Statistics II
Introduction & Correlation & Regression

Çağrı Çöltekin

University of Groningen
Information Science

April 15, 2014

* with ideas/examples/slides from John Nerbonne & Hartmut Fitz

When, where, who?

- Lectures: Tuesday 11:00–13:00, Zernikezaal
- Computer Labs:

Group 2	Wed	11:00–13:001312.0119A	Noortje Hemmen
Group 4	Thu	11:00–13:001312.0119A	Lena Rampula
Group 3	Fri	11:00–13:001312.0119A	Lena Rampula
Group 1	Fri	13:00–15:001312.0119A	Noortje Hemmen

- Office Hours: Wednesday 09:00–11:00, or by appointment (email c.coltekin@rug.nl).
- Course web page: `http://www.let.rug.nl/coltekin/statII/`
- Textbook: Andy Field (2009). *Discovering statistics using SPSS.*. 3rd ed. London: Sage

Evaluation

- Exam (80%)
- Lab exercises (10%): you will get
  - 2 if complete and in time
  - 1 if incomplete or late (less than one week)
  - 0 otherwise
- Quizzes (6%): quiz scores count only if you get 60% or higher, otherwise you get a 0.
- Attendance (5%): if you are present at five or more lectures (0% otherwise).

Easter eggs (1% bonus): for the first person realizing intentional statistics-related mistakes on the slides.

The plan

1. Simple regression
2. Multiple regression
3. ANOVA (+ general linear models)
4. Factorial ANOVA
5. Repeated measures ANOVA (+ mixed-effect models)
6. Logistic regression
7. Summary & (possibly) some advanced topics

What you should already know

- Descriptive statistics
- Sampling: how to obtain data
- Basics of probability
- Basics of hypothesis testing

Some terms/concepts you should know

If you are not familiar with the following, it is time to go back to your Statistics I course, and get a good understanding of them

- mean
- median
- mode
- variance
- standard deviation
- standard error
- normal (or Gaussian) distribution
- z-score
- t distribution
- t-score
- variable types: numeric, categorical, …
- histogram
- box-and-whisker plot
- confidence intervals
- Q-Q (or P-P) plot for normality
- null hypothesis (H_0) and alternative hypothesis (H_a)
- parametric/non-parametric tests

What is statistics about?

- Descriptive statistics is about making sense of data.
  - statistics like mean and median, graphics like histograms, scatter plots or box-and-whisker plots help us to understand the data at hand better.
- Inferential statistics is about making sense *out* of data.
  - We do not stop with understanding the sample at hand, we want generalizations about the population that the sample came from.
- Statistics is a collection of tools for converting data into information.

But, I have no interest in becoming a researcher

…why should I care?
Maybe not, but you will need to make decisions, based on statistics:

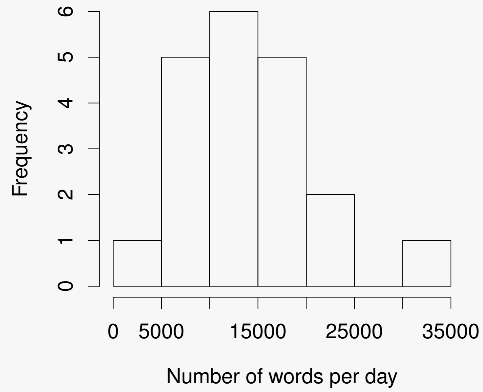
- Whether to decide in favor of a proposed change in education.
- Whether spending on advertisements in a new media/channel would be beneficial for your company.
- Whether (or to what extend) you should allow your child to watch TV, play video games, eat junk food.
- Whether to buy the expensive anti-aging cream or anti-hair-loss shampoo with claims of 'clinically proven' effect.

All of these will be presented to you in some form of statistics.

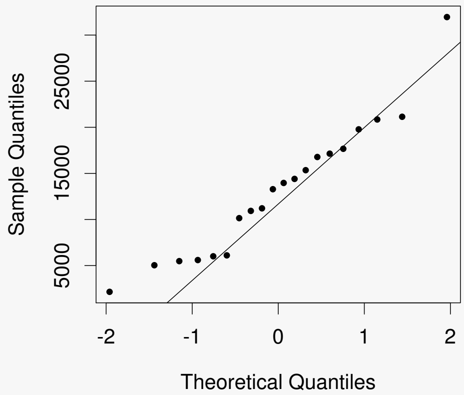
An example: how much do we speak on average?

- We recruit 20 university students, record everything they say during a day, and count the number of words.
- $x_{1..n} = 17667,\ 15347,\ 14401,\ 5037,\ 20845\ldots$
- The mean is, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 13248.1$.
- Estimated variance is $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 52,518,951$
- Estimated standard deviation is $s = \sqrt{52,518,951} = 7246.996$.
- Based on this data what is our best estimate of number of words a person speaks a day?
- Is this estimate reliable?

Visualizing data: histograms



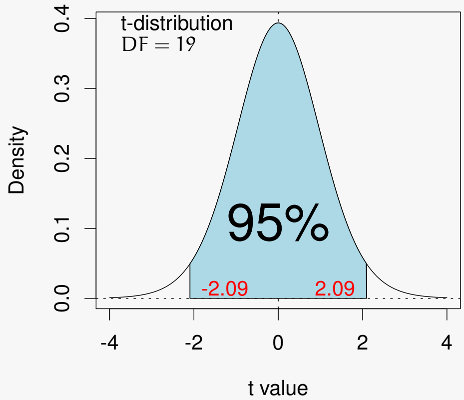
Is the distribution normal?



Confidence intervals: accounting for uncertainty

- A confidence interval is an interval specified around known sample mean. The interval is typically set to 95% or 99% (by convention).
- The question is: *if we did this experiment many times, in how many of them the true mean would fall within the interval?*
- The estimated standard deviation of the sample means (called *standard error of the mean*) is $SE_{\bar{x}} = \frac{s_x}{\sqrt{n}}$.
- We use *Student's t-distribution* to which the interval covers the true mean with a given probability (e.g., 95%).

