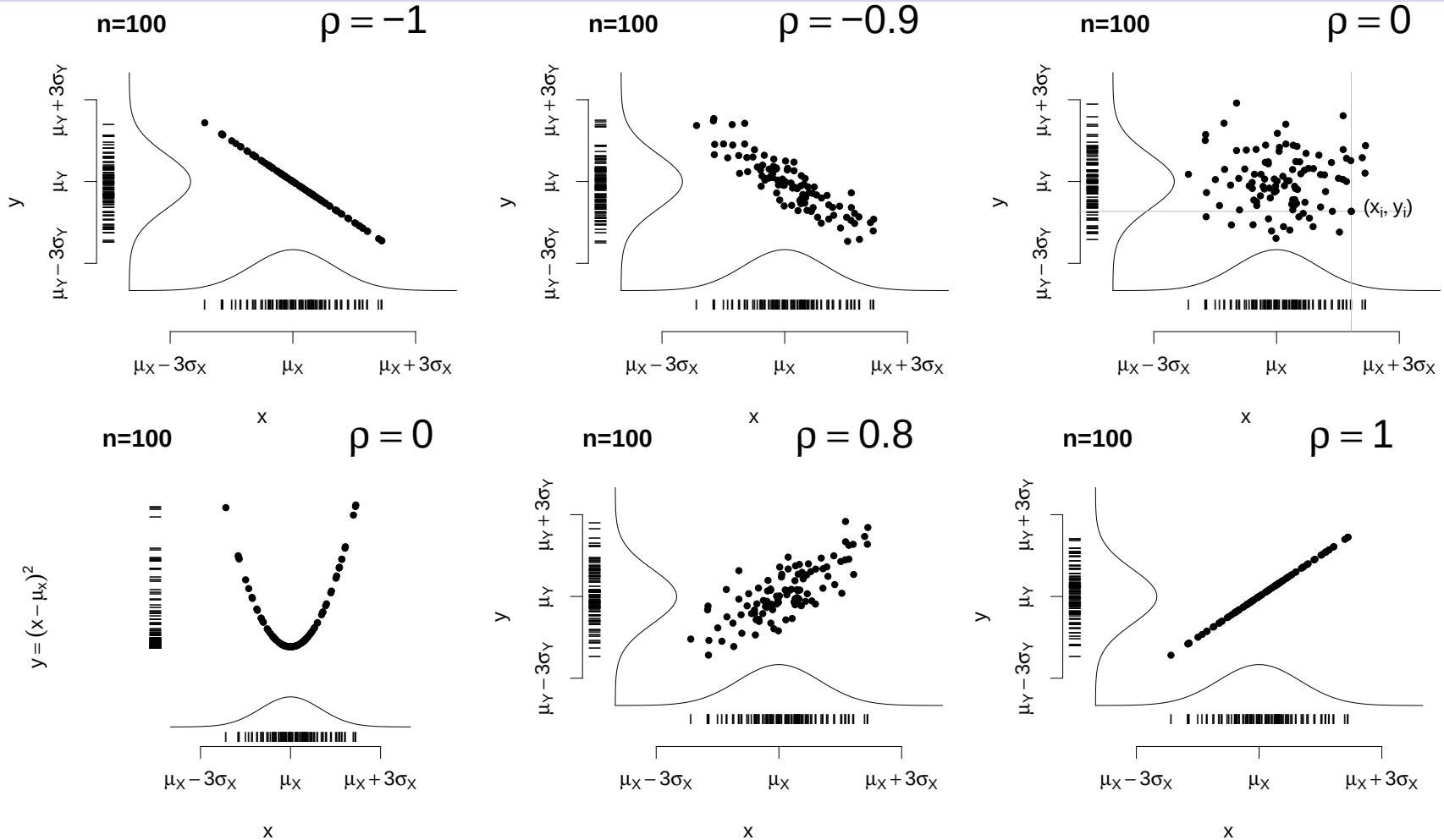
Examples: Correlation

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



- 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

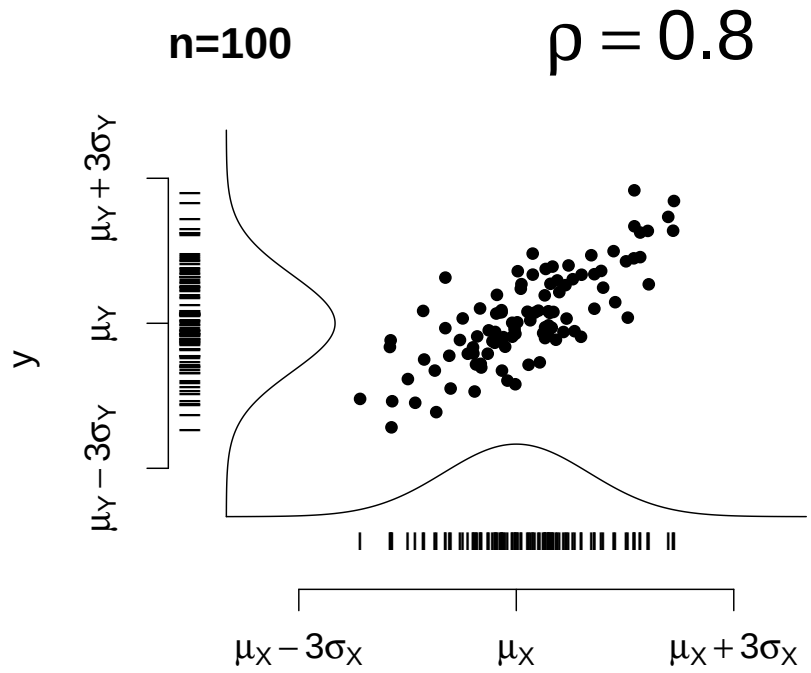


- The correlation ρ is a measure for the degree of the *linear* relation
- $\rho = 0 \Leftrightarrow$ no linear relation (say: X and Y are *uncorrelated*)
- $\rho > 0 \Leftrightarrow$ positive linear relation
