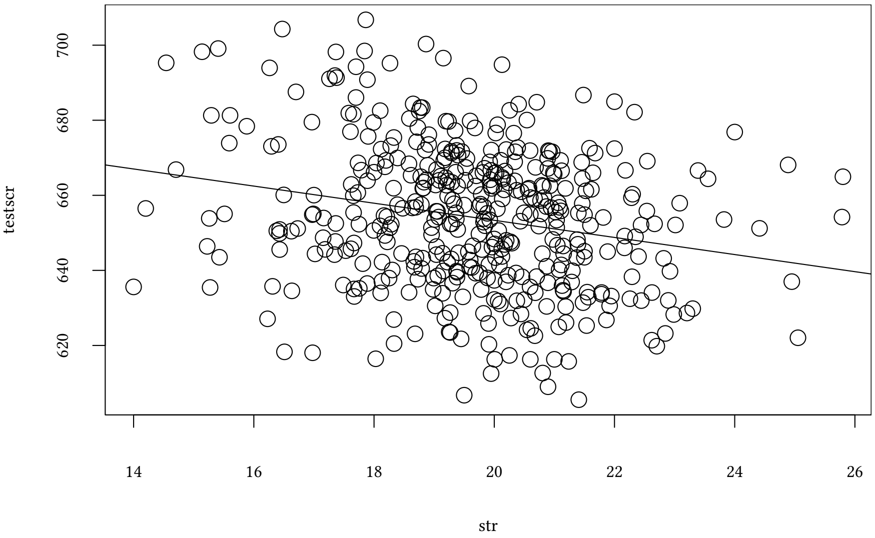
Empirical Methods



```
Call:
lm(formula = testscr ~ str)

Residuals:
    Min      1Q  Median      3Q     Max
-47.727 -14.251   0.483  12.822  48.540

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept) 698.9330     9.4675  73.825    < 2e-16 ***
str          -2.2798     0.4798  -4.751 0.00000278 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.58 on 418 degrees of freedom
Multiple R-squared:  0.05124,Adjusted R-squared:  0.04897
F-statistic: 22.58 on 1 and 418 DF,  p-value: 0.000002783
```

Oliver Kirchkamp

This handout is a summary of the slides we use in the lecture. Please expect slides and small parts of the lecture to change from time to time and print only the material you currently need.

The handout is perhaps not very helpful unless you also attend the lecture. This handout is also not supposed to replace a book. At the end of each chapter you find recommendations which textbooks you can use to revise the chapter.

Contents

Components:

- Lecture
- Exercise
- Homework (counts for final grade)
- Discussion board
- (Q+A meeting, if necessary)

Homepage:
https://www.kirchkamp.de/mw241/

Literature:

- William H. Greene. “Econometric Analysis”. Pearson.
- James H. Stock and Mark W. Watson. “Introduction to Econometrics”. Pearson.
- John Kruschke. “Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan”. Academic Press.

Routines to support your learning experience

- Learn together with others. Form a learning group.
- Follow a routine. Learn each week at the same time.

Homework

- Problem: The skills we learn in this course require (regular) practice.
- How can you distribute your practice over the entire term? How do you get feedback?

→ Weekly homework, weekly feedback.

Diskussion board

- As long as your questions could be relevant for the other students, please use the discussion board.

Empirical methods — Where are we?

Statistics:

- Descriptive statistics
- Inferential statistics
  - $\vdots$
  - Biometrics
  - Psychometrics
  - Econometrics / Empirical Methods
    - Frequentist methods in Econometrics
    - Bayesian methods in Econometrics

1 Review of Probability and Statistics

1.1 How do economists work?

- Develop theories
- Test theories
- Use theories to predict

(Theories can be more or less specific and formalised)

Desirable properties of economic theories

$$\begin{array}{ll} & \left.\begin{array}{l}\text{Congruence with reality}\\ \text{Generality}\\ \text{Tractability}\end{array}\right\}\ \textit{(Stigler, 1965)} \\ \hline + & \left.\begin{array}{l}\text{Parsimony}\\ \text{Falsifiability}\\ \text{Predictive precision}\\ \text{Conceptually insightful}\end{array}\right\}\ \begin{array}{l}\textit{(Gabaix,}\\ \textit{Laibson,}\\ \textit{2008)}\end{array}\end{array}$$

Claim

- For each economic theory there is an alternative theory predicting the opposite.

  Often economic theories *suggest* relationships — often with policy implications — but these relationships are rarely *quantified.*
  - How large is the increase in the performance of students when courses are smaller?

  - How large is the increase in your income when you study for another year?
  - What is the elasticity of demand for cigarettes?
  - By how much does the GDP increase/decrease if the ECB raises interest rates by 1%?

1.2 Econometrics uses data to measure causal relationships

- Ideal approach: controlled experiment (control group/treatment group)
  - By how much does the performance of students increase when courses are smaller?
  - How much more do you earn if you spend an additional year at university.
  - What is the price elasticity of cigarettes?
  - By how much does the GDP increase if the ECB reduces the interest by one percentage point?
- Controlled experiments can be hard.
- Most of the data we have are from uncontrolled processes.
  - Student test scores
  - Incomes of alumni
  - Time series data about monetary policy
- Problems related to data from uncontrolled processes
  - Unobserved factors
  - Simultaneous causalities
  - Coincidence ↮ causality

→ Application of econometric methods

  - Quantifying causal effects using observational data from uncontrolled processes
  - Extrapolating times series

→ Evaluating the econometric work of others

1.3 Working with R

Software:

- R
  - free
  - wide range of applications
  - In the lecture we will illustrate many things with R. You should try these examples on you own computer. Use the online help to look up unknown commands.
- SAS, STATA, EViews, TSP, SPSS,...
  - expensive
  - more specialised
  - more heterogeneous syntax

- Why can't we use spreadsheet software (Libreoffice, Gnumeric, Microsoft Excel...) to structure our data?
- Using spreadsheet software makes it harder to document our work.

Spreadsheet: Hard to document, error prone:

- In the menu File/Open choose the file `D.csv`.
- In the next dialog specify options how the file should be imported.
- Select the menu Data/Statistics/Regression
- In the next dialog specify...
  - Independent variable range: `$D06.$A$2:$A$41`
  - Dependent variable range: `$D06.$B$2:$B$41`
  - ⋮

Notation like `$D06.$A$2:$A$41` is not obvious and is error prone.

Statistical software (e.g. R): Easy to document:

```
lm(y ~ x,data=read.csv("D.csv"))
```

Almost always we will return to our work:

- Reproducable work saves time.

- If our work is well documented, it is easy to continue our own work, even after a break.

- Claim: It is obvious how statistical analysis should be carried out.

- Everything follows from the data and the question.

Unfortunately this is not the case:

E.g.: Silberzahn, R., Uhlman, D.,...Nosek, B. (2018). Many analysts, one data set: making transparent how variations in analytic choices affect results. *Advances in Methods and practices in Psychological Science* (1). 337-356.

Authors give the same data and the same question to 29 teams of researchers.

- Question: Are soccer referees likely to give red cards to dark-skin-toned players than to light-skin-toned players?

- Answer (depending on the team): dark-skin-toned players receive...
  - 11% fewer yellow and red cards.

    ⋮
  - 2.93 × more yellow and red cards.

In this course we will use R. You should install two pieces of software:

- R

  Instructions can be found on `https://cran.r-project.org/`

- A frontend for R. Here I use RStudio.

  `https://rstudio.com/products/rstudio/download/`

- R organises its functions and data in "libraries". Before we can use a library, we should tell R:

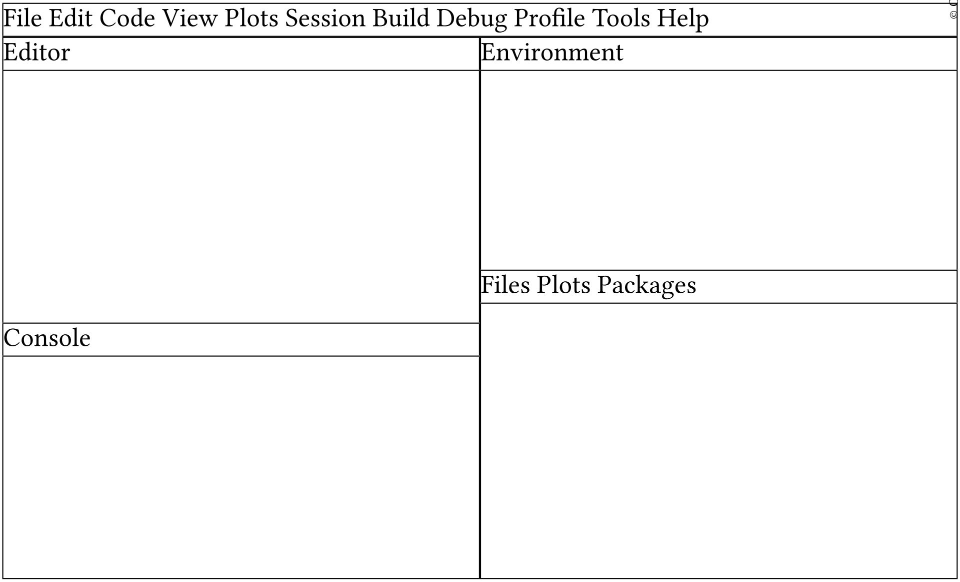
  ```
  library(Ecdat)
  ```

  This succeeds if the library is installed. If we get an error...

  ```
  Error in library(Ecdat): there is no package called 'Ecdat'
  ```

  ...then we must install the library (e.g. from RStudio, bottom right, tab »Packages / Install«).

A frontend (like RStudio) simplifies working with R.


File Edit Code View Plots Session Build Debug Profile Tools Help
Editor
Environment
Files Plots Packages
Console

- You might find yourself confronted with intimidating R syntax. You should not be intimidated.

- You should learn to understand the main results (text and graphs) of statistical software.

1.4 An example for statistical inference

- How does learning success change when class size is reduced by one student? What, if class size is reduced by eight students?

- Can we answer this question without using data?

- E.g. test scores from 420 school districts in California from 1998-1999
  - `str` = student teacher ratio number of students in the district / full time equivalent teachers
  - `testscr` = 5th-grade test score (Stanford-9 achievement test)

For our examples we use the statistical software R.

R organises data and methods in libraries. To use them, you need the command `library`:

```
library(Ecdat)
library(car)
```

To access a canned data set we use the command `data`:

```
data(Caschool)
```

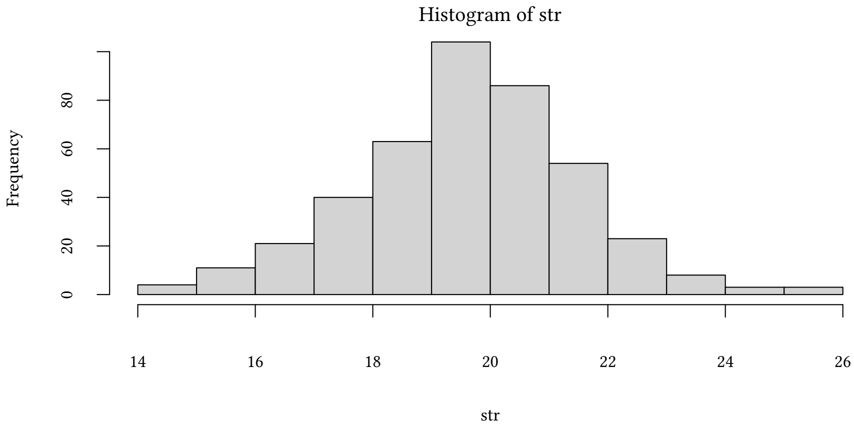
We can now use components of this data set:

```
head(Caschool)

  distcod  county                        district grspan enrltot
1   75119 Alameda              Sunol Glen Unified  KK-08     195
2   61499   Butte            Manzanita Elementary  KK-08     240
3   61549   Butte     Thermalito Union Elementary  KK-08    1550
4   61457   Butte Golden Feather Union Elementary  KK-08     243
5   61523   Butte        Palermo Union Elementary  KK-08    1335
6   62042  Fresno         Burrel Union Elementary  KK-08     137
  teachers calwpct mealpct computer testscr   compstu  expnstu
1    10.90  0.5102  2.0408       67  690.80 0.3435898 6384.911
2    11.15 15.4167 47.9167      101  661.20 0.4208333 5099.381
3    82.90 55.0323 76.3226      169  643.60 0.1090323 5501.955
4    14.00 36.4754 77.0492       85  647.70 0.3497942 7101.831
5    71.50 33.1086 78.4270      171  640.85 0.1280899 5235.988
6     6.40 12.3188 86.9565       25  605.55 0.1824818 5580.147
       str    avginc     elpct readscr mathscr
1 17.88991 22.690001  0.000000   691.6   690.0
2 21.52466  9.824000  4.583333   660.5   661.9
3 18.69723  8.978000 30.000002   636.3   650.9
4 17.35714  8.978000  0.000000   651.9   643.5
5 18.67133  9.080333 13.857677   641.8   639.9
6 21.40625 10.415000 12.408759   605.7   605.4
```

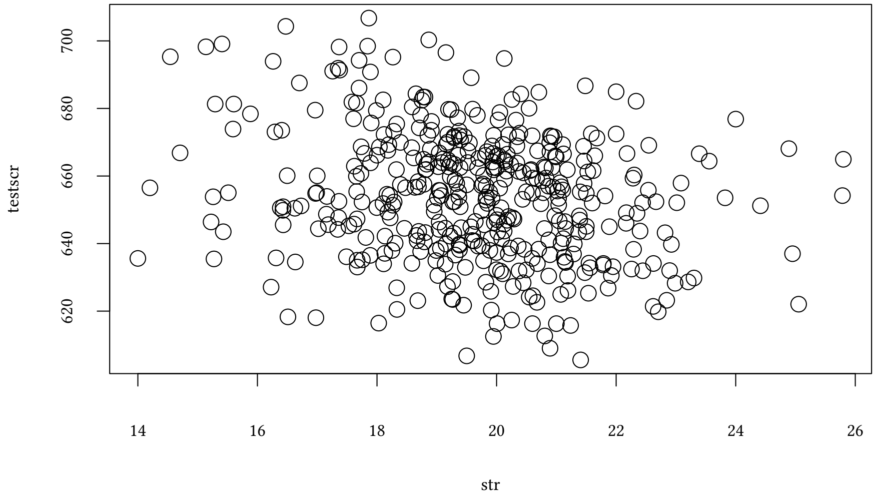
It is cumbersome to write the name of a data set - here `Caschool` - time and time again. Whenever we intend to work with the same data set for a while we can use `attach (Caschool)`. This will tell R to look at `Caschool` first, whenever we ask for a variable.

```
attach(Caschool)
```

```
hist(str)
```



```
plot(testscr ~ str)
```



In our sample test results *testscr* seem to be getting worse as student teacher ratios *str* are getting higher.

Is it possible to show that districts with low student teacher ratios `str` have systematically higher test scores `testscr`?

- Compare average test scores in districts with small `str` to test scores in districts with high `str` (estimation)
- Test the null hypothesis that mean test scores are the same against the alternative hypothesis that they are not (hypothesis testing)
- Estimate an interval for the difference of the mean test scores (confidence interval / credible interval)
- Is the difference large enough
  - for a school reform
  - to convince parents
  - to convince the school authority

In the following example we want to split up the data set into two pieces — schools with a student teacher ratio above and below 20. In other words, we will introduce a nominal variable. In R a nominal variable is called a factor. `factor` converts a continuous variable (`str`) into a factor.

`t.test` performs a student-t test to compare mean values.

```
head(testscr)

[1] 690.80 661.20 643.60 647.70 640.85 605.55

head(str)

[1] 17.88991 21.52466 18.69723 17.35714 18.67133 21.40625

head(str>20)

[1] FALSE  TRUE FALSE FALSE FALSE  TRUE

large <- str>20
head(large)

[1] FALSE  TRUE FALSE FALSE FALSE  TRUE
```

```
t.test(testscr ~ large)


Welch Two Sample t-test

data:  testscr by large
t = 3.9231, df = 393.72, p-value = 0.0001031
```

```
alternative hypothesis: true difference in means between group FALSE and group TRUE is not
95 percent confidence interval:
  3.584445 10.785813
sample estimates:
mean in group FALSE  mean in group TRUE
           657.1846            649.9994
```

We first calculate a categorical variable `large`. Then we describe a relationship `testsrc ~ large` and use `t.test` to test this relationship. The variable to be tested is given before the tilde in the relationship. The factor describing the two groups to be compared is given after the tilde.

This simple test tells us that there is a significant difference of the test scores `testscr` between large and small school groups.

We can *estimate* the difference between the two groups, we can *test* a hypothesis, and we can calculate a *confidence interval.*

Plan

- Estimates, hypotheses tests and confidence intervals.
- Generalize these concepts to regressions.

1.5 Populations and Samples

1.5.1 Population

- The set of all entities which could theoretically be observed (e.g. all imaginable school districts at all points in time under all imaginable conditions)
- Often we assume that the *population* is of infinite size (or at least very large)
- Usually we know something about A (our sample) and we want to say something about B. We can do this, if we assume that both A and B are drawn from the same *population.*

1.5.2 Random variables and distributions

- random variable (RV) = *numerical* summary of random event
  - discrete (categorial, factor) variable / continuous variable
  - one-dimensional variable / multi-dimensional variable
- Describing random variables using distributions
  - Probabilities of events $\Pr(x) = \Pr(X = x)$
    - discrete RV

- Cumulative distribution function $F(x) = P(X < x)$
  - * one-dimensional RV
- Probability density function $f(x) = \frac{dF(x)}{dx}$
  - * continuous RV

Properties of random variables / distributions

- Expected value $E(X)$, μ_X, (theoretical) mean value of X

  mean value of X for the entire population

- Variance $E((X - \mu_X)^2) = \sigma_X^2$

  Measure of the mean of the squared deviation from the mean of the distribution

- Standard deviation $\sqrt{\text{variance}} = \sigma_X$

Generally,

- $X \sim F(\theta_1, \theta_2, \ldots)$

where θ_1, θ_2, ... are parameters of the distribution.

E.g. the normal distribution is characterised by μ and σ^2:

- $X \sim N(\mu, \sigma^2)$

1.5.3 Sample

- A part of the population that we observe (e.g. Californian school districts in 1998 (and under the conditions of this year))

Population

- Random variable, distribution
- Moments of a distribution (mean value, variance, standard deviation, covariance, correlation)
- Conditional distribution, conditional mean values
- Distribution of a random sample

Sample

- Realisation of a random variable, empirical distribution
- Empirical moments (mean value, variance, standard deviation, covariance, correlation)
- Conditional empirical distribution, conditional empirical mean values

Independence

- Consider the sample $Y_1 \ldots Y_n$ of a population Y
- Before the sample is drawn, the $Y_1 \ldots Y_n$ are random.
- After the sample was drawn, the values of $Y_1 \ldots Y_n$ are "realised", they are fixed numbers — they are not random anymore.
- $Y_1 \ldots Y_n$ is the data set. Y_i is the value of Y for observation i (person i, district i etc.)
- If we draw a sample *randomly*, it is true that
  - If Y_i is drawn randomly, we *can't predict* Y_i before it is drawn.
  - Y_i and Y_j are *independently distributed*
  - Y_i and Y_j were drawn from the same distribution. Thus, they are *identically distributed*
  - We say that Y_i and Y_j are independently and identically distributed (= *i.i.d.*).
  - More generally speaking: Y_i are i.i.d. for $i = 1, \ldots, n$.

1.6 Estimations

- Populations are characterised by parameters (θ)
- We draw a sample $Y_1 \ldots Y_n$ from the population. The sample gives us some information about the population Y.
- We use the sample to *estimate* properties (parameters θ) of the population.

We start with a simple problem: How can we estimate the mean value of Y (not the mean value of $Y_1 \ldots Y_n$)?

Idea:

- We could simply use the mean value $\bar{Y}$ of the sample $Y_1 \ldots Y_n$
- We could simply use the first observation Y_1
- We could use the median of the sample $Y_1 \ldots Y_n$
- $\cdots$

Which criteria should we use to pick an estimator?

1.6.1 Desirable properties of estimators

Unbiasedness A point estimate $\hat{\theta}(X_1,\ldots,X_n)$ to estimate θ is an *unbiased* estimate of θ iff

$$\forall\theta : E\big(\hat{\theta}(X_1,\ldots,X_n)\big) = \theta$$

Efficiency Let us compare two *unbiased* estimators $\hat{\theta}_1$, $\hat{\theta}_2$:
$\hat{\theta}_1$ is *more efficient* than $\hat{\theta}_2$ ($\hat{\theta}_1$ dominates $\hat{\theta}_2$) iff

$$\forall\theta : \text{var}(\hat{\theta}_1(X_1,\ldots,X_n)) < \text{var}(\hat{\theta}_2(X_1,\ldots,X_n))$$

The unbiased estimator is called *most efficient* estimator if for all unbiased estimators $\hat{\theta}'$ we have

$$\forall\theta : \text{var}(\hat{\theta}(X_1,\ldots,X_n)) \leq \text{var}(\hat{\theta}'(X_1,\ldots,X_n))$$

Consistency An estimator is *consistent* iff

$$\forall\epsilon > 0 : \lim_{n\to\infty} \Pr(|\hat{\theta}(X_1,\ldots,X_n) - \theta| < \epsilon) = 1$$

1.6.2 Characteristics of sampling distributions

Sampling distribution = distribution of a statistic of a random sample (Y)
An interesting statistic is the sample *mean*: $\bar{Y}$.

Expected value of $\bar{Y}$
$E(\bar{Y}) = \mu_Y$, i.e. $\bar{Y}$ is an *unbiased* estimator of μ_Y

Variance of $\bar{Y}$
How does the variance of $\bar{Y}$ depend on the size of the sample n?

$$\text{var}(\bar{Y}) = \frac{\sigma_Y^2}{n}$$

Question: Does $\bar{Y}$ converge to μ_Y if n is large?

Law of large numbers:
$\bar{Y}$ is a *consistent* estimator of μ_Y.
Formally: If $Y_1,\ldots,Y_n$ i.i.d. and $\sigma_Y^2 < \infty$, then $\bar{Y}$ is a consistent estimator of μ_Y, i.e.

$$\forall_{\epsilon>0} : \lim_{n\to\infty} \Pr(|\bar{Y} - \mu_Y| < \epsilon) = 1 \qquad \text{we can also say} \qquad \bar{Y} \xrightarrow{p} \mu_Y$$

Central Limit Theorem:

If $Y_1, \ldots, Y_n$ i.i.d. and $0 < \sigma_Y^2 < \infty$ and n is large, then the distribution of $\bar{Y}$ approximates a normal distribution

$$\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$$

1.7 Literature

- Stock, Watson. "Introduction to Econometrics". Chapters 1-3.

Appendix 1.A Examples for the lecture

Example 1: Your sample contains 4 independently and identically distributed observations: $X_1, \ldots, X_4$. Which estimators for $E(X)$ are unbiased?

- $\frac{1}{3}X_1 + \frac{2}{3}X_2$
- $X_1 - X_2$
- $\frac{1}{3}X_1 + \frac{1}{3}X_3$
- $\frac{1}{4}\sum_{i=1}^{4} \sqrt[3]{X_i^3}$
- $\frac{1}{4}\sum_{i=1}^{4} \sqrt{X_i^2}$

Now assume that the variance of X is 10. What can you say about the variance of $X_1 + 2X_2 - X_4$?

Example 2: Your sample contains 5 independently and identically distributed observations: $X_1, \ldots, X_5$. Which estimators for $E(X)$ are unbiased.

- $\sum_{i=1}^{5} X_i$
- X_3
- $\frac{X_2}{2} + \frac{X_3}{2}$
- $X_1 + X_2 - X_3$
- $\frac{1}{5}\sum_{i=1}^{5} X_i$

Now $\text{var}(X) = 90$. What can you say about the variance of $\frac{1}{3}(X_1 + X_5)$?

Example 3: Which estimator for $\mathsf{E}(X)$ is more efficient?

- $\frac{X_2}{2} + \frac{X_3}{2}$ is more efficient than X_3 ?
- X_3 is more efficient than $X_1 + X_2 - X_3$
- $\frac{1}{5}\sum_{i=1}^{5} X_i$ is more efficient than $\sum_{i=1}^{5} X_i$
- $\frac{1}{5}\sum_{i=1}^{5} X_i$ is more efficient than X_3
- $\frac{X_2}{2} + \frac{X_3}{2}$ is more efficient than $X_3 + X_4 - X_5$

Appendix 1.B Exercises

Econometrics:

Exercise 1.1 • *What is econometrics?*

- *What is the aim of econometrics?*
- *Which kind of questions can be answered with econometric tools? Give some examples of questions that can possibly be answered with econometrics. Give examples of different fields, e.g. policy advice, science, marketing, ...*

Data sources:

Exercise 1.2 • *What are typical data sources for econometric analysis?*

- *Which problems might be related to the different data sources?*

Revision I

Exercise 1.3 *In the following task we will refresh some basic concepts:*

You have the following data about children's age (`a`*) and the pocket money (*`pm`*) they receive from their parents on children in elementary school.*

age in years (`a`*)*	*pocket money/week in \$ (*`pm`*)*
6	*0*
7	*2*
6	*2*
7	*3*
8	*4*
8	*2*
9	*5*
10	*4*
9	*2*
10	*4*

Compute the following items:

- *Mean (`pm`)*
- *Median (`pm`)*
- *First/third quartile (`pm`)*
- *Variance (`pm`)*
- *Standard deviation (`pm`)*
- *Correlation (`a, pm`)*

Revision II

Exercise 1.4 *Define the following items:*

- *Confidence interval*
- *Histogram*
- *Scatter plot*
- *Box plot*

First steps in R To solve the next exercises the following commands for R might help: *`<-`, `$`, `?`, `??`, `c`, `library`, `data`, `str`, `names`, `plot`, `boxplot`, `abline`, `summary`, `mean`, `quantile`, `qnorm`, `qt`, `cor`, `lm`, `with`, `within`, `t.test`, `set.seed`, `replicate`, `sample`, `subset`, `density`.*

Exercise 1.5 *Do the following tasks using R and the data from the exercise above on children's pocket money.*

- *Read the data into R assigning the names `age` and `pm` to the variables age and pocket money, respectively.*
- *Compute the descriptive statistics that you calculated for the exercise above in R.*
- *Visualize the data with a scatter plot.*
- *Do you see a problem in visualizing the two variables with a scatter plot?*

Exercise 1.6 *Do the following tasks using R and the library Ecdat which contains economic data sets. Use the data set `Schooling` on wage and education:*

- *What is the data set about? What does it contain?*

- *Give a summary statistic about the hourly wage in cents in 1976 (*`wage76`*).*

- *Draw a histogram on the hourly wage in 1976 (*`wage76`*).*

- *Draw a box plot on the years of education in 1976 (*`ed76`*).*

- *Draw a scatter plot on the hourly wage (*`wage76`*) and the years of education (*`ed76`*) both in the year 1976.*

- *Are the wage (*`wage76`*) and the years of education received (*`ed76`*) correlated? What does this result mean?*

Female labor supply

Exercise 1.7 *Do the following tasks using R and the library Ecdat. Use the data set* `Workinghours` *on female labor supply:*

- *Are the hours worked by wives (*`hours`*) related to the other income of the household (*`income`*)? Calculate the answer and illustrate it with a graph.*

- *Are the hours worked by wives (*`hours`*) related to the education they received (*`education`*)? Calculate the answer and illustrate it with a graph.*

- *Does the number of hours worked by wives (*`hours`*) who have at least one child below 6 differ compared to wives without children under 6? Illustrate your answer with a graph.*

- *Do wives who live in a home owned by the household work more hours?*

- *How many hours do wives below 26 years of age work on average? Is this significantly more than wives of age 26 and above work?*

Students' test scores

Exercise 1.8 *Do the following tasks using R and the library Ecdat. Use the data set* `Caschool` *on results of test scores in Californian schools:*

- *Is the average test score in math (*`mathscr`*) equal to 652? First, phrase your hypothesis and your alternative hypothesis, then do the computation in R.*

- *Are the average score for math and reading (*`mathscr`*) and (*`readscr`*) equal? First, phrase your hypothesis and your alternative hypothesis, then do the computation in R.*

- *Are the results of the two test scores (*`mathscr`*) and (*`readscr`*) related to each other? Compute the answer and illustrate it with a graph.*

- *Look at the outputs you obtained in R. Explain the information you have received from R.*

Sampling distributions:

Exercise 1.9

- *Have another look at* `Workinghours` *and study the sampling distribution of the mean of* `hours`*, i.e. you take a large number of samples and study the distribution of all these means.*
- *Given your (approximate) sampling distribution, how likely is it to obtain a mean value for* `hours` *smaller or equal then 1100? Compare your result with a* `t.test`
- *Now consider the difference in* `hours` *separately for wives with and without children under 5. Use your sampling distribution to calculate a 95% confidence interval. Compare this interval with the interval you get from a* `t.test`

Exercise 1.10 *The average weight of a bag of potatoes is 50 kg with a standard deviation of 2 kg. You assume the weight to follow a normal distribution. You buy 4 bags and determine the average weight* $\bar{x}$ *of these 4 bags.*

1. *How probable is it you find* $\bar{x} > 50$ *kg?*
2. *How probable is it that you find* $\bar{x} > 52$ *kg?*
3. *What is the weight* x^**, such that in 99% of all cases* $\bar{x} > x^*$.

Exercise 1.11 *A firm produces chains. The length of each link is independent of each other and normally distributed. The mean length of a link is 10. 95% of all links have a length between 9 and 11. The total length of each chain is the sum of the lengths of its links. You consider chains with 100 links*

1. *How probable is it that such a chain has a total length of more than 1000?*
2. *How probable is it that such a chain has a total length between 900 and 1100?*
3. *How probable is it that such a chain has a total length between 990 and 1010?*
4. *How probable is it that such a chain has a total length between 999 and 1001?*

Exercise 1.12 *You acquire a new machine. The producer claims that, on average, the machine can be used for 10000 hours with a standard deviation of 400 hours. You assume that the lifetime of a machine follows a normal distribution.*

1. *How probable is it that your machine lives for more than 11000 hours?*
2. *How probable is it that your machine lives between 9900 and 11000 hours?*
3. *How many machines do you have to buy to have a 99% chance that at least one of them lives for 10000 hours.*

Exercise 1.13 *You are interested in the expected value of* X. *Your sample contains nine independent observations:* $X_1, \ldots, X_9$.

1. *Which estimators for* $E(X)$ *are unbiased?*

   - X_9
   - $X_1 + X_7 - X_9$
   - $2X_1 + 2X_7 - 3X_9$
   - $X_1 - X_7 - X_9$
   - $2X_1 + 3X_7 - 2X_9$

   Which estimators for $E(X)$ *are efficient?*

   - $X_9 X_1 + X_7 - X_9$
   - $2X_1 + 2X_7 - 3X_9$
   - $\frac{1}{9}\sum_i^9 X_i$
   - $\sum_i^9 X_i$

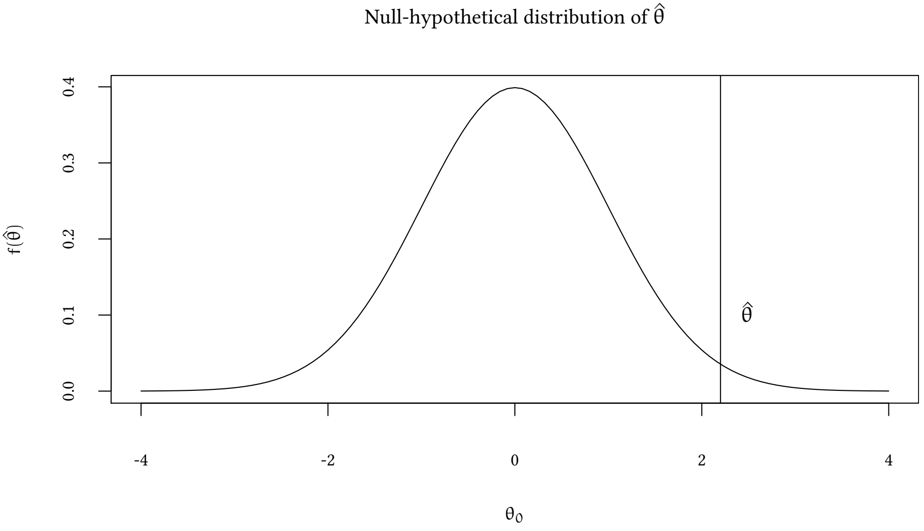
2. *Which estimators for* $E(X)$ *are more efficient?*

   - X_9 *is more efficient than* $X_1 + X_7 - X_9$
   - $X_1 + X_7 - X_9$ *is more efficient than* $2 \cdot X_9 - X_8$
   - X_1 *is more efficient than* $(X_2 + X_3)/2$
   - $\frac{1}{9}\sum_i^9 X_i$ *is more efficient than* $\frac{1}{7}\sum_i^7 X_i$
   - X_3 *is more efficient than* $\frac{1}{7}\sum_i^7 X_i$

2 Review of frequentist inference

2.1 Testing hypotheses

2.1.1 General idea

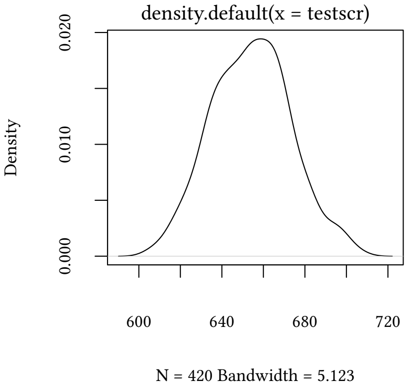


Null hypothesis: The population parameter is θ_0. We took only a (small) sample, hence we observed something else ($\hat{\theta}$) in our sample.

- Is it *significant* that we observed $\hat{\theta}$, although we hypothesized θ_0? — significance test at a given *significance level.*

- If the population parameter is, indeed, θ_0, how *probable* is it to draw a sample like ours (or a sample even more adverse to our hypothesis)? — p-*value* (marginal significance).

```
data(Caschool,package="Ecdat")
attach(Caschool)
plot(density(testscr))
```



Is 652 the mean *testscr*?

```
t.test(testscr, mu=652)


One Sample t-test

data:  testscr
t = 2.3196, df = 419, p-value = 0.02084
alternative hypothesis: true mean is not equal to 652
95 percent confidence interval:
 652.3291 655.9840
sample estimates:
mean of x
 654.1565
```

- $H_0 : E(Y) = \mu_{Y,0}$ versus $H_1 : E(Y) \neq \mu_{Y,0}$ (two-sided test)
- $H_0 : E(Y) = \mu_{Y,0}$ versus $H_1 : E(Y) > \mu_{Y,0}$ one-sided test
- $H_0 : E(Y) = \mu_{Y,0}$ versus $H_1 : E(Y) < \mu_{Y,0}$ one-sided test

	H_0 false	H_0 true
fail to reject H_0	type II error false negative	
reject H_0		type I error false positive

Level of significance of a test
Predefined probability of rejecting the null hypothesis, despite it being true.

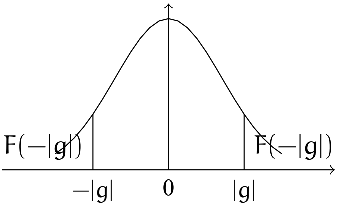
p-value of a statistic (e.g. for $\bar{Y}$)

Probability of drawing a sample $Y_1, \dots, Y_N$, that is at least as adverse to the null hypothesis as our data — given that the null hypothesis is true.

e.g. with $\bar{Y}$: p-value$= \Pr_{H_0}(|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{\text{sample}} - \mu_{Y,0}|)$
($\bar{Y}^{\text{sample}} =$ the value of $\bar{Y}$ for *our* data).

- To calculate this, we have to know the sampling distribution of $\bar{Y}$ (complicated, if n is small).
- If n is large → Central Limit Theorem → $\bar{Y}$ follows approximately a normal distribution.

$$p = \Pr_{H_0}\left(\left|\underbrace{\frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}}_{g}\right| > \left|\underbrace{\frac{\bar{Y}^{\text{sample}} - \mu_{Y,0}}{\sigma_{\bar{Y}}}}_{g^{\text{sample}}}\right|\right)$$

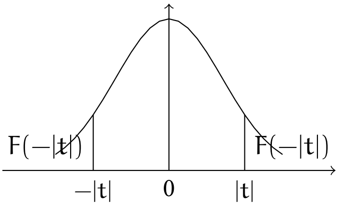


- In practice $\sigma_{\bar{Y}}$ is unknown — we can estimate $\hat{\sigma}_{\bar{Y}}$.

2.1.2 Calculating the p-value with the help of an estimated $\hat{\sigma}^2_{\bar{Y}}$

$$\begin{aligned}
p\text{-value} &= \Pr_{H_0}\left(|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{\text{sample}} - \mu_{Y,0}|\right) \\
&= \Pr_{H_0}\left(\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right| > \left|\frac{\bar{Y}^{\text{sample}} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right|\right) \\
&\approx \Pr_{H_0}\left(\left|\underbrace{\frac{\bar{Y} - \mu_{Y,0}}{\hat{\sigma}_{\bar{Y}}}}_{t}\right| > \left|\underbrace{\frac{\bar{Y}^{\text{sample}} - \mu_{Y,0}}{\hat{\sigma}_{\bar{Y}}}}_{t^{\text{sample}}}\right|\right)
\end{aligned}$$

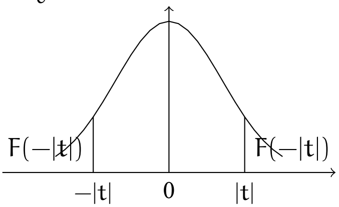
Distribution of t under H_0:



H_0 is rejected if $p < \alpha$
Testing the hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$:

- Determine the t-statistic:
$$t = \frac{\bar{Y}^{\text{sample}} - \mu_{Y,0}}{\hat{\sigma}_{\bar{Y}}}$$
- The p-value is $p = \Pr\left(|t| > \left|t^{\text{sample}}\right|\right) = 2\,F(-|t^{\text{sample}}|)$

  Distribution of t under H_0:



2.1.3 Relation between p-value and the level of significance

The level of significance is given. E.g. if the given level of significance is 5%,...

- ...the null hypothesis is rejected if $|t| > -Q(0.025) \approx 1.96$,
- ...equivalently the null hypothesis is rejected if $p < 0.05$.
- The p-value is also called *marginal level of significance.*
- In many situations we will provide more information by calculating a p-value, rather than by telling whether we rejected the null hypothesis or not.

2.1.4 What happened to the t-table and the degrees of freedom?

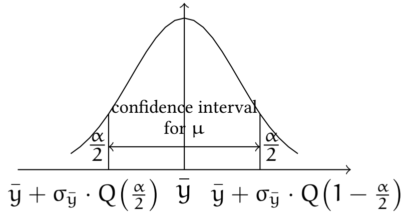
- If $Y_1, \ldots, Y_n$ are i.i.d. and *normally distributed* $Y_i \sim N(\mu_Y, \sigma_Y^2)$, the t-statistic follows the Student-t distribution with $n - 1$ degrees of freedom.
- Some critical values of the t-distribution can be found in all old statistics books. The recipe is simple:
  1. Calculate the t-statistic

2. Calculate the degrees of freedom: $n-1$
3. Look up the 2.5% critical value
4. If the t-statistic is greater (in absolute terms) than the critical value, reject the null hypothesis.

2.2 Confidence intervals

A 95% *confidence interval* for μ_Y is the interval that contains the true value of μ_Y in 95% percent of all repeated samples.

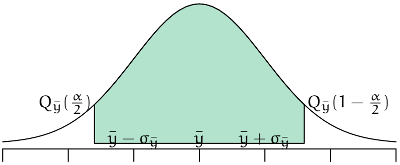
(Note: Usually we are *not* repeating samples. The above is a *very* abstract statement about the procedure, *not* about μ_Y!)

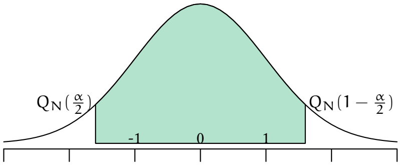

confidence interval for μ
$\frac{\alpha}{2}$
$\frac{\alpha}{2}$
$\bar{y}+\sigma_{\bar{y}}\cdot Q\left(\frac{\alpha}{2}\right)$
$\bar{y}$
$\bar{y}+\sigma_{\bar{y}}\cdot Q\left(1-\frac{\alpha}{2}\right)$

$H_0:\bar{Y}=\mu_0$ is rejected, if μ_0 is outside the confidence interval.
Required:

- A confidence level (e.g. $P=95\%=1-\alpha$)
- A sample mean $\bar{y}$
- A standard error of the sample mean $\sigma_{\bar{y}}$.
- Determine the area under the left/right part of the distribution $\frac{\alpha}{2}=\frac{1-P}{2}$
- Determine the quantile of the standard normal distribution $Q_N\left(\frac{\alpha}{2}\right)$ or $Q_N\left(1-\frac{\alpha}{2}\right)$
- If necessary, determine $\sigma_{\bar{y}}=\frac{\sigma_Y}{\sqrt{n}}$

$$\left[\bar{y}+\sigma_{\bar{y}}\cdot Q_N\left(\frac{\alpha}{2}\right),\bar{y}-\sigma_{\bar{y}}\cdot Q_N\left(\frac{\alpha}{2}\right)\right]=\left[\bar{y}-\sigma_{\bar{y}}\cdot Q_N\left(1-\frac{\alpha}{2}\right),\bar{y}+\sigma_{\bar{y}}\cdot Q_N\left(1-\frac{\alpha}{2}\right)\right]$$


$Q_{\bar{y}}(\frac{\alpha}{2})$
$Q_{\bar{y}}(1-\frac{\alpha}{2})$
$\bar{y}-\sigma_{\bar{y}}$
$\bar{y}$
$\bar{y}+\sigma_{\bar{y}}$



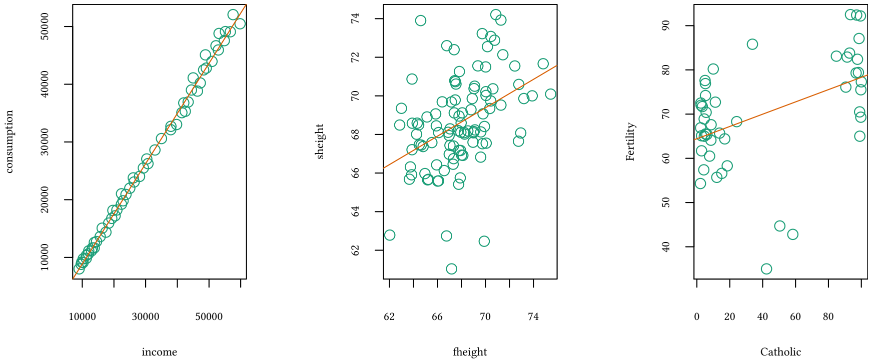
Note:

- The confidence interval is based on the *random* sample $Y_1, \ldots, Y_n$. Thus, the confidence interval is *random* itself.
- The parameter μ_Y of the population is (in the frequentist school) *not random* — but we do not know it.

```
t.test(testscr)


One Sample t-test

data:  testscr
t = 703.61, df = 419, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 652.3291 655.9840
sample estimates:
mean of x
 654.1565
```

2.3 Maximum likelihood method

Estimation of linear relationships Often a straight line is a good approximation of a relationship:



- OLS estimates straight lines, however OLS comes with several assumptions.
- What can we do when these assumptions are not satisfied? Is there a constructive way to find estimators?
- Consider a random variable $X \sim f(x|\theta)$ with unknown parameter θ.

- We draw a sample $X_1, \ldots, X_n$.
- Likelihood to draw a specific sample $\Pr(X_1, \ldots, X_n|\theta)$.

  Consider a single observation X_i:

$$\text{discrete:} \qquad \Pr(X_i = x_i|\theta) = \begin{cases} \text{Probability, that } X_i = x_i \text{ if} \\ \theta \text{ was the true parameter} \end{cases}$$

$$\text{continuous:} \qquad f(x_i|\theta) = \begin{cases} \text{Density of } X_i \text{ if} \\ \theta \text{ is the true parameter} \end{cases}$$

  For simplicity we write $f(x_i|\theta)$ in the discrete and the continuous case. We call $f(x_i|\theta)$ a “likelihood”.

Terminology: $f(x|\theta)$ is a density, not a probability. With a continuous distribution the *probability* $\Pr(X = x) = 0$.

However, the probability $\Pr(X \in [x - \epsilon, x + \epsilon])$ is proportional to $f(x|\theta)$ for small ϵ. For us it is sufficient to work with something that is only *proportional* to the probability to observe our sample $x_1, \ldots, x_n$. We call this expression *likelihood.*

Since likelihood is proportional to probability we know that, if we maximise likelihood we also maximise probability.

2.3.1 Likelihood function

What is the likelihood, for a given θ, to draw a specific x_1? (Here we use lowercase x_i to indicate that x_i has already realised.)

$$f(x_1|\theta)$$

In a next step we ask: What is the likelihood of the entire sample: What is the likelihood, for a given θ, to draw a sample $x_1, x_2, x_3, \ldots, x_n$?

$$f(x_1|\theta) \cdot f(x_2|\theta) \cdot f(x_3|\theta) \cdots f(x_n|\theta)$$

Now given a sample $x_1, x_2, x_3, \ldots, x_n$ we call

$$L(x_1, \ldots, x_n|\theta) \equiv f(x_1|\theta) \cdot f(x_2|\theta) \cdot f(x_3|\theta) \cdots f(x_n|\theta)$$

the *likelihood function.*

Two interpretations:

- θ is given: $L(x_1, \ldots, x_n|\theta)$ is the likelihood that $(x_1, \ldots, x_n)$ realise.

  Since θ is usually unknown, this is not helpful.

- $(x_1, \ldots, x_n)$ is given: $L(x_1, \ldots, x_n|\theta)$ is the likelihood, that $(x_1, \ldots, x_n)$ realise, if θ was the true parameter.

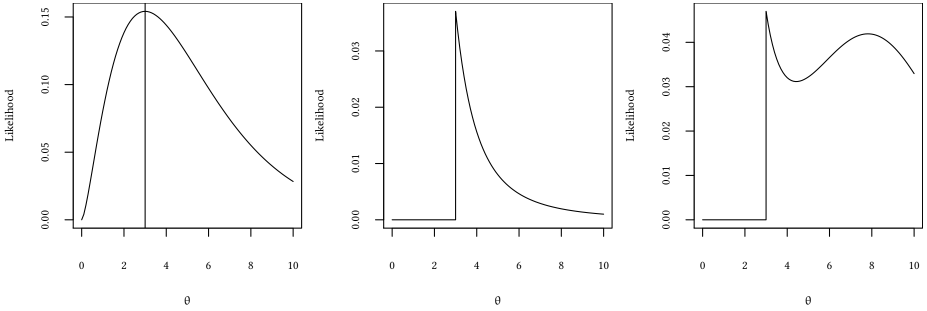
  Since we know $(x_1, \ldots, x_n)$ we can use this approach to estimate θ.

The maximum-likelihood (ML) method

- $f(X|\theta)$ conditional density (likelihood)
- Θ the set of possible parameter values for θ
- Definition: $\hat{\theta}$ is maximum-likelihood (ML) estimator iff $\forall\theta \in \Theta$:

  $$L(x_1, \dots, x_n|\hat{\theta}) \geq L(x_1, \dots, x_n|\theta)$$

  We use as *estimator* for θ the value $\hat{\theta}$, for which, given the assumed likelihood function $f(X|\theta)$, realisations $x_1, \dots, x_n$ are most probable.



2.3.2 Log-likelihood Function

So far, we maximise the likelihood function:

$$L(x_1, \dots, x_n|\theta) \equiv f(x_1|\theta) \cdots f(x_n|\theta)$$

The larger our sample, the longer this product, the harder it is to take a derivative. It is easier to maximise the log-likelihood.

$$\begin{aligned} \log L(x_1, \dots, x_n|\theta) &= \log\left(f(x_1|\theta) \cdots f(x_n|\theta)\right) = \\ & \log f(x_1|\theta) + \cdots + \log f(x_n|\theta) \end{aligned}$$

Now we only take the derivative of a sum. This is often easier.

$$\frac{d}{d\theta} \log L(x_1, \dots, x_n|\theta) = \frac{f'(x_1|\theta)}{f(x_1|\theta)} + \cdots + \frac{f'(x_n|\theta)}{f(x_n|\theta)}$$

Using maximum likelihood in practice Whenever the assumptions of simpler models (OLS) are not fulfilled, e.g.:

- Discrete choice models

  - Logistic regresssion (unemployment, insolvency,...)
  - Count data (number of patents per firm, affairs per person, children per household,...)
  - Interval data (income relevant for social insurance is censored from above, bids in auctions,...)
  - $\vdots$

- Random effects models
  - e.g. repeated measurements of the same unit
- Quantile regression
  - Outliers

An example Here we use ML to find the mean of a small sample:

$$X \sim N(\mu, \sigma^2)$$

$$\Pr(\mathbf{X}|\mu, \sigma) = \prod \text{dnorm}(X_i|\mu, \sigma)$$

$$LL = \sum \log\left(\text{dnorm}(X_i|\mu, \sigma)\right)$$

```
X <- c(1,2,3)
LL <- function(theta) -sum(log(dnorm(X,mean=theta[1],sd=theta[2])))
optim(c(0,1),LL)

$par
[1] 1.9999124 0.8164934

$value
[1] 3.648618

$counts
function gradient
      77       NA

$convergence
[1] 0

$message
NULL
```

Use the sample moments instead:

```
mean(X)

[1] 2
```

```
sd(X)

[1] 1
```

Note: The ML estimator yields (almost) *the same* result for μ, but a different result for σ.

Properties of the ML estimator (given some conditions):

- consistency ($\hat{\theta}_{\text{ML}} \xrightarrow{p} \theta$)
- asymptotic normality ($\hat{\theta}_{\text{ML}} \xrightarrow{p} N(\theta, \frac{1}{n} I^{-1})$)

An alternative method to estimate θ is the method of moments:

2.4 Method of moments

Reminder: Moments

$$\begin{aligned} n\text{-th } \textit{non}\text{-central moment of } X: \quad & \mu'_n = E(X^n) \\ n\text{-th central moment of } X: \quad & \mu_n = E\left((X - \mu'_1)^n\right) \end{aligned}$$

We already know the first four moments as properties of a distribution (just as an illustration):

mean:	$\mu'_1 = E(X)$
variance:	$\mu_2 = E\left((X - E(X))^2\right)$
skewness:	$\frac{\mu_3}{(\mu_2)^{(3/2)}}$
kurtosis:	$\frac{\mu_4}{(\mu_2)^2}$

We distinguish:

Moments of a population (unknown): μ_1, μ_2, μ_3, ...

Moments of a sample (known): $m_1 = \bar{Y}$, m_2, m_3, ...

Population has a density $f(X|\theta)$ with unknown parameter θ:

- $f(X|\theta)$ has moments $\mu_1(\theta)$, $\mu_2(\theta)$, $\mu_3(\theta)$,...
- Method of moments:
  - Determine the sample moments m_1, m_2, m_3...(we usually start with the mean m_1)
  - Equate sample moments with population moments...

    $$\mu_i(\hat{\theta}) = m_i$$

    ...and solve for $\hat{\theta}$.

2.5 Literature

- Greene. "Econometric Analysis". Chapter 12-14.
- Stock, Watson. "Introduction to Econometrics". Chapter 11.

Appendix 2.A Examples for the lecture

Example 1: In your sample with 4 observations of a normally distributed random variable with unknown variance you determine a mean of 10 and a variance of 100. Your confidence level is 95%. How do you determine the width of a confidence interval for the mean?

- None of the following.
- `qt(10,4)*0.95`
- `qnorm(10)*5`
- `qt(.95,3)*10`
- `qnorm(.95)*10`

Example 2: In your sample with 25 observations of a normally distributed random variable with known variance 25 you determine a mean of 10. Your confidence level is 95%. How do you determine a lower boundary of the confidence interval for the mean?

- None of the following.
- `10 + qnorm(0.05)`
- `10 + 5 * qnorm(0.05)`
- `10 + 5 * qnorm(0.025)`
- `10 - qnorm(0.975)`

Example 3: A random variable is distributed as follows: $\Pr(X=-1)=\theta$, $\Pr(X=1)=\theta$, $\Pr(X=0)=1-2\theta$. Your sample is $X=\{-1,0\}$:

What is the ML estimator for θ?

Example 4: The random variable X follows a distribution $\mathcal{X}_\theta$ with expected value $E(X)=1/\theta$ and variance θ^2. We have $\theta>0$. The variable x contains your sample. Use R to calculate the method of moments estimator for θ based on the second moment.

Example 5: A random variable is distributed as follows: $\Pr(X = 1) = \theta$, $\Pr(X = 2) = \theta$, $\Pr(X = 3) = 1 - 2\theta$ where $\theta \in [0, 1/2]$. You sample is $X = \{1, 2, 1, 1, 1, 1\}$.

What is the ML estimator for θ?

Example 6: A random variable follows a uniform distribution over $[a, b]$. Your sample is $X = \{1, 2, 3\}$.

1. What is the ML estimator for a and for b.
2. You sample is now $X = \{1, 3, 5\}$. You also know that $b - a = 10$. What is the method of moment estimator based on the first moment.

Example 7: X is uniformly distributed over $[a, b]$. We are looking for b. The first moment of the uniform distribution is $E(X) = (a + b)/2$. The variance of the uniform distribution is $\text{var}(X) = (b-a)^2/12$ The sample contains nine observations: $X_1, \ldots, X_9$.

1. Which estimator for b is unbiased?
   - X_9
   - $X_1 + X_7 - a$
   - $2X_1 + 2X_7 - 2X_9 - a$
   - $X_1 - X_7 - X_9 + 2 \cdot a$
   - $a + X_1 + X_7 - X_9$
2. Which estimator for b is more efficient?
   - $X_2 - a + X_3$ is more efficient than $\frac{1}{2}(4X_1 - 2a)$
   - $X_2 - a + X_3$ is more efficient than $2 \cdot X_1 + X_7 - X_6 - a$
   - $\frac{1}{2}(4X_1 - 2a)$ is more efficient than $2 \cdot X_1 + X_7 - X_6 - a$
   - $2 \cdot X_1 - a$ is more efficient than $2 \cdot X_1 + X_7 - X_6 - a$
   - $-a + \frac{2}{9} \sum_i^9 X_i$ is more efficient than $2 \cdot X_1 + X_7 - X_6 - a$

Appendix 2.B Exercises

Exercise 2.1 *Your sample of 87 observations has a mean of 150. You know that the estimated mean has a standard error of 10. What is a 99% confidence interval for the mean? What is a 95% confidence interval for the mean?*

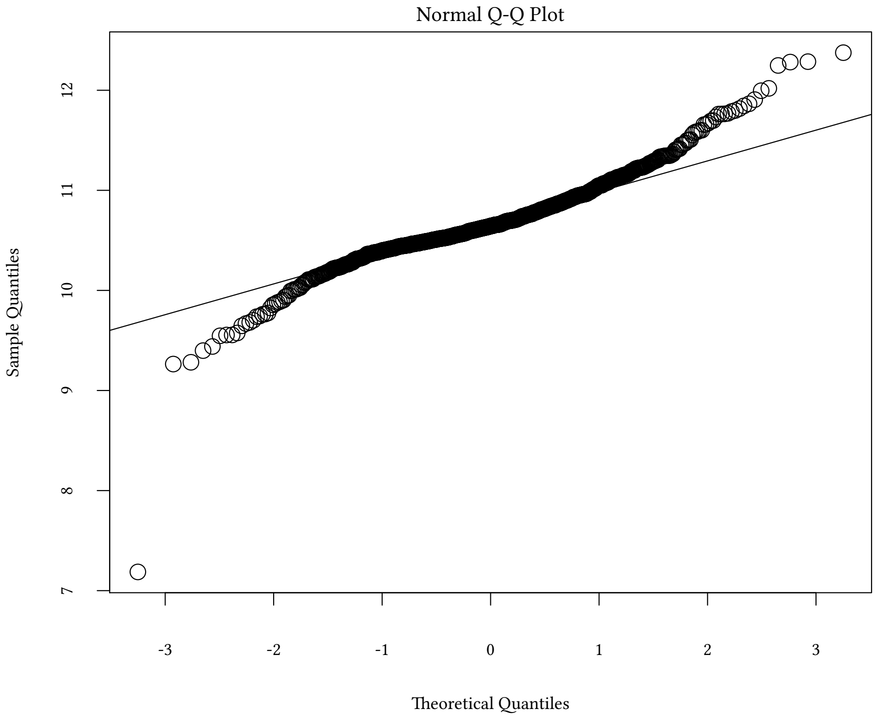
Exercise 2.2 *Your sample with 100 observations has a mean of 200. You know that each observations has a standard deviation of 10. What is a 99% confidence interval for the mean? What is a 95% confidence interval for the mean?*

Exercise 2.3 *Your sample is* $X = \{1, 2, 3\}$. *Assume that* X *follows a normal distribution. Use the* t*-distribution to find a 95% confidence interval for the mean. Determine a 99% confidence interval.*

Exercise 2.4 *Consider the data set* `Participation` *from* `Ecdat`. *Use the* t*-distribution to determine a confidence interval for the average log of non-labour income* `lnlinc`. *Is it reasonable to assume this distribution.*

The t-test assumes that our variable X follows a normal distribution. How can we test this assumption. A Q-Q plot allows us to compare a sample distribution with a theoretical distribution (here the normal distribution). The horizontal axis shows theoretical quantiles, the vertical axis shows sample quantiles. If the sample distribution is only a linear transformation of the standard normal distribution then all sample observations are on one straight line..

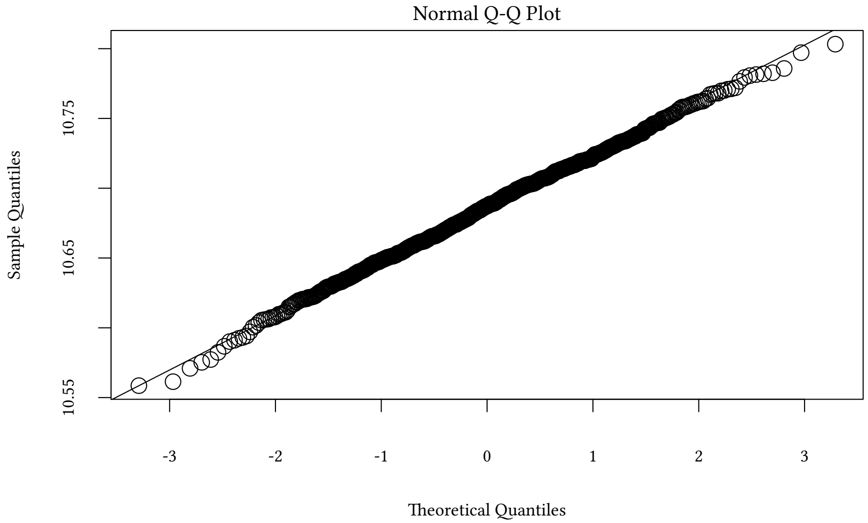
```
qqnorm(lnnlinc)
qqline(lnnlinc)
```



We see that our sample distribution does not look like a normal distribution. The assumptions of a t-test are seemingly not satisfied.

Things look better for means. Here is a plot of 1000 means of samples from this distribution:

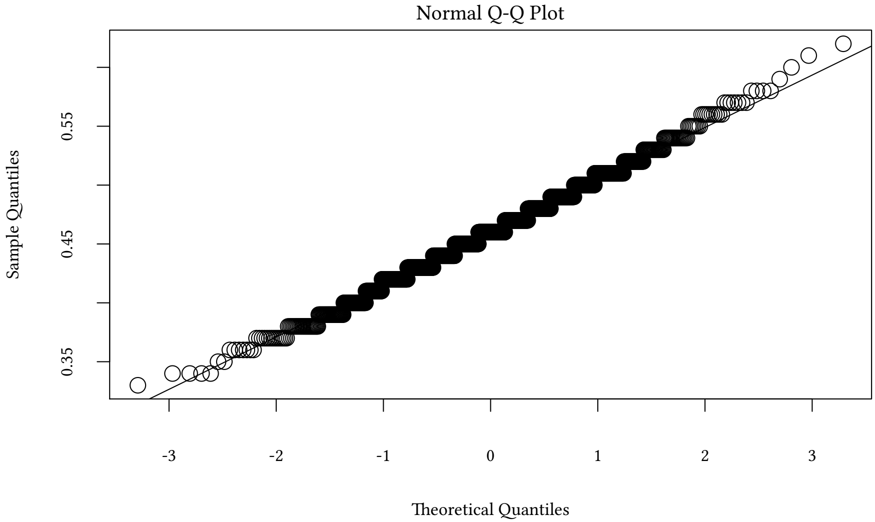
```
xx<-replicate(1000,mean(sample(lnnlinc,100)))
qqnorm(xx)
qqline(xx)
```



Exercise 2.5 *Consider the data set* `Participation` *from* `Ecdat`*. Use the* `t`*-distribution to determine a confidence interval for the mean of the average participation in the labour force* `lfp`*. Is it reasonable to assume this distribution?*

Here is again a plot for means:

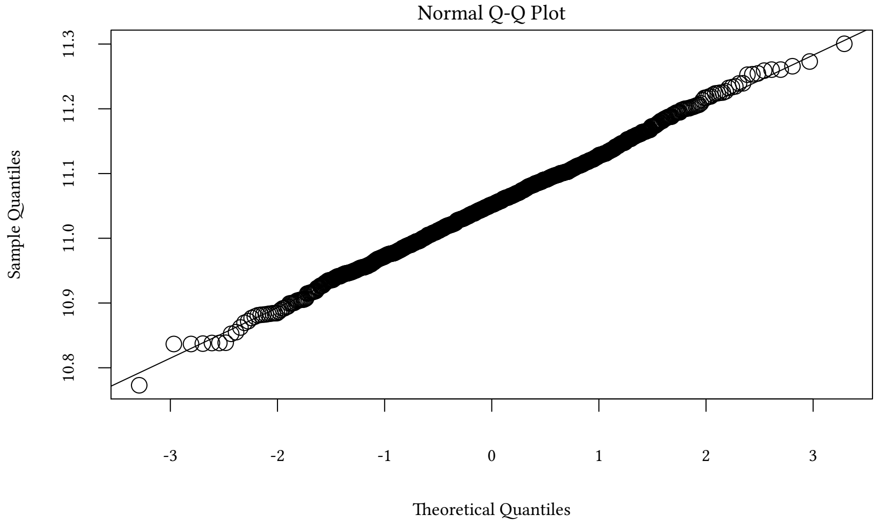
```
xx<-replicate(1000,mean(sample(lfp,100)=="yes"))
qqnorm(xx)
qqline(xx)
```



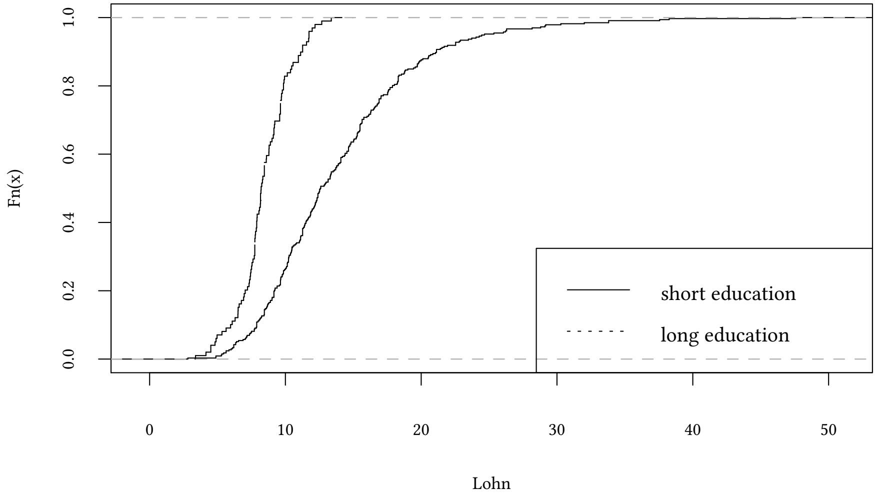
Exercise 2.6 *Consider the data set* `Bwages` *from* `Ecdat`*. Use a* `t`*-distribution to determine a confidence interval for the mean of the gross hourly wage* `wage`*. Compare now a confidence interval for workers with lower education* `educ==1` *and with higher education* `educ==5`*. Use a graph to compare the distribution of wages.*

Here is again a plot of means:

```
xx<-replicate(1000,mean(sample(wage,1000)))
qqnorm(xx)
qqline(xx)
```



```
plot(ecdf(wage[educ==5]),do.points=FALSE,verticals=TRUE,xlab="Lohn",main="")
plot(ecdf(wage[educ==1]),do.points=FALSE,verticals=TRUE,lty='dashed',add=TRUE)
legend("bottomright",c("short education","long education"),lty=c(1,3))
```



Exercise 2.7 *The content of a carton of milk follows a normal distribution. You sample 10 cartons and obtain the following values (in ml):* $(999, 1000, 997, 1005, 1001, 1000, 998, 999, 1000, 1000)$

1. *Use R to find a 99%-confidence interval for the average content.*
2. *Which of the following statements is correct?*
   a) *With a probability of 99% the interval contains the true value.*
   b) *If the procedure is used over and over again, then in 99% of all cases the interval will contain the true value.*
   c) *The interval is $[995, 6204; 1004, 3796]$.*
   d) *The 95%-confidence interval is larger.*
   e) *The 95%-confidence interval is smaller.*

Exercise 2.8 *The length of widgets follows a normal distribution with a variance of 100 and unknown mean. A sample of 100 widgets yields an average length of 2000. How do you find the lower boundary of a 95% confidence interval for the mean?*

Exercise 2.9 *The length of widgets follows a normal distribution with standard deviation 100 and unknown mean. A sample of 25 widgets yields an average length of 2000.*

1. *Determine the upper boundary of a 99% confidence interval for the mean.*
2. *How do you determine the lower boundary of a 90% confidence interval for the mean?*

Exercise 2.10 *The lifetime of batteries follows a normal distribution with a standard deviation of 30 hours. 100 randomly sampled batteries have a total lifetime of 6935.75 hours.*

1. *What is the 90% confidence interval for the average lifetime? Use the following quantiles:*

	0.001	0.0025	0.005	0.01	0.025	0.05	0.1
$Q^N(x)$	-3.090	-2.807	-2.576	-2.326	-1.960	-1.645	-1.282
$Q^t_{19}(x)$	-3.579	-3.174	-2.861	-2.539	-2.093	-1.729	-1.328
$Q^t_{20}(x)$	-3.552	-3.153	-2.845	-2.528	-2.086	-1.725	-1.325
$Q^t_{21}(x)$	-3.527	-3.135	-2.831	-2.518	-2.080	-1.721	-1.323

2. *How large should your sample be such that the length of the confidence interval is smaller or equal than 5 hours?*
3. *Your sample of size 20 contains the following observations: (60.5, 80, 71, 73.7, 65, 68, 64.4, 62.9, 74, 78, 72.9, 74, 67.5, 72.8, 61.9, 71, 58, 61, 72.8, 73). Use the sample standard deviation to determine your 90% confidence interval.*

Exercise 2.11 $X \sim D(0,a)$ *is distributed as follows*

$$f(X,a) = \begin{cases} \frac{2}{a} - \frac{2x}{a^2} & \textit{if } x \in [0,a] \\ 0 & \textit{otherwise} \end{cases}$$

You sample is $X = \{0,2\}$. *What is the ML-estimator for* a*?*

Exercise 2.12 *A random number follows the distribution Pr*$(X = 1) = \theta$, *Pr*$(X = 2) = 1 - \theta - \theta^2$, *Pr*$(X = 3) = \theta^2$ *with* $0 \leq \theta \leq \frac{6}{10}$. *Your sample is* $X = \{1,1,2,3\}$. *Find the ML estimator for* θ*!*

Exercise 2.13 $X \sim B(1,p)$ *follows a binomial distribution with sample size 1 and unknown* p. *Three successive samples yield* $X = \{0,1,1\}$. *Estimate* p *using the method of moments. The mean of the binomial distribution is* $n \cdot p$.

