



A Guide on Data Analysis

Mike Nguyen

2022-09-17

Contents

Preface	11
1 Introduction	13
2 Prerequisites	21
2.1 Matrix Theory	21
2.2 Probability Theory	26
2.3 General Math	51
2.4 Data Import/Export	56
2.5 Data Manipulation	62
I. BASIC	77
3 Descriptive Statistics	77
3.1 Numerical Measures	77
3.2 Graphical Measures	79
3.3 Normality Assessment	83
3.4 Bivariate Statistics	88
3.5 Two Continuous	89
3.6 Categorical and Continuous	90
3.7 Two Discrete	91

4 Basic Statistical Inference	109
4.1 One Sample Inference	111
4.2 Two Sample Inference	122
4.3 Categorical Data Analysis	132
4.4 Divergence Metrics and Test for Comparing Distributions	140
II. REGRESSION	147
5 Linear Regression	147
5.1 Ordinary Least Squares	147
5.2 Feasible Generalized Least Squares	193
5.3 Weighted Least Squares	201
5.4 Generalized Least Squares	202
5.5 Feasible Prais Winsten	202
5.6 Feasible group level Random Effects	203
5.7 Ridge Regression	204
5.8 Principal Component Regression	205
5.9 Robust Regression	205
5.10 Maximum Likelihood	207
6 Non-linear Regression	217
6.1 Inference	217
6.2 Non-linear Least Squares	222
7 Generalized Linear Models	249
7.1 Logistic Regression	249
7.2 Probit Regression	259
7.3 Binomial Regression	260
7.4 Poisson Regression	263
7.5 Negative Binomial Regression	267
7.6 Multinomial	268
7.7 Generalization	275

CONTENTS	5
8 Linear Mixed Models	293
8.1 Dependent Data	293
8.2 Estimation	301
8.3 Inference	307
8.4 Information Criteria	309
8.5 Split-Plot Designs	311
8.6 Repeated Measures in Mixed Models	316
8.7 Unbalanced or Unequally Spaced Data	318
8.8 Application	318
9 Nonlinear and Generalized Linear Mixed Models	335
9.1 Estimation	338
9.2 Application	342
9.3 Summary	363
III. RAMIFICATIONS	367
10 Model Specification	367
10.1 Nested Model	367
10.2 Non-Nested Model	368
10.3 Heteroskedasticity	370
11 Imputation (Missing Data)	373
11.1 Assumptions	374
11.2 Solutions to Missing data	376
11.3 Criteria for Choosing an Effective Approach	397
11.4 Another Perspective	397
11.5 Diagnosing the Mechanism	399
11.6 Application	400

12 Data	413
12.1 Cross-Sectional	413
12.2 Time Series	413
12.3 Repeated Cross Sections	420
12.4 Panel Data	421
13 Hypothesis Testing	445
13.1 Types of hypothesis testing	446
13.2 Wald test	448
13.3 The likelihood ratio test	455
13.4 Lagrange Multiplier (Score)	456
14 Prediction and Estimation	457
15 Moderation	459
15.1 emmeans package	460
15.2 probmod package	468
15.3 interactions package	468
15.4 interactionR package	487
15.5 sjPlot package	487
IV. CAUSAL INFERENCE	491
16 Causal Inference	491
16.1 Treatment effect types	496
A. EXPERIMENTAL DESIGN	507
17 Experimental Design	507
17.1 Semi-random Experiment	510
17.2 Rerandomization	512

CONTENTS	7
18 Sampling	517
18.1 Simple Sampling	517
18.2 Stratified Sampling	519
18.3 Unequal Probability Sampling	521
18.4 Balanced Sampling	521
19 Analysis of Variance (ANOVA)	525
19.1 Completely Randomized Design (CRD)	526
19.2 Nonparametric ANOVA	561
19.3 Sample Size Planning for ANOVA	563
19.4 Randomized Block Designs	565
19.5 Nested Designs	571
19.6 Single Factor Covariance Model	575
20 Multivariate Methods	579
20.1 MANOVA	602
20.2 Principal Components	621
20.3 Factor Analysis	631
20.4 Discriminant Analysis	648
B. QUASI-EXPERIMENTAL DESIGN	675
21 Quasi-experimental	675
22 Regression Discontinuity	677
22.1 Specification Checks	681
22.2 Steps for Sharp RD	689
22.3 Steps for Fuzzy RD	689
22.4 Steps for RDiT (Regression Discontinuity in Time)	689
22.5 Evaluation of an RD	693
22.6 Applications	695

23 Difference-in-differences	703
23.1 Simple Dif-n-dif	704
23.2 Multiple periods and variation in treatment timing	713
23.3 Staggered Dif-n-dif	713
23.4 Two-way Fixed-effects	716
24 Synthetic Control	719
24.1 Synthetic Difference-in-differences	745
25 Event Studies	747
25.1 Other Issues	749
25.2 Aggregation	753
25.3 Heterogeneity in the event effect	754
25.4 Expected Return Calculation	754
25.5 Application	756
26 Matching Methods	761
26.1 MatchIt	773
26.2 MatchingFrontier	781
26.3 Propensity Scores	793
26.4 Mahalanobis Distance	795
26.5 Coarsened Exact Matching	795
26.6 Genetic Matching	799
26.7 Matching for time series-cross-section data	809
27 Interrupted Time Series	811
C. OTHER CONCERNS	815
28 Endogeneity	815
28.1 Measurement Error	817
28.2 Simultaneity	825
28.3 Endogenous Treatment	826
28.4 Endogenous Sample Selection	851

CONTENTS	9
29 Mediation	865
29.1 Traditional	865
29.2 Model-based causal mediation analysis	869
30 Directed Acyclic Graph	877
V. MISCELLANEOUS	881
31 Report	881
31.1 One summary table	882
31.2 Model Comparison	884
31.3 Changes in an estimate	890
32 Exploratory Data Analysis	893
33 Sensitivity Analysis/ Robustness Check	895
33.1 Specification curve	895
33.2 Coefficient stability	911
A Appendix	913
A.1 Git	913
A.2 Short-cut	915
A.3 Function short-cut	915
A.4 Citation	917
A.5 Install all necessary packages/libaries on your local machine . . .	917
B Bookdown cheat sheet	921
B.1 Operation	921
B.2 Math Expresssion/ Syntax	922
B.3 Table	926

Preface

This guide is an attempt to streamline and demystify the data analysis process. By no means this is an ultimate guide, or I am a great source of knowledge, or I claim myself to be a statistician/ econometrician, but I am a strong proponent of learning by teaching, and doing. Hence, this is more like a learning experience for both you and me. This book is completely free. My target audiences are those who have little to no experience in statistics and data science to those that have some interests in these fields and want to dive deeper and have a more holistic method. Even though my substantive domain of interest is marketing, this book can also be used for other disciplines that use scientific methods or data analysis.



More books by the author can be found here:

- Advanced Data Analysis: the second book in the data analysis series, which covers machine learning models (with a focus on prediction)
- Marketing Research
- Communication Theory

Chapter 1

Introduction

Since the beginning of the century, we have been bombarded with amazing advancements and inventions, especially in the field of statistics, information technology, computer science, or a new emerging field - data science. However, I believe the downside of this introduction is that we use **big** and **trendy** words too often (i.e., big data, machine learning, deep learning).

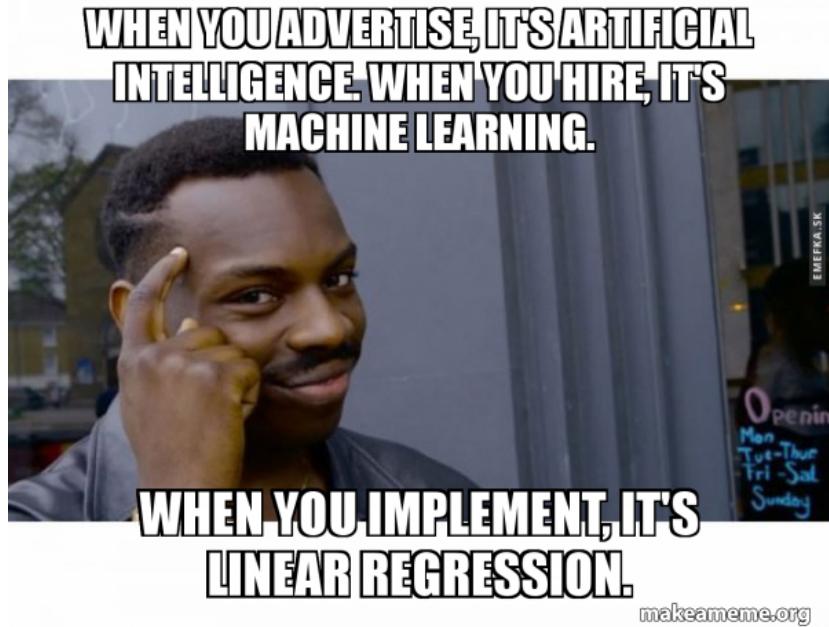
Each substantive field will have a metric subfield:

- Econometrics in economics
- Psychometrics in psychology
- Chemometrics in chemistry
- Sabermetrics in sports
- Biostatistics in public health and medicine

But to laymen, these are known as:

- Data Science
- Applied Statistics
- Computational Social Science

It's all fun and exciting when I learned these new tools. But I have to admit that I hardly retain any of these new ideas. However, writing down from the beginning till the end of a data analysis process is the solution that I came up with. Accordingly, let's dive right in.



Some general recommendations:

- The more you practice/habituize/condition, more line of codes that you write, more function that you memorize, I think the more you will like this journey.
- Readers can follow this book several ways:
 - If you are interested in particular methods/tools, you can jump to that section by clicking the section name.
 - If you want to follow a traditional path of data analysis, read the Linear Regression section.
 - If you want to create your experiment and test your hypothesis, read the Analysis of Variance (ANOVA) section.
- Alternatively, if you rather see the application of models, and disregard any theory or underlying mechanisms, you can skip to summary and application portion of each section.
- If you don't understand a part, search the title of that part of that part on Google, and read more into that subject. This is just a general guide.
- If you want to customize your code beyond the ones provided in this book, run in the console `help(code)` or `?code`. For example, I want more information on `hist` function, I'll type in the console `?hist` or `help(hist)`.

- Another way is that you can search on Google. Different people will use different packages to achieve the same result in R. Accordingly, if you want to create a histogram, search on Google `histogram in R`, then you should be able to find multiple ways to create histogram in R.

Information in this book are from various sources, but most of the content is based on several courses that I have taken formally. I'd like to give professors credit accordingly.

Course	Professor
Data Analysis I	Erin M. Schliep
Data Analysis II	Christopher Wikle
Applied Econometric	Alyssa Carlson

Tools of statistics

- Probability Theory
- Mathematical Analysis
- Computer Science
- Numerical Analysis
- Database Management

Code Replication

This book was built with R version 4.0.4 (2021-02-15) and the following packages:

package	version	source
abind	1.4-5	CRAN (R 4.0.3)
agridat	1.20	CRAN (R 4.0.5)
ape	5.6-1	CRAN (R 4.0.5)
assertthat	0.2.1	CRAN (R 4.0.3)
backports	1.4.1	CRAN (R 4.0.5)
bookdown	0.24	CRAN (R 4.0.5)
boot	1.3-28	CRAN (R 4.0.5)
broom	0.7.12	CRAN (R 4.0.5)
bslib	0.3.1	CRAN (R 4.0.5)
cachem	1.0.6	CRAN (R 4.0.5)
callr	3.7.0	CRAN (R 4.0.5)
car	3.0-12	CRAN (R 4.0.5)
carData	3.0-5	CRAN (R 4.0.5)
cellranger	1.1.0	CRAN (R 4.0.3)
cli	3.2.0	CRAN (R 4.0.4)

package	version	source
coda	0.19-4	CRAN (R 4.0.3)
colorspace	2.0-3	CRAN (R 4.0.4)
corpcor	1.6.10	CRAN (R 4.0.5)
crayon	1.5.0	CRAN (R 4.0.5)
cubature	2.0.4.2	CRAN (R 4.0.5)
curl	4.3.2	CRAN (R 4.0.5)
data.table	1.14.2	CRAN (R 4.0.5)
DBI	1.1.2	CRAN (R 4.0.5)
dbplyr	2.1.1	CRAN (R 4.0.5)
desc	1.4.0	CRAN (R 4.0.5)
devtools	2.4.3	CRAN (R 4.0.5)
digest	0.6.29	CRAN (R 4.0.5)
dplyr	1.0.8	CRAN (R 4.0.4)
ellipsis	0.3.2	CRAN (R 4.0.5)
evaluate	0.15	CRAN (R 4.0.4)
extrafont	0.17	CRAN (R 4.0.3)
extrafontdb	1.0	CRAN (R 4.0.3)
fansi	1.0.2	CRAN (R 4.0.5)
faraway	1.0.7	CRAN (R 4.0.3)
fastmap	1.1.0	CRAN (R 4.0.3)
forcats	0.5.1	CRAN (R 4.0.3)
foreign	0.8-82	CRAN (R 4.0.5)
fs	1.5.2	CRAN (R 4.0.5)
generics	0.1.2	CRAN (R 4.0.5)
ggplot2	3.3.5	CRAN (R 4.0.5)
glue	1.6.1	CRAN (R 4.0.5)
gttable	0.3.0	CRAN (R 4.0.3)
haven	2.4.3	CRAN (R 4.0.5)
Hmisc	4.6-0	CRAN (R 4.0.5)
hms	1.1.1	CRAN (R 4.0.5)
htmltools	0.5.2	CRAN (R 4.0.5)
htmlwidgets	1.5.4	CRAN (R 4.0.5)
httr	1.4.2	CRAN (R 4.0.3)
investr	1.4.0	CRAN (R 4.0.3)
jpeg	0.1-9	CRAN (R 4.0.5)
jquerylib	0.1.4	CRAN (R 4.0.5)
jsonlite	1.7.3	CRAN (R 4.0.5)
kableExtra	1.3.4	CRAN (R 4.0.4)
knitr	1.37	CRAN (R 4.0.5)
lattice	0.20-45	CRAN (R 4.0.5)
latticeExtra	0.6-29	CRAN (R 4.0.3)
lifecycle	1.0.1	CRAN (R 4.0.5)
lme4	1.1-28	CRAN (R 4.0.4)
lmerTest	3.1-3	CRAN (R 4.0.3)

package	version	source
lsr	0.5.2	CRAN (R 4.0.5)
ltm	1.2-0	CRAN (R 4.0.4)
lubridate	1.8.0	CRAN (R 4.0.5)
magrittr	2.0.2	CRAN (R 4.0.5)
MASS	7.3-55	CRAN (R 4.0.5)
matlib	0.9.5	CRAN (R 4.0.5)
Matrix	1.4-0	CRAN (R 4.0.5)
MCMCglmm	2.33	CRAN (R 4.0.5)
memoise	2.0.1	CRAN (R 4.0.5)
mgcv	1.8-38	CRAN (R 4.0.5)
minqa	1.2.4	CRAN (R 4.0.3)
modelr	0.1.8	CRAN (R 4.0.3)
munsell	0.5.0	CRAN (R 4.0.3)
nlme	3.1-155	CRAN (R 4.0.5)
nloptr	2.0.0	CRAN (R 4.0.5)
nlstools	2.0-0	CRAN (R 4.0.5)
nnet	7.3-17	CRAN (R 4.0.5)
numDeriv	2016.8-1.1	CRAN (R 4.0.3)
openxlsx	4.2.5	CRAN (R 4.0.5)
pbkrtest	0.5.1	CRAN (R 4.0.5)
pillar	1.7.0	CRAN (R 4.0.5)
pkgbuild	1.3.1	CRAN (R 4.0.5)
pkgconfig	2.0.3	CRAN (R 4.0.3)
pkgload	1.2.4	CRAN (R 4.0.5)
png	0.1-7	CRAN (R 4.0.3)
ppsr	0.0.2	CRAN (R 4.0.5)
prettyunits	1.1.1	CRAN (R 4.0.3)
processx	3.5.2	CRAN (R 4.0.5)
ps	1.6.0	CRAN (R 4.0.5)
pscl	1.5.5	CRAN (R 4.0.3)
purrr	0.3.4	CRAN (R 4.0.3)
R6	2.5.1	CRAN (R 4.0.5)
RColorBrewer	1.1-2	CRAN (R 4.0.3)
Rcpp	1.0.8	CRAN (R 4.0.5)
readr	2.1.2	CRAN (R 4.0.5)
readxl	1.3.1	CRAN (R 4.0.3)
remotes	2.4.2	CRAN (R 4.0.5)
reprex	2.0.1	CRAN (R 4.0.5)
rgl	0.108.3	CRAN (R 4.0.5)
rio	0.5.29	CRAN (R 4.0.5)
rlang	1.0.1	CRAN (R 4.0.5)
RLRsim	3.1-6	CRAN (R 4.0.4)
rmarkdown	2.11	CRAN (R 4.0.5)
rprojroot	2.0.2	CRAN (R 4.0.3)

package	version	source
rstudioapi	0.13	CRAN (R 4.0.3)
Rttf2pt1	1.3.10	CRAN (R 4.0.5)
vest	1.0.2	CRAN (R 4.0.5)
sass	0.4.0	CRAN (R 4.0.5)
scales	1.1.1	CRAN (R 4.0.3)
sessioninfo	1.2.2	CRAN (R 4.0.5)
stringi	1.7.6	CRAN (R 4.0.5)
stringr	1.4.0	CRAN (R 4.0.3)
svglite	2.1.0	CRAN (R 4.0.5)
systemfonts	1.0.4	CRAN (R 4.0.5)
tensorA	0.36.2	CRAN (R 4.0.3)
testthat	3.1.2	CRAN (R 4.0.5)
tibble	3.1.6	CRAN (R 4.0.5)
tidyverse	1.2.0	CRAN (R 4.0.5)
tidyselect	1.1.2	CRAN (R 4.0.4)
tzdb	1.3.1	CRAN (R 4.0.5)
usethis	0.2.0	CRAN (R 4.0.5)
utf8	2.1.5	CRAN (R 4.0.5)
vctrs	0.3.8	CRAN (R 4.0.5)
viridisLite	0.4.0	CRAN (R 4.0.5)
webshot	0.5.2	CRAN (R 4.0.5)
withr	2.4.3	CRAN (R 4.0.5)
xfun	0.29	CRAN (R 4.0.5)
xml2	1.3.3	CRAN (R 4.0.5)
xtable	1.8-4	CRAN (R 4.0.3)
yaml	2.3.4	CRAN (R 4.0.4)
zip	2.2.0	CRAN (R 4.0.5)

```
#> - Session info -----
#>   setting  value
#>   version R version 4.0.4 (2021-02-15)
#>   os        Windows 10 x64 (build 19043)
#>   system    x86_64, mingw32
#>   ui        RTerm
#>   language  (EN)
#>   collate   English_United States.1252
#>   ctype     English_United States.1252
#>   tz        America/Chicago
#>   date      2022-02-22
#>   pandoc   2.17.1.1 @ C:/Program Files/RStudio/bin/quarto/bin/ (via rmarkdown)
#>
#> - Packages -----
#>   package     * version date (UTC) lib source
```

```
#> assertthat    0.2.1   2019-03-21 [2] CRAN (R 4.0.3)
#> backports     1.4.1   2021-12-13 [1] CRAN (R 4.0.5)
#> bookdown      0.24    2021-09-02 [2] CRAN (R 4.0.5)
#> brio          1.1.3   2021-11-30 [1] CRAN (R 4.0.5)
#> broom          0.7.12  2022-01-28 [1] CRAN (R 4.0.5)
#> cachem         1.0.6   2021-08-19 [2] CRAN (R 4.0.5)
#> callr          3.7.0   2021-04-20 [2] CRAN (R 4.0.5)
#> cellranger    1.1.0   2016-07-27 [2] CRAN (R 4.0.3)
#> cli            3.2.0   2022-02-14 [1] CRAN (R 4.0.4)
#> codetools      0.2-18  2020-11-04 [2] CRAN (R 4.0.4)
#> colorspace     2.0-3   2022-02-21 [1] CRAN (R 4.0.4)
#> crayon         1.5.0   2022-02-14 [2] CRAN (R 4.0.5)
#> DBI            1.1.2   2021-12-20 [1] CRAN (R 4.0.5)
#> dbplyr         2.1.1   2021-04-06 [2] CRAN (R 4.0.5)
#> desc           1.4.0   2021-09-28 [1] CRAN (R 4.0.5)
#> devtools       2.4.3   2021-11-30 [1] CRAN (R 4.0.5)
#> digest          0.6.29  2021-12-01 [1] CRAN (R 4.0.5)
#> dplyr          * 1.0.8  2022-02-08 [1] CRAN (R 4.0.4)
#> ellipsis        0.3.2   2021-04-29 [2] CRAN (R 4.0.5)
#> evaluate        0.15    2022-02-18 [2] CRAN (R 4.0.4)
#> fansi           1.0.2   2022-01-14 [1] CRAN (R 4.0.5)
#> fastmap         1.1.0   2021-01-25 [2] CRAN (R 4.0.3)
#> forcats         * 0.5.1  2021-01-27 [2] CRAN (R 4.0.3)
#> fs              1.5.2   2021-12-08 [1] CRAN (R 4.0.5)
#> generics         0.1.2  2022-01-31 [1] CRAN (R 4.0.5)
#> ggplot2         * 3.3.5  2021-06-25 [2] CRAN (R 4.0.5)
#> glue             1.6.1   2022-01-22 [1] CRAN (R 4.0.5)
#> gtable           0.3.0   2019-03-25 [2] CRAN (R 4.0.3)
#> haven            2.4.3   2021-08-04 [2] CRAN (R 4.0.5)
#> highr            0.9    2021-04-16 [2] CRAN (R 4.0.5)
#> hms              1.1.1   2021-09-26 [1] CRAN (R 4.0.5)
#> htmltools        0.5.2   2021-08-25 [2] CRAN (R 4.0.5)
#> httr              1.4.2   2020-07-20 [2] CRAN (R 4.0.3)
#> jpeg             * 0.1-9  2021-07-24 [2] CRAN (R 4.0.5)
#> jsonlite         1.7.3   2022-01-17 [1] CRAN (R 4.0.5)
#> knitr            1.37    2021-12-16 [1] CRAN (R 4.0.5)
#> lifecycle        1.0.1   2021-09-24 [2] CRAN (R 4.0.5)
#> lubridate        1.8.0   2021-10-07 [1] CRAN (R 4.0.5)
#> magrittr         2.0.2   2022-01-26 [1] CRAN (R 4.0.5)
#> memoise          2.0.1   2021-11-26 [1] CRAN (R 4.0.5)
#> modelr           0.1.8   2020-05-19 [2] CRAN (R 4.0.3)
#> munsell          0.5.0   2018-06-12 [2] CRAN (R 4.0.3)
#> pillar            1.7.0   2022-02-01 [1] CRAN (R 4.0.5)
#> pkgbuild         1.3.1   2021-12-20 [1] CRAN (R 4.0.5)
#> pkgconfig        2.0.3   2019-09-22 [2] CRAN (R 4.0.3)
#> pkgload           1.2.4   2021-11-30 [1] CRAN (R 4.0.5)
```

```
#> prettyunits    1.1.1    2020-01-24 [2] CRAN (R 4.0.3)
#> processx      3.5.2    2021-04-30 [2] CRAN (R 4.0.5)
#> ps             1.6.0    2021-02-28 [2] CRAN (R 4.0.5)
#> purrr          * 0.3.4   2020-04-17 [2] CRAN (R 4.0.3)
#> R6              2.5.1    2021-08-19 [2] CRAN (R 4.0.5)
#> Rcpp            1.0.8    2022-01-13 [2] CRAN (R 4.0.5)
#> readr           * 2.1.2   2022-01-30 [1] CRAN (R 4.0.5)
#> readxl          1.3.1    2019-03-13 [2] CRAN (R 4.0.3)
#> remotes         2.4.2    2021-11-30 [1] CRAN (R 4.0.5)
#> reprex          2.0.1    2021-08-05 [2] CRAN (R 4.0.5)
#> rlang            1.0.1    2022-02-03 [1] CRAN (R 4.0.5)
#> rmarkdown        2.11     2021-09-14 [2] CRAN (R 4.0.5)
#> rprojroot       2.0.2    2020-11-15 [2] CRAN (R 4.0.3)
#> rstudioapi      0.13     2020-11-12 [2] CRAN (R 4.0.3)
#> rvest            1.0.2    2021-10-16 [1] CRAN (R 4.0.5)
#> scales           * 1.1.1   2020-05-11 [2] CRAN (R 4.0.3)
#> sessioninfo     1.2.2    2021-12-06 [1] CRAN (R 4.0.5)
#> stringi          1.7.6    2021-11-29 [1] CRAN (R 4.0.5)
#> stringr          * 1.4.0   2019-02-10 [2] CRAN (R 4.0.3)
#> testthat         3.1.2    2022-01-20 [1] CRAN (R 4.0.5)
#> tibble           * 3.1.6   2021-11-07 [1] CRAN (R 4.0.5)
#> tidyverse         * 1.3.1   2021-04-15 [2] CRAN (R 4.0.5)
#> tzdb              0.2.0    2021-10-27 [1] CRAN (R 4.0.5)
#> usethis          2.1.5    2021-12-09 [1] CRAN (R 4.0.5)
#> utf8              1.2.2    2021-07-24 [2] CRAN (R 4.0.5)
#> vctrs              0.3.8    2021-04-29 [2] CRAN (R 4.0.5)
#> withr              2.4.3    2021-11-30 [1] CRAN (R 4.0.5)
#> xfun              0.29     2021-12-14 [1] CRAN (R 4.0.5)
#> xml2              1.3.3    2021-11-30 [1] CRAN (R 4.0.5)
#> yaml              2.3.4    2022-02-17 [1] CRAN (R 4.0.4)
#>
#> [1] C:/Users/tn9k4/OneDrive - University of Missouri/Documents/R/win-library/4.0
#> [2] C:/Program Files/R/R-4.0.4/library
#>
#> -----
```

Chapter 2

Prerequisites

This chapter is just a quick review of Matrix Theory and Probability Theory

If you feel you do not need to brush up on these theories, you can jump right into Descriptive Statistics

2.1 Matrix Theory

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad (2.1)$$

$$A' = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{bmatrix} \quad (2.2)$$

$$(\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}'\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{ACAB} \neq \mathbf{BA}(\mathbf{A}')' = \mathbf{A}(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}\mathbf{B}$$

If A has an inverse, it is called **invertible**. If A is not invertible it is called **singular**.

$$\begin{aligned} \mathbf{A} &= \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} \\ &= \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & \sum_{i=1}^3 a_{1i}b_{i2} & \sum_{i=1}^3 a_{1i}b_{i3} \\ \sum_{i=1}^3 a_{2i}b_{i1} & \sum_{i=1}^3 a_{2i}b_{i2} & \sum_{i=1}^3 a_{2i}b_{i3} \end{pmatrix} \end{aligned} \quad (2.3)$$

Let \mathbf{a} be a 3×1 vector, then the quadratic form is

$$\mathbf{a}'\mathbf{B}\mathbf{a} = \sum_{i=1}^3 \sum_{j=1}^3 a_i b_{ij} a_j$$

Length of a vector

Let \mathbf{a} be a vector, $\|\mathbf{a}\|$ (the 2-norm of the vector) is the length of vector \mathbf{a} , is the square root of the inner product of the vector with itself:

$$\|\mathbf{a}\| = \sqrt{\mathbf{a}'\mathbf{a}}$$

2.1.1 Rank

- Dimension of space spanned by its columns (or its rows).
- Number of linearly independent columns/rows

For a $n \times k$ matrix \mathbf{A} and $k \times k$ matrix \mathbf{B}

- $\text{rank}(A) \leq \min(n, k)$
- $\text{rank}(A) = \text{rank}(A') = \text{rank}(A'A) = \text{rank}(AA')$
- $\text{rank}(AB) = \min(\text{rank}(A), \text{rank}(B))$
- \mathbf{B} is invertible if and only if $\text{rank}(\mathbf{B}) = k$ (non-singular)

2.1.2 Inverse

In scalar, $a = 0$ then $1/a$ does not exist. In matrix, a matrix is invertible when it's a non-zero matrix.

A non-singular square matrix A is invertible if there exists a non-singular square matrix B such that,

$$AB = I$$

Then $A^{-1} = B$. For a 2×2 matrix,

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

$$A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

For the partition matrix,

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \end{bmatrix} \quad (2.4)$$

Properties for a non-singular square matrix

- $\mathbf{A}^{-1} = A$
- for a non-zero scalar b , $(b\mathbf{A})^{-1} = b^{-1}\mathbf{A}^{-1}$
- for a matrix B , $(BA)^{-1} = B^{-1}A^{-1}$ only if B is non-singular
- $(\mathbf{A}^{-1})' = (\mathbf{A}')^{-1}$
- Never notate $1/\mathbf{A}$

2.1.3 Definiteness

A symmetric square $k \times k$ matrix, \mathbf{A} , is Positive Semi-Definite if for any non-zero $k \times 1$ vector \mathbf{x} ,

$$\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$$

A symmetric square $k \times k$ matrix, \mathbf{A} , is Negative Semi-Definite if for any non-zero $k \times 1$ vector \mathbf{x}

$$\mathbf{x}'\mathbf{A}\mathbf{x} \leq 0$$

\mathbf{A} is indefinite if it is neither positive semi-definite or negative semi-definite.

The identity matrix is positive definite

Example Let $\mathbf{x} = (x_1 x_2)'$, then for a 2×2 identity matrix,

$$\begin{aligned} \mathbf{x}'\mathbf{I}\mathbf{x} &= (x_1 x_2) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= (x_1 x_2) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= x_1^2 + x_2^2 > 0 \end{aligned} \quad (2.5)$$

Definiteness gives us the ability to compare matrices $\mathbf{A} - \mathbf{B}$ is PSD This property also helps us show efficiency (which variance covariance matrix of one estimator is smaller than another)

Properties

- any variance matrix is PSD
- a matrix \mathbf{A} is PSD if and only if there exists a matrix \mathbf{B} such that $\mathbf{A} = \mathbf{B}'\mathbf{B}$

- if \mathbf{A} is PSD, then $\mathbf{B}'\mathbf{AB}$ is PSD
- if \mathbf{A} and \mathbf{C} are non-singular, then $\mathbf{A}-\mathbf{C}$ is PSD if and only if $\mathbf{C}^{-1} - \mathbf{A}^{-1}$
- if \mathbf{A} is PD (ND) then \mathbf{A}^{-1} is PD (ND)

Note

- Indefinite A is neither PSD nor NSD. There is no comparable concept in scalar.
- If a square matrix is PSD and invertible then it is PD

Example:

1. Invertible / Indefinite

$$\begin{bmatrix} -1 & 0 \\ 0 & 10 \end{bmatrix}$$

2. Non-invertible/ Indefinite

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

3. Invertible / PSD

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

4. Non-Invertible / PSD

$$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

2.1.4 Matrix Calculus

$y = f(x_1, x_2, \dots, x_k) = f(\mathbf{x})$ where \mathbf{x} is a $1 \times k$ row vector. The Gradient (first order derivative with respect to a vector) is,

$$\frac{\partial f(x)}{\partial x} = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_k} \end{pmatrix}$$

The **Hessian** (second order derivative with respect to a vector) is,

$$\frac{\partial^2 f(x)}{\partial x \partial x'} = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_k} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_k \partial x_1} & \frac{\partial^2 f(x)}{\partial x_k \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_k \partial x_k} \end{pmatrix}$$

Define the derivative of $f(\mathbf{X})$ with respect to $\mathbf{X}_{(n \times p)}$ as the matrix

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \left(\frac{\partial f(\mathbf{X})}{\partial x_{ij}} \right)$$

Define \mathbf{a} to be a vector and \mathbf{A} to be a matrix which does not depend upon \mathbf{y} . Then

$$\frac{\partial \mathbf{a}' \mathbf{y}}{\partial \mathbf{y}} = \mathbf{a}$$

$$\frac{\partial \mathbf{y}' \mathbf{y}}{\partial \mathbf{y}} = 2\mathbf{y}$$

$$\frac{\partial \mathbf{y}' \mathbf{A} \mathbf{y}}{\partial \mathbf{y}} = (\mathbf{A} + \mathbf{A}')\mathbf{y}$$

If \mathbf{X} is a symmetric matrix then

$$\frac{\partial |\mathbf{X}|}{\partial x_{ij}} = \begin{cases} X_{ii}, i = j \\ X_{ij}, i \neq j \end{cases}$$

where X_{ij} is the (i,j)th cofactor of \mathbf{X}

If \mathbf{X} is symmetric and \mathbf{A} is a matrix which does not depend upon \mathbf{X} then

$$\frac{\partial \text{tr} \mathbf{X} \mathbf{A}}{\partial \mathbf{X}} = \mathbf{A} + \mathbf{A}' - \text{diag}(\mathbf{A})$$

If \mathbf{X} is symmetric and we let \mathbf{J}_{ij} be a matrix which has a 1 in the (i,j)th position and 0s elsewhere, then

$$\frac{\partial \mathbf{X}^6 - 1}{\partial x_{ij}} = \begin{cases} -\mathbf{X}^{-1} \mathbf{J}_{ii} \mathbf{X}^{-1}, & i = j \\ -\mathbf{X}^{-1} (\mathbf{J}_{ij} + \mathbf{J}_{ji}) \mathbf{X}^{-1}, & i \neq j \end{cases}$$

2.1.5 Optimization

	Scalar Optimization	Vector Optimization
First Order Condition	$\frac{\partial f(x_0)}{\partial x} = 0$	$\frac{\partial f(x_0)}{\partial x} = \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}$
Second Order Condition	$\frac{\partial^2 f(x_0)}{\partial x^2} > 0$	$\frac{\partial^2 f(x_0)}{\partial x x'} > 0$
Convex → Min		
Concave → Max	$\frac{\partial^2 f(x_0)}{\partial x^2} < 0$	$\frac{\partial^2 f(x_0)}{\partial x x'} < 0$

2.2 Probability Theory

2.2.1 Axiom and Theorems of Probability

1. Let S denote a sample space of an experiment $P[S]=1$
2. $P[A] \geq 0$ for every event A
3. Let A_1, A_2, A_3, \dots be a finite or an infinite collection of mutually exclusive events. Then $P[A_1 \cup A_2 \cup A_3 \dots] = P[A_1] + P[A_2] + P[A_3] + \dots$
4. $P[\emptyset] = 0$
5. $P[A'] = 1 - P[A]$
6. $P[A_1 \cup A_2] = P[A_1] + P[A_2] - P[A_1 \cap A_2]$

Conditional Probability

$$P[A|B] = \frac{A \cap B}{P[B]}$$

Independent Events Two events A and B are independent if and only if:

1. $P[A \cap B] = P[A]P[B]$

- 2. $P[A|B] = P[A]$
- 3. $P[B|A] = P[B]$

A finite collection of events A_1, A_2, \dots, A_n is independent if and only if any subcollection is independent.

Multiplication Rule $P[A \cap B] = P[A|B]P[B] = P[B|A]P[A]$

Bayes' Theorem Let A_1, A_2, \dots, A_n be a collection of mutually exclusive events whose union is S.

Let b be an event such that $P[B] \neq 0$

Then for any of the events A_j , $j = 1, 2, \dots, n$

$$P[A|B] = \frac{P[B|A_j]P[A_j]}{\sum_{i=1}^n P[B|A_i]P[A_i]}$$

Jensen's Inequality

- If $g(x)$ is convex $E(g(X)) \geq g(E(X))$
- If $g(x)$ is concave $E(g(X)) \leq g(E(X))$

2.2.1.1 Law of Iterated Expectations

$$E(Y) = E(E(Y|X))$$

2.2.1.2 Correlation and Independence

Independence

- $f(x, y) = f_X(x)f_Y(y)$
- $f_{Y|X}(y|x) = f_Y(y)$ and $f_{X|Y}(x|y) = f_X(x)$
- $E(g_1(X)g_2(Y)) = E(g_1(X))E(g_2(Y))$

Mean Independence (implied by independence)

- Y is mean independent of X if and only if $E(Y|X) = E(Y)$
- $E(Xg(Y)) = E(X)E(g(Y))$

Uncorrelated (implied by independence and mean independence)

- $Cov(X, Y) = 0$
- $Var(X + Y) = Var(X) + Var(Y)$
- $E(XY) = E(X)E(Y)$

Strongest ↓ Independence ↓ Mean Independence ↓ Uncorrelated ↓ Weakest

2.2.2 Central Limit Theorem

Let X_1, X_2, \dots, X_n be a random sample of size n from a distribution (not necessarily normal) X with mean μ and variance σ^2 . then for large n ($n \geq 25$),

1. \bar{X} is approximately normal with mean $\mu_{\bar{X}} = \mu$ and variance $\sigma_{\bar{X}}^2 = \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$
2. \hat{p} is approximately normal with $\mu_{\hat{p}} = p, \sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}$
3. $\hat{p}_1 - \hat{p}_2$ is approximately normal with $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2, \sigma_{\hat{p}_1 - \hat{p}_2}^2 = \frac{p_1(1-p)}{n_1} + \frac{p_2(1-p)}{n_2}$
4. $\bar{X}_1 - \bar{X}_2$ is approximately normal with $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2, \sigma_{\bar{X}_1 - \bar{X}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
5. The following random variables are approximately standard normal:

- $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$
- $\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$
- $\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$
- $\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

If $\{x_i\}_{i=1}^n$ is an iid random sample from a probability distribution with finite mean μ and finite variance σ^2 then the sample mean $\bar{x} = n^{-1} \sum_{i=1}^n x_i$ scaled by \sqrt{n} has the following limiting distribution

$$\sqrt{n}(\bar{x} - \mu) \xrightarrow{d} N(0, \sigma^2)$$

or if we were to standardize the sample mean,

$$\frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} \xrightarrow{d} N(0, 1)$$

- holds for most random sample from any distribution (continuous, discrete, unknown).
- extends to multivariate case: random sample of a random vector converges to a multivariate normal.

- Variance from the limiting distribution is the asymptotic variance (Avar)

$$Avar(\sqrt{n}(\bar{x} - \mu)) = \sigma^2 \lim_{n \rightarrow \infty} Var(\sqrt{n}(\bar{x} - \mu)) = \sigma^2 Avar(.) \neq \lim_{n \rightarrow \infty} Var(.)$$

2.2.3 Random variable

	Discrete Variable	Continuous Variable
Definition	A random variable is discrete if it can assume at most a finite or countably infinite number of possible values	A random variable is continuous if it can assume any value in some interval or intervals of real numbers and the probability that it assumes any specific value is 0
Density Function	A function f is called a density for X if: (1) $f(x) \geq 0$ (2) $\sum_{\text{all } x} f(x) = 1$ (3) $f(x) = P(X = x)$ for x real	A function f is called a density for X if: (1) $f(x) \geq 0$ for x real (2) $\int_{-\infty}^{\infty} f(x) dx = 1$ (3) $P[a \leq X \leq b] = \int_a^b f(x) dx$ for a and b real
Cumulative Distribution Function for x real	$F(x) = P[X \leq x]$	$F(x) = P[X \leq x] = \int_{-\infty}^x f(t) dt$
$E[H(X)]$ $\mu = E[X]$	$\sum_{\text{all } x} H(x)f(x)$ $\sum_{\text{all } x} xf(x)$ $\sum_{\text{all } x \in X} (x^k f(x))$	$\int_{-\infty}^{\infty} H(x)f(x)$ $\int_{-\infty}^{\infty} xf(x)$ $\int_{-\infty}^{\infty} (x^k f(x))$
Ordinary Moments the k th ordinary moment for variable X is defined as: $E[X^k]$		
Moment generating function (mgf) $m_X(t) = E[e^{tX}]$	$\sum_{\text{all } x \in X} (e^{tx} f(x))$	$\int_{-\infty}^{\infty} (e^{tx} f(x) dx)$

Expected Value Properties:

- $E[c] = c$ for any constant c

- $E[cX] = cE[X]$ for any constant c
- $E[X+Y] = E[X] + E[Y]$
- $E[XY] = E[X]E[Y]$ (if X and Y are independent)

Expected Variance Properties:

- $Var(c) = 0$ for any constant c
- $Var(cX) = c^2Var(X)$ for any constant c
- $Var(X) \geq 0$
- $Var(X) = E(X^2) - (E(X))^2$
- $Var(X + c) = Var(X)$
- $Var(X + Y) = Var(X) + Var(Y)$ (if X and Y are independent)

Standard deviation $\sigma = \sqrt(\sigma^2) = \sqrt(Var X)$

Suppose y_1, \dots, y_p are possibly correlated random variables with means μ_1, \dots, μ_p . then

$$\mathbf{y} = (y_1, \dots, y_p)' E(\mathbf{y}) = (\mu_1, \dots, \mu_p)' =$$

Let $\sigma_{ij} = cov(y_i, y_j)$ for $i, j = 1, \dots, p$.

Define

$$= (\sigma_{ij}) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{pmatrix}$$

Hence, Σ is the variance-covariance or dispersion matrix. And Σ is symmetric with $(p+1)p/2$ unique parameters.

Alternatively, let $u_{p \times 1}$ and $v_{v \times 1}$ be random vectors with means \mathbf{u} and \mathbf{v} . then

$$\mathbf{u}\mathbf{v}' = cov(\mathbf{u}, \mathbf{v}) = E[(\mathbf{u} - \mathbf{u})(\mathbf{v} - \mathbf{v})']$$

$$\Sigma_{uv} \neq \Sigma_{vu} \text{ (but } \Sigma_{uv} = \Sigma'_{vu})$$

Properties of Covariance Matrices

1. Symmetric: $\Sigma' = \Sigma$
2. Eigendecomposition (spectral decomposition, symmetric decomposition): $\Sigma = V\Lambda V'$, where V is a matrix of eigenvectors such that $V'V = I$ (orthonormal), and Λ is a diagonal matrix with eigenvalues $(\lambda_1, \dots, \lambda_p)$ on the diagonal.

3. Non-negative definite, $\mathbf{a}^T \mathbf{a} \geq 0$ for any $\mathbf{a} \in R^p$. Equivalently, the eigenvalues of Σ , $\lambda_1 \geq \dots \geq \lambda_p \geq 0$
4. $\|\Sigma\| = \lambda_1, \dots, \lambda_p \geq 0$ (generalized variance)
5. $\text{trace}(\Sigma) = \text{tr}(\Sigma) = \lambda_1 + \dots + \lambda_p = \sigma_{11} + \dots + \sigma_{pp}$ = sum of variances (total variance)

Note: Σ is usually required to be positive definite. This implies that all eigenvalues are positive, and Σ has an inverse Σ^{-1} , such that $\Sigma^{-1} = \mathbf{I}_{p \times p} = \Sigma^{-1}$

Correlation Matrices

Define the correlation ρ_{ij} and the correlation matrix by

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

$$\mathbf{R} = \begin{pmatrix} \rho_{11} & \rho_{12} & \dots & \rho_{1p} \\ \rho_{21} & \rho_{22} & \dots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \dots & \rho_{pp} \end{pmatrix}$$

where $\rho_{ii} = 1$ for all i.

Let \mathbf{x} and \mathbf{y} be random vectors with means μ_x and μ_y and variance-covariance matrices Σ_x and Σ_y . Let \mathbf{A} and \mathbf{B} be matrices of constants and \mathbf{c} and \mathbf{d} be vectors of constants. Then,

- $E(\mathbf{A}\mathbf{y} + \mathbf{c}) = \mathbf{A}\mathbf{y} + \mathbf{c}$
- $\text{var}(\mathbf{A}\mathbf{y} + \mathbf{c}) = \mathbf{A}\text{var}(\mathbf{y})\mathbf{A}' = \mathbf{A}\mathbf{y}\mathbf{A}'$
- $\text{cov}(\mathbf{A}\mathbf{y} + \mathbf{c}, \mathbf{B}\mathbf{y} + \mathbf{d}) = \mathbf{A}\mathbf{y}\mathbf{B}'$

2.2.4 Moment generating function

Moment generating function properties:

$$(a) \frac{d^k(m_X(t))}{dt^k}|_{t=0} = E[X^k]$$

$$(b) \mu = E[X] = m'_X(0)$$

$$(c) E[X^2] = m''_X(0)$$

mgf Theorems

Let X_1, X_2, \dots, X_n, Y be random variables with moment-generating functions $m_{X_1}(t), m_{X_2}(t), \dots, m_{X_n}(t), m_Y(t)$

1. If $m_{X_1}(t) = m_{X_2}(t)$ for all t in some open interval about 0, then X_1 and X_2 have the same distribution
2. If $Y = \alpha + \beta X_1$, then $m_Y(t) = e^{\alpha t}m_{X_1}(\beta t)$
3. If X_1, X_2, \dots, X_n are independent and $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_n X_n$ (where $\alpha_0, \dots, \alpha_n$ are real numbers), then $m_Y(t) = e^{\alpha_0 t}m_{X_1}(\alpha_1 t)m_{X_2}(\alpha_2 t)\dots m_{X_n}(\alpha_n t)$
4. Suppose X_1, X_2, \dots, X_n are independent normal random variables with means $\mu_1, \mu_2, \dots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$. If $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_n X_n$ (where $\alpha_0, \dots, \alpha_n$ are real numbers), then Y is normally distributed with mean $\mu_Y = \alpha_0 + \alpha_1 \mu_1 + \alpha_2 \mu_2 + \dots + \alpha_n \mu_n$ and variance $\sigma_Y^2 = \alpha_1^2 \sigma_1^2 + \alpha_2^2 \sigma_2^2 + \dots + \alpha_n^2 \sigma_n^2$

2.2.5 Moment

Moment	Uncentered	Centered
1st	$E(X) = \mu = Mean(X)$	
2nd	$E(X^2)$	$E((X - \mu)^2) = Var(X) = \sigma^2$
3rd	$E(X^3)$	$E((X - \mu)^3)$
4th	$E(X^4)$	$E((X - \mu)^4)$

$$\text{Skewness}(X) = E((X - \mu)^3)/\sigma^3$$

$$\text{Kurtosis}(X) = E((X - \mu)^4)/\sigma^4$$

Conditional Moments

$$E(Y|X = x) = \begin{cases} \sum_y y f_Y(y|x) & \text{for discrete RV} \\ \int_y y f_Y(y|x) dy & \text{for continuous RV} \end{cases}$$

$$Var(Y|X = x) = \begin{cases} \sum_y (y - E(Y|x))^2 f_Y(y|x) & \text{for discrete RV} \\ \int_y (y - E(Y|x))^2 f_Y(y|x) dy & \text{for continuous RV} \end{cases}$$

2.2.5.1 Multivariate Moments

$$E = \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} E(X) \\ E(Y) \end{pmatrix} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \quad (2.6)$$

$$\begin{aligned} Var \begin{pmatrix} X \\ Y \end{pmatrix} &= \begin{pmatrix} Var(X) & Cov(X, Y) \\ Cov(X, Y) & Var(Y) \end{pmatrix} \\ &= \begin{pmatrix} E((X - \mu_X)^2) & E((X - \mu_X)(Y - \mu_Y)) \\ E((X - \mu_X)(Y - \mu_Y)) & E((Y - \mu_Y)^2) \end{pmatrix} \end{aligned} \quad (2.7)$$

Properties

- $E(aX + bY + c) = aE(X) + bE(Y) + c$
- $Var(aX + bY + c) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$
- $Cov(aX + bY, cX + bY) = acVar(X) + bdVar(Y) + (ad + bc)Cov(X, Y)$
- Correlation: $\rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$

2.2.6 Distributions

Conditional Distributions

$$f_{X|Y}(X|Y = y) = \frac{f(X, Y)}{f_Y(y)}$$

$f_{X|Y}(X|Y = y) = f_X(X)$ if X and Y are independent

2.2.6.1 Discrete

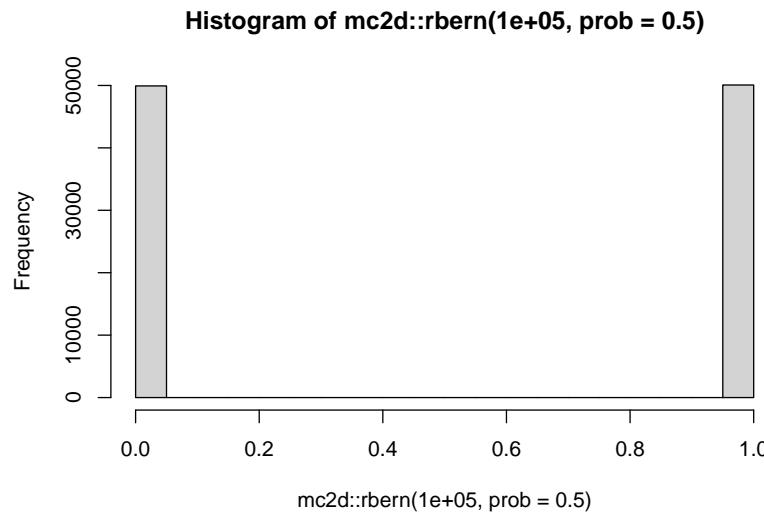
CDF: Cumulative Density Function

MGF: Moment Generating Function

2.2.6.1.1 Bernoulli $Bernoulli(p)$

PDF

```
hist(mc2d::rbern(100000, prob=.5))
```



2.2.6.1.2 Binomial $B(n, p)$

- the experiment consists of a fixed number (n) of Bernoulli trials, each of which results in a success (s) or failure (f)
- The trials are identical and independent, and probability of success (p) and probability of failure ($q = 1 - p$) remains the same for all trials.
- The random variable X denotes the number of successes obtained in the n trials.

Density

$$f(x) = \binom{n}{x} p^x q^{n-x}$$

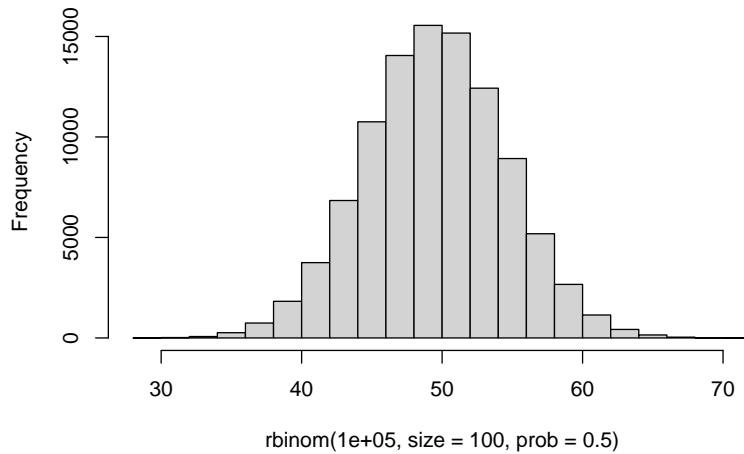
CDF

You have to use table

PDF

```
# Histogram of 100000 random values from a sample of 100 with probability of 0.5
hist(rbinom(100000, size = 100, prob = 0.5))
```

Histogram of $rbinom(1e+05, size = 100, prob = 0.5)$



MGF

$$m_X(t) = (q + pe^t)^n$$

Mean

$$\mu = E(x) = np$$

Variance

$$\sigma^2 = Var(X) = npq$$

2.2.6.1.3 Poisson $Pois(\lambda)$

- Arises with Poisson process, which involves observing discrete events in a continuous “interval” of time, length, or space.
- The random variable X is the number of occurrences of the event within an interval of s units
- The parameter λ is the average number of occurrences of the event in question per measurement unit. For the distribution, we use the parameter $k = \lambda s$

Density

$$f(x) = \frac{e^{-k} k^x}{x!}$$

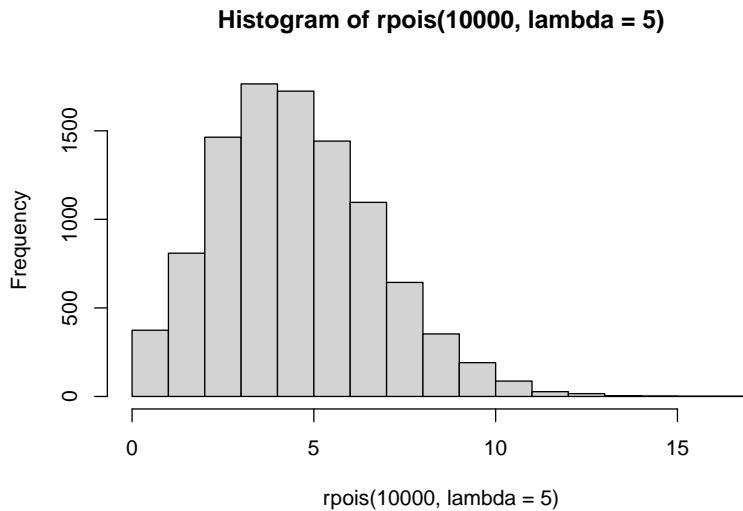
, $k > 0$, $x = 0, 1, \dots$

CDF

Use table

PDF

```
# Poisson dist with mean of 5 or Poisson(5)
hist(rpois(10000, lambda = 5))
```

**MGF**

$$m_X(t) = e^{k(e^t - 1)}$$

Mean

$$\mu = E(X) = k$$

Variance

$$\sigma^2 = Var(X) = k$$

2.2.6.1.4 Geometric

- The experiment consists of a series of trials. The outcome of each trial can be classed as being either a “success” (s) or “failure” (f). (This is called a Bernoulli trial).
- The trials are identical and independent in the sense that the outcome of one trial has no effect on the outcome of any other. The probability of success (p) and probability of failure (q=1-p) remains the same from trial to trial.
- lack of memory
- X: the number of trials needed to obtain the first success.

Density

$$f(x) = pq^{x-1}$$

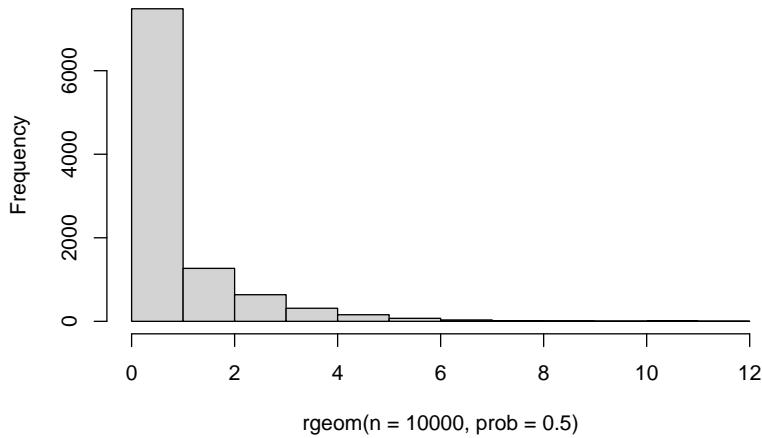
CDF

$$F(x) = 1 - q^x$$

PDF

```
# hist of Geometric distribution with probability of success = 0.5
hist(rgeom(n = 10000, prob = 0.5))
```

Histogram of rgeom(n = 10000, prob = 0.5)

**MGF**

$$m_X(t) = \frac{pe^t}{1 - qe^t}$$

for $t < -\ln(q)$

Mean

$$\mu = \frac{1}{p}$$

Variance

$$\sigma^2 = \text{Var}(X) = \frac{q}{p^2}$$

2.2.6.1.5 Hypergeometric

- The experiment consists of drawing a random sample of size n without replacement and without regard to order from a collection of N objects.
- Of the N objects, r have a trait of interest; $N-r$ do not have the trait
- X is the number of objects in the sample with the trait.

Density

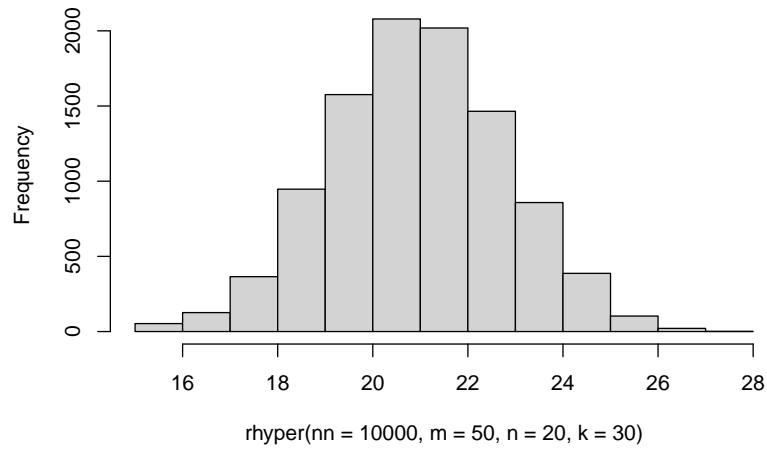
$$f(x) = \frac{\binom{r}{x} \binom{N-r}{n-x}}{\binom{N}{n}}$$

where $\max[0, n - (N - r)] \leq x \leq \min(n, r)$

PDF

```
# hist of hypergeometric distribution with the number of white balls = 50, and the number of black balls = 30
hist(rhyper(nn = 10000 , m=50, n=20, k=30))
```

Histogram of rhyper(nn = 10000, m = 50, n = 20, k = 30)



Mean

$$\mu = E(x) = \frac{nr}{N}$$

Variance

$$\sigma^2 = var(X) = n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right)$$

Note For large N (if $\frac{n}{N} \leq 0.05$), this distribution can be approximated using a Binomial distribution with $p = \frac{r}{N}$

2.2.6.1.6

2.2.6.2 Continuous

2.2.6.2.1 Uniform

- Defined over an interval (a,b) in which the probabilities are “equally likely” for subintervals of equal length.

Density

$$f(x) = \frac{1}{b-a}$$

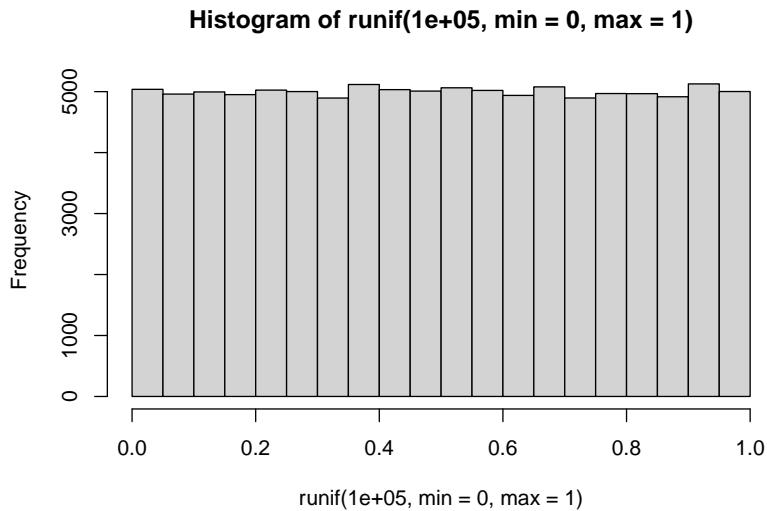
for $a < x < b$

CDF

$$\begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } x > b \end{cases}$$

PDF

```
hist(runif(100000, min = 0, max = 1))
```



MGF

$$\begin{cases} \frac{e^{tb} - e^{ta}}{t(b-a)} & \text{if } t \neq 0 \\ 1 & \text{if } t = 0 \end{cases}$$

Mean

$$\mu = E(X) = \frac{a+b}{2}$$

Variance

$$\sigma^2 = Var(X) = \frac{(b-a)^2}{12}$$

2.2.6.2.2 Gamma

- is used to define the exponential and chi-squared distributions
- The gamma function is defined as:

$$\Gamma(\alpha) = \int_0^\infty z^{\alpha-1} e^{-z} dz$$

where $\alpha > 0$

- Properties of The Gamma function:

– $\Gamma(1) = 1$ + For $\alpha > 1$, $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ + If n is an integer and $n > 1$, then $\Gamma(n) = (n - 1)!$

Density

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$

CDF

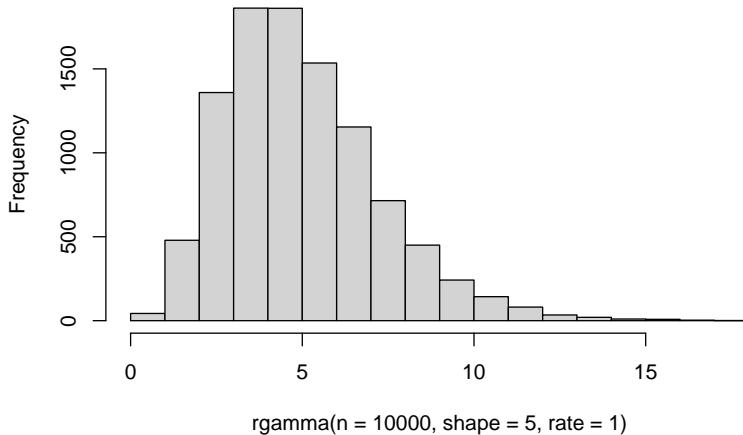
$$F(x, n, \beta) = 1 - \sum_{k=0}^{n-1} \frac{(\frac{x}{\beta})^k e^{-x/\beta}}{k!}$$

for $x > 0$, and $\alpha = n$ (a positive integer)

PDF

```
hist(rgamma(n = 10000, shape = 5, rate = 1))
```

Histogram of rgamma(n = 10000, shape = 5, rate = 1)



MGF

$$m_X(t) = (1 - \beta t)^{-\alpha}$$

where $t < \frac{1}{\beta}$

Mean

$$\mu = E(X) = \alpha\beta$$

Variance

$$\sigma^2 = Var(X) = \alpha\beta^2$$

2.2.6.2.3 Normal $N(\mu, \sigma^2)$

- is symmetric, bell-shaped curve with parameters μ and σ^2
- also known as Gaussian.

Density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

for $-\infty < x, \mu < \infty, \sigma > 0$

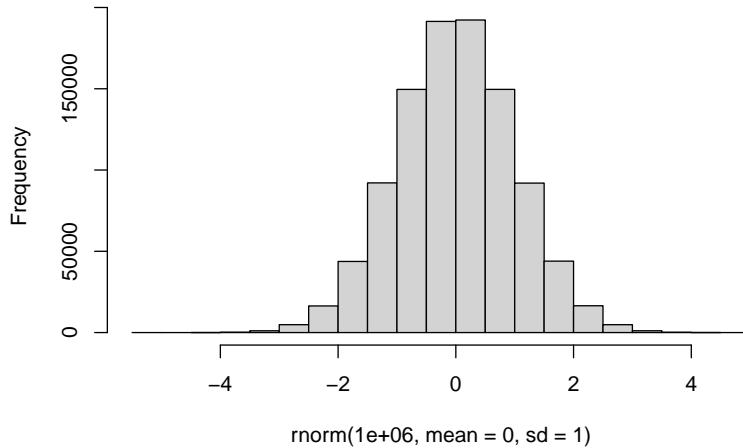
CDF

Use table

PDF

```
hist(rnorm(1000000, mean = 0, sd = 1))
```

Histogram of rnorm(1e+06, mean = 0, sd = 1)



MGF

$$m_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$$

Mean

$$\mu = E(X)$$

Variance

$$\sigma^2 = Var(X)$$

Standard Normal Random Variable

- The normal random variable Z with mean $\mu = 0$ and standard deviation $\sigma = 1$ is called standard normal
- Any normal random variable X with mean μ and standard deviation σ can be converted to the standard normal random variable $Z = \frac{X-\mu}{\sigma}$

Normal Approximation to the Binomial Distribution

Let X be binomial with parameters n and p. For large n (so that (A) $p \leq .5$ and $np > 5$ or (B) $p > .5$ and $nq > 5$), X is approximately normally distributed with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$

When using the normal approximation, add or subtract 0.5 as needed for the continuity correction

Discrete	Approximate Normal (corrected)
$P(X = c)$	$P(c - 0.5 < Y < c + 0.5)$
$P(X < c)$	$P(Y < c - 0.5)$
$P(X \geq c)$	$P(Y < c + 0.5)$
$P(X > c)$	$P(Y > c + 0.5)$
$P(X \leq c)$	$P(Y > c - 0.5)$

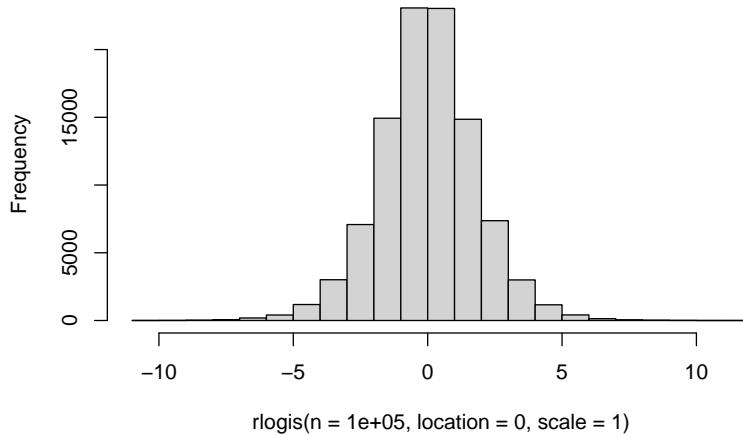
Normal Probability Rule

If X is normally distributed with parameters μ and σ , then

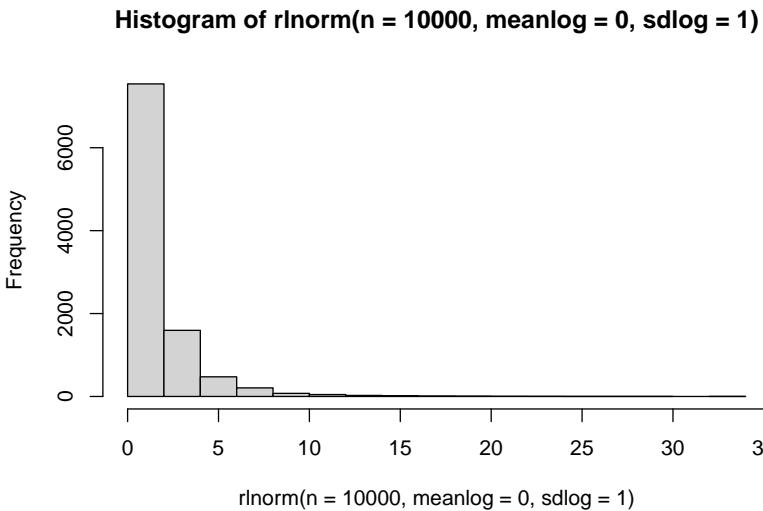
- $P(-\sigma < X - \mu < \sigma) \approx .68 * P(-2\sigma < X - \mu < 2\sigma) \approx .95 * P(-3\sigma < X - \mu < 3\sigma) \approx .997$

2.2.6.2.4 Logistic $\text{Logistic}(\mu, s)$ **PDF**

```
hist(rlogis(n = 100000, location = 0, scale = 1))
```

Histogram of rlogis(n = 1e+05, location = 0, scale = 1)**2.2.6.2.5 Lognomral** $\text{lognormal}(\mu, \sigma^2)$ **PDF**

```
hist(rlnorm(n = 10000, meanlog = 0, sdlog = 1))
```



2.2.6.2.6 Exponential $Exp(\lambda)$

- A special case of the gamma distribution with $\alpha = 1$
- Lack of memory
- λ = rate Within a Poisson process with parameter λ , if W is the waiting time until the occurrence of the first event, then W has an exponential distribution with $\beta = 1/\lambda$

Density

$$f(x) = \frac{1}{\beta} e^{-x/\beta}$$

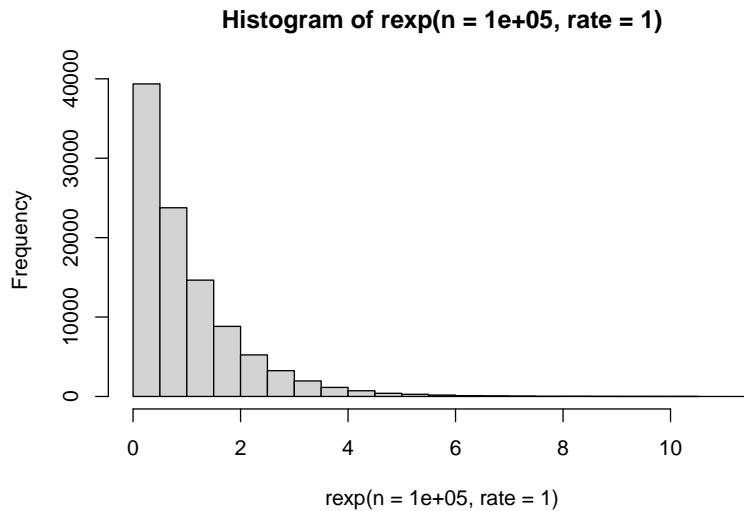
for $x, \beta > 0$

CDF

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 - e^{-x/\beta} & \text{if } x > 0 \end{cases} \quad (2.8)$$

PDF

```
hist(rexp(n = 100000, rate = 1))
```

**MGF**

$$m_X(t) = (1 - \beta t)^{-1}$$

for $t < 1/\beta$ **Mean**

$$\mu = E(X) = \beta$$

Variance

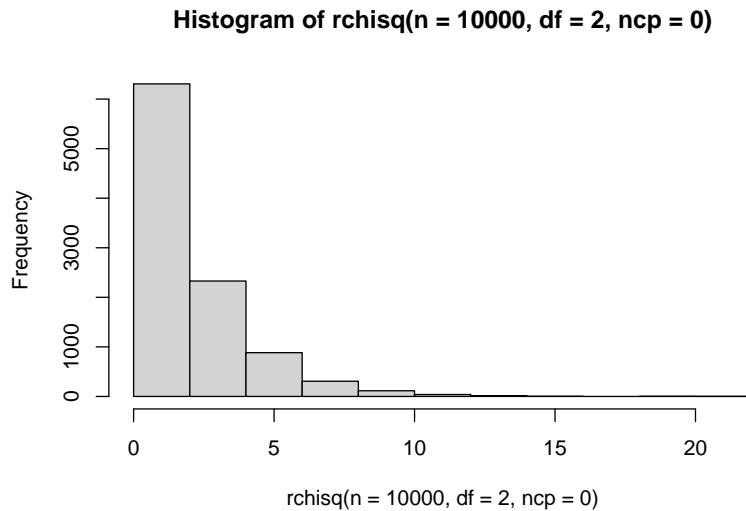
$$\sigma^2 = Var(X) = \beta^2$$

2.2.6.2.7 Chi-squared $\chi^2 = \chi^2(k)$

- A special case of the gamma distribution with $\beta = 2$, and $\alpha = \gamma/2$ for a positive integer γ
- The random variable X is denoted χ^2_γ and is said to have a chi-squared distribution with γ degrees of freedom.

Density Use density for Gamma Distribution with $\beta = 2$ and $\alpha = \gamma/2$ **CDF** Use table**PDF**

```
hist(rchisq(n = 10000, df=2, ncp = 0))
```

**MGF**

$$m_X(t) = (1 - 2t)^{-\gamma/2}$$

Mean

$$\mu = E(X) = \gamma$$

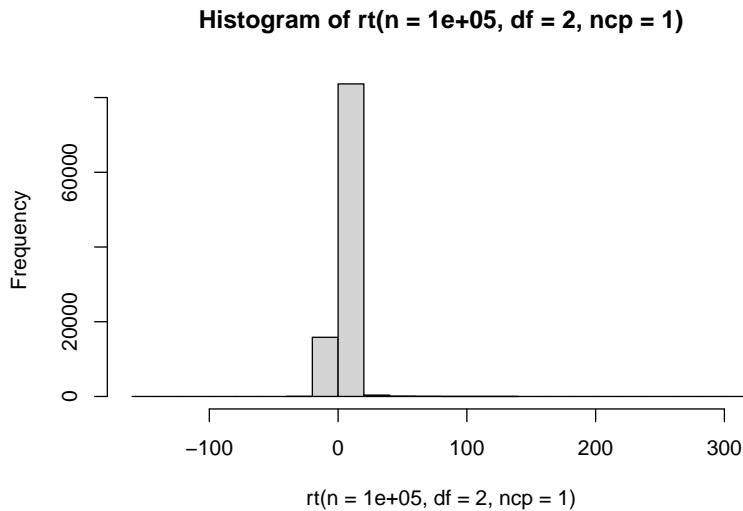
Variance

$$\sigma^2 = Var(X) = 2\gamma$$

2.2.6.2.8 Student T $T(v)$

- $T = \frac{Z}{\sqrt{\chi^2_{\gamma}/\gamma}}$, where Z is standard normal follows a student-t distribution with γ dof
- The distribution is symmetric, bell-shaped , with a mean of $\mu = 0$

```
hist(rt(n = 100000, df=2, ncp =1))
```

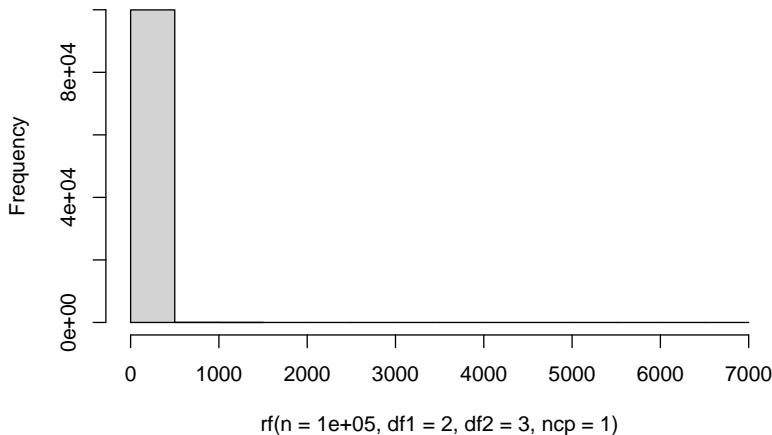


2.2.6.2.9 F-Distribution $F(d_1, d_2)$

- F distribution is strictly positive
- $F = \frac{\chi_{\gamma_1}^2 / \gamma_1}{\chi_{\gamma_2}^2 / \gamma_2}$ follows an F distribution with dof γ_1 and γ_2 , where $\chi_{\gamma_1}^2$ and $\chi_{\gamma_2}^2$ are independent chi-squared random variables.
- The distribution is asymmetric and never negative.

PDF

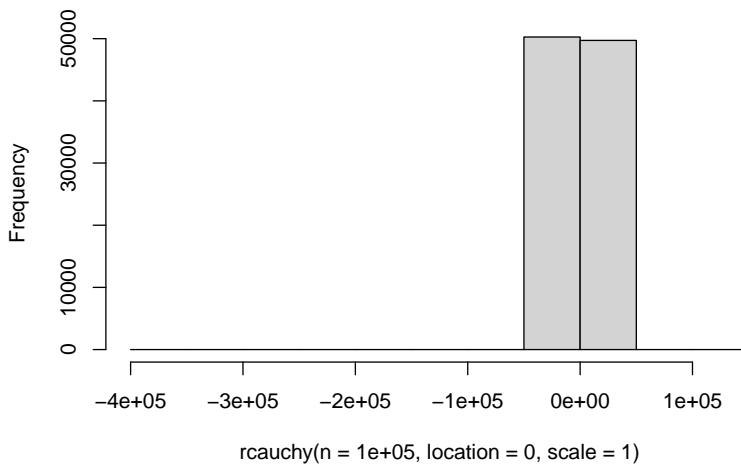
```
hist(rf(n = 100000, df1=2, df2=3, ncp=1))
```

Histogram of rf(n = 1e+05, df1 = 2, df2 = 3, ncp = 1)`rf(n = 1e+05, df1 = 2, df2 = 3, ncp = 1)`

2.2.6.2.10 Cauchy Central Limit Theorem and Weak Law do not apply to Cauchy because it does not have finite mean and finite variance

PDF

```
hist(rcauchy(n = 100000, location = 0, scale = 1))
```

Histogram of rcauchy(n = 1e+05, location = 0, scale = 1)`rcauchy(n = 1e+05, location = 0, scale = 1)`

2.2.6.2.11 Multivariate Normal Distribution Let \mathbf{y} be a p-dimensional multivariate normal (MVN) rv with mean μ and variance Σ . Then, the density of \mathbf{y} is

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mu)' \Sigma^{-1}(\mathbf{y} - \mu)\right)$$

We have $\mathbf{y} \sim N_p(\mu, \Sigma)$

Properties:

- Let $\mathbf{A}_{r \times p}$ be a fixed matrix. then $\mathbf{A}\mathbf{y} \sim N_r(\mathbf{A}\mu, \mathbf{A}\mathbf{A}')$. Note that $r \leq p$ and all rows of \mathbf{A} must be linearly independent to guarantee that $\mathbf{A}\mathbf{A}'$ is non-singular.
- Let \mathbf{G} be a matrix such that $\mathbf{G}'\mathbf{G} = \mathbf{I}$. then, $\mathbf{G}'\mathbf{y} \sim N_p(\mathbf{G}'\mu, \mathbf{I})$ and $\mathbf{G}'(\mathbf{y} - \mu) \sim N_p(\mathbf{0}, \mathbf{I})$.
- Any fixed linear combination of y_1, \dots, y_p say $\mathbf{c}'\mathbf{y}$, follows $\mathbf{c}'\mathbf{y} \sim N_1(\mathbf{c}'\mu, \mathbf{c}'\mathbf{c})$

Large Sample Properties

Suppose that y_1, \dots, y_n are a random sample from some population with mean μ and variance-variance matrix Σ

$$\mathbf{Y} \sim MVN(\mu, \Sigma)$$

Then

- $\bar{\mathbf{y}} = \frac{1}{n} \sum_{i=1}^n \mathbf{y}_i$ is a consistent estimator for μ
- $\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'$ is a consistent estimator for Σ
- Multivariate Central Limit Theorem: Similar to the univariate case, $\sqrt{n}(\bar{\mathbf{y}} - \mu) \sim N_p(\mathbf{0}, \Sigma)$ when n is large relative to p (e.g., $n \geq 25p$), which is equivalent to $\bar{\mathbf{y}} \sim N_p(\mu, \Sigma/n)$
- Wald's Theorem: $n(\bar{\mathbf{y}} - \mu)' \mathbf{S}^{-1}(\bar{\mathbf{y}} - \mu) \sim \chi^2_{(p)}$ when n is large relative to p .

2.2.6.2.12

2.3 General Math

Chebyshev's Inequality Let X be a random variable with mean μ and standard deviation σ . Then for any positive number k :

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

Chebyshev's Inequality does not require that X be normally distributed

Maclaurin series expansion for

$$e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \dots$$

Geometric series:

$$s_n = \sum_{k=1}^n ar^{n-1} = \frac{a(1 - r^n)}{1 - r}$$

if $|r| < 1$

$$s = \sum_{k=1}^{\infty} ar^{n-1} = \frac{a}{1 - r}$$

2.3.1 Law of large numbers

Let X_1, X_2, \dots be an infinite sequence of independent and identically distributed (i.i.d) Then, the sample average is

$$\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$$

converges to the expected value ($\bar{X}_n \rightarrow \mu$) as $n \rightarrow \infty$

$$Var(X_i) = Var\left(\frac{1}{n}(X_1 + \dots + X_n)\right) = \frac{1}{n^2}Var(X_1 + \dots + X_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

The difference between Weak Law and Strong Law regards the mode of convergence

2.3.1.1 Weak Law

The sample average converges in probability towards the expected value

$$\bar{X}_n \rightarrow^p \mu$$

when $n \rightarrow \infty$

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

The sample mean from a iid random sample ($\{x_i\}_{i=1}^n$) from any population with a finite mean and finite variance σ^2 is a consistent estimation for the population mean μ

$$plim(\bar{x}) = plim(n^{-1} \sum_{i=1}^n x_i) = \mu$$

2.3.1.2 Strong Law

The sample average converges almost surely to the expected value

$$\bar{X}_n \rightarrow^{a.s} \mu$$

when $n \rightarrow \infty$

Equivalently,

$$P(\lim_{n \rightarrow \infty} \bar{X}_n = \mu) = 1$$

2.3.2 Law of Iterated Expectation

Let X, Y be random variables. Then,

$$E(X) = E(E(X|Y))$$

means that the expected value of X can be calculated from the probability distribution of $X|Y$ and Y

2.3.3 Convergence

2.3.3.1 Convergence in Probability

- $n \rightarrow \infty$, an estimator (random variable) that is close to the true value.
- The random variable θ_n converges in probability to a constant c if

$$\lim_{n \rightarrow \infty} P(|\theta_n - c| \geq \epsilon) = 0$$

for any positive ϵ

Notation

$$plim(\theta_n) = c$$

Equivalently,

$$\theta_n \xrightarrow{p} c$$

Properties of Convergence in Probability

- Slutsky's Theorem: for a continuous function $g(\cdot)$, if $plim(\theta_n) = \theta$ then $plim(g(\theta_n)) = g(\theta)$
- if $\gamma_n \xrightarrow{p} \gamma$ then
 - $plim(\theta_n + \gamma_n) = \theta + \gamma + plim(\theta_n \gamma_n) = \theta\gamma + plim(\theta_n/\gamma_n) = \theta/\gamma$ if $\gamma \neq 0$
 - Also hold for random vectors/ matrices

2.3.3.2 Convergence in Distribution

- As $n \rightarrow \infty$, the distribution of a random variable may converge towards another ("fixed") distribution.
- The random variable X_n with CDF $F_n(x)$ converges in distribution to a random variable X with CDF $F(X)$ if

$$\lim_{n \rightarrow \infty} |F_n(x) - F(x)| = 0$$

at all points of continuity of $F(X)$

Notation $F(x)$ is the limiting distribution of X_n or $X_n \xrightarrow{d} X$

- $E(X)$ is the limiting mean (asymptotic mean)
- $\text{Var}(X)$ is the limiting variance (asymptotic variance)

Note

$$E(X) \neq \lim_{n \rightarrow \infty} E(X_n) \text{Avar}(X_n) \neq \lim_{n \rightarrow \infty} \text{Var}(X_n)$$

Properties of Convergence in Distribution

- Continuous Mapping Theorem: for a continuous function $g(\cdot)$, if $X_n \xrightarrow{d} g(X)$ then $g(X_n) \xrightarrow{d} g(X)$
- if $Y_n \xrightarrow{d} c$, then
 - $X_n + Y_n \xrightarrow{d} X + c$
 - $Y_n X_n \xrightarrow{d} cX$
 - $X_n Y_n \xrightarrow{d} X/c$ if $c \neq 0$
- also hold for random vectors/matrices

2.3.3.3 Summary

Properties of Convergence

Probability	Distribution
Slutsky's Theorem: for a continuous function $g(\cdot)$, if $\text{plim}(\theta_n) = \theta$ then $\text{plim}(g(\theta_n)) = g(\theta)$ if $\gamma_n \xrightarrow{p} \gamma$ then $\text{plim}(\theta_n + \gamma_n) = \theta + \gamma$ $\text{plim}(\theta_n \gamma_n) = \theta \gamma$ $\text{plim}(\theta_n / \gamma_n) = \theta / \gamma$ if $\gamma \neq 0$	Continuous Mapping Theorem: for a continuous function $g(\cdot)$, if $X_n \xrightarrow{d} g(X)$ then $g(X_n) \xrightarrow{d} g(X)$ if $Y_n \xrightarrow{d} c$, then $X_n + Y_n \xrightarrow{d} X + c$ $Y_n X_n \xrightarrow{d} cX$ $X_n Y_n \xrightarrow{d} X/c$ if $c \neq 0$

Convergence in Probability is stronger than Convergence in Distribution. However, Convergence in Distribution does not guarantee Convergence in Probability

2.3.4 Sufficient Statistics

Likelihood

- describes the extent to which the sample provides support for any particular parameter value.
- Higher support corresponds to a higher value for the likelihood
- The exact value of any likelihood is **meaningless**,
- The relative value, (i.e., comparing two values of θ), is **informative**.

$$L(\theta_0; y) = P(Y = y | \theta = \theta_0) = f_Y(y; \theta_0)$$

Likelihood Ratio

$$\frac{L(\theta_0; y)}{L(\theta_1; y)}$$

Likelihood Function

For a given sample, you can create likelihoods for all possible values of θ , which is called *likelihood function*

$$L(\theta) = L(\theta; y) = f_Y(y; \theta)$$

In a sample of size n , the likelihood function takes the form of a product

$$L(\theta) = \prod_{i=1}^n f_i(y_i; \theta)$$

Equivalently, the log likelihood function

$$l(\theta) = \sum_{i=1}^n \log f_i(y_i; \theta)$$

Sufficient statistics

- A statistic, $T(y)$, is any quantity that can be calculated purely from a sample (independent of θ)
- A statistic is **sufficient** if it conveys all the available information about the parameter.

$$L(\theta; y) = c(y)L^*(\theta; T(y))$$

Nuisance parameters If we are interested in a parameter (e.g., mean). Other parameters requiring estimation (e.g., standard deviation) are **nuisance** parameters. We can replace nuisance parameters in likelihood function with their estimates to create a **profile likelihood**.

2.3.5 Parameter transformations

log-odds transformation

$$\text{Logodds} = g(\theta) = \ln\left[\frac{\theta}{1-\theta}\right]$$

log transformation

2.4 Data Import/Export

Extended Manual by R

Table 2.6: Table by Rio Vignette

Format	Typical Extension	Import Package	Export Package	Installed by Default
Comma-separated data	.csv	data.table	data.table	Yes
Pipe-separated data	.psv	data.table	data.table	Yes
Tab-separated data	.tsv	data.table	data.table	Yes
CSVY (CSV + YAML metadata header)	.csvy	data.table	data.table	Yes
SAS	.sas7bdat	haven	haven	Yes
SPSS	.sav	haven	haven	Yes
SPSS (compressed)	.zsav	haven	haven	Yes
Stata	.dta	haven	haven	Yes
SAS XPORT	.xpt	haven	haven	Yes

Format	Typical Extension	Import Package	Export Package	Installed by Default
SPSS Portable	.por	haven		Yes
Excel	.xls	readxl		Yes
Excel	.xlsx	readxl	openxlsx	Yes
R syntax	.R	base	base	Yes
Saved R objects	.RData, .rda	base	base	Yes
Serialized R objects	.rds	base	base	Yes
Epiinfo	.rec	foreign		Yes
Minitab	.mtp	foreign		Yes
Systat	.syd	foreign		Yes
“XBASE” database files	.dbf	foreign	foreign	Yes
Weka	.arff	foreign	foreign	Yes
Attribute-Relation File Format				
Data Interchange Format	.dif	utils		Yes
Fortran data	no recognized extension	utils		Yes
Fixed-width format data	.fwf	utils	utils	Yes
gzip comma-separated data	.csv.gz	utils	utils	Yes
Apache Arrow (Parquet)	.parquet	arrow	arrow	No
EViews	.wfl	hexView		No
Feather	.feather	feather	feather	No
R/Python interchange format				
Fast Storage	.fst	fst	fst	No
JSON	.json	jsonlite	jsonlite	No
Matlab	.mat	rmatio	rmatio	No
OpenDocument Spreadsheet	.ods	readODS	readODS	No

Format	Typical Extension	Import Package	Export Package	Installed by Default
HTML Tables	.html	xml2	xml2	No
Shallow XML documents	.xml	xml2	xml2	No
YAML	.yml	yaml	yaml	No
Clipboard	default is tsv	clipr	clipr	No
Google Sheets	as Comma-separated data			

R limitations:

- By default, R use 1 core in CPU
- R puts data into memory (limit around 2-4 GB), while SAS uses data from files on demand
- Categorization
 - Medium-size file: within RAM limit, around 1-2 GB
 - Large file: 2-10 GB, there might be some workaround solution
 - Very large file > 10 GB, you have to use distributed or parallel computing

Solutions:

- buy more RAM
- HPC packages
 - Explicit Parallelism
 - Implicit Parallelism
 - Large Memory
 - Map/Reduce
- specify number of rows and columns, typically including command `nrow =`
- Use packages that store data differently

- `bigrmemory`, `biganalytics`, `bigtabulate`, `synchronicity`, `bigalgebra`, `bigvideo` use C++ to store matrices, but also support one class type
- For multiple class types, use `ff` package
- Very Large datasets use
 - `RHadoop` package
 - `HadoopStreaming`
 - `Rhipe`

2.4.1 Medium size

```
library("rio")
```

To import multiple files in a directory

```
str(import_list(dir()), which = 1)
```

To export a single data file

```
export(data, "data.csv")
export(data, "data.dta")
export(data, "data.txt")
export(data, "data_cyl.rds")
export(data, "data.rdata")
export(data, "data.R")
export(data, "data.csv.zip")
export(data, "list.json")
```

To export multiple data files

```
export(list(mtcars = mtcars, iris = iris), "data_file_type") # where data_file_type should substi
```

To convert between data file types

```
# convert Stata to SPSS
convert("data.dta", "data.sav")
```

2.4.2 Large size

Use R on a cluster

- Amazon Web Service (AWS): \$1/hr

Import files as chunks

```
file_in    <- file("in.csv","r")
chunk_size <- 100000 # choose the best size for you
x          <- readLines(file_in, n=chunk_size)
```

`data.table` method

```
require(data.table)
mydata = fread("in.csv", header = T)
```

`ff` package: this method does not allow you to pass connections

```
library("ff")
x <- read.csv.ffdf(
  file = "file.csv",
  nrow = 10,
  header = TRUE,
  VERBOSE = TRUE,
  first.rows = 10000,
  next.rows = 50000,
  colClasses = NA
)
```

`bigmemory` package

```
my_data <- read.big.matrix('in.csv', header = T)
```

`sqldf` package

```
library(sqldf)
my_data <- read.csv.sql('in.csv')

iris2 <- read.csv.sql("iris.csv",
  sql = "select * from file where Species = 'setosa' ")
```

```
library(RMySQL)
```

RQLite package

- Download SQLite, pick “A bundle of command-line tools for managing SQLite database files” for Window 10
- Unzip file, and open `sqlite3.exe`.
- Type in the prompt
 - `sqlite> .cd 'C:\Users\data'` specify path to your desired directory
 - `sqlite> .open database_name.db` to open a database
 - To import the CSV file into the database
 - * `sqlite> .mode csv` specify to SQLite that the next file is .csv file
 - * `sqlite> .import file_name.csv database_name` to import the csv file to the database
 - `sqlite> .exit` After you’re done, exit the sqlite program

```
library(DBI)
library(dplyr)
library("RSQLite")
setwd("")
con <- dbConnect(RSQLite::SQLite(), "data_base.db")
tbl <- tbl(con, "data_table")
tbl %>%
  filter() %>%
  select() %>%
  collect() # to actually pull the data into the workspace
dbDisconnect(con)
```

arrow package

```
library("arrow")
read_csv_arrow()
```

vroom package

```
library(vroom)
spec(vroom(file_path))
compressed <- vroom_example("mtcars.csv.zip")
vroom(compressed)
```

data.table package

```
s = fread("sample.csv")
```

Comparisons regarding storage space

```
test = ff::read.csv.ffdf(file = "")  
object.size(test) # worst  
  
test1 = data.table::fread(file = "")  
object.size(test1) # best  
  
test2 = readr::read_csv("")  
object.size(test2) # 2nd  
  
test3 = vroom(file = "")  
object.size(test3) # equal to read_csv
```

To work with big data, you can convert it to `csv.gz`, but since typically, R would require you to load the whole data then export it. With data greater than 10 GB, we have to do it sequentially. Even though `read.csv` is much slower than `readr::read_csv`, we still have to use it because it can pass connection, and it allows you to loop sequentially. On the other, because currently `readr::read_csv` does not have the `skip` function, and even if we can use the `skip`, we still have to read and skip lines in previous loop.

For example, say you `read_csv(, n_max = 100, skip = 0)` and then `read_csv(, n_max = 200, skip = 100)` you actually have to read again the first 100 rows. However, `read.csv` without specifying anything, will continue at the 100 mark.

Notice, sometimes you might have error looking like this

“Error in (function (con, what, n = 1L, size = NA_integer_, signed = TRUE,
: can only read from a binary connection”

then you can change it instead of `r` in the connection into `rb`. Even though an author of the package suggested that `file` should be able to recognize the appropriate form, so far I did not prevail.

2.5 Data Manipulation

```
# load packages  
library(tidyverse)  
library(lubridate)
```

```

x <- c(1, 4, 23, 4, 45)
n <- c(1, 3, 5)
g <- c("M", "M", "F")
df <- data.frame(n, g)
df
#>   n  g
#> 1 1  M
#> 2 3  M
#> 3 5  F
str(df)
#> 'data.frame':   3 obs. of  2 variables:
#> $ n: num  1 3 5
#> $ g: chr  "M" "M" "F"

#Similarly
df <- tibble(n, g)
df
#> # A tibble: 3 x 2
#>       n  g
#>   <dbl> <chr>
#> 1     1  M
#> 2     3  M
#> 3     5  F
str(df)
#> tibble [3 x 2] (S3:tbl_df/tbl/data.frame)
#> $ n: num [1:3] 1 3 5
#> $ g: chr [1:3] "M" "M" "F"

# list form
lst <- list(x, n, g, df)
lst
#> [[1]]
#> [1] 1 4 23 4 45
#>
#> [[2]]
#> [1] 1 3 5
#>
#> [[3]]
#> [1] "M" "M" "F"
#>
#> [[4]]
#> # A tibble: 3 x 2
#>       n  g
#>   <dbl> <chr>
#> 1     1  M

```

```

#> 2      3 M
#> 3      5 F

# Or
lst2 <- list(num = x, size = n, sex = g, data = df)
lst2
#> $num
#> [1] 1 4 23 4 45
#>
#> $size
#> [1] 1 3 5
#>
#> $sex
#> [1] "M" "M" "F"
#>
#> $data
#> # A tibble: 3 x 2
#>       n   g
#>     <dbl> <chr>
#> 1     1   M
#> 2     3   M
#> 3     5   F

# Or
lst3 <- list(x = c(1, 3, 5, 7),
             y = c(2, 2, 2, 4, 5, 5, 5, 6),
             z = c(22, 3, 3, 3, 5, 10))
lst3
#> $x
#> [1] 1 3 5 7
#>
#> $y
#> [1] 2 2 2 4 5 5 5 6
#>
#> $z
#> [1] 22 3 3 3 5 10

# find the means of x, y, z.

# can do one at a time
mean(lst3$x)
#> [1] 4
mean(lst3$y)
#> [1] 3.875
mean(lst3$z)

```

```
#> [1] 7.666667

# list apply
lapply(lst3, mean)
#> $x
#> [1] 4
#>
#> $y
#> [1] 3.875
#>
#> $z
#> [1] 7.666667

# OR
sapply(lst3, mean)
#>      x      y      z
#> 4.000000 3.875000 7.666667

# Or, tidyverse function map()
map(lst3, mean)
#> $x
#> [1] 4
#>
#> $y
#> [1] 3.875
#>
#> $z
#> [1] 7.666667

# The tidyverse requires a modified map function called map_dbl()
map_dbl(lst3, mean)
#>      x      y      z
#> 4.000000 3.875000 7.666667

# Binding
dat01 <- tibble(x = 1:5, y = 5:1)
dat01
#> # A tibble: 5 x 2
#>       x     y
#>   <int> <int>
#> 1     1     5
#> 2     2     4
#> 3     3     3
#> 4     4     2
```

```
#> 5      5      1
dat02 <- tibble(x = 10:16, y = x/2)
dat02
#> # A tibble: 7 x 2
#>   x     y
#>   <int> <dbl>
#> 1 10     5
#> 2 11     5.5
#> 3 12     6
#> 4 13     6.5
#> 5 14     7
#> 6 15     7.5
#> 7 16     8
dat03 <- tibble(z = runif(5)) # 5 random numbers from interval (0,1)
dat03
#> # A tibble: 5 x 1
#>   z
#>   <dbl>
#> 1 0.356
#> 2 0.699
#> 3 0.844
#> 4 0.589
#> 5 0.384

# row binding
bind_rows(dat01, dat02, dat03)
#> # A tibble: 17 x 2
#>   x     y
#>   <int> <dbl>
#> 1 1     5
#> 2 2     4
#> 3 3     3
#> 4 4     2
#> 5 5     1
#> 6 10    5
#> 7 11    5.5
#> 8 12    6
#> 9 13    6.5
#> 10 14   7
#> 11 15   7.5
#> 12 16   8
#> 13 1     5
#> 14 2     4
#> 15 3     3
#> 16 4     2
```

```
#> 17      5   1

# use ".id" argument to create a new column that contains an identifier for the original data.
bind_rows(dat01, dat02, .id = "id")
#> # A tibble: 12 x 3
#>   id     x     y
#>   <chr> <int> <dbl>
#> 1 1      1     5
#> 2 1      2     4
#> 3 1      3     3
#> 4 1      4     2
#> 5 1      5     1
#> 6 2     10    5
#> 7 2     11   5.5
#> 8 2     12    6
#> 9 2     13   6.5
#> 10 2    14    7
#> 11 2    15   7.5
#> 12 2    16    8

# with name
bind_rows("dat01" = dat01, "dat02" = dat02, .id = "id")
#> # A tibble: 12 x 3
#>   id     x     y
#>   <chr> <int> <dbl>
#> 1 dat01  1     5
#> 2 dat01  2     4
#> 3 dat01  3     3
#> 4 dat01  4     2
#> 5 dat01  5     1
#> 6 dat02  10    5
#> 7 dat02  11   5.5
#> 8 dat02  12    6
#> 9 dat02  13   6.5
#> 10 dat02 14    7
#> 11 dat02 15   7.5
#> 12 dat02 16    8

# bind_rows() also works on lists of data frames
list01 <- list("dat01" = dat01, "dat02" = dat02)
list01
#> $dat01
#> # A tibble: 5 x 2
#>   x     y
#>   <int> <int>
```

```
#> 1     1     5
#> 2     2     4
#> 3     3     3
#> 4     4     2
#> 5     5     1
#>
#> #> $dat02
#> # A tibble: 7 x 2
#>       x     y
#>   <int> <dbl>
#> 1     10    5
#> 2     11   5.5
#> 3     12    6
#> 4     13   6.5
#> 5     14    7
#> 6     15   7.5
#> 7     16    8
bind_rows(list01)
#> # A tibble: 12 x 2
#>       x     y
#>   <int> <dbl>
#> 1     1     5
#> 2     2     4
#> 3     3     3
#> 4     4     2
#> 5     5     1
#> 6     10    5
#> 7     11   5.5
#> 8     12    6
#> 9     13   6.5
#> 10    14    7
#> 11    15   7.5
#> 12    16    8
bind_rows(list01, .id = "source")
#> # A tibble: 12 x 3
#>   source     x     y
#>   <chr> <int> <dbl>
#> 1 dat01     1     5
#> 2 dat01     2     4
#> 3 dat01     3     3
#> 4 dat01     4     2
#> 5 dat01     5     1
#> 6 dat02    10    5
#> 7 dat02    11   5.5
#> 8 dat02    12    6
```

```

#> 9 dat02     13   6.5
#> 10 dat02    14   7
#> 11 dat02    15  7.5
#> 12 dat02    16   8

# The extended example below demonstrates how this can be very handy.

# column binding
bind_cols(dat01, dat03)
#> # A tibble: 5 x 3
#>       x     y     z
#>   <int> <int> <dbl>
#> 1     1     5  0.356
#> 2     2     4  0.699
#> 3     3     3  0.844
#> 4     4     2  0.589
#> 5     5     1  0.384

# Regular expressions -----
names <- c("Ford, MS", "Jones, PhD", "Martin, Phd", "Huck, MA, MLS")

# pattern: first comma and everything after it
str_remove(names, pattern = ", [[:print:]]+")
#> [1] "Ford"    "Jones"   "Martin"  "Huck"

# [[:print:]]+ = one or more printable characters

# Reshaping -----
# Example of a wide data frame. Notice each person has multiple test scores
# that span columns.
wide <- data.frame(name=c("Clay","Garrett","Addison"),
                    test1=c(78, 93, 90),
                    test2=c(87, 91, 97),
                    test3=c(88, 99, 91))
wide
#>      name test1 test2 test3
#> 1 Clay    78    87    88
#> 2 Garrett  93    91    99
#> 3 Addison  90    97    91

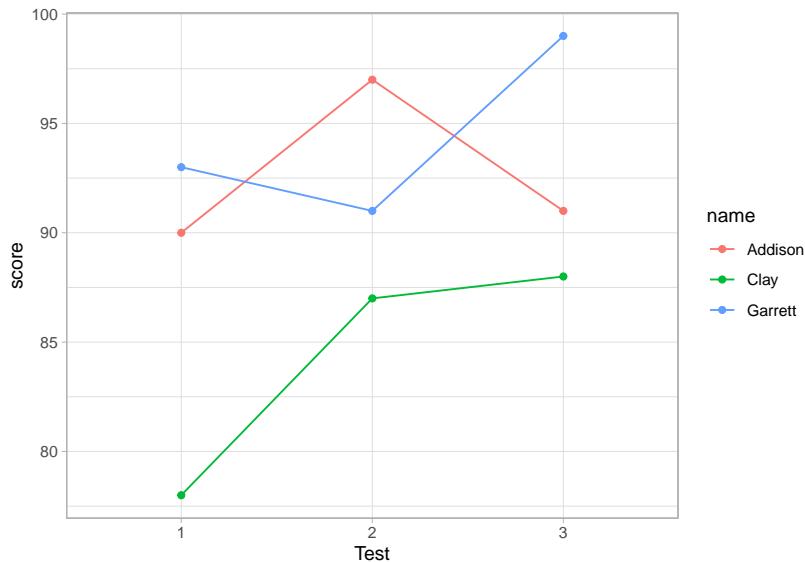
# Example of a long data frame. This is the same data as above, but in long
# format. We have one row per person per test.

```

```
long <- data.frame(name=rep(c("Clay", "Garrett", "Addison"), each=3),
                     test=rep(1:3, 3),
                     score=c(78, 87, 88, 93, 91, 99, 90, 97, 91))
long
#>      name test score
#> 1    Clay   1    78
#> 2    Clay   2    87
#> 3    Clay   3    88
#> 4 Garrett  1    93
#> 5 Garrett  2    91
#> 6 Garrett  3    99
#> 7 Addison  1    90
#> 8 Addison  2    97
#> 9 Addison  3    91

# mean score per student
aggregate(score ~ name, data = long, mean)
#>      name    score
#> 1 Addison 92.66667
#> 2 Clay    84.33333
#> 3 Garrett 94.33333
aggregate(score ~ test, data = long, mean)
#>    test    score
#> 1     1 87.00000
#> 2     2 91.66667
#> 3     3 92.66667

# line plot of scores over test, grouped by name
ggplot(long, aes(x = factor(test), y = score, color = name, group = name)) +
  geom_point() +
  geom_line() +
  xlab("Test")
```



```
##### reshape wide to long
pivot_longer(wide, test1:test3, names_to = "test", values_to = "score")
#> # A tibble: 9 x 3
#>   name    test  score
#>   <chr>   <chr> <dbl>
#> 1 Clay    test1    78
#> 2 Clay    test2    87
#> 3 Clay    test3    88
#> 4 Garrett test1    93
#> 5 Garrett test2    91
#> 6 Garrett test3    99
#> 7 Addison test1    90
#> 8 Addison test2    97
#> 9 Addison test3    91

# Or
pivot_longer(wide, -name, names_to = "test", values_to = "score")
#> # A tibble: 9 x 3
#>   name    test  score
#>   <chr>   <chr> <dbl>
#> 1 Clay    test1    78
#> 2 Clay    test2    87
#> 3 Clay    test3    88
#> 4 Garrett test1    93
#> 5 Garrett test2    91
```

```

#> 6 Garrett test3    99
#> 7 Addison test1   90
#> 8 Addison test2   97
#> 9 Addison test3   91

# drop "test" from the test column with names_prefix argument
pivot_longer(wide, -name, names_to = "test", values_to = "score",
             names_prefix = "test")
#> # A tibble: 9 x 3
#>   name    test  score
#>   <chr>   <chr> <dbl>
#> 1 Clay     1      78
#> 2 Clay     2      87
#> 3 Clay     3      88
#> 4 Garrett  1      93
#> 5 Garrett  2      91
#> 6 Garrett  3      99
#> 7 Addison  1      90
#> 8 Addison  2      97
#> 9 Addison  3      91

##### reshape long to wide
pivot_wider(long, name, names_from = test, values_from = score)
#> # A tibble: 3 x 4
#>   name     `1`   `2`   `3`
#>   <chr>   <dbl> <dbl> <dbl>
#> 1 Clay      78    87    88
#> 2 Garrett   93    91    99
#> 3 Addison   90    97    91

# using the names_prefix argument lets us prepend text to the column names.
pivot_wider(long, name, names_from = test, values_from = score,
            names_prefix = "test")
#> # A tibble: 3 x 4
#>   name    test1 test2 test3
#>   <chr>   <dbl> <dbl> <dbl>
#> 1 Clay      78    87    88
#> 2 Garrett   93    91    99
#> 3 Addison   90    97    91

```

The verbs of data manipulation

- select: selecting (or not selecting) columns based on their names (eg: select columns Q1 through Q25)

- slice: selecting (or not selecting) rows based on their position (eg: select rows 1:10)
- mutate: add or derive new columns (or variables) based on existing columns (eg: create a new column that expresses measurement in cm based on existing measure in inches)
- rename: rename variables or change column names (eg: change “GraduationRate100” to “grad100”)
- filter: selecting rows based on a condition (eg: all rows where gender = Male)
- arrange: ordering rows based on variable(s) numeric or alphabetical order (eg: sort in descending order of Income)
- sample: take random samples of data (eg: sample 80% of data to create a “training” set)
- summarize: condense or aggregate multiple values into single summary values (eg: calculate median income by age group)
- group_by: convert a tbl into a grouped tbl so that operations are performed “by group”; allows us to summarize data or apply verbs to data by groups (eg, by gender or treatment)
- the pipe: %>%
- Use Ctrl + Shift + M (Win) or Cmd + Shift + M (Mac) to enter in RStudio
- The pipe takes the output of a function and “pipes” into the first argument of the next function.

I. BASIC

Chapter 3

Descriptive Statistics

When you have an area of interest that you want to research, a problem that you want to solve, a relationship that you want to investigate, theoretical and empirical processes will help you.

Estimand is defined as “a quantity of scientific interest that can be calculated in the population and does not change its value depending on the data collection design used to measure it (i.e., it does not vary with sample size and survey design, or the number of nonrespondents, or follow-up efforts).” (Rubin, 1996)

Estimands include:

- population means
- Population variances
- correlations
- factor loading
- regression coefficients

3.1 Numerical Measures

There are differences between a population and a sample

Measures of	Category	Population	Sample
-	What is it?	Reality	A small fraction of reality (inference)
-	Characteristics described by	Parameters	Statistics

Measures of Category	Population	Sample
Central Mean	$\mu = E(Y)$	$\hat{\mu} = \bar{y}$
Ten-dency		
Central Median	50-th percentile	$y_{(\frac{n+1}{2})}$
Ten-dency		
Dispersion	$\sigma^2 = var(Y) = E(Y - \mu)^2$	$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i^2 - n\bar{y}^2)$
Coefficient of Variation	$\frac{\sigma}{\mu}$	$\frac{s}{\bar{y}}$
Dispersion	Interquartile Range	difference between 25th and 75th percentiles. Robust to outliers
Shape	Skewness Standardized 3rd central moment (unitless)	$g_1 = \frac{\mu_3}{\mu_2^{3/2}}$ $\hat{g}_1 = \frac{m_3}{m_2 \sqrt{m_2}}$
Shape	Central moments	$\mu = E(Y)$ $m_2 = \sum_{i=1}^n (y_i - \bar{y})^2 / n$ $\mu_2 = \sigma^2 = E(Y - \mu)^2$ $m_3 = \sum_{i=1}^n (y_i - \bar{y})^3 / n$ $\mu_3 = E(Y - \mu)^3$ $\mu_4 = E(Y - \mu)^4$
Shape	Kurtosis (peakedness and tail thickness) Standardized 4th central moment	$g_2^* = \frac{E(Y - \mu)^4}{\sigma^4}$ $\hat{g}_2 = \frac{m_4}{m_2^2} - 3$

Note:

- Order Statistics: $y_{(1)}, y_{(2)}, \dots, y_{(n)}$ where $y_{(1)} < y_{(2)} < \dots < y_{(n)}$
- Coefficient of variation: standard deviation over mean. This metric is stable, dimensionless statistic for comparison.
- Symmetric: mean = median, skewness = 0
- Skewed right: mean > median, skewness > 0
- Skewed left: mean < median, skewness < 0
- Central moments: $\mu = E(Y)$, $\mu_2 = \sigma^2 = E(Y - \mu)^2$, $\mu_3 = E(Y - \mu)^3$, $\mu_4 = E(Y - \mu)^4$

- For normal distributions, $\mu_3 = 0$, so $g_1 = 0$
- \hat{g}_1 is distributed approximately as $N(0,6/n)$ if sample is from a normal population. (valid when $n > 150$)
 - For large samples, inference on skewness can be based on normal tables with 95% confidence interval for g_1 as $\hat{g}_1 \pm 1.96\sqrt{6/n}$
 - For small samples, special tables from Snedecor and Cochran 1989, Table A 19(i) or Monte Carlo test

Kurtosis > 0 (leptokurtic)	heavier tail	compared to a normal distribution with the same σ (e.g., t-distribution)
Kurtosis < 0 (platykurtic)	lighter tail	compared to a normal distribution with the same σ

- For a normal distribution, $g_2^* = 3$. Kurtosis is often redefined as: $g_2 = \frac{E(Y-\mu)^4}{\sigma^4} - 3$ where the 4th central moment is estimated by $m_4 = \sum_{i=1}^n (y_i - \bar{y})^4/n$
 - the asymptotic sampling distribution for \hat{g}_2 is approximately $N(0,24/n)$ (with $n > 1000$)
 - large sample on kurtosis uses standard normal tables
 - small sample uses tables by Snedecor and Cochran, 1989, Table A 19(ii) or Geary 1936

```
data = rnorm(100)
library(e1071)
skewness(data)
#> [1] -0.3133111

kurtosis(data)
#> [1] -0.3036789
```

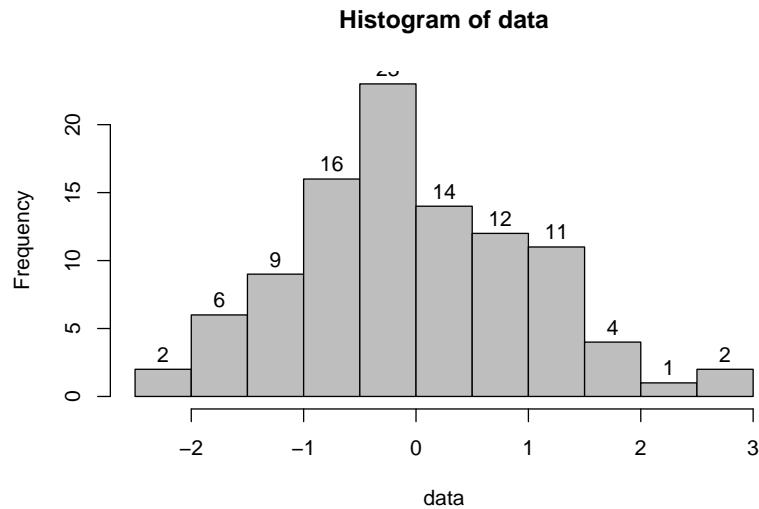
3.2 Graphical Measures

3.2.1 Shape

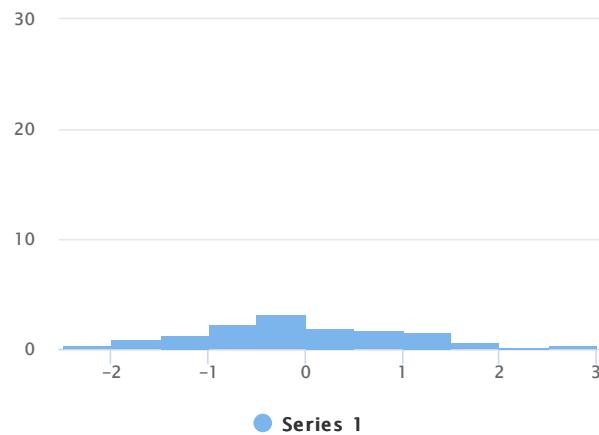
It's a good habit to label your graph, so others can easily follow.

```
data = rnorm(100)

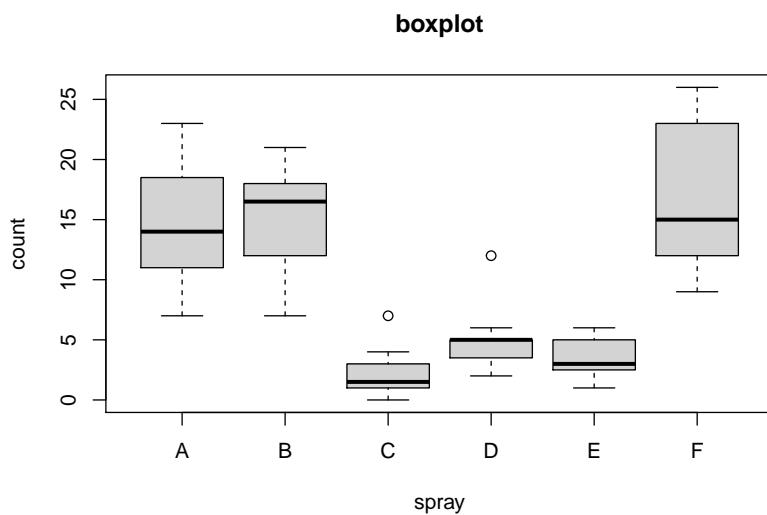
# Histogram
hist(data, labels = T, col="grey", breaks = 12)
```



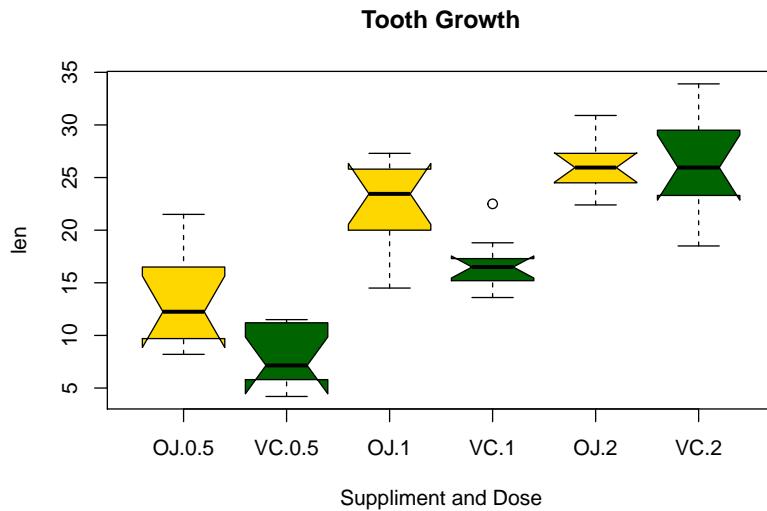
```
# Interactive histogram
pacman::p_load("highcharter")
hchart(data)
```



```
# Box-and-Whisker plot  
boxplot(count ~ spray, data = InsectSprays, col = "lightgray", main="boxplot")
```



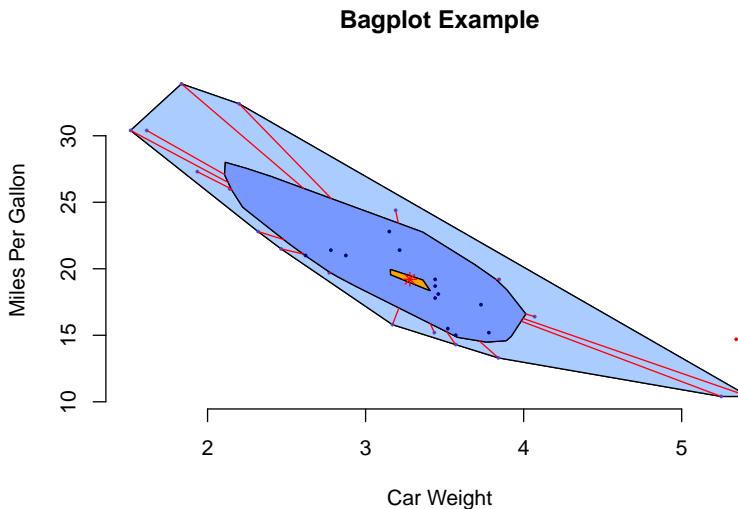
```
# Notched Boxplot  
boxplot(len~supp*dose, data=ToothGrowth, notch=TRUE,  
       col=(c("gold","darkgreen")),  
       main="Tooth Growth", xlab="Suppliment and Dose")
```



```
# If notches differ -> medians differ

# Stem-and-Leaf Plots
stem(data)
#>
#>   The decimal point is at the /
#>
#>   -2 | 3100
#>   -1 | 88765442211000
#>   -0 | 9998887776665555443333332222111111100
#>    0 | 011122223344556688899999
#>    1 | 0222222233356778
#>    2 | 368

# Bagplot - A 2D Boxplot Extension
pacman::p_load(aplpack)
attach(mtcars)
bagplot(wt,mpg, xlab="Car Weight", ylab="Miles Per Gallon",
        main="Bagplot Example")
```



Others more advanced plots

```
# boxplot.matrix()  #library("sfsmisc")
# boxplot.n()       #library("gplots")
# vioplot()         #library("vioplot")
```

3.2.2 Scatterplot

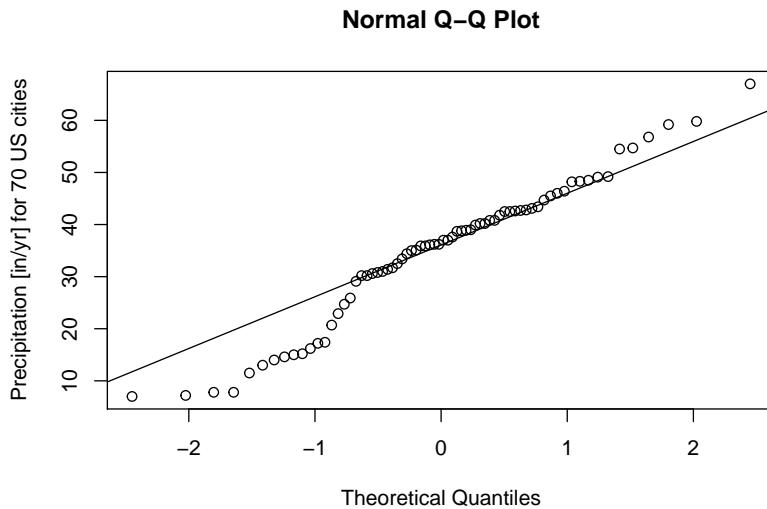
```
# pairs(mtcars)
```

3.3 Normality Assessment

Since Normal (Gaussian) distribution has many applications, we typically want/wish our data or our variable is normal. Hence, we have to assess the normality based on not only Numerical Measures but also Graphical Measures

3.3.1 Graphical Assessment

```
pacman::p_load("car")
qqnorm(precip, ylab = "Precipitation [in/yr] for 70 US cities")
qqline(precip)
```



The straight line represents the theoretical line for normally distributed data. The dots represent real empirical data that we are checking. If all the dots fall on the straight line, we can be confident that our data follow a normal distribution. If our data wiggle and deviate from the line, we should be concerned with the normality assumption.

3.3.2 Summary Statistics

Sometimes it's hard to tell whether your data follow the normal distribution by just looking at the graph. Hence, we often have to conduct statistical test to aid our decision. Common tests are

- Methods based on normal probability plot
 - Correlation Coefficient with Normal Probability Plots
 - Shapiro-Wilk Test
- Methods based on empirical cumulative distribution function
 - Anderson-Darling Test
 - Kolmogorov-Smirnov Test
 - Cramer-von Mises Test
 - Jarque-Bera Test

3.3.2.1 Methods based on normal probability plot

3.3.2.1.1 Correlation Coefficient with Normal Probability Plots
 (Looney and Gullledge, 1985) (Shapiro and Francia, 1972) The correlation coefficient between $y_{(i)}$ and m_i^* as given on the normal probability plot:

$$W^* = \frac{\sum_{i=1}^n (y_{(i)} - \bar{y})(m_i^* - 0)}{(\sum_{i=1}^n (y_{(i)} - \bar{y})^2 \sum_{i=1}^n (m_i^* - 0)^2) \cdot 5}$$

where $\bar{m}^* = 0$

Pearson product moment formula for correlation:

$$\hat{p} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{(\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (x_i - \bar{x})^2) \cdot 5}$$

- When the correlation is 1, the plot is exactly linear and normality is assumed.
- The closer the correlation is to zero, the more confident we are to reject normality
- Inference on W^* needs to be based on special tables (Looney and Gullledge, 1985)

```
library("EnvStats")
gofTest(data,test="ppcc")$p.value #Probability Plot Correlation Coefficient
#> [1] 0.6831237
```

3.3.2.1.2 Shapiro-Wilk Test (Shapiro and Wilk, 1965)

$$W = \left(\frac{\sum_{i=1}^n a_i (y_{(i)} - \bar{y})(m_i^* - 0)}{(\sum_{i=1}^n a_i^2 (y_{(i)} - \bar{y})^2 \sum_{i=1}^n (m_i^* - 0)^2) \cdot 5} \right)^2$$

where a_1, \dots, a_n are weights computed from the covariance matrix for the order statistics.

- Researchers typically use this test to assess normality. ($n < 2000$) Under normality, W is close to 1, just like W^* . Notice that the only difference between W and W^* is the “weights”.

```
gofTest(data,test="sw")$p.value #Shapiro-Wilk is the default.
#> [1] 0.6023442
```

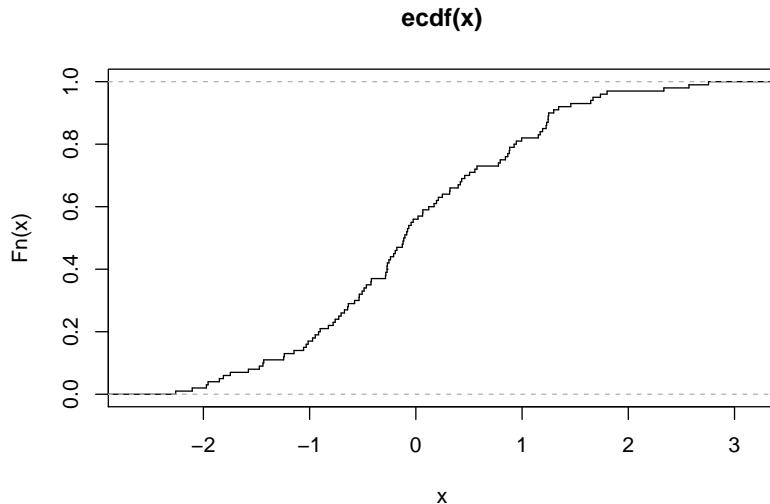
3.3.2.2 Methods based on empirical cumulative distribution function

The formula for the empirical cumulative distribution function (CDF) is:

$F_n(t)$ = estimate of probability that an observation $\leq t$ = (number of observations $\leq t$)/n

This method requires large sample sizes. However, it can apply to distributions other than the normal (Gaussian) one.

```
# Empirical CDF hand-code
plot.ecdf(data, verticals = T, do.points=F)
```



3.3.2.2.1 Anderson-Darling Test (Anderson and Darling, 1952)

The Anderson-Darling statistic:

$$A^2 = \int_{-\infty}^{\infty} (F_n(t) - F(t))^2 \frac{dF(t)}{F(t)(1-F(t))}$$

- a weight average of squared deviations (it weights small and large values of t more)

For the normal distribution,

$$A^2 = -(\sum_{i=1}^n (2i-1)(\ln(p_i) + \ln(1-p_{n+1-i}))) / n - n$$

where $p_i = \Phi\left(\frac{y_{(i)} - \bar{y}}{s}\right)$, the probability that a standard normal variable is less than $\frac{y_{(i)} - \bar{y}}{s}$

- Reject normal assumption when A^2 is too large
- Evaluate the null hypothesis that the observations are randomly selected from a normal population based on the critical value provided by (Marsaglia and Marsaglia, 2004) and (Stephens, 1974)
- This test can be applied to other distributions:
 - Exponential
 - Logistic
 - Gumbel
 - Extreme-value
 - Weibull: $\log(\text{Weibull}) = \text{Gumbel}$
 - Gamma
 - Logistic
 - Cauchy
 - von Mises
 - Log-normal (two-parameter)

Consult (Stephens, 1974) for more detailed transformation and critical values.

```
gofTest(data,test="ad")$p.value #Anderson-Darling
#> [1] 0.5751952
```

3.3.2.2.2 Kolmogorov-Smirnov Test

- Based on the largest absolute difference between empirical and expected cumulative distribution
- Another deviation of K-S test is Kuiper's test

```
gofTest(data,test="ks")$p.value #Kolmogorov-Smirnov
#> [1] 0.8196354
```

3.3.2.2.3 Cramer-von Mises Test

- Based on the average squared discrepancy between the empirical distribution and a given theoretical distribution. Each discrepancy is weighted equally (unlike Anderson-Darling test weights end points more heavily)

```
gofTest(data,test="cvm")$p.value #Cramer-von Mises
#> [1] 0.4706012
```

3.3.2.2.4 Jarque–Bera Test (Bera and Jarque, 1981)

Based on the skewness and kurtosis to test normality.

$JB = \frac{n}{6}(S^2 + (K - 3)^2/4)$ where S is the sample skewness and K is the sample kurtosis

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\sum_{i=1}^n (x_i - \bar{x})^3 / n}{(\sum_{i=1}^n (x_i - \bar{x})^2 / n)^{3/2}}$$

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\sum_{i=1}^n (x_i - \bar{x})^4 / n}{(\sum_{i=1}^n (x_i - \bar{x})^2 / n)^2}$$

recall $\hat{\sigma}^2$ is the estimate of the second central moment (variance) $\hat{\mu}_3$ and $\hat{\mu}_4$ are the estimates of third and fourth central moments.

If the data comes from a normal distribution, the JB statistic asymptotically has a chi-squared distribution with two degrees of freedom.

The null hypothesis is a joint hypothesis of the skewness being zero and the excess kurtosis being zero.

3.4 Bivariate Statistics

Correlation between

- Two Continuous variables
- Two Discrete variables
- Categorical and Continuous

	Categorical	Continuous
Categorical	Phi coefficient Cramer's V Tschuprow's T Freeman's Theta Epsilon-squared $GoodmanKruskal's\Lambda$ Somers' D Kendall's Tau-b Yule's Q and Y Tetrachoric Correlation Polychoric Correlation	
Continuous	Point-Biserial Correlation Logistic Regression	Pearson Correlation Spearman Correlation

Questions to keep in mind:

1. Is the relationship linear or non-linear?
2. If the variable is continuous, is it normal and homoskedastic?
3. How big is your dataset?

3.5 Two Continuous

```
n = 100 # (sample size)

data = data.frame(A = sample(1:20, replace = TRUE, size = n),
                  B = sample(1:30, replace = TRUE, size = n))
```

3.5.1 Pearson Correlation

- Good with linear relationship

```
library(Hmisc)
rcorr(data$A, data$B, type="pearson")
#>      x     y
#> x  1.00 -0.05
#> y -0.05  1.00
#>
#> n= 100
#>
```

```
#>
#> P
#>   x      y
#>   x      0.6237
#>   y  0.6237
```

3.5.2 Spearman Correlation

```
library(Hmisc)
rcorr(data$A, data$B, type="spearman")
#>   x      y
#> x  1.00 -0.05
#> y -0.05  1.00
#>
#> n= 100
#>
#>
#> P
#>   x      y
#>   x      0.6203
#>   y  0.6203
```

3.6 Categorical and Continuous

3.6.1 Point-Biserial Correlation

Similar to the Pearson correlation coefficient, the point-biserial correlation coefficient is between -1 and 1 where:

- -1 means a perfectly negative correlation between two variables
- 0 means no correlation between two variables
- 1 means a perfectly positive correlation between two variables

```
x <- c(0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0)
y <- c(12, 14, 17, 17, 11, 22, 23, 11, 19, 8, 12)

#calculate point-biserial correlation
cor.test(x, y)
#>
```

```
#> Pearson's product-moment correlation
#>
#> data: x and y
#> t = 0.67064, df = 9, p-value = 0.5193
#> alternative hypothesis: true correlation is not equal to 0
#> 95 percent confidence interval:
#> -0.4391885 0.7233704
#> sample estimates:
#>       cor
#> 0.2181635
```

Alternatively

```
ltm::biserial.cor(y,x, use = c("all.obs"), level = 2)
#> [1] 0.2181635
```

3.6.2 Logistic Regression

See 3.6.2

3.7 Two Discrete

3.7.1 Distance Metrics

Some consider distance is not a correlation metric because it isn't unit independent (i.e., if you scale the distance, the metrics will change), but it's still a useful proxy. Distance metrics are more likely to be used for similarity measure.

- Euclidean Distance
- Manhattan Distance
- Chessboard Distance
- Minkowski Distance
- Canberra Distance
- Hamming Distance
- Cosine Distance
- Sum of Absolute Distance
- Sum of Squared Distance
- Mean-Absolute Error

3.7.2 Statistical Metrics

3.7.2.1 Chi-squared test

3.7.2.1.1 Phi coefficient

- 2 binary

```
dt = matrix(c(1,4,3,5), nrow = 2)
dt
#>      [,1] [,2]
#> [1,]    1    3
#> [2,]    4    5
psych::phi(dt)
#> [1] -0.18
```

3.7.2.1.2 Cramer's V

- between nominal categorical variables (no natural order)

$$\text{Cramer's V} = \sqrt{\frac{\chi^2/n}{\min(c-1, r-1)}}$$

where

- χ^2 = Chi-square statistic
- n = sample size
- r = # of rows
- c = # of columns

```
library('lsr')
n = 100 # (sample size)
set.seed(1)
data = data.frame(A = sample(1:5, replace = TRUE, size = n),
                  B = sample(1:6, replace = TRUE, size = n))

cramersV(data$A, data$B)
#> [1] 0.1944616
```

Alternatively,

- `ncchisq` noncentral Chi-square
- `nchisqadj` Adjusted noncentral Chi-square
- `fisher` Fisher Z transformation
- `fisheradj` bias correction Fisher z transformation

```
DescTools::CramerV(data, conf.level = 0.95, method = "ncchisqadj")
#> Cramer V      lwr.ci      upr.ci
#> 0.3472325 0.3929964 0.4033053
```

3.7.2.1.3 Tschuprow's T

- 2 nominal variables

```
DescTools::TschuprowT(data)
#> [1] 0.1100808
```

3.7.3 Ordinal Association (Rank correlation)

- Good with non-linear relationship

3.7.3.1 Ordinal and Nominal

```
n = 100 # (sample size)
set.seed(1)
dt = table(data.frame(
  A = sample(1:4, replace = TRUE, size = n), # ordinal
  B = sample(1:3, replace = TRUE, size = n) # nominal
))
dt
#>      B
#> A   1 2 3
#>   1 7 11 9
#>   2 11 6 14
#>   3 7 11 4
#>   4 6 4 10
```

3.7.3.1.1 Freeman's Theta

- Ordinal and nominal

```
# this package is not available for R >= 4.0.0
rcompanion::freemanTheta(dt,group = "column" ) # because column is the grouping variab
```

3.7.3.1.2 Epsilon-squared

- Ordinal and nominal

```
# this package is not available for R >= 4.0.0
rcompanion::epsilonSquared(dt,group = "column" ) # because column is the grouping vari
```

3.7.3.2 Two Ordinal

```
n = 100 # (sample size)
set.seed(1)
dt = table(data.frame(
  A = sample(1:4, replace = TRUE, size = n), # ordinal
  B = sample(1:3, replace = TRUE, size = n) # ordinal
))
dt
#>      B
#> A   1 2 3
#>   1 7 11 9
#>   2 11 6 14
#>   3 7 11 4
#>   4 6 4 10
```

3.7.3.2.1 Goodman Kruskal's Gamma

- 2 ordinal variables

```
DescTools::GoodmanKruskalGamma(dt, conf.level = 0.95)
#>      gamma      lwr.ci      upr.ci
#>  0.006781013 -0.229032069  0.242594095
```

3.7.3.2.2 Somers' D

- or Somers' Delta
- 2 ordinal variables

```
DescTools::SomersDelta(dt, conf.level = 0.95)
#>      somers      lwr.ci      upr.ci
#>  0.005115859 -0.172800185  0.183031903
```

3.7.3.2.3 Kendall's Tau-b

- 2 ordinal variables

```
DescTools::KendallTauB(dt, conf.level = 0.95)
#>      tau_b      lwr.ci      upr.ci
#>  0.004839732 -0.163472443  0.173151906
```

3.7.3.2.4 Yule's Q and Y

- 2 ordinal variables

Special version (2 x 2) of the Goodman Kruskal's Gamma coefficient.

		Variable 1	
		a	b
Variable 2		c	d

$$\text{Yule's Q} = \frac{ad - bc}{ad + bc}$$

We typically use Yule's Q in practice while Yule's Y has the following relationship with Q.

$$\text{Yule's Y} = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

$$Q = \frac{2Y}{1 + Y^2}$$

$$Y = \frac{1 - \sqrt{1 - Q^2}}{Q}$$

```

n = 100 # (sample size)
set.seed(1)
dt = table(data.frame(A = sample(c(0, 1), replace = TRUE, size = n),
                      B = sample(c(0, 1), replace = TRUE, size = n)))
dt
#>      B
#> A    0  1
#>   0 25 24
#>   1 28 23

DescTools:::YuleQ(dt)
#> [1] -0.07778669

```

3.7.3.2.5 Tetrachoric Correlation

- is a special case of Polychoric Correlation when both variables are binary

```

library(psych)

n = 100 # (sample size)

data = data.frame(A = sample(c(0, 1), replace = TRUE, size = n),
                   B = sample(c(0, 1), replace = TRUE, size = n))

#view table
head(data)
#>   A B
#> 1 1 0
#> 2 1 0
#> 3 0 0
#> 4 1 0
#> 5 1 0
#> 6 1 0

table(data)
#>      B
#> A    0  1
#>   0 21 23
#>   1 34 22

#calculate tetrachoric correlation
tetrachoric(data)
#> Call: tetrachoric(x = data)

```

```
#> tetrachoric correlation
#>   A      B
#> A  1.0
#> B -0.2  1.0
#>
#> with tau of
#>   A      B
#> -0.15  0.13
```

3.7.3.2.6 Polychoric Correlation

- between ordinal categorical variables (natural order).
- Assumption: Ordinal variable is a discrete representation of a latent normally distributed continuous variable. (Income = low, normal, high).

```
library(polykor)

n = 100 # (sample size)

data = data.frame(A = sample(1:4, replace = TRUE, size = n),
                  B = sample(1:6, replace = TRUE, size = n))

head(data)
#>   A B
#> 1 1 3
#> 2 1 1
#> 3 3 5
#> 4 2 3
#> 5 3 5
#> 6 4 4

#calculate polychoric correlation between ratings
polychor(data$A, data$B)
#> [1] 0.01607982
```

3.7.4 Summary

```
library(tidyverse)

data("mtcars")
df = mtcars
```

```

df_factor = df %>%
  mutate(cyl = factor(cyl),
        vs = factor(vs),
        am = factor(am),
        gear = factor(gear),
        carb = factor(carb))
# summary(df)
str(df)
#> 'data.frame':   32 obs. of  11 variables:
#> $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
#> $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
#> $ disp: num  160 160 108 258 360 ...
#> $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
#> $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
#> $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
#> $ qsec: num  16.5 17 18.6 19.4 17 ...
#> $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
#> $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
#> $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
#> $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
str(df_factor)
#> 'data.frame':   32 obs. of  11 variables:
#> $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
#> $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
#> $ disp: num  160 160 108 258 360 ...
#> $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
#> $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
#> $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
#> $ qsec: num  16.5 17 18.6 19.4 17 ...
#> $ vs  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
#> $ am  : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
#> $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
#> $ carb: Factor w/ 6 levels "1","2","3","4",...: 4 4 1 1 2 1 4 2 2 4 ...

```

Get the correlation table for continuous variables only

```

cor(df)
#>          mpg         cyl         disp        hp       drat        wt
#> mpg  1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594
#> cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958
#> disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799
#> hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479
#> drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406
#> wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000

```

```
#> qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159
#> vs    0.6640389 -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157
#> am    0.5998324 -0.5226070 -0.5912270 -0.2432043  0.71271113 -0.6924953
#> gear   0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870
#> carb   -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059
#>          qsec           vs            am            gear            carb
#> mpg   0.41868403  0.6640389  0.59983243  0.4802848 -0.55092507
#> cyl   -0.59124207 -0.8108118 -0.52260705 -0.4926866  0.52698829
#> disp  -0.43369788 -0.7104159 -0.59122704 -0.5555692  0.39497686
#> hp    -0.70822339 -0.7230967 -0.24320426 -0.1257043  0.74981247
#> drat  0.09120476  0.4402785  0.71271113  0.6996101 -0.09078980
#> wt    -0.17471588 -0.5549157 -0.69249526 -0.5832870  0.42760594
#> qsec  1.00000000  0.7445354 -0.22986086 -0.2126822 -0.65624923
#> vs    0.74453544  1.0000000  0.16834512  0.2060233 -0.56960714
#> am    -0.22986086  0.1683451  1.00000000  0.7940588  0.05753435
#> gear  -0.21268223  0.2060233  0.79405876  1.0000000  0.27407284
#> carb  -0.65624923 -0.5696071  0.05753435  0.2740728  1.00000000

# only complete obs
# cor(df, use = "complete.obs")
```

Alternatively, you can also have the

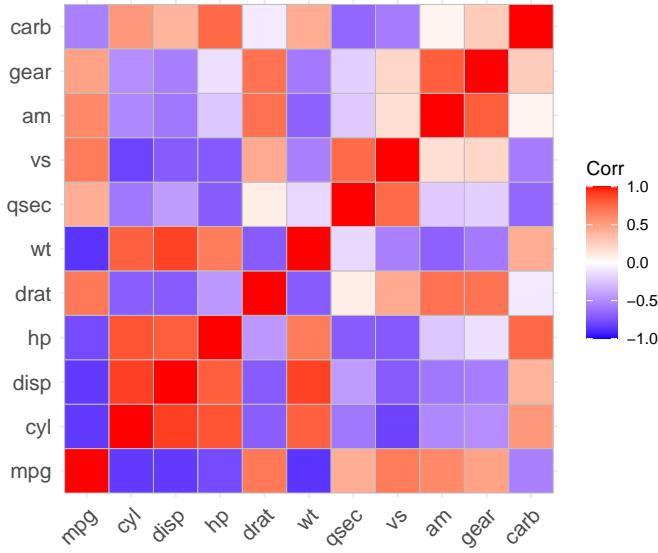
```
Hmisc:::rcorr(as.matrix(df), type = "pearson")
#>      mpg cyl disp hp drat wt qsec vs am gear carb
#> mpg  1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
#> cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
#> disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
#> hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
#> drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
#> wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
#> qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00  0.74 -0.23 -0.21 -0.66
#> vs    0.66 -0.81 -0.71 -0.72  0.44 -0.55  0.74  1.00  0.17  0.21 -0.57
#> am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
#> gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.21  0.79  1.00  0.27
#> carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66 -0.57  0.06  0.27  1.00
#>
#> n= 32
#>
#>
#> P
#>      mpg cyl disp hp drat wt qsec vs am gear
#> mpg      0.0000 0.0000 0.0000 0.0000 0.0171 0.0000 0.0003 0.0054
#> cyl     0.0000          0.0000 0.0000 0.0000 0.0004 0.0000 0.0022 0.0042
#> disp    0.0000 0.0000          0.0000 0.0000 0.0131 0.0000 0.0004 0.0010
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1
cyl	-0.85	1
disp	-0.85	0.90	1
hp	-0.78	0.83	0.79	1
drat	0.68	-0.70	-0.71	-0.45	1
wt	-0.87	0.78	0.89	0.66	-0.71	1
qsec	0.42	-0.59	-0.43	-0.71	0.09	-0.17	1
vs	0.66	-0.81	-0.71	-0.72	0.44	-0.55	0.74	1	.	.	.
am	0.60	-0.52	-0.59	-0.24	0.71	-0.69	-0.23	0.17	1	.	.
gear	0.48	-0.49	-0.56	-0.13	0.70	-0.58	-0.21	0.21	0.79	1	.
carb	-0.55	0.53	0.39	0.75	-0.09	0.43	-0.66	-0.57	0.06	0.27	1

```
#> hp 0.0000 0.0000 0.0000 0.0100 0.0000 0.0000 0.0000 0.1798 0.4930
#> drat 0.0000 0.0000 0.0000 0.0100 0.0000 0.6196 0.0117 0.0000 0.0000
#> wt 0.0000 0.0000 0.0000 0.0000 0.0000 0.3389 0.0010 0.0000 0.0005
#> qsec 0.0171 0.0004 0.0131 0.0000 0.6196 0.3389 0.0000 0.2057 0.2425
#> vs 0.0000 0.0000 0.0000 0.0000 0.0117 0.0010 0.0000 0.3570 0.2579
#> am 0.0003 0.0022 0.0004 0.1798 0.0000 0.0000 0.2057 0.3570 0.0000
#> gear 0.0054 0.0042 0.0010 0.4930 0.0000 0.0005 0.2425 0.2579 0.0000
#> carb 0.0011 0.0019 0.0253 0.0000 0.6212 0.0146 0.0000 0.0007 0.7545 0.1290
#> carb
#> mpg 0.0011
#> cyl 0.0019
#> disp 0.0253
#> hp 0.0000
#> drat 0.6212
#> wt 0.0146
#> qsec 0.0000
#> vs 0.0007
#> am 0.7545
#> gear 0.1290
#> carb
```

```
modelsummary::datasummary_correlation(df)
```

```
ggcorrplot::ggcorrplot(cor(df))
```



Different comparison between different correlation between different types of variables (i.e., continuous vs. categorical) can be problematic. Moreover, the problem of detecting non-linear vs. linear relationship/correlation is another one. Hence, a solution is that using mutual information from information theory (i.e., knowing one variable can reduce uncertainty about the other).

To implement mutual information, we have the following approximations

$$\downarrow \text{prediction error} \approx \downarrow \text{uncertainty} \approx \downarrow \text{association strength}$$

More specifically, following the X2Y metric, we have the following steps:

1. Predict y without x (i.e., baseline model)
 1. Average of y when y is continuous
 2. Most frequent value when y is categorical
2. Predict y with x (e.g., linear, random forest, etc.)
3. Calculate the prediction error difference between 1 and 2

To have a comprehensive table that could handle

- continuous vs. continuous
- categorical vs. continuous

- continuous vs. categorical
- categorical vs. categorical

the suggested model would be Classification and Regression Trees (CART). But we can certainly use other models as well.

The downfall of this method is that you might suffer

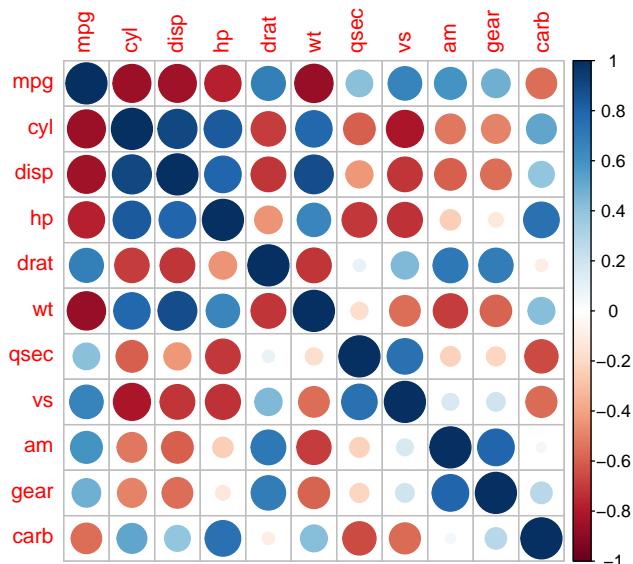
1. Symmetry: $(x, y) \neq (y, x)$
2. Comparability : Different pair of comparison might use different metrics (e.g., misclassification error vs. MAE)

```
library(ppsr)
# ppsr::score_df(iris) # if you want a dataframe
ppsr::score_matrix(iris, do_parallel = TRUE, n_cores = parallel::detectCores()/2) # i
#>           Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
#> Sepal.Length  1.00000000  0.04632352  0.5491398  0.4127668 0.4075487
#> Sepal.Width   0.06790301  1.00000000  0.2376991  0.2174659 0.2012876
#> Petal.Length  0.61608360  0.24263851  1.0000000  0.7917512 0.7904907
#> Petal.Width   0.48735314  0.20124105  0.7437845  1.0000000 0.7561113
#> Species       0.55918638  0.31344008  0.9167580  0.9398532 1.0000000
ppsr::score_matrix(df, do_parallel = TRUE, n_cores = parallel::detectCores()/2)
#>           mpg          cyl         disp        hp      drat       wt
#> mpg  1.0000000  0.32362397  0.25436628  0.210509478  0.20883649  0.24609235
#> cyl   0.3861810  1.00000000  0.57917897  0.537257954  0.33458867  0.38789293
#> disp  0.3141056  0.54883158  1.00000000  0.485122916  0.35317905  0.23669306
#> hp    0.2311418  0.37853515  0.35542647  1.000000000  0.24544714  0.12721154
#> drat  0.1646116  0.19540490  0.38730966  0.165537791  1.00000000  0.35076928
#> wt    0.2075760  0.11113261  0.20447239  0.155585827  0.12978458  1.00000000
#> qsec  0.1521642  0.10498746  0.07192679  0.134441221  0.08171630  0.05880165
#> vs    0.2000000  0.02514286  0.02514286  0.025142857  0.06862112  0.02514286
#> am    0.0615873  0.12825397  0.24409373  0.004444444  0.30608113  0.17742706
#> gear  0.1785968  0.43293014  0.43554416  0.154438566  0.54542788  0.31526071
#> carb  0.3472565  0.30798148  0.21228704  0.151523221  0.01103355  0.15957673
#>           qsec          vs          am          gear          carb
#> mpg  0.11030342  0.17957228  0.13297202  0.1752449  0.25426760
#> cyl   0.32753721  0.39827893  0.13263224  0.2877488  0.20925329
#> disp  0.31714642  0.35324790  0.23897094  0.4231630  0.15461337
#> hp    0.33941571  0.37794795  0.03821570  0.2159412  0.24105326
#> drat  0.16134068  0.17783324  0.30379298  0.4475122  0.03137800
#> wt    0.09367580  0.12214824  0.24118900  0.1590473  0.14181111
#> qsec  1.00000000  0.24973489  0.02334953  0.0000000  0.07539415
#> vs    0.40000000  1.00000000  0.10000000  0.1251429  0.20000000
#> am    0.15936508  0.11250000  1.00000000  0.3972789  0.00000000
```

```
#> gear 0.04791667 0.08012053 0.30341155 1.0000000 0.01486068
#> carb 0.21944241 0.25373093 0.00000000 0.00000000 1.00000000
```

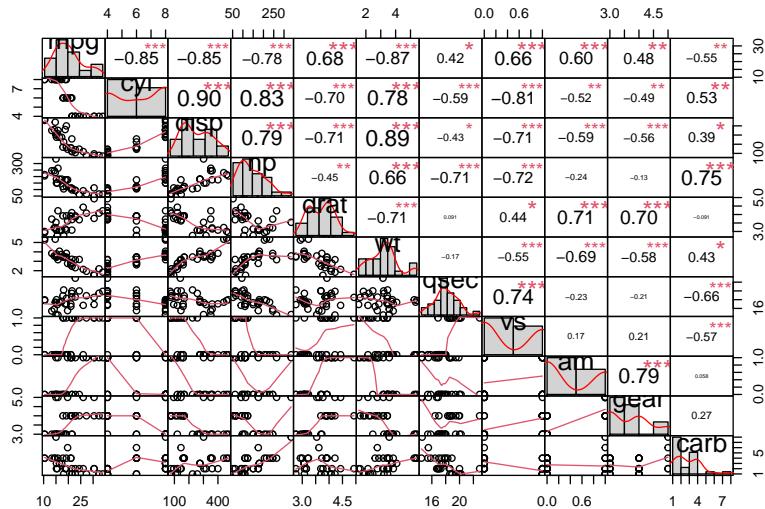
3.7.5 Visualization

```
corrplot::corrplot(cor(df))
```

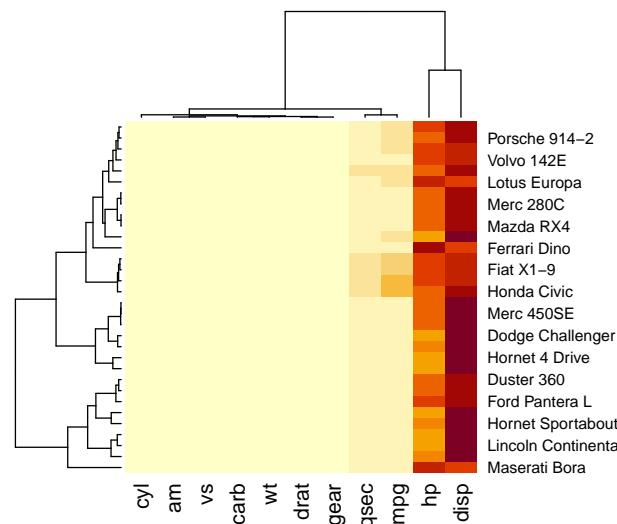


Alternatively,

```
PerformanceAnalytics::chart.Correlation(df, histogram = T, pch = 19)
```

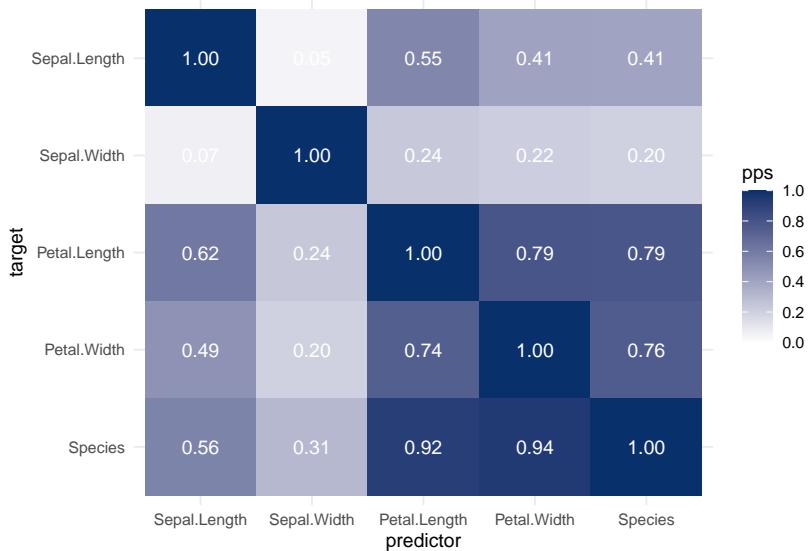


```
heatmap(as.matrix(df))
```

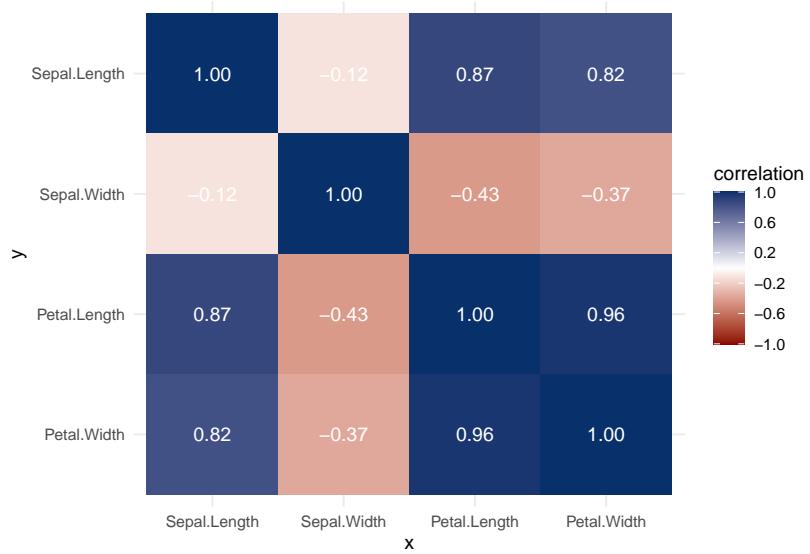


More general form,

```
ppsr::visualize_pps(df = iris, do_parallel = TRUE, n_cores = parallel::detectCores() / 2)
```

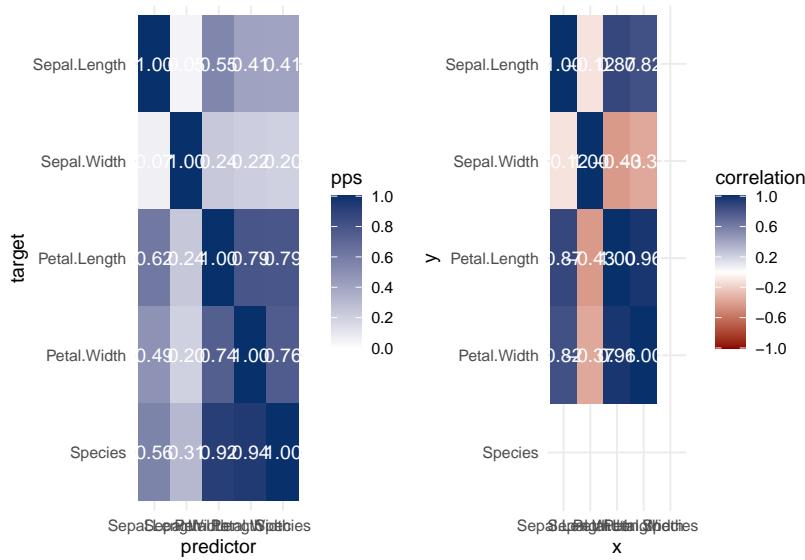


```
ppsr::visualize_correlations(
  df = iris
)
```



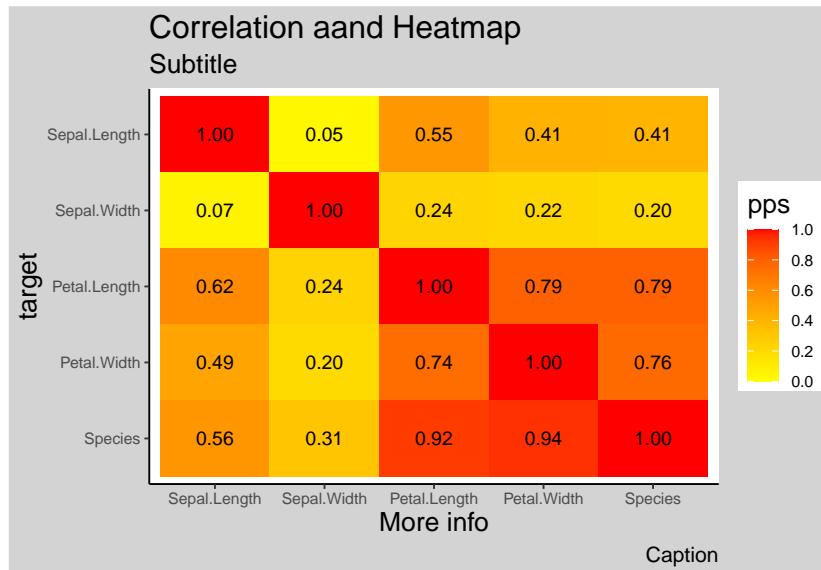
Both heatmap and correlation at the same time

```
ppsr::visualize_both(
  df = iris,
  do_parallel = TRUE,
  n_cores = parallel::detectCores() / 2
)
```



More elaboration with ggplot2

```
ppsr::visualize_pps(df = iris,
                     color_value_high = 'red',
                     color_value_low = 'yellow',
                     color_text = 'black') +
  ggplot2::theme_classic() +
  ggplot2::theme(plot.background = ggplot2::element_rect(fill = "lightgrey")) +
  ggplot2::theme(title = ggplot2::element_text(size = 15)) +
  ggplot2::labs(title = 'Correlation aand Heatmap',
                subtitle = 'Subtitle',
                caption = 'Caption',
                x = 'More info')
```



Chapter 4

Basic Statistical Inference

- One Sample Inference
- Two Sample Inference
- Categorical Data Analysis
- Divergence Metrics and Test for Comparing Distributions
- Make **inferences** (an interpretation) about the true parameter value β based on our estimator/estimate
- Test whether our underlying assumptions (about the true population parameters, random variables, or model specification) hold true.

Testing does not

- Confirm with 100% a hypothesis is true
- Confirm with 100% a hypothesis is false
- Tell you how to interpret the estimate value (Economic vs. Practical vs. Statistical Significance)

Hypothesis: Translate an objective in better understanding the results in terms of specifying a value (or sets of values) in which our population parameters should/should not lie.

- **Null hypothesis (H_0)**: A statement about the population parameter that we take to be true in which we would need the data to provide substantial evidence that against it.
 - Can be either a single value (ex: $H_0 : \beta = 0$) or a set of values (ex: $H_0 : \beta_1 \geq 0$)

- Will generally be the value you would not like the population parameter to be (subjective)
 - * $H_0 : \beta_1 = 0$ means you would like to see a non-zero coefficient
 - * $H_0 : \beta_1 \geq 0$ means you would like to see a negative effect
- “Test of Significance” refers to the two-sided test: $H_0 : \beta_j = 0$
- **Alternative hypothesis** (H_a or H_1) (Research Hypothesis): All other possible values that the population parameter may be if the null hypothesis does not hold.

Type I Error

Error made when H_0 is rejected when, in fact, H_0 is true.

The probability of committing a Type I error is α (known as **level of significance** of the test)

Type I error (α): probability of rejecting H_0 when it is true.

Legal analogy: In U.S. law, a defendant is presumed to be “innocent until proven guilty”.

If the null hypothesis is that a person is innocent, the Type I error is the probability that you conclude the person is guilty when he is innocent.

Type II Error

Type II error level (β): probability that you fail to reject the null hypothesis when it is false.

In the legal analogy, this is the probability that you fail to find the person guilty when he or she is guilty.

Error made when H_0 is not rejected when, in fact, H_1 is true

The probability of committing a Type II error is β (known as the **power** of the test)

Random sample of size n: A collection of n independent random variables taken from the distribution X, each with the same distribution as X.

Sample mean

$$\bar{X} = (\sum_{i=1}^n X_i)/n$$

Sample Median

\tilde{x} = the middle observation in a sample of observation order from smallest to largest (or vice versa).

If n is odd, \tilde{x} is the middle observation,

If n is even, \tilde{x} is the average of the two middle observations.

Sample variance

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{n \sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2}{n(n-1)}$$

Sample standard deviation

$$S = \sqrt{S^2}$$

Sample proportions

$$\hat{p} = \frac{X}{n} = \frac{\text{number in the sample with trait}}{\text{sample size}}$$

$$\widehat{p_1 - p_2} = \hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2} = \frac{n_2 X_1 - n_1 X_2}{n_1 n_2}$$

Estimators**Point Estimator**

$\hat{\theta}$ is a statistic used to approximate a population parameter θ

Point estimate

The numerical value assumed by $\hat{\theta}$ when evaluated for a given sample

Unbiased estimator

If $E(\hat{\theta}) = \theta$, then $\hat{\theta}$ is an unbiased estimator for θ

1. \bar{X} is an unbiased estimator for μ
2. S^2 is an unbiased estimator for σ^2
3. \hat{p} is an unbiased estimator for p
4. $\hat{p}_1 - \hat{p}_2$ is an unbiased estimator for $p_1 - p_2$
5. $\hat{X}_1 - \hat{X}_2$ is an unbiased estimator for $\mu_1 - \mu_2$

Note: S is a biased estimator for σ

Distribution of the sample mean

If \bar{X} is the sample mean based on a random sample of size n drawn from a normal distribution X with mean μ and standard deviation σ , the \bar{X} is normally distributed, with mean $\mu_{\bar{X}} = \mu$ and variance $\sigma_{\bar{X}}^2 = \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$. Then the **standard error of the mean** is: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

4.1 One Sample Inference

$$Y_i \sim i.i.d. N(\mu, \sigma^2)$$

i.i.d. standards for “independent and identically distributed”

Hence, we have the following model:

$$Y_i = \mu + \epsilon_i \text{ where}$$

- $\epsilon_i \sim^{iid} N(0, \sigma^2)$
- $E(Y_i) = \mu$
- $Var(Y_i) = \sigma^2$
- $\bar{y} \sim N(\mu, \sigma^2/n)$

4.1.1 The Mean

When σ^2 is estimated by s^2 , then

$$\frac{\bar{y} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

Then, a $100(1 - \alpha)\%$ confidence interval for μ is obtained from:

$$1 - \alpha = P\left(-t_{\alpha/2; n-1} \leq \frac{\bar{y} - \mu}{s/\sqrt{n}} \leq t_{\alpha/2; n-1}\right) = P\left(\bar{y} - (t_{\alpha/2; n-1})s/\sqrt{n} \leq \mu \leq \bar{y} + (t_{\alpha/2; n-1})s/\sqrt{n}\right)$$

And the interval is

$$\bar{y} \pm (t_{\alpha/2; n-1})s/\sqrt{n}$$

and s/\sqrt{n} is the standard error of \bar{y}

If the experiment were repeated many times, $100(1 - \alpha)\%$ of these intervals would contain μ

	Confidence Interval 100(1 - α)	Sample Sizes Confidence α , Error d	Hypothesis Testing Test Statistic
When σ^2 is known, X is normal (or $n \geq 25$)	$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$	$n \approx \frac{z_{\alpha/2}^2 \sigma^2}{d^2}$	$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

	Confidence Interval 100(1 - α)	Sample Sizes Confidence α , Error d	Hypothesis Testing Test Statistic
When σ^2 is unknown, X is normal (or $n \geq 25$)	$\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$	$n \approx \frac{z_{\alpha/2}^2 s^2}{d^2}$	$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$

4.1.1.1 For Difference of Means ($\mu_1 - \mu_2$), Independent Samples

	100(1 - α) Confidence Interval	Hypothesis Testing Test Statistic	
When σ^2 is known	$\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	
When σ^2 is unknown, Variances Assumed EQUAL	$\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$	$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$	Pooled Variance: $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ Degrees of Freedom: $\gamma = n_1 + n_2 - 2$
When σ^2 is unknown, Variances Assumed UNEQUAL	$\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2} \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)}$	$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)}}$	Degrees of Freedom: $\gamma = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\left(\frac{s_1^2}{n_1} \right)^2 + \left(\frac{s_2^2}{n_2} \right)^2}$

4.1.1.2 For Difference of Means ($\mu_1 - \mu_2$), Paired Samples ($D = X - Y$)

100(1 - α) Confidence Interval

$$\bar{D} \pm t_{\alpha/2} \frac{s_d}{\sqrt{n}}$$

Hypothesis Testing Test Statistic

$$t = \frac{\bar{D} - D_0}{s_d / \sqrt{n}}$$

4.1.1.3 Difference of Two Proportions

Mean

$$\hat{p}_1 - \hat{p}_2$$

Variance

$$\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

100(1 - α) Confidence Interval

$$\hat{p}_1 - \hat{p}_2 + z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Sample Sizes, Confidence α , Error d

(Prior Estimate fo \hat{p}_1, \hat{p}_2)

$$n \approx \frac{z_{\alpha/2}^2 [p_1(1-p_1) + p_2(1-p_2)]}{d^2}$$

(No Prior Estimates for \hat{p})

$$n \approx \frac{z_{\alpha/2}^2}{2d^2}$$

Hypothesis Testing - Test Statistics

Null Value $(p_1 - p_2)_0 \neq 0$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)_0}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

Null Value $(p_1 - p_2)_0 = 0$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

where

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

4.1.2 Single Variance

$$1-\alpha = P(\chi^2_{1-\alpha/2; n-1}) \leq (n-1)s^2/\sigma^2 \leq \chi^2_{\alpha/2; n-1} = P\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right)$$

and a $100(1 - \alpha)\%$ confidence interval for σ^2 is:

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2; n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2; n-1}}\right)$$

Confidence limits for σ^2 are obtained by computing the positive square roots of these limits

Equivalently,

100(1 - α) Confidence Interval

$$L_1 = \frac{(n-1)s^2}{\chi^2_{\alpha/2}} L_1 = \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

Hypothesis Testing Test Statistic

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

4.1.3 Single Proportion (p)

	Sample Sizes		
Confidence Interval	Confidence α , Error d (prior estimate for \hat{p})	(No prior estimate for \hat{p})	Hypothesis Testing Test Statistic
$100(1 - \alpha)$			
$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$n \approx \frac{z_{\alpha/2}^2 \hat{p}(1-\hat{p})}{d^2}$	$n \approx \frac{z_{\alpha/2}^2}{4d^2}$	$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

4.1.4 Power

Formally, power (for the test of the mean) is given by:

$$\pi(\mu) = 1 - \beta = P(\text{test rejects } H_0 | \mu)$$

To evaluate the power, one needs to know the distribution of the test statistic if the null hypothesis is false.

For 1-sided z-test where $H_0 : \mu \leq \mu_0$
 $H_A : \mu > 0$

The power is:

$$\begin{aligned}\pi(\mu) &= P(\bar{y} > \mu_0 + z_\alpha \sigma / \sqrt{n} | \mu) \\ &= P\left(Z = \frac{\bar{y} - \mu}{\sigma / \sqrt{n}} > z_\alpha + \frac{\mu_0 - \mu}{\sigma / \sqrt{n}} | \mu\right) \\ &= 1 - \Phi\left(z_\alpha + \frac{(\mu_0 - \mu)\sqrt{n}}{\sigma}\right) \\ &= \Phi\left(-z_\alpha + \frac{(\mu - \mu_0)\sqrt{n}}{\sigma}\right)\end{aligned}$$

where $1 - \Phi(x) = \Phi(-x)$ since the normal pdf is symmetric

Power is correlated to the difference in $\mu - \mu_0$, sample size n, variance σ^2 , and the α -level of the test (through z_α)

Equivalently, power can be increased by making α large, σ^2 smaller, or n larger.

For 2-sided z-test is:

$$\pi(\mu) = \Phi\left(-z_{\alpha/2} + \frac{(\mu_0 - \mu)\sqrt{n}}{\sigma}\right) + \Phi\left(-z_{\alpha/2} + \frac{(\mu - \mu_0)\sqrt{n}}{\sigma}\right)$$

4.1.5 Sample Size

4.1.5.1 1-sided Z-test

Example: to show that the mean response μ under the treatment is higher than the mean response μ_0 without treatment (show that the treatment effect $\delta = \mu - \mu_0$ is large)

Because power is an increasing function of $\mu - \mu_0$, it is only necessary to find n that makes the power equal to $1 - \beta$ at $\mu = \mu_0 + \delta$

Hence, we have

$$\pi(\mu_0 + \delta) = \Phi\left(-z_\alpha + \frac{\delta\sqrt{n}}{\sigma}\right) = 1 - \beta$$

Since $\Phi(z_\beta) = 1 - \beta$, we have

$$-z_\alpha + \frac{\delta\sqrt{n}}{\sigma} = z_\beta$$

Then n is

$$n = \left(\frac{(z_\alpha + z_\beta)\sigma}{\delta} \right)^2$$

Then, we need larger samples, when

- the sample variability is large (σ is large)
- α is small (z_α is large)
- power $1 - \beta$ is large (z_β is large)
- The magnitude of the effect is smaller (δ is small)

Since we don't know δ and σ . We can base σ on previous studies, pilot studies. Or, obtain an estimate of σ by anticipating the range of the observation (without outliers). divide this range by 4 and use the resulting number as an approximate estimate of σ . For normal (distribution) data, this is reasonable.

4.1.5.2 2-sided Z-test

We want to know the min n, required to guarantee $1 - \beta$ power when the treatment effect $\delta = |\mu - \mu_0|$ is at least greater than 0. Since the power function for the 2-sided is increasing and symmetric in $|\mu - \mu_0|$, we only need to find n that makes the power equal to $1 - \beta$ when $\mu = \mu_0 + \delta$

$$n = \left(\frac{(z_{\alpha/2} + z_\beta)\sigma}{\delta} \right)^2$$

We could also use the confidence interval approach. If we require that an α -level two-sided CI for μ be

$$\bar{y} \pm D$$

where $D = z_{\alpha/2}\sigma/\sqrt{n}$ gives

$$n = \left(\frac{z_{\alpha/2}\sigma}{D} \right)^2$$

(round up to the nearest integer)

```
data = rnorm(100)
t.test(data, conf.level=0.95)
#>
#> One Sample t-test
#>
#> data: data
#> t = -0.55586, df = 99, p-value = 0.5796
```

```
#> alternative hypothesis: true mean is not equal to 0
#> 95 percent confidence interval:
#> -0.2124625 0.1194737
#> sample estimates:
#> mean of x
#> -0.04649437
```

$$H_0 : \mu \geq 30 \quad H_a : \mu < 30$$

```
t.test(data, mu=30, alternative="less")
#>
#> One Sample t-test
#>
#> data: data
#> t = -359.22, df = 99, p-value < 2.2e-16
#> alternative hypothesis: true mean is less than 30
#> 95 percent confidence interval:
#> -Inf 0.09238761
#> sample estimates:
#> mean of x
#> -0.04649437
```

4.1.6 Note

For t-tests, the sample and power are not as easy as z-test.

$$\pi(\mu) = P\left(\frac{\bar{y} - \mu_0}{s/\sqrt{n}} > t_{n-1;\alpha} | \mu\right)$$

when $\mu > \mu_0$ (i.e., $\mu - \mu_0 = \delta$), the random variable $(\bar{y} - \mu_0)/(s/\sqrt{n})$ does not have a Student's t distribution, but rather is distributed as a non-central t-distribution with non-centrality parameter $\delta\sqrt{n}/\sigma$ and d.f. of $n - 1$

- The power is an increasing function of this non-centrality parameter (note, when $\delta = 0$ the distribution is usual Student's t-distribution).
- To evaluate power, one must consider numerical procedure or use special charts

Approximate Sample Size Adjustment for t-test. We use an adjustment to the z-test determination for sample size.

Let $v = n - 1$, where n is sample size derived based on the z-test power. Then the 2-sided t-test sample size (approximate) is given:

$$n^* = \frac{(t_{v;\alpha/2} + t_{v;\beta})^2 \sigma^2}{\delta^2}$$

4.1.7 One-sample Non-parametric Methods

```
lecture.data=c(0.76, 0.82, 0.80, 0.79, 1.06, 0.83, -0.43, -0.34, 3.34, 2.33)
```

4.1.7.1 Sign Test

If we want to test $H_0 : \mu_{(0.5)} = 0; H_a : \mu_{(0.5)} > 0$ where $\mu_{(0.5)}$ is the population median. We can

- (1) Count the number of observation (y_i 's) that exceed 0. Denote this number by s_+ , called the number of plus signs. Let $s_- = n - s_+$, which is the number of minus signs.
- (2) Reject H_0 if s_+ is large or equivalently, if s_- is small.

To determine how large s_+ must be to reject H_0 at a given significance level, we need to know the distribution of the corresponding random variable S_+ under the null hypothesis, which is a binomial with $p = 1/2$, when the null is true.

To work out the null distribution using the binomial formula, we have α -level test rejects H_0 if $s_+ \geq b_{n,\alpha}$, where $b_{n,\alpha}$ is the upper α critical point of the $Bin(n, 1/2)$ distribution. Both S_+ and S_- have this same distribution ($S = S_+ + S_- = n$).

$$\text{p-value} = P(S \geq s_+) = \sum_{i=s_+}^n \binom{n}{i} \left(\frac{1}{2}\right)^n$$

equivalently,

$$P(S \leq s_-) = \sum_{i=0}^{s_-} \binom{n}{i} \left(\frac{1}{2}\right)^n$$

For large sample sizes, we could use the normal approximation for the binomial, in which case reject H_0 if

$$s_+ \geq n/2 + 1/2 + z_\alpha \sqrt{n/4}$$

For the 2-sided test, we use the tests statistic $s_{max} = \max(s_+, s_-)$ or $s_{min} = \min(s_+, s_-)$. An α -level test rejects H_0 if the p-value is $\leq \alpha$, where the p-value is computed from:

$$p-value = 2 \sum_{i=s_{max}}^n \binom{n}{i} \left(\frac{1}{2}\right)^n = s \sum_{i=0}^{s_{min}} \binom{n}{i} \left(\frac{1}{2}\right)^n$$

Equivalently, rejecting H_0 if $s_{max} \geq b_{n,\alpha/2}$

A large sample normal approximation can be used, where

$$z = \frac{s_{max} - n/2 - 1/2}{\sqrt{n/4}}$$

and reject H_0 at α if $z \geq z_{\alpha/2}$

However, treatment of 0 is problematic for this test.

- Solution 1: randomly assign 0 to the positive or negative (2 researchers might get different results).
- Solution 2: count each 0 as a contribution 1/2 toward s_+ and s_- (but then could not apply the binomial distribution)
- Solution 3: ignore 0 (reduces the power of test due to decreased sample size).