- $\rho < 0 \Leftrightarrow$ negative linear relation
- $|\rho| = 1 \Leftrightarrow$ perfect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

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

Empirical correlation

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

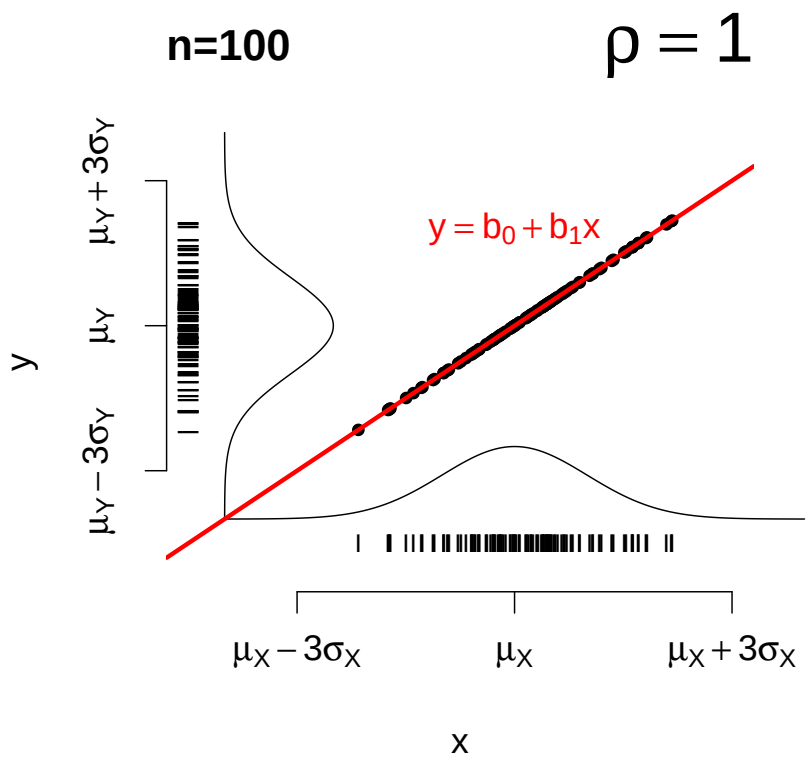


- For realizations $(x_i, y_i)_{i=1,2,\ldots,n}$ estimate ρ through the *empirical correlation*