How certain are we about these measurements?



Confidence intervals: how to calculate it

$$t = \frac{\bar{x} - \mu}{SE_{\bar{x}}}$$

$$\begin{aligned}
-2 < &\ \frac{13248.1 - \mu}{\frac{7246.996}{\sqrt{20}}} < 2 \\
-2 \times 1620.478 < &\ 13248.1 - \mu < 2 \times 1620.478 \\
-2 \times 1620.478 - 13248.1 < &\ -\mu < 2 \times 1620.478 - 13248.1 \\
-16489.06 < &\ -\mu < -10007.14 \\
10007.14 < &\ \mu < 16489.06
\end{aligned}$$

We are 95% confident that the true mean is in the range $[10007.14, 16489.06]$.

Basic hypothesis testing: one sample t-test

Assuming we know that an average person utters 20,000 words per day, do university students talk more or less than an average person?

H_0 : The population mean (of university students) is 20,000 word per day.

H_a : Population mean is different than 20,000 words per day (two-tailed hypothesis).

Since 95% confidence interval $[10007.14, 16489.06]$ does not include 20,000, we reject the null hypothesis, and conclude that we found a statistically significant difference at α-level $= 0.05$.

One sample t-test: looking at it another way

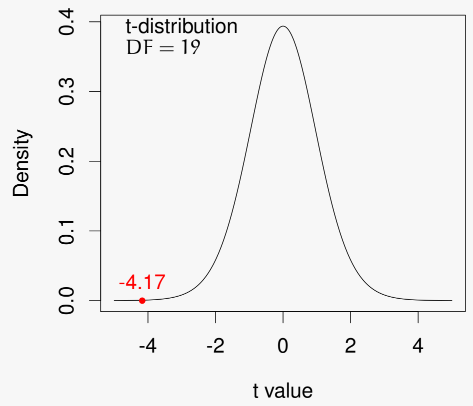
- Calculate the t-score, given the null hypothesis is true ($\mu = 20,000$):

$$\begin{aligned}
t &= \frac{\bar{x} - \mu}{SE_{\bar{x}}} \\
&= \frac{13248.1 - 20000}{\frac{7246.996}{\sqrt{20}}} \\
&= -4.16661
\end{aligned}$$

- Calculate the probability of $t \leq -4.16661$ under the t-distribution with $DF = 19$ (e.g., check via probability tables).

$$p = 0.0003$$

One sample t-test: visualization



Two-samples t-test

Half of our word counts come from women and the other half from men. The question is whether women talk more than men.
This time we let the software do it for us:

```
data: words by gender
t = -0.0367, df = 18, p-value = 0.4856
alternative hypothesis: true difference in means is greater than
    0
95 percent confidence interval:
 -5651.608      Inf
sample estimates:
mean in group F mean in group M
        13309.2        13187.0
```

t-score reported in the listing above should not be negative.

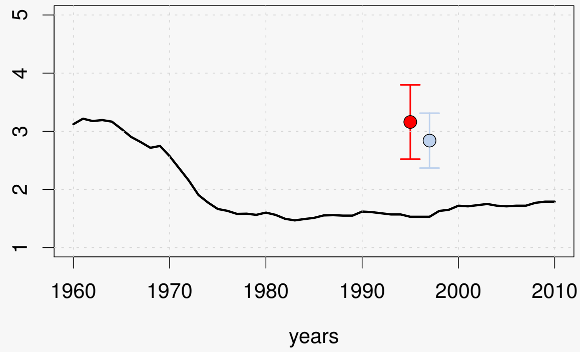
Your turn: calculating confidence intervals

$n = 22$ people responded to the class survey (excluding non-Dutch speakers):

- Mean number of siblings in the data is $\bar{x} = 3.16$.
- The standard deviation is $sd(x) = 1.55$.
- $t(21)$ for $p < 0.025$ is -2.08.

Calculate the 95% confidence interval for the birth rate (around the time you were born).
Use the approximation $\bar{x} \pm 2 \times SE$, and $n = 16$ if you do not have a calculator.
My calculation: $[2.521858, 3.798142]$.

Birth rate in NL according to the World Bank



Did the World Bank make a mistake?

Null-hypothesis significance testing

- Null-hypothesis significance testing (NHST) is probably most widely used scientific tool.
- It is important to get a fair understanding of it.
- If you are confused, you are not alone. Hypothesis testing is confusing.

Typical NHST procedure

- Define a null hypothesis (H_0) that expresses when your hypothesis is wrong.
- Define an alternative hypothesis (H_a, or H_1) as what you expect to find. (well...depending on which NHST procedure you follow.)
- Choose a significance level (α-level) at which to reject the H_0. Typical values are 0.05, 0.01, 0.001.
- Apply the appropriate test, say t-test, which will yield a p-value, of obtaining the sample you have, if H_0 was true.
- If $p < \alpha$, we reject the H_0, otherwise, we fail to reject the H_0.

NHST: problems/suggestions

Beware:

- The p-value is not the probability of null-hypothesis being true.
- Not finding a significant difference does not mean there is none: you can never accept the null hypothesis.
- Statistical significance does not warrant practical importance.

Suggestions:

- Whenever you see a p-value insert 'if null hypothesis was true' in your conclusions.
- Report value of the p (not just $p < .05$).
- Always look for effect sizes, interpret along with (confidence) interval estimates around the effect sizes.

Effect sizes: what are they?

A few examples:

- The estimate of the mean.
- The estimate of the difference between two means. Or, *Cohen's d* ($\frac{\bar{x}_1 - \bar{x}_2}{s}$), if you like standardized measures.
- Ratio or percentage of change (say, in a year, or after treatment).
- Correlation coefficient r (or r^2).
- Slope values in a regression analysis.
- Proportion of variance explained by a model: multiple-r^2 (or adjusted-r^2), η^2 (or ω^2).