```
binom.test(sum(lecture.data > 0), length(lecture.data)) # alternative = "greater" or a
#>
#> Exact binomial test
#>
#> data: sum(lecture.data > 0) and length(lecture.data)
#> number of successes = 8, number of trials = 10, p-value = 0.1094
#> alternative hypothesis: true probability of success is not equal to 0.5
#> 95 percent confidence interval:
#> 0.4439045 0.9747893
#> sample estimates:
#> probability of success
#> 0.8
```

4.1.7.2 Wilcoxon Signed Rank Test

Since the Sign Test could not consider the magnitude of each observation from 0, the Wilcoxon Signed Rank Test improves by taking account the ordered magnitudes of the observation, but it will impose the requirement of symmetric to this test (while Sign Test does not)

$$H_0 : \mu_{0.5} = 0 H_a : \mu_{0.5} > 0$$

(assume no ties or same observations)

The signed rank test procedure:

1. rank order the observation y_i in terms of their absolute values. Let r_i be the rank of y_i in this ordering. Since we assume no ties, the ranks r_i are uniquely determined and are a permutation of the integers 1,2,...,n.
2. Calculate w_+ , which is the sum of the ranks of the positive values, and w_- , which is the sum of the ranks of the negative values. Note that $w_+ + w_- = r_1 + r_2 + \dots = 1 + 2 + \dots + n = n(n+1)/2$
3. Reject H_0 if w_+ is large (or if w_- is small)

To know what is large or small with regard to w_+ and w_- , we need the distribution of W_+ and W_- when the null is true.

Since these null distributions are identical and symmetric, the p-value is $P(W \geq w_+) = P(W \leq w_-)$

An α -level test rejects the null if the p-value is $\leq \alpha$, or if $w_+ \geq w_{n,\alpha}$, where $w_{n,\alpha}$ is the upper α critical point of the null distribution of W.

This distribution of W has a special table. For large n, the distribution of W is approximately normal.

$$z = \frac{w_+ - n(n+1)/4 - 1/2}{\sqrt{n(n+1)(2n+1)/24}}$$

The test rejects H_0 at level α if

$$w_+ \geq n(n+1)/4 + 1/2 + z_\alpha \sqrt{n(n+1)(2n+1)/24} \approx w_{n,\alpha}$$

For the 2-sided test, we use $w_{max} = \max(w_+, w_-)$ or $w_{min} = \min(w_+, w_-)$, with p-value given by:

$$p-value = 2P(W \geq w_{max}) = 2P(W \leq w_{min})$$

Same as Sign Test, we ignore 0. In some cases where some of the $|y_i|$'s may be tied for the same rank, we simply assign each of the tied ranks the average rank (or "midrank").

Example, if $y_1 = -1$, $y_3 = 3$ and $y_3 = -3$, and $y_4 = 5$, then $r_1 = 1$, $r_2 = r_3 = (2+3)/2 = 2.5$, $r_4 = 4$

```
wilcox.test(lecture.data) #does not use normal approximation (using the underlying W d
#>
#> Wilcoxon signed rank exact test
#>
#> data: lecture.data
#> V = 52, p-value = 0.009766
#> alternative hypothesis: true location is not equal to 0

wilcox.test(lecture.data,exact=F) #uses normal approximation
#>
#> Wilcoxon signed rank test with continuity correction
#>
#> data: lecture.data
#> V = 52, p-value = 0.01443
#> alternative hypothesis: true location is not equal to 0
```

4.2 Two Sample Inference

4.2.1 Means

Suppose we have 2 sets of observations,

- y_1, \dots, y_{n_y}
- x_1, \dots, x_{n_x}

that are random samples from two independent populations with means μ_y and μ_x and variances σ_y^2, σ_x^2 . Our goal is to compare μ_x and μ_y or $\sigma_y^2 = \sigma_x^2$

4.2.1.1 Large Sample Tests

Assume that n_y and n_x are large (≥ 30). Then,

$$E(\bar{y} - \bar{x}) = \mu_y - \mu_x \text{Var}(\bar{y} - \bar{x}) = \sigma_y^2/n_y + \sigma_x^2/n_x$$

Then,

$$Z = \frac{\bar{y} - \bar{x} - (\mu_y - \mu_x)}{\sqrt{\sigma_y^2/n_y + \sigma_x^2/n_x}} \sim N(0, 1)$$

(according to Central Limit Theorem). For large samples, we can replace variances by their unbiased estimators (s_y^2, s_x^2), and get the same large sample distribution.

An approximate $100(1 - \alpha)\%$ CI for $\mu_y - \mu_x$ is given by:

$$\bar{y} - \bar{x} \pm z_{\alpha/2} \sqrt{s_y^2/n_y + s_x^2/n_x}$$

$$H_0 : \mu_y - \mu_x = \delta_0 \quad H_A : \mu_y - \mu_x \neq \delta_0$$

at the α -level with the statistic:

$$z = \frac{\bar{y} - \bar{x} - \delta_0}{\sqrt{s_y^2/n_y + s_x^2/n_x}}$$

and reject H_0 if $|z| > z_{\alpha/2}$

If $\delta = 0$, it means that we are testing whether two means are equal.

4.2.1.2 Small Sample Tests

If the two samples are from normal distribution, iid $N(\mu_y, \sigma_y^2)$ and iid $N(\mu_x, \sigma_x^2)$ and the two samples are independent, we can do inference based on the t-distribution

Then we have 2 cases

- Equal Variance
- Unequal Variance

4.2.1.2.1 Equal variance Assumptions

- iid: so that $\text{var}(\bar{y}) = \sigma_y^2/n_y$; $\text{var}(\bar{x}) = \sigma_x^2/n_x$
- Independence between samples: No observation from one sample can influence any observation from the other sample, to have

$$\begin{aligned} \text{var}(\bar{y} - \bar{x}) &= \text{var}(\bar{y}) + \text{var}\bar{x} - 2\text{cov}(\bar{y}, \bar{x}) \\ &= \text{var}(\bar{y}) + \text{var}\bar{x} \\ &= \sigma_y^2/n_y + \sigma_x^2/n_x \end{aligned}$$

- Normality: Justifies the use of the t-distribution

Let $\sigma^2 = \sigma_y^2 = \sigma_x^2$. Then, s_y^2 and s_x^2 are both unbiased estimators of σ^2 . We then can pool them.

Then the pooled variance estimate is

$$s^2 = \frac{(n_y - 1)s_y^2 + (n_x - 1)s_x^2}{(n_y - 1) + (n_x - 1)}$$

has $n_y + n_x - 2$ df.

Then the test statistic

$$T = \frac{\bar{y} - \bar{x} - (\mu_y - \mu_x)}{s \sqrt{1/n_y + 1/n_x}} \sim t_{n_y+n_x-2}$$

100(1 - α)% CI for $\mu_y - \mu_x$ is

$$\bar{y} - \bar{x} \pm (t_{n_y+n_x-2})s \sqrt{1/n_y + 1/n_x}$$

Hypothesis testing:

$$H_0 : \mu_y - \mu_x = \delta_0 \quad H_1 : \mu_y - \mu_x \neq \delta_0$$

we reject H_0 if $|t| > t_{n_y+n_x-2; \alpha/2}$

4.2.1.2.2 Unequal Variance Assumptions

1. Two samples are independent
 1. Scatter plots
 2. Correlation coefficient (if normal)
2. Independence of observation in each sample
 1. Test for serial correlation
3. For each sample, homogeneity of variance
 1. Scatter plots
 2. Formal tests
4. Normality
5. Equality of variances (homogeneity of variance between samples)
 1. F-test
 2. Barlett test
 3. [Modified Levene Test]

To compare 2 normal $\sigma_y^2 \neq \sigma_x^2$, we use the test statistic:

$$T = \frac{\bar{y} - \bar{x} - (\mu_y - \mu_x)}{\sqrt{s_y^2/n_y + s_x^2/n_x}}$$

In this case, T does not follow the t-distribution (its distribution depends on the ratio of the unknown variances σ_y^2, σ_x^2). In the case of small sizes, we can approximate tests by using the Welch-Satterthwaite method (Satterthwaite, 1946). We assume T can be approximated by a t-distribution, and adjust the degrees of freedom.

Let $w_y = s_y^2/n_y$ and $w_x = s_x^2/n_x$ (the w's are the square of the respective standard errors)

Then, the degrees of freedom are

$$v = \frac{(w_y + w_x)^2}{w_y^2/(n_y - 1) + w_x^2/(n_x - 1)}$$

Since v is usually fractional, we truncate down to the nearest integer.

100(1 - α)% CI for $\mu_y - \mu_x$ is

$$\bar{y} - \bar{x} \pm t_{v,\alpha/2} \sqrt{s_y^2/n_y + s_x^2/n_x}$$

Reject H_0 if $|t| > t_{v,\alpha/2}$, where

$$t = \frac{\bar{y} - \bar{x} - \delta_0}{\sqrt{s_y^2/n_y + s_x^2/n_x}}$$

4.2.2 Variances

$$F_{ndf,ddf} = \frac{s_1^2}{s_2^2}$$

where $s_1^2 > s_2^2, ndf = n_1 - 1, ddf = n_2 - 1$

4.2.2.1 F-test

Test

$$H_0 : \sigma_y^2 = \sigma_x^2 H_a : \sigma_y^2 \neq \sigma_x^2$$

Consider the test statistic,

$$F = \frac{s_y^2}{s_x^2}$$

Reject H_0 if

- $F > f_{n_y-1, n_x-1, \alpha/2}$ or
- $F < f_{n_y-1, n_x-1, 1-\alpha/2}$

Where $F > f_{n_y-1, n_x-1, \alpha/2}$ and $F < f_{n_y-1, n_x-1, 1-\alpha/2}$ are the upper and lower $\alpha/2$ critical points of an F-distribution, with $n_y - 1$ and $n_x - 1$ degrees of freedom.

Note

- This test depends heavily on the assumption Normality.
- In particular, it could give many significant results when observations come from long-tailed distributions (i.e., positive kurtosis).
- If we cannot find support for normality, then we can use nonparametric tests such as the [Modified Levene Test]

```
data(iris)
irisVe=iris$Petal.Width[iris$Species=="versicolor"]
irisVi=iris$Petal.Width[iris$Species=="virginica"]

var.test(irisVe,irisVi)
#>
#> F test to compare two variances
#>
#> data: irisVe and irisVi
#> F = 0.51842, num df = 49, denom df = 49, p-value = 0.02335
#> alternative hypothesis: true ratio of variances is not equal to 1
#> 95 percent confidence interval:
#> 0.2941935 0.9135614
#> sample estimates:
#> ratio of variances
#> 0.5184243
```

4.2.2.2 Modified Levene Test (Brown-Forsythe Test)

- considers averages of absolute deviations rather than squared deviations.
Hence, less sensitive to long-tailed distributions.

- This test is still good for normal data

For each sample, we consider the absolute deviation of each observation from the median:

$$d_{y,i} = |y_i - y_{.5}| d_{x,i} = |x_i - x_{.5}|$$

Then,

$$t_L^* = \frac{\bar{d}_y - \bar{d}_x}{s \sqrt{\frac{1}{n_y} + \frac{1}{n_x}}}$$

The pooled variance s^2 is given by:

$$s^2 = \frac{\sum_i^{n_y} (d_{y,i} - \bar{d}_y)^2 + \sum_j^{n_x} (d_{x,i} - \bar{d}_x)^2}{n_y + n_x - 2}$$

- If the error terms have constant variance and n_y and n_x are not extremely small, then $t_L^* \sim t_{n_x+n_y-2}$
- We reject the null hypothesis when $|t_L^*| > t_{n_y+n_x-2;\alpha/2}$
- This is just the two-sample t-test applied to the absolute deviations.

```
dVe=abs(irisVe-median(irisVe))
dVi=abs(irisVi-median(irisVi))
t.test(dVe,dVi,var.equal=T)
#>
#> Two Sample t-test
#>
#> data: dVe and dVi
#> t = -2.5584, df = 98, p-value = 0.01205
#> alternative hypothesis: true difference in means is not equal to 0
#> 95 percent confidence interval:
#> -0.12784786 -0.01615214
#> sample estimates:
#> mean of x mean of y
#>      0.154      0.226

# small samples t-test
t.test(irisVe,irisVi,var.equal=F)
#>
#> Welch Two Sample t-test
```

```
#>
#> data: irisVe and irisVi
#> t = -14.625, df = 89.043, p-value < 2.2e-16
#> alternative hypothesis: true difference in means is not equal to 0
#> 95 percent confidence interval:
#> -0.7951002 -0.6048998
#> sample estimates:
#> mean of x mean of y
#> 1.326 2.026
```

4.2.3 Power

Consider $\sigma_y^2 = \sigma_x^2 = \sigma^2$

Under the assumption of equal variances, we take size samples from both groups ($n_y = n_x = n$)

For 1-sided testing,

$$H_0 : \mu_y - \mu_x \leq 0 \quad H_a : \mu_y - \mu_x > 0$$

α -level z-test rejects H_0 if

$$z = \frac{\bar{y} - \bar{x}}{\sigma \sqrt{2/n}} > z_\alpha$$

$$\pi(\mu_y - \mu_x) = \Phi(-z_\alpha + \frac{\mu_y - \mu_x}{\sigma} \sqrt{n/2})$$

We need sample size n that give at least $1 - \beta$ power when $\mu_y - \mu_x = \delta$, where δ is the smallest difference that we want to see.

Power is given by:

$$\Phi(-z_\alpha + \frac{\delta}{\sigma} \sqrt{n/2}) = 1 - \beta$$

4.2.4 Sample Size

Then, the sample size is

$$n = 2 \left(\frac{\sigma(z_\alpha + z_\beta)}{\delta} \right)^2$$

For 2-sided test, replace z_α with $z_{\alpha/2}$.

As with the one-sample case, to perform an exact 2-sample t-test sample size calculation, we must use a non-central t-distribution.

A correction that gives the approximate t-test sample size can be obtained by using the z-test n value in the formula:

$$n^* = 2 \left(\frac{\sigma(t_{2n-2;\alpha} + t_{2n-2;\beta})}{\delta} \right)^2$$

where we use $\alpha/2$ for the two-sided test

4.2.5 Matched Pair Designs

We have two treatments

Subject	Treatment A	Treatment B	Difference
1	y_1	x_1	$d_1 = y_1 - x_1$
2	y_2	x_2	$d_2 = y_2 - x_2$
.	.	.	.
n	y_n	x_n	$d_n = y_n - x_n$

we assume $y_i \sim^{iid} N(\mu_y, \sigma_y^2)$ and $x_i \sim^{iid} N(\mu_x, \sigma_x^2)$, but since y_i and x_i are measured on the same subject, they are correlated.

Let

$$\mu_D = E(y_i - x_i) = \mu_y - \mu_x \sigma_D^2 = var(y_i - x_i) = Var(y_i) + Var(x_i) - 2cov(y_i, x_i)$$

If the matching induces **positive** correlation, then the variance of the difference of the measurements is reduced as compared to the independent case. This is the point of Matched Pair Designs. Although covariance can be negative, giving a larger variance of the difference than the independent sample case, usually the covariance is positive. This means both y_i and x_i are large for many of the same subjects, and for others, both measurement are small. (we still assume that various subjects respond independently of each other, which is necessary for the iid assumption within groups).

Let $d_i = y_i - x_i$, then

- $\bar{d} = \bar{y} - \bar{x}$ is the sample mean of the d_i
- $s_d^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2$ is the sample variance of the difference

Once the data are converted to differences, we are back to One Sample Inference and can use its tests and CIs.

4.2.6 Nonparametric Tests for Two Samples

For Matched Pair Designs, we can use the One-sample Non-parametric Methods.

Assume that Y and X are random variables with CDF F_y and F_x . then, Y is **stochastically** larger than X for all real number u, $P(Y > u) \geq P(X > u)$.

Equivalently, $P(Y \leq u) \leq P(X \leq u)$, which is $F_Y(u) \leq F_X(u)$, same thing as $F_Y < F_X$

If two distributions are identical, except that one is shifted relative to the other, then each of distribution can be indexed by a location parameter, say θ_y and θ_x . In this case, $Y > X$ if $\theta_y > \theta_x$

Consider the hypotheses,

$$H_0 : F_Y = F_X \quad H_a : F_Y < F_X$$

where the alternative is an upper one-sided alternative.

- We can also consider the lower one-sided alternative

$$H_a : F_Y > F_X \text{ or } H_a : F_Y < F_X \text{ or } F_Y > F_X$$

- In this case, we don't use $H_a : F_Y \neq F_X$ as that allows arbitrary differences between the distributions, without requiring one be stochastically larger than the other.

If the distributions only differ in terms of their location parameters, we can focus hypothesis tests on the parameters (e.g., $H_0 : \theta_y = \theta_x$ vs. $\theta_y > \theta_x$)

We have 2 equivalent nonparametric tests that consider the hypothesis mentioned above

1. Wilcoxon rank test
2. Mann-Whitney U test

4.2.6.1 Wilcoxon rank test

1. Combine all $n = n_y + n_x$ observations and rank them in ascending order.
2. Sum the ranks of the y's and x's separately. Let w_y and w_x be these sums. ($w_y + w_x = 1 + 2 + \dots + n = n(n+1)/2$)

3. Reject H_0 if w_y is large (equivalently, w_x is small)

Under H_0 , any arrangement of the y's and x's is equally likely to occur, and there are $(n_y + n_x)!/(n_y!n_x!)$ possible arrangements.

- Technically, for each arrangement we can compute the values of w_y and w_x , and thus generate the distribution of the statistic under the null hypothesis.
- This could lead to computationally intensive.

```
wilcox.test(irisVe,irisVi,alternative="two.sided",conf.level=0.95, exact=F,correct=T)
#>
#> Wilcoxon rank sum test with continuity correction
#>
#> data: irisVe and irisVi
#> W = 49, p-value < 2.2e-16
#> alternative hypothesis: true location shift is not equal to 0
```

4.2.6.2 Mann-Whitney U test

The Mann-Whitney test is computed as follows:

1. Compare each y_i with each x_i .
Let u_y be the number of pairs in which $y_i > x_i$. Let u_x be the number of pairs in which $y_i < x_i$. (assume there are no ties). There are $n_y n_x$ such comparisons and $u_y + u_x = n_y n_x$.
2. Reject H_0 if u_y is large (or u_x is small)

Mann-Whitney U test and Wilcoxon rank test are related:

$$u_y = w_y - n_y(n_y + 1)/2 \quad u_x = w_x - n_x(n_x + 1)/2$$

An α -level test rejects H_0 if $u_y \geq u_{n_y, n_x, \alpha}$, where $u_{n_y, n_x, \alpha}$ is the upper α critical point of the null distribution of the random variable, U.

The p-value is defined to be $P(Y \geq u_y) = P(U \leq u_x)$. One advantage of Mann-Whitney U test is that we can use either u_y or u_x to carry out the test.

For large n_y and n_x , the null distribution of U can be well approximated by a normal distribution with mean $E(U) = n_y n_x / 2$ and variance $var(U) = n_y n_x (n+1) / 12$. A large sample z-test can be based on the statistic:

$$z = \frac{u_y - n_y n_x / 2 - 1/2}{\sqrt{n_y n_x (n+1) / 12}}$$

The test rejects H_0 at level α if $z \geq z_\alpha$ or if $u_y \geq u_{n_y, n_x, \alpha}$ where

$$u_{n_y, n_x, \alpha} \approx n_y n_x / 2 + 1/2 + z_\alpha \sqrt{n_y n_x (n+1) / 12}$$

For the 2-sided test, we use the test statistic $u_{max} = \max(u_y, u_x)$ and $u_{min} = \min(u_y, u_x)$ and p-value is given by

$$p-value = 2P(U \geq u_{max}) = 2P(U \leq u_{min})$$

Since we assume there are no ties (when $y_i = x_j$), we count 1/2 towards both u_y and u_x . Even though the sampling distribution is not the same, but large sample approximation is still reasonable,

4.3 Categorical Data Analysis

Categorical Data Analysis when we have categorical outcomes

- Nominal variables: no logical ordering (e.g., sex)
- Ordinal variables: logical order, but relative distances between values are not clear (e.g., small, medium, large)

The distribution of one variable changes when the level (or values) of the other variable change. The row percentages are different in each column.

4.3.1 Inferences for Small Samples

The approximate tests based on the asymptotic normality of $\hat{p}_1 - \hat{p}_2$ do not apply for small samples.

Using **Fisher's Exact Test** to evaluate $H_0 : p_1 = p_2$

- Assume X_1 and X_2 are independent Binomial
- Let x_1 and x_2 be the corresponding observed values.
- Let $n = n_1 + n_2$ be the total sample size
- $m = x_1 + x_2$ be the observed number of successes.

- By assuming that m (total successes) is fixed, and conditioning on this value, one can show that the conditional distribution of the number of successes from sample 1 is Hypergeometric
- If we want to test $H_0 : p_1 = p_2$ and $H_a : p_1 \neq p_2$, we have

$$Z^2 = \left(\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}} \right)^2 \sim \chi^2_{1,\alpha}$$

where $\chi^2_{1,\alpha}$ is the upper α percentage point for the central Chi-squared with one d.f.

This extends to the contingency table setting: whether the observed frequencies are equal to those expected under a null hypothesis of no association.

4.3.2 Test of Association

Pearson Chi-square test statistic is

$$\chi^2 = \sum_{\text{all categories}} \frac{(observed - expected)^2}{expected}$$

Comparison of proportions for several independent surveys or experiments

	Experiment 1	Experiment 2	...	Experiment k
Number of successes	x_1	x_2	...	x_k
Number of failures	$n_1 - x_1$	$n_2 - x_2$...	$n_k - x_k$
	n_1	n_2	...	n_k

$H_0 : p_1 = p_2 = \dots = p_k$ vs. the alternative that the null is not true (at least one pair are not equal).

We estimate the common value of the probability of success on a single trial assuming H_0 is true:

$$\hat{p} = \frac{x_1 + x_2 + \dots + x_k}{n_1 + n_2 + \dots + n_k}$$

we use table of expected counts when H_0 is true:

success	$n_1 \hat{p}$	$n_2 \hat{p}$...	$n_k \hat{p}$
failure	$n_1(1 - \hat{p})$	$n_2(1 - \hat{p})$...	$n_k(1 - \hat{p})$

	n_1	n_2	...	n_k
--	-------	-------	-----	-------

$$\chi^2 = \sum_{\text{all cells in table}} \frac{(observed - expected)^2}{expected}$$

with $k-1$ degrees of freedom

4.3.2.1 Two-way Count Data

	1	2	...	j	...	c	Row Total
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2j}	...	n_{2c}	$n_{2.}$
.
r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rc}	$n_{r.}$
Column Total	$n_{.1}$	$n_{.2}$...	$n_{.j}$...	$n_{.c}$	n
Total							

Design 1

total sample size fixed $n = \text{constant}$ (e.g., survey on job satisfaction and income); both row and column totals are random variables

Design 2

Fix the sample size in each group (in each row) (e.g., Drug treatments success or failure); fixed number of participants for each treatment; independent random samples from the two row populations.

These different sampling designs imply two different probability models.

4.3.2.2 Total Sample Size Fixed

Design 1

random sample of size n drawn from a single population, and sample units are cross-classified into r row categories and c column

This results in an $r \times c$ table of observed counts

$$n_{ij} = 1, \dots, r; j = 1, \dots, c$$

Let p_{ij} be the probability of classification into cell (i,j) and $\sum_{i=1}^r \sum_{j=1}^c p_{ij} = 1$.

Let N_{ij} be the random variable corresponding to n_{ij}

The joint distribution of the N_{ij} is multinomial with unknown parameters p_{ij}

Denote the row variable by X and column variable by Y, then $p_{ij} = P(X = i, Y = j)$ and $p_{i\cdot} = P(X = i)$ and $p_{\cdot j} = P(Y = j)$ are the marginal probabilities. The null hypothesis that X and Y are statistically independent (i.e., no association) is just:

$$H_0 : p_{ij} = P(X = i, Y = j) = P(X = i)P(Y = j) = p_{i\cdot}p_{\cdot j} \quad H_a : p_{ij} \neq p_{i\cdot}p_{\cdot j}$$

for all i,j.

4.3.2.3 Row Total Fixed

Design 2

Random samples of sizes n_1, \dots, n_r are drawn independently from $r \geq 2$ row populations. In this case, the 2-way table row totals are $n_{i\cdot} = n_i$ for $i = 1, \dots, r$.

The counts from each row are modeled by independent multinomial distributions.

X is fixed, Y is observed.

Then, p_{ij} represent conditional probabilities $p_{ij} = P(Y = j|X = i)$

The null hypothesis is the probability of response j is the same, regardless of the row population (i.e., no association):

$$H_0 : p_{ij} = P(Y = j|X = i) = p_j \text{ for all } i, j = 1, 2, \dots, \text{cor } H_0 : (p_{i1}, p_{i2}, \dots, p_{ic}) = (p_1, p_2, \dots, p_c) \text{ for all } i \quad H_a : (p_{i1}, p_{i2}, \dots, p_{ic}) \neq (p_1, p_2, \dots, p_c) \text{ for all } i$$

Although the hypotheses to be tested are different for two sampling designs,
The chi-square test is identical

We have estimated expected frequencies:

$$\hat{e}_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n}$$

The Chi-square statistic is

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \sim \chi_{(r-1)(c-1)}^2$$

α -level test rejects H_0 if $\chi^2 > \chi_{(r-1)(c-1), \alpha}^2$

4.3.2.4 Pearson Chi-square Test

- Determine whether an association exists
- Sometimes, H_0 represents the model whose validity is to be tested. Contrast this with the conventional formulation of H_0 as the hypothesis that is to be disproved. The goal in this case is not to disprove the model, but to see whether data are consistent with the model and if deviation can be attributed to chance.
- These tests do not measure the strength of an association.
- These tests depend on and reflect the sample size - double the sample size by copying each observation, double the χ^2 statistic even though the strength of the association does not change.
- The Pearson Chi-square Test is not appropriate when more than about 20% of the cells have an expected cell frequency of less than 5 (large-sample p-values not appropriate).
- When the sample size is small the exact p-values can be calculated (this is prohibitive for large samples); calculation of the exact p-values assumes that the column totals and row totals are fixed.

```
july.x=480
july.n=1000
sept.x=704
sept.n=1600
```

$$H_0 : p_J = 0.5 \quad H_a : p_J < 0.5$$

```
prop.test(x=july.x,n=july.n,p=0.5,alternative="less",correct=F)
#>
#> 1-sample proportions test without continuity correction
#>
#> data: july.x out of july.n, null probability 0.5
#> X-squared = 1.6, df = 1, p-value = 0.103
#> alternative hypothesis: true p is less than 0.5
#> 95 percent confidence interval:
#> 0.0000000 0.5060055
#> sample estimates:
#> p
#> 0.48
```

$$H_0 : p_J = p_S \quad H_a : p_j \neq p_S$$

```
prop.test(x=c(july.x,sept.x),n=c(july.n,sept.n),correct=F)
#>
#> 2-sample test for equality of proportions without continuity
#> correction
#>
#> data: c(july.x, sept.x) out of c(july.n, sept.n)
#> X-squared = 3.9701, df = 1, p-value = 0.04632
#> alternative hypothesis: two.sided
#> 95 percent confidence interval:
#> 0.0006247187 0.0793752813
#> sample estimates:
#> prop 1 prop 2
#> 0.48 0.44
```

4.3.3 Ordinal Association

- An ordinal association implies that as one variable increases, the other tends to increase or decrease (depending on the nature of the association).
- For tests for variables with two or more levels, the levels must be in a logical ordering.

4.3.3.1 Mantel-Haenszel Chi-square Test

The Mantel-Haenszel Chi-square Test is more powerful for testing ordinal associations, but does not test for the strength of the association.

This test is presented in the case where one has a series of 2×2 tables that examine the same effects under different conditions (If there are K such tables, we have $2 \times 2 \times K$ table)

In stratum k, given the marginal totals $(n_{1k}, n_{2k}, n_{1..k}, n_{2..k})$, the sampling model for cell counts is the Hypergeometric (knowing n_{11k} determines $(n_{12k}, n_{21k}, n_{22k})$, given the marginal totals)

Assuming conditional independence, the Hypergeometric mean and variance of n_{11k} are

$$m_{11k} = E(n_{11k}) = \frac{n_{1..k}n_{.1k}}{n_{..k}} \quad var(n_{11k}) = \frac{n_{1..k}n_{2..k}n_{.1k}n_{.2k}}{n_{..k}^2(n_{..k}-1)}$$

To test conditional independence, Mantel and Haenszel proposed

$$M^2 = \frac{(|\sum_k n_{11k} - \sum_k m_{11k}| - .5)^2}{\sum_k var(n_{11k})} \sim \chi_1^2$$

This method can be extended to general I x J x K tables.

(2 x 2 x 3) table

```
Bron=array(c(20, 9, 382, 214, 10, 7, 172, 120, 12, 6, 327, 183), dim = c(2, 2, 3), dimnames = list(c("Bronchitis", "No Bronchitis"), c("Yes", "No"), c("High", "Low")))
margin.table(Bron,c(1,2))
#>          Bronchitis
#> Particulate Yes  No
#>           High 42 881
#>           Low  22 517
# assess whether the relationship between Bronchitis by Particulate Level varies by Age
library(samplesizeCMH)
marginal_table=margin.table(Bron,c(1,2))
odds.ratio(marginal_table)
#> [1] 1.120318

# whether these odds vary by age. The conditional odds can be calculated using the odds.ratio function
apply(Bron,3,odds.ratio)
#>    15-24      25-39      40+
#> 1.2449098 0.9966777 1.1192661

# Mantel-Haenszel Test
mantelhaen.test(Bron,correct=T)
#>
#> Mantel-Haenszel chi-squared test with continuity correction
#>
#> data: Bron
#> Mantel-Haenszel X-squared = 0.11442, df = 1, p-value = 0.7352
#> alternative hypothesis: true common odds ratio is not equal to 1
#> 95 percent confidence interval:
#> 0.6693022 1.9265813
#> sample estimates:
#> common odds ratio
#> 1.135546
```

4.3.3.1.1 McNemar's Test special case of Mantel-Haenszel Chi-square Test

```
vote=cbind(c(682,22),c(86,810))
mcnemar.test(vote,correct=T)
#>
#> McNemar's Chi-squared test with continuity correction
```

```
#>
#> data: vote
#> McNemar's chi-squared = 36.75, df = 1, p-value = 1.343e-09
```

4.3.3.2 Spearman Rank Correlation

To test for the strength of association between two ordinally scaled variables, we can use Spearman Rank Correlation statistic

Let X and Y be two random variables measured on an ordinal scale. Consider n pairs of observations (x_i, y_i) , $i = 1, \dots, n$

The Spearman Rank Correlation coefficient (denoted by r_S) is calculated using the Pearson correlation formula, but based on the ranks of x_i and y_i .

Spearman Rank Correlation be calculated

1. Assign ranks to x_i 's and y_i 's separately. Let $u_i = rank(x_i)$ and $v_i = rank(y_i)$
2. Calculate r_S using the formula for the Pearson correlation coefficient, but applied to the ranks:

$$r_S = \frac{\sum_{i=1}^n (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{(\sum_{i=1}^n (u_i - \bar{u})^2)(\sum_{i=1}^n (v_i - \bar{v})^2)}}$$

r_S ranges between -1 and +1 , with

- $r_S = -1$ if there is a perfect negative monotone association
- $r_S = +1$ if there is a perfect positive monotone association between X and Y.

To test

H_0 : X and Y independent

H_a : X and Y positively associated

For large n (e.g., $n \geq 10$),

$$r_S \sim N(0, 1/(n-1))$$

Then,

$$Z = r_s \sqrt{n-1} \sim N(0, 1)$$

4.4 Divergence Metrics and Test for Comparing Distributions

Similarity among distributions using divergence statistics, which is different from

- Deviation statistics: difference between the realization of a variable and some value (e.g., mean). Statistics of the deviation distributions consist of standard deviation, average absolute deviation, median absolute deviation , maximum absolute deviation.
- Deviance statistics: goodness-of-fit statistic for statistical models (comparable to the sum of squares of residuals in OLS to cases that use ML estimation). Usually used in generalized linear models.

Divergence statistics is a statistical distance (different from metrics)

- Divergences do not require symmetry
- Divergences generalize squared distance (instead of linear distance). Hence, fail the triangle inequality

Can be used for

- Detecting data drift in machine learning
- Feature selections
- Variational Auto Encoder
- Detect similarity between policies (i.e., distributions) in reinforcement learning
- To see consistency in two measured variables of two constructs.

Techniques

- Kullback-Leibler Divergence
- Jensen-Shannon Divergence
- Kolmogorov-Smirnov Test

Packages

- `entropy`
- `phileentropy`

4.4.1 Kullback-Leibler Divergence

- Also known as relative entropy
- Not a metric (does not satisfy the triangle inequality)
- Can be generalized to the multivariate case
- Measure the similarity between two discrete probability distributions
 - P = true data distribution
 - Q = predicted data distribution
- It quantifies info loss when moving from P to Q (i.e., information loss when P is approximated by Q)

Discrete

$$D_{KL}(P||Q) = \sum_i P_i \log\left(\frac{P_i}{Q_i}\right)$$

Continuous

$$D_{KL}(P||Q) = \int P(x) \log\left(\frac{P(x)}{Q(x)}\right) dx$$

where

- $K \in [0, \infty)$ from similar to diverge
- Non-symmetric between two distributions: $D_{KL}(P|Q) \neq D_{KL}(Q|P)$

```
library(philentropy)
# philentropy::dist.diversity(rbind(X = 1:10 / sum(1:10), Y = 1:20 / sum(1:20)),
#                               p = 2,
#                               unit = "log2")

# continuous
KL(rbind(X = 1:10/sum(1:10), Y = 1:10/sum(1:10)), unit = "log2")
#> kullback-leibler
#>          0

# discrete
KL(rbind(X = 1:10, Y = 1:10), est.prob = "empirical")
#> kullback-leibler
#>          0
```

4.4.2 Jensen-Shannon Divergence

- Also known as info radius or total divergence to the average

$$D_{JS}(P||Q) = \frac{1}{2}(D_{KL}(P||M) + D_{KL}(Q||M))$$

where

- $M = \frac{1}{2}(P + Q)$ is a mixed distribution
- $D_{JS} \in [0, 1]$ for \log_2 and $D_{JS} \in [0, \ln(2)]$ for \log_e

```
library(philentropy)
# continuous
JSD(rbind(X = 1:10, Y = 1:20), unit = "log2")
#> jensen-shannon
#>      20.03201

# discrete
JSD(rbind(X = 1:10, Y = 1:20), est.prob = "empirical")
#> jensen-shannon
#>      0.06004756
```

4.4.3 Wasserstein Distance

- measure the distance between two empirical cdfs

$$W = \int_{x \in R} |E(x) - F(X)|^p$$

- This is also a test statistics

```
set.seed(1)
transport::wasserstein1d(rnorm(100), rnorm(100, mean = 1))
#> [1] 0.8533046

set.seed(1)
# Wasserstein metric
twosamples::wass_stat(rnorm(100), rnorm(100, mean = 1))
#> [1] 0.8533046
```

```
set.seed(1)
# permutation-based tw sample test using Wasserstein metric
twosamples::wass_test(rnorm(100), rnorm(100, mean = 1))
#> Test Stat   P-Value
#> 0.8533046 0.0002500
#> attr(",details")
#>      n1      n2 n.boo
#>    100     100 2000
```

4.4.4 Kolmogorov-Smirnov Test

- Can be used for continuous distribution

H_0 : Empirical distribution follows a specified distribution

H_1 : Empirical distribution does not follow a specified distribution

- Using non-parametric

$$D = \max |P(X) - Q(X)|$$

- $D \in [0,1]$ from the densities are evenly distributed to not evenly distributed

```
library(entropy)
library(tidyverse)

lst = list(sample_1 = c(1:20), sample_2 = c(2:30), sample_3 = c(3:30))

expand.grid(1:length(lst), 1:length(lst)) %>%
  rowwise() %>%
  mutate(KL = KL.empirical(lst[[Var1]], lst[[Var2]]))
#> # A tibble: 9 x 3
#> # Rowwise:
#>   Var1  Var2    KL
#>   <int> <int>  <dbl>
#> 1     1     1  0
#> 2     2     1 0.150
#> 3     3     1 0.183
#> 4     1     2 0.704
#> 5     2     2  0
#> 6     3     2 0.0679
```

```
#> 7     1     3 0.622
#> 8     2     3 0.0870
#> 9     3     3 0
```

To use the test for discrete data, use bootstrap version of the KS test (bypass the continuity requirement)

```
Matching::ks.boot(Tr = c(0:10), Co = c(0:10))
#> $ks.boot$pvalue
#> [1] 1
#>
#> $ks
#>
#> Two-sample Kolmogorov-Smirnov test
#>
#> data: Tr and Co
#> D = 0, p-value = 1
#> alternative hypothesis: two-sided
#>
#>
#> $nboots
#> [1] 1000
#>
#> attr(", "class")
#> [1] "ks.boot"
```

II. REGRESSION

Chapter 5

Linear Regression

Estimator Desirable Properties

1. Unbiased
2. Consistency
 - $\text{plim} \hat{\beta}_n = \beta$
 - based on the law of large numbers, we can derive consistency
 - More observations means more precise, closer to the true value.
3. Efficiency
 - Minimum variance in comparison to another estimator.
 - OLS is BLUE (best linear unbiased estimator) means that OLS is the most efficient among the class of linear unbiased estimator Gauss-Markov Theorem
 - If we have correct distributional assumptions, then the Maximum Likelihood is asymptotically efficient among consistent estimators.

5.1 Ordinary Least Squares

The most fundamental model in statistics or econometric is a OLS linear regression. OLS = Maximum likelihood when the error term is assumed to be normally distributed.

Regression is still great if the underlying CEF (conditional expectation function) is not linear. Because regression has the following properties:

1. For $E[Y_i|X_{1i}, \dots, X_{Ki}] = a + \sum_{k=1}^K b_k X_{ki}$ (i.e., the CEF of Y_i on X_{1i}, \dots, X_{Ki} is linear, then the regression of Y_i on X_{1i}, \dots, X_{Ki} is the CEF
2. For $E[Y_i|X_{1i}, \dots, X_{Ki}]$ is a nonlinear function of the conditioning variables, the regression of Y_i on X_{1i}, \dots, X_{Ki} will give you the best linear approximation to the nonlinear CEF (i.e., minimize the expected squared deviation between the fitted values from the linear model and the CEF).

5.1.1 Simple Regression (Basic Model)

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

- Y_i : response (dependent) variable at i-th observation
- β_0, β_1 : regression parameters for intercept and slope.
- X_i : known constant (independent or predictor variable) for i-th observation
- ϵ_i : random error term

$$E(\epsilon_i) = 0, var(\epsilon_i) = \sigma^2 cov(\epsilon_i, \epsilon_j) = 0 \text{ for all } i \neq j$$

Y_i is random since ϵ_i is:

$$\begin{aligned} E(Y_i) &= E(\beta_0 + \beta_1 X_i + \epsilon_i) \\ &= E(\beta_0) + E(\beta_1 X_i) + E(\epsilon) \\ &= \beta_0 + \beta_1 X_i \end{aligned}$$

$$\begin{aligned} var(Y_i) &= var(\beta_0 + \beta_1 X_i + \epsilon_i) \\ &= var(\epsilon_i) \\ &= \sigma^2 \end{aligned}$$

Since $cov(\epsilon_i, \epsilon_j) = 0$ (uncorrelated), the outcome in any one trial has no effect on the outcome of any other. Hence, Y_i, Y_j are uncorrelated as well (conditioned on the X's)

Note

Least Squares does not require a distributional assumption

Relationship between bivariate regression and covariance

Covariance between 2 variables:

$$C(X_i, Y_i) = E[(X_i - E[X_i])(Y_i - E[Y_i])]$$

Which has the following properties

1. $C(X_i, X_i) = \sigma_X^2$
2. If either $E(X_i) = 0 | E(Y_i) = 0$, then $Cov(X_i, Y_i) = E[X_i Y_i]$
3. Given $W_i = a + bX_i$ and $Z_i = c + dY_i$, then $Cov(W_i, Z_i) = bdC(X_i, Y_i)$

For the bivariate regression, the slope is

$$\beta = \frac{Cov(Y_i, X_i)}{Var(X_i)}$$

To extend this to a multivariate case

$$\beta_k = \frac{C(Y_i, \tilde{X}_{ki})}{Var(\tilde{X}_{ki})}$$

Where \tilde{X}_{ki} is the residual from a regression of X_{ki} on the $K - 1$ other covariates included in the model

And intercept

$$\alpha = E[Y_i] - \beta E(X_i)$$

5.1.1.1 Estimation

Deviation of Y_i from its expected value:

$$Y_i - E(Y_i) = Y_i - (\beta_0 + \beta_1 X_i)$$

Consider the sum of the square of such deviations:

$$Q = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$

$$b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} b_0 = \frac{1}{n} \left(\sum_{i=1}^n Y_i - b_1 \sum_{i=1}^n X_i \right) = \bar{Y} - b_1 \bar{X}$$

5.1.1.2 Properties of Least Least Estimators

$$E(b_1) = \beta_1 E(b_0) = E(\bar{Y}) - \bar{X} \beta_1 E(\bar{Y}) = \beta_0 + \beta_1 \bar{X} E(b_0) = \beta_0 var(b_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} var(b_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right)$$

$\text{var}(b_1)$ approaches 0 as more measurements are taken at more X_i values (unless X_i is at its mean value)

$\text{var}(b_0)$ approaches 0 as n increases when the X_i values are judiciously selected.

Mean Square Error

$$MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}$$

Unbiased estimator of MSE:

$$E(MSE) = \sigma^2$$

$$s^2(b_1) = \widehat{\text{var}(b_1)} = \frac{MSE}{\sum_{i=1}^n (X_i - \bar{X})^2} s^2(b_0) = \widehat{\text{var}(b_0)} = MSE \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)$$

$$E(s^2(b_1)) = \text{var}(b_1) E(s^2(b_0)) = \text{var}(b_0)$$

5.1.1.3 Residuals

$$e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$$

- e_i is an estimate of $\epsilon_i = Y_i - E(Y_i)$
- ϵ_i is always unknown since we don't know the true β_0, β_1

$$\sum_{i=1}^n e_i = 0 \quad \sum_{i=1}^n X_i e_i = 0$$

Residual properties

1. $E[e_i] = 0$
2. $E[X_i e_i] = 0$ and $E[\hat{Y}_i e_i] = 0$

5.1.1.4 Inference

Normality Assumption

- Least Squares estimation does not require assumptions of normality.
- However, to do inference on the parameters, we need distributional assumptions.

- Inference on β_0, β_1 and Y_h are not extremely sensitive to moderate departures from normality, especially if the sample size is large
- Inference on Y_{pred} is very sensitive to the normality assumptions.

Normal Error Regression Model

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

5.1.1.4.1 β_1 Under the normal error model,

$$b_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2})$$

A linear combination of independent normal random variable is normally distributed

Hence,

$$\frac{b_1 - \beta_1}{s(b_1)} \sim t_{n-2}$$

A $(1 - \alpha)100\%$ confidence interval for β_1 is

$$b_1 \pm t_{t-\alpha/2; n-2} s(b_1)$$

5.1.1.4.2 β_0 Under the normal error model, the sampling distribution for b_0 is

$$b_0 \sim N(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right))$$

Hence,

$$\frac{b_0 - \beta_0}{s(b_0)} \sim t_{n-2}$$

A $(1 - \alpha)100\%$ confidence interval for β_0 is

$$b_0 \pm t_{1-\alpha/2; n-2} s(b_0)$$

5.1.1.4.3 Mean Response Let X_h denote the level of X for which we wish to estimate the mean response

- We denote the mean response when $X = X_h$ by $E(Y_h)$
- A point estimator of $E(Y_h)$ is \hat{Y}_h :

$$\hat{Y}_h = b_0 + b_1 X_h$$

Note

$$E(\bar{Y}_h) = E(b_0 + b_1 X_h) = \beta_0 + \beta_1 X_h = E(Y_h)$$

(unbiased estimator)

$$\begin{aligned} var(\hat{Y}_h) &= var(b_0 + b_1 X_h) \\ &= var(\bar{Y} + b_1(X_h - \bar{X})) \\ &= var(\bar{Y}) + (X_h - \bar{X})^2 var(b_1) + 2(X_h - \bar{X}) cov(\bar{Y}, b_1) \\ &= \frac{\sigma^2}{n} + (X_h - \bar{X})^2 \frac{\sigma^2}{\sum(X_i - \bar{X})^2} \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \end{aligned}$$

Since $cov(\bar{Y}, b_1) = 0$ due to the iid assumption on ϵ_i

An estimate of this variance is

$$s^2(\hat{Y}_h) = MSE \left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)$$

the sampling distribution for the mean response is

$$\hat{Y}_h \sim N(E(Y_h), var(\hat{Y}_h)) \frac{\hat{Y}_h - E(Y_h)}{s(\hat{Y}_h)} \sim t_{n-2}$$

A $100(1 - \alpha)\%$ CI for $E(Y_h)$ is

$$\hat{Y}_h \pm t_{1-\alpha/2; n-2} s(\hat{Y}_h)$$

5.1.1.4.4 Prediction of a new observation Regarding the Mean Response, we are interested in estimating **mean** of the distribution of Y given a certain X.

Now, we want to **predict** an individual outcome for the distribution of Y at a given X. We call \hat{Y}_{pred}

Estimation of mean response versus prediction of a new observation:

- the point estimates are the same in both cases: $\hat{Y}_{pred} = \hat{Y}_h$
- It is the variance of the prediction that is different; hence, prediction intervals are different than confidence intervals. The prediction variance must consider:
 - Variation in the mean of the distribution of Y
 - Variation within the distribution of Y

We want to predict: mean response + error

$$\beta_0 + \beta_1 X_h + \epsilon$$

Since $E(\epsilon) = 0$, use the least squares predictor:

$$\hat{Y}_h = b_0 + b_1 X_h$$

The variance of the predictor is

$$\begin{aligned} var(b_0 + b_1 X_h + \epsilon) &= var(b_0 + b_1 X_h) + var(\epsilon) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) + \sigma^2 \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \end{aligned}$$

An estimate of the variance is given by

$$s^2(pred) = MSE \left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \frac{\hat{Y}_h - \bar{Y}_h}{s(pred)} \sim t_{n-2}$$

100(1 - α)% prediction interval is

$$\bar{Y}_h \pm t_{1-\alpha/2; n-2} s(pred)$$

The prediction interval is very sensitive to the distributional assumption on the errors, ϵ

5.1.1.4.5 Confidence Band We want to know the confidence interval for the entire regression line, so we can draw conclusions about any and all mean response for the entire regression line $E(Y) = \beta_0 + \beta_1 X$ rather than for a given response Y

Working-Hotelling Confidence Band

For a given X_h , this band is

$$\hat{Y}_h \pm W s(\hat{Y}_h)$$

where $W^2 = 2F_{1-\alpha;2,n-2}$, which is just 2 times the F-stat with 2 and n-2 degrees of freedom

- the interval width will change with each X_h (since $s(\hat{Y}_h)$ changes)
- the boundary values for this confidence band will always define a hyperbole containing the regression line
- will be smallest at $X = \bar{X}$

5.1.1.5 ANOVA

Partitioning the Total Sum of Squares: Consider the corrected Total sum of squares:

$$SSTO = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

Measures the overall dispersion in the response variable

We use the term corrected because we correct for mean, the uncorrected total sum of squares is given by $\sum Y_i^2$

use $\hat{Y}_i = b_0 + b_1 X_i$ to estimate the conditional mean for Y at X_i

$$\begin{aligned} \sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\ &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + 2 \sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \\ &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\bar{Y}_i - \bar{Y})^2 \end{aligned}$$

$$STTO = SSE + SSR$$

where SSR is the regression sum of squares, which measures how the conditional mean varies about a central value.

The cross-product term in the decomposition is 0:

$$\begin{aligned}
 \sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) &= \sum_{i=1}^n (Y_i - \bar{Y} - b_1(X_i - \bar{X}))(\bar{Y} + b_1(X_i - \bar{X}) - \bar{Y}) \\
 &= b_1 \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}) - b_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 \\
 &= b_1 \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \sum_{i=1}^n (X_i - \bar{X})^2 - b_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 \\
 &= b_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 - b_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 \\
 &= 0
 \end{aligned}$$

$$\begin{aligned}
 SSTO &= SSR + SSE \\
 (n - 1d.f.) &= (1d.f.) + (n - 2d.f.)
 \end{aligned}$$

Source of Variation	Sum of Squares	df	Mean Square	F
Regression (model)	SSR	1	MSR = SSR/df	MSR/MSE
Error	SSE	n-2	MSE = SSE/df	
Total (Corrected)	SSTO	n-1		

$$E(MSE) = \sigma^2 E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^n (X_i - \bar{X})^2$$

- If $\beta_1 = 0$, then these two expected values are the same
- if $\beta_1 \neq 0$ then $E(MSR)$ will be larger than $E(MSE)$

which means the ratio of these two quantities, we can infer something about β_1

Distribution theory tells us that if $\epsilon_i \sim iidN(0, \sigma^2)$ and assuming $H_0 : \beta_1 = 0$ is true,

$$\frac{MSE}{\sigma^2} \sim \chi_{n-2}^2 \frac{MSR}{\sigma^2} \sim \chi_1^2 \text{ if } \beta_1 = 0$$

where these two chi-square random variables are independent.

Since the ratio of 2 independent chi-square random variable follows an F distribution, we consider:

$$F = \frac{MSR}{MSE} \sim F_{1,n-2}$$

when $\beta_1 = 0$. Thus, we reject $H_0 : \beta_1 = 0$ (or $E(Y_i) = \text{constant}$) at α if

$$F > F_{1-\alpha;1,n-2}$$

this is the only null hypothesis that can be tested with this approach.

Coefficient of Determination

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

where $0 \leq R^2 \leq 1$

Interpretation: The proportionate reduction of the total variation in Y after fitting a linear model in X.

It is not really correct to say that R^2 is the “variation in Y explained by X”.

R^2 is related to the correlation coefficient between Y and X:

$$R^2 = (r)^2$$

where $r = \text{corr}(x, y)$ is an estimate of the Pearson correlation coefficient. Also, note

$$b_1 = \left(\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)^{1/2} r = \frac{s_y}{s_x} r$$

Lack of Fit

$Y_{11}, Y_{21}, \dots, Y_{n_1,1}$: n_1 repeat obs at X_1

$Y_{1c}, Y_{2c}, \dots, Y_{n_c,c}$: n_c repeat obs at X_c

So, there are c distinct X values.

Let \bar{Y}_j be the mean over replicates for X_j

Partition the Error Sum of Squares:

$$\begin{aligned}
\sum_i \sum_j (Y_{ij} - \hat{Y}_{ij})^2 &= \sum_i \sum_j (Y_{ij} - \bar{Y}_j + \bar{Y}_j - \hat{Y}_{ij})^2 \\
&= \sum_i \sum_j (Y_{ij} - \bar{Y}_j)^2 + \sum_i \sum_j (\bar{Y}_j - \hat{Y}_{ij})^2 + \text{cross product term} \\
&= \sum_i \sum_j (Y_{ij} - \bar{Y}_j)^2 + \sum_j n_j (\bar{Y}_j - \hat{Y}_{ij})^2 \\
SSE &= SSPE + SSLF
\end{aligned}$$

- SSPE: “pure error sum of squares” has $n-c$ degrees of freedom since we need to estimate c means
- SSLF: “lack of fit sum of squares” has $c - 2$ degrees of freedom (the number of unique X values - number of parameters used to specify the conditional mean regression model)

$$MSPE = \frac{SSPE}{df_{pe}} = \frac{SSPE}{n-c} MSLF = \frac{SSLF}{df_{lf}} = \frac{SSLF}{c-2}$$

The **F-test for Lack-of-Fit** tests

$$H_0 : Y_{ij} = \beta_0 + \beta_1 X_i + \epsilon_{ij}, \epsilon_{ij} \sim iidN(0, \sigma^2) H_a : Y_{ij} = \alpha_0 + \alpha_1 X_i + f(X_i, Z_1, \dots) + \epsilon_{ij}^*, \epsilon_{ij}^* \sim iidN(0, \sigma^2)$$

$$E(MSPE) = \sigma^2 \text{ under either } H_0, H_a$$

$$E(MSLF) = \sigma^2 + \frac{\sum n_j(f(X_i, \dots))^2}{n-2} \text{ in general and}$$

$$E(MSLF) = \sigma^2 \text{ when } H_0 \text{ is true}$$

We reject H_0 (i.e., the model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ is not adequate) if

$$F = \frac{MSLF}{MSPE} > F_{1-\alpha; c-2, n-c}$$

Failing to reject H_0 does not imply that $H_0 : Y_{ij} = \beta_0 + \beta_1 X_i + \epsilon_{ij}$ is exactly true, but it suggests that this model may provide a reasonable approximation to the true model.

Source of Variation	Sum of Squares	df	Mean Square	F
Regression	SSR	1	MSR	MSR / MSE
Error	SSE	n-2	MSE	

Source of Variation	Sum of Squares	df	Mean Square	F
Lack of fit	SSLF	c-2	MSLF	MSLF / MSPE
Pure Error	SSPE	n-c	MSPE	
Total(Corrected)	SSTO	n-1		

Repeat observations have an effect on R^2 :

- It is impossible for R^2 to attain 1 when repeat obs. exist (SSE can't be 0)
- The maximum R^2 attainable in this situation:

$$R_{max}^2 = \frac{SSTO - SSPE}{SSTO}$$

- Not all levels of X need have repeat observations.
- Typically, when H_0 is appropriate, one still uses MSE as the estimate for σ^2 rather than MSPE, Since MSE has more degrees of freedom, sometimes people will pool these estimates.

Joint Inference

The confidence coefficient for both β_0 and β_1 considered simultaneously is $\leq \alpha$

Let

- \bar{A}_1 be the event that the first interval covers β_0
- \bar{A}_2 be the event that the second interval covers β_1

$$P(\bar{A}_1) = 1 - \alpha P(\bar{A}_2) = 1 - \alpha$$

The probability that both \bar{A}_1 and \bar{A}_2

$$\begin{aligned} P(\bar{A}_1 \cap \bar{A}_2) &= 1 - P(\bar{A}_1 \cup \bar{A}_2) \\ &= 1 - P(A_1) - P(A_2) + P(A_1 \cap A_2) \\ &\geq 1 - P(A_1) - P(A_2) \\ &= 1 - 2\alpha \end{aligned}$$

If β_0 and β_1 have separate 95% confidence intervals, the joint (family) confidence coefficient is at least $1 - 2(0.05) = 0.9$. This is called a **Bonferroni Inequality**. We could use a procedure in which we obtained $1 - \alpha/2$ confidence intervals

for the two regression parameters separately, then the joint (Bonferroni) family confidence coefficient would be at least $1 - \alpha$

The $1 - \alpha$ joint Bonferroni confidence interval for β_0 and β_1 is given by calculating:

$$b_0 \pm Bs(b_0)b_1 \pm Bs(b_1)$$

where $B = t_{1-\alpha/4;n-2}$

Interpretation: If repeated samples were taken and the joint $(1 - \alpha)$ intervals for β_0 and β_1 were obtained, $(1 - \alpha)100\%$ of the joint intervals would contain the true pair (β_0, β_1) . That is, in $\alpha \times 100\%$ of the samples, one or both intervals would not contain the true value.

- The Bonferroni interval is **conservative**. It is a lower bound and the joint intervals will tend to be correct more than $(1 - \alpha)100\%$ of the time (lower power). People usually consider a larger α for the Bonferroni joint tests (e.g., $\alpha = 0.1$)
- The Bonferroni procedure extends to testing more than 2 parameters. Say we are interested in testing $\beta_0, \beta_1, \dots, \beta_{g-1}$ (g parameters to test). Then, the joint Bonferroni interval is obtained by calculating the $(1 - \alpha/g) 100\%$ level interval for each separately.
- For example, if $\alpha = 0.05$ and $g = 10$, each individual test is done at the $1 - \frac{.05}{10}$ level. For 2-sided intervals, this corresponds to using $t_{1-\frac{.05}{2(10)};n-p}$ in the CI formula. This procedure works best if g is relatively small, otherwise the intervals for each individual parameter are very wide and the test is way too conservative.
- b_0, b_1 are usually correlated (negatively if $\bar{X} > 0$ and positively if $\bar{X} < 0$)
- Other multiple comparison procedures are available.

5.1.1.6 Assumptions

- Linearity of the regression function
- Error terms have constant variance
- Error terms are independent
- No outliers
- Error terms are normally distributed
- No Omitted variables

5.1.1.7 Diagnostics

Constant Variance
Plot residuals vs. X

Outliers
 plot residuals vs. X
 box plots
 stem-leaf plots
 scatter plots

we could use standardize the residuals to have unit variance. These standardized residuals are called studentized residuals:

$$r_i = \frac{e_i - \bar{e}}{s(e_i)} = \frac{e_i}{s(e_i)}$$

A simplified standardization procedure gives semi-studentized residuals:

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$$

Non-independent of Error Terms
 plot residuals vs. time

Residuals e_i are not independent random variables because they involve the fitted values \hat{Y}_i , which are based on the same fitted regression function.

If the sample size is large, the dependency among e_i is relatively unimportant.

To detect non-independence, it helps to plot the residual for the i-th response vs. the (i-1)-th

Non-normality of Error Terms

to detect non-normality (distribution plots of residuals, box plots of residuals, stem-leaf plots of residuals, normal probability plots of residuals)

- Need relatively large sample sizes.
- Other types of departure affect the distribution of the residuals (wrong regression function, non-constant error variance,...)

5.1.1.7.1 Objective Tests of Model Assumptions

- Normality
 - Use Methods based on empirical cumulative distribution function to test on residuals.
- Constancy of error variance
 - Brown-Forsythe Test (Modified Levene Test)
 - Breusch-Pagan Test (Cook-Weisberg Test)

5.1.1.8 Remedial Measures

If the simple linear regression is not appropriate, one can:

- more complicated models
- transformations on X and/or Y (may not be “optimal” results)

Remedial measures based on deviations:

- Non-linearity:
 - Transformations
 - more complicated models
- Non-constant error variance:
 - Weighted Least Squares
 - Transformations
- Correlated errors:
 - serially correlated error models (times series)
- Non-normality
- Additional variables: multiple regression.
- Outliers:
 - Robust estimation.

5.1.1.8.1 Transformations use transformations of one or both variables before performing the regression analysis.

The properties of least-squares estimates apply to the transformed regression, not the original variable.

If we transform the Y variable and perform regression to get:

$$g(Y_i) = b_0 + b_1 X_i$$

Transform back:

$$\hat{Y}_i = g^{-1}(b_0 + b_1 X_i)$$

\hat{Y}_i will be biased. we can correct this bias.

Box-Cox Family Transformations

$$Y' = Y^\lambda$$

where λ is a parameter to be determined from the data.

λ	Y'
2	Y^2
0.5	\sqrt{Y}
0	$\ln(Y)$
-0.5	$1/\sqrt{Y}$
-1	$1/Y$

To pick λ , we can do estimation by:

- trial and error
- maximum likelihood
- numerical search

Variance Stabilizing Transformations

A general method for finding a variance stabilizing transformation, when the standard deviation is a function of the mean, is the **delta method** - an application of a Taylor series expansion.

$$\sigma = \sqrt{\text{var}(Y)} = f(\mu)$$

where $\mu = E(Y)$ and $f(\mu)$ is some smooth function of the mean.

Consider the transformation $h(Y)$. Expand this function in a Taylor series about μ . Then,

$$h(Y) = h(\mu) + h'(\mu)(Y - \mu) + \text{small terms}$$

we want to select the function $h(\cdot)$ so that the variance of $h(Y)$ is nearly constant for all values of $\mu = E(Y)$:

$$\begin{aligned} \text{const} &= \text{var}(h(Y)) \\ &= \text{var}(h(\mu) + h'(\mu)(Y - \mu)) \\ &= (h'(\mu))^2 \text{var}(Y - \mu) \\ &= (h'(\mu))^2 \text{var}(Y) \\ &= (h'(\mu))^2 (f(\mu))^2 \end{aligned}$$

we must have,

$$h'(\mu) \propto \frac{1}{f(\mu)}$$

then,

$$h(\mu) = \int \frac{1}{f(\mu)} d\mu$$

Example: For the Poisson distribution: $\sigma^2 = \text{var}(Y) = E(Y) = \mu$

Then,

$$\sigma = f(\mu) = \sqrt{\mu} h'(\mu) \propto \frac{1}{\mu} = \mu^{-0.5}$$

Then, the variance stabilizing transformation is:

$$h(\mu) = \int \mu^{-0.5} d\mu = \frac{1}{2}\sqrt{\mu}$$

hence, \sqrt{Y} is used as the variance stabilizing transformation.

If we don't know $f(\mu)$

1. Trial and error. Look at residuals plots
2. Ask researchers about previous studies or find published results on similar experiments and determine what transformation was used.
3. If you have multiple observations Y_{ij} at the same X values, compute \bar{Y}_i and s_i and plot them
If $s_i \propto \bar{Y}_i^\lambda$ then consider $s_i = a\bar{Y}_i^\lambda$ or $\ln(s_i) = \ln(a) + \lambda \ln(\bar{Y}_i)$. So regression the natural log of s_i on the natural log of \bar{Y}_i gives \hat{a} and $\hat{\lambda}$ and suggests the form of $f(\mu)$ If we don't have multiple obs, might still be able to "group" the observations to get \bar{Y}_i and s_i .

Transformation	Situation	Comments
\sqrt{Y}	$\text{var}(\epsilon_i) = kE(Y_i)$	counts from Poisson dist
$\sqrt{Y+1}$	$\text{var}(\epsilon_i) = kE(Y_i)$	small counts or zeroes
$\log(Y)$	$\text{var}(\epsilon_i) = k(E(Y_i))^2$	positive integers with wide range
$\log(Y+1)$	$\text{var}(\epsilon_i) = k(E(Y_i))^2$	some counts zero
$1/Y$	$\text{var}(\epsilon_i) = k(E(Y_i))^4$	most responses near zero, others large

Transformation	Situation	Comments
$\arcsin(\sqrt{Y})$	$\text{var}(\epsilon_i) = kE(Y_i)(1 - E(Y_i))$	data are binomial proportions or %

5.1.2 Multiple Linear Regression

Geometry of Least Squares

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\mathbf{b} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{H}\mathbf{y}\end{aligned}$$

sometimes H is denoted as P.

H is the projection operator.

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

is the projection of y onto the linear space spanned by the columns of X (model space). The dimension of the model space is the rank of X.

Facts:

1. H is symmetric (i.e., $H = H'$)

2. $HH = H$

$$\begin{aligned}\mathbf{HH} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{I}\mathbf{X}' \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\end{aligned}$$

3. H is an $n \times n$ matrix with $\text{rank}(H) = \text{rank}(X)$

4. $(\mathbf{I} - \mathbf{H}) = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is also a projection operator. It projects onto the $n - k$ dimensional space that is orthogonal to the k dimensional space spanned by the columns of X

5. $\mathbf{H}(\mathbf{I} - \mathbf{H}) = (\mathbf{I} - \mathbf{H})\mathbf{H} = \mathbf{0}$

Partition of uncorrected total sum of squares:

$$\begin{aligned}
\mathbf{y}'\mathbf{y} &= \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{e}'\mathbf{e} \\
&= (\mathbf{H}\mathbf{y})'(\mathbf{H}\mathbf{y}) + ((\mathbf{I} - \mathbf{H})\mathbf{y})'((\mathbf{I} - \mathbf{H})\mathbf{y}) \\
&= \mathbf{y}'\mathbf{H}'\mathbf{H}\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{H})'(\mathbf{I} - \mathbf{H})\mathbf{y} \\
&= \mathbf{y}'\mathbf{H}\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}
\end{aligned}$$

or partition for the corrected total sum of squares:

$$\mathbf{y}'(\mathbf{I} - \mathbf{H}_1)\mathbf{y} = \mathbf{y}'(\mathbf{H} - \mathbf{H}_1)\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}$$

where $H_1 = \frac{1}{n}J = 1'(1'1)1$

Source	SS	df	MS	F
Regression	$SSR = \mathbf{y}'(\mathbf{H} - \frac{1}{n}\mathbf{J})\mathbf{y}$	$p - 1$	$SSR/(p-1)$	MSR/MSE
Error	$SSE = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}$	$n - p$	$SSE/(n-p)$	
Total	$\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{J}\mathbf{y}/n$	$n - 1$		

Equivalently, we can express

$$\mathbf{Y} = \hat{\mathbf{X}} + (\mathbf{Y} - \hat{\mathbf{X}})$$

where

- $\hat{\mathbf{Y}} = \hat{\mathbf{X}}$ = sum of a vector of fitted values
- $\mathbf{e} = (\mathbf{Y} - \hat{\mathbf{X}})$ = residual
- \mathbf{Y} is the $n \times 1$ vector in a n -dimensional space R^n
- $\hat{\mathbf{X}}$ is an $n \times p$ full rank matrix. and its columns generate a p -dimensional subspace of R^n . Hence, any estimator $\hat{\mathbf{X}}$ is also in this subspace.

We choose least squares estimator that minimize the distance between \mathbf{Y} and $\hat{\mathbf{X}}$, which is the **orthogonal projection** of \mathbf{Y} onto $\hat{\mathbf{X}}$.

$$\begin{aligned}
\|\mathbf{Y} - \hat{\mathbf{X}}\|^2 &= \|\mathbf{Y} - \hat{\mathbf{X}}\|^2 + \|\hat{\mathbf{X}}\|^2 \\
&= (\mathbf{Y} - \hat{\mathbf{X}})'(\mathbf{Y} - \hat{\mathbf{X}}) + (\hat{\mathbf{X}})'(\hat{\mathbf{X}}) \\
&= (\mathbf{Y} - \hat{\mathbf{X}})'(\mathbf{Y} - \hat{\mathbf{X}}) + \hat{\mathbf{X}}'(\mathbf{Y} - \hat{\mathbf{X}}) \\
&= (\mathbf{Y} - \hat{\mathbf{X}})'(\mathbf{Y} - \hat{\mathbf{X}}) + \hat{\mathbf{X}}'(\mathbf{Y} - \hat{\mathbf{X}}) \\
&= \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{X}}'\mathbf{Y} + \hat{\mathbf{X}}'\hat{\mathbf{X}}
\end{aligned}$$

where the norm of a $(p \times 1)$ vector \mathbf{a} is defined by:

$$\|\mathbf{a}\| = \sqrt{\mathbf{a}'\mathbf{a}} = \sqrt{\sum_{i=1}^p a_i^2}$$

Coefficient of Multiple Determination

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

Adjusted Coefficient of Multiple Determination

$$R_a^2 = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)} = 1 - \frac{(n-1)SSE}{(n-p)SSTO}$$

Sequential and Partial Sums of Squares:

In a regression model with coefficients $\beta = (\beta_0, \beta_1, \dots, \beta_{p-1})'$, we denote the uncorrected and corrected SS by

$$SSM = SS(\beta_0, \beta_1, \dots, \beta_{p-1}) SSM_m = SS(\beta_0, \beta_1, \dots, \beta_{p-1} | \beta_0)$$

There are 2 decompositions of SSM_m :

- **Sequential SS:** (not unique - depends on order, also referred to as Type I SS, and is the default of `anova()` in R)

$$SSM_m = SS(\beta_1 | \beta_0) + SS(\beta_2 | \beta_0, \beta_1) + \dots + SS(\beta_{p-1} | \beta_0, \dots, \beta_{p-2})$$

- **Partial SS:** (use more in practice - contribution of each given all of the others)

$$SSM_m = SS(\beta_1 | \beta_0, \beta_2, \dots, \beta_{p-1}) + \dots + SS(\beta_{p-1} | \beta_0, \beta_1, \dots, \beta_{p-2})$$

5.1.3 OLS Assumptions

- A1 Linearity
- A2 Full rank
- A3 Exogeneity of Independent Variables
- A4 Homoskedasticity
- A5 Data Generation (random Sampling)
- A6 Normal Distribution

5.1.3.1 A1 Linearity

$$A1 : y = \mathbf{x}\beta + \epsilon \quad (5.1)$$

Not restrictive

- x can be nonlinear transformation including interactions, natural log, quadratic

With A3 (Exogeneity of Independent), linearity can be restrictive

5.1.3.1.1 Log Model

Model	Form	Interpretation of β	
			In words
Level- Level	$y = \beta_0 + \beta_1 x + \epsilon$	$\Delta y = \beta_1 \Delta x$	A unit change in x will result in β_1 unit change in y
Log- Level	$\ln(y) = \beta_0 + \beta_1 x + \text{epsilon}$	$\% \Delta y = 100 \beta_1 \Delta x$	A unit change in x result in $100 \beta_1$ % change in y
Level- Log	$y = \beta_0 + \beta_1 \ln(x) + \epsilon$	$\Delta y = (\beta_1 / 100) \% \Delta x$	One percent change in x result in $\beta_1 / 100$ units change in y
Log- Log	$\ln(y) = \beta_0 + \beta_1 \ln(x) + \epsilon$	$\% \Delta y = \beta_1 \% \Delta x$	One percent change in x result in β_1 percent change in y

5.1.3.1.2 Higher Orders $y = \beta_0 + x_1\beta_1 + x_1^2\beta_2 + \epsilon$

$$\frac{\partial y}{\partial x_1} = \beta_1 + 2x_1\beta_2$$

- The effect of x_1 on y depends on the level of x_1
- The partial effect at the average = $\beta_1 + 2E(x_1)\beta_2$
- Average Partial Effect = $E(\beta_1 + 2x_1\beta_2)$

5.1.3.1.3 Interactions $y = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_1x_2\beta_3 + \epsilon$

- β_1 is the average effect on y for a unit change in x_1 when $x_2 = 0$
- $\beta_1 + x_2\beta_3$ is the partial effect of x_1 on y which depends on the level of x_2

5.1.3.2 A2 Full rank

$$A2 : \text{rank}(E(x'x)) = k \quad (5.2)$$

also known as **identification condition**

- columns of \mathbf{x} cannot be written as a linear function of the other columns
- which ensures that each parameter is unique and exists in the population regression equation

5.1.3.3 A3 Exogeneity of Independent Variables

$$A3 : E[\epsilon|x_1, x_2, \dots, x_k] = E[\epsilon|\mathbf{x}] = 0 \quad (5.3)$$

strict exogeneity

- also known as **mean independence** check back on Correlation and Independence
- by the Law of Iterated Expectations $E(\epsilon) = 0$, which can be satisfied by always including an intercept.
- independent variables do not carry information for prediction of ϵ
- A3 implies $E(y|x) = x\beta$, which means the conditional mean function must be a linear function of \mathbf{x} A1 Linearity

5.1.3.3.1 A3a Weaker Exogeneity Assumption

Exogeneity of Independent variables

$$\text{A3a: } E(\mathbf{x}'_i \epsilon_i) = 0$$

- x_i is **uncorrelated** with ϵ_i Correlation and Independence
- Weaker than **mean independence** A3
 - A3 implies A3a, not the reverse
 - No causality interpretations
 - Cannot test the difference

5.1.3.4 A4 Homoskedasticity

$$A4 : \text{Var}(\epsilon|x) = \text{Var}(\epsilon) = \sigma^2 \quad (5.4)$$

- Variation in the disturbance to be the same over the independent variables

5.1.3.5 A5 Data Generation (random Sampling)

$$A5 : y_i, x_{i1}, \dots, x_{ik-1} : i = 1, \dots, n \quad (5.5)$$

is a random sample

- random sample mean samples are independent and identically distributed (iid) from a joint distribution of (y, \mathbf{x})
- with A3 and A4, we have
 - **Strict Exogeneity:** $E(\epsilon_i | x_1, \dots, x_n) = 0$. independent variables do not carry information for prediction of ϵ
 - **Non-autocorrelation:** $E(\epsilon_i \epsilon_j | x_1, \dots, x_n) = 0$ The error term is uncorrelated across the draws conditional on the independent variables
 $\rightarrow A4 : \text{Var}(\epsilon | \mathbf{X}) = \text{Var}(\epsilon) = \sigma^2 I_n$
- In times series and spatial settings, A5 is less likely to hold.

5.1.3.5.1 A5a A stochastic process $\{x_t\}_{t=1}^T$ is **stationary** if for every collection fo time indices $\{t_1, t_2, \dots, t_m\}$, the joint distribution of

$$x_{t_1}, x_{t_2}, \dots, x_{t_m}$$

is the same as the joint distribution of

$$x_{t_1+h}, x_{t_2+h}, \dots, x_{t_m+h}$$

for any $h \geq 1$

- The joint distribution for the first ten observation is the same for the next ten, etc.
- Independent draws automatically satisfies this

A stochastic process $\{x_t\}_{t=1}^T$ is **weakly stationary** if x_t and x_{t+h} are “almost independent” as h increases without bounds.

- two observation that are very far apart should be “almost independent”

Common Weakly Dependent Processes

1. Moving Average process of order 1 (MA(1))

MA(1) means that there is only one period lag.

$$y_t = u_t + \alpha_1 u_{t-1} E(y_t) = E(u_t) + \alpha_1 E(u_{t-1}) = 0 \\ Var(y_t) = var(u_t) + \alpha_1 var(u_{t-1}) = \sigma^2 + \alpha_1^2 \sigma^2 = \sigma^2(1 + \alpha_1^2)$$

where u_t is drawn iid over t with variance σ^2

An increase in the absolute value of α_1 increases the variance

When the MA(1) process can be **inverted** ($|\alpha| < 1$) then

$$u_t = y_t - \alpha_1 u_{t-1}$$

called the autoregressive representation (express current observation in term of past observation).

We can expand it to more than 1 lag, then we have MA(q) process

$$y_t = u_t + \alpha_1 u_{t-1} + \dots + \alpha_q u_{t-q}$$

where $u_t \sim WN(0, \sigma^2)$

- Covariance stationary: irrespective of the value of the parameters.
- Invertibility when $\alpha < 1$
- The conditional mean of MA(q) depends on the q lags (long-term memory).
- In MA(q), all autorcorrelations beyond q are 0.

$$\begin{aligned} Cov(y_t, y_{t-1}) &= Cov(u_t + \alpha_1 u_{t-1}, u_{t-1} + \alpha_1 u_{t-2}) \\ &= \alpha_1 var(u_{t-1}) \\ &= \alpha_1 \sigma^2 \end{aligned}$$

$$\begin{aligned} Cov(y_t, y_{t-2}) &= Cov(u_t + \alpha_1 u_{t-1}, u_{t-2} + \alpha_1 u_{t-3}) \\ &= 0 \end{aligned}$$

An MA models a linear relationship between the dependent variable and the current and past values of a stochastic term.

2. Auto regressive process of order 1 (AR(1))

$$y_t = \rho y_{t-1} + u_t, |\rho| < 1$$

where u_t is drawn iid over t with variance σ^2

$$Cov(y_t, y_{t-1}) = Cov(\rho y_{t-1} + u_t - t, y_{t-1})$$

$$= \rho Var(y_{t-1})$$

$$= \rho \frac{\sigma^2}{1 - \rho^2}$$

$$Cov(y_t, y_{t-h}) = \rho^h \frac{\sigma^2}{1 - \rho^2}$$

Stationarity: in the continuum of t, the distribution of each t is the same

$$E(y_t) = E(y_{t-1}) = \dots = E(y_0) = \rho y_0 + u_1$$

where the initial observation $y_0 = 0$

Assume $E(y_t) = 0$

$$y_t = \rho^t y_{t-t} + \rho^{t-1} u_1 + \rho^{t-2} u_2 + \dots + \rho u_{t-1} + u_t = \rho^t y_0 + \rho^{t-1} u_1 + \rho^{t-2} u_2 + \dots + \rho u_{t-1} + u_t$$

Hence, y_t is the weighted of all of the u_t time observations before. y will be correlated with all the previous observations as well as future observations.

$$Var(y_t) = Var(\rho y_{t-1} + u_t) = \rho^2 Var(y_{t-1}) + Var(u_t) + 2\rho Cov(y_{t-1}, u_t) = \rho^2 Var(y_{t-1}) + \sigma^2$$

Hence,

$$Var(y_t) = \frac{\sigma^2}{1 - \rho^2}$$

to have Variance constantly over time, then $\rho \neq 1$ or -1 .

Then stationarity requires $\rho \neq 1$ or -1 . weakly dependent process $|\rho| < 1$

To estimate the AR(1) process, we use **Yule-Walker Equation**

$$y_t = \epsilon_t + \phi y_{t-1} y_{t-\tau} = \epsilon_t y_{t-\tau} + \phi y_{t-1} y_{t-\tau}$$

For $\tau \geq 1$, we have

$$\gamma\tau = \phi\gamma(\tau - 1)\rho_t = \phi^t$$

when you generalize to pth order autoregressive process, AR(p):

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \epsilon_t$$

AR(p) process is **covariance stationary**, and decay in autocorrelations.

When we combine MA(q) and AR(p), we have ARMA(p,q) process, where you can see seasonality. For example, ARMA(1,1)

$$y_t = \phi y_{t-1} + \epsilon_t + \alpha \epsilon_{t-1}$$

Random Walk process

$$y_t = y_0 + \sum_{s=1}^t u_s$$

- not stationary : when $y_0 = 0$ then $E(y_t) = 0$, but $Var(y_t) = t\sigma^2$. Further along in the spectrum, the variance will be larger
- not weakly dependent: $Cov(\sum_{s=1}^t u_s, \sum_{s=1}^{t-h} u_s) = (t-h)\sigma^2$. So the covariance (fixed) is not diminishing as h increases

Assumption A5a: $\{y_t, x_{t1}, \dots, x_{tk-1}\}$

where $t = 1, \dots, T$ are **stationary and weakly dependent processes**.

Alternative Weak Law, Central Limit Theorem

If z_t is a weakly dependent stationary process with a finite first absolute moment and $E(z_t) = \mu$, then

$$T^{-1} \sum_{t=1}^T z_t \xrightarrow{p} \mu$$

If additional regulatory conditions hold (Greene, 1990), then

$$\sqrt{T}(\bar{z} - \mu) \xrightarrow{d} N(0, B)$$

where $B = Var(z_t) + 2 \sum_{h=1}^{\infty} Cov(z_t, z_{t-h})$

5.1.3.6 A6 Normal Distribution

$$A6 : \epsilon | \mathbf{x} \sim N(0, \sigma^2 I_n) \quad (5.6)$$

The error term is normally distributed

From A1-A3, we have **identification** (also known as **Orthogonality Condition**) of the population parameter β

$$\begin{aligned} y &= x\beta + \epsilon && \text{A1} \\ x'y &= x'x\beta + x'\epsilon \\ E(x'y) &= E(x'x)\beta + E(x'\epsilon) \\ E(x'y) &= E(x'x)\beta && \text{A3} \\ [E(x'x)]^{-1}E(x'y) &= [E(x'x)]^{-1}E(x'x)\beta && \text{A2} \\ [E(x'x)]^{-1}E(x'y) &= \beta \end{aligned}$$

β is the row vector of parameters that produces the best predictor of y we choose the min of γ :

$$\underset{\gamma}{\operatorname{argmin}} E((y - x\gamma)^2)$$

First Order Condition

$$\begin{aligned} \frac{\partial((y - x\gamma)^2)}{\partial\gamma} &= 0 \\ -2E(x'(y - x\gamma)) &= 0 \\ E(x'y) - E(x'x\gamma) &= 0 \\ E(x'y) &= E(x'x)\gamma \\ (E(x'x))^{-1}E(x'y) &= \gamma \end{aligned}$$

Second Order Condition

$$\begin{aligned} \frac{\partial^2 E((y - x\gamma)^2)}{\partial\gamma\partial\gamma'} &= 0 \\ E\left(\frac{\partial(y - x\gamma)^2}{\partial\gamma\partial\gamma'}\right) &= 2E(x'x) \end{aligned}$$

If A3 holds, then $2E(x'x)$ is PSD \rightarrow minimum

5.1.4 Theorems

5.1.4.1 Frisch-Waugh-Lovell Theorem

$$\mathbf{y} = \mathbf{X} + = \mathbf{X}_1 \mathbf{x}_1 + \mathbf{X}_2 \mathbf{x}_2 +$$

Equivalently,

$$\begin{pmatrix} X'_1 X_1 & X'_1 X_2 \\ X'_2 X_1 & X'_2 X_2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} X'_1 y \\ X'_2 y \end{pmatrix}$$

Hence,

$$\hat{\mathbf{x}}_1 = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{y} - (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{X}_2 \hat{\mathbf{x}}_2$$

1. Betas from the multiple regression are not the same as the betas from each of the individual simple regression
2. Different set of \mathbf{X} will affect all the coefficient estimates.
3. If $X'_1 X_2 = 0$ or $\hat{\beta}_2 = 0$, then 1 and 2 do not hold.

5.1.4.2 Gauss-Markov Theorem

For a linear regression model

$$\mathbf{y} = \mathbf{X} +$$

Under A1, A2, A3, A4, OLS estimator defined as

$$\hat{\beta} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}$$

is the minimum variance linear (in \mathbf{y}) unbiased estimator of β

Let $\tilde{\beta} = \mathbf{C} \mathbf{y}$, be another linear estimator where \mathbf{C} is $k \times n$ and only function of \mathbf{X} , then for it be unbiased,

$$\begin{aligned} E(\tilde{\beta} | \mathbf{X}) &= E(\mathbf{C} \mathbf{y} | \mathbf{X}) \\ &= E(\mathbf{C} \mathbf{X} + \mathbf{C} \epsilon | \mathbf{X}) \\ &= \mathbf{C} \mathbf{X} \end{aligned}$$

which equals the true parameter β only if $\mathbf{C} \mathbf{X} = \mathbf{I}$

Equivalently, $\tilde{\beta} = \beta + \mathbf{C} \epsilon$ and the variance of the estimator is $Var(\tilde{\beta} | \mathbf{X}) = \sigma^2 \mathbf{C} \mathbf{C}'$

To show minimum variance,

$$\begin{aligned}
 &= \sigma^2(\mathbf{C} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{C} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X})' \\
 &= \sigma^2(\mathbf{C}\mathbf{C}' - \mathbf{C}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) \\
 &= \sigma^2(\mathbf{C}\mathbf{C}' - (\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}) \\
 &= \sigma^2\mathbf{C}\mathbf{C}' - \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \\
 &= Var(\tilde{\beta}|\mathbf{X}) - Var(\hat{\beta}|\mathbf{X})
 \end{aligned}$$

Hierarchy of OLS Assumptions

			Classical LM (BUE)
Identification	Unbiasedness	Gauss-Markov (BLUE)	Small-sample
Data Description	Consistency	Asymptotic Inference (z and Chi-squared)	Inference (t and F)
Variation in X	Variation in X	Variation in X	Variation in X
	Random Sampling	Random Sampling	Random Sampling
	Linearity in Parameters	Linearity in Parameters	Linearity in Parameters
	Zero Conditional Mean	Zero Conditional Mean	Zero Conditional Mean
	Heteroskedasticity	Heteroskedasticity	Normality of Errors

5.1.5 Variable Selection

depends on

- Objectives or goals
- Previously acquired expertise
- availability of data

- availability of computer software

Let P - 1 be the number of possible X variables

5.1.5.1 Mallows's C_p Statistic

(Mallows, 1973, Technometrics, 15, 661-675)

A measure of the predictive ability of a fitted model

Let \hat{Y}_{ip} be the predicted value of Y_i using the model with p parameters. The total standardized mean square error of prediction is:

$$\Gamma_p = \frac{\sum_{i=1}^n E(\hat{Y}_{ip} - E(Y_i))^2}{\sigma^2} = \frac{\sum_{i=1}^n [E(\hat{Y}_{ip}) - E(Y_i)]^2 + \sum_{i=1}^n var(\hat{Y}_{ip})}{\sigma^2}$$

the first term in the numerator is the (bias) $\hat{2}$ term and the 2nd term is the prediction variance term.

- bias term decreases as more variables are added to the model.
- if we assume the full model ($p=P$) is the true model, then $E(\hat{Y}_{ip}) - E(Y_i) = 0$ and the bias is 0.
- Prediction variance increase as more variables are added to the model $\sum var(\hat{Y}_{ip}) = p\sigma^2$
- thus, a tradeoff between bias and variance terms is achieved by minimizing Γ_p .
- Since Γ_p is unknown (due to β). we use an estimate: $C_p = \frac{SSE_p}{\hat{\sigma}^2} - (n-2p)$ which is an unbiased estimate of Γ_p
- As more variables are added to the model, the SSE_p decreases but $2p$ increases. where $\hat{\sigma}^2 = MSE(X_1, \dots, X_{P-1})$ the MSE with all possible X variables in the model.
- when there is no bias then $E(C_p) \approx p$. Thus, good models have C_p close to p.
- Prediction: consider models with $C_p \leq p$
- Parameter estimation: consider models with $C_p \leq 2p - (P - 1)$. Fewer variables should be eliminated from the model to avoid excess bias in the estimates.

5.1.5.2 Akaike Information Criterion (AIC)

$$AUC = n \ln\left(\frac{SSE_p}{n}\right) + 2p$$

- increasing p (number of parameters) leads first-term decreases, and second-term increases.
- We want model with small values of AIC. If the AIC increases when a parameter is added to the model, that parameter is not needed.
- AIC represents a tradeoff between precision of fit against the number of parameters used.

5.1.5.3 Bayes (or Schwarz) Information Criterion

$$BIC = n \ln\left(\frac{SSE_p}{n}\right) + (\ln n)p$$

The coefficient in front of p tends to penalize more heavily models with a larger number of parameters (as compared to AIC).

5.1.5.4 Prediction Error Sum of Squares (PRESS)

$$PRESS_p = \sum_{i=1}^n (Y_i - \hat{Y}_{i(i)})^2$$

where $\hat{Y}_{i(i)}$ is the prediction of the i-th response when the i-th observation is not used, obtained for the model with p parameters.

- evaluates the predictive ability of a postulated model by omitting one observation at a time.
- we want small $PRESS_p$ values
- It can be computationally intensive when you have large p.

5.1.5.5 Best Subsets Algorithm

- “leap and bounds” algorithm of (Furnival and Wilson, 1974) combines comparison of SSE for different subset models with control over the sequence in which the subset regression are computed.

- Guarantees finding the best m subset regressions within each subset size with less computational burden than all possible subsets.

```
library("leaps")
regsubsets()
```

5.1.5.6 Stepwise Selection Procedures

The **forward stepwise** procedure:

- finds a plausible subset sequentially.
- at each step, a variable is added or deleted.
- criterion for adding or deleting is based on SSE, R^2 , T, or F-statistic.

Note:

- Instead of using exact F-values, computer packages usually specify the equivalent “significance” level. For example, SLE is the “significance” level to enter, and SLS is the “significance” level to stay. The SLE and SLS are guides rather than true tests of significance.
- The choice of SLE and SLS represents a balancing of opposing tendencies. Use of large SLE values tends to result in too many predictor variables; models with small SLE tend to be under-specified resulting in σ^2 being badly overestimated.
- As for choice of SLE, can choose between 0.05 and 0.5.
- If $SLE > SLS$ then a cycling pattern may occur. Although most computer packages can detect can stop when it happens. A quick fix: $SLS = SLE /2$ (Bendel and Afifi, 1977).
- If $SLE < SLS$ then the procedure is conservative and may lead variables with low contribution to be retained.
- Order of variable entry does not matter.

Automated Selection Procedures:

- Forward selection: Same idea as forward stepwise except it doesn’t test if variables should be dropped once enter. (not as good as forward stepwise).
- Backward Elimination: begin with all variables and identifies the one with the smallest F-value to be dropped.

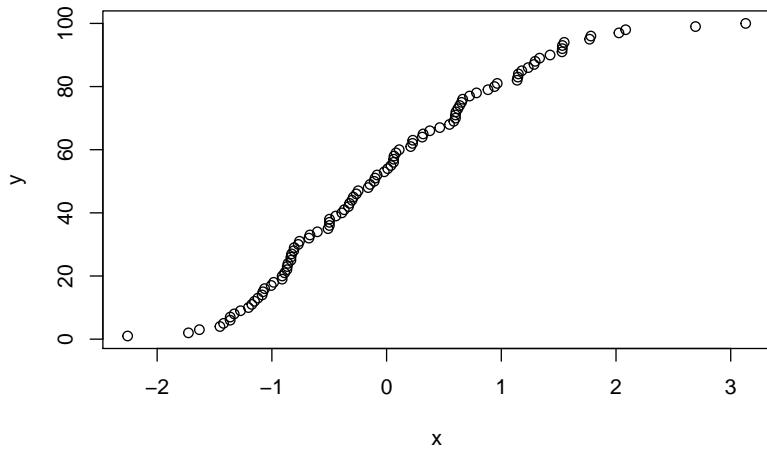
5.1.6 Diagnostics

5.1.6.1 Normality of errors

could use Methods based on normal probability plot or Methods based on empirical cumulative distribution function

or plots such as

```
y = 1:100
x = rnorm(100)
qqplot(x,y)
```



5.1.6.2 Influential observations/outliers

5.1.6.2.1 Hat matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

where $\hat{\mathbf{Y}} = \mathbf{HY}$, $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$ and $var(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$

- $\sigma^2(e_i) = \sigma^2(1 - h_{ii})$, where h_{ii} is the i-th element of the main diagonal of \mathbf{H} (must be between 0 and 1).
- $\sum_{i=1}^n h_{ii} = p$

- $\text{cov}(e_i, e_j) = -h_{ii}\sigma^2$ where $i \neq j$
- Estimate: $s^2(e_i) = \text{MSE}(1 - h_{ii})$
- Estimate: $c\hat{\text{ov}}(e_i, e_j) = -h_{ij}(\text{MSE})$; if model assumption are correct, this covariance is very small for large data sets.
- If $\mathbf{x}_i = [1 X_{i,1} \dots X_{i,p-1}]'$ (the vector of X-values for a given response), then $h_{ii} = \mathbf{x}'_i (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i$ (depends on relative positions of the design points $X_{i,1}, \dots, X_{i,p-1}$)

5.1.6.2.2 Studentized Residuals

$$r_i = \frac{e_i}{s(e_i)} r_i \sim N(0, 1)$$

where $s(e_i) = \sqrt{\text{MSE}(1 - h_{ii})}$. r_i is called the studentized residual or standardized residual.

- you can use the semi-studentized residual before, $e_i^* = e_i \sqrt{\text{MSE}}$. This doesn't take into account the different variances for each e_i .

We would want to see the model without a particular value. You delete the i-th case, fit the regression to the remaining n-1 cases, get estimated responses for the i-th case, $\hat{Y}_{i(i)}$, and find the difference, called the **deleted residual**:

$$d_i = Y_i - \hat{Y}_{i(i)} = \frac{e_i}{1 - h_{ii}}$$

we don't need to recompute the regression model for each case

As h_{ii} increases, d_i increases.

$$s^2(d_i) = \frac{\text{MSE}_{(i)}}{1 - h_{ii}}$$

where $\text{MSE}_{(i)}$ is the mean square error when the i-th case is omitted.

Let

$$t_i = \frac{d_i}{s(d_i)} = \frac{e_i}{\sqrt{\text{MSE}_{(i)}(1 - h_{ii})}}$$

be the **studentized deleted residual**, which follows a t-distribution with $n-p-1$ df.

$$(n - p)MSE = (n - p - 1)MSE_{(i)} + \frac{e_i^2}{1 - h_{ii}}$$

hence, we do not need to fit regressions for each case and

$$t_i = e_i \left(\frac{n - p - 1}{SSE(1 - h_{ii}) - e_i^2} \right)^{1/2}$$

The outlying Y-observations are those cases whose studentized deleted residuals are large in absolute value. If there are many residuals to consider, a Bonferroni critical value can be can ($t_{1-\alpha/2n; n-p-1}$)

Outlying X Observations

Recall, $0 \leq h_{ii} \leq 1$ and $\sum_{i=1}^n h_{ii} = p$ (the total number of parameters)

A large h_{ii} indicates that the i-th case is distant from the center of all X observations (the **leverage** of the i-th case). That is, a large value suggests that the observation exercises substantial leverage in determining the fitted value \hat{Y}_i

We have $\hat{\mathbf{Y}} = \mathbf{HY}$, a linear combination of Y-values; h_{ii} is the weight of the observation Y_i ; so h_{ii} measures the role of the X values in determining how important Y_i is in affecting the \hat{Y}_i .

Large h_{ii} implies $var(e_i)$ is small, so larger h_{ii} implies that \hat{Y}_i is close to Y_i

- small data sets: $h_{ii} > .5$ suggests “large”.
- large data sets: $h_{ii} > \frac{2p}{n}$ is “large.”

Using the hat matrix to identify extrapolation:

- Let \mathbf{x}_{new} be a vector containing the X values for which an inference about a mean response or a new observation is to be made.
- Let \mathbf{X} be the data design matrix used to fit the data. Then, if $h_{new,new} = \mathbf{x}_{\text{new}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{\text{new}}$ is within the range of leverage values (h_{ii}) for cases in the data set, no extrapolation is involved; otherwise; extrapolation is indicated.

Identifying Influential Cases:

by influential we mean that exclusion of an observation causes major changes int he fitted regression. (not all outliers are influential)

- Influence on Single Fitted Values: DFFITS

- Influence on All Fitted Values: Cook's D
- Influence on the Regression Coefficients: DFBETAS

5.1.6.2.3 DFFITS Influence on Single Fitted Values: DFFITS

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}} = t_i \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2}$$

- the standardized difference between the i-th fitted value with all observations and with the i-th case removed.
- studentized deleted residual multiplied by a factor that is a function of the i-th leverage value.
- influence if:
 - small to medium data sets: $|DFFITS| > 1$
 - large data sets: $|DFFITS| > 2\sqrt{p/n}$

5.1.6.2.4 Cook's D Influence on All Fitted Values: Cook's D

$$D_i = \frac{\sum_{j=1}^n (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p(MSE)} = \frac{e_i^2}{p(MSE)} \left(\frac{h_{ii}}{(1 - h_{ii})^2} \right)$$

gives the influence of i-th case on all fitted values.

If e_i increases or h_{ii} increases, then D_i increases.

D_i is a percentile of an $F_{(p,n-p)}$ distribution. If the percentile is greater than .5(50%) then the i-th case has major influence. In practice, if $D_i > 4/n$, then the i-th case has major influence.

5.1.6.2.5 DFBETAS Influence on the Regression Coefficients: DFBETAS

$$(DFBETAS)_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)} c_{kk}}}$$

for $k = 0, \dots, p - 1$ and c_{kk} si the k-th diagonal element of $\mathbf{X}'\mathbf{X}^{-1}$

Influence of the i -th case on each regression coefficient b_k ($k=0,\dots,p-1$) is the difference between the estimated regression coefficients based on all n cases and the regression coefficients obtained when the i -th case is omitted ($b_{k(i)}$)

- small data sets: $|DFBETA| > 1$
- large data sets: $|DFBETA| > 2\sqrt{n}$
- Sign of DFBETA inculcates whether inclusion of a case leads to an increase or a decrease in estimates of the regression coefficient.

5.1.6.3 Collinearity

Multicollinearity refers to correlation among explanatory variables.

- large changes in the estimated regression coefficient when a predictor variable is added or deleted, or when an observation is altered or deleted.
- noninsignificant results in individual tests on regression coefficients for important predictor variables.
- estimated regression coefficients with an algebraic sign that is the opposite of that expected from theoretical consideration or prior experience.
- large coefficients of simple correlation between pairs of predictor variables in the correlation matrix.
- wide confidence intervals for the regression coefficients representing important predictor variables.

When some of X variables are so highly correlated that the inverse $(X'X)^{-1}$ does not exist or is very computationally unstable.

Correlated Predictor Variables: if some X variables are “perfectly” correlated, the system is undetermined and there are an infinite number of models that fit the data. That is, if $X'X$ is singular, then $(X'X)^{-1}$ doesn’t exist. Then,

- parameters cannot be interpreted ($\mathbf{b} = (X'X)^{-1}X'y$)
- sampling variability is infinite ($s^2(\mathbf{b}) = \text{MSE}(X'X)^{-1}$)

5.1.6.3.1 VIFs Let R_k^2 be the coefficient of multiple determination when X_k is regressed on the $p - 2$ other X variables in the model. Then,

$$VIF_k = \frac{1}{1 - R_k^2}$$

- large values indicate that a near collinearity is causing the variance of b_k to be inflated, $var(b_k) \propto \sigma^2(VIF_k)$
- Typically, the rule of thumb is that $VIF > 4$ mean you should see why this is the case, and $VIF_k > 10$ indicates a serious problem collinearity problem that could result in poor parameters estimates.
- the mean of all VIF's provide an estimate of the ratio of the true multicollinearity to a model where the X variables are uncorrelated
- serious multicollinearity if $avg(VIF) >> 1$

5.1.6.3.2 Condition Number Condition Number

spectral decomposition

$$\mathbf{X}'\mathbf{X} = \sum_{i=1}^p \lambda_i \mathbf{u}_i \mathbf{u}_i'$$

where λ_i is the eigenvalue and \mathbf{u}_i is the eigenvector. $\lambda_1 > \dots > \lambda_p$ and the eigenvecotrs are orthogonal:

$$\begin{cases} \mathbf{u}_i' \mathbf{u}_j = 0 & \text{for } i \neq j \\ 1 & \text{for } i = j \end{cases}$$

The condition number is then

$$k = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

- values $k > 30$ are cause for concern
- values $30 < k < 100$ imply moderate dependencies.
- values $k > 100$ imply strong collinearity

Condition index

$$\delta_i = \sqrt{\frac{\lambda_{max}}{\lambda_i}}$$

where $i = 1, \dots, p$

we can find the proportion of the total variance associated with the k-th regression coefficient and the i-th eigen mode:

$$\frac{u_{ik}^2/\lambda_i}{\sum_j(u_{jk}^2/\lambda_j)}$$

These variance proportions can be helpful for identifying serious collinearity

- the condition index must be large
- the variance proportions must be large ($>,5$) for at least two regression coefficients.

5.1.6.4 Constancy of Error Variance

5.1.6.4.1 Brown-Forsythe Test (Modified Levene Test)

- Does not depend on normality
- Applicable when error variance increases or decreases with X
- relatively large sample size needed (so we can ignore dependency between residuals)
- Split residuals into 2 groups ($e_{i1}, i = 1, \dots, n_1; e_{i2}, j = 1, \dots, n_2$)
- Let $d_{i1} = |e_{i1} - \tilde{e}_1|$ where \tilde{e}_1 is the median of group 1.
- Let $d_{j2} = |e_{j2} - \tilde{e}_2|$.
- Then, a 2-sample t-test:

$$t_L = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{1/n_1 + 1/n_2}}$$

where

$$s^2 = \frac{\sum_i(d_{i1} - \bar{d}_1)^2 + \sum_j(d_{j2} - \bar{d}_2)^2}{n - 2}$$

If $|t_L| > t_{1-\alpha/2; n-2}$ conclude the error variance is not constant.

5.1.6.4.2 Breusch-Pagan Test (Cook-Weisberg Test) Assume the error terms are independent and normally distributed, and

$$\sigma_i^2 = \gamma_0 + \gamma_1 X_i$$

Constant error variance corresponds to $\gamma_1 = 0$, i.e., test

- $H_0 : \gamma_1 = 0$
- $H_1 : \gamma_1 \neq 0$

by regressing the squared residuals on X in the usual manner. Obtain the regression sum of squares from this: SSR^* (the SSR from the regression of e_i^2 on X_i). Then, define

$$X_{BP}^2 = \frac{SSR^*/2}{(SSE/n)^2}$$

where SSE is the error sum of squares from the regression of Y on X.

If $H_0 : \gamma_1 = 0$ holds and n is reasonably large, X_{BP}^2 follows approximately the χ^2 distribution with 1 d.f. We reject H_0 (Homogeneous variance) if $X_{BP}^2 > \chi^2_{1-\alpha;1}$

5.1.6.5 Independence

5.1.6.5.1 Plots

5.1.6.5.2 Durbin-Watson

5.1.6.5.3 Time-series

5.1.6.5.4 Spatial Statistics

5.1.7 Model Validation

- split data into 2 groups: training (model building) sample and validation (prediction) sample.

- the model MSE will tend to underestimate the inherent variability in making future predictions. to consider actual predictive ability, consider mean squared prediction error (MSPE):

$$MSPE = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n^*}$$

- where Y_i is the known value of the response variable in the i-th validation case.
- \hat{Y}_i is the predicted value based on a model fit with the training data set.
- n^* is the number of cases in the validation set.
- we want MSPE to be close to MSE (in which MSE is not biased); so look at the ratio MSPE / MSE (closer to 1, the better).

5.1.8 Finite Sample Properties

- n is fixed
- Bias** On average, how close is our estimate to the true value
 - $Bias = E(\hat{\beta}) - \beta$ where β is the true parameter value and $\hat{\beta}$ is the estimator for β
 - An estimator is **unbiased** when
 - * $Bias = E(\hat{\beta}) - \beta = 0$ or $E(\hat{\beta}) = \beta$
 - * means that the estimator will produce estimates that are, on average, equal to the value it is trying to estimate
- Distribution of an estimator:** An estimator is a function of random variables (data)
- Standard Deviation:** the spread of the estimator.

OLS

Under A1 A2 A3, OLS is unbiased

$$\begin{aligned}
E(\hat{\beta}) &= E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) && \text{A2} \\
&= E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X} +)) && \text{A1} \\
&= E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\
&= E(\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\
&= \beta + E((\mathbf{X}'\mathbf{X}^{-1})) \\
&= \beta + E(E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' | \mathbf{X})) && \text{LIE} \\
&= \beta + E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(|\mathbf{X}|)) \\
&= \beta + E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'0)) && \text{A3} \\
&= \beta
\end{aligned}$$

where LIE stands for Law of Iterated Expectation

If A3 does not hold, then OLS will be **biased**

From **Frisch-Waugh-Lovell Theorem**, if we have the omitted variable $\hat{\beta}_2 \neq 0$ and $\mathbf{X}'_1\mathbf{X}_2 \neq 0$, then the omitted variable will cause OLS estimator to be biased.

Under A1 A2 A3 A4, we have the conditional variance of the OLS estimator as follows]

$$\begin{aligned}
Var(\hat{\beta}|\mathbf{X}) &= Var(\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' | \mathbf{X}) && \text{A1-A2} \\
&= Var((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' | \mathbf{X}) \\
&= \mathbf{X}'\mathbf{X}^{-1}\mathbf{X}'Var(\epsilon | \mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \mathbf{X}'\mathbf{X}^{-1}\mathbf{X}'\sigma^2 I\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} && \text{A4} \\
&= \sigma^2 \mathbf{X}'\mathbf{X}^{-1}\mathbf{X}'I\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

Sources of variation

1. $\sigma^2 = Var(\epsilon_i | \mathbf{X})$

- The amount of unexplained variation ϵ_i is large relative to the explained \mathbf{x}_i variation

2. “Small” $Var(x_{i1}), Var(x_{i2}), \dots$

- Not a lot of variation in \mathbf{X} (no information)
- small sample size

3. “Strong” correlation between the explanatory variables

- x_{i1} is highly correlated with a linear combination of 1, x_{i2}, x_{i3}, \dots
- include many irrelevant variables will contribute to this.

- If x_1 is perfectly determined in the regression → **Perfect Collinearity** → A2 is violated.
- If x_1 is highly correlated with a linear combination of other variables, then we have **Multicollinearity**

5.1.8.1 Check for Multicollinearity

Variance Inflation Factor (VIF) Rule of thumb $VIF \geq 10$ is large

$$VIF = \frac{1}{1 - R_1^2}$$

5.1.8.2 Standard Errors

- $Var(\hat{\beta}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ is the variance of the estimate $\hat{\beta}$
- **Standard Errors** are estimators/estimates of the standard deviation (square root of the variance) of the estimator $\hat{\beta}$
- Under A1-A5, then we can estimate $\sigma^2 = Var(\epsilon^2|\mathbf{X})$ the standard errors as

$$s^2 = \frac{1}{n-k} \sum_{i=1}^n e_i^2 = \frac{1}{n-k} SSR$$

- degrees of freedom adjustment: because $e_i \neq \epsilon_i$ and are estimated using k estimates for β , we lose degrees of freedom in our variance estimate.
- $s = \sqrt{s^2}$ is a biased estimator for the standard deviation ([Jensen's Inequality])

Standard Errors for $\hat{\beta}$

$$SE(\hat{\beta}_{j-1}) = s \sqrt{[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}} = \frac{s}{\sqrt{SST_{j-1}(1 - R_{j-1}^2)}}$$

where SST_{j-1} and R_{j-1}^2 from the following regression

x_{j-1} on $1, x_1, \dots, x_{j-2}, x_j, x_{j+1}, \dots, x_{k-1}$

Summary of Finite Sample Properties

- Under A1-A3: OLS is unbiased
- Under A1-A4: The variance of the OLS estimator is $Var(\hat{\beta}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

- Under A1-A4, A6: OLS estimator $\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$
- Under A1-A4, Gauss-Markov Theorem holds \rightarrow OLS is BLUE
- Under A1-A5, the above standard errors are unbiased estimator of standard deviation for $\hat{\beta}$

5.1.9 Large Sample Properties

- let $n \rightarrow \infty$
- A perspective that allows us to evaluate the “quality” of estimators when finite sample properties are not informative, or impossible to compute
- consistency, asymptotic distribution, asymptotic variance

Motivation

- Finite Sample Properties need strong assumption A1 A3 A4 A6
- Other estimation such as GLS, MLE need to be analyzed using Large Sample Properties

Let $\mu(\mathbf{X}) = E(y|\mathbf{X})$ be the **Conditional Expectation Function**

- $\mu(\mathbf{X})$ is the minimum mean squared predictor (over all possible functions)

$$\min E((y - f(\mathbf{X}))^2)$$

under A1 and A3,

$$\mu(\mathbf{X}) = \mathbf{X}\beta$$

Then the **linear projection**

$$L(y|1, \mathbf{X}) = \gamma_0 + \mathbf{X}Var(X)^{-1}Cov(X, Y)$$

where $\mathbf{X}Var(X)^{-1}Cov(X, Y) = \gamma$

is the minimum mean squared linear approximation to be conditional mean function

$$(\gamma_0, \gamma) = argmin E((E(y|\mathbf{X}) - (a + \mathbf{X}\mathbf{b}))^2)$$

- OLS is always **consistent** for the linear projection, but not necessarily unbiased.

- Linear projection has no causal interpretation
- Linear projection does not depend on assumption A1 and A3

Evaluating an estimator using large sample properties:

- Consistency: measure of centrality
- Limiting Distribution: the shape of the scaled estimator as the sample size increases
- Asymptotic variance: spread of the estimator with regards to its limiting distribution.

An estimator $\hat{\theta}$ is consistent for θ if $\hat{\theta}_n \xrightarrow{p} \theta$

- As n increases, the estimator converges to the population parameter value.
- Unbiased does not imply consistency and consistency does not imply unbiased.

Based on Weak Law of Large Numbers

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \left(\sum_{i=1}^n \mathbf{x}'_i \mathbf{x}_i\right)^{-1} \sum_{i=1}^n \mathbf{x}'_i \mathbf{y}_i \\ &= (n^{-1} \sum_{i=1}^n \mathbf{x}'_i \mathbf{x}_i)^{-1} n^{-1} \sum_{i=1}^n \mathbf{x}'_i \mathbf{y}_i\end{aligned}$$

$$\begin{aligned}plim(\hat{\beta}) &= plim((n^{-1} \sum_{i=1}^n \mathbf{x}'_i \mathbf{x}_i)^{-1} n^{-1} \sum_{i=1}^n \mathbf{x}'_i \mathbf{y}_i) \\ &= plim((n^{-1} \sum_{i=1}^n \mathbf{x}'_i \mathbf{x}_i)^{-1}) plim(n^{-1} \sum_{i=1}^n \mathbf{x}'_i \mathbf{y}_i) \\ &= (plim(n^{-1} \sum_{i=1}^n \mathbf{x}'_i \mathbf{x}_i)^{-1}) plim(n^{-1} \sum_{i=1}^n \mathbf{x}'_i \mathbf{y}_i) \text{ due to A2, A5} \\ &= E(\mathbf{x}'_i \mathbf{x}_i)^{-1} E(\mathbf{x}'_i \mathbf{y}_i)\end{aligned}$$

$$E(\mathbf{x}'_i \mathbf{x}_i)^{-1} E(\mathbf{x}'_i \mathbf{y}_i) = \beta + E(\mathbf{x}'_i \mathbf{x}_i)^{-1} E(\mathbf{x}'_i \epsilon_i)$$

Under A1, A2, A3a, A5 OLS is consistent, but not guarantee unbiased.

Under A1, A2, A3a, A5, and $\mathbf{x}'_i \mathbf{x}_i$ has finite first and second moments (CLT),
 $Var(\mathbf{x}'_i \epsilon_i) = \mathbf{B}$

- $(n^{-1} \sum_{i=1}^n \mathbf{x}'_i \mathbf{x}_i)^{-1} \xrightarrow{p} (E(\mathbf{x}'_i \mathbf{x}_i))^{-1}$
- $\sqrt{n}(n^{-1} \sum_{i=1}^n \mathbf{x}'_i \epsilon_i) \xrightarrow{d} N(0, \mathbf{B})$

$$\sqrt{n}(\hat{\beta} - \beta) = (n^{-1} \sum_{i=1}^n \mathbf{x}'_i \mathbf{x}_i)^{-1} \sqrt{n}(n^{-1} \sum_{i=1}^n \mathbf{x}'_i \epsilon_i) \xrightarrow{d} N(0, \Sigma)$$

where $\Sigma = (E(\mathbf{x}'_i \mathbf{x}_i))^{-1} \mathbf{B} (E(\mathbf{x}'_i \mathbf{x}_i))^{-1}$

- holds under A3a
- Do not need A4 and A6 to apply CLT
 - If A4 does not hold, then $\mathbf{B} = Var(\mathbf{x}'_i \epsilon_i) = \sigma^2 E(x'_i x_i)$ which means $\Sigma = \sigma^2 (E(\mathbf{x}'_i \mathbf{x}_i))^{-1}$, use standard errors

Heteroskedasticity can be from

- Limited dependent variable
- Dependent variables with large/skewed ranges

Solving Asymptotic Variance

$$\begin{aligned} \Sigma &= (E(\mathbf{x}'_i \mathbf{x}_i))^{-1} \mathbf{B} (E(\mathbf{x}'_i \mathbf{x}_i))^{-1} \\ &= (E(\mathbf{x}'_i \mathbf{x}_i))^{-1} Var(\mathbf{x}'_i \epsilon_i) (E(\mathbf{x}'_i \mathbf{x}_i))^{-1} \\ &= (E(\mathbf{x}'_i \mathbf{x}_i))^{-1} E[(\mathbf{x}'_i \epsilon_i - 0)(\mathbf{x}'_i \epsilon_i - 0)] (E(\mathbf{x}'_i \mathbf{x}_i))^{-1} \quad \text{A3a} \\ &= (E(\mathbf{x}'_i \mathbf{x}_i))^{-1} E[E(\frac{2}{i} | \mathbf{x}_i) \mathbf{x}'_i \mathbf{x}_i] (E(\mathbf{x}'_i \mathbf{x}_i))^{-1} \quad \text{LIE} \\ &= (E(\mathbf{x}'_i \mathbf{x}_i))^{-1} \sigma^2 E(\mathbf{x}'_i \mathbf{x}_i) (E(\mathbf{x}'_i \mathbf{x}_i))^{-1} \quad \text{A4} \\ &= \sigma^2 (E(\mathbf{x}'_i \mathbf{x}_i)) \end{aligned}$$

Under A1, A2, A3a, A4, A5:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2 (E(\mathbf{x}'_i \mathbf{x}_i))^{-1})$$

- The Asymptotic variance is approximation for the variance in the scaled random variable for $\sqrt{n}(\hat{\beta} - \beta)$ when n is large.
- use $Avar(\sqrt{n}(\hat{\beta} - \beta))/n$ as an approximation for finite sample variance for large n:

$$Avar(\sqrt{n}(\hat{\beta} - \beta)) \approx Var(\sqrt{n}(\hat{\beta} - \beta)) Avar(\sqrt{n}(\hat{\beta} - \beta))/n \approx Var(\sqrt{n}(\hat{\beta} - \beta))/n = Var(\hat{\beta})$$

- $\text{Avar}(\cdot)$ does not behave the same way as $\text{Var}(\cdot)$

$$\text{Avar}(\sqrt{n}(\hat{\beta} - \beta))/n \neq \text{Avar}(\sqrt{n}(\hat{\beta} - \beta)/\sqrt{n}) \neq \text{Avar}(\hat{\beta})$$

In Finite Sample Properties, we calculate standard errors as an estimate for the conditional standard deviation:

$$SE_{fs}(\hat{\beta}_{j-1}) = \sqrt{\text{Var}(\hat{\beta}_{j-1}|\mathbf{X})} = \sqrt{s^2[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}$$

In Large Sample Properties, we calculate standard errors as an estimate for the square root of asymptotic variance

$$SE_{ls}(\hat{\beta}_{j-1}) = \sqrt{\text{Avar}(\sqrt{n}\hat{\beta}_{j-1})/n} = \sqrt{s^2[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}$$

Hence, the standard error estimator is the same for finite sample and large sample.

- Same estimator, but conceptually estimating two different things.
- Valid under weaker assumptions: the assumptions needed to produce a consistent estimator for the finite sample conditional variance (A1-A5) are stronger than those needed to produce a consistent estimator for the asymptotic variance (A1,A2,A3a,A4,A5)

Suppose that y_1, \dots, y_n are a random sample from some population with mean μ and variance-covariance matrix Σ

- $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ is a consistent estimator for μ
- $S = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})'$ is a consistent estimator for Σ .
- Multivariate Central limit Theorem: Similar to the univariate case, $\sqrt{n}(\bar{y} - \mu) \sim N_p(0, \Sigma)$, when n is large relative to p (e.g., $n \geq 25p$). Equivalently, $\bar{y} \sim N_p(\mu, \Sigma/n)$.
- Wald's Theorem: $n(\bar{y} - \mu)'S^{-1}(\bar{y} - \mu) \sim \chi^2_{(p)}$ when n is large relative to p .

5.2 Feasible Generalized Least Squares

Motivation for a more efficient estimator

- Gauss-Markov Theorem holds under A1-A4

- A4: $\text{Var}(\epsilon|\mathbf{X}) = \sigma^2 I_n$
 - Heteroskedasticity: $\text{Var}(\epsilon_i|\mathbf{X}) \neq \sigma^2 I_n$
 - Serial Correlation: $\text{Cov}(\epsilon_i, \epsilon_j|\mathbf{X}) \neq 0$
- Without A4, how can we know which unbiased estimator is the most efficient?

Original (unweighted) model:

$$\mathbf{y} = \mathbf{X} +$$

Suppose A1-A3 hold, but A4 does not hold,

$$\mathbf{V}\text{ar}(\epsilon|\mathbf{X}) = \neq 2\mathbf{I}_n$$

We will try to use OLS to estimate the transformed (weighted) model

$$\mathbf{w}\mathbf{y} = \mathbf{w}\mathbf{X} + \mathbf{w}$$

We need to choose \mathbf{w} so that

$$\mathbf{w}'\mathbf{w} = -1$$

then \mathbf{w} (full-rank matrix) is the **Cholesky decomposition** of Ω^{-1} (full-rank matrix)

In other words, \mathbf{w} is the squared root of Ω (squared root version in matrix)

$$\Omega = \text{var}(\epsilon|X)\Omega^{-1} = \text{var}(\epsilon|X)^{-1}$$

Then, the transformed equation (IGLS) will have the following properties.

$$\begin{aligned}\hat{\mathbf{y}}_{\text{IGLS}} &= (\mathbf{X}'\mathbf{w}'\mathbf{w}\mathbf{X})^{-1}\mathbf{X}'\mathbf{w}'\mathbf{w}\mathbf{y} \\ &= (\mathbf{X}'^{-1}\mathbf{X})^{-1}\mathbf{X}'^{-1}\mathbf{y} \\ &= +\mathbf{X}'^{-1}\mathbf{X}'^{-1}\mathbf{y}\end{aligned}$$

Since A1-A3 hold for the unweighted model

$$\begin{aligned}\mathbf{E}(\hat{\mathbf{y}}_{\text{IGLS}}|\mathbf{X}) &= E(+\mathbf{X}'^{-1}\mathbf{X}'^{-1}\mathbf{y}|X) \\ &= +\mathbf{E}(\mathbf{X}'^{-1}\mathbf{X}'^{-1}\mathbf{y})|\mathbf{X} \\ &= +\mathbf{X}'^{-1}\mathbf{X}'^{-1}\mathbf{E}(\mathbf{y}|\mathbf{X}) \quad \text{since A3 : } E(\epsilon|X) = 0 \\ &= \end{aligned}$$

→ IGLS estimator is unbiased

$$\begin{aligned}
 \mathbf{Var}(\mathbf{w} | \mathbf{X}) &= \mathbf{w} \mathbf{Var}(\mathbf{|X|}) \mathbf{w}' \\
 &= \mathbf{w} \mathbf{w}' \\
 &= \mathbf{w} (\mathbf{w}' \mathbf{w})^{-1} \mathbf{w}' \quad \text{since } \mathbf{w} \text{ is a full-rank matrix} \\
 &= \mathbf{w} \mathbf{w}^{-1} (\mathbf{w}')^{-1} \mathbf{w}' \\
 &= \mathbf{I}_n
 \end{aligned}$$

→ A4 holds for the transformed (weighted) equation

Then, the variance for the estimator is

$$\begin{aligned}
 Var(\hat{\beta}_{IGLS} | \mathbf{X}) &= \mathbf{Var}(\hat{\beta} + (\mathbf{X}'^{-1} \mathbf{X})^{-1} \mathbf{X}'^{-1} | \mathbf{X}) \\
 &= \mathbf{Var}((\mathbf{X}'^{-1} \mathbf{X})^{-1} \mathbf{X}'^{-1} | \mathbf{X}) \\
 &= (\mathbf{X}'^{-1} \mathbf{X})^{-1} \mathbf{X}'^{-1} \mathbf{Var}(\mathbf{|X|})^{-1} \mathbf{X} (\mathbf{X}'^{-1} \mathbf{X})^{-1} \quad \text{because A4 holds} \\
 &= (\mathbf{X}'^{-1} \mathbf{X})^{-1} \mathbf{X}'^{-1} \mathbf{X}'^{-1} \mathbf{X} (\mathbf{X}'^{-1} \mathbf{X})^{-1} \\
 &= (\mathbf{X}'^{-1} \mathbf{X})^{-1}
 \end{aligned}$$

Let $A = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' - (\mathbf{X}'^{-1} \mathbf{X}) \mathbf{X}'^{-1}$ then

$$Var(\hat{\beta}_{OLS} | X) - Var(\hat{\beta}_{IGLS} | X) = A \Omega A'$$

And Ω is Positive Semi Definite, then $A \Omega A'$ also PSD, then IGLS is more efficient

The name **Infeasible** comes from the fact that it is impossible to compute this estimator.

$$\mathbf{w} = \begin{pmatrix} w_{11} & 0 & 0 & \dots & 0 \\ w_{21} & w_{22} & 0 & \dots & 0 \\ w_{31} & w_{32} & w_{33} & \dots & \dots \\ w_{n1} & w_{n2} & w_{n3} & \dots & w_{nn} \end{pmatrix}$$

With $n(n+1)/2$ number of elements and n observations → infeasible to estimate.
(number of equation > data)

Hence, we need to make assumption on Ω to make it feasible to estimate \mathbf{w} :

1. Heteroskedasticity : multiplicative exponential model
2. AR(1)
3. Cluster

5.2.1 Heteroskedasticity

$$\begin{aligned} \text{Var}(\epsilon_i | x_i) &= E(\epsilon^2 | x_i) \neq \sigma^2 \\ &= h(x_i) = \sigma_i^2 (\text{variance of the error term is a function of } x) \end{aligned} \quad (5.7)$$

For our model,

$$y_i = x_i\beta + \epsilon_i(1/\sigma_i)y_i = (1/\sigma_i)x_i\beta + (1/\sigma_i)\epsilon_i$$

then, from (5.7)

$$\begin{aligned} \text{Var}((1/\sigma_i)\epsilon_i | X) &= (1/\sigma_i^2)\text{Var}(\epsilon_i | X) \\ &= (1/\sigma_i^2)\sigma_i^2 \\ &= 1 \end{aligned}$$

then the weight matrix \mathbf{w} in the matrix equation

$$\mathbf{wy} = \mathbf{wX} + \mathbf{w}$$

$$\mathbf{w} = \begin{pmatrix} 1/\sigma_1 & 0 & 0 & \dots & 0 \\ 0 & 1/\sigma_2 & 0 & \dots & 0 \\ 0 & 0 & 1/\sigma_3 & \dots & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & 1/\sigma_n \end{pmatrix}$$

Infeasible Weighted Least Squares

1. Assume we know σ_i^2 (Infeasible)
2. The IWLS estimator is obtained as the least squared estimated for the following weighted equation

$$(1/\sigma_i)y_i = (1/\sigma_i)\mathbf{x}_i\beta + (1/\sigma_i)\epsilon_i$$

- Usual standard errors for the weighted equation are valid if $\text{Var}(\epsilon | \mathbf{X}) = \sigma_i^2$
- If $\text{Var}(\epsilon | \mathbf{X}) \neq \sigma_i^2$ then heteroskedastic robust standard errors are valid.

Problem: We do not know $\sigma_i^2 = \text{Var}(\epsilon_i | \mathbf{x}_i) = E(\epsilon_i^2 | \mathbf{x}_i)$

- One observation ϵ_i cannot estimate a sample variance estimate σ_i^2

- Model ϵ_i^2 as reasonable (strictly positive) function of x_i and independent error v_i (strictly positive)

$$\epsilon_i^2 = v_i \exp(\mathbf{x}_i)$$

Then we can apply a log transformation to recover a linear in parameters model,

$$\ln(\epsilon_i^2) = \mathbf{x}_i + \ln(v_i)$$

where $\ln(v_i)$ is independent \mathbf{x}_i

We do not observe ϵ_i * OLS residual (e_i) as an approximate

5.2.2 Serial Correlation

$$Cov(\epsilon_i, \epsilon_j | \mathbf{X}) \neq 0$$

Under covariance stationary,

$$Cov(\epsilon_i, \epsilon_j | \mathbf{X}) = Cov(\epsilon_i, \epsilon_{i+h} | \mathbf{x}_i, \mathbf{x}_{i+h}) = \gamma_h$$

And the variance covariance matrix is

$$Var(\epsilon | \mathbf{X}) = \Omega = \begin{pmatrix} \sigma^2 & \gamma_1 & \gamma_2 & \dots & \gamma_{n-1} \\ \gamma_1 & \sigma^2 & \gamma_1 & \dots & \gamma_{n-2} \\ \gamma_2 & \gamma_1 & \sigma^2 & \dots & \dots \\ \vdots & \vdots & \vdots & \ddots & \gamma_1 \\ \gamma_{n-1} & \gamma_{n-2} & \ddots & \gamma_1 & \sigma^2 \end{pmatrix}$$

There n parameters to estimate - need some sort fo structure to reduce number of parameters to estimate.

- Time Series
 - Effect of inflation and deficit on Treasury Bill interest rates
- Cross-sectional
 - Clustering

5.2.2.1 AR(1)

$$y_t = \beta_0 + x_t\beta_1 + \epsilon_t\epsilon_t = \rho\epsilon_{t-1} + u_t$$

and the variance covariance matrix is

$$Var(\epsilon|\mathbf{X}) = \frac{\sigma_u^2}{1-\rho} \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \ddots & \cdot \\ \cdot & \cdot & \cdot & \ddots & \rho \\ \rho^{n-1} & \rho^{n-2} & \cdot & \rho & 1 \end{pmatrix}$$

Hence, there is only 1 parameter to estimate: ρ

- Under A1, A2, A3a, A5a, OLS is consistent and asymptotically normal
- Use Newey West Standard Errors for valid inference.
- Apply Infeasible Cochrane Orcutt (as if we knew ρ)
- Because

$$u_t = \epsilon_t - \rho\epsilon_{t-1}$$

satisfies A3, A4, A5 we'd like to transform the above equation to one that has u_t as the error.

$$\begin{aligned} y_t - \rho y_{t-1} &= (\beta_0 + x\beta_1 + \epsilon_t) - \rho(\beta_0 + x_{t-1}\beta_1 + \epsilon_{t-1}) \\ &= (1 - \rho)\beta_0 + (x_t - \rho x_{t-1})\beta_1 + u_t \end{aligned}$$

5.2.2.1.1 Infeasible Cochrane Orcutt

1. Assume that we know ρ (Infeasible)
2. The ICO estimator is obtained as the least squared estimated for the following weighted first difference equation

$$y_t - \rho y_{t-1} = (1 - \rho)\beta_0 + (x_t - \rho x_{t-1})\beta_1 + u_t$$

- Usual standard errors for the weighted first difference equation are valid if the errors truly follow an AR(1) process
- If the serial correlation is generated from a more complex dynamic process then Newey-West HAC standard errors are valid

Problem We do not know ρ

- ρ is the correlation between ϵ_t and ϵ_{t-1} : estimate using OLS residuals (e_i) as proxy

$$\hat{\rho} = \frac{\sum_{t=1}^T e_t e_{t-1}}{\sum_{t=1}^T e_t^2}$$

which can be obtained from the OLS regression of

$$e_t = \rho e_{t-1} + u_t$$

where we suppress the intercept.

- We are losing an observation
- By taking the first difference we are dropping the first observation

$$y_1 = \beta_0 + x_1 \beta_1 + \epsilon_1$$

- Feasible Prais Winsten Transformation applies the Infeasible Cochrane Orcutt but includes a weighted version of the first observation

$$(\sqrt{1 - \rho^2})y_1 = \beta_0 + (\sqrt{1 - \rho^2})x_1 \beta_1 + (\sqrt{1 - \rho^2})\epsilon_1$$

5.2.2.2 Cluster

$$y_{gi} = \mathbf{x}_{gi}\beta + \epsilon_{gi}$$

$$Cov(\epsilon_{gi}, \epsilon_{hj}) \begin{cases} = 0 & \text{for } g \neq h \text{ and any pair (i,j)} \\ \neq 0 & \text{for any (i,j) pair} \end{cases}$$

Intra-group Correlation

Each individual in a single group may be correlated but independent across groups.

- A4 is violated. usual standard errors for OLS are valid.
- Use **cluster robust standard errors** for OLS.

Suppose there are 3 groups with different n

$$Var(\epsilon|\mathbf{X}) = \Omega = \begin{pmatrix} \sigma^2 & \delta_{12}^1 & \delta_{13}^1 & 0 & 0 & 0 \\ \delta_{12}^1 & \sigma^2 & \delta_{23}^1 & 0 & 0 & 0 \\ \delta_{13}^1 & \delta_{23}^1 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & \delta_{12}^2 & 0 \\ 0 & 0 & 0 & \delta_{12}^2 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2 \end{pmatrix}$$

where $Cov(\epsilon_{gi}, \epsilon_{gj}) = \delta_{ij}^g$ and $Cov(\epsilon_{gi}, \epsilon_{hj}) = 0$ for any i and j

Infeasible Generalized Least Squares (Cluster)

1. Assume that σ^2 and δ_{ij}^g are known, plug into Ω and solve for the inverse Ω^{-1} (infeasible)
2. The Infeasible Generalized Least Squares Estimator is

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

Problem * We do not know σ^2 and δ_{ij}^g + Can make assumptions about data generating process that is causing the clustering behavior. - Will give structure to $Cov(\epsilon_{gi}, \epsilon_{gj}) = \delta_{ij}^g$ which makes it feasible to estimate - if the assumptions are wrong then we should use cluster robust standard errors.

Solution Assume **group level random effects** specification in the error

$$y_{gi} = \mathbf{g}_i\beta + c_g + u_{gi} Var(c_g|\mathbf{x}_i) = \sigma_c^2 Var(u_{gi}|\mathbf{x}_i) = \sigma_u^2$$

where c_g and u_{gi} are independent of each other, and mean independent of \mathbf{x}_i

- c_g captures the common group shocks (independent across groups)
- u_{gi} captures the individual shocks (independent across individuals and groups)

Then the error variance is

$$Var(\epsilon|\mathbf{X}) = \Omega = \begin{pmatrix} \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & \sigma_c^2 & 0 & 0 & 0 \\ \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & 0 & 0 & 0 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & 0 \\ 0 & 0 & 0 & \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_c^2 + \sigma_u^2 \end{pmatrix}$$

Use Feasible group level Random Effects

5.3 Weighted Least Squares

1. Estimate the following equation using OLS

$$y_i = \mathbf{x}_i\beta + \epsilon_i$$

and obtain the residuals $e_i = y_i - \mathbf{x}_i\hat{\beta}$

2. Transform the residual and estimate the following by OLS,

$$\ln(e_i^2) = \mathbf{x}_i\gamma + \ln(v_i)$$

and obtain the predicted values $g_i = \mathbf{x}_i\hat{\gamma}$

3. The weights will be the untransformed predicted outcome,

$$\hat{\sigma}_i = \sqrt{\exp(g_i)}$$

4. The FWLS (Feasible WLS) estimator is obtained as the least squared estimated for the following weighted equation

$$(1/\hat{\sigma}_i)y_i = (1/\hat{\sigma}_i)\mathbf{x}_i\beta + (1/\hat{\sigma}_i)\epsilon_i$$

Properties of the FWLS

- The infeasible WLS estimator is unbiased under A1-A3 for the unweighted equation.
- The FWLS estimator is NOT an unbiased estimator.
- The FWLS estimator is consistent under A1, A2, (for the unweighted equation), A5, and $E(\mathbf{x}'_i\epsilon_i/\sigma_i^2) = 0$
 - A3a is not sufficient for the above equation
 - A3 is sufficient for the above equation.
- The FWLS estimator is asymptotically more efficient than OLS if the errors have multiplicative exponential heteroskedasticity.
 - If the errors are truly multiplicative exponential heteroskedasticity, then usual standard errors are valid
 - If we believe that there may be some mis-specification with the **multiplicative exponential model**, then we should report heteroskedastic robust standard errors.

5.4 Generalized Least Squares

Consider

$$\mathbf{y} = \mathbf{X} +$$

where,

$$var(\epsilon) = \mathbf{G} = \begin{pmatrix} g_{11} & g_{12} & \dots & g_{1n} \\ g_{21} & g_{22} & \dots & g_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ g_{n1} & \ddots & \ddots & g_{nn} \end{pmatrix}$$

The variances are heterogeneous, and the errors are correlated.

$$\hat{\mathbf{b}}_G = (\mathbf{X}' \mathbf{G}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{G}^{-1} \mathbf{Y}$$

if we know \mathbf{G} , we can estimate \mathbf{b} just like OLS. However, we do not know \mathbf{G} . Hence, we model the structure of \mathbf{G} .

5.5 Feasible Prais Winsten

Weighting Matrix

$$\mathbf{w} = \begin{pmatrix} \sqrt{1 - \hat{\rho}^2} & 0 & 0 & \dots & 0 \\ -\hat{\rho} & 1 & 0 & \dots & 0 \\ 0 & -\hat{\rho} & 1 & & \cdot \\ \vdots & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & 0 & -\hat{\rho} & 1 \end{pmatrix}$$

1. Estimate the following equation using OLS

$$y_t = \mathbf{x}_t \beta + \epsilon_t$$

and obtain the residuals $e_t = y_t - \mathbf{x}_t \hat{\beta}$

2. Estimate the correlation coefficient for the AR(1) process by estimating the following by OLS (without no intercept)

$$e_t = \rho e_{t-1} + u_t$$

3. Transform the outcome and independent variables \mathbf{wy} and \mathbf{wX} respectively (weight matrix as stated).
4. The FPW estimator is obtained as the least squared estimated for the following weighted equation

$$\mathbf{wy} = \mathbf{wX} + \mathbf{w}$$

Properties of Feasible Prais Winsten Estimator

- The Infeasible PW estimator is under A1-A3 for the unweighted equation
- The FPW estimator is biased
- The FPW is consistent under A1 A2 A5 and

$$E((\mathbf{x}_t - \mathbf{x}_{t-1})')(\epsilon_t - \rho\epsilon_{t-1}) = 0$$

- A3a is not sufficient for the above equation
- A3 is sufficient for the above equation
- The FPW estimator is asymptotically more efficient than OLS if the errors are truly generated as AR(1) process
 - If the errors are truly generated as AR(1) process then usual standard errors are valid
 - If we are concerned that there may be a more complex dependence structure of heteroskedasticity, then we use Newey West Standard Errors

5.6 Feasible group level Random Effects

1. Estimate the following equation using OLS

$$y_{gi} = \mathbf{x}_{gi}\beta + \epsilon_{gi}$$

and obtain the residuals $e_{gi} = y_{gi} - \mathbf{x}_{gi}\hat{\beta}$ 2. Estimate the variance using the usual s^2 estimator

$$s^2 = \frac{1}{n-k} \sum_{i=1}^n e_i^2$$

as an estimator for $\sigma_c^2 + \sigma_u^2$ and estimate the within group correlation,

$$\hat{\sigma}_c^2 = \frac{1}{G} \sum_{g=1}^G \left(\frac{1}{\sum_{i=1}^{n_g-1} i} \sum_{i \neq j} \sum_j e_{gi} e_{gj} \right)$$

and plug in the estimates to obtain $\hat{\Omega}$

- 3. The feasible group level RE estimator is obtained as

$$\hat{\beta} = (\mathbf{X}^\wedge)^{-1} \mathbf{X}^\wedge \mathbf{y}$$

Properties of the Feasible group level Random Effects Estimator

- The infeasible group RE estimator is a linear estimator and is unbiased under A1-A3 for the unweighted equation
 - A3 requires $E(\epsilon_{gi}|\mathbf{x}_i) = E(c_g|\mathbf{x}_i) + (u_{gi}|\mathbf{x}_i) = 0$ so we generally assume $E(c_g|\mathbf{x}_i) + (u_{gi}|\mathbf{x}_i) = 0$. The assumption $E(c_g|\mathbf{x}_i) = 0$ is generally called **random effects assumption**
- The Feasible group level Random Effects is biased
- The Feasible group level Random Effects is consistent under A1-A3a, and A5a for the unweighted equation.
 - A3a requires $E(\mathbf{x}'_i \epsilon_{gi}) = E(\mathbf{x}'_i c_g) + (\mathbf{x}'_i u_{gi}) = 0$
- The Feasible group level Random Effects estimator is asymptotically more efficient than OLS if the errors follow the random effects specification
 - If the errors do follow the random effects specification than the usual standard errors are consistent
 - If there might be a more complex dependence structure or heteroskedasticity, then we need cluster robust standard errors.

5.7 Ridge Regression

When we have the Collinearity problem, we could use the Ridge regression.

The main problem with multicollinearity is that $\mathbf{X}'\mathbf{X}$ is “ill-conditioned”. The idea for ridge regression: adding a constant to the diagonal of $\mathbf{X}'\mathbf{X}$ improves the conditioning

$$\mathbf{X}'\mathbf{X} + c\mathbf{I} (c > 0)$$

The choice of c is hard. The estimator

$$\mathbf{b}^R = (\mathbf{X}'\mathbf{X} + c\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

is **biased**.

- it has smaller variance than the OLS estimator; as c increases, the bias increases but the variance decreases.
- always exists some value of c for which the ridge regression estimator has a smaller total MSE than the OLS
- the optimal c varies with application and data set.
- to find the “optimal” c we could use “ridge trace”.

we plot the values of the $p - 1$ parameter estimates for different values of c , simultaneously.

- typically, as c increases toward 1 the coefficients decreases to 0.
- the values of the VIF tend to decrease rapidly as c gets bigger than 0. The VIF values begin to change slowly as c approaches 1.
- then we can examine the ridge trace and VIF values and chooses the smallest value of c where the regression coefficients first become stable in the ridge trace and the VIF values have become sufficiently small (which is very subjective).
- typically, this procedure is applied to the standardized regression model.

5.8 Principal Component Regression

This also addresses the problem of multicollinearity

5.9 Robust Regression

- To address the problem of influential cases.
- can be used when a known functional form is to be fitted, and when the errors are not normal due to a few outlying cases.

5.9.1 Least Absolute Residuals (LAR) Regression

also known as minimum L_1 -norm regression.

$$L_1 = \sum_{i=1}^n |Y_i - (\beta_0 + \beta_1 X_{i1} + \dots + \beta_{p-1} X_{ip-1})|$$

which is not sensitive to outliers and inadequacies of the model specification.

5.9.2 Least Median of Squares (LMS) Regression

$$\text{median}\{[Y_i - (\beta_0 - \beta_1 X_{i1} + \dots + \beta_{p-1} X_{ip-1})]^2\}$$

5.9.3 Iteratively Reweighted Least Squares (IRLS) Robust Regression

- uses Weighted Least Squares to lessen the influence of outliers.
- the weights w_i are inversely proportional to how far an outlying case is (e.g., based on the residual)
- the weights are revised iteratively until a robust fit

Process:

Step 1: Choose a weight function for weighting the cases.

Step 2: obtain starting weights for all cases.

Step 3: Use the starting weights in WLS and obtain the residuals from the fitted regression function.

Step 4: use the residuals in Step 3 to obtain revised weights.

Step 5: continue until convergence.

Note:

- If you don't know the form of the regression function, consider using non-parametric regression (e.g., locally weighted regression, regression trees, projection pursuit, neural networks, smoothing splines, loess, wavelets).
- could use to detect outliers or confirm OLS.

5.10 Maximum Likelihood

Premise: find values of the parameters that maximize the probability of observing the data In other words, we try to maximize the value of theta in the likelihood function

$$L(\theta) = \prod_{i=1}^n f(y_i|\theta)$$

$f(y|\theta)$ is the probability density of observing a single value of Y given some value of θ $f(y|\theta)$ can be specified as various type of distributions. You can review back section Distributions. For example If y is a dichotomous variable, then

$$L(\theta) = \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1-y_i}$$

$\hat{\theta}$ is the Maximum Likelihood estimate if $L(\hat{\theta}) > L(\theta_0)$ for all values of θ_0 in the parameter space.

5.10.1 Motivation for MLE

Suppose we know the conditional distribution of y given x:

$$f_{Y|X}(y, x; \theta)$$

where θ is the unknown parameter of distribution. Sometimes we are only concerned with the unconditional distribution $f_Y(y; \theta)$

Then given a sample of iid data, we can calculate the joint distribution of the entire sample,

$$f_{Y_1, \dots, Y_n | X_1, \dots, X_n}(y_1, \dots, y_n, x_1, \dots, x_n; \theta) = \prod_{i=1}^n f_{Y|X}(y_i, x_i; \theta)$$

The joint distribution evaluated at the sample is the likelihood (probability) that we observed this particular sample (depends on θ)

Idea for MLE: Given a sample, we choose our estimates of the parameters that gives the highest likelihood (probability) of observing our particular sample

$$\max_{\theta} \prod_{i=1}^n f_{Y|X}(y_i, x_i; \theta)$$

Equivalently,

$$\max_{\theta} \prod_{i=1}^n \ln(f_{Y|X}(y_i, x_i; \theta))$$

Solving for the Maximum Likelihood Estimator

1. Solve First Order Condition

$$\frac{\partial}{\partial \theta} \sum_{i=1}^n \ln(f_{Y|X}(y_i, x_i; \hat{\theta}_{MLE})) = 0$$

where $\hat{\theta}_{MLE}$ is defined.

2. Evaluate Second Order Condition

$$\frac{\partial^2}{\partial \theta^2} \sum_{i=1}^n \ln(f_{Y|X}(y_i, x_i; \hat{\theta}_{MLE})) < 0$$

where the above condition ensures we can solve for a maximum

Examples:

Unconditional Poisson Distribution: Number of products ordered on Amazon within an hour, number of website visits a day for a political campaign.

Exponential Distribution: Length of time until an earthquake occurs, length of time a car battery lasts.

$$f_{Y|X}(y, x; \theta) = \exp(-y/x\theta)/x\theta f_{Y_1, \dots, Y_n | X_1, \dots, X_n}(y_1, \dots, y_n, x_1, \dots, x_n; \theta) = \prod_{i=1}^n \exp(-y_i/x_i\theta)/x_i\theta$$

5.10.2 Assumption

- **High Level Regulatory Assumptions** is the sufficient condition used to show large sample properties
 - Hence, for each MLE, we will need to either assume or verify if the regulatory assumptions holds.
- observations are independent and have the same density function.

- Under multivariate normal assumption, ML yields consistent estimates of the means and the covariance matrix for multivariate distribution with finite fourth moments (Little and Smith, 1987)

To find the MLE, we usually differentiate the **log-likelihood** function and set it equal to 0.

$$\frac{d}{d\theta} l(\theta) = 0$$

This is the **score** equation

Our confidence in the MLE is quantified by the “pointedness” of the log-likelihood

$$I_O(\theta) = \frac{d^2}{d\theta^2} l(\theta) = 0$$

called the **observed information**

while

$$I(\theta) = E[I_O(\theta; Y)]$$

is the expected information. (also known as Fisher Information). which we base our variance of the estimator.

$$V(\hat{\Theta}) \approx I(\theta)^{-1}$$

Consistency of MLE

Suppose that y_i and x_i are iid drawn from the true conditional pdf $f_{Y|X}(y_i, x_i; \theta_0)$. If the following regulatory assumptions hold,

- R1: If $\theta \neq \theta_0$ then $f_{Y|X}(y_i, x_i; \theta) \neq f_{Y|X}(y_i, x_i; \theta_0)$
- R2: The set Θ that contains the true parameters θ_0 is compact
- R3: The log-likelihood $\ln(f_{Y|X}(y_i, x_i; \theta_0))$ is continuous at each θ with probability 1
- R4: $E(\sup_{\theta \in \Theta} |\ln(f_{Y|X}(y_i, x_i; \theta_0))|)$

then the MLE estimator is consistent,

$$\hat{\theta}_{MLE} \rightarrow^p \theta_0$$

Asymptotic Normality of MLE

Suppose that y_i and x_i are iid drawn from the true conditional pdf $f_{Y|X}(y_i, x_i; \theta)$. If R1-R4 and the following hold

R5: θ_0 is in the interior of the set Θ

R6: $f_{Y|X}(y_i, x_i; \theta)$ is twice continuously differentiable in θ and $f_{Y|X}(y_i, x_i; \theta) > 0$ for a neighborhood $N \in \Theta$ around θ_0

R7: $\int sup_{\theta \in N} ||\partial f_{Y|X}(y_i, x_i; \theta) / \partial \theta|| d(y, x) < \infty$, $\int sup_{\theta \in N} ||\partial^2 f_{Y|X}(y_i, x_i; \theta) / \partial \theta \partial \theta'|| d(y, x) < \infty$ and $E(sup_{\theta \in N} ||\partial^2 ln(f_{Y|X}(y_i, x_i; \theta)) / \partial \theta \partial \theta'||) < \infty$

R8: The information matrix $I(\theta_0) = Var(\partial f_{Y|X}(y, x_i; \theta_0) / \partial \theta)$ exists and is non-singular

then the MLE estimator is asymptotically normal,

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1})$$

5.10.3 Properties

(EJD et al., 1998)

- (1) Consistent: estimates are approximately unbiased in large samples
- (2) Asymptotically efficient: approximately smaller standard errors compared to other estimator
- (3) Asymptotically normal: with repeated sampling, the estimates will have an approximately normal distribution. Suppose that $\hat{\theta}_n$ is the MLE for θ based on n independent observations. then $\hat{\theta}_n \sim N(\theta, H^{-1})$.

- where H is called the Fisher information matrix. It contains the expected values of the second partial derivatives of the log-likelihood function. The (i,j)th element of H is $-E(\frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j})$
- We can estimate H by finding the form determined above, and evaluating it at $\theta = \hat{\theta}_n$

- (4) Invariance: MLE for $g(\theta) = g(\hat{\theta})$ for any function $g(\cdot)$

$$\hat{\Theta} \approx^d (\theta, I(\theta)^{-1})$$

Explicit vs Implicit MLE

- If we solve the score equation to get an expression of MLE, then it's called **explicit**
- If there is no closed form for MLE, and we need some algorithms to derive its expression, it's called **implicit**

Large Sample Property of MLE

Implicit in these theorems is the assumption that we know what the conditional distribution,

$$f_{Y|X}(y_i, x_i; \theta_0)$$

but just do now know the exact parameter value.

- Any Distributional mis-specification will result in inconsistent parameter estimates.
- Quasi-MLE: Particular settings/ assumption that allow for certain types of distributional mis-specification (Ex: as long as the distribution is part of particular class or satisfies a particular assumption, then estimating with a wrong distribution will not lead to inconsistent parameter estimates).
- non-parametric/ Semi-parametric estimation: no or very little distributional assumption are made. (hard to implement, derive properties, and interpret)

The asymptotic variance of the MLE achieves the **Cramer-Rao Lower Bound**

- The **Cramer-Rao Lower Bound** is a lower bound for the asymptotic variance of a consistent and asymptotically normally distributed estimator.
- If an estimator achieves the lower bound then it is the most efficient estimator.

The maximum Likelihood estimator (assuming the distribution is correctly specified and R1-R8 hold) is the most efficient consistent and asymptotically normal estimator. * most efficient among ALL consistent estimators (not limited to unbiased or linear estimators).

Note

- ML is better choice for binary, strictly positive, count, or inherent heteroskedasticity than linear model.
- ML will assume that we know the conditional distribution of the outcome, and derive an estimator using that information.
 - Adds an assumption that we know the distribution (which is similar to A6 Normal Distribution in linear model)
 - will produce a more efficient estimator.

5.10.4 Compare to OLS

MLE is not a cure for most of OLS problems:

- To do joint inference in MLE, we typically use log-likelihood calculation, instead of F-score
- Functional form affects estimation of MLE and OLS.
- Perfect Collinearity/Multicollinearity: highly correlated are likely to yield large standard errors.
- Endogeneity (Omitted variables bias, Simultaneous equations): Like OLS, MLE is also biased against this problem

5.10.5 Application

Other applications of MLE

- Corner Solution
 - Ex: hours worked, donations to charity
 - Estimate with Tobit
- Non-negative count
 - Ex: Numbers of arrest, Number of cigarettes smoked a day
 - Estimate with Poisson regression
- Multinomial Choice
 - Ex: Demand for cars, votes for primary election
 - Estimate with multinomial probit or logit
- Ordinal Choice
 - Ex: Levels of Happiness, Levels of Income
 - Ordered Probit

Model for binary Response

A binary variable will have a Bernoulli distribution:

$$f_Y(y_i; p) = p^{y_i} (1 - p)^{(1-y_i)}$$

where p is the probability of success. The conditional distribution is:

$$f_{Y|X}(y_i, x_i; p(.)) = p(x_i)^{y_i} (1 - p(x_i))^{(1-y_i)}$$

So choose $p(x_i)$ to be a reasonable function of x_i and unknown parameters θ

We can use **latent variable model** as probability functions

$$y_i = 1\{y_i^* > 0\} y_i^* = x_i \beta - \epsilon_i$$

- y_i^* is a latent variable (unobserved) that is not well-defined in terms of units/magnitudes
- ϵ_i is a mean 0 unobserved random variable.

We can rewrite the model without the latent variable,

$$y_i = 1\{x_i \beta > \epsilon_i\}$$

Then the probability function,

$$p(x_i) = P(y_i = 1|x_i) = P(x_i \beta > \epsilon_i|x_i) = F_{\epsilon|X}(x_i \beta|x_i)$$

then we need to choose a conditional distribution for ϵ_i . Hence, we can make additional strong independence assumption

ϵ_i is independent of x_i

Then the probability function is simply,

$$p(x_i) = F_{\epsilon}(x_i \beta)$$

The probability function is also the conditional expectation function,

$$E(y_i|x_i) = P(y_i = 1|x_i) = F_{\epsilon}(x_i \beta)$$

so we allow the conditional expectation function to be non-linear.

Common distributional assumption

1. **Probit:** Assume ϵ_i is standard normally distributed, then $F_\epsilon(\cdot) = \Phi(\cdot)$ is the standard normal CDF.
2. **Logit:** Assume ϵ_i is standard logistically distributed, then $F_\epsilon(\cdot) = \Lambda(\cdot)$ is the standard normal CDF.

Step to derive

1. Choose a distribution (normal or logistic) and plug into the following log likelihood,

$$\ln(f_{Y|X}(y_i, x_i; \beta)) = y_i \ln(F_\epsilon(x_i \beta)) + (1 - y_i) \ln(1 - F_\epsilon(x_i \beta))$$

2. Solve the MLE by finding the Maximum of

$$\hat{\beta}_{MLE} = \operatorname{argmax} \sum_{i=1}^n \ln(f_{Y|X}(y_i, x_i; \beta))$$

Properties of the Probit and Logit Estimators

- Probit or Logit is consistent and asymptotically normal if
 - [A2] holds: $E(x_i' x_i)$ exists and is non-singular
 - [A5] (or A5a) holds: $\{y_i, x_i\}$ are iid (or stationary and weakly dependent).
 - Distributional assumptions on ϵ_i hold: Normal/Logistic and independent of x_i
- Under the same assumptions, Probit or Logit is also asymptotically efficient with asymptotic variance,

$$I(\beta_0)^{-1} = [E(\frac{(f_\epsilon(x_i \beta_0))^2}{F_\epsilon(x_i \beta_0)(1 - F_\epsilon(x_i \beta_0))} x_i' x_i)]^{-1}$$

where $F_\epsilon(x_i \beta_0)$ is the probability density function (derivative of the CDF)

5.10.5.1 Interpretation

β is the average response in the latent variable associated with a change in x_i

- Magnitudes do not have meaning
- Direction does have meaning

The **partial effect** for a Non-linear binary response model

$$E(y_i|x_i) = F_\epsilon(x_i\beta)PE(x_{ij}) = \frac{\partial E(y_i|x_i)}{\partial x_{ij}} = f_\epsilon(x_i\beta)\beta_j$$

- The partial effect is the coefficient parameter β_j multiplied by a scaling factor $f_\epsilon(x_i\beta)$
- The scaling factor depends on x_i so the partial effect changes depending on what x_i is

Single value for the partial effect

- **Partial Effect at the Average (PEA)** is the partial effect for an average individual

$$f_\epsilon(\bar{x}\hat{\beta})\hat{\beta}_j$$

- **Average Partial Effect (APE)** is the average of all partial effect for each individual.

$$\frac{1}{n} \sum_{i=1}^n f_\epsilon(x_i\hat{\beta})\hat{\beta}_j$$

In the linear model, APE = PEA.

In a non-linear model (e.g., binary response), APE \neq PEA

Chapter 6

Non-linear Regression

Definition: models in which the derivatives of the mean function with respect to the parameters depend on one or more of the parameters.

To approximate data, we can approximate the function

- by a high-order polynomial
- by a linear model (e.g., a Taylor expansion around X's)
- a collection of locally linear models or basis function

but it would not be easy to interpret, or not enough data, or can't interpret them globally.

intrinsically nonlinear models:

$$Y_i = f(\mathbf{x}_i; \theta) + \epsilon_i$$

where $f(\mathbf{x}_i; \theta)$ is a nonlinear function relating $E(Y_i)$ to the independent variables x_i

- \mathbf{x}_i is a $k \times 1$ vector of independent variables (fixed).
- θ is a $p \times 1$ vector of parameters.
- ϵ_i s are iid variables mean 0 and variance σ^2 . (sometimes it's normal).

6.1 Inference

Since $Y_i = f(\mathbf{x}_i, \theta) + \epsilon_i$, where $\epsilon_i \sim iid(0, \sigma^2)$. We can obtain $\hat{\theta}$ by minimizing $\sum_{i=1}^n (Y_i - f(x_i, \theta))^2$ and estimate $s^2 = \hat{\sigma}_\epsilon^2 = \frac{\sum_{i=1}^n (Y_i - f(x_i, \theta))^2}{n-p}$

6.1.1 Linear Function of the Parameters

If we assume $\epsilon_i \sim N(0, \sigma^2)$, then

$$\hat{\theta} \sim AN(\cdot, \sigma^2 [\mathbf{F}(\theta)'\mathbf{F}(\theta)]^{-1})$$

where An = asymptotic normality

Asymptotic means we have enough data to make inference (As your sample size increases, this becomes more and more accurate (to the true value)).

Since we want to do inference on linear combinations of parameters or contrasts.

If we have $\mathbf{a} = (\theta_0, \theta_1, \theta_2)'$ and we want to look at $\theta_1 - \theta_2$; we can define vector $\mathbf{a} = (0, 1, -1)'$, consider inference for \mathbf{a}'

Rules for expectation and variance of a fixed vector \mathbf{a} and random vector \mathbf{Z} ;

$$E(\mathbf{a}'\mathbf{Z}) = \mathbf{a}'E(\mathbf{Z})var(\mathbf{a}'\mathbf{Z}) = \mathbf{a}'var(\mathbf{Z})\mathbf{a}$$

Then,

$$\hat{\mathbf{a}}' \sim AN(\mathbf{a}', \sigma^2 \mathbf{a}' [\mathbf{F}(\cdot)'\mathbf{F}(\cdot)]^{-1}\mathbf{a})$$

and $\hat{\mathbf{a}}'$ is asymptotically independent of s^2 (to order $1/n$) then

$$\frac{\hat{\mathbf{a}}' - \mathbf{a}'}{s(\mathbf{a}' [\mathbf{F}(\cdot)'\mathbf{F}(\cdot)]^{-1}\mathbf{a})^{1/2}} \sim t_{n-p}$$

to construct $100(1 - \alpha)\%$ confidence interval for \mathbf{a}'

$$\mathbf{a}' \pm t_{(1-\alpha/2, n-p)} s(\mathbf{a}' [\mathbf{F}(\cdot)'\mathbf{F}(\cdot)]^{-1}\mathbf{a})^{1/2}$$

Suppose $\mathbf{a}' = (0, \dots, j, \dots, 0)$. Then, a confidence interval for the jth element of \mathbf{a}' is

$$\hat{\theta}_j \pm t_{(1-\alpha/2, n-p)} s \sqrt{\hat{c}^j}$$

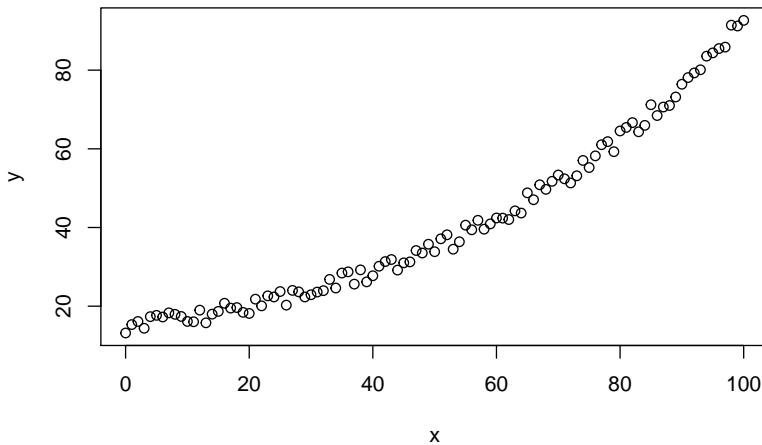
where \hat{c}^j is the jth diagonal element of $[\mathbf{F}(\cdot)'\mathbf{F}(\cdot)]^{-1}$

```
#set a seed value
set.seed(23)

#Generate x as 100 integers using seq function
x<-seq(0,100,1)
```

```
#Generate y as a*e^(bx)+c
y<-runif(1,0,20)*exp(runif(1,0.005,0.075)*x)+runif(101,0,5)

# visualize
plot(x,y)
```



```
#define our data frame
datf = data.frame(x,y)

#define our model function
mod =function(a,b,x) a*exp(b*x)
```

In this example, we can get the starting values by using linearized version of the function $\log y = \log a + bx$. Then, we can fit a linear regression to this and use our estimates as starting values

```
#get starting values by linearizing
lin_mod=lm(log(y)~x,data=datf)

#define the a parameter back from the log scale; b is ok
astrt = exp(as.numeric(lin_mod$coef[1]))
bstrt = as.numeric(lin_mod$coef[2])
print(c(astrt,bstrt))
#> [1] 14.07964761 0.01855635
```

with `nls`, we can fit the nonlinear model via least squares

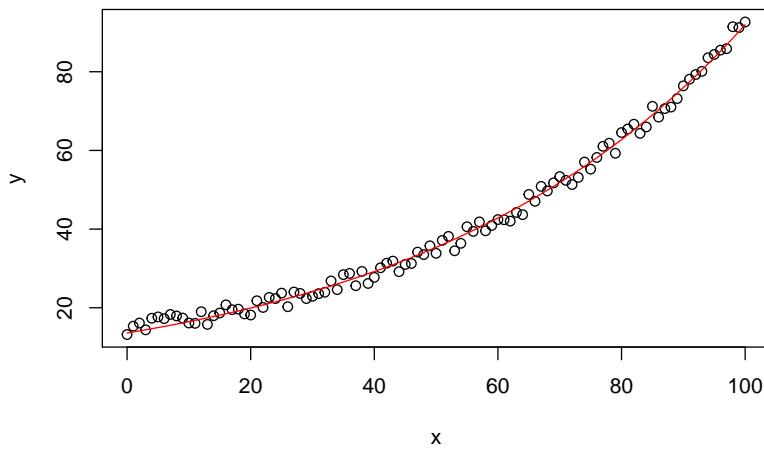
```

nlin_mod = nls(y ~ mod(a, b, x),
                start = list(a = astrt, b = bstrt),
                data = datf)

#look at model fit summary
summary(nlin_mod)
#>
#> Formula: y ~ mod(a, b, x)
#>
#> Parameters:
#>     Estimate Std. Error t value Pr(>|t|)
#> a 13.603909  0.165390   82.25 <2e-16 ***
#> b  0.019110  0.000153  124.90 <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 1.542 on 99 degrees of freedom
#>
#> Number of iterations to convergence: 3
#> Achieved convergence tolerance: 7.006e-07

#add prediction to plot
plot(x, y)
lines(x, predict(nlin_mod), col = "red")

```



6.1.2 Nonlinear

Suppose that $h(\theta)$ is a nonlinear function of the parameters. We can use Taylor series about θ

$$h(\hat{\theta}) \approx h(\theta) + \mathbf{h}'[\hat{\theta} - \theta]$$

where $\mathbf{h} = (\frac{\partial h}{\partial \theta_1}, \dots, \frac{\partial h}{\partial \theta_p})'$

with

$$E(\hat{\theta}) \approx \theta var(\hat{\theta}) \approx \sigma^2 [\mathbf{F}(\cdot)' \mathbf{F}(\cdot)]^{-1} E(h(\hat{\theta})) \approx h(\theta) var(h(\hat{\theta})) \approx \sigma^2 \mathbf{h}' [\mathbf{F}(\cdot)' \mathbf{F}(\cdot)]^{-1} \mathbf{h}$$

Thus,

$$h(\hat{\theta}) \sim AN(h(\theta), \sigma^2 \mathbf{h}' [\mathbf{F}(\cdot)' \mathbf{F}(\cdot)]^{-1} \mathbf{h})$$

and an approximate $100(1 - \alpha)\%$ confidence interval for $h(\theta)$ is

$$h(\hat{\theta}) \pm t_{(1-\alpha/2; n-p)} s(\mathbf{h}' [\mathbf{F}(\cdot)' \mathbf{F}(\cdot)]^{-1} \mathbf{h})^{1/2}$$

where \mathbf{h} and $\mathbf{F}(\theta)$ are evaluated at $\hat{\theta}$

Regarding **prediction interval** for Y at $x = x_0$

$$Y_0 = f(x_0; \theta) + \epsilon_0, \epsilon_0 \sim N(0, \sigma^2) \hat{Y}_0 = f(x_0, \hat{\theta})$$

As $n \rightarrow \infty$, $\hat{\theta} \rightarrow \theta$, so we

$$f(x_0, \hat{\theta}) \approx f(x_0, \theta) + \mathbf{f}_0(\cdot)'[\hat{\theta} - \theta]$$

where

$$\mathbf{f}_0(\theta) = (\frac{\partial f(x_0, \theta)}{\partial \theta_1}, \dots, \frac{\partial f(x_0, \theta)}{\partial \theta_p})'$$

(note: this $f_0(\theta)$ is different from $f(\theta)$).

$$Y_0 - \hat{Y}_0 \approx Y_0 - f(x_0, \theta) - f_0(\theta)'[\hat{\theta} - \theta] = \epsilon_0 - f_0(\theta)'[\hat{\theta} - \theta]$$

$$E(Y_0 - \hat{Y}_0) \approx E(\epsilon_0) E(\hat{\theta} - \theta) = 0 var(Y_0 - \hat{Y}_0) \approx var(\epsilon_0 - (\mathbf{f}_0(\cdot)' \mathbf{f}_0(\cdot))') = \sigma^2 + \sigma^2 \mathbf{f}_0(\cdot)' [\mathbf{F}(\cdot)' \mathbf{F}(\cdot)]^{-1} \mathbf{f}_0(\cdot) = \sigma^2 (1 + \mathbf{f}_0(\cdot)' [\mathbf{F}(\cdot)' \mathbf{F}(\cdot)]^{-1} \mathbf{f}_0(\cdot))$$

Hence, combining

$$Y_0 - \hat{Y}_0 \sim AN(0, \sigma^2(1 + \mathbf{f}_0(\theta)'[\mathbf{F}(\theta)' \mathbf{F}(\theta)]^{-1} \mathbf{f}_0(\theta)))$$

Note:

Confidence intervals for the mean response Y_i (which is different from prediction intervals) can be obtained similarly.

6.2 Non-linear Least Squares

- The LS estimate of θ , $\hat{\theta}$ is the set of parameters that minimizes the residual sum of squares:

$$S(\hat{\theta}) = SSE(\hat{\theta}) = \sum_{i=1}^n \{Y_i - f(\mathbf{x}_i; \hat{\theta})\}^2$$

- to obtain the solution, we can consider the partial derivatives of $S(\theta)$ with respect to each θ_j and set them to 0, which gives a system of p equations. Each normal equation is

$$\frac{\partial S(\theta)}{\partial \theta_j} = -2 \sum_{i=1}^n \{Y_i - f(\mathbf{x}_i; \theta)\} \left[\frac{\partial f(\mathbf{x}_i; \theta)}{\partial \theta_j} \right] = 0$$

- but we can't obtain a solution directly/analytically for this equation.

Numerical Solutions

- Grid search
 - A “grid” of possible parameter values and see which one minimize the residual sum of squares.
 - finer grid = greater accuracy
 - could be inefficient, and hard when p is large.
- Gauss-Newton Algorithm
 - we have an initial estimate of θ denoted as $\hat{\theta}^{(0)}$
 - use a Taylor expansions of $f(\mathbf{x}_i; \theta)$ as a function of θ about the point $\hat{\theta}^{(0)}$

$$\begin{aligned} Y_i &= f(x_i; \theta) + \epsilon_i \\ &= f(x_i; \theta) + \sum_{j=1}^p \left\{ \frac{\partial f(x_i; \theta)}{\partial \theta_j} \right\}_{\theta=\hat{\theta}^{(0)}} (\theta_j - \hat{\theta}_j^{(0)}) + \text{remainder} + \epsilon_i \end{aligned}$$

Equivalently,

In matrix notation,

$$\begin{aligned}\mathbf{Y} &= \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \\ \mathbf{f}(\hat{\theta}^{(0)}) &= \begin{bmatrix} f(\mathbf{x}_1; \hat{\theta}^{(0)}) \\ \vdots \\ f(\mathbf{x}_n; \hat{\theta}^{(0)}) \end{bmatrix} \\ &= \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix} \\ \mathbf{F}(\hat{\theta}^{(0)}) &= \left[\begin{array}{ccc} \frac{\partial f(x_1,)}{\partial \theta_1} & \cdots & \frac{\partial f(x_1,)}{\partial \theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(x_n,)}{\partial \theta_1} & \cdots & \frac{\partial f(x_n,)}{\partial \theta_p} \end{array} \right]_{\theta=\hat{\theta}^{(0)}}\end{aligned}$$

Hence,

$$\mathbf{Y} = \mathbf{f}(\hat{\theta}^{(0)}) + \mathbf{F}(\hat{\theta}^{(0)})(\theta - \hat{\theta}^{(0)}) + \epsilon + \text{remainder}$$

where we assume that the remainder is small and the error term is only assumed to be iid with mean 0 and variance σ^2 .

We can rewrite the above equation as

$$\mathbf{Y} - \mathbf{f}(\hat{\theta}^{(0)}) \approx \mathbf{F}(\hat{\theta}^{(0)})(\theta - \hat{\theta}^{(0)}) + \epsilon$$

where it is in the form of linear model. After we solve for $(\theta - \hat{\theta}^{(0)})$ and let it equal to $\hat{\delta}^{(1)}$

Then we new estimate is given by adding the Gauss increment adjustment to the initial estimate $\hat{\theta}^{(1)} = \hat{\theta}^{(0)} + \hat{\delta}^{(1)}$

We can repeat this process.

Gauss-Newton Algorithm Steps:

1. initial estimate $\hat{\theta}^{(0)}$, set $j = 0$
2. Taylor series expansion and calculate $\mathbf{f}(\hat{\theta}^{(j)})$ and $\mathbf{F}(\hat{\theta}^{(j)})$
3. Use OLS to get $\hat{\delta}^{(j+1)}$
4. get the new estimate $\hat{\theta}^{(j+1)}$, return to step 2
5. continue until "convergence"

6. With the final parameter estimate $\hat{\theta}$, we can estimate σ^2 if $\epsilon \sim (\mathbf{0}, \sigma^2 \mathbf{I})$ by

$$\hat{\sigma}^2 = \frac{1}{n-p}(\mathbf{Y} - \mathbf{f}(x; \hat{\theta}))'(\mathbf{Y} - \mathbf{f}(x; \hat{\theta}))$$

Criteria for convergence

1. Minor change in the objective function (SSE = residual sum of squares)

$$\frac{|SSE(\hat{\theta}^{(j+1)}) - SSE(\hat{\theta}^{(j)})|}{SSE(\hat{\theta}^{(j)})} < \gamma_1$$

2. Minor change in the parameter estimates

$$|\hat{\theta}^{(j+1)} - \hat{\theta}^{(j)}| < \gamma_2$$

3. “residual projection” criterion of (Bates and Watts, 1981)

6.2.1 Alternative of Gauss-Newton Algorithm

6.2.1.1 Gauss-Newton Algorithm

Normal equations:

$$\frac{\partial SSE(\theta)}{\partial \theta} = 2\mathbf{F}(\theta)'[\mathbf{Y} - \mathbf{f}(\theta)]$$

$$\begin{aligned}\hat{\theta}^{(j+1)} &= \hat{\theta}^{(j)} + \hat{\delta}^{(j+1)} \\ &= \hat{\theta}^{(j)} + [\mathbf{F}((\hat{\theta})^{(j)})'\mathbf{F}(\hat{\theta}^{(j)})]^{-1}\mathbf{F}(\hat{\theta})^{(j)} \\ &= \hat{\theta}^{(j)} - \frac{1}{2}[\mathbf{F}(\hat{\theta}^{(j)})'\mathbf{F}(\hat{\theta}^{(j)})]^{-1}\frac{\partial SSE(\hat{\theta}^{(j)})}{\partial \theta}\end{aligned}$$

where

- $\frac{\partial SSE(\hat{\theta}^{(j)})}{\partial \theta}$ is a gradient vector (points in the direction in which the SSE increases most rapidly). This path is known as steepest ascent.
- $[\mathbf{F}(\hat{\theta}^{(j)})'\mathbf{F}(\hat{\theta}^{(j)})]^{-1}$ indicates how far to move
- $-1/2$: indicator of the direction of steepest descent.

6.2.1.2 Modified Gauss-Newton Algorithm

To avoid overstepping (the local min), we can use the modified Gauss-Newton Algorithm. We define a new proposal for θ

$$\hat{\theta}^{(j+1)} = \hat{\theta}^{(j)} + \alpha_j \hat{\delta}^{(j+1)}, 0 < \alpha_j < 1$$

where

- α_j (called the “learning rate”): is used to modify the step length.

We could also have $\alpha * 1/2$, but typically it is assumed to be absorbed into the learning rate.

A way to choose α_j , we can use **step halving**

$$\hat{\theta}^{(j+1)} = \hat{\theta}^{(j)} + \frac{1}{2^k} \hat{\delta}^{(j+1)}$$

where

- k is the smallest non-negative integer such that

$$SSE(\hat{\theta}^{(j)} + \frac{1}{2^k} \hat{\delta}^{(j+1)}) < SSE(\hat{\theta}^{(j)})$$

which means we try $\hat{\delta}^{(j+1)}$, then $\hat{\delta}^{(j+1)}/2$, $\hat{\delta}^{(j+1)}/4$, etc.

The most general form of the convergence algorithm is

$$\hat{\theta}^{(j+1)} = \hat{\theta}^{(j)} - \alpha_j \mathbf{A}_j \frac{\partial Q(\hat{\theta}^{(j)})}{\partial \theta}$$

where

- \mathbf{A}_j is a positive definite matrix
- α_j is the learning rate
- $\frac{\partial Q(\hat{\theta}^{(j)})}{\partial \theta}$ is the gradient based on some objective function Q (a function of θ), which is typically the SSE in nonlinear regression applications (e.g., cross-entropy for classification).

Refer back to the **Modified Gauss-Newton Algorithm**, we can see it is in this form

$$\hat{\theta}^{(j+1)} = \hat{\theta}^{(j)} - \alpha_j [\mathbf{F}(\hat{\theta}^{(j)})' \mathbf{F}(\hat{\theta}^{(j)})]^{-1} \frac{\partial SSE(\hat{\theta}^{(j)})}{\partial \theta}$$

where $\mathbf{Q} = SSE$, $[\mathbf{F}(\hat{\theta}^{(j)})' \mathbf{F}(\hat{\theta}^{(j)})]^{-1} = \mathbf{A}$

6.2.1.3 Steepest Descent

(also known just “gradient descent”)

$$\hat{\theta}^{(j+1)} = \hat{\theta}^{(j)} - \alpha_j \mathbf{I}_{p \times p} \frac{\partial \mathbf{Q}(\hat{\theta}^{(j)})}{\partial \theta}$$

- slow to converge, moves rapidly initially.
- could be used for starting values

6.2.1.4 Levenberg -Marquardt

$$\hat{\theta}^{(j+1)} = \hat{\theta}^{(j)} - \alpha_j [\mathbf{F}(\hat{\theta}^{(j)})' \mathbf{F}(\hat{\theta}^{(j)}) + \tau \mathbf{I}_{p \times p}] \frac{\partial \mathbf{Q}(\hat{\theta}^{(j)})}{\partial \theta}$$

which is a compromise between the Gauss-Newton Algorithm and the Steepest Descent.

- best when $\mathbf{F}(\hat{\theta}^{(j)})' \mathbf{F}(\hat{\theta}^{(j)})$ is nearly singular ($\mathbf{F}(\hat{\theta}^{(j)})$ isn't of full rank)
- similar to ridge regression
- If $SSE(\hat{\theta}^{(j+1)}) < SSE(\hat{\theta}^{(j)})$, then $\tau = \tau/10$ for the next iteration. Otherwise, $\tau = 10\tau$

6.2.1.5 Newton-Raphson

$$\hat{\theta}^{(j+1)} = \hat{\theta}^{(j)} - \alpha_j [\frac{\partial^2 Q(\hat{\theta}^{(j)})}{\partial \theta \partial \theta'}]^{-1} \frac{\partial \mathbf{Q}(\hat{\theta}^{(j)})}{\partial \theta}$$

The **Hessian matrix** can be rewritten as:

$$\frac{\partial^2 Q(\hat{\theta}^{(j)})}{\partial \theta \partial \theta'} = 2\mathbf{F}((\hat{\theta})^{(j)})' \mathbf{F}(\hat{\theta}^{(j)}) - 2 \sum_{i=1}^n [Y_i - f(x_i; \theta)] \frac{\partial^2 f(x_i; \theta)}{\partial \theta \partial \theta'}$$

which contains the same term that Gauss-Newton Algorithm, combined with one containing the second partial derivatives of $f()$. (methods that require the second derivatives of the objective function are known as “second-order methods”.) However, the last term $\frac{\partial^2 f(x_i; \theta)}{\partial \theta \partial \theta'}$ can sometimes be nonsingular.

6.2.1.6 Quasi-Newton

update θ according to

$$\hat{\theta}^{(j+1)} = \hat{\theta}^{(j)} - \alpha_j \mathbf{H}_j^{-1} \frac{\partial \mathbf{Q}(\hat{\theta}^{(j)})}{\partial \theta}$$

where H_j is a symmetric positive definite approximation to the Hessian, which gets closer as $j \rightarrow \infty$.

- \mathbf{H}_j is computed iteratively
- Among first-order methods (where only first derivatives are required), this method performs best.

6.2.1.7 Derivative Free Methods

- **secant Method:** like Gauss-Newton Algorithm, but calculates the derivatives numerically from past iterations.
- **Simplex Methods**
- **Genetic Algorithm**
- **Differential Evolution Algorithms**
- **Particle Swarm Optimization**
- **Ant Colony Optimization**

6.2.2 Practical Considerations

To converge, algorithm need good initial estimates.

- Starting values:
 - Prior or theoretical info

- A grid search or a graph of $SSE(\theta)$
- could also use OLS to get starting values.
- Model interpretation: if you have some idea regarding the form of the objective function, then you can try to guess the initial value.
- Expected Value Parameterization
- Constrained Parameters: (constraints on parameters like $\theta_i > a, a < \theta_i < b$)
 - fit the model first to see if the converged parameter estimates satisfy the constraints.
 - if they don't satisfy, then try re-parameterizing

6.2.2.1 Failure to converge

- $SSE(\theta)$ may be “flat” in a neighborhood of the minimum.
- You can try different or “better” starting values.
- Might suggest the model is too complex for the data, might consider simpler model.

6.2.2.2 Convergence to a Local Minimum

- Linear least squares has the property that $SSE(\theta) = (\mathbf{Y} - \mathbf{X})'(\mathbf{Y} - \mathbf{X})$, which is quadratic and has a unique minimum (or maximum).
- Nonlinear least squares need not have a unique minimum
- Using different starting values can help
- If the dimension of θ is low, graph $SSE(\theta)$ as a function of θ_i
- Different algorithm can help (e.g., genetic algorithm, particle swarm)

To converge, algorithms need good initial estimates.

- Starting values:
 - prior or theoretical info
 - A grid search or a graph
 - OLS estimates as starting values
 - Model interpretation
 - Expected Value Parameterization

- Constrained Parameters:

- try the model without the constraints first.
- If the resulted parameter estimates does not satisfy the constraint, try re-parameterizing

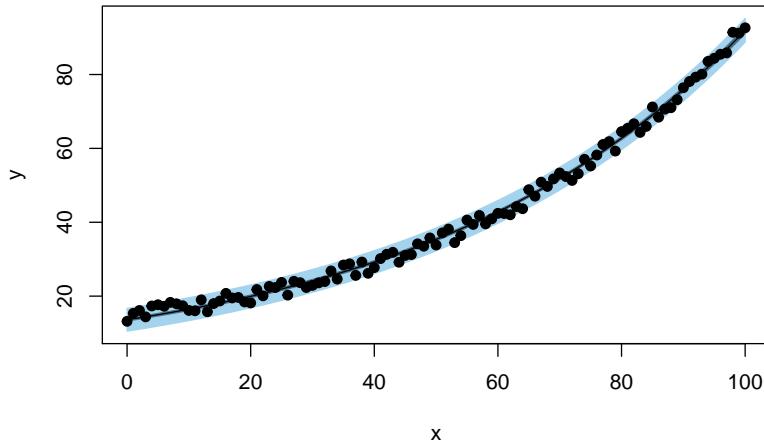
```
# Grid search
#choose grid of a and b values
aseq = seq(10,18,.2)
bseq = seq(.001,.075,.001)

na = length(aseq)
nb = length(bseq)
SSout = matrix(0,na*nb,3) #matrix to save output
cnt = 0
for (k in 1:na){
  for (j in 1:nb){
    cnt = cnt+1
    ypred = mod(aseq[k],bseq[j],x) #evaluate model w/ these parms
    ss = sum((y-ypred)^2) #this is our SSE objective function
    #save values of a, b, and SSE
    SSout[cnt,1]=aseq[k]
    SSout[cnt,2]=bseq[j]
    SSout[cnt,3]=ss
  }
}
#find minimum SSE and associated a,b values
mn_indx = which.min(SSout[,3])
astrt = SSout[mn_indx,1]
bstrt = SSout[mn_indx,2]
#now, run nls function with these starting values
nlin_modG=nls(y~mod(a,b,x),start=list(a=astrt,b=bstrt))

nlin_modG
#> Nonlinear regression model
#>   model: y ~ mod(a, b, x)
#>   data: parent.frame()
#>       a          b
#> 13.60391  0.01911
#>   residual sum-of-squares: 235.5
#>
#> Number of iterations to convergence: 3
#> Achieved convergence tolerance: 2.293e-07
# Note, the package `nls_multstart` will allow you to do a grid search without programming your own
```

For prediction interval

```
plotFit(
  nlin_modG,
  interval = "both",
  pch = 19,
  shade = TRUE,
  col.conf = "skyblue4",
  col.pred = "lightskyblue2",
  data = datf
)
```



Based on the forms of your function, you can also have programmed starting values from `nls` function (e.g., logistic growth, asymptotic regression, etc).

