Statistics II
Regression & Correlation

Çağrı Çöltekin

ideas/examples/slides from
John Nerbonne & Hartmut Fitz

University of Groningen, Dept of Information Science

university of groningen

April 17, 2013

When, where, who?

- Lectures: Wednesday 13:00–15:00, Boeringzaal
- Computer Labs:

Group 1	Tue	09:00–11:001312.0107A	Mets Visser
Group 3	Thu	11:00–13:001312.0119A	Mets Visser
Group 4	Thu	13:00–15:001312.0119A	Carmen Klaussner
Group 2	Fri	11:00–13:001312.0119A	Carmen Klaussner

- Office Hours: Wednesday 10:00–12:00, or by appointment (email c.coltekin@rug.nl).
- Course web page: `http://www.let.rug.nl/coltekin/statII/`

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 (5%): 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.

The plan

1. Simple regression
2. Multiple regression
3. ANOVA
4. Factorial ANOVA
5. Repeated measures ANOVA
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

Why do (inferential) statistics?

If your experiment needs statistics, you ought to have done a better experiment. — Ernest Rutherford

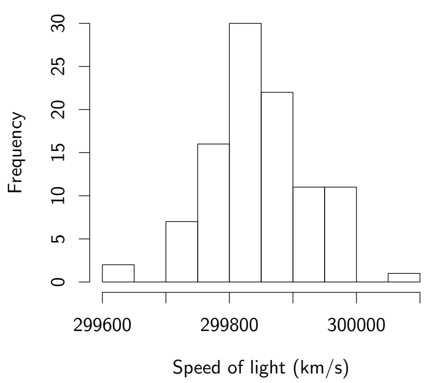
- Our results are based on a **sample**, we want to generalize to the **population** the sample was drawn from.
- The values we obtain include **measurement error**.

Even a very precise experiment cannot account for all sources of **variation**.

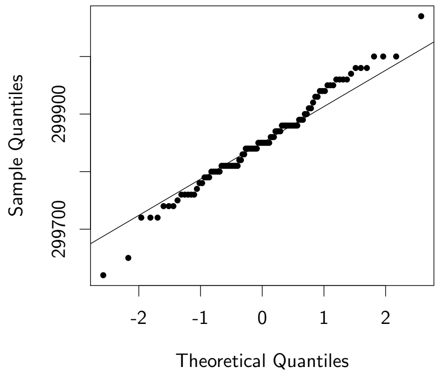
The speed of light

- In 1879, A. Michelson took 100 measurements of the speed of light ($n = 100$).
- The data looks like $x_{1..n} = 299850, 299740, 299900, 300070, 299930 \ldots$
- The mean is, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 299852.4$.
- Estimated variance is $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 6242.67$
- Estimated standard deviation is $s = \sqrt{6242.66} = 79.01$.
- Based on this data what is our best estimate of the speed of light?
- Why do individual measurements differ?

Speed of light: histogram



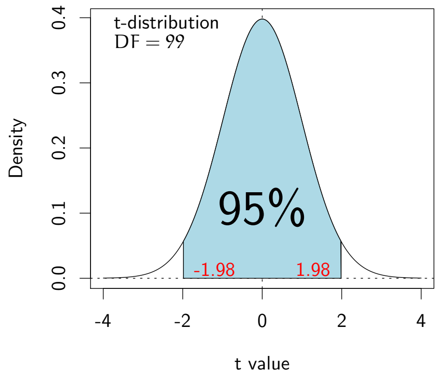
Speed of light: is the distribution normal?



Confidence intervals: accounting for uncertainty

- A confidence 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 given probability (e.g., 95%).

How certain are we about these measurements?



Confidence intervals: how to calculate it

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

$$\begin{aligned}
-2 < & \frac{299852.4 - \mu}{\frac{79.01}{\sqrt{100}}} < 2 \\
-2 \times 7.9 < & \ 299852.4 - \mu < 2 \times 7.9 \\
-2 \times 7.9 - 299852.4 < & -\mu < 2 \times 7.9 - 299852.4 \\
-299868.2 < & -\mu < -299836.6 \\
299836.6 < & \ \mu < 299868.2
\end{aligned}$$

We are 95% confident that the true mean is in the range $[299836.6, 299868.2]$.

