Multivariate statistics

2024-08-08

Table of contents

III EXAMPREPARATION 208

IV ADDITIONAL RESSOURCES 212

Chapter 1

Multivariate Analysis

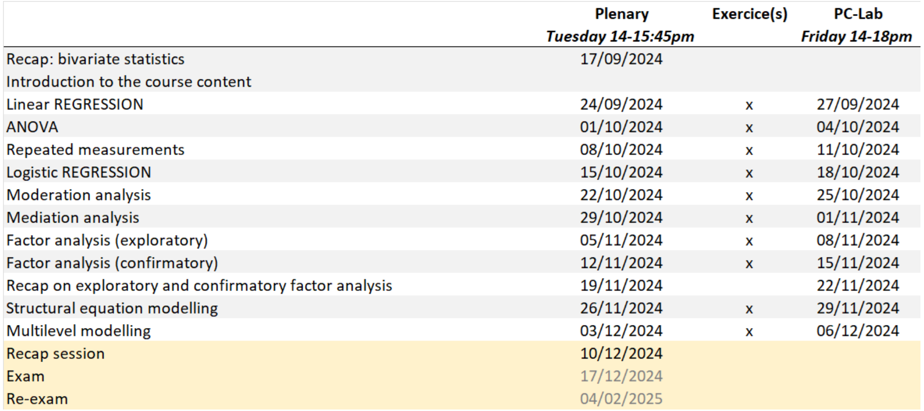
Chapter 2

Course content

The course will cover the following topics and entails PC-Lab sessions with practical examples to familiarize with analysis of variance and regression analysis, which serve as a basis for many relevant statistical methods. The main goals relate to being able to understand the procedures in the foreground of each method (learn-oriented) and to address specific research questions (competence-oriented). More specifically, the course aims to provide you with the following abilities:

- In-depth knowledge of the most important methods of multivariate statistics
- Knowledge of the limits and requirements of the procedures
- Comprehending and calculating tasks/cases with R
- Interpretation of the empirical parameters
- Translation of questions/hypotheses into appropriate analysis methods

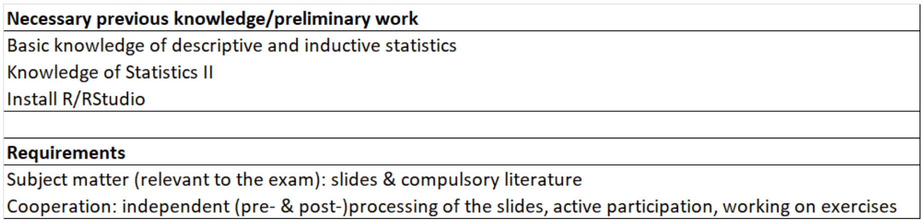
These abilities will enable us to assess the quality of scientific publications that are associated with multivariate methods, but also to attend special lectures and to acquire further multivariate methods autonomously.



	Plenary	Exercice(s)	PC-Lab
	Tuesday 14-15:45pm		*Friday 14-18pm*
Recap: bivariate statistics	17/09/2024		
Introduction to the course content			
Linear REGRESSION	24/09/2024	x	27/09/2024
ANOVA	01/10/2024	x	04/10/2024
Repeated measurements	08/10/2024	x	11/10/2024
Logistic REGRESSION	15/10/2024	x	18/10/2024
Moderation analysis	22/10/2024	x	25/10/2024
Mediation analysis	29/10/2024	x	01/11/2024
Factor analysis (exploratory)	05/11/2024	x	08/11/2024
Factor analysis (confirmatory)	12/11/2024	x	15/11/2024
Recap on exploratory and confirmatory factor analysis	19/11/2024		22/11/2024
Structural equation modelling	26/11/2024	x	29/11/2024
Multilevel modelling	03/12/2024	x	06/12/2024
Recap session	10/12/2024		
Exam	17/12/2024		
Re-exam	04/02/2025		

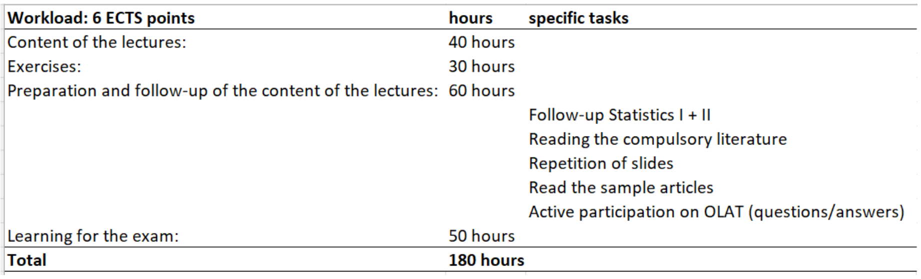
Figure 2.1: Semester schedule

The excel file also gives you an idea of the necessary prior knowledge and of the workload:



Necessary previous knowledge/preliminary work
Basic knowledge of descriptive and inductive statistics
Knowledge of Statistics II
Install R/RStudio
Requirements
Subject matter (relevant to the exam): slides & compulsory literature
Cooperation: independent (pre- & post-)processing of the slides, active participation, working on exercises

Figure 2.2: Prior knowledge



Workload: 6 ECTS points	**hours**	**specific tasks**
Content of the lectures:	40 hours	
Exercises:	30 hours	
Preparation and follow-up of the content of the lectures:	60 hours	
		Follow-up Statistics I + II
		Reading the compulsory literature
		Repetition of slides
		Read the sample articles
		Active participation on OLAT (questions/answers)
Learning for the exam:	50 hours	
Total	**180 hours**	

Figure 2.3: Workload

Chapter 3

R programming language and RStudio

The course is taught using the R programming language. R offers a wide variety of statistics-related libraries and provides a favorable environment for statistical computing and design. It is used by many quantitative analysts as a programming tool since it's useful for data importing and cleaning.

RStudio integrates with R as an IDE (Integrated Development Environment) to provide further functionality. RStudio combines a source code editor, build automation tools and a debugger.

3.1 Install R and RStudio

For this course, you are required to install both R and RStudio on your personal computer.

Install R:

- Download the last version of R for Windows or Mac (https://cran.r-project.org/)
- Save the installation file and execute it.
- Verify that the installation is completed and is working as expected.

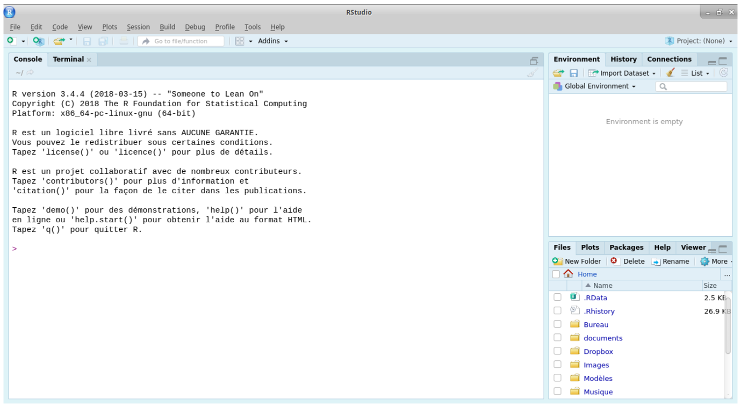
Install RStudio:

- Download RStudio: open source version compatible with Windows or Mac (https://posit.co/download/rstudio-desktop)
- Save the installation file and execute it.
- Verify that the installation is completed and is working as expected.

You should see this logo on your Desktop:

Figure 3.1: RStudio logo

When opening RStudio, the following panes should be accessible to you:



```
R version 3.4.4 (2018-03-15) -- "Someone to Lean On"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R est un logiciel libre livré sans AUCUNE GARANTIE.
Vous pouvez le redistribuer sous certaines conditions.
Tapez 'license()' ou 'licence()' pour plus de détails.

R est un projet collaboratif avec de nombreux contributeurs.
Tapez 'contributors()' pour plus d'information et
'citation()' pour la façon de le citer dans les publications.

Tapez 'demo()' pour des démonstrations, 'help()' pour l'aide
en ligne ou 'help.start()' pour obtenir l'aide au format HTML.
Tapez 'q()' pour quitter R.

>
```

Figure 3.2: RStudio view

You can verify that everything works as expected by typing the following code in the console:

```
print("Hello, welcome!")
## [1] "Hello, welcome!"
```

3.2 Recommendations

In order to easily share scripts with other people using operating systems different from yours (without having problems, for example, with accents and other special characters), it is recommended to ask RStudio to encode the default scripts in UTF-8 (universal encoding):

Tools > Global Options > Code > Saving > Default text encoding and choose UTF-8

This manipulation is not useful for Linux users, who encode in UTF-8 by default.

3.3 R packages

In R, you have to load extensions (packages) that allow you to perform specific statistical operations. Each package is a sort of directory of functions, which you can install (only once is enough) and then call (as many times as you want) when you need it. We will be using different packages that will be most useful to follow the course. First, generic packages dedicated to basic statistical operations, to the manipulation and representation of data, and to the creation of automated reports are presented. Then, a short list of packages per family of methods will be provided.

3.4 Databases and external data sources

Several databases will be used during the exercise sessions (as well as for the demonstrations during the lectures). Most databases are data collected in the framework of (inter)national opinion surveys.

Concerning the Swiss data, you will be able to access the data immediately and free of charge on SWISSUbase. If you are not already registered with SWISSUbase, click here to register.

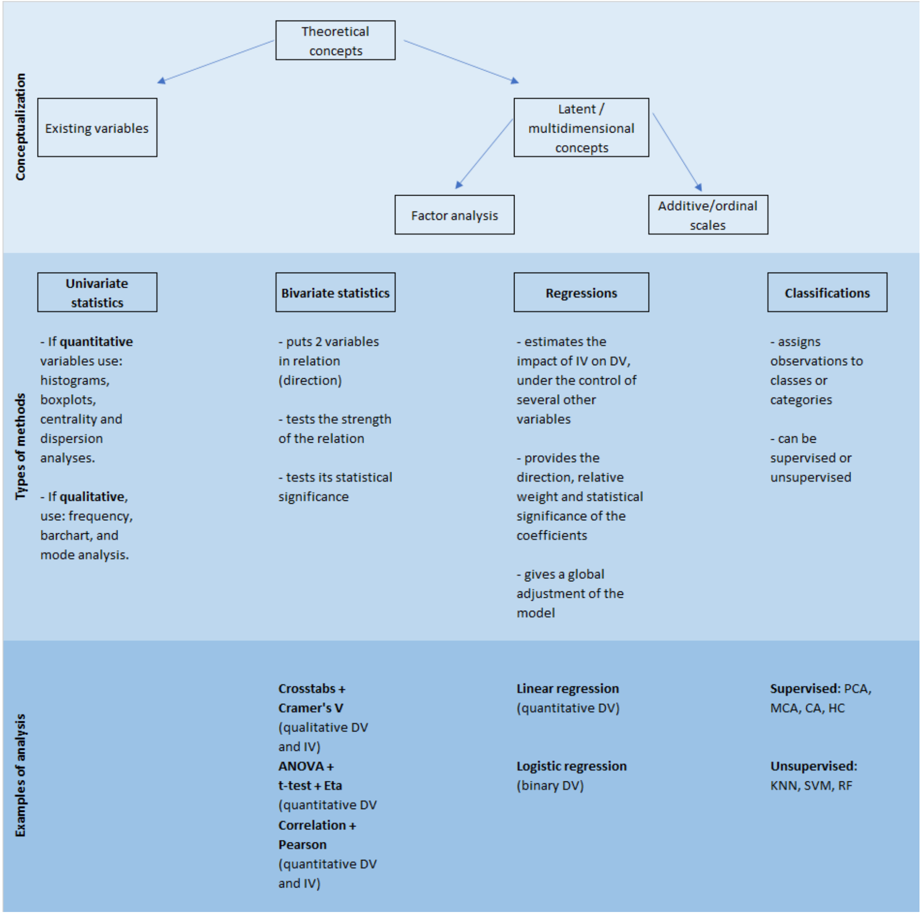
Cross-national opinion surveys are also available under the following websites:

- European Social Survey (https://www.europeansocialsurvey.org)
- World Value Survey (https://www.worldvaluessurvey.org)
- International Social Survey Program (https://issp.org/)
- Eurobarometer (https://europa.eu/eurobarometer/surveys/browse/all)

3.5 What's next? Learning more stats

Univariate analysis consists of the analysis of only one variable. It thus deals with one quantity that changes, but not with causes or relationships. The main purpose is to describe the data and find patterns that exist within it. Furthermore, bivariate analysis involves two different variables. The analysis of this type of data deals with relationships among the two variables.

When the analysis involves three or more variables, it is categorized under multivariate. It is similar to bivariate but contains more than one dependent variable. The ways to perform analysis on this data depends on the goals to be achieved. Some of the techniques are regression analysis, path analysis, factor analysis, multivariate analysis of variance, and more.



	Univariate statistics	**Bivariate statistics**	**Regressions**	**Classifications**
Types of methods	- If **quantitative** variables use: histograms, boxplots, centrality and dispersion analyses. - If **qualitative**, use: frequency, barchart, and mode analysis.	- puts 2 variables in relation (direction) - tests the strength of the relation - tests its statistical significance	- estimates the impact of IV on DV, under the control of several other variables - provides the direction, relative weight and statistical significance of the coefficients - gives a global adjustment of the model	- assigns observations to classes or categories - can be supervised or unsupervised
Examples of analysis		**Crosstabs + Cramer's V** (qualitative DV and IV) **ANOVA + t-test + Eta** (quantitative DV **Correlation + Pearson** (quantitative DV and IV)	**Linear regression** (quantitative DV) **Logistic regression** (binary DV)	**Supervised:** PCA, MCA, CA, HC **Unsupervised:** KNN, SVM, RF

Figure 3.3: Types of methods and analyses

Part I

RECAP

Chapter 4

Quick recap: focus on bivariate statistics

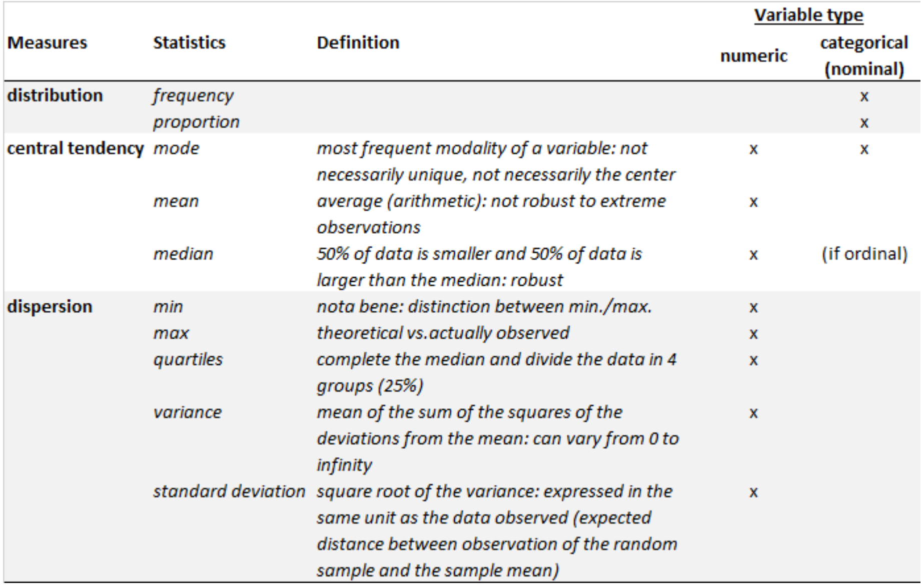
4.1 Recap: bivariate statistic and hypothesis testing

When doing research, we go from the simplest to the most complex types of analysis. Each stage of the analysis brings information for the next step. However, this is an iterative process as sometimes it is necessary go back to simpler steps.

- Descriptive / univariate analysis
- Bivariate analysis
- Multivariate analysis (final model)

4.1.1 Univariate statistics

At the univariate stage, the objective is to become familiar with the variables (terms, missing data, atypical cases, etc.) and to propose a diagnosis of the variables (e.g. any filtering or recoding needed?). At this stage, we describe the distribution of observations over a variable (e.g. frequencies of all modalities of a variable) and we characterize the distribution of a variable (e.g. central tendency and dispersion) using appropriate tools (e.g. summaries, graphics, etc.).



Measures	Statistics	Definition	Variable type numeric	 categorical (nominal)
distribution	*frequency*			x
	proportion			x
central tendency	*mode*	*most frequent modality of a variable: not necessarily unique, not necessarily the center*	x	x
	mean	*average (arithmetic): not robust to extreme observations*	x	
	median	*50% of data is smaller and 50% of data is larger than the median: robust*	x	(if ordinal)
dispersion	*min*	*nota bene: distinction between min./max.*	x	
	max	*theoretical vs.actually observed*	x	
	quartiles	*complete the median and divide the data in 4 groups (25%)*	x	
	variance	*mean of the sum of the squares of the deviations from the mean: can vary from 0 to infinity*	x	
	standard deviation	*square root of the variance: expressed in the same unit as the data observed (expected distance between observation of the random sample and the sample mean)*	x	

Figure 4.1: Univariate statistics

4.1.2 Reporting descriptive statistics

Below are some basic commands to calculate descriptive statistics. R provides packages/functions, which makes calculating and automatically generating a table of summary statistics for categorical and numeric variables. Let's write an example with the "world" dataset from the "poliscidata" package by focusing on the following variables: dem_level4, literacy, pop_age, gender_unequal, religion. We will generate an example of a summary statistics table using the "table1" package. A full description of the dataset can be found here.

```
# install the needed packages
# install.packages("poliscidata")
# install.packages("table1")
# load the data and select the variables
db=poliscidata::world
db=db[,c("country","dem_level4","literacy","pop_age","gender_unequal","religoin")]
# make sure to declare categorical variables as factors
db$country = factor(db$country)
db$dem_level4 = factor(db$dem_level4)
db$religoin = factor(db$religoin)
# get the descriptive table by type of democracy level
table1::table1(~literacy + pop_age + gender_unequal + religoin | dem_level4,
               data = db,
               na.rm = TRUE,
               digits = 1,
               format.number = TRUE)
```

	Full Democ	Part Democ	Hybrid	Authoritarian	Overall
	(N=24)	(N=52)	(N=39)	(N=52)	(N=167)
literacy					
Mean (SD)	100 (3)	90 (10)	70 (20)	70 (20)	80 (20)
Median [Min, Max]	100 [80, 100]	90 [40, 100]	70 [20, 100]	80 [30, 100]	90 [20, 100]
Missing	2 (8.3%)	4 (7.7%)	4 (10.3%)	0 (0%)	10 (6.0%)
pop_age					
Mean (SD)	40 (4)	30 (8)	20 (7)	20 (6)	30 (9)
Median [Min, Max]	40 [30, 40]	30 [20, 40]	20 [20, 40]	20 [20, 40]	30 [20, 40]
Missing	0 (0%)	1 (1.9%)	0 (0%)	0 (0%)	1 (0.6%)
gender_unequal					
Mean (SD)	0.3 (0.09)	0.5 (0.2)	0.7 (0.09)	0.6 (0.1)	0.5 (0.2)
Median [Min, Max]	0.3 [0.2, 0.5]	0.5 [0.2, 0.8]	0.7 [0.5, 0.8]	0.7 [0.4, 0.9]	0.6 [0.2, 0.9]
Missing	0 (0%)	7 (13.5%)	9 (23.1%)	16 (30.8%)	32 (19.2%)
religoin					
Catholic	12 (50.0%)	24 (46.2%)	8 (20.5%)	8 (15.4%)	52 (31.1%)
Orthodox Christian	0 (0%)	9 (17.3%)	4 (10.3%)	3 (5.8%)	16 (9.6%)
Other Christian	7 (29.2%)	8 (15.4%)	7 (17.9%)	5 (9.6%)	27 (16.2%)
Muslim	0 (0%)	4 (7.7%)	14 (35.9%)	28 (53.8%)	46 (27.5%)
Buddhist	1 (4.2%)	2 (3.8%)	4 (10.3%)	4 (7.7%)	11 (6.6%)
Other	3 (12.5%)	5 (9.6%)	2 (5.1%)	3 (5.8%)	13 (7.8%)
Missing	1 (4.2%)	0 (0%)	0 (0%)	1 (1.9%)	2 (1.2%)

4.1.3 Fiability and validity

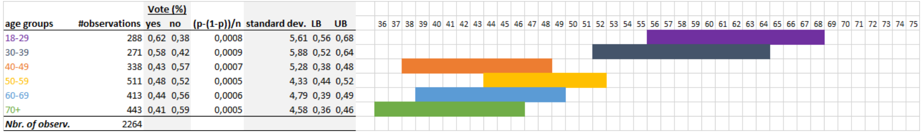
The Total Survey Error (TSE) framework (Biemer, 2010) accounts for the different sources of errors that occur at each stage of an investigation (e.g. coverage errors, selection errors, and measurement errors). Two important concepts are: reliability and validity. Reliability refers to the idea of replicability. It accounts for the degree of consistency of a measurement (by different observers, at different times). Validity addresses the conclusions we can draw from of a measure. For instance, internal validity aims at measuring the fit between the concept and the measure.

Sample statistics are estimators of population parameters. A particular point estimate cannot be expected to be exactly equal to the value of the parameter in the population. Therefore, the estimate always contains a margin of error. For instance, the normal law suggests that the distribution of a variable is around an average value and the other values increase and decrease in a homogeneous and symmetrical way around this average value. The distribution is thus in the form of a bell curve (or Gaussian curve).

The margin of error is also called the confidence interval. It corresponds to the area where we know for a given probability that the average or the percentage of a value will be found.

For a mean: $\overline{x} \pm 1.96(\frac{\sigma(X)}{\sqrt{n}})$

For a proportion: $Z_\alpha \sqrt{\frac{p(1-p)}{n}}$



		Vote (%)					
age groups	#observations	yes	no	(p-(1-p))/n	standard dev.	LB	UB
18-29	288	0,62	0,38	0,0008	5,61	0,56	0,68
30-39	271	0,58	0,42	0,0009	5,88	0,52	0,64
40-49	338	0,43	0,57	0,0007	5,28	0,38	0,48
50-59	511	0,48	0,52	0,0005	4,33	0,44	0,52
60-69	413	0,44	0,56	0,0006	4,79	0,39	0,49
70+	443	0,41	0,59	0,0005	4,58	0,36	0,46
Nbr. of observ.	2264						

Figure 4.2: Confidence intervals for proportions

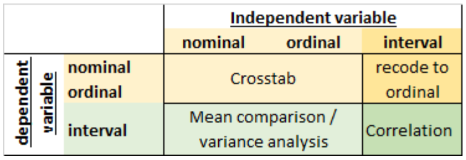
4.1.4 Bivariate analysis

Inferential statistics makes use of probability theory. It indicates to the researcher the probability of being wrong by generalizing the findings of his study to the entire population, assuming a certain risk of error (generally 0.05, or 5%). To do so, two tools are used: estimation and hypothesis testing.

Three steps are generally applied:

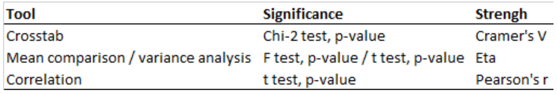
- linking two variables: what values does DV take (e.g. use of social networks) according to an IV (e.g. political experience)?
- determining the statistical significance of this relationship (= hypothesis test)
- quantifying the strength of the relationship (and possibly determine its direction)

Statistical tools and tests in bivariate analysis depend on the level of measurement of the variables:



		Independent variable		
		nominal	ordinal	interval
dependent variable	nominal ordinal	Crosstab		recode to ordinal
	interval	Mean comparison / variance analysis		Correlation

Figure 4.3: Types of bivariate analyses



Tool	Significance	Strengh
Crosstab	Chi-2 test, p-value	Cramer's V
Mean comparison / variance analysis	F test, p-value / t test, p-value	Eta
Correlation	t test, p-value	Pearson's r

Figure 4.4: Types of bivariate tests

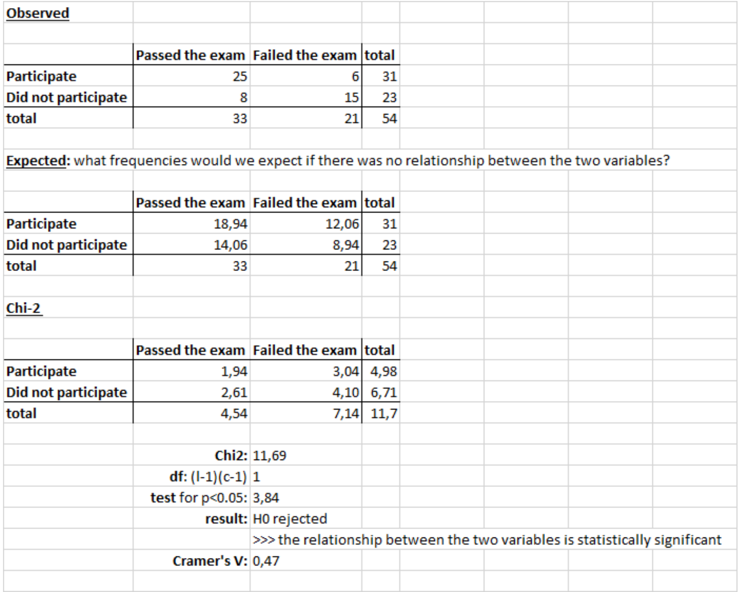
4.1.4.1 Relationship between two categorical variables

We can examine the relationship between two categorical variables based on a Chi-square test. The tested hypothesis reads as: what is the probability of obtaining the frequencies observed in the sample if there was no relationship between the two variables in the population? The null hypothesis states that the two variables are independent. The Chi-2 statistic is calculated as follows:

$$Chi^2 = \sum \frac{(o-e)^2}{e}$$

where o is the observed value and e is the expected value.

This calculated Chi-2 statistic is compared to the critical value (obtained from statistical tables) with df=(r-1)(c-1) degrees of freedom and p = 0.05. If the calculated Chi-2 statistic is greater than the critical value, then the variables are not independent of each other and are, thus, significantly associated.



Observed

	Passed the exam	Failed the exam	total
Participate	25	6	31
Did not participate	8	15	23
total	33	21	54

Expected: what frequencies would we expect if there was no relationship between the two variables?

	Passed the exam	Failed the exam	total
Participate	18,94	12,06	31
Did not participate	14,06	8,94	23
total	33	21	54

Chi-2

	Passed the exam	Failed the exam	total
Participate	1,94	3,04	4,98
Did not participate	2,61	4,10	6,71
total	4,54	7,14	11,7

Chi2:	11,69
df: (l-1)(c-1)	1
test for p<0.05:	3,84
result:	H0 rejected
	>>> the relationship between the two variables is statistically significant
Cramer's V:	0,47

Figure 4.5: Chi-2 test and Cramer's V

We can reproduce this example with R code:

```r
my_data=data.frame(passed=c(25,8),
                   failed=c(6,15))
rownames(my_data)=c("participated","not_participated")
chisq=chisq.test(my_data, correct = F) # No Yates correction (conservative)
round(chisq$expected,2)
```

```
##                  passed failed
## participated      18.94  12.06
## not_participated  14.06   8.94
chisq
##
##  Pearson's Chi-squared test
##
## data:  my_data
## X-squared = 11.686, df = 1, p-value = 0.0006297
```

4.1.4.2 Relationship between one interval variable and one categorical variable

We can compare the distribution of values on a quantitative variable (central tendency, dispersion) for different subpopulations (e.g. different modalities of the categorical independent variable). To see if the two variables are independent or if there is a relationship between them, we compare the variation between groups (mean) and the variation within groups (standard deviation). There is a relationship between two variables when the different groups are distinct and homogeneous.

We need to determine the statistical significance of the relationship to see if it can be generalized to the population. As for the crosstables, it is necessary to carry out a test of hypotheses. The null hypothesis states that the two variables are independent (if p<0.05 for the F-test, reject H0).

As with crosstables, there is also a measure of association: Eta ranging from 0 (perfect independence) to 1 (maximum association). In addition, Eta-2 informs us about the explanatory power of the model: percentage of the variance of the DV explained by the IV.

4.1.4.3 Relationship between two interval variables

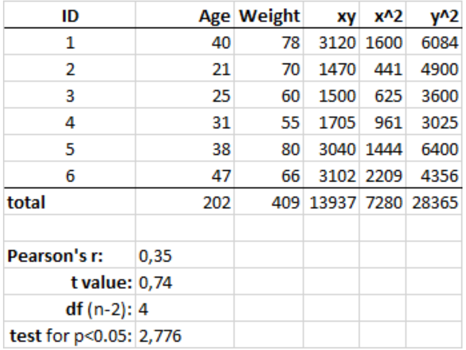
Correlation is a measure of association that provides information on the (linear) relationship between two quantitative variables, namely its direction and strength. The correlation coefficient, Pearson's r, ranges from -1 (negative relationship) to +1 (positive relationship):

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}}$$

The p-value can be determined by using the correlation coefficient table for the degrees of freedom (df = n−2) and calculating the t-value (and determining the corresponding p-value using the t-table):

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

If the p-value is $< 5\%$, then the correlation between x and y is significant.

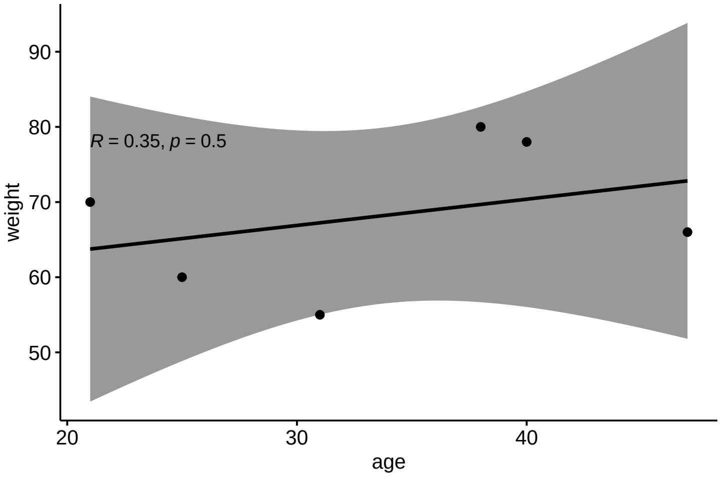


ID	Age	Weight	xy	x^2	y^2
1	40	78	3120	1600	6084
2	21	70	1470	441	4900
3	25	60	1500	625	3600
4	31	55	1705	961	3025
5	38	80	3040	1444	6400
6	47	66	3102	2209	4356
total	202	409	13937	7280	28365
Pearson's r:	0,35				
t value:	0,74				
df (n-2):	4				
test for p<0.05:	2,776				

Figure 4.6: Pearson's r calculation

We can reproduce this example with R code:

```
# create a dataset
my_data <- data.frame(age=c(40,21,25,31,38,47),
                      weight=c(78,70,60,55,80,66))
ggpubr::ggscatter(my_data, x = "age", y = "weight",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "age", ylab = "weight")
```



```
cor.test(my_data$age, my_data$weight, method = "pearson")
##
##  Pearson's product-moment correlation
##
```

```
## data:  my_data$age and my_data$weight
## t = 0.74025, df = 4, p-value = 0.5002
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6465987  0.9040106
## sample estimates:
##       cor
## 0.3471102
```

Nota bene: Do not forget to conduct preliminary test to check the test assumptions. In particular, is the covariation linear? Verify it using a scatter plot. Are the variables following a normal distribution? Use Shapiro-Wilk normality test using the function shapiro.test() and look at the normality plot using the function ggqqplot(). For the Shapiro-Wilk test, the null hypothesis states that the data are normally distributed.

4.2 How it works in R?

See the lecture slides on bivariate statistics:

You can also download the PDF of the slides here:

4.3 Quiz

True

False

Statement

Cross tabulation assesses the relationship between two variables with nominal measurement.

The p-value can be defined as the probability of observing a result when the null hypothesis is false.

The null hypothesis for the difference of means test posits that one group will have a higher average than another on the dependent variable.

It is important to assess the magnitude of the result whenever conducting a statistical test because a higher coefficient always means that it will be statistically significant.

View Results

My results will appear here

4.4 Time to practice on your own

4.4.1 Chi-square test

Try to produce a crosstab in R between the variables standing for democratic decentralization (decent08) and the type of regime (dem_level4). What can you derive from the crosstab and its associated Chi-squared test? You can use the CrossTable() function from the descr package.

```
db=poliscidata::world
db=db[,c("decent08","dem_level4")]
db$country = factor(db$decent08)
db$dem_level4 = factor(db$dem_level4)
descr::CrossTable(db$decent08, db$dem_level4, expected = F, chisq = T, prop.chisq = F)
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
## =========================================================================================
##                                    db$dem_level4
## db$decent08                        Full Democ   Part Democ   Hybrid   Authoritarian   Tot
## -----------------------------------------------------------------------------------------
## No local elections                          0            3        4              16      2
##                                         0.000        0.130    0.174           0.696   0.2
##                                         0.000        0.097    0.160           0.432
##                                         0.000        0.026    0.035           0.140
## -----------------------------------------------------------------------------------------
## Lgsltr is elctd bt exctv is ap              5            7        5              14      3
##                                         0.161        0.226    0.161           0.452   0.2
##                                         0.238        0.226    0.200           0.378
##                                         0.044        0.061    0.044           0.123
## -----------------------------------------------------------------------------------------
## Lgsltr and exctv ar lclly elct             16           21       16               7      6
##                                         0.267        0.350    0.267           0.117   0.5
##                                         0.762        0.677    0.640           0.189
##                                         0.140        0.184    0.140           0.061
## -----------------------------------------------------------------------------------------
## Total                                      21           31       25              37     1
##                                         0.184        0.272    0.219           0.325
## =========================================================================================
##
## Statistics for All Table Factors
```

```
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 = 30.41647      d.f. = 6      p = 0.0000328
```

Solution: interpretation

The Chi-squared test result indicates a statistically significant association between the type of local government elections and the level of democracy ($p < 0.05$). The observed distribution shows that countries without local elections are predominantly authoritarian, while those with both the legislature and executive locally elected are more likely to be full or partial democracies. The test's Chi-squared value of 30.41647 with 6 degrees of freedom suggests that the differences in the distributions are unlikely to be due to random chance, pointing towards a meaningful relationship between local election practices and the overall level of democracy in a country.
Nota bene: The degrees of freedom (df) for a Chi-squared test are calculated based on the number of categories in the variables being analyzed. In this case, the degrees of freedom are 6. The formula for calculating degrees of freedom in a Chi-squared test for a contingency table is: $df = (r - 1) * (c - 1) = (3 - 1) * (4 - 1) = 2 * 3 = 6$. The degrees of freedom determine which Chi-squared distribution to use for calculating the p-value. They reflect the complexity of the data and are used to compare the calculated Chi-squared value against the critical value from the Chi-squared distribution table to determine significance.

4.4.2 t-test

Let's now assess whether there is a statistically significant difference between the mean literacy level (literacy) and whether the government is a democracy (democ). You can rely on the t.test() function. What are your conclusions?

```
db=poliscidata::world
db=db[,c("democ","literacy")]
db$country = factor(db$democ)
t.test(db$literacy ~ db$democ)
##
##  Welch Two Sample t-test
##
## data:  db$literacy by db$democ
## t = -3.4616, df = 138.73, p-value = 0.0007142
## alternative hypothesis: true difference in means between group No and group Yes is not e
## 95 percent confidence interval:
##  -17.593909  -4.801764
## sample estimates:
##  mean in group No mean in group Yes 
##          74.12239          85.32022
```

Solution: interpretation

The Welch Two Sample t-test results indicate a statistically significant difference in literacy rates between countries categorized as “No” and “Yes” for democracy (p-value < 0.05). The test statistic (t = -3.4616) and degrees of freedom (138.73) support rejecting the null hypothesis, suggesting that the mean literacy rate is not equal between the two groups. Specifically, the mean literacy rate is lower in non-democratic countries (74.12) compared to democratic countries (85.32). The 95% confidence interval for the difference in means ranges from -17.59 to -4.80, further indicating that democratic countries tend to have significantly higher literacy rates.

We can also compare the difference between the means for a categorical variables with more than 2 categories. In this case, we use the ANOVA and the function aov(). Let’s compare the mean literacy level (literacy) for each type of regime (dem_level4). What are your conclusions?

```
db=poliscidata::world
db=db[,c("dem_level4","literacy")]
db$country = factor(db$dem_level4)
summary(aov(db$literacy ~ db$dem_level4))
##                Df Sum Sq Mean Sq F value  Pr(>F)    
## db$dem_level4   3  16942    5647   17.67 6.7e-10 ***
## Residuals     153  48900     320                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 10 observations effacées parce que manquantes
```

Solution: interpretation

The ANOVA results indicate a highly significant effect of the type of regime on literacy rates (F = 17.67, p < 0.05). With 3 degrees of freedom for the factor and 153 degrees of freedom for residuals, the results show that differences in types of regime are associated with significant variations in mean literacy rates.

4.4.3 Pearson correlation test

Finally, let’s correlate the variables for literacy and Gender Inequality Index (gender_unequal). You can use the cor.test function. What is your conclusion?

```
db=poliscidata::world
db=db[,c("gender_unequal","literacy")]
cor.test(db$gender_unequal, db$literacy, method=c("pearson"))
##
##  Pearson's product-moment correlation
##
## data:  db$gender_unequal and db$literacy
## t = -12.695, df = 126, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.81653 -0.66163
## sample estimates:
##        cor
## -0.7491485
```

i Solution: interpretation

The Pearson's product-moment correlation test results indicate a strong and statistically significant negative correlation between gender inequality and literacy rates (p-value < 0.05). The correlation coefficient (r = -0.749) suggests that higher levels of gender inequality are associated with lower literacy rates. The 95% confidence interval for the correlation ranges from -0.816 to -0.662, reinforcing the strength and direction of this relationship. The test statistic (t = -12.695) and degrees of freedom (126) further support the conclusion that the observed correlation is highly unlikely to be due to random chance, indicating a meaningful inverse relationship between these variables.

Nota bene: It is crucial to remember that correlation does not imply causation. This means that although gender inequality and lower literacy rates are related, we cannot conclude that one directly causes the other based solely on this correlation. Other factors could be influencing both variables. Establishing causation would require further research, including experimental or longitudinal studies to determine the direction and nature of the relationship.

Part II

MULTIVARIATE CONTENT

Chapter 5

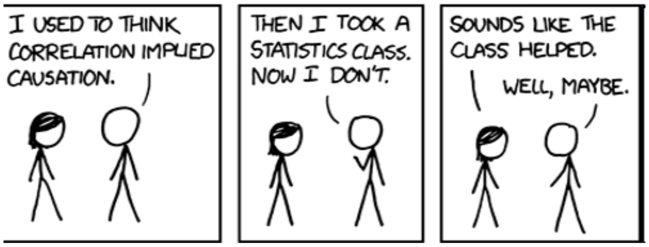
Linear regression

5.1 Linear regression analyis

Linear regression is a central method of analysis in the social sciences. In its simplest form, it resembles bivariate analysis of two metric variables. It is also an important method as it allows an easy transition to multivariate data analysis (multivariate regression) and as it allows an extension to categorical data analysis (logistic, ordinal logit, and multinomial logit models), as well as hierarchical models.

Linear regression helps us in answering typical questions, such as the existence of a relationship (dependence) between an independent variable (e.g., education) and a dependent variable (e.g., income). It also provides us with information about the strength and the robustness of this relationship (e.g., by controlling for other variables in the model).

5.2 Excurse: correlation is no proof of causality


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Social sciences often rely on (survey) data that were collected for a population of interest. When data are collected at a single point in time (rather than repetitively for the same individuals), the main focus of the research is about whether two phenomena are correlated. Correlation examines association of two metric variables (interval/ratio level) and measures the direction and strength to which two variables are related (between -1 and +1), most often using Pearson's correlation coefficient.

Assessing causality between two phenomena requires additional features beyond correlation. It also requires that a phenomenon X comes before a phenomenon Y, thus relying on longitudinal data or, with more theoretical argumentation, cross-sectional data. It also requires that the correlation between X and Y remains if we control for further relevant variables.

5.3 Simple linear regression

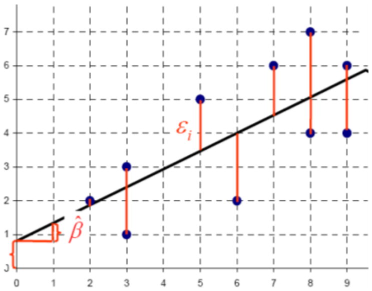
In its simplest form, linear regression is a way of predicting value of one variable Y through another known variable X. It is a hypothetical model which uses the the equation of a straight line (linear model) based on the method of least squares which minimize the sum of squared errors (SSE). What does this mean?

Let's take a basic example covering the relationship between years of education (X) and income (Y). Here, a research question could be: Which income can be attained with ten years of education? Answering this question requires finding the best fitting line between the two variables and provides us with the formula:

$$\mathbf{\hat{y}_i = a + b\, x_i}$$

where $\mathbf{\hat{y}_i}$ is the estimated income for an individual, **a** is the constant (or intercept: value where education equals 0), $\mathbf{x_i}$ is the value of education for an individual, and **b** is the effect size of education on income (or slope: change in income by a change in education). The equation for the slope is represented by dividing the covariance over the variance:

$$\mathbf{b} = \frac{\sum(x-\overline{x}) \times (y-\overline{y})}{\sum(x-\overline{x})^2} = \frac{s_{xy}}{s_x^2}$$



In principle, there are plenty of lines through the data and the line with the lowest SSE represents the line which best fit the data. In our example, the predicted income deviates from the observed income, because all values are not lying precisely on a line. Therefore, we need to estimate a regression line where distances (errors) between the predicted and observed values are minimized. The residuals read as follows:

$$\hat{} = \mathbf{y} - \hat{\mathbf{y}}_{\mathbf{i}} = \mathbf{y}_{\mathbf{i}} - \hat{\alpha} - \hat{\beta}\mathbf{x}_{\mathbf{i}}$$

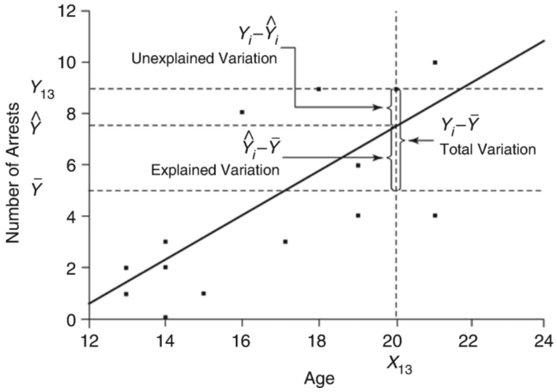
Up to here, we are able to interpret the slope coefficient, which gives us the direction of the effect (+ of **-**) and the extent to which Y changes if X increases by 1. However, we also need to assess the statistical significance of the effect. Indeed, the slope coefficient represents the point estimator (sample) for the "true" population value (β). Here, the standard error provides us with a measure for the variability (dispersion) of **b** and, thereby, the confidence intervals to test significance of the effect. In this case, the null hypothesis states that there is no relation between X and Y (using the t-statistic).

5.4 Total variation

How much variance is there on the Y? We are interested in the overall variation. Therefore, without knowing X, the mean of Y is the best estimate for all Y values. This is also called the "baseline model" and the resulting error is called total variation: $TSS = \sum(Y_i - \overline{Y})^2$.

However, knowing X, the regression line is a better estimate for Y. The resulting error is the residual variation ("residuals"): $SSE = \sum(Y_i - \hat{Y}_i)^2$.

Therefore, the explained variation in the regression model is: $MSS = \sum(\hat{Y}_i - \overline{Y})^2$.



The original figure can be found here.

5.5 Multivariate model

In real world, the relations between two variables are often influenced by "third variables". In this case, the bivariate association is not informative if not controlling for the disturbing influence of a third, fourth, fifth, etc. variable (with different measurement levels). Multivariate regression takes the influence of other variables on the relation between two variables into account. Therefore, it is powerful to detect spurious- and suppressor effects, indirect effects, and interaction effects. This goal is to estimate the "net" influence of several independent variables:

$$\mathbf{\hat{y} = a + b_1\ x_1 + b_2\ x_2 + b_k\ x_k}$$

In multivariate models, the b coefficient represents the "raw" regression coefficient. It is important to note that the comparison of b coefficients is not meaningful if the variables are measured in different units (e.g. working time in hours & family size in number of individuals). A possible solution is the standardization of the variables (mean=0 & standard deviation=1) in view of assessing which independent variable has greater effect on Y (relative contribution). The standardized b (written by convention β) are often referred as the beta coefficients. In a simple linear model, β corresponds to Pearson r. The interpretation of the β coefficient indicates by how many standard deviations Y varies when X varies by one standard deviation.

5.6 Explained proportion of the variance

R^2 provides us with a measure of accuracy of the regression model, that is how well does the regression line approximate the real data points. It represents the share of explained variance (how much variation in Y is explained through X) and varies between 0 and +1. R^2 tends to increase as we increase the number of variables in the model.

R^2 does not tell us whether the independent variables are true cause of changes in Y (no causality), whether omitted-variable bias exists (third variable problem), whether the correct regression was used, and whether appropriate independent variables have been chosen. In research, our focus is often only one part of a complex story.

ℹ R^2 explained

In addition to the direction, strength and significance of the effect of the **b** coefficient, multivariate model also provide a coefficient of determination R^2, which can be expressed as follows:

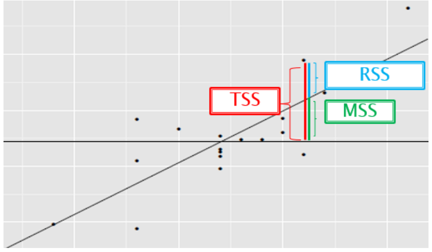
$$R^2 = \frac{TSS - SSE}{TSS} = \frac{MSS}{TSS}$$

$$TSS = \sum (y_i - \overline{y})^2$$

$$MSS = \sum (\hat{y}_i - \overline{y})^2$$

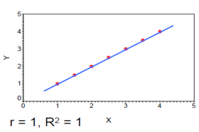
$$SSE = \sum (y_i - \hat{y}_i)^2$$

MSS is the model sum of squares (also known as ESS, or explained sum of squares). It is the sum of the squares of the prediction from the linear regression minus the mean for that variable.
TSS is the total sum of squares associated with the outcome variable, which is the sum of the squares of the measurements minus their mean.
RSS is the residual sum of squares, which is the sum of the squares of the measurements minus the prediction from the linear regression.

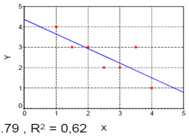


5.7 Relationship between r and R^2

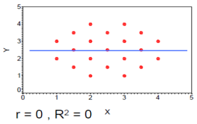
If X and Y are two metric variables, R^2 is a measure of a linear relation between X and Y, while r (Pearson correlation) is the empirical correlation between X and Y. Importantly, in bivariate regression, $R^2 = r^2_{xy}$ (where r is a standardized regression coefficient). In multiple regression, r represents the correlation between the observed y values and the predicted $\hat{y}$ values.



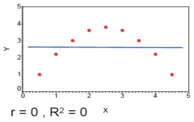
r = 1, R^2 = 1



r = -0.79 , R^2 = 0,62



r = 0 , R^2 = 0



r = 0 , R^2 = 0

5.8 F-test for the regression model

F-test is a global check of the regression model. It tests the hypothesis whether in the population there is a connection between the Y and the X:

- H0: no connection (or $R^2 = 0$)
- H1: there is a connection (or R^2 not equal 0)

F-test explained

$$MS_M = \frac{MSS}{k}$$

$$MS_R = \frac{SSE}{N-k-1}$$

, where k is the number of independent variables and N is the number of individuals.

$$F = \frac{MS_M}{MS_R}$$

5.9 Standard estimation error

The standard estimation error characterizes the spread of the y-values around the regression line and is therefore a quality measure for the accuracy of the regression prediction:

$$s_e = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n-k-1}}$$

The standard error of estimation comes from the fact that:

- Part of the estimation error can be attributed to chance (e.g. “human randomness”)
- Part is determined by predictors that are not included in the model
- Part of the estimation error arises from measurement errors

FAQ on the standard error of estimation

What contributes to the standard error of estimation? The standard error of estimation increases with:

- large k, since k is subtracted in the denominator
- if the residuals are large.

Nota bene: if n is large, the estimation error does not automatically decrease! This is because for every observation there is also a deviation that is included in the calculation of the total.