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

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

Perfect linear relation for $\rho = 1$

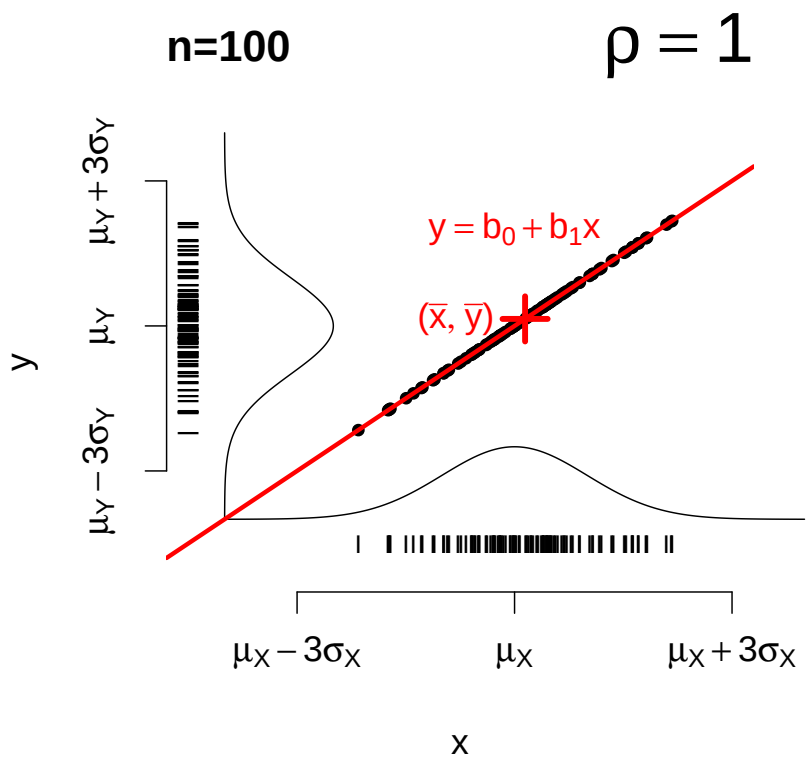


- $\rho = 1$: the points lie on a line $y = b_0 + b_1 x$.
- For the slope b_1 and the intercept b_0 it holds

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

Perfect linear relation for $\rho = 1$

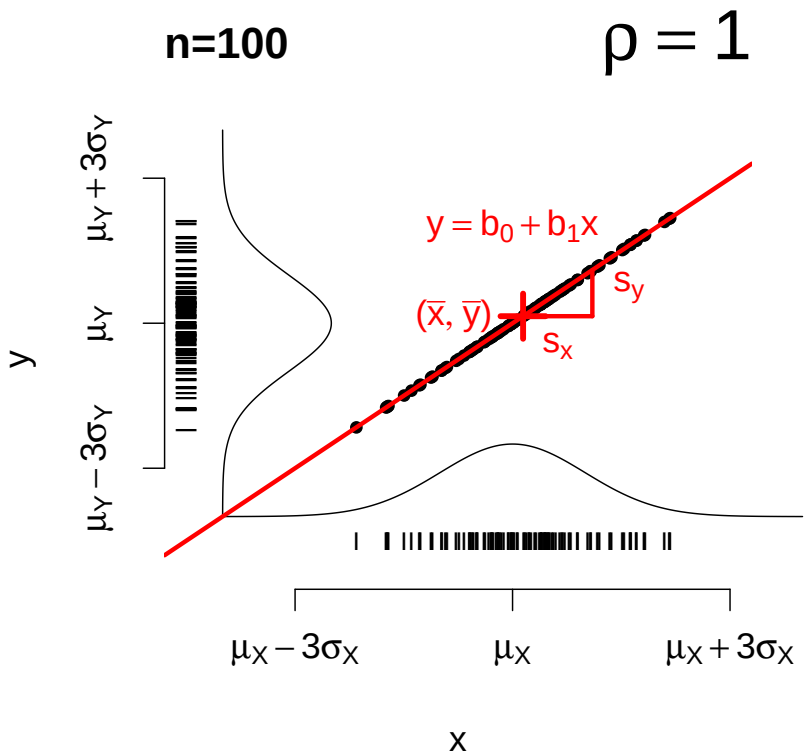
$$b_1 = \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$