It is best to interpret effect sizes with respect to the problem studied.

Statistical models

All statistical analyses can be cast into a model:

$$\text{response} = \text{model} + \text{error}$$

- model is what we are interested in.
- error effects the precision (and certainty) of our estimates.
- we prefer models with smaller error.
- we prefer simpler models.

What are the models?

- Model of the mean (sometime called the null model):

$$y = \mu + e$$

- Model with multiple group means (like in ANOVA):

$$y = \mu + \delta_1 + \delta_2 + e$$

- Model with a single predictor (regression, but also t-test):

$$y = a + bx + e$$

- Model with multiple predictors (regression, ANOVA):

$$y = a + b_1x_1 + b_2x_2 + \ldots + e$$

Correlation and Regression

Two common methods of analyzing relationship between two (numeric) variables are *correlation* and *regression*. For example,

- Education and income.
- Height and weight.
- Age and ability (e.g., language skills, cognitive functions, eye sight, ...)
- Speed and accuracy.

Correlation

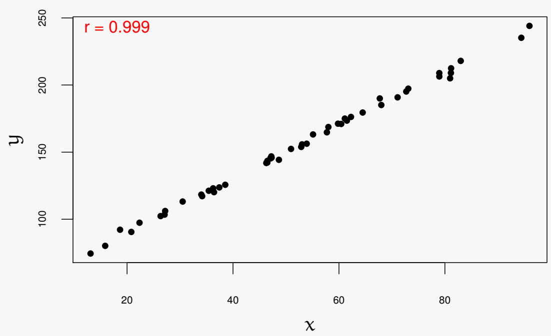
Correlation coefficient is a standardized measure of covariance between two variables. It takes values between −1 and 1

- 1 Perfect positive correlation.
- $(0, 1)$ positive correlation: x increases as y increases.
- 0 No correlation, variables are independent.
- $(-1, 0)$ negative correlation: x decreases as y increases.
- -1 Perfect negative correlation.

Note: correlation is a symmetric measure.

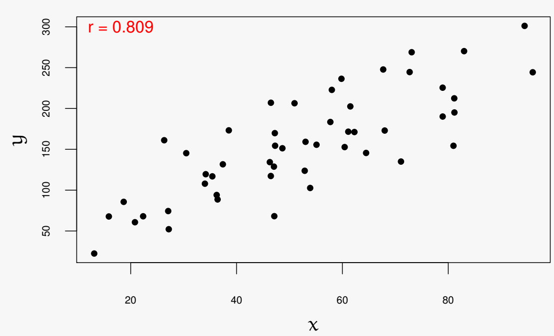
Scatter plots

Scatterplots are a good way to visualize the relationship between two variables:



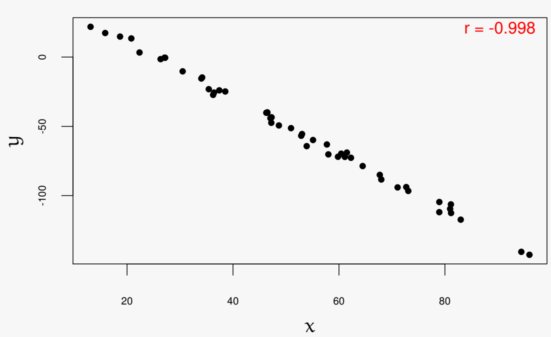
Scatter plots

Scatterplots are a good way to visualize the relationship between two variables:



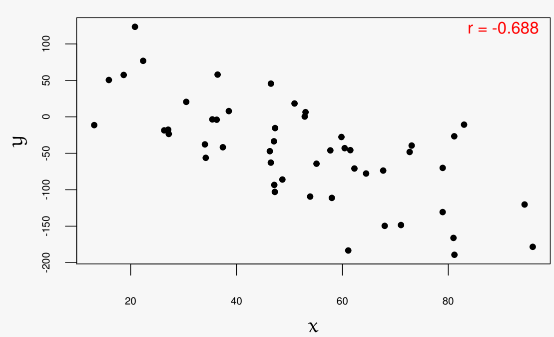
Scatter plots

Scatterplots are a good way to visualize the relationship between two variables:



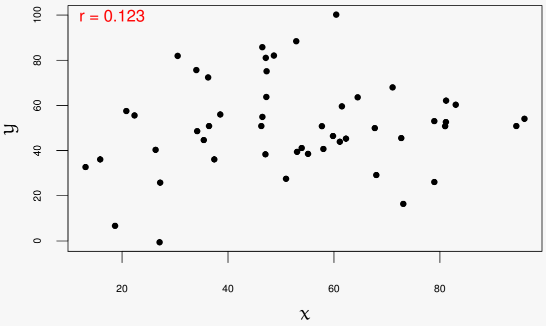
Scatter plots

Scatterplots are a good way to visualize the relationship between two variables:



Scatter plots

Scatterplots are a good way to visualize the relationship between two variables:



Pearson product-moment correlation coefficient

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} z_{x_i} z_{y_i}$$

- Reminder: $z_x = \frac{x - \mu_x}{\sigma_x}$
- If z_{x_i} and z_{y_i} have the same sign, the result is positive.
- If z_{x_i} and z_{y_i} have the opposite signs, the result is negative.
- Pearson's r has the same assumption of linear regression (we'll discuss it soon).
- When assumptions do not hold, use non-parametric alternatives: *Spearman's ρ (rho)* or *Kendall's τ (tau)*.

Inference for correlation

Correlation coefficient shows the association of values within the sample, if we want to know whether the results hold for the population,

- We can calculate a confidence interval (e.g., 95%).
- Do a single-sample t-test with null hypothesis that $r = 0$.