Basic hypothesis testing: one sample t-test

The known the value of the speed of light in vacuum is $299,792.458 km/s$. Assuming the previous example was testing a special case, we set our hypotheses:

H_0 : The speed of light in the experiment condition is $299,792.458 km/s$.

H_a : The speed of light in the experiment condition is different than $299,792.458 km/s$ (two-tailed hypothesis).

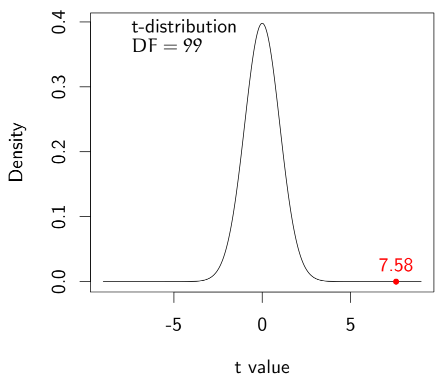
Since 95% confidence interval $[299,836.6, 299,868.2]$ does not include $299,792.458$, we would reject the null hypothesis, and conclude that we found a difference with α-level $= 0.05$.

Basic hypothesis testing: looking it another way

- Calculate the t-score for the mean, given the null hypothesis is true:
$$\begin{aligned} t &= \frac{\bar{x} - \mu}{SE_{\bar{x}}} \\ &= \frac{299852.4 - 299792.458}{7.9} \\ &= 7.59 \end{aligned}$$
- Calculate the probability a value this extreme under the t-distribution with $DF = 99$ (or check via probability tables).

$$\begin{aligned} p &= 1.9 \times 10^{-11} \\ &= 0.0000000000019 \end{aligned}$$

Basic hypothesis testing: visualizaiton



Some terms you should know

If you are not familiar with the following, it is time to go back to your Statistcs 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

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

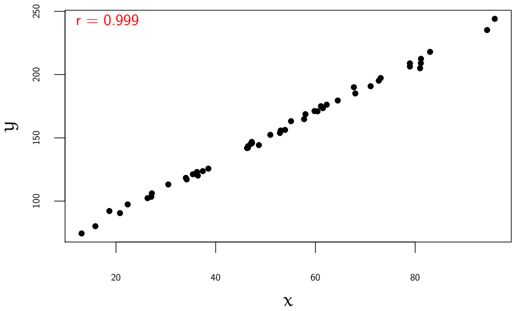
Correlation coefficient is a standardized measure of covariance between two variables, x and y. It takes values between -1 and 1

- 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.

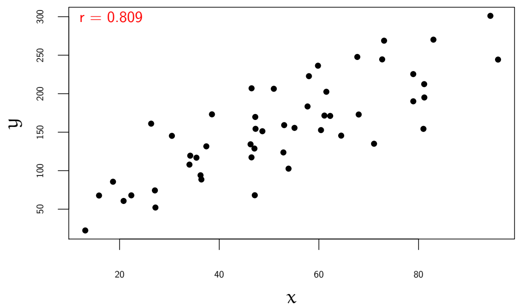
Scatter plots

*Scatterplot*s are a good way to visualize the relationship between two variables:



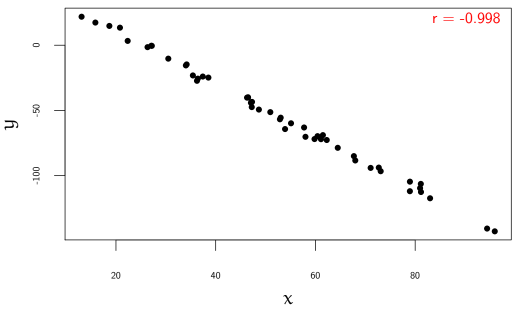
Scatter plots

*Scatterplot*s are a good way to visualize the relationship between two variables:



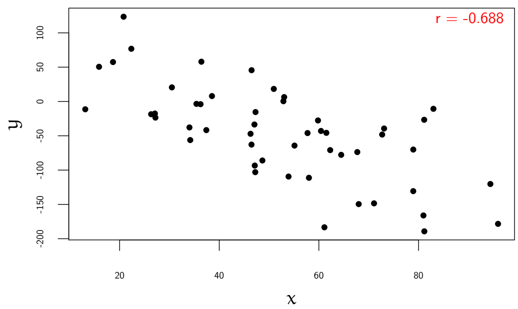
Scatter plots

*Scatterplot*s are a good way to visualize the relationship between two variables:



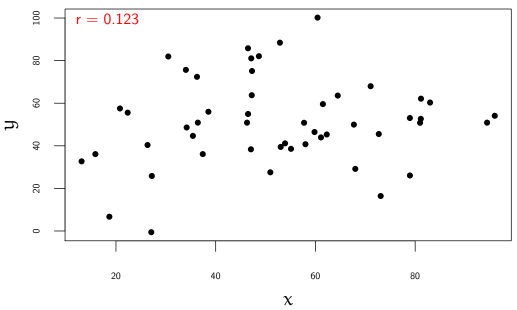
Scatter plots

*Scatterplot*s are a good way to visualize the relationship between two variables:



Scatter plots

*Scatterplot*s 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 and weaknesses 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 to multiple predictor variables easily (next week).

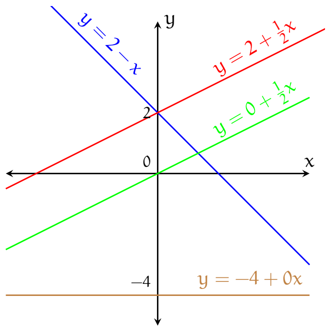
The linear equation

$y = a + bx$

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 regression equation

$$y_i = a + bx_i + e_i$$

- 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.
- $a + bx$ is the *deterministic* part of the model (we sometimes use $\hat{y}$).
- 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).

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:

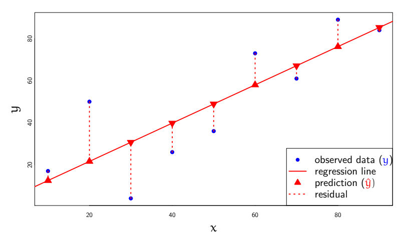
$$\sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - (a + bx_i))^2$$

- This minimization problem can be solved analytically, yielding:

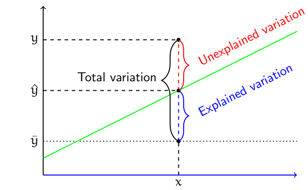
$$b = r\frac{\sigma_y}{\sigma_x}$$
$$a = \bar{y} - b\bar{x}$$

* See appendix for the derivation.

Visualization of regression procedure



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

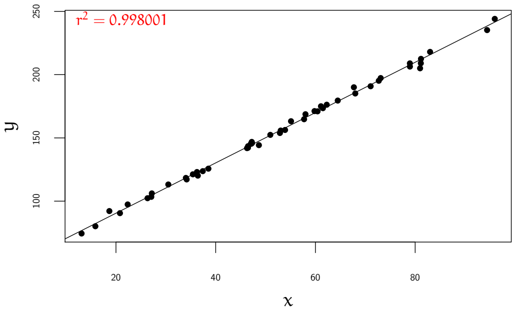
We can express the variation explained by a regression model as:

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

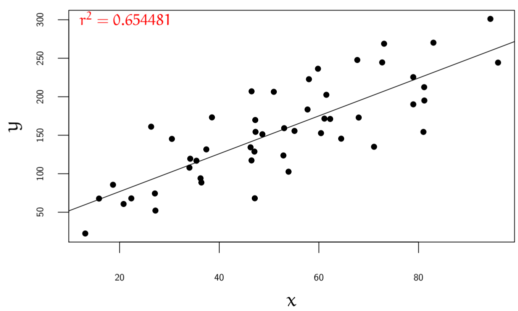
It can be shown that this value is the square of the correlation coefficient, r^2, also called the coefficient of determination.

- $100 \times r^2$ 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.