- ad intercept b_0:
  - 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 $\rho = 1$

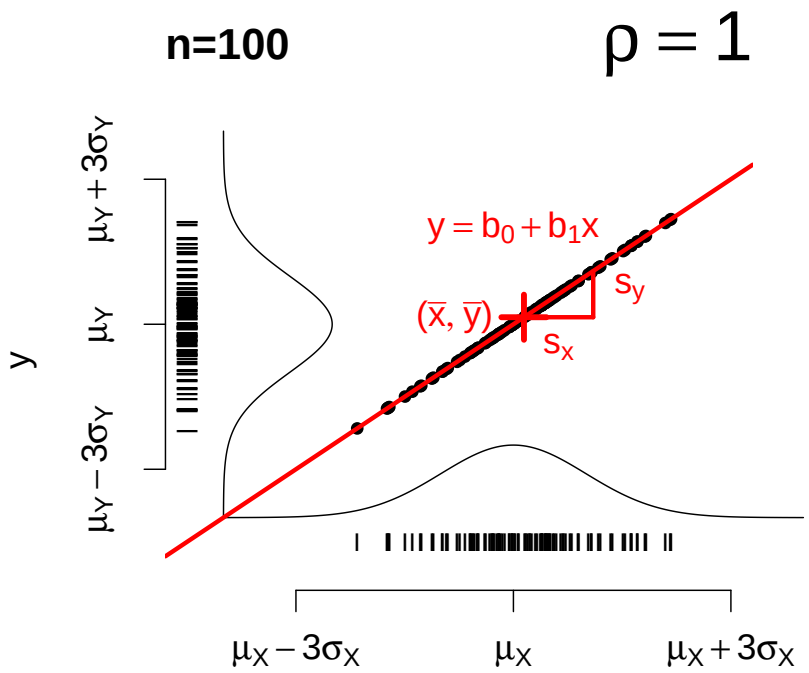
$$b_1 = \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$



- ad slope b_1:
  - 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 $\rho = 1$

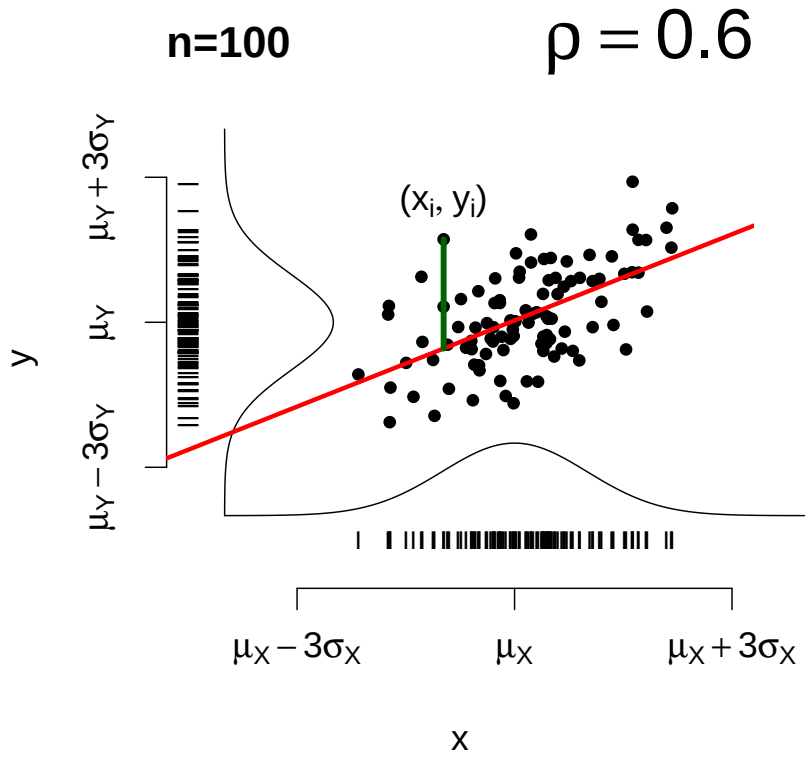
$$b_1 = \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$