Note: The inference is based on the following statistic which is t-distributed with $DF = n - 2$.

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

Correlation is not causation

- Shoe size correlates highly with reading ability.
- Chocolate consumption in a country correlates with number of Nobel prize winners.
- Weight of a person correlates with the daily amount of calorie intake.
- Number of police station in a neighborhood correlates with the rate of crime.
- Decrease in number of pirates or ratio of people wearing hats is correlated with global warming.

Regression

Regression analysis is about finding the best linear equation that describes the relationship between two variables.

- Regression is closely related to correlation: higher the correlation between two variables, better the fit of regression line.
- Simple regression can be extended for multiple predictors easily (next week).

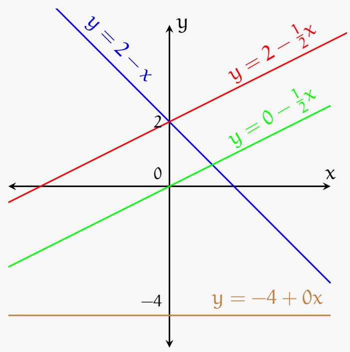
The linear equation

$$y = a + bx$$

a (intercept) is where the line crosses the y axis.

b (slope) is the change in y as x is increased one unit.

What is the correlation between x and y for each line (relation)?



The sign of the slope in the equations for red and green lines are wrong.

The simple linear model

$$y_i = a + bx_i + e_i$$

y is the *outcome* (or response, or dependent) variable. The index i represent each unit observation/measurement (sometimes called a 'case').

x is the *predictor* (or explanatory, or independent) variable.

a is the intercept.

b is the slope of the regression line.

a and b are called *coefficients*.

$a + bx$ is the *deterministic* part of the model. It is the model's prediction of y ($\hat{y}$), given x.

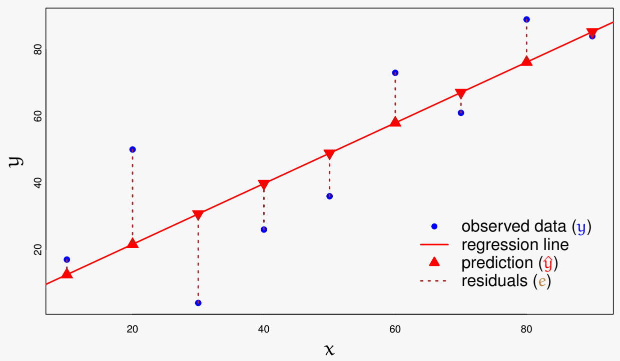
e is the residual, error, or the variation that is not accounted for by the model. Assumed to be (approximately) normally distributed with 0 mean (e_i are assumed to be i.i.d).

Notation differences for the regression equation

$$y_i = a + bx_i + e_i$$

- Sometimes, Greek letters α and β are used for intercept and the slope, respectively.
- Another common notation to use only b or β, but use subscripts, 0 indicating the intercept and 1 indicating the slope.
- It is also common to use ϵ for the error term (residuals).
- Sometimes coefficients wear hats, to emphasize that they are estimates.

Visualization of regression procedure



Least-squares regression

Least-squares regression is the method of determining regression coefficients that minimizes the sum of squared residuals (SS_R).

$$y_i = \underbrace{a + bx_i}_{\hat{y}_i} + e_i$$

- We try to find a and b, that minimizes the prediction error:

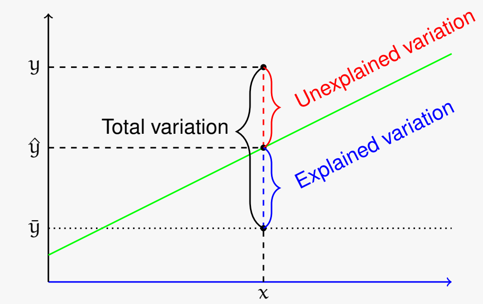
$$\sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - (a + bx_i))^2$$

- This minimization problem can be solved analytically, yielding:

$$b = r\frac{\sigma_y}{\sigma_x}$$
$$a = \bar{y} - b\bar{x}$$

* See appendix for the derivation.

Variation explained by regression



Total variation = Unexplained variation + Explained variation

$y - \bar{y}$ = $y - \hat{y}$ + $\hat{y} - \bar{y}$

Assessing the model fit: r^2

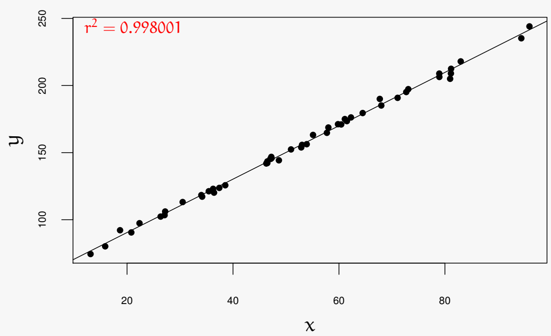
We can express the variation explained by a regression model as:

$$\frac{\text{Explained variation}}{\text{Total variation}} = \frac{\sum_i^n (\hat{y}_i - \bar{y})^2}{\sum_i^n (y_i - \bar{y})^2} = \frac{SS_M}{SS_T}$$

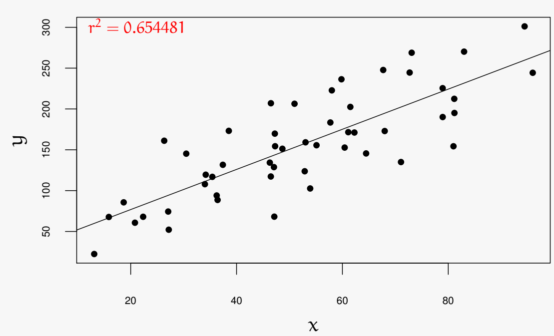
It can be shown that this value is the square of the correlation coefficient, r^2, also called the coefficient of determination.