Exercise 2.14 *Consider a binomially distributed random variable with* $X \sim B(100,\theta)$. *Your sample is* $X = \{20,35,14\}$.

1. *Use R to estimate* θ *with the method of moments.*
2. *What is your estimate for* θ *if your sample is* $X = \{1,19,4\}$*?*

Exercise 2.15 $X \sim D(0,a)$ *has the following distribution*

$$f(X,a) = \begin{cases} \frac{2}{a} - \frac{2x}{a^2} & \textit{if } x \in [0,a] \\ 0 & \textit{otherwise} \end{cases}$$

Your sample is $X = \{0,2\}$. *Use the method of moments to get an estimate for* a. *The first moment of* D *is* $\mu = \frac{a}{3}$. *The second moment of* D *is* $\sigma^2 = \frac{1}{18}a^2$.

Exercise 2.16 $X \sim N(\mu,\sigma^2)$ *follows a normal distribution with unknown* μ *and* σ. *Your sample is* $X = \{0,1,1\}$. *Estimate* μ *using the method of moments.*

Exercise 2.17 *The density of the random variable* X *is given by*

$$f(x) = \begin{cases} 1 - \dfrac{x-\theta}{2} & \textit{if } x \in [\theta,\theta+2] \\ 0 & \textit{otherwise} \end{cases}$$

We have $E(X) = \frac{2}{3} + \theta$. *The parameter* θ *is unknown.*

1. *What is the ML estimator for* θ *if your sample is* $\{2,3\}$.
2. *What is the method of moment estimator for* θ *(based on the first moment) if your sample is* $X = \{2, 3\frac{1}{3}\}$.

Exercise 2.18 *The distribution function of a discrete random variable is*

$$Pr(X = x) = \begin{cases} \frac{1}{4} & \textit{if } x = \theta \\ \frac{1}{2} & \textit{if } x = \theta + 1 \\ \frac{1}{4} & \textit{if } x = \theta + 2 \\ 0 & \textit{otherwise} \end{cases}$$

1. *What is the ML estimator for θ, if your sample is $X = \{2, 2, 3\}$.*

2. *Now use the method of moments (based on the first moment).*

Exercise 2.19 *The distribution of X is given by $Pr(X = 1) = \theta$, $Pr(X = 3) = \theta$, $Pr(X = 5) = 1 - 2\theta$. We have $0 \leq \theta \leq \frac{1}{2}$.*

1. *Your sample is $X = \{1, 3\}$. What is the ML estimator for θ?*

2. *Your sample is $X = \{1, 3, 5\}$. Use the method of moments based on the first moment to obtain an estimator for θ?*

3 Review of Linear Regression

- draw a line through two-dimensional data Y and X

- estimate a causal dependence between Y and X

Lines have a *slope* and an *axis intercept.*
⟶ Estimation, hypothesis test, confidence interval

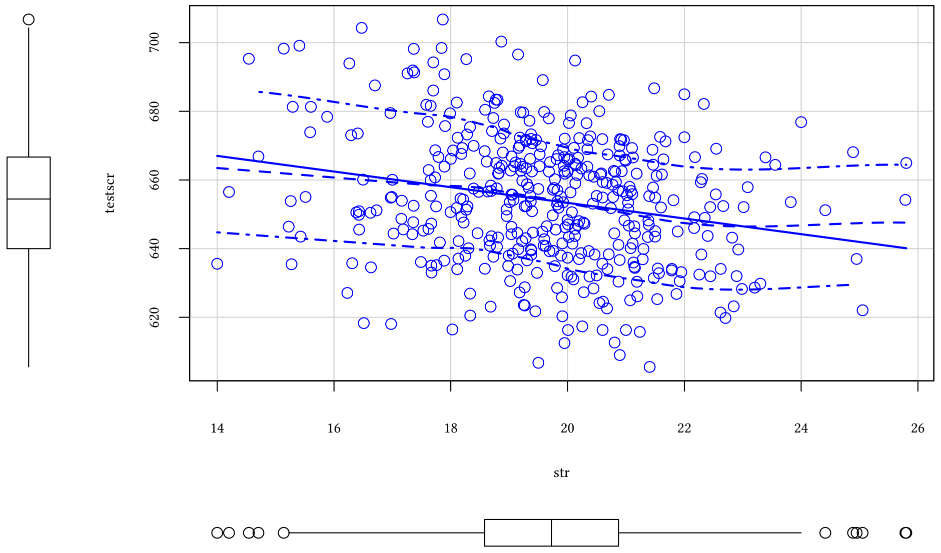
$$\texttt{testscr} = \beta_1 \texttt{str} + \beta_0$$

- β_0 and β_1 are parameters of the *population*

- we do not not know them — hence, we have to estimate them (like μ)

In the following diagram such a line is drawn through a scatterplot

```r
data(Caschool)
attach(Caschool)
scatterplot(testscr~str)
```



$$Y_i = \beta_1 X_i + \beta_0 + u_i \qquad i = 1, \ldots, n$$

- Y dependent variable
- X independent variable
- β_1 slope
- β_0 axis intercept
- u error term (other factors that impact Y)

How can we estimate β_0 and β_1?

- OLS (ordinary least squares)
- ML (maximum likelihood)
- Bayesian inference

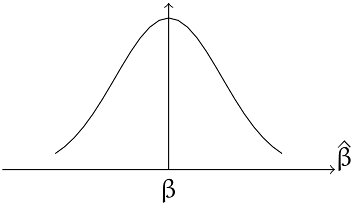
3.1 OLS assumptions

1. $E(u_i|X_i = x) = 0$ (strict exogeneity)
2. (X_i, Y_i) are i.i.d.
3. Large outliers in X and Y are rare (the fourth moments of X and Y exist)
4. $\text{var}(u|X = x)$ ist constant, u is homoscedastic
5. u is normally distributed $u \sim N(0, \sigma^2)$

Assumptions 4 and 5 are more restrictive.

- $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are calculated by means of the sample. A different sample results in different values for $\widehat{\beta}_0$ and $\widehat{\beta}_1$.
- The sample is random.

→ $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are random variables (they have a distribution, similar to $\bar{Y}$).



  - Is $E(\widehat{\beta}_1) = \beta_1$? (OLS is unbiased)
  - Is $\text{var}(\widehat{\beta}_1)$ small?
  - How do we test hypotheses? (e.g. $\beta_1 = 0$)
  - How do we calculate a confidence interval for β_0 and β_1?

Unbiasedness
Assuming 1-3, $E(\widehat{\beta}) = \beta$

Asymptotic consistency
Assuming 1-3, $\widehat{\beta} \xrightarrow{p} N(\beta, \Sigma_{\widehat{\beta}})$

Gauss Markov
Assuming 1-4, $\widehat{\beta}_1$ has the smallest variance of *all linear estimators* (of all estimators which are linear functions of Y).

Efficiency of OLS-II
Assuming 1-5, $\widehat{\beta}_1$ has the smallest variance of *all consistent estimators*, if $n \to \infty$ (regardless of whether the estimators are linear or non-linear)

3.2 Hypotheses tests and confidence intervals

- two-sided test:

  $H_0 : \beta_1 = \beta_{1,0}$ versus $H_1 : \beta_1 \neq \beta_{1,0}$.

- one-sided test:

  $H_0 : \beta_1 = \beta_{1,0}$ versus $H_1 : \beta_1 > \beta_{1,0}$

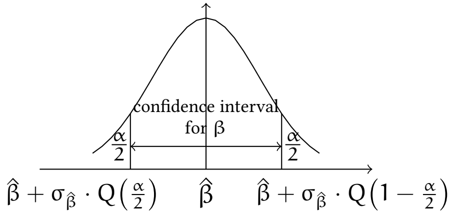
  $H_0 : \beta_1 = \beta_{1,0}$ versus $H_1 : \beta_1 < \beta_{1,0}$

with $\beta_{1,0}$ being the hypothetical value of the null hypothesis.

Approach (provided the sample is large): Build t-statistic und determine the p-value (or compare the t-statistic with the critical $N(0,1)$ value .

$$t = \frac{\hat\beta_1 - \beta_{1,0}}{\sigma_{\hat\beta_1}}$$

A 95% *confidence interval* for β is the interval that contains the true value of β in 95% percent of all repeated samples.



confidence interval
for β
$\frac{\alpha}{2}$
$\frac{\alpha}{2}$
$\hat\beta + \sigma_{\hat\beta} \cdot Q\left(\frac{\alpha}{2}\right)$
$\hat\beta$
$\hat\beta + \sigma_{\hat\beta} \cdot Q\left(1 - \frac{\alpha}{2}\right)$

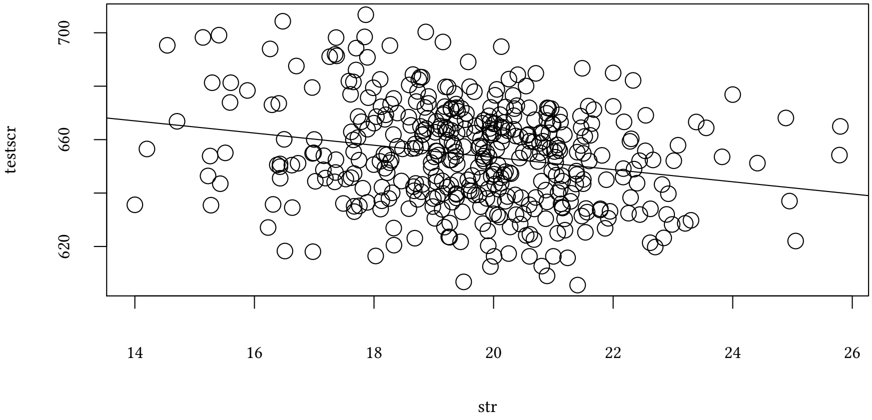
$H_0 : \hat\beta = \beta_0$ is rejected, if β_0 is outside the confidence interval.

3.3 Continuous and nominal variables

- Continuous:
  - gross domestic product
  - income in Euro
  - *str*
- Nominal / discrete
  - sex
  - profession
  - sector of a firm
  - income in categories
- Binary variable / dummy-variables are a special case of nominal variables

- sex male/female
- income higher than 40 000 Euro per year Yes/No
- unemployed Yes/No
- university degree Yes/No

```
plot(testscr ~ str)
abline(lm(testscr ~ str))
```



```
lm(testscr ~ str)


Call:
lm(formula = testscr ~ str)

Coefficients:
(Intercept)          str
     698.93        -2.28
```

$$\texttt{testscr} = 698.93 + -2.28\,\texttt{str} + u$$

In

$$\texttt{testsrc} = \beta_0 + \beta_1 \texttt{str} + u$$

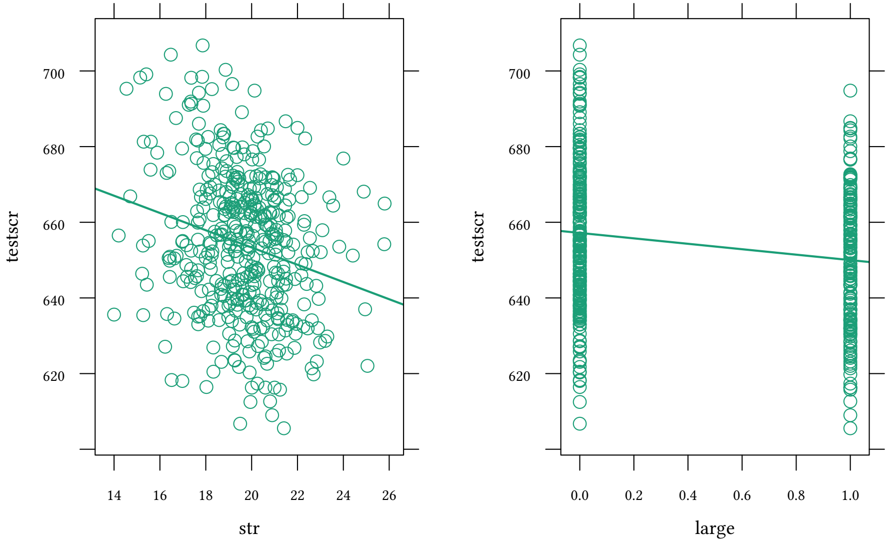
we used a continous independent variable *str*. But what if we only had binary data for *str*?

$$\mathtt{large} = \begin{cases} 1 & \text{if } \mathtt{str} > 20 \\ 0 & \text{else} \end{cases}$$

Now estimate

$$\mathtt{testsrc} = \beta_0 + \beta_1 \mathtt{large} + u$$

```
attach(Caschool)
large <- str>20
est<-lm(testscr ~ large)
```



In general (when X is a binary / dummy-variable)

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Interpretation:

If $X_i = 0$: $Y_i = \beta_0 + u_i$

The mean value $\bar{Y} = \beta_0$

$E(Y_i | X_i = 0) = \beta_0$

If $X_i = 1$: $Y_i = \beta_0 + \beta_1 + u_i$

The mean value $\bar{Y} = \beta_0 + \beta_1$

$E(Y_i | X_i = 1) = \beta_0 + \beta_1$

$\beta_1 = E(Y_i|X_i = 1) - E(Y_i|X_i = 0)$ is the difference between the mean values of the two groups of the population.

```
est<-lm(testscr ~ large)
summary(est)


Call:
lm(formula = testscr ~ large)

Residuals:
    Min      1Q  Median      3Q     Max 
-50.435 -14.071  -0.285  12.778  49.565 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  657.185      1.202  546.62  < 2e-16 ***
largeTRUE     -7.185      1.852   -3.88 0.000121 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.74 on 418 degrees of freedom
Multiple R-squared:  0.03476,Adjusted R-squared:  0.03245 
F-statistic: 15.05 on 1 and 418 DF,  p-value: 0.0001215
```

```
confint(est)


                2.5 %     97.5 %
(Intercept) 654.82130 659.547833
largeTRUE   -10.82554  -3.544715

t.test(testscr ~ large)


Welch Two Sample t-test

data:  testscr by large
t = 3.9231, df = 393.72, p-value = 0.0001031
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  3.584445 10.785813
sample estimates:
mean in group FALSE  mean in group TRUE 
           657.1846            649.9994 
```

- It does not matter (much), whether we use a Student t-test to compare the mean values of groups,
- or whether we calculate a regression with a single binary variable.

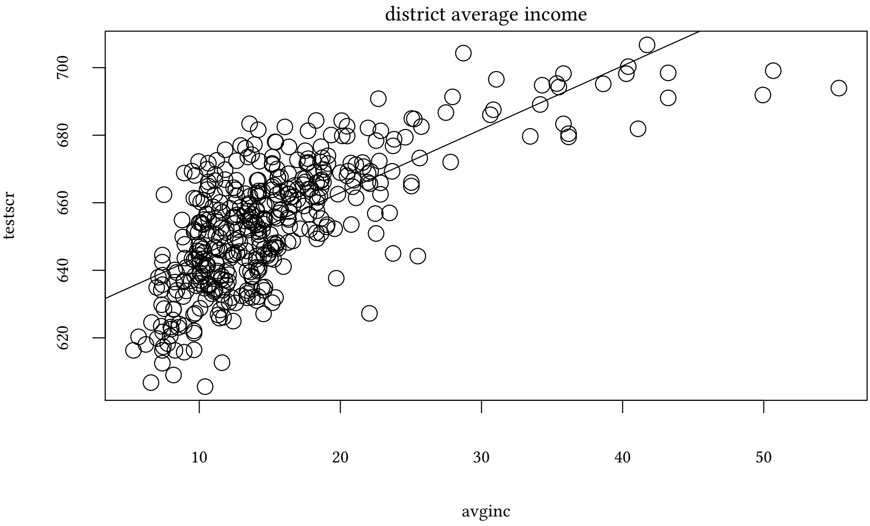
A regression can be useful if we want to control for additional regressors.

3.4 Non-linear regression functions

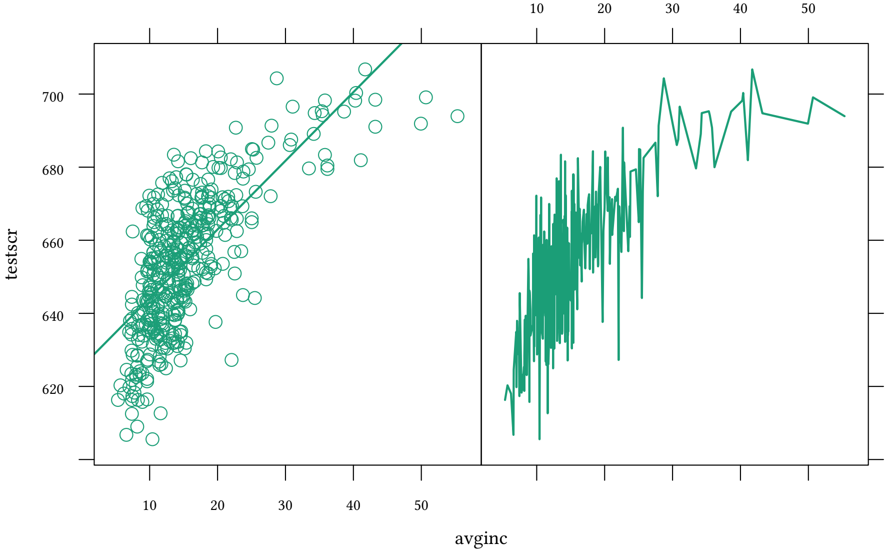
```
data(Caschool)
attach(Caschool)
est1 <- lm(testscr ~ avginc)
summary(est1)
```

```
Call:
lm(formula = testscr ~ avginc)

Residuals:
    Min      1Q  Median      3Q     Max 
-39.574  -8.803   0.603   9.032  32.530 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 625.3836     1.5324  408.11   <2e-16 ***
avginc        1.8785     0.0905   20.76   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.39 on 418 degrees of freedom
Multiple R-squared:  0.5076,Adjusted R-squared:  0.5064 
F-statistic: 430.8 on 1 and 418 DF,  p-value: < 2.2e-16
```

```
plot(testscr ~ avginc,main="district average income")
abline(est1)
```



```
z<-xyplot(testscr ~ avginc,data=Caschool,type=c("p","r"))
c(z,update(z,type='a'))
```

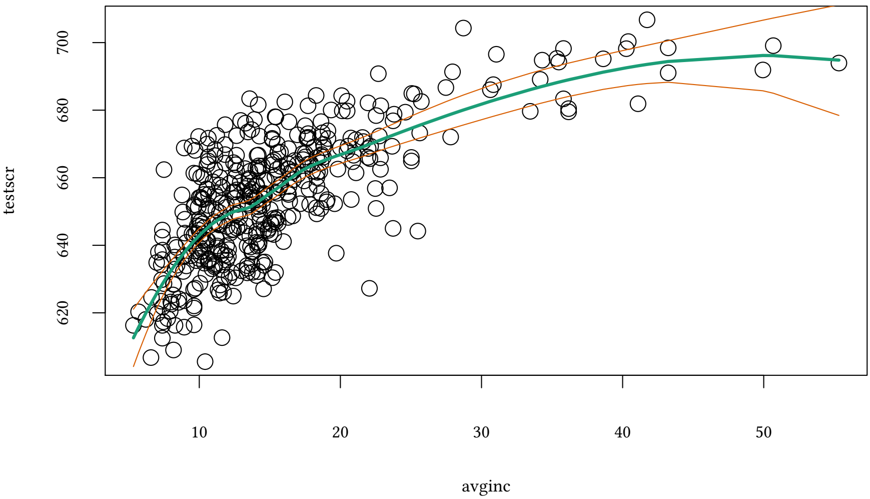


Clearly, the above relationship allows for too much flexibility. We have to restrict flexibility somehow:

Here we have two options: Either we specify a precise functional form for the relation between *avginc* and *testscr* or we leave the precise form open, requiring only a "smooth" relationship, like the following:

A semiparametric fit of the relation between *avginc* and *testscr*:

```
lo <- loess(testscr ~ avginc,data=Caschool[order(avginc),])
plot(lo,xlab="avginc",ylab="testscr")
spline.lo<-predict(lo,se=TRUE)
lines(lo$x,lo$fitted,lwd=3,col=1)
with(spline.lo,{
    lines(lo[["x"]],fit+qnorm(.025)*se.fit,col=2)
    lines(lo[["x"]],fit+qnorm(.975)*se.fit,col=2)
})
```



- Semiparametric
  - does not specify an (economic) model,
  - does not restrict the shape of the fitted functions (except imposing “smoothness”).

  + Great to get an idea of the general shape.

  − Difficult to interpret.
- Parametric
  - specifies an (economic) model.
  - May misrepresent the data (since it is restrictive).

  + Easier to interpret.

  (but also easy to be misled)

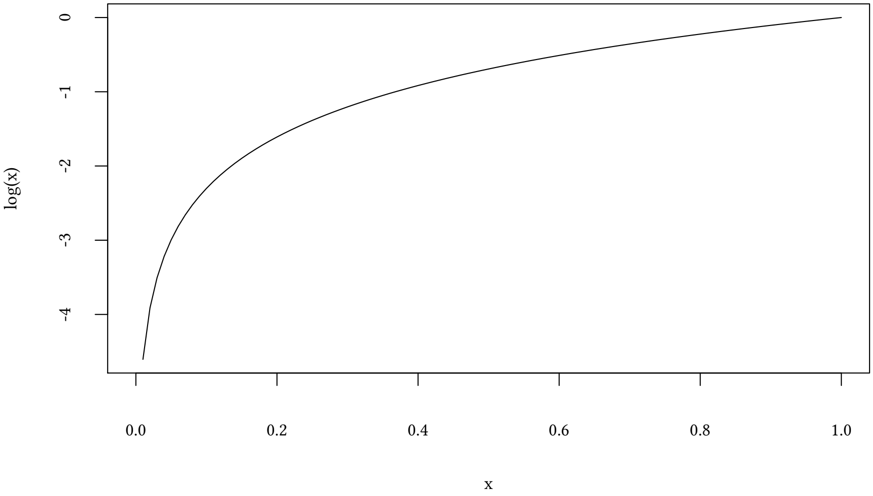
  Non-linear functions of a single independent variable

  ∗ Polynomials in X

  ∗ Logarithmic transformation

3.4.1 Logarithmic Models

```
curve(log(x))
```



- $Y_i = \beta_0 + \beta_1 \cdot \log X_i + u_i$ linear-log
- $\log Y_i = \beta_0 + \beta_1 \cdot X_i + u_i$ log-linear
- $\log Y_i = \beta_0 + \beta_1 \cdot \log X_i + u_i$ log-log

A minor technicality: In the following we draw a number of graphs. Since it is easier to draw a line of a function if the `avginc`-values are ordered, we calculate an ordering of `avginc`:

```
or <- order(avginc)
```

3.4.2 Logarithmic Models: linear-log

$$Y_i = \beta_0 + \beta_1 \cdot \log X_i + u_i$$

marginal effects:

$$\frac{\partial Y_i}{\partial X_i} = \beta_1 \frac{1}{X_i}$$

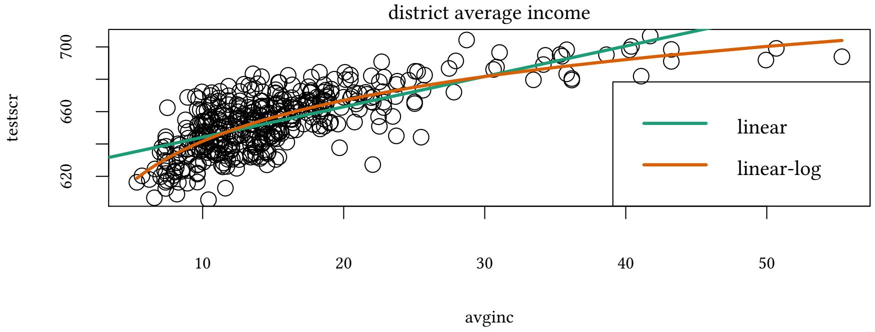
$$\Delta Y_i \approx \frac{\Delta X_i}{X_i} \cdot \beta_1$$

if X_i changes by 1% ($\Delta X_i = 0.01 \cdot X_i$) ...

$$\Delta Y_i \approx \frac{0.01 X_i}{X_i} \cdot \beta_1$$

...Y_i changes by $0.01 \cdot \beta_1$

```
(estL <- lm(testscr ~ log(avginc)))


Call:
lm(formula = testscr ~ log(avginc))

Coefficients:
(Intercept)  log(avginc)
     557.83        36.42

plot(testscr ~ avginc,main="district average income")
abline(est1,col=1,lwd=3)
lines(avginc[or],fitted(estL)[or],col=2,lwd=3)
legend("bottomright",c("linear","linear-log"),lwd=3,col=c(1,2))
```



```
X<-c(10,40,60)
(maL<-coef(estL)[2]/X)

[1] 3.6419683 0.9104921 0.6069947
```

3.4.3 Logarithmic Models: log-linear

$$\log Y_i = \beta_0 + \beta_1 \cdot X_i + u_i \quad \rightarrow \quad Y_i = e^{\beta_0 + \beta_1 \cdot X_i + u_i}$$

marginal effects: $\downarrow$

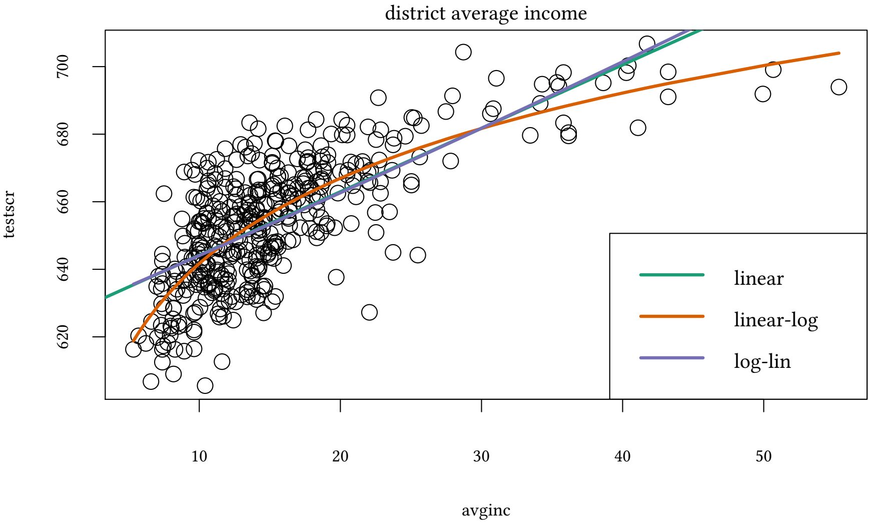
$$\frac{\partial \log Y_i}{\partial X_i} = \beta_1 \qquad\qquad \frac{\partial Y_i}{\partial X_i} = \beta_1 \cdot e^{\beta_0 + \beta_1 \cdot X_i}$$

$$\frac{\Delta \log Y_i}{\Delta X_i} \approx \beta_1 \quad \rightarrow \quad \Delta \log Y_i \approx \beta_1 \Delta X_i$$

$$\frac{1}{Y_i} \approx \frac{\Delta \log Y_i}{\Delta Y_i} \rightarrow \frac{\Delta Y_i}{Y_i} \approx \Delta \log Y_i \quad \rightarrow \quad \frac{\Delta Y_i}{Y_i} \approx \Delta \log Y_i \approx \beta_1 \cdot \Delta X_i$$

A change of X_i by one unit translates into a relative change of Y_i by the share β_1

```
(estLL <- lm(log(testscr) ~ avginc))


Call:
lm(formula = log(testscr) ~ avginc)

Coefficients:
(Intercept)       avginc
   6.439362     0.002844

plot(testscr ~ avginc,main="district average income")
abline(est1,col=1,lwd=3)
lines(avginc[or],fitted(estL)[or],col=2,lwd=3)
lines(avginc[or],exp(fitted(estLL))[or],col=3,lwd=3)
legend("bottomright",c("linear","linear-log","log-lin"),lwd=3,col=c(1,2,3))
```



```
coef(estLL)[2]

    avginc
0.00284407
```

$$\frac{\partial Y_i}{\partial X_i} = \beta_1 \cdot e^{\beta_0 + \beta_1 \cdot X_i}$$

```
X<-c(10,40,60)
(maLL<-coef(estLL)[2]*exp(predict(estLL,
                   newdata=list(avginc=X))))

       1        2        3
1.831772 1.994924 2.111688
```

$$\log Y_i = \beta_0 + \beta_1 \cdot \log X_i + u_i \quad \rightarrow \quad Y_i = e^{\beta_0} \cdot X_i^{\beta_1} \cdot e^{u_i}$$

marginal effect:

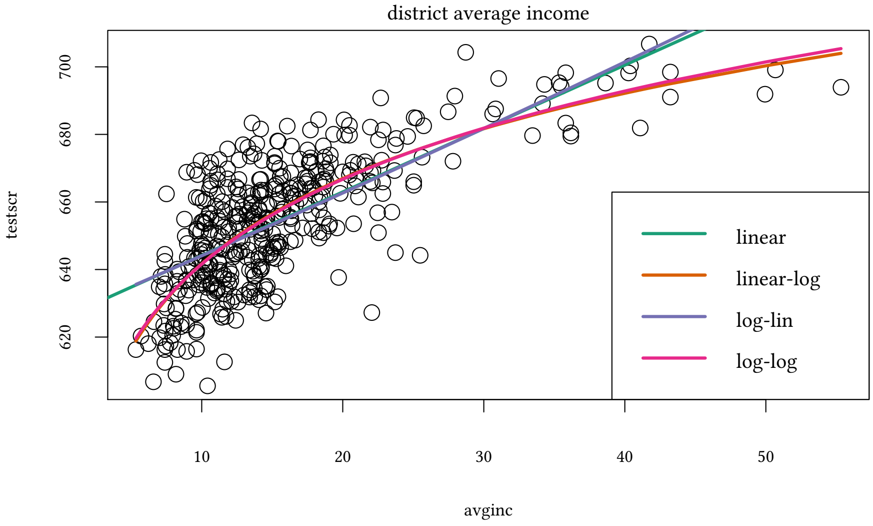
$$\frac{\partial Y_i}{\partial X_i} = e^{\beta_0} \cdot \beta_1 X_i^{\beta_1 - 1} = \beta_1 \frac{Y_i}{X_i}$$

$$\frac{\partial Y_i}{\partial X_i} \cdot \frac{X_i}{Y_i} = \beta_1 \quad \rightarrow \quad \frac{\partial Y_i}{Y_i} = \beta_1 \cdot \frac{\partial X_i}{X_i}$$

β_1 is the elasticity of Y_i with respect to X_i.

```
(estLLL<- lm(log(testscr) ~ log(avginc)))


Call:
lm(formula = log(testscr) ~ log(avginc))

Coefficients:
(Intercept)  log(avginc)
    6.33635      0.05542
```

```
plot(testscr ~ avginc,main="district average income")
abline(est1,col=1,lwd=3)
lines(avginc[or],fitted(estL)[or],col=2,lwd=3)
lines(avginc[or],exp(fitted(estLL))[or],col=3,lwd=3)
lines(avginc[or],exp(fitted(estLLL))[or],col=4,lwd=3)
legend("bottomright",c("linear","linear-log","log-lin","log-log"),lwd=3,col=c(1,2,3,4))
```



$$\frac{\partial Y_i}{\partial X_i} = e^{\beta_0} \cdot \beta_1 X_i^{\beta_1 - 1} \approx \beta_1 \frac{Y_i}{X_i}$$

```
X<-c(10,40,60)
(maLLL<-exp(coef(estLLL)[1])*
     coef(estLLL)[2]*X^(coef(estLLL)[2]-1))

[1] 3.5556504 0.9598969 0.6544736
```

3.4.4 Comparison of the three logarithmic models

- X and/or Y are transformed
- The regression equation is linear in the transformed variables
- Hypothesis tests and confidence intervals can be calculated in the usual way
- The interpretation of β is different in each case
- R^2 and AIC can be used to compare log-log and log-linear
- R^2 and AIC can be used to compare linear-log and a linear model
- Comparing $\log Y_i$ and Y_i is impossible.

→ We need economic theory to motivate one of the four specifications.

X	10.00	40.00	60.00
linLog	3.64	0.91	0.61
logLin	1.83	1.99	2.11
logLog	3.56	0.96	0.65

3.5 Summary of the linear model

$$Y_i = \beta_1 X_i + \beta_0 + u_i \qquad i = 1, \ldots, n$$

Assumptions:

- $E(u_i|X_i = x) = 0$ (strict exogeneity)
- (X_i, Y_i) are i.i.d.
- Large outliers in X and Y are rare (the fourth moments of X and Y exist)
- The sample is large (n is large) or residuals $u_i \sim N(0, \sigma)$

Results:

- Estimate (sample distribution of $\hat{\beta}$).
- Frequentist Null-hypothesis test:
  - $\hat{\beta}$ is t-distributed if $u_i \sim N(0, \sigma)$
  - $\hat{\beta}$ is approximately normally distributed (if n is large)

  → p-value.
- Confidence intervals

3.6 Literature

- Stock, Watson. “Introduction to Econometrics”. Chapters 4, 5, 8.

Appendix 3.A Examples for the lecture

Dummies Suppose you have estimated the following relationship:

$$Y = 15 + 2D + u$$

D is a dummy variable which is 0 for category A and 1 for category B.

1. You redefine D so that it is 1 for A and 0 for B. What relationship do you expect now?
2. You redefine D so that it is -1 for A and $+1$ for B. What can you say now about the relation between D and Y?

Rescaling Suppose you have estimated the following relationship:

$$Y = -5 + 3X + u$$

where X is a distance, measured in km.

- How large would the marginal effect of X be if you measured X in m (1 km = 1000 m)?
- The mean of $Y = 3$. By how much do you expect Y to change on average if X increases from a mean of $X = 5$ to $X = 6$?

Significance Which probability is given by the p-value?

- The probability to reject the null-hypothesis if it is false.
- The probability of a type 1 error
- The probability of a type 2 error
- The probability to reject the alternative hypothesis if it is false.
- None of the above

Confidence intervals and significance Your estimate for $\hat{\theta}$ follows a normal distribution. Your 99%-confidence interval for θ is $CI_\theta^{99\%} = [-1, 2]$. Which of the following hypotheses can be rejected at a 5% significance level?

- $\theta = 3$
- $\theta < -1$
- $\theta > 3$
- $\theta = 0$
- $\theta > 0$

Non-linear relationships You estimate the following model:

$$\log Y = \beta_0 + \beta_1 \log X$$

You estimate $\beta_0 = 3$, $\beta_1 = -1$. Which of the following statements is correct?

- If X decreases by 0.01, Y decreases by 0.01
- If X increases by 1%, Y decreases by 1%.
- If X increases by 0.01, Y decreases by 1%.
- If X increases by 1%, Y decreases by 0.01.
- None of the above

Appendix 3.B Exercises

To solve the next exercises the following commands for R might help: `<-`, `data`, `library`, `getwd`, `setwd`, `rm`, `write.table`, `read.table`, `str`, `names`, `order`, `with`, `tapply`, `ifelse`, `plot`, `scatterplot`, `abline`, `lines`, `cor`, `mean`, `summary`, `confint`, `log`, `exp`, `subset`, `lm`, `t.test`. Note that the `scatterplot` command is provided by the `car` library.

Regressions I

Exercise 3.1

- *Define the following items:*
  - *Regression*
  - *Independent variable*
  - *Dependent variable*
- *Give the formula for a linear regression with a single regressor.*
- *What does β_1 indicate?*
- *What does u indicate?*

Regressions II

Exercise 3.2 *Use the data set `Crime` of the library `Ecdat` in R.*

- *How do you interpret positive (negative) coefficients in simple linear models?*
- *What is the influence on the number of police men per capita (`polpc`) on the crime rate in crimes committed per person (`crmrte`)? Interpret your result. Do you have an explanation for this result?*

- *How can we visualize regressions? Draw the respective graph in R.*
- *What is the correlation coefficient of the number of police men per capital (`polpc`) and the crime rate (`crmrte`)? Interpret the result.*
- *What do the standard errors tell you?*
- *What does R^2 indicate?*
- *What do the p-values indicate?*

Regressions III

Exercise 3.3 *List and explain the assumptions that have to be fulfilled to be able to use OLS.*

Classes of variables

Exercise 3.4 *Explain and give examples of the following types of variables:*

- *continuous*
- *discrete*
- *binary*

Dummies

Exercise 3.5 *Use the data set `BudgetFood` of the library `Ecdat` in R.*

- *What is a dummy variable? Which values can it take?*
- *Name some typical examples of variables which are often coded as dummy variables.*
- *You have heard that older people care more about the quality of food they eat and thus spend more on food than younger people. Test whether Spanish citizens from the age of 30 on spend a higher percentage of their available income on food (`wfood`) than younger do. Do you have an explanation for these findings?*
- *Interpret the result of your regression. What does β_1 specify in this case?*
- *Could you apply a different test for the same question? Use this test in R and compare the results.*
- *Check the relation of age (`age`) and percentage of income spent on food (`wfood`) with a graph.*

Reading data into R

Exercise 3.6 • *Create a data file with the data on pocket money and age that we used in chapters 1 and 2. Use the headers* `age` *and* `pm` *and save it in the format* `csv` *under the name* `pocketmoney.csv`*.*

- *Read the file* `pocketmoney.csv` *into R.*
- *Draw a scatter plot with* `age` *on the x-axis and* `pm` *on the y-axis.*
- *Draw the same scatter plot, this time without box plots and without the lowess spline, but with a linear regression line.*
- *Label your scatter plot with* `age` *on the* `x`*-axis and* `pocket money` *on the* `y`*-axis. Give your graph the title "Children's pocket money".*

Exercise 3.7 *Your task is to work on a hypothetical data set in R.*

The variable names `A`*,* `B`*,* `C`*,* `D`*,* `E`*, and* `Year` *are in the header of your data file* `file.csv`*. The data set contains 553 observations in the format* `csv` *(comma separated values). Explain what the following commands do and choose the correct one (with explanation).*

- *First, read your data set into R.*
  1. `daten = read.csv("file.csv", header=YES, sep=';')`
  2. `daten = read.csv("file.csv", header=TRUE, sep=';')`
  3. `daten = read.table("file.csv")`
  4. `daten = read.table("file.csv", header=YES, sep=',')`
- *Further, you would like to know the correlation between the variables* `B`*,* `C`*, and* `D`*. How can you find this out?*
  1. `corr(B,C, D)`
  2. `corr(daten)`
  3. `cor(daten)`
  4. `corr(B,C, D)`

Using and generating dummy variables

Exercise 3.8 *Use the data set* `Fatality` *of the library* `Ecdat` *in R.*

- *What is the data set about?*
- *Which of the variables are nominal variables?*
- *Which of the variables are discrete variables?*

- *Which of the variables are continuous variables?*

- *Which of the variables are dummy variables?*

Exercise 3.9 *Use the data set* `Clothing` *of the library* `Ecdat` *in R.*

- *Create a dummy variable which takes the value 1 if the size of the sales floor is* > 120 *sqm and* 0 *otherwise.*

- *Draw a graph with separate box plots for large and small sales floors on the sales per square meter.*

- *Measure the influence of the size of the sales floor on the sales per square meter.*

- *Do the same task as above, this time using your variable for large sales floors.*

Advantages and disadvantages of OLS

Exercise 3.10 • *What are the advantages and disadvantages of using OLS?*

- *What can you do to fix these problems in OLS estimations?*

Logarithmic Regression

Exercise 3.11 • *What is a logarithm?*

- *Where do logarithms occur in nature or science?*

- *Give some examples for the use of logarithmic functions in economic contexts.*

Exercise 3.12 • *Which different types of logarithmic regressions do you know? Give the formulae for each of them.*

- *How do you interpret the coefficients of the different logarithmic models?*

- *Give economic examples for each of them.*

Exercise 3.13 *Use the data set* `Wages` *of the library* `Ecdat` *in R on wages in the United States.*

- *Estimate the effect of years of experience (*`exp`*), whether the employee has a blue collar job (*`bluecol`*), whether the employee lives in a standard metropolitan area (*`smsa`*), gender (*`sex`*), years of education (*`ed`*), and whether the employee is black (*`black`*) on the logarithm of wage (*`lwage`*). Do you think it makes sense to use this model? Would you rather suggest a different model? Which one would you suggest?*

- *Estimate both models in R. Interpret and compare the outputs.*

- *Visualize the relationship of experience and wage with a graph. Does this graph support the choice of your model?*

Exercise 3.14 • *How do you decide which model (linear model or one of the nonlinear models) to use?*

Exercise 3.15 *Use the data set* `Housing` *from the library* `Ecdat` *to study the effect of the lot size (*`lotsize`*) on its price (*`price`*).*

Estimate the following four specifications: linear, linear-log, log-linear, and log-log. Interpret the estimated coefficients and plot the respective model predictions. Which specification should you prefer based on R^2*?*

4 Review of Models with Multiple Regressors

$$\texttt{testsrc} = \beta_1 \texttt{str} + \beta_0 + u$$

- *testscr* test score
- *str* student / teacher ratio

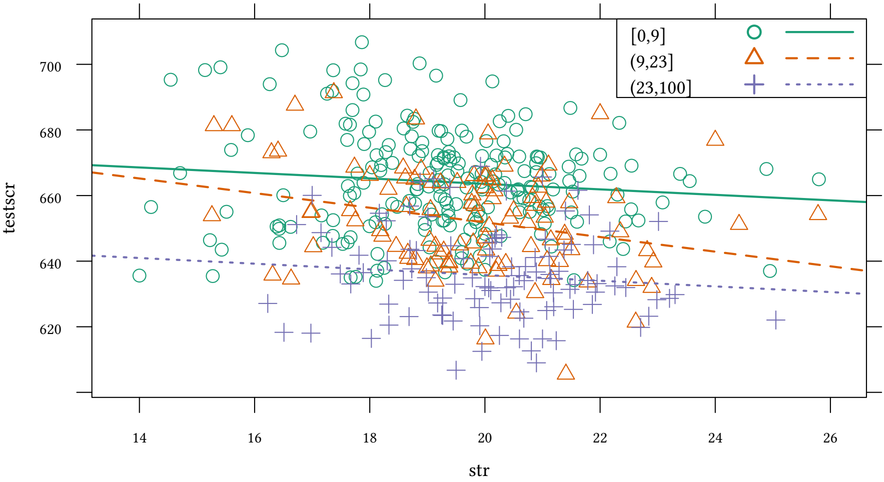
How can we include more than one factor at the same time?

- Keep one factor "constant" by only looking at a small group (e.g. all students with a very similar *elpct* (english learner percentage))

```
data(Caschool,package="Ecdat")
attach(Caschool)
summary(elpct)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   1.941   8.778  15.768  22.970  85.540 
```

```
lm(testscr ~ str ,subset=(elpct<9))


Call:
lm(formula = testscr ~ str, subset = (elpct < 9))

Coefficients:
(Intercept)          str  
    680.252       -0.835  

lm(testscr ~ str ,subset=(elpct>=9 & elpct<23))


Call:
lm(formula = testscr ~ str, subset = (elpct >= 9 & elpct < 23))
```

```
Coefficients:
(Intercept)          str
    696.445       -2.231

lm(testscr ~ str ,subset=(elpct>=23))


Call:
lm(formula = testscr ~ str, subset = (elpct >= 23))

Coefficients:
(Intercept)          str
   653.0746      -0.8656
```



depending on *elpct* the estimated relationships are very different.

→ extend the regression model

$$\mathtt{testsrc} = \beta_1 \mathtt{str} + \beta_2 \mathtt{elpct} + \beta_0 + u$$

generally:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

for each observation:

$$\left.\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + u_1\\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + u_2\\
y_3 &= \beta_0 + \beta_1 x_{31} + \beta_2 x_{32} + \cdots + \beta_k x_{3k} + u_3\\
&\vdots\\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + u_n
\end{aligned}\right\}\quad \mathbf{y} = X\beta + \mathbf{u}$$

```
lm(testscr ~ str + elpct)


Call:
lm(formula = testscr ~ str + elpct)

Coefficients:
(Intercept)          str        elpct
   686.0322      -1.1013      -0.6498
```

4.1 Assumptions for the multiple regression model

1. $E(u_i|X_i = x) = 0$
2. (X_i, Y_i) are i.i.d.
3. Large outliers in X and Y are rare (the fourth moments of X and Y exist)
4. X has the same rank as its number of columns (no multicollinearity)
5. $\text{var}(u|X = \mathbf{x})$ is constant, u is homoscedastic
6. u is normally distributed $u \sim N(0, \sigma^2)$

→ under assumptions 1–4 the OLS estimator $\widehat{\beta} = (X'X)^{-1}X'\mathbf{y}$ is unbiased and consistent.

→ For large samples $\widehat{\beta}$ is jointly normally distributed.

4.2 Specification errors

What happens if we don't include a variable into our model?

$$\text{E.g.:}\quad \texttt{testsrc} = \beta_1 \texttt{str} + \beta_0 + u$$

- *testscr* test score
- *str* student / teacher ratio

What else could have an influence on `testscr`?

	corr.with regressor `str`	influence on dep. var. `testscr`
percent of English learners	x	x
time of day of the test		x
parking lot space per student	x	

If we do *not* include a variable in our estimation equation, but

- this variable is correlated with the regressor
- *and* this variable is has an influence on the dependent variable

our estimation for β is biased (omitted variable bias).

4.2.1 Examples:

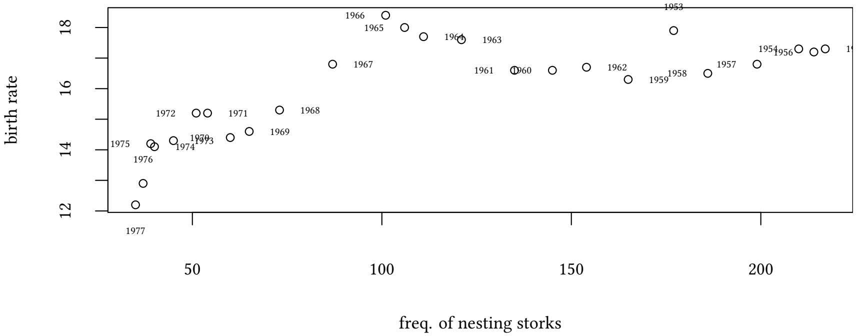
- Classical music → intelligence of children (Rauscher, Shaw, Ky; Nature; 1993)

  (missing variable: income)
- French paradox: Red wine, foie gras → less illnesses of the coronary blood vessels (Samuel Black, 1819)

  (missing variable: percentage of fish and sugar in the diet,...)
- Storks in Lower Saxony → birth rate

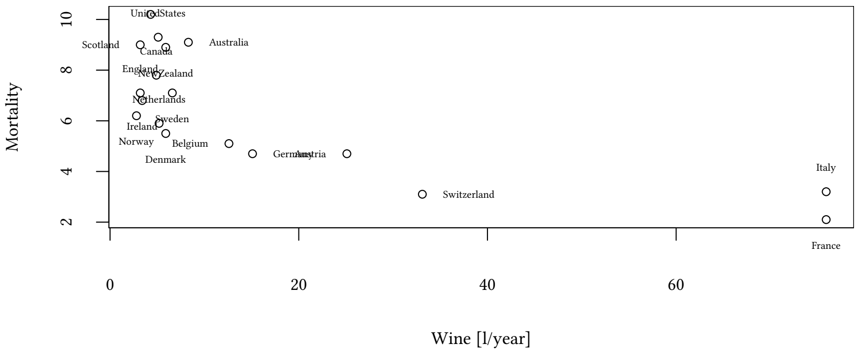
  (missing variable: industrialisation)

Birth rate and nesting storks



Gabriel, K. R. and Odoroff, C. L. (1990) Biplots in biomedical research. *Statistics in Medicine* 9(5): pp. 469-485.

French paradox



Mortality due to coronary heart disease (per 1000 men, 55 – 64 years). St. Leger A.S., Cochrane, A.L. and Moore, F. (1979). Factors Associated with Cardiac Mortality in Developed Countries with Particular Reference to the Consumption of Wine, *Lancet*: 1017–1020.

4.2.2 Specification errors —generalization

Let the true model be

$$y = X_1\beta_1 + X_2\beta_2 + u$$

what happens if we forget to include X_2 into the specification of our model?

$$\hat{\beta} = \underbrace{(X'X)^{-1}X'}_{X^+}y$$

$$\begin{aligned}
b_1 &= (X_1'X_1)^{-1}X_1'y \\
&= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + u) \\
&= (X_1'X_1)^{-1}X_1'X_1\beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + \\
&\qquad (X_1'X_1)^{-1}X_1'u \\
E(b_1) &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2
\end{aligned}$$

Hence, $E(b_1) = \beta_1$ only if

- $\beta_2 = 0$
- or $X_1'X_2 = 0$, i.e. X_1 and X_2 are orthogonal

Specification errors — another example The true model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

X_2 is correlated with X_1, e.g.:

$$X_2 = \alpha X_1 + \zeta$$

Omitting X_2 means:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2(\alpha X_1 + \zeta) + u \\ &= \beta_0 + \underbrace{(\beta_1 + \beta_2\alpha)}_{b_1} X_1 + \underbrace{\beta_2\zeta + u}_{u'} \end{aligned}$$

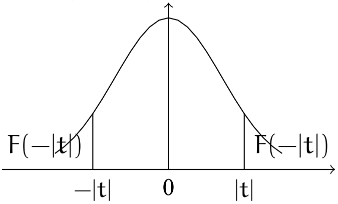
- We overestimate β_1 if $\beta_2\alpha > 0$
- We underestimate β_1 if $\beta_2\alpha < 0$
- underspecified model, a regressor β_2 is missing:
  - $\hat{\beta}$ is unbiased only if $\beta_2 = 0$ or $X_1'X_2 = 0$.
- overspecified model, regressors are collinear:
- - $\hat{\beta}$ cannot be estimated ($X'X$ cannot be inverted)
- overspecified model, regressors are almost collinear:
  - $\hat{\beta}$ can only be estimated inexactly

4.3 Hypothesis tests

4.3.1 Simple Hypotheses

Testing the hypothesis $H_0 : \beta_j = \beta_{j,0}$ against $H_1 : \beta_j \neq \beta_{j,0}$:

- Determine the t statistic:
$$t = \frac{\hat{\beta}_j - \beta_{j,0}}{\hat{\sigma}_{\hat{\beta}_j}}$$
- The p-value is $p = \Pr\left(|t| > \left|t^{\text{sample}}\right|\right) = 2\Phi(-|t^{\text{sample}}|)$



4.3.2 Joint Hypotheses

```r
summary(lm(testscr ~ str + elpct + expnstu))
```

```
Call:
lm(formula = testscr ~ str + elpct + expnstu)

Residuals:
    Min      1Q  Median      3Q     Max 
-51.340 -10.111   0.293  10.318  43.181 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 649.577947  15.205719  42.719  < 2e-16 ***
str          -0.286399   0.480523  -0.596  0.55149    
elpct        -0.656023   0.039106 -16.776  < 2e-16 ***
expnstu       0.003868   0.001412   2.739  0.00643 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.35 on 416 degrees of freedom
Multiple R-squared:  0.4366,Adjusted R-squared:  0.4325 
F-statistic: 107.5 on 3 and 416 DF,  p-value: < 2.2e-16
```

We might want to test the following:

$$\beta_{\texttt{str}} = 0 \text{ and } \beta_{\texttt{expnstu}} = 0 \,.$$

$$\begin{array}{ll} \text{Formally} & H_0 : \beta_1 = \beta_{1,0} \wedge \beta_2 = \beta_{2,0} \\ & \qquad\qquad \text{versus} \\ & H_1 : \beta_1 \neq \beta_{1,0} \vee \beta_2 \neq \beta_{2,0} \end{array}$$

Idea: We could just test $\beta_1 = 0$ and $\beta_2 = 0$ independently of each other.

$$t_1 = \frac{\hat\beta_1 - \beta_{1,0}}{\hat\sigma_{\hat\beta_1}} \qquad t_2 = \frac{\hat\beta_2 - \beta_{2,0}}{\hat\sigma_{\hat\beta_2}}$$

In that case the null hypothesis $H_0 : \beta_1 = \beta_{1,0} \wedge \beta_2 = \beta_{2,0}$ would be rejected if either $\beta_1 = 0$ or $\beta_2 = 0$ are rejected.

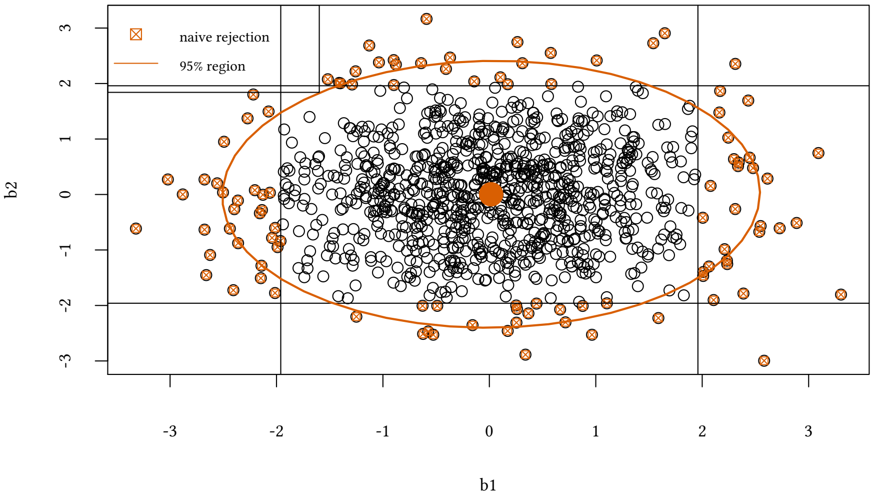
We can easily see that this does not even work for uncorrelated βs:

```
set.seed(100)
N<-1000
p<-0.05
qcrit<- -qnorm(p/2)
b1<-rnorm(N)
mean(abs(b1)>qcrit)*100

[1] 5.9

b2<-rnorm(N)
mean(abs(b2)>qcrit)*100
```

```
[1] 4.6

reject<-abs(b1)>qcrit | abs(b2)>qcrit
mean(reject)*100

[1] 10.3
```

In the example 10.3 % of the values are rejected by the joint test, not 5%. This is not a coincidence. The next diagram shows that we are not only cutting off on the left and on the right, but also at the top and at the bottom.

```
plot(b2 ~ b1,cex=.7)
points(b2 ~ b1,subset=reject,col=2,pch=7,cex=.5)
abline(v=c(qcrit,-qcrit),h=c(qcrit,-qcrit))
dataEllipse(b1,b2,levels=1-p,plot.points=FALSE,col=2)
legend("topleft",c("naive rejection","95\\% region"),pch=c(7,NA),col=2,lty=c(NA,1),cex=.7)
```

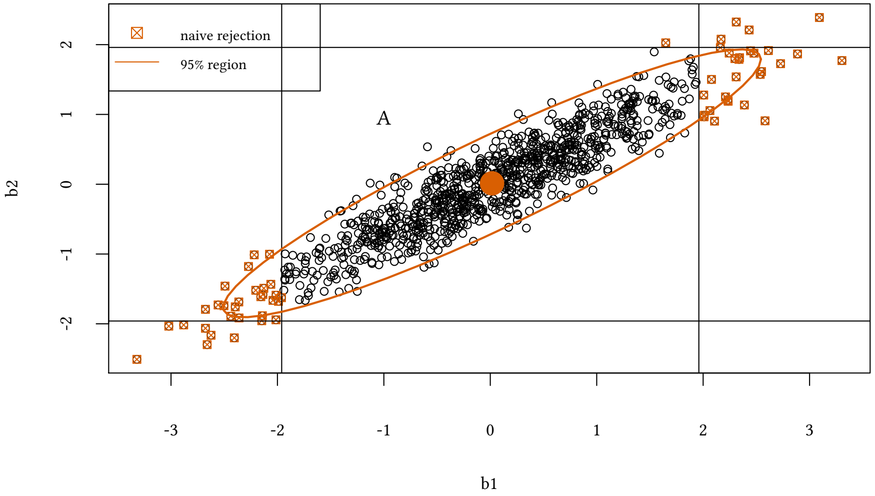


Additionally we can see that this naïve approach only takes the maximum deviation of the variables into account. It would be more sensible to exclude all observations outside of the (elliptical) 95% region.

The second problem becomes even more annoying if the random variables are correlated:

```
b1<-rnorm(N)
b2<-.3* rnorm(N)  + .7*b1
reject<-abs(b1)>qcrit | abs(b2)>qcrit
```

```
plot(b2 ~ b1,cex=.5)
points(b2 ~ b1,subset=reject,col=2,pch=7,cex=.5)
abline(v=c(qcrit,-qcrit),h=c(qcrit,-qcrit))
dataEllipse(b1,b2,levels=1-p,plot.points=FALSE,col=2)
text(-1,1,"A")
legend("topleft",c("naive rejection","95\\% region"),pch=c(7,NA),col=2,lty=c(NA,1),cex=.7)
```



For example, ”‘A”’ in the diagram is clearly outside the confidence ellipse, but none of its single coordinates are conspicious.

4.3.3 F statistic for two restrictions

$$t_1 = \frac{\hat\beta_1 - \beta_{1,0}}{\hat\sigma_{\hat\beta_1}} \qquad t_2 = \frac{\hat\beta_2 - \beta_{2,0}}{\hat\sigma_{\hat\beta_2}}$$

$$F = \frac{1}{2} \cdot \frac{t_1^2 + t_2^2 - 2\hat\rho_{t_1 t_2} \cdot t_1 \cdot t_2}{1 - \hat\rho_{t_1 t_2}^2}$$

with $\hat\rho_{t_1 t_2}$ being the estimated correlation between t_1 and t_2.

4.3.4 More than two restrictions

Write restrictions as

$$R\beta = \mathbf{r}$$

e.g.

$$(0,1,0,\cdots,0)\begin{pmatrix}\beta_0\\\beta_1\\\beta_2\\\vdots\\\beta_k\end{pmatrix}=0$$

$$\begin{pmatrix}0&1&0&\cdots&0\\0&0&1&\cdots&0\end{pmatrix}\begin{pmatrix}\beta_0\\\beta_1\\\beta_2\\\vdots\\\beta_k\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix}$$

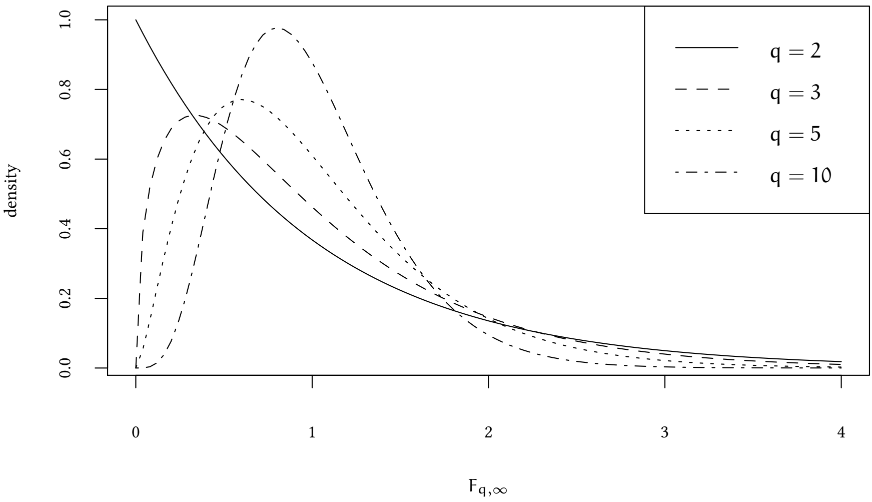
$$\begin{pmatrix}1&0&-1&\cdots&0\\0&1&0&\cdots&0\\0&1&2&\cdots&0\end{pmatrix}\begin{pmatrix}\beta_0\\\beta_1\\\beta_2\\\vdots\\\beta_k\end{pmatrix}=\begin{pmatrix}0\\7\\0\end{pmatrix}$$

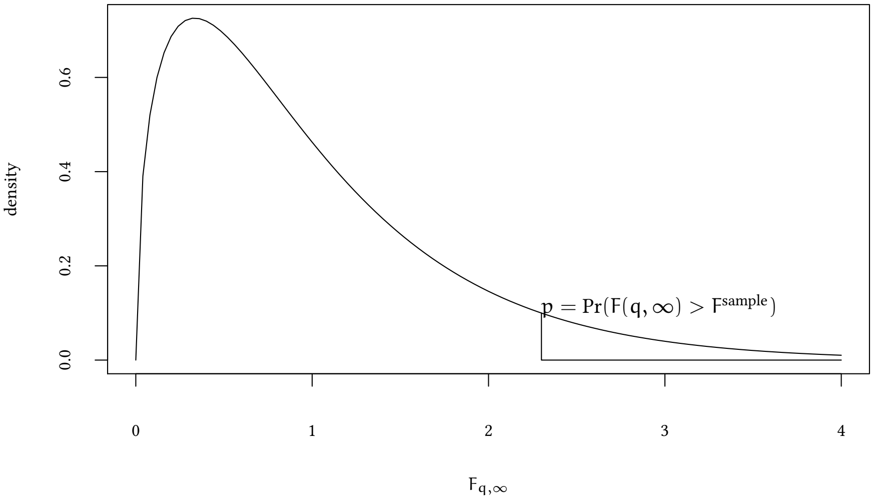
$$F=\frac{1}{q}(R\widehat{\boldsymbol{\beta}}-\mathbf{r})'\left(R\hat{\Sigma}_{\widehat{\beta}\widehat{\beta}}R'\right)^{-1}(R\widehat{\boldsymbol{\beta}}-\mathbf{r})$$

with q being the number of restrictions.