How do standard estimation errors, intercept and slope behave to each other? The intercept and slope have nothing to do with explained or unexplained variance, so nothing to do with the standard error of estimation either.

5.10 Overfitting

The regression equation is always exactly identified if the number of observations is only one higher than the number of independent variables (if: n-1 = k). There are only sufficient "degrees of freedom" if there are more measuring points (observations) than variables. Only if the deviations are small, then we are talking about a high quality of fit of the regression calculation.

However, specific sample features only apply to one sample, not to the total population and not to other samples either. They should not be allowed to contribute to model quality. As a consequence, the estimate always tends to be too high. That means:

- The estimated Y values and the empirical Y values correlate (probably) stronger than in reality
- The residuals are smaller than they should actually be (according to the conditions in the population)
- The explained variance is consequently greater than it should be

5.11 Shrinkage

The determination coefficient often turns out to be smaller with a new sample. This phenomenon/principle is called shrinkage. Therefore, a corrected R^2 (adjusted R^2) takes this into account by reducing the simple R^2 by a correction value that is larger: the larger the number of predictors and the smaller the number of cases (whereas the simple R^2 increases with each added predictor, the corrected R^2 decreases as more predictors are added).

ℹ Adjusted R^2 explained

Adjusted R^2 is a corrected goodness-of-fit (model accuracy) measure for linear models. Because R^2 tends to optimistically estimate the fit of the linear regression (as it typically increases as the number of effects are included in the model), adjusted R^2 attempts to correct for this overestimation. Adjusted R^2 might even decrease if a specific effect does not improve the model.
Adjusted R^2 is calculated as follows:

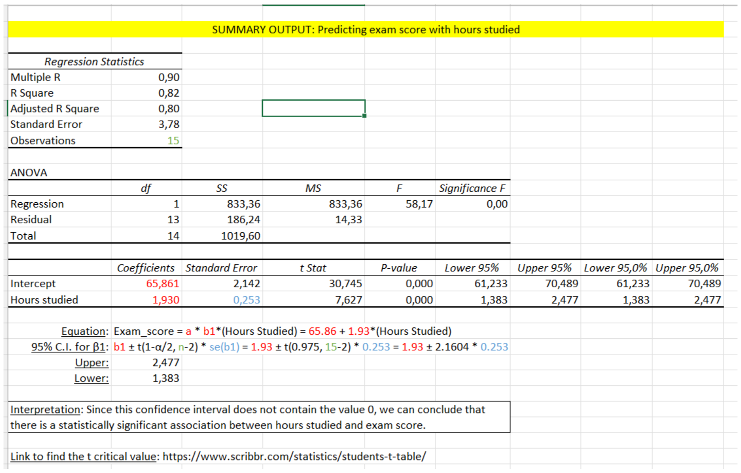
$$R^2_{adj} = 1 - [\frac{(1-R^2)(n-1)}{(n-k-1)}]$$

where n represents the number of data points and k the number of independent variables.
So, if R^2 does not increase significantly on the addition of a new independent variable, then the value of adjusted R^2 will actually decrease. But, if adding the new independent variable leads to a significant increase in R^2 value, then the adjusted R^2 value will also increase.

5.12 Statistical significance

A t-test is used to determine whether the regression coefficients deviate significantly from zero (H0). To calculate the t-value per coefficient, the standard error (S_b) of each coefficient is also necessary. The S_b characterizes the spread of the regression coefficients around the population parameter and is therefore a quality measure for the accuracy of the parameter estimation. The larger the SE, the smaller the empirical t-value and the less likely it is that the H0 will be rejected.

The significance of the effect is based on the ratio between a coefficient and its standard error. This is the empirical value of the t-test ($t = b/S_b$), which is compared with a critical value (c.v.=1.96). The null hypothesis (H0) is accepted if the critical value is $<$ 1.96. The confidence interval should not contain 0.



SUMMARY OUTPUT: Predicting exam score with hours studied

Regression Statistics	
Multiple R	0,90
R Square	0,82
Adjusted R Square	0,80
Standard Error	3,78
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	833,36	833,36	58,17	0,00
Residual	13	186,24	14,33		
Total	14	1019,60			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	65,861	2,142	30,745	0,000	61,233	70,489	61,233	70,489
Hours studied	1,930	0,253	7,627	0,000	1,383	2,477	1,383	2,477

Equation: Exam_score = a * b1*(Hours Studied) = 65.86 + 1.93*(Hours Studied)
95% C.I. for β1: b1 ± t(1-α/2, n-2) * se(b1) = 1.93 ± t(0.975, 15-2) * 0.253 = 1.93 ± 2.1604 * 0.253
Upper: 2,477
Lower: 1,383

Interpretation: Since this confidence interval does not contain the value 0, we can conclude that there is a statistically significant association between hours studied and exam score.

Link to find the t critical value: https://www.scribbr.com/statistics/students-t-table/

Figure 5.1: Confidence interval for the regression slope

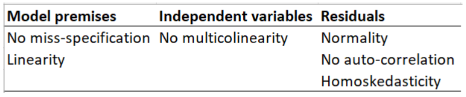
5.13 Dummy variables

Categorical independent variables are problematic because the numeric codes assigned to their categories are arbitrary (when the code changes, the estimate changes). A solution suggests associating a binary dummy variable with each modality (coded 0 and 1). A categorical variable with c modalities is replaced by c-1 dummies.

The omitted modality serves as a reference category. The choice of the reference category is made on the basis of empirical considerations (e.g., modality with the largest number of cases or modality that makes sense from a theoretical point of view). The coefficients b are interpreted with respect to the reference category.

5.14 Postulates and assumptions

Linear regression requires using numerical variables (or dichotomous for the independent variables). A dichotomous variable is always coded 1 and 0. Linear regression entails several premises:



Model premises	Independent variables	Residuals
No miss-specification	No multicolinearity	Normality
Linearity		No auto-correlation
		Homoskedasticity

A regression model should contain exactly those independent variables that are relevant for the dependent variable, no more and no less. Bivariate regression models, for example, are always misspecified. The risk with having too few variables is underfitting, thus leading to too high regression coefficients insufficient explanations. The risk with too many/irrelevant variables is overfitting, thus leading relevant significances to disappear and/or random significance to arise. To conduct linear regression, we must have sufficient degrees of freedom. In general, this corresponds to at least 10 (or 20) times more cases than variables.

Linear regression only works well for (almost) normally distributed variables. When normality does not apply, we might think about transforming the variables (e.g., log, powers, etc.).

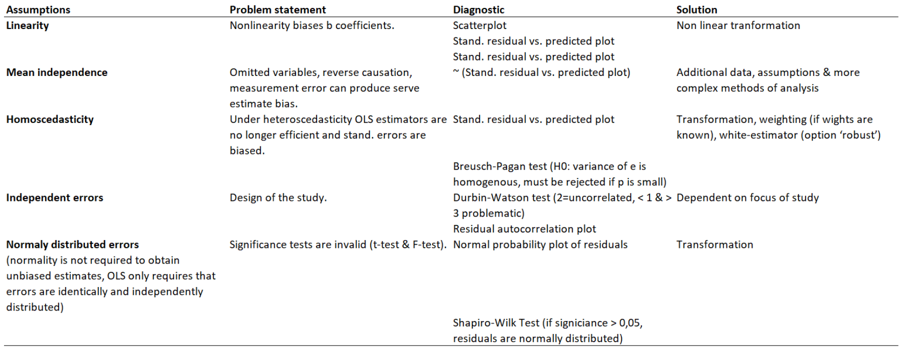
There must be an independence of the independent variables, that is an absence of multicollinearity between the independent variables. Regression coefficients can be biased if there is a strong correlation between two independent variables.

Furthermore, the error terms must be distributed according to a normal law and should be zero on average.

There must also be an independence of error terms (to be checked when we have time series).

Linear regression also requires homoscedasticity, that is the fact that the variance of the error terms is constant for all values of the independent variables.

The premises can be tested as follows:



Assumptions	**Problem statement**	**Diagnostic**	**Solution**
Linearity	Nonlinearity biases b coefficients.	Scatterplot Stand. residual vs. predicted plot Stand. residual vs. predicted plot	Non linear tranformation
Mean independence	Omitted variables, reverse causation, measurement error can produce serve estimate bias.	~ (Stand. residual vs. predicted plot)	Additional data, assumptions & more complex methods of analysis
Homoscedasticity	Under heteroscedasticity OLS estimators are no longer efficient and stand. errors are biased.	Stand. residual vs. predicted plot	Transformation, weighting (if wights are known), white-estimator (option ‘robust’)
		Breusch-Pagan test (H0: variance of e is homogenous, must be rejected if p is small)	
Independent errors	Design of the study.	Durbin-Watson test (2=uncorrelated, < 1 & > 3 problematic) Residual autocorrelation plot	Dependent on focus of study
Normaly distributed errors (normality is not required to obtain unbiased estimates, OLS only requires that errors are identically and independently distributed)	Significance tests are invalid (t-test & F-test).	Normal probability plot of residuals	Transformation
		Shapiro-Wilk Test (if signiciance > 0,05, residuals are normally distributed)	

Figure 5.2: Linear regression diagnostics

5.14.1 Multicollinearity

Multicollinearity is not a severe violation of linear regression (only in extreme cases). In the presence of multicollinearity, the estimates are still consistent, but there is an increase in standard errors (estimates are less precise). The problem occurs when researchers include many highly correlated variables (e.g., age and birth year). Therefore, there should be a theory-driven, careful, and intelligent variable selection.

Collinearity statistics include the Tolerance which varies from 0 to 1 and must be >0.4. The VIF (variance inflation factor) measure can also be used which varies from 1 to ∞ and must be <2.5 (which corresponds to a Tolerance of 0.4). Note that there is an inverse relationship between Tolerance and VIF values. As such, Tolerance values >0.4 and VIF values <2.5 are signs of no serious multicollinearity among the variables.

If there is multicollinearity, there are basically two solutions: either eliminating one (or more) problematic variables, or merging the problematic variables (e.g., in a scale).

- Tolerance $= 1 - R^2$
- VIF $= 1/(1 - R^2)$

5.14.2 Normal distribution of the residuals

The residuals must be normally distributed. For instance, the deviations between estimated and observed values should be zero or close to zero in most cases. Risks of non-normal residuals result in biased results, thus the significance tests are biased. As a reminder:

- for every empirical value Y there is an estimated value Y and a residual

- the larger the residual for an estimate, the worse the regression estimate for this estimate
- the larger the residuals are overall, the larger the standard estimation error (“mean” of the squared residuals), and the poorer the overall goodness of fit of the regression estimate

5.14.3 Autocorrelation of the residuals (independent errors)

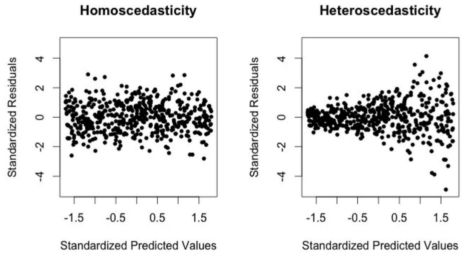
The non-independence of the residuals occurs when sampling is not independent of one another, i.e. systematically related (time series data, periodic surveys). In these cases, the risks are:

- Bias in the standard error of the regression coefficients
- Bias in the confidence interval for the regression coefficients

We can verify this using the Durbin-Watson statistics.

5.14.4 Homoscedasticity

Heteroscedasticity is present when there is a relationship between the estimated y-values and the residuals of the observation values. This can be verified by a visual inspection - no pattern should be discernible.



5.15 Outliers

The impact of an observation is depending on two factors. First, the discrepancy (or outlierness) which assesses observations with large residual (observations whose Y value is unusual given its values on X). Second, the leverage which assesses observations with extreme value on X with the general rule:

$$lev > (2k + 2)/n)$$

- A “high leverage” outlier impacts the model fit, may not have big residuals, and can increase R^2.
- A “low leverage” outlier has a lower impact on the model fit, has usually a big residual, inflates standard errors, and decreases R^2.

There are several strategies to identify outliers, such as conducting frequency analysis one variable, using the scatterplots for two variables, looking at the n highest and lowest values, and investigating the residuals (through partial residual plots or added value plots), as well as relying on influential measures. For instance:

- Cook's Distance with the general rule: $d > 4/(n - k - 1)$
- dfbeta with the general rule: $2/\sqrt{n})$

Note that there might be cases where we want to keep outliers. This may include cases where we have meaningful cluster (e.g., important subgroup in the data) or cases where outliers might reflect a "real" pattern in the data. In these cases, it is important to present the results both with and without outliers.

5.16 Interaction effects

Interaction effects enable us to model non-additive effects. In additive models, the effects of independent variables are the same for the complete sample (e.g., effect of education on income is the same for women and men). But, theoretically, effect should differ for different groups (e.g., women and men).

Interaction effects suggest that the strength of the association between x_1 and$y$ is dependent on the level of a third variable x_2:

$$y = a + b_1 x_1 + b_2 x_2 + b_3(x_1 \times x_2) + \epsilon$$

5.16.1 Recap on variable types

Categorical variables (qualitative variables) refer to a characteristic that can't be quantifiable. They can be either nominal or ordinal. Nominal variables describe a name, label or category without natural order (e.g., sex, marital status, race, etc.). Ordinal variables are defined by an order relation between the different categories (e.g., ranking of students, sport clubs, social class, etc.).

Numeric variables (quantitative variables) is a quantifiable characteristic whose values are numbers. Numeric variables may be either continuous or discrete. Discrete (or interval) variables assume only a finite number of real values within a given interval (e.g., people in a household, temperature, etc.). Continuous (or ratio) variables can assume an infinite number of real values within a given interval (e.g., height/weight of a person).

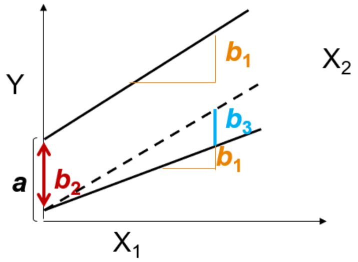
5.16.2 Interaction between interval variables

For instance, we might be interested in the effect of age (x_1) on income (y) depending on years of schooling (x_2). In this case, the interpretation of the main effect of x_1 is the effect of x_1 if x_2 equals 0 (and conversely for x_2). The interpretation of the interaction effect is done by calculating the effect of x_1 on y for different levels of x_2 (and conversely for x_2).

Note that the interpretation of interval variables might be problematic if the zero value is not meaningful (e.g., age=0). A better interpretation of the main effects can be obtained by centering the variables (through subtracting the mean score from each data-point). In this case, the interpretation of the intercept (a) also changes and represents the predicted value of y if x is the average.

5.16.3 Interaction between dummy and interval variables

For instance, we might be interested in how the association between years of education (x_1) and income (y) varies by gender (x_2, where 1=women and 0=men). In this case, the slope (b) and intercept (a) of the regression of x_1 on y are dependent on specific values of x_2. Therefore, for each individual value of x_2 (e.g., 0/1) there is a regression line. The general interpretation rules suggest that the main effect of x_1 represents the effect of x_1 if x_2 equals 0 (and conversely for x_2). Thus, the interaction effect indicates how much the effect of x_1 changes with an unit change in x_2 (and conversely for x_2).



5.16.4 Interaction between ordinal and nominal variables

For instance, we might be interested in how the association between the fact of having a child (x_1, where 1=having (at least) a child and 0=no child) and income (y) varies by gender (x_2, where 1=women and 0=men).

5.17 How it works in R?

See the lecture slides on linear regression:

You can also download the PDF of the slides here:

5.18 Quiz

True

False

Statement

The coefficients of the least squares regression line are determined by minimizing the sum of the squares of the residuals.

A residual is computed as an x-coordinate from the data minus an x-coordinate predicted by the line.

The critical value for a confidence interval for the slope of the least squares regression line for all pairs in the population does not depend on the confidence level.

The variance inflation factor (VIF) can be used to identify this issue of multicollinearity.

View Results

My results will appear here

5.19 Example from the literature

The following article relies on linear regression as a method of analysis:

Rauchfleisch, A., & Metag, J. (2016). The special case of Switzerland: Swiss politicians on Twitter. New Media & Society, 18(10), 2413-2431. Available here.

Please reflect on the following questions:

- What is the research question of the study?
- What are the research hypotheses?
- Is linear regression an appropriate method of analysis to answer the research question?
- What are the main findings of the linear regression analysis?

5.20 Time to practice on your own

You can download the PDF of the exercises here:

The exercises 1 and 2 will use the data from the round 10 of the European Social Survey (ESS). You can download the data directly on the ESS website.

The objective is to conduct linear regression to explain the consumption of news about politics and current affairs (‘nwspol’: watching, reading or listening in minutes). Explanatory variables include a set of political and sociodemographic variables. Political variables can include: political interest (‘polintr’), confidence in ability to participate in politics (‘cptppola’), and self-placement on the left-right political scale (‘lrscale’). Sociodemographic variables can include: gender (‘gndr’), age (‘agea’), and years of education (‘eduyrs’). You can also decide to rely on additional or alternative variables.

Let’s start by loading the dataset, selecting the relevant variables, and filtering the data for Switzerland only:

```
library(foreign)
db <- read.spss(file=paste0(getwd(),
                  "/data/ESS10.sav"),
                use.value.labels = F,
                to.data.frame = T)
sel <- db |>
  dplyr::select(cntry, nwspol, polintr, cptppola, lrscale, gndr, agea, eduyrs) |>
  stats::na.omit() |>
  dplyr::filter(cntry=="CH") # select respondents from Switzerland
# verify the class and range of the variables
sel$nwspol=as.numeric(sel$nwspol)
sel = sel[sel$nwspol<=180,] # maximally 3hours of news consumption
sel$gndr = as.factor(sel$gndr)
sel$gndr = ifelse(sel$gndr=="2", "female", "male")
sel$gndr = as.factor(as.character(sel$gndr))
```

5.20.1 Example 1: stepwise regression

Conduct your own linear regression analysis by following a stepwise logic:

- The variable that shows the highest correlation with the dependent variable is selected (in absolute terms)

```
round(cor(sel[,c("nwspol", "polintr", "cptppola", "lrscale", "agea", "eduyrs")]),2)
##          nwspol polintr cptppola lrscale  agea eduyrs
## nwspol     1.00   -0.32     0.11    0.05  0.31  -0.01
## polintr   -0.32    1.00    -0.47    0.07 -0.19  -0.20
## cptppola   0.11   -0.47     1.00   -0.04 -0.06   0.19
## lrscale    0.05    0.07    -0.04    1.00  0.16  -0.20
## agea       0.31   -0.19    -0.06    0.16  1.00  -0.12
## eduyrs    -0.01   -0.20     0.19   -0.20 -0.12   1.00
reg1 =lm(nwspol ~ polintr, data=sel)
summary(reg1)
##
## Call:
## lm(formula = nwspol ~ polintr, data = sel)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -73.89 -29.59 -13.89  16.11 134.71
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   88.196      2.826   31.21   <2e-16 ***
## polintr      -14.301      1.162  -12.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.12 on 1346 degrees of freedom
## Multiple R-squared:  0.1011, Adjusted R-squared:  0.1005
## F-statistic: 151.4 on 1 and 1346 DF,  p-value: < 2.2e-16
```

> **Interpretation**
>
> The variable measuring political interest displays the highest correlation (in absolute term) with the news consumption. Note that the correlation is negative because political interest in measured with the item “How interested would you say you are in politics - are you…” where the response scale is as follows: 1 for “Very interested” and 4 for “Not at all interested”. Ideally, we would reverse the scale before running the analyses and interpreting the results.
> Concerning the regression model, the coefficient for political interest is -14.301 (so, +14.301 with a reversed scale) and its effect is significantly impacting news consumption ($p < 0.001$). The proportion of explained variance is given by an R^2 of 0.10 (or 10%).

- The variable that has the highest semi-partial correlation with the dependent variable from the remaining variables is then selected (in absolute terms). The semi-partial correlation coefficient is the correlation between all of Y and that part of X which is independent of Z.

```
res1 = resid(reg1)
round(cor(res1, sel[,c("cptppola", "lrscale", "agea", "eduyrs")]),2)
##      cptppola lrscale agea eduyrs
## [1,]    -0.05    0.07 0.26  -0.08
summary(lm(nwspol ~ polintr + agea, data=sel))
##
## Call:
## lm(formula = nwspol ~ polintr + agea, data = sel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -93.076 -24.208  -8.404  17.573 139.423
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 55.94452    4.16625   13.43   <2e-16 ***
## polintr    -12.05596    1.14124  -10.56   <2e-16 ***
## agea         0.54653    0.05343   10.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 35.77 on 1345 degrees of freedom
## Multiple R-squared:  0.166,  Adjusted R-squared:  0.1648 
## F-statistic: 133.9 on 2 and 1345 DF,  p-value: < 2.2e-16
```

Interpretation

The variable with the highest semi-partial correlation is age. Therefore, we include it in the regression model. We see that the effect of political interest remains significant. Furthermore, the effect of age is positively impacting news consumption (coef = 0.54) and that it is also significant (p < 0.001). Note that we cannot compare the effect of political interest and age as the variables have not been standardized. The proportion of explained variance is given by R^2 and is now 0.16 (or 16%). Each time check whether adding more variables significantly improves the model.

5.20.2 Example 2: deductive approach

Conduct your own linear regression analysis by following a deductive logic:

- Include the independent variables simultaneously in the regression equation

```
regbt = lm(nwspol ~
             polintr +
             cptppola +
             lrscale +
             agea +
             eduyrs +
             relevel(gndr,"male"),
           data=sel)
summary(regbt)
## 
## Call:
## lm(formula = nwspol ~ polintr + cptppola + lrscale + agea + eduyrs + 
##     relevel(gndr, "male"), data = sel)
## 
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max 
## -93.288 -24.295  -7.841  17.618 139.605 
## 
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                  62.29255    7.80393   7.982 3.06e-15 ***
## polintr                     -12.57088    1.33517  -9.415  < 2e-16 ***
## cptppola                     -0.30173    1.06016  -0.285    0.776    
## lrscale                       0.28449    0.49602   0.574    0.566    
## agea                          0.52845    0.05564   9.498  < 2e-16 ***
## eduyrs                       -0.32865    0.26416  -1.244    0.214    
## relevel(gndr, "male")female  -2.18405    2.00285  -1.090    0.276    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 35.77 on 1341 degrees of freedom
## Multiple R-squared:  0.1685, Adjusted R-squared:  0.1648 
## F-statistic:  45.3 on 6 and 1341 DF,  p-value: < 2.2e-16
```

Interpretation

Adding all the variables together in the model shows that only political interest and age are significantly impacting on news consumption. The proportion of explained variance is given by R^2 and remains at 0.16 (or 16%). Adding more variables did not significantly improve the model.

- Include hierarchically (blockwise) the independent variables into groups and include them in the regression equation group by group (e.g. sociodemographic variables first and political variables then).

```
regbs = lm(nwspol ~ 
             agea + 
             eduyrs + 
             relevel(gndr,"male"), 
           data=sel)
summary(regbs)
## 
## Call:
## lm(formula = nwspol ~ agea + eduyrs + relevel(gndr, "male"), 
##     data = sel)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -81.489 -27.006  -8.368  19.546 149.590 
## 
```

```
## Coefficients:
##                              Estimate Std. Error t value   Pr(>|t|)    
## (Intercept)                  21.32329    4.49269   4.746 0.00000229 ***
## agea                          0.66480    0.05493  12.103    < 2e-16 ***
## eduyrs                        0.29096    0.26232   1.109     0.2676    
## relevel(gndr, "male")female -4.16850    2.03222  -2.051     0.0404 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.17 on 1344 degrees of freedom
## Multiple R-squared:  0.1002, Adjusted R-squared:  0.09822
## F-statistic:  49.9 on 3 and 1344 DF,  p-value: < 2.2e-16
```

Interpretation

Including only sociodemographic variables (age, gender and years of education) shows that age and gender (with "male" as the reference category) are statistically significant (p $<$ 0.001 and p = 0.04 respectively). The proportion of explained variance is 0.10 (or 10%).

Each time check whether the (groups of) variables significantly improve the model.

```
cat(paste0("R2 from the model with sociodemo variables is: ",
             round(summary(regbs)$adj.r.squared,3),
             "\n",
             "R2 from the model including all the variables is: ",
             round(summary(regbt)$adj.r.squared,3)))
## R2 from the model with sociodemo variables is: 0.098
## R2 from the model including all the variables is: 0.165
```

Interpretation

The proportion of explained variance is 0.10 (or 10%) for the model with only sociodemographic variables, while it is 0.16 (16%) for the model containing all the variables. However, we have seen that only age and political interest are significantly related to news consumption (gender looses its significance in the model containing all the variables).

5.20.3 Example 3: postulates and assumptions

Based on the previous regression model, evaluate the assumptions of linear regression: no multicolinearity, normality of the residuals, no auto-correlation of the residuals (especially for time series and panel data), homoscedasticity. It is also important to check for outliers.

First, we want to check for multicollinearity. In case of multicollinearity issue, regression model is not able to accurately associate variance in the outcome variable with the correct predictor variable, leading to incorrect inferences. Beyond theoretical reflections, there are several steps to test for multicollinearity issues, such as correlation, VIF (and tolerance):

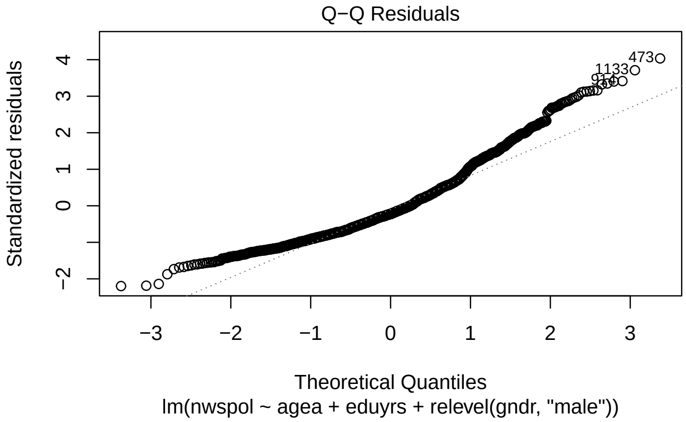
```
c <- sel[,-c(1,6)] # removes cntry and gndr
round(cor(c),3)
##          nwspol polintr cptppola lrscale   agea eduyrs
## nwspol    1.000  -0.318    0.105   0.046  0.311 -0.014
## polintr  -0.318   1.000   -0.471   0.072 -0.192 -0.202
## cptppola  0.105  -0.471    1.000  -0.037 -0.057  0.187
## lrscale   0.046   0.072   -0.037   1.000  0.158 -0.203
## agea      0.311  -0.192   -0.057   0.158  1.000 -0.124
## eduyrs   -0.014  -0.202    0.187  -0.203 -0.124  1.000
olsrr::ols_vif_tol(regbs)
##                     Variables Tolerance      VIF
## 1                        agea 0.9838041 1.016463
## 2                      eduyrs 0.9789747 1.021477
## 3 relevel(gndr, "male")female 0.9939485 1.006088
```

Interpretation

The correlation matrix does not point to very high correlations (e.g., > 0.7). Furthermore, the VIF (<2.5) and Tolerance (>0.4) show no sign of multicollinearity.

Second, we can test for the normality of the residuals. We can examine the normal Predicted Probability (P-P) plot to determine if the residuals are normally distributed (ideally, they they will conform to the diagonal normality line):

```
plot(regbs, 2)
```

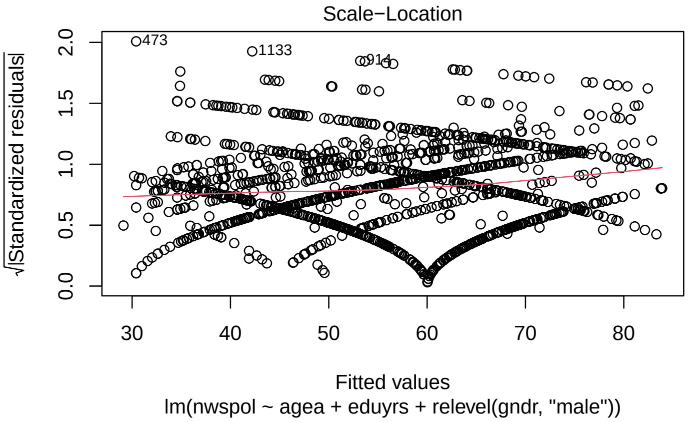


i Interpretation

In order to make valid inferences, the residuals of the regression (differences between the observed value of the dependent variable and the predicted value) should follow a normal distribution. This is approximately the case here.

Third, we can check for homoscedasticity (variance of the error terms should be constant for all values of the independent variables). In the context of t-tests and ANOVAs, the same concept is referred to as equality (or homogeneity) of variances. We can check this by plotting the predicted values and residuals on a scatterplot:

```r
plot(regbs, 3)
```

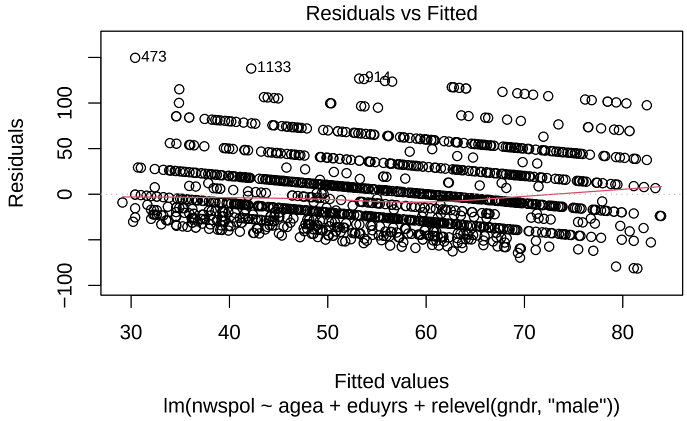


ℹ Interpretation

The residuals need to be spread equally along the ranges of predictors (ideally, there should be a horizontal line with equally spread points). This is the case here.

We can also check for linearity of the residuals:

```
plot(regbs, 1)
```



ℹ Interpretation

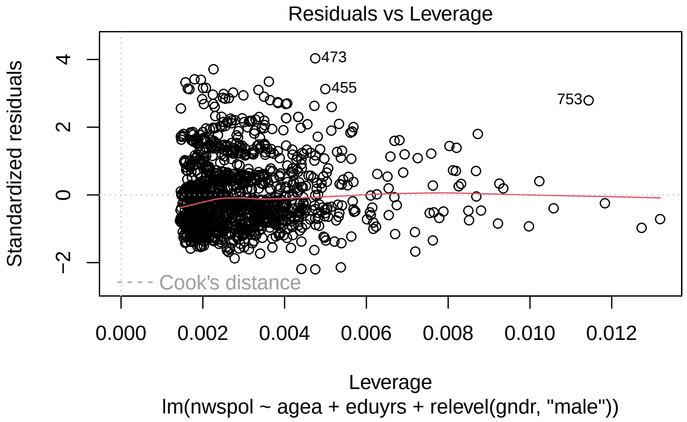
The predictor variables in the regression should have a straight-line relationship with the outcome variable (ideally, the plot would not have a pattern where the red line is approximately horizontal at zero). This is the case here.

Nota bene: Using the Durbin-Watson test, we can test the null hypothesis stating that the errors are not auto-correlated with themselves (if p-value > 0.05, we would fail to reject the null hypothesis).

```
# car::durbinWatsonTest(regbs)
```

Fourth, we can also verify if we have outliers. A value $> 2(p+1)/n$ indicates an observation with high leverage, where p is the number of predictors and n is the number of observations (in our case: 2*(3+1)/1348=0.006).

```
plot(regbs, 5)
```



To extract outliers, you might want to flag observations whose leverage score is more than three times greater than the mean leverage value as a high leverage point.

```
model_data <- broom::augment(regbs)
high_lev <- dplyr::filter(model_data,.hat>3*mean(model_data$.hat))
high_lev
## # A tibble: 10 x 11
##    .rownames nwspol  agea eduyrs `relevel(gndr, "male")` .fitted .resid     .hat .sigma
##    <chr>      <dbl> <dbl>  <dbl> <fct>                     <dbl>  <dbl>    <dbl>  <dbl>
## 1 40             20    45     25 female                     54.3 -34.3  0.00998   37.2
## 2 140            60    31     25 female                     45.0  15.0  0.0102    37.2
## 3 345            15    39     27 female                         50.9 -35.9  0.0127    37.2
## 4 425            20    20      0 male                           34.6 -14.6  0.0106    37.2
## 5 536            60    37     24 male                           52.9   7.10 0.00935   37.2
## 6 725            80    76      0 female                         67.7  12.3  0.00925   37.2
## # i 4 more rows
## # i 2 more variables: .cooksd <dbl>, .std.resid <dbl>
```

5.20.4 Example 4: suppressor effect

Suppose a new variable X2 is added to the regression equation in addition to X1. Suppose, X1 explains substantial variance of Y because both variables capture well a certain phenomenon (here: B). Under which circumstances does this increase the model quality (R^2)?

Normally, an increase in R^2 can only be expected if:

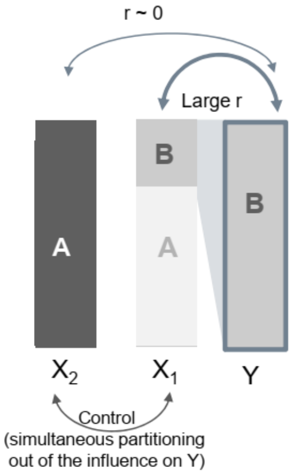
- X2 is correlated with Y: when both X2 and Y capture a particular phenomenon (here: C)
- X1 and X2 are only weakly correlated because they predominantly capture different phenomena (here: B and C)

In cases where this ideal scenario is not or only limited valid, R2 will hardly increase, and may even weaken the influence of X1 on Y when X2 is added.

Now, suppose a predictor variable X1 captures 70% of phenomenon A and 30% of another phenomenon B. Y, on the other hand, captures phenomenon B dominantly. Then the variable X1 will correlate only very moderately with Y since the dominance of phenomenon B in Y has virtually prevented a higher correlation. Suppose a new variable X2 is added to the regression equation. Under what circumstances does this increase the model quality?

ℹ Response: suppressor effect

Assume that a second predictor variable X2 also dominantly captures phenomenon A. Phenomenon B is only weakly or not at all captured in X2. Since the predictors are controlled simultaneously and reciprocally, the influence of the first predictor variable X1 is “freed” from the dominant influence of phenomenon A and suddenly phenomenon B dominates, which in turn is also dominant in Y. Suddenly there is a strong influence of the predictor variable X1 on Y! In this case the second variable X2 is suppressor for the influence of X1 on Y. Why? Only the residual variance of X1 remains for correlations with Y, but this has very large common variance components with Y.


r ~ 0
Large r
A
B
A
B
X2
X1
Y
Control
(simultaneous partitioning
out of the influence on Y)

Chapter 6

ANOVA

6.1 What is ANOVA?

An ANOVA (Analysis of Variance) is used to determine whether or not there is a statistically significant difference between the means of three or more independent groups. The two most common types of ANOVA are the one-way ANOVA and two-way ANOVA.

The one-way analysis of variance (ANOVA), also known as one-factor ANOVA, is an extension of the independent two-sample t-test for comparing means when more than two groups are present. You can use the t-test if you just have two groups. The F-test and the t-test are equivalent in this scenario.

6.1.1 One-way and two-way ANOVA

One-way ANOVA is used to determine how one factor (or independent variable, IV) impacts a response variable (or dependent variable, DV).

- Example: political affiliation impacts on media coverage.

Two-way ANOVA is used to determine how two factors impact a response variable, and to determine whether or not there is an interaction between the two factors on the response variable.

- Example: political affiliation and gender impact on media coverage.

6.1.2 Regression versus variance analyses

Regression analysis:

- Relationships between two variables
- Metric independent variables
- Logic: "The bigger X, the bigger Y"

- Useful to evaluate survey data

Analysis of variance:

- Differences between two or more groups
- Categorical independent variables
- Logic: “More in group A than in group B”
- Useful for evaluating experimental data

But, both methods are based on the same principle, the general linear model and the F-test.

Regression: The quality of the model results from the deviation of the individual values from the regression line

ANOVA: Model quality results from the deviation of the individual values from the group mean

6.1.3 Basic hypotheses of ANOVA

The data is divided into several sub-groups using one single grouping variable (also called factor variable). The basic hypotheses for ANOVA tests are:

- Null hypothesis: the means of the various groups are identical.
- Alternative hypothesis: at least one sample mean differs from the rest.

ANOVA can only be used when the observations are gathered separately and randomly from the population described by the factor levels. That is, each factor level’s data has to be normally distributed and the variance in the sub-populations must be similar (this can be verified using Levene’s test).

6.1.4 T-test and F-test

If you want to compare the means of two groups, you can do a t-test.

If you want to compare the means of more than two groups, one would have to compare each group with each other. However, the alpha error cumulation arises as the error probability of 5% is reclaimed each time for pairwise comparisons:

$$\alpha = 1 - (1 - \alpha)n$$

where n is the number of tests.

The solution consists in using a test that checks whether at least 2 groups differ significantly from each other: F-test, which has no alpha-error accumulation.

6.1.5 Decomposition of the variance

To find the total amount of variation within our data we calculate the difference between each observed data point and the grand mean. We then square these differences and add them together to give us the total sum of squares (TSS, also called SST on the next figure).

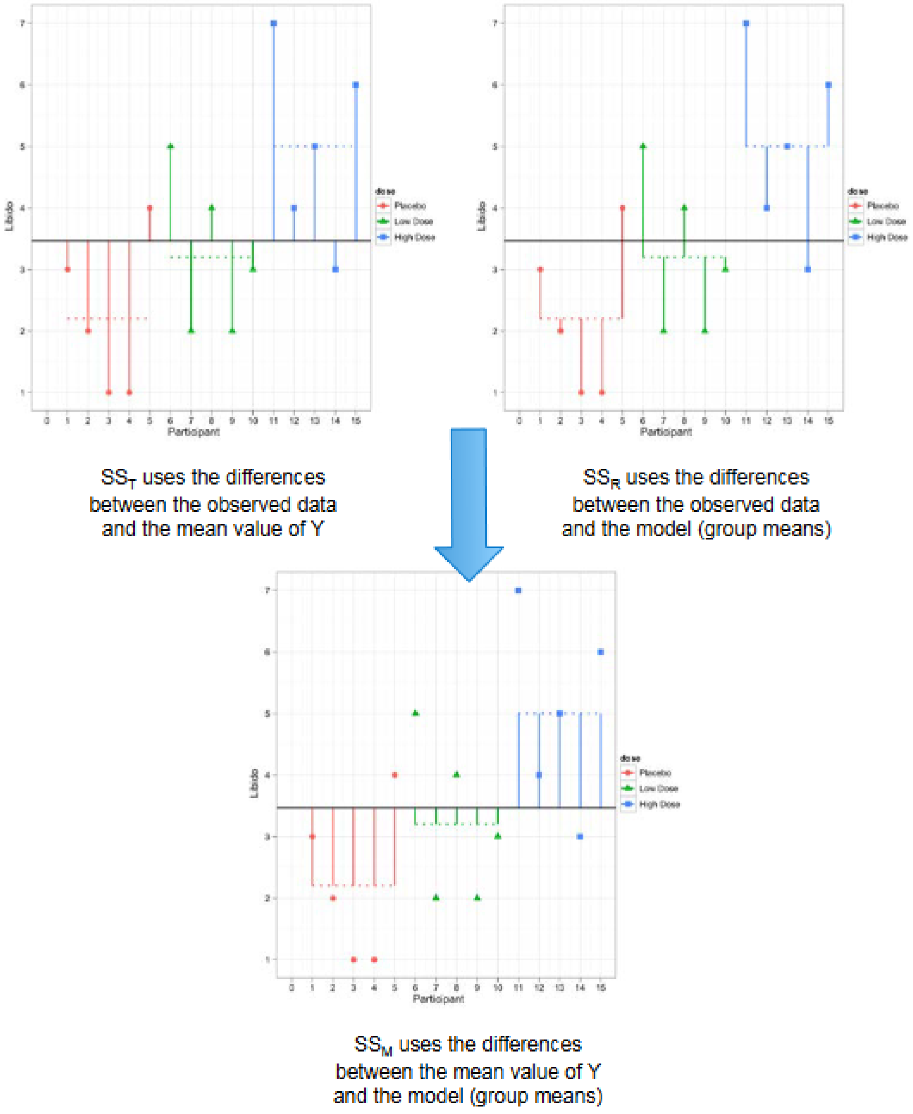
$$TSS = \sum (\hat{y}_{ij} - \hat{y}_{Grand})^2$$

The model sum of squares (MSS, also called SSM on the next figure) requires us to calculate the differences between each participant's predicted value and the grand mean: it is the sum of the squared distances between what the model predicts for each data point (i.e., the dotted horizontal line for the group to which the data point belongs) and the overall mean of the data (the solid horizontal line).

$$MSS = \sum n_j (\hat{y}_j - \hat{y}_{Grand})^2$$

The final sum of squares is the residual sum of squares (SSE, also called SSR on the next figure), which tells us how much of the variation cannot be explained by the model. The simplest way to calculate SSE is to subtract MSS from TSS. It is calculated by looking at the difference between the score obtained by a person and the mean of the group to which the person belongs. These distances between each data point and the group mean are squared and then added together to give the SSE.

$$SSE = \sum (y_{ij} - \hat{y}_i)^2$$



SS_T uses the differences between the observed data and the mean value of Y

SS_R uses the differences between the observed data and the model (group means)

SS_M uses the differences between the mean value of Y and the model (group means)

The original figure can be found here.

6.1.6 Mean Squares (MS)

The sum of squares depends on the number of groups and the number of cases per group. MS are variances calculated by dividing the SS by the corresponding number of degrees of freedom.

- Model Mean Sum of Squares: $MS_M = \frac{MSS}{df_M}$, $df_M = k - 1$.

- Mean sum of squares of the residuals: $MS_R = \frac{RSS}{df_R}$, $df_R = k(n_j - 1) = n - k$.

The test variable F_{Ratio} is calculated as follows: $F = \frac{MS_M}{MS_R}$.

Note that n is the number of observations by group and k is the number of groups.

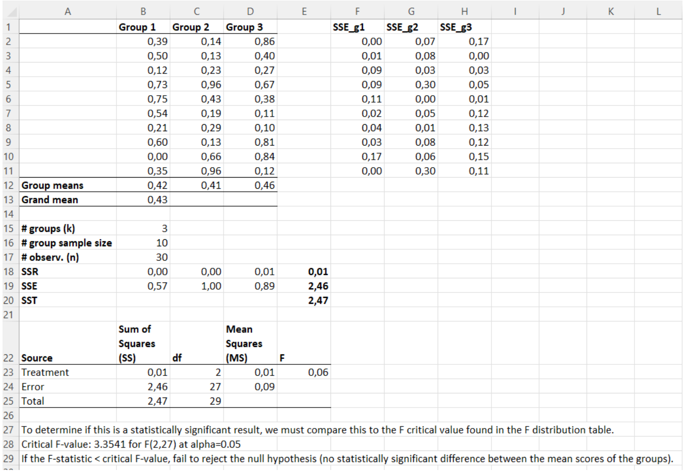
Calculate the F-statistics

Let's assume that we have three groups to compare (A, B, and C). The following steps must be undertaken in one-way ANOVA:

- Calculate the common variance, often known as residual variance or variance within samples: SS_w
- Calculate the difference in sample means as follows:
  - Calculate the average of each group
  - Calculate the difference in sample means: SS_b

The F-statistic is calculated based on the ratio of SS_b/SS_w. Note that a lower ratio suggests that the means of the samples being compared are not significantly different. A greater ratio, on the other hand, indicates that the differences in group means are significant.

The figure below shows the calculation of the F-statistic for one-way ANOVA offer. You can click on the figure to download the excel file.



	A	B	C	D	E	F	G	H	I	J	K	L
1		Group 1	Group 2	Group 3		SSE_g1	SSE_g2	SSE_g3				
2		0,39	0,14	0,86		0,00	0,07	0,17				
3		0,50	0,13	0,40		0,01	0,08	0,00				
4		0,12	0,23	0,27		0,09	0,03	0,03				
5		0,73	0,96	0,67		0,09	0,30	0,05				
6		0,75	0,43	0,38		0,11	0,00	0,01				
7		0,54	0,19	0,11		0,02	0,05	0,12				
8		0,21	0,29	0,10		0,04	0,01	0,13				
9		0,60	0,13	0,81		0,03	0,08	0,12				
10		0,00	0,66	0,84		0,17	0,06	0,15				
11		0,35	0,96	0,12		0,00	0,30	0,11				
12	Group means	0,42	0,41	0,46								
13	Grand mean	0,43										
14												
15	# groups (k)	3										
16	# group sample size	10										
17	# observ. (n)	30										
18	SSR	0,00	0,00	0,01	0,01							
19	SSE	0,57	1,00	0,89	2,46							
20	SST				2,47							
21												
22	Source	Sum of Squares (SS)	df	Mean Squares (MS)	F							
23	Treatment	0,01	2	0,01	0,06							
24	Error	2,46	27	0,09								
25	Total	2,47	29									
26												
27	To determine if this is a statistically significant result, we must compare this to the F critical value found in the F distribution table.											
28	Critical F-value: 3.3541 for F(2,27) at alpha=0.05											
29	If the F-statistic < critical F-value, fail to reject the null hypothesis (no statistically significant difference between the mean scores of the groups).											

Figure 6.1: F-statistic calculation

6.1.7 Effect of two or more factors

To compare the influence of several different factors, it is necessary to test the interaction hypotheses: influence of a factor depending on one or more other factors (A and B).

- MSS_A: Calculation with the groups of the factor A (we pretend that B does not exist)
- MSS_B: Calculation with the groups of the factor B (we pretend that A does not exist)
- MSS_{AB}: $MSS - MSS_A - MSS_B$ indicates what the model can explain beyond factors A and B.

A separate F is used for each factor and interaction term and $MS = \frac{SS}{df}$ as in single ANOVA.

- Factor A: $F = \frac{MS_{MA}}{MS_R}$
- Factor B: $F = \frac{MS_{MB}}{MS_R}$
- Interaction AxB: $F = \frac{MS_{MAB}}{MS_R}$

6.1.8 Effect size

R^2 gives the effect size of the entire model: proportion of variance (TSS) that can be explained by the model (MSS).

$$R^2 = \frac{MSS_{overall}}{TSS}$$

Eta^2 is a measure of effect size that is commonly used in ANOVA models. It measures the proportion of variance associated with each main effect and interaction effect in an ANOVA model.

$$Eta^2 = \frac{SS_{effect}}{SS_{total}}$$

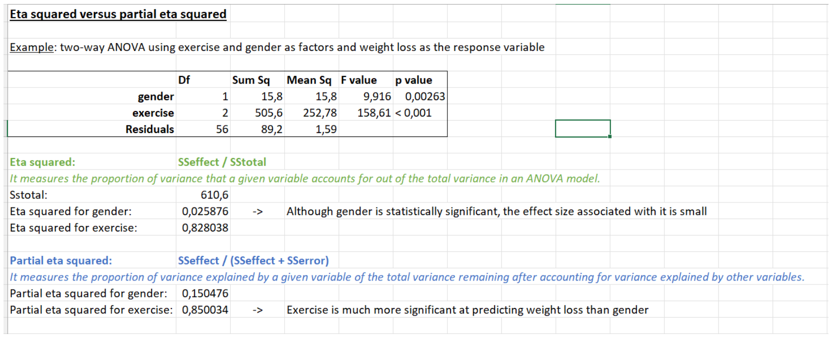
where SS_{effect} is the sum of squares of an effect for one variable and SS_{total} is the total sum of squares in the ANOVA model.

The value for Eta^2 ranges from 0 to 1, where values closer to 1 indicate a higher proportion of variance that can be explained by a given variable in the model.

Partial Eta^2 (or $Eta^2_{partial}$) is the respective effect size of the individual factors: reflects the proportion of the variance that can be explained by the respective factors or interaction terms.

- Eta^2 for factor A: $\frac{SS_A}{TSS}$
- Eta^2 for factor B: $\frac{SS_B}{TSS}$
- Eta^2 for interaction AxB: $\frac{SS_{AB}}{TSS}$

The value for $Eta^2_{partial}$ also ranges from 0 to 1, where values closer to 1 indicate a higher proportion of variance that can be explained by a given variable in the model after accounting for variance explained by other variables in the model.



