Linear Regression

18.11.2015

General information

- Lecture website
  `stat.ethz.ch/~muellepa`

- Script, slides and other important information are on the website.

Introduction - Why Statistics?

- There is a **fast growing amount of data** these days, in nearly all research (and applied) areas.
- We want to **extract useful information** from data or **check our hypotheses**.
- E.g., among a large set of variables (temperature, pressure, . . .): which have an effect on the yield of a process and how do the relationships look like?
- We need to be able to **quantify uncertainty**, because "the data could have been different".

- Instead of simply determining a plain numerical estimate for a model parameter, we typically have the following goals:
  - Determine **other plausible values** of the parameter.
  - **Test** whether a specific parameter value is **compatible** with the data.

- Moreover, we want to be able to **understand** and **challenge** the statistical methodology that is applied in current research papers.

Course Outline

Outline of the content

- Linear Regression
- Nonlinear Regression
- Design of Experiments
- Multivariate Statistics

Comments

- Due to time-constraints we will not be able to do "all the details" but you should get the **main idea** of the different topics.
- The lecture notes contain more material than we will be able to discuss in class!
- The relevant parts are those that we discuss in class.

Goals of Today's Lecture

- Get (again) familiar with the statistical concepts:
  - tests
  - confidence intervals
  - p-values

- Understand the difference between a standard numerical analysis of the least squares problem and the statistical approach.

- Be able to interpret a simple or a multiple regression model (e.g., meaning of parameters). Understand the most important model outputs (tests, coefficient of determination, . . .).

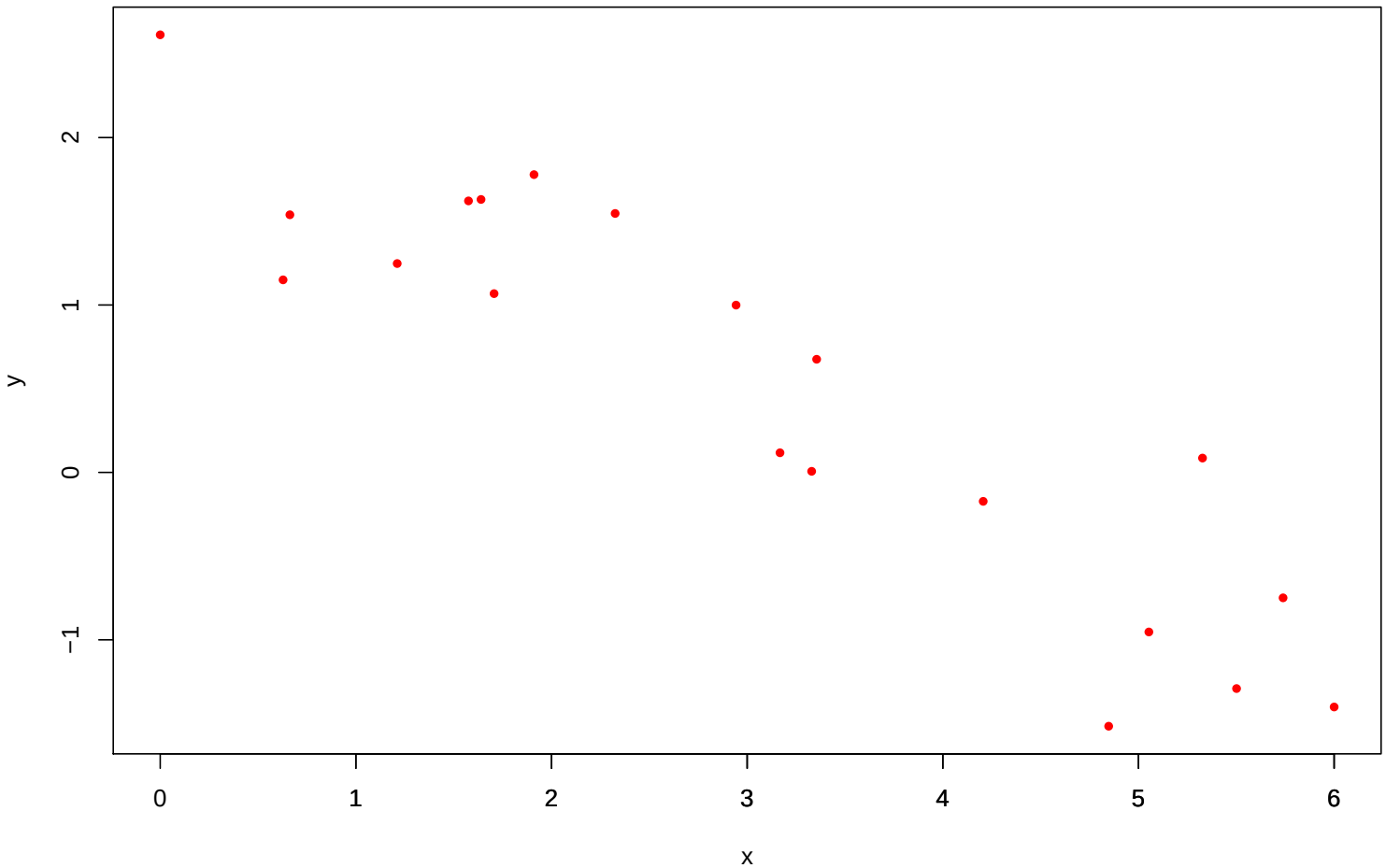
Simple Linear Regression

Introduction

Linear regression is a "nice" statistical modeling approach in the sense that:

- It is a good example to **illustrate statistical concepts** and to learn about the **3 basic questions of statistical inference**:
  - Estimation
  - Tests
  - Confidence intervals
- It is simple, powerful and used very often.
- It is the basis of many other approaches.

Possible (artificial) data set



Goal

Model the relationship between a **response variable** Y and **one predictor variable** x.

- E.g. height of tree (Y) vs. pH-value of soil (x).
- Simplest relation one can think of is

$$Y = \beta_0 + \beta_1 x + \text{Error}.$$

This is called the **simple linear regression model**. It consists of

- an intercept β_0,
- a slope β_1,
- and an error term (e.g., measurement error).

The error term accounts for the fact that the model does **not** give an exact fit to the data.

Simple Linear Regression: Parameter Estimation

- We have a data set of n points $(x_i, Y_i), i = 1, \ldots, n$ and want to **estimate** the unknown parameters.