```
apropos("SS")
#> [1] "ss"           "SSasymp"      "SSasympOff"   "SSasympOrig"  "SSbiexp"
#> [6] "SSD"          "SSfol"        "SSfpl"        "SSgompertz"   "SSlogis"
#> [11] "SSmicmen"    "SSout"        "SSweibull"
```

For example, a logistic growth model:

$$P = \frac{K}{1 + \exp(P_0 + rt)} + \epsilon$$

where

- P = population at time t

- K = carrying capacity
- r = population growth rate

but in R you have slight different parameterization:

$$P = \frac{asym}{1 + exp(\frac{xmid-t}{scal})}$$

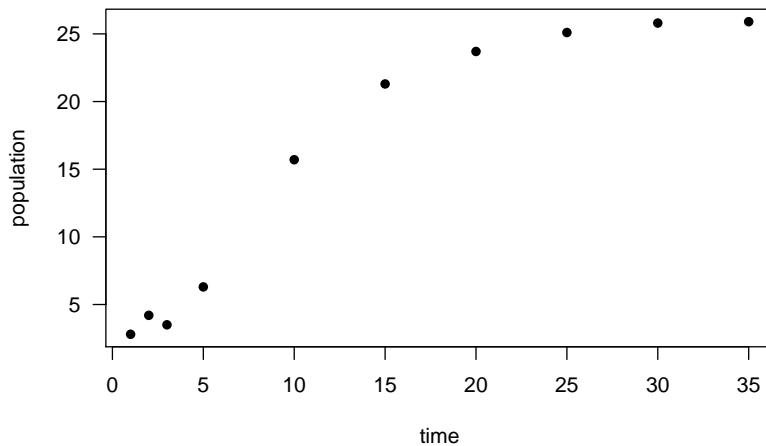
where

- asym = carrying capacity
- xmid = the x value at the inflection point of the curve
- scal = scaling parameter.

Hence, you have

- K = asym
- r = -1/scal
- $P_0 = -rxmid$

```
# simulated data
time <- c(1, 2, 3, 5, 10, 15, 20, 25, 30, 35)
population <- c(2.8, 4.2, 3.5, 6.3, 15.7, 21.3, 23.7, 25.1, 25.8, 25.9)
plot(time, population, las = 1, pch = 16)
```



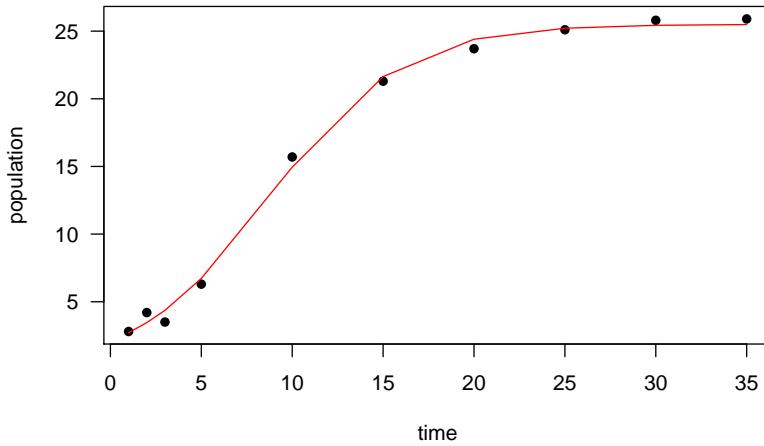
```
# model fitting
logisticModelSS <- nls(population ~ SSlogis(time, Asym, xmid, scal))
summary(logisticModelSS)
#>
#> Formula: population ~ SSlogis(time, Asym, xmid, scal)
#>
#> Parameters:
#>     Estimate Std. Error t value Pr(>|t|)
#> Asym   25.5029    0.3666   69.56 3.34e-11 ***
#> xmid   8.7347    0.3007   29.05 1.48e-08 ***
#> scal    3.6353    0.2186   16.63 6.96e-07 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.6528 on 7 degrees of freedom
#>
#> Number of iterations to convergence: 1
#> Achieved convergence tolerance: 1.908e-06
coef(logisticModelSS)
#>      Asym      xmid      scal
#> 25.502890  8.734698  3.635333
```

Other parameterization

```
#convert to other parameterization
Ks = as.numeric(coef(logisticModelSS)[1])
rs = -1/as.numeric(coef(logisticModelSS)[3])
Pos = - rs * as.numeric(coef(logisticModelSS)[2])
#let's refit with these parameters
logisticModel <- nls(population ~ K / (1 + exp(Po + r * time)), start=list(Po=Pos,r=rs,1))
summary(logisticModel)
#>
#> Formula: population ~ K/(1 + exp(Po + r * time))
#>
#> Parameters:
#>     Estimate Std. Error t value Pr(>|t|)
#> Po   2.40272    0.12702   18.92 2.87e-07 ***
#> r    -0.27508    0.01654  -16.63 6.96e-07 ***
#> K    25.50289    0.36665   69.56 3.34e-11 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.6528 on 7 degrees of freedom
#>
```

```
#> Number of iterations to convergence: 0
#> Achieved convergence tolerance: 1.924e-06

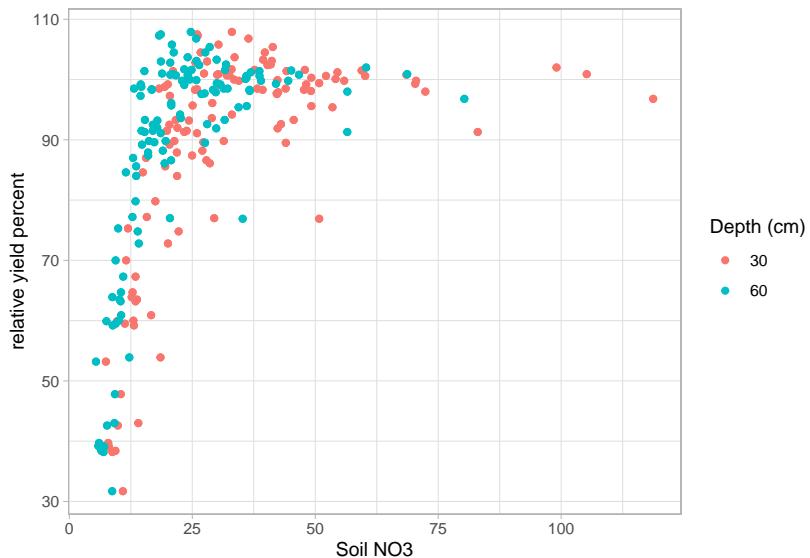
#note: initial values = solution (highly unusual, but ok)
plot(time, population, las = 1, pch = 16)
lines(time, predict(logisticModel), col = "red")
```



If can also define your own self-starting function if your models are uncommon (built in `nls`)

Example is based on (Schabenberger and Pierce, 2001)

```
#Load data
dat <- read.table("images/dat.txt", header = T)
# plot
dat.plot <-
  ggplot(dat) + geom_point(aes(
    x = no3,
    y = ryp,
    color = as.factor(depth)
  )) +
  labs(color = 'Depth (cm)') + xlab('Soil NO3') + ylab('relative yield percent')
dat.plot
```



The suggested model (known as plateau model) is

$$E(Y_{ij}) = (\beta_{0j} + \beta_{1j}N_{ij})I_{N_{ij} \leq \alpha_j} + (\beta_{0j} + \beta_{1j}\alpha_j)I_{N_{ij} > \alpha_j}$$

where

- N is an observation
- i is a particular observation
- j = 1,2 corresponding to depths (30,60)

```
#First define model as a function
nonlinModel <- function(predictor,b0,b1,alpha){
  ifelse(predictor<=alpha,
         b0+b1*predictor, #if observation less than cutoff simple linear model
         b0+b1*alpha) #otherwise flat line
}
```

define `selfStart` function. Because we defined our model to be linear in the first part and then plateau (remain constant) we can use the first half of our predictors (sorted by increasing value) to get an initial estimate for the slope and intercept of the model, and the last predictor value (alpha) can be the starting value for the plateau parameter.

```
nonlinModelInit <- function(mCall,LHS,data){
  #sort data by increasing predictor value -
  #done so we can just use the low level no3 conc to fit a simple model
  xy <- sortedXyData(mCall[['predictor']],LHS,data)
  n <- nrow(xy)
  #For the first half of the data a simple linear model is fit
  lmFit <- lm(xy[1:(n/2), 'y']~xy[1:(n/2), 'x'])
  b0 <- coef(lmFit)[1]
  b1 <- coef(lmFit)[2]
  #for the cut off to the flat part select the last x value used in creating linear model
  alpha <- xy[(n/2), 'x']
  value <- c(b0,b1,alpha)
  names(value) <- mCall[c('b0','b1','alpha')]
  value
}
```

combine model and custom function to calculate starting values.

```
SS_nonlinModel <- selfStart(nonlinModel,nonlinModelInit,c('b0','b1','alpha'))

#Above code defined model and selfStart now just need to call it for each of the depths
sep30_nls <-
  nls(ryp ~ SS_nonlinModel(predictor = no3, b0, b1, alpha), data = dat[dat$depth ==
    30, ])

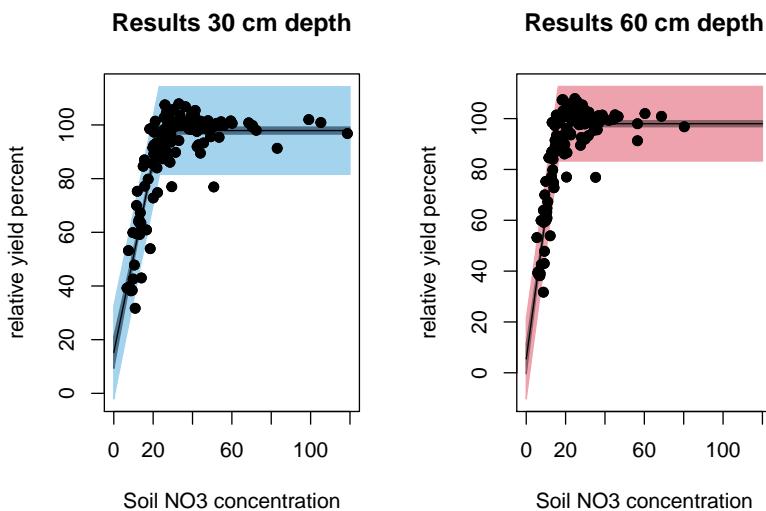
sep60_nls <-
  nls(ryp ~ SS_nonlinModel(predictor = no3, b0, b1, alpha), data = dat[dat$depth ==
    60, ])

par(mfrow = c(1, 2))
plotFit(
  sep30_nls,
  interval = "both",
  pch = 19,
  shade = TRUE,
  col.conf = "skyblue4",
  col.pred = "lightskyblue2",
  data = dat[dat$depth == 30, ],
  main = 'Results 30 cm depth',
  ylab = 'relative yield percent',
  xlab = 'Soil NO3 concentration',
  xlim = c(0, 120)
)
plotFit(
  sep60_nls,
```

```

interval = "both",
pch = 19,
shade = TRUE,
col.conf = "lightpink4",
col.pred = "lightpink2",
data = dat[dat$depth == 60, ],
main = 'Results 60 cm depth',
ylab = 'relative yield percent',
xlab = 'Soil NO3 concentration',
xlim = c(0, 120)
)

```



```

summary(sep30_nls)
#>
## Formula: ryp ~ SS_nonlinModel(predictor = no3, b0, b1, alpha)
#>
## Parameters:
##             Estimate Std. Error t value Pr(>|t|)
## b0      15.1943    2.9781   5.102 6.89e-07 ***
## b1       3.5760    0.1853  19.297 < 2e-16 ***
## alpha   23.1324    0.5098  45.373 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 8.258 on 237 degrees of freedom
## 
```

```
#> Number of iterations to convergence: 6
#> Achieved convergence tolerance: 3.608e-09
summary(sep60_nls)
#>
#> Formula: ryp ~ SS_nonlinModel(predictor = no3, b0, b1, alpha)
#>
#> Parameters:
#>      Estimate Std. Error t value Pr(>|t|)
#> b0     5.4519    2.9785   1.83  0.0684 .
#> b1     5.6820    0.2529  22.46 <2e-16 ***
#> alpha  16.2863   0.2818  57.80 <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 7.427 on 237 degrees of freedom
#>
#> Number of iterations to convergence: 5
#> Achieved convergence tolerance: 8.571e-09
```

Instead of modeling the depths model separately we model them together - so there is a common slope, intercept, and plateau.

```
red_nls <-
  nls(ryp ~ SS_nonlinModel(predictor = no3, b0, b1, alpha), data = dat)

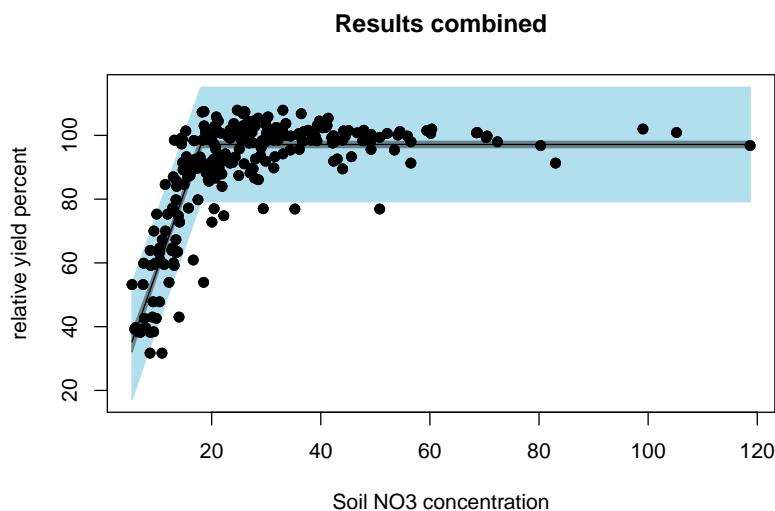
summary(red_nls)
#>
#> Formula: ryp ~ SS_nonlinModel(predictor = no3, b0, b1, alpha)
#>
#> Parameters:
#>      Estimate Std. Error t value Pr(>|t|)
#> b0     8.7901    2.7688   3.175  0.0016 **
#> b1     4.8995    0.2207  22.203 <2e-16 ***
#> alpha  18.0333   0.3242  55.630 <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 9.13 on 477 degrees of freedom
#>
#> Number of iterations to convergence: 7
#> Achieved convergence tolerance: 7.126e-09

par(mfrow = c(1, 1))
plotFit(
  red_nls,
```

```

interval = "both",
pch = 19,
shade = TRUE,
col.conf = "lightblue4",
col.pred = "lightblue2",
data = dat,
main = 'Results combined',
ylab = 'relative yield percent',
xlab = 'Soil NO3 concentration'
)

```

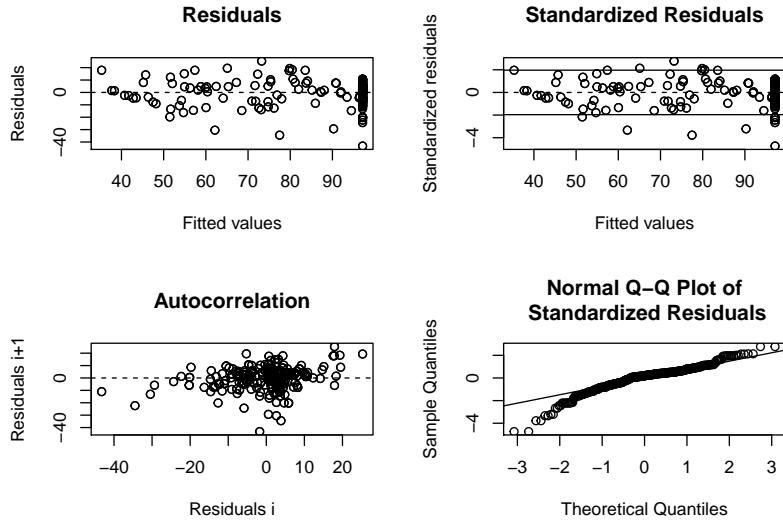


Examine residual values for the combined model.

```

library(nlstools)
#using nlstools nlsResiduals function to get some quick residual plots
#can also use test.nlsResiduals(resid)
# https://www.rdocumentation.org/packages/nlstools/versions/1.0-2
resid <- nlsResiduals(red_nls)
plot(resid)

```



can we test whether the parameters for the two soil depth fits are significantly different? To know if the combined model is appropriate, we consider a parameterization where we let the parameters for the 60cm model be equal to the parameters from the 30cm model plus some increment:

$$\beta_{02} = \beta_{01} + d_0 \quad \beta_{12} = \beta_{11} + d_1 \quad \alpha_2 = \alpha_1 + d_a$$

We can implement this in the following function:

```
nonlinModelF <- function(predictor,soildep,b01,b11,a1,d0,d1,da){
  b02 = b01 + d0 #make 60cm parms = 30cm parms + increment
  b12 = b11 + d1
  a2 = a1 + da

  y1 = ifelse(predictor<=a1,
             b01+b11*predictor, #if observation less than cutoff simple linear model
             b01+b11*a1) #otherwise flat line
  y2 = ifelse(predictor<=a2,
             b02+b12*predictor,
             b02+b12*a2)
  y = y1*(soildep == 30) + y2*(soildep == 60) #combine models
  return(y)
}
```

Starting values are easy now because we fit each model individually.

```

Soil_full=nls(ryp~nonlinModelF(predictor=no3,soildep=depth,b01,b11,a1,d0,d1,da),
              data=dat,
              start=list(b01=15.2,b11=3.58,a1=23.13,d0=-9.74,d1=2.11,da=-6.85))

summary(Soil_full)
#>
#> Formula: ryp ~ nonlinModelF(predictor = no3, soildep = depth, b01, b11,
#>      a1, d0, d1, da)
#>
#> Parameters:
#>   Estimate Std. Error t value Pr(>|t|)
#> b01    15.1943   2.8322   5.365 1.27e-07 ***
#> b11     3.5760   0.1762  20.291  < 2e-16 ***
#> a1     23.1324   0.4848  47.711  < 2e-16 ***
#> d0     -9.7424   4.2357  -2.300   0.0219 *
#> d1      2.1060   0.3203   6.575 1.29e-10 ***
#> da     -6.8461   0.5691 -12.030  < 2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 7.854 on 474 degrees of freedom
#>
#> Number of iterations to convergence: 1
#> Achieved convergence tolerance: 3.742e-06

```

So, the increment parameters, d_1, d_2, d_a are all significantly different from 0, suggesting that we should have two models here.

6.2.3 Model/Estiamtion Adequcy

(Bates and Watts, 1980) assess nonlinearity in terms of 2 components of curvature:

- **Intrinsic nonlinearity:** the degree of bending and twisting in $f(\theta)$; our estimation approach assumes that hte true function is relatively flat (planar) in the neighborhood fo $\hat{\theta}$, which would not be true if $f()$ has a lot of “bending” int he neighborhood of $\hat{\theta}$ (independent of parameterizaiton)
 - If bad, the distribution of residuals will be seriously distorted
 - slow to converge
 - difficult to identify (could use this function `rms.curve`)
 - Solution:

- * could use higher order Taylor expansions estimation
- * Bayesian method
- **Parameter effects nonlinearity:** degree to which curvature (nonlinearity) is affected by choice of θ (data dependent; dependent on parameterization)
 - leads to problems with inference on $\hat{\theta}$
 - `rms.curve` in MASS can identify
 - bootstrap-based inference can also be used
 - Solution: try to reparameterize.

```
#check parameter effects and intrinsic curvature

modD = deriv3(~ a*exp(b*x), c("a","b"),function(a,b,x) NULL)

nlin_modD=nls(y~modD(a,b,x),start=list(a=astrt,b=bstrt),data=datf)

rms.curv(nlin_modD)
#> Parameter effects: c^theta x sqrt(F) = 0.0626
#>           Intrinsic: c^iota x sqrt(F) = 0.0062
```

In linear model, we have Linear Regression, we have goodness of fit measure as R^2 :

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

but not valid in the nonlinear case because the error sum of squares and model sum of squares do not add to the total corrected sum of squares

$$SSR + SSE \neq SST$$

but we can use pseudo- R^2 :

$$R^2_{pseudo} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

But we can't interpret this as the proportion of variability explained by the model. We should use as a relative comparison of different models.

Residual Plots: standardize, similar to OLS. useful when the intrinsic curvature is small:

The studentized residuals

$$r_i = \frac{e_i}{s\sqrt{1 - \hat{c}_i}}$$

where \hat{c}_i is the i-th diagonal of $\hat{\mathbf{H}} = \mathbf{F}(\boldsymbol{\theta})[\mathbf{F}(\boldsymbol{\theta})'\mathbf{F}(\boldsymbol{\theta})]^{-1}\mathbf{F}(\boldsymbol{\theta})'$

We could have problems of

- Collinearity: the condition number of $[\mathbf{F}(\boldsymbol{\theta})'\mathbf{F}(\boldsymbol{\theta})]^{-1}$ should be less than 30. Follow (Magel and Hertsgaard, 1987); reparameterize if possible
- Leverage: Like OLS, but consider $\hat{\mathbf{H}} = \mathbf{F}(\boldsymbol{\theta})[\mathbf{F}(\boldsymbol{\theta})'\mathbf{F}(\boldsymbol{\theta})]^{-1}\mathbf{F}(\boldsymbol{\theta})'$ (also known as “tangent plant hat matrix”) (Laurent and Cook, 1992)
- Heterogeneous Errors: weighted Non-linear Least Squares
- Correlated Errors:
 - Generalized Nonlinear Least Squares
 - Nonlinear Mixed Models
 - Bayesian methods

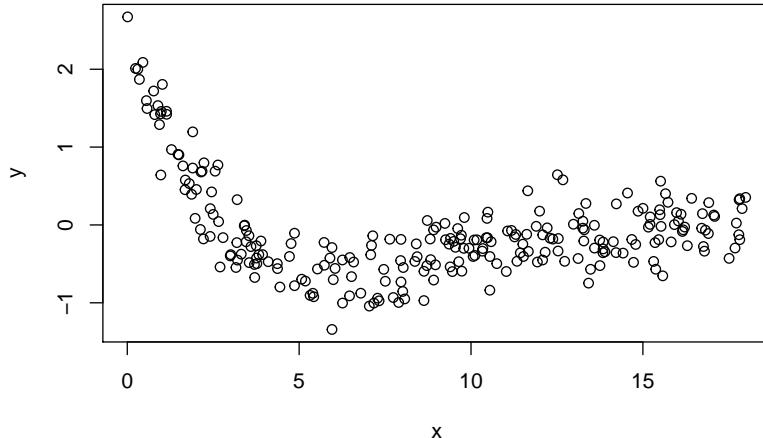
6.2.4 Application

$$y_i = \frac{\theta_0 + \theta_1 x_i}{1 + \theta_2 \exp(0.4x_i)} + \epsilon_i$$

where $i = 1, \dots, n$

Get the starting values

```
plot(my_data)
```



We notice that $Y_{max} = \theta_0 + \theta_1 x_i$ in which we can find x_i from data

```
max(my_data$y)
#> [1] 2.6722
my_data$x[which.max(my_data$y)]
#> [1] 0.0094
```

hence, $x = 0.0094$ when $y = 2.6722$ when we have the first equation as

$$2.6722 = \theta_0 + 0.0094\theta_1 + 0.0094\theta_1 + 0\theta_2 = 2.6722$$

Secondly, we notice that we can obtain the “average” of y when

$$1 + \theta_2 \exp(0.4x) = 2$$

then we can find this average numbers of x and y

```
mean(my_data$y) #find mean y
#> [1] -0.0747864
my_data$y[which.min(abs(my_data$y-(mean(my_data$y))))] # find y closest to its mean
#> [1] -0.0773

my_data$x[which.min(abs(my_data$y-(mean(my_data$y))))] #find x closest to the mean y
#> [1] 11.0648
```

we have the second equation

$$1 + \theta_2 \exp(0.4 * 11.0648) = 20\theta_1 + 0\theta_1 + 83.58967\theta_2 = 1$$

Thirdly, we can plug in the value of x closest to 1 to find the value of y

```
my_data$x[which.min(abs(my_data$x-1))] # find value of x closest to 1
#> [1] 0.9895
match(my_data$x[which.min(abs(my_data$x-1))], my_data$x) # find index of x closest to 1
#> [1] 14
my_data$y[match(my_data$x[which.min(abs(my_data$x-1))], my_data$x)] # find y value
#> [1] 1.4577
```

hence we have

$$1.457 = \frac{\theta_0 + \theta_1 * 0.9895}{1 + \theta_2 \exp(0.4 * 0.9895)} 1.457 + 2.164479 * \theta_2 = \theta_0 + \theta_1 * 0.9895 \theta_0 + \theta_1 * 0.9895 - 2.164479 * \theta_2 = 1.457$$

with 3 equations, we can solve them to get the starting value for $\theta_0, \theta_1, \theta_2$

$$\theta_0 + 0.0094\theta_1 + 0\theta_2 = 2.67220\theta_1 + 0\theta_1 + 83.58967\theta_2 = 1\theta_0 + \theta_1 * 0.9895 - 2.164479 * \theta_2 = 1.457$$

```
library(matlib)
A = matrix(c(0,0.0094, 0, 0, 0, 83.58967, 1, 0.9895, -2.164479), nrow = 3, ncol = 3, byrow = TRUE)
b = c(2.6722, 1, 1.457 )
showEqn(A, b)
#> 0*x1 + 0.0094*x2 + 0*x3 = 2.6722
#> 0*x1 + 0*x2 + 83.58967*x3 = 1
#> 1*x1 + 0.9895*x2 - 2.164479*x3 = 1.457
Solve(A, b, fractions = F)
#> x1 = -279.80879739
#> x2 = 284.27659574
#> x3 = 0.0119632
```

Construct manually Gauss-Newton Algorithm

```
#starting value
theta_0 strt = -279.80879739
theta_1 strt = 284.27659574
theta_2 strt = 0.0119632
```

```

#model
mod_4 = function(theta_0,theta_1,theta_2,x){
  (theta_0 + theta_1*x)/(1+ theta_2*exp(0.4*x))
}

#define a function
f_4 = expression((theta_0 + theta_1*x)/(1+ theta_2*exp(0.4*x)))

#take the first derivative
df_4.d_theta_0=D(f_4,'theta_0')

df_4.d_theta_1=D(f_4,'theta_1')

df_4.d_theta_2=D(f_4,'theta_2')

# save the result of all iterations
theta_vec = matrix(c(theta_0 strt,theta_1 strt,theta_2 strt))
delta= matrix(NA, nrow=3,ncol = 1)

f_theta = as.matrix(eval(f_4,list(x=my_data$x,theta_0 = theta_vec[1,1],theta_1 = theta_vec[2,1],theta_2 = theta_vec[3,1])))

i = 1

repeat {
  F_theta_0 = as.matrix(cbind(
    eval(
      df_4.d_theta_0,
      list(
        x = my_data$x,
        theta_0 = theta_vec[1, i],
        theta_1 = theta_vec[2, i],
        theta_2 = theta_vec[3, i]
      )
    ),
    eval(
      df_4.d_theta_1,
      list(
        x = my_data$x,
        theta_0 = theta_vec[1, i],
        theta_1 = theta_vec[2, i],
        theta_2 = theta_vec[3, i]
      )
    ),
    eval(
      df_4.d_theta_2,
      list(
        x = my_data$x,
        theta_0 = theta_vec[1, i],
        theta_1 = theta_vec[2, i],
        theta_2 = theta_vec[3, i]
      )
    )
  ))
}

```

```

list(
  x = my_data$x,
  theta_0 = theta_vec[1, i],
  theta_1 = theta_vec[2, i],
  theta_2 = theta_vec[3, i]
)
)
))
delta[, i] = (solve(t(F_theta_0) %*% F_theta_0)) %*% t(F_theta_0) %*% (my_data$y -
theta_vec = cbind(theta_vec, matrix(NA, nrow = 3, ncol = 1))
theta_vec[, i+1] = theta_vec[, i] + delta[, i]
i = i + 1

f_theta = cbind(f_theta, as.matrix(eval(
  f_4,
  list(
    x = my_data$x,
    theta_0 = theta_vec[1, i],
    theta_1 = theta_vec[2, i],
    theta_2 = theta_vec[3, i]
)
)))
delta = cbind(delta, matrix(NA, nrow = 3, ncol = 1))

#convergence criteria based on SSE
if (abs(sum((my_data$y - f_theta[,i])^2)-sum((my_data$y - f_theta[,i-1])^2))/(sum(
  break
}
}
delta
#> [1,] 2.811840e+02 -0.03929013 0.43160654 0.6904856 0.6746748 0.4056460
#> [2,] -2.846545e+02 0.03198446 -0.16403964 -0.2895487 -0.2933345 -0.1734087
#> [3,] -1.804567e-05 0.01530258 0.05137285 0.1183271 0.1613129 0.1160404
#> [4,] 0.09517681 NA
#> [5,] -0.03928239 NA
#> [6,] 0.03004911 NA
theta_vec
#> [1,] -279.8087974 1.37521388 1.33592375 1.76753029 2.4580158 3.1326907
#> [2,] 284.2765957 -0.37788712 -0.34590266 -0.50994230 -0.7994910 -1.0928255
#> [3,] 0.0119632 0.01194515 0.02724773 0.07862059 0.1969477 0.3582607
#> [4,] 3.5383367 3.6335135

```

```
#> [2,] -1.2662342 -1.3055166
#> [3,] 0.4743011 0.5043502

head(f_theta)
#>      [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]     [,8]
#> [1,] -273.8482 1.355410 1.297194 1.633802 2.046023 2.296554 2.389041 2.404144
#> [2,] -209.0859 1.268192 1.216738 1.514575 1.863098 2.059505 2.126009 2.135969
#> [3,] -190.3323 1.242916 1.193433 1.480136 1.810629 1.992095 2.051603 2.060202
#> [4,] -177.1891 1.225196 1.177099 1.456024 1.774000 1.945197 1.999945 2.007625
#> [5,] -148.5872 1.186618 1.141549 1.403631 1.694715 1.844154 1.888953 1.894730
#> [6,] -119.9585 1.147980 1.105961 1.351301 1.615968 1.744450 1.779859 1.783866

# estimate sigma^2

sigma2 = 1 / (nrow(my_data) - 3) * (t(my_data$y - (f_theta[, ncol(f_theta)]))) %*%
  (my_data$y - (f_theta[, ncol(f_theta)])) # p = 3
sigma2
#>      [,1]
#> [1,] 0.0801686
```

After 8 iterations, my function has converged. And objective function value at convergence is

```
sum((my_data$y - f_theta[,i])^2)
#> [1] 19.80165
```

and the parameters of θ s are

```
theta_vec[,ncol(theta_vec)]
#> [1] 3.6335135 -1.3055166 0.5043502
```

and the asymptotic variance covariance matrix is

```
as.numeric(sigma2)*as.matrix(solve(crossprod(F_theta_0)))
#>      [,1]     [,2]     [,3]
#> [1,] 0.11552571 -0.04817428 0.02685848
#> [2,] -0.04817428  0.02100861 -0.01158212
#> [3,]  0.02685848 -0.01158212  0.00703916
```

Issue that I encounter in this problem was that it was very sensitive to starting values. when I tried the value of 1 for all θ s, I have vastly different parameter estimates. Then, I try to use the model interpretation to try to find reasonable starting values.

Check with predefined function in nls

```
nlin_4 = nls(y ~ mod_4(theta_0,theta_1, theta_2, x), start = list(theta_0=-279.80879738,
nlin_4
#> Nonlinear regression model
#>   model: y ~ mod_4(theta_0, theta_1, theta_2, x)
#>   data: my_data
#>   theta_0 theta_1 theta_2
#>   3.6359 -1.3064  0.5053
#>   residual sum-of-squares: 19.8
#>
#> Number of iterations to convergence: 9
#> Achieved convergence tolerance: 2.294e-07
```

Chapter 7

Generalized Linear Models

Even though we call it generalized linear model, it is still under the paradigm of non-linear regression, because the form of the regression model is non-linear. The name generalized linear model derived from the fact that we have \mathbf{x}'_i (which is linear form) in the model.

7.1 Logistic Regression

$$p_i = f(\mathbf{x}_i; \beta) = \frac{\exp(\mathbf{x}'_i)}{1 + \exp(\mathbf{x}'_i)}$$

Equivalently,

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 + p_i}\right) = \mathbf{x}'_i$$

where $\frac{p_i}{1 + p_i}$ is the **odds**.

In this form, the model is specified such that **a function of the mean response is linear**. Hence, **Generalized Linear Models**

The likelihood function

$$L(p_i) = \prod_{i=1}^n p_i^{Y_i} (1 - p_i)^{1 - Y_i}$$

where $p_i = \frac{\mathbf{x}'_i}{1 + \mathbf{x}'_i}$ and $1 - p_i = (1 + \exp(\mathbf{x}'_i))^{-1}$

Hence, our objective function is

$$Q(\beta) = \log(L(\beta)) = \sum_{i=1}^n Y_i \mathbf{x}'_i - \sum_{i=1}^n \log(1 + \exp(\mathbf{x}'_i))$$

we could maximize this function numerically using the optimization method above, which allows us to find numerical MLE for $\hat{\beta}$. Then we can use the standard asymptotic properties of MLEs to make inference.

Property of MLEs is that parameters are asymptotically unbiased with sample variance-covariance matrix given by the **inverse Fisher information matrix**

$$\hat{\beta} \sim AN(\beta, [\mathbf{I}(\beta)]^{-1})$$

where the **Fisher Information matrix**, $\mathbf{I}(\beta)$ is

$$\begin{aligned} \mathbf{I}(\beta) &= E\left[\frac{\partial \log(L(\beta))}{\partial \beta} \frac{\partial \log(L(\beta))}{\partial \beta'}\right] \\ &= E\left[\left(\frac{\partial \log(L(\beta))}{\partial \beta_i} \frac{\partial \log(L(\beta))}{\partial \beta_j}\right)_{ij}\right] \end{aligned}$$

Under **regularity conditions**, this is equivalent to the negative of the expected value of the Hessian Matrix

$$\begin{aligned} \mathbf{I}(\beta) &= -E\left[\frac{\partial^2 \log(L(\beta))}{\partial \beta \partial \beta'}\right] \\ &= -E\left[\left(\frac{\partial^2 \log(L(\beta))}{\partial \beta_i \partial \beta_j}\right)_{ij}\right] \end{aligned}$$

Example:

$$x'_i \beta = \beta_0 + \beta_1 x_i$$

$$-\frac{\partial^2 \ln(L(\beta))}{\partial \beta_0^2} = \sum_{i=1}^n \frac{\exp(x'_i \beta)}{1 + \exp(x'_i \beta)} - \left[\frac{\exp(x'_i \beta)}{1 + \exp(x'_i \beta)} \right]^2 = \sum_{i=1}^n p_i(1-p_i) - \frac{\partial^2 \ln(L(\beta))}{\partial \beta_1^2} = \sum_{i=1}^n \frac{x_i^2 \exp(x'_i \beta)}{1 + \exp(x'_i \beta)} - \left[\frac{\exp(x'_i \beta)}{1 + \exp(x'_i \beta)} \right]^2$$

Hence,

$$\mathbf{I}(\beta) = \begin{bmatrix} \sum_i p_i(1-p_i) & \sum_i x_i p_i(1-p_i) \\ \sum_i x_i p_i(1-p_i) & \sum_i x_i^2 p_i(1-p_i) \end{bmatrix}$$

Inference

Likelihood Ratio Tests

To formulate the test, let $\beta = [\beta'_1, \beta'_2]'$. If you are interested in testing a hypothesis about β_1 , then we leave β_2 unspecified (called **nuisance parameters**). β_1 and β_2 can either a **vector** or **scalar**, or β_2 can be null.

Example: $H_0 : \beta_1 = \beta_{1,0}$ (where $\beta_{1,0}$ is specified) and $\hat{\beta}_{2,0}$ be the MLE of β_2 under the restriction that $\beta_1 = \beta_{1,0}$. The likelihood ratio test statistic is

$$-2 \log \Lambda = -2[\log(L(\beta_{1,0}, \hat{\beta}_{2,0})) - \log(L(\hat{\beta}_1, \hat{\beta}_2))]$$

where

- the first term is the value fo the likelihood for the fitted restricted model
- the second term is the likelihood value of the fitted unrestricted model

Under the null,

$$-2 \log \Lambda \sim \chi_v^2$$

where v is the dimension of β_1

We reject the null when $-2 \log \Lambda > \chi_{v,1-\alpha}^2$

Wald Statistics

Based on

$$\hat{\beta} \sim AN(\beta, [\mathbf{I}(\beta)^{-1}])$$

$$H_0 : \mathbf{L}\hat{\beta} = 0$$

where \mathbf{L} is a $q \times p$ matrix with q linearly independent rows. Then

$$W = (\mathbf{L}')([\mathbf{I}(\beta)]^{-1}\mathbf{L}')^{-1}(\mathbf{L})$$

under the null hypothesis

Confidence interval

$$\hat{\beta}_i \pm 1.96 \hat{s}_{ii}^2$$

where \hat{s}_{ii}^2 is the i-th diagonal of $[\mathbf{I}(\beta)]^{-1}$

If you have

- large sample size, the likelihood ratio and Wald tests have similar results.
- small sample size, the likelihood ratio test is better.

Logistic Regression: Interpretation of β

For single regressor, the model is

$$\text{logit}\{\hat{p}_{x_i}\} \equiv \text{logit}(\hat{p}_i) = \log\left(\frac{\hat{p}_i}{1 - \hat{p}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

When $x = x_i + 1$

$$\text{logit}\{\hat{p}_{x_i+1}\} = \hat{\beta}_0 + \hat{\beta}(x_i + 1) = \text{logit}\{\hat{p}_{x_i}\} + \hat{\beta}_1$$

Then,

$$\text{logit}\{\hat{p}_{x_i+1}\} - \text{logit}\{\hat{p}_{x_i}\} = \log\{\text{odds}[\hat{p}_{x_i+1}]\} - \log\{\text{odds}[\hat{p}_{x_i}]\} = \log\left(\frac{\text{odds}[\hat{p}_{x_i+1}]}{\text{odds}[\hat{p}_{x_i}]}\right) = \hat{\beta}_1$$

and

$$\exp(\hat{\beta}_1) = \frac{\text{odds}[\hat{p}_{x_i+1}]}{\text{odds}[\hat{p}_{x_i}]}$$

the estimated **odds ratio**

the estimated odds ratio, when there is a difference of c units in the regressor x , is $\exp(c\hat{\beta}_1)$. When there are multiple covariates, $\exp(\hat{\beta}_k)$ is the estimated odds ratio for the variable x_k , assuming that all of the other variables are held constant.

Inference on the Mean Response

Let $x_h = (1, x_{h1}, \dots, x_{hp-1})'$. Then

$$\hat{p}_h = \frac{\exp(\mathbf{x}_h' \hat{\beta})}{1 + \exp(\mathbf{x}_h' \hat{\beta})}$$

and $s^2(\hat{p}_h) = \mathbf{x}_h' [\mathbf{I}(\hat{\beta})]^{-1} \mathbf{x}_h$

For new observation, we can have a cutoff point to decide whether $y = 0$ or 1 .

7.1.1 Application

```
library(kableExtra)
library(dplyr)
library(pscl)
library(ggplot2)
library(faraway)
library(nnet)
library(agridat)
library(nlstools)
```

Logistic Regression

$x \sim Unif(-0.5, 2.5)$. Then $\eta = 0.5 + 0.75x$

```
set.seed(23) #set seed for reproducibility
x <- runif(1000,min = -0.5,max = 2.5)
eta1 <- 0.5 + 0.75*x
```

Passing η 's into the inverse-logit function, we get

$$p = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

where $p \in [0, 1]$

Then, we generate $y \sim Bernoulli(p)$

```
p <- exp(eta1)/(1+exp(eta1))
y <- rbinom(1000,1,p)
BinData <- data.frame(X = x, Y = y)
```

Model Fit

```
Logistic_Model <- glm(formula = Y ~ X,
                      family = binomial, # family = specifies the response distribution
                      data = BinData)
summary(Logistic_Model)
#>
#> Call:
#> glm(formula = Y ~ X, family = binomial, data = BinData)
#>
#> Deviance Residuals:
#>      Min        1Q    Median        3Q       Max
#> -2.2317   0.4153   0.5574   0.7922   1.1469
#>
```

```
#> Coefficients:
#>             Estimate Std. Error z value Pr(>|z|)
#> (Intercept) 0.46205   0.10201  4.530 5.91e-06 ***
#> X            0.78527   0.09296  8.447 < 2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#> Null deviance: 1106.7 on 999 degrees of freedom
#> Residual deviance: 1027.4 on 998 degrees of freedom
#> AIC: 1031.4
#>
#> Number of Fisher Scoring iterations: 4
nlstools::confint2(Logistic_Model)
#>           2.5 %    97.5 %
#> (Intercept) 0.2618709 0.6622204
#> X           0.6028433 0.9676934
OddsRatio <- coef(Logistic_Model) %>% exp
OddsRatio
#> (Intercept)          X
#>     1.587318      2.192995
```

Based on the odds ratio, when

- $x = 0$, the odds of success of 1.59
- $x = 1$, the odds of success increase by a factor of 2.19 (i.e., 119.29% increase).

Deviance Tests

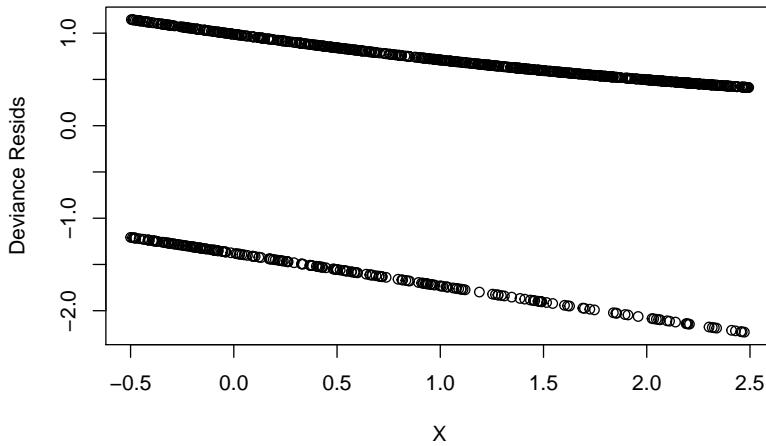
H_0 : No variables are related to the response (i.e., model with just the intercept) H_1 : at least one variable is related to the response

```
Test_Dev = Logistic_Model>null.deviance - Logistic_Model$deviance
p_val_dev <- 1-pchisq(q = Test_Dev, df = 1)
```

Since we see the p-value of 0, we reject the null that no variables are related to the response

Deviance residuals

```
Logistic_Resids <- residuals(Logistic_Model, type = "deviance")
plot(
  y = Logistic_Resids,
  x = BinData$X,
  xlab = 'X',
  ylab = 'Deviance Resids'
)
```



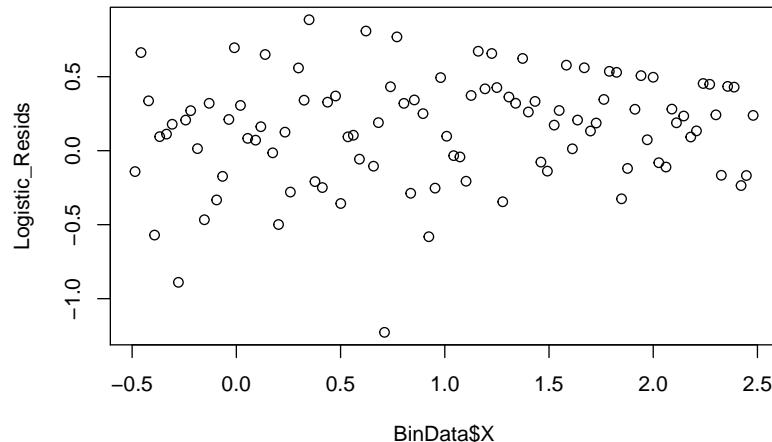
However, this plot is not informative. Hence, we can see the residuals plots that are grouped into bins based on prediction values.

```
plot_bin <- function(Y,
                      X,
                      bins = 100,
                      return.DF = FALSE) {
  Y_Name <- deparse(substitute(Y))
  X_Name <- deparse(substitute(X))
  Binned_Plot <- data.frame(Plot_Y = Y, Plot_X = X)
  Binned_Plot$bin <-
    cut(Binned_Plot$Plot_X, breaks = bins) %>% as.numeric
  Binned_Plot_summary <- Binned_Plot %>%
    group_by(bin) %>%
    summarise(
      Y_ave = mean(Plot_Y),
      X_ave = mean(Plot_X),
```

```

Count = n()
) %>% as.data.frame
plot(
  y = Binned_Plot_summary$Y_ave,
  x = Binned_Plot_summary$X_ave,
  ylab = Y_Name,
  xlab = X_Name
)
if (return.DF)
  return(Binned_Plot_summary)
}
plot_bin(Y = Logistic_Resids,
         X = BinData$X,
         bins = 100)

```

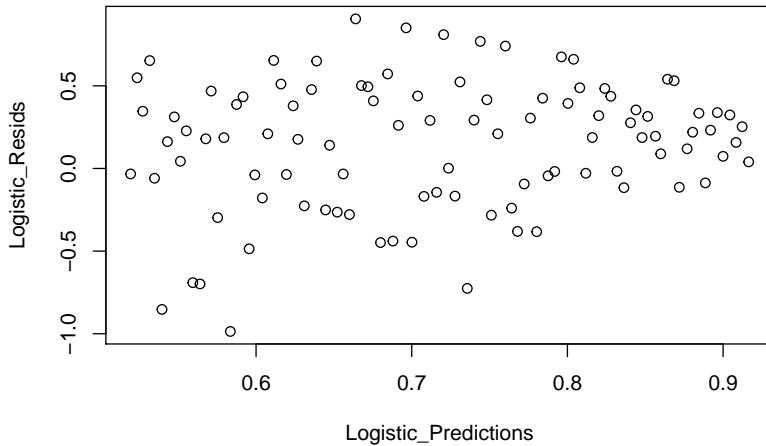


We can also see the predicted value against the residuals.

```

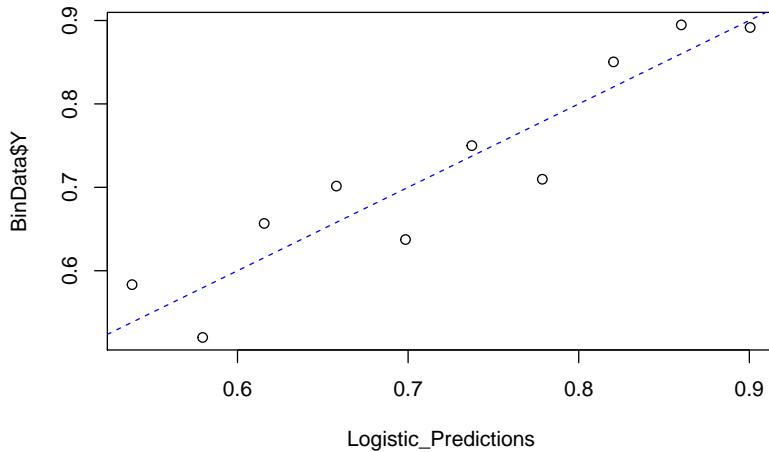
Logistic_Predictions <- predict(Logistic_Model, type = "response")
plot_bin(Y = Logistic_Resids, X = Logistic_Predictions, bins = 100)

```



We can also look at a binned plot of the logistic prediction versus the true category

```
NumBins <- 10
Binned_Data <- plot_bin(
  Y = BinData$Y,
  X = Logistic_Predictions,
  bins = NumBins,
  return.DF = TRUE
)
Binned_Data
#>   bin      Y_ave      X_ave Count
#> 1  1  0.5833333  0.5382095    72
#> 2  2  0.5200000  0.5795887    75
#> 3  3  0.6567164  0.6156540    67
#> 4  4  0.7014925  0.6579674    67
#> 5  5  0.6373626  0.6984765    91
#> 6  6  0.7500000  0.7373341    72
#> 7  7  0.7096774  0.7786747    93
#> 8  8  0.8503937  0.8203819   127
#> 9  9  0.8947368  0.8601232   133
#> 10 10  0.8916256  0.9004734   203
abline(0, 1, lty = 2, col = 'blue')
```

**Formal deviance test****Hosmer-Lemeshow test**

Null hypothesis: the observed events match the expected events

$$X_{HL}^2 = \sum_{j=1}^J \frac{(y_j - m_j \hat{p}_j)^2}{m_j \hat{p}_j (1 - \hat{p}_j)}$$

where

- within the j-th bin, y_j is the number of successes
- m_j = number of observations
- \hat{p}_j = predicted probability

Under the null hypothesis, $X_{HLL}^2 \sim \chi_{J-1}^2$

```
HL_BinVals <-
  (Binned_Data$Count * Binned_Data$Y_ave - Binned_Data$Count * Binned_Data$X_ave) ^
  2 /
  Binned_Data$Count * Binned_Data$X_ave * (1 - Binned_Data$X_ave)
HLpval <-
  pchisq(q = sum(HL_BinVals),
         df = NumBins,
         lower.tail = FALSE)
HLpval
#> [1] 0.9999989
```

Since p-value = 0.99, we do not reject the null hypothesis (i.e., the model is fitting well).

7.2 Probit Regression

$$E(Y_i) = p_i = \Phi(\mathbf{x}'_i)$$

where $\Phi()$ is the CDF of a $N(0,1)$ random variable.

Other models (e.g, t-distribution; log-log; I complimentary log-log)

We let $Y_i = 1$ success, $Y_i = 0$ no success. We assume $Y \sim Ber$ and $p_i = P(Y_i = 1)$, the success probability. We consider a logistic regression with the response function $logit(p_i) = \mathbf{x}'_i \beta$

Confusion matrix

		Predicted	
		1	0
Truth	1	True Positive (TP)	False Negative (FN)
	0	False Positive (FP)	True Negative (TN)

Sensitivity: ability to identify positive results

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity: ability to identify negative results

$$\text{Specificity} = \frac{TN}{TN + FP}$$

False positive rate: Type I error (1- specificity)

$$\text{False Positive Rate} = \frac{FP}{TN + FP}$$

False Negative Rate: Type II error (1-sensitivity)

$$\text{False Negative Rate} = \frac{FN}{TP + FN}$$

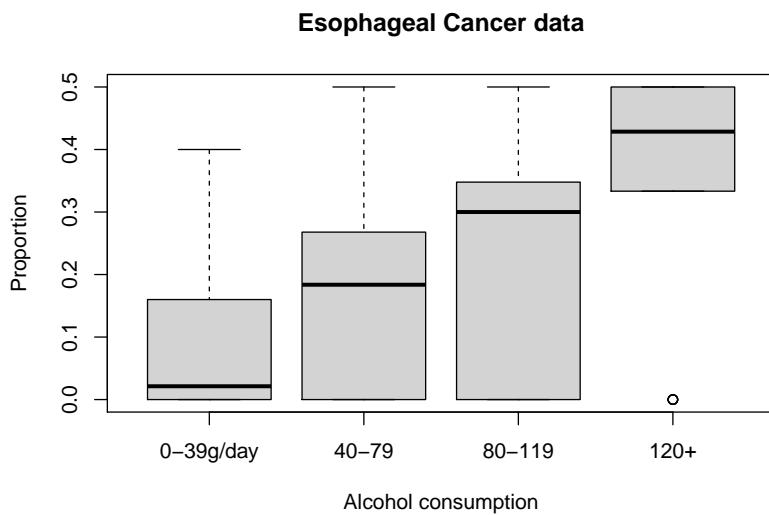
	Predicted	
Truth	1	0
1	Sensitivity	False Negative Rate
0	False Positive Rate	Specificity

7.3 Binomial Regression

Binomial

Here, cancer case = successes, and control case = failures.

```
data("esoph")
head(esoph, n = 3)
#>   agegp      alcgp      tobgp ncases ncontrols
#> 1 25-34 0-39g/day 0-9g/day      0        40
#> 2 25-34 0-39g/day    10-19      0        10
#> 3 25-34 0-39g/day    20-29      0        6
plot(
  esoph$ncases / (esoph$ncases + esoph$ncontrols) ~ esoph$alcgp,
  ylab = "Proportion",
  xlab = 'Alcohol consumption',
  main = 'Esophageal Cancer data'
)
```



```

class(esoph$agegp) <- "factor"
class(esoph$alcp) <- "factor"
class(esoph$tobgp) <- "factor"

# only the alcohol consumption as a predictor
model <- glm(cbind(ncases, ncontrols) ~ alcgp, data = esoph, family = binomial)
summary(model)
#>
#> Call:
#> glm(formula = cbind(ncases, ncontrols) ~ alcgp, family = binomial,
#>       data = esoph)
#>
#> Deviance Residuals:
#>      Min        1Q    Median        3Q       Max
#> -3.6629  -1.0478  -0.0081   0.6307   3.0296
#>
#> Coefficients:
#>             Estimate Std. Error z value Pr(>|z|)
#> (Intercept) -2.6610     0.1921 -13.854 < 2e-16 ***
#> alcgp40-79   1.1064     0.2303   4.804 1.56e-06 ***
#> alcgp80-119   1.6656     0.2525   6.597 4.20e-11 ***
#> alcgp120+    2.2630     0.2721   8.317 < 2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#> Null deviance: 227.24 on 87 degrees of freedom
#> Residual deviance: 138.79 on 84 degrees of freedom
#> AIC: 294.27
#>
#> Number of Fisher Scoring iterations: 5

#Coefficient Odds
coefficients(model) %>% exp
#> (Intercept) alcgp40-79 alcgp80-119 alcgp120+
#> 0.06987952 3.02331229 5.28860570 9.61142563
deviance(model)/df.residual(model)
#> [1] 1.652253
model$aic
#> [1] 294.27

# alcohol consumption and age as predictors
better_model <- glm(cbind(ncases, ncontrols) ~ agegp + alcgp, data = esoph, family = binomial)

```

```

summary(better_model)
#>
#> Call:
#> glm(formula = cbind(ncases, ncontrols) ~ agegp + alcgp, family = binomial,
#>       data = esoph)
#>
#> Deviance Residuals:
#>      Min        1Q    Median        3Q       Max
#> -1.8979  -0.5592  -0.1995   0.5029   2.6250
#>
#> Coefficients:
#>             Estimate Std. Error z value Pr(>|z|)
#> (Intercept) -5.6180    1.0217 -5.499 3.82e-08 ***
#> agegp35-44   1.5376    1.0646  1.444 0.148669
#> agegp45-54   2.9470    1.0217  2.884 0.003922 **
#> agegp55-64   3.3116    1.0172  3.255 0.001132 **
#> agegp65-74   3.5774    1.0209  3.504 0.000458 ***
#> agegp75+     3.5858    1.0620  3.377 0.000734 ***
#> alcgp40-79   1.1392    0.2367  4.814 1.48e-06 ***
#> alcgp80-119  1.4951    0.2600  5.749 8.97e-09 ***
#> alcgp120+    2.2228    0.2843  7.820 5.29e-15 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#> Null deviance: 227.241 on 87 degrees of freedom
#> Residual deviance: 64.572 on 79 degrees of freedom
#> AIC: 230.05
#>
#> Number of Fisher Scoring iterations: 6

```

```

better_model$aic #smaller AIC is better
#> [1] 230.0526
coefficients(better_model) %>% exp
#> (Intercept) agegp35-44 agegp45-54 agegp55-64 agegp65-74 agegp75+
#> 0.003631855 4.653273722 19.047899816 27.428640745 35.780787582 36.082010052
#> alcgp40-79 alcgp80-119 alcgp120+
#> 3.124334222 4.459579378 9.233256747
pchisq(
  q = model$deviance - better_model$deviance,
  df = model$df.residual - better_model$df.residual,
  lower = FALSE
)
#> [1] 1.354906e-14

```

```

# specify link function as probit
Prob_better_model <- glm(
  cbind(ncases, ncontrols) ~ agegp + alcgp,
  data = esoph,
  family = binomial(link = probit)
)
summary(Prob_better_model)
#>
#> Call:
#> glm(formula = cbind(ncases, ncontrols) ~ agegp + alcgp, family = binomial(link = probit),
#>       data = esoph)
#>
#> Deviance Residuals:
#>      Min        1Q    Median        3Q        Max
#> -1.8676  -0.5938  -0.1802   0.4852   2.6056
#>
#> Coefficients:
#>             Estimate Std. Error z value Pr(>|z|)
#> (Intercept) -2.9800    0.4291 -6.945 3.79e-12 ***
#> agegp35-44   0.6991    0.4491  1.557 0.119520
#> agegp45-54   1.4212    0.4292  3.311 0.000929 ***
#> agegp55-64   1.6512    0.4262  3.874 0.000107 ***
#> agegp65-74   1.8039    0.4297  4.198 2.69e-05 ***
#> agegp75+     1.8025    0.4613  3.908 9.32e-05 ***
#> alcgp40-79   0.6224    0.1247  4.990 6.03e-07 ***
#> alcgp80-119  0.8256    0.1418  5.823 5.80e-09 ***
#> alcgp120+    1.2839    0.1596  8.043 8.77e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#> Null deviance: 227.241 on 87 degrees of freedom
#> Residual deviance: 61.938 on 79 degrees of freedom
#> AIC: 227.42
#>
#> Number of Fisher Scoring iterations: 6

```

7.4 Poisson Regression

From the Poisson distribution

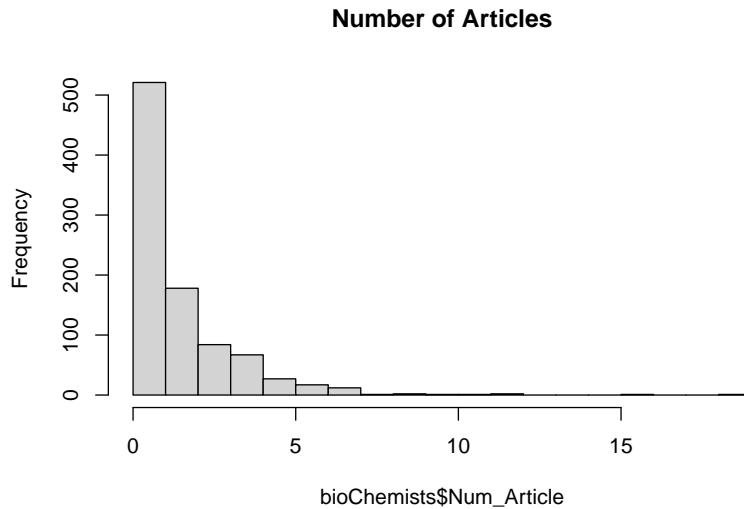
$$f(Y_i) = \frac{\mu_i^{Y_i} \exp(-\mu_i)}{Y_i!}, Y_i = 0, 1, \dots E(Y_i) = \mu_i \text{var}(Y_i) = \mu_i$$

which is a natural distribution for counts. We can see that the variance is a function of the mean. If we let $\mu_i = f(\mathbf{x}_i)$, it would be similar to Logistic Regression since we can choose $f()$ as $\mu_i = \mathbf{x}'_i$, $\mu_i = \exp(\mathbf{x}'_i)$, $\mu_i = \log(\mathbf{x}'_i)$

7.4.1 Application

Count Data and Poisson regression

```
data(bioChemists, package = "pscl")
bioChemists <- bioChemists %>%
  rename(
    Num_Article = art, #articles in last 3 years of PhD
    Sex = fem, #coded 1 if female
    Married = mar, #coded 1 if married
    Num_Kid5 = kid5, #number of children under age 6
    PhD_Quality = phd, #prestige of PhD program
    Num_MentArticle = ment #articles by mentor in last 3 years
  )
hist(bioChemists$Num_Article, breaks = 25, main = 'Number of Articles')
```



```
Poisson_Mod <- glm(Num_Article ~ ., family=poisson, bioChemists)
summary(Poisson_Mod)

#>
#> Call:
#> glm(formula = Num_Article ~ ., family = poisson, data = bioChemists)
#>
#> Deviance Residuals:
#>      Min        1Q    Median        3Q        Max
#> -3.5672  -1.5398  -0.3660   0.5722   5.4467
#>
#> Coefficients:
#>             Estimate Std. Error z value Pr(>|z|)
#> (Intercept) 0.304617  0.102981  2.958  0.0031 **
#> SexWomen     -0.224594  0.054613 -4.112 3.92e-05 ***
#> MarriedMarried 0.155243  0.061374  2.529  0.0114 *
#> Num_Kid5     -0.184883  0.040127 -4.607 4.08e-06 ***
#> PhD_Quality    0.012823  0.026397  0.486  0.6271
#> Num_MentArticle 0.025543  0.002006 12.733 < 2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for poisson family taken to be 1)
#>
#> Null deviance: 1817.4 on 914 degrees of freedom
#> Residual deviance: 1634.4 on 909 degrees of freedom
#> AIC: 3314.1
#>
#> Number of Fisher Scoring iterations: 5
```

Residual of 1634 with 909 df isn't great.

We see Pearson χ^2

```
Predicted_Means <- predict(Poisson_Mod,type = "response")
X2 <- sum((bioChemists$Num_Article - Predicted_Means)^2/Predicted_Means)
X2
#> [1] 1662.547
pchisq(X2,Poisson_Mod$df.residual, lower.tail = FALSE)
#> [1] 7.849882e-47
```

With interaction terms, there are some improvements

```
Poisson_Mod_All2way <- glm(Num_Article ~ .^2, family=poisson, bioChemists)
Poisson_Mod_All3way <- glm(Num_Article ~ .^3, family=poisson, bioChemists)
```

Consider the $\hat{\phi} = \frac{\text{deviance}}{df}$

```
Poisson_Mod$deviance / Poisson_Mod$df.residual
#> [1] 1.797988
```

This is evidence for over-dispersion. Likely cause is missing variables. And remedies could either be to include more variables or consider random effects.

A quick fix is to force the Poisson Regression to include this value of ϕ , and this model is called “Quasi-Poisson”.

```
phi_hat = Poisson_Mod$deviance/Poisson_Mod$df.residual
summary(Poisson_Mod, dispersion = phi_hat)
#>
#> Call:
#> glm(formula = Num_Article ~ ., family = poisson, data = bioChemists)
#>
#> Deviance Residuals:
#>      Min        1Q    Median        3Q       Max
#> -3.5672  -1.5398  -0.3660   0.5722   5.4467
#>
#> Coefficients:
#>              Estimate Std. Error z value Pr(>|z|)
#> (Intercept) 0.30462   0.13809   2.206  0.02739 *
#> SexWomen    -0.22459   0.07323  -3.067  0.00216 **
#> MarriedMarried 0.15524   0.08230   1.886  0.05924 .
#> Num_Kid5     -0.18488   0.05381  -3.436  0.00059 ***
#> PhD_Quality   0.01282   0.03540   0.362  0.71715
#> Num_MentArticle 0.02554   0.00269   9.496 < 2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for poisson family taken to be 1.797988)
#>
#> Null deviance: 1817.4 on 914 degrees of freedom
#> Residual deviance: 1634.4 on 909 degrees of freedom
#> AIC: 3314.1
#>
#> Number of Fisher Scoring iterations: 5
```

Or directly rerun the model as

```
quasiPoisson_Mod <- glm(Num_Article ~ ., family=quasipoisson, bioChemists)
```

Quasi-Poisson is not recommended, but Negative Binomial Regression that has an extra parameter to account for over-dispersion is.

7.5 Negative Binomial Regression

```

library(MASS)
NegBinom_Mod <- MASS::glm.nb(Num_Article ~ ., bioChemists)
summary(NegBinom_Mod)
#>
#> Call:
#> MASS::glm.nb(formula = Num_Article ~ ., data = bioChemists, init.theta = 2.264387695,
#>     link = log)
#>
#> Deviance Residuals:
#>      Min        1Q    Median        3Q       Max
#> -2.1678  -1.3617  -0.2806   0.4476   3.4524
#>
#> Coefficients:
#>             Estimate Std. Error z value Pr(>|z|)
#> (Intercept) 0.256144  0.137348  1.865 0.062191 .
#> SexWomen    -0.216418  0.072636 -2.979 0.002887 **
#> MarriedMarried 0.150489  0.082097  1.833 0.066791 .
#> Num_Kid5    -0.176415  0.052813 -3.340 0.000837 ***
#> PhD_Quality  0.015271  0.035873  0.426 0.670326
#> Num_MentArticle 0.029082  0.003214  9.048 < 2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for Negative Binomial(2.264) family taken to be 1)
#>
#> Null deviance: 1109.0 on 914 degrees of freedom
#> Residual deviance: 1004.3 on 909 degrees of freedom
#> AIC: 3135.9
#>
#> Number of Fisher Scoring iterations: 1
#>
#>
#>          Theta:  2.264
#> Std. Err.:  0.271
#>
#> 2 x log-likelihood:  -3121.917

```

We can see the dispersion is 2.264 with SE = 0.271, which is significantly different from 1, indicating overdispersion. Check Over-Dispersion for more detail

7.6 Multinomial

If we have more than two categories or groups that we want to model relative to covariates (e.g., we have observations $i = 1, \dots, n$ and groups/ covariates $j = 1, 2, \dots, J$), multinomial is our candidate model

Let

- p_{ij} be the probability that the i -th observation belongs to the j -th group
- Y_{ij} be the number of observations for individual i in group j ; An individual will have observations $Y_{i1}, Y_{i2}, \dots, Y_{iJ}$
- assume the probability of observing this response is given by a multinomial distribution in terms of probabilities p_{ij} , where $\sum_{j=1}^J p_{ij} = 1$. For interpretation, we have a baseline category $p_{i1} = 1 - \sum_{j=2}^J p_{ij}$

The link between the mean response (probability) p_{ij} and a linear function of the covariates

$$\eta_{ij} = \mathbf{x}'_i j = \log \frac{p_{ij}}{p_{i1}}, j = 2, \dots, J$$

We compare p_{ij} to the baseline p_{i1} , suggesting

$$p_{ij} = \frac{\exp(\eta_{ij})}{1 + \sum_{i=2}^J \exp(\eta_{ij})}$$

which is known as **multinomial logistic** model.

Note:

- Softmax coding for multinomial logistic regression: rather than selecting a baseline class, we treat all K class symmetrically - equally important (no baseline).

$$P(Y = k | X = x) = \frac{\exp(\beta_{k1} + \dots + \beta_{kp}x_p)}{\sum_{l=1}^K \exp(\beta_{l0} + \dots + \beta_{lp}x_p)}$$

then the log odds ratio between k -th and k' -th classes is

$$\log\left(\frac{P(Y = k | X = x)}{P(Y = k' | X = x)}\right) = (\beta_{k0} - \beta_{k'0}) + \dots + (\beta_{kp} - \beta_{k'p})x_p$$

```
library(faraway)
library(dplyr)
data(nes96, package="faraway")
head(nes96, 3)
#>   popul TVnews selfLR ClinLR DoleLR      PID age   educ income   vote
#> 1     0    7 extCon extLib   Con strRep  36   HS $3Kminus   Dole
#> 2   190    1 sliLib sliLib sliCon weakDem 20 Coll $3Kminus Clinton
#> 3    31    7   Lib   Lib   Con weakDem 24 BAdeg $3Kminus Clinton
```

We try to understand their political strength

```
table(nes96$PID)
#>
#>   strDem weakDem indDem indind indRep weakRep strRep
#>   200     180    108     37     94    150    175
nes96$Political_Strength <- NA
nes96$Political_Strength[nes96$PID %in% c("strDem", "strRep")] <-
  "Strong"
nes96$Political_Strength[nes96$PID %in% c("weakDem", "weakRep")] <-
  "Weak"
nes96$Political_Strength[nes96$PID %in% c("indDem", "indind", "indRep")] <-
  "Neutral"
nes96 %>% group_by(Political_Strength) %>% summarise(Count = n())
#> # A tibble: 3 x 2
#>   Political_Strength Count
#>   <chr>              <int>
#> 1 Neutral             239
#> 2 Strong              375
#> 3 Weak                330
```

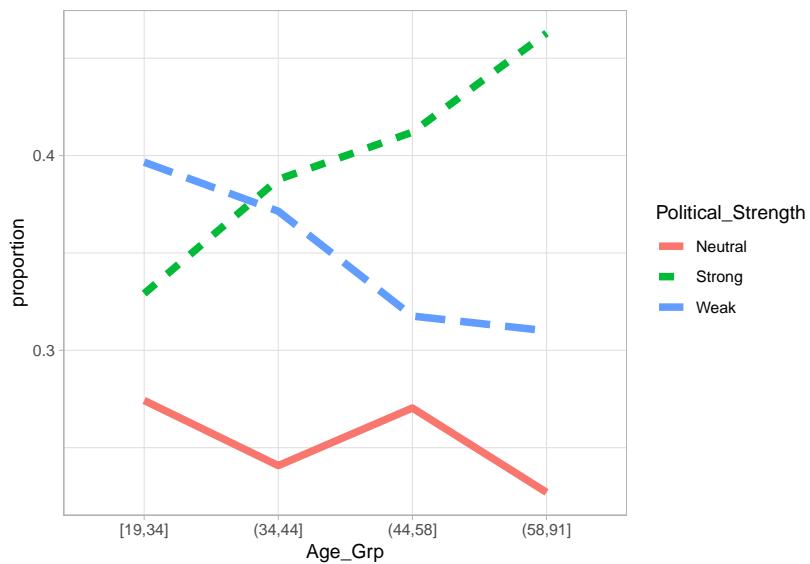
visualize the political strength variable

```
library(ggplot2)
Plot_DF <- nes96 %>%
  mutate(Age_Grp = cut_number(age, 4)) %>%
  group_by(Age_Grp, Political_Strength) %>%
  summarise(count = n()) %>%
  group_by(Age_Grp) %>%
  mutate(etotal = sum(count), proportion = count / etotal)
Age_Plot <- ggplot(
  Plot_DF,
  aes(
    x = Age_Grp,
    y = proportion,
```

```

group = Political_Strength,
linetype = Political_Strength,
color = Political_Strength
)
) +
  geom_line(size = 2)
Age_Plot

```



Fit the multinomial logistic model:

model political strength as a function of age and education

```

library(nnet)
Multinomial_Model <-
  multinom(Political_Strength ~ age + educ, nes96, trace = F)
summary(Multinomial_Model)
#> Call:
#> multinom(formula = Political_Strength ~ age + educ, data = nes96,
#>   trace = F)
#>
#> Coefficients:
#> (Intercept)      age      educ.L      educ.Q      educ.C      educ^4
#> Strong -0.08788729  0.010700364 -0.1098951 -0.2016197 -0.1757739 -0.02116307
#> Weak   0.51976285 -0.004868771 -0.1431104 -0.2405395 -0.2411795  0.18353634
#>      educ^5      educ^6
#> Strong -0.16643777 -0.1359449

```

```
#> Weak -0.1489030 -0.2173144
#>
#> Std. Errors:
#> (Intercept) age educ.L educ.Q educ.C educ^4
#> Strong 0.3017034 0.005280743 0.4586041 0.4318830 0.3628837 0.2964776
#> Weak 0.3097923 0.005537561 0.4920736 0.4616446 0.3881003 0.3169149
#> educ^5 educ^6
#> Strong 0.2515012 0.2166774
#> Weak 0.2643747 0.2199186
#>
#> Residual Deviance: 2024.596
#> AIC: 2056.596
```

Alternatively, stepwise model selection based AIC

```
Multinomial_Step <- step(Multinomial_Model,trace = 0)
#> trying - age
#> trying - educ
#> trying - age
Multinomial_Step
#> Call:
#> multinom(formula = Political_Strength ~ age, data = nes96, trace = F)
#>
#> Coefficients:
#> (Intercept) age
#> Strong -0.01988977 0.009832916
#> Weak 0.59497046 -0.005954348
#>
#> Residual Deviance: 2030.756
#> AIC: 2038.756
```

compare the best model to the full model based on deviance

```
pchisq(q = deviance(Multinomial_Step) - deviance(Multinomial_Model),
df = Multinomial_Model$edf-Multinomial_Step$edf,lower=F)
#> [1] 0.9078172
```

We see no significant difference

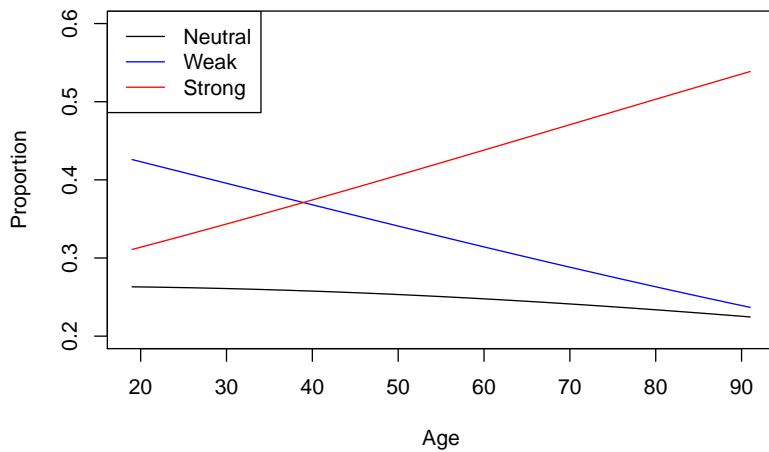
Plot of the fitted model

```
PlotData <- data.frame(age = seq(from = 19, to = 91))
Preds <-
  PlotData %>% bind_cols(data.frame(predict(
    object = Multinomial_Step,
```

```

    PlotData, type = "probs"
  )))
plot(
  x = Preds$age,
  y = Preds$Neutral,
  type = "l",
  ylim = c(0.2, 0.6),
  col = "black",
  ylab = "Proportion",
  xlab = "Age"
)
lines(x = Preds$age,
      y = Preds$Weak,
      col = "blue")
lines(x = Preds$age,
      y = Preds$Strong,
      col = "red")
legend(
  'topleft',
  legend = c('Neutral', 'Weak', 'Strong'),
  col = c('black', 'blue', 'red'),
  lty = 1
)

```



```

predict(Multinomial_Step,data.frame(age = 34)) # predicted result (category of political strength)
#> [1] Weak
#> Levels: Neutral Strong Weak
predict(Multinomial_Step,data.frame(age = c(34,35)),type="probs") # predicted result of the probabilities
#> Neutral      Strong      Weak
#> 1 0.2597275 0.3556910 0.3845815
#> 2 0.2594080 0.3587639 0.3818281

```

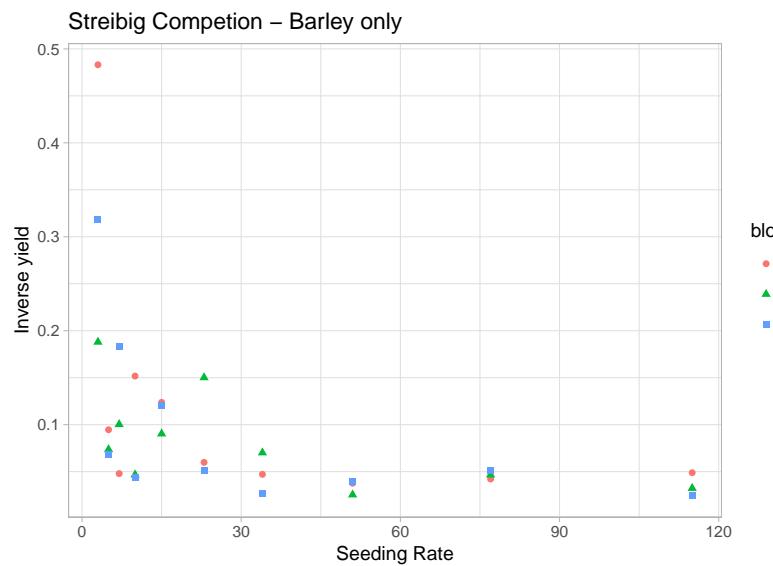
If categories are ordered (i.e., ordinal data), we must use another approach (still multinomial, but use cumulative probabilities).

Another example

```

library(agridat)
dat <- agridat::streibig.competition
# See Schabberger and Pierce, pages 370+
# Consider only the mono-species barley data (no competition from Sinapis)
gammaDat <- subset(dat, sseeds < 1)
gammaDat <-
  transform(gammaDat,
    x = bseeds,
    y = bdwt,
    block = factor(block))
# Inverse yield looks like it will be a good fit for Gamma's inverse link
ggplot(gammaDat, aes(x = x, y = 1 / y)) + geom_point(aes(color = block, shape =
  block)) +
  xlab('Seeding Rate') + ylab('Inverse yield') + ggtitle('Streibig Competition - Barley only')

```



$$Y \sim \text{Gamma}$$

because Gamma is non-negative as opposed to Normal. The canonical Gamma link function is the inverse (or reciprocal) link

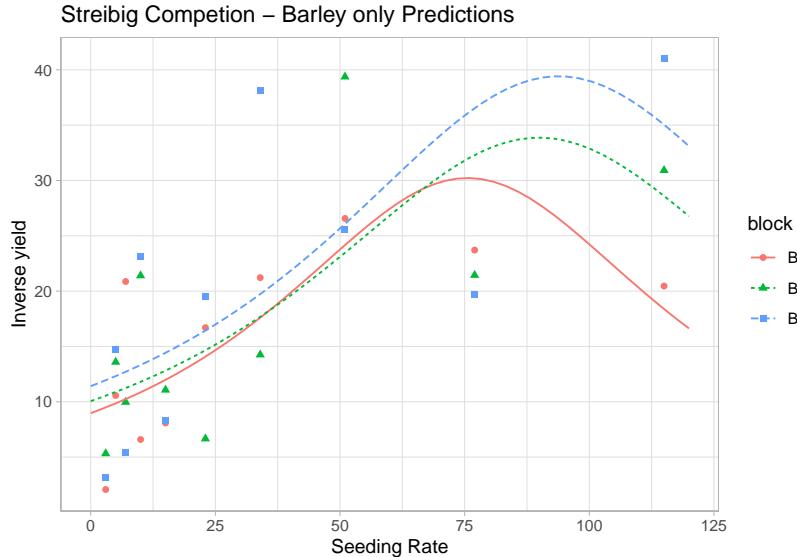
$$\eta_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \beta_2x_{ij}^2Y_{ij} = \eta_{ij}^{-1}$$

The linear predictor is a quadratic model fit to each of the j-th blocks. A different model (not fitted) could be one with common slopes: `glm(y ~ x + I(x^2), ...)`

```
# linear predictor is quadratic, with separate intercept and slope per block
m1 <- glm(y ~ block + block*x + block*I(x^2), data=gammaDat,family=Gamma(link="inverse")
summary(m1)
#>
#> Call:
#> glm(formula = y ~ block + block * x + block * I(x^2), family = Gamma(link = "inverse",
#>     data = gammaDat)
#>
#> Deviance Residuals:
#>      Min        1Q    Median        3Q       Max
#> -1.21708 -0.44148  0.02479  0.17999  0.80745
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 1.115e-01 2.870e-02  3.886 0.000854 ***
#> blockB2     -1.208e-02 3.880e-02 -0.311 0.758630
#> blockB3     -2.386e-02 3.683e-02 -0.648 0.524029
#> x           -2.075e-03 1.099e-03 -1.888 0.072884 .
#> I(x^2)       1.372e-05 9.109e-06  1.506 0.146849
#> blockB2:x   5.198e-04 1.468e-03  0.354 0.726814
#> blockB3:x   7.475e-04 1.393e-03  0.537 0.597103
#> blockB2:I(x^2) -5.076e-06 1.184e-05 -0.429 0.672475
#> blockB3:I(x^2) -6.651e-06 1.123e-05 -0.592 0.560012
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for Gamma family taken to be 0.3232083)
#>
#> Null deviance: 13.1677  on 29  degrees of freedom
#> Residual deviance: 7.8605  on 21  degrees of freedom
#> AIC: 225.32
#>
#> Number of Fisher Scoring iterations: 5
```

For predict new value of x

```
newdf <-
  expand.grid(x = seq(0, 120, length = 50), block = factor(c('B1', 'B2', 'B3')))
newdf$pred <- predict(m1, new = newdf, type = 'response')
ggplot(gammaDat, aes(x = x, y = y)) + geom_point(aes(color = block, shape =
  block)) +
  xlab('Seeding Rate') + ylab('Inverse yield') + ggtitle('Streibig Competition - Barley only Predictions')
  geom_line(data = newdf, aes(
    x = x,
    y = pred,
    color = block,
    linetype = block
  ))
```



7.7 Generalization

We can see that Poisson regression looks similar to logistic regression. Hence, we can generalize to a class of modeling. Thanks to (Nelder and Wedderburn, 1972), we have the **generalized linear models** (GLMs). Estimation is generalized in these models.

Exponential Family

The theory of GLMs is developed for data with distribution given by the **exponential family**.

The form of the data distribution that is useful for GLMs is

$$f(y; \theta, \phi) = \exp\left(\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

where

- θ is called the natural parameter
- ϕ is called the dispersion parameter

Note:

This family includes the

Gamma

,

Normal

,

Poisson

, and other. For all parameterization of the exponential family, check this link

Example

if we have $Y \sim N(\mu, \sigma^2)$

$$\begin{aligned} f(y; \mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(y - \mu)^2\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}(y^2 - 2y\mu + \mu^2) - \frac{1}{2}\log(2\pi\sigma^2)\right) \\ &= \exp\left(\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right) \\ &= \exp\left(\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right) \end{aligned}$$

where

- $\theta = \mu$
- $b(\theta) = \frac{\mu^2}{2}$
- $a(\phi) = \sigma^2 = \phi$
- $c(y, \phi) = -\frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\sigma^2)\right)$

Properties of GLM exponential families

1. $E(Y) = b'(\theta)$ where $b'(\theta) = \frac{\partial b(\theta)}{\partial \theta}$ (here ' is “prime”, not transpose)
2. $var(Y) = a(\phi)b''(\theta) = a(\phi)V(\mu)$.

- $V(\mu)$ is the *variance function*; however, it is only the variance in the case that $a(\phi) = 1$
3. If $a(), b(), c()$ are identifiable, we will derive expected value and variance of Y .

Example

Normal distribution

$$b'(\theta) = \frac{\partial b(\mu^2/2)}{\partial \mu} = \mu V(\mu) = \frac{\partial^2(\mu^2/2)}{\partial \mu^2} = 1 \rightarrow var(Y) = a(\phi) = \sigma^2$$

Poisson distribution

$$\begin{aligned} f(y, \theta, \phi) &= \frac{\mu^y \exp(-\mu)}{y!} \\ &= \exp(y \log(\mu) - \mu - \log(y!)) \\ &= \exp(y\theta - \exp(\theta) - \log(y!)) \end{aligned}$$

where

- $\theta = \log(\mu)$
- $a(\phi) = 1$
- $b(\theta) = \exp(\theta)$
- $c(y, \phi) = \log(y!)$

Hence,

$$E(Y) = \frac{\partial b(\theta)}{\partial \theta} = \exp(\theta) = \mu var(Y) = \frac{\partial^2 b(\theta)}{\partial \theta^2} = \mu$$

Since $\mu = E(Y) = b'(\theta)$

In GLM, we take some monotone function (typically nonlinear) of μ to be linear in the set of covariates

$$g(\mu) = g(b'(\theta)) = \mathbf{x}'$$

Equivalently,

$$\mu = g^{-1}(\mathbf{x}')$$

where $g(.)$ is the **link function** since it links mean response ($\mu = E(Y)$) and a linear expression of the covariates

Some people use $\eta = \mathbf{x}'$ where η = the “linear predictor”

GLM is composed of 2 components

The random component:

- is the distribution chosen to model the response variables Y_1, \dots, Y_n
- is specified by the choice of $a(), b(), c()$ in the exponential form
- Notation:
 - Assume that there are n **independent** response variables Y_1, \dots, Y_n with densities

$$f(y_i; \theta_i, \phi) = \exp\left(\frac{\theta_i y_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right)$$

notice each observation might have different densities

- Assume that ϕ is constant for all $i = 1, \dots, n$, but θ_i will vary. $\mu_i = E(Y_i)$ for all i .

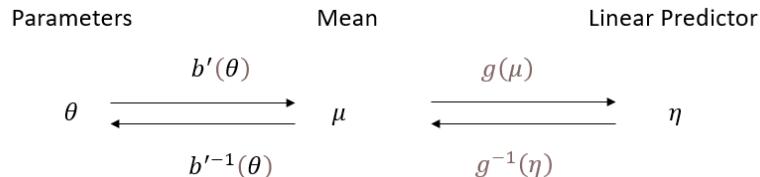
The systematic component

- is the portion of the model that gives the relation between μ and the covariates \mathbf{x}
- consists of 2 parts:
 - the *link function*, $g(\cdot)$
 - the *linear predictor*, $\eta = \mathbf{x}'$
- Notation:
 - assume $g(\mu_i) = \mathbf{x}' = \eta_i$ where $= (\beta_1, \dots, \beta_p)'$
 - The parameters to be estimated are $\beta_1, \dots, \beta_p, \phi$

The Canonical Link

To choose $g(\cdot)$, we can use **canonical link function** (Remember: Canonical link is just a special case of the link function)

If the link function $g(\cdot)$ is such $g(\mu_i) = \eta_i = \theta_i$, the natural parameter, then $g(\cdot)$ is the canonical link.



- $b(\theta)$ = cumulant moment generating function
- $g(\mu)$ is the link function, which relates the linear predictor to the mean and is required to be monotone increasing, continuously differentiable and invertible.

Equivalently, we can think of canonical link function as

$$\gamma^{-1} \circ g^{-1} = I$$

which is the identity. Hence,

$$\theta = \eta$$

The inverse link

$g^{-1}(.)$ is also known as the mean function, take linear predictor output (ranging from $-\infty$ to ∞) and transform it into a different scale.

- **Exponential:** converts \mathbf{X} into a curve that is restricted between 0 and ∞ (which you can see that is useful in case you want to convert a linear predictor into a non-negative value). $\lambda = \exp(y) = \mathbf{X}$
- **Inverse Logit** (also known as logistic): converts \mathbf{X} into a curve that is restricted between 0 and 1, which is useful in case you want to convert a linear predictor to a probability. $\theta = \frac{1}{1+\exp(-y)} = \frac{1}{1+\exp(-\mathbf{X})}$
 - y = linear predictor value
 - θ = transformed value

The **identity link** is that

$$\eta_i = g(\mu_i) = \mu_i \mu_i = g^{-1}(\eta_i) = \eta_i$$

Link	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
Identity	μ_i	η_i
Log	$\log_e \mu_i$	e^{η_i}
Inverse	μ_i^{-1}	η_i^{-1}
Inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
Square-root	$\sqrt{\mu_i}$	η_i^{2}
Logit	$\log_e \frac{\mu_i}{1 - \mu_i}$	$\frac{1}{1 + e^{-\eta_i}}$
Probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
Log-log	$-\log_e[-\log_e(\mu_i)]$	$\exp[-\exp(-\eta_i)]$
Complementary log-log	$\log_e[-\log_e(1 - \mu_i)]$	$1 - \exp[-\exp(\eta_i)]$

NOTE: μ_i is the expected value of the response; η_i is the linear predictor; and $\Phi(\cdot)$ is the cumulative distribution function of the standard-normal distribution.

Table 15.1 Generalized Linear Models 15.1 the Structure of Generalized Linear Models

More example on the link functions and their inverses can be found on page 380

Example

Normal random component

- Mean Response: $\mu_i = \theta_i$
- Canonical Link: $g(\mu_i) = \mu_i$ (the identity link)

Binomial random component

- Mean Response: $\mu_i = \frac{n_i \exp(\theta)}{1 + \exp(\theta)}$ and $\theta(\mu_i) = \log(\frac{p_i}{1-p_i}) = \log(\frac{\mu_i}{n_i - \mu_i})$
- Canonical link: $g(\mu_i) = \log(\frac{\mu_i}{n_i - \mu_i})$ (logit link)

Poisson random component

- Mean Response: $\mu_i = \exp(\theta_i)$
- Canonical Link: $g(\mu_i) = \log(\mu_i)$

Gamma random component:

- Mean response: $\mu_i = -\frac{1}{\theta_i}$ and $\theta(\mu_i) = -\mu_i^{-1}$
- Canonical Link: $g(\mu_i) = -\frac{1}{\mu_i}$

Inverse Gaussian random

- Canonical Link: $g(\mu_i) = \frac{1}{\mu_i^2}$

7.7.1 Estimation

- MLE for parameters of the **systematic component** (β)
- Unification of derivation and computation (thanks to the exponential forms)
- No unification for estimation of the dispersion parameter (ϕ)

7.7.1.1 Estimation of β

We have

$$f(y_i; \theta_i, \phi) = \exp\left(\frac{\theta_i y_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right) E(Y_i) = \mu_i = b'(\theta) var(Y_i) = b''(\theta) a(\phi) = V(\mu_i) a(\phi) g(\mu_i) = \mathbf{x}'_i \beta = \eta_i$$

If the log-likelihood for a single observation is $l_i(\beta, \phi)$. The log-likelihood for all n observations is

$$\begin{aligned} l(\beta, \phi) &= \sum_{i=1}^n l_i(\beta, \phi) \\ &= \sum_{i=1}^n \left(\frac{\theta_i y_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right) \end{aligned}$$

Using MLE to find β , we use the chain rule to get the derivatives

$$\begin{aligned} \frac{\partial l_i(\beta, \phi)}{\partial \beta_j} &= \frac{\partial l_i(\beta, \phi)}{\partial \theta_i} \times \frac{\partial \theta_i}{\partial \mu_i} \times \frac{\partial \mu_i}{\partial \eta_i} \times \frac{\partial \eta_i}{\partial \beta_j} \\ &= \sum_{i=1}^n \left(\frac{y_i - \mu_i}{a(\phi)} \times \frac{1}{V(\mu_i)} \times \frac{\partial \mu_i}{\partial \eta_i} \times x_{ij} \right) \end{aligned}$$

If we let

$$w_i \equiv \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^2 V(\mu_i)^{-1}$$

Then,

$$\frac{\partial l_i(\beta, \phi)}{\partial \beta_j} = \sum_{i=1}^n \left(\frac{y_i \mu_i}{a(\phi)} \times w_i \times \frac{\partial \eta_i}{\partial \mu_i} \times x_{ij} \right)$$

We can also get the second derivatives using the chain rule.

Example:

For the

Newton – Raphson

algorithm, we need

$$-E\left(\frac{\partial^2 l(\beta, \phi)}{\partial \beta_j \partial \beta_k}\right)$$

where (j, k) th element of the **Fisher information matrix** $\mathbf{I}(\beta)$

Hence,

$$-E\left(\frac{\partial^2 l(\beta, \phi)}{\partial \beta_j \partial \beta_k}\right) = \sum_{i=1}^n \frac{w_i}{a(\phi)} x_{ij} x_{ik}$$

for the (j, k) th element

If Bernoulli model with logit link function (which is the canonical link)

$$b(\theta) = \log(1 + \exp(\theta)) = \log(1 + \exp(\mathbf{x}')) a(\phi) = 1 c(y_i, \phi) = 0 E(Y) = b'(\theta) = \frac{\exp(\theta)}{1 + \exp(\theta)} = \mu = p\eta = g(\mu)$$

For Y_i , $i = 1, \dots$, the log-likelihood is

$$l_i(\beta, \phi) = \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) = y_i \mathbf{x}'_i \beta - \log(1 + \exp(\mathbf{x}'))$$

Additionally,

$$V(\mu_i) = \mu_i(1 - \mu_i) = p_i(1 - p_i) \frac{\partial \mu_i}{\partial \eta_i} = p_i(1 - p_i)$$

Hence,

$$\begin{aligned} \frac{\partial l(\beta, \phi)}{\partial \beta_j} &= \sum_{i=1}^n \left[\frac{y_i - \mu_i}{a(\phi)} \times \frac{1}{V(\mu_i)} \times \frac{\partial \mu_i}{\partial \eta_i} \times x_{ij} \right] \\ &= \sum_{i=1}^n (y_i - p_i) \times \frac{1}{p_i(1 - p_i)} \times p_i(1 - p_i) \times x_{ij} \\ &= \sum_{i=1}^n (y_i - p_i) x_{ij} \\ &= \sum_{i=1}^n \left(y_i - \frac{\exp(\mathbf{x}'_i)}{1 + \exp(\mathbf{x}'_i)} \right) x_{ij} \end{aligned}$$

then

$$w_i = \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^2 V(\mu_i))^{-1} = p_i(1-p_i)$$

$$\mathbf{I}_{jk}(\phi) = \sum_{i=1}^n \frac{w_i}{a(\phi)} x_{ij} x_{ik} = \sum_{i=1}^n p_i(1-p_i) x_{ij} x_{ik}$$

The **Fisher-scoring** algorithm for the MLE of β is

$$\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}^{(m+1)} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}^{(m)} + \mathbf{I}^{-1}(\phi) \begin{pmatrix} \frac{\partial l(\beta, \phi)}{\partial \beta_1} \\ \frac{\partial l(\beta, \phi)}{\partial \beta_2} \\ \vdots \\ \frac{\partial l(\beta, \phi)}{\partial \beta_p} \end{pmatrix} \Big|_{\beta=\beta^{(m)}}$$

Similar to

Newton-Raphson

expect the matrix of second derivatives by the expected value of the second derivative matrix.

In matrix notation,

$$\begin{aligned} \frac{\partial l}{\partial \beta} &= \frac{1}{a(\phi)} \mathbf{X}' \mathbf{W} (\mathbf{y} - \mathbf{X}\beta) \\ &= \frac{1}{a(\phi)} \mathbf{F}' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\beta) \end{aligned}$$

$$\mathbf{I}(\beta) = \frac{1}{a(\phi)} \mathbf{X}' \mathbf{W} \mathbf{X} = \frac{1}{a(\phi)} \mathbf{F}' \mathbf{V}^{-1} \mathbf{F}$$

where

- \mathbf{X} is an $n \times p$ matrix of covariates
- \mathbf{W} is an $n \times n$ diagonal matrix with (i,i) th element given by w_i
- an $n \times n$ diagonal matrix with (i,i) th element given by $\frac{\partial \eta_i}{\partial \mu_i}$
- $\mathbf{F} = -\mathbf{X}'$ an $n \times p$ matrix with i th row $\frac{\partial \mu_i}{\partial \beta} = \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \mathbf{x}_i'$
- \mathbf{V} an $n \times n$ diagonal matrix with (i,i) th element given by $V(\mu_i)$

Setting the derivative of the log-likelihood equal to 0, ML estimating equations are

$$\mathbf{F}'\mathbf{V}^{-1}\mathbf{y} = \mathbf{F}'\mathbf{V}^{-1}$$

where all components of this equation expect \mathbf{y} depends on the parameters β

Special Cases

If one has a canonical link, the estimating equations reduce to

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'$$

If one has an identity link, then

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$$

which gives the generalized least squares estimator

Generally, we can rewrite the Fisher-scoring algorithm as

$$\beta^{(m+1)} = \beta^{(m)} + (\hat{\mathbf{F}}'\hat{\mathbf{V}}^{-1}\hat{\mathbf{F}})^{-1}\hat{\mathbf{F}}'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \hat{\mathbf{y}})$$

Since $\hat{\mathbf{F}}$, $\hat{\mathbf{V}}$, $\hat{\mu}$ depend on β , we evaluate at $\beta^{(m)}$

From starting values $\beta^{(0)}$, we can iterate until convergence.

Notes:

- if $a(\phi)$ is a constant or of the form $m_i\phi$ with known m_i , then ϕ cancels.

7.7.1.2 Estimation of ϕ

2 approaches:

1. MLE

$$\frac{\partial l_i}{\partial \phi} = \frac{(\theta_i y_i - b(\theta_i)a'(\phi))}{a^2(\phi)} + \frac{\partial c(y_i, \phi)}{\partial \phi}$$

the MLE of ϕ solves

$$\frac{a^2(\phi)}{a'(\phi)} \sum_{i=1}^n \frac{\partial c(y_i, \phi)}{\partial \phi} = \sum_{i=1}^n (\theta_i y_i - b(\theta_i))$$

- Situation others than normal error case, expression for $\frac{\partial c(y, \phi)}{\partial \phi}$ are not simple
- Even for the canonical link and $a(\phi)$ constant, there is no nice general expression for $-E(\frac{\partial^2 l}{\partial \phi^2})$, so the unification GLMs provide for estimation of β breaks down for ϕ

2. Moment Estimation (“Bias Corrected χ^2 ”)

- The MLE is not conventional approach to estimation of ϕ in GLMS.
- For the exponential family $\text{var}(Y) = V(\mu)a(\phi)$. This implies

$$a(\phi) = \frac{\text{var}(Y)}{V(\mu)} = \frac{E(Y - \mu)^2}{V(\mu)} a(\hat{\phi}) = \frac{1}{n-p} \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu})}$$

where p is the dimension of β

- GLM with canonical link function $g(.) = (b'(.))^{-1}$

$$g(\mu) = \theta = \eta = \mathbf{x}' \mu = g^{-1}(\eta) = b'(\eta)$$

- so the method estimator for $a(\phi) = \phi$ is

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^n \frac{(y_i - g^{-1}(\hat{\eta}_i))^2}{V(g^{-1}(\hat{\eta}_i))}$$

7.7.2 Inference

We have

$$\hat{v}\hat{a}r(\beta) = a(\phi)(\hat{\mathbf{F}}'\hat{\mathbf{V}}\hat{\mathbf{F}})^{-1}$$

where

- \mathbf{V} is an $n \times n$ diagonal matrix with diagonal elements given by $V(\mu_i)$
- \mathbf{F} is an $n \times p$ matrix given by $\mathbf{F} = \frac{\partial \mu}{\partial \beta}$
- Both \mathbf{V}, \mathbf{F} are dependent on the mean μ , and thus β . Hence, their estimates $(\hat{\mathbf{V}}, \hat{\mathbf{F}})$ depend on $\hat{\beta}$.

$$H_0 : \mathbf{L} = \mathbf{d}$$

where \mathbf{L} is a $q \times p$ matrix with a **Wald** test

$$W = (\mathbf{L} - \mathbf{d})'(\mathbf{a}(\cdot)\mathbf{L}'\hat{\mathbf{F}}'\hat{\mathbf{V}}^{-1}\hat{\mathbf{F}})\mathbf{L}'^{-1}(\mathbf{L} - \mathbf{d})$$

which follows χ_q^2 distribution (asymptotically), where q is the rank of \mathbf{L}

In the simple case $H_0 : \beta_j = 0$ gives $W = \frac{\hat{\beta}_j^2}{\hat{v}ar(\hat{\beta}_j)} \sim \chi_1^2$ asymptotically

Likelihood ratio test

$$\Lambda = 2(l(\hat{\beta}_f) - l(\hat{\beta}_r)) \sim \chi_q^2$$

where

- q is the number of constraints used to fit the reduced model $\hat{\beta}_r$, and $\hat{\beta}_r$ is the fit under the full model.

Wald test is easier to implement, but likelihood ratio test is better (especially for small samples).

7.7.3 Deviance

Deviance is necessary for goodness of fit, inference and for alternative estimation of the dispersion parameter. We define and consider Deviance from a likelihood ratio perspective.

- Assume that ϕ is known. Let $\tilde{\theta}$ denote the full and $\hat{\theta}$ denote the reduced model MLEs. Then, the likelihood ratio (2 times the difference in log-likelihoods) is

$$2 \sum_{i=1}^n \frac{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)}{a_i(\phi)}$$

- For exponential families, $\mu = E(y) = b'(\theta)$, so the natural parameter is a function of $\mu : \theta = \theta(\mu) = b'^{-1}(\mu)$, and the likelihood ratio turns into

$$2 \sum_{i=1}^m \frac{y_i\{\theta(\tilde{\mu}_i - \theta(\hat{\mu}_i)\} - b(\theta(\tilde{\mu}_i)) + b(\theta(\hat{\mu}_i))}{a_i(\phi)}$$

- Comparing a fitted model to “the fullest possible model”, which is the **saturated model**: $\tilde{\mu}_i = y_i$, $i = 1, \dots, n$. If $\tilde{\theta}_i^* = \theta(y_i)$, $\hat{\theta}_i^* = \theta(\hat{\mu}_i)$, the likelihood ratio is

$$2 \sum_{i=1}^n \frac{y_i(\tilde{\theta}_i^* - \hat{\theta}_i^* + b(\hat{\theta}_i^*))}{a_i(\phi)}$$

- (McCullagh and Nelder, 2019) specify $a(\phi) = \phi$, then the likelihood ratio can be written as

$$D^*(\mathbf{y}; \hat{\theta}) = \frac{2}{\phi} \sum_{i=1}^n \{y_i(\tilde{\theta}_i^* - \hat{\theta}_i^*) - b(\tilde{\theta}_i^*) + b(\hat{\theta}_i^*)\}$$

where

- $D^*(\mathbf{y}; \hat{\theta})$ = scaled deviance
- $D(\mathbf{y}; \hat{\theta}) = \phi D^*(\mathbf{y}; \hat{\theta})$ = deviance

Note:

- in some random component distributions, we can write $a_i(\phi) = \phi m_i$, where
 - m_i is some known scalar that may change with the observations.
 - Then, the scaled deviance components are divided by m_i :

$$D^*(\mathbf{y}; \hat{\theta}) \equiv 2 \sum_{i=1}^n \{y_i(\tilde{\theta}_i^* - \hat{\theta}_i^*) - b(\tilde{\theta}_i^*) + b(\hat{\theta}_i^*)\}/(\phi m_i)$$

- $D^*(\mathbf{y}; \hat{\theta}) = \sum_{i=1}^n d_i$ where d_i is the deviance contribution from the i th observation.
- D is used in model selection
- D^* is used in goodness of fit tests (as it is a likelihood ratio statistic).

$$D^*(\mathbf{y}; \hat{\theta}) = 2\{l(\mathbf{y}; \tilde{\theta}) - l(\mathbf{y}; \hat{\theta})\}$$

- d_i are used to form **deviance residuals**

Example:

Normal

We have

$$\theta = \mu\phi = \sigma^2 b(\theta) = \frac{1}{2}\theta^2 a(\phi) = \phi$$

Hence,

$$\tilde{\theta}_i = y_i \hat{\theta}_i = \hat{\mu}_i = g^{-1}(\hat{\eta}_i)$$

And

$$\begin{aligned}
D &= 2 \sum_{i=1}^n Y_i^2 - y_i \hat{\mu}_i - \frac{1}{2} y_i^2 + \frac{1}{2} \hat{\mu}_i^2 \\
&= \sum_{i=1}^n y_i^2 - 2y_i \hat{\mu}_i + \hat{\mu}_i^2 \\
&= \sum_{i=1}^n (y_i - \hat{\mu}_i)^2
\end{aligned}$$

which is the **residual sum of squares**

Poisson

$$f(y) = \exp\{y \log(\mu) - \mu - \log(y!)\} \theta = \log(\mu) b(\theta) = \exp(\theta) a(\phi) = 1 \tilde{\theta}_i = \log(y_i) \hat{\theta}_i = \log(\hat{\mu}_i) \hat{\mu}_i = g^{-1}(\hat{\eta}_i)$$

Then,

$$\begin{aligned}
D &= 2 \sum_{i=1}^n y_i \log(y_i) - y_i \log(\hat{\mu}_i) - y_i + \hat{\mu}_i \\
&= 2 \sum_{i=1}^n y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)
\end{aligned}$$

and

$$d_i = 2\left\{y_i \log\left(\frac{y_i}{\hat{\mu}}\right) - (y_i - \hat{\mu}_i)\right\}$$

7.7.3.1 Analysis of Deviance

The difference in deviance between a reduced and full model, where q is the difference in the number of free parameters, has an asymptotic χ_q^2 . The likelihood ratio test

$$D^*(\mathbf{y}; \mathbf{r}) - D^*(\mathbf{y}; \mathbf{f}) = 2\{l(\mathbf{y}; \mathbf{f}) - l(\mathbf{y}; \mathbf{r})\}$$

this comparison of models is **Analysis of Deviance**. GLM uses this analysis for model selection.

An estimation of ϕ is

$$\hat{\phi} = \frac{D(\mathbf{y}; \mathbf{r})}{n - p}$$

where p = number of parameters fit.

Excessive use of χ^2 test could be problematic since it is asymptotic (McCullagh and Nelder, 2019)

7.7.3.2 Deviance Residuals

We have $D = \sum_{i=1}^n d_i$. Then, we define **deviance residuals**

$$r_{D_i} = \text{sign}(y_i - \hat{\mu}_i) \sqrt{d_i}$$

Standardized version of deviance residuals is

$$r_{s,i} = \frac{y_i - \hat{\mu}}{\hat{\sigma}(1 - h_{ii})^{1/2}}$$

Let $\mathbf{H}^{\text{GLM}} = \mathbf{W}^{1/2} \mathbf{X} (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}^{-1/2}$, where \mathbf{W} is an n x n diagonal matrix with (i,i)th element given by w_i (see Estimation of β). Then Standardized deviance residuals is equivalently

$$r_{s,D_i} = \frac{r_{D_i}}{\{\hat{\phi}(1 - h_{ii}^{glm})\}^{1/2}}$$

where h_{ii}^{glm} is the ith diagonal of \mathbf{H}^{GLM}

7.7.3.3 Pearson Chi-square Residuals

Another χ^2 statistic is **Pearson** χ^2 statistics: (assume $m_i = 1$)

$$X^2 = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

where $\hat{\mu}_i$ is the fitted mean response fo the model of interest.

The **Scaled Pearson** χ^2 statistic is given by $\frac{X^2}{\phi} \sim \chi^2_{n-p}$ where p is the number of parameters estimated. Hence, the **Pearson** χ^2 residuals are

$$X_i^2 = \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

If we have the following assumptions:

- Independent samples
- No over-dispersion: If $\phi = 1$, $\frac{D(\hat{\mathbf{y}})}{n-p}$ and $\frac{X^2}{n-p}$ have a value substantially larger 1 indicates **improperly specified model** or **overdispersion**
- Multiple groups

then $\frac{X^2}{\phi}$ and $D^*(\hat{\mathbf{y}})$ both follow χ^2_{n-p}

7.7.4 Diagnostic Plots

- Standardized residual Plots:
 - $\text{plot}(r_{s,D_i}, \hat{\mu}_i)$ or $\text{plot}(r_{s,D_i}, T(\hat{\mu}_i))$ where $T(\hat{\mu}_i)$ is transformation($\hat{\mu}_i$) called **constant information scale**:
 - $\text{plot}(r_{s,D_i}, \hat{\eta}_i)$

Random Component	$T(\hat{\mu}_i)$
Normal	$\hat{\mu}$
Poisson	$2\sqrt{\hat{\mu}}$
Binomial	$2 \sin^{-1}(\sqrt{\hat{\mu}})$
Gamma	$2 \log(\hat{\mu})$
Inverse Gaussian	$-2\hat{\mu}^{-1/2}$

- If we see:
 - Trend, it means we might have a wrong link function, or choice of scale
 - Systematic change in range of residuals with a change in $T(\hat{\mu})$ (incorrect random component) (systematic \neq random)
- $\text{plot}(|r_{D_i}|, \hat{\mu}_i)$ to check **Variance Function**.

7.7.5 Goodness of Fit

To assess goodness of fit, we can use

- Deviance

- Pearson Chi-square Residuals

In nested model, we could use likelihood-based information measures:

$$AIC = -2l(\hat{\mu}) + 2p \quad AICC = -2l(\hat{\mu}) + 2p\left(\frac{n}{n-p-1}\right) \quad BIC = 2l(\hat{\mu}) + p \log(n)$$

where

- $l(\hat{\mu})$ is the log-likelihood evaluated at the parameter estimates
- p is the number of parameters
- n is the number of observations.

Note: you have to use the same data with the same model (i.e., same link function, same random underlying random distribution). but you can have different number of parameters.

Even though statisticians try to come up with measures that are similar to R^2 , in practice, it is not so appropriate. For example, they compare the log-likelihood of the fitted model against the that of a model with just the intercept:

$$R_p^2 = 1 - \frac{l(\hat{\mu})}{l(\hat{\mu}_0)}$$

For certain specific random components such as binary response model, we have **rescaled generalized R^2 :

$$\bar{R}^2 = \frac{R_*^2}{\max(R_*^2)} = \frac{1 - \exp\{-\frac{2}{n}(l(\hat{\mu}) - l(\hat{\mu}_0))\}}{1 - \exp\{\frac{2}{n}l(\hat{\mu}_0)\}}$$

7.7.6 Over-Dispersion

Random Components	$var(Y)$	$V(\mu)$
Binomial	$var(Y) = n\mu(1-\mu)$	$V(\mu) = \phi n\mu(1-\mu)$ where $m_i = n$
Poisson	$var(Y) = \mu$	$V(\mu) = \phi\mu$

In both cases $\phi = 1$. Recall $b''(\theta) = V(\mu)$ check Estimation of ϕ .

If we find

- $\phi > 1$: over-dispersion (i.e., too much variation for an independent binomial or Poisson distribution).
- $\phi < 1$: under-dispersion (i.e., too little variation for an independent binomial or Poisson distribution).

If we have either over or under-dispersion, it means we might have unspecified random component, we could

- Select a different random component distribution that can accommodate over or under-dispersion (e.g., negative binomial, Conway-Maxwell Poisson)
- use

GeneralizedLinearMixedModels

NonlinearandGeneralizedLinearMixedModels

to handle random effects in generalized linear models.

Chapter 8

Linear Mixed Models

8.1 Dependent Data

Forms of dependent data:

- Multivariate measurements on different individuals: (e.g., a person's blood pressure, fat, etc are correlated)
- Clustered measurements: (e.g., blood pressure measurements of people in the same family can be correlated).
- Repeated measurements: (e.g., measurement of cholesterol over time can be correlated) “If data are collected repeatedly on experimental material to which treatments were applied initially, the data is a repeated measure.” (Schabenberger and Pierce, 2001)
- Longitudinal data: (e.g., individual's cholesterol tracked over time are correlated): “data collected repeatedly over time in an observational study are termed longitudinal.” (Schabenberger and Pierce, 2001)
- Spatial data: (e.g., measurement of individuals living in the same neighborhood are correlated)

Hence, we like to account for these correlations.

Linear Mixed Model (LMM), also known as **Mixed Linear Model** has 2 components:

- **Fixed effect** (e.g, gender, age, diet, time)
- **Random effects** representing individual variation or auto correlation/spatial effects that imply **dependent (correlated) errors**

Review Two-Way Mixed Effects ANOVA

We choose to model the random subject-specific effect instead of including dummy subject covariates in our model because:

- reduction in the number of parameters to estimate
- when you do inference, it would make more sense that you can infer from a population (i.e., random effect).

LLM Motivation

In a repeated measurements analysis where Y_{ij} is the response for the i-th individual measured at the j-th time,

$$i = 1, \dots, N ; j = 1, \dots, n_i$$

$$\mathbf{Y}_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ \vdots \\ Y_{in_i} \end{pmatrix}$$

is all measurements for subject i.

Stage 1: (Regression Model) how the response changes over time for the ith subject

$$\mathbf{Y}_i = \mathbf{Z}_i \beta_i + \epsilon_i$$

where

- Z_i is an $n_i \times q$ matrix of known covariates
- β_i is an unknown $q \times 1$ vector of subject-specific coefficients (regression coefficients different for each subject)
- ϵ_i are the random errors (typically $\sim N(0, \sigma^2 I)$)

We notice that there are two many β to estimate here. Hence, this is the motivation for the second stage

Stage 2: (Parameter Model)

$$\beta_i = \mathbf{K}_i b_i + \mathbf{b}_i$$

where

- K_i is a $q \times p$ matrix of known covariates

- β is a $p \times 1$ vector of unknown parameter
- \mathbf{b}_i are independent $N(0, D)$ random variables

This model explain the observed variability between subjects with respect to the subject-specific regression coefficients, β_i . We model our different coefficient (β_i) with respect to β .

Example:

Stage 1:

$$Y_{ij} = \beta_{1i} + \beta_{2i}t_{ij} + \epsilon_{ij}$$

where

- $j = 1, \dots, n_i$

In the matrix notation,

$$\mathbf{Y}_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix}$$

$$\mathbf{Z}_i = \begin{pmatrix} 1 & t_{i1} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix}$$

$$\beta_i = \begin{pmatrix} \beta_{1i} \\ \beta_{2i} \end{pmatrix}$$

$$\epsilon_i = \begin{pmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{in_i} \end{pmatrix}$$

Thus,

$$\mathbf{Y}_i = \mathbf{Z}_i \beta_i + \epsilon_i$$

Stage 2:

$$\beta_{1i} = \beta_0 + b_{1i}\beta_{2i} = \beta_1 L_i + \beta_2 H_i + \beta_3 C_i + b_{2i}$$

where L_i, H_i, C_i are indicator variables defined to 1 as the subject falls into different categories.

Subject specific intercepts do not depend upon treatment, with β_0 (the average response at the start of treatment), and $\beta_1, \beta_2, \beta_3$ (the average time effects for each of three treatment groups).

$$\mathbf{K}_i = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & L_i & H_i & C_i \end{pmatrix} \beta = (\beta_0, \beta_1, \beta_2, \beta_3)' \mathbf{b}_i = \begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} \beta_i = \mathbf{K}_i + \mathbf{b}_i$$

To get $\hat{\beta}$, we can fit the model sequentially:

1. Estimate $\hat{\beta}_i$ in the first stage
2. Estimate $\hat{\beta}$ in the second stage by replacing β_i with $\hat{\beta}_i$

However, problems arise from this method:

- information is lost by summarizing the vector \mathbf{Y}_i solely by $\hat{\beta}_i$
- we need to account for variability when replacing β_i with its estimate
- different subjects might have different number of observations.

To address these problems, we can use **Linear Mixed Model (Laird and Ware, 1982)**

Substituting stage 2 into stage 1:

$$\mathbf{Y}_i = \mathbf{Z}_i \mathbf{K}_i \beta + \mathbf{Z}_i \mathbf{b}_i + \epsilon_i$$

Let $\mathbf{X}_i = \mathbf{Z}_i \mathbf{K}_i$ be an $n_i \times p$ matrix. Then, the LMM is

$$\mathbf{Y}_i = \mathbf{X}_i \beta + \mathbf{Z}_i \mathbf{b}_i + \epsilon_i$$

where

- $i = 1, \dots, N$
- β are the fixed effects, which are common to all subjects
- \mathbf{b}_i are the subject specific random effects. $\mathbf{b}_i \sim N_q(\mathbf{0}, \mathbf{D})$
- $\epsilon_i \sim N_{n_i}(\mathbf{0}, \mathbf{I})$
- \mathbf{b}_i and ϵ_i are independent
- $\mathbf{Z}_{i(n_i \times q)}$ and $\mathbf{X}_{i(n_i \times p)}$ are matrices of known covariates.

Equivalently, in the hierarchical form, we call **conditional** or **hierarchical** formulation of the linear mixed model

$$\mathbf{Y}_i | \mathbf{b}_i \sim N(\mathbf{X}_i \beta + \mathbf{Z}_i \mathbf{b}_i, \sigma^2 \mathbf{I}) \quad \mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$$

for $i = 1, \dots, N$. denote the respective functions by $f(\mathbf{Y}_i | \mathbf{b}_i)$ and $f(\mathbf{b}_i)$

In general,

$$f(A, B) = f(A|B)f(B)f(A) = \int f(A, B)dB = \int f(A|B)f(B)dB$$

In the LMM, the marginal density of \mathbf{Y}_i is

$$f(\mathbf{Y}_i) = \int f(\mathbf{Y}_i | \mathbf{b}_i) f(\mathbf{b}_i) d\mathbf{b}_i$$

which can be shown

$$\mathbf{Y}_i \sim N(\mathbf{X}_i \beta + \mathbf{Z}_i \mathbf{D} \mathbf{Z}'_i + \epsilon_i)$$

This is the **marginal** formulation of the linear mixed model

Notes:

We no longer have $Z_i b_i$ in the mean, but add error in the variance (marginal dependence in Y). kinda of averaging out the common effect. Technically, we shouldn't call it averaging the error b (adding it to the variance covariance matrix), it should be called adding random effect

Continue with our example

$$Y_{ij} = (\beta_0 + b_{1i}) + (\beta_1 L_i + \beta_2 H_i + \beta_3 C_i + b_{2i})t_{ij} + \epsilon_{ij}$$

for each treatment group

$$Y_{ik} = \begin{cases} \beta_0 + b_{1i} + (\beta_1 + b_{2i})t_{ij} + \epsilon_{ij} & L \\ \beta_0 + b_{1i} + (\beta_2 + b_{2i})t_{ij} + \epsilon_{ij} & H \\ \beta_0 + b_{1i} + (\beta_3 + b_{2i})t_{ij} + \epsilon_{ij} & C \end{cases}$$

- Intercepts and slopes are all subject specific
- Different treatment groups have different slopes, but the same intercept.

In the hierarchical model form

$$\mathbf{Y}_i | \mathbf{b}_i \sim N(\mathbf{X}_i \beta + \mathbf{Z}_i \mathbf{b}_i, \sigma^2_i) \quad \mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$$

X will be in the form of

$$\mathbf{X}_i = \mathbf{Z}_i \mathbf{K}_i = \begin{bmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \cdot & \cdot \\ 1 & t_{in_i} \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & L_i & H_i & C_i \end{bmatrix} = \begin{bmatrix} 1 & t_{i1}L_i & t_{i1}H_i & T_{i1}C_i \\ 1 & t_{i2}L_i & t_{i2}H_i & T_{i2}C_i \\ \cdot & \cdot & \cdot & \cdot \\ 1 & t_{in_i}L_i & t_{in_i}H_i & T_{in_i}C_i \end{bmatrix}$$

$$\beta = (\beta_0, \beta_1, \beta_2, \beta_3)' \mathbf{b}_i = \begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix}, D = \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix}$$

Assuming $\sigma^2 \mathbf{I}_{n_i}$, which is called **conditional independence**, meaning the response on subject i are independent conditional on \mathbf{b}_i and β

In the marginal model form

$$Y_{ij} = \beta_0 + \beta_1 L_i t_{ij} + \beta_2 H_i t_{ij} + \beta_3 C_i t_{ij} + \eta_{ij}$$

$$\text{where } \eta_i \sim N(\mathbf{0}, \mathbf{Z}_i \mathbf{D} \mathbf{Z}'_i + \sigma^2 \mathbf{I}_{n_i})$$

Equivalently,

$$\mathbf{Y}_i \sim N(\mathbf{X}_i, \mathbf{Z}_i \mathbf{D} \mathbf{Z}'_i + \sigma^2 \mathbf{I}_{n_i})$$

In this case that $n_i = 2$

$$\mathbf{Z}_i \mathbf{D} \mathbf{Z}'_i = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \end{pmatrix} \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ t_{i1} & t_{i2} \end{pmatrix} = \begin{pmatrix} d_{11} + 2d_{12}t_{i1} + d_{22}t_{i1}^2 & d_{11} + d_{12}(t_{i1} + t_{i2}) \\ d_{11} + d_{12}(t_{i1} + t_{i2}) + d_{22}t_{i1}t_{i2} & d_{11} + 2d_{12}t_{i2} \end{pmatrix}$$

$$var(Y_{i1}) = d_{11} + 2d_{12}t_{i1} + d_{22}t_{i1}^2 + \sigma^2$$

On top of correlation in the errors, the marginal implies that the variance function of the response is quadratic over time, with positive curvature d_{22}

8.1.1 Random-Intercepts Model

If we remove the random slopes,

- the assumption is that all variability in subject-specific slopes can be attributed to treatment differences
- the model is random-intercepts model. This has subject specific intercepts, but the same slopes within each treatment group.

$$\mathbf{Y}_i | b_i \sim N(\mathbf{X}_i \beta + 1b_i, \Sigma_i) b_i \sim N(0, d_{11})$$

The marginal model is then ($\sigma^2 = \sigma^2 \mathbf{I}$)

$$\mathbf{Y}_i \sim N(\mathbf{X}_i \beta, 11' d_{11} + \sigma^2 \mathbf{I})$$

The marginal covariance matrix is

$$cov(\mathbf{Y}_i) = 11' d_{11} + \sigma^2 I = \begin{pmatrix} d_{11} + \sigma^2 & d_{11} & \dots & d_{11} \\ d_{11} & d_{11} + \sigma^2 & d_{11} & \dots \\ \vdots & \vdots & \ddots & \vdots \\ d_{11} & \dots & \dots & d_{11} + \sigma^2 \end{pmatrix}$$

the associated correlation matrix is

$$corr(\mathbf{Y}_i) = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \dots & \dots & 1 \end{pmatrix}$$

where $\rho \equiv \frac{d_{11}}{d_{11} + \sigma^2}$

Thus, we have

- constant variance over time
- equal, positive correlation between any two measurements from the same subject
- a covariance structure that is called **compound symmetry**, and ρ is called the **intra-class correlation**
- that when ρ is large, the **inter-subject variability** (d_{11}) is large relative to the intra-subject variability (σ^2)

8.1.2 Covariance Models

If the conditional independence assumption, ($\epsilon_i = \sigma^2 \mathbf{I}_{n_i}$). Consider, $\epsilon_i = \epsilon_{(1)i} + \epsilon_{(2)i}$, where

- $\epsilon_{(1)i}$ is a “serial correlation” component. That is, part of the individual’s profile is a response to time-varying stochastic processes.
- $\epsilon_{(2)i}$ is the measurement error component, and is independent of $\epsilon_{(1)i}$

Then

$$\mathbf{Y}_i = \mathbf{X}_i + \mathbf{Z}_i \mathbf{b}_i + \epsilon_{(1)i} + \epsilon_{(2)i}$$

where

- $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$
- $\epsilon_{(2)i} \sim N(\mathbf{0}, \mathbf{2I}_{n_i})$
- $\epsilon_{(1)i} \sim N(\mathbf{0}, \mathbf{2H}_i)$
- \mathbf{b}_i and ϵ_i are mutually independent

To model the structure of the $n_i \times n_i$ correlation (or covariance) matrix \mathbf{H}_i . Let the (j,k)th element of \mathbf{H}_i be $h_{ijk} = g(t_{ij} - t_{ik})$, that is a function of the times t_{ij} and t_{ik} , which is assumed to be some function of the “distance” between the times.

$$h_{ijk} = g(|t_{ij} - t_{ik}|)$$

for some decreasing function $g(\cdot)$ with $g(0) = 1$ (for correlation matrices).

Examples of this type of function:

- Exponential function: $g(|t_{ij} - t_{ik}|) = \exp(-\phi|t_{ij} - t_{ik}|)$
- Gaussian function: $g(|t_{ij} - t_{ik}|) = \exp(-\phi(t_{ij} - t_{ik})^2)$

Similar structures could also be used for \mathbf{D} matrix (of \mathbf{b})

Example: Autoregressive Covariance Structure

A first order Autoregressive Model (AR(1)) has the form

$$\alpha_t = \phi \alpha_{t-1} + \eta_t$$

where $\eta_t \sim iidN(0, \sigma_\eta^2)$

Then, the covariance between two observations is

$$cov(\alpha_t, \alpha_{t+h}) = \frac{\sigma_\eta^2 \phi^{|h|}}{1 - \phi^2}$$

for $h = 0, \pm 1, \pm 2, \dots; |\phi| < 1$

Hence,

$$\text{corr}(\alpha_t, \alpha_{t+h}) = \phi^{|h|}$$

If we let $\alpha_T = (\alpha_1, \dots, \alpha_T)'$, then

$$\text{corr}(\alpha_T) = \begin{bmatrix} 1 & \phi^1 & \phi^2 & \dots & \phi^2 \\ \phi^1 & 1 & \phi^1 & \dots & \phi^{T-1} \\ \phi^2 & \phi^1 & 1 & \dots & \phi^{T-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^T & \phi^{T-1} & \phi^{T-2} & \dots & 1 \end{bmatrix}$$

Notes:

- The correlation decreases as time lag increases
- This matrix structure is known as a **Toeplitz** structure
- More complicated covariance structures are possible, which is critical component of spatial random effects models and time series models.
- Often, we don't need both random effects \mathbf{b} and $\epsilon_{(1)i}$

More in the Time Series section

8.2 Estimation

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i$$

where $\beta, \mathbf{b}_i, \mathbf{D}, \epsilon_i$ we must obtain estimation from the data

- \mathbf{D}, ϵ_i are unknown, but fixed, parameters, and must be estimated from the data
- \mathbf{b}_i is a random variable. Thus, we can't estimate these values, but we can predict them. (i.e., you can't estimate a random thing).

If we have

- $\hat{\beta}$ as an estimator of β
- \mathbf{b}_i as a predictor of \mathbf{b}_i

Then,

- The population average estimate of \mathbf{Y}_i is $\hat{\mathbf{Y}}_i = \mathbf{X}_i\hat{\beta}$
- The subject-specific prediction is $\hat{\mathbf{Y}}_i = \mathbf{X}_i\hat{\beta} + \mathbf{Z}_i\hat{b}_i$

According to (Henderson, 1950), estimating equations known as the mixed model equations:

$$\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'^{-1}\mathbf{X} & \mathbf{X}'^{-1}\mathbf{Z} \\ \mathbf{Z}'^{-1}\mathbf{X} & \mathbf{Z}'^{-1}\mathbf{Z} + \mathbf{B}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{X}'^{-1}\mathbf{Y} \\ \mathbf{Z}'^{-1}\mathbf{Y} \end{bmatrix}$$

where

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{bmatrix}; \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_N \end{bmatrix}; \mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_N \end{bmatrix}; \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_N \end{bmatrix} cov(\boldsymbol{\epsilon}) = , \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{Z}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{Z}_n \end{bmatrix}, \mathbf{B} = \begin{bmatrix}$$

The model has the form

$$\mathbf{Y} = \mathbf{X} + \mathbf{Z}\mathbf{b} + \mathbf{Y} \sim N(\mathbf{X}, \mathbf{Z}\mathbf{B}\mathbf{Z}' +)$$

If $\mathbf{V} = \mathbf{Z}\mathbf{B}\mathbf{Z}' +$, then the solutions to the estimating equations can be

$$\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}\hat{\mathbf{b}} = \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X})$$

The estimate $\hat{\beta}$ is a generalized least squares estimate.

The predictor, $\hat{\mathbf{b}}$ is the best linear unbiased predictor (BLUP), for \mathbf{b}

$$E(\hat{\beta}) = \beta var(\hat{\beta}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}E(\hat{\mathbf{b}}) = 0var(\hat{\mathbf{b}} - \mathbf{b}) = \mathbf{B} - \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B} + \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}$$

The variance here is the variance of the prediction error (mean squared prediction error, MSPE), which is more meaningful than $var(\hat{\mathbf{b}})$, since MSPE accounts for both variance and bias in the prediction.

To derive the mixed model equations, consider

$$= \mathbf{Y} - \mathbf{X} - \mathbf{Z}\mathbf{b}$$

Let $T = \sum_{i=1}^N n_i$ be the total number of observations (i.e., the length of $\mathbf{Y}, \boldsymbol{\epsilon}$) and Nq the length of \mathbf{b} . The joint distribution of $\mathbf{b}, \boldsymbol{\epsilon}$ is

$$f(\mathbf{b}, \boldsymbol{\epsilon}) = \frac{1}{(2\pi)^{(T+Nq)/2}} \left| \begin{bmatrix} \mathbf{B} & 0 \\ 0 & \mathbf{I}_{Nq} \end{bmatrix} \right|^{-1/2} \exp \left(-\frac{1}{2} \left[\begin{bmatrix} \mathbf{b} \\ \mathbf{Y} - \mathbf{X} - \mathbf{Z}\mathbf{b} \end{bmatrix} \right]' \left[\begin{bmatrix} \mathbf{B} & 0 \\ 0 & \mathbf{I}_{Nq} \end{bmatrix} \right]^{-1} \left[\begin{bmatrix} \mathbf{b} \\ \mathbf{Y} - \mathbf{X} - \mathbf{Z}\mathbf{b} \end{bmatrix} \right] \right)$$

Maximization of $f(\mathbf{b}, \epsilon)$ with respect to \mathbf{b} and β requires minimization of

$$Q = \begin{bmatrix} \mathbf{b} \\ \mathbf{Y} - \mathbf{X} - \mathbf{Z}\mathbf{b} \end{bmatrix}' \begin{bmatrix} \mathbf{B} & 0 \\ 0 & \mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{b} \\ \mathbf{Y} - \mathbf{X} - \mathbf{Z}\mathbf{b} \end{bmatrix} = \mathbf{b}'\mathbf{B}^{-1}\mathbf{b} + (\mathbf{Y} - \mathbf{X} - \mathbf{Z}\mathbf{b})'(\mathbf{Y} - \mathbf{X} - \mathbf{Z}\mathbf{b})^{-1}(\mathbf{Y} - \mathbf{X} - \mathbf{Z}\mathbf{b})$$

Setting the derivatives of Q with respect to \mathbf{b} and β to zero leads to the system of equations:

$$\begin{aligned} \mathbf{X}'^{-1}\mathbf{X} + \mathbf{X}'^{-1}\mathbf{Z}\mathbf{b} &= \mathbf{X}'^{-1}\mathbf{Y} \\ (\mathbf{Z}'^{-1}\mathbf{Z} + \mathbf{B}^{-1})\mathbf{b} + \mathbf{Z}'^{-1}\mathbf{X} &= \mathbf{Z}'^{-1}\mathbf{Y} \end{aligned}$$

Rearranging

$$\begin{bmatrix} \mathbf{X}'^{-1}\mathbf{X} & \mathbf{X}'^{-1}\mathbf{Z} \\ \mathbf{Z}'^{-1}\mathbf{X} & \mathbf{Z}'^{-1}\mathbf{Z} + \mathbf{B}^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'^{-1}\mathbf{Y} \\ \mathbf{Z}'^{-1}\mathbf{Y} \end{bmatrix}$$

Thus, the solution to the mixed model equations give:

$$\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'^{-1}\mathbf{X} & \mathbf{X}'^{-1}\mathbf{Z} \\ \mathbf{Z}'^{-1}\mathbf{X} & \mathbf{Z}'^{-1}\mathbf{Z} + \mathbf{B}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'^{-1}\mathbf{Y} \\ \mathbf{Z}'^{-1}\mathbf{Y} \end{bmatrix}$$

Equivalently,

Bayes' theorem

$$f(\mathbf{b}|\mathbf{Y}) = \frac{f(\mathbf{Y}|\mathbf{b})f(\mathbf{b})}{\int f(\mathbf{Y}|\mathbf{b})f(\mathbf{b})d\mathbf{b}}$$

where

- $f(\mathbf{Y}|\mathbf{b})$ is the “likelihood”
- $f(\mathbf{b})$ is the prior
- the denominator is the “normalizing constant”
- $f(\mathbf{b}|\mathbf{Y})$ is the posterior distribution

In this case

$$\mathbf{Y}|\mathbf{b} \sim N(\mathbf{X} + \mathbf{Z}\mathbf{b}, \sigma^2\mathbf{I}) \quad \mathbf{b} \sim N(\mathbf{0}, \mathbf{B})$$

The posterior distribution has the form

$$\mathbf{b}|\mathbf{Y} \sim N(\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}), (\mathbf{Z}'^{-1}\mathbf{Z} + \mathbf{B}^{-1})^{-1})$$

Hence, the best predictor (based on squared error loss)

$$E(\mathbf{b}|\mathbf{Y}) = \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X})$$

8.2.1 Estimating \mathbf{V}

If we have $\tilde{\mathbf{V}}$ (estimate of \mathbf{V}), then we can estimate:

$$\hat{\beta} = (\mathbf{X}'\tilde{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\tilde{\mathbf{V}}^{-1}\mathbf{Y}\hat{\mathbf{b}} = \mathbf{B}\mathbf{Z}'\tilde{\mathbf{V}}^{-1}(\mathbf{Y} - \mathbf{X})$$

where \mathbf{b} is **EBLUP** (estimated BLUP) or **empirical Bayes estimate**

Note:

- $\hat{v}ar(\hat{\beta})$ is a consistent estimator of $var(\hat{\beta})$ if $\tilde{\mathbf{V}}$ is a consistent estimator of \mathbf{V}
- However, $\hat{v}ar(\hat{\beta})$ is biased since the variability arises from estimating \mathbf{V} is not accounted for in the estimate.
- Hence, $\hat{v}ar(\hat{\beta})$ underestimates the true variability

Ways to estimate \mathbf{V}

- Maximum Likelihood Estimation (MLE)
- Restricted Maximum Likelihood (REML)
- Estimated Generalized Least Squares
- Bayesian Hierarchical Models (BHM)

8.2.1.1 Maximum Likelihood Estimation (MLE)

Grouping unknown parameters in Σ and B under a parameter vector θ . Under MLE, $\hat{\theta}$ and $\hat{\beta}$ maximize the likelihood $\mathbf{y} \sim N(\mathbf{X}, \mathbf{V}(\theta))$. Synonymously, $-2\log L(\mathbf{y}; \theta)$:

$$-2l(\theta, \beta, \mathbf{y}) = \log |\mathbf{V}(\theta)| + (\mathbf{y} - \mathbf{X})'\mathbf{V}(\theta)^{-1}(\mathbf{y} - \mathbf{X}) + N \log(2\pi)$$

- Step 1: Replace β with its maximum likelihood (where θ is known $\hat{\beta} = (\mathbf{X}'\mathbf{V}(\theta)^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}(\theta)^{-1}\mathbf{y}$)
- Step 2: Minimize the above equation with respect to θ to get the estimator $\hat{\theta}_{MLE}$
- Step 3: Substitute $\hat{\theta}_{MLE}$ back to get $\hat{\beta}_{MLE} = (\mathbf{X}'\mathbf{V}(\hat{\theta}_{MLE})^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}(\hat{\theta}_{MLE})^{-1}\mathbf{y}$
- Step 4: Get $\hat{\mathbf{b}}_{MLE} = \mathbf{B}(\hat{\theta}_{MLE})\mathbf{Z}'\mathbf{V}(\hat{\theta}_{MLE})^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}_{MLE})$

Note:

- $\hat{\theta}$ are typically negatively biased due to unaccounted fixed effects being estimated, which we could try to account for.

8.2.1.2 Restricted Maximum Likelihood (REML)

REML accounts for the number of estimated mean parameters by adjusting the objective function. Specifically, the likelihood of linear combination of the elements of \mathbf{y} is accounted for.

We have $\mathbf{K}'\mathbf{y}$, where \mathbf{K} is any $N \times (N - p)$ full-rank contrast matrix, which has columns orthogonal to the \mathbf{X} matrix (that is $\mathbf{K}'\mathbf{X} = 0$). Then,

$$\mathbf{K}'\mathbf{y} \sim N(0, \mathbf{K}'\mathbf{V}(\beta)\mathbf{K})$$

where β is no longer in the distribution

We can proceed to maximize this likelihood for the contrasts to get $\hat{\theta}_{REML}$, which does not depend on the choice of \mathbf{K} . And $\hat{\beta}$ are based on $\hat{\theta}$

Comparison REML and MLE

- Both methods are based upon the likelihood principle, and have desired properties for the estimates:
 - consistency
 - asymptotic normality
 - efficiency
- ML estimation provides estimates for fixed effects, while REML can't
- In balanced models, REML is identical to ANOVA
- REML accounts for df for the fixed effects in the model, which is important when \mathbf{X} is large relative to the sample size
- Changing β has no effect on the REML estimates of θ
- REML is less sensitive to outliers than MLE
- MLE is better than REML regarding model comparisons (e.g., AIC or BIC)

8.2.1.3 Estimated Generalized Least Squares

MLE and REML rely upon the Gaussian assumption. To overcome this issue, EGLS uses the first and second moments.

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i$$

where

- $\epsilon_i \sim (\mathbf{0}, \mathbf{V}_i)$
- $\mathbf{b}_i \sim (\mathbf{0}, \mathbf{D})$
- $cov(\epsilon_i, \mathbf{b}_i) = 0$

Then the EGLS estimator is

$$\begin{aligned}\hat{\beta}_{GLS} &= \left\{ \sum_{i=1}^n \mathbf{X}'_i \mathbf{V}_i^{-1} \mathbf{X}_i \right\}^{-1} \sum_{i=1}^n \mathbf{X}'_i \mathbf{V}_i^{-1} \mathbf{Y}_i \\ &= \{\mathbf{X}' \mathbf{V}^{-1} \mathbf{X}\}^{-1} \mathbf{X}' \mathbf{V}^{-1} \mathbf{Y}\end{aligned}$$

depends on the first two moments

- $E(\mathbf{Y}_i) = \mathbf{X}_i\beta$
- $var(\mathbf{Y}_i) = \mathbf{V}_i$

EGLS use $\hat{\mathbf{V}}$ for $\mathbf{V}(\cdot)$

$$\hat{\beta}_{EGLS} = \{\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X}\}^{-1} \mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{Y}$$

Hence, the fixed effects estimators for the MLE, REML, and EGLS are of the same form, except for the estimate of \mathbf{V}

In case of non-iterative approach, EGLS can be appealing when \mathbf{V} can be estimated without much computational burden.

8.2.1.4 Bayesian Hierarchical Models (BHM)

Joint distribution can be decomposed hierarchically in terms of the product of conditional distributions and a marginal distribution

$$f(A, B, C) = f(A|B, C)f(B|C)f(C)$$

Applying to estimate \mathbf{V}

$$\begin{aligned} f(\mathbf{Y}, \mathbf{b}, \boldsymbol{\theta}) &= f(\mathbf{Y} | \mathbf{b}, \boldsymbol{\theta})f(\mathbf{b} | \mathbf{Y}, \boldsymbol{\theta})f(\boldsymbol{\theta} | \mathbf{Y}, \mathbf{b}) \quad \text{based on probability decomposition} \\ &= f(\mathbf{Y} | \mathbf{b}, \boldsymbol{\theta})f(\mathbf{b} | \boldsymbol{\theta})f(\boldsymbol{\theta} | \mathbf{Y}) \quad \text{based on simplifying modeling assumptions} \end{aligned}$$

elaborate on the second equality, if we assume conditional independence (e.g., given $\boldsymbol{\theta}$, no additional info about \mathbf{b} is given by knowing β), then we can simply from the first equality

Using Bayes' rule

$$f(\boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\beta} | \mathbf{Y}) \propto f(\mathbf{Y} | \boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\beta})f(\mathbf{b} | \boldsymbol{\theta})f(\boldsymbol{\beta} | \mathbf{Y})$$

where

$$\mathbf{Y} | \boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\beta} \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \mathbf{V}(\boldsymbol{\theta})) \quad \mathbf{b} | \boldsymbol{\theta} \sim \mathbf{N}(\mathbf{0}, \mathbf{B}(\boldsymbol{\theta}))$$

and we also have to have prior distributions for $f(\boldsymbol{\beta}), f(\boldsymbol{\theta})$

With normalizing constant, we can obtain the posterior distribution. Typically, we can't get analytical solution right away. Hence, we can use Markov Chain Monte Carlo (MCMC) to obtain samples from the posterior distribution.

Bayesian Methods:

- account for the uncertainty in parameters estimates and accommodate the propagation of that uncertainty through the model
- can adjust prior information (i.e., priori) in parameters
- Can extend beyond Gaussian distributions
- but hard to implement algorithms and might have problem converging

8.3 Inference

8.3.1 Parameters $\boldsymbol{\beta}$

8.3.1.1 Wald test

We have

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \{\mathbf{X}'\mathbf{V}^{-1}(\boldsymbol{\theta})\mathbf{X}\}^{-1}\mathbf{X}'\mathbf{V}^{-1}(\boldsymbol{\theta})\mathbf{Y} \quad var(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})) = \{\mathbf{X}'\mathbf{V}^{-1}(\boldsymbol{\theta})\mathbf{X}\}^{-1}$$

We can use $\hat{\boldsymbol{\beta}}$ in place of $\boldsymbol{\beta}$ to approximate Wald test

$$H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{d}$$

With

$$W = (\mathbf{A} - \mathbf{d})' [\mathbf{A}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A} - \mathbf{d})$$

where $W \sim \chi^2_{rank(A)}$ under H_0 is true. However, it does not take into account variability from using $\hat{\theta}$ in place of θ , hence the standard errors are underestimated

8.3.1.2 F-test

Alternatively, we can use the modified F-test, suppose we have $var(\mathbf{Y}) = \sigma^2 \mathbf{V}(\theta)$, then

$$F^* = \frac{(\mathbf{A} - \mathbf{d})' [\mathbf{A}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A} - \mathbf{d})}{\hat{\sigma}^2 \text{rank}(A)}$$

where $F^* \sim f_{rank(A), den(df)}$ under the null hypothesis. And $den(df)$ needs to be approximated from the data by either:

- Satterthwaite method
- Kenward-Roger approximation

Under balanced cases, the Wald and F tests are similar. But for small sample sizes, they can differ in p-values. And both can be reduced to t-test for a single β

8.3.1.3 Likelihood Ratio Test

$$H_0 : \beta \in \Theta_{\beta,0}$$

where $\Theta_{\beta,0}$ is a subspace of the parameter space, Θ_β of the fixed effects β . Then

$$-2 \log \lambda_N = -2 \log \left\{ \frac{\hat{L}_{ML,0}}{\hat{L}_{ML}} \right\}$$

where

- $\hat{L}_{ML,0}, \hat{L}_{ML}$ are the maximized likelihood obtained from maximizing over $\Theta_{\beta,0}$ and Θ_β
- $-2 \log \lambda_N \sim \chi^2_{df}$ where df is the difference in the dimension (i.e., number of parameters) of $\Theta_{\beta,0}$ and Θ_β

This method is not applicable for REML. But REML can still be used to test for covariance parameters between nested models.

8.3.2 Variance Components

- For ML and REML estimator, $\hat{\theta} \sim N(\theta, I(\theta))$ for large samples
- Wald test in variance components is analogous to the fixed effects case (see 8.3.1.1)
 - However, the normal approximation depends largely on the true value of θ . It will fail if the true value of θ is close to the boundary of the parameter space Θ_θ (i.e., $\sigma^2 \approx 0$)
 - Typically works better for covariance parameter, than variance parameters.
- The likelihood ratio tests can also be used with ML or REML estimates. However, the same problem of parameters

8.4 Information Criteria

- account for the likelihood and the number of parameters to assess model comparison.

8.4.1 Akaike's Information Criteria (AIC)

Derived as an estimator of the expected Kullback discrepancy between the true model and a fitted candidate model

$$AIC = -2l(\hat{\theta}, \hat{\beta}) + 2q$$

where

- $l(\hat{\theta}, \hat{\beta})$ is the log-likelihood
- q = the effective number of parameters; total of fixed and those associated with random effects (variance/covariance; those not estimated to be on a boundary constraint)

Note:

- In comparing models that differ in their random effects, this method is not advised to due the inability to get the correct number of effective parameters).
- We prefer smaller AIC values.
- If your program uses $l - q$ then we prefer larger AIC values (but rarely).

- can be used for mixed model section, (e.g., selection of the covariance structure), but the sample size must be very large to have adequate comparison based on the criterion
- Can have a large negative bias (e.g., when sample size is small but the number of parameters is large) due to the penalty term can't approximate the bias adjustment adequately

8.4.2 Corrected AIC (AICC)

- developed by (HURVICH and TSAI, 1989)
- correct small-sample adjustment
- depends on the candidate model class
- Only if you have fixed covariance structure, then AICC is justified, but not general covariance structure

8.4.3 Bayesian Information Criteria (BIC)

$$BIC = -2l(\hat{\theta}, \hat{\beta}) + q \log n$$

where n = number of observations.

- we prefer smaller BIC value
- BIC and AIC are used for both REML and MLE if we have the same mean structure. Otherwise, in general, we should prefer MLE

With our example presented at the beginning of Linear Mixed Models,

$$Y_{ik} = \begin{cases} \beta_0 + b_{1i} + (\beta_1 + b_{2i})t_{ij} + \epsilon_{ij} & L \\ \beta_0 + b_{1i} + (\beta_2 + b_{2i})t_{ij} + \epsilon_{ij} & H \\ \beta_0 + b_{1i} + (\beta_3 + b_{2i})t_{ij} + \epsilon_{ij} & C \end{cases}$$

where

- $i = 1, \dots, N$
- $j = 1, \dots, n_i$ (measures at time t_{ij})

Note:

- we have subject-specific intercepts,

$$\mathbf{Y}_i | b_i \sim N(\mathbf{X}_i \beta + 1b_i, \sigma^2 \mathbf{I}) b_i \sim N(0, d_{11})$$

here, we want to estimate β, σ^2, d_{11} and predict b_i

8.5 Split-Plot Designs

- Typically used in the case that you have two factors where one needs much larger units than the other.

Example:

A: 3 levels (large units)

B: 2 levels (small units)

- A and B levels are randomized into 4 blocks.
- But it differs from Randomized Block Designs. In each block, both have one of the 6 (3x2) treatment combinations. But Randomized Block Designs assign in each block randomly, while split-plot does not randomize this step.
- Moreover, because A needs to be applied in large units, factor A is applied only once in each block while B can be applied multiple times.

Hence, we have our model

If A is our factor of interest

$$Y_{ij} = \mu + \rho_i + \alpha_j + e_{ij}$$

where

- i = replication (block or subject)
- j = level of Factor A
- μ = overall mean
- ρ_i = variation due to the i-th block
- $e_{ij} \sim N(0, \sigma_e^2)$ = whole plot error

If B is our factor of interest

$$Y_{ijk} = \mu + \phi_{ij} + \beta_k + \epsilon_{ijk}$$

where

- ϕ_{ij} = variation due to the ij-th main plot
- β_k = Factor B effect
- $\epsilon_{ijk} \sim N(0, \sigma_e^2)$ = subplot error
- $\phi_{ij} = \rho_i + \alpha_j + e_{ij}$

Together, the split-plot model

$$Y_{ijk} = \mu + \rho_i + \alpha_j + e_{ij} + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk}$$

where

- i = replicate (blocks or subjects)
- j = level of factor A
- k = level of factor B
- μ = overall mean
- ρ_i = effect of the block
- α_j = main effect of factor A (fixed)
- $e_{ij} = (\rho\alpha)_{ij}$ = block by factor A interaction (the whole plot error, random)
- β_k = main effect of factor B (fixed)
- $(\alpha\beta)_{jk}$ = interaction between factors A and B (fixed)
- ϵ_{ijk} = subplot error (random)

We can approach sub-plot analysis based on

- the ANOVA perspective
 - Whole plot comparisons
 - * Compare factor A to the whole plot error (i.e., α_j to e_{ij})
 - * Compare the block to the whole plot error (i.e., ρ_i to e_{ij})
 - Sub-plot comparisons:
 - * Compare factor B to the subplot error (β to ϵ_{ijk})
 - * Compare the AB interaction to the subplot error ($(\alpha\beta)_{jk}$ to ϵ_{ijk})
- the mixed model perspective

$$\mathbf{Y} = \mathbf{X} + \mathbf{Z}\mathbf{b} +$$

8.5.1 Application

8.5.1.1 Example 1

$$y_{ijk} = \mu + i_i + v_j + (iv)_{ij} + f_k + \epsilon_{ijk}$$

where

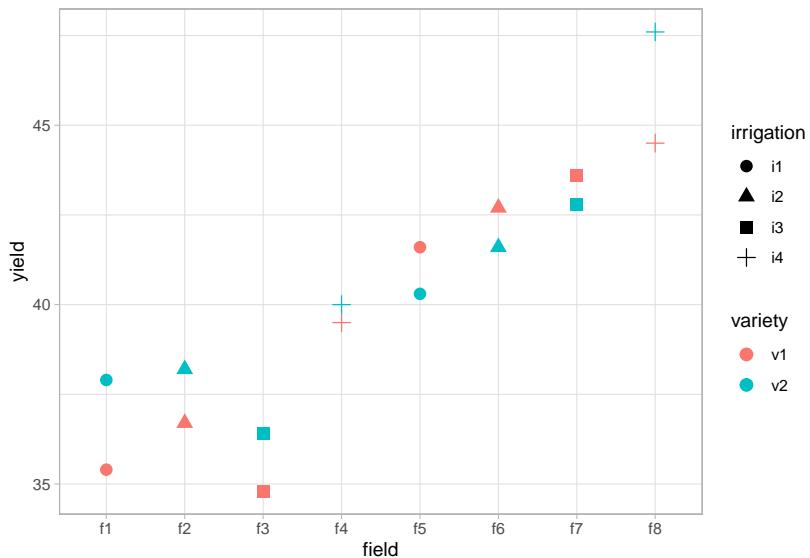
- y_{ijk} = observed yield

- μ = overall average yield
- i_i = irrigation effect
- v_j = variety effect
- $(iv)_{ij}$ = irrigation by variety interaction
- f_k = random field (block) effect
- ϵ_{ijk} = residual
- because variety-field combination is only observed once, we can't have the random interaction effects between variety and field

```

library(ggplot2)
data(irrigation, package = "faraway")
summary(irrigation)
#>      field  irrigation variety     yield
#>    f1      :2  i1:4        v1:8   Min.  :34.80
#>    f2      :2  i2:4        v2:8   1st Qu.:37.60
#>    f3      :2  i3:4          Median :40.15
#>    f4      :2  i4:4          Mean   :40.23
#>    f5      :2          3rd Qu.:42.73
#>    f6      :2          Max.   :47.60
#>  (Other):4
head(irrigation, 4)
#>      field  irrigation variety yield
#> 1    f1        i1        v1  35.4
#> 2    f1        i1        v2  37.9
#> 3    f2        i2        v1  36.7
#> 4    f2        i2        v2  38.2
ggplot(irrigation,
       aes(
         x = field,
         y = yield,
         shape = irrigation,
         color = variety
       )) +
  geom_point(size = 3)

```



```

sp_model <- lmerTest::lmer(yield ~ irrigation * variety + (1 | field), irrigation)
summary(sp_model)
#> Linear mixed model fit by REML. t-tests use Satterthwaite's method [
#> lmerModLmerTest]
#> Formula: yield ~ irrigation * variety + (1 | field)
#>   Data: irrigation
#>
#> REML criterion at convergence: 45.4
#>
#> Scaled residuals:
#>    Min      1Q  Median      3Q     Max
#> -0.7448 -0.5509  0.0000  0.5509  0.7448
#>
#> Random effects:
#> Groups   Name        Variance Std.Dev.
#> field    (Intercept) 16.200   4.025
#> Residual           2.107   1.452
#> Number of obs: 16, groups: field, 8
#>
#> Fixed effects:
#>             Estimate Std. Error    df t value Pr(>|t|)    
#> (Intercept) 38.500    3.026 12.725 0.000109 ***
#> irrigationi2  1.200    4.279 4.487  0.280 0.791591  
#> irrigationi3  0.700    4.279 4.487  0.164 0.877156  
#> irrigationi4  3.500    4.279 4.487  0.818 0.454584  
#> varietyv2     0.600    1.452 4.000  0.413 0.700582  

```

```

#> irrigationi2:varietyv2 -0.400    2.053 4.000 -0.195 0.855020
#> irrigationi3:varietyv2 -0.200    2.053 4.000 -0.097 0.927082
#> irrigationi4:varietyv2  1.200    2.053 4.000  0.584 0.590265
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Correlation of Fixed Effects:
#>           (Intr) irrigt2 irrigt3 irrigt4 vrtyv2 irr2:2 irr3:2
#> irrigation2 -0.707
#> irrigation3 -0.707  0.500
#> irrigation4 -0.707  0.500  0.500
#> varietyv2   -0.240  0.170  0.170  0.170
#> irrigtn2:vr2 0.170 -0.240 -0.120 -0.120 -0.707
#> irrigtn3:vr2 0.170 -0.120 -0.240 -0.120 -0.707  0.500
#> irrigtn4:vr2 0.170 -0.120 -0.120 -0.240 -0.707  0.500  0.500

anova(sp_model,ddf = c("Kenward-Roger"))
#> Type III Analysis of Variance Table with Kenward-Roger's method
#>           Sum Sq Mean Sq NumDF DenDF F value Pr(>F)
#> irrigation      2.4545 0.81818     3      4  0.3882 0.7685
#> variety         2.2500 2.25000     1      4  1.0676 0.3599
#> irrigation:variety 1.5500 0.51667     3      4  0.2452 0.8612

```

Since p-value of the interaction term is insignificant, we consider fitting without it.

```

library(lme4)
sp_model_additive <- lmer(yield ~ irrigation + variety + (1 | field), irrigation)
anova(sp_model_additive,sp_model,ddf = "Kenward-Roger")
#> Data: irrigation
#> Models:
#> sp_model_additive: yield ~ irrigation + variety + (1 | field)
#> sp_model: yield ~ irrigation * variety + (1 | field)
#>           npar   AIC   BIC logLik deviance Chisq Df Pr(>Chisq)
#> sp_model_additive  7 83.959 89.368 -34.980   69.959
#> sp_model          10 88.609 96.335 -34.305   68.609 1.3503  3      0.7172

```

Since p-value of Chi-square test is insignificant, we can't reject the additive model is already sufficient. Looking at AIC and BIC, we can also see that we would prefer the additive model

Random Effect Examination

`exactRLRT` test

- H_0 : $\text{Var}(\text{random effect})$ (i.e., σ^2) = 0

- H_a : $\text{Var}(\text{random effect})$ (i.e., σ^2) > 0

```
sp_model <- lme4::lmer(yield ~ irrigation * variety + (1 | field), irrigation)
library(RLRsim)
exactRLRT(sp_model)
#>
#> simulated finite sample distribution of RLRT.
#>
#> (p-value based on 10000 simulated values)
#>
#> data:
#> RLRT = 6.1118, p-value = 0.0096
```

Since the p-value is significant, we reject H_0

8.6 Repeated Measures in Mixed Models

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_{i(k)} + \epsilon_{ijk}$$

where

- i-th group (fixed)
- j-th (repeated measure) time effect (fixed)
- k-th subject
- $\delta_{i(k)} \sim N(0, \sigma_\delta^2)$ (k-th subject in the i-th group) and $\epsilon_{ijk} \sim N(0, \sigma^2)$ (independent error) are random effects ($i = 1, \dots, n_A, j = 1, \dots, n_B, k = 1, \dots, n_i$)

hence, the variance-covariance matrix of the repeated observations on the k-th subject of the i-th group, $\mathbf{Y}_{ik} = (Y_{i1k}, \dots, Y_{in_B k})'$, will be

$$\begin{aligned} \text{subject} &= \begin{pmatrix} \sigma_\delta^2 + \sigma^2 & \sigma_\delta^2 & \dots & \sigma_\delta^2 \\ \sigma_\delta^2 & \sigma_\delta^2 + \sigma^2 & \dots & \sigma_\delta^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\delta^2 & \sigma_\delta^2 & \dots & \sigma_\delta^2 + \sigma^2 \end{pmatrix} \\ &= (\sigma_\delta^2 + \sigma^2) \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix} \quad \text{product of a scalar and a correlation matrix} \end{aligned}$$

where $\rho = \frac{\sigma_\delta^2}{\sigma_\delta^2 + \sigma^2}$, which is the compound symmetry structure that we discussed in Random-Intercepts Model

But if you only have repeated measurements on the subject over time, AR(1) structure might be more appropriate

Mixed model for a repeated measure

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where

- ϵ_{ijk} combines random error of both the whole and subplots.

In general,

$$\mathbf{Y} = \mathbf{X} +$$

where

- $\epsilon \sim N(0, \sigma^2)$ where Σ is block diagonal if the random error covariance is the same for each subject

The variance covariance matrix with AR(1) structure is

$$\text{subject} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n_B-1} \\ \rho & 1 & \rho & \dots & \rho^{n_B-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n_B-1} & \rho^{n_B-2} & \rho^{n_B-3} & \dots & 1 \end{pmatrix}$$

Hence, the mixed model for a repeated measure can be written as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where

- ϵ_{ijk} = random error of whole and subplots

Generally,

$$\mathbf{Y} = \mathbf{X} +$$

where $\epsilon \sim N(0, \Sigma)$ and Σ = block diagonal if the random error covariance is the same for each subject.

8.7 Unbalanced or Unequally Spaced Data

Consider the model

$$Y_{ikt} = \beta_0 + \beta_{0i} + \beta_1 t + \beta_{1i} t + \beta_2 t^2 + \beta_{2i} t^2 + \epsilon_{ikt}$$

where

- $i = 1, 2$ (groups)
- $k = 1, \dots, n_i$ (individuals)
- $t = (t_1, t_2, t_3, t_4)$ (times)
- β_{2i} = common quadratic term
- β_{1i} = common linear time trends
- β_{0i} = common intercepts

Then, we assume the variance-covariance matrix of the repeated measurements collected on a particular subject over time has the form

$$_{ik} = \sigma^2 \begin{pmatrix} 1 & \rho^{t_2-t_1} & \rho^{t_3-t_1} & \rho^{t_4-t_1} \\ \rho^{t_2-t_1} & 1 & \rho^{t_3-t_2} & \rho^{t_4-t_2} \\ \rho^{t_3-t_1} & \rho^{t_3-t_2} & 1 & \rho^{t_4-t_3} \\ \rho^{t_4-t_1} & \rho^{t_4-t_2} & \rho^{t_4-t_3} & 1 \end{pmatrix}$$

which is called “power” covariance model

We can consider $\beta_{2i}, \beta_{1i}, \beta_{0i}$ accordingly to see whether these terms are needed in the final model

8.8 Application

R Packages for mixed models

- **nlme**
 - has nested structure
 - flexible for complex design
 - not user-friendly
- **lme4**
 - computationally efficient
 - user-friendly

- can handle nonnormal response
- for more detailed application, check Fitting Linear Mixed-Effects Models Using lme4
- Others
 - Bayesian setting: MCMCglmm, brms
 - For genetics: ASReml

8.8.1 Example 1 (Pulps)

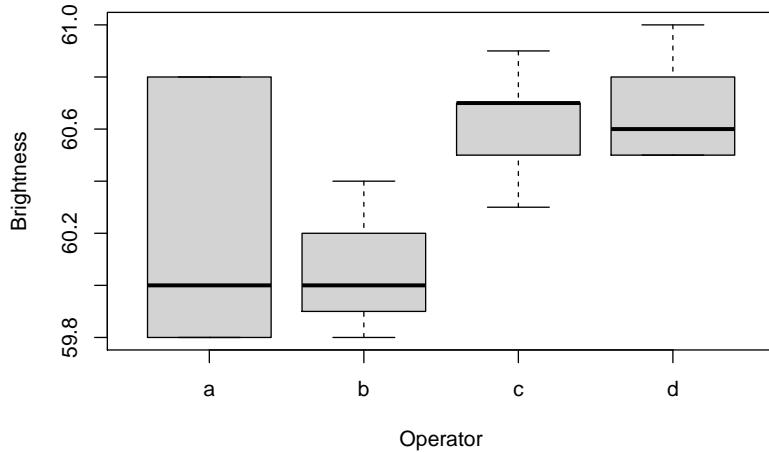
Model:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

where

- $i = 1, \dots, a$ groups for random effect α_i
- $j = 1, \dots, n$ individuals in each group
- $\alpha_i \sim N(0, \sigma_\alpha^2)$ is random effects
- $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ is random effects
- Imply compound symmetry model where the intraclass correlation coefficient is: $\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}$
- If factor a does not explain much variation, low correlation within the levels: $\sigma_\alpha^2 \rightarrow 0$ then $\rho \rightarrow 0$
- If factor a explain much variation, high correlation within the levels $\sigma_\alpha^2 \rightarrow \infty$ hence, $\rho \rightarrow 1$

```
data(pulp, package = "faraway")
plot(
  y = pulp$bright,
  x = pulp$operator,
  xlab = "Operator",
  ylab = "Brightness"
)
```



```
pulp %>% dplyr::group_by(operator) %>% dplyr::summarise(average = mean(bright))
#> # A tibble: 4 x 2
#>   operator     average
#>   <fct>       <dbl>
#> 1 a            60.2
#> 2 b            60.1
#> 3 c            60.6
#> 4 d            60.7
```

lmer application

```
library(lme4)
mixed_model <- lmer(formula = bright ~ 1 + (1 | operator), # pipe (i.e., / ) denotes r
                     data = pulp)
summary(mixed_model)
#> Linear mixed model fit by REML. t-tests use Satterthwaite's method [ 
#> lmerModLmerTest]
#> Formula: bright ~ 1 + (1 | operator)
#> Data: pulp
#>
#> REML criterion at convergence: 18.6
#>
#> Scaled residuals:
#>      Min      1Q  Median      3Q     Max
#> -1.4666 -0.7595 -0.1244  0.6281  1.6012
#>
```

```

#> Random effects:
#> Groups Name Variance Std.Dev.
#> operator (Intercept) 0.06808 0.2609
#> Residual 0.10625 0.3260
#> Number of obs: 20, groups: operator, 4
#>
#> Fixed effects:
#> Estimate Std. Error df t value Pr(>|t|)
#> (Intercept) 60.4000 0.1494 3.0000 404.2 3.34e-08 ***
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
coef(mixed_model)
#> $operator
#> (Intercept)
#> a 60.27806
#> b 60.14088
#> c 60.56767
#> d 60.61340
#>
#> attr(", "class")
#> [1] "coef.mer"
fixef(mixed_model) # fixed effects
#> (Intercept)
#> 60.4
confint(mixed_model) # confidence interval
#> 2.5 % 97.5 %
#> .sig01 0.000000 0.6178987
#> .sigma 0.238912 0.4821845
#> (Intercept) 60.071299 60.7287012
ranef(mixed_model) # random effects
#> $operator
#> (Intercept)
#> a -0.1219403
#> b -0.2591231
#> c 0.1676679
#> d 0.2133955
#>
#> with conditional variances for "operator"
VarCorr(mixed_model) # random effects standard deviation
#> Groups Name Std.Dev.
#> operator (Intercept) 0.26093
#> Residual 0.32596
re_dat = as.data.frame(VarCorr(mixed_model))
rho = re_dat[1, 'vcov']/(re_dat[1, 'vcov'] + re_dat[2, 'vcov']) # rho based on the above formula
rho

```

```
#> [1] 0.3905354
```

To Satterthwaite approximation for the denominator df, we use `lmerTest`

```
library(lmerTest)
summary(lmerTest::lmer(bright ~ 1 + (1 | operator), pulp))$coefficients
#>             Estimate Std. Error df t value    Pr(>|t|)
#> (Intercept)   60.4  0.1494434 3 404.1664 3.340265e-08
confint(mixed_model)[3,]
#> 2.5 % 97.5 %
#> 60.0713 60.7287
```

In this example, we can see that the confidence interval computed by `confint` in `lmer` package is very close is `confint` in `lmerTest` model.

`MCMglmm` application

under the Bayesian framework

```
library(MCMCglmm)
mixed_model_bayes <- MCMCglmm(bright~1,random=~operator, data=pulp, verbose=FALSE)
summary(mixed_model_bayes)$solutions
#>           post.mean l-95% CI u-95% CI eff.samp pMCMC
#> (Intercept) 60.39767 60.15776 60.66305      1000 0.001
```

this method offers the confidence interval slightly more positive than `lmer` and `lmerTest`

8.8.1.1 Prediction

```
# random effects prediction (BLUPs)
ranef(mixed_model)$operator
#> (Intercept)
#> a -0.1219403
#> b -0.2591231
#> c 0.1676679
#> d 0.2133955
fixef(mixed_model) + ranef(mixed_model)$operator #prediction for each categories
#> (Intercept)
#> a 60.27806
#> b 60.14088
#> c 60.56767
#> d 60.61340
```

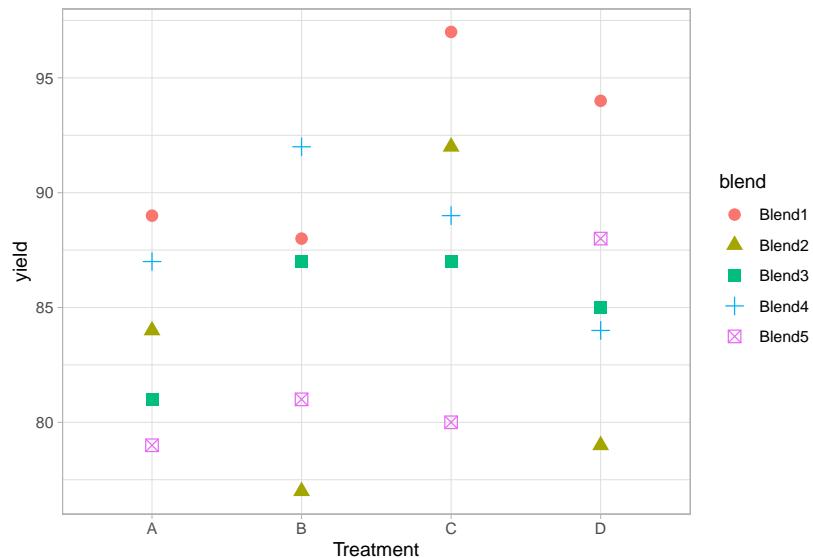
```
predict(mixed_model, newdata=data.frame(operator=c('a','b','c','d'))) # equivalent to the above
#>      1      2      3      4
#> 60.27806 60.14088 60.56767 60.61340
```

use `bootMer()` to get bootstrap-based confidence intervals for predictions.

Another example using GLMM in the context of blocking

Penicillin data

```
data(penicillin, package = "faraway")
summary(penicillin)
#>   treat    blend      yield
#>   A:5    Blend1:4   Min.   :77
#>   B:5    Blend2:4   1st Qu.:81
#>   C:5    Blend3:4   Median :87
#>   D:5    Blend4:4   Mean   :86
#>          Blend5:4   3rd Qu.:89
#>                      Max.   :97
library(ggplot2)
ggplot(penicillin, aes(
  y = yield,
  x = treat,
  shape = blend,
  color = blend
)) + # treatment = fixed effect, blend = random effects
  geom_point(size = 3) +
  xlab("Treatment")
```



```

library(lmerTest) # for p-values
mixed_model <- lmerTest::lmer(yield ~ treat + (1 | blend),
                             data = penicillin)
summary(mixed_model)
#> Linear mixed model fit by REML. t-tests use Satterthwaite's method [ 
#> lmerModLmerTest]
#> Formula: yield ~ treat + (1 / blend)
#> Data: penicillin
#>
#> REML criterion at convergence: 103.8
#>
#> Scaled residuals:
#>      Min     1Q Median     3Q    Max
#> -1.4152 -0.5017 -0.1644  0.6830  1.2836
#>
#> Random effects:
#> Groups   Name        Variance Std.Dev.
#> blend    (Intercept) 11.79    3.434
#> Residual           18.83    4.340
#> Number of obs: 20, groups: blend, 5
#>
#> Fixed effects:
#>             Estimate Std. Error    df t value Pr(>|t|)    
#> (Intercept) 84.000     2.475 11.075 33.941 1.51e-12 ***
#> treatB       1.000     2.745 12.000  0.364  0.7219    
#> treatC       5.000     2.745 12.000  1.822  0.0935 .  

```

```
#> treatD      2.000     2.745 12.000   0.729   0.4802
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Correlation of Fixed Effects:
#>          (Intr) treatB treatC
#> treatB -0.555
#> treatC -0.555  0.500
#> treatD -0.555  0.500  0.500

#The BLUPs for the each blend
ranef(mixed_model)$blend
#>           (Intercept)
#> Blend1    4.2878788
#> Blend2   -2.1439394
#> Blend3   -0.7146465
#> Blend4    1.4292929
#> Blend5   -2.8585859
```

Examine treatment effect

```
anova(mixed_model) # p-value based on lmerTest
#> Type III Analysis of Variance Table with Satterthwaite's method
#>           Sum Sq Mean Sq NumDF DenDF F value Pr(>F)
#> treat      70  23.333     3     12  1.2389 0.3387
```

Since the p-value is greater than 0.05, we can't reject the null hypothesis that there is no treatment effect.

```
library(pbkrtest)
full_model <- lmer(yield ~ treat + (1 | blend), penicillin, REML=FALSE) #REML is not appropriate
null_model <- lmer(yield ~ 1 + (1 | blend), penicillin, REML=FALSE)
KRmodcomp(full_model, null_model) # use Kenward-Roger approximation for df
#> large : yield ~ treat + (1 / blend)
#> small : yield ~ 1 + (1 / blend)
#>           stat      ndf      ddf F.scaling p.value
#> Ftest   1.2389  3.0000 12.0000      1  0.3387
```

Since the p-value is greater than 0.05, and consistent with our previous observation, we conclude that we can't reject the null hypothesis that there is no treatment effect.

8.8.2 Example 2 (Rats)

```

rats <- read.csv(
  "images/rats.dat",
  header = F,
  sep = ' ',
  col.names = c('Treatment', 'rat', 'age', 'y')
)
rats$t <- log(1 + (rats$age - 45)/10) #log transformed age

```

We are interested in whether treatment effect induces changes over time.

```

rat_model <- lmerTest::lmer(y~t:Treatment+(1|rat),data=rats) #treatment = fixed effect
summary(rat_model)
#> Linear mixed model fit by REML. t-tests use Satterthwaite's method [
#> lmerModLmerTest]
#> Formula: y ~ t:Treatment + (1 / rat)
#>   Data: rats
#>
#> REML criterion at convergence: 932.4
#>
#> Scaled residuals:
#>      Min     1Q Median     3Q    Max
#> -2.25574 -0.65898 -0.01163  0.58356  2.88309
#>
#> Random effects:
#> Groups   Name        Variance Std.Dev.
#> rat      (Intercept) 3.565    1.888
#> Residual           1.445    1.202
#> Number of obs: 252, groups:  rat, 50
#>
#> Fixed effects:
#>             Estimate Std. Error       df t value Pr(>|t|)
#> (Intercept) 68.6074   0.3312  89.0275 207.13 <2e-16 ***
#> t:Treatmentcon 7.3138   0.2808 247.2762  26.05 <2e-16 ***
#> t:Treatmenthig 6.8711   0.2276 247.7097  30.19 <2e-16 ***
#> t:Treatmentlow 7.5069   0.2252 247.5196  33.34 <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Correlation of Fixed Effects:
#>            (Intr) t:Trtmntc t:Trtmnth
#> t:Trtmntcn -0.327
#> t:Trtmnthg -0.340  0.111

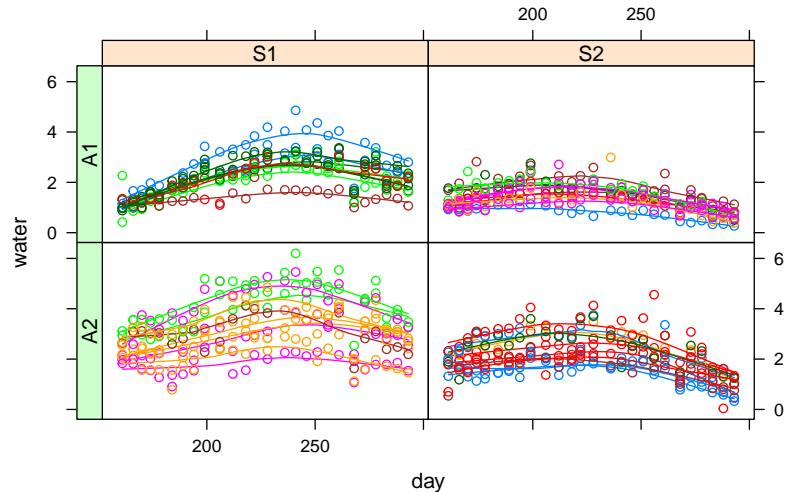
```

```
#> t:Tretmntlw -0.351  0.115     0.119
anova(rat_model)
#> Type III Analysis of Variance Table with Satterthwaite's method
#>           Sum Sq Mean Sq NumDF DenDF F value    Pr(>F)
#> t:Treatment 3181.9 1060.6      3 223.21 734.11 < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is significant, we can be confident concluding that there is a treatment effect

8.8.3 Example 3 (Agridat)

```
library(agridat)
library(latticeExtra)
dat <- harris.wateruse
# Compare to Schabenberger & Pierce, fig 7.23
useOuterStrips(
  xyplot(
    water ~ day | species * age,
    dat,
    as.table = TRUE,
    group = tree,
    type = c('p', 'smooth'),
    main = "harris.wateruse 2 species, 2 ages (10 trees each)"
  )
)
```

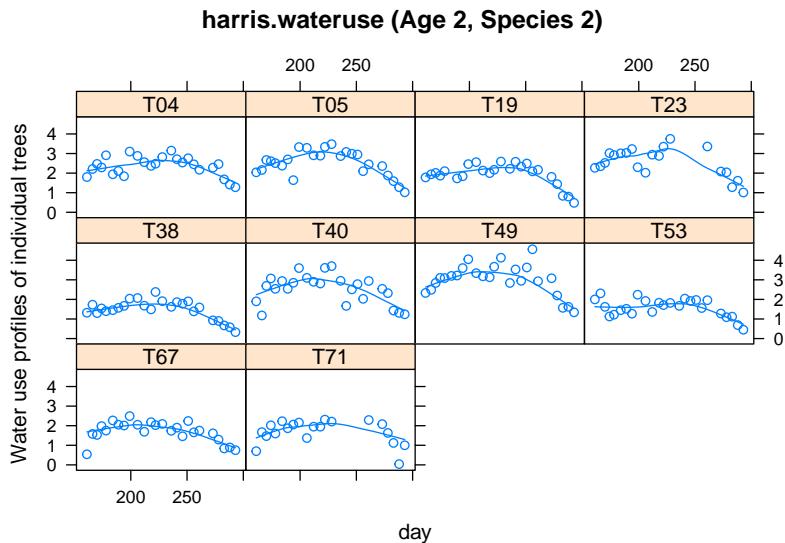
harris.wateruse 2 species, 2 ages (10 trees each)

Remove outliers

```
dat <- subset(dat, day!=268)
```

Plot between age and species

```
xyplot(
  water ~ day | tree,
  dat,
  subset = age == "A2" & species == "S2",
  as.table = TRUE,
  type = c('p', 'smooth'),
  ylab = "Water use profiles of individual trees",
  main = "harris.wateruse (Age 2, Species 2)"
)
```



```
# Rescale day for nicer output, and convergence issues, add quadratic term
dat <- transform(dat, ti = day / 100)
dat <- transform(dat, ti2 = ti * ti)
# Start with a subgroup: age 2, species 2
d22 <- droplevels(subset(dat, age == "A2" & species == "S2"))
```

lme function from nlme package

```
library(nlme)
## We use pdDiag() to get uncorrelated random effects
m1n <- lme(
  water ~ 1 + ti + ti2, #intercept, time and time-squared = fixed effects
  data = d22,
  na.action = na.omit,
  random = list(tree = pdDiag( ~ 1 + ti + ti2)) # random intercept, time and time squared per tree
)
ranef(m1n)
#>   (Intercept)          ti          ti2
#> T04  0.1985796  1.609864e-09  4.990101e-10
#> T05  0.3492827  2.487690e-10 -4.845287e-11
#> T19 -0.1978989 -7.681202e-10 -1.961453e-10
#> T23  0.4519003 -3.270426e-10 -2.413583e-10
#> T38 -0.6457494 -1.608770e-09 -3.298010e-10
#> T40  0.3739432  3.264705e-10 -2.543109e-11
#> T49  0.8620648  9.021831e-10 -5.402247e-12
#> T53 -0.5655049 -8.279040e-10 -4.579291e-11
```

```
#> T67 -0.4394623 -3.485113e-10 2.147434e-11
#> T71 -0.3871552 7.930610e-10 3.718993e-10

fixef(m1n)
#> (Intercept) ti ti2
#> -10.798799 12.346704 -2.838503
summary(m1n)
#> Linear mixed-effects model fit by REML
#> Data: d22
#> AIC BIC logLik
#> 276.5142 300.761 -131.2571
#>
#> Random effects:
#> Formula: ~1 + ti + ti2 | tree
#> Structure: Diagonal
#> (Intercept) ti ti2 Residual
#> StdDev: 0.5187869 1.438333e-05 3.864019e-06 0.3836614
#>
#> Fixed effects: water ~ 1 + ti + ti2
#> Value Std.Error DF t-value p-value
#> (Intercept) -10.798799 0.8814666 227 -12.25094 0
#> ti 12.346704 0.7827112 227 15.77428 0
#> ti2 -2.838503 0.1720614 227 -16.49704 0
#> Correlation:
#> (Intr) ti
#> ti -0.979
#> ti2 0.970 -0.997
#>
#> Standardized Within-Group Residuals:
#> Min Q1 Med Q3 Max
#> -3.07588246 -0.58531056 0.01210209 0.65402695 3.88777402
#>
#> Number of Observations: 239
#> Number of Groups: 10
```

lmer function from lme4 package

```
m1lmer <- lmer(water~1+ti+ti2+(ti+ti2||tree), data = d22, na.action = na.omit)
ranef(m1lmer)
#> $tree
#> (Intercept) ti ti2
#> T04 0.1985796 0 0
#> T05 0.3492827 0 0
#> T19 -0.1978989 0 0
#> T23 0.4519003 0 0
```

```
#> T38 -0.6457494 0 0
#> T40 0.3739432 0 0
#> T49 0.8620648 0 0
#> T53 -0.5655049 0 0
#> T67 -0.4394623 0 0
#> T71 -0.3871552 0 0
#>
#> with conditional variances for "tree"
```

Notes:

- || double pipes= uncorrelated random effects
- To remove the intercept term:
 - (0+ti|tree)
 - (ti-1|tree)

```
fixef(m1lmer)
#> (Intercept) ti ti2
#> -10.798799 12.346704 -2.838503
m1l <- lmer(water ~ 1 + ti + ti2 + (1 | tree) + (0 + ti | tree) + (0 + ti2 | tree), data = d22)
ranef(m1l)
#> $tree
#> (Intercept) ti ti2
#> T04 0.1985796 0 0
#> T05 0.3492827 0 0
#> T19 -0.1978989 0 0
#> T23 0.4519003 0 0
#> T38 -0.6457494 0 0
#> T40 0.3739432 0 0
#> T49 0.8620648 0 0
#> T53 -0.5655049 0 0
#> T67 -0.4394623 0 0
#> T71 -0.3871552 0 0
#>
#> with conditional variances for "tree"
fixef(m1l)
#> (Intercept) ti ti2
#> -10.798799 12.346704 -2.838503
```

To include structured covariance terms, we can use the following way

```

m2n <- lme(
  water ~ 1 + ti + ti2,
  data = d22,
  random = ~ 1 | tree,
  cor = corExp(form = ~ day | tree),
  na.action = na.omit
)
ranef(m2n)
#>   (Intercept)
#> T04    0.1929971
#> T05    0.3424631
#> T19   -0.1988495
#> T23    0.4538660
#> T38   -0.6413664
#> T40    0.3769378
#> T49    0.8410043
#> T53   -0.5528236
#> T67   -0.4452930
#> T71   -0.3689358
fixef(m2n)
#> (Intercept)      ti      ti2
#> -11.223310  12.712094 -2.913682
summary(m2n)
#> Linear mixed-effects model fit by REML
#> Data: d22
#>      AIC      BIC logLik
#> 263.3081 284.0911 -125.654
#>
#> Random effects:
#> Formula: ~1 / tree
#>   (Intercept) Residual
#> StdDev:  0.5154042 0.3925777
#>
#> Correlation Structure: Exponential spatial correlation
#> Formula: ~day / tree
#> Parameter estimate(s):
#>   range
#> 3.794624
#> Fixed effects: water ~ 1 + ti + ti2
#>             Value Std.Error DF t-value p-value
#> (Intercept) -11.223310 1.0988725 227 -10.21348     0
#> ti          12.712094 0.9794235 227 12.97916     0
#> ti2         -2.913682 0.2148551 227 -13.56115     0
#> Correlation:
#>   (Intr) ti

```

```
#> ti -0.985
#> ti2 0.976 -0.997
#>
#> Standardized Within-Group Residuals:
#>      Min        Q1        Med        Q3        Max
#> -3.04861039 -0.55703950  0.00278101  0.62558762  3.80676991
#>
#> Number of Observations: 239
#> Number of Groups: 10
```


Chapter 9

Nonlinear and Generalized Linear Mixed Models

- NLMMs extend the nonlinear model to include both fixed effects and random effects
- GLMMs extend the generalized linear model to include both fixed effects and random effects.