Eta squared versus partial eta squared

Example: two-way ANOVA using exercise and gender as factors and weight loss as the response variable

	Df	Sum Sq	Mean Sq	F value	p value
gender	1	15,8	15,8	9,916	0,00263
exercise	2	505,6	252,78	158,61	< 0,001
Residuals	56	89,2	1,59		

Eta squared: **SSeffect / SStotal**

It measures the proportion of variance that a given variable accounts for out of the total variance in an ANOVA model.

Sstotal:	610,6		
Eta squared for gender:	0,025876	->	Although gender is statistically significant, the effect size associated with it is small
Eta squared for exercise:	0,828038		

Partial eta squared: **SSeffect / (SSeffect + SSerror)**

It measures the proportion of variance explained by a given variable of the total variance remaining after accounting for variance explained by other variables.

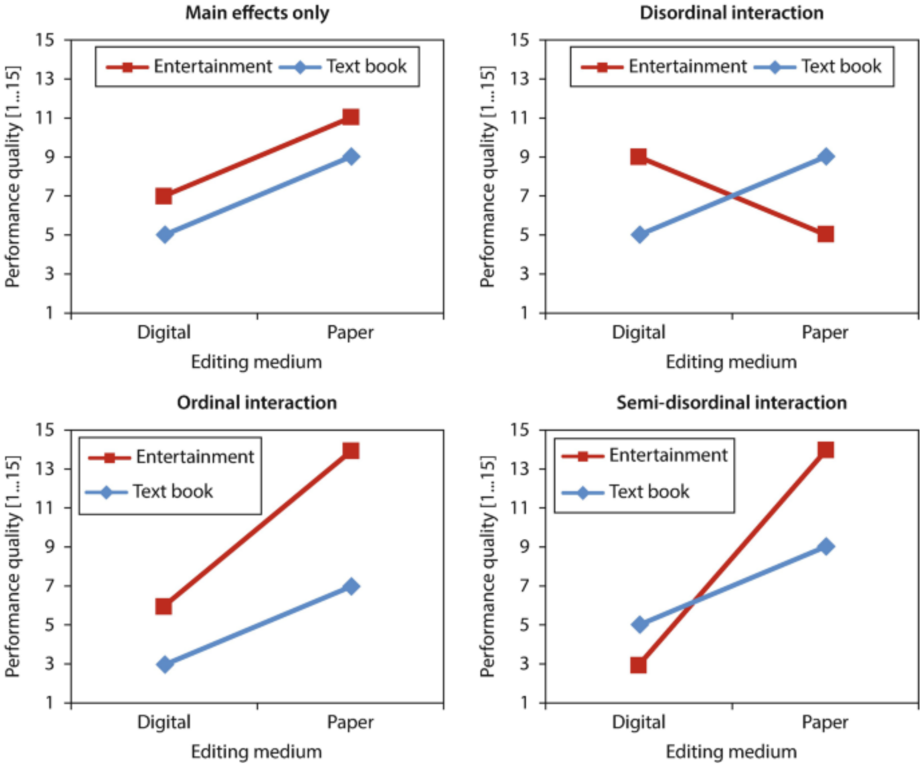
Partial eta squared for gender:	0,150476		
Partial eta squared for exercise:	0,850034	->	Exercise is much more significant at predicting weight loss than gender

Figure 6.2: Eta squared versus partial eta squared

6.1.9 Types of interaction

There exist different types of interaction:

- A null interaction means there is no interaction: effects of factor A on the dependent variable are therefore the same at all stages of the factor B (parallel lines)
- An ordinal interaction suggests that the interaction of two factors A and B is significant: effects of A are not the same at all factor levels of B (lines do not run parallel but do not cross)
- A disordinal interaction suggests that the interaction of A and B is significant, but a disordinal interaction is shown by crossing lines (any main effect that may be present must not be interpreted)
- A semi-disordinal interaction means that the interaction of A and B is significant, and it can be shown by crossing lines or by lines that do not cross (only one of the main effects that may be present may be interpreted)



The original figure can be found here.

6.1.10 Which groups differ significantly?

The F-test only tells us that there is a significant difference exists, but it does not tell us which groups differ. Therefore, we need to conduct post-hoc tests or contrasts within a factor. Between the factors test pair comparisons can be conducted to check for significant differences.

6.1.11 Statistical assumptions

There are two assumptions for the ANOVA:

- Normal distribution within the groups: dependent variable must be normally distributed in each group (but not necessarily in the overall sample)
- Homogeneity of variance between groups: variances in the dependent variable must be the same in each group

The normal distribution assumption can be assessed visually. Furthermore, the Kolmogorov-Smirnov test (KS test) and Shapiro-Wilk test (more suitable from small samples < 50) can assess the distribution of the dependent variable against normal distribution, with deviations that are too large becoming significant (not desirable!). For serious violations of the normal distribution assumption:

- the dependent variable can be normalized by transformations (e.g. logarithm, squaring, etc.)
- problematic cases can be excluded
- it is also possible to rely on a non-parametric analysis of variance (e.g. Kruskall-Wallis test)

Levene's test can be used to test the null hypothesis that variances in the groups are equal (so rejecting this null hypothesis is undesirable). The analysis of variance (F-test) is quite robust against violations of the homogeneity of variance if the group sizes are (approximately) identical. However, post-hoc tests are not robust! In the case of serious violations of homogeneity of variance, there are two possibilities:

- Correction of the F-ratio with Welch test (best choice)
- Complete renunciation of F-test and calculation with Kruskall-Wallis test (non-parametric tests)

However, both are only possible with a one-factor analysis of variance! Therefore, as an approximation, it is possible to calculate the F-ratio for each factor individually using the Welch test and the Kruskall Wallis test. If the result is the same (sign. or not sign.), the result of the ANOVA can be trusted.

6.1.12 In a nutshell

ANOVA is based on the principle of scattering into explained and unexplained variance (F-test). In multi-factor ANOVA, variance is explained both by the individual factors and by their interaction (F-value for each factor and the interaction term).

Eta^2 provides information about the effect size of the factors and interaction terms (compared to R^2, which refers to the entire model).

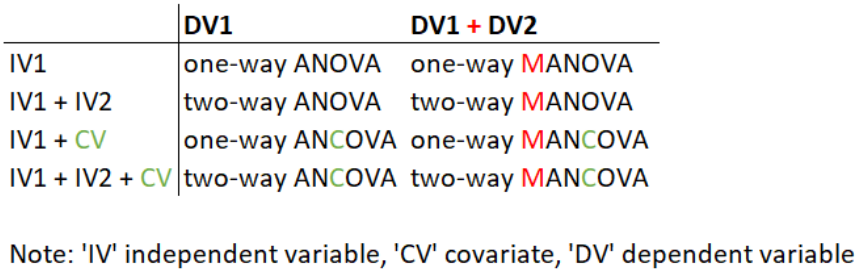
There are different interaction types: ordinal, disordinal and semi-disordinal. The main effect of a factor can only be interpreted if it is not affected by a disordinal interaction.

Post-hoc tests show which factor levels differ significantly from each other. The Simple Effects Analysis shows for which factor levels of a factor A there is a significant difference due to factor B (interaction).

The ANOVA is based on two assumptions: normal distribution of the dependent variables and homogeneity of variances in the groups. It is generally quite robust to violations of these assumptions. There are tests to test these assumptions as well as procedures to report the results when they are violated.

6.2 Beyond one-way ANOVA

Below we explain the differences between the statistical methods ANOVA, ANCOVA, MANOVA, and MANCOVA.



	DV1	DV1 + DV2
IV1	one-way ANOVA	one-way MANOVA
IV1 + IV2	two-way ANOVA	two-way MANOVA
IV1 + CV	one-way ANCOVA	one-way MANCOVA
IV1 + IV2 + CV	two-way ANCOVA	two-way MANCOVA

Note: 'IV' independent variable, 'CV' covariate, 'DV' dependent variable

6.2.1 ANCOVA

An ANCOVA (Analysis of Covariance) is also used to determine whether or not there is a statistically significant difference between the means of three or more independent groups. Unlike an ANOVA, however, an ANCOVA includes one or more covariates (CV), which can help us better understand how a factor impacts a response variable after accounting for some covariate(s).

Covariates are control variables whose influence on the dependent variable should be controlled. They are often variables that cannot be manipulated at all for theoretical or practical research reasons (e.g. age, gender, personality variables). They are included in the modeling as metric variables. Their influence on the dependent variable is calculated before the influence of the manipulated independent variables is calculated.

Covariates help us with reducing unexplained variance and within-group variance (SSE). The effect of the factors on the dependent variable can thus be determined more accurately (MSS). This improves the explanatory power of the overall model (R^2). The explanatory power of the factors (η^2) also improves.

The main advantage is that influences of other variables on the dependent variable can be “controlled” without being manipulated in the experiment. However, there are also disadvantages. For instance, the strict cause-and-effect logic of an experimental design does not apply to covariates. Futhermore, covariates must have a measurement (and they must be scale). Moreover, one can never include all relevant covariates. Note that we loses one degree of freedom of the residuals per covariate (SSE).

- Example: political affiliation impacts on media coverage controlling for the number of social media followers (covariate).

6.2.2 ANCOVA assumptions

ANCOVA makes several assumptions about the data, such as:

- Linearity between the covariate and the response variable at each level of the grouping variable.
  - This can be checked by creating a clustered scatterplot of the covariate and the response variable.
- Homogeneity of regression slopes: the slopes of the regression lines, formed by the covariate and the response variable, should be the same for each group.
  - This can be verified visually as there should be no interaction between the outcome and the covariate (regression lines drawn by groups should be parallel).
- The response variable should be approximately normally distributed.
  - This can be verified using the Shapiro-Wilk normality test on the model residuals.
- Homoscedasticity or homogeneity of variance of residuals for all groups.
  - Residuals are assumed to have constant variance (homoscedasticity)
- There should be no significant outliers within groups.

Nota bene: If several covariates are used, we also need low correlations between the covariates (avoidance of multicollinearity). In the case of multicollinearity, this would lead to a reduction of the degrees of freedom without a simultaneous reduction of the SS.

6.2.3 ANCOVA Post-hoc tests

"Normal" post-hoc tests cannot be called up for covariance analyses. However, there are three alternatives:

- Post hoc tests on "estimated marginal means"
- Contrasts
- Simple effects analysis

6.2.4 MANOVA

A MANOVA (Multivariate Analysis of Variance) is identical to an ANOVA, except it uses two or more response variables. Similar to the ANOVA, it can also be one-way or two-way.

Example of one-way MANOVA: political affiliation impacts on media coverage and also on budget.

Example of two-way MANOVA: political affiliation and gender impact on media coverage and also on budget.

6.2.5 MANCOVA

A MANCOVA (Multivariate Analysis of Covariance) is identical to MANOVA, except it also includes one or more covariates. Similar to a MANOVA, a MANCOVA can also be one-way or two-way.

Example of one-way MANCOVA: political affiliation impacts on media coverage and also on budget while controlling for the number of social media followers (covariate).

Example of two-way MANCOVA: political affiliation and gender impact on media coverage and also on budget while controlling for the number of social media followers (covariate).

6.3 How it works in R?

See the lecture slides on one-way (M)AN(C)OVA:

You can also download the PDF of the slides here: PDF

6.4 Quiz

True

False

Statement

A p-value of .03 means that in only 3 in 100 times would a result as large as the one observed in a sample of data be observed if the null hypothesis were actually true for the population.

A 2-way ANOVA compares the levels of two or more factors on a continuous variable.

A covariate is a continuous independent variable in an ANCOVA testing and a factor is a categorical independent variable in an ANCOVA testing.

In a 2-way MANCOVA, the 2 (or more) dependent variables should be unrelated.

View Results

My results will appear here

6.5 Example from the literature

The following article relies on ANOVA as a method of analysis:

Otto, L. P., Lecheler, S., & Schuck, A. R. (2020). Is context the key? The (non-) differential effects of mediated incivility in three European countries. Political Communication, 37(1), 88-107. Available here.

Please reflect on the following questions:

- What is the research question of the study?
- What are the research hypotheses?
- Is ANOVA an appropriate method of analysis to answer the research question?
- What are the main findings of the ANOVA analysis?

6.6 Time to practice on your own

You can download the PDF of the exercises here:

6.6.1 Exercice 1: Gapminder

6.6.1.1 Bivariate recap: mean comparisons and correlations

We will be using the Gapminder data to in the context of t-test, ANOVA and Chi-squared test. The goal is to train your understanding of the underlying principles of hypothesis testing and p-values (when to reject the null hypothesis or accept the alternative hypothesis when making inference about the population from sample data).

Let's start by loading the Gapminder data and look at the mean life expectancy by countries:

```
data = gapminder::gapminder
aggregate(lifeExp ~ continent, data, mean)
##   continent  lifeExp
## 1    Africa 48.86533
## 2  Americas 64.65874
## 3      Asia 60.06490
## 4    Europe 71.90369
## 5   Oceania 74.32621
```

Let's presume that the mean life expectancy in Africa is 50 years old. We would like to test whether the observed mean from the sample (here: 48.86) is statistically different from the "true/presumed" mean:

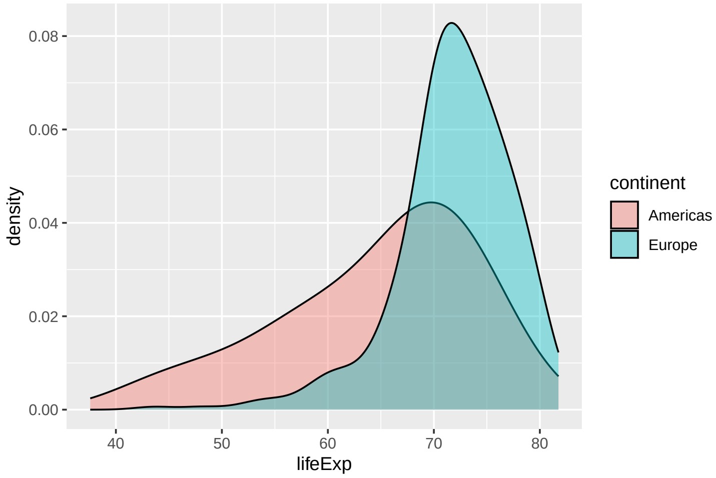
```
pop_mean=50 # according to H0
smp_mean=data$lifeExp[data$continent=="Africa"]
t.test(smp_mean, mu=pop_mean, alternative = "two.sided", conf.level = 0.95)
##
```

```
##  One Sample t-test
##
## data:  smp_mean
## t = -3.0976, df = 623, p-value = 0.002038
## alternative hypothesis: true mean is not equal to 50
## 95 percent confidence interval:
##  48.14599 49.58467
## sample estimates:
## mean of x
##  48.86533
```

> ℹ Interpretation
>
> Here, we are testing a single sample against a known value. The results show that the sample mean (e.g., Africa) significantly differs from the presumed value (e.g., 50). The t-value is -3.09 and the associated p-value is 0.002. The confidence interval is given and does not contain the presumed value (e.g., 50).

Now, we would like to test the null hypothesis stating that there is no difference in the mean life expectancy between Europe and America:

```
sub=data[data$continent=="Europe" | data$continent=="Americas",]
# visualize the distributions
ggplot2::ggplot(data=sub, ggplot2::aes(
  x=lifeExp, group=continent, fill=continent)) +
  ggplot2::geom_density(adjust=1.5, alpha=.4)
# conduct t.test
t.test(lifeExp ~ continent, data=sub, mu=0,
       alt="two.sided", conf=0.95, var.eq=F, paired=F)
##
##  Welch Two Sample t-test
##
## data:  lifeExp by continent
## t = -11.861, df = 460.72, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Americas and group Europe
## 95 percent confidence interval:
##  -8.445287 -6.044612
## sample estimates:
## mean in group Americas   mean in group Europe
##               64.65874               71.90369
```



Interpretation

Here, we are comparing the mean life expectancy for two continents (e.g., Americas and Europe). The results show that there is a statistically significant difference in mean life expectancy between the continents (t-value = -11.86 and p-value < 0.001).

Now, we will add additional variables to the Gapminder data to test the relationship between levels of GDP and life expectancy. Start by creating two dichotomous variables (high vs low GDP and below vs above average life expectancy):

```
library(plyr)
library(dplyr)
sel <- data %>%
  dplyr::filter(continent == "Europe") %>%
  plyr::mutate(gdp_factor = if_else(gdpPercap > mean(gdpPercap),
                              "high gdp", "low gdp"),
         below_avg_life_exp = if_else(lifeExp < mean(lifeExp),
         "low life expectancy", "high life expectancy"))
sel$gdp_factor <- as.factor(sel$gdp_factor)
sel$below_avg_life_exp <- as.factor(sel$below_avg_life_exp)
table(sel$gdp_factor, sel$below_avg_life_exp)
##
##            high life expectancy low life expectancy
##   high gdp                  134                  14
##   low gdp                    55                 157
```

Now, we want to look at the relationship between the two newly created dichotomous variables (high vs low GDP and below vs above average life expectancy).

We can run a chi square test with the chisq.test() function:

```
chisq.test(table(sel$gdp_factor, sel$below_avg_life_exp))
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(sel$gdp_factor, sel$below_avg_life_exp)
## X-squared = 143.26, df = 1, p-value < 2.2e-16
```

Interpretation

The chi square test shows a significant relationship between high vs low GDP and below vs above average life expectancy (chi square = 143.26 and p-value < 0.001).

Finally, we can look at the relationship between original variables for GDP and life expectancy using the cor.test() function:

```
cor.test(data$lifeExp,data$gdpPercap)
##
##  Pearson's product-moment correlation
##
## data:  data$lifeExp and data$gdpPercap
## t = 29.658, df = 1702, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5515065 0.6141690
## sample estimates:
##       cor
## 0.5837062
```

Interpretation

The correlation test shows a significant relationship between GDP and life expectancy (pearson = 0.58, t-value = 29.65 and p-value < 0.001). Note that most likely we would actually use a different correlation test, as the data on life expectancy are non-normal.

Graphical representations

Optionally, we can offer a static visualization of the data for the year 2007:

```
data$year <- as.numeric(data$year)
data$country <- as.character(data$country)
library(plotly)
sub = data %>%
  filter(year=="2007") %>%
  dplyr::select(-year)
p <- sub %>%
      arrange(desc(pop)) %>%
      mutate(country = factor(country, country)) %>%
      ggplot(aes(x=gdpPercap, y=lifeExp, size=pop, color=continent)) +
        geom_point(alpha=0.5) +
        scale_size(range = c(.1, 24), name=" ") +
  ggtitle("Relationship between gdpPerCap and lifeExp in 2007")
# ggplotly(p)
htmlwidgets::saveWidget(
                widget = ggplotly(p), #the plotly object
                file = "images/Plotly2007_GDPbyLifeExp.html",
                selfcontained = TRUE #creates a single html file
                )
```

Optionally, a dynamic visualization is also possible using the following code:

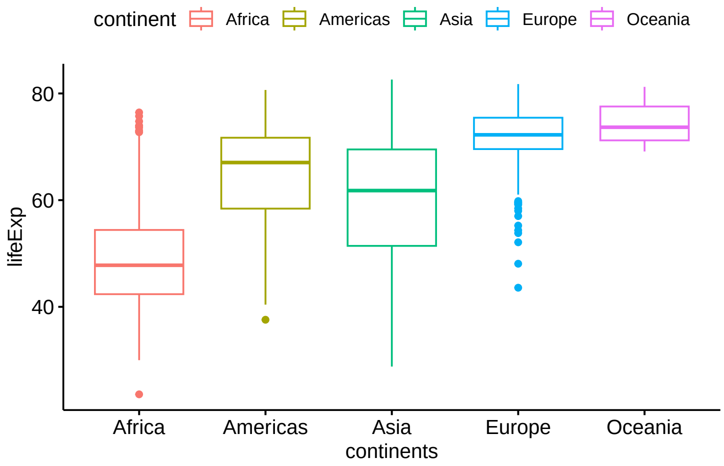
```
colors <- c('#C61951','#A4C6B0','#4AC6B7', '#1972A4', '#965F8A')
r3 <- plot_ly(
  data,
  x = ~gdpPercap,
  y = ~lifeExp,
  frame=~year,
  color = ~continent,
  type = "scatter",
  mode="markers",
  colors=~colors,
  size=~pop,
  marker = list(symbol = 'circle', sizemode = 'diameter',
                      line = list(width = 2, color = '#FFFFFF'), opacity=0.4)) %>%
  layout(
        title="Relationship between gdpPerCap and lifeExp over time",
        xaxis = list(title = 'gdpPercap',
                      # gridcolor = 'rgb(255, 255, 255)',
                      # range=c(10,50),
                      zerolinewidth = 1,
                      ticklen = 5,
                      gridwidth = 2),
         yaxis = list(title = 'lifeExp',
                      # gridcolor = 'rgb(255, 255, 255)',
                      # range=c(40,80),
                      zerolinewidth = 1,
                      ticklen = 5,
                      gridwith = 2),
        paper_bgcolor = 'rgb(243, 243, 243)',
        plot_bgcolor = 'rgb(243, 243, 243)'
  )%>%
  animation_opts(
    2000, redraw = FALSE
  ) %>%
  animation_slider(
    currentvalue = list(prefix = "YEAR ", font = list(color="black"))
  )
# r3
htmlwidgets::saveWidget(
                widget = r3, #the plotly object
                file = "images/PlotlyDynamic_GDPbyLifeExp.html",
                selfcontained = TRUE #creates a single html file
                )
```

6.6.1.2 ANOVA

Let's see if the different continents really do have different life expectancy. In order to do this, we will perform an ANOVA using our continents as different groups, then test for differences between the groups.

Therefore, we would like to test the null hypothesis stating that there is no difference in the mean life expectancy between more continents (Europe, America, Asia, Africa, and Oceania). Let's start by visualizing the differences using boxplots:

```
ggpubr::ggboxplot(data, x = "continent",
          y = "lifeExp",
          color = "continent",
          ylab = "lifeExp", xlab = "continents")
```



Now, we want to formally compute ANOVA test:

```
res.aov <- aov(lifeExp ~ continent, data = data)
summary(res.aov)
##               Df Sum Sq Mean Sq F value Pr(>F)    
## continent      4 139343   34836   408.7 <2e-16 ***
## Residuals   1699 144805      85                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

> ℹ Interpretation
>
> ANOVA tells us that continent does have a significant effect on life expectancy (F-value = 408.7 and p-value < 0.001). What we want now is to be able to tell what those differences are.

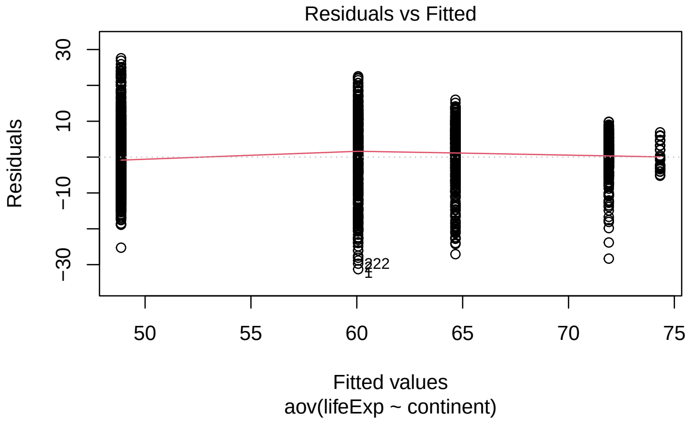
In ANOVA test, a significant p-value indicates that some of the group means are different, but we do not know which pairs of groups are different. To do so, we need to perform multiple pairwise-comparison:

```
TukeyHSD(res.aov)
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lifeExp ~ continent, data = data)
##
## $continent
##                        diff        lwr       upr     p adj
## Americas-Africa   15.793407 14.022263 17.564550 0.0000000
## Asia-Africa       11.199573  9.579887 12.819259 0.0000000
## Europe-Africa     23.038356 21.369862 24.706850 0.0000000
## Oceania-Africa    25.460878 20.216908 30.704848 0.0000000
## Asia-Americas     -4.593833 -6.523432 -2.664235 0.0000000
## Europe-Americas    7.244949  5.274203  9.215696 0.0000000
## Oceania-Americas   9.667472  4.319650 15.015293 0.0000086
## Europe-Asia       11.838783 10.002952 13.674614 0.0000000
## Oceania-Asia      14.261305  8.961718 19.560892 0.0000000
## Oceania-Europe     2.422522 -2.892185  7.737230 0.7250559
```

> Interpretation
>
> The output shows that Africa has a significantly lower life expectancy than every other continent, followed by Asia, and then the Americas. Oceania and Europe, meanwhile, have about equal life expectancy.

Remember that the ANOVA test assumes that, the data are normally distributed and the variance across groups are homogeneous. The residuals versus fits plot can be used to check the homogeneity of variances:

```
plot(res.aov, 1)
```



Interpretation

The plot shows no evident relationships between residuals and fitted values (the mean of each groups), which is good. So, we can assume the homogeneity of variances.

It is also possible to use Levene' test (or Bartlett's test) to check the homogeneity of variances (if the p-value is less than the significance level of 0.05, it means that there is evidence that the variance across groups is statistically significantly different):

```
car::leveneTest(lifeExp ~ continent, data = data)
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    4  51.568 < 2.2e-16 ***
##       1699
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation

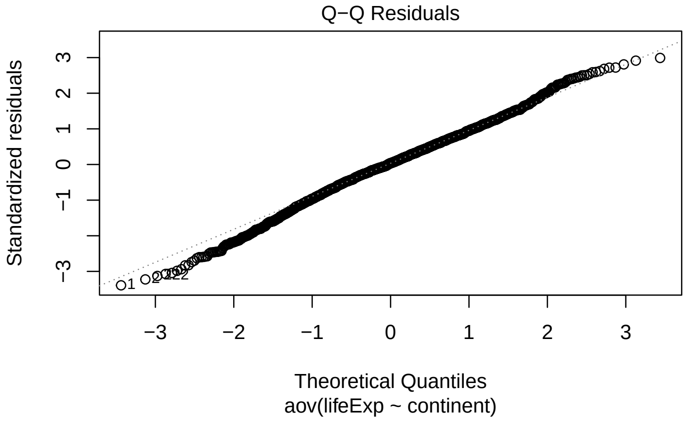
From the output of Levene's test, we can see that the p-value is less than 0.05. This means that there is evidence to suggest that the variance across groups is statistically significantly different. Therefore, we cannot assume the homogeneity of variances.
The classical one-way ANOVA test requires an assumption of equal variances for all groups. An alternative procedure (Welch one-way test), that does not require that assumption have been implemented in the function

oneway.test():

```
oneway.test(lifeExp ~ continent, data = data)
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  lifeExp and continent
## F = 664.9, num df = 4.00, denom df = 178.95, p-value < 2.2e-16
pairwise.t.test(data$lifeExp, data$continent,
                p.adjust.method = "BH", pool.sd = FALSE)
##
##  Pairwise comparisons using t tests with non-pooled SD
##
## data:  data$lifeExp and data$continent
##
##          Africa  Americas Asia    Europe
## Americas < 2e-16 -        -       -
## Asia     < 2e-16 1.8e-08  -       -
## Europe   < 2e-16 < 2e-16  < 2e-16 -
## Oceania  < 2e-16 9.4e-14  < 2e-16 0.0064
##
## P value adjustment method: BH
```

To test for the normality of residuals, we can plot the quantiles of the residuals against the quantiles of the normal distribution:

```
plot(res.aov, 2)
```



Interpretation

As all the points fall approximately along this reference line, we can assume normality.

The Shapiro-Wilk test on the ANOVA residuals can be used to assess whether normality is violated:

```
aov_residuals <- residuals(object = res.aov)
shapiro.test(aov_residuals)
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.9954, p-value = 0.000044
```

Interpretation

The Shapiro-Wilk test on the ANOVA residuals does not support the normality assumption.
A non-parametric alternative to one-way ANOVA is Kruskal-Wallis rank sum test, which can be used when ANOVA assumptions are not met:

```
kruskal.test(lifeExp ~ continent, data = data)
##
##  Kruskal-Wallis rank sum test
##
## data:  lifeExp by continent
## Kruskal-Wallis chi-squared = 843.38, df = 4, p-value < 2.2e-16
```

6.6.2 Exercice 2: ANCOVA

In this exercise, you will learn how to:

- Calculate and interpret one-way and two-way ANCOVA in R
- Check ANCOVA assumptions
- Perform post-hoc testing (multiple pairwise comparisons between groups to identify groups that are different)

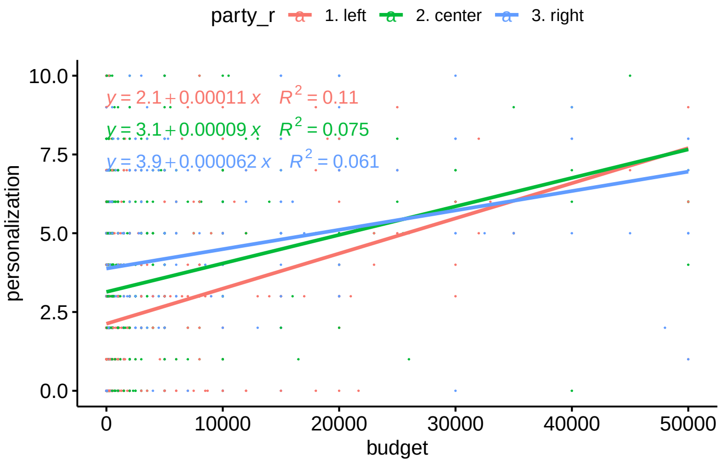
For instance, we are interested in measuring the personalized style of campaigning (Y) explained by the political affiliation (X) and the available budget (CV). To do so, we will rely on the data covering the Swiss part of the Comparative Candidate Survey. We will be using the Selects 2019 Candidate Survey.

Let's start by selecting the relevant variables and by recoding them (also select politicians with a budget equal or below CHF 50,000.-):

```r
library(foreign)
db <- read.spss(file=paste0(getwd(),
"/data/1186_Selects2019_CandidateSurvey_Data_v1.1.0.sav"),
                use.value.labels = F,
                to.data.frame = T)
sel <- db |>
  dplyr::select(B12,T9a,B6) |>
  stats::na.omit() |>
  dplyr::rename("budget"="B12",
                "party"="T9a",
                "personalization"="B6")
# budget without outliers
sel$budget <- as.numeric(as.character(sel$budget))
sel <- sel[sel$budget<=50000,]
# party recoded into left-center-right
sel$party_r <- NA
sel$party_r <- ifelse(sel$party==3 |
                        sel$party==5, "1. left", as.character(sel$party_r))
sel$party_r <- ifelse(sel$party==2 |
                        sel$party==6, "2. center", as.character(sel$party_r))
sel$party_r <- ifelse(sel$party==1 |
                        sel$party==4, "3. right", as.character(sel$party_r))
sel <- sel[!is.na(sel$party_r),]
# personalization (invert scale)
sel$personalization <- as.numeric(as.character(sel$personalization))
sel$personalization <- (sel$personalization-10)*(-1)
```

Now, we can create a scatterplot between the covariate (budget) and the response variable (personalization), and add regression lines showing the equations and R^2 by groups (party affiliations):

```r
p = ggpubr::ggscatter(
  sel, x = "budget", y = "personalization",
  color = "party_r", add = "reg.line", size=0
  )+
  ggpubr::stat_regline_equation(
    ggplot2::aes(label =  paste(..eq.label.., ..rr.label.., sep = "~~~~"),
                 color = party_r)
    )
plot(p)
```



Interpretation

The figure displays a linear relationship between budget and personalization for each training party group.
Furthermore, we can see an interaction between the covariate and the independent variable (the party trends start to cross from a budget of ~25'000 Swiss francs). This violates the assumption of homogeneity of the regression slopes (there should be no significant interaction between the covariate and the grouping variable).

We can now verify whether there is a significant interaction between the covariate and the grouping variable:

```
sel |> rstatix::anova_test(personalization ~ party_r*budget)
## ANOVA Table (type II tests)
##
##           Effect DFn  DFd       F        p p<.05   ges
## 1        party_r   2 1382  39.858 1.48e-17     * 0.055
## 2         budget   1 1382 120.642 5.75e-27     * 0.080
## 3 party_r:budget   2 1382   3.434 3.30e-02     * 0.005
```

Interpretation

The output shows that there is no homogeneity of regression slopes as the interaction term is statistically significant (F(2, 1382) = 3.43, p = 0.005). Nota bene: this does not mean that the model is "wrong" per se, as the fact that the interaction is significant tells us something interesting about the relationship between the variables.

Nota bene: The order of the variables is important in the calculation of the ANCOVA. You want to remove the effect of the covariate first and before entering your primary variable or interest:

```
res.aov <- sel |> rstatix::anova_test(personalization ~ budget + party_r)
rstatix::get_anova_table(res.aov)
## ANOVA Table (type II tests)
##
##    Effect DFn  DFd       F        p p<.05   ges
## 1  budget   1 1384 120.220 6.97e-27     * 0.080
## 2 party_r   2 1384  39.718 1.69e-17     * 0.054
```

Pairwise comparisons can be made to identify groups that are statistically different. Bonferroni's multiple test correction is applied using the emmeans_test() function from the rstatix package:

```
compa = sel |>
  rstatix::emmeans_test(
    personalization ~ party_r, covariate = budget,
    p.adjust.method = "bonferroni"
    )
compa[,c(3,4,6,7)]
## # A tibble: 3 x 4
##   group1    group2    statistic        p
##   <chr>     <chr>         <dbl>    <dbl>
## 1 1. left   2. center     -6.04 1.93e- 9
## 2 1. left   3. right      -8.54 3.36e-17
## 3 2. center 3. right      -3.26 1.16e- 3
```

To estimate the normality of the residuals, you first need to calculate the model using lm(). In R, you can easily augment your data to add predictors and residuals using the function augment() from the broom package and then use the Shapiro-Wilk test.

```
model <- lm(personalization ~ budget + party_r, data = sel)
model.metrics <- broom::augment(model)
rstatix::shapiro_test(model.metrics$.resid)
## # A tibble: 1 x 3
##   variable              statistic  p.value
##   <chr>                     <dbl>    <dbl>
## 1 model.metrics$.resid      0.973 1.35e-15
```

Interpretation

The Shapiro Wilk test is significant (p < 0.001), so we cannot assume normality of residuals.

ANCOVA assumes that the variance of the residuals is equal for all groups. This can be checked using Levene's test:

```
model.metrics |> rstatix::levene_test(.resid ~ party_r)
## # A tibble: 1 x 4
##     df1   df2 statistic       p
##   <int> <int>     <dbl>   <dbl>
## 1     2  1385      6.70 0.00127
```

Interpretation

The Levene's test is significant (p < 0.01), so we cannot assume homogeneity of the residual variances for all groups.

Furthermore, we need to check for outliers. These can be identified by looking at the normalized residuals (or studentized residuals), which is the residual divided by its estimated standard error. The normalized residuals can be interpreted as the number of standard errors outside the regression line.

```
outliers = model.metrics |>
  dplyr::filter(abs(.std.resid) > 3) |>
  as.data.frame()
outliers[,c(1:4,5:6,9)]
##   .rownames personalization budget party_r  .fitted   .resid     .cooksd
## 1      1696              10      0 1. left 2.229553 7.770447 0.005059324
## 2      1934              10    200 1. left 2.246654 7.753346 0.005000313
```

Chapter 7

Repeated Measurements

7.1 Repeated measurements analyis

The repeated measures ANOVA is used for analyzing data where same subjects are measured more than once on the same outcome variable under different time points or conditions. This test is also referred to as a within-subjects ANOVA (or ANOVA with repeated measures).

Repeated measures ANOVA is the equivalent of the one-way ANOVA, but for related, not independent groups, and is the extension of the dependent t-test.

In repeated measures ANOVA, the independent variable (also referred as the within-subject factor) has categories (called levels or groups). We can analyse data using a repeated measures ANOVA for two types of study design:

- investigating changes in mean scores over three or more time points (e.g., pre-, midway and post-intervention)
- investigating differences in mean scores under three or more different conditions

7.1.1 Within- and between-subject variables

For one withing-subject factor, we can consider the example where time is used to evaluate whether there is any difference in social media reliance across the several times of data.

We can also consider adding age cohort as a between-subject factor in order to test effect of age cohort on social media reliance, as well as the interaction effect between time and age cohort.

7.1.2 One-way and two-way repeated measurements analyis

An ANOVA with repeated measures is used to compare three or more group means where the participants are the same in each group. This usually occurs in situations when participants are measured multiple times to see changes to an intervention or when participants are subjected to more than one condition/trial and the response to each of these conditions wants to be compared.

One-way repeated measures ANOVA is an extension of the paired-samples t-test for comparing the means of three or more levels of a within-subjects variable.

Two-way repeated measures ANOVA is used to evaluate simultaneously the effect of two within-subject factors on a continuous outcome variable.

7.2 Assumptions

The repeated measures ANOVA makes the following assumptions:

- No significant outliers
- Normality of the dependent variable (at each time point)
  - We can use histograms and normality tests
- Variance of the differences between groups should be equal (sphericity assumption)
  - We can use Mauchly's test of Sphericity (if p>0.05, sphericity can be assumed).

Note that, if the above assumptions are not met there are a non-parametric alternative (Friedman test) to the one-way repeated measures ANOVA. However, there are no non-parametric alternatives to the two-way (and the three-way) repeated measures ANOVA. Thus, in the situation where the assumptions are not met, you could consider running the two-way repeated measures ANOVA on the transformed and non-transformed data to see if there are any meaningful differences.

7.3 Mauchly's test of sphericity

Mauchly's test of sphericity tests the null hypothesis (H0) that all variances can be considered homogeneous. A significant result leads to the rejection of H0 since at least two variances are unequal. Violations of sphericity lead to a biased F-test and to biased post-hoc test results!

There are two different ways of calculating sphericity violation:

- Greenhouser & Geisser (1959): Rather conservative (the injuries are overestimated, especially in the case of minor injuries to the sphericity)
- Huynh & Feldt (1976): Rather liberal (thus rather underestimated)

Stephen (2002) therefore recommends taking the arithmetic mean of the two estimates. The F value remains the same, but if the sphericity is violated, the interpretation of the F value must be adjusted. The degrees of freedom are corrected downwards, which means that the F value becomes significant less quickly.

7.4 Logic of repeated measures ANOVA

The logic behind a repeated measures ANOVA is very similar to that of a between-subjects ANOVA. A between-subjects ANOVA partitions total variability into between-groups variability (SS_b) and within-groups variability (SS_w). Within-group variability (SS_w) is defined as the error variability (SS_{error}). Following division by the appropriate degrees of freedom, a mean sum of squares for between-groups (MS_b) and within-groups (MS_w) is determined and an F-statistic is calculated as the ratio of MS_b to MS_w (or MS_{error}).

$$F = \frac{MS_b}{MS_w} = \frac{MS_b}{MS_{error}}$$

A repeated measures ANOVA calculates an F-statistic in a similar way:

$$F = \frac{MS_{conditions}}{MS_{error}} = \frac{MS_{time}}{MS_{error}}$$

A repeated measures ANOVA can further partition the error term, thus reducing its size:

$$SS_{error} = SS_w - SS_{subjects} = SS_T - SS_{conditions} - SS_{subjects}$$

This has the effect of increasing the value of the F-statistic due to the reduction of the denominator and leading to an increase in the power of the test to detect significant differences between means. With a repeated measures ANOVA, as we are using the same subjects in each group, we can remove the variability due to the individual differences between subjects, referred to as $SS_{subjects}$, from the within-groups variability (SSw) by treating each subject as a block. That is, each subject becomes a level of a factor called subjects. The ability to subtract $SS_{subjects}$ will leave us with a smaller SS_{error} term.

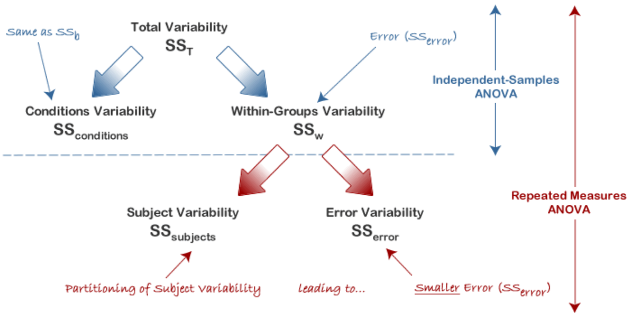


Figure 7.1: Source: https://statistics.laerd.com/statistical-guides/repeated-measures-anova-statistical-guide.php

Now that we have removed the between-subjects variability, our new SS_{error} only reflects individual variability to each condition. You might recognize this as the interaction effect of subject by conditions (how subjects react to the different conditions).

ℹ Computation example for one-way repeated measures ANOVA

The calculation of SS_{time} is the same as for SS_b in an independent ANOVA, and can be expressed as:

$$SS_{time} = SS_b = \sum n_i(\overline{x}_i - \overline{x})^2$$

where k = number of conditions, n_i = number of subjects under each (ith) condition, $\overline{x}_i$ = mean score for each (ith) condition, and $\overline{x}$ = grand mean. Within-groups variation (SS_w) is also calculated in the same way as in an independent ANOVA, expressed as follows:

$$SS_w = \sum(x_{i1} - \overline{x}_1)^2 + \sum(x_{i2} - \overline{x}_2)^2 + ... + \sum(x_{ik} - \overline{x}_k)^2$$

where k = number of conditions, and x_{i1} = score of the ith subject in group 1.
We treat each subject as a level of an independent factor called subjects. We can then calculate $SS_{subjects}$ as follows:

$$SS_{subjects} = k\sum(\overline{x}_i - \overline{x})^2$$

where k = number of conditions, $\overline{x}_i$ = mean subject i, and $\overline{x}$ = grand mean.
To determine the mean sum of squares for time (MS_{time}) we divide SS_{time} by its associated degrees of freedom $(k-1)$:

$$MS_{time} = \frac{SS_{time}}{(k-1)}$$

We do the same for the mean sum of squares for error (MS_{error}), this time dividing by $(n-1)(k-1)$ degrees of freedom:

$$MS_{time} = \frac{SS_{error}}{(n-1)(k-1)}$$

7.5 F-statistic

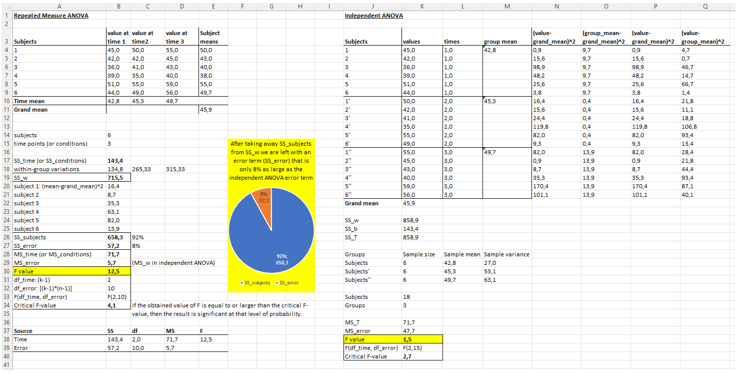
The sums of squares depend on the number of measurement times and cases, so the variance (standardized SS) is used to calculate the F-statistic.

We can calculate the F-statistic as:

$$F = \frac{MS_{time}}{MS_{error}}$$

We can then ascertain the critical F-statistic for our F-distribution with our degrees of freedom for condition and error, and determine whether our F-statistic indicates a statistically significant result.

The figure below shows that the calculation of the F-statistic for repeated measures ANOVA offers very different results compared to the results that would be obtained using the calculation for independent ANOVA. You can click on the figure to download the excel file.



	A	B	C	D	E
1	**Repeated Measure ANOVA**				
2					
3	**Subjects**	**value at time 1**	**value at time2**	**value at time 3**	**Subject means**
4	1	45,0	50,0	55,0	50,0
5	2	42,0	42,0	45,0	43,0
6	3	36,0	41,0	43,0	40,0
7	4	39,0	35,0	40,0	38,0
8	5	51,0	55,0	59,0	55,0
9	6	44,0	49,0	56,0	49,7
10	**Time mean**	42,8	45,3	49,7	
11	**Grand mean**				45,9
12					
13					
14	subjects	6			
15	time points (or conditions)	3			
16					
17	SS_time (or SS_conditions)	**143,4**			
18	within-group variations	134,8	265,33	315,33	
19	SS_w	**715,5**			
20	subject 1: (mean-grand_mean)^2	16,4			
21	subject 2	8,7			
22	subject 3	35,3			
23	subject 4	63,1			
24	subject 5	82,0			
25	subject 6	13,9			
26	SS_subjects	**658,3**	92%		
27	SS_error	**57,2**	8%		
28	MS_time (or MS_conditions)	**71,7**			
29	MS_error	**5,7**	(MS_w in independent ANOVA)		
30	F value	**12,5**			
31	df_time: (k-1)	2			
32	df_error: [(k-1)*(n-1)]	10			
33	F(df_time, df_error)	F(2,10)			
34	Critical F-value	**4,1**	If the obtained value of F is equal to or larger than the critical F-value, then the result is significant at that level of probability.		
35					
36					
37	**Source**	**SS**	**df**	**MS**	**F**
38	Time	143,4	2,0	71,7	12,5
39	Error	57,2	10,0	5,7	
40					
41					

After taking away SS_subjects from SS_w we are left with an error term (SS_error) that is only 8% as large as the independent ANOVA error term

	J	K	L	M	N	O	P	Q
1	**Independent ANOVA**							
2								
3	**Subjects**	**values**	**times**	**group mean**	**(value-grand_mean)^2**	**(group_mean-grand_mean)^2**	**(value-grand_mean)^2**	**(value-group_mean)^2**
4	1	45,0	1,0	42,8	0,9	9,7	0,9	4,7
5	2	42,0	1,0		15,6	9,7	15,6	0,7
6	3	36,0	1,0		98,9	9,7	98,9	46,7
7	4	39,0	1,0		48,2	9,7	48,2	14,7
8	5	51,0	1,0		25,6	9,7	25,6	66,7
9	6	44,0	1,0		3,8	9,7	3,8	1,4
10	1'	50,0	2,0	45,3	16,4	0,4	16,4	21,8
11	2'	42,0	2,0		15,6	0,4	15,6	11,1
12	3'	41,0	2,0		24,4	0,4	24,4	18,8
13	4'	35,0	2,0		119,8	0,4	119,8	106,8
14	5'	55,0	2,0		82,0	0,4	82,0	93,4
15	6'	49,0	2,0		9,3	0,4	9,3	13,4
16	1''	55,0	3,0	49,7	82,0	13,9	82,0	28,4
17	2''	45,0	3,0		0,9	13,9	0,9	21,8
18	3''	43,0	3,0		8,7	13,9	8,7	44,4
19	4''	40,0	3,0		35,3	13,9	35,3	93,4
20	5''	59,0	3,0		170,4	13,9	170,4	87,1
21	6''	56,0	3,0		101,1	13,9	101,1	40,1
22	**Grand mean**	45,9						
23								
24	SS_w	858,9						
25	SS_b	143,4						
26	SS_T	858,9						
27								
28	Groups	Sample size	Sample mean	Sample variance				
29	Subjects	6	42,8	27,0				
30	Subjects'	6	45,3	53,1				
31	Subjects''	6	49,7	63,1				
32								
33	Subjects	18						
34	Groups	3						
35								
36	MS_T	71,7						
37	MS_error	47,7						
38	F value	**1,5**						
39	F(df_time, df_error)	F(2,15)						
40	Critical F-value	**2,7**						
41								

Figure 7.2: F-statistic calculation

7.6 Two-way repeated measures ANOVA

A two-way repeated measures ANOVA compares the mean differences between groups that have been split on two within-subjects factors (also known as independent variables). A two-way repeated measures ANOVA is often used in studies where you have measured a dependent variable over two or more time points, or when subjects have undergone two or more conditions (e.g., "time" and "conditions").

Remember that the two-way (repeated measures) ANOVA is an omnibus test statistic and cannot tell you which specific groups within each factor were significantly different from each other. It only tells you that at least two of the groups were different. To determine which groups differ from each other, you can use post hoc tests.

7.7 Effect size

The partial eta-squared is specific to the factor i, but if there are several factors, you cannot add the individual partial eta-squares to form a total value because the denominator does not contain the total sum of squares (total variance)!

