Empirical Methods—MW24.1

Call:
\texttt{lm(formula = testscr ~ str)}

Residuals:
\begin{tabular}{llllll}
  Min & 1Q & Median & 3Q & Max \\
-47.727 & -14.251 & 0.483 & 12.822 & 48.540 \\
\end{tabular}

Coefficients:
\begin{tabular}{llllllll}
  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 & *** \\
\end{tabular}

---
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

1 Review of Probability and Statistics

1.1 What is the purpose of economics

1.2 Econometrics uses data to measure causal relationships

1.3 Example

1.4 Populations and Samples

1.4.1 Population

1.4.2 Random variables and distributions

1.4.3 Sample

1.5 Estimations

1.5.1 Desirable properties of estimators

1.5.2 Characteristics of sampling distributions

1.5.3 Why should we use \( \bar{Y} \) to estimate \( \mu_Y \)?

1.6 Literature

1.A Examples for the lecture

1.B Exercises

2 Review of frequentist inference

2.1 Testing hypotheses

2.1.1 General idea

2.1.2 Calculating the p-value with the help of an estimated \( \hat{\sigma} \)

2.1.3 Relation between p-value and the level of significance

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

2.2 Confidence intervals

2.3 Maximum likelihood method

2.3.1 Motivation

2.3.2 Likelihood function

2.3.3 Log-likelihood Function

2.4 Method of moments

2.5 Literature

2.A Examples for the lecture

2.B Exercises
## 3 Review of Linear Regression

3.1 OLS assumptions .................................................. 46
3.2 Hypotheses tests and confidence intervals ................. 47
3.3 Continuous and nominal variables ......................... 47
3.4 Non-linear regression functions ....................... 51
   3.4.1 Logarithmic Models .................................. 53
   3.4.2 Logarithmic Models: linear-log .................. 54
   3.4.3 Logarithmic Models: log-linear .................. 56
   3.4.4 Comparison of the three logarithmic models .... 59
3.5 Summary .............................................................. 59
3.6 Literature .............................................................. 60
3.A Examples for the lecture ..................................... 60
3.B Exercises ............................................................. 61

## 4 Review of Models with Multiple Regressors

4.1 Assumptions for the multiple regression model ........ 68
4.2 Specification errors ................................................. 68
   4.2.1 Examples ................................................ 69
   4.2.2 Specification errors — generalization ............ 70
4.3 Hypothesis tests ..................................................... 71
   4.3.1 Simple Hypotheses .................................. 71
   4.3.2 Joint Hypotheses .................................... 72
   4.3.3 F statistic for two restrictions ................. 74
   4.3.4 More than two restrictions .................... 74
   4.3.5 Specials cases: ....................................... 77
4.4 Interactions .......................................................... 78
   4.4.1 Interactions between binary variables .......... 79
4.5 Literature .............................................................. 80
4.A Examples for the lecture ..................................... 80
4.B Exercises ............................................................. 82

## 5 Bootstrap

5.1 Motivation ............................................................ 89
5.2 Approximate permutation test ................................ 96
5.3 Parameters, distributions and the plug-in principle .... 98
   5.3.1 The bootstrap algorithm for standard errors ... 99
   5.3.2 Sampling multivariate data ....................... 100
5.4 The parametric bootstrap ...................................... 103
5.5 Literature .............................................................. 105
5.A Examples for the lecture ..................................... 106
5.B Exercises ............................................................. 106
## Bayesian methods

- **6.1 Motivation** .............................................................................. 107
- **6.2 The intention of the researcher — p-hacking** ......................... 111
- **6.3 Probabilities** ....................................................................... 113
- **6.4 Decision making** .................................................................. 115
- **6.5 Prior information** ................................................................. 116
- **6.6 Objectivity and subjectivity** .................................................. 116
- **6.7 Issues** .................................................................................. 117
- **6.8 Literature** .............................................................................. 118
- **6.A Examples for the lecture** ....................................................... 118
- **6.B Exercises** ............................................................................. 118

## Bayes in practice

- **7.1 Example: The distribution of the population mean** ............... 120
- **7.2 Conjugate Priors** ................................................................. 122
- **7.3 An alternative: MCMC** .......................................................... 125
- **7.4 Accumulating evidence** ......................................................... 127
- **7.5 Priors** .................................................................................. 129
- **7.6 Finding posteriors** ................................................................. 129
  - **7.6.1 Overview** ...................................................................... 129
  - **7.6.2 Rejection sampling** ......................................................... 129
  - **7.6.3 Metropolis-Hastings** ...................................................... 130
  - **7.6.4 Gibbs sampling** .............................................................. 131
- **7.7 Technical issues** .................................................................... 132
  - **7.7.1 Introduction** .................................................................. 132
  - **7.7.2 Demeaning** .................................................................. 134
  - **7.7.3 Autocorrelation** .............................................................. 140
  - **7.7.4 Convergence and mixing of chains** ................................. 142
  - **7.7.5 A better vague prior for \( \tau \)** ....................................... 145
- **7.8 Literature** .............................................................................. 146
- **7.A Examples for the lecture** ....................................................... 147
- **7.B Exercises** ............................................................................. 147

## Binary choice

- **8.1 Example** ............................................................................... 149
  - **8.1.1 A binary dependent variable** ............................................ 149
  - **8.1.2 Try to estimate the relationship with OLS** ....................... 151
  - **8.1.3 A problem with OLS** ...................................................... 152
- **8.2 Families and links** ................................................................. 156
  - **8.2.1 A list of common families and link functions** ................... 157
- **8.3 Marginal effects** ................................................................. 157
  - **8.3.1 Marginal effects with OLS** ............................................ 157
### Contents

8.3.2 Marginal effects with binary choice models ........................................... 158
8.3.3 The marginal effect is not constant ...................................................... 159
8.3.4 The normal link function ................................................................. 161
8.3.5 Marginal effects with the normal link function ................................. 163
8.3.6 Marginal effects and interactions ....................................................... 165
8.4 Odds ratios in the logit model ............................................................... 166
8.4.1 Odds ratios with interactions ............................................................ 167
8.4.2 Confidence intervals ............................................................................ 167
8.5 Literature .................................................................................................... 168
8.A Examples for the lecture ........................................................................... 168
8.B Exercises ................................................................................................... 169

9 More on discrete choice .............................................................................. 170

9.1 Logistic regression: ................................................................................... 170
9.1.1 The Maximum Likelihood method .................................................... 170
9.1.2 Misspecification in OLS and in Binary choice .................................. 172
9.1.3 Confidence intervals ........................................................................... 173
9.1.4 Bayesian logistic regression ............................................................... 174
9.2 Count data .................................................................................................. 176
9.2.1 Model ................................................................................................... 176
9.2.2 Illustration — the Fair data ............................................................... 177
9.2.3 Poisson versus negative binomial ....................................................... 179
9.2.4 Bayesian count data ............................................................................ 183
9.3 Multinomial (polytomous) logit ............................................................... 186
9.3.1 Motivation and background ............................................................... 186
9.3.2 Example ............................................................................................... 189
9.3.3 Bayesian multinomial ......................................................................... 192
9.4 Literature .................................................................................................... 193
9.A Examples for the lecture ........................................................................... 193
9.B Exercises ................................................................................................... 195
9.C Ordered probit .......................................................................................... 195
9.C.1 Model ................................................................................................... 195
9.C.2 Illustration — the Fair data ............................................................... 197
9.C.3 Illustration II — a simulated dataset .................................................. 202

10 Mixed effects models ................................................................................. 204

10.1 Terminology ............................................................................................... 204
10.2 Examples .................................................................................................... 208
10.2.1 Initial commands .............................................................................. 208
10.2.2 A small example ............................................................................... 208
10.3 5 different methods - 5 different results .............................................. 217
10.3.1 A larger example ............................................................................... 217
10.3.2 Pooled OLS ............................................. 217
10.3.3 Between OLS ........................................... 218
10.3.4 Fixed effects .......................................... 219
10.3.5 Mixed effects ......................................... 219
10.3.6 More general mixed effects ......................... 220
10.3.7 Bayes and mixed effects ............................. 220
10.4 Estimation results ......................................... 221
10.4.1 Residuals ............................................... 221
10.4.2 Estimated standard errors ......................... 222
10.4.3 Testing random effects .............................. 223
10.4.4 Confidence intervals for fixed effects (in a ME model) ............................................. 225
10.5 Literature .................................................. 227
10.A Examples for the lecture .............................. 228
10.B Exercises .................................................. 230

11 Instrumental variables ...................................... 231
11.1 Instruments ............................................... 231
11.2 Demand and Supply ...................................... 237
11.3 The general Instrumental Variables Model .......... 246
11.4 An example for (potential) underidentification .... 247
11.5 Weak instruments ........................................ 253
11.6 Discrete endogeneous variables ....................... 254
11.7 Literature .................................................. 258
11.A Examples for the lecture .............................. 258
11.B Exercises .................................................. 260

12 Model Specification and Model Comparison .................. 261
12.1 Scaling coefficients ...................................... 262
12.2 Which coefficients ........................................ 263
12.3 Measure $R^2$ ............................................. 265
12.4 Nevertheless: Measure contribution to $R^2$ ......... 266
12.5 Analysis of Variance ..................................... 270
  12.5.1 t-test versus F test .................................. 272
12.6 Information criteria ....................................... 273
12.7 Hypothesis tests for individual coefficients .......... 279
12.8 Comparing GLM models .................................. 280
  12.8.1 Goodness of fit ....................................... 280
  12.8.2 Likelihood ratio index and AIC .................. 280
  12.8.3 An example: Labour market participation ....... 281
  12.8.4 z-test versus $\chi^2$ test ............................ 283
12.9 Bayesian Model Comparison ........................... 283
  12.9.1 Preliminaries ......................................... 283
12.9.2 Odds in the simple model .................................................. 284
12.9.3 Compare different models ................................................. 284
12.9.4 Result of the Bayesian Model Comparison ......................... 287
12.9.5 Model uncertainty ............................................................. 288
12.9.6 Bayes factors ................................................................. 288
12.10 Literature ........................................................................ 290
12.A Examples for the lecture .................................................... 290
12.B Exercises ........................................................................ 292
12.C GLM - A larger model ....................................................... 293
12.D Technical details for model comparison ............................... 296

Homepage: https://www.kirchkamp.de/master/em.html

Schedule: Lecture: Thu, 12-14 (not on “dies academicus”, 24.10.)

Exercise: Tue, 12-14
          Fri, 10-12

Midterm (1/3): Thu. 28.11., 18:00, HS 1
              Resit: 12.12., 18:00, HS 3
              Register 10 days before the midterm exam through Friedolin !!!

Final exam (2/3): 10.2., 16:00, HS 1
                  Register for the exam through Friedolin!

Literature:


Empirical methods — Where are we?
Statistics:

• Descriptive statistics

• Inferential statistics

  –
  – Biometrics
  – Psychometrics
  – Econometrics / Empirical Methods

  * Frequentist methods in Econometrics
* Bayesian methods in Econometrics

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

1 Review of Probability and Statistics

**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 if the ECB raises interest rates by 1%?

1.1 What is the purpose of economics

- Developing theories
- Testing theories
- Using theories for prediction

Desirable properties of economic theories

- Congruence with reality
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 become smaller?
  - How much more do you earn if you decide to 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?

- ↑ hard to do

- 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 $\leftrightarrow$ 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 Example

- 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
  - \(testscr\) = 5\(^{th}\)-grade test score (Stanford-9 achievement test)

For our examples we use the statistical software \(\mathbf{R}\). 
\(\mathbf{R}\) organises data and methods in libraries. To use them, you need the command `library`:

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

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

```r
data(Caschool)
```

We can now use components of this data set:

```r
head(Caschool)
```

<table>
<thead>
<tr>
<th>distcod</th>
<th>county</th>
<th>district</th>
<th>grspan</th>
<th>enrltot</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>75119 Alameda</td>
<td>Sunol Glen Unified</td>
<td>KK-08</td>
<td>195</td>
</tr>
<tr>
<td>2</td>
<td>61499 Butte</td>
<td>Manzanita Elementary</td>
<td>KK-08</td>
<td>240</td>
</tr>
<tr>
<td>3</td>
<td>61549 Butte</td>
<td>Thermalito Union Elementary</td>
<td>KK-08</td>
<td>1550</td>
</tr>
<tr>
<td>4</td>
<td>61457 Butte</td>
<td>Golden Feather Union Elementary</td>
<td>KK-08</td>
<td>243</td>
</tr>
<tr>
<td>5</td>
<td>61523 Butte</td>
<td>Palermo Union Elementary</td>
<td>KK-08</td>
<td>1335</td>
</tr>
<tr>
<td>6</td>
<td>62042 Fresno</td>
<td>Burrel Union Elementary</td>
<td>KK-08</td>
<td>137</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>teachers</th>
<th>calwpct</th>
<th>mealpct</th>
<th>computer</th>
<th>testscr</th>
<th>compstu</th>
<th>expnstu</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.90</td>
<td>0.5102</td>
<td>2.0408</td>
<td>67</td>
<td>690.80</td>
<td>0.3435898 6384.911</td>
</tr>
<tr>
<td>2</td>
<td>11.15</td>
<td>15.4167</td>
<td>47.9167</td>
<td>101</td>
<td>661.20</td>
<td>0.4208333 5099.381</td>
</tr>
<tr>
<td>3</td>
<td>82.90</td>
<td>55.0323</td>
<td>76.3226</td>
<td>169</td>
<td>643.60</td>
<td>0.1090323 5501.955</td>
</tr>
<tr>
<td>4</td>
<td>14.00</td>
<td>36.4754</td>
<td>77.0492</td>
<td>85</td>
<td>647.70</td>
<td>0.3497942 7101.831</td>
</tr>
<tr>
<td>5</td>
<td>71.50</td>
<td>33.1086</td>
<td>78.4270</td>
<td>171</td>
<td>640.85</td>
<td>0.1280899 5235.988</td>
</tr>
<tr>
<td>6</td>
<td>6.40</td>
<td>12.3188</td>
<td>86.9565</td>
<td>25</td>
<td>605.55</td>
<td>0.1824818 5580.147</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>str</th>
<th>avginc</th>
<th>elpct</th>
<th>readscr</th>
<th>mathscr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17.88991</td>
<td>22.690001</td>
<td>0.000000</td>
<td>691.6</td>
</tr>
<tr>
<td>2</td>
<td>21.52466</td>
<td>9.824000</td>
<td>4.583333</td>
<td>660.5</td>
</tr>
<tr>
<td>3</td>
<td>18.69723</td>
<td>8.978000</td>
<td>30.000002</td>
<td>636.3</td>
</tr>
<tr>
<td>4</td>
<td>17.35714</td>
<td>8.978000</td>
<td>0.000000</td>
<td>651.9</td>
</tr>
<tr>
<td>5</td>
<td>18.67133</td>
<td>9.080333</td>
<td>13.857677</td>
<td>641.8</td>
</tr>
<tr>
<td>6</td>
<td>21.40625</td>
<td>10.415000</td>
<td>12.408759</td>
<td>605.7</td>
</tr>
</tbody>
</table>
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.

```r
attach(Caschool)
```

```r
hist(str)
```

![Histogram of str](image)

```r
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.
We first calculate a categorical variable `large`. Then we describe a relationship `testscr ~ 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.
1.4 Populations and Samples

1.4.1 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.

1.4.2 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 $Pr(x) = Pr(X = x)$
  - Cumulative distribution function $F(x) = P(X < x)$
  - Probability density function $f(x) = \frac{dF(x)}{dx}$

Properties of random variables

- Expected value $E(X)$, $\mu_X$, (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 $\theta_1, \theta_2, \ldots$ are parameters of the distribution. E.g. the normal distribution is characterised by $\mu$ and $\sigma^2$:

\[ X \sim N(\mu, \sigma^2) \]
1.4.3 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 $i = 1, \ldots, n$. 
1.5 Estimations

- Populations are characterised by parameters (\( \theta \))
- 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 \( \theta \)) 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 \( Y_1 \ldots Y_n \)
- \( \ldots \)

Which criteria should we use to pick an estimator?

1.5.1 Desirable properties of estimators

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

\[
\forall \theta : E(\hat{\theta}(X_1, \ldots, X_n)) = \theta
\]

**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

\[
\forall \epsilon > 0 : \lim_{n \to \infty} \Pr(|\hat{\theta}(X_1, \ldots, X_n) - \theta| < \epsilon) = 1
\]
1.5.2 Characteristics of sampling distributions

Expected value of \( \bar{Y} \)
\[ E(\bar{Y}) = \mu_Y, \text{ i.e. } \bar{Y} \text{ is an unbiased estimator of } \mu_Y \]

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 \( \mu_Y \) if \( n \) is large?

Law of large numbers:
\( \bar{Y} \) is a consistent estimator of \( \mu_Y \).
Formally: If \( Y_1, \ldots, Y_n \) i.i.d. and \( \sigma_Y^2 < \infty \), then \( \bar{Y} \) is a consistent estimator of \( \mu_Y \), i.e.
\[ \forall \epsilon > 0 : \lim_{n \to \infty} \Pr(|\bar{Y} - \mu_Y| < \epsilon) = 1 \quad \text{we can also say} \quad \bar{Y} \xrightarrow{p} \mu_Y \]

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

1.5.3 Why should we use \( \bar{Y} \) to estimate \( \mu_Y \)?
- \( \bar{Y} \) is unbiased: \( E(\bar{Y}) = \mu_Y \)
- \( \bar{Y} \) is consistent: \( \bar{Y} \xrightarrow{p} \mu_Y \)
- \( \bar{Y} \) is the “least squares” estimator for \( \mu_Y \)
  \( \bar{Y} \) is the solution of \( \min_x \sum_{i=1}^{n} (Y_i - x)^2 \)
- \( \bar{Y} \) has a smaller variance than all other linear unbiased estimators.
  For any estimator \( \hat{\mu}_Y = \frac{1}{n} \sum_{i=1}^{n} \alpha_i Y_i \) with \( \{\alpha_i\} \) so that \( \hat{\mu}_Y \) is unbiased, it is true that \( \text{var}(\bar{Y}) \leq \text{var}(\hat{\mu}_Y) \)
- However, there are non-linear estimators, too...
1.6 Literature


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?

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

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 $E(X)$ 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.

<table>
<thead>
<tr>
<th>age in years (a)</th>
<th>pocket money/week in $ (pm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
</tr>
</tbody>
</table>

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?

• $X_9X_1 + X_7 - X_9$
• $2X_1 + 2X_7 - 3X_9$
• $\frac{1}{9} \sum_{i=1}^{9} X_i$
• $\sum_{i=1}^{9} X_i$

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 $(X_2 + X_3)/2$
• $\frac{1}{9} \sum_{i=1}^{9} X_i$ is more efficient than $\frac{1}{7} \sum_{i=1}^{7} X_i$
• $X_3$ is more efficient than $\frac{1}{7} \sum_{i=1}^{7} X_i$

2 Review of frequentist inference

2.1 Testing hypotheses

2.1.1 General idea

![Graph of probability density function](image)
Null hypothesis: The population parameter is $\theta_0$. 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 $\theta_0$? — significance test at a given significance level.

- If the population parameter is, indeed, $\theta_0$, how probable is it to draw a sample like ours (or a sample even more adverse to our hypothesis)? — p-value (marginal significance).

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

![Density plot with sample density]

Is 652 the mean `testscr`?

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

<table>
<thead>
<tr>
<th></th>
<th>$H_0$ false</th>
<th>$H_0$ true</th>
</tr>
</thead>
<tbody>
<tr>
<td>fail to</td>
<td>type II error</td>
<td>type I error</td>
</tr>
<tr>
<td>reject $H_0$</td>
<td>false negative</td>
<td>false positive</td>
</tr>
</tbody>
</table>

**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 $Y_1, \ldots, Y_N$, that is at least as adverse to the null hypothesis as our data — given that the null hypothesis is true.

$$ \text{e.g. with } \bar{Y} : \text{p-value} = Pr_{H_0}(|\bar{Y} - \mu_{Y,0}| > |\bar{Y}_{\text{sample}} - \mu_{Y,0}|) $$

($\bar{Y}_{\text{sample}}$ = 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 $\rightarrow$ Central Limit Theorem $\rightarrow$ $\bar{Y}$ follows approximately a normal distribution.

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

- 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 $\hat{\sigma}_Y^2$

$$p\text{-value} = \Pr_{H_0} \left( |\bar{Y} - \mu_{Y,0}| > |\bar{Y}_{\text{sample}} - \mu_{Y,0}| \right)$$

$$= \Pr_{H_0} \left( \frac{|\bar{Y} - \mu_{Y,0}|}{\sigma_Y} > \frac{|\bar{Y}_{\text{sample}} - \mu_{Y,0}|}{\sigma_Y} \right)$$

$$\approx \Pr_{H_0} \left( \frac{|\bar{Y} - \mu_{Y,0}|}{\hat{\sigma}_Y} > \frac{|\bar{Y}_{\text{sample}} - \mu_{Y,0}|}{\hat{\sigma}_Y_{\text{sample}}} \right)$$

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:
  $$t = \frac{\bar{Y}_{\text{sample}} - \mu_{Y,0}}{\hat{\sigma}_Y}$$

- The $p$-value is $p = \Pr \left( |t| > |t_{\text{sample}}| \right) = 2 F(-|t_{\text{sample}}|)$

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 $\mu_Y$ is the interval that contains the true value of $\mu_Y$ 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 $\mu_Y$!)

\[
\bar{y} - \frac{\sigma_{\bar{y}}}{\sqrt{n}} Q\left(\frac{\alpha}{2}\right) \leq \mu_Y \leq \bar{y} + \frac{\sigma_{\bar{y}}}{\sqrt{n}} Q\left(1 - \frac{\alpha}{2}\right)
\]

$H_0 : \bar{Y} = \mu_0$ is rejected, if $\mu_0$ 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 \( \bar{y} = \frac{\sigma_y}{\sqrt{n}} \)

• The interval is

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

Note:

• The confidence interval is based on the random sample \( Y_1, \ldots, Y_n \). Thus, the confidence interval is random itself.

• The parameter \( \mu_Y \) 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

2.3.1 Motivation

Motivation: bids in English auctions  In the previous chapter we discussed desirable properties of estimators. Here we discuss how estimators can be constructed. We will compare two methods:
• Maximum likelihood

• Method of moments

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

To estimate, we assume that all elements of the population can be observed with the same probability. We rule out a *selection bias*.

Sometimes we have a systematic selection bias. One example are bids in an English auction.

In “An experimental analysis of auctions with interdependent valuation”, (Games and Economic Behavior, 2004, Vol. 48/1, pp. 54-85) Kirchkamp and Moldovanu study, how bidders in an English auction learn from the bids of the other participants in the auction.

One problem we face here is that bids are only observed for losers of the auction, not for winners. For winners we only know that their (highest) bid must be higher or equal than the losing bid.

The following graph show the maximum bids of some players. Bids of the losers are shows as +. Bids of winners can usually not be observed. Let us assume we could observe...
these bids. In the graph these hypothetical bids are indicated with ◦.

- Question: Do bidders learn in the English auction from their competitors’ bids?
- Problem: We only observe losers’ bids.
- Solution: Maximum likelihood: We try different bid functions until we find the one with the highest likelihood.


- Consider a random variable $X \sim f(x|\theta)$ with unknown parameter $\theta$.
- 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 $X_i$:

- discrete: $Pr(X_i = x_i|\theta) = \begin{cases} \text{Probability, that } X_i = x_i \text{ if } \\ \theta \text{ was the true parameter} \end{cases}$
- continuous: $f(x_i|\theta) = \begin{cases} \text{Density of } X_i \text{ if } \\ \theta \text{ is the true parameter} \end{cases}$

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 \( \epsilon \). 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.

### 2.3.2 Likelihood function

What is the likelihood, for a given \( \theta \), 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 \( \theta \), to draw a sample \( x_1, x_2, x_3, \ldots, x_n \)?

\[
f(x_1|\theta) \cdot f(x_2|\theta) \cdot f(x_3|\theta) \ldots f(x_n|\theta)
\]

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

\[
L(x_1, \ldots, x_n|\theta) \equiv f(x_1|\theta) \cdot f(x_2|\theta) \cdot f(x_3|\theta) \ldots f(x_n|\theta)
\]

the **likelihood function**.

Two interpretations:

- \( \theta \) is given: \( L(x_1, \ldots, x_n|\theta) \) is the likelihood that \( (x_1, \ldots, x_n) \) realise. Since \( \theta \) 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 \( \theta \) was the true parameter. Since we know \( (x_1, \ldots, x_n) \) we can use this approach to estimate \( \theta \).

### The maximum-likelihood (ML) method

- \( f(X|\theta) \) conditional density (likelihood)

- \( \Theta \) the set of possible parameter values for \( \theta \)

- Definition: \( \hat{\theta} \) is maximum-likelihood (ML) estimator iff \( \forall \theta \in \Theta : \)

\[
L(x_1, \ldots, x_n|\hat{\theta}) \geq L(x_1, \ldots, x_n|\theta)
\]

We use as estimator for \( \theta \) the value \( \hat{\theta} \), for which, given the assumed likelihood function \( f(X|\theta) \), realisations \( x_1, \ldots, x_n \) are most probable.
2.3.3 Log-likelihood Function

So far, we maximise the likelihood function:

\[ L(x_1, \ldots, x_n|\theta) \equiv f(x_1|\theta) \cdots f(x_n|\theta) \]

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

\[
\log L(x_1, \ldots, x_n|\theta) = \log (f(x_1|\theta) \cdots f(x_n|\theta)) = \\
\log f(x_1|\theta) + \cdots + \log f(x_n|\theta)
\]

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

\[
\frac{d}{d\theta} \log L(x_1, \ldots, x_n|\theta) = \frac{f'(x_1|\theta)}{f(x_1|\theta)} + \cdots + \frac{f'(x_n|\theta)}{f(x_n|\theta)}
\]

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

- Discrete choice models
  - Logistic regression (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, ...)
- 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:

\[ X \sim N(\mu, \sigma^2) \]

\[ Pr(X|\mu, \sigma) = \prod dnorm(X_i|\mu, \sigma) \]

\[ LL = \sum \log(dnorm(X_i|\mu, \sigma)) \]

```r
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:

```r
mean(X)
```

[1] 2

```r
sd(X)
```

[1] 1

Note: The ML estimator yields (almost) the same result for \( \mu \), but a different result for \( \sigma \).