- $100 \times r^2$ can be interpreted as 'the percentage of variance explained by the model'.
- r^2 shows how well the model fits to the data: closer the data points to the regression line, higher the value of r^2.
- r^2 is also a way of characterizing the effect size.

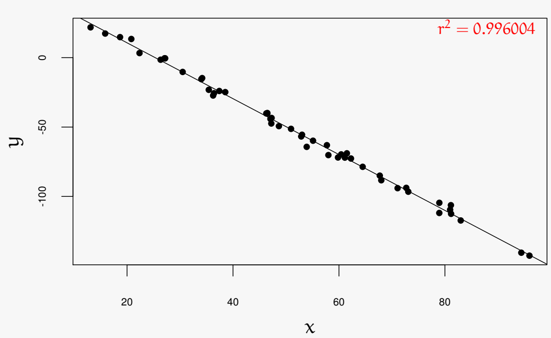
r^2: examples



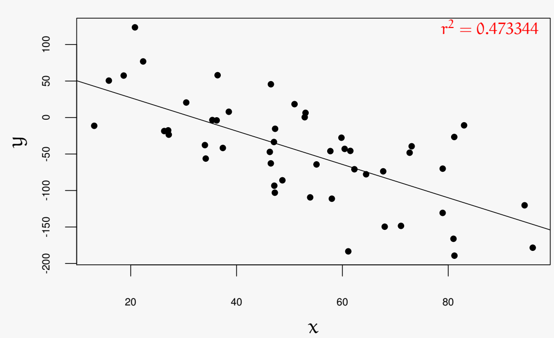
r^2: examples



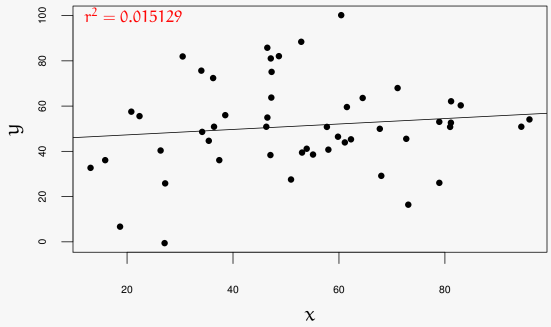
r^2: examples



r^2: examples



r^2: examples



Inference for coefficients

We calculate standard errors for coefficients, SE_b and SE_a (see appendix for the formulas).

- We can construct confidence intervals for a and b as usual using t-distribution with $n-2$ degrees of freedom.
- If corresponding confidence interval does not contain 0, we state that the estimate of the parameter is statistically significant.
- In most cases inference about the intercept is not very informative. It indicates whether intercept is different from 0 or not.
- If the estimate of the slope (b) is statistically significant, we conclude that the effect of predictor on the response variable is not due to chance.

Inference for overall model fit

We can also test whether the overall model fit is significant. To do this, we use the ratio,

$$F = \frac{\text{Explained variance}}{\text{Unexplained variance}} = \frac{MS_M}{MS_R} = \frac{\sum_i^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n-2}\sum_i^n (y_i - \hat{y}_i)^2}$$

- This ratio follows an F-distribution with $DF = (1, n-2)$.
- Note: $MS_M = SS_M/DF_M$ and $MS_R = SS_R/DF_R$.
- If variance explained is larger than the unexplained variance, then the model is doing something useful. So, we test for $F > 1$.
- This test is equivalent to the t-test for the slope for simple regression.

* More on F-distribution later.

Interpreting a regression model fit

a the expected value of the response, when predictor is zero.

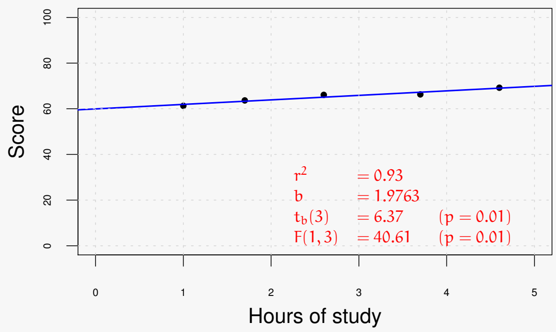
b the expected difference in the response for one-unit difference in the predictor.

r^2 variation explained by the model.

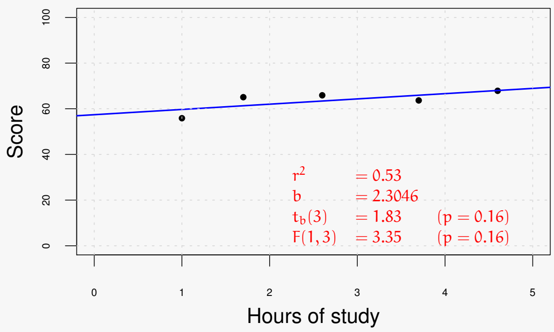
t-test for coefficients: is my coefficient significantly different than 0?

F-test for model fit: is the variation explained by the model larger than the unexplained variation?