Partial eta-squared is where the the $SS_{subjects}$ has been removed from the denominator:

$$\eta^2_{partial} = \frac{SS_{conditions}}{SS_{conditions} + SS_{error}}$$

> **i** Nota bene: critics about Eta-square
>
> Eta-square overestimates the effect (i.e. the explained variance shares are assumed to be too large), because in samples there are very likely to be group differences even if there are no differences in the total population (especially with small n)!
> Therefrore, a recommendation of many authors is not to use the measured variances (SS) as the basis for the calculation of effect sizes, but their estimates in the total population (then no more bias). This parameter is called omega-square.

7.8 Pairwise comparisons

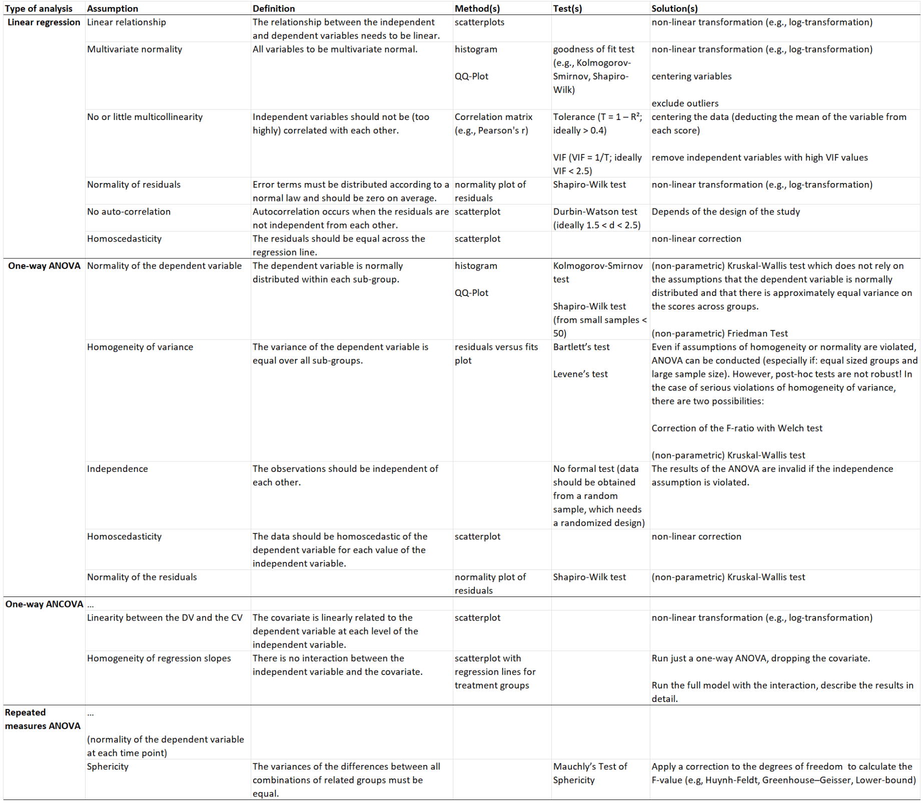
There are several methods to conduct pairwise comparisons:

- Contrasts: particularly useful for repeated measurement designs: each measurement time point is tested against the previous one: t2 vs. t1, t3 vs. t2, etc.

- Post-hoc tests (Bonferroni correction is recommended for violations of sphericity)
- Simple Effects also possible

7.9 Recap on the models' assumptions

Below is a recap of the assumptions of the different models that we have seen so far (linear regression, ANOVA, repeated measures ANOVA). The assumptions are defined, methods and tests are highlighted, as well as solutions proposed.



Type of analysis	Assumption	Definition	Method(s)	Test(s)	Solution(s)
Linear regression	Linear relationship	The relationship between the independent and dependent variables needs to be linear.	scatterplots		non-linear transformation (e.g., log-transformation)
	Multivariate normality	All variables to be multivariate normal.	histogram QQ-Plot	goodness of fit test (e.g., Kolmogorov-Smirnov, Shapiro-Wilk)	non-linear transformation (e.g., log-transformation) centering variables exclude outliers
	No or little multicollinearity	Independent variables should not be (too highly) correlated with each other.	Correlation matrix (e.g., Pearson's r)	Tolerance ($T = 1 – R^2$; ideally > 0.4) VIF ($VIF = 1/T$; ideally $VIF < 2.5$)	centering the data (deducting the mean of the variable from each score) remove independent variables with high VIF values
	Normality of residuals	Error terms must be distributed according to a normal law and should be zero on average.	normality plot of residuals	Shapiro-Wilk test	non-linear transformation (e.g., log-transformation)
	No auto-correlation	Autocorrelation occurs when the residuals are not independent from each other.	scatterplot	Durbin-Watson test (ideally $1.5 < d < 2.5$)	Depends of the design of the study
	Homoscedasticity	The residuals should be equal across the regression line.	scatterplot		non-linear correction
One-way ANOVA	Normality of the dependent variable	The dependent variable is normally distributed within each sub-group.	histogram QQ-Plot	Kolmogorov-Smirnov test Shapiro-Wilk test (from small samples < 50)	(non-parametric) Kruskal-Wallis test which does not rely on the assumptions that the dependent variable is normally distributed and that there is approximately equal variance on the scores across groups. (non-parametric) Friedman Test
	Homogeneity of variance	The variance of the dependent variable is equal over all sub-groups.	residuals versus fits plot	Bartlett's test Levene's test	Even if assumptions of homogeneity or normality are violated, ANOVA can be conducted (especially if: equal sized groups and large sample size). However, post-hoc tests are not robust! In the case of serious violations of homogeneity of variance, there are two possibilities: Correction of the F-ratio with Welch test (non-parametric) Kruskal-Wallis test
	Independence	The observations should be independent of each other.		No formal test (data should be obtained from a random sample, which needs a randomized design)	The results of the ANOVA are invalid if the independence assumption is violated.
	Homoscedasticity	The data should be homoscedastic of the dependent variable for each value of the independent variable.	scatterplot		non-linear correction
	Normality of the residuals		normality plot of residuals	Shapiro-Wilk test	(non-parametric) Kruskal-Wallis test
One-way ANCOVA	...				
	Linearity between the DV and the CV	The covariate is linearly related to the dependent variable at each level of the independent variable.	scatterplot		non-linear transformation (e.g., log-transformation)
	Homogeneity of regression slopes	There is no interaction between the independent variable and the covariate.	scatterplot with regression lines for treatment groups		Run just a one-way ANOVA, dropping the covariate. Run the full model with the interaction, describe the results in detail.
Repeated measures ANOVA	... (normality of the dependent variable at each time point)				
	Sphericity	The variances of the differences between all combinations of related groups must be equal.		Mauchly's Test of Sphericity	Apply a correction to the degrees of freedom to calculate the F-value (e.g, Huynh-Feldt, Greenhouse–Geisser, Lower-bound)

Figure 7.3: Models' assumptions

7.10 In a nutshell

In the ANOVA with repeated measures, a distinction is made between the between-participant and the within-participant variance.

The within-participant variance can be further subdivided into model explained variance and unexplained (error) variance. The F-value is calculated from their ratio.

A prerequisite for the ANOVA with repeated measurements is sphericity (i.e. homogeneous variances between the measurement times). The Mauchly's test checks this requirement, it should not be significant.

If there is no sphericity, the degrees of freedom for the critical F-value must be corrected. For small violations (epsilon $> .75$) with the Huynh & Feldt correction, for larger violations (epsilon $< .75$) with the Greenhouse & Geisser correction.

7.11 How it works in R?

Important note: For repeated measures ANOVA in R, it requires the long format of data. For the long format, we would need to stack the data from each individual into a vector (data at each time in a single column). If the database is in short format (data at each time in different columns), the function melt() from the R package reshape2 can be used. See the lecture slides on repeated measures ANOVA:

You can also download the PDF of the slides here: PDF

7.12 Quiz

True

False

Statement

When Mauchly's test for equality of variances fails to show significance, you have evidence that the data are suitable for the application of the One-way ANOVA repeated measures test.

You have conducted the One-way ANOVA repeated measures test, and results indicate an F value of 29.34 with a Sig. (significance) level of .506. Does it suggest that the means are equal?

If you find overall significance for five measurements, you may then test for significance for 10 pairwise comparisons.

In a two-way repeated measures ANOVA the distribution of the dependent variable in each combination of the related groups should be approximately normally distributed.

View Results

My results will appear here

7.13 Example from the literature

The following article relies on repeated measurements ANOVA as a method of analysis:

Hameleers, M., Brosius, A., & de Vreese, C. H. (2021). Where's the fake news at? European news consumers' perceptions of misinformation across information sources and topics. Harvard Kennedy School Misinformation Review. Available here.

Please reflect on the following questions:

- What is the research question of the study?
- What are the research hypotheses?
- Is repeated measurements ANOVA an appropriate method of analysis to answer the research question?
- What are the main findings of the repeated measurements ANOVA analysis?

7.14 Time to practice on your own

You can download the PDF of the exercises here:

7.14.1 Exercise 1: strengthening environmental protection over time

Use the data from the Selects 2019 Panel Survey and assess whether respondents' stance towards strengthening environmental protection has increased over the first three waves (before, during and after the campaign).

Start by downloading the data and by selecting the variables.

```
library(foreign)
db <- read.spss(file=paste0(getwd(),
                "/data/1184_Selects2019_Panel_Data_v4.0.sav"),
                use.value.labels = F,
                to.data.frame = T)
sel <- db |>
  dplyr::select(id,
    # wave 1
    W1_f15340d,
    # wave 2
    W2_f15340d,
    # wave 3
    W3_f15340d) |>
  stats::na.omit()
# inverse the scale
sel$W1_f15340d=(sel$W1_f15340d-6)*(-1)
sel$W2_f15340d=(sel$W2_f15340d-6)*(-1)
sel$W3_f15340d=(sel$W3_f15340d-6)*(-1)
```

Next, reshape the data so that there are in a long format.

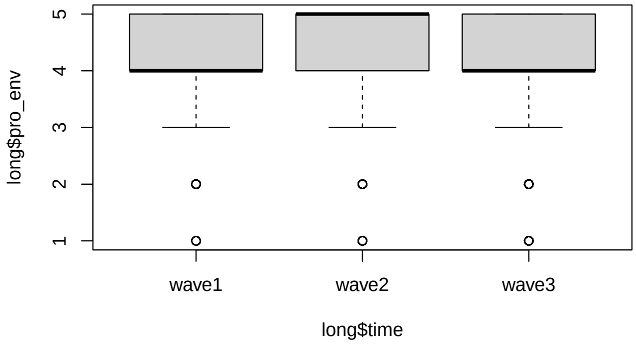
```
long <- reshape(as.data.frame(sel),
                direction="long",
                varying = c("W1_f15340d","W2_f15340d","W3_f15340d"),
                v.names = "pro_env",
                times =c("wave1","wave2","wave3"))
```

Then, we can check the normality of the dependent variable using the Shapiro-Wilk normality test:

```
# Shapiro-Wilk test
long |>
  dplyr::group_by(time) |>
  rstatix::shapiro_test(pro_env)
## # A tibble: 3 x 4
##   time  variable statistic        p
##   <chr> <chr>        <dbl>    <dbl>
## 1 wave1 pro_env      0.773 8.31e-45
## 2 wave2 pro_env      0.743 8.74e-47
## 3 wave3 pro_env      0.786 6.03e-44
```

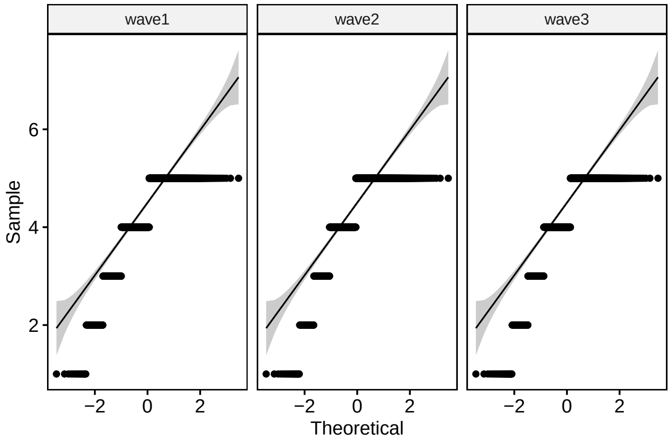
Interpretation

If the data is normally distributed, the p-value should be greater than 0.05., which is not the case here. Note that we need to test if the data was normally distributed at each time point. You can also visualize the distribution over time using boxplots.



Nota bene: If your sample size is greater than 50, the normal QQ plot is preferred because at larger sample sizes the Shapiro-Wilk test becomes very sensitive even to a minor deviation from normality.

```
# QQ plot
ggpubr::ggqqplot(long, "pro_env", facet.by = "time")
```



The assumption of sphericity will be automatically checked during the computation of the ANOVA test using the R function anova_test(). By using the function get_anova_table() to extract the ANOVA table, the Greenhouse-Geisser sphericity correction is automatically applied to factors violating the sphericity assumption. Now, we can check whether there are group differences:

```
# group differences
res.aov <- rstatix::anova_test(data = long,
                      dv = pro_env,
                      wid = id,
                      within = time)
res.aov
## ANOVA Table (type III tests)
##
## $ANOVA
##   Effect DFn  DFd      F        p p<.05   ges
## 1   time   2 3682 37.718 6.09e-17     * 0.004
##
## $`Mauchly's Test for Sphericity`
##   Effect     W     p p<.05
## 1   time 0.993 0.001     *
##
## $`Sphericity Corrections`
##   Effect   GGe        DF[GG]    p[GG] p[GG]<.05   HFe        DF[HF]    p[HF] p[HF]<.05
## 1   time 0.993 1.99, 3655.01 7.76e-17         * 0.994 1.99, 3658.94 7.49e-17         *
```

The sphericity test is violated (W=0.993 and p<0.05). Therefore, we need to look at the Greenhouser-Geisser (GG) Huynh-Feldt (HF) corrections. The p-values (p[GG] and p[HF]) are significant, thus indicating that the observed F values are significant and accepting the hypothesis that we have different means.

> ℹ Interpretation
>
> The pro-environment score was statistically significantly different at the different time points: F = 37.71, p < 0.05. Furthermore, the value for “ges” (generalized effect size) gives us the amount of variability due to the within-subjects factor.

Finally, we can assess which group (or time) differences are statistically significant:

```
# Post-hoc test to assess differences
pwc <- long |>
  rstatix::pairwise_t_test(
    pro_env ~ time,
    paired = TRUE,
    p.adjust.method = "bonferroni"
    )
pwc[,c(2,3,6,8,10)]
## # A tibble: 3 x 5
##   group1 group2 statistic        p p.adj.signif
##   <chr>  <chr>      <dbl>    <dbl> <chr>
## 1 wave1  wave2      -3.15 2   e- 3 **
## 2 wave1  wave3       5.26 1.58e- 7 ****
## 3 wave2  wave3       8.80 3.19e-18 ****
```

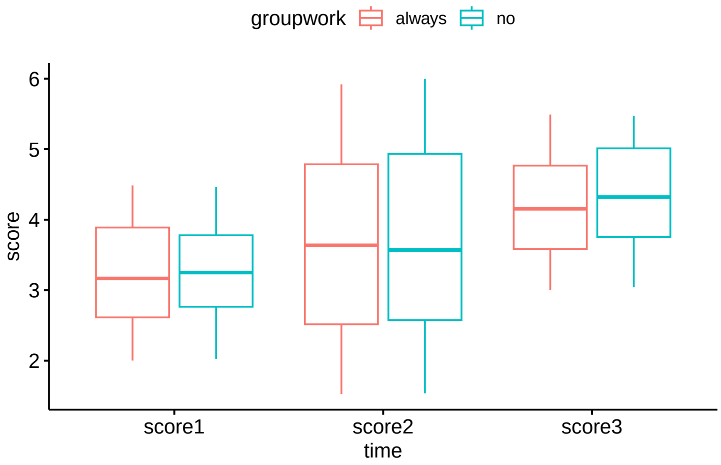
7.14.2 Exercise 2: two-way repeated measure ANOVA

Let's create a dataset containing a score measured at three points in time. In a second step, we will investigate if (frequently) working in group can induce a significant increase of the score over time.

```
data <- data.frame(matrix(nrow = 200, ncol = 0))
set.seed(123)
data$score1 <- runif(nrow(data), min=2, max=4.5)
data$score2 <- runif(nrow(data), min=1.5, max=6)
data$score3 <- runif(nrow(data), min=3, max=5.5)
# assign id
data$id = rep(seq(1:100),2)
# assign group work variable
data$groupwork = c(rep(c("always"),100), rep(c("no"),100))
# copy of the data
copy = data
# re-arrange the data
data <- data |>
  tidyr::gather(key = "time", value = "score", score1, score2, score3) |>
  rstatix::convert_as_factor(id, time)
```

Now, we will test whether there is significant interaction between working in group and time on the score. We can use boxplots of the score colored by working in group:

```
ggpubr::ggboxplot(
  data, x = "time",
  y = "score",
  color = "groupwork"
  )
```



We can check whether there are outliers:

```r
data |>
  dplyr::group_by(groupwork, time) |>
  rstatix::identify_outliers(score)
## [1] groupwork  time       id          score      is.outlier is.extreme
## <0 lignes> (ou 'row.names' de longueur nulle)
```

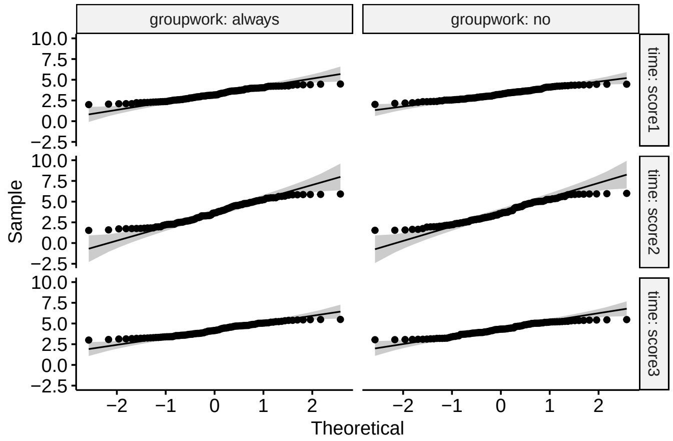
We next compute Shapiro-Wilk test to test for the normality assumption for each combinations of factor levels:

```r
# Shapiro
data |>
  dplyr::group_by(groupwork, time) |>
  rstatix::shapiro_test(score)
## # A tibble: 6 x 5
##   groupwork time   variable statistic        p
##   <chr>     <fct>  <chr>        <dbl>    <dbl>
## 1 always    score1 score        0.952 0.00119
## 2 always    score2 score        0.945 0.000418
## 3 always    score3 score        0.948 0.000592
## 4 no        score1 score        0.964 0.00736
## 5 no        score2 score        0.950 0.000789
## 6 no        score3 score        0.945 0.000379
```

Interpretation

The output shows that the score is generally not normally distributed (p < 0.05).
Note that for large sample size, we can also use the QQ plot:

```r
# QQ plot
ggpubr::ggqqplot(data, "score") +
  ggplot2::facet_grid(time ~ groupwork, labeller = "label_both")
```



We can assess whether there is a statistically significant two-way interactions between group work and time:

```
# We also need to convert id and time into factor variables
# data$groupwork <- as.factor(data$groupwork)
data$time <- as.factor(data$time)
data$id <- as.factor(data$id)
res.aov <- rstatix::anova_test(
  data = data,
  dv = score,
  wid = id,
  within = c(groupwork, time)
  )
# rstatix::get_anova_table(res.aov)
res.aov
## ANOVA Table (type III tests)
##
## $ANOVA
##           Effect DFn DFd      F        p p<.05      ges
## 1      groupwork   1  99  0.454 5.02e-01       0.000696
## 2           time   2 198 46.979 2.01e-17     * 0.153000
## 3 groupwork:time   2 198  0.050 9.51e-01       0.000163
##
## $`Mauchly's Test for Sphericity`
##           Effect     W         p p<.05
## 1           time 0.791 0.0000101     *
## 2 groupwork:time 0.809 0.0000317     *
##
## $`Sphericity Corrections`
##           Effect   GGe       DF[GG]    p[GG] p[GG]<.05   HFe       DF[HF]    p[HF]
## 1           time 0.827 1.65, 163.74 7.74e-15         * 0.839 1.68, 166.17 5.07e-15
## 2 groupwork:time 0.840 1.68, 166.31 9.28e-01           0.853 1.71, 168.85 9.30e-01
##   p[HF]<.05
## 1         *
## 2
```

ℹ Interpretation

There is a statistically no significant two-way interactions between group work and time, F = 0.05, p > 0.05. Furthermore, the shericity test is violated, thus suggesting to look at the GG and HF corrections.

Procedure for post-hoc test:

A significant two-way interaction indicates that the impact that one factor (e.g., group work) has on the outcome variable (e.g., score) depends on the level of the other factor (e.g., time), and vice versa. So, you can decompose a significant two-way interaction into:

- i) simple main effect (one-way model of the first variable at each level of the second variable: e.g., group work at each time point);
- ii) simple pairwise comparisons if the simple main effect is significant (pairwise comparisons to determine which groups are different: e.g., pairwise comparisons between categories of group work).

For a non-significant two-way interaction, you need to determine whether you have any statistically significant main effects from the ANOVA output (e.g., comparisons for group work and time variable).

```
# Comparisons for group work
res1 = data |>
  rstatix::pairwise_t_test(
    score ~ groupwork,
    paired = TRUE,
    p.adjust.method = "bonferroni"
)
res1[,c(2,3,6,8,10)]
## # A tibble: 1 x 5
##   group1 group2 statistic     p p.adj.signif
##   <chr>  <chr>      <dbl> <dbl> <chr>
## 1 always no        -0.661 0.509 ns
# Comparisons for the time variable
res2 = data |>
  rstatix::pairwise_t_test(
    score ~ time,
    paired = TRUE,
    p.adjust.method = "bonferroni"
)
res2[,c(2,3,6,8,10)]
## # A tibble: 3 x 5
##   group1 group2 statistic        p p.adj.signif
##   <chr>  <chr>      <dbl>    <dbl> <chr>
## 1 score1 score2     -4.05 7.31e- 5 ***
## 2 score1 score3    -13.5  5.28e-30 ****
## 3 score2 score3     -5.04 1.02e- 6 ****
```

Chapter 8

Logistic regression

8.1 Logistic regression analyis

Logistic regression is a model with a binary dependent variable (e.g., 0 = not elected versus 1 = elected). In this case, we are interested in knowing the probability of a phenomenon (e.g., get elected, lose a job, becoming sick, etc) to occur:

$$Y(P = 1) = a + bX + e$$

also referred as Linear Probability Model (LPM). However, the interpretation is complicated be that fact that applying LPM may return values outside the [0,1] interval (probabilities are necessary between 0 and 1!).

Example: An increase in political experience (e.g., number of years a candidate has joined his party) by 1 year increases/decreases the probability of being elected as a national representative by $b * 100$ percentage points.

Simple regression: professional consultant example

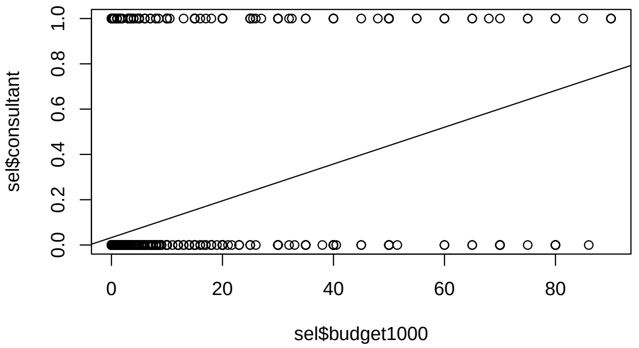
Let's predict the probability that a candidate employs professional consultants explained by the available campaigning budget in a linear regression framework.
We can first look at the distribution of political candidates employing a consultant:

```r
library(foreign)
db <- read.spss(file=paste0(getwd(),
                "/data/1186_Selects2019_CandidateSurvey_Data_v1.1.0.sav"),
                use.value.labels = F,
                to.data.frame = T)
sel <- db |>
  dplyr::select(B11,B12,B6) |>
  stats::na.omit() |>
  dplyr::rename("consultant"="B11",
                "budget"="B12",
                "personalization"="B6") |>
  plyr::mutate(budget=as.numeric(as.character(as.character(budget))))
sel$consultant <- ifelse(sel$consultant==1,1,0)
# keep candidates with a budget<100'000
sel <- sel[sel$budget<100000,]
sel$budget1000 <- sel$budget/1000
# distribution
nb <- table(sel$consultant)
prop <- round(table(sel$consultant)/nrow(sel), 2)
prop <- cbind(nb,prop)
rownames(prop) <- c("without_consultant","with_consultant")
```

```r
# show table
as.data.frame(prop)
##                       nb prop
## without_consultant 1712 0.92
## with_consultant     143 0.08
```

We see that ~8% of candidates employ a professional consultant, the probability to employ a consultant being calculated as follows: 143/(143+1712)=8%.
We can plot the relationship between hiring a consultant and the budget:

```r
plot(sel$budget1000,sel$consultant)
abline(lm(consultant ~ budget1000, data=sel))
```



We see that the linear line is not ideal and does not fit the data. A dichotomous dependent variable does not display a normal distribution. It cannot have a linear relationship with an explanatory variable. Therefore, linear regression could make predictions smaller than 0 and bigger than 1, which is not possible!

8.2 The problem with LPM

LPM poses several issues:

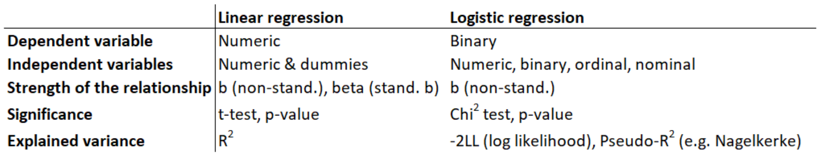
- non-linearity
- non-normal distribution of the errors (only two possible outcomes)
- heteroscedasticity
- difficulty in interpreting the results

Therefore, we need a non-linear model predicting probability of an event which remains in [0,1] bounds. The question is how to constrain all possible outcomes between 0 and 1.

Non-linear probability models are essential in social sciences which often deal with categorical dependent variables (e.g., binary, ordinal, nominal). There exists several models depending of the measurement level of the dependent variable. Logistic regression allows to predict models for categorical dependent variables in the simplest binary form and is part of the Generalized Linear Model (GLM) framework:

- Binomial: binary variable
- Multinomial: categorical variable
- Gaussian: interval variable
- Poisson: count data (discrete)
- Gamma: >0 (non-discrete)

8.3 Differences between linear and logistic regressions



	Linear regression	Logistic regression
Dependent variable	Numeric	Binary
Independent variables	Numeric & dummies	Numeric, binary, ordinal, nominal
Strength of the relationship	b (non-stand.), beta (stand. b)	b (non-stand.)
Significance	t-test, p-value	Chi^2 test, p-value
Explained variance	R^2	-2LL (log likelihood), Pseudo-R^2 (e.g. Nagelkerke)

8.4 Odds ratio (OR)

OR is the relative chance of an event happening under two different conditions. In other words, it is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group (relative odds).

Let's take the following games:

- Game 1: profit of 20, loss of 100, odds of 20/100=0.2
- Game 2: profit of 10, loss of 100, odds of 10/100=0.1

Here, the OR is 0.2/0.1=2. Since the OR is >1, it suggests that the odds in game 1 is higher than in game 2: the relative chance of winning instead of loosing is 2-times higher in game 1.

OR: further examples

- An OR of 1.3 means there is a 30% increase in the odds of an outcome.
- An OR of 2 means there is a 100% increase in the odds of an outcome (same as saying that there is a doubling of the odds of the outcome).
- An OR of 0.3 means there is an 70% decrease in the odds of an outcome.

8.5 Constraining all possible outcomes

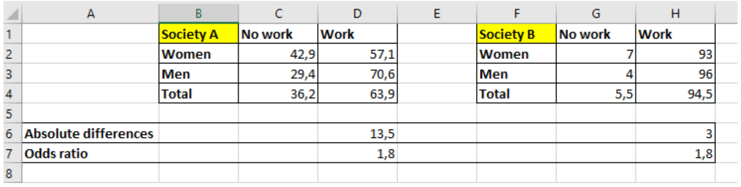
8.5.1 Step 1: p to odds

First, we need to transform Y into $p/(1-p)$ so that the binary dependent variable can be expressed as a function of continuous positive values ranging from 0 to $+\infty$.

The formal interpretation suggests that "a unit increase in X changes the odds that Y=1 instead of Y=0 by a factor of e^z (antilog of the odds), all else equal.

- OR less than 1= negative effect (log-odds)
- OR greater than 1= positive effect (log-odds)

But, OR give no indication about the magnitude of the implied change in the probabilities. Let's look at the following example of two societies:



	A	B	C	D	E	F	G	H
1		Society A	No work	Work		Society B	No work	Work
2		Women	42,9	57,1		Women	7	93
3		Men	29,4	70,6		Men	4	96
4		Total	36,2	63,9		Total	5,5	94,5
5								
6	Absolute differences			13,5				3
7	Odds ratio			1,8				1,8
8								

Figure 8.1: Absolute differences versus odds ratio

In the above scenario, different differences in probabilities can be associated to the same odds ratio. The social **mechanisms** responsible for gender effect on having work are the same in the two societies, but the **intensity** of the effect resulting from those mechanisms is much stronger in A than B.

Note that OR can be expressed as % change in the odds $100 * (OR - 1)$.

8.5.2 Step 2: odds to log odds

Up to now, we have $odds = p/(1-p)$, which we now transform into log odds: $ln(p/(1-p))$.

- If p=0.5, odds=1 and log odds: 0, which suggests no effect
- If p=0.8 of success (0.2 of failure), odds=4 and log odds: 1.38, which suggests a positive effect
- If p=0.8 of failure (0.2 of success), odds=1/4 and log odds: -1.38, which suggests a negative effect

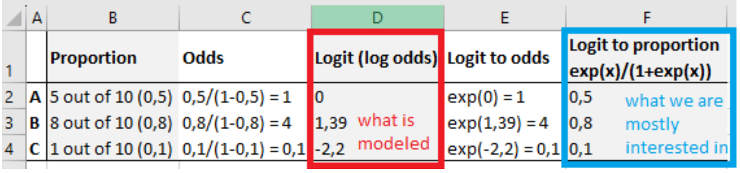
So, for log odds, only the sign (direction of the coefficient) can be interpreted.

8.5.3 Step 3: back to probabilities

The problem with log odds (also called logit) is that they are not easy to interpret: "logarithmic odds of an increase in budget by one franc on being elected". Therefore, we need to transform log odds into probabilities:

$$P_i(y=1) = \frac{e^{a+b_1x_1+b_2x_2}}{1+e^{a+b_1x_1+b_2x_2}}$$

where $e^1 = 2.71828$ and $e^{ln(x)} = x$.

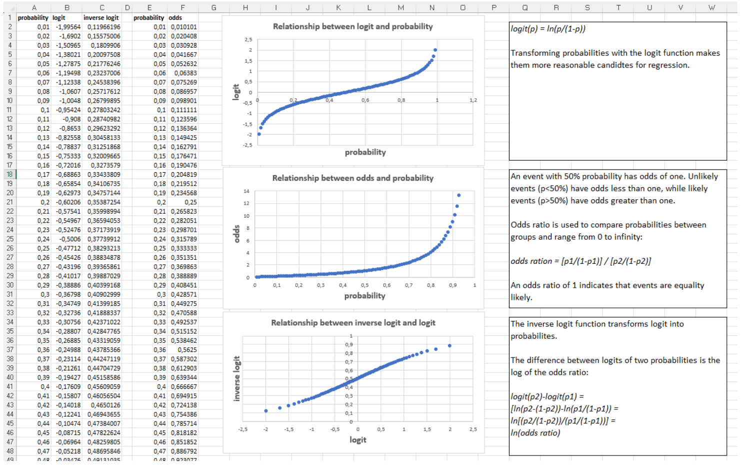


	Proportion	Odds	Logit (log odds)	Logit to odds	Logit to proportion exp(x)/(1+exp(x))
A	5 out of 10 (0,5)	0,5/(1-0,5) = 1	0	exp(0) = 1	0,5 what we are
B	8 out of 10 (0,8)	0,8/(1-0,8) = 4	1,39 what is	exp(1,39) = 4	0,8 mostly
C	1 out of 10 (0,1)	0,1/(1-0,1) = 0,1	-2,2 modeled	exp(-2,2) = 0,1	0,1 interested in

Figure 8.2: Logistic transformations

8.6 Relationship between probability, logit and odds ratio

If p is a probability, then p/(1 − p) is the corresponding odds. Furthermore, the logit of the probability is the logarithm of the odds:



probability	logit	inverse logit	probability	odds
0,01	-1,99564	0,11966196	0,01	0,010101
0,02	-1,6902	0,15575006	0,02	0,020408
0,03	-1,50965	0,1809906	0,03	0,030928
0,04	-1,38021	0,20097508	0,04	0,041667
0,05	-1,27875	0,21776246	0,05	0,052632
0,06	-1,19498	0,23237006	0,06	0,06383
0,07	-1,12338	0,24538396	0,07	0,075269
0,08	-1,0607	0,25717612	0,08	0,086957
0,09	-1,0048	0,26799895	0,09	0,098901
0,1	-0,95424	0,27803242	0,1	0,111111
0,11	-0,908	0,28740982	0,11	0,123596
0,12	-0,8653	0,29623292	0,12	0,136364
0,13	-0,82558	0,30458133	0,13	0,149425
0,14	-0,78837	0,31251868	0,14	0,162791
0,15	-0,75333	0,32009665	0,15	0,176471
0,16	-0,72016	0,3273579	0,16	0,190476
0,17	-0,68863	0,33433809	0,17	0,204819
0,18	-0,65854	0,34106735	0,18	0,219512
0,19	-0,62973	0,34757144	0,19	0,234568
0,2	-0,60206	0,35387254	0,2	0,25
0,21	-0,57541	0,35998994	0,21	0,265823
0,22	-0,54967	0,36594053	0,22	0,282051
0,23	-0,52476	0,37173919	0,23	0,298701
0,24	-0,5006	0,37739912	0,24	0,315789
0,25	-0,47712	0,38293213	0,25	0,333333
0,26	-0,45426	0,38834878	0,26	0,351351
0,27	-0,43196	0,39365861	0,27	0,369863
0,28	-0,41017	0,39887029	0,28	0,388889
0,29	-0,38886	0,40399168	0,29	0,408451
0,3	-0,36798	0,40902999	0,3	0,428571
0,31	-0,34749	0,41399185	0,31	0,449275
0,32	-0,32736	0,41888337	0,32	0,470588
0,33	-0,30756	0,42371022	0,33	0,492537
0,34	-0,28807	0,42847765	0,34	0,515152
0,35	-0,26885	0,43319059	0,35	0,538462
0,36	-0,24988	0,43785366	0,36	0,5625
0,37	-0,23114	0,44247119	0,37	0,587302
0,38	-0,21261	0,44704729	0,38	0,612903
0,39	-0,19427	0,45158586	0,39	0,639344
0,4	-0,17609	0,45609059	0,4	0,666667
0,41	-0,15807	0,46056504	0,41	0,694915
0,42	-0,14018	0,4650126	0,42	0,724138
0,43	-0,12241	0,46943655	0,43	0,754386
0,44	-0,10474	0,47384007	0,44	0,785714
0,45	-0,08715	0,47822624	0,45	0,818182
0,46	-0,06964	0,48259805	0,46	0,851852
0,47	-0,05218	0,48695846	0,47	0,886792
0,48	-0,03476	0,49131035	0,48	0,923077

logit(p) = ln(p/(1-p))

Transforming probabilities with the logit function makes them more reasonable candidtes for regression.

An event with 50% probability has odds of one. Unlikely events (p<50%) have odds less than one, while likely events (p>50%) have odds greater than one.

Odds ratio is used to compare probabilities between groups and range from 0 to infinity:

odds ration = [p1/(1-p1)] / [p2/(1-p2)]

An odds ratio of 1 indicates that events are equality likely.

The inverse logit function transforms logit into probabilites.

The difference between logits of two probabilities is the log of the odds ratio:

logit(p2)-logit(p1) =
[ln(p2-(1-p2))-ln(p1/(1-p1)) =
ln[(p2/(1-p2))/(p1/(1-p1))] =
ln(odds ratio)

Figure 8.3: Relationship between probability, logit and odds ratio

8.7 Marginal predictions

The main benefit of marginal predictions is that it leaves everything at the mean, except for the variable that you are interested in. Therefore, one variable is changing while the others are not.

For a continuous covariate, the margins compute how P(Y=1) changes as X changes from 0 to 1, controlling for other variables in the model. For a dichotomous independent variable, the marginal effect equates the difference in the adjusted predictions for two groups (e.g., for women and men). For discrete covariate, the margins compute the effect of a discrete change of the covariate (discrete change effects).

8.8 Model fit

R^2 in OLS is a measure for model fit. It is based on differences between real observations and the regression line. It tries to remove as much error as possible.

In logistic regression, there is no comparable measure since all values on Y are either 1 or 0 . We are explaining probabilities (not explained variance). The model fit is based on Maximum Likelihood Estimation (MLE) which tries to guess parameters that have highest likelihood of producing observed sample patterns. Likelihood is the probability that our statistical model is actually found in a sample, thus the better the model fits the data, the more likelier it is.

The smaller the Log likelihood the better the model fit: by how much Log likelihood has decreased by adding (a) variable(s), and by calculating the significance of this difference. It is possible to compare models statistically by a Chi-squared test.

Note 1: Pseudo $R^2 = (-2LL0 - -2LL1)/-2LL0$ has no clear interpretation, but provides a good way to compare models (rather than assessing fit).

Note 2: Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are alternative measures (the smaller AIC or BIC, the better the model fit).

8.9 How it works in R?

See the lecture slides on logistic regression:

You can also download the PDF of the slides here: PDF

8.10 Quiz

True

False

Statement

Logistic regression is used to make predictions about a dichotomous dependent variable.

Odds can be defined as the number of times something occurs relative to the number of times it does not occur.

If the odds ratio of a dummy variable is greater than 1, then the group captured in the dummy variable is predicted to be more likely than the reference group to have something occur.

When there is exactly a 0.5 probability of something occurring, the log odds are 1.

View Results

My results will appear here

8.11 Example from the literature

The following article relies on logistic regression as a method of analysis:

Vogler, D., & Schäfer, M. S. (2020). Growing influence of university PR on science news coverage? A longitudinal automated content analysis of university media releases and newspaper coverage in Switzerland, 2003 2017. International Journal of Communication, 14, 22. Available here.

Please reflect on the following questions:

- What is the research question of the study?
- What are the research hypotheses?
- Is logistic regression an appropriate method of analysis to answer the research question?
- What are the main findings of the logistic regression analysis?

8.12 Time to practice on your own

You can download the PDF of the exercises here:

8.12.1 Exercise 1: probability of hiring a consultant according to campaign personalization

For instance, we are interested in measuring the likelihood of hiring a consultant (Y) explained by personalized style of campaigning (X). To do so, we will rely on the data covering the Swiss part of the Comparative Candidate Survey. We will be using the Selects 2019 Candidate Survey.

We can look at the likelihood of hiring of consultant (B11) by the level of campaign personalization (where B6 is recoded as 0=attention to the party and 10=attention to the candidate):

```
library(foreign)
db <- read.spss(file=paste0(getwd(),
                "/data/1186_Selects2019_CandidateSurvey_Data_v1.1.0.sav"),
                use.value.labels = F,
                to.data.frame = T)
sel <- db |>
```

```
  dplyr::select(B11,B12,B6) |>
  stats::na.omit() |>
  dplyr::rename("consultant"="B11",
                "budget"="B12",
                "personalization"="B6") |>
  plyr::mutate(budget=as.numeric(as.character(as.character(budget))))
sel$consultant <- ifelse(sel$consultant==1,1,0)
# keep candidates with a budget<100'000
sel <- sel[sel$budget<100000,]
# reverse the scale: higher values = higher personaliz.
sel$personalization <- as.numeric(as.character(sel$personalization))
sel$personalization <- (sel$personalization-11)*(-1)
# mean by level of personalization
p = aggregate(sel$consultant, by=list(sel$personalization), FUN=mean)
colnames(p) = c("personalization","mean")
p
##    personalization        mean
## 1                1 0.035398230
## 2                2 0.006896552
## 3                3 0.031690141
## 4                4 0.107279693
## 5                5 0.086956522
## 6                6 0.117391304
## 7                7 0.120000000
## 8                8 0.119047619
## 9                9 0.186440678
## 10              10 0.173913043
## 11              11 0.200000000
```

Now, we can calculate the odds of hiring a consultant for a very personalized campaign (personalization = 10):

Interpretation

$$\frac{0.17}{(1-0.17)} = 0.2$$

This suggests that for each candidate without a consultant, there are 0.2 candidates hiring a consultant. Alternatively:

$$\frac{(1-0.17)}{(1-(1-0.17))} = 4.9$$

This suggests that for each candidate hiring a consultant, there are 4.9 candidates without a consultant.

Now, calculate the odds of hiring a consultant for a very low personalized campaign (personalization = 0):

Interpretation

$$\frac{0.03}{(1-0.03)} = 0.03$$

Therefore, the odds ratio is: 0.2/0.03 = 6.7, suggesting that the odds of hiring a consultant are 6.7 higher for candidates with a very high personalized campaign than candidates with a very low personalized campaign.

The logit of the dependent variable (Y) is estimated by the following equation:

$$logit(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \epsilon$$

The logit does not indicate the probability that an event occurs. Apply the necessary transformation to know this probability (prob(Y=1)):

Answer

$$proba = \frac{exp^{logit}}{1 + exp^{logit}}$$

Let's go back to our example and run the logistic regression:

```
model2 <- glm(consultant ~ personalization,
              data=sel,
              family="binomial")
summary(model2)
##
## Call:
## glm(formula = consultant ~ personalization, family = "binomial",
##     data = sel)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -3.54005    0.19321  -18.32  < 2e-16 ***
## personalization  0.22052    0.03132    7.04 1.92e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1007.64  on 1854  degrees of freedom
## Residual deviance:  957.86  on 1853  degrees of freedom
## AIC: 961.86
##
## Number of Fisher Scoring iterations: 5
```

Coefficients in the above output are log odds: 0.22 means that by augmenting the personalization of one point, log odds change by 0.22.

Now, assess the odds of hiring a consultant for a very personalized campaign (personalization=10):

> **Interpretation**
>
> $$Logit = -3.29 + 0.23 * 10 = -0.99$$
>
> The odds ratio for the personalization variable is exp(0.22)=1.24. This suggests that, for each unit increase on the personalization scale, the odds increase by a factor of 1.24, which is equivalent to an increase of 24%. Beware that the odds ratio does not provide information about the probability of hiring a consultant. We can calculate the probability as follows:
>
> $$Probability = \frac{exp(logit)}{1 + exp(logit)} = \frac{e^{-0.99}}{(1 + e^{-0.99})} = 0.37$$

8.12.2 Exercise 2: predict the reliance of social media as campaigning tool

Using the same dataset, let's investigate the following question: how does the level of campaign personalization and the fact of being affiliated to a governmental party, and being an incumbent affect the reliance of social media as campaigning tool?

In this scenario, the binary outcome is whether politicians rely on social media (combination of B4m and B4p) and the predictors are personalization (B6), being affiliated to a governmental party (based on T9), and being an incumbent (T11c).

Let's prepare the data, including the selection and recoding of the relevant variables:

```
library(foreign)
db <- read.spss(file=paste0(getwd(),
                  "/data/1186_Selects2019_CandidateSurvey_Data_v1.1.0.sav"),
                use.value.labels = F,
                to.data.frame = T)
sel <- db |>
  dplyr::select(B4m,B4p,T9,B6,T11c) |>
  stats::na.omit() |>
  dplyr::rename("facebook"="B4m",
                "twitter"="B4p",
                "party"="T9",
                "personalization"="B6",
                "incumbentNC"="T11c")
# reliance on social media
sel$twitter=ifelse(sel$twitter>0,1,0)
sel$facebook=ifelse(sel$facebook>0,1,0)
sel$SMuse=ifelse(sel$facebook==1 | sel$twitter==1, 1, 0)
sel$SMuse=as.factor(sel$SMuse)
# party in government
sel$in_gov=ifelse(sel$party %in% c(1,2,3,4,7), 1, 0)
sel$in_gov=as.factor(sel$in_gov)
# personalization (invert scale)
sel$personalization <- as.numeric(as.character(sel$personalization))
sel$personalization <- (sel$personalization-10)*(-1)
# incumbent
sel$incumbentNC <- as.factor(sel$incumbentNC)
# head
head(sel[,c(3:ncol(sel))])
##   party personalization incumbentNC SMuse in_gov
## 1    11               0           0     0      0
## 2    11               5           0     1      0
## 3    11               3           1     1      0
## 4    11               5           0     1      0
## 5    11               0           0     0      0
## 6    11               0           0     0      0
```

Now, we can conduct logistic regression and interpret the findings. Recall that, for log odds, we interpret only the sign of the coefficients (positive/negative). Coefficients smaller than 1 suggests a negative effect (negative log odds) and coefficients larger than 1 suggest positive effect (positive log odds). You can also transform to percentages using the formula 100*(OR-1):

```
mod <- glm(SMuse ~
             personalization +
             in_gov +
             incumbentNC,
           data=sel,
           family = "binomial")
summary(mod)
##
## Call:
## glm(formula = SMuse ~ personalization + in_gov + incumbentNC,
##     family = "binomial", data = sel)
##
```

```
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)    
## (Intercept)      0.26890    0.08100   3.320 0.000901 ***
## personalization  0.16152    0.02046   7.896 2.89e-15 ***
## in_gov1          0.08618    0.09972   0.864 0.387475    
## incumbentNC1     0.96787    0.30413   3.182 0.001461 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2585.2  on 2094  degrees of freedom
## Residual deviance: 2487.0  on 2091  degrees of freedom
## AIC: 2495
##
## Number of Fisher Scoring iterations: 4
# transformation
exp(coef(mod))
##     (Intercept) personalization         in_gov1    incumbentNC1 
##        1.308526        1.175299        1.090003        2.632324 
```

ℹ Interpretation

In our example: when personalization goes up by one, the odds of relying on social media increase by a factor of 1.16, controlling for the other variables in the model. In other terms, when personalization goes up by one, the odds of using social media increase by 16% (100(1.16-1)).

The marginal effects indicate a change in predicted probability as X increases by 1. For categorical predictors, you have to take the predicted probability of the group A minus the predicted probability of the group B.

There are different ways of calculating predicted probabilities. In the social sciences, the most commonly used are Adjusted Predictions at the Means (APMs). For instance, we can assess the predicted probabilities of using social media for political incumbents, when the personalization level is at the mean and for incumbent not affiliated to a party in government.

```
newdata = data.frame(personalization=5,
                     in_gov="0",
                     incumbentNC="1")
predict(mod, newdata, type="response")
##         1 
## 0.8853785 
```

Nota bene: Marginal Effects at the Means (MEMs) are calculated by taking the difference of two APMs. Let's also calculate the predicted probabilities of using social media for political non-incumbents, when the personalization level is at the mean and for politicians not affiliated to a party in government. Then, calculate the difference between both predicted probabilities:

```
newdata2 = data.frame(personalization=5,
                      in_gov="0",
                      incumbentNC="0")
print(paste0("for incumbents: ",
             round(predict(mod, newdata, type="response"),2),
             "; for non-incumbents: ",
             round(predict(mod, newdata2, type="response"),2)))
## [1] "for incumbents: 0.89; for non-incumbents: 0.75"
```

In logistic regressions, there is no such R-squared value for general linear models. Instead, we can calculate a metric known as McFadden's R-Squared, which ranges from 0 to just under 1, with higher values indicating a better model fit. We use the following formula to calculate McFadden's R-Squared:

$$1 - (\frac{LL_{model}}{LL_{null}})$$

```
with(summary(mod), 1 - deviance/null.deviance)
## [1] 0.03798762
```

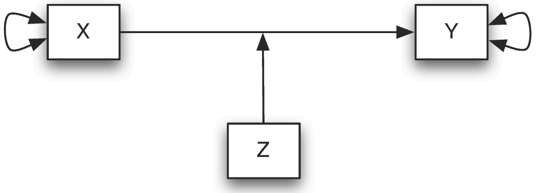
Chapter 9

Moderation analysis

9.1 Moderation analysis

Moderation analysis allows us to test for the influence of a third variable, Z, on the relationship between variables X and Y. Rather than testing a causal link between these other variables, moderation tests for when or under what conditions an effect occurs.