Properties of the ML estimator (given some conditions):

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

An alternative method to estimate \( \theta \) is the method of moments:
2.4 Method of moments

Reminder: Moments

\[ n\text{-th non-central moment of } X: \mu'_n = E(X^n) \]
\[ n\text{-th central moment of } X: \mu_n = E((X - \mu'_1)^n) \]

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

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>mean:</td>
<td>\mu'_1 = E(X)</td>
</tr>
<tr>
<td>variance:</td>
<td>\mu_2 = E((X - E(X))^2)</td>
</tr>
<tr>
<td>skewness:</td>
<td>\frac{\mu_3}{(\mu_2)^{3/2}}</td>
</tr>
<tr>
<td>kurtosis:</td>
<td>\frac{\mu_4}{(\mu_2)^2}</td>
</tr>
</tbody>
</table>

We distinguish:

**Moments of a population** (unknown): \( \mu_1, \mu_2, \mu_3, \ldots \)

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

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

- \( f(X|\theta) \) has moments \( \mu_1(\theta), \mu_2(\theta), \mu_3(\theta), \ldots \)

- Method of moments:
  - Determine the sample moments \( m_1, m_2, m_3 \ldots \) (we usually start with the mean \( m_1 \))
  - Equate sample moments with population moments...

\[ \mu_i(\hat{\theta}) = m_i \]

...and solve for \( \hat{\theta} \).
Back to our motivating example:

Using a likelihood estimator one can show that bidders in an English auction use bids of their competitors to adjust their own bids. The allocation under the English auction is more efficient than, e.g., a second price auction (where bids of competitors are not visible).


2.5 Literature


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.

- $qt(10, 4) \times 0.95$
• \( \text{qnorm}(10) \ast 5 \)
• \( \text{qt}(0.95, 3) \ast 10 \)
• \( \text{qnorm}(0.95) \ast 10 \)

**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.
• \( 10 + \text{qnorm}(0.05) \)
• \( 10 + 5 \ast \text{qnorm}(0.05) \)
• \( 10 + 5 \ast \text{qnorm}(0.025) \)
• \( 10 - \text{qnorm}(0.975) \)

**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 \( \theta \)?

**Example 4:** The random variable \( X \) follows a distribution \( X_\theta \) with expected value \( E(X) = 1/\theta \) and variance \( \theta^2 \). We have \( \theta > 0 \). The variable \( x \) contains your sample. Use R to calculate the method of moments estimator for \( \theta \) 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 \( \theta \)?

**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 \(2 \cdot X_1 + X_7 - X_6 - a\)
   - \(-a + \frac{2}{9} \sum_i X_i\) 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 \(\ln \text{lninc}\). 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.
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:

```r
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:

```r
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:

```r
xx<-replicate(1000,mean(sample(wage,1000)))
qqnorm(xx)
qqline(xx)
```
Normal Q-Q Plot

```
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:

<table>
<thead>
<tr>
<th></th>
<th>0.001</th>
<th>0.0025</th>
<th>0.005</th>
<th>0.01</th>
<th>0.025</th>
<th>0.05</th>
<th>0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q_{0.001}^N$</td>
<td>-3.090</td>
<td>-2.807</td>
<td>-2.576</td>
<td>-2.326</td>
<td>-1.960</td>
<td>-1.645</td>
<td>-1.282</td>
</tr>
<tr>
<td>$Q_{0.0025}^N$</td>
<td>-3.579</td>
<td>-3.174</td>
<td>-2.861</td>
<td>-2.539</td>
<td>-2.093</td>
<td>-1.729</td>
<td>-1.328</td>
</tr>
<tr>
<td>$Q_{0.005}^N$</td>
<td>-3.552</td>
<td>-3.153</td>
<td>-2.845</td>
<td>-2.528</td>
<td>-2.086</td>
<td>-1.725</td>
<td>-1.325</td>
</tr>
<tr>
<td>$Q_{0.01}^N$</td>
<td>-3.527</td>
<td>-3.135</td>
<td>-2.831</td>
<td>-2.518</td>
<td>-2.080</td>
<td>-1.721</td>
<td>-1.323</td>
</tr>
</tbody>
</table>

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

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

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 $\theta$.

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 $\theta$ with the method of moments.

2. What is your estimate for $\theta$ if your sample is $X = \{1, 19, 4\}$?

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

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

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 $\sigma^2 = \frac{118}{1} a^2$.

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

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

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

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

1. What is the ML estimator for $\theta$ if your sample is $\{2, 3\}$.

2. What is the method of moment estimator for $\theta$ (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

\[
Pr(X = x) = \begin{cases} 
\frac{1}{4} & \text{if } x = 0 \\
\frac{1}{2} & \text{if } x = 0 + 1 \\
\frac{1}{4} & \text{if } x = 0 + 2 \\
0 & \text{otherwise}
\end{cases}
\]

1. What is the ML estimator for \(\theta\), 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 \(\theta\)?

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

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.

\(\rightarrow\) Estimation, hypothesis test, confidence interval

\[
testscr = \beta_1 str + \beta_0
\]

- \(\beta_0\) and \(\beta_1\) are parameters of the population
- we do not know them — hence, we have to estimate them (like \(\mu\))

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

data(Caschool)
attach(Caschool)
scatterplot(testscr-str)
$Y_i = \beta_1 X_i + \beta_0 + u_i \quad i = 1, \ldots, n$

- $Y$ dependent variable
- $X$ independent variable
- $\beta_1$ slope
- $\beta_0$ axis intercept
- $u$ error term (other factors that impact $Y$)

**How can we estimate $\beta_0$ and $\beta_1$?**

- 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.

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

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

- Is \( E(\hat{\beta}_1) = \beta_1 \)? (OLS is unbiased)
- Is \( \text{var}(\hat{\beta}_1) \) small?
- How do we test hypotheses? (e.g. \( \beta_1 = 0 \))
- How do we calculate a confidence interval for \( \beta_0 \) and \( \beta_1 \)?

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

Asymptotic consistency
Assuming 1-3, \( \hat{\beta} \overset{p}{\rightarrow} N(\beta, \Sigma_\beta) \)

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

Efficiency of OLS-II
Assuming 1-5, \( \hat{\beta}_1 \) has the smallest variance of all consistent estimators, if \( n \rightarrow \infty \) (regardless of whether the estimators are linear or non-linear)
3.2 Hypotheses tests and confidence intervals

- two-sided test:
  \[ H_0 : \beta_1 = \beta_{1,0} \text{ versus } H_1 : \beta_1 \neq \beta_{1,0} . \]

- one-sided test:
  \[ H_0 : \beta_1 = \beta_{1,0} \text{ versus } H_1 : \beta_1 > \beta_{1,0} \]
  \[ H_0 : \beta_1 = \beta_{1,0} \text{ versus } H_1 : \beta_1 < \beta_{1,0} \]

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 \( \beta \) is the interval that contains the true value of \( \beta \) in 95% percent of all repeated samples.

\[ \hat{\beta} + \sigma_{\hat{\beta}} \cdot Q\left(\frac{a}{2}\right) \quad \hat{\beta} \quad \hat{\beta} + \sigma_{\hat{\beta}} \cdot Q\left(1 - \frac{a}{2}\right) \]

\( H_0 : \bar{Y} = \beta_0 \) is rejected, if \( \beta_0 \) 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

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

```
lm(testscr ~ str)
Call:
lm(formula = testscr ~ str)
Coefficients:
(Intercept) str
    698.93    -2.28
```

\[
\text{testscr} = 698.93 - 2.28 \text{str} + u
\]

In

\[
\text{testsrc} = \beta_0 + \beta_1 \text{str} + u
\]
we used a continuous independent variable \( \text{str} \). But what if we only had binary data for \( \text{str} \)?

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

Now estimate

\[
\text{testsrc} = \beta_0 + \beta_1 \text{large} + u
\]

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.

```r
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

```r
confint(est)
```

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

```r
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

```r
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 408.111  < 2e-16 ***
avginc      1.8785      0.0905  20.761  < 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)
```

![Graph showing the relationship between test scores and district average income](image)
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 \texttt{avginc} and \texttt{testscr} or we leave the precise form open, requiring only a “smooth” relationship, like the following:

\textbf{A semiparametric fit of the relation between \texttt{avginc} and \texttt{testscr}:}

```r
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
• $Y_i = \beta_0 + \beta_1 \cdot \log X_i + u_i$  \hspace{1cm} \text{linear-log}

• $\log Y_i = \beta_0 + \beta_1 \cdot X_i + u_i$  \hspace{1cm} \text{log-linear}

• $\log Y_i = \beta_0 + \beta_1 \cdot \log X_i + u_i$  \hspace{1cm} \text{log-log}

### 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% ($\Delta X_i = 0.01 \cdot X_i$) ...

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

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

Since it is easier to draw a line of a function if the \textit{avginc}-values are ordered, we calculate an ordering of \textit{avginc}:
or <- order(avginc)

(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

\[ \log Y_i = \beta_0 + \beta_1 \cdot X_i + u_i \rightarrow Y_i = e^{\beta_0 + \beta_1 \cdot X_i + u_i} \]

marginal effects:

\[ \frac{\partial \log Y_i}{\partial X_i} = \beta_1 \]

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

\[ \frac{\Delta \log Y_i}{\Delta X_i} \approx \beta_1 \rightarrow \Delta \log Y_i \approx \beta_1 \Delta X_i \]

\[ \frac{1}{Y_i} \approx \frac{\Delta \log Y_i}{\Delta Y_i} \rightarrow \frac{\Delta Y_i}{Y_i} \approx \Delta \log Y_i \rightarrow \frac{\Delta Y_i}{Y_i} \approx \Delta \log Y_i \approx \beta_1 \cdot \Delta X_i \]

A change of \( X_i \) by one unit translates into a relative change of \( Y_i \) by the share \( \beta_1 \)

```r
(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))
```
$\frac{\partial Y_i}{\partial X_i} = \beta_1 \cdot e^{\beta_0 + \beta_1 \cdot X_i}$

$X <- c(10,40,60)$

(coef(estLL)[2]*exp(predict(estLL,newdata=list(avginc=X))))

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.831772</td>
<td>1.994924</td>
<td>2.111688</td>
</tr>
</tbody>
</table>

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

marginal effect:

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

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

$\beta_1$ is the elasticity of $Y_i$ with respect to $X_i$. 
$$\frac{\partial Y_i}{\partial X_i} = e^{\beta_0} \cdot \beta_1 X_i^{\beta_1 - 1} \approx \beta_1 \frac{Y_i}{X_i}$$

\[ X \leftarrow c(10, 40, 60) \]

\[ \text{maLLL} \leftarrow \exp(\text{coef}(\text{estLLL})[1]) \times \text{coef}(\text{estLLL})[2] \times X^{-(\text{coef}(\text{estLLL})[2] - 1)} \]

[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 $\beta$ 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.

$\rightarrow$ We need economic theory to motivate one of the four specifications.

<table>
<thead>
<tr>
<th></th>
<th>10.00</th>
<th>40.00</th>
<th>60.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>linLog</td>
<td>3.64</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>logLin</td>
<td>1.83</td>
<td>1.99</td>
</tr>
<tr>
<td></td>
<td>logLog</td>
<td>3.56</td>
<td>0.96</td>
</tr>
</tbody>
</table>

3.5 Summary

Having started from these 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 $u_i \sim N(0, \sigma)$

we now know how to

- estimate (sample distribution of $\hat{\beta}$)
- test hypotheses
  - $\hat{\beta}$ is t-distributed if $u_i \sim N(0, \sigma)$
  - $\hat{\beta}$ is approximately normally distributed (if $n$ is large)
  This allows us to calculate the $p$-value.
- calculate confidence intervals
3.6 Literature


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 $\theta$ is $\text{CI}_99\% = [-1, 2]$. 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 $\beta_1$ 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 ($cmrte$)? 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 ($cmrte$)? 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 $\beta_1$ 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=TRUE, 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

\[ \text{testscr} = \beta_1 \text{str} + \beta_0 + u \]

- 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 e1pct (english learner percentage))

```r
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 \( \text{elpct} \) the estimated relationships are very different.

→ extend the regression model

\[
\text{testsrc} = \beta_1 \text{str} + \beta_2 \text{elpct} + \beta_0 + u
\]
generally:

\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u
\]
for each observation:

\[
\begin{align*}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + u_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + u_2 \\
y_3 &= \beta_0 + \beta_1 x_{31} + \beta_2 x_{32} + \cdots + \beta_k x_{3k} + u_3 \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + u_n
\end{align*}
\]

\[
\text{lm(testscr} \sim \text{str } + \text{elpct})
\]

Call:

\[
\text{lm(formula = testscr} \sim \text{str } + \text{elpct})
\]

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. \( \text{var}(u|X = x) \) is constant, \( u \) is homoscedastic

6. \( u \) is normally distributed \( u \sim N(0, \sigma^2) \)

\[ \text{→ under assumptions 1–4 the OLS estimator } \hat{\beta} = (X'X)^{-1}X'y \text{ is unbiased and consistent.} \]

\[ \text{→ For large samples } \hat{\beta} \text{ is jointly normally distributed.} \]

4.2 Specification errors

What can happen if we forget to include a variable into our model?

E.g.: \( \text{tests}rc = \beta_1 \text{str} + \beta_0 + u \)

- \( \text{tests}rc \) test score
- \( \text{str} \) student / teacher ratio

<table>
<thead>
<tr>
<th>What else could have an influence on ( \text{tests}rc )?</th>
<th>corr. with regressor</th>
<th>influence on dep. var.</th>
</tr>
</thead>
<tbody>
<tr>
<td>percent of English learners</td>
<td>( \text{str} )</td>
<td>( \text{tests}rc )</td>
</tr>
<tr>
<td>time of day of the test</td>
<td>( x )</td>
<td></td>
</tr>
<tr>
<td>parking lot space per student</td>
<td>( x )</td>
<td></td>
</tr>
</tbody>
</table>

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

- this variable is correlated with the regressor
- \( \text{and} \) this variable is has an influence on the dependent variable

our estimation for \( \beta \) is biased (omitted variable bias).
4.2.1 Examples:

- Classical music $\rightarrow$ intelligence of children (Rauscher, Shaw, Ky; Nature; 1993)

  (missing variable: income)

- French paradox: Red wine, foie gras $\rightarrow$ less illnesses of the coronary blood vessels (Samuel Black, 1819)

  (missing variable: percentage of fish and sugar in the diet,...)

- Storks in Lower Saxony $\rightarrow$ birth rate

  (missing variable: industrialisation)


### 4.2.2 Specification errors — generalization

Let the true model be

\[ \mathbf{y} = \mathbf{X}_1 \beta_1 + \mathbf{X}_2 \beta_2 + \mathbf{u} \]

what happens if we forget to include \( \mathbf{X}_2 \) into the specification of our model?

\[ \hat{\beta} = \left( \mathbf{X}' \mathbf{X}_1 \right)^{-1} \mathbf{X}' \mathbf{y} \]

\[
\begin{align*}
\mathbf{b}_1 &= \left( \mathbf{X}'_1 \mathbf{X}_1 \right)^{-1} \mathbf{X}'_1 \mathbf{y} \\
&= \left( \mathbf{X}'_1 \mathbf{X}_1 \right)^{-1} \mathbf{X}'_1 (\mathbf{X}_1 \beta_1 + \mathbf{X}_2 \beta_2 + \mathbf{u}) \\
&= \left( \mathbf{X}'_1 \mathbf{X}_1 \right)^{-1} \mathbf{X}'_1 \mathbf{x}_1 \beta_1 + \left( \mathbf{X}'_1 \mathbf{X}_1 \right)^{-1} \mathbf{X}'_1 \mathbf{X}_2 \beta_2 + \left( \mathbf{X}'_1 \mathbf{X}_1 \right)^{-1} \mathbf{X}'_1 \mathbf{u} \\
\mathbb{E}(\mathbf{b}_1) &= \beta_1 + \left( \mathbf{X}'_1 \mathbf{X}_1 \right)^{-1} \mathbf{X}'_1 \mathbf{X}_2 \beta_2
\end{align*}
\]

Hence, \( \mathbb{E}(\mathbf{b}_1) = \beta_1 \) only if

- \( \beta_2 = 0 \)
- or \( \mathbf{X}'_1 \mathbf{X}_2 = 0 \), i.e. \( \mathbf{X}_1 \) and \( \mathbf{X}_2 \) are orthogonal
Specification errors — another example  

The correct 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:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 (\alpha X_1 + \zeta) + u \\
= \beta_0 + (\beta_1 + \beta_2 \alpha) X_1 + \beta_2 \zeta + u \]

- We overestimate \( \beta_1 \) if \( \beta_2 \alpha > 0 \)
- We underestimate \( \beta_1 \) if \( \beta_2 \alpha < 0 \)

- underspecified model, a regressor \( \beta_2 \) 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:

\[ t = \frac{\hat{\beta}_j - \beta_{j,0}}{\hat{\sigma}_{\beta_j}} \]

- The p-value is \( p = Pr \left( |t| > |t_{\text{sample}}| \right) = 2\Phi(-|t_{\text{sample}}|) \)

\[ F(-|t|) \]

\[ F(|t|) \]

\[-|t| \quad 0 \quad |t|\]
4.3.2 Joint Hypotheses

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

Formally $H_0: \beta_1 = \beta_{1,0} \land \beta_2 = \beta_{2,0}$

versus

$H_1: \beta_1 \neq \beta_{1,0} \lor \beta_2 \neq \beta_{2,0}$

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

$$
t_1 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\hat{\sigma}_{\beta_1}} \quad t_2 = \frac{\hat{\beta}_2 - \beta_{2,0}}{\hat{\sigma}_{\beta_2}}
$$

In that case the null hypothesis $H_0: \beta_1 = \beta_{1,0} \land \beta_2 = \beta_{2,0}$ 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 $\beta$s:

```r
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
```
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.

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:

```r
set.seed(100)
b1<-rnorm(N)
```
4.3.3 \( F \) statistic for two restrictions

\[
\begin{align*}
t_1 &= \frac{\hat{\beta}_1 - \beta_{1,0}}{\hat{\sigma}_{\beta_1}} \hspace{1cm} t_2 = \frac{\hat{\beta}_2 - \beta_{2,0}}{\hat{\sigma}_{\beta_2}} \\
F &= \frac{1}{2} \cdot \frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1 t_2} \cdot t_1 \cdot t_2}{1 - \hat{\rho}_{t_1 t_2}^2}
\end{align*}
\]

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

\[
R \beta = r
\]
e.g.

\[
\begin{pmatrix}
0, 1, 0, \ldots, 0
\end{pmatrix}
\begin{pmatrix}
\beta_0 \\
\beta_1 \\
\beta_2 \\
\vdots \\
\beta_k
\end{pmatrix} = 0
\]

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

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

\[
F = \frac{1}{q} (R\hat{\beta} - r)' \left( R\Sigma_{\hat{\beta}} R' \right)^{-1} (R\hat{\beta} - r)
\]

with \( q \) being the number of restrictions.

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

\[
F \xrightarrow{p} F_{q,\infty}
\]
\[ q = 2 \]
\[ q = 3 \]
\[ q = 5 \]
\[ q = 10 \]

\[ p = \Pr(F(q, \infty) > F_{\text{sample}}) \]
4.3.5 Specials cases:

The last line of a regression output

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

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

Residuals:

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>-51.340</td>
<td>1Q</td>
<td>0.293</td>
<td>10.318</td>
</tr>
<tr>
<td>1Q</td>
<td>-10.111</td>
<td>Median</td>
<td>Max</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3Q</td>
<td></td>
<td>43.181</td>
<td></td>
</tr>
</tbody>
</table>

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:

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

\[t_1 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\hat{\sigma}_{\hat{\beta}_1}}\]

\[X \sim t_{(k)} \iff X^2 \sim F_{(1,k)}\]

\[\text{rbind} \] constructs a vector by joining the arguments together. \[\text{rbind} \] joins the arguments of the function (vectors, matrices) line-wise. \[\text{cbind} \] 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.

```r
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

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8900</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>8570</td>
<td>2</td>
<td>3300.3</td>
<td>8.01</td>
<td>0.000386 ***</td>
</tr>
</tbody>
</table>

---
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.

- \( \frac{\partial testscr}{\partial str} \) 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:

\[
testsrc = \beta_1 \text{str} + \beta_2 \text{elpct} + \beta_0 + u
\]

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

Example 2:

\[
lwage = \beta_1 \text{ed} + \beta_0 + u
\]
Min. 1st Qu. Median Mean 3rd Qu. Max.
4.00 12.00 12.00 12.85 16.00 17.00

\texttt{lm(lwage \sim ed)}

Call:
\texttt{lm(formula = lwage \sim ed)}

Coefficients:
(Intercept) ed
5.8388 0.0652

Example 3:
\[ lwage = \beta_1 \text{college} + \beta_2 \text{sex} + \beta_0 + u \]

\texttt{college=ed>16}

\texttt{lm(lwage \sim college + sex)}

Call:
\texttt{lm(formula = lwage \sim college + sex)}

Coefficients:
(Intercept) collegeTRUE sexmale
6.2254 0.3340 0.4626

\subsection{4.4.1 Interactions between binary variables}

\[ lwage = \beta_0 + \beta_1 \text{college} + \beta_2 \text{sex} + \beta_3 \text{sex} \cdot \text{college} + u \]

\texttt{(est<- lm(lwage \sim college + sex + sex:college))}

Call:
\texttt{lm(formula = lwage \sim 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:

\texttt{aggregate(lwage \sim college + sex,FUN=mean)}
### 4.5 Literature


### 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:

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

where \( \text{male} = 1 \) if it is a boy, \( \text{age} \) denotes the child’s age in years and \( \text{rain} = 1 \) 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 \( \text{male} \) (1 if male, 0 otherwise) and added a variable \( \text{female} \) (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?