- We can write the model as

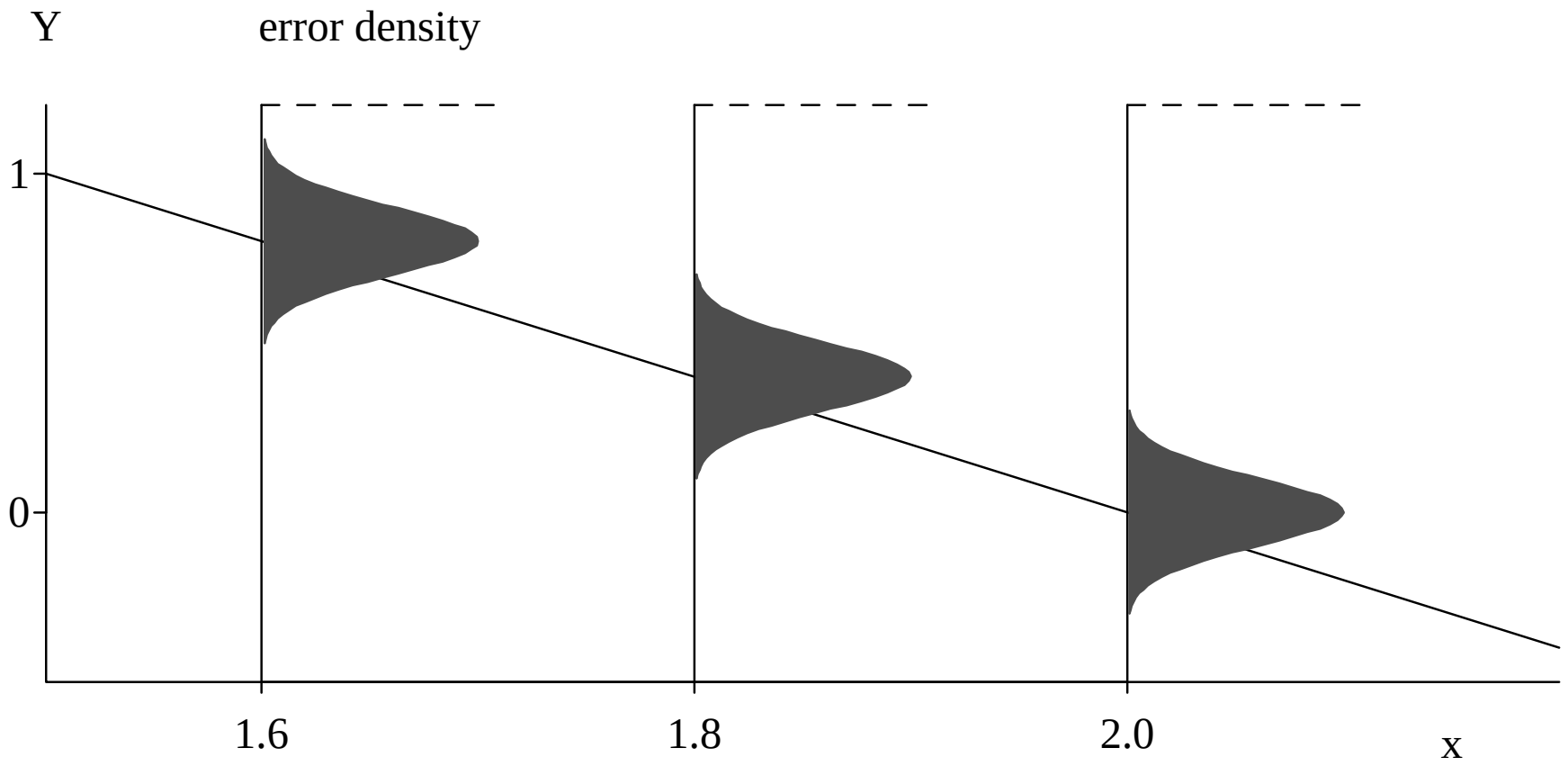
$$Y_i = \beta_0 + \beta_1 x_i + E_i, \quad i = 1, \ldots, n,$$

  where E_i are the errors (that cannot be observed).

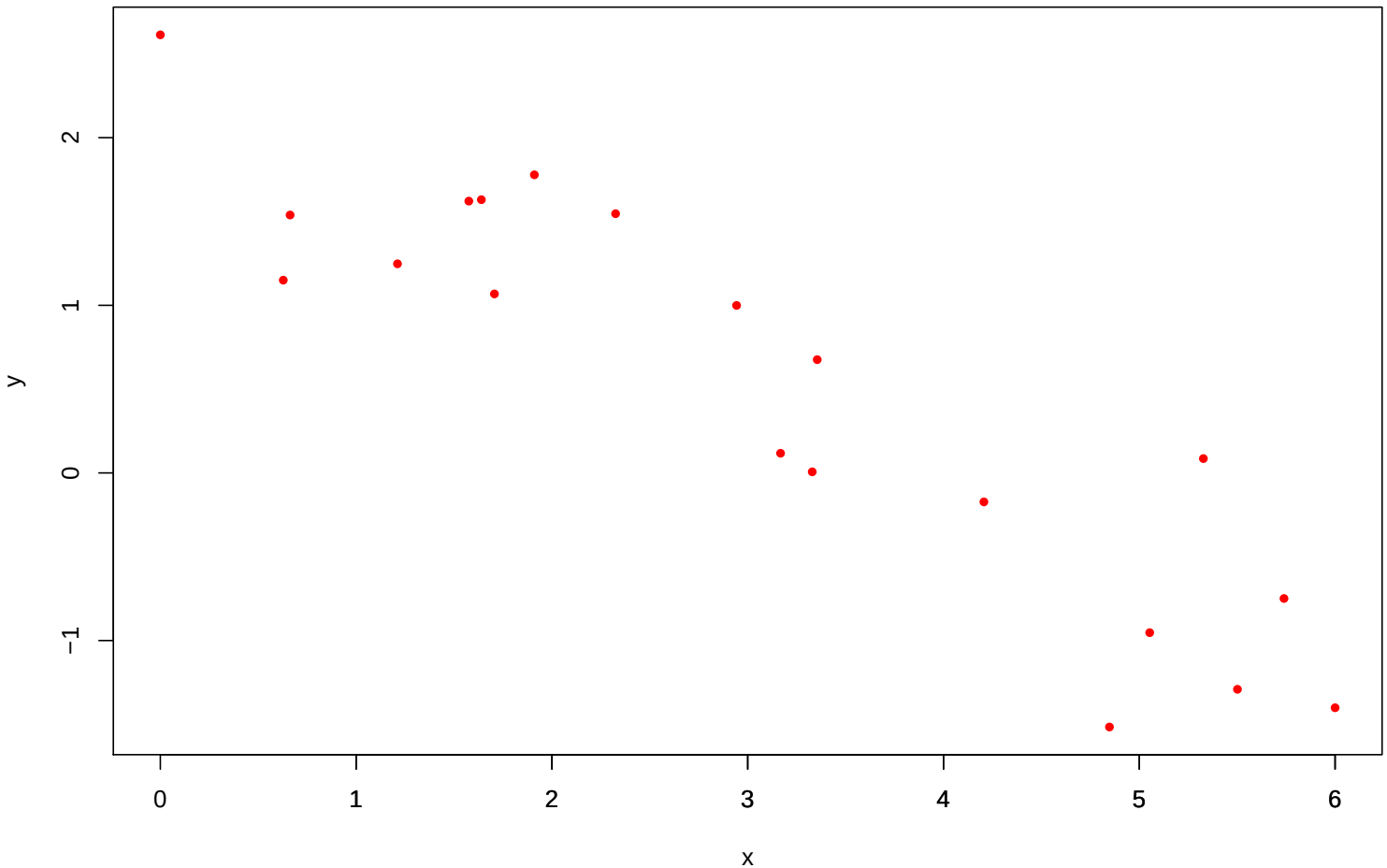
- Usual assumptions are $E_i \sim \mathcal{N}(0, \sigma^2)$, independent.

- Hence, in total we have the following **unknown parameters**
  - intercept β_0
  - slope β_1
  - error variance σ^2 (nuisance parameter).