In summary:

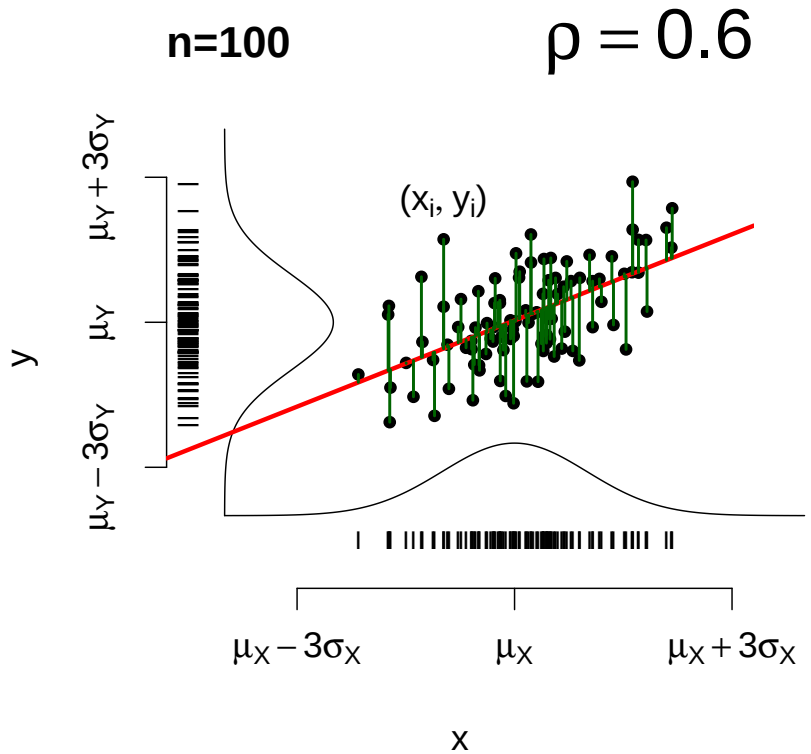
- y_i is 'explained' by x_i
- the line passes the center of mass $(\bar{x}, \bar{y})$
- regarding the slope think in the standard deviations
  - one step to the right of size s_x results in an increase of size s_y
  - but this particular slope is a consequence of the special case $\rho = 1$
  - general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

General: linear relation plus error



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



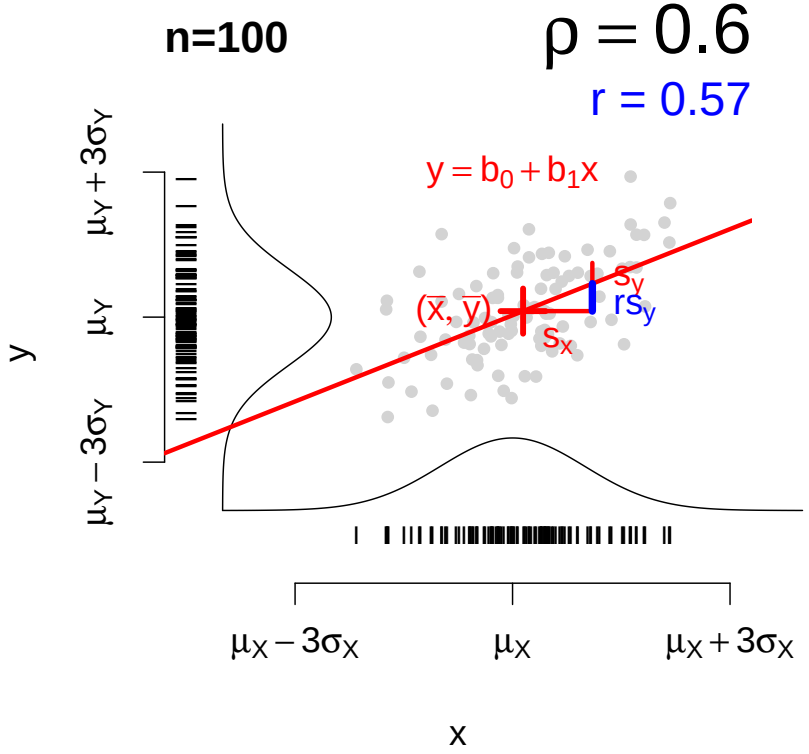
- $y_i = \beta_0 + \beta_1 x_i + e_i$
- in principle there are many possible lines
- Definition: the line, which minimizes the sum of squares of the residuals, is called the *regression line*
- i.e., search β_0 and β_1 such that $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_i])^2$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