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

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

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):
\[ \hat{Y}_i = 300 + 10 \cdot X_1 + 20 \cdot X_2 \]

(1.0) (2.5)

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 lot-size (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 house. 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 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 downhill 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:

\[
\text{days}_i = 8 - 1 \cdot \text{iceSkating} - 2.5 \cdot \text{alpineSkiing} - 1.5 \cdot \text{iceSkating} \cdot \text{alpineSkiing}
\]

• 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: a comparison of two treatments:

```r
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:

```r
mean(mouse.t)
```

```
[1] 86.85714
```

```r
mean(mouse.c)
```

```
[1] 56.22222
```

Question: Is the difference significant? Traditionally we estimate

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

and hence

\[
\hat{\sigma}_x = \sqrt{\frac{\hat{\sigma}_x^2}{n}}
\]

The estimated standard error of the difference between the two means $\bar{x}$ and $\bar{y}$ is
\[
\hat{\sigma}_{\bar{x} - \bar{y}} = \sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}
\]

\(sdDiff <- \text{sqr}(\text{var}(\text{mouse.t})/\text{length}(\text{mouse.t})+\text{var}(\text{mouse.c})/\text{length}(\text{mouse.c}))\)

[1] 28.92647

Hence \(t = \frac{\bar{x} - \bar{y}}{\hat{\sigma}_{\bar{x} - \bar{y}}}\) is

\(t <- (\text{mean}(\text{mouse.t}) - \text{mean}(\text{mouse.c})) / \text{sdDiff} \)

[1] 1.059062

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

\[2*(1-\text{pnorm}(t))\]

[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 \(\sigma_X^2\), then

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

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:

\(\text{mean}(\text{mouse.t}) - \text{mean}(\text{mouse.c})\)

[1] 30.63492

\(\text{median}(\text{mouse.t}) - \text{median}(\text{mouse.c})\)

[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: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

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

![Empirical CDF](image)

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* $x$. 
Then our estimate of the distribution of $s$ is given by the distribution of the bootstrap replications $s(x^*)_1, s(x^*)_2, \ldots, s(x^*)_B$.

Our estimate of the standard error of $s(x)$ is given by the standard deviation of the bootstrap replications $s(x^*)_1, s(x^*)_2, \ldots, s(x^*)_B$.

\[
\hat{s}_{BS} = se_{b=1}^B (s(x^*_b)) = \sqrt{\frac{1}{B-1} \sum_{b=1}^B \left( s(x^*_b) - \frac{1}{B} \sum_{b=1}^B s(x^*_b) \right)^2}
\]
```r
median(sample(mouse.t, replace=TRUE)) - median(sample(mouse.c, replace=TRUE))
[1] 44
```

```r
median(sample(mouse.t, replace=TRUE)) - median(sample(mouse.c, replace=TRUE))
[1] 54
```

```r
median(sample(mouse.t, replace=TRUE)) - median(sample(mouse.c, replace=TRUE))
[1] 59
```

```r
replicate(20, median(sample(mouse.t, replace=TRUE)) - median(sample(mouse.c, replace=TRUE)))
[1] 48 48 -27 54 91 59 -2 -23 -17 54 95 68 48 48 91 95
[17] 53 53 -8 47
```

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

![Density plot](image)

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

```r
(sdDiff2 <- sd(replicate(100, mean(sample(mouse.t, replace=TRUE)) -
mean(sample(mouse.c, replace=TRUE)))))))
[1] 30.10606
```

```r
(sdMedianDiff2 <- sd(replicate(100, median(sample(mouse.t, replace=TRUE)) -
median(sample(mouse.c, replace=TRUE)))))))
[1] 40.91847
```
Is 100 large enough? Try different sizes of the bootstrap:

```r
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 $Y$).

```r
(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

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>94</td>
</tr>
<tr>
<td>2</td>
<td>197</td>
</tr>
</tbody>
</table>
Now we resample:

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

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>94</td>
</tr>
<tr>
<td>2</td>
<td>197</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>38</td>
</tr>
<tr>
<td>5</td>
<td>99</td>
</tr>
<tr>
<td>6</td>
<td>141</td>
</tr>
<tr>
<td>7</td>
<td>23</td>
</tr>
<tr>
<td>8</td>
<td>52</td>
</tr>
<tr>
<td>9</td>
<td>104</td>
</tr>
<tr>
<td>10</td>
<td>146</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>12</td>
<td>50</td>
</tr>
<tr>
<td>13</td>
<td>31</td>
</tr>
<tr>
<td>14</td>
<td>40</td>
</tr>
<tr>
<td>15</td>
<td>27</td>
</tr>
<tr>
<td>16</td>
<td>46</td>
</tr>
</tbody>
</table>

```r
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))})
```
Non-parametric test

Can be constructed always when $H_0$ is that $X$ does not affect $Y$ (observations are exchangeable).

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 $\theta$ a parameter of $F$.

**Parameters and statistics:**
\theta = t(F) \quad \text{with } t(\cdot) \text{ being a statistic of } F

The plug-in principle

\hat{\theta} = t(\hat{F})

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

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

Glivenko-Cantelli’s theorem (1933)

\[
\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{n \to \infty} 0 \quad \text{almost surely}
\]

if \( t(\hat{F}) \) is continuous then

\[
t(\hat{F}_n) \xrightarrow{n \to \infty} t(F) = \theta \quad \text{almost surely}
\]

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 \( x^{*1}, x^{*2}, \ldots, x^{*B} \).

   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 \( \hat{\theta}^{*b} = s(x^{*b}) \).

3. \[
\hat{\sigma}_{s,BS} = \text{se}_{b=1}^{B} (\hat{\theta}^{*b}) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left( \hat{\theta}^{*b} - \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{*b} \right)^2}
\]

\[
\lim_{B \to \infty} \hat{\sigma}_{s,BS} = \text{se}_{\hat{F}}(\hat{\theta}^*)
\]
5.3.2 Sampling multivariate data

```r
with(law82, plot(100 * GPA ~ LSAT, ylab = "GPA"))
points(law, 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 $\theta$ and the $t(\cdot)$ we are talking about is the correlation.)
The true correlation:

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

[1] 0.7599979

The plug-in estimate:

```r
(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]  5 12  7 14 15  1  8 14  9  7 15  7 11  9  2

law[i2,]

LSAT  GPA
   5   666 3.44
  12   575 2.74
  14   572 2.88
  15   594 2.96
    1   576 3.39
    8   661 3.43
  14.1  572 2.88
    9   651 3.36
  7.1   555 3.00
 15.1   594 2.96
  7.2   555 3.00
  11   653 3.12
  9.1   651 3.36
    2   635 3.30

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

[1] 0.1373078
```
How quickly does the estimate of $\hat{\sigma}$ converge?

```r
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 81 110 148 200 270 365 493 665 898 1212 1636 2208 2981

BSestimates <- sapply(BSsizes,lawBS)

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

![Plot showing the distribution of BS estimates](image)

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

```r
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))})
```
Comparison of the bootstrap estimate $se_F(\hat{\theta})$ with the standard error $se_F(\hat{\theta})$:

```r
sd(law.boot)
[1] 0.1343748
sd(law82.boot)
[1] 0.125938
```

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 2.748516
[11,] 549.1056 3.122109
[12,] 585.2290 2.898061
[13,] 583.5167 2.955358
[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

```r
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?

```r
sd(pBoot)
```

[1] 0.1168818

Remember:

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

```r
sd(law82.boot)
```

sd(law.boot)

5.5 Literature


Appendix 5.A  Examples for the lecture

Example 1  The dataset \textit{x1.csv} contains two variables, \textit{x1} and \textit{x2}.

- What is the interquartile range of \textit{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 vertical axis. How quick does your estimation converge?

Example 2  Look again at the dataset \textit{x1.csv}.

- What is the average value of the ratio \textit{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 \texttt{R} might help: \texttt{data, str, names, length, dim, plot, density, median, summary, sd, quantile, cor, set.seed, sample, replicate, rnorm, lm, vcov, coeftest}. Note that \texttt{coeftest} is provided by the \texttt{car} library.

Exercise 5.1  Consider the dataset \textit{Caschool} from the package \textit{Ecdat}.

- What is the median value of \textit{testscr}?
- Give a standard deviation for the median value.
- Give a 95%-confidence interval.
- Use a parametric bootstrap.

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

- What is \( \hat{\sigma} \).
• What is the standard deviation of $\hat{\alpha}$?
• What is the ratio $\hat{\beta}/\hat{\alpha}$?
• 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 e1pct 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, $\theta$ 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.)

• $X \rightarrow \theta$, $X$ is fixed, $\theta$ 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.

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. $\theta = \text{guilty}$ or $\theta = \text{not guilty}$. Alternatively, $\beta = 0$ or $\beta \neq 0$. 

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

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

with $A = \underline{\theta}$ parameter and $B = \underline{X}$ data:

$$
\Pr(\theta) \cdot \Pr(X|\theta) \cdot \frac{1}{\Pr(X)} = \Pr(\theta|X)
$$

$$
\int \Pr(\theta) \Pr(X|\theta) \, d\theta
$$

posterior
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|θ₀)) is equal to the probability that the defendant is innocent (Pr(θ₀|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(evidence|not guilty) = 1/342 million
  - Pr(evidence|not guilty) = 1/25
- Sally Clark
  - Pr(evidence|not guilty) = 1/73 million
  - Pr(not guilty|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
- →: Pr(evidence|not guilty) = (1/8543)² ≈ 1/73 million

Problems:

- Correlation of SIDS within a family. Pr(2nd child) = (1/8543) × 5...10
- SIDS is actually more likely in this case: P = 1/8543 → P = 1/1300
  Pr(evidence|1 not guilty mother) = 1/(1300 · 130) = 0.000592 %
- Intention of the researcher/multiple testing: 750 000 births in England and Wales / year. How likely is it to find two successive SIDS or more among 750 000 mothers.
  Pr(evidence|750 000 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?

\[
\frac{\text{Pr}(\theta) \cdot \text{Pr}(X|\theta) \cdot \frac{1}{\text{Pr}(X)}}{\text{Pr}(\theta|X)} = \frac{\text{Pr}(\theta)}{\text{Pr}(X)}
\]

\[
\frac{\text{Pr} (g) \cdot \text{Pr}(X|g)}{\text{Pr}(g) \cdot \text{Pr}(X|g) + (1-\text{Pr}(g)) \cdot \text{Pr}(X|\neg g)} \cdot \frac{1}{\text{Pr}(X)} = \frac{\text{Pr}(g|X)}{\text{Pr}(X)}
\]

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) = \frac{1}{600,000} + \frac{599,999}{600,000} \cdot \frac{1}{1300.130}

Pr(g|evidence) = 22%

If Pr(g) = 1/18800 then Pr(g|evidence) = 90%
If Pr(g) = 1/1710 then Pr(g|evidence) = 99%

\[
\text{Pr}(X|\theta) \neq \text{Pr}(\theta|X)
\]
• The interpretation of \( \Pr(X|\theta) \) as a p-value is affected by the intention of the researcher (e.g. multiple testing).

\( \rightarrow \) The same \( X \) could have been collected with different intentions \( \rightarrow \) 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

\( \phi_j \) test procedure

• choice of control variables
• data exclusion
• coding
• analysis
• interactions
• predictors

\( T(X, \phi_j) \) test result

**p-hacking**

• perform \( J \) tests: \( \{\ldots, T(X, \phi_j), \ldots\} \)

• report the best result, given the data: \( T(X, \phi_{\text{best}}) \)

\( \rightarrow \) to correct for multiple testing we need to know \( J \)

\( \rightarrow \) 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 \( \beta \): \( \beta_i \sim N(0,1) \)
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:
  $$\Pr(guilt) : \Pr(\text{innocence}) = \frac{1}{2} : \frac{1}{2}.$$  
  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 committed 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 \( \alpha \in \{0, 1\} \). Payoffs are \( \pi_{\alpha \theta} \).

We have \( \pi_{11} > \pi_{01} \) and \( \pi_{00} > \pi_{10} \), i.e. it is better to choose \( \alpha = \theta \). Expected payoffs:

\[
E(\pi|\alpha) = \pi_{\alpha 1} \cdot \Pr(\theta = 1) + \pi_{\alpha 0} \cdot \Pr(\theta = 0)
\]

Optimal decision: choose \( \alpha = 1 \) iff

\[
\frac{\pi_{11} \cdot \Pr(\theta = 1) + \pi_{10} \cdot \Pr(\theta = 0)}{\Pr(X)} > \frac{\pi_{01} \cdot \Pr(\theta = 1) + \pi_{00} \cdot \Pr(\theta = 0)}{\Pr(X)}.
\]

Rearrange: choose \( \alpha = 1 \) iff

\[
\Pr(\theta = 1) \left( \frac{\pi_{11} - \pi_{01}}{g_1} \right) > \Pr(\theta = 0) \left( \frac{\pi_{00} - \pi_{10}}{g_0} \right).
\]

Here \( g_\alpha \) can be seen as the gain from choosing the correct action (or the loss from choosing the wrong action) if \( \theta = \alpha \).

If we have some data \( X \):

\[
\Pr(\theta = 1|X)g_1 > \Pr(\theta = 0|X)g_0.
\]

Bayes’ rule:

\[
\Pr(\theta) \cdot \Pr(X|\theta) \cdot \frac{1}{\Pr(X)} = \Pr(\theta|X)
\]

choose \( \alpha = 1 \) iff

\[
\frac{g_1}{g_0} \cdot \frac{\Pr(\theta = 0|X)}{\Pr(\theta = 1|X)} = \frac{\Pr(\theta = 0) \cdot \Pr(X|\theta = 0)}{\Pr(\theta = 1) \cdot \Pr(X|\theta = 1)} = \frac{\Pr(\theta = 0) \cdot \Pr(X|\theta = 0)}{\Pr(\theta = 1) \cdot \Pr(X|\theta = 1)}.
\]

choose \( \alpha = 1 \) iff

\[
\frac{g_1 \Pr(\theta = 1)}{g_0 \Pr(\theta = 0)} \cdot \Pr(X|\theta = 1) > \Pr(X|\theta = 0)
\]

Bayesian chooses \( \alpha = 1 \) iff \( \Pr(X|\theta = 0) < \frac{g_1 \Pr(\theta = 1)}{g_0 \Pr(\theta = 0)} \cdot \Pr(X|\theta = 1) \)

Frequentist chooses \( \alpha = 1 \) 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 $\frac{g_1 \Pr(\theta = 1)}{g_0 \Pr(\theta = 0)} \cdot \Pr(X|\theta = 1)$ 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 $\to$ 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


- Hoff, A First Course in Bayesian Statistical Methods.


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 $\zeta$, 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

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 \( \sigma \), and if \( \theta = 2 \), then \( y \) has a normal distribution with mean 2 and standard deviation \( \sigma \). 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 \( \theta \) changes in shape as \( \sigma \) 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. \( n = 10, p \text{ value} = 0.032 \).
2. \( n = 100, p \text{ value} = 0.032 \).

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, \( p(1) = p(2) = p(3) = \frac{1}{3} \). According to theory B we have \( p(1) = \frac{1}{2} \), \( p(2) = p(3) = \frac{1}{4} \). According to theory C we have \( p(1) = p(2) = \frac{1}{4}, p(3) = \frac{1}{2} \).

- 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. What 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**

```r
library(Ecdat)
data(Crime)
xyplot(crmrte ~ prbarr, data=Crime, subset=year==81)
```
We can have a look at a part of the data with `head`:

```r
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`:

```r
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

\[
\frac{\Pr(\theta) \cdot \Pr(X|\theta)}{\Pr(X)} \cdot \frac{1}{\int \Pr(\theta) \cdot \Pr(X|\theta) \, d\theta} = \Pr(\theta|X)
\]

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

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 \(D\) the distribution of parameter \(\theta\).

\[
D_0 \xrightarrow{X_1} D_1 \xrightarrow{X_2} D_{12} \\
D_0 \xrightarrow{X_2} D_2 \xrightarrow{X_1} D_{12}
\]

This is easier if \(D_0, D_1, D_2, D_{12}\) 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**

<table>
<thead>
<tr>
<th>Likelihood</th>
<th>known</th>
<th>model parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X \sim N(\mu, \sigma^2))</td>
<td>(\tau = 1/\sigma^2)</td>
<td>(\mu \sim N(\mu_0, \sigma_0^2))</td>
</tr>
<tr>
<td>(X \sim N(\mu, \tau))</td>
<td>(\mu)</td>
<td>(\tau \sim \Gamma(\alpha_0, \beta_0))</td>
</tr>
<tr>
<td>(X \sim \text{bern}(p))</td>
<td>(p)</td>
<td>(p \sim \text{Beta}(\alpha_0, \beta_0))</td>
</tr>
</tbody>
</table>

If the prior model parameter follows the conjugate prior, then the posterior model parameter is in the same family.
Conjugate priors, example: normal likelihood $\mu$

- Likelihood: $X \sim N(\mu, \sigma^2)$ with known $\tau = 1/\sigma^2$.
- Model parameter: $\mu$
- Conjugate prior distribution: $\mu \sim N(\mu_0, \sigma_0^2)$
- Prior hyperparameter: $\mu_0, \sigma_0^2$ i.e. prior $\mu \sim N(\mu_0, \sigma_0^2)$.
- Posterior hyperparameter:

$$
\mu_{\text{post}} = \left( \frac{\mu_0}{\sigma_0^2} + \frac{n \cdot \bar{x}}{\sigma^2} \right) / \left( \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right) = \frac{\tau_0 \mu_0 + n \tau \bar{x}}{\tau_0 + n \tau}
$$

$$
\tau_{\text{post}} = 1/\sigma_{\text{post}}^2 = \left( \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right) = \tau_0 + n \tau
$$

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: $\mu_0, \tau_0$ (they determine the distribution of $\mu$)
- Parameters: $\mu, \tau$
- Posterior hyperparameters: $\mu_{\text{post}}, \tau_{\text{post}}$

Conjugate Priors, example: Normal Likelihood $\tau$

- Likelihood: $X \sim N(\mu, \tau)$ with known $\mu$.
- Model parameter: $\tau = 1/\sigma^2$
- Conjugate prior distribution: $\tau \sim \Gamma(\alpha_0, \beta_0)$
- Prior hyperparameter: $\alpha_0, \beta_0$
- Posterior hyperparameter:
shape \quad \alpha_{\text{post}} = \alpha_0 + \frac{n}{2}

rate \quad \beta_{\text{post}} = \beta_0 + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2}

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: \( \alpha_0, \beta_0 \) (they determine the distribution of \( \tau \))
- Parameters: \( \mu, \tau \)
- 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: \( \alpha_0, \beta_0 \)
- Posterior hyperparameter:

\[
\alpha_{\text{post}} = \alpha_0 + \sum x_i \\
\beta_{\text{post}} = \beta_0 + n - \sum x_i
\]

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: \( \alpha_0, \beta_0 \) (they determine the distribution of \( p \))
- Parameters: \( p \)
- Posterior hyperparameters: \( \alpha_{\text{post}}, \beta_{\text{post}} \)
Problems with the analytical approach

- Restrictive for...
  - priors
  - likelihood (“the model” in the frequentist world)
- For many relevant cases we have no analytical solution.
- → numerical methods, 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

Required:

- Priors for $\mu$ and $\tau$.
- Likelihood: $y \sim N(\mu, \tau)$ with $\tau = 1/\sigma^2$

→ Posterior distribution of $\mu$ and $\tau$.

We will here just “use” our software got get a result. Below we will explain what the software actually does.

```r
library(runjags)
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

- Stochastic nodes (discrete/continuous univariate/multivariate distributed):
  
  $y[i] \sim dnorm(\mu,\tau)$

  ... 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:

\[ sd \leftarrow \sqrt{\frac{1}{\tau}} \]

JAGS samples from the posterior of \( \mu \) and \( \tau \). Here is a distribution for \( \mu \):

```
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
mu    0.272130  0.299  0.3253  0.29914  0.013618  0.29942  0.000096296
sd    0.110260  0.12831  0.1487  0.12888  0.0098465  0.12753  0.000069557

       MC%ofSD     SSeff     AC.10         psrf
mu 0.7  200000  0.0065149  1.0003
sd 0.7  20040  0.0050537  0.99996

Total time taken: 0.9 seconds
Testing a point prediction in the \( t.\text{test} \), 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 \( \mu \in \{0.29, 0.31\}? \)

```r
X.df <- data.frame(as.mcmc(X.jags))
str(X.df)

'data.frame': 20000 obs. of 2 variables:
$ mu: num 0.311 0.295 0.298 0.289 0.305 ...
$ sd: num 0.12 0.122 0.151 0.122 0.139 ...
```

We can now say, how probable it is, ex post, that \( \mu \in [0.29, 0.31] \):

```r
100*mean(with(X.df, mu > 0.29 & mu < 0.31))
[1] 53.92
```

...or in a more narrow interval:

```r
100*mean(with(X.df, mu > 0.299 & mu < 0.301))
[1] 5.725
```

```r
100*mean(with(X.df, mu > 0.2999 & mu < 0.3001))
[1] 0.57
```

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 \( \mu > 0.25? \)

```r
100*mean(with(X.df, mu > 0.25))
[1] 99.975
```

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 \( \mu \) (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 \( \sigma_{\mu} \) from our data, but a different \( \mu \)).
    
    \( \sigma_{\mu} = 0.014 \) is equivalent to \( \tau_{\mu} = 1/\sigma_{\mu}^2 = 5102 \)
Now we combine the data, i.e. we use a prior $\mu \sim \mathcal{N}(0.4, 5102)$

```r
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)
}'
XA.jags <- run.jags(model=XA.model, data=list(y=y), monitor=c("mu","sd"))
summary(XA.jags)
```

<table>
<thead>
<tr>
<th></th>
<th>Lower95</th>
<th>Median</th>
<th>Upper95</th>
<th>Mean</th>
<th>SD</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>mu</td>
<td>0.330308</td>
<td>0.3514715</td>
<td>0.372552</td>
<td>0.3516435</td>
<td>0.01085470</td>
<td>0.3516392</td>
</tr>
<tr>
<td>sd</td>
<td>0.118612</td>
<td>0.1382475</td>
<td>0.162210</td>
<td>0.1389645</td>
<td>0.01124677</td>
<td>0.1378229</td>
</tr>
</tbody>
</table>

- 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.”

```r
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)
```

![Graph showing three normal distributions](image)
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

\[
\frac{\Pr(\theta) \cdot \Pr(\theta) \cdot \Pr(X|\theta)}{\Pr(X)} \cdot \frac{1}{\int \Pr(\theta) \cdot \Pr(X|\theta) \, d\theta} = \Pr(\theta|X)
\]

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 \( \theta \) from joint distribution \( \Pr(X|\theta) \).
  - Gibbs sampling: quicker, samples are correlated, requires sampling of \( \theta_i \) from conditional (on \( \theta_{-i} \)) distribution \( \Pr(X|\theta_i, \theta_{-i}) \).
    → this is easy! (at least much easier than \( \Pr(X|\theta) \))

7.6.2 Rejection sampling

\[
\frac{\Pr(\theta) \cdot \Pr(\theta) \cdot \Pr(X|\theta)}{\Pr(X)} \cdot \frac{1}{\int \Pr(\theta) \cdot \Pr(X|\theta) \, d\theta} = \Pr(\theta|X)
\]

How it works: Iterate the following:

- Sample a candidate \( \theta \) and a uniformly distributed random number \( r \).
- If \( \Pr(\theta) \cdot \Pr(X|\theta) > r \) then \( \theta \) 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

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

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 \( \alpha \geq 1 \): \( \theta_{t+1} = \theta' \) \( \text{j} \) \( \text{ump} \)
  - If \( \alpha < 1 \): with probability \( \alpha \) we have \( \theta_{t+1} = \theta' \), otherwise \( \theta_{t+1} = \theta_t \) \( \text{stay} \).