For instance, we can start with a bivariate relationship between an input variable X (e.g. training) and an outcome variable Y (e.g. test score). We can hypothesize that there is a relationship between them. A moderator variable Z (e.g. gender) is a variable that alters the strength of the relationship between X and Y.



Moderators can strengthen, weaken, or reverse the nature of a relationship. Moderation can be tested by looking for significant interactions between the moderating variable (Z) and the independent variable (X).

Nota bene: Moderators (when) are conceptually different from mediators (how/why) but some variables may be a moderator or a mediator depending on your question.

9.2 Similarity with ANOVA

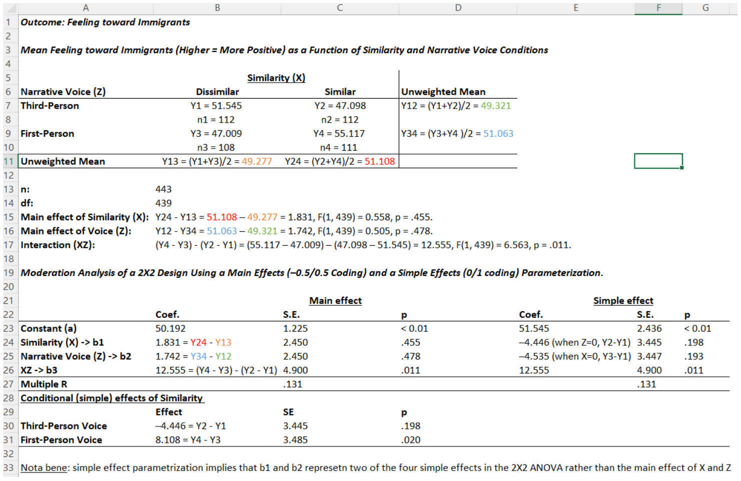
The moderation is an interaction and therefore comparable to the interaction in ANOVA. If X and moderator are dichotomous, the moderation corresponds to a 2x2 ANOVA.

However, a moderator moderates the causal relationship from X to Y. The scale level can be dichotomous, categorical or metric. Furthermore, a moderator must be causally independent of X.

Technically, moderations (interactions) are linked multiplicatively in the regression analysis: A x B.

Statistically, the moderator and X must always be considered "in isolation" (not just as moderation or interaction).

An illustration of the similarity between a simple moderation analysis and a 2x2 ANOVA can be found below (this example is taken from Igartua and Hayes (2021)):



	A	B	C	D	E	F	G
1	***Outcome: Feeling toward Immigrants***						
2							
3	***Mean Feeling toward Immigrants (Higher = More Positive) as a Function of Similarity and Narrative Voice Conditions***						
4							
5		**Similarity (X)**					
6	**Narrative Voice (Z)**	**Dissimilar**	**Similar**	**Unweighted Mean**			
7	**Third-Person**	Y1 = 51.545	Y2 = 47.098	Y12 = (Y1+Y2)/2 = 49.321			
8		n1 = 112	n2 = 112				
9	**First-Person**	Y3 = 47.009	Y4 = 55.117	Y34 = (Y3+Y4)/2 = 51.063			
10		n3 = 108	n4 = 111				
11	**Unweighted Mean**	Y13 = (Y1+Y3)/2 = 49.277	Y24 = (Y2+Y4)/2 = 51.108				
12							
13	**n:**	443					
14	**df:**	439					
15	**Main effect of Similarity (X):**	Y24 - Y13 = 51.108 – 49.277 = 1.831, F(1, 439) = 0.558, p = .455.					
16	**Main effect of Voice (Z):**	Y12 - Y34 = 51.063 – 49.321 = 1.742, F(1, 439) = 0.505, p = .478.					
17	**Interaction (XZ):**	(Y4 - Y3) - (Y2 - Y1) = (55.117 – 47.009) – (47.098 – 51.545) = 12.555, F(1, 439) = 6.563, p = .011.					
18							
19	***Moderation Analysis of a 2X2 Design Using a Main Effects (–0.5/0.5 Coding) and a Simple Effects (0/1 coding) Parameterization.***						
20							
21			**Main effect**		**Simple effect**		
22		**Coef.**	**S.E.**	**p**	**Coef.**	**S.E.**	**p**
23	**Constant (a)**	50.192	1.225	< 0.01	51.545	2.436	< 0.01
24	**Similarity (X) -> b1**	1.831 = Y24 - Y13	2.450	.455	–4.446 (when Z=0, Y2-Y1)	3.445	.198
25	**Narrative Voice (Z) -> b2**	1.742 = Y34 - Y12	2.450	.478	–4.535 (when X=0, Y3-Y1)	3.447	.193
26	**XZ -> b3**	12.555 = (Y4 - Y3) - (Y2 - Y1)	4.900	.011	12.555	4.900	.011
27	**Multiple R**		.131			.131	
28	**Conditional (simple) effects of Similarity**						
29		**Effect**	**SE**	**p**			
30	**Third-Person Voice**	–4.446 = Y2 - Y1	3.445	.198			
31	**First-Person Voice**	8.108 = Y4 - Y3	3.485	.020			
32							
33	Nota bene: simple effect parametrization implies that b1 and b2 represetn two of the four simple effects in the 2X2 ANOVA rather than the main effect of X and Z						

Figure 9.1: Similarity between a simple moderation analysis and a 2x2 ANOVA

9.3 The linear regression framework

Moderation analysis can be conducted by adding one (or multiple) interaction terms in a regression analysis. For example, if Z is a moderator for the relation between X and Y, we can fit a regression model:

$$Y = \beta_0 + \beta_1 * X + \beta_2 * Z + \beta_3 * XZ + \epsilon$$

$$Y = \beta_0 + \beta_2 * Z + (\beta_1 + \beta_3 * Z) * X + \epsilon$$

Thus, if β_3 is not equal to 0, the relationship between X and Y depends on the value of Z, which indicates a moderation effect.

If Z is a dichotomous/binary variable (e.g. gender) the above equation can be written as:

$$\beta_0 + \beta_1 * X + \epsilon$$

for male (Z=0)

$$\beta_0 + \beta_2 + (\beta_1 + \beta_3) * X + \epsilon$$

for female (Z=1)

9.4 Interpretation and centering

If X and/or moderator become significant, main effects are present. If the moderation term becomes significant, there is a moderation effect. The (possibly significant) influences of X and the moderator are then so-called "conditional" effects.

The value 0 usually has no meaningful meaning (e.g. in rating scales 1 to 5 there is no zero at all). Therefore, it is a good practice to centering means subtracting the overall mean from each value. The centering changes the interpretation decisively:

- if has the value 0, the moderator is moderately developed
- if has negative values, below-average characteristics are present
- if has positive values, above-average characteristics are present

The change in meaning must be taken into account in the interpretation:

- influence of the predictor on Y with an average expression of the moderator
- influence of the moderator on Y with an average expression of the predictor

9.5 Multicollinearity

The moderation term is formed with moderator and X. However, the moderator and X are also contained individually in the regression equation. This often leads to multicollinearity (i.e. low tolerances or high VIF values).

If moderator and X are centered, the symptoms of multicollinearity are superficially defused. However, the multicollinearity itself remains.

Reminder: Multicollinearity means that predictors are (too strongly) related to each other. Since the moderation term also consists of A (or B), it can easily correlate with A (or B).

9.6 Simple slopes

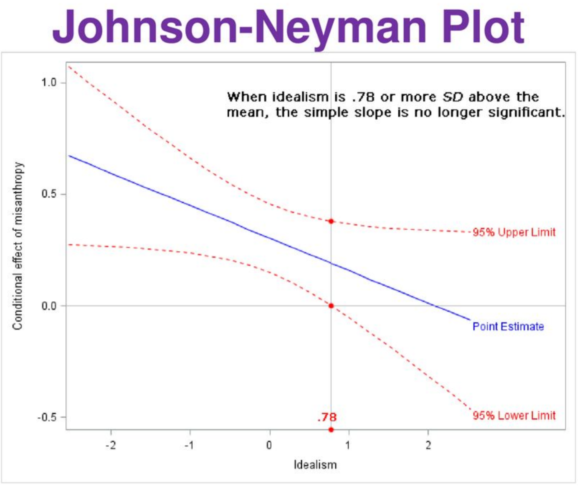
When the moderation effect becomes significant, it needs to be "illustrated" in order to make it interpretable. To do so, we can rely on simple slope analysis: comparison of the regression lines for low, medium and high levels of the moderator.

Typically, we use the mean value of the moderator, as well as the values + and - 1 SD are used, but theoretically any values can be considered.

9.7 Significance range according to Johnson & Neyman (only if moderation metric)

This method suggests to conduct comparison of the regression equation for many characteristics of the moderator to identify areas of significance.

It is more useful than simple slopes for metric moderators, since illustrations result in less loss of information in the moderator's levels. The effect (b) of the X on Y is now illustrated in the diagram as a function of the moderator (not to be confused with a regression line!), as well as the confidence interval for this effect:



9.8 Steps in the analysis

A moderation analysis typically consists of the following steps:

- compute the interaction term X*Z
- fit a multiple regression model with X, Z, and XZ as predictors
- test whether the regression coefficient for XZ is significant or not
- interpret the moderation effect
- display the moderation effect graphically.

9.9 Important remarks: variable's effect as a function of a moderator

A conceptual representation of a simple moderation model with a single moderating variable Z modifying the relationship between X and Y.

Let's assume that X and Z are either dichotomous or continuous and the outcome variable Y is a continuous dimension suitable for analysis with linear regression, we have the following equations:

$$\hat{Y} = a + b_1 * X + b_2 * Z + b_3 * XZ = a + (b_1 + b_3 * Z)X + b_2 * Z$$

In this representation, the weight for X is not a single number but, rather, a function of Z: $b_1 + b_3 * Z$. The output is sometimes called the simple slope of X or the conditional effect of X. The coefficients b_1 and b_2 may or may not have a substantive interpretation, depending on how X and Z are coded or, in the case of dichotomous X and Z, what two numbers are used to represent the groups in the data.

Important remarks:

- It is never correct to interpret b_1 as the effect of X on Y "controlling for" the effect of Z and XZ.
- It usually is not correct to interpret b_1 as the "average effect" of X or the "main effect" of X (a term from analysis of variance that applies only when X and W are categorical)
- b_1 can be the main effect of X from a 2X2 ANOVA if X and Z are coded appropriately.
- Similar arguments apply to the interpretation of b_2.

Correct interpretation:

- b_1 is the conditional effect of X on Y when Z = 0.
- b_1 is the estimated difference in Y between two cases in the data that differ by one unit in X but have a value of 0 for Z.
- b_2 is the conditional effect of Z on Y when X = 0.
- When X = 0, the conditional effect of Z reduces to b_2.

- When Z = 0, X's effect equals b_1; when Z = 1, X's effect equals $b_1 + b_3$; when Z = 2, X's effect is $b_1 + 2b_3$, and so forth.

9.10 Reporting the results

The following must be observed to report moderation analysis:

- always shows both (un)standardized effects
- in the case of a statistically significant interaction, the conditional effects (aka main effects) can no longer be interpreted unquestioningly
- in the case of a statistically significant interaction, the difference between several slopes in the model (e.g. those represented as simple slopes) is significant

9.11 Further resources

For a long time the PROCESS macro has been one of the best ways of testing moderations (interactions) when using SPSS. In December 2020, Hayes has released the PROCESS function for R.

To download Hayes' PROCESS macro for R go here.

A tutorial about how to use the process() function in R can be found here.

9.12 In a nutshell

A moderation analysis is a regression analysis in which the dependent variable (Y) is predicted from an independent variable (X) and a moderator (Z) and the interaction of both (XZ).

X and Z should be centered prior to analysis. On the one hand, this simplifies the interpretation of the results and, on the other hand, helps against the effects of multicollinearity in the model.

If the interaction term becomes significant, there is moderation.

The moderation effect can be illustrated by simple slopes and significance areas according to Johnson-Neyman.

9.13 How it works in R?

See the lecture slides on moderation analysis:

You can also download the PDF of the slides here:

9.14 Example from the literature

The following article relies explains moderation (and mediation) analysis with an illustrative example:

Igartua, J. J., & Hayes, A. F. (2021). Mediation, moderation, and conditional process analysis: Concepts, computations, and some common confusions. The Spanish Journal of Psychology, 24, e49. Available here.

Please reflect on the following questions:

- How is simple moderation analysis similar to a 2x2 ANOVA?
- What is the difference between "main effects" and "simple effects" models?
- What is the advantage of using Johnson-Neyman technique to illustrate conditional effects?

9.15 Quiz

True

False

Statement

Moderation occurs when the relationship between two variables (X and Y) changes as a function of a third variable (Z).

Moderation occurs When a third variable (Z) affects the relation between an independent variable (X) and the dependent variable (Y).

The simple main (conditional) effect is when the effect of X (or Z) on Y when Z (or X) is equal to 0.

Standardization of the variables is useful to mitigate multicollinearity.

View Results

My results will appear here

9.16 Time to practice on your own

You can download the PDF of the exercises here: PDF

In this exercise, we will use the data "protest.sav" (Hayes, 2022) which can be downloaded here under "data files and code". Especially, we will focus on the following variables:

- Protest (independent variable): A lawyer protests against gender discrimination (experimental group, dichotomous 0 = no and 1 = yes)

- Like (dependent variable): assessment of the lawyer (scale 1-7)
- Sexism (moderator): perception of sexism as a ubiquitous problem in society (scale 1-7)

9.16.1 Exercise 1: protest with a continuous moderator

We want to test the assumption that when women believe that sexism is a problem in society, they like the lawyer more when he protests sexism than when he doesn't protest.

Start by drawing the regression equations.

Solution: equation

The regression equation go as:

$$Y_i = \beta_0 + \beta_1 * Protest_i + \beta_2 * Sexismus_i + \beta_3 * (Protest * Sexismus_i) + \epsilon$$

Now, we want to know if the overall model is significant? Start by importing the data:

```
# load the data
library(foreign)
db <- read.spss(file=paste0(getwd(),
                "/data/protest.sav"),
                use.value.labels = F,
                to.data.frame = T)
```

Note that there are several ways to center the variables when creating the interaction term.

```
# interaction term
# without centering
db$ProtestXSexism1 = db$protest*db$sexism
# with centering
db$ProtestXSexism2 = (db$protest-mean(db$protest)) * (db$sexism-mean(db$sexism))
# z-standardization
# db$ProtestXSexism3 = scale(db$protest)*scale(db$sexism)
# view
head(db)
##   sexism liking respappr protest x Sexism_h_t ProtestXSexism1 ProtestXSexism2
## 1   4.25   4.50     5.75       0 4          1               0       0.5914260
## 2   4.62   6.83     5.75       0 6          1               0       0.3390229
## 3   4.62   4.83     5.25       0 4          1               0       0.3390229
## 4   4.37   4.83     4.25       0 5          1               0       0.5095655
## 5   4.25   5.50     2.50       0 3          1               0       0.5914260
## 6   4.00   6.83     4.75       0 3          1               0       0.7619686
```

Now, run the regression model:

```
# regression model (with centering)
m.cent = lm(liking ~ protest + sexism + ProtestXSexism2, data=db)
summary(m.cent)
##
## Call:
## lm(formula = liking ~ protest + sexism + ProtestXSexism2, data = db)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9894 -0.6381  0.0478  0.7404  2.3650
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.79659    0.58679   8.174 2.83e-13 ***
## protest          0.49262    0.18722   2.631  0.00958 **
## sexism           0.09613    0.11169   0.861  0.39102
## ProtestXSexism2  0.83355    0.24356   3.422  0.00084 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9888 on 125 degrees of freedom
## Multiple R-squared:  0.1335, Adjusted R-squared:  0.1127
## F-statistic: 6.419 on 3 and 125 DF,  p-value: 0.0004439

# regression model (without centering)
m.roh = lm(liking ~ protest + sexism + ProtestXSexism1, data=db)
summary(m.roh)
##
## Call:
## lm(formula = liking ~ protest + sexism + ProtestXSexism1, data = db)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9894 -0.6381  0.0478  0.7404  2.3650
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       7.7062     1.0449   7.375 1.99e-11 ***
## protest          -3.7727     1.2541  -3.008  0.00318 **
## sexism           -0.4725     0.2038  -2.318  0.02205 *
## ProtestXSexism1   0.8336     0.2436   3.422  0.00084 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9888 on 125 degrees of freedom
## Multiple R-squared:  0.1335, Adjusted R-squared:  0.1127
## F-statistic: 6.419 on 3 and 125 DF,  p-value: 0.0004439
```

How much variance does the model explain? Are there main effects or conditional effects? If yes, what do they look like?

> Solution: interpretation
>
> The overall model is significant (p < .001) and explains 13.3% of the variance. We are allowed to interpret the results of the regression.
> There is a significant conditional effect of protest when sexism = 0. It is not a main effect since interaction is also included.
> Protest has an effect on the assessment if the moderator (sexism) has the value zero (= a medium level, since mean centered)

Is there a moderation effect?

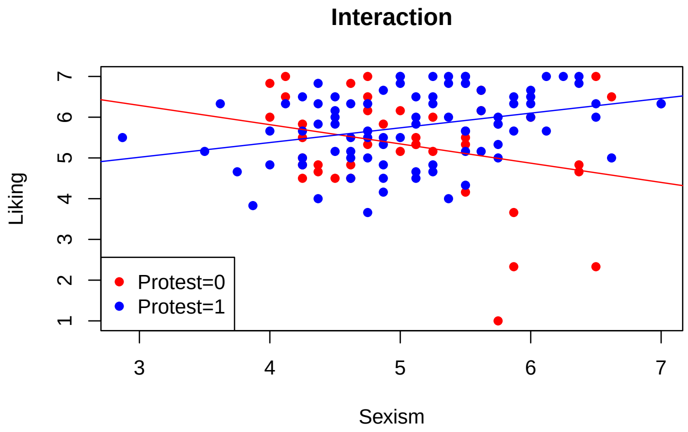
```
# run the model without the interaction term
m0 = lm(liking ~ protest + sexism, data=db)
# compare the R2
summary(m.cent)$r.squared - summary(m0)$r.squared
## [1] 0.08119242
# get EtaSq
DescTools::EtaSq(m.cent)
##                      eta.sq eta.sq.part
## protest         0.047991546 0.052478399
## sexism          0.005136019 0.005892326
## ProtestXSexism2 0.081192415 0.085672955
```

If so, how much variance does this explain and what does this effect mean in general?

> Solution: interpretation
>
> The interaction of protest and sexism perception is significant. There is a moderation effect.
> The effect of the protest on the lawyer's assessment varies depending on how strongly the subjects perceive sexism as a problem.
> The interaction contributes 8.1% to explaining the variance. The dependent variable is thus better explained if moderation is taken into account.

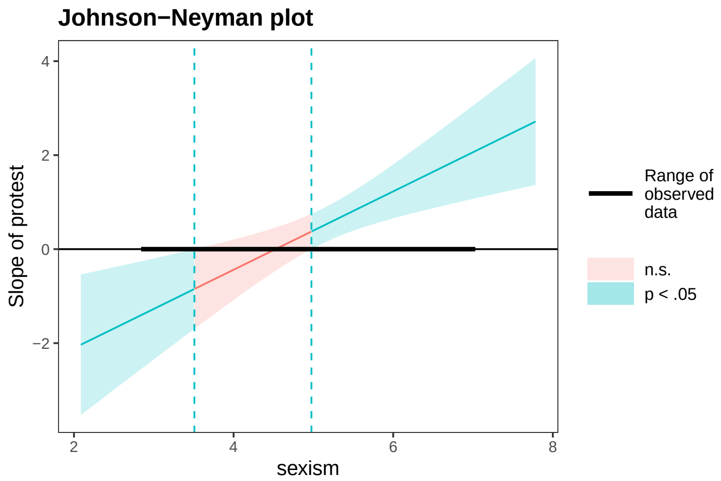
Illustrate the moderation effect graphically and interpret it. First, we can create an interaction plot:

```r
# extract the needed coefficients
intercept.p0 = coefficients(m.roh)[1]
intercept.p1 = coefficients(m.roh)[1] + coefficients(m.roh)[2]
slope.p0 = coefficients(m.roh)[3]
slope.p1 = coefficients(m.roh)[3] + coefficients(m.roh)[4]
# interaction plot
par(mfrow=c(1,1))
farben = c("red","blue")
plot(db$sexism, db$liking, main="Interaction",
     col=farben[db$protest+1],pch=16,
     xlab="Sexism",ylab="Liking")
abline(intercept.p0,slope.p0,col="red")
abline(intercept.p1,slope.p1,col="blue")
legend("bottomleft",
       c("Protest=0","Protest=1"),
       col=c("red","blue"),pch=16)
```



Second, we can provide a Johnson-Neyman plot:

```r
library(interactions)
m.simplified = lm(liking ~ protest*sexism, data=db)
johnson_neyman(m.simplified,"protest","sexism")
## JOHNSON-NEYMAN INTERVAL
##
## When sexism is OUTSIDE the interval [3.51, 4.98], the slope of protest is p < .05.
##
## Note: The range of observed values of sexism is [2.87, 7.00]
```



9.16.2 Exercise 2: protest with a dichotomous moderator

Now, divide the moderator into a dichotomous variable (sexism low vs. high) with a median split and recalculate the moderation analysis.

```
# Median split
db$sexism.ms = as.integer(db$sexism>=median(db$sexism))
```

What changes in the output?

```
# new model
m3 = lm(liking ~ protest*sexism.ms, data=db)
summary(m3)
##
## Call:
## lm(formula = liking ~ protest * sexism.ms, data = db)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9179 -0.6815  0.1263  0.7963  2.0821
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.6491     0.2125  26.584  < 2e-16 ***
## protest            -0.1154     0.2634  -0.438  0.66199
## sexism.ms          -0.7312     0.3122  -2.342  0.02074 *
## protest:sexism.ms   1.2090     0.3779   3.199  0.00175 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.9967 on 125 degrees of freedom
## Multiple R-squared:  0.1195, Adjusted R-squared:  0.09839
## F-statistic: 5.656 on 3 and 125 DF,  p-value: 0.001148
# get EtaSq
DescTools::EtaSq(m3)
##                        eta.sq eta.sq.part
## protest           0.043987025 0.047581194
## sexism.ms         0.002000014 0.002266368
## protest:sexism.ms 0.072096985 0.075686621
# get the coeff
meanvalues = tapply(db$liking, list(db$protest,db$sexism.ms),FUN=mean)
meanvalues
##          0        1
## 0 5.649091 4.917895
## 1 5.533659 6.011489
```

Solution: interpretation

R-square significantly worse (probably due to loss of information during dichotomization) and the coefficients change slightly.
There are only two Simple Slopes because there are only two moderator levels. There are no Johnson-Neyman values.

Calculate the moderation analysis again with the variable «x» as the independent variable (it measure the lawyer protests to varying degrees on a scale of 1-7) and the metric moderator. What changes in the output?

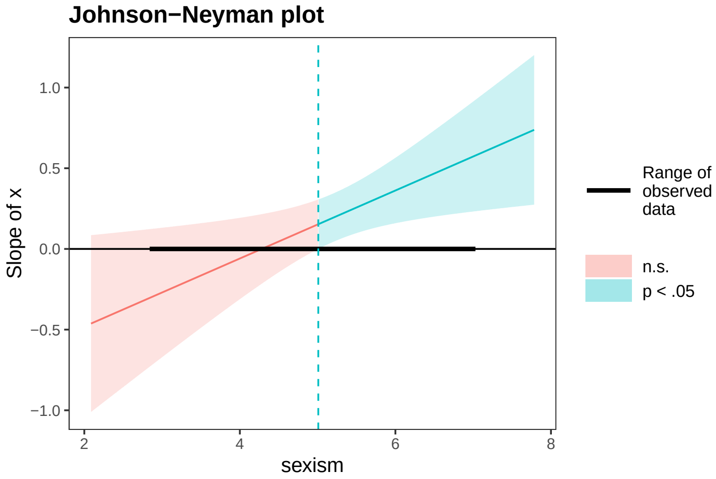
```
# new model
m4 = lm(liking ~ x*sexism, data=db)
summary(m4)
##
## Call:
## lm(formula = liking ~ x * sexism, data = db)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0451 -0.6128  0.1029  0.7720  1.6583
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  8.65856    1.90366   4.548 0.0000126 ***
## x           -0.90210    0.45129  -1.999    0.0478 *
## sexism      -0.73604    0.36256  -2.030    0.0445 *
## x:sexism     0.21069    0.08563   2.460    0.0152 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.009 on 125 degrees of freedom
## Multiple R-squared:  0.09858,    Adjusted R-squared:  0.07695
## F-statistic: 4.557 on 3 and 125 DF,  p-value: 0.004584
# get EtaSq
DescTools::EtaSq(m4)
##                 eta.sq eta.sq.part
## x         0.046574596  0.04912969
## sexism    0.006844541  0.00753586
## x:sexism  0.043654300  0.04619148
```

Solution: interpretation

Now the hypothesis is that the more women believe that sexism is a problem, the more they like the lawyer, the more she protests.
The coefficients change, the explanation of variance overall and through the interaction alone is weaker in each case (but this is simply because it is a completely different and only a simulated variable).

What changes in the graphics?

```
# plots
johnson_neyman(m4,"x","sexism")
## JOHNSON-NEYMAN INTERVAL
##
## When sexism is OUTSIDE the interval [0.20, 5.01], the slope of x is p < .05.
##
## Note: The range of observed values of sexism is [2.87, 7.00]
```



ℹ Solution: interpretation

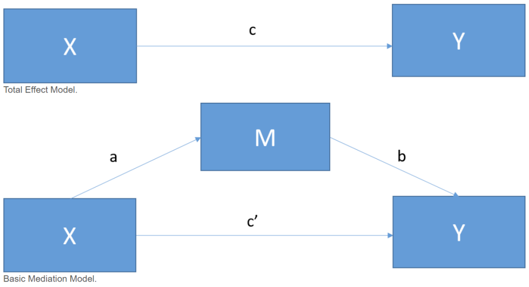
The Johnson-Neyman graph hardly changes. The range of significance shifts only slightly.

Chapter 10

Mediation analysis

10.1 Mediation analsyis

Mediation analysis tests a hypothetical causal chain where one variable X affects a second variable M and, in turn, that variable affects a third variable Y. Mediators describe the how or why of a relationship between two other variables. Mediators describe the process through which an effect occurs. This is also sometimes called an indirect effect. We therefore have the following properties:



- Total effect of X on Y: c = c' + ab
- Indirect effect of X on Y: ab
- Direct effect of X on Y after controlling for M: c' = c - ab

Perfect mediation occurs when the effect of X on Y decreases to 0 with M in the model.

Partial mediation occurs when the effect of X on Y decreases by a nontrivial amount (the actual amount is up for debate) with M in the model.

10.2 Investigation of the direct and indirect effects

Mediation was previously considered statistically significant if a and b were significant and c' was less than c. One calculated c-c' (difference method) and regarded a mediation as statistically significant if c was significant and c' was not. The mediation was thus validated by a logical conclusion without considering a*b (the indirect effect) specifically.

However, this approach might not be suitable for the following cases:

- c just significant, c' just no longer significant: Would be explained as statistically significant mediation using the difference method, but mediation explains too little.
- c significant, c' still significant: Would be explained as insignificant mediation using the difference method, but mediation explains enough, but not all mediation processes were found (therefore c' still becomes significant).
- c is not significant: Would be explained as insignificant mediation using the difference method, but mediation effect is possible if indirect and direct effects are in opposite directions.

10.3 Significance

In order to be able to speak of mediation, the following must be statistically reliable: ab not equal 0 (i.e. a not equal to 0 and b not equal to 0!).

To determine the significance of the indirect effect, the indirect effect is backed up statistically using the so-called Sobel test. Due to unrealistic requirements, however, it is no longer recommended.

Instead, a bootstrap method (also "resampling method") is preferred. Empirical sample is considered as a pseudo-population, and a large number of random samples (each with replacement) are drawn from this pseudo-population (1000 or more). The indirect effect is calculated in each case. The result is a distribution of the indirect effects (which are normally distributed!)

10.4 Two approaches

We will cover two ways to conduct mediation analysis:

- Baron & Kenny's (1986) 4-step indirect effect method
- mediation package (Tingley et al., 2014)

10.4.1 Method with 4-steps from Baron & Kenny

This method includes 4-steps:

- estimate the relationship between X on Y (c must be significantly different from 0)
- estimate the relationship between X on M (a must be significantly different from 0)
- estimate the relationship between M on Y controlling for X (b must be significantly different from 0)
- estimate the relationship between Y on X controlling for M (should be non-significant and nearly 0)

10.4.2 Method with the R mediation package

This package uses the more recent bootstrapping method of Preacher & Hayes (2004) to address the power limitations of the Sobel test. This method computes the point estimate of the indirect effect (ab) over a large number of random sample (typically 1000) so it does not assume that the data are normally distributed and is especially more suitable for small sample sizes than the Barron & Kenny method.

This method includes 2-steps:

- estimate the relationship between X on M
- estimate the relationship between X on Y controlling for M

The mediate function gives us our Average Causal Mediation Effects (ACME), our Average Direct Effects (ADE), our combined indirect and direct effects (Total Effect), and the ratio of these estimates (Prop. Mediated). The ACME here is the indirect effect of M (total effect - direct effect) and thus this value tells us if our mediation effect is significant.

10.5 Inconsistent mediation

In general, complete mediation implies that these step 1-2-3-4 (in the 4-steps from Baron & Kenny) be met. However, in practice, some researchers also consider that the essential steps in establishing mediation are steps 2 (X on M) and 3 (Z on Y, controlling for X).

Furthermore, if c (direct effect of X on Y after controlling for M) had the opposite sign to that of ab (indirect effect of X on Y), then there would still be mediation (even if step 1 could not be met). In this case the mediator acts like a suppressor variable (see MacKinnon, Fairchild & Fritz, 2007).

For instance, we could have the following model: participation in strikes (or social demonstrations) and satisfaction with government mediated by media consumption. Presumably, we have the following relations:

- the direct effect is negative, with more participation in (anti-government) demonstration(s) suggesting less satisfaction with government

- however, likely the effect of participation in demonstration(s) on consuming media is positive
- furthermore, the effect of media consumption on satisfaction with government is likely to be positive

Thus, we obtain a positive indirect effect, and the total effect of participation in demonstration(s) on satisfaction with government is likely to be very small because the direct and indirect effects will tend to cancel each other out.

10.6 Sobel test

Given the above example, it is recommended to perform a single test of ab (rather than two separated tests of a and b). The test was first proposed by Sobel (1982). The Sobel test uses the following standard error estimate of ab:

$$\sqrt{b^2 * s_a^2 + a^2 * s_b^2}$$

The test of the indirect effect is given by dividing ab by the above standard error estimate and treating the ratio as a Z test:

$$z = \frac{ab}{\sqrt{b^2 * s_a^2 + a^2 * s_b^2}}$$

.

If a z-score is larger than 1.96 in absolute value, the mediation effect is significant at the .05 level.

10.7 Bootstraping

The Sobel test is easy to conduct but presumes that a and b are independent (which may not always be true). Furthermore, it assumes that ab is normally distributed (which might not work well for small sample sizes).

One solution is to use the bootstrap method (see Bollen & Stine, 1990). This method has no distribution assumption on the indirect effect ab, but rather approximates the distribution of ab using its bootstrap distribution.

Using the original data set (Sample size = n) as the population, the model draws a bootstrap sample of n individuals with paired (Y, X, M) scores randomly from the data set with replacement. From the bootstrap sample, estimate ab based on a set of regression models. The steps are repeated for a total of N times (N=number of bootstraps).

10.8 Total mediation versus partial mediation

One speaks of total mediation when the direct effect disappears as a result of mediation: c != 0, and thus ab = c.

One speaks of partial mediation when the direct effect does not disappear as a result of the mediation, but a residue remains: c!= 0 (i.e. if ab != c).

In the social science context, mediations are mostly partial, because one process rarely explains the full influence of X on Y.

10.9 Why consider ab instead of c-c'?

The difference method (c-c') originally assumed that c must be significant (that is, that the X must have a total effect on the Y, which is to be explained by mediation).

However, mediation can also exist if c = 0. This is possible because there may be several significant mediator effects that cancel each other out (but not all may have been measured).

The targeted examination of ab thus provides a more precise picture of individual significant mediation processes, even if there is no total effect.

10.10 More recent methodology: lavaan

The R package that handles measurement and analysis models most easily is lavaan. For the calculation, the model formula and the data are passed to the sem() function.

The result is a complex object with many methods of its own. The most important are: summary(), fitMeasures(), parameterEstimates(), and inspect().

10.11 In a nutshell

A mediation analysis is a regression analysis in which the influence of an independent variable (X) on a dependent variable Y is fully or partially explained by a mediator (M).

The indirect effect via the mediator is formed by multiplying two effects: effect of X on M and effect of M on Y. The significance of the indirect effect is determined by bootstrapping.

Significant mediation occurs when the confidence interval for the indirect effect is not 0.

The size of the effect can be described by the unstandardized or standardized coefficients.

A mediation can exist even if there is no significant total effect.

10.12 How it works in R?

See the lecture slides on mediation analysis:

You can also download the PDF of the slides here:

10.13 Quiz

True

False

Statement

Perfect mediation occurs when the relationship between the predictor (X) and the outcome (Y) is completely wiped out when the mediator (M) is included in the model.

To measure the effect size of mediation, we rely on the ratio of the indirect to total effect; mediator accounts for % of shared variance.

The Sobel test works well for small sample sizes.

To test for a mediation effect, we assess the size of the indirect effect and ensure that the confidence interval does not contains 0.

View Results

My results will appear here

10.14 Example from the literature

The following article relies on moderated mediation as a method of analysis:

Matthes, J. (2013). Do hostile opinion environments harm political participation? The moderating role of generalized social trust. International Journal of Public Opinion Research, 25(1), 23-42. Available here.

Please reflect on the following questions:

- What is the research question of the study?
- What are the research hypotheses?
- Is moderated mediation an appropriate method of analysis to answer the research question?
- What are the main findings of the moderated mediation analysis?

10.15 Time to practice on your own

You can download the PDF of the exercises here:

10.15.1 Exercise 1: Mediation analysis (using lavaan)

In this exercise, we will use the data "protest.sav" (Hayes, 2022) which can be downloaded here under "data files and code". Especially, we will focus on the following variables:

- Protest (independent variable): A lawyer protests against gender discrimination (experimental group, dichotomous 0 = no and 1 = yes)
- Respappr (mediator): Perceived adequacy of response (scale 1-7)
- Like (dependent variable): assessment of the lawyer (scale 1-7)

Hypothesis: If the lawyer protests, her reaction will be perceived as more appropriate, and therefore the lawyer will be evaluated more favorably.

Start by drawing the regression equations.

ℹ Solution: equation

The regression equations go as:

$$Y_i = \beta_0 + \beta_1 * Protest_i + \beta_2 * Respappr_i + \epsilon_i = c' + b$$

$$M_i = \beta_0 + \beta_1 * Protest_i + \epsilon_i = a$$

$$Y_i = \beta_0 + \beta_1 * Protest_i + \epsilon_i = c = c' + ab$$

Now, we want to calculate the mediation model and to answer the following questions:

- Is the a-path significant? If so, how much variance does it explain?
- Are the b-path and the c'-path significant? If so, how much variance do they explain together?
- Is there a total effect of X on Y?
- Is there a significant mediation effect? If so, is there partial or total mediation?

Start by loading and selecting the data:

```r
# load the data
library(foreign)
db <- read.spss(file=paste0(getwd(),
                "/data/protest.sav"),
                use.value.labels = F,
                to.data.frame = T)
# get the data
sel <- db |>
  dplyr::select(protest, respappr, liking) |>
  stats::na.omit()
```

Now, define the model in laavan and give the results for the c path:

```
# define the model
modell.c = "liking ~ protest"
# get the complete output
fit.c = lavaan::sem(modell.c, data=sel)
lavaan::parameterestimates(fit.c, standardized=T)[1:8]
##       lhs op     rhs   est    se     z pvalue ci.lower
## 1  liking  ~ protest 0.479 0.193 2.478  0.013    0.100
## 2  liking ~~  liking 1.044 0.130 8.031  0.000    0.789
## 3 protest ~~ protest 0.217 0.000    NA     NA    0.217
lavaan::inspect(fit.c,"r2")
## liking
##  0.045
lavaan::fitMeasures(fit.c)[c("chisq","df","aic","cfi","rmsea")]
##    chisq       df      aic      cfi    rmsea
##   0.0000   0.0000 375.5983   1.0000   0.0000
```

ℹ Solution: Interpretation

The C path is significant and explains 4.5% of the variance of liking. Linear regression makes sense for this data.

Now get the model for the c' path:

```
# define the model
modell.mediation = "
## Direct effect
liking ~ protest
## mediation path
respappr ~ protest
liking ~ respappr
"
# get the complete output
fit.med = lavaan::sem(modell.mediation, data=sel)
lavaan::parameterestimates(fit.med, standardized=T)[1:8]
##        lhs op      rhs    est    se      z pvalue ci.lower
## 1   liking  ~  protest -0.101 0.198 -0.508  0.611   -0.489
## 2 respappr  ~  protest  1.440 0.220  6.544  0.000    1.008
## 3   liking  ~ respappr  0.402 0.069  5.857  0.000    0.268
## 4   liking ~~   liking  0.824 0.103  8.031  0.000    0.623
## 5 respappr ~~ respappr  1.354 0.169  8.031  0.000    1.024
## 6  protest ~~  protest  0.217 0.000     NA     NA    0.217
lavaan::inspect(fit.med,"r2")
##   liking respappr
##    0.246    0.249
lavaan::fitMeasures(fit.med)[c("chisq","df","aic","cfi","rmsea")]
##   chisq      df     aic     cfi   rmsea
##   0.000   0.000 756.355   1.000   0.000
```

Solution: Interpretation

The c' path is not significant, but both a and b are significant and the variance explained by liking is significantly higher (24.6%) than in the last model.

The output is a little better understandable if you take the indirect and the overall effect on Y with the names of the paths:

```
# define the model
modell.complete = "
## direct effect
liking ~ c*protest
## mediation path
respappr ~ a*protest
liking ~ b*respappr
## indirect effect (a*b)
ab := a*b
## total effect (c+a*b)
total := c+a*b
"
# get the complete output
fit.complete = lavaan::sem(modell.complete, data=sel)
lavaan::parameterestimates(fit.complete, standardized=T)[1:8]
##        lhs op      rhs label    est    se      z pvalue
## 1   liking  ~  protest     c -0.101 0.198 -0.508  0.611
## 2 respappr  ~  protest     a  1.440 0.220  6.544  0.000
## 3   liking  ~ respappr     b  0.402 0.069  5.857  0.000
## 4   liking ~~   liking        0.824 0.103  8.031  0.000
## 5 respappr ~~ respappr        1.354 0.169  8.031  0.000
## 6  protest ~~  protest        0.217 0.000     NA     NA
## 7       ab :=      a*b    ab  0.579 0.133  4.364  0.000
## 8    total :=    c+a*b total  0.479 0.193  2.478  0.013
lavaan::inspect(fit.complete,"r2")
##   liking respappr
##    0.246    0.249
```

```
lavaan::fitMeasures(fit.complete)[c("chisq","df","aic","cfi","rmsea")]
##   chisq      df     aic     cfi   rmsea
##   0.000   0.000 756.355   1.000   0.000
```

Give the full interpretation of the model:

Solution: Interpretation

- The indirect effect (ab) of X via the mediator on Y is significant and positively directed.
- If the lawyer protests, her reaction will be perceived as more appropriate than if she does not protest. An appropriate perception of the reaction then leads to a better assessment of the lawyer.
- The direct effect (c), however, is not significant. Since the indirect effect is significant at the same time, there is complete mediation.
- The connection between the lawyer's protesting and the lawyer's evaluation is fully mediated by the lawyer's reaction being perceived as more appropriate when she protests.

10.15.2 Exercise 2: Moderated mediation analysis (using lavaan)

In this exercise, we will use the data "protest.sav" (Hayes, 2022) which can be downloaded here under "data files and code". Especially, we will focus on the following variables:

- Protest (independent variable): A lawyer protests against gender discrimination (experimental group, dichotomous 0 = no and 1 = yes)
- Respappr (mediator): Perceived adequacy of response (scale 1-7)
- Like (dependent variable): assessment of the lawyer (scale 1-7)
- Sexism (moderator): perception of sexism as a ubiquitous problem in society (scale 1-7)

We want to test the assumption that if the lawyer protests, her response will be judged more appropriate by women who perceive sexism as a problem (moderator: dichotomous variable "sexism"), and therefore the lawyer will be judged better.

We want to test the following hypothesis: If the lawyer protests against gender discrimination, her response is perceived as more appropriate and therefore the lawyer is judged better.

Start by drawing the regression equations.

Solution: equation

The regression equations go as:

$$Y_i = \beta_0 + \beta_1 * Protest_i + \beta_2 * Respappr_i + \epsilon_i = c' + b$$

$$M_i = \beta_0 + \beta_1 * Protest_i + \beta_2 * Sexism_i + \beta_3 * (Protest_i * Sexism_i)\epsilon_i = a$$

$$Y_i = \beta_0 + \beta_1 * Protest_i + \epsilon_i = c = c' + ab$$

Now, we want to calculate the mediation model and to answer to following questions:

- Is the a-path moderated? How much mediator variance does regression explain for the overall a-path and how much of that is explained by moderation? Illustrate the moderation of the a-path. What do the results mean in terms of content?
- Are the b-path and the c'-path significant? If so, how much variance do they explain together?

Start by loading and selecting the data, and also construct the interaction variable:

```r
# load the data
library(foreign)
db <- read.spss(file=paste0(getwd(),
                "/data/protest.sav"),
                use.value.labels = F,
                to.data.frame = T)
# get the data
sel <- db |>
  dplyr::select(protest, respappr, liking, sexism) |>
  stats::na.omit()
# construct the interaction variable
sel$protest.sexism = sel$protest*sel$sexism
```

Now, define the model in laavan:

```r
# define the model
modell.mod = "
  liking ~ c*protest ## direct effect
  respappr ~ a*protest + sexism + protest.sexism ## moderation/mediation paths
  liking ~ b*respappr
  protest ~~ sexism ## covariances
  protest ~~ protest.sexism
  sexism ~~ protest.sexism
  respappr ~1 ## Intercepts
  liking ~1
  ab := a*b ## indirect effect
  total := c+a*b ## total effect
"
# get the complete output
fit.mod = lavaan::sem(modell.mod, data=sel)
lavaan::parameterestimates(fit.mod, standardized=T)[1:8]
##               lhs op            rhs label    est    se      z pvalue
## 1          liking  ~        protest     c -0.101 0.198 -0.508  0.611
## 2        respappr  ~        protest     a -2.687 1.429 -1.880  0.060
## 3        respappr  ~         sexism       -0.529 0.232 -2.278  0.023
## 4        respappr  ~ protest.sexism        0.810 0.278  2.919  0.004
## 5          liking  ~       respappr     b  0.402 0.069  5.857  0.000
## 6         protest ~~         sexism        0.015 0.032  0.456  0.648
## 7         protest ~~ protest.sexism        1.114 0.141  7.886  0.000
## 8          sexism ~~ protest.sexism        0.501 0.176  2.846  0.004
## 9        respappr ~1                       6.567 1.191  5.516  0.000
## 10         liking ~1                       3.747 0.302 12.400  0.000
## 11         liking ~~         liking        0.824 0.103  8.031  0.000
## 12       respappr ~~       respappr        1.269 0.158  8.031  0.000
## 13        protest ~~        protest        0.217 0.027  8.031  0.000
## 14         sexism ~~         sexism        0.610 0.076  8.031  0.000
## 15 protest.sexism ~~ protest.sexism        6.151 0.766  8.031  0.000
## 16        protest ~1                       0.682 0.041 16.640  0.000
## 17         sexism ~1                       5.117 0.069 74.441  0.000
## 18 protest.sexism ~1                       3.505 0.218 16.053  0.000
## 19             ab :=            a*b    ab -1.081 0.604 -1.790  0.073
## 20          total :=          c+a*b total -1.182 0.664 -1.781  0.075
lavaan::inspect(fit.mod,"r2")
##   liking respappr
##    0.246    0.296
lavaan::fitMeasures(fit.mod)[c("chisq","df","aic","cfi","rmsea")]
##        chisq           df          aic          cfi        rmsea
##    6.5581814    2.0000000 1345.5708600    0.9921174    0.1329187
```

What do the results mean in terms of content?

Solution: Interpretation

The overall model for the mediator (respappr) significantly explains 29.6% of the mediator's variance.
There is a significant conditional effect of the independent variable on the mediator when the moderator has a value of 0 (= with a moderate level of sexism, the lawyer's protest leads to her reaction being perceived as more appropriate).

The a-path from the independent variable to the mediator is significantly moderated by the sexism attitude. This moderation alone explains 4.8% (0.294-0.246=0.048) of the variance of the mediator.
The extent to which protesting influences the perceived appropriateness of the reaction thus varies depending on the subjects' sexism attitude.
However, the strong correlations between the independent variables are a problem for the model! The standardized coefficients are outside the natural limits of -1 to +1 and RMSEA has risen above 0.1. The model represents the data poorly. Moderation is better studied independently of mediation.

Finally, we want to know whether there is a significant moderated mediation effect? If so, how can this be described and interpreted in terms of content?

ℹ Solution: Interpretation

Interaction is significant: The moderator has an influence on the a-path.
Path ab is not significant: There is not an indirect effect for all values of the moderator.
RMSEA is too high and there are beta values above 1.0: The data are not suitable for this evaluation and the estimators cannot be fully trusted.

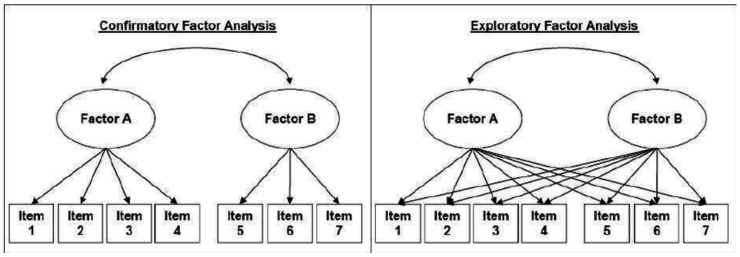
Chapter 11

EFA

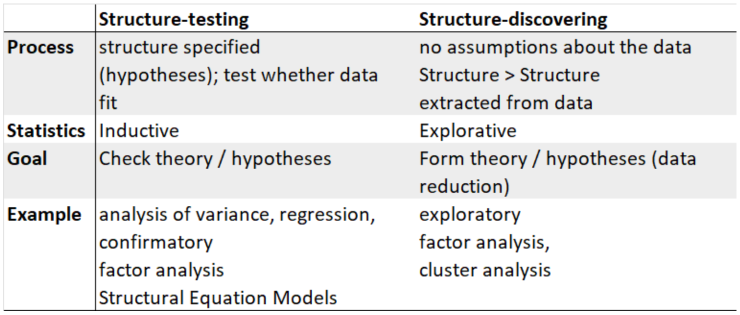
11.1 Factor analysis: exploratory vs confirmatory

In exploratory factor analysis (EFA), all measured variables are related to every latent variable. It is used to reduce data to a smaller set of summary variables and to explore the underlying theoretical structure of the phenomena. It asks what factors are given in observed data and, thereby, requires interpretation of usefulness of a model (it should be confirmed with confirmatory factor analysis).

In confirmatory factor analysis (CFA), researchers can specify the number of factors required in the data and which measured variable is related to which latent variable. It asks how well a proposed model fits a given data. However, it does not give a definitive answer, but is rather useful to compare models (and data).



The differences in both approaches can be summarized as follows:



	Structure-testing	Structure-discovering
Process	structure specified (hypotheses); test whether data fit	no assumptions about the data Structure > Structure extracted from data
Statistics	Inductive	Explorative
Goal	Check theory / hypotheses	Form theory / hypotheses (data reduction)
Example	analysis of variance, regression, confirmatory factor analysis Structural Equation Models	exploratory factor analysis, cluster analysis

11.1.1 Prerequisites

Prerequisites of factor analysis include:

- Condition: There are several interval-scaled characteristics (items).
- Rule of thumb: At least 50 people and 3x more people as variables (ideally: 5x more people as variables!).

11.1.2 Dimensionality reduction

Dimensionality reduction transforms a data set from a high-dimensional space into a low-dimensional space, and can be a good choice when you suspect there are too many variables which can be a problem because it is difficult to understand (or visualize) data in higher dimensions.