If assumptions 1–4 (see 4.1) are satisfied:

$$F\xrightarrow{p}F_{q,\infty}$$





4.3.5 Specials cases:

The last line of a regression output

```
summary(lm(testscr ~ str + elpct + expnstu))


Call:
lm(formula = testscr ~ str + elpct + expnstu)

Residuals:
    Min      1Q  Median      3Q     Max 
-51.340 -10.111   0.293  10.318  43.181 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 649.577947  15.205719  42.719  < 2e-16 ***
str          -0.286399   0.480523  -0.596  0.55149    
elpct        -0.656023   0.039106 -16.776  < 2e-16 ***
expnstu       0.003868   0.001412   2.739  0.00643 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.35 on 416 degrees of freedom
Multiple R-squared:  0.4366,Adjusted R-squared:  0.4325 
F-statistic: 107.5 on 3 and 416 DF,  p-value: < 2.2e-16
```

Restrictions for “the” F-statistic of an estimation (β_0 is not being tested)

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Testing a single coefficient:

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} = 0$$

$$t_1 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\hat{\sigma}_{\hat{\beta}_1}} \qquad X \sim t_{(k)} \Leftrightarrow X^2 \sim F_{(1,k)}$$

`c` constructs a vector by joining the arguments together. `rbind` joins the arguments of the function (vectors, matrices) line-wise. `cbind` joins the arguments of the function (vectors, matrices) column-wise. `linearHypothesis` tests linear hypotheses. `pf` calculated the distribution function of the F-distribution, *`df`* calculates the density function, *`qf`* calculates quantiles of the F-distribution, *`rf`* calculates an F-distributed random variable.

```
est<-lm(testscr ~ str + elpct + expnstu)
linearHypothesis(est,c("str=0","expnstu=0"))

Linear hypothesis test

Hypothesis:
str = 0
expnstu = 0

Model 1: restricted model
Model 2: testscr ~ str + elpct + expnstu

  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1    418 89000
2    416 85700  2    3300.3 8.0101 0.000386 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.4 Interactions

- Maybe the effect of group sizes on test scores depends on further circumstances.
- Maybe small groups sizes have a particularity large effect if groups have a lot of foreign students.
- $\frac{\partial\texttt{testscr}}{\partial\texttt{str}}$ depends on *elpct*.
- Generally: $\frac{\partial Y}{\partial X_1}$ depends on X_2.
- How can we include this interaction into a model?
- First look at binary X, later consider continuous X.

Example 1:

$$\texttt{testsrc} = \beta_1\texttt{str} + \beta_2\texttt{elpct} + \beta_0 + u$$

in this model the effect of *str* is (assumed to be) independent of *elpct*

Example 2:

$$\texttt{lwage} = \beta_1\texttt{ed} + \beta_0 + u$$

```
attach(Wages)
summary(ed)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.00   12.00   12.00   12.85   16.00   17.00
```

$$\mathtt{lwage} = \beta_1\mathtt{college} + \beta_2\mathtt{sex} + \beta_0 + u$$

```
college=ed>16
lm(lwage ~ college + sex)


Call:
lm(formula = lwage ~ college + sex)

Coefficients:
(Intercept)  collegeTRUE      sexmale
     6.2254       0.3340       0.4626
```

4.4.1 Interactions between binary variables

$$\mathtt{lwage} = \underbrace{\beta_0}_{6.21} + \underbrace{\beta_1}_{0.55}\mathtt{college} + \underbrace{\beta_2}_{0.49}\mathtt{sex} + \underbrace{\beta_3}_{-0.24}\mathtt{sex}\cdot\mathtt{college} + u$$

```
(est<- lm(lwage ~ college + sex + sex:college))


Call:
lm(formula = lwage ~ college + sex + sex:college)

Coefficients:
        (Intercept)          collegeTRUE              sexmale
             6.2057               0.5543               0.4850
collegeTRUE:sexmale
            -0.2412
```

Instead of regression coefficients we can calculate mean values for the individual categories:

```
aggregate(lwage ~ college + sex,FUN=mean)

  college    sex    lwage
1   FALSE female 6.205665
2    TRUE female 6.760007
3   FALSE   male 6.690634
4    TRUE   male 7.003751
```

mean(lwage)		sex	
		female	male
college	FALSE	6.21 β_0	6.69 $\beta_0 + \beta_2$
	TRUE	6.76 $\beta_0 + \beta_1$	7.00 $\beta_0 + \beta_1 + \beta_2 + \beta_3$

Effect of college education for women: β_1 Effect of college education for men: $\beta_1+\beta_3$

4.5 Literature

- Stock, Watson. "Introduction to Econometrics". Chapters 6-9.

Appendix 4.A Examples for the lecture

Example 1 You are the trainer of a mixed team in track and field of children aged 10–13. You have estimated the following model for the time (in seconds) that it takes your team members to run 100 meters:

$\widehat{\texttt{time}}_i = 29.8 - 0.8 \cdot \texttt{male} - 1.2 \cdot \texttt{age} + 0.5 \cdot \texttt{rain}$

where *male*=1 if it is a boy, *age* denotes the child's age in years and *rain*=1 if it has been raining during the day so that the ground is muddy.

- Which of the variables are dummy variables? Explain the meaning of the coefficients of the dummy variables.
- How would the estimated model look like if we dropped the variable *male* (1 if male, 0 otherwise) and added a variable *female* (1 if female, 0 otherwise) instead?
- You would like to predict the time that it takes the next girl to run 100 meters. She is 10 years old. You know that it has not been raining today. What is the predicted time?
- What would be the prediction for a 13-years-old boy on a day when it has been raining?
- Assume that you use the estimation above for a 20 year old man on a rainy day. What would be the estimated time? Is that realistic? Why?
- In your prediction of the time that the members of your team need to run 100 meters you would like to add a measure for ability. Unfortunately, you have never classified your team members according to ability. Which other measures could help you to approximate the ability of each child?
- Do you think the model is well specified? What would you change if you could? Why might you not be able to specify the model the way you want?

Example 2 Use the data set *Wages1* of the library *Ecdat* in R.

- Estimate the effect of experience (*exper*), years of schooling (*school*), and gender (*sex*) on hourly wages (*wage*).
- Has experience or education a higher impact on wage?

- Is the impact of one more year of education higher for employees who have never attended college (`school` ≤ 12) or for those who have received some college education (`school` > 12)? Visualize this with a graph.

- Do you see a similar effect for experience?

Example 3 In the following regression you estimate the impact of a and b on Y. The variables a and b can each take values 0 or 1. You obtain the following output:

```
Call: lm(formula = y ~ a * b)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.0000     0.7071   5.657 1.55e-05
a            4.0000     1.0000   4.000 0.000704
b           -4.0000     1.0000  -4.000 0.000704
a:b         -2.0000     1.4142  -1.414 0.172678
---
Residual standard error: 1.732 on 20 degrees of
freedom
Multiple R-squared: 0.7778
F-statistic: 23.3 on 3 and 20 DF,p-value: 9.7e-07
```

Which values for Y do you expect for the following combinations of a and b?

a	b	$\hat{Y}$
0	0	
1	0	
1	1	

What is the marginal effect...

- of b on Y if $a = 1$?

- of a on Y if $b = 0$?

How do you determine for the above regression the lower limit of the 95%-confidence interval for the coefficient of b?

- `4-2*qt(.975,df=20)`

- `4-qt(.975,df=20)`

- `-4-qt(.975,df=20)`

- `4+qt(.975,df=20)`

- none of the above

You still consider the above regression. You assume a significance level of 5%. Furthermore you assume normally distributed residuals.

- a has a significant impact on b.
- b has a significant impact on Y.
- The hypothesis the coefficient of b is 5 is rejected.
- The hypothesis the coefficient of b is -1 is rejected.
- The model explains 23.3% of the variance of Y.

Appendix 4.B Exercises

Exercise 4.1

- *What is a multiple regression?*
- *Give the formula for a regression with multiple regressors. Explain what each part of the formula means.*
- *Which other methods do you know to include more than one factor into an analysis?*

Exercise 4.2 *You would like to find out which factors influence the sales of downhill skis in your friend's store.*

- *How would you specify your model? Which regressors would you include?*
- *Which signs would you expect for the coefficients of the different regressors?*

Exercise 4.3 *You would like to conduct a survey to find out which factors influence the income of employees.*

- *Which variables do you think have an influence on income and should be included in your model?*
- *You are only allowed to ask your respondents about their age, their gender, and the number of years of education they have obtained. Build a model with these variables as regressors. Which signs do you expect for the coefficients of each of these variables?*
- *Your assistant has estimated the following equation for monthly incomes:* $\hat{I} = 976.9 + 38.2 \cdot a + 80.5 \cdot b - 350.7 \cdot c$ *with* $N = 250$. *Unfortunately, he has not noted which variables indicate what. Look at the regression. Can you tell which variable stands for which factor?*
- *What is the estimated income of a woman aged 27 who has obtained 17 years of education?*
- *One employee wonders whether she should pursue a MBA program. The one-year program costs 6 000 in tuition fees. During this year she will forego a monthly salary of 3 200 (assume 12.5 salaries per year, for simplicity assume that you live in a world without taxation and where future income is not discounted). Will this degree pay out during the next ten years?*

- *Do you think that the above model is a good one? What would you change if you could?*

Exercise 4.4 *Use the data set `Icecream` of the library `Ecdat` in R.*

- *You want to estimate the effect of average weekly family income (`income`), price of ice cream (`price`), and the average temperature (`temp`) on ice cream consumption (`cons`). Formulate your model.*
- *Which variables do you expect to have an influence on ice cream consumption? In which direction do you expect the effect?*
- *Check your assumptions in R.*

Exercise 4.5 *Use the data set `Computers` of the library `Ecdat` in R.*

- *What does the data set contain?*
- *Classify the variables of the data set.*
- *Estimate the price of a computer (`price`). Which regressors would you include? Why? Which sign would you expect for the coefficients of your regressors?*
- *Interpret your results.*

Exercise 4.6 *Use the data set `RetSchool` of the library `Ecdat` in R.*

- *You want to estimate the effect of grades (`grade76`) and experience (`exp76`) on wages (`wage76`) separately for younger (≤ 30 years) and older (> 30 years) employees. Is the effect of grades and experience different for these two age groups?*
- *Now you estimate the same model as above, but this time using a dummy variable for employees who are older than 30 years in 1976. Explain the difference of this model and the model above. How do you interpret the results?*

Specification error I

Exercise 4.7
- *What happens if you forget to include a variable? When is this problematic? How is this specification error called? Give some examples for this problem.*
- *What could you do to fix this problem? Why might one not be able to fix this problem?*

Exercise 4.8
- *What are the assumptions that have to be fulfilled for using multiple regressions? List and explain them.*

Exercise 4.9 *Multicollinearity is one of the problems that might occur in an estimation.*

- *Describe what multicollinearity means.*
- *How can one detect multicollinearity?*
- *How can one avoid multicollinearity?*

Rents in Jena

Exercise 4.10 *A company rents apartments for students in Jena. The manager would like to estimate a model for rents for apartments. He has information on the size of the apartment, the number of bedrooms, the number of bathrooms, whether the kitchen is large, whether the apartment has a balcony, whether there is a tub in the bath room, and the location measured as the distance to the ThULB.*

- *Specify a model to estimate the rents in Jena. Which of the above variables would you include? Which signs do you expect the coefficients of these variables to take? Explain your answer.*
- *Do you think the model is well specified? Are there any other variables you would like to add?*

Exercise 4.11 *Product Z of your company has been advertised during the last year on two different TV channels: EURO1 and SAT5. Prices for spots are the same on both channels. A study with data on the last available periods has provided the following model (standard errors in parentheses):*

$$\hat{Y}_i = 300 + \underset{(1.0)}{10} \cdot X_1 + \underset{(2.5)}{20} \cdot X_2$$

You have 44 observations, $R^2 = 0.9$. Y *stands for the sales amount of your product* Z *(in 1000 €),* X_1 *stands for expenses on commercials at EURO1 (in 1000 €),* X_2 *for expenses on SAT5 (in 1000 €).*

- *Which advertisement method should you prefer according to your regression results (all other factors constant)? Explain your answer.*

Exercise 4.12 *Use the data set* `Housing` *of the library* `Ecdat` *in R.*

- *Build a model predicting the price of a house (*`price`*) depending on the lotsize (*`lotsize`*), the number of bathrooms (*`bathrms`*), the number of bedrooms (*`bedrooms`*), whether the house has air condition (*`airco`*), and whether the house is located in a preferred neighbourhood (*`prefarea`*). Estimate this model in R.*
- *Create a dummy which takes the value 1 if the house has at least one bathroom. Estimate the same model as above, this time using the dummy for bathroom instead of the number of bathrooms. What happens? Why?*
- *Construct a variable which indicates the prices in €. Assume an exchange rate of 0.74 € for each Canadian Dollar. Estimate the same model as above, this time estimating the price of the house in €. Interpret your result.*
- *Create a dummy for houses with fewer than 3 bedrooms. Estimate the same model as above, this time including the dummy variable for a small hourse. Explain your result.*

Specification errors

Exercise 4.13 *Use the data set* `BudgetItaly` *of the library* `Ecdat` *in R.*

- *Estimate the effect of the price of food (*`pfood`*), and of the size of the household (*`size`*) on the share of expenditures for food (*`wfood`*) in R.*
- *Look at the results of your estimation and interpret it. Do you think you could estimate a better model?*
- *Now add the share of expenditures for housing and fuel (*`whouse`*). Interpret your result.*
- *Do you think there is any multicollinearity in your model? Test this.*
- *Create a dummy variable for the price of food in € (*`Europrice`*). Every Lira is worth 0.0005 €. Add this variable to your second model. What happens? Why?*

Exercise 4.14

- *What specification errors do you know? List and explain them.*
- *What does this mean for your estimations?*
- *How should you proceed in modeling?*

Exercise 4.15 *Use the data set* `Housing` *of the library* `Ecdat` *in R.*

- *Draw a scatter plot on the the lot size and the price of a house. Look at the graph. From your visual impression, would you say that the error terms are homo- or heteroscedastic?*
- *What does the data set contain?*
- *Look at the variables the data set contains. Formulate some sensible hypotheses and test them in R.*

Exercise 4.16

- *What is heteroscedasticity?*
- *Give an example for data where residual variances of differ along the dimension of a second variable.*
- *What is homoscedasticity?*
- *Which advantage does homoscedasticity have for econometric analysis?*
- *Imagine you had data with heteroscedastic error terms. You perform a data analysis under the assumption of homoscedasticity. Is your estimator consistent?*
- *Now you have data with homoscedastic error terms and you perform a data analysis under the assumption of heteroscedasticity. Is your estimator consistent?*
- *How would the answers to the last two questions be if you had a large sample?*

Linear and Non-linear Regressions

Exercise 4.17 *You want to estimate different models for the following problem sets of one dependent and one independent variable (assume that the models are otherwise specified correctly, i.e. all other important variables are included, no high correlation between two independent variables).*

Name the appropriate model (linear, quadratic, log-lin, lin-log, or log-log) for each of the problems and explain your choice (exercise adapted from Studenmund's "Using econometrics", chapter 7, exercise 2).

- *Dependent variable: time it takes to walk from A to B, independent variable: distance from A to B*
- *Dependent variable: total amount spent on food, independent variable: income*
- *Dependent variable: monthly wage, independent variable: age*
- *Dependent variable: number of ski lift tickets sold, independent variable: whether there is snow*
- *Dependent variable: GDP growth rate , independent variable: years passed since beginning of transformation to an industrialized country*
- *Dependent variable: CO2 emission, independent variable: kilometers driven with car*
- *Dependent variable: hourly wage, independent variable: number of years of job experience*
- *Dependent variable: physical ability, independent variable: age*

Exercise 4.18 *A group of athletes prepares for a competition. You have the following information about the athletes: age (A), gender (G; 1 if female, 0 otherwise), daily training (T; 1 if true, 0 otherwise), healthy diet (E; 1 if true, 0 otherwise), and ranking list scores (R). Age and gender are not correlated with the other variables. You assume that athletes only do especially well in the ranking list if they practice daily and if they follow a healthy diet; a daily training is only effective in combination with a healthy diet. What would be possible specifications of your model to test your assumption? (Here we* ***don't*** *ask for the "best" specification.)*

1. $R = \beta_0 + \beta_1 \cdot T + \beta_2 \cdot E + u$
2. $R = \beta_0 + \beta_1 \cdot A + \beta_2 \cdot G + \beta_3 \cdot T + \beta_4 \cdot E + \beta_5 \cdot T \cdot E + u$
3. $R = \beta_0 + \beta_1 \cdot T + \beta_2 \cdot E + \beta_3 \cdot T \cdot E + u$
4. $T \cdot E = \beta_0 + \beta_1 \cdot R + u$
5. $R = \beta_0 + \beta_1 \cdot A + \beta_2 \cdot G + \beta_3 \cdot T + \beta_4 \cdot E + u$

Interaction terms

Exercise 4.19

- *What is an interaction term?*
- *How do you construct an interaction term?*
- *Write down a regression function including an interaction term.*
- *How do you interpret interaction terms?*
- *Why do you have to include not only the interaction of two variables into your regression function, but also each of the individual variables? What would happen if you would not include the individual variables?*
- *Give some examples of situations where you think that interactions play a role.*

Exercise 4.20 *You are a teacher of a cross country skiing school. Each year you teach students who have never done cross country skiing before the free style technique (also called skating technique). You realize that your students differ in how fast they learn this new technique. You think that the two things that matter are whether a student knows how to ice skate and whether a student is familiar with down hill skiing.*

You estimate the following model for the number of days it takes them to learn the new technique so well that they are able to do their first tours:

$$\widehat{\texttt{days}}_i = 8 - 1 \cdot \texttt{iceSkating} - 2.5 \cdot \texttt{alpineSkiing} - 1.5 \cdot \texttt{iceSkating} \cdot \texttt{alpineSkiing}$$

- *How many day needs a person ...to learn the skating technique with cross country skis?*
  - *who has never done any ice skating nor downhill skiing*
  - *who has never done any ice skating, but some downhill skiing*
  - *who knows how to ice skate, but has never done any downhill skiing*
  - *who is familiar with both ice skating and downhill skiing*

Exercise 4.21 *You would like to estimate students' school achievements measured in test scores (testscore) in a developing country. You think that gender (female) and educational background of the parents (eduparents; measured in years) have an impact. In particular, you think that poor people cannot afford that their children spend all their time on learning, because they also need their help to earn money. This might be especially true for girls, because their parents might think that education is less important for them. How would you test this assumption in R? Which if the following commands is correct (multiple correct answers possible)?*

1. `summary(lm(testscore ~ female+eduparents+eduparents*female))`

2. *`summary(lm(testscore=female+eduparents+eduparents:female))`*

3. *`summary(lm(testscore ~ female*eduparents))`*

4. *`summary(lm(testscore ~ female+eduparents+eduparents:female))`*

5. *`eduparentsfemale <- eduparents*female summary(lm(testscore ~ female+eduparents+eduparentsfemale))`*

6. *`summary(lm(testscore <- eduparents*female))`*

7. *`summary(lm(testscore ~ eduparents:female))`*

Exercise 4.22 *Use the data set `RetSchool` of the library `Ecdat` in R on returns to schooling in the United States.*

- *You are interested whether people considered as "black" (`black`) and people living in the south (`south76`) earn less than others. Further, you are interested whether Afro Americans (black) who live in the south (`south76`) earn even less. You control for years of experience (`exp76`) and grades (`grade76`). Solve this problem using R.*

Exercise 4.23 *Use the data set `DoctorContacts` of the library `Ecdat` in R on contacts to medical doctors.*

- *How do gender (`sex`), age (`age`), income (`linc`), the education of the head of the household (`educdec`), health (`health`), physical limitations (`physlim`), and the number of chronic diseases (`ndisease`) effect the number of visits to a medical doctor? In which direction do you expect the effects to go?*

- *Is there an interaction between gender and physical limitations?*

Non-linear functions

Exercise 4.24

- *Which non-linear functions do you know? List them.*

- *Note the formula for each of them.*

- *Give examples for each of them.*

5 Bootstrap

5.1 Motivation

Let us start with a simple example[1]: a comparison of two treatments:

[1] The example is based on B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, 1994.

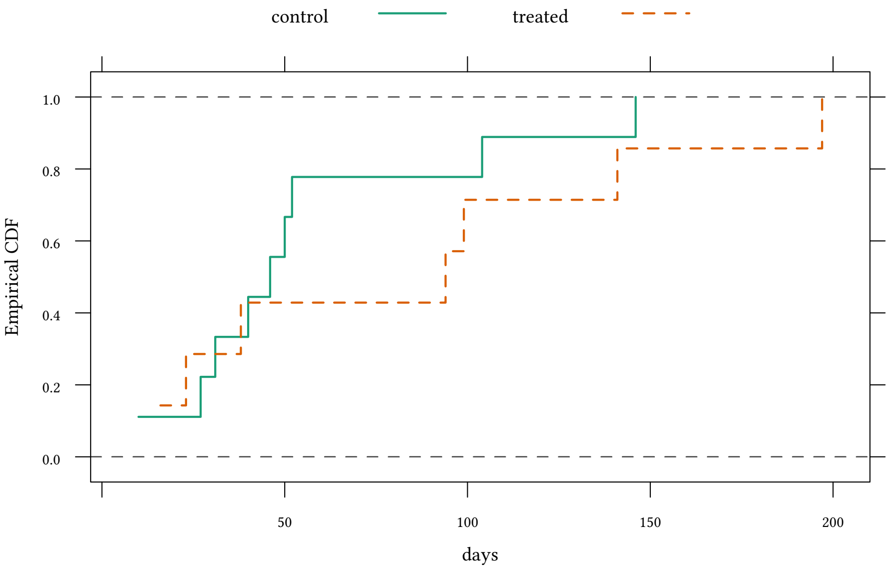
```
library(bootstrap)
library(boot)
mouse.t

[1]  94 197  16  38  99 141  23

mouse.c

[1]  52 104 146  10  50  31  40  27  46
```

Treated mice have a larger mean survival time:

```
mean(mouse.t)

[1] 86.85714

mean(mouse.c)

[1] 56.22222
```

Question: Is the difference significant?

```
with(rbind.fill(data.frame(type="treated",mouse=mouse.t),
                data.frame(type="control",mouse=mouse.c)),
     ecdfplot(~mouse,group=type,auto.key=list(columns=2),xlab="days"))
```



Traditionally we estimate

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

and hence

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{\hat{\sigma}_x^2}{n}}$$

The estimated standard error of the difference between the two means $\bar{x}$ and $\bar{y}$ is

$$\hat{\sigma}_{\bar{x}-\bar{y}} = \sqrt{\frac{\hat{\sigma}_x^2}{n} + \frac{\hat{\sigma}_y^2}{m}}$$

```
(sdDiff <- sqrt(var(mouse.t)/length(mouse.t)+
                var(mouse.c)/length(mouse.c)))

[1] 28.92647
```

Hence $t = \frac{\bar{x}-\bar{y}}{\sigma_{\bar{x}-\bar{y}}}$ is

```
(t <- (mean(mouse.t) - mean(mouse.c)) / sdDiff )

[1] 1.059062
```

Is this is large number? Does t follow (approximately) a Normal distribution under the Null?

```
2*(1-pnorm(t))

[1] 0.2895715
```

Why are we allowed to do this?

Central limit theorem

Be $X_1, \ldots, X_n$ a sequence of independently and identically (i.i.d.) distributed random variables, then with expected value $E(X)$ and variance σ_X^2, then

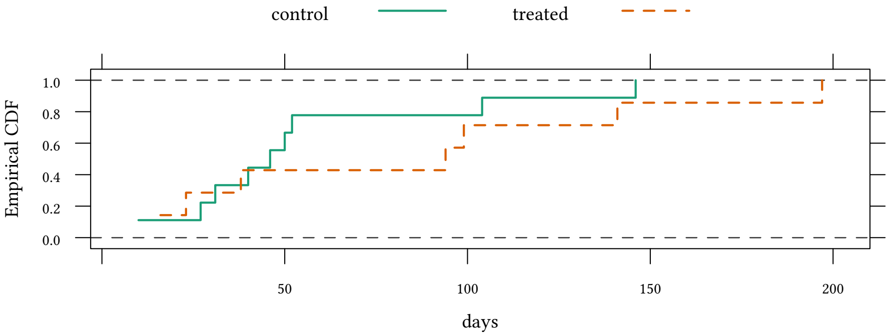
$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i - nE(X)}{\sigma\sqrt{n}} = \lim_{n\to\infty}\left(\frac{\sum_{i=1}^{n} X_i}{n} - E(X)\right)\cdot\frac{\sqrt{n}}{\sigma} \sim N(0,1)\,.$$

When do we have a problem?

- When n is *small*.
- When we are interested in a statistic other than the *mean*.

Example: Let us assume we are interested in the median:

```
mean(mouse.t) - mean(mouse.c)

[1] 30.63492

median(mouse.t) - median(mouse.c)

[1] 48
```



- In this example the difference in medians is larger than the difference in means. But what is the standard error of the difference in medians?

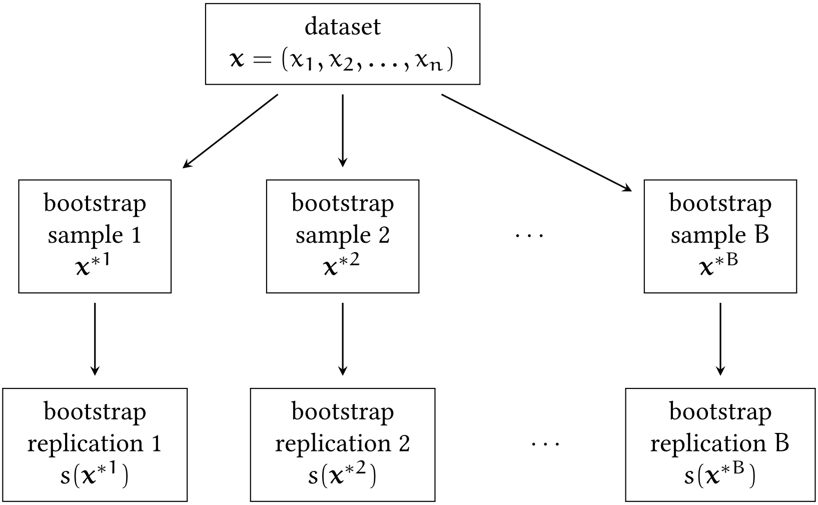
- There are formulae for the standard error of medians, but they require specific distributions of X.

- Do we know the distribution of X? — No, but we can estimate it!

- Above we have estimated $E(X)$ by using our sample to calculate the statistic: $\bar{x} = \widehat{E(X)}$

- Now we use our sample as an estimation of the distribution of X.

Let us assume that we are interested in the statistic s. Our *estimation of the distribution* of X is given by the *distribution of the dataset* $\mathbf{x}$.


dataset
$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$
bootstrap sample 1 $\boldsymbol{x}^{*1}$
bootstrap sample 2 $\boldsymbol{x}^{*2}$
...
bootstrap sample B $\boldsymbol{x}^{*B}$
bootstrap replication 1 $s(\boldsymbol{x}^{*1})$
bootstrap replication 2 $s(\boldsymbol{x}^{*2})$
...
bootstrap replication B $s(\boldsymbol{x}^{*B})$

Then our *estimate of the distribution of* s is given by the *distribution of the bootstrap replications* $s(\boldsymbol{x}^{*1}), s(\boldsymbol{x}^{*2}), \ldots, s(\boldsymbol{x}^{*B})$.

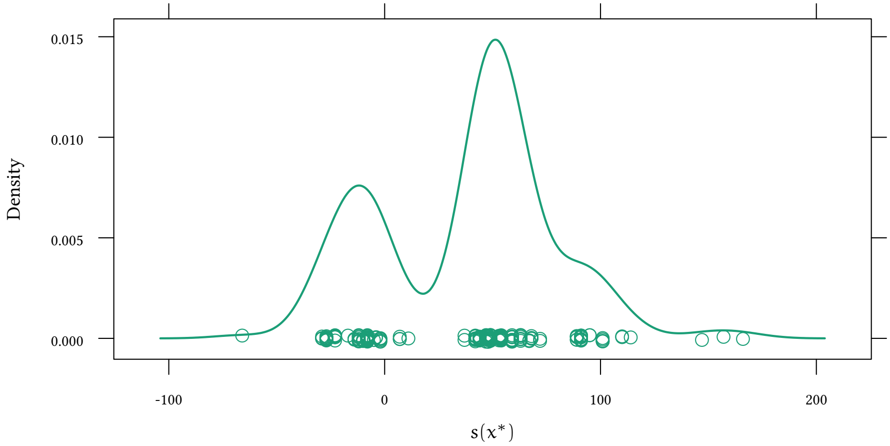
Our *estimate of the standard error of* $s(\boldsymbol{x})$ is given by the *standard deviation of the bootstrap replications* $s(\boldsymbol{x}^{*1}), s(\boldsymbol{x}^{*2}), \ldots, s(\boldsymbol{x}^{*B})$.

$$\hat{\sigma}_{s,BS} = \mathrm{se}_{b=1}^{B}(s(x^{*b})) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(s(x^{*b}) - \frac{1}{B}\sum_{b=1}^{B} s(x^{*b})\right)^2}$$

```
mouse.t

[1]  94 197  16  38  99 141  23

sample(mouse.t)

[1]  23  16 141 197  38  99  94

sample(mouse.t,replace=TRUE)

[1]  99  38 141 141  94 197  16

sample(mouse.t,replace=TRUE)

[1] 99 16 16 94 38 94 94

sample(mouse.t,replace=TRUE)

[1]  99  16 197  23 197  94 141
```

```
median(sample(mouse.t,replace=TRUE))-
    median(sample(mouse.c,replace=TRUE))

[1] -8

median(sample(mouse.t,replace=TRUE))-
    median(sample(mouse.c,replace=TRUE))

[1] 53

median(sample(mouse.t,replace=TRUE))-
    median(sample(mouse.c,replace=TRUE))

[1] 95

replicate(20,median(sample(mouse.t,replace=TRUE))-
             median(sample(mouse.c,replace=TRUE)))

 [1]  48 166 -17  67  -2  53  48  -8  53  44  44  44 -23  42 166 -14
[17]   7  -8  -4  -2
```

```
densityplot(replicate(200,median(sample(mouse.t,replace=TRUE))-
             median(sample(mouse.c,replace=TRUE))),xlab="$s(x^*)$")
```



```
set.seed(123)
```

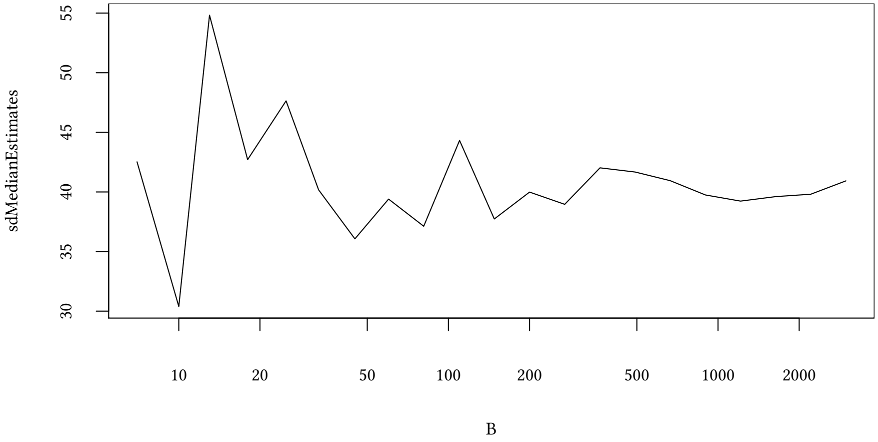
```
(sdDiff2 <- sd(replicate(100,mean(sample(mouse.t,replace=TRUE))-
                         mean(sample(mouse.c,replace=TRUE)))))

[1] 27.57955
```

```
(sdMedianDiff2 <- sd(replicate(100,median(sample(mouse.t,replace=TRUE))-
                                    median(sample(mouse.c,replace=TRUE)))))

[1] 39.22875
```

Is 100 large enough? Try different sizes of the bootstrap:

```
sdMedian <- function(B) sd(replicate(B,median(sample(mouse.t,replace=TRUE))-
                                         median(sample(mouse.c,replace=TRUE))))
set.seed(123)
(B<-round(exp(seq(2,8,.3))))

 [1]    7   10   13   18   25   33   45   60   81  110  148  200  270
[14]  365  493  665  898 1212 1636 2208 2981

sdMedianEstimates <- sapply(B,sdMedian)
```

```
plot(sdMedianEstimates ~ B,log="x",t="l")
```



- A number of 50 ...200 bootstraps is usually sufficient to estimate standard errors.

→ we increase computational complexity by a factor of 50 ...200.

→ we do not have to worry about “unusual” statistics or distributions.

We replace complicated derivations based on specific distributions with computing power.

5.2 Approximate permutation test

- Situation:
  - X may or may not influence Y.
  - Null hypothesis: X plays no role (values of X are exchangable).
  - A test statistic t which captures the influence.
- Idea: Compare the test statistic t from the sample with the distribution under the Null. Create the distribution under the Null through permutations of X (but not Y).

```
(all.mice<-data.frame(rbind(cbind(X=1,Y=mouse.t),
                            cbind(X=0,Y=mouse.c))))

   X   Y
1  1  94
2  1 197
3  1  16
4  1  38
5  1  99
6  1 141
7  1  23
8  0  52
9  0 104
10 0 146
11 0  10
12 0  50
13 0  31
14 0  40
15 0  27
16 0  46
```

```
with(all.mice,aggregate(Y~X,FUN=median))

  X  Y
1 0 46
2 1 94

testStat<-function(data)
    c(with(data,aggregate(Y~X,
                          FUN=median))[["Y"]]
      %*% c(-1,1))
testStat(all.mice)

[1] 48
```

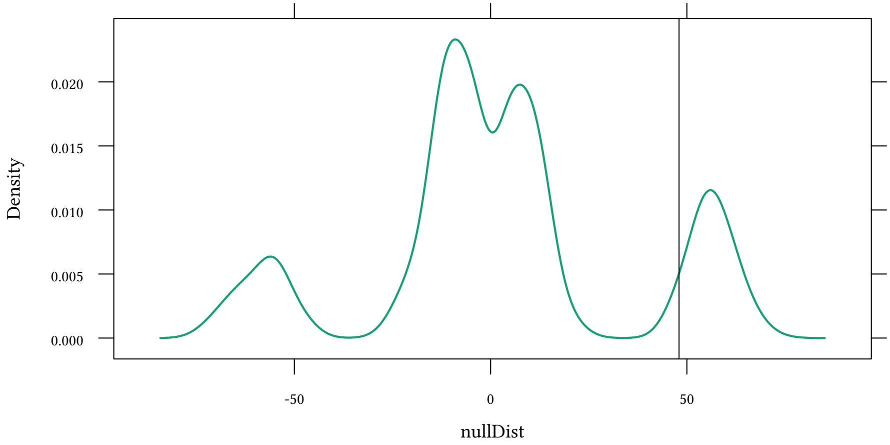
```
all.mice

   X   Y
```

```
1   1  94
2   1 197
3   1  16
4   1  38
5   1  99
6   1 141
7   1  23
8   0  52
9   0 104
10  0 146
11  0  10
12  0  50
13  0  31
14  0  40
15  0  27
16  0  46
```

Now we resample:

```
within(all.mice,X<-sample(X))

   X   Y
1  1  94
2  1 197
3  0  16
4  0  38
5  0  99
6  0 141
7  0  23
8  1  52
9  1 104
10 1 146
11 0  10
12 1  50
13 0  31
14 0  40
15 0  27
16 1  46
```

```
nullDist<-replicate(1000,
   testStat(within(all.mice,X<-sample(X))))
densityplot(nullDist,plot.points=FALSE,panel=function(...) {
        panel.densityplot(...);panel.abline(v=testStat(all.mice))})
```



```
testStat(all.mice)

[1] 48

mean(nullDist >= testStat(all.mice) )

[1] 0.187

mean(abs(nullDist) >= abs(testStat(all.mice)))

[1] 0.306
```

- Non-parametric test
- Can be constructed always when H_0 is that X does not affect Y (observations are exchangable).
- Exist for any test statistic (test statistic can be chosen efficiently)

Example: Fisher's exact test

5.3 Parameters, distributions and the plug-in principle

F is a distribution of X iff

$$F(x, \theta, \ldots) \equiv \Pr(X < x|\theta, \ldots)$$

We call θ a *parameter* of F.

Parameters and statistics:

$$\theta = t(F) \qquad \text{with } t(\cdot) \text{ being a } \textit{statistic} \text{ of } F$$

The plug-in principle

$$\hat\theta = t(\hat F)$$

We (simply) apply the statistic $t(\cdot)$ to obtain an *estimator* for θ.

Convergence Why does it make sense to use a plug-in estimator?

Glivenko-Cantelli's theorem (1933)

$$\sup_x |\hat F_n(x) - F(x)| \overset{n\to\infty}{\to} 0 \qquad \text{almost surely}$$

if $t(\hat F)$ is continuous then

$$t(\hat F_n) \overset{n\to\infty}{\to} t(F) = \theta \qquad \text{almost surely}$$

We can use the bootstrap to measure...

- ...the bias of the plug-in estimate
- ...the standard error of the plug-in estimate

5.3.1 The bootstrap algorithm for standard errors

The bootstrap algorithm for standard errors

1. Draw B independent bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$.

   Each sample has size n and is drawn with replacement from $x_1, x_2, \ldots, x_n$.

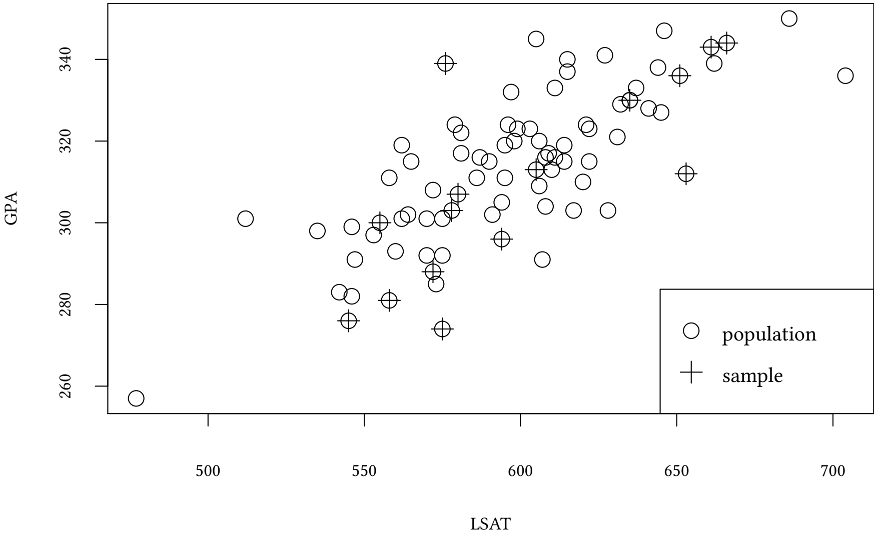
2. For bootstrap sample $b \in \{1, 2, \ldots, B\}$ determine $\hat\theta^{*b} = s(\mathbf{x}^{*b})$.

3. $\hat\sigma_{s,BS} = \mathrm{se}_{b=1}^{B}(\hat\theta^{*b}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat\theta^{*b} - \frac{1}{B}\sum_{b=1}^{B}\hat\theta^{*b}\right)^2}$

$$\lim_{B\to\infty} \hat\sigma_{s,BS} = \underbrace{\mathrm{se}_{\hat F}(\hat\theta^*)}_{\text{ideal bootstrap}}$$

5.3.2 Sampling multivariate data

```
with(law82,plot(LSAT,100*GPA,ylab="GPA"))
with(law,points(LSAT,100*GPA,pch=3))
legend("bottomright",c("population","sample"),pch=c(1,3))
```



Our universe are 82 law schools.

Our population contains data about 82 law schools (dataset *law82*).

Our sample (`law`) only contains 15 observations.

We are interested in the *correlation* of GPA (Grade Point Average, undergraduate score) and LSAT (Law School Admission Test Score).

(i.e. the θ and the $t(\cdot)$ we are talking about is the correlation.)
The true correlation:

```
with(law82,cor(GPA,LSAT))

[1] 0.7599979
```

The plug-in estimate:

```
(lawCorEst <- with(law,cor(GPA,LSAT)))

[1] 0.7763745
```

If F is bivariate normal, then the estimated correlation coefficient $\hat{\rho}$ has a standard deviation of about

$$\hat{\sigma}_{\text{normal}} = \frac{1-\hat{\rho}^2}{\sqrt{n-3}} \approx 0.115$$

With a bootstrap we can live *without* the assumption that F is normal.

```
law

   LSAT  GPA
1   576 3.39
2   635 3.30
3   558 2.81
4   578 3.03
5   666 3.44
6   580 3.07
7   555 3.00
8   661 3.43
9   651 3.36
10  605 3.13
11  653 3.12
12  575 2.74
13  545 2.76
14  572 2.88
15  594 2.96
```

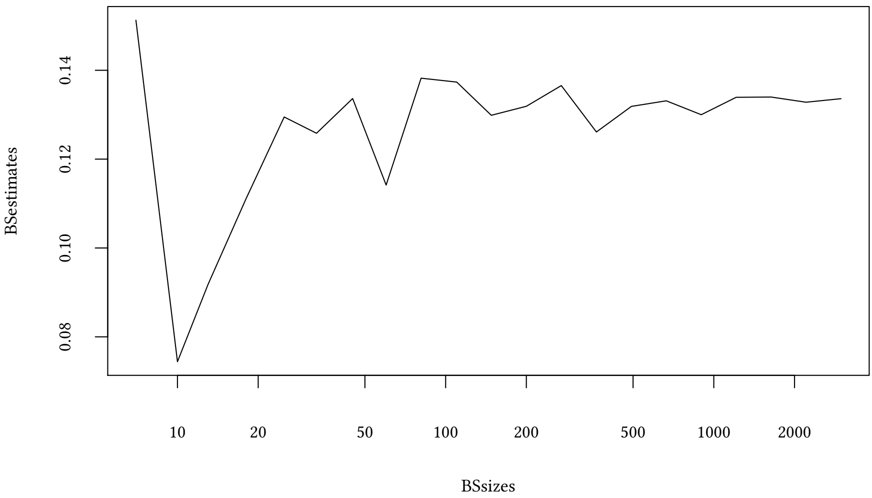
```
set.seed(123)
samplesize<-nrow(law)
ind<-1:samplesize
(i2<-sample(ind,replace=TRUE))

 [1] 15 15  3 14  3 10  2  6 11  5  4 14  6  9 10

law[i2,]

     LSAT  GPA
15    594 2.96
15.1  594 2.96
3     558 2.81
14    572 2.88
3.1   558 2.81
10    605 3.13
2     635 3.30
6     580 3.07
11    653 3.12
5     666 3.44
4     578 3.03
14.1  572 2.88
6.1   580 3.07
9     651 3.36
10.1  605 3.13
```

```
law.boot <- replicate(200,{i2 <- sample(ind,replace=TRUE);
                           with(law[i2,],cor(GPA,LSAT))})
sd(law.boot)

[1] 0.1197215
```

How quickly does the estimate of $\hat\sigma$ converge?

```
set.seed(123)
lawBS <- function(B) sd(replicate(B,{i2 <- sample(ind,replace=TRUE);
                                     with(law[i2,],cor(GPA,LSAT))}))
(BSsizes<-round(exp(seq(2,8,.3))))

 [1]    7   10   13   18   25   33   45   60   81  110  148  200  270
[14]  365  493  665  898 1212 1636 2208 2981

BSestimates <- sapply(BSsizes,lawBS)
```

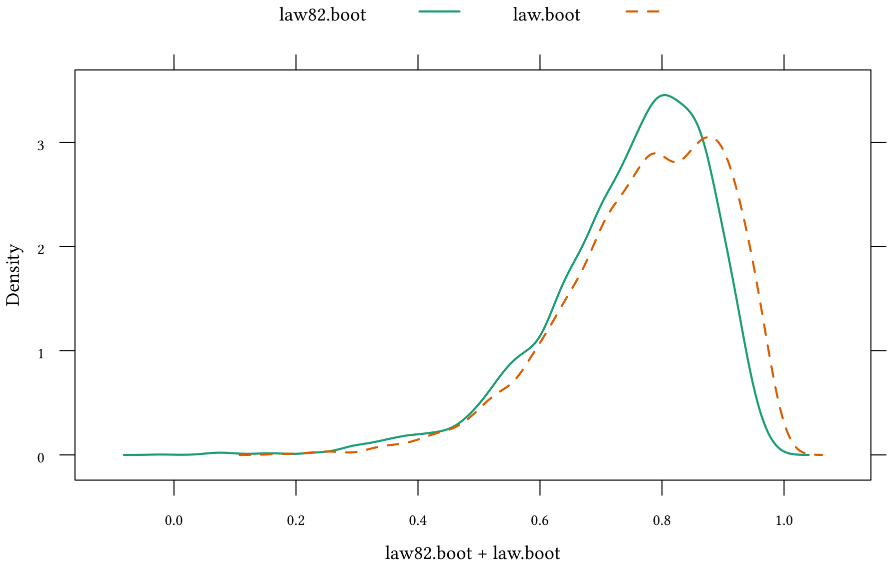
```
plot(BSestimates ~ BSsizes,log="x",t="l")
```



The distribution of $\hat\theta^*$ What is the distribution of the bootstrap replications $\hat F(\hat\theta^*)$ and how does this compare to $F(\hat\theta^*)$?

```
set.seed(123)
ind<-1:dim(law)[1]
law.boot <- replicate(5000,{i2 <- sample(ind,size=samplesize,replace=TRUE);
                            with(law[i2,],cor(GPA,LSAT))})
ind<-1:dim(law82)[1]
law82.boot <- replicate(5000,{i2 <- sample(ind,size=samplesize,replace=TRUE);
                              with(law82[i2,],cor(GPA,LSAT))})
```

```
densityplot(~ law82.boot + law.boot,plot.points=FALSE,
            auto.key=list(columns=2,size=3,between=1))
```



Comparison of the bootstrap estimate $\mathrm{se}_{\hat{F}}(\hat{\sigma})$ with the standard error $\mathrm{se}_F(\hat{\sigma})$:

```
sd(law.boot)

[1] 0.1323102

sd(law82.boot)

[1] 0.1317328
```

5.4 The parametric bootstrap

Now we assume that we know a lot about F, e.g. that F follows a (multivariate) normal distribution.

- $x_1, x_2, \ldots, x_n$ are n *observed values.*
- estimate the parameters of F (e.g. mean, variance, covariance in the case of a normal distribution) and obtain F_{par}.
- Apply the bootstrap algorithm for standard errors to F_{par}.

```
library(MASS)
paraBoot <- function(XY,B)  {
  Sigma <- cov(XY)
  mu <- rapply(XY,mean)
  mvrnorm(n=B,mu=mu,Sigma=Sigma)
}
```

```
cor(paraBoot(law,samplesize))
```

```
          LSAT       GPA
LSAT 1.0000000 0.4829548
GPA  0.4829548 1.0000000
```

```
cor(paraBoot(law,samplesize))[2,1]
```

```
[1] 0.4829548
```

```
paraBoot(law,samplesize)
```

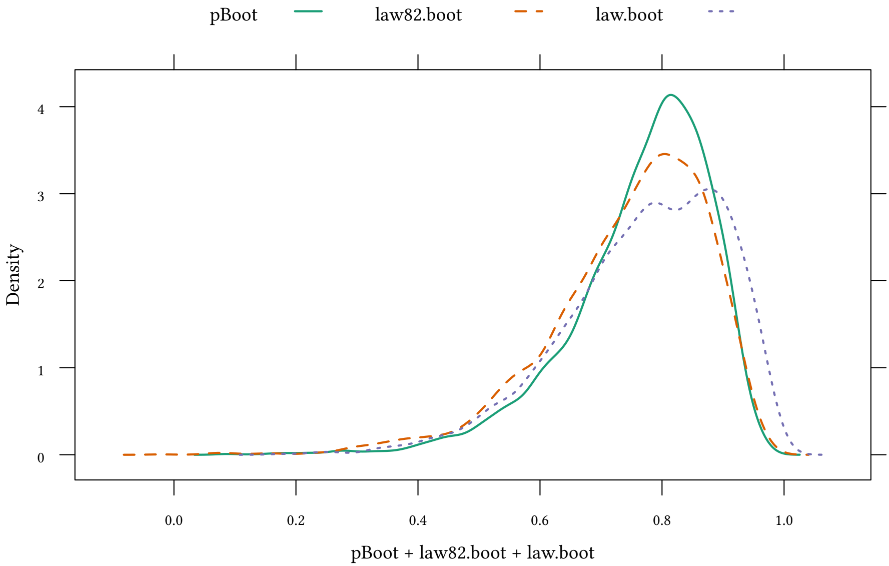
```
          LSAT      GPA
 [1,] 623.6927 2.926382
 [2,] 609.8872 3.061776
 [3,] 535.1199 3.101807
 [4,] 597.3203 2.973695
 [5,] 594.8628 3.142786
 [6,] 528.5856 2.934303
 [7,] 581.0027 3.040980
 [8,] 653.1386 3.491305
 [9,] 628.9728 3.336389
[10,] 618.8925 3.274851
[11,] 549.1056 3.122109
[12,] 585.2290 2.898061
[13,] 583.5167 2.995358
[14,] 595.6399 3.248417
[15,] 623.4986 3.007323
```

```
pBoot <- replicate(5000,cor(
   paraBoot(law,samplesize))[2,1])
head(pBoot)
```

```
[1] 0.4829548 0.7598232 0.8823887 0.8611649 0.6253162 0.9020786
```

Example: The law school data again

```
densityplot(~pBoot+law82.boot + law.boot,plot.points=FALSE,
            auto.key=list(columns=3,size=2,between=1))
```



And what is the parametric bootstrap estimation of the standard deviation?

```
sd(pBoot)

[1] 0.1168818
```

Remember:

$$\hat{\sigma}_{\text{normal}} = \frac{1-\hat{\rho}^2}{\sqrt{n-3}} \approx 0.115$$

```
sd(law82.boot)
sd(law.boot)
```

5.5 Literature

- Bradley Efron and Robert J. Tibshirani. “An Introduction to the Bootstrap”. Chapman & Hall, 1994.
- A. C. Davison, D. V. Hinkley. “Bootstrap Methods and their Application”. Cambridge University Press, 1997.
- William H. Greene. “Econometric Analysis”. 2012. Chapter 15.

Appendix 5.A Examples for the lecture

Example 1 The dataset `x1.csv` contains two variables, `x1` and `x2`.

- What is the interquartile range of `x1`?
- What is the standard deviation of your estimate?
- Draw a graph with the number of bootstrap replications on the horizontal axis and the estimated standard deviation on the vertial axis. How quick does your estimation converge?

Example 2 Look again at the dataset `x1.csv`.

- What is the average value of the ratio `x1/x2`.
- What is the standard deviation we would obtain with the standard formula?
- Use a bootstrap to estimate the standard deviation.
- How does the standard deviation converge? Show a graph!

Example 3 Have another look at the previous exercise. What result would you obtain with a parametric bootstrap?

Appendix 5.B Exercises

To solve the next exercises the following commands for R might help: `data, str, names, length, dim, plot, density, median, summary, sd, quantile, cor, set.seed, sample, replicate, rnorm, lm, vcov, coeftest`. Note that `coeftest` is provided byt the `car` library.

Exercise 5.1 *Consider the dataset `Caschool` from the package `Ecdat`.*

- *What is the median value of `testscr`?*
- *Give a standard deviation for the median value.*
- *Give a 95%-confidence interval.*
- *Use a parametric bootstrap.*

Exercise 5.2 *Consider the dataset `Caschool` from the package `Ecdat`. When you explain `testscr` as a function of `str` you obtain a regression coefficient $\widehat{\beta}$ for `str`. Call the standard deviation of this coefficient $\hat{\sigma}$.*

- *What is $\hat{\sigma}$.*

- *What is the standard deviation of $\hat{\sigma}$?*
- *What is the ratio $\hat{\beta}/\hat{\sigma}$?*
- *What is the standard deviation of this ratio?*
- *Give a 95%-confidence interval for this ratio.*
- *Now use a parametric bootstrap.*

Exercise 5.3 *You still consider the dataset* `Caschool` *from the package* `Ecdat`*. When you explain* `testscr` *as a function of* `str` *you obtain a regression coefficient $\hat{\beta}$ for* `str`*. Your Null hypothesis is $\beta = 0$.*

- *What is the p-value from an approximate permutation test for this hypothesis?*
- *You now include* `elpct` *as a regressor. How does your p-value change?*
- *Use an approximate permutation test to test whether the Spearman rank correlation coefficient of* `str` *and* `testscr` *is zero.*

6 Bayesian methods

6.1 Motivation

- Compare Bayesian with frequentist methods.

  Two schools of statistical inference: Bayesian / Frequentist

  - Frequentist: Standard hypothesis testing, p-values, confidence intervals. Well known.
  - Bayesian: beliefs conditional on data.

- Learn to apply Bayesian methods.
  - What is the equivalent of frequentist method X in the Bayesian world?
  - How to put Bayesian methods into practice?

Frequentist: Null Hypothesis Significance Testing (Ronald A. Fisher, Statistical Methods for Research Workers, 1925, p. 43)

- $X \leftarrow \theta$, X is random, θ is fixed.
- Confidence intervals and p-values are easy to calculate.
- Interpretation of confidence intervals and p-values is awkward.

- p-values depend on the intention of the researcher.
- We can test “Null-hypotheses” (but where do these Null-hypotheses come from).
- Not good at accumulating knowledge.
- More restrictive modelling.

Bayesian: (Thomas Bayes, 1702-1761; Metropolis et al., “Equations of State Calculations by Fast Computing Machines”. Journal of Chemical Physics, 1953.)

- $X \to \theta$, X is fixed, θ is random.
- Requires more computational effort.
- “Credible intervals” are easier to interpret.
- Can work with “uninformed priors” (similar results as with frequentist statistics)
- Efficient at accumulating knowledge.
- Flexible modelling.

Most people are still used to the frequentist approach. Although the Bayesian approach might have clear advantages it is important that we are able to understand research that is done in the context of the frequentist approach.

$$\Pr(A \wedge B) = \Pr(A) \cdot \Pr(B|A) = \Pr(B) \cdot \Pr(A|B)$$

$$\text{rewrite:} \qquad \Pr(A) \cdot \Pr(B|A) \frac{1}{\Pr(B)} = \Pr(A|B)$$

$$\text{with } A = \underbrace{\theta}_{\text{parameter}} \text{ and } B = \underbrace{X}_{\text{data}}:$$

$$\underbrace{\Pr(\theta)}_{\text{prior}} \cdot \underbrace{\Pr(X|\theta)}_{\text{likelihood}} \cdot \frac{1}{\underbrace{\Pr(X)}_{\int \Pr(\theta)\Pr(X|\theta)\, d\theta}} = \underbrace{\Pr(\theta|X)}_{\text{posterior}}$$

Before we come to a more formal comparison, let us compare the two approaches, frequentist versus Bayesian, with the help of an example.

I will use an example from the legal profession. Courts have to decide whether a defendant is guilty or innocent. Scientists have to decide whether a hypothesis is correct or not correct. Statistically, in both cases we are talking about the value of a parameter. $\theta =$ guilty or $\theta =$ not guilty. Alternatively, $\beta = 0$ or $\beta \neq 0$.

My hope is that the legal context makes it more obvious how the decision process fails or succeeds.

The prosecutors' fallacy

Assuming that the prior probability of a random match ($\Pr(X|\theta_0)$) is equal to the probability that the defendant is innocent ($\Pr(\theta_0|X)$).

Two problems:

- p-values depend on the researcher's intention. E.g. multiple testing (several suspects, perhaps the entire population, is "tested", only one suspect is brought to trial)
- Conditional probability (neglecting prior probabilities of the crime)
- Lucia de Berk:
  - $\Pr(\text{evidence}|\text{not guilty}) = 1/342$ million
  - $\Pr(\text{evidence}|\text{not guilty}) = 1/25$
- Sally Clark
  - $\Pr(\text{evidence}|\text{not guilty}) = 1/73$ million
  - $\Pr(\text{not guilty}|\text{evidence}) = 78\%$

The Sally Clark case

- 1996: First child dies from SIDS (sudden infant death syndrome): $P = 1/8543$
- 1998: Second child dies from SIDS: $P = 1/8543$
- $\to$: $\Pr(\text{evidence}|\text{not guilty}) = (1/8543)^2 \approx 1/73$ million
- 1999: life imprisonment, upheld at appeal in 2000.

Problems:

- Correlation of SIDS within a family. $\Pr(\text{2nd child}) = (1/8543) \times 5 \ldots 10$
- SIDS is actually more likely in this case: $P = 1/8543 \to P = 1/1300$

  $\Pr(\text{evidence}|1 \text{ not guilty mother}) = 1/(1300 \cdot 130) = 0.000592\ \%$
- Intention of the researcher/multiple testing: $\approx$ 750 000 births in England and Wales / year. How likely is it to find two successive SIDS or more among 750 000 mothers.

  $\Pr(\text{evidence}|750\,000 \text{ not guilty mothers}) = 98.8\ \%$.

But what is the (posterior) probability of *guilt*? Here we need prior information.

- What is the prior probability of a mother murdering her child?

$$\underbrace{\Pr(\theta)}_{\text{prior}} \cdot \underbrace{\Pr(X|\theta)}_{\text{likelihood}} \cdot \frac{1}{\Pr(X)} = \underbrace{\Pr(\theta|X)}_{\text{posterior}}$$

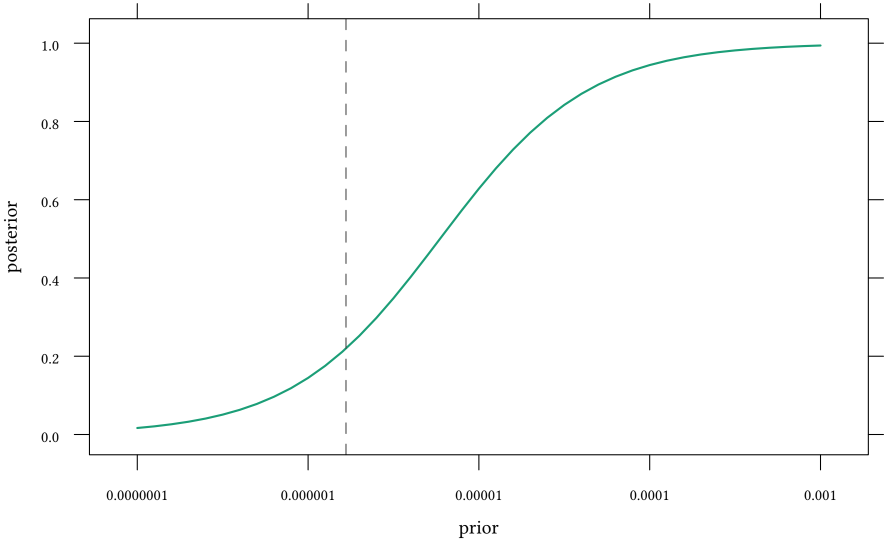
$$\underbrace{\Pr(g)}_{\text{prior}} \cdot \underbrace{\Pr(X|g)}_{\text{likelihood}} \cdot \frac{1}{\underbrace{\Pr(g)\cdot\Pr(X|g)+(1-\Pr(g))\cdot\Pr(X|\text{not g})}_{\Pr(X)}} = \underbrace{\Pr(g|X)}_{\text{posterior}}$$

Data from the U.S.A. (Miller, Oberman, 2004): per 600 000 mothers 1 killed child, $\Pr(g) = 1/600\,000$.

$\Pr(X|g) = 1$, $\Pr(X) = \underbrace{\frac{1}{600\,000}}_{\text{guilty}} + \underbrace{\frac{599\,999}{600\,000}\cdot\frac{1}{1300\cdot 130}}_{\text{not guilty}}$

$$\Pr(g|\text{evidence}) = 22\%$$

If $\Pr(g) = 1/18800$ then $\Pr(g|\text{evidence}) = 90\%$
If $\Pr(g) = 1/1710$ then $\Pr(g|\text{evidence}) = 99\%$



$$\Pr(X|\theta) \quad \neq \quad \Pr(\theta|X)$$

- The interpretation of $\Pr(X|\theta)$ as a p-value is affected by the intention of the researcher (e.g. multiple testing).

→ The same X could have been collected with different intentions → different p-values.

- $\Pr(\theta|X)$ is not affected by the intention of the researcher (e.g. by multiple testing).

- $\Pr(\theta|X)$ forces us to think about a (subjective) *prior*.

6.2 The intention of the researcher — p-hacking

X data

ϕ_j test procedure

- choice of control variables
- data exclusion
- coding
- analysis
- interactions
- predictors
- $\vdots$

$T(X, \phi_j)$ test result

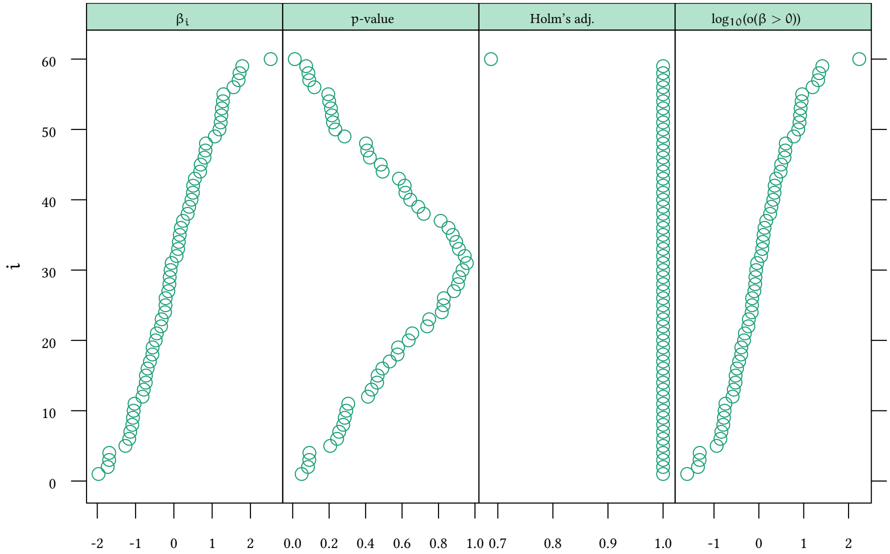
p-hacking

- perform J tests: $\{\ldots, T(X, \phi_j), \ldots\}$
- report the best result, given the data: $T(X, \phi_{\text{best}})$

→ to correct for multiple testing we need to know J ↓

→ robustness checks (for all J ↓)

An example: A researcher uses 60 explanatory variables to explain one dependent variable. Here we assume (for simplicity) that they all have the same standard error $\sigma = 1$.



smallest p-value:	no correction	$p = 0.011$
	Holm's adjustment	$p = 0.69$

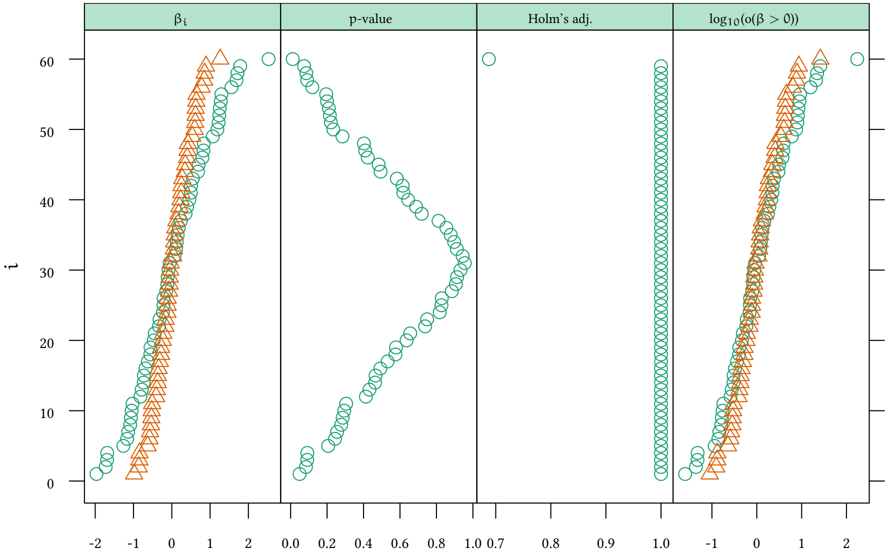
A statement about the p-value depends on the intention of the researcher. It is affected by multiple testing.

A statement about the posterior odds does not depend on the intention of the researcher. It does, though, depend, on a prior.

Above we assumed a flat prior. Is this reasonable? Perhaps, if we have already studied dozens of these variables, and they all seem to be drawn from a distribution with $\mu = 0$ and $\sigma = 1$, it is no longer reasonable to have a flat prior.

Above we pretended to be ignorant. We used a flat prior in each study.

Now we use a prior for β: $\beta_i \sim N(0, 1)$



largest odds: flat prior $\beta_i > 0/\beta_i < 0$ odds=170 : 1

informed prior $\beta_i > 0/\beta_i < 0$ odds=26 : 1

Pretending to be ignorant and assuming a flat prior can be misleading.

- Flat prior in the Sally Clark case:

  $\Pr(\text{guilt}) : \Pr(\text{innocence}) = \frac{1}{2} : \frac{1}{2}$.

  This is absurd.

- Also $H_0 : \forall_i \beta_i = 0$ could be absurd.

6.3 Probabilities

Consider the following statements:

Frequentist probability

- The probability to throw two times a six is 1/36.
- The probability to win the state lottery is about 1:175 000 000.
- The probability of rainfall on a given day in August is 1/3.
- The probability that a male develops lung or bronchus cancer is 7.43%.

Subjective probability

- The probability of rainfall tomorrow is 1/3.
- The probability that a Mr. Smith develops lung or bronchus cancer is 7.43%.
- The probability that Ms. X commited a crime is 20%.

Frequentist

- $P =$ objective probability (sampling of the data X is infinite).

→ but what if the event occurs only once (rainfall tomorrow, Mr. Smith's health,...)?

  → von Mises: event has no probability

  → Popper: invent a fictitious population from which the event is a random sample (propensity probability).

- Parameters θ are unknown but fixed during repeated sampling.

Bayesian

- $P =$ subjective probability of an event (de Finetti/Ramsey/Savage)

  ≈ betting quotients

- Parameters θ follow a (subjective) distribution.

Fixed quantities:

Frequentist

- Parameters θ are fixed (but unknown).

Bayesian

- Data X are fixed.

Probabilistic statements:

Frequentist

- ...about the frequency of errors p.
- Data X are a random sample and could potentially be resampled infinitely often.

Bayesian

- ...about the distribution of parameters θ.

6.4 Decision making

Which decision rule, Bayesian or frequentist, uses information more efficiently?

$$\Pr(\theta) \cdot \Pr(X|\theta) \cdot \frac{1}{\Pr(X)} = \Pr(\theta|X)$$

Assume $\theta \in \{0,1\}$. Implement an action $a \in \{0,1\}$. Payoffs are $\pi_{a\theta}$.

We have $\pi_{11} > \pi_{01}$ and $\pi_{00} > \pi_{10}$, i.e. it is better to choose $a = \theta$. Expected payoffs:

$$E(\pi|a) = \pi_{a1} \cdot \Pr(\theta = 1) + \pi_{a0} \cdot \Pr(\theta = 0)$$

Optimal decision: choose $a = 1$ iff

$$\underbrace{\pi_{11} \cdot \Pr(\theta = 1) + \pi_{10} \cdot \Pr(\theta = 0)}_{E(\pi|a=1)} > \underbrace{\pi_{01} \cdot \Pr(\theta = 1) + \pi_{00} \cdot \Pr(\theta = 0)}_{E(\pi|a=0)} .$$

Rearrange: choose $a = 1$ iff

$$\Pr(\theta = 1)\underbrace{(\pi_{11} - \pi_{01})}_{g_1} > \Pr(\theta = 0)\underbrace{(\pi_{00} - \pi_{10})}_{g_0} .$$

Here g_a can be seen as the gain from choosing the correct action (or the loss from choosing the wrong action) if $\theta = a$.

If we have some data X:

$$\Pr(\theta = 1|X)g_1 > \Pr(\theta = 0|X)g_0.$$

Bayes' rule:

$$\Pr(\theta) \cdot \Pr(X|\theta) \cdot \frac{1}{\Pr(X)} = \Pr(\theta|X)$$

$$\text{choose } a = 1 \text{ iff } \frac{g_1}{g_0} > \frac{\Pr(\theta = 0|X)}{\Pr(\theta = 1|X)} = \frac{\frac{\Pr(\theta=0)\cdot\Pr(X|\theta=0)}{\Pr(X)}}{\frac{\Pr(\theta=1)\cdot\Pr(X|\theta=1)}{\Pr(X)}} = \frac{\Pr(\theta = 0) \cdot \Pr(X|\theta = 0)}{\Pr(\theta = 1) \cdot \Pr(X|\theta = 1)}$$

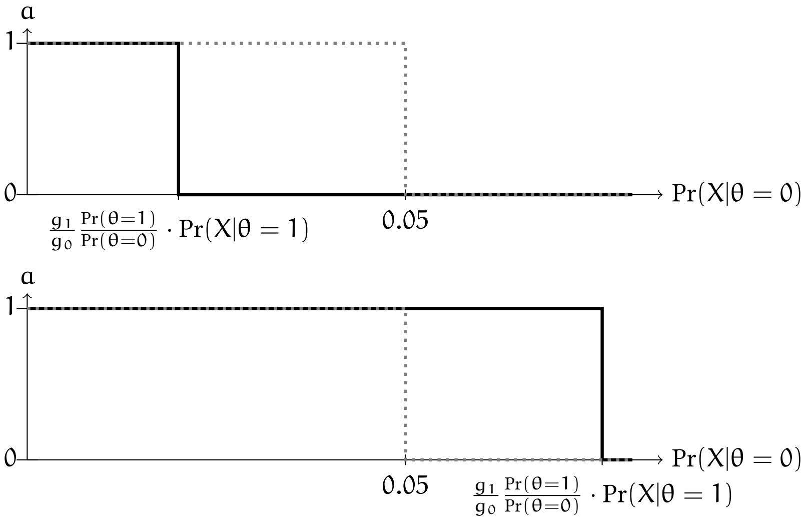
$$\text{choose } a = 1 \text{ iff } \frac{g_1}{g_0}\frac{\Pr(\theta = 1)}{\Pr(\theta = 0)} \cdot \Pr(X|\theta = 1) > \Pr(X|\theta = 0)$$

$$\text{Bayesian chooses } a = 1 \text{ iff } \Pr(X|\theta = 0) < \frac{g_1}{g_0}\frac{\Pr(\theta = 1)}{\Pr(\theta = 0)} \cdot \Pr(X|\theta = 1)$$

$$\text{Frequentist chooses } a = 1 \text{ iff } \Pr(X|\theta = 0) < 0.05 .$$

(Here we assume that H_0 is $\theta = 0$.)

When do Bayesians and Frequentists disagree?



For very small and for very large values of $\Pr(X|\theta=0)$ both Bayesians and frequentists make the same choice. Only in the range between $\frac{g_1}{g_0}\frac{\Pr(\theta=1)}{\Pr(\theta=0)}\cdot\Pr(X|\theta=1)$ and 0.05 choices differ. In that range the Bayesian choice maximises expected payoffs while the frequentist does not.

6.5 Prior information

- Prior research (published / unpublished)
- Intuition (of researcher / audience)
- Convenience (conjugate priors, vague priors).

Prior information is *not* the statistician's personal opinion. Prior information is the result of and subject to scientific debate.

6.6 Objectivity and subjectivity

- Bayesian decision making requires assumptions about...
  - $\Pr(\theta)$ (prior information)
  - g_0, g_0 (cost and benefits)

  Scientists might disagree about this information.

→ Bayesian decision making is therefore accused of being "subjective".

Bayesian's might "choose" priors, cost and benefits, to subjectively determine the result. E.g. in the Sally Clark case, the researcher might "choose" the prior probability of a mother to kill her child to be $1/1710$ to conclude guilt with $\Pr(g|\text{evidence}) = 99\%$.

The Bayesian's answer:

- Prior information, cost and benefits are relevant information. Disregarding them (as the frequentists do) is a strange concept of "objectivity".
- Priors, cost and benefits are subject to scientific debate, like any other assumption. We have to talk about priors, not assume them away.
- Subjectivity exists in both worlds:
  - B.+F. make assumptions about the model → more dangerous than priors.
  - In F. the intention of the researcher has a major influence on p-values and confidence intervals.

6.7 Issues

- Probability: frequentist vs. subjective.
- Prior information, how to obtain?
- Results, objective / subjective.
- Flexible modelling: F. has only a limited number of models.

  F: precise method, using a tool which is sometimes not such a good representation of the problem.

  B: approximate method, using a tool which can give a more precise representation of the problem.
- Interpretation: p-values versus posteriors.

  B. predicts (posterior) probability of a hypothesis.

  F. writes carefully worded statements which are wrong 5% of the time (or any other probability) provided H_0 is true.
- Quality of decisions: p-values are only a heuristic for a decision rule.

  B.'s decisions are better in expectation.

6.8 Literature

- John H. Kruschke. “Doing Bayesian Data Analysis. A Tutorial with R, JAGS, and Stan”. Academic Press. 2014.
- Hoff, A First Course in Bayesian Statistical Methods.
- C. Andrieu, de Freitas, N., Doucet, A, Jordan, M. “An Introduction to MCMC for Machine Learning.” Machine Learning. 2003, 50(1-2), pp 5-43. (To learn more about MCMC sampling)
- William H. Greene. “Econometric Analysis”. Chapter 16.

Appendix 6.A Examples for the lecture

A researcher is measuring the performance P_i of n different market mechanisms $i \in \{1, \ldots, n\}$. For each mechanism i she plans to run several experiments in the lab. To compare the performance of each mechanism with a known and constant reference performance $\bar{P}$ she plans to use a t-test for each mechanism i, testing always the alternative hypothesis $P_i \neq \bar{P}$. Her plan is to report only those results which are significant. After running the experiments she finds that only one mechanism, mechanism ζ, has a p-value smaller than 5%. She concludes that she can reject $P_\zeta = \bar{P}$. What can you say about the type 1 error of this procedure?

- The type 1 error is larger than the p-value reported by the t-test
- The type 1 error is identical to the p-value reported by the t-test
- The type 1 error is independent of n
- The type 1 error increases with n
- None of the above

Consider the following problem: A taxi was invoveld in an accident. This city has 100 taxis. 85 taxis belong to the green company. 15 taxis belong to the blue company. An eyewitness claims the taxi was blue. However, this eyewitness is wong 20% of the time.

What is the probability that the taxi in the accident was blue?

(Tversky and Kahneman, 1982)

Appendix 6.B Exercises

To solve the next exercises the following commands for R might help: *`curve`*, *`integrate`*, *`dnorm`*, *`dbeta`*, *`function`*, *`sapply`*, *`choose`*, *`sum`*, *`prod`*.

Exercise 6.1 *Approximately 1/125 of all births are fraternal twins and 1/300 of births are identical twins. Elvis Presley had a twin brother (who died at birth).*

What is the probability that Elvis was an identical twin? (You may approximate the probability of a boy or girl birth as $\frac{1}{2}$.)

(This exercise is from Gelman et al.'s "Bayesian Data Analysis", 3rd edition, 1.12, exercise 6.)

Exercise 6.2 *Suppose that if $\theta = 1$, then y has a normal distribution with mean 1 and standard deviation σ, and if $\theta = 2$, then y has a normal distribution with mean 2 and standard deviation σ. Also, suppose $\Pr(\theta = 1) = 0.5$ and $\Pr(\theta = 2) = 0.5$.*

- *For $\sigma = 2$, write the formula for the marginal probability density for y and sketch it.*
- *What is $\Pr(\theta = 1|y = 1)$, again supposing $\sigma = 2$?*
- *Describe how the posterior density of θ changes in shape as σ is increased and as it is decreased.*

(This exercise is from Gelman et al.'s "Bayesian Data Analysis", 3rd edition, 1.12, exercise 1.)

Exercise 6.3 *Suppose that you toss a coin twice and observe "heads" twice.*

1. *What is the maximum likelihood estimate that the third toss will result in "heads" again?*
2. *Using the Beta(1,1) distribution for the prior, what is the Bayesian estimate that the third toss will result in "heads" again?*

Exercise 6.4 *Suppose that a student has correctly answered 9 out of 12 exam questions.*

1. *What is the probability that he was answering randomly?*
2. *What is the p-value (two-tailed) of the hypothesis that he was answering randomly? What does it mean for the respective hypothesis at 5%-significance level?*
3. *Assume that at the outset it is equally likely that the student can be answering the questions randomly and that he can be answering them in a non-random manner. Based on the data observed, which hypothesis is more likely to be true?*

Exercise 6.5 *With the only difference between the studies being the number of observations, which of the two provides more evidence against the null hypothesis?*

1. *n = 10, p value = 0.032.*
2. *n = 100, p value = 0.032.*

Exercise 6.6 *An event can have three possible outcomes, 1, 2, and 3. There are three theories, A, B, and C. According to theory A all events have the same probability,* $p(1) = p(2) = p(3) = 1/3$*. According to theory B we have* $p(1) = 1/2$*,* $p(2) = p(3) = 1/4$*. According to theory C we have* $p(1) = p(2) = 1/4$*,* $p(3) = 1/2$*.*

- *A priori you consider all theories equally probable. You observe 30 realisations of the event and you find 10 times 1, 10 times 2, 10 times 3. Whan can you say about posterior probabilities of A, B and C?*

- *Now assume that you have instead observed 40 realisations of the event and found 20 times 1, 10 times 2, and 10 times 3. Whan can you now say about posterior probabilities of A, B and C?*

Exercise 6.7 *You order two coins to be manufactured to physical specifications guaranteeing the probability of "heads" to be 0.55 (coin A) and 0.9 (coin B). The first coin arrives without labeling and as a test, you flip it 20 times to observe "heads" 17 times.*

1. *What is the probability of such outcome of the test?*

2. *If there is no particular reason why one coin or the other would arrive earlier, what is the probability that what you have is coin A? Which coin are you more likely to have?*

3. *If you contact the manufacturer and they are 95% sure to have shipped coin A, what is the probability that what you have is coin A? Which coin are you more likely to have?*

(This problem is taken from Chapter 1 of 'Bayesian Inference: with ecological applications' by William A. Link and Richard J. Barker, 2010.)

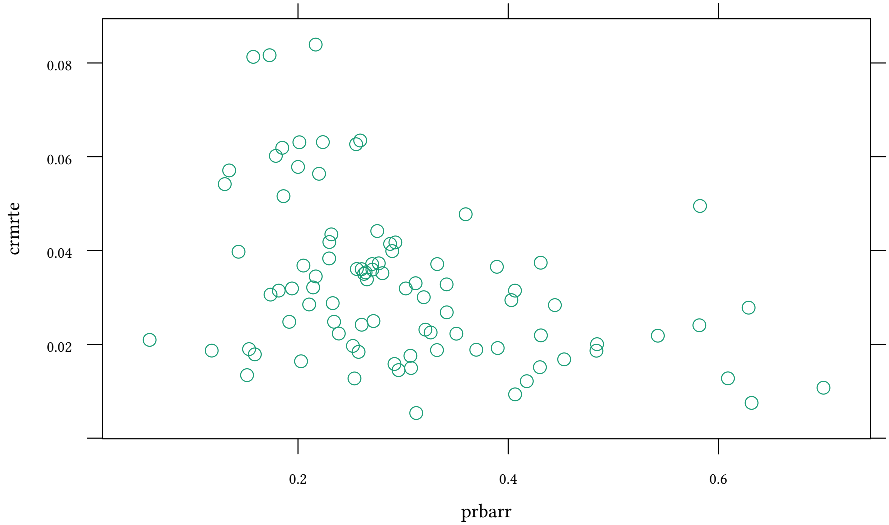
7 Bayes in practice

7.1 Example: The distribution of the population mean

Here we ask the question: "What is the probability to be arrested in North Carolina in 1981 (conditional on a crime committed)?"

Example: Crime in North Carolina counties in 1981