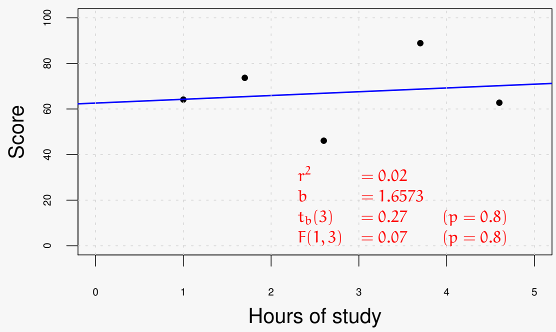
Visualizing regression results



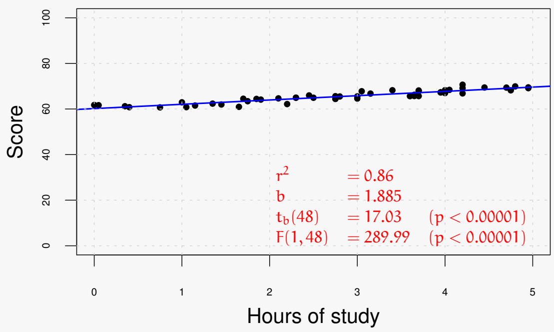
Visualizing regression results



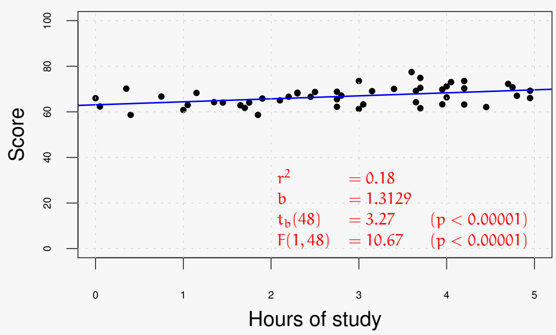
Visualizing regression results



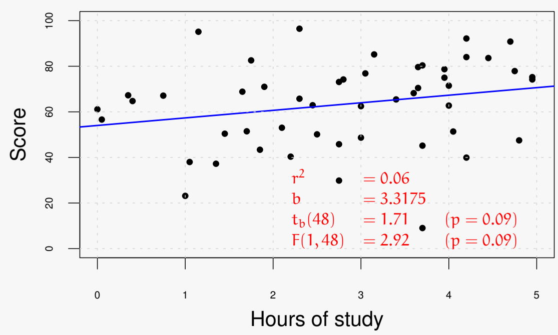
Visualizing regression results



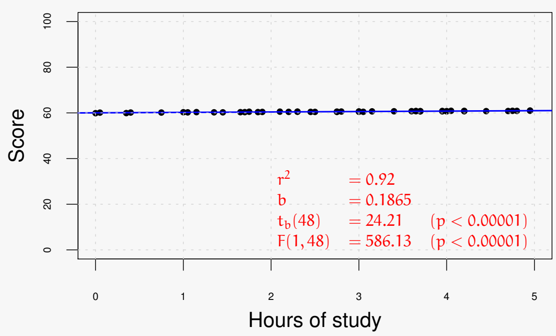
Visualizing regression results



Visualizing regression results



Visualizing regression results



Checking the validity of the model

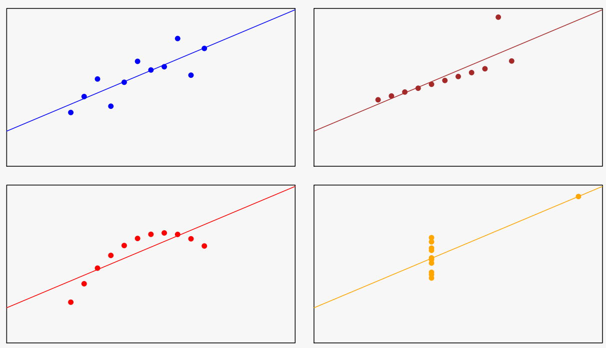
Before arriving at any conclusions from a model fit we need to do a few checks.

- The relationship between the response variable and the predictor should be *linear*.
- The residuals should be distributed normally with $mean = 0$.
- Residual variance should be constant.
- The residuals should be independent and identically distributed (i.i.d.).
- Least-squares regression is sensitive to *outliers*, more importantly *influential* observations.

How to detect influential observations?

- Influential observations affect the regression line.
- Outliers are easy to spot on a scatter plot for single predictor.
- Not all outliers are influential, an outlier is more likely to be influential if it has high leverage (having an extreme x value).
- One (of many) statistics that are used for detecting influential cases is Cook's distance, which measures the effect of removing a case from the regression estimation.
- The values for large (above 1) Cook's distance are typically considered influential.

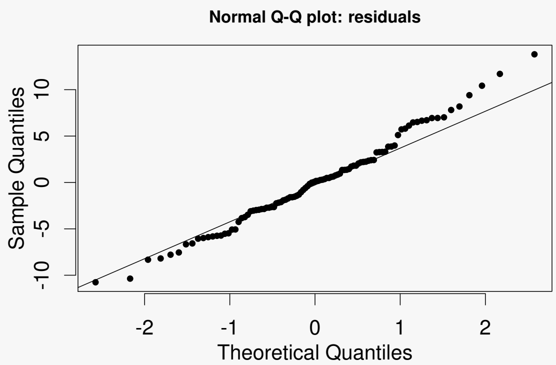
Always plot your data



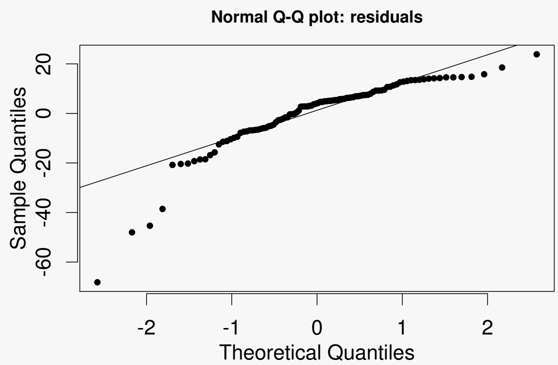
* This data set is known as Anscombe's quartet (Anscombe, 1973).

All four sets have the same mean, variance and fitted regression line.