A nonlinear mixed model has the form of

$$Y_{ij} = f(\mathbf{x}_{ij}, \beta_i) + \epsilon_{ij}$$

for the j -th response from cluster (or subject) i ($i = 1, \dots, n$), where

- $j = 1, \dots, n_i$
- β_i are the fixed effects
- ϵ_i are the random effects for cluster i
- \mathbf{x}_{ij} are the regressors or design variables
- $f(\cdot)$ is nonlinear mean response function

A GLMM can be written as:

we assume

$$y_i | \alpha_i \sim \text{indep } f(y_i | \alpha)$$

and $f(y_i | \alpha)$ is an exponential family distribution,

$$f(y_i|\alpha) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} - c(y_i, \phi)\right]$$

The conditional mean of y_i is related to θ_i

$$\mu_i = \frac{\partial b(\theta_i)}{\partial \theta_i}$$

The transformation of this mean will give us the desired linear model to model both the fixed and random effects.

$$E(y_i|\alpha) = \mu_i g(\mu_i) = \mathbf{x}'_i + \mathbf{z}'_i$$

where $g()$ is a known link function and μ_i is the conditional mean. We can see similarity to GLM

We also have to specify the random effects distribution

$$\alpha \sim f(\alpha)$$

which is similar to the specification for mixed models.

Moreover, law of large number applies to fixed effects so that you know it is a normal distribution. But here, you can specify α subjectively.

Hence, we can show NLMM is a special case of the GLMM

$$\mathbf{Y}_i = \mathbf{f}(\mathbf{x}_i, \alpha_i) + \epsilon_i \quad \mathbf{Y}_i = \mathbf{g}^{-1}(\mathbf{x}'_i \beta + \mathbf{z}'_i \alpha_i) + \epsilon_i$$

where the inverse link function corresponds to a nonlinear transformation of the fixed and random effects.

Note:

- we can't derive the analytical formulation of the marginal distribution because nonlinear combination of normal variables is not normally distributed, even in the case of additive error (ϵ_i) and random effects (α_i) are both normal.

Consequences of having random effects

The marginal mean of y_i is

$$E(y_i) = E_\alpha(E(y_i|\alpha)) = E_\alpha(\mu_i) = E(g^{-1}(\mathbf{x}'_i + \mathbf{z}'_i))$$

Because $g^{-1}()$ is nonlinear, this is the most simplified version we can go for.

In special cases such as log link ($g(\mu) = \log \mu$ or $g^{-1}() = \exp()$) then

$$E(y_i) = E(\exp(\mathbf{x}'_i + \mathbf{z}'_i)) = \exp(\mathbf{x}'_i) E(\exp(\mathbf{z}'_i \alpha))$$

which is the moment generating function of α evaluated at \mathbf{z}_i

Marginal variance of y_i

$$\begin{aligned} var(y_i) &= var_\alpha(E(y_i|\alpha)) + E_\alpha(var(y_i|\alpha)) \\ &= var(\mu_i) + E(a(\phi)V(\mu_i)) \\ &= var(g^{-1}(\mathbf{x}'_i + \mathbf{z}'_i)) + E(a(\phi)V(g^{-1}(\mathbf{x}'_i + \mathbf{z}'_i))) \end{aligned}$$

Without specific assumption about $g()$ and/or the conditional distribution of \mathbf{y} , this is the most simplified version.

Marginal covariance of \mathbf{y}

In a linear mixed model, random effects introduce a dependence among observations which share any random effect in common

$$\begin{aligned} cov(y_i, y_j) &= cov_\alpha(E(y_i|), E(y_j|)) + E_\alpha(cov(y_i, y_j|)) \\ &= cov(\mu_i, \mu_j) + E(0) \\ &= cov(g^{-1}(\mathbf{x}'_i \beta + \mathbf{z}'_i), g^{-1}(\mathbf{x}'_j \beta + \mathbf{z}'_j)) \end{aligned}$$

- Important: conditioning to induce the covariability

Example:

Repeated measurements on the subjects. Let y_{ij} be the j-th count taken on the i-th subject.

then, the model is $y_{ij}| \sim \text{indep Pois}(\mu_{ij})$. Here

$$\log(\mu_{ij}) = \mathbf{x}'_{ij}\beta + \alpha_i$$

where $\alpha_i \sim iidN(0, \sigma_\alpha^2)$

which is a log-link with a random patient effect.

9.1 Estimation

In linear mixed models, the marginal likelihood for \mathbf{y} is the integration of the random effects from the hierarchical formulation

$$f(\mathbf{y}) = \int f(\mathbf{y}|\alpha) f(\alpha) d\alpha$$

For linear mixed models, we assumed that the 2 component distributions were Gaussian with linear relationships, which implied the marginal distribution was also linear and Gaussian and allows us to solve this integral analytically.

On the other hand, GLMMs, the distribution for $f(\mathbf{y}|\alpha)$ is not Gaussian in general, and for NLMMS, the functional form between the mean response and the random (and fixed) effects is nonlinear. In both cases, we can't perform the integral analytically, which means we have to solve it

- numerically and/or
- linearize the inverse link function.

9.1.1 Estimation by Numerical Integration

The marginal likelihood is

$$L(\beta; \mathbf{y}) = \int f(\mathbf{y}|\alpha) f(\alpha) d\alpha$$

Estimation fo the fixed effects requires $\frac{\partial l}{\partial \beta}$, where l is the log-likelihood

One way to obtain the marginal inference is to numerically integrate out the random effects through

- numerical quadrature
- Laplace approximation
- Monte Carlo methods

When the dimension of α is relatively low, this is easy. But when the dimension of α is high, additional approximation is required.

9.1.2 Estimation by Linearization

Idea: Linearized version of the response (known as working response, or pseudo-response) called \tilde{y}_i and then the conditional mean is

$$E(\tilde{y}_i|\alpha) = \mathbf{x}'_i\beta + \mathbf{z}'_i\alpha$$

and also estimate $\text{var}(\tilde{y}_i|\alpha)$. then, apply Linear Mixed Models estimation as usual.

The difference is only in how the linearization is done (i.e., how to expand $f(\mathbf{x}, \cdot)$ or the inverse link function

9.1.2.1 Penalized quasi-likelihood

(PQL)

This is the more popular method

$$\tilde{y}_i^{(k)} = \hat{\eta}_i^{(k-1)} + (y_i - \hat{\mu}_i^{(k-1)}) \frac{d\eta}{d\mu} |_{\hat{\eta}_i^{(k-1)}}$$

where

- $\eta_i = g(\mu_i)$ is the linear predictor
- k = iteration of the optimization algorithm

The algorithm updates \tilde{y}_i after each linear mixed model fit using $E(\tilde{y}_i|\alpha)$ and $\text{var}(\tilde{y}_i|\alpha)$

Comments:

- Easy to implement
- Inference is only asymptotically correct due to the linearizaton
- Biased estimates are likely for binomial response with small groups and worst for Bernoulli response. Similarly for Poisson models with small counts. (Faraway, 2016)
- Hypothesis testing and confidence intervals also have problems.

9.1.2.2 Generalized Estimating Equations

(GEE)

Let a marginal generalized linear model for the mean of y as a function of the predictors, which means we linearize the mean response function and assume a dependent error structure

Example

Binary data:

$$\text{logit}(E(\mathbf{y})) = \mathbf{X}\beta$$

If we assume a “working covariance matrix”, \mathbf{V} the the elements of \mathbf{y} , then the maximum likelihood equations for estimating β is

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} = \mathbf{X}'\mathbf{V}^{-1}E(\mathbf{y})$$

If \mathbf{V} is correct, then unbiased estimating equations

We typically define $\mathbf{V} = \mathbf{I}$. Solutions to unbiased estimating equation give consistent estimators.

In practice, we assume a covariance structure, and then do a logistic regression, and calculate its large sample variance

Let $y_{ij}, j = 1, \dots, n_i, i = 1, \dots, K$ be the j -th measurement on the i -th subject.

$$\mathbf{y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{pmatrix}$$

with mean

$$_i = \begin{pmatrix} \mu_{i1} \\ \vdots \\ \mu_{in_i} \end{pmatrix}$$

and

$$\mathbf{x}_{ij} = \begin{pmatrix} X_{ij1} \\ \vdots \\ X_{ijp} \end{pmatrix}$$

Let $\mathbf{V}_i = \text{cov}(\mathbf{y}_i)$, then based on(LIANG and ZEGER, 1986) GEE estimates for β can be obtained from solving the equation:

$$S(\beta) = \sum_{i=1}^K \frac{\partial_i'}{\partial \beta} \mathbf{V}^{-1} (\mathbf{y}_i - \mu_i) = 0$$

Let $\mathbf{R}_i(\mathbf{c})$ be an $n_i \times n_i$ “working” correlation matrix specified up to some parameters \mathbf{c} . Then, $\mathbf{V}_i = a(\phi) \mathbf{B}_i^{1/2} \mathbf{R}(\mathbf{c}) \mathbf{B}_i^{1/2}$, where \mathbf{B}_i is an $n_i \times n_i$ diagonal matrix with $V(\mu_{ij})$ on the j-th diagonal

If $\mathbf{R}(\mathbf{c})$ is the true correlation matrix of \mathbf{y}_i , then \mathbf{V}_i is the true covariance matrix

The working correlation matrix must be estimated iteratively by a fitting algorithm:

1. Compute the initial estimate of β (using GLM under the independence assumption)
2. Compute the working correlation matrix \mathbf{R} based upon studentized residuals
3. Compute the estimate covariance $\hat{\mathbf{V}}_i$
4. Update β according to

$$\beta_{r+1} = \beta_r + \left(\sum_{i=1}^K \frac{\partial_i'}{\partial \beta} \hat{\mathbf{V}}_i^{-1} \frac{\partial_i}{\partial \beta} \right)$$

5. Iterate until the algorithm converges

Note: Inference based on likelihoods is not appropriate because this is not a likelihood estimator

9.1.3 Estimation by Bayesian Hierarchical Models

Bayesian Estimation

$$f(\beta | \mathbf{y}) \propto f(\mathbf{y} | \beta) f(\beta)$$

Numerical techniques (e.g., MCMC) can be used to find posterior distribution. This method is best in terms of not having to make simplifying approximation and fully accounting for uncertainty in estimation and prediction, but it could be complex, time-consuming, and computationally intensive.

Implementation Issues:

- No valid joint distribution can be constructed from the given conditional model and random parameters

- The mean/ variance relationship and the random effects lead to constraints on the marginal covariance model
- Difficult to fit computationally

2 types of estimation approaches:

1. Approximate the objective function (marginal likelihood) through integral approximation
 1. Laplace methods
 2. Quadrature methods
 3. Monte Carlo integration
2. Approximate the model (based on Taylor series linearizations)

Packages in R

- GLMM: MASS::glmmPQL lme4::glmer glmmTMB
- NLMM: nlme::nlme; lme4::nlmer brms::brm
- Bayesian: MCMCglmm ; brms::brm

Example: Non-Gaussian Repeated measurements

- When the data are Gaussian, then Linear Mixed Models
- When the data are non-Gaussian, then Nonlinear and Generalized Linear Mixed Models

9.2 Application

9.2.1 Binomial (CBPP Data)

```
data(cbpp, package = "lme4")
head(cbpp)
#>   herd incidence size period
#> 1     1         2    14      1
#> 2     1         3    12      2
#> 3     1         4     9      3
#> 4     1         0     5      4
#> 5     2         3    22      1
#> 6     2         1    18      2
```

PQL

Pro:

- Linearizes the response to have a pseudo-response as the mean response (like LMM)
- computationally efficient

Cons:

- biased for binary, Poisson data with small counts
- random effects have to be interpreted on the link scale
- can't interpret AIC/BIC value

```
library(MASS)
pql_cbpp <-
  glmmPQL(
    cbind(incidence, size - incidence) ~ period,
    random = ~ 1 | herd,
    data = cbpp,
    family = binomial(link = "logit"),
    verbose = F
  )
summary(pql_cbpp)
#> Linear mixed-effects model fit by maximum likelihood
#>   Data: cbpp
#>   AIC BIC logLik
#>   NA  NA      NA
#>
#> Random effects:
#>   Formula: ~1 / herd
#>             (Intercept) Residual
#> StdDev:  0.5563535 1.184527
#>
#> Variance function:
#>   Structure: fixed weights
#>   Formula: ~invwt
#> Fixed effects: cbind(incidence, size - incidence) ~ period
#>                  Value Std.Error DF t-value p-value
#> (Intercept) -1.327364 0.2390194 38 -5.553372 0.0000
#> period2     -1.016126 0.3684079 38 -2.758156 0.0089
#> period3     -1.149984 0.3937029 38 -2.920944 0.0058
#> period4     -1.605217 0.5178388 38 -3.099839 0.0036
```

```
#> Correlation:
#>           (Intr) period2 period3
#> period2 -0.399
#> period3 -0.373  0.260
#> period4 -0.282  0.196  0.182
#>
#> Standardized Within-Group Residuals:
#>      Min       Q1       Med       Q3       Max
#> -2.0591168 -0.6493095 -0.2747620  0.5170492  2.6187632
#>
#> Number of Observations: 56
#> Number of Groups: 15

exp(0.556)
#> [1] 1.743684
```

is how the herd specific outcome odds varies.

We can interpret the fixed effect coefficients just like in GLM. Because we use logit link function here, we can say that the log odds of the probability of having a case in period 2 is -1.016 less than period 1 (baseline).

```
summary(pql_cbpp)$tTable
#>          Value Std.Error DF   t-value     p-value
#> (Intercept) -1.327364 0.2390194 38 -5.553372 2.333216e-06
#> period2     -1.016126 0.3684079 38 -2.758156 8.888179e-03
#> period3     -1.149984 0.3937029 38 -2.920944 5.843007e-03
#> period4     -1.605217 0.5178388 38 -3.099839 3.637000e-03
```

Numerical Integration

Pro:

- more accurate

Con:

- computationally expensive
- won't work for complex models.

```
library(lme4)
numint_cbpp <-
glmer(
```

```

cbind(incidence, size - incidence) ~ period + (1 | herd),
  data = cbpp,
  family = binomial(link = "logit")
)
summary(numint_cbpp)
#> Generalized linear mixed model fit by maximum likelihood (Laplace
#> Approximation) [glmerMod]
#> Family: binomial ( logit )
#> Formula: cbind(incidence, size - incidence) ~ period + (1 / herd)
#>   Data: cbpp
#>
#>     AIC      BIC  logLik deviance df.resid
#>   194.1    204.2   -92.0    184.1      51
#>
#> Scaled residuals:
#>   Min     1Q Median     3Q    Max
#> -2.3816 -0.7889 -0.2026  0.5142  2.8791
#>
#> Random effects:
#> Groups Name        Variance Std.Dev.
#> herd   (Intercept) 0.4123   0.6421
#> Number of obs: 56, groups: herd, 15
#>
#> Fixed effects:
#>             Estimate Std. Error z value Pr(>|z|)
#> (Intercept) -1.3983    0.2312 -6.048 1.47e-09 ***
#> period2     -0.9919    0.3032 -3.272 0.001068 **
#> period3     -1.1282    0.3228 -3.495 0.000474 ***
#> period4     -1.5797    0.4220 -3.743 0.000182 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Correlation of Fixed Effects:
#>           (Intr) period2 period3
#> period2 -0.363
#> period3 -0.340  0.280
#> period4 -0.260  0.213  0.198

```

For small data set, the difference between two approaches are minimal

```

library(rbenchmark)
benchmark(
  "MASS" = {
    pql_cbpp <-
      glmmPQL(

```

```

            cbind(incidence, size - incidence) ~ period,
            random = ~ 1 | herd,
            data = cbpp,
            family = binomial(link = "logit"),
            verbose = F
        )
    },
    "lme4" = {
        glmer(
            cbind(incidence, size - incidence) ~ period + (1 | herd),
            data = cbpp,
            family = binomial(link = "logit")
        )
    },
    replications = 50,
    columns = c("test", "replications", "elapsed", "relative"),
    order = "relative"
)
#>   test replications elapsed relative
#> 1 MASS          50     2.39   1.000
#> 2 lme4          50     5.17   2.163

```

In numerical integration, we can set `nAGQ > 1` to switch the method of likelihood evaluation, which might increase accuracy

```

library(lme4)
numint_cbpp_GH <-
  glmer(
    cbind(incidence, size - incidence) ~ period + (1 | herd),
    data = cbpp,
    family = binomial(link = "logit"),
    nAGQ = 20
)
summary(numint_cbpp_GH)$coefficients[, 1] - summary(numint_cbpp)$coefficients[, 1]
#>   (Intercept)      period2      period3      period4
#> -0.0008808634  0.0005160912  0.0004066218  0.0002644629

```

Bayesian approach to GLMMs

- assume the fixed effects parameters have distribution
- can handle models with intractable result under traditional methods
- computationally expensive

```

library(MCMCglmm)
Bayes_cbpp <-
  MCMCglmm(
    cbind(incidence, size - incidence) ~ period,
    random = ~ herd,
    data = cbpp,
    family = "multinomial2",
    verbose = FALSE
  )
summary(Bayes_cbpp)
#>
#> Iterations = 3001:12991
#> Thinning interval = 10
#> Sample size = 1000
#>
#> DIC: 537.8536
#>
#> G-structure: ~herd
#>
#> post.mean l-95% CI u-95% CI eff.samp
#> herd 0.02245 7.94e-17 0.129 47.74
#>
#> R-structure: ~units
#>
#> post.mean l-95% CI u-95% CI eff.samp
#> units 1.13 0.2532 2.151 315.4
#>
#> Location effects: cbind(incidence, size - incidence) ~ period
#>
#> post.mean l-95% CI u-95% CI eff.samp pMCMC
#> (Intercept) -1.5462 -2.2018 -0.8965 1000.0 <0.001 ***
#> period2 -1.2338 -2.3042 -0.1553 811.2 0.022 *
#> period3 -1.3671 -2.4137 -0.2875 742.6 0.010 **
#> period4 -1.9972 -3.2697 -0.6881 566.4 0.002 **
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

- MCMCglmm fits a residual variance component (useful with dispersion issues)

```

apply(Bayes_cbpp$VCV, 2, sd) #explains less variability
#>      herd      units
#> 0.08910401 0.51628335

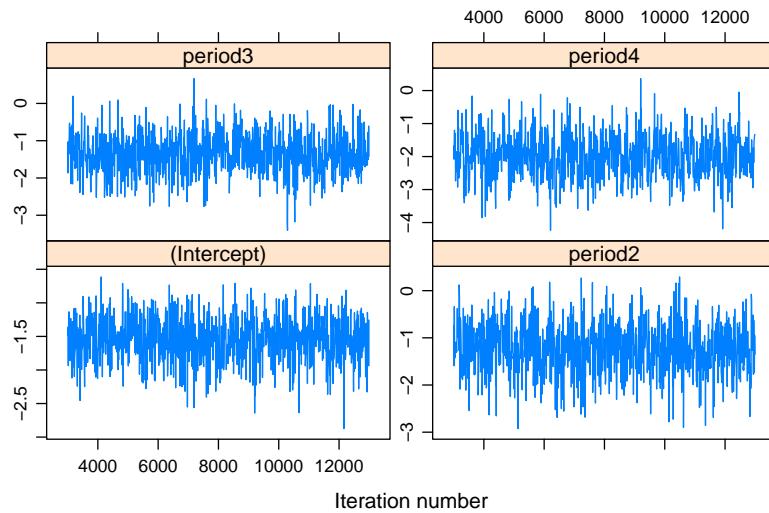
```

```
summary(Bayes_cbpp)$solutions
#>           post.mean   l-95% CI   u-95% CI   eff.samp pMCMC
#> (Intercept) -1.546216 -2.201782 -0.8965055 1000.0000 0.001
#> period2     -1.233829 -2.304225 -0.1552787  811.2456 0.022
#> period3     -1.367094 -2.413671 -0.2875085  742.6031 0.010
#> period4     -1.997218 -3.269653 -0.6881274  566.3967 0.002
```

interpret Bayesian “credible intervals” similarly to confidence intervals

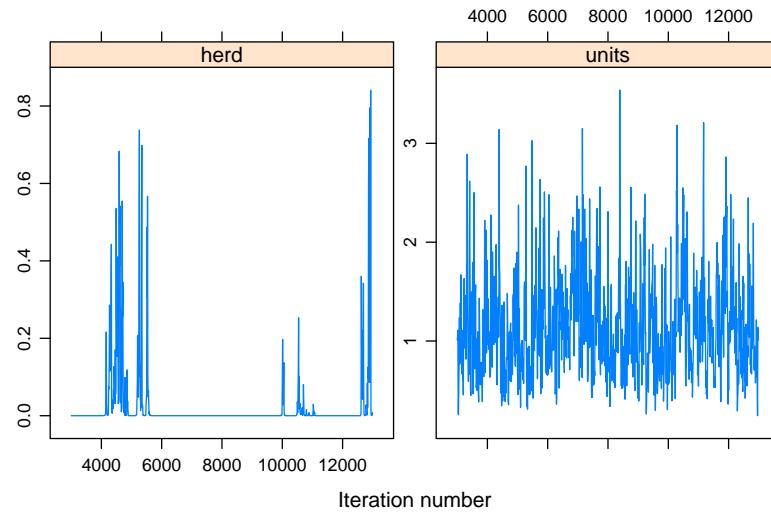
Make sure you make post-hoc diagnoses

```
library(lattice)
xyplot(as.mcmc(Bayes_cbpp$Sol), layout = c(2, 2))
```



There is no trend, well-mixed

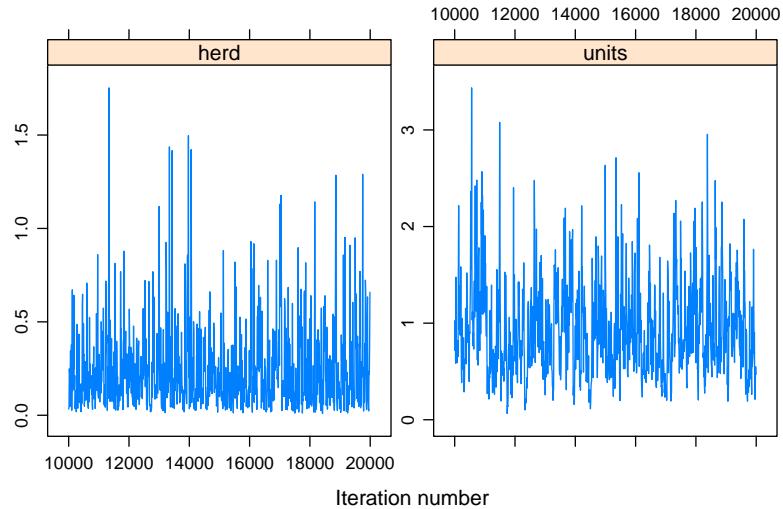
```
xyplot(as.mcmc(Bayes_cbpp$VCV), layout=c(2,1))
```



For the herd variable, a lot of them are 0, which suggests problem. To fix the instability in the herd effect sampling, we can either

- modify the prior distribution on the herd variation
- increases the number of iteration

```
library(MCMCglmm)
Bayes_cbpp2 <-
  MCMCglmm(
    cbind(incidence, size - incidence) ~ period,
    random = ~ herd,
    data = cbpp,
    family = "multinomial2",
    nitt = 20000,
    burnin = 10000,
    prior = list(G = list(list(
      V = 1, nu = .1
    ))),
    verbose = FALSE
  )
xyplot(as.mcmc(Bayes_cbpp2$VCV), layout = c(2, 1))
```



To change the shape of priors, in `MCMCglmm` use:

- `V` controls for the location of the distribution (default = 1)
- `nu` controls for the concentration around `V` (default = 0)

9.2.2 Count (Owl Data)

```
library(glmmTMB)
library(dplyr)
data(Owls, package = "glmmTMB")
Owls <- Owls %>% rename(Ncalls = SiblingNegotiation)
```

In a typical Poisson model, λ (Poisson mean), is model as $\log(\lambda) = \mathbf{x}'$. But if the response is the rate (e.g., counts per BroodSize), we could model it as $\log(\lambda/b) = \mathbf{x}'$, equivalently $\log(\lambda) = \log(b) + \mathbf{x}'$ where b is BroodSize. Hence, we “offset” the mean by the log of this variable.

```
owls_glmer <-
  glmer(
    Ncalls ~ offset(log(BroodSize)) + FoodTreatment * SexParent +
      (1 | Nest),
    family = poisson,
    data = Owls
```

```

)
summary(owls_glmer)
#> Generalized linear mixed model fit by maximum likelihood (Laplace
#> Approximation) [glmerMod]
#> Family: poisson  ( log )
#> Formula: Ncalls ~ offset(log(BroodSize)) + FoodTreatment * SexParent +
#>           (1 | Nest)
#> Data: Owls
#>
#>      AIC      BIC  logLik deviance df.resid
#>  5212.8  5234.8 -2601.4   5202.8     594
#>
#> Scaled residuals:
#>    Min     1Q Median     3Q    Max
#> -3.5529 -1.7971 -0.6842  1.2689 11.4312
#>
#> Random effects:
#> Groups Name        Variance Std.Dev.
#> Nest   (Intercept) 0.2063   0.4542
#> Number of obs: 599, groups: Nest, 27
#>
#> Fixed effects:
#>                               Estimate Std. Error z value Pr(>|z|)
#> (Intercept)                  0.65585  0.09567  6.855 7.12e-12 ***
#> FoodTreatmentSatiated       -0.65612  0.05606 -11.705 < 2e-16 ***
#> SexParentMale                 -0.03705  0.04501 -0.823  0.4104
#> FoodTreatmentSatiated:SexParentMale 0.13135  0.07036  1.867  0.0619 .
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Correlation of Fixed Effects:
#>          (Intr) FdTrtsS SxPrnM
#> FdTTrmntStt -0.225
#> SexParentMl -0.292  0.491
#> FdTTrtmS:SPM  0.170 -0.768 -0.605

```

- nest explains a relatively large proportion of the variability (its standard deviation is larger than some coefficients)
- the model fit isn't great (deviance of 5202 on 594 df)

```

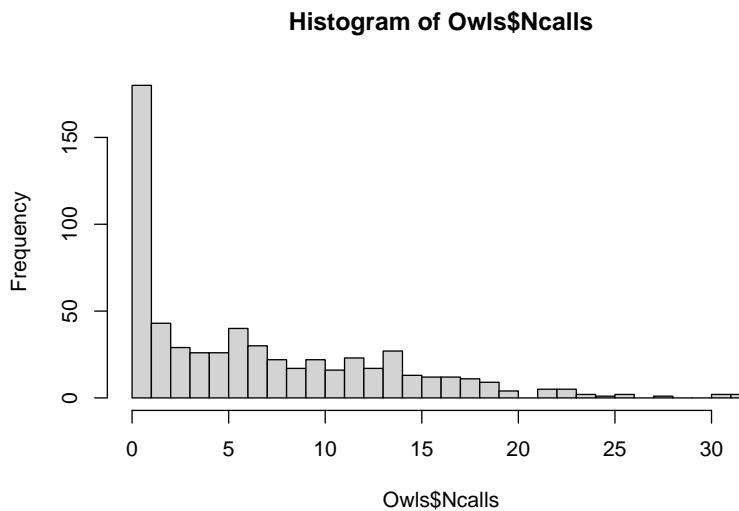
# Negative binomial model
owls_glmerNB <-
  glmer.nb(Ncalls ~ offset(log(BroodSize)) + FoodTreatment * SexParent
           + (1 | Nest), data = Owls)

```

```
c(Deviance = round(summary(owls_glmerNB)$AICtab["deviance"], 3),
  df = summary(owls_glmerNB)$AICtab["df.resid"])
#> Deviance.deviance      df.df.resid
#>           3483.616          593.000
```

There is an improvement using negative binomial considering overdispersion

```
hist(Owls$Ncalls, breaks=30)
```



To account for too many 0s in these data, we can use zero-inflated Poisson (ZIP) model.

- `glmmTMB` can handle ZIP GLMMs since it adds automatic differentiation to existing estimation strategies.

```
library(glmmTMB)
owls_glmm <-
  glmmTMB(
    Ncalls ~ FoodTreatment * SexParent + offset(log(BroodSize)) +
      (1 | Nest),
    ziformula = ~ 0,
    family = nbinom2(link = "log"),
    data = Owls
  )
```

```

owls_glmm_zi <-
  glmmTMB(
    Ncalls ~ FoodTreatment * SexParent + offset(log(BroodSize)) +
      (1 | Nest),
    ziformula = ~ 1,
    family = nbinom2(link
      = "log"),
    data = Owls
  )
# Scale Arrival time to use as a covariate for zero-inflation parameter
Owls$ArrivalTime <- scale(Owls$ArrivalTime)
owls_glmm_zi_cov <- glmmTMB(
  Ncalls ~ FoodTreatment * SexParent +
    offset(log(BroodSize)) +
    (1 | Nest),
  ziformula = ~ ArrivalTime,
  family = nbinom2(link
    = "log"),
  data = Owls
)
as.matrix(anova(owls_glmm, owls_glmm_zi))
#>           Df     AIC     BIC   logLik deviance    Chisq Chi Df
#> owls_glmm     6 3495.610 3521.981 -1741.805 3483.610      NA    NA
#> owls_glmm_zi  7 3431.646 3462.413 -1708.823 3417.646 65.96373      1
#>             Pr(>Chisq)
#> owls_glmm          NA
#> owls_glmm_zi 4.592983e-16
as.matrix(anova(owls_glmm_zi,owls_glmm_zi_cov))
#>           Df     AIC     BIC   logLik deviance    Chisq Chi Df
#> owls_glmm_zi     7 3431.646 3462.413 -1708.823 3417.646      NA    NA
#> owls_glmm_zi_cov  8 3422.532 3457.694 -1703.266 3406.532 11.11411      1
#>             Pr(>Chisq)
#> owls_glmm_zi          NA
#> owls_glmm_zi_cov 0.0008567362
summary(owls_glmm_zi_cov)
#> Family: nbinom2 ( log )
#> Formula:
#> Ncalls ~ FoodTreatment * SexParent + offset(log(BroodSize)) + (1 / Nest)
#> Zero inflation: ~ArrivalTime
#> Data: Owls
#>
#>           AIC     BIC   logLik deviance df.resid
#> 3422.5 3457.7 -1703.3  3406.5      591
#>
#> Random effects:
```

```

#>
#> Conditional model:
#> Groups Name           Variance Std.Dev.
#> Nest    (Intercept) 0.07487  0.2736
#> Number of obs: 599, groups: Nest, 27
#>
#> Dispersion parameter for nbinom2 family (): 2.22
#>
#> Conditional model:
#>                                         Estimate Std. Error z value Pr(>/z|)
#> (Intercept)                      0.84778   0.09961  8.511 < 2e-16 ***
#> FoodTreatmentSatiated          -0.39529   0.13742 -2.877  0.00402 **
#> SexParentMale                  -0.07025   0.10435 -0.673  0.50079
#> FoodTreatmentSatiated:SexParentMale 0.12388   0.16449  0.753  0.45138
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Zero-inflation model:
#>                                         Estimate Std. Error z value Pr(>/z|)
#> (Intercept)      -1.3018     0.1261 -10.32 < 2e-16 ***
#> ArrivalTime      0.3545     0.1074   3.30 0.000966 ***
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

We can see ZIP GLMM with an arrival time covariate on the zero is best.

- arrival time has a positive effect on observing a nonzero number of calls
- interactions are non significant, the food treatment is significant (fewer calls after eating)
- nest variability is large in magnitude (without this, the parameter estimates change)

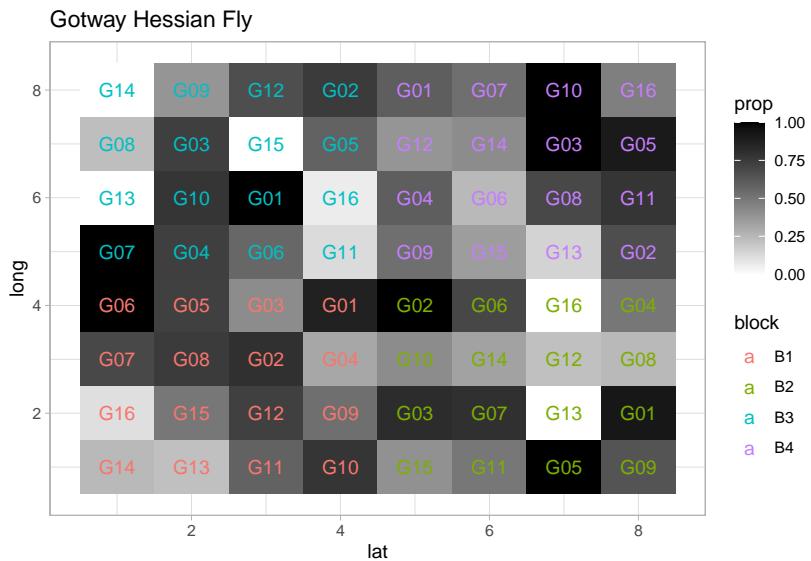
9.2.3 Binomial

```

library(agridat)
library(ggplot2)
library(lme4)
library(spAMM)
data(gotway.hessianfly)
dat <- gotway.hessianfly
dat$prop <- dat$y / dat$n

```

```
ggplot(dat, aes(x = lat, y = long, fill = prop)) +
  geom_tile() +
  scale_fill_gradient(low = 'white', high = 'black') +
  geom_text(aes(label = gen, color = block)) +
  ggtitle('Gotway Hessian Fly')
```



- Fixed effects (β) = genotype
- Random effects (α) = block

```
flymodel <-
  glmer(
    cbind(y, n - y) ~ gen + (1 | block),
    data = dat,
    family = binomial,
    nAGQ = 5
  )
summary(flymodel)
#> Generalized linear mixed model fit by maximum likelihood (Adaptive
#>   Gauss-Hermite Quadrature, nAGQ = 5) [glmerMod]
#> Family: binomial ( logit )
#> Formula: cbind(y, n - y) ~ gen + (1 / block)
#>   Data: dat
#>
#>      AIC        BIC    logLik deviance df.resid
#>    162.2     198.9    -64.1    128.2       47
```

```

#>
#> Scaled residuals:
#>      Min       1Q   Median      3Q      Max
#> -2.38644 -1.01188  0.09631  1.03468  2.75479
#>
#> Random effects:
#> Groups Name           Variance Std.Dev.
#> block  (Intercept) 0.001022 0.03196
#> Number of obs: 64, groups: block, 4
#>
#> Fixed effects:
#>             Estimate Std. Error z value Pr(>|z|)
#> (Intercept)  1.5035    0.3914   3.841 0.000122 ***
#> genG02      -0.1939    0.5302  -0.366 0.714644
#> genG03      -0.5408    0.5103  -1.060 0.289260
#> genG04      -1.4342    0.4714  -3.043 0.002346 **
#> genG05      -0.2037    0.5429  -0.375 0.707486
#> genG06      -0.9783    0.5046  -1.939 0.052533 .
#> genG07      -0.6041    0.5111  -1.182 0.237235
#> genG08      -1.6774    0.4907  -3.418 0.000630 ***
#> genG09      -1.3984    0.4725  -2.960 0.003078 **
#> genG10      -0.6817    0.5333  -1.278 0.201181
#> genG11      -1.4630    0.4843  -3.021 0.002522 **
#> genG12      -1.4591    0.4918  -2.967 0.003010 **
#> genG13      -3.5528    0.6600  -5.383 7.31e-08 ***
#> genG14      -2.5073    0.5264  -4.763 1.90e-06 ***
#> genG15      -2.0872    0.4851  -4.302 1.69e-05 ***
#> genG16      -2.9697    0.5383  -5.517 3.46e-08 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

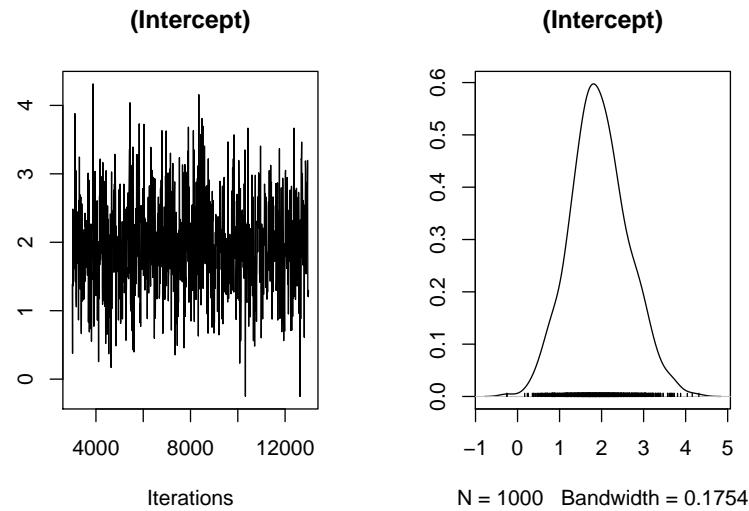
```

Equivalently, we can use `MCMCglmm`, for a Bayesian approach

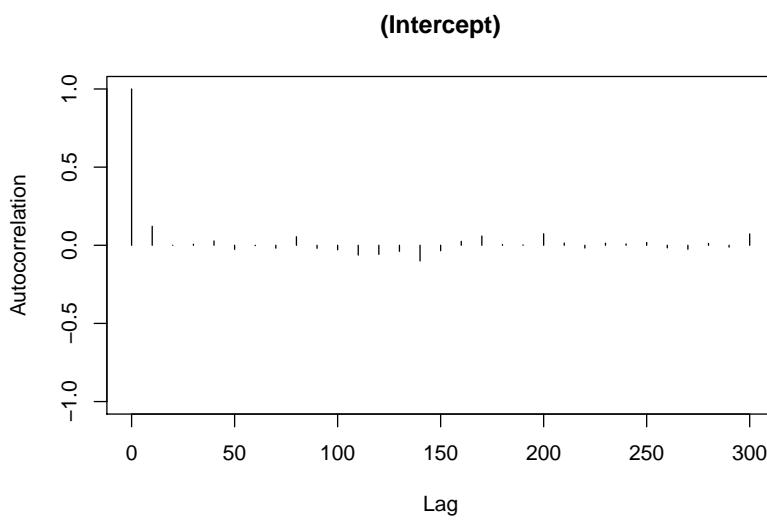
```

library(coda)
Bayes_flymodel <- MCMCglmm(
  cbind(y, n - y) ~ gen ,
  random = ~ block,
  data = dat,
  family = "multinomial2",
  verbose = FALSE
)
plot(Bayes_flymodel$Sol[, 1], main = dimnames(Bayes_flymodel$Sol)[[2]][1])

```



```
autocorr.plot(Bayes_flymodel$Sol[,1], main=dimnames(Bayes_flymodel$Sol)[[2]][1])
```

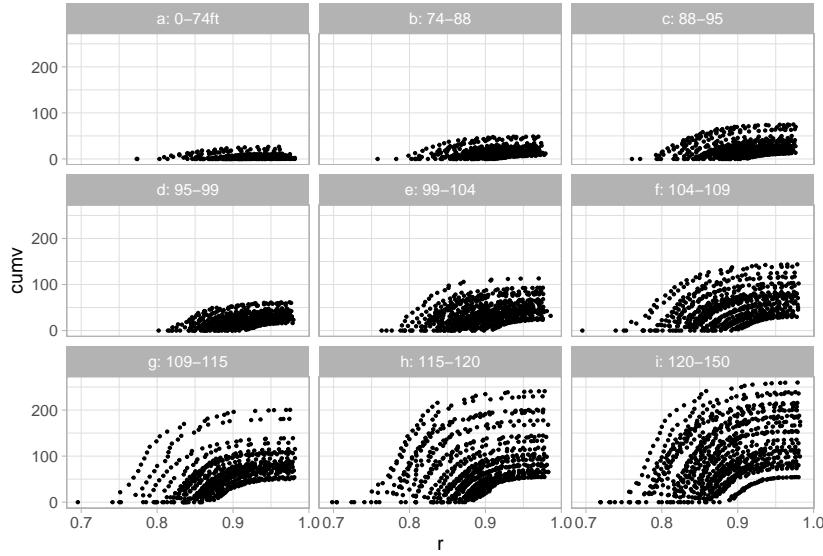


9.2.4 Example from (Schabenberger and Pierce, 2001) section 8.4.1

```
dat2 <- read.table("images/YellowPoplarData_r.txt")
names(dat2) <- c('tn', 'k', 'dbh', 'totht', 'dob', 'ht', 'maxd', 'cumv')
dat2$t <- dat2$dob / dat2$dbh
dat2$r <- 1 - dat2$dob / dat2$totht
```

The cumulative volume relates to the complementary diameter (subplots were created based on total tree height)

```
library(ggplot2)
library(dplyr)
dat2 <- dat2 %>% group_by(tn) %>% mutate(
  z = case_when(
    totht < 74 & totht >= 0 ~ 'a: 0-74ft',
    totht < 88 & totht >= 74 ~ 'b: 74-88',
    totht < 95 & totht >= 88 ~ 'c: 88-95',
    totht < 99 & totht >= 95 ~ 'd: 95-99',
    totht < 104 & totht >= 99 ~ 'e: 99-104',
    totht < 109 & totht >= 104 ~ 'f: 104-109',
    totht < 115 & totht >= 109 ~ 'g: 109-115',
    totht < 120 & totht >= 115 ~ 'h: 115-120',
    totht < 140 & totht >= 120 ~ 'i: 120-150',
  )
)
ggplot(dat2, aes(x = r, y = cumv)) + geom_point(size = 0.5) + facet_wrap(vars(z))
```



The proposed non-linear model:

$$V_{id_j} = (\beta_0 + (\beta_1 + b_{1i}) \frac{D_i^2 H_i}{1000}) (\exp[-(\beta_2 + b_{2i}) t_{ij} \exp(\beta_3 t_{ij})]) + e_{ij}$$

where

- b_{1i}, b_{2i} are random effects
- e_{ij} are random errors

```
library(nlme)
tmp <-
  nlme(
    cumv ~ (b0 + (b1 + u1) * (dbh * dbh * totht / 1000)) * (exp(-(b2 + u2) *
      (t / 1000) * exp(b3 * t)))
    data = dat2,
    fixed = b0 + b1 + b2 + b3 ~ 1,
    # 1 on the right hand side of the formula indicates a single fixed effects for the corresponding coefficients
    random = list(pdDiag(u1 + u2 ~ 1)),
    #uncorrelated random effects
    groups = ~ tn,
    #group on trees so each tree w/ have u1 and u2
    start = list(fixed = c(
      b0 = 0.25,
      b1 = 2.3,
```

```

        b2 = 2.87,
        b3 = 6.7
    ))
)
summary(tmp)
#> Nonlinear mixed-effects model fit by maximum likelihood
#>   Model: cumu ~ (b0 + (b1 + u1) * (dbh * dbh * toht/1000)) * (exp(-(b2 +
#>   Data: dat2
#>       AIC      BIC      logLik
#>   31103.73 31151.33 -15544.86
#>
#> #> Random effects:
#>   Formula: list(u1 ~ 1, u2 ~ 1)
#>   Level: tn
#>   Structure: Diagonal
#>           u1      u2 Residual
#> StdDev: 0.1508094 0.447829 2.226361
#>
#> #> Fixed effects: b0 + b1 + b2 + b3 ~ 1
#>           Value Std.Error DF t-value p-value
#> b0 0.249386 0.12894686 6297 1.9340 0.0532
#> b1 2.288832 0.01266805 6297 180.6776 0.0000
#> b2 2.500497 0.05606685 6297 44.5985 0.0000
#> b3 6.848871 0.02140677 6297 319.9395 0.0000
#> Correlation:
#>   b0     b1     b2
#> b1 -0.639
#> b2  0.054  0.056
#> b3 -0.011 -0.066 -0.850
#>
#> #> Standardized Within-Group Residuals:
#>           Min            Q1            Med            Q3            Max
#> -6.694575e+00 -3.081861e-01 -8.910696e-05  3.469469e-01  7.855665e+00
#>
#> #> Number of Observations: 6636
#> #> Number of Groups: 336
nlme::intervals(tmp)
#> Approximate 95% confidence intervals
#>
#> #> Fixed effects:
#>           lower      est.      upper
#> b0 -0.003318095 0.2493855 0.5020891
#> b1  2.264006138 2.2888323 2.3136585
#> b2  2.390619987 2.5004970 2.6103740
#> b3  6.806919317 6.8488712 6.8908232

```

```
#>
#> Random Effects:
#>   Level: tn
#>           lower      est.      upper
#> sd(u1) 0.1376080 0.1508094 0.1652772
#> sd(u2) 0.4056135 0.4478290 0.4944382
#>
#> Within-group standard error:
#>   lower      est.      upper
#> 2.187260 2.226361 2.266161
```

- Little different from the book because of different implementation of non-linear mixed models.

```
library(cowplot)
nlmmfn <- function(fixed,rand,dbh,totht,t){
  b0 <- fixed[1]
  b1 <- fixed[2]
  b2 <- fixed[3]
  b3 <- fixed[4]
  u1 <- rand[1]
  u2 <- rand[2]
  #just made so we can predict w/o random effects
  return((b0+(b1+u1)*(dbh*dbh*totht/1000))*(exp(-(b2+u2)*(t/1000)*exp(b3*t))))
}

#Tree 1
pred1 <- data.frame(seq(1,24,length.out=100))
names(pred1) <- 'dob'
pred1$tn <- 1
pred1$dbh <- unique(dat2[dat2$tn==1,]$dbh)
pred1$t <- pred1$dob/pred1$dbh
pred1$totht <- unique(dat2[dat2$tn==1,]$totht)
pred1$r <- 1-pred1$dob/pred1$totht

pred1$test <- predict(tmp,pred1)
pred1$testno <- nlmmfn(fixed=tmp$coefficients$fixed, rand = c(0,0),pred1$dbh,pred1$totht,pred1$t)

p1 <- ggplot(pred1)+geom_line(aes(x=r,y=test,color='with random'))+geom_line(aes(x=r,y=testno,color='without random'))

#Tree 151
pred151 <- data.frame(seq(1,21,length.out=100))
```

```

names(pred151) <- 'dob'
pred151$tn <- 151
pred151$dbh <- unique(dat2[dat2$tn==151,]$dbh)
pred151$t <- pred151$dob/pred151$dbh
pred151$toht <- unique(dat2[dat2$tn==151,]$toht)
pred151$r <- 1-pred151$dob/pred151$toht

pred151$test <- predict(tmp,pred151)
pred151$testno <- nlmmfn(fixed=tmp$coefficients$fixed, rand = c(0,0),pred151$dbh,pred151$toht)

p2 <- ggplot(pred151)+geom_line(aes(x=r,y=test,color='with random'))+geom_line(aes(x=r,toht))

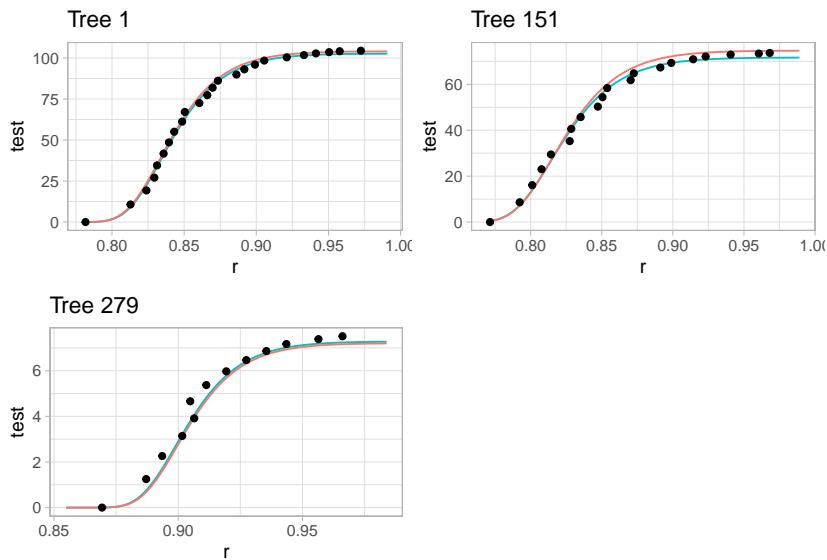
#Tree 279
pred279 <- data.frame(seq(1,9,length.out=100))
names(pred279) <- 'dob'
pred279$tn <- 279
pred279$dbh <- unique(dat2[dat2$tn==279,]$dbh)
pred279$t <- pred279$dob/pred279$dbh
pred279$toht <- unique(dat2[dat2$tn==279,]$toht)
pred279$r <- 1-pred279$dob/pred279$toht

pred279$test <- predict(tmp,pred279)
pred279$testno <- nlmmfn(fixed=tmp$coefficients$fixed, rand = c(0,0),pred279$dbh,pred279$toht)

p3 <- ggplot(pred279)+geom_line(aes(x=r,y=test,color='with random'))+geom_line(aes(x=r,toht))

plot_grid(p1,p2,p3)

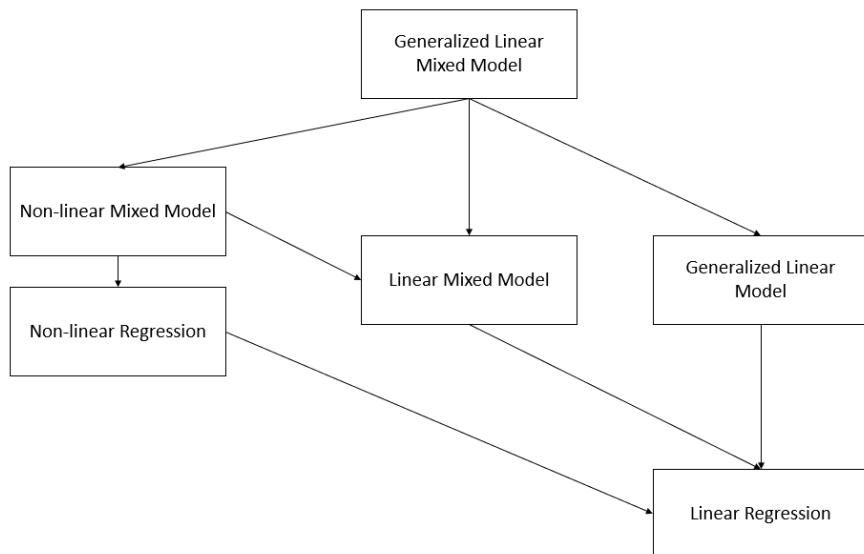
```



red line = predicted observations based on the common fixed effects

teal line = tree-specific predictions with random effects

9.3 Summary



III. RAMIFICATIONS

Chapter 10

Model Specification

Test whether underlying assumptions hold true

- Nested Model (A1/A3)
- Non-Nested Model (A1/A3)
- Heteroskedasticity (A4)

10.1 Nested Model

$$\begin{aligned} y &= \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + \epsilon && \text{unrestricted model} \\ y &= \beta_0 + x_1\beta_1 + \epsilon && \text{restricted model} \end{aligned}$$

Unrestricted model is always longer than the restricted model

The restricted model is “nested” within the unrestricted model

To determine which variables should be included or exclude, we could use the same Wald Test

Adjusted R^2

- R^2 will always increase with more variables included
- Adjusted R^2 tries to correct by penalizing inclusion of unnecessary variables.

$$R^2 = 1 - \frac{SSR/n}{SST/n} R_{adj}^2 = 1 - \frac{SSR/(n-k)}{SST/(n-1)} = 1 - \frac{(n-1)(1-R^2)}{(n-k)}$$

- R_{adj}^2 increases if and only if the t-statistic on the additional variable is greater than 1 in absolute value.

- R^2_{adj} is valid in models where there is no heteroskedasticity
- therefore it **should not** be used in determining which variables should be included in the model (the t or F-tests are more appropriate)

10.1.1 Chow test

Should we run two different regressions for two groups?

10.2 Non-Nested Model

compare models with different non-nested specifications

10.2.1 Davidson-Mackinnon test

10.2.1.1 Independent Variable

should the independent variables be logged? decide between non-nested alternatives

$$\begin{aligned} y &= \beta_0 + x_1\beta_1 + x_2\beta_2 + \epsilon && (\text{level eq}) \\ y &= \beta_0 + \ln(x_1)\beta_1 + x_2\beta_2 + \epsilon && (\text{log eq}) \end{aligned}$$

1. Obtain predict outcome when estimating the model in log equation \check{y} and then estimate the following auxiliary equation,

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \check{y}\gamma + error$$

and evaluate the t-statistic for the null hypothesis $H_0 : \gamma = 0$

2. Obtain predict outcome when estimating the model in the level equation \hat{y} , then estimate the following auxiliary equation,

$$y = \beta_0 + \ln(x_1)\beta_1 + x_2\beta_2 + \check{y}\gamma + error$$

and evaluate the t-statistic for the null hypothesis $H_0 : \gamma = 0$

- If you reject the null in the (1) step but fail to reject the null in the second step, then the log equation is preferred.

- If fail to reject the null in the (1) step but reject the null in the (2) step then, level equation is preferred.
- If reject in both steps, then you have statistical evidence that neither model should be used and should re-evaluate the functional form of your model.
- If fail to reject in both steps, you do not have sufficient evidence to prefer one model over the other. You can compare the R^2_{adj} to choose between the two models.

$$y = \beta_0 + \ln(x)\beta_1 + \epsilon \\ y = \beta_0 + x(\beta_1) + x^2\beta_2 + \epsilon$$

- Compare which better fits the data
- Compare standard R^2 is unfair because the second model is less parsimonious (more parameters to estimate)
- The R^2_{adj} will penalize the second model for being less parsimonious + Only valid when there is no heteroskedasticity (A4 holds)
- Should only compare after a Davidson-Mackinnon test

10.2.1.2 Dependent Variable

$$y = \beta_0 + x_1\beta_1 + \epsilon \quad \text{level eq} \\ \ln(y) = \beta_0 + x_1\beta_1 + \epsilon \quad \text{log eq}$$

- In the level model, regardless of how big y is, x has a constant effect (i.e., one unit change in x_1 results in a β_1 unit change in y)
- In the log model, the larger in y is, the effect of x is stronger (i.e., one unit change in x_1 could increase y from 1 to $1+\beta_1$ or from 100 to $100+100x\beta_1$)
- Cannot compare R^2 or R^2_{adj} because the outcomes are complement different, the scaling is different (SST is different)

We need to “un-transform” the $\ln(y)$ back to the same scale as y and then compare,

1. Estimate the model in the log equation to obtain the predicted outcome $\hat{\ln}(y)$
2. “Un-transform” the predicted outcome

$$\hat{m} = \exp(\hat{\ln}(y))$$

3. Estimate the following model (without an intercept)

$$y = \alpha \hat{m} + error$$

and obtain predicted outcome \hat{y}

4. Then take the square of the correlation between \hat{y} and y as a scaled version of the R^2 from the log model that can now compare with the usual R^2 in the level model.

10.3 Heteroskedasticity

- Using robust standard errors are always valid
- If there is significant evidence of heteroskedasticity implying A4 does not hold
 - Gauss-Markov Theorem no longer holds, OLS is not BLUE.
 - Should consider using a better linear unbiased estimator (Weighted Least Squares or Generalized Least Squares)

10.3.1 Breusch-Pagan test

A4 implies

$$E(\epsilon_i^2 | \mathbf{x}_i) = \sigma^2$$

$$\epsilon_i^2 = \gamma_0 + x_{i1}\gamma_1 + \dots + x_{ik-1}\gamma_{k-1} + error$$

and determining whether or not \mathbf{x}_i has any predictive value

- if \mathbf{x}_i has predictive value, then the variance changes over the levels of \mathbf{x}_i which is evidence of heteroskedasticity
- if \mathbf{x}_i does not have predictive value, the variance is constant for all levels of \mathbf{x}_i

The Breusch-Pagan test for heteroskedasticity would compute the F-test of total significance for the following model

$$\epsilon_i^2 = \gamma_0 + x_{i1}\gamma_1 + \dots + x_{ik-1}\gamma_{k-1} + error$$

A low p-value means we reject the null of homoskedasticity

However, Breusch-Pagan test cannot detect heteroskedasticity in non-linear form

10.3.2 White test

test heteroskedasticity would allow for a non-linear relationship by computing the F-test of total significance for the following model (assume there are three independent random variables)

$$e_i^2 = \gamma_0 + x_i\gamma_1 + x_{i2}\gamma_2 + x_{i3}\gamma_3 + x_{i1}^2\gamma_4 + x_{i2}^2\gamma_5 + x_{i3}^2\gamma_6 + (x_{i1} \times x_{i2})\gamma_7 + (x_{i1} \times x_{i3})\gamma_8 + (x_{i2} \times x_{i3})\gamma_9 + error$$

A low p-value means we reject the null of homoskedasticity

Equivalently, we can compute LM as $LM = nR_{e^2}^2$ where the $R_{e^2}^2$ come from the regression with the squared residual as the outcome

- The LM statistic has a χ_k^2 distribution

Chapter 11

Imputation (Missing Data)

Imputation is a statistical procedure where you replace missing data with *some values*

- Unit imputation = single data point
- Item imputation = single feature value

Imputation is usually seen as the illegitimate child of statistical analysis. Several reasons that contribute to this negative views could be:

- (1) Peoples hardly do imputation correctly (which will introduce bias to your estimates)
- (2) Imputation can only be applied to a small range of problems correctly

If you have missing data on y (dependent variable), you probability would not be able to do any imputation appropriately. However, if you have certain type of missing data (e.g., non-random missing data) in the x 's variable (independent variables), then you can still salvage your collected data points with imputation.

We also need to talk why you would want to do imputation in the first place. If your purpose is inference/ explanation (valid statistical inference not optimal point prediction), then imputation would not offer much help (Rubin, 1996). However, if your purpose is prediction, you would want your standard error to be reduced by including information (non-missing data) on other variables of a data point. Then imputation could be the tool that you're looking for.

For most software packages, it will use listwise deletion or casewise deletion to have complete case analysis (analysis with only observations with all information). Not until recently that statistician can propose some methods that are

a bit better than listwise deletion which are maximum likelihood and multiple imputation.

“Judging the quality of missing data procedures by their ability to recreate the individual missing values (according to hit rate, mean square error, etc) does not lead to choosing procedures that result in valid inference”, (Rubin, 1996)

Missing data can make it more challenging to big datasets.

11.1 Assumptions

11.1.1 Missing Completely at Random (MCAR)

Missing Completely at Random, MCAR, means there is no relationship between the missingness of the data and any values, observed or missing. Those missing data points are a random subset of the data. There is nothing systematic going on that makes some data more likely to be missing than others.

The probability of missing data on a variable is unrelated to the value of it or to the values of any other variables in the data set.

Note: the “missingness” on Y can be correlated with the “missingness” on X. We can compare the value of other variables for the observations with missing data, and observations without missing data. If we reject the t-test for mean difference, we can say there is evidence that the data are not MCAR. But we cannot say that our data are MCAR if we fail to reject the t-test.

- the propensity for a data point to be missing is completely random.
- There’s no relationship between whether a data point is missing and any values in the data set, missing or observed.
- The missing data are just a random subset of the data.

11.1.2 Missing at Random (MAR)

Missing at Random, MAR, means there is a systematic relationship between the propensity of missing values and the observed data, but not the missing data. Whether an observation is missing has nothing to do with the missing values, but it does have to do with the values of an individual’s observed variables. So, for example, if men are more likely to tell you their weight than women, weight is MAR.

MAR is weaker than MCAR

$$P(Y_{missing}|Y, X) = P(Y_{missing}|X)$$

The probability of Y missing given Y and X equal to the probability of Y missing given X. However, it is impossible to provide evidence to the MAR condition.

- the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data. In another word, there is a systematic relationship between the propensity of missing values and the observed data, but not the missing data.
 - For example, if men are more likely to tell you their weight than women, weight is MAR
- MAR requires that the cause of the missing data is unrelated to the missing values but may be related to the observed values of other variables.
- MAR means that the missing values are related to observed values on other variables. As an example of CD missing data, missing income data may be unrelated to the actual income values but are related to education. Perhaps people with more education are less likely to reveal their income than those with less education

11.1.3 Ignorable

The missing data mechanism is ignorable when

- (1) The data are MAR
- (2) the parameters in the function of the missing data process are unrelated to the parameters (of interest) that need to be estimated.

In this case, you actually don't need to model the missing data mechanisms unless you would like to improve on your accuracy, in which case you still need to be very rigorous about your approach to improve efficiency in your parameters.

11.1.4 Nonignorable

Missing Not at Random, MNAR, means there is a relationship between the propensity of a value to be missing and its values.

Example: people with the lowest education are missing on education or the sickest people are most likely to drop out of the study.

MNAR is called Nonignorable because the missing data mechanism itself has to be modeled as you deal with the missing data. You have to include some model for why the data are missing and what the likely values are.

Hence, in the case of nonignorable, the data are not MAR. Then, your parameters of interest will be biased if you do not model the missing data mechanism. One of the most widely used approach for nonignorable missing data is (Heckman, 1976)

- Another name: Missing Not at Random (MNAR): there is a relationship between the propensity of a value to be missing and its values
 - For example, people with low education will be less likely to report it.
- We need to model why the data are missing and what the likely values are.
- the missing data mechanism is related to the missing values
- It commonly occurs when people do not want to reveal something very personal or unpopular about themselves
- Complete case analysis can give highly biased results for NI missing data. If proportionally more low and moderate income individuals are left in the sample because high income people are missing, an estimate of the mean income will be lower than the actual population mean.

11.2 Solutions to Missing data

11.2.1 Listwise Deletion

Also known as complete case deletion only where you only retain cases with complete data for all features.

Advantages:

- Can be applied to any statistical test (SEM, multi-level regression, etc.)
- In the case of MCAR, both the parameters estimates and its standard errors are unbiased.
- In the case of MAR among independent variables (not depend on the values of dependent variables), then listwise deletion parameter estimates can still be unbiased. (Little, 1992) For example, you have a model $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ if the probability of missing data on X_1 is independent of Y , but dependent on the value of X_1 and X_2 , then the model estimates are still unbiased.

- The missing data mechanism depends on the values of the independent variables are the same as stratified sampling. And stratified sampling does not bias your estimates
- In the case of logistic regression, if the probability of missing data on any variable depends on the value of the dependent variable, but independent of the value of the independent variables, then the listwise deletion will yield biased intercept estimate, but consistent estimates of the slope and their standard errors (Vach, 1994). However, logistic regression will still fail if the probability of missing data is dependent on both the value of the dependent and independent variables.
- Under regression analysis, listwise deletion is more robust than maximum likelihood and multiple imputation when MAR assumption is violated.

Disadvantages:

- It will yield a larger standard errors than other more sophisticated methods discussed later.
- If the data are not MCAR, but MAR, then your listwise deletion can yield biased estimates.
- In other cases than regression analysis, other sophisticated methods can yield better estimates compared to listwise deletion.

11.2.2 Pairwise Deletion

This method could only be used in the case of linear models such as linear regression, factor analysis, or SEM. The premise of this method based on that the coefficient estimates are calculated based on the means, standard deviations, and correlation matrix. Compared to listwise deletion, we still utilized as many correlation between variables as possible to compute the correlation matrix.

Advantages:

- If the true missing data mechanism is MCAR, pair wise deletion will yield consistent estimates, and unbiased in large samples
- Compared to listwise deletion: (Glasser, 1964)
 - If the correlation among variables are low, pairwise deletion is more efficient estimates than listwise
 - If the correlations among variables are high, listwise deletion is more efficient than pairwise.

Disadvantages:

- If the data mechanism is MAR, pairwise deletion will yield biased estimates.
- In small sample, sometimes covariance matrix might not be positive definite, which means coefficients estimates cannot be calculated.

Note: You need to read carefully on how your software specify the sample size because it will alter the standard errors.

11.2.3 Dummy Variable Adjustment

Also known as Missing Indicator Method or Proxy Variable

Add another variable in the database to indicate whether a value is missing.

Create 2 variables

$$D = \begin{cases} 1 & \text{data on } X \text{ are missing} \\ 0 & \text{otherwise} \end{cases} \quad (11.1)$$

$$X^* = \begin{cases} X & \text{data are available} \\ c & \text{data are missing} \end{cases} \quad (11.2)$$

Note: A typical choice for c is usually the mean of X

Interpretation:

- Coefficient of D is the difference in the expected value of Y between the group with data and the group without data on X .
- Coefficient of X^* is the effect of the group with data on Y

Disadvantages:

- This method yields bias estimates of the coefficient even in the case of MCAR (Jones, 1996)

11.2.4 Imputation

11.2.4.1 Mean, Mode, Median Imputation

- Bad:
 - Mean imputation does not preserve the relationships among variables
 - Mean imputation leads to An Underestimate of Standard Errors → you're making Type I errors without realizing it.

- Biased estimates of variances and covariances (Haitovsky, 1968)
- In high-dimensions, mean substitution cannot account for dependence structure among features.

11.2.4.2 Maximum Likelihood

When missing data are MAR and monotonic (such as in the case of panel studies), ML can be adequately in estimating coefficients.

Monotonic means that if you are missing data on X₁, then that observation also has missing data on all other variables that come after it.

ML can generally handle linear models, log-linear model, but beyond that, ML still lacks both theory and software to implement.

11.2.4.2.1 Expectation-Maximization Algorithm (EM Algorithm)

An iterative process:

- (1) Other variables are used to impute a value (Expectation).
- (2) Check whether the value is most likely (Maximization).
- (3) If not, it re-imputes a more likely value.

You start your regression with your estimates based on either listwise deletion or pairwise deletion. After regressing missing variables on available variables, you obtain a regression model. Plug the missing data back into the original model, with modified variances and covariances. For example, if you have missing data on X_{ij} you would regress it on available data of $X_{i(j)}$, then plug the expected value of X_{ij} back with its X_{ij}^2 turn into $X_{ij}^2 + s_{j(j)}^2$ where $s_{j(j)}^2$ stands for the residual variance from regressing X_{ij} on $X_{i(j)}$. With the new estimated model, you rerun the process until the estimates converge.

Advantages:

- (1) easy to use
- (2) preserves the relationship with other variables (important if you use Factor Analysis or Linear Regression later on), but best in the case of Factor Analysis, which doesn't require standard error of individuals item.

Disadvantages:

- (1) Standard errors of the coefficients are incorrect (biased usually downward - underestimate)
- (2) Models with overidentification, the estimates will not be efficient

11.2.4.2.2 Direct ML (raw maximum likelihood) Advantages

- (1) efficient estimates and correct standard errors.

Disadvantages:

- (1) Hard to implement

11.2.4.3 Multiple Imputation

MI is designed to use “the Bayesian model-based approach to *create* procedures, and the frequentist (randomization-based approach) to *evaluate* procedures”. (Rubin, 1996)

MI estimates have the same properties as ML when the data is MAR

- Consistent
- Asymptotically efficient
- Asymptotically normal

MI can be applied to any type of model, unlike Maximum Likelihood that is only limited to a small set of models.

A drawback of MI is that it will produce slightly different estimates every time you run it. To avoid such problem, you can set seed when doing your analysis to ensure its reproducibility.

11.2.4.3.1 Single Random Imputation

Random draws from the residual distribution of each imputed variable and add those random numbers to the imputed values.

For example, if we have missing data on X, and it's MCAR, then

- (1) regress X on Y (Listwise Deletion method) to get its residual distribution.
- (2) For every missing value on X, we substitute with $\tilde{x}_i = \hat{x}_i + \rho u_i$ where
 - u_i is a random draw from a standard normal distribution
 - \hat{x}_i is the predicted value from the regression of X and Y
 - ρ is the standard deviation of the residual distribution of X regressed on Y.

However, the model you run with the imputed data still thinks that your data are collected, not imputed, which leads your standard error estimates to be too low and test statistics too high.

To address this problem, we need to repeat the imputation process which leads us to repeated imputation or multiple random imputation.

11.2.4.3.2 Repeated Imputation “Repeated imputations are draws from the posterior predictive distribution of the missing values under a specific model , a particular Bayesian model for both the data and the missing mechanism”(Rubin, 1996)

Repeated imputation, also known as, multiple random imputation, allows us to have multiple “completed” data sets. The variability across imputations will adjust the standard errors upward.

The estimate of the standard error of \bar{r} (mean correlation estimates between X and Y) is

$$SE(\bar{r}) = \sqrt{\frac{1}{M} \sum_k s_k^2 + (1 + \frac{1}{M})(\frac{1}{M-1}) \sum_k (r_k - \bar{r})^2}$$

where M is the number of replications, r_k is the the correlation in replication k, s_k is the estimated standard error in replication k.

However, this method still considers the parameter in predicting \tilde{x} is still fixed, which means we assume that we are using the true parameters to predict \tilde{x} . To overcome this challenge, we need to introduce variability into our model for \tilde{x} by treating the parameters as a random variables and use Bayesian posterior distribution of the parameters to predict the parameters.

However, if your sample is large and the proportion of missing data is small, the extra Bayesian step might not be necessary. If your sample is small or the proportion of missing data is large, the extra Bayesian step is necessary.

Two algorithms to get random draws of the regression parameters from its posterior distribution:

- Data Augmentation
- Sampling importance/resampling (SIR)

Authors have argued for SIR superiority due to its computer time (King et al., 2001)

11.2.4.3.2.1 Data Augmentation Steps for data augmentation:

- (1) Choose starting values for the parameters (e.g., for multivariate normal, choose means and covariance matrix). These values can come from previous values, expert knowledge, or from listwise deletion or pairwise deletion or EM estimation.
- (2) Based on the current values of means and covariances calculate the coefficients estimates for the equation that variable with missing data is regressed on all other variables (or variables that you think will help predict the missing values, could also be variables that are not in the final estimation model)

- (3) Use the estimates in step (2) to predict values for missing values. For each predicted value, add a random error from the residual normal distribution for that variable.
- (4) From the “complete” data set, recalculate the means and covariance matrix. And take a random draw from the posterior distribution of the means and covariances with Jeffreys’ prior.
- (5) Using the random draw from step (4), repeat step (2) to (4) until the means and covariances stabilize (converged).

The iterative process allows us to get random draws from the joint posterior distribution of both data and parameters, given the observed data.

Rules of thumb regarding convergence:

- The higher the proportion of missing, the more iterations
- the rate of convergence for EM algorithm should be the minimum threshold for DA.
- You can also check if your distribution has been converged by diagnostic statistics Can check Bayesian Diagnostics for some introduction.

Types of chains

1. **Parallel:** Run a separate chain of iterations for each of data set. Different starting values are encouraged. For example, one could use bootstrap to generate different data set with replacement, and for each data set, calculate the starting values by EM estimates.
 - Pro: Run faster, and less likely to have dependence in the resulting data sets.
 - Con: Sometimes it will not converge
2. **Sequential** one long chain of data augmentation cycles. After burn-in and thinning, you will have to data sets
 - Pro: Converged to the true posterior distribution is more likely.
 - Con: The resulting data sets are likely to be dependent. Remedies can be thinning and burn-in.

Note on Non-normal or categorical data The normal-based methods still work well, but you will need to do some transformation. For example,

- If the data is skewed, then log-transform, then impute, then exponentiate to have the missing data back to its original metric.
- If the data is proportion, logit-transform, impute, then de-transform the missing data.

If you want to impute non-linear relationship, such as interaction between 2 variables and 1 variable is categorical. You can do separate imputation for different levels of that variable separately, then combined for the final analysis.

- If all variables that have missing data are categorical, then **unrestricted multinomial model** or **log-linear model** is recommended.
- If a single categorical variable, **logistic (logit) regression** would be sufficient.

11.2.4.4 Nonparametric/ Semiparametric Methods

11.2.4.4.1 Hot Deck Imputation

- Used by U.S. Census Bureau for public datasets
- approximate Bayesian bootstrap
- A randomly chosen value from an individual in the sample who has similar values on other variables. In other words, find all the sample subjects who are similar on other variables, then randomly choose one of their values on the missing variable.

When we have n_1 cases with complete data on Y and n_0 cases with missing data on Y

- Step 1: From n_1 , take a random sample (with replacement) of n_1 cases
- Step 2: From the retrieved sample take a random sample (with replacement) of n_0 cases
- Step 3: Assign the n_0 cases in step 2 to n_0 missing data cases.
- Step 4: Repeat the process for every variable.
- Step 5: For multiple imputation, repeat the four steps multiple times.

Note:

- If we skip step 1, it reduce variability for estimating standard errors.
- Good:
 - Constrained to only possible values.
 - Since the value is picked at random, it adds some variability, which might come in handy when calculating standard errors.
- Challenge: how can you define “similar” here.

11.2.4.4.2 Cold Deck Imputation Contrary to Hot Deck, Cold Deck choose value systematically from an observation that has similar values on other variables, which remove the random variation that we want.

Donor samples of “cold-deck” imputation come from a different data set.

11.2.4.4.3 Predictive Mean Matching Steps:

1. Regress Y on X (matrix of covariates) for the n_1 (i.e., non-missing cases) to get coefficients b (a $k \times 1$ vector) and residual variance estimates s^2
2. Draw randomly from the posterior predictive distribution of the residual variance (assuming a noninformative prior) by calculating $\frac{(n_1-k)s^2}{\chi^2}$, where χ^2 is a random draw from a $\chi^2_{n_1-k}$ and let $s_{[1]}^2$ be an i-th random draw
3. Randomly draw from the posterior distribution of the coefficients b , by drawing from $MVN(b, s_{[1]}^2(X'X)^{-1})$, where X is an $n_1 \times k$ matrix of X values. Then we have b_1
4. Using step 1, we can calculate standardized residuals for n_1 cases: $e_i = \frac{y_i - bx_i}{\sqrt{s^2(1-k/n_1)}}$
5. Randomly draw a sample (with replacement) of n_0 from the n_1 residuals in step 4
6. With n_0 cases, we can calculate imputed values of Y: $y_i = b_{[1]}x_i + s_{[1]}e_i$ where e_i are taken from step 5, and $b_{[1]}$ taken from step 3, and $s_{[1]}$ taken from step 2.
7. Repeat steps 2 through 6 except for step 4.

Notes:

- can be used for multiple variables where each variable is imputed using all other variables as predictor.
- can also be used for heteroskedasticity in imputed values.

Example from Statistics Globble

```

set.seed(918273)                                     # Seed
N <- 3000                                         # Sample size
y <- round(runif(N, -10, 10))                     # Target variable Y
x1 <- y + round(runif(N, 0, 50))                  # Auxiliary variable 1
x2 <- round(y + 0.25 * x1 + rnorm(N, -3, 15))    # Auxiliary variable 2
x3 <- round(0.1 * x1 + rpois(N, 2))              # Auxiliary variable 3
x4 <- as.factor(round(0.02 * y + runif(N)))       # Auxiliary variable 4 (categorical variable)
y[rbinom(N, 1, 0.2) == 1] <- NA                   # Insert 20% missing data in Y
data <- data.frame(y, x1, x2, x3, x4)             # Store data in dataset
head(data)                                         # First 6 rows of our data

#>   y  x1  x2  x3  x4
#> 1  8 38  -3   6   1
#> 2  1 50  -9   5   0
#> 3  5 43  20   5   1
#> 4 NA  9  13   3   0
#> 5 -4 40 -10   6   0
#> 6 NA 29  -6   5   1

```

```
library("mice")                                # Load mice package

##### Impute data via predictive mean matching (single imputation)#####

imp_single <- mice(data, m = 1, method = "pmm") # Impute missing values
#>
#> iter imp variable
#> 1 1 y
#> 2 1 y
#> 3 1 y
#> 4 1 y
#> 5 1 y
data_imp_single <- complete(imp_single)          # Store imputed data
# head(data_imp_single)

# Since single imputation underestimates standard errors, we use multiple imputation

##### Predictive mean matching (multiple imputation)#####

imp_multi <- mice(data, m = 5, method = "pmm") # Impute missing values multiple times
#>
#> iter imp variable
#> 1 1 y
#> 1 2 y
#> 1 3 y
#> 1 4 y
#> 1 5 y
#> 2 1 y
#> 2 2 y
#> 2 3 y
#> 2 4 y
#> 2 5 y
#> 3 1 y
#> 3 2 y
#> 3 3 y
#> 3 4 y
#> 3 5 y
#> 4 1 y
#> 4 2 y
#> 4 3 y
#> 4 4 y
#> 4 5 y
#> 5 1 y
#> 5 2 y
#> 5 3 y
```

```

#>   5   4   y
#>   5   5   y
data_imp_multi_all <- complete(imp_multi,
                                "repeated",
                                include = TRUE) # Store multiply imputed data

data_imp_multi <- data.frame(
  data_imp_multi_all[, 1:6], data[, 2:5]) # Combine imputed Y and X1-X4 (for con

head(data_imp_multi) # First 6 rows of our multiply imputed data
#>   y.0 y.1 y.2 y.3 y.4 y.5 x1  x2 x3 x4
#> 1   8   8   8   8   8   8 38  -3   6   1
#> 2   1   1   1   1   1   1 50  -9   5   0
#> 3   5   5   5   5   5   5 43  20   5   1
#> 4   NA  -6  -4  -4  -1  -3   9  13   3   0
#> 5   -4  -4  -4  -4  -4  -4 40  -10  6   0
#> 6   NA  -8   5  -4   1   4 29  -6   5   1

```

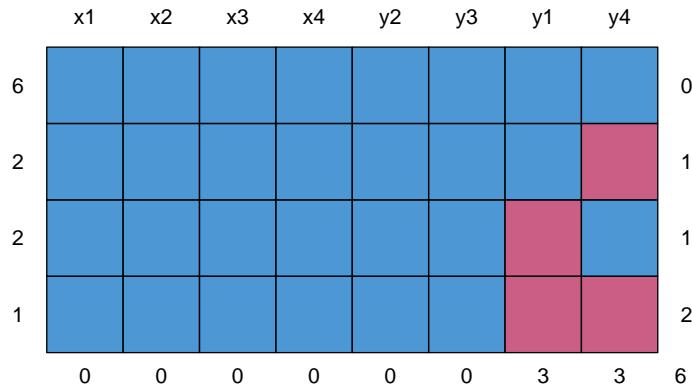
Example from UCLA Statistical Consulting (Bruin, 2011)

```

library(mice)
library(VIM)
library(lattice)
library(ggplot2)
## set observations to NA
anscombe <- within(anscombe, {
  y1[1:3] <- NA
  y4[3:5] <- NA
})
## view
head(anscombe)
#>   x1 x2 x3 x4    y1    y2    y3    y4
#> 1 10 10 10  8    NA 9.14  7.46 6.58
#> 2  8   8   8   8    NA 8.14  6.77 5.76
#> 3 13 13 13  8    NA 8.74 12.74    NA
#> 4  9   9   9   8 8.81 8.77  7.11    NA
#> 5 11 11 11  8 8.33 9.26  7.81    NA
#> 6 14 14 14  8 9.96 8.10  8.84 7.04

## check missing data patterns
md.pattern(anscombe)

```



```

#>   x1 x2 x3 x4 y2 y3 y1 y4
#> 6  1  1  1  1  1  1  1  1  0
#> 2  1  1  1  1  1  1  1  0  1
#> 2  1  1  1  1  1  1  0  1  1
#> 1  1  1  1  1  1  1  0  0  2
#>   0  0  0  0  0  0  3  3  6

## Number of observations per patterns for all pairs of variables
p <- md.pairs(anscombe)
p # rr = number of observations where both pairs of values are observed
#> $rr
#>   x1 x2 x3 x4 y1 y2 y3 y4
#> x1 11 11 11 11  8 11 11  8
#> x2 11 11 11 11  8 11 11  8
#> x3 11 11 11 11  8 11 11  8
#> x4 11 11 11 11  8 11 11  8
#> y1  8  8  8  8  8  8  8  6
#> y2 11 11 11 11  8 11 11  8
#> y3 11 11 11 11  8 11 11  8
#> y4  8  8  8  8  6  8  8  8
#>
#> $rm
#>   x1 x2 x3 x4 y1 y2 y3 y4
#> x1  0  0  0  0  3  0  0  3
#> x2  0  0  0  0  3  0  0  3
#> x3  0  0  0  0  3  0  0  3
#> x4  0  0  0  0  3  0  0  3

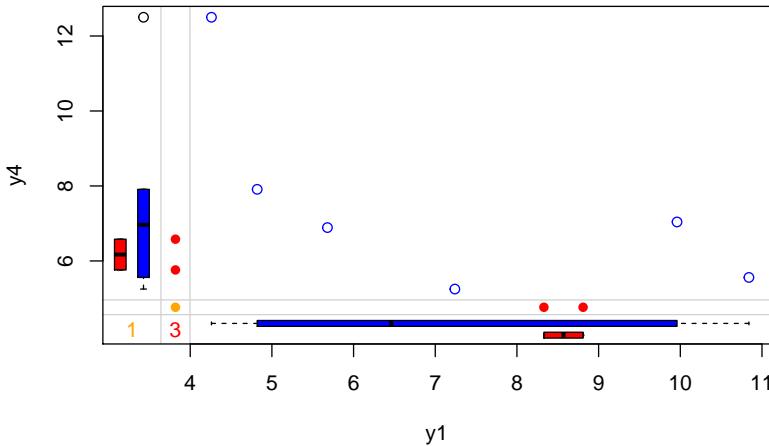
```

```

#> y1  0  0  0  0  0  0  0  0  2
#> y2  0  0  0  0  3  0  0  0  3
#> y3  0  0  0  0  3  0  0  0  3
#> y4  0  0  0  0  2  0  0  0  0
#>
#> $mr
#>   x1 x2 x3 x4 y1 y2 y3 y4
#> x1  0  0  0  0  0  0  0  0  0
#> x2  0  0  0  0  0  0  0  0  0
#> x3  0  0  0  0  0  0  0  0  0
#> x4  0  0  0  0  0  0  0  0  0
#> y1  3  3  3  3  0  3  3  2
#> y2  0  0  0  0  0  0  0  0  0
#> y3  0  0  0  0  0  0  0  0  0
#> y4  3  3  3  3  2  3  3  0
#>
#> $mm
#>   x1 x2 x3 x4 y1 y2 y3 y4
#> x1  0  0  0  0  0  0  0  0  0
#> x2  0  0  0  0  0  0  0  0  0
#> x3  0  0  0  0  0  0  0  0  0
#> x4  0  0  0  0  0  0  0  0  0
#> y1  0  0  0  0  3  0  0  0  1
#> y2  0  0  0  0  0  0  0  0  0
#> y3  0  0  0  0  0  0  0  0  0
#> y4  0  0  0  0  1  0  0  0  3
# rm = the number of observations where both variables are missing values
# mr = the number of observations where the first variable's value (e.g. the row variab
# mm = the number of observations where the second variable's value (e.g. the col variab

## Margin plot of y1 and y4
marginplot(anscombe[c(5, 8)], col = c("blue", "red", "orange"))

```



```
## 5 imputations for all missing values
imp1 <- mice(anscombe, m = 5)
#>
#>   iter imp variable
#>   1   1   y1   y4
#>   1   2   y1   y4
#>   1   3   y1   y4
#>   1   4   y1   y4
#>   1   5   y1   y4
#>   2   1   y1   y4
#>   2   2   y1   y4
#>   2   3   y1   y4
#>   2   4   y1   y4
#>   2   5   y1   y4
#>   3   1   y1   y4
#>   3   2   y1   y4
#>   3   3   y1   y4
#>   3   4   y1   y4
#>   3   5   y1   y4
#>   4   1   y1   y4
#>   4   2   y1   y4
#>   4   3   y1   y4
#>   4   4   y1   y4
#>   4   5   y1   y4
#>   5   1   y1   y4
#>   5   2   y1   y4
```

```

#>   5   3   y1   y4
#>   5   4   y1   y4
#>   5   5   y1   y4

## linear regression for each imputed data set - 5 regression are run
fitm <- with(imp1, lm(y1 ~ y4 + x1))
summary(fitm)
#> # A tibble: 15 x 6
#>   term      estimate std.error statistic p.value    nobs
#>   <chr>     <dbl>     <dbl>     <dbl>     <dbl> <int>
#> 1 (Intercept) 6.44      1.77      3.63  0.00664    11
#> 2 y4        -0.381     0.162     -2.35  0.0468    11
#> 3 x1        0.458      0.101      4.56  0.00186    11
#> 4 (Intercept) 6.17      2.02      3.06  0.0156    11
#> 5 y4        -0.351     0.183     -1.92  0.0914    11
#> 6 x1        0.443      0.117      3.79  0.00535    11
#> 7 (Intercept) 6.47      2.02      3.21  0.0125    11
#> 8 y4        -0.360     0.186     -1.93  0.0893    11
#> 9 x1        0.430      0.114      3.77  0.00549    11
#> 10 (Intercept) 6.51      2.63      2.48  0.0384    11
#> 11 y4        -0.358     0.242     -1.48  0.178    11
#> 12 x1        0.433      0.149      2.91  0.0195    11
#> 13 (Intercept) 6.15      2.37      2.60  0.0318    11
#> 14 y4        -0.362     0.219     -1.65  0.137    11
#> 15 x1        0.478      0.138      3.46  0.00855   11

## pool coefficients and standard errors across all 5 regression models
pool(fitm)
#> Class: mipo   m = 5
#>          term m  estimate      ubar         b          t dfcom      df
#> 1 (Intercept) 5 6.3487398 4.76263432 0.0300416474 4.79868430  8 6.495687
#> 2           y4 5 -0.3623970 0.04025189 0.0001255017 0.04040250  8 6.520908
#> 3           x1 5  0.4485592 0.01560878 0.0003990336 0.01608762  8 6.341712
#>          riv lambda      fmi
#> 1 0.007569335 0.007512471 0.2165521
#> 2 0.003741490 0.003727544 0.2130085
#> 3 0.030677620 0.029764516 0.2374856

## output parameter estimates
summary(pool(fitm))
#>          term  estimate std.error statistic      df  p.value
#> 1 (Intercept) 6.3487398 2.1905899 2.898187 6.495687 0.02504369
#> 2           y4 -0.3623970 0.2010037 -1.802937 6.520908 0.11751237
#> 3           x1  0.4485592 0.1268370 3.536502 6.341712 0.01119463

```

11.2.4.4.4 Stochastic Imputation

Regression imputation + random residual = Stochastic Imputation

Most multiple imputation is based off of some form of stochastic regression imputation.

Good:

- Has all the advantage of Regression Imputation
- and also has the random components

Bad:

- might lead to implausible values (e.g. negative values)
- can't handle heteroskedastic data

Note

Multiple Imputation usually based on some form of stochastic regression imputation.

```
# Income data

set.seed(91919)                      # Set seed
N <- 1000                             # Sample size

income <- round(rnorm(N, 0, 500))      # Create some synthetic income data
income[income < 0] <- income[income < 0] * (- 1)

x1 <- income + rnorm(N, 1000, 1500)    # Auxiliary variables
x2 <- income + rnorm(N, - 5000, 2000)

income[rbinom(N, 1, 0.1) == 1] <- NA    # Create 10% missingness in income

data_inc_miss <- data.frame(income, x1, x2)
```

Single stochastic regression imputation

```
imp_inc_sri <- mice(data_inc_miss, method = "norm.nob", m = 1)
#>
#> iter imp variable
#>   1   1   income
#>   2   1   income
#>   3   1   income
#>   4   1   income
#>   5   1   income
data_inc_sri <- complete(imp_inc_sri)
```

Single predictive mean matching

```
imp_inc_pmm <- mice(data_inc_miss, method = "pmm", m = 1)
#>
#> iter imp variable
#> 1 1 income
#> 2 1 income
#> 3 1 income
#> 4 1 income
#> 5 1 income
data_inc_pmm <- complete(imp_inc_pmm)
```

Stochastic regression imputation contains negative values

```
data_inc_sri$income[data_inc_sri$income < 0]
#> [1] -66.055957 -96.980053 -28.921432 -4.175686 -54.480798 -27.207102
#> [7] -143.603500 -80.960488
data_inc_pmm$income[data_inc_pmm$income < 0] # No values below 0
#> numeric(0)
```

Proof for heteroskedastic data

```
# Heteroscedastic data

set.seed(654654)                                # Set seed
N <- 1:5000                                     # Sample size

a <- 0
b <- 1
sigma2 <- N^2
eps <- rnorm(N, mean = 0, sd = sqrt(sigma2))

y <- a + b * N + eps                           # Heteroscedastic variable
x <- 30 * N + rnorm(N[length(N)], 1000, 200) # Correlated variable

y[rbinom(N[length(N)], 1, 0.3) == 1] <- NA    # 30% missings

data_het_miss <- data.frame(y, x)
```

Single stochastic regression imputation

```
imp_het_sri <- mice(data_het_miss, method = "norm.nob", m = 1)
#>
#> iter imp variable
```

```
#> 1 1 y
#> 2 1 y
#> 3 1 y
#> 4 1 y
#> 5 1 y
data_het_sri <- complete(imp_het_sri)
```

Single predictive mean matching

```
imp_het_pmm <- mice(data_het_miss, method = "pmm", m = 1)
#>
#> iter imp variable
#> 1 1 y
#> 2 1 y
#> 3 1 y
#> 4 1 y
#> 5 1 y
data_het_pmm <- complete(imp_het_pmm)
```

Comparison between predictive mean matching and stochastic regression imputation

```
par(mfrow = c(1, 2))                                # Both plots in one graphic

plot(x[!is.na(data_het_sri$y)],                  # Plot of observed values
      data_het_sri$y[!is.na(data_het_sri$y)],
      main = "",                                     # Title of plot
      xlab = "X", ylab = "Y")
points(x[is.na(y)], data_het_sri$y[is.na(y)],    # Plot of missing values
       col = "red")
title("Stochastic Regression Imputation",        # Title of plot
      line = 0.5)
abline(lm(y ~ x, data_het_sri),                  # Regression line
       col = "#1b98e0", lwd = 2.5)
legend("topleft",                                 # Legend
       c("Observed Values", "Imputed Values", "Regression Y ~ X"),
       pch = c(1, 1, NA),
       lty = c(NA, NA, 1),
       col = c("black", "red", "#1b98e0"))

plot(x[!is.na(data_het_pmm$y)],                 # Plot of observed values
      data_het_pmm$y[!is.na(data_het_pmm$y)],
      main = "",                                     # Title of plot
      xlab = "X", ylab = "Y")
```

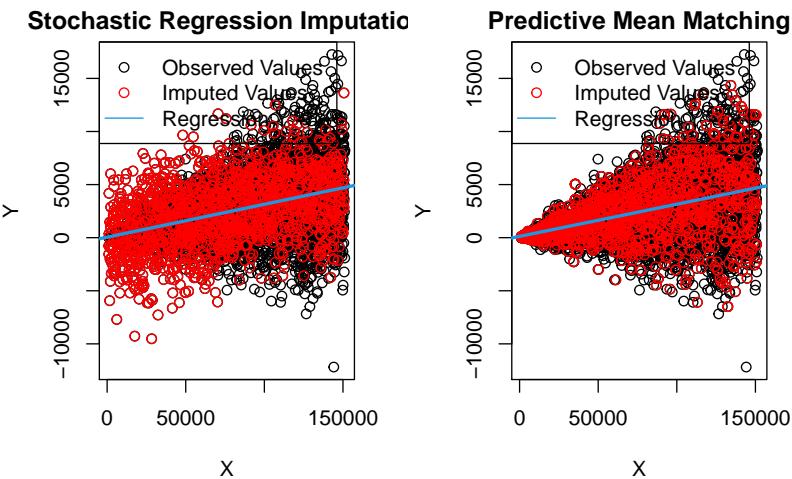
```

points(x[is.na(y)], data_het_pmm$y[is.na(y)],      # Plot of missing values
       col = "red")
title("Predictive Mean Matching",                  # Title of plot
      line = 0.5)
abline(lm(y ~ x, data_het_pmm),
       col = "#1b98e0", lwd = 2.5)
legend("topleft",                                # Legend
       c("Observed Values", "Imputed Values", "Regression Y ~ X"),
       pch = c(1, 1, NA),
       lty = c(NA, NA, 1),
       col = c("black", "red", "#1b98e0"))

mtext("Imputation of Heteroscedastic Data",      # Main title of plot
      side = 3, line = - 1.5, outer = TRUE, cex = 2)

```

Imputation of Heteroscedastic Data



11.2.4.5 Regression Imputation

Also known as conditional mean imputation Missing value is based (regress) on other variables.

- Good:
 - Maintain the relationship with other variables (i.e., preserve dependence structure among features, unlike 11.2.4.1).

- If the data are MCAR, least-squares coefficients estimates will be consistent, and approximately unbiased in large samples (Gourieroux and Monfort, 1981)
 - * Can have improvement on efficiency by using weighted least squares (Beale and Little, 1975) or generalized least squares (Gourieroux and Monfort, 1981).
- Bad:
 - No variability left. treated data as if they were collected.
 - Underestimate the standard errors and overestimate test statistics

11.2.4.6 Interpolation and Extrapolation

An estimated value from other observations from the same individual. It usually only works in longitudinal data.

11.2.4.7 K-nearest neighbor (KNN) imputation

The above methods are model-based imputation (regression).
 This is an example of neighbor-based imputation (K-nearest neighbor).

For every observation that needs to be imputed, the algorithm identifies ‘k’ closest observations based on some types distance (e.g., Euclidean) and computes the weighted average (weighted based on distance) of these ‘k’ obs.

For a discrete variable, it uses the most frequent value among the k nearest neighbors.

- Distance metrics: Hamming distance.

For a continuous variable, it uses the mean or mode.

- Distance metrics:
 - Euclidean
 - Mahalanobis
 - Manhattan

11.2.4.8 Bayesian Ridge regression implementation

11.2.4.9 Matrix Completion

Impute items missing at random while accounting for dependence between features by using principal components, which is known as **matrix completion** (James et al., 2013, Sec 12.3)

Consider an $n \times p$ feature matrix, \mathbf{X} , with element x_{ij} , some of which are missing.

Similar to 20.2, we can approximate the matrix \mathbf{X} in terms of its leading PCs.

We consider the M principal components that optimize

$$\min_{\mathbf{A} \in R^{n \times M}, \mathbf{B} \in R^{p \times M}} \left\{ \sum_{(i,j) \in O} (x_{ij} - \sum_{m=1}^M a_{im} b_{jm})^2 \right\}$$

where O is the set of all observed pairs indices (i, j) , a subset of the possible $n \times p$ pairs

Once this minimization is solved,

- One can impute a missing observation, x_{ij} , with $\hat{x}_{ij} = \sum_{m=1}^M \hat{a}_{im} \hat{b}_{jm}$ where $\hat{a}_{im}, \hat{b}_{jm}$ are the (i, m) and (j, m) elements, respectively, of the matrices $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ from the minimization, and
- One can approximately recover the M principal component scores and loadings, as we did when the data were complete

The challenge here is to solve this minimization problem: the eigen-decomposition no longer applies (as in 20.2)

Hence, we have to use iterative algorithm (James et al., 2013, Alg 12.1)

1. Create a complete data matrix $\tilde{\mathbf{X}}$ of dimension $n \times p$ of which the (i, j) element equals

$$\tilde{x}_{ij} = \begin{cases} x_{ij} & \text{if } (i, j) \in O \\ \bar{x}_j & \text{if } (i, j) \notin O \end{cases}$$

where \bar{x}_j is the average of the observed values for the j th variable in the incomplete data matrix \mathbf{X}

O indexes the observations that are observed in \mathbf{X}

2. Repeat these 3 steps until some objectives are met

a. Solve

$$\min_{\mathbf{A} \in R^{n \times M}, \mathbf{B} \in R^{p \times M}} \left\{ \sum_{(i,j) \in O} (x_{ij} - \sum_{m=1}^M a_{im} b_{jm})^2 \right\}$$

by computing the principal components of $\tilde{\mathbf{X}}$

- b. For each element $(i, j) \notin O$, set $\tilde{x}_{ij} \leftarrow \sum_{m=1}^M \hat{a}_{im} \hat{b}_{jm}$
- c. Compute the objective

$$\sum_{(i,j) \in O} (x_{ij} - \sum_{m=1}^M \hat{a}_{im} \hat{b}_{jm})^2$$

- 3. Return the estimated missing entries $\tilde{x}_{ij}, (i, j) \notin O$

11.2.5 Other methods

- For panel data, or clustered data, use `pan` package by Schafer (1997)

11.3 Criteria for Choosing an Effective Approach

Criteria for an ideal technique in treating missing data:

1. Unbiased parameter estimates
2. Adequate power
3. Accurate standard errors (p-values, confidence intervals)

The Multiple Imputation and Full Information Maximum Likelihood are the most ideal candidate. Single imputation will generally lead to underestimation of standard errors.

11.4 Another Perspective

Model bias can arise from various factors including:

- Imputation method
- Missing data mechanism (MCAR vs. MAR)
- Proportion of the missing data
- Information available in the data set

Since the imputed observations are themselves estimates, their values have corresponding random error. But when you put in that estimate as a data point, your software doesn't know that. So it overlooks the extra source of error, resulting in too-small standard errors and too-small p-values. So multiple imputation comes up with multiple estimates.

Because multiple imputation have a random component, the multiple estimates are slightly different. This re-introduces some variation that your software can incorporate in order to give your model accurate estimates of standard error. Multiple imputation was a huge breakthrough in statistics about 20 years ago. It solves a lot of problems with missing data (though, unfortunately not all) and if done well, leads to unbiased parameter estimates and accurate standard errors. If your rate of missing data is very, very small (2-3%) it doesn't matter what technique you use.

Remember that there are three goals of multiple imputation, or any missing data technique:

- Unbiased parameter estimates in the final analysis (regression coefficients, group means, odds ratios, etc.)
- accurate standard errors of those parameter estimates, and therefore, accurate p-values in the analysis
- adequate power to find meaningful parameter values significant.

Hence,

1. Don't round off imputations for dummy variables. Many common imputation techniques, like MCMC, require normally distributed variables. Suggestions for imputing categorical variables were to dummy code them, impute them, then round off imputed values to 0 or 1. Recent research, however, has found that rounding off imputed values actually leads to biased parameter estimates in the analysis model. You actually get better results by leaving the imputed values at impossible values, even though it's counter-intuitive.
2. Don't transform skewed variables. Likewise, when you transform a variable to meet normality assumptions before imputing, you not only are changing the distribution of that variable but the relationship between that variable and the others you use to impute. Doing so can lead to imputing outliers, creating more bias than just imputing the skewed variable.

3. Use more imputations. The advice for years has been that 5-10 imputations are adequate. And while this is true for unbiasedness, you can get inconsistent results if you run the multiple imputation more than once. (Bodner, 2008) recommends having as many imputations as the percentage of missing data. Since running more imputations isn't any more work for the data analyst, there's no reason not to.
4. Create multiplicative terms before imputing. When the analysis model contains a multiplicative term, like an interaction term or a quadratic, create the multiplicative terms first, then impute. Imputing first, and then creating the multiplicative terms actually biases the regression parameters of the multiplicative term (von Hippel, 2009)

11.5 Diagnosing the Mechanism

11.5.1 MAR vs. MNAR

The only true way to distinguish between MNAR and MAR is to measure some of that missing data.

It's a common practice among professional surveyors to, for example, follow-up on a paper survey with phone calls to a group of the non-respondents and ask a few key survey items. This allows you to compare respondents to non-respondents.

If their responses on those key items differ by very much, that's good evidence that the data are MNAR.

However in most missing data situations, we can't get a hold of the missing data. So while we can't test it directly, we can examine patterns in the data get an idea of what's the most likely mechanism.

The first thing in diagnosing randomness of the missing data is to use your substantive scientific knowledge of the data and your field. The more sensitive the issue, the less likely people are to tell you. They're not going to tell you as much about their cocaine usage as they are about their phone usage.

Likewise, many fields have common research situations in which non-ignorable data is common. Educate yourself in your field's literature.

11.5.2 MCAR vs. MAR

There is a very useful test for MCAR, Little's test.

A second technique is to create dummy variables for whether a variable is missing.

1 = missing 0 = observed

You can then run t-tests and chi-square tests between this variable and other variables in the data set to see if the missingness on this variable is related to the values of other variables.

For example, if women really are less likely to tell you their weight than men, a chi-square test will tell you that the percentage of missing data on the weight variable is higher for women than men.

11.6 Application

How many imputation:

Usually 5. (unless you have extremely high portion of missing, in which case you probably need to check your data again)

According to Rubin, the relative efficiency of an estimate based on m imputations to infinity imputation is approximately

$$(1 + \frac{\lambda}{m})^{-1}$$

where λ is the rate of missing data

Example 50% of missing data means an estimate based on 5 imputation has standard deviation that is only 5% wider compared to an estimate based on infinity imputation
 $(\sqrt{1 + 0.5/5} = 1.049)$

```
library(missForest)

#load data
data <- iris

#Generate 10% missing values at Random
set.seed(1)
iris.mis <- prodNA(iris, noNA = 0.1)

#remove categorical variables
iris.mis.cat <- iris.mis
iris.mis <- subset(iris.mis, select = -c(Species))
```

11.6.1 Imputation with mean / median / mode

```
# whole data set
e1071::impute(iris.mis, what = "mean") # replace with mean
e1071::impute(iris.mis, what = "median") # replace with median

# by variables
Hmisc::impute(iris.mis$Sepal.Length, mean) # mean
Hmisc::impute(iris.mis$Sepal.Length, median) # median
Hmisc::impute(iris.mis$Sepal.Length, 0) # replace specific number
```

check accuracy

```
library(DMwR)
actuals <- iris$Sepal.Width[is.na(iris.mis$Sepal.Width)]
predicteds <- rep(mean(iris$Sepal.Width, na.rm=T), length(actuals))
regr.eval(actuals, predicteds)
#>      mae      mse      rmse      mape
#> 0.2870303 0.1301598 0.3607767 0.1021485
```

11.6.2 KNN

```
library(DMwR)
# iris.mis[, !names(iris.mis) %in% c("Sepal.Length")]
# data should be this line. But since knn cant work with 3 or less variables, we need to use at least 4

# knn is not appropriate for categorical variables
knnOutput <-
  knnImputation(data = iris.mis.cat,
                 #k = 10,
                 meth = "median" # could use "median" or "weighAvg"
                 ) # should exclude the dependent variable: Sepal.Length
anyNA(knnOutput)
#> [1] FALSE

library(DMwR)
actuals <- iris$Sepal.Width[is.na(iris.mis$Sepal.Width)]
predicteds <- knnOutput[is.na(iris.mis$Sepal.Width), "Sepal.Width"]
regr.eval(actuals, predicteds)
#>      mae      mse      rmse      mape
#> 0.2318182 0.1038636 0.3222788 0.0823571
```

Compared to mape (mean absolute percentage error) of mean imputation, we see almost always see improvements.

11.6.3 rpart

For categorical (factor) variables, rpart can handle

```
library(rpart)
class_mod <- rpart(Species ~ . - Sepal.Length, data=iris.mis.cat[!is.na(iris.mis.cat$Species),])
anova_mod <- rpart(Sepal.Width ~ . - Sepal.Length, data=iris.mis[!is.na(iris.mis$Sepal.Width),])
species_pred <- predict(class_mod, iris.mis.cat[is.na(iris.mis.cat$Species),])
width_pred <- predict(anova_mod, iris.mis[is.na(iris.mis$Sepal.Width),])
```

11.6.4 MICE (Multivariate Imputation via Chained Equations)

Assumption: data are MAR

It imputes data per variable by specifying an imputation model for each variable

Example

We have X_1, X_2, \dots, X_k . If X_1 has missing data, then it is regressed on the rest of the variables. Same procedure applies if X_2 has missing data. Then, predicted values are used in place of missing values.

By default,

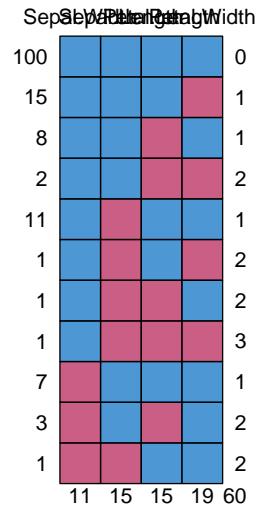
- **Continuous variables** use linear regression.
- **Categorical Variables** use logistic regression.

Methods in MICE:

- PMM (Predictive Mean Matching) – For numeric variables
- logreg(Logistic Regression) – For Binary Variables(with 2 levels)
- polyreg(Bayesian polytomous regression) – For Factor Variables (≥ 2 levels)
- Proportional odds model (ordered, ≥ 2 levels)

```
# load package
library(mice)
library(VIM)

# check missing values
md.pattern(iris.mis)
```

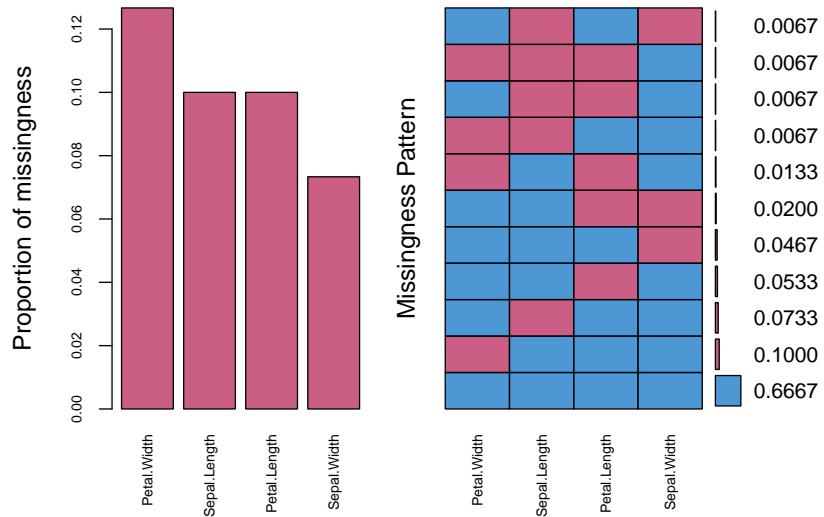


```

#>      Sepal.Width Sepal.Length Petal.Length Petal.Width
#> 100          1           1           1          1  0
#> 15           1           1           1          0  1
#> 8            1           1           0          1  1
#> 2            1           1           0          0  2
#> 11          1           0           1          1  1
#> 1            1           0           1          0  2
#> 1            1           0           0          1  2
#> 1            1           0           0          0  3
#> 7            0           1           1          1  1
#> 3            0           1           0          1  2
#> 1            0           0           1          1  2
#>           11          15          15          19 60

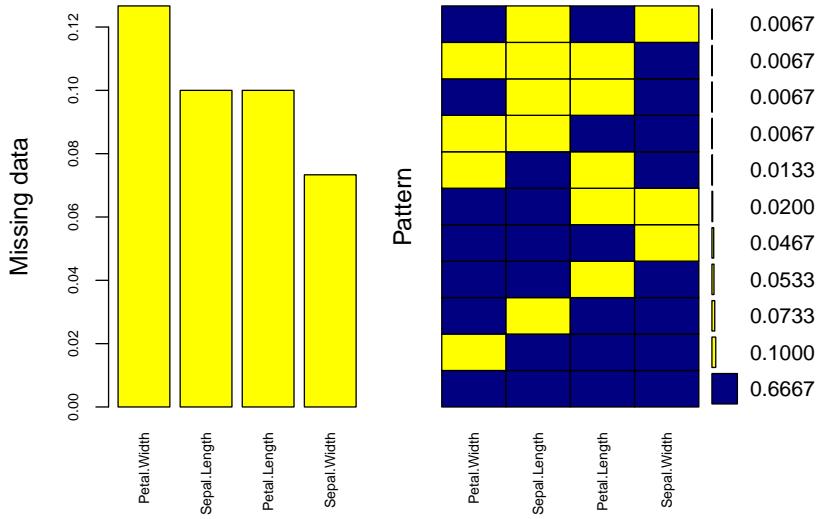
#plot the missing values
aggr(iris.mis, col=mdc(1:2), numbers=TRUE, sortVars=TRUE, labels=names(iris.mis), cex.axis=.7, ga

```



```
#>
#> Variables sorted by number of missings:
#>      Variable      Count
#>      Petal.Width 0.12666667
#>      Sepal.Length 0.10000000
#>      Petal.Length 0.10000000
#>      Sepal.Width 0.07333333

mice_plot <- agrgr(iris.mis, col=c('navyblue','yellow'),
                    numbers=TRUE, sortVars=TRUE,
                    labels=names(iris.mis), cex.axis=.7,
                    gap=3, ylab=c("Missing data","Pattern"))
```



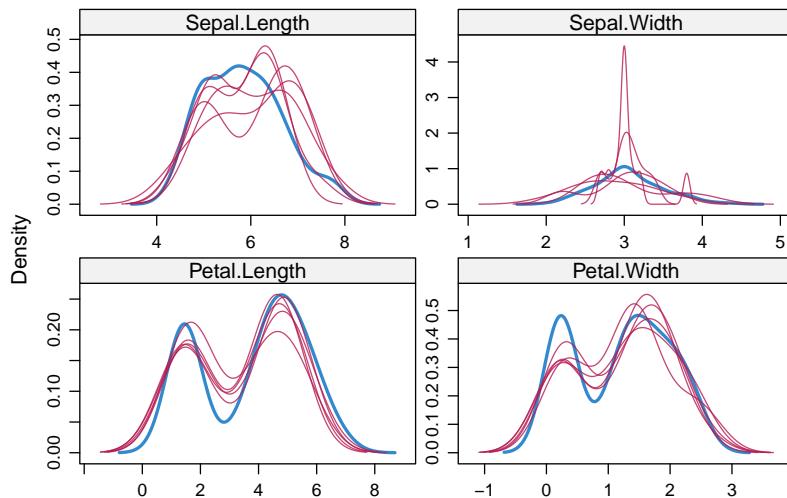
```
#>
#> Variables sorted by number of missings:
#>   Variable      Count
#>   Petal.Width 0.12666667
#>   Sepal.Length 0.10000000
#>   Petal.Length 0.10000000
#>   Sepal.Width 0.07333333
```

Impute Data

```
imputed_Data <-
  mice(
    iris.mis,
    m = 5, # number of imputed datasets
    maxit = 50, # number of iterations taken to impute missing values
    method = 'pmm', # method used in imputation. Here, we used predictive mean matching
    # other methods can be
    # "pmm": Predictive mean matching
    # "midastouch" : weighted predictive mean matching
    # "sample": Random sample from observed values
    # "cart": classification and regression trees
    # "rf": random forest imputations.
    # "2lonly.pmm": Level-2 class predictive mean matching
    # Other methods based on whether variables are (1) numeric, (2) binary, (3) ordered, (4),
    seed = 500
  )
```

```
summary(imputed_Data)
#> Class: mids
#> Number of multiple imputations: 5
#> Imputation methods:
#> Sepal.Length Sepal.Width Petal.Length Petal.Width
#> "pmm"       "pmm"      "pmm"      "pmm"
#> PredictorMatrix:
#> Sepal.Length Sepal.Width Petal.Length Petal.Width
#> Sepal.Length      0          1          1          1
#> Sepal.Width        1          0          1          1
#> Petal.Length       1          1          0          1
#> Petal.Width        1          1          1          0

#make a density plot
densityplot(imputed_Data)
```



```
#the red (imputed values) should be similar to the blue (observed)
```

Check imputed dataset

```
# 1st dataset
completeData <- complete(imputed_Data, 1)

# 2nd dataset
complete(imputed_Data, 2)
```

Regression model using imputed datasets

```
# regression model
fit <- with(data = imputed_Data, exp = lm(Sepal.Width ~ Sepal.Length + Petal.Width))

#combine results of all 5 models
combine <- pool(fit)
summary(combine)

#>           term   estimate  std.error statistic      df     p.value
#> 1  (Intercept) 1.8963130 0.32453912 5.843095 131.0856 3.838556e-08
#> 2 Sepal.Length  0.2974293 0.06679204 4.453066 130.2103 1.802241e-05
#> 3 Petal.Width -0.4811603 0.07376809 -6.522608 108.8253 2.243032e-09
```

11.6.5 Amelia

- Use bootstrap based EMB algorithm (faster and robust to impute many variables including cross sectional, time series data etc)
- Use parallel imputation feature using multicore CPUs.

Assumptions

- All variables follow Multivariate Normal Distribution (MVN). Hence, this package works best when data is MVN, or transformation to normality.
- Missing data is Missing at Random (MAR)

Steps:

1. m bootstrap samples and applies EMB algorithm to each sample. Then we have m different estimates of mean and variances.
2. the first set of estimates are used to impute first set of missing values using regression, then second set of estimates are used for second set and so on.

However, Amelia is different from MICE

- MICE imputes data on variable by variable basis whereas MVN uses a joint modeling approach based on multivariate normal distribution.
- MICE can handle different types of variables while the variables in MVN need to be normally distributed or transformed to approximate normality.
- MICE can manage imputation of variables defined on a subset of data whereas MVN cannot.

```

library(Amelia)
data("iris")
#seed 10% missing values
iris.mis <- prodNA(iris, noNA = 0.1)

# idvars - keep all ID variables and other variables which you don't want to impute
# noms - keep nominal variables here

#specify columns and run amelia
amelia_fit <- amelia(iris.mis, m=5, parallel = "multicore", noms = "Species")
#> -- Imputation 1 --
#>
#>   1  2  3  4  5  6  7  8
#>
#> -- Imputation 2 --
#>
#>   1  2  3  4  5  6  7  8  9 10
#>
#> -- Imputation 3 --
#>
#>   1  2  3  4  5  6
#>
#> -- Imputation 4 --
#>
#>   1  2  3  4  5  6  7  8
#>
#> -- Imputation 5 --
#>
#>   1  2  3  4  5  6  7

# access imputed outputs
# amelia_fit$imputations[[1]]

```

11.6.6 missForest

- an implementation of random forest algorithm (a non parametric imputation method applicable to various variable types). Hence, no assumption about function form of f. Instead, it tries to estimate f such that it can be as close to the data points as possible.
- builds a random forest model for each variable. Then it uses the model to predict missing values in the variable with the help of observed values.
- It yields out of bag imputation error estimate. Moreover, it provides high level of control on imputation process.
- Since bagging works well on categorical variable too, we don't need to

remove them here.

```
library(missForest)
#impute missing values, using all parameters as default values
iris.imp <- missForest(iris.mis)
#> missForest iteration 1 in progress...done!
#> missForest iteration 2 in progress...done!
#> missForest iteration 3 in progress...done!
#> missForest iteration 4 in progress...done!
#> missForest iteration 5 in progress...done!
#> missForest iteration 6 in progress...done!
#> missForest iteration 7 in progress...done!
# check imputed values
# iris.imp$ximp

# check imputation error
# NRMSE is normalized mean squared error. It is used to represent error derived from imputing continuous variables
# PFC (proportion of falsely classified) is used to represent error derived from imputing categorical variables
iris.imp$OOBerror
#>      NRMSE      PFC
#> 0.14004198 0.03053435

#comparing actual data accuracy
iris.err <- mixError(iris.imp$ximp, iris.mis, iris)
iris.err
#>      NRMSE      PFC
#> 0.11567322 0.05263158
```

This means categorical variables are imputed with 5% error and continuous variables are imputed with 14% error.

This can be improved by tuning the values of `mtry` and `ntree` parameter.

- `mtry` refers to the number of variables being randomly sampled at each split.
- `ntree` refers to number of trees to grow in the forest.

11.6.7 Hmisc

- `impute()` function imputes missing value using user defined statistical method (mean, max, median). It's default is median.
- `aregImpute()` allows mean imputation using additive regression, bootstrapping, and predictive mean matching.

1. In bootstrapping, different bootstrap resamples are used for each of multiple imputations. Then, a flexible additive model (non parametric regression method) is fitted on samples taken with replacements from original data and missing values (acts as dependent variable) are predicted using non-missing values (independent variable).
2. it uses predictive mean matching (default) to impute missing values. Predictive mean matching works well for continuous and categorical (binary & multi-level) without the need for computing residuals and maximum likelihood fit.

Note

- For predicting categorical variables, Fisher's optimum scoring method is used.
- **Hmisc** automatically recognizes the variables types and uses bootstrap sample and predictive mean matching to impute missing values.
- **missForest** can outperform **Hmisc** if the observed variables have sufficient information.

Assumption

- linearity in the variables being predicted.

```
library(Hmisc)
# impute with mean value
iris.mis$imputed_age <- with(iris.mis, impute(Sepal.Length, mean))

# impute with random value
iris.mis$imputed_age2 <- with(iris.mis, impute(Sepal.Length, 'random'))

# could also use min, max, median to impute missing value

# using argImpute
impute_arg <- aregImpute(~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width +
Species, data = iris.mis, n.impute = 5) # argImpute() automatically identifies the var
#> Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5
Iteration 6
Iteration 7
Iteration 8
```

```

impute_arg # R-squares are for predicted missing values.
#>
#> Multiple Imputation using Bootstrap and PMM
#>
#> aregImpute(formula = ~Sepal.Length + Sepal.Width + Petal.Length +
#>     Petal.Width + Species, data = iris.mis, n.impute = 5)
#>
#> n: 150   p: 5   Imputations: 5      nk: 3
#>
#> Number of NAs:
#> Sepal.Length Sepal.Width Petal.Length Petal.Width      Species
#>          14         14         13         15         19
#>
#>           type d.f.
#> Sepal.Length    s    2
#> Sepal.Width     s    2
#> Petal.Length    s    2
#> Petal.Width     s    2
#> Species         c    2
#>
#> Transformation of Target Variables Forced to be Linear
#>
#> R-squares for Predicting Non-Missing Values for Each Variable
#> Using Last Imputations of Predictors
#> Sepal.Length Sepal.Width Petal.Length Petal.Width      Species
#>          0.884       0.606       0.983       0.955       0.989

# check imputed variable Sepal.Length
impute_arg$imputed$Sepal.Length
#> [,1] [,2] [,3] [,4] [,5]
#> 3  5.0  5.0  4.7  4.8  5.4
#> 25 5.4  5.0  4.8  5.1  5.0
#> 49  5.1  5.1  5.0  5.1  5.0
#> 56  6.5  5.8  6.0  6.9  6.3
#> 60  6.1  6.0  5.5  5.6  5.7
#> 81  5.5  5.7  5.6  5.2  5.7
#> 83  5.7  5.5  5.7  5.6  5.5
#> 104 6.4  6.5  6.7  6.4  6.7
#> 108 6.3  7.2  7.7  7.7  7.7
#> 121 7.2  6.3  6.3  6.4  6.5
#> 127 5.6  6.3  6.0  6.3  6.3
#> 133 6.8  6.7  5.9  6.4  6.9
#> 148 5.7  5.8  6.8  6.4  6.1
#> 150 6.1  6.4  6.7  6.3  5.9

```

11.6.8 mi

1. allows graphical diagnostics of imputation models and convergence of imputation process.
2. uses Bayesian version of regression models to handle issue of separation.
3. automatically detects irregularities in data (e.g., high collinearity among variables).
4. adds noise to imputation process to solve the problem of additive constraints.

```
library(mi)
# default values of parameters
# 1. rand.imp.method as "bootstrap"
# 2. n.imp (number of multiple imputations) as 3
# 3. n.iter ( number of iterations) as 30
mi_data <- mi(iris.mis, seed = 335)
summary(mi_data)
```

Chapter 12

Data

There are multiple ways to categorize data. For example,

- Qualitative vs. Quantitative:

Qualitative	Quantitative
in-depth interviews, documents, focus groups, case study, ethnography.	experiments, observation in words, survey with closed-end questions, structured interviews
open-ended questions. observations in words	
language, descriptive	quantities, numbers
Text-based	Numbers-based
Subjective	Objectivity

12.1 Cross-Sectional

12.2 Time Series

$$y_t = \beta_0 + x_{t1}\beta_1 + x_{t2}\beta_2 + \dots + x_{t(k-1)}\beta_{k-1} + \epsilon_t$$

Examples

- Static Model

$$- y_t = \beta_0 + x_1\beta_1 + x_2\beta_2 - x_3\beta_3 - \epsilon_t$$

- Finite Distributed Lag model

$$\begin{aligned} & - y_t = \beta_0 + p e_t \delta_0 + p e_{t-1} \delta_1 + p e_{t-2} \delta_2 + \epsilon_t \\ & - \textbf{Long Run Propensity (LRP)} \text{ is } LRP = \delta_0 + \delta_1 + \delta_2 \end{aligned}$$

- Dynamic Model

$$- GDP_t = \beta_0 + \beta_1 GDP_{t-1} - \epsilon_t$$

Finite Sample Properties for Time Series:

- A1-A3: OLS is unbiased
- A1-A4: usual standard errors are consistent and Gauss-Markov Theorem holds (OLS is BLUE)
- A1-A6, A6: Finite Sample Wald Test (t-test and F-test) are valid

A3 might not hold under time series setting

- Spurious Time Trend - solvable
- Strict vs Contemporaneous Exogeneity - not solvable

In time series data, there are many processes:

- Autoregressive model of order p: AR(p)
- Moving average model of order q: MA(q)
- Autoregressive model of order p and moving average model of order q: ARMA(p,q)
- Autoregressive conditional heteroskedasticity model of order p: ARCH(p)
- Generalized Autoregressive conditional heteroskedasticity of orders p and q; GARCH(p,q)

12.2.1 Deterministic Time trend

Both the dependent and independent variables are trending over time

Spurious Time Series Regression

$$y_t = \alpha_0 + t\alpha_1 + v_t$$

and x takes the form

$$x_t = \lambda_0 + t\lambda_1 + u_t$$

- $\alpha_1 \neq 0$ and $\lambda_1 \neq 0$

- v_t and u_t are independent
- there is no relationship between y_t and x_t

If we estimate the regression,

$$y_t = \beta_0 + x_t\beta_1 + \epsilon_t$$

so the true $\beta_1 = 0$

- Inconsistent: $\text{plim}(\hat{\beta}_1) = \frac{\alpha_1}{\lambda_1}$
- Invalid Inference: $|t| \rightarrow^d \infty$ for $H_0 : \beta_1 = 0$, will always reject the null as $n \rightarrow \infty$
- Uninformative R^2 : $\text{plim}(R^2) = 1$ will be able to perfectly predict as $n \rightarrow \infty$

We can rewrite the equation as

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t \epsilon_t = \alpha_1 t + v_t$$

where $\beta_0 = \alpha_0$ and $\beta_1 = 0$. Since x_t is a deterministic function of time, ϵ_t is correlated with x_t and we have the usual omitted variable bias.

Even when y_t and x_t are related ($\beta_1 \neq 0$) but they are both trending over time, we still get spurious results with the simple regression on y_t on x_t

Solutions to Spurious Trend

1. Include time trend t as an additional control
 - consistent parameter estimates and valid inference
2. Detrend both dependent and independent variables and then regress the detrended outcome on detrended independent variables (i.e., regress residuals \hat{u}_t on residuals \hat{v}_t)
 - Detrending is the same as partialling out in the Frisch-Waugh-Lovell Theorem
 - Could allow for non-linear time trends by including t , t^2 , and $\exp(t)$
 - Allow for seasonality by including indicators for relevant “seasons” (quarters, months, weeks).

A3 does not hold under:

- Feedback Effect

- ϵ_t influences next period's independent variables
- Dynamic Specification
 - include last time period outcome as an explanatory variable
- Dynamically Complete
 - For finite distributed lag model, the number of lags needs to be absolutely correct.

12.2.2 Feedback Effect

$$y_t = \beta_0 + x_t\beta_1 + \epsilon_t$$

A3

$$E(\epsilon_t | \mathbf{X}) = E(\epsilon_t | x_1, x_2, \dots, x_t, x_{t+1}, \dots, x_T)$$

will not equal 0, because y_t will likely influence x_{t+1}, \dots, x_T

- A3 is violated because we require the error to be uncorrelated with all time observation of the independent regressors (**strict exogeneity**)

12.2.3 Dynamic Specification

$$y_t = \beta_0 + y_{t-1}\beta_1 + \epsilon_t$$

$$E(\epsilon_t | \mathbf{X}) = E(\epsilon_t | y_1, y_2, \dots, y_t, y_{t+1}, \dots, y_T)$$

will not equal 0, because y_t and ϵ_t are inherently correlated

- A3 is violated because we require the error to be uncorrelated with all time observation of the independent regressors (**strict exogeneity**)
- Dynamic Specification is not allowed under A3

12.2.4 Dynamically Complete

$$y_t = \beta_0 + x_t \delta_0 + x_{t-1} \delta_1 + \epsilon_t$$

$$E(\epsilon_t | \mathbf{X}) = E(\epsilon_t | x_1, x_2, \dots, x_t, x_{t+1}, \dots, x_T)$$

will not equal 0, because if we did not include enough lags, x_{t-2} and ϵ_t are correlated

- A3 is violated because we require the error to be uncorrelated with all time observation of the independent regressors (strict exogeneity)
- Can be corrected by including more lags (but when stop?)

Without A3

- OLS is biased
- Gauss-Markov Theorem
- Finite Sample Properties are invalid

then, we can

- Focus on Large Sample Properties
- Can use A3a instead of A3

A3a in time series become

$$A3a : E(\mathbf{x}'_t \epsilon_t) = 0$$

only the regressors in this time period need to be independent from the error in this time period (**Contemporaneous Exogeneity**)

- ϵ_t can be correlated with ..., $x_{t-2}, x_{t-1}, x_{t+1}, x_{t+2}, \dots$
- can have a dynamic specification $y_t = \beta_0 + y_{t-1} \beta_1 + \epsilon_t$

Deriving Large Sample Properties for Time Series

- Assumptions A1, A2, A3a
- Weak Law and Central Limit Theorem depend on A5
 - x_t and ϵ_t are dependent over t
 - without Weak Law or Central Limit Theorem depend on A5, we cannot have Large Sample Properties for OLS

- Instead of A5, we consider A5a
- Derivation of the Asymptotic Variance depends on A4
 - time series setting introduces **Serial Correlation**: $Cov(\epsilon_t, \epsilon_s) \neq 0$

under A1, A2, A3a, and A5a, OLS estimator is **consistent**, and **asymptotically normal**

12.2.5 Highly Persistent Data

If y_t, \mathbf{x}_t are not weakly dependent stationary process

* y_t and y_{t-h} are not almost independent for large h * A5a does not hold and OLS is not **consistent** and does not have a limiting distribution. * Example + Random Walk $y_t = y_{t-1} + u_t$ + Random Walk with a drift: $y_t = \alpha + y_{t-1} + u_t$

Solution First difference is a stationary process

$$y_t - y_{t-1} = u_t$$

- If u_t is a weakly dependent process (also called integrated of order 0) then y_t is said to be difference-stationary process (integrated of order 1)
- For regression, if $\{y_t, \mathbf{x}_t\}$ are random walks (integrated at order 1), can consistently estimate the first difference equation

$$\begin{aligned} y_t - y_{t-1} &= (\mathbf{x}_t - \mathbf{x}_{t-1}\beta + \epsilon_t - \epsilon_{t-1}) \\ \Delta y_t &= \Delta \mathbf{x}\beta + \Delta u_t \end{aligned}$$

Unit Root Test

$$y_t = \alpha + \alpha y_{t-1} + u_t$$

tests if $\rho = 1$ (integrated of order 1)

- Under the null $H_0 : \rho = 1$, OLS is not consistent or asymptotically normal.
- Under the alternative $H_a : \rho < 1$, OLS is consistent and asymptotically normal.
- usual t-test is not valid, will need to use the transformed equation to produce a valid test.

Dickey-Fuller Test

$$\Delta y_t = \alpha + \theta y_{t-1} + v_t$$

where $\theta = \rho - 1$

- $H_0 : \theta = 0$ and $H_a : \theta < 0$
- Under the null, Δy_t is weakly dependent but y_{t-1} is not.
- Dickey and Fuller derived the non-normal asymptotic distribution. If you reject the null then y_t is not a random walk.

Concerns with the standard Dickey Fuller Test

1. Only considers a fairly simplistic dynamic relationship

$$\Delta y_t = \alpha + \theta y_{t-1} + \gamma_1 \Delta_{t-1} + \dots + \gamma_p \Delta_{t-p} + v_t$$

- with one additional lag, under the null Δy_t is an AR(1) process and under the alternative y_t is an AR(2) process.
- Solution: include lags of Δy_t as controls.

2. Does not allow for time trend

$$\Delta y_t = \alpha + \theta y_{t-1} + \delta t + v_t$$

- allows y_t to have a quadratic relationship with t
- Solution: include time trend (changes the critical values).

Adjusted Dickey-Fuller Test

$$\Delta y_t = \alpha + \theta y_{t-1} + \delta t + \gamma_1 \Delta y_{t-1} + \dots + \gamma_p \Delta y_{t-p} + v_t$$

where $\theta = 1 - \rho$

- $H_0 : \theta_1 = 0$ and $H_a : \theta_1 < 0$
- Under the null, Δy_t is weakly dependent but y_{t-1} is not
- Critical values are different with the time trend, if you reject the null then y_t is not a random walk.

12.2.5.0.1 Newey West Standard Errors If A4 does not hold, we can use Newey West Standard Errors (HAC - Heteroskedasticity Autocorrelation Consistent)

$$\hat{B} = T^{-1} \sum_{t=1}^T e_t^2 \mathbf{x}'_t \mathbf{x}_t + \sum_{h=1}^g \left(1 - \frac{h}{g+1}\right) T^{-1} \sum_{t=h+1}^T e_t e_{t-h} (\mathbf{x}'_t \mathbf{x}_{t-h} + \mathbf{x}'_{t-h} \mathbf{x}_t)$$

- estimates the covariances up to a distance g part
- downweights to insure \hat{B} is PSD
- How to choose g :
 - For yearly data: $g = 1$ or 2 is likely to account for most of the correlation
 - For quarterly or monthly data: g should be larger ($g = 4$ or 8 for quarterly and $g = 12$ or 14 for monthly)
 - can also take integer part of $4(T/100)^{2/9}$ or integer part of $T^{1/4}$

Testing for Serial Correlation

1. Run OLS regression of y_t on \mathbf{x}_t and obtain residuals e_t
2. Run OLS regression of e_t on \mathbf{x}_t, e_{t-1} and test whether coefficient on e_{t-1} is significant.
3. Reject the null of no serial correlation if the coefficient is significant at the 5% level.
 - Test using heteroskedastic robust standard errors
 - can include e_{t-2}, e_{t-3}, \dots in step 2 to test for higher order serial correlation (t-test would now be an F-test of joint significance)

12.3 Repeated Cross Sections

For each time point (day, month, year, etc.), a set of data is sampled. This set of data can be different among different time points.

For example, you can sample different groups of students each time you survey.

Allowing structural change in pooled cross section

$$y_i = \mathbf{x}_i\beta + \delta_1 y_1 + \dots + \delta_T y_T + \epsilon_i$$

Dummy variables for all but one time period

- allows different intercept for each time period
- allows outcome to change on average for each time period

Allowing for structural change in pooled cross section

$$y_i = \mathbf{x}_i\beta + \mathbf{x}_i y_1 \gamma_1 + \dots + \mathbf{x}_i y_T \gamma_T + \delta_1 y_1 + \dots + \delta_T y_T + \epsilon_i$$

Interact x_i with time period dummy variables

- allows different slopes for each time period
- allows effects to change based on time period (**structural break**)
- Interacting all time period dummies with x_i can produce many variables
- use hypothesis testing to determine which structural breaks are needed.

12.3.1 Pooled Cross Section

$$y_i = \mathbf{x}_i + \mathbf{x}_i \times \mathbf{y}_1 \mathbf{1}_1 + \dots + \mathbf{x}_i \times \mathbf{y}_T \mathbf{1}_T + \mathbf{y}_1 \mathbf{1}_1 + \dots + \mathbf{y}_T \mathbf{1}_T + \epsilon_i$$

Interact x_i with time period dummy variables

- allows different slopes for each time period
- allows effect to change based on time period (structural break)
 - interacting all time period dummies with x_i can produce many variables
- use hypothesis testing to determine which structural breaks are needed.

12.4 Panel Data

Detail notes in R can be found here

Follows an individual over T time periods.

Panel data structure is like having n samples of time series data

Characteristics

- Information both across individuals and over time (cross-sectional and time-series)
- N individuals and T time periods
- Data can be either
 - Balanced: all individuals are observed in all time periods
 - Unbalanced: all individuals are not observed in all time periods.
- Assume correlation (clustering) over time for a given individual, with independence over individuals.

Types

- Short panel: many individuals and few time periods.

- Long panel: many time periods and few individuals
- Both: many time periods and many individuals

Time Trends and Time Effects

- Nonlinear
- Seasonality
- Discontinuous shocks

Regressors

- Time-invariant regressors $x_{it} = x_i$ for all t (e.g., gender, race, education) have zero within variation
- Individual-invariant regressors $x_{it} = x_t$ for all i (e.g., time trend, economy trends) have zero between variation

Variation for the dependent variable and regressors

- Overall variation: variation over time and individuals.
- Between variation: variation between individuals
- Within variation: variation within individuals (over time).

Estimate	Formula
Individual mean	$\bar{x}_i = \frac{1}{T} \sum_t x_{it}$
Overall mean	$\bar{x} = \frac{1}{NT} \sum_i \sum_t x_{it}$
Overall Variance	$s_O^2 = \frac{1}{NT-1} \sum_i \sum_t (x_{it} - \bar{x})^2$
Between variance	$s_B^2 = \frac{1}{N-1} \sum_i (\bar{x}_i - \bar{x})^2$
Within variance	$s_W^2 = \frac{1}{NT-1} \sum_i \sum_t (x_{it} - \bar{x}_i)^2 = \frac{1}{NT-1} \sum_i \sum_t (x_{it} - \bar{x}_i + \bar{x})^2$

Note: $s_O^2 \approx s_B^2 + s_W^2$

Since we have n observation for each time period t, we can control for each time effect separately by including time dummies (time effects)

$$y_{it} = \mathbf{x}_{it} + d_1\delta_1 + \dots + d_{T-1}\delta_{T-1} + \epsilon_{it}$$

Note: we cannot use these many time dummies in time series data because in time series data, our n is 1. Hence, there is no variation, and sometimes not enough data compared to variables to estimate coefficients.

Unobserved Effects Model Similar to group clustering, assume that there is a random effect that captures differences across individuals but is constant in time.

$$y_{it} = \mathbf{x}'_{it} + d_1\delta_1 + \dots + d_{T-1}\delta_{T-1} + c_i + u_{it}$$

where

- $c_i + u_{it} = \epsilon_{it}$
- c_i unobserved individual heterogeneity (effect)
- u_{it} idiosyncratic shock
- ϵ_{it} unobserved error term.

12.4.1 Pooled OLS Estimator

If c_i is uncorrelated with x_{it}

$$E(\mathbf{x}'_{it}(c_i + u_{it})) = 0$$

then A3a still holds. And we have Pooled OLS consistent.

If A4 does not hold, OLS is still consistent, but not efficient, and we need cluster robust SE.

Sufficient for A3a to hold, we need

- **Exogeneity** for u_{it} A3a (contemporaneous exogeneity): $E(\mathbf{x}'_{it}u_{it}) = 0$ time varying error
- **Random Effect Assumption** (time constant error): $E(\mathbf{x}'_{it}c_i) = 0$

Pooled OLS will give you consistent coefficient estimates under A1, A2, A3a (for both u_{it} and RE assumption), and A5 (randomly sampling across i).

12.4.2 Individual-specific effects model

- If we believe that there is unobserved heterogeneity across individual (e.g., unobserved ability of an individual affects y), If the individual-specific effects are correlated with the regressors, then we have the Fixed Effects Estimator. and if they are not correlated we have the Random Effects Estimator.

12.4.2.1 Random Effects Estimator

Random Effects estimator is the Feasible GLS estimator that assumes u_{it} is serially uncorrelated and homoskedastic

- Under A1, A2, A3a (for both u_{it} and RE assumption) and A5 (randomly sampling across i), RE estimator is consistent.
 - If A4 holds for u_{it} , RE is the most efficient estimator
 - If A4 fails to hold (may be heteroskedasticity across i, and serial correlation over t), then RE is not the most efficient, but still more efficient than pooled OLS.

12.4.2.2 Fixed Effects Estimator

also known as **Within Estimator** uses within variation (over time)

If the **RE assumption** is not hold ($E(\mathbf{x}'_{it} c_i) \neq 0$), then A3a does not hold ($E(\mathbf{x}'_{it} \epsilon_i) \neq 0$).

Hence, the OLS and RE are inconsistent/biased (because of omitted variable bias)

However, FE can only fix bias due to time-invariant factors (both observables and unobservables) correlated with treatment (not time-variant factors that correlated with the treatment).

The traditional FE technique is flawed when lagged dependent variables are included in the model. (Nickell, 1981) (Narayanan and Nair, 2013)

With measurement error in the independent, FE will exacerbate the errors-in-the-variables bias.

12.4.2.2.1 Demean Approach

To deal with violation in c_i , we have

$$y_{it} = \mathbf{x}_{it} + c_i + u_{it}$$

$$\bar{y}_i = \bar{\mathbf{x}}_i \beta + c_i + \bar{u}_i$$

where the second equation is the time averaged equation

using **within transformation**, we have

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i) \beta + u_{it} - \bar{u}_i$$

because c_i is time constant.

The Fixed Effects estimator uses POLS on the transformed equation

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\beta + d_1\delta_1 + \dots + d_{T-2}\delta_{T-2} + u_{it} - \bar{u}_i$$

- we need A3 (strict exogeneity) ($E((\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'(u_{it} - \bar{u}_i)) = 0$) to have FE consistent.
- Variables that are time constant will be absorbed into c_i . Hence we cannot make inference on time constant independent variables.
 - If you are interested in the effects of time-invariant variables, you could consider the OLS or **between estimator**
- It's recommended that you should still use cluster robust standard errors.

12.4.2.2.2 Dummy Approach Equivalent to the within transformation (i.e., mathematically equivalent to Demean Approach), we can have the fixed effect estimator be the same with the dummy regression

$$y_{it} = x_{it}\beta + d_1\delta_1 + \dots + d_{T-2}\delta_{T-2} + c_1\gamma_1 + \dots + c_{n-1}\gamma_{n-1} + u_{it}$$

where

$$c_i = \begin{cases} 1 & \text{if observation is } i \\ 0 & \text{otherwise} \end{cases} \quad (12.1)$$

- The standard error is incorrectly calculated.
- the FE within transformation is controlling for any difference across individual which is allowed to correlated with observables.

12.4.2.2.3 First-difference Approach Economists typically use this approach

$$y_{it} - y_{i(t-1)} = (\mathbf{x}_{it} - \mathbf{x}_{i(t-1)})\beta + (u_{it} - u_{i(t-1)})$$

12.4.2.2.4 Fixed Effects Summary

- The three approaches are **almost** equivalent.
 - Demean Approach is mathematically equivalent to Dummy Approach
 - If you have only 1 period, all 3 are the same.

- Since fixed effect is a within estimator, only **status changes** can contribute to β variation.
 - Hence, with a small number of changes then the standard error for β will explode
- Status changes mean subjects change from (1) control to treatment group or (2) treatment to control group. Those who have status change, we call them **switchers**. (more on this in
 - Treatment effect is typically **non-directional**.
 - You can give a parameter for the direction if needed.
- Issues:
 - You could have fundamental difference between switchers and non-switchers. Even though we can't definitive test this, but providing descriptive statistics on switchers and non-switchers can give us confidence in our conclusion.
 - Because fixed effects focus on bias reduction, you might have larger variance (typically, with fixed effects you will have less df)
- If the true model is random effect, economists typically don't care, especially when c_i is the random effect and $c_i \perp x_{it}$ (because RE assumption is that it is unrelated to x_{it}). The reason why economists don't care is because RE wouldn't correct bias, it only improves efficiency over OLS.
- You can estimate FE for different units (not just individuals).
- FE removes bias from time invariant factors but not without costs because it uses within variation, which imposes strict exogeneity assumption on u_{it} :

$$E[(x_{it} - \bar{x}_i)(u_{it} - \bar{u}_{it})] = 0$$

Recall

$$Y_{it} = \beta_0 + X_{it}\beta_1 + \alpha_i + u_{it}$$

where $\epsilon_{it} = \alpha_i + u_{it}$

$$\hat{\sigma}_\epsilon^2 = \frac{SSR_{OLS}}{NT - K}$$

$$\hat{\sigma}_u^2 = \frac{SSR_{FE}}{NT - (N + K)} = \frac{SSR_{FE}}{N(T - 1) - K}$$

It's ambiguous whether your variance of error changes up or down because SSR can increase while the denominator decreases.

FE can be unbiased, but not consistent (i.e., not converging to the true effect)

12.4.2.2.5 FE Examples

12.4.2.2.6 Blau (1999)

- Intergenerational mobility
- If we transfer resources to low income family, can we generate upward mobility (increase ability)?

Mechanisms for intergenerational mobility

1. Genetic (policy can't affect) (i.e., ability endowment)
2. Environmental indirect
3. Environmental direct

$$\frac{\% \Delta \text{Human capital}}{\% \Delta \text{income}}$$

4. Financial transfer

Income measures:

1. Total household income
2. Wage income
3. Non-wage income
4. Annual versus permanent income

Core control variables:

Bad controls are those jointly determined with dependent variable

Control by mother = choice by mother

Uncontrolled by mothers:

- mother race
- location of birth
- education of parents
- household structure at age 14

$$Y_{ijt} = X_{jt}\beta_i + I_{jt}\alpha_i + \epsilon_{ijt}$$

where

- $i = \text{test}$
- $j = \text{individual (child)}$
- $t = \text{time}$

Grandmother's model

Since child is nested within mother and mother nested within grandmother, the fixed effect of child is included in the fixed effect of mother, which is included in the fixed-effect of grandmother

$$Y_{ijgmt} = X_{it}\beta_i + I_{jt}\alpha_i + \gamma_g + u_{ijgmt}$$

where

- i = test, j = kid, m = mother, g = grandmother
- where γ_g includes γ_m includes γ_j

Grandma fixed-effect

Pros:

- control for some genetics + fixed characteristics of how mother are raised
- can estimate effect of parameter income

Con:

- Might not be a sufficient control

Common to cluster at the fixed-effect level (common correlated component)

Fixed effect exaggerates attenuation bias

Error rate on survey can help you fix this (plug in the number only , but not the uncertainty associated with that number).

12.4.2.2.7 BABCOCK (2010)

$$T_{ijct} = \alpha_0 + S_{jct}\alpha_1 + X_{ijct}\alpha_2 + u_{ijct}$$

where

- S_{jct} is the average class expectation

- $X_{ijct}\alpha_2$ is the individual characteristics
- i student
- j instructor
- c course
- t time

$$T_{ijct} = \beta_0 + S_{jct}\beta_1 + X_{ijct}\beta_2 + \mu_{jc} + \epsilon_{ijct}$$

where μ_{jc} is instructor by course fixed effect (unique id), which is different from $(\theta_j + \delta_c)$

1. Decrease course shopping because conditioned on available information (μ_{ja}) (class grade and instructor's info).
2. Grade expectation change even though class materials stay the same

Identification strategy is

- Under (fixed) time-varying factor that could bias my coefficient (simultaneity)

$$Y_{ijt} = X_{it}\beta_1 + \text{Teacher Experience}_{jt}\beta_2 + \text{Teacher education}_{jt}\beta_3 + \text{Teacher score}_{it}\beta_4 + \dots + \epsilon_{ijt}$$

Drop teacher characteristics, and include teacher dummy effect

$$Y_{ijt} = X_{it}\alpha + \Gamma_{it}\theta_j + u_{ijt}$$

where α is the within teacher (conditional on teacher fixed effect) and $j = 1 \rightarrow (J - 1)$

Nuisance in the sense that we don't about the interpretation of α

The least we can say about θ_j is the teacher effect conditional on student test score.

$$Y_{ijt} = X_{it}\gamma + \epsilon_{ijt}$$

γ is between within (unconditional) and e_{ijt} is the prediction error

$$e_{ijt} = T_{it}\delta_j + \tilde{e}_{ijt}$$

where δ_j is the mean for each group

$$Y_{ijkt} = Y_{ijkt-1} + X_{it}\beta + T_{it}\tau_j + (W_i + P_k + \epsilon_{ijkt})$$

where

- Y_{ijkt-1} = lag control
- τ_j = teacher fixed time
- W_i is the student fixed effect
- P_k is the school fixed effect
- $u_{ijkt} = W_i + P_k + \epsilon_{ijkt}$

And we worry about selection on class and school

Bias in τ (for 1 teacher) is

$$\frac{1}{N_j} \sum_{i=1}^{N_j} (W_i + P_k + \epsilon_{ijkt})$$

where N_j = the number of student in class with teacher j

then we can $P_k + \frac{1}{N_j} \sum_{i=1}^{N_j} (W_i + \epsilon_{ijkt})$

Shocks from small class can bias τ

$$\frac{1}{N_j} \sum_{i=1}^{N_j} \epsilon_{ijkt} \neq 0$$

which will inflate the teacher fixed effect

Even if we create random teacher fixed effect and put it in the model, it still contains bias mentioned above which can still affect τ (but we do not know the way it will affect - whether more positive or negative).

If teachers switch schools, then we can estimate both teacher and school fixed effect (**mobility web** thin vs. thick)

Mobility web refers to the web of switchers (i.e., from one status to another).

$$Y_{ijkt} = Y_{ijk(t-1)}\alpha + X_{it}\beta + T_{it}\tau + P_k + \epsilon_{ijkt}$$

If we demean (fixed-effect), τ (teacher fixed effect) will go away

If you want to examine teacher fixed effect, we have to include teacher fixed effect

Control for school, the article argues that there is no selection bias

For $\frac{1}{N_j} \sum_{i=1}^{N_j} \epsilon_{ijk}^t$ (teacher-level average residuals), $var(\tau)$ does not change with N_j (Figure 2 in the paper). In words, the quality of teachers is not a function of the number of students

If $var(\tau) = 0$ it means that teacher quality does not matter

Spin-off of Measurement Error: Sampling error or estimation error

$$\hat{\tau}_j = \tau_j + \lambda_j$$

$$var(\hat{\tau}) = var(\tau + \lambda)$$

Assume $cov(\tau_j, \lambda_j) = 0$ (reasonable) In words, your randomness in getting children does not correlation with teacher quality.

Hence,

$$\begin{aligned} var(\hat{\tau}) &= var(\tau) + var(\lambda) \\ var(\tau) &= var(\hat{\tau}) - var(\lambda) \end{aligned}$$

We have $var(\hat{\tau})$ and we need to estimate $var(\lambda)$

$$var(\lambda) = \frac{1}{J} \sum_{j=1}^J \hat{\sigma}_j^2$$

where $\hat{\sigma}_j^2$ is the squared standard error of the teacher j (a function of n)

Hence,

$$\frac{var(\tau)}{var(\hat{\tau})} = \text{reliability} = \text{true variance signal}$$

also known as how much noise in $\hat{\tau}$ and

$$1 - \frac{var(\tau)}{var(\hat{\tau})} = \text{noise}$$

Even in cases where the true relationship is that τ is a function of N_j , then our recovery method for λ is still not affected

To examine our assumption

$$\hat{\tau}_j = \beta_0 + X_j\beta_1 + \epsilon_j$$

Regressing teacher fixed-effect on teacher characteristics should give us R^2 close to 0, because teacher characteristics cannot predict sampling error ($\hat{\tau}$ contain sampling error)

12.4.3 Tests for Assumptions

We typically don't test heteroskedasticity because we will use robust covariance matrix estimation anyway.

Dataset

```
library("plm")
data("EmplUK", package="plm")
data("Produc", package="plm")
data("Grunfeld", package="plm")
data("Wages", package="plm")
```

12.4.3.1 Poolability

also known as an F test of stability (or Chow test) for the coefficients

H_0 : All individuals have the same coefficients (i.e., equal coefficients for all individuals).

H_a Different individuals have different coefficients.

Notes:

- Under a within (i.e., fixed) model, different intercepts for each individual are assumed
- Under random model, same intercept is assumed

```
library(plm)
plm::pooltest(inv~value+capital, data=Grunfeld, model="within")
#>
#> F statistic
#>
#> data: inv ~ value + capital
#> F = 5.7805, df1 = 18, df2 = 170, p-value = 1.219e-10
#> alternative hypothesis: instability
```

Hence, we reject the null hypothesis that coefficients are stable. Then, we should use the random model.

12.4.3.2 Individual and time effects

use the Lagrange multiplier test to test the presence of individual or time or both (i.e., individual and time).

Types:

- **honda**: (Honda, 1985) Default
- **bp**: (Breusch and Pagan, 1980) for unbalanced panels
- **kw**: (King and Wu, 1997) unbalanced panels, and two-way effects
- **ghm**: (Gourieroux et al., 1982): two-way effects

```
pFtest(inv~value+capital, data=Grunfeld, effect="twoways")
#>
#> F test for twoways effects
#>
#> data: inv ~ value + capital
#> F = 17.403, df1 = 28, df2 = 169, p-value < 2.2e-16
#> alternative hypothesis: significant effects
pFtest(inv~value+capital, data=Grunfeld, effect="individual")
#>
#> F test for individual effects
#>
#> data: inv ~ value + capital
#> F = 49.177, df1 = 9, df2 = 188, p-value < 2.2e-16
#> alternative hypothesis: significant effects
pFtest(inv~value+capital, data=Grunfeld, effect="time")
#>
#> F test for time effects
#>
#> data: inv ~ value + capital
#> F = 0.23451, df1 = 19, df2 = 178, p-value = 0.9997
#> alternative hypothesis: significant effects
```

12.4.3.3 Cross-sectional dependence/contemporaneous correlation

- Null hypothesis: residuals across entities are not correlated.

```
pcdtest(inv~value+capital, data=Grunfeld, model="within")
#>
#> Pesaran CD test for cross-sectional dependence in panels
#>
```

```
#> data: inv ~ value + capital
#> z = 4.6612, p-value = 3.144e-06
#> alternative hypothesis: cross-sectional dependence
```

12.4.3.3.1 Global cross-sectional dependence

12.4.3.3.2 Local cross-sectional dependence use the same command, but supply matrix w to the argument.

```
pcdtest(inv~value+capital, data=Grunfeld, model="within")
#>
#> Pesaran CD test for cross-sectional dependence in panels
#>
#> data: inv ~ value + capital
#> z = 4.6612, p-value = 3.144e-06
#> alternative hypothesis: cross-sectional dependence
```

12.4.3.4 Serial Correlation

- Null hypothesis: there is no serial correlation
- usually seen in macro panels with long time series (large N and T), not seen in micro panels (small T and large N)
- Serial correlation can arise from individual effects(i.e., time-invariant error component), or idiosyncratic error terms (e.g, in the case of AR(1) process). But typically, when we refer to serial correlation, we refer to the second one.
- Can be
 - **marginal** test: only 1 of the two above dependence (but can be biased towards rejection)
 - **joint** test: both dependencies (but don't know which one is causing the problem)
 - **conditional** test: assume you correctly specify one dependence structure, test whether the other departure is present.

12.4.3.4.1 Unobserved effect test

- semi-parametric test (the test statistic $W \sim N$ regardless of the distribution of the errors) with $H_0 : \sigma_\mu^2 = 0$ (i.e., no unobserved effects in the residuals), favors pooled OLS.

- Under the null, covariance matrix of the residuals = its diagonal (off-diagonal = 0)
- It is robust against both **unobserved effects** that are constant within every group, and any kind of **serial correlation**.

```
pwtest(log(gsp)~log(pcap)+log(pc)+log(emp)+unemp, data=Produc)
#>
#> Wooldridge's test for unobserved individual effects
#>
#> data: formula
#> z = 3.9383, p-value = 8.207e-05
#> alternative hypothesis: unobserved effect
```

Here, we reject the null hypothesis that the no unobserved effects in the residuals. Hence, we will exclude using pooled OLS.

12.4.3.4.2 Locally robust tests for random effects and serial correlation

- A joint LM test for **random effects** and **serial correlation** assuming normality and homoskedasticity of the idiosyncratic errors [Baltagi and Li (1991)][Baltagi and Li, 1995]

```
pbsytest(log(gsp)~log(pcap)+log(pc)+log(emp)+unemp, data=Produc, test="j")
#>
#> Baltagi and Li AR-RE joint test - balanced panel
#>
#> data: formula
#> chisq = 4187.6, df = 2, p-value < 2.2e-16
#> alternative hypothesis: AR(1) errors or random effects
```

Here, we reject the null hypothesis that there is no presence of **serial correlation**, and **random effects**. But we still do not know whether it is because of serial correlation, of random effects or of both

To know the departure from the null assumption, we can use (Bera et al., 2001)'s test for first-order serial correlation or random effects (both under normality and homoskedasticity assumption of the error).

BSY for serial correlation

```
pbsytest(log(gsp)~log(pcap)+log(pc)+log(emp)+unemp, data=Produc)
#>
#> Bera, Sosa-Escudero and Yoon locally robust test - balanced panel
```

```
#>
#> data: formula
#> chisq = 52.636, df = 1, p-value = 4.015e-13
#> alternative hypothesis: AR(1) errors sub random effects
```

BSY for random effects

```
pbsytest(log(gsp)~log(pcap)+log(pc)+log(emp)+unemp, data=Produc, test="re")
#>
#> Bera, Sosa-Escudero and Yoon locally robust test (one-sided) -
#> balanced panel
#>
#> data: formula
#> z = 57.914, p-value < 2.2e-16
#> alternative hypothesis: random effects sub AR(1) errors
```

Since BSY is only locally robust, if you “know” there is no serial correlation, then this test is based on LM test is more superior:

```
plmtest(inv ~ value + capital, data = Grunfeld, type = "honda")
#>
#> Lagrange Multiplier Test - (Honda) for balanced panels
#>
#> data: inv ~ value + capital
#> normal = 28.252, p-value < 2.2e-16
#> alternative hypothesis: significant effects
```

On the other hand, if you know there is no random effects, to test for serial correlation, use (BREUSCH, 1978)-(Godfrey, 1978)’s test

```
lmtest::bgtest()
```

If you “know” there are random effects, use (Baltagi and Li, 1995)’s. to test for serial correlation in both AR(1) and MA(1) processes.

H_0 : Uncorrelated errors.

Note:

- one-sided only has power against positive serial correlation.
- applicable to only balanced panels.

```

pbtest(log(gsp)~log(pcap)+log(pc)+log(emp)+unemp,
       data=Produc, alternative="onesided")
#>
#> Baltagi and Li one-sided LM test
#>
#> data: log(gsp) ~ log(pcap) + log(pc) + log(emp) + unemp
#> z = 21.69, p-value < 2.2e-16
#> alternative hypothesis: AR(1)/MA(1) errors in RE panel model

```

General serial correlation tests

- applicable to random effects model, OLS, and FE (with large T, also known as long panel).
- can also test higher-order serial correlation

```

plm::pbgttest(plm::plm(inv~value+capital, data = Grunfeld, model = "within"), order = 2)
#>
#> Breusch-Godfrey/Wooldridge test for serial correlation in panel models
#>
#> data: inv ~ value + capital
#> chisq = 42.587, df = 2, p-value = 5.655e-10
#> alternative hypothesis: serial correlation in idiosyncratic errors

```

in the case of short panels (small T and large n), we can use

```

pwartest(log(emp) ~ log(wage) + log(capital), data=EmplUK)
#>
#> Wooldridge's test for serial correlation in FE panels
#>
#> data: plm.model
#> F = 312.3, df1 = 1, df2 = 889, p-value < 2.2e-16
#> alternative hypothesis: serial correlation

```

12.4.3.5 Unit roots/stationarity

- Dickey-Fuller test for stochastic trends.
- Null hypothesis: the series is non-stationary (unit root)
- You would want your test to be less than the critical value (p<.5) so that there is evidence there is not unit roots.

12.4.3.6 Heteroskedasticity

- Breusch-Pagan test

- Null hypothesis: the data is homoskedastic
- If there is evidence for heteroskedasticity, robust covariance matrix is advised.
- To control for heteroskedasticity: Robust covariance matrix estimation (Sandwich estimator)
 - “white1” - for general heteroskedasticity but no serial correlation (check serial correlation first). Recommended for random effects.
 - “white2” - is “white1” restricted to a common variance within groups. Recommended for random effects.
 - “arellano” - both heteroskedasticity and serial correlation. Recommended for fixed effects

12.4.4 Model Selection

12.4.4.1 POLS vs. RE

The continuum between RE (used FGLS which more assumption) and POLS check back on the section of FGLS

Breusch-Pagan LM test

- Test for the random effect model based on the OLS residual
- Null hypothesis: variances across entities is zero. In another word, no panel effect.
- If the test is significant, RE is preferable compared to POLS

12.4.4.2 FE vs. RE

- RE does not require strict exogeneity for consistency (feedback effect between residual and covariates)

Hypothesis	If true
$H_0 : Cov(c_i, \mathbf{x}_{it}) = 0$	$\hat{\beta}_{RE}$ is consistent and efficient, while $\hat{\beta}_{FE}$ is consistent
$H_0 : Cov(c_i, \mathbf{x}_{it}) \neq 0$	$\hat{\beta}_{RE}$ is inconsistent, while $\hat{\beta}_{FE}$ is consistent

Hausman Test

For the Hausman test to run, you need to assume that

- strict exogeneity hold
- A4 to hold for u_{it}

Then,

- Hausman test statistic: $H = (\hat{\beta}_{RE} - \hat{\beta}_{FE})'(V(\hat{\beta}_{RE}) - V(\hat{\beta}_{FE}))(\hat{\beta}_{RE} - \hat{\beta}_{FE}) \sim \chi^2_{n(X)}$ where $n(X)$ is the number of parameters for the time-varying regressors.
- A low p-value means that we would reject the null hypothesis and prefer FE
- A high p-value means that we would not reject the null hypothesis and consider RE estimator.

```
gw <- plm(inv~value+capital, data=Grunfeld, model="within")
gr <- plm(inv~value+capital, data=Grunfeld, model="random")
phtest(gw, gr)
#>
#> Hausman Test
#>
#> data: inv ~ value + capital
#> chisq = 2.3304, df = 2, p-value = 0.3119
#> alternative hypothesis: one model is inconsistent
```

Violation	Basic	Instrument	Variable	Generalized	General	Means	CCEMG	imator
Esti-	Es-	vari-	Coeffi-	Method	FGLS	groups	for	
ma-	ti-	able	cient	of	esti-	esti-	limited	
tor	ma-	Esti-	estima-	Moments	ma-	ma-	depen-	
	tor	mator	tor	estima-	tor	tor	dent	
				tor			variables	

12.4.5 Summary

- All three estimators (POLS, RE, FE) require A1, A2, A5 (for individuals) to be consistent. Additionally,
 - If A4 does not hold, use cluster robust SE but POLS is not efficient
- POLS is consistent under A3a(for u_{it}): $E(\mathbf{x}'_{it} u_{it}) = 0$, and RE Assumption $E(\mathbf{x}'_{it} c_i) = 0$
 - If A4 (for u_{it}) holds then usual SE are valid and RE is most efficient
- RE is consistent under A3a(for u_{it}): $E(\mathbf{x}'_{it} u_{it}) = 0$, and RE Assumption $E(\mathbf{x}'_{it} c_i) = 0$
 - If A4 (for u_{it}) holds then usual SE are valid and RE is most efficient

- If A4 (for u_{it}) does not hold, use cluster robust SE ,and RE is no longer most efficient (but still more efficient than POLS)
- FE is consistent under A3 $E((\mathbf{x}_{it} - \bar{\mathbf{x}}_{it})'(\mathbf{u}_{it} - \bar{\mathbf{u}}_{it})) = 0$
 - Cannot estimate effects of time constant variables
 - A4 generally does not hold for $\mathbf{u}_{it} - \bar{\mathbf{u}}_{it}$ so cluster robust SE are needed

Note: A5 for individual (not for time dimension) implies that you have A5a for the entire data set.

Estimator / True Model	POLS	RE	FE
POLS	Consistent	Consistent	Inconsistent
FE	Consistent	Consistent	Consistent
RE	Consistent	Consistent	Inconsistent

Based on table provided by Ani Katchova

12.4.6 Application

Recommended application of `plm` can be found here and here by Yves Croissant

```
#install.packages("plm")
library("plm")

library(foreign)
Panel <- read.dta("http://dss.princeton.edu/training/Panel101.dta")

attach(Panel)
Y <- cbind(y)
X <- cbind(x1, x2, x3)

# Set data as panel data
pdata <- pdata.frame(Panel, index=c("country", "year"))

# Pooled OLS estimator
pooling <- plm(Y ~ X, data=pdata, model= "pooling")
summary(pooling)

# Between estimator
between <- plm(Y ~ X, data=pdata, model= "between")
summary(between)
```

```

# First differences estimator
firstdiff <- plm(Y ~ X, data=pdata, model= "fd")
summary(firstdiff)

# Fixed effects or within estimator
fixed <- plm(Y ~ X, data=pdata, model= "within")
summary(fixed)

# Random effects estimator
random <- plm(Y ~ X, data=pdata, model= "random")
summary(random)

# LM test for random effects versus OLS
# Accept Null, then OLS, Reject Null then RE
plmtest(pooling, effect = "individual", type = c("bp")) # other type: "honda", "kw", "ghm"; other

# B-P/LM and Pesaran CD (cross-sectional dependence) test
pcdtest(fixed, test = c("lm")) # Breusch and Pagan's original LM statistic
pcdtest(fixed, test = c("cd")) # Pesaran's CD statistic

# Serial Correlation
pbgttest(fixed)

# stationary
library("tseries")
adf.test(pdata$y, k = 2)

# LM test for fixed effects versus OLS
pFtest(fixed, pooling)

# Hausman test for fixed versus random effects model
phptest(random, fixed)

# Breusch-Pagan heteroskedasticity
library(lmtest)
bptest(y ~ x1 + factor(country), data = pdata)

# If there is presence of heteroskedasticity
## For RE model
coeftest(random) #orginal coef
coeftest(random, vcovHC) # Heteroskedasticity consistent coefficients

t(sapply(c("HCO", "HC1", "HC2", "HC3", "HC4"), function(x) sqrt(diag(vcovHC(random, type = x)))))

# HCO - heteroskedasticity consistent. The default.

```

```
# HC1, HC2, HC3 - Recommended for small samples. HC3 gives less weight to influential observations
# HC4 - small samples with influential observations
# HAC - heteroskedasticity and autocorrelation consistent

## For FE model
coeftest(fixed) # Original coefficients
coeftest(fixed, vcovHC) # Heteroskedasticity consistent coefficients
coeftest(fixed, vcovHC(fixed, method = "arellano")) # Heteroskedasticity consistent covariances
t(sapply(c("HCO", "HC1", "HC2", "HC3", "HC4"), function(x) sqrt(diag(vcovHC(fixed, type = "HC1"))[x])))
```

Advanced

Other methods to estimate the random model:

- "swar": *default* (Swamy and Arora, 1972)
- "walhus": (Wallace and Hussain, 1969)
- "amemiya": (Fuller and Battese, 1974)
- "nerlove": (Nerlove, 1971)

Other effects:

- Individual effects: *default*
- Time effects: "time"
- Individual and time effects: "twoways"

Note: no random two-ways effect model for `random.method = "nerlove"`

```
amemiya <- plm(Y ~ X, data=pdata, model= "random", random.method = "amemiya", effect = "twoways")
```

To call the estimation of the variance of the error components

```
ercomp(Y~X, data=pdata, method = "amemiya", effect = "twoways")
```

Check for the unbalancedness. Closer to 1 indicates balanced data (Ahrens and Pincus, 1981)

```
punbalancedness(random)
```

Instrumental variable

- "bvk": default (Balestra and Varadharajan-Krishnakumar, 1987)
- "baltagi": (Baltagi, 1981)
- "am" (Amemiya and MacCurdy, 1986)
- "bms": (Breusch et al., 1989)

```
instr <- plm(Y ~ X | X_ins, data = pdata, random.method = "ht", model = "random", inst.method = "
```

12.4.7 Other Estimators

12.4.7.1 Variable Coefficients Model

```
fixed_pvcm <- pvcm(Y~X, data=pdata, model="within")
random_pvcm <- pvcm(Y~X, data=pdata, model="random")
```

More details can be found here

12.4.7.2 Generalized Method of Moments Estimator

Typically use in dynamic models. Example is from plm package

```
z2 <- pgmm(log(emp) ~ lag(log(emp), 1)+ lag(log(wage), 0:1) +
            lag(log(capital), 0:1) | lag(log(emp), 2:99) +
            lag(log(wage), 2:99) + lag(log(capital), 2:99),
            data = EmplUK, effect = "twoways", model = "onestep",
            transformation = "ld")
summary(z2, robust = TRUE)
```

12.4.7.3 General Feasible Generalized Least Squares Models

Assume there is no cross-sectional correlation Robust against intragroup heteroskedasticity and serial correlation. Suited when n is much larger than T (long panel) However, inefficient under groupwise heteroskedasticity.

```
# Random Effects
zz <- pggls(log(emp)~log(wage)+log(capital), data=EmplUK, model="pooling")

# Fixed
zz <- pggls(log(emp)~log(wage)+log(capital), data=EmplUK, model="within")
```


Chapter 13

Hypothesis Testing

Error types:

- Type I Error (False Positive):
 - Reality: nope
 - Diagnosis/Analysis: yes
- Type II Error (False Negative):
 - Reality: yes
 - Diagnosis/Analysis: nope

Power: The probability of rejecting the null hypothesis when it is actually false

Note:

- Always written in terms of the population parameter (β) not the estimator/estimate ($\hat{\beta}$)
- Sometimes, different disciplines prefer to use β (i.e., standardized coefficient), or \mathbf{b} (i.e., unstandardized coefficient)
 - β and \mathbf{b} are similar in interpretation; however, β is scale free. Hence, you can see the relative contribution of β to the dependent variable.
On the other hand, \mathbf{b} can be more easily used in policy decisions.
 - $$\beta_j = \mathbf{b} \frac{s_{x_j}}{s_y}$$
- Assuming the null hypothesis is true, what is the (asymptotic) distribution of the estimator

- Two-sided

$$H_0 : \beta_j = 0 \quad H_1 : \beta_j \neq 0$$

then under the null, the OLS estimator has the following distribution

$$A1 - A3a, A5 : \sqrt{n}\hat{\beta}_j \sim N(0, Avar(\sqrt{n}\hat{\beta}_j))$$

- For the one-sided test, the null is a set of values, so now you choose the worst case single value that is hardest to prove and derive the distribution under the null
- One-sided

$$H_0 : \beta_j \geq 0 \quad H_1 : \beta_j < 0$$

then the hardest null value to prove is $H_0 : \beta_j = 0$. Then under this specific null, the OLS estimator has the following asymptotic distribution

$$A1 - A3a, A5 : \sqrt{n}\hat{\beta}_j \sim N(0, Avar(\sqrt{n}\hat{\beta}_j))$$

13.1 Types of hypothesis testing

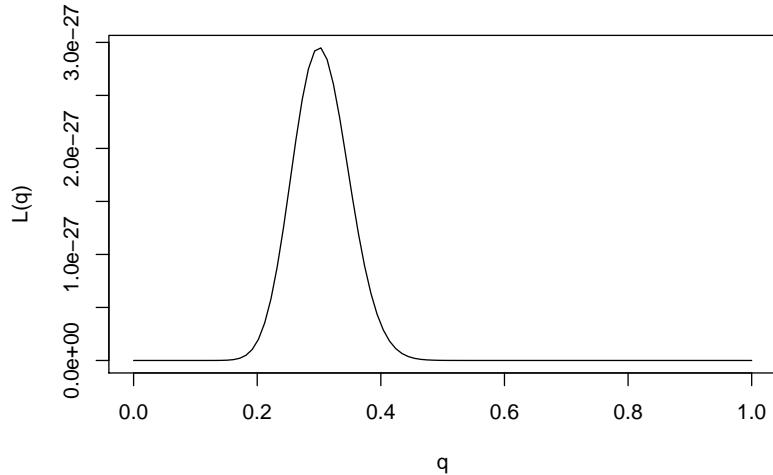
$$H_0 : \theta = \theta_0$$

$$H_1 : \theta \neq \theta_0$$

How far away / extreme θ can be if our null hypothesis is true

Assume that our likelihood function for q is $L(q) = q^{30}(1-q)^{70}$ **Likelihood function**

```
q = seq(0,1,length=100)
L= function(q){q^30 * (1-q)^70}
plot(q,L(q),ylab="L(q)",xlab="q",type="l")
```

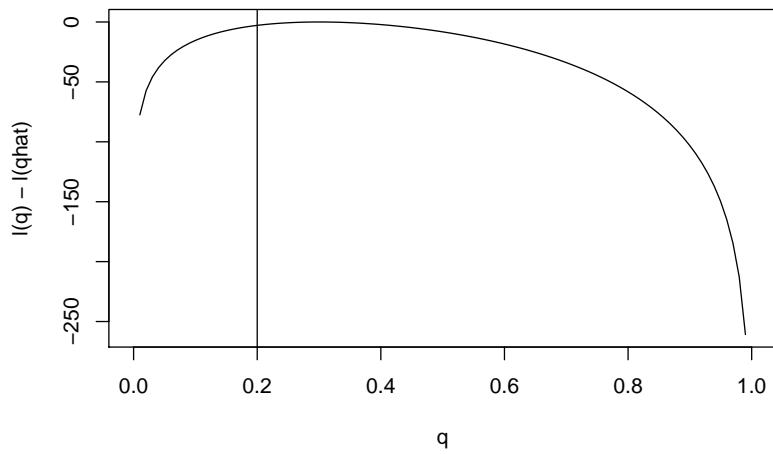


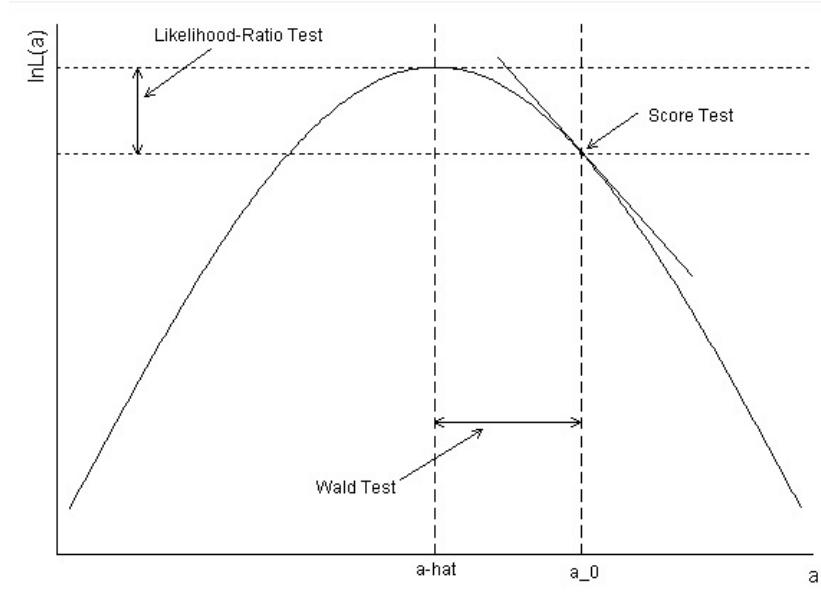
Log-Likelihood function

```

q = seq(0,1,length=100)
l= function(q){30*log(q) + 70 * log(1-q)}
plot(q,l(q)-l(0.3),ylab="l(q) - l(qhat)",xlab="q",type="l")
abline(v=0.2)

```





(Fox, 1991)

typically, The likelihood ratio test (and Lagrange Multiplier (Score)) performs better with small to moderate sample sizes, but the Wald test only requires one maximization (under the full model).

13.2 Wald test

$$W = (\hat{\theta} - \theta_0)' [cov(\hat{\theta})]^{-1} (\hat{\theta} - \theta_0) W \sim \chi_q^2$$

where $cov(\hat{\theta})$ is given by the inverse Fisher Information matrix evaluated at $\hat{\theta}$ and q is the rank of $cov(\hat{\theta})$, which is the number of non-redundant parameters in θ

Alternatively,

$$t_W = \frac{(\hat{\theta} - \theta_0)^2}{I(\theta_0)^{-1}} \sim \chi_{(v)}^2$$

where v is the degree of freedom.

Equivalently,

$$s_W = \frac{\hat{\theta} - \theta_0}{\sqrt{I(\hat{\theta})^{-1}}} \sim Z$$

How far away in the distribution your sample estimate is from the hypothesized population parameter.

For a null value, what is the probability you would have obtained a realization “more extreme” or “worse” than the estimate you actually obtained?

Significance Level (α) and Confidence Level ($1 - \alpha$)

- The significance level is the benchmark in which the probability is so low that we would have to reject the null
- The confidence level is the probability that sets the bounds on how far away the realization of the estimator would have to be to reject the null.

Test Statistics

- Standardized (transform) the estimator and null value to a test statistic that always has the same distribution
- Test Statistic for the OLS estimator for a single hypothesis

$$T = \frac{\sqrt{n}(\hat{\beta}_j - \beta_{j0})}{\sqrt{n}SE(\hat{\beta}_j)} \sim^a N(0, 1)$$

Equivalently,

$$T = \frac{(\hat{\beta}_j - \beta_{j0})}{SE(\hat{\beta}_j)} \sim^a N(0, 1)$$

the test statistic is another random variable that is a function of the data and null hypothesis.

- T denotes the random variable test statistic
- t denotes the single realization of the test statistic

Evaluating Test Statistic: determine whether or not we reject or fail to reject the null hypothesis at a given significance / confidence level

Three equivalent ways

1. Critical Value
2. P-value

- 3. Confidence Interval
- 4. Critical Value

For a given significance level, will determine the critical value (c)

* One-sided: $H_0 : \beta_j \geq \beta_{j0}$

$$P(T < c | H_0) = \alpha$$

Reject the null if $t < c$

- One-sided: $H_0 : \beta_j \leq \beta_{j0}$

$$P(T > c | H_0) = \alpha$$

Reject the null if $t > c$

- Two-sided: $H_0 : \beta_j \neq \beta_{j0}$

$$P(|T| > c | H_0) = \alpha$$

Reject the null if $|t| > c$

2. p-value

Calculate the probability that the test statistic was worse than the realization you have

- One-sided: $H_0 : \beta_j \geq \beta_{j0}$

$$\text{p-value} = P(T < t | H_0)$$

- One-sided: $H_0 : \beta_j \leq \beta_{j0}$

$$\text{p-value} = P(T > t | H_0)$$

- Two-sided: $H_0 : \beta_j \neq \beta_{j0}$

$$\text{p-value} = P(|T| < t | H_0)$$

reject the null if p-value < α

3. Confidence Interval

Using the critical value associated with a null hypothesis and significance level, create an interval

$$CI(\hat{\beta}_j)_\alpha = [\hat{\beta}_j - (c \times SE(\hat{\beta}_j)), \hat{\beta}_j + (c \times SE(\hat{\beta}_j))]$$

If the null set lies outside the interval then we reject the null.

- We are not testing whether the true population value is close to the estimate, we are testing that given a field true population value of the parameter, how like it is that we observed this estimate.
- Can be interpreted as we believe with $(1 - \alpha) \times 100\%$ probability that the confidence interval captures the true parameter value.