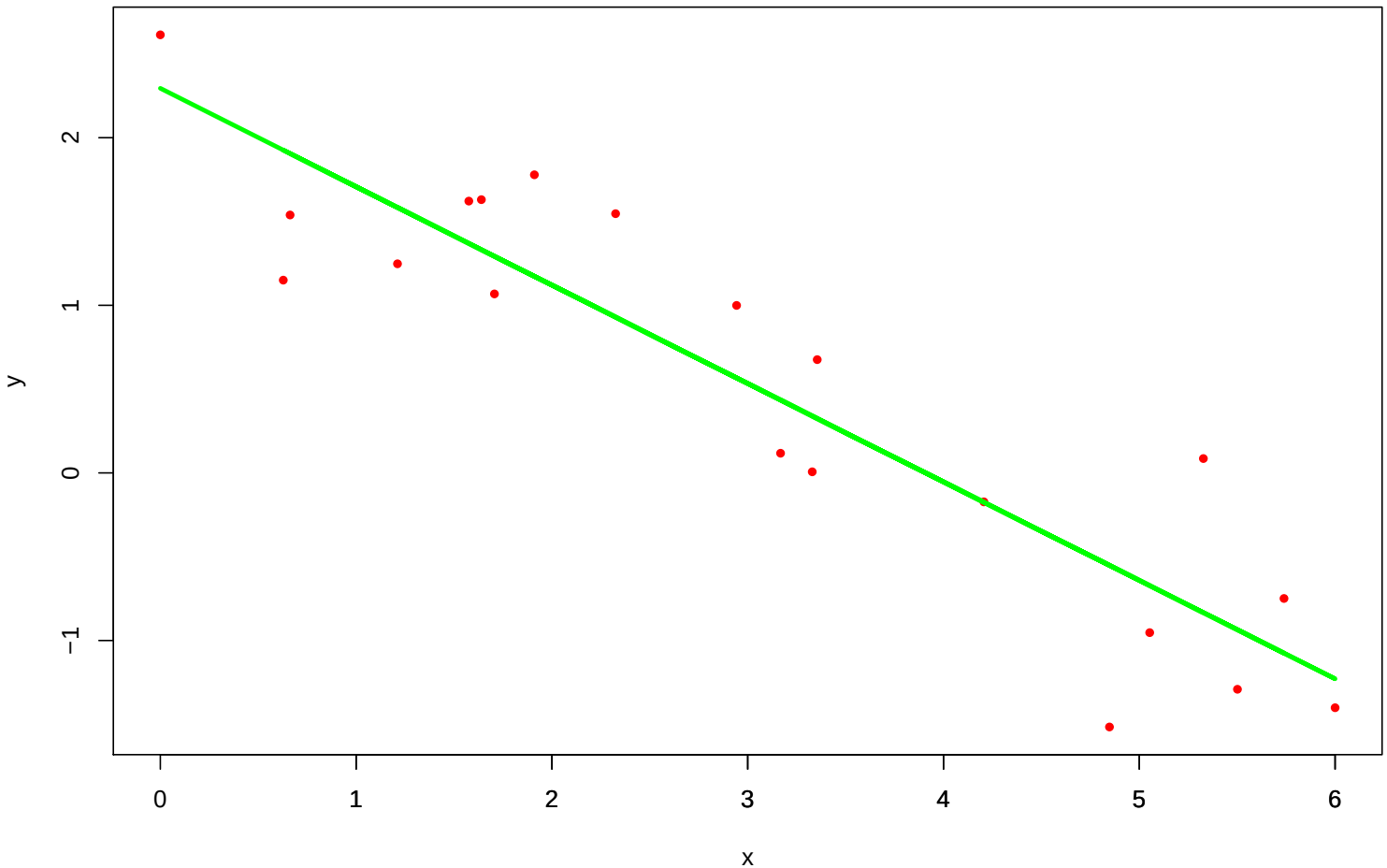
Visualization of data generating process



Possible (artificial) data set



Regression line



- The (unknown) parameters β_0 and β_1 are **estimated** using the principle of **least squares**.
- The idea is to minimize the sum of the squared distances of the observed data-points from the regression line

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 ,$$

  the so called **sum of squares**.
- This leads to parameter estimates

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ \widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{x}. \end{aligned}$$

- This is what you have learned in numerical analysis.
- Moreover

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} R_i^2,$$

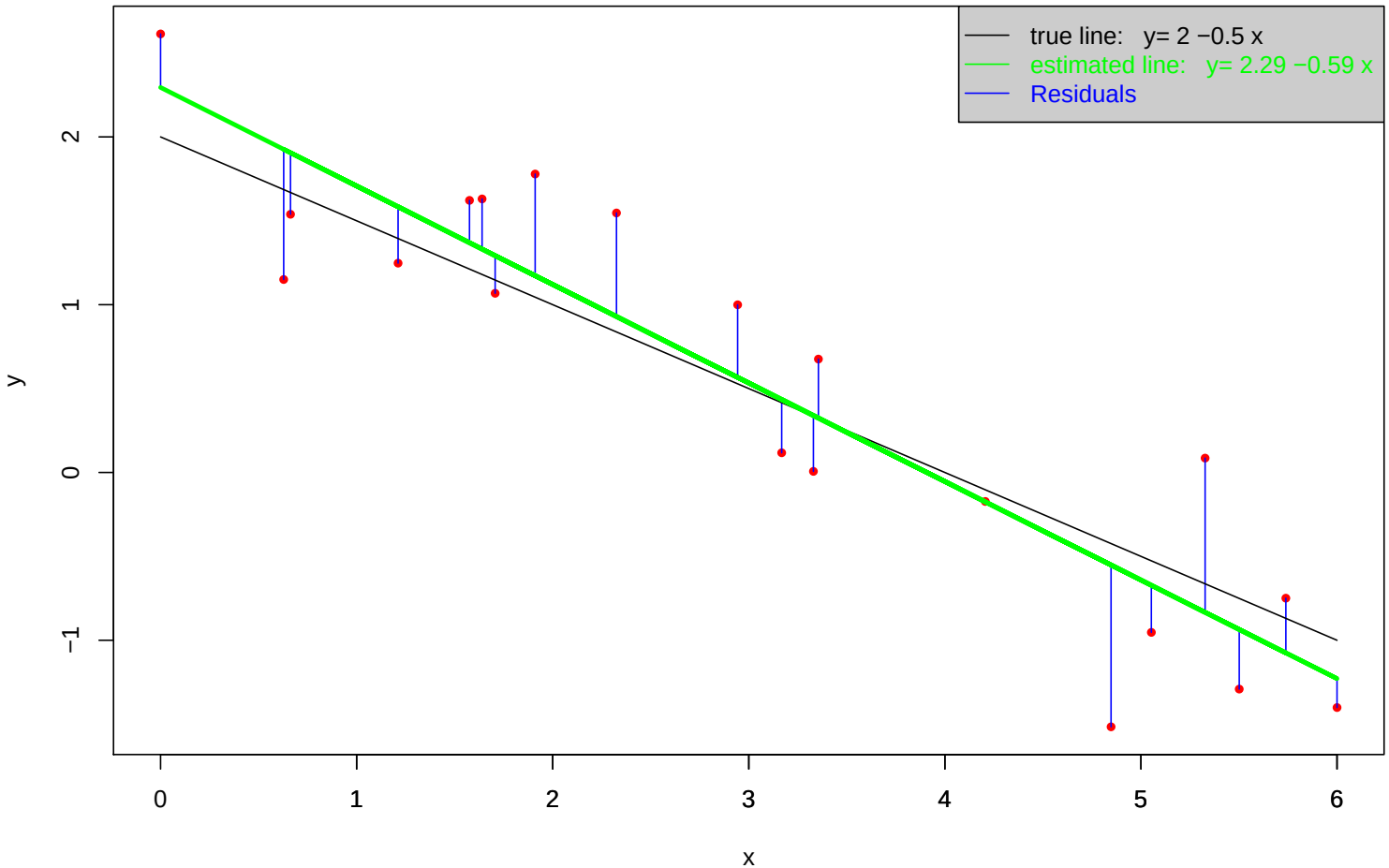
  where

$$R_i = Y_i - \widehat{y}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$$

  are the (observable) **residuals**.