Another potential consequence of having a multitude of predictors is possible harm to a model. The simplest example is a method like ordinary linear regression where the number of predictors should be less than the number of data points used to fit the model.

Another issue is multicollinearity, where between-predictor correlations can negatively impact the mathematical operations used to estimate a model. If there are an extremely large number of predictors, it is fairly unlikely that there are an equal number of real underlying effects. Predictors may be measuring the same latent effect(s), and thus such predictors will be highly correlated.

There are several dimensionality reduction methods that can be used with different types of data for different requirements:

- combinating features:
  - linear: Principal component analysis (PCA), Factor analysis (FA), Multiple correspondence analysis (MCA), Linear discriminant analysis (LDA) or Singular value decomposition (SVD)
  - non-linear: Kernel PCA, t-distributed Stochastic Neighbor Embedding (t-SNE) or Multidimensional scaling (MDS)
- keeping most important features: Random forests, Forward or Backward selection

11.2 General procedure of EFA

Several steps need to be undertaken to conduct exploratorty factor analysis:

- Setup and evaluate data set
- Choose number of factors to extract
- Extract (and rotate) factors
- Evaluate what you have and possibly repeat the second and third steps
- Interpret and write-up results

There are also general guidelines to follow:

- It is better to select only the variables of interest from the data set.
- As factor analysis (and PCA) does not play well with missing data, it is better to remove cases that have missing data.
- Remember that factor analysis is designed for continuous data, although it is possible to include categorical data in a factor analysis.

11.3 Suitability of the data

Only relevant items may be included in the factor analysis. For instance, items must correlate significantly. Here, Bartlett Test is useful to assess the hypothesis that the sample came from a population in which the variables are uncorrelated. It checks whether the correlation matrix is an identity matrix or not:

- H0: The variables are uncorrelated in the population.
- H1: The variables are correlated in the population.

Kaiser-Meyer-Olkin criterion (KMO) and Measure of Sampling Adequacy (MSA) further test to what extent the variance of one variable is explained by the other variables. This indicates whether a data set is suitable for a factor analysis:

- value range 0-1
- values from .8 desirable
- values below .5 unacceptable

11.4 PCA versus EFA

In Principal Component Analysis (PCA) each variable can be fully explained by a linear combination of r factors:

$$Z_{ij} = f_{i1}a_{1j} + f_{i2}a_{2j} + ... + f_{ip}a_{pj}$$

where z is the standard variable x.

This approach should be used when data set structuring and data reduction is the primary goal.

In Exploratory Factor Analysis (EFA) each variable cannot be fully explained by a linear combination of r factors: e.g

$$Z_{ij} = f_{i1}a_{1j} + f_{i2}a_{2j} + ... + f_{ip}a_{pj} + e_j$$

This apporach should be used when trying to identify latent variables that are crucial to answering the items.

11.5 Spatial representation

Each of the given vectors (items) can be written exactly using the basis vectors (factors). The angle between the base vector (factor) and a given vector (item) is the correlation coefficient between item and factor: the factor loading.

Fundamental theorem: every observed value of a variable x_j can be described as a linear combination of several (hypothetical) factors (f_{jn}). That means that the answer to an item can be traced back to the sum of the factor values (f_{jn}), which are weighted with the factor loadings (a_{jn}).

11.6 Extraction of the factors

To extract the factors, we can z-standardize all variables (mean = 0, standard deviation = 1, variance = 1). The factor that explains the variance of all variables (x_1 to x_j) at most is searched for step by step. Then - similar to multiple regression - a linear combination of the items is formed:

$$F_1 = b_{11}x_1 + b_{12}x_2 + ... + b_{1j}x_j$$

To measure the explanation of the variance of a variable, the coefficient of determination represents the square of the factor loading (see R^2 in the regression). The first factor is the z-standardized variable for which the sum of the determination coefficients is maximum.

11.7 Decision about the number of factors

There are 4 decision aids to decide about the number of factors:

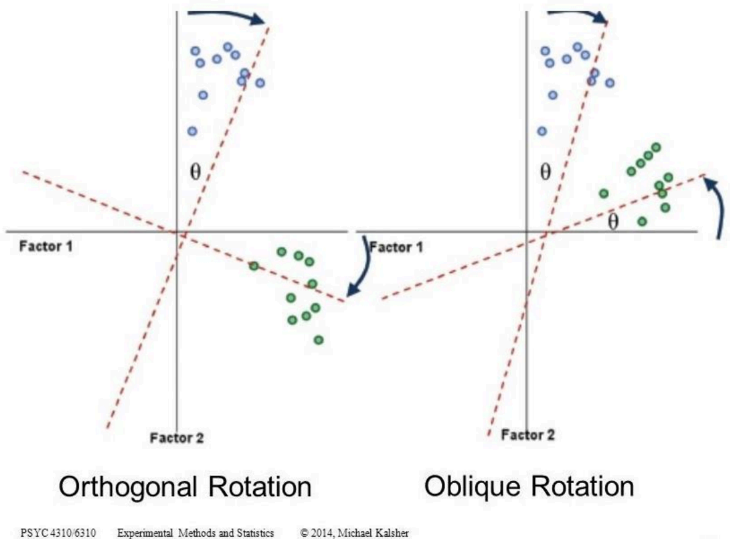
- Kaiser criterion: significant factor explains more variance than any of the original variables and this criterion is fulfilled from an eigenvalue of 1 onwards
- Scree plot: eigenvalues are plotted in the graphic and the factors that lie above a “threshold” are extracted
- Content plausibility: so many factors are accepted that a plausible interpretation results

- A priori criterion: it is theoretically determined in advance how many factors there should be

11.8 Factor rotation

By rotation one obtains simple structure, the factor interpretation is relieved. There are 2(3) types of rotation:

- Varimax (Right Angle): the factor axes remain at right angles, so that they are uncorrelated
- Oblique: the factor axes do not remain at right angles, so that they are correlated
- Combination (first right angles and then oblique)



PSYC 4310/6310 Experimental Methods and Statistics © 2014, Michael Kalsher

Which form of factor rotation is chosen in a specific case often depends on the theory behind it:

- In the case of orthogonal rotations, the results are easier to interpret because the factors are uncorrelated.
- Orthogonal rotations are usually also appropriate for pure dimension reduction.

However, orthogonal models often do not do justice to the underlying relationships:

- The assumption of zero correlation between the factors is often too strict and does not reflect the complexity of the data.
- If highly correlated factors are assumed in the data, an oblique rotation appears to make more sense than an orthogonal one.

11.9 Interpretation and naming

Factor interpretation is based on the factor loadings. Loadings from .5 are usually interpreted. Ideally, variables load exactly high on one factor and low on another.

For factor naming, variables with higher loadings should be given more consideration. A negative charge does not mean that the item does not belong to the factor. However, the sign must be taken into account in the interpretation.

Possible problems:

- Loading several variables on several factors poses interpretation difficulties
- Negative factor, or one-item factor

11.10 Lexicon

- Factor: latent dimension "behind" the variables, which is responsible for the manifestation of a directly measured variable
- Component: Collection of variables that have things in common
- Indicator (Item): a directly measured variable that, together with other variables, makes up a component/factor
- Factor loading: Correlation between a variable and a factor (should be $>$ .5 on one factor and preferably $<$.3 on other factors)
- Squared factor loadings: indicates the proportion of variance (=determination coefficient) of the variable that is explained by the factor
- Eigenvalue of the factor: proportion of the variance of all directly measured variables that is explained by a factor (=sum of the squared factor loadings [column by column]). For a factor to be relevant enough to be extracted, its eigenvalue should be > 1
- Commonality of a variable: proportion of the variance of an item that is explained by all factors (= sum of the squared factor loadings [row by row])
- Factor value: calculated value based on the observed values on the measured variables and the factor loading of the variable on the factor
- Rotation: optimization of the factor solution, perpendicular (orthogonal) vs oblique rotation

11.11 How it works in R?

See the lecture slides exploratory factor analysis:

You can also download the PDF of the slides here: PDF

11.12 Quiz

True

False

Statement

The larger the model chi-square test statistic, the larger the residual covariance.

Both in EFA and CFA we specify the pattern of indicator-factor loadings.

In CFA measurement error of indicators is removed during the estimation.

Kaiser's criterion and scree plot are alternative methods for determining how many factors to retain.

View Results

My results will appear here

11.13 Example from the literature

The following article relies on EFA as a method of analysis:

Schulz, A., Müller, P., Schemer, C., Wirz, D. S., Wettstein, M., & Wirth, W. (2018). Measuring populist attitudes on three dimensions. International Journal of Public Opinion Research, 30(2), 316-326. Available here.

Please reflect on the following questions:

- What is the research question of the study?
- What are the research hypotheses?
- Is EFA an appropriate method of analysis to answer the research question?
- What are the main findings of the EFA?

11.14 Time to practice on your own

You can download the PDF of the EFA exercises here:

11.14.1 Exercise 1: Big-5

To illustrate EFA, let us use the International Personality Item Pool data available in the psych package. It includes 25 personality self report items following the big 5 personality structure.

The first step is to test if the dataset is suitable for conducting factor analysis. To do so, run the Bartlett's Test of Sphericity and the Kaiser Meyer Olkin (KMO) measure.

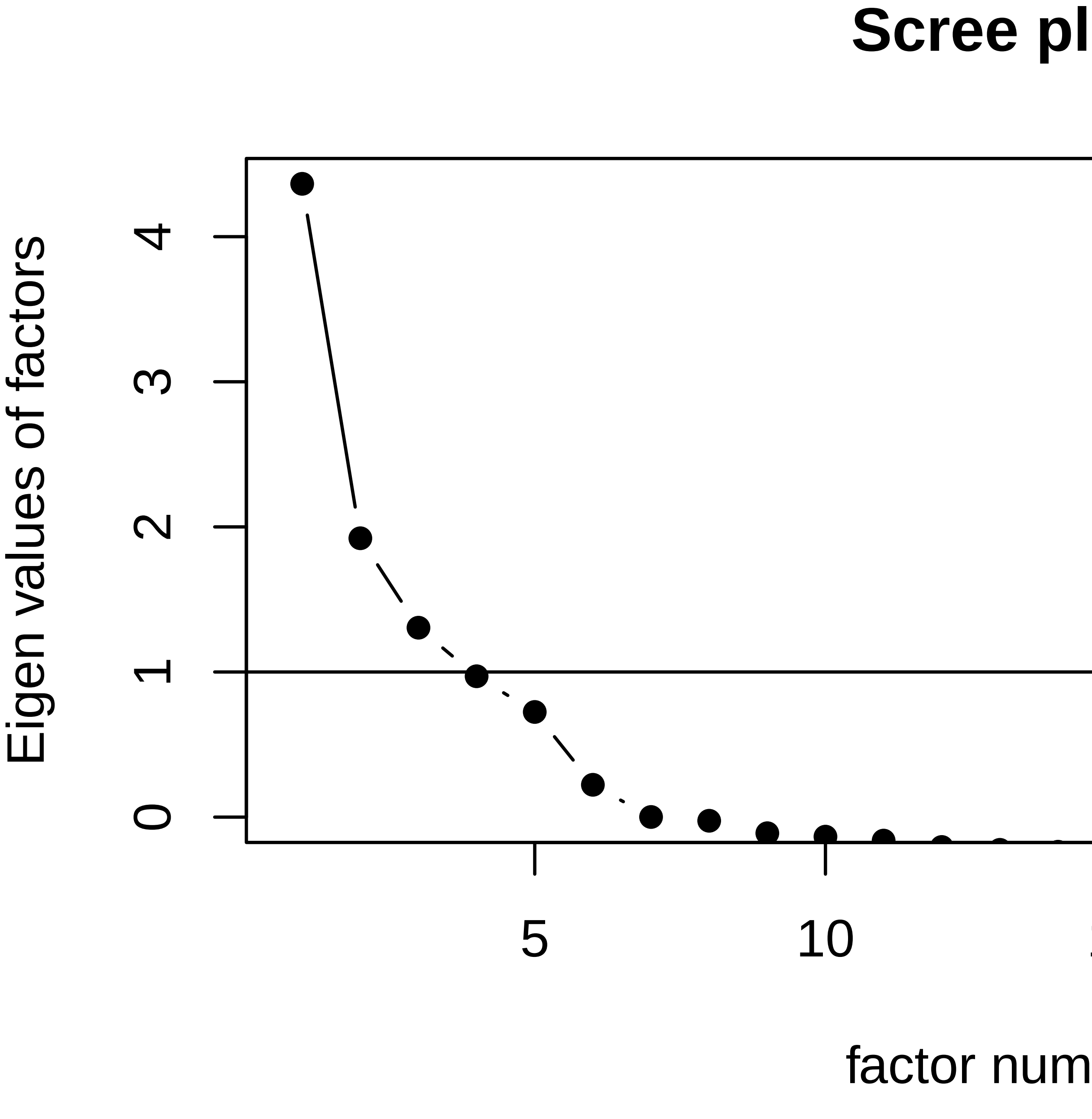
Reminder: Bartlett's Test of Sphericity tests whether a matrix (of correlations) is significantly different from an identity matrix (probability that the correlation matrix has significant correlations among at least some of the variables in a dataset). KMO measure indicates the degree to which each variable in a set is predicted without error by the other variables (a KMO value close to 1 indicates that the sum of partial correlations is not large relative to the sum of correlations and so factor analysis should yield distinct and reliable factors).

```
# load the data
data <- psych::bfi[, 1:25] # 25 first columns corresponding to the items
data <- na.omit(data)
# option 1: check suitability
psych::KMO(data)
## Kaiser-Meyer-Olkin factor adequacy
## Call: psych::KMO(r = data)
## Overall MSA =  0.85
## MSA for each item =
##   A1   A2   A3   A4   A5   C1   C2   C3   C4   C5   E1   E2   E3   E4   E5   N1   N2   N3
## 0.75 0.84 0.87 0.88 0.90 0.84 0.80 0.85 0.83 0.86 0.84 0.88 0.90 0.88 0.89 0.78 0.78 0.86
##   N4   N5   O1   O2   O3   O4   O5
## 0.89 0.86 0.86 0.78 0.84 0.77 0.76
bartlett = psych::cortest.bartlett(data)
## R was not square, finding R from data
print(paste0("Chi-2: ", round(bartlett[["chisq"]],2),
             "; p-value: ", bartlett[["p.value"]]))
## [1] "Chi-2: 18146.07; p-value: 0"
# option 2: check suitability
# performance::check_factorstructure(data)
```

Once you are confident that the dataset is appropriate for factor analysis, you can explore a factor structure made of 5 (theoretically motivated) latent variables. Start by defining the model.

```
# optional: eigenvalues
ev <- eigen(cor(data))
print(ev$values)
##  [1] 5.1343112 2.7518867 2.1427020 1.8523276 1.5481628 1.0735825 0.8395389 0.7992062
##  [9] 0.7189892 0.6880888 0.6763734 0.6517998 0.6232530 0.5965628 0.5630908 0.5433053
## [17] 0.5145175 0.4945031 0.4826395 0.4489210 0.4233661 0.4006715 0.3878045 0.3818568
## [25] 0.2625390
psych::scree(data, pc=FALSE)
# fit an EFA
efa <- psych::fa(data, nfactors = 5, rotate="oblimin")
## Le chargement a nécessité le package : GPArotation
efa_para <- psych::fa(data, nfactors = 5, rotate="oblimin") |>
  parameters::model_parameters(sort = TRUE, threshold = "max")
efa_para
```

```
## # Rotated loadings from Factor Analysis (oblimin-rotation)
##
## Variable | MR2  |  MR1  |  MR3  |  MR5  |  MR4  | Complexity | Uniqueness
## -------------------------------------------------------------------------
## N1       | 0.83 |       |       |       |       |    1.07    |    0.32
## N2       | 0.78 |       |       |       |       |    1.03    |    0.39
## N3       | 0.70 |       |       |       |       |    1.08    |    0.46
## N5       | 0.48 |       |       |       |       |    2.00    |    0.65
## N4       | 0.47 |       |       |       |       |    2.33    |    0.49
## E2       |      | 0.67  |       |       |       |    1.08    |    0.45
## E4       |      | -0.59 |       |       |       |    1.52    |    0.46
## E1       |      | 0.55  |       |       |       |    1.22    |    0.65
## E5       |      | -0.42 |       |       |       |    2.68    |    0.59
## E3       |      | -0.41 |       |       |       |    2.65    |    0.56
## C2       |      |       | 0.67  |       |       |    1.18    |    0.55
## C4       |      |       | -0.64 |       |       |    1.13    |    0.52
## C3       |      |       | 0.57  |       |       |    1.10    |    0.68
## C5       |      |       | -0.56 |       |       |    1.41    |    0.56
## C1       |      |       | 0.55  |       |       |    1.20    |    0.65
## A3       |      |       |       | 0.68  |       |    1.06    |    0.46
## A2       |      |       |       | 0.66  |       |    1.03    |    0.54
## A5       |      |       |       | 0.54  |       |    1.48    |    0.53
## A4       |      |       |       | 0.45  |       |    1.74    |    0.70
## A1       |      |       |       | -0.44 |       |    1.88    |    0.80
## O3       |      |       |       |       | 0.62  |    1.16    |    0.53
## O5       |      |       |       |       | -0.54 |    1.21    |    0.70
## O1       |      |       |       |       | 0.52  |    1.10    |    0.68
## O2       |      |       |       |       | -0.47 |    1.68    |    0.73
## O4       |      |       |       |       | 0.36  |    2.65    |    0.75
##
## The 5 latent factors (oblimin rotation) accounted for 42.36% of the total variance of the
```


Scree pl
Eigen values of factors
4
3
2
1
0
5
10
factor num

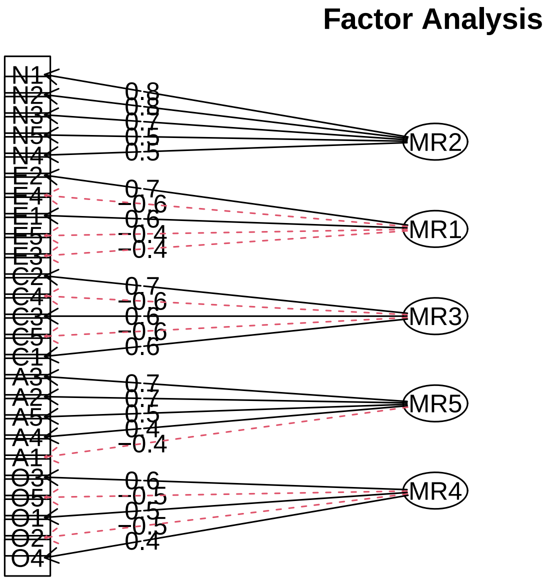
What do you see? How can you interpret the output?

Solution: Interpretation

The 25 items spread on the 5 latent factors nicely - the famous big 5.

It is possible to visualize the results to ease the interpretation:

```r
loads <- efa$loadings
psych::fa.diagram(loads)
```



Based on this model, you could predict back the scores for each individual for these new variables. This could be useful for further analysis (e.g. regression analysis).

Tips: use the function predict() and give labels to the latent factors. Based on this model, you could predict back the scores for each individual for these new variables. This could be useful for further analysis (e.g. regression analysis).

Tips: use the function predict() and give labels to the latent factors.

```r
# predictions <- predict(
#   efa_para,
#   names = c("Neuroticism", "Conscientiousness", "Extraversion", "Agreeableness", "Opennnes
#   verbose = FALSE
# )
# head(predictions)
```

11.14.2 Exercise 2: Environmental concerns

We will use survey data from the World Values Survey (WVS) website to investigate human belief and values, especially about environmental (EC) We will analyse Swiss data derived from the much larger WVS cross-country database (2007). The data can be downloaded here.

EC has been measured by a set of 8 items on a four-step Likert Scale. These items are thought to load on three different facets of EC: concerns about one's own community (water quality, air quality and sanitation), concerns about the world at large (fears about global warming, loss of biodiversity and ocean pollution), willingness to pay/do more. We want to answer the following question: Is the three-factor model proposed by the structure of items (or can we assume a one-factor structure)?

Let's first prepare the data and get the table of correlations:

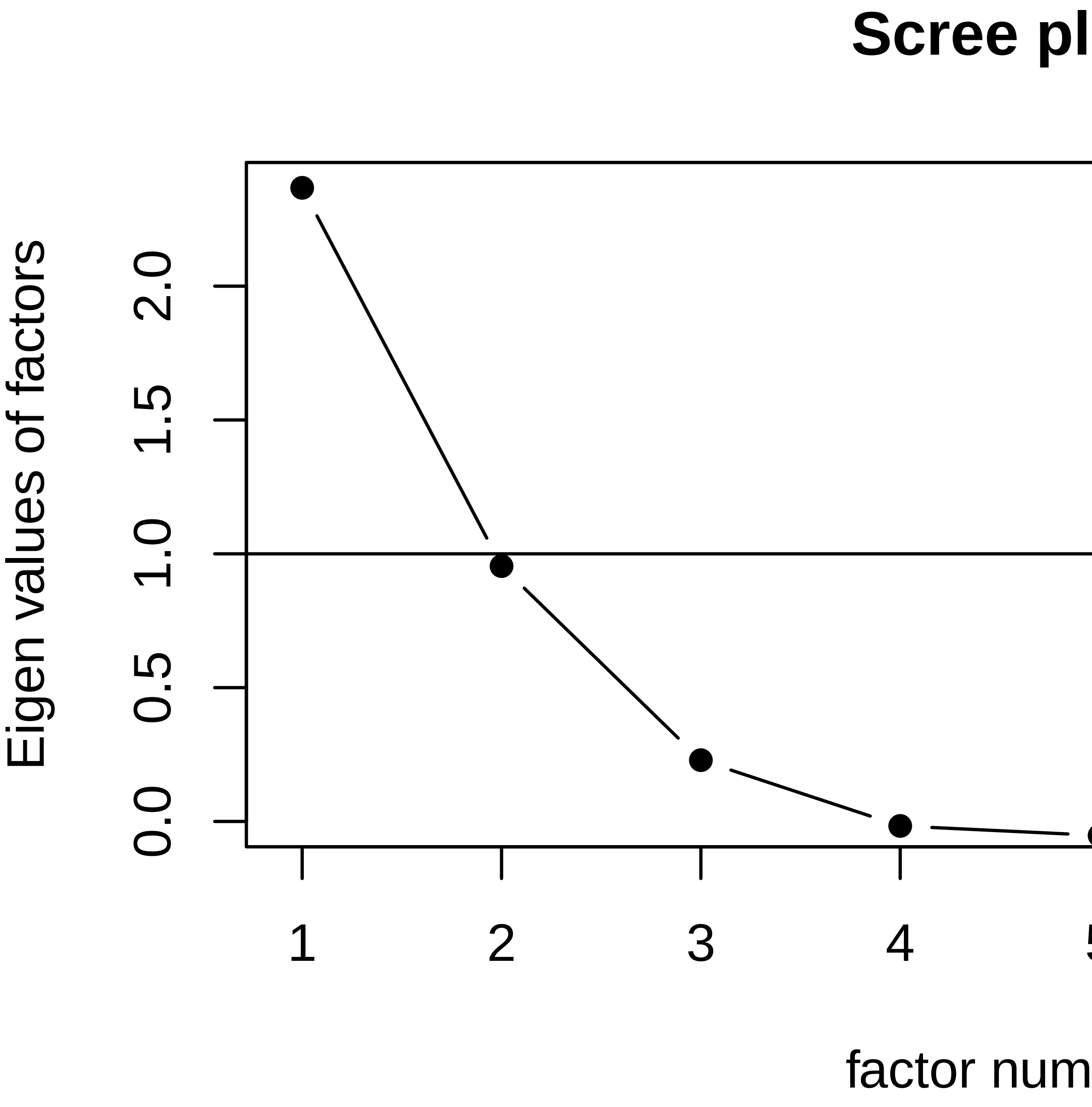
```
db <- openxlsx::read.xlsx(paste0(getwd(),
                  "/data/WV5_Data_Switzerland_Excel_v20201117.xlsx"))
colnames(db) <- gsub(":.*","",colnames(db))
sel <- db |>
  dplyr::select(V108,V109,V110,
                V111,V112,V113,
                V106,V107
                ) |>
  dplyr::rename("water"="V108",
                "air"="V109",
                "sanitation"="V110",
                "warming"="V111",
                "biodiv"="V112",
                "pollution"="V113",
                "taxes"="V106",
                "gov"="V107") |>
  stats::na.omit()
sel = replace(sel, sel==-1, NA)
sel = sel[complete.cases(sel),]
# reverse scale
for(i in 1:ncol(sel)){sel[,i] <- (sel[,i]-5)*(-1)}
# correlation
round(cor(sel),2)
##            water  air sanitation warming biodiv pollution taxes  gov
## water       1.00 0.73       0.84    0.04   0.06      0.10  0.00 0.09
## air         0.73 1.00       0.74    0.14   0.14      0.15  0.07 0.03
## sanitation  0.84 0.74       1.00    0.06   0.08      0.12  0.00 0.11
## warming     0.04 0.14       0.06    1.00   0.49      0.39  0.20 0.01
## biodiv      0.06 0.14       0.08    0.49   1.00      0.45  0.13 0.09
## pollution   0.10 0.15       0.12    0.39   0.45      1.00  0.14 0.00
## taxes       0.00 0.07       0.00     0.20    0.13       0.14  1.00 -0.23
## gov         0.09 0.03       0.11     0.01    0.09       0.00 -0.23  1.00
```

The first step is to test if the dataset is suitable for conducting factor analysis. To do so, run the Bartlett's Test of Sphericity and the Kaiser Meyer Olkin (KMO) measure.

```
# option 1: check suitability
psych::KMO(sel)
## Kaiser-Meyer-Olkin factor adequacy
## Call: psych::KMO(r = sel)
## Overall MSA =  0.72
## MSA for each item =
##      water        air sanitation    warming     biodiv  pollution      taxes        gov
##       0.71       0.84       0.70       0.69       0.66       0.74       0.62       0.50
bartlett = psych::cortest.bartlett(sel)
## R was not square, finding R from data
print(paste0("Chi-2: ", round(bartlett[["chisq"]],2),
             "; p-value: ", bartlett[["p.value"]]))
## [1] "Chi-2: 3331.64; p-value: 0"
# option 2: check suitability
# performance::check_factorstructure(sel)
```

Now, you can explore a factor structure made of 3 (theoretically motivated) latent variables. Start by defining the model.

```
# optional: eigenvalues
ev <- eigen(cor(sel))
print(ev$values)
## [1] 2.6757147 1.8568003 1.2061976 0.7281643 0.6039547 0.4827758 0.2920477 0.1543450
psych::scree(sel, pc=FALSE)
# fit an EFA
efa <- psych::fa(sel, nfactors = 3, rotate="oblimin")
efa_para <- psych::fa(sel, nfactors = 3, rotate="oblimin") |>
  parameters::model_parameters(sort = TRUE, threshold = "max")
efa_para
## # Rotated loadings from Factor Analysis (oblimin-rotation)
##
## Variable   | MR1  | MR2  |  MR3  | Complexity | Uniqueness
## -----------------------------------------------------------
## sanitation | 0.93 |      |       |    1.01    |    0.14
## water      | 0.92 |      |       |    1.00    |    0.17
## air        | 0.79 |      |       |    1.04    |    0.34
## biodiv     |      | 0.79 |       |    1.02    |    0.40
## warming    |      | 0.64 |       |    1.05    |    0.56
## pollution  |      | 0.57 |       |    1.04    |    0.65
## gov        |      |      | -0.53 |    1.14    |    0.72
## taxes       |       |       | 0.48  |    1.20    |    0.73
##
## The 3 latent factors (oblimin rotation) accounted for 53.70% of the total variance of the
```


Scree pl
Eigen values of factors
0.0
0.5
1.0
1.5
2.0
1
2
3
4
5
factor num

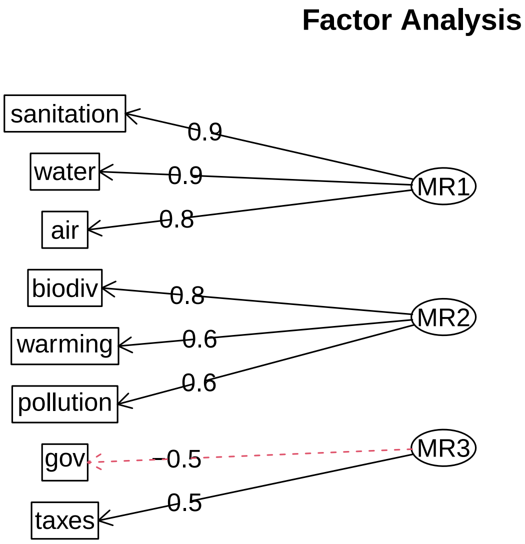
What do you see? How can you interpret the output?

Solution: Interpretation

The 8 items spread on the 3 latent factors nicely. Note that the 'gov' item has a negative loading.

It is possible to visualize the results to ease the interpretation:

```r
loads <- efa$loadings
psych::fa.diagram(loads)
```



Chapter 12

CFA

12.1 Confirmatory factor analysis

CFA is a model-data fit test based on multivariate regression. Outputs are coefficients of paths and fit indices. If the paths are significant and indices indicate acceptable or high degree of fit, that means the structural model is confirmed by data.

12.1.1 Reminder: latent constructs

Latent constructs are not “directly” measurable (e.g. media usage motives, media or brand ratings, attitudes, emotions and empathy in the media reception, trust, etc.).

Furthermore, definitions can be understood differently or terms are completely unknown. Some concepts also have several dimensions/facets, so that one question alone is not enough to fully understand the concept.

Therefore, we should ensure several “indicator variables” that allow conclusions to be drawn about the latent variable.

12.1.2 Reminder: fundamental theorem

Like the EFA, the CFA is also based on the fundamental theorem of factor analysis:

$$x_{ij} = a_{j1}f_{j1} + a_{j2}f_{j2} + ... + a_{jn}f_{jn}$$

with n factors.

- x_{ij} = value of person i on observed variable j

- a_{j1} factor loading = correlation of a variable j and factor 1
- f_{i1} = factor value of person i on factor 1

Based on a given (hypothesized) relationship structure between indicator variables and factors (i.e. whether a correlation is assumed or not), to estimate the factor loadings in such a way that the empirical value covariance structure between items can be reproduced as well as possible with their help.

It is about a comparison (better: an adjustment) between empirically found correlations between indicators and the theoretically assumed correlations between factors and items.

12.1.3 Recap: EFA versus CFA

In EFA:

- the number of factors is searched for exploratively with the help of more or less fixed criteria (Kaiser criterion, scree plot, content plausibility).
- the assignment of the indicators to the construct is exploratory (via the factor loadings and the rotation).
- all variables load on all factors

In CFA:

- the number of factors is determined a priori.
- the indicators are assigned to the construct on the basis of theoretical considerations.
- all aspects of the factor model must be specified in advance: number of factors, relationship patterns between items and factors and between factors, etc.

Thus, CFA is one of the structure-testing methods of multivariate data analysis. Its central goal is testing of measurement models for hypothetical construct.

12.1.4 Concrete research applications

CFA enables us to tackle research questions such as:

- Is the assumed factor structure confirmed?
- Does my theoretically specified measurement model match my data?
- Is there a good fit between the theoretical model and the empirical model?
- What is the quality of “competing” measurement models?
- Do my data confirm the assumed dimensionality of my construct?
- Do my data confirm the assumed hierarchy of my construct?
- Can I use the same measuring instrument in different groups (e.g. countries)?
- Etc…

12.2 Requirements of CFA

Mathematical requirements:

- There are several interval-scaled characteristics (items), each of which is (roughly) normally distributed.
- Items that should theoretically load on a factor should correlate empirically (if not, they are probably not determined by the same factor)
- Sufficient directly measured variables (items) must be available in order to be able to test the assumed model structure made up of items and factors (model identification, more on this later)

Theoretical assumptions:

- Theoretical or at least "logical" justification for the previously expected model structure made up of items and factors
- Reflective measurement model

Recommendation: It is best to have an EFA before the CFA - either with the same or with a different sample.

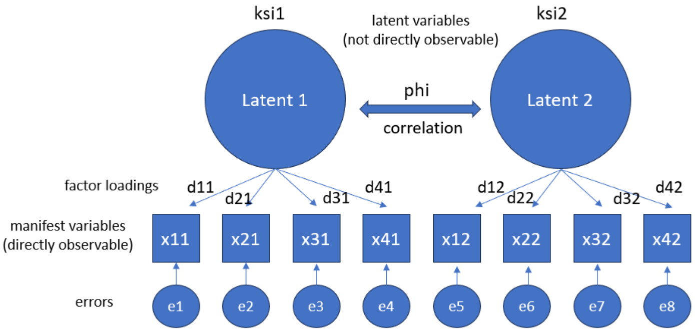
To this end, the theoretical/latent construct is first precisely defined: describe exactly which aspects a theoretical term contains. Then indicators for the latent construct are developed (item formulation). These should depict the theoretical construct (or its individual dimensions) as precisely as possible. Indicators themselves should correlate with one another, since they depend on the same hypothetical construct (reflective models!).

12.3 Model structure

The problem of identifying models describes the question of whether a system of equations can be solved mathematically. The model parameters (free parameters) must be estimated from the empirical variances and covariances of the manifest variables.

Accordingly, all parameters to be estimated in the model (factor loadings, error terms and, if applicable, permitted correlations between latent variables) must be calculable/representable with the help of empirical parameters.

If this is the case, then the model is identified. If this is not the case, there is (so to speak) an equation with too many unknowns. Such an equation is "unsolvable", or such a model is "unidentified". It is about determining the degrees of freedom of the model.



$$\begin{bmatrix} X11 \\ X21 \\ X31 \\ X41 \\ X12 \\ X22 \\ X32 \\ x42 \end{bmatrix} = \begin{bmatrix} d11 & 0 \\ d21 & 0 \\ d31 & 0 \\ d41 & 0 \\ 0 & d12 \\ 0 & d22 \\ 0 & d32 \\ 0 & d42 \end{bmatrix} * \begin{bmatrix} ksi1 \\ ksi2 \end{bmatrix} + \begin{bmatrix} e1 \\ e2 \\ e3 \\ e4 \\ e5 \\ e6 \\ e7 \\ e8 \end{bmatrix}$$

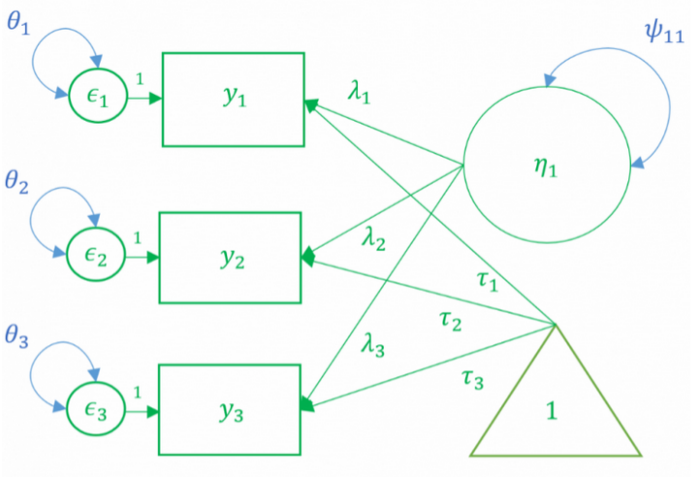
The identification of the model structure consists of two “tasks”:

- defining a metric for the latent constructs
- checking whether there is enough information to estimate the model

Both steps influence the number of degrees of freedom of the model. This aspect also relates to the complexity of the model. The more complex a model is, the more parameters have to be estimated, and the greater the model’s degrees of freedom must be in order to make the estimation possible.

12.4 Identification of a three-item one-factor CFA

Identification for the one-factor CFA with three items is necessary. The total parameters include 3 factor loadings, 3 residual variances and 1 factor variance (which comes from the fact that we do not observe the factor but are estimating its variance).



In order to identify a factor in a CFA model with three or more items, there are two options known respectively as the marker method and the variance standardization method:

- marker method fixes the first loading of each factor to 1, where the model-implied covariance matrix is as follows:

$$\psi_{11}\begin{pmatrix}1\\ \lambda_2\\ \lambda_3\end{pmatrix}\begin{pmatrix}1 & \lambda_2 & \lambda_3\end{pmatrix}+\begin{pmatrix}\theta_{11} & 0 & 0\\ 0 & \theta_{22} & 0\\ 0 & 0 & \theta_{33}\end{pmatrix}$$

- variance standardization method fixes the variance of each factor to 1 but freely estimates all loadings, where the model-implied covariance matrix is as follows:

$$1\begin{pmatrix}\lambda_1\\ \lambda_2\\ \lambda_3\end{pmatrix}\begin{pmatrix}\lambda_1 & \lambda_2 & \lambda_3\end{pmatrix}+\begin{pmatrix}\theta_{11} & 0 & 0\\ 0 & \theta_{22} & 0\\ 0 & 0 & \theta_{33}\end{pmatrix}$$

Notice in both models that the residual covariances stay freely estimated.

By fixing $\lambda_1 = 1$ and by setting the unique residual covariances to zero (e.g. $\theta_{12} = \theta_{21} = \theta_{13} = \theta_{31} = \theta_{23} = \theta_{32} = 0$), we can demonstrate that we obtain a just-identified model. Indeed, we start with 10 total (unique) parameters, but we fix 1 loading, and 3 unique residual covariances (resulting in 4 fixed parameters). Furthermore, the number of known values is $(3(3+1))/2 = 6$. Thus, the number of free parameters is 10-4=6. Since we have 6 known values, our degrees of freedom is 6-6=0, which is defined to be saturated. This is known as the marker method.

Similarly, we can go through the process of calculating the degrees of freedom for the variance standardization method. We also start with 10 total (unique) parameters, but we fix 1 factor variance, and 3 unique residual covariances (resulting in 4 fixed parameters). Furthermore, the number of known values is $(3(3+1))/2 = 6$. Thus, the number of free parameters is 10-4=6. Since we have 6 known values, our degrees of freedom is 6-6=0, which is defined to be saturated.

ℹ Degree of freedom: calculation

The formula for calculation degrees of freedom is as follows:

$$df = \frac{m(m+1)}{2} - 2m - \frac{X(X-1)}{2}$$

where m represent the number of indicators and X the number of independent latent constructs. We can differentiate between different parts:

- $\frac{m(m+1)}{2}$ gives us the maximum degree of freedom in the model
- $2m$ specifies the number of parameters to be estimated
- $\frac{X(X-1)}{2}$ represents the free off-diagonal covariances of the constructs

In the above example, we have 8 indicators (m) and 2 latent constructs (X). Applying the formula, we obtain:

$$df = \frac{8(9)}{2} - 16 - \frac{2(1)}{2} = 19$$

12.5 Types of parameters

Fixed parameters:

- Parameters that are assigned a specific constant value a priori.
- Mostly when it is assumed on the basis of theoretical considerations that there are no causal relationships between certain variables, the corresponding parameters are set to zero
- However, a value greater than zero can also be specified, to which a parameter is then fixed, for example if the exact stronger of a relationship between two variables is already known in advance

Constrained parameters:

- Parameters that should be estimated in the model, but whose value should correspond exactly to the value of one or more other parameters.
- E.g. if the influence of two independent variables on a dependent variable is considered to be equal.
- If two parameters are defined as restricted, only one parameter has to be estimated instead of two.

Free parameters:

- All other parameters whose values are considered unknown and should be estimated from the empirical data.

12.6 Defining a metric for the latent constructs

The latent variables and error variables to be estimated initially have no metric. In order to be able to interpret the variable later, a scale must be assigned.

To do so, you can choose a reference variable and fix the factor loading to 1. With regard to the latent variable, the best indicator variable should be selected. It then means that the latent variable is identical to the selected indicator variable except for the measurement error. In addition, all relationships between the measurement error terms to be estimated and the measured indicator variables are fixed at 1. It then means that the measurement errors to be estimated correspond to the observed values, except for the influence of the latent variables.

12.7 Checking whether there is enough information to estimate the model

If n manifest variables are collected within the framework of a project, then the empirical variances and covariances of these variables can be calculated:

$$p = \frac{n(n+1)}{2}$$

where p corresponds to the number of non-redundant values in the variance-covariance matrix (it thus corresponds to the available "information" that we use as the basis for calculating all free parameters of our model).

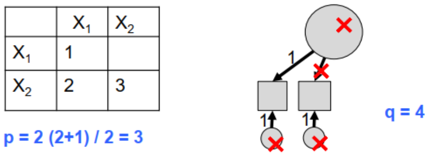
The difference between the available empirical information (p) and the number of (free) parameters to be estimated (q) gives the degrees of freedom of the model (df_M):

$$df_M = p - q$$

where p is the number of empirical information available from the n empirical indicators) and q corresponds to the number of free parameters to be estimated in each case.

If the empirically available information is the same as the number of parameters to be estimated, the number of model degrees of freedom corresponds to zero. The number of model parameters to be estimated must not exceed the number of empirical information given, otherwise a model is not identified (i.e., not solvable). Furthermore, the degrees of freedom of the model must be superior or equal to 0 for a specified model to be identified.

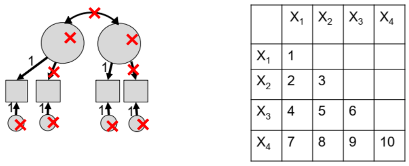
In the following example, the model is under-identified (df $<$ 0): the information available from the empirical data is not sufficient to calculate the parameters.



	X_1	X_2
X_1	1	
X_2	2	3

What about the next example?

Is the model under- or over-identified?



	X_1	X_2	X_3	X_4
X_1	1			
X_2	2	3		
X_3	4	5	6	
X_4	7	8	9	10

A positive number of degrees of freedom is necessary for the solvability of SEM/CFA! For empirical surveys, ensure that at least as many indicator variables are surveyed as are necessary to achieve a positive number of degrees of freedom. In the example above:

$$p = (4(4+1))/2 = (4*5)/2 = 10$$

$$q = 9$$

$$df_M = p - q = 10 - 9 = 1$$

Thus, the model is over-identified.

12.8 Parameters determination

The CFA first calculates the empirical variance-covariance matrix (in short: covariance matrix) with the collected data (= empirical relationships in the data). Based on the defined measurement model, the model-theoretical covariance matrix is calculated (= expected relationships in the data).

The covariance matrix should be reproduced as accurately as possible by the model. This means that the model-theoretical covariance matrix should resemble the empirical covariance matrix as closely as possible.

The better it is possible to reproduce the empirical covariance matrix with the model-theoretical covariance matrix, the more reliable the parameter estimates are and the better the model is.

To assess the model quality so-called fit-indices/fit-measures describe how good the model is.

Exactly one solution is possible for exactly identified models, several solutions are possible for over-identified models, the best solution must be found. The solution is improved with each (iterative) step by improving the fit of the model (i.e. the fit of the model to the empirical data/the empirical covariance matrix). In other words, the difference between the empirical and model-theoretical covariance matrix should be minimal.

The parameters can actually be estimated using various functions. The most common method is the so-called Maximum Likelihood (ML) method. Here the estimated parameter is selected that is most likely to reproduce the observed data. This method assumes metric data.

12.9 Checking the quality

Once a solution has been found, it must be checked for quality. The reliability and validity of the measurement must therefore be assessed.

- Reliability: Repeated measurements must always produce the same result (items must all show a high loading with the latent construct)
- Validity: the measuring instrument should measure what it is supposed to measure

The assessment of the results follows a multi-stage process:

- Checking at indicator level
- Construct level testing
- Examination on model level

Concerning the examination at indicator level (items), we must ensure that only "good" indicators are included in a model. Thus, there should be sufficiently high correlations between the items (e.g. EFA, Cronbach's Alpha, etc). Furthermore, the plausibility of the factor loadings, the significance of the factor loadings, and the strength of the factor loadings (and of the squared factor loadings) should be determined.

- Standardized solution corresponds to factor loadings (to be interpreted as in EFA, values >.5 are desirable)
- Squared factor loadings indicate the percent variance of a variable that is explained by the factor behind it (should reach at least .3)

At the construct level (factors), the question is answered as to whether the constructs/factors are reliably and validly measured (e.g. factor reliability, average extracted variance of the factors, and discriminant validity).

- Factor reliability (should be > .5)
- Average extracted variance of factors (DEV, should be > .5)
- Discriminant validity (Fornell/Larcker criterion)
    - DEV > squared correlation of the constructs (=common variance of the factors)

  - Measure for the fact that the explanation of the variance of the factors on the indicators is greater than the connection between the factors

The examination at model level (overall model) suggests to check whether the empirical variance-covariance matrix is reproduced as well as possible by the model-theoretical variance-covariance matrix (e.g. Fit-indices, Chi-Square test statistic, RMSEA and SRMR).

- Chi-Square Test (H0: empirical covariance matrix = model-theoretical covariance matrix)
  - The smaller the difference between the two matrices, the smaller the chi-square value (the smaller the chi-square, the better).
  - The test should not turn out to be significant, since equality between the empirical and model-theoretical variance-covariance matrix is desirable.

Recommendations for model model evaluation

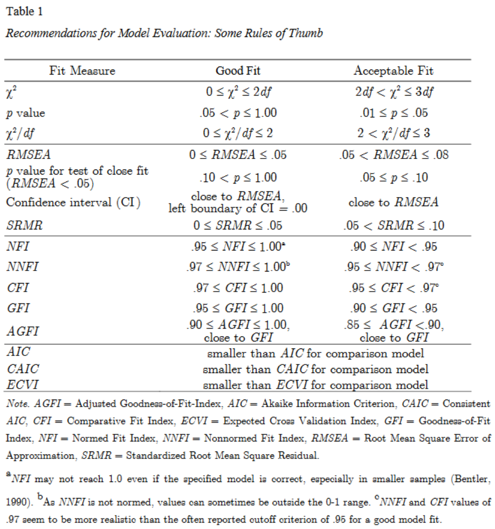


Table 1

Recommendations for Model Evaluation: Some Rules of Thumb

Fit Measure	Good Fit	Acceptable Fit
χ^2	$0 \leq \chi^2 \leq 2df$	$2df < \chi^2 \leq 3df$
p value	$.05 < p \leq 1.00$	$.01 \leq p \leq .05$
χ^2/df	$0 \leq \chi^2/df \leq 2$	$2 < \chi^2/df \leq 3$
RMSEA	$0 \leq RMSEA \leq .05$	$.05 < RMSEA \leq .08$
p value for test of close fit ($RMSEA < .05$)	$.10 < p \leq 1.00$	$.05 \leq p \leq .10$
Confidence interval (CI)	close to *RMSEA*, left boundary of CI = .00	close to *RMSEA*
SRMR	$0 \leq SRMR \leq .05$	$.05 < SRMR \leq .10$
NFI	$.95 \leq NFI \leq 1.00$[a]	$.90 \leq NFI < .95$
NNFI	$.97 \leq NNFI \leq 1.00$[b]	$.95 \leq NNFI < .97$[c]
CFI	$.97 \leq CFI \leq 1.00$	$.95 \leq CFI < .97$[c]
GFI	$.95 \leq GFI \leq 1.00$	$.90 \leq GFI < .95$
AGFI	$.90 \leq AGFI \leq 1.00$, close to *GFI*	$.85 \leq AGFI < .90$, close to *GFI*
AIC	smaller than *AIC* for comparison model	
CAIC	smaller than *CAIC* for comparison model	
ECVI	smaller than *ECVI* for comparison model	

Note. *AGFI* = Adjusted Goodness-of-Fit-Index, *AIC* = Akaike Information Criterion, *CAIC* = Consistent *AIC*, *CFI* = Comparative Fit Index, *ECVI* = Expected Cross Validation Index, *GFI* = Goodness-of-Fit Index, *NFI* = Normed Fit Index, *NNFI* = Nonnormed Fit Index, *RMSEA* = Root Mean Square Error of Approximation, *SRMR* = Standardized Root Mean Square Residual.

[a] *NFI* may not reach 1.0 even if the specified model is correct, especially in smaller samples (Bentler, 1990). [b] As *NNFI* is not normed, values can sometimes be outside the 0-1 range. [c] *NNFI* and *CFI* values of .97 seem to be more realistic than the often reported cutoff criterion of .95 for a good model fit.

The original table from Schermelleh-Engel et al. (2003) can be found here.

12.10 Model Chi-square

The model chi-square is defined as either nF_{ML} or $(n-1)F_{ML}$ depending on the statistical package where n is the sample size and F_{ML} is the fit function from maximum likelihood.