With stronger assumption (A1-A6), we could consider Finite Sample Properties

$$T = \frac{\hat{\beta}_j - \beta_{j0}}{SE(\hat{\beta}_j)} \sim T(n - k)$$

- This above distributional derivation is strongly dependent on A4 and A5
- T has a student t-distribution because the numerator is normal and the denominator is χ^2 .
- Critical value and p-values will be calculated from the student t-distribution rather than the standard normal distribution.
- $n \rightarrow \infty$, $T(n - k)$ is asymptotically standard normal.

Rule of thumb

- if $n - k > 120$: the critical values and p-values from the t-distribution are (almost) the same as the critical values and p-values from the standard normal distribution.
- if $n - k < 120$
 - if (A1-A6) hold then the t-test is an exact finite distribution test
 - if (A1-A3a, A5) hold, because the t-distribution is asymptotically normal, computing the critical values from a t-distribution is still a valid asymptotic test (i.e., not quite the right critical values and p-values, the difference goes away as $n \rightarrow \infty$)

13.2.1 Multiple Hypothesis

- test multiple parameters at the same time
 - $H_0 : \beta_1 = 0 \& \beta_2 = 0$
 - $H_0 : \beta_1 = 1 \& \beta_2 = 0$
- perform a series of simple hypothesis does not answer the question (joint distribution vs. two marginal distributions).
- The test statistic is based on a restriction written in matrix form.

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + \epsilon$$

Null hypothesis is $H_0 : \beta_1 = 0 \& \beta_2 = 0$ can be rewritten as $H_0 : \mathbf{R}\beta - \mathbf{q} = 0$ where

- \mathbf{R} is a $m \times k$ matrix where m is the number of restrictions and k is the number of parameters. \mathbf{q} is a $k \times 1$ vector
- \mathbf{R} “picks up” the relevant parameters while \mathbf{q} is the null value of the parameter

$$\mathbf{R} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \mathbf{q} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Test Statistic for OLS estimator for a multiple hypothesis

$$F = \frac{(\mathbf{R} - \mathbf{q})\hat{\Sigma}^{-1}(\mathbf{R} - \mathbf{q})^T}{m} \sim^a F(m, n - k)$$

- $\hat{\Sigma}^{-1}$ is the estimator for the asymptotic variance-covariance matrix
 - if A4 holds, both the homoskedastic and heteroskedastic versions produce valid estimator
 - If A4 does not hold, only the heteroskedastic version produces valid estimators.
- When $m = 1$, there is only a single restriction, then the F-statistic is the t-statistic squared.
- F distribution is strictly positive, check F-Distribution for more details.

13.2.2 Linear Combination

Testing multiple parameters at the same time

$$H_0 : \beta_1 - \beta_2 = 0 \quad H_0 : \beta_1 - \beta_2 > 0 \quad H_0 : \beta_1 - 2 * \beta_2 = 0$$

Each is a single restriction on a function of the parameters.

Null hypothesis:

$$H_0 : \beta_1 - \beta_2 = 0$$

can be rewritten as

$$H_0 : \mathbf{R}\beta - \mathbf{q} = 0$$

where $\mathbf{R}=(0 \ 1 \ -1 \ 0 \ 0)$ and $\mathbf{q} = 0$

13.2.3 Estimate Difference in Coefficients

There is no package to estimate for the difference between two coefficients and its CI, but a simple function created by Katherine Zee can be used to calculate this difference. Some modifications might be needed if you don't use standard `lm` model in R.

```
diffest_lm <- function(x1, x2, model) {
  diffest <-
    summary(model)$coef[x1, "Estimate"] - summary(model)$coef[x2, "Estimate"]
  vardiff <- (summary(model)$coef[x1, "Std. Error"] ^ 2 +
    summary(model)$coef[x2, "Std. Error"] ^ 2) - (2 * (vcov(model)[x1, x2]))
  # variance of x1 + variance of x2 - 2*covariance of x1 and x2
  diffse <- sqrt(vardiff)
  tdiff <- (diffest) / (diffse)
  ptdiff <- 2 * (1 - pt(abs(tdiff), model$df, lower.tail = T))
  upr <-
    diffest + qt(.975, df = model$df) * diffse # will usually be very close to 1.96
  lwr <- diffest + qt(.025, df = model$df) * diffse
  df <- model$df
  return(list(
    est = round(diffest, digits = 2),
    t = round(tdiff, digits = 2),
    p = round(ptdiff, digits = 4),
    lwr = round(lwr, digits = 2),
    upr = round(upr, digits = 2),
  ))
}
```

```
        df = df
    ))
}
```

13.2.4 Application

```
library("car")

# Multiple hypothesis
mod.davis <- lm(weight ~ repwt, data=Davis)
linearHypothesis(mod.davis, c("(Intercept) = 0", "repwt = 1"), white.adjust = TRUE)
#> Linear hypothesis test
#>
#> Hypothesis:
#> (Intercept) = 0
#> repwt = 1
#>
#> Model 1: restricted model
#> Model 2: weight ~ repwt
#>
#> Note: Coefficient covariance matrix supplied.
#>
#>   Res.Df Df      F Pr(>F)
#> 1     183
#> 2     181  2 3.3896 0.03588 *
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Linear Combination
mod.duncan <- lm(prestige ~ income + education, data=Duncan)
linearHypothesis(mod.duncan, "1*income - 1*education = 0")
#> Linear hypothesis test
#>
#> Hypothesis:
#> income - education = 0
#>
#> Model 1: restricted model
#> Model 2: prestige ~ income + education
#>
#>   Res.Df    RSS Df Sum of Sq      F Pr(>F)
#> 1     43 7518.9
#> 2     42 7506.7  1     12.195 0.0682 0.7952
```

13.2.5 Nonlinear

Suppose that we have q nonlinear functions of the parameters

$$\mathbf{h}(\theta) = \{h_1(\theta), \dots, h_q(\theta)\}'$$

The, n, the Jacobian matrix ($\mathbf{H}(\theta)$), of rank q is

$$\mathbf{H}_{q \times p}(\theta) = \begin{pmatrix} \frac{\partial h_1(\theta)}{\partial \theta_1} & \dots & \frac{\partial h_1(\theta)}{\partial \theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_q(\theta)}{\partial \theta_1} & \dots & \frac{\partial h_q(\theta)}{\partial \theta_p} \end{pmatrix}$$

where the null hypothesis $H_0 : \mathbf{h}(\theta) = 0$ can be tested agiasnt the 2-sided alternative with the Wald statistic

$$W = \frac{\mathbf{h}(\hat{\theta})' \{ \mathbf{H}(\hat{\theta}) [\mathbf{F}(\hat{\theta})' \mathbf{F}(\hat{\theta})]^{-1} \mathbf{H}(\hat{\theta}) \}' \}^{-1} \mathbf{h}(\hat{\theta})}{s^2 q} \sim F_{q, n-p}$$

13.3 The likelihood ratio test

$$t_{LR} = 2[l(\hat{\theta}) - l(\theta_0)] \sim \chi_v^2$$

where v is the degree of freedom.

Compare the height of the log-likelihood of the sample estimate in relation to the height of log-likelihood of the hypothesized population parameter

Alternatively,

This test considers a ratio of two maximizations,

L_r = maximized value of the likelihood under H_0 (the reduced model)
 L_f = maximized value of the likelihood under

Then, the likelihood ratio is:

$$\Lambda = \frac{L_r}{L_f}$$

which can't exceed 1 (since L_f is always at least as large as $L - r$ because L_r is the result of a maximization under a restricted set of the parameter values).

The likelihood ratio statistic is:

$$-2\ln(\Lambda) = -2\ln(L_r/L_f) = -2(l_r - l_f) \lim_{n \rightarrow \infty} (-2\ln(\Lambda)) \sim \chi_v^2$$

where v is the number of parameters in the full model minus the number of parameters in the reduced model.

If L_r is much smaller than L_f (the likelihood ratio exceeds $\chi_{\alpha,v}^2$), then we reject the reduced model and accept the full model at $\alpha \times 100\%$ significance level

13.4 Lagrange Multiplier (Score)

$$t_S = \frac{S(\theta_0)^2}{I(\theta_0)} \sim \chi_v^2$$

where v is the degree of freedom.

Compare the slope of the log-likelihood of the sample estimate in relation to the slope of the log-likelihood of the hypothesized population parameter

Chapter 14

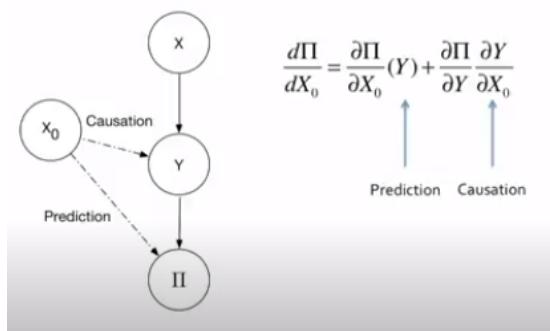
Prediction and Estimation

Prediction and Estimation (Inference) have been the two fundamental pillars in statistics.

- You cannot have both. You can either have high prediction or high estimation.
 - In prediction, you minimize the loss function.
 - In estimation, you try to best fit the data. Because the goal of estimation is to best fit the data, you always run the risk of not predicting well.

In high dimension, you always have weak to strong collinearity. Hence, your estimation can be undesirable. And you can't pick which one variable to stay in the model, but all these troubles would not affect your prediction. In Plateau problem

- If two functions are similar in output space, you can still do prediction, but you can't do estimation because of exploded standard errors.



(SICSS 2018 - Sendhil Mullainathan's presentation slide)

Selective Labels Problem (The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables)

Recall Linear Regression 5 OLS estimates

$$\begin{aligned}\hat{\beta}_{OLS} &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'(\mathbf{X} + \epsilon)) \\ &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\beta + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\epsilon) \\ \hat{\beta}_{OLS} &\rightarrow \beta + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\epsilon)\end{aligned}$$

Hence, OLS estimates will be unbiased (i.e., get rid of $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\epsilon)$) if we have the following 2 conditions:

1. $E(\epsilon|\mathbf{X}) = 0$ With an intercept, we can usually solve this problem
2. $Cov(\mathbf{X}, \epsilon) = 0$

Problem with estimation usually stems from the second condition.

Tools to combat this problem can be found in causal inference 16

Chapter 15

Moderation

- Spotlight Analysis: Compare the mean of the dependent of the two groups (treatment and control) at every value (Simple Slopes Analysis)
- Floodlight Analysis: is spotlight analysis on the whole range of the moderator (Johnson-Neyman intervals)

Other Resources:

- **BANOVAL** : floodlight analysis on Bayesian ANOVA models
- **cSEM** : `doFloodlightAnalysis` in SEM model
- (Spiller et al., 2013)

Terminology:

- Main effects (slopes): coefficients that do no involve interaction terms
- Simple slope: when a continuous independent variable interact with a moderating variable, its slope at a particular level of the moderating variable
- Simple effect: when a categorical independent variable interacts with a moderating variable, its effect at a particular level of the moderating variable.

Example:

$$Y = \beta_0 + \beta_1 X + \beta_2 M + \beta_3 X \times M$$

where

- β_0 = intercept
- β_1 = simple effect (slope) of X (independent variable)
- β_2 = simple effect (slope) of M (moderating variable)
- β_3 = interaction of X and M

Three types of interactions:

1. Continuous by continuous
2. Continuous by categorical
3. Categorical by categorical

15.1 emmeans package

```
install.packages("emmeans")
```

```
library(emmeans)
```

Data set is from UCLA seminar where `gender` and `prog` are categorical

```
dat <- readRDS("data/exercise.rds") %>%
  mutate(prog = factor(prog, labels = c("jog", "swim", "read"))) %>%
  mutate(gender = factor(gender, labels = c("male", "female")))
```

15.1.1 Continuous by continuous

```
contcont <- lm(loss~hours*effort,data=dat)
summary(contcont)
#>
#> Call:
#> lm(formula = loss ~ hours * effort, data = dat)
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -29.52 -10.60  -1.78  11.13  34.51
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
```

```
#> (Intercept) 7.79864 11.60362 0.672 0.5017
#> hours -9.37568 5.66392 -1.655 0.0982 .
#> effort -0.08028 0.38465 -0.209 0.8347
#> hours:effort 0.39335 0.18750 2.098 0.0362 *
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 13.56 on 896 degrees of freedom
#> Multiple R-squared: 0.07818, Adjusted R-squared: 0.07509
#> F-statistic: 25.33 on 3 and 896 DF, p-value: 9.826e-16
```

Simple slopes for a continuous by continuous model

Spotlight analysis (Aiken and West, 2005): usually pick 3 values of moderating variable:

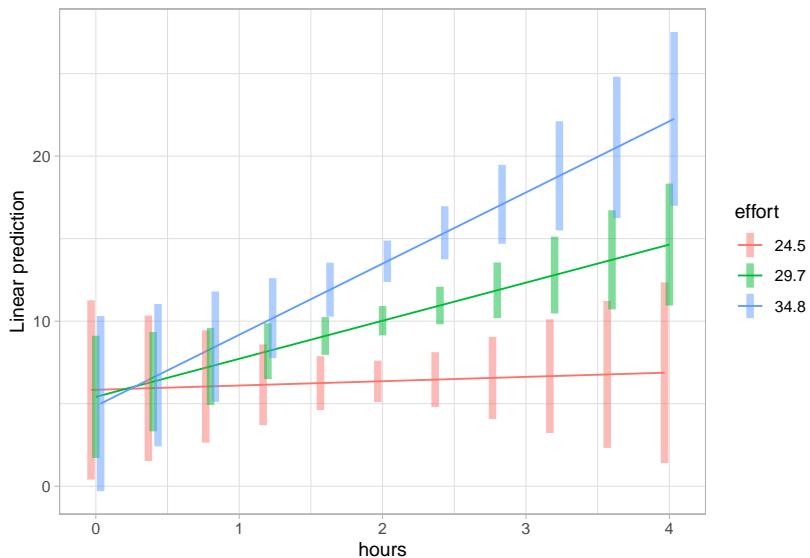
- Mean Moderating Variable + $\sigma \times$ (Moderating variable)
- Mean Moderating Variable
- Mean Moderating Variable - $\sigma \times$ (Moderating variable)

```
effar <- round(mean(dat$effort) + sd(dat$effort), 1)
effr <- round(mean(dat$effort), 1)
effbr <- round(mean(dat$effort) - sd(dat$effort), 1)

# specify list of points
mylist <- list(effort=c(effbr,effr,effar))

# get the estimates
emtrends(contcont, ~effort, var="hours", at=mylist)
#> effort hours.trend SE df lower.CL upper.CL
#> 24.5 0.261 1.352 896 -2.392 2.91
#> 29.7 2.307 0.915 896 0.511 4.10
#> 34.8 4.313 1.308 896 1.745 6.88
#>
#> Confidence level used: 0.95

# plot
mylist <- list(hours=seq(0,4,by=0.4),effort=c(effbr,effr,effar))
emmip(contcont,effort~hours,at=mylist, CIs=TRUE)
```



```
# statistical test for slope difference
emtrends(contcont, pairwise ~effort, var="hours", at=list(mylist), adjust="none")
#> $emtrends
#>   effort hours.trend    SE  df lower.CL upper.CL
#>   24.5      0.261 1.352 896   -2.392    2.91
#>   29.7      2.307 0.915 896     0.511    4.10
#>   34.8      4.313 1.308 896     1.745    6.88
#>
#> Results are averaged over the levels of: hours
#> Confidence level used: 0.95
#>
#> $contrasts
#>   contrast   estimate    SE  df t.ratio p.value
#>   24.5 - 29.7    -2.05 0.975 896   -2.098  0.0362
#>   24.5 - 34.8    -4.05 1.931 896   -2.098  0.0362
#>   29.7 - 34.8    -2.01 0.956 896   -2.098  0.0362
#>
#> Results are averaged over the levels of: hours
```

The 3 p-values are the same as the interaction term.

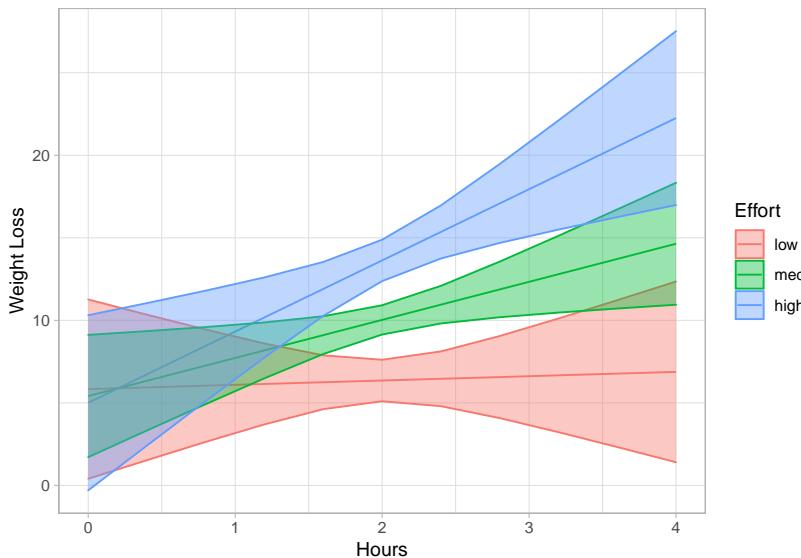
For publication, we use

```
library(ggplot2)

# data
```

```
(mylist                                <- list(hours=seq(0,4,by=0.4),effort=c(effbr,effr,effar)))
#> $hours
#> [1] 0.0 0.4 0.8 1.2 1.6 2.0 2.4 2.8 3.2 3.6 4.0
#>
#> $effort
#> [1] 24.5 29.7 34.8
contcontdat                         <- emmip(contcont,effort~hours,at=mylist, CIs=TRUE, plotit=FALSE)
contcontdat$feffort                  <- factor(contcontdat$effort)
levels(contcontdat$feffort) <- c("low","med","high")

# plot
p <- ggplot(data = contcontdat, aes(x = hours, y = yvar, color = feffort)) + geom_line()
p1 <- p + geom_ribbon(aes(ymax=UCL, ymin=LCL, fill=feffort), alpha=0.4)
p1 + labs(x="Hours", y="Weight Loss", color="Effort", fill="Effort")
```



15.1.2 Continuous by categorical

```
# use Female as baseline
dat$gender <- relevel(dat$gender, ref="female")

contcat <- lm(loss ~ hours * gender, data = dat)
summary(contcat)
#>
#> Call:
```

```
#> lm(formula = loss ~ hours * gender, data = dat)
#>
#> Residuals:
#>    Min      1Q  Median      3Q     Max
#> -27.118 -11.350 -1.963 10.001 42.376
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   3.335     2.731   1.221   0.222
#> hours         3.315     1.332   2.489   0.013 *
#> gendermale    3.571     3.915   0.912   0.362
#> hours:gendermale -1.724     1.898  -0.908   0.364
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 14.06 on 896 degrees of freedom
#> Multiple R-squared:  0.008433, Adjusted R-squared:  0.005113
#> F-statistic:  2.54 on 3 and 896 DF, p-value: 0.05523
```

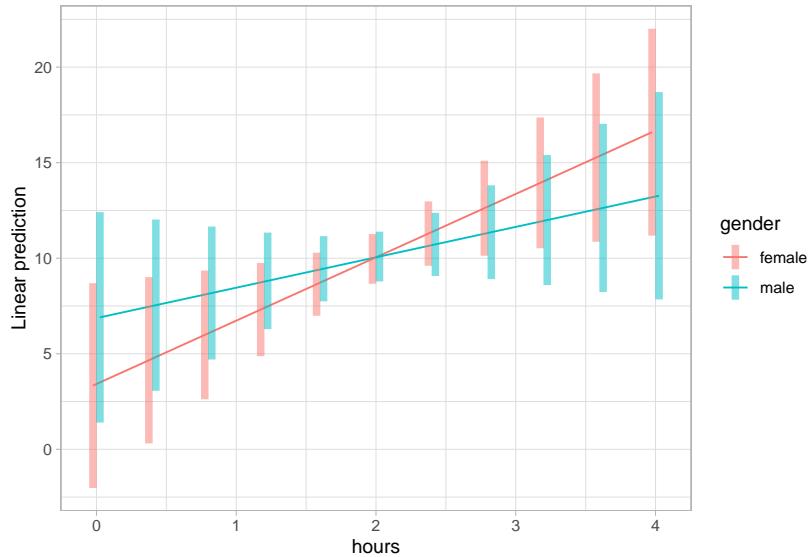
Get simple slopes by each level of the categorical moderator

```
emtrends(concat, ~ gender, var="hours")
#> gender hours.trend SE df lower.CL upper.CL
#> female      3.32 1.33 896    0.702    5.93
#> male        1.59 1.35 896   -1.063    4.25
#>
#> Confidence level used: 0.95

# test difference in slopes
emtrends(concat, pairwise ~ gender, var="hours") # which is the same as the interaction
#> $emtrends
#> gender hours.trend SE df lower.CL upper.CL
#> female      3.32 1.33 896    0.702    5.93
#> male        1.59 1.35 896   -1.063    4.25
#>
#> Confidence level used: 0.95
#>
#> $contrasts
#> contrast      estimate SE df t.ratio p.value
#> female - male     1.72 1.9 896    0.908   0.3639
```

```
# plot
(mylist <- list(hours=seq(0,4,by=0.4),gender=c("female","male")))
#> $hours
#> [1] 0.0 0.4 0.8 1.2 1.6 2.0 2.4 2.8 3.2 3.6 4.0
```

```
#>
#> $gender
#> [1] "female" "male"
emmmip(concat, gender ~hours, at=mylist, CIs=TRUE)
```



15.1.3 Categorical by categorical

```
# relevel baseline
dat$prog <- relevel(dat$prog, ref="read")
dat$gender <- relevel(dat$gender, ref="female")

catcat <- lm(loss ~ gender * prog, data = dat)
summary(catcat)
#>
#> Call:
#> lm(formula = loss ~ gender * prog, data = dat)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -19.1723  -4.1894  -0.0994  3.7506  27.6939
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
```

```
#> (Intercept)      -3.6201    0.5322  -6.802 1.89e-11 ***
#> gendermale       -0.3355    0.7527  -0.446   0.656
#> progjog          7.9088    0.7527  10.507 < 2e-16 ***
#> progsim          32.7378    0.7527  43.494 < 2e-16 ***
#> gendermale:progjog 7.8188    1.0645   7.345 4.63e-13 ***
#> gendermale:progsim -6.2599    1.0645  -5.881 5.77e-09 ***
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 6.519 on 894 degrees of freedom
#> Multiple R-squared:  0.7875, Adjusted R-squared:  0.7863
#> F-statistic: 662.5 on 5 and 894 DF, p-value: < 2.2e-16
```

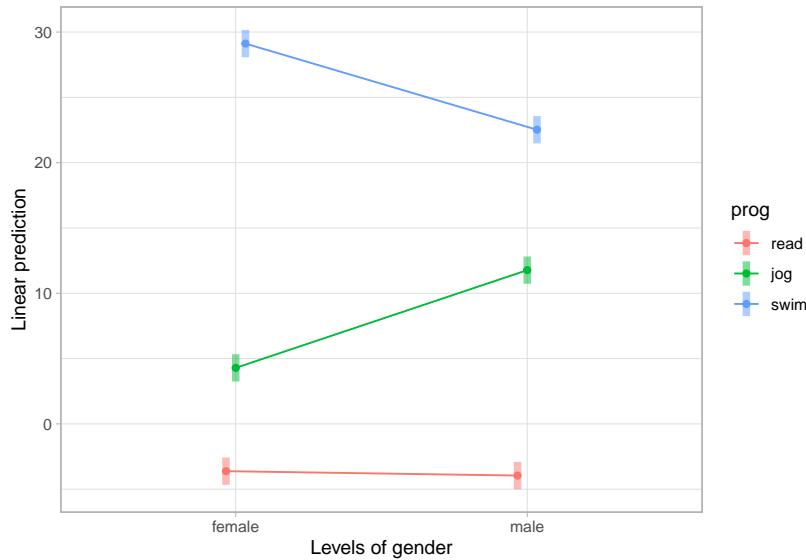
Simple effects

```
emcatcat <- emmeans(catcat, ~ gender*prog)

# differences in predicted values
contrast(emcatcat, "revpairwise", by = "prog", adjust = "bonferroni")
#> prog = read:
#> contrast estimate SE df t.ratio p.value
#> male - female -0.335 0.753 894 -0.446 0.6559
#>
#> prog = jog:
#> contrast estimate SE df t.ratio p.value
#> male - female 7.483 0.753 894 9.942 <.0001
#>
#> prog = swim:
#> contrast estimate SE df t.ratio p.value
#> male - female -6.595 0.753 894 -8.762 <.0001
```

Plot

```
emmip(catcat, prog ~ gender, CIs=TRUE)
```



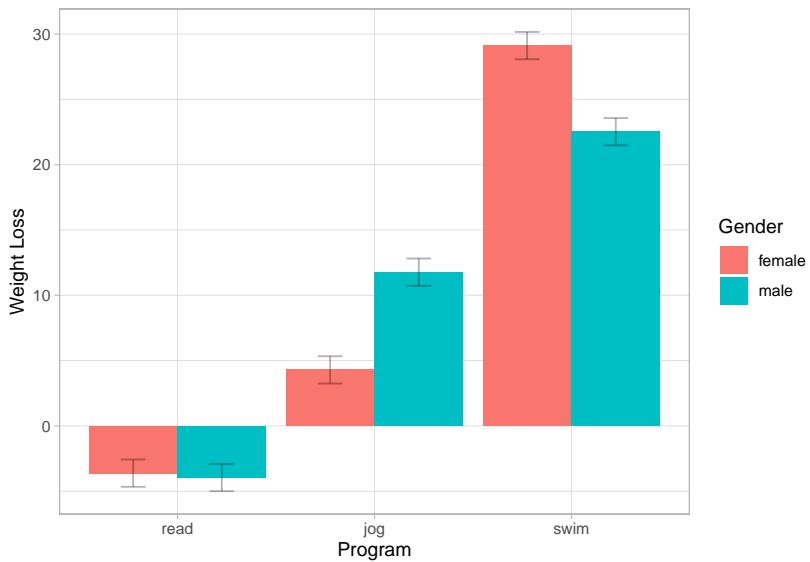
Bar graph

```

catcatdat <- emmip(catcat, gender ~ prog, CIs = TRUE, plotit = FALSE)
p <-
  ggplot(data = catcatdat, aes(x = prog, y = yvar, fill = gender)) +
  geom_bar(stat = "identity", position = "dodge")

p1 <-
  p + geom_errorbar(
    position = position_dodge(.9),
    width = .25,
    aes(ymax = UCL, ymin = LCL),
    alpha = 0.3
  )
p1 + labs(x = "Program", y = "Weight Loss", fill = "Gender")

```



15.2 probmod package

- Not recommend: package has serious problem with subscript.

```
install.packages("probemod")

library(probemod)

myModel <-
  lm(loss ~ hours * gender, data = dat %>% select(loss, hours, gender))
jnresults <- jn(myModel, dv='loss', iv='hours', mod='gender')

pickapoint(myModel, dv='loss', iv='hours', mod='gender', alpha=.01)
plot(jnresults)
```

15.3 interactions package

- Recommend

```
install.packages("interactions")
```

15.3.1 Continuous interaction

- (at least one of the two variables is continuous)

```
library(interactions)
library(jtools) # for summ()
states <- as.data.frame(state.x77)
fiti <- lm(Income ~ Illiteracy * Murder + `HS Grad`, data = states)
summ(fiti)
```

Observations	50
Dependent variable	Income
Type	OLS linear regression

F(4,45)	10.65
R ²	0.49
Adj. R ²	0.44

	Est.	S.E.	t val.	p
(Intercept)	1414.46	737.84	1.92	0.06
Illiteracy	753.07	385.90	1.95	0.06
Murder	130.60	44.67	2.92	0.01
'HS Grad'	40.76	10.92	3.73	0.00
Illiteracy:Murder	-97.04	35.86	-2.71	0.01

Standard errors: OLS

For continuous moderator, the three values chosen are:

- -1 SD above the mean
- The mean
- -1 SD below the mean

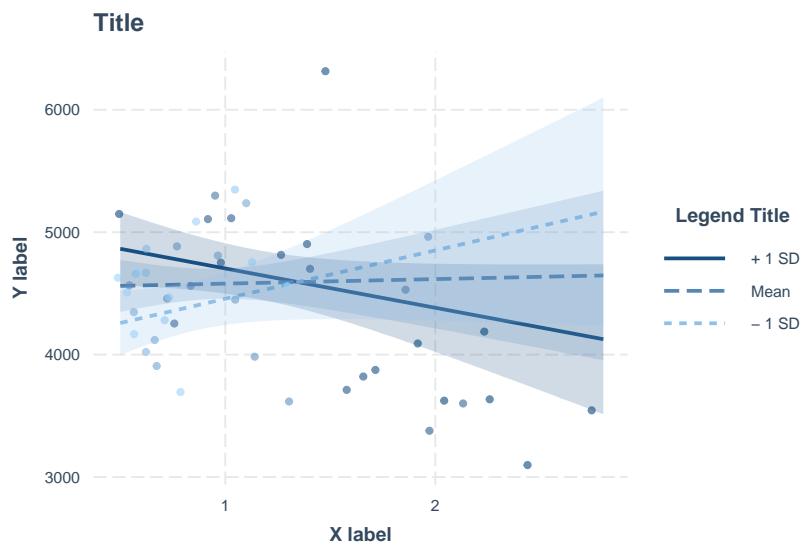
```
interact_plot(fiti,
              pred = Illiteracy,
              modx = Murder,
              # centered = "none", # if you don't want the plot to mean-center
```

```

# modx.values = "plus-minus", # exclude the mean value of the moderator
# modx.values = "terciles" # split moderator's distribution into 3 group
plot.points = T, # overlay data
point.shape = T, # different shape for differennt levels of the moderator
jitter = 0.1, # if two data points are on top one another, this moves them
# other appearance option
x.label = "X label",
y.label = "Y label",
main.title = "Title",
legend.main = "Legend Title",
colors = "blue",

# include confidence band
interval = TRUE,
int.width = 0.9,
robust = TRUE # use robust SE
)

```



To include weights from the regression inn the plot

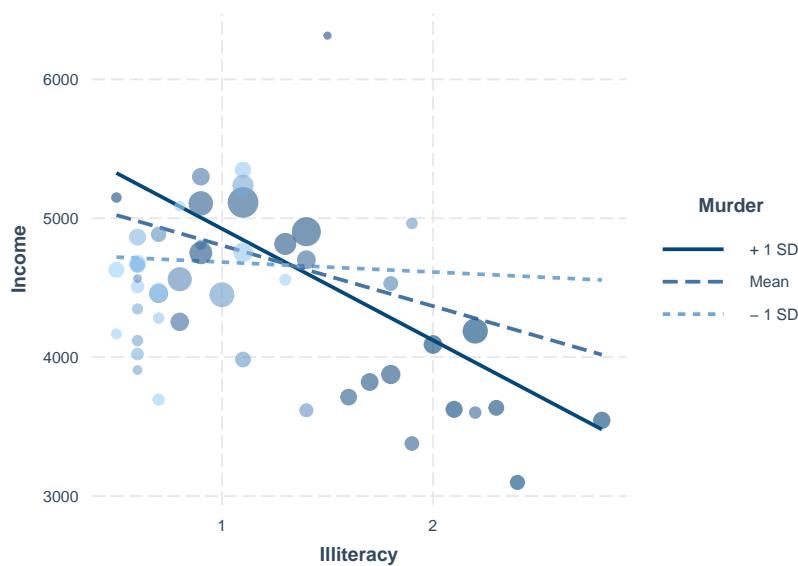
```

fiti <- lm(Income ~ Illiteracy * Murder,
            data = states,
            weights = Population)

interact_plot(fiti,

```

```
pred = Illiteracy,
modx = Murder,
plot.points = TRUE)
```



Partial Effect Plot

```
library(ggplot2)
data(cars)
fitc <- lm(cty ~ year + cyl * displ + class + fl + drv, data = mpg)
summ(fitc)
```

Observations	234
Dependent variable	cty
Type	OLS linear regression

F(16,217)	99.73
R ²	0.88
Adj. R ²	0.87

```
interact_plot(
  fitc,
  pred = displ,
```

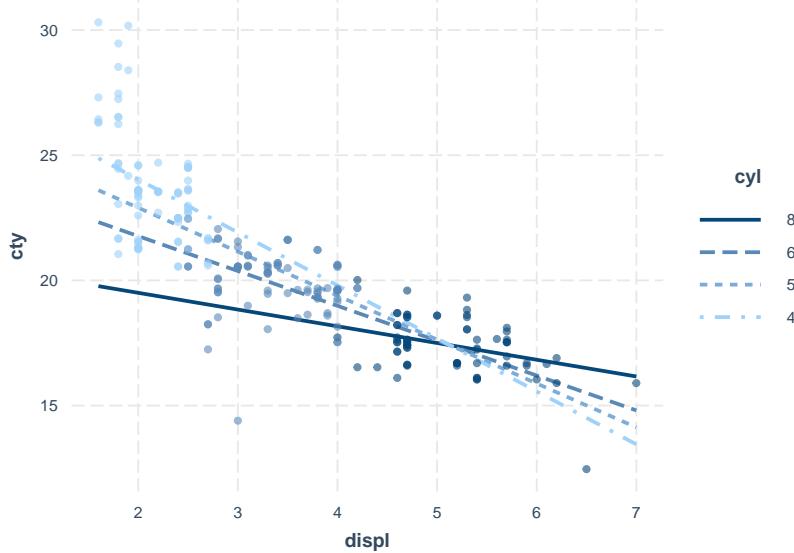
	Est.	S.E.	t val.	p
(Intercept)	-200.98	47.01	-4.28	0.00
year	0.12	0.02	5.03	0.00
cyl	-1.86	0.28	-6.69	0.00
displ	-3.56	0.66	-5.41	0.00
classcompact	-2.60	0.93	-2.80	0.01
classmidsize	-2.63	0.93	-2.82	0.01
classminivan	-4.41	1.04	-4.24	0.00
classpickup	-4.37	0.93	-4.68	0.00
classsubcompact	-2.38	0.93	-2.56	0.01
classsuv	-4.27	0.87	-4.92	0.00
fld	6.34	1.69	3.74	0.00
fle	-4.57	1.66	-2.75	0.01
flp	-1.92	1.59	-1.21	0.23
flr	-0.79	1.57	-0.50	0.61
drvf	1.40	0.40	3.52	0.00
drvrv	0.49	0.46	1.06	0.29
cyl:displ	0.36	0.08	4.56	0.00

Standard errors: OLS

```

modx = cyl,
partial.residuals = TRUE, # the observed data is based on displ, cyl, and model error
modx.values = c(4, 5, 6, 8)
)

```



Check linearity assumption in the model

Plot the lines based on the subsample (red line), and whole sample (black line)

```
x_2 <- runif(n = 200, min = -3, max = 3)
w <- rbinom(n = 200, size = 1, prob = 0.5)
err <- rnorm(n = 200, mean = 0, sd = 4)
y_2 <- 2.5 - x_2 ^ 2 - 5 * w + 2 * w * (x_2 ^ 2) + err

data_2 <- as.data.frame(cbind(x_2, y_2, w))

model_2 <- lm(y_2 ~ x_2 * w, data = data_2)
summ(model_2)
```

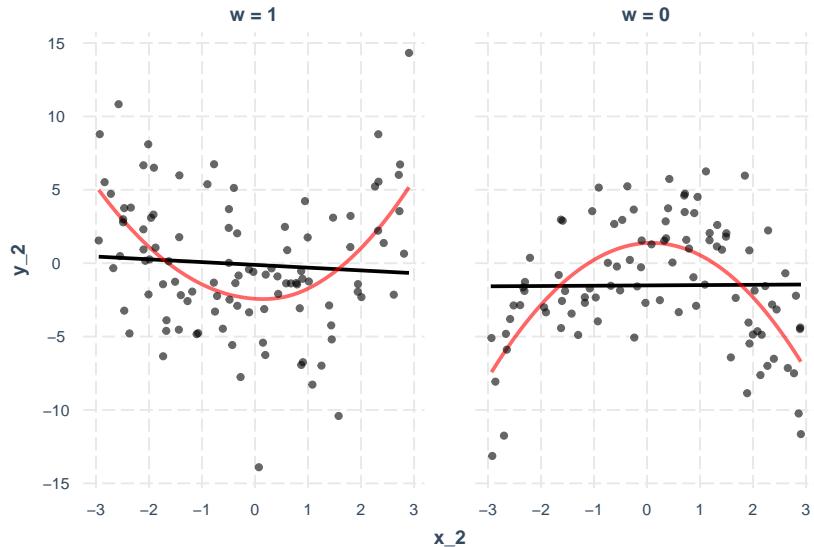
Observations	200
Dependent variable	y_2
Type	OLS linear regression

F(3,196)	1.95
R ²	0.03
Adj. R ²	0.01

	Est.	S.E.	t val.	p
(Intercept)	-1.51	0.45	-3.36	0.00
x_2	0.02	0.26	0.08	0.94
w	1.40	0.63	2.22	0.03
x_2:w	-0.21	0.37	-0.56	0.57

Standard errors: OLS

```
interact_plot(
  model_2,
  pred = x_2,
  modx = w,
  linearity.check = TRUE,
  plot.points = TRUE
)
```



15.3.1.1 Simple Slopes Analysis

- continuous by continuous variable interaction (still work for binary)
- conditional slope of the variable of interest (i.e., the slope of X when we hold M constant at a value)

Using `sim_slopes` it will

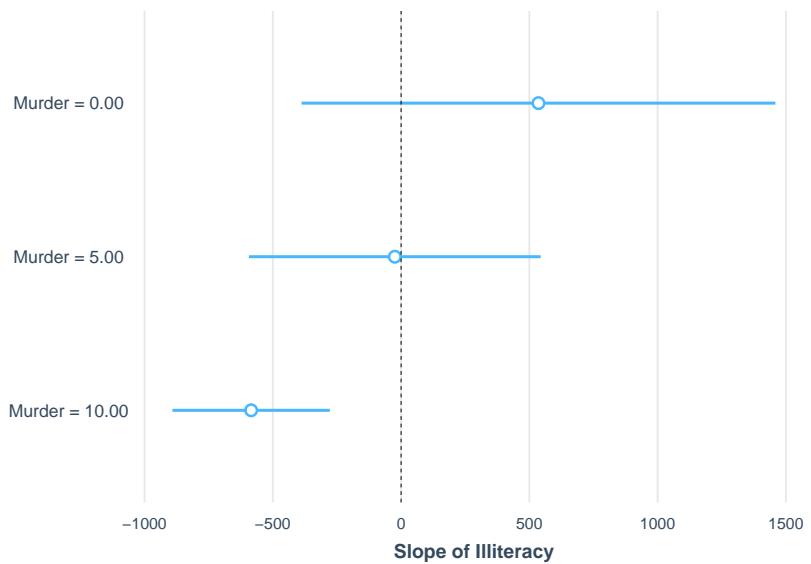
- mean-center all variables except the variable of interest
- For moderator that is
 - Continuous, it will pick mean, and plus/minus 1 SD
 - Categorical, it will use all factor

`sim_slopes` requires

- A regression model with an interaction term)
- Variable of interest (`pred =`)
- Moderator: (`modx =`)

```
sim_slopes(fiti,
            pred = Illiteracy,
            modx = Murder,
            johnson_neyman = FALSE)
#> SIMPLE SLOPES ANALYSIS
#>
#> Slope of Illiteracy when Murder = 5.420973 (- 1 SD):
#>
#>   Est.     S.E.    t val.     p
#>   -----
#>   -71.59   268.65    -0.27   0.79
#>
#> Slope of Illiteracy when Murder = 8.685043 (Mean):
#>
#>   Est.     S.E.    t val.     p
#>   -----
#>   -437.12   175.82    -2.49   0.02
#>
#> Slope of Illiteracy when Murder = 11.949113 (+ 1 SD):
#>
#>   Est.     S.E.    t val.     p
#>   -----
#>   -802.66   145.72    -5.51   0.00

# plot the coefficients
ss <- sim_slopes(fiti,
                  pred = Illiteracy,
                  modx = Murder,
                  modx.values = c(0, 5, 10))
plot(ss)
```



```
# table
ss <- sim_slopes(fiti,
                   pred = Illiteracy,
                   modx = Murder,
                   modx.values = c(0, 5, 10))
library(huxtable)
as_huxtable(ss)
```

Table 15.1

Value of Murder	Slope of Illiteracy
Value of Murder	slope
0.00	535.50 (458.77)
5.00	-24.44 (282.48)
10.00	-584.38 (152.37)***

15.3.1.2 Johnson-Neyman intervals

To know all the values of the moderator for which the slope of the variable of interest will be statistically significant, we can use the Johnson-Neyman interval (Johnson and Neyman, 1936)

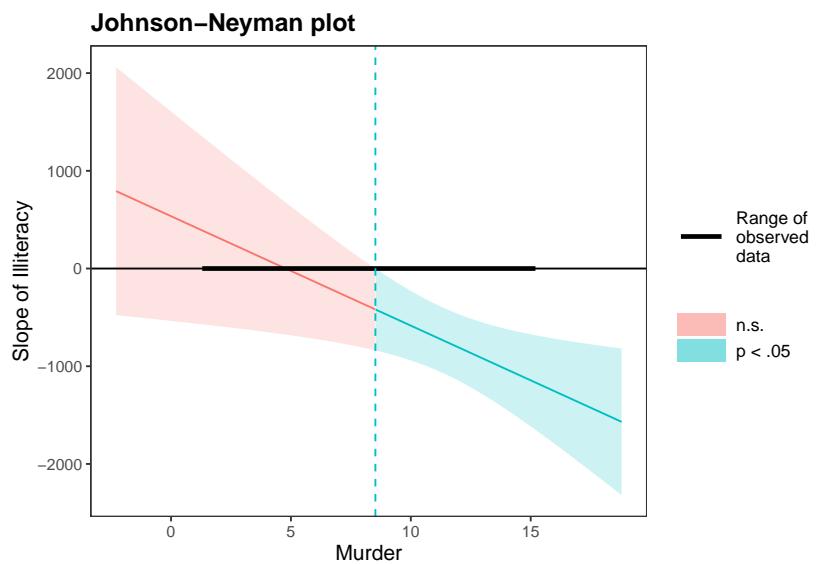
Even though we kind of know that the alpha level when implementing the Johnson-Neyman interval is not correct (Bauer and Curran, 2005), not until recently that there is a correction for the type I and II errors (Esarey and Sumner, 2017).

Since Johnson-Neyman inflates the type I error (comparisons across all values of the moderator)

```
sim_slopes(fiti,
            pred = Illiteracy,
            modx = Murder,
            johnson_neyman = TRUE,
            control.fdr = TRUE, # correction for type I and II
            # cond.int = TRUE, # include conditional intercepts
            robust = "HC3", # robust SE
            # centered = "none", # don't mean-centered non-focal variables
            jnalpha = 0.05)
#> JOHNSON-NEYMAN INTERVAL
#>
#> When Murder is OUTSIDE the interval [-11.70, 8.75], the slope of Illiteracy
#> is p < .05.
#>
#> Note: The range of observed values of Murder is [1.40, 15.10]
#>
#> Interval calculated using false discovery rate adjusted t = 2.33
#>
#> SIMPLE SLOPES ANALYSIS
#>
#> Slope of Illiteracy when Murder = 5.420973 (- 1 SD):
#>
#>   Est.     S.E.    t val.      p
#> -----
#> -71.59   256.60    -0.28   0.78
#>
#> Slope of Illiteracy when Murder = 8.685043 (Mean):
#>
#>   Est.     S.E.    t val.      p
#> -----
#> -437.12   191.07    -2.29   0.03
#>
#> Slope of Illiteracy when Murder = 11.949113 (+ 1 SD):
#>
#>   Est.     S.E.    t val.      p
#> -----
#> -802.66   178.75    -4.49   0.00
```

For plotting, we can use `johnson_neyman`

```
johnson_neyman(fiti,
  pred = Illiteracy,
  modx = Murder,
  control.fdr = TRUE, # correction for type I and II
  alpha = .05)
#> JOHNSON-NEYMAN INTERVAL
#>
#> When Murder is OUTSIDE the interval [-22.57, 8.52], the slope of Illiteracy
#> is p < .05.
#>
#> Note: The range of observed values of Murder is [1.40, 15.10]
#>
#> Interval calculated using false discovery rate adjusted t = 2.33
```



Note:

- y-axis is the **conditional slope** of the variable of interest

15.3.1.3 3-way interaction

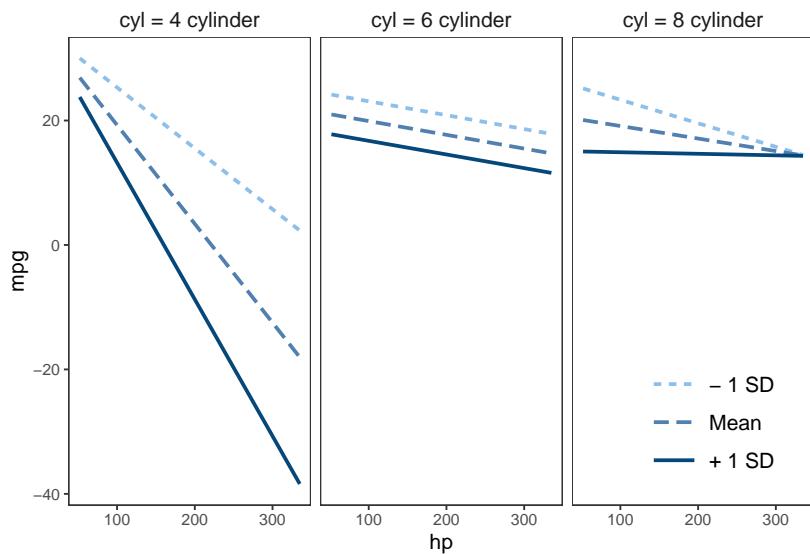
```
# fita3 <-
#   lm(rating ~ privileges * critical * learning, data = attitude)
```

```

# 
# probe_interaction(
#   fita3,
#   pred = critical,
#   modx = learning,
#   mod2 = privileges,
#   alpha = .1
# )

mtcars$cyl <- factor(mtcars$cyl,
  labels = c("4 cylinder", "6 cylinder", "8 cylinder"))
fitc3 <- lm(mpg ~ hp * wt * cyl, data = mtcars)
interact_plot(fitc3,
  pred = hp,
  modx = wt,
  mod2 = cyl) +
  theme_apa(legend.pos = "bottomright")

```



Johnson-Neyman 3-way interaction

```

library(survey)
data(api)

dstrat <- svydesign(

```

```

    id = ~ 1,
    strata = ~ stype,
    weights = ~ pw,
    data = apistrat,
    fpc = ~ fpc
)

regmodel3 <-
  survey::svyglm(api00 ~ avg.ed * growth * enroll, design = dstrat)

sim_slopes(
  regmodel3,
  pred = growth,
  modx = avg.ed,
  mod2 = enroll,
  jnplot = TRUE
)
#> ##### While enroll (2nd moderator) = 153.0518 (- 1 SD) #####
#>
#> JOHNSON-NEYMAN INTERVAL
#>
#> When avg.ed is OUTSIDE the interval [2.75, 3.82], the slope of growth is p
#> < .05.
#>
#> Note: The range of observed values of avg.ed is [1.38, 4.44]
#>
#> SIMPLE SLOPES ANALYSIS
#>
#> Slope of growth when avg.ed = 2.085002 (- 1 SD):
#>
#>   Est.   S.E.   t val.      p
#>   ----- -----
#>   1.25   0.32     3.86   0.00
#>
#> Slope of growth when avg.ed = 2.787381 (Mean):
#>
#>   Est.   S.E.   t val.      p
#>   ----- -----
#>   0.39   0.22     1.75   0.08
#>
#> Slope of growth when avg.ed = 3.489761 (+ 1 SD):
#>
#>   Est.   S.E.   t val.      p
#>   ----- -----
#>   -0.48   0.35    -1.37   0.17

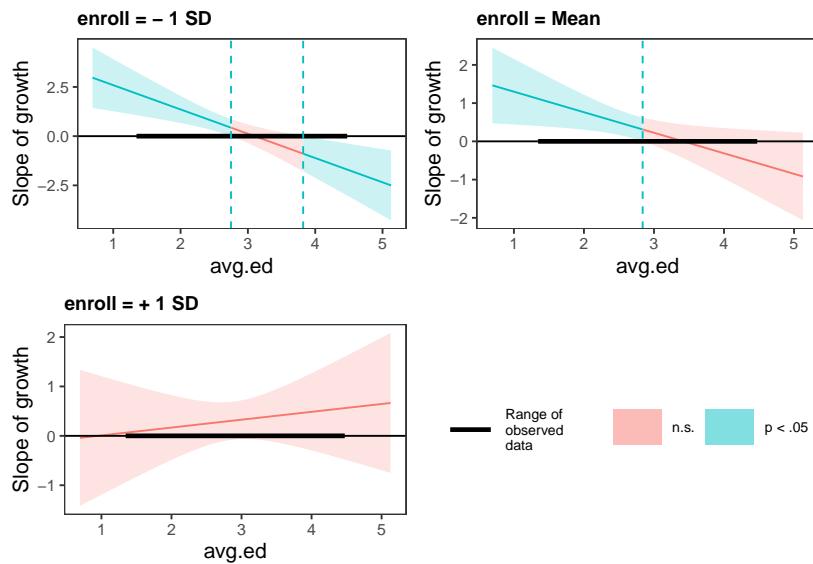
```

```

#>
#> ##### While enroll (2nd moderator) = 595.2821 (Mean) #####
#>
#> JOHNSON-NEYMAN INTERVAL
#>
#> When avg.ed is OUTSIDE the interval [2.84, 7.83], the slope of growth is p
#> < .05.
#>
#> Note: The range of observed values of avg.ed is [1.38, 4.44]
#>
#> SIMPLE SLOPES ANALYSIS
#>
#> Slope of growth when avg.ed = 2.085002 (- 1 SD):
#>
#>   Est.   S.E.   t val.      p
#>   ----- ----- -----
#>   0.72   0.22    3.29   0.00
#>
#> Slope of growth when avg.ed = 2.787381 (Mean):
#>
#>   Est.   S.E.   t val.      p
#>   ----- ----- -----
#>   0.34   0.16    2.16   0.03
#>
#> Slope of growth when avg.ed = 3.489761 (+ 1 SD):
#>
#>   Est.   S.E.   t val.      p
#>   ----- ----- -----
#>   -0.04   0.24   -0.16   0.87
#>
#> ##### While enroll (2nd moderator) = 1037.5125 (+ 1 SD) #####
#>
#> JOHNSON-NEYMAN INTERVAL
#>
#> The Johnson-Neyman interval could not be found. Is the p value for your
#> interaction term below the specified alpha?
#>
#> SIMPLE SLOPES ANALYSIS
#>
#> Slope of growth when avg.ed = 2.085002 (- 1 SD):
#>
#>   Est.   S.E.   t val.      p
#>   ----- ----- -----
#>   0.18   0.31    0.58   0.56
#>

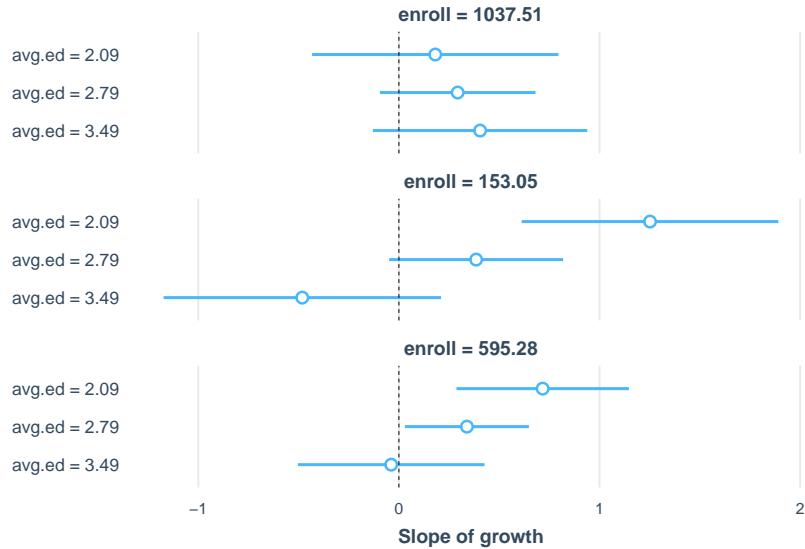
```

```
#> Slope of growth when avg.ed = 2.787381 (Mean):
#>
#>   Est.    S.E.    t val.      p
#>   -----
#>   0.29    0.20     1.49    0.14
#>
#> Slope of growth when avg.ed = 3.489761 (+ 1 SD):
#>
#>   Est.    S.E.    t val.      p
#>   -----
#>   0.40    0.27     1.49    0.14
```



Report

```
ss3 <-
  sim_slopes(regmodel3,
              pred = growth,
              modx = avg.ed,
              mod2 = enroll)
plot(ss3)
```



```
as_huxtable(ss3)
```

15.3.2 Categorical interaction

```
library(ggplot2)
mpg2 <- mpg %>%
  mutate(cyl = factor(cyl))

mpg2["auto"] <- "auto"
mpg2$auto[mpg2$trans %in% c("manual(m5)", "manual(m6)")] <- "manual"
mpg2$auto <- factor(mpg2$auto)
mpg2["fwd"] <- "2wd"
mpg2$fwd[mpg2$drv == "4"] <- "4wd"
mpg2$fwd <- factor(mpg2$fwd)
## Drop the two cars with 5 cylinders (rest are 4, 6, or 8)
mpg2 <- mpg2[mpg2$cyl != "5",]
## Fit the model
fit3 <- lm(cty ~ cyl * fwd * auto, data = mpg2)

library(jtools) # for summ()
summ(fit3)
```

Table 15.2

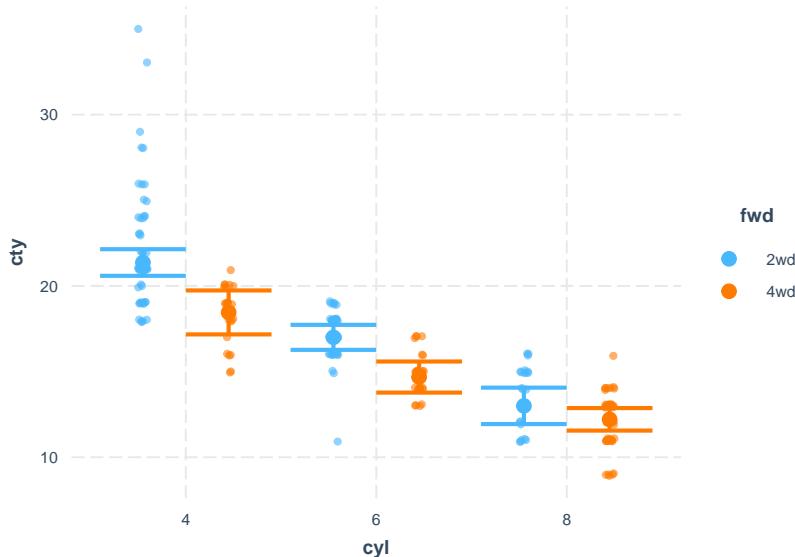
<i>enroll = 153</i>	
Value of avg.ed	Slope of growth
Value of avg.ed	slope
2.09	1.25 (0.32)***
2.79	0.39 (0.22)##
<i>enroll = 595.28</i>	
Value of avg.ed	Slope of growth
3.49	-0.48 (0.35)
2.09	0.72 (0.22)**
2.79	0.34 (0.16)*
<i>enroll = 1037.51</i>	
Value of avg.ed	Slope of growth
3.49	-0.04 (0.24)
2.09	0.18 (0.31)
2.79	0.29 (0.20)
3.49	0.40 (0.27)
Observations	230
Dependent variable	cty
Type	OLS linear regression
F(11,218)	61.37
R ²	0.76
Adj. R ²	0.74

```
cat_plot(fit3,
         pred = cyl,
         modx = fwd,
```

	Est.	S.E.	t val.	p
(Intercept)	21.37	0.39	54.19	0.00
cyl6	-4.37	0.54	-8.07	0.00
cyl8	-8.37	0.67	-12.51	0.00
fwd4wd	-2.91	0.76	-3.83	0.00
automanual	1.45	0.57	2.56	0.01
cyl6:fwd4wd	0.59	0.96	0.62	0.54
cyl8:fwd4wd	2.13	0.99	2.15	0.03
cyl6:automanual	-0.76	0.90	-0.84	0.40
cyl8:automanual	0.71	1.18	0.60	0.55
fwd4wd:automanual	-1.66	1.07	-1.56	0.12
cyl6:fwd4wd:automanual	1.29	1.52	0.85	0.40
cyl8:fwd4wd:automanual	-1.39	1.76	-0.79	0.43

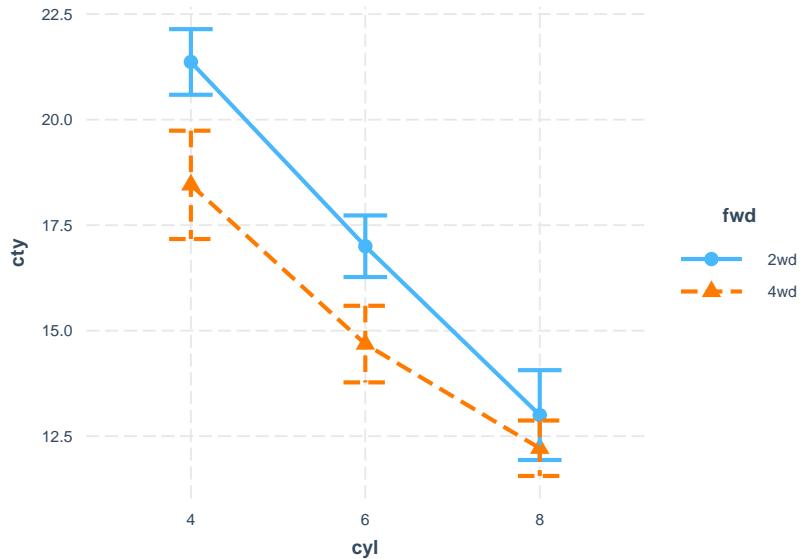
Standard errors: OLS

```
plot.points = T)
```

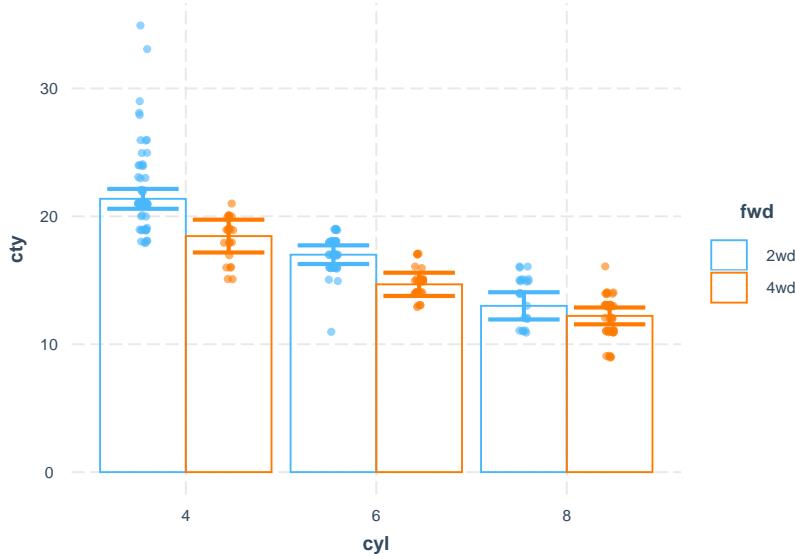


```
#line plots
cat_plot(
  fit3,
  pred = cyl,
  modx = fwd,
  geom = "line",
```

```
  point.shape = TRUE,  
  # colors = "Set2", # choose color  
  vary.lty = TRUE  
)
```



```
# bar plot  
cat_plot(  
  fit3,  
  pred = cyl,  
  modx = fwd,  
  geom = "bar",  
  interval = T,  
  plot.points = TRUE  
)
```



15.4 interactionR package

- For publication purposes
- Following
 - (Knol and VanderWeele, 2012) for presentation
 - (Hosmer and Lemeshow, 1992) for confidence intervals based on the delta method
 - (Zou, 2008) for variance recovery “mover” method
 - (Longnecker et al., 1996) for bootstrapping

```
install.packages("interactionR")
```

15.5 sjPlot package

- For publication purposes (recommend, but more advanced)
- link

IV. CAUSAL INFERENCE

Chapter 16

Causal Inference

After all of the mambo jumbo that we have learned so far, I want to now talk about the concept of causality. We usually say that correlation is not causation. Then, what is causation?

One of my favorite books has explained this concept beautifully (Mackenzie and Pearl, 2018). And I am just going to quickly summarize the gist of it from my understanding. I hope that it can give you an initial grasp on the concept so that later you can continue to read up and develop a deeper understanding.

It's important to have a deep understanding regarding the method research. However, one needs to be aware of its limitation. As mentioned in various sections throughout the book, we see that we need to ask experts for number as our baseline or visit literature to gain insight from past research.

Here, we dive in a more conceptual side statistical analysis as a whole, regardless of particular approach.

You probably heard scientists say correlation doesn't mean causation. There are ridiculous spurious correlations that give a firm grip on what the previous phrase means. The pioneer who tried to use regression to infer causation in social science was Yule (1899) (but it was a fatal attempt where he found relief policy increases poverty). To make a causal inference from statistics, **the equation (function form) must be stable** under intervention (i.e., variables are manipulated). Statistics is used to be a causality-free enterprise in the past.

Not until the development of path analysis by Sewall Wright in the 1920s that the discipline started to pay attention to causation. Then, it remained dormant until the Causal Revolution (quoted Judea Pearl's words). This revolution introduced the calculus of causation which includes (1) causal diagrams), and (2) a symbolic language

The world has been using $P(Y|X)$ (statistics use to derive this), but what we want is to compare the difference between

- $P(Y|do(X))$: treatment group
- $P(Y|do(not - X))$: control group

Hence, we can see a clear difference between $P(Y|X) \neq P(Y|do(X))$

The conclusion we want to make from data is counterfactuals: **What would have happened had we not do X?**

To teach a robot to make inference, we need inference engine

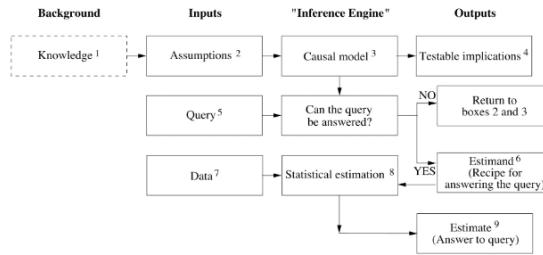


FIGURE I How an “inference engine” combines data with causal knowledge to produce answers to queries of interest. The dashed box is not part of the engine but is required for building it. Arrows could also be drawn from boxes 4 and 9 to box 1, but I have opted to keep the diagram simple.

Figure 16.1: p. 12 (Mackenzie and Pearl, 2018)

Levels of cognitive ability to be a causal learner:

1. Seeing
2. Doing
3. Imagining

Ladder of causation (associated with levels of cognitive ability as well):

1. Association: conditional probability, correlation, regression
2. Intervention
3. Counterfactuals

Level	Activity	Questions	Examples
Association $P(y x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease?
Intervention $P(y do(x), z)$	Doing	What if? Intervening	What if I spend more time learning, will my result change?

Level	Activity	Questions	Examples
Counterfactuals	Imagining Why? $P(y_x x', y')$	was it X that caused Y? What if I had acted differently	What if I stopped smoking a year ago?

Table by (Pearl, 2019, p. 2)

You cannot define causation from probability alone

If you say X causes Y if X raises the probability of Y." On the surface, it might sound intuitively right. But when we translate it to probability notation: $P(Y|X) > P(Y)$, it can't be more wrong. Just because you are seeing X (1st level), it **doesn't mean** the probability of Y increases.

It could be either that (1) X causes Y, or (2) Z affects both X and Y. Hence, people might use **control variables**, which translate: $P(Y|X, Z = z) > P(Y|Z = z)$, then you can be more confident in your probabilistic observation. However, the question is how can you choose Z

With the invention of the do-operator, now you can represent X causes Y as

$$P(Y|do(X)) > P(Y)$$

and with the help of causal diagram, now you can answer questions at the 2nd level (Intervention)

Note: people under econometrics might still use "Granger causality" and "vector autoregression" to use the probability language to represent causality (but it's not).

The 7 tools for Structural Causal Model framework (Pearl, 2019):

1. Encoding Causal Assumptions - transparency and testability (with graphical representation)
2. Do-calculus and the control of confounding: "back-door"
3. The algorithmization of Counterfactuals
4. Mediation Analysis and the Assessment of Direct and Indirect Effects
5. Adaptability, External validity and Sample Selection Bias: are still researched under "domain adaptation", "transfer learning"
6. Recovering from missing data
7. Causal Discovery:

1. d-separation
2. Functional decomposition (Hoyer et al., 2008; Shimizu et al., 2009; Chen and Chan, 2012)
3. Spontaneous local changes (Pearl, 2013)

Simpson's Paradox:

- A statistical association seen in an entire population is reversed in sub-population.

Structural Causal Model accompanies graphical causal model to create more efficient language to represent causality

Structural Causal Model is the solution to the curse of dimensionality (i.e., large numbers of variable p , and small dataset n) thanks to product decomposition. It allows us to solve problems without knowing the function, parameters, or distributions of the error terms.

Suppose you have a causal chain $X \rightarrow Y \rightarrow Z$:

$$P(X = x, Y = y, Z = z) = P(X = x)P(Y = y|X = x)P(Z = z|Y = y)$$

Experimental Design	Quasi-experimental Design
Experimentalist	Observationalist
Experimental Data	Observational Data
Random Assignment (reduce treatment imbalance)	Random Sampling (reduce sample selection error)

Tools in a hierarchical order

1. Experimental Design: Randomized Control Trials (Gold standard): Tier 1
2. Quasi-experimental
 1. Regression Discontinuity Tier 1A
 2. Difference-In-Differences Tier 2
 3. Synthetic Control Tier 2A
 4. Event Studies Tier 2B
 5. Fixed Effects Estimator 12.4.2.2: Tier 3
 6. Endogenous Treatment: mostly Instrumental Variable: Tier 3A

7. Matching Methods Tier 4
8. Interrupted Time Series Tier 4A
9. Endogenous Sample Selection 28.4: mostly Heckman's correction

Internal vs. External Validity

- Internal Validity: Economists and applied scientists mostly care about
- External Validity: Localness might affect your external validity

For many economic policies, there is a difference between **treatment** and **intention to treat**.

For example, we might have an effective vaccine (i.e., intention to treat), but it does not mean that everybody will take it (i.e., treatment).

There are four types of subjects that we deal with:

- **Non-switchers:** we don't care about non-switchers because even if we introduce or don't introduce the intervention, it won't affect them.
 - Always takers
 - Never takers
- **Switchers**
 - **Compliers:** defined as those who respect the intervention.
 - * We only care about compliers because when we introduce the intervention, they will do something. When we don't have any interventions, they won't do it.
 - * Tools above are used to identify the causal impact of an intervention on compliers
 - * If we have only **compliers** in our dataset, then **intention to treatment = treatment effect**.
 - **Defiers:** those who will go to the opposite direction of your treatment.
 - * We typically aren't interested in defiers because they will do the opposite of what we want them to do. And they are typically a small group; hence, we just assume they don't exist.

	Treatment Assignment	Control Assignment
Compliers	Treated	No Treated

	Treatment Assignment	Control Assignment
Always-takers	Treated	Treated
Never-takers	Not treated	No treated
Defiers	Not treated	Treated

Directional Bias due to selection into treatment comes from 2 general opposite sources

1. **Mitigation-based:** select into treatment to combat a problem
2. **Preference-based:** select into treatment because units like that kind of treatment.

16.1 Treatment effect types

This section is based on Paul Testa's note on egap

Terminology:

- Quantities of causal interest (i.e., treatment effect types)
- Estimands: parameters of interest
- Estimators: procedures to calculate estimates for the parameters of interest

Sources of bias (according to prof. Luke Keele)

$$\begin{aligned} \text{Estimator - True Causal Effect} \\ = & \text{Hidden bias} + \text{Misspecification bias} + \text{Statistical Noise} \\ = & \text{Due to design} + \text{Due to modeling} + \text{Due to finite sample} \end{aligned}$$

16.1.1 Average Treatment Effects

Average treatment effect (ATE) is the difference in means of the treated and control groups

Randomization under Experimental Design can provide an unbiased estimate of ATE.

Let $Y_i(1)$ denote the outcome of individual i under treatment and
 $Y_i(0)$ denote the outcome of individual i under control

Then, the treatment effect for individual i is the difference between her outcome under treatment and control

$$\tau_i = Y_i(1) - Y_i(0)$$

Without a time machine or dimension portal, we can only observe one of the two events: either individual i experiences the treatment or she doesn't.

Then, the ATE as a quantity of interest can come in handy since we can observe across all individuals

$$ATE = \frac{1}{N} \sum_{i=1}^N \tau_i = \frac{\sum_1^N Y_i(1)}{N} - \frac{\sum_1^N Y_i(0)}{N}$$

With random assignment (i.e., treatment assignment is independent of potential outcome and observables and unobservables), the observed means difference between the two groups is an unbiased estimator of the average treatment effect

$$E(Y_i(1)|D = 1) = E(Y_i(1)|D = 0) = E(Y_i(1))E(Y_i(0)|D = 1) = E(Y_i(0)|D = 0) = E(Y_i(0))$$

$$ATE = E(Y_i(1)) - E(Y_i(0))$$

16.1.2 Conditional Average Treatment Effects

Treatment effects can be different for different groups of people. In words, treatment effects can vary across subgroups.

To examine the heterogeneity across groups (e.g., men vs. women), we can estimate the conditional average treatment effects (CATE) for each subgroup

$$CATE = E(Y_i(1) - Y_i(0)|D_i, X_i))$$

16.1.3 Intent-to-treat Effects

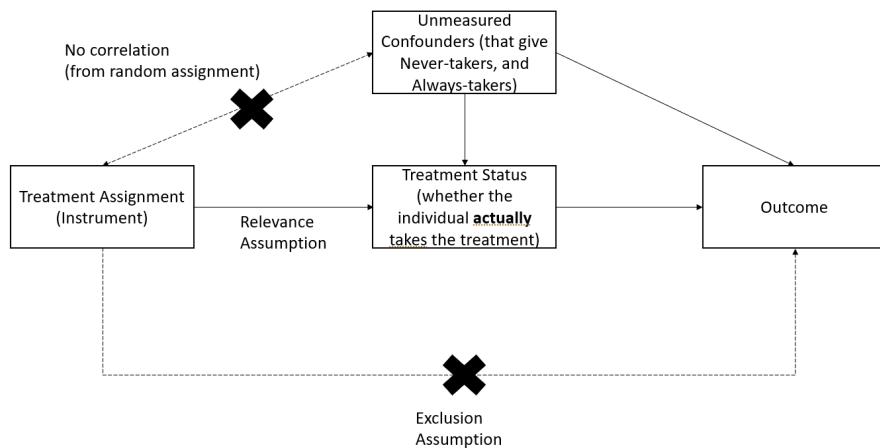
When we encounter non-compliance (either people suppose to receive treatment don't receive it, or people suppose to be in the control group receive the treatment), treatment receipt is not independent of potential outcomes and confounders.

In this case, the difference in observed means between the treatment and control groups is not Average Treatment Effects, but Intent-to-treat Effects (ITT). In words, ITT is the treatment effect on those who **receive** the treatment

16.1.4 Local Average Treatment Effects

Instead of estimating the treatment effects of those who **receive** the treatment (i.e., Intent-to-treat Effects), you want to estimate the treatment effect of those who actually **comply** with the treatment. This is the local average treatment effects (LATE) or complier average causal effects (CACE). I assume we don't use CATE to denote complier average treatment effect because it was reserved for conditional average treatment effects.

- Using random treatment assignment as an instrument, we can recover the effect of treatment on compliers.



- As the percent of compliers increases, Intent-to-treat Effects and Local Average Treatment Effects converge
- Rule of thumb: $SE(LATE) = SE(ITT)/(share of compliers)$
- LATE estimate is always greater than the ITT estimate
- LATE can also be estimated using a pure placebo group (Gerber et al., 2010).
- Partial compliance is hard to study, and IV/2SLS estimator is biased, we have to use Bayesian (Long et al., 2010; Jin and Rubin, 2009, 2008).

16.1.4.1 One-sided noncompliance

- One-sided noncompliance is when in the sample, we only have compliers and never-takers

- With the exclusion restriction (i.e., excludability), never-takers have the same results in the treatment or control group (i.e., never treated)
- With random assignment, we can have the same number of never-takers in the treatment and control groups
- Hence,

$$LATE = \frac{ITT}{\text{share of compliers}}$$

16.1.4.2 Two-sided noncompliance

- Two-sided noncompliance is when in the sample, we have compliers, never-takers, and always-takers
- To estimate LATE, beyond excludability like in the One-sided noncompliance case, we need to assume that there is no defiers (i.e., monotonicity assumption) (this is excusable in practical studies)

$$LATE = \frac{ITT}{\text{share of compliers}}$$

16.1.5 Population vs. Sample Average Treatment Effects

See (Imai et al., 2008) for when the sample average treatment effect (SATE) diverges from the population average treatment effect (PATE).

To stay consistent, this section uses notations from (Imai et al., 2008)'s paper.

In a finite population N , we observe n observations ($N \gg n$), where half is in the control and half is in the treatment group.

With unknown data generating process, we have

$$I_i = \begin{cases} 1 & \text{if unit } i \text{ is in the sample} \\ 0 & \text{otherwise} \end{cases}$$

$$T_i = \begin{cases} 1 & \text{if unit } i \text{ is in the treatment group} \\ 0 & \text{if unit } i \text{ is in the control group} \end{cases}$$

$$\text{potential outcome} = \begin{cases} Y_i(1) & \text{if } T_i = 1 \\ Y_i(0) & \text{if } T_i = 0 \end{cases}$$

Observed outcome is

$$Y_i|I_i = 1 = T_i Y_i(1) + (1 - T_i) Y_i(0)$$

Since we can never observe both outcome for the same individual, the treatment effect is always unobserved for unit i

$$TE_i = Y_i(1) - Y_i(0)$$

Sample average treatment effect is

$$SATE = \frac{1}{n} \sum_{i \in \{I_i=1\}} TE_i$$

Population average treatment effect is

$$PATE = \frac{1}{N} \sum_{i=1}^N TE_i$$

Let X_i be observables and U_i be unobservables for unit i

The baseline estimator for SATE and PATE is

$$\begin{aligned} D &= \frac{1}{n/2} \sum_{i \in (I_i=1, T_i=1)} Y_i - \frac{1}{n/2} \sum_{i \in (I_i=1, T_i=0)} Y_i \\ &= \text{observed sample mean of the treatment group} \\ &\quad - \text{observed sample mean of the control group} \end{aligned}$$

Let Δ be the estimation error (deviation from the truth), under an additive model

$$Y_i(t) = g_t(X_i) + h_t(U_i)$$

The decomposition of the estimation error is

$$\begin{aligned} PATE - D &= \Delta = \Delta_S + \Delta_T \\ &= (PATE - SATE) + (SATE - D) \\ &= \text{sample selection} + \text{treatment imbalance} \\ &= (\Delta_{S_X} + \Delta_{S_U}) + (\Delta_{T_X} + \Delta_{T_U}) \\ &= (\text{selection on observed} + \text{selection on unobserved}) \\ &\quad + (\text{treatment imbalance in observed} + \text{unobserved}) \end{aligned}$$

16.1.5.1 Estimation Error from Sample Selection

Also known as sample selection error

$$\Delta_S = PATE - SATE = \frac{N-n}{N}(NATE - SATE)$$

where NATE is the non-sample average treatment effect (i.e., average treatment effect for those in the population but not in your sample):

$$NATE = \sum_{i \in (I_i=0)} \frac{TE_i}{N-n}$$

From the equation, to have zero sample selection error (i.e., $\Delta_S = 0$), we can either

- Get $N = n$ by redefining your sample as the population of interest
- $NATE = SATE$ (e.g., TE_i is constant over i in both your selected sample, and those in the population that you did not select)

Note

- When you have heterogenous treatment effects, **random sampling** can only warrant **sample selection bias**, not **sample selection error**.
- Since we can rarely know the true underlying distributions of the observables (X) and unobservables (U), we cannot verify whether the empirical distributions of your observables and unobservables for those in your sample is identical to that of your population (to reduce Δ_S). For special case,
 - Say you have census of your population, you can adjust for the observables X to reduce Δ_{S_X} , but still you cannot adjust your unobservables (U)
 - Say you are willing to assume TE_i is constant over
 - * X_i , then $\Delta_{S_X} = 0$
 - * U_i , then $\Delta_U = 0$

16.1.5.2 Estimation Error from Treatment Imbalance

Also known as treatment imbalance error

$$\Delta_T = SATE - D$$

$\Delta_T \rightarrow 0$ when treatment and control groups are balanced (i.e., identical empirical distributions) for both observables (X) and unobservables (U)

However, in reality, we can only readjust for observables, not unobservables.

	[Blocking][**Randomized Block Designs]**	[Matching] Matching Methods
Definition	Random assignment within strata based on pre-treatment observables	Dropping, repeating or grouping observations to balance covariates between the treatment and control group (Rubin, 1973)
Time	Before randomization of treatments	After randomization of treatments
What if the set of covariates used to adjust is irrelevant?	Nothing happens	In the worst case scenario (e.g., these variables are uncorrelated with the treatment assignment, but correlated with the post-treatment variables), matching induces bias that is greater than just using the unadjusted difference in means
Benefits	$\Delta_{T_X} = 0$ (no imbalance on observables). But we don't know its effect on unobservables imbalance (might reduce if the unobservables are correlated with the observables)	Reduce model dependence, bias, variance, mean-square error

16.1.6 Average Treatment Effects on the Treated and Control

Average Effect of treatment on the Treated (ATT) is

$$\begin{aligned} ATT &= E(Y_i(1) - Y_i(0)|D_i = 1) \\ &= E(Y_i(1)|D_i = 1) - E(Y_i(0)|D_i = 1) \end{aligned}$$

Average Effect of treatment on the Control (ATC) (i.e., the effect **would be** for those weren't treated) is

$$\begin{aligned} ATC &= E(Y_i(1) - Y_i(0)|D_i = 0) \\ &= E(Y_i(1)|D_i = 0) - E(Y_i(0)|D_i = 0) \end{aligned}$$

Under random assignment and full compliance,

$$ATE = ATT = ATC$$

Sample average treatment effect on the treated is

$$SATT = \frac{1}{n} \sum_i TE_i$$

where

- TE_i is the treatment effect for unit i
- n is the number of treated units in the sample
- i belongs the subset (i.e., sample) of the population of interest that is treated.

Population average treatment effect on the treated is

$$PATT = \frac{1}{N} \sum_i TE_i$$

where

- TE_i is the treatment effect for unit i
- N is the number of treated units in the population
- i belongs to the population of interest that is treated.

16.1.7 Quantile Average Treatment Effects

Instead of the middle point estimate (ATE), we can also understand the changes in the distribution the outcome variable due to the treatment.

Using quantile regression and more assumptions (Abadie et al., 2002; Chernozhukov and Hansen, 2005), we can have consistent estimate of quantile treatment effects (QTE), with which we can make inference regarding a given quantile.

16.1.8 Mediation Effects

With additional assumptions (i.e., sequential ignorability (Imai et al., 2010b; Bullock et al., 2010)), we can examine the mechanism of the treatment on the outcome.

Under the causal framework,

- the indirect effect of treatment via a mediator is called average causal mediation effect (ACME)
- the direct effect of treatment on outcome is the average direct effect (ADE)

More in the Mediation Section 29

16.1.9 Log-odds Treatment Effects

For binary outcome variable, we might be interested in the log-odds of success. See (Freedman, 2008) on how to estimate a consistent causal effect.

Alternatively, attributable effects (Rosenbaum, 2002) can also be appropriate for binary outcome.

A. EXPERIMENTAL DESIGN

Chapter 17

Experimental Design

- Randomized Control Trials (RCT) or Experiments have always been and are likely to continue in the future to be the holy grail of causal inference, because of
 - unbiased estimates
 - elimination of confounding factors on average (covariate imbalance is always possible. Hence, you want to do Rerandomization to achieve platinum standard set by (Tukey, 1993))
- RCT means you have two group treatment (or experimental) group and control group. Hence, as you introduce the treatment (your exogenous variable) to the treatment group, the only expected difference in the outcomes of the two group should be due to the treatment.
- Subjects from the same population will be **randomly assigned** to either treatment or control group. This random assignment give us the confidence that changes in the outcome variable will be due only the treatment, not any other source (variable).
- It can be easier for hard science to have RCT because they can introduce the treatment, and have control environments. But it's hard for social scientists because their subjects are usually human, and some treatment can be hard to introduce, or environments are uncontrollable. Hence, social scientists have to develop different tools (Quasi-experimental) to recover causal inference or to recreate the treatment and control group environment.
- With RCT, you can easily establish internal validity
- Even though random assignment is not the same thing as *ceteris paribus* (i.e., holding everything else constant), it should have the same effect (i.e., under random manipulation, *other things equal* can be observed, on average, across treatment and control groups).

Selection Problem

Assume we have

- binary treatment $D_i = (0, 1)$
- outcome of interest Y_i for individual i
 - Y_{0i} are those were **not treated**
 - Y_{1i} are those were **treated**

$$\text{Potential Outcome} = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

Then, what we observe in the outcome variable is

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$$

It's likely that Y_{1i} and Y_{0i} both have their own distributions (i.e., different treatment effect for different people). Since we can't see both outcomes for the same individual (unless we have a time machine), then we can only make inference regarding the average outcome of those who were treated and those who were not.

$$\begin{aligned} E[Y_i|D_i = 1] - E[Y_i|D_i = 0] &= (E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]) + (E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]) \\ &= (E[Y_{1i} - Y_{0i}|D_i = 1]) + (E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]) \end{aligned}$$

Observed difference in treatment = Average treatment effect on the treated + Selection bias

- **The average treatment effect** is the average between between a person who is treated and the same person (in another parallel universe) who is not treated
- **The selection bias** is the difference between those who were treated and those who weren't treated

With **random assignment** of treatment (D_i) under Experimental Design, we can have D_i independent of potential outcomes

$$\begin{aligned} E[Y_i|D_i = 1] - E[Y_i|D_i = 0] &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0)] \\ &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0)] \quad D_i \perp Y_i \\ &= E[Y_{1i} - Y_{0i}|D_i = 1] \\ &= E[Y_{1i} - Y_{0i}] \end{aligned}$$

Another representation under regression

Suppose that you know the effect is

$$Y_{1i} - Y_{0i} = \rho$$

The observed outcome variable (for an individual) can be rewritten as

$$\begin{aligned} Y_i &= E(Y_{0i}) + (Y_{1i} - Y_{0i})D_i + [Y_{0i} - E(Y_{0i})] \\ &= \alpha + \rho D_i + \eta_i \end{aligned}$$

where η_i = random variation of Y_{0i}

Hence, the conditional expectation of an individual outcome on treatment status is

$$\begin{aligned} E[Y_i|D_i = 1] &= \alpha + \rho + E[\eta_i|D_i = 1] \\ E[Y_i|D_i = 0] &= \alpha + E[\eta_i|D_i = 0] \end{aligned}$$

Thus,

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = \rho + E[\eta_i|D_i = 1] - E[\eta_i|D_i = 0]$$

where $E[\eta_i|D_i = 1] - E[\eta_i|D_i = 0]$ is the selection bias - correlation between the regression error term (η_i), and the regressor (D_i)

Under regression, we have

$$E[\eta_i|D_i = 1] - E[\eta_i|D_i = 0] = E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$$

which is the difference in outcomes between **those who weren't treated get treated** and **those who weren't treated stay untreated**

Say you have control variables (X_i), that is **uncorrelated with the treatment** (D_i), then you can include in your model, and it won't (in principle) affect your estimate of the treatment effect (ρ) with an added benefit of reducing the residual variance, which subsequently reduces the standard error of other estimates.

$$Y_i = \alpha + \rho D_i + X'_i \gamma + \eta_i$$

Examples in marketing:

- (Gordon et al., 2019)
- (Lewis and Rao, 2015)

17.1 Semi-random Experiment

Chicago Open Enrollment Program (Cullen et al., 2005)

- Students can apply to “choice” schools
- Many schools are oversubscribed (Demand > Supply)
- Resolve scarcity via random lotteries
- Non-random enrollment, we only have random lottery which mean the above

Let

$$\delta_j = E(Y_i|Enroll_{ij} = 1; Apply_{ij} = 1) - E(Y_i|Enroll_{ij} = 0; Apply_{ij} = 1)$$

and

$$\theta_j = E(Y_i|Win_{ij} = 1; Apply_{ij} = 1) - E(Y_i|Win_{ij} = 0; Apply_{ij} = 1)$$

Hence, we can clearly see that $\delta_j \neq \theta_j$ because you can only enroll, but you cannot ensure that you will win. Thus, **intention to treat is different from treatment effect.**

Non-random enrollment, we only have random lottery which means we can only estimate θ_j

To recover the true treatment effect, we can use

$$\delta_j = \frac{E(Y_i|W_{ij} = 1; A_{ij} = 1) - E(Y_i|W_{ij} = 0; A_{ij} = 1)}{P(Enroll_{ij} = 1|W_{ij} = 1; A_{ij} = 1) - P(Enroll_{ij} = 1|W_{ij} = 0; A_{ij} = 1)}$$

where

- δ_j = treatment effect
- W = Whether students win the lottery
- A = Whether student apply for the lottery
- i = application
- j = school

Say that we have

10 win

Number students	Type	Selection effect	Treatment effect	Total effect
1	Always Takers	+0.2	+1	+1.2
2	Compliers	0	+1	+1
7	Never Takers	-0.1	0	-0.1

10 lose

Number students	Type	Selection effect	Treatment effect	Total effect
1	Always Takers	+0.2	+1	+1.2
2	Compliers	0	0	0
7	Never Takers	-0.1	0	-0.1

Intent to treatment = Average effect of who you give option to choose

$$E(Y_i|W_{ij} = 1; A_{ij} = 1) = \frac{1 * (1.2) + 2 * (1) + 7 * (-0.1)}{10} \\ = 0.25$$

$$E(Y_i|W_{ij} = 0; A_{ij} = 1) = \frac{1 * (1.2) + 2 * (0) + 7 * (-0.1)}{10} \\ = 0.05$$

Hence,

$$\text{Intent to treatment} = 0.25 - 0.05 = 0.2$$

$$\text{Treatment effect} = 1$$

$$P(\text{Enroll}_{ij} = 1|W_{ij} = 1; A_{ij} = 1) = \frac{1+2}{10} = 0.3$$

$$P(\text{Enroll}_{ij} = 1|W_{ij} = 0; A_{ij} = 1) = \frac{1}{10} = 0.1$$

$$\text{Treatment effect} = \frac{0.2}{0.3 - 0.1} = 1$$

After knowing how to recover the treatment effect, we turn our attention to the main model

$$Y_{ia} = \delta W_{ia} + \lambda L_{ia} + e_{ia}$$

where

- W = whether a student wins a lottery
- L = whether student enrolls in the lottery
- δ = intent to treat

Hence,

- Conditional on lottery, the δ is valid
- But without lottery, your δ is not random
- Winning and losing are only identified within lottery
- Each lottery has multiple entries. Thus, we can have within estimator

We can also include other control variables ($X_i\theta$)

$$Y_{ia} = \delta_1 W_{ia} + \lambda_1 L_{ia} + X_i\theta + u_{ia}$$

$$E(\delta) = E(\delta_1)$$

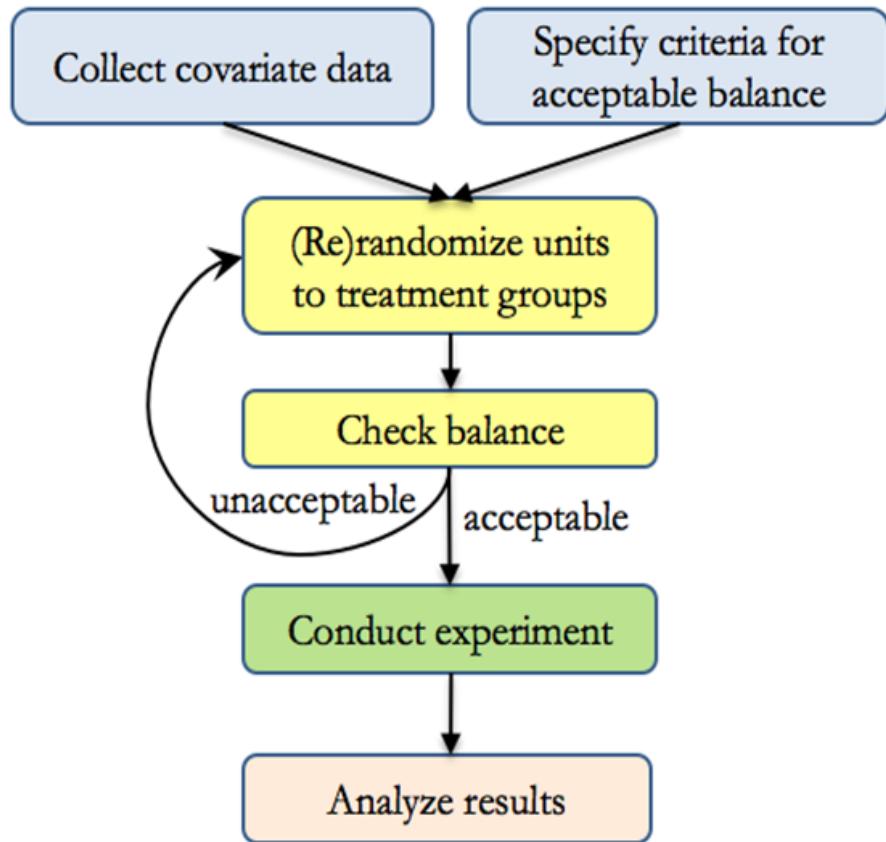
$E(\lambda) \neq E(\lambda_1)$ because choosing a lottery is not random

Including ($X_i\theta$) just shifts around control variables (i.e., reweighting of lottery), which would not affect your treatment effect $E(\delta)$

17.2 Rerandomization

- Since randomization only balances baseline covariates on average, imbalance in covariates due to random chance can still happen.
- In case that you have a “bad” randomization (i.e., imbalance for important baseline covariates), (Morgan and Rubin, 2012) introduce the idea of rerandomization.

- Rerandomization is checking balance during the randomization process (before the experiment), to eliminate bad allocation (i.e., those with unacceptable balance).
- The greater the number of variables, the greater the likelihood that at least one covariate would be imbalanced across treatment groups.
 - Example: For 10 covariates, the probability of a significant difference at $\alpha = .05$ for at least one covariate is $1 - (1 - .05)^{10} = 0.4 = 40\%$
- Rerandomization increase treatment effect estimate precision if the covariates are correlated with the outcome.
 - Improvement in precision for treatment effect estimate depends on (1) improvement in covariate balance and (2) correlation between covariates and the outcome.
- You also need to take into account rerandomization into your analysis when making inference.
- Rerandomization is equivalent to increasing our sample size.
- Alternatives include
 - Stratified randomization (Johansson and Schultzberg, 2022)
 - Matched randomization (Greevy, 2004; Kapelner and Krieger, 2014)
 - Minimization (Pocock and Simon, 1975; cor, 1976)



Graph from USC Schaeffer Center

Rerandomization Criterion

- Acceptable randomization is based on a function of covariate matrix \mathbf{X} and vector of treatment assignments \mathbf{W}

$$W_i = \begin{cases} 1 & \text{if treated} \\ 0 & \text{if control} \end{cases}$$

- Mahalanobis Distance, M , can be used as criteria for acceptable balance

Let M be the multivariate distance between groups means

$$M = (\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C)' \text{cov}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C)^{-1} (\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C) = \left(\frac{1}{n_T} + \frac{1}{n_C} \right)^{-1} (\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C)' \text{cov}(\mathbf{X})^{-1} (\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C)$$

With large sample size and “pure” randomization $M \sim \chi_k^2$ where k is the number of covariates to be balanced

Then let p_a be the probability of accepting a randomization. Choosing appropriate p_a is a tradeoff between balance and time.

Then the rule of thumb is rerandomize when $M > a$

Chapter 18

Sampling

18.1 Simple Sampling

Simple (random) Sampling

```
library(dplyr)
iris_df <- iris
set.seed(1)
sample_n(iris_df, 10)
#>   Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
#> 1         5.8        2.7       4.1        1.0 versicolor
#> 2         6.4        2.8       5.6        2.1 virginica
#> 3         4.4        3.2       1.3        0.2   setosa
#> 4         4.3        3.0       1.1        0.1   setosa
#> 5         7.0        3.2       4.7        1.4 versicolor
#> 6         5.4        3.0       4.5        1.5 versicolor
#> 7         5.4        3.4       1.7        0.2   setosa
#> 8         7.6        3.0       6.6        2.1 virginica
#> 9         6.1        2.8       4.7        1.2 versicolor
#> 10        4.6        3.4       1.4        0.3   setosa
```

```

#> [75] 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 1 0
#> [112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
#> [149] 0 0

# Simple random sampling without replacement (sequential method)
srswor1(10, length(iris_df$id))
#> [1] 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
#> [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
#> [75] 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
#> [112] 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0
#> [149] 0 0

# Simple random sampling with replacement
srswr(10, length(iris_df$id))
#> [1] 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 1 1 0 0 0 0
#> [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
#> [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
#> [112] 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
#> [149] 0 0

```

```
library(survey)
data("api")
srs_design <- svydesign(data = apistrat,
                        weights = ~pw,
                        fpc = ~fpc,
                        id = ~1)
```

```
library(sampler)
rsamp(albania,
      n = 260,
      over = 0.1, # desired oversampling proportion
      rep = F)
```

Identify missing points between sample and collected data

```
alsample <- rsamp(df = albania, 544)
alreceived <- rsamp(df = alsample, 390)
rmissing(sampdf = alsample,
          colldf = alreceived,
          col_name = qvKod)
```

18.2 Stratified Sampling

A stratum is a subset of the population that has at least one common characteristic.

Steps:

1. Identify relevant strata and their representation in the population.
2. Randomly sample to select a sufficient number of subjects from each stratum.

Stratified sampling reduces sampling error.

```
library(dplyr)
# by number of rows
sample_iris <- iris %>%
  group_by(Species) %>%
  sample_n(5)
sample_iris
#> # A tibble: 15 x 5
#> # Groups:   Species [3]
#>   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
#>       <dbl>     <dbl>      <dbl>      <dbl> <fct>
#> 1       4.4      3          1.3      0.2  setosa
#> 2       5.2      3.5        1.5      0.2  setosa
#> 3       5.1      3.8        1.5      0.3  setosa
#> 4       5.2      3.4        1.4      0.2  setosa
#> 5       4.5      2.3        1.3      0.3  setosa
#> 6       5.5      2.5        4        1.3  versicolor
#> 7       7         3.2        4.7      1.4  versicolor
#> 8       6.7      3          5        1.7  versicolor
#> 9       6.1      2.9        4.7      1.4  versicolor
#> 10      5.5      2.4        3.8      1.1  versicolor
#> 11      6.4      2.7        5.3      1.9  virginica
#> 12      6.4      2.8        5.6      2.1  virginica
#> 13      6.4      3.2        5.3      2.3  virginica
#> 14      6.8      3.2        5.9      2.3  virginica
#> 15      7.2      3.6        6.1      2.5  virginica

# by fraction
sample_iris <- iris %>%
  group_by(Species) %>%
  sample_frac(size = .15)
sample_iris
#> # A tibble: 24 x 5
```

```
#> # Groups: Species [3]
#>   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
#>   <dbl>       <dbl>       <dbl>       <dbl> <fct>
#> 1 5.5         4.2         1.4         0.2 setosa
#> 2 5            3           1.6         0.2 setosa
#> 3 5.2         4.1         1.5         0.1 setosa
#> 4 4.6         3.1         1.5         0.2 setosa
#> 5 5.1         3.7         1.5         0.4 setosa
#> 6 4.8         3.4         1.9         0.2 setosa
#> 7 5.1         3.3         1.7         0.5 setosa
#> 8 5.5         3.5         1.3         0.2 setosa
#> 9 5            2.3         3.3         1     versicolor
#> 10 5.6        2.9         3.6         1.3 versicolor
#> # ... with 14 more rows
```

```
library(sampler)
# Stratified sample using proportional allocation without replacement
ssamp(df=albania, n=360, strata=qarku, over=0.1)
#> # A tibble: 395 x 45
#>   qarku Q_ID bashkia      BAS_ID zaz    njesiaAdministr~ COM_ID qvKod zgjedhes
#>   <fct> <int> <fct>      <int> <fct> <fct> <int> <fct> <int>
#> 1 Berat    1 Berat      11 ZAZ ~ "Berat "    1101 "\"3~    558
#> 2 Berat    1 Berat      11 ZAZ ~ "Berat "    1101 "\"3~    815
#> 3 Berat    1 Berat      11 ZAZ ~ "Singe"    1108 "\"3~    419
#> 4 Berat    1 Kucove     13 ZAZ ~ "Lumas"    1104 "\"3~    237
#> 5 Berat    1 Kucove     13 ZAZ ~ "Kucove"   1201 "\"3~    562
#> 6 Berat    1 Skrapar    17 ZAZ ~ "Corovode"  1303 "\"3~    829
#> 7 Berat    1 Berat      11 ZAZ ~ "Roshnik"  1107 "\"3~    410
#> 8 Berat    1 Ura Vajgurore 19 ZAZ ~ "Ura Vajgurore" 1110 "\"3~    708
#> 9 Berat    1 Kucove     13 ZAZ ~ "Perondi"   1203 "\"3~    835
#> 10 Berat   1 Kucove     13 ZAZ ~ "Kucove"   1201 "\"3~    907
#> # ... with 385 more rows, and 36 more variables: meshkuj <int>, femra <int>,
#> #   totalSeats <int>, vendndodhja <fct>, ambienti <fct>, totalVoters <int>,
#> #   femVoters <int>, maleVoters <int>, unusedBallots <int>,
#> #   damagedBallots <int>, ballotsCast <int>, invalidVotes <int>,
#> #   validVotes <int>, lsi <int>, ps <int>, pdk <int>, sfida <int>, pr <int>,
#> #   pd <int>, pbdksh <int>, adk <int>, psd <int>, ad <int>, frd <int>,
#> #   pds <int>, pdiu <int>, aak <int>, mega <int>, pksh <int>, apd <int>, ...
```

Identify number of missing points by strata between sample and collected data

```
alsample <- rsamp(df = albania, 544)
alreceived <- rsamp(df = alsample, 390)
smissing(
```

```

    sampdf = alsample,
    colddf = alreceived,
    strata = qarku,
    col_name = qvKod
)

```

18.3 Unequal Probability Sampling

```

UPbrewer()
UPmaxentropy()
UPmidzuno()
UPmidzunopi2()
UPmultinomial()
UPpivotal()
UPrandompivotal()
UPpoisson()
UPSampford()
UPsystematic()
UPrandomsystematic()
UPsystematicpi2()
UPtille()
UPtillepi2()

```

18.4 Balanced Sampling

- Purpose: to get the same means in the population and the sample for all the auxiliary variables
- Balanced sampling is different from purposive selection

Balancing equations

$$\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k$$

where \mathbf{x}_k is a vector of auxiliary variables

18.4.1 Cube

- flight phase

- landing phase

```
samplecube()
fastflightcube()
landingcube()
```

18.4.2 Stratification

- Try to replicate the population based on the original multivariate histogram

```
library(survey)
data("api")
srs_design <- svydesign(data = apistrat,
                        weights = ~pw,
                        fpc = ~fpc,
                        strata = ~stype,
                        id = ~1)
```

```
balancedstratification()
```

18.4.3 Cluster

```
library(survey)
data("api")
srs_design <- svydesign(data = apiclus1,
                        weights = ~pw,
                        fpc = ~fpc,
                        id = ~dnum)
```

```
balancedcluster()
```

18.4.4 Two-stage

```
library(survey)
data("api")
srs_design <- svydesign(data = apiclus2,
                        fpc = ~fpc1 + fpc2,
                        id = ~ dnum + snum)
```

```
balancedtwostage()
```


Chapter 19

Analysis of Variance (ANOVA)

ANOVA is using the same underlying mechanism as linear regression. However, the angle that ANOVA chooses to look at is slightly different from the traditional linear regression. It can be more useful in the case with **qualitative variables** and **designed experiments**.

Experimental Design

- **Factor:** explanatory or predictor variable to be studied in an investigation
- **Treatment** (or Factor Level): “value” of a factor applied to the experimental unit
- **Experimental Unit:** person, animal, piece of material, etc. that is subjected to treatment(s) and provides a response
- **Single Factor Experiment:** one explanatory variable considered
- **Multifactor Experiment:** more than one explanatory variable
- **Classification Factor:** A factor that is not under the control of the experimenter (observational data)
- **Experimental Factor:** assigned by the experimenter

Basics of experimental design:

- Choices that a statistician has to make:
 - set of treatments
 - set of experimental units
 - treatment assignment (selection bias)
 - measurement (measurement bias, blind experiments)

- Advancements in experimental design:
 1. **Factorial Experiments:**
consider multiple factors at the same time (interaction)
 2. **Replication:** repetition of experiment
 - assess mean squared error
 - control over precision of experiment (power)
 3. **Randomization**
 - Before R.A. Fisher (1900s), treatments were assigned systematically or subjectively
 - randomization: assign treatments to experimental units at random, which averages out systematic effects that cannot be controlled by the investigator
 4. **Local control:** Blocking or Stratification
 - Reduce experimental errors and increase power by placing restrictions on the randomization of treatments to experimental units.

Randomization may also eliminate correlations due to time and space.

19.1 Completely Randomized Design (CRD)

Treatment factor A with $a \geq 2$ treatment levels. Experimental units are randomly assigned to each treatment. The number of experimental units in each group can be

- equal (balanced): n
- unequal (unbalanced): n_i for the i -th group ($i = 1, \dots, a$).

The total sample size is $N = \sum_{i=1}^a n_i$

Possible assignments of units to treatments are $k = \frac{N!}{n_1!n_2!\dots n_a!}$

Each has probability $1/k$ of being selected. Each experimental unit is measured with a response Y_{ij} , in which j denotes unit and i denotes treatment.

Treatment

	1	2	...	a
Y_{11}	Y_{21}	...	Y_{a1}	
Y_{12}	

	1	2	...	a
Sample Mean	$\bar{Y}_{1.}$	$\bar{Y}_{2.}$	\dots	$\bar{Y}_{a.}$
Sample SD	s_1	s_2	\dots	s_a

$$\text{where } \bar{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$$

And the grand mean is $\bar{Y}_{..} = \frac{1}{N} \sum_i \sum_j Y_{ij}$

19.1.1 Single Factor Fixed Effects Model

also known as Single Factor (One-Way) ANOVA or ANOVA Type I model.

Partitioning the Variance

The total variability of the Y_{ij} observation can be measured as the deviation of Y_{ij} around the overall mean $\bar{Y}_{..}$: $Y_{ij} - \bar{Y}_{..}$

This can be rewritten as:

$$\begin{aligned} Y_{ij} - \bar{Y}_{..} &= Y_{ij} - \bar{Y}_{..} + \bar{Y}_{i.} - \bar{Y}_{i.} \\ &= (\bar{Y}_{i.} - \bar{Y}_{..}) + (Y_{ij} - \bar{Y}_{i.}) \end{aligned}$$

where

- the first term is the *between* treatment differences (i.e., the deviation of the treatment mean from the overall mean)
- the second term is *within* treatment differences (i.e., the deviation of the observation around its treatment mean)

$$\sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2 = \sum_i n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 + \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})^2$$

$$SSTO = SSTR + SSE$$

$$\text{total SS} = \text{treatment SS} + \text{error SS}$$

$$(N - 1) \text{ d.f.} = (a - 1) \text{ d.f.} + (N - a) \text{ d.f.}$$

we lose a d.f. for the total corrected SSTO because of the estimation of the mean ($\sum_i \sum_j (Y_{ij} - \bar{Y}_{..}) = 0$)

And, for the SSTR $\sum_i n_i (\bar{Y}_{i.} - \bar{Y}_{..}) = 0$

Accordingly, $MSTR = \frac{SST}{a-1}$ and $MSR = \frac{SSE}{N-a}$

ANOVA Table

Source of Variation	SS	df	MS
Between Treatments	$\sum_i n_i (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2$	a-1	SSTR/(a-1)
Error (within treatments)	$\sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2$	N-a	SSE/(N-a)
Total (corrected)	$\sum_i n_i (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2$	N-1	

Linear Model Explanation of ANOVA

19.1.1.1 Cell means model

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

where

- Y_{ij} response variable in j-th subject for the i-th treatment
- μ_i : parameters (fixed) representing the unknown population mean for the i-th treatment
- ϵ_{ij} independent $N(0, \sigma^2)$ errors
- $E(Y_{ij}) = \mu_i$ $var(Y_{ij}) = var(\epsilon_{ij}) = \sigma^2$
- All observations have the same variance

Example:

$a = 3$ (3 treatments) $n_1 = n_2 = n_3 = 2$

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}$$

$\mathbf{y} = \mathbf{X} +$

$X_{k,ij} = 1$ if the k-th treatment is used

$X_{k,ij} = 0$ Otherwise

Note: no intercept term.

$$\begin{aligned}\mathbf{b} &= \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} \\ &= \begin{bmatrix} n_1 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & n_3 \end{bmatrix}^{-1} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} \\ &= \begin{bmatrix} \bar{Y}_1 \\ \bar{Y}_2 \\ \bar{Y}_3 \end{bmatrix}\end{aligned}\tag{19.1}$$

is the BLUE (best linear unbiased estimator) for $\beta = [\mu_1 \mu_2 \mu_3]'$

$$E(\mathbf{b}) = \beta$$

$$var(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \begin{bmatrix} 1/n_1 & 0 & 0 \\ 0 & 1/n_2 & 0 \\ 0 & 0 & 1/n_3 \end{bmatrix}$$

$$var(b_i) = var(\hat{\mu}_i) = \sigma^2/n_i \text{ where } \mathbf{b} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

$$\begin{aligned}MSE &= \frac{1}{N-a} \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})^2 \\ &= \frac{1}{N-a} \sum_i [(n_i - 1) \frac{\sum_i (Y_{ij} - \bar{Y}_{i.})^2}{n_i - 1}] \\ &= \frac{1}{N-a} \sum_i (n_i - 1)s_i^2\end{aligned}$$

We have $E(s_i^2) = \sigma^2$

$$E(MSE) = \frac{1}{N-a} \sum_i (n_i - 1)\sigma^2 = \sigma^2$$

Hence, MSE is an unbiased estimator of σ^2 , regardless of whether the treatment means are equal or not.

$$E(MSTR) = \sigma^2 + \frac{\sum_i n_i (\mu_i - \mu_{\cdot})^2}{a-1}$$

$$\text{where } \mu_{\cdot} = \frac{\sum_{i=1}^a n_i \mu_i}{\sum_{i=1}^a n_i}$$

If all treatment means are equals ($=\mu_{\cdot}$), $E(MSTR) = \sigma^2$.

Then we can use an F-test for the equality of all treatment means:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_a$$

$$H_a : \text{not all } \mu_i \text{ are equal}$$

$$F = \frac{MSTR}{MSE}$$

where large values of F support H_a (since MSTR will tend to exceed MSE when H_a holds)

and F near 1 support H_0 (upper tail test)

Equivalently, when H_0 is true, $F \sim f_{(a-1, N-a)}$

- If $F \leq f_{(a-1, N-a; 1-\alpha)}$, we cannot reject H_0
- If $F \geq f_{(a-1, N-a; 1-\alpha)}$, we reject H_0

Note: If $a = 2$ (2 treatments), F-test = two sample t-test

19.1.1.2 Treatment Effects (Factor Effects)

Besides Cell means model, we have another way to formalize one-way ANOVA:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

where

- Y_{ij} is the j-th response for the i-th treatment
- τ_i i-th treatment effect
- μ constant component, common to all observations
- ϵ_{ij} independent random errors $\sim N(0, \sigma^2)$

For example, $a = 3$, $n_1 = n_2 = n_3 = 2$

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix} \quad (19.2)$$

$\mathbf{y} = \mathbf{X} +$

However,

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \sum_i n_i & n_1 & n_2 & n_3 \\ n_1 & n_1 & 0 & 0 \\ n_2 & 0 & n_2 & 0 \\ n_3 & 0 & 0 & n_3 \end{pmatrix}$$

is **singular** thus does not exist, \mathbf{b} is insolvable (infinite solutions)

Hence, we have to impose restrictions on the parameters to a model matrix \mathbf{X} of full rank.

Whatever restriction we use, we still have:

$E(Y_{ij}) = \mu + \tau_i = \mu_i = \text{mean response for } i\text{-th treatment}$

19.1.1.2.1 Restriction on sum of tau $\sum_{i=1}^a \tau_i = 0$

implies

$$\mu = \mu + \frac{1}{a} \sum_{i=1}^a (\mu + \tau_i)$$

is the average of the treatment mean (grand mean) (overall mean)

$$\begin{aligned} \tau_i &= (\mu + \tau_i) - \mu = \mu_i - \mu \\ &= \text{treatment mean} - \text{grand mean} \\ &= \text{treatment effect} \end{aligned}$$

$$\tau_a = -\tau_1 - \tau_2 - \dots - \tau_{a-1}$$

Hence, the mean for the a -th treatment is

$$\mu_a = \mu + \tau_a = \mu - \tau_1 - \tau_2 - \dots - \tau_{a-1}$$

Hence, the model need only “a” parameters:

$$\mu, \tau_1, \tau_2, \dots, \tau_{a-1}$$

Equation (19.2) becomes

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix} \quad (19.3)$$

$$\mathbf{y} = \mathbf{X} +$$

where $\beta \equiv [\mu, \tau_1, \tau_2]'$

Equation (19.1) with $\sum_i \tau_i = 0$ becomes

$$\begin{aligned} \mathbf{b} &= \begin{bmatrix} \hat{\mu} \\ \hat{\tau}_1 \\ \hat{\tau}_2 \end{bmatrix} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} \\ &= \begin{bmatrix} \sum_i n_i & n_1 - n_3 & n_2 - n_3 \\ n_1 - n_3 & n_1 + n_3 & n_3 \\ n_2 - n_3 & n_3 & n_2 - n_3 \end{bmatrix}^{-1} \begin{bmatrix} Y_{..} \\ Y_{1.} - Y_{3.} \\ Y_{2.} - Y_{3.} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{3} \sum_{i=1}^3 \bar{Y}_i \\ \bar{Y}_{1.} - \frac{1}{3} \sum_{i=1}^3 \bar{Y}_i \\ \bar{Y}_{2.} - \frac{1}{3} \sum_{i=1}^3 \bar{Y}_i \end{bmatrix} \\ &= \begin{bmatrix} \hat{\mu} \\ \hat{\tau}_1 \\ \hat{\tau}_2 \end{bmatrix} \end{aligned}$$

and $\hat{\tau}_3 = -\hat{\tau}_1 - \hat{\tau}_2 = \bar{Y}_3 - \frac{1}{3} \sum_i \bar{Y}_i$.

19.1.1.2.2 Restriction on first tau In R, `lm()` uses the restriction $\tau_1 = 0$. For the previous example, for $n_1 = n_2 = n_3 = 2$, and $\tau_1 = 0$. Then the treatment means can be written as:

$$\mu_1 = \mu + \tau_1 = \mu + 0 = \mu \mu_2 = \mu + \tau_2 \mu_3 = \mu + \tau_3$$

Hence, μ is the mean response for the first treatment

In the matrix form,

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_2 \\ \tau_3 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}$$

$\mathbf{y} = \mathbf{X} +$

$$\beta = [\mu, \tau_2, \tau_3]'$$

$$\begin{aligned} \mathbf{b} &= \begin{bmatrix} \hat{\mu} \\ \hat{\tau}_2 \\ \hat{\tau}_3 \end{bmatrix} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} \\ &= \begin{bmatrix} \sum_i n_i & n_2 & n_3 \\ n_2 & n_2 & 0 \\ n_3 & 0 & n_3 \end{bmatrix}^{-1} \begin{bmatrix} \bar{Y}_{..} \\ \bar{Y}_{2..} \\ \bar{Y}_{3..} \end{bmatrix} \\ &= \begin{bmatrix} \bar{Y}_{1..} \\ \bar{Y}_{2..} - \bar{Y}_{1..} \\ \bar{Y}_{3..} - \bar{Y}_{1..} \end{bmatrix} \end{aligned}$$

$$E(\mathbf{b}) = \beta = \begin{bmatrix} \mu \\ \tau_2 \\ \tau_3 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 - \mu_1 \\ \mu_3 - \mu_1 \end{bmatrix}$$

$$var(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}var(\hat{\mu}) = var(\bar{Y}_{1..}) = \sigma^2/n_1 var(\hat{\tau}_2) = var(\bar{Y}_{2..} - \bar{Y}_{1..}) = \sigma^2/n_2 + \sigma^2/n_1 var(\hat{\tau}_3) = var(\bar{Y}_{3..} - \bar{Y}_{1..}) = \sigma^2$$

Note For all three parameterization, the ANOVA table is the same

- Model 1: $Y_{ij} = \mu_i + \epsilon_{ij}$
- Model 2: $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$ where $\sum_i \tau_i = 0$
- Model 3: $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$ where $\tau_1 = 0$

All models have the same calculation for \hat{Y} as

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{P}\mathbf{Y} = \mathbf{X}\mathbf{b}$$

ANOVA Table

Source of Variation	SS	df	MS	F
Between Treatments	$\sum_i n_i (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2 = \mathbf{Y}'(\mathbf{P} - \mathbf{P}_1)\mathbf{Y}$	a-1	$\frac{SSTR}{a-1}$	$\frac{MSTR}{MSE}$
Error (within treatments)	$\sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2 = \mathbf{e}'\mathbf{e}$	N-a	$\frac{SSE}{N-a}$	
Total (corrected)	$\sum_i n_i (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2 = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{P}_1\mathbf{Y}$	N-1		

where $\mathbf{P}_1 = \frac{1}{n}\mathbf{J}$

The F-statistic here has (a-1,N-a) degrees of freedom, which gives the same value for all three parameterization, but the hypothesis test is written a bit different:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_a \quad H_0 : \mu + \tau_1 = \mu + \tau_2 = \dots = \mu + \tau_a$$

The F-test here serves as a preliminary analysis, to see if there is any difference at different factors. For more in-depth analysis, we consider different testing of treatment effects.

19.1.1.3 Testing of Treatment Effects

- A Single Treatment Mean μ_i
- A Differences Between Treatment Means
- A Contrast Among Treatment Means
- A Linear Combination of Treatment Means

19.1.1.3.1 Single Treatment Mean We have $\hat{\mu}_i = \bar{Y}_{i\cdot}$ where

- $E(\bar{Y}_{i\cdot}) = \mu_i$
- $var(\bar{Y}_{i\cdot}) = \sigma^2/n_i$ estimated by $s^2(\bar{Y}_{i\cdot}) = MSE/n_i$

Since $\frac{\bar{Y}_{i\cdot} - \mu_i}{s(\bar{Y}_{i\cdot})} \sim t_{N-a}$ and the confidence interval for μ_i is $\bar{Y}_{i\cdot} \pm t_{1-\alpha/2; N-a} s(\bar{Y}_{i\cdot})$, then we can do a t-test for the means difference with some constant c

$$H_0 : \mu_i = c \quad H_1 : \mu_i \neq c$$

where

$$T = \frac{\bar{Y}_{i\cdot} - c}{s(\bar{Y}_{i\cdot})}$$

follows t_{N-a} when H_0 is true.

If $|T| > t_{1-\alpha/2; N-a}$, we can reject H_0

19.1.1.3.2 Differences Between Treatment Means Let $D = \mu_i - \mu'_i$, also known as **pairwise comparison**

D can be estimated by $\hat{D} = \bar{Y}_i - \bar{Y}'_i$ is unbiased ($E(\hat{D}) = \mu_i - \mu'_i$)

Since \bar{Y}_i and \bar{Y}'_i are independent, then

$$\text{var}(\hat{D}) = \text{var}(\bar{Y}_i) + \text{var}(\bar{Y}'_i) = \sigma^2(1/n_i + 1/n'_i)$$

can be estimated with

$$s^2(\hat{D}) = \text{MSE}(1/n_i + 1/n'_i)$$

With the single treatment inference,

$$\frac{\hat{D} - D}{s(\hat{D})} \sim t_{N-a}$$

hence,

$$\hat{D} \pm t_{(1-\alpha/2; N-a)} s(\hat{D})$$

Hypothesis tests:

$$H_0 : \mu_i = \mu'_i \quad H_a : \mu_i \neq \mu'_i$$

can be tested by the following statistic

$$T = \frac{\hat{D}}{s(\hat{D})} \sim t_{1-\alpha/2; N-a}$$

reject H_0 if $|T| > t_{1-\alpha/2; N-a}$

19.1.1.3.3 Contrast Among Treatment Means generalize the comparison of two means, we have **contrasts**

A contrast is a linear combination of treatment means:

$$L = \sum_{i=1}^a c_i \mu_i$$

where each c_i is non-random constant and sum to 0:

$$\sum_{i=1}^a c_i = 0$$

An unbiased estimator of a contrast L is

$$\hat{L} = \sum_{i=1}^a c_i \bar{Y}_i.$$

and $E(\hat{L}) = L$. Since the \bar{Y}_i , $i = 1, \dots, a$ are independent.

$$var(\hat{L}) = var\left(\sum_{i=1}^a c_i \bar{Y}_i\right) = \sum_{i=1}^a var(c_i \bar{Y}_i) = \sum_{i=1}^a c_i^2 var(\bar{Y}_i) = \sum_{i=1}^a c_i^2 \sigma^2 / n_i = \sigma^2 \sum_{i=1}^a c_i^2 / n_i$$

Estimation of the variance:

$$s^2(\hat{L}) = MSE \sum_{i=1}^a \frac{c_i^2}{n_i}$$

\hat{L} is normally distributed (since it is a linear combination of independent normal random variables).

Then, since SSE/σ^2 is χ^2_{N-a}

$$\frac{\hat{L} - L}{s(\hat{L})} \sim t_{N-a}$$

A $1 - \alpha$ confidence limits are given by

$$\hat{L} \pm t_{1-\alpha/2; N-a} s(\hat{L})$$

Hypothesis testing

$$H_0 : L = 0 \quad H_a : L \neq 0$$

with

$$T = \frac{\hat{L}}{s(\hat{L})}$$

reject H_0 if $|T| > t_{1-\alpha/2; N-a}$

19.1.1.3.4 Linear Combination of Treatment Means just like contrast $L = \sum_{i=1}^a c_i \mu_i$ but no restrictions on the c_i coefficients.

Tests of a single treatment mean, two treatment means, and contrasts can all be considered from the same perspective.

$$H_0 : \sum c_i \mu_i = c \quad H_a : \sum c_i \mu_i \neq c$$

The test statistics (t-stat) can be considered equivalently as F-tests; $F = (T)^2$ where $F \sim F_{1, N-a}$. Since the numerator degrees of freedom is always 1 in these cases, we refer to them as single-degree-of-freedom tests.

Multiple Contrasts

To test simultaneously $k \geq 2$ contrasts, let T_1, \dots, T_k be the t-stats. The joint distribution of these random variables is a multivariate t-distribution (the tests are dependent since they're based on the same data).

Limitations for comparing multiple contrasts:

1. The confidence coefficient $1 - \alpha$ only applies to a particular estimate, not a series of estimates; similarly, the Type I error rate, α , applies to a particular test, not a series of tests. Example: 3 t-tests at $\alpha = 0.05$, if tests are independent (which they are not), $0.95^3 = 0.857$ (thus $\alpha = 0.143$ not 0.05)
2. The confidence coefficient $1 - \alpha$ and significance level α are appropriate only if the test was not suggested by the data.
 - often, the results of an experiment suggest important (i.e.g, potential significant) relationships.
 - the process of studying effects suggested by the data is called **data snooping**

Multiple Comparison Procedures:

- Tukey
- Scheffe
- Bonferroni

19.1.1.3.4.1 Tukey All pairwise comparisons of factor level means. All pairs $D = \mu_i - \mu'_i$ or all tests of the form:

$$H_0 : \mu_i - \mu'_i = 0 \quad H_a : \mu_i - \mu'_i \neq 0$$

- When all sample sizes are equal ($n_1 = n_2 = \dots = n_a$) then the Tukey method family confidence coefficient is exactly $1 - \alpha$ and the significance level is exactly α
- When the sample sizes are not equal, the family confidence coefficient is greater than $1 - \alpha$ (i.e., the significance level is less than α) so the test is **conservative**
- Tukey considers the **studentized range distribution**. If we have Y_1, \dots, Y_r , observations from a normal distribution with mean α and variance σ^2 . Define:

$$w = \max(Y_i) - \min(Y_i)$$

as the range of the observations. Let s^2 be an estimate of σ^2 with v degrees of freedom. Then,

$$q(r, v) = \frac{w}{s}$$

is called the studentized range. The distribution of q uses a special table.

Notes

- when we are not interested in testing all pairwise comparisons the confidence coefficient for the family of comparisons under consideration will be greater than $1 - \alpha$ (with the significance level less than α)
- Tukey can be used for “data snooping” as long as the effects to be studied on the basis of preliminary data analysis are pairwise comparisons.

19.1.1.3.4.2 Scheffe This method applies when the family of interest is the set of possible contrasts among the treatment means:

$$L = \sum_{i=1}^a c_i \mu_i$$

where $\sum_{i=1}^a c_i = 0$

That is, the family of all possible contrasts L or

$$H_0 : L = 0 \quad H_a : L \neq 0$$

The family confidence level for the Scheffe procedure is exactly $1 - \alpha$ (i.e., significance level = α) whether the sample sizes are equal or not.

For simultaneous confidence intervals,

$$\hat{L} \pm Ss(\hat{L})$$

where $\hat{L} = \sum c_i \bar{Y}_{i.}$, $s^2(\hat{L}) = MSE \sum c_i^2/n_i$ and $S^2 = (a-1)f_{(1-\alpha; a-1, N-a)}$

The Scheffe procedure considers

$$F = \frac{\hat{L}^2}{(a-1)s^2(\hat{L})}$$

where we reject H_0 at the family significance level α if $F > f_{(1-\alpha; a-1, N-a)}$

Note

- Since applications of the Scheffe never involve all conceivable contrasts, the **finite family** confidence coefficient will be larger than $1 - \alpha$, so $1 - \alpha$ is a lower bound. Thus, people often consider a larger α (e.g., 90% confidence interval)
- Scheffe can be used for “data scooping” since the family of statements contains all possible contrasts.
- If only pairwise comparisons are to be considered, The Tukey procedure gives narrower confidence limits.

19.1.1.3.4.3 Bonferroni Applicable whether the sample sizes are equal or unequal.

For the confidence intervals,

$$\hat{L} \pm Bs(\hat{L})$$

where $B = t_{(1-\alpha/(2g); N-a)}$ and g is the number of comparisons in the family.

Hypothesis testing

$$H_0 : L = 0 \quad H_a : L \neq 0$$

Let $T = \frac{\hat{L}}{s(\hat{L})}$ and reject H_0 if $|T| > t_{1-\alpha/(2g), N-a}$

Notes

- If all pairwise comparisons are of interest, the Tukey procedure is superior (narrower confidence intervals). If not, Bonferroni may be better.
- Bonferroni is better than Scheffe when the number of contrasts is about the same as the treatment levels (or less).
- Recommendation: compute all threes and pick the smallest.
- Bonferroni can't be used for **data snooping**

19.1.1.3.4.4 Fisher's LSD does not control for family error rate
use t-stat for testing

$$H_0 : \mu_i = \mu_j$$

t-stat

$$t = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_j})}}$$

19.1.1.3.4.5 Newman-Keuls Do not recommend using this test since it has less power than ANOVA.

19.1.1.3.5 Multiple comparisons with a control

19.1.1.3.5.1 Dunnett We have a groups where the last group is the control group, and the $a - 1$ treatment groups.

Then, we compare treatment groups to the control group. Hence, we have $a - 1$ contrasts (i.e., $a - 1$ pairwise comparisons)

19.1.1.3.6 Summary When choosing a multiple contrast method:

- Pairwise
 - Equal groups sizes: Tukey

- Unequal groups sizes: Tukey, Scheffe
- Not pairwise
 - with control: Dunnett
 - general: Bonferroni, Scheffe

19.1.2 Single Factor Random Effects Model

Also known as ANOVA Type II models.

Treatments are chosen at from larger population. We extend inference to all treatments in the population and not restrict our inference to those treatments that happened to be selected for the study.

19.1.2.1 Random Cell Means

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

where

- $\mu_i \sim N(\mu, \sigma_\mu^2)$ and independent
- $\epsilon_{ij} \sim N(0, \sigma^2)$ and independent

μ_i and ϵ_{ij} are mutually independent for $i = 1, \dots, a; j = 1, \dots, n$

With all treatment sample sizes are equal

$$E(Y_{ij}) = E(\mu_i) = \mu \text{var}(Y_{ij}) = \text{var}(\mu_i) + \text{var}(\epsilon_i) = \sigma_\mu^2 + \sigma^2$$

Since Y_{ij} are not independent

$$\begin{aligned} \text{cov}(Y_{ij}, Y_{ij'}) &= E(Y_{ij}Y_{ij'}) - E(Y_{ij})E(Y_{ij'}) \\ &= E(\mu_i^2 + \mu_i\epsilon_{ij'} + \mu_i\epsilon_{ij} + \epsilon_{ij}\epsilon_{ij'}) - \mu^2 \\ &= \sigma_\mu^2 + \mu^2 - \mu^2 && \text{if } j \neq j' \\ &= \sigma_\mu^2 && \text{if } j = j' \end{aligned}$$

$$\begin{aligned} \text{cov}(Y_{ij}, Y_{i'j'}) &= E(\mu_i\mu_{i'} + \mu_i\epsilon_{i'j'} + \mu_{i'}\epsilon_{ij} + \epsilon_{ij}\epsilon_{i'j'}) - \mu^2 \\ &= \mu^2 - \mu^2 && \text{if } i \neq i' \\ &= 0 \end{aligned}$$

Hence,

- all observations have the same variance
- any two observations from the same treatment have covariance σ_μ^2
- The correlation between any two responses from the same treatment:

$$\rho(Y_{ij}, Y_{ij'}) = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma^2} \quad j \neq j'$$

Inference

Intraclass Correlation Coefficient

$$\frac{\sigma_\mu^2}{\sigma^2 + \sigma_\mu^2}$$

which measures the proportion of total variability of Y_{ij} accounted for by the variance of μ_i

$$H_0 : \sigma_\mu^2 = 0 \quad H_a : \sigma_\mu^2 \neq 0$$

H_0 implies $\mu_i = \mu$ for all i , which can be tested by the F-test in ANOVA.

The understandings of the Single Factor Fixed Effects Model and the Single Factor Random Effects Model are different, the ANOVA is same for the one factor model. The difference is in the expected mean squares

Random Effects Model	Fixed Effects Model
$E(MSE) = \sigma^2$	$E(MSE) = \sigma^2$
$E(MSTR) = \sigma^2 - n\sigma_\mu^2$	$E(MSTR) = \sigma^2 + \frac{\sum_i n_i (\mu_i - \mu)^2}{a-1}$

If $\sigma_\mu^2 = 0$, then MSE and MSTR have the same expectation (σ^2). Otherwise, $E(MSTR) > E(MSE)$. Large values of the statistic

$$F = \frac{MSTR}{MSE}$$

suggest we reject H_0 .

Since $F \sim F_{(a-1, a(n-1))}$ when H_0 holds. If $F > f_{(1-\alpha; a-1, a(n-1))}$ we reject H_0 . If sample sizes are not equal, F-test can still be used, but the df are $a-1$ and $N-a$.

19.1.2.1.1 Estimation of μ An unbiased estimator of $E(Y_{ij}) = \mu$ is the grand mean: $\hat{\mu} = \bar{Y}_{..}$.

The variance of this estimator is

$$\begin{aligned} var(\bar{Y}_{..}) &= var\left(\sum_i \bar{Y}_{i.}/a\right) \\ &= \frac{1}{a^2} \sum_i var(\bar{Y}_{i.}) \\ &= \frac{1}{a^2} \sum_i (\sigma_\mu^2 + \sigma^2/n) \\ &= \frac{1}{a^2} (\sigma_\mu^2 + \sigma^2/n) \\ &= \frac{n\sigma_\mu^2 + \sigma^2}{an} \end{aligned}$$

An unbiased estimator of this variance is $s^2(\bar{Y}) = \frac{MSTR}{an}$. Thus $\frac{\bar{Y}_{..} - \mu}{s(\bar{Y}_{..})} \sim t_{a-1}$

A $1 - \alpha$ confidence interval is $\bar{Y}_{..} \pm t_{(1-\alpha/2;a-1)} s(\bar{Y}_{..})$

19.1.2.1.2 Estimation of $\sigma_\mu^2/(\sigma_\mu^2 + \sigma^2)$ In the random and fixed effects model, MSTR and MSE are independent. When the sample sizes are equal ($n_i = n$ for all i),

$$\frac{\frac{MSTR}{n\sigma_\mu^2 + \sigma^2}}{\frac{MSE}{\sigma^2}} \sim f_{(a-1, a(n-1))}$$

$$P(f_{(\alpha/2; a-1, a(n-1))} \leq \frac{\frac{MSTR}{n\sigma_\mu^2 + \sigma^2}}{\frac{MSE}{\sigma^2}} \leq f_{(1-\alpha/2; a-1, a(n-1))}) = 1 - \alpha$$

$$L = \frac{1}{n} \left(\frac{MSTR}{MSE} \left(\frac{1}{f_{(1-\alpha/2; a-1, a(n-1))}} \right) - 1 \right) U = \frac{1}{n} \left(\frac{MSTR}{MSE} \left(\frac{1}{f_{(\alpha/2; a-1, a(n-1))}} \right) - 1 \right)$$

The lower and upper (L^*, U^*) confidence limits for $\frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma^2}$

$$L^* = \frac{L}{1+L} U^* = \frac{U}{1+U}$$

If the lower limit for $\frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma^2}$ is negative, it is customary to set $L = 0$.

19.1.2.1.3 Estimation of σ^2 $a(n-1)MSE/\sigma^2 \sim \chi^2_{a(n-1)}$, the $(1-\alpha)$ confidence interval for σ^2 :

$$\frac{a(n-1)MSE}{\chi^2_{1-\alpha/2;a(n-1)}} \leq \sigma^2 \leq \frac{a(n-1)MSE}{\chi^2_{\alpha/2;a(n-1)}}$$

can also be used in case sample sizes are not equal - then df is N-a.

19.1.2.1.4 Estimation of σ_μ^2 $E(MSE) = \sigma^2$ $E(MSTR) = \sigma^2 + n\sigma_\mu^2$.
Hence,

$$\sigma_\mu^2 = \frac{E(MSTR) - E(MSE)}{n}$$

An unbiased estimator of σ_μ^2 is given by

$$s_\mu^2 = \frac{MSTR - MSE}{n}$$

if $s_\mu^2 < 0$, set $s_\mu^2 = 0$

If sample sizes are not equal,

$$s_\mu^2 = \frac{MSTR - MSE}{n'}$$

where $n' = \frac{1}{a-1}(\sum_i n_i - \frac{\sum_i n_i^2}{\sum_i n_i})$

no exact confidence intervals for σ_μ^2 , but we can approximate intervals.

Satterthwaite Procedure can be used to construct approximate confidence intervals for linear combination of expected mean squares

A linear combination:

$$\sigma_\mu^2 = \frac{1}{n}E(MSTR) + (-\frac{1}{n})E(MSE)$$

$$S = d_1E(MS_1) + \dots + d_hE(MS_h)$$

where d_i are coefficients.

An unbiased estimator of S is

$$\hat{S} = d_1MS_1 + \dots + d_hMS_h$$

Let df_i be the degrees of freedom associated with the mean square MS_i . The **Satterthwaite** approximation:

$$\frac{(df)\hat{S}}{S} \sim \chi_{df}^2$$

where

$$df = \frac{(d_1 MS_1 + \dots + d_h MS_h)^2}{(d_1 MS_1)^2 / df_1 + \dots + (d_h MS_h)^2 / df_h}$$

An approximate $1 - \alpha$ confidence interval for S:

$$\frac{(df)\hat{S}}{\chi_{1-\alpha/2;df}^2} \leq S \leq \frac{(df)\hat{S}}{\chi_{\alpha/2;df}^2}$$

For the single factor random effects model

$$\frac{(df)s_\mu^2}{\chi_{1-\alpha/2;df}^2} \leq \sigma_\mu^2 \leq \frac{(df)s_\mu^2}{\chi_{\alpha/2;df}^2}$$

where

$$df = \frac{(sn_\mu^2)^2}{\frac{(MSTR)^2}{a-1} + \frac{(MSE)^2}{a(n-1)}}$$

19.1.2.2 Random Treatment Effects Model

$$\tau_i = \mu_i - E(\mu_i) = \mu_i - \mu$$

we have $\mu_i = \mu + \tau_i$ and

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

where

- μ = constant, common to all observations
- $\tau_i \sim N(0, \sigma_\tau^2)$ independent (random variables)
- $\epsilon_{ij} \sim N(0, \sigma^2)$ independent.

- τ_i, ϵ_{ij} are independent ($i=1,\dots,a; j =1,\dots,n$)
- our model is concerned with only balanced single factor ANOVA.

Diagnostics Measures

- Non-constant error variance (plots, Levene test, Hartley test).
- Non-independence of errors (plots, Durban-Watson test).
- Outliers (plots, regression methods).
- Non-normality of error terms (plots, Shapiro-Wilk, Anderson-Darling).
- Omitted Variable Bias (plots)

Remedial

- Weighted Least Squares
- Transformations
- Non-parametric Procedures.

Note

- Fixed effect ANOVA is relatively robust to
 - non-normality
 - unequal variances when sample sizes are approximately equal; at least the F-test and multiple comparisons. However, single comparisons of treatment means are sensitive to unequal variances.
- Lack of independence can seriously affect both fixed and random effect ANVOA.

19.1.3 Two Factor Fixed Effect ANOVA

The multi-factor experiment is

- more efficient
- provides more info
- gives more validity to the findings.

19.1.3.1 Balanced

Assumption:

- All treatment sample sizes are equal
- All treatment means are of equal importance

Assume:

- Factor A has a levels and Factor B has b levels. All $a \times b$ factor levels are considered.
- The number of treatments for each level is n . $N = abn$ observations in the study.

19.1.3.1.1 Cell Means Model

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk}$$

where

- μ_{ij} are fixed parameters (cell means)
- $i = 1, \dots, a$ = the levels of Factor A
- $j = 1, \dots, b$ = the levels of Factor B.
- $\epsilon_{ijk} \sim \text{indep } N(0, \sigma^2)$ for $i = 1, \dots, a$, $j = 1, \dots, b$ and $k = 1, \dots, n$

And

$$E(Y_{ijk}) = \mu_{ij} \quad var(Y_{ijk}) = var(\epsilon_{ijk}) = \sigma^2$$

Hence,

$$Y_{ijk} \sim \text{indep } N(\mu_{ij}, \sigma^2)$$

And the model is

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

Thus,

$$E(\mathbf{Y}) = \mathbf{X}\beta var(\mathbf{Y}) = \sigma^2 \mathbf{I}$$

Interaction

$$(\alpha\beta)_{ij} = \mu_{ij} - (\mu_{..} + \alpha_i + \beta_j)$$

where

- $\mu_{..} = \sum_i \sum_j \mu_{ij}/ab$ is the grand mean
- $\alpha_i = \mu_{i..} - \mu_{..}$ is the main effect for factor A at the i-th level
- $\beta_j = \mu_{.j} - \mu_{..}$ is the main effect for factor B at the j-th level
- $(\alpha\beta)_{ij}$ is the interaction effect when factor A is at the i-th level and factor B is at the j-th level.
- $(\alpha\beta)_{ij} = \mu_{ij} - \mu_{i..} - \mu_{.j} + \mu_{..}$

Examine interactions:

- Examine whether all μ_{ij} can be expressed as the sums $\mu_{..} + \alpha_i + \beta_j$
- Examine whether the difference between the mean responses for any two levels of factor B is the same for all levels of factor A.
- Examine whether the difference between the mean response for any two levels of factor A is the same for all levels of factor B
- Examine whether the treatment mean curves for the different factor levels in a treatment plot are parallel.

For $j = 1, \dots, b$

$$\begin{aligned} \sum_i (\alpha\beta)_{ij} &= \sum_i (\mu_{ij} - \mu_{..} - \alpha_i - \beta_j) \\ &= \sum_i \mu_{ij} - a\mu_{..} - \sum_i \alpha_i - a\beta_j \\ &= a\mu_{.j} - a\mu_{..} - \sum_i (\mu_{i..} - \mu_{..}) - a(\mu_{.j} - \mu_{..}) \\ &= a\mu_{.j} - a\mu_{..} - a\mu_{..} + a\mu_{..} - a(\mu_{.j} - \mu_{..}) \\ &= 0 \end{aligned}$$

Similarly, $\sum_j (\alpha\beta)_{ij} = 0, i = 1, \dots, a$ and $\sum_i \sum_j (\alpha\beta)_{ij} = 0, \sum_i \alpha_i = 0, \sum_j \beta_j = 0$

19.1.3.1.2 Factor Effects Model

$$\mu_{ij} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where

- $\mu_{..}$ is a constant
- α_i are constants subject to the restriction $\sum_i \alpha_i = 0$
- β_j are constants subject to the restriction $\sum_j \beta_j = 0$
- $(\alpha\beta)_{ij}$ are constants subject to the restriction $\sum_i (\alpha\beta)_{ij} = 0$ for $j = 1, \dots, b$ and $\sum_j (\alpha\beta)_{ij} = 0$ for $i = 1, \dots, a$
- $\epsilon_{ijk} \sim \text{indep } N(0, \sigma^2)$ for $k = 1, \dots, n$

We have

$$E(Y_{ijk}) = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} \quad \text{var}(Y_{ijk}) = \sigma^2 Y_{ijk} \sim N(\mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij}, \sigma^2)$$

We have $1 + a + b + ab$ parameters. But there are ab parameters in the Cell Means Model. In the Factor Effects Model, the restrictions limit the number of parameters that can be estimated:

$$1 \text{ for } \mu_{..} \quad (a - 1) \text{ for } \alpha_i \quad (b - 1) \text{ for } \beta_j \quad (a - 1)(b - 1) \text{ for } (\alpha\beta)_{ij}$$

Hence, there are

$$1 + a - 1 + b - 1 + ab - a - b + 1 = ab$$

parameters in the model.

We can have several restrictions when considering the model in the form $\mathbf{Y} = \mathbf{X}\beta + \epsilon$

One way:

$$\alpha_a = \alpha_1 - \alpha_2 - \dots - \alpha_{a-1} \beta_b = -\beta_1 - \beta_2 - \dots - \beta_{b-1} (\alpha\beta)_{ib} = -(\alpha\beta)_{i1} - (\alpha\beta)_{i2} - \dots - (\alpha\beta)_{i,b-1}; i = 1, \dots, a \quad (\alpha\beta)_{aj} = -(\alpha\beta)_{1j} -$$

We can fit the model by least squares or maximum likelihood

Cell Means Model

minimize

$$Q = \sum_i \sum_j \sum_k (Y_{ijk} - \mu_{ij})^2$$

estimators

$$\hat{\mu}_{ij} = \bar{Y}_{ij} \hat{Y}_{ijk} = \bar{Y}_{ij} e_{ijk} = Y_{ijk} - \hat{Y}_{ijk} = Y_{ijk} - \bar{Y}_{ij}$$

Factor Effects Model

$$Q = \sum_i \sum_j \sum_k (Y_{ijk} - \mu_{..} - \alpha_i - \beta_j - (\alpha\beta)_{ij})^2$$

subject to the restrictions

$$\sum_i \alpha_i = 0 \sum_j \beta_j = 0 \sum_i (\alpha\beta)_{ij} = 0 \sum_j (\alpha\beta)_{ij} = 0$$

estimators

$$\hat{\mu}_{..} = \bar{Y}_{...} \hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...} \hat{\beta}_j = \bar{Y}_{.j.} - \bar{Y}_{...} (\hat{\alpha}\beta)_{ij} = \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}$$

The fitted values

$$\hat{Y}_{ijk} = \bar{Y}_{...} + (\bar{Y}_{i..} - \bar{Y}_{...}) + (\bar{Y}_{.j.} - \bar{Y}_{...}) + (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}) = \bar{Y}_{ij.}$$

where

$$e_{ijk} = Y_{ijk} - \bar{Y}_{ij.} e_{ijk} \sim \text{indep } (0, \sigma^2)$$

and

$$s_{\hat{\mu}_{..}}^2 = \frac{MSE}{nab} s_{\hat{\alpha}_i}^2 = MSE\left(\frac{1}{nb} - \frac{1}{nab}\right) s_{\hat{\beta}_j}^2 = MSE\left(\frac{1}{na} - \frac{1}{nab}\right) s_{(\hat{\alpha}\beta)_{ij}}^2 = MSE\left(\frac{1}{n} - \frac{1}{na} - \frac{1}{nb} + \frac{1}{nab}\right)$$

19.1.3.1.2.1 Partitioning the Total Sum of Squares

$$Y_{ijk} - \bar{Y}_{...} = \bar{Y}_{ij.} - \bar{Y}_{...} + Y_{ijk} - \bar{Y}_{ij.}$$

$Y_{ijk} - \bar{Y}_{...}$: Total deviation

$\bar{Y}_{ij.} - \bar{Y}_{...}$: Deviation of treatment mean from overall mean

$Y_{ijk} - \bar{Y}_{ij.}$: Deviation of observation around treatment mean (residual).

$$\sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{...})^2 = n \sum_i \sum_j (\bar{Y}_{ij.} - \bar{Y}_{...})^2 + \sum_i \sum_j sum_k (Y_{ijk} - \bar{Y}_{ij.})^2$$

$$SSTO = SSTR + SSE$$

(cross product terms are 0)

$$\bar{Y}_{ij.} - \bar{Y}_{...} = \bar{Y}_{i..} - \bar{Y}_{...} + \bar{Y}_{.j.} - \bar{Y}_{...} + \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}$$

squaring and summing:

$$n \sum_i \sum_j (\bar{Y}_{ij.} - \bar{Y}_{...})^2 = nb \sum_i (\bar{Y}_{i..} - \bar{Y}_{...})^2 + na \sum_j (\bar{Y}_{.j.} - \bar{Y}_{...})^2 + n \sum_i \sum_j (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2$$

$$SSTR = SSA + SSB + SSAB$$

The interaction term from

$$SSAB = SSTO - SSE - SSA - SSB - SSAB = SSTR - SSA - SSB$$

where

- SSA is the factor A sum of squares (measures the variability of the estimated factor A level means $\bar{Y}_{i..}$)- the more variable, the larger SSA
- SSB is the factor B sum of squares
- SSAB is the interaction sum of squares, measuring the variability of the estimated interactions.

19.1.3.1.2.2 Partitioning the df $N = abn$ cases and ab treatments.

For one-way ANOVA and regression, the partition has df:

$$SS : SSTO = SSTR + SSEdf : N - 1 = (ab - 1) + (N - ab)$$

we must further partition the $ab - 1$ df with SSTR

$$SSTR = SSA + SSB + SSABab - 1 = (a - 1) + (b - 1) + (a - 1)(b - 1)$$

- $df_{SSA} = a - 1$: a treatment deviations but 1 df is lost due to the restriction $\sum(\bar{Y}_{i..} - \bar{Y}_{...}) = 0$
- $df_{SSB} = b - 1$: b treatment deviations but 1 df is lost due to the restriction $\sum(\bar{Y}_{.j.} - \bar{Y}_{...}) = 0$
- $df_{SSAB} = (a - 1)(b - 1) = (ab - 1) - (a - 1) - (b - 1)$: ab interactions, there are $(a+b-1)$ restrictions, so df = ab-a-(b-1) = (a-1)(b-1)

19.1.3.1.2.3 Mean Squares

$$MSA = \frac{SSA}{a - 1} MSB = \frac{SSB}{b - 1} MSAB = \frac{SSAB}{(a - 1)(b - 1)}$$

The expected mean squares are

$$E(MSE) = \sigma^2 E(MSA) = \sigma^2 + nb \frac{\sum \alpha_i^2}{a - 1} = \sigma^2 + nb \frac{\sum (\sum_{i..} - \mu_{...})^2}{a - 1} E(MSB) = \sigma^2 + na \frac{\sum \beta_j^2}{b - 1} = \sigma^2 + na \frac{\sum (\sum_{.j.} - \mu_{...})^2}{b - 1}$$

If there are no factor A main effects (all $\mu_{i..} = 0$ or $\alpha_i = 0$) the MSA and MSE have the same expectation; otherwise $MSA > MSE$. Same for factor B, and interaction effects. which case we can examine F-statistics.

Interaction

$$\begin{aligned} H_0 : \mu_{ij} - \mu_{i..} - \mu_{.j.} + \mu_{...} &= 0 && \text{for all } i,j \\ H_a : \mu_{ij} - \mu_{i..} - \mu_{.j.} + \mu_{...} &\neq 0 && \text{for some } i,j \end{aligned}$$

or

$$H_0 : \text{All}(\alpha\beta)_{ij} = 0 H_a : \text{Not all}(\alpha\beta) = 0$$

Let $F = \frac{MSAB}{MSE}$. When H_0 is true $F \sim f_{((a-1)(b-1), ab(n-1))}$. So reject H_0 when $F > f_{((a-1)(b-1), ab(n-1))}$

Factor A main effects:

$$H_0 : \mu_{1.} = \mu_{2.} = \dots = \mu_{a.} H_a : \text{Not all } \mu_i. \text{ are equal}$$

or

$$H_0 : \alpha_1 = \dots = \alpha_a = 0 H_a : \text{Not all } \alpha_i \text{ are equal to 0}$$

$F = \frac{MSA}{MSE}$ and reject H_0 if $F > f_{(1-\alpha; a-1, ab(n-1))}$

19.1.3.1.2.4 Two-way ANOVA

Source of Variation	SS	df	MS	F
Factor A	SSA	a-1	MSA = SSA/(a-1)	MSA/MSE
Factor B	SSB	b-1	MSB = SSB/(b-1)	MSB/MSE
AB interactions	SSAB	(a-1)(b-1)	MSAB = SSAB /MSE	
Error	SSE	ab(n-1)	MSE = SSE/ab(n-1)	
Total (corrected)	SSTO	abn - 1		

Doing 2-way ANOVA means you always check interaction first, because if there are significant interactions, checking the significance of the main effects becomes moot.

The main effects concern the mean responses for levels of one factor averaged over the levels of the other factor. When interaction is present, we can't conclude that a given factor has no effect, even if these averages are the same. It means that the effect of the factor depends on the level of the other factor.

On the other hand, if you can establish that there is no interaction, then you can consider inference on the factor main effects, which are then said to be **additive**.

And we can also compare factor means like the Single Factor Fixed Effects Model using Tukey, Scheffe, Bonferroni.

We can also consider contrasts in the 2-way model

$$L = \sum c_i \mu_i$$

where $\sum c_i = 0$
which is estimated by

$$\hat{L} = \sum c_i \bar{Y}_{i..}$$

with variance

$$\sigma^2(\hat{L}) = \frac{\sigma^2}{bn} \sum c_i^2$$

and variance estimate

$$\frac{MSE}{bn} \sum c_i^2$$

Orthogonal Contrasts

$$L_1 = \sum c_i \mu_i, \sum c_i = 0 \\ L_2 = \sum d_i \mu_i, \sum d_i = 0$$

these contrasts are said to be **orthogonal** if

$$\sum \frac{c_i d_i}{n_i} = 0$$

in balanced case $\sum c_i d_i = 0$

$$\begin{aligned} cov(\hat{L}_1, \hat{L}_2) &= cov\left(\sum_i c_i \bar{Y}_{i..}, \sum_l d_l \bar{Y}_{l..}\right) \\ &= \sum_i \sum_l c_i d_l cov(\bar{Y}_{i..}, \bar{Y}_{l..}) \\ &= \sum_i c_i d_i \frac{\sigma^2}{bn} = 0 \end{aligned}$$

Orthogonal contrasts can be used to further partition the model sum of squares. There are many sets of orthogonal contrasts and thus, many ways to partition the sum of squares.

A special set of orthogonal contrasts that are used when the levels of a factor can be assigned values on a metric scale are called **orthogonal polynomials**

Coefficients can be found for the special case of

- equal spaced levels (e.g., (0 15 30 45 60))
- equal sample sizes ($n_1 = n_2 = \dots = n_{ab}$)

We can define the SS for a given contrast:

$$SS_L = \frac{\hat{L}^2}{\sum_{i=1}^a (c_i^2/bn_i)}$$

$$T = \frac{\hat{L}}{\sqrt{MSE \sum_{i=1}^a (c_i^2/bn_i)}} \sim t$$

Moreover,

$$t_{(1-\alpha/2; df)}^2 = F_{(1-\alpha; 1, df)}$$

So,

$$\frac{SS_L}{MSE} \sim F_{(1-\alpha; 1, df_{MSE})}$$

all contrasts have d.f = 1

19.1.3.2 Unbalanced

We could have unequal numbers of replications for all treatment combinations:

- observational studies
- dropouts in designed studies
- larger sample sizes for inexpensive treatments
- Sample sizes to match population makeup.

Assume that each factor combination has at least 1 observation (no empty cells)

Consider the same model as:

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where sample sizes are: n_{ij} :

$$n_{i.} = \sum_j n_{ij} n_{.j} = \sum_i n_{ij} n_T = \sum_i \sum_j n_{ij}$$

Problem here is that

$$SSTO \neq SSA + SSB + SSAB + SSE$$

(the design is **non-orthogonal**)

- For $i = 1, \dots, a - 1$,
 $\begin{cases} u_i = 1 & \text{if the obs is from the } i\text{-th level of Factor 1} \\ -1 & \text{if the obs is from the } a\text{-th level of Factor 1} \\ 0 & \text{otherwise} \end{cases}$
- For $j = 1, \dots, b - 1$
 $\begin{cases} v_j = 1 & \text{if the obs is from the } j\text{-th level of Factor 1} \\ -1 & \text{if the obs is from the } b\text{-th level of Factor 1} \\ 0 & \text{otherwise} \end{cases}$

We can use these indicator variables as predictor variables and $\mu_{..}, \alpha_i, \beta_j, (\alpha\beta)_{ij}$ as unknown parameters.

$$Y = \mu_{..} + \sum_{i=1}^{a-1} \alpha_i u_i + \sum_{j=1}^{b-1} \beta_j v_j + \sum_{i=1}^{a-1} \sum_{j=1}^{b-1} (\alpha\beta)_{ij} u_i v_j + \epsilon$$

To test hypotheses, we use the extra sum of squares idea.

For interaction effects

$$H_0 : \text{all } (\alpha\beta)_{ij} = 0 \quad H_a : \text{not all } (\alpha\beta)_{ij} = 0$$

Or to test

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad H_a : \text{not all } \beta_j = 0$$

Analysis of Factor Means

(e.g., contrasts) is analogous to the balanced case, with modifications in the formulas for means and standard errors to account for unequal sample sizes.

Or , we can fit the cell means model and consider it from a regression perspective

If you have empty cells (i.e., some factor combinations have no observation), then the equivalent regression approach can't be used. But you can still do partial analyses

19.1.4 Two-Way Random Effects ANOVA

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where

- $\mu_{..}$: constant
- $\alpha_i \sim N(0, \sigma_\alpha^2), i = 1, \dots, a$ (independent)
- $\beta_j \sim N(0, \sigma_\beta^2), j = 1, \dots, b$ (independent)
- $(\alpha\beta)_{ij} \sim N(0, \sigma_{\alpha\beta}^2), i = 1, \dots, a, j = 1, \dots, b$ (independent)
- $\epsilon_{ijk} \sim N(0, \sigma^2)$ (independent)

All $\alpha_i, \beta_j, (\alpha\beta)_{ij}$ are pairwise independent

Theoretical means, variances, and covariances are

$$E(Y_{ijk}) = \mu_{..}, \text{var}(Y_{ijk}) = \sigma_Y^2 = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma^2$$

So

$$Y_{ijk} \sim N(\mu_{..}, \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma^2)$$

$$\text{cov}(Y_{ijk}, Y_{ij'k'}) = \sigma_\alpha^2, j \neq j' \text{cov}(Y_{ijk}, Y_{i'jk'}) = \sigma_\beta^2, i \neq i' \text{cov}(Y_{ijk}, Y_{ijk'}) = \sigma_{\alpha\beta}^2, k \neq k' \text{cov}(Y_{ijk}, Y_{i'j'k'}) = 0, i \neq i', j \neq j', k \neq k'$$

19.1.5 Two-Way Mixed Effects ANOVA

19.1.5.1 Balanced

One fixed factor, while other is random treatment levels, we have a **mixed effects model** or a **mixed model**

Restricted mixed model for 2-way ANOVA:

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where

- $\mu_{..}$: constant

- α_i : fixed effects with constraints subject to restriction $\sum \alpha_i = 0$
- $\beta_j \sim \text{indepN}(0, \sigma_\beta^2)$
- $(\alpha\beta)_{ij} \sim N(0, \frac{a-1}{a} \sigma_{\alpha\beta}^2)$ subject to restriction $\sum_i (\alpha\beta)_{ij} = 0$ for all j, the variance here is written as the proportion for convenience; it makes the expected mean squares simpler (other assumed $\text{var}((\alpha\beta)_{ij}) = \sigma_{\alpha\beta}^2$)
- $\text{cov}((\alpha\beta)_{ij}, (\alpha\beta)_{i'j'}) = -\frac{1}{a} \sigma_{\alpha\beta}^2, i \neq i'$
- $\epsilon_{ijk} \sim \text{indepN}(0, \sigma^2)$
- $\beta_j, (\alpha\beta)_{ij}, \epsilon_{ijk}$ are pairwise independent

Two-way mixed models are written in an “unrestricted” form, with no restrictions on the interaction effects $(\alpha\beta)_{ij}$, they are pairwise independent. Let $\beta^*, (\alpha\beta)_{ij}^*$ be the unrestricted random effects, and $(\bar{\alpha}\beta)_{ij}^*$ the means averaged over the fixed factor for each level of random factor B.

$$\beta_j = \beta_j^* + (\bar{\alpha}\beta)_{ij}^* (\alpha\beta)_{ij} = (\alpha\beta)_{ij}^* - (\bar{\alpha}\beta)_{ij}^*$$

Some consider the restricted model to be more general. but here we consider the restricted form.

$$E(Y_{ijk}) = \mu_{..} + \alpha_i \text{var}(Y_{ijk}) = \sigma_\beta^2 + \frac{a-1}{a} \sigma_{\alpha\beta}^2 + \sigma^2$$

Responses from the same random factor (B) level are correlated

$$\text{cov}(Y_{ijk}, Y_{ijk'}) = E(Y_{ijk}Y_{ijk'}) - E(Y_{ijk})E(Y_{ijk'}) = \sigma_\beta^2 + \frac{a-1}{a} \sigma_{\alpha\beta}^2, k \neq k'$$

Similarly,

$$\text{cov}(Y_{ijk}, Y_{i'jk'}) = \sigma_\beta^2 - \frac{1}{a} \sigma_{\alpha\beta}^2, i \neq i' \text{ and } j \neq j'$$

Hence, you can see that the only way you don't have dependence in the Y is when they don't share the same random effect.

An advantage of the **restricted mixed model** is that 2 observations from the same random factor b level can be positively or negatively correlated. In the **unrestricted model**, they can only be positively correlated.

Fixed ANOVA		
Mean Square	(A, B Fixed)	Random ANOVA (A,B random)
MSA	a - 1	$\sigma^2 + nb \sum_{i=1}^a \alpha_i^2$
MSB	b-1	$\sigma^2 + na \sum_{j=1}^b \beta_j^2$
MSAB	(a-1)(b-1)	$\sigma^2 + n \sum_{i=1}^a \sum_{j=1}^b (\alpha_i \beta_j)^2$
MSE	(n-1)ab	σ^2

For fixed, random, and mixed models (balanced), the ANOVA table sums of squares calculations are identical. (also true for df and mean squares). The only difference is with the expected mean squares, thus the test statistics.

In Random ANOVA, we test

$$H_0 : \sigma^2 = 0 \quad H_a : \sigma^2 > 0$$

by considering $F = \frac{MSA}{MSAB} \sim F_{a-1;(a-1)(b-1)}$

The same test statistic is used for mixed models, but in that case we are testing null hypothesis that all of the $\alpha_i = 0$

The test statistic different for the same null hypothesis under the fixed effects model.

Test for effects of	Fixed ANOVA (A&B fixed)	Random	Mixed ANOVA
		ANOVA (A&B random)	(A fixed, B random)
Factor A	$\frac{MSA}{MSE}$	$\frac{MSA}{MSB}$	$\frac{MSA}{MSAB}$
Factor B	$\frac{MSB}{MSE}$	$\frac{MSB}{MSAB}$	$\frac{MSB}{MSAB}$
AB interactions	$\frac{MSAB}{MSE}$	$\frac{MSAB}{MSE}$	$\frac{MSAB}{MSE}$

Estimation Of Variance Components

In random and mixed effects models, we are interested in estimating the **variance components**

Variance component σ_β^2 in the mixed ANOVA.

$$E(\sigma_\beta^2) = \frac{E(MSB) - E(MSE)}{na} = \frac{\sigma^2 + na\sigma_\beta^2 - \sigma^2}{na} = \sigma_\beta^2$$

which can be estimated with

$$\hat{\sigma}_\beta^2 = \frac{MSB - MSE}{na}$$

Confidence intervals for variance components can be constructed (approximately) by using the **Satterthwaite** procedure or the MLS procedure (like the 1-way random effects)

Estimation of Fixed Effects in Mixed Models

$$\hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...}\hat{\mu}_{i..} = \bar{Y}_{...} + (\bar{Y}_{i..} - \bar{Y}_{...}) = \bar{Y}_{i..}\sigma^2(\hat{\alpha}_i) = \frac{\sigma^2 + n\sigma_{\alpha\beta}^2}{bn} = \frac{E(MSAB)}{bn}s^2(\hat{\alpha}_i) = \frac{MSAB}{bn}$$

Contrasts on the Fixed Effects

$$L = \sum c_i \alpha_i \sum c_i = 0 \hat{L} = \sum c_i \hat{\alpha}_i \sigma^2(\hat{L}) = \sum c_i^2 \sigma^2(\hat{\alpha}_i) s^2(\hat{L}) = \frac{MSAB}{bn} \sum c_i^2$$

Confidence intervals and tests can be constructed as usual

19.1.5.2 Unbalanced

For a mixed model with $a = 2, b = 4$

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \quad var(\beta_j) = \sigma_\beta^2 var((\alpha\beta)_{ij}) = \frac{2-1}{2} \sigma_{\alpha\beta}^2 = \frac{\sigma_{\alpha\beta}^2}{2} var(\epsilon_{ijk}) = \sigma^2 E(Y_{ijk}) = \mu_{..} + \epsilon_{ijk}$$

assume

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, M)$$

where M is block diagonal

density function

$$f(\mathbf{Y}) = \frac{1}{(2\pi)^{N/2}|M|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{Y} - \mathbf{X})' \mathbf{M}^{-1} (\mathbf{Y} - \mathbf{X})\right)$$

if we knew the variance components, we could use GLS:

$$\hat{\beta}_{GLS} = (\mathbf{X}' \mathbf{M}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{M}^{-1} \mathbf{Y}$$

but we usually don't know the variance components $\sigma^2, \sigma_\beta^2, \sigma_{\alpha\beta}^2$ that make up M

Another way to get estimates is by **Maximum likelihood estimation**

we try to maximize its log

$$\ln L = -\frac{N}{2} \ln(2\pi) - \frac{1}{2} \ln |M| - \frac{1}{2} (\mathbf{Y} - \mathbf{X})' M^{-1} (\mathbf{Y} - \mathbf{X})$$

19.2 Nonparametric ANOVA

19.2.1 Kruskal-Wallis

Generalization of independent samples Wilcoxon Rank sum test for 2 independent samples (like F-test of one-way ANOVA is a generalization to several independent samples of the two sample t-test)

Consider the one-way case:

We have

- $a \geq 2$ treatments
- n_i is the sample size for the i th treatment
- Y_{ij} is the j -th observation from the i th treatment.
- we make **no** assumption of normality
- We only assume that observations on the i th treatment are a random sample from the continuous CDF F_i , $i = 1, \dots, n$, and are mutually independent.

$$H_0 : F_1 = F_2 = \dots = F_a \quad H_a : F_i < F_j \text{ for some } i \neq j$$

or if distribution is from the location-scale family, $H_0 : \theta_1 = \theta_2 = \dots = \theta_a$)

Procedure

- Rank all $N = \sum_{i=1}^a n_i$ observations in ascending order. Let $r_{ij} = rank(Y_{ij})$, note $\sum_i \sum_j r_{ij} = 1 + 2 + \dots + N = \frac{N(N+1)}{2}$

- Calculate the rank sums and averages:

$$r_{i\cdot} = \sum_{j=1}^{n_i} r_{ij}$$

and

$$\bar{r}_{i\cdot} = \frac{r_{i\cdot}}{n_i}, i = 1, \dots, a$$

- Calculate the test statistic on the ranks:

$$\chi^2_{KW} = \frac{SSTR}{\frac{SSTO}{N-1}}$$

where $SSTR = \sum n_i (\bar{r}_{i\cdot} - \bar{r}_{..})^2$ and $SSTO = \sum \sum (\bar{r}_{ij} - \bar{r}_{..})^2$

- For large n_i (≥ 5 observations) the Kruskal-Wallis statistic is approximated by a χ^2_{a-1} distribution when all the treatment means are equal. Hence, reject H_0 if $\chi^2_{KW} > \chi^2_{(1-\alpha;a-1)}$.
- If sample sizes are small, one can exhaustively work out all possible distinct ways of assigning N ranks to the observations from a treatments and calculate the value of the KW statistic in each case ($\frac{N!}{n_1! \dots n_a!}$ possible combinations). Under H_0 all of these assignments are equally likely.

19.2.2 Friedman Test

When the responses $Y_{ij} = 1, \dots, n, j = 1, \dots, r$ in a randomized complete block design are not normally distributed (or do not have constant variance), a non-parametric test is more helpful.

A distribution-free rank-based test for comparing the treatments in this setting is the Friedman test. Let F_{ij} be the CDF of random Y_{ij} , corresponding to the observed value y_{ij}

Under the null hypothesis, F_{ij} are identical for all treatments j separately for each block i .

$$H_0 : F_{i1} = F_{i2} = \dots = F_{ir} \text{ for all } i \quad H_a : F_{ij} < F_{ij'} \text{ for some } j \neq j' \text{ for all } i$$

For location parameter distributions, treatment effects can be tested:

$$H_0 : \tau_1 = \tau_2 = \dots = \tau_r \quad H_a : \tau_j > \tau_{j'} \text{ for some } j \neq j'$$

Procedure

- Rank observations from the r treatments separately within each block (in ascending order; if ties, each tied observation is given the mean of ranks involved). Let the ranks be called r_{ij}
- Calculate the Friedman test statistic

$$\chi_F^2 = \frac{SSTR}{\frac{SSTR + SSE}{n(r-1)}}$$

where

$$SSTR = n \sum (\bar{r}_{.j} - \bar{r}_{..})^2 SSE = \sum \sum (r_{ij} - \bar{r}_{.j})^2 \bar{r}_{.j} = \frac{\sum_i r_{ij} \bar{r}_{..}}{n} = \frac{r+1}{2}$$

If there is no ties, it can be rewritten as

$$\chi_F^2 = \left[\frac{12}{nr(n+1)} \sum_j r_{.j}^2 \right] - 3n(r+1)$$

with large number of blocks, χ_F^2 is approximately χ_{r-1}^2 under H_0 . Hence, we reject H_0 if $\chi_F^2 > \chi_{(1-\alpha;r-1)}^2$

The exact null distribution for χ_F^2 can be derived since there are $r!$ possible ways of assigning ranks 1,2,...,r to the r observations within each block. There are n blocks and thus $(r!)^n$ possible assignments to the ranks, which are equally likely when H_0 is true.

19.3 Sample Size Planning for ANOVA

19.3.1 Balanced Designs

19.3.1.1 Single Factor Studies

19.3.1.1.1 Fixed cell means

$$P(F > f_{(1-\alpha;a-1,N-a)} | \phi) = 1 - \beta$$

where ϕ is the **noncentrality parameter** (measures how unequal the treatment means μ_i are)

$$\phi = \frac{1}{\sigma} \sqrt{\frac{n}{a} \sum_i (\mu_i - \mu_{.})^2}, (n_i \equiv n)$$

and

$$\mu_{..} = \frac{\sum \mu_i}{a}$$

To decide on the power probabilities we use the noncentral F distribution.

We could use the power table directly when effects are fixed and design is balanced by using **minimum range** of factor level means for your desired differences

$$\Delta = \max(\mu_i) - \min(\mu_i)$$

Hence, we need

- α level
- Δ
- σ
- β

Notes:

- When Δ/σ is small greatly affects sample size, but if Δ/σ is large.
- Reducing α or β increases the required sample sizes.
- Error in estimating σ can make a large difference.

19.3.1.2 Multi-factor Studies

The same noncentral F tables can be used here

For two-factor fixed effect model

Test for interactions:

$$\phi = \frac{1}{\sigma} \sqrt{\frac{n \sum \sum (\alpha \beta_{ij})^2}{(a-1)(b-1) + 1}} = \frac{1}{\sigma} \sqrt{\frac{n \sum \sum (\mu_{ij} - \mu_{i..} - \mu_{.j} + \mu_{...})^2}{(a-1)(b-1) + 1}} v_1 = (a-1)(b-1)v_2 = ab(n-1)$$

Test for Factor A main effects:

$$\phi = \frac{1}{\sigma} \sqrt{\frac{nb \sum \alpha_i^2}{a}} = \frac{1}{\sigma} \sqrt{\frac{nb \sum (\mu_{i..} - \mu_{...})^2}{a}} v_1 = a - 1v_2 = ab(n-1)$$

Test for Factor B main effects:

$$\phi = \frac{1}{\sigma} \sqrt{\frac{na \sum \beta_j^2}{b}} = \frac{1}{\sigma} \sqrt{\frac{na \sum (\mu_{.j} - \mu_{..})^2}{b}} v_1 = b - 1 v_2 = ab(n - 1)$$

Procedure:

1. Specify the minimum range of Factor A means
2. Obtain sample sizes with $r = a$. The resulting sample size is b_n , from which n can be obtained.
3. Repeat the first 2 steps for Factor B minimum range.
4. Choose the greater number of sample size between A and B.

19.3.2 Randomized Block Experiments

Analogous to completely randomized designs. The power of the F-test for treatment effects for randomized block design uses the same non-centrality parameter as completely randomized design:

$$\phi = \frac{1}{\sigma} \sqrt{\frac{n}{r} \sum (\mu_i - \mu_{..})^2}$$

However, the power level is different from the randomized block design because

- error variance σ^2 is different
- $\text{df}(\text{MSE})$ is different.

19.4 Randomized Block Designs

To improve the precision of treatment comparisons, we can reduce variability among the experimental units. We can group experimental units into **blocks** so that each block contains relatively homogeneous units.

- Within each block, random assignment treatments to units (separate random assignment for each block)
- The number of units per block is a multiple of the number of factor combinations.
- Commonly, use each treatment once in each block.

Benefits of Blocking

- Reduction in variability of estimators for treatment means

- Improved power for t-tests and F-tests
- Narrower confidence intervals
- Smaller MSE
- Compare treatments under different conditions (related to different blocks).

Loss from **Blocking** (little to lose)

- If you don't do blocking well, you waste df on negligible block effects that could have been used to estimate σ^2
- hence, the df for t-tests and denominator df for F-tests will be reduced without reducing MSE and small loss of power for both tests.

Consider

$$Y_{ij} = \mu_{..} + \rho_i + \tau_j + \epsilon_{ij} \quad i = 1, 2, \dots, n; j = 1, 2, \dots, r$$

where

- $\mu_{..}$: overall mean response, averaging across all blocks and treatments
- ρ_i : block effect, average difference in response for i-th block ($\sum \rho_i = 0$)
- τ_j treatment effect, average across blocks ($\sum \tau_j = 0$)
- $\epsilon_{ij} \sim iidN(0, \sigma^2)$: random experimental error.

Here, we assume that the block and treatment effects are additive. The difference in average response for any pair of treatments is the same **within** each block

$$(\mu_{..} + \rho_i + \tau_j) - (\mu_{..} + \rho_i + \tau'_j) = \tau_j - \tau'_j$$

for all $i = 1, \dots, n$ blocks

$$\hat{\mu} = \bar{Y}_{..} \hat{\rho}_i = \bar{Y}_{i.} - \bar{Y}_{..} \hat{\tau}_j = \bar{Y}_{.j} - \bar{Y}_{..}$$

Hence,

$$\hat{Y}_{ij} = \bar{Y}_{..} + (\bar{Y}_{i.} - \bar{Y}_{..}) + (\bar{Y}_{.j} - \bar{Y}_{..}) = \bar{Y}_{i.} + \bar{Y}_{.j} - \bar{Y}_{..} e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}$$

ANOVA table

Source of Varia- tion	SS	df	Fixed Treatments E(MS)	Random Treatments E(MS)
Blocks	$r \sum_i (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2$	n - 1	$\sigma^2 + r \frac{\sum \rho_i^2}{n-1}$	$\sigma^2 + r \frac{\sum \rho_i^2}{n-1}$
Treatments	$\sum_j (\bar{Y}_{.j} - \bar{Y}_{..})^2$	r - 1	$\sigma^2 + n \frac{\sum \tau_j^2}{r-1}$	$\sigma^2 + n \sigma_\tau^2$
Error	$\sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{.j} + \bar{Y}_{..})^2$	(n-1)(r-1)	σ^2	σ^2
Total	SSTO	nr-1		

F-tests

$$H_0 : \tau_1 = \tau_2 = \dots = \tau_r = 0 \quad \text{Fixed Treatment Effects}$$

$$H_a : \text{not all } \tau_j = 0$$

$$H_0 : \sigma_\tau^2 = 0 \quad \text{Random Treatment Effects}$$

$$H_a : \sigma_\tau^2 \neq 0$$

In both cases $F = \frac{MSTR}{MSE}$, reject H_0 if $F > f_{(1-\alpha; r-1, (n-1)(r-1))}$
we don't use F-test to compare blocks, because

- We have a priori that blocks are different
- Randomization is done “within” block.

To estimate the efficiency that was gained by blocking (relative to completely randomized design).

$$\hat{\sigma}_{CR}^2 = \frac{(n-1)MSBL + n(r-1)MSE}{nr-1}$$

$$\hat{\sigma}_{RB}^2 = MSE$$

$$\frac{\hat{\sigma}_{CR}^2}{\hat{\sigma}_{RB}^2} = \text{above 1}$$

then a completely randomized experiment would

$$(\frac{\hat{\sigma}_{CR}^2}{\hat{\sigma}_{RB}^2} - 1)\%$$

more observations than the randomized block design to get the same MSE

If batches are randomly selected then they are random effects. That is , if the experiment was repeated, a new sample of i batches would be selected,d yielding new values for $\rho_1, \rho_2, \dots, \rho_i$ then.

$$\rho_1, \rho_2, \dots, \rho_j \sim N(0, \sigma_\rho^2)$$

Then,

$$Y_{ij} = \mu_{..} + \rho_i + \tau_j + \epsilon_{ij}$$

where

- $\mu_{..}$ fixed
- ρ_i : random iid $N(0, \sigma_\rho^2)$
- τ_j fixed (or random) $\sum \tau_j = 0$
- $\epsilon_{ij} \sim iidN(0, \sigma^2)$

**Fixed Treatment&&

$$E(Y_{ij}) = \mu_{..} + \tau_j var(Y_{ij}) = \sigma_\rho^2 + \sigma^2 cov(Y_{ij}, Y_{ij'}) = \sigma^2, j \neq j' \text{ treatments within same block are correlated}$$

Correlation between 2 observations in the same block

$$\frac{\sigma_\rho^2}{\sigma^2 + \sigma_\rho^2}$$

The expected MS for the additive fixed treatment effect, random block effect is

Source	SS	E(MS)
Blocks	SSBL	$\sigma^2 + r\sigma_\rho^2$
Treatment	SSTR	$\sigma^2 + n \frac{\sum \tau_j^2}{r-1}$
Error	SSE	σ^2

Interactions and Blocks

without replications within each block for each treatment, we can't consider

interaction between block and treatment when the block effect is fixed. Hence, only in the random block effect, we have

$$Y_{ij} = \mu_{..} + \rho_i + \tau_j + (\rho\tau)_{ij} + \epsilon_{ij}$$

where

- $\mu_{..}$ constant
- $\rho_i \sim iidN(0, \sigma_\rho^2)$ random
- τ_j fixed ($\sum \tau_j = 0$)
- $(\rho\tau)_{ij} \sim N(0, \frac{r-1}{r} \sigma_{\rho\tau}^2)$ with $\sum_j (\rho\tau)_{ij} = 0$ for all i
- $cov((\rho\tau)_{ij}, (\rho\tau)_{ij'}) = -\frac{1}{r} \sigma_{\rho\tau}^2$ for $j \neq j'$
- $\epsilon_{ij} \sim iidN(0, \sigma^2)$ random

Note: a special case of mixed 2-factor model with 1 observation per “cell”

$$E(Y_{ij}) = \mu_{..} + \tau_j var(Y_{ij}) = \sigma_\rho^2 + \frac{r-1}{r} \sigma_{\rho\tau}^2 + \sigma^2 cov(Y_{ij}, Y_{ij'}) = \sigma_\rho^2 - \frac{1}{r} \sigma_{\rho\tau}^2, j \neq j' \text{ obs from the same block are correlated}$$

The sum of squares and degrees of freedom for interaction model are the same as those for the additive model. The difference exists only with the expected mean squares

Source	SS	df	E(MS)
Blocks	SSBL	n-1	$\sigma^2 + r\sigma_\rho^2$
Treatment	SSTR	r-1	$\sigma^2 + \sigma_{\rho\tau}^2 + n \frac{\sum \tau_j^2}{r-1}$
Error	SSE	(n-1)(r-1)	$\sigma^2 + \sigma_{\rho\tau}^2$

- No exact test is possible for block effects when interaction is present (Not important if blocks are used primarily to reduce experimental error variability)
- $E(MSE) = \sigma^2 + \sigma_{\rho\tau}^2$ the error term variance and interaction variance $\sigma_{\rho\tau}^2$. We can't estimate these components separately with this model. The two are **confounded**.
- If more than 1 observation per treatment block combination, one can consider interaction with fixed block effects, which is called **generalized randomized block designs** (multifactor analysis).

19.4.1 Tukey Test of Additivity

(Tukey's 1 df test for additivity)

formal test of interaction effects between blocks and treatments for a randomized block design. can also be considered for testing additivity in 2-way analyses when there is only one observation per cell.

we consider a less restricted interaction term

$$(\rho\tau)_{ij} = D\rho_i\tau_j (\text{D: Constant})$$

So,

$$Y_{ij} = \mu_{..} + \rho_i + \tau_j + D\rho_i\tau_j + \epsilon_{ij}$$

the least square estimate or MLE for D

$$\hat{D} = \frac{\sum_i \sum_j \rho_i \tau_j Y_{ij}}{\sum_i \rho_i^2 \sum_j \tau_j^2}$$

replacing the parameters by their estimates

$$\hat{D} = \frac{\sum_i \sum_j (\bar{Y}_{i.} - \bar{Y}_{..})(\bar{Y}_{.j} - \bar{Y}_{..})Y_{ij}}{\sum_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 \sum_j (\bar{Y}_{.j} - \bar{Y}_{..})^2}$$

Thus, the interaction sum of squares

$$SSint = \sum_i \sum_j \hat{D}^2 (\bar{Y}_{i.} - \bar{Y}_{..})^2 (\bar{Y}_{.j} - \bar{Y}_{..})^2$$

The ANOVA decomposition

$$SSTO = SSBL + SSTR + SSint + SSRem$$

where $SSRem$: remainder sum of squares

$$SSRem = SSTO - SSBL - SSTR - SSint$$

if $D = 0$ (i.e., no interactions of the type $D\rho_i\tau_j$). SSint and SSRem are independent $\chi^2_{1, rn-r-n}$.