**Advantages:**
- Faster than rejection sampling (in particular if \( \theta \) 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 \( \theta \) sequentially.

- determine \( \theta_1^{t+1} \) with \( f(\theta_1|\theta_2^t, \theta_3^t, \theta_4^t, \ldots, \theta_n^t) \)
- determine \( \theta_2^{t+1} \) with \( f(\theta_2|\theta_1^{t+1}, \theta_3^t, \theta_4^t, \ldots, \theta_n^t) \)
- determine \( \theta_3^{t+1} \) with \( f(\theta_3|\theta_1^{t+1}, \theta_2^{t+1}, \theta_4^t, \ldots, \theta_n^t) \)
- ... 
- determine \( \theta_n^{t+1} \) with \( f(\theta_n|\theta_1^{t+1}, \theta_2^{t+1}, \ldots, \theta_{n-1}^{t+1}) \)

**Advantages:**
- Requires only conditional distributions. \( f(\theta_1|\theta_{-1}) \), 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.

```r
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:

```r
est <- lm(crmrte ~ pbarr, data = Crime, subset = year == 81)
summary(est)
```

Call:
```
lm(formula = crmrte ~ pbarr, 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

\[ Y = \beta_0 + \beta_1 X + u \quad \text{where } u \sim N(0, \sigma^2) \]
\[ Y \sim N(\beta_0 + \beta_1 X, \sigma^2) \]
\[ Y \sim N(\beta_0 + \beta_1 X, \tau) \]

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:

```r
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,.001)
  beta1 ~ dnorm (0,.001)
  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):

<table>
<thead>
<tr>
<th>Lower95</th>
<th>Median</th>
<th>Upper95</th>
<th>Mean</th>
<th>SD</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta0</td>
<td>0.037053</td>
<td>0.048573</td>
<td>0.06003</td>
<td>0.048601</td>
<td>0.0058882</td>
</tr>
<tr>
<td>beta1</td>
<td>-0.08893</td>
<td>-0.053022</td>
<td>-0.017757</td>
<td>-0.053063</td>
<td>0.018197</td>
</tr>
</tbody>
</table>

MCerr MC%ofSD SS eff AC.10 psrf
beta0 0.00013929 2.4 1787 0.17357 1.0006
beta1 0.00044478 2.4 1674 0.17632 1.0007

Total time taken: 0.4 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 $\beta_0$ and $\beta_1$).

We see that, to fit the data, $\beta_0$ and $\beta_1$ must be correlated. This is also what we see in the distribution of the sampled posterior:

```r
reg.df <- as.data.frame(combine.mcmc(reg.jags))
xyplot(bet1 ~ bet0, 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.

\begin{align}
Y &= \beta_0 + \beta_1 X \\
Y - \bar{Y} &= \beta_0 - \bar{Y} + \beta_1 \bar{X} + \beta_1 (X - \bar{X}) 
\end{align}

Now we demean the data:

```r
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"))
```

```r
summary(reg2.jags)
```

<table>
<thead>
<tr>
<th></th>
<th>Lower95</th>
<th>Median</th>
<th>Upper95</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta0</td>
<td>-0.004726</td>
<td>-0.000027</td>
<td>0.004494</td>
<td>-0.000021</td>
</tr>
<tr>
<td>beta1</td>
<td>-0.088722</td>
<td>-0.052886</td>
<td>-0.016754</td>
<td>-0.052907</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>Mode</td>
<td>MCerr</td>
<td>MC%ofSD</td>
</tr>
<tr>
<td>beta0</td>
<td>0.002343</td>
<td>0.000012</td>
<td>0.000016</td>
<td>0.7 20404</td>
</tr>
<tr>
<td>beta1</td>
<td>0.018464</td>
<td>-0.053449</td>
<td>0.000130</td>
<td>0.7 20000</td>
</tr>
<tr>
<td></td>
<td>AC.10</td>
<td>psrf</td>
<td></td>
<td></td>
</tr>
<tr>
<td>beta0</td>
<td>0.005122</td>
<td>1.000028</td>
<td></td>
<td></td>
</tr>
<tr>
<td>beta1</td>
<td>0.003685</td>
<td>1.000036</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Compare with not demeaned estimation:

```
summary(reg.jags)
```

<table>
<thead>
<tr>
<th></th>
<th>Lower95</th>
<th>Median</th>
<th>Upper95</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta0</td>
<td>0.0370532</td>
<td>0.0485726</td>
<td>0.0600297</td>
<td>0.04860075</td>
<td>0.005888166</td>
</tr>
<tr>
<td>beta1</td>
<td>-0.0889300</td>
<td>-0.05302175</td>
<td>-0.0177575</td>
<td>-0.05306305</td>
<td>0.018196589</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mode</th>
<th>MCerr</th>
<th>MC%ofSD</th>
<th>SSef</th>
<th>AC.10</th>
<th>psrf</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta0</td>
<td>0.04891597</td>
<td>0.0001392896</td>
<td>2.4</td>
<td>1787</td>
<td>0.1735746</td>
</tr>
<tr>
<td>beta1</td>
<td>-0.05241875</td>
<td>0.0004447758</td>
<td>2.4</td>
<td>1674</td>
<td>0.1763241</td>
</tr>
</tbody>
</table>

The estimate for \( beta_1 \) does not change (here we assume that we are mainly interested in the marginal effect, i.e. in \( beta_1 \)).

Now \( beta_0 \) and \( beta_1 \) 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:

<table>
<thead>
<tr>
<th>demeaned</th>
<th>raw</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-8</td>
</tr>
<tr>
<td>1</td>
<td>-6</td>
</tr>
<tr>
<td>0</td>
<td>-4</td>
</tr>
<tr>
<td>2</td>
<td>-2</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
</tr>
</tbody>
</table>

Here are 100 samples:

<table>
<thead>
<tr>
<th>demeaned</th>
<th>raw</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-8</td>
</tr>
<tr>
<td>1</td>
<td>-6</td>
</tr>
<tr>
<td>0</td>
<td>-4</td>
</tr>
<tr>
<td>2</td>
<td>-2</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
</tr>
</tbody>
</table>

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

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.

```r
acfplot(as.mcmc(reg.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:
In the demeaned model `reg2`, the value of AC.10 is smaller:

```
summary(reg2.jags)

<table>
<thead>
<tr>
<th></th>
<th>Lower95</th>
<th>Median</th>
<th>Upper95</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta0</td>
<td>-0.00472617</td>
<td>-0.000002706275</td>
<td>0.000449406</td>
<td>-0.00002115496</td>
<td></td>
</tr>
<tr>
<td>beta1</td>
<td>-0.08872240</td>
<td>-0.05288575000</td>
<td>-0.01675430</td>
<td>-0.05290716645</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>0.002343448</td>
<td>0.00001269985</td>
<td>0.0001640579</td>
<td>0.7</td>
<td>20404</td>
</tr>
<tr>
<td>AC.10</td>
<td>0.005122056</td>
<td>1.0000280</td>
<td>0.003685244</td>
<td>1.000036</td>
<td></td>
</tr>
</tbody>
</table>
```

• A sample of 10 000 can, thus, not be treated as 10 000 independent observations.

```
effectiveSize(as.mcmc.list(reg.jags))
```

```
beta0  beta1
1786.992  1673.778
```

```
effectiveSize(as.mcmc.list(reg2.jags))
```

```
beta0  beta1
20404.05  20000.00
```

• 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 $z$:

```r
(x <- rbinom(9, 1, .5))
```

```
[1] 0 0 0 1 0 1 0 0 0
```
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 \textit{psrf} for a “nice” case:

```
gelman.diag(reg.jags)
```

```
Potential scale reduction factors:

<table>
<thead>
<tr>
<th></th>
<th>Point est.</th>
<th>Upper C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta0</td>
<td>1</td>
<td>1.01</td>
</tr>
<tr>
<td>beta1</td>
<td>1</td>
<td>1.01</td>
</tr>
</tbody>
</table>
```

```
Multivariate psrf
1
```

And now the island case:

```
gelman.diag(island.jags)
```

```
Potential scale reduction factors:

<table>
<thead>
<tr>
<th></th>
<th>Point est.</th>
<th>Upper C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>z</td>
<td>1.24</td>
<td>1.87</td>
</tr>
</tbody>
</table>
```

```
acfplot(as.mcmc(island.jags), aspect="fill", layout=c(2,1), ylim=c(-1,1))
```
Autocorrelation in the island case

As a result of autocorrelation, the “effective size” is smaller than the sample size.

```r
effectiveSize(as.mcmc.list(island.jags))

    z
7.127945

effectiveSize(as.mcmc.list(reg.jags))

    beta0   beta1
1786.992 1673.778

effectiveSize(as.mcmc.list(reg2.jags))

    beta0   beta1
20404.05 20000.00
```

7.7.5 A better vague prior for \( \tau \)

When we specify a regression model we need a precision parameter \( \tau \). So far we did this:

```r
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)
}'
A more stable way to model $\tau$ might be the following:

```r
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:

- $\tau \sim \Gamma\left(\frac{m^2}{d^2}, \frac{m}{d^2}\right)$
- $m \sim \Gamma(1,1)$
- $d \sim \Gamma(1,1)$

$\rightarrow E(\tau) = m$, $\text{var}(\tau) = d^2$

- Speed: no substantial loss
- Convergence: often faster

### 7.8 Literature

- Hoff, A First Course in Bayesian Statistical Methods.


### 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):

<table>
<thead>
<tr>
<th>Lower95</th>
<th>Median</th>
<th>Upper95</th>
<th>Mean</th>
<th>SD</th>
<th>Mode</th>
<th>MCerr</th>
</tr>
</thead>
<tbody>
<tr>
<td>mu</td>
<td>4.2</td>
<td>6.5</td>
<td>9.0</td>
<td>6.5</td>
<td>1.0</td>
<td>--</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MC%ofSD</th>
<th>SSeff</th>
<th>AC.10</th>
<th>psrf</th>
</tr>
</thead>
<tbody>
<tr>
<td>mu</td>
<td>0.7</td>
<td>20045</td>
<td>-0.0055416</td>
</tr>
</tbody>
</table>

- Did the sampler converge?
- Are the samples autocorrelated?
- What is a 95% credible interval?
- How probable is it that $6.5 \leq \mu \leq 9$?

**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 $\mu$ and known variance $\sigma^2 = 1$. The mean of your sample of 10 observations is $\bar{x} = 2$.

You are interested in $\mu$. Your prior for $\mu$ is that it follows a normal distribution with mean 40 and variance 4.

What is the posterior distribution for $\mu$?

### 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 \( \alpha \) and \( \beta \).

- 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 $\alpha$ and $\beta$? 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 $\alpha$ and $\beta$? Draw the density of $p$. Can you give an 95%-credible interval for $p$?

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 $\text{qual}$ depends on the effort and a random component $e$ which is $\mathcal{N}(0, \sigma)$.

Students pass the final exam ($Y = 1$) when their qualification is larger than a critical value $\text{crit}$.

$$Y = 1 \iff x + e > \text{crit}$$

True relationship:

$$\Pr(Y = 1| x) = F(x, \beta)$$
$$\Pr(Y = 0| x) = 1 - F(x, \beta)$$

The teacher knows the structure of the process $F()$, but not its parameters $\beta$. 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 $x$.

```r
N <- 100
sigma <- .1
crit <- .3
set.seed(453)
x <- sort(runif(N))
```

Using a sorted 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.

```r
histogram(x)
ecdfplot(x)
```
Let us next add some random noise \(e\) to the effort \(x\) of the teacher and determine the (unobservable) qualification \(\text{qual}\) as well as the (observable) exam result \(y\):

\[
e \leftarrow \sigma * \text{rnorm}(N)
\]
\[
\text{qual} \leftarrow x + e
\]
\[
\text{plot}(x, \text{qual}, \text{pch}=4, \text{col}=4)
\]
\[
\text{abline}(h=\text{crit}, \text{col}=4, \text{lt}\text{y}=\text{"dotted"})
\]
\[
y \leftarrow \text{ifelse(qual}>\text{crit}, 1, 0)
\]
\[
\text{points}(y \sim x, \text{pch}=1, \text{col}=1)
\]
\[
\text{legend}(\text{"bottomright"}, \text{c(\"latent", \"observed\")}, \text{pch}=\text{c(4,1), col}=\text{c(4,1), bg}=\text{"white"})
\]
8.1.2 Try to estimate the relationship with OLS

\[ \Pr(Y = 1|x) = F(x, \beta) \]
\[ \Pr(Y = 0|x) = 1 - F(x, \beta) \]

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.

- $\text{E}(u|X) \neq 0$ 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:

\[
\begin{align*}
\Pr(Y = 1|x) &= F(x, \beta) \\
\Pr(Y = 0|x) &= 1 - F(x, \beta)
\end{align*}
\]

Trick: find a monotonic $\hat{F}$, such that $F(x, \beta) = \hat{F}(x'\beta)$ with

\[
\lim_{x'\beta \to +\infty} \Pr(Y = 1|x) = 1 \\
\lim_{x'\beta \to -\infty} \Pr(Y = 1|x) = 0
\]

Any continuous distribution function would do.

A common distribution function is, e.g. the standard normal distribution. In R we call this function \texttt{pnorm}. Another possible function is the logistic function $\frac{e^x}{1+e^x}$. 

```r
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

```r
plot(plogis,-5,5,main="logistic",ylab="F(x)"
```

Less common functions are the Weibull model and the the Weibull distribution...

```r
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 ...

```r
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.

```r
(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:

```r
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

```r
lm( y ~ x)
```

Call:
```
  lm(formula = y ~ x)
```

Coefficients:
```
(Intercept)    x
  0.06257  1.20008
```

Now we do the same with glm

```r
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

\[ Y \sim N(X\beta, \sigma^2) \]

Similarly, with the logistic model we said

\[ \Pr(Y = 1|X) = F(X\beta) \]

but we can also write

\[ Y \sim \text{binom}(F(X\beta)) \]

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
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):

<table>
<thead>
<tr>
<th>family</th>
<th>link (F())</th>
</tr>
</thead>
<tbody>
<tr>
<td>gaussian (N)</td>
<td>identity, log, inverse</td>
</tr>
<tr>
<td>binomial</td>
<td>logit, probit, cauchit, log, cloglog</td>
</tr>
<tr>
<td>gamma</td>
<td>identity, log, inverse</td>
</tr>
<tr>
<td>poisson</td>
<td>identity, log, sqrt</td>
</tr>
<tr>
<td>inverse,gaussian</td>
<td>identity, log, inverse, 1/μ²</td>
</tr>
<tr>
<td>quasi</td>
<td>logit, probit, log, cloglog, identity, inverse, sqrt, 1/μ²</td>
</tr>
</tbody>
</table>

8.2.1 A list of common families and link functions

8.3 Marginal effects

8.3.1 Marginal effects with OLS

\[ E[y|x] = E[x'β + ε] = x'β \]

\[ \frac{\partial E[y|x]}{\partial x} = β \]

Here the marginal effect is simple to determine, it is just β. The marginal effect does not depend on x.
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)

\[
\text{myScale} = 0.1
\]

\[
\text{abline}(h=\text{coef(or)}["x"] \times \text{myScale}, \text{lty}=5, \text{col}=5)
\]

8.3.2 Marginal effects with binary choice models

\[
\Pr(Y = 1|x) = F(x'\beta) \quad \rightarrow \quad E[y|x] = F(x'\beta)
\]

\[
\frac{\partial E[y|x]}{\partial x} = f(x'\beta) \cdot \beta
\]

\[
\text{lmarginal} <- \text{dlogis(predict(lr))} \times \text{coef(lr)}["x"]
\]

We can calculate the marginal effects and show them in our diagram:
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 \(\text{mean}(x)\), hence \(x'\beta\) is

\[
\begin{align*}
\text{xbmean} & \leftarrow \text{coef(lr)} \%\% \text{c(1,mean(x))} \\
\text{xbmean} & \left[
\begin{array}{c}
1,1 \\
1, \\
\end{array}
\right] 3.928001
\end{align*}
\]

hence, the marginal effect \(f(x'\beta) \cdot \beta\) is

\[
\begin{align*}
\text{dlogis(xbmean)} & \times \text{coef(lr)}["x"] \\
\left[
\begin{array}{c}
1,1 \\
1, \\
\end{array}
\right] 0.349894
\end{align*}
\]
• We can also evaluate the marginal effect for all observations and then use the average. Remember that \texttt{lmarginal} still contains all the marginal effects, so we can simply take the mean.

\[
\text{mean(lmarginal)}
\]

\[
[1] 0.8585523
\]

As we see, the result depends heavily on the method. The first method gave us the marginal effect on the \textit{average subject}, the second the \textit{average marginal effect} (of all subjects).

\[
d\text{logis}(\text{xbmean}) \cdot \text{coef(lr)}["x"]
\]

\[
[,1]
\]

\[
[1,] 0.349894
\]

\[
\text{mean(lmarginal)}
\]

\[
[1] 0.8585523
\]

Furthermore, we could argue that a marginal change should, ideally, first target at the subjects which can be influenced most easily:

\[
\text{max(lmarginal)}
\]

\[
[1] 4.612386
\]

When \(X_d\) is a dummy variable, derivatives make no sense. Instead, we use (for the average marginal effect):

\[
\Pr(Y = 1|x_{-d}, d = 1) - \Pr(Y = 1|x_{-d}, d = 0)
= \frac{1}{n} \sum_{i=1}^{n} \left( F((x_0,i,\ldots,x_{d-1,i},1,x_{d+1,i},\ldots,x_k,i)' \beta) - F((x_0,i,\ldots,x_{d-1,i},0,x_{d+1,i},\ldots,x_k,i)' \beta) \right)
\]

and for the marginal effect at the average \(x\):

\[
\Pr(Y = 1|\bar{x}_{-d}, d = 1) - \Pr(Y = 1|\bar{x}_{-d}, d = 0)
= F((\bar{x}_0,i,\ldots,\bar{x}_{d-1,i},1,\bar{x}_{d+1,i},\ldots,\bar{x}_k,i)' \beta) -
- F((\bar{x}_0,i,\ldots,\bar{x}_{d-1,i},0,\bar{x}_{d+1,i},\ldots,\bar{x}_k,i)' \beta)
\]
8.3.4 The normal link function

\[
\Pr(Y = 1|x) = F(x, \beta) = \Phi(x'\beta)
\]

\[
\Pr(Y = 0|x) = 1 - F(x, \beta) = 1 - \Phi(x'\beta)
\]

or

\[
Y \sim \text{binom}\left(\frac{\Phi(X\beta)}{1}\right)
\]

(pr <- glm( y ~ x, family=binomial(link="probit")))

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

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):

<table>
<thead>
<tr>
<th></th>
<th>Lower95</th>
<th>Median</th>
<th>Upper95</th>
<th>Mean</th>
<th>SD</th>
<th>Mode</th>
<th>MCerr</th>
<th>MC%ofSD</th>
<th>SSeff</th>
</tr>
</thead>
<tbody>
<tr>
<td>b1</td>
<td>-5.3592</td>
<td>-3.47</td>
<td>-1.9202</td>
<td>-3.5625</td>
<td>0.89921</td>
<td>--</td>
<td>-0.036674</td>
<td>4.1</td>
<td>601</td>
</tr>
<tr>
<td>b2</td>
<td>6.6925</td>
<td>11.483</td>
<td>17.975</td>
<td>11.84</td>
<td>2.9615</td>
<td>--</td>
<td>0.12332</td>
<td>4.2</td>
<td>577</td>
</tr>
</tbody>
</table>

AC.10  psrf
b1 0.5495 1.0015
b2 0.55569 1.0012

Total time taken: 3.2 seconds

Here is the figure again, now with the estimation result for the normal link function added:
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:

```r
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 (probabilities 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 cannot say too much about ratios of probabilities for extreme values).

### 8.3.5 Marginal effects with the normal link function

\[
\begin{align*}
\Pr(Y = 1|x) &= F(x, \beta) = \Phi(x'\beta) \\
\Pr(Y = 0|x) &= 1 - F(x, \beta) = 1 - \Phi(x'\beta) \\
E(y|x) &= F(x'\beta) = \Phi(x'\beta)
\end{align*}
\]

Marginal effects are:

\[
\frac{\partial E[y|x]}{\partial x} = f(x'\beta) \cdot \beta = \phi(x'\beta) \cdot \beta
\]

```r
pmarginal <- dnorm(pr$logit.predictors) * coef(pr)"x"
```

Now let us draw the familiar figure again, now with the marginal effects of the probit model added:

```r
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?

```r
xbmeanP <- coef(pr) %*% c(1, mean(x))
dnorm(xbmeanP) * coef(pr)["x"]
```

```
[,1]
[1,] 0.3353767
```

- Effect for the average unit:

```r
mean(pmargin)
```

```
[1] 0.8420524
```

- Effect for the most efficient unit:

```r
max(pmargin)
```

```
[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**

\[ y = \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon \]
\[ E[y|x] = \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 \]
\[ \frac{\partial E[y|x]}{\partial x_1} = \beta_1 + \beta_{12} x_2 \]
\[ \frac{\partial^2 E[y|x]}{\partial x_1 \partial x_2} = \beta_{12} \]

**probit model**

\[ E[y|x] = \Phi(\beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2) \]
\[ \frac{\partial E[y|x]}{\partial x_1} = (\beta_1 + \beta_{12} x_2) \cdot \phi(u) \]
\[ \frac{\partial^2 E[y|x]}{\partial x_1 \partial x_2} = \beta_{12} \cdot \phi(u) + (\beta_1 + \beta_{12} x_2) \cdot \frac{\partial}{\partial x_2} \phi(u) \]
\[ = \beta_{12} \cdot \phi(u) + (\beta_1 + \beta_{12} x_2) \cdot (\beta_2 + \beta_{12} x_1) \cdot \phi'(u) \]

Note: The marginal effect is not \( \beta_{12} \phi(u) \)
Note 2: Assume \( \beta_{12} = 0 \)

\[ \frac{\partial^2 E[y|x]}{\partial x_1 \partial x_2} = \beta_{12} \cdot \phi(u) + (\beta_1 + \beta_{12} x_2) \cdot (\beta_2 + \beta_{12} x_1) \cdot \phi'(u) \]
\[ = \beta_1 \beta_2 \phi'(u) \]

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 \( \beta_{12} \) 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

\[
\begin{align*}
E[y|x] &= \Phi(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3) \\
\frac{\partial^3 E[y|x]}{\partial x_1 \partial x_2 \partial x_3} &= \phi''(u) \cdot (\beta_2 + x_1 \beta_{12} + x_3 \beta_{23} + x_1 x_3 \beta_{123}) \cdot \\
&\quad (\beta_1 + x_2 \beta_{12} + x_3 \beta_{13} + x_2 x_3 \beta_{123}) + \\
&\phi'(u) \cdot (\beta_{12} + x_3 \beta_{123}) \cdot (\beta_3 + x_1 \beta_{13} + x_2 \beta_{23} + x_1 x_2 \beta_{123}) + \\
&\phi'(u) \cdot (\beta_{13} + x_2 \beta_{123}) \cdot (\beta_2 + x_1 \beta_{12} + x_3 \beta_{23} + x_1 x_3 \beta_{123}) + \\
&\phi'(u) \cdot (\beta_{23} + x_1 \beta_{123}) \cdot (\beta_1 + x_2 \beta_{12} + x_3 \beta_{13} + x_2 x_3 \beta_{123}) + \\
&\phi(u) \cdot \beta_{123}
\end{align*}
\]

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

\[
\log(o) = \log \left( \frac{p}{1-p} \right) = \log \left( \frac{e^{x' \beta}}{1 + e^{x' \beta}} \right) = \log \left( \frac{\frac{e^{x' \beta}}{1 + e^{x' \beta}}}{1 - \frac{e^{x' \beta}}{1 + e^{x' \beta}}} \right) = \\
\log \left( \frac{\frac{e^{x' \beta}}{1 + e^{x' \beta}}}{1 - \frac{e^{x' \beta}}{1 + e^{x' \beta}}} \right) = \log e^{x' \beta} = x' \beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots
\]

or

\[o = e^{\beta_0} \cdot e^{\beta_1 x_1} \cdot e^{\beta_2 x_2} \ldots\]

i.e. an increase of \(x_i\) by one unit increases the odds \(o\) by a factor of \(e^{\beta_i}\).