- However, we have made some assumptions about the stochastic behavior of the error term.
- Can we get some extra information based on these assumptions?

Visualization of residuals



- The parameter estimates $\widehat{\beta}_0$, $\widehat{\beta}_1$ are **random variables**!
- Why? Because they depend on the Y_i's that have a random error component.
- Or in other words: **"The data could have been different"**.
- For other realizations of the error term we get slightly different parameter estimates ($\rightsquigarrow$ see animation!).

- The **stochastic model** allows us to quantify uncertainties. It can be shown that

$$\begin{aligned} \widehat{\beta}_1 &\sim \mathcal{N}\left(\beta_1, \sigma^2/\mathrm{SS}_X\right) \\ \widehat{\beta}_0 &\sim \mathcal{N}\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SS}_X}\right)\right), \end{aligned}$$

  where $\mathrm{SS}_X = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

- See animation for illustration of empirical distribution.

- This information can now be used to perform **tests** and to derive **confidence intervals**.

Statistical Tests: General Concepts

- First we recall the basics about statistical testing (restricting ourselves to two-sided tests).
- We have to specify a **null-hypothesis** H_0 and an **alternative** H_A about a model parameter.
- H_0 is typically of the form "no effect", "no difference", "status quo" etc.
- It is the position of a critic who doesn't believe you.
- H_A is the complement of H_0 (what you want to show).
- We want to **reject** H_0 in favor of H_A.
- In order to judge about H_0 and H_A we need some quantity that is based on our data. We call it a **test statistic** and denote it by T.

- As T is stochastic there is a chance to do wrong decisions:
  - Reject H_0 even though it is true (**type I error**)
  - Do not reject H_0 even though H_A holds (**type II error**).

- How can we convince a critic? We assume that he is right, i.e. we assume that H_0 really holds.

- Assume that we know the distribution of T under H_0. We are nice and allow the critic to control the type I error-rate.

- This means that we choose a rejection region such that T falls in that region only with probability (e.g.) 5% (**significance level**) if H_0 holds.

- We reject H_0 in favor of H_A if T falls in the **rejection region**.
- If we can reject H_0 we have "statistically proven" H_A.
- If we **cannot** reject H_0 we can basically say **nothing**, because **absence of evidence is not evidence of absence**.
- Of course we try to use a test statistic T that falls in the rejection region with high probability if H_0 does not hold (**power of the test**).

- Assume that we want to test whether $\beta_1 = 0$. Or in words: "The predictor x has no influence on the response Y"
- This means we have the null hypothesis $H_0 : \beta_1 = 0$ vs. the alternative $H_A : \beta_1 \neq 0$.
- Intuitively we should reject H_0 if we observe a large absolute value of $\widehat{\beta}_1$. But what does large mean here? Use distribution under H_0 to quantify!

Distribution of $\widehat{\beta}_1$

For the true (but unknown) β_1 it holds that

$$T = \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}/\sqrt{\mathrm{SS}_X}} \sim t_{n-2}.$$

Hence, under H_0: $\frac{\widehat{\beta}_1}{\widehat{\sigma}/\sqrt{\mathrm{SS}_X}} \sim t_{n-2}$ (null-distribution).

Remarks

- $\widehat{\sigma}/\sqrt{\mathrm{SS}_X}$ is also called the **estimated standard error** of $\widehat{\beta}_1$.
- We have a t-distribution because we use $\widehat{\sigma}$ instead of σ.
- We reject H_0 if the test statistic T lies in the "extreme regions" of the t_{n-2} distribution.
- If we test at the 5% significance level we reject H_0 if

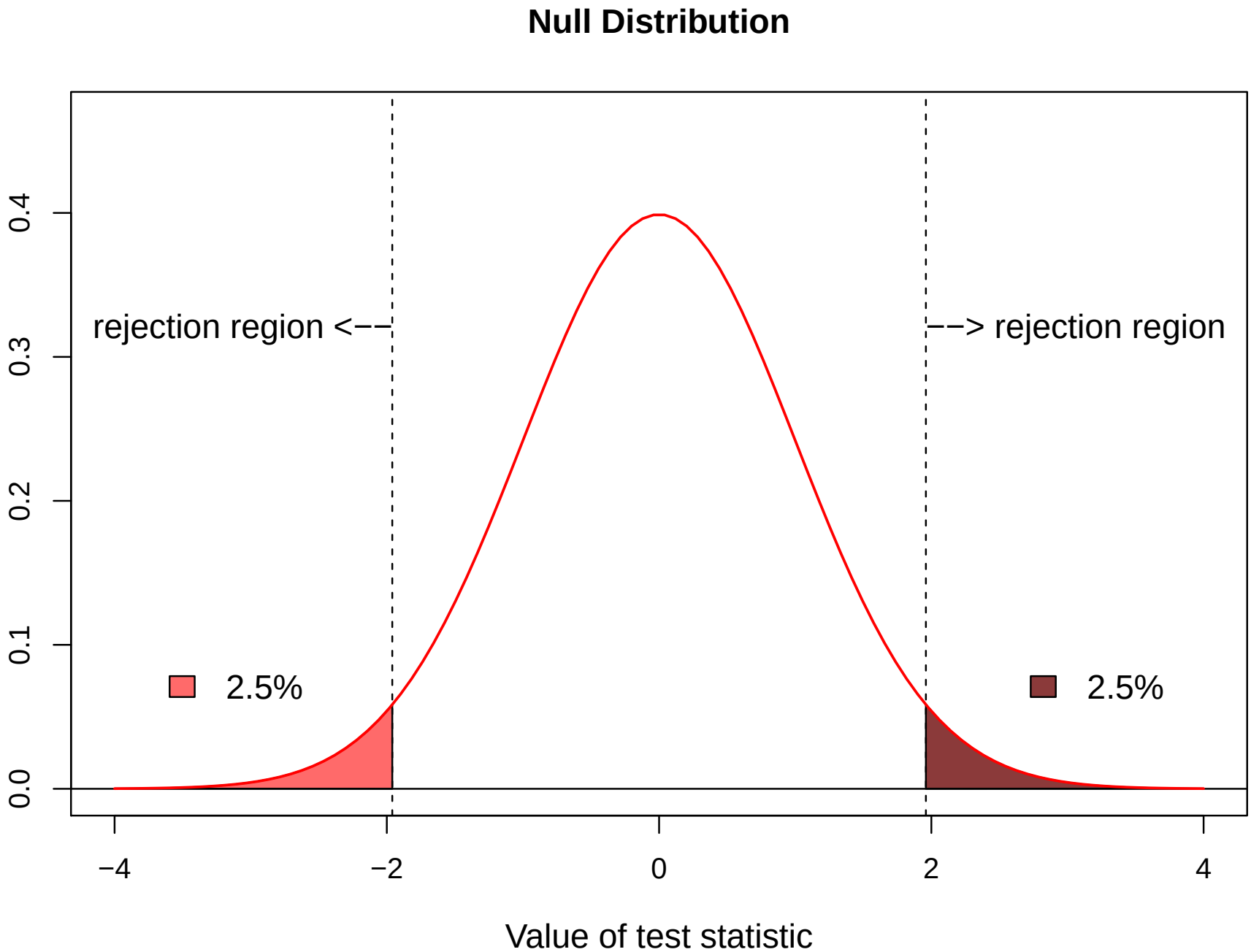
$$|T| \geq q_{0.975}^{t_{n-2}},$$

where $q_{0.975}^{t_{n-2}}$ is the 97.5%-quantile of the t_{n-2} distribution.