If $D = 0$,

$$F = \frac{SSint/1}{SSRem/(rn - r - n)} \sim f_{(1-\alpha; rn-r-n)}$$

if

$H_0 : D = 0$ no interaction present $H_a : D \neq 0$ interaction of form $D\rho_i\tau_j$ present

we reject H_0 if $F > f_{(1-\alpha; 1, nr-r-n)}$

19.5 Nested Designs

Let μ_{ij} be the mean response when factor A is at the i-th level and factor B is at the j-th level.

If the factors are crossed, the jth level of B is the same for all levels of A.

If factor B is nested within A, the j-th level of B when A is at level 1 has nothing in common with the j-th level of B when A is at level 2.

Factors that can't be manipulated are designated as **classification factors**, as opposed to **experimental factors** (i.e., you assign to the experimental units).

19.5.1 Two-Factor Nested Designs

- Consider B is nested within A.
- both factors are fixed
- All treatment means are equally important.

Mean responses

$$\mu_{i\cdot} = \sum_j \mu_{ij}/b$$

Main effect factor A

$$\alpha_i = \mu_{i\cdot} - \mu_{\cdot\cdot}$$

$$\text{where } \mu_{\cdot\cdot} = \frac{\mu_{ij}}{ab} = \frac{\sum_i \mu_{i\cdot}}{a} \text{ and } \sum_i \alpha_i = 0$$

Individual effects of B is denoted as $\beta_{j(i)}$ where $j(i)$ indicates the j-th level of factor B is nested within the it-h level of factor A

$$\beta_{j(i)} = \mu_{ij} - \mu_{i..} = \mu_{ij} - \alpha_i - \mu_{...} \sum_j \beta_{j(i)} = 0, i = 1, \dots, a$$

$\beta_{j(i)}$ is the **specific effect** of the jth level of factor B nested within the ith level of factor A. Hence,

$$\mu_{ij} \equiv \mu_{...} + \alpha_i + \beta_{j(i)} \equiv \mu_{...} + (\mu_{i..} - \mu_{...}) + (\mu_{ij} - \mu_{i..})$$

Model

$$Y_{ijk} = \mu_{...} + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$$

where

- Y_{ijk} response for the kth treatment when factor A is at the i-th level and factor B is at the jth level ($i = 1, \dots, a$; $j = 1, \dots, b$; $k = 1, \dots, n$)
- $\mu_{...}$ constant
- α_i constants subject to restriction $\sum_i \alpha_i = 0$
- $\beta_{j(i)}$ constants subject to restriction $\sum_j \beta_{j(i)} = 0$ for all i
- $\epsilon_{ijk} \sim iidN(0, \sigma^2)$

$$E(Y_{ijk}) = \mu_{...} + \alpha_i + \beta_{j(i)} var(Y_{ijk}) = \sigma^2$$

there is no interaction term in a nested model

ANOVA for Two-Factor Nested Designs

Least Squares and MLE estimates

Parameter	Estimator
$\mu_{...}$	$\bar{Y}_{...}$
α_i	$\bar{Y}_{i..} - \bar{Y}_{...}$
$\beta_{j(i)}$	$\bar{Y}_{ij.} - \bar{Y}_{i..}$
\hat{Y}_{ijk}	$\bar{Y}_{ij.}$

$$\text{residual } e_{ijk} = Y_{ijk} - \bar{Y}_{ijk}$$

$$SSTO = SSA + SSB(A) + SSE$$

$$\sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{...})^2 = bn \sum_i (\bar{Y}_{i..} - \bar{Y}_{...})^2 + n \sum_i \sum_j (\bar{Y}_{ij.} - \bar{Y}_{i..})^2 + \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij.})^2$$

ANOVA Table

Source of Variation	SS	df	MS	E(MS)
Factor A	SSA	a-1	MSA	$\sigma^2 + bn \frac{\sum \alpha_i^2}{a-1}$
Factor B	SSB(A)	a(b-1)	MSB(A)	$\sigma^2 + n \frac{\sum um \sum eta_i^2}{a(b-1)}$
Error	SSE	ab(n-1)	MSE	σ^2
Total	SSTO	abn-1		

Tests For Factor Effects

$$H_0 : \text{All } \alpha_i = 0 H_a : \text{not all } \alpha_i = 0$$

$$F = \frac{MSA}{MSE} \sim f_{(1-\alpha; a-1, (n-1)ab)} \text{ reject if } F > f$$

$$H_0 : \text{All } \beta_{j(i)} = 0 H_a : \text{not all } \beta_{j(i)} = 0$$

$$F = \frac{MSB(A)}{MSE} \sim f_{(1-\alpha; a(b-1), (n-1)ab)} \text{ reject } F > f$$

Testing Factor Effect Contrasts

$$L = \sum c_i \mu_i \text{ where } \sum c_i = 0$$

$$\hat{L} = \sum c_i \bar{Y}_{i..} \hat{L} \pm t_{(1-\alpha/2; df)} s(\hat{L})$$

where $s^2(\hat{L}) = \sum c_i^2 s^2(\bar{Y}_{i..})$, where $s^2(\bar{Y}_{i..}) = \frac{MSE}{bn}$, $df = ab(n-1)$

Testing Treatment Means

$L = \sum c_i \mu_{.j}$ estimated by $\hat{L} = \sum c_i \bar{Y}_{ij}$ with confidence limits:

$$\hat{L} \pm t_{(1-\alpha/2; (n-1)ab)} s(\hat{L})$$

where

$$s^2(\hat{L}) = \frac{MSE}{n} \sum c_i^2$$

Unbalanced Nested Two-Factor Designs

If there are different number of levels of factor B for different levels of factor A, then the design is called **unbalanced**

The model

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_{j(i)} + \epsilon_{ijk} \quad i = 1, 2; j = 1, \dots, b_i; k = 1, \dots, n_{ij} \quad n_{i1} = 3, n_{i2} = 2, n_{11} = n_{13} = 2, n_{12} = 1, n_{21} = n_{22}$$

where $\alpha_1, \beta_{1(1)}, \beta_{2(1)}, \beta_{1(2)}$ are parameters. And constraints: $\alpha_2 = -\alpha_1, \beta_{3(1)} = -\beta_{1(1)} - \beta_{2(1)}, \beta_{2(2)} = -\beta_{1(2)}$

4 indicator variables

$$X_1 = \begin{cases} 1 & \text{if obs from school 1} \\ -1 & \text{if obs from school 2} \end{cases} \quad (19.4)$$

$$X_2 = \begin{cases} 1 & \text{if obs from instructor 1 in school 1} \\ -1 & \text{if obs from instructor 3 in school 1} \\ 0 & \text{otherwise} \end{cases} \quad (19.5)$$

$$X_3 = \begin{cases} 1 & \text{if obs from instructor 2 in school 1} \\ -1 & \text{if obs from instructor 3 in school 1} \\ 0 & \text{otherwise} \end{cases} \quad (19.6)$$

$$X_4 = \begin{cases} 1 & \text{if obs from instructor 1 in school 1} \\ -1 & \text{if obs from instructor 2 in school 1} \\ 0 & \text{otherwise} \end{cases} \quad (19.7)$$

Regression Full Model

$$Y_{ijk} = \mu_{..} + \alpha_1 X_{ijk1} + \beta_{1(1)} X_{ijk2} + \beta_{2(1)} X_{ijk3} + \beta_{1(2)} X_{ijk4} + \epsilon_{ijk}$$

Random Factor Effects

If

$$\alpha_1 \sim iidN(0, \sigma_\alpha^2) \quad \beta_{j(i)} \sim iidN(0, \sigma_\beta^2)$$

Mean Square	Expected Mean Squares A fixed, B random	Expected Mean Squares A random, B random
MSA	$\sigma^2 + n\sigma_\beta^2 + bn \sum_{a=1}^n \alpha_i^2$	$\sigma^2 + bn\sigma_\alpha^2 + n\sigma_\beta^2$
MSB(A)	$\sigma^2 + n\sigma_\beta^2$	$\sigma^2 + n\sigma_\beta^2$
MSE	σ^2	σ^2

Test Statistics

Factor A	$\frac{MSA}{MSB(A)}$	$\frac{MSA}{MSB(A)}$
Factor B(A)	$\frac{MSB(A)}{MSE}$	$\frac{MSB(A)}{MSE}$

Another way to increase the precision of treatment comparisons by reducing variability is to use regression models to adjust for differences among experimental units (also known as **analysis of covariance**).

19.6 Single Factor Covariance Model

$$Y_{ij} = \mu_{..} + \tau_i + \gamma(X_{ij} - \bar{X}_{..}) + \epsilon_{ij}$$

for $i = 1, \dots, r; j = 1, \dots, n_i$

where

- $\mu_{..}$ overall mean
- τ_i : fixed treatment effects ($\sum \tau_i = 0$)
- γ : fixed regression coefficient effect between X and Y
- X_{ij} covariate (not random)
- $\epsilon_{ij} \sim iidN(0, \sigma^2)$: random errors

If we just use γX_{ij} as the regression term (rather than $\gamma(X_{ij} - \bar{X}_{..})$), then $\mu_{..}$ is no longer the overall mean; thus we need to centered mean.

$$E(Y_{ij}) = \mu_{..} + \tau_i + \gamma(X_{ij} - \bar{X}_{..}) var(Y_{ij}) = \sigma^2$$

$Y_{ij} \sim N(\mu_{ij}, \sigma^2)$, where

$$\mu_{ij} = \mu_+ + \tau_i + \gamma(X_{ij} - \bar{X}_{..}) \sum \tau_i = 0$$

Thus, the mean response (μ_{ij}) is a regression line with intercept $\mu_+ + \tau_i$ and slope γ for each treatment i .

Assumption:

- All treatment regression lines have the same slope
- when treatment interact with covariate X (non-parallel slopes), covariance analysis is **not** appropriate. in which case we should use separate regression lines.

More complicated regression features (e.g., quadratic, cubic) or additional covariates e.g.,

$$Y_{ij} = \mu_+ + \tau_i + \gamma_1(X_{ij1} - \bar{X}_{..2}) + \gamma_2(X_{ij2} - \bar{X}_{..2}) + \epsilon_{ij}$$

Regression Formulation

We can use indicator variables for treatments

$$l_1 = \begin{cases} 1 & \text{if case is from treatment 1} \\ -1 & \text{if case is from treatment r} \\ 0 & \text{otherwise} \end{cases} \dots l_{r-1} = \begin{cases} 1 & \text{if case is from treatment r-1} \\ -1 & \text{if case is from treatment r} \\ 0 & \text{otherwise} \end{cases}$$

Let $x_{ij} = X_{ij} - \bar{X}_{..}$. the regression model is

$$Y_{ij} = \mu_+ + \tau_1 l_{ij,1} + \dots + \tau_{r-1} l_{ij,r-1} + \gamma x_{ij} + \epsilon_{ij}$$

where $I_{ij,1}$ is the indicator variable l_1 for the j -th case from treatment i . The treatment effect $\tau_1, \dots, \tau_{r-1}$ are just regression coefficients for the indicator variables.

We could use the same diagnostic tools for this case.

Inference

Treatment effects

$$H_0 : \tau_1 = \tau_2 = \dots = 0 \quad H_a : \text{not all } \tau_i = 0$$

Full Model : $Y_{ij} = \mu_+ + \tau_i + \gamma X_{ij} + \epsilon_{ij}$ Reduced Model : $Y_{ij} = \mu_+ + \gamma X_{ij} + \epsilon_{ij}$

$$F = \frac{SSE(R) - SSE(F)}{(N-2) - (N-(r+1))} / \frac{SSE(F)}{N-(r+1)} \sim F_{(r-1, N-(r+1))}$$

If we are interested in comparisons of treatment effects.

For example, r = 3. We estimate $\tau_1, \tau_2, \tau_3 = -\tau_1 - \tau_2$

Comparison	Estimate	Variance of Estimator
$\tau_1 - \tau_2$	$\hat{\tau}_1 - \hat{\tau}_2$	$var(\hat{\tau}_1) + var(\hat{\tau}_2) - 2cov(\hat{\tau}_1 \hat{\tau}_2)$
$\tau_1 - \tau_3$	$2\hat{\tau}_1 + \hat{\tau}_2$	$4var(\hat{\tau}_1) + var(\hat{\tau}_2) - 4cov(\hat{\tau}_1 \hat{\tau}_2)$
$\tau_2 - \tau_3$	$\hat{\tau}_1 + 2\hat{\tau}_2$	$var(\hat{\tau}_1) + 4var(\hat{\tau}_2) - 4cov(\hat{\tau}_1 \hat{\tau}_2)$

Testing for Parallel Slopes

Example:

r = 3

$$Y_{ij} = \mu_+ + \tau_1 I_{ij,1} + \tau_2 I_{ij,2} + \gamma X_{ij} + \beta_1 I_{ij,1} X_{ij} + \beta_2 I_{ij,2} X_{ij} + \epsilon_{ij}$$

where β_1, β_2 : interaction coefficients.

$$H_0 : \beta_1 = \beta_2 = 0 \quad H_a : \text{at least one } \beta \neq 0$$

If we can't reject H_0 using F-test then we have evidence that the slopes are parallel

Adjusted Means

The means in response after adjusting for the covariate effect

$$Y_{i.}(adj) = \bar{Y}_{i.} - \hat{\gamma}(\bar{X}_{i.} - \bar{X}_{..})$$

Chapter 20

Multivariate Methods

y_1, \dots, y_p are possibly correlated random variables with means μ_1, \dots, μ_p

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix}$$

$$E(\mathbf{y}) = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}$$

Let $\sigma_{ij} = cov(y_i, y_j)$ for $i, j = 1, \dots, p$

$$= (\sigma_{ij}) = \begin{pmatrix} \sigma_{11} & \sigma_{22} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{pmatrix}$$

where (symmetric) is the variance-covariance or dispersion matrix

Let $\mathbf{u}_{p \times 1}$ and $\mathbf{v}_{q \times 1}$ be random vectors with means μ_u and μ_v . Then

$$uv = cov(\mathbf{u}, \mathbf{v}) = E[(\mathbf{u} - \mu_u)(\mathbf{v} - \mu_v)']$$

in which $uv \neq vu$ and $uv = vu'$

Properties of Covariance Matrices

1. Symmetric $' =$

2. Non-negative definite $\mathbf{a}' \mathbf{a} \geq 0$ for any $\mathbf{a} \in R^p$, which is equivalent to eigenvalues of $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$
3. $\| = \lambda_1 \lambda_2 \dots \lambda_p \geq 0$ (**generalized variance**) (the bigger this number is, the more variation there is)
4. $\text{trace}(\mathbf{A}) = \text{tr}(\mathbf{A}) = \lambda_1 + \dots + \lambda_p = \sigma_{11} + \dots + \sigma_{pp} = \text{sum of variance}$ (**total variance**)

Note:

- \mathbf{A} is typically required to be positive definite, which means all eigenvalues are positive, and \mathbf{A} has an inverse \mathbf{A}^{-1} such that $\mathbf{A}^{-1} = \mathbf{I}_{p \times p} = \mathbf{A}^{-1}$

Correlation Matrices

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

$$\mathbf{R} = \begin{pmatrix} \rho_{11} & \rho_{12} & \dots & \rho_{1p} \\ \rho_{21} & \rho_{22} & \dots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \dots & \rho_{pp} \end{pmatrix}$$

where ρ_{ij} is the correlation, and $\rho_{ii} = 1$ for all i

Alternatively,

$$\mathbf{R} = [\text{diag}(\mathbf{A})]^{-1/2} [\text{diag}(\mathbf{A})]^{-1/2}$$

where $\text{diag}(\mathbf{A})$ is the matrix which has the σ_{ii} 's on the diagonal and 0's elsewhere and $\mathbf{A}^{1/2}$ (the square root of a symmetric matrix) is a symmetric matrix such as $\mathbf{A} = \mathbf{A}^{1/2} \mathbf{A}^{1/2}$

Equalities

Let

- \mathbf{x} and \mathbf{y} be random vectors with means μ_x and μ_y and variance -variance matrices Σ_x and Σ_y .
- \mathbf{A} and \mathbf{B} be matrices of constants and \mathbf{c} and \mathbf{d} be vectors of constants

Then

- $E(\mathbf{A}\mathbf{y} + \mathbf{c}) = \mathbf{A}\mu_y + \mathbf{c}$

- $\text{var}(\mathbf{A}\mathbf{y} + \mathbf{c}) = \mathbf{A}\text{var}(\mathbf{y})\mathbf{A}' = \mathbf{A}_{\mathbf{y}}\mathbf{A}'$
- $\text{cov}(\mathbf{A}\mathbf{y} + \mathbf{c}, \mathbf{B}\mathbf{y} + \mathbf{d}) = \mathbf{A}_{\mathbf{y}}\mathbf{B}'$
- $E(\mathbf{A}\mathbf{y} + \mathbf{B}\mathbf{x} + \mathbf{c}) = \mathbf{A}_{\mathbf{y}} + \mathbf{B}_{\mathbf{x}} + \mathbf{c}$
- $\text{var}(\mathbf{A}\mathbf{y} + \mathbf{B}\mathbf{x} + \mathbf{c}) = \mathbf{A}_{\mathbf{y}}\mathbf{A}' + \mathbf{B}_{\mathbf{x}}\mathbf{B}' + \mathbf{A}_{\mathbf{y}\mathbf{x}}\mathbf{B}' + \mathbf{B}'_{\mathbf{y}\mathbf{x}}\mathbf{A}'$

Multivariate Normal Distribution

Let \mathbf{y} be a multivariate normal (MVN) random variable with mean μ and variance Σ . Then the density of \mathbf{y} is

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp(-\frac{1}{2}(\mathbf{y} - \mu)' \Sigma^{-1} (\mathbf{y} - \mu))$$

$$\mathbf{y} \sim N_p(\mu, \Sigma)$$

20.0.1 Properties of MVN

- Let $\mathbf{A}_{r \times p}$ be a fixed matrix. Then $\mathbf{A}\mathbf{y} \sim N_r(\mathbf{A}\mu, \mathbf{A}\mathbf{A}')$. $r \leq p$ and all rows of \mathbf{A} must be linearly independent to guarantee that $\mathbf{A}\mathbf{A}'$ is non-singular.
- Let \mathbf{G} be a matrix such that $\Sigma^{-1} = \mathbf{G}\mathbf{G}'$. Then $\mathbf{G}'\mathbf{y} \sim N_p(\mathbf{G}'\mu, \mathbf{I})$ and $\mathbf{G}'(\mathbf{y} - \mu) \sim N_p(0, \mathbf{I})$
- Any fixed linear combination of y_1, \dots, y_p (say $\mathbf{c}'\mathbf{y}$) follows $\mathbf{c}'\mathbf{y} \sim N_1(\mathbf{c}'\mu, \mathbf{c}'\Sigma\mathbf{c})$
- Define a partition, $[\mathbf{y}_1', \mathbf{y}_2']'$ where
 - \mathbf{y}_1 is $p_1 \times 1$
 - \mathbf{y}_2 is $p_2 \times 1$,
 - $p_1 + p_2 = p$
 - $p_1, p_2 \geq 1$ Then

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 11 & 12 \\ 21 & 22 \end{pmatrix} \right)$$

- The marginal distributions of \mathbf{y}_1 and \mathbf{y}_2 are $\mathbf{y}_1 \sim N_{p_1}(\mu_1, \Sigma_{11})$ and $\mathbf{y}_2 \sim N_{p_2}(\mu_2, \Sigma_{22})$
- Individual components y_1, \dots, y_p are all normally distributed $y_i \sim N_1(\mu_i, \sigma_{ii})$

- The conditional distribution of \mathbf{y}_1 and \mathbf{y}_2 is normal

$$-\mathbf{y}_1|\mathbf{y}_2 \sim N_{p1}(\mathbf{1} + \mathbf{1}\mathbf{2}^{-1}\mathbf{2}\mathbf{2}^{-1}(\mathbf{y}_2 - \mathbf{2}), \mathbf{1}\mathbf{1} - \mathbf{1}\mathbf{2}^{-1}\mathbf{2}\mathbf{2}^{-1}\mathbf{2}\mathbf{1})$$

* In this formula, we see if we know (have info about) \mathbf{y}_2 , we can re-weight \mathbf{y}_1 's mean, and the variance is reduced because we know more about \mathbf{y}_1 because we know \mathbf{y}_2

- which is analogous to $\mathbf{y}_2|\mathbf{y}_1$. And \mathbf{y}_1 and \mathbf{y}_2 are independently distributed only if $\mathbf{1}_2 = 0$

- If $\mathbf{y} \sim N(\mu, \Sigma)$ and Σ is positive definite, then $(\mathbf{y} - \mu)' \Sigma^{-1}(\mathbf{y} - \mu) \sim \chi^2_{(p)}$

- If \mathbf{y}_i are independent $N_p(\mu_i, \Sigma_i)$ random variables, then for fixed matrices $\mathbf{A}_{i(m \times p)}$, $\sum_{i=1}^k \mathbf{A}_i \mathbf{y}_i \sim N_m(\sum_{i=1}^k \mathbf{A}_i \mu_i, \sum_{i=1}^k \mathbf{A}_i \Sigma_i \mathbf{A}_i')$

Multiple Regression

$$\begin{pmatrix} Y \\ \mathbf{x} \end{pmatrix} \sim N_{p+1} \left(\begin{bmatrix} \mu_y \\ x \end{bmatrix}, \begin{bmatrix} \sigma_Y^2 & yx \\ yx & xx \end{bmatrix} \right)$$

The conditional distribution of Y given \mathbf{x} follows a univariate normal distribution with

$$\begin{aligned} E(Y|\mathbf{x}) &= \mu_y + yx \Sigma_{xx}^{-1}(\mathbf{x} - \mu_x) \\ &= \mu_y - \Sigma_{yx} \Sigma_{xx}^{-1} \mu_x + \Sigma_{yx} \Sigma_{xx}^{-1} \mathbf{x} \\ &= \beta_0 + \mathbf{x}' \mathbf{\beta} \end{aligned}$$

where $\beta = (\beta_1, \dots, \beta_p)' = \Sigma_{xx}^{-1} yx$ (e.g., analogous to $(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$ but not the same if we consider Y_i and \mathbf{x}_i , $i = 1, \dots, n$ and use the empirical covariance formula: $var(Y|\mathbf{x}) = \sigma_Y^2 - yx \Sigma_{xx}^{-1} yx'$)

Samples from Multivariate Normal Populations

A random sample of size n , $\mathbf{y}_1, \dots, \mathbf{y}_n$ from $N_p(\mu, \Sigma)$. Then

- Since $\mathbf{y}_1, \dots, \mathbf{y}_n$ are iid, their sample mean, $\bar{\mathbf{y}} = \sum_{i=1}^n \mathbf{y}_i / n \sim N_p(\mu, \Sigma/n)$, that is, $\bar{\mathbf{y}}$ is an unbiased estimator of μ
- The $p \times p$ sample variance-covariance matrix, \mathbf{S} is $\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})' = \frac{1}{n-1} (\sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i' - n \bar{\mathbf{y}} \bar{\mathbf{y}}')$
 - where \mathbf{S} is symmetric, unbiased estimator of Σ and has $p(p+1)/2$ random variables.
- $(n-1)\mathbf{S} \sim W_p(n-1, \Sigma)$ is a Wishart distribution with $n-1$ degrees of freedom and expectation $(n-1)\Sigma$. The Wishart distribution is a multivariate extension of the Chi-squared distribution.

- $\bar{\mathbf{y}}$ and \mathbf{S} are independent
- $\bar{\mathbf{y}}$ and \mathbf{S} are sufficient statistics. (All of the info in the data about μ and Σ is contained in $\bar{\mathbf{y}}$ and \mathbf{S} , regardless of sample size).

Large Sample Properties

$\mathbf{y}_1, \dots, \mathbf{y}_n$ are a random sample from some population with mean μ and variance-covariance matrix Σ

- $\bar{\mathbf{y}}$ is a consistent estimator for μ
- \mathbf{S} is a consistent estimator for Σ
- **Multivariate Central Limit Theorem:** Similar to the univariate case, $\sqrt{n}(\bar{\mathbf{y}} - \mu) \xrightarrow{D} N_p(\mathbf{0}, \Sigma)$ where n is large relative to p ($n \geq 25p$), which is equivalent to $\bar{\mathbf{y}} \xrightarrow{D} N_p(\mu, \Sigma/n)$
- **Wald's Theorem:** $n(\bar{\mathbf{y}} - \mu)' \mathbf{S}^{-1} (\bar{\mathbf{y}} - \mu)$ when n is large relative to p .

Maximum Likelihood Estimation for MVN

Suppose iid $\mathbf{y}_1, \dots, \mathbf{y}_n \sim N_p(\mu, \Sigma)$, the likelihood function for the data is

$$\begin{aligned} L(\mu, \Sigma) &= \prod_{j=1}^n \left(\frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}_j - \mu)' \Sigma^{-1} (\mathbf{y}_j - \mu)\right) \right) \\ &= \frac{1}{(2\pi)^{np/2} |\Sigma|^{n/2}} \exp\left(-\frac{1}{2} \sum_{j=1}^n (\mathbf{y}_j - \mu)' \Sigma^{-1} (\mathbf{y}_j - \mu)\right) \end{aligned}$$

Then, the MLEs are

$$\hat{\mu} = \bar{\mathbf{y}}$$

$$\hat{\Sigma} = \frac{n-1}{n} \mathbf{S}$$

using derivatives of the log of the likelihood function with respect to μ and Σ

Properties of MLEs

- Invariance: If $\hat{\theta}$ is the MLE of θ , then the MLE of $h(\theta)$ is $h(\hat{\theta})$ for any function $h(\cdot)$
- Consistency: MLEs are consistent estimators, but they are usually biased

- Efficiency: MLEs are efficient estimators (no other estimator has a smaller variance for large samples)
- Asymptotic normality: Suppose that $\hat{\theta}_n$ is the MLE for θ based upon n independent observations. Then $\hat{\theta}_n \sim N(\theta, \mathbf{H}^{-1})$
 - \mathbf{H} is the Fisher Information Matrix, which contains the expected values of the second partial derivatives of the log-likelihood function. the (i,j)th element of \mathbf{H} is $-E\left(\frac{\partial^2 l(\cdot)}{\partial \theta_i \partial \theta_j}\right)$
 - we can estimate \mathbf{H} by finding the form determined above, and evaluate it at $\theta = \hat{\theta}_n$
- Likelihood ratio testing: for some null hypothesis, H_0 we can form a likelihood ratio test
 - The statistic is: $\Lambda = \frac{\max_{H_0} l(\cdot, | \mathbf{Y})}{\max l(\mu, | \mathbf{Y})}$
 - For large n, $-2 \log \Lambda \sim \chi^2_{(v)}$ where v is the number of parameters in the unrestricted space minus the number of parameters under H_0

Test of Multivariate Normality

- Check univariate normality for each trait (X) separately
 - Can check
Normality Assessment
 - The good thing is that if any of the univariate trait is not normal, then the joint distribution is not normal (see again [m]). If a joint multivariate distribution is normal, then the marginal distribution has to be normal.
 - However, marginal normality of all traits does not imply joint MVN
 - Easily rule out multivariate normality, but not easy to prove it
- Mardia's tests for multivariate normality
 - Multivariate skewness is
$$\beta_{1,p} = E[(\mathbf{y} - \bar{\mathbf{y}})' \mathbf{A}^{-1} (\mathbf{x} - \bar{\mathbf{x}})]^3$$
 - where \mathbf{x} and \mathbf{y} are independent, but have the same distribution (note: β here is not regression coefficient)
 - Multivariate kurtosis is defined as
 -
 - $$\beta_{2,p} = E[(\mathbf{y} - \bar{\mathbf{y}})' \mathbf{A}^{-1} (\mathbf{x} - \bar{\mathbf{x}})]^2$$

- For the MVN distribution, we have $\beta_{1,p} = 0$ and $\beta_{2,p} = p(p+2)$
- For a sample of size n , we can estimate

$$\hat{\beta}_{1,p} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n g_{ij}^2$$

$$\hat{\beta}_{2,p} = \frac{1}{n} \sum_{i=1}^n g_{ii}^2$$

* where $g_{ij} = (\mathbf{y}_i - \bar{\mathbf{y}})' \mathbf{S}^{-1} (\mathbf{y}_j - \bar{\mathbf{y}})$. Note: $g_{ii} = d_i^2$ where d_i^2 is the Mahalanobis distance

- (MARDIA, 1970) shows for large n

$$\kappa_1 = \frac{n\hat{\beta}_{1,p}}{6} \sim \chi_{p(p+1)(p+2)/6}^2$$

$$\kappa_2 = \frac{\hat{\beta}_{2,p} - p(p+2)}{\sqrt{8p(p+2)/n}} \sim N(0, 1)$$

- * Hence, we can use κ_1 and κ_2 to test the null hypothesis of MVN.
- * When the data are non-normal, normal theory tests on the mean are sensitive to $\beta_{1,p}$, while tests on the covariance are sensitive to $\beta_{2,p}$

- Alternatively, Doornik-Hansen test for multivariate normality (Doornik and Hansen, 2008)
- Chi-square Q-Q plot
 - Let $\mathbf{y}_i, i = 1, \dots, n$ be a random sample from $N_p(\mu, \Sigma)$
 - Then $\mathbf{z}_i = \Sigma^{-1/2}(\mathbf{y}_i - \mu), i = 1, \dots, n$ are iid $N_p(\mathbf{0}, \mathbf{I})$. Thus, $d_i^2 = \mathbf{z}_i' \mathbf{z}_i \sim \chi_p^2, i = 1, \dots, n$
 - plot the ordered d_i^2 values against the quantiles of the χ_p^2 distribution. When normality holds, the plot should approximately resemble a straight line passing through the origin at a 45 degree
 - it requires large sample size (i.e., sensitive to sample size). Even if we generate data from a MVN, the tail of the Chi-square Q-Q plot can still be out of line.
- If the data are not normal, we can
 - ignore it
 - use nonparametric methods
 - use models based upon an approximate distribution (e.g., GLMM)

- try performing a transformation

```

library(heplots)
library(ICSNP)
library(MVN)
library(tidyverse)

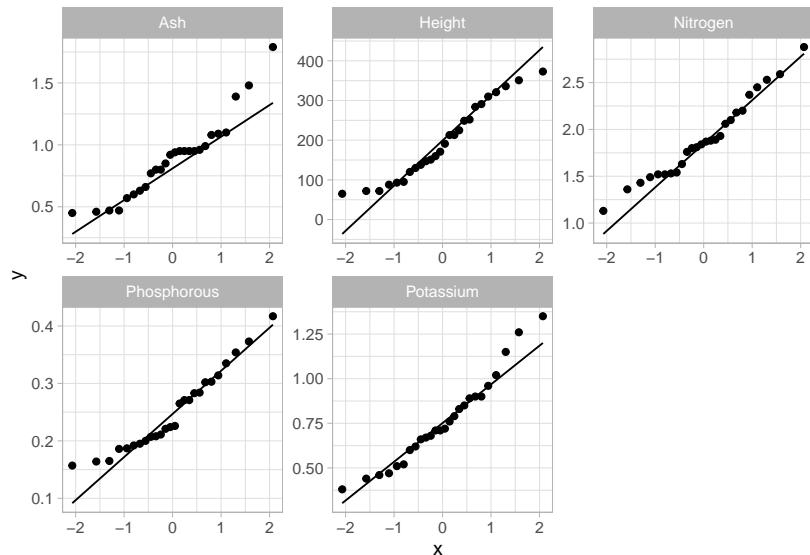
trees = read.table("images/trees.dat")
names(trees) <- c("Nitrogen", "Phosphorous", "Potassium", "Ash", "Height")
str(trees)
#> 'data.frame': 26 obs. of 5 variables:
#> $ Nitrogen : num 2.2 2.1 1.52 2.88 2.18 1.87 1.52 2.37 2.06 1.84 ...
#> $ Phosphorous: num 0.417 0.354 0.208 0.335 0.314 0.271 0.164 0.302 0.373 0.265 ...
#> $ Potassium : num 1.35 0.9 0.71 0.9 1.26 1.15 0.83 0.89 0.79 0.72 ...
#> $ Ash        : num 1.79 1.08 0.47 1.48 1.09 0.99 0.85 0.94 0.8 0.77 ...
#> $ Height     : int 351 249 171 373 321 191 225 291 284 213 ...

summary(trees)
#>      Nitrogen      Phosphorous      Potassium       Ash
#> Min.   :1.130   Min.   :0.1570   Min.   :0.3800   Min.   :0.4500
#> 1st Qu.:1.532   1st Qu.:0.1963   1st Qu.:0.6050   1st Qu.:0.6375
#> Median :1.855   Median :0.2250   Median :0.7150   Median :0.9300
#> Mean   :1.896   Mean   :0.2506   Mean   :0.7619   Mean   :0.8873
#> 3rd Qu.:2.160   3rd Qu.:0.2975   3rd Qu.:0.8975   3rd Qu.:0.9825
#> Max.   :2.880   Max.   :0.4170   Max.   :1.3500   Max.   :1.7900
#> 
#>      Height
#> Min.   : 65.0
#> 1st Qu.:122.5
#> Median :181.0
#> Mean   :196.6
#> 3rd Qu.:276.0
#> Max.   :373.0
cor(trees, method = "pearson") # correlation matrix
#>           Nitrogen Phosphorous Potassium      Ash      Height
#> Nitrogen    1.0000000  0.6023902 0.5462456 0.6509771 0.8181641
#> Phosphorous 0.6023902  1.0000000 0.7037469 0.6707871 0.7739656
#> Potassium   0.5462456  0.7037469 1.0000000 0.6710548 0.7915683
#> Ash         0.6509771  0.6707871 0.6710548 1.0000000 0.7676771
#> Height      0.8181641  0.7739656 0.7915683 0.7676771 1.0000000

# qq-plot
gg <- trees %>%
  pivot_longer(everything(), names_to = "Var", values_to = "Value") %>%
  ggplot(aes(sample = Value)) +
  geom_qq()

```

```
geom_qq_line() +
facet_wrap("Var", scales = "free")
gg
```



```
# Univariate normality
sw_tests <- apply(trees, MARGIN = 2, FUN = shapiro.test)
sw_tests
#> $Nitrogen
#>
#> Shapiro-Wilk normality test
#>
#> data: newX[, i]
#> W = 0.96829, p-value = 0.5794
#>
#> $Phosphorous
#>
#> Shapiro-Wilk normality test
#>
#> data: newX[, i]
#> W = 0.93644, p-value = 0.1104
#>
#> $Potassium
#>
```

```
#> Shapiro-Wilk normality test
#>
#> data: newX[, i]
#> W = 0.95709, p-value = 0.3375
#>
#> $Ash
#>
#> Shapiro-Wilk normality test
#>
#> data: newX[, i]
#> W = 0.92071, p-value = 0.04671
#>
#> $Height
#>
#> Shapiro-Wilk normality test
#>
#> data: newX[, i]
#> W = 0.94107, p-value = 0.1424
# Kolmogorov-Smirnov test
ks_tests <- map(trees, ~ ks.test(scale(.x), "pnorm"))
ks_tests
#> $Nitrogen
#>
#> One-sample Kolmogorov-Smirnov test
#>
#> data: scale(.x)
#> D = 0.12182, p-value = 0.8351
#> alternative hypothesis: two-sided
#>
#> $Phosphorous
#>
#> One-sample Kolmogorov-Smirnov test
#>
#> data: scale(.x)
#> D = 0.17627, p-value = 0.3944
#> alternative hypothesis: two-sided
#>
#> $Potassium
#>
#> One-sample Kolmogorov-Smirnov test
#>
```

```

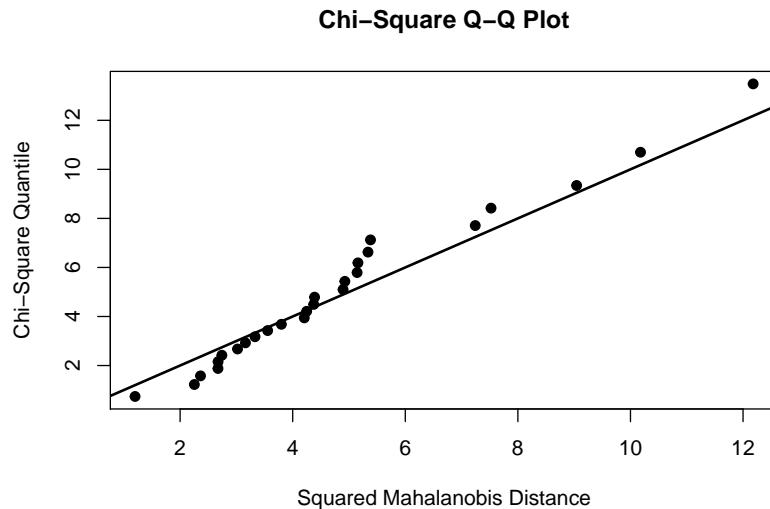
#> data: scale(.x)
#> D = 0.10542, p-value = 0.9348
#> alternative hypothesis: two-sided
#>
#>
#> $Ash
#>
#> One-sample Kolmogorov-Smirnov test
#>
#> data: scale(.x)
#> D = 0.14503, p-value = 0.6449
#> alternative hypothesis: two-sided
#>
#>
#> $Height
#>
#> One-sample Kolmogorov-Smirnov test
#>
#> data: scale(.x)
#> D = 0.1107, p-value = 0.9076
#> alternative hypothesis: two-sided

# Mardia's test, need large sample size for power
mardia_test <-
  mvn(
    trees,
    mvnTest = "mardia",
    covariance = FALSE,
    multivariatePlot = "qq"
  )

mardia_test$multivariateNormality
#>           Test      Statistic      p value Result
#> 1 Mardia Skewness 29.7248528871795 0.72054426745778 YES
#> 2 Mardia Kurtosis -1.67743173185383 0.0934580886477281 YES
#> 3            MVN             <NA>          <NA>     YES

# Doornik-Hansen's test
dh_test <-
  mvn(
    trees,
    mvnTest = "dh",
    covariance = FALSE,
    multivariatePlot = "qq"
  )

```



```

dh_test$multivariateNormality
#>           Test      E df      p value MVN
#> 1 Doornik-Hansen 161.9446 10 1.285352e-29 NO

# Henze-Zirkler's test
hz_test <-
  mvn(
    trees,
    mvnTest = "hz",
    covariance = FALSE,
    multivariatePlot = "qq"
  )
hz_test$multivariateNormality
#>           Test      HZ      p value MVN
#> 1 Henze-Zirkler 0.7591525 0.6398905 YES
# The last column indicates whether dataset follows a multivariate normality or not (i.e., NO or YES)

# Royston's test
# can only apply for 3 < obs < 5000 (because of Shapiro-Wilk's test)
royston_test <-
  mvn(
    trees,
    mvnTest = "royston",
    covariance = FALSE,
    multivariatePlot = "qq"
  )

```

```

royston_test$multivariateNormality
#>      Test      H      p value MVN
#> 1 Royston 9.064631 0.08199215 YES

# E-statistic
estat_test <-
  mvn(
    trees,
    mvnTest = "energy",
    covariance = FALSE,
    multivariatePlot = "qq"
  )
estat_test$multivariateNormality
#>      Test Statistic p value MVN
#> 1 E-statistic 1.091101 0.554 YES

```

20.0.2 Mean Vector Inference

In the univariate normal distribution, we test $H_0 : \mu = \mu_0$ by using

$$T = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

under the null hypothesis. And reject the null if $|T|$ is large relative to $t_{(1-\alpha/2, n-1)}$ because it means that seeing a value as large as what we observed is rare if the null is true

Equivalently,

$$T^2 = \frac{(\bar{y} - \mu_0)^2}{s^2/n} = n(\bar{y} - \mu_0)(s^2)^{-1}(\bar{y} - \mu_0) \sim f_{(1, n-1)}$$

20.0.2.1 Natural Multivariate Generalization

$$H_0 : \mathbf{y}_0 = \mathbf{H}_a : \mathbf{y} \neq \mathbf{y}_0$$

Define **Hotelling's T^2** by

$$T^2 = n(\bar{\mathbf{y}} - \mathbf{y}_0)' \mathbf{S}^{-1} (\bar{\mathbf{y}} - \mathbf{y}_0)$$

which can be viewed as a generalized distance between $\bar{\mathbf{y}}$ and \mathbf{y}_0

Under the assumption of normality,

$$F = \frac{n-p}{(n-1)p} T^2 \sim f_{(p,n-p)}$$

and reject the null hypothesis when $F > f_{(1-\alpha,p,n-p)}$

- The T^2 test is invariant to changes in measurement units.
 - If $\mathbf{z} = \mathbf{C}\mathbf{y} + \mathbf{d}$ where \mathbf{C} and \mathbf{d} do not depend on \mathbf{y} , then $T^2(\mathbf{z}) = T^2(\mathbf{y})$
- The T^2 test can be derived as a **likelihood ratio** test of $H_0 : \mu = \mu_0$

20.0.2.2 Confidence Intervals

20.0.2.2.1 Confidence Region An “exact” $100(1-\alpha)\%$ confidence region for μ is the set of all vectors, \mathbf{v} , which are “close enough” to the observed mean vector, $\bar{\mathbf{y}}$ to satisfy

$$n(\bar{\mathbf{y}} - \mu_0)' \mathbf{S}^{-1} (\bar{\mathbf{y}} - \mu_0) \leq \frac{(n-1)p}{n-p} f_{(1-\alpha,p,n-p)}$$

- \mathbf{v} are just the mean vectors that are not rejected by the T^2 test when $\bar{\mathbf{y}}$ is observed.

In case that you have 2 parameters, the confidence region is a “hyper-ellipsoid”.

In this region, it consists of all μ_0 vectors for which the T^2 test would not reject H_0 at significance level α

Even though the confidence region better assesses the joint knowledge concerning plausible values of μ , people typically include confidence statement about the individual component means. We’d like all of the separate confidence statements to hold **simultaneously** with a specified high probability. Simultaneous confidence intervals: intervals **against** any statement being incorrect

20.0.2.2.1.1 Simultaneous Confidence Statements

- Intervals based on a rectangular confidence region by projecting the previous region onto the coordinate axes:

$$\bar{y}_i \pm \sqrt{\frac{(n-1)p}{n-p} f_{(1-\alpha,p,n-p)} \frac{s_{ii}}{n}}$$

for all $i = 1, \dots, p$

which implied confidence region is conservative; it has at least $100(1 - \alpha)\%$

Generally, simultaneous $100(1 - \alpha)\%$ confidence intervals for all linear combinations $\mathbf{a}'\mathbf{y}$ of the elements of the mean vector are given by

$$\mathbf{a}'\hat{\mathbf{y}} \pm \sqrt{\frac{(n-1)p}{n-p} f_{(1-\alpha, p, n-p)} \frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{n}}$$

- works for any arbitrary linear combination $\mathbf{a}' = a_1\mu_1 + \dots + a_p\mu_p$, which is a projection onto the axis in the direction of \mathbf{a}
- These intervals have the property that the probability that at least one such interval does not contain the appropriate \mathbf{a}' is no more than α
- These types of intervals can be used for “data snooping” (like

Scheffé

)

20.0.2.2.1.2 One μ at a time

- One at a time confidence intervals:

$$\bar{y}_i \pm t_{(1-\alpha/2, n-1)} \sqrt{\frac{s_{ii}}{n}}$$

- Each of these intervals has a probability of $1 - \alpha$ of covering the appropriate μ_i
- But they ignore the covariance structure of the p variables
- If we only care about k simultaneous intervals, we can use “one at a time” method with the

Bonferroni

correction.

- This method gets more conservative as the number of intervals k increases.

20.0.3 General Hypothesis Testing

20.0.3.1 One-sample Tests

$$H_0 : \mathbf{C} = \mathbf{0}$$

where

- \mathbf{C} is a $c \times p$ matrix of rank c where $c \leq p$

We can test this hypothesis using the following statistic

$$F = \frac{n - c}{(n - 1)c} T^2$$

where $T^2 = n(\mathbf{C}\bar{\mathbf{y}})'(\mathbf{C}\mathbf{S}\mathbf{C}')^{-1}(\mathbf{C}\bar{\mathbf{y}})$

Example:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_p$$

Equivalently,

$$\mu_1 - \mu_2 = 0 : \mu_{p-1} - \mu_p = 0$$

a total of $p - 1$ tests. Hence, we have \mathbf{C} as the $p - 1 \times p$ matrix

$$\mathbf{C} = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & -1 \end{pmatrix}$$

number of rows = $c = p - 1$

Equivalently, we can also compare all of the other means to the first mean. Then, we test $\mu_1 - \mu_2 = 0, \mu_1 - \mu_3 = 0, \dots, \mu_1 - \mu_p = 0$, the $(p - 1) \times p$ matrix \mathbf{C} is

$$\mathbf{C} = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 \\ -1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & 0 & \dots & 0 & 1 \end{pmatrix}$$

The value of T^2 is invariant to these equivalent choices of \mathbf{C}

This is often used for **repeated measures designs**, where each subject receives each treatment once over successive periods of time (all treatments are administered to each unit).

Example:

Let y_{ij} be the response from subject i at time j for $i = 1, \dots, n, j = 1, \dots, T$. In this case, $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})'$, $i = 1, \dots, n$ are a random sample from $N_T(\mathbf{\mu}, \mathbf{S})$

Let $n = 8$ subjects, $T = 6$. We are interested in μ_1, \dots, μ_6

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_6$$

Equivalently,

$$\mu_1 - \mu_2 = 0, \mu_2 - \mu_3 = 0, \dots, \mu_5 - \mu_6 = 0$$

We can test orthogonal polynomials for 4 equally spaced time points. To test for example the null hypothesis that quadratic and cubic effects are jointly equal to 0, we would define \mathbf{C}

$$\mathbf{C} = \begin{pmatrix} 1 & -1 & -1 & 1 \\ -1 & 3 & -3 & 1 \end{pmatrix}$$

20.0.3.2 Two-Sample Tests

Consider the analogous two sample multivariate tests.

Example: we have data on two independent random samples, one sample from each of two populations

$$\mathbf{y}_{1i} \sim N_p(\boldsymbol{\mu}_1, \mathbf{S}_1), \mathbf{y}_{2j} \sim N_p(\boldsymbol{\mu}_2, \mathbf{S}_2)$$

We **assume**

- normality
- equal variance-covariance matrices
- independent random samples

We can summarize our data using the **sufficient statistics** $\bar{\mathbf{y}}_1, \mathbf{S}_1, \bar{\mathbf{y}}_2, \mathbf{S}_2$ with respective sample sizes, n_1, n_2

Since we assume that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}$, compute a pooled estimate of the variance-covariance matrix on $n_1 + n_2 - 2$ df

$$\mathbf{S} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{(n_1 - 1) + (n_2 - 1)}$$

$$H_0 : {}_1 = {}_2 H_a : {}_1 \neq {}_2$$

At least one element of the mean vectors is different

We use

- $\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2$ to estimate $\mu_1 - \mu_2$
- \mathbf{S} to estimate

Note: because we assume the two populations are independent, there is no covariance

$$\text{cov}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) = \text{var}(\bar{\mathbf{y}}_1) + \text{var}(\bar{\mathbf{y}}_2) = \frac{1}{n_1} + \frac{1}{n_2} = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Reject H_0 if

$$\begin{aligned} T^2 &= (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \{ \mathbf{S} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \}^{-1} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) \\ &= \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \{ \mathbf{S} \}^{-1} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) \\ &\geq \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} f_{(1-\alpha, n_1 + n_2 - p - 1)} \end{aligned}$$

or equivalently, if

$$F = \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2 \geq f_{(1-\alpha, p, n_1 + n_2 - p - 1)}$$

A $100(1-\alpha)\%$ confidence region for $\mu_1 - \mu_2$ consists of all vector δ which satisfy

$$\frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2 - \delta)' \mathbf{S}^{-1} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2 - \delta) \leq \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} f_{(1-\alpha, p, n_1 + n_2 - p - 1)}$$

The simultaneous confidence intervals for all linear combinations of $\mu_1 - \mu_2$ have the form

$$\mathbf{a}' (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) \pm \sqrt{\frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} f_{(1-\alpha, p, n_1 + n_2 - p - 1)}} \times \sqrt{\mathbf{a}' \mathbf{S} \mathbf{a} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

Bonferroni intervals, for k combinations

$$(\bar{y}_{1i} - \bar{y}_{2i}) \pm t_{(1-\alpha/2k, n_1+n_2-2)} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)s_{ii}}$$

20.0.3.3 Model Assumptions

If model assumption are not met

- Unequal Covariance Matrices
 - If $n_1 = n_2$ (large samples) there is little effect on the Type I error rate and power fo the two sample test
 - If $n_1 > n_2$ and the eigenvalues of Σ_1^{-1} are less than 1, the Type I error level is inflated
 - If $n_1 > n_2$ and some eigenvalues of Σ_2^{-1} are greater than 1, the Type I error rate is too small, leading to a reduction in power
- Sample Not Normal
 - Type I error level of the two sample T^2 test isn't much affect by moderate departures from normality if the two populations being sampled have similar distributions
 - One sample T^2 test is much more sensitive to lack of normality, especially when the distribution is skewed.
 - Intuitively, you can think that in one sample your distribution will be sensitive, but the distribution of the difference between two similar distributions will not be as sensitive.
 - Solutions:
 - * Transform to make the data more normal
 - * Large large samples, use the χ^2 (Wald) test, in which populations don't need to be normal, or equal sample sizes, or equal variance-covariance matrices
 - $H_0 : \mu_1 = \mu_2 = \dots = \mu_k = H_a : \text{at least 2 are different}$

20.0.3.3.1 Equal Covariance Matrices Tests With independent random samples from k populations of p-dimensional vectors. We compute the sample covariance matrix for each, \mathbf{S}_i , where $i = 1, \dots, k$

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k = H_a : \text{at least 2 are different}$$

Assume H_0 is true, we would use a pooled estimate of the common covariance matrix,

$$\mathbf{S} = \frac{\sum_{i=1}^k (n_i - 1) \mathbf{S}_i}{\sum_{i=1}^k (n_i - 1)}$$

with $\sum_{i=1}^k (n_i - 1)$

20.0.3.3.1.1 Bartlett's Test (a modification of the likelihood ratio test). Define

$$N = \sum_{i=1}^k n_i$$

and (note: || are determinants here, not absolute value)

$$M = (N - k) \log |\mathbf{S}| - \sum_{i=1}^k (n_i - 1) \log |\mathbf{S}_i|$$

$$C^{-1} = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(k-1)} \left\{ \sum_{i=1}^k \left(\frac{1}{n_i - 1} \right) - \frac{1}{N - k} \right\}$$

- Reject H_0 when $MC^{-1} > \chi^2_{1-\alpha, (k-1)p(p+1)/2}$
- If not all samples are from normal populations, MC^{-1} has a distribution which is often shifted to the right of the nominal χ^2 distribution, which means H_0 is often rejected even when it is true (the Type I error level is inflated). Hence, it is better to test individual normality first, or then multivariate normality before you do Bartlett's test.

20.0.3.4 Two-Sample Repeated Measurements

- Define $\mathbf{y}_{hi} = (y_{hi1}, \dots, y_{hit})'$ to be the observations from the i-th subject in the h-th group for times 1 through T
- Assume that $\mathbf{y}_{11}, \dots, \mathbf{y}_{1n_1}$ are iid $N_t(\mu_1, \Sigma_1)$ and that $\mathbf{y}_{21}, \dots, \mathbf{y}_{2n_2}$ are iid $N_t(\mu_2, \Sigma_2)$
- $H_0 : \mathbf{C}(\mu_1 - \mu_2) = \mathbf{0}_c$ where \mathbf{C} is a $c \times t$ matrix of rank c where $c \leq t$
- The test statistic has the form

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \mathbf{C}' (\mathbf{C} \mathbf{S} \mathbf{C}')^{-1} \mathbf{C} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$$

where \mathbf{S} is the pooled covariance estimate. Then,

$$F = \frac{n_1 + n_2 - c - 1}{(n_1 + n_2 - 2)c} T^2 \sim f_{(c, n_1 + n_2 - c - 1)}$$

when H_0 is true

If the null hypothesis $H_0 : \mu_1 = \mu_2$ is rejected. A weaker hypothesis is that the profiles for the two groups are parallel.

$$\mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} : \mu_{1t-1} - \mu_{2t-1} = \mu_{1t} - \mu_{2t}$$

The null hypothesis matrix term is then

$H_0 : \mathbf{C}(\mu_1 - \mu_2) = \mathbf{0}_c$, where $c = t - 1$ and

$$\mathbf{C} = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -1 \end{pmatrix}_{(t-1) \times t}$$

```
# One-sample Hotelling's T^2 test
# Create data frame
plants <- data.frame(
  y1 = c(2.11, 2.36, 2.13, 2.78, 2.17),
  y2 = c(10.1, 35.0, 2.0, 6.0, 2.0),
  y3 = c(3.4, 4.1, 1.9, 3.8, 1.7)
)

# Center the data with the hypothesized means and make a matrix
plants_ctr <- plants %>%
  transmute(y1_ctr = y1 - 2.85,
            y2_ctr = y2 - 15.0,
            y3_ctr = y3 - 6.0) %>%
  as.matrix()

# Use anova.mlm to calculate Wilks' lambda
onesamp_fit <- anova(lm(plants_ctr ~ 1), test = "Wilks")
onesamp_fit # can't reject the null of hypothesized vector of means
#> Analysis of Variance Table
#>
#> Df      Wilks approx F num Df den Df Pr(>F)
```

```

#> (Intercept) 1 0.054219   11.629      3      2 0.08022 .
#> Residuals    4
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Paired-Sample Hotelling's T^2 test
library(ICSNP)

# Create data frame
waste <- data.frame(
  case = 1:11,
  com_y1 = c(6, 6, 18, 8, 11, 34, 28, 71, 43, 33, 20),
  com_y2 = c(27, 23, 64, 44, 30, 75, 26, 124, 54, 30, 14),
  state_y1 = c(25, 28, 36, 35, 15, 44, 42, 54, 34, 29, 39),
  state_y2 = c(15, 13, 22, 29, 31, 64, 30, 64, 56, 20, 21)
)

# Calculate the difference between commercial and state labs
waste_diff <- waste %>%
  transmute(y1_diff = com_y1 - state_y1,
            y2_diff = com_y2 - state_y2)
# Run the test
paired_fit <- HotellingsT2(waste_diff)
paired_fit # value T.2 in the output corresponds to the approximate F-value in the output
#>
#> Hotelling's one sample T2-test
#>
#> data: waste_diff
#> T.2 = 6.13777, df1 = 2, df2 = 9, p-value = 0.02083
#> alternative hypothesis: true location is not equal to c(0,0)
#> reject the null that the two labs' measurements are equal

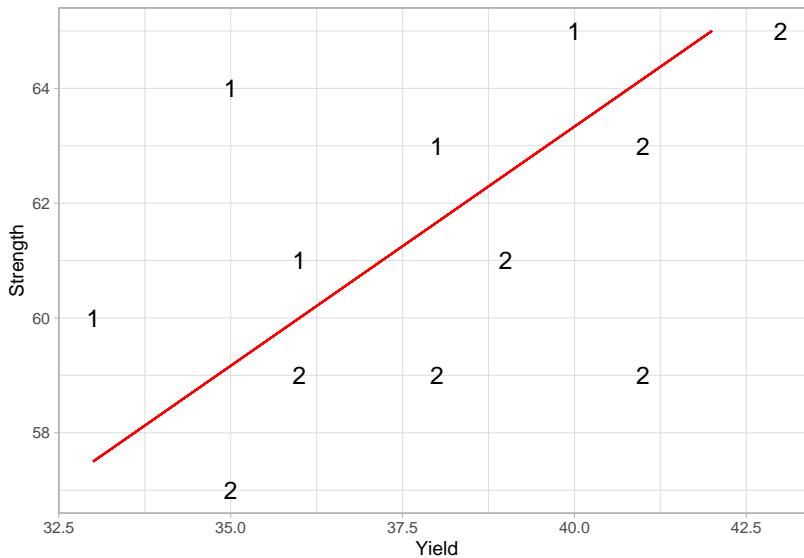
# Independent-Sample Hotelling's T^2 test with Bartlett's test

# Read in data
steel <- read.table("images/steel.dat")
names(steel) <- c("Temp", "Yield", "Strength")
str(steel)
#> 'data.frame': 12 obs. of 3 variables:
#> $ Temp : int 1 1 1 1 1 2 2 2 2 ...
#> $ Yield : int 33 36 35 38 40 35 36 38 39 41 ...
#> $ Strength: int 60 61 64 63 65 57 59 59 61 63 ...

# Plot the data
ggplot(steel, aes(x = Yield, y = Strength)) +

```

```
geom_text(aes(label = Temp), size = 5) +
  geom_segment(aes(
    x = 33,
    y = 57.5,
    xend = 42,
    yend = 65
  ), col = "red")
```



```
# Bartlett's test for equality of covariance matrices
# same thing as Box's M test in the multivariate setting
bart_test <- boxM(steel[, -1], steel$Temp)
bart_test # fail to reject the null of equal covariances
#>
#> Box's M-test for Homogeneity of Covariance Matrices
#>
#> data: steel[, -1]
#> Chi-Sq (approx.) = 0.38077, df = 3, p-value = 0.9442

# anova.mlm
twosamp_fit <-
  anova(lm(cbind(Yield, Strength) ~ factor(Temp), data = steel), test = "Wilks")
twosamp_fit
#> Analysis of Variance Table
#>
```

```

#>           Df    Wilks approx F num Df   den Df   Pr(>F)
#> (Intercept) 1 0.001177  3818.1      2      9 6.589e-14 ***
#> factor(Temp) 1 0.294883     10.8      2      9  0.004106 **
#> Residuals    10
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# ICSNP package
twosamp_fit2 <-
  HotellingsT2(cbind(steel$Yield, steel$Strength) ~ factor(steel$Temp))
twosamp_fit2
#>
#> Hotelling's two sample T2-test
#>
#> data: cbind(steel$Yield, steel$Strength) by factor(steel$Temp)
#> T.2 = 10.76, df1 = 2, df2 = 9, p-value = 0.004106
#> alternative hypothesis: true location difference is not equal to c(0,0)

# reject null. Hence, there is a difference in the means of the bivariate normal distr

```

20.1 MANOVA

Multivariate Analysis of Variance

One-way MANOVA

Compare treatment means for h different populations

Population 1: $\mathbf{y}_{11}, \mathbf{y}_{12}, \dots, \mathbf{y}_{1n_1} \sim iddN_p(1,)$

\vdots

Population h : $\mathbf{y}_{h1}, \mathbf{y}_{h2}, \dots, \mathbf{y}_{hn_h} \sim iddN_p(h,)$

Assumptions

1. Independent random samples from h different populations
2. Common covariance matrices
3. Each population is multivariate **normal**

Calculate the summary statistics $\bar{\mathbf{y}}_i$, \mathbf{S} and the pooled estimate of the covariance matrix \mathbf{S}

Similar to the univariate one-way ANOVA, we can use the effects model formulation $\mathbf{y}_i = \mu + \mathbf{u}_i + \mathbf{e}_i$, where

- μ_i is the population mean for population i

- is the overall mean effect
- τ_i is the treatment effect of the i -th treatment.

For the one-way model: $\mathbf{y}_{ij} = \mu + \tau_i + \epsilon_{ij}$ for $i = 1, \dots, h; j = 1, \dots, n_i$ and $\epsilon_{ij} \sim N_p(\mathbf{0}, \sigma^2 I)$

However, the above model is over-parameterized (i.e., infinite number of ways to define μ and the τ_i 's such that they add up to μ). Thus we can constrain by having

$$\sum_{i=1}^h n_i \tau_i = 0$$

or

$$\sum_{i=1}^h \tau_i = 0$$

The observational equivalent of the effects model is

$$\begin{aligned} \mathbf{y}_{ij} &= \bar{\mathbf{y}} + (\bar{\mathbf{y}}_i - \bar{\mathbf{y}}) + (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i) \\ &= \text{overall sample mean} + \text{treatment effect} + \text{residual (under univariate ANOVA)} \end{aligned}$$

After manipulation

$$\sum_{i=1}^h \sum_{j=1}^{n_i} (\bar{\mathbf{y}}_{ij} - \bar{\mathbf{y}})(\bar{\mathbf{y}}_{ij} - \bar{\mathbf{y}})' = \sum_{i=1}^h n_i (\bar{\mathbf{y}}_i - \bar{\mathbf{y}})(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})' + \sum_{i=1}^h \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}})(\mathbf{y}_{ij} - \bar{\mathbf{y}})'$$

LHS = Total corrected sums of squares and cross products (SSCP) matrix

RHS =

- 1st term = Treatment (or between subjects) sum of squares and cross product matrix (denoted H;B)
- 2nd term = residual (or within subject) SSCP matrix denoted (E;W)

Note:

$$\mathbf{E} = (n_1 - 1)\mathbf{S}_1 + \dots + (n_h - 1)\mathbf{S}_h = (n - h)\mathbf{S}$$

MANOVA table

Table 20.1: MONOVA table

Source	SSCP	df
Treatment	\mathbf{H}	$h - 1$
Residual (error)	\mathbf{E}	$\sum_{i=1}^h n_i - h$
Total Corrected	$\mathbf{H} + \mathbf{E}$	$\sum_{i=1}^h n_i - 1$

$$H_0 : \tau_1 = \tau_2 = \dots = \tau_h = \mathbf{0}$$

We consider the relative “sizes” of \mathbf{E} and $\mathbf{H} + \mathbf{E}$

Wilk’s Lambda

Define Wilk’s Lambda

$$\Lambda^* = \frac{|\mathbf{E}|}{|\mathbf{H} + \mathbf{E}|}$$

Properties:

1. Wilk’s Lambda is equivalent to the F-statistic in the univariate case
2. The exact distribution of Λ^* can be determined for especial cases.
3. For large sample sizes, reject H_0 if

$$-(\sum_{i=1}^h n_i - 1 - \frac{p+h}{2}) \log(\Lambda^*) > \chi^2_{(1-\alpha, p(h-1))}$$

20.1.1 Testing General Hypotheses

- h different treatments
- with the i -th treatment
- applied to n_i subjects that
- are observed for p repeated measures.

Consider this a p dimensional obs on a random sample from each of h different treatment populations.

$$\mathbf{y}_{ij} = \mu + \epsilon_i + \eta_{ij}$$

for $i = 1, \dots, h$ and $j = 1, \dots, n_i$

Equivalently,

$$\mathbf{Y} = \mathbf{XB} +$$

where $n = \sum_{i=1}^h n_i$ and with restriction $\mathbf{h} = 0$

$$\mathbf{Y}_{(n \times p)} = \begin{bmatrix} \mathbf{y}'_{11} \\ \vdots \\ \mathbf{y}'_{1n_1} \\ \vdots \\ \mathbf{y}'_{hn_h} \end{bmatrix}, \mathbf{B}_{(h \times p)} = \begin{bmatrix} ' \\ ' \\ 1 \\ \vdots \\ ' \\ h-1 \end{bmatrix}, \mathbf{\epsilon}_{(n \times p)} = \begin{bmatrix} \epsilon'_{11} \\ \vdots \\ \epsilon'_{1n_1} \\ \vdots \\ \epsilon'_{hn_h} \end{bmatrix}$$

$$\mathbf{X}_{(n \times h)} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 0 \end{bmatrix}$$

Estimation

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Rows of \mathbf{Y} are independent (i.e., $\text{var}(\mathbf{Y}) = \mathbf{I}_n \otimes \mathbf{C}$, an $np \times np$ matrix, where \otimes is the Kronecker product).

$$H_0 : \mathbf{LBM} = 0H_a : \mathbf{LBM} \neq 0$$

where

- \mathbf{L} is a $g \times h$ matrix of full row rank ($g \leq h$) = comparisons across groups
- \mathbf{M} is a $p \times u$ matrix of full column rank ($u \leq p$) = comparisons across traits

The general treatment corrected sums of squares and cross product is

$$\mathbf{H} = \mathbf{M}'\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1}\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{M}$$

or for the null hypothesis $H_0 : \mathbf{LBM} = \mathbf{D}$

$$\mathbf{H} = (\mathbf{L}\hat{\mathbf{B}}\mathbf{M} - \mathbf{D})'[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}]^{-1}(\mathbf{L}\hat{\mathbf{B}}\mathbf{M} - \mathbf{D})$$

The general matrix of residual sums of squares and cross product

$$\mathbf{E} = \mathbf{M}'\mathbf{Y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}\mathbf{M} = \mathbf{M}'[\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}'(\mathbf{X}'\mathbf{X})^{-1}\hat{\mathbf{B}}]\mathbf{M}$$

We can compute the following statistic eigenvalues of $\mathbf{H}\mathbf{E}^{-1}$

- Wilk's Criterion: $\Lambda^* = \frac{|\mathbf{E}|}{|\mathbf{H} + \mathbf{E}|}$. The df depend on the rank of $\mathbf{L}, \mathbf{M}, \mathbf{X}$
- Lawley-Hotelling Trace: $U = \text{tr}(\mathbf{H}\mathbf{E}^{-1})$
- Pillai Trace: $V = \text{tr}(\mathbf{H}(\mathbf{H} + \mathbf{E}^{-1})$
- Roy's Maximum Root: largest eigenvalue of $\mathbf{H}\mathbf{E}^{-1}$

If H_0 is true and n is large, $-(n - 1 - \frac{p+h}{2}) \ln \Lambda^* \sim \chi^2_{p(h-1)}$. Some special values of p and h can give exact F-dist under H_0

```
# One-way MANOVA

library(car)
library(emmeans)
library(profileR)
library(tidyverse)

## Read in the data
gpagmat <- read.table("images/gpagmat.dat")

## Change the variable names
names(gpagmat) <- c("y1", "y2", "admit")

## Check the structure
str(gpagmat)
#> 'data.frame':   85 obs. of  3 variables:
#> $ y1 : num  2.96 3.14 3.22 3.29 3.69 3.46 3.03 3.19 3.63 3.59 ...
#> $ y2 : int  596 473 482 527 505 693 626 663 447 588 ...
#> $ admit: int  1 1 1 1 1 1 1 1 1 1 ...

## Plot the data
gg <- ggplot(gpagmat, aes(x = y1, y = y2)) +
  geom_text(aes(label = admit, col = as.character(admit))) +
  scale_color_discrete(name = "Admission",
                        labels = c("Admit", "Do not admit", "Borderline")) +
```

```

scale_x_continuous(name = "GPA") +
scale_y_continuous(name = "GMAT")

## Fit one-way MANOVA
oneway_fit <- manova(cbind(y1, y2) ~ admit, data = gpagmat)
summary(oneway_fit, test = "Wilks")
#>      Df Wilks approx F num Df den Df    Pr(>F)
#> admit     1 0.6126   25.927      2     82 1.881e-09 ***
#> Residuals 83
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# reject the null of equal multivariate mean vectors between the three admission groups

# Repeated Measures MANOVA

## Create data frame
stress <- data.frame(
  subject = 1:8,
  begin = c(3, 2, 5, 6, 1, 5, 1, 5),
  middle = c(3, 4, 3, 7, 4, 7, 1, 2),
  final = c(6, 7, 4, 7, 6, 7, 3, 5)
)

# If independent = time with 3 levels -> univariate ANOVA (require sphericity assumption (i.e., t
# If each level of indepedenet time as a separate variable -> MANOVA (does not require sphericity

## MANOVA
stress_mod <- lm(cbind(begin, middle, final) ~ 1, data = stress)
idata <-
  data.frame(time = factor(
    c("begin", "middle", "final"),
    levels = c("begin", "middle", "final")
  ))
repeat_fit <-
  Anova(
    stress_mod,
    idata = idata,
    idesign = ~ time,
    icontrasts = "contr.poly"
  )
summary(repeat_fit) # can't reject the null hypothesis of sphericity, hence univariate ANOVA is used
#>

```

```

#> Type III Repeated Measures MANOVA Tests:
#>
#> -----
#>
#> Term: (Intercept)
#>
#> Response transformation matrix:
#>      (Intercept)
#> begin          1
#> middle         1
#> final          1
#>
#> Sum of squares and products for the hypothesis:
#>      (Intercept)
#> (Intercept)    1352
#>
#> Multivariate Tests: (Intercept)
#>           Df test stat approx F num Df den Df Pr(>F)
#> Pillai        1  0.896552 60.66667     1      7 0.00010808 ***
#> Wilks         1  0.103448 60.66667     1      7 0.00010808 ***
#> Hotelling-Lawley 1  8.666667 60.66667     1      7 0.00010808 ***
#> Roy           1  8.666667 60.66667     1      7 0.00010808 ***
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> -----
#>
#> Term: time
#>
#> Response transformation matrix:
#>      time.L   time.Q
#> begin -7.071068e-01  0.4082483
#> middle -7.850462e-17 -0.8164966
#> final  7.071068e-01  0.4082483
#>
#> Sum of squares and products for the hypothesis:
#>      time.L   time.Q
#> time.L 18.062500 6.747781
#> time.Q  6.747781 2.520833
#>
#> Multivariate Tests: time
#>           Df test stat approx F num Df den Df Pr(>F)
#> Pillai        1  0.7080717 7.276498     2      6 0.024879 *
#> Wilks         1  0.2919283 7.276498     2      6 0.024879 *
#> Hotelling-Lawley 1  2.4254992 7.276498     2      6 0.024879 *

```

```

#> Roy              1 2.4254992 7.276498      2      6 0.024879 *
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Univariate Type III Repeated-Measures ANOVA Assuming Sphericity
#>
#>           Sum Sq num Df Error SS den Df F value    Pr(>F)
#> (Intercept) 450.67     1   52.00      7 60.6667 0.0001081 ***
#> time         20.58     2   24.75     14  5.8215 0.0144578 *
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Mauchly Tests for Sphericity
#>
#>       Test statistic p-value
#> time      0.7085 0.35565
#>
#>
#> Greenhouse-Geisser and Huynh-Feldt Corrections
#> for Departure from Sphericity
#>
#>       GG eps Pr(>F[GG])
#> time 0.77429  0.02439 *
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#>       HF eps Pr(>F[HF])
#> time 0.9528433 0.01611634
# we also see linear significant time effect, but no quadratic time effect

## Polynomial contrasts
# What is the reference for the marginal means?
ref_grid(stress_mod, mult.name = "time")
#> 'emmGrid' object with variables:
#>   1 = 1
#>   time = multivariate response levels: begin, middle, final

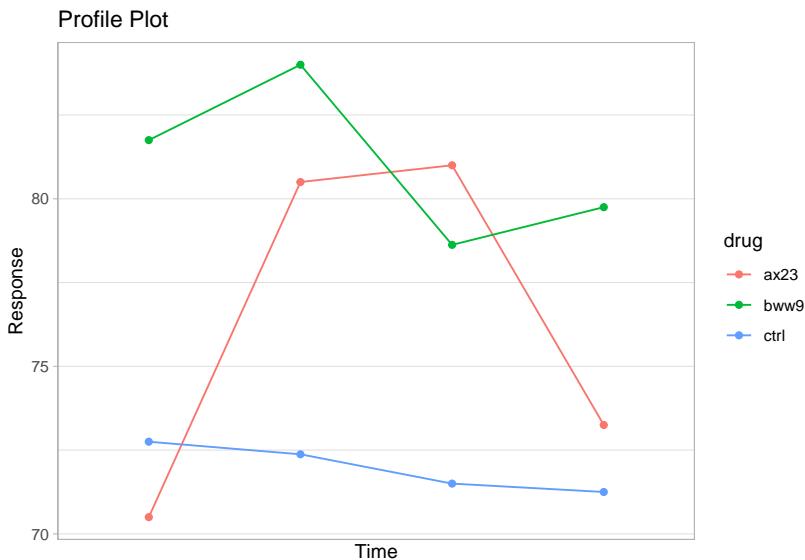
# marginal means for the levels of time
contr_means <- emmeans(stress_mod, ~ time, mult.name = "time")
contrast(contr_means, method = "poly")
#> contrast estimate   SE df t.ratio p.value
#> linear      2.12 0.766 7  2.773 0.0276
#> quadratic   1.38 0.944 7  1.457 0.1885

```

```
# MANOVA

## Read in Data
heart <- read.table("images/heart.dat")
names(heart) <- c("drug", "y1", "y2", "y3", "y4")
## Create a subject ID nested within drug
heart <- heart %>%
  group_by(drug) %>%
  mutate(subject = row_number()) %>%
  ungroup()
str(heart)
#> #> tibble [24 x 6] (S3:tbl_df/tbl/data.frame)
#> #> $ drug    : chr [1:24] "ax23" "ax23" "ax23" "ax23" ...
#> #> $ y1      : int [1:24] 72 78 71 72 66 74 62 69 85 82 ...
#> #> $ y2      : int [1:24] 86 83 82 83 79 83 73 75 86 86 ...
#> #> $ y3      : int [1:24] 81 88 81 83 77 84 78 76 83 80 ...
#> #> $ y4      : int [1:24] 77 82 75 69 66 77 70 70 80 84 ...
#> #> $ subject: int [1:24] 1 2 3 4 5 6 7 8 1 2 ...

## Create means summary for profile plot, pivot longer for plotting with ggplot
heart_means <- heart %>%
  group_by(drug) %>%
  summarize_at(vars(starts_with("y")), mean) %>%
  ungroup() %>%
  pivot_longer(-drug, names_to = "time", values_to = "mean") %>%
  mutate(time = as.numeric(as.factor(time)))
gg_profile <- ggplot(heart_means, aes(x = time, y = mean)) +
  geom_line(aes(col = drug)) +
  geom_point(aes(col = drug)) +
  ggtitle("Profile Plot") +
  scale_y_continuous(name = "Response") +
  scale_x_discrete(name = "Time")
gg_profile
```



```
## Fit model
heart_mod <- lm(cbind(y1, y2, y3, y4) ~ drug, data = heart)
man_fit <- car::Anova(heart_mod)
summary(man_fit)
#>
#> Type II MANOVA Tests:
#>
#> Sum of squares and products for error:
#>      y1      y2      y3      y4
#> y1 641.00 601.750 535.250 426.00
#> y2 601.75 823.875 615.500 534.25
#> y3 535.25 615.500 655.875 555.25
#> y4 426.00 534.250 555.250 674.50
#>
#> -----
#>
#> Term: drug
#>
#> Sum of squares and products for the hypothesis:
#>      y1      y2      y3      y4
#> y1 567.00 335.2500 42.7500 387.0
#> y2 335.25 569.0833 404.5417 367.5
#> y3 42.75 404.5417 391.0833 171.0
#> y4 387.00 367.5000 171.0000 316.0
#>
#> Multivariate Tests: drug
```

```

#>                  Df test stat  approx F num Df den Df      Pr(>F)
#> Pillai            2  1.283456 8.508082      8     38 1.5010e-06 ***
#> Wilks             2  0.079007 11.509581      8     36 6.3081e-08 ***
#> Hotelling-Lawley 2  7.069384 15.022441      8     34 3.9048e-09 ***
#> Roy               2  6.346509 30.145916      4     19 5.4493e-08 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# reject the null hypothesis of no difference in means between treatments

```

reject the null hypothesis of no difference in means between treatments

```

## Contrasts
heart$drug <- factor(heart$drug)
L <- matrix(c(0, 2,
              1, -1,-1, -1), nrow = 3, byrow = T)
colnames(L) <- c("bww9:ctrl", "ax23:rest")
rownames(L) <- unique(heart$drug)
contrasts(heart$drug) <- L
contrasts(heart$drug)
#>      bww9:ctrl ax23:rest
#> ax23          0         2
#> bww9           1        -1
#> ctrl          -1        -1

# do not set contrast L if you do further analysis (e.g., Anova, lm)
# do M matrix instead

M <- matrix(c(1, -1, 0, 0,
              0, 1, -1, 0,
              0, 0, 1, -1), nrow = 4)
## update model to test contrasts
heart_mod2 <- update(heart_mod)
coef(heart_mod2)
#>                 y1          y2          y3          y4
#> (Intercept) 75.00 78.9583333 77.041667 74.75
#> drugbww9:ctrl 4.50 5.8125000 3.562500 4.25
#> drugax23:rest -2.25 0.7708333 1.979167 -0.75

# Hypothesis test for bww9 vs control after transformation M
# same as linearHypothesis(heart_mod, hypothesis.matrix = c(0,1,-1), P = M)
bww9vctrl <-
  car::linearHypothesis(heart_mod2,
                        hypothesis.matrix = c(0, 1, 0),
                        P = M)
bww9vctrl

```

```

#>
#> Response transformation matrix:
#> [,1] [,2] [,3]
#> y1     1     0     0
#> y2    -1     1     0
#> y3     0    -1     1
#> y4     0     0    -1
#>
#> Sum of squares and products for the hypothesis:
#>          [,1]      [,2]      [,3]
#> [1,] 27.5625 -47.25 14.4375
#> [2,] -47.2500  81.00 -24.7500
#> [3,] 14.4375 -24.75   7.5625
#>
#> Sum of squares and products for error:
#>          [,1]      [,2]      [,3]
#> [1,] 261.375 -141.875 28.000
#> [2,] -141.875  248.750 -19.375
#> [3,] 28.000  -19.375 219.875
#>
#> Multivariate Tests:
#>           Df test stat approx F num Df den Df Pr(>F)
#> Pillai        1 0.2564306 2.184141      3     19 0.1233
#> Wilks         1 0.7435694 2.184141      3     19 0.1233
#> Hotelling-Lawley 1 0.3448644 2.184141      3     19 0.1233
#> Roy           1 0.3448644 2.184141      3     19 0.1233

bww9vctrl <-
  car::linearHypothesis(heart_mod,
                        hypothesis.matrix = c(0, 1, -1),
                        P = M)

bww9vctrl
#>
#> Response transformation matrix:
#> [,1] [,2] [,3]
#> y1     1     0     0
#> y2    -1     1     0
#> y3     0    -1     1
#> y4     0     0    -1
#>
#> Sum of squares and products for the hypothesis:
#>          [,1]      [,2]      [,3]
#> [1,] 27.5625 -47.25 14.4375
#> [2,] -47.2500  81.00 -24.7500
#> [3,] 14.4375 -24.75   7.5625

```

```

#>
#> Sum of squares and products for error:
#>      [,1]      [,2]      [,3]
#> [1,] 261.375 -141.875 28.000
#> [2,] -141.875 248.750 -19.375
#> [3,] 28.000  -19.375 219.875
#>
#> Multivariate Tests:
#>          Df test stat approx F num Df den Df Pr(>F)
#> Pillai           1 0.2564306 2.184141   3     19 0.1233
#> Wilks            1 0.7435694 2.184141   3     19 0.1233
#> Hotelling-Lawley 1 0.3448644 2.184141   3     19 0.1233
#> Roy              1 0.3448644 2.184141   3     19 0.1233

```

there is no significant difference in means between the control and bww9 drug

```

# Hypothesis test for ax23 vs rest after transformation M
axx23vrest <-
  car::linearHypothesis(heart_mod2,
                        hypothesis.matrix = c(0, 0, 1),
                        P = M)
axx23vrest
#>
#> Response transformation matrix:
#>      [,1] [,2] [,3]
#> y1     1    0    0
#> y2    -1    1    0
#> y3     0   -1    1
#> y4     0    0   -1
#>
#> Sum of squares and products for the hypothesis:
#>      [,1]      [,2]      [,3]
#> [1,] 438.0208 175.20833 -395.7292
#> [2,] 175.2083  70.08333 -158.2917
#> [3,] -395.7292 -158.29167 357.5208
#>
#> Sum of squares and products for error:
#>      [,1]      [,2]      [,3]
#> [1,] 261.375 -141.875 28.000
#> [2,] -141.875 248.750 -19.375
#> [3,] 28.000  -19.375 219.875
#>
#> Multivariate Tests:
#>          Df test stat approx F num Df den Df Pr(>F)
#> Pillai           1 0.855364 37.45483   3     19 3.5484e-08 ***

```

```

#> Wilks              1  0.144636 37.45483      3    19 3.5484e-08 ***
#> Hotelling-Lawley  1  5.913921 37.45483      3    19 3.5484e-08 ***
#> Roy                1  5.913921 37.45483      3    19 3.5484e-08 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

axx23vrest <-
  car::linearHypothesis(heart_mod,
                        hypothesis.matrix = c(2, -1, 1),
                        P = M)

axx23vrest
#>
#> Response transformation matrix:
#>   [,1] [,2] [,3]
#> y1     1     0     0
#> y2    -1     1     0
#> y3     0    -1     1
#> y4     0     0    -1
#>
#> Sum of squares and products for the hypothesis:
#>   [,1]      [,2]      [,3]
#> [1,] 402.5208 127.41667 -390.9375
#> [2,] 127.4167  40.33333 -123.7500
#> [3,] -390.9375 -123.75000 379.6875
#>
#> Sum of squares and products for error:
#>   [,1]      [,2]      [,3]
#> [1,] 261.375 -141.875  28.000
#> [2,] -141.875  248.750 -19.375
#> [3,]  28.000 -19.375 219.875
#>
#> Multivariate Tests:
#>           Df test stat approx F num Df den Df Pr(>F)
#> Pillai          1  0.842450 33.86563      3    19 7.9422e-08 ***
#> Wilks            1  0.157550 33.86563      3    19 7.9422e-08 ***
#> Hotelling-Lawley 1  5.347205 33.86563      3    19 7.9422e-08 ***
#> Roy              1  5.347205 33.86563      3    19 7.9422e-08 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

there is a significant difference in means between ax23 drug treatment and the rest of the treatments

20.1.2 Profile Analysis

Examine similarities between the treatment effects (between subjects), which is useful for longitudinal analysis. Null is that all treatments have the same average effect.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_h$$

Equivalently,

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_h$$

The exact nature of the similarities and differences between the treatments can be examined under this analysis.

Sequential steps in profile analysis:

1. Are the profiles **parallel**? (i.e., is there no interaction between treatment and time)
2. Are the profiles **coincidental**? (i.e., are the profiles identical?)
3. Are the profiles **horizontal**? (i.e., are there no differences between any time points?)

If we reject the null hypothesis that the profiles are parallel, we can test

- Are there differences among groups within some subset of the total time points?
- Are there differences among time points in a particular group (or groups)?
- Are there differences within some subset of the total time points in a particular group (or groups)?

Example

- 4 times ($p = 4$)
- 3 treatments ($h=3$)

20.1.2.1 Parallel Profile

Are the profiles for each population identical except for a mean shift?

$$H_0 : \mu_{11} - \mu_{21} - \mu_{12} - \mu_{22} = \cdots = \mu_{1t} - \mu_{2t} \mu_{11} - \mu_{31} - \mu_{12} - \mu_{32} = \cdots = \mu_{1t} - \mu_{3t} \cdots$$

for $h - 1$ equations

Equivalently,

$$H_0 : \mathbf{LBM} = \mathbf{0}$$

$$\mathbf{LBM} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \mu_{11} & \dots & \mu_{14} \\ \mu_{21} & \dots & \mu_{24} \\ \mu_{31} & \dots & \mu_{34} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} = \mathbf{0}$$

where this is the cell means parameterization of \mathbf{B}

The multiplication of the first 2 matrices \mathbf{LB} is

$$\begin{bmatrix} \mu_{11} - \mu_{21} & \mu_{12} - \mu_{22} & \mu_{13} - \mu_{23} & \mu_{14} - \mu_{24} \\ \mu_{11} - \mu_{31} & \mu_{12} - \mu_{32} & \mu_{13} - \mu_{33} & \mu_{14} - \mu_{34} \end{bmatrix}$$

which is the differences in treatment means at the same time

Multiplying by \mathbf{M} , we get the comparison across time

$$\begin{bmatrix} (\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22}) & (\mu_{11} - \mu_{21}) - (\mu_{13} - \mu_{23}) & (\mu_{11} - \mu_{21}) - (\mu_{14} - \mu_{24}) \\ (\mu_{11} - \mu_{31}) - (\mu_{12} - \mu_{32}) & (\mu_{11} - \mu_{31}) - (\mu_{13} - \mu_{33}) & (\mu_{11} - \mu_{31}) - (\mu_{14} - \mu_{34}) \end{bmatrix}$$

Alternatively, we can also use the effects parameterization

$$\mathbf{LBM} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \mu' \\ \tau'_1 \\ \tau'_2 \\ \tau'_3 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} = \mathbf{0}$$

In both parameterizations, $\text{rank}(\mathbf{L}) = h - 1$ and $\text{rank}(\mathbf{M}) = p - 1$

We could also choose \mathbf{L} and \mathbf{M} in other forms

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

and

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}$$

and still obtain the same result.

20.1.2.2 Coincidental Profiles

After we have evidence that the profiles are parallel (i.e., fail to reject the parallel profile test), we can ask whether they are identical?

Given profiles are **parallel**, then if the sums of the components of μ_i are identical for all the treatments, then the profiles are **identical**.

$$H_0 : \mathbf{1}'_p \mu_1 = \mathbf{1}'_p \mu_2 = \cdots = \mathbf{1}'_p \mu_h$$

Equivalently,

$$H_0 : \mathbf{LBM} = \mathbf{0}$$

where for the cell means parameterization

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}$$

and

$$\mathbf{M} = [1 \ 1 \ 1 \ 1]'$$

multiplication yields

$$\begin{bmatrix} (\mu_{11} + \mu_{12} + \mu_{13} + \mu_{14}) - (\mu_{31} + \mu_{32} + \mu_{33} + \mu_{34}) \\ (\mu_{21} + \mu_{22} + \mu_{23} + \mu_{24}) - (\mu_{31} + \mu_{32} + \mu_{33} + \mu_{34}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Different choices of \mathbf{L} and \mathbf{M} can yield the same result

20.1.2.3 Horizontal Profiles

Given that we can't reject the null hypothesis that all h profiles are the same, we can ask whether all of the elements of the common profile equal? (i.e., horizontal)

$$H_0 : \mathbf{LBM} = \mathbf{0}$$

$$\mathbf{L} = [\begin{array}{ccc} 1 & 0 & 0 \end{array}]$$

and

$$\mathbf{M} = \left[\begin{array}{ccc} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{array} \right]$$

hence,

$$[(\mu_{11} - \mu_{12}) \quad (\mu_{12} - \mu_{13}) \quad (\mu_{13} + \mu_{14})] = [\begin{array}{ccc} 0 & 0 & 0 \end{array}]$$

Note:

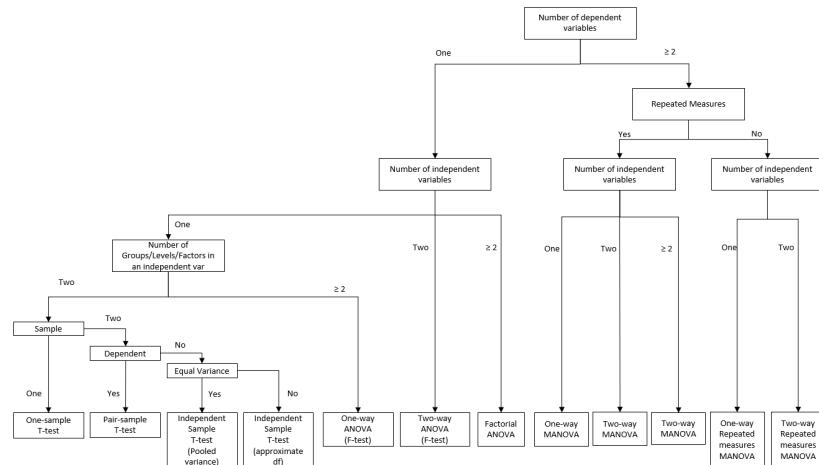
- If we fail to reject all 3 hypotheses, then we fail to reject the null hypotheses of both no difference between treatments and no differences between traits.

Test	Equivalent test for
Parallel profile	Interaction
Coincidental profile	main effect of between-subjects factor
Horizontal profile	main effect of repeated measures factor

```
profile_fit <-
pbg(
  data = as.matrix(heart[, 2:5]),
  group = as.matrix(heart[, 1]),
  original.names = TRUE,
  profile.plot = FALSE
)
summary(profile_fit)
#> Call:
#> pbga(data = as.matrix(heart[, 2:5]), group = as.matrix(heart[,
```

```
#>      1]), original.names = TRUE, profile.plot = FALSE)
#>
#> Hypothesis Tests:
#> $`Ho: Profiles are parallel`
#>   Multivariate.Test Statistic Approx.F num.df den.df      p.value
#> 1           Wilks 0.1102861 12.737599      6     38 7.891497e-08
#> 2           Pillai 1.0891707  7.972007      6     40 1.092397e-05
#> 3 Hotelling-Lawley 6.2587852 18.776356      6     36 9.258571e-10
#> 4          Roy 5.9550887 39.700592      3     20 1.302458e-08
#>
#> $`Ho: Profiles have equal levels`
#>   Df Sum Sq Mean Sq F value Pr(>F)
#> group      2 328.7 164.35  5.918 0.00915 **
#> Residuals  21 583.2  27.77
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> $`Ho: Profiles are flat`
#>   F df1 df2      p-value
#> 1 14.30928   3 19 4.096803e-05
#> # reject null hypothesis of parallel profiles
#> # reject the null hypothesis of coincidental profiles
#> # reject the null hypothesis that the profiles are flat
```

20.1.3 Summary



20.2 Principal Components

- Unsupervised learning
- find important features
- reduce the dimensions of the data set
- “decorrelate” multivariate vectors that have dependence.
- uses eigenvector/eigenvalue decomposition of covariance (correlation) matrices.

According to the “spectral decomposition theorem”, if $p \times p$ is a positive semi-definite, symmetric, real matrix, then there exists an orthogonal matrix \mathbf{A} such that $\mathbf{A}' \mathbf{A} = \Lambda$ where Λ is a diagonal matrix containing the eigenvalues

$$= \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{pmatrix}$$

$$\mathbf{A} = (\mathbf{a}_1 \quad \mathbf{a}_2 \quad \dots \quad \mathbf{a}_p)$$

the i -th column of \mathbf{A} , \mathbf{a}_i , is the i -th $p \times 1$ eigenvector of Σ that corresponds to the eigenvalue, λ_i , where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$. Alternatively, express in matrix decomposition:

$$= \mathbf{A} \mathbf{A}'$$

$$= \mathbf{A} \left(\begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{pmatrix} \mathbf{A}' \right) = \sum_{i=1}^p \lambda_i \mathbf{a}_i \mathbf{a}_i'$$

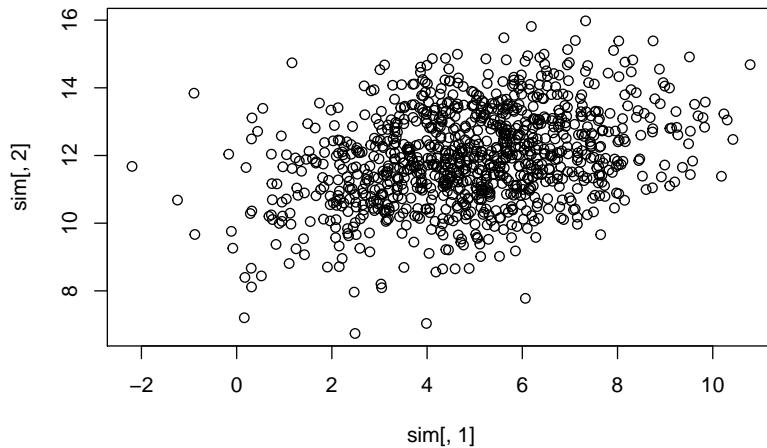
where the outer product $\mathbf{a}_i \mathbf{a}_i'$ is a $p \times p$ matrix of rank 1.

For example,

$$\mathbf{x} \sim N_2(\mu, \Sigma)$$

$$= \begin{pmatrix} 5 \\ 12 \end{pmatrix}; \quad = \begin{pmatrix} 4 & 1 \\ 1 & 2 \end{pmatrix}$$

```
library(MASS)
mu = as.matrix(c(5,12))
Sigma = matrix(c(4,1,1,2), nrow = 2, byrow = T)
sim <- mvrnorm(n = 1000, mu = mu, Sigma = Sigma)
plot(sim[,1], sim[,2])
```



Here,

$$\mathbf{A} = \begin{pmatrix} 0.9239 & -0.3827 \\ 0.3827 & 0.9239 \end{pmatrix}$$

Columns of \mathbf{A} are the eigenvectors for the decomposition

Under matrix multiplication ($\mathbf{A}' \mathbf{A}$ or $\mathbf{A}' \mathbf{A}$), the off-diagonal elements equal to 0

Multiplying data by this matrix (i.e., projecting the data onto the orthogonal axes); the distribution of the resulting data (i.e., “scores”) is

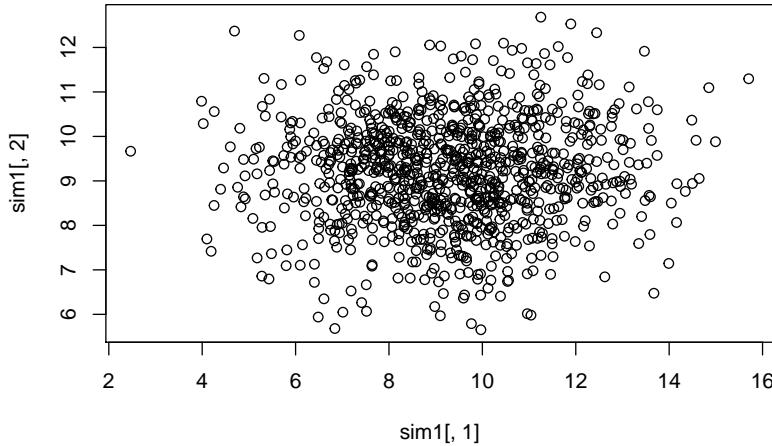
$$N_2(\mathbf{A}', \mathbf{A}' \mathbf{A}) = N_2(\mathbf{A}',)$$

Equivalently,

$$\mathbf{y} = \mathbf{A}' \mathbf{x} \sim N \left[\begin{pmatrix} 9.2119 \\ 9.1733 \end{pmatrix}, \begin{pmatrix} 4.4144 & 0 \\ 0 & 1.5859 \end{pmatrix} \right]$$

```
A_matrix = matrix(c(0.9239,-0.3827,0.3827,0.9239),nrow = 2, byrow = T)
t(A_matrix) %*% A_matrix
#>      [,1]     [,2]
#> [1,] 1.000051 0.000000
#> [2,] 0.000000 1.000051
```

```
sim1 <- mvrnorm(n = 1000, mu = t(A_matrix) %*% mu, Sigma = t(A_matrix) %*% Sigma %*% A_matrix)
plot(sim1[,1],sim1[,2])
```



No more dependence in the data structure, plot

Notes:

- The i -th eigenvalue is the variance of a linear combination of the elements of \mathbf{x} ; $\text{var}(y_i) = \text{var}(\mathbf{a}'_i \mathbf{x}) = \lambda_i$
- The values on the transformed set of axes (i.e., the y_i 's) are called the scores. These are the orthogonal projections of the data onto the “new principal component axes”
- Variances of y_1 are greater than those for any other possible projection

Covariance matrix decomposition and projection onto orthogonal axes = PCA

20.2.1 Population Principal Components

$p \times 1$ vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ which are iid with $\text{var}(\mathbf{x}_i) =$

- The first PC is the linear combination $y_1 = \mathbf{a}'_1 \mathbf{x} = a_{11}x_1 + \dots + a_{1p}x_p$ with $\mathbf{a}'_1 \mathbf{a}_1 = 1$ such that $\text{var}(y_1)$ is the maximum of all linear combinations of \mathbf{x} which have unit length

- The second PC is the linear combination $y_1 = \mathbf{a}'_2 \mathbf{x} = a_{21}x_1 + \dots + a_{2p}x_p$ with $\mathbf{a}'_2 \mathbf{a}_2 = 1$ such that $\text{var}(y_1)$ is the maximum of all linear combinations of \mathbf{x} which have unit length and uncorrelated with y_1 (i.e., $\text{cov}(\mathbf{a}'_1 \mathbf{x}, \mathbf{a}'_2 \mathbf{x}) = 0$)
- continues for all y_i to y_p

\mathbf{a}_i 's are those that make up the matrix \mathbf{A} in the symmetric decomposition $\mathbf{A}' \mathbf{A} = \Sigma$, where $\text{var}(y_1) = \lambda_1, \dots, \text{var}(y_p) = \lambda_p$. And the total variance of \mathbf{x} is

$$\begin{aligned}\text{var}(x_1) + \dots + \text{var}(x_p) &= \text{tr}(\Sigma) = \lambda_1 + \dots + \lambda_p \\ &= \text{var}(y_1) + \dots + \text{var}(y_p)\end{aligned}$$

Data Reduction

To reduce the dimension of data from p (original) to k dimensions without much “loss of information”, we can use properties of the population principal components

- Suppose $\Sigma \approx \sum_{i=1}^k \lambda_i \mathbf{a}_i \mathbf{a}'_i$. Even though the true variance-covariance matrix has rank p , it can be well approximated by a matrix of rank k ($k < p$)
- New “traits” are linear combinations of the measured traits. We can attempt to make meaningful interpretation of the combinations (with orthogonality constraints).
- The proportion of the total variance accounted for by the j -th principal component is

$$\frac{\text{var}(y_j)}{\sum_{i=1}^p \text{var}(y_i)} = \frac{\lambda_j}{\sum_{i=1}^p \lambda_i}$$

- The proportion of the total variation accounted for by the first k principal components is $\frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^p \lambda_i}$
- Above example, we have $4.4144/(4+2) = .735$ of the total variability can be explained by the first principal component

20.2.2 Sample Principal Components

Since Σ is unknown, we use

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$$

Let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_p \geq 0$ be the eigenvalues of \mathbf{S} and $\hat{\mathbf{a}}_1, \hat{\mathbf{a}}_2, \dots, \hat{\mathbf{a}}_p$ denote the eigenvectors of \mathbf{S}

Then, the i -th sample principal component score (or principal component or score) is

$$\hat{y}_{ij} = \sum_{k=1}^p \hat{a}_{ik} x_{kj} = \hat{\mathbf{a}}'_i \mathbf{x}_j$$

Properties of Sample Principal Components

- The estimated variance of $y_i = \hat{\mathbf{a}}'_i \mathbf{x}_j$ is $\hat{\lambda}_i$
- The sample covariance between \hat{y}_i and $\hat{y}_{i'}$ is 0 when $i \neq i'$
- The proportion of the total sample variance accounted for by the i -th sample principal component is $\frac{\hat{\lambda}_i}{\sum_{k=1}^p \hat{\lambda}_k}$
- The estimated correlation between the i -th principal component score and the l -th attribute of \mathbf{x} is

$$r_{x_l, \hat{y}_i} = \frac{\hat{a}_{il} \sqrt{\lambda_i}}{\sqrt{s_{ll}}}$$

- The correlation coefficient is typically used to interpret the components (i.e., if this correlation is high then it suggests that the l -th original trait is important in the i -th principle component). According to

@Johnson_1988, pp.433 – 434

, r_{x_l, \hat{y}_i} only measures the univariate contribution of an individual X to a component Y without taking into account the presence of the other X 's. Hence, some prefer \hat{a}_{il} coefficient to interpret the principal component.

- $r_{x_l, \hat{y}_i}; \hat{a}_{il}$ are referred to as “loadings”

To use k principal components, we must calculate the scores for each data vector in the sample

$$\mathbf{y}_j = \begin{pmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{kj} \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{a}}'_1 \mathbf{x}_j \\ \hat{\mathbf{a}}'_2 \mathbf{x}_j \\ \vdots \\ \hat{\mathbf{a}}'_k \mathbf{x}_j \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{a}}'_1 \\ \hat{\mathbf{a}}'_2 \\ \vdots \\ \hat{\mathbf{a}}'_k \end{pmatrix} \mathbf{x}_j$$

Issues:

- Large sample theory exists for eigenvalues and eigenvectors of sample covariance matrices if inference is necessary. But we do not do inference with PCA, we only use it as exploratory or descriptive analysis.
- PC is not invariant to changes in scale (Exception: if all trait are rescaled by multiplying by the same constant, such as feet to inches).
 - PCA based on the correlation matrix \mathbf{R} is different than that based on the covariance matrix
 - PCA for the correlation matrix is just rescaling each trait to have unit variance
 - Transform \mathbf{x} to \mathbf{z} where $z_{ij} = (x_{ij} - \bar{x}_i)/\sqrt{s_{ii}}$ where the denominator affects the PCA
 - After transformation, $cov(\mathbf{z}) = \mathbf{R}$
 - PCA on \mathbf{R} is calculated in the same way as that on \mathbf{S} (where $\hat{\lambda}_1 + \dots + \hat{\lambda}_p = p$)
 - The use of \mathbf{R}, \mathbf{S} depends on the purpose of PCA.
 - * If the scale of the observations if different, covariance matrix is more preferable. but if they are dramatically different, analysis can still be dominated by the large variance traits.
 - How many PCs to use can be guided by
 - * Scree Graphs: plot the eigenvalues against their indices. Look for the “elbow” where the steep decline in the graph suddenly flattens out; or big gaps.
 - * minimum Percent of total variation (e.g., choose enough components to have 50% or 90%). can be used for interpretations.
 - * Kaiser’s rule: use only those PC with eigenvalues larger than 1 (applied to PCA on the correlation matrix) - ad hoc
 - * Compare to the eigenvalue scree plot of data to the scree plot when the data are randomized.

20.2.3 Application

PCA on the covariance matrix is usually not preferred due to the fact that PCA is not invariant to changes in scale. Hence, PCA on the correlation matrix is more preferred

This also addresses the problem of multicollinearity

The eigenvectors may differ by a multiplication of -1 for different implementation, but same interpretation.

```

library(tidyverse)
## Read in and check data
stock <- read.table("images/stock.dat")
names(stock) <- c("allied", "dupont", "carbide", "exxon", "texaco")
str(stock)
#> 'data.frame':   100 obs. of  5 variables:
#> $ allied : num  0 0.027 0.1228 0.057 0.0637 ...
#> $ dupont : num  0 -0.04485 0.06077 0.02995 -0.00379 ...
#> $ carbide: num  0 -0.00303 0.08815 0.06681 -0.03979 ...
#> $ exxon  : num  0.0395 -0.0145 0.0862 0.0135 -0.0186 ...
#> $ texaco : num  0 0.0435 0.0781 0.0195 -0.0242 ...

## Covariance matrix of data
cov(stock)
#>          allied      dupont      carbide      exxon      texaco
#> allied  0.0016299269 0.0008166676 0.0008100713 0.0004422405 0.0005139715
#> dupont  0.0008166676 0.0012293759 0.0008276330 0.0003868550 0.0003109431
#> carbide 0.0008100713 0.0008276330 0.0015560763 0.0004872816 0.0004624767
#> exxon   0.0004422405 0.0003868550 0.0004872816 0.0008023323 0.0004084734
#> texaco  0.0005139715 0.0003109431 0.0004624767 0.0004084734 0.0007587370

## Correlation matrix of data
cor(stock)
#>          allied      dupont      carbide      exxon      texaco
#> allied  1.0000000 0.5769244 0.5086555 0.3867206 0.4621781
#> dupont  0.5769244 1.0000000 0.5983841 0.3895191 0.3219534
#> carbide 0.5086555 0.5983841 1.0000000 0.4361014 0.4256266
#> exxon   0.3867206 0.3895191 0.4361014 1.0000000 0.5235293
#> texaco  0.4621781 0.3219534 0.4256266 0.5235293 1.0000000

# cov(scale(stock)) # give the same result

## PCA with covariance
cov_pca <- prcomp(stock) # uses singular value decomposition for calculation and an N - 1 divisor
# alternatively, princomp can do PCA via spectral decomposition, but it has worse numerical accuracy

# eigen values
cov_results <- data.frame(eigen_values = cov_pca$sdev ^ 2)
cov_results %>%
  mutate(proportion = eigen_values / sum(eigen_values),
         cumulative = cumsum(proportion)) # first 2 PCs account for 73% variance in the data
#>   eigen_values proportion cumulative
#> 1 0.0035953867 0.60159252  0.6015925
#> 2 0.0007921798 0.13255027  0.7341428
#> 3 0.0007364426 0.12322412  0.8573669

```

```

#> 4 0.0005086686 0.08511218 0.9424791
#> 5 0.0003437707 0.05752091 1.0000000

# eigen vectors
cov_pca$rotation # prcomp calls rotation
#>          PC1         PC2         PC3         PC4         PC5
#> allied  0.5605914  0.73884565 -0.1260222  0.28373183 -0.20846832
#> dupont  0.4698673 -0.09286987 -0.4675066 -0.68793190  0.28069055
#> carbide 0.5473322 -0.65401929 -0.1140581  0.50045312 -0.09603973
#> exxon   0.2908932 -0.11267353  0.6099196 -0.43808002 -0.58203935
#> texaco  0.2842017  0.07103332  0.6168831  0.06227778  0.72784638
# princomp calls loadings.

# first PC = overall average
# second PC compares Allied to Carbide

## PCA with correlation
#same as scale(stock) %>% prcomp
cor_pca <- prcomp(stock, scale = T)

# eigen values
cor_results <- data.frame(eigen_values = cor_pca$sdev ^ 2)
cor_results %>%
  mutate(proportion = eigen_values / sum(eigen_values),
         cumulative = cumsum(proportion))
#>   eigen_values proportion cumulative
#> 1    2.8564869  0.57129738  0.5712974
#> 2    0.8091185  0.16182370  0.7331211
#> 3    0.5400440  0.10800880  0.8411299
#> 4    0.4513468  0.09026936  0.9313992
#> 5    0.3430038  0.06860076  1.0000000

# first eigen values corresponds to less variance than PCA based on the covariance matrix

# eigen vectors
cor_pca$rotation
#>          PC1         PC2         PC3         PC4         PC5
#> allied  0.4635405 -0.2408499  0.6133570 -0.3813727  0.4532876
#> dupont  0.4570764 -0.5090997 -0.1778996 -0.2113068 -0.6749814
#> carbide 0.4699804 -0.2605774 -0.3370355  0.6640985  0.3957247
#> exxon   0.4216770  0.5252647 -0.5390181 -0.4728036  0.1794482
#> texaco  0.4213291  0.5822416  0.4336029  0.3812273 -0.3874672
# interpretation of PC2 is different from above: it is a comparison of Allied, Dupont, Carbide, Exxon, Texaco

```

Covid Example

To reduce collinearity problem in this dataset, we can use principal components as regressors.

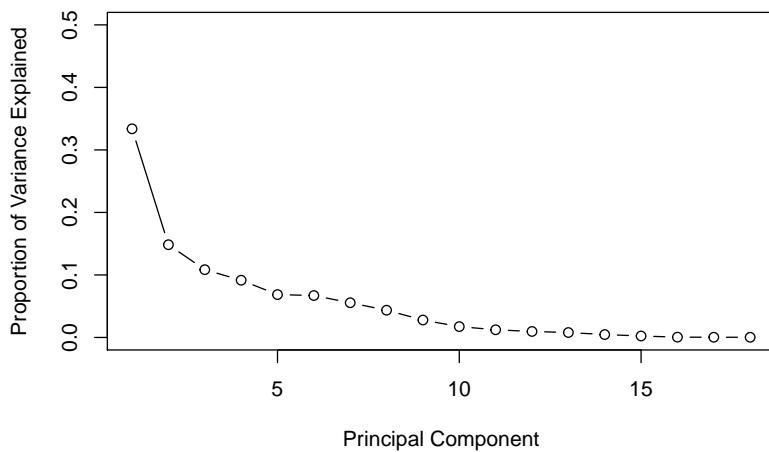
```

load('images/M0covid.RData')
covidpca <- prcomp(ndat[,-1],scale = T,center = T)

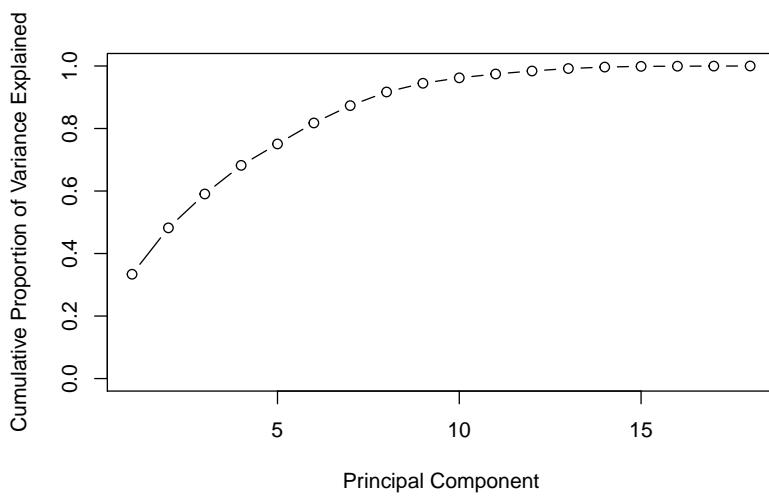
covidpca$rotation[,1:2]
#> 
#> X..Population.in.Rural.Areas          PC1      PC2
#> 0.32865838  0.05090955
#> Area..sq..miles.          0.12014444 -0.28579183
#> Population.density..sq..miles.       -0.29670124  0.28312922
#> Literacy.rate          -0.12517700 -0.08999542
#> Families          -0.25856941  0.16485752
#> Area.of.farm.land..sq..miles.        0.02101106 -0.31070363
#> Number.of.farms          -0.03814582 -0.44809679
#> Average.value.of.all.property.per.farm..dollars. -0.05410709  0.14404306
#> Estimation.of.rurality..          -0.19040210  0.12089501
#> Male..          0.02182394 -0.09568768
#> Number.of.Physicians.per.100.000     -0.31451606  0.13598026
#> average.age          0.29414708  0.35593459
#> X0.4.age.proportion        -0.11431336 -0.23574057
#> X20.44.age.proportion      -0.32802128 -0.22718550
#> X65.and.over.age.proportion  0.30585033  0.32201626
#> prop..White..nonHisp        0.35627561 -0.14142646
#> prop..Hispanic          -0.16655381 -0.15105342
#> prop..Black          -0.33333359  0.24405802

# Variability of each principal component: pr.var
pr.var <- covidpca$sdev ^ 2
# Variance explained by each principal component: pve
pve <- pr.var / sum(pr.var)
plot(
  pve,
  xlab = "Principal Component",
  ylab = "Proportion of Variance Explained",
  ylim = c(0, 0.5),
  type = "b"
)

```



```
plot(  
  cumsum(pve),  
  xlab = "Principal Component",  
  ylab = "Cumulative Proportion of Variance Explained",  
  ylim = c(0, 1),  
  type = "b"  
)
```



```
# the first six principle account for around 80% of the variance.

#using base lm function for PC regression
pcadat <- data.frame(covidpca$x[, 1:6])
pcadat$y <- ndat$Y
pcr.man <- lm(log(y) ~ ., pcadat)
mean(pcr.man$residuals ^ 2)
#> [1] 0.03453371

#comparison to lm w/o prin comps
lm.fit <- lm(log(Y) ~ ., data = ndat)
mean(lm.fit$residuals ^ 2)
#> [1] 0.02335128
```

MSE for the PC-based model is larger than regular regression, because models with a large degree of collinearity can still perform well.

`pcr` function in `pls` can be used for fitting PC regression (it will select the optimal number of components in the model).

20.3 Factor Analysis

Purpose

- Using a few linear combinations of underlying unobservable (latent) traits, we try to describe the covariance relationship among a large number of measured traits
- Similar to PCA, but factor analysis is **model based**

More details can be found on PSU stat or UMN stat

Let \mathbf{y} be the set of p measured variables

$$E(\mathbf{y}) =$$

$$\text{var}(\mathbf{y}) =$$

We have

$$\begin{aligned} \mathbf{y} - \bar{\mathbf{y}} &= \mathbf{Lf} + \boldsymbol{\epsilon} \\ &= \begin{pmatrix} l_{11}f_1 + l_{12}f_2 + \cdots + l_{1m}f_m \\ \vdots \\ l_{p1}f_1 + l_{p2}f_2 + \cdots + l_{pm}f_m \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_p \end{pmatrix} \end{aligned}$$

where

- $\mathbf{y} - \bar{\mathbf{y}}$ = the p centered measurements
- $\mathbf{L} = p \times m$ matrix of factor loadings
- \mathbf{f} = unobserved common factors for the population
- ϵ = random errors (i.e., variation that is not accounted for by the common factors).

We want m (the number of factors) to be much smaller than p (the number of measured attributes)

Restrictions on the model

- $E(\epsilon) = \mathbf{0}$
- $var(\epsilon) = \Psi_{p \times p} = diag(\psi_1, \dots, \psi_p)$
- ϵ, \mathbf{f} are independent
- Additional assumption could be $E(\mathbf{f}) = \mathbf{0}, var(\mathbf{f}) = \mathbf{I}_{m \times m}$ (known as the orthogonal factor model), which imposes the following covariance structure on \mathbf{y}

$$\begin{aligned} var(\mathbf{y}) &= var(\mathbf{Lf} + \epsilon) \\ &= var(\mathbf{Lf}) + var(\epsilon) \\ &= \mathbf{L} var(\mathbf{f}) \mathbf{L}' + \\ &= \mathbf{L} \mathbf{I} \mathbf{L}' + \\ &= \mathbf{L} \mathbf{L}' + \end{aligned}$$

Since ϵ is diagonal, the off-diagonal elements of $\mathbf{L} \mathbf{L}'$ are σ_{ij} , the covariances in ϵ , which means $cov(y_i, y_j) = \sum_{k=1}^m l_{ik} l_{jk}$ and the covariance of \mathbf{y} is completely determined by the m factors ($m \ll p$)

$var(y_i) = \sum_{k=1}^m l_{ik}^2 + \psi_i$ where ψ_i is the **specific variance** and the summation term is the i -th **communality** (i.e., portion of the variance of the i -th variable contributed by the m common factors ($h_i^2 = \sum_{k=1}^m l_{ik}^2$))

The factor model is only uniquely determined up to an orthogonal transformation of the factors.

Let $\mathbf{T}_{m \times m}$ be an orthogonal matrix $\mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$ then

$$\begin{aligned}
 \mathbf{y} - &= \mathbf{Lf} + \epsilon \\
 &= \mathbf{LTT}'\mathbf{f} + \epsilon \\
 &= \mathbf{L}^*(\mathbf{T}'\mathbf{f}) + \epsilon \quad \text{where } \mathbf{L}^* = \mathbf{LT}
 \end{aligned}$$

and

$$\begin{aligned}
 &= \mathbf{LL}' + \\
 &= \mathbf{LTT}'\mathbf{L} + \\
 &= (\mathbf{L}^*)(\mathbf{L}^*)' +
 \end{aligned}$$

Hence, any orthogonal transformation of the factors is an equally good description of the correlations among the observed traits.

Let $\mathbf{y} = \mathbf{Cx}$, where \mathbf{C} is any diagonal matrix, then $\mathbf{L}_y = \mathbf{CL}_x$ and $\mathbf{y} = \mathbf{C}_x\mathbf{C}$

Hence, we can see that factor analysis is also invariant to changes in scale

20.3.1 Methods of Estimation

To estimate \mathbf{L}

1. Principal Component Method
2. Principal Factor Method
3. 20.3.1.3

20.3.1.1 Principal Component Method

Spectral decomposition

$$\begin{aligned}
 &= \lambda_1 \mathbf{a}_1 \mathbf{a}'_1 + \cdots + \lambda_p \mathbf{a}_p \mathbf{a}'_p \\
 &= \mathbf{A} \mathbf{A}' \\
 &= \sum_{k=1}^m \lambda_k \mathbf{a}_k \mathbf{a}'_k + \sum_{k=m+1}^p \lambda_k \mathbf{a}_k \mathbf{a}'_k \\
 &= \sum_{k=1}^m l_k l'_k + \sum_{k=m+1}^p \lambda_k \mathbf{a}_k \mathbf{a}'_k
 \end{aligned}$$

where $l_k = \sqrt{\lambda_k}$ and the second term is not diagonal in general.

Assume

$$\psi_i = \sigma_{ii} - \sum_{k=1}^m l_{ik}^2 = \sigma_{ii} - \sum_{k=1}^m \lambda_i a_{ik}^2$$

then

$$\approx \mathbf{L}\mathbf{L}' +$$

To estimate \mathbf{L} and Ψ , we use the expected eigenvalues and eigenvectors from \mathbf{S} or \mathbf{R}

- The estimated factor loadings don't change as the number of actors increases
- The diagonal elements of $\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\Psi}$ are equal to the diagonal elements of \mathbf{S} and \mathbf{R} , but the covariances may not be exactly reproduced
- We select m so that the off-diagonal elements close to the values in \mathbf{S} (or to make the off-diagonal elements of $\mathbf{S} - \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\Psi}$ small)

20.3.1.2 Principal Factor Method

Consider modeling the correlation matrix, $\mathbf{R} = \mathbf{L}\mathbf{L}' + \mathbf{\Psi}$. Then

$$\mathbf{L}\mathbf{L}' = \mathbf{R} - \mathbf{\Psi} = \begin{pmatrix} h_1^2 & r_{12} & \dots & r_{1p} \\ r_{21} & h_2^2 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & h_p^2 \end{pmatrix}$$

where $h_i^2 = 1 - \psi_i$ (the communality)

Suppose that initial estimates are available for the communalities, $(h_1^*)^2, (h_2^*)^2, \dots, (h_p^*)^2$, then we can regress each trait on all the others, and then use the r^2 as h^2

The estimate of $\mathbf{R} - \mathbf{\Psi}$ at step k is

$$(\mathbf{R} - \mathbf{\Psi})_k = \begin{pmatrix} (h_1^*)^2 & r_{12} & \dots & r_{1p} \\ r_{21} & (h_2^*)^2 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & (h_p^*)^2 \end{pmatrix} = \mathbf{L}_k^*(\mathbf{L}_k^*)'$$

where

$$\mathbf{L}_k^* = (\sqrt{\hat{\lambda}_1^* \hat{\mathbf{a}}_1^*}, \dots, \sqrt{\hat{\lambda}_m^* \hat{\mathbf{a}}_m^*})$$

and

$$\hat{\psi}_{i,k}^* = 1 - \sum_{j=1}^m \hat{\lambda}_i^* (\hat{a}_{ij}^*)^2$$

we used the spectral decomposition on the estimated matrix $(\mathbf{R} -)$ to calculate the $\hat{\lambda}_i^*$ s and the $\hat{\mathbf{a}}_i^*$ s

After updating the values of $(\hat{h}_i^*)^2 = 1 - \hat{\psi}_{i,k}^*$ we will use them to form a new \mathbf{L}_{k+1}^* via another spectral decomposition. Repeat the process

Notes:

- The matrix $(\mathbf{R} -)_k$ is not necessarily positive definite
- The principal component method is similar to principal factor if one considers the initial communalities are $h^2 = 1$
- if m is too large, some communalities may become larger than 1, causing the iterations to terminate. To combat, we can
 - fix any communality that is greater than 1 at 1 and then continues.
 - continue iterations regardless of the size of the communalities. However, results can be outside fo the parameter space.

20.3.1.3 Maximum Likelihood Method

Since we need the likelihood function, we make the additional (critical) assumption that

- $\mathbf{y}_j \sim N(\mathbf{f}, \mathbf{\Psi})$ for $j = 1, \dots, n$
- $\mathbf{f} \sim N(\mathbf{0}, \mathbf{I})$
- $\epsilon_j \sim N(\mathbf{0}, \mathbf{\Psi})$

and restriction

- $\mathbf{L}'^{-1}\mathbf{L} = \mathbf{\Psi}$ where $\mathbf{\Psi}$ is a diagonal matrix. (since the factor loading matrix is not unique, we need this restriction).

Notes:

- Finding MLE can be computationally expensive
- we typically use other methods for exploratory data analysis
- Likelihood ratio tests could be used for testing hypotheses in this framework (i.e., Confirmatory Factor Analysis)

20.3.2 Factor Rotation

$\mathbf{T}_{m \times m}$ is an orthogonal matrix that has the property that

$$\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{I}} = \hat{\mathbf{L}}^*(\hat{\mathbf{L}}^*)' + \hat{\mathbf{I}}$$

where $\mathbf{L}^* = \mathbf{LT}$

This means that estimated specific variances and communalities are not altered by the orthogonal transformation.

Since there are an infinite number of choices for \mathbf{T} , some selection criterion is necessary

For example, we can find the orthogonal transformation that maximizes the objective function

$$\sum_{j=1}^m \left[\frac{1}{p} \sum_{i=1}^p \left(\frac{l_{ij}^{*2}}{h_i} \right)^2 - \left\{ \frac{\gamma}{p} \sum_{i=1}^p \left(\frac{l_{ij}^{*2}}{h_i} \right)^2 \right\}^2 \right]$$

where $\frac{l_{ij}^{*2}}{h_i}$ are “scaled loadings”, which gives variables with small communalities more influence.

Different choices of γ in the objective function correspond to different orthogonal rotation found in the literature;

1. Varimax $\gamma = 1$ (rotate the factors so that each of the p variables should have a high loading on only one factor, but this is not always possible).
2. Quartimax $\gamma = 0$
3. Equimax $\gamma = m/2$
4. Parsimax $\gamma = \frac{p(m-1)}{p+m-2}$
5. Promax: non-orthogonal or oblique transformations
6. Harris-Kaiser (HK): non-orthogonal or oblique transformations

20.3.3 Estimation of Factor Scores

Recall

$$(\mathbf{y}_j - \bar{\mathbf{y}}) = \mathbf{L}_{p \times m} \mathbf{f}_j + \epsilon_j$$

If the factor model is correct then

$$\text{var}(\epsilon_j) = \text{diag}(\psi_1, \dots, \psi_p)$$

Thus we could consider using weighted least squares to estimate $\hat{\mathbf{f}}_j$, the vector of factor scores for the j -th sampled unit by

$$\begin{aligned}\hat{\mathbf{f}} &= (\mathbf{L}'^{-1}\mathbf{L})^{-1}\mathbf{L}'^{-1}(\mathbf{y}_j - \bar{\mathbf{y}}) \\ &\approx (\mathbf{L}'^{-1}\mathbf{L})^{-1}\mathbf{L}'^{-1}(\mathbf{y}_j - \bar{\mathbf{y}})\end{aligned}$$

20.3.3.1 The Regression Method

Alternatively, we can use the regression method to estimate the factor scores

Consider the joint distribution of $(\mathbf{y}_j - \bar{\mathbf{y}})$ and \mathbf{f}_j assuming multivariate normality, as in the maximum likelihood approach. then,

$$\begin{pmatrix} \mathbf{y}_j - \bar{\mathbf{y}} \\ \mathbf{f}_j \end{pmatrix} \sim N_{p+m} \left(\begin{bmatrix} \mathbf{L}\mathbf{L}' + \mathbf{I}_{m \times m} & \mathbf{L} \\ \mathbf{L}' & \mathbf{I}_{m \times m} \end{bmatrix} \right)$$

when the m factor model is correct

Hence,

$$E(\mathbf{f}_j | \mathbf{y}_j - \bar{\mathbf{y}}) = \mathbf{L}'(\mathbf{L}\mathbf{L}' + \mathbf{I}_{m \times m})^{-1}(\mathbf{y}_j - \bar{\mathbf{y}})$$

notice that $\mathbf{L}'(\mathbf{L}\mathbf{L}' + \mathbf{I}_{m \times m})^{-1}$ is an $m \times p$ matrix of regression coefficients

Then, we use the estimated conditional mean vector to estimate the factor scores

$$\hat{\mathbf{f}}_j = \hat{\mathbf{L}}'(\hat{\mathbf{L}}\hat{\mathbf{L}}' + \mathbf{I}_{m \times m})^{-1}(\mathbf{y}_j - \bar{\mathbf{y}})$$

Alternatively, we could reduce the effect of possible incorrect determination fo the number of factors m by using \mathbf{S} as a substitute for $\hat{\mathbf{L}}\hat{\mathbf{L}}' + \mathbf{I}_{m \times m}$ then

$$\hat{\mathbf{f}}_j = \hat{\mathbf{L}}'\mathbf{S}^{-1}(\mathbf{y}_j - \bar{\mathbf{y}})$$

where $j = 1, \dots, n$

20.3.4 Model Diagnostic

- Plots
- Check for outliers (recall that $\mathbf{f}_j \sim iidN(\mathbf{0}, \mathbf{I}_{m \times m})$)
- Check for multivariate normality assumption
- Use univariate tests for normality to check the factor scores
- **Confirmatory Factor Analysis:** formal testing of hypotheses about loadings, use MLE and full/reduced model testing paradigm and measures of model fit

20.3.5 Application

In the `psych` package,

- `h2` = the communalities
- `u2` = the uniqueness
- `com` = the complexity

```
library(psych)
library(tidyverse)
## Load the data from the psych package
data(Harman.5)
Harman.5
#>      population schooling employment professional housevalue
#> Tract1       5700     12.8      2500        270     25000
#> Tract2       1000     10.9       600         10     10000
#> Tract3      3400      8.8      1000        10      9000
#> Tract4      3800     13.6      1700        140     25000
#> Tract5      4000     12.8      1600        140     25000
#> Tract6      8200      8.3      2600        60      12000
#> Tract7      1200     11.4       400         10     16000
#> Tract8      9100     11.5      3300        60     14000
#> Tract9      9900     12.5      3400        180     18000
#> Tract10     9600     13.7      3600        390     25000
#> Tract11     9600      9.6      3300        80      12000
#> Tract12     9400     11.4      4000        100     13000

# Correlation matrix
cor_mat <- cor(Harman.5)
cor_mat
```

```

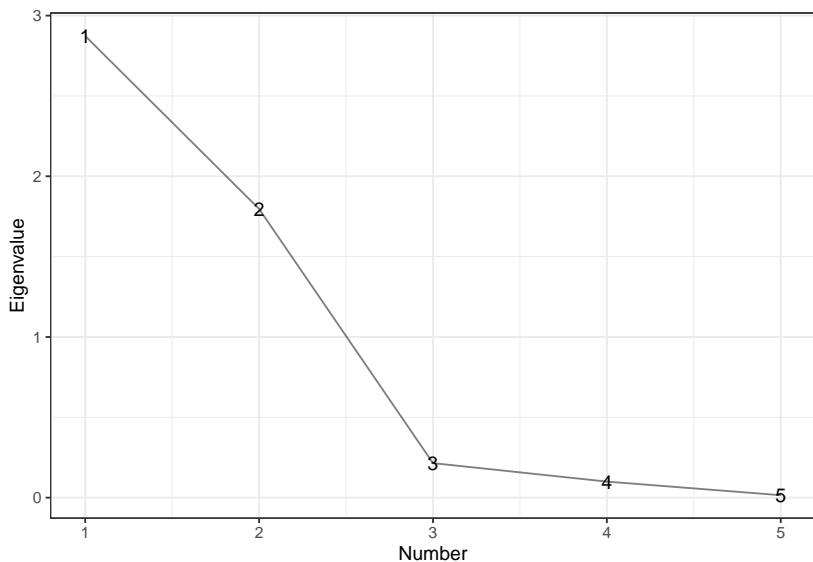
#>           population schooling employment professional housevalue
#> population   1.0000000 0.00975059  0.9724483   0.4388708 0.02241157
#> schooling    0.00975059 1.00000000  0.1542838   0.6914082 0.86307009
#> employment   0.97244826 0.15428378  1.0000000   0.5147184 0.12192599
#> professional 0.43887083 0.69140824  0.5147184   1.0000000 0.77765425
#> housevalue   0.02241157 0.86307009  0.1219260   0.7776543 1.00000000

## Principal Component Method with Correlation
cor_pca <- prcomp(Harman.5, scale = T)
# eigen values
cor_results <- data.frame(eigen_values = cor_pca$sdev ^ 2)

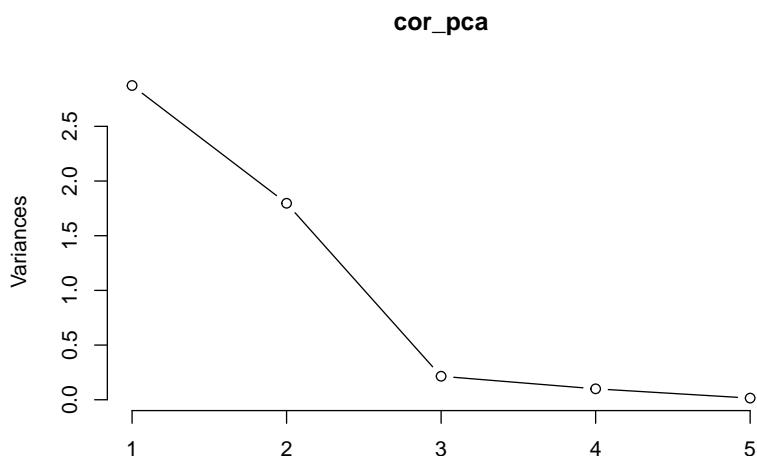
cor_results <- cor_results %>%
  mutate(
    proportion = eigen_values / sum(eigen_values),
    cumulative = cumsum(proportion),
    number = row_number()
  )
cor_results
#>   eigen_values proportion cumulative number
#> 1  2.87331359 0.574662719  0.5746627     1
#> 2  1.79666009 0.359332019  0.9339947     2
#> 3  0.21483689 0.042967377  0.9769621     3
#> 4  0.09993405 0.019986811  0.9969489     4
#> 5  0.01525537 0.003051075  1.0000000     5

# Scree plot of Eigenvalues
scree_gg <- ggplot(cor_results, aes(x = number, y = eigen_values)) +
  geom_line(alpha = 0.5) +
  geom_text(aes(label = number)) +
  scale_x_continuous(name = "Number") +
  scale_y_continuous(name = "Eigenvalue") +
  theme_bw()
scree_gg

```



```
screeplot(cor_pca, type = 'lines')
```



```
## Keep 2 factors based on scree plot and eigenvalues
factor_pca <- principal(Harman.5, nfactors = 2, rotate = "none")
factor_pca
#> Principal Components Analysis
```

```

#> Call: principal(r = Harman.5, nfactors = 2, rotate = "none")
#> Standardized loadings (pattern matrix) based upon correlation matrix
#>          PC1    PC2    h2   u2 com
#> population  0.58  0.81  0.99  0.012 1.8
#> schooling   0.77 -0.54  0.89  0.115 1.8
#> employment  0.67  0.73  0.98  0.021 2.0
#> professional 0.93 -0.10  0.88  0.120 1.0
#> housevalue   0.79 -0.56  0.94  0.062 1.8
#>
#>          PC1    PC2
#> SS loadings     2.87 1.80
#> Proportion Var  0.57 0.36
#> Cumulative Var 0.57 0.93
#> Proportion Explained 0.62 0.38
#> Cumulative Proportion 0.62 1.00
#>
#> Mean item complexity = 1.7
#> Test of the hypothesis that 2 components are sufficient.
#>
#> The root mean square of the residuals (RMSR) is 0.03
#> with the empirical chi square 0.29 with prob < 0.59
#>
#> Fit based upon off diagonal values = 1

# factor 1 = overall socioeconomic health
# factor 2 = contrast of the population and employment against school and house value

## Ssquared multiple correlation (SMC) prior, no rotation
factor_pca_smcl <- fa(
  Harman.5,
  nfactors = 2,
  fm = "pa",
  rotate = "none",
  SMC = TRUE
)
factor_pca_smcl
#> Factor Analysis using method = pa
#> Call: fa(r = Harman.5, nfactors = 2, rotate = "none", SMC = TRUE, fm = "pa")
#> Standardized loadings (pattern matrix) based upon correlation matrix
#>          PA1    PA2    h2   u2 com
#> population  0.62  0.78  1.00 -0.0027 1.9
#> schooling   0.70 -0.53  0.77  0.2277 1.9
#> employment  0.70  0.68  0.96  0.0413 2.0
#> professional 0.88 -0.15  0.80  0.2017 1.1

```

```

#> housevalue  0.78 -0.60 0.96  0.0361 1.9
#>
#>                               PA1  PA2
#> SS loadings            2.76 1.74
#> Proportion Var        0.55 0.35
#> Cumulative Var        0.55 0.90
#> Proportion Explained  0.61 0.39
#> Cumulative Proportion 0.61 1.00
#>
#> Mean item complexity = 1.7
#> Test of the hypothesis that 2 factors are sufficient.
#>
#> The degrees of freedom for the null model are 10 and the objective function was 10.34
#> The degrees of freedom for the model are 1 and the objective function was 0.34
#>
#> The root mean square of the residuals (RMSR) is 0.01
#> The df corrected root mean square of the residuals is 0.03
#>
#> The harmonic number of observations is 12 with the empirical chi square 0.02 with 1 degrees of freedom
#> The total number of observations was 12 with Likelihood Chi Square = 2.44 with 1 degrees of freedom
#>
#> Tucker Lewis Index of factoring reliability = 0.596
#> RMSEA index = 0.336 and the 90 % confidence intervals are 0.0967 to 0.5967
#> BIC = -0.04
#> Fit based upon off diagonal values = 1

## SMC prior, Promax rotation
factor_pca_smc_pro <- fa(
  Harman.5,
  nfactors = 2,
  fm = "pa",
  rotate = "Promax",
  SMC = TRUE
)
factor_pca_smc_pro
#> Factor Analysis using method = pa
#> Call: fa(r = Harman.5, nfactors = 2, rotate = "Promax", SMC = TRUE,
#>          fm = "pa")
#> Standardized loadings (pattern matrix) based upon correlation matrix
#>                               PA1  PA2   h2    u2 com
#> population      -0.11 1.02 1.00 -0.0027 1.0
#> schooling       0.90 -0.11 0.77  0.2277 1.0
#> employment      0.02  0.97 0.96  0.0413 1.0
#> professional    0.75  0.33 0.80  0.2017 1.4
#> housevalue      1.01 -0.14 0.96  0.0361 1.0

```

```

#>
#>          PA1  PA2
#> SS loadings    2.38 2.11
#> Proportion Var 0.48 0.42
#> Cumulative Var 0.48 0.90
#> Proportion Explained 0.53 0.47
#> Cumulative Proportion 0.53 1.00
#>
#> With factor correlations of
#>      PA1  PA2
#> PA1 1.00 0.25
#> PA2 0.25 1.00
#>
#> Mean item complexity = 1.1
#> Test of the hypothesis that 2 factors are sufficient.
#>
#> The degrees of freedom for the null model are 10 and the objective function was 6.38 with 0
#> The degrees of freedom for the model are 1 and the objective function was 0.34
#>
#> The root mean square of the residuals (RMSR) is 0.01
#> The df corrected root mean square of the residuals is 0.03
#>
#> The harmonic number of observations is 12 with the empirical chi square 0.02 with prob < 0.999999999999999
#> The total number of observations was 12 with Likelihood Chi Square = 2.44 with prob < 0.15
#>
#> Tucker Lewis Index of factoring reliability = 0.596
#> RMSEA index = 0.336 and the 90 % confidence intervals are 0 0.967
#> BIC = -0.04
#> Fit based upon off diagonal values = 1

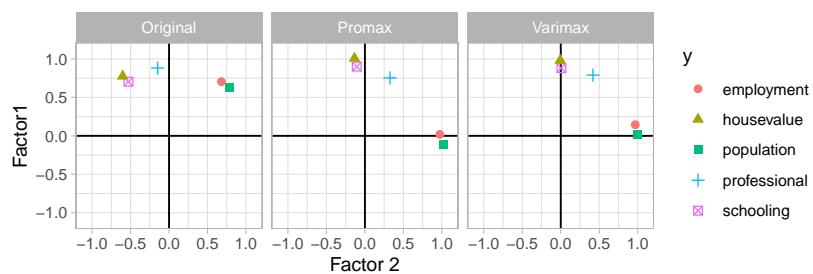
## SMC prior, varimax rotation
factor_pca_smc_var <- fa(
  Harman.5,
  nfactors = 2,
  fm = "pa",
  rotate = "varimax",
  SMC = TRUE
)
## Make a data frame of the loadings for ggplot2
factors_df <-
  bind_rows(
    data.frame(
      y = rownames(factor_pca_smc$loadings),
      unclass(factor_pca_smc$loadings)
    ),

```

```

data.frame(
  y = rownames(factor_pca_smc_pro$loadings),
  unclass(factor_pca_smc_pro$loadings)
),
data.frame(
  y = rownames(factor_pca_smc_var$loadings),
  unclass(factor_pca_smc_var$loadings)
),
.id = "Rotation"
)
flag_gg <- ggplot(factors_df) +
  geom_vline(aes(xintercept = 0)) +
  geom_hline(aes(yintercept = 0)) +
  geom_point(aes(
    x = PA2,
    y = PA1,
    col = y,
    shape = y
  ), size = 2) +
  scale_x_continuous(name = "Factor 2", limits = c(-1.1, 1.1)) +
  scale_y_continuous(name = "Factor1", limits = c(-1.1, 1.1)) + facet_wrap("Rotation",
  labeller =
  "1" =
  )) +
  coord_fixed(ratio = 1) # make aspect ratio of each facet 1
flag_gg

```



```
# promax and varimax did a good job to assign trait to a particular factor

factor_mle_1 <- fa(
  Harman.5,
  nfactors = 1,
  fm = "mle",
  rotate = "none",
  SMC = TRUE
)
factor_mle_1
#> Factor Analysis using method = ml
#> Call: fa(r = Harman.5, nfactors = 1, rotate = "none", SMC = TRUE, fm = "mle")
#> Standardized loadings (pattern matrix) based upon correlation matrix
#>           ML1    h2   u2 com
#> population  0.97 0.950 0.0503  1
#> schooling   0.14 0.021 0.9791  1
#> employment  1.00 0.995 0.0049  1
#> professional 0.51 0.261 0.7388  1
#> housevalue   0.12 0.014 0.9864  1
#>
#>           ML1
#> SS loadings  2.24
#> Proportion Var 0.45
#>
#> Mean item complexity = 1
#> Test of the hypothesis that 1 factor is sufficient.
#>
#> The degrees of freedom for the null model are 10 and the objective function was 6.38 with 0
#> The degrees of freedom for the model are 5 and the objective function was 3.14
#>
#> The root mean square of the residuals (RMSR) is 0.41
#> The df corrected root mean square of the residuals is 0.57
#>
#> The harmonic number of observations is 12 with the empirical chi square 39.41 with prob <
#> The total number of observations was 12 with Likelihood Chi Square = 24.56 with prob < 0
#>
#> Tucker Lewis Index of factoring reliability = 0.022
#> RMSEA index = 0.564 and the 90 % confidence intervals are 0.374 0.841
#> BIC = 12.14
#> Fit based upon off diagonal values = 0.5
#> Measures of factor score adequacy
#>           ML1
#> Correlation of (regression) scores with factors      1.00
#> Multiple R square of scores with factors            1.00
#> Minimum correlation of possible factor scores     0.99
```

```

factor_mle_2 <- fa(
  Harman.5,
  nfactors = 2,
  fm = "mle",
  rotate = "none",
  SMC = TRUE
)
factor_mle_2
#> Factor Analysis using method = ml
#> Call: fa(r = Harman.5, nfactors = 2, rotate = "none", SMC = TRUE, fm = "mle")
#> Standardized loadings (pattern matrix) based upon correlation matrix
#>           ML2   ML1   h2   u2 com
#> population    -0.03 1.00 1.00 0.005 1.0
#> schooling      0.90 0.04 0.81 0.193 1.0
#> employment     0.09 0.98 0.96 0.036 1.0
#> professional   0.78 0.46 0.81 0.185 1.6
#> housevalue     0.96 0.05 0.93 0.074 1.0
#>
#>           ML2   ML1
#> SS loadings      2.34 2.16
#> Proportion Var   0.47 0.43
#> Cumulative Var   0.47 0.90
#> Proportion Explained  0.52 0.48
#> Cumulative Proportion 0.52 1.00
#>
#> Mean item complexity = 1.1
#> Test of the hypothesis that 2 factors are sufficient.
#>
#> The degrees of freedom for the null model are 10 and the objective function was 0.31
#> The degrees of freedom for the model are 1 and the objective function was 0.31
#>
#> The root mean square of the residuals (RMSR) is 0.01
#> The df corrected root mean square of the residuals is 0.05
#>
#> The harmonic number of observations is 12 with the empirical chi square 0.05 with
#> The total number of observations was 12 with Likelihood Chi Square = 2.22 with p = 0.945
#>
#> Tucker Lewis Index of factoring reliability = 0.658
#> RMSEA index = 0.307 and the 90 % confidence intervals are 0.0945 to 0.945
#> BIC = -0.26
#> Fit based upon off diagonal values = 1
#> Measures of factor score adequacy
#>
#> Correlation of (regression) scores with factors      ML2   ML1
#> Multiple R square of scores with factors            0.98 1.00
#>                                         0.95 1.00

```

```

#> Minimum correlation of possible factor scores      0.91 0.99

factor_mle_3 <- fa(
  Harman.5,
  nfactors = 3,
  fm = "mle",
  rotate = "none",
  SMC = TRUE
)
factor_mle_3
#> Factor Analysis using method = ml
#> Call: fa(r = Harman.5, nfactors = 3, rotate = "none", SMC = TRUE, fm = "mle")
#> Standardized loadings (pattern matrix) based upon correlation matrix
#>          ML2   ML1   ML3   h2   u2 com
#> population -0.12  0.98 -0.11  0.98  0.0162 1.1
#> schooling   0.89  0.15  0.29  0.90  0.0991 1.3
#> employment   0.00  1.00  0.04  0.99  0.0052 1.0
#> professional 0.72  0.52 -0.10  0.80  0.1971 1.9
#> housevalue   0.97  0.13 -0.09  0.97  0.0285 1.1
#>
#>          ML2   ML1   ML3
#> SS loadings    2.28 2.26 0.11
#> Proportion Var 0.46 0.45 0.02
#> Cumulative Var 0.46 0.91 0.93
#> Proportion Explained 0.49 0.49 0.02
#> Cumulative Proportion 0.49 0.98 1.00
#>
#> Mean item complexity = 1.2
#> Test of the hypothesis that 3 factors are sufficient.
#>
#> The degrees of freedom for the null model are 10 and the objective function was 6.38 with 0
#> The degrees of freedom for the model are -2 and the objective function was 0
#>
#> The root mean square of the residuals (RMSR) is 0
#> The df corrected root mean square of the residuals is NA
#>
#> The harmonic number of observations is 12 with the empirical chi square 0 with prob < NA
#> The total number of observations was 12 with Likelihood Chi Square = 0 with prob < NA
#>
#> Tucker Lewis Index of factoring reliability = 1.318
#> Fit based upon off diagonal values = 1
#> Measures of factor score adequacy
#>          ML2   ML1   ML3
#> Correlation of (regression) scores with factors 0.99 1.00 0.82
#> Multiple R square of scores with factors       0.98 1.00 0.68

```

#> Minimum correlation of possible factor scores	0.96 0.99 0.36
--	----------------

The output info for the null hypothesis of no common factors is in the statement
 “The degrees of freedom for the null model ..”

The output info for the null hypothesis that number of factors is sufficient is in the statement “The total number of observations was ...”

One factor is not enough, two is sufficient, and not enough data for 3 factors (df of -2 and NA for p-value). Hence, we should use 2-factor model.

20.4 Discriminant Analysis

Suppose we have two or more different populations from which observations could come from. Discriminant analysis seeks to determine which of the possible population an observation comes from while making as few mistakes as possible

- This is an alternative to logistic approaches with the following advantages:
 - when there is clear separation between classes, the parameter estimates for the logic regression model can be **surprisingly** unstable, while discriminant approaches do not suffer
 - If X is normal in each of the classes and the sample size is small, then discriminant approaches can be more accurate

Notation

Similar to MANOVA, let $\mathbf{y}_{j1}, \mathbf{y}_{j2}, \dots, \mathbf{y}_{jn_j} \sim iid f_j(\mathbf{y})$ for $j = 1, \dots, h$

Let $f_j(\mathbf{y})$ be the density function for population j . Note that each vector \mathbf{y} contain measurements on all p traits

1. Assume that each observation is from one of h possible populations.
2. We want to form a discriminant rule that will allocate an observation \mathbf{y} to population j when \mathbf{y} is in fact from this population

20.4.1 Known Populations

The maximum likelihood discriminant rule for assigning an observation \mathbf{y} to one of the h populations allocates \mathbf{y} to the population that gives the largest likelihood to \mathbf{y}

Consider the likelihood for a single observation \mathbf{y} , which has the form $f_j(\mathbf{y})$ where j is the true population.

Since j is unknown, to make the likelihood as large as possible, we should choose the value j which causes $f_j(\mathbf{y})$ to be as large as possible

Consider a simple univariate example. Suppose we have data from one of two binomial populations.

- The first population has $n = 10$ trials with success probability $p = .5$
- The second population has $n = 10$ trials with success probability $p = .7$
- to which population would we assign an observation of $y = 7$
- Note:
 - $f(y = 7|n = 10, p = .5) = .117$
 - $f(y = 7|n = 10, p = .7) = .267$ where $f(\cdot)$ is the binomial likelihood.
 - Hence, we choose the second population

Another example

We have 2 populations, where

- First population: $N(\mu_1, \sigma_1^2)$
- Second population: $N(\mu_2, \sigma_2^2)$

The likelihood for a single observation is

$$f_j(y) = (2\pi\sigma_j^2)^{-1/2} \exp\left\{-\frac{1}{2}\left(\frac{y - \mu_j}{\sigma_j}\right)^2\right\}$$

Consider a likelihood ratio rule

$$\begin{aligned} \Lambda &= \frac{\text{likelihood of } y \text{ from pop 1}}{\text{likelihood of } y \text{ from pop 2}} \\ &= \frac{f_1(y)}{f_2(y)} \\ &= \frac{\sigma_2}{\sigma_1} \exp\left\{-\frac{1}{2}\left[\left(\frac{y - \mu_1}{\sigma_1}\right)^2 - \left(\frac{y - \mu_2}{\sigma_2}\right)^2\right]\right\} \end{aligned}$$

Hence, we classify into

- pop 1 if $\Lambda > 1$
- pop 2 if $\Lambda < 1$

- for ties, flip a coin

Another way to think:

we classify into population 1 if the “standardized distance” of y from μ_1 is less than the “standardized distance” of y from μ_2 which is referred to as a **quadratic discriminant rule**.

(Significant simplification occurs in the special case where $\sigma_1 = \sigma_2 = \sigma^2$)

Thus, we classify into population 1 if

$$(y - \mu_2)^2 > (y - \mu_1)^2$$

or

$$|y - \mu_2| > |y - \mu_1|$$

and

$$-2 \log(\Lambda) = -2y \frac{(\mu_1 - \mu_2)}{\sigma^2} + \frac{(\mu_1^2 - \mu_2^2)}{\sigma^2} = \beta y + \alpha$$

Thus, we classify into population 1 if this is less than 0.

Discriminant classification rule is linear in y in this case.

20.4.1.1 Multivariate Expansion

Suppose that there are 2 populations

- $N_p(\mu_1, \Sigma_1)$
- $N_p(\mu_2, \Sigma_2)$

$$\begin{aligned} -2 \log\left(\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right) &= \log|\Sigma_1| + (\mathbf{x} - \mu_1)' \Sigma_1^{-1} (\mathbf{x} - \mu_1) \\ &\quad - [\log|\Sigma_2| + (\mathbf{x} - \mu_2)' \Sigma_2^{-1} (\mathbf{x} - \mu_2)] \end{aligned}$$

Again, we classify into population 1 if this is less than 0, otherwise, population 2. And like the univariate case with non-equal variances, this is a quadratic discriminant rule.

And if the covariance matrices are equal: $\Sigma_1 = \Sigma_2 = \Sigma$ classify into population 1 if

$$(\mu_1 - \mu_2)' \Sigma^{-1} \mathbf{x} - \frac{1}{2} (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) \geq 0$$

This linear discriminant rule is also referred to as **Fisher's linear discriminant function**

By assuming the covariance matrices are equal, we assume that the shape and orientation fo the two populations must be the same (which can be a strong restriction)

In other words, for each variable, it can have different mean but the same variance.

- Note: LDA Bayes decision boundary is linear. Hence, quadratic decision boundary might lead to better classification. Moreover, the assumption of same variance/covariance matrix across all classes for Gaussian densities imposes the linear rule, if we allow the predictors in each class to follow MVN distribution with class-specific mean vectors and variance/covariance matrices, then it is **Quadratic Discriminant Analysis**. But then, you will have more parameters to estimate (which gives more flexibility than LDA) at the cost of more variance (bias -variance tradeoff).

When μ_1, μ_2 , Σ are known, the probability of misclassification can be determined:

$$\begin{aligned} P(2|1) &= P(\text{calssify into pop 2} | \mathbf{x} \text{ is from pop 1}) \\ &= P((\mu_1 - \mu_2)' \Sigma^{-1} \mathbf{x} \leq \frac{1}{2} (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) | \mathbf{x} \sim N(\mu_1, \Sigma)) \\ &= \Phi(-\frac{1}{2}\delta) \end{aligned}$$

where

- $\delta^2 = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)$
- Φ is the standard normal cdf

Suppose there are h possible populations, which are distributed as $N_p(\mu_p, \Sigma_p)$. Then, the maximum likelihood (linear) discriminant rule allocates \mathbf{y} to population j where j minimizes the squared Mahalanobis distance

$$(\mathbf{y} - \mu_j)' \Sigma_j^{-1} (\mathbf{y} - \mu_j)$$

20.4.1.2 Bayes Discriminant Rules

If we know that population j has prior probabilities π_j (assume $\pi_j > 0$) we can form the Bayes discriminant rule.

This rule allocates an observation \mathbf{y} to the population for which $\pi_j f_j(\mathbf{y})$ is maximized.

Note:

- **Maximum likelihood discriminant rule** is a special case of the **Bayes discriminant rule**, where it sets all the $\pi_j = 1/h$

Optimal Properties of Bayes Discriminant Rules

- let p_{ii} be the probability of correctly assigning an observation from population i
- then one rule (with probabilities p_{ii}) is as good as another rule (with probabilities p'_{ii}) if $p_{ii} \geq p'_{ii}$ for all $i = 1, \dots, h$
- The first rule is better than the alternative if $p_{ii} > p'_{ii}$ for at least one i .
- A rule for which there is no better alternative is called admissible
- Bayes Discriminant Rules are admissible
- If we utilized prior probabilities, then we can form the posterior probability of a correct allocation, $\sum_{i=1}^h \pi_i p_{ii}$
- Bayes Discriminant Rules have the largest possible posterior probability of correct allocation with respect to the prior
- These properties show that **Bayes Discriminant rule is our best approach.**

Unequal Cost

- We want to consider the cost misallocation Define c_{ij} to be the cost associated with allocation a member of population j to population i .
- Assume that
 - $c_{ij} > 0$ for all $i \neq j$
 - $c_{ij} = 0$ if $i = j$
- We could determine the expected amount of loss for an observation allocated to population i as $\sum_j c_{ij} p_{ij}$ where the p_{ij} s are the probabilities of allocating an observation from population j into population i

- We want to minimize the amount of loss expected for our rule. Using a Bayes Discrimination, allocate \mathbf{y} to the population j which minimizes $\sum_{k \neq j} c_{ij} \pi_k f_k(\mathbf{y})$
- We could assign equal probabilities to each group and get a maximum likelihood type rule. here, we would allocate \mathbf{y} to population j which minimizes $\sum_{k \neq j} c_{jk} f_k(\mathbf{y})$

Example:

Two binomial populations, each of size 10, with probabilities $p_1 = .5$ and $p_2 = .7$

And the probability of being in the first population is .9

However, suppose the cost of inappropriately allocating into the first population is 1 and the cost of incorrectly allocating into the second population is 5.

In this case, we pick population 1 over population 2

In general, we consider two regions, R_1 and R_2 associated with population 1 and 2:

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c_{12}\pi_2}{c_{21}\pi_1}$$

$$R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{c_{12}\pi_2}{c_{21}\pi_1}$$

where c_{12} is the cost of assigning a member of population 2 to population 1.

20.4.1.3 Discrimination Under Estimation

Suppose we know the form of the distributions for populations of interests, but we still have to estimate the parameters.

Example:

we know the distributions are multivariate normal, but we have to estimate the means and variances

The maximum likelihood discriminant rule allocates an observation \mathbf{y} to population j when j maximizes the function

$$f_j(\mathbf{y}|\hat{\theta})$$

where $\hat{\theta}$ are the maximum likelihood estimates of the unknown parameters

For instance, we have 2 multivariate normal populations with distinct means, but common variance covariance matrix

MLEs for μ_1 and μ_2 are $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$ and common S is \mathbf{S} .

Thus, an estimated discriminant rule could be formed by substituting these sample values for the population values

20.4.1.4 Native Bayes

- The challenge with classification using Bayes' is that we don't know the (true) densities, f_k , $k = 1, \dots, K$, while LDA and QDA make **strong multivariate normality assumptions** to deal with this.
- Naive Bayes makes only one assumption: **within the kth class, the p predictors are independent (i.e., for $k = 1, \dots, K$)**

$$f_k(x) = f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_p)$$

where f_{kj} is the density function of the j-th predictor among observation in the k-th class.

This assumption allows the use of joint distribution without the need to account for dependence between observations. However, this (native) assumption can be unrealistic, but still works well in cases where the number of sample (n) is not large relative to the number of features (p).

With this assumption, we have

$$P(Y = k | X = x) = \frac{\pi_k \times f_{k1}(x_1) \times \cdots \times f_{kp}(x_p)}{\sum_{l=1}^K \pi_l \times f_{l1}(x_1) \times \cdots \times f_{lp}(x_p)}$$

we only need to estimate the one-dimensional density function f_{kj} with either of these approaches:

- When X_j is quantitative, assume it has a univariate normal distribution (with independence): $X_j | Y = k \sim N(\mu_{jk}, \sigma_{jk}^2)$ which is more restrictive than QDA because it assumes predictors are independent (e.g., a diagonal covariance matrix)
- When X_j is quantitative, use a kernel density estimator Kernel Methods ; which is a smoothed histogram
- When X_j is qualitative, we count the promotion of training observations for the j-th predictor corresponding to each class.

20.4.1.5 Comparison of Classification Methods

Assuming we have K classes and K is the baseline from (James , Witten, Hastie, and Tibshirani book)

Comparing the log odds relative to the K class

20.4.1.5.1 Logistic Regression

$$\log\left(\frac{P(Y = k|X = x)}{P(Y = K|X = x)}\right) = \beta_{k0} + \sum_{j=1}^p \beta_{kj}x_j$$

20.4.1.5.2 LDA

$$\log\left(\frac{P(Y = k|X = x)}{P(Y = K|X = x)}\right) = a_k + \sum_{j=1}^p b_{kj}x_j$$

where a_k and b_{kj} are functions of $\pi_k, \pi_K, \mu_k, \mu_K$,

Similar to logistic regression, LDA assumes the log odds is linear in x

Even though they look like having the same form, the parameters in logistic regression are estimated by MLE, whereas LDA linear parameters are specified by the prior and normal distributions

We expect LDA to outperform logistic regression when the normality assumption (approximately) holds, and logistic regression to perform better when it does not

20.4.1.5.3 QDA

$$\log\left(\frac{P(Y = k|X = x)}{P(Y = K|X = x)}\right) = a_k + \sum_{j=1}^p b_{kj}x_j + \sum_{j=1}^p \sum_{l=1}^p c_{kjl}x_jx_l$$

where a_k, b_{kj}, c_{kjl} are functions $\pi_k, \pi_K, \mu_k, \mu_K, k, K$

20.4.1.5.4 Naive Bayes

$$\log\left(\frac{P(Y = k|X = x)}{P(Y = K|X = x)}\right) = a_k + \sum_{j=1}^p g_{kj}(x_j)$$

where $a_k = \log(\pi_k/\pi_K)$ and $g_{kj}(x_j) = \log(\frac{f_{kj}(x_j)}{f_{Kj}(x_j)})$ which is the form of generalized additive model

20.4.1.5.5 Summary

- LDA is a special case of QDA
- LDA is robust when it comes to high dimensions
- Any classifier with a linear decision boundary is a special case of naive Bayes with $g_{kj}(x_j) = b_{kj}x_j$, which means LDA is a special case of naive Bayes. LDA assumes that the features are normally distributed with a common within-class covariance matrix, and naive Bayes assumes independence of the features.
- Naive bayes is also a special case of LDA with restricted to a diagonal matrix with diagonals, σ^2 (another notation $diag(\cdot)$) assuming $f_{kj}(x_j) = N(\mu_{kj}, \sigma_j^2)$
- QDA and naive Bayes are not special case of each other. In principle, naive Bayes can produce a more flexible fit by the choice of $g_{kj}(x_j)$, but it's restricted to only purely additive fit, but QDA includes multiplicative terms of the form $c_{kjl}x_jx_l$
- None of these methods uniformly dominates the others: the choice of method depends on the true distribution of the predictors in each of the K classes, n and p (i.e., related to the bias-variance tradeoff).

Compare to the non-parametric method (KNN)

- KNN would outperform both LDA and logistic regression when the decision boundary is highly nonlinear, but can't say which predictors are most important, and requires many observations
- KNN is also limited in high-dimensions due to the curse of dimensionality
- Since QDA is a special type of nonlinear decision boundary (quadratic), it can be considered as a compromise between the linear methods and KNN classification. QDA can have fewer training observations than KNN but not as flexible.

From simulation:

True decision boundary	Best performance
Linear	LDA + Logistic regression
Moderately nonlinear	QDA + Naive Bayes
Highly nonlinear (many training, p is not large)	KNN

- like linear regression, we can also introduce flexibility by including transformed features \sqrt{X}, X^2, X^3

20.4.2 Probabilities of Misclassification

When the distribution are exactly known, we can determine the misclassification probabilities exactly. however, when we need to estimate the population parameters, we have to estimate the probability of misclassification

- Naive method
 - Plugging the parameters estimates into the form for the misclassification probabilities results to derive at the estimates of the misclassification probability.
 - But this will tend to be optimistic when the number of samples in one or more populations is small.
- Resubstitution method
 - Use the proportion of the samples from population i that would be allocated to another population as an estimate of the misclassification probability
 - But also optimistic when the number of samples is small
- Jack-knife estimates:
 - The above two methods use observation to estimate both parameters and also misclassification probabilities based upon the discriminant rule
 - Alternatively, we determine the discriminant rule based upon all of the data except the k -th observation from the j -th population
 - then, determine if the k -th observation would be misclassified under this rule
 - perform this process for all n_j observation in population j . An estimate fo the misclassfication probability would be the fraction of n_j observations which were misclassified
 - repeat the process for other $i \neq j$ populations
 - This method is more reliable than the others, but also computationally intensive
- Cross-Validation

Summary

Consider the group-specific densities $f_j(\mathbf{x})$ for multivariate vector \mathbf{x} .

Assume equal misclassification costs, the Bayes classification probability of \mathbf{x} belonging to the j -th population is

$$p(j|\mathbf{x}) = \frac{\pi_j f_j(\mathbf{x})}{\sum_{k=1}^h \pi_k f_k(\mathbf{x})}$$

$$j = 1, \dots, h$$

where there are h possible groups.

We then classify into the group for which this probability of membership is largest

Alternatively, we can write this in terms of a **generalized squared distance** formation

$$D_j^2(\mathbf{x}) = d_j^2(\mathbf{x}) + g_1(j) + g_2(j)$$

where

- $d_j^2(\mathbf{x}) = (\mathbf{x} - \mathbf{z}_j)' \mathbf{V}_j^{-1} (\mathbf{x} - \mathbf{z}_j)$ is the squared Mahalanobis distance from \mathbf{x} to the centroid of group j , and
 - $\mathbf{V}_j = \mathbf{S}_j$ if the within group covariance matrices are not equal
 - $\mathbf{V}_j = \mathbf{S}_p$ if a pooled covariance estimate is appropriate

and

$$g_1(j) = \begin{cases} \ln |\mathbf{S}_j| & \text{within group covariances are not equal} \\ 0 & \text{pooled covariance} \end{cases}$$

$$g_2(j) = \begin{cases} -2 \ln \pi_j & \text{prior probabilities are not equal} \\ 0 & \text{prior probabilities are equal} \end{cases}$$

then, the posterior probability of belonging to group j is

$$p(j|\mathbf{x}) = \frac{\exp(-.5D_j^2(\mathbf{x}))}{\sum_{k=1}^h \exp(-.5D_k^2(\mathbf{x}))}$$

$$\text{where } j = 1, \dots, h$$

and \mathbf{x} is classified into group j if $p(j|\mathbf{x})$ is largest for $j = 1, \dots, h$ (or, $D_j^2(\mathbf{x})$ is smallest).

20.4.2.1 Assessing Classification Performance

For binary classification, confusion matrix

		Predicted class		
True Class	- or Null	- or Null	+ or Null	Total
	+ or Null	True Neg (TN)	False Pos (FP)	N
	Total	False Neg (FN)	True Pos (TP)	P

	N*	P*	
--	----	----	--

and table 4.6 from (James et al., 2013)

Name	Definition	Synonyms
False Pos rate	FP/N	Type I error, 1 - Specificity
True Pos. rate	TP/P	1 - Type II error, power, sensitivity, recall
Pos Pred. value	TP/P*	Precision, 1 - false discovery promotion
Neg. Pred. value	TN/N*	

ROC curve (receiver Operating Characteristics) is a graphical comparison between **sensitivity** (true positive) and **specificity** ($= 1 - \text{false positive}$)

y-axis = true positive rate

x-axis = false positive rate

as we change the threshold rate for classifying an observation as from 0 to 1

AUC (area under the ROC) ideally would equal to 1, a bad classifier would have AUC = 0.5 (pure chance)

20.4.3 Unknown Populations/ Nonparametric Discrimination

When your multivariate data are not Gaussian, or known distributional form at all, we can use the following methods

20.4.3.1 Kernel Methods

We approximate $f_j(\mathbf{x})$ by a kernel density estimate

$$\hat{f}_j(\mathbf{x}) = \frac{1}{n_j} \sum_{i=1}^{n_j} K_j(\mathbf{x} - \mathbf{x}_i)$$

where

- $K_j(\cdot)$ is a kernel function satisfying $\int K_j(\mathbf{z})d\mathbf{z} = 1$
- $\mathbf{x}_i, i = 1, \dots, n_j$ is a random sample from the j -th population.

Thus, after finding $\hat{f}_j(\mathbf{x})$ for each of the h populations, the posterior probability of group membership is

$$p(j|\mathbf{x}) = \frac{\pi_j \hat{f}_j(\mathbf{x})}{\sum_{k=1}^h \pi_k \hat{f}_k(\mathbf{x})}$$

where $j = 1, \dots, h$

There are different choices for the kernel function:

- Uniform
- Normal
- Epanechnikov
- Biweight
- Triweight

With these kernels, we have to pick the “radius” (or variance, width, window width, bandwidth) of the kernel, which is a smoothing parameter (the larger the radius, the more smooth the kernel estimate of the density).

To select the smoothness parameter, we can use the following method

If we believe the populations were close to multivariate normal, then

$$R = \left(\frac{4/(2p+1)}{n_j} \right)^{1/(p+1)}$$

But since we do not know for sure, we might choose several different values and select one that vies the best out of sample or cross-validation discrimination.

Moreover, you also have to decide whether to use different kernel smoothness for different populations, which is similar to the individual and pooled covariances in the classical methodology.

20.4.3.2 Nearest Neighbor Methods

The nearest neighbor (also known as k-nearest neighbor) method performs the classification of a new observation vector based on the group membership of its nearest neighbors. In practice, we find

$$d_{ij}^2(\mathbf{x}, \mathbf{x}_i) = (\mathbf{x}, \mathbf{x}_i) V_j^{-1}(\mathbf{x}, \mathbf{x}_i)$$

which is the distance between the vector \mathbf{x} and the i-th observation in group j

We consider different choices for V_j

For example,

$$\mathbf{V}_j = \mathbf{S}_p \mathbf{V}_j = \mathbf{S}_j \mathbf{V}_j = \mathbf{I} \mathbf{V}_j = \text{diag}(\mathbf{S}_p)$$

We find the k observations that are closest to \mathbf{x} (where users pick k). Then we classify into the most common population, weighted by the prior.

20.4.3.3 Modern Discriminant Methods

Note:

Logistic regression (with or without random effects) is a flexible model-based procedure for classification between two populations.

The extension of logistic regression to the multi-group setting is polychotomous logistic regression (or, multinomial regression).

The machine learning and pattern recognition are growing with strong focus on nonlinear discriminant analysis methods such as:

- radial basis function networks
- support vector machines
- multiplayer perceptrons (neural networks)

The general framework

$$g_j(\mathbf{x}) = \sum_{l=1}^m w_{jl} \phi_l(\mathbf{x}; \boldsymbol{\theta}_l) + w_{j0}$$

where

- $j = 1, \dots, h$

- m nonlinear basis functions ϕ_l , each of which has n_m parameters given by $\theta_l = \{\theta_{lk} : k = 1, \dots, n_m\}$

We assign \mathbf{x} to the j -th population if $g_j(\mathbf{x})$ is the maximum for all $j = 1, \dots, h$

Development usually focuses on the choice and estimation of the basis functions, ϕ_l and the estimation of the weights w_{jl}

More details can be found (Webb, 2002)

20.4.4 Application

```
library(class)
library(klaR)
library(MASS)
library(tidyverse)

## Read in the data
crops <- read.table("images/crops.txt")
names(crops) <- c("crop", "y1", "y2", "y3", "y4")
str(crops)
#> 'data.frame':   36 obs. of  5 variables:
#> $ crop: chr  "Corn" "Corn" "Corn" "Corn" ...
#> $ y1  : int  16 15 16 18 15 15 12 20 24 21 ...
#> $ y2  : int  27 23 27 20 15 32 15 23 24 25 ...
#> $ y3  : int  31 30 27 25 31 32 16 23 25 23 ...
#> $ y4  : int  33 30 26 23 32 15 73 25 32 24 ...

## Read in test data
crops_test <- read.table("images/crops_test.txt")
names(crops_test) <- c("crop", "y1", "y2", "y3", "y4")
str(crops_test)
#> 'data.frame':   5 obs. of  5 variables:
#> $ crop: chr  "Corn" "Soybeans" "Cotton" "Sugarbeets" ...
#> $ y1  : int  16 21 29 54 32
#> $ y2  : int  27 25 24 23 32
#> $ y3  : int  31 23 26 21 62
#> $ y4  : int  33 24 28 54 16
```

20.4.4.1 LDA

Default prior is proportional to sample size and `lda` and `qda` do not fit a constant or intercept term

```

## Linear discriminant analysis
lda_mod <- lda(crop ~ y1 + y2 + y3 + y4,
                 data = crops)
lda_mod
#> Call:
#> lda(crop ~ y1 + y2 + y3 + y4, data = crops)
#>
#> Prior probabilities of groups:
#>      Clover      Corn      Cotton      Soybeans      Sugarbeets
#> 0.3055556 0.1944444 0.1666667 0.1666667 0.1666667
#>
#> Group means:
#>
#>      y1      y2      y3      y4
#> Clover 46.36364 32.63636 34.18182 36.63636
#> Corn   15.28571 22.71429 27.42857 33.14286
#> Cotton 34.50000 32.66667 35.00000 39.16667
#> Soybeans 21.00000 27.00000 23.50000 29.66667
#> Sugarbeets 31.00000 32.16667 20.00000 40.50000
#>
#> Coefficients of linear discriminants:
#>          LD1         LD2         LD3         LD4
#> y1 -6.147360e-02 0.009215431 -0.02987075 -0.014680566
#> y2 -2.548964e-02 0.042838972 0.04631489 0.054842132
#> y3 1.642126e-02 -0.079471595 0.01971222 0.008938745
#> y4 5.143616e-05 -0.013917423 0.05381787 -0.025717667
#>
#> Proportion of trace:
#>    LD1    LD2    LD3    LD4
#> 0.7364 0.1985 0.0576 0.0075

## Look at accuracy on the training data
lda_fitted <- predict(lda_mod,newdata = crops)
# Contingency table
lda_table <- table(truth = crops$crop, fitted = lda_fitted$class)
lda_table
#>           fitted
#> truth      Clover Corn Cotton Soybeans Sugarbeets
#> Clover        6   0     3     0      2
#> Corn          0   6     0     1      0
#> Cotton        3   0     1     2      0
#> Soybeans      0   1     1     3      1
#> Sugarbeets    1   1     0     2      2
# accuracy of 0.5 is just random (not good)

## Posterior probabilities of membership

```

```

crops_post <- cbind.data.frame(crops,
                                crop_pred = lda_fitted$class,
                                lda_fitted$posterior)
crops_post <- crops_post %>%
  mutate(missed = crop != crop_pred)
head(crops_post)
#>   crop y1 y2 y3 y4 crop_pred      Clover      Corn      Cotton  Soybeans
#> 1 Corn  16 27 31 33      Corn 0.08935164 0.4054296 0.1763189 0.2391845
#> 2 Corn  15 23 30 30      Corn 0.07690181 0.4558027 0.1420920 0.2530101
#> 3 Corn  16 27 27 26      Corn 0.09817815 0.3422454 0.1365315 0.3073105
#> 4 Corn  18 20 25 23      Corn 0.10521511 0.3633673 0.1078076 0.3281477
#> 5 Corn  15 15 31 32      Corn 0.05879921 0.5753907 0.1173332 0.2086696
#> 6 Corn  15 32 32 15  Soybeans 0.09723648 0.3278382 0.1318370 0.3419924
#>   Sugarbeets missed
#> 1 0.08971545 FALSE
#> 2 0.07219340 FALSE
#> 3 0.11573442 FALSE
#> 4 0.09546233 FALSE
#> 5 0.03980738 FALSE
#> 6 0.10109590 TRUE
# posterior shows that posterior of corn membership is much higher than the prior

## LOOCV
# leave-one-out cross validation for linear discriminant analysis
# cannot run the prdecit function using the object with CV = TRUE because it returns t
lda_cv <- lda(crop ~ y1 + y2 + y3 + y4,
               data = crops, CV = TRUE)
# Contingency table
lda_table_cv <- table(truth = crops$crop, fitted = lda_cv$class)
lda_table_cv
#>           fitted
#> truth      Clover  Corn  Cotton  Soybeans  Sugarbeets
#>   Clover        4     3     1      0       3
#>   Corn          0     4     1      2       0
#>   Cotton         3     0     0      2       1
#>   Soybeans       0     1     1      3       1
#>   Sugarbeets     2     1     0      2       1

## Predict the test data
lda_pred <- predict(lda_mod, newdata = crops_test)

## Make a contingency table with truth and most likely class
table(truth=crops_test$crop, predict=lda_pred$class)
#>           predict
#> truth      Clover  Corn  Cotton  Soybeans  Sugarbeets

```

```
#>   Clover      0      0      1      0      0
#>   Corn       0      1      0      0      0
#>   Cotton     0      0      0      1      0
#>   Soybeans   0      0      0      1      0
#>   Sugarbeets 1      0      0      0      0
```

LDA didn't do well on both within sample and out-of-sample data.

20.4.4.2 QDA

```
## Quadratic discriminant analysis
qda_mod <- qda(crop ~ y1 + y2 + y3 + y4,
                 data = crops)

## Look at accuracy on the training data
qda_fitted <- predict(qda_mod, newdata = crops)
# Contingency table
qda_table <- table(truth = crops$crop, fitted = qda_fitted$class)
qda_table
#>           fitted
#>   truth      Clover Corn Cotton Soybeans Sugarbeets
#>   Clover      9     0     0     0      2
#>   Corn        0     7     0     0      0
#>   Cotton      0     0     6     0      0
#>   Soybeans    0     0     0     6      0
#>   Sugarbeets  0     0     1     1      4

## LOOCV
qda_cv <- qda(crop ~ y1 + y2 + y3 + y4,
                  data = crops, CV = TRUE)
# Contingency table
qda_table_cv <- table(truth = crops$crop, fitted = qda_cv$class)
qda_table_cv
#>           fitted
#>   truth      Clover Corn Cotton Soybeans Sugarbeets
#>   Clover      9     0     0     0      2
#>   Corn        3     2     0     0      2
#>   Cotton      3     0     2     0      1
#>   Soybeans    3     0     0     2      1
#>   Sugarbeets  3     0     1     1      1

## Predict the test data
qda_pred <- predict(qda_mod, newdata = crops_test)
```

```
## Make a contingency table with truth and most likely class
table(truth = crops_test$crop, predict = qda_pred$class)
#>           predict
#> truth      Clover Corn Cotton Soybeans Sugarbeets
#> Clover       1    0     0      0      0
#> Corn         0    1     0      0      0
#> Cotton       0    0     1      0      0
#> Soybeans     0    0     0      1      0
#> Sugarbeets   0    0     0      0      1
```

20.4.4.3 KNN

knn uses design matrices of the features.

```
## Design matrices
X_train <- crops %>%
  dplyr::select(-crop)
X_test <- crops_test %>%
  dplyr::select(-crop)
Y_train <- crops$crop
Y_test <- crops_test$crop

## Nearest neighbors with 2 neighbors
knn_2 <- knn(X_train, X_train, Y_train, k = 2)
table(truth = Y_train, fitted = knn_2)
#>           fitted
#> truth      Clover Corn Cotton Soybeans Sugarbeets
#> Clover      10    0     1      0      0
#> Corn        0    7     0      0      0
#> Cotton      1    0     3      0      2
#> Soybeans    0    0     0      4      2
#> Sugarbeets   1    0     0      0      5

## Accuracy
mean(Y_train==knn_2)
#> [1] 0.8055556

## Performance on test data
knn_2_test <- knn(X_train, X_test, Y_train, k = 2)
table(truth = Y_test, predict = knn_2_test)
#>           predict
#> truth      Clover Corn Cotton Soybeans Sugarbeets
#> Clover       1    0     0      0      0
#> Corn         0    1     0      0      0
```

```

#> Cotton      0  0  0  0  1
#> Soybeans    0  0  0  1  0
#> Sugarbeets  0  0  0  0  1

## Accuracy
mean(Y_test==knn_2_test)
#> [1] 0.8

## Nearest neighbors with 3 neighbors
knn_3 <- knn(X_train, X_train, Y_train, k = 3)
table(truth = Y_train, fitted = knn_3)
#>          fitted
#> truth      Clover Corn Cotton Soybeans Sugarbeets
#> Clover      8   0    3     0      0
#> Corn        0   4    1     2      0
#> Cotton      1   0    4     1      0
#> Soybeans    0   0    0     5      1
#> Sugarbeets  0   1    1     1      3

## Accuracy
mean(Y_train==knn_3)
#> [1] 0.6666667

## Performance on test data
knn_3_test <- knn(X_train, X_test, Y_train, k = 3)
table(truth = Y_test, predict = knn_3_test)
#>          predict
#> truth      Clover Corn Cotton Soybeans Sugarbeets
#> Clover      1   0    0     0      0
#> Corn        0   1    0     0      0
#> Cotton      0   0    1     0      0
#> Soybeans    0   0    0     1      0
#> Sugarbeets  0   0    0     0      1

## Accuracy
mean(Y_test==knn_3_test)
#> [1] 1

```

20.4.4.4 Stepwise

Stepwise discriminant analysis using the `stepclass` in function in the `klaR` package.