In particular if \(x_i\) is a binary variable \((x_i \in \{0, 1\})\)

\[
\frac{\Pr(Y=1|x_i=1)}{\Pr(Y=1|x_i=0)} = \frac{o(Y=1|x_i=1)}{o(Y=1|x_i=0)} \equiv o_{x_i} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i + \ldots} = e^{\beta_i}
\]
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

\[
\frac{\Pr(Y = 1| x_i = 1)}{1 - \Pr(Y = 1| x_i = 1)} \approx \frac{\Pr(Y = 1| x_i = 0)}{1 - \Pr(Y = 1| x_i = 0)}
\]

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

\[
\frac{0.9}{0.1} = 9 \quad \text{(risk-ratio)}
\]

\[
\frac{\Pr(Y = 1| x_i = 1)}{1 - \Pr(Y = 1| x_i = 1)} = \frac{0.9}{0.1} = 81 \quad \text{(odds-ratio)}
\]

### 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

\[
o_{x_1} = \frac{o(Y = 1| x_1 = 1)}{o(Y = 1| x_1 = 0)} = \frac{e^{\beta_1 + \beta_{12} x_2}}{e^{\beta_2 x_2}} = e^{\beta_1 + \beta_{12} x_2}
\]

hence

\[
o_{x_1| x_2=1} = \frac{o_{x_1| x_2=1}}{o_{x_1| x_2=0}} = \frac{e^{\beta_1 + \beta_{12}}}{e^{\beta_1}} = e^{\beta_{12}}
\]

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

\[
\exp(\text{coef}(lr))
\]

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>3.767015e-03</td>
<td>1.064459e+08</td>
<td></td>
</tr>
</tbody>
</table>

Confidence intervals can be calculated similarly:

\[
\exp(\text{confint}(lr))
\]

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5 %</td>
<td>97.5 %</td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>0.00007483879</td>
<td>3.998610e-02</td>
</tr>
<tr>
<td>x</td>
<td>56340.31979050605</td>
<td>4.418568e+13</td>
</tr>
</tbody>
</table>

Note: This interpretation is possible for the logistic link function only.
8.5 Literature


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 ($\alpha$ and $\beta$ are placeholders):

```r
for (i in 1:length(y)) { \alpha ; \beta }
for(k in 1:2) {b[k] ~ dnorm (0,.0001) }
```

Assume that $\alpha$ has the value $y[i] \sim dbern(p[i])$
What should be the value of $\beta$:  

• None of the following is correct.
  • $y[i] \sim \text{plogis}(b[1]+b[2]*x[i],0,1)$
  • $y[i] \leftarrow \text{plogis}(b[1]+b[2]*x[i],0,1)$
  • $p[i] \leftarrow \text{plogis}(b[1]+b[2]*x[i],0,1)$
  • $p[i] \sim \text{plogis}(b[1]+b[2]*x[i],0,1)$

**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 average woman? Use both the logit and probit specifications.
  2. What is the value of $f(x’\beta)$ 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) \( \text{wageh} = 7.5, \text{agew} = 42.5, \text{educw} = 12, \text{experience} = 10.5, \text{child6} = 3, \text{child618} = 1. \)

5. If you suspect that treating \( \text{child6} \) 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 \( \text{wageh} \) and \( \text{child618} \) are jointly significant? Use both the logit and probit specifications.

**Exercise 8.3** Let \( \text{grad} \) be a dummy variable for whether a student-athlete at a large university graduates in five years. Let \( \text{hsGPA} \) and \( \text{SAT} \) be high school grade point average and SAT score, respectively. Let \( \text{study} \) 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:

\[
\hat{P}(\text{grad} = 1|\text{hsGPA, SAT, study}) = \mathcal{L}(-1.17 + .24\text{hsGPA} + .00058\text{SAT} + .073\text{study})
\]

where \( \mathcal{L}(z) = \exp(z)/(1 + \exp(z)) \) is the logistic function.

Holding \( \text{hsGPA} \) fixed at 3.0 and \( \text{SAT} \) fixed at 1200, 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**

\[
\Pr(Y = 1|x) = F(X'\beta)
\]

\[
F^{-1}(\Pr(Y = 1|x)) = X'\beta
\]

\[
Y \sim \text{binom}(\frac{F(X'\beta)}{\text{family}}, \text{link})
\]

```r
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

\[ \Pr(Y = 1|x) = F(x'\beta) \]

Let \(-u \sim F\).
By definition: \(\Pr(Y = 1|x) = F(x'\beta) \equiv \Pr(x'\beta > -u) = \Pr(x'\beta + u > 0)\)

Note: we can normalise the variance of \(u\) to any number — we only have to adjust \(\beta\). \(y^1/\sigma = x'\beta/\sigma + u/\sigma\) is the same model.
Different \(F\) (probit, logit) have different \(\sigma\), hence different \(\beta\).

• Logistic distribution has standard deviation \(\pi/\sqrt{3}\).
• Normal distribution has standard deviation 1.

\[ \beta_{logit} \approx \frac{\pi}{\sqrt{3}} \cdot \beta_{probit} \]

\[
\text{glm}(\text{lfp}=="yes" \sim \text{lnnlinc}, \text{data}=\text{Participation}, \text{family}=\text{binomial}(\text{link}=\text{logit}))
\]

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

\[
\text{glm}(\text{lfp}=="yes" \sim \text{lnnlinc}, \text{data}=\text{Participation}, \text{family}=\text{binomial}(\text{link}=\text{probit}))
\]

Call: glm(formula = lfp == "yes" ~ lnnlinc, family = binomial(link = probit),
data = Participation)
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

$$L(\beta|y, X) = \prod_i (F(x_i' \beta))^{y_i} (1 - F(x_i' \beta))^{1-y_i}$$

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 $F(x) = \frac{e^x}{1 + e^x}$)

$$\frac{d \log L}{d \beta} = \sum_i \left( \frac{y_i f(x_i' \beta)}{F(x_i' \beta)} + (1 - y_i) \frac{-f(x_i' \beta)}{1 - F(x_i' \beta)} \right) x_i = \ldots$$

$$= \sum_i (y_i - F(x_i' \beta)) x_i = 0$$

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 $\beta_1$ is

$$E[\hat{\beta}_1] = \beta_1 + (X'_1 X_1)^{-1} X'_1 X_2 \beta_2$$

which is $\beta_1$ 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.

```r
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

\[
\text{Pr}(Y = 1|x) = F(X^\prime \beta) \\
F^{-1}(\text{Pr}(Y = 1|x)) = X^\prime \beta \\
Y \sim \text{binom} \left(\frac{F(X^\prime \beta)}{\text{family}}\right)
\]

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 \texttt{glm} module to make the sampler faster.

```r
library(runjags)
bayesL <- run.jags(model=modelL, modules="glm", inits=initJags, data=dataL, monitor=c("beta0", "beta1"))
plot(bayesL, var="beta1", plot.type=c("trace", "density"), layout=c(1,2))
```
The same for the probit model

```r
modelP <- 'model {
  for (i in 1:length(y)) {
```
\[ y[i] \sim \text{dbern}(p[i]) \]
\[
p[i] \leftarrow \text{pnorm}(\beta_0 + \beta_1 x[i], 0, 1) \]
\}
\begin{align*}
\beta_0 & \sim \text{dnorm}(0, 0.0001) \\
\beta_1 & \sim \text{dnorm}(0, 0.0001)
\end{align*}
}'

\begin{verbatim}
bayesP <- run.jags(model=modelP, modules="glm", inits=initJags, data=dataL, monitor=c("beta0", "beta1"))
\end{verbatim}

\begin{verbatim}
bayesP$summary$quantiles[, c("2.5\%", "97.5\%")]
\end{verbatim}

\begin{verbatim}
2.5\%      97.5\%
beta0 -0.1893351 -0.02152194 
beta1 -0.7910998 -0.35416947
\end{verbatim}

\begin{verbatim}
confint(glm(lfp ~ lnnlinc, family=binomial(link="probit"), data=Participation))
\end{verbatim}

\begin{verbatim}
2.5 \%      97.5 \%
(Intercept) 3.655146 8.3364864 
lnnlinc -0.790139 -0.3518155
\end{verbatim}

9.2 Count data

9.2.1 Model

Poisson process
(birth-only process):

- During one unit of time on average \( \lambda \) arrivals occur.
- During a time span of length \( \tau \) on average \( \tau \lambda \) arrivals occur.

Note that this process has no memory!

Be \( N(t) \) the number of arrivals until time \( t \)
\[
\Pr(N(t + \tau) - N(t) = y) = \frac{e^{-\lambda \tau} (\lambda \tau)^y}{y!}
\]

Poisson distribution:
\[
\Pr(Y_i = y_i | \lambda) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}
\]
Link function

It is convenient to assume a logarithmic link:

\[ \text{link } x_i' \beta \text{ to } \lambda_i: \quad \log \lambda_i = x_i' \beta \quad \Leftrightarrow \quad \lambda_i = e^{x_i' \beta} \quad (> 0) \]

\[ \Pr(Y_i = y_i|x_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} = \frac{e^{-e^{x_i' \beta}} (e^{x_i' \beta})^{y_i}}{y_i!} \]

\[ \log L = \sum_{i=1}^{n} -e^{x_i' \beta} + y_i x_i' \beta - \log y_i! \]

\[ \frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{n} (y_i - e^{x_i' \beta}) x_i = 0 \]

The Hessian is negative definite and Newton’s method converges quickly

9.2.2 Illustration — the Fair data

- \textit{nbaffairs} number of affairs in past year

We will assume that this variable follows a Poisson distribution.
library(MASS)
data(Fair)
table(Fair$nbaffairs)


<table>
<thead>
<tr>
<th>nbaffairs</th>
<th>number of affairs in past year</th>
</tr>
</thead>
<tbody>
<tr>
<td>rate</td>
<td>self rating of marriage, from 1 (very unhappy) to 5 (very happy)</td>
</tr>
<tr>
<td>religious</td>
<td>how religious, from 1 (anti) to 5 (very)</td>
</tr>
<tr>
<td>ym</td>
<td>number of years married</td>
</tr>
</tbody>
</table>

```
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)

Deviance Residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-4.2037</td>
<td>-1.5983</td>
<td>-1.1287</td>
<td>-0.7448</td>
<td>7.4076</td>
</tr>
</tbody>
</table>

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:

cconfint(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:

```r
summary(est0 <- glm(nbaffairs ~ 1, family = poisson(link = log), data = Fair))
```

Call:
glm(formula = nbaffairs ~ 1, family = poisson(link = log), data = Fair)

Deviance Residuals:
  Min 1Q Median 3Q Max
-1.706 -1.706 -1.706 -1.706 5.435

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))
```

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>7</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>451</td>
<td>34</td>
<td>17</td>
<td>19</td>
<td>42</td>
<td>38</td>
</tr>
<tr>
<td>nba</td>
<td>140</td>
<td>204</td>
<td>149</td>
<td>72</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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.
Poisson distribution: Pois(λ)

Mean: λ
Variance: λ

Negative binomial distribution: NB(μ, θ)

Mean: μ
Variance: μ + μ²/θ

Poisson is a special case of NB:

$$\lim_{{\theta \to \infty}} NB(\mu, \theta) = Pois(\mu)$$

$$\mu = 1.46:$$

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","NB \theta=2","NB \theta=0.1")),
    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)

Deviance Residuals:
Min 1Q Median 3Q Max
-0.7688 -0.7688 -0.7688 -0.7688 1.0476

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

\[ \theta = 0.1120 \]
\[ \text{Std. Err.} = 0.0119 \]
\[ 2 \times \log\text{-likelihood: } -1502.3460 \]

\[ \mu = 1.46, \theta = 0.112: \]

```
fTab \leftarrow \text{within}(fTab, nb \leftarrow \text{dnbinom}(nbaffairs, mu = \exp(\text{coef}(est.nb)), size = \text{est.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)), ylab = "Density", scales = list(y = list(log = 10)), yscale.components = yscale.components.log10ticks)
```

```
summary(est.nb \leftarrow \text{glm.nb}(nbaffairs ~ rate + religious + ym, data = Fair))
```

Call:
```
\text{glm.nb(formula = nbaffairs ~ rate + religious + ym, data = Fair,}
\hspace{1cm} \text{init.theta = 0.1418819265, link = log})
```

Deviance Residuals:
```
          Min       1Q   Median       3Q      Max
-1.1073 -0.8034  -0.6884  -0.4537  2.4451
```

Coefficients:
```
                Estimate Std. Error z value  Pr(>|z|)
(Intercept) -1.423180  0.159686  -8.953  1.81e-19
rate        0.620725  0.014306  43.141  < 2e-16
religious   0.162186  0.033583   4.825  1.53e-06
ym          0.447673  0.033583  13.343  < 2e-16
```
(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:

\[ 2(\log L_1 - \log L_0) \overset{\text{approx.}}{\sim} X_{k_1-k_0}^2 \]

\texttt{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

\texttt{reg.model<-'model \{ for (i in 1:length(y)) \{ y[i] \sim \text{dnorm}(beta0 + beta1*x[i],tau) \} beta0 \sim \text{dnorm}(0,0.0001) beta1 \sim \text{dnorm}(0,0.0001) tau \sim \text{dgamma}(0.01,0.01) \}'}

Bayesian Poisson
```r
pois.model <- 'model {
    for (i in 1:length(y)) {
        y[i] ~ dpois(lambda[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:
\[ \text{NB}_{\text{glm}}(\mu, \theta) = \text{NB}_{\text{JAGS}}(p, \theta) = \text{NB}(\mu, \sigma^2) = \ldots \]

\[ p = \frac{\theta / \theta + \mu}{\theta} \]

\[ \sigma^2 = \mu + \frac{\mu^2}{\theta} \]

```r
negbin.model <- 'model {
  for(i in 1:length(y)) {
    y[i] ~ dnegbin(p[i],theta)
    p[i] <- theta/(theta+mu[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)]

<table>
<thead>
<tr>
<th></th>
<th>2.5%</th>
<th>50%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta[1]</td>
<td>-0.1639</td>
<td>0.063</td>
<td>0.3081</td>
</tr>
<tr>
<td>beta[2]</td>
<td>-0.6444</td>
<td>-0.43</td>
<td>-0.22</td>
</tr>
<tr>
<td>beta[3]</td>
<td>-0.6637</td>
<td>-0.44</td>
<td>-0.22</td>
</tr>
<tr>
<td>beta[4]</td>
<td>0.0381</td>
<td>0.09</td>
<td>0.13</td>
</tr>
<tr>
<td>theta</td>
<td>0.1109</td>
<td>0.138</td>
<td>0.17</td>
</tr>
</tbody>
</table>

confint(est.nb)

<table>
<thead>
<tr>
<th></th>
<th>2.5 %</th>
<th>97.5 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.253</td>
<td>3.547</td>
</tr>
<tr>
<td>rate</td>
<td>-0.64</td>
<td>-0.22</td>
</tr>
<tr>
<td>religious</td>
<td>-0.65</td>
<td>-0.23</td>
</tr>
<tr>
<td>ym</td>
<td>0.04</td>
<td>0.13</td>
</tr>
</tbody>
</table>

est.nb["theta"]

[1] 0.1418819

gelman.diag(negbin.jags)

Potential scale reduction factors:

<table>
<thead>
<tr>
<th></th>
<th>Point est. Upper C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta[1]</td>
<td>1</td>
</tr>
</tbody>
</table>
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:

\[ \begin{align*}
\eta_1 &= x' \beta_1 + \xi_1 \\
\eta_2 &= x' \beta_2 + \xi_2 \\
\eta_3 &= x' \beta_3 + \xi_3 \\
\vdots & \end{align*} \]
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 $\xi_j$ are i.i.d. we often normalise their variance to a convenient value.
  - different distributions for $\xi$ 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 $v_{ki}$ follows a logistic distribution

$$
\Pr(y = k|\xi_k) = \prod_{j \neq k} \frac{F_{\text{Gumbel}}(x'(\beta_k - \beta_j) + \xi_k)}{\Pr(n_j < n_k)}
$$

average over $\xi_k$

$$
\Pr(y = k) = \int f_{\text{Gumbel}}(\xi_k) \prod_{j \neq k} F_{\text{Gumbel}}(x'(\beta_k - \beta_j) + \xi_k) \, d\xi_k
$$

$$
\Pr(y = k) = \frac{e^{x'\beta_k}}{\sum_{i=1}^{m} e^{x'\beta_i}}
$$

→ we get the following multinomial logit (McFadden)

$$
\begin{align*}
\Pr(y = 1) &= \frac{e^{x'\beta_1}}{\sum_{k=1}^{m} e^{x'\beta_k}} \\
\Pr(y = 2) &= \frac{e^{x'\beta_2}}{\sum_{k=1}^{m} e^{x'\beta_k}} \\
\Pr(y = 3) &= \frac{e^{x'\beta_3}}{\sum_{k=1}^{m} e^{x'\beta_k}} \\
& \vdots
\end{align*}
$$

• $0 < \Pr(y = k) < 1$

• $\sum_k \Pr(y = k) = 1$

↑ $\beta_k$ are not identified

Normalise:

$$
\begin{align*}
\Pr(y = 1) &= \frac{1}{1 + \sum_{k=2}^{m} e^{x'\beta_k}} \\
\Pr(y = 2) &= \frac{e^{x'\beta_2}}{1 + \sum_{k=2}^{m} e^{x'\beta_k}} \\
\Pr(y = 3) &= \frac{e^{x'\beta_3}}{1 + \sum_{k=2}^{m} e^{x'\beta_k}} \\
& \vdots
\end{align*}
$$
the odds ratios are:

\[
\frac{\Pr(y = k)}{\Pr(y = 1)} = e^{x'\beta_k}
\]

This is a strong assumption on the error terms.

**Independence from irrelevant alternatives — IIA**

Example:

- Dependent = choice of travel mode
- unobservable = personal preference for/against means of mass transportation (tube/train).

\[
\frac{\Pr(y = \text{tube})}{\Pr(y = 1)} = e^{x'\beta_{\text{tube}}}
\]

\[
\frac{\Pr(y = \text{train})}{\Pr(y = 1)} = e^{x'\beta_{\text{train}}}
\]

- \(\rightarrow\) choices/error terms are correlated.

\(\rightarrow\) multinomial logit can represent systematic variation of choices (explained by observed characteristics) but not systematic individual (unobserved) variation of choices.

**The log-likelihood:**

With \(I_k(y_i) = \begin{cases} 1 & \text{if } y_i = i \\ 0 & \text{otherwise} \end{cases}\)

\[
\log L = \sum_i \sum_{k=1}^m I_k(y_i) \log \Pr(y_i = k) = \sum_i \sum_{k=1}^m I_k(y_i) \log \frac{e^{x_i'\beta_k}}{1 + \sum_{k=2}^m e^{x_i'\beta_k}}
\]

this function \(\log L\) is globally concave in \(\beta\) (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 (\(\text{mat}\)) to generate the choices but the estimator gives us a different set of parameters back (\(\text{coef(est)}\)). We also see how these two sets of parameters are related.