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



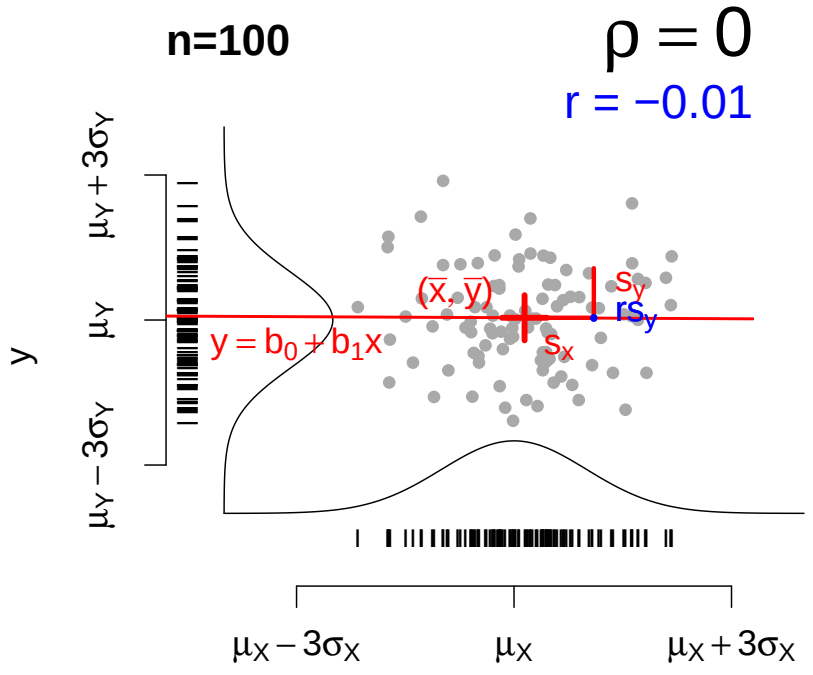
Meaning:

- 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

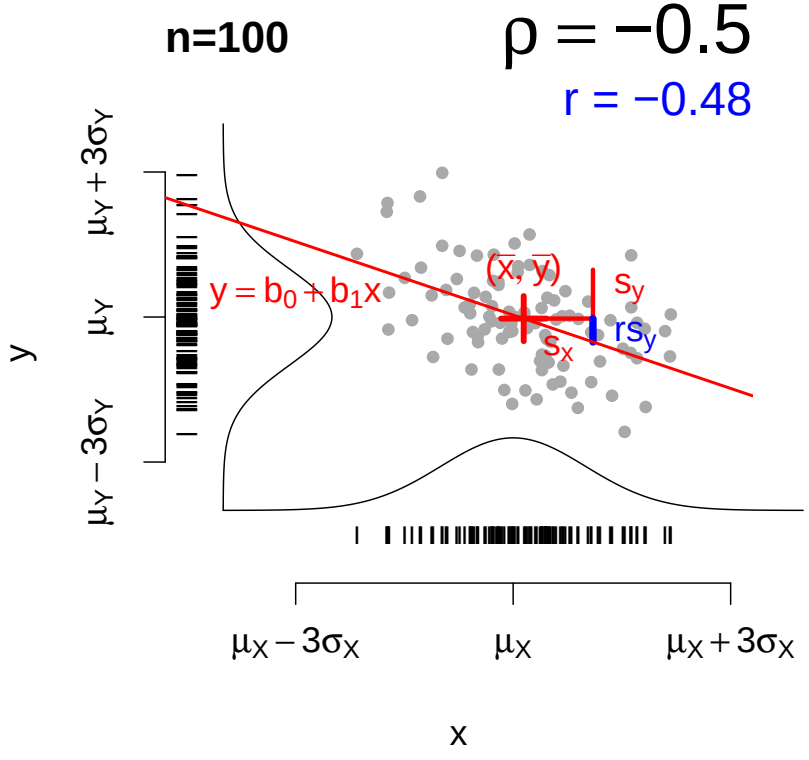
$$b_1 = r \cdot \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$