- Or in other words:

  "We reject H_0 if T falls either in the region of the 2.5% extreme cases on the left side or the 2.5% extreme cases on the right side of the distribution under H_0" (see picture on next slide).

- Remember: $q_{0.975}^{t_{n-2}} \approx 1.96$ for large n.

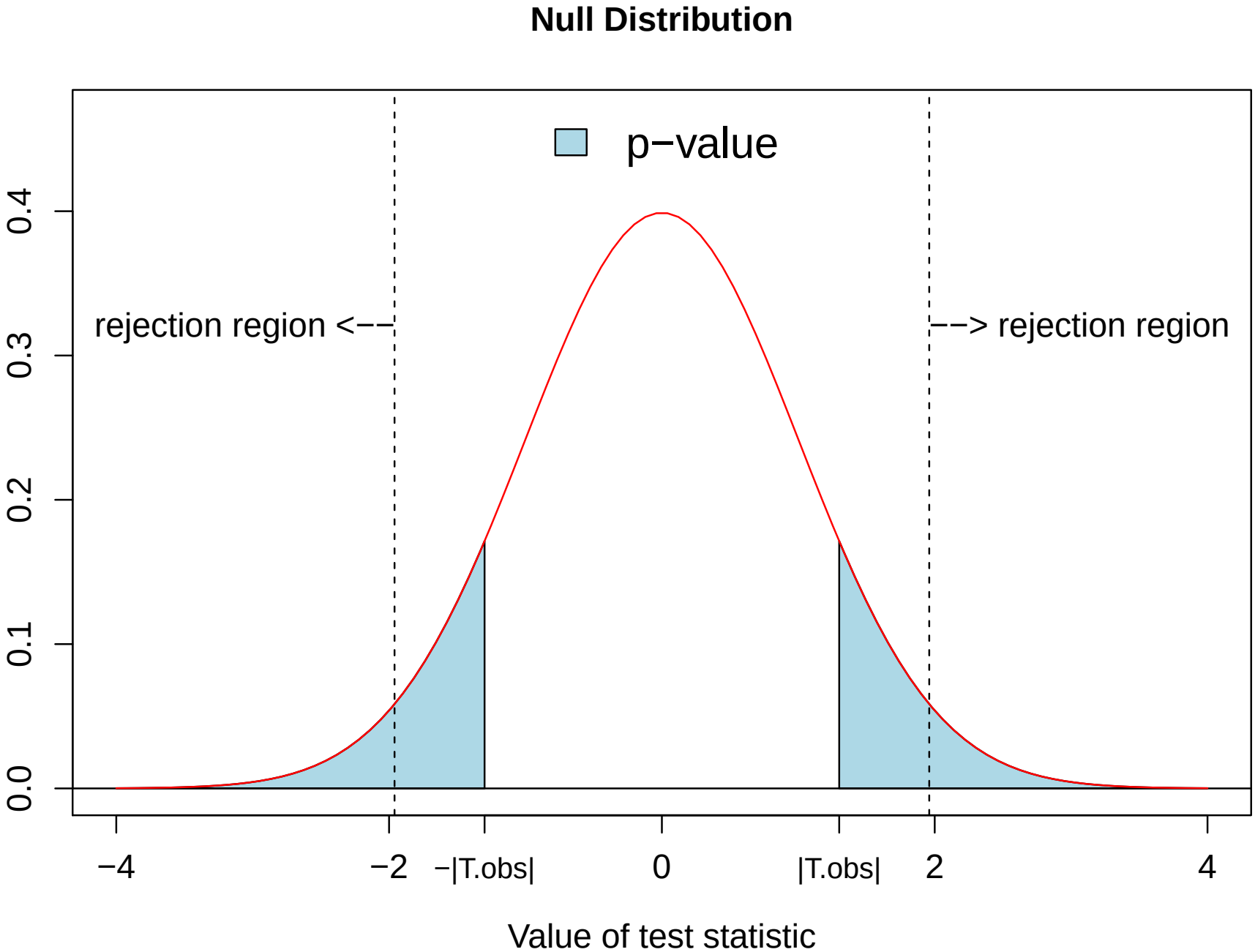

Null Distribution
rejection region <−−
−−> rejection region
2.5%
2.5%
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
Value of test statistic

P-Value

- The p-value is the probability of observing an at least as extreme event if the null-hypothesis is true.

$$p = P_{H_0}(|T| \geq |T_{\text{observed}}|).$$

- Here: "Given that x has no effect on Y, what is the probability of observing a test-statistic T at least as extreme as the observed one?".
- The p-value tells us how extreme our observed T is with respect to the null-distribution.
- If the p-value is less than the significance-level (5%), we reject H_0.
- The p-value contains more information than the test decision alone.


Null Distribution
p−value
rejection region <−−
−−> rejection region
0.0
0.1
0.2
0.3
0.4
−4
−2
−|T.obs|
0
|T.obs|
2
4
Value of test statistic

Confidence Intervals

- A confidence interval (CI) for the parameter β_1 contains all **"plausible values"** for β_1.

- Construction: A 95%-CI consists of all parameter values β_1 that cannot be rejected using the 5%-test above.

$$
\begin{aligned}
CI &= \{\text{all parameter values that are not rejected}\} \\
&= \{\beta_1;\ |T| \leq q_{0.975}^{t_{n-2}}\} \\
&= \widehat{\beta}_1 \pm \widehat{\sigma}/\sqrt{SS_X} \cdot q_{0.975}^{t_{n-2}} \\
&\approx \widehat{\beta}_1 \pm 2 \cdot \widehat{\sigma}/\sqrt{SS_X} \text{ (for large } n\text{)} \\
&= \text{estimate} \pm 2 \cdot \text{estimated standard error (for large } n\text{).}
\end{aligned}
$$

- Alternative interpretation:

  A 95%-confidence interval covers the true parameter value with probability 0.95.