```

step <- stepclass(
  crop ~ y1 + y2 + y3 + y4,
  data = crops,
  method = "qda",
  improvement = 0.15
)
#> correctness rate: 0.46667;  in: "y1";  variables (1): y1
#>
#> hr.elapsed min.elapsed sec.elapsed
#>      0.00        0.00       0.12

step$process
#>   step var varname result.pm
#> 0 start 0      -- 0.0000000
#> 1 in    1      y1 0.4666667

step$performance.measure
#> [1] "correctness rate"

```

Iris Data

```

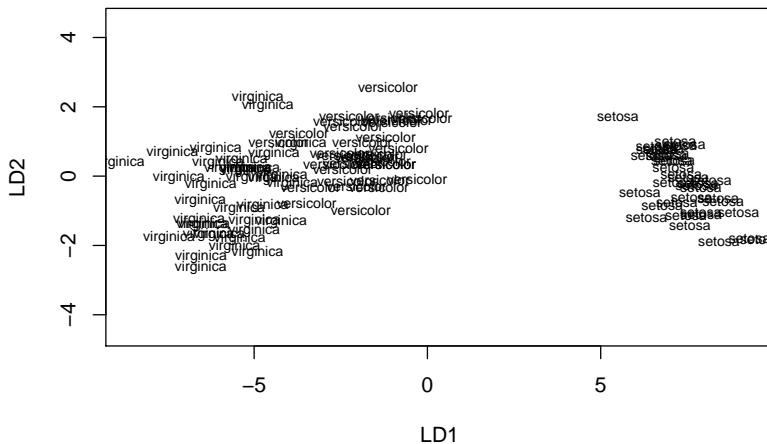
library(dplyr)
data('iris')
set.seed(1)
samp <-
  sample.int(nrow(iris), size = floor(0.70 * nrow(iris)), replace = F)

train.iris <- iris[samp, ] %>% mutate_if(is.numeric, scale)
test.iris <- iris[-samp, ] %>% mutate_if(is.numeric, scale)

library(ggplot2)
iris.model <- lda(Species ~ ., data = train.iris)
#pred
pred.lda <- predict(iris.model, test.iris)
table(truth = test.iris$Species, prediction = pred.lda$class)
#>           prediction
#> truth      setosa versicolor virginica
#> setosa      15      0      0
#> versicolor     0     17      0
#> virginica     0      0     13

plot(iris.model)

```



```

iris.model.qda <- qda(Species~., data=train.iris)
#pred
pred.qda <- predict(iris.model.qda, test.iris)
table(truth=test.iris$Species, prediction=pred.qda$class)
#>           prediction
#>   truth      setosa versicolor virginica
#>   setosa       15      0        0
#>   versicolor    0     16        1
#>   virginica     0      0       13

```

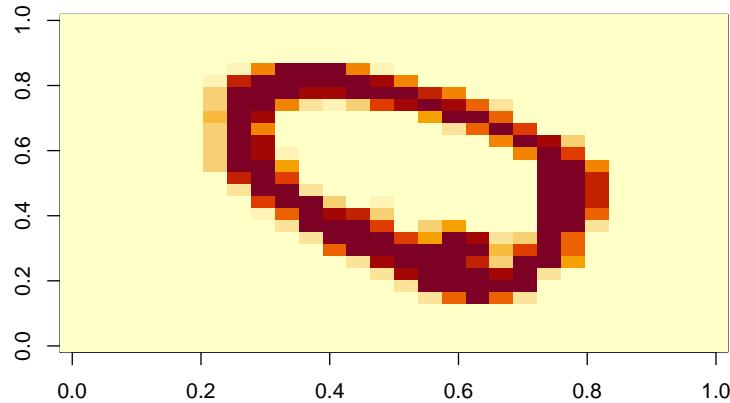
20.4.4.5 PCA with Discriminant Analysis

we can use both PCA for dimension reduction in discriminant analysis

```

zeros <- as.matrix(read.table("images/mnist0_train_b.txt"))
nines <- as.matrix(read.table("images/mnist9_train_b.txt"))
train <- rbind(zeros[1:1000, ], nines[1:1000, ])
train <- train / 255 #divide by 255 per notes (so ranges from 0 to 1)
train <- t(train) #each column is an observation
image(matrix(train[, 1], nrow = 28), main = 'Example image, unrotated')

```

Example image, unrotated

```

test <- rbind(zeros[2501:3000, ], nines[2501:3000, ])
test <- test / 255
test <- t(test)
y.train <- c(rep(0, 1000), rep(9, 1000))
y.test <- c(rep(0, 500), rep(9, 500))

library(MASS)
pc <- prcomp(t(train))
train.large <- data.frame(cbind(y.train, pc$x[, 1:10]))
large <- lda(y.train ~ ., data = train.large)
#the test data set needs to be constructed w/ the same 10 princomps
test.large <- data.frame(cbind(y.test, predict(pc, t(test))[, 1:10]))
pred.lda <- predict(large, test.large)
table(truth = test.large$y.test, prediction = pred.lda$class)
#>      prediction
#> truth    0    9
#>       0 491    9
#>       9   5 495

large.qda <- qda(y.train~., data=train.large)
#prediction
pred.qda <- predict(large.qda, test.large)
table(truth=test.large$y.test, prediction=pred.qda$class)
#>      prediction

```

```
#> truth    0    9  
#>      0 493    7  
#>      9    3 497
```


B.
QUASI-EXPERIMENTAL
DESIGN

Chapter 21

Quasi-experimental

In most cases, it means that you have pre- and post-intervention data.

A great resource for causal inference is Causal Inference Mixtape, especially if you like to read about the history of causal inference as a field as well (codes for Stata, R, and Python).

Identification strategy for any quasi-experiment (No ways to prove or formal statistical test, but you can provide plausible argument and evidence)

1. Where the exogenous variation comes from (by argument and institutional knowledge)
2. Exclusion restriction: Evidence that the variation in the exogenous shock and the outcome is due to no other factors
 1. Stable unit treatment value assumption (SUTVA) state that the treatment of unit i affect only the outcome of unit i (i.e., no spillover to the control groups)

All quasi-experimental methods involve tradeoff between power and support for the exogeneity assumption (i.e., discard variation in the data that is not exogenous).

Consequently, we don't usually look at R^2 (Ebbes et al., 2011). And it can even be misleading to use R^2 as the basis for model comparison.

Clustering should be based on the design, not the expectations of correlation (Abadie et al., 2017). With **small sample**, you should use **wild bootstrap procedure** (Cameron et al., 2008) to correct for the downward bias (see (Canay et al., 2021) for additional assumptions).

Typical robustness check: recommended by (Goldfarb et al., 2022)

- Different controls: show models with and without controls. Typically, we want to see the change in the estimate of interest. See (Altonji et al., 2005) for formal assessment based on Rosenbaum bounds (i.e., changes in the estimate and threat of Omitted variables on the estimate). For specific application in marketing, see (Manchanda et al., 2015) (Shin et al., 2012)
- Different functional forms
- Different window of time (in longitudinal setting)
- Different dependent variables (those that are related) or different measure of the dependent variables
- Different control group size (matched vs. un-matched samples)
- Placebo tests: see each placebo test for each setting below.

Showing the mechanism:

- Mediation analysis
- Moderation analysis
 - Estimate the model separate (for different group)
 - Assess whether the three-way interaction between the source of variation (e.g., under DID, cross-sectional and time series) and group membership is significant.

External Validity:

- Assess how representative your sample is
- Explain limitation of the design
- Use quasi-experimental results in conjunction with structural models: see (Anderson et al., 2015) [Einav et al. (2010)](Chung et al., 2014)

Limitation

1. What is your identifying assumptions or identification strategy
2. What are threats to the validity of your assumptions?
3. What you do to address it? And maybe how future research can do to address it.

Chapter 22

Regression Discontinuity

- A regression discontinuity occurs when there is a discrete change (jump) in treatment likelihood in the distribution of a continuous (or roughly continuous) variable (i.e., **running/forcing/assignment variable**).
 - Running variable can also be time, but the argument for time to be continuous is hard to argue because usually we do not see increment of time (e.g., quarterly or annual data). Unless we have minute or hour data, then we might be able to argue for it.
- Review paper (Imbens and Lemieux, 2007; Lee and Lemieux, 2010)
- Other readings:
 - https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_rd.pdf
 - https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_rdd_standards_122315.pdf
- (Thistlethwaite and Campbell, 1960): first paper to use RD in the context of merit awards on future academic outcomes.
- RD is a localized experiment at the cutoff point
 - Hence, we always have to qualify (perfunctory) our statement in research articles that “our research might not generalize to beyond the bandwidth.”
- In reality, RD and experimental (from random assignment) estimates are very similar ((Chaplin et al., 2018; Bertanha and Imbens, 2014); Mathematica). But still, it’s hard to prove empirically for every context (there might be future study that finds a huge difference between local estimate - causal - and overall estimate - random assignment).

- Threats: only valid near threshold: inference at threshold is valid on average. Interestingly, random experiment showed the validity already.
- Tradeoff between efficiency and bias
- Regression discontinuity is under the framework of Instrumental Variable argued by (Angrist and Krueger, 1999) and a special case of the Matching Methods (matching at one point) argued by (Heckman et al., 1999).
- The hard part is to find a setting that can apply, but once you find one, it's easy to apply
- We can also have multiple cutoff lines. However, for each cutoff line, there can only be one breakup point
- RD can have multiple coinciding effects (i.e., joint distribution or bundled treatment), then RD effect in this case would be the joint effect.
- As the running variable becomes more discrete your framework should be Interrupted Time Series, but more granular levels you can use RD. When you have infinite data (or substantially large) the two frameworks are identical. RD is always better than Interrupted Time Series
- Multiple alternative model specifications that produce consistent result are more reliable (parametric - linear regression with polynomials terms, and non-parametric - local linear regression)
- RD should be viewed more as a description of a data generating process, rather than a method or approach (similar to randomized experiment)
- RD is close to
 - other quasi-experimental methods in the sense that it's based on the discontinuity at a threshold
 - randomized experiments in the sense that it's local randomization.

There are several types of Regression Discontinuity:

1. Sharp RD: Change in treatment probability at the cutoff point is 1
 - Kink design: Instead of a discontinuity in the level of running variable, we have a discontinuity in the slope of the function (while the function/level can remain continuous) (Nielsen et al., 2010). See (Böckerman et al., 2018) for application, and (Card et al., 2012, 2015) for theory.
2. Kink RD
3. Fuzzy RD: Change in treatment probability less than 1

4. Fuzzy Kink RD
5. RDiT: running variable is time.

Others:

- Multiple cutoff
- Multiple Scores
- Geographic RD
- Dynamic Treatments
- Continuous Treatments

Consider

$$D_i = 1_{X_i > c}$$

$$D_i = \begin{cases} D_i = 1 & \text{if } X_i > C \\ D_i = 0 & \text{if } X_i < C \end{cases}$$

where

- D_i = treatment effect
- X_i = score variable (continuous)
- c = cutoff point

Identification (Identifying assumptions) of RD:

Average Treatment Effect at the cutoff (continuity-based)

$$\begin{aligned} \alpha_{SRDD} &= E[Y_{1i} - Y_{0i} | X_i = c] \\ &= E[Y_{1i} | X_i = c] - E[Y_{0i} | X_i = c] \\ &= \lim_{x \rightarrow c^+} E[Y_{1i} | X_i = c] - \lim_{x \rightarrow c^-} E[Y_{0i} | X_i = c] \end{aligned}$$

Average Treatment Effect in a neighborhood (Local Randomization-based):

$$\begin{aligned} \alpha_{LR} &= E[Y_{1i} - Y_{0i} | X_i \in W] \\ &= \frac{1}{N_1} \sum_{X_i \in W, T_i=1} Y_i - \frac{1}{N_0} \sum_{X_i \in W, T_i=0} Y_i \end{aligned}$$

Local Randomization Approach assumes that inside the chosen window $W = [c - w, c + w]$ are assigned to treatment as good as random:

1. Joint probability distribution of scores for units inside the chosen window W is known
2. Potential outcomes are not affected by value of the score

$$Y_i(0, x) = Y_i(0)Y_i(1, x) = Y_i(1)$$

This approach is stronger than the continuity-based because we assume the regressions are continuously at c and unaffected by the running variable inside the chosen window W

RDD estimates the local average treatment effect (LATE), at the cutoff point which is not at the individual or population levels.

Since researchers typically care more about the internal validity, than external validity, localness affects only external validity.

Assumptions:

- Independent assignment
- Continuity of conditional regression functions
 - $E[Y(0)|X = x]$ and $E[Y(1)|X = x]$ are continuous in x .
- RD is valid if cutpoint is **exogenous** (i.e., **no endogenous selection**) and running variable is **not manipulable**
- Only treatment(s) (e.g., could be joint distribution of multiple treatments) cause discontinuity or jump in the outcome variable
- All other factors are **smooth** through the cutoff (i.e., threshold) value. (we can also test this assumption by seeing no discontinuity in other factors). If they “jump”, they will bias your causal estimate

Threats to RD

- Variables (other than treatment) change discontinuously at the cutoff
 - We can test for jumps in these variables (including pre-treatment outcome)
- Multiple discontinuities for the assignment variable
- Manipulation of the assignment variable
 - At the cutoff point, check for continuity in the density of the assignment variable.

22.1 Specification Checks

1. Balance Checks
2. Sorting/Bunching/Manipulation
3. Placebo Tests
4. Sensitivity to Bandwidth Choice

22.1.1 Balance Checks

- Also known as checking for Discontinuities in Average Covariates
- Null Hypothesis: The average effect of covariates on pseudo outcomes (i.e., those qualitatively cannot be affected by the treatment) is 0.
- If this hypothesis is rejected, you better have a good reason to why because it can cast serious doubt on your RD design.

22.1.2 Sorting/Bunching/Manipulation

- Also known as checking for A Discontinuity in the Distribution of the Forcing Variable
- Also known as clustering or density test
- Formal test is McCrary sorting test (McCrary, 2008) or (Cattaneo et al., 2019)
- Since human subjects can manipulate the running variable to be just above or below the cutoff (assuming that the running variable is manipulable), especially when the cutoff point is known in advance for all subjects, this can result in a discontinuity in the distribution of the running variable at the cutoff (i.e., we will see “bunching” behavior right before or after the cutoff)>
 - People would like to sort into treatment if it’s desirable. The density of the running variable would be 0 just below the threshold
 - People would like to be out of treatment if it’s undesirable
- (McCrary, 2008) proposes a density test (i.e., a formal test for manipulation of the assignment variable).
 - H_0 : The continuity of the density of the running variable (i.e., the covariate that underlies the assignment at the discontinuity point)
 - H_a : A jump in the density function at that point

- Even though it's not a requirement that the density of the running must be continuous at the cutoff, but a discontinuity can suggest manipulations.
- (Lee and Lemieux, 2009) offers a guide to know when you should warrant the manipulation
- Usually it's better to know your research design inside out so that you can suspect any manipulation attempts.
 - We would suspect the direction of the manipulation. And typically, it's one-way manipulation. In cases where we might have both ways, theoretically they would cancel each other out.
- We could also observe partial manipulation in reality (e.g., when subjects can only imperfectly manipulate). But typically, as we treat it like fuzzy RD, we would not have identification problems. But complete manipulation would lead to serious identification issues.
- Remember: even in cases where we fail to reject the null hypothesis for the density test, we could not rule out completely that identification problem exists (just like any other hypotheses)
- Bunching happens when people self-select to a specific value in the range of a variable (e.g., key policy thresholds).
- Review paper (Kleven, 2016)
- **This test can only detect manipulation that changes the distribution of the running variable.** If you can choose the cutoff point or you have 2-sided manipulation, this test will fail to detect it.
- Histogram in bunching is similar to a density curve (we want narrower bins, wider bins bias elasticity estimates)
- We can also use bunching method to study individuals' or firm's responsiveness to changes in policy.
- Under RD, we assume that we don't have any manipulation in the running variable. However, bunching behavior is a manipulation by firms or individuals. Thus, violating this assumption.
 - Bunching can fix this problem by estimating what densities of individuals would have been without manipulation (i.e., manipulation-free counterfactual).
 - **The fraction of persons who manipulated** is then calculated by comparing the observed distribution to manipulation-free counterfactual distributions.

- Under RD, we do not need this step because the observed and manipulation-free counterfactual distributions are assumed to be the same. RD assume there is no manipulation (i.e., assume the manipulation-free counterfactual distribution)

When running variable and outcome variable are simultaneously determined, we can use a modified RDD estimator to have consistent estimate. (Bajari et al., 2011)

- **Assumptions:**

- Manipulation is **one-sided**: People move one way (i.e., either below the threshold to above the threshold or vice versa, but not to or away the threshold), which is similar to the monotonicity assumption under instrumental variable 28.3.1
- Manipulation is **bounded** (also known as regularity assumption): so that we can use people far away from this threshold to derive at our counterfactual distribution (Blomquist et al., 2017)

Steps:

1. Identify the window in which the running variable contains bunching behavior. We can do this step empirically based on data (Bosch et al., 2020). Additionally robustness test is needed (i.e., varying the manipulation window).
2. Estimate the manipulation-free counterfactual
3. Calculating the standard errors for inference can follow (Chetty et al., 2011) where we bootstrap resampling residuals in the estimation of the counts of individuals within bins (large data can render this step unnecessary).

If we pass the bunching test, we can move on to the Placebo Test

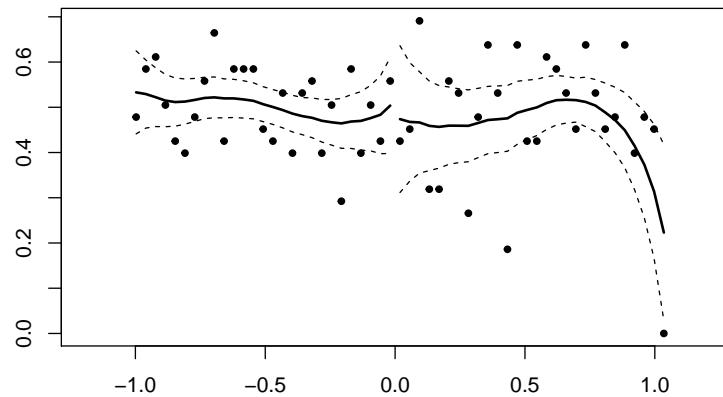
McCrary (2008) test

A jump in the density at the threshold (i.e., discontinuity) hold can serve as evidence for sorting around the cutoff point

```
library(rdd)

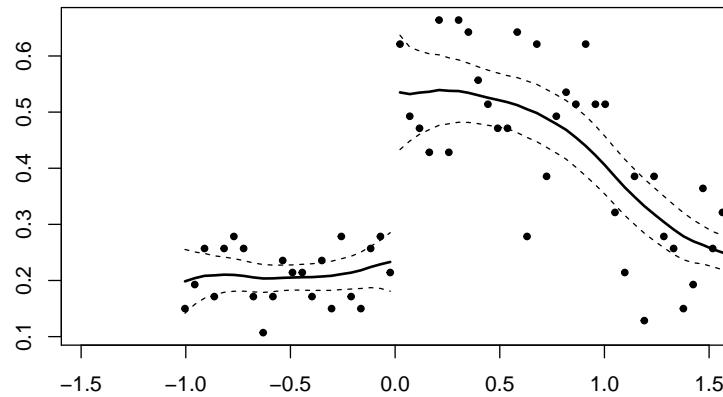
# you only need the running variable and the cutoff point

# Example by the package's authors
#No discontinuity
x<-runif(1000,-1,1)
DCdensity(x,0)
```



```
#> [1] 0.7540493
```

```
#Discontinuity
x<-runif(1000,-1,1)
x<-x+2*(runif(1000,-1,1)>0&x<0)
DCdensity(x,0)
```



```
#> [1] 0.0004090731
```

Cattaneo et al. (2019) test

```
library(rddensity)

# Example by the package's authors
# Continuous Density
set.seed(1)
x <- rnorm(2000, mean = -0.5)
rdd <- rddensity(X = x, vce = "jackknife")
summary(rdd)
#>
#> Manipulation testing using local polynomial density estimation.
#>
#> Number of obs = 2000
#> Model = unrestricted
#> Kernel = triangular
#> BW method = estimated
#> VCE method = jackknife
#>
#> c = 0           Left of c      Right of c
#> Number of obs 1376          624
#> Eff. Number of obs 354          345
#> Order est. (p) 2              2
#> Order bias (q) 3              3
#> BW est. (h)    0.514         0.609
#>
#> Method          T            P > |T|
#> Robust          -0.6798       0.4966
#>
#> P-values of binomial tests (H0: p=0.5).
#>
#> Window Length / 2      <c     >=c   P>|T|
#> 0.013                 11      9      0.8238
#> 0.026                 19      15     0.6076
#> 0.038                 29      21     0.3222
#> 0.051                 42      26     0.0681
#> 0.064                 44      33     0.2543
#> 0.077                 48      45     0.8358
#> 0.090                 55      51     0.7709
#> 0.102                 66      59     0.5917
#> 0.115                 74      67     0.6135
#> 0.128                 82      71     0.4189

# you have to specify your own plot (read package manual)
```

22.1.3 Placebo Tests

- Also known as Discontinuities in Average Outcomes at Other Values
- We should not see any jumps at other values (either $X_i < c$ or $X_i \geq c$)
 - Use the same bandwidth you use for the cutoff, and move it along the running variable: testing for a jump in the conditional mean of the outcome at the median of the running variable.
- Also known as falsification checks
- Before and after the cutoff point, we can run the placebo test to see whether X's are different).
- The placebo test is where you expect your coefficients to be not different from 0.
- This test can be used for
 - Testing no discontinuity in predetermined variables:
 - Testing other discontinuities
 - Inclusion and exclusion of covariates: RDD parameter estimates should not be sensitive to the inclusion or exclusion of other covariates.
- This is analogous to Experimental Design where we cannot only test whether the observables are similar in both treatment and control groups (if we reject this, then we don't have random assignment), but we cannot test unobservables.

Balance on observable characteristics on both sides

$$Z_i = \alpha_0 + \alpha_1 f(x_i) + [I(x_i \geq c)]\alpha_2 + [f(x_i) \times I(x_i \geq c)]\alpha_3 + u_i$$

where

- x_i is the running variable
- Z_i is other characteristics of people (e.g., age, etc)

Theoretically, Z_i should not be affected by treatment. Hence, $E(\alpha_2) = 0$

Moreover, when you have multiple Z_i , you typically have to simulate joint distribution (to avoid having significant coefficient based on chance).

The only way that you don't need to generate joint distribution is when all Z_i 's are independent (unlikely in reality).

Under RD, you shouldn't have to do any Matching Methods. Because just like when you have random assignment, there is no need to make balanced dataset before and after the cutoff. If you have to do balancing, then your RD assumptions are probably wrong in the first place.

22.1.4 Sensitivity to Bandwidth Choice

- Methods for bandwidth selection
 - Ad hoc or substantively driven
 - Data driven: cross validation
 - Conservative approach: (Cattaneo et al., 2020)
- The objective is to minimize the mean squared error between the estimated and actual treatment effects.
- Then, we need to see how sensitive our results will be dependent on the choice of bandwidth.
- In some cases, the best bandwidth for testing covariates may not be the best bandwidth for treating them, but it may be close.

```
# find optimal bandwidth by Imbens-Kalyanaraman
rdd::IKbandwidth(running_var, outcome_var, cutpoint = "", kernel = "triangular") # can also pick
```

22.1.5 Fuzzy RD Design

When you have cutoff that does not perfectly determine treatment, but creates a discontinuity in the likelihood of receiving the treatment, you need another instrument

For those that are close to the cutoff, we create an instrument for D_i

$$Z_i = \begin{cases} 1 & \text{if } X_i \geq c \\ 0 & \text{if } X_i < c \end{cases}$$

Then, we can estimate the effect of the treatment for compliers only (i.e., those treatment D_i depends on Z_i)

The LATE parameter

$$\lim_{c-\epsilon \leq X \leq c+\epsilon, \epsilon \rightarrow 0} \left(\frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)} \right)$$

equivalently, the canonical parameter:

$$\frac{\lim_{x \downarrow c} E(Y|X=x) - \lim_{x \uparrow c} E(Y|X=x)}{\lim_{x \downarrow c} E(D|X=x) - \lim_{x \uparrow c} E(D|X=x)}$$

Two equivalent ways to estimate

1. First
 1. Sharp RDD for Y
 2. Sharp RDD for D
 3. Take the estimate from step 1 divide by that of step 2
2. Second: Subset those observations that are close to c and run instrumental variable Z

22.1.6 Regression Kink Design

- If the slope of the treatment intensity changes at the cutoff (instead of the level of treatment assignment), we can have regression kink design
- Example: unemployment benefits

Sharp Kink RD parameter

$$\alpha_{KRD} = \frac{\lim_{x \downarrow c} \frac{d}{dx} E[Y_i|X_i=x] - \lim_{x \uparrow c} \frac{d}{dx} E[Y_i|X_i=x]}{\lim_{x \downarrow c} \frac{d}{dx} b(x) - \lim_{x \uparrow c} \frac{d}{dx} b(x)}$$

where $b(x)$ is a known function inducing “kink”

Fuzzy Kink RD parameter

$$\alpha_{KRD} = \frac{\lim_{x \downarrow c} \frac{d}{dx} E[Y_i|X_i=x] - \lim_{x \uparrow c} \frac{d}{dx} E[Y_i|X_i=x]}{\lim_{x \downarrow c} \frac{d}{dx} E[D_i|X_i=x] - \lim_{x \uparrow c} \frac{d}{dx} E[D_i|X_i=x]}$$

22.1.7 Mutli-cutoff, Multi-score, geographic RD

Multi-cutoff

$$E[Y_{1i} - Y_{0i}|X_i=x, C_i=c]$$

Multi-score (in multiple dimensions) (e.g., math and English cutoff for certain honor class):

$$E[Y_{1i} - Y_{0i}|X_{1i}=x_1, X_{2i}=x]$$

22.2 Steps for Sharp RD

1. Graph the data by computing the average value of the outcome variable over a set of bins (large enough to see a smooth graph, and small enough to make the jump around the cutoff clear).
2. Run regression on both sides of the cutoff to get the treatment effect
3. Robustness checks:
 1. Assess possible jumps in other variables around the cutoff
 2. Hypothesis testing for bunching
 3. Placebo tests
 4. Varying bandwidth

22.3 Steps for Fuzzy RD

1. Graph the data by computing the average value of the outcome variable over a set of bins (large enough to see a smooth graph, and small enough to make the jump around the cutoff clear).
2. Graph the probability of treatment
3. Estimate the treatment effect using 2SLS
4. Robustness checks:
 1. Assess possible jumps in other variables around the cutoff
 2. Hypothesis testing for bunching
 3. Placebo tests
 4. Varying bandwidth

22.4 Steps for RDiT (Regression Discontinuity in Time)

Notes:

- Additional assumption: Time-varying confounders change smoothly across the cutoff date
- Typically used in policy implementation in the same date for all subjects, but can also be used for cases where implementation dates are different between subjects. In the second case, researchers typically use different RDiT specification for each time series.

- Sometimes the date of implementation is not randomly assigned by chosen strategically. Hence, RDiT should be thought of as the “discontinuity at a threshold” interpretation of RD (not as “local randomization”). (Hausman and Rapson, 2018, p. 8)
- Normal RD uses variation in the N dimension, while RDiT uses variation in the T dimension
- Choose polynomials based on BIC typically. And can have either global polynomial or pre-period and post-period polynomial for each time series (but usually the global one will perform better)
- Could use **augmented local linear** outlined by (Hausman and Rapson, 2018, p. 12), where estimate the model with all the control first then take the residuals to include in the model with the RDiT treatment (remember to use bootstrapping method to account for the first-stage variance in the second stage).

Pros:

- can overcome cases where there is no cross-sectional variation in treatment implementation (dif-n-dif is not feasible)
 - There are papers that use both RDiT and DiD to (1) see the differential treatment effects across individuals/ space (Auffhammer and Kellogg, 2011) or (2) compare the 2 estimates where the control group’s validity is questionable (Gallego et al., 2013).
- Better than pre/post comparison because it can include flexible controls
- Better than event studies because it can use long-time horizons (may not be too relevant now since the development long-time horizon event studies), and it can use higher-order polynomials time control variables.

Cons:

- Taking observation for from the threshold (in time) can bias your estimates because of unobservables and time-series properties of the data generating process.
- (McCrary, 2008) test is not possible (see Sorting/Bunching/Manipulation) because when the density of the running (time) is uniform, you can’t use the test.
- Time-varying unobservables may impact the dependent variable discontinuously
- Error terms are likely to include persistence (serially correlated errors)
- Researchers cannot model time-varying treatment under RDiT

- In a small enough window, the local linear specification is fine, but the global polynomials can either be too big or too small (Hausman and Rapson, 2018)

Biases

- Time-Varying treatment Effects
 - increase sample size either by
 - * more granular data (greater frequency): will not increase power because of the problem of serial correlation
 - * increasing time window: increases bias from other confounders
 - 2 additional assumption:
 - * Model is correctly specified (with all confounders or global polynomial approximation)
 - * Treatment effect is correctly specified (whether it's smooth and constant, or varies)
 - * These 2 assumptions do not interact (we don't want them to interact - i.e., we don't want the polynomial correlated with the unobserved variation in the treatment effect)
 - There usually a difference between short-run and long-run treatment effects, but it's also possibly that the bias can stem from the overfitting problem of the polynomial specification. (Hausman and Rapson, 2018, p. 544)
- Autoregression (serial dependence)
 - Need to use **clustered standard errors** to account for serial dependence in the residuals
 - In the case of serial dependence in ϵ_{it} , we don't have a solution, including a lagged dependent variable would misspecify the model (probably find another research project)
 - In the case of serial dependence in y_{it} , with long window, it becomes fuzzy to what you try to recover. You can include the **lagged dependent variable** (bias can still come from the time-varying treatment or over-fitting of the global polynomial)
- Sorting and Anticipation Effects
 - Cannot run the (McCrory, 2008) because the density of the time running variable is uniform
 - Can still run tests to check discontinuities in other covariates (you want no discontinuities) and discontinuities in the outcome variable at other placebo thresholds (you don't want discontinuities)

- Hence, it's hard to argue for the causal effect here because it could be the total effect of the causal treatment and the unobserved sorting/anticipation/adaptation/avoidance effects. You can only argue that there is no such behavior

Recommendations for robustness check following (Hausman and Rapson, 2018, p. 549)

1. Plot the raw data and residuals (after removing confounders or trend). With varying polynomial and local linear controls, inconsistent results can be a sign of time-varying treatment effects.
2. Using global polynomial, you could overfit, then show polynomial with different order and alternative local linear bandwidths. If the results are consistent, you're ok
3. Placebo Tests: estimate another RD (1) on another location or subject (that did not receive the treatment) or (2) use another date.
4. Plot RD discontinuity on continuous controls
5. Donut RD to see if avoiding the selection close to the cutoff would yield better results (Barreca et al., 2011)
6. Test for auto-regression (using only pre-treatment data). If there is evidence for autoregression, include the lagged dpednetn variable
7. Augmented local linear (no need to use global polynomial and avoid overfitting)
 1. Use full sample to exclude the effect of important predictors
 2. Estimate the conditioned second stage on a smaller sample bandwidth

Examples from (Hausman and Rapson, 2018, p. 534) in

econ

- (Davis, 2008): Air quality
- (Auffhammer and Kellogg, 2011): Air quality
- (Chen and Whalley, 2012): Air quality
- (De Paola et al., 2012): car accidents
- (Gallego et al., 2013): air quality
- (Bento et al., 2014): Traffic
- (Anderson, 2014): Traffic
- (Burger et al., 2014): Car accidents

- (Brodeur et al., 2021): covid 19 lockdowns on well-being

marketing

- (Busse et al., 2006): Vehicle prices
- (Chen et al., 2009): Customer Satisfaction
- (Busse et al., 2010): Vehicle prices
- (Davis and Kahn, 2010): vehicle prices

22.5 Evaluation of an RD

- Evidence for (either formal tests or graphs)
 - Treatment and outcomes change discontinuously at the cutoff, while other variables and pre-treatment outcomes do not.
 - No manipulation of the assignment variable.
- Results are robust to various functional forms of the forcing variable
- Is there any other (unobserved) confound that could cause the discontinuous change at the cutoff (i.e., multiple forcing variables / bundling of institutions)?
- External Validity: How likely the result at the cutoff will generalize?

General Model

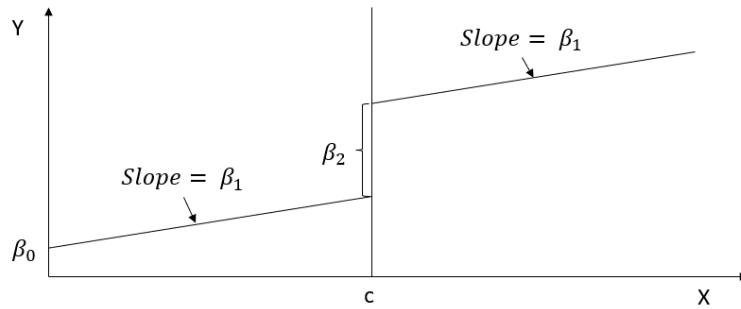
$$Y_i = \beta_0 + f(x_i)\beta_1 + [I(x_i \geq c)]\beta_2 + \epsilon_i$$

where $f(x_i)$ is any functional form of x_i

Simple case

When $f(x_i) = x_i$ (linear function)

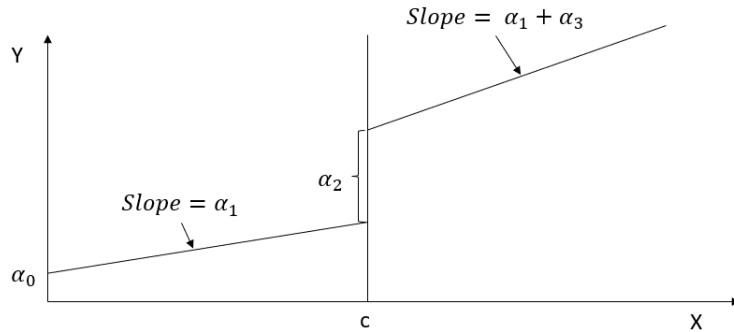
$$Y_i = \beta_0 + x_i\beta_1 + [I(x_i \geq c)]\beta_2 + \epsilon_i$$



RD gives you β_2 (causal effect) of X on Y at the cutoff point

In practice, everyone does

$$Y_i = \alpha_0 + f(x_i)\alpha_1 + [I(x_i \geq c)]\alpha_2 + [f(x_i) \times I(x_i \geq c)]\alpha_3 + u_i$$



where we estimate different slope on different sides of the line

and if you estimate α_3 to be no different from 0 then we return to the simple case

Notes:

- Sparse data can make α_3 large differential effect
- People are very skeptical when you have complex $f(x_i)$, usual simple function forms (e.g., linear, squared term, etc.) should be good. However, if you still insist, then **non-parametric estimation** can be your best bet.

Bandwidth of c (window)

- Closer to c can give you lower bias, but also efficiency
- Wider c can increase bias, but higher efficiency.
- Optimal bandwidth is very controversial, but usually we have to do it in the appendix for research article anyway.
- We can either
 - drop observations outside of bandwidth or
 - weight depends on how far and close to c

22.6 Applications

Examples in marketing:

- (Narayanan and Kalyanam, 2015)
- (Hartmann et al., 2011): nonparametric estimation and guide to identifying causal marketing mix effects

Packages in R (see (Thoemmes et al., 2016) for detailed comparisons): all can handle both sharp and fuzzy RD

- **rdd**
- **rdrobust**
- **rddensity** discontinuity in density tests (Sorting/Bunching/Manipulation) using local polynomials and binomial test
- **rdlocrand** covariate balance, binomial tests, window selection
- **rdmulti** multiple cutoffs and multiple scores
- **rdpower** power, sample selection
- **rddtools**

Package	rdd	rdrobust	rddtools
Coefficient estimator	Local linear regression	local polynomial regression	local polynomial regression

Package	rdd	rdrobust	rddtools
bandwidth selectors	(Imbens and Kalyanara- man, 2009)	(Calonico et al., 2014) (Imbens and Kalyanaraman, 2010) (Ludwig and Miller, 2007) (Calonico et al., 2019a,b)	(Imbens and Kalyanaraman, 2010)
Kernel functions	Epanechnikov Gaussian	Epanechnikov	Gaussian
	• Triangular		
	• Rectangular		
Bias Correction		Local polynomial regression	
Covariate options	Include	Include	Include Residuals
Assumptions testing	McCrary sorting		McCrary sorting Equality of covariates distribution and mean

based on table 1 (Thoemmes et al., 2016) (p. 347)

22.6.1 Example 1

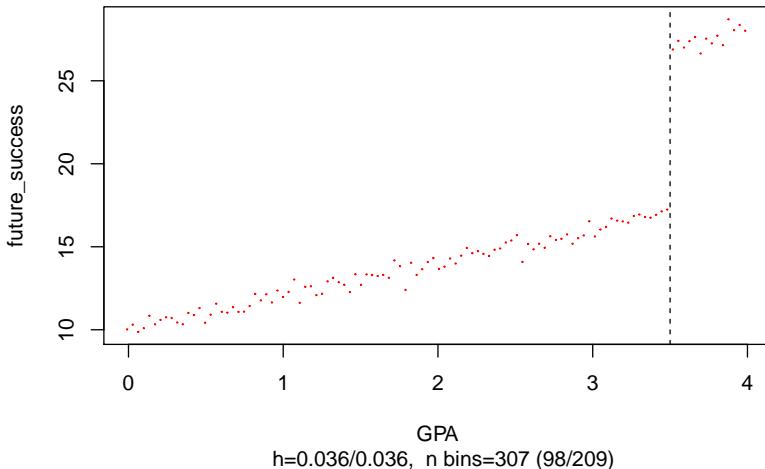
Example by Leihua Ye

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$$

$$X_i = \begin{cases} 1, & W_i \geq c \\ 0, & W_i < c \end{cases}$$

```
#cutoff point = 3.5
GPA <- runif(1000, 0, 4)
future_success <- 10 + 2 * GPA + 10 * (GPA >= 3.5) + rnorm(1000)
#install and load the package 'rddtools'
#install.packages("rddtools")
```

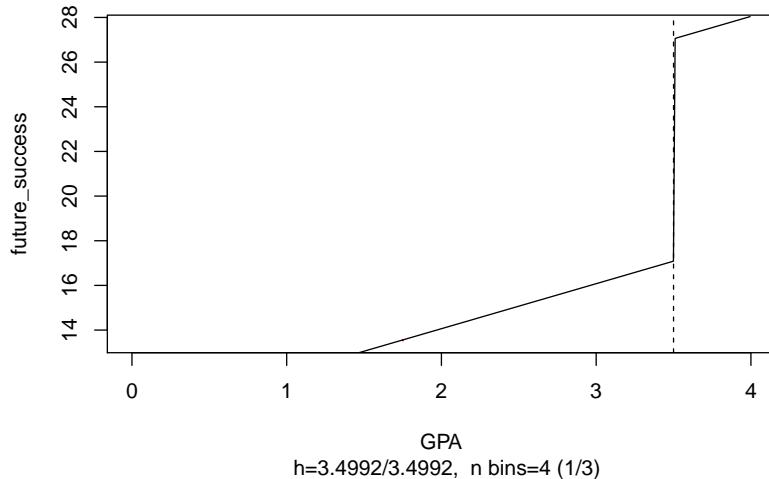
```
library(rddtools)
data <- rdd_data(future_success, GPA, cutpoint = 3.5)
# plot the dataset
plot(
  data,
  col = "red",
  cex = 0.1,
  xlab = "GPA",
  ylab = "future_success"
)
```



```
# estimate the sharp RDD model
rdd_mod <- rdd_reg_lm(rdd_object = data, slope = "same")
summary(rdd_mod)
#>
#> Call:
#> lm(formula = y ~ ., data = dat_step1, weights = weights)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -2.93235 -0.66786 -0.00799  0.69991  3.01768
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 17.08582    0.07178 238.03 <2e-16 ***
#> D           9.95513    0.11848   84.03 <2e-16 ***
```

```
#> x           2.01615   0.03546   56.85   <2e-16 ***  
#> ---  
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
#>  
#> Residual standard error: 1.046 on 997 degrees of freedom  
#> Multiple R-squared:  0.9617, Adjusted R-squared:  0.9616  
#> F-statistic: 1.253e+04 on 2 and 997 DF,  p-value: < 2.2e-16
```

```
# plot the RDD model along with binned observations  
plot(  
  rdd_mod,  
  cex = 0.1,  
  col = "red",  
  xlab = "GPA",  
  ylab = "future_success"  
)
```



22.6.2 Example 2

Bowblis and Smith (2019)

Occupational licensing can either increase or decrease market efficiency:

- More information means more efficiency
- Increased entry barriers (i.e., friction) increase efficiency

Components of RD

- Running variable
- Cutoff: 120 beds or above
- Treatment: you have to have the treatment before the cutoff point.

Under OLS

$$Y_i = \alpha_0 + X_i\alpha_1 + LW_i\alpha_2 + \epsilon_i$$

where

- LW_i Licensed/certified workers (in fraction format for each center).
- Y_i = Quality of service

Bias in α_2

- Mitigation-based: terrible quality can lead to more hiring, which negatively bias α_2
- Preference-based: places that have higher quality staff want to keep high quality staffs.

Under RD

$$Y_{ist} = \beta_0 + [I(Bed \geq 121)_{ist}] \beta_1 + f(Size_{ist}) \beta_2 + [f(Size_{ist}) \times I(Bed \geq 121)_{ist}] \beta_3 + X_{it}\delta + \gamma_s + \theta_t + \epsilon_{ist}$$

where

- s = state
- t = year
- i = hospital

This RD is fuzzy

- If right near the threshold (bandwidth), we have states with different sorting (i.e., non-random), then we need the fixed-effect for state s . But then your RD assumption wrong anyway, then you won't do it in the first place

- Technically, we could also run the fixed-effect regression, but because it's lower in the causal inference hierarchy. Hence, we don't do it.
- Moreover, in the RD framework, we don't include t before treatment (but in the FE we have to include before and after)
- If we include π_i for each hospital, then we don't have variation in the causal estimates (because hardly any hospital changes their bed size in the panel)
- When you have β_1 as the intent to treat (because the treatment effect does not coincide with the intent to treat)
- You cannot take those fuzzy cases out, because it will introduce the selection bias.
- Note that we cannot drop cases based on behavioral choice (because we will exclude non-compliers), but we can drop when we have particular behaviors ((e.g., people like round numbers)).

Thus, we have to use Instrument variable 28.3.1

Stage 1:

$$QSW_{ist} = \alpha_0 + [I(Bed \geq 121)_{ist}] \alpha_1 + f(Size_{ist}) \alpha_2 + [f(Size_{ist}) \times I(Bed \geq 121)_{ist}] \alpha_3 + X_{it} \delta + \gamma_s + \theta_t + \epsilon_{ist}$$

(Note: you should have different fixed effects and error term - $\delta, \gamma_s, \theta_t, \epsilon_{ist}$ from the first equation, but I ran out of Greek letters)

Stage 2:

$$Y_{ist} = \gamma_0 + \gamma_1 Q\hat{W}S_{ist} + f(Size_{ist}) \delta_2 + [f(Size_{ist}) \times I(Bed \geq 121)] \delta_3 + X_{it} \lambda + \eta_s + \tau_t + u_{ist}$$

- The bigger the jump (discontinuity), the more similar the 2 coefficients ($\gamma_1 \approx \beta_1$) where γ_1 is the average treatment effect (of exposing to the policy)
- β_1 will always be closer to 0 than γ_1
- Figure 1 shows bunching at every 5 units cutoff, but 120 is still out there.
- If we have manipulable bunching, there should be decrease at 130
- Since we have limited number of mass points (at the round numbers), we should cluster standard errors by the mass point

22.6.3 Example 3

Replication of (Carpenter and Dobkin, 2009) by Philipp Leppert, dataset from here

22.6.4 Example 4

For a detailed application, see (Thoemmes et al., 2016) where they use `rdd`, `rdrobust`, `rddtools`

Chapter 23

Difference-in-differences

Examples in marketing

- (Liaukonyte et al., 2015): TV ad on online shopping
- (Akca and Rao, 2020): aggregators for airlines business effect
- (Pattabhiramaiah et al., 2018): paywall affects readership
- (Wang et al., 2018): political ad source and message tone on vote shares and turnout using discontinuities in the level of political ads at the borders
- (Datta et al., 2018): streaming service on total music consumption using timing of users adoption of a music streaming service
- (Janakiraman et al., 2018): data breach announcement affect customer spending using timing of data breach and variation whether customer info was breached in that event
- (Lim et al., 2020): nutritional labels on nutritional quality for other brands in a category using variation in timing of adoption of nutritional labels across categories
- (Guo et al., 2020): payment disclosure laws effect on physician prescription behavior using Timing of the Massachusetts open payment law as the exogenous shock
- (Israeli, 2018): digital monitoring and enforcement on violations using enforcement of min ad price policies
- (Ramani and Srinivasan, 2019): firms respond to foreign direct investment liberalization using India's reform in 1991.
- (He et al., 2022): using Amazon policy change to examine the causal impact of fake reviews on sales, average ratings.
- (Peukert et al., 2022): using European General data protection Regulation, examine the impact of policy change on website usage.

Show the mechanism via

- Mediation analysis: see (Habel et al., 2021)

- Moderation analysis: see (Goldfarb and Tucker, 2011)

23.1 Simple Dif-n-dif

- A tool developed intuitively to study “natural experiment”, but its uses are much broader.
- Fixed Effects Estimator is the foundation for DID
- Why is dif-in-dif attractive? Identification strategy: Inter-temporal variation between groups
 - **Cross-sectional estimator** helps avoid omitted (unobserved) **common trends**
 - **Time-series estimator** helps overcome omitted (unobserved) **cross-sectional differences**

Consider

- $D_i = 1$ treatment group
- $D_i = 0$ control group
- $T = 1$ After the treatment
- $T = 0$ Before the treatment

	After ($T = 1$)	Before ($T = 0$)
Treated $D_i = 1$	$E[Y_{1i}(1) D_i = 1]$	$E[Y_{0i}(0) D_i = 1]$
Control $D_i = 0$	$E[Y_{0i}(1) D_i = 0]$	$E[Y_{0i}(0) D_i = 0]$

missing $E[Y_{0i}(1)|D = 1]$

The Average Treatment Effect on Treated (ATT)

$$E[Y_1(1) - Y_0(1)|D = 1] = \{E[Y(1)|D = 1] - E[Y(1)|D = 0]\} - \{E[Y(0)|D = 1] - E[Y(0)|D = 0]\}$$

More elaboration:

- For the treatment group, we isolate the difference between being treated and not being treated. If the untreated group would have been affected in a different way, the DiD design and estimate would tell us nothing.

Extension

More than 2 groups (multiple treatments and multiple controls), and more than 2 period (pre and post)

$$Y_{igt} = \alpha_g + \gamma_t + \beta I_{gt} + \delta X_{igt} + \epsilon_{igt}$$

where

- α_g is the group-specific fixed effect
- γ_t = time specific fixed effect
- β = dif-in-dif effect
- I_{gt} = interaction terms (n treatment indicators x n post-treatment dummies) (capture effect heterogeneity over time)

This specification is the “two-way fixed effects DiD” - **TWFE** (i.e., 2 sets of fixed effects: group + time).

- However, if you have Staggered Dif-n-dif (i.e., treatment is applied at different times to different groups). TWFE is really bad.

Matching Methods

- Match treatment and control based on pre-treatment observables
- Modify SEs appropriately (Heckman et al., 1997)

23.1.1 Assumptions

- **Parallel Trends:** Difference between the treatment and control groups remain constant if there were no treatment.
 - should be used in cases where
 - * you observe before and after an event
 - * you have treatment and control groups
 - not in cases where
 - * treatment is not random
 - * confounders.
 - To support we use
 - * Placebo test

* Prior trends test

- **Linear additive effects** (of group/unit specific and time-specific):
 - If they are not additively interact, we have to use the weighted 2FE estimator (Imai and Kim, 2020)
 - Typically seen in the Staggered Dif-n-dif
- No anticipation: There is no causal effect of the treatment before its implementation.

23.1.1.1 Prior Trends Test

1. Plot the average outcomes over time for both treatment and control group before and after the treatment in time.
2. Statistical test for difference in trends (using data from before the treatment period)

$$Y = \alpha_g + \beta_1 T + \beta_2 T \times G + \epsilon$$

where

- Y = the outcome variable
- α_g = group fixed effects
- T = time (e.g., specific year, or month)
- β_2 = different time trend for each group

Hence, if $\beta_2 = 0$ provides evidence that there are no difference in the trend for the two groups prior the time treatment.

You can also use different functional forms (e.g, polynomial or nonlinear).

If $\beta_2 \neq 0$ statistically, possible reasons can be:

- Statistical significance can be driven by large sample
- Or the trends are so consistent, and just one period deviation can throw off the trends. Hence, statistical statistical significance.

Technically, we can still salvage the research by including time fixed effects, instead of just the before-and-after time fixed effect (actually, most researchers do this mechanically anyways nowadays). However, a side effect can be that the time fixed effects can also absorb some part your treatment effect as well, especially in cases where the treatment effects vary with time (i.e., stronger or weaker over time) (Wolfers, 2003).

23.1.1.2 Placebo Test

Procedure:

1. Sample data only in the period before the treatment in time.
2. Consider different fake cutoff in time, either
 1. Try the whole sequence in time
 2. Generate random treatment period, and use **randomization inference** to account for sampling distribution of the fake effect.
3. Estimate the DiD model but with the post-time = 1 with the fake cutoff
4. A significant DiD coefficient means that you violate the parallel trends!
You have a big problem.

Alternatively,

- When data have multiple control groups, drop the treated group, and assign another control group as a “fake” treated group. But even if it fails (i.e., you find a significant DiD effect) among the control groups, it can still be fine. However, this method is used under Synthetic Control

Possible issues

- Estimate dependent on functional form:
 - When the size of the response depends (nonlinearly) on the size of the intervention, we might want to look at the difference in the group with high intensity vs. low.
- Long-term effects
 - Parallel trends are more likely to be observed over shorter period (window of observation)
- Heterogeneous effects
 - Different intensity (e.g., doses) for different groups.
- Ashenfelter dip
 - Participants are systemically different from nonparticipants before the treatment, leading to the question of permanent or transitory changes.
- Response to event might not be immediate (can't be observed right away in the dependent variable)

- Using lagged dependent variable Y_{it-1} might be more appropriate (Blundell and Bond, 1998)
- Other factors that affect the difference in trends between the two groups (i.e., treatment and control) will bias your estimation.
- Correlated observations within a group or time.
- Incidental parameters problems (Lancaster, 2000): it's always better to use individual and time fixed effect.
- When examining the effects of variation in treatment timing, we have to be careful because negative weights (per group) can be negative if there is a heterogeneity in the treatment effects over time. Example: [Athey and Imbens (2022)][Borusyak et al., 2021](Goodman-Bacon, 2021). In this case you should use new estimands proposed by [Callaway and Sant'Anna (2021)][de Chaisemartin and D'Haultfœuille, 2020], in the `did` package. If you expect lags and leads, see (Sun and Abraham, 2021)
- (Gibbons et al., 2018) caution when we suspect the treatment effect and treatment variance vary across groups

Robustness Check

- Placebo DiD (if the DiD estimate $\neq 0$, parallel trend is violated, and original DiD is biased):
 - Group: Use fake treatment groups: A population that was **not** affected by the treatment
 - Time: Redo the DiD analysis for period before the treatment (expected treatment effect is 0) (e.g., for previous year or period).
- Possible alternative control group: Expected results should be similar
- Try different windows (further away from the treatment point, other factors can creep in and nullify your effect).
- Treatment Reversal (what if we don't see the treatment event)
- Higher-order polynomial time trend (to relax linearity assumption)
- Test whether other dependent variables that should be affected by the event are indeed unaffected.
 - Use the same control and treatment period (DiD $\neq 0$, there is a problem)

23.1.2 Examples

23.1.2.1 Example from Princeton

```
library(foreign)
mydata = read.dta("http://dss.princeton.edu/training/Panel101.dta")
```

create a dummy variable to indicate the time when the treatment started

```
mydata$time = ifelse(mydata$year >= 1994, 1, 0)
```

create a dummy variable to identify the treatment group

```
mydata$treated = ifelse(mydata$country == "E" |
                        mydata$country == "F" | mydata$country == "G" ,
                        1,
                        0)
```

create an interaction between time and treated

```
mydata$did = mydata$time * mydata$treated
```

estimate the DID estimator

```
didreg = lm(y ~ treated + time + did, data = mydata)
summary(didreg)
#>
#> Call:
#> lm(formula = y ~ treated + time + did, data = mydata)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -9.768e+09 -1.623e+09  1.167e+08  1.393e+09  6.807e+09
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 3.581e+08 7.382e+08  0.485  0.6292
#> treated     1.776e+09 1.128e+09  1.575  0.1200
#> time        2.289e+09 9.530e+08  2.402  0.0191 *
#> did         -2.520e+09 1.456e+09 -1.731  0.0882 .
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#>
#> Residual standard error: 2.953e+09 on 66 degrees of freedom
#> Multiple R-squared:  0.08273,   Adjusted R-squared:  0.04104
#> F-statistic: 1.984 on 3 and 66 DF,  p-value: 0.1249
```

The `did` coefficient is the differences-in-differences estimator. Treat has a negative effect

23.1.2.2 Example by Card and Krueger (1993)

found that increase in minimum wage increases employment

Experimental Setting:

- New Jersey (treatment) increased minimum wage
- Penn (control) did not increase minimum wage

		After	Before	
Treatment	NJ	A	B	A - B
Control	PA	C	D	C - D
		A - C	B - D	(A - B) - (C - D)

where

- $A - B$ = treatment effect + effect of time (additive)
- $C - D$ = effect of time
- $(A - B) - (C - D)$ = dif-n-dif

The identifying assumptions:

- Can't have **switchers**
- PA is the control group
 - is a good counter factual
 - is what NJ would look like if they hadn't had the treatment

$$Y_{jt} = \beta_0 + NJ_j\beta_1 + POST_t\beta_2 + (NJ_j \times POST_t)\beta_3 + X_{jt}\beta_4 + \epsilon_{jt}$$

where

- j = restaurant
- NJ = dummy where 1 = NJ, and 0 = PA
- $POST$ = dummy where 1 = post, and 0 = pre

Notes:

- We don't need β_4 in our model to have unbiased β_3 , but including it would give our coefficients efficiency
- If we use ΔY_{jt} as the dependent variable, we don't need $POST_t \beta_2$ anymore
- Alternative model specification is that the authors use NJ high wage restaurant as control group (still choose those that are close to the border)
- The reason why they can't control for everything (PA + NJ high wage) is because it's hard to interpret the causal treatment
- Dif-n-dif utilizes similarity in pretrend of the dependent variables. However, this is neither a necessary nor sufficient for the identifying assumption.
 - It's not sufficient because they can have multiple treatments (technically, you could include more control, but your treatment can't interact)
 - It's not necessary because trends can be parallel after treatment
- However, we can't never be certain; we just try to find evidence consistent with our theory so that dif-n-dif can work.
- Notice that we don't need before treatment the **levels of the dependent variable** to be the same (e.g., same wage average in both NJ and PA), dif-n-dif only needs **pre-trend (i.e., slope)** to be the same for the two groups.

23.1.2.3 Example by Butcher et al. (2014)

Theory:

- Highest achieving students are usually in hard science. Why?
 - Hard to give students the benefit of doubt for hard science
 - How unpleasant and how easy to get a job. Degrees with lower market value typically want to make you feel more pleasant

Under OLS

$$E_{ij} = \beta_0 + X_i\beta_1 + G_j\beta_2 + \epsilon_{ij}$$

where

- X_i = student attributes
- β_2 = causal estimate (from grade change)
- E_{ij} = Did you choose to enroll in major j
- G_j = grade given in major j

Examine $\hat{\beta}_2$

- Negative bias: Endogenous response because department with lower enrollment rate will give better grade
- Positive bias: hard science is already having best students (i.e., ability), so if they don't their grades can be even lower

Under dif-n-dif

$$Y_{idt} = \beta_0 + POST_t\beta_1 + Treat_d\beta_2 + (POST_t \times Treat_d)\beta_3 + X_{idt} + \epsilon_{idt}$$

where

- Y_{idt} = grade average

	Intercept	Treat	Post	Treat*Post
Treat Pre	1	1	0	0
Treat Post	1	1	1	1
Control Pre	1	0	0	0
Control Post	1	0	1	0
	Average for pre-control β_0			

A more general specification of the dif-n-dif is that

$$Y_{idt} = \alpha_0 + (POST_t \times Treat_d)\alpha_1 + \theta_d + \delta_t + X_{idt} + u_{idt}$$

where

- $(\theta_d + \delta_t)$ richer, more df than $Treat_d\beta_2 + Post_t\beta_1$ (because fixed effects subsume Post and treat)
- α_1 should be equivalent to β_3 (if your model assumptions are correct)

Under causal inference, R^2 is not so important.

23.2 Multiple periods and variation in treatment timing

This is an extension of the DiD framework to settings where you have

- more than 2 time periods
- different treatment timing

When treatment effects are heterogeneous across time or units,

23.3 Staggered Dif-n-dif

- When subjects are treated at different point in time, we have to use staggered DiD
- For design where a treatment is applied and units are exposed to this treatment at all time afterward, see (Athey and Imbens, 2022)

Basic design (Stevenson and Wolfers, 2006)

$$Y_{it} = \alpha_i + \delta_t + \sum_k \gamma_k D_{ik} + \epsilon_{it}$$

where

- α_i are the country dummies
- δ_t are the time dummies
- $D_{ik} = 1$ for country i with treatment in all $k \geq t$ periods

Assumptions

- **Rollout Exogeneity:** if the treatment is randomly implemented over time (i.e., unrelated to variables that could also affect our dependent variables)

- **No confounding events**
- **Exclusion restrictions**
 - *No-anticipation assumption*: future treatment time do not affect current outcomes
 - *Invariance-to-history assumption*: the time a unit under treatment does not affect the outcome (i.e., the time exposed does not matter, just whether exposed or not). This presents causal effect of early or late adoption on the outcome.
- Auxiliary assumptions:
 - Contant treatment effects across units
 - Constat treatment effect over time
 - Random sampling
 - Effect Additivity

23.3.1 Example by Doleac and Hansen (2020)

- The purpose of banning a checking box for ex-criminal was banned because we thought that it gives more access to felons
- Even if we ban the box, employers wouldn't just change their behaviors. But then the unintended consequence is that employers statistically discriminate based on race

3 types of ban the box

1. Public employer only
2. Private employer with government contract
3. All employers

Main identification strategy

- If any county in the Metropolitan Statistical Area (MSA) adopts ban the box, it means the whole MSA is treated. Or if the state adopts “ban the ban,” every county is treated

Under Simple Dif-n-dif

$$Y_{it} = \beta_0 + \beta_1 Post_t + \beta_2 treat_i + \beta_3 (Post_t \times Treat_i) + \epsilon_{it}$$

But if there is no common post time, then we should use Staggered Dif-n-dif

$$E_{imrt} = \alpha + \beta_1 BTB_{imt} W_{imt} + \beta_2 BTB_{mt} + \beta_3 BTB_{mt} H_{imt} + \delta_m + D_{imt} \beta_5 + \lambda_{rt} + \delta_m \times f(t) \beta_7 + e_{imrt}$$

where

- i = person; m = MSA; r = region (US regions e.g., midwest) ; r = region; t = year
- W = White; B = Black; H = Hispanic
- $\beta_1 BTB_{imt} W_{imt} + \beta_2 BTB_{mt} + \beta_3 BTB_{mt} H_{imt}$ are the 3 dif-n-dif variables (BTB = “ban the box”)
- δ_m = dummy for MSI
- D_{imt} = control for people
- λ_{rt} = region by time fixed effect
- $\delta_m \times f(t)$ = linear time trend within MSA (but we should not need this if we have good pre-trend)

If we put $\lambda_r - \lambda_t$ (separately) we will get more broad fixed effect, while λ_{rt} will give us deeper and narrower fixed effect.

Before running this model, we have to drop all other races. And $\beta_1, \beta_2, \beta_3$ are not collinear because there are all interaction terms with BTB_{mt}

If we just want to estimate the model for black men, we will modify it to be

$$E_{imrt} = \alpha + BTB_{mt} \beta_1 + \delta_m + D_{imt} \beta_5 + \lambda_{rt} + (\delta_m \times f(t)) \beta_7 + e_{imrt}$$

$$E_{imrt} = \alpha + BTB_{m(t-3t)} \theta_1 + BTB_{m(t-2)} \theta_2 + BTB_{mt} \theta_4 + BTB_{m(t+1)} \theta_5 + BTB_{m(t+2)} \theta_6 + BTB_{m(t+3t)} \theta_7 + [\delta_m + D_{imt} \beta_5 + \lambda_r]$$

We have to leave $BTB_{m(t-1)} \theta_3$ out for the category would not be perfect collinearity

So the year before BTB ($\theta_1, \theta_2, \theta_3$) should be similar to each other (i.e., same pre-trend). Remember, we only run for places with BTB.

If θ_2 is statistically different from θ_3 (baseline), then there could be a problem, but it could also make sense if we have pre-trend announcement.

Example by Philipp Leppert replicating Card and Krueger (1994)

Example by Anthony Schmidt

23.4 Two-way Fixed-effects

A generalization of the dif-n-dif model is the two-way fixed-effects models where you have multiple groups and time effects.

This package is based on (Somaini and Wolak, 2016)

```
# dataset
library(bacondecomp)
df <- bacondecomp::castle

library(xtreg2way)
# output <- xtreg2way(y,
#                      data.frame(x1, x2),
#                      iid,
#                      tid,
#                      w,
#                      noise = "1",
#                      se = "1")

# equivalently
output <- xtreg2way(l_homicide ~ post,
                     df,
                     iid = df$state, # group id
                     tid = df$year, # time id
                     # w, # vector of weight
                     se = "1")
output$betaHat
#>      [,1]
#> l_homicide 0.08181162
output$aVarHat
#>      [,1]
#> [1,] 0.003751709

# to save time, you can use your structure in the last output for a new set of variables
# output2 <- xtreg2way(y, x1, struc=output$struc)
```

Standard errors estimation options

Set	Estimation
<code>se = "0"</code>	Assume homoskedasticity and no within group correlation or serial correlation
<code>se = "1"</code> (default)	robust to heteroskedasticity and serial correlation (Arellano et al., 1987)

Set	Estimation
<code>se = "2"</code>	robust to heteroskedasticity, but assumes no correlation within group or serial correlation
<code>se = "11"</code>	Aerllano SE with df correction performed by Stata xtreg (Somaini and Wolak, 2021)

Alternatively, you can also do it manually or with the `plm` package, but you have to be careful with how the SEs are estimated

```
library(multiwayvcov) # get vcov matrix
library(lmtest) # robust SEs estimation

# manual
output3 <- lm(l_homicide ~ post + factor(state) + factor(year),
               data = df)

# get variance-covariance matrix
vcov_tw <- multiwayvcov::cluster.vcov(output3,
                                         cbind(df$state, df$year),
                                         use_white = F,
                                         df_correction = F)

# get coefficients
coeftest(output3, vcov_tw)[2,]
#>   Estimate Std. Error t value Pr(>/t|)
#> 0.08181162 0.05671410 1.44252696 0.14979397

# using the plm package
library(plm)

output4 <- plm(l_homicide ~ post,
                data = df,
                index = c("state", "year"),
                model = "within",
                effect = "twoways")

# get coefficients
coeftest(output4, vcov = vcovHC, type = "HC1")
#>
#> t test of coefficients:
#>
#>   Estimate Std. Error t value Pr(>/t|)
#> post 0.081812 0.057748 1.4167 0.1572
```

As you can see, differences stem from SE estimation, not the coefficient estimate.

Chapter 24

Synthetic Control

Examples in marketing:

- (Tirunillai and Tellis, 2017): offline TV ad on Online Chatter
- (Guo et al., 2020): payment disclosure laws effect on physician prescription behavior using Timing of the Massachusetts open payment law as the exogenous shock
- (Wang et al., 2019): mobile hailing technology adoption on drivers' hourly earnings

Notes

- Synthetic control method (SCM) is a generalization of the Difference-in-differences model
- For a review of the method, see (Abadie, 2021)
- SCMs can also be used under the Bayesian framework where we do not have to impose any restrictive priori (Kim et al., 2020)
- Different from Matching Methods because SCMs match on the pre-treatment outcomes in each period while Matching Methods match on the number of covariates.
- A data driven procedure to construct more comparable control groups (i.e., black box).
- To do causal inference with control and treatment group using Matching Methods, you typically have to have similar covariates in the control and the treated groups. However, if you don't methods like Propensity Scores and DID can perform rather poorly (i.e., large bias).

Advantages over Difference-in-differences

1. Maximization of the observable similarity between control and treatment (maybe also unobservables)
2. Can also be used in cases where no untreated case with similar on matching dimensions with treated cases
3. Objective selection of controls.

Advantages over linear regression

- Regression weights for the estimator will be outside of [0,1] (because regression allows extrapolation), and it will not be sparse (i.e., can be less than 0).
- No extrapolation under SCMs
- Explicitly state the fit (i.e., the weight)
- Can be estimated without the post-treatment outcomes for the control group (can't p-hack)

Advantages:

1. From the selection criteria, researchers can understand the relative importance of each candidate
2. Post-intervention outcomes are not used in synthetic. Hence, you can't retro-fit.
3. Observable similarity between control and treatment cases is maximized

Disadvantages:

1. It's hard to argue for the weights you use to create the "synthetic control"

SCM is recommended when

1. Social events to evaluate large-scale program or policy
2. Only one treated case with several control candidates.

Assumptions

- Donor subject is a good match for the synthetic control (i.e., gap between the dependent of the donor subject and that of the synthetic control should be 0 before treatment)

- Only the treated subject undergoes the treatment and not any of the subjects in the donor pool.
- No other changes to the subjects during the whole window.
- The counterfactual outcome of the treatment group can be imputed in a **linear combination** of control groups.

Identification: The exclusion restriction is met conditional on the pre-treatment outcomes.

Synth provides an algorithm that finds weighted combination of the comparison units where the weights are chosen such that it best resembles the values of predictors of the outcome variable for the affected units before the intervention

Setting (notation followed professor Alberto Abadie)

- $J + 1$ units in periods $1, \dots, T$
- The first unit is the treated one during $T_0 + 1, \dots, T$
- J units are called a donor pool
- Y_{it}^I is the outcome for unit i if it's exposed to the treatment during $T_0 + 1, \dots, T$
- Y_{it}^N is the outcome for unit i if it's not exposed to the treatment

We try to estimate the effect of treatment on the treated unit

$$\tau_{1t} = Y_{1t}^I - Y_{1t}^N$$

where we observe the first treated unit already $Y_{1t}^I = Y_{1t}$

To construct the synthetic control unit, we have to find appropriate weight for each donor in the donor pool by finding $\mathbf{W} = (w_2, \dots, w_{J+1})'$ where

- $w_j \geq 0$ for $j = 2, \dots, J + 1$
- $w_2 + \dots + w_{J+1} = 1$

The “appropriate” vector \mathbf{W} here is constrained to

$$\min ||\mathbf{X}_1 - \mathbf{X}_0 \mathbf{W}||$$

where

- \mathbf{X}_1 is the $k \times 1$ vector of pre-treatment characteristics for the treated unit

- \mathbf{X}_0 is the $k \times J$ matrix of pre-treatment characteristics for the untreated units

For simplicity, researchers usually use

$$\min ||\mathbf{X}_1 - \mathbf{X}_0 \mathbf{W}|| = (\sum_{h=1}^k v_h (X_{h1} - w_2 X - h2 - \dots - w_{J+1} X_{hJ+1}))^{1/2}$$

where

- v_1, \dots, v_k is a vector positive constants that represent the predictive power of the k predictors on Y_{1t}^N (i.e., the potential outcome of the treated without treatment) and it can be chosen either explicitly by the researcher or by data-driven methods

For penalized synthetic control (Abadie and L'Hour, 2021), the minimization problem becomes

$$\min_{\mathbf{W}} ||\mathbf{X}_1 - \sum_{j=2}^{J+1} W_j \mathbf{X}_j||^2 + \lambda \sum_{j=2}^{J+1} W_j ||\mathbf{X}_1 - \mathbf{X}_j||^2$$

where

- $W_j \geq 0$ and $\sum_{j=2}^{J+1} W_j = 1$
- $\lambda > 0$ balances over-fitting of the treated and minimize the sum of pairwise distances
 - $\lambda \rightarrow 0$: pure synthetic control (i.e solution for the unpenalized estimator)
 - $\lambda \rightarrow \infty$: nearest neighbor matching

Advantages:

- For $\lambda > 0$, you have unique and sparse solution
- Reduces the interpolation bias when averaging dissimilar units
- Penalized SC never uses dissimilar units

Then the synthetic control estimator is

$$\hat{\tau}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}$$

where Y_{jt} is the outcome for unit j at time t

Consideration

Under the factor model (Abadie et al., 2010)

$$Y_{it}^N = {}_t\mathbf{Z}_i + {}_t\mu_i + \epsilon_{it}$$

where

- Z_i = observables
- μ_i = unobservables
- ϵ_{it} = unit-level transitory shock (i.e., random noise)

with assumptions of \mathbf{W}^* such that

$$\begin{aligned} \sum_{j=2}^{J+1} w_j^* \mathbf{Z}_j &= \mathbf{Z}_1 \\ &\dots \\ \sum_{j=2}^{J+1} w_j^* Y_{j1} &= Y_{11} \\ \sum_{j=2}^{J+1} w_j^* Y_{jT_0} &= Y_{1T_0} \end{aligned}$$

Basically, we assume that the synthetic control is a good counterfactual when the treated unit is not exposed to the treatment.

Then,

- the bias bound depends on close fit, which is controlled by the ratio between ϵ_{it} (transitory shock) and T_0 (the number of pre-treatment periods). In other words, you should have good fit for Y_{1t} for pre-treatment period (i.e., T_0 should be large while small variance in ϵ_{it})
- When you have poor fit, you have to use bias correction version of the synthetic control. See [Arkhangelsky et al. (2019); Abadie and L'Hour (2021)](Ben-Michael et al., 2020)

- Overfitting can be the result of small T_0 (the number of pre-treatment periods), large J (the number of units in the donor pool), and large ϵ_{it} (noise)
 - Mitigation: put only similar units (to the treated one) in the donor pool

To make inference, we have to create a permutation distribution (by iteratively reassigned the treatment to the units in the donor pool and estimate the placebo effects in each iteration). We say there is an effect of the treatment when the magnitude of value of the treatment effect on the treated unit is extreme relative to the permutation distribution.

It's recommended to use one-sided inference. And the permutation distribution is superior to the p-values alone (because sampling-based inference is hard under SCMs either because of undefined sampling mechanism or the sample is the population).

For benchmark (permutation) distribution (e.g., uniform), see (Firpo and Possebom, 2018)

Applications

24.0.1 Example 1

by Danilo Freire

```
# install.packages("Synth")
# install.packages("gsynth")
library("Synth")
library("gsynth")
```

simulate data for 10 states and 30 years. State A receives the treatment $T = 20$ after year 15.

```
set.seed(1)
year      <- rep(1:30, 10)
state     <- rep(LETTERS[1:10], each = 30)
X1        <- round(rnorm(300, mean = 2, sd = 1), 2)
X2        <- round(rbinom(300, 1, 0.5) + rnorm(300), 2)
Y         <- round(1 + 2 * X1 + rnorm(300), 2)
df        <- as.data.frame(cbind(Y, X1, X2, state, year))
df$Y      <- as.numeric(as.character(df$Y))
df$X1    <- as.numeric(as.character(df$X1))
df$X2    <- as.numeric(as.character(df$X2))
df$year   <- as.numeric(as.character(df$year))
```

```

df$state.num <- rep(1:10, each = 30)
df$state      <- as.character(df$state)
df$`T`        <- ifelse(df$state == "A" & df$year >= 15, 1, 0)
df$Y          <- ifelse(df$state == "A" & df$year >= 15, df$Y + 20, df$Y)

str(df)
#> 'data.frame':   300 obs. of  7 variables:
#> $ Y       : num  2.29 4.51 2.07 8.87 4.37 1.32 8 7.49 6.98 3.72 ...
#> $ X1      : num  1.37 2.18 1.16 3.6 2.33 1.18 2.49 2.74 2.58 1.69 ...
#> $ X2      : num  1.96 0.4 -0.75 -0.56 -0.45 1.06 0.51 -2.1 0 0.54 ...
#> $ state    : chr  "A" "A" "A" "A" ...
#> $ year     : num  1 2 3 4 5 6 7 8 9 10 ...
#> $ state.num: int  1 1 1 1 1 1 1 1 1 ...
#> $ T        : num  0 0 0 0 0 0 0 0 0 0 ...

dataprep.out <-
  dataprep(
    df,
    predictors = c("X1", "X2"),
    dependent   = "Y",
    unit.variable = "state.num",
    time.variable = "year",
    unit.names.variable = "state",
    treatment.identifier = 1,
    controls.identifier = c(2:10),
    time.predictors.prior = c(1:14),
    time.optimize.ssr = c(1:14),
    time.plot         = c(1:30)
  )

synth.out <- synth(dataprep.out)
#>
#> X1, X0, Z1, Z0 all come directly from dataprep object.
#>
#>
#> ****
#> searching for synthetic control unit
#>
#>
#> ****
#> ****
#> ****
#>
#> MSPE (LOSS V): 9.831789

```

```

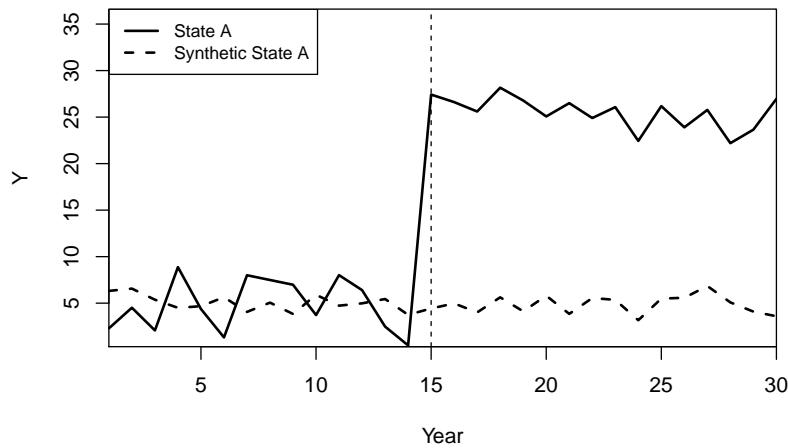
#>
#> solution.v:
#> 0.3888387 0.6111613
#>
#> solution.w:
#> 0.1115941 0.1832781 0.1027237 0.312091 0.06096758 0.03509706 0.05893735 0.05746256

print(synth.tables <- synth.tab(
  dataprep.res = dataprep.out,
  synth.res    = synth.out)
)
#> $tab.pred
#>   Treated Synthetic Sample Mean
#> X1  2.028      2.028      2.017
#> X2  0.513      0.513      0.394
#>
#> $tab.v
#>   v.weights
#> X1 0.389
#> X2 0.611
#>
#> $tab.w
#>   w.weights unit.names unit.numbers
#> 2    0.112      B      2
#> 3    0.183      C      3
#> 4    0.103      D      4
#> 5    0.312      E      5
#> 6    0.061      F      6
#> 7    0.035      G      7
#> 8    0.059      H      8
#> 9    0.057      I      9
#> 10   0.078      J     10
#>
#> $tab.loss
#>   Loss W  Loss V
#> [1,] 9.761708e-12 9.831789

path.plot(synth.res    = synth.out,
          dataprep.res = dataprep.out,
          Ylab         = c("Y"),
          Xlab         = c("Year"),
          Legend       = c("State A", "Synthetic State A"),
          Legend.position = c("topleft")
)

```

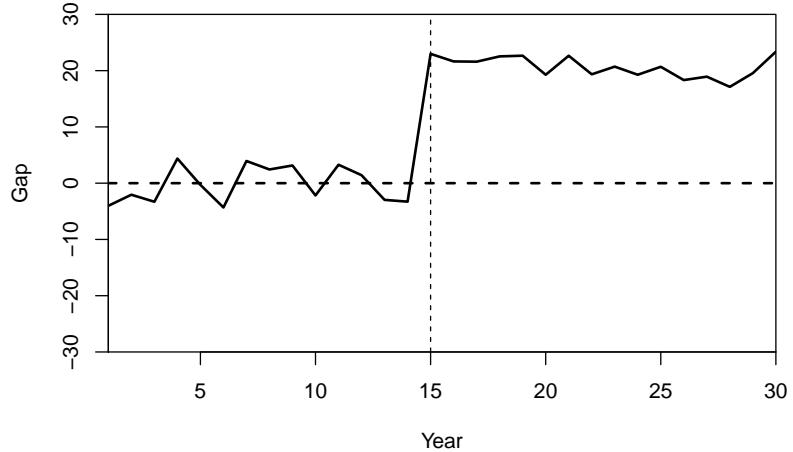
```
abline(v = 15,
       lty = 2)
```



Gaps plot:

```
gaps.plot(synth.res    = synth.out,
           dataprep.res = dataprep.out,
           Ylab         = c("Gap"),
           Xlab         = c("Year"),
           Ylim         = c(-30, 30),
           Main         = "")
)

abline(v = 15,
       lty = 2)
```



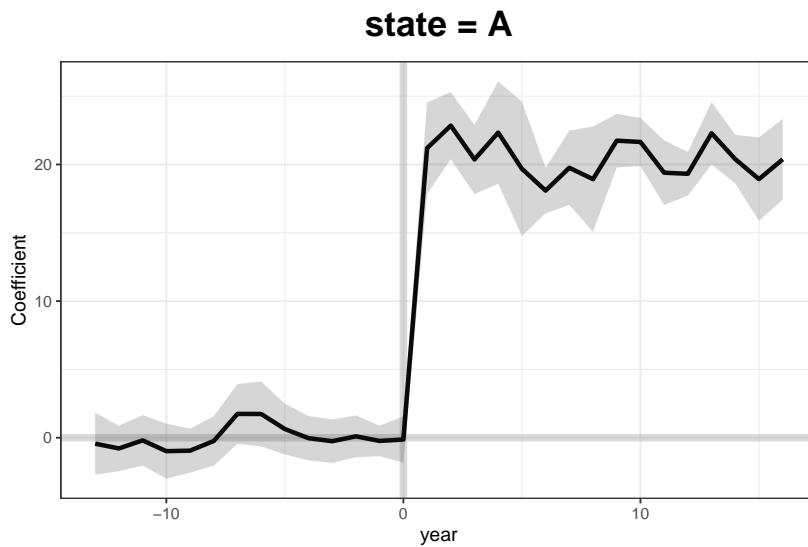
Alternatively, `gsynth` provides options to estimate iterative fixed effects, and handle multiple treated units at tat time.

Here, we use two-way fixed effects and bootstrapped standard errors

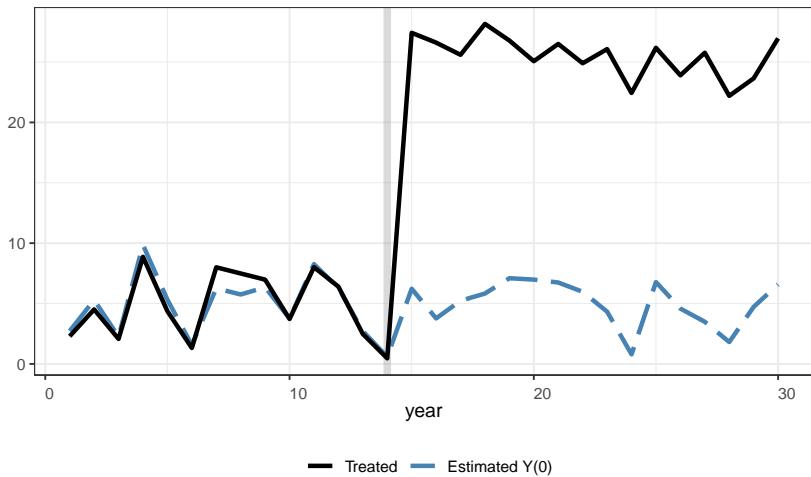
```
gsynth.out <- gsynth(
  Y ~ `T` + X1 + X2,
  data = df,
  index = c("state", "year"),
  force = "two-way",
  CV = TRUE,
  r = c(0, 5),
  se = TRUE,
  inference = "parametric",
  nboots = 1000,
  parallel = F # TRUE
)
#> Cross-validating ...
#>   r = 0; sigma2 = 1.13533; IC = 0.95632; PC = 0.96713; MSPE = 1.65502
#>   r = 1; sigma2 = 0.96885; IC = 1.54420; PC = 4.30644; MSPE = 1.33375
#>   r = 2; sigma2 = 0.81855; IC = 2.08062; PC = 6.58556; MSPE = 1.27341*
#>   r = 3; sigma2 = 0.71670; IC = 2.61125; PC = 8.35187; MSPE = 1.79319
#>   r = 4; sigma2 = 0.62823; IC = 3.10156; PC = 9.59221; MSPE = 2.02301
#>   r = 5; sigma2 = 0.55497; IC = 3.55814; PC = 10.48406; MSPE = 2.79596
#>
#>   r* = 2
#>
```

```
#>  
Simulating errors .....  
Bootstrapping ...  
#> .....
```

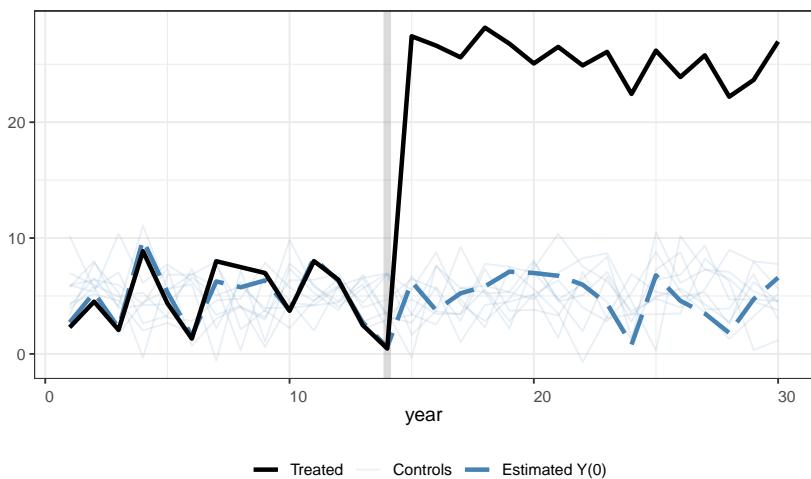
```
plot(gsynth.out)
```



```
plot(gsynth.out, type = "counterfactual")
```

Treated and Counterfactual (A)

```
plot(gsynth.out, type = "counterfactual", raw = "all") # shows estimations for the controls
```

Treated and Counterfactual (A)

24.0.2 Example 2

by Leihua Ye

```

library(Synth)
data("basque")
dim(basque) #774*17
#> [1] 774 17
head(basque)
#>   regionno    regionname year   gdpcap sec.agriculture sec.energy sec.industry
#> 1       1 Spain (Espana) 1955 2.354542          NA          NA          NA
#> 2       1 Spain (Espana) 1956 2.480149          NA          NA          NA
#> 3       1 Spain (Espana) 1957 2.603613          NA          NA          NA
#> 4       1 Spain (Espana) 1958 2.637104          NA          NA          NA
#> 5       1 Spain (Espana) 1959 2.669880          NA          NA          NA
#> 6       1 Spain (Espana) 1960 2.869966          NA          NA          NA
#>   sec.construction sec.services.venta sec.services.nonventa school.illit
#> 1           NA           NA           NA          NA          NA
#> 2           NA           NA           NA          NA          NA
#> 3           NA           NA           NA          NA          NA
#> 4           NA           NA           NA          NA          NA
#> 5           NA           NA           NA          NA          NA
#> 6           NA           NA           NA          NA          NA
#>   school.prim school.med school.high school.post.high popdens invest
#> 1           NA           NA           NA          NA          NA          NA
#> 2           NA           NA           NA          NA          NA          NA
#> 3           NA           NA           NA          NA          NA          NA
#> 4           NA           NA           NA          NA          NA          NA
#> 5           NA           NA           NA          NA          NA          NA
#> 6           NA           NA           NA          NA          NA          NA

```

transform data to be used in `synth()`

```

dataprep.out <- dataprep(
  foo = basque,
  predictors = c(
    "school.illit",
    "school.prim",
    "school.med",
    "school.high",
    "school.post.high",
    "invest"
  ),
  predictors.op = "mean",
  # the operator
  time.predictors.prior = 1964:1969,
  #the entire time frame from the #beginning to the end
  special.predictors = list(
    list("gdpcap", 1960:1969, "mean"),

```

```

list("sec.agriculture", seq(1961, 1969, 2), "mean"),
list("sec.energy", seq(1961, 1969, 2), "mean"),
list("sec.industry", seq(1961, 1969, 2), "mean"),
list("sec.construction", seq(1961, 1969, 2), "mean"),
list("sec.services.venta", seq(1961, 1969, 2), "mean"),
list("sec.services.nonventa", seq(1961, 1969, 2), "mean"),
list("popdens", 1969, "mean")
),
dependent = "gdpcap",
# dv
unit.variable = "regionno",
#identifying unit numbers
unit.names.variable = "regionname",
#identifying unit names
time.variable = "year",
#time-periods
treatment.identifier = 17,
#the treated case
controls.identifier = c(2:16, 18),
#the control cases; all others #except number 17
time.optimize.ssr = 1960:1969,
#the time-period over which to optimize
time.plot = 1955:1997
)#the entire time period before/after the treatment

```

where

- X1 = the control case before the treatment
- X0 = the control cases after the treatment
- Z1: the treatment case before the treatment
- Z0: the treatment case after the treatment

```

synth.out = synth(data.prep.obj = dataprep.out, method = "BFGS")
#>
#> X1, X0, Z1, Z0 all come directly from dataprep object.
#>
#>
#> ****
#> searching for synthetic control unit
#>
#>
#> ****

```

```
#> ****
#> ****
#>
#> MSPE (LOSS V): 0.008864606
#>
#> solution.v:
#> 0.02773094 1.194e-07 1.60609e-05 0.0007163836 1.486e-07 0.002423908 0.0587055 0.2651997 0.028
#>
#> solution.w:
#> 2.53e-08 4.63e-08 6.44e-08 2.81e-08 3.37e-08 4.844e-07 4.2e-08 4.69e-08 0.8508145 9.75e-08 3.
```

Calculate the difference between the real basque region and the synthetic control

```
gaps = dataprep.out$Y1plot - (dataprep.out$Y0plot
                                %*% synth.out$solution.w)
gaps[1:3,1]
#>      1955      1956      1957
#> 0.15023473 0.09168035 0.03716475

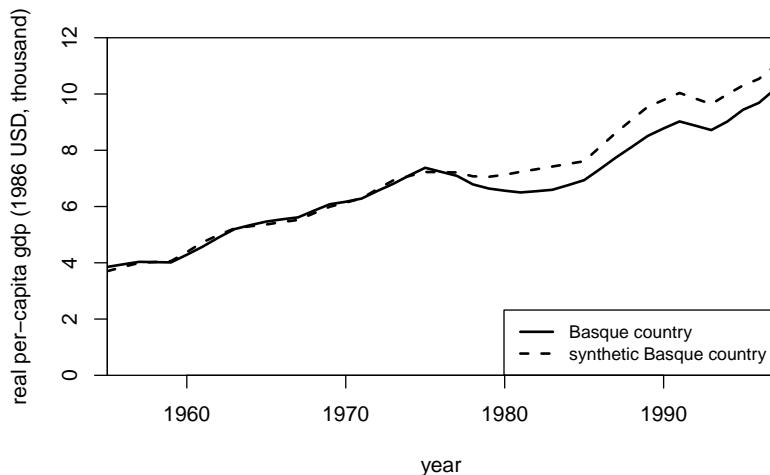
synth.tables = synth.tab(dataprep.res = dataprep.out,
                        synth.res = synth.out)
names(synth.tables)
#> [1] "tab.pred"    "tab.v"       "tab.w"       "tab.loss"
synth.tables$tab.pred[1:13,]
#>                                              Treated Synthetic Sample Mean
#> school.illit                         39.888   256.337   170.786
#> school.prim                          1031.742  2730.104  1127.186
#> school.med                           90.359   223.340   76.260
#> school.high                           25.728   63.437   24.235
#> school.post.high                     13.480   36.153   13.478
#> invest                               24.647   21.583   21.424
#> special.gdpcap.1960.1969            5.285    5.271    3.581
#> special.sec.agriculture.1961.1969  6.844    6.179   21.353
#> special.sec.energy.1961.1969        4.106    2.760    5.310
#> special.sec.industry.1961.1969     45.082   37.636   22.425
#> special.sec.construction.1961.1969  6.150    6.952    7.276
#> special.sec.services.venta.1961.1969 33.754   41.104   36.528
#> special.sec.services.nonventa.1961.1969 4.072    5.371    7.111
```

Relative importance of each unit

```
synth.tables$tab.w[8:14, ]
#> w.weights          unit.names unit.numbers
#> 9      0.000  Castilla-La Mancha         9
```

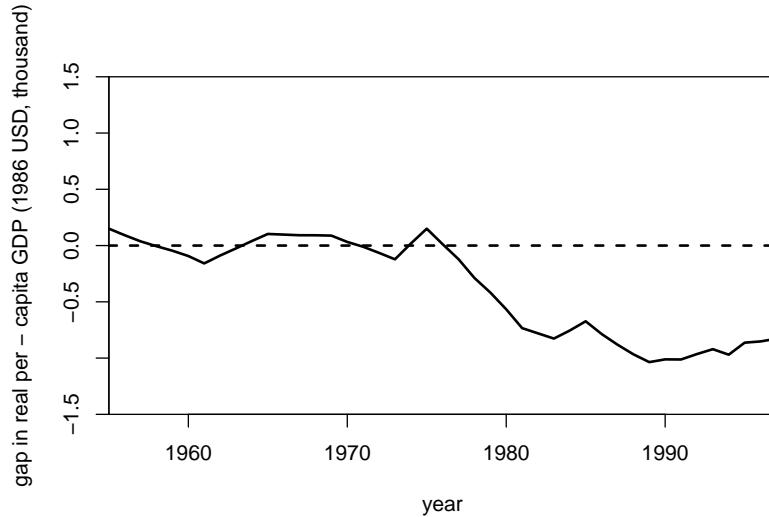
```
#> 10      0.851          Cataluna        10
#> 11      0.000  Comunidad Valenciana   11
#> 12      0.000          Extremadura    12
#> 13      0.000          Galicia       13
#> 14      0.149 Madrid (Comunidad De) 14
#> 15      0.000  Murcia (Region de)   15
```

```
# plot the changes before and after the treatment
path.plot(
  synth.res = synth.out,
  dataprep.res = dataprep.out,
  Ylab = "real per-capita gdp (1986 USD, thousand)",
  Xlab = "year",
  Ylim = c(0, 12),
  Legend = c("Basque country",
            "synthetic Basque country"),
  Legend.position = "bottomright"
)
```



```
gaps.plot(
  synth.res = synth.out,
  dataprep.res = dataprep.out,
  Ylab = "gap in real per - capita GDP (1986 USD, thousand)",
  Xlab = "year",
  Ylim = c(-1.5, 1.5),
```

```
Main = NA
)
```



Doubly Robust Difference-in-Differences

Example from DRDID package

```
library(DRDID)
data(nsw_long)
# Form the Lalonde sample with CPS comparison group
eval_lalonde_cps <- subset(nsw_long, nsw_long$treated == 0 | nsw_long$sample == 2)
```

Estimate Average Treatment Effect on Treated using Improved Locally Efficient Doubly Robust DID estimator

```
out <-
drdid(
  yname = "re",
  tname = "year",
  idname = "id",
  dname = "experimental",
  xformula = ~ age + educ + black + married + nodegree + hisp + re74,
  data = eval_lalonde_cps,
  panel = TRUE
)
summary(out)
```

```
#> Call:
#> drdid(yname = "re", tname = "year", idname = "id", dname = "experimental",
#>         xformula = ~age + educ + black + married + nodegree + hisp +
#>                     re74, data = eval_lalonde_cps, panel = TRUE)
#> -----
#> Further improved locally efficient DR DID estimator for the ATT:
#>
#>   ATT      Std. Error   t value   Pr(>|t|) [95% Conf. Interval]
#> -901.2703  393.6247   -2.2897    0.022   -1672.7747 -129.766
#> -----
#> Estimator based on panel data.
#> Outcome regression est. method: weighted least squares.
#> Propensity score est. method: inverse prob. tilting.
#> Analytical standard error.
#> -----
#> See Sant'Anna and Zhao (2020) for details.
```

24.0.3 Example 3

by Synth package's authors

```
library(Synth)
data("basque")
```

`synth()` requires

- X_1 vector of treatment predictors
- X_0 matrix of same variables for control group
- Z_1 vector of outcome variable for treatment group
- Z_0 matrix of outcome variable for control group

use `dataprep()` to prepare data in the format that can be used throughout the Synth package

```
dataprep.out <- dataprep(
  foo = basque,
  predictors = c(
    "school.illit",
    "school.prim",
    "school.med",
    "school.high",
```

```

    "school.post.high",
    "invest"
),
predictors.op = "mean",
time.predictors.prior = 1964:1969,
special.predictors = list(
  list("gdpcap", 1960:1969 , "mean"),
  list("sec.agriculture", seq(1961, 1969, 2), "mean"),
  list("sec.energy", seq(1961, 1969, 2), "mean"),
  list("sec.industry", seq(1961, 1969, 2), "mean"),
  list("sec.construction", seq(1961, 1969, 2), "mean"),
  list("sec.services.venta", seq(1961, 1969, 2), "mean"),
  list("sec.services.nonventa", seq(1961, 1969, 2), "mean"),
  list("popdens", 1969, "mean")
),
dependent = "gdpcap",
unit.variable = "regionno",
unit.names.variable = "regionname",
time.variable = "year",
treatment.identifier = 17,
controls.identifier = c(2:16, 18),
time.optimize.ssr = 1960:1969,
time.plot = 1955:1997
)

```

find optimal weights that identifies the synthetic control for the treatment group

```

synth.out <- synth(data.prep.obj = dataprep.out, method = "BFGS")
#>
#> X1, X0, Z1, Z0 all come directly from dataprep object.
#>
#> ****
#> searching for synthetic control unit
#>
#> ****
#> ****
#> ****
#> MSPE (LOSS V): 0.008864606
#>
#> solution.v:
#> 0.02773094 1.194e-07 1.60609e-05 0.0007163836 1.486e-07 0.002423908 0.0587055 0.2651997 0.028
#>

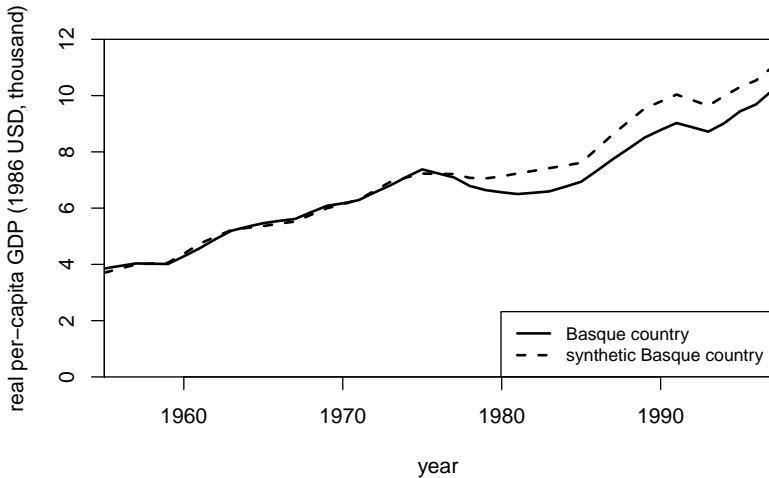
```

```
#> solution.w:
#> 2.53e-08 4.63e-08 6.44e-08 2.81e-08 3.37e-08 4.844e-07 4.2e-08 4.69e-08 0.8508145

gaps <- dataprep.out$Y1plot - (dataprep.out$Y0plot %*% synth.out$solution.w)
gaps[1:3, 1]
#> 1955      1956      1957
#> 0.15023473 0.09168035 0.03716475

synth.tables <-
  synth.tab(dataprep.res = dataprep.out, synth.res = synth.out)
names(synth.tables) # you can pick tables to see
#> [1] "tab.pred" "tab.v"    "tab.w"    "tab.loss"

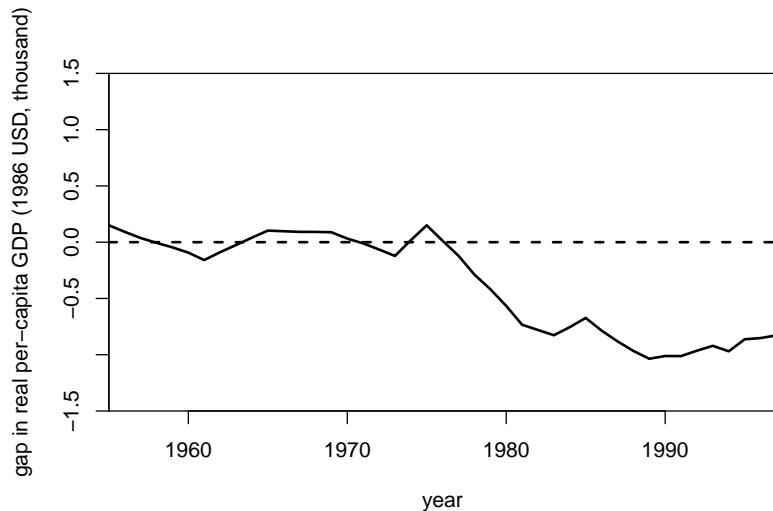
path.plot(
  synth.res = synth.out,
  dataprep.res = dataprep.out,
  Ylab = "real per-capita GDP (1986 USD, thousand)",
  Xlab = "year",
  Ylim = c(0, 12),
  Legend = c("Basque country",
            "synthetic Basque country"),
  Legend.position = "bottomright"
)
```



```

gaps.plot(
  synth.res = synth.out,
  dataprep.res = dataprep.out,
  Ylab = "gap in real per-capita GDP (1986 USD, thousand)",
  Xlab = "year",
  Ylim = c(-1.5, 1.5),
  Main = NA
)

```



You could also run placebo tests

24.0.4 Example 4

by Michael Robbins and Steven Davenport who are authors of **MicroSynth** with the following improvements:

- Standardization `use.survey = TRUE` and permutation (`perm = 250` and `jack = TRUE`) for placebo tests
- Omnibus statistic (set to `omnibus.var`) for multiple outcome variables
- incorporate multiple follow-up periods `end.post`

Notes:

- Both predictors and outcome will be used to match units before intervention
 - Outcome variable has to be **time-variant**
 - Predictors are **time-invariant**
-

```

library(microsynth)
data("seattledmi")

cov.var <- c("TotalPop", "BLACK", "HISPANIC", "Males_1521", "HOUSEHOLDS",
           "FAMILYHOU", "FEMALE_HOU", "RENTER_HOU", "VACANT_HOU")
match.out <- c("i_felony", "i_misdemea", "i_drugs", "any_crime")

seal1 <- microsynth(
  seattledmi,
  idvar = "ID",
  timevar = "time",
  intvar = "Intervention",
  start.pre = 1,
  end.pre = 12,
  end.post = 16,
  match.out = match.out, # outcome variable will be matched on exactly
  match.covar = cov.var, # specify covariates will be matched on exactly
  result.var = match.out, # used to report results
  omnibus.var = match.out, # feature in the omnibus p-value
  test = "lower",
  n.cores = min(parallel::detectCores(), 2)
)
seal1
#> microsynth object
#>
#> Scope:
#> Units:           Total: 9642 Treated: 39 Untreated: 9603
#> Study Period(s): Pre-period: 1 - 12 Post-period: 13 - 16
#> Constraints:    Exact Match: 58     Minimized Distance: 0
#> Time-variant outcomes:
#> Exact Match: i_felony, i_misdemea, i_drugs, any_crime (4)
#> Minimized Distance: (0)
#> Time-invariant covariates:
#> Exact Match: TotalPop, BLACK, HISPANIC, Males_1521, HOUSEHOLDS, FAMILYHOU, FEMALE_
#> Minimized Distance: (0)
#>
#> Results:
```

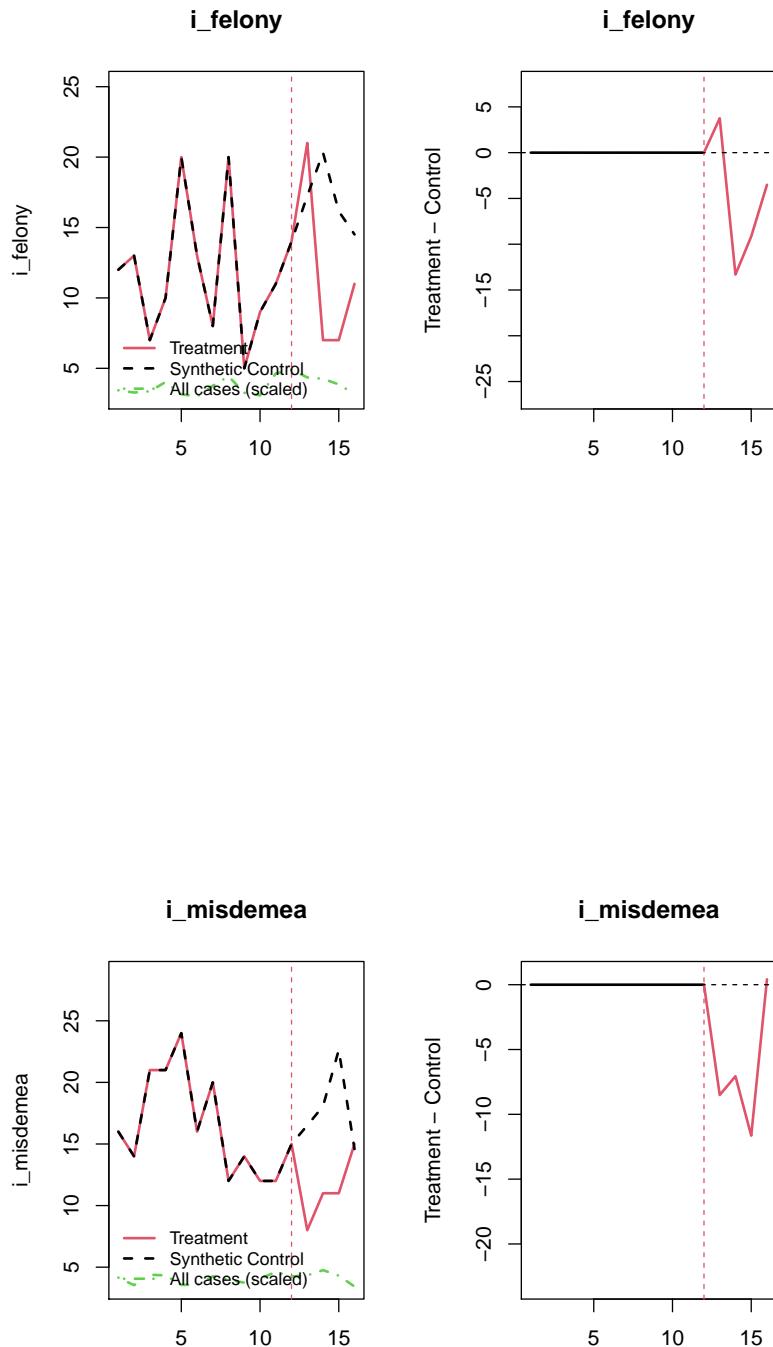
```
#> end.post = 16
#>          Trt    Con Pct.Chng Linear.pVal Linear.Lower Linear.Upper
#> i_felony   46  68.22  -32.6%      0.0109      -50.3%      -8.4%
#> i_misdemea 45  71.80  -37.3%      0.0019      -52.8%     -16.7%
#> i_drugs     20  23.76  -15.8%      0.2559      -46.4%      32.1%
#> any_crime  788 986.44 -20.1%      0.0146      -32.9%     -4.9%
#> Omnibus     --   --    --      0.0006      --        --
summary(sea1)
#> Weight Balance Table:
#>
#>          Targets Weighted.Control All.scaled
#> Intercept           39      39.000239 39.0000000
#> TotalPop            2994    2994.051921 2384.7476665
#> BLACK                173     173.000957 190.5224020
#> HISPANIC              149     149.002632 159.2682016
#> Males_1521             49      49.000000 97.3746111
#> HOUSEHOLDS           1968    1968.033976 1113.5588052
#> FAMILYHOU              519     519.010767 475.1876167
#> FEMALE_HOU              101     101.000957 81.1549471
#> RENTER_HOU              1868    1868.020338 581.9340386
#> VACANT_HOU              160     160.011485 98.4222153
#> i_felony.12             14      14.000000 4.9023024
#> i_felony.11             11      11.000239 4.6313006
#> i_felony.10              9      9.000000 3.0740510
#> i_felony.9               5      5.000000 3.2641568
#> i_felony.8               20     20.000000 4.4331052
#> i_felony.7               8      8.000000 3.7616677
#> i_felony.6               13     13.000000 3.0012446
#> i_felony.5               20     20.000718 3.1549471
#> i_felony.4               10     10.000000 4.0245800
#> i_felony.3               7      7.000000 3.3693217
#> i_felony.2               13     13.000239 3.2803360
#> i_felony.1               12     12.000000 3.4380834
#> i_misdemea.12             15     15.000239 4.2470442
#> i_misdemea.11             12     12.000000 4.6070317
#> i_misdemea.10             12     12.000000 4.0771624
#> i_misdemea.9               14     14.000000 3.7414437
#> i_misdemea.8               12     12.000000 3.9679527
#> i_misdemea.7               20     20.000000 4.2551338
#> i_misdemea.6               16     16.000479 3.5594275
#> i_misdemea.5               24     24.000000 3.5634723
#> i_misdemea.4               21     21.000239 4.3360299
#> i_misdemea.3               21     21.000000 4.3845675
#> i_misdemea.2               14     14.000000 3.5351587
#> i_misdemea.1               16     16.000000 4.1540137
```

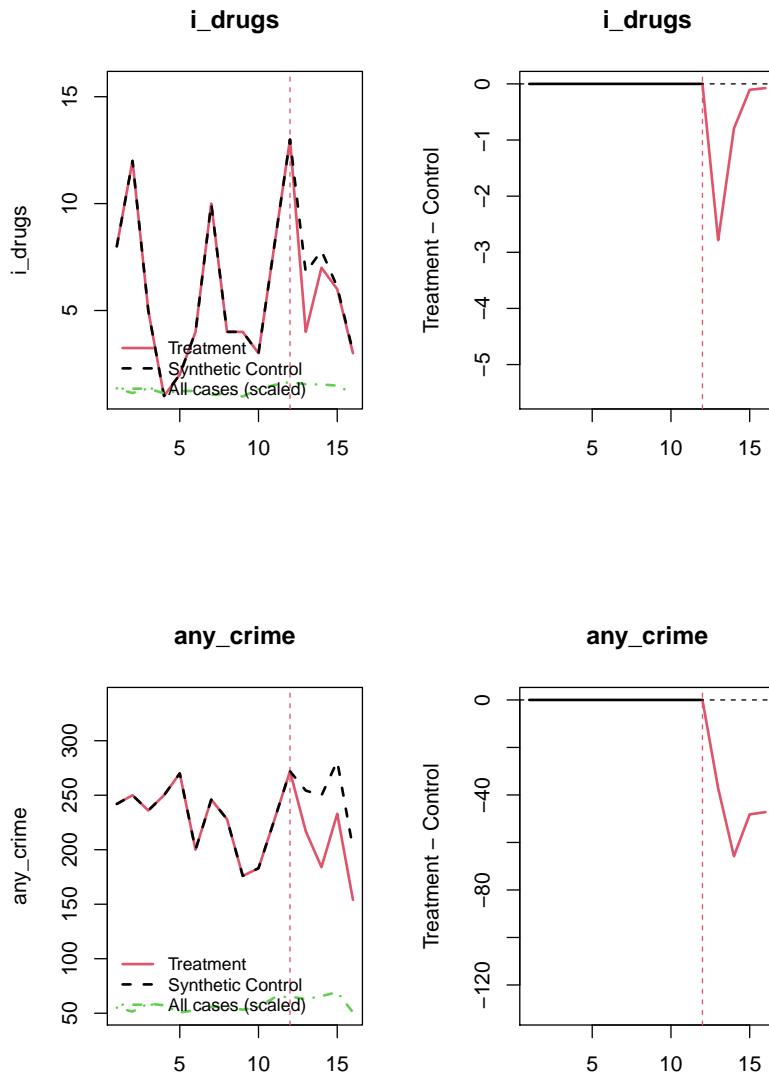
```

#> i_drugs.12      13    13.000000  1.6543248
#> i_drugs.11      8     8.000000  1.5127567
#> i_drugs.10      3     3.000000  1.3226509
#> i_drugs.9       4     4.000000  0.9788426
#> i_drugs.8       4     4.000000  1.1123211
#> i_drugs.7       10    10.000000 1.0516490
#> i_drugs.6       4     4.000000  1.2377100
#> i_drugs.5       2     2.000000  1.2296204
#> i_drugs.4       1     1.000000  1.1244555
#> i_drugs.3       5     5.000000  1.3550093
#> i_drugs.2       12    12.000000 1.1365899
#> i_drugs.1       8     8.000239  1.3590541
#> any_crime.12    272   272.001196 65.3397635
#> any_crime.11    227   227.001675 64.2395769
#> any_crime.10    183   183.000957 55.6929060
#> any_crime.9     176   176.000479 53.2377100
#> any_crime.8     228   228.000479 55.8142502
#> any_crime.7     246   246.002393 55.8061605
#> any_crime.6     200   200.000957 52.8291848
#> any_crime.5     270   270.001436 50.6530803
#> any_crime.4     250   250.000957 57.2946484
#> any_crime.3     236   236.000957 58.8680772
#> any_crime.2     250   250.001196 51.5429371
#> any_crime.1     242   242.000957 55.1144991
#>
#> Results:
#>
#> end.post = 16
#>          Trt  Con Pct.Chng Linear.pVal Linear.Lower Linear.Upper
#> i_felony    46  68.22 -32.6%    0.0109    -50.3%    -8.4%
#> i_misdemea 45  71.80 -37.3%    0.0019    -52.8%   -16.7%
#> i_drugs     20  23.76 -15.8%    0.2559    -46.4%    32.1%
#> any_crime   788 986.44 -20.1%    0.0146    -32.9%   -4.9%
#> Omnibus    --   --   --    0.0006    --      --

```

```
plot_microsynth(sea1)
```





```
sea2 <- microsynth(seattledmi,
  idvar="ID", timevar="time", intvar="Intervention",
  start.pre=1, end.pre=12, end.post=c(14, 16),
  match.out=match.out, match.covar=cov.var,
  result.var=match.out, omnibus.var=match.out,
  test="lower",
  perm=250, jack=TRUE,
```

```
n.cores = min(parallel::detectCores(), 2))
```

24.1 Synthetic Difference-in-differences

reference: (Arkhangelsky et al., 2021)

Chapter 25

Event Studies

The earliest paper that used event study was (Dolley, 1933)

(Campbell et al., 1997) introduced this method, which based on the efficient markets theory by (Fama, 1970)

Previous marketing studies:

- (Horsky and Swyngedouw, 1987): name change
- (Chaney et al., 1991) new product announcements
- (Agrawal and Kamakura, 1995): celebrity endorsement
- (Lane and Jacobson, 1995): brand extensions
- (Johnson and Houston, 2000): joint venture
- (Geyskens et al., 2002): Internet channel (for newspapers)
- (Wiles et al., 2010): Regulatory Reports of Deceptive Advertising
- (Sood and Tellis, 2009): innovation payoff

Potential avenues:

- Ad campaigns
- Market entry
- product failure/recalls
- Patents

Pros:

- Better than accounting based measures (e.g., profits) because managers can manipulate profits (Benston, 1985)
- Easy to do

Events can be

- Internal (e.g., stock repurchase)
- External (e.g., macroeconomic variables)

Assumptions:

1. Efficient market theory
2. Shareholders are the most important group among stakeholders
3. The event sharply affects share price
4. Expected return is calculated appropriately

Steps:

1. Event Identification: (e.g., dividends, M&A, stock buyback, laws or regulation, privatization vs. nationalization, celebrity endorsements, name changes, or brand extensions etc.)
1. Estimation window: Normal return expected return ($T_0 \rightarrow T_1$) (sometimes include days before to capture leakages).
 - Recommendation by (Moorman and Lehmann, 2004, p. 13) is to use 250 days before the event (and 45-day between the estimation window and the event window).
 - Similarly, (McWilliams and Siegel, 1997) and (Fornell et al., 2006) 255 days ending 46 days before the event date
2. Event window: contain the event date ($T_1 \rightarrow T_2$) (have to argue for the event window and can't do it empirically)
3. Post Event window: $T_2 \rightarrow T_3$
2. Normal vs. Abnormal returns

$$\epsilon_{it}^* = R_{it} - E(R_{it}|X_t)$$

where

- ϵ_{it}^* = abnormal return
- R_{it} = realized (actual) return
- $E(R_{it}|X_t)$ normal expected return

There are several model to calculate the expected return

A. Statistical Models: assumes jointly multivariate normal and iid over time (need distributional assumptions for valid finite-sample estimation) rather robust (hence, recommended)

1. Constant Mean Return Model
2. Market Model
3. Adjusted Market Return Model
4. Factor Model

B. Economic Models (strong assumption regarding investor behavior)

1. Capital Asset Pricing Model
2. Arbitrage pricing theory

25.1 Other Issues

25.1.1 Economic significance

Total wealth gain (loss) from the event

$$\Delta W_t = CAR_t \times MKTVAL_0$$

where

- ΔW_t = gain (loss)
- CAR_t = cumulative residuals to date t
- $MKTVAL_0$ market value of the firm before the event window

25.1.2 Statistical Power

increases with

- more firms
- less days in the event window (avoiding potential contamination from confounds)

25.1.3 Testing

25.1.3.1 Parametric Test

Applying CLT

$$t_{CAR} = \frac{\bar{CAR}_{it}}{\sigma(CAR_{it})/\sqrt{n}} t_{BHAR} = \frac{\bar{BHAR}_{it}}{\sigma(BHAR_{it})/\sqrt{n}}$$

25.1.3.2 Non-parametric Test

- No assumptions about return distribution
- Sign Test (assumes symmetry in returns)
- Rank Test (allows for non-symmetry in returns)

25.1.4 Confounders

According to (Fornell et al., 2006), need to control:

- one-day event period = day when Wall Street Journal publish ACSI announcement.
- 5 days before and after event to rule out other news (PR Newswires, Dow Jones, Business Wires)
 - M&A, Spin-offs, stock splits
 - CEO or CFO changes,
 - Layoffs, restructurings, earnings announcements, lawsuits
 - Capital IQ - Key Developments: covers almost all important events so you don't have to search on news.

25.1.5 Biases

- Different closing time obscure estimation of the abnormal returns, check (Campbell et al., 1998)
- Upward bias in aggregating CAR + transaction prices (bid and ask)
- Cross-sectional dependence in the returns bias the standard deviation estimates downward, which inflates the test statistics when events share common dates (MacKinlay, 1997). Hence, (Jaffe, 1974) Portfolio method should be used to correct for this bias.

25.1.6 Long-run event studies

- 12 - 60 months event window: (LOUGHAN and RITTER, 1995) (BRAV and GOMPERS, 1997) (Desai and Jain, 1999)
- Exemplar: (Dutta et al., 2018)

```
library(crseEventStudy)

# example by the package's author
data(demo_returns)
SAR <-
  sar(event = demo_returns$EON,
      control = demo_returns$RWE,
      logret = FALSE)
mean(SAR)
#> [1] 0.006870196
```

25.1.6.1 Buy and Hold Abnormal Returns (BHAR)

- Classic references: (LOUGHAN and RITTER, 1995) (Barber and Lyon, 1997) (Lyon et al., 1999)

$$AR_{it} = R_{it} - E(R_{it}|X_t)$$

$$BHAR_{t \rightarrow t+K}^i = \Pi_k (1 + AR_{i(t+k)})$$

Where as CAR is the arithmetic sum, BHAR is the geometric sum

25.1.6.2 Portfolio method

This section follows strictly the procedure in (Wiles et al., 2010)

A portfolio for every day in calendar time (including all securities which experience an event that time).

For each portfolio, the securities and their returns are equally weighted

1. For all portfolios, the average abnormal return are calculated as

$$AAR_{Pt} = \frac{\sum_{i=1}^S AR_i}{S}$$

where

- S is the number of securities in portfolio P
 - AR_i is the abnormal return for the stock i in the portfolio
2. For every portfolio P , a time series estimate of $\sigma(AAR_{Pt})$ is calculated for the preceding k days, assuming that the AAR_{Pt} are independent over time.
 3. Each portfolio's average abnormal return is standardized

$$SAAR_{Pt} = \frac{AAR_{Pt}}{SD(AAR_{Pt})}$$

4. Average standardized residual across all portfolio's in calendar time

$$ASAAR = \frac{1}{n} \sum_{i=1}^{255} SAAR_{Pt} \times D_t$$

where

- $D_t = 1$ when there is at least one security in portfolio t
 - $D_t = 0$ when there are no security in portfolio t
 - n is the number of days in which the portfolio have at least one security
 $n = \sum_{i=1}^{255} D_t$
5. The cumulative average standardized average abnormal returns is

$$CASSAR_{S_1, S_2} = \sum_{i=S_1}^{S_2} ASAAR$$

If the ASAAR are independent over time, then standard deviation for the above estimate is $\sqrt{S_2 - S_1 + 1}$

then, the test statistics is

$$t = \frac{CASSAR_{S_1, S_2}}{\sqrt{S_2 - S_1 + 1}}$$

25.2 Aggregation

25.2.1 Over Time

We calculate the cumulative abnormal (CAR) for the event windows

H_0 : Standardized cumulative abnormal return for stock i is 0 (no effect of events on stock performance)

H_1 : SCAR is not 0 (there is an effect of events on stock performance)

25.2.2 Across Firms + Over Time

Additional assumptions: Abnormal returns of different stocks are uncorrelated (rather strong), but it's very valid if event windows for different stocks do not overlap. If the windows for different overlap, follow (?) and (?)

H_0 : The mean of the abnormal returns across all firms is 0 (no effect)

H_1 : The mean of the abnormal returns across all firms is different from 0 (there is an effect)

Parametric (empirically either one works fine) (assume abnormal returns is normally distributed) :

1. Aggregate the CAR of all stocks (Use this if the true abnormal variance is greater for stocks with higher variance)
2. Aggregate the SCAR of all stocks (Use this if the true abnormal return is constant across all stocks)

Non-parametric (no parametric assumptions):

1. Sign test:
 - Assume both the abnormal returns and CAR to be independent across stocks
 - Assume 50% with positive abnormal returns and 50% with negative abnormal return
 - The null will be that there is a positive abnormal return correlated with the event (if you want the alternative to be there is a negative relationship)
 - With skewed distribution (likely in daily stock data), the size test is not trustworthy. Hence, rank test might be better
2. Rank test
 - Null: there is no abnormal return during the event window

25.3 Heterogeneity in the event effect

$$y = X\theta + \eta$$

where

- y = CAR
- X = Characteristics that lead to heterogeneity in the event effect (i.e., abnormal returns) (e.g., firm or event specific)
- η = error term

Note:

- In cases with selection bias (firm characteristics and investor anticipation of the event: larger firms might enjoy great positive effect of an event, and investors endogenously anticipate this effect and overvalue the stock), we have to use the White's t -statistics to have the lower bounds of the true significance of the estimates.

25.4 Expected Return Calculation

25.4.1 Statistical Models

- based on statistical assumptions about the behavior of returns (e.g., multivariate normality)
- we only need to assume stable distributions (OWEN and RABINOVITCH, 1983)

25.4.1.1 Constant Mean Return Model

The expected normal return is the mean of the real returns

$$Ra_{it} = R_{it} - \bar{R}_i$$

Assumption:

- returns revert to its mean (very questionable)

The basic mean returns model generally delivers similar findings to more complex models since the variance of abnormal returns is not decreased considerably (Brown and Warner, 1985)

25.4.1.2 Market Model

$$R_{it} = \alpha_i + \beta R_{mt} + \epsilon_{it}$$

where

- R_{it} = stock return i in period t
- R_{mt} = market return
- ϵ_{it} = zero mean ($E(\epsilon_{it}) = 0$) error term with its own variance σ^2

Notes:

- People typically use S&P 500, CRSP value-weighted or equal-weighted index as the market portfolio.
- When $\beta = 0$, the Market Model is the Constant Mean Return Model
- better fit of the market-model, the less variance in abnormal return, and the more easy to detect the event's effect
- recommend generalized method of moments to be robust against autocorrelation and heteroskedasticity

25.4.1.3 FF3

(Fama and French, 1993)

$$E(R_{it}|X_t) - r_{ft} = \alpha_i + \beta_{1i}(E(R_{mt}|X_t) - r_{ft}) + b_{2i}SML_t + b_{3i}HML_t$$

where

- r_{ft} risk-free rate (e.g., 3-month Treasury bill)
- R_{mt} is the market-rate (e.g., S&P 500)
- SML: returns on small (size) portfolio minus returns on big portfolio
- HML: returns on high (B/M) portfolio minus returns on low portfolio.

25.4.2 Economic Model

The only difference between CAPM and APT is that APT has multiple factors (including factors beyond the focal company)

Economic models put limits on a statistical model that come from assumed behavior that is derived from theory.

25.4.2.1 Capital Asset Pricing Model (CAPM)

$$E(R_i) = R_f + \beta_i(E(R_m) - R_f)$$

where

- $E(R_i)$ = expected firm return
- R_f = risk free rate
- $E(R_m - R_f)$ = market risk premium
- β_i = firm sensitivity

25.4.2.2 Arbitrage Pricing Theory (APT)

$$R = R_f + \Lambda f + \epsilon$$

where

- $\epsilon \sim N(0, \Psi)$
- Λ = factor loadings
- $f \sim N(\mu, \Omega)$ = general factor model
 - μ = expected risk premium vector
 - Ω = factor covariance matrix

25.5 Application

Packages:

- `eventstudies`
- `erer`
- `EventStudy`
- `AbnormalReturns`
- Event Study Tools
- `estudy2`
- `PerformanceAnalytics`

In practice, people usually sort portfolio because they are not sure whether the FF model is specified correctly.

Steps:

1. Sort all returns in CRSP into 10 deciles based on size.
2. In each decile, sort returns into 10 deciles based on BM
3. Get the average return of the 100 portfolios for each period (i.e., expected returns of stocks given decile - characteristics)
4. For each stock in the event study: Compare the return of the stock to the corresponding portfolio based on size and BM.

Notes:

- Sorting produces outcomes that are often more conservative (e.g., FF abnormal returns can be greater than those that used sorting).
- If the results change when we do B/M first then size or vice versa, then the results are not robust (this extends to more than just two characteristics - e.g., momentum).
-

Examples:

Forestry:

- (Mei and Sun, 2008) M&A on financial performance (forest product)
- (Sun and Liao, 2011) litigation on firm values

```
library(erer)

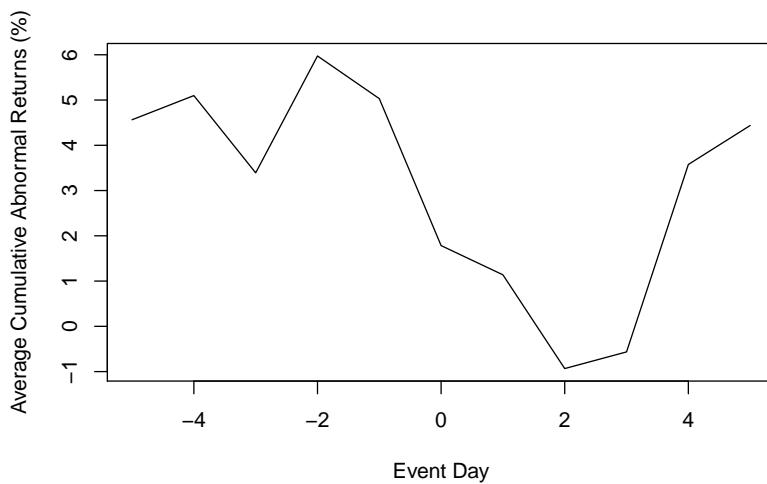
# example by the package's author
data(daEsa)
hh <- evReturn(
  y = daEsa,          # dataset
  firm = "wpp",       # firm name
  y.date = "date",    # date in y
  index = "sp500",    # index
  est.win = 250,      # estimation window width in days
  digits = 3,
  event.date = 19990505, # firm event dates
  event.win = 5        # one-side event window width in days (default = 3, where 3 before + 1)
)
hh; plot(hh)
```

```

#>
#> === Regression coefficients by firm =====
#>   N firm event.date alpha.c alpha.e alpha.t alpha.p alpha.s beta.c beta.e
#> 1 1 upp    19990505 -0.135   0.170  -0.795   0.428        0.665  0.123
#>   beta.t beta.p beta.s
#> 1  5.419  0.000 ***

#>
#> === Abnormal returns by date =====
#>   day Ait.wpp      Hnt
#> 1  -5  4.564  4.564
#> 2  -4  0.534  5.098
#> 3  -3 -1.707  3.391
#> 4  -2  2.582  5.973
#> 5  -1 -0.942  5.031
#> 6   0 -3.247  1.784
#> 7   1 -0.646  1.138
#> 8   2 -2.071 -0.933
#> 9   3  0.368 -0.565
#> 10  4  4.141  3.576
#> 11  5  0.861  4.437
#>
#> === Average abnormal returns across firms ===
#>   name estimate error t.value p.value sig
#> 1 CiT.wpp    4.437 8.888  0.499  0.618
#> 2 GNT       4.437 8.888  0.499  0.618

```



Example by [Ana Julia Akaishi Padula, Pedro Albuquerque (posted on LAMFO)]([https://lamfo-unb.github.io/2017/08/17/Teste-de-Eventos-en/#:~:text=The%20abnormal%20return%20\(Ra,regression%20in%20the%20estimation%20window.\)](https://lamfo-unb.github.io/2017/08/17/Teste-de-Eventos-en/#:~:text=The%20abnormal%20return%20(Ra,regression%20in%20the%20estimation%20window.)))

Example in **AbnormalReturns** package

Chapter 26

Matching Methods

Assumption: Observables can identify the selection into the treatment and control groups

Identification: The exclusion restriction can be met conditional on the observables

Motivation

Effect of college quality on earnings

They ultimately estimate the treatment effect on the treated of attending a top (high ACT) versus bottom (low ACT) quartile college

Example

Aaronson et al. (2007)

Do teachers qualifications (causally) affect student test scores?

Step 1:

$$Y_{ijt} = \delta_0 + Y_{ij(t-1)}\delta_1 + X_{it}\delta_2 + Z_{jt}\delta_3 + \epsilon_{ijt}$$

There can always be another variable

Any observable sorting is imperfect

Step 2:

$$Y_{ijst} = \alpha_0 + Y_{ij(t-1)}\alpha_1 + X_{it}\alpha_2 + Z_{jt}\alpha_3 + \gamma_s + u_{isjt}$$

- $\delta_3 > 0$
- $\delta_3 > \alpha_3$

- γ_s = school fixed effect

Sorting is less within school. Hence, we can introduce the school fixed effect

Step 3:

Find schools that look like they are putting students in class randomly (or as good as random) + we run step 2

$$Y_{isjt} = Y_{isj(t-1)}\lambda + X_{it}\alpha_1 + Z_{jt}\alpha_{21} + (Z_{jt} \times D_i)\alpha_{22} + \gamma_5 + u_{isjt}$$

- D_{it} is an element of X_{it}
- Z_{it} = teacher experience

$$D_{it} = \begin{cases} 1 & \text{if high poverty} \\ 0 & \text{otherwise} \end{cases}$$

$H_0 : \alpha_{22} = 0$ test for effect heterogeneity whether the effect of teacher experience (Z_{jt}) is different

- For low poverty is α_{21}
- For high poverty effect is $\alpha_{21} + \alpha_{22}$

Matching is **selection on observables** and only works if you have good observables.

Sufficient identification assumption under Selection on observable/ back-door criterion (based on Bernard Koch's presentation)

- Strong conditional ignorability
 - $Y(0), Y(1) \perp T|X$
 - No hidden confounders
- Overlap
 - $\forall x \in X, t \in \{0, 1\} : p(T = t|X = x) > 0$
 - All treatments have non-zero probability of being observed
- SUTVA/ Consistency
 - Treatment and outcomes of different subjects are independent

Relative to OLS

1. Matching makes the **common support** explicit (and changes default from “ignore” to “enforce”)
2. Relaxes linear function form. Thus, less parametric.

It also helps if you have high ratio of controls to treatments.

For detail summary (Stuart, 2010)

Matching is defined as “any method that aims to equate (or “balance”) the distribution of covariates in the treated and control groups.” (Stuart, 2010, pp. 1)

Equivalently, matching is a selection on observables identifications strategy.

If you think your OLS estimate is biased, a matching estimate (almost surely) is too.

Unconditionally, consider

$$E(Y_i^T|T) - E(Y_i^C|C) + E(Y_i^C|T) - E(Y_i^C|T) = E(Y_i^T - Y_i^C|T) + [E(Y_i^C|T) - E(Y_i^C|C)] = E(Y_i^T - Y_i^C|T) + \text{selection bias}$$

where $E(Y_i^T - Y_i^C|T)$ is the causal inference that we want to know.

Randomization eliminates the selection bias.

If we don't have randomization, then $E(Y_i^C|T) \neq E(Y_i^C|C)$

Matching tries to do selection on observables $E(Y_i^C|X, T) = E(Y_i^C|X, C)$

Propensity Scores basically do $E(Y_i^C|P(X), T) = E(Y_i^C|P(X), C)$

Matching standard errors will exceed OLS standard errors

The treatment should have larger predictive power than the control because you use treatment to pick control (not control to pick treatment).

The average treatment effect (ATE) is

$$\frac{1}{N_T} \sum_{i=1}^{N_T} (Y_i^T - \frac{1}{N_{C_T}} \sum_{i=1}^{N_{C_T}} Y_i^C)$$

Since there is no closed-form solution for the standard error of the average treatment effect, we have to use bootstrapping to get standard error.

Professor Gary King advocates instead of using the word “matching”, we should use “**pruning**” (i.e., deleting observations). It is a preprocessing step where it prunes nonmatches to make control variables less important in your analysis.

Without Matching

- **Imbalance data** leads to **model dependence** lead to a lot of **researcher discretion** leads to **bias**

With Matching

- We have balance data which essentially erase human discretion

Table 26.1: Table @ref(tab:Gary King - International Methods Colloquium talk 2015)

Balance Covariates	Complete Randomization	Fully Exact
Observed	On average	Exact
Unobserved	On average	On average

Fully blocked is superior on

- imbalance
- model dependence
- power
- efficiency
- bias
- research costs
- robustness

Matching is used when

- Outcomes are not available to select subjects for follow-up
- Outcomes are available to improve precision of the estimate (i.e., reduce bias)

Hence, we can only observe one outcome of a unit (either treated or control), we can think of this problem as missing data as well. Thus, this section is closely related to Imputation (Missing Data)

In observational studies, we cannot randomize the treatment effect. Subjects select their own treatments, which could introduce selection bias (i.e., systematic differences between group differences that confound the effects of response variable differences).

Matching is used to

- reduce model dependence
- diagnose balance in the dataset

Assumptions of matching:

1. treatment assignment is independent of potential outcomes given the covariates
 - $T \perp (Y(0), Y(1))|X$
 - known as ignorability, or ignorable, no hidden bias, or unconfounded.
 - You typically satisfy this assumption when unobserved covariates correlated with observed covariates.
 - But when unobserved covariates are unrelated to the observed covariates, you can use sensitivity analysis to check your result, or use “design sensitivity” (Heller et al., 2009)
2. positive probability of receiving treatment for all X
 - $0 < P(T = 1|X) < 1 \forall X$
3. Stable Unit Treatment value Assumption (SUTVA)
 - Outcomes of A are not affected by treatment of B.
 - Very hard in cases where there is “spillover” effects (interactions between control and treatment). To combat, we need to reduce interactions.

Generalization

- P_t : treated population -> N_t : random sample from treated
- P_c : control population -> N_c : random sample from control
- μ_i = means ; Σ_i = variance covariance matrix of the p covariates in group i ($i = t, c$)
- X_j = p covariates of individual j
- T_j = treatment assignment
- Y_j = observed outcome
- Assume: $N_t < N_c$
- Treatment effect is $\tau(x) = R_1(x) - R_0(x)$ where

- $R_1(x) = E(Y(1)|X)$
- $R_0(x) = E(Y(0)|X)$
- Assume: parallel trends hence $\tau(x) = \tau \forall x$
 - If the parallel trends are not assumed, an average effect can be estimated.
- Common estimands:
 - Average effect of the treatment on the treated (ATT): effects on treatment group
 - Average treatment effect (ATE): effect on both treatment and control

Steps:

1. Define “closeness”: decide distance measure to be used
 1. Which variables to include:
 1. Ignorability (no unobserved differences between treatment and control)
 1. Since cost of including unrelated variables is small, you should include as many as possible (unless sample size/power doesn't allow you to because of increased variance)
 2. Do not include variables that were affected by the treatment.
 3. Note: if a matching variable (i.e., heavy drug users) is highly correlated to the outcome variable (i.e., heavy drinkers), you will be better to exclude it in the matching set.
 2. Which distance measures: more below
 2. Matching methods
 1. Nearest neighbor matching
 1. Simple (greedy) matching: performs poorly when there is competition for controls.
 2. Optimal matching: considers global distance measure
 3. Ratio matching: to combat increase bias and reduced variation when you have k:1 matching, one can use approximations by Rubin and Thomas (1996).
 4. With or without replacement: with replacement is typically better, but one needs to account for dependence in the matched sample when doing later analysis (can use frequency weights to combat).

2. Subclassification, Full Matching and Weighting

Nearest neighbor matching assigns 0 (control) or 1 (treated), while these methods use weights between 0 and 1.

1. Subclassification: distribution into multiple subclass (e.g., 5-10)
2. Full matching: optimally minimize the average of the distances between each treated unit and each control unit within each matched set.
3. Weighting adjustments: weighting technique uses propensity scores to estimate ATE. If the weights are extreme, the variance can be large not due to the underlying probabilities, but due to the estimation procedure. To combat this, use (1) weight trimming, or (2) doubly robust methods when propensity scores are used for weighing or matching.
 1. Inverse probability of treatment weighting (IPTW) $w_i = \frac{T_i}{\hat{e}_i} + \frac{1-T_i}{1-\hat{e}_i}$
 2. Odds $w_i = T_i + (1-T_i) \frac{\hat{e}_i}{1-\hat{e}_i}$
 3. Kernel weighting (e.g., in economics) averages over multiple units in the control group.

3. Assessing Common Support

- common support means overlapping of the propensity score distributions in the treatment and control groups. Propensity score is used to discard control units from the common support. Alternatively, convex hull of the covariates in the multi-dimensional space.

3. Assessing the quality of matched samples (Diagnose)

- Balance = similarity of the empirical distribution of the full set of covariates in the matched treated and control groups. Equivalently, treatment is unrelated to the covariates
 - $\tilde{p}(X|T=1) = \tilde{p}(X|T=0)$ where \tilde{p} is the empirical distribution.
- Numerical Diagnostics
 1. standardized difference in means of each covariate (most common), also known as "standardized bias", "standardized difference in means".
 2. standardized difference of means of the propensity score (should be < 0.25) (Rubin, 2001)
 3. ratio of the variances of the propensity score in the treated and control groups (should be between 0.5 and 2). (Rubin, 2001)
 4. For each covariate, the ratio of the variance of the residuals orthogonal to the propensity score in the treated and control groups.

Note: can't use hypothesis tests or p-values because of (1) in-sample property (not population), (2) conflation of changes in balance with changes in statistical power.

- Graphical Diagnostics
 - QQ plots
 - Empirical Distribution Plot

4. Estimate the treatment effect

1. After k:1
 1. Need to account for weights when use matching with replacement.
2. After Subclassification and Full Matching
 1. weighting the subclass estimates by the number of treated units in each subclass for ATT
 2. Weighting by the overall number of individual in each subclass for ATE.
3. Variance estimation: should incorporate uncertainties in both the matching procedure (step 3) and the estimation procedure (step 4)

Notes:

- With missing data, use generalized boosted models, or multiple imputation (Qu and Lipkovich, 2009)
- Violation of ignorable treatment assignment (i.e., unobservables affect treatment and outcome). control by
 - measure pre-treatment measure of the outcome variable
 - find the difference in outcomes between multiple control groups. If there is a significant difference, there is evidence for violation.
 - find the range of correlations between unobservables and both treatment assignment and outcome to nullify the significant effect.
- Choosing between methods
 - smallest standardized difference of mean across the largest number of covariates
 - minimize the standardized difference of means of a few particularly prognostic covariates
 - fest number of large standardized difference of means (> 0.25)
 - (Diamond and Sekhon, 2013) automates the process

- In practice
 - If ATE, ask if there is enough overlap of the treated and control groups' propensity score to estimate ATE, if not use ATT instead
 - If ATT, ask if there are controls across the full range of the treated group
- Choose matching method
 - If ATE, use IPTW or full matching
 - If ATT, and more controls than treated (at least 3 times), k:1 nearest neighbor without replacement
 - If ATT, and few controls , use subclassification, full matching, and weighting by the odds
- Diagnostic
 - If balance, use regression on matched samples
 - If imbalance on few covariates, treat them with Mahalanobis
 - If imbalance on many covariates, try k:1 matching with replacement

Ways to define the distance D_{ij}

1. Exact

$$D_{ij} = \begin{cases} 0, & \text{if } X_i = X_j, \\ \infty, & \text{if } X_i \neq X_j \end{cases}$$

An advanced is Coarsened Exact Matching

2. Mahalanobis

$$D_{ij} = (X_i - X_j)' \Sigma^{-1} (X_i - X_j)$$

where

Σ = variance covariance matrix of X in the

- control group if ATT is interested
- polled treatment and control groups if ATE is interested

3. Propensity score:

$$D_{ij} = |e_i - e_j|$$

where e_k = the propensity score for individual k

An advanced is Prognosis score (Hansen, 2008), but you have to know (i.e., specify) the relationship between the covariates and outcome.

4. Linear propensity score

$$D_{ij} = |\text{logit}(e_i) - \text{logit}(e_j)|$$

The exact and Mahalanobis are not good in high dimensional or non normally distributed X's cases.

We can combine Mahalanobis matching with propensity score calipers (Rubin and Thomas, 2000)

Other advanced methods for longitudinal settings

- marginal structural models (Robins et al., 2000)
- balanced risk set matching (Li et al., 2001)

Most matching methods are based on (ex-post)

- propensity score
- distance metric
- covariates

Packages

- **cem** Coarsened exact matching
- **Matching** Multivariate and propensity score matching with balance optimization
- **MatchIt** Nonparametric preprocessing for parametric causal inference.
Have nearest neighbor, Mahalanobis, caliper, exact, full, optimal, sub-classification

- `MatchingFrontier` optimize balance and sample size (King et al., 2016)
- `optmatch` optimal matching with variable ratio, optimal and full matching
- `PSAgraphics` Propensity score graphics
- `rbounds` sensitivity analysis with matched data, examine ignorable treatment assignment assumption
- `twang` weighting and analysis of non-equivalent groups
- `CBPS` covariate balancing propensity score. Can also be used in the longitudinal setting with marginal structural models.
- `PanelMatch` based on Imai, Kim, and Wang (2018)

Matching	Regression
Not as sensitive to the functional form of the covariates	can estimate the effect of a continuous treatment
Easier to assess whether it's working	estimate the effect of all the variables (not just the treatment)
Easier to explain	can estimate interactions of treatment with covariates
allows a nice visualization of an evaluation	
If your treatment is fairly rare, you may have a lot of control observations that are obviously not comparable	
Less parametric	More parametric
Enforces common support (i.e., space where treatment and control have the same characteristics)	

However, the problem of **omitted variables** (i.e., those that affect both the outcome and whether observation was treated) - unobserved confounders is still present in matching methods.

Difference between matching and regression following Jörn-Stephan Pischke's lecture

Suppose we want to estimate the effect of treatment on the treated

$$\begin{aligned}\delta_{TOT} &= E[Y_{1i} - Y_{0i} | D_i = 1] \\ &= E\{E[Y_{1i}|X_i, D_i = 1] - E[Y_{0i}|X_i, D_i = 1]\} | D_i = 1\} \quad \text{law of iterated expectations}\end{aligned}$$

Under conditional independence

$$E[Y_{0i}|X_i, D_i = 0] = E[Y_{0i}|X_i, D_i = 1]$$

then

$$\begin{aligned}\delta_{TOT} &= E\{E[Y_1|X_i, D_i = 1] - E[Y_0|X_i, D_i = 0]|D_i = 1\} \\ &= E\{E[y_i|X_i, D_i = 1] - E[y_i|X_i, D_i = 0]|D_i = 1\} \\ &= E[\delta_X|D_i = 1]\end{aligned}$$

where δ_X is an X-specific difference in means at covariate value X_i

When X_i is discrete, the matching estimand is

$$\delta_M = \sum_x \delta_x P(X_i = x|D_i = 1)$$

where $P(X_i = x|D_i = 1)$ is the probability mass function for X_i given $D_i = 1$

According to Bayes rule,

$$P(X_i = x|D_i = 1) = \frac{P(D_i = 1|X_i = x) \times P(X_i = x)}{P(D_i = 1)}$$

hence,

$$\begin{aligned}\delta_M &= \frac{\sum_x \delta_x P(D_i = 1|X_i = x)P(X_i = x)}{\sum_x P(D_i = 1|X_i = x)P(X_i = x)} \\ &= \sum_x \delta_x \frac{P(D_i = 1|X_i = x)P(X_i = x)}{\sum_x P(D_i = 1|X_i = x)P(X_i = x)}\end{aligned}$$

On the other hand, suppose we have regression

$$y_i = \sum_x d_{ix} \beta_x + \delta_R D_i + \epsilon_i$$

where

- d_{ix} = dummy that indicates $X_i = x$
- β_x = regression-effect for $X_i = x$
- δ_R = regression estimand where

$$\begin{aligned}\delta_R &= \frac{\sum_x \delta_x [P(D_i = 1|X_i = x)(1 - P(D_i = 1|X_i = x))]P(X_i = x)}{\sum_x [P(D_i = 1|X_i = x)(1 - P(D_i = 1|X_i = x))]P(X_i = x)} \\ &= \sum_x \delta_x \frac{[P(D_i = 1|X_i = x)(1 - P(D_i = 1|X_i = x))]P(X_i = x)}{\sum_x [P(D_i = 1|X_i = x)(1 - P(D_i = 1|X_i = x))]P(X_i = x)}\end{aligned}$$

the difference between the regression and matching estimand is the weights they use to combine the covariate specific treatment effect δ_x

Type	depends on	uses weights which	interpretation	makes sense because
MatchIt	$D_i = 1 X_i = x)$ the fraction of treated observa- tions in a covariate cell (i.e., or the mean of D_i)	This is larger in cells with many treated observations.		we want the effect of treatment on the treated
Regression	$D_i = 1 X_i = x)(1 - P(D_i = 1 X_i))$ the variance of D_i in the covariate cell	This weight is largest in cells where there are half treated and half untreated observations. (this is the reason why we want to treat our sample so it is balanced, before running regular regression model, as mentioned above).		these cells will produce the lowest variance estimates of δ_x . If all the δ_x are the same, the most efficient estimand uses the lowest variance cells most heavily.

The goal of matching is to produce covariate balance (i.e., distributions of covariates in treatment and control groups are approximately similar as they would be in a successful randomized experiment).

26.1 MatchIt

Procedure typically involves (proposed by Noah Freifer using **MatchIt**)

1. planning
2. matching
3. checking (balance)
4. estimating the treatment effect

```
library(MatchIt)
data("lalonde")
```

examine **treat** on **re78**

1. Planning

- select type of effect to be estimated (e.g., mediation effect, conditional effect, marginal effect)
- select the target population
- select variables to match/balance (Austin, 2011) (VanderWeele, 2019)

2. Check Initial Imbalance

```
# No matching; constructing a pre-match matchit object
m.out0 <- matchit(
  formula(treat ~ age + educ + race + married + nodegree + re74 + re75, env = lalonde),
  data = data.frame(lalonde),
  method = NULL,
  # assess balance before matching
  distance = "glm" # logistic regression
)

# Checking balance prior to matching
summary(m.out0)
```

3. Matching

```
# 1:1 NN PS matching w/o replacement
m.out1 <- matchit(
  treat ~ age + educ, #+ race + married + nodegree + re74 + re75,
  data = lalonde,
  method = "nearest",
  distance = "glm"
)
m.out1
#> A matchit object
#> - method: 1:1 nearest neighbor matching without replacement
#> - distance: Propensity score
#>           - estimated with logistic regression
#> - number of obs.: 614 (original), 370 (matched)
#> - target estimand: ATT
#> - covariates: age, educ
```

4. Check balance

Sometimes you have to make trade-off between balance and sample size.

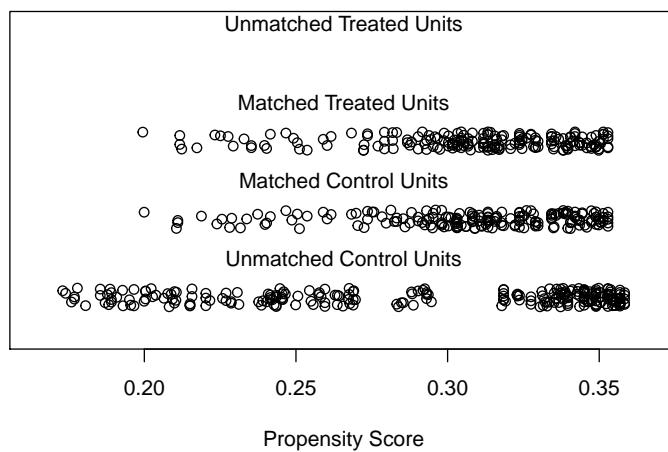
```

# Checking balance after NN matching
summary(m.out1, un = FALSE)
#>
#> Call:
#> matchit(formula = treat ~ age + educ, data = lalonde, method = "nearest",
#>           distance = "glm")
#>
#> Summary of Balance for Matched Data:
#>          Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
#> distance      0.3080      0.3077      0.0094  0.9963  0.0033
#> age          25.8162     25.8649     -0.0068  1.0300  0.0050
#> educ         10.3459     10.2865      0.0296  0.5886  0.0253
#>          eCDF Max Std. Pair Dist.
#> distance    0.0432      0.0146
#> age          0.0162      0.0597
#> educ         0.1189      0.8146
#>
#> Sample Sizes:
#>          Control Treated
#> All          429       185
#> Matched      185       185
#> Unmatched    244        0
#> Discarded     0        0

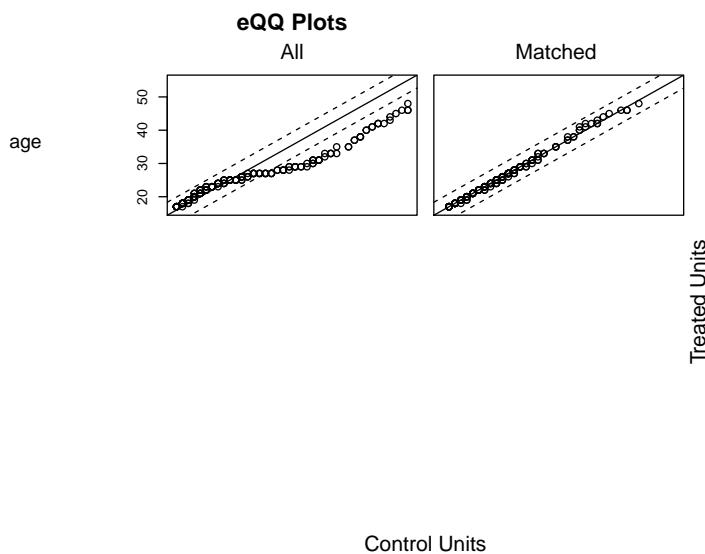
# examine visually
plot(m.out1, type = "jitter", interactive = FALSE)

```

Distribution of Propensity Scores



```
plot(
  m.out1,
  type = "qq",
  interactive = FALSE,
  which.xs = c("age")#, "married", "re75")
)
```



Try Full Match (i.e., every treated matches with one control, and every control with one treated).

```
# Full matching on a probit PS
m.out2 <- matchit(treat ~ age + educ, # + race + married + nodegree + re74 + re75,
                    data = lalonde,
                    method = "full",
                    distance = "glm",
                    link = "probit")
m.out2
#> A matchit object
#> - method: Optimal full matching
#> - distance: Propensity score
#>           - estimated with probit regression
#> - number of obs.: 614 (original), 614 (matched)
#> - target estimand: ATT
#> - covariates: age, educ
```

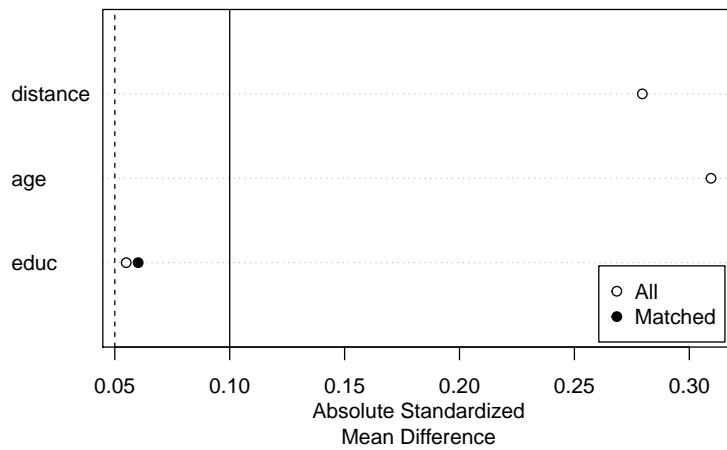
Checking balance again

```

# Checking balance after full matching
summary(m.out2, un = FALSE)
#>
#> Call:
#> matchit(formula = treat ~ age + educ, data = lalonde, method = "full",
#>           distance = "glm", link = "probit")
#>
#> Summary of Balance for Matched Data:
#>          Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
#> distance      0.3082      0.3082      0.0015   0.9837   0.0042
#> age          25.8162     25.8670     -0.0071   0.9966   0.0064
#> educ         10.3459     10.4669     -0.0602   0.4703   0.0416
#>          eCDF Max Std. Pair Dist.
#> distance    0.0270      0.0453
#> age          0.0264      0.1369
#> educ         0.1613      1.3057
#>
#> Sample Sizes:
#>          Control Treated
#> All        429.       185
#> Matched (ESS) 216.09     185
#> Matched      429.       185
#> Unmatched      0.         0
#> Discarded      0.         0

plot(summary(m.out2))

```



Exact Matching

```
# Full matching on a probit PS
m.out3 <-
  matchit(
    treat ~ age + educ, # + race + married + nodegree + re74 + re75,
    data = lalonde,
    method = "exact"
  )
m.out3
#> A matchit object
#> - method: Exact matching
#> - number of obs.: 614 (original), 332 (matched)
#> - target estimand: ATT
#> - covariates: age, educ
```

Subclassification

```
m.out4 <- matchit(
  treat ~ age + educ, # + race + married + nodegree + re74 + re75,
  data = lalonde,
  method = "subclass"
)
m.out4
#> A matchit object
#> - method: Subclassification (6 subclasses)
#> - distance: Propensity score
#>           - estimated with logistic regression
#> - number of obs.: 614 (original), 614 (matched)
#> - target estimand: ATT
#> - covariates: age, educ

# Or you can use in conjunction with "nearest"
m.out4 <- matchit(
  treat ~ age + educ, # + race + married + nodegree + re74 + re75,
  data = lalonde,
  method = "nearest",
  option = "subclass"
)
m.out4
#> A matchit object
#> - method: 1:1 nearest neighbor matching without replacement
#> - distance: Propensity score
#>           - estimated with logistic regression
#> - number of obs.: 614 (original), 370 (matched)
```

```
#> - target estimand: ATT
#> - covariates: age, educ
```

Optimal Matching

```
m.out5 <- matchit(
  treat ~ age + educ, # + race + married + nodegree + re74 + re75,
  data = lalonde,
  method = "optimal",
  ratio = 2
)
m.out5
#> A matchit object
#> - method: 2:1 optimal pair matching
#> - distance: Propensity score
#>           - estimated with logistic regression
#> - number of obs.: 614 (original), 555 (matched)
#> - target estimand: ATT
#> - covariates: age, educ
```

Genetic Matching

```
m.out6 <- matchit(
  treat ~ age + educ, # + race + married + nodegree + re74 + re75,
  data = lalonde,
  method = "genetic"
)
m.out6
#> A matchit object
#> - method: 1:1 genetic matching without replacement
#> - distance: Propensity score
#>           - estimated with logistic regression
#> - number of obs.: 614 (original), 370 (matched)
#> - target estimand: ATT
#> - covariates: age, educ
```

4. Estimating the Treatment Effect

```
# get matched data
m.data1 <- match.data(m.out1)

head(m.data1)
#>   treat age educ   race married nodegree re74 re75          re78  distance
```

```
#> NSW1      1  37   11  black      1      1    0    0  9930.0460 0.2536942
#> NSW2      1  22    9 hispan     0      1    0    0  3595.8940 0.3245468
#> NSW3      1  30   12  black     0      0    0    0  24909.4500 0.2881139
#> NSW4      1  27   11  black     0      1    0    0  7506.1460 0.3016672
#> NSW5      1  33    8  black     0      1    0    0  289.7899 0.2683025
#> NSW6      1  22    9  black     0      1    0    0  4056.4940 0.3245468
#>       weights subclass
#> NSW1      1        1
#> NSW2      1       98
#> NSW3      1      109
#> NSW4      1      120
#> NSW5      1      131
#> NSW6      1      142
```

```
library("lmtest") #coeftest
library("sandwich") #vcovCL

# imbalance matched dataset
fit1 <- lm(re78 ~ treat + age + educ , #+ race + married + nodegree + re74 + re75,
            data = m.data1,
            weights = weights)

coeftest(fit1, vcov. = vcovCL, cluster = ~subclass)
#>
#> t test of coefficients:
#>
#>           Estimate Std. Error t value Pr(>|t|)
#> (Intercept) -106.555   2464.875 -0.0432 0.965542
#> treat        -1221.832    785.492 -1.5555 0.120691
#> age          152.118     53.241  2.8571 0.004519 **
#> educ         362.502    171.893  2.1089 0.035634 *
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

treat coefficient = estimated ATT

```
# balance matched dataset
m.data2 <- match.data(m.out2)

fit2 <- lm(re78 ~ treat + age + educ , #+ race + married + nodegree + re74 + re75,
            data = m.data2, weights = weights)

coeftest(fit2, vcov. = vcovCL, cluster = ~subclass)
#>
#> t test of coefficients:
```

```
#>
#>           Estimate Std. Error t value Pr(>|t|)
#> (Intercept) -588.890   2890.500 -0.2037 0.838630
#> treat        -380.494    685.482 -0.5551 0.579047
#> age          169.097     53.592  3.1553 0.001682 **
#> educ         285.433    206.986  1.3790 0.168400
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When reporting, remember to mention

1. the matching specification (method, and additional options)
2. the distance measure (e.g., propensity score)
3. other methods, and rationale for the final chosen method.
4. balance statistics of the matched dataset.
5. number of matched, unmatched, discarded
6. estimation method for treatment effect.

26.2 MatchingFrontier

As mentioned in `MatchIt`, you have to make trade-off (also known as bias-variance trade-off) between balance and sample size. An automated procedure to optimize this trade-off is implemented in `MatchingFrontier` (King et al., 2016), which solves this joint optimization problem.

I follow `MatchingFrontier` guide

```
# library(devtools)
# install_github('ChristopherLucas/MatchingFrontier')
library(MatchingFrontier)
data("lalonde")
# choose var to match on
match.on <- colnames(lalonde)[!(colnames(lalonde) %in% c('re78', 'treat'))]
match.on
#> [1] "age"      "education" "black"      "hispanic"  "married"   "nodegree"
#> [7] "re74"     "re75"
# Mahalanobis frontier (default)
mahal.frontier <-
  makeFrontier(
    dataset = lalonde,
    treatment = "treat",
    match.on = match.on
```

```

    )
#> Calculating Mahalanobis distances...
#> Calculating theoretical frontier...
#> Calculating information for plotting the frontier...
mahal.frontier
#> An imbalance frontier with 997 points.

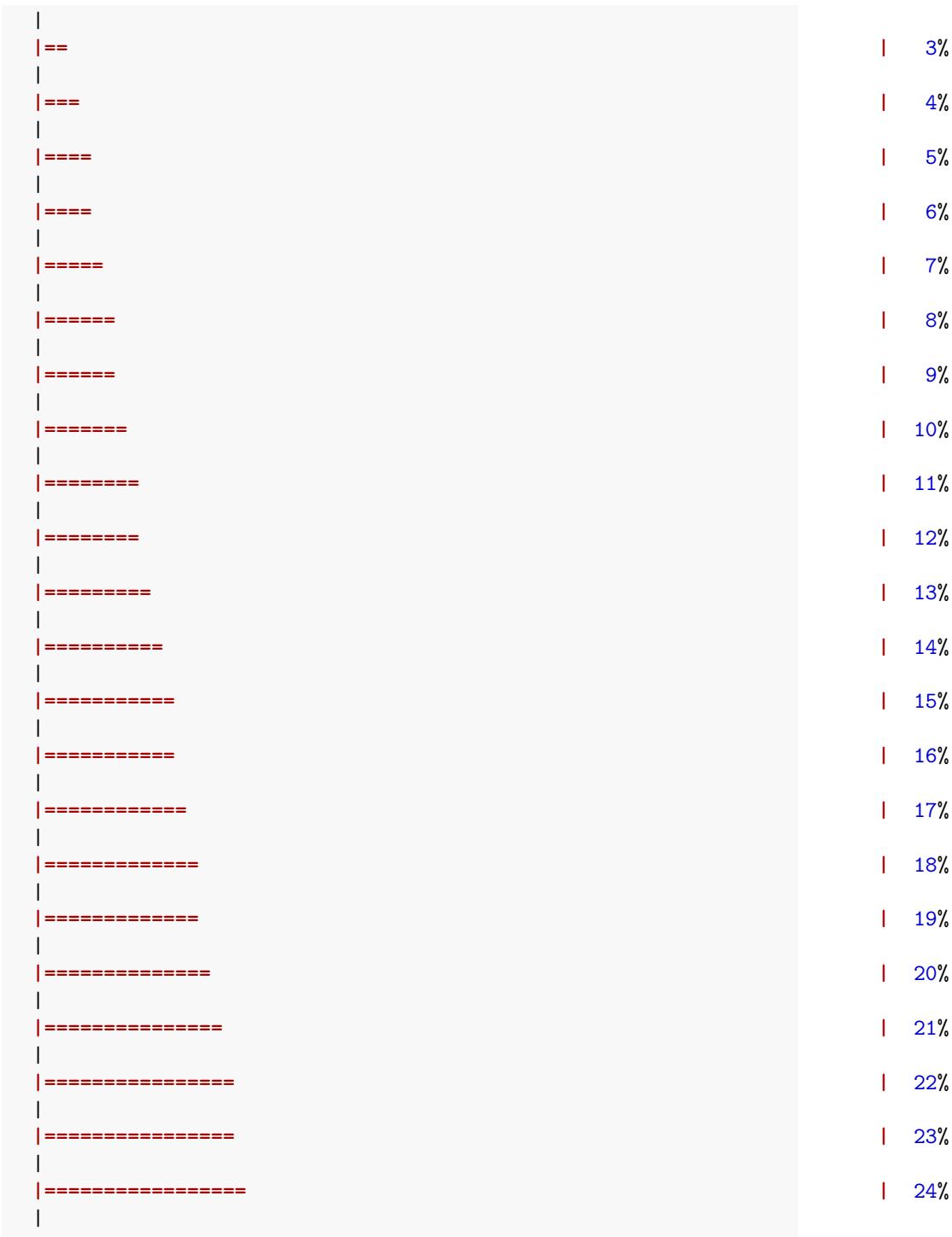
# L1 frontier
L1.frontier <-
  makeFrontier(
    dataset = lalonde,
    treatment = 'treat',
    match.on = match.on,
    QOI = 'SATT',
    metric = 'L1',
    ratio = 'fixed'
  )
#> Calculating L1 binnings...
#> Calculating L1 frontier... This may take a few minutes...
L1.frontier
#> An imbalance frontier with 976 points.