The model chi-square is a meaningful test only when you have an over-identified model (there are still degrees of freedom left over after accounting for all the free parameters in your model).

Note on the sample size

Model chi-square is sensitive to large sample sizes. So how big of a sample do we need? Kline (2016) notes the $N : q$ rule, which states that the sample size should be determined by the number of q parameters in your model, and the recommended ratio is $20 : 1$. This means that if you have 10 parameters, you should have n=200.
Reference: Kline, R. B. (2016). Principles and Practice of Structural Equation Modelling, 4th edn. New York. NY: The Guilford Press.

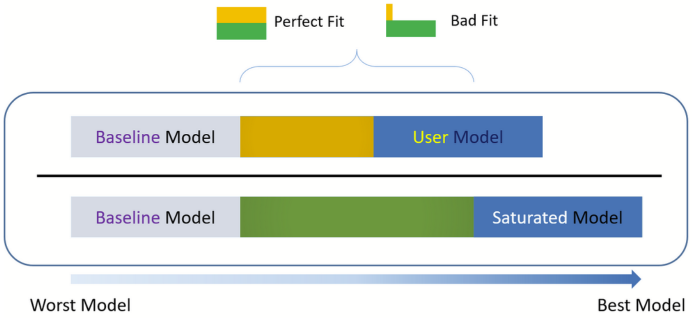
12.11 Baseline model

The model test baseline is also known as the null model, where all covariances are set to zero and freely estimates variances. Rather than estimate the factor loadings, we only estimate the observed means and variances (removing all the covariances). Recall that we have $\frac{p(p+1)}{2}$ covariances. Since we are only estimating the p variances we have $\frac{p(p+1)}{2} - p$ degrees of freedom. You can think of the baseline (or null) model as the worst model you can come up with and the saturated model as the best model. Theoretically, the baseline model is useful for understanding how other fit indices are calculated.

12.12 Approximate fit indexes

To resolve the problem related to the Chi-square sensitivity under large samples, approximate fit indexes that are not based on accepting or rejecting the null hypothesis were developed. These approximate fit indexes can be classified into:

- incremental or relative fit indexes (e.g. CFI and TLI): assesses the ratio of the deviation of the user model from the baseline model against the deviation of the saturated model (best fitting model) from the baseline model



- absolute fit indexes (e.g. RMSEA): compares the user model to the observed data

12.12.1 CFI and TLI

The CFI or comparative fit index is a popular fit index as a supplement to the model chi-square. Let $\delta = \chi^2 - df$ where df is the degrees of freedom for that particular model. The closer δ is to zero, the more the model fits the data. The formula for the CFI is:

$$CFI = \frac{\delta(Baseline) - \delta(User)}{\delta(Baseline)}$$

TLI is also an incremental fit index. The term used in the TFI is the relative chi-square (a.k.a. normed chi-square) defined as χ^2/df. Compared to the model chi-square, relative chi-square is less sensitive to sample size. Whereas $\chi^2/df = 1$ indicates perfect fit, some researchers say that a relative chi-square greater than 2 indicates poor fit. The TLI is defined as:

$$TLI = \frac{min(\chi^2_{Baseline}/df_{Baseline}, 1) - min(\chi^2_{User}/df_{User}, 1)}{min(\chi^2_{Baseline}/df_{Baseline}, 1) - 1}$$

Acceptable range of Chi-square values

Suppose you ran a CFA with 20 degrees of freedom. What would be the acceptable range of chi-square values based on the criteria that the relative chi-square greater than 2 indicates poor fit?
The range of acceptable chi-square values ranges between 20 (indicating perfect fit) and 40, since $40/20 = 2$.

12.12.2 RMSEA

RMSEA (Root Mean Squared Error of Approximation):

- This measure uses inferential statistics to check whether a model can approximate reality well (= approximation test).
- It is therefore not about the absolute correctness of the model, as with the Chi-Square test, but about evaluating the best possible approximation.
- The measure also includes the model complexity.

$$RMSEA = \sqrt{\frac{\delta}{df(n-1)}} = \sqrt{\frac{\chi^2 - df}{df(n-1)}}$$

Recommendations:

- RMSEA .05 good fit
- RMSEA .08 acceptable fit

SRMR (Standardized Root Mean Squared Residual):

- Usually there is a discrepancy between the empirical and model-theoretical covariance matrix. Therefore, descriptive measures of discrepancy are often used.
- These give an answer to the question of whether an existing discrepancy can be neglected and also includes the model complexity.
- If the empirical and model-theoretical covariance matrix are completely identical, this measure assumes a value of zero.

Recommendations:

- SRMR .05 good fit
- SRMR .10 more acceptable

12.13 In case of bad model fit

There are several things we can do if the model fit is too bad:

- View Modification indices (MI) and adjust model if necessary
- MI indicate how the model fit can be improved (e.g. how the theoretical model can better approximate the empirical data)
- MI usually propose to release certain fixed parameters
- Many MI proposals do not necessarily make sense with regard to the theoretical assumptions (find a balance between mathematical fit and theoretical meaning)

12.14 Comparison of different models

Comparing different models can be useful when we have competing theoretical models.

Comparison rules:

- lower RMSEA = descriptively better model
- lower SRMR = descriptively better model
- larger CFI = descriptively better model
- smaller AIC = descriptively better model

Statistically verified comparisons between competing models only possible with nested models (= models are exactly identical except that one path is more/less estimated).

To do so, we can rely on the Chi-Square Difference Test. If the test is not significant, then the model with more fixed parameters (i.e. the more economical and therefore theoretically clearer model) is no worse than the model in which more paths are allowed. Beware that, like the chi-square test, the chi-square difference test is also very fast significant.

12.15 In a nutshell

In CFA, a previously accepted measurement model (number of factors, assignment of indicators to factors, etc.) is checked on the basis of collected data.

The mathematical specification of measurement models works with fixed and free parameters. Path diagrams can be converted into systems of equations.

The factor analysis or measurement model is essentially a linear regression model where the main predictor, the factor, is latent. For a single subject, the simple linear regression equation is defined as: $y = b_0 + b_1 x + e$. Similarly, for a single item, the factor analysis model is: $y_1 = \tau_1 + \lambda_1 \eta + e$. We can represent this multivariate model (i.e., multiple outcomes, items, or indicators) as a matrix equation:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix} + \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} \eta_1 + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}$$

The variance-covariance matrix can be described using the model-implied covariance matrix. This is in contrast to the observed population covariance matrix which comes only from the data.

$$\sum(\theta) = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} (\psi_{11}) \begin{pmatrix} \lambda_1 & \lambda_2 & \lambda_3 \end{pmatrix} + \begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ \theta_{31} & \theta_{32} & \theta_{33} \end{pmatrix}$$

The basic assumption of factor analysis is that for a collection of observed variables there are a set of underlying factors that can explain the interrelationships among those variables. These interrelationships are measured by the covariances.

Traditionally, the intercepts (τ) are not estimated, which means that all the parameters we need can come directly from the covariance model.

In the model-implied covariance, we assume that the residuals are independent which means that for example, the covariance between the second and first residual ($\theta 23$), is set to zero. As such the only covariance terms to be estimated are the variance of the latent factors (ψ) and the variances of the residuals ($\theta 11$, $\theta 22$ and $\theta 33$).

Models must have sufficient degrees of freedom (relationship between empirical information and parameters to be estimated) in order to be mathematically solvable (model identification).

In order to identify a factor in a CFA model with three or more items, there are two options known respectively as the marker method and the variance standardization method:

- marker method fixes the first loading of each factor to 1,
- variance standardization method fixes the variance of each factor to 1 but freely estimates all loadings.

In matrix notation, the marker method can be written as:

$$\begin{pmatrix} 1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} (\psi_{11}) \begin{pmatrix} 1 & \lambda_2 & \lambda_3 \end{pmatrix} + \begin{pmatrix} \theta_{11} & 0 & 0 \\ 0 & \theta_{22} & 0 \\ 0 & 0 & \theta_{33} \end{pmatrix}$$

In matrix notation, the variance standardization method looks like:

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} (1) \begin{pmatrix} \lambda_1 & \lambda_2 & \lambda_3 \end{pmatrix} + \begin{pmatrix} \theta_{11} & 0 & 0 \\ 0 & \theta_{22} & 0 \\ 0 & 0 & \theta_{33} \end{pmatrix}$$

The results of a CFA are assessed on 3 levels:

- Indicator level: strength and significance of the factor loadings as well as the size of the squared factor loadings
- Construct level: factor reliability, average extracted variance of factors, discriminant validity
- Model Level: Chi-Square Test Statistic, RMSEA, SRMR

Four commonly used model fit measures are:

- Model chi-square is the chi-square statistic obtained from the maximum likelihood statistic (in lavaan: Test Statistic for the Model Test User Model)
- CFI is the Comparative Fit Index – values can range between 0 and 1 (values greater than 0.90, conservatively 0.95 indicate good fit)
- TLI Tucker Lewis Index which also ranges between 0 and 1 (if greater than 1 it should be rounded to 1; values greater than 0.90 indicating good fit). The CFI is always greater than the TLI.
- RMSEA is the root mean square error of approximation (in lavaan: p-value of close fit, that the RMSEA < 0.05. If rejected, it means that the model is not a close fitting model).

Historically, model chi-square was the only measure of fit but in practice the null hypothesis was often rejected due to the chi-square's heightened sensitivity under large samples. To resolve this problem, approximate fit indexes that were not based on accepting or rejecting the null hypothesis were developed. These can be further classified into absolute (e.g. CFI and TLI) and incremental (e.g. RMSEA) fit indexes:

- incremental fit indexes: assess the ratio of the deviation of the user model from the worst fitting model (a.k.a. the baseline model) against the deviation of the saturated model from the baseline model.
- absolute fit indexes: compare the user model to the observed data.

12.16 How it works in R?

See the lecture slides confirmatory factor analysis:

You can also download the PDF of the slides here:

12.17 Quiz

True

False

Statement

The elements of the Factor Matrix represent correlations of each item with a factor.

In common factor analysis, the sum of squared loadings is the eigenvalue (e.g., assume a two-factor solution with 3+ items).

The communality is unique to each factor or component.

In oblique rotation, an element of a factor pattern matrix is the unique contribution of the factor to the item.

View Results

My results will appear here

12.18 Time to practice on your own

You can download the PDF of the CFA exercises here:

12.18.1 Exercise 1: Big-5

To illustrate CFA, let us use the same International Personality Item Pool data available in the psych package to confirm the big 5 personality structure.

Start by defining the model using the lavaan syntax and extract the parameters.

```r
# load the data
data <- psych::bfi[, 1:25] # 25 first columns corresponding to the items
data <- na.omit(data)
# write the model
model_measurement <- "
  Neuroticism =~ N1 + N2 + N3 + N4 + N5
  Conscientiousness =~ C1 + C2 + C3 + C4 + C5
  Extraversion =~ E1 + E2 + E3 + E4 + E5
  Agreeableness =~ A1 + A2 + A3 + A4 + A5
  Opennness =~ O1 + O2 + O3 + O4 + O5
"
fit_measurement <- lavaan::sem(model_measurement, data = data)
# summary(fit_measurement, fit.measures = TRUE, standardized = TRUE)
coef <- lavaan::parameterestimates(fit_measurement)
# output: only 10 first rows
lavaan::parameterEstimates(fit_measurement, standardized=TRUE) |>
  dplyr::select('Latent Factor'=lhs,
         Indicator=rhs,
         B=est,
         SE=se,
         Z=z,
         'p-value'=pvalue,
         Beta=std.all) |>
  dplyr::slice_head(n = 10) |> # select only the 10 first rows
  knitr::kable(digits = 3, booktabs=TRUE)
```

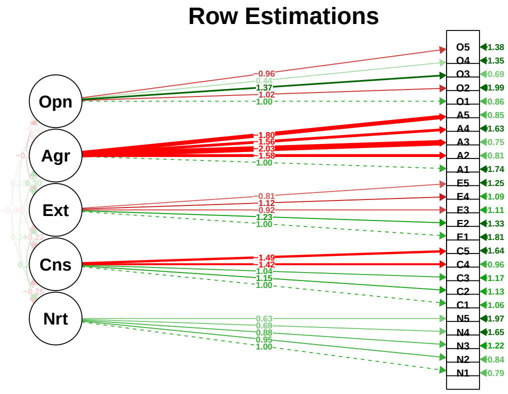
Latent Factor	Indicator	B	SE	Z	p-value	Beta
Neuroticism	N1	1.000	0.000	NA	NA	0.825
Neuroticism	N2	0.947	0.024	39.899	0	0.803
Neuroticism	N3	0.884	0.025	35.919	0	0.721
Neuroticism	N4	0.692	0.025	27.753	0	0.573
Neuroticism	N5	0.628	0.026	24.027	0	0.503
Conscientiousness	C1	1.000	0.000	NA	NA	0.551
Conscientiousness	C2	1.148	0.057	20.152	0	0.592
Conscientiousness	C3	1.036	0.054	19.172	0	0.546
Conscientiousness	C4	-1.421	0.065	-21.924	0	-0.702
Conscientiousness	C5	-1.489	0.072	-20.694	0	-0.620

Also get the fit statistics of the model. Is it a good model?

```
fits <- lavaan::fitMeasures(fit_measurement)
data.frame(fits = round(fits[c("ntotal","df",
                               "chisq","pvalue",
                               "rmsea","rmsea.pvalue","srmr")], 2))
##                  fits
## ntotal        2436.00
## df             265.00
## chisq         4165.47
## pvalue           0.00
## rmsea            0.08
## rmsea.pvalue     0.00
## srmr             0.08
```

Lastly, make a visualization of the output using the semPaths() function.

```
semPlot::semPaths(fit_measurement, what = "est", rotation = 2,
                  style = "lisrel", font = 2)
title("Row Estimations")
```



12.18.2 Exercise 2: Calculate the degree of freedom for one-factor CFA with more than 3 items

The benefit of performing a one-factor CFA with more than three items is that:

- your model is automatically identified (there will be more than 6 free parameters)
- your model will not be saturated (there will be degrees of freedom left over to assess model fit).

Imagine that we have specified the following model in lavaan with 8 items.

```
#one factor eight items, variance std
mod <- 'f =~ q01 + q02 + q03 + q04 + q05 + q06 + q07 + q08'
onefac8items <- cfa(mod, data=dat,std.lv=TRUE)
summary(onefac8items, fit.measures=TRUE, standardized=TRUE)
```

From this model, explain how to obtain 20 degrees of freedom from the 8-item one factor CFA by first calculating the number of free parameters and comparing that to the number of known values.

ℹ Solution

The number of elements in the variance-covariance matrix is:

$$\frac{n*(n+1)}{2} = \frac{8*(8+1)}{2} = 36$$

We also have 8 loadings (λ_i), 8 residual variances (θ_i) and 1 variance of the factor (ψ_i). Thus, in total we have 17 unique parameters.
This gives us 17-1=16 free parameters, where we have fixed 1 parameter (using the variance standardization method).
Therefore, the degrees of freedom is 36-16=20. This suggests that we have an over-identified model (degrees of freedom above 0).

12.18.3 Exercise 3: Calculate the degree of freedom for two-factor CFA with more than 3 items

Imagine that we have specified the following model in lavaan with two factors:

```
#one factor eight items, variance std
mod <- '
  f1 =~ q01 + q03 + q04 + q05 + q08
  f2 =~ a*q06 + a*q07
  ## equate the 2 items on the same factors
  ## while setting the factor variance at 1
  f1 ~~ 0*f2' ## orthogonal factors
twofac7items <- cfa(mod, data=dat,std.lv=TRUE)
summary(twofac7items, fit.measures=TRUE, standardized=TRUE)
```

From this model, explain how to obtain 15 degrees of freedom from the 7-item two-factor CFA by first calculating the number of free parameters and comparing that to the number of known values. We also make the assumption of uncorrelated (orthogonal) factors.

ℹ Solution

The number of elements in the variance-covariance matrix is:

$$\frac{n*(n+1)}{2} = \frac{7*(7+1)}{2} = 28$$

We also have 7 loadings (λ_i), 7 residual variances (θ_i) and 2 variance of the factors (ψ_i). Thus, in total we have 16 unique parameters. But we make the assumption of uncorrelated factors, thus we have 14 free parameters. Therefore, the degrees of freedom is 28-14=14. However, as we constrained the loadings of q06 and q07 to be equal, it frees up a parameter and thus we have 14+1=15 degrees of freedom.

Now, we make the assumption of correlated (oblique) factors, which gives us the following model:

```
#one factor eight items, variance std
mod <- '
  f1 =~ q01 + q03 + q04 + q05 + q08
  f2 =~ q06 + q07'
twofac7items <- cfa(mod, data=dat,std.lv=TRUE)
summary(twofac7items, fit.measures=TRUE, standardized=TRUE)
```

Explain how to obtain 13 degrees of freedom from the 7-item two-factor CFA by first calculating the number of free parameters and comparing that to the number of known values.

ℹ Solution

The number of elements in the variance-covariance matrix is:

$$\frac{n*(n+1)}{2} = \frac{7*(7+1)}{2} = 28$$

We also have 7 loadings (λ_i), 7 residual variances (θ_i) and 1 covariance of the factors (ψ_{21}). Thus, in total we have 15 unique parameters. Therefore, the degrees of freedom is 28-15=13.

Chapter 13

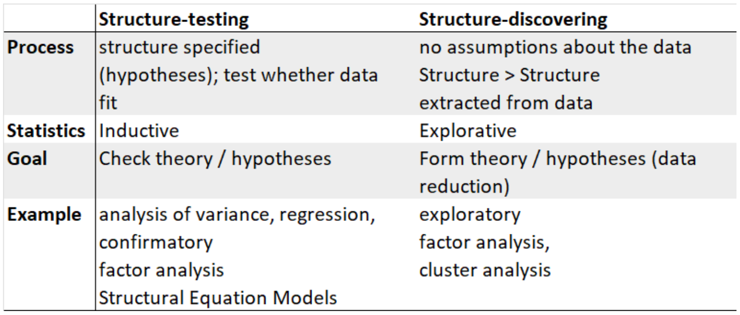
Recap EFA and CFA

13.1 Recap on factor analysis: exploratory and confirmatory

In exploratory factor analysis (EFA), all measured variables are related to every latent variable. It is used to reduce data to a smaller set of summary variables and to explore the underlying theoretical structure of the phenomena. It asks what factors are given in observed data and, thereby, requires interpretation of usefulness of a model (it should be confirmed with confirmatory factor analysis).

In confirmatory factor analysis (CFA), researchers can specify the number of factors required in the data and which measured variable is related to which latent variable. It asks how well a proposed model fits a given data.

The differences in both approaches can be summarized as follows:



	Structure-testing	**Structure-discovering**
Process	structure specified (hypotheses); test whether data fit	no assumptions about the data Structure > Structure extracted from data
Statistics	Inductive	Explorative
Goal	Check theory / hypotheses	Form theory / hypotheses (data reduction)
Example	analysis of variance, regression, confirmatory factor analysis Structural Equation Models	exploratory factor analysis, cluster analysis

13.1.1 Prerequisites

Prerequisites of factor analysis include:

- Condition: There are several interval-scaled characteristics (items).

- Rule of thumb: At least 50 people and 3x more people as variables (ideally: 5x more people as variables!).

13.2 Recap slides on EFA and CFA

See the recap slides EFA and EFA:

You can also download the PDF of the slides here:

13.3 Additional optional exercices on CFA

13.3.1 Optional exercise: Anxiety (inspired from J. Lin's UCLA presentation)

We will use a real world example of a questionnaire which Andy Field terms the SPSS Anxiety Questionnaire (SAQ). The first eight items consist of the following:

- Statistics makes me cry
- My friends will think I'm stupid for not being able to cope with SPSS
- Standard deviations excite me
- I dream that Pearson is attacking me with correlation coefficients
- I don't understand statistics
- I have little experience with computers
- All computers hate me
- I have never been good at mathematics

```
library(foreign)
## Warning: le package 'foreign' a été compilé avec la version R 4.3.2
dat <- read.spss(paste0("https://stats.idre.ucla.edu/wp-content/uploads/2018/05/SAQ.sav"),to
# correlations
round(cor(dat[,1:8]),2)
##       q01   q02   q03   q04   q05   q06   q07   q08
## q01  1.00 -0.10 -0.34  0.44  0.40  0.22  0.31  0.33
## q02 -0.10  1.00  0.32 -0.11 -0.12 -0.07 -0.16 -0.05
## q03 -0.34  0.32  1.00 -0.38 -0.31 -0.23 -0.38 -0.26
## q04  0.44 -0.11 -0.38  1.00  0.40  0.28  0.41  0.35
## q05  0.40 -0.12 -0.31  0.40  1.00  0.26  0.34  0.27
## q06  0.22 -0.07 -0.23  0.28  0.26  1.00  0.51  0.22
## q07  0.31 -0.16 -0.38  0.41  0.34  0.51  1.00  0.30
## q08  0.33 -0.05 -0.26  0.35  0.27  0.22  0.30  1.00
# covariances
round(cov(dat[,1:8]),2)
##       q01   q02   q03   q04   q05   q06   q07   q08
## q01  0.69 -0.07 -0.30  0.34  0.32  0.20  0.28  0.24
## q02 -0.07  0.72  0.29 -0.09 -0.10 -0.07 -0.15 -0.04
## q03 -0.30  0.29  1.16 -0.39 -0.32 -0.27 -0.45 -0.24
## q04  0.34 -0.09 -0.39  0.90  0.37  0.30  0.43  0.29
## q05  0.32 -0.10 -0.32  0.37  0.93  0.28  0.36  0.23
## q06  0.20 -0.07 -0.27  0.30  0.28  1.26  0.64  0.22
## q07  0.28 -0.15 -0.45  0.43  0.36  0.64  1.22  0.29
## q08  0.24 -0.04 -0.24  0.29  0.23  0.22  0.29  0.76
```

> interpretation
>
> The interpretation of the correlation table are the standardized covariances between a pair of items. In a correlation table, the diagonal elements are always one because an item is always perfectly correlated with itself.
> In a typical variance-covariance matrix, the diagonals constitute the variances of the item and the off-diagonals the covariances.

We decide the use only Items 1, 3, 4, 5, and 8 as indicators of SPSS Anxiety and Items 6 and 7 as indicators of Attribution Bias. Thus, we will now proceed with a two-factor CFA where we assume uncorrelated (or orthogonal) factors. Having a two-item factor presents a special problem for identification. In order to identify a two-item factor there are two options:

- Freely estimate the loadings of the two items on the same factor but equate them to be equal while setting the variance of the factor at 1
- Freely estimate the variance of the factor, using the marker method for the first item, but covary (correlate) the two-item factor with another factor

Since we are doing an uncorrelated two-factor solution here, we are relegated to the first option.

How does this model compare to a one-factor model?

```
library(lavaan)
## Warning: le package 'lavaan' a été compilé avec la version R 4.3.2
## This is lavaan 0.6-17
## lavaan is FREE software! Please report any bugs.
# one factor model
m1 <- 'f1 =~ q01+ q03 + q04 + q05 + q08
        f2 =~ a*q06 + a*q07
        f1 ~~ 0*f2 '
onefac7items <- cfa(m1, data=dat, std.lv=TRUE)
summary(onefac7items, fit.measures=TRUE, standardized=TRUE)
## lavaan 0.6.17 ended normally after 14 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                        14
##   Number of equality constraints                     1
##
##   Number of observations                          2571
##
## Model Test User Model:
##
##   Test statistic                               841.205
##   Degrees of freedom                                15
##   P-value (Chi-square)                           0.000
##
## Model Test Baseline Model:
##
##   Test statistic                              3876.345
##   Degrees of freedom                                21
##   P-value                                        0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                    0.786
##   Tucker-Lewis Index (TLI)                       0.700
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)             -23684.164
##   Loglikelihood unrestricted model (H1)     -23263.562
##
##   Akaike (AIC)                               47394.328
##   Bayesian (BIC)                             47470.405
##   Sample-size adjusted Bayesian (SABIC)      47429.101
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.146
##   90 Percent confidence interval - lower         0.138
##   90 Percent confidence interval - upper         0.155
##   P-value H_0: RMSEA <= 0.050                    0.000
##   P-value H_0: RMSEA >= 0.080                    1.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.180
##
## Parameter Estimates:
##
```

```
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
## Latent Variables:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   f1 =~
##     q01               0.539    0.017   31.135    0.000    0.539    0.651
##     q03              -0.573    0.023  -24.902    0.000   -0.573   -0.533
##     q04               0.652    0.020   33.032    0.000    0.652    0.687
##     q05               0.567    0.020   27.812    0.000    0.567    0.588
##     q08               0.431    0.019   22.862    0.000    0.431    0.494
##   f2 =~
##     q06        (a)    0.797    0.017   46.329    0.000    0.797    0.710
##     q07        (a)    0.797    0.017   46.329    0.000    0.797    0.723
##
## Covariances:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   f1 ~~
##     f2                0.000                               0.000    0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##    .q01               0.395    0.015   26.280    0.000    0.395    0.576
##    .q03               0.827    0.027   30.787    0.000    0.827    0.716
##    .q04               0.474    0.020   24.230    0.000    0.474    0.527
##    .q05               0.608    0.021   29.043    0.000    0.608    0.654
##    .q08               0.575    0.018   31.760    0.000    0.575    0.756
##    .q06               0.623    0.027   22.916    0.000    0.623    0.495
##    .q07               0.580    0.026   21.925    0.000    0.580    0.477
##     f1                1.000                               1.000    1.000
##     f2                1.000                               1.000    1.000
#uncorrelated two factor solution, var std method
m <- 'f1 =~ q01+ q03 + q04 + q05 + q08
        f2 =~ a*q06 + a*q07
        f1 ~~ 0*f2 '
twofac7items <- cfa(m, data=dat, std.lv=TRUE)
summary(twofac7items, fit.measures=TRUE, standardized=TRUE)
## lavaan 0.6.17 ended normally after 14 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                        14
##   Number of equality constraints                     1
##
```

```
##   Number of observations                          2571
##
## Model Test User Model:
##
##   Test statistic                               841.205
##   Degrees of freedom                                15
##   P-value (Chi-square)                           0.000
##
## Model Test Baseline Model:
##
##   Test statistic                              3876.345
##   Degrees of freedom                                21
##   P-value                                        0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                    0.786
##   Tucker-Lewis Index (TLI)                       0.700
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)             -23684.164
##   Loglikelihood unrestricted model (H1)     -23263.562
##
##   Akaike (AIC)                               47394.328
##   Bayesian (BIC)                             47470.405
##   Sample-size adjusted Bayesian (SABIC)      47429.101
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.146
##   90 Percent confidence interval - lower         0.138
##   90 Percent confidence interval - upper         0.155
##   P-value H_0: RMSEA <= 0.050                    0.000
##   P-value H_0: RMSEA >= 0.080                    1.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.180
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
```

```
##
## Latent Variables:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   f1 =~
##     q01               0.539    0.017   31.135    0.000    0.539    0.651
##     q03              -0.573    0.023  -24.902    0.000   -0.573   -0.533
##     q04               0.652    0.020   33.032    0.000    0.652    0.687
##     q05               0.567    0.020   27.812    0.000    0.567    0.588
##     q08               0.431    0.019   22.862    0.000    0.431    0.494
##   f2 =~
##     q06        (a)    0.797    0.017   46.329    0.000    0.797    0.710
##     q07        (a)    0.797    0.017   46.329    0.000    0.797    0.723
##
## Covariances:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   f1 ~~
##     f2                0.000                               0.000    0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##    .q01               0.395    0.015   26.280    0.000    0.395    0.576
##    .q03               0.827    0.027   30.787    0.000    0.827    0.716
##    .q04               0.474    0.020   24.230    0.000    0.474    0.527
##    .q05               0.608    0.021   29.043    0.000    0.608    0.654
##    .q08               0.575    0.018   31.760    0.000    0.575    0.756
##    .q06               0.623    0.027   22.916    0.000    0.623    0.495
##    .q07               0.580    0.026   21.925    0.000    0.580    0.477
##     f1                1.000                               1.000    1.000
##     f2                1.000                               1.000    1.000
```

Interpretation

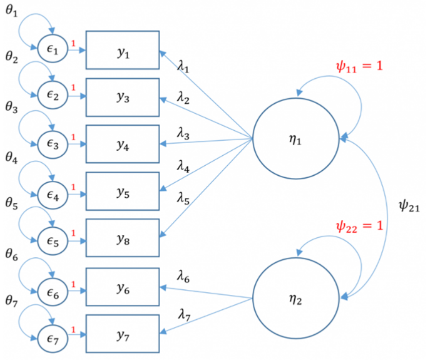
Since we have 7 items, the total elements in our variance covariance matrix is 7(7+1)/2=28. The number of free parameters to be estimated include 7 residual variances , 7 loadings for a total of 14. Then we have 28-14=14 degrees of freedom. However for identification of the two indicator factor model, we constrained the loadings of Item 6 and Item 7 to be equal, which frees up a parameter and hence we end up with 14+1=15 degrees of freedom.
We can see that the uncorrelated two factor CFA solution gives us a higher chi-square (lower is better), higher RMSEA and lower CFI/TLI than the one-factor model, which means overall it is a poorer fitting model.

We decide to go with a correlated (oblique) two factor model:

```
#correlated two factor solution, marker method
m <- 'f1 =~ q01+ q03 + q04 + q05 + q08
        f2 =~ q06 + q07'
twofac7items_n <- cfa(m, data=dat, std.lv=TRUE)
summary(twofac7items_n, fit.measures=TRUE, standardized=TRUE)
## lavaan 0.6.17 ended normally after 18 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                        15
##
##   Number of observations                          2571
##
## Model Test User Model:
##
##   Test statistic                                66.768
##   Degrees of freedom                                13
##   P-value (Chi-square)                           0.000
##
## Model Test Baseline Model:
##
##   Test statistic                              3876.345
##   Degrees of freedom                                21
##   P-value                                        0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                    0.986
##   Tucker-Lewis Index (TLI)                       0.977
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)             -23296.945
##   Loglikelihood unrestricted model (H1)     -23263.562
##
##   Akaike (AIC)                               46623.891
##   Bayesian (BIC)                             46711.672
##   Sample-size adjusted Bayesian (SABIC)      46664.013
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.040
##   90 Percent confidence interval - lower         0.031
##   90 Percent confidence interval - upper         0.050
##   P-value H_0: RMSEA <= 0.050                    0.952
##   P-value H_0: RMSEA >= 0.080                    0.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.021
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
## Latent Variables:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   f1 =~
##     q01               0.513    0.017   30.460    0.000    0.513    0.619
##     q03              -0.599    0.022  -26.941    0.000   -0.599   -0.557
##     q04               0.658    0.019   34.876    0.000    0.658    0.694
##     q05               0.567    0.020   28.676    0.000    0.567    0.588
##     q08               0.435    0.018   23.701    0.000    0.435    0.498
##   f2 =~
##     q06               0.669    0.025   27.001    0.000    0.669    0.596
##     q07               0.949    0.027   35.310    0.000    0.949    0.861
##
## Covariances:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   f1 ~~
##     f2                0.676    0.020   33.023    0.000    0.676    0.676
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##    .q01               0.423    0.014   29.157    0.000    0.423    0.617
##    .q03               0.796    0.026   31.025    0.000    0.796    0.689
##    .q04               0.466    0.018   25.824    0.000    0.466    0.518
##    .q05               0.608    0.020   30.173    0.000    0.608    0.654
##    .q08               0.572    0.018   32.332    0.000    0.572    0.752
##    .q06               0.811    0.030   27.187    0.000    0.811    0.644
##    .q07               0.314    0.040    7.815    0.000    0.314    0.258
##     f1                1.000                               1.000    1.000
##     f2                1.000                               1.000    1.000
```

Interpretation

Compared to the uncorrelated two-factor solution, the chi-square and RMSEA are both lower. The test of RMSEA is not significant which means that we do not reject the null hypothesis that the RMSEA is less than or equal to 0.05. Additionally the CFI and TLI are both higher and pass the 0.95 threshold. This is even better fitting than the one-factor solution. We then choose the final two correlated factor CFA model as shown below:



Chapter 14

Structural Equation Modelling

14.1 Structural equation model (SEM)

CFA is a structure-testing procedure which enable us to answer questions such as: Do my data confirm my previously made hypothetical assumptions about their (factorial) structure?

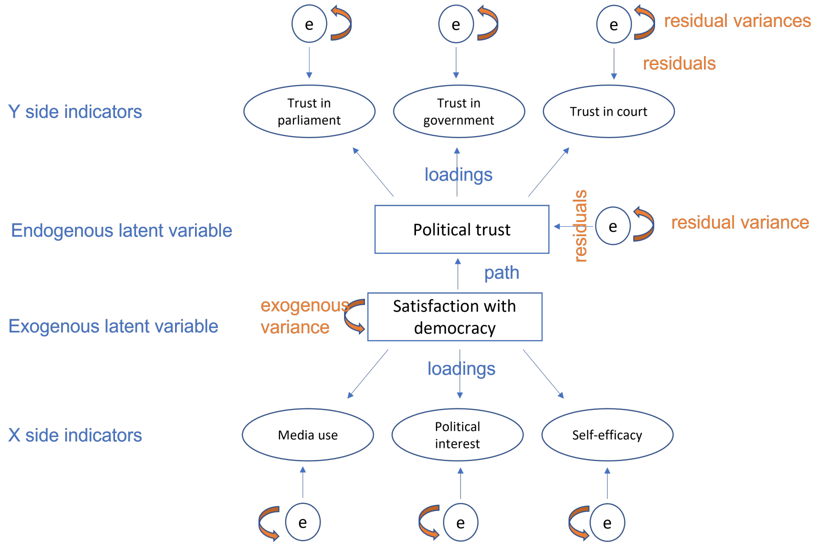
SEM is a structure testing technique. It enables us to answer questions such as: Do my data confirm my previously made hypothetical assumptions about the "causal" structure between variables? Causal structures can be diverse and SEM can flexibly map this diversity. Structural equation models are about testing (sometimes very complex) "causal" structures between (potentially many) variables.

Simply speaking, a structural equation model (SEM) is a combination of confirmatory factor analysis and path analysis.

Structural equation modeling includes two sets of models:

- the measurement model
- the structural model

The measurement model can be expressed as a factor model. The figure below displays a model to measure political trust using three variables - trust in parliament, trust in government and trust in court. It also measures satisfaction with democracy (SWD) using three variables - media use, political interest and feeling of self-efficacy. If one believes that satisfaction with democracy influences political trust, then one can fit a path model. Therefore, a structural model is actually a path model.



14.2 Logic of SEM and similarity with CFA

SEM are hypothesis-testing methods. In this context, previously theoretically deduced considerations on relationships between variables are tested using empirical data. It is thus determined in advance which variables may/should be related and which should not (similar to CFA).

As with the CFA, it is a comparison of covariance matrices that are checked as a whole:

- empirical variance-covariance matrix (= actual correlations between all examined variables in the data)
- model-theoretical variance-covariance matrix (= expected relationships between all variables examined)

The aim is that the model-theoretical matrix is able to reproduce the empirical matrix. SEM allow the testing of entire hypotheses systems in one model.

14.3 Comparison to "normal" regression

SEM follow the intuitive logic of causal models. We can thus assess simultaneous and overall testing of several "causal" relationships in one model instead of testing individual hypotheses in different analyzes.

For instance, linear regression models unrealistically assumes that constructs are perfectly measured, thus confusing measurement error with statistical error (e.g. unexplained variance).

SEM with latent variables allow the measurement error to be extracted from the statistical error. Therefore, SEM come with potentially more exact estimation of the significance and strength of the connections between latent variables.

14.4 Construct validity, model fit, and better explanation of relationships

SEM seeks to find a balance between maximizing the variance between constructs while ensuring the model remains parsimonious and generalizable.

Maximizing the sum of squares between constructs can be useful in several respects:

- Reliability and Validity Assessment: higher variance suggests that the latent variable accounts for a larger proportion of the variation in the observed indicators, which contributes to the construct's reliability and validity.
- Model Fit and Explanation of Relationships: a model that explains a higher proportion of the variance in the observed variables tends to fit the data better.
- Precision in Estimation (robustness of conclusions): higher variance between constructs often results in more precise estimates of the relationships among constructs.

14.5 Quick recap of the concepts

A SEM thus specify causal relationships between exogenous and endogenous variables. Below is a quick recap of the core concepts:

- Observed variable: Exists in data
- Indicator: Observed (exogenous or endogenous)
- Latent variable: Constructed in model
- Factor: Latent (exogenous or endogenous)
- Exogenous: Independent that explains endogenous (observed or latent) (predictor)
- Endogenous: Dependent that has a causal path leading to it (observed or latent) (response)
- Measurement model: Links observed and latent variables
- Loading: Path between indicator and factor
- Regression model: Path between exogenous and endogenous variables

14.6 Latent variables

In the structural model, the ("causal") relationships between different variables are formulated (directed relationships!). These variables can be both latent

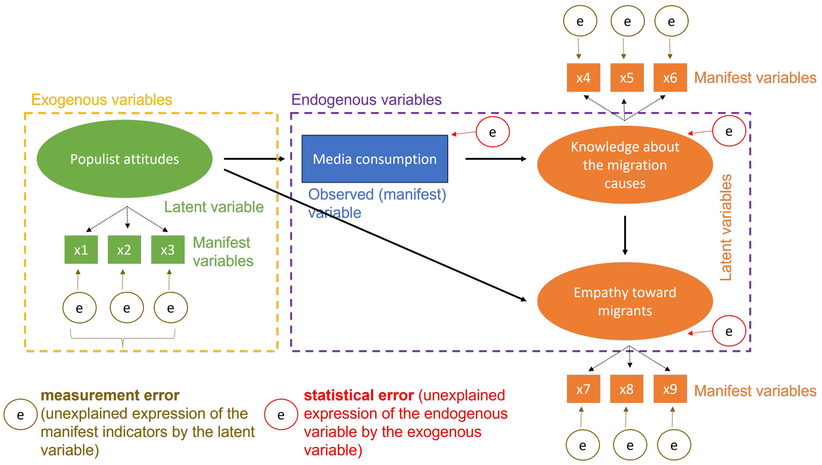
(= hypothetical, not directly observable) and manifest (= directly observable) constructs.

Latent variables are specified in the measurement model (like CFA) and are estimated in such a way that they best represent their respective manifest indicators (items). Manifest constructs (not indicators!) of the structural model are not part of the latent variable measurement model.

14.7 Statistical error versus measurement error

SEM with latent variables allow the distinction between statistical error in estimating the "causal" influences of the exogenous on the endogenous variables (unexplained expression of the endogenous variable by the exogenous variable) and the measurement error of the variable (unexplained expression of the manifest indicators by the latent variable).

Analyzes with only manifest variables (e.g., ANOVA, multiple regression, mediation, path models) mix up the statistical error and the measurement error and implicitly assume perfectly reliable measurements. Since this is unrealistic and there should be measurement errors even with manifest variables (e.g., not perfect reliability), the statistical error (unexplained variance) may be larger than it actually is if the measurement error were included.


e
e
e
x4
x5
x6
Manifest variables
Exogenous variables
Endogenous variables
Populist attitudes
Media consumption
e
Knowledge about the migration causes
e
Observed (manifest) variable
Latent variable
x1
x2
x3
Manifest variables
Latent variables
e
e
e
Empathy toward migrants
e
x7
x8
x9
Manifest variables
e
e
e
e
measurement error (unexplained expression of the manifest indicators by the latent variable)
e
statistical error (unexplained expression of the endogenous variable by the exogenous variable)

14.8 Path models

A path model is a special case of a SEM in which the structural model is present but the measurement model is omitted (exclusively manifest constructs in the structural model).

Path models can also be calculated as SEM in the covariance analysis approach (including model quality measures, etc.), but for better differentiation they are not called SEM but path models.

ℹ More vocabulary

SEM without latent variables (i.e. with structural model but without measurement model), synonymous terms:

- Path analysis
- Path model

IMPORTANT: Do not confuse the terms with the "path diagram" (path diagram = only the graphical representation of relationship structures between variables; path diagrams can be used for SEM with and without latent variables created)
SEM with at least one latent variable (i.e. with a structural model and at least one measurement model) has synonymous terms:

- Structural equation model
- Causal analysis (caution: despite SEM, causality still depends on the data collection method, more later)
- Covariance structure analysis

14.9 Prerequisites

There are several mathematical prerequisites:

- Rule of thumb: sample size K at least 200 (also K - q > 50, where K is the sample size and q is the number of free parameters to be estimated)
- If only latent variables in the structural model:
  - Prerequisites as for CFA
  - Reflective measurement model
  - Interval-scaled and reasonably normally distributed manifest indicator variables
- If partially latent and partially manifest constructs in the structural model:
  - SEM can handle a mixture of manifest and latent constructs in the structural model very well
  - Structural equation system must be identifiable, i.e. the equation system must be mathematically solvable (similar to CFA)

There are also theoretical prerequisites:

- Theoretically secured hypotheses about the connection between constructs in the SEM
  - Especially with cross-sectional data: theoretically reasonable deduced "causal direction" (which constructs are exogenous, which endogenous?)
- Adhere to the confirmatory character of the analysis or, if this is not done, then deal with it transparently

  - Strictly confirmatory testing: If we proceeded strictly confirmatory, we could only specify one model and accept or reject it with a test (hypothesis testing).
  - Alternative models testing: Testing different plausible theories against each other
  - Generating adaptation of the model until the data "fit"

14.10 Identification of the structure

The identification of the model structure consists of two "tasks":

- Establishing a metric for the latent constructs.
- Checking whether there is enough information to estimate the model.

Both steps influence the number of degrees of freedom of the model (model degrees of freedom). This aspect also relates to the complexity of the model. The more complex a model is, the more parameters have to be estimated, and the greater the model's degrees of freedom must be in order to make the estimation possible.

Important: This is not about the sample size, which can also be used to determine "degrees of freedom". Observations do not mean individual cases here, but the number of manifest variables (number whether observed variables).

14.10.1 Establishing metric

The latent variables and error variables to be estimated initially have no metric. In order to be able to interpret the variable later, a scale must be assigned.

One possibility is to choose a reference variable and fix the factor loading to 1. With regard to the latent variable, the "best" indicator variable should be selected (i.e. the variable that is assumed to best index the latent variable). It means that the latent variable is identical to the selected indicator variable except for the measurement error.

In addition, all relationships between the measurement error terms to be estimated and the measured indicator variables are fixed at 1. It means that the measurement errors to be estimated correspond to the observed values except for the influence of the latent variables (everything that is not affected by the latent variable in the indicator variable is explained is a measurement error).

Nota bene: For each fixed parameter, we gain one model degree of freedom (no longer needs to be estimated).

14.10.2 Checking model information

If n manifest variables are collected as part of a project, empirical variances and covariances can be calculated from these variables:

$$p = \frac{n(n+1)}{2}$$

, where p corresponds to the number of non-redundant values in the variance-covariance matrix. It thus corresponds to the available "information" that we use as the basis for calculating all free parameters of our model.

The difference between the available empirical information (p) and the number of (free) parameters to be estimated (q) gives the degrees of freedom of the model (df_M).

$$df_M = p - q$$

If the empirically available information is the same as the number of parameters to be estimated, the number of model degrees of freedom corresponds to zero. The number of model parameters to be estimated must not be less than the number of empirical information given, otherwise a model is not identified (i.e., not solvable). The degrees of freedom of the model must be $>$ or equal to zero for a specified model to be considered identified.

ℹ Degree of freedom: calculation for SEM

Recall that the formula for calculation degrees of freedom is as follows:

$$df = \frac{m(m+1)}{2} - 2m - \frac{X(X-1)}{2}$$

where m represent the number of indicators and X the number of independent latent constructs.
For SEM, this formula includes two additional parameters:

$$df = \frac{m(m+1)}{2} - 2m - \frac{X(X-1)}{2} - g - b$$

where m represent the number of indicators, X the number of independent latent constructs, g the structural relationships from independent to dependent constructs, and b structural relationships from dependent to dependent constructs.
In the above example, we have 9 indicators (m) and 1 independent latent construct (X). Furthermore, we have 2 gamma relationships and 1 beta relationship. Applying the formula, we obtain:

$$df = \frac{9(10)}{2} - 18 - \frac{1(0)}{2} - 2 - 1 = 24$$

14.11 Analytical steps for SEM

- Model Specification - defines hypothetical relationships

- Model Identification:
    - Over identified - more known than free parameters
    - Just-identified - the number of unknown equals the number of free parameters
    - Under-identified - the number of unknown is greater than the number of parameters (model coefficients cannot be estimated)
- Parameter Estimation - comparing actual and estimated covariance (i.e. maximum likelihood estimate)
- Model Evaluation- goodness of fit (i.e. Chi-square, Akaike Information Criterion, Comparative Fit Index)
- Model Modification - post hoc model modification

14.11.1 Chi-square test

The Chi-Square test poses that:

- H0: empirical covariance matrix is equal to the model-theoretical covariance matrix
- H1: empirical covariance matrix is not equal to the model-theoretical covariance matrix

The smaller the difference between the two matrices, the smaller the chi-square value. So, the smaller the chi-square is, the better it is. The test should not turn out to be significant, since equality between the empirical and model-theoretical variance-covariance matrix is desirable.

14.11.2 RMSEA (Root Mean Squared Error of Approximation)

RMSEA uses inferential statistics to check whether a model can approximate reality well (approximation test). It is therefore not about the absolute correctness of the model, as with the Chi-Square test, but about evaluating the best possible approximation.

Recommendations:

- RMSEA $<$ .05 good fit (also test of the characteristic value: p close $>$ .1, means that RMSEA is clearly not significantly larger than RMSEA $<$ .05)
- RMSEA $<$ .08 acceptable fit (also test of the parameter: .05 $<$ p close $<$ .1 means that RMSEA does not tend to be significantly larger than RMSEA $<$ .05)

14.11.3 SRMR (Standardized Root Mean Squared Residual)

Usually there is a discrepancy between the empirical and model-theoretical covariance matrix. Therefore, descriptive measures of discrepancy are often used. These give an answer to the question of whether an existing discrepancy can be neglected and also includes the model complexity. If the empirical and model-theoretical covariance matrix are completely identical, this measure assumes a value of zero.

Recommendations:

- SRMR $< .05$ good fit
- SRMR $< .10$ acceptable fit

14.12 Individual components

Have the individual components of the SEM (entire measurement model AND structural model) also been identified? For structural models, this depends heavily on the structure:

- Recursive models (= models without repercussions between endogenous variables, i.e. directed models) are always identified
- Non-recursive models (= models with repercussions between endogenous variables, i.e. undirected models) are difficult to determine by hand, you should leave that to the statistics program

14.13 Critics to SEM

SEM seems to test causal relationships between exogenous and endogenous variables and is even called causal analysis. But, ultimately, only relationships between variables become statistical examined and the researcher decides (often) what are exogenous and what are endogenous variables.Directions of impact could be directed in the opposite direction as long as the method of data collection is cross-sectional. Causal evidence using SEM is only possible with panel data or in an experiment (and then not mandatory for all paths of the structural model).

In practice, SEM are too often not tested for confirmation, but adjusted until the model fit is “good”, without this being reported transparently. The more paths between variables are freely estimated in the structural model, the less SEM test theoretical ideas. More parsimonious models in which more parameters are fixed a priori (e.g., direct effects in SEM mediation models set to 0) test assumptions, but if everything is allowed to be related to everything (i.e., freely estimated), then there are no assumptions more that are tested, just a SEM that fits the data.

Interesting resource: Shah, R., & Goldstein, S. M. (2006). Use of structural equation modeling in operations management research: Looking back and forward. Journal of Operations management, 24(2), 148-169. https://doi.org/10.1016/j.jom.2005.05.001

14.14 How it works in R?

See the lecture slides on SEM:

You can also download the PDF of the slides here:

14.15 Example from the literature

The following article relies on SEM as a method of analysis:

Gil de Zúñiga, H., González-González, P., & Goyanes, M. (2021). Pathways to political persuasion: Linking online, social media, and fake news with political attitude change through political discussion. American Behavioral Scientist, 00027642221118272. Available here.

Please reflect on the following questions:

- What is the research question of the study?
- What are the research hypotheses?
- Is SEM an appropriate method of analysis to answer the research question?
- What are the main findings of the SEM?

14.16 Time to practice on your own

You can download the PDF of the exercises here:

14.16.1 SEM with multiple latent variables

We want to create a structural model that relates multiple latent variables using social science data describing the effect of student background on academic achievement.