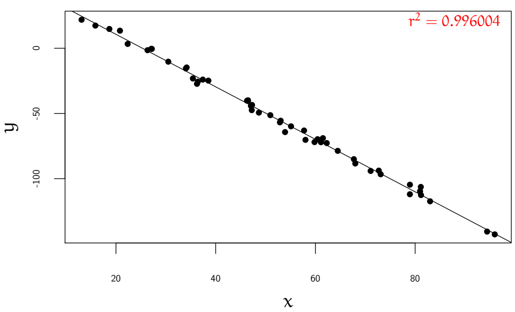
r^2: examples



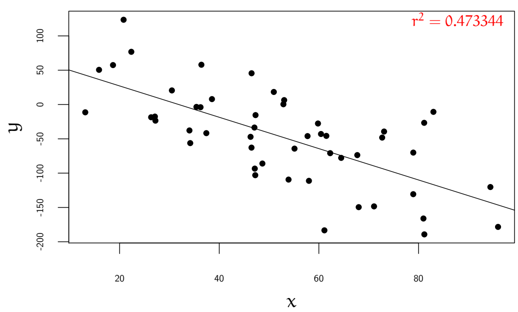
r^2: examples



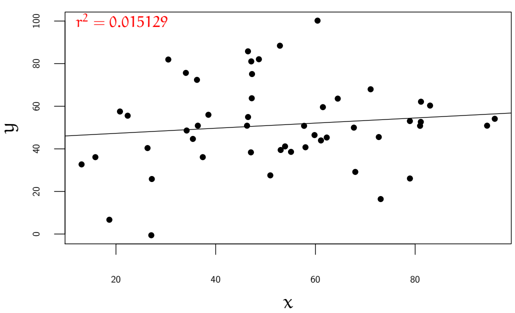
r^2: examples



r^2: examples



r^2: examples



Inference for regression

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.
- If the estimate of the slope (b) is statistically significant, the effect of predictor on the response variable is not due to chance. In other words: we are confident about the direction (sign) of the effect.

F-test for regression

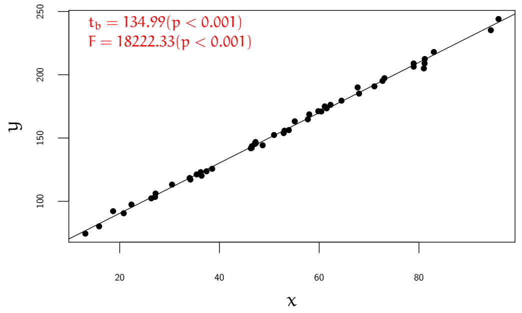
We can also test whether the overall model fit is significant. To do this, we use the ratio,

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

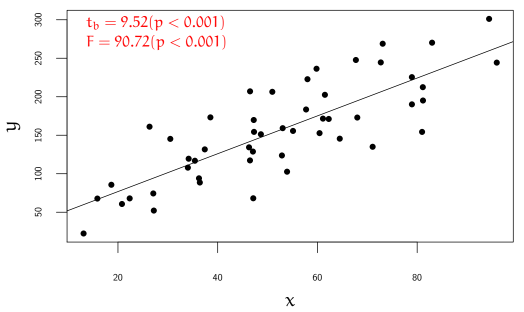
- This ratio follows an F-distribution with $DF = (1, n-2)$.
- Note: MS_M is the variance explained by the regression line in comparison to the mean of y, the null model.
- We require variance explained to be larget than the unexplained variance. So, we test for $F > 1$.
- This test is equivalent to the t-test for the slope for simple regression.

* More on F-distribution later.