```
library(Ecdat)
data(Crime)
xyplot(crmrte ~ prbarr,data=Crime,subset=year==81)
```



```
y <- subset(Crime,year==81)[["prbarr"]]
```

We can have a look at a part of the data with *head*:

```
head(y)

[1] 0.289696 0.202899 0.406593 0.431095 0.631579 0.369650
```

If we suspect that average rate to be arrested to be 0.3, we use a *t.test*:

```
t.test(y,mu=.3)


One Sample t-test

data:  y
t = -0.070496, df = 89, p-value = 0.944
alternative hypothesis: true mean is not equal to 0.3
95 percent confidence interval:
 0.2724894 0.3256254
sample estimates:
mean of x
0.2990574
```

How should we interpret this result?

7.2 Conjugate Priors

$$\underbrace{\Pr(\theta)}_{\text{prior}} \cdot \underbrace{\Pr(X|\theta)}_{\text{likelihood}} \cdot \frac{1}{\underbrace{\Pr(X)}_{\int \Pr(\theta)\cdot\Pr(X|\theta)\,d\theta}} = \underbrace{\Pr(\theta|X)}_{\text{posterior}}$$

Find $\Pr(\theta|X)$:

- Exact: but $\int \Pr(\theta) \cdot \Pr(X|\theta)\, d\theta$ can be hard (except for specific priors and likelihoods).
- MCMC Sampling

  $\vdots$

Exact Bayesian inference is possible for (few) special cases.

- Present examples for exact inference.
- Present more general (approximate) method (MCMC).

Accumulating evidence Independence → Exchangability
When we accumulate data X_1 and X_2 it should not matter, whether we first observe X_1 and then add X_2 or vice versa.

Call $\mathcal{D}$ the distribution of parameter θ.

$$\begin{tikzcd}
\mathcal{D}_0 \arrow[r, "X_1"] & \mathcal{D}_1 \arrow[r, "X_2"] & \mathcal{D}_{12} \\
\mathcal{D}_0 \arrow[r, "X_2"] & \mathcal{D}_2 \arrow[r, "X_1"] & \mathcal{D}_{12}
\end{tikzcd}$$

This is easier if $\mathcal{D}_0, \mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_{12}$ belong to one family.

For a some combinations of prior distributions and likelihoods we can actually calculate analytically the posterior distribution.

Conjugate priors for a likelihood function

Likelihood	known	model parameter
$X \sim N(\mu, \tau)$	$\tau = 1/\sigma^2.$	$\mu \sim N(\mu_0, \sigma_0^2)$
$X \sim N(\mu, \tau)$	μ	$\tau \sim \Gamma(\alpha_0, \beta_0)$
$X \sim \text{bern}(p)$		$p \sim \text{Beta}(\alpha_0, \beta_0)$
$\vdots$		

If the prior model parameter follows the conjugate prior, then the posterior model parameter is in the same family.

Conjugate priors, example: normal likelihood μ

- Likelihood: $X \sim N(\mu, \sigma^2)$ with known $\tau = 1/\sigma^2$.
- Model parameter: μ
- Conjugate prior distribution: $\mu \sim N(\mu_0, \sigma_0^2)$
- Prior hyperparameter: μ_0, σ_0^2 i.e. prior $\mu \sim N(\mu_0, \sigma_0^2)$.
- Posterior hyperparameter:

$$\mu_{\text{post}} = \left(\frac{\mu_0}{\sigma_0^2} + \frac{n \cdot \bar{x}}{\sigma^2}\right) \Big/ \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right) = \frac{\tau_0 \mu_0 + n\tau\bar{x}}{\tau_0 + n\tau}$$

$$\tau_{\text{post}} = 1/\sigma_{\text{post}}^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right) = \tau_0 + n\tau$$

i.e. posterior $\mu \sim N(\mu_{\text{post}}, \sigma_{\text{post}}^2)$.

In other words:

- Prior: $\mu \sim N(\mu_0, \tau_0)$
- Likelihood: $X \sim N(\mu, \tau)$
- Posterior: $\mu \sim N(\mu_{\text{post}}, \tau_{\text{post}})$.

Terminology:

- Hyperparameters: μ_0, τ_0 (they determine the distribution of μ)
- Parameters: μ, τ
- Posterior hyperparameters: $\mu_{\text{post}}, \tau_{\text{post}}$

Conjugate Priors, example: Normal Likelihood τ

- Likelihood: $X \sim N(\mu, \tau)$ with known μ.
- Model parameter: $\tau = 1/\sigma^2$
- Conjugate prior distribution: $\tau \sim \Gamma(\alpha_0, \beta_0)$
- Prior hyperparameter: α_0, β_0
- Posterior hyperparameter:

$$\begin{aligned} \text{shape} \quad \alpha_{\text{post}} &= \alpha_0 + \frac{n}{2} \\ \text{rate} \quad \beta_{\text{post}} &= \beta_0 + \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2} \end{aligned}$$

In other words:

- Prior: $\tau \sim \Gamma(\alpha_0, \beta_0)$
- Likelihood: $X \sim N(\mu, \tau)$
- Posterior: $\tau \sim \Gamma(\alpha_{\text{post}}, \beta_{\text{post}})$.

Terminology:

- Hyperparameters: α_0, β_0 (they determine the distribution of τ)
- Parameters: μ, τ
- Posterior hyperparameters: $\alpha_{\text{post}}, \beta_{\text{post}}$

Conjugate Priors, example: Bernoulli Likelihood

- Likelihood: $X \sim \text{bern}(p)$.
- Model parameter: p
- Conjugate prior distribution: $p \sim \text{Beta}(\alpha_0, \beta_0)$
- Prior hyperparameter: α_0, β_0
- Posterior hyperparameter:

$$\begin{aligned} \alpha_{\text{post}} &= \alpha_0 + \sum x_i \\ \beta_{\text{post}} &= \beta_0 + n - \sum x_i \end{aligned}$$

In other words:

- Prior: $p \sim \text{Beta}(\alpha_0, \beta_0)$
- Likelihood: $X \sim \text{bern}(p)$
- Posterior: $p \sim \text{Beta}(\alpha_{\text{post}}, \beta_{\text{post}})$

Terminology:

- Hyperparameters: α_0, β_0 (they determine the distribution of p)
- Parameters: p
- Posterior hyperparameters: $\alpha_{\text{post}}, \beta_{\text{post}}$

Problems with the analytical approach Analytical approach

- Restrictive for...
  - priors
  - likelihood (“the model” in the frequentist world)
- For many relevant cases we have no analytical solution.

→

Numerical approach

- Markov Chain Monte Carlo (MCMC) methods, Metropolis-Hastings sampling, Gibbs sampling,...

Construct a Markov Chain that has the posterior distribution as its equilibrium distribution.

7.3 An alternative: MCMC

In the following we need another piece of software:

- JAGS (Just Another Gibbs Sampler)

  You can obtain JAGS at *`https://mcmc-jags.sourceforge.io/`*

  (please do, we will use JAGS frequently)

- JAGS allows us to flexibly model complex statistical models with the help of Markov chain Monte Carlo (MCMC) methods.

Required:

- Priors for μ and τ (need not be conjugate).
- Likelihood: $y \sim N(\mu, \tau)$ with $\tau = 1/\sigma^2$

→ Posterior distribution of μ and τ.

We will here just “use” our software got get a result. Below we will explain what the software actually does.

```
library(runjags)
library(coda)
X.model <- 'model {
 for (i in 1:length(y)) {
        y[i] ~ dnorm(mu,tau)
    }
    mu   ~ dnorm (0,.0001)
    tau  ~ dgamma(.01,.01)
    sd  <- sqrt(1/tau)
}'
X.jags<-run.jags(model=X.model,
                 data=list(y=y),
                 monitor=c("mu","sd"))
```

Notation for nodes (BUGS/JAGS/Stan/...)

- Stochastic nodes (discrete/continuous univariate/multivariate distributed):

  ```
  y[i] ~ dnorm(mu,tau)
  ...
  mu   ~ dnorm (0,.0001)
  ```

  - ...can be specified by *data* (have always this value)
  - ...can be unspecified
  - ...can be specified by *inits* (have this value before the first sample)

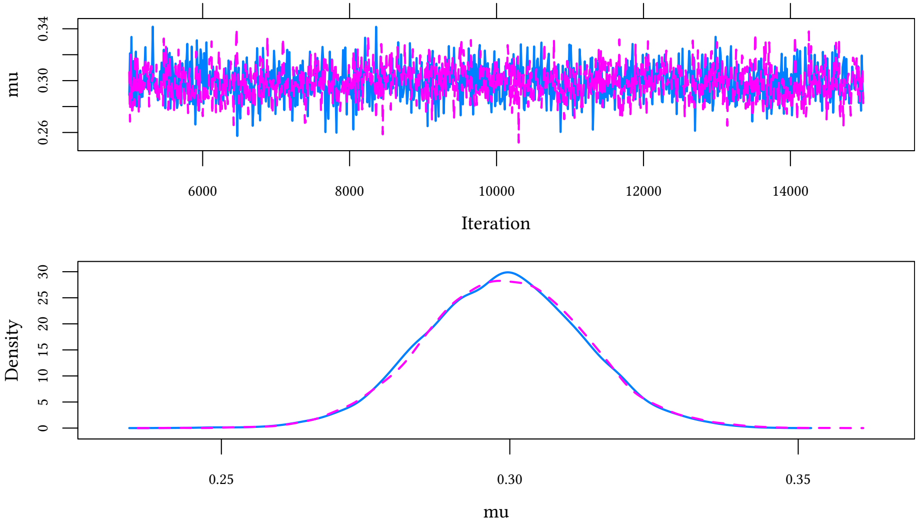
  Note: if *data* or *inits* sets a value to NA, this means “unspecified”.

- Deterministic nodes:

  ```
  sd  <- sqrt(1/tau)
  ```

JAGS samples from the posterior of μ and τ. Here is a distribution for μ:

```
plot(X.jags,var="mu",plot.type=c("trace","density"),layout=c(1,2))
```



Here is a *summary* of our estimation results:

```
X.jags


JAGS model summary statistics from 20000 samples (chains = 2; adapt+burnin = 5000):

   Lower95  Median Upper95    Mean        SD  Mode       MCerr MC%ofSD
mu 0.27243 0.29905 0.32637   0.299  0.013734    -- 0.000097879     0.7
sd 0.11051 0.12821 0.14839 0.12882 0.0097974    -- 0.000069278     0.7

   SSeff      AC.10    psrf
mu 19687 -0.0024031  1.0002
sd 20000 -0.0067856 0.99997

Total time taken: 0.3 seconds
```

Testing a point prediction in the *t.test*, as in $\mu = 0.3$, is a bit strange, at least from the Bayesian perspective. It might be more interesting to make a statement about the probability of an interval.

First, we convert our jags-object into a dataframe: How probable is $\mu \in (0.29, 0.31)$?

```
X.df<-data.frame(as.mcmc(X.jags))
str(X.df)

'data.frame': 20000 obs. of  2 variables:
 $ mu: num  0.289 0.306 0.291 0.285 0.292 ...
 $ sd: num  0.128 0.11 0.123 0.141 0.125 ...
```

We can now say how probable it is, ex post, that $\mu \in [0.29, .31]$:

```
100*mean(with(X.df,mu > 0.29 & mu < 0.31))

[1] 53.29
```

...or in a more narrow interval:

```
100*mean(with(X.df,mu > 0.299 & mu < 0.301))

[1] 5.78

100*mean(with(X.df,mu > 0.2999 & mu < 0.3001))

[1] 0.615
```

If, say, a government target is to have an average arrest rate of at least 0.25, we can now calculate the probability that $\mu > 0.25$.

How probable is $\mu > 0.25$?

```
100*mean(with(X.df,mu > 0.25))

[1] 99.94
```

In the Section 7.6 we will explain how all this works.

7.4 Accumulating evidence

↑ Above we used non-informative priors. ($\mu \sim N(0, 0.0001)$)

- Assume that we know something about μ (or that we talk to somebody who knows).
  - E.g. we ran a similar study in a different state.

    We found $\mu = 0.4$ and $\sigma_\mu = 0.014$ (i.e. the same σ_μ from our data, but a different μ).

    ($\sigma_\mu = 0.014$ is equivalent to $\tau_\mu = 1/\sigma_\mu^2 = 5102$)
  - Now we combine the data, i.e. we use a prior $\mu \sim N(0.4, 5102)$

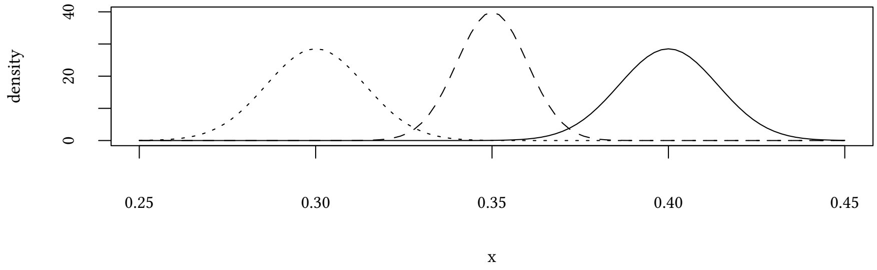
```
XA.model <- 'model {
 for (i in 1:length(y)) {
        y[i] ~ dnorm(mu,tau)
    }
    mu   ~ dnorm (0.4,1/0.014^2)
    tau  ~ dgamma(.01,.01)
    sd  <- sqrt(1/tau)
}'
```

```
summary(X.jags)

    Lower95    Median  Upper95      Mean          SD Mode
mu 0.272430 0.2990540 0.326372 0.2990049 0.01373359   NA
sd 0.110511 0.1282105 0.148395 0.1288232 0.00979741   NA
          MCerr MC%ofSD SSeff        AC.10     psrf
mu 0.00009787891     0.7 19687 -0.002403105 1.000191
sd 0.00006927815     0.7 20000 -0.006785630 0.999975

XA.jags<-run.jags(model=XA.model,data=list(y=y),monitor=c("mu","sd"))
summary(XA.jags)

    Lower95    Median  Upper95      Mean          SD Mode
mu 0.329557 0.3515650 0.372902 0.3516538 0.01101915   NA
sd 0.117699 0.1383025 0.161668 0.1390772 0.01135810   NA
          MCerr MC%ofSD SSeff       AC.10      psrf
mu 0.00008866214     0.8 15446 0.004111965 0.9999514
sd 0.00009282035     0.8 14974 0.002732966 1.0000881
```

- Prior mean: 0.4
- Sample mean: 0.3
- Posterior mean: 0.35

"A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule."

```
curve(dnorm(x,mean=.35,sd=.01),from=.25,to=.45,ylab='density',lty=2)
curve(dnorm(x,mean=.3,sd=.014),add=TRUE,lty=3)
curve(dnorm(x,mean=.4,sd=.014),add=TRUE)
```



7.5 Priors

- noninformative, flat, vague, diffuse

- weakly informative: intentionally weaker than the available prior knowledge, to keep the parameter within “reasonable bounds”.

- informative: available prior knowledge.

7.6 Finding posteriors

7.6.1 Overview

$$\underbrace{\Pr(\theta)}_{\text{prior}} \cdot \underbrace{\Pr(X|\theta)}_{\text{likelihood}} \cdot \frac{1}{\underbrace{\Pr(X)}_{\int \Pr(\theta)\cdot\Pr(X|\theta)\,d\theta}} = \underbrace{\Pr(\theta|X)}_{\text{posterior}}$$

Find $\Pr(\theta|X)$:

- Exact: but $\int \Pr(\theta) \cdot \Pr(X|\theta)\,d\theta$ can be hard (except for specific priors and likelihoods).

- MCMC Sampling
  - Rejection sampling: can be very slow (for a high-dimensional problem, and our problems are high-dimensional).
  - Metropolis–Hastings: quicker, samples are correlated, requires sampling of θ from joint distribution $\Pr(X|\theta)$.
  - Gibbs sampling: quicker, samples are correlated, requires sampling of θ_i from conditional (on θ_{-i}) distribution $\Pr(X|\{\theta_i, \theta_{-i}\})$.

    $\rightarrow$ this is easy! (at least much easier than $\Pr(X|\theta)$)

7.6.2 Rejection sampling

$$\underbrace{\Pr(\theta)}_{\text{prior}} \cdot \underbrace{\Pr(X|\theta)}_{\text{likelihood}} \cdot \frac{1}{\underbrace{\Pr(X)}_{\int \Pr(\theta)\cdot\Pr(X|\theta)\,d\theta}} = \underbrace{\Pr(\theta|X)}_{\text{posterior}}$$

How it works: Iterate the following:

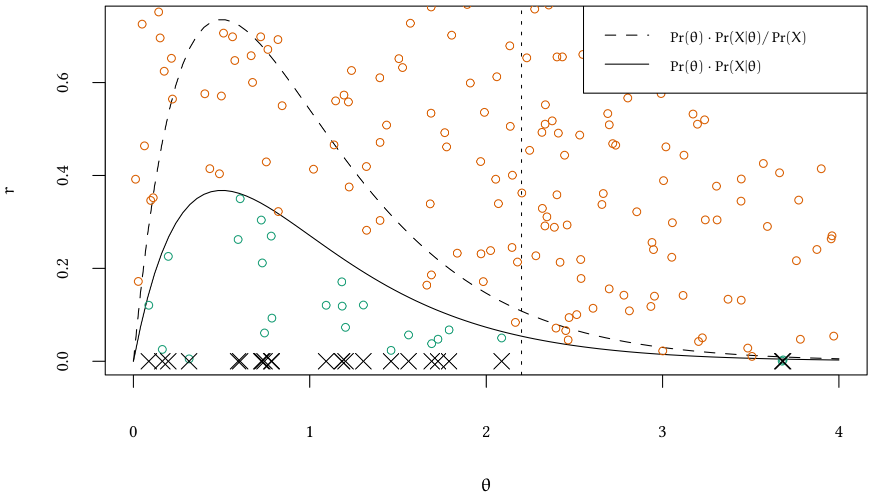
- Sample a candidate θ and a uniformly distributed random number r.

- If $\Pr(\theta) \cdot \Pr(X|\theta) > r$ then θ goes into the sample.

Problems:

- Slow (reject most of the time)

- $\max(r) > \max(\Pr(\theta) \cdot \Pr(X|\theta))$



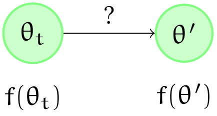
- The more dimensions we have, the more rejections. It would be nice to sample mainly in the posterior.

7.6.3 Metropolis-Hastings

$$\underbrace{\Pr(\theta)}_{\text{prior}}\cdot\underbrace{\Pr(X|\theta)}_{\text{likelihood}}\cdot\frac{1}{\underbrace{\Pr(X)}_{\int \Pr(\theta)\cdot\Pr(X|\theta)\,d\theta}}=\underbrace{\Pr(\theta|X)}_{\text{posterior}}$$

Metropolis Hastings

- Generates a sample of $\Pr(\theta|X)$.
- Needs only $f(\theta)=\Pr(\theta)\cdot\Pr(X|\theta)$



- Arbitrary symmetric PDF $Q(\theta|\eta)$, e.g. $Q=N$.
- Starting point $\theta=\theta_0$.

- Iterate:
  - Sample a candidate $\theta' \sim Q(\theta'|\theta_t)$.
  - Acceptance ratio is $\alpha = f(\theta')/f(\theta_t)$.
  - If $\alpha \geq 1$: $\underbrace{\theta_{t+1} = \theta'}_{\text{jump}}$
  - If $\alpha < 1$: with probability α we have $\underbrace{\theta_{t+1} = \theta'}_{\text{jump}}$, otherwise $\underbrace{\theta_{t+1} = \theta_t}_{\text{stay}}$.

Advantages:

- Faster than rejection sampling (in particular if θ is from a higher dimension).

Disadvantages:

- Samples are correlated (depending on Q).
  - If Q makes wide jumps: more rejections but less correlation.
  - If Q makes small jumps: fewer rejections but more correlation.
- Initial samples are from a different distribution. “burn-in” required.
- Finding a “good” jumping distribution $Q(x|y)$ can be tricky.

7.6.4 Gibbs sampling

Essentially as in Metropolis-Hastings, except that sampling is performed for each component of θ sequentially.

- determine θ_1^{t+1} with $f(\theta_1|\theta_2^t, \theta_3^t, \theta_4^t, \ldots, \theta_n^t)$
- determine θ_2^{t+1} with $f(\theta_2|\theta_1^{t+1}, \theta_3^t, \theta_4^t, \ldots, \theta_n^t)$
- determine θ_3^{t+1} with $f(\theta_3|\theta_1^{t+1}, \theta_2^{t+1}, \theta_4^t, \ldots, \theta_n^t)$
- $\vdots$
- determine θ_n^{t+1} with $f(\theta_n|\theta_1^{t+1}, \theta_2^{t+1}, \ldots, \theta_{n-1}^{t+1})$

Advantages:

- Requires only conditional distributions. $f(\theta_i|\theta_{-1})$, not joint distributions.

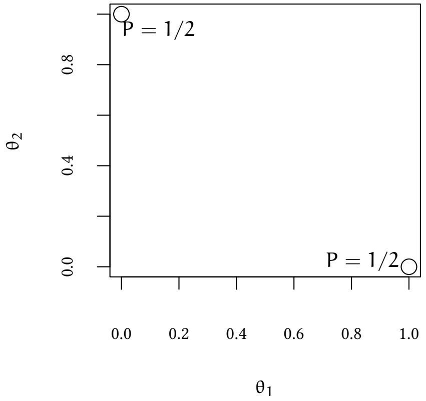
  (Humans often think about problems in terms of conditional distributions.)
- Finding a “good” jumping distribution $Q(x|y)$ is easier.

Disadvantages:

- Samples are correlated (potentially more than in MH if the number of dimensions is large).

- Initial samples are from a different distribution. “burn-in” required.

- Can get stuck on “disconnected islands”.



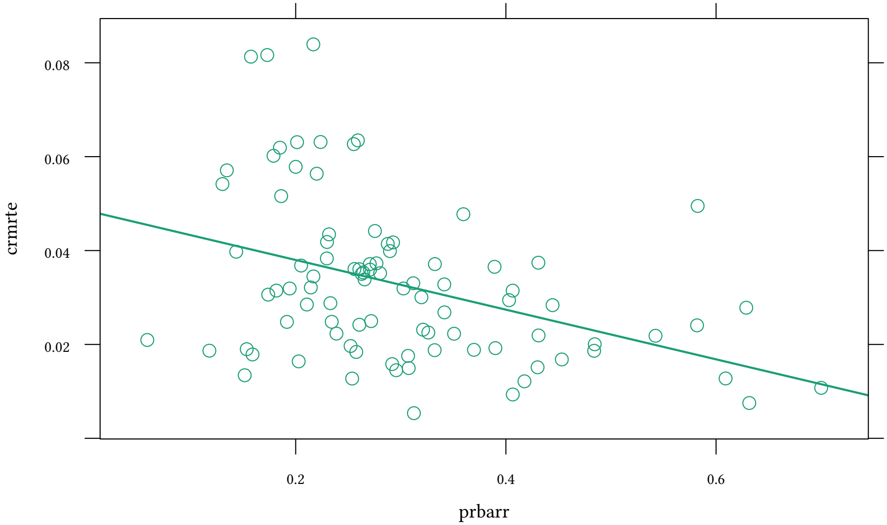
7.7 Technical issues

7.7.1 Introduction

We use linear regression as an example to illustrate some issues of the mechanics behind the MCMC sampling mentioned in the previous section.

Example: Crime in North Carolina in 1981 Let us have another look at the crime rate and the arrest rate in North Carolina.

```
xyplot(crmrte ~ prbarr,data=Crime,subset=year==81,type=c("p","r"))
```



We suspect that the crime rate is a linear function of the arrest rate. The standard tool would be OLS:

```
est<-lm(crmrte ~ prbarr,data=Crime,subset=year==81)
summary(est)



Call:
lm(formula = crmrte ~ prbarr, data = Crime, subset = year ==
    81)

Residuals:
      Min        1Q    Median        3Q       Max
-0.027125 -0.009932 -0.000848  0.007013  0.046819

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.048577   0.004261  11.400  < 2e-16 ***
prbarr      -0.052924   0.013129  -4.031 0.000118 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01571 on 88 degrees of freedom
Multiple R-squared:  0.1559,Adjusted R-squared:  0.1463
F-statistic: 16.25 on 1 and 88 DF,  p-value: 0.0001177
```

OLS

$$
\begin{aligned}
&Y = \beta_0 + \beta_1 X + u \quad \text{where } u \sim N(0, \sigma^2)\\
&Y \sim N(\beta_0 + \beta_1 X, \sigma^2)\\
&Y \sim N(\beta_0 + \beta_1 X, \tau)
\end{aligned}
$$

Both notations are equivalent. The former is more common in the frequentist context, the latter more common in the Bayesian context.

Now we do the same exercise in JAGS:

```
data<-with(subset(Crime,year==81),list(y=crmrte,x=prbarr))
reg.model<-'model {
 for (i in 1:length(y)) {
        y[i] ~ dnorm(beta0 + beta1*x[i],tau)
    }
    beta0 ~ dnorm (0,.0001)
    beta1 ~ dnorm (0,.0001)
    tau   ~ dgamma(.01,.01)
}'
reg.jags<-run.jags(model=reg.model,data=data,monitor=c("beta0","beta1"))
```

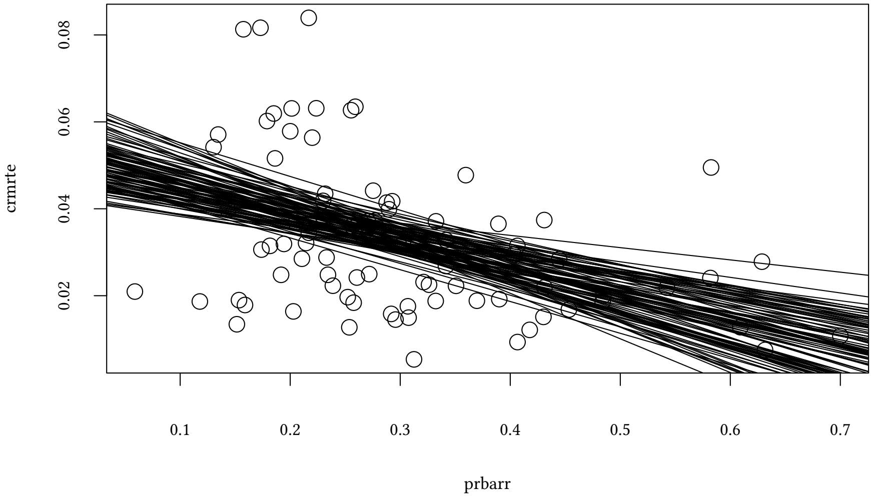
```
JAGS model summary statistics from 20000 samples (chains = 2; adapt+burnin = 5000):

        Lower95    Median   Upper95      Mean        SD Mode
beta0  0.037576  0.048829   0.06044  0.048832 0.0058477   --
beta1 -0.089112 -0.053883 -0.018401 -0.053766  0.018021   --

           MCerr MC%ofSD SSeff   AC.10   psrf
beta0 0.00014361     2.5  1658  0.1903 1.0004
beta1 0.00044062     2.4  1673 0.19073 1.0006

Total time taken: 0.2 seconds

coef(est)

(Intercept)      prbarr
 0.04857749 -0.05292384
```

The distribution we get here is very similar to the distribution parameters from the simple OLS.

7.7.2 Demeaning

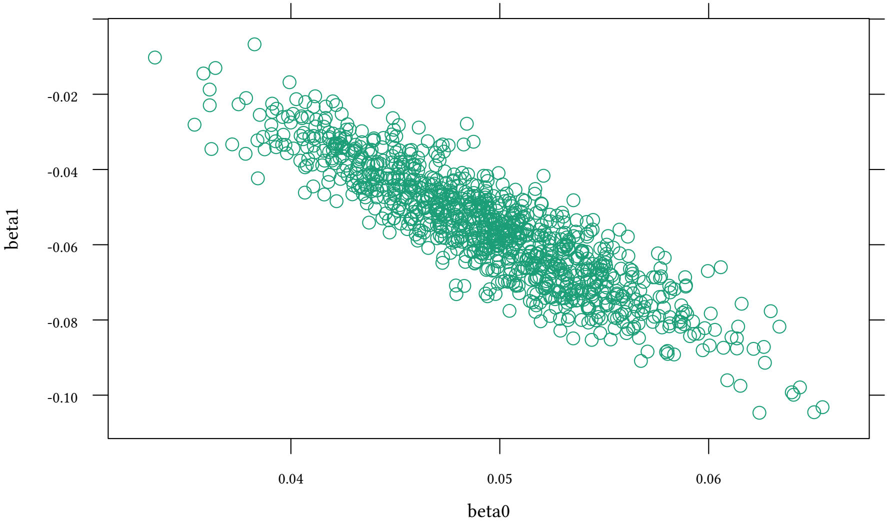
This is a technical issue. Demeaning might help improving the performance of our sampler.

Here is again a plot of the data the linear model has to fit. Added to the plot are the first 100 regressions (i.e. pairs of *beta0* and *beta1*).



We see that, to fit the data, *beta0* and *beta1* must be correlated. This is also what we see in the distribution of the sampled posterior:

```
reg.df<-as.data.frame(combine.mcmc(reg.jags))
xyplot(beta1~beta0,data=head(reg.df,1000))
```



As we will see below, this correlation makes the Gibbs sampler slower. One solution is to demean the data.

Demeaning does not change the estimate of the coefficient of X, it does change the constant, though.

$$Y = \beta_0 + \beta_1 X \tag{1}$$

$$Y - \bar{Y} = \underbrace{\beta_0 - \bar{Y} + \beta_1 \bar{X}}_{\beta_0'} + \beta_1 (X - \bar{X}) \tag{2}$$

Now we demean the data:

```
data2<-with(data,list(y=y-mean(y),x=x-mean(x)))
reg2.jags<-run.jags(model=reg.model,data=data2,monitor=c("beta0","beta1"))
```

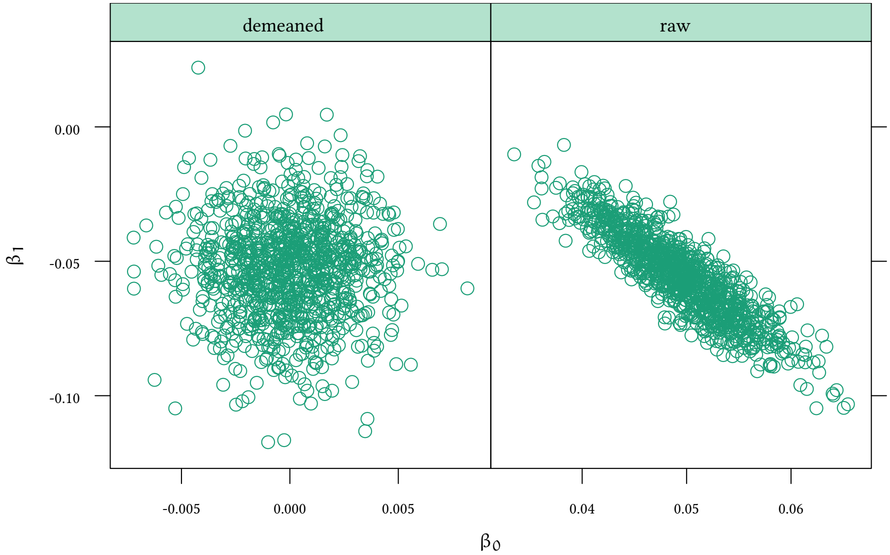
```
summary(reg2.jags)

          Lower95          Median     Upper95               Mean
beta0 -0.00434161 -0.000005983955  0.00468955  0.000001777658
beta1 -0.08971150 -0.052542750000 -0.01758430 -0.052632509443
               SD Mode         MCerr MC%ofSD SSeff        AC.10
beta0 0.002297572    NA 0.00001624629     0.7 20000 -0.012070859
beta1 0.018375021    NA 0.00013105835     0.7 19657 -0.002665037
          psrf
beta0 1.000063
beta1 1.000102
```

Compare with not demeaned estimation:

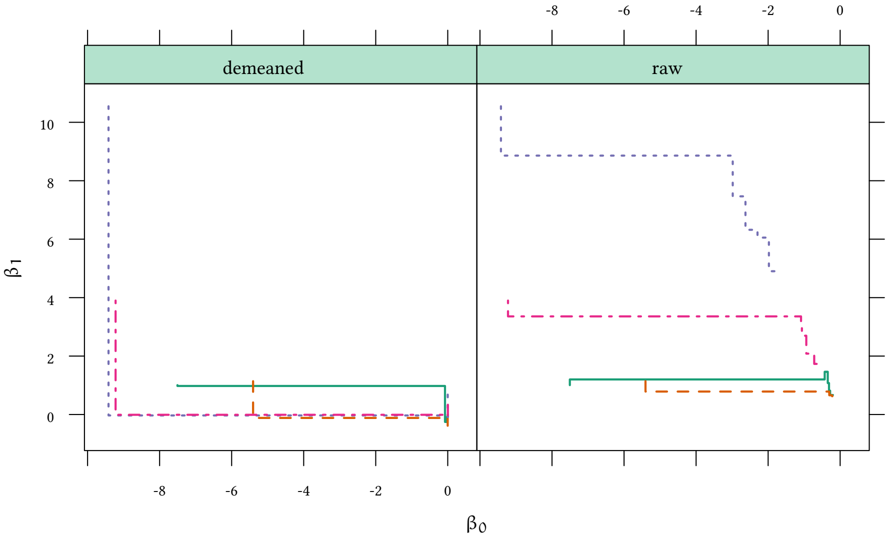
```
summary(reg.jags)

         Lower95      Median     Upper95        Mean           SD Mode
beta0  0.0375756  0.0488293  0.0604395  0.04883211 0.005847734   NA
beta1 -0.0891124 -0.0538834 -0.0184010 -0.05376627 0.018021130   NA
             MCerr MC%ofSD SSeff     AC.10     psrf
beta0 0.0001436095     2.5  1658 0.1902952 1.000353
beta1 0.0004406158     2.4  1673 0.1907322 1.000600
```

The estimate for *beta1* does not change (here we assume that we are mainly interested in the marginal effect, i.e. in *beta1*).

Now *beta0* and *beta1* are no longer correlated:

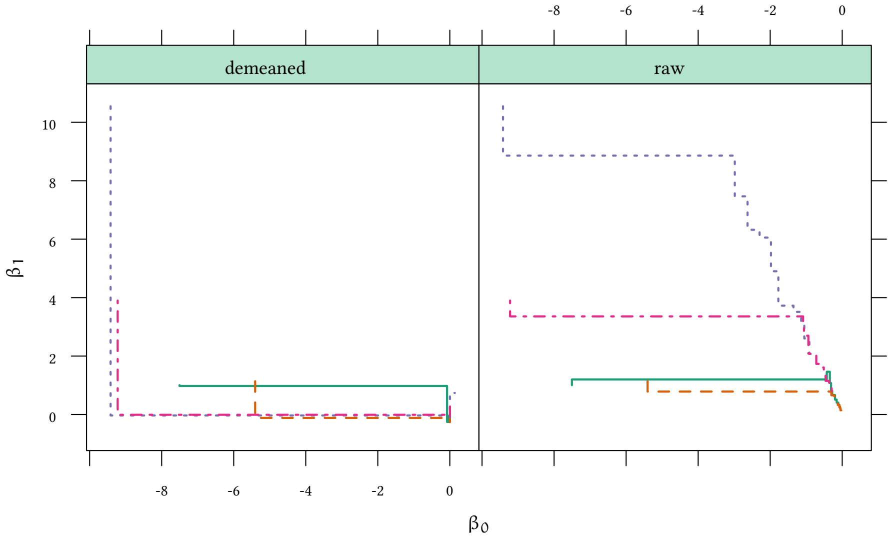


Convergence with raw and demeaned data To better understand convergence, we look at the first few samples in each case. Let us look at 5 chains with 5 samples each:

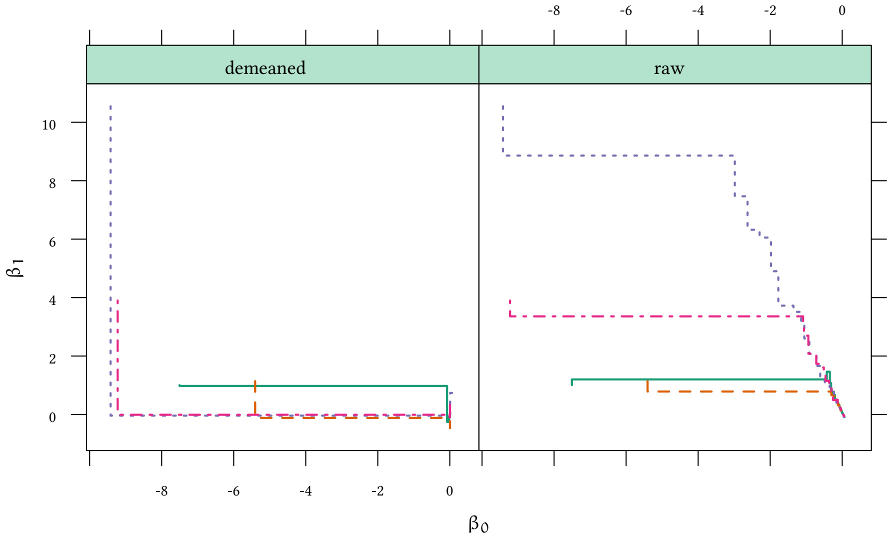


In the demeaned case, the Gibbs sampler jumps almost immediately to the center of the distribution. Convergence is reached within a small number of steps. In the not-demeaned case the Gibbs sampler walks slowly along the joint distribution of `beta0` and `beta1`. It takes a longer number of steps to reach the center of the distribution and to converge.

Here are 10 samples:



Here are 100 samples:



The Gibbs sampler can only increase the probability of *one* single posterior parameter in one step. In the posterior distribution the sampler, therefore, can only move parallel to one of the axes. If the posterior distribution is asymmetric (as in the raw data) convergence is slow.

The three steps of the Gibbs sampler **The three steps of the Gibbs sampler**

adaptation: optimise the algorithm (find a good jump distribution Q)

burnin: converge to the approximate shape of the distribution

sample: use a fixed algorithm to sample from posterior

Our problem:

- make sure that the sampler has converged

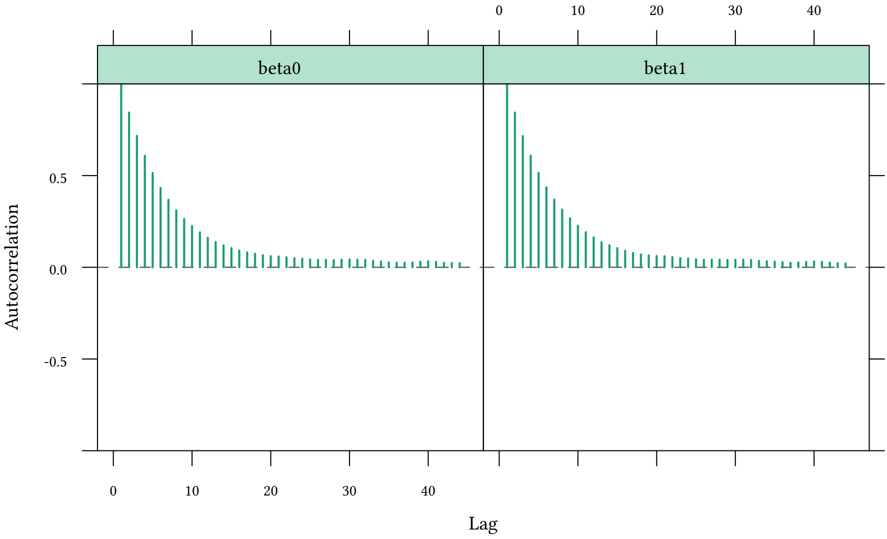
Solution:

- Demeaning (converge quickly to posterior)
- Good init values (start already from within the posterior)

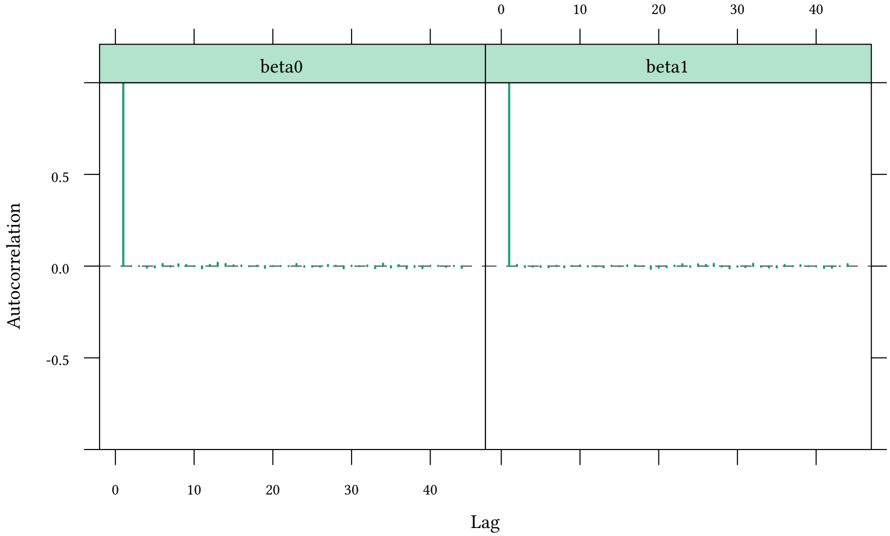
7.7.3 Autocorrelation

A related problem of the Gibbs sampler is that two successive samples may be (auto-) correlated.

```
acfplot(as.mcmc(reg.jags),aspect="fill",layout=c(2,1),ylim=c(-1,1))
```



```
acfplot(as.mcmc(reg2.jags),aspect="fill",layout=c(2,1),ylim=c(-1,1))
```



Another summary statistic in this context is the autocorrelation at a lag of 10, *AC.10*. Here for the not-demeaned model:

```
summary(reg.jags)

         Lower95      Median     Upper95         Mean          SD Mode
beta0  0.0375756  0.0488293  0.0604395  0.04883211 0.005847734   NA
beta1 -0.0891124 -0.0538834 -0.0184010 -0.05376627 0.018021130   NA
             MCerr MC%ofSD SSeff     AC.10     psrf
beta0 0.0001436095     2.5  1658 0.1902952 1.000353
beta1 0.0004406158     2.4  1673 0.1907322 1.000600
```

In the demeaned model *reg2*, the value of *AC.10* is smaller:

```
summary(reg2.jags)

          Lower95          Median     Upper95            Mean
beta0 -0.00434161 -0.000005983955  0.00468955  0.000001777658
beta1 -0.08971150 -0.052542750000 -0.01758430 -0.052632509443
               SD Mode          MCerr MC%ofSD SSeff        AC.10
beta0 0.002297572   NA 0.00001624629     0.7 20000 -0.012070859
beta1 0.018375021   NA 0.00013105835     0.7 19657 -0.002665037
          psrf
beta0 1.000063
beta1 1.000102
```

- A sample of 10 000 can, thus, not be treated as 10 000 independent observations.

```
effectiveSize(as.mcmc.list(reg.jags))


   beta0    beta1
1658.092 1672.800


effectiveSize(as.mcmc.list(reg2.jags))


   beta0    beta1
20000.00 19657.41
```

- Thinning (take only every n-th sample) does not lose much information.

7.7.4 Convergence and mixing of chains

In the following example we create on purpose a situation with two (almost) disconnected islands:

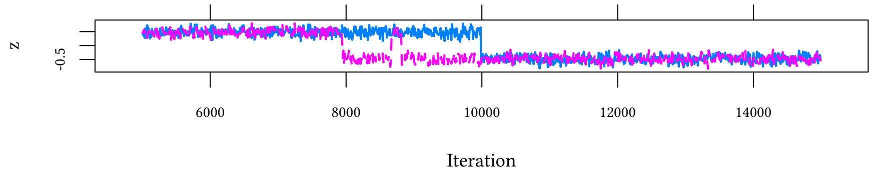
Example: Almost disconnected islands

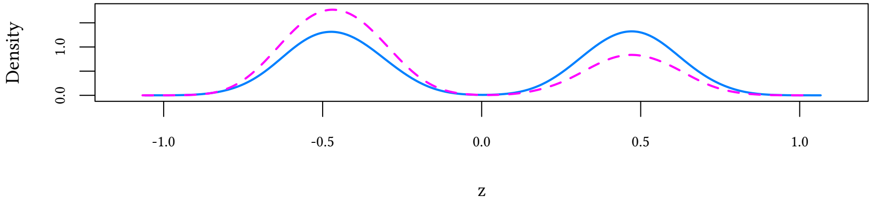
Create some data and estimate parameter z:

```
(x<-rbinom(9,1,.5))

[1] 0 0 0 1 0 1 0 0 0

island.mod<-'model {
 for (i in 1:length(x)) {
        x[i] ~ dbern(z^2)
    }
  z ~ dunif(-1,1)
}'
island.jags<-run.jags(model=island.mod,data=list(x=x),monitor=c("z"),
                      inits=ini)
```

(Since this is the first time I use the *inits* parameter for *run.jags*: In Appendix ....

Both $z = \sqrt{1/2}$ and $z = -\sqrt{1/2}$ fit the original process.





Gelman, Rubin (1992): potential scale reduction factor Idea: take k chains, discard "warm-up", split remaining chains, so that we have $2k$ sequences $\{\psi\}$, each of length n.

$$B = \text{between sequence variance}$$
$$W = \text{within sequence variance}$$

Variance of all chains combined:

$$\hat{\sigma}^2 = \frac{n-1}{n} W + \frac{B}{n}$$

Potential scale reduction:

$$\hat{R} = \sqrt{\frac{\hat{\sigma}^2}{W}}$$

Let us first look at the *psrf* for a "nice" case:

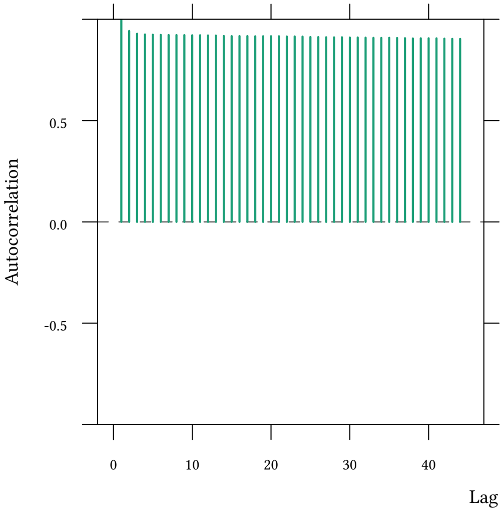
```
gelman.diag(reg.jags)

Potential scale reduction factors:
      Point est. Upper C.I.
beta0          1          1
beta1          1          1

Multivariate psrf

1
```

And now the island case:

```
gelman.diag(island.jags)

Potential scale reduction factors:

  Point est. Upper C.I.
z       1.24       1.87
```

Autocorrelation in the island case

```
acfplot(as.mcmc(island.jags),aspect="fill",layout=c(2,1),ylim=c(-1,1))
```



As a result of autocorrelation, the “effective size” is smaller than the sample size.

```
effectiveSize(as.mcmc.list(island.jags))

       z
7.127945
```

```
effectiveSize(as.mcmc.list(reg.jags))

   beta0    beta1
1658.092 1672.800

effectiveSize(as.mcmc.list(reg2.jags))

   beta0    beta1
20000.00 19657.41
```

7.7.5 A better vague prior for τ

When we specify a regression model we need a precision parameter τ. So far we did this:

```
reg.model<-'model {
 for (i in 1:length(y)) {
        y[i] ~ dnorm(beta0 + beta1*x[i],tau)
    }
    beta0 ~ dnorm (0,.0001)
    beta1 ~ dnorm (0,.0001)
    tau   ~ dgamma(.01,.01)
}'
```

A more stable way to model τ might be the following:

```
reg2.model<-'model {
 for (i in 1:length(y)) {
        y[i] ~ dnorm(beta0 + beta1*x[i],tau)
    }
    beta0 ~ dnorm (0,.0001)
    beta1 ~ dnorm (0,.0001)
    tau   ~ dgamma(m^2/d^2,m/d^2)
    m     ~ dgamma(1,1)
    d     ~ dgamma(1,1)
}'
```

- $\tau \sim \Gamma(0.01, 0.01)$

Remember:

- If $\tau \sim \Gamma(\alpha, \beta)$ then $E(\tau) = \alpha/\beta$ and $\text{var}(\tau) = \alpha/\beta^2$.
- $\alpha = 0.01$, $\beta = 0.01$ works well if $E(\tau) \approx 1$ and $\text{var}(\tau) \approx 100$.

Alternative:

- $\tau \sim \Gamma\left(\frac{m^2}{d^2}, \frac{m}{d^2}\right)$
- $m \sim \Gamma(1, 1)$

- $d \sim \Gamma(1,1)$

$\rightarrow E(\tau) = m, \mathrm{var}(\tau) = d^2$

- Speed: no substantial loss

- Convergence: often faster

7.8 Literature

- John H. Kruschke. “Doing Bayesian Data Analysis. A Tutorial with R, JAGS, and Stan”. Academic Press. 2014.

- Hoff, A First Course in Bayesian Statistical Methods.

- C. Andrieu, de Freitas, N., Doucet, A, Jordan, M. “An Introduction to MCMC for Machine Learning.” Machine Learning. 2003, 50(1-2), pp 5-43. (To learn more about MCMC sampling)

- Christian Robert and George Casella (2011). A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data. *Statistical Science.* 26(1), 102–115. (To learn more about the history of the field)

- William H. Greene. “Econometric Analysis”. Chapter 16.

Appendix 7.A Examples for the lecture

Example 1 You obtain the following output from JAGS (based on two chains with 10 000 samples each):

```
   Lower95 Median Upper95   Mean     SD Mode     MCerr
mu  4.2    6.5     9.0      6.5      1.0  -- 0.0083945

   MC%ofSD SSeff      AC.10    psrf
mu     0.7 20045 -0.0055416 0.99998
```

- Did the sampler converge?

- Are the samples autocorrelated?

- What is a 95% credible interval?

- How probable is it that $6.5 \leq \texttt{mu} \leq 9$?

Example 2 Have another look at the dataset `Caschool` from `Ecdat`.

- Explain `testscr` as a linear function of `str`. Don't use demeaning. What can you say about convergence?
- Now use demeaning. Compare your results.

Example 3 Your sample is drawn from a normal distribution with unknown mean μ and known variance $\sigma^2 = 1$. The mean of your sample of 10 observations is $\bar{x} = 2$.

You are interested in μ. Your prior for μ is that it follows a normal distribution with mean 40 and variance 4.

What is the posterior distribution for μ.

Appendix 7.B Exercises

To solve the next exercises the following commands for R might help: `curve`, `abline`, `dunif`, `dbeta`, `dnormal`, `dgamma`, `dbern`, `run.jags`, `gelman.plot`. The plot method for `runjags` objects has an option `plot(plot.type = c("trace", "density", "autocorr", "crosscorr"))`. The `runjags` command is provided by the library `runjags`. The `gelman.plot` command is provided by the library `coda`.

Exercise 7.1 *An event can have two possible outcomes, success or failure. You are interested in estimating the probability of success,* p*. You assume that* p *follows a Beta distribution with parameters* a *and* b*.*

Ex ante, you believe that any value of p *is equally likely. You observe two successes and one failure.*

1. *Find the Bayesian estimate of* p *analytically.*
2. *Find the Bayesian estimate of* p *using the Gibbs sampler (use the default settings).*

Exercise 7.2 *You are interested in estimating the linear relationship between the price of a computer and its CPU speed and RAM capacity (use 'Computers' from Ecdat).*

Inspect the convergence of the two coefficient estimates as well as their 95% CDIs using the sets of parameters below. Plot the respective trace, density, autocorrelation and cross-correlation graphs, as well as the evolution of PSRF as the number of samples increases. Also compare the actual time it takes to run the simulations.

1. *5 chains, 40 burn-in samples, 10 adaptation samples, 100 total (additional) samples;*
2. *5 chains, 40 burn-in samples, 10 adaptation samples, 100 total (additional) samples; demeaned data;*
3. *5 chains, 400 burn-in samples, 100 adaptation samples, 1000 total (additional) samples.*

4. *5 chains, 4000 burn-in samples, 1000 adaptation samples, 10000 total (additional) samples.*

Exercise 7.3 *Consider the general case of tossing a coin* n *times and observing "heads"* k *times.*

- *What is the maximum likelihood estimate that the next toss will result in "heads"?*
- *Using the Beta*$(1, 1)$ *distribution for the prior, what is the Bayesian estimate that the next toss will result in "heads"?*
- *When does the Bayesian estimate (while using the Beta*(a, b) *distribution for the prior) converge to the maximum likelihood estimate?*

Exercise 7.4 *An event can have two possible outcomes, success and failure. You are interested in the probability of a success,* p*. You assume that* p *follows a Beta distribution with parameters* α *and* β*.*

- *You think that all values of* $p \in [0, 1]$ *are equally probable. How would you model an uninformed prior? Draw the prior density of* p*.*
- *Now you observe 3 successes and no failure. What is your posterior for* α *and* β*? Draw the density of* p*.*
- *How probable is it that* $p < 1/2$*?*
- *How probable is it that* $p > 3/4$*?*
- *Thereafter, you observe 3 more failures and no success. What is now your posterior for* α *and* β*? Draw the density of* p*. Can you give an 95%-credible interval for* p*?*

Appendix 7.C Initial values for *run.jags*

To describe the posterior distribution, JAGS generates a sample from that distribution. Since this sample is "random", the exact sample depends on "initial values", i.e. on the values JAGS uses to start the MCMC chain. To make sure that my results are reproducible, I use the *inits* parameter of *run.jags* to provide always the same initial values to JAGS. Then I always get exactly the same results.

If you use different initial values, you may get slightly different results. If you use no initial values, JAGS will choose own values and, again, you will get slightly different results.

To keep things simple, I tried to avoid this topic in the lecture. Since this is not a lecture on the stability of MCMC algorithms, you can simply drop the *inits* parameter for *run.jags*. If you do this, your results will not be exactly my results, but they should be reasonably close. However, when you try my code, you may wonder what the *inits* parameter is doing. If you want to obtain exactly the same results, you can do the following:

```
initJags<-list()
initJags[[1]]<-list(.RNG.seed=1,.RNG.name="base::Mersenne-Twister")
initJags[[2]]<-list(.RNG.seed=2,.RNG.name="base::Super-Duper")
initJags[[3]]<-list(.RNG.seed=3,.RNG.name="base::Wichmann-Hill")
initJags[[4]]<-list(.RNG.seed=4,.RNG.name="base::Marsaglia-Multicarry")

genInit <- function(nChains,fn=NULL) {
    x<-list()
    for (i in 1:nChains) {
        x[[i]]<-initJags[[i]]
        if(!is.null(fn)) {
            vals<-fn(i)
            lapply(1:length(vals),function(j)
                x[[i]][[names(vals)[j]]]<<-vals[[j]])
        }
    }
    x
}
```

I first define a list *initJags* with four elements (because I am not using more than four MCMC chains). Each element of this list provides an initial value for the random number generator (e.g. *.RNG.seed=1*) and a name of the algorithm for the random numbers (e.g. *.RNG.name="base::Mersenne-Twister"*). Sometimes I also want to add initial values for parameters. This is what the *fn* parameter can do in genInit. If *fn* is not specified (i.e. it has the default value of *NULL*), then only the seed and the name of the random number generator are returned. If *fn* has a value, then it should be a function that provides further initial values.

So, again: If you need reproducible results, then you should use the *inits* parameter of run.jags. This makes sure that JAGS always generates the same MCMC chain and always generates the same posterior. If reproducible results are not your top priority, then you can drop the parameter.

8 Binary choice

8.1 Example

8.1.1 A binary dependent variable

A teacher has 100 students and devotes a random effort $x \in [0, 1]$ to each of them.

Student qualification *qual* depends on the effort and a random component e with $e \sim N(0, \sigma^2)$.

Students pass the final exam ($Y = 1$) when their qualification is larger than a critical value *crit*.

$$Y = 1 \Leftrightarrow x + e > \texttt{crit}$$

True relationship:

$$\begin{array}{rcl} \Pr(Y=1|x) & = & F(x,\beta) \\ \Pr(Y=0|x) & = & 1-F(x,\beta) \end{array}$$

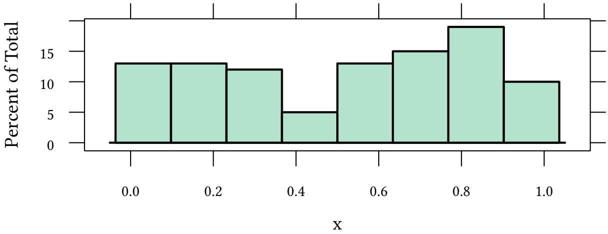
The teacher knows the structure of the process $F()$, but not its parameters β. The teacher is interested in knowing the critical effort that must be invested in the student such that, net of the random component, the student passes.

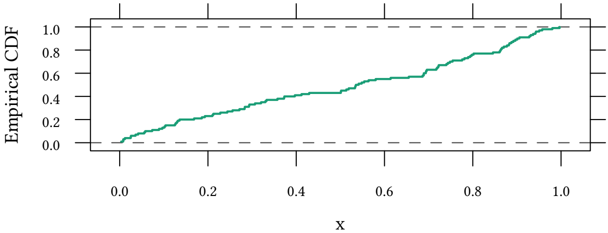
Let us first define a few variables, in particular our vector of random efforts x.

```
N <- 100
sigma <- .1
crit <- .3
set.seed(453)
x <- sort(runif(N))
```

Using a `sort`ed vector `x` will help us in our plots later. Below we see a histogram of x. An alternative way to look at the distribution of x is the cumulative distribution function.

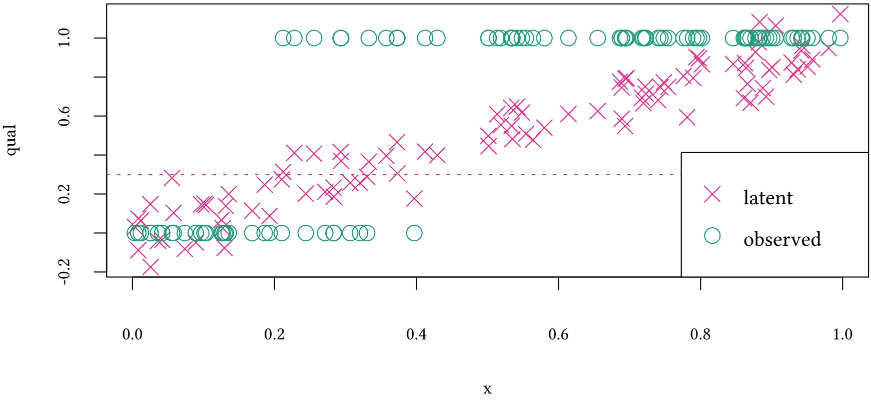
```
histogram(x)
ecdfplot(x)
```





Let us next add some random noise e to the effort x of the teacher and determine the (unobservable) qualification `qual` as well as the (observable) exam result y:

```
e <- sigma * rnorm(N)
qual <- x + e
plot (x,qual,pch=4,col=4)
abline (h=crit,col=4,lty="dotted")
y <- ifelse(qual>crit,1,0)
points(y ~ x, pch=1,col=1)
legend("bottomright",c("latent","observed"),pch=c(4,1),col=c(4,1),bg="white")
```



8.1.2 Try to estimate the relationship with OLS

$$\begin{aligned}\Pr(Y=1|x) &= F(x,\beta)\\ \Pr(Y=0|x) &= 1-F(x,\beta)\end{aligned}$$

With OLS we might be tempted to use

$$\Pr(Y=1|x)=F(x,\beta)=x'\beta$$

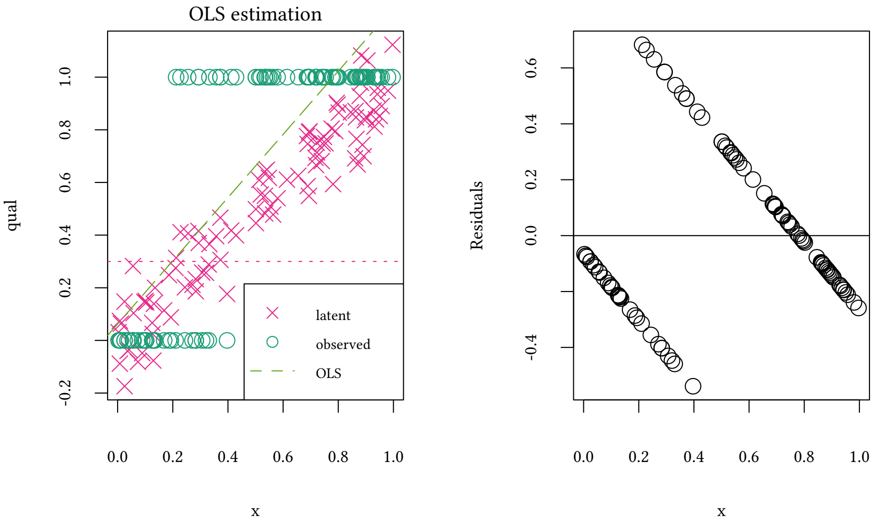
and then estimate

$$y=x'\beta+u$$

Why is this problematic?

- $E(u|X)\neq 0$ for most X
- predictions will not always look like probabilities

```
or <- lm ( y ~ x)
plot (qual ~ x,pch=4,col=4,main="OLS estimation")
abline(h=crit,col=4,lty="dotted")
points(y ~ x, pch=1,col=1)
abline(or,lty=5,col=5)
legend("bottomright",c("latent","observed","OLS"),pch=c(4,1,-1),col=c(4,1,5),
       cex=.7,lty=c(0,0,2),bg="white")
plot(or$residuals ~ x,ylab="Residuals");abline(h=0)
```



8.1.3 A problem with OLS

We see two problems with OLS.

- $\mathrm{E}(u|X) \neq 0$ for most X
- $x'\beta$ does not stay in $[0,1]$, predictions do not always look like probabilities

Let us start with the second problem: How to ensure $x'\beta \in [0,1]$?

Remember:

$$\begin{aligned} \Pr(Y=1|x) &= F(x,\beta) \\ \Pr(Y=0|x) &= 1-F(x,\beta) \end{aligned}$$

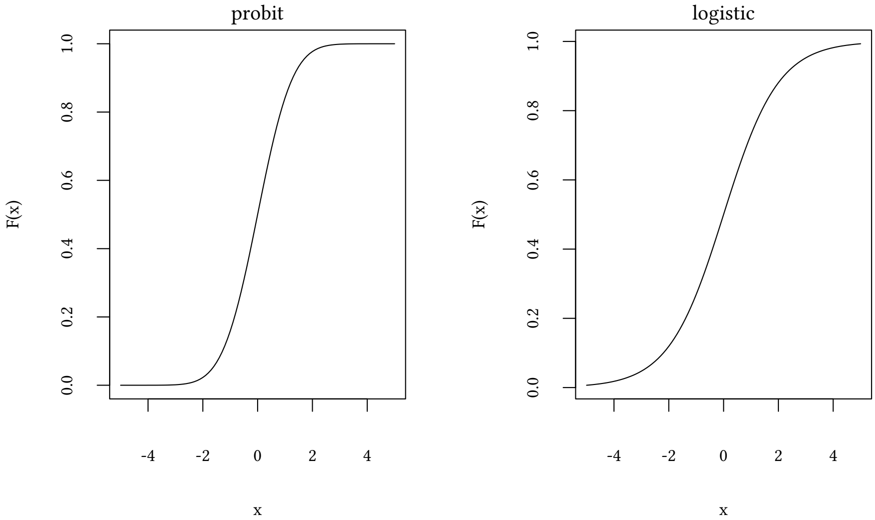
Trick: find a monotonic $\hat{F}$, such that $F(x, \beta) = \hat{F}(x'\beta)$ with

$$
\begin{aligned}
\lim_{x'\beta \to +\infty} \Pr(Y = 1|x) &= 1 \\
\lim_{x'\beta \to -\infty} \Pr(Y = 1|x) &= 0
\end{aligned}
$$

Any continuous distribution function would do.

A common distribution function is, e.g. the standard normal distribution. In R we call this function `pnorm`. Another possible function is the logistic function $\frac{e^x}{1+e^x}$

```
plot(pnorm,-5,5,main="probit",ylab="F(x)")
plot(function(x) {exp(x)/(1+exp(x))},-5,5,main="logistic",ylab="F(x)")
```

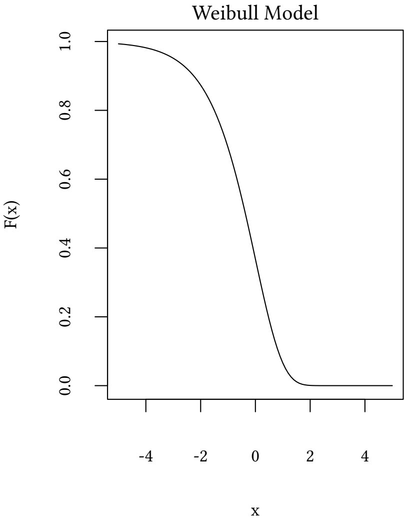


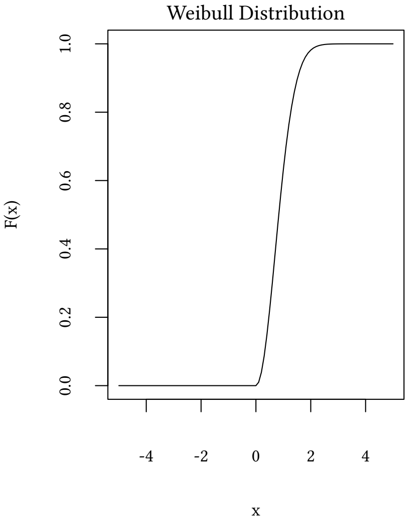
Since the logistic function has a couple of convenient properties, we will use it frequently and, hence, it also has a special name. In R we call it `plogis`. Instead of writing down the complicated formula, we could have said

```
plot(function(x) {exp(x)/(1+exp(x))},-5,5,main="logistic",ylab="F(x)")
plot(plogis,-5,5,main="logistic",ylab="F(x)")
```

Less common functions are the Weibull model and the the Weibull distribution...

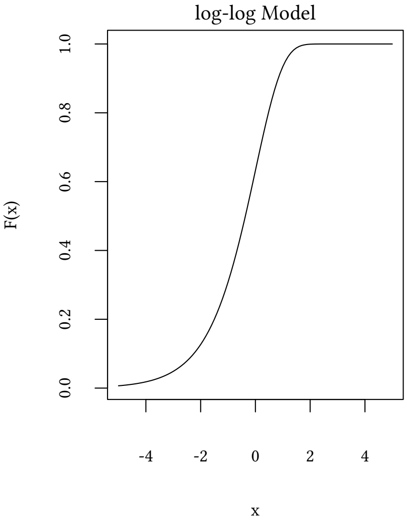
```
plot(function(x) {exp(-exp(x))},-5,5,main="Weibull Model",ylab="F(x)")
plot(function(x) {pweibull(x,shape=2)},-5,5, main="Weibull Distribution",ylab="F(x)")
```

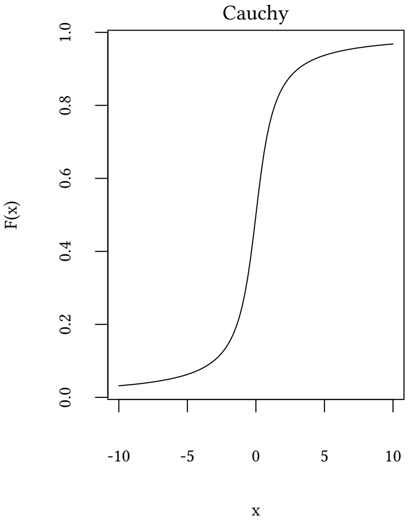




...the log-log Model or the Cauchy distribution ...