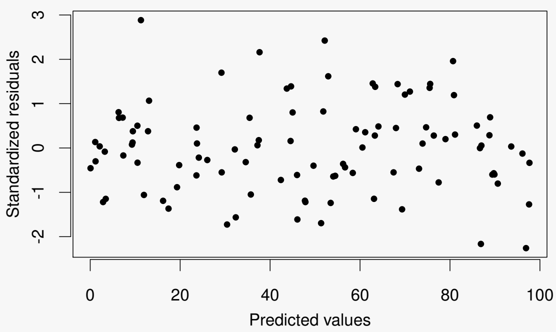
Normality of residuals: not bad



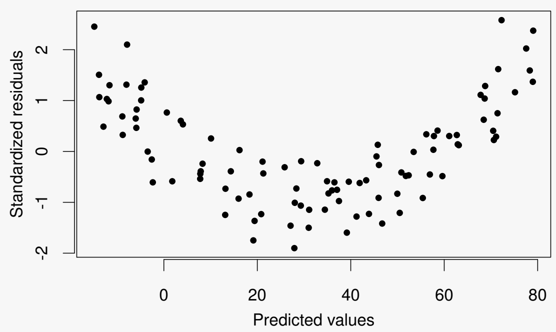
Normality of residuals: bad



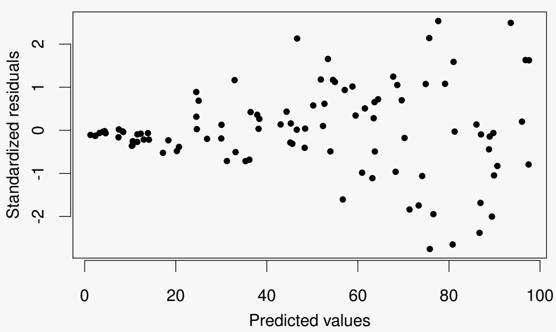
Checking residual distribution: good



Checking residual distribution: non-linear



Checking residual distribution: non-constant variance



Example: the data

We want to see the effect of mother's IQ to four-year-old children's cognitive test scores (Fake data, based on analysis presented in Gelman&Hill 2007).

Case	Kid's Score	Mom's IQ
1	109	91
2	99	102
3	96	88
...		
43	108	101
44	110	78
45	97	67

Example regression analysis (R output)

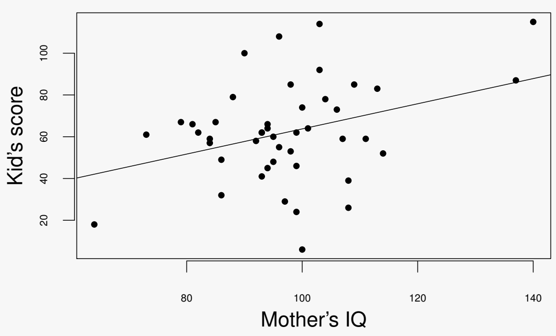
```
lm(formula = kid.score ~ mother.iq)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.5174   24.2375  0.145   0.885
mother.iq   0.6023    0.2471  2.437   0.019 *
---
Residual standard error: 22.59 on 43 degrees of freedom
Multiple R-squared: 0.1214, Adjusted R-squared: 0.101
F-statistic: 5.941 on 1 and 43 DF, p-value: 0.019
```

$b = 0.6$ Expected score difference between two children whose mother's IQ differs one unit.

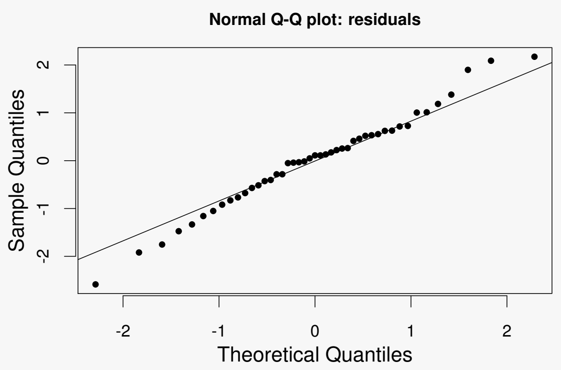
$r^2 = 0.12$ Mother's IQ explains 12% of the variation in test scores.

$p = 0.02$ Given the sample size, probability of finding a b value that far from 0 (two-tailed t-test with null hypothesis $b = 0$).

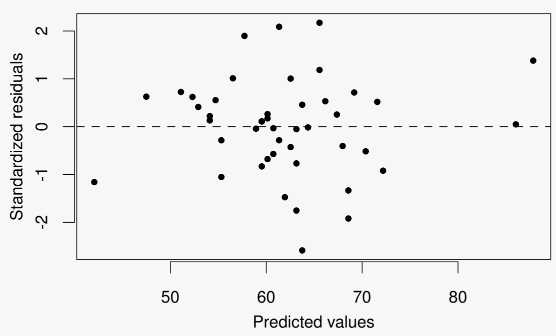
Example: scatter plot and the regression line



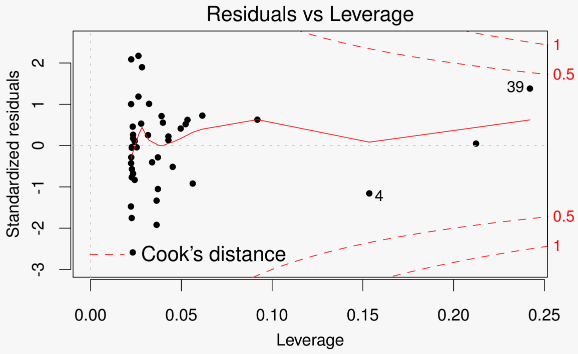
Example: normality of the residuals



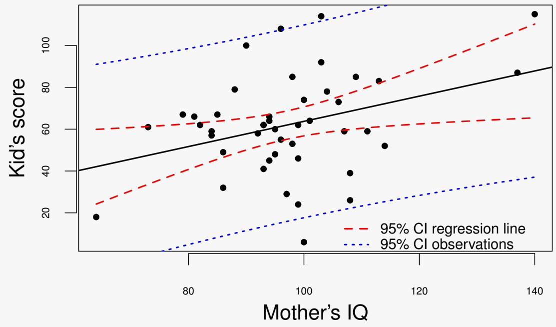
Example: residuals



Example: residuals vs. leverage



Example: prediction with the fitted model



Summary and Next week

Today:

- Some preliminaries: confidence intervals, hypothesis testing..
- Correlation
- Single regression

Next week:

- Multiple regression (sections 7.5–7.10 in 3rd edition, 8.5-8.9).