Let us first create individual explanatory variables, \(x1, x2\).

```r
N <- 100
sd <- 10
ex <- cbind(x1 = runif(N), x2 = runif(N))
head(ex)
```
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 1 3 2 1 2 1 1 3 3 1 3 3 1 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 3 1 3 1 3 1 3 2
[65] 1 2 2 1 1 1 2 1 2 3 2 2 3 3 3 1 3 3 2 1 1 2 3 2 1 3 2 3 3
[97] 1 3 1 2
```

```
choice <- max.col(latent)
library(nnet)
est <- multinom(choice ~ x1 + x2, as.data.frame(ex))
```

```
# weights: 12 (6 variable)
initial value 109.861229
iter 10 value 18.323213
iter 20 value 16.923568
iter 30 value 16.881715
iter 40 value 16.880637
iter 50 value 16.880332
```
iter  60 value 16.880044
iter  70 value 16.879931
final  value 16.879896
converged

est

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 em-
ployed above. However, they are a projection. We are expecting this:

mat

    [,1] [,2]
[1,]  400  0
[2,]  250  200
[3,]  100  300

but we got that:

coef(est)

   (Intercept)    x1     x2
2  0.9444688 -25.80588  31.72050
3  0.1557040 -58.59718  52.66552

The estimator normalises the first category to zero

mat

    [,1] [,2]
[1,]   0   0
[2,] -150  200
[3,] -300  300

mat - cbind(c(1,1,1)) %*% mat[1,]
and sets the variance to one:

\[
(mat - cbind(c(1,1,1)) \times mat[1,]) \times \pi / \sqrt{6} / 10
\]

\[
[,1] [,2] 
[1,] 0.00000 0.00000 \\
[2,] -19.23825 25.65100 \\
[3,] -38.47649 38.47649
\]

To access estimation results we have the usual extractor functions:

```r
coef(est)
```

<table>
<thead>
<tr>
<th></th>
<th>x1</th>
<th>x2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.944688</td>
<td>-25.80588</td>
</tr>
<tr>
<td>3</td>
<td>0.1557040</td>
<td>-58.59718</td>
</tr>
</tbody>
</table>

```r
confint(est)
```

<table>
<thead>
<tr>
<th></th>
<th>x1</th>
<th>x2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5 %</td>
<td>-43.459195</td>
<td>-8.152558</td>
</tr>
<tr>
<td>97.5 %</td>
<td>11.420190</td>
<td>52.020819</td>
</tr>
<tr>
<td>2.5 %</td>
<td>-89.59278</td>
<td>-27.601578</td>
</tr>
<tr>
<td>97.5 %</td>
<td>26.23746</td>
<td>79.093575</td>
</tr>
</tbody>
</table>

### 9.3.3 Bayesian multinomial

```r
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%")]

<table>
<thead>
<tr>
<th></th>
<th>2.5%</th>
<th>50%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta[1,2]</td>
<td>-3.279832</td>
<td>1.180415</td>
<td>8.111798</td>
</tr>
<tr>
<td>beta[2,2]</td>
<td>-75.747138</td>
<td>-34.115950</td>
<td>-17.223420</td>
</tr>
<tr>
<td>beta[3,2]</td>
<td>21.681665</td>
<td>42.334950</td>
<td>86.289170</td>
</tr>
<tr>
<td>beta[1,3]</td>
<td>-5.593608</td>
<td>0.042870</td>
<td>7.799049</td>
</tr>
<tr>
<td>beta[2,3]</td>
<td>-119.114450</td>
<td>-78.265450</td>
<td>-47.038998</td>
</tr>
<tr>
<td>beta[3,3]</td>
<td>42.391963</td>
<td>71.952100</td>
<td>110.543175</td>
</tr>
</tbody>
</table>

confint(est)

<table>
<thead>
<tr>
<th></th>
<th>2.5 %</th>
<th>97.5 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-3.098834</td>
<td>4.987771</td>
</tr>
<tr>
<td>x1</td>
<td>-43.459195</td>
<td>-8152558</td>
</tr>
<tr>
<td>x2</td>
<td>11.420190</td>
<td>52.020819</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>2.5 %</th>
<th>97.5 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-4.74507</td>
<td>5.056478</td>
</tr>
<tr>
<td>x1</td>
<td>-89.59278</td>
<td>-27.601578</td>
</tr>
<tr>
<td>x2</td>
<td>26.23746</td>
<td>79.093575</td>
</tr>
</tbody>
</table>

9.4 Literature


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):

```r
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.
• \( y[i] \leftarrow \text{dpois}(p[i]) \)
• \( y[i] \sim \text{dlogis}(p[i]) \)
• \( y[i] \sim \text{dprobit}(p[i]) \)
• \( y[i] \sim \text{dpois}(p[i]) \)

**Example 2**  
What should in this case be the value of \( \beta \)?

• None of the following is correct.
• \( p[i] \leftarrow \exp(b[1]+b[2]*x) \)
• \( p[i] \sim \exp(b[1]+b[2]*x) \)
• \( p[i] \leftarrow \log(b[1]+b[2]*x) \)
• \( p[i] \sim b[1]+b[2]*x \)

**Example 3**  
You estimate the following multinomial logit model:

\[
\begin{align*}
\eta_1 &= x' \beta_1 + \xi_1 \\
\eta_2 &= x' \beta_2 + \xi_2 \\
\eta_3 &= x' \beta_3 + \xi_3
\end{align*}
\]

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 (\( \alpha_1 \) and \( \alpha_2 \) are placeholders):

```r
multinom(formula = y \sim x1+x2-1)  
Coefficients:
   x1  x2
  2  1 -2
  3 \alpha_1 \alpha_2
```

• What do you expect for \( \alpha_1 \)?
• What do you expect for \( \alpha_2 \)?
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 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 `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:

\[
\begin{align*}
\Pr(y_i = 1) &= \Pr(\kappa_0 < x_i'\beta + u \leq \kappa_1) \\
\Pr(y_i = 2) &= \Pr(\kappa_1 < x_i'\beta + u \leq \kappa_2) \\
\Pr(y_i = 3) &= \Pr(\kappa_2 < x_i'\beta + u \leq \kappa_3) \\
&\quad \vdots
\end{align*}
\]
or (solving for $u$)

$$\Pr(y_i = 1) = \Pr(\kappa_0 - x_i'\beta < u \leq \kappa_1 - x_i'\beta)$$
$$\Pr(y_i = 2) = \Pr(\kappa_1 - x_i'\beta < u \leq \kappa_2 - x_i'\beta)$$
$$\Pr(y_i = 3) = \Pr(\kappa_2 - x_i'\beta < u \leq \kappa_3 - x_i'\beta)$$

: 

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

\[
\begin{align*}
\Pr(y_i = 1) &= \Pr(\kappa_0 - x_i'\beta < u \leq \kappa_1 - x_i'\beta) \\
\Pr(y_i = 2) &= \Pr(\kappa_1 - x_i'\beta < u \leq \kappa_2 - x_i'\beta) \\
\Pr(y_i = 3) &= \Pr(\kappa_2 - x_i'\beta < u \leq \kappa_3 - x_i'\beta) \\
\end{align*}
\]

\[
\log L = \sum_i \sum_{k=1}^m I_k(y_i) \log Pr(y_i = k)
\]

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 marriage (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.

```r
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...

```r
(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:

```r
(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 $\kappa_i$:

```r
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")
```

![Graph illustrating the estimated thresholds $\kappa_i$.](image)
- Dependent variable $y[i]$
- Latent variable $t[i]$
- Independent variable $x[i]$
- Parameters $\beta$, $\kappa[j]$

**JAGS notation for intervals**

$$y[i] \sim \text{dinterval}(t[i], \kappa)$$

where

- $t[i]$ realisation of latent variable
- $y[i]$ observable rating
- $\kappa$ thresholds

If $Y \sim \text{dinterval}(t, \kappa)$ then

- $Y = 0$ if $t \leq \kappa[1]$
- $Y = m$ if $\kappa[m] < t \leq \kappa[m + 1]$ for $1 \leq m < M$
- $Y = M$ if $\kappa[M] < t$

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))
bayesO <- run.jags(model = model0, data = dataList, inits = list(initList, initList),
  monitor = c("beta", "kappa"))

bayesO$summary$quantiles[, c("2.5\%", "50\%", "97.5\%")]

2.5% 50% 97.5%
beta -0.06249314 -0.04941700 -0.01508178
kappa[1] -2.67677300 -2.40821000 -1.97988975
kappa[2] -1.69587550 -1.54138500 -1.09638975
kappa[3] -1.12315075 -0.97184000 -0.54379553
kappa[4] -0.21635610 -0.09625535 0.28168908

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

bayesO

JAGS model summary statistics from 20000 samples (chains = 2; adapt+burnin = 5000):
### 9.C.3 Illustration II — a simulated dataset

In the following we generate a latent variable \textit{latent}. The problem is that we can not observe \textit{latent}. We only see \textit{y} which indicates a range of possible values for \textit{latent}. These ranges are defined by \textit{cuts}.

```r
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.
polr estimates the ordered model. By default the logistic method is used.

\[
\text{polr(factor}(y) - x)\]

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

\[
b/\text{sd}\times\pi/\sqrt{3}
\]

\[
[1] 0.3627599
\]

and as intercepts
which is both close to the estimated coefficients. The standard normal distribution has $\sigma = 1$, hence we get different results with the probit method:

```r
cuts/sd*pi/sqrt(3)
```

```
```

As the coefficient for $x$ we should expect $b/sd$

```
[1] 0.2
```

and as intercepts

```r
cuts/sd
```

```
[1] 4 8 10 11
```

### 10 Mixed effects models

#### 10.1 Terminology

**Fixed-Effects, Random-Effects, Mixed-Models**

FE-only-model:

$$y_j = \beta X_j + u_j \quad \text{with } u_j \sim N(0, \sigma^2 I_n)$$

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.

$$y_{ij} = \beta X_{ij} + \gamma_i Z_{ij} + u_{ij} \quad \text{with } \gamma_i \sim N(0, \Psi), u_{ij} \sim N(0, \Sigma_i)$$
• Fixed effects: The researcher is interested in specific effects $\beta$. The researcher does not care/is ignorant about the type of distribution of $\beta$. If $\beta$ is a big object, estimation can be expensive.

• Random effects: The researcher only cares about the distribution of $\gamma$. It is sufficient to estimate only the parameters of the distribution of $\gamma$ (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 equilibrium,...)
  – 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 $\epsilon$

Solution A (inefficient):

- 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 . . . 12
- k group 1 . . . 10
- t time 1 . . . 8
- Pooled OLS (pretend $\epsilon_{ikt}$ 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 $\nu_{ik}$

+ unbiased if $\nu_{ik}$ 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 $\nu_{ik}$ and $X$ (if it can be assumed)

– biased if $\nu_{ik}$ 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 $\beta_0$ 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

```r
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.

```r
simple <- as.data.frame(cbind(x=c(1,2,3,4),y=c(3,0,6,6.8744),i=c(1,1,2,2)))
```

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>i</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.0000</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0.0000</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>6.0000</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>6.8744</td>
<td>2</td>
</tr>
</tbody>
</table>
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:**
\[ y_{ik} = \beta_0 + \beta_1 x_{ik} + \epsilon_{ik} \text{ with } \epsilon_{ik} \sim N(0, \sigma) \]

\[
\text{ol <- lm(y ~ x, data=simple)}
\]
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

```r
  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

```r
  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   NA   NA   NA
          x       2.469   NA   NA   NA

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
3.9689034871 -0.0001213948

ranef(ranef)
with conditional variances for "i"

coeff(ranef)$i

(Intercept) x
1 1.765852 -0.0001213948
2 6.171955 -0.0001213948

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)  3.9689035    5.0114522    0.792
       x       -0.0001214    1.6197091   0.000
Correlation of Fixed Effects:

(Intr)  
\[
x = -0.808
\]

Here is a picture of the different estimated regression lines:

```r
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 ($\sigma_v = 0$), all the variance goes into $u$.
Fixed effects puts no restriction on $\sigma_v$, hence $\sigma_u$ can be minimal.
Random effects is in between.

\[
\begin{align*}
\text{between OLS} & \quad y_i = \beta_0 + \beta_1 x_i + \nu_i \\
\text{pooled OLS} & \quad y_{ik} = \beta_0 + \beta_1 x_{ik} + \nu_i + \epsilon_{ik} \\
\text{mixed effects} & \quad y_{ik} = \beta_0 + \beta_1 x_{ik} + \nu_i + \epsilon_{ik} \\
\text{fixed effects} & \quad y_{ik} = \beta_0 + \beta_1 x_{ik} + \sum_i \gamma_i d_i + \epsilon_{ik}
\end{align*}
\]

<table>
<thead>
<tr>
<th></th>
<th>(\nu_i)</th>
<th>(\epsilon_{ik})</th>
</tr>
</thead>
<tbody>
<tr>
<td>between OLS</td>
<td>finitely expensive</td>
<td>cheap (no cost)</td>
</tr>
<tr>
<td>pooled OLS</td>
<td>0 (infinitely expensive)</td>
<td>finitely expensive</td>
</tr>
<tr>
<td>mixed effects</td>
<td>finitely expensive</td>
<td>finitely expensive</td>
</tr>
<tr>
<td>fixed effects</td>
<td>cheap (no cost)</td>
<td>finitely expensive</td>
</tr>
</tbody>
</table>

**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 \(\nu_i\) (setting them to 0) and, under this constraint, minimizes \(\epsilon_{ikt}\).

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 \(\nu_i\) and \(\epsilon_{ikt}\)
- If coefficients were close to the fixed effects model, then \(\nu_i\) are large (in absolute terms) and the \(\epsilon_{ikt}\) are small.
- If the coefficients were close to the OLS model, then \(\nu_i\) are very small but the \(\epsilon_{ikt}\) 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

```r
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 \( y_2 \) without an individual specific random effect.

```r
y2 <- x + 6*uerr
```

We put them all in one dataset.

```r
data <- data.frame(cbind(y,y2,x,i,ierr,uerr))
```

10.3.2 Pooled OLS

\[ y_{it} = \beta_0 + \beta_1 x_{it} + \epsilon_{it} \quad \text{with} \quad \epsilon_{ik} \sim N(0, \sigma) \]

```r
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 $\beta$ is consistent if residuals $\epsilon_{it}$ are uncorrelated with $X$.
- With repeated observations (as in our case), estimation of $\Sigma_{\beta\beta}$ is generally not consistent.

10.3.3 Between OLS

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{with } \nu_i \sim N(0, \sigma)$$

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:

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9423</td>
<td>-16.8802</td>
<td>-5.4127</td>
<td>8.0874</td>
<td>9.9221</td>
</tr>
</tbody>
</table>

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 $\beta$ is consistent if residuals ($\nu_i$) are uncorrelated with $X$.
- $\Sigma_{\beta\beta}$ can not be estimated.
- The method is inefficient since the variance within a group is not exploited.
### 10.3.4 Fixed effects

\[ y_{it} = \beta_0 + \beta_1 x_{it} + \sum_i \gamma_i d_i + \epsilon_{it} \]

```r
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 -2.8790      0.5033  -5.720  2.62e-08 ***
 as.factor(i)2 -20.6976     0.5053 -40.962 < 2e-16 ***
 as.factor(i)3 -9.2534      0.4936  -18.747 < 2e-16 ***
 as.factor(i)4  4.2317      0.4864   8.700  2.44e-16 ***
 as.factor(i)5  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 \( \beta \) is consistent if residuals (\( \epsilon_{it} \)) are uncorrelated with \( X \). This is a weaker requirement, since, with fixed effects, residuals are only \( \epsilon_{ikt} \), not \( v_i \).

- 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

\[ y_{it} = \beta_0 + \beta_1 x_{it} + v_i + \epsilon_{it} \]

```r
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:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-2.61559</td>
<td>-0.74732</td>
<td>0.02794</td>
<td>0.75531</td>
<td>2.44665</td>
</tr>
</tbody>
</table>

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>(Intercept)</td>
<td>97.861</td>
<td>9.892</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>8.442</td>
<td>2.905</td>
</tr>
</tbody>
</table>

Number of obs: 300, groups: i, 6

Fixed effects:

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-3.8219</td>
<td>4.0525</td>
<td>-0.943</td>
</tr>
<tr>
<td>x</td>
<td>1.0626</td>
<td>0.5763</td>
<td>1.844</td>
</tr>
</tbody>
</table>

Correlation of Fixed Effects:

<table>
<thead>
<tr>
<th>(Intr)</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.072</td>
</tr>
</tbody>
</table>

- Estimation of $\beta$ is consistent if $\nu_i$ and $\epsilon_{it}$ are uncorrelated with $X$. 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:

$$y_{it} = \beta_0 + \nu_i + \beta_1 x_{it} + \epsilon_{it}$$

\texttt{lmer( y ~ x + (1|g) )}

Random effect for intercept and slope:

$$y_{it} = \beta_0 + \nu_i + (\beta_1 + \nu'_i)x_{it} + \epsilon_{it}$$

\texttt{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):

<table>
<thead>
<tr>
<th></th>
<th>Lower95</th>
<th>Median</th>
<th>Upper95</th>
<th>Mean</th>
<th>SD</th>
<th>Mode</th>
<th>MCerr</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta[1]</td>
<td>-8.0354</td>
<td>-0.1792</td>
<td>11.116</td>
<td>0.40936</td>
<td>4.74</td>
<td>-1.07</td>
<td>0.765</td>
</tr>
<tr>
<td>beta[2]</td>
<td>-0.0442</td>
<td>1.061</td>
<td>2.211</td>
<td>1.0635</td>
<td>0.58</td>
<td>1.038</td>
<td>0.002</td>
</tr>
<tr>
<td>sd[1]</td>
<td>2.6825</td>
<td>2.9086</td>
<td>3.152</td>
<td>2.912</td>
<td>0.29</td>
<td>2.897</td>
<td>0.006</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>MC%ofSD</th>
<th>SS eff</th>
<th>AC.10</th>
<th>psrf</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta[1]</td>
<td>16.1</td>
<td>38</td>
<td>0.98046</td>
<td>1.086</td>
</tr>
<tr>
<td>beta[2]</td>
<td>0.5</td>
<td>39704</td>
<td>0.00003</td>
<td>1.000</td>
</tr>
<tr>
<td>sd[1]</td>
<td>0.5</td>
<td>38425</td>
<td>-0.00096</td>
<td>1</td>
</tr>
<tr>
<td>sd[2]</td>
<td>1.8</td>
<td>3127</td>
<td>0.12393</td>
<td>1.021</td>
</tr>
</tbody>
</table>

Total time taken: 1.7 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:

```r
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**

```r
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 \( \epsilon_{ikt} \)
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

\[
2 \cdot (\log(L_{\text{large}}) - \log(L_{\text{small}})) \sim \chi^2_k \quad \text{with} \quad k = df_{\text{large}} - df_{\text{small}}
\]

here

\[
2 \cdot (\log(L_{\text{RE}}) - \log(L_{\text{OLS}})) \sim \chi^2_k \quad \text{with} \quad k = 1
\]

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.

```r
mixedML <- update(mixed, REML=FALSE)
```

```r
teststat <- 2*(logLik(mixedML)-logLik(ols))
```

\[
\text{teststat} \approx \chi^2 \quad \text{unless we are at the boundary of the parameter space.}
\]

Testing \( \sigma_v^2 = 0 \) we are at the boundary. Nevertheless...

```r
plot(function(x) dchisq(x,1),0,2,ylab="$\chi^2$")
```
The p-value of the $\chi^2$-test would be

```r
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:

```r
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 $\chi^2$ distribution:

```r
plot(ecdf(dev[, "LL", 1]), 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)
```
• The assumption of a $\chi^2$ distribution is rather conservative.

• If we manage to reject $H_0$ (that $\sigma^2 = 0$) based on the $\chi^2$ 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 $\chi^2$ 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:

```r
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 $\beta$ 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    3.839424 0.5950676

\[
\text{bias} = \bar{\theta}_\text{BS} - \hat{\theta}
\]

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:

<table>
<thead>
<tr>
<th>Level</th>
<th>Normal</th>
<th>Basic</th>
<th>Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>95%</td>
<td>(-0.105, 2.227)</td>
<td>(-0.149, 2.224)</td>
<td>(-0.098, 2.274)</td>
</tr>
</tbody>
</table>

Calculations and Intervals on Original Scale

Here is a plot of the different bootstrapped estimates:

```r
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:

```
bayes$m$summary$m$quantiles[2,]
```

<table>
<thead>
<tr>
<th></th>
<th>2.5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5%</td>
<td>-0.07416123</td>
<td>0.67500175</td>
<td>1.06134000</td>
<td>1.45387750</td>
<td>2.18903050</td>
</tr>
</tbody>
</table>

10.5 Literature

Literature

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. \( \epsilon_{it} \) is the residual, \( \nu_i \) is the individual specific random effect, \( \nu_j' \) 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 (\( \alpha \) is a placeholder):

```r
for(k in 1:length(y)){
}
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 \( \alpha \) ?

- \( nu[k] + nu2[k] \)
- \( nu[i[k]] + nu2[j[i[k]]] \)
- \( nu[j[i[k]]] + nu2[j[i[k]]] \)
- \( nu[i[k]] + nu2[i[j[k]]] \)
- other value

Example 2  You estimated a mixed effects model with \texttt{lmer} and you have stored the result in the variable \texttt{mer}. You have also estimated the same model, though without the random effects, with \texttt{lm} and stored the result in the variable \texttt{ols}. To find out whether the mixed effects model is justified, you run the following approximate permutation test:
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, \( \nu_i \) is the individual specific random effect. You use the following model in JAGS (\( \alpha \) is a placeholder):

\[
\text{for(k in 1:length(y)){ } for(j in 1:max(i)){nu[j]~dnorm(0,tau[2])} for(k in 1:2){beta[k]~dnorm(0,.0001);tau[k]~dgamma(.01,.01)}}
\]

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 \( \alpha \)?

- \( y[k] \sim \text{dnorm}(\text{beta}[1]+\text{beta}[2]\times x[k]+\nu[k],\text{tau}[1]) \)
- \( y[k] \sim \text{dnorm}(\text{beta}[1]+\text{beta}[2]\times x[k]+\nu[i[k]],\text{tau}[1]) \)
- \( y[k] \sim \text{dnorm}(\text{beta}[1]+\text{beta}[2]\times x[k]+i[\nu[k]],\text{tau}[1]) \)
- \( y[k] \sim \text{dnorm}(\text{beta}[1]+\text{beta}[2]\times x[k]+i[k],\text{tau}[1]) \)
- 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:

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Ch.2</th>
<th>Df</th>
<th>Pr(&gt; Ch.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>m.ols</td>
<td>3</td>
<td>319</td>
<td>327</td>
<td>-156</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>m.mer</td>
<td>4</td>
<td>317</td>
<td>327</td>
<td>-154</td>
<td>4.21</td>
<td>1</td>
<td>0.0401</td>
</tr>
</tbody>
</table>

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

\[
E(u_i | X_i = x) = 0 \quad \rightarrow \quad X \not\perp u
\]

The problem:

\[
Y = X\beta + u \quad \text{but} \quad X \not\perp u
\]

Solution (Wright, 1928), use instrument \(Z\) (we “instrument” \(X\) with \(Z\)):

- \(Z\) is relevant: \(\text{cor}(Z, X) \neq 0\) \(\quad (Z \not\perp X)\)
- \(Z\) is exogeneous: \(\text{cor}(Z, u) = 0\) \(\quad (Z \perp u)\)

1st stage: \(X = Z\gamma + \nu\) \(\quad \) (works, since \(Z \not\perp X\))
\[
\rightarrow \quad \hat{X} = Z\gamma
\]

2nd stage: \(Y = \hat{X}\beta + u\) \(\quad \) (works, since \(Z \perp u\), and, hence \(\hat{X} \perp u\))

Here we construct \(X\) as correlated with \(u\).

- \(Z\) is relevant, \(\text{cor}(Z, X) \neq 0\).
- \(Z\) is exogeneous, \(\text{cor}(Z, u) = 0\).

\[
\text{set.seed(123)}
\]
\[
N \leftarrow 100
\]
\[
u \leftarrow \text{rnorm}(N)
\]
\[
Z \leftarrow \text{rnorm}(N) \quad \# Z \text{ is exogeneous}
\]
\[
X \leftarrow Z + 0.5*\text{rnorm}(N) - u \quad \# X \text{ is endogenous, } Z \text{ is relevant}
\]
\[
Y \leftarrow X + u
\]

When we estimate

\[
Y = \beta_0 + \beta_1 X + u
\]

we should expect \(\hat{\beta}_1 \approx 1\).