Simple Linear Regression: Computer Output

Example

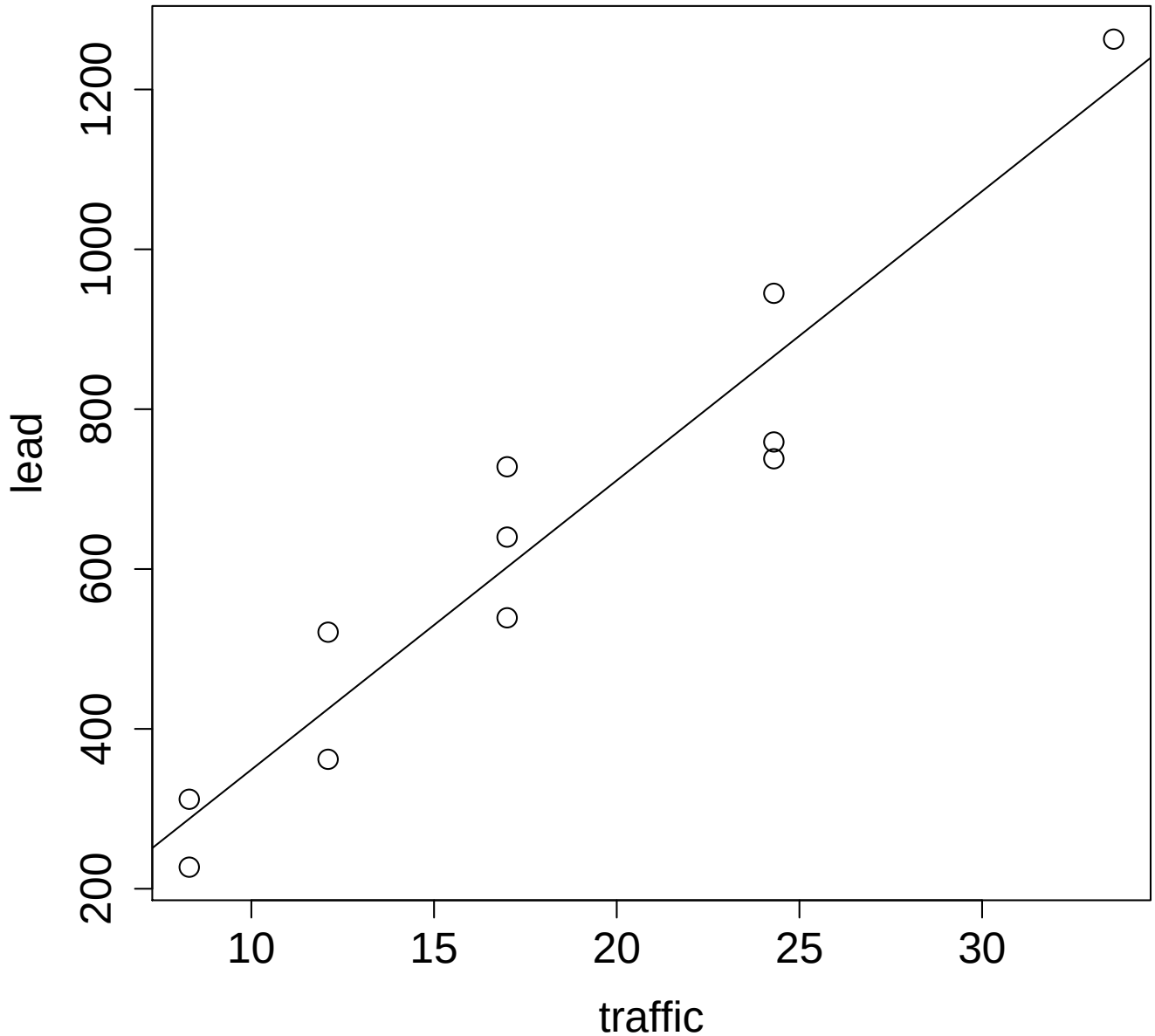
Model the lead content of tree barks ($\mu g/g$) using the traffic amount (in 1000 cars per day). Data was collected at different streets.

Computer Output

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -12.842     72.143  -0.178    0.863
traffic       36.184      3.693   9.798 4.24e-06 ***

Residual standard error: 92.19 on 9 degrees of freedom
Multiple R-squared: 0.9143, Adjusted R-squared: 0.9048
F-statistic: 96.01 on 1 and 9 DF,  p-value: 4.239e-06
```

Data and fitted model



Multiple Linear Regression

- Now we have **more than one predictor**.
- **Model**

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_m x_i^{(m)} + E_i, \quad i = 1, \ldots, n$$

$$E_i \sim \mathcal{N}(0, \sigma^2), \text{ independent.}$$

- Unknown parameters: $\beta_0, \beta_1, \ldots, \beta_m, \sigma^2$.
- The model is called **linear** because it is **linear in the parameters**.
- Note: There are **no assumptions regarding the predictors**!
- **Interpretation of Coefficients**

  β_j measures the efffect of $x^{(j)}$ on Y **after** having subtracted all other effects from $x^{(k)}$ on Y, $k \neq j$.

- In matrix form we have

$$Y = X\beta + E,$$

where

$$X = \begin{pmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(m)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(m)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(m)} \end{pmatrix}, \quad Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix},$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \quad E = \begin{pmatrix} E_1 \\ E_2 \\ \vdots \\ E_n \end{pmatrix}.$$

- The matrix X is called the **design matrix**. It consists of n rows (the different observations) and $p = m + 1$ columns (the different predictors).

- Again, the model is fitted using **least squares**, leading to

$$\widehat{\beta} = (X^T X)^{-1} X^T Y,$$

  i.e., we have a **closed form solution**.

- Moreover

$$\widehat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} (Y_i - \widehat{y}_i)^2 .$$

Multiple Linear Regression: Inference

- Again, it can be shown that

$$\widehat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2 \left((X^T X)^{-1}\right)_{jj}\right).$$

- Again, the estimator $\widehat{\beta}_j$ fluctuates around the true value β_j.
- The standard error is given by

$$\sigma\sqrt{\left((X^T X)^{-1}\right)_{jj}}$$

- This leads to the test statistic

$$T_j = \frac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}\sqrt{\left((X^T X)^{-1}\right)_{jj}}} \sim t_{n-p}.$$

- This is very similar to simple linear regression, with the exception that we now have a t_{n-p} instead of a t_{n-2} distribution.

- Tests and CI for individual parameters are constructed as in the simple linear regression case.

- Here we can also do **simultaneous tests.**
  $H_0 : \beta_1 = \beta_2 = \ldots = \beta_m = 0$ (no effect from any of the predictors).
  H_A : at least on $\beta_j \neq 0$.

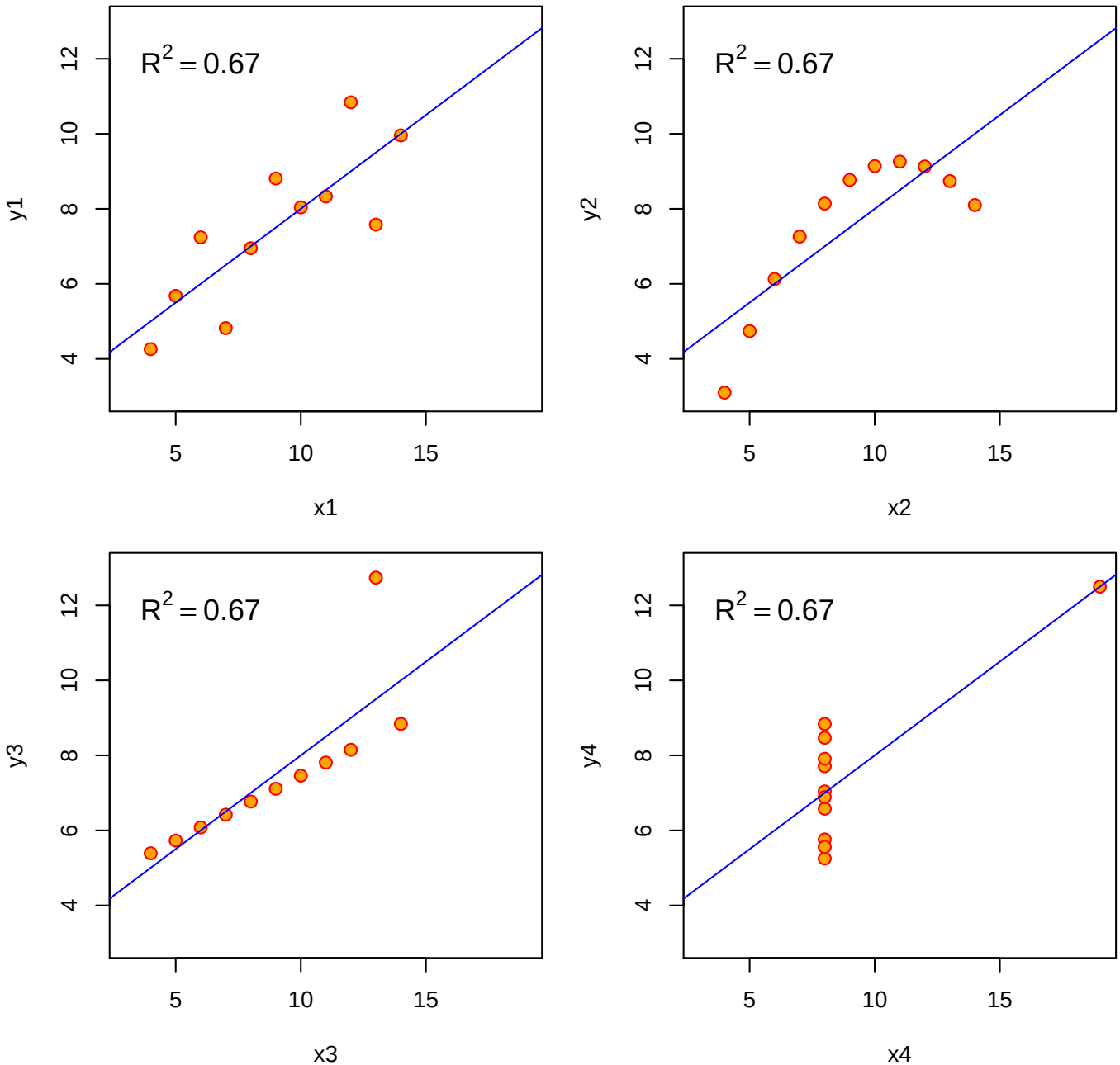
- This can be tested using an F-test (see computer output).

Coefficient of Determination R^2

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

- R^2 is the proportion of the variance that can be explained by the regression model.
- $R^2 = 1$ is equivalent to a perfect fit (all residuals equal 0).
- Simple linear regression: $R^2 = \text{Corr}(x, y)^2$.
- Multiple linear regression: $R^2 = \text{Corr}(\widehat{y}, y)^2$.
- R^2 **does not tell you how well your model fits the data** (see next slide). E.g, it does not tell you whether the relationships are really linear or not.

Anscombe's 4 Regression data sets



Multiple Linear Regression: Computer Output

Example

Model the "monthly steam demand" of a factory using the predictors "operating days" and "average outside temperature".

Computer Output

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     9.126885   1.102801   8.276 3.35e-08 ***
Operating.Days  0.202815   0.045768   4.431  0.00021 ***
Temperature    -0.072393   0.007999  -9.050 7.19e-09 ***

Residual standard error: 0.6616 on 22 degrees of freedom
Multiple R-squared: 0.8491, Adjusted R-squared: 0.8354
F-statistic:  61.9 on 2 and 22 DF,  p-value: 9.226e-10
```