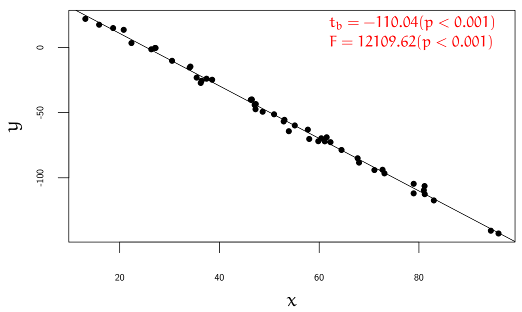
Significance: examples



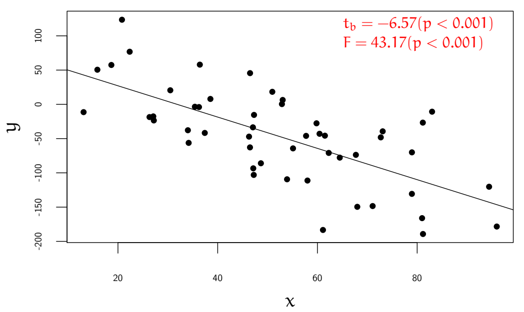
Significance: examples



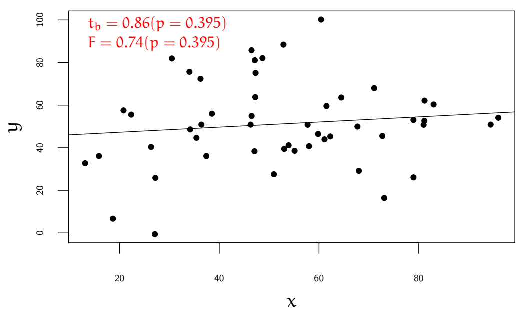
Significance: examples



Significance: examples



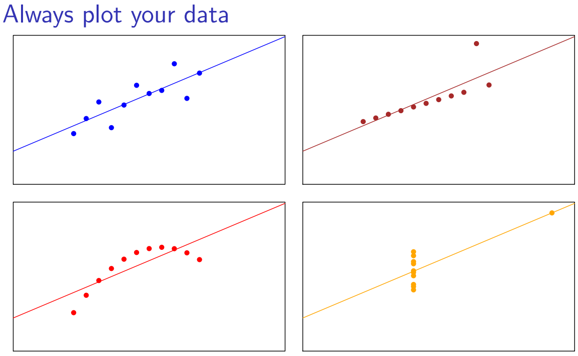
Significance: examples



Checking the validity of the model

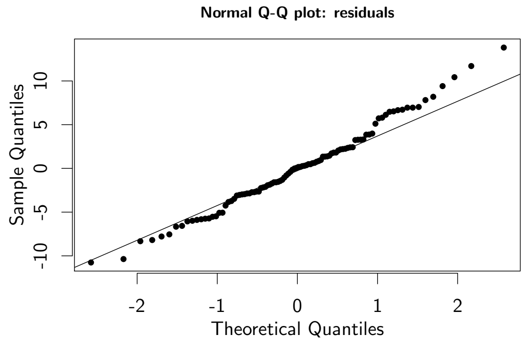
- The relationship between the response variable and the predictor should be *linear*.
- The residuals should be distributed normally with $mean = 0$. (As a result, the response variable should also be normally distributed).
- The residuals should be independent for any two observation.
- Least-squares regression is sensitive to *outliers*, more importantly *influential* observations.

Always plot your data

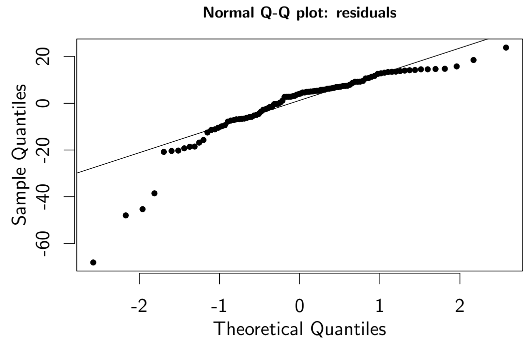


* This data set is known as Anscombe's quartet (Anscombe, 1973). All four sets have the same mean, variance and fitted regression line.