```
plot(function(x) {1-exp(-exp(x))},-5,5,main="log-log Model",ylab="F(x)")
plot(pcauchy,-10,10,main="Cauchy",ylab="F(x)")
```



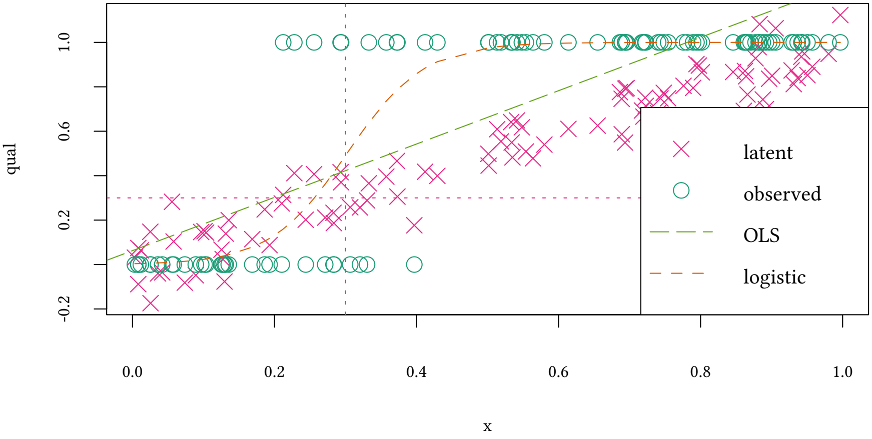


So let us apply, e.g. the logistic function to our problem. This is no longer done with lm, now we need a generalised linear model: glm.

```
(lr <- glm( y ~ x,family=binomial(link="logit")))


Call:  glm(formula = y ~ x, family = binomial(link = "logit"))

Coefficients:
(Intercept)            x
     -5.581       18.483

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:      125.4
Residual Deviance: 30.04  AIC: 34.04
```

Let us compare, in a graph, the OLS model and the logistic model:

```
plot (qual ~ x,pch=4,col=4)
abline(h=crit,col=4,v=crit,lty="dotted")
points(y ~ x, pch=1,col=1)
abline(or,lty=5,col=5)
lines(fitted(lr) ~ x, lty=2,col=2)
legend("bottomright",c("latent","observed","OLS","logistic"),
       pch=c(4,1,-1,-1),col=c(4,1,5,2),lty=c(0,0,5,2),bg="white")
```



The glm function is rather flexible. We used to do OLS with lm, but we can also do this with glm. Let us first do OLS with lm

```
lm( y ~ x)


Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    0.06257      1.20008
```

Now we do the same with glm

```
glm( y ~ x)


Call:  glm(formula = y ~ x)

Coefficients:
(Intercept)            x
    0.06257      1.20008

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:     21.76
Residual Deviance: 7.499  AIC: 30.75
```

The estimated coefficients are the same. But what happened to the long list of parameters we had previously in glm when we estimated the logistic model. Actually, glm always has *default* values for its parameters. We could as well have said

```
glm( y ~ x,family=gaussian(link="identity"))


Call:  glm(formula = y ~ x, family = gaussian(link = "identity"))

Coefficients:
(Intercept)            x
    0.06257      1.20008

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:     21.76
Residual Deviance: 7.499  AIC: 30.75
```

and obtain the same result. The *family* argument indicates the type of the distribution of the dependent variable.

8.2 Families and links

Remember that we could write the OLS model with a standard normal distributed error term

$$Y = X\beta + \epsilon \text{ and } \epsilon \sim N(0, \sigma^2)$$

alternatively as

$$Y \sim \underbrace{N}_{\text{family}}(\underbrace{X\beta}_{\text{link}}, \sigma^2)$$

Similarly, with the logistic model we said

$$\Pr(Y = 1|x) = F(X\beta)$$

but we can also write

$$Y \sim \underbrace{\text{binom}}_{\text{family}}(\underbrace{F}_{\text{link}}(X\beta))$$

```
glm(y ~ x,family=binomial(link=logit))


Call:  glm(formula = y ~ x, family = binomial(link = logit))

Coefficients:
(Intercept)            x
     -5.581       18.483

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:      125.4
Residual Deviance: 30.04  AIC: 34.04
```

Remember that we could write the OLS model with a standard normal distributed error term

$$Y = X\beta + \epsilon \text{ and } \epsilon \sim N(0, \sigma^2)$$

alternatively as

$$Y \sim \underbrace{N}_{\text{family}}(\underbrace{X\beta}_{\text{link}}, \sigma^2)$$

```
OLS.model <- 'model {
   for (i in 1:length(y)) {
      y[i] ~ dnorm(beta[1]+beta[2]*x[i],tau)
   }
   for (k in 1:2) {
     beta[k] ~ dnorm(0,.0001)
   }
   tau ~ dgamma(.01,.01)
}'
OLS.jags<-run.jags(model=OLS.model,
                 data=list(y=y,x=x),
                 monitor=c("beta"))
```

Similarly, with the logistic model we said

$$\Pr(Y = 1|x) = F(X\beta)$$

but we can also write

$$Y \sim \underbrace{\text{binom}}_{\text{family}}(\underbrace{F}_{\text{link}}(X\beta))$$

```
logit.model <- 'model {
for (i in 1:length(y)) {
    y[i] ~ dbern(p[i]);
    p[i] <- plogis(b1+b2*x[i],0,1)
}
b1 ~ dnorm (0,.0001)
b2 ~ dnorm (0,.0001)
}'
est.jags<-run.jags(logit.model,
                   data=list(y=y,x=x),
                   monitor=c("b1","b2"))
```

```
JAGS model summary statistics from 20000 samples (chains = 2; adapt+burnin = 5000):

   Lower95  Median Upper95    Mean     SD Mode    MCerr MC%ofSD SSeff
b1 -9.9975 -6.1745  -3.159 -6.4184  1.806   -- 0.083155     4.6   472
b2  10.826  20.495  32.956  21.325 5.9069   --  0.26624     4.5   492

     AC.10   psrf
b1 0.61469 1.0429
b2  0.6169  1.035

Total time taken: 1.7 seconds
```

Compare with *glm*:

```
lr


Call:  glm(formula = y ~ x, family = binomial(link = "logit"))

Coefficients:
(Intercept)            x
     -5.581       18.483

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:     125.4
Residual Deviance: 30.04  AIC: 34.04
```

8.2.1 A list of common families and link functions

family	link ($F()$)
gaussian (N)	identity, log, inverse
binomial	logit, probit, cauchit, log, cloglog
gamma	identity, log, inverse
poisson	identity, log, sqrt
inverse.gaussian	identity, log, inverse, $1/\mu^2$
quasi	logit, probit, log, cloglog, identity, inverse, sqrt, $1/\mu^2$

8.3 Marginal effects

8.3.1 Marginal effects with OLS

$$E[y|x] = E[x'\beta + \epsilon] = x'\beta$$

$$\frac{\partial E[y|x]}{\partial x} = \beta$$

Here the marginal effect is simple to determine, it is just β.

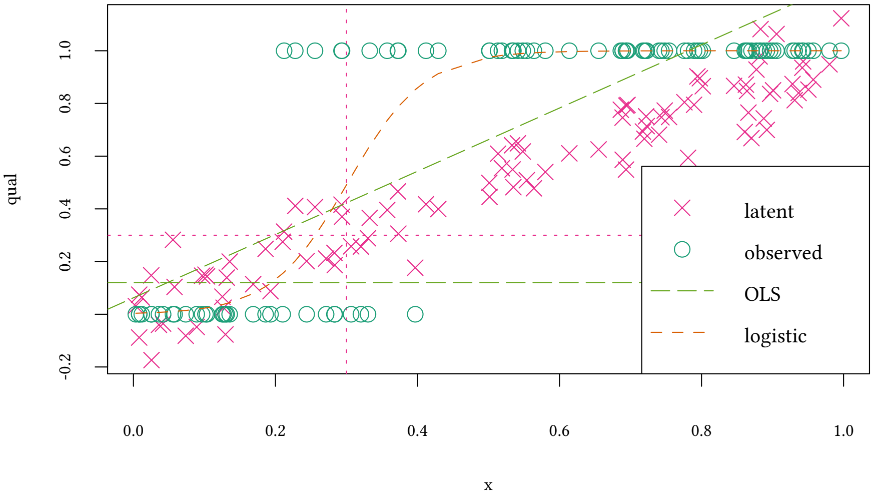
The marginal effect does not depend on x.

```
coef(or)

(Intercept)           x 
 0.06256655  1.20007843 
```

If we want to show the marginal effect in our diagram, we can say (after rescaling the effect so that it fits into our diagram)

```
myScale=.1
```

```
abline(h=coef(or)["x"]*myScale,lty=5,col=5)
```



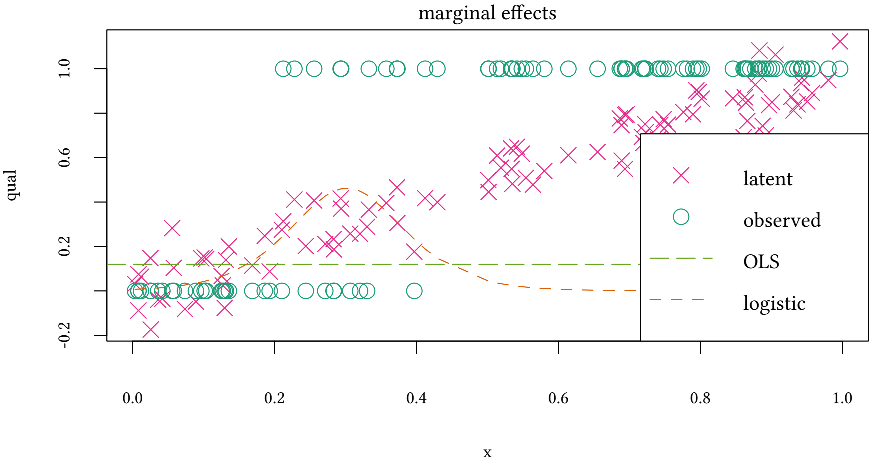
8.3.2 Marginal effects with binary choice models

$$\Pr(Y = 1|x) = F(x'\beta) \qquad \rightarrow \qquad \begin{aligned} E[y|x] &= F(x'\beta) \\ \frac{\partial E[y|x]}{\partial x} &= f(x'\beta)\cdot\beta \end{aligned}$$

```
lmarginal <- dlogis(predict(lr)) * coef(lr)["x"]
```

We can calculate the marginal effects and show them in our diagram:

```
plot (qual ~ x,pch=4,col=4,main="marginal effects")
points(y ~ x, pch=1,col=1)
abline(h=coef(or)["x"]*myScale,lty=5,col=5)
lines(lmarginal * myScale ~ x,lty=2,col=2)
legend("bottomright",c("latent","observed","OLS","logistic"),
       pch=c(4,1,-1,-1),col=c(4,1,5,2),lty=c(0,0,5,2),bg="white")
```



8.3.3 The marginal effect is not constant

$$\frac{\partial E[y|x]}{\partial x} = f(x'\beta) \cdot \beta$$

It is highest for the critical value and smaller for the extreme values of x. If our teacher is interested in increasing the number of students who pass the test, then she might concentrate her efforts on the students in the middle, since there the expected marginal return is highest.

Now assume, somebody is interested in a somehow *aggregate* measure of the marginal effect. Where do we evaluate the marginal effect?

- We can evaluate the marginal effect at the sample mean. In this case the sample mean is *mean(x)*, hence $x'\beta$ is

  ```
  xbmean <- coef(lr) %*% c(1,mean(x))
  xbmean
  ```

  ```
             [,1]
  [1,] 3.928001
  ```

  hence, the marginal effect $f(x'\beta) \cdot \beta$ is

  ```
  dlogis(xbmean)*coef(lr)["x"]
  ```

  ```
             [,1]
  [1,] 0.349894
  ```

- We can also evaluate the marginal effect for all observations and then use the average. Remember that `lmarginal` still contains all the marginal effects, so we can simply take the mean.

  ```
    mean(lmarginal)

  [1] 0.8585523
  ```

  As we see, the result depends heavily on the method. The first method gave us the marginal effect on the *average subject*, the second the *average marginal effect* (of all subjects).

```
dlogis(xbmean)*coef(lr)["x"]

         [,1]
[1,] 0.349894

mean(lmarginal)

[1] 0.8585523
```

Furthermore, we could argue that a marginal change should, ideally, first target at the subjects which can be influenced most easily:

```
  max(lmarginal)

[1] 4.612386
```

When X_d is a dummy variable, derivatives make no sense. Instead, we use (for the average marginal effect):

$$
\begin{aligned}
&\Pr(Y=1|x_{-d},d=1)-\Pr(Y=1|x_{-d},d=0)\\
&=\frac{1}{n}\sum_{i=1}^{n}\Big(F((x_{0,i},\ldots,x_{d-1,i},1,x_{d+1,i},\ldots,x_{k,i})'\beta)-\\
&\qquad -F((x_{0,i},\ldots,x_{d-1,i},0,x_{d+1,i},\ldots,x_{k,i})'\beta)\Big)
\end{aligned}
$$

and for the marginal effect at the average x:

$$
\begin{aligned}
&\Pr(Y=1|\bar{x}_{-d},d=1)-\Pr(Y=1|\bar{x}_{-d},d=0)\\
&=F((\bar{x}_{0,i},\ldots,\bar{x}_{d-1,i},1,\bar{x}_{d+1,i},\ldots,\bar{x}_{k,i})'\beta)-\\
&-F((\bar{x}_{0,i},\ldots,\bar{x}_{d-1,i},0,\bar{x}_{d+1,i},\ldots,\bar{x}_{k,i})'\beta))
\end{aligned}
$$

8.3.4 The normal link function

$$\begin{aligned}\Pr(Y=1|x) &= F(x,\beta)=\Phi(x'\beta)\\ \Pr(Y=0|x) &= 1-F(x,\beta)=1-\Phi(x'\beta)\end{aligned}$$

or

$$Y \sim \underbrace{\text{binom}}_{\text{family}}(\underbrace{\Phi}_{\text{link}}(X\beta))$$

```
probit.model <- 'model {
for (i in 1:length(y)) {
    y[i] ~ dbern(p[i]);
    p[i] <- pnorm(b1+b2*x[i],0,1)
}
b1 ~ dnorm (0,.0001)
b2 ~ dnorm (0,.0001)
}'
est.jags<-run.jags(probit.model,
                   data=list(y=y,x=x),
                   monitor=c("b1","b2"))
```

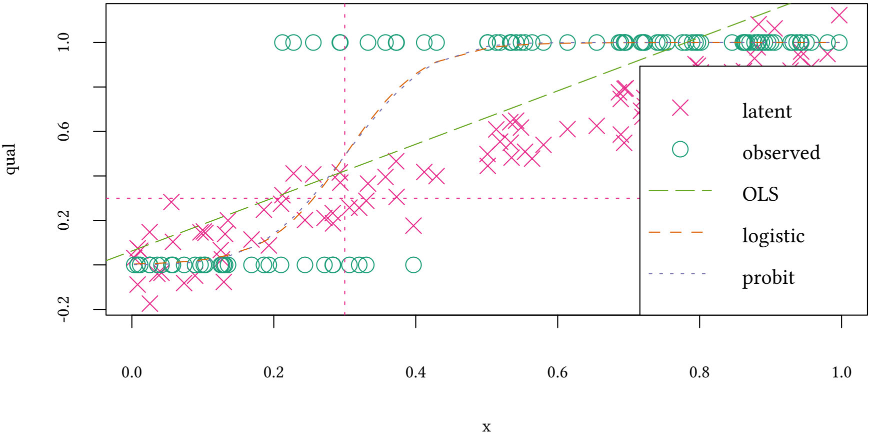
```
JAGS model summary statistics from 20000 samples (chains = 2; adapt+burnin = 5000):

   Lower95  Median Upper95    Mean      SD Mode    MCerr MC%ofSD
b1  -5.357 -3.4068  -1.924 -3.5348 0.90483   -- 0.039553     4.4
b2  6.7886  11.317  18.109  11.748  2.9855   --  0.12997     4.4

   SSeff   AC.10   psrf
b1   523 0.58459 1.0082
b2   528 0.58601 1.0064

Total time taken: 1.6 seconds
```

Compare with *glm*:

```
pr


Call:  glm(formula = y ~ x, family = binomial(link = "probit"))

Coefficients:
(Intercept)            x
     -3.203       10.601

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:      125.4
Residual Deviance: 29.37  AIC: 33.37
```

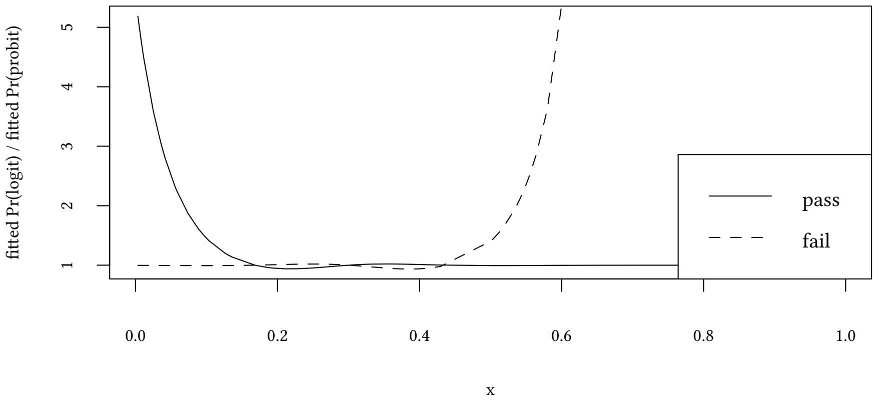
Here is the figure again, now with the estimation result for the normal link function added:

```
plot (qual ~ x,pch=4,col=4)
points(y ~ x, pch=1,col=1)
abline(h=crit,col=4,v=crit,lty="dotted")
abline(or,lty=5,col=5)
lines(fitted(lr) ~ x, lty=2,col=2)
lines(fitted(pr) ~ x, lty=3,col=3)
legend("bottomright",c("latent","observed","OLS","logistic","probit"),
       pch=c(4,1,-1,-1,-1),col=c(4,1,5,2,3),lty=c(0,0,5,2,3),bg="white")
```



We see that the logistic model and the probit model does not seem to be far apart. Note, however, that for very small or very large probabilities the distribution does matter. To see that, we have a look at the ratio of the predicted values:

```
lp <- fitted(lr)/fitted(pr)
lp1 <- (1-fitted(lr))/(1-fitted(pr))
plot (lp ~ x,type="l",ylab="fitted $\\Pr$(logit) / fitted $\\Pr$(probit)")
lines (lp1 ~ x,lty=2)
legend("bottomright",c("pass","fail"),lty=c(1,2),bg="white")
```



To predict probabilities for extreme values (probabilitites close to zero or close to one) the choice of the link function is essential. (Since, usually, we do not know much about the "correct" link function, we can not say too much about ratios of probabilities for extreme values).

8.3.5 Marginal effects with the normal link function

$$\begin{aligned}\Pr(Y=1|x) &= F(x,\beta)=\Phi(x'\beta)\\ \Pr(Y=0|x) &= 1-F(x,\beta)=1-\Phi(x'\beta)\end{aligned}$$

$$E(y|x)=F(x'\beta)=\Phi(x'\beta)$$

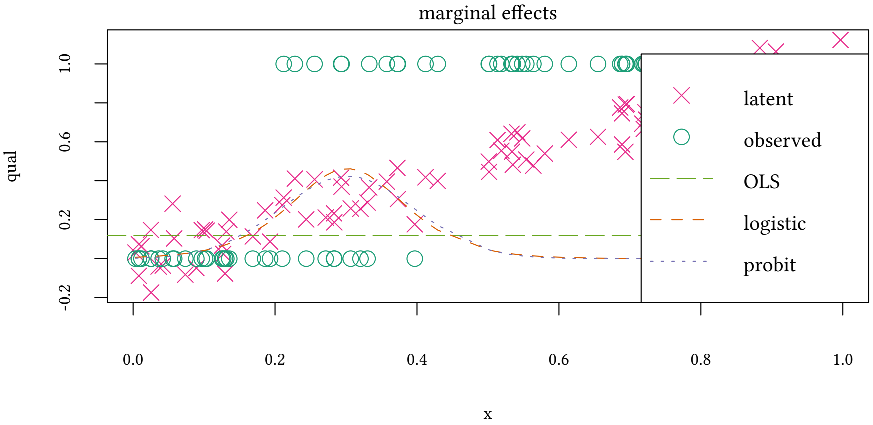
Marginal effects are:

$$\frac{\partial E[y|x]}{\partial x}=f(x'\beta)\cdot\beta=\phi(x'\beta)\cdot\beta$$

```
pmarginal<-dnorm(pr$linear.predictors)*coef(pr)["x"]
```

Now let us draw the familiar figure again, now with the marginal effects of the probit model added:

```
plot (qual ~ x,pch=4,col=4,main="marginal effects")
points(y ~ x, pch=1,col=1)
abline(h=coef(or)["x"]*myScale,lty=5,col=5)
lines(lmarginal*myScale ~ x,lty=2,col=2)
lines(pmarginal*myScale ~ x,lty=3,col=3)
legend("bottomright",c("latent","observed","OLS","logistic","probit"),
       pch=c(4,1,-1,-1,-1),col=c(4,1,5,2,3),lty=c(0,0,5,2,3),bg="white")
```



Again, we have different options to calculate aggregate marginal effects.

$$\frac{\partial E[y|x]}{\partial x} = f(x'\beta)\cdot\beta = \phi(x'\beta)\cdot\beta$$

- Average effect at the sample mean?

```
xbmeanP <- coef(pr) %*% c(1,mean(x))
dnorm(xbmeanP)*coef(pr)["x"]
```

```
          [,1]
[1,] 0.3353767
```

- Effect for the average unit:

```
mean(pmarginal)
```

```
[1] 0.8420524
```

- Effect for the most efficient unit:

```
max(pmarginal)
```

```
[1] 4.224432
```

As with the logistic regression, also with the probit regression the result depends heavily on the method.

8.3.6 Marginal effects and interactions

linear model

$$\begin{aligned}
y &= \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon \\
E[y|x] &= \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 \\
\frac{\partial E[y|x]}{\partial x_1} &= \beta_1 + \beta_{12} x_2 \\
\frac{\partial^2 E[y|x]}{\partial x_1 \partial x_2} &= \beta_{12}
\end{aligned}$$

probit model

$$\begin{aligned}
E[y|x] &= \Phi(\underbrace{\beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2}_{u}) \\
\frac{\partial E[y|x]}{\partial x_1} &= (\beta_1 + \beta_{12} x_2) \cdot \phi(u) \\
\frac{\partial^2 E[y|x]}{\partial x_1 \partial x_2} &= \beta_{12} \cdot \phi(u) + (\beta_1 + \beta_{12} x_2) \cdot \frac{\partial}{\partial x_2} \phi(u) \\
&= \beta_{12} \cdot \phi(u) + (\beta_1 + \beta_{12} x_2) \cdot (\beta_2 + \beta_{12} x_1) \cdot \phi'(u)
\end{aligned}$$

Note: The marginal effect is *not* $\beta_{12}\phi(u)$

Note 2: Assume $\beta_{12} = 0$

$$\begin{aligned}
\frac{\partial^2 E[y|x]}{\partial x_1 \partial x_2} &= \beta_{12} \cdot \phi(u) + (\beta_1 + \beta_{12} x_2) \cdot (\beta_2 + \beta_{12} x_1) \cdot \phi'(u) \\
&= \beta_1 \beta_2 \phi'(u)
\end{aligned}$$

i.e. even if the estimated interaction on the level of the latent variable is zero, the marginal effect could be (significantly) different from zero.

- Obviously: Significance of the coefficient β_{12} has nothing to do with significance of the marginal effect.

- Should we study the model (the interactions) on the level of the latent variable, or on the level of y?

Interactions of three expressions

$$\begin{aligned}
E[y|x] &= \Phi(\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_{12}x_1x_2+\beta_{13}x_1x_3+\beta_{23}x_2x_3)\\
\frac{\partial^3E[y|x]}{\partial x_1\partial x_2\partial x_3} &= \varphi''(u)\cdot(\beta_3+x_1\beta_{13}+x_2\beta_{23}+x_1x_2\beta_{123})\cdot\\
&\qquad(\beta_2+x_1\beta_{12}+x_3\beta_{23}+x_1x_3\beta_{123})\cdot\\
&\qquad(\beta_1+x_2\beta_{12}+x_3\beta_{13}+x_2x_3\beta_{123})+\\
&\quad\varphi'(u)\cdot(\beta_{12}+x_3\beta_{123})\cdot(\beta_3+x_1\beta_{13}+x_2\beta_{23}+x_1x_2\beta_{123})+\\
&\quad\varphi'(u)\cdot(\beta_{13}+x_2\beta_{123})\cdot(\beta_2+x_1\beta_{12}+x_3\beta_{23}+x_1x_3\beta_{123})+\\
&\quad\varphi'(u)\cdot(\beta_{23}+x_1\beta_{123})\cdot(\beta_1+x_2\beta_{12}+x_3\beta_{13}+x_2x_3\beta_{123})+\\
&\quad\varphi(u)\cdot\beta_{123}
\end{aligned}$$

8.4 Odds ratios in the logit model

$$\Pr(Y=1|x)=F(x'\beta)\quad\text{with}\quad F(x'\beta)=\frac{e^{x'\beta}}{1+e^{x'\beta}}$$

Let us define the odds $o=\frac{p}{1-p}$, then

$$\log(o)=\log\frac{p}{1-p}=\log\frac{\frac{e^{x'\beta}}{1+e^{x'\beta}}}{1-\frac{e^{x'\beta}}{1+e^{x'\beta}}}=\log\frac{\frac{e^{x'\beta}}{1+e^{x'\beta}}}{\frac{1+e^{x'\beta}-e^{x'\beta}}{1+e^{x'\beta}}}=$$

$$\log\frac{e^{x'\beta}}{1+e^{x'\beta}-e^{x'\beta}}=\log e^{x'\beta}=x'\beta=\beta_0+\beta_1x_1+\beta_2x_2+\ldots$$

or

$$o=e^{\beta_0}\cdot e^{\beta_1x_1}\cdot e^{\beta_2x_2}\ldots$$

i.e. an increase of x_i by one unit increases the *odds* o by a factor of e^{β_i}.

In particular if x_i is a binary variable ($x_i\in\{0,1\}$)

$$\begin{aligned}
\frac{\frac{\Pr(Y=1|x_i=1)}{1-\Pr(Y=1|x_i=1)}}{\frac{\Pr(Y=1|x_i=0)}{1-\Pr(Y=1|x_i=0)}} &= \frac{o(Y=1|x_i=1)}{o(Y=1|x_i=0)}\equiv o_{x_i}\\
&= \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_i+\ldots}}{e^{\beta_0+\beta_1x_1+\beta_2x_2+\ldots+0+\ldots}}\\
&= e^{\beta_i}
\end{aligned}$$

Note: The odds-ratio is a ratio of a ratio of probabilities. This is not necessarily intuitive.

Note 2: Only if $\Pr(Y=1|x_i=1) \approx \Pr(Y=1|x_i=0)$ we have

$$\frac{\frac{\Pr(Y=1|x_i=1)}{1-\Pr(Y=1|x_i=1)}}{\frac{\Pr(Y=1|x_i=0)}{1-\Pr(Y=1|x_i=0)}} \approx \frac{\Pr(Y=1|x_i=1)}{\Pr(Y=1|x_i=0)}$$

i.e. the *odds-ratio* approximates the *risk-ratio*. The risk-ratio might be more intuitive.

E.g. $\Pr(Y=1|x_i=1)=0.9$ and $\Pr(Y=1|x_i=0)=0.1$ then

$$\frac{0.9}{0.1} = 9 \quad \text{(risk-ratio)}$$

$$\frac{\frac{\Pr(Y=1|x_i=1)}{1-\Pr(Y=1|x_i=1)}}{\frac{\Pr(Y=1|x_i=0)}{1-\Pr(Y=1|x_i=0)}} = \frac{\frac{0.9}{0.1}}{\frac{0.1}{0.9}} = 81 \quad \text{(odds-ratio)}$$

8.4.1 Odds ratios with interactions

Assume $\Pr(Y=1|x) = F(\beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2)$ with F the logistic function. Then

$$o_{x_1} = \frac{o(Y=1|x_1=1)}{o(Y=1|x_1=0)} = \frac{e^{\beta_1+\beta_2 x_2+\beta_{12}x_2}}{e^{\beta_2 x_2}} = e^{\beta_1+\beta_{12}x_2}$$

hence

$$\frac{o_{x_1|x_2=1}}{o_{x_1|x_2=0}} = \frac{e^{\beta_1+\beta_{12}}}{e^{\beta_1}} = e^{\beta_{12}}$$

i.e. $e^{\beta_{12}}$ is a ratio of two odds ratios (which, in turn, are ratios of ratios of probabilities).

8.4.2 Confidence intervals

```
exp(coef(lr))

 (Intercept)            x
3.767015e-03 1.064459e+08
```

Confidence intervals can be calculated similarly:

```
exp(confint(lr))

                        2.5 %       97.5 %
(Intercept)     0.00007483879 3.998610e-02
x           56340.31979050595 4.418568e+13
```

Note: This interpretation is possible for the logistic link function only.

8.5 Literature

- Stock, Watson. “Introduction to Econometrics”. Chapter 11.
- William H. Greene. “Econometric Analysis”. Chapter 17.

Appendix 8.A Examples for the lecture

Example 1 You want to measure the effect of a continuous effort *x* on success *y* where *y=1* in case of a success and *y=0* in case of a non-success. With R which command yields the desired result?

- None of the following is correct.
- `glm(y ~ x)`
- `glm(y ~ x,family=binomial(link=logit))`
- `glm(y ~ x,family=poisson(link=log))`
- `glm(y ~ logit(x)`

Example 2 You obtain the following output from a *glm* model:

```
Call: glm(formula=y~x,family=binomial(link=logit))
              Estimate   Std. Error   z value   Pr(>|z|)
(Intercept)        -10        40000     -0.00       1.00
          x        120       200000      0.00       1.00
```

Which statements are correct?

- The effect of *x* on *y* is not significant.
- The marginal effect of *x* on *y* is 120.
- The marginal effect of *x* on $\Pr(y=1)$ is $e^{120x-10}$.
- The marginal effect of *x* on $\Pr(y=1)/\Pr(y=0)$ is $e^{120x-10}$.
- The marginal effect of *x* on $\Pr(y=1)/\Pr(y=0)$ is 120.

Example 3 You use JAGS to estimate a logit model. Consider the following model (α and β are placeholders):

```
for (i in 1:length(y)) { α ; β }
for(k in 1:2) {b[k] ~ dnorm (0,.0001) }
```

Assume that α has the value *y[i] ~ dbern(p[i])*
What should be the value of β:

- None of the following is correct.
- `y[i] ~ plogis(b[1]+b[2]*x[i],0,1)`
- `y[i] <- plogis(b[1]+b[2]*x[i],0,1)`
- `p[i] <- plogis(b[1]+b[2]*x[i],0,1)`
- `p[i] ~ plogis(b[1]+b[2]*x[i],0,1)`

Appendix 8.B Exercises

To solve the next exercises the following commands for R might help: *`data, str, names, with, glm, lrtest, confint, summary, mean, dlogis, dnorm`*. The command *`lrtest`* is provided by the *`lmtest`* library.

Exercise 8.1 *Consider the dataset `Benefits` from the `Ecdat` package.*

- *Explain whether a worker received ui Benefits as a function of `stateur` and `statemb`.*
- *What is the marginal effect of `statemb` at the average?*
- *What is the average marginal effect of `statemb`?*
- *Include the interaction between `statemb` and `statemb`.*
- *Give a 95%-confidence interval for the effect of the interaction.*

Exercise 8.2 *Use the `Mroz` dataset from `Ecdat` to explain whether a woman chooses to work or not as a function of her age, education, experience, her husband's wage as well as the number of children that they have.*

1. *What is the marginal effect of `child6` for an everage woman? Use both the logit and probit specifications.*
2. *What is the value of $f(\bar{x}'\beta)$ for either specification? Use it to approximate the marginal effect of `wageh` for an average woman. Compare to the actual estimate.*
3. *What is the average marginal effect of `child6`? Use both the logit and probit specifications.*
4. *What is the marginal effect of `child6` estimated by the linear probability model? Compare it to the marginal effect of `child6` (again, using both the logit and probit specifications) for the following two women:*

   a) *`wageh = 7.5, agew = 42.5, educw = 12, experience = 10.5, child6 = 0, child618 = 1;`*

b) `wageh = 7.5, agew = 42.5, educw = 12, experience = 10.5, child6 = 3, child618 = 1.`

5. *If you suspect that treating `child6` as a continuous variable may not be a good idea, how would you evaluate its marginal effect for woman 4a above? Use both the logit and probit specifications.*

6. *How can you test if `wageh` and `child618` are jointly significant? Use both the logit and probit specifications.*

Exercise 8.3 *Let `grad` be a dummy variable for whether a student-athlete at a large university graduates in five years. Let `hsGPA` and `SAT` be high school grade point average and SAT score, respectively. Let `study` be the number of hours spent per week in an organized study hall.*

Suppose that, using data on 420 student-athletes, the following logit model is obtained:

$$\hat{\mathrm{P}}(\mathtt{grad}=1|\mathtt{hsGPA},\mathtt{SAT},\mathtt{study})=\mathcal{L}(-1.17+.24\mathtt{hsGPA}+.00058\mathtt{SAT}+.073\mathtt{study})$$

where $\mathcal{L}(z)=\exp(z)/(1+\exp(z))$ *is the logistic function.*

Holding `hsGPA` fixed at 3.0 and `SAT` fixed at 1 200, compute the estimated difference in the graduation probability for someone who spent 10 hours per week in study hall and someone who spent 5 hours per week.

(This exercise is taken from Wooldridge, Introductory Econometrics, 17.2).

9 More on discrete choice

9.1 Logistic regression:

9.1.1 The Maximum Likelihood method

$$\begin{aligned}\Pr(Y=1|x)&=F(X'\beta)\\ F^{-1}(\Pr(Y=1|x))&=X'\beta\\ Y\sim\underbrace{\text{binom}}_{\text{family}}&(\underbrace{F}_{\text{link}}(X'\beta))\end{aligned}$$

```
data(Participation,package="Ecdat")
glm(lfp=="yes" ~ lnnlinc,data=Participation,family=binomial(link=logit))


Call:  glm(formula = lfp == "yes" ~ lnnlinc, family = binomial(link = logit),
    data = Participation)
```

```
Coefficients:
(Intercept)      lnnlinc
     9.6276      -0.9165

Degrees of Freedom: 871 Total (i.e. Null);  870 Residual
Null Deviance:      1203
Residual Deviance: 1176   AIC: 1180
```

Different coefficients in logit and probit

$$\Pr(Y = 1|x) \quad = \quad F(x'\beta)$$

Let $-u \sim F$.
By definition: $\Pr(Y = 1|x) = F(x'\beta) \equiv \Pr(x'\beta > -u) = \Pr(\underbrace{x'\beta + u}_{\text{latent } y^l} > 0)$

Note: we can normalise the variance of u to any number — we only have to adjust β.

$y^l/\sigma = x'\beta/\sigma + u/\sigma$ is the same model.

Different F (probit, logit) have different σ, hence different β.

```
glm(lfp=="yes" ~ lnnlinc,data=Participation,
    family=binomial(link=logit))


Call:  glm(formula = lfp == "yes" ~ lnnlinc, family = binomial(link = logit),
    data = Participation)

Coefficients:
(Intercept)      lnnlinc
     9.6276      -0.9165

Degrees of Freedom: 871 Total (i.e. Null);  870 Residual
Null Deviance:      1203
Residual Deviance: 1176   AIC: 1180
```

Logistic distribution has standard deviation $\pi/\sqrt{3}$.

```
glm(lfp=="yes" ~ lnnlinc,data=Participation,
    family=binomial(link=probit))


Call:  glm(formula = lfp == "yes" ~ lnnlinc, family = binomial(link = probit),
    data = Participation)

Coefficients:
(Intercept)      lnnlinc
```

```
     5.9706      -0.5686

Degrees of Freedom: 871 Total (i.e. Null);  870 Residual
Null Deviance:     1203
Residual Deviance: 1176  AIC: 1180
```

Normal distribution has standard deviation 1.

$$\beta_{\text{logit}} \approx \pi/\sqrt{3} \cdot \beta_{\text{probit}}$$

Maximum Likelihood What the ML estimator actually does:

$$\Pr(\forall i : Y_i = y_i|X) = \prod_{y_i=1} F(x_i'\beta) \cdot \prod_{y_i=0} (1 - F(x_i'\beta))$$

thus, the likelihood function is

$$L(\beta|\{y, X\}) = \prod_i (F(x_i'\beta))^{y_i} \left(1 - F(x_i'\beta)\right)^{1-y_i}$$

take logs

$$\log L = \sum_i (y_i \log F(x_i'\beta) + (1 - y_i) \log (1 - F(x_i'\beta)))$$

take the first derivative

$$\frac{d \log L}{d\beta} = \sum_i \left(\frac{y_i f(x_i'\beta)}{F(x_i'\beta)} + (1 - y_i) \frac{-f(x_i'\beta)}{1 - F(x_i'\beta)} \right) x_i = 0$$

in the logit case, e.g. (since in the logit case $F(x) = \dfrac{e^x}{1 + e^x}$)

$$\begin{aligned} \frac{d \log L}{d\beta} &= \sum_i \left(\frac{y_i f(x_i'\beta)}{F(x_i'\beta)} + (1 - y_i) \frac{-f(x_i'\beta)}{1 - F(x_i'\beta)} \right) x_i = \ldots \\ &= \sum_i (y_i - F(x_i'\beta))\, x_i = 0 \end{aligned}$$

For the logit and the probit model the Hessian is always negative definite.
The numerical optimisation is well behaved.

9.1.2 Misspecification in OLS and in Binary choice

OLS: If the true model is

$$y = X_1\beta_1 + X_2\beta_2 + u$$

and $X_2\beta_2$ is omitted then the expected estimate for β_1 is

$$E[\hat{\beta}_1] = \beta_1 + (X_1'X_1)^{-1} X_1'X_2\beta_2$$

which is β_1 if

- $\beta_2 = 0$
- or $X_1'X_2 = 0$, i.e. X_1 and X_2 are orthogonal.

Binary choice: Only the first holds!

This is bad news. With OLS we could drop a variable, even if the variable did matter. We only had to assume that variables we dropped are not correlated with variables which are still in the model.

Now everything that matters must be in the model.

9.1.3 Confidence intervals

As with linear models we can calculate confidence intervals for our estimated coefficients (and as many people prefer confidence intervals over p-values, we might want to do this). *confint* from the standard `stats` package provides (symmetric) Wald confidence intervals. With generalised linear models the likelihood function need not be symmetric. The *MASS* library contains a more precise method to determine confidence intervals for generalised linear models.

```
library(MASS)
(est<-glm(lfp=="yes" ~ lnnlinc,data=Participation,family=binomial(link=logit)))


Call:  glm(formula = lfp == "yes" ~ lnnlinc, family = binomial(link = logit),
    data = Participation)

Coefficients:
(Intercept)      lnnlinc
     9.6276      -0.9165

Degrees of Freedom: 871 Total (i.e. Null);  870 Residual
Null Deviance:     1203
Residual Deviance: 1176   AIC: 1180

confint(est)


                2.5 %     97.5 %
(Intercept)  5.839321 13.5788118
lnnlinc     -1.286580 -0.5619307
```

9.1.4 Bayesian logistic regression

Remember: OLS

```
reg.model<-'model {
for (i in 1:length(y)) {
        y[i] ~ dnorm(beta0 + beta1*x[i],tau)
    }
    beta0 ~ dnorm (0,.0001)
    beta1 ~ dnorm (0,.0001)
    tau   ~ dgamma(.01,.01)
}'
```

The following specifications are equivalent:

$$\Pr(Y = 1|x) = F(X'\beta)$$

$$F^{-1}(\Pr(Y = 1|x)) = X'\beta$$

$$Y \sim \underbrace{\text{binom}}_{\text{family}}(\underbrace{F}_{\text{link}}(X'\beta))$$

```
modelL <- 'model {
 for (i in 1:length(y)) {
        y[i] ~ dbern(p[i])
        p[i] <- plogis(beta0+beta1*x[i],0,1)
    }
    beta0    ~ dnorm (0,.0001)
    beta1    ~ dnorm (0,.0001)
}'
```

When we define the data in our Bayesian model, we have to make sure that data can only contain numbers. Here we have to translate “yes” and “no” into 1 and 0.

We also include the *glm* module to make the sampler faster.

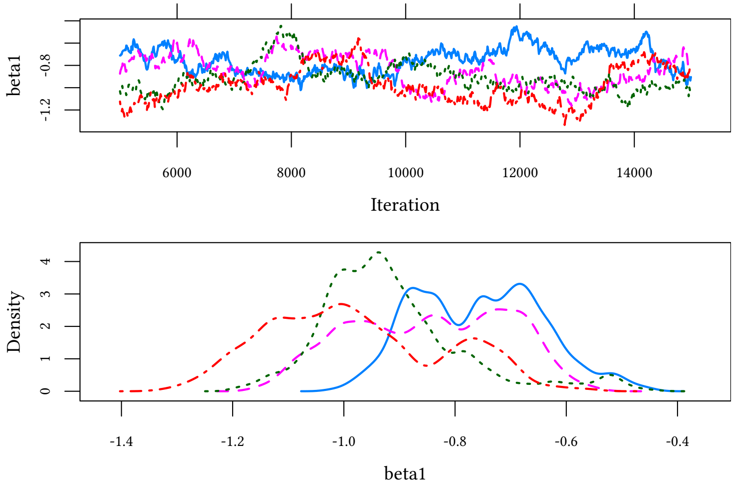
No demeaning

```
library(runjags)
dataL0<-with(Participation,
             list(y=ifelse(lfp=="yes",1,0),
                  x=lnnlinc))
bayesL0<-run.jags(model=modelL,
                  modules="glm",inits=initJags,
                  data=dataL0,
                  monitor=c("beta0","beta1"))
```

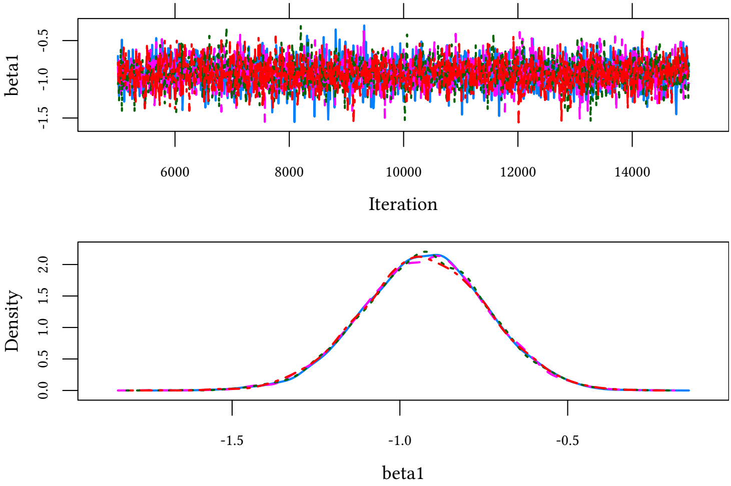
Demeaning

```
library(runjags)
dataL<-with(Participation,
            list(y=ifelse(lfp=="yes",1,0),
                  x=lnnlinc-mean(lnnlinc)))
bayesL<-run.jags(model=modelL,
                 modules="glm",inits=initJags,
                 data=dataL,
                 monitor=c("beta0","beta1"))
```

```
plot(bayesL0,var="beta1",plot.type=c("trace","density"),layout=c(1,2))
```



```
plot(bayesL,var="beta1",plot.type=c("trace","density"),layout=c(1,2))
```



```
bayesL0


JAGS model summary statistics from 40000 samples (chains = 4; adapt+burnin = 5000):

      Lower95  Median  Upper95     Mean       SD Mode    MCerr MC%ofSD
beta0  6.0837  9.2659    12.53   9.1787   1.7045   --   0.3005    17.6
beta1 -1.1798 -0.8824 -0.57554 -0.87449 0.15963   -- 0.028243    17.7

      SSeff   AC.10   psrf
beta0    32 0.98397 1.3286
beta1    32 0.98386 1.3283

Total time taken: 14.7 seconds
```

```
bayesL


JAGS model summary statistics from 40000 samples (chains = 4; adapt+burnin = 5000):

        Lower95   Median Upper95     Mean       SD Mode      MCerr
beta0 -0.29949 -0.16666 -0.0298  -0.1668 0.069125   -- 0.00043812
beta1   -1.285 -0.92365 -0.5558 -0.92524  0.18648   --   0.001175

      MC%ofSD SSeff      AC.10   psrf
beta0     0.6 24894  0.0018092 1.0001
beta1     0.6 25188 -0.0020817      1

Total time taken: 15.5 seconds
```

```
confint(glm(lfp=="yes" ~ lnnlinc,data=Participation,family=binomial(link=logit)))

                2.5 %     97.5 %
(Intercept)  5.839321 13.5788118
lnnlinc     -1.286580 -0.5619307
```

The same for the probit model

```
modelP <- 'model {
 for (i in 1:length(y)) {
        y[i] ~ dbern(p[i])
        p[i] <- pnorm(beta0+beta1*x[i],0,1)
    }
    beta0   ~ dnorm (0,.0001)
    beta1   ~ dnorm (0,.0001)
}'
bayesP<-run.jags(model=modelP,modules="glm",
                 inits=initJags,
                 data=dataL,
                 monitor=c("beta0","beta1"))
```

```
bayesP$summary$quantiles[,c("2.5%","97.5%")]

            2.5%       97.5%
beta0 -0.1893351 -0.02152194
beta1 -0.7910998 -0.35416947
```

```
confint(glm(lfp ~ lnnlinc,
            family=binomial(link="probit"),
            data=Participation))

                2.5 %     97.5 %
(Intercept)  3.655146  8.3364864
lnnlinc     -0.790139 -0.3518155
```

9.2 Count data

9.2.1 Model

Poisson process

(birth-only process):

- During one unit of time on average λ arrivals occur.
- During a time span of length τ on average $\tau\lambda$ arrivals occur.

  Note that this process has no memory!

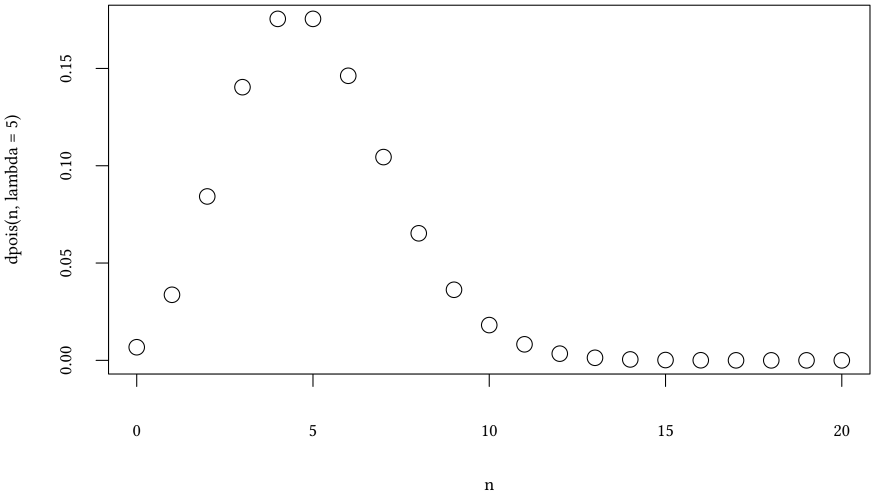
Be $N(t)$ the number of arrivals until time t

$$\Pr\left(N(t+\tau) - N(t) = y\right) = \frac{e^{-\lambda\tau}(\lambda\tau)^y}{y!}$$

Poisson distribution:

$$\Pr(Y_i = y_i|\lambda) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

```r
n<-0:20;
plot(n,dpois(n,lambda=5),t="p")
```



Link function

It is convenient to assume a logarithmic link:

$$\begin{aligned}
\text{link } x_i'\beta \text{ to } \lambda_i: \qquad \log\lambda_i = x_i'\beta \quad &\Leftrightarrow \quad \lambda_i = e^{x_i'\beta} \quad (>0) \\
\Pr(Y_i = y_i|x_i) &= \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} = \frac{e^{-e^{x_i'\beta}}\left(e^{x_i'\beta}\right)^{y_i}}{y_i!} \\
\log L &= \sum_{i=1}^{n} -e^{x_i'\beta} + y_i x_i'\beta - \log y_i! \\
\frac{\partial \log L}{\partial\beta} &= \sum_{i=1}^{n}\left(y_i - e^{x_i'\beta}\right)x_i = 0
\end{aligned}$$

The Hessian is negative definite and Newton's method converges quickly

9.2.2 Illustration — the Fair data

- *nbaffairs* number of affairs in past year

We will assume that this variable follows a Poisson distribution.

```
library(MASS)
data(Fair)
table(Fair$nbaffairs)


  0   1   2   3   7  12 
451  34  17  19  42  38 
```

nbaffairs number of affairs in past year

rate self rating of marriage, from 1 (very unhappy) to 5 (very happy)

religious how religious, from 1 (anti) to 5 (very)

ym number of years married

```
est<-glm(nbaffairs ~ rate + religious + ym ,
    family=poisson(link=log),data=Fair)
summary(est)


Call:
glm(formula = nbaffairs ~ rate + religious + ym, family = poisson(link = log), 
    data = Fair)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
```

```
(Intercept)  2.174078   0.144487   15.05   <2e-16 ***
rate        -0.402186   0.027337  -14.71   <2e-16 ***
religious   -0.363778   0.030737  -11.84   <2e-16 ***
ym           0.075631   0.006832   11.07   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2925.5  on 600  degrees of freedom
Residual deviance: 2401.6  on 597  degrees of freedom
AIC: 2903.6

Number of Fisher Scoring iterations: 6
```

As with other models, we have the usual extractor functions:

```
confint(est)
```

```
                  2.5 %      97.5 %
(Intercept)  1.88862169  2.45506542
rate        -0.45568022 -0.34849566
religious   -0.42418295 -0.30366524
ym           0.06230564  0.08909739
```

9.2.3 Poisson versus negative binomial

A simpler model where *nbaffairs* is only explained by a constant:

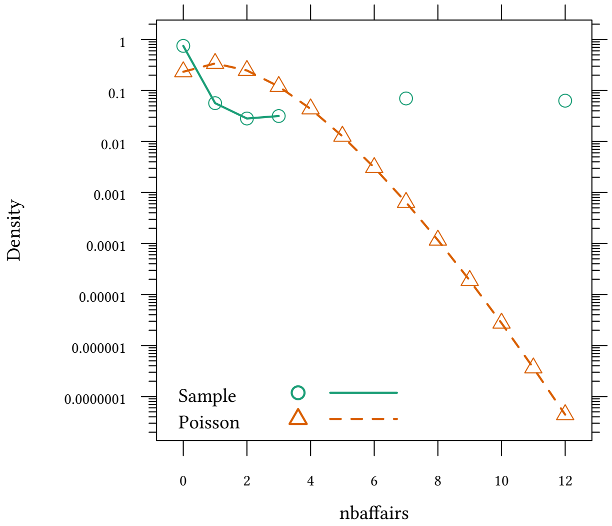
```
summary(est0<-glm(nbaffairs ~ 1,family=poisson(link=log),data=Fair))
```

```


Call:
glm(formula = nbaffairs ~ 1, family = poisson(link = log), data = Fair)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3756     0.0338   11.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2925.5  on 600  degrees of freedom
Residual deviance: 2925.5  on 600  degrees of freedom
AIC: 3421.4

Number of Fisher Scoring iterations: 6
```

```
nba<-sort(unique(Fair$nbaffairs))
rbind(table(Fair$nbaffairs),
   nba,
   round(dpois(nba,exp(coef(est0)))*nrow(Fair),0))

      0   1   2  3  7 12
    451  34  17 19 42 38
nba   0   1   2  3  7 12
    140 204 149 72  0  0
```

Our model predicts too few extreme values (0, 7, 12) and too many intermediate values (1,2,3). Compared with the poisson distribution we have too many zeroes in our data. We also have too many observations where *nbaffairs* is large: Overdispersion.

```
fTab<-within(with(Fair,aggregate(nbaffairs,list(nbaffairs=nbaffairs),length)),
             x<-x/sum(x))
fTab<-merge(as.data.frame(list(nbaffairs=0:max(fTab$nbaffairs))),fTab,all.x=TRUE)
fTab<-within(fTab,pois<-dpois(nbaffairs,exp(coef(est0))))
xyplot(x+pois~nbaffairs,data=fTab,t=c("p","a"),
       auto.key=list(lines=TRUE,x=.02,y=.02,corner=c(0,0),text=c("Sample","Poisson")),
       ylab="Density",
       scales=list(y = list(log=10)),yscale.components = yscale.components.log10ticks)
```



Poisson distribution: Pois(λ)

Mean: λ

Variance: λ

Negative binomial distribution: $\mathrm{NB}(\mu, \theta)$

Mean: μ

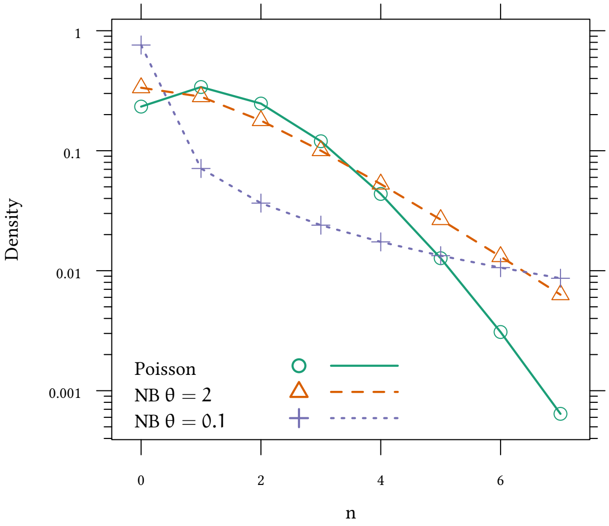
Variance: $\mu + \mu^2/\theta$

Poisson is a special case of NB:

$$\lim_{\theta\to\infty} \mathrm{NB}(\mu, \theta) = \mathrm{Pois}(\mu)$$

$\mu = 1.46$:

```
mu<-exp(coef(est0))
n<-0:7
theta<-c(2,1/10)
p<-dpois(n,mu)
nb1<-dnbinom(n,mu=mu,size=theta[1])
nb5<-dnbinom(n,mu=mu,size=theta[2])
xyplot(p+nb1+nb5~n,type=c("p","l"),
       auto.key=list(lines=TRUE,x=.02,y=.02,corner=c(0,0),
                     text=c("Poisson",paste("NB $\\theta=",theta,"$"))),
       ylab="Density",scales=list(y = list(log=10)),
       yscale.components = yscale.components.log10ticks)
```

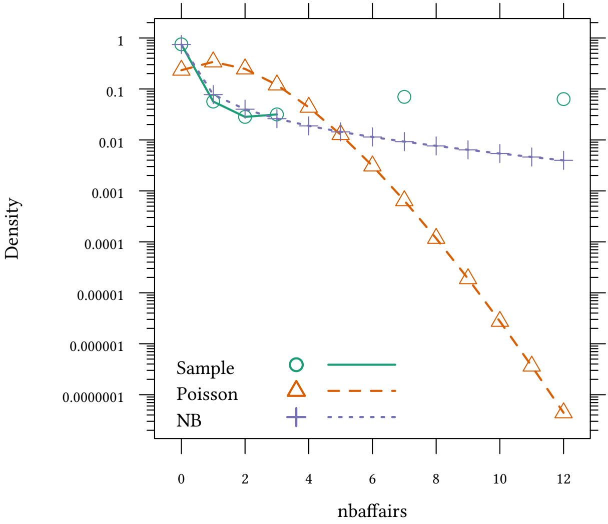


```
summary(est0.nb<-glm.nb(nbaffairs ~ 1,data=Fair ))


Call:
```

```
glm.nb(formula = nbaffairs ~ 1, data = Fair, init.theta = 0.1119600337,
    link = log)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3756     0.1265   2.969  0.00299 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.112) family taken to be 1)

    Null deviance: 330.24  on 600  degrees of freedom
Residual deviance: 330.24  on 600  degrees of freedom
AIC: 1506.3

Number of Fisher Scoring iterations: 1


              Theta:  0.1120
          Std. Err.:  0.0119

 2 x log-likelihood:  -1502.3460
```

$\mu = 1.46, \theta = 0.112$:

```
fTab<-within(fTab,nb<-dnbinom(nbaffairs,mu=exp(coef(est0.nb)),size=est0.nb$theta))
xyplot(x+pois+nb~nbaffairs,data=fTab,t=c("p","a"),
       auto.key=list(lines=TRUE,x=.02,y=.02,corner=c(0,0),
                     text=c("Sample","Poisson","NB")),
       ylab="Density",
       scales=list(y = list(log=10)),
       yscale.components = yscale.components.log10ticks)
```



```
summary(est.nb<-glm.nb(nbaffairs ~ rate + religious + ym,data=Fair ))


Call:
glm.nb(formula = nbaffairs ~ rate + religious + ym, data = Fair,
    init.theta = 0.1418819265, link = log)

Coefficients:
            Estimate Std. Error z value  Pr(>|z|)
(Intercept)  2.36871    0.55987   4.231 0.0000233 ***
rate        -0.42355    0.10748  -3.941 0.0000813 ***
religious   -0.43516    0.10408  -4.181 0.0000290 ***
ym           0.08531    0.02242   3.806  0.000141 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.1419) family taken to be 1)

    Null deviance: 389.62  on 600  degrees of freedom
Residual deviance: 339.33  on 597  degrees of freedom
AIC: 1467.2

Number of Fisher Scoring iterations: 1


              Theta:  0.1419
          Std. Err.:  0.0158

 2 x log-likelihood:  -1457.2180
```

Likelihood ratio test:

$$2(\log L_1 - \log L_0) \overset{\text{approx.}}{\sim} \chi^2_{k_1 - k_0}$$

```
pchisq(2*(logLik(est.nb)-logLik(est)),lower.tail=FALSE,df=1)

'log Lik.' 9.706931e-315 (df=5)
```

9.2.4 Bayesian count data

Remember: OLS

```
reg.model<-'model {
for (i in 1:length(y)) {
        y[i] ~ dnorm(beta0 + beta1*x[i],tau)
    }
    beta0 ~ dnorm (0,.0001)
    beta1 ~ dnorm (0,.0001)
    tau   ~ dgamma(.01,.01)
}'
```

Bayesian Poisson

$$\Pr(Y_i = y_i|\lambda) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

$$\log \lambda_i = x_i'\beta$$

```
pois.model <- 'model {
 for (i in 1:length(y)) {
        y[i] ~ dpois(lambda[i])
        lambda[i] <- exp(beta[1]+beta[2]*rate[i]+
                         beta[3]*religious[i]+beta[4]*ym[i])
    }
 for (k in 1:4) {
    beta[k]   ~ dnorm (0,.0001)
  }
}'
pois.data<-with(Fair,list(y=nbaffairs,rate=rate-mean(rate),
           religious=religious-mean(religious),ym=ym-mean(ym)))
pois.jags<-run.jags(model=pois.model,data=pois.data,inits=initJags,
                    monitor="beta")
```

```
pois.jags$summary$quantiles[,c(1,3,5)]
```

```
               2.5%         50%       97.5%
beta[1] -0.01226990  0.07538280  0.15774748
beta[2] -0.45606613 -0.40241000 -0.34855152
beta[3] -0.42447120 -0.36384600 -0.30374590
beta[4]  0.06235428  0.07568285  0.08901146

confint(est)

                  2.5 %      97.5 %
(Intercept)  1.88862169  2.45506542
rate        -0.45568022 -0.34849566
religious   -0.42418295 -0.30366524
ym           0.06230564  0.08909739
```

```
gelman.diag(pois.jags)

Potential scale reduction factors:

        Point est. Upper C.I.
beta[1]          1          1
beta[2]          1          1
beta[3]          1          1
beta[4]          1          1

Multivariate psrf

1
```

Bayesian Negative Binomial

Different ways to parameterise NB:

$$\begin{aligned}\mathrm{NB}_{\mathtt{glm}}(\mu,\theta) =& \mathrm{NB}_{\mathrm{JAGS}}(p,\theta) = \\ =& \mathrm{NB}(\mu,\sigma^2) = \ldots \\ p =& \frac{\theta}{\theta+\mu} \\ \sigma^2 =& \mu + \frac{\mu^2}{\theta}\end{aligned}$$

```
negbin.model <- 'model {
  for(i in 1:length(y)) {
    y[i] ~ dnegbin(p[i],theta)
    p[i] <- theta/(theta+mu[i])
    mu[i] <- exp(beta[1]+beta[2]*rate[i]+
                 beta[3]*religious[i]+beta[4]*ym[i])
  }
  for (k in 1:4) {
    beta[k] ~ dnorm(0,.0001)
    }
  theta ~ dgamma(0.01,0.01)
}'
negbin.jags<-run.jags(model=negbin.model,data=pois.data,
           inits=initJags,monitor=c("beta","theta"))
```

```
negbin.jags$summary$quantiles[,c(1,3,5)]

              2.5%       50%      97.5%
beta[1] -0.1638697  0.062788  0.3081551
beta[2] -0.6443793 -0.427715 -0.2175550
beta[3] -0.6637210 -0.437648 -0.2239360
beta[4]  0.0381666  0.086401  0.1346971
theta    0.1109419  0.138245  0.1718422
```

```
confint(est.nb)

                  2.5 %     97.5 %
(Intercept)  1.25335497  3.5470148
rate        -0.64004901 -0.2176862
religious   -0.65176584 -0.2250606
ym           0.03875986  0.1330084

est.nb[["theta"]]

[1] 0.1418819
```

```
gelman.diag(negbin.jags)

Potential scale reduction factors:

        Point est. Upper C.I.
beta[1]          1          1
beta[2]          1          1
beta[3]          1          1
beta[4]          1          1
theta            1          1

Multivariate psrf

1
```

9.3 Multinomial (polytomous) logit

9.3.1 Motivation and background

Multinomial logit

- choices are mutually exclusive

- choices are exhaustive
- choices are finite

Problems:

- one can map problems that do not look mutually exclusive or not exhaustive into a problem that is

  E.g.: heating modes: gas / oil / wood / electricity

  What about households which use, e.g., gas + electricity →

  - introduce an additional category
  - ask for 'primary source of heating'

  Some households do not use any of the above:

  - introduce an additional category

Random utility models
Can we tell a story like in the logit/probit case?

Random utility model
Latent variables:

$$\begin{array}{rcl} \eta_1 & = & x'\beta_1 + \xi_1 \\ \eta_2 & = & x'\beta_2 + \xi_2 \\ \eta_3 & = & x'\beta_3 + \xi_3 \\ & \vdots & \end{array}$$

The decision maker chooses alternative k if $\eta_k \geq \eta_j$ for all j

Normalisations
Note: these models are equivalent to their affine transformations.

- Normalise $x'\beta_1 = 0$.
- If ξ_j are i.i.d. we often normalise their variance to a convenient value.

  → different distributions for ξ lead to different coefficients

  $\beta_{\text{logit}} \approx \pi/\sqrt{6} \times \beta_{\text{probit}}$

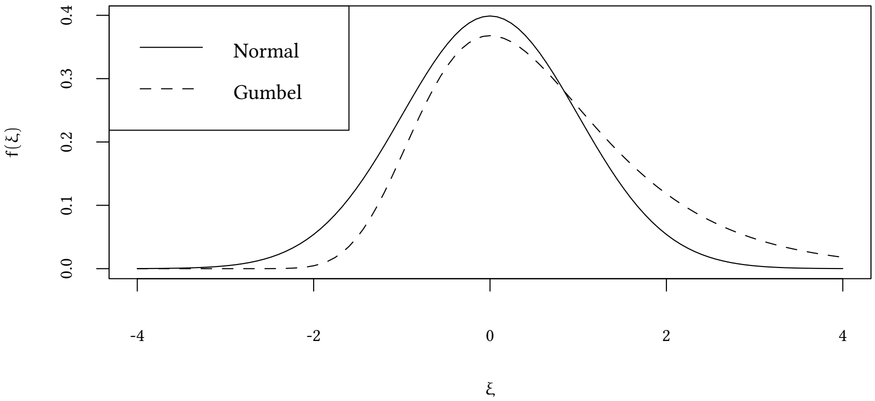
Differences

Let us look at the differences between two alternatives:

$$\nu_{kj} = \eta_k - \eta_j = x'(\beta_k - \beta_j) + \xi_k - \xi_j$$

- $\xi \sim N(0,1)$: $\xi_k - \xi_j$ has variance 2 and covariance 1 (for $k \neq j$)

```r
dgumbel<-function(x) exp(-exp(-x)-x)
plot(dnorm,-4,4,ylab="$f(\\xi)$",xlab="$\\xi$")
curve(dgumbel,add=TRUE,lty=2)
legend("topleft",c("Normal","Gumbel"),lty=1:2)
```



- $\xi \sim \text{Gumbel}\left(F_{\text{Gumbel}}(\xi) = e^{-e^{-\xi}}\right)$ then
  - the difference ν_{ki} follows a logistic distribution

$$\Pr(y = k|\xi_k) = \prod_{j \neq k} \underbrace{F_{\text{Gumbel}}(x'(\beta_k - \beta_j) + \xi_k)}_{\Pr(\eta_j < \eta_k)}$$

average over ξ_k

$$\Pr(y = k) = \int f_{\text{Gumbel}}(\xi_k) \prod_{j \neq k} F_{\text{Gumbel}}(x'(\beta_k - \beta_j) + \xi_k)\, d\xi_k$$

$$\Pr(y = k) = \frac{e^{x'\beta_k}}{\sum_{i=1}^{m} e^{x'\beta_i}}$$

→ we get the following multinomial logit (McFadden)

$$\begin{aligned}
\Pr(y=1) &= \frac{e^{x'\beta_1}}{\sum_{k=1}^m e^{x'\beta_k}}\\
\Pr(y=2) &= \frac{e^{x'\beta_2}}{\sum_{k=1}^m e^{x'\beta_k}}\\
\Pr(y=3) &= \frac{e^{x'\beta_3}}{\sum_{k=1}^m e^{x'\beta_k}}\\
&\vdots
\end{aligned}$$

- $0 < \Pr(y=k) < 1$
- $\sum_k \Pr(y=k) = 1$

$\uparrow$ β_k are not identified

Normalise: $\rightarrow$

$$\begin{aligned}
\Pr(y=1) &= \frac{1}{1+\sum_{k=2}^m e^{x'\beta_k}}\\
\Pr(y=2) &= \frac{e^{x'\beta_2}}{1+\sum_{k=2}^m e^{x'\beta_k}}\\
\Pr(y=3) &= \frac{e^{x'\beta_3}}{1+\sum_{k=2}^m e^{x'\beta_k}}\\
&\vdots
\end{aligned}$$

the odds ratios are:

$$\frac{\Pr(y=k)}{\Pr(y=1)} = e^{x'\beta_k}$$

This is a strong assumption on the error terms.

Indenpendence from irrelevant alternatives — IIA

Example:

- Dependent = choice of travel mode
- unobservable = personal preference for/against means of mass transportation (tube/train).

$\frac{\Pr(y=\text{tube})}{\Pr(y=1)} = e^{x'\beta_\text{tube}}$ $\frac{\Pr(y=\text{train})}{\Pr(y=1)} = e^{x'\beta_\text{train}}$

- → choices/error terms are correlated.

→ multinomial logit can represent systematic variation of choices (explained by observed characteristics) but *not* systematic individual (unobserved) variation of choices.

The log-likelihood:

With $I_k(y_i) = \begin{cases} 1 & \text{if } y_i = i \\ 0 & \text{otherwise} \end{cases}$

$$\begin{aligned} \log L &= \sum_i \sum_{k=1}^m I_k(y_i) \log \Pr(y_i = k) \\ &= \sum_i \sum_{k=1}^m I_k(y_i) \log \frac{e^{x_i' \beta_k}}{1 + \sum_{k=2}^m e^{x_i' \beta_k}} \end{aligned}$$

this function $\log L$ is globally concave in β (McFadden, 1974)

9.3.2 Example

The purpose of this example is to illustrate an identification problem in the context of multinomial logit. There are different ways to describe the same choices. In the example we see that we use one set of parameters (`mat`) to generate the choices but the estimator gives us a different set of parameters back (`coef(est)`). We also see how these two sets of parameters are related.

Let us first create individual explanatory variables, *x1*, *x2*.