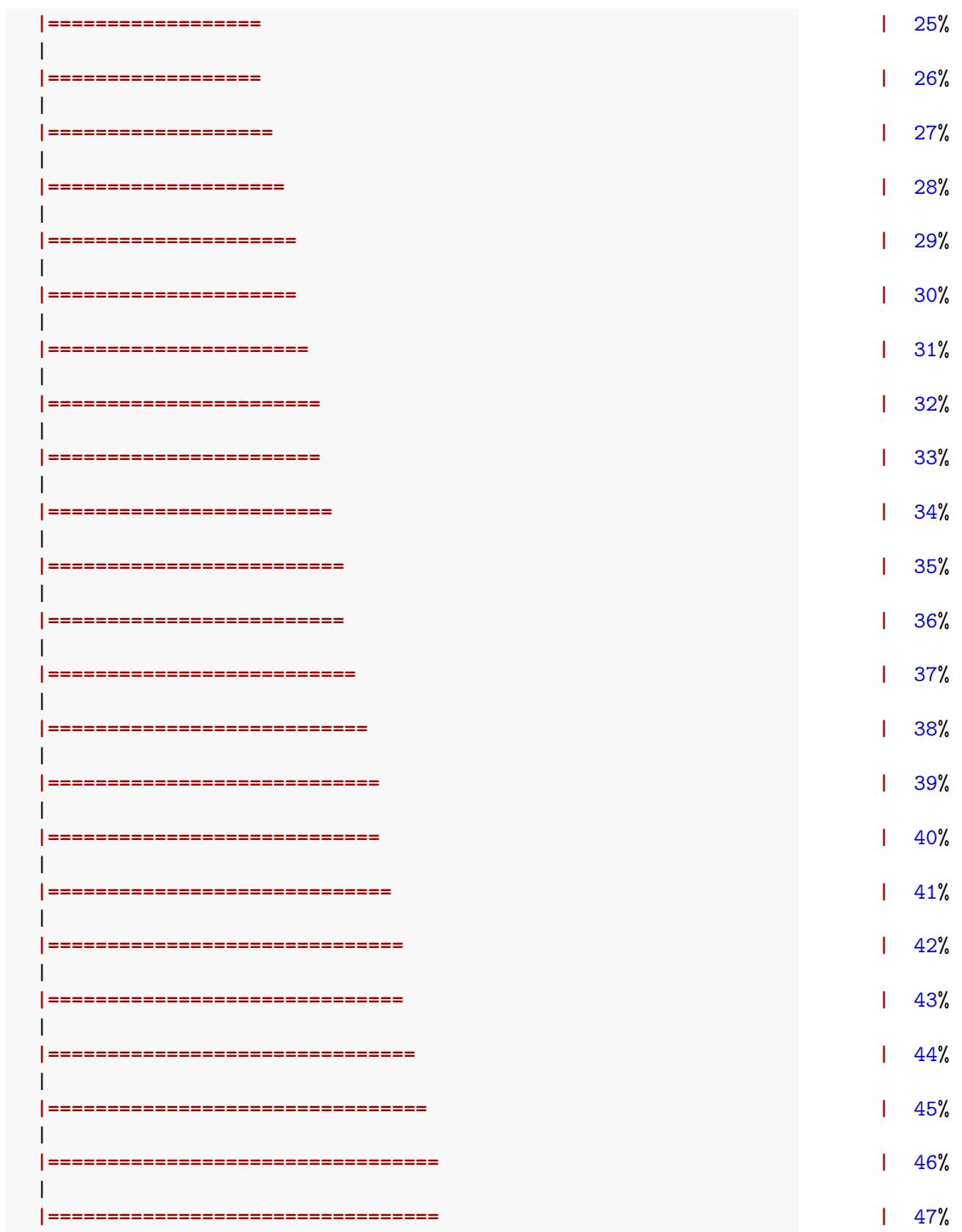
# estimate effects along the frontier

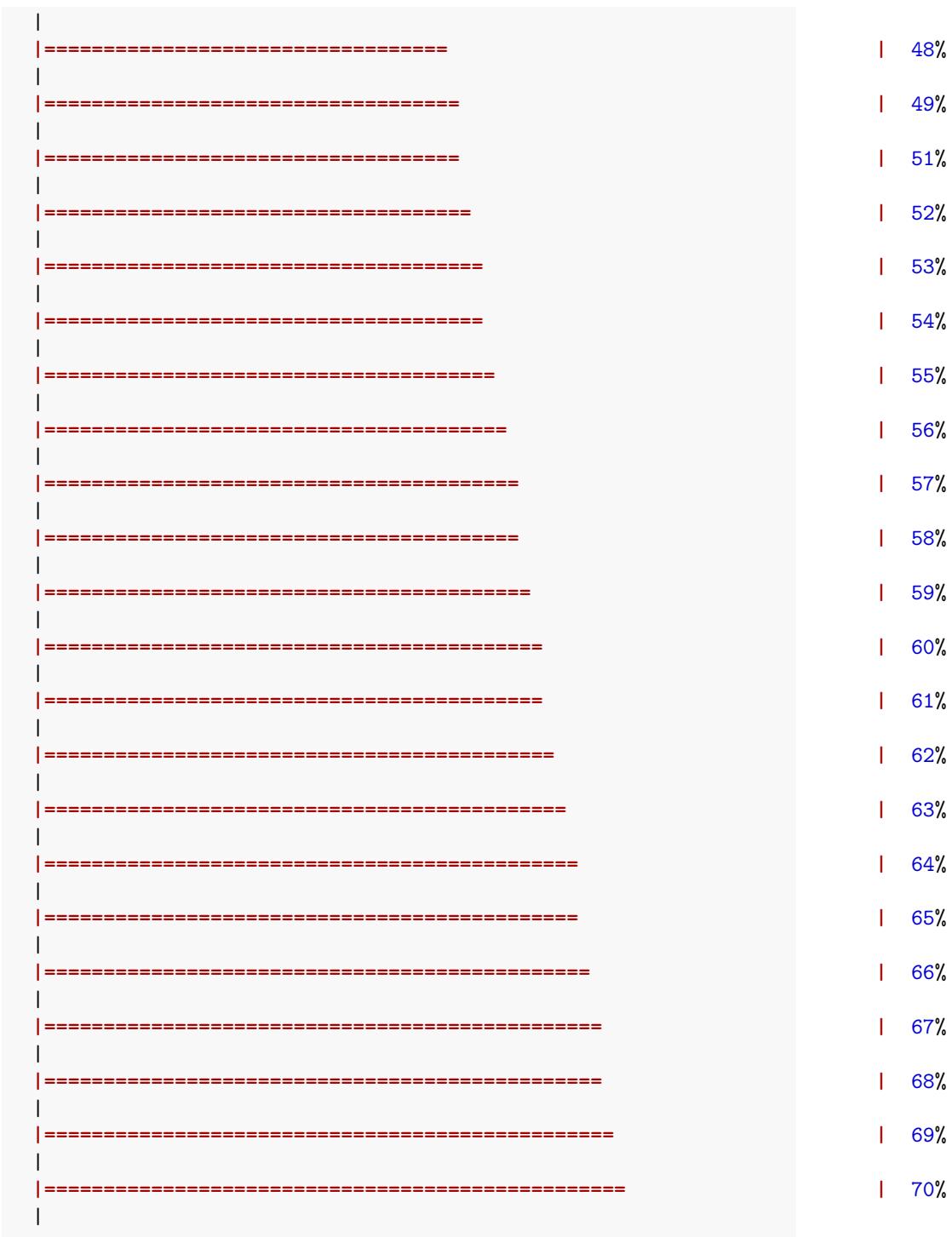
# Set base form
my.form <-
  as.formula(re78 ~ treat + age + black + education + hispanic + married + nodegree - 1)

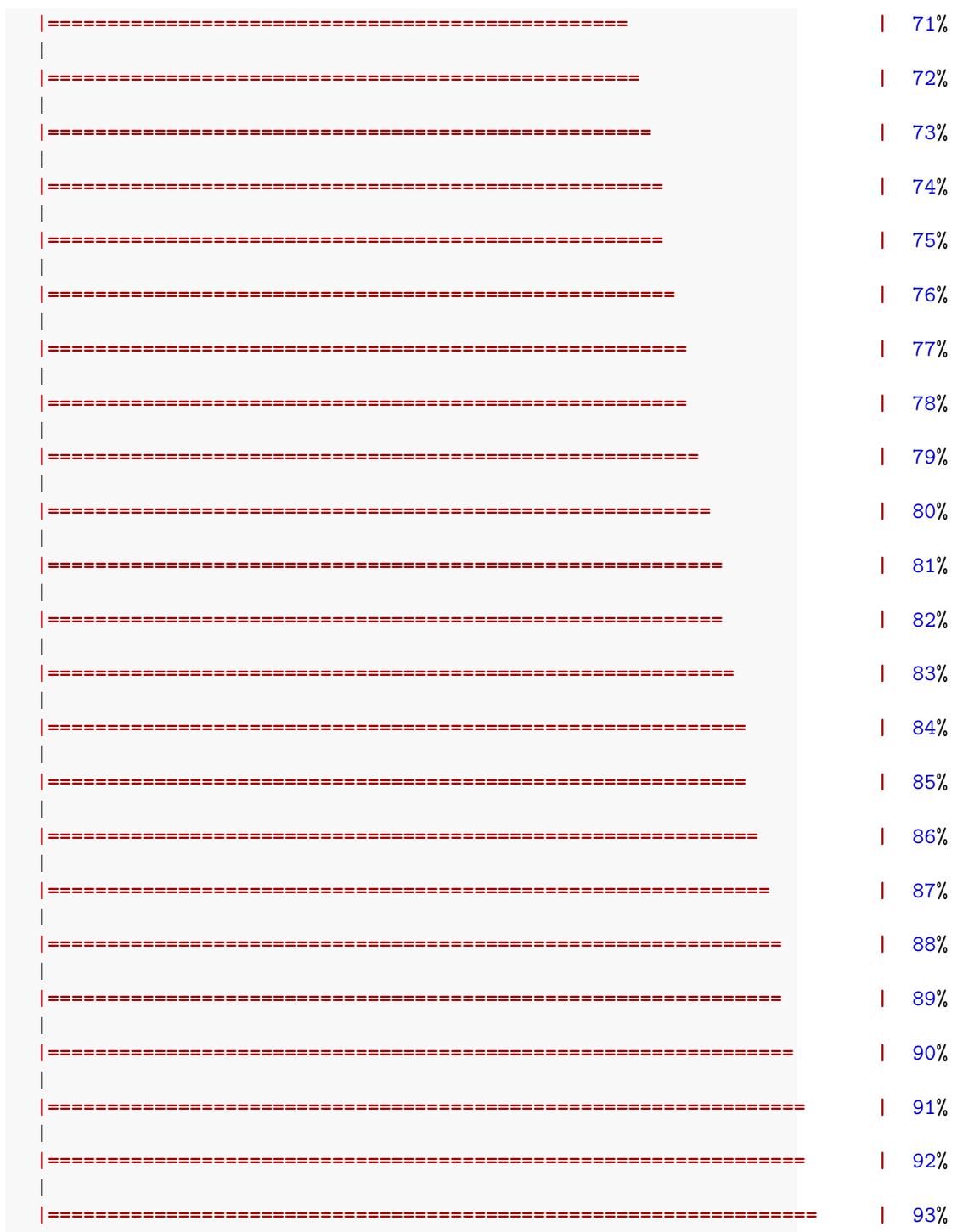
# Estimate effects for the mahalanobis frontier
mahal.estimates <-
  estimateEffects(
    mahal.frontier,
    're78 ~ treat',
    mod.dependence.formula = my.form,
    continuous.vars = c('age', 'education', 're74', 're75'),
    prop.estimated = .1,
    means.as.cutpoints = TRUE
  )
#>
| |
| |
|= | 0%
| |
|= | 1%
| |
|= | 2%

```



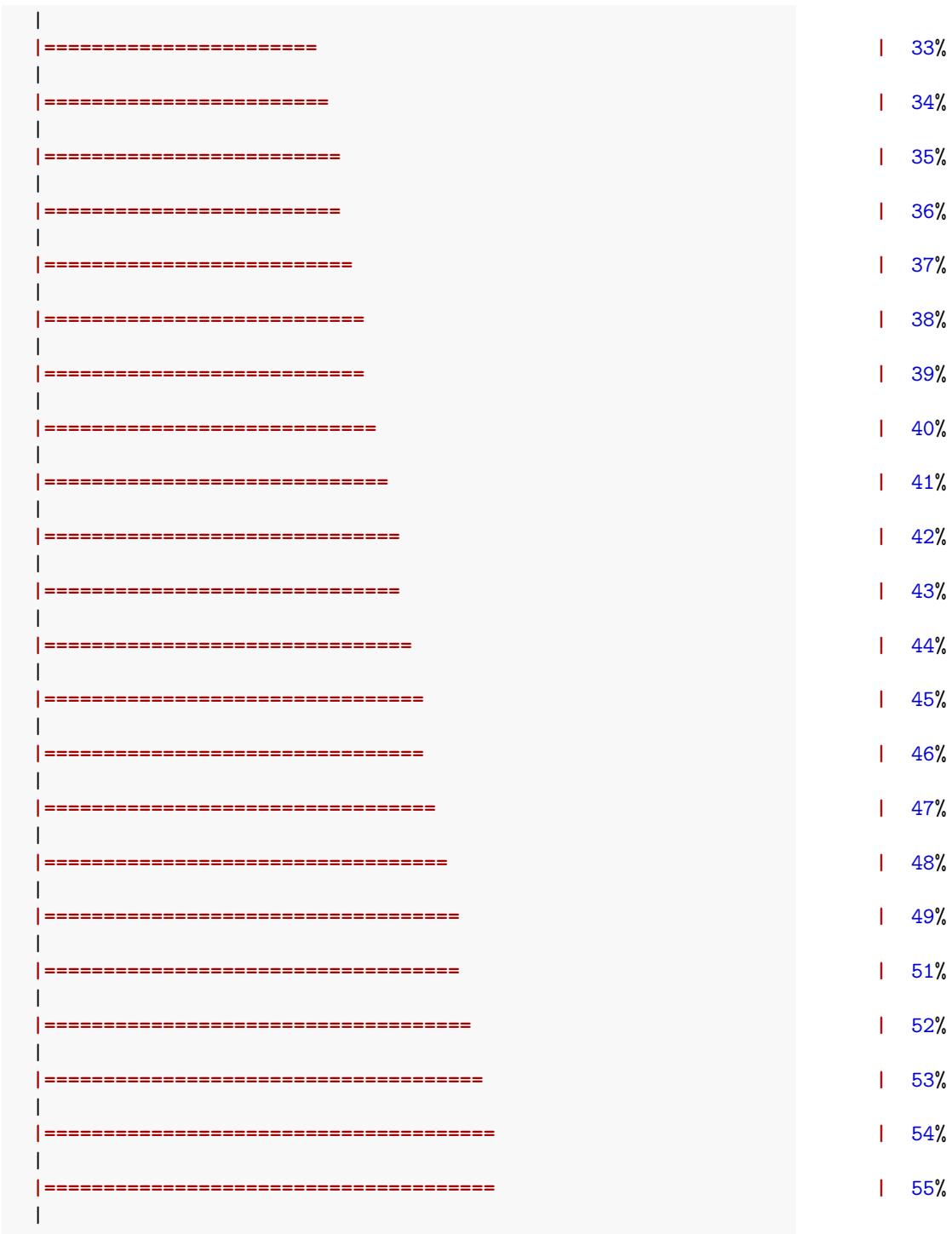


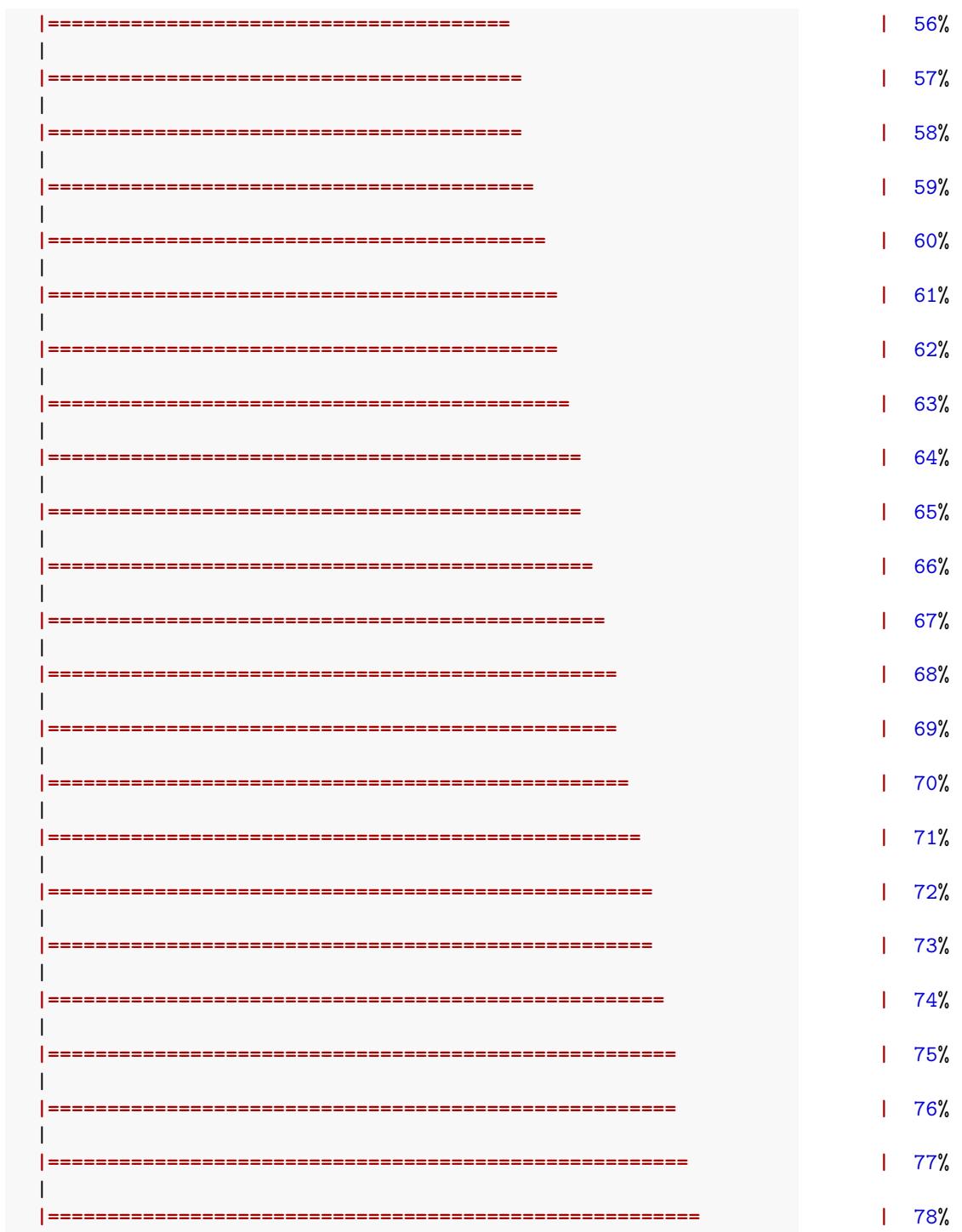


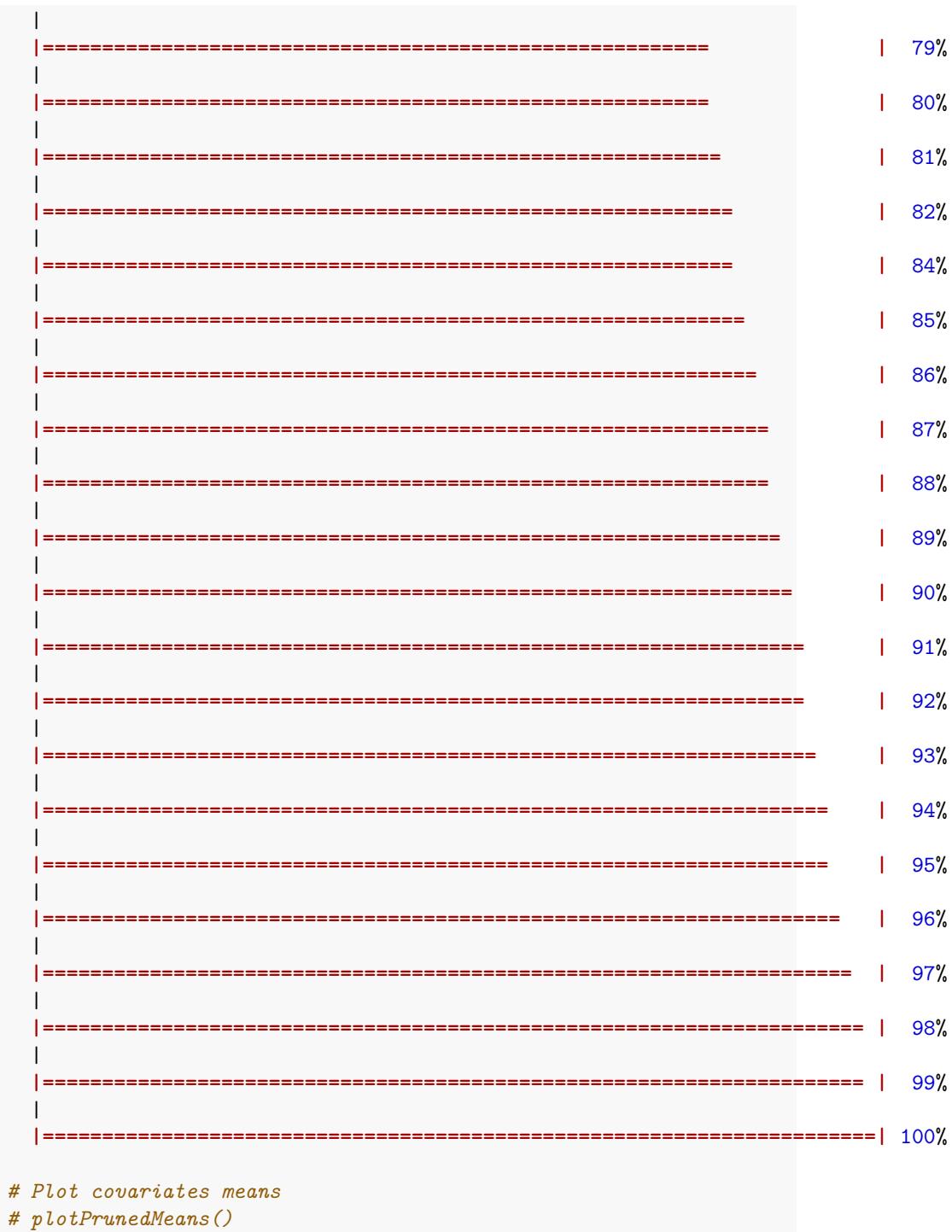


```
|  
|=====| 94%  
|  
|=====| 95%  
|  
|=====| 96%  
|  
|=====| 97%  
|  
|=====| 98%  
|  
|=====| 99%  
|  
|=====| 100%  
  
# Estimate effects for the L1 frontier  
L1.estimates <-  
  estimateEffects(  
    L1.frontier,  
    're78 ~ treat',  
    mod.dependence.formula = my.form,  
    continuous.vars = c('age', 'education', 're74', 're75'),  
    prop.estimated = .1,  
    means.as.cutpoints = TRUE  
  )  
#>  
|  
|=====| 0%  
|  
|=| 1%  
|  
|=| 2%  
|  
==| 3%  
|  
====| 4%  
|  
====| 5%  
|  
====| 6%  
|  
=====| 7%  
|  
=====| 8%  
|
```







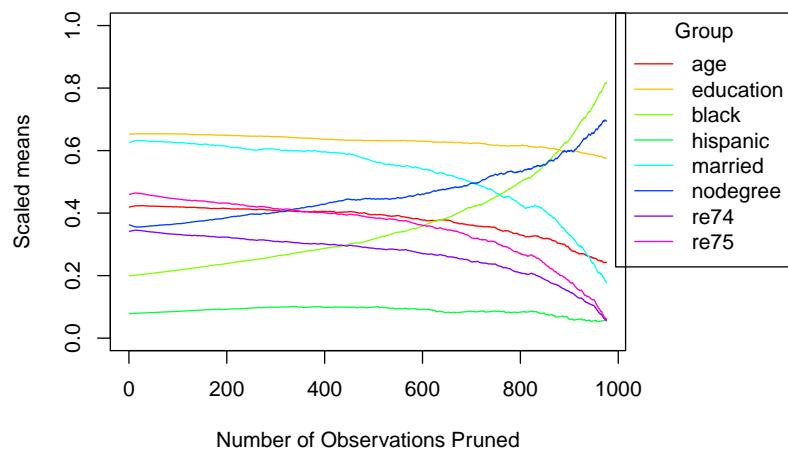


```

# Plot estimates (deprecated)
# plotEstimates(
#   L1.estimates,
#   ylim = c(-10000, 3000),
#   cex.lab = 1.4,
#   cex.axis = 1.4,
#   panel.first = grid(NULL, NULL, lwd = 2, )
# )

# Plot estimates
plotMeans(L1.frontier)

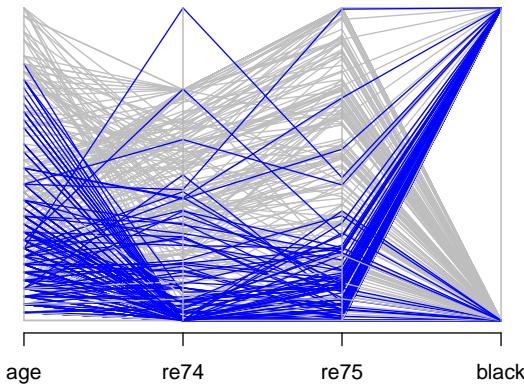
```



```

# parallel plot
parallelPlot(
  L1.frontier,
  N = 400,
  variables = c('age', 're74', 're75', 'black'),
  treated.col = 'blue',
  control.col = 'gray'
)

```



```
# export matched dataset
matched.data <- generateDataset(L1.frontier, N = 400) # take 400 units
```

26.3 Propensity Scores

Even though I mention the propensity scores matching method here, it is no longer recommended to use such method in research and publication (King and Nielsen, 2019) because it increases

- imbalance
- inefficiency
- model dependence: small changes in the model specification lead to big changes in model results
- bias

PSM tries to accomplish complete randomization while other methods try to achieve fully blocked. Hence, you probably better off use any other methods.

Propensity is “the probability of receiving the treatment given the observed covariates.” (Rosenbaum and Rubin, 1985)

Equivalently, it can be understood as the probability of being treated.

$$e_i(X_i) = P(T_i = 1|X_i)$$

Estimation using

- logistic regression
- Non parametric methods:
 - boosted CART
 - generalized boosted models (gbm)

Steps by Gary King's slides

- reduce k elements of X to scalar
- $\pi_i \equiv P(T_i = 1|X) = \frac{1}{1+e^{X_i\beta}}$
- Distance $(X_c, X_t) = |\pi_c - \pi_t|$
- match each treated unit to the nearest control unit
- control units: not reused; pruned if unused
- prune matches if distances > caliper

In the best case scenario, you randomly prune, which increases imbalance

Other methods dominate because they try to match exactly hence

- $X_c = X_t \rightarrow \pi_c = \pi_t$ (exact match leads to equal propensity scores) but
- $\pi_c = \pi_t \nrightarrow X_c = X_t$ (equal propensity scores do not necessarily lead to exact match)

Do not include/control for irrelevant covariates because it leads your PSM to be more random, hence more imbalance

What you left with after pruning is more important than what you start with then throw out.

Diagnostics:

- balance of the covariates
- no need to concern about collinearity
- can't use c-stat or stepwise because those model fit stat do not apply

26.4 Mahalanobis Distance

Approximates fully blocked experiment

$$\text{Distance } (X_c, X_t) = \sqrt{(X_c - X_t)' S^{-1} (X_c - X_t)}$$

where S^{-1} standardize the distance

In application we use Euclidean distance.

Prune unused control units, and prune matches if distance > caliper

26.5 Coarsened Exact Matching

Steps from Gray King's slides International Methods Colloquium talk 2015

- Temporarily coarsen X
- Apply exact matching to the coarsened X, C(X)
 - sort observation into strata, each with unique values of C(X)
 - prune stratum with 0 treated or 0 control units
- Pass on original (uncoarsened) units except those pruned

Properties:

- Monotonic imbalance bounding (MIB) matching method
 - maximum imbalance between the treated and control chosen ex ante
- meets congruence principle
- robust to measurement error
- can be implemented with multiple imputation
- works well for multi-category treatments

Assumptions:

- Ignorability (i.e., no omitted variable bias)

More detail in (Iacus et al., 2012)

Example by package's authors

```

library(cem)
data(LeLonde)

Le <- data.frame(na.omit(LeLonde)) # remove missing data
# treated and control groups
tr <- which(Le$treated==1)
ct <- which(Le$treated==0)
ntr <- length(tr)
nct <- length(ct)

# unadjusted, biased difference in means
mean(Le$re78[tr]) - mean(Le$re78[ct])
#> [1] 759.0479

# pre-treatment covariates
vars <-
  c(
    "age",
    "education",
    "black",
    "married",
    "nodegree",
    "re74",
    "re75",
    "hispanic",
    "u74",
    "u75",
    "q1"
  )

# overall imbalance statistics
imbalance(group=Le$treated, data=Le[vars]) # L1 = 0.902
#>
#> Multivariate Imbalance Measure: L1=0.902
#> Percentage of local common support: LCS=5.8%
#>
#> Univariate Imbalance Measures:
#>
#>          statistic   type      L1 min 25%      50%      75%
#> age      -0.252373042 (diff) 5.102041e-03  0  0  0.0000 -1.0000
#> education 0.153634710 (diff) 8.463851e-02  1  0  1.0000  1.0000
#> black     -0.010322734 (diff) 1.032273e-02  0  0  0.0000  0.0000
#> married    -0.009551495 (diff) 9.551495e-03  0  0  0.0000  0.0000
#> nodegree   -0.081217371 (diff) 8.121737e-02  0 -1  0.0000  0.0000
#> re74       -18.160446880 (diff) 5.551115e-17  0  0 284.0715 806.3452

```

```

#> re75      101.501761679 (diff) 5.551115e-17   0   0 485.6310 1238.4114
#> hispanic  -0.010144756 (diff) 1.014476e-02   0   0 0.0000 0.0000
#> u74       -0.045582186 (diff) 4.558219e-02   0   0 0.0000 0.0000
#> u75       -0.065555292 (diff) 6.555529e-02   0   0 0.0000 0.0000
#> q1        7.494021189 (Chi2) 1.067078e-01 NA NA NA NA
#>           max
#> age        -6.0000
#> education   1.0000
#> black       0.0000
#> married     0.0000
#> nodegree    0.0000
#> re74       -2139.0195
#> re75        490.3945
#> hispanic    0.0000
#> u74         0.0000
#> u75         0.0000
#> q1          NA

# drop other variables that are not pre-treatmentt matching variables
todrop <- c("treated", "re78")
imbalance(group=Le$treated, data=Le, drop=todrop)
#>
#> Multivariate Imbalance Measure: L1=0.902
#> Percentage of local common support: LCS=5.8%
#>
#> Univariate Imbalance Measures:
#>
#>           statistic type      L1 min 25%      50%      75%
#> age        -0.252373042 (diff) 5.102041e-03   0   0 0.0000 -1.0000
#> education  0.153634710 (diff) 8.463851e-02   1   0 1.0000 1.0000
#> black      -0.010322734 (diff) 1.032273e-02   0   0 0.0000 0.0000
#> married    -0.009551495 (diff) 9.551495e-03   0   0 0.0000 0.0000
#> nodegree   -0.081217371 (diff) 8.121737e-02   0  -1 0.0000 0.0000
#> re74       -18.160446880 (diff) 5.551115e-17   0   0 284.0715 806.3452
#> re75       101.501761679 (diff) 5.551115e-17   0   0 485.6310 1238.4114
#> hispanic  -0.010144756 (diff) 1.014476e-02   0   0 0.0000 0.0000
#> u74       -0.045582186 (diff) 4.558219e-02   0   0 0.0000 0.0000
#> u75       -0.065555292 (diff) 6.555529e-02   0   0 0.0000 0.0000
#> q1        7.494021189 (Chi2) 1.067078e-01 NA NA NA NA
#>           max
#> age        -6.0000
#> education   1.0000
#> black       0.0000
#> married     0.0000
#> nodegree    0.0000

```

```
#> re74      -2139.0195
#> re75      490.3945
#> hispanic   0.0000
#> u74       0.0000
#> u75       0.0000
#> q1        NA
```

automated coarsening

```
mat <- cem(treatment = "treated", data = Le, drop = "re78", keep.all=TRUE)
#>
#> Using 'treated'='1' as baseline group
mat
#>          G0  G1
#> All      392 258
#> Matched   95  84
#> Unmatched 297 174

# mat$w
```

coarsening by explicit user choice

```
# categorial variables
levels(Le$q1) # grouping option
#> [1] "agree"           "disagree"         "neutral"
#> [4] "no opinion"      "strongly agree"   "strongly disagree"
q1.grp <- list(c("strongly agree", "agree"), c("neutral", "no opinion"), c("strongly d

# continuous variables
table(Le$education)
#>
#>    3   4   5   6   7   8   9   10  11  12  13  14  15
#>    1   5   4   6  12  55 106 146 173 113  19   9   1
educut <- c(0, 6.5, 8.5, 12.5, 17) # use cutpoints

mat1 <- cem(treatment = "treated", data = Le, drop = "re78", cutpoints = list(education
#>
#> Using 'treated'='1' as baseline group
mat1
#>          G0  G1
#> All      392 258
#> Matched   158 115
#> Unmatched 234 143
```

- Can also use progressive coarsening method to control the number of matches.
- `cem` can also handle some missingness.

26.6 Genetic Matching

- GM uses iterative checking process of propensity scores, which combines propensity scores and Mahalanobis distance.
- GM is arguably “superior” method than nearest neighbor or full matching in imbalanced data
- Use a genetic search algorithm to find weights for each covariate such that we have optimal balance.
- Implementation
 - could use *with replacement*
 - balance can be based on
 - * paired t-tests (dichotomous variables)
 - * Kolmogorov-Smirnov (multinomial and continuous)

Packages

`Matching`

```
library(Matching)
data(lalonde)
attach(lalonde)

#The covariates we want to match on
X = cbind(age, educ, black, hisp, married, nodegr, u74, u75, re75, re74)

#The covariates we want to obtain balance on
BalanceMat <- cbind(age, educ, black, hisp, married, nodegr, u74, u75, re75, re74,
                      I(re74*re75))

#
#Let's call GenMatch() to find the optimal weight to give each
#covariate in 'X' so as we have achieved balance on the covariates in
#'BalanceMat'. This is only an example so we want GenMatch to be quick
#so the population size has been set to be only 16 via the 'pop.size'
#option. This is *WAY* too small for actual problems.
```

```
#For details see http://sekhon.berkeley.edu/papers/MatchingJSS.pdf.
#
genout <- GenMatch(Tr=treat, X=X, BalanceMatrix=BalanceMat, estimand="ATE", M=1,
                     pop.size=16, max.generations=10, wait.generations=1)
#>
#>
#> Thu Jun 09 22:36:19 2022
#> Domains:
#> 0.000000e+00 <= X1 <= 1.000000e+03
#> 0.000000e+00 <= X2 <= 1.000000e+03
#> 0.000000e+00 <= X3 <= 1.000000e+03
#> 0.000000e+00 <= X4 <= 1.000000e+03
#> 0.000000e+00 <= X5 <= 1.000000e+03
#> 0.000000e+00 <= X6 <= 1.000000e+03
#> 0.000000e+00 <= X7 <= 1.000000e+03
#> 0.000000e+00 <= X8 <= 1.000000e+03
#> 0.000000e+00 <= X9 <= 1.000000e+03
#> 0.000000e+00 <= X10 <= 1.000000e+03
#>
#> Data Type: Floating Point
#> Operators (code number, name, population)
#> (1) Cloning..... 1
#> (2) Uniform Mutation..... 2
#> (3) Boundary Mutation..... 2
#> (4) Non-Uniform Mutation..... 2
#> (5) Polytope Crossover..... 2
#> (6) Simple Crossover..... 2
#> (7) Whole Non-Uniform Mutation..... 2
#> (8) Heuristic Crossover..... 2
#> (9) Local-Minimum Crossover..... 0
#>
#> SOFT Maximum Number of Generations: 10
#> Maximum Nonchanging Generations: 1
#> Population size : 16
#> Convergence Tolerance: 1.000000e-03
#>
#> Not Using the BFGS Derivative Based Optimizer on the Best Individual Each Generation
#> Not Checking Gradients before Stopping.
#> Using Out of Bounds Individuals.
#>
#> Maximization Problem.
#> GENERATION: 0 (initializing the population)
#> Lexical Fit..... 1.747466e-01 1.795176e-01 1.795176e-01 2.755332e-01 3.173114e-01
#> #unique..... 16, #Total UniqueCount: 16
#> var 1:
```

```

#> best..... 3.758091e+02
#> mean..... 4.897395e+02
#> variance..... 9.320437e+04
#> var 2:
#> best..... 4.944768e+02
#> mean..... 5.218593e+02
#> variance..... 6.830606e+04
#> var 3:
#> best..... 1.880547e+02
#> mean..... 4.838327e+02
#> variance..... 1.030481e+05
#> var 4:
#> best..... 8.125678e+02
#> mean..... 4.445995e+02
#> variance..... 8.135038e+04
#> var 5:
#> best..... 9.058067e+02
#> mean..... 4.753729e+02
#> variance..... 1.184631e+05
#> var 6:
#> best..... 1.733063e+02
#> mean..... 4.782400e+02
#> variance..... 8.948808e+04
#> var 7:
#> best..... 5.766096e+02
#> mean..... 4.722599e+02
#> variance..... 7.199369e+04
#> var 8:
#> best..... 3.736603e+02
#> mean..... 4.108310e+02
#> variance..... 8.454007e+04
#> var 9:
#> best..... 5.987977e+02
#> mean..... 3.762504e+02
#> variance..... 7.462591e+04
#> var 10:
#> best..... 5.352480e+02
#> mean..... 4.491692e+02
#> variance..... 8.739694e+04
#>
#> GENERATION: 1
#> Lexical Fit..... 1.747466e-01 1.795176e-01 1.795176e-01 2.755332e-01 3.173114e-01 3.173114e-01
#> #unique..... 12, #Total UniqueCount: 28
#> var 1:
#> best..... 3.758091e+02

```

```

#> mean..... 3.556052e+02
#> variance..... 6.953381e+03
#> var 2:
#> best..... 4.944768e+02
#> mean..... 5.087018e+02
#> variance..... 3.156408e+03
#> var 3:
#> best..... 1.880547e+02
#> mean..... 1.798786e+02
#> variance..... 4.539894e+03
#> var 4:
#> best..... 8.125678e+02
#> mean..... 6.669235e+02
#> variance..... 5.540181e+04
#> var 5:
#> best..... 9.058067e+02
#> mean..... 9.117080e+02
#> variance..... 3.397508e+03
#> var 6:
#> best..... 1.733063e+02
#> mean..... 2.041563e+02
#> variance..... 3.552940e+04
#> var 7:
#> best..... 5.766096e+02
#> mean..... 4.995566e+02
#> variance..... 2.585541e+04
#> var 8:
#> best..... 3.736603e+02
#> mean..... 5.068097e+02
#> variance..... 4.273372e+04
#> var 9:
#> best..... 5.987977e+02
#> mean..... 4.917595e+02
#> variance..... 3.084008e+04
#> var 10:
#> best..... 5.352480e+02
#> mean..... 5.445554e+02
#> variance..... 8.587358e+03
#>
#> GENERATION: 2
#> Lexical Fit..... 1.747466e-01 1.795176e-01 1.795176e-01 2.755332e-01 3.173114e-01
#> #unique..... 8, #Total UniqueCount: 36
#> var 1:
#> best..... 3.758091e+02
#> mean..... 3.562379e+02

```

```

#> variance..... 2.385358e+03
#> var 2:
#> best..... 4.944768e+02
#> mean..... 4.956205e+02
#> variance..... 1.566415e+02
#> var 3:
#> best..... 1.880547e+02
#> mean..... 2.166244e+02
#> variance..... 6.801191e+03
#> var 4:
#> best..... 8.125678e+02
#> mean..... 8.059555e+02
#> variance..... 6.565662e+02
#> var 5:
#> best..... 9.058067e+02
#> mean..... 8.852994e+02
#> variance..... 3.056774e+03
#> var 6:
#> best..... 1.733063e+02
#> mean..... 2.210856e+02
#> variance..... 2.369334e+04
#> var 7:
#> best..... 5.766096e+02
#> mean..... 5.482967e+02
#> variance..... 4.613957e+03
#> var 8:
#> best..... 3.736603e+02
#> mean..... 3.943396e+02
#> variance..... 1.895797e+03
#> var 9:
#> best..... 5.987977e+02
#> mean..... 6.005851e+02
#> variance..... 3.308459e+02
#> var 10:
#> best..... 5.352480e+02
#> mean..... 5.444285e+02
#> variance..... 4.950778e+02
#>
#> 'wait.generations' limit reached.
#> No significant improvement in 1 generations.
#>
#> Solution Lexical Fitness Value:
#> 1.747466e-01 1.795176e-01 1.795176e-01 2.755332e-01 3.173114e-01 3.173114e-01 3.857502e-01
#>
#> Parameters at the Solution:

```

```

#>
#> X[ 1] : 3.758091e+02
#> X[ 2] : 4.944768e+02
#> X[ 3] : 1.880547e+02
#> X[ 4] : 8.125678e+02
#> X[ 5] : 9.058067e+02
#> X[ 6] : 1.733063e+02
#> X[ 7] : 5.766096e+02
#> X[ 8] : 3.736603e+02
#> X[ 9] : 5.987977e+02
#> X[10] : 5.352480e+02
#>
#> Solution Found Generation 1
#> Number of Generations Run 2
#>
#> Thu Jun 09 22:36:20 2022
#> Total run time : 0 hours 0 minutes and 1 seconds

#The outcome variable
Y=re78/1000

#
# Now that GenMatch() has found the optimal weights, let's estimate
# our causal effect of interest using those weights
#
mout <- Match(Y=Y, Tr=treat, X=X, estimand="ATE", Weight.matrix=genout)
summary(mout)
#>
#> Estimate... 1.9046
#> AI SE..... 0.82198
#> T-stat..... 2.3171
#> p.val..... 0.020498
#>
#> Original number of observations..... 445
#> Original number of treated obs..... 185
#> Matched number of observations..... 445
#> Matched number of observations (unweighted). 597

#
#Let's determine if balance has actually been obtained on the variables of interest
#
mb <- MatchBalance(treat~age +educ+black+ hisp+ married+ nodegr+ u74+ u75+
re75+ re74+ I(re74*re75),
match.out=mout, nboots=500)
#>
```

```

#> ***** (V1) age *****
#>                               Before Matching      After Matching
#> mean treatment.....      25.816          25.112
#> mean control.....       25.054          24.902
#> std mean diff.....     10.655          3.1286
#>
#> mean raw eQQ diff..... 0.94054        0.36181
#> med  raw eQQ diff.....  1                  0
#> max  raw eQQ diff..... 7                  8
#>
#> mean eCDF diff.....    0.025364       0.0099025
#> med  eCDF diff.....    0.022193       0.0075377
#> max  eCDF diff.....    0.065177       0.028476
#>
#> var ratio (Tr/Co)..... 1.0278         0.98286
#> T-test p-value.....    0.26594        0.27553
#> KS Bootstrap p-value.. 0.504           0.842
#> KS Naive p-value..... 0.7481          0.96884
#> KS Statistic.....     0.065177       0.028476
#>
#>
#> ***** (V2) educ *****
#>                               Before Matching      After Matching
#> mean treatment.....     10.346          10.189
#> mean control.....      10.088          10.222
#> std mean diff.....     12.806          -1.9653
#>
#> mean raw eQQ diff..... 0.40541        0.098827
#> med  raw eQQ diff.....  0                  0
#> max  raw eQQ diff..... 2                  2
#>
#> mean eCDF diff.....    0.028698       0.0070591
#> med  eCDF diff.....    0.012682       0.0033501
#> max  eCDF diff.....    0.12651         0.033501
#>
#> var ratio (Tr/Co)..... 1.5513          1.0458
#> T-test p-value.....    0.15017        0.38575
#> KS Bootstrap p-value.. 0.02             0.454
#> KS Naive p-value..... 0.062873        0.89105
#> KS Statistic.....     0.12651         0.033501
#>
#>
#> ***** (V3) black *****
#>                               Before Matching      After Matching
#> mean treatment.....     0.84324        0.8382

```

```

#> mean control..... 0.82692      0.8382
#> std mean diff..... 4.4767       0
#>
#> mean raw eQQ diff.... 0.016216     0
#> med  raw eQQ diff.... 0           0
#> max  raw eQQ diff.... 1           0
#>
#> mean eCDF diff..... 0.0081601    0
#> med  eCDF diff..... 0.0081601    0
#> max  eCDF diff..... 0.01632      0
#>
#> var ratio (Tr/Co).... 0.92503      1
#> T-test p-value..... 0.64736      1
#>
#>
#> ***** (V4) hisp *****
#>                                     Before Matching   After Matching
#> mean treatment..... 0.059459      0.08764
#> mean control..... 0.10769       0.08764
#> std mean diff..... -20.341       0
#>
#> mean raw eQQ diff.... 0.048649     0
#> med  raw eQQ diff.... 0           0
#> max  raw eQQ diff.... 1           0
#>
#> mean eCDF diff..... 0.024116     0
#> med  eCDF diff..... 0.024116     0
#> max  eCDF diff..... 0.048233     0
#>
#> var ratio (Tr/Co).... 0.58288      1
#> T-test p-value..... 0.064043     1
#>
#>
#> ***** (V5) married *****
#>                                     Before Matching   After Matching
#> mean treatment..... 0.18919       0.16854
#> mean control..... 0.15385       0.16854
#> std mean diff..... 8.9995        0
#>
#> mean raw eQQ diff.... 0.037838     0
#> med  raw eQQ diff.... 0           0
#> max  raw eQQ diff.... 1           0
#>
#> mean eCDF diff..... 0.017672     0
#> med  eCDF diff..... 0.017672     0

```

```

#> max eCDF diff..... 0.035343          0
#>
#> var ratio (Tr/Co).... 1.1802           1
#> T-test p-value..... 0.33425          1
#>
#>
#> ***** (V6) nodegr *****
#>                               Before Matching      After Matching
#> mean treatment..... 0.70811          0.78876
#> mean control..... 0.83462          0.78652
#> std mean diff..... -27.751          0.54991
#>
#> mean raw eQQ diff.... 0.12432          0.001675
#> med raw eQQ diff.... 0                   0
#> max raw eQQ diff.... 1                   1
#>
#> mean eCDF diff..... 0.063254         0.00083752
#> med eCDF diff..... 0.063254         0.00083752
#> max eCDF diff..... 0.12651          0.001675
#>
#> var ratio (Tr/Co).... 1.4998           0.9923
#> T-test p-value..... 0.0020368        0.73901
#>
#>
#> ***** (V7) u74 *****
#>                               Before Matching      After Matching
#> mean treatment..... 0.70811          0.73258
#> mean control..... 0.75                0.73034
#> std mean diff..... -9.1895          0.50714
#>
#> mean raw eQQ diff.... 0.037838         0.001675
#> med raw eQQ diff.... 0                   0
#> max raw eQQ diff.... 1                   1
#>
#> mean eCDF diff..... 0.020946         0.00083752
#> med eCDF diff..... 0.020946         0.00083752
#> max eCDF diff..... 0.041892          0.001675
#>
#> var ratio (Tr/Co).... 1.1041           0.99472
#> T-test p-value..... 0.33033          0.31731
#>
#>
#> ***** (V8) u75 *****
#>                               Before Matching      After Matching
#> mean treatment..... 0.6                 0.64494

```

```

#> mean control..... 0.68462 0.65169
#> std mean diff..... -17.225 -1.4072
#>
#> mean raw eQQ diff.... 0.081081 0.0050251
#> med raw eQQ diff.... 0 0
#> max raw eQQ diff.... 1 1
#>
#> mean eCDF diff..... 0.042308 0.0025126
#> med eCDF diff..... 0.042308 0.0025126
#> max eCDF diff..... 0.084615 0.0050251
#>
#> var ratio (Tr/Co).... 1.1133 1.0088
#> T-test p-value..... 0.068031 0.17952
#>
#>
#> ***** (V9) re75 *****
#> Before Matching After Matching
#> mean treatment..... 1532.1 1300.3
#> mean control..... 1266.9 1317.3
#> std mean diff..... 8.2363 -0.5676
#>
#> mean raw eQQ diff.... 367.61 104.16
#> med raw eQQ diff.... 0 0
#> max raw eQQ diff.... 2110.2 2510.6
#>
#> mean eCDF diff..... 0.050834 0.0080924
#> med eCDF diff..... 0.061954 0.0067002
#> max eCDF diff..... 0.10748 0.021776
#>
#> var ratio (Tr/Co).... 1.0763 0.97978
#> T-test p-value..... 0.38527 0.8025
#> KS Bootstrap p-value.. 0.044 0.824
#> KS Naive p-value..... 0.16449 0.99891
#> KS Statistic..... 0.10748 0.021776
#>
#>
#> ***** (V10) re74 *****
#> Before Matching After Matching
#> mean treatment..... 2095.6 2019.8
#> mean control..... 2107 2106.4
#> std mean diff..... -0.23437 -1.768
#>
#> mean raw eQQ diff.... 487.98 243.25
#> med raw eQQ diff.... 0 0
#> max raw eQQ diff.... 8413 7870.3

```

```

#>
#> mean eCDF diff..... 0.019223      0.0083159
#> med eCDF diff..... 0.0158       0.0067002
#> max eCDF diff..... 0.047089      0.025126
#>
#> var ratio (Tr/Co).... 0.7381      0.85755
#> T-test p-value..... 0.98186      0.39373
#> KS Bootstrap p-value.. 0.556       0.596
#> KS Naive p-value..... 0.97023      0.99172
#> KS Statistic..... 0.047089      0.025126
#>
#>
#> ***** (V11) I(re74 * re75) *****
#>                               Before Matching      After Matching
#> mean treatment..... 13118591      12261673
#> mean control..... 14530303      14240665
#> std mean diff..... -2.7799      -4.2761
#>
#> mean raw eQQ diff.... 3278733      2602379
#> med raw eQQ diff.... 0             0
#> max raw eQQ diff.... 188160151     223801517
#>
#> mean eCDF diff..... 0.022723      0.0043797
#> med eCDF diff..... 0.014449      0.0033501
#> max eCDF diff..... 0.061019      0.011725
#>
#> var ratio (Tr/Co).... 0.69439      0.58474
#> T-test p-value..... 0.79058      0.17475
#> KS Bootstrap p-value.. 0.266       0.994
#> KS Naive p-value..... 0.81575      1
#> KS Statistic..... 0.061019      0.011725
#>
#>
#> Before Matching Minimum p.value: 0.0020368
#> Variable Name(s): nodegr Number(s): 6
#>
#> After Matching Minimum p.value: 0.17475
#> Variable Name(s): I(re74 * re75) Number(s): 11

```

26.7 Matching for time series-cross-section data

Examples: (SCHEVE and STASAVAGE, 2012) and (Acemoglu et al., 2014)

Materials from Imai et al.'s slides

Identification strategy:

- Within-unit over-time variation
- within-time across-units variation

Chapter 27

Interrupted Time Series

- Regression Discontinuity in Time
- Control for
 - Seasonable trends
 - Concurrent events
- Pros (Penfold and Zhang, 2013)
 - control for long-term trends
- Cons
 - Min of 8 data points before and 8 after an intervention
 - Multiple events hard to distinguish

Notes:

- For subgroup analysis (heterogeneity in effect size), see (Harper and Bruckner, 2017)

Example by Leihua Ye

```
# data preparation
set.seed(1)
CaseID = rep(1:100, 6)

# intervention
Intervention = c(rep(0, 300), rep(1, 300))
```

```
Outcome_Variable = c(rnorm(300), abs(rnorm(300) * 4))

mydata = cbind(CaseID, Intervention, Outcome_Variable)

mydata = as.data.frame(mydata)

#construct a simple OLS model
model = lm(Outcome_Variable ~ Intervention, data = mydata)
summary(model) # there is a signficant effect
#>
#> Call:
#> lm(formula = Outcome_Variable ~ Intervention, data = mydata)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -3.3050 -1.2315 -0.1734  0.8691 11.9185
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  0.03358   0.11021   0.305   0.761
#> Intervention 3.28903   0.15586  21.103 <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 1.909 on 598 degrees of freedom
#> Multiple R-squared:  0.4268, Adjusted R-squared:  0.4259
#> F-statistic: 445.3 on 1 and 598 DF,  p-value: < 2.2e-16
```

C. OTHER CONCERNS

Chapter 28

Endogeneity

Refresher

A general model framework

$$\mathbf{Y} = \mathbf{X} + \epsilon$$

where

- $\mathbf{Y} = n \times 1$
- $\mathbf{X} = n \times k$
- $\beta = k \times 1$
- $\epsilon = n \times 1$

Then, OLS estimates of coefficients are

$$\begin{aligned}\hat{\beta}_{OLS} &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'(\mathbf{X} + \epsilon)) \\ &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\beta + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\epsilon) \\ \hat{\beta}_{OLS} &\rightarrow \beta + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\epsilon)\end{aligned}$$

To have unbiased estimates, we have to get rid of the second part $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\epsilon)$

There are 2 conditions to achieve unbiased estimates:

1. $E(\epsilon|X) = 0$ (This is easy, putting an intercept can solve this issue)
2. $Cov(\mathbf{X}, \epsilon) = 0$ (This is the hard part)

We only care about omitted variable

Usually, the problem will stem Omitted Variables Bias, but we only care about omitted variable bias when

1. Omitted variables correlate with the variables we care about (X). If OMV does not correlate with X , we don't care, and random assignment makes this correlation goes to 0)
2. Omitted variables correlates with outcome/ dependent variable

There are more types of endogeneity listed below.

Types of endogeneity

1. Endogenous Treatment
 - Omitted Variables Bias
 - Motivation/choice
 - Ability/talent
 - Self-selection
 - Feedback Effect (Simultaneity): also known as bidirectionality
 - Reverse Causality: Substle difference from Simultaneity: Technically, two variables affect each other sequentially, but in a big enough time frame, (e.g., monthly, or yearly), our coefficient will be biased just like simultaneity.
 - Measurement Error
2. Endogenous Sample Selection

To deal with this problem, we have a toolbox (that has been mentioned in previous chapter 16)

Tools in a hierarchical order

1. Experimental Design: Randomized Control Trials (Gold standard): Tier 1
2. Quasi-experimental
 1. Regression Discontinuity Tier 1A
 2. Difference-In-Differences Tier 2
 3. Synthetic Control Tier 2A

4. Event Studies Tier 2B
5. Fixed Effects Estimator 12.4.2.2: Tier 3
6. Endogenous Treatment: mostly Instrumental Variable: Tier 3A
7. Matching Methods Tier 4
8. Interrupted Time Series Tier 4A
9. Endogenous Sample Selection 28.4: mostly Heckman's correction

Using control variables in regression is a “selection on observables” identification strategy.

In other words, if you believe you have an omitted variable, and you can measure it, including it in the regression model solves your problem. These uninterested variables are called control variables in your model.

However, this is rarely the case (because the problem is we don't have their measurements). Hence, we need more elaborate methods:

- Endogenous Treatment
- Endogenous Sample Selection

Before we get to methods that deal with bias arises from omitted variables, we consider cases where we do have measurements of a variable, but there is measurement error (bias).

28.1 Measurement Error

- Data error can stem from
 - Coding errors
 - Reporting errors

Two forms of measurement error:

1. Random (stochastic) (indeterminate error) (Classical Measurement Errors): noise or measurement errors do not show up in a consistent or predictable way.
2. Systematic (determinate error) (Non-classical Measurement Errors): When measurement error is consistent and predictable across observations.
 1. Instrument errors (e.g., faulty scale) -> calibration or adjustment

2. Method errors (e.g., sampling errors) -> better method development
+ study design
3. Human errors (e.g., judgement)

Usually the systematic measurement error is a bigger issue because it introduces “bias” into our estimates, while random error introduces noise into our estimates

- Noise -> regression estimate to 0
- Bias -> can pull estimate to upward or downward.

28.1.1 Classical Measurement Errors

28.1.1.1 Right-hand side

- Right-hand side measurement error: When the measurement is in the covariates, then we have the endogeneity problem.

Say you know the true model is

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

But you don't observe X_i , but you observe

$$\tilde{X}_i = X_i + e_i$$

which is known as classical measurement errors where we **assume** e_i is uncorrelated with X_i (i.e., $E(X_i e_i) = 0$)

Then, when you estimate your observed variables, you have (substitute X_i with $\tilde{X}_i - e_i$):

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 (\tilde{X}_i - e_i) + u_i \\ &= \beta_0 + \beta_1 \tilde{X}_i + u_i - \beta_1 e_i \\ &= \beta_0 + \beta_1 \tilde{X}_i + v_i \end{aligned}$$

In words, the measurement error in X_i is now a part of the error term in the regression equation v_i . Hence, we have an endogeneity bias.

Endogeneity arises when

$$\begin{aligned} E(\tilde{X}_i v_i) &= E((X_i + e_i)(u_i - \beta_1 e_i)) \\ &= -\beta_1 Var(e_i) \neq 0 \end{aligned}$$

Since \tilde{X}_i and e_i are positively correlated, then it leads to

- a negative bias in $\hat{\beta}_1$ if the true β_1 is positive
- a positive bias if β_1 is negative

In other words, measurement errors cause **attenuation bias**, which in turn pushes the coefficient towards 0

As $Var(e_i)$ increases or $\frac{Var(e_i)}{Var(\tilde{X})} \rightarrow 1$ then e_i is a random (noise) and $\beta_1 \rightarrow 0$
(random variable \tilde{X} should not have any relation to Y_i)

Technical note:

The size of the bias in the OLS-estimator is

$$\hat{\beta}_{OLS} = \frac{cov(\tilde{X}, Y)}{var(\tilde{X})} = \frac{cov(X + e, \beta X + u)}{var(X + e)}$$

then

$$plim \hat{\beta}_{OLS} = \beta \frac{\sigma_X^2}{\sigma_X^2 + \sigma_e^2} = \beta \lambda$$

where λ is **reliability** or signal-to-total variance ratio or attenuation factor

Reliability affect the extent to which measurement error attenuates $\hat{\beta}$. The attenuation bias is

$$\hat{\beta}_{OLS} - \beta = -(1 - \lambda)\beta$$

Thus, $\hat{\beta}_{OLS} < \beta$ (unless $\lambda = 1$, in which case we don't even have measurement error).

Note:

Data transformation worsen (magnify) the measurement error

$$y = \beta x + \gamma x^2 + \epsilon$$

then, the attenuation factor for $\hat{\gamma}$ is the square of the attenuation factor for $\hat{\beta}$
(i.e., $\lambda_{\hat{\gamma}} = \lambda_{\hat{\beta}}^2$)

Adding covariates increases attenuation bias

To fix classical measurement error problem, we can

1. Find estimates of either σ_X^2, σ_e^2 or λ from validation studies, or survey data.
2. Endogenous Treatment Use instrument Z correlated with X but uncorrelated with ϵ
3. Abandon your project

28.1.1.2 Left-hand side

When the measurement is in the outcome variable, econometricians or causal scientists do not care because they still have an unbiased estimate of the coefficients (the zero conditional mean assumption is not violated, hence we don't have endogeneity). However, statisticians might care because it might inflate our uncertainty in the coefficient estimates (i.e., higher standard errors).

$$\tilde{Y} = Y + v$$

then the model you estimate is

$$\tilde{Y} = \beta X + u + v$$

Since v is uncorrelated with X , then $\hat{\beta}$ is consistently estimated by OLS
If we have measurement error in Y_i , it will pass through β_1 and go to u_i

28.1.2 Non-classical Measurement Errors

Relaxing the assumption that X and ϵ are uncorrelated

Recall the true model we have true estimate is

$$\hat{\beta} = \frac{\text{cov}(X + \epsilon, \beta X + u)}{\text{var}(X + \epsilon)}$$

then without the above assumption, we have

$$\begin{aligned} plim \hat{\beta} &= \frac{\beta(\sigma_X^2 + \sigma_{X\epsilon})}{\sigma_X^2 + \sigma_\epsilon^2 + 2\sigma_{X\epsilon}} \\ &= \left(1 - \frac{\sigma_\epsilon^2 + \sigma_{X\epsilon}}{\sigma_X^2 + \sigma_\epsilon^2 + 2\sigma_{X\epsilon}}\right)\beta \\ &= (1 - b_{\epsilon\tilde{X}})\beta \end{aligned}$$

where $b_{\epsilon\tilde{X}}$ is the covariance between \tilde{X} and ϵ (also the regression coefficient of a regression of ϵ on \tilde{X})

Hence, the Classical Measurement Errors is just a special case of Non-classical Measurement Errors where $b_{\epsilon\tilde{X}} = 1 - \lambda$

So when $\sigma_{X\epsilon} = 0$ (Classical Measurement Errors), increasing this covariance $b_{\epsilon\tilde{X}}$ increases the covariance increases the attenuation factor if more than half

of the variance in \tilde{X} is measurement error, and decreases the attenuation factor otherwise. This is also known as **mean reverting measurement error** (Bound and Krueger, 1989)

A general framework for both right-hand side and left-hand side measurement error is (Bound et al., 1994):

consider the true model

$$\mathbf{Y} = \mathbf{X} +$$

then

$$\begin{aligned}\hat{\beta} &= (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{Y}} \\ &= (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'(\tilde{\mathbf{X}} - \mathbf{U} + \mathbf{v} +) \\ &= + (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'(-\mathbf{U} + \mathbf{v} +) \\ p\lim \hat{\beta} &= \beta + p\lim(\tilde{X}'\tilde{X})^{-1}\tilde{X}'(-U\beta + v) \\ &= \beta + p\lim(\tilde{X}'\tilde{X})^{-1}\tilde{X}'W \begin{bmatrix} -\beta \\ 1 \end{bmatrix}\end{aligned}$$

Since we collect the measurement errors in a matrix $W = [U|v]$, then

$$(-U\beta + v) = W \begin{bmatrix} -\beta \\ 1 \end{bmatrix}$$

Hence, in general, biases in the coefficients β are regression coefficients from regressing the measurement errors on the mis-measured \tilde{X}

Notes:

- Instrumental Variable can help fix this problem
- There can also be measurement error in dummy variables and you can still use Instrumental Variable to fix it.

28.1.3 Solution to Measurement Errors

28.1.3.1 Correlation

$$P(\rho|data) = \frac{P(data|\rho)P(\rho)}{P(data)}$$

Posterior Probability \propto Likelihood \times Prior Probability

where

- ρ is a correlation coefficient
- $P(data|\rho)$ is the likelihood function evaluated at ρ
- $P(\rho)$ prior probability
- $P(data)$ is the normalizing constant

With sample correlation coefficient r :

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Then the posterior density approximation of ρ is (Schisterman et al., 2003, pp.3)

$$P(\rho|x, y) \propto P(\rho) \frac{(1 - \rho^2)^{(n-1)/2}}{(1 - \rho \times r)^{n-(3/2)}}$$

where

- $\rho = \tanh \xi$ where $\xi \sim N(z, 1/n)$
- $r = \tanh z$

Then the posterior density follow a normal distribution where

Mean

$$\mu_{posterior} = \sigma_{posterior}^2 \times (n_{prior} \times \tanh^{-1} r_{prior} + n_{likelihood} \times \tanh^{-1} r_{likelihood})$$

variance

$$\sigma_{posterior}^2 = \frac{1}{n_{prior} + n_{Likelihood}}$$

To simplify the integration process, we choose prior that is

$$P(\rho) \propto (1 - \rho^2)^c$$

where

- c is the weight the prior will have in estimation (i.e., $c = 0$ if no prior info, hence $P(\rho) \propto 1$)

Example:

Current study: $r_{xy} = 0.5, n = 200$

Previous study: $r_{xy} = 0.2765, (n = 50205)$

Combining two, we have the posterior following a normal distribution with the **variance** of

$$\sigma_{posterior}^2 = \frac{1}{n_{prior} + n_{Likelihood}} = \frac{1}{200 + 50205} = 0.0000198393$$

Mean

$$\begin{aligned}\mu_{Posterior} &= \sigma_{Posterior}^2 \times (n_{prior} \times \tanh^{-1} r_{prior} + n_{likelihoood} \times \tanh^{-1} r_{likelihoood}) \\ &= 0.0000198393 \times (50205 \times \tanh^{-1} 0.2765 + 200 \times \tanh^{-1} 0.5) \\ &= 0.2849415\end{aligned}$$

Hence, $Posterior \sim N(0.691, 0.0009)$, which means the correlation coefficient is $\tanh(0.691) = 0.598$ and 95% CI is

$$\mu_{posterior} \pm 1.96 \times \sqrt{\sigma_{Posterior}^2} = 0.2849415 \pm 1.96 \times (0.0000198393)^{1/2} = (0.2762115, 0.2936714)$$

Hence, the interval for posterior ρ is $(0.2693952, 0.2855105)$

If future authors suspect that they have

1. Large sampling variation
2. Measurement error in either measures in the correlation, which attenuates the relationship between the two variables

Applying this Bayesian correction can give them a better estimate of the correlation between the two.

To implement this calculation in R, see below

```
n_new           <- 200
r_new           <- 0.5
alpha           <- 0.05

update_correlation <- function(n_new, r_new, alpha) {
  n_meta          <- 50205
  r_meta          <- 0.2765
```

```

# Variance
var_xi           <- 1 / (n_new + n_meta)
format(var_xi, scientific = FALSE)

# mean
mu_xi            <- var_xi * (n_meta * atanh(r_meta) + n_new * (atanh(r_new)))
format(mu_xi, scientific = FALSE)

# confidence interval
upper_xi          <- mu_xi + qnorm(1 - alpha / 2) * sqrt(var_xi)
lower_xi          <- mu_xi - qnorm(1 - alpha / 2) * sqrt(var_xi)

# rho
mean_rho          <- tanh(mu_xi)
upper_rho          <- tanh(upper_xi)
lower_rho          <- tanh(lower_xi)

# return a list
return(
  list(
    "mu_xi" = mu_xi,
    "var_xi" = var_xi,
    "upper_xi" = upper_xi,
    "lower_xi" = lower_xi,
    "mean_rho" = mean_rho,
    "upper_rho" = upper_rho,
    "lower_rho" = lower_rho
  )
)
}

# Old confidence interval
r_new + qnorm(1 - alpha / 2) * sqrt(1/n_new)
#> [1] 0.6385904
r_new - qnorm(1 - alpha / 2) * sqrt(1/n_new)
#> [1] 0.3614096

testing = update_correlation(n_new = n_new, r_new = r_new, alpha = alpha)

# Updated rho
testing$mean_rho
#> [1] 0.2774723

```

```
# Updated confidence interval
testing$upper_rho
#> [1] 0.2855105
testing$lower_rho
#> [1] 0.2693952
```

28.2 Simultaneity

- When independent variables (X 's) are jointly determined with the dependent variable Y , typically through an equilibrium mechanism, violates the second condition for causality (i.e., temporal order).
- Examples: quantity and price by demand and supply, investment and productivity, sales and advertisement

General Simultaneous (Structural) Equations

$$Y_i = \beta_0 + \beta_1 X_i + u_i X_i = \alpha_0 + \alpha_1 Y_i + v_i$$

Hence, the solutions are

$$Y_i = \frac{\beta_0 + \beta_1 \alpha_0}{1 - \alpha_1 \beta_1} + \frac{\beta_1 v_i + u_i}{1 - \alpha_1 \beta_1} X_i = \frac{\alpha_0 + \alpha_1 \beta_0}{1 - \alpha_1 \beta_1} + \frac{v_i + \alpha_1 u_i}{1 - \alpha_1 \beta_1}$$

If we run only one regression, we will have biased estimators (because of **simultaneity bias**):

$$\begin{aligned} Cov(X_i, u_i) &= Cov\left(\frac{v_i + \alpha_1 u_i}{1 - \alpha_1 \beta_1}, u_i\right) \\ &= \frac{\alpha_1}{1 - \alpha_1 \beta_1} Var(u_i) \end{aligned}$$

In an even more general model

$$\begin{cases} Y_i = \beta_0 + \beta_1 X_i + \beta_2 T_i + u_i \\ X_i = \alpha_0 + \alpha_1 Y_i + \alpha_2 Z_i + v_i \end{cases}$$

where

- X_i, Y_i are **endogenous** variables determined within the system
- T_i, Z_i are **exogenous** variables

Then, the reduced form of the model is

$$\begin{cases} Y_i = \frac{\beta_0 + \beta_1 \alpha_0}{1 - \alpha_1 \beta_1} + \frac{\beta_1 \alpha_2}{1 - \alpha_1 \beta_1} Z_i + \frac{\beta_2}{1 - \alpha_1 \beta_1} T_i + \tilde{u}_i \\ \quad = B_0 + B_1 Z_i + B_2 T_i + \tilde{u}_i \\ X_i = \frac{\alpha_0 + \alpha_1 \beta_0}{1 - \alpha_1 \beta_1} + \frac{\alpha_2}{1 - \alpha_1 \beta_1} Z_i + \frac{\alpha_1 \beta_2}{1 - \alpha_1 \beta_1} T_i + \tilde{v}_i \\ \quad = A_0 + A_1 Z_i + A_2 T_i + \tilde{v}_i \end{cases}$$

Then, now we can get consistent estimates of the reduced form parameters

And to get the original parameter estimates

$$\frac{B_1}{A_1} = \beta_1 B_2 \left(1 - \frac{B_1 A_2}{A_1 B_2}\right) = \beta_2 \frac{A_2}{B_2} = \alpha_1 A_1 \left(1 - \frac{B_1 A_2}{A_1 B_2}\right) = \alpha_2$$

Rules for Identification

Order Condition (necessary but not sufficient)

$$K - k \geq m - 1$$

where

- M = number of endogenous variables in the model
- K = number of exogenous variables in the model
- m = number of endogenous variables in a given
- k = is the number of exogenous variables in a given equation

This is actually the general framework for instrumental variables

28.3 Endogenous Treatment

Using the OLS estimates as a reference point

```
library(AER)
library(REndo)
set.seed(421)
data("CASchools")
school <- CASchools
school$stratio <- with(CASchools, students / teachers)
```

```

m1.ols <-
  lm(read ~ stratio + english + lunch + grades + income + calworks + county,
  data = school)
summary(m1.ols)$coefficients[1:7, ]
#>             Estimate Std. Error    t value    Pr(>|t|)
#> (Intercept) 683.45305948 9.56214469 71.4748711 3.011667e-218
#> stratio      -0.30035544 0.25797023 -1.1643027 2.450536e-01
#> english      -0.20550107 0.03765408 -5.4576041 8.871666e-08
#> lunch         -0.38684059 0.03700982 -10.4523759 1.427370e-22
#> gradesKK-08 -1.91291321 1.35865394 -1.4079474 1.599886e-01
#> income        0.71615378 0.09832843 7.2832829 1.986712e-12
#> calworks     -0.05273312 0.06154758 -0.8567863 3.921191e-01

```

28.3.1 Instrumental Variable

A3a requires ϵ_i to be uncorrelated with \mathbf{x}_i

Assume A1 , A2, A5

$$plim(\hat{\beta}_{OLS}) = \beta + [E(\mathbf{x}'_i \mathbf{x}_i)]^{-1} E(\mathbf{x}'_i \epsilon_i)$$

A3a is the weakest assumption needed for OLS to be **consistent**

A3 fails when x_{ik} is correlated with ϵ_i

- Omitted Variables Bias: ϵ_i includes any other factors that may influence the dependent variable (linearly)
- Simultaneity Demand and prices are simultaneously determined.
- Endogenous Sample Selection we did not have iid sample
- Measurement Error

Note

- Omitted Variable: an omitted variable is a variable, omitted from the model (but is in the ϵ_i) and unobserved has predictive power towards the outcome.
- Omitted Variable Bias: is the bias (and inconsistency when looking at large sample properties) of the OLS estimator when the omitted variable.
- We can have both positive and negative selection bias (it depends on what our story is)

The **structural equation** is used to emphasize that we are interested understanding a **causal relationship**

$$y_{i1} = \beta_0 + \mathbf{z}_{i1}\beta_1 + y_{i2}\beta_2 + \epsilon_i$$

where

- y_{it} is the outcome variable (inherently correlated with ϵ_i)
- y_{i2} is the endogenous covariate (presumed to be correlated with ϵ_i)
- β_1 represents the causal effect of y_{i2} on y_{i1}
- \mathbf{z}_{i1} is exogenous controls (uncorrelated with ϵ_i) ($E(z'_{1i}\epsilon_i) = 0$)

OLS is an inconsistent estimator of the causal effect β_2

If there was no endogeneity

- $E(y'_{i2}\epsilon_i) = 0$
- the exogenous variation in y_{i2} is what identifies the causal effect

If there is endogeneity

- Any wiggle in y_{i2} will shift simultaneously with ϵ_i

$$plim(\hat{\beta}_{OLS}) = \beta + [E(\mathbf{x}'_{i1}\mathbf{x}_{i1})]^{-1}E(\mathbf{x}'_{i1}\epsilon_i)$$

where

- β is the causal effect
- $[E(\mathbf{x}'_{i1}\mathbf{x}_{i1})]^{-1}E(\mathbf{x}'_{i1}\epsilon_i)$ is the endogenous effect

Hence $\hat{\beta}_{OLS}$ can be either more positive and negative than the true causal effect.

Motivation for **Two Stage Least Squares (2SLS)**

$$y_{i1} = \beta_0 + \mathbf{z}_{i1}\beta_1 + y_{i2}\beta_2 + \epsilon_i$$

We want to understand how movement in y_{i2} effects movement in y_{i1} , but whenever we move y_{i2} , ϵ_i also moves.

Solution

We need a way to move y_{i2} independently of ϵ_i , then we can analyze the response in y_{i1} as a causal effect

- Find an **instrumental variable(s)** z_{i2}
 - Instrument **Relevance**: when** z_{i2} moves then y_{i2} also moves

- Instrument **Exogeneity**: when z_{i2} moves then ϵ_i does not move.
- z_{i2} is the **exogenous variation that identifies** the causal effect β_2

Finding an Instrumental variable:

- Random Assignment: + Effect of class size on educational outcomes: instrument is initial random
- Relation's Choice + Effect of Education on Fertility: instrument is parent's educational level
- Eligibility + Trade-off between IRA and 401K retirement savings: instrument is 401k eligibility

Example

Return to College

- education is correlated with ability - endogenous
- **Near 4year** as an instrument
 - Instrument Relevance: when **near** moves then education also moves
 - Instrument Exogeneity: when **near** moves then ϵ_i does not move.
- Other potential instruments; near a 2-year college. Parent's Education. Owning Library Card

$$y_{i1} = \beta_0 + z_{i1}\beta_1 + y_{i2}\beta_2 + \epsilon_i$$

First Stage (Reduced Form) Equation:

$$y_{i2} = \pi_0 + z_{i1} \pi_1 + z_{i2} \pi_2 + v_i$$

where

- $\pi_0 + z_{i1} \pi_1 + z_{i2} \pi_2$ is exogenous variation v_i is endogenous variation

This is called a **reduced form equation**

- Not interested in the causal interpretation of π_1 or π_2
- A linear projection of z_{i1} and z_{i2} on y_{i2} (simple correlations)
- The projections π_1 and π_2 guarantee that $E(z'_{i1}v_i) = 0$ and $E(z'_{i2}v_i) = 0$

Instrumental variable z_{i2}

- **Instrument Relevance:** $\pi_2 \neq 0$
- **Instrument Exogeneity:** $E(\mathbf{z}_{i2} | \mathbf{x}) = 0$

Moving only the exogenous part of y_{i2} is moving

$$\tilde{y}_{i2} = \pi_0 + \mathbf{z}_{i1} \cdot \mathbf{1} + \mathbf{z}_{i2} \cdot \mathbf{2}$$

two Stage Least Squares (2SLS)

$$y_{i1} = \beta_0 + \mathbf{z}_{i1} \cdot \mathbf{1} + y_{i2}\beta_2 + \epsilon_i$$

$$y_{i2} = \pi_0 + \mathbf{z}_{i2} \cdot \mathbf{2} + \mathbf{v}_i$$

Equivalently,

$$y_{i1} = \beta_0 + \mathbf{z}_{i1}\beta_1 + \tilde{y}_{i2}\beta_2 + u_i \quad (28.1)$$

where

- $\tilde{y}_{i2} = \pi_0 + \mathbf{z}_{i2} \cdot \mathbf{2}$
- $u_i = v_i\beta_2 + \epsilon_i$

The (28.1) holds for A1, A5

- A2 holds if the instrument is relevant $\pi_2 \neq 0$ + $y_{i1} = \beta_0 + \mathbf{z}_{i1} \cdot \mathbf{1} + (\mathbf{0} + \mathbf{z}_{i1} \cdot \mathbf{1} + \mathbf{z}_{i2} \cdot \mathbf{2})\beta_2 + u_i$
- A3a holds if the instrument is exogenous $E(\mathbf{z}_{i2}\epsilon_i) = 0$

$$\begin{aligned} E(\tilde{y}'_{i2}u_i) &= E((\pi_0 + \mathbf{z}_{i1} \cdot \mathbf{1} + \mathbf{z}_{i2})(v_i\beta_2 + \epsilon_i)) \\ &= E((\pi_0 + \mathbf{z}_{i1} \cdot \mathbf{1} + \mathbf{z}_{i2})(\epsilon_i)) \\ &= E(\epsilon_i)\pi_0 + E(\epsilon_i z_{i1})\pi_1 + E(\epsilon_i z_{i2}) \\ &= 0 \end{aligned}$$

Hence, (28.1) is consistent

The 2SLS Estimator

1. Estimate the first stage using OLS

$$y_{i2} = \pi_0 + \mathbf{z}_{i2} \cdot \mathbf{2} + \mathbf{v}_i$$

and obtained estimated value \hat{y}_{i2}

2. Estimate the altered equation using OLS

$$y_{i1} = \beta_0 + \mathbf{z}_{i1}' \mathbf{1} + \hat{y}_{i2}\beta_2 + \epsilon_i$$

Properties of the 2SLS Estimator

- Under A1, A2, A3a (for z_{i1}), A5 and if the instrument satisfies the following two conditions, + **Instrument Relevance**: $\pi_2 \neq 0$ + **Instrument Exogeneity**: $E(\mathbf{z}'_{i2}\epsilon_i) = 0$ then the 2SLS estimator is consistent
- Can handle more than one endogenous variable and more than one instrumental variable

$$y_{i1} = \beta_0 + z_{i1}\beta_1 + y_{i2}\beta_2 + y_{i3}\beta_3 + \epsilon_i \\ y_{i2} = \pi_0 + z_{i1}\pi_1 + z_{i2}\pi_2 + z_{i3}\pi_3 + z_{i4}\pi_4 + v_{i2} \\ y_{i3} = \gamma_0 + z_{i1}\gamma_1 + z_{i2}\gamma_2 + z_{i3}\gamma_3 + z_{i4}\gamma_4 + v_{i3}$$

+ **IV estimator**: one endogenous variable with a single instrument
+ **2SLS estimator**: one endogenous variable with multiple instruments
+ **GMM estimator**: multiple endogenous variables with multiple instruments

- Standard errors produced in the second step are not correct
 - Because we do not know \tilde{y} perfectly and need to estimate it in the first step, we are introducing additional variation
 - We did not have this problem with FGLS because “the first stage was orthogonal to the second stage.” This is generally not true for most multi-step procedure.
 - If A4 does not hold, need to report robust standard errors.
- 2SLS is less efficient than OLS and will always have larger standard errors.
 - First, $Var(u_i) = Var(v_i\beta_2 + \epsilon_i) > Var(\epsilon_i)$
 - Second, \hat{y}_{i2} is generally highly collinear with \mathbf{z}_{i1}
- The number of instruments need to be at least as many or more the number of endogenous variables.

Note

- 2SLS can be combined with FGLS to make the estimator more efficient: You have the same first-stage, and in the second-stage, instead of using OLS, you can use FLGS with the weight matrix \hat{w}
- Generalized Method of Moments can be more efficient than 2SLS.
- In the second-stage of 2SLS, you can also use MLE, but then you are making assumption on the distribution of the outcome variable, the endogenous variable, and their relationship (joint distribution).

28.3.1.1 Testing Assumption

1. Test of Endogeneity: Is y_{i2} truly endogenous (i.e., can we just use OLS instead of 2SLS)?
2. Testing Instrument's assumptions
 - Exogeneity (Cannot always test “and when you can it might not be informative”)
 - Relevancy (need to avoid “weak instruments”)

28.3.1.1.1 Test of Endogeneity

- 2SLS is generally so inefficient that we may prefer OLS if there is not much endogeneity
- Biased but inefficient vs efficient but biased
- Want a sense of “how endogenous” y_{i2} is
 - if “very” endogeneous - should use 2SLS
 - if not “very” endogenous - perhaps prefer OLS

Invalid Test of Endogeneity: y_{i2} is endogenous if it is correlated with ϵ_i ,

$$\epsilon_i = \gamma_0 + y_{i2}\gamma_1 + error_i$$

where $\gamma_1 \neq 0$ implies that there is endogeneity

- ϵ_i is not observed, but using the residuals

$$e_i = \gamma_0 + y_{i2}\gamma_1 + error_i$$

is **NOT** a valid test of endogeneity + The OLS residual, e is mechanically uncorrelated with y_{i2} (by FOC for OLS) + In every situation, γ_1 will be essentially 0 and you will never be able to reject the null of no endogeneity

Valid test of endogeneity

- If y_{i2} is not endogenous then ϵ_i and v are uncorrelated

$$y_{i1} = \beta_0 + \mathbf{z}_{i1}\beta_1 + y_{i2}\beta_2 + \epsilon_i y_{i2} = \pi_0 + \mathbf{z}_{i1}\pi_1 + z_{i2}\pi_2 + v_i$$

variable Addition test: include the first stage residuals as an additional variable,

$$y_{i1} = \beta_0 + \mathbf{z}_{i1}\beta_1 + y_{i2}\beta_2 + \hat{v}_i\theta + error_i$$

Then the usual t-test of significance is a valid test to evaluate the following hypothesis. **note** this test requires your instrument to be valid instrument.

$$\begin{aligned} H_0 : \theta &= 0 && (\text{not endogenous}) \\ H_1 : \theta &\neq 0 && (\text{endogenous}) \end{aligned}$$

28.3.1.1.2 Testing Instrument's assumptions The instrumental variable must satisfy

1. Exogeneity (Cannot always test “and when you can it might not be informative”)
2. Relevancy (need to avoid “weak instruments”)

28.3.1.1.2.1 Exogeneity Why exogeneity matter?

$$E(\mathbf{z}'_{i2}\epsilon_i) = 0$$

- If A3a fails - 2SLS is also inconsistent
- If instrument is not exogenous, then we need to find a new one.
- Similar to Test of Endogeneity, when there is a single instrument

$$e_i = \gamma_0 + \mathbf{z}_{i2}\gamma_1 + error_i H_0 : \gamma_1 = 0$$

is **NOT** a valid test of endogeneity

- the OLS residual, e is mechanically uncorrelated with z_{i2} : $\hat{\gamma}_1$ will be essentially 0 and you will never be able to determine if the instrument is endogenous.

Solution

Testing Instrumental Exogeneity in an Over-identified Model

- When there is more than one exogenous instrument (per endogenous variable), we can test for instrument exogeneity.
 - When we have multiple instruments, the model is said to be over-identified.
 - Could estimate the same model several ways (i.e., can identify/ estimate β_1 more than one way)
- Idea behind the test: if the controls and instruments are truly exogenous then OLS estimation of the following regression,

$$\epsilon_i = \gamma_0 + z_{i1}\gamma_1 + z_{i2}\gamma_2 + error_i$$

should have a very low R^2

- if the model is **just identified** (one instrument per endogenous variable) then the $R^2 = 0$

Steps:

- (1) Estimate the structural equation by 2SLS (using all available instruments) and obtain the residuals e
- (2) Regress e on all controls and instruments and obtain the R^2
- (3) Under the null hypothesis (all IV's are uncorrelated), $nR^2 \sim \chi^2(q)$, where q is the number of instrumental variables minus the number of endogenous variables
 - if the model is just identified (one instrument per endogenous variable) then $q = 0$, and the distribution under the null collapses.

low p-value means you reject the null of exogenous instruments. Hence you would like to have high p-value in this test.

Pitfalls for the Overid test

- the overid test is essentially compiling the following information.
 - Conditional on first instrument being exogenous is the other instrument exogenous?
 - Conditional on the other instrument being exogenous, is the first instrument exogenous?
- If all instruments are endogenous than neither test will be valid
- really only useful if one instrument is thought to be truly exogenous (randomly assigned). even f you do reject the null, the test does not tell you which instrument is exogenous and which is endogenous.

Result	Implication
reject the null	you can be pretty sure there is an endogenous instrument, but don't know which one.
fail to reject	could be either (1) they are both exogenous, (2) they are both endogenous.

28.3.1.1.2.2 Relevancy Why Relevance matter?

$$\pi_2 \neq 0$$

- * used to show A2 holds + If $\pi_2 = 0$ (instrument is not relevant) then A2 fails
- perfect multicollinearity
- + If π_2 is close to 0 (**weak instrument**) then there is near perfect multicollinearity - 2SLS is highly inefficient (Large standard errors).
- * A weak instrument will exacerbate any inconsistency due to an instrument being (even slightly) endogenous.
- + In the simple case with no controls and a single endogenous variable and single instrumental variable,

$$plim(\hat{\beta}_{2SLS}) = \beta_2 + \frac{E(z_{i2}\epsilon_i)}{E(z_{i2}y_{i2})}$$

Testing Weak Instruments

- can use t-test (or F-test for over-identified models) in the first stage to determine if there is a weak instrument problem.

- (Stock and Yogo, 2005): a statistical rejection of the null hypothesis in the first stage at the 5% (or even 1%) level is not enough to insure the instrument is not weak
 - Rule of Thumb: need a F-stat of at least 10 (or a t-stat of at least 3.2) to reject the null hypothesis that the instrument is weak.

Summary of the 2SLS Estimator

$$y_{i1} = \beta_0 + \mathbf{z}_{i1}\beta_1 + y_{i2}\beta_2 + \epsilon_i y_{i2} = \pi_0 + \mathbf{z}_{i1_1} + \mathbf{z}_{i2_2} + v_i$$

- when A3a does not hold

$$E(y'_{i2}\epsilon_i) \neq 0$$

- Then the OLS estimator is no longer unbiased or consistent.
- If we have valid instruments \mathbf{z}_{i2}
- Relevancy (need to avoid “weak instruments”): $\pi_2 \neq 0$ Then the 2SLS estimator is consistent under A1, A2, A5a, and the above two conditions.
+ If A4 also holds, then the usual standard errors are valid. + If A4 does not hold then use the robust standard errors.

$$y_{i1} = \beta_0 + \mathbf{z}_{i1}\beta_1 + y_{i2}\beta_2 + \epsilon_i y_{i2} = \pi_0 + \mathbf{z}_{i1_1} + \mathbf{z}_{i2_2} + v_i$$

- When A3a does hold

$$E(y'_{i2}\epsilon_i) = 0$$

and we have valid instruments, then both the OLS and 2SLS estimators are consistent.

- The OLS estimator is always more efficient
- can use the variable addition test to determine if 2SLS is needed (A3a does hold) or if OLS is valid (A3a does not hold)

Sometimes we can test the assumption for instrument to be valid:

- Exogeneity : Only table when there are more instruments than endogenous variables.
- Relevancy (need to avoid “weak instruments”): Always testable, need the F-stat to be greater than 10 to rule out a weak instrument

Application

Expenditure as observed instrument

```
m2.2sls <-
  ivreg(
    read ~ stratio + english + lunch + grades + income + calworks +
        county | expenditure + english + lunch + grades + income + calworks +
        county ,
    data = school
  )
summary(m2.2sls)$coefficients[1:7, ]
#>           Estimate Std. Error     t value   Pr(>|t|)
#> (Intercept) 700.47891593 13.58064436 51.5792106 8.950497e-171
#> stratio      -1.13674002  0.53533638 -2.1234126 3.438427e-02
#> english       -0.21396934  0.03847833 -5.5607753 5.162571e-08
#> lunch         -0.39384225  0.03773637 -10.4366757 1.621794e-22
#> gradesKK-08  -1.89227865  1.37791820 -1.3732881 1.704966e-01
#> income         0.62487986  0.11199008  5.5797785 4.668490e-08
#> calworks      -0.04950501  0.06244410 -0.7927892 4.284101e-01
```

28.3.1.2 Checklist

1. Regress the dependent variable on the instrument (reduced form). Since under OLS, we have unbiased estimate, the coefficient estimate should be significant (make sure the sign makes sense)
2. Report F-stat on the excluded instruments. F-stat < 10 means you have a weak instrument (Stock et al., 2002).
3. Present R^2 before and after including the instrument (Rossi, 2014)
4. For models with multiple instrument, present first and second-stage result for each instrument separately. Overid test should be conducted (e.g., Sargan-Hansen J)
5. Hausman test between OLS and 2SLS (don't confuse this test for evidence that endogeneity is irrelevant - under invalid IV, the test is useless)
6. Compare the 2SLS with the limited information ML. If they are different, you have evidence for weak instruments.

28.3.1.3 Good Instruments

Exogeneity and Relevancy are necessary but not sufficient for IV to produce consistent estimates.

Without theory or possible explanation, you can always create a new variable that is correlated with X and uncorrelated with ϵ

For example, we want to estimate the effect of price on quantity (Reiss, 2011, p. 960)

$$Q = \beta_1 P + \beta_2 X + \epsilon P = \pi_1 X + \eta$$

where ϵ and η are jointly determined, $X \perp \epsilon, \eta$

Without theory, we can just create a new variable $Z = X + u$ where $E(u) = 0; u \perp X, \epsilon, \eta$

Then, Z satisfied both conditions:

- Relevancy: X correlates $P \rightarrow Z$ correlates P
- Exogeneity: $u \perp \epsilon$ (random noise)

But obviously, it's not a valid instrument (intuitively). But theoretically, relevance and exogeneity are not sufficient to identify β because of unsatisfied rank condition for identification.

Moreover, the functional form of the instrument also plays a role when choosing a good instrument. Hence, we always need to check for the robustness of our instrument.

IV methods even with valid instruments can still have poor sampling properties (finite sample bias, large sampling errors) (Rossi, 2014)

When you have a weak instrument, it's important to report it appropriately (Lee et al., 2021). This problem will be exacerbated if you have multiple instruments.

28.3.1.3.1 Lagged dependent variable In time series data sets, we can use lagged dependent variable as an instrument because it is not influenced by current shocks.

Citations for lagged dependent variable in econ (Chetty et al., 2013),

28.3.2 Internal instrumental variable

- (also known as **instrument free methods**). This section is based on Raluca Gui's guide
- alternative to external instrumental variable approaches
- All approaches here assume a **continuous dependent variable**

28.3.2.1 Non-hierarchical Data (Cross-classified)

$$Y_t = \beta_0 + \beta_1 P_t + \beta_2 X_t + \epsilon_t$$

where

- $t = 1, \dots, T$ (indexes either time or cross-sectional units)
- Y_t is a $k \times 1$ response variable
- X_t is a $k \times n$ exogenous regressor
- P_t is a $k \times 1$ continuous endogenous regressor
- ϵ_t is a structural error term with $\mu_\epsilon = 0$ and $E(\epsilon^2) = \sigma^2$
- β are model parameters

The endogeneity problem arises from the correlation of P_t and ϵ_t :

$$P_t = \gamma Z_t + v_t$$

where

- Z_t is a $l \times 1$ vector of internal instrumental variables
- v_t is a random error with $\mu_{v_t} = 0, E(v^2) = \sigma_v^2, E(\epsilon v) = \sigma_{\epsilon v}$
- Z_t is assumed to be stochastic with distribution G
- v_t is assumed to have density $h(\cdot)$

28.3.2.1.1 Latent Instrumental Variable (Ebbes et al., 2005)

assume Z_t (unobserved) to be uncorrelated with ϵ_t , which is similar to Instrumental Variable. Hence, Z_t and v_t can't be identified without distributional assumptions

The distributions of Z_t and v_t need to be specified such that:

- (1) endogeneity of P_t is corrected
- (2) the distribution of P_t is empirically close to the integral that expresses the amount of overlap of Z as it is shifted over (= the convolution between Z_t and ν_t).

When the density $h(\cdot)$ = Normal, then G cannot be normal because the parameters would not be identified (Ebbes et al., 2005) .

Hence,

- in the LIV model the distribution of Z_t is discrete
- in the Higher Moments Method and Joint Estimation Using Copula methods, the distribution of Z_t is taken to be skewed.

Z_t are assumed **unobserved, discrete and exogenous**, with

- an unknown number of groups m
- γ is a vector of group means.

Identification of the parameters relies on the distributional assumptions of

- P_t : a non-Gaussian distribution
- Z_t discrete with $m \geq 2$

Note:

- If Z_t is continuous, the model is unidentified
- If $P_t \sim N$, you have inefficient estimates.

```
m3.liv <- latentIV(read ~ stratio, data=school)
summary(m3.liv)$coefficients[1:7,]
#>              Estimate     Std. Error      z-score    Pr(>/z/)
#> (Intercept) 6.996014e+02 2.686186e+02 2.604441e+00 9.529597e-03
#> stratio     -2.272673e+00 1.367757e+01 -1.661605e-01 8.681108e-01
#> pi1         -4.896363e+01 5.526907e-08 -8.859139e+08 0.0000000e+00
#> pi2          1.963920e+01 9.225351e-02 2.128830e+02 0.0000000e+00
#> theta5       6.939432e-152 3.354672e-160 2.068587e+08 0.0000000e+00
#> theta6       3.787512e+02 4.249457e+01 8.912932e+00 1.541524e-17
#> theta7       -1.227543e+00 4.885276e+01 -2.512741e-02 9.799653e-01
```

it will return a coefficient very different from the other methods since there is only one endogenous variable.

28.3.2.1.2 Joint Estimation Using Copula assume Z_t (unobserved) to be uncorrelated with ϵ_t , which is similar to Instrumental Variable. Hence, Z_t and ν_t can't be identified without distributional assumptions

(Park and Gupta, 2012) allows joint estimation of the continuous P_t and ϵ_t using Gaussian copulas, where a copula is a function that maps several conditional distribution functions (CDF) into their joint CDF).

The underlying idea is that using information contained in the observed data, one selects marginal distributions for P_t and ϵ_t . Then, the copula model constructs a flexible multivariate joint distribution that allows a wide range of correlations between the two marginals.

The method allows both continuous and discrete P_t .

In the special case of **one continuous** P_t , estimation is based on MLE
Otherwise, based on Gaussian copulas, augmented OLS estimation is used.

Assumptions:

- skewed P_t
- the recovery of the correct parameter estimates
- $\epsilon_t \sim$ normal marginal distribution. The marginal distribution of P_t is obtained using the **Epanechnikov kernel density estimator**

$$\hat{h}_p = \frac{1}{T.b} \sum_{t=1}^T K\left(\frac{p - P_t}{b}\right)$$

where

- P_t = endogenous variables
- $K(x) = 0.75(1 - x^2)I(|x| \leq 1)$
- $b = 0.9T^{-1/5} \times \min(s, IQR/1.34)$ suggested by (Silverman, 1969)
 - IQR = interquartile range
 - s = sample standard deviation
 - T = n of time periods observed in the data

In augmented OLS and MLE, the inference procedure occurs in two stages:

- (1): the empirical distribution of P_t is computed
 - (2) used in it constructing the likelihood function)
- Hence, the standard errors would not be correct.

So we use the sampling distributions (from bootstrapping) to get standard errors and the variance-covariance matrix. Since the distribution of the bootstrapped parameters is highly skewed, we report the percentile confidence intervals is preferable.

```

set.seed(110)
m4.cc <-
  copulaCorrection(
    read ~ stratio + english + lunch + calworks +
      grades + income + county | continuous(stratio),
    data = school,
    optimx.args = list(method = c("Nelder-Mead"), itnmax = 60000),
    num.boots = 2,
    verbose = FALSE
  )
summary(m4.cc)$coefficients[1:7, ]
#>           Point Estimate   Boots SE Lower Boots CI (95%) Upper Boots CI (95%)
#> (Intercept) 683.06900891 2.80554212          NA          NA
#> stratio     -0.32434608 0.02075999          NA          NA
#> english      -0.21576110 0.01450666          NA          NA
#> lunch        -0.37087664 0.01902052          NA          NA
#> calworks     -0.05569058 0.02076781          NA          NA
#> gradesKK-08 -1.92286128 0.25684614          NA          NA
#> income        0.73595353 0.04725700          NA          NA

```

we run this model with only one endogenous continuous regressor (`stratio`). Sometimes, the code will not converge, in which case you can use different

- optimization algorithm
- starting values
- maximum number of iterations

28.3.2.1.3 Higher Moments Method suggested by (Lewbel, 1997) to identify ϵ_t caused by **measurement error**.

Identification is achieved by using third moments of the data, with no restrictions on the distribution of ϵ_t

The following instruments can be used with 2SLS estimation to obtain consistent estimates:

$$\begin{aligned}
 q_{1t} &= (G_t - \bar{G}) \\
 q_{2t} &= (G_t - \bar{G})(P_t - \bar{P}) \\
 q_{3t} &= (G_t - \bar{G})(Y_t - \bar{Y}) \\
 q_{4t} &= (Y_t - \bar{Y})(P_t - \bar{P}) \\
 q_{5t} &= (P_t - \bar{P})^2 \\
 q_{6t} &= (Y_t - \bar{Y})^2
 \end{aligned}$$

where

- $G_t = G(X_t)$ for any given function G that has finite third own and cross moments
- X = exogenous variable

q_{5t}, q_{6t} can be used only when the measurement and ϵ_t are symmetrically distributed. The rest of the instruments does not require any distributional assumptions for ϵ_t .

Since the regressors $G(X) = X$ are included as instruments, $G(X)$ can't be a linear function of X in q_{1t}

Since this method has very strong assumptions, Higher Moments Method should only be used in case of overidentification

```
set.seed(111)
m5.hetEr <-
  hetErrorsIV(
    read ~ stratio + english + lunch + calworks + income +
      grades + county | stratio | IIV(income, english),
    data = school
  )
summary(m5.hetEr)$coefficients[1:7, ]
#>           Estimate Std. Error   t value Pr(>|t|)    
#> (Intercept) 662.78791557 27.90173069 23.7543657 2.380436e-76
#> stratio      0.71480686  1.31077325  0.5453322 5.858545e-01
#> english     -0.19522271  0.04057527 -4.8113717 2.188618e-06
#> lunch        -0.37834232  0.03927793 -9.6324402 9.760809e-20
#> calworks    -0.05665126  0.06302095 -0.8989273 3.692776e-01
#> income       0.82693755  0.17236557  4.7975797 2.335271e-06
#> gradesKK-08 -1.93795843  1.38723186 -1.3969968 1.632541e-01
```

recommend using this approach to create additional instruments to use with external ones for better efficiency.

28.3.2.1.4 Heteroskedastic Error Approach

- using means of variables that are uncorrelated with the product of heteroskedastic errors to identify structural parameters.
- This method can be used either when you don't have external instruments or you want to use additional instruments to improve the efficiency of the IV estimator (Lewbel, 2012)
- The instruments are constructed as simple functions of data
- Model's assumptions:

$$E(X\epsilon) = 0E(Xv) = 0\text{cov}(Z, \epsilon v) = 0\text{cov}(Z, v^2) \neq 0 \text{ (for identification)}$$

Structural parameters are identified by 2SLS regression of Y on X and P, using X and $[Z - E(Z)]$ as instruments.

$$\text{instrument's strength} \propto \text{cov}((Z - \bar{Z})v, v)$$

where $\text{cov}((Z - \bar{Z})v, v)$ is the degree of heteroskedasticity of v with respect to Z (Lewbel, 2012), which can be empirically tested.

If it is zero or close to zero (i.e., the instrument is weak), you might have imprecise estimates, with large standard errors.

- Under homoskedasticity, the parameters of the model are unidentified.
- Under heteroskedasticity related to at least some elements of X, the parameters of the model are identified.

28.3.2.2 Hierarchical Data

Multiple independent assumptions involving various random components at different levels mean that any moderate correlation between some predictors and a random component or error term can result in a significant bias of the coefficients and of the variance components. (Kim and Frees, 2007) proposed a generalized method of moments which uses both, the between and within variations of the exogenous variables, but only assumes the within variation of the variables to be endogenous.

Assumptions

- the errors at each level $\sim iidN$
- the slope variables are exogenous
- the level-1 $\epsilon \perp X, P$. If this is not the case, additional, external instruments are necessary

Hierarchical Model

$$\begin{aligned} Y_{cst} &= Z_{cst}^1 \beta_{cs}^1 + X_{cst}^1 \beta_1 + \epsilon_{cst}^1 \\ \beta_{cs}^1 &= Z_{cs}^2 \beta_c^2 + X_{cst}^2 \beta_2 + \epsilon_{cst}^2 \\ \beta_c^2 &= X_c^3 \beta_3 + \epsilon_c^3 \end{aligned}$$

Bias could stem from:

- errors at the higher two levels ($\epsilon_c^3, \epsilon_{cst}^2$) are correlated with some of the regressors
- only third level errors (ϵ_c^3) are correlated with some of the regressors

(Kim and Frees, 2007) proposed

- When all variables are assumed exogenous, the proposed estimator equals the random effects estimator
- When all variables are assumed endogenous, it equals the fixed effects estimator
- also use omitted variable test (based on the Hausman-test (Hausman, 1978) for panel data), which allows the comparison of a robust estimator and an estimator that is efficient under the null hypothesis of no omitted variables or the comparison of two robust estimators at different levels.

```
set.seed(113)
school$gr08 <- school$grades == "KK-06"
m7.multilevel <-
  multilevelIV(read ~ stratio + english + lunch + income + gr08 +
    calworks + (1 | county) | endo(stratio),
    data = school)
summary(m7.multilevel)$coefficients[1:7, ]
#>           Estimate Std. Error      z-score     Pr(>|z|)
#> (Intercept) 675.8228656 5.58008680 121.1133248 0.000000e+00
#> stratio      -0.4956054 0.23922638  -2.0717005 3.829339e-02
#> english       -0.2599777 0.03413530   -7.6160948 2.614656e-14
#> lunch         -0.3692954 0.03560210  -10.3728537 3.295342e-25
#> income        0.6723141 0.08862012    7.5864728 3.287314e-14
#> gr08TRUE      2.1590333 1.28167222    1.6845440 9.207658e-02
#> calworks     -0.0570633 0.05711701   -0.9990596 3.177658e-01
```

Another example using simulated data

- level-1 regressors: $X_{11}, X_{12}, X_{13}, X_{14}, X_{15}$, where X_{15} is correlated with the level-2 error (i.e., endogenous).
- level-2 regressors: $X_{21}, X_{22}, X_{23}, X_{24}$
- level-3 regressors: X_{31}, X_{32}, X_{33}

We estimate a three-level model with X_{15} assumed endogenous. Having a three-level hierarchy, `multilevelIV()` returns five estimators, from the most robust to omitted variables (FE_L2), to the most efficient (REF) (i.e. lowest mean squared error).

- The random effects estimator (REF) is efficient assuming no omitted variables
- The fixed effects estimator (FE) is unbiased and asymptotically normal even in the presence of omitted variables.

- Because of the efficiency, the random effects estimator is preferable if you think there is no omitted variables
- The robust estimator would be preferable if you think there is omitted variables.

```

data(dataMultilevelIV)
set.seed(114)
formula1 <-
  y ~ X11 + X12 + X13 + X14 + X15 + X21 + X22 + X23 + X24 +
  X31 + X32 + X33 + (1 | CID) + (1 | SID) | endo(X15)
m8.multilevel <-
  multilevelIV(formula = formula1, data = dataMultilevelIV)
coef(m8.multilevel)
#>           REF      FE_L2      FE_L3      GMM_L2      GMM_L3
#> (Intercept) 64.3168856  0.0000000  0.0000000 64.3485944 64.3168868
#> X11          3.0213405  3.0459605  3.0214255  3.0146686 3.0213403
#> X12          8.9522160  8.9839088  8.9524723  8.9747533 8.9522169
#> X13         -2.0194178 -2.0145054 -2.0193321 -2.0021426 -2.0194171
#> X14          1.9651420  1.9791437  1.9648317  1.9658681 1.9651421
#> X15         -0.5647915 -0.9777361 -0.5647621 -0.9750309 -0.5648070
#> X21         -2.3316225  0.0000000 -2.2845297 -2.3052516 -2.3316215
#> X22         -3.9564944  0.0000000 -3.9553644 -4.0130975 -3.9564966
#> X23         -2.9779887  0.0000000 -2.9756848 -2.9488487 -2.9779876
#> X24          4.9078293  0.0000000  4.9084694  4.7933756 4.9078250
#> X31          2.1142348  0.0000000  0.0000000  2.1164477 2.1142349
#> X32          0.3934770  0.0000000  0.0000000  0.3799626 0.3934764
#> X33          0.1082086  0.0000000  0.0000000  0.1108386 0.1082087

summary(m8.multilevel, "REF")
#>
#> Call:
#> multilevelIV(formula = formula1, data = dataMultilevelIV)
#>
#> Number of levels: 3
#> Number of observations: 2824
#> Number of groups: L2(CID): 1368 L3(SID): 40
#>
#> Coefficients for model REF:
#>           Estimate Std. Error z-score Pr(>/z|)
#> (Intercept) 64.31689    7.87332   8.169 3.11e-16 ***
#> X11          3.02134    0.02576  117.306 < 2e-16 ***
#> X12          8.95222    0.02572 348.131 < 2e-16 ***
#> X13         -2.01942    0.02409 -83.835 < 2e-16 ***
#> X14          1.96514    0.02521  77.937 < 2e-16 ***
#> X15         -0.56479    0.01950 -28.962 < 2e-16 ***

```

```

#> X21      -2.33162   0.16228 -14.368 < 2e-16 ***
#> X22      -3.95649   0.13119 -30.160 < 2e-16 ***
#> X23      -2.97799   0.06611 -45.044 < 2e-16 ***
#> X24      4.90783    0.19796  24.792 < 2e-16 ***
#> X31      2.11423    0.10433  20.264 < 2e-16 ***
#> X32      0.39348    0.30426   1.293  0.1959
#> X33      0.10821    0.05236   2.067  0.0388 *
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Omitted variable tests for model REF:
#>           df     Chisq p-value
#> GMM_L2_vs_REF 7     18.74 0.009040 **
#> GMM_L3_vs_REF 13   -12872.98 1.000000
#> FE_L2_vs_REF  13    39.99 0.000139 ***
#> FE_L3_vs_REF  13    39.99 0.000138 ***
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

True $\beta_{X_{15}} = -1$. We can see that some estimators are bias because X_{15} is correlated with the level-two error, to which only FE_L2 and GMM_L2 are robust

To select the appropriate estimator, we use the omitted variable test.

In a three-level setting, we can have different estimator comparisons:

- Fixed effects vs. random effects estimators: Test for omitted level-two and level-three omitted effects, simultaneously, one compares FE_L2 to REF. But we will not know at which omitted variables exist.
- Fixed effects vs. GMM estimators: Once the existence of omitted effects is established but not sure at which level, we test for level-2 omitted effects by comparing FE_L2 vs GMM_L3. If you reject the null, the omitted variables are at level-2. The same is accomplished by testing FE_L2 vs. GMM_L2, since the latter is consistent only if there are no omitted effects at level-2.
- Fixed effects vs. fixed effects estimators: We can test for omitted level-2 effects, while allowing for omitted level-3 effects by comparing FE_L2 vs. FE_L3 since FE_L2 is robust against both level-2 and level-3 omitted effects while FE_L3 is only robust to level-3 omitted variables.

Summary, use the omitted variable test comparing REF vs. FE_L2 first.

- If the null hypothesis is rejected, then there are omitted variables either at level-2 or level-3

- Next, test whether there are level-2 omitted effects, since testing for omitted level three effects relies on the assumption there are no level-two omitted effects. You can use any of these pair of comparisons:
 - FE_L2 vs. FE_L3
 - FE_L2 vs. GMM_L2
- If no omitted variables at level-2 are found, test for omitted level-3 effects by comparing either
 - FE_L3 vs. GMM_L3
 - GMM_L2 vs. GMM_L3

```
summary(m8.multilevel, "REF")
#>
#> Call:
#> multilevelIV(formula = formula1, data = dataMultilevelIV)
#>
#> Number of levels: 3
#> Number of observations: 2824
#> Number of groups: L2(CID): 1368 L3(SID): 40
#>
#> Coefficients for model REF:
#>             Estimate Std. Error z-score Pr(>/z|)
#> (Intercept) 64.31689   7.87332   8.169 3.11e-16 ***
#> X11         3.02134   0.02576 117.306 < 2e-16 ***
#> X12         8.95222   0.02572 348.131 < 2e-16 ***
#> X13        -2.01942   0.02409 -83.835 < 2e-16 ***
#> X14         1.96514   0.02521  77.937 < 2e-16 ***
#> X15        -0.56479   0.01950 -28.962 < 2e-16 ***
#> X21        -2.33162   0.16228 -14.368 < 2e-16 ***
#> X22        -3.95649   0.13119 -30.160 < 2e-16 ***
#> X23        -2.97799   0.06611 -45.044 < 2e-16 ***
#> X24         4.90783   0.19796  24.792 < 2e-16 ***
#> X31         2.11423   0.10433  20.264 < 2e-16 ***
#> X32         0.39348   0.30426   1.293   0.1959
#> X33         0.10821   0.05236   2.067   0.0388 *
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Omitted variable tests for model REF:
#>             df    Chisq p-value
#> GMM_L2_vs_REF  7    18.74 0.009040 **
#> GMM_L3_vs_REF 13 -12872.98 1.000000
#> FE_L2_vs_REF  13    39.99 0.000139 ***
#> FE_L3_vs_REF  13    39.99 0.000138 ***
```

```
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# compare REF with all the other estimators. Testing REF (the most efficient estimator) against F
```

Since the null hypothesis is rejected ($p = 0.000139$), there is bias in the random effects estimator.

To test for level-2 omitted effects (regardless of level-3 omitted effects), we compare FE_L2 versus FE_L3

```
summary(m8.multilevel, "FE_L2")
#>
#> Call:
#> multilevelIV(formula = formula1, data = dataMultilevelIV)
#>
#> Number of levels: 3
#> Number of observations: 2824
#> Number of groups: L2(CID): 1368  L3(SID): 40
#>
#> Coefficients for model FE_L2:
#>             Estimate Std. Error z-score Pr(>/z|)
#> (Intercept) 0.000e+00 4.275e-19   0.00      1
#> X11         3.046e+00 2.978e-02 102.30 <2e-16 ***
#> X12         8.984e+00 3.360e-02 267.41 <2e-16 ***
#> X13        -2.015e+00 3.107e-02 -64.83 <2e-16 ***
#> X14         1.979e+00 3.203e-02  61.80 <2e-16 ***
#> X15        -9.777e-01 3.364e-02 -29.06 <2e-16 ***
#> X21         0.000e+00 1.824e-18   0.00      1
#> X22         0.000e+00 1.303e-18   0.00      1
#> X23         0.000e+00 4.389e-18   0.00      1
#> X24         0.000e+00 1.724e-18   0.00      1
#> X31         0.000e+00 1.468e-17   0.00      1
#> X32         0.000e+00 8.265e-18   0.00      1
#> X33         0.000e+00 2.793e-17   0.00      1
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Omitted variable tests for model FE_L2:
#>             df Chisq p-value
#> FE_L2_vs_REF    13 39.99 0.000139 ***
#> FE_L2_vs_FE_L3   9 36.02 3.92e-05 ***
#> FE_L2_vs_GMM_L2 12 39.99 7.21e-05 ***
#> FE_L2_vs_GMM_L3 13 39.99 0.000139 ***
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis of no omitted level-2 effects is rejected ($p = 3.92e - 05$). Hence, there are omitted effects at level-two. We should use FE_L2 which is consistent with the underlying data that we generated (level-2 error correlated with X_{15} , which leads to biased FE_L3 coefficients).

The omitted variable test between FE_L2 and GMM_L2 should reject the null hypothesis of no omitted level-2 effects (p-value is 0).

If we assume an endogenous variable as exogenous, the RE and GMM estimators will be biased because of the wrong set of internal instrumental variables. To increase our confidence, we should compare the omitted variable tests when the variable is considered endogenous vs. exogenous to get a sense whether the variable is truly endogenous.

28.3.3 Proxy Variables

- Can be in place of the omitted variable
- will not be able to estimate the effect of the omitted variable
- will be able to reduce some endogeneity caused by the omitted variable
- but it can have Measurement Error. Hence, you have to be extremely careful when using proxies.

Criteria for a proxy variable:

1. The proxy is correlated with the omitted variable.
2. Having the omitted variable in the regression will solve the problem of endogeneity
3. The variation of the omitted variable unexplained by the proxy is uncorrelated with all independent variables, including the proxy.

IQ test can be a proxy for ability in the regression between wage explained education.

For the third requirement

$$\text{ability} = \gamma_0 + \gamma_1 \text{IQ} + \epsilon$$

where ϵ is uncorrelated with education and IQ test.

28.4 Endogenous Sample Selection

sample selection or self-selection problem

the omitted variable is how people were selected into the sample

Some disciplines consider nonresponse bias and selection bias as sample selection.

- When unobservable factors that affect who is in the sample are independent of unobservable factors that affect the outcome, the sample selection is not endogenous. Hence, the sample selection is ignorable and estimator that ignores sample selection is still consistent.
- when the unobservable factors that affect who is included in the sample are correlated with the unobservable factors that affect the outcome, the sample selection is endogenous and not ignorable, because estimators that ignore endogenous sample selection are not consistent (we don't know which part of the observable outcome is related to the causal relationship and which part is due to different people were selected for the treatment and control groups).

Assumptions: - The unobservables that affect the treatment selection and the outcome are jointly distributed as bivariate normal.

Notes: - If you don't have strong exclusion restriction, identification is driven by the assumed non linearity in the functional form (through inverse Mills ratio). E.g., the estimate depend on the bivariate normal distribution of the error structure:

- + With strong exclusion restriction for the covariate in the correction equation, the variation is
- + With weak exclusion restriction, and the covariate exists in both steps, it's the assumed error

To combat Sample selection, we can

- Randomization: participants are randomly selected into treatment and control.
- Instruments that determine the treatment status (i.e., treatment vs. control) but not the outcome (Y)
- Functional form of the selection and outcome processes: originated from (Heckman, 1976), later on generalize by (Amemiya, 1984)

We have our main model

$$\mathbf{y}^* = \mathbf{x}\mathbf{b} +$$

However, the pattern of missingness (i.e., censored) is related to the unobserved (latent) process:

$$\mathbf{z}^* = \mathbf{w} + \mathbf{u}$$

and

$$z_i = \begin{cases} 1 & \text{if } z_i^* > 0 \\ 0 & \text{if } z_i^* \leq 0 \end{cases}$$

Equivalently, $z_i = 1$ (y_i is observed) when

$$u_i \geq -w_i\gamma$$

Hence, the probability of observed y_i is

$$\begin{aligned} P(u_i \geq -w_i\gamma) &= 1 - \Phi(-w_i\gamma) \\ &= \Phi(w_i\gamma) \quad \text{symmetry of the standard normal distribution} \end{aligned}$$

We will **assume**

- the error term of the selection $\mathbf{u} \sim \mathbf{N}(\mathbf{0}, \mathbf{I})$
- $Var(u_i) = 1$ for identification purposes

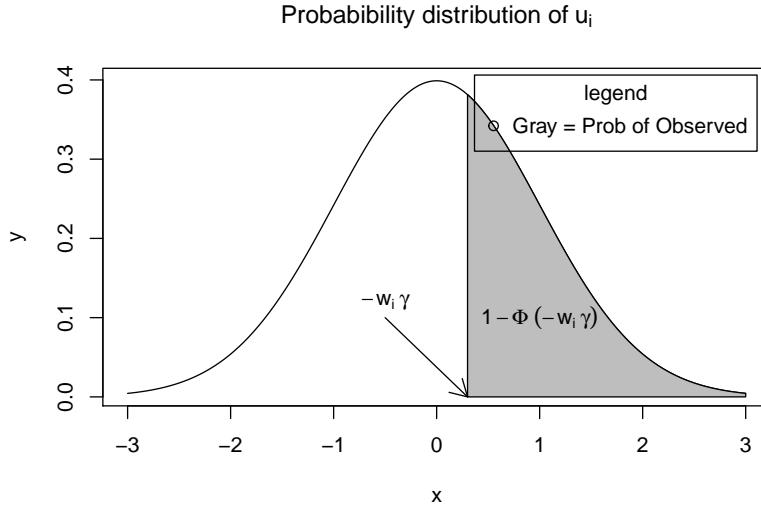
Visually, $P(u_i \geq -w_i\gamma)$ is the shaded area.

```
x = seq(-3, 3, length = 200)
y = dnorm(x, mean = 0, sd = 1)
plot(x,
      y,
      type = "l",
      main = bquote("Probability distribution of" ~ u[i]))
x = seq(0.3, 3, length = 100)
y = dnorm(x, mean = 0, sd = 1)
polygon(c(0.3, x, 3), c(0, y, 0), col = "gray")
text(1, 0.1, bquote(1 - Phi ~ (-w[i] ~ gamma)))
arrows(-0.5, 0.1, 0.3, 0, length = .15)
text(-0.5, 0.12, bquote(-w[i] ~ gamma))
legend(
  "topright",
  "Gray = Prob of Observed",
  pch = 1,
```

```

    title = "legend",
    inset = .02
)

```



Hence in our observed model, we see

$$y_i = x_i\beta + \epsilon_i \text{ when } z_i = 1 \quad (28.2)$$

and the joint distribution of the selection model (u_i), and the observed equation (ϵ_i) as

$$\begin{bmatrix} u \\ \epsilon \end{bmatrix} \sim^{iid} N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & \sigma_\epsilon^2 \end{bmatrix} \right)$$

The relation between the observed and selection models:

$$\begin{aligned}
 E(y_i | y_i \text{ observed}) &= E(y_i | z^* > 0) \\
 &= E(y_i | -w_i \gamma) \\
 &= \mathbf{x}_i \beta + E(\epsilon_i | u_i > -w_i \gamma) \\
 &= \mathbf{x}_i \beta + \rho \sigma_\epsilon \frac{\phi(w_i \gamma)}{\Phi(w_i \gamma)}
 \end{aligned}$$

where $\frac{\phi(w_i \gamma)}{\Phi(w_i \gamma)}$ is the Inverse Mills Ratio. and $\rho \sigma_\epsilon \frac{\phi(w_i \gamma)}{\Phi(w_i \gamma)} \geq 0$

Great visualization of special cases of correlation patterns amongst data and errors by professor Rob Hick

Note:

(Bareinboim et al., 2014) is an excellent summary of cases that we can still do causal inference in case of selection bias. I'll try to summarize their idea here:

Let X be an action, Y be an outcome, and S be a binary indicator of entry into the data pool where ($S = 1$ = in the sample, $S = 0$ =out of sample) and Q be the conditional distribution $Q = P(y|x)$.

Usually we want to understand , but because of S , we only have $P(y, x|S = 1)$. Hence, we'd like to recover $P(y|x)$ from $P(y, x|S = 1)$

- If both X and Y affect S , we can't unbiasedly estimate $P(y|x)$

In the case of Omitted variable bias (U) and sample selection bias (S), you have unblocked extraneous “flow” of information between X and Y , which causes spurious correlation for X and Y . Traditionally, we would recover Q by parametric assumption of

- (1) the data generating process (e.g., Heckman 2-step)
- (2) type of data-generating model (e.g., treatment-dependent or outcome-dependent)
- (3) selection's probability $P(S = 1|Pa_s)$ with non-parametrically based causal graphical models, the authors proposed more robust way to model misspecification regardless of the type of data-generating model, and do not require selection's probability. Hence, you can recover Q
 - without external data
 - with external data
 - causal effects with the Selection-backdoor criterion

28.4.1 Tobit-2

also known as Heckman's standard sample selection model
Assumption: joint normality of the errors

Data here is taken from (Mroz, 1987)'s paper.

We want to estimate the log(wage) for married women, with education, experience, experience squared, and a dummy variable for living in a big city. But we can only observe the wage for women who are working, which means a lot

of married women in 1975 who were out of the labor force are unaccounted for. Hence, an OLS estimate of the wage equation would be bias due to sample selection. Since we have data on non-participants (i.e., those who are not working for pay), we can correct for the selection process.

The Tobit-2 estimates are consistent

28.4.1.1 Example 1

```
library(sampleSelection)
library(dplyr)
data("Mroz87") #1975 data on married women's pay and labor-force participation from the Panel Study
head(Mroz87)
#>   lfp hours kids5 kids618 age educ    wage repwage hushrs husage huseduc huswage
#> 1   1 1610     1      0 32 12 3.3540    2.65  2708     34     12 4.0288
#> 2   1 1656     0      2 30 12 1.3889    2.65  2310     30      9 8.4416
#> 3   1 1980     1      3 35 12 4.5455    4.04  3072     40     12 3.5807
#> 4   1  456     0      3 34 12 1.0965    3.25  1920     53     10 3.5417
#> 5   1 1568     1      2 31 14 4.5918    3.60  2000     32     12 10.0000
#> 6   1 2032     0      0 54 12 4.7421    4.70  1040     57     11 6.7106
#>   faminc    mtr motheduc fatheduc unem city exper nwifeinc wifecoll huscoll
#> 1 16310 0.7215      12      7 5.0    0    14 10.910060 FALSE FALSE
#> 2 21800 0.6615      7      7 11.0   1    5 19.499981 FALSE FALSE
#> 3 21040 0.6915      12      7 5.0    0    15 12.039910 FALSE FALSE
#> 4  7300 0.7815      7      7 5.0    0    6 6.799996 FALSE FALSE
#> 5 27300 0.6215      12     14 9.5    1    7 20.100058 TRUE FALSE
#> 6 19495 0.6915      14      7 7.5    1    33 9.859054 FALSE FALSE
Mroz87 = Mroz87 %>%
  mutate(kids = kids5+kids618)

library(nnet)
library(ggplot2)
library(reshape2)
```

2-stage Heckman's model:

- (1) probit equation estimates the selection process (who is in the labor force?)
- (2) the results from 1st stage are used to construct a variable that captures the selection effect in the wage equation. This correction variable is called the **inverse Mills ratio**.

```
# OLS: log wage regression on LF participants only
ols1 = lm(log(wage) ~ educ + exper + I(exper^2) + city, data=subset(Mroz87, lfp==1))
```

```
# Heckman's Two-step estimation with LFP selection equation
heck1 = heckit( lfp ~ age + I( age^2 ) + kids + huswage + educ, # the selection process
                 log(wage) ~ educ + exper + I( exper^2 ) + city, data=Mroz87 )
```

Use only variables that affect the selection process in the selection equation. Technically, the selection equation and the equation of interest could have the same set of regressors. But it is not recommended because we should only use variables (or at least one) in the selection equation that affect the selection process, but not the wage process (i.e., instruments). Here, variable `kids` fulfill that role: women with kids may be more likely to stay home, but working moms with kids would not have their wages change.

Alternatively,

```
# ML estimation of selection model
ml1 = selection( lfp ~ age + I( age^2 ) + kids + huswage + educ,
                  log(wage) ~ educ + exper + I( exper^2 ) + city, data=Mroz87 )
```

```
library("stargazer")
library("Mediana")
library("plm")
# function to calculate corrected SEs for regression
cse = function(reg) {
  rob = sqrt(diag(vcovHC(reg, type = "HC1")))
  return(rob)
}

# stargazer table
stargazer(ols1, heck1, ml1,
           se=list(cse(ols1),NULL,NULL),
           title="Married women's wage regressions", type="text",
           df=FALSE, digits=4, selection.equation = T)
#>
#> Married women's wage regressions
#> =====
#>                               Dependent variable:
#> -----
#>                               log(wage)          lfp
#>                         OLS        Heckman      selection
#>                               selection
#>                         (1)          (2)          (3)
#> -----
#> age                           0.1861***    0.1842*** 
#>                               (0.0652)    (0.0658)
```

```

#> I(age2)           -0.0024***   -0.0024***  

#>                   (0.0008)    (0.0008)  

#>  

#> kids              -0.1496***   -0.1488***  

#>                   (0.0383)    (0.0385)  

#>  

#> huswage           -0.0430***   -0.0434***  

#>                   (0.0122)    (0.0123)  

#>  

#> educ              0.1057***   0.1250***   0.1256***  

#>                   (0.0130)    (0.0228)    (0.0229)  

#>  

#> exper              0.0411***  

#>                   (0.0154)  

#>  

#> I(exper2)         -0.0008*  

#>                   (0.0004)  

#>  

#> city               0.0542  

#>                   (0.0653)  

#>  

#> Constant           -0.5308***   -4.1815***   -4.1484***  

#>                   (0.2032)    (1.4024)    (1.4109)  

#>  

#> -----
#> Observations       428          753          753  

#> R2                 0.1581        0.1582  

#> Adjusted R2        0.1501        0.1482  

#> Log Likelihood     -914.0777  

#> rho                0.0830        0.0505 (0.2317)  

#> Inverse Mills Ratio 0.0551 (0.2099)  

#> Residual Std. Error 0.6667  

#> F Statistic         19.8561***  

#> ======  

#> Note:               *p<0.1; **p<0.05; ***p<0.01

```

Rho is an estimate of the correlation of the errors between the selection and wage equations. In the lower panel, the estimated coefficient on the inverse Mills ratio is given for the Heckman model. The fact that it is not statistically different from zero is consistent with the idea that selection bias was not a serious problem in this case.

If the estimated coefficient of the inverse Mills ratio in the Heckman model is not statistically different from zero, then selection bias was not a serious problem.

28.4.1.2 Example 2

This code is from R package sampleSelection

```
set.seed(0)
library("sampleSelection")
library("mvtnorm")
eps <- rmvnorm(500, c(0,0), matrix(c(1,-0.7,-0.7,1), 2, 2)) # bivariate normal disturbance
xs <- runif(500) # uniformly distributed explanatory variable (vectors of explanatory variables)
ys <- xs + eps[,1] > 0 # probit data generating process
xo <- runif(500) # vectors of explanatory variables for outcome equation
yoX <- xo + eps[,2] # latent outcome
yo <- yoX*(ys > 0) # observable outcome
# true intercepts = 0 and our true slopes = 1
# xs and xo are independent. Hence, exclusion restriction is fulfilled
summary(selection(ys~xs, yo ~xo))
#> -----
#> Tobit 2 model (sample selection model)
#> Maximum Likelihood estimation
#> Newton-Raphson maximisation, 5 iterations
#> Return code 1: gradient close to zero (gradtol)
#> Log-Likelihood: -712.3163
#> 500 observations (172 censored and 328 observed)
#> 6 free parameters (df = 494)
#> Probit selection equation:
#>           Estimate Std. Error t value Pr(>|t|)
#> (Intercept) -0.2228     0.1081  -2.061   0.0399 *
#> xs            1.3377     0.2014   6.642 8.18e-11 ***
#> Outcome equation:
#>           Estimate Std. Error t value Pr(>|t|)
#> (Intercept) -0.0002265  0.1294178  -0.002    0.999
#> xo            0.7299070  0.1635925   4.462 1.01e-05 ***
#> Error terms:
#>           Estimate Std. Error t value Pr(>|t|)
#> sigma      0.9190     0.0574 16.009 < 2e-16 ***
#> rho        -0.5392     0.1521 -3.544 0.000431 ***
#> ---
#> Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> -----
```

without the exclusion restriction, we generate yo using xs instead of xo.

```
yoX <- xs + eps[,2]
yo <- yoX*(ys > 0)
summary(selection(ys ~ xs, yo ~ xs))
```

```

#> -----
#> Tobit 2 model (sample selection model)
#> Maximum Likelihood estimation
#> Newton-Raphson maximisation, 14 iterations
#> Return code 8: successive function values within relative tolerance limit (reltol)
#> Log-Likelihood: -712.8298
#> 500 observations (172 censored and 328 observed)
#> 6 free parameters (df = 494)
#> Probit selection equation:
#>           Estimate Std. Error t value Pr(>|t|)
#> (Intercept) -0.1984     0.1114  -1.781   0.0756 .
#> xs          1.2907     0.2085   6.191 1.25e-09 ***
#> Outcome equation:
#>           Estimate Std. Error t value Pr(>|t|)
#> (Intercept) -0.5499     0.5644  -0.974  0.33038
#> xs          1.3987     0.4482   3.120  0.00191 **
#> Error terms:
#>           Estimate Std. Error t value Pr(>|t|)
#> sigma    0.85091    0.05352 15.899 <2e-16 ***
#> rho     -0.13226    0.72684 -0.182   0.856
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> -----

```

We can see that our estimates are still unbiased but standard errors are substantially larger. The exclusion restriction (i.e., independent information about the selection process) has a certain identifying power that we desire. Hence, it's better to have different set of variable for the selection process from the interested equation. Without the exclusion restriction, we solely rely on the functional form identification.

28.4.2 Tobit-5

Also known as the switching regression model

Condition: There is at least one variable in X in the selection process not included in the observed process. Used when there are separate models for participants, and non-participants.

```

set.seed(0)
vc <- diag(3)
vc[lower.tri(vc)] <- c(0.9, 0.5, 0.1)
vc[upper.tri(vc)] <- vc[lower.tri(vc)]
eps <- rmvnorm(500, c(0,0,0), vc) # 3 disturbance vectors by a 3-dimensional normal distribution
xs <- runif(500) # uniformly distributed on [0, 1]

```

```

ys <- xs + eps[,1] > 0
xo1 <- runif(500) # uniformly distributed on [0, 1]
yo1 <- xo1 + eps[,2]
xo2 <- runif(500) # uniformly distributed on [0, 1]
yo2 <- xo2 + eps[,3]

```

exclusion restriction is fulfilled when x's are independent.

```

summary(selection(ys~xs, list(yo1 ~ xo1, yo2 ~ xo2))) # one selection equation and a l
#> -----
#> Tobit 5 model (switching regression model)
#> Maximum Likelihood estimation
#> Newton-Raphson maximisation, 11 iterations
#> Return code 1: gradient close to zero (gradtol)
#> Log-Likelihood: -895.8201
#> 500 observations: 172 selection 1 (FALSE) and 328 selection 2 (TRUE)
#> 10 free parameters (df = 490)
#> Probit selection equation:
#>           Estimate Std. Error t value Pr(>|t|)
#> (Intercept) -0.1550    0.1051  -1.474   0.141
#> xs           1.1408    0.1785   6.390 3.86e-10 ***
#> Outcome equation 1:
#>           Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  0.02708   0.16395   0.165   0.869
#> xo1          0.83959   0.14968   5.609  3.4e-08 ***
#> Outcome equation 2:
#>           Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  0.1583    0.1885   0.840   0.401
#> xo2          0.8375    0.1707   4.908 1.26e-06 ***
#> Error terms:
#>           Estimate Std. Error t value Pr(>|t|)
#> sigma1      0.93191   0.09211  10.118 <2e-16 ***
#> sigma2      0.90697   0.04434  20.455 <2e-16 ***
#> rho1        0.88988   0.05353  16.623 <2e-16 ***
#> rho2        0.17695   0.33139   0.534   0.594
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> -----

```

All the estimates are close to the true values.

Example of functional form misspecification

```

set.seed(5)
eps <- rmvnorm(1000, rep(0, 3), vc)
eps <- eps^2 - 1 # subtract 1 in order to get the mean zero disturbances
xs <- runif(1000, -1, 0) # interval [-1, 0] to get an asymmetric distribution over observed choice
ys <- xs + eps[,1] > 0
xo1 <- runif(1000)
yo1 <- xo1 + eps[,2]
xo2 <- runif(1000)
yo2 <- xo2 + eps[,3]
summary(selection(ys~xs, list(yo1 ~ xo1, yo2 ~ xo2), iterlim=20))
#> -----
#> Tobit 5 model (switching regression model)
#> Maximum Likelihood estimation
#> Newton-Raphson maximisation, 4 iterations
#> Return code 3: Last step could not find a value above the current.
#> Boundary of parameter space?
#> Consider switching to a more robust optimisation method temporarily.
#> Log-Likelihood: -1665.936
#> 1000 observations: 760 selection 1 (FALSE) and 240 selection 2 (TRUE)
#> 10 free parameters (df = 990)
#> Probit selection equation:
#>           Estimate Std. Error t value Pr(>/t|)
#> (Intercept) -0.53698    0.05808  -9.245  < 2e-16 ***
#> xs          0.31268    0.09395   3.328  0.000906 ***
#> Outcome equation 1:
#>           Estimate Std. Error t value Pr(>/t|)
#> (Intercept) -0.70679    0.03573  -19.78   <2e-16 ***
#> xo1         0.91603    0.05626   16.28   <2e-16 ***
#> Outcome equation 2:
#>           Estimate Std. Error t value Pr(>/t|)
#> (Intercept)  0.1446      NA       NA       NA
#> xo2         1.1196    0.5014    2.233   0.0258 *
#> Error terms:
#>           Estimate Std. Error t value Pr(>/t|)
#> sigma1     0.67770    0.01760   38.50   <2e-16 ***
#> sigma2     2.31432    0.07615   30.39   <2e-16 ***
#> rho1      -0.97137      NA       NA       NA
#> rho2      0.17039      NA       NA       NA
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> -----

```

Although we still have an exclusion restriction (xo1 and xo2 are independent), we now have problems with the intercepts (i.e., they are statistically significantly different from the true values zero), and convergence problems.

If we don't have the exclusion restriction, we will have a larger variance of xs

```
set.seed(6)
xs <- runif(1000, -1, 1)
ys <- xs + eps[,1] > 0
yo1 <- xs + eps[,2]
yo2 <- xs + eps[,3]
summary(tmp <- selection(ys~xs, list(yo1 ~ xs, yo2 ~ xs), iterlim=20))
#> -----
#> Tobit 5 model (switching regression model)
#> Maximum Likelihood estimation
#> Newton-Raphson maximisation, 16 iterations
#> Return code 8: successive function values within relative tolerance limit (reltol)
#> Log-Likelihood: -1936.431
#> 1000 observations: 626 selection 1 (FALSE) and 374 selection 2 (TRUE)
#> 10 free parameters (df = 990)
#> Probit selection equation:
#>           Estimate Std. Error t value Pr(>/t|)
#> (Intercept) -0.3528     0.0424  -8.321 2.86e-16 ***
#> xs          0.8354     0.0756   11.050  < 2e-16 ***
#> Outcome equation 1:
#>           Estimate Std. Error t value Pr(>/t|)
#> (Intercept) -0.55448    0.06339  -8.748  < 2e-16 ***
#> xs          0.81764    0.06048   13.519  < 2e-16 ***
#> Outcome equation 2:
#>           Estimate Std. Error t value Pr(>/t|)
#> (Intercept)  0.6457     0.4994   1.293    0.196
#> xs          0.3520     0.3197   1.101    0.271
#> Error terms:
#>           Estimate Std. Error t value Pr(>/t|)
#> sigma1     0.59187    0.01853  31.935  < 2e-16 ***
#> sigma2     1.97257    0.07228  27.289  < 2e-16 ***
#> rho1      0.15568    0.15914   0.978    0.328
#> rho2     -0.01541    0.23370  -0.066    0.947
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> -----
```

Usually it will not converge. Even if it does, the results may be seriously biased.

Note

The log-likelihood function of the models might not be globally concave. Hence, it might not converge, or converge to a local maximum. To combat this, we can use

- Different starting value

- Different maximization methods.
- refer to Non-linear Least Squares for suggestions.

28.4.2.0.1 Pattern-Mixture Models

- compared to the Heckman's model where it assumes the value of the missing data is predetermined, pattern-mixture models assume missingness affect the distribution of variable of interest (e.g., Y)
- To read more, you can check NCSU, stefvanbuuren.

Chapter 29

Mediation

29.1 Traditional

(Baron and Kenny, 1986) is outdated because of step 1, but we could still see the original idea.

3 regressions

- Step 1: $X \rightarrow Y$
- Step 2: $X \rightarrow M$
- Step 3: $X + M \rightarrow Y$

where

- X = independent variable
- Y = dependent variable
- M = mediating variable

1. Originally, the first path from $X \rightarrow Y$ suggested by (Baron and Kenny, 1986) needs to be significant. But there are cases that you could have indirect effect of X on Y without significant direct effect of X on Y (e.g., when the effect is absorbed into M , or there are two counteracting effects M_1, M_2 that cancel out each other effect).

Mathematically,

$$Y = b_0 + b_1 X + \epsilon$$

b_1 does **not** need to be **significant**.

2. We examine the effect of X on M . This step requires that there is a significant effect of X on M to continue with the analysis

Mathematically,

$$M = b_0 + b_2 X + \epsilon$$

where b_2 needs to be **significant**.

3. In this step, we want to the effect of M on Y “absorbs” most of the direct effect of X on Y (or at least makes the effect smaller).

Mathematically,

$$Y = b_0 + b_4 X + b_3 M + \epsilon$$

b_4 needs to be either smaller or insignificant.

The effect of X on Y	then, M ... mediates between X and Y
completely disappear (b_4 insignificant)	Fully (i.e., full mediation)
partially disappear (b_4 smaller than in step 1)	Partially (i.e., partial mediation)

4. Examine the mediation effect (i.e., whether it is significant)

- First approach: Sobel’s test (Sobel, 1982)
- Second approach: bootstrapping (Preacher and Hayes, 2004) (preferable)

More details can be found here

29.1.1 Example 1

from Virginia’s library

```

myData <-  

  read.csv('http://static.lib.virginia.edu/statlab/materials/data/mediationData.csv')  
  

# Step 1 (no longer necessary)  

model.0 <- lm(Y ~ X, myData)  
  

# Step 2  

model.M <- lm(M ~ X, myData)  
  

# Step 3  

model.Y <- lm(Y ~ X + M, myData)  
  

# Step 4 (bootstrapping)  

library(mediation)  

results <- mediate(  

  model.M,  

  model.Y,  

  treat = 'X',  

  mediator = 'M',  

  boot = TRUE,  

  sims = 500
)  

summary(results)
#>
#> Causal Mediation Analysis
#>
#> Nonparametric Bootstrap Confidence Intervals with the Percentile Method
#>
#>           Estimate 95% CI Lower 95% CI Upper p-value
#> ACME          0.3565    0.2039     0.53 <2e-16 ***
#> ADE           0.0396   -0.2084     0.31  0.828
#> Total Effect  0.3961    0.1481     0.63  0.004 **
#> Prop. Mediated 0.9000    0.4768     2.07  0.004 **
#> ---  

#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Sample Size Used: 100
#>
#> Simulations: 500

```

- Total Effect = $0.3961 = b_1$ (step 1) = total effect of X on Y without M
- Direct Effect = ADE = $0.0396 = b_4$ (step 3) = direct effect of X on Y accounting for the indirect effect of M

- ACME = Average Causal Mediation Effects = $b_1 - b_4 = 0.3961 - 0.0396 = 0.3565 = b_2 \times b_3 = 0.56102 * 0.6355 = 0.3565$

Using `mediation` package suggested by (Imai et al., 2010a) (Imai et al., 2010b). More on details of the package can be found [here](#)

2 types of Inference in this package:

1. Model-based inference:

- Assumptions:
 - Treatment is randomized (could use matching methods to achieve this).
 - Sequential Ignorability: conditional on covariates, there is other confounders that affect the relationship between (1) treatment-mediator, (2) treatment-outcome, (3) mediator-outcome. Typically hard to argue in observational data. This assumption is for the identification of ACME (i.e., average causal mediation effects).

2. Design-based inference

Notations: we stay consistent with package instruction

- $M_i(t)$ = mediator
- T_i = treatment status (0,1)
- $Y_i(t, m)$ = outcome where t = treatment, and m = mediating variables.
- X_i = vector of observed pre-treatment confounders
- Treatment effect (per unit i) = $\tau_i = Y_i(1, M_i(1)) - Y_i(0, M_i(0))$ which has 2 effects
 - Causal mediation effects: $\delta_i(t) \equiv Y_i(t, M_i(1)) - Y_i(t, M_i(0))$
 - Direct effects: $\zeta_i(t) \equiv Y_i(1, M_i(1)) - Y_i(0, M_i(0))$
 - summing up to the treatment effect: $\tau_i = \delta_i(t) + \zeta_i(1 - t)$

More on sequential ignorability

$$\{Y_i(t', m), M_i(t)\} \perp T_i | X_i = x$$

$$Y_i(t', m) \perp M_i(t) | T_i = t, X_i = x$$

where

- $0 < P(T_i = t|X_i = x)$
- $0 < P(M_i = m|T_i = t, X_i = x)$

First condition is the standard strong ignorability condition where treatment assignment is random conditional on pre-treatment confounders.

Second condition is stronger where the mediators is also random given the observed treatment and pre-treatment confounders. This condition is satisfied only when there is no unobserved pre-treatment confounders, and post-treatment confounders, and multiple mediators that are correlated.

My understanding is that until the moment I write this note, there is **no way to test the sequential ignorability assumption**. Hence, researchers can only do sensitivity analysis to argue for their result.

29.2 Model-based causal mediation analysis

I only put my understanding of model-based causal mediation analysis because I do not encounter design-based. Maybe in the future when I have to use it, I will start reading on it.

Fit 2 models

- mediator model: conditional distribution of the mediators $M_i|T_i, X_i$
- Outcome model: conditional distribution of $Y_i|T_i, M_i, X_i$

`mediation` can accommodate almost all types of model for both mediator model and outcome model except Censored mediator model.

The update here is that estimation of ACME does not rely on product or difference of coefficients (see 29.1.1 ,

which requires very strict assumption: (1) linear regression models of mediator and outcome, (2) T_i and M_i effects are additive and no interaction

```
library(mediation)
set.seed(2014)
data("framing", package = "mediation")

med.fit <-
  lm(emo ~ treat + age + educ + gender + income, data = framing)
out.fit <-
  glm(
    cong_mesg ~ emo + treat + age + educ + gender + income,
```

```

    data = framing,
    family = binomial("probit")
  )

# Quasi-Bayesian Monte Carlo
med.out <-
  mediate(
    med.fit,
    out.fit,
    treat = "treat",
    mediator = "emo",
    robustSE = TRUE,
    sims = 1000 # should be 10000 in practice
  )
summary(med.out)
#>
#> Causal Mediation Analysis
#>
#> Quasi-Bayesian Confidence Intervals
#>
#>                               Estimate 95% CI Lower 95% CI Upper p-value
#> ACME (control)           0.0826   0.0356   0.14 <2e-16 ***
#> ACME (treated)          0.0831   0.0348   0.14 <2e-16 ***
#> ADE (control)            0.0137  -0.0967  0.13  0.82
#> ADE (treated)           0.0142  -0.1101  0.14  0.82
#> Total Effect             0.0968  -0.0290  0.23  0.14
#> Prop. Mediated (control) 0.7706  -6.3968  4.70  0.14
#> Prop. Mediated (treated) 0.7938  -5.7506  4.52  0.14
#> ACME (average)           0.0829   0.0351   0.14 <2e-16 ***
#> ADE (average)             0.0140  -0.1047  0.13  0.82
#> Prop. Mediated (average) 0.7822  -6.0737  4.61  0.14
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Sample Size Used: 265
#>
#> Simulations: 1000

```

Nonparametric bootstrap version

```

med.out <-
  mediate(
    med.fit,
    out.fit,

```

```

    boot = TRUE,
    treat = "treat",
    mediator = "emo",
    sims = 1000, # should be 10000 in practice
    boot.ci.type = "bca" # bias-corrected and accelerated intervals
)
summary(med.out)
#>
#> Causal Mediation Analysis
#>
#> Nonparametric Bootstrap Confidence Intervals with the BCa Method
#>
#>                               Estimate 95% CI Lower 95% CI Upper p-value
#> ACME (control)          0.0833   0.0386   0.15  0.002 **
#> ACME (treated)         0.0844   0.0374   0.15  0.002 **
#> ADE (control)           0.0114  -0.0875   0.13  0.792
#> ADE (treated)          0.0125  -0.1033   0.14  0.792
#> Total Effect            0.0958  -0.0291   0.23  0.124
#> Prop. Mediated (control) 0.8696 -97.4552   1.00  0.126
#> Prop. Mediated (treated) 0.8806 -82.9081   1.02  0.126
#> ACME (average)          0.0839   0.0381   0.15  0.002 **
#> ADE (average)            0.0120  -0.0961   0.14  0.792
#> Prop. Mediated (average) 0.8751 -90.6217   1.01  0.126
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Sample Size Used: 265
#>
#>
#> Simulations: 1000

```

If theoretically understanding suggests that there is treatment and mediator interaction

```

med.fit <-
  lm(emo ~ treat + age + educ + gender + income, data = framing)
out.fit <-
  glm(
    cong_mesg ~ emo * treat + age + educ + gender + income,
    data = framing,
    family = binomial("probit")
  )
med.out <-
  mediate(
    med.fit,

```

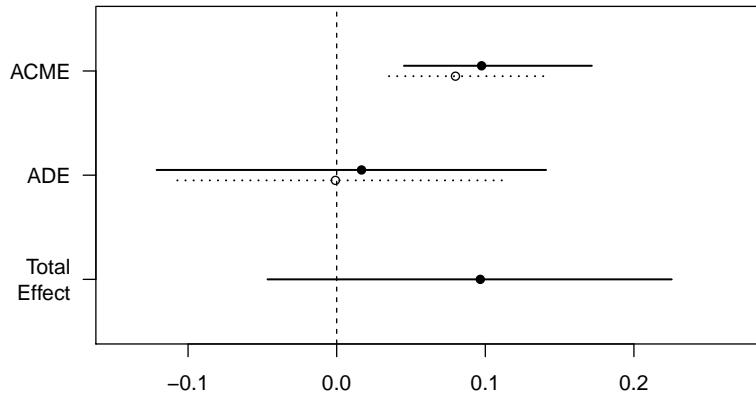
```

    out.fit,
    treat = "treat",
    mediator = "emo",
    robustSE = TRUE,
    sims = 100
)
summary(med.out)
#>
#> Causal Mediation Analysis
#>
#> Quasi-Bayesian Confidence Intervals
#>
#>                               Estimate 95% CI Lower 95% CI Upper p-value
#> ACME (control)          0.079925   0.035230      0.14 <2e-16 ***
#> ACME (treated)          0.097504   0.045279      0.17 <2e-16 ***
#> ADE (control)           -0.000865  -0.107228     0.11  0.98
#> ADE (treated)           0.016714   -0.121163     0.14  0.76
#> Total Effect            0.096640   -0.046523     0.23  0.26
#> Prop. Mediated (control) 0.672278   -5.266859    3.40  0.26
#> Prop. Mediated (treated) 0.860650   -6.754965    3.60  0.26
#> ACME (average)          0.088715   0.040207      0.15 <2e-16 ***
#> ADE (average)           0.007925   -0.111833     0.14  0.88
#> Prop. Mediated (average) 0.766464   -5.848496    3.43  0.26
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Sample Size Used: 265
#>
#> Simulations: 100

test.TMint(med.out, conf.level = .95) # test treatment-mediator interaction effect
#>
#> Test of ACME(1) - ACME(0) = 0
#>
#> data: estimates from med.out
#> ACME(1) - ACME(0) = 0.017579, p-value = 0.44
#> alternative hypothesis: true ACME(1) - ACME(0) is not equal to 0
#> 95 percent confidence interval:
#> -0.02676143  0.06257828

```

```
plot(med.out)
```



`mediation` can be used in conjunction with any of your imputation packages.

And it can also handle **mediated moderation** or **non-binary treatment variables**, or **multi-level data**

Sensitivity Analysis for sequential ignorability

- test for unobserved pre-treatment covariates
- ρ = correlation between the residuals of the mediator and outcome regressions.
- If ρ is significant, we have evidence for violation of sequential ignorability (i.e., there is unobserved pre-treatment confounders).

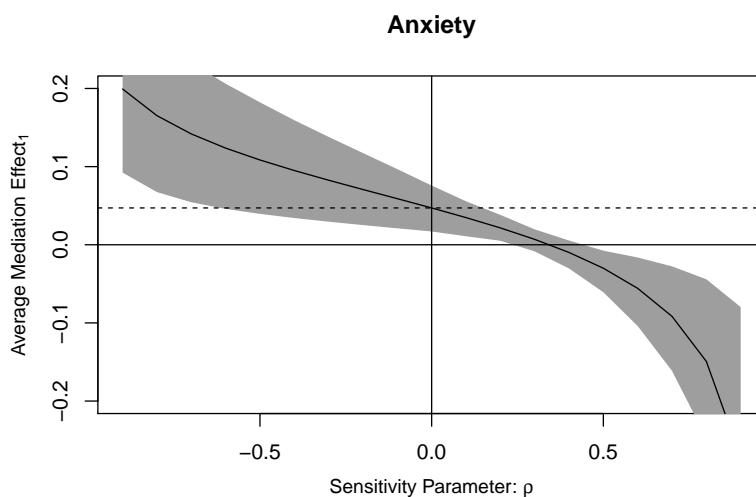
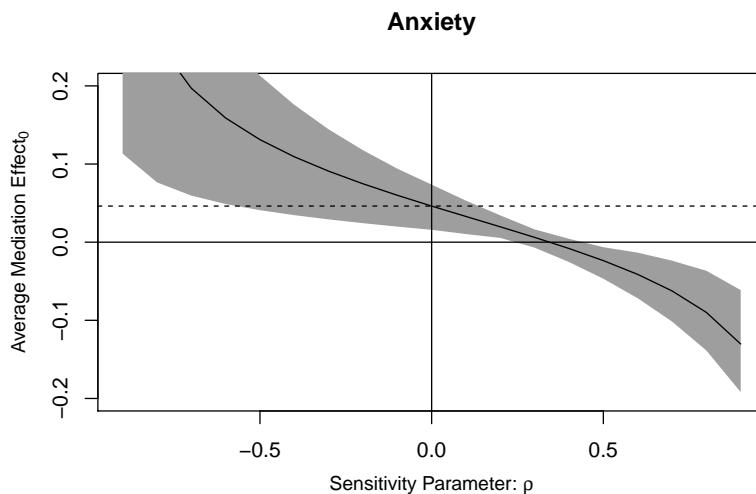
```
med.fit <-
  lm(emo ~ treat + age + educ + gender + income, data = framing)
out.fit <-
  glm(
    cong_mesg ~ emo + treat + age + educ + gender + income,
    data = framing,
    family = binomial("probit")
  )
med.out <-
  mediate(
    med.fit,
    out.fit,
    treat = "treat",
```

```

    mediator = "emo",
    robustSE = TRUE,
    sims = 100
)
sens.out <-
  medsens(med.out,
    rho.by = 0.1, # \rho varies from -0.9 to 0.9 by 0.1
    effect.type = "indirect", # sensitivity on ACME
    # effect.type = "direct", # sensitivity on ADE
    # effect.type = "both", # sensitivity on ACME and ADE
    sims = 100)
summary(sens.out)
#>
#> Mediation Sensitivity Analysis: Average Mediation Effect
#>
#> Sensitivity Region: ACME for Control Group
#>
#>   Rho ACME(control) 95% CI Lower 95% CI Upper R^2_M*R^2_Y* R^2_M~R^2_Y~
#> [1,] 0.3      0.0061     -0.0070      0.0163      0.09      0.0493
#> [2,] 0.4     -0.0081     -0.0254      0.0043      0.16      0.0877
#>
#> Rho at which ACME for Control Group = 0: 0.3
#> R^2_M*R^2_Y* at which ACME for Control Group = 0: 0.09
#> R^2_M~R^2_Y~ at which ACME for Control Group = 0: 0.0493
#>
#> Sensitivity Region: ACME for Treatment Group
#>
#>   Rho ACME(treated) 95% CI Lower 95% CI Upper R^2_M*R^2_Y* R^2_M~R^2_Y~
#> [1,] 0.3      0.0069     -0.0085      0.0197      0.09      0.0493
#> [2,] 0.4     -0.0099     -0.0304      0.0054      0.16      0.0877
#>
#> Rho at which ACME for Treatment Group = 0: 0.3
#> R^2_M*R^2_Y* at which ACME for Treatment Group = 0: 0.09
#> R^2_M~R^2_Y~ at which ACME for Treatment Group = 0: 0.0493

plot(sens.out, sens.par = "rho", main = "Anxiety", ylim = c(-0.2, 0.2))

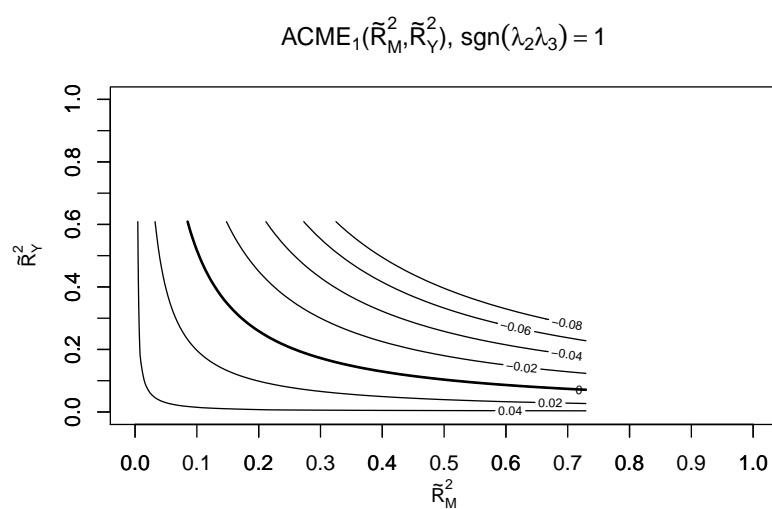
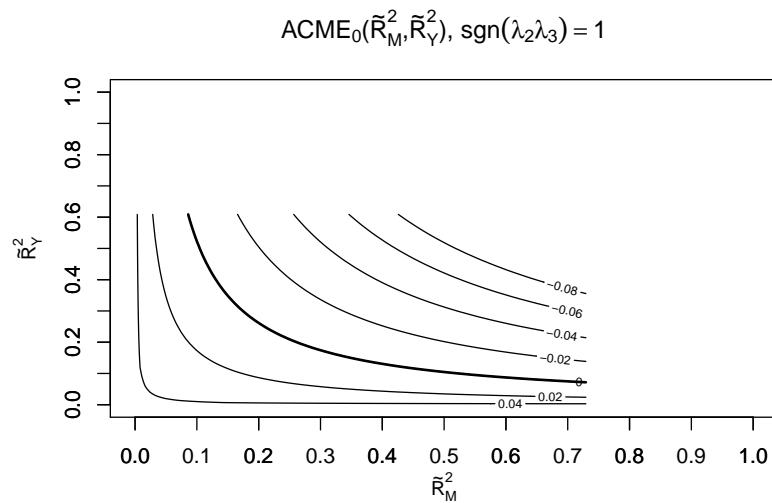
```



ACME confidence intervals contains 0 when $\rho \in (0.3, 0.4)$

Alternatively, using R^2 interpretation, we need to specify the direction of confounder that affects the mediator and outcome variables in plot using `sign.prod = "positive"` (i.e., same direction) or `sign.prod = "negative"` (i.e., opposite direction).

```
plot(sens.out, sens.par = "R2", r.type = "total", sign.prod = "positive")
```



Chapter 30

Directed Acyclic Graph

Native R:

- `dagitty`
- `ggdag`
- `dagR`
- `r-causal`: by Center for Causal Discovery. Also available in Python

Publication-ready (with R and Latex): shinyDAG

Standalone program: DAG program by Sven Knuppel

V. MISCELLANEOUS

Chapter 31

Report

Structure

- Exploratory analysis
 - plots
 - preliminary results
 - interesting structure/features in the data
 - outliers
- Model
 - Assumptions
 - Why this model/ How is this model the best one?
 - Consideration: interactions, collinearity, dependence
- Model Fit
 - How well does it fit?
 - Are the model assumptions met?
 - * Residual analysis
- Inference/ Prediction
 - Are there different way to support your inference?
- Conclusion
 - Recommendation
 - Limitation of the analysis
 - How to correct those in the future

This chapter is based on the `jtools` package. More information can be found here.

31.1 One summary table

Packages for reporting:

Summary Statistics Table:

- qwraps2
- vtable
- gtsummary
- apaTables
- stargazer

Regression Table

- gtsummary
- sjPlot, sjmisc, sjlabelled
- stargazer: recommended (Example)
- modelsummary

```
library(jtools)
data(movies)
fit <- lm(metascore ~ budget + us_gross + year, data = movies)
summ(fit)
```

Observations	831 (10 missing obs. deleted)
Dependent variable	metascore
Type	OLS linear regression

F(3,827)	26.23
R ²	0.09
Adj. R ²	0.08

	Est.	S.E.	t val.	p
(Intercept)	52.06	139.67	0.37	0.71
budget	-0.00	0.00	-5.89	0.00
us_gross	0.00	0.00	7.61	0.00
year	0.01	0.07	0.08	0.94

Standard errors: OLS

```
summ(fit, scale = TRUE, vifs = TRUE, part.corr = TRUE, confint = TRUE, pvals = FALSE) #notice the
```

Observations	831 (10 missing obs. deleted)
Dependent variable	metascore
Type	OLS linear regression

F(3,827)	26.23
R ²	0.09
Adj. R ²	0.08

	Est.	2.5%	97.5%	t val.	VIF	partial.r	part.r
(Intercept)	63.01	61.91	64.11	112.23	NA	NA	NA
budget	-3.78	-5.05	-2.52	-5.89	1.31	-0.20	-0.20
us_gross	5.28	3.92	6.64	7.61	1.52	0.26	0.25
year	0.05	-1.18	1.28	0.08	1.24	0.00	0.00

Standard errors: OLS; Continuous predictors are mean-centered and scaled by 1 s.d.

```
#obtain cluster-robust SE
data("PetersenCL", package = "sandwich")
fit2 <- lm(y ~ x, data = PetersenCL)
summ(fit2, robust = "HC3", cluster = "firm")
```

Observations	5000
Dependent variable	y
Type	OLS linear regression

F(1,4998)	1310.74
R ²	0.21
Adj. R ²	0.21

	Est.	S.E.	t val.	p
(Intercept)	0.03	0.07	0.44	0.66
x	1.03	0.05	20.36	0.00

Standard errors: Cluster-robust, type = HC3

Model to Equation

```
# install.packages("equatiomatic")
fit <- lm(metascore ~ budget + us_gross + year, data = movies)
# show the theoretical model
equatiomatic::extract_eq(fit)
```

$$\text{metascore} = \alpha + \beta_1(\text{budget}) + \beta_2(\text{us_gross}) + \beta_3(\text{year}) + \epsilon \quad (31.1)$$

```
# display the actual coefficients
equatiomatic::extract_eq(fit, use_coefs = TRUE)
```

$$\hat{\text{metascore}} = 52.06 + 0(\text{budget}) + 0(\text{us_gross}) + 0.01(\text{year}) \quad (31.2)$$

31.2 Model Comparison

```
fit <- lm(metascore ~ log(budget), data = movies)
fit_b <- lm(metascore ~ log(budget) + log(us_gross), data = movies)
fit_c <- lm(metascore ~ log(budget) + log(us_gross) + runtime, data = movies)
coef_names <- c("Budget" = "log(budget)", "US Gross" = "log(us_gross)",
                 "Runtime (Hours)" = "runtime", "Constant" = "(Intercept)")
export_summs(fit, fit_b, fit_c, robust = "HC3", coefs = coef_names)
```

Another package is `modelsummary`

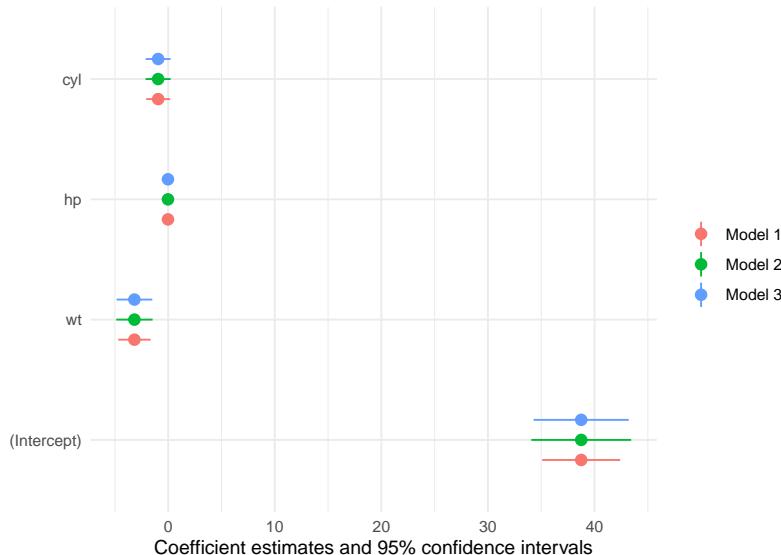
```
library(modelsummary)
lm_mod <- lm(mpg ~ wt + hp + cyl, mtcars)
msummary(lm_mod, vcov = c("iid", "robust", "HC4"))
```

```
modelplot(lm_mod, vcov = c("iid", "robust", "HC4"))
```

Table 31.1

	Model 1	Model 2	Model 3
Budget	-2.43 *** (0.44)	-5.16 *** (0.62)	-6.70 *** (0.67)
US Gross		3.96 *** (0.51)	3.85 *** (0.48)
Runtime (Hours)			14.29 *** (1.63)
Constant	105.29 *** (7.65)	81.84 *** (8.66)	83.35 *** (8.82)
N	831	831	831
R2	0.03	0.09	0.17

Standard errors are heteroskedasticity robust. *** p < 0.001; ** p < 0.01; * p < 0.05.



Another package is **stargazer**

	Model 1	Model 2	Model 3
(Intercept)	38.752 (1.787)	38.752 (2.286)	38.752 (2.177)
wt	-3.167 (0.741)	-3.167 (0.833)	-3.167 (0.819)
hp	-0.018 (0.012)	-0.018 (0.010)	-0.018 (0.013)
cyl	-0.942 (0.551)	-0.942 (0.573)	-0.942 (0.572)
Num.Obs.	32	32	32
R2	0.843	0.843	0.843
R2 Adj.	0.826	0.826	0.826
AIC	155.5	155.5	155.5
BIC	162.8	162.8	162.8
Log.Lik.	-72.738	-72.738	-72.738
F	50.171	31.065	32.623
Std.Errors	IID	Robust	HC4

```

library("stargazer")
stargazer(attitude)
#>
## Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: h
## Date and time: Thu, Jun 09, 2022 - 10:37:34 PM
## \begin{table}[,htbp] \centering
##   \caption{ }
##   \label{ }
##   \begin{tabular}{@{\extracolsep{5pt}}lcccccc}
##     \hline
##     & \multicolumn{1}{c}{N} & \multicolumn{1}{c}{Mean} & \multicolumn{1}{c}{SD} \\
##     \hline
##     rating & 30 & 64.633 & 12.173 & 40 & 58.8 & 71.8 & 85 \\
##     complaints & 30 & 66.600 & 13.315 & 37 & 58.5 & 77 & 90 \\
##     privileges & 30 & 53.133 & 12.235 & 30 & 45 & 62.5 & 83 \\
##     learning & 30 & 56.367 & 11.737 & 34 & 47 & 66.8 & 75 \\
##     raises & 30 & 64.633 & 10.397 & 43 & 58.2 & 71 & 88 \\
##     critical & 30 & 74.767 & 9.895 & 49 & 69.2 & 80 & 92 \\
##     advance & 30 & 42.933 & 10.289 & 25 & 35 & 47.8 & 72 \\
##     \hline
##     \end{tabular}
##   \end{table}
## 2 OLS models
linear.1 <- lm(rating ~ complaints + privileges + learning + raises + critical,data =

```

```

linear.2 <- lm(rating ~ complaints + privileges + learning, data = attitude)
## create an indicator dependent variable, and run a probit model
attitude$high.rating <- (attitude$rating > 70)
probit.model <-
  glm(
    high.rating ~ learning + critical + advance,
    data = attitude,
    family = binomial(link = "probit")
  )
stargazer(linear.1,
          linear.2,
          probit.model,
          title = "Results",
          align = TRUE)
#>
#> % Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at f
#> % Date and time: Thu, Jun 09, 2022 - 10:37:34 PM
#> % Requires LaTeX packages: dcolumn
#> \begin{table}[\!htbp] \centering
#>   \caption{Results}
#>   \label{f}
#> \begin{tabular}{@{\extracolsep{5pt}}lD{.}{.}{-3} D{.}{.}{-3} D{.}{.}{-3} }
#> \hline
#> \hline
#>   & \multicolumn{3}{c}{\textit{Dependent variable:}} \\
#> \cline{2-4}
#> \hline
#>   & \multicolumn{2}{c}{rating} & \multicolumn{1}{c}{high.rating} \\
#> \hline
#>   & \multicolumn{2}{c}{\textit{OLS}} & \multicolumn{1}{c}{\textit{probit}} \\
#> \hline
#>   & \multicolumn{1}{c}{(1)} & \multicolumn{1}{c}{(2)} & \multicolumn{1}{c}{(3)} \\
#> \hline
#>   complaints & 0.692^{***} & 0.682^{***} & \\
#>   & (0.149) & (0.129) & \\
#>   & & & \\
#>   & privileges & -0.104 & -0.103 \\
#>   & & (0.135) & (0.129) \\
#>   & & & \\
#>   & learning & 0.249 & 0.238^{*} \\
#>   & & (0.160) & (0.139) \\
#>   & & & (0.053) \\
#>   & & & \\
#>   & raises & -0.033 & \\
#>   & & (0.202) & \\
#>   & & & \\
#>   & critical & 0.015 & -0.001 \\
#>   & & (0.147) & (0.044) \\
#>   & & & \\
#> \hline

```

```

#> advance & & & -0.062 \\
#> & & & (0.042) \\
#> & & & \\
#> Constant & 11.011 & 11.258 & -7.476^{**} \\
#> & (11.704) & (7.318) & (3.570) \\
#> & & & \\
#> \hline \\[-1.8ex]
#> Observations & \multicolumn{1}{c}{30} & \multicolumn{1}{c}{30} & \multicolumn{1}{c}{c}\\
#> R$^2\$ & \multicolumn{1}{c}{0.715} & \multicolumn{1}{c}{c}{0.715} & \\
#> Adjusted R$^2\$ & \multicolumn{1}{c}{0.656} & \multicolumn{1}{c}{c}{0.682} & \\
#> Log Likelihood & & & \multicolumn{1}{c}{-9.087} \\
#> Akaike Inf. Crit. & & & \multicolumn{1}{c}{26.175} \\
#> Residual Std. Error & \multicolumn{1}{c}{7.139 (df = 24)} & \multicolumn{1}{c}{c}{6.86} \\
#> F Statistic & \multicolumn{1}{c}{12.063\$^{***} (df = 5; 24)} & \multicolumn{1}{c}{c}{12.063\$^{***} (df = 5; 24)} & \\
#> \hline
#> \hline \\[-1.8ex]
#> \textit{[Note:]} & \multicolumn{3}{r}{\$^{*}\$p\$<\$0.1; \$^{**}\$p\$<\$0.05; \$^{***}\$p\$<\$0.01}
#> \end{tabular}
#> \end{table}

# Latex
stargazer(
  linear.1,
  linear.2,
  probit.model,
  title = "Regression Results",
  align = TRUE,
  dep.var.labels = c("Overall Rating", "High Rating"),
  covariate.labels = c(
    "Handling of Complaints",
    "No Special Privileges",
    "Opportunity to Learn",
    "Performance-Based Raises",
    "Too Critical",
    "Advancement"
  ),
  omit.stat = c("LL", "ser", "f"),
  no.space = TRUE
)

# ASCII text output
stargazer(
  linear.1,
  linear.2,
  type = "text",

```

```

title = "Regression Results",
dep.var.labels = c("Overall Rating", "High Rating"),
covariate.labels = c(
  "Handling of Complaints",
  "No Special Privileges",
  "Opportunity to Learn",
  "Performance-Based Raises",
  "Too Critical",
  "Advancement"
),
omit.stat = c("LL", "ser", "f"),
ci = TRUE,
ci.level = 0.90,
single.row = TRUE
)
#>
#> Regression Results
#> =====
#> Dependent variable:
#> -----
#> Overall Rating
#> (1) (2)
#> -----
#> Handling of Complaints 0.692*** (0.447, 0.937) 0.682*** (0.470, 0.894)
#> No Special Privileges -0.104 (-0.325, 0.118) -0.103 (-0.316, 0.109)
#> Opportunity to Learn 0.249 (-0.013, 0.512) 0.238* (0.009, 0.467)
#> Performance-Based Raises -0.033 (-0.366, 0.299)
#> Too Critical 0.015 (-0.227, 0.258)
#> Advancement 11.011 (-8.240, 30.262) 11.258 (-0.779, 23.296)
#> -----
#> Observations 30 30
#> R2 0.715 0.715
#> Adjusted R2 0.656 0.682
#> =====
#> Note: *p<0.1; **p<0.05; ***p<0.01

stargazer(
  linear.1,
  linear.2,
  probit.model,
  title = "Regression Results",
  align = TRUE,
  dep.var.labels = c("Overall Rating", "High Rating"),
  covariate.labels = c(
    "Handling of Complaints",

```

```

    "No Special Privileges",
    "Opportunity to Learn",
    "Performance-Based Raises",
    "Too Critical",
    "Advancement"
),
omit.stat = c("LL", "ser", "f"),
no.space = TRUE
)

```

Correlation Table

```

correlation.matrix <- cor(attitude[,c("rating","complaints","privileges")])
stargazer(correlation.matrix, title="Correlation Matrix")

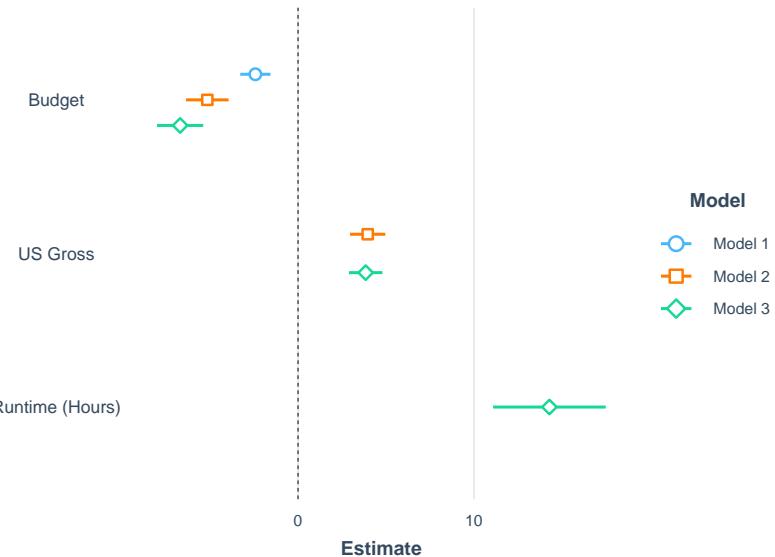
```

31.3 Changes in an estimate

```

coef_names <- coef_names[1:3] # Dropping intercept for plots
plot_summs(fit, fit_b, fit_c, robust = "HC3", coefs = coef_names)

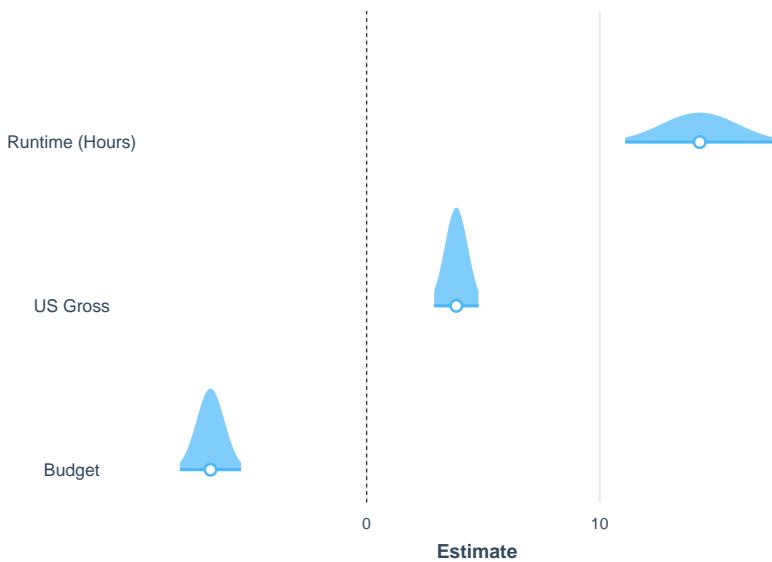
```



```

plot_summs(fit_c, robust = "HC3", coefs = coef_names, plot.distributions = TRUE)

```



Chapter 32

Exploratory Data Analysis

```
# load to get txhousing data
library(ggplot2)
```

Data Report

Feature Engineering

Missing Data

```
# install.packages("DataExplorer")
library(DataExplorer)

# creat a html file that contain all reports
create_report(txhousing)

introduce() # see basic info

dummify() # create binary columns from discrete variables
split_columns() # split data into discrete and continuous parts

plot_correlation() # heatmap for discrete var
plot_intro()

plot_missing() # plot missing value
profile_missing() # profile missing values
```

```
plot_prcmp() # plot PCA
```

Error Identification

```
# install.packages("dataReporter")
library(dataReporter)
makeDataReport() # detailed report like DataExplorer
```

Summary statistics

```
library(skimr)
skim() # give only few quick summary stat, not as detailed as the other two packages
```

Not so code-y process

Quick and dirty way to look at your data

```
# install.packages("rpivotTable")
library(rpivotTable)
# give set up just like Excel table
data %>%
  rpivotTable::rpivotTable()
```

Code generation and wrangling

Shiny-app based Tableau style

```
# install.packages("esquisse")
library(esquisse)
esquisse::esquisser()
```

Customized your daily/automatic report

```
# install.packages("chronicle")
library(chronicle)
```

```
# install.packages("dlookr")
# install.packages("descriptr")
```

Chapter 33

Sensitivity Analysis/ Robustness Check

33.1 Specification curve

- also known as Specification robustness graph or coefficient stability plot

Resources

- In Stata or speccurve
- (Simonsohn et al., 2020)

33.1.1 starbility

- Recommend

Installation

```
devtools::install_github('https://github.com/AakaashRao/starbility')
library(starbility)
```

Example by the package's author

```
library(tidyverse)
library(starbility)
library(lfe)
```

```
data("diamonds")
set.seed(43)
indices = sample(1:nrow(diamonds),
                 replace = F,
                 size = round(nrow(diamonds) / 20))
diamonds = diamonds[indices, ]
```

Plot different combinations of controls

```
# If you want to make the diamond dimensions as base control
base_controls = c(
  'Diamond dimensions' = 'x + y + z' # include all variables under 1 dimension
)

perm_controls = c(
  'Depth' = 'depth',
  'Table width' = 'table'
)

nonperm_fe_controls = c(
  'Clarity FE (granular)' = 'clarity',
  'Clarity FE (binary)' = 'high_clarity'
)

# Adding fixed effects
nonperm_fe_controls = c(
  'Clarity FE (granular)' = 'clarity',
  'Clarity FE (binary)' = 'high_clarity'
)

# Adding instrumental variables
instruments = 'x+y+z'

# clustering and weights
diamonds$sample_weights = runif(n = nrow(diamonds))

# robust standard errors
starb_felm_custom = function(spec, data, rhs, ...) {
  spec = as.formula(spec)
  model = lfe::felm(spec, data=data) %>% broom::tidy()

  row = which(model$term==rhs)
  coef = model[row, 'estimate'] %>% as.numeric()
```

```

se    = model[row, 'std.error'] %>% as.numeric()
p    = model[row, 'p.value'] %>% as.numeric()

# 99% confidence interval
z = qnorm(0.995)
# one-tailed test
return(c(coef, p/2, coef+z*se, coef-z*se))
}

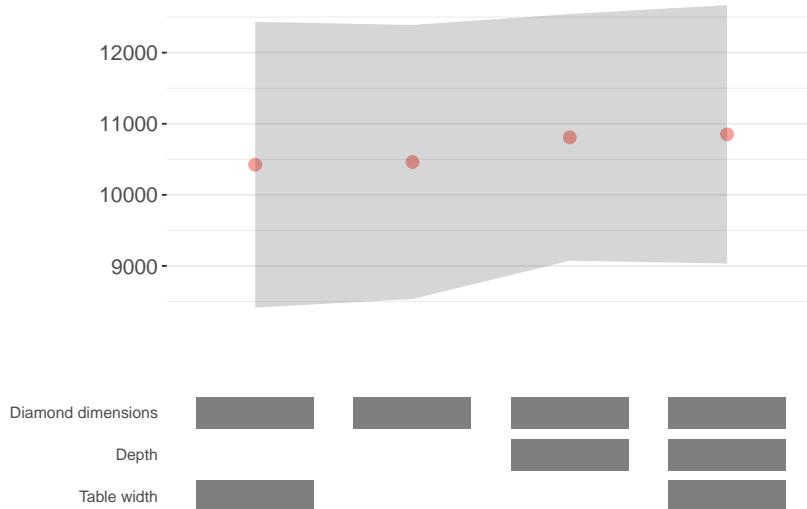
plots = stability_plot(
  data = diamonds,
  lhs = 'price',
  rhs = 'carat',
  error_geom = 'ribbon', # make the plot more aesthetics
  # error_geom = 'none', # if you don't want ribbon (i.e., error bar)
  model = starb_felm_custom,
  cluster = 'cut',
  weights = 'sample_weights',
  # iv = instruments,
  perm = perm_controls,
  base = base_controls,
  # perm_fe = perm_fe_controls,
  # nonperm_fe = nonperm_fe_controls, # if you want to include fixed effects sequentially (not
  # fe_always = F, # if you want to have a model without any Fixed Effects
  sort = "asc-by-fe", # sort "asc", "desc", or by fixed effects: "asc-by-fe" or "desc-by-fe"

  # if you have less variables and want more aesthetics
  # control_geom = 'circle',
  # point_size = 2,
  # control_spacing = 0.3,

  # error_alpha = 0.2, # change alpha of the error geom
  # point_size = 1.5, # change the size of the coefficient points
  # control_text_size = 10, # change the size of the control labels
  # coef_ylim = c(-5000, 35000), # change the endpoints of the y-axis
  # trip_top = 3, # change the spacing between the two panels

  rel_height = 0.6
)
plots

```



```
# add comments
# replacement_coef_panel = plots[[1]] +
#   scale_y_reverse() +
#   theme(panel.grid.minor = element_blank()) +
#   geom_vline(xintercept = 41, linetype = 'dashed', alpha = 0.4) +
#   annotate(geom = 'label', x = 52, y = 30000, label = 'What a great\ncspecification!')
#
# combine_plots(replacement_coef_panel,
#               plots[[2]],
#               rel_height = 0.6)
```

Note:

- $p < 0.01$: red
- $p < 0.05$: green
- $p < 0.1$: blue
- $p > 0.1$: black

More Advanced Stuff

```
# Step 1: Control Grid
diamonds$high_clarity = diamonds$clarity %in% c('VS1', 'VVS2', 'VVS1', 'IF')
base_controls = c(
```

```

'Diamond dimensions' = 'x + y + z'
)

perm_controls = c(
  'Depth' = 'depth',
  'Table width' = 'table'
)

perm_fe_controls = c(
  'Cut FE' = 'cut',
  'Color FE' = 'color'
)
nonperm_fe_controls = c(
  'Clarity FE (granular)' = 'clarity',
  'Clarity FE (binary)' = 'high_clarity'
)

grid1 = stability_plot(data = diamonds,
                       lhs = 'price',
                       rhs = 'carat',
                       perm = perm_controls,
                       base = base_controls,
                       perm_fe = perm_fe_controls,
                       nonperm_fe = nonperm_fe_controls,
                       run_to=2)

knitr::kable(grid1 %>% head(10))

```

Diamond dimensions	Depth	Table width	Cut FE	Color FE	np_fe
1	0	0	0	0	
1	1	0	0	0	
1	0	1	0	0	
1	1	1	0	0	
1	0	0	1	0	
1	1	0	1	0	
1	0	1	1	0	
1	1	1	1	0	
1	0	0	0	1	
1	1	0	0	1	

Step 2: Get model expression

```

grid2 = stability_plot(grid = grid1,
                       data=diamonds,

```

```

    lhs='price',
    rhs='carat',
    perm=perm_controls,
    base=base_controls,
    run_from=2,
    run_to=3)

```

```
knitr:::kable(grid2 %>% head(10))
```

Diamond dimensions	Depth	Table width	np_fe	expr
1	0	0	0	price~carat+x+y+z 0 0 0
1	1	0	0	price~carat+x+y+z+depth 0 0 0
1	0	1	0	price~carat+x+y+z+table 0 0 0
1	1	1	0	price~carat+x+y+z+depth+table 0 0 0
1	0	0	0	price~carat+x+y+z 0 0 0
1	1	0	0	price~carat+x+y+z+depth 0 0 0
1	0	1	0	price~carat+x+y+z+table 0 0 0
1	1	1	0	price~carat+x+y+z+depth+table 0 0 0
1	0	0	0	price~carat+x+y+z 0 0 0
1	1	0	0	price~carat+x+y+z+depth 0 0 0

```

# Step 3: Estimate models
grid3 = stability_plot(grid = grid2,
                       data=diamonds,
                       lhs='price',
                       rhs='carat',
                       perm=perm_controls,
                       base=base_controls,
                       run_from=3,
                       run_to=4)

knitr:::kable(grid3 %>% head(10))

```

Diamond dimensions	Depth	Table width	np_fe	expr	coef	p
1	0	0	0	price~carat+x+y+z 0 0 0	10461.86	p<0
1	1	0	0	price~carat+x+y+z+depth 0 0 0	10808.25	p<0
1	0	1	0	price~carat+x+y+z+table 0 0 0	10423.42	p<0
1	1	1	0	price~carat+x+y+z+depth+table 0 0 0	10851.31	p<0
1	0	0	0	price~carat+x+y+z 0 0 0	10461.86	p<0
1	1	0	0	price~carat+x+y+z+depth 0 0 0	10808.25	p<0
1	0	1	0	price~carat+x+y+z+table 0 0 0	10423.42	p<0
1	1	1	0	price~carat+x+y+z+depth+table 0 0 0	10851.31	p<0
1	0	0	0	price~carat+x+y+z 0 0 0	10461.86	p<0
1	1	0	0	price~carat+x+y+z+depth 0 0 0	10808.25	p<0

```
# Step 4: Get dataframe to draw
dfs = stability_plot(grid = grid3,
                      data=diamonds,
                      lhs='price',
                      rhs='carat',
                      perm=perm_controls,
                      base=base_controls,
                      run_from=4,
                      run_to=5)

coef_grid = dfs[[1]]
control_grid = dfs[[2]]

knitr::kable(coef_grid %>% head(10))
```

Diamond dimensions	Depth	Table width	np_fe	expr	coef	p
1	0	0	0	price~carat+x+y+z 0 0 0	10461.86	p<0
1	1	0	0	price~carat+x+y+z+depth 0 0 0	10808.25	p<0
1	0	1	0	price~carat+x+y+z+table 0 0 0	10423.42	p<0
1	1	1	0	price~carat+x+y+z+depth+table 0 0 0	10851.31	p<0
1	0	0	0	price~carat+x+y+z 0 0 0	10461.86	p<0
1	1	0	0	price~carat+x+y+z+depth 0 0 0	10808.25	p<0
1	0	1	0	price~carat+x+y+z+table 0 0 0	10423.42	p<0
1	1	1	0	price~carat+x+y+z+depth+table 0 0 0	10851.31	p<0
1	0	0	0	price~carat+x+y+z 0 0 0	10461.86	p<0
1	1	0	0	price~carat+x+y+z+depth 0 0 0	10808.25	p<0

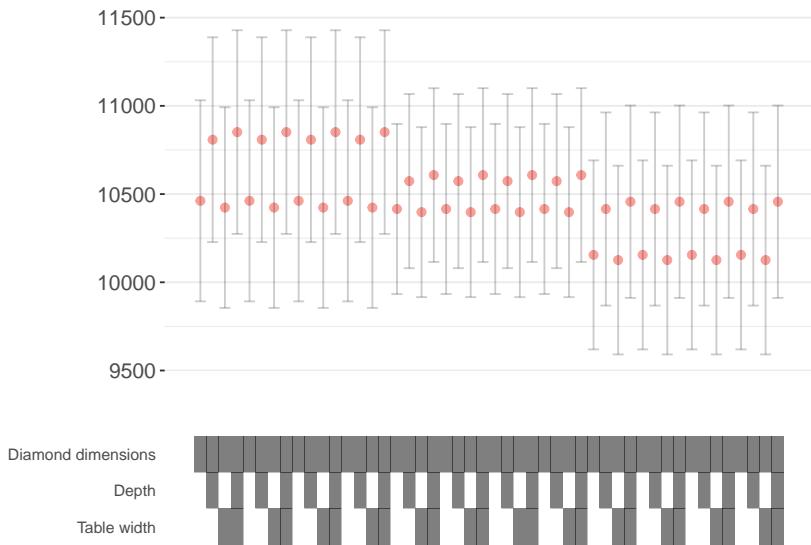
```
# Step 5: plot the sensitivity graph
panels = stability_plot(data = diamonds,
                        lhs='price',
                        rhs='carat',
```

```

coef_grid = coef_grid,
control_grid = control_grid,
run_from=5,
run_to=6)

stability_plot(data = diamonds,
lhs='price',
rhs='carat',
coef_panel = panels[[1]],
control_panel = panels[[2]],
run_from = 6,
run_to = 7)

```



In step 2, we can modify to use other function (e.g., `glm`)

```

diamonds$above_med_price = as.numeric(diamonds$price > median(diamonds$price))

base_controls = c('Diamond dimensions' = 'x + y + z')

perm_controls = c('Depth' = 'depth',
                  'Table width' = 'table',
                  'Clarity' = 'clarity')
lhs_var = 'above_med_price'
rhs_var = 'carat'

```

```

grid1 = stability_plot(
  data = diamonds,
  lhs = lhs_var,
  rhs = rhs_var,
  perm = perm_controls,
  base = base_controls,
  fe_always = F,
  run_to = 2
)

# Create control part of formula
base_perm = c(base_controls, perm_controls)
grid1$expr = apply(grid1[, 1:length(base_perm)], 1,
  function(x)
    paste(base_perm[names(base_perm)[which(x == 1)]], collapse = '+'))

# Complete formula with LHS and RHS variables
grid1$expr = paste(lhs_var, '~', rhs_var, '+', grid1$expr, sep = '')

knitr::kable(grid1 %>% head(10))

```

Diamond dimensions	Depth	Table width	Clarity	np_fe	expr
1	0	0	0		above_med_price~carat+x + y + z
1	1	0	0		above_med_price~carat+x + y + z+depth
1	0	1	0		above_med_price~carat+x + y + z+table
1	1	1	0		above_med_price~carat+x + y + z+depth+ta
1	0	0	1		above_med_price~carat+x + y + z+clarity
1	1	0	1		above_med_price~carat+x + y + z+depth+cla
1	0	1	1		above_med_price~carat+x + y + z+table+cla
1	1	1	1		above_med_price~carat+x + y + z+depth+ta

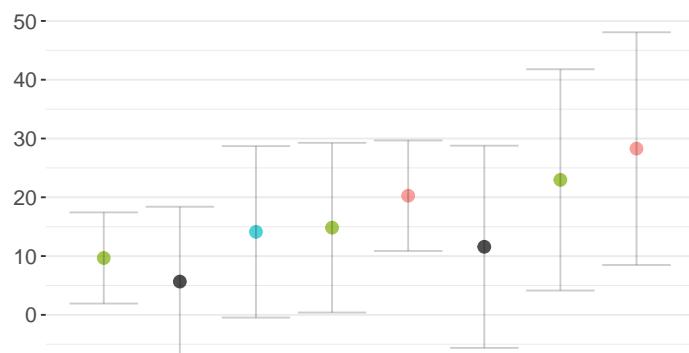
```

# customer function for the logit model
starb_logit = function(spec, data, rhs, ...) {
  spec = as.formula(spec)
  model = glm(spec, data=data, family='binomial', weights=data$weight) %>%
    broom::tidy()
  row = which(model$term==rhs)
  coef = model[row, 'estimate'] %>% as.numeric()
  se = model[row, 'std.error'] %>% as.numeric()
  p = model[row, 'p.value'] %>% as.numeric()

  return(c(coef, p, coef+1.96*se, coef-1.96*se))
}

```

```
stability_plot(grid = grid1,
               data = diamonds,
               lhs = lhs_var,
               rhs = rhs_var,
               model = starb_logit,
               perm = perm_controls,
               base = base_controls,
               fe_always = F,
               run_from=3)
```



For getting other specification (e.g., different CI)

```
library(margins)
starb_logit_enhanced = function(spec, data, rhs, ...) {
  # Unpack ...
  l = list(...)
  get_mfx = ifelse(is.null(l$get_mfx), F, T) # Set a default to F

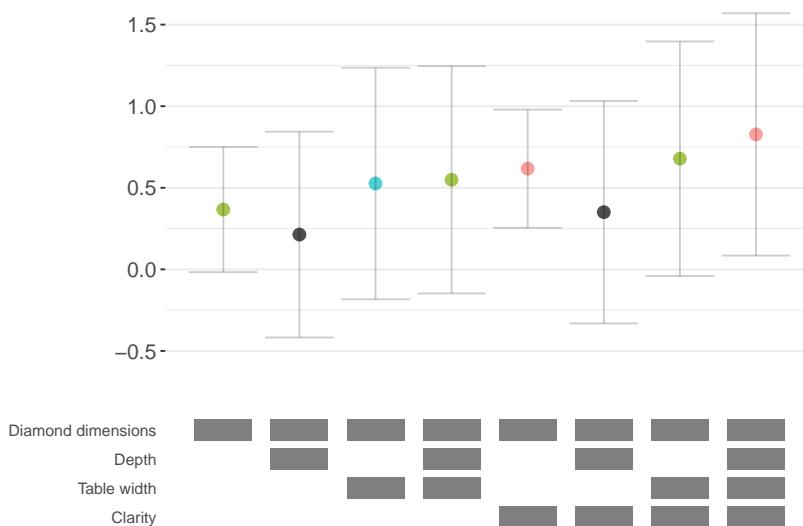
  spec = as.formula(spec)
  if (get_mfx) {
    model = glm(spec, data=data, family='binomial', weights=data$weight) %>%
      margins() %>%
      summary
    row = which(model$factor==rhs)
    coef = model[row, 'AME'] %>% as.numeric()
```

```

se    = model[row, 'SE'] %>% as.numeric()
p     = model[row, 'p'] %>% as.numeric()
} else {
  model = glm(spec, data=data, family='binomial', weights=data$weight) %>%
    broom::tidy()
  row = which(model$term==rhs)
  coef = model[row, 'estimate'] %>% as.numeric()
  se   = model[row, 'std.error'] %>% as.numeric()
  p    = model[row, 'p.value'] %>% as.numeric()
}
z = qnorm(0.995)
return(c(coef, p, coef+z*se, coef-z*se))
}

stability_plot(grid = grid1,
               data = diamonds,
               lhs = lhs_var,
               rhs = rhs_var,
               model = starb_logit_enhanced,
               get_mfx = T,
               perm = perm_controls,
               base = base_controls,
               fe_always = F,
               run_from = 3)

```



To get your customized plot

```
dfs = stability_plot(grid = grid1,
                      data = diamonds,
                      lhs = lhs_var,
                      rhs = rhs_var,
                      model = starb_logit_enhanced,
                      get_mfx = T,
                      perm = perm_controls,
                      base = base_controls,
                      fe_always = F,
                      run_from = 3,
                      run_to = 5)

coef_grid_logit = dfs[[1]]
control_grid_logit = dfs[[2]]

min_space = 0.5

coef_plot = ggplot2::ggplot(coef_grid_logit, aes(x = model, y = coef, shape=p, group=p))
  geom_linerange(aes(ymin = error_low, ymax = error_high), alpha=0.75) +
  geom_point(size=5