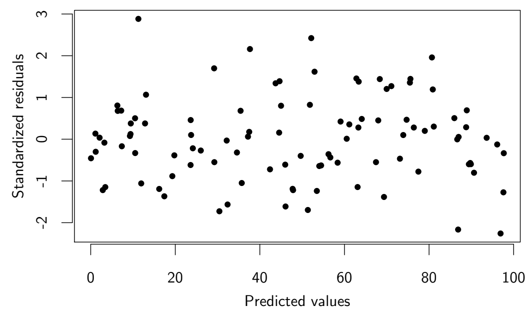
Normality of residuals: not bad



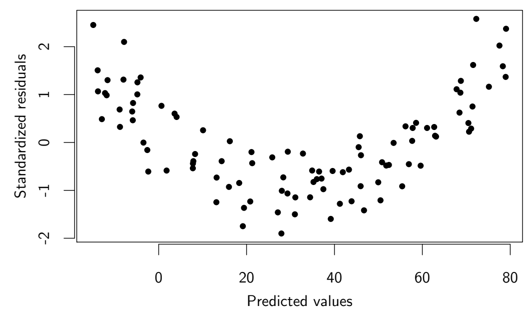
Normality of residuals: bad



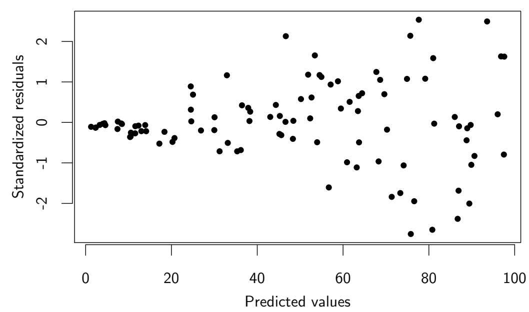
Checking residual distribution: good



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).

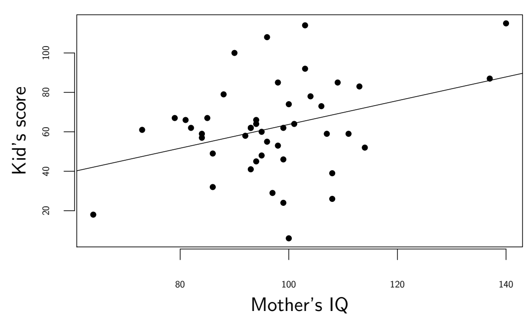
Case	Kid's Score	Mom's IQ
1	109	91
2	99	102
3	96	88
...		
43	108	101
44	110	78
45	97	67

Example: regression analysis in R

```
> lm(kid.score ~ mother.iq)
Call:
lm(formula = kid.score ~ mother.iq)
Coefficients:
(Intercept)   mother.iq
     3.5174      0.6023
```

How do we interpret the intercept and the slope? (assuming our model assumptions are correct)

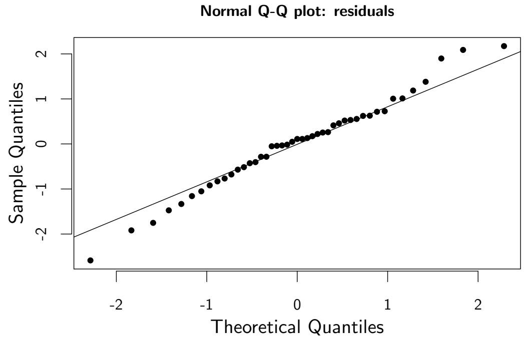
Example: scatter plot and the regression line



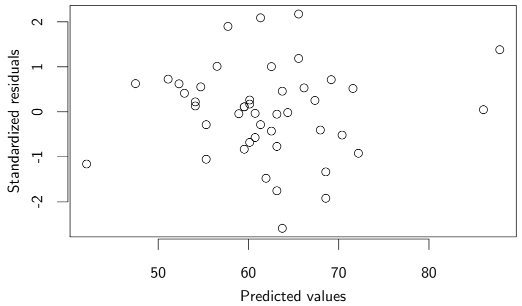
Example: inference and the model fit

```
> summary(lm(kid.score ~ mother.iq)
Call:
lm(formula = kid.score ~ mother.iq)
Residuals:
    Min      1Q  Median      3Q     Max
-57.749 -12.737 2.467 12.286 48.444
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 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
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
```

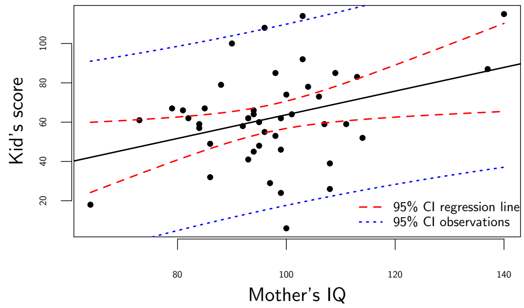

Example: normality of the residuals



Example: residuals



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).

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:

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

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

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

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

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

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