Naïve model: ignore that \(X \not\perp u\)
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:

```r
cor(X, residuals(lm(Y~X)))
```

```
[1] 1.383238e-17
```
We can do the same exercise with Bayes:

```
instNaive.model <- 'model {
  for(i in 1:length(y)) {
  }
  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)
```

JAGS model summary statistics from 40000 samples (chains = 4; adapt+burnin = 5000):

<table>
<thead>
<tr>
<th></th>
<th>Lower95</th>
<th>Median</th>
<th>Upper95</th>
<th>Mean</th>
<th>SD</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta[1]</td>
<td>-0.10248</td>
<td>0.03218</td>
<td>0.16471</td>
<td>0.032067</td>
<td>0.068043</td>
<td>0.033791</td>
</tr>
<tr>
<td>beta[2]</td>
<td>0.48975</td>
<td>0.58024</td>
<td>0.66977</td>
<td>0.57992</td>
<td>0.045798</td>
<td>0.57934</td>
</tr>
<tr>
<td>tau</td>
<td>1.6405</td>
<td>2.2258</td>
<td>2.8898</td>
<td>2.2405</td>
<td>0.32111</td>
<td>2.1891</td>
</tr>
</tbody>
</table>
Instrument \( Z \not\Rightarrow u \)

**Problem:**

\[
X \text{ is endogeneous: } \text{cor}(X, u) \neq 0 \quad (X \not\Leftrightarrow u)
\]

To solve the problem, we use \( Z \) as an instrument:

1st stage: \( X = Z Y + \nu \) \quad (works, since \( Z \not\Leftrightarrow X \))

\[
\hat{X} = Z Y
\]

2nd stage: \( Y = \hat{X} \beta + u \) \quad (works, since \( Z \not\Rightarrow u \), and, hence \( \hat{X} \not\Rightarrow u \))

Remember:

\[
Z \text{ is relevant: } \text{cor}(Z, X) \neq 0 \quad (Z \not\Leftrightarrow X)
\]

\[
Z \text{ is exogeneous: } \text{cor}(Z, u) = 0 \quad (Z \not\Leftrightarrow u)
\]

We first use the Gibbs sampler to estimate this model:

```r
inst.model <- 'model {
  for(i in 1:length(y)) {
    x[i] ~ dnorm( xHat[i], tau[2]) # 1st stage
    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):

<table>
<thead>
<tr>
<th></th>
<th>Lower95</th>
<th>Median</th>
<th>Upper95</th>
<th>Mean</th>
<th>SD</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta[1]</td>
<td>-0.14196</td>
<td>0.085526</td>
<td>0.33267</td>
<td>0.087691</td>
<td>0.12105</td>
<td>0.080594</td>
</tr>
<tr>
<td>beta[2]</td>
<td>0.75038</td>
<td>1.2207</td>
<td>0.97997</td>
<td>0.12285</td>
<td>0.95087</td>
<td></td>
</tr>
<tr>
<td>sd[1]</td>
<td>0.4155</td>
<td>0.47913</td>
<td>0.5508</td>
<td>0.48124</td>
<td>0.034862</td>
<td>0.47855</td>
</tr>
<tr>
<td>sd[2]</td>
<td>0.94344</td>
<td>1.09</td>
<td>1.2505</td>
<td>1.095</td>
<td>0.079045</td>
<td>1.0831</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>MCerr</th>
<th>MC%ofSD</th>
<th>SSeff</th>
<th>AC.10</th>
<th>psrf</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta[1]</td>
<td>0.0020677</td>
<td>1.7</td>
<td>3427</td>
<td>0.17659</td>
<td>1.0009</td>
</tr>
<tr>
<td>beta[2]</td>
<td>0.0021715</td>
<td>1.8</td>
<td>3201</td>
<td>0.20065</td>
<td>1.0011</td>
</tr>
<tr>
<td>sd[1]</td>
<td>0.00017456</td>
<td>0.5</td>
<td>39884</td>
<td>-0.0005609</td>
<td>1</td>
</tr>
<tr>
<td>sd[2]</td>
<td>0.00041293</td>
<td>0.5</td>
<td>36644</td>
<td>0.0092112</td>
<td>0.99999</td>
</tr>
</tbody>
</table>

Total time taken: 2.2 seconds

est1 <- lm(X ~ Z)
est2 <- lm(Y ~ predict(est1))
summary(est2)
Call:
\texttt{lm(formula = Y ~ predict(est1))}

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>-0.9328</td>
<td>-0.3228</td>
<td>-0.0424</td>
<td>0.3128</td>
<td>1.0910</td>
</tr>
</tbody>
</table>

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:

\texttt{library(AER)}

\texttt{est.2sls <- ivreg(Y ~ X | Z)}

\texttt{summary(est.2sls)}

Call:
\texttt{ivreg(formula = Y ~ X | Z)}

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>-2.26775</td>
<td>-0.57731</td>
<td>-0.05521</td>
<td>0.53982</td>
<td>2.07423</td>
</tr>
</tbody>
</table>

Coefficients:

|                      | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------------|----------|------------|---------|---------|
| (Intercept)          |         |            |         |         |
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.

```r
data("CigarettesSW", package = "AER")
head(CigarettesSW, n=4)

# state year cpi population packs income tax price
tables(CigarettesSW, year)

1 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
tax 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:

```r
Cig <- within(subset(CigarettesSW, year == 1995), {
  rprice <- price/cpi
  rincome <- income/population/cpi
  tdiff <- (taxs - tax)/cpi
  rtax <- tax/cpi
})
```
We are interested in

\[
\log(\text{packs}) = \beta_0 + \beta_1 \log(\text{rprice}) + u
\]

However, \text{rprice} is endogenous, correlated with \(u\).
We can use \text{tdiff}, the sales tax on cigarettes, as an instrument for \(\log(\text{rprice})\).

1st stage: \[ \log(\text{rprice}) = \gamma_0 + \gamma_1 \text{tdiff} + \nu \]

2nd stage: \[ \log(\text{packs}) = \beta_0 + \beta_1 \log(\text{rprice}) + u \]

\begin{align*}
\text{tdiff} &= Z \\
\log(\text{rprice}) &= X \\
\log(\text{packs}) &= Y \\
\end{align*}

\(u\) → \text{tdiff} = Z

\[ \downarrow \]

\[ \log(\text{rprice}) = X \]

\(\log(\text{packs}) = Y\)

\textbf{with}(\text{Cig,} \{\text{plot(packs} - \text{rprice}); \text{plot(tdiff, rprice)})\}))
The naïve approach

```r
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:

```r
cor(residuals(est0.lm), log(Cig$rprice))
```

```
[1] -1.957944e-16
```

2 Stage Least Squares (2SLS)

The 1st stage:

```r
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
-0.64676 -0.09030  0.01787  0.11245  0.40779
```
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)  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
Note that the standard errors with this (manual) two stage procedure are not correct. The regression in the second stage cannot take into account the standard errors of the first stage.

The **ivreg** command:

```r
est1.iv <- ivreg(log(packs) ~ log(rprice) | tdiff, data = Cig)
summary(est1.iv)
```

Comparison of estimated price elasticities:
We use the same model as before:

```r
inst.model <- 'model {
for(i in 1:length(y)) {
  x[i]~dnorm(xHat[i], tau[2]) # 1st stage
  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):

<table>
<thead>
<tr>
<th></th>
<th>Lower 95</th>
<th>Median</th>
<th>Upper 95</th>
<th>Mean</th>
<th>SD</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta[1]</td>
<td>-0.074083</td>
<td>0.000479</td>
<td>0.075199</td>
<td>0.000333</td>
<td>0.03787</td>
<td>-0.001335</td>
</tr>
<tr>
<td>beta[2]</td>
<td>-2.1209</td>
<td>-1.1158</td>
<td>-0.29165</td>
<td>-1.1538</td>
<td>0.47205</td>
<td>-1.0608</td>
</tr>
<tr>
<td>sd[1]</td>
<td>0.18351</td>
<td>0.22911</td>
<td>0.27851</td>
<td>0.23112</td>
<td>0.02458</td>
<td>0.22597</td>
</tr>
<tr>
<td>sd[2]</td>
<td>0.078461</td>
<td>0.096956</td>
<td>0.11886</td>
<td>0.097944</td>
<td>0.010522</td>
<td>0.095617</td>
</tr>
</tbody>
</table>

MCerr | MC%ofSD| SSeff | AC.10 | psrf
beta[1]| 0.00023624 | 0.6 | 25700 | 0.00035278 | 1.0003 |
Could there be an omitted variable bias in our second stage equation. Perhaps demand is not only affected by price, but also by income?

→ include \( \log(\text{rincome}) \) in the second stage:

\[
\begin{align*}
\text{log}(\text{rincome}) &= W \\
\text{log}(\text{rprice}) &= X \\
\text{log}(\text{packs}) &= Y \\
\end{align*}
\]

\( u \quad \text{tdiff} = Z \)

\[
est2.iv <- \text{ivreg}(\text{log(packs)} ~ \text{log(rprice)} + \text{log(rincome)} \mid \text{log(rincome)} + \text{tdiff}, \text{data} = \text{Cig})
\]

\text{summary(est2.iv)}

Call:
\text{ivreg(formula = log(packs) ~ log(rprice) + log(rincome) \mid log(rincome) + tdiff, data = Cig)}

Residuals:
\[
\begin{array}{rrrrr}
\text{Min} & \text{1Q} & \text{Median} & \text{3Q} & \text{Max} \\
-0.611000 & -0.086072 & 0.009423 & 0.106912 & 0.393159
\end{array}
\]

Coefficients:
\[
\begin{array}{llllll}
\text{Estimate} & \text{Std. Error} & \text{t value} & \text{Pr(>|t|)} \\
\hline
(\text{Intercept}) & 9.4307 & 1.3584 & 6.943 & 0.00000000124 & *** \\
\text{log(rprice)} & -1.1434 & 0.3595 & -3.181 & 0.00266 & ** \\
\text{log(rincome)} & 0.2145 & 0.2686 & 0.799 & 0.42867 & \\
\end{array}
\]
Note that we include here $\log(\text{rincome})$ as an instrument for itself. Technically this means, that $\log(\text{rincome})$ will be perfectly predicted, i.e. not instrumented.

```r
lm(\log(\text{rincome}) \sim I(\log(\text{rincome})) + \text{tdiff}, \text{data} = \text{Cig})
```

Call:
```
lm(formula = log(rincome) ~ I(log(rincome)) + tdiff, data = Cig)
```

Coefficients:
```
(Intercept) I(log(rincome)) tdiff
0.000e+00 1.000e+00 3.043e-19
```

(we have to use $I(\ldots)$ if we insist on including the response on the right hand side)

**More instruments**

Above we used only $\text{tdiff}$ as an instrument. Perhaps we can do better.

```r
est.ts1b <- lm(log(rprice)~\text{tdiff} + \text{rtax}, \text{data}=\text{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 ***
```

Indeed, compared to the equation where we used only $\text{tdiff}$, 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) 9.89505  1.05861   9.348  4.12e-12 ***
log(rprice) -1.27742  0.26323  -4.853  1.50e-05 ***
log(rincome)  0.28043  0.23862   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

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \\
\quad + \beta_{k+1} W_1 + \ldots + \beta_{k+r} W_r + u \\
\quad + \gamma_{1i} Z_1 + \gamma_{2i} Z_2 + \ldots + \gamma_{mi} Z_m + \nu_i
\]

\( \forall i \in 1, \ldots k: \ X_i = \gamma_{0i} + \gamma_{1i} Z_1 + \gamma_{2i} Z_2 + \ldots + \gamma_{mi} Z_m + \nu_i \)
11 INSTRUMENTAL VARIABLES

- \( X_1, \ldots, X_k \): \( k \) 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:

```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:

```r
summary(lm( Y ~ X + I(X^2) ))
```

Call:
`lm(formula = Y ~ X + I(X^2))`

Residuals:
```
            Min         1Q    Median         3Q        Max
-1.012080  -0.220840  -0.002212  0.235993  0.899389
```
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:

```
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:

<table>
<thead>
<tr>
<th>(Intercept)</th>
<th>X</th>
<th>I(X^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.664</td>
<td>1.321</td>
<td>NA</td>
</tr>
</tbody>
</table>

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):

\[
\begin{align*}
x^2 & \quad u & \quad Z_1 \\
\downarrow & \quad \downarrow & \quad \downarrow \\
y & \leftarrow & x
\end{align*}
\]

summary(ivreg(Y ~ X + I(X^2) | I(X^2) + Z1))

Call:
ivreg(formula = Y ~ X + I(X^2) | I(X^2) + Z1)

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-5.1118</td>
<td>-1.0691</td>
<td>0.1049</td>
<td>1.2330</td>
<td>4.7750</td>
</tr>
</tbody>
</table>

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 | 0.001866   | -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
Solution 2: Linear instruments

\[
\begin{array}{c}
\text{ivreg}(Y \sim X + I(X^2) \mid Z1 + Z2)
\end{array}
\]

Call:
ivreg(formula = Y ~ X + I(X^2) | Z1 + Z2)

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>-2.9241</td>
<td>-1.9393</td>
<td>-1.1785</td>
<td>0.7575</td>
<td>25.1905</td>
</tr>
</tbody>
</table>

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

xHat <- predict(lm(X ~ Z1+Z2))
x2Hat <- predict(lm(X^2 ~ Z1+Z2))
plot(xHat,x2Hat)
### Solution 3: Quadratic instruments

\[
\begin{align*}
  & u \\
  & Z_1, Z_1^2, Z_2, Z_2^2 \\
  \downarrow & \quad \downarrow \\
  Y & \leftarrow X, X^2
\end{align*}
\]

```r
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

```r
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 $Z$

\[ \begin{align*}
\downarrow & \quad Z_1, Z_1^2 \\
\downarrow & \quad Y \leftarrow X, X^2
\end{align*} \]

```
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 0.9233
X  1.15788  0.16785   6.898 9.34e-12 ***
I(X^2) -0.09161  0.07720  -1.187 0.2363
```

---
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
11.5 Weak instruments

1st stage: $X = Z \gamma + \nu$

$\hat{X} = Z \gamma$

2nd stage: $Y = \hat{X} \beta + u$

- 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 $< 10$)

  $\rightarrow$ 2SLS estimator is no longer reliable.

  $\rightarrow$ find better instruments, or use Bayesian inference.

Are the instruments in our example weak?
summary(lm(X ~ Z1+I(Z1^2)))$fstatistic

<table>
<thead>
<tr>
<th>value</th>
<th>numdf</th>
<th>dendf</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.725538</td>
<td>2.000000</td>
<td>997.000000</td>
</tr>
</tbody>
</table>

summary(lm(X^2 ~ Z1+I(Z1^2)))$fstatistic

<table>
<thead>
<tr>
<th>value</th>
<th>numdf</th>
<th>dendf</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6406691</td>
<td>2.0000000</td>
<td>997.0000000</td>
</tr>
</tbody>
</table>

summary(lm(X ~ Z1+I(Z1^2)+Z2+I(Z2^2)))$fstatistic

<table>
<thead>
<tr>
<th>value</th>
<th>numdf</th>
<th>dendf</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.318</td>
<td>4.000000</td>
<td>995.000000</td>
</tr>
</tbody>
</table>

summary(lm(X^2 ~ Z1+I(Z1^2)+Z2+I(Z2^2)))$fstatistic

<table>
<thead>
<tr>
<th>value</th>
<th>numdf</th>
<th>dendf</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.395981</td>
<td>4.0000000</td>
<td>995.0000000</td>
</tr>
</tbody>
</table>

Apparently we do not have a good instrument for $X^2$.

### 11.6 Discrete endogeneous variables

Remember:

1st stage: $X = Z \gamma + \nu$ (works, since $Z \not\perp X$)

$\hat{X} = Z \gamma$

2nd stage: $Y = \hat{X} \beta + u$ (works, since $Z \vdash u$, and, hence $\hat{X} \vdash u$)

Now consider the case where $X$ is not linear in $Z$:

```r
set.seed(123)
N <- 1000
u <- rnorm(N)
nu <- rnorm(N)
Z <- rnorm(N)
X <- as.numeric((Z + u + nu) > 0)
Y <- X + u
```

```latex
\begin{align*}
u & \quad Z \\
\text{↓} & \quad \text{↓} \\
Y & \quad \leftarrow X
\end{align*}
```

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)
Without Bayes we could also use the 2SLS: Caution! We estimate a non-linear process with a linear model.

\texttt{summary(ivreg(Y ~ X | Z))}

Call:
\texttt{ivreg(formula = Y ~ X | Z)}

Residuals:
\begin{center}
\begin{tabular}{ccccc}
Min & 1Q & Median & 3Q & Max \\
-2.86937 & -0.66214 & -0.02271 & 0.66499 & 3.26873 \\
\end{tabular}
\end{center}

Coefficients:
\begin{center}
\begin{tabular}{lccccc}
 & 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 *** \\
\end{tabular}
\end{center}

---
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
We could also use two stage non-linear estimation:

(biased standard errors)

```r
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
-3.324000 -0.923100 -0.001900  0.927700  3.791600
```

Coefficients:
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.05804    0.10057   0.577  0.56400
Xhat        0.91584    0.18448   4.964 7e-05 ***
```

---

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: 7e-05
11.7 Literature

Literature


Appendix 11.A Examples for the lecture

Example 1 The variables \(x_1\), \(x_2\), and \(x_3\) are independent and from a continuous distribution. Which of the following instrumental variable models are exactly identified?

- \(\text{ivreg}(y \sim x_1 | x_1)\)
- \(\text{ivreg}(y \sim x_1 | x_2)\)
- \(\text{ivreg}(y \sim x_1 + x_3 | x_2)\)
- \(\text{ivreg}(y \sim x_1 + x_3 | x_2 + I(x_2^2))\)
- \(\text{ivreg}(y \sim x_1 | x_2 + x_3)\)

Example 2 You want to estimate an equivalent to the following instrumental variables model:

\[
\text{ivreg}(y \sim x_1 | x_2 + x_3)
\]

Consider the following model for JAGS (\(\alpha\) and \(\beta\) are a placeholders):
for(i in 1:length(y)) {
  x1[i] ~ dnorm( x1H[i], tau[2])
  x1H[i] <- a
  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]
• 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:

\[
\text{ivreg}(Y \sim X + I(X^2) / I(X^2) + Z1))
\]

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 \((Z_1 \text{ and } X^2)\) are used as instruments for two variables \(X\) and \(X^2\).

If \(Z_1\) 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:

```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*?

- \(\text{ivreg}(y \sim x + I(x^2) \mid I(x^2) + z)\)
- \(\text{ivreg}(y \sim x \mid z)\)
- \(\text{ivreg}(y \sim x + xH \mid z)\)
- \(\text{ivreg}(y \sim x \mid g1 + g2)\)
- 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

```r
attach(Caschool)
coef(lm(testscr ~ str + elpct + expnstu))
```

```
(Intercept) str elpct expnstu
649.577947257 -0.286399240 -0.656022660 0.003867902
```

```r
elratio <- elpct/100
c coef(lm(testscr ~ str + elratio + expnstu))
```

```
(Intercept) str elratio expnstu
649.577947257 -0.2863992 -65.602266008 0.003867902
```

```r
expnstuTSD <- expnstu/1000
c coef(lm(testscr ~ str + elpct + expnstuTSD))
```

```
(Intercept) str elpct expnstuTSD
649.5779473 -0.28639922 -0.6560227 3.8679018
```
12.2 Which coefficients

- \texttt{testscr} \sim \texttt{distcod} + \texttt{county} + \texttt{district} + \texttt{grspan} + \texttt{enrltot} + \texttt{teachers} + \texttt{calwpct} + \texttt{mealpct} + \texttt{computer} + \texttt{compstu} + \texttt{expnstu} + \texttt{str} + \texttt{avginc} + \texttt{elpct} + \texttt{readscr} + \texttt{mathscr}  
  This model contains perhaps too many coefficients → multicollinearity)

- \texttt{testscr} \sim \texttt{str}  
  This model contains perhaps too few coefficients → omitted variable bias)

- Omitted variable bias

\[ E(\hat{b}_1) = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 \]

Only when \( X_1 \) is orthogonal to \( X_2 \) or \( \beta_2 \) is zero we have no bias

- Overfitting (multicollinearity)

\[ \Sigma_{\hat{\beta}\hat{\beta}} = (X'X)^{-1}X'\text{Iu}^2X(X'X)^{-1} \]

When \( X \) is (almost) collinear, then \( (X'X)^{-1} \) is large, and then \( \Sigma \) is large, hence our estimates are not precise.

Model specification:

\rightarrow start with a base specification

\rightarrow 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 \texttt{mtable} from the \texttt{memisc} package presents a table with an overview of different models. To adjust this table to our needs, we can adjust \texttt{getSummary.lm}. Mainly we use the version of \texttt{getSummary.lm} from the \texttt{memisc} package, but here we replace the (homoscedastic) standard error with the heteroscedasticity-consistent standard error:

```r
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:
\[ \text{est1} \leftarrow \text{lm} (\text{testscr} \sim \text{str}) \]
\[ \text{est2} \leftarrow \text{lm} (\text{testscr} \sim \text{str} + \text{elpct}) \]
\[ \text{est3} \leftarrow \text{lm} (\text{testscr} \sim \text{str} + \text{elpct} + \text{mealpct}) \]
\[ \text{est4} \leftarrow \text{lm} (\text{testscr} \sim \text{str} + \text{elpct} + \text{calwpct}) \]
\[ \text{est5} \leftarrow \text{lm} (\text{testscr} \sim \text{str} + \text{elpct} + \text{mealpct} + \text{calwpct}) \]

\[
\begin{array}{c|c|c|c|c|c}
(1) & (2) & (3) & (4) & (5) \\
\hline
(\text{Intercept}) & 698.933^{***} & 686.032^{***} & 700.150^{***} & 697.999^{***} & 700.392^{***} \\
 & (9.467) & (7.411) & (4.686) & (6.024) & (4.698) \\
\text{str} & -2.280^{***} & -1.101^{**} & -0.998^{**} & -1.308^{***} & -1.014^{***} \\
 & (0.480) & (0.380) & (0.239) & (0.307) & (0.240) \\
\text{elpct} & -0.650^{**} & -0.122^{***} & -0.488^{***} & -0.130^{***} & \\
 & (0.039) & (0.032) & (0.033) & & \\
\text{mealpct} & -0.547^{***} & & & & \\
 & (0.022) & & & & \\
\text{calwpct} & -0.790^{***} & -0.048 & & & \\
 & (0.053) & (0.061) & & & \\
\hline
\text{R-squared} & 0.051 & 0.426 & 0.775 & 0.629 & 0.775 \\
\text{AIC} & 3650.499 & 3441.123 & 3050.999 & 3260.656 & 3052.376 \\
\text{N} & 420 & 420 & 420 & 420 & 420 \\
\end{array}
\]

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 $\rightarrow$ 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:

<table>
<thead>
<tr>
<th></th>
<th>lmg</th>
<th>pmvd</th>
<th>last</th>
<th>first</th>
</tr>
</thead>
<tbody>
<tr>
<td>str</td>
<td>0.03119231</td>
<td>0.0148176134</td>
<td>0.059126952</td>
<td>0.03175031</td>
</tr>
<tr>
<td>elpct</td>
<td>0.22371548</td>
<td>0.0242703918</td>
<td>0.048159854</td>
<td>0.25708495</td>
</tr>
<tr>
<td>mealpct</td>
<td>0.53343971</td>
<td>0.9600101671</td>
<td>0.890678586</td>
<td>0.46768098</td>
</tr>
<tr>
<td>calwpct</td>
<td>0.21165250</td>
<td>0.0009018276</td>
<td>0.002034608</td>
<td>0.24348376</td>
</tr>
</tbody>
</table>

Average coefficients for different model sizes:

<table>
<thead>
<tr>
<th></th>
<th>1X</th>
<th>2Xs</th>
<th>3Xs</th>
<th>4Xs</th>
</tr>
</thead>
<tbody>
<tr>
<td>str</td>
<td>-2.2798083</td>
<td>-1.4612232</td>
<td>-1.1371224</td>
<td>-1.01435328</td>
</tr>
<tr>
<td>elpct</td>
<td>-0.6711562</td>
<td>-0.4347537</td>
<td>-0.2510901</td>
<td>-0.12982189</td>
</tr>
<tr>
<td>mealpct</td>
<td>-0.6102858</td>
<td>-0.5922408</td>
<td>-0.5645062</td>
<td>-0.52861908</td>
</tr>
<tr>
<td>calwpct</td>
<td>-1.0426750</td>
<td>-0.5863541</td>
<td>-0.2639020</td>
<td>-0.04785371</td>
</tr>
</tbody>
</table>
Relative importances for testscr

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.
Now consider a situation where $x_1 \rightarrow x_2 \rightarrow y$:

```r
set.seed(123)
x1 <- rnorm(100)
x2 <- rnorm(100) + 2 * x1
u <- .01*rnorm(100)
y <- x2 + u
est <- lm(y ~ x1 + x2)
```

```r
plot(calc.relimp(est, type=c("lmg","pmvd")))
```

### 95% CI

```r
plot(calc.relimp(est, type=c("lmg","pmvd")))
```
If $x_1$ and $x_2$ are independent, both \texttt{lmg} and \texttt{pmvd} allocate response variance in the same way.

However, if $x_1$ and $x_2$ are dependent it is possible that the coefficient of $x_1$ is close to zero but still correlated with $y$. In such a case \texttt{lmg} still allocates response variance to $x_1$ while \texttt{pmvd} does not.

12.5 Analysis of Variance

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).

$$\frac{\text{RSS}_1 - \text{RSS}_2}{\text{RSS}_2} \frac{n - k_2}{k_2 - k_1} \sim F_{(k_2 - k_1, n - k_2)}$$

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

$$-2 \cdot L = n \cdot \log \frac{\text{RSS}}{n} + C$$

But then

$$2 \cdot (L_2 - L_1) = n \cdot \left( \log \frac{\text{RSS}_1}{n} - \log \frac{\text{RSS}_2}{n} \right) = n \log \frac{\text{RSS}_1}{\text{RSS}_2} \sim \chi^2_{k_2 - k_1}$$
We can, thus, also use the $\chi^2$ statistic to compare two models.