Residual Analysis

- We made assumptions about the error term in the model. We assumed that the errors
  - have expectation 0 (i.e. the regression model is correct),
  - have constant variance,
  - are normally distributed,
  - are independent.

- Tests and confidence intervals are based on these assumptions. They can be (substantially) wrong if they are not fulfilled!
- Results can be "worthless" if assumptions are not met!
- **Residual Analysis** is the visual inspection of the model fit to verify assumptions.
- Among the most popular tools are the **Tukey-Anscombe plot** and **QQ-plots** (and many more...).

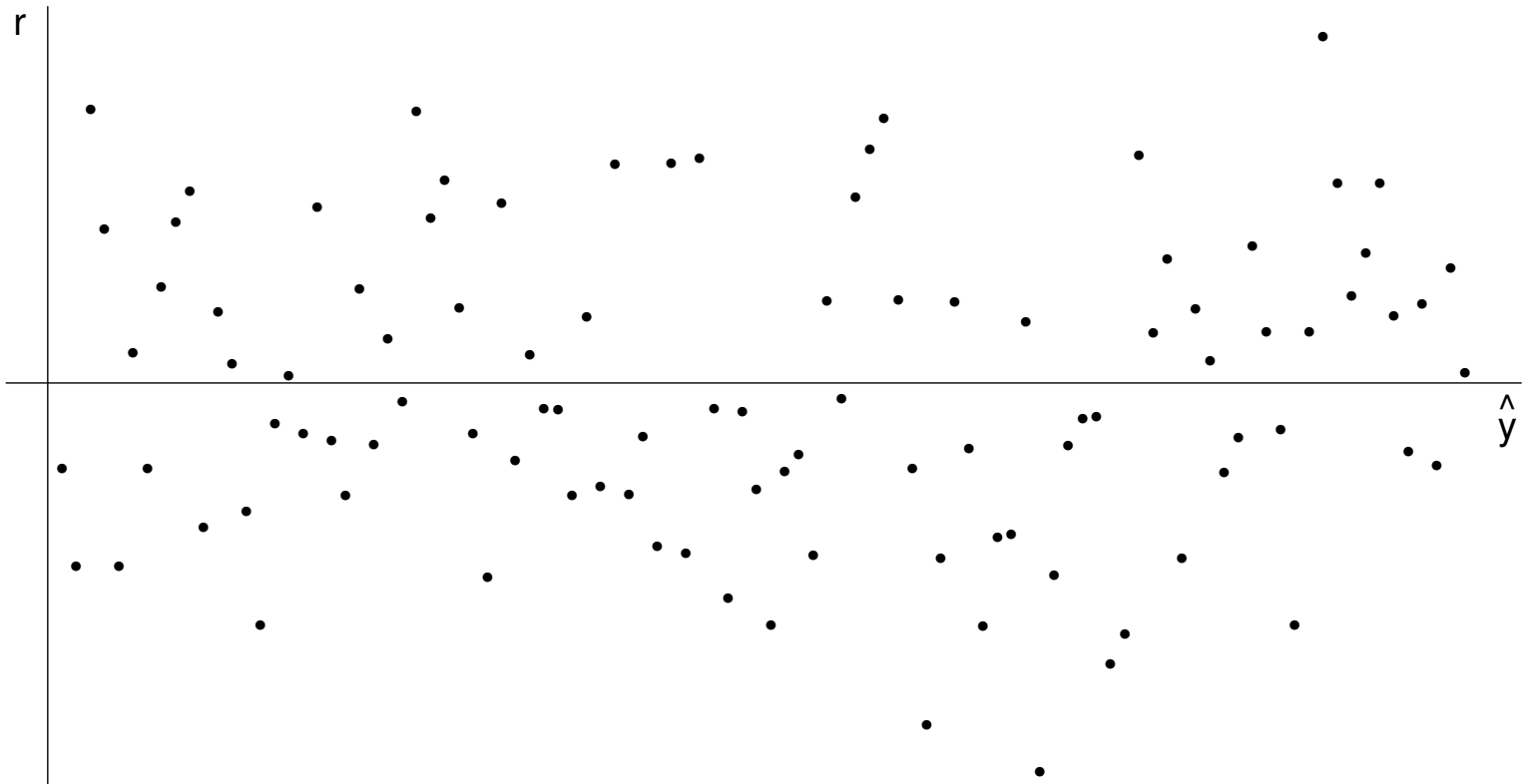
Tukey-Anscombe Plot

Plot (standardized) residuals R_i against fitted values $\widehat{y}_i$.