```
N<-100
sd<-10
ex <- cbind(x1=runif(N),
            x2=runif(N))
head(ex)
```

```
            x1        x2
[1,] 0.2875775 0.5999890
[2,] 0.7883051 0.3328235
[3,] 0.4089769 0.4886130
[4,] 0.8830174 0.9544738
[5,] 0.9404673 0.4829024
[6,] 0.0455565 0.8903502
```

The following matrix determines how individual characteristics translate into preferences for three choices:

```
mat<-rbind(c(400,0),
          c(250,200),
          c(100,300))
mat
```

```
     [,1] [,2]
[1,]  400    0
[2,]  250  200
[3,]  100  300
```

```
latent<-(ex %*% t(mat)) +
    sd * cbind(rnorm(N),rnorm(N),
                rnorm(N))
head(latent)

          [,1]     [,2]     [,3]
[1,] 107.92694 213.8803 201.6020
[2,] 317.89089 276.7651 171.1507
[3,] 161.12385 197.3154 178.0962
[4,] 349.73154 417.0811 364.1188
[5,] 366.67073 327.5539 234.5459
[6,]  17.77232 184.6967 274.9725

max.col(latent)

  [1] 2 1 2 2 1 3 2 1 2 1 2 1 1 3 3 1 3 3 3 1 1 1 1 1 1 2 1 1 2 3 1 2
 [33] 2 2 3 2 2 3 3 3 3 2 1 1 3 2 2 1 1 2 3 2 1 3 1 3 3 1 1 1 1 3 3 2
 [65] 1 2 2 1 1 1 2 1 2 3 2 3 2 2 3 3 3 1 3 1 3 2 1 1 2 3 2 1 3 2 3 3
 [97] 1 3 1 2
```

We are expecting this:

```
mat

     [,1] [,2]
[1,]  400    0
[2,]  250  200
[3,]  100  300
```

```
choice <- max.col(latent)
library(nnet)
est<-multinom(choice ~ x1 + x2,as.data.frame(ex))
```

```
Call:
multinom(formula = choice ~ x1 + x2, data = as.data.frame(ex))

Coefficients:
  (Intercept)        x1       x2
2   0.9444688 -25.80588 31.72050
3   0.1557040 -58.59718 52.66552

Residual Deviance: 33.75979
AIC: 45.75979
```

Note that the estimated coefficients are not the matrix of coefficients `mat` that we employed above. However, they are a projection. The estimator normalises the first category to zero

```
mat - cbind(c(1,1,1)) %*% mat[1,]

     [,1] [,2]
[1,]    0    0
[2,] -150  200
[3,] -300  300
```

and sets the variance to one:

```
(mat -  cbind(c(1,1,1)) %*% mat[1,])*pi /
    sqrt(6) / 10

          [,1]     [,2]
[1,]   0.00000  0.00000
[2,] -19.23825 25.65100
[3,] -38.47649 38.47649
```

To access estimation results we have the usual extractor functions:

```
coef(est)

  (Intercept)        x1       x2
2   0.9444688 -25.80588 31.72050
3   0.1557040 -58.59718 52.66552

confint(est)

, , 2

                 2.5 %    97.5 %
(Intercept)  -3.098834  4.987771
x1          -43.459195 -8.152558
x2           11.420190 52.020819

, , 3

                2.5 %     97.5 %
(Intercept)  -4.74507   5.056478
x1          -89.59278 -27.601578
x2           26.23746  79.093575
```

9.3.3 Bayesian multinomial

```
modelM <- 'model {
 for (i in 1:length(y)) {
        for (j in 1:3) { # three different choices
          exb[i,j] <- exp(inprod(beta[,j],ex[i,]))
        }
        y[i] ~ dcat(exb[i,1:3])
    }
    for (k in 1:K) {
        beta[k,1] <- 0 # identifying restriction
    }
    for (j in 2:3) {
      for (k in 1:K) {
         beta[k,j] ~ dnorm(0,.0001)
      }
    }
}'
dataList<-list(y=choice,ex=cbind(1,ex),K=dim(ex)[2]+1)
bayesM <-run.jags(model=modelM,data=dataList,monitor=c("beta"))
```

```
bayesM$summary$quantiles[-c(1,2,3),
                         c("2.5%","50%","97.5%")]

                 2.5%         50%      97.5%
beta[1,2]   -2.916810   1.0036950   5.485993
beta[2,2]  -51.165215 -27.8410000 -13.741013
beta[3,2]   17.886848  34.2814000  59.503305
beta[1,3]   -4.896983  -0.1411975   5.400235
beta[2,3] -114.122250 -69.8986000 -41.405530
beta[3,3]   37.173045  61.8531000  95.136755
```

```
confint(est)

, , 2

                 2.5 %    97.5 %
(Intercept)  -3.098834  4.987771
x1          -43.459195 -8.152558
x2           11.420190 52.020819

, , 3

                2.5 %     97.5 %
(Intercept)  -4.74507   5.056478
x1          -89.59278 -27.601578
x2           26.23746  79.093575
```

9.4 Literature

- John H. Kruschke. “Doing Bayesian Data Analysis. A Tutorial with R, JAGS, and Stan”. Academic Press. 2014. Chapter 21.
- William H. Greene. “Econometric Analysis”. Chapter 16-18.

Appendix 9.A Examples for the lecture

Example 1 You use JAGS to estimate a probit model. Consider the following model (α and β are placeholders):

```
for (i in 1:length(y)) { α ; β }
for(k in 1:2) {b[k] ~ dnorm (0,.0001) }
```

Now *y* is a count variable. You want to use a Poisson model. What should be the value of α?

- None of the following is correct.
- `y[i] <- dpois(p[i])`
- `y[i] ~ dlogis(p[i])`
- `y[i] ~ dprobit(p[i])`
- `y[i] ~ dpois(p[i])`

Example 2 What should in this case be the value of β?

- None of the following is correct.
- `p[i] <- exp(b[1]+b[2]*x)`
- `p[i] ~ exp(b[1]+b[2]*x)`
- `p[i] <- log(b[1]+b[2]*x)`
- `p[i] ~ b[1]+b[2]*x`

Example 3 You estimate the following multinomial logit model:

$$\begin{aligned}\eta_1 &= x'\beta_1 + \xi_1\\ \eta_2 &= x'\beta_2 + \xi_2\\ \eta_3 &= x'\beta_3 + \xi_3\end{aligned}$$

The decision maker chooses alternative k if $\eta_k \geq \eta_j$ for all j. You expect the following:

$$(\beta_1, \beta_2, \beta_3) = \begin{pmatrix} 3 & 4 & 1 \\ 2 & 0 & 6 \end{pmatrix}$$

You obtain the following output in R (α_1 and α_2 are placeholders):

```
multinom(formula = y ~ x1+x2-1)
Coefficients:
   x1   x2
2   1   -2
3  α1   α2
```

- What do you expect for α_1?
- What do you expect for α_2?

Appendix 9.B Exercises

To solve the next exercises the following commands for R might help: `data, str, summary, table, mean, quantile, density, confint, coef, plot, abline, glm`. Furthermore, `coeftest` and `lrtest` from by the `lmtest` library, `run.jags` from`runjags`, `as.mcmcm` from `coda`, `glm.nb` from `MASS`, and `multinom` from `nnet`. You might also need from JAGS the functions `dpois, dnegbin, dcat`.

Exercise 9.1 *Consider the dataset `Doctor` from the package `Ecdat`.*

1. *Use a Bayesian model to explain the number of doctor visits as a function of `access`.*
2. *Give a credible interval for the effect of `access`.*
3. *How probable is it that the coefficient of* $\texttt{access} > 1$*?*
4. *Compare your result with a frequentist model.*
5. *How can you take into account overdispersion?*

Exercise 9.2 *Use the Fishing dataset from Ecdat to model the choice of the fishing mode as a function of income. Use "beach" as the base category.*

1. *Fit the model using the Bayesian framework.*
2. *Fit the model using the frequentist framework.*
3. *What does the model predict for the first person from the sample?*
4. *Which fishing mode was actually chosen by said person?*
5. *What is the marginal effect of `income` for said person?*
6. *What is the average marginal effect of `income`?*
7. *What is the average effect of increasing `income` by 1000?*

Appendix 9.C Ordered probit

9.C.1 Model

We observe whether latent variables $x'\beta$ are in an interval

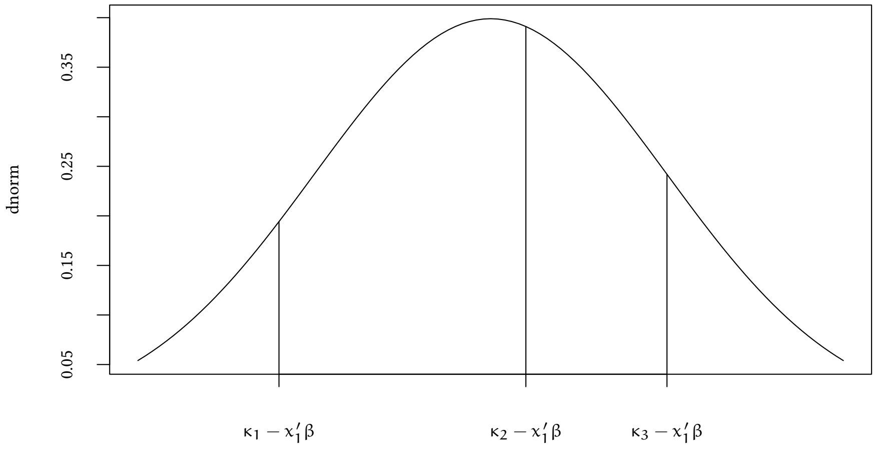
$$\begin{array}{rcl}
\Pr(y_i = 1) & = & \Pr(\kappa_0 < x_i'\beta + u \leq \kappa_1) \\
\Pr(y_i = 2) & = & \Pr(\kappa_1 < x_i'\beta + u \leq \kappa_2) \\
\Pr(y_i = 3) & = & \Pr(\kappa_2 < x_i'\beta + u \leq \kappa_3) \\
 & \vdots &
\end{array}$$

or (solving for u)

$$\begin{aligned}
\Pr(y_i = 1) &= \Pr(\kappa_0 - x_i'\beta < u \leq \kappa_1 - x_i'\beta) \\
\Pr(y_i = 2) &= \Pr(\kappa_1 - x_i'\beta < u \leq \kappa_2 - x_i'\beta) \\
\Pr(y_i = 3) &= \Pr(\kappa_2 - x_i'\beta < u \leq \kappa_3 - x_i'\beta) \\
&\vdots
\end{aligned}$$

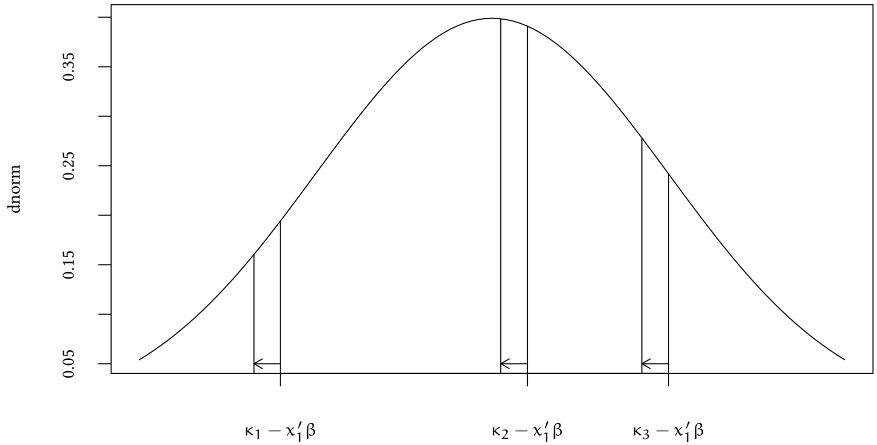
The u can follow any (standard) distribution (logistic, normal, ...)

```
plot(dnorm,-2,2,xaxt="n",xlab=NA)
kappas<-c(-1.2,.2,1)
for(i in 1:length(kappas)) {x<-kappas[i];lines(c(x,x),c(0,dnorm(x)))}
axis(1,kappas,sapply(1:length(kappas),function(d) sprintf("$\\kappa_%d - x_1'\\beta$",d)))
```



Marginal effects:

```
plot(dnorm,-2,2,xaxt="n",xlab=NA)
kappas<-c(-1.2,.2,1)
for(i in 1:length(kappas)) {
    x<-kappas[i];lines(c(x,x),c(0,dnorm(x)))
    y<-kappas[i]-.15;lines(c(y,y),c(0,dnorm(y)))
    arrows(x,.05,y,.05,length=.05)
}
axis(1,kappas,sapply(1:length(kappas),function(d) sprintf("$\\kappa_%d - x_1'\\beta$",d)))
```



The maximum likelihood problem

$$\begin{aligned}
\Pr(y_i = 1) &= \Pr(\kappa_0 - x_i'\beta < u \leq \kappa_1 - x_i'\beta) \\
\Pr(y_i = 2) &= \Pr(\kappa_1 - x_i'\beta < u \leq \kappa_2 - x_i'\beta) \\
\Pr(y_i = 3) &= \Pr(\kappa_2 - x_i'\beta < u \leq \kappa_3 - x_i'\beta) \\
&\vdots
\end{aligned}$$

$$\log L = \sum_i \sum_{k=1}^{m} I_k(y_i) \log \Pr(y_i = k)$$

with $I_k(y_i) = \begin{cases} 1 & \text{if } y_i = i \\ 0 & \text{otherwise} \end{cases}$

9.C.2 Illustration — the Fair data

As an illustration, let us look at a dataset on extramarital affairs, collected by Ray Fair. Two variables from the dataset are

- *ym* number of years married

- `rate` self rating of mariage (unhappy=1...5=happy)

Does the rating of marriage change over time? A naïve approach would be to use OLS and to explain `rate` as a linear function of *ym*.

```
library(MASS)
library(Ecdat)
data(Fair)
lm(rate ~ ym,data=Fair)


Call:
lm(formula = rate ~ ym, data = Fair)

Coefficients:
(Intercept)           ym
    4.32546     -0.04814
```

This approach would assume that all ratings are equidistant. More appropriate is, perhaps, an ordered logistic model...

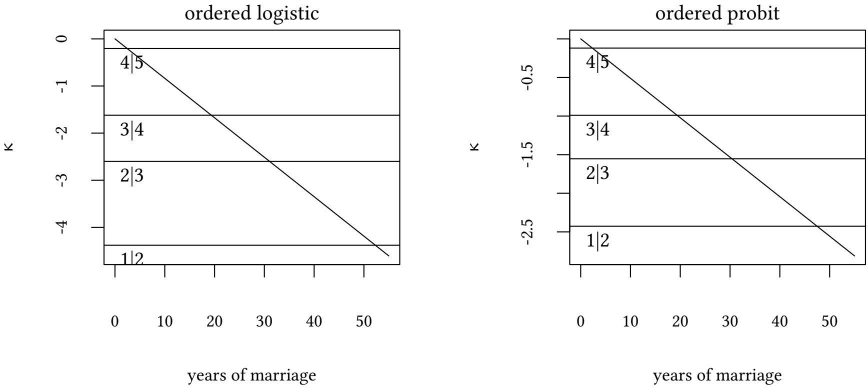
```
(estL<-polr(factor(rate) ~ ym,data=Fair))

Call:
polr(formula = factor(rate) ~ ym, data = Fair)

Coefficients:
         ym
-0.08371391

Intercepts:
       1|2        2|3        3|4        4|5
-4.3786529 -2.5996956 -1.6207810 -0.2043441

Residual Deviance: 1597.27
AIC: 1607.27
```

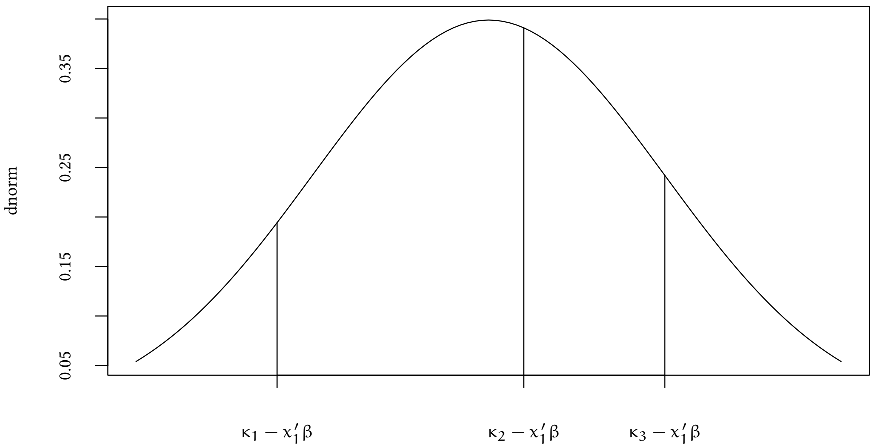
...or an ordered probit:

```
(estP<-polr(factor(rate) ~ ym,data=Fair,method="probit"))

Call:
polr(formula = factor(rate) ~ ym, data = Fair, method = "probit")

Coefficients:
         ym
-0.05110974

Intercepts:
      1|2       2|3       3|4       4|5
-2.427247 -1.552900 -0.990142 -0.119791

Residual Deviance: 1594.99
AIC: 1604.99
```

The following graph illustrates the estimated thresholds κ_i:

```
probFig <- function (est,main) {
  plot(function(x) {x * est$coef},0,55,ylab="$\\kappa$",xlab="years of marriage",main=main)
  for (a in est$zeta) {
    abline(h=a)
    lab=names(est$zeta)[which(est$zeta==a)]
    text(1,a,labels=lab,adj=c(0,1))
  }
}
probFig(estL,main="ordered logistic")
probFig(estP,main="ordered probit")
```





- Dependent variable *y[i]*
- Latent variable *t[i]*
- Independent variable *x[i]*
- Parameters *beta, kappa[j]*

JAGS notation for intervals

```
y[i] ~ dinterval(t[i],kappa)
```

where

t[i]	realisation of latent variable
y[i]	observable rating
kappa	thresholds

If $Y \sim \texttt{dinterval}(t, \kappa)$ then

$$\begin{array}{lll} Y = 0 & \text{if } t \leq \kappa[1] & \\ Y = m & \text{if } \kappa[m] < t \leq \kappa[m+1] & \text{for } 1 \leq m < M \\ Y = M & \text{if } \kappa[M] < t & \end{array}$$

Note: We have to give JAGS possible initial values:

```
dataList<-list(y=Fair$rate-1,x=Fair$ym,K=max(Fair$rate)-1)
initList<-with(dataList,list(t=y+1/2,kappa0=1:K))
```

```
model0 <- 'model {
 for (i in 1:length(y)) {
    y[i] ~ dinterval(t[i],kappa)
    t[i] ~ dnorm(beta*x[i],1)
 }
 for (j in 1:K) {
    kappa0[j] ~ dnorm(0,.0001)
 }
 kappa[1:4] <- sort(kappa0)
 beta ~ dnorm(0,.0001)
}'
dataList<-list(y=Fair$rate-1,x=Fair$ym,K=max(Fair$rate)-1)
initList<-with(dataList,list(t=y+1/2,kappa0=1:K))
bayes0 <-run.jags(model=model0,data=dataList,inits=list(initList,initList),
                  monitor=c("beta","kappa"))
```

```
bayes0$summary$quantiles[,c("2.5%","50%","97.5%")]

               2.5%         50%        97.5%
beta     -0.0617288 -0.04027465 -0.005994167
kappa[1] -2.6400820 -2.28460000 -1.792639500
kappa[2] -1.6685503 -1.41552000 -0.982247425
kappa[3] -1.0998930 -0.83911250 -0.449412175
kappa[4] -0.2153431 -0.00285062  0.399073250

estP

Call:
polr(formula = factor(rate) ~ ym, data = Fair, method = "probit")

Coefficients:
         ym
-0.05110974

Intercepts:
      1|2       2|3       3|4       4|5
-2.427247 -1.552900 -0.990142 -0.119791

Residual Deviance: 1594.99
AIC: 1604.99
```

Convergence is not too exciting

```
bayes0


JAGS model summary statistics from 20000 samples (chains = 2; adapt+burnin = 5000):
```

```
            Lower95      Median    Upper95       Mean       SD  Mode
beta      -0.062806   -0.040275 -0.0075228  -0.039193  0.01464    --
kappa[1]     -2.682     -2.2846    -1.8701    -2.2804  0.21023    --
kappa[2]    -1.6914     -1.4155    -1.0225    -1.3975  0.18359    --
kappa[3]      -1.11    -0.83911   -0.46484   -0.83729  0.17726    --
kappa[4]   -0.24848  -0.0028506    0.33923   0.012609  0.15895    --

              MCerr  MC%ofSD  SSeff     AC.10   psrf
beta      0.0050188     34.3      9   0.86849 1.5322
kappa[1]   0.026083     12.4     65   0.93933 1.4325
kappa[2]   0.053617     29.2     12   0.98852 1.6002
kappa[3]    0.06039     34.1      9   0.99152 1.6347
kappa[4]   0.056333     35.4      8   0.99215   1.51

Total time taken: 5.7 seconds
```

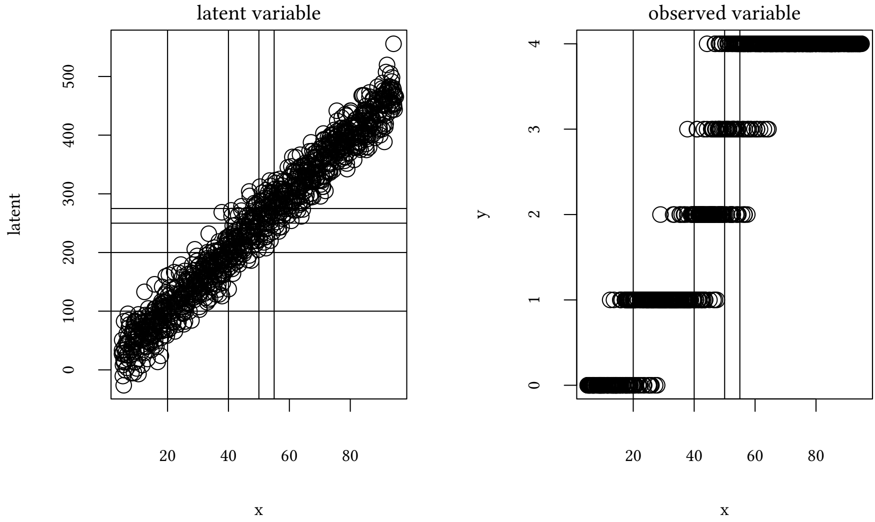
9.C.3 Illustration II — a simulated dataset

In the following we generate a latent variable *latent*. The problem is that we can not observe *latent*. We only see y which indicates a range of possible values for *latent*. These ranges are defined by *cuts*.

```
set.seed(123)
N <-1000
sd<-25
b <-5
x <-runif(N,min=5,max=95)
latent<-b*x+sd*rnorm(N)
cuts<-sort(c(100,200,250,275))
y <- sapply(latent,function(x) sum(x>cuts))
```

The following graph illustrates the problem. On the left we have the latent variable on the vertical axis. Our observed variable is shown in the right diagram. This is the data we use in our regression.

```
plot(latent ~ x,main="latent variable")
abline(h=cuts)
abline(v=cuts/b)
plot (y ~ x,main="observed variable")
abline(v=cuts/b)
```



polr estimates the ordered model. By default the logistic method is used.

```
polr(factor(y) ~ x)

Call:
polr(formula = factor(y) ~ x)

Coefficients:
        x
0.3656912

Intercepts:
     0|1      1|2      2|3      3|4
 7.32400 14.65210 18.03400 19.85187

Residual Deviance: 748.3983
AIC: 758.3983
```

Is this the value we should expect? Remember that the logistic distribution has $\sigma = \pi/\sqrt{3}$. As the coefficient for x we should, hence, expect

```
b/sd*pi/sqrt(3)

[1] 0.3627599
```

and as intercepts

```
cuts/sd*pi/sqrt(3)

[1]  7.255197 14.510395 18.137994 19.951793
```

which is both close to the estimated coefficients.

The standard normal distribution has $\sigma = 1$, hence we get different results with the probit method:

```
polr(factor(y) ~ x,method="probit")

Call:
polr(formula = factor(y) ~ x, method = "probit")

Coefficients:
        x
0.2037362

Intercepts:
      0|1       1|2       2|3       3|4
 4.102613  8.161292 10.050343 11.069073

Residual Deviance: 744.0376
AIC: 754.0376
```

As the coefficient for *x* we should expect

```
b/sd

[1] 0.2
```

and as intercepts

```
cuts/sd

[1]  4  8 10 11
```

10 Mixed effects models

10.1 Terminology

Fixed-Effects, Random-Effects, Mixed-Models FE-only-model:

$$y_j = \beta X_j + u_j \qquad \text{with } u_j \sim N(0, \sigma^2 \mathbf{I}_n)$$

ME-model: There are $i \in \{1 \ldots M\}$ groups of data, each with $j \in \{1 \ldots n_i\}$ observations. Groups i represent one or several factors.

$$y_{ij} = \beta X_{ij} + \gamma_i Z_{ij} + u_{ij} \qquad \text{with } \gamma_i \sim N(0, \Psi), u_{ij} \sim N(0, \Sigma_i)$$

- Fixed effects: The researcher is interested in specific effects β. The researcher does not care/is ignorant about the type of distribution of β. If β is a big object, estimation can be expensive.
- Random effects: The researcher only cares about the distribution of γ. It is sufficient to estimate only the parameters of the distribution of γ (e.g. its standard deviation). This is less expensive/makes more efficient use of the data.
- Mixed model: includes a mixture of fixed and random factors

Terminology Depending on the field, mixed effects models are known under different names:

- Mixed effects models
- Random effects models
- Hierarchical models
- Multilevel models
- Panel models

Why mixed effects models?

- Repeated observation of the same unit:
  - as part of a panel outside the lab
  - participant in the experiment
  - group of participants in an experiments
- Reasons for repeated observations:
  - within observational unit (participant/group) comparisons
  - study the dynamics of a process (market behaviour, convergence to equilibium,...)
  - allow “learning of the game”

A possible experiment Example: Repeated public good game

Question: is there a decay of contributions over time?

- participants in a group of four can contribute to a public good
- 8 repetitions
- random matching in groups of 12
- observe 10 matching groups (120 participants)

In our raw data we have $12 \times 8 \times 10 = 960$ observations.

Problems

- Repeated measurements
- always the same participants
- always the same matching groups

Observations are correlated — OLS requires uncorrelated ϵ

Solution A: Aggregation

- Aggregate 960 observations over 10 matching groups.

Disadvantage:

- Loss of power (only 10 matching groups, not 960 observations)
- Control of individual effects only through randomisation
  (groups/participants might have different and known (even controlled) properties)

It would be nice to know:

- What is the treatment effect (in the example: the effect of time)
- What is an effect due to other observables (e.g. gender, risk aversion, social preferences)
- What is the heterogeneity of participants (due to unobservable differences)
- What is the heterogeneity of groups (e.g. due to contamination in the experiment)

Alternative (more efficient):

- Models with mixed effects

This example: OLS, fixed effects and random effects Indices:

- i individuals $1 \ldots 12$
- k group $1 \ldots 10$
- t time $1 \ldots 8$
- Pooled OLS (pretend ϵ_{ikt} are independent):
  $y_{ikt} = \beta_0 + \beta_1 x_{1,ikt} + \beta_2 x_{2,ikt} + \epsilon_{ikt}$
  with $\epsilon_{ikt} \sim N(0, \sigma)$

- Fixed effects for participants $i \times k$:

  $y_{ikt} = \beta_0 + \beta_1 x_{1,ikt} + \beta_2 x_{2,ikt} + \sum_{i,k} \gamma_{ik} d_{ik} + \epsilon_{ikt}$

  with $\epsilon_{ikt} \sim N(0, \sigma)$

- Random effects for participants $i \times k$:

  $y_{ikt} = \beta_0 + \beta_1 x_{1,ikt} + \beta_2 x_{2,ikt} + \nu_{ik} + \epsilon_{ikt}$

  with $\nu_{ik} \sim N(0, \sigma_\nu)$ and $\epsilon_{ikt} \sim N(0, \sigma)$

Fixed effects

+ captures individual heterogeneity

− only describes heterogeneity in the *sample* (this is not a problem if sample heterogeneity is experimentally controlled, e.g. fixed effect for treatment 1 vs. treatment 2)

− less stable since many coefficients have to be estimated

+ makes no distributional assumptions on heterogeneity

− can be fooled by spurious correlation among X and ν_{ik}

+ unbiased if ν_{ik} and X are dependent

Random effects

+ captures individual heterogeneity

+ estimates heterogeneity in the population

+ more stable since fewer coefficients have to be estimated

− makes distributional assumptions on heterogeneity

+ exploits independence of ν_{ik} and X (if it can be assumed)

− biased if ν_{ik} and X are dependent

Terminology

$$y_{ikt} = \beta_0 + \beta_1 x_{1,ikt} + \beta_2 x_{2,ikt} + \nu_{ik} + \epsilon_{ikt}$$

- Random effects — units ik are selected *randomly* from a population. An *effect* on β_0 means that the mean y depends on the choice of ik (we could also have random effects for $\beta_1, \beta_2, \ldots$).

- Hierarchical/multilevel model — first we explain variance on the level of ik, then on the level of ikt.

  Here we add another level: first we explain variance on the level of k, then ik, then on the level of ikt.

$$y_{ikt} = \beta_0 + \beta_1 x_{1,ikt} + \beta_2 x_{2,ikt} + \nu'_k + \nu_{ik} + \epsilon_{ikt}$$

10.2 Examples

10.2.1 Initial commands

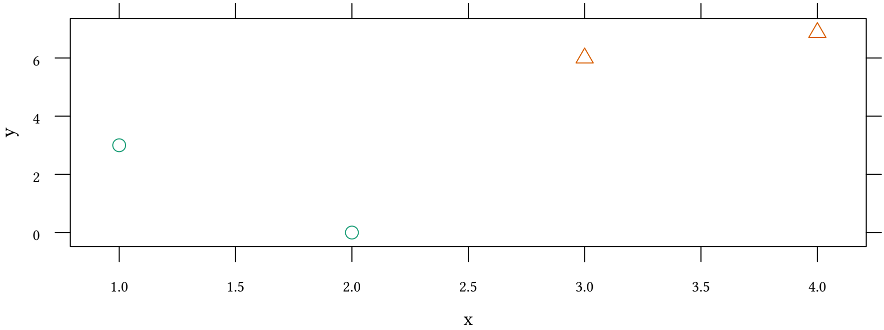
```
bootstrapsize<-100
library(Ecdat)
library(car)
library(Hmisc)
load(file="data/me.Rdata")
```

10.2.2 A small example

The following figure shows the predicted relationship for the various methods. The dataset is very simple. We have only four observations, two from two groups each. The first groups (shown as circles) are at the bottom left of the diagram, the second group (triangles) are top right.

```
simple <- data.frame(cbind(x=c(1,2,3,4),y=c(3,0,6,6.8744),i=c(1,1,2,2)))
```

```
simple

  x      y i
1 1 3.0000 1
2 2 0.0000 1
3 3 6.0000 2
4 4 6.8744 2
```



Let us now look at the following models:

- Pooled OLS:

$$y_{ik} = \beta_0 + \beta_1 x_{ik} + \epsilon_{ik} \text{ with } \epsilon_{ik} \sim N(0, \sigma)$$

- Between OLS (aggregate for each group i):

$$y_i = \beta_0 + \beta_1 x_i + \nu_i \text{ with } \nu_i \sim N(0, \sigma)$$

- Fixed effects for groups i ("within", since only variance within the same group i matters):

$$y_{ik} = \beta_0 + \beta_1 x_{ik} + \sum_i \gamma_i d_i + \epsilon_{ik} \text{ with } \epsilon_{ik} \sim N(0, \sigma)$$

- Random effects for groups i:

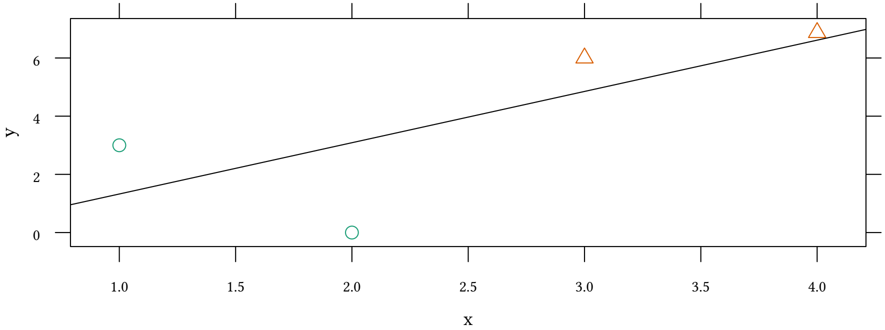
$$y_{ik} = \beta_0 + \beta_1 x_{ik} + \nu_i + \epsilon_{ik} \text{ with } \nu_i \sim N(0, \sigma_\nu) \text{ and } \epsilon_{ik} \sim N(0, \sigma)$$

Let us now try to use these models in R:

Pooled OLS:

$$y_{ik} = \beta_0 + \beta_1 x_{ik} + \epsilon_{ik} \text{ with } \epsilon_{ik} \sim N(0, \sigma)$$

```
ol <- lm(y ~ x,data=simple)
```



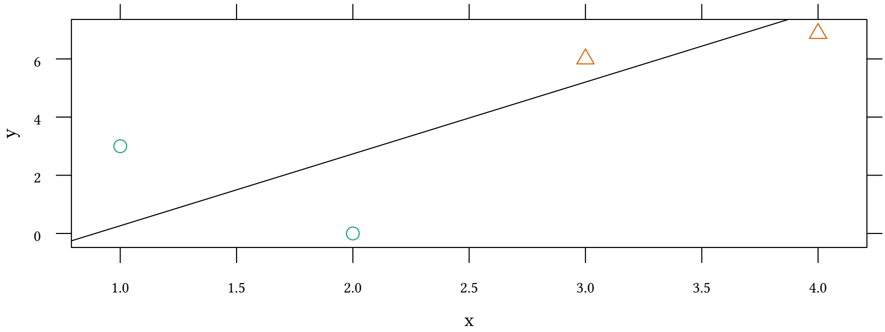
```
summary(ol)


Call:
lm(formula = y ~ x, data = simple)

Residuals:
      1       2       3       4 
 1.6749 -3.0874  1.1502  0.2623 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.4372     3.2089  -0.136    0.904
x             1.7623     1.1717   1.504    0.271

Residual standard error: 2.62 on 2 degrees of freedom
Multiple R-squared:  0.5308,Adjusted R-squared:  0.2961 
F-statistic: 2.262 on 1 and 2 DF,  p-value: 0.2715
```

Between OLS:

$$y_i = \beta_0 + \beta_1 x_i + \nu_i \text{ with } \nu_i \sim N(0, \sigma)$$

```
simple


  x      y i
1 1 3.0000 1
2 2 0.0000 1
3 3 6.0000 2
4 4 6.8744 2

betweenSimple <- aggregate( cbind(x,y) ~ i,
                           FUN=mean,data=simple)
betweenSimple
```

```
  i   x      y
1 1 1.5 1.5000
2 2 3.5 6.4372

betweenOLS <- lm (y ~ x,data=betweenSimple)
```



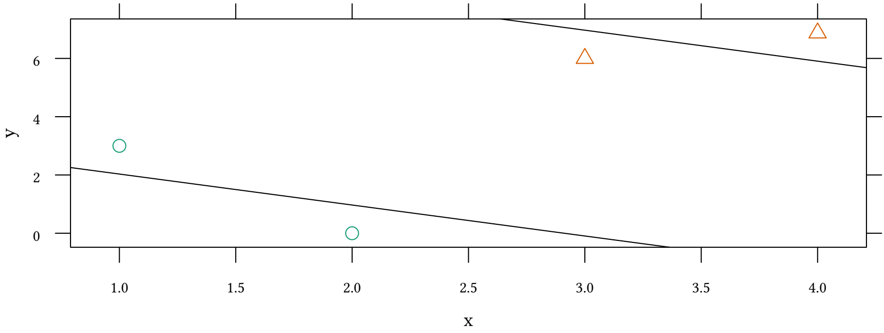
```
summary(betweenOLS)


Call:
lm(formula = y ~ x, data = betweenSimple)

Residuals:
ALL 2 residuals are 0: no residual degrees of freedom!

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.203        NaN     NaN      NaN
x              2.469        NaN     NaN      NaN

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1,Adjusted R-squared:    NaN
F-statistic:   NaN on 1 and 0 DF,  p-value: NA
```

Fixed effects for groups i**:**

$$y_{ik} = \beta_0 + \beta_1 x_{ik} + \sum_i \gamma_i d_i + \epsilon_{ik} \text{ with } \epsilon_{ik} \sim N(0, \sigma)$$

We also call the fixed effects model a “within” model, since only variance within the same group i matters.

```
fixef <- lm(y ~ x + as.factor(i),data=simple)
```



```
summary(fixef)


Call:
lm(formula = y ~ x + as.factor(i), data = simple)

Residuals:
      1       2       3       4 
 0.9686 -0.9686 -0.9686  0.9686 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.094      3.212   0.963    0.512
x               -1.063      1.937  -0.549    0.681
as.factor(i)2    7.063      4.332   1.630    0.350

Residual standard error: 1.937 on 1 degrees of freedom
Multiple R-squared:  0.8717,Adjusted R-squared:  0.6152 
F-statistic: 3.398 on 2 and 1 DF,  p-value: 0.3581
```

Random effects for groups i:

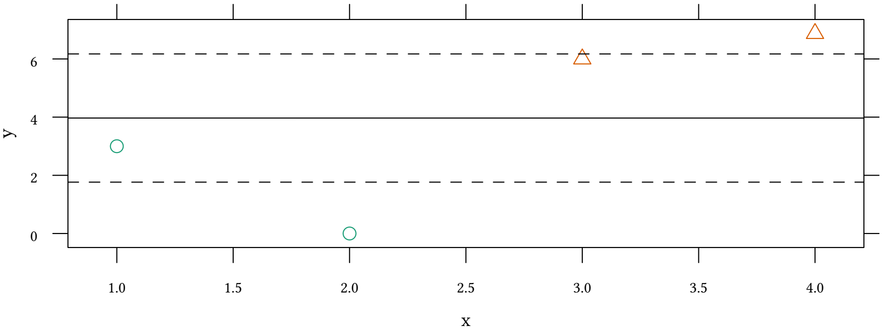
$$y_{ik} = \beta_0 + \beta_1 x_{ik} + \nu_i + \epsilon_{ik} \text{ with } \nu_i \sim N(0, \sigma_\nu) \text{ and } \epsilon_{ik} \sim N(0, \sigma)$$

```
library(lme4)
ranef <- lmer(y ~ x + (1|i),data=simple)
```

```
fixef(ranef)


  (Intercept)             x 
 3.9689030682 -0.0001212273 
```

```
ranef(ranef)

$i
  (Intercept)
1   -2.203052
2    2.203052

with conditional variances for "i"
```

```
coef(ranef)$i

  (Intercept)             x
1    1.765852 -0.0001212273
2    6.171955 -0.0001212273
```

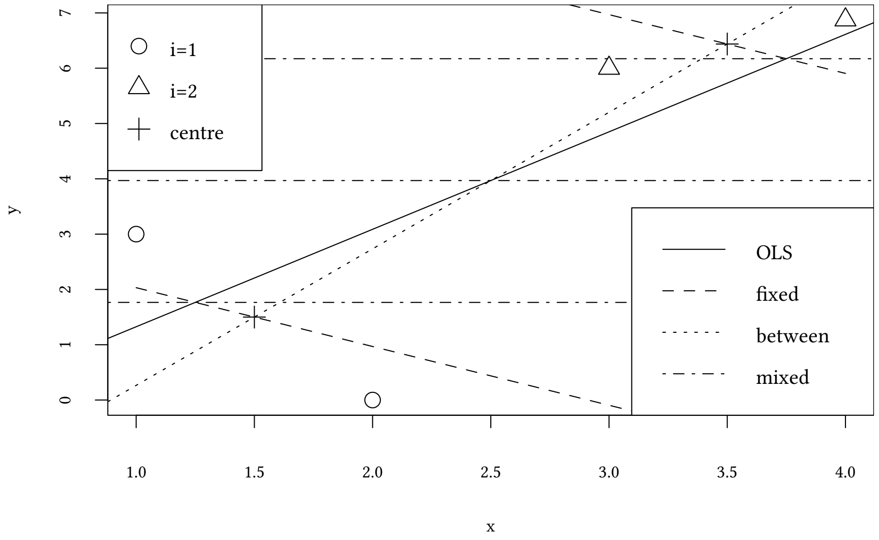


```
summary(ranef)

Linear mixed model fit by REML ['lmerMod']
Formula: y ~ x + (1 | i)
   Data: simple

REML criterion at convergence: 12.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-0.9114 -0.2943  0.1371  0.4314  0.6371

Random effects:
 Groups   Name        Variance Std.Dev.
 i        (Intercept) 15.560   3.945
 Residual              3.753   1.937
Number of obs: 4, groups:  i, 2

Fixed effects:
              Estimate Std. Error t value
```

```
(Intercept)  3.9689031  5.0114521   0.792
x           -0.0001212  1.6197091   0.000

Correlation of Fixed Effects:
  (Intr)
x -0.808
```

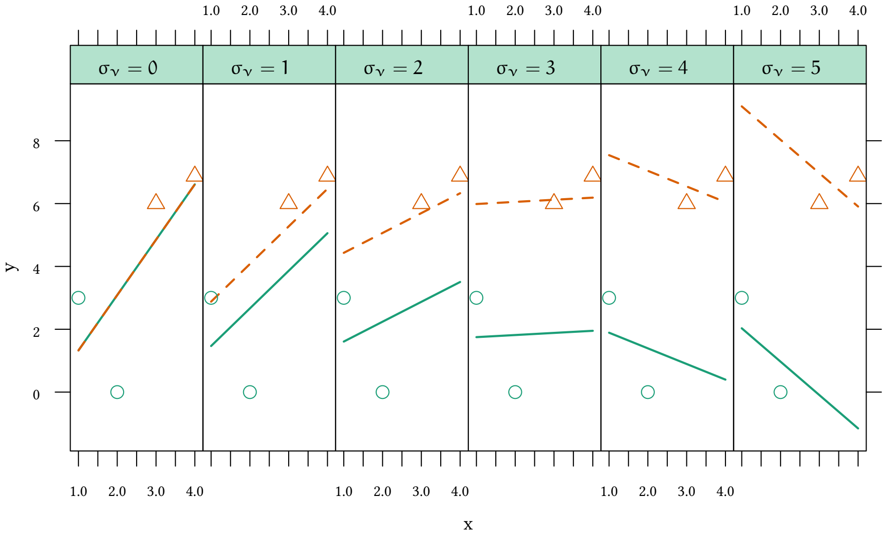
Here is a picture of the different estimated regression lines:

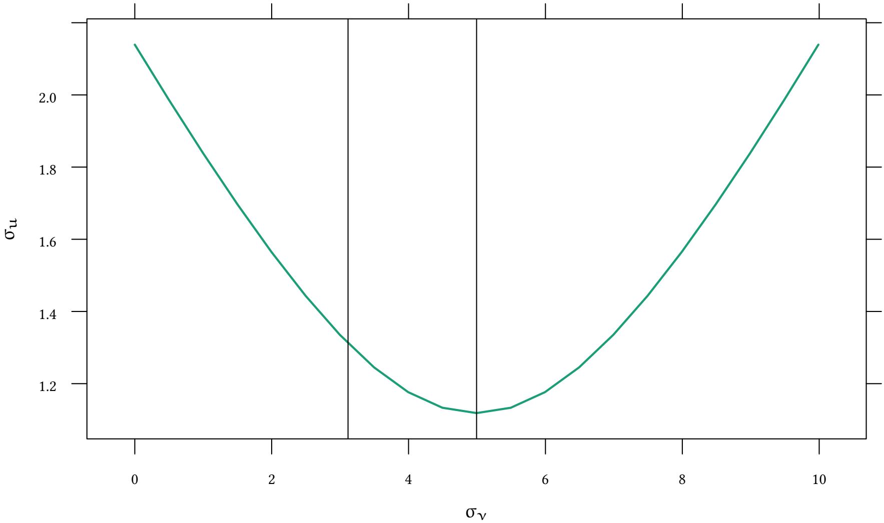
```
plot(y ~ x,data=simple,pch=i)
points(betweenSimple[,c("x","y")],pch=3)
abline(ol)
abline(betweenOLS,lty=3)
qq <- sapply(unique(simple$i),function(g)
             lines(predict(fixef,newdata=within(simple,i<-g)) ~ x,
                   data=simple,lty=2))
abline(fixef(ranef),lty=4)
qq<-apply(coef(ranef)$i,1,abline,lty=4)
legend("topleft",c("i=1","i=2","centre"),pch=1:3,bg="white",box.lty=1)
legend("bottomright",c("OLS","fixed","between","mixed"),bg="white",lty=1:4)
```



We see that, depending on the model, we can get anything between a positive and a negative relationship.

The following graph illustrates the transition from the pure OLS model (left) to the pure fixed effects model (right):





OLS allows no fixed effect ($\sigma_\nu = 0$), all the variance goes into u.

Fixed effects puts no restriction on σ_ν, hence σ_u can be minimal.

Random effects is in between.

$$\begin{array}{llcllll}
\text{between OLS} & y_i & = & \beta_0 + \beta_1 x_i & +\nu_i & \\
\text{pooled OLS} & y_{ik} & = & \beta_0 + \beta_1 x_{ik} & & +\epsilon_{ik} \\
\text{mixed effects} & y_{ik} & = & \beta_0 + \beta_1 x_{ik} & +\nu_i & +\epsilon_{ik} \\
\text{fixed effects} & y_{ik} & = & \beta_0 + \beta_1 x_{ik} & +\sum_i \gamma_i d_i & +\epsilon_{ik}
\end{array}$$

	ν_i	ϵ_{ik}
between OLS	finitely expensive	cheap (no cost)
pooled OLS	0 (infinitely expensive)	finitely expensive
mixed effects	finitely expensive	finitely expensive
fixed effects	cheap (no cost)	finitely expensive

between OLS

- neglects any variance within groups
- fits a line through the center of each group

pooled OLS

- neglects any group specific effect
- imposes an infinitely high cost on the fixed effect ν_i (setting them to 0) and, under this constraint, minimizes ϵ_{ikt}.

Pooled OLS yields an estimation between the *between OLS* and the *fixed effects* estimator.

fixed effect

- neglects all the variance across groups
- does not impose any cost on fixed effects

In the example, the relationship within the two groups is decreasing on average, hence a negative slope is estimated.

random effect

- balances ν_i and ϵ_{ikt}
- If coefficients were close to the fixed effects model, then ν_i are large (in absolute terms) and the ϵ_{ikt} are small.
- If the coefficients were close to the OLS model, then ν_i are very small but the ϵ_{ikt} are getting larger.
- mixed effects yields an estimation between the fixed effect and the (pooled) OLS estimation.

10.3 5 different methods - 5 different results

10.3.1 A larger example

Consider the following relationship:

$y_{it} = x_{it} + \nu_i + \epsilon_{it}$

with $\nu_i \sim N(0, \sigma_\nu)$ and $\epsilon_{ikt} \sim N(0, \sigma)$

We simulate and test now the following methods

- between OLS
- pooled OLS
- Fixed effects
- Mixed effects

```
set.seed(10)
I <- 6
T <- 50
i <- as.factor(rep(1:I,each=T))
ierr <- 15*rep(rnorm(I),each=T)
uerr <- 3*rnorm(I*T)
x <- runif(I*T)
y <- x + ierr + uerr
```

For comparison we will also construct a dependent variable *y2* without an individual specific random effect.

```
y2 <- x + 6*uerr
```

We put them all in one dataset.

```
data <- data.frame(cbind(y,y2,x,i,ierr,uerr))
```

10.3.2 Pooled OLS

$$y_{it} = \beta_0 + \beta_1 x_{it} + \epsilon_{it} \qquad \text{with } \epsilon_{ik} \sim N(0, \sigma)$$

```
ols <- lm (y ~ x,data=data)
summary(ols)


Call:
lm(formula = y ~ x, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-23.742  -4.895   2.283   7.343  16.896
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -3.826      1.092  -3.504  0.00053 ***
x              1.071      1.875   0.571  0.56828    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.515 on 298 degrees of freedom
Multiple R-squared:  0.001094,Adjusted R-squared:  -0.002258 
F-statistic: 0.3263 on 1 and 298 DF,  p-value: 0.5683
```

- Estimation of β is consistent if residuals ϵ_{it} are uncorrelated with $\mathbf{X}$.
- With repeated observations (as in our case), estimation of $\Sigma_{\beta\beta}$ is generally not consistent.

10.3.3 Between OLS

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \text{with } \nu_i \sim N(0, \sigma)$$

```
data.between <- aggregate(data,list(data$i),mean)
ols.between <- lm(y ~ x,data=data.between)
summary(ols.between)


Call:
lm(formula = y ~ x, data = data.between)

Residuals:
       1        2        3        4        5        6 
  3.3411   0.9423 -16.8802  -5.4127   8.0874   9.9221 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -4.150     68.155  -0.061    0.954
x              1.715    135.100   0.013    0.990

Residual standard error: 11.07 on 4 degrees of freedom
Multiple R-squared:  4.03e-05,Adjusted R-squared:  -0.2499 
F-statistic: 0.0001612 on 1 and 4 DF,  p-value: 0.9905
```

- Estimation of β is consistent if residuals (ν_i) are uncorrelated with $\mathbf{X}$.
- $\Sigma_{\beta\beta}$ can not be estimated.
- The method is inefficient since the variance within a group is not exploited.

10.3.4 Fixed effects

$$y_{it} = \beta_0 + \beta_1 x_{it} + \sum_i \gamma_i d_i + \epsilon_{it}$$

```
fixed <- lm(y ~ x + as.factor(i) - 1,data=data)
summary(fixed)


Call:
lm(formula = y ~ x + as.factor(i) - 1, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6167 -2.1856  0.0641  2.1945  7.1029

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
x               1.0626     0.5763   1.844   0.0662 .
as.factor(i)1  -0.4466     0.5210  -0.857   0.3920
as.factor(i)2  -2.8790     0.5033  -5.720 2.62e-08 ***
as.factor(i)3 -20.6976     0.5053 -40.962  < 2e-16 ***
as.factor(i)4  -9.2534     0.4936 -18.747  < 2e-16 ***
as.factor(i)5   4.2317     0.4864   8.700 2.44e-16 ***
as.factor(i)6   6.1136     0.5099  11.990  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.905 on 293 degrees of freedom
Multiple R-squared:  0.9182,Adjusted R-squared:  0.9163
F-statistic: 470.1 on 7 and 293 DF,  p-value: < 2.2e-16
```

- Estimation of β is consistent if residuals (ϵ_{it}) are uncorrelated with $\mathbf{X}$. This is a weaker requirement, since, with fixed effects, residuals are only ϵ_{ikt}, not ν_i.
- Estimation of $\Sigma_{\beta\beta}$ is consistent.
- The procedure loses some efficiency, since all the d_i are exactly estimated (although we are not interested in d_i).

10.3.5 Mixed effects

$$y_{it} = \beta_0 + \beta_1 x_{it} + \nu_i + \epsilon_{it}$$

```
mixed <- lmer(y ~ x + (1|i),data=data)
summary(mixed)

Linear mixed model fit by REML ['lmerMod']
Formula: y ~ x + (1 | i)
   Data: data
```

```
REML criterion at convergence: 1522.1

Scaled residuals: 
     Min       1Q   Median       3Q      Max 
-2.61559 -0.74732  0.02794  0.75531  2.44665 

Random effects:
 Groups   Name        Variance Std.Dev.
 i        (Intercept) 97.861   9.892   
 Residual              8.442   2.905   
Number of obs: 300, groups:  i, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  -3.8219     4.0525  -0.943
x             1.0626     0.5763   1.844

Correlation of Fixed Effects:
  (Intr)
x -0.072
```

- Estimation of β is consistent if ν_i and ϵ_{it} are uncorrelated with $\mathbf{X}$. This is a stronger requirement than with fixed effects.
- Estimation of $\Sigma_{\beta\beta}$ is consistent.
- The procedure is more efficient than fixed effects.

10.3.6 More general mixed effects

Random effect for only the intercept:

$$y_{it} = \beta_0 + \nu_i + \beta_1 x_{it} + \epsilon_{it}$$

```
lmer( y ~ x + (1|g) )
```

Random effect for intercept and slope:

$$y_{it} = \beta_0 + \nu_i + (\beta_1 + \nu_i')x_{it} + \epsilon_{it}$$

```
lmer( y ~ x + (1+x|g) )
```

10.3.7 Bayes and mixed effects

$$y_{it} = \beta_0 + \beta_1 x_{it} + \nu_i + \epsilon_{it} \quad \text{or} \quad y_{it} \sim N(\beta_0 + \beta_1 x_{it} + \nu_i, \tau_1)$$

```
modelM <- 'model {
 for (i in 1:length(y)) {
        y[i] ~ dnorm(beta[1]+beta[2]*x[i]+nu[group[i]],tau[1])
 }
 for (j in 1:max(group)) {
    nu[j] ~ dnorm(0,tau[2])
 }
 for (k in 1:2) {
    beta[k]    ~ dnorm (0,.0001)
    tau[k]     ~ dgamma(.01,.01)
    sd[k]     <- sqrt(1/tau[k])
 }
}'
dataList<-with(data,list(y=y-mean(y),x=x-mean(x),group=as.numeric(data$i)))
bayesM<-run.jags(model=modelM,data=dataList,inits=genInit(4),
                 monitor=c("beta","sd"))
```

```
bayesM


JAGS model summary statistics from 40000 samples (chains = 4; adapt+burnin = 5000):

          Lower95   Median Upper95    Mean       SD    Mode      MCerr
beta[1]   -8.0354 -0.17929  11.116 0.40936   4.7418 -1.0715    0.76541
beta[2] -0.044229   1.0613   2.211  1.0635 0.57794  1.0382  0.0029004
sd[1]      2.6825   2.9086  3.1524  2.9129 0.12007  2.8978 0.00061253
sd[2]       5.192   10.509  20.364  11.578  4.6721   9.297   0.083549

        MC%ofSD SSeff       AC.10   psrf
beta[1]    16.1    38     0.98046 1.0864
beta[2]     0.5 39704 0.000028797 1.0001
sd[1]       0.5 38425 -0.00096291      1
sd[2]       1.8  3127     0.12393 1.0217

Total time taken: 1.3 seconds
```

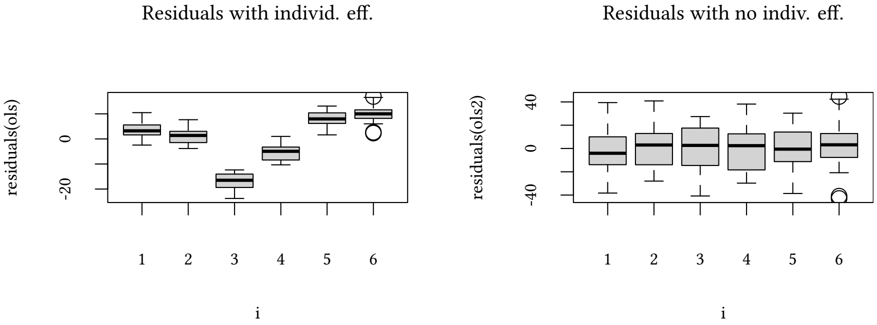
10.4 Estimation results

10.4.1 Residuals

OLS residuals In the above exercise we actually *knew* the correct relationship. How can we discover the need for a fixed- or mixed effects model from our data?

Let us have a look at the residuals of the OLS estimation:

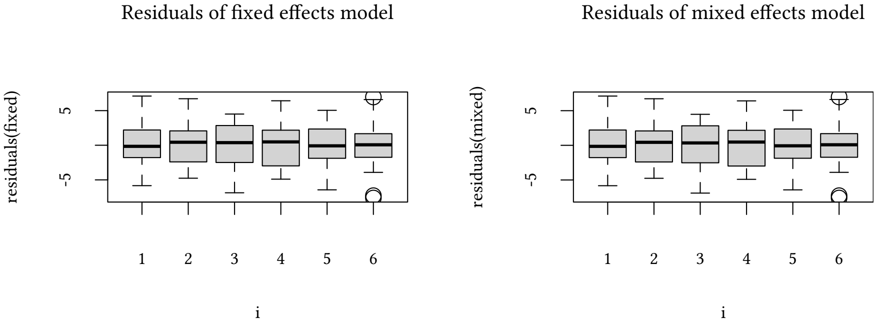
```
ols2 <- lm (y2 ~ x,data=data)
with(data,boxplot(residuals(ols) ~ i,main="Residuals with individ. eff."))
with(data,boxplot(residuals(ols2) ~ i,main="Residuals with no indiv. eff."))
```



The left graph shows the residuals for the model where we do have individual specific effects, the right graph shows residuals for the $y2$ model without such effects.

Fixed- and mixed effects residuals

```
with(data,boxplot (residuals(fixed) ~ i,main="Residuals of fixed effects model"))
with(data,boxplot (residuals(mixed) ~ i,main="Residuals of mixed effects model"))
```



10.4.2 Estimated standard errors

$$y_{it} = \beta_0 + \beta_1 x_{it} + \nu_i + \epsilon_{it}$$

Let us compare the estimated standard errors of the residuals ϵ_{ikt}

```
summary(ols)$sigma

[1] 9.515232

summary(fixed)$sigma

[1] 2.905489

sigma(mixed)

[1] 2.905489
```

Here the estimated standard errors of the mixed and fixed effects model are similar. This need not be the case.

10.4.3 Testing random effects

- Do we really have a random effect? How can we test this?

Idea: Likelihood ratio test (the following procedure works (overconservatively) for testing random effects, but does not work very well if we want to test fixed effects).

generally

$$2 \cdot (\log(L_{\text{large}}) - \log(L_{\text{small}})) \sim \chi^2_k \qquad \text{with } k = df_{\text{large}} - df_{\text{small}}$$

here

$$2 \cdot (\log(L_{RE}) - \log(L_{OLS})) \sim \chi^2_k \qquad \text{with } k = 1$$

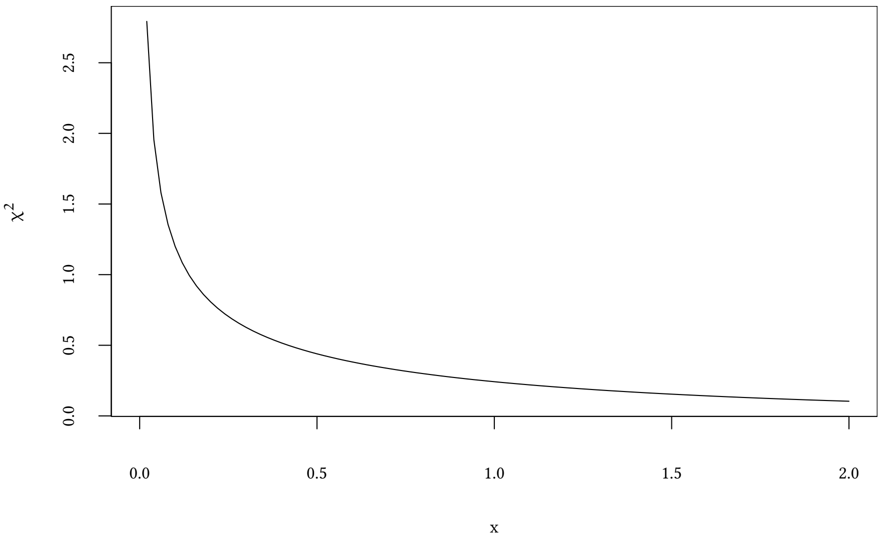
Note that we have to set *REML=FALSE* in the mixed model if we want to compare the likelihood of the mixed with a fixed effects model.

```
mixedML <- update(mixed,REML=FALSE)
teststat <- 2*(logLik(mixedML)-logLik(ols))[1]
```

`teststat` $\overset{\text{approx.}}{\sim}$ χ^2 unless we are at the boundary of the parameter space.

Testing $\sigma^2_\nu = 0$ we are at the boundary. Nevertheless...

```
plot(function(x) dchisq(x,1),0,2,ylab="$\\chi^2$")
```



The p-value of the χ^2-test would be

```
teststat

[1] 673.6798

pchisq(teststat,1,lower=FALSE)

[1] 1.582576e-148
```

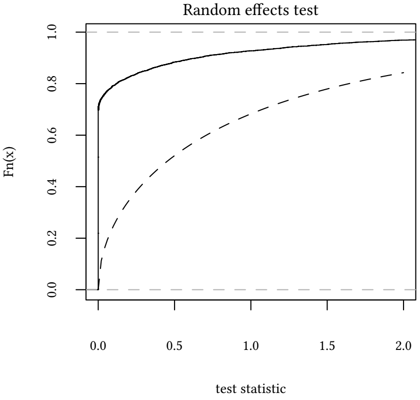
Let us bootstrap the distribution of *teststat* under the Null that there is no mixed effect:

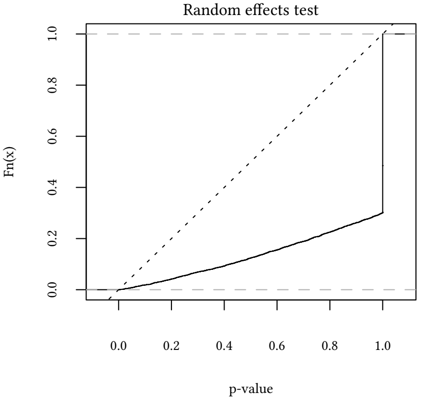
```
set.seed(125)
dev <- replicate(5000,{
  by <- unlist(simulate(ols))
  bols <- lm (by ~ x,data=data)
  bmixed <- refit(mixedML,by)
  LL <- 2*(logLik(bmixed)-logLik(bols))[1]
  c(LL=LL,pchisq=pchisq(LL,1,lower=FALSE))
})
```

The bootstrapped distribution differs from the χ^2 distribution:

```
plot(ecdf(dev["LL",]),do.points=FALSE,verticals=TRUE,xlim=c(0,2),
     xlab="test statistic",main="Random effects test")
plot(function(x) pchisq(x,df=1,lower=TRUE),xlim=c(0,2),lty=2,add=TRUE)
#
```

```
plot(ecdf(dev["pchisq",]),do.points=FALSE,verticals=TRUE,
     xlab="p-value",main="Random effects test")
abline(a=0,b=1,lty=3)
```





- The assumption of a χ^2 distribution is rather conservative.
- If we manage to reject H_0 (that $\sigma_\nu^2 = 0$) based on the χ^2 distribution, then we can definitely reject it based on the bootstrapped distribution.
- We might actually accept H_0 too often.

  $\rightarrow$ If we find a test statistic which is still acceptable according to the χ^2 distribution (pooled OLS is ok), chances are that we could reject this statistic with the bootstrapped distribution.

We can, of course, use the bootstrapped value of the test statistic and compare it with the value from our test:

```
mean(teststat < dev["LL",])

[1] 0
```

Note that we need many bootstrap replications to get reliable estimates for p-values.

10.4.4 Confidence intervals for fixed effects (in a ME model)

The distribution of estimated coefficients β does not follow a t distribution with a number of degrees of freedom one can determine easily.

To determine confidence intervals for estimated coefficients we have to bootstrap a sample of coefficients.