Estimating the regression line

We express the sum of squared residuals as a function of the (unknown) regression line:

$$\begin{aligned}\sum_{i=1}^{n} \epsilon_i^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n}(y_i - (a + bx_i))^2 \\ &= \sum_{i=1}^{n}(y_i - a - bx_i)^2 \\ &= \sum_{i=1}^{n}(a^2 + 2abx_i - 2ay_i + b^2x_i^2 - 2bx_iy_i + y_i^2)\end{aligned}$$

Thus, $\sum_{i=1}^{n} \epsilon_i^2$ is function f in x, y with unknown parameters a, b.

Estimating the regression line

For a fixed sample $\mathcal{S} = (x, y)$, we want to minimize $f_{ab}(x, y)$ with

$$f_{ab}(x, y) = \sum_{i=1}^{n}(a^2 + 2abx_i - 2ay_i + b^2x_i^2 - 2bx_iy_i + y_i^2)$$

To minimize this function, find a and b such that $f'_{ab}(x, y) = 0$.

Treat a and b as variables and find partial derivatives $\frac{\partial}{\partial a}f$, $\frac{\partial}{\partial b}f$

$$\begin{aligned}\frac{\partial}{\partial a}f = f'_{xyb}(a) &= \sum_{i=1}^{n}(2a + 2bx_i - 2y_i) \\ \frac{\partial}{\partial b}f = f'_{xya}(b) &= \sum_{i=1}^{n}(2ax_i + 2bx_i^2 - 2x_iy_i)\end{aligned}$$

Relationship between correlation and regression

Recall we obtained two partial derivatives (when minimizing sum of squared residuals):

$$f'_{xyb}(a) = \sum_{i=1}^{n}(2a + 2bx_i - 2y_i) \quad (1)$$

$$f'_{xya}(b) = \sum_{i=1}^{n}(2ax_i + 2bx_i^2 - 2x_iy_i) \quad (2)$$

Set (1) to zero:

$$\begin{aligned}& f'_{xyb}(a) = 0 \\ \Leftrightarrow\ & n \cdot 2a + \sum_{i=1}^{n}(2bx_i - 2y_i) = 0 \\ \Leftrightarrow\ & n \cdot 2a + 2b\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} y_i = 0 \\ \Leftrightarrow\ & n \cdot a = n \cdot \overline{y} - n \cdot b\overline{x} \\ \Leftrightarrow\ & a = \overline{y} - b\overline{x}\end{aligned}$$

Relationship between correlation and regression

Plug $a = \overline{y} - b\overline{x}$ into (2) and set to zero:

$$\begin{aligned}& f'_{xya}(b) = 0 \\ \Leftrightarrow\ & \sum_{i=1}^{n}(2(\overline{y} - b\overline{x})x_i + 2bx_i^2 - 2x_iy_i) = 0 \\ \Leftrightarrow\ & (\overline{y} - b\overline{x})(n\overline{x}) + b\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_iy_i = 0 \\ \Leftrightarrow\ & n\overline{x}\overline{y} - b\overline{x}^2n + b\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_iy_i = 0 \\ \Leftrightarrow\ & b(\sum_{i=1}^{n} x_i^2 - \overline{x}^2n) = \sum_{i=1}^{n} x_iy_i - n\overline{x}\overline{y} \\ \Leftrightarrow\ & b = \frac{\sum_{i=1}^{n} x_iy_i - n\overline{x}\overline{y}}{\sum_{i=1}^{n} x_i^2 - \overline{x}^2n}\end{aligned}$$

Relationship between correlation and regression

$$\begin{aligned}b = \frac{\sum_{i=1}^{n} x_iy_i - n\overline{x}\overline{y}}{\sum_{i=1}^{n} x_i^2 - \overline{x}^2n} \quad & \Leftrightarrow\ b = \frac{\sum_{i=1}^{n} x_iy_i - n\overline{x}\overline{y}}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ & \Leftrightarrow\ b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ & \Leftrightarrow\ b = \frac{1}{n-1}\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2)} \\ & \Leftrightarrow\ b = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - \overline{x})(y_i - \overline{y})}{\sigma_x^2} \\ & \Leftrightarrow\ b = \left(\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\sigma_x}\right)\left(\frac{y_i - \overline{y}}{\sigma_y}\right)\right) \cdot \frac{\sigma_y}{\sigma_x} \\ & \Leftrightarrow\ b = r\frac{\sigma_y}{\sigma_x}\end{aligned}$$

Another relation between correlation and regression

$$
\begin{aligned}
\frac{\text{explained variance}}{\text{total variance}} &= \frac{\sum_{i=1}^{n}((a + bx_i) - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \\
&= \frac{\sum_{i=1}^{n}((\overline{y} - b\overline{x} + bx_i) - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \\
&= \frac{\sum_{i=1}^{n} b^2(x_i - \overline{x})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \\
&= b^2 \cdot \left(\frac{\sigma_x}{\sigma_y}\right)^2 \\
&= r^2 \left(\frac{\sigma_y}{\sigma_x}\right)^2 \cdot \left(\frac{\sigma_x}{\sigma_y}\right)^2 \\
&= r^2
\end{aligned}
$$

Standard error for the regression slope and intercept

$$SE_b = \frac{s_r}{\sqrt{\sum(x_i - \bar{x})^2}}$$

$$SE_a = s_r \times \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$$