```r
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:

```r
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)

```r
n* log(RSS1 / RSS2) ~ \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

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>416</td>
<td>35451</td>
<td>1</td>
<td>1203.3</td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

F-test for ANOVA

```r
pf((RSS1 - RSS2)/RSS2 * (n-k2)/(k2-k1),k2-k1,n-k2,lower=FALSE)
```

[1] 0.0001547027

```r
anova(est1,est2)
```

Analysis of Variance Table

<table>
<thead>
<tr>
<th>Model 1: testscr ~ str + mealpct + calwpct</th>
<th>Model 2: testscr ~ str + elpct + mealpct + calwpct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Res.Df</td>
<td>RSS</td>
</tr>
<tr>
<td>--------</td>
<td>-----</td>
</tr>
<tr>
<td>1</td>
<td>416</td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(p-Value from $\chi^2$ test was slightly different (Pr > $\chi^2=0.0001342$))

12.5.1 t-test versus F test

The previous anova tests $\beta_{elpct} = 0$. This can also be done with a t-test:

```r
summary(est2)
```

Call:

```r
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 |

---
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.

Hirotugo 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

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>calwpct</td>
<td>70.1</td>
<td>34239</td>
</tr>
<tr>
<td>-</td>
<td>enrltot</td>
<td>78.9</td>
<td>34247</td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td>34169</td>
<td>1859.5</td>
</tr>
<tr>
<td>-</td>
<td>elpct</td>
<td>1262.6</td>
<td>35431</td>
</tr>
<tr>
<td>-</td>
<td>str</td>
<td>1552.8</td>
<td>35721</td>
</tr>
<tr>
<td>-</td>
<td>mealpct</td>
<td>20702.3</td>
<td>54871</td>
</tr>
</tbody>
</table>

Step: AIC=1858.36

testscr ~ str + elpct + mealpct + enrltot

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>enrltot</td>
<td>60</td>
<td>34298</td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td>34239</td>
<td>1858.4</td>
</tr>
<tr>
<td>-</td>
<td>elpct</td>
<td>1208</td>
<td>35446</td>
</tr>
<tr>
<td>-</td>
<td>str</td>
<td>1496</td>
<td>35734</td>
</tr>
<tr>
<td>-</td>
<td>mealpct</td>
<td>51150</td>
<td>85388</td>
</tr>
</tbody>
</table>

Step: AIC=1857.09

testscr ~ str + elpct + mealpct

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td>34298</td>
<td>1857.1</td>
</tr>
<tr>
<td>-</td>
<td>elpct</td>
<td>1167</td>
<td>35465</td>
</tr>
<tr>
<td>-</td>
<td>str</td>
<td>1441</td>
<td>35740</td>
</tr>
<tr>
<td>-</td>
<td>mealpct</td>
<td>52947</td>
<td>87245</td>
</tr>
</tbody>
</table>

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

Coefficients:

<table>
<thead>
<tr>
<th>(Intercept)</th>
<th>str</th>
<th>elpct</th>
<th>mealpct</th>
</tr>
</thead>
<tbody>
<tr>
<td>700.1500</td>
<td>-0.9983</td>
<td>-0.1216</td>
<td>-0.5473</td>
</tr>
</tbody>
</table>

**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.

```r
set.seed(123)
N<-nrow(Caschool)
```
within sample a smaller model must have a larger deviance.

Now do the same out of sample:

\[
\text{sum}((\text{CaOut}\$\text{testscr} - \text{predict}(\text{est}, \text{newdata}\!=\!\text{CaOut}))^2)
\]

[1] 17831.88

\[
\text{sum}((\text{CaOut}\$\text{testscr} - \text{predict}(\text{estSm}, \text{newdata}\!=\!\text{CaOut}))^2)
\]

[1] 17382.6

Out of sample a smaller model can have a smaller deviance.
Of course, the above graph depends on the specific sample of \textit{CaIn} and \textit{CaOut}. We could repeat this exercise for many samples.

\section*{The same idea with polynomial functions}

Within sample deviance:

\begin{verbatim}
plot(sapply(1:15, function(r) deviance(lm(testscr~poly(avginc,r),data=CaIn))),
    xlab="degree of polynomial $r$",ylab="within sample deviance")
\end{verbatim}
Out of sample deviance:

```r
plot(sapply(1:9, function(r) sum((CaOut$testscr - predict(lm(testscr ~ poly(avginc, r), data=CaIn), newdata=CaOut))^2)),
     xlab="degree of polynomial $r$", ylab="out of sample deviance")
```
Of course, the above graph depends on the specific sample of \( \text{CaIn} \) and \( \text{CaOut} \). We could repeat this exercise for many samples.

AIC tries to capture the quality of out of sample prediction:

\[
\begin{align*}
\text{plot}(\text{sapply}(1:15, \text{function}(r) \text{ AIC}(\text{lm}(\text{testscr} \sim \text{poly}(\text{avginc}, r)))), \\
\text{xlab} = "\text{degree of polynomial } r\", \text{ylab} = "\text{AIC}\")
\end{align*}
\]

If we want to model the relation between \( \text{avginc} \) and \( \text{testscr} \), perhaps the best is a polynomial of degree 5:

\[
\text{newdata} <- \text{list}(\text{avginc}=5:55) \\
\text{plot}(\text{testscr} - \text{avginc}) \\
\text{lines(\text{predict(\text{lm}(\text{testscr} \sim \text{poly(\text{avginc}, 5)}),\text{newdata=\text{newdata})-\text{newdata}\$\text{avginc},lty=2,lwd=4))})}
\]
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):

$$LRI = 1 - \frac{\log L}{\log L_0}$$

Pseudo $R^2$ is different from $R^2$: ...
More parameters (K) will give a better fit. We introduce a penalty for “overfitting”.

Akaike “An Information Criterion”: \( \text{AIC} = -2 \log L + 2K \)

Bayesian information criterion (BIC): \( \text{BIC} = -2 \log L + K \log N \)

We will prefer the model with the lower AIC or BIC.

### 12.8.3 An example: Labour market participation

```r
data(Participation)
reg<-glm ( lfp ~ lnnlinc, family=binomial(link="logit"),data=Participation)
```

Call:
```r
glm(formula = lfp ~ lnnlinc, family = binomial(link = "logit"),
    data = Participation)
```

Deviance Residuals:
```
 Min 1Q Median 3Q Max
-1.6823 -1.1188 -0.8713 1.1983 1.7232
```

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
`

```r
reg0<-glm ( lfp ~ 1,family=binomial(link="logit"),data=Participation)
summary(reg0)
```

Call:
```r
glm(formula = lfp ~ 1, family = binomial(link = "logit"), data = Participation)
```

Deviance Residuals:
```
    Min 1Q Median 3Q Max
-1.110 -1.110 -1.110 1.246 1.246
```

Coefficients:
```
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.16090   0.06795  -2.368  0.0179 *
```

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

```
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_1}{L_s} \sim \chi^2_{l-s}
```

where \( L_s \) and \( L_1 \) 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
```r
anova(reg0, reg, test = "LRT")
```

**Analysis of Deviance Table**

<table>
<thead>
<tr>
<th>Resid. Df</th>
<th>Resid. Dev</th>
<th>Df</th>
<th>Deviance</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>871</td>
<td>1203.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>870</td>
<td>1176.0</td>
<td>1</td>
<td>27.251</td>
</tr>
</tbody>
</table>

---

**Signif. codes:** 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

or even with

```r
anova(reg, test = "LRT")
```

**Analysis of Deviance Table**

Model: binomial, link: logit

Response: lfp

Terms added sequentially (first to last)

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td></td>
<td>871</td>
<td>1203.2</td>
<td></td>
</tr>
<tr>
<td>lnnlinc</td>
<td>1</td>
<td>27.251</td>
<td>870</td>
<td>1176.0</td>
</tr>
</tbody>
</table>

---

**Signif. codes:** 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### 12.8.4 z-test versus \( \chi^2 \) test

Note that z-test and \( \chi^2 \) are not equivalent:

- p-value from z-test (from summary table): 0.000000698
- p-value from \( \chi^2 \)-test (from ANOVA): 0.000000179

LR-tests only work asymptotically.

Appendix 12.C presents an extension to this model.

### 12.9 Bayesian Model Comparison

#### 12.9.1 Preliminaries

Consider \( \beta_2 \) in the following relationship:
**12.9.2 Odds in the simple model**

\[ \text{testscr} = \beta_0 + \beta_1 \text{elpct} + \beta_2 \text{str} + u \]

- What kind of probabilistic statements can we make about \( \beta_2 \)?
- We can, e.g., compare \( \Pr(\beta_2 \leq 0) \) with \( \Pr(\beta_2 > 0) \).

\[
\text{mean}(\text{pre.beta3} > 0) = 0.002025
\]
\[
\text{mean}(\text{pre.beta3} \leq 0) = 0.997975
\]

- \( \Pr(\beta_2 > 0) = 0.002025 \)
- \( \Pr(\beta_2 \leq 0) = 0.997975 \)

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:

```r
set.seed(123)
x <- rnorm(5,5,1)
```

- \( x \sim \text{N}(\mu, 1) \) where \( \text{N} \) is the normal distribution.

We compare three models:
\[
\mu_1 = 3.8, \mu_2 = 4, \mu_3 = 6.
\]

```r
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
1 641
2 2484
3 16875

with(data.frame(as.mcmc(c.jags)),
     table(m)/length(m))

   m
1 0.03205
2 0.12420
3 0.84375

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 \( \text{dcat} \)?

```r
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))
```
(ddirch denotes the Dirichlet distribution, a generalisation of the Beta distribution.)

```r
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")]

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SSeff</th>
<th>psrf</th>
</tr>
</thead>
<tbody>
<tr>
<td>mSel[1]</td>
<td>0.03245</td>
<td>19676</td>
<td>1.002214053</td>
</tr>
<tr>
<td>mSel[2]</td>
<td>0.11850</td>
<td>20066</td>
<td>1.000030424</td>
</tr>
<tr>
<td>mSel[3]</td>
<td>0.84905</td>
<td>20289</td>
<td>0.999964926</td>
</tr>
</tbody>
</table>
```

**Example 2**

- A binary process selects (randomly) among the two models we want to compare.
  1. `testscr = \beta_0 + \beta_1elpct + \beta_2str + u`
  2. `testscr = \beta_0 + \beta_1elpct + 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)

```r
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) {
  }
}
```
12.9.4 Result of the Bayesian Model Comparison

Interesting is the result for $m_I$. This variable has the value 0 or 1 for model 0 or model 1 respectively.

$\text{mean}(m_I) = 0.02205$ 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 $\frac{\text{Pr}(\text{model 0})}{\text{Pr}(\text{model 1})} = 44.4$.

Compare with p-values from the frequentist analysis:

```r
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 92.565555 < 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=$\text{Pr}(\text{data}|H_0)$

Here $H_0 =$ “model 1 is true”
\[ p\text{-value}=Pr(\text{data|model 1}) = 0.003978056 \]

- In our Bayesian model comparison we have calculated
  - \( Pr(\text{model 0|data}) = 0.97795 \)
  - \( Pr(\text{model 1|data}) = 0.02205. \)

**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 Fairy. 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

\[
K = \frac{\Pr(X|H_1)}{\Pr(X|H_2)} = \left( \int \Pr(\theta_1|H_1) \Pr(X|\theta_1, H_1) \, d\theta_1 \right) \frac{\Pr(X|\theta_1^*, H_1)}{\Pr(X|\theta_2^*, H_2)}
\]

Bayes factor LR-test

Interpreting \(K\): Harold Jeffreys (1961):

<table>
<thead>
<tr>
<th>(10^0)</th>
<th>(10^{0.5})</th>
<th>(10^1)</th>
<th>(10^{1.5})</th>
<th>(10^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>barely worth mentioning</td>
<td>substantial</td>
<td>strong</td>
<td>very strong</td>
<td>decisive</td>
</tr>
</tbody>
</table>


\[
\begin{align*}
& e^0 \approx 2.7 & e^3 \approx 20 & e^5 \approx 150 \\
& barely worth mentioning & positive & strong & very strong \\
\end{align*}
\]

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

- Hoff, A First Course in Bayesian Statistical Methods.

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:

\[
\text{for } (i \text{ in } 1: \text{length}(y)) \{
\text{y[i] ~ } \alpha
\}
\]
\[
\text{j ~ dbern(1/3)}
\]

What is the value of \( \alpha \)?

- \( \text{dnorm(ifeelse(equals(j,1),0,3),1)} \)
- \( \text{ifeelse(dnorm(equals(j,1),0,3),1)} \)
- \( \text{dnorm(equals(ifeelse(j,1),0,3),1)} \)
- \( \text{equals(dnorm(ifeelse(j,1),0,3),1)} \)
- 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 \( k \):

\[
\begin{array}{cccccccc}
  k & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
  \text{AIC} & 192.78 & 165.21 & 147.83 & 90.07 & 78.05 & 5.93 & 7.76 \\
\end{array}
\]

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 (\( \alpha \) is a placeholder):

```r
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 ~ dunif(0,1) 
```

What should you fill in for \( \alpha \)?

- `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

```r
```
Example 5  In your output from JAGS you obtain a 95%-credible interval for $mI$ of $CI_{95\%} = [0, 1]$ 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. \( \text{price} \sim \text{speed} + \text{hd} + \text{ram} + \text{screen} + \text{cd} \) and \( \text{price} \sim \text{speed} + \text{hd} + \text{ram} + \text{screen} + \text{multi} \).

2. Can you use, e.g., a t-, F-, or a likelihood ratio test to discriminate between the two specifications?

3. Can you use R\(^2\) or information criteria to discriminate between the two specifications?

4. 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

```r
reg4<-glm(lfp ~ lnnlinc + age + educ + nyc + noc ,
          family=binomial(link="logit"),data=Participation)
summary(reg4)
```

Call:
\[
glm(formula = \text{lfp} \sim \text{lnnlinc} + \text{age} + \text{educ} + \text{nyc} + \text{noc}, \text{family} = \text{binomial(link = "logit")}, \text{data} = \text{Participation})
\]

Deviance Residuals:
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:

```r
reg3 <- update ( reg4 , ~ . - educ)
```

Call:
```
glm(formula = lfp ~ lnnlinc + age + nyc + noc, family = binomial(link = "logit"),
     data = Participation)
```

Deviance Residuals:
```
     Min      1Q  Median      3Q     Max
-1.794798 -1.060691 -0.618491  1.110100  2.436777
```

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

```r
anova(reg3, reg4, test = "LRT")
```

Analysis of Deviance Table

<table>
<thead>
<tr>
<th>Resid. Df</th>
<th>Resid. Dev</th>
<th>Df</th>
<th>Deviance</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>867</td>
<td></td>
<td>1102.044</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>866</td>
<td>1</td>
<td>1098.914</td>
<td>1.329689</td>
</tr>
</tbody>
</table>

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Here z-test and $\chi^2$ are more similar (still different):

- p-value from z-test (from summary table):
  
  0.0775

- p-value from $\chi^2$-test (from ANOVA):
  
  0.0769

AIC with the larger model

```r
step(reg4)
```

Start: AIC=1110.91

lfp ~ lnnlinc + age + educ + nyc + noc

```
Df Deviance  AIC
- noc  1 1098.965 1108.965
<none>  1 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>  1 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:

<table>
<thead>
<tr>
<th></th>
<th>lnnlinc</th>
<th>age</th>
<th>educ</th>
<th>nyc</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>12.4641761</td>
<td>-0.9018712</td>
<td>-0.5576145</td>
<td>-0.0452272</td>
</tr>
</tbody>
</table>

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. \( \text{testscr} = \beta_0 + \beta_1 \text{elpct} + \beta_2 \text{str} + u \)
  2. \( \text{testscr} = \beta_0 + \beta_1 \text{elpct} + u \)

- 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:
- \( \beta_i \sim \text{dnorm}(0, 0.0001) \)
- \( \tau \sim \text{dgamma}(0.01, 0.01) \)

Below we want to give some (informed) help to the Bayesian model:

- Pseudoprior for \( \beta \) is trivial.
  - Mean of \( \beta \) can be taken from the estimate of \( \beta \).
  - Precision \( \tau = 1/\sigma^2 \), and we can find \( \sigma \) from the estimate for \( \beta \).

- To obtain the prior for \( \tau \) we use properties of the \( \Gamma \)-distribution:
  If \( \tau \sim \Gamma(\alpha, \beta) \) then \( E(\tau) = \alpha/\beta \) and \( \text{var}(\tau) = \alpha/\beta^2 \)

  We estimate parameters of the \( \Gamma \)-distribution as follows:
  \( \hat{\alpha} = \bar{\tau}^2/\text{var}(\tau), \hat{\beta} = \bar{\tau}/\text{var}(\tau) \).
Bayesian Model Comparison: 1st model

```r
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 \(\beta\) and \(\tau\):

```r
(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
```

```r
(priTau<-matrix(c(1,10),nrow=2,ncol=2,byrow=TRUE))

[,1] [,2]
[1,]  1  10
[2,]  1  10
```

Estimate the model:

```r
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):

<table>
<thead>
<tr>
<th></th>
<th>Lower95</th>
<th>Median</th>
<th>Upper95</th>
<th>Mean</th>
<th>SD</th>
<th>Mode</th>
</tr>
</thead>
</table>
Apparentley 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 com-
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):

<table>
<thead>
<tr>
<th></th>
<th>Lower95</th>
<th>Median</th>
<th>Upper95</th>
<th>Mean</th>
<th>SD</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta[1]</td>
<td>-1.4186</td>
<td>0.003578</td>
<td>1.3566</td>
<td>0.00075669</td>
<td>0.70749</td>
<td>--</td>
</tr>
<tr>
<td>beta[2]</td>
<td>-0.72803</td>
<td>-0.65037</td>
<td>-0.57322</td>
<td>-0.65036</td>
<td>0.039496</td>
<td>--</td>
</tr>
<tr>
<td>beta[3]</td>
<td>-1.8534</td>
<td>-1.1017</td>
<td>-0.36435</td>
<td>-1.1012</td>
<td>0.37932</td>
<td>--</td>
</tr>
<tr>
<td>beta[4]</td>
<td>-194.63</td>
<td>-1.046</td>
<td>196.91</td>
<td>-0.30985</td>
<td>99.673</td>
<td>--</td>
</tr>
<tr>
<td>beta[5]</td>
<td>-201.79</td>
<td>0.55934</td>
<td>193.51</td>
<td>0.26679</td>
<td>100.46</td>
<td>--</td>
</tr>
<tr>
<td>tau[1]</td>
<td>0.0041457</td>
<td>0.0047721</td>
<td>0.0054357</td>
<td>0.004779</td>
<td>0.00033008</td>
<td>--</td>
</tr>
<tr>
<td>tau[2]</td>
<td>0 0.0602e-29</td>
<td>0.338</td>
<td>1.0039</td>
<td>9.5414</td>
<td>--</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>MCerr</th>
<th>MC%ofSD</th>
<th>SS eff</th>
<th>AC.10</th>
<th>psrf</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta[1]</td>
<td>0.003504</td>
<td>0.5</td>
<td>0.0095198</td>
<td>1.0001</td>
<td></td>
</tr>
<tr>
<td>beta[2]</td>
<td>0.00020404</td>
<td>0.5</td>
<td>0.0048292</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>beta[3]</td>
<td>0.0019639</td>
<td>0.5</td>
<td>0.0061509</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>beta[4]</td>
<td>0.49767</td>
<td>0.5</td>
<td>0.0051776</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>beta[5]</td>
<td>0.50408</td>
<td>0.5</td>
<td>0.0065203</td>
<td>0.99997</td>
<td></td>
</tr>
<tr>
<td>tau[1]</td>
<td>1.6512e-06</td>
<td>0.5</td>
<td>0.0074455</td>
<td>0.99997</td>
<td></td>
</tr>
<tr>
<td>tau[2]</td>
<td>0.047707</td>
<td>0.5</td>
<td>0.0017028</td>
<td>1.0063</td>
<td></td>
</tr>
</tbody>
</table>

Total time taken: 5.4 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):

<table>
<thead>
<tr>
<th></th>
<th>Lower95</th>
<th>Median</th>
<th>Upper95</th>
<th>Mean</th>
<th>SD</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta[1]</td>
<td>-194.02</td>
<td>0.49833</td>
<td>198.37</td>
<td>0.10098</td>
<td>100.05</td>
<td>--</td>
</tr>
<tr>
<td>beta[2]</td>
<td>-197.28</td>
<td>-1.6256</td>
<td>194.03</td>
<td>-1.4997</td>
<td>100.25</td>
<td>--</td>
</tr>
<tr>
<td>beta[3]</td>
<td>-196.97</td>
<td>-0.051012</td>
<td>193.68</td>
<td>-0.26462</td>
<td>99.619</td>
<td>--</td>
</tr>
<tr>
<td>beta[4]</td>
<td>-1.391</td>
<td>-0.0073152</td>
<td>1.4125</td>
<td>-0.00229</td>
<td>0.71152</td>
<td>--</td>
</tr>
</tbody>
</table>
Bayesian Model Comparison: Pseudopriors

beta0.pri<-select0.jags$summary$statistics[1:3,c("Mean","SD")]
beta1.pri<-select1.jags$summary$statistics[4:5,c("Mean","SD")]

Estimate with Pseudopriors

selectP.data<-list(y=testscr-mean(testscr),
  x1=elpct-mean(elpct),x2=str-mean(str),
  priBeta=priBetaP,priTau=priTauP)

JAGS model summary statistics from 40000 samples (chains = 4; adapt+burnin = 5000):
<table>
<thead>
<tr>
<th></th>
<th>Lower95</th>
<th>Median</th>
<th>Upper95</th>
<th>Mean</th>
<th>SD</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta[1]</td>
<td>-0.97842</td>
<td>0.0025589</td>
<td>0.98714</td>
<td>0.0013213</td>
<td>0.50438</td>
<td>--</td>
</tr>
<tr>
<td>beta[2]</td>
<td>-0.70519</td>
<td>-0.65002</td>
<td>-0.5953</td>
<td>-0.65011</td>
<td>0.028051</td>
<td>--</td>
</tr>
<tr>
<td>beta[3]</td>
<td>-1.6311</td>
<td>-1.0993</td>
<td>-0.56962</td>
<td>-1.0997</td>
<td>0.27063</td>
<td>--</td>
</tr>
<tr>
<td>beta[4]</td>
<td>-1.3888</td>
<td>-0.0085613</td>
<td>1.3797</td>
<td>-0.0075432</td>
<td>0.70524</td>
<td>--</td>
</tr>
<tr>
<td>beta[5]</td>
<td>-0.74814</td>
<td>-0.67116</td>
<td>-0.59525</td>
<td>-0.67126</td>
<td>0.039147</td>
<td>--</td>
</tr>
<tr>
<td>tau[1]</td>
<td>0.0043113</td>
<td>0.0047839</td>
<td>0.0052412</td>
<td>0.0047876</td>
<td>0.0002363</td>
<td>--</td>
</tr>
<tr>
<td>tau[2]</td>
<td>0.0040696</td>
<td>0.0046924</td>
<td>0.005324</td>
<td>0.0046979</td>
<td>0.00032108</td>
<td>--</td>
</tr>
<tr>
<td>mI</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02205</td>
<td>0.14685</td>
<td>0</td>
</tr>
</tbody>
</table>

```
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: 6.4 seconds

```r
acfplot(as.mcmc(selectPseudoprior.jags), aspect="fill")
```