- 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

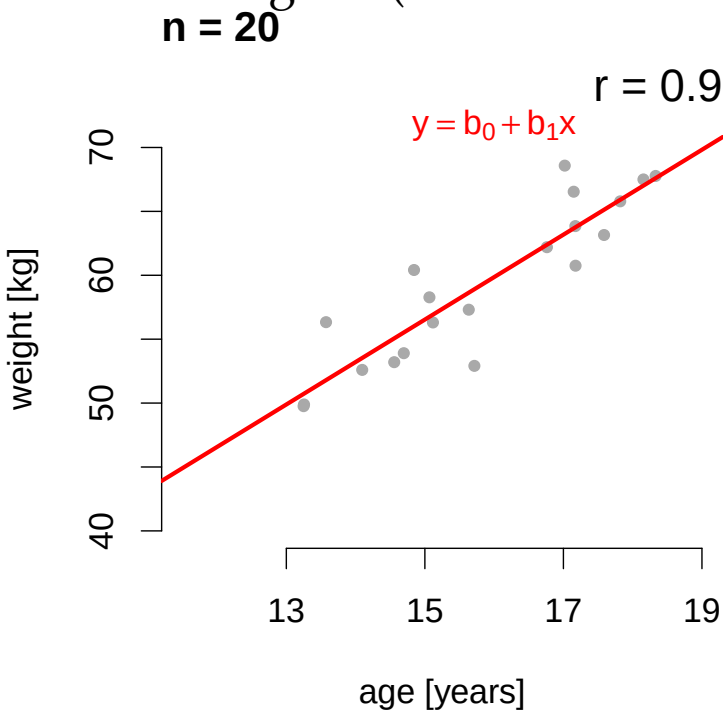
Regressionsgerade: examples

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



Data analysis: regression line

- Is there a relation between age and weight in teenage years?
- $n = 20$ 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 $b_0 = \bar{y} - b_1 \cdot \bar{x}$.

Multiple-choice questions

(1) If there is a very strong correlation between two random variables then the correlation coefficient must be

a. any value larger than 1
b. much smaller than 0, if the correlation is negative
c. much larger than 0, regardless of whether the correlation is negative or positive
d. none of the rest

Multiple-choice questions

(2) In a linear regression model ('y_i modeled as a linear function of x_i plus error') the parameters are estimated via least squares. For the mean and the empirical variance of the x and y values we obtain $\bar{x} = 5$, $s_x^2 = 4$, $\bar{y} = 7$ and $s_y^2 = 9$. It holds that

a. the regression line goes through $(5, 6)$
b. the regression line goes through $(7, 7)$
c. the slope of the regression line is smaller or equals 1.5
d. the slope of the regression line is larger than 1.5

Multiple-choice questions

(3) The relationship between number of beers consumed (x) and blood alcohol content (y) was studied in 20 students by using linear regression. The regression line obtained from the study

$$y = -0.013 + 0.018 \cdot x$$

implies that

a. each beer consumed increases blood alcohol by 1.3%.
b. on average it takes 1.8 beers to increase blood alcohol content by 1%.
c. each beer consumed increases blood alcohol by an average of amount of 1.8%.
d. each beer consumed increases blood alcohol by exactly 0.018.