```
summary(mixed)

Linear mixed model fit by REML ['lmerMod']
Formula: y ~ x + (1 | i)
   Data: data

REML criterion at convergence: 1522.1

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.61559 -0.74732  0.02794  0.75531  2.44665

Random effects:
 Groups   Name        Variance Std.Dev.
 i        (Intercept) 97.861   9.892
 Residual              8.442   2.905
Number of obs: 300, groups:  i, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  -3.8219     4.0525  -0.943
x             1.0626     0.5763   1.844

Correlation of Fixed Effects:
  (Intr)
x -0.072
```

```
set.seed(123)
require(boot)
```

```
mixed.boot <- bootMer(mixed,function(.) fixef(.),nsim=1000)
```

The *t* component of the *boot* object contains all the estimated fixed effects. We can also calculate manually *mean* and *sd*:

```
apply(mixed.boot$t,2,function(x) c(mean=mean(x),sd=sd(x)))

     (Intercept)         x
mean   -3.765866 1.0643228
sd      3.839421 0.5950676
```

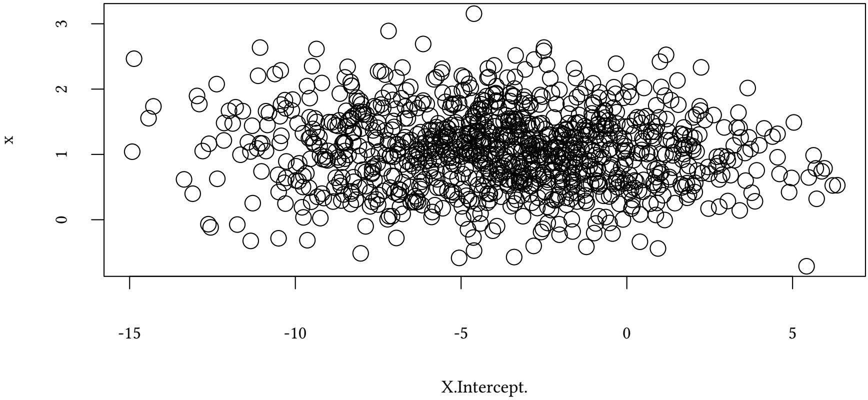
$$\left(\text{bias} = \bar{\theta}_{BS} - \hat{\theta}\right)$$

```
boot.ci(mixed.boot,index=2,type=c("norm","basic","perc"))

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
```

```
boot.ci(boot.out = mixed.boot, type = c("norm", "basic", "perc"),
    index = 2)

Intervals :
Level      Normal              Basic              Percentile
95%   (-0.105,  2.227 )   (-0.149,  2.224 )   (-0.098,  2.274 )
Calculations and Intervals on Original Scale
```

Here is a plot of the different bootstrapped estimates:

```
plot(data.frame(mixed.boot))
```



The credible interval from the Bayesian estimate (with flat priors) are (as expected) very similar to the confidence interval from the bootstrap:

```
bayesM$summary$quantiles[2,]

       2.5%         25%         50%         75%       97.5% 
-0.07416123  0.67500175  1.06134000  1.45387750  2.18903050 
```

10.5 Literature

Literature

- Jose C. Pinheiro and Douglas M. Bates. “Mixed Effects Models in S and S-Plus”. Springer, 2002.
- Julian J. Faraway. “Extending the Linear Model with R”. Chapman & Hall, 2006.

- John K. Kruschke. “Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan”. Academic Press, 2nd Edition, 2014.
- Stock, Watson. “Introduction to Econometrics”. Chapter 10.
- William H. Greene. “Econometric Analysis“. Chapter 11.

Appendix 10.A Examples for the lecture

Example 1 You estimate the following mixed effects model

$$y_{it} = \beta_1 + \beta_2 x_{it} + \nu_i + \nu'_j + \epsilon_{it}$$

where $i \in \{1, \ldots, I\}$ denotes the individual, $j \in \{1, \ldots, J\}$ denotes a group, and $t \in \{1, \ldots, T\}$ denotes time. ϵ_{it} is the residual, ν_i is the individual specific random effect, ν'_j is the group specific random effect. The vector `j[i]` tells you to which group individual `i` belongs. You use the following model in JAGS (α is a placeholder):

```
for(k in 1:length(y)){
    y[k] ~ dnorm( beta[1] + beta[2]*x[k] + α, tau[1] )
}
for(k in 1:max(i)) { nu[k] ~ dnorm(0,tau[2]) }
for(k in 1:max(j)) { nu2[k] ~ dnorm(0,tau[3]) }
for(k in 1:2) { beta[k] ~ dnorm(0,.0001) }
for(k in 1:3) { tau[k] ~ dgamma(.01,.01) }
```

Here *y* and *x* are two $I \times T$ long vectors denoting the dependent variable y and the independent x. *i* is a vector with length $I \times T$ denoting to which individual i each observation k belongs. *j* is a vector of length I denoting to which group j each individual i belongs.

What should be the value of α ?

- `nu[k] + nu2[k]`
- `nu[i[k]] + nu2[j[i[k]]]`
- `nu[j[i[k]]] + nu2[j[i[k]]]`
- `nu[i[k]]] + nu2[i[j[k]]]`
- other value

Example 2 You estimated a mixed effects model with *lmer* and you have stored the result in the variable *mer*. You have also estimated the same model, though without the random effects, with *lm* and stored the result in the variable *ols*. To find out whether the mixed effects model is justified, you run the following approximate permuatation test:

```
lls <- replicate(500,{
    y1 <- simulate(ols)[[1]]
    l.ols <- logLik(lm (y1 ~ x))[1]
    l.mer <- logLik(refit(mer,y1))[1]
    2*(l.mer-l.ols)
})
mean(lls < 2*(logLik(mer)-logLik(ols))[1])
```

The last statement returns the value 0.98. Your Null hypothesis is that there is no random effect. What is the p-value?

- .02
- .04
- .96
- .98
- other value

Example 3 You estimate the following mixed effects model

$$y_{i,t} = \beta_1 + \beta_2 x_{i,t} + \nu_i + \epsilon_{i,t}$$

where $i \in \{1, \ldots, I\}$ denotes the individual, and $t \in \{1, \ldots, T\}$ denotes time. $\epsilon_{i,t}$ is the residual, ν_i is the individual specific random effect. You use the following model in JAGS (α is a placeholder):

```
for(k in 1:length(y)){ α } for(j in 1:max(i)){nu[j]~dnorm(0,tau[2])} for(k in
1:2){beta[k]~dnorm(0,.0001);tau[k]~dgamma(.01,.01)}
```

Here *y* and *x* are two $I \times T$ long vectors denoting the dependent variable y and the independent x. *i* is a $I \times T$ long vector which denotes the individual i.

What should be the value of α ?

- `y[k] ~ dnorm(beta[1]+beta[2]*x[k]+nu[k],tau[1])`
- `y[k] ~ dnorm(beta[1]+beta[2]*x[k]+nu[i[k]],tau[1])`
- `y[k] ~ dnorm(beta[1]+beta[2]*x[k]+i[nu[k]],tau[1])`
- `y[k] ~ dnorm(beta[1]+beta[2]*x[k]+i[k],tau[1])`
- other value

Example 4 You use *lmer* from the *lme4* library to estimate a mixed effects model *m.mer*. You also use *lm* to estimate a simple OLS model. A comparison of the two models with the help of *anova* yields the following result:

```
         Df  AIC  BIC  logLik    χ^2  Df  Pr(> χ^2)
 m.ols    3  319  327    -156
 m.mer    4  317  327    -154  4.21   1     0.0401
```

Assume a level of significance of 5%. Which of the following statements is true?

- The random effect is significant.
- The method used by *anova* here is often too conservative. The true p-value may well be smaller than 0.0401.
- The method used by *anova* here is often anti-conservative. The true p-value may well be larger than 0.0401.
- It is possible to obtain a better estimate for the p-value of this test with the help of a bootstrap.
- *anova* does a Likelihood ratio test here.

Appendix 10.B Exercises

To solve the next exercises the following commands for R might help: *read.table, str, table, list, as.factor, aggregate, summary, mean, residuals, fitted, sigma, logLik, lm, ranef, anova, boxplot*. Furthermore *lrtest* from *lmtest* and *lmer* from *lme4*. For *lme4* you might want to use the *REML* options.

Exercise 10.1 *The file ex1.csv contains observations on x1, x2, y and a group variable group. You are interested in how x1 and x2 influence y. Estimate the following models, compare their coefficients and standard errors:*

- *Pooled OLS*
- *Between OLS*
- *Fixed Effects*

Exercise 10.2 *Have another look at the data from ex1.csv. Now also estimate a model with a random effect for groups.*

Exercise 10.3 *What can you say about the distribution of residuals of your estimates for ex1.csv?*

Exercise 10.4 *The file ex2.csv contains observations on x1, x2, y and a group variable group. You are interested in how x1 and x2 influence y.*

- *In the fixed effects model: Is the group specific effect significant?*
- *In the mixed effects model: Is the group specific effect significant?*

11 Instrumental variables

11.1 Instruments

Remember from OLS assumptions: We require

$$\text{strict exogeneity:} \qquad E(u_i|X_i = x) = 0 \quad \to \quad X \vdash u$$

The problem:

$$Y = X\beta + u \qquad \text{but } X \nvdash u$$

Solution (Wright, 1928), use instrument Z (we "instrument" X with Z):

$$\begin{array}{lll} Z \text{ is relevant:} & \text{cor}(Z, X) \neq 0 & (Z \nvdash X) \\ Z \text{ is exogeneous:} & \text{cor}(Z, u) = 0 & (Z \vdash u) \end{array}$$

$$\begin{tikzcd} u \arrow[d] \arrow[dr] & Z \arrow[d] \\ Y & X \arrow[l] \end{tikzcd}$$

$$\begin{array}{rll} \text{1st stage: } & X = Z\gamma + \nu & (\text{works, since } Z \nvdash X) \\ \to & \hat{X} = Z\gamma & \\ \text{2nd stage: } & Y = \hat{X}\beta + u & (\text{works, since } Z \vdash u, \text{ and, hence } \hat{X} \vdash u) \end{array}$$

Here we construct X as correlated with u.

- Z is relevant, $\text{cor}(Z, X) \neq 0$.
- Z is exogeneous, $\text{cor}(Z, u) = 0$.

$$\begin{tikzcd} u \arrow[d] \arrow[dr] & Z \arrow[d] \\ Y & X \arrow[l] \end{tikzcd}$$

```
set.seed(123)
N <- 100
u <- rnorm(N)
Z <- rnorm(N)              # Z is exogeneous
X <- Z + .5*rnorm(N) - u # X is endogenous, Z is relevant
Y <- X + u
```

When we estimate

$$Y = \beta_0 + \beta_1 X + u$$

we should expect $\hat{\beta}_1 \approx 1$.

Naïve model: ignore that $X \nvdash u$

```
summary(lm( Y ~ X ))
```

```
Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.59416 -0.42295 -0.00768  0.45972  1.88043 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.03257    0.06706   0.486    0.628    
X            0.58001    0.04503  12.881   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6678 on 98 degrees of freedom
Multiple R-squared:  0.6287,Adjusted R-squared:  0.6249 
F-statistic: 165.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

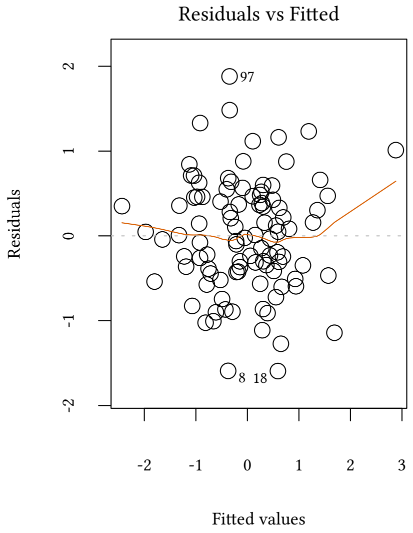
Note that we have – without an economic model – no chance to find only from X and Y that exogeneity is violated.

Estimated residuals look innocent:

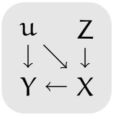
```
cor(X, residuals(lm(Y~X)))
```

```
[1] 1.383238e-17
```

```
plot(lm(Y~X),which=1,sub.caption="")
```



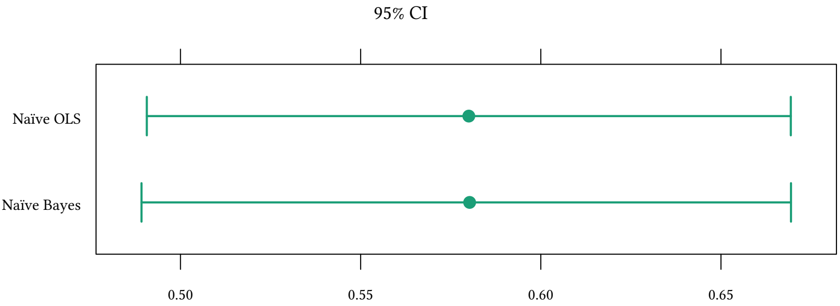
We can do the same exercise with Bayes:



```
instNaive.model <- 'model {
  for(i in 1:length(y)) {
    y[i] ~ dnorm( beta[1] + beta[2] * x[i], tau)
    }
    beta[1] ~ dnorm(0,.0001)
    beta[2] ~ dnorm(0,.0001)
    tau     ~ dgamma(0.01,.01)
}'
ini <- genInit(4)
instNaive.jags <- run.jags(model=instNaive.model,data=list(y=Y,x=X,z=Z),
                    monitor=c("beta","tau"),inits=ini)
```

```
instNaive.jags


JAGS model summary statistics from 40000 samples (chains = 4; adapt+burnin = 5000):

         Lower95    Median Upper95     Mean       SD Mode      MCerr
beta[1] -0.10248 0.032187 0.16471 0.032067 0.068043   -- 0.00034075
beta[2]  0.48975  0.58024 0.66977  0.57992 0.045798   -- 0.00023016
tau       1.6405   2.2258  2.8898   2.2405  0.32111   --  0.0016154
```

```
        MC%ofSD SSeff      AC.10    psrf
beta[1]     0.5 39874   0.013368       1
beta[2]     0.5 39594 -0.0023739  1.0001
tau         0.5 39514  0.0010761 0.99996

Total time taken: 0.5 seconds
```



Instrument $Z \vdash u$

Problem:

$$X \text{ is endogeneous:} \quad \mathrm{cor}(X,u) \neq 0 \qquad (X \nvdash u)$$

$$\begin{tikzcd} u \arrow[d] \arrow[dr] & Z \arrow[d] \\ Y & X \arrow[l] \end{tikzcd}$$

To solve the problem, we use Z as an instrument:

$$\begin{aligned} \text{1st stage: } X &= Z\gamma + \nu && \text{(works, since } Z \nvdash X) \\ \hat{X} &= Z\gamma \\ \text{2nd stage: } Y &= \hat{X}\beta + u && \text{(works, since } Z \vdash u\text{, and, hence } \hat{X} \vdash u) \end{aligned}$$

Remember:

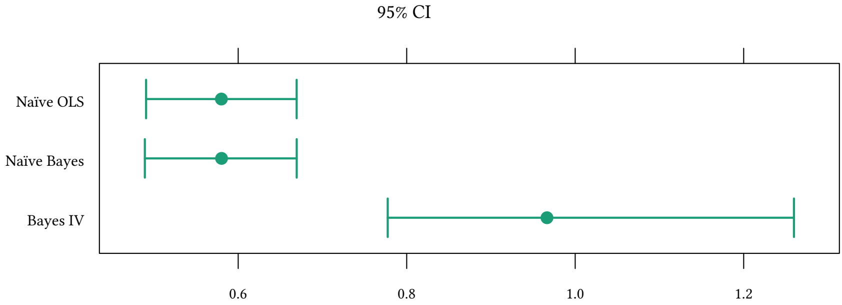
$$\begin{aligned} &Z \text{ is relevant:} && \mathrm{cor}(Z,X) \neq 0 && (Z \nvdash X) \\ &Z \text{ is exogeneous:} && \mathrm{cor}(Z,u) = 0 && (Z \vdash u) \end{aligned}$$

We first use the Gibbs sampler to estimate this model:

```
inst.model <- 'model {
  for(i in 1:length(y)) {
    x[i]    ~  dnorm( xHat[i], tau[2])                    # 1st stage
    xHat[i]<- gamma[1] + gamma[2] * z[i]
    y[i]    ~  dnorm(beta[1] + beta[2] * xHat[i], tau[1]) # 2nd stage
  }
  for(k in 1:2) {
    beta[k] ~ dnorm(0,.0001)
    gamma[k]~ dnorm(0,.0001)
    tau[k]  ~ dgamma(0.01,.01)
    sd[k]   <- sqrt(1/tau[k])
    }
}'
```

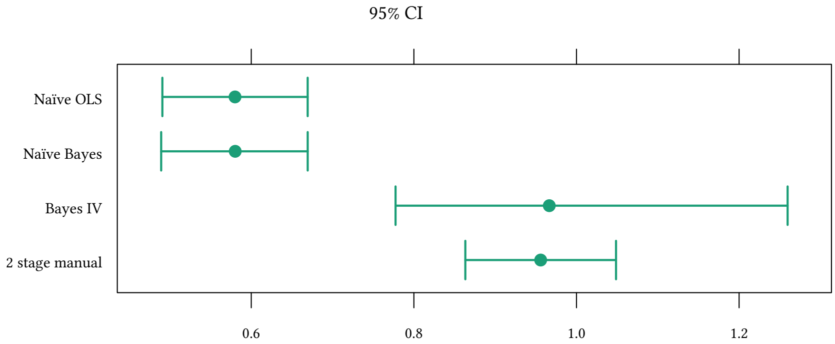
```
inst.jags <- run.jags(model=inst.model,data=list(y=Y,x=X,z=Z),
                    monitor=c("beta","sd"),inits=ini)
```

```
inst.jags


JAGS model summary statistics from 40000 samples (chains = 4; adapt+burnin = 5000):

          Lower95   Median Upper95     Mean       SD Mode      MCerr
beta[1] -0.14196 0.085526 0.33267 0.087691  0.12105   --  0.0020677
beta[2]  0.75038  0.96653  1.2207  0.97997  0.12285   --  0.0021715
sd[1]     0.4155  0.47913  0.5508  0.48124 0.034862   -- 0.00017456
sd[2]    0.94344     1.09  1.2505    1.095 0.079045   -- 0.00041293

        MC%ofSD SSeff      AC.10    psrf
beta[1]     1.7  3427    0.17659  1.0009
beta[2]     1.8  3201    0.20065  1.0011
sd[1]       0.5 39884 -0.0005609       1
sd[2]       0.5 36644  0.0092112 0.99999

Total time taken: 1.4 seconds
```



```
est1 <- lm(X ~ Z)
est2 <- lm(Y ~ predict(est1))
summary(est2)
```

```
Call:
lm(formula = Y ~ predict(est1))

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9328 -0.3228 -0.0424  0.3128  1.0910 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    0.08434    0.04814   1.752   0.0829 .  
predict(est1)  0.95596    0.04671  20.468   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4771 on 98 degrees of freedom
Multiple R-squared:  0.8104,Adjusted R-squared:  0.8085 
F-statistic: 418.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

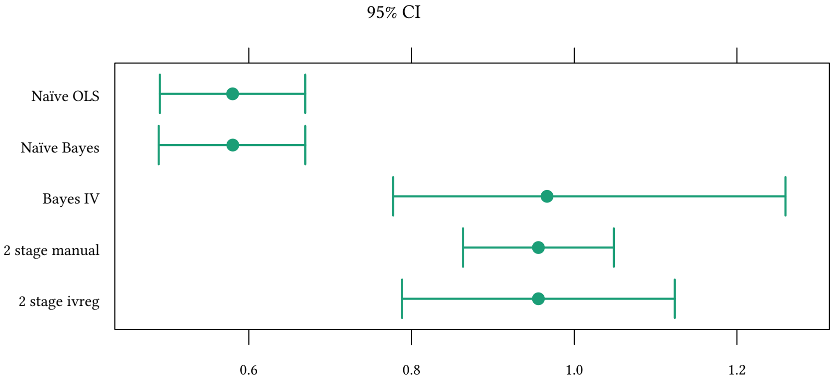


Of course, a similar result can also be obtained with a specialised frequentist tool:

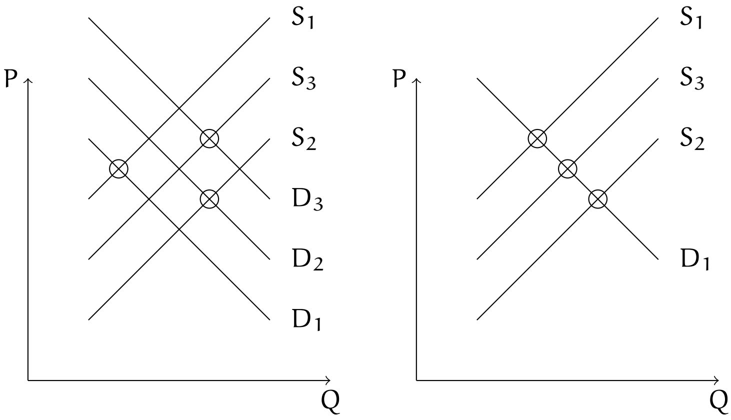
```
library(AER)
```

```
est.2sls <- ivreg( Y ~ X | Z)
summary(est.2sls)


Call:
ivreg(formula = Y ~ X | Z)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.26775 -0.57731 -0.05521  0.53982  2.07423 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  0.08434    0.08814   0.957    0.341
X            0.95596    0.08551  11.180   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8735 on 98 degrees of freedom
Multiple R-Squared: 0.3645,Adjusted R-squared: 0.3581
Wald test:   125 on 1 and 98 DF,  p-value: < 2.2e-16
```



11.2 Demand and Supply



How can we estimate the slope of D, if D and S are moving simultaneously?

The Demand for Cigarettes

The dataset

- `state:` Factor indicating state.
- `year:` Factor indicating year.
- `cpi:` Consumer price index.
- `population:` State population.
- `packs:` Number of packs per capita.
- `income:` State personal income (total, nominal).
- `tax:` Average state, federal and average local excise taxes for fiscal year.
- `taxs:` Average excise taxes for fiscal year, *including sales tax*.
- `price:` Average price during fiscal year, including sales tax.

```
data("CigarettesSW", package = "AER")
head(CigarettesSW,n=4)

  state year   cpi population    packs    income  tax    price
1    AL 1985 1.076    3973000 116.4863  46014968 32.5 102.1817
2    AR 1985 1.076    2327000 128.5346  26210736 37.0 101.4750
3    AZ 1985 1.076    3184000 104.5226  43956936 31.0 108.5788
4    CA 1985 1.076   26444000 100.3630 447102816 26.0 107.8373
      taxs
1 33.34834
2 37.00000
3 36.17042
4 32.10400

table(CigarettesSW$year)


1985 1995
  48   48
```

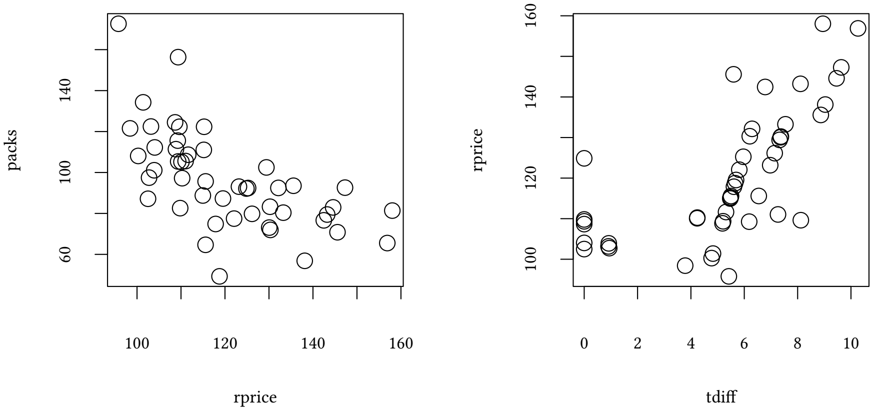
We have to construct some variables:
Clean the data:

```
Cig <- within(subset(CigarettesSW,year=="1995"),{
    rprice <- price/cpi
    rincome <- income/population/cpi
    tdiff <- (taxs - tax)/cpi
    rtax  <- tax/cpi
})
```

$$\begin{aligned}\texttt{tdiff} &= Z\\ \log(\texttt{rprice}) &= X\\ \log(\texttt{packs}) &= Y\end{aligned}$$

$$\begin{tikzcd}
u \arrow[dr] \arrow[dd] & \texttt{tdiff} = Z \arrow[d] \\
 & \log(\texttt{rprice}) = X \arrow[dl] \\
\log(\texttt{packs}) = Y &
\end{tikzcd}$$

```r
with(Cig,{plot(packs ~ rprice); plot(tdiff, rprice)})
```

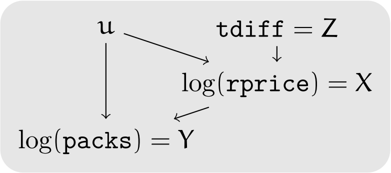


We are interested in

$$\log(\texttt{packs}) = \beta_0 + \beta_1 \log(\texttt{rprice}) + u$$

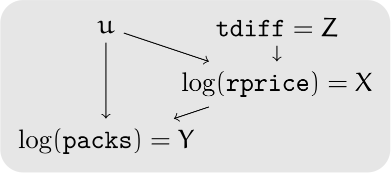
However, `rprice` is endogeneous, correlated with u.
We can use `tdiff`, the sales tax on cigarettes, as an instrument for $\log(\texttt{rprice})$.

$$\begin{aligned}\text{1st stage:}\;\; \log(\texttt{rprice}) &= \gamma_0 + \gamma_1 \texttt{tdiff} + \nu\\ \log\widehat{(\texttt{rprice})} &= \gamma_0 + \gamma_1 \texttt{tdiff}\\ \text{2nd stage:}\;\; \log(\texttt{packs}) &= \beta_0 + \beta_1 \log\widehat{(\texttt{rprice})} + u\end{aligned}$$



The naïve approach

```
est0.lm <- lm(log(packs) ~ log(rprice), data = Cig)
summary(est0.lm)


Call:
lm(formula = log(packs) ~ log(rprice), data = Cig)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.64676 -0.09030  0.01787  0.11245  0.40779 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  10.3389     1.0353   9.986 4.25e-13 ***
log(rprice)  -1.2131     0.2164  -5.604 1.13e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1896 on 46 degrees of freedom
Multiple R-squared:  0.4058,Adjusted R-squared:  0.3928 
F-statistic: 31.41 on 1 and 46 DF,  p-value: 0.00000113
```



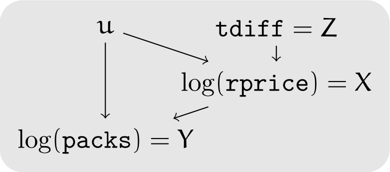
The estimated residuals do not tell us anything about the problem:

```
cor(residuals(est0.lm),
    log(Cig$rprice))

[1] -1.835831e-16
```

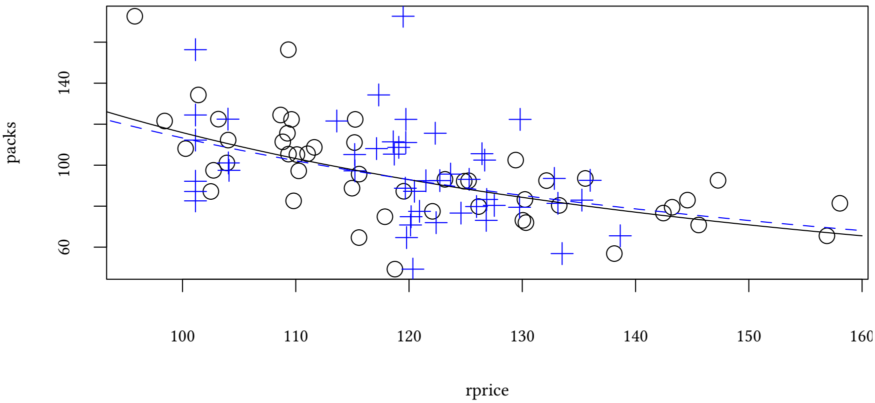
2 Stage Least Squares (2SLS)

The 1st stage:

```
est.ts1<-lm(log(rprice) ~ tdiff, data=Cig)
summary(est.ts1)


Call:
lm(formula = log(rprice) ~ tdiff, data = Cig)

Residuals:
      Min        1Q    Median        3Q       Max
-0.221027 -0.044324  0.000111  0.063730  0.210717

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)
(Intercept) 4.616546   0.029108   158.6      < 2e-16 ***
tdiff       0.030729   0.004802     6.4 0.0000000727 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09394 on 46 degrees of freedom
Multiple R-squared:  0.471,Adjusted R-squared:  0.4595
F-statistic: 40.96 on 1 and 46 DF,  p-value: 0.00000007271
```



```
with(Cig,{plot(packs ~ rprice);
          points(exp(predict(est.ts1)), packs, pch=3, col="blue")})
```



The 2nd stage:

```
est.ts2<-lm(log(packs) ~ predict(est.ts1), data=Cig)
summary(est.ts2)


Call:
lm(formula = log(packs) ~ predict(est.ts1), data = Cig)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.63180 -0.15802  0.00524  0.13574  0.61434 

Coefficients:
                 Estimate Std. Error t value  Pr(>|t|)    
(Intercept)        9.7199     1.8012   5.396 0.0000023 ***
predict(est.ts1)  -1.0836     0.3766  -2.877   0.00607 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2264 on 46 degrees of freedom
Multiple R-squared:  0.1525,Adjusted R-squared:  0.1341 
F-statistic: 8.277 on 1 and 46 DF,  p-value: 0.006069
```

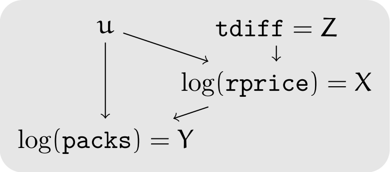
$$\begin{tikzcd}
u \arrow[dr] \arrow[dd] & \mathtt{tdiff} = Z \arrow[d] \\
 & \log(\mathtt{rprice}) = X \arrow[dl] \\
\log(\mathtt{packs}) = Y &
\end{tikzcd}$$

Note that the standard errors with this (manual) two stage procedure are not correct.

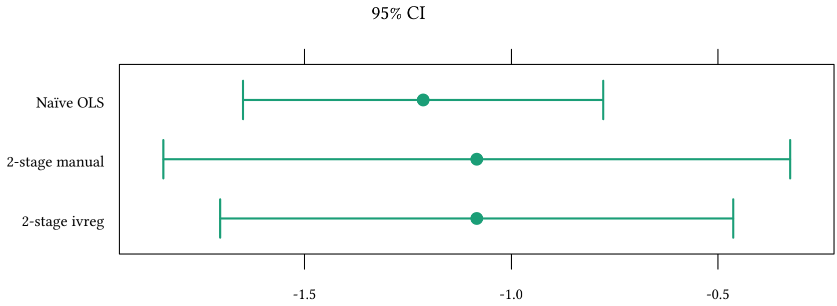
The regression in the second stage can not take into account the standard errors of the first stage.

The *ivreg* command:

```
est1.iv <- ivreg(log(packs) ~ log(rprice) | tdiff, data = Cig)
summary(est1.iv)


Call:
ivreg(formula = log(packs) ~ log(rprice) | tdiff, data = Cig)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.64619 -0.07732  0.02981  0.11283  0.41904 

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)    
(Intercept)   9.7199     1.5141   6.420 0.0000000679 ***
log(rprice)  -1.0836     0.3166  -3.422      0.00131 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1904 on 46 degrees of freedom
Multiple R-Squared: 0.4011,Adjusted R-squared: 0.3881 
Wald test: 11.71 on 1 and 46 DF,  p-value: 0.001313
```

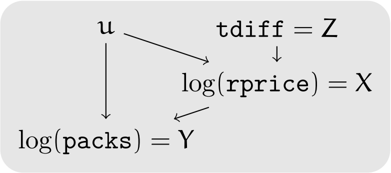


Comparison of estimated price elasticities:



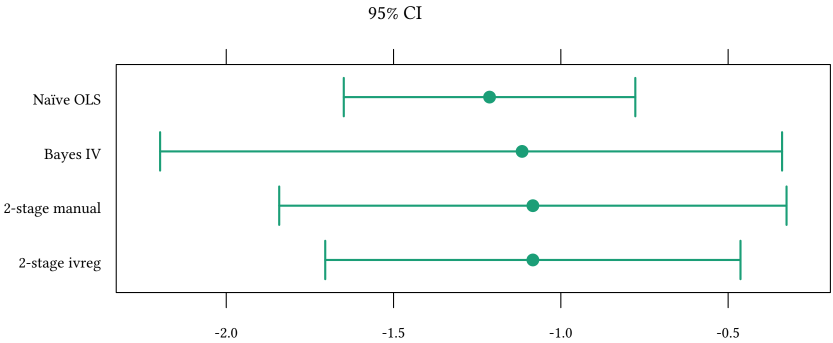
We use the same model as before:

```
inst.model <- 'model {
  for(i in 1:length(y)) {
    x[i]    ~  dnorm( xHat[i], tau[2])                  # 1st stage
    xHat[i]<- gamma[1] + gamma[2] * z[i]
    y[i]    ~  dnorm(beta[1] + beta[2] * xHat[i], tau[1]) # 2nd stage
  }
  for(k in 1:2) {
    beta[k] ~ dnorm(0,.0001)
    gamma[k]~ dnorm(0,.0001)
    tau[k]  ~ dgamma(0.01,.01)
    sd[k]   <- sqrt(1/tau[k])
    }
}'
```



```
cig.data <- with(Cig,list(y=log(packs)-mean(log(packs)),
                         x=log(rprice)-mean(log(rprice)),
                         z=tdiff-mean(tdiff)))
cig.jags <- run.jags(model=inst.model,
                    data=cig.data,
                    monitor=c("beta","sd"),
                    inits=ini)
```

```
cig.jags


JAGS model summary statistics from 40000 samples (chains = 4; adapt+burnin = 5000):

          Lower95     Median  Upper95       Mean       SD  Mode
beta[1] -0.074083 0.00047877 0.075199 0.00033328 0.037873    --
beta[2]   -2.1209    -1.1158 -0.29165    -1.1538  0.47205    --
sd[1]     0.18351    0.22911  0.27851    0.23112 0.024658    --
sd[2]    0.078461   0.096956  0.11886   0.097944 0.010522    --

              MCerr MC%ofSD SSeff      AC.10   psrf
beta[1]  0.00023624     0.6 25700 0.00035278 1.0003
beta[2]   0.0030556     0.6 23866  0.0058074 1.0001
sd[1]    0.00012855     0.5 36795 -0.0016919      1
sd[2]   0.000055055     0.5 36525 -0.0016729      1

Total time taken: 0.7 seconds
```

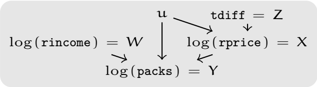


Could there be an ommitted variable bias in our second stage equation. Perhaps demand is not only affected by price, but also by income?

→ include $\log(\texttt{rincome})$ in the second stage:

$$\begin{tikzcd}
 & & u \arrow[drr] \arrow[dd] & & \texttt{tdiff}=Z \arrow[d] \\
\log(\texttt{rincome})=W \arrow[drr] & & & & \log(\texttt{rprice})=X \arrow[dll] \\
 & & \log(\texttt{packs})=Y & &
\end{tikzcd}$$

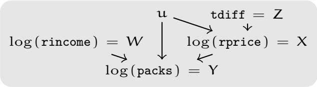
```
est2.iv <- ivreg(log(packs) ~ log(rprice) + log(rincome) | log(rincome) + tdiff,
                 data = Cig)
summary(est2.iv)


Call:
ivreg(formula = log(packs) ~ log(rprice) + log(rincome) | log(rincome) +
    tdiff, data = Cig)

Residuals:
      Min        1Q    Median        3Q       Max
-0.611000 -0.086072  0.009423  0.106912  0.393159

Coefficients:
             Estimate Std. Error t value     Pr(>|t|)
(Intercept)    9.4307     1.3584   6.943 0.0000000124 ***
log(rprice)   -1.1434     0.3595  -3.181      0.00266 **
log(rincome)   0.2145     0.2686   0.799      0.42867
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1896 on 45 degrees of freedom
Multiple R-Squared: 0.4189,Adjusted R-squared: 0.3931
Wald test: 6.534 on 2 and 45 DF,  p-value: 0.003227
```



Note that we include here *log(rincome)* as an instrument for itself. Technically this means, that *log(rincome)* will be perfectly predicted, i.e. not instrumented.

```
lm(log(rincome) ~ I(log(rincome)) + tdiff,data=Cig)


Call:
lm(formula = log(rincome) ~ I(log(rincome)) + tdiff, data = Cig)

Coefficients:
    (Intercept)  I(log(rincome))            tdiff
      5.128e-16        1.000e+00        3.043e-19
```

(we have to use *I(...)* if we insist on including the response on the right hand side)

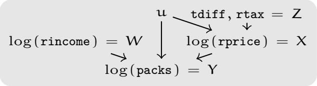


More instruments

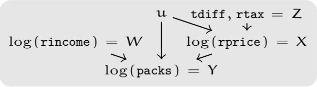
Above we used only *tdiff* as an instrument. Perhaps we can do better.

```
est.ts1b <- lm(log(rprice)~tdiff + rtax,data=Cig)
summary(est.ts1b)


Call:
lm(formula = log(rprice) ~ tdiff + rtax, data = Cig)

Residuals:
      Min        1Q    Median        3Q       Max
-0.062214 -0.019435 -0.002044  0.021219  0.096169

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.3687412  0.0179294  243.66  < 2e-16 ***
tdiff       0.0101778  0.0021291    4.78 0.000019 ***
rtax        0.0101627  0.0005904   17.21  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03449 on 45 degrees of freedom
Multiple R-squared:  0.9302,Adjusted R-squared:  0.9271
F-statistic: 300.1 on 2 and 45 DF,  p-value: < 2.2e-16
```

Indeed, compared to the equation where we used only *tdiff*, the R^2 increased from 0.471 to 0.93.



```
est3.iv <- ivreg(log(packs) ~ log(rprice) + log(rincome) | log(rincome) + tdiff + rtax,
                 data = Cig)
summary(est3.iv)


Call:
ivreg(formula = log(packs) ~ log(rprice) + log(rincome) | log(rincome) +
    tdiff + rtax, data = Cig)

Residuals:
       Min         1Q     Median         3Q        Max
-0.6006931 -0.0862222 -0.0009999  0.1164699  0.3734227

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.8950     1.0586   9.348 4.12e-12 ***
log(rprice)   -1.2774     0.2632  -4.853 1.50e-05 ***
log(rincome)   0.2804     0.2386   1.175    0.246
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1879 on 45 degrees of freedom
Multiple R-Squared: 0.4294,Adjusted R-squared: 0.4041
Wald test: 13.28 on 2 and 45 DF,  p-value: 0.00002931
```



```
anova(est3.iv,est1.iv)

Analysis of Variance Table

Model 1: log(packs) ~ log(rprice) + log(rincome) | log(rincome) + tdiff +
    rtax
Model 2: log(packs) ~ log(rprice) | tdiff
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     45 1.5880
2     46 1.6668 -1 -0.078748 1.3815  0.246
```

11.3 The general Instrumental Variables Model

$$
\begin{aligned}
Y = \beta_0 + \underbrace{\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k}_{\text{endogeneous}} + \\
+ \underbrace{\beta_{k+1} W_1 + \ldots + \beta_{k+r} W_r}_{\text{exogenous}} + u \\
\forall i \in 1, \ldots k: \quad X_i = \gamma_{0i} + \underbrace{\gamma_{1i} Z_1 + \gamma_{2i} Z_2 + \ldots + \gamma_{mi} Z_m}_{\text{instruments}} + \nu_i
\end{aligned}
$$

$$\begin{tikzcd}
 & u \arrow[dd] \arrow[dr] & & Z_1, \ldots, Z_m \arrow[dl] \\
W_1, \ldots, W_r \arrow[dr] & & X_1, \ldots, X_k \arrow[dl] & \\
 & Y & &
\end{tikzcd}$$

- $X_1, \ldots, X_k$: k endogenous variables (perhaps correlated with u)
- $W_1, \ldots, W_r$: r exogenous variables (not correlated with u)
- $Z_1, \ldots, Z_m$: m exogenous instruments (not correlated with u)

Requirement: $m \geq k$

- $m < k$: underidentified model, we can not distinguish the different X_j
- $m = k$: exactly identified model
- $m > k$: overidentified model (even better)

11.4 An example for (potential) underidentification

We first generate some data:

$$\begin{tikzcd}
u \arrow[d] \arrow[dr] & Z_1 \arrow[d] & Z_2 \arrow[dl] \\
Y & X, X^2 \arrow[l] &
\end{tikzcd}$$

```r
set.seed(123)
N <- 1000
u <- rnorm(N)
Z1 <- rnorm(N)
Z2 <- rnorm(N)
X <- -5*u + Z1 + Z2 + rnorm(N)
Y <- X - .1*X^2 + u
```

The following would be too naïve:

```
summary(lm( Y ~ X + I(X^2) ))


Call:
lm(formula = Y ~ X + I(X^2))

Residuals:
     Min       1Q   Median       3Q      Max 
-1.01208 -0.22084 -0.00221  0.23599  0.89939 

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)    
(Intercept)  0.0049042  0.0132344    0.371    0.711    
X            0.8204981  0.0020501  400.229   <2e-16 ***
I(X^2)      -0.1000327  0.0002928 -341.666   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3358 on 997 degrees of freedom
Multiple R-squared:  0.9967,Adjusted R-squared:  0.9967 
F-statistic: 1.506e+05 on 2 and 997 DF,  p-value: < 2.2e-16
```

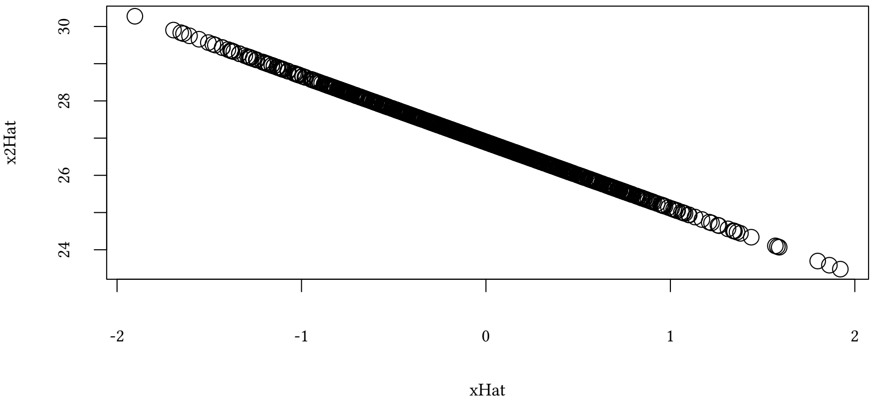
But using a single Z_1 as an instrument does not work. Underidentified IV:

$$\begin{tikzcd}
u \arrow[d] \arrow[dr] & Z_1 \arrow[d] \\
Y & X, X^2 \arrow[l]
\end{tikzcd}$$

```
ivreg(Y ~ X + I(X^2) | Z1)

Warning in ivreg.fit(X, Y, Z, weights, offset, ...): more regressors than
instruments


Call:
ivreg(formula = Y ~ X + I(X^2) | Z1)

Coefficients:
(Intercept)            X       I(X^2)  
     -2.664        1.321           NA  
```

Why is this underidentified?

```
xHat  <- predict(lm(X   ~ Z1))
x2Hat <- predict(lm(X^2 ~ Z1))
plot(xHat,x2Hat)
```



“Solution” 1: No instrument for X^2 (this is clearly not a perfect idea):

```
summary(ivreg(Y ~ X + I(X^2) | I(X^2) + Z1))


Call:
ivreg(formula = Y ~ X + I(X^2) | I(X^2) + Z1)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.1118 -1.0691  0.1049  1.2330  4.7750 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.076312   0.072341  -1.055    0.292    
X            1.149723   0.093705  12.270   <2e-16 ***
I(X^2)      -0.096204   0.001864 -51.606   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.741 on 997 degrees of freedom
Multiple R-Squared: 0.9114,Adjusted R-squared: 0.9112 
Wald test:  2699 on 2 and 997 DF,  p-value: < 2.2e-16
```

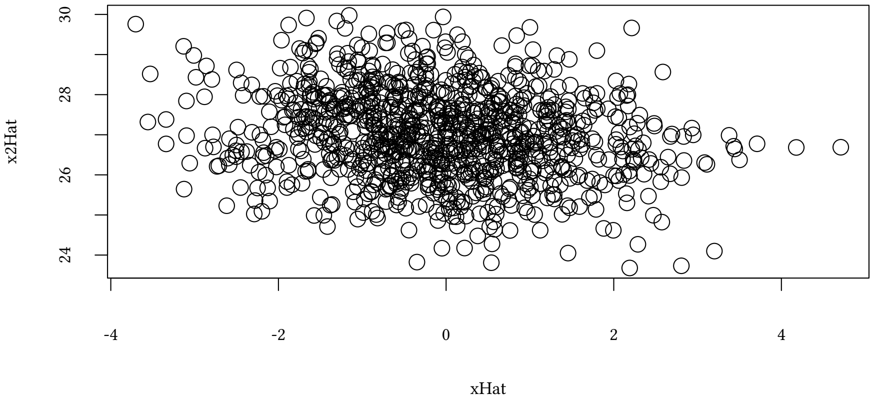
$$\begin{tikzcd}
X^2 \arrow[dr] & u \arrow[d] \arrow[dr] & Z_1 \arrow[d] \\
 & Y & X \arrow[l]
\end{tikzcd}$$

Solution 2: Linear instruments

```
summary(ivreg(Y ~ X + I(X^2) | Z1 + Z2))


Call:
ivreg(formula = Y ~ X + I(X^2) | Z1 + Z2)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9241 -1.9393 -1.1785  0.7575 25.1905 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.13493    2.42351   0.881   0.3786    
X            1.00362    0.07782  12.896   <2e-16 ***
I(X^2)      -0.17843    0.08968  -1.990   0.0469 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.103 on 997 degrees of freedom
Multiple R-Squared: 0.7183,Adjusted R-squared: 0.7177 
Wald test: 93.25 on 2 and 997 DF,  p-value: < 2.2e-16 
```

$$\begin{tikzcd}
u \arrow[d] \arrow[dr] & Z_1, Z_2 \arrow[d] \\
Y & X, X^2 \arrow[l]
\end{tikzcd}$$

```
xHat  <- predict(lm(X   ~ Z1+Z2))
x2Hat <- predict(lm(X^2 ~ Z1+Z2))
plot(xHat,x2Hat)
```

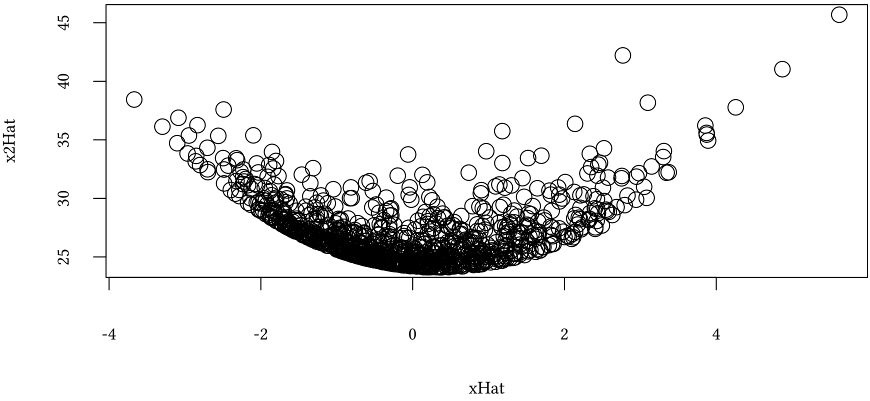


Solution 3: Quadratic instruments

```
summary(ivreg(Y ~ X + I(X^2) | Z1 + Z2 + I(Z1^2) + I(Z2^2)))


Call:
ivreg(formula = Y ~ X + I(X^2) | Z1 + Z2 + I(Z1^2) + I(Z2^2))

Residuals:
    Min      1Q  Median      3Q     Max 
-2.0272 -0.8524 -0.4031  0.4599  8.7182 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.52363    0.41328   1.267    0.205    
X            1.01473    0.03209  31.619  < 2e-16 ***
I(X^2)      -0.11875    0.01522  -7.801 1.54e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.309 on 997 degrees of freedom
Multiple R-Squared: 0.9499,Adjusted R-squared: 0.9498 
Wald test: 531.1 on 2 and 997 DF,  p-value: < 2.2e-16
```

$$\begin{tikzcd}
u \arrow[d] \arrow[dr] & Z_1, Z_1^2, Z_2, Z_2^2 \arrow[d] \\
Y & X, X^2 \arrow[l]
\end{tikzcd}$$

```
xHat  <- predict(lm(X   ~ Z1+Z2+ I(Z1^2) + I(Z2^2)))
x2Hat <- predict(lm(X^2 ~ Z1+Z2+ I(Z1^2) + I(Z2^2)))
plot(xHat,x2Hat)
```

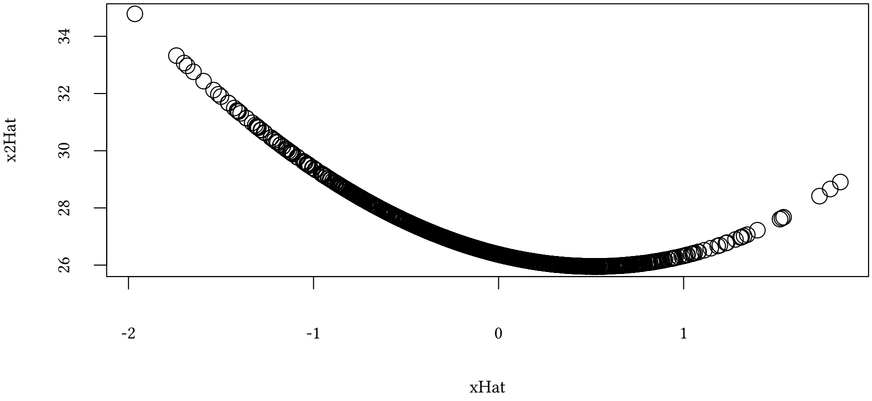


Solution 4: Quadratic instruments, based on one Z

```
summary(ivreg(Y ~ X + I(X^2) | Z1 + I(Z1^2)))


Call:
ivreg(formula = Y ~ X + I(X^2) | Z1 + I(Z1^2))

Residuals:
    Min      1Q  Median      3Q     Max 
-5.8396 -1.0147  0.2301  1.3320  3.6346 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.19983    2.07678  -0.096    0.923    
X            1.15788    0.16785   6.898 9.34e-12 ***
I(X^2)      -0.09161    0.07720  -1.187    0.236    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.79 on 997 degrees of freedom
Multiple R-Squared: 0.9063,Adjusted R-squared: 0.9061 
Wald test: 98.89 on 2 and 997 DF,  p-value: < 2.2e-16
```

$$\begin{tikzcd}
u \arrow[d] \arrow[dr] & Z_1, Z_1^2 \arrow[d] \\
Y & X, X^2 \arrow[l]
\end{tikzcd}$$

```
xHat  <- predict(lm(X   ~ Z1+I(Z1^2)))
x2Hat <- predict(lm(X^2 ~ Z1+I(Z1^2)))
plot(xHat,x2Hat)
```



11.5 Weak instruments

$$\begin{tikzcd} u \arrow[d] \arrow[dr] & Z \arrow[d] \\ Y & X \arrow[l] \end{tikzcd}$$

$$\begin{aligned} \text{1st stage: } X &= Z\gamma + \nu \\ \hat{X} &= Z\gamma \\ \text{2nd stage: } Y &= \hat{X}\beta + u \end{aligned}$$

- Weak instruments = instruments that explain not much of the variation in X (when there is only one endogeneous X:
  Z is weak if the F-statistic of the first stage equation is < 10)

⟶ 2SLS estimator is no longer reliable.

⟶ find better instruments, or use Bayesian inference.

Are the instruments in our example weak?

```
summary(lm(X ~ Z1+I(Z1^2)))$fstatistic

     value      numdf      dendf
  6.725538   2.000000 997.000000
```

```
summary(lm(X^2 ~ Z1+I(Z1^2)))$fstatistic

      value        numdf        dendf
  0.6406691    2.0000000  997.0000000

summary(lm(X ~ Z1+I(Z1^2)+Z2+I(Z2^2)))$fstatistic

  value    numdf    dendf
 16.318    4.000  995.000

summary(lm(X^2 ~ Z1+I(Z1^2)+Z2+I(Z2^2)))$fstatistic

     value      numdf      dendf
  1.395981   4.000000 995.000000
```

Apparently we do not have a good instrument for X^2.

11.6 Discrete endogeneous variables

Remember:

$$
\begin{aligned}
\text{1st stage: } X &= Z\gamma + \nu && (\text{works, since } Z \nvdash X) \\
\hat{X} &= Z\gamma && \\
\text{2nd stage: } Y &= \hat{X}\beta + u && (\text{works, since } Z \vdash u\text{, and, hence } \hat{X} \vdash u)
\end{aligned}
$$

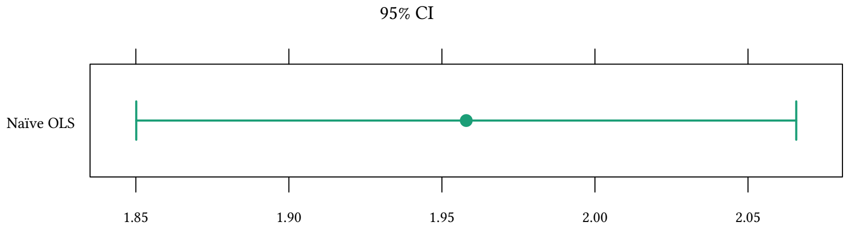
Now consider the case where X is not linear in Z:

```
set.seed(123)
N   <- 1000
u <- rnorm(N)
nu  <- rnorm(N)
Z   <- rnorm(N)
X   <- as.numeric((Z + u + nu )>0)
Y   <- X + u
```

$$\begin{tikzcd}
u \arrow[d] \arrow[dr] & Z \arrow[d] \\
Y & X \arrow[l]
\end{tikzcd}$$

```
summary(lm( Y ~ X ))


Call:
lm(formula = Y ~ X)

Residuals:
     Min        1Q    Median        3Q       Max
```

```
-2.45873 -0.59440  0.00077  0.56316  2.74402

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.46093    0.03877  -11.89   <2e-16 ***
X            1.95795    0.05494   35.64   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8687 on 998 degrees of freedom
Multiple R-squared:   0.56,Adjusted R-squared:   0.5595
F-statistic:  1270 on 1 and 998 DF,  p-value: < 2.2e-16
```

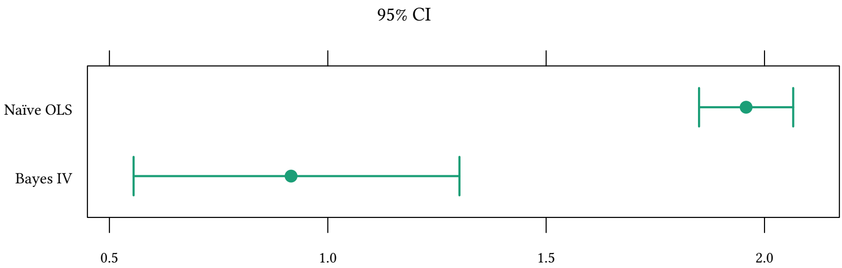


Bayesian inference

```
discrete.model <- 'model {
  for(i in 1:length(y)) {
    x[i]    ~  dbern(xHat[i])                 # 1st stage
    xHat[i]<- pnorm(gamma[1]+gamma[2]*z[i],0,1)
    y[i]    ~  dnorm(beta[1]+beta[2]*xHat[i],tau) # 2nd stage
  }
  for (k in 1:2) {
    beta[k]~dnorm(0,.0001)
    gamma[k]~dnorm(0,.0001)
  }
  tau~dgamma(0.01,.010)
}'
disc.jags <- run.jags(model=discrete.model,
                      data=list(y=Y,x=X,z=Z),
                      modules="glm",monitor=c("beta"),inits=ini)
```

```
disc.jags


JAGS model summary statistics from 40000 samples (chains = 4; adapt+burnin = 5000):

        Lower95   Median Upper95     Mean      SD Mode     MCerr
beta[1] -0.1438 0.056849 0.26483 0.055345 0.10425   -- 0.0005668
```

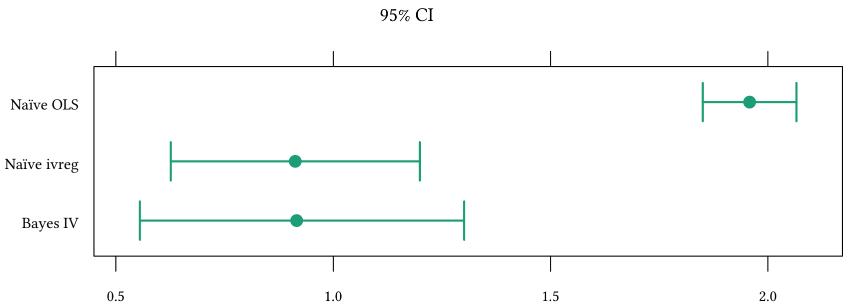
```
beta[2] 0.55419  0.91595  1.3002  0.91947 0.19118   -- 0.0010297

        MC%ofSD SSeff      AC.10   psrf
beta[1]     0.5 33831  0.0012897 1.0001
beta[2]     0.5 34474 0.00081462 1.0002

Total time taken: 40.4 seconds
```



Without Bayes we could also use the 2SLS:
Caution! We estimate a non-linear process with a linear model.

```
summary(ivreg(Y ~ X |Z))


Call:
ivreg(formula = Y ~ X | Z)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.86937 -0.66214 -0.02271  0.66499  3.26873 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.0596     0.0795   0.750    0.454    
X             0.9127     0.1461   6.248 6.15e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.014 on 998 degrees of freedom
Multiple R-Squared: 0.4004,Adjusted R-squared: 0.3998 
Wald test: 39.04 on 1 and 998 DF,  p-value: 6.15e-10 
```

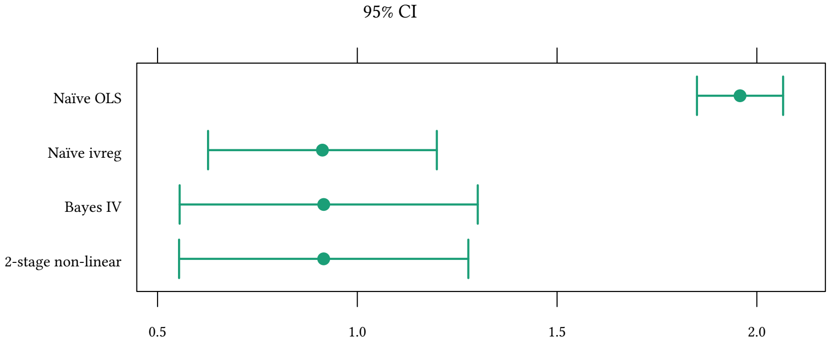
(see Chesher and Rosen, 2015, for a discussion)



We could also use two stage non-linear estimation:

(biased standard errors)

```
step1 <- glm(X ~ Z, family=binomial(link=logit))
Xhat  <- plogis(predict(step1))
summary(lm (Y ~ Xhat))


Call:
lm(formula = Y ~ Xhat)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.3240 -0.9231 -0.0019  0.9277  3.7916 

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)    
(Intercept)  0.05804    0.10057   0.577      0.564    
Xhat         0.91584    0.18448   4.964 0.00000081 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.294 on 998 degrees of freedom
Multiple R-squared:  0.0241,Adjusted R-squared:  0.02312 
F-statistic: 24.65 on 1 and 998 DF,  p-value: 0.00000081
```



11.7 Literature

Literature

- John K. Kruschke. “Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan”. Academic Press, 2nd Edition, 2014.
- Stock, Watson. “Introduction to Econometrics”. Chapter 12.
- William H. Greene. “Econometric Analysis“. Chapter 8.

Appendix 11.A Examples for the lecture

Example 1 The variables *x1*, *x2*, and *x3* are independent and from a continuous distribution. Which of the following instrumental variable models are exactly identified?

- `ivreg(y~x1|x1)`
- `ivreg(y~x1|x2)`
- `ivreg(y~x1+x3|x2)`
- `ivreg(y~x1+x3|x2+I(x2^2))`
- `ivreg(y~x1|x2+x3)`

Example 2 You want to estimate an equivalent to the following instrumental variables model:

```
ivreg(y~x1|x2+x3)
```

Consider the following model for JAGS (α and β are a placeholders):

```
for(i in 1:length(y)) {
  x1[i] ~ dnorm( x1H[i], tau[2])
  x1H[i] <- α
  y[i] ~ dnorm(β, tau[1])
}
for(k in 1:3) {
  b[k] ~ dnorm(0,.0001)
  c[k] ~ dnorm(0,.0001)
  tau[k] ~ dgamma(0.01,.01)
}
```

What is the value of β?

- `b[1]+b[2]*x1H[i]`
- `b[1]+b[2]*x1[i]`
- `x1[i]`
- `x1H[i]`
- other value

What is the value of α?

- `c[1]+c[2]*x2[i]+c[3]*x3[i]`
- `c[1]+c[2]*x[2]+c[3]*x[3]`
- `x2+x3`
- `c[1]*x2+c[2]*x3`
- other value

Example 3 You want to estimate the equation $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + u$. You fear, however, that X and u may be correlated. You decide to use *Z1* as an instrument. You estimate the following regression:

ivreg(Y ~ X + I(X^2) | I(X^2) + Z1))

Which statements are true?

- *Z1* can only be good instrument for X if *Z1* is not correlated with u.
- *Z1* is only a good instrument if it is not correlated with *X*.
- The above approach is underidentified, since only one instrument *Z1* is used as an instrument for both *X* and *X^2*.

- The above approach is exactly identified, since two instruments (`Z1` and `X^2`) are used as instruments for two variables `X` and `X^2`.
- If `Z1` is independent of u, the above approach does provide a good instrument for `X` and for `X^2`.

Example 4 Consider the following model for JAGS:

```
  for(i in 1:length(y)) {    x[i] ~ dnorm( xH[i], tau[2])    xH[i] <- g[1] + g[2]
* z[i]    y[i] ~ dnorm(b[1] + b[2] * xH[i], tau[1]) } for(k in 1:2) {    b[k] ~
dnorm(0,.0001)    g[k] ~ dnorm(0,.0001)    tau[k] ~ dgamma(0.01,.01) }
```

What is the equivalent of this model with *`ivreg`*?

- `ivreg(y ~ x + I(x^2) | I(x^2) + z))`
- `ivreg(y ~ x | z))`
- `ivreg(y ~ x + xH | z))`
- `ivreg(y ~ x | g1 + g2))`
- other value

Appendix 11.B Exercises

To solve the next exercises the following commands for R might help: *`data`*, *`lm`*, *`summary`*, *`ivreg`*, *`runjags`*, *`mean`*, *`coeftest`*, *`confint`*, *`I`*, *`residuals`*.

Exercise 11.1 *Use the `Mroz` dataset from `Ecdat` to explain `hearnw`, the hourly wage of married women, as a function of `educw`, their education. Use the log-linear specification to model the relation. Use the subsample for whom the value of `work` is equal to `"no"`.*

1. *What is the estimated effect of an extra year of education on the hourly wage?*

   You suspect that education could be correlated with the error term in the OLS regression and would like to instrument it with `educwf`, father's education.

2. *What are the necessary conditions for `educwf` to be a valid instrument for `educw`? How can you check them?*
3. *Use 2SLS to instrument `educw` with `educwf`. What is the estimated effect of an extra year of education on the hourly wage?*
4. *Now, use a ready-made IV estimator instead. What is the estimated effect of an extra year of education on the hourly wage?*
5. *Use Bayesian inference to estimate the effect of an extra year of education on the hourly wage while instrumenting `educw` with `educwf`.*

6. *Compare the standard errors between the 2SLS and IV estimates as well as between the IV and OLS estimates.*

7. *Compare the 95% confidence intervals between the OLS and IV estimates. Can you deduce whether they are statistically different or not?*

Exercise 11.2 *Use the `Mroz` dataset from `Ecdat` to explain `hearnw`, the hourly wage of married women, as a function of `educw`, their education, `experience` and `experience^2`. Use the log-linear specification to model the relation. Use the subsample for whom the value of `work` is equal to `"no"`.*

1. *What is the estimated effect of an extra year of education on the hourly wage?*

   You suspect that education could be correlated with the error term in the OLS regression and would like to instrument it.

2. *Use `educwm`, mother's education as an instrument for `educw`. Do they covariate? What is the estimated effect of an extra year of education on the hourly wage?*

3. *Use `educwf`, father's education as an instrument for `educw`. Do they covariate? What is the estimated effect of an extra year of education on the hourly wage?*

4. *Now, use both instruments. Do they covariate with `educw`? What is the estimated effect of an extra year of education on the hourly wage?*

5. *Compare the standard errors among the three IV specifications.*

6. *Compare the 95(using both instruments). Can you deduce whether they are statistically different or not?*

   To check if education is actually correlated with the error term in the original OLS regression, re-estimate it while adding the residuals from fitting the reduced form equation (using both instruments) as an extra regressor.

7. *What is the estimated effect of an extra year of education on the hourly wage? Compare to the IV estimate.*

8. *Is the coefficient estimate at the residuals significant at the 5% level? Between the original OLS and IV, which estimator should you prefer based on the result of this test?*

12 Model Specification and Model Comparison

Inference:

- Frequentist

- Bayesian

Estimation:

- Method of Moments
- Maximum Likelihood
- MCMC sampling
- Bootstrap
- ⋮

Models:

- Large + small
- Nonlinear
- Discrete choice
- Mixed effects
- Instruments
- ⋮

12.1 Scaling coefficients

How should we add a new variable to the regression? — Scaling might simplify readability

```
attach(Caschool)
coef(lm(testscr ~ str + elpct + expnstu))

  (Intercept)           str         elpct       expnstu
649.577947257  -0.286399240  -0.656022660   0.003867902

elratio <- elpct/100
coef(lm(testscr ~ str + elratio + expnstu))

  (Intercept)           str       elratio       expnstu
649.577947257  -0.286399240 -65.602266008   0.003867902

expnstuTSD <- expnstu/1000
coef(lm(testscr ~ str + elpct + expnstuTSD))

(Intercept)         str       elpct  expnstuTSD
649.5779473  -0.2863992  -0.6560227   3.8679018
```

12.2 Which coefficients

- *testscr* ∼ *distcod + county + district + grspan + enrltot + teachers + calwpct + mealpct + computer + compstu + expnstu + str + avginc + elpct + readscr + mathscr* This model contains perhaps too many coefficients → multicollinearity)
- *testscr* ∼ *str* This model contains perhaps too few coefficients → omitted variable bias)
- Omitted variable bias

  $$\mathrm{E}(\mathbf{b}_1) = \boldsymbol{\beta}_1 + (X_1'X_1)^{-1}X_1'X_2\boldsymbol{\beta}_2$$

  Only when X_1 is orthogonal to X_2 or $\boldsymbol{\beta}_2$ is zero we have no bias
- Overfitting (multicollinearity)

  $$\Sigma_{\hat{\beta}\hat{\beta}} = (X'X)^{-1}X'\mathbf{Iu}^2X(X'X)^{-1}$$

  When X is (almost) collinear, then $(X'X)^{-1}$ is large, and then Σ is large, hence our estimates are not precise.

Model specification:

→ start with a base specification

→ building on these, develop alternative specifications

When coefficients change in an alternative specification, this can be a sign of omitted variable bias.

An example: The function *mtable* from the *memisc* package presents a table with an overview of different models. To adjust this table to our needs, we can adjust *getSummary.lm*. Mainly we use the version of *getSummary.lm* from the *memisc* package, but here we replace the (homoscedastic) standard error with the heteroscedasticity-consistent standard error:

```
library(memisc)
getSummary.lm <- function (est,...) {
  z <- memisc::getSummary.lm(est,...)
  z$coef[,"se"] <- sqrt(diag(hccm(est)))
  z
}
```

Now let us estimate a few models:

```
est1 <- lm(testscr ~ str)
est2 <- lm(testscr ~ str + elpct)
est3 <- lm(testscr ~ str + elpct + mealpct)
est4 <- lm(testscr ~ str + elpct + calwpct)
est5 <- lm(testscr ~ str + elpct + mealpct + calwpct)
```

```
mtable("(1)"=est1,"(2)"=est2,"(3)"=est3,"(4)"=est4,"(5)"=est5,
       summary.stats=c("R-squared","AIC","N"))
```

	(1)	(2)	(3)	(4)	(5)
(Intercept)	698.933***	686.032***	700.150***	697.999***	700.392***
	(9.467)	(7.411)	(4.686)	(6.024)	(4.698)
str	−2.280***	−1.101**	−0.998***	−1.308***	−1.014***
	(0.480)	(0.380)	(0.239)	(0.307)	(0.240)
elpct		−0.650***	−0.122***	−0.488***	−0.130***
		(0.039)	(0.032)	(0.033)	(0.034)
mealpct			−0.547***		−0.529***
			(0.022)		(0.032)
calwpct				−0.790***	−0.048
				(0.053)	(0.061)
R-squared	0.051	0.426	0.775	0.629	0.775
AIC	3650.499	3441.123	3050.999	3260.656	3052.376
N	420	420	420	420	420

Significance: $*** \equiv p < 0.001$; $** \equiv p < 0.01$; $* \equiv p < 0.05$

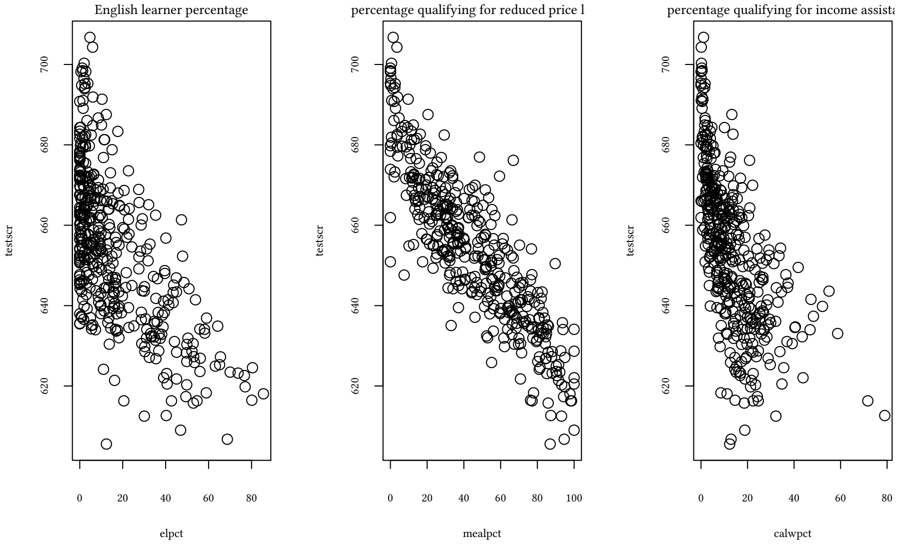
→ creates trust!

What is the benefit of one model over another

- measure R^2
- measure contribution to R^2
- look at p-value of the t-statistic
- look at p-value of the variance analysis
- measure AIC
- Bayesian model comparison

Perhaps it is helpful to control for *wealth* in the school district. Which variables could, in the Caschool example, be a good indicator for wealth?