We create a measurement model by defining each latent variable:

- Adjustment: measured by motivation, harmony, and stability.
- Risk: measured by verbal, negative parent psychology, and socioeconomic status.
- Achievement: measured by reading, arithmetic, and spelling.

We also define the regression path where achievement will be the combination of adjustment and risk.

```
dat <- read.csv("https://stats.idre.ucla.edu/wp-content/uploads/2021/02/worland5.csv")
mod <- '
```

```
  # measurement model
  adjust =~ motiv + harm + stabi
  risk =~ verbal + ppsych + ses
  achieve =~ read + arith + spell
  # regression
  achieve ~ adjust + risk
'
```

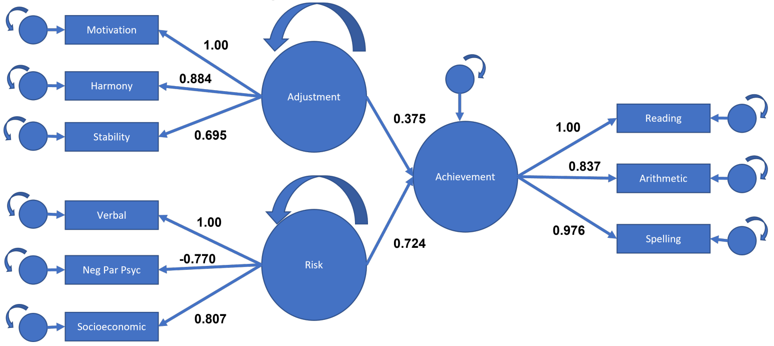
Next, we fit the model to the data using sem() from lavaan. Followed by a summary, including model fit.

```
fit <- lavaan::sem(mod, data=dat)
summary(fit, fit.measures=TRUE)
## Length   Class    Mode
##       1 lavaan      S4
```

How would you illustrate this analysis using a path diagram?

Solution

We obtain the following results:



Note that further analysis using these data can be found in the Introduction to structural equation modeling (SEM) in R with Lavaan from the UCLA: Statistical Consulting Group.

14.16.2 SEM framework

Based on the following flowchart (original can be found here), state whether the following statements are true or false:

- SEM encompasses a broad range of linear models and combines simultaneous linear equations with latent variable modeling.

Solution

True: Multivariate regression and path analysis are simultaneous equations of observed variables; factor analysis is a latent variable model, and structural regression combines the concepts of path analysis with factor analysis.

- Multivariate regression means that there is always more than one exogenous predictor in my model.

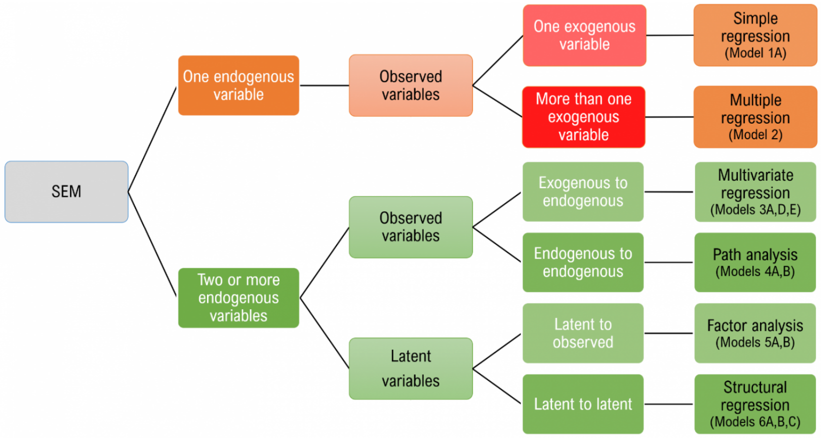
Solution

False: Multivariate regression indicates more than one endogenous variable. You can certainly have only one exogenous predictor of multiple endogenous variables.

- Structural regression models the regression paths only among latent variables.

Solution

True: Structural regression defines relationships between latent variables and path analysis defines relationships between observed variables.


SEM
One endogenous variable
Observed variables
One exogenous variable
Simple regression (Model 1A)
More than one exogenous variable
Multiple regression (Model 2)
Two or more endogenous variables
Observed variables
Exogenous to endogenous
Multivariate regression (Models 3A,D,E)
Endogenous to endogenous
Path analysis (Models 4A,B)
Latent variables
Latent to observed
Factor analysis (Models 5A,B)
Latent to latent
Structural regression (Models 6A,B,C)

Chapter 15

Multilevel modelling

15.1 Multilevel regression analysis

Many research topics have multilevel structured data which consist of multiple macro and micro units within each macro unit (e.g. individuals within countries, individuals within occupations, children within classes within schools, etc). Therefore, at each level there are both mean characteristics (fixed effects) and differences (random effects).

Single level regression modelling is not the appropriate method to use here. For instance, the effect of an X within clusters can be different from the effect between clusters. Indeed, single level regression does not consider the nested data structure, which may violate the uncorrelated errors assumption. It is likely that in a "rich" country all people have higher wages than in "poor" countries? If errors are correlated, this is likely to cause the following problems:

- efficiency of estimators is low
- standard errors are low
- coefficients are often significant

It will also be unclear:

- how much variation in Y is at each level?
- how much the context impacts on individual level after controlling for other relevant factors?

The purpose of multilevel modelling (MLM) is to correct for biased estimates resulting from clustering and to provide correct standard errors, confidence intervals, as well as significance tests. It will thus decompose the total variance of Y into portions associated with each level (e.g. individual vs country).

We might want to know: How much variation is there in individual wages between and within countries? Which countries have particularly low and high wages ?

15.1.1 Fixed versus random effects

In order to understand MLM we need to understand random effects, and to understand random effects we need to understand variance. We often need to think more about where the variance in our system is showing up in our model. It allows us to decompose the variance of the dependent variable into:

- within-context variance
- between-context variance

So far, we are familiar with the residual variance from OLS, but might that residual variance be better attributed to within a group? Or between a group?

Recall that a factor is a categorical predictor that has two or more levels. Up to this point (e.g. in ANOVAs) we have only talked about fixed factors which assumes that the levels are separate, independent, and not similar. Fixed effects estimate separate levels with no relationship assumed between the levels. Fixed effects also assume a common variance known as homoscedasticity. Post-hoc adjustments are needed to do pairwise comparisons of the different factor levels. Random effects means that each level can be thought of as a random variable from an underlying process or distribution. Estimation of random effects provides inference about the specific levels (similar to a fixed effect), but also population level information (think about it as if each level of the effect is a draw from a random variable).

15.1.2 Example

Let's image that we have 10 people (10 levels in our model) in a study about time spent to read the news on a daily basis over a 5-day period. Each day, we ask people to report how much time they spent reading the news (n=10*5=50) so that each person has 5 observations.

- A fixed effect model enables us to estimate the means of the 10 individuals and assumes that each of the individuals has a common variance around their news reading time (variance is the same or similar for everyone in the study).
- A random effect model enables us to estimate the mean and variance of the participants and to make a reasonable prediction about others that were not enrolled in the study (amount of time spent to read the news within a given individual is much likely to be similar than compared to someone else).

15.1.3 When not to use random effects

You might not want to use random effects when the number of factor levels is very low (there is not definitive recommendation here). Furthermore, it is commonly reported that you may want five or more factor levels for a random effect in order to really benefit from what the random effect can do. Note that a group with a large sample size and/or strong information (e.g. a strong relationship) will have very little influence of the grand mean and largely reflect the information contained entirely within the group.

Another case in which you may not want to random effect is when you do not assume that your factor levels come from a common distribution.

We already know about the OLS model (labelled "fixed effects" in the figure below). The next figure displays different types of random effects:

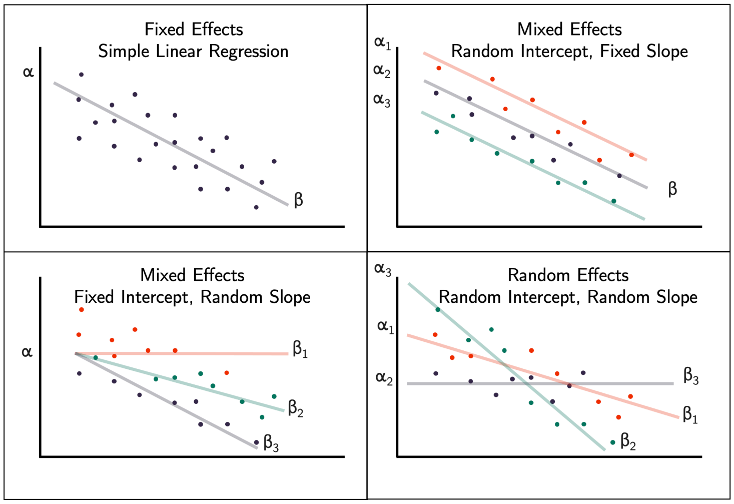


Figure 15.1: Source: https://bookdown.org/steve_midway/DAR/random-effects.html

15.2 Fixed effect regression approach

One might account for the nested structured of the data by including a "grouping variable" for individuals (dummy for each country or year) or inclusion of contextual explanatory variables (e.g. gender equality index per country).

$$y_{ij} = \beta_0 + \beta_1 * X_{ij} + \sum_{C-1} (\beta_C * CD_j + \epsilon_{ij})$$

where, β_0 is the overall intercept (e.g. reference country or year) and β_C is the intercept for one country. The problem is that grouping parameters are treated as fixed effects ignoring random variability associated with macro-level characteristics. This "dummy variable approach" thus suggests that group differences are fixed effects. What if number of groups is very large? What about including group-level predictors as all degrees of freedom at group-level have been consumed by country dummy?

15.3 Random intercept model (with fixed slope)

The next model we will examine is the MLM with a random intercept and fixed slope.

15.3.1 Null model

In its simplest form, a MLM is in the form of a "null model" which contains no explanatory variables. It thus includes one regression constant (intercept) and assumes that this varies across contexts. The intercept is a random variable.

So far, we are able to answer questions like: Are there countries differences with respect to happiness? How much of this variation is due to these country differences?

15.3.2 Interclass Correlation Coefficient (ICC)

A MLM with a random intercept is a point where we can pause and run a diagnostic called the Interclass Correlation Coefficient (ICC).

$$ICC = \frac{\sigma_\alpha^2}{\sigma^2 + \sigma_\alpha^2}$$

The ICC tells us how much similarity is within contexts (i.e. countries). Basically, it accounts for the closeness of observations in same context relative to closeness of observations in different contexts.

For instance, it can tell us how much % of variance in Y (e.g. happiness) can be explained by context (belonging to a different country).

The ICC ranges from 0 (no clustering/single level data structure) and 1 (maximum clustering). In practice, 0 or 1 rarely occur. A general recommendation is that if the ICC is small, then use of a single level model.

15.3.3 Random intercept model (with fixed slope) including one predictor

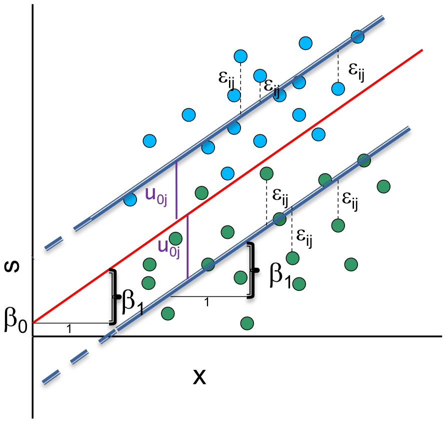
At this stage, it is unclear which factors account for the variation. Are there more reasons why individuals and countries might differ with respect to happiness?

Random intercept model is a combination of variance component and a linear model (for a person i in country j). We thus have a fixed part ($\beta_0 + \beta_x *$

X_{ij}), where the estimated parameters are the beta coefficients and are the same for each observation in the sample. We also have a random part ($u_{0j} + \epsilon_{ij}$), where estimated parameters are variances which are allowed to vary (e.g. across countries).

$$y_{ij} = \beta_0 + \beta_x * X_{ij} + u_{0j} + \epsilon_{ij}$$

In this model, there are two random terms and therefore two types of residuals: u_j a the level-2 and e_{ij} at the level-1. There are also an overall (average) line β_0 and group (average) lines $\beta_0 + u_j$.



So far, we are able to answer questions as: Do country differences in happiness remain after controlling for gender? How much of variation in happiness is due to country differences after controlling for gender? What is the relationship between an individual's happiness and their gender?

15.3.4 Model assumptions

The model assumptions are that individual and group-level residuals are normally distributed. Furthermore, residuals at the same/different level are uncorrelated.

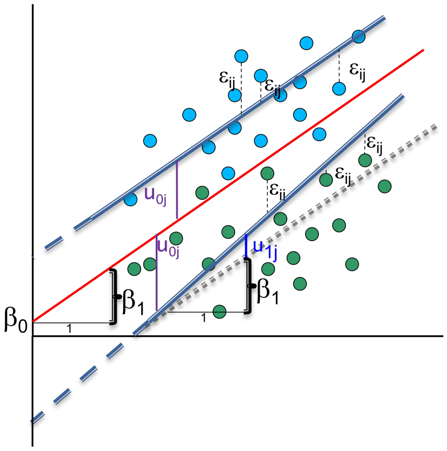
15.4 Random slope model (with random intercept)

The assumption of random intercept model suggests that group lines have all same slope as overall regression line in every group (effect of X on Y is the same for every country). Is this always valid? It can be argued that X is not a fixed but a random effect and that its slope can vary across groups.

We can account for the random slope by adding random term to coefficient of x_{1ij} (e.g. working hours), so it can be different for each group (e.g. country):

$$y_{ij} = \beta_0 + (\beta_1 + u_{1j}) * x_{1ij} + u_{0j} + \epsilon_{0ij}$$

Note that one extra parameter u_{1ij} leads to two extra parameters σ^2_{u1} (variance in slopes between groups) and σ^2_{u01} (covariance between intercepts and slopes).



With a random slope model, we could now answer questions as: Does the effect of being employed has the same impact on happiness across countries?

Useful reference:

- Sommet, N. & Morselli, D. (2021). Keep Calm and Learn Multilevel Linear Modeling: A Three-Step Procedure Using SPSS, Stata, R, and Mplus. *(International Review of Social Psychology)*, 34(1). Available here.

15.4.1 Note on fixed intercept with random slope model

It might be the case that a model requires a fixed intercept but a random slope specification. This could be the case when we hypothesize that the effect of our predictor differs among groups, but when we want to fix the intercept because we know that all groups start in a similar place.

15.4.2 Testing the random part

The likelihood ratio test (LRtest) can be used to compare the random slope model to a random intercept model. The null hypothesis states that σ^2_{u1} and σ^2_{u0} should equal 0. If this is true, the random intercept model is more appropriate than a random slope model. If random intercept/slope coefficients are not significantly different from zero, this suggests that there is not much random variability in the slope/intercept. Therefore, there is no need to specify random parameter.

15.5 Cross-level interactions

Adding cross-level interactions allows us to assess whether contextual factors (Z) influence the effect of level-1 X variables. For instance, we can answer questions such as: Is the effect of gender on happiness stronger/weaker in countries with a high gender equality index?

$$y_{ij} = \beta_0 + \beta_1 * x_{ij} + \beta_2 * Z_j + \beta_3 x_{ij} * Z_j + u_{0j} + u_{1j} x_{ij} + \epsilon_{ij}$$

The cross-level interaction should be included in the fixed part of model. Therefore, the direct main and interaction effects have to be interpreted together (similar to OLS): the β_3 coefficient indicates the impact of each unit change in Z on the slope β_1.

Useful links to external data sources to link at the macro-level are:

- Eurostat: http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/
- OECD: http://www.oecd.org/home
- Worldbank: http://data.worldbank.org/
- ILO: http://www.ilo.org/global/statistics-and-databases/lang--en/index.htm
- Databanks: http://www.gapminder.org/ and http://www.statsilk.com/

15.6 Variance decomposition approach

To assess how much variance is explained by a model, simple OLS regression entail the R^2 statistic (proportion of explained variance). In MLM, there are several several (co)variances.

Hox (2010, pp.70) proposes to examine residual error variance in a sequence of models. This approach suggests examining the residual error variances in a sequence of models:

- intercept-only model (since there are no explanatory variables in the model, it is reasonable to interpret variances as the error variances)
- model including the level-1 predictors
- model including the level-2 predictors
- model with random coefficient
- model with cross-level interaction

Other useful references:

- Clarke, P., Crawford, C., Steele, F. & Vignoles, A. (2010). The choice between fixed and random effects models: some considerations for educational research. Institute of Education DoQSS, Working Paper No. 10. Available here.

- Moehring, K. (2012). The fixed effect as an alternative to multilevel analysis for cross-national analyses. GK Soclife working paper. Available here

15.7 Centering and standardizing

The regression intercept equals the expected value of Y if all X are 0. However, what if 0 has no useful meaning or is not possible? (e.g. age of 0). In this case, it is useful to transform the X variables, for instance by centering them. There are two options:

- Grand mean centering: computing variables as deviations from the overall mean (the constant in model reflects the mean of all cases)
- Group mean centering: computing variables as deviation from the group mean (decomposes within vs. between effects)

15.8 Concluding remarks

Ultimately the application of MLM with fixed or random effects is done to capture more realism in the phenomenon we are seeking to describe with our model. While all models are to some degree incorrect, inaccurate estimation or pooling dissimilar information may extend the degree to which the model is inaccurate or misleading.

15.9 How it works in R?

See the lecture slides on MLM:

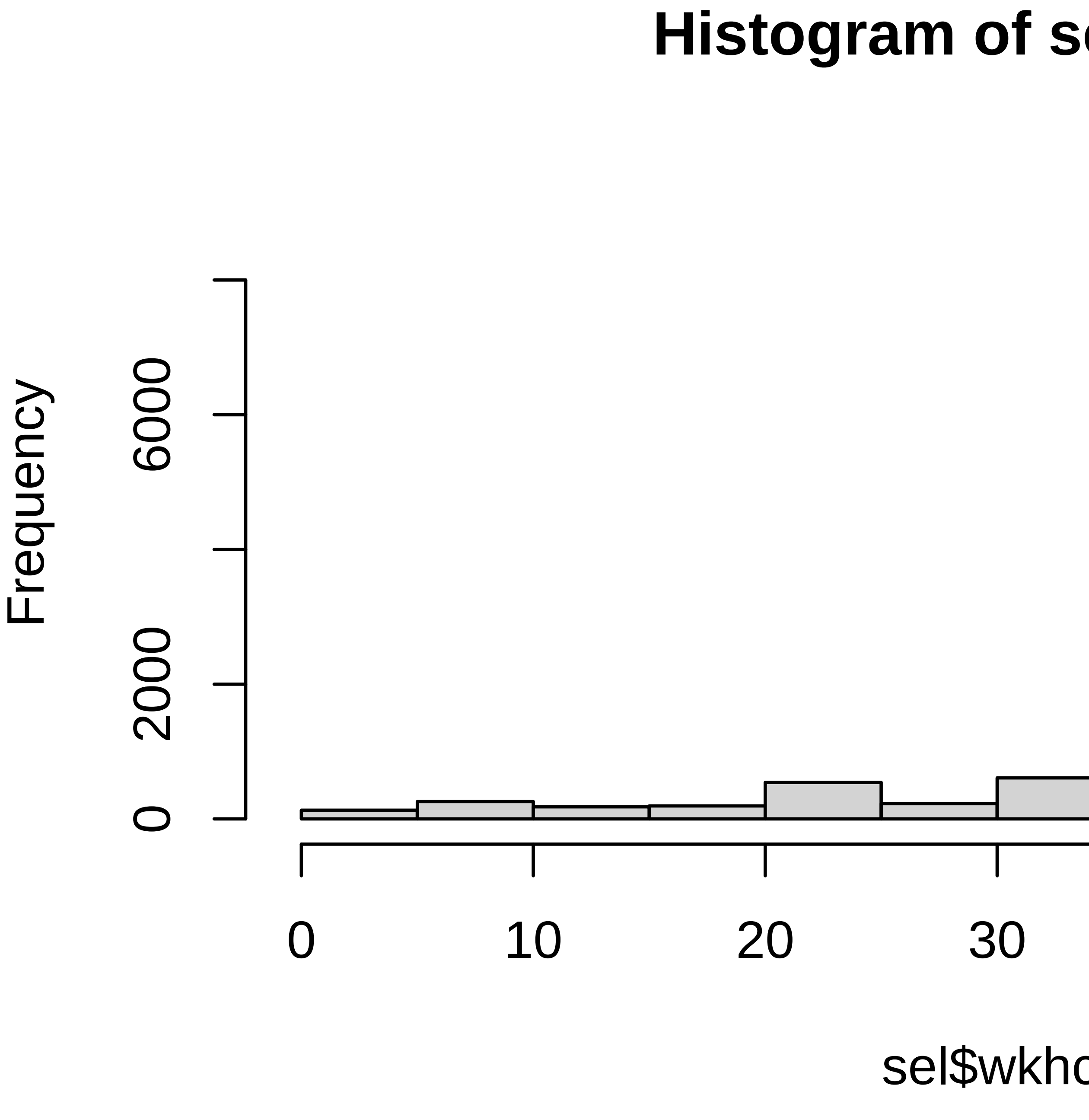
You can also download the PDF of the slides here: PDF

15.10 Time to practice on your own

In the following example, we would like to assess cross-national differences in the gender gap in working time.

First thing you want is to download the data from the European Social Survey and select the variables: country information, working time, gender and additional explanatory variables (e.g. level of education, having children, etc). Then, you want to know how the variables are coded and to apply the necessary changes. You can also filter outliers (e.g. very high working hours).

```
library(foreign)
db <- read.spss(file=paste0(getwd(),"/data/ESS10.sav"),
                use.value.labels = T,
                to.data.frame = T)
sel <- db |>
  dplyr::select(cntry,wkhct,gndr,eduyrs,chldhhe) |>
  stats::na.omit()
# drop levels
sel$cntry <- droplevels(sel$cntry)
sel <- sel[complete.cases(sel),]
# recodings
sel$gndr <- ifelse(sel$gndr=="Female",1,0)
sel$wkhct <- as.numeric(sel$wkhct)
sel$eduyrs <- as.numeric(sel$eduyrs)
sel$cntry <- as.factor(sel$cntry)
# filtering
sel <- sel[sel$wkhct<60,]
hist(sel$wkhct)
```


Histogram of se
Frequency
0
2000
6000
0
10
20
30
sel$wkho

Second, you start by conducting a fixed effects model.

```
dummy_model <- lm(wkhct ~
                    gndr +
                    cntry,
                  data=sel)
summary(dummy_model)
##
## Call:
## lm(formula = wkhct ~ gndr + cntry, data = sel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.958  -0.958   1.287   3.350  24.145
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          41.89245    0.22431 186.762  < 2e-16 ***
## gndr                 -2.31631    0.13146 -17.620  < 2e-16 ***
## cntrySwitzerland     -3.28260    0.35072  -9.360  < 2e-16 ***
## cntryCzechia         -0.56063    0.32625  -1.718 0.085746 .
## cntryEstonia         -1.55739    0.34830  -4.471 7.83e-06 ***
## cntryFinland         -4.52794    0.32602 -13.889  < 2e-16 ***
## cntryFrance          -4.72122    0.33860 -13.943  < 2e-16 ***
## cntryGreece           1.15618    0.33480   3.453 0.000555 ***
## cntryCroatia          0.13675    0.37672   0.363 0.716614
## cntryHungary          0.49649    0.33865   1.466 0.142649
## cntryIceland         -5.13347    0.45526 -11.276  < 2e-16 ***
## cntryItaly           -2.60405    0.33088  -7.870 3.78e-15 ***
## cntryLithuania       -0.99921    0.34479  -2.898 0.003761 **
## cntryMontenegro       0.06515    0.90505   0.072 0.942617
## cntryNorth Macedonia  0.25775    0.41787   0.617 0.537364
## cntryNetherlands     -8.17255    0.36731 -22.250  < 2e-16 ***
## cntryNorway          -6.24267    0.37639 -16.586  < 2e-16 ***
## cntryPortugal        -1.20709    0.35048  -3.444 0.000575 ***
## cntrySlovenia        -1.14729    0.39832  -2.880 0.003978 **
## cntrySlovakia        -0.14967    0.37599  -0.398 0.690593
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.059 on 15183 degrees of freedom
## Multiple R-squared:  0.1014, Adjusted R-squared:  0.1002
## F-statistic: 90.14 on 19 and 15183 DF,  p-value: < 2.2e-16
```

Interpretation

From the output, we see that gender has a negative effect on working hours (women work on average less hours than men). We also note some country differences. The R-squared is 0.09, which means that the model explains ~9% of the overall variance.

Third, you conduct a “null” model with only working hours. This allows you to assess how much variation in working hours can be attributed to individual differences and country differences.

```
empty_model <- lme4::lmer(wkhct ~ 1 + (1 | cntry),
                  data=sel)
stargazer::stargazer(empty_model, type="text", single.row = T)
##
## ===============================================
##                          Dependent variable:
##                      ---------------------------
##                                 wkhct
## -----------------------------------------------
## Constant                  38.691*** (0.599)
## -----------------------------------------------
## Observations                     15,203
## Log Likelihood                 -53,490.590
## Akaike Inf. Crit.              106,987.200
## Bayesian Inf. Crit.            107,010.100
## ===============================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
# ICC
performance::icc(empty_model)
## # Intraclass Correlation Coefficient
##
##     Adjusted ICC: 0.092
##   Unadjusted ICC: 0.092
```

Interpretation

The ICC can be interpreted as the proportion of the variance explained by the grouping structure in the population. Here, the ICC tells us that ~8% of the variance can be attributed to country differences. The average expected working hours for a person is ~39 hrs across all countries.

Fourth, you can conduct a random intercept model with gender, and compare it with a random intercept with individual level variables (having children, years of education, etc.).

```
# gender only
random <- lme4::lmer(wkhct ~ gndr + (1 | cntry),
                  data=sel)
stargazer::stargazer(random, type="text", single.row = T)
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                                wkhct
## -----------------------------------------------
## gndr                     -2.315*** (0.131)
## Constant                 39.874*** (0.606)
## -----------------------------------------------
## Observations                  15,203
## Log Likelihood              -53,338.130
## Akaike Inf. Crit.           106,684.300
## Bayesian Inf. Crit.         106,714.800
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
# add more variables
random2 <- lme4::lmer(wkhct ~ gndr + eduyrs + chldhhe + (1 | cntry),
                  data=sel)
stargazer::stargazer(random2, type="text", single.row = T)
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                                wkhct
## -----------------------------------------------
## gndr                     -2.382*** (0.132)
## eduyrs                   -0.061*** (0.018)
## chldhheNo                -1.013*** (0.137)
## Constant                 41.221*** (0.643)
## -----------------------------------------------
## Observations                  15,203
## Log Likelihood              -53,302.970
## Akaike Inf. Crit.           106,617.900
## Bayesian Inf. Crit.         106,663.700
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

You can also assess the variance between countries:

```
as.data.frame(lme4::VarCorr(random))
##        grp        var1 var2      vcov    sdcor
## 1    cntry (Intercept) <NA>  6.778696 2.603593
## 2 Residual        <NA> <NA> 64.944980 8.058845
```

Interpretation

The average working hours is ~40 hrs across all countries. It varies by a score of ~7 between countries.

You can also test whether it makes sense to rely on a random intercept instead of the fixed effects model.

```
anova(random, dummy_model)
## refitting model(s) with ML (instead of REML)
## Data: sel
## Models:
## random: wkhct ~ gndr + (1 | cntry)
## dummy_model: wkhct ~ gndr + cntry
##             npar    AIC    BIC logLik deviance  Chisq Df Pr(>Chisq)
## random         4 106683 106713 -53337   106675
## dummy_model   21 106617 106777 -53287   106575 100.25 17      8e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation

The comparison of both models using the anova() function tells us that the random intercept model is better than the fixed effects model to explain working hours.

Fifth, we can conduct a random slope model with gender:

```
randomslope <- lme4::lmer(wkhct ~ 1 + gndr +
                            (1 + gndr| cntry),
                          data=sel)
stargazer::stargazer(randomslope, type="text", single.row = T)
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                                wkhct
## -----------------------------------------------
## gndr                     -2.317*** (0.467)
## Constant                 39.849*** (0.443)
## -----------------------------------------------
## Observations                  15,203
## Log Likelihood              -53,256.100
## Akaike Inf. Crit.          106,524.200
## Bayesian Inf. Crit.         106,570.000
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

You can also assess the variance between countries:

```
as.data.frame(lme4::VarCorr(randomslope))
##        grp        var1 var2      vcov     sdcor
## 1    cntry (Intercept) <NA>  3.532110 1.8793908
## 2    cntry        gndr <NA>  3.748614 1.9361337
## 3    cntry (Intercept) gndr  2.320980 0.6378506
## 4 Residual        <NA> <NA> 64.133868 8.0083624
```

Interpretation

Average working hours for a person is ~40 hrs across all countries (varies by a score of ~3.5 between countries) Gender decreases the working hours on average by ~2.3 (varies by a score of ~3.7 between countries). Furthermore, the correlation between the intercept and slope is positive (~2.3).

You can also test whether it makes sense to rely on a random slope instead of a fixed slope model.

```
# anova(randomslope, random)
```

Finally, you can add a cross-level interaction using the information provided by the Gender equality Index.

```
sel$cntry = droplevels(sel$cntry)
sel$cntry = as.numeric(sel$cntry)
sel$GEI = NA
sel$GEI[sel$cntry==1] = 58.8
sel$GEI[sel$cntry==2] = NA
sel$GEI[sel$cntry==3] = 55.7
sel$GEI[sel$cntry==4] = 59.8
sel$GEI[sel$cntry==5] = 73.4
sel$GEI[sel$cntry==6] = 74.6
sel$GEI[sel$cntry==7] = 51.2
sel$GEI[sel$cntry==8] = NA
sel$GEI[sel$cntry==9] = 55.6
sel$GEI[sel$cntry==10] = NA
sel$GEI[sel$cntry==11] = 63
sel$GEI[sel$cntry==12] = 55.5
sel$GEI[sel$cntry==13] = NA
sel$GEI[sel$cntry==14] = NA
sel$GEI[sel$cntry==15] = 72.1
```

```
sel$GEI[sel$cntry==16] = NA
sel$GEI[sel$cntry==17] = 59.9
sel$GEI[sel$cntry==18] = 68.3
sel$GEI[sel$cntry==19] = 54.1
# cross level interaction term
# random slope with cross-level interaction
randomslope2 <- lme4::lmer(wkhct ~ 1 + gndr + GEI +
                             gndr*GEI +
                            (1 + gndr| cntry),
                          data=sel)
stargazer::stargazer(randomslope2, type="text", single.row = T)
##
## ===============================================
##                          Dependent variable:
##                      ---------------------------
##                                 wkhct
## -----------------------------------------------
## gndr                      6.325** (3.095)
## GEI                      -0.208*** (0.035)
## gndr:GEI                 -0.135*** (0.050)
## Constant                 52.664*** (2.164)
## -----------------------------------------------
## Observations                  12,020
## Log Likelihood              -41,711.130
## Akaike Inf. Crit.           83,438.260
## Bayesian Inf. Crit.         83,497.420
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Interpretation

The GEI has a negative impact on working hours. Furthermore, the cross-level interaction of between GEI and gender has also a negative impact.

Part III

EXAM PREPARATION

Chapter 16

Exam preparation

16.1 Mock-exam

On this page, you will find a mock-exam:

Please find here the mock-exam: PDF

Please also find here the solutions: PDF

16.2 Additional examples of questions

Here are additional examples to train for the exam:

Please find additional questions here: PDF

Please also find here the solutions: PDF

According to the semester schedule, questions about the exam (as well as any question related to the course content) will be discussed during the "Recap" session on December 12th.

16.3 SEB mock-exam

Please consider that on December 12th (from 7:00 to 22:00) you have the opportunity to take the mock-exam on site. To do so, open the SEB mock-exam link and complete the mock-exam. The SEB mock-exam link will be published here a few days before the mock-exam. It is recommended to take the mock-exam in order to:

- do the short-term functional test
- test your access to the exam (whether the login works or the access to the exam course is possible)
- check the examination environment (exam navigation, layout, etc.)

16.4 How to prepare for the exam?

The exam will focus on the content of the slides of each chapter. The examples from the literature and the practical exercises will also help you to gain a better understanding of how to apply the statistical methods.

The exam covers the material from the course on "Multivariate Statistics" and lasts 50 minutes.

The exam consists of 7 sections:

- linear regression
- ANOVA and repeated measures ANOVA
- logistic regression
- moderation analysis and mediation analysis
- EFA and CFA
- SEM
- multilevel modelling

16.5 Question types

The exam entails three types of question:

- single choice questions: there are four answers to each question, one of which is correct (max 1 point awarded, and should take between 1 and 2 minutes to answer)
- Kprim questions: there are four statements/answers for every Kprim question and you must decide for each of these statements/answers whether it is correct or false (4 right decisions per question will give you 2 points, 3 right answers will lead to 1 point, 2 or less right answers will give you 0 points (max 2 points awarded, and should take between 1.5 and 3 minutess to answer)
- open-questions (between 1 and 3 points awarded, and should take twice the time as the number of points to answer)

16.6 What is not part of the exam

The following aspects will not be part of the exam:

- You will not write R code/syntax during the exam
- You will not be asked to interpret R code/syntax during the exam
- You will not write mathematical demonstrations during the exam

Part IV

ADDITIONAL RESSOURCES

Chapter 17

Longitudinal data analysis

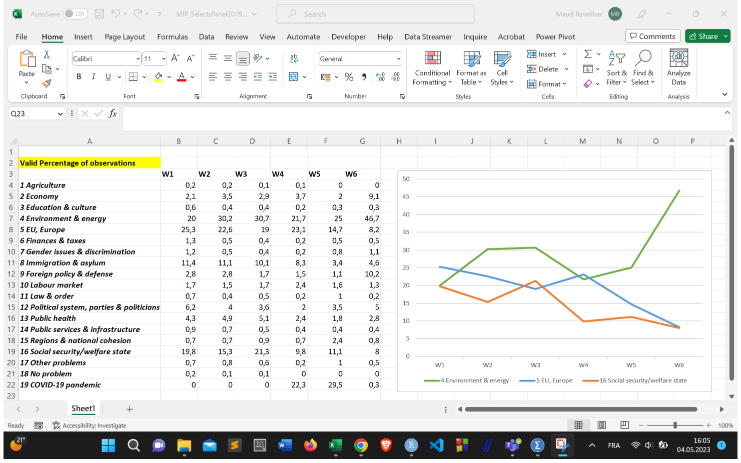
17.1 Longitudial data analysis with panel data

A panel is usually denoted by having multiple entries (rows) for the same entity (e.g. respondent, company, etc) in a dataset. The multiple entries are due to different time periods at which the entity was observed.

17.2 The problem with OLS

OLS can be used to pool observations of the same entity recorded at different time points. However, observations of the same entity are then treated as if they originate from other entities. Important influences like serial correlation of observations within the same entity cannot be considered, leading to biased estimates.

For instance, we might be interested in knowing what factors affect people's change in their opinion about what is the most important problem facing the country. The next figure displays the change in the proportion of citizens' concerns about the most important policy issues during the Swiss 2019 federal election campaign.



Valid Percentage of observations

	W1	W2	W3	W4	W5	W6
1 Agriculture	0,2	0,2	0,1	0,1	0	0
2 Economy	2,1	3,5	2,9	3,7	2	9,1
3 Education & culture	0,6	0,4	0,4	0,2	0,3	0,3
4 Environment & energy	20	30,2	30,7	21,7	25	46,7
5 EU, Europe	25,3	22,6	19	23,1	14,7	8,2
6 Finances & taxes	1,3	0,5	0,4	0,2	0,5	0,5
7 Gender issues & discrimination	1,2	0,5	0,4	0,2	0,8	1,1
8 Immigration & asylum	11,4	11,1	10,1	8,3	3,4	4,6
9 Foreign policy & defense	2,8	2,8	1,7	1,5	1,1	10,2
10 Labour market	1,7	1,5	1,7	2,4	1,6	1,3
11 Law & order	0,7	0,4	0,5	0,2	1	0,2
12 Political system, parties & politicians	6,2	4	3,6	2	3,5	5
13 Public health	4,3	4,9	5,1	2,4	1,8	2,8
14 Public services & infrastructure	0,9	0,7	0,5	0,4	0,4	0,4
15 Regions & national cohesion	0,7	0,7	0,9	0,7	2,4	0,8
16 Social security/welfare state	19,8	15,3	21,3	9,8	11,1	8
17 Other problems	0,7	0,8	0,6	0,2	1	0,5
18 No problem	0,2	0,1	0,1	0	0	0
19 COVID-19 pandemic	0	0	0	22,3	29,5	0,3

17.3 Fixed effect model

Fixed effect models (e.g., including time dummy variables) are sometimes applied to remove omitted variable bias. By estimating changes within a specific group (over time) all time-invariant differences between entities are controlled for.

The assumption behind the fixed effect model is that something influences the independent variables and one needs to control for it. The model removes characteristics that do not change over time, leading to unbiased estimates of the remaining independent variables on the dependent variable. Hence, if unobserved characteristics do not change over time, each change in the dependent variable must be due to influences not related to the fixed effects, which are controlled for.

i Time-invariant variables

Note that the influence of time-invariant independent variables on the dependent variable cannot be examined with a fixed effect model. In this case, Generalized Estimating Equation (GEE) models are more suitable for estimating a nonvarying (or average) coefficient in the presence of clustering.
GEEs are based on the quasi-likelihood theory and no assumption is made about the distribution of response observations (Liang and Zeger, 1986). Furthermore, with GEEs, there is no need to make hypotheses on the structure of residuals compared to other models. This makes it possible to manage the panel data with a dichotomous dependent variable.

17.4 Multilevel model

Random effect models (with random intercept and/or slope) assume that any variation between entities is random and not correlated with the independent variables used in the estimation model. This also means that time-invariant variables (like a person's gender) can be taken into account as independent variables. The entity's error term (unobserved heterogeneity) is hence not correlated with the independent variables.

A decision between a fixed and random effects model can be made with the Hausman test assessing whether the individual error terms are correlated with the independent variables. The null hypothesis states that there is no such correlation (thus, one should go with a random effect model).The alternative hypothesis is that a correlation exists (thus, one should opt for a fixed effect model). The test is implemented in function phtest().

17.5 plm package

A useful ressource to conduct longitudinal data analysis is the plm R package. For more information about the plm package, refer to the following article:

Croissant, Y., & Millo, G. (2008). Panel Data Econometrics in R: The plm Package. *Journal of Statistical Software*, 27(2), 1-43. Available here

17.6 Analytical steps

As with MLM, we can start by conducting a fixed effect model by including dummy variables (e.g. for years). Then, the function pFtest() tests for fixed effects with the null hypothesis that pooled linear model is better than fixed effects.

In contrast to the fixed effect model, the random effect model assumes that an individual (entity) specific effect is not correlated with the independent variables. A decision between a fixed and random effects model can be made with the Hausman test. The null hypothesis states that there is no such correlation (thus, one should prefer the random effect model). The alternative hypothesis is that a correlation exists (thus, one should go for the fixed effect model).

The Breusch-Pagan Lagrange multiplier Test further helps to decide between a random effects model and a simple linear regression. This test is implemented in function plmtest() with the null hypothesis that the variance across entities is zero (thus, this means that there is no panel effect).

17.7 Heteroskedasticity and serial correlation

Testing for the presence of heteroskedasticity can be implemented in the function bptest().

For long time-series, a test for serial correlation of the residuals should be performed because serial correlation can lead to an underestimation of standard errors (too small) and an overestimation of R2 (too large). The test is implemented in function pbgtest() with the null hypothesis that there is no serial correlation.

In order to solve the issue of serial correlation, clustered standard errors have to be used. Clustered standard errors estimate the variance of the coefficient when independent variables are correlated within the entity. Correcting the standard errors can be done with the function vcovHC().

17.8 How it works in R?

See the lecture slides on longitudinal data analysis:

Chapter 18

Latent Growth Modelling

18.1 Latent Growth Modelling (LGM)

LGM allows us to investigate longitudinal trends or group differences in measurement (also called growth trajectories). For example, suppose we want to model the political (or their political political engagement) polarization of panel participants over time. The objective of this chapter is to understand the concept of LGM and to be able to implement these models in lavaan.

18.2 Recap: lavaan syntax

Let's first recall the typical lavaan syntax:

- ~ predict, used for regression of observed outcome to observed predictors
- =~ indicator, used for latent variable to observed indicator in factor analysis measurement models
- ~~ covariance
- ~1 intercept or mean (e.g., q01 ~ 1 estimates the mean of variable q01)
- 1* fixes parameter or loading to one
- NA* frees parameter or loading (useful to override default marker method)
- a* labels the parameter 'a', used for model constraints
- c(a,b)* specifically for multigroup models, see note below

18.3 Understanding the concept of LGM

LGM are thus a special class of CFA models used to model trajectories over time. LGM models can be considered as an extension of the CFA model where the intercepts are freely estimated. Most commonly:

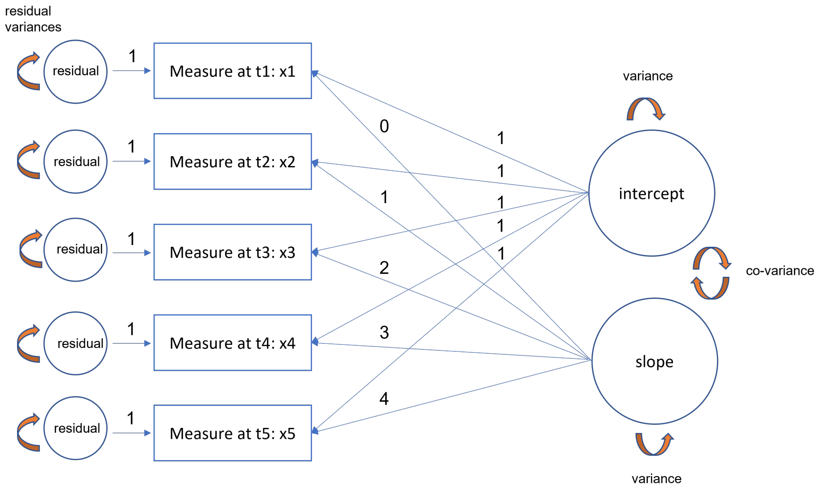
- the loadings for the intercept factor are all fixed to 1

- the loadings for the linear slope factor can be any ordered progression
- the first loading of the slope factor is fixed to 0
- the progression proceeds ordinally in increments of 1
- a one-unit increase in time is interpreted as an increase in the predicted outcome for a fixed time interval increase

Unlike traditional CFA models where the interpretation focuses on the loadings, the loadings are fixed in GLM which implies that the focus is on the latent intercept and linear slope factors.

For GLM, the dataset should be structured in wide format: each column represents the outcome (or dependent variable) at a specific time point. An assumption is that each observation or row must be independent of another. However, columns indicating outcomes are time dependent.

Here is a path diagram representing our particular LGM:



The equation for each time point defined for a person states as:

$$x_{it} = \tau_i + \xi_1 + (t)\xi_2 + \delta_{it}$$

where the observed intercepts are constrained to be zero ($\tau_i = 0$). Furthermore, the (symmetric) variance covariance matrix of the latent intercept and slope is defined as:

$$\Phi = \begin{bmatrix} \phi_{11} & \\ \phi_{21} & \phi_{22} \end{bmatrix}$$

where ϕ_{11} is the variance of the latent intercept, ϕ_{22} is the variance of the latent slope and $\phi 21$ is the covariance of the intercept and slope. The population parameters correspond to the latent intercept and linear slope means.

18.4 Ordinal versus measured time in an LGM

The fixed loadings for the slope factor can be defined in different way depending on how time is modeled. For instance, it can be assumed that time is measured ordinally (in increments of year or semester), regardless of the actual measured time (the loadings will be 0,1,2,3,…). In other situations, it may be more beneficial to use measured time, such as for a study design where assessments are implemented at baseline, 3-, 6-, 9-, and 12-month follow-up (the loadings will be fixed at 0, 3, 6, 9 and 12), thus corresponding to the actual time of assessment. The choice of using ordinal versus measured time impacts on the mean of the intercept and slope and, therefore, the interpretation of these terms. A recommendation is that, if time intervals are unequal, it may be better to use measured time.

18.5 Quiz

True

False

Statement

In a latent growth model, the observed intercepts are constrained to be zero but the latent intercepts are unconstrained to be free.

Suppose the mean of the slope is positive. A positive covariance between the intercept and slope means that for lower values of the starting X, the weaker the linear increase in X over time.

Suppose the last semester was four months long instead of three. If we continue to use ordinal time (i.e., 0,1,2,3,4), the mean of the slope factor would still be interpreted as the increase in X for every one semester increase in time.

Suppose the last semester was four months long instead of three. Using measured time would be more representative.

View Results

My results will appear here

18.6 Equivalence of the LGM to the hierarchical linear model (HLM)

LGM can be re-specified as an equivalent hierarchical linear model (HLM). The first modification is the format of the dataset: it should be in long format with repeated observations of time spanning multiple rows of data for every subject. For instance, HLM consists of repeated observations at Level 1 nested within individuals (index for time points: i) at Level 2 (index for individuals: j).

This suggests that the LGM model is equivalent to a HLM model with time nested within individuals with the following specifications:

- inclusion of a fixed effect of time at Level 1
- addition of a random intercept clustered by student and a random slope of time (indicated as time|student with the lmer syntax)

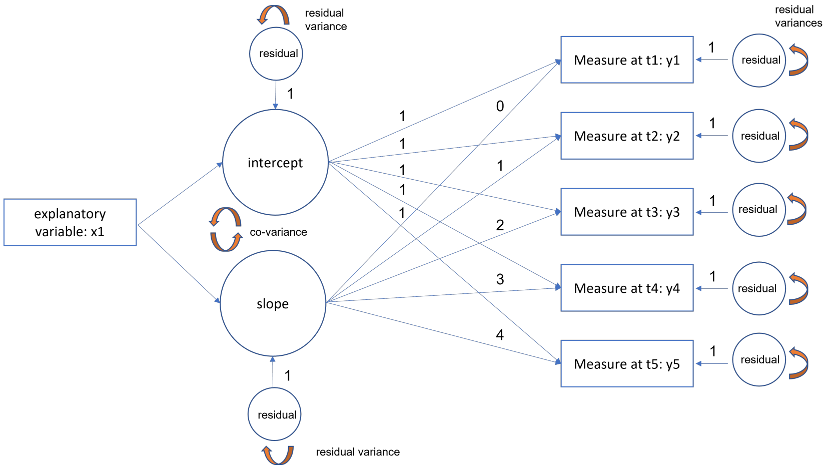
$$
\begin{aligned}
y_{ij} = \beta_{0j} + \beta_{1j} * TIME_{ij} + r_{ij} \\
\beta_{0j} = \gamma_{00} + u_{0j} \\
\beta_{1j} = \gamma_{10} + u_{1j}
\end{aligned}
$$

The important difference between the default LGM and HLM is that the residual variances of the LGM are unconstrained across timepoints but are constrained to be the same across time in a HLM by default. In order to constrain the residual variances in an LGM, we can use the a* notation in the lavaan syntax. In that case, the interpretation of the output is exactly the same for the LGM as it is for the HLM.

18.7 Adding a predictor to the LGM

Until now, the fundamental latent growth model was characterized an intercept and linear growth factor, thus enabling us to answer the question of whether the linear trajectory is increasing, decreasing or flat over time. Now, we would like to find out what are potential predictors of this linear trajectory. For example, if the trajectory of political involvement increases over time, how does political affiliation predict this trend? Are there ideological differences in either the starting political involvement (intercept factor) or the growth trajectory (slope factor)? We thus would have the following specifications:

- latent factors now have a predictor
- latent intercept and slope factors become endogenous (y-side variables)
- there is a residual term which was not in the exogenous model (replace the variance of factor with the residual factor term)
- additional of an exogenous predictor (x1)



18.8 How it works in R?

See the lecture slides on LGM:

Part V

TOOLKIT

Chapter 19

Regression toolkit

19.1 ShinyApp to visualise (simple/multiple) regressions

We can visualise the result of a (simple/multiple) regression modelling with a ShinyApp that relies on a subset of the European Social Survey data and data from the Gender Equality Index (the ShinyApp is also available here).

Note: the source code is inspired from E. Harrison' shinyfit application which can be found on GitHub.