Multiple-choice questions

(4) Regression analysis was used to study the relationship between return rate x (the percentage of birds that return to the colony in a given year) and immigration rate y (the percentage of new adults that join the colony per year). Based on the obtained regression line

$$y = 30 - 0.35 \cdot x,$$

if the return rate were to decrease by 10% then the rate of immigration to the colony would

a. increase by 35%
b. increase by 3.5%
c. decrease by 0.35%
d. decrease by 3.5%

Multiple-choice questions

(5) If the correlation coefficient is a positive value, then the slope b_1 of the regression line $y = b_0 + b_1 \cdot x$

a. must also be positive
b. can be either negative or positive
c. can be zero
d. can not be zero

Multiple-choice questions

(6) Linear regression was applied to obtain a relationship between the sales y (in euro) and advertising x (in euro) across all the branches of an international company. The obtained regression line is

$$y = 5000 + 7.25 \cdot x.$$

If the advertising budgets of two branches of the company differ by 30000 euro then what is the predicted difference in their sales (in euro)?

a. 217500
b. 222500
c. 5000
d. 7.25

Multiple-choice questions

(7) Anna studied the impact of the dose of a new drug treatment for high blood pressure. She thinks that the drug might be more effective in people with very high blood pressure. Because she expects a bigger change in those patients who start the treatment with high blood pressure, she uses linear regression to analyze the relationship between the initial blood pressure of a patient x and the change in blood pressure after treatment with the new drug y. If Anna obtained a very strong positive relationship between these variables, then:

a. there is evidence that the higher the patients initial blood pressure, the bigger the impact of the new drug.

b. there is evidence that the higher the patients initial blood pressure, the smaller the impact of the new drug.

c. there is evidence for an association of some kind between the patients initial blood pressure and the impact of the new drug on the patients blood pressure

d. none of the previous are correct.

Multiple-choice questions

(8) Data for two variables x and y were collected and a regression line $y = -2.3 + 1.7 \cdot x$ was obtained. The difference between the predicted y-value (from the regression line) and the actual y-value is called residual (error of approximation). What is the residual for point $(5, 6)$?

a. 2.9
b. 0.2
c. 6.2
d. 7.9

Multiple-choice questions

(9) Given a data set with points (x, y), Anna computed the empirical stanadard deviations $s_x = 2.5$ and $s_y = 1.5$, and the empirical correlation coefficient $r = 0.63$ between x and y. What is the slope b_1 of the regression line $y = b_0 + b_1 \cdot x$ she obtained?

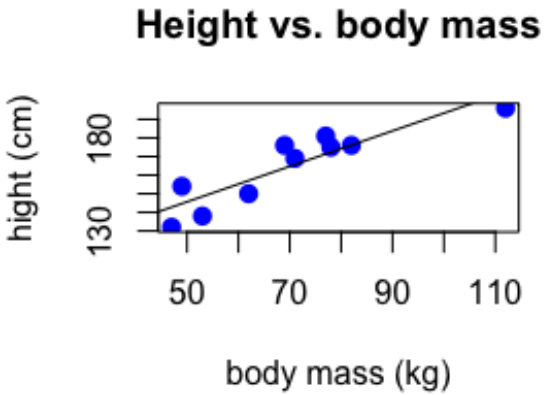
a. 1.05
b. 0.378
c. 3.75
d. 1.67

Multiple-choice questions

(10) Anna interviewed ten members of her family asking them their heights (in cm) and body masses (in kilograms). She then computed the empirical correlation between the heights and body masses $r = 0.9$ and used the following R code

```
plot(bodymass, height, pch = 16, cex = 1.3, col = "blue",main = "Height vs.  body mass", xlab = "body mass (kg)",
ylab = "hight (cm)")
abline(lm(height~bodymass))
lm(height~bodymass)
```

to get the output




```
Call:
lm(formula = height ~ bodymass)

Coefficients:
(Intercept)     bodymass
    98.0054       0.9528
```

Which one of the following statements is correct?

a. the slope of the regression line is 98.0054
b. there is a quadratic relationship between the `height` and `body mass`
c. the regression line is $\texttt{hight} = 98.0054 + 0.9528 \cdot \texttt{body mass}$
d. a linear model is not suitable in this example.

Thank you for your attention!