```
plot(testscr ~ elpct,main="English learner percentage")
plot(testscr ~ mealpct,main="percentage qualifying for reduced price lunch")
plot(testscr ~ calwpct,main="percentage qualifying for income assistance")
```

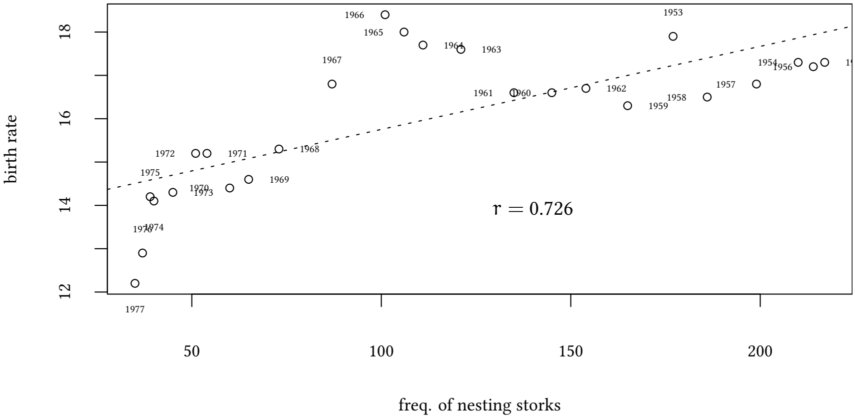


12.3 Measure R^2

$$R^2 = 1 - \frac{RSS}{TSS}$$

- R^2 only measures the "fit" of the regression
- R^2 does not measure causality (e.g. parkings lots → testscr)
- R^2 does not measure the absence of omitted variable bias

Remember: Birth rate and nesting storks

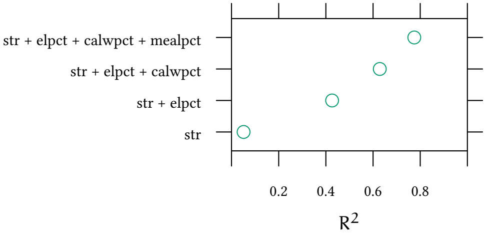


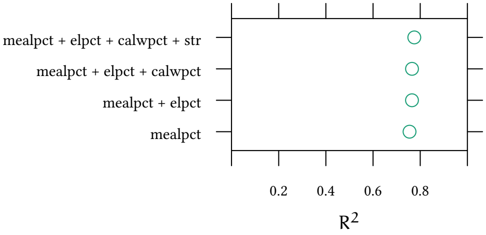
12.4 Nevertheless: Measure contribution to R^2

There are different ways to measure contribution to R^2:

Study R^2 when the variable we are analysing...

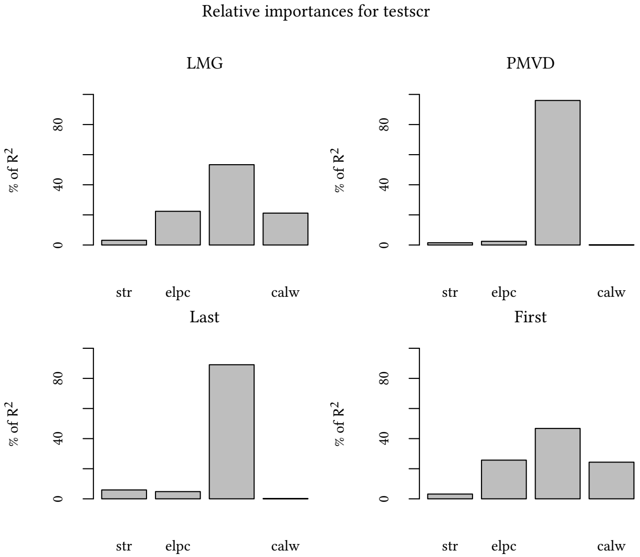
- is the only one (first) in the model
- when it is the last one
- all other feasible sequences
  - lmg (Lindeman, Merenda, Gold, 1980)
  - pmvd (proportional marginal variance decomposition, Feldman, 2005)





```
library(relaimpo)
est <- lm(testscr ~ str + elpct + mealpct + calwpct)
calc.relimp(est,type=c("first","last","lmg","pmvd"),
            rela=TRUE)

Response variable: testscr
Total response variance: 363.0301
Analysis based on 420 observations

4 Regressors:
str elpct mealpct calwpct
Proportion of variance explained by model: 77.49%
Metrics are normalized to sum to 100% (rela=TRUE).

Relative importance metrics:

               lmg         pmvd        last      first
str     0.03119231 0.0148176134 0.059126952 0.03175031
elpct   0.22371548 0.0242703918 0.048159854 0.25708495
mealpct 0.53343971 0.9600101671 0.890678586 0.46768098
calwpct 0.21165250 0.0009018276 0.002034608 0.24348376

Average coefficients for different model sizes:

                1X        2Xs        3Xs         4Xs
str     -2.2798083 -1.4612232 -1.1371224 -1.01435328
elpct   -0.6711562 -0.4347537 -0.2510901 -0.12982189
mealpct -0.6102858 -0.5922408 -0.5645062 -0.52861908
calwpct -1.0426750 -0.5863541 -0.2639020 -0.04785371
```

```
plot(calc.relimp(est,type=c("first","last","lmg","pmvd"),rela=TRUE))
```

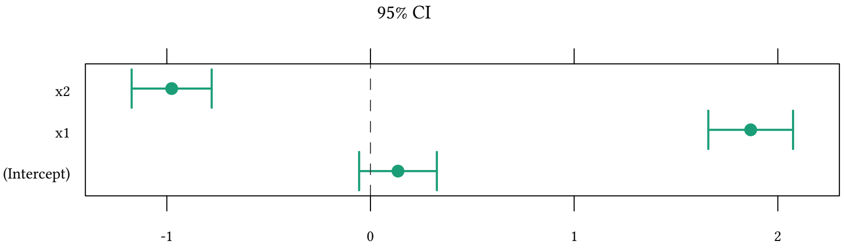


The contribution to the R^2 can often be more informative than coefficients or p-values.

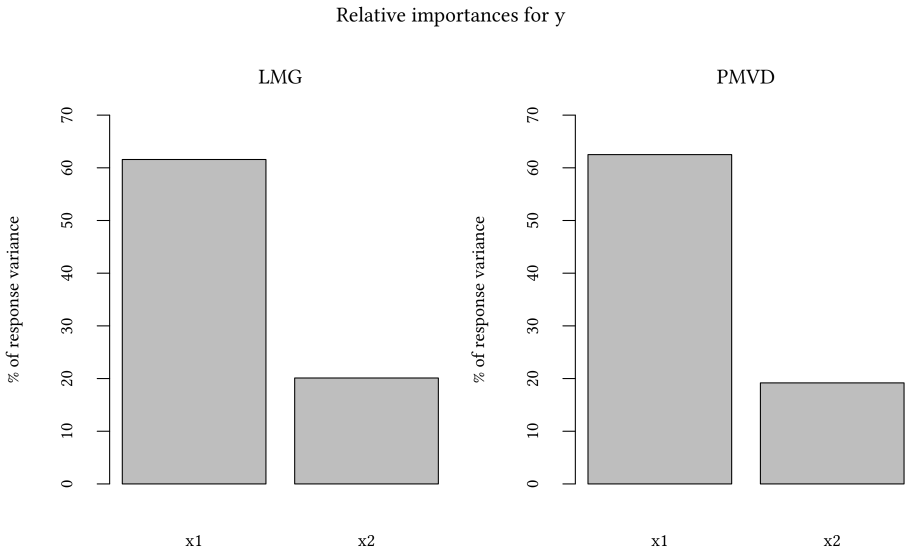
More on lmg versus pmvd

Let us start with a situation where explanatory variables x_1 and x_2 are independent. Simple situation, x_1 and x_2 are independent:

```
set.seed(123)
x1 <- rnorm(100)
x2 <- rnorm(100)
u  <- rnorm(100)
y  <- 2 * x1 - x2 + u
est<- lm(y ~ x1 + x2)
```

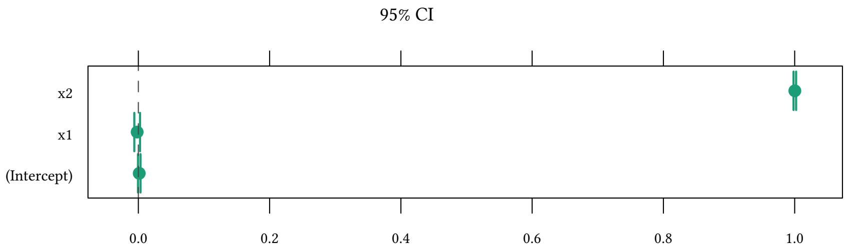


```
plot(calc.relimp(est,type=c("lmg","pmvd")))
```

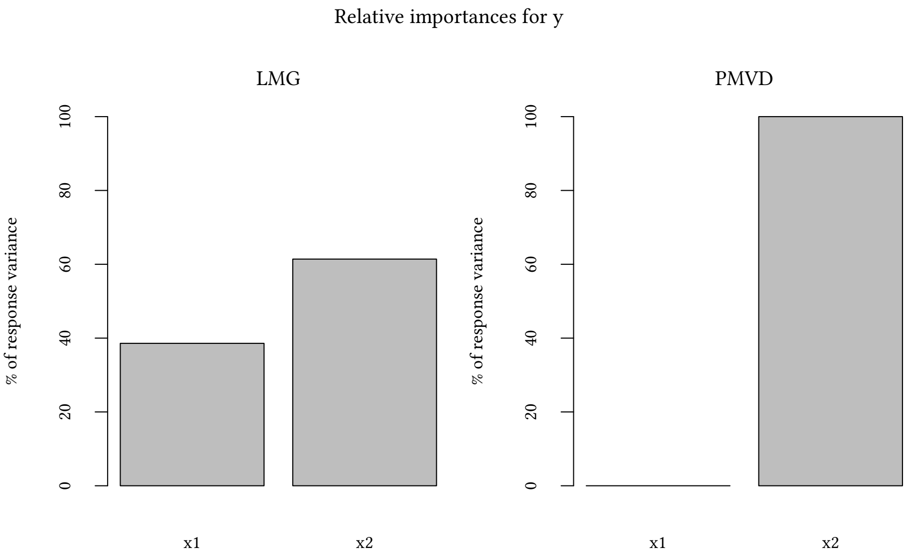


Now consider a situation where $x_1 \to x_2 \to y$:

```
set.seed(123)
x1 <- rnorm(100)
x2 <- rnorm(100) + 2 * x1
u  <- .01*rnorm(100)
y  <- x2 + u
est<- lm(y ~ x1 + x2)
```



```
plot(calc.relimp(est,type=c("lmg","pmvd")))
```



If *x1* and *x2* are independent, both *lmg* and *pmvd* allocate response variance in the same way.

However, if *x1* and *x2* are dependent it is possible that the coefficient of x_1 is close to zero but still correlated with y. In such a case *lmg* still allocates response variance to x_1 while *pmvd* does not.

12.5 Analysis of Variance

```
est <- lm(testscr ~ 1)
summary(est)


Call:
lm(formula = testscr ~ 1)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.606 -14.107   0.293  12.506  52.593 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 654.1565     0.9297   703.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.05 on 419 degrees of freedom
```

```
est <- lm(testscr ~ str + elpct + mealpct + calwpct)
summary(est)


Call:
lm(formula = testscr ~ str + elpct + mealpct + calwpct)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.179  -5.239  -0.185   5.171  31.308 

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)    
(Intercept) 700.39184    4.69797 149.084   < 2e-16 ***
str          -1.01435    0.23974  -4.231 0.0000286 ***
elpct        -0.12982    0.03400  -3.819  0.000155 ***
mealpct      -0.52862    0.03219 -16.422   < 2e-16 ***
calwpct      -0.04785    0.06097  -0.785  0.432974    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.084 on 415 degrees of freedom
Multiple R-squared:  0.7749,Adjusted R-squared:  0.7727 
F-statistic: 357.1 on 4 and 415 DF,  p-value: < 2.2e-16
```

Instead of looking at the p-value of a coefficient, we compare the variance of the residuals of two different models: model 2 (with the coefficient(s)), model 1 (without).

$$\frac{RSS_1 - RSS_2}{RSS_2} \frac{n - k_2}{k_2 - k_1} \sim F_{(k_2-k_1, n-k_2)}$$

We can, hence, use the F statistic to compare two models.

Let L be the log-likelihood of the estimated model.

One can show that for linear models

$$-2 \cdot L = n \cdot \log \frac{RSS}{n} + C$$

But then

$$2 \cdot (L_2 - L_1) = n \cdot \left(\log \frac{RSS_1}{n} - \log \frac{RSS_2}{n} \right) =$$

$$n \log \frac{RSS_1}{RSS_2} \sim \chi^2_{k_2-k_1}$$

We can, thus, also use the χ^2 statistic to compare two models.

```
est2 <- lm(testscr ~ str + elpct + mealpct + calwpct)
est1 <- lm(testscr ~ str +  mealpct + calwpct)
```

```
sum(est2$residuals^2)

[1] 34247.46

sum(est1$residuals^2)

[1] 35450.8
```

An easier way to obtain the RSS is `deviance`:

```
deviance(est2)

[1] 34247.46

RSS2<-deviance(est2)
RSS1<-deviance(est1)
L2 <- logLik(est2)
L1 <- logLik(est1)
n  <- length(est1$residuals)
k2 <- est2$rank
k1 <- est1$rank
```

$$2 \cdot (L_2 - L_1) \sim \chi^2_{k_2 - k_1}$$

```
pchisq(2 *(L2 - L1),k2-k1,lower=FALSE)

'log Lik.' 0.0001398651 (df=6)
```

$$n \log \frac{RSS_1}{RSS_2} \sim \chi^2_{k_2 - k_1}$$

```
pchisq(n * log(RSS1 / RSS2),k2-k1,lower=FALSE)

[1] 0.0001398651
```

```
anova(est1,est2,test="LRT")

Analysis of Variance Table

Model 1: testscr ~ str + mealpct + calwpct
Model 2: testscr ~ str + elpct + mealpct + calwpct
  Res.Df   RSS Df Sum of Sq  Pr(>Chi)
1    416 35451
2    415 34247  1    1203.3 0.0001342 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-test for ANOVA

```
pf((RSS1 - RSS2)/RSS2 * (n-k2)/(k2-k1),k2-k1,n-k2,lower=FALSE)

[1] 0.0001547027
```

```
anova(est1,est2)

Analysis of Variance Table

Model 1: testscr ~ str + mealpct + calwpct
Model 2: testscr ~ str + elpct + mealpct + calwpct
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)    
1    416 35451                                  
2    415 34247  1    1203.3 14.582 0.0001547 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(p-Value from χ^2 test was slightly different ($\Pr > \chi^2$=0.0001342))

12.5.1 t-test versus F test

The previous anova tests $\beta_{\texttt{elpct}} = 0$. This can also be done with a t-test:

```
summary(est2)


Call:
lm(formula = testscr ~ str + elpct + mealpct + calwpct)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.179  -5.239  -0.185   5.171  31.308 

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)    
(Intercept) 700.39184    4.69797 149.084   < 2e-16 ***
str          -1.01435    0.23974  -4.231 0.0000286 ***
elpct        -0.12982    0.03400  -3.819  0.000155 ***
mealpct      -0.52862    0.03219 -16.422   < 2e-16 ***
calwpct      -0.04785    0.06097  -0.785  0.432974    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.084 on 415 degrees of freedom
Multiple R-squared:  0.7749,Adjusted R-squared:  0.7727 
F-statistic: 357.1 on 4 and 415 DF,  p-value: < 2.2e-16
```

```
summary(est2)[["coefficients"]]["elpct","Pr(>|t|)"]

[1] 0.0001547027
```

```
anova(est2,est1)[["Pr(>F)"]][2]

[1] 0.0001547027
```

As long as we are testing

- only a single coefficient
- and use the same variance-covariance matrix

F-test (ANOVA) and t-test are equivalent.

- Advantage of ANOVA: Can test more than a single restriction.

12.6 Information criteria

Goal: Find a model that explains the data well but has as few parameters as possible. (prevent overfitting)

Let L be the log-likelihood of the estimated model.
Hirotsugo Akaike (1971): An Information Criterion:

$$AIC = -2 \cdot L + 2 \cdot k$$

Gideon E. Schwarz (1978): Bayesian Information Criterion

$$BIC = -2 \cdot L + k \cdot \log n$$

```
est <- lm(testscr ~ str + elpct + mealpct + calwpct + enrltot)
extractAIC(est)

[1]    6.000 1859.498
```

step now looks automatically for a model with a good AIC

```
step(est)

Start:  AIC=1859.5
testscr ~ str + elpct + mealpct + calwpct + enrltot

          Df Sum of Sq   RSS    AIC
- calwpct  1      70.1 34239 1858.4
- enrltot  1      78.9 34247 1858.5
<none>                 34169 1859.5
- elpct    1    1262.6 35431 1872.7
- str      1    1552.8 35721 1876.2
- mealpct  1   20702.3 54871 2056.4

Step:  AIC=1858.36
```

```
testscr ~ str + elpct + mealpct + enrltot

          Df Sum of Sq   RSS    AIC
- enrltot  1        60 34298 1857.1
<none>                 34239 1858.4
- elpct    1      1208 35446 1870.9
- str      1      1496 35734 1874.3
- mealpct  1     51150 85388 2240.2

Step:  AIC=1857.09
testscr ~ str + elpct + mealpct

          Df Sum of Sq   RSS    AIC
<none>                 34298 1857.1
- elpct    1      1167 35465 1869.1
- str      1      1441 35740 1872.4
- mealpct  1     52947 87245 2247.2

Call:
lm(formula = testscr ~ str + elpct + mealpct)

Coefficients:
(Intercept)          str        elpct      mealpct
   700.1500      -0.9983      -0.1216      -0.5473
```

What the AIC tries to achieve

- within sample prediction
  - including more variables always improve the likelihood
- out of sample prediction
  - including more variables may decrease the likelihood

AIC tries to maximise out of sample performance.

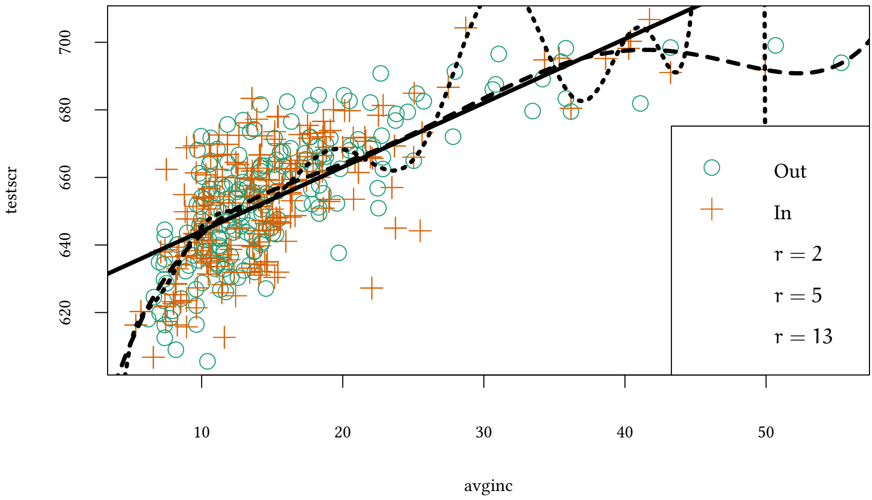
```
set.seed(123)
N<-nrow(Caschool)
mySamp<-sample(1:N,N/2)
CaIn<-Caschool[mySamp,]
CaOut<-Caschool[-mySamp,]
est  <-lm(testscr ~ str + elpct + mealpct + calwpct + enrltot,data=CaIn)
estSm<-lm(testscr ~ str + elpct + mealpct           + enrltot,data=CaIn)
deviance(est)

[1] 16727.67

deviance(estSm)

[1] 16953.53
```

→ within sample a smaller model must have a *larger* deviance.
Now do the same out of sample:

```
sum((CaOut$testscr-predict(est,newdata=CaOut))^2)

[1] 17831.88

sum((CaOut$testscr-predict(estSm,newdata=CaOut))^2)

[1] 17382.6
```

Out of sample a smaller model can have a *smaller* deviance.

```
newdata<-list(avginc=seq(1,60,length.out=150))
pplot <- function(deg,lty) with(CaIn,
   lines(predict(lm(testscr ~ poly(avginc,deg)),
                 newdata=newdata)~newdata$avginc,lty=lty,lwd=4))
plot(testscr ~ avginc,pch=pch,col=pcol)
with(CaIn,{
     points(avginc,testscr,pch=3,col=2);pplot(1,1);pplot(5,2);pplot(13,3)})
legend("bottomright",c("Out","In","$r=2$","$r=5$","$r=13$"),lty=c(NA,NA,1:3),
       pch=c(1,3,NA,NA,NA),col=c(1,2,NA,NA,NA),lwd=c(1,1,4,4,4),bg="white")
```

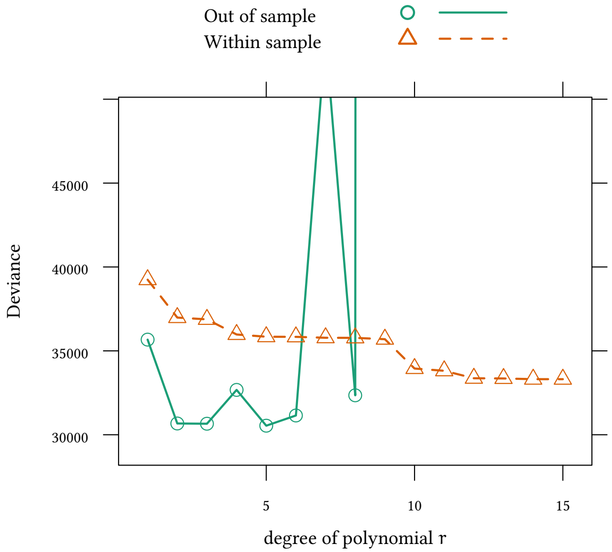


Of course, the above graph depends on the specific sample of `CaIn` and `CaOut`. We could repeat this exercise for many samples.

The same idea with polynomial functions

Within sample deviance:

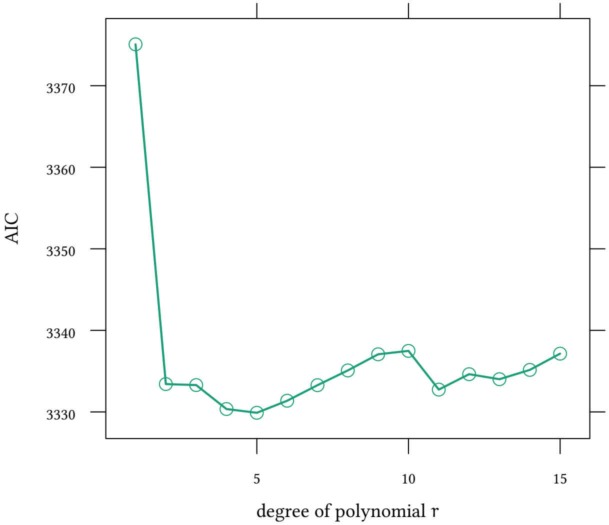
```
r <- 1:15
dev.data<-rbind(
    data.frame(type="Within sample", r=r, dev=sapply(r,function(r)
        deviance(lm(testscr~poly(avginc,r), data=CaIn)))),
    data.frame(type="Out of sample",r=r,dev=sapply(r,function(r)
        sum((CaOut$testscr-predict(lm(testscr~poly(avginc,r), data=CaIn),
                                   newdata=CaOut))^2))))
xyplot(dev ~ r,group=type,data=dev.data,
       ylim=c(10^4.45,10^4.7),
       type=c("p","l"),
       xlab="degree of polynomial $r$",ylab="Deviance",
       auto.key=list(space="top",lines=TRUE))
##plot(sapply(1:15,function(r) deviance(lm(testscr~poly(avginc,r),data=CaIn))),
##     xlab="degree of polynomial $r$",ylab="within sample deviance",log="y")
```



Of course, the above graph depends on the specific sample of *CaIn* and *CaOut*. We could repeat this exercise for many samples.

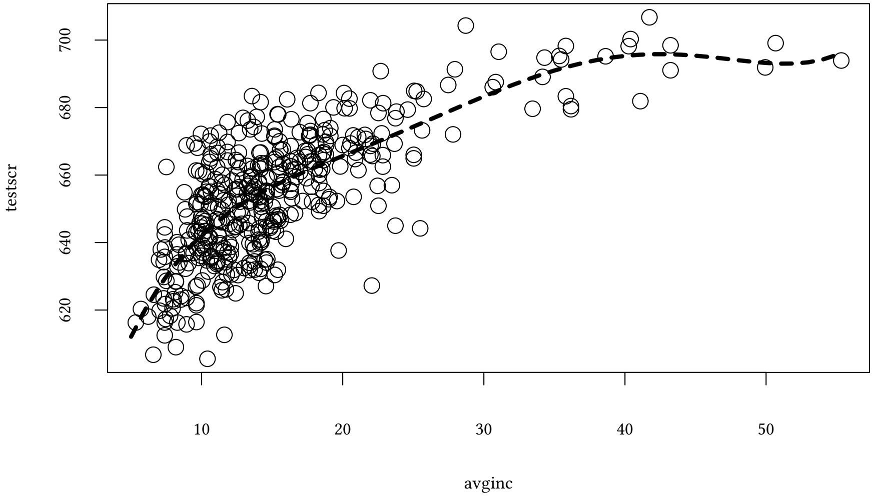
AIC tries to capture the quality of out of sample prediction:

```
aics<-sapply(r,function(r) AIC(lm(testscr~poly(avginc,r))))
xyplot(aics ~ r,xlab="degree of polynomial $r$",ylab="AIC",type=c("p","l"))
```



If we want to model the relation between *avginc* and *testscr*, perhaps the best is a polynomial of degree 5:

```
newdata<-list(avginc=5:55)
plot(testscr ~ avginc)
lines(predict(lm(testscr ~ poly(avginc,5)),newdata=newdata)~newdata$avginc,lty=2,lwd=4)
```



Frequentist Null Hypothesis Testing
(ANOVA, t-test)

- What is the correct model?

Information criteria
(AIC)

- What is a good prediction?
  (out of sample)

12.7 Hypothesis tests for individual coefficients

```r
summary(lm(testscr ~ str + elpct + mealpct + calwpct))
```

```

Call:
lm(formula = testscr ~ str + elpct + mealpct + calwpct)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.179  -5.239  -0.185   5.171  31.308 

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)    
(Intercept) 700.39184    4.69797 149.084   < 2e-16 ***
str          -1.01435    0.23974  -4.231 0.0000286 ***
elpct        -0.12982    0.03400  -3.819  0.000155 ***
mealpct      -0.52862    0.03219 -16.422   < 2e-16 ***
calwpct      -0.04785    0.06097  -0.785  0.432974    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.084 on 415 degrees of freedom
Multiple R-squared:  0.7749,Adjusted R-squared:  0.7727 
F-statistic: 357.1 on 4 and 415 DF,  p-value: < 2.2e-16
```

- p-values are for hypothesis tests $H_0 : \beta_i = 0$, $H_1 : \beta_i \neq 0$.
- p-values do not take into account out-of sample performance (AIC does).
- p-values do not take into account explanatory power (contribution to R^2 does).
- p-values do not take into account economic relevance.
- p-values do not take into account economic plausibility.

12.8 Comparing GLM models

Labour market participation

Let us look at an example from labour market participation. The dataset `Participation` contains 872 observations from Switzerland. The variable `lfp` is a binary variable which is "yes" for a participating individual and "no" otherwise. The variable `lnnlinc` describes the nonlabour income. A hypothesis might be that the higher the nonlabour income, the less likely is it that the individual participates.

12.8.1 Goodness of fit

We can no longer use R^2 or RSS to assess goodness of fit. Instead we use likelihood based statistics.

12.8.2 Likelihood ratio index and AIC

- Mac Fadden's likelihood ratio index (similar to R^2, Pseudo R^2 in Stata):

$$\mathrm{LRI} = 1 - \frac{\log L}{\log L_0}$$

  Pseudo R^2 is different from R^2:

```
1-logLik(lm(testscr ~ str))/logLik(lm(testscr ~ 1))

'log Lik.' 0.006025159 (df=3)

summary(lm(testscr ~ str))[["r.squared"]]

[1] 0.0512401
```

- More parameters (K) will give a better fit. We introduce a penalty for "overfitting".

  Akaike "An Information Criterion": $\mathrm{AIC} = -2\log L + 2K$

- Bayesian information criterion (BIC): $\mathrm{BIC} = -2\log L + K\log N$

  We will prefer the model with the lower AIC or BIC.

12.8.3 An example: Labour market participation

```
data(Participation)
reg<-glm ( lfp ~ lnnlinc , family=binomial(link="logit"), data=Participation )
```

```

Call:
glm(formula = lfp ~ lnnlinc, family = binomial(link = "logit"), 
    data = Participation)

Coefficients:
            Estimate Std. Error z value    Pr(>|z|)    
(Intercept)   9.6276     1.9728   4.880 0.000001060 ***
lnnlinc      -0.9165     0.1847  -4.962 0.000000698 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1203.2  on 871  degrees of freedom
Residual deviance: 1176.0  on 870  degrees of freedom
AIC: 1180

Number of Fisher Scoring iterations: 4
```

```
reg0<-glm ( lfp ~ 1 , family=binomial(link="logit"), data=Participation )
summary(reg0)


Call:
glm(formula = lfp ~ 1, family = binomial(link = "logit"), data = Participation)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)  
(Intercept) -0.16090    0.06795  -2.368   0.0179 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1203.2  on 871  degrees of freedom
Residual deviance: 1203.2  on 871  degrees of freedom
AIC: 1205.2

Number of Fisher Scoring iterations: 3
```

AIC:

```
extractAIC(reg)

[1]    2.000 1179.972

extractAIC(reg0)

[1]    1.000 1205.223
```

Hence, we should prefer *reg* over *reg0*.

Mac Fadden's likelihood ratio index:

```
1-logLik(reg)/logLik(reg0)

'log Lik.' 0.02264865 (df=2)
```

The likelihood ratio:

$$2\log\frac{L_l}{L_s}\sim\chi^2_{l-s}$$

where L_s and L_l are likelihoods of a small model (with s parameters) and a large model (with l parameters), respectively.

```
pchisq (2 * (logLik(reg) - logLik(reg0))[1] , 1 , lower.tail=FALSE)

[1] 0.0000001786474
```

Apparently, *`lnnlinc`* significantly improves the fit of the model.
We can obtain the same information more conveniently with

```
anova(reg0,reg,test="LRT")

Analysis of Deviance Table

Model 1: lfp ~ 1
Model 2: lfp ~ lnnlinc
  Resid. Df Resid. Dev Df Deviance     Pr(>Chi)
1       871     1203.2
2       870     1176.0  1   27.251 0.0000001786 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

or even with

```
anova(reg,test="LRT")

Analysis of Deviance Table

Model: binomial, link: logit

Response: lfp

Terms added sequentially (first to last)


        Df Deviance Resid. Df Resid. Dev     Pr(>Chi)
NULL                      871     1203.2
lnnlinc  1   27.251       870     1176.0 0.0000001786 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

12.8.4 z-test versus χ^2 test

Appendix 12.C presents an extension to this model.

12.9 Bayesian Model Comparison

12.9.1 Preliminaries

Consider β_2 in the following relationship:

$$\mathtt{testscr} = \beta_0 + \beta_1 \mathtt{elpct} + \beta_2 \mathtt{str} + u$$

- What kind of probabilistic statements can we make about β_2?
- We can, e.g., compare $\Pr(\beta_2 \leq 0)$ with $\Pr(\beta_2 > 0)$.

```
pre.mod <- 'model {
   for (i in 1:length(y)) {
     y[i] ~ dnorm(beta[1]+beta[2]*x1[i]+beta[3]*x2[i],tau)
   }
   for(k in 1:3) {
      beta[k] ~ dnorm(0,.0001)
   }
   tau ~ dgamma(.01,.01)
}'
pre.data<-list(y=testscr-mean(testscr),
               x1=elpct-mean(elpct),x2=str-mean(str))
pre.jags<-run.jags(pre.mod,data=pre.data,monitor="beta",inits=genInit(4))
pre.beta3<-as.mcmc(pre.jags)[,"beta[3]"]
```

12.9.2 Odds in the simple model

$$\mathtt{testscr} = \beta_0 + \beta_1 \mathtt{elpct} + \beta_2 \mathtt{str} + u$$

```
mean(pre.beta3 > 0)

[1] 0.002025

mean(pre.beta3 <= 0)

[1] 0.997975
```

- $\Pr(\beta_2 > 0) =$0.002025
- $\Pr(\beta_2 \leq 0) =$0.997975

Odds: $\frac{\Pr(\beta_2 \leq 0)}{\Pr(\beta_2 > 0)} =$493

12.9.3 Compare different models

↑ So far we estimate only a single coefficient. Can we generalise this idea to comparing different models?

- Idea: A binary process selects (randomly) among the two models we want to compare.

Example 1: select among three models

- Data:

```
set.seed(123)
x <- rnorm(5,5,1)
```

- $x \sim N(\mu, 1)$ where N is the normal distribution.

  We compare three models:

  $\mu_1 = 3.8$, $\mu_2 = 4$, $\mu_3 = 6$.

```
c.model <- 'model {
for (i in 1:length(x)) {
   x[i] ~ dnorm(mu[m],1)
}
 m ~ dcat(modelProb)
}'
c.data<-list(x=x,modelProb=c(1,1,1),
             mu=c(3.8,4,6))
c.jags<-run.jags(c.model,c.data,
                 monitor=c("m"))
```

```
with(data.frame(as.mcmc(c.jags)),
     table(m))

m
    1     2     3
  671  2411 16918
```

```
with(data.frame(as.mcmc(c.jags)),
    table(m)/length(m))

m
      1       2       3
0.03355 0.12055 0.84590
```

What is, actually, a vague prior in this context? Can we give more flexibility to the prior for *m*?

How can we motivate the parameter of `dcat`?

```
c.model2 <- 'model {
for (i in 1:length(x)) {
   x[i] ~ dnorm(mu[m],1)
}
 m ~ dcat(mP)
 mP ~ ddirch(modelProb)
}'
c.jags<-run.jags(c.model2,c.data,monitor=c("m"))
with(data.frame(as.mcmc(c.jags)),table(m)/length(m))

m
     1      2      3
0.0344 0.1306 0.8350
```

(`ddirch` denotes the Dirichlet distribution, a generalisation of the Beta distribution.)

```
c.model3 <- 'model {
for (i in 1:length(x)) {
   x[i] ~ dnorm(mu[m],1)
}
m ~ dcat(modelProb)
for (i in 1:length(modelProb)) {
   mSel[i] <- ifelse(m==i,1,0)
}
}'
c.jags<-run.jags(c.model3,c.data,monitor=c("mSel"))
summary(c.jags)[,c("Mean","SSeff","psrf")]

           Mean SSeff        psrf
mSel[1] 0.03405 20000 0.999960055
mSel[2] 0.12070 20000 1.000078543
mSel[3] 0.84525 20000 1.000046582
```

Example 2

- A binary process selects (randomly) among the two models we want to compare.

  1. $\texttt{testscr} = \beta_0 + \beta_1\texttt{elpct} + \beta_2\texttt{str} + u$
  2. $\texttt{testscr} = \beta_0 + \beta_1\texttt{elpct} + u$

- Problem: While one of the two models is “not selected”, parameters of this model can take any value (and do not reduce likelihood).

  → convergence is slow!

- Solution: “Pseudopriors” (the binary process already has informed priors about the two models)

```
select.model <- 'model {
    for (i in 1:length(y)) {
        y[i] ~ dnorm(ifelse(equals(mI,0),
                        beta[1]+beta[2]*x1[i]+beta[3]*x2[i],
                        beta[4]+beta[5]*x1[i]),
                    tau[mI+1])
    }
    for (j in 1:5) {
       beta[j] ~ dnorm(priBeta[j,1],1/priBeta[j,2]^2)
    }
    for (j in 1:2) {
       tau[j] ~ dgamma(priTau[j,1]^2/priTau[j,2]^2,priTau[j,1]/
                                                    priTau[j,2]^2)
    }
    mI ~ dbern(p)
    p ~ dunif(0,1)
}'
```

The technical details of the implementation can be found in Appendix 12.D.

12.9.4 Result of the Bayesian Model Comparison

Interesting is the result for *mI*. This variable has the value 0 or 1 for model 0 or model 1 respectively.

mean(mI) =0.02205 means the following:

- In 0.97795 of all cases we have model 0 (the larger model)
- In 0.02205 of all cases we have model 1 (the smaller model).

The odds $\frac{\Pr(\text{model } 0)}{\Pr(\text{model } 1)}$ =44.4.

Compare with p-values from the frequentist analysis:

```
summary(lm(testscr ~ elpct + str ))



Call:
lm(formula = testscr ~ elpct + str)

Residuals:
      Min        1Q    Median        3Q       Max
-48.84466 -10.24036  -0.30784   9.81535  43.46070

Coefficients:
               Estimate  Std. Error   t value   Pr(>|t|)
(Intercept) 686.0322487   7.4113125  92.56555 < 2.22e-16 ***
elpct        -0.6497768   0.0393425 -16.51588 < 2.22e-16 ***
```

```
str          -1.1012959   0.3802783  -2.89603  0.0039781 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.4645 on 417 degrees of freedom
Multiple R-squared:  0.426431,Adjusted R-squared:  0.42368
F-statistic: 155.014 on 2 and 417 DF,  p-value: < 2.22e-16
```

Compare with p-values Why are these numbers so different from p-values?

- p-value=$\Pr(\text{data}|H_0)$
  Here H_0 = “model 1 is true”
  - p-value=$\Pr(\text{data}|\text{model 1}) = 0.003978056$
- In our Bayesian model comparison we have calculated
  - $\Pr(\text{model 0}|\text{data}) = 0.97795$
  - $\Pr(\text{model 1}|\text{data}) = 0.02205$.

Model comparison: Words of caution Whenever we compare models, we have to ask ourselves:

Are the models we are comparing plausible?

Example: Returning from the lecture, I find on my desk a book which I long wanted to read.

- Model 1: The book has been provided by Santa Claus.
- Model 2: The book has been provided by the Tooth Fairy.

A careful model comparison finds Santa Claus 50 times more likely than the Tooth Fary. Does this mean that Santa Claus, indeed, brought the book? (Possibly not!)
The same applies, of course, to interpreting p-values from hypothesis tests.

12.9.5 Model uncertainty

Frequentist Null Hypothesis Testing
Inference is based on one model (this is restrictive).

Bayesian
Composite inference from model posteriors (more flexible).
What if there is a large number of possible models?

- Occam’s window: consider a subset of plausible and not too complex models.
- Markov Chain Monte Carlo Model Composition (MC^3).

12.9.6 Bayes factors

Posterior probability of Model H_1:

$$\Pr(H_1|X) = \Pr(H_1) \cdot \Pr(X|H_1) \frac{1}{\Pr(X)}$$

Hence

$$\frac{\Pr(H_1|X)}{\Pr(H_2|X)} = \frac{\Pr(H_1) \cdot \Pr(X|H_1)}{\Pr(H_2) \cdot \Pr(X|H_2)}$$

For uninformed priors $\Pr(H_1) = \Pr(H_2)$ we have the *Bayes factor*

$$K = \frac{\Pr(X|H_1)}{\Pr(X|H_2)} = \underbrace{\frac{\int \Pr(\theta_1|H_1) \Pr(X|\theta_1, H_1)\, d\theta_1}{\int \Pr(\theta_2|H_2) \Pr(X|\theta_2, H_2)\, d\theta_2}}_{\text{Bayes factor}} \neq \underbrace{\frac{\Pr(X|\theta_1^*, H_1)}{\Pr(X|\theta_2^*, H_2)}}_{\text{LR-test}}$$

Interpreting K: Harold Jeffreys (1961):

$10^0 = 1$	$10^{0.5} \approx 3$	10^1	$10^{1.5} \approx 30$	10^2
barely worth mentioning	substantial	strong	very strong	decisive

Robert E. Kass and Adrian E. Raftery (1995):

$e^0 = 1$	$e^1 \approx 2.7$	$e^3 \approx 20$	$e^5 \approx 150$
barely worth mentioning	positive	strong	very strong

Frequentist Null Hypothesis Testing
(ANOVA, t-test)

- What is the correct model?

  (Problem: many models are “not rejected”, they can’t be all “correct”)

Information criteria
(AIC)

- What is a good prediction?

  (out of sample)

Bayes ratio

- How probable is a model?

12.10 Literature

- John H. Kruschke. “Doing Bayesian Data Analysis. A Tutorial with R, JAGS, and Stan”. Academic Press. 2014.
- Hoff, A First Course in Bayesian Statistical Methods.
- William H. Greene. “Econometric Analysis”. Chapter 16.

Appendix 12.A Examples for the lecture

Example 1 Your sample contains a random variable y which is drawn from one of two possible distributions. Model A assumes a normal distribution with mean 0 and precision 1. Model B assumes a normal distribution with mean 3 and precision 1. You use JAGS to determine the posterior probabilities. Consider the following model:

```
for (i in 1:length(y)) {
  y[i] ~ α
}
j ~ dbern(1/3)
```

What is the value of α?

- `dnorm(ifelse(equals(j,1),0,3),1)`
- `ifelse(dnorm(equals(j,1),0,3),1)`
- `dnorm(equals(ifelse(j,1),0,3),1)`
- `equals(dnorm(ifelse(j,1),0,3),1)`
- other value

Example 2 What was your prior probability of Model A in the above model?

- 1/3
- 1/2
- 2/3
- 1
- other value

Example 3 You estimate the following equation:

$$Y = \sum_{i=0}^{k} \beta_i X^i + \epsilon$$

The following table shows the AIC of the estimated model for different values of k:

k	1	2	3	4	5	6	7
AIC	192.78	165.21	147.83	90.07	78.05	5.93	7.76

Which value of k should you choose, based on the AIC?

- 1
- 2
- 4
- 6
- other value

Example 4 You use the following model for model selection in JAGS (α is a placeholder):

```
  for (i in 1:length(y)) {    y[i] ~ dnorm(ifelse(equals(mI,0),
      b[1]+b[2]*x1[i],        b[3]+b[4]*x2[i]),      tau[mI+1]) } for (j in 1:4) {
b[j] ~ dnorm(0,.001) } for (j in 1:2) { tau[j] ~ dexp(0.01)} mI ~ α p ~ dunif(0,1)
```

What should you fill in for α?

- `dpois(p)`
- `dbern(p)`
- `p`
- `exp(p)`
- other value

Which econometric models does the above JAGS model compare?

- $Y = \beta_0 + \beta_1 X_1 + \epsilon$ against $Y = \beta_0 + \beta_1 X_2 + \epsilon$
- $Y = \beta_0 + \beta_1 X_1 + \epsilon$ against $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
- $Y = \beta_0 + \beta_1 X_1 + \epsilon$ against $Y = \beta_0 + \beta_1 X_1^2 + \epsilon$
- $Y = \beta_0 + \beta_1 X_1 + \epsilon$ against $Y = \beta_0 + \epsilon$
- other value

Example 5 In your output from JAGS you obtain a 95%-credible interval for `mI` of $\mathrm{CI}_{95\%} = [0, 1]$ and a mean for `mI` of 0.3. What can you conclude?

- The probability of the model $Y = \beta_0 + \beta_1 X_1 + \epsilon$ is 0.3.
- The probability of the model $Y = \beta_0 + \beta_1 X_1 + \epsilon$ is 0.7.
- The probability of the model $Y = \beta_0 + \beta_1 X_1 + \epsilon$ is 0.5.
- Since the credible interval for `mI` includes the extremes 0 and 1 one can not make a statement how probable the models are.

Appendix 12.B Exercises

To solve the next exercises the following commands for R might help: `data, order, plot, abline, lines, fitted, lm, glm, summary, anova, step, lmtest, lht, mtable, runjags` and `calc.relimp` from `relaimpo`.

Exercise 12.1 *Use the* `Housing` *dataset from* `Ecdat` *to estimate the effect of* `lotsize` *(x) on* `price` *(y). Compare/interpret the coefficient estimates, predictions, and fit (R^2, adj. R^2, AIC, BIC) across a number of model specifications. Which specification tracks the data the best? Which specification is the correct one?*

The model specifications are as follows:

1. $y = \beta_0 + \beta_1 \cdot x + u$
2. $y = \beta_0 + \beta_1 \cdot \log(x) + u$
3. $\log(y) = \beta_0 + \beta_1 \cdot x + u$
4. $\log(y) = \beta_0 + \beta_1 \cdot \log(x) + u$
5. $y = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2 + u$
6. $y = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2 + \beta_3 \cdot x^3 + u$

Exercise 12.2 *Use the* `Computers` *dataset from* `Ecdat` *to explain the price of a personal computer as a function of its characteristics. Consider the following variables as potential regressors:* `speed`, `hd`, `ram`, `screen`, `cd`, `multi`.

1. *How much of the variation in* `price` *can be explained by all the regressors combined?*
2. *Use ANOVA to measure the contribution of individual regressors to the total R-squared. Which appear the most/least important?*
3. *Again, use ANOVA but have the order of the regressors in your specification reversed. What happens to their individual contributions to the total R-squared? Explain.*

4. *Use the* `first`*,* `last` *and* `lmg` *metrics from the* `relaimpo` *library to measure the contribution of individual regressors to the total R-squared. Which regressor(s) should you include if you want your specification to have e.g., only one, three, or five regressors?*

5. *Use the in-built* `step` *procedure to determine which variables to include in the model. What specification does the procedure suggest? Why is it better than using all six variables? Compare to what the* `last` *metric suggests.*

6. *Can you discriminate between the specification suggested in (e) and one that includes all six variables using alternative tests?*

Exercise 12.3 *Use the* `Computers` *dataset from* `Ecdat` *to explain the price of a personal computer as a function of its characteristics. You consider two specifications:*

1. `price ~ speed + hd + ram + screen + cd` *and*

2. `price ~ speed + hd + ram + screen + multi`*.*

1. *Can you use, e.g., a* t*-,* F*-, or a likelihood ratio test to discriminate between the two specifications?*

2. *Can you use* R^2 *or information criteria to discriminate between the two specifications?*

3. *Can you use Bayesian inference to discriminate between the two specifications (assuming convergence)?*

Appendix 12.C GLM - A larger model

Now consider an extension to Section 12.8. Add more parameters to the model:

lfp	labour force participation
lnnlinc	the log of nonlabour income
age	age in years divided by 10
educ	years of formal education
nyc	the number of young children (younger than 7)
noc	number of older children

```
reg4<-glm(lfp ~ lnnlinc + age + educ + nyc + noc ,
          family=binomial(link="logit"),data=Participation)
summary(reg4)


Call:
glm(formula = lfp ~ lnnlinc + age + educ + nyc + noc, family = binomial(link = "logit"),
    data = Participation)

Coefficients:
```

```
               Estimate Std. Error  z value   Pr(>|z|)
(Intercept) 12.4332220  2.1440371  5.79898 6.6721e-09 ***
lnnlinc     -0.8941444  0.2040576 -4.38182 1.1769e-05 ***
age         -0.5639514  0.0889128 -6.34275 2.2571e-10 ***
educ        -0.0458492  0.0259706 -1.76543   0.077492 .
nyc         -1.2210057  0.1757883 -6.94589 3.7609e-12 ***
noc         -0.0163471  0.0722887 -0.22614   0.821095
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1203.223  on 871  degrees of freedom
Residual deviance: 1098.914  on 866  degrees of freedom
AIC: 1110.914

Number of Fisher Scoring iterations: 4
```

We see that the parameter *educ* is not really significant. We can use the likelihood ratio to test whether *educ* contributes significantly to the model:

```
reg3 <- update ( reg4 , ~ . - educ)


Call:
glm(formula = lfp ~ lnnlinc + age + nyc + noc, family = binomial(link = "logit"),
    data = Participation)

Coefficients:
                Estimate  Std. Error  z value   Pr(>|z|)
(Intercept) 13.19940863  2.12457675  6.21272 5.2074e-10 ***
lnnlinc     -1.01664147  0.19467228 -5.22232 1.7669e-07 ***
age         -0.53862389  0.08745047 -6.15919 7.3119e-10 ***
nyc         -1.20847373  0.17506732 -6.90291 5.0948e-12 ***
noc         -0.00286754  0.07183633 -0.03992    0.96816
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1203.223  on 871  degrees of freedom
Residual deviance: 1102.044  on 867  degrees of freedom
AIC: 1112.044

Number of Fisher Scoring iterations: 4
```

ANOVA with the larger model

```
anova(reg3,reg4,test="LRT")

Analysis of Deviance Table
```

```
Model 1: lfp ~ lnnlinc + age + nyc + noc
Model 2: lfp ~ lnnlinc + age + educ + nyc + noc
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       867   1102.044
2       866   1098.914  1 3.129689 0.076878 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here z-test and χ^2 are more similar (still different):

- p-value from z-test (from summary table):

  0.0775

- p-value from χ^2-test (from ANOVA):

  0.0769

AIC with the larger model

```
step(reg4)

Start:  AIC=1110.91
lfp ~ lnnlinc + age + educ + nyc + noc

          Df Deviance      AIC
- noc      1 1098.965 1108.965
<none>       1098.914 1110.914
- educ     1 1102.044 1112.044
- lnnlinc  1 1120.063 1130.063
- age      1 1142.217 1152.217
- nyc      1 1157.589 1167.589

Step:  AIC=1108.97
lfp ~ lnnlinc + age + educ + nyc

          Df Deviance      AIC
<none>       1098.965 1108.965
- educ     1 1102.045 1110.045
- lnnlinc  1 1121.089 1129.089
- age      1 1145.938 1153.938
- nyc      1 1164.962 1172.962

Call:  glm(formula = lfp ~ lnnlinc + age + educ + nyc, family = binomial(link = "logit"),
    data = Participation)

Coefficients:
(Intercept)      lnnlinc          age         educ          nyc
 12.4641761   -0.9018712   -0.5576145   -0.0452272   -1.2071992

Degrees of Freedom: 871 Total (i.e. Null);  867 Residual
```

```
Null Deviance:     1203.22
Residual Deviance: 1098.97  AIC: 1108.97
```

Note that...

- p-values from ANOVA would suggest to remove *educ* from the model.
- AIC would suggest to keep *educ* in the model.

Appendix 12.D Technical details for model comparison

- Idea: A binary process selects (randomly) among the two models we want to compare.
    1. $\texttt{testscr} = \beta_0 + \beta_1\texttt{elpct} + \beta_2\texttt{str} + u$
    2. $\texttt{testscr} = \beta_0 + \beta_1\texttt{elpct} + u$
- Problem: While one of the two models is “not selected”, parameters of this model can take any value (and do not reduce likelihood).

  $\rightarrow$ convergence is slow!
- Solution: “Pseudopriors” (the binary process already has informed priors about the two models)

Bayesian Model Comparison: Specifying priors

Remember: the “uninformed” priors we used above were similar to the following:
$\beta_i \sim \texttt{dnorm}(0, 0.0001)$
$\tau \sim \texttt{dgamma}(.01, .01)$
Below we want to give some (informed) help to the Bayesian model:

- Pseudoprior for β is trivial.
    - Mean of β can be taken from the estimate of β.
    - Precision $\tau = 1/\sigma^2$, and we can find σ from the estimate for β.
- To obtain the prior for τ we use properties of the Γ-distribution:

  If $\tau \sim \Gamma(\alpha, \beta)$ then $E(\tau) = \alpha/\beta$ and $\operatorname{var}(\tau) = \alpha/\beta^2$

  We estimate parameters of the Γ-distribution as follows:

  $\hat{\alpha} = \bar{\tau}^2/\operatorname{var}(\tau)$, $\hat{\beta} = \bar{\tau}/\operatorname{var}(\tau)$.

Bayesian Model Comparison: 1st model

```
select.model <- 'model {
    for (i in 1:length(y)) {
        y[i] ~ dnorm(ifelse(equals(mI,0),
                        beta[1]+beta[2]*x1[i]+beta[3]*x2[i],
                        beta[4]+beta[5]*x1[i]),
                     tau[mI+1])
    }
    for (j in 1:5) {
       beta[j] ~ dnorm(priBeta[j,1],1/priBeta[j,2]^2)
    }
    for (j in 1:2) {
       tau[j] ~ dgamma(priTau[j,1]^2/priTau[j,2]^2,priTau[j,1]/
                                                   priTau[j,2]^2)
    }
    mI ~ dbern(p)
    p ~ dunif(0,1)
}'
```

No pseudopriors. Construct the vague prior for β and τ:

```
(priBeta<-matrix(c(0,100),nrow=5,ncol=2,byrow=TRUE))

     [,1] [,2]
[1,]    0  100
[2,]    0  100
[3,]    0  100
[4,]    0  100
[5,]    0  100

(priTau<-matrix(c(1,10),nrow=2,ncol=2,byrow=TRUE))

     [,1] [,2]
[1,]    1   10
[2,]    1   10
```

Estimate the model:

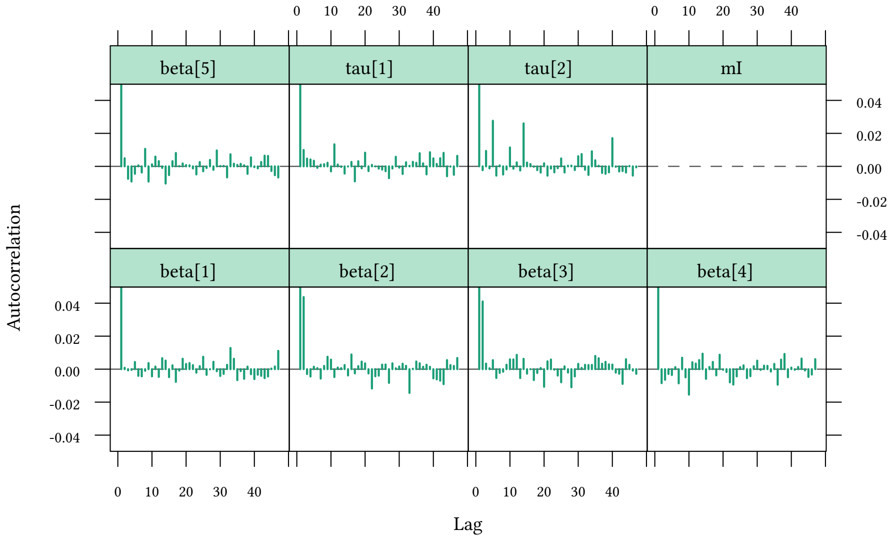
```
select.data<-list(y=testscr-mean(testscr),
                  x1=elpct-mean(elpct),x2=str-mean(str),
                  priBeta=priBeta,priTau=priTau)
select.jags<-run.jags(select.model,data=select.data,
     monitor=c("beta","tau","mI"),inits=genInit(4))
```

Bayesian Model Comparison: Result of the 1st model

```
JAGS model summary statistics from 40000 samples (chains = 4; adapt+burnin = 5000):

          Lower95      Median  Upper95          Mean           SD  Mode
beta[1]   -1.4301  0.00056115   1.3592    -0.0024293      0.70784    --
beta[2]  -0.72793    -0.64938  -0.5735      -0.64963     0.039362    --
```

```
beta[3]   -1.8383    -1.1016  -0.34679    -1.1012    0.38058   --
beta[4]    -192.5    -0.4787    196.85   -0.33221     99.941   --
beta[5]    -197.5   -0.36493    192.78   -0.17792     100.17   --
tau[1]  0.0041393  0.0047697 0.0054392  0.0047783 0.00033204   --
tau[2]          0 1.4045e-29   0.27308    0.99347     9.8156   --
mI              0          0         0          0          0    0

             MCerr MC%ofSD SSeff      AC.10    psrf
beta[1]  0.0035521     0.5 39710   0.001459       1
beta[2] 0.00020561     0.5 36651 -0.0049776       1
beta[3]  0.0019746     0.5 37150  0.0058367       1
beta[4]     0.4879     0.5 41959  0.0043681 0.99998
beta[5]    0.49703     0.5 40616  0.0058656 0.99999
tau[1]  1.6907e-06     0.5 38570   0.013324 0.99999
tau[2]     0.05216     0.5 35413 -0.0021811  1.0137
mI              --      --    --         --      --

Total time taken: 6.6 seconds
```

```
acfplot(as.mcmc(select.jags),aspect="fill")
```



Apparently the indicator *mI* does not mix too well.

Bayesian Model Comparison: Restrict the model

Force the model to estimate the two cases separately (this can be done with the complete model, just set `mI=0`):

```
select0.data<-list(y=testscr-mean(testscr),
                   x1=elpct-mean(elpct),x2=str-mean(str),
                   priBeta=priBeta,priTau=priTau,mI=0)
select0.jags<-run.jags(select.model,data=select0.data,
monitor=c("beta","tau"),inits=genInit(4))
```

```
select0.jags


JAGS model summary statistics from 40000 samples (chains = 4; adapt+burnin = 5000):

          Lower95     Median   Upper95       Mean         SD Mode
beta[1]   -1.4186  0.0035783    1.3566 0.00075669    0.70749   --
beta[2]  -0.72803   -0.65037  -0.57322   -0.65036   0.039496   --
beta[3]   -1.8534    -1.1017  -0.36435    -1.1012    0.37932   --
beta[4]   -194.63     -1.046    196.91   -0.30985     99.673   --
beta[5]   -201.79    0.55934    193.51    0.26679     100.46   --
tau[1]  0.0041457  0.0047721 0.0054357   0.004779 0.00033008   --
tau[2]          0 4.0602e-29     0.338     1.0039     9.5414   --

             MCerr MC%ofSD SSeff       AC.10    psrf
beta[1]   0.003504     0.5 40768   0.0095198  1.0001
beta[2] 0.00020404     0.5 37467  -0.0048292       1
beta[3]  0.0019639     0.5 37306   0.0061509       1
beta[4]    0.49767     0.5 40112  -0.0051776  1.0001
beta[5]    0.50408     0.5 39715 -0.00065203 0.99997
tau[1]  1.6512e-06     0.5 39962  -0.0074455 0.99997
tau[2]    0.047707     0.5 40000   0.0017028  1.0063

Total time taken: 5.8 seconds
```

Now set *mI=1*:

```
select1.data<-list(y=testscr-mean(testscr),
                   x1=elpct-mean(elpct),x2=str-mean(str),
                   priBeta=priBeta,priTau=priTau,mI=1)
select1.jags<-run.jags(select.model,data=select1.data,
monitor=c("beta","tau"),inits=genInit(4))
```

```
select1.jags


JAGS model summary statistics from 40000 samples (chains = 4; adapt+burnin = 5000):

          Lower95     Median   Upper95      Mean         SD Mode
beta[1]   -194.02    0.49833    198.37   0.10098     100.05   --
beta[2]   -197.28    -1.6256    194.03   -1.4997     100.25   --
beta[3]   -196.97  -0.051012    193.68  -0.26462     99.619   --
beta[4]    -1.391 -0.0073152    1.4125  -0.00229    0.71152   --
beta[5]  -0.75145   -0.67094   -0.5971  -0.67106   0.039257   --
tau[1]          0 4.0602e-29     0.338    1.0039     9.5414   --
```

```
tau[2]  0.0040686  0.0046888 0.0053396 0.0046959 0.00032399   --

             MCerr MC%ofSD SSeff       AC.10    psrf
beta[1]    0.49534     0.5 40794   0.0095587  1.0001
beta[2]    0.49884     0.5 40390   -0.004714       1
beta[3]     0.4981     0.5 40000   0.0064915 0.99999
beta[4]   0.003551     0.5 40149  -0.0052715  1.0001
beta[5] 0.00019701     0.5 39704 -0.00080976 0.99997
tau[1]    0.047707     0.5 40000   0.0017028  1.0063
tau[2]  1.6212e-06     0.5 39941  -0.0068089 0.99997

Total time taken: 5.9 seconds
```

Bayesian Model Comparison: Pseudopriors

```
beta0.pri<-select0.jags$summary$statistics[1:3,c("Mean","SD")]
tau0.pri<-select0.jags$summary$statistics[6,c("Mean","SD")]
beta1.pri<-select1.jags$summary$statistics[4:5,c("Mean","SD")]
tau1.pri<-select1.jags$summary$statistics[7,c("Mean","SD")]
(priBetaP<-rbind(beta0.pri,beta1.pri))

                   Mean           SD
beta[1]  0.000756685519 0.7074931872
beta[2] -0.650355923225 0.0394956681
beta[3] -1.101214551765 0.3793161321
beta[4] -0.002289955175 0.7115205788
beta[5] -0.671057176575 0.0392566271

(priTauP<-rbind(tau0.pri,tau1.pri))

                  Mean             SD
tau0.pri 0.00477901491 0.000330076635
tau1.pri 0.00469593902 0.000323994639
```

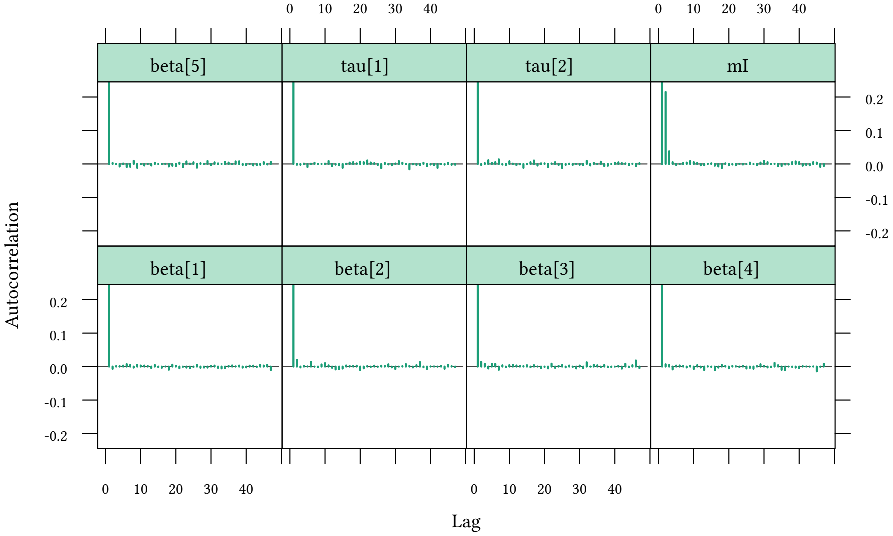
Estimate with Pseudopriors

```
selectP.data<-list(y=testscr-mean(testscr),
                   x1=elpct-mean(elpct),x2=str-mean(str),
                   priBeta=priBetaP,priTau=priTauP)
selectPseudoprior.jags<-run.jags(select.model,data=selectP.data,
   monitor=c("beta","tau","mI"),inits=genInit(4))
```

```
selectPseudoprior.jags


JAGS model summary statistics from 40000 samples (chains = 4; adapt+burnin = 5000):

          Lower95     Median   Upper95       Mean         SD Mode
beta[1]  -0.97842  0.0025589   0.98714  0.0013213    0.50438   --
```

```
beta[2]  -0.70519   -0.65002   -0.5953   -0.65011   0.028051   --
beta[3]   -1.6311    -1.0993  -0.56962    -1.0997    0.27063   --
beta[4]   -1.3888 -0.0085613    1.3797 -0.0075432    0.70524   --
beta[5]  -0.74814   -0.67116  -0.59525   -0.67126   0.039147   --
tau[1]  0.0043113  0.0047839 0.0052412  0.0047876  0.0002363   --
tau[2]  0.0040696  0.0046924  0.005324  0.0046979 0.00032108   --
mI              0          0         0    0.02205    0.14685    0

             MCerr MC%ofSD SSeff      AC.10    psrf
beta[1]  0.0025219     0.5 40000   0.003068  1.0001
beta[2] 0.00014233     0.5 38841  0.0033865  1.0001
beta[3]  0.0013868     0.5 38082  0.0048813       1
beta[4]  0.0035464     0.5 39546  0.0039226  1.0001
beta[5] 0.00019574     0.5 40000 -0.0046747 0.99998
tau[1]  1.1815e-06     0.5 40000   0.009035 0.99998
tau[2]  1.6082e-06     0.5 39863  0.0011391       1
mI      0.00092482     0.6 25213  0.0029701  1.0001

Total time taken: 6 seconds
```

```
acfplot(as.mcmc(selectPseudoprior.jags),aspect="fill")
```


beta[5]
tau[1]
tau[2]
mI
beta[1]
beta[2]
beta[3]
beta[4]
Autocorrelation
Lag