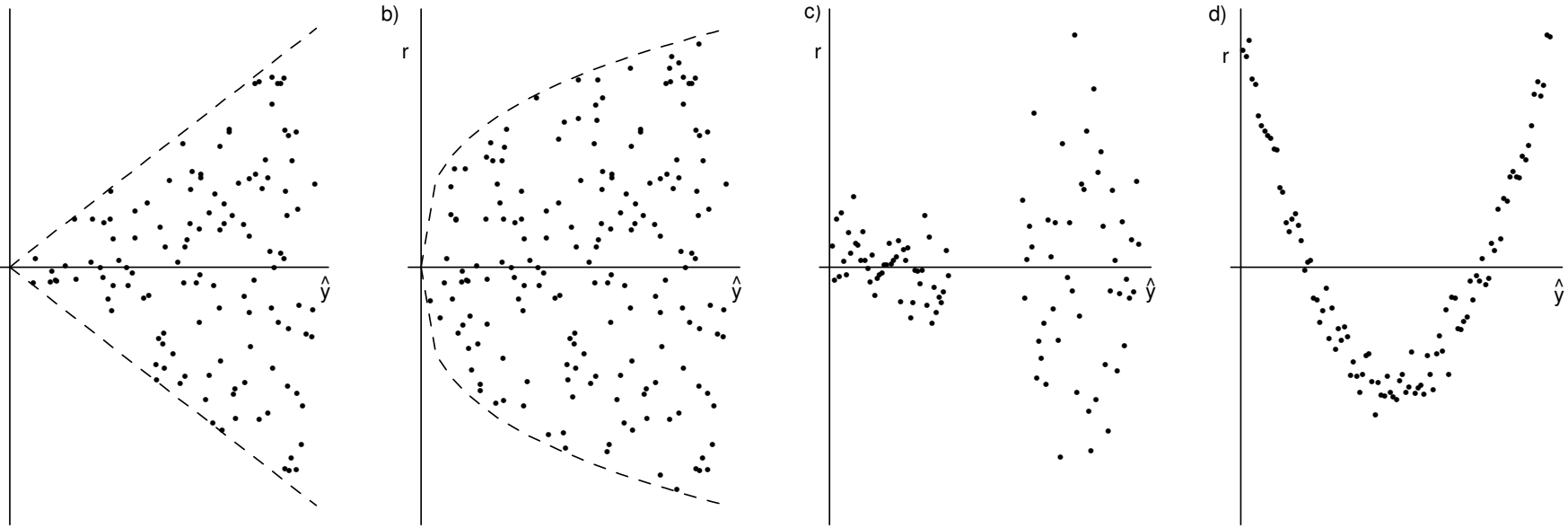
Checks

- $E[E_i] = 0$?
- $\text{Var}(E_i) = \sigma^2$ constant?

TA-plot should show random scatter around the zero line:



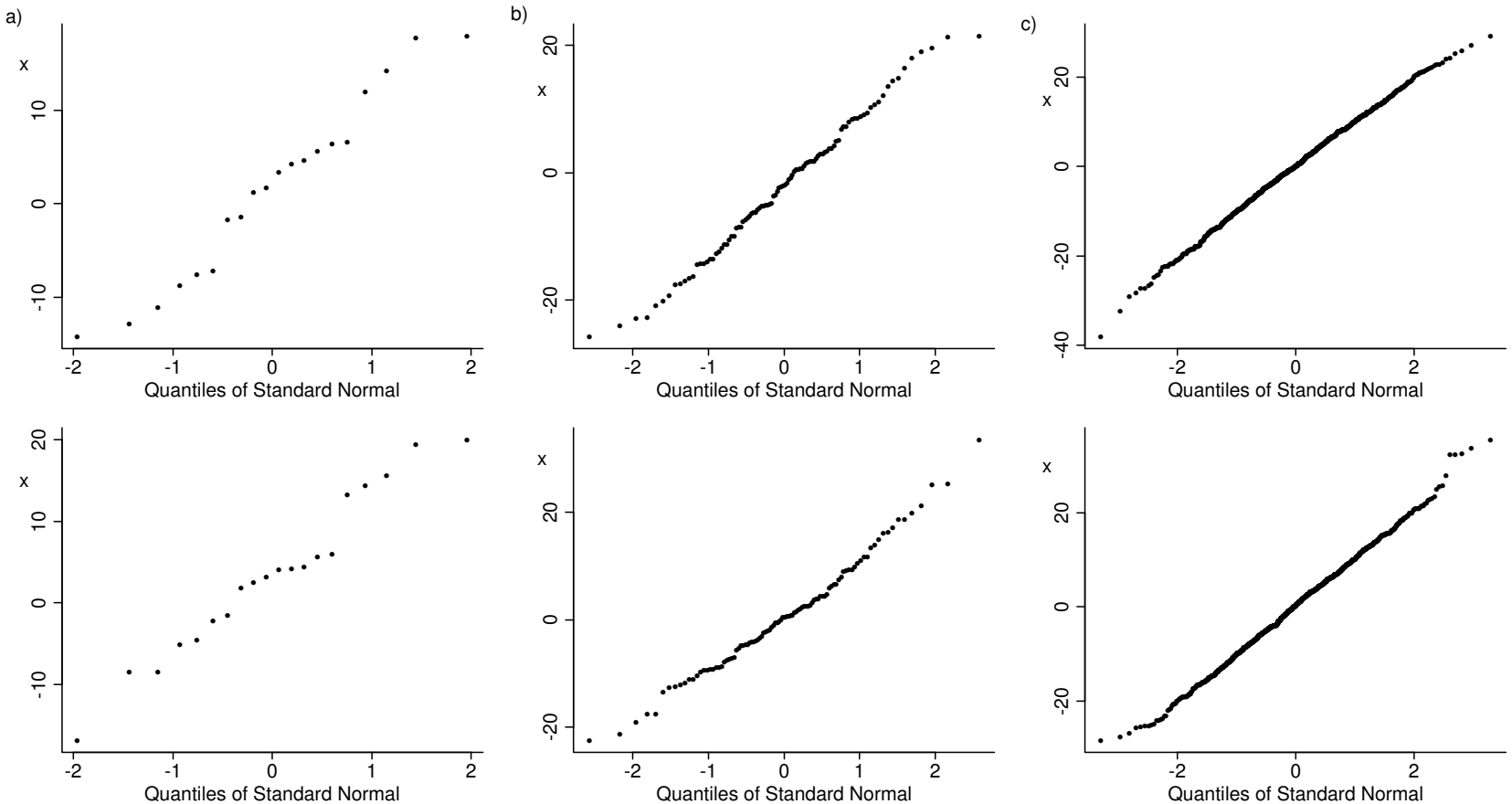
TA-Plots: Regression Assumptions Violated



QQ-Plot

Plot empirical quantiles of residuals against quantiles of standard normal distribution $\rightsquigarrow$ Should show a more or less **straight line**.

Good examples for various sample sizes



QQ-Plots: Non-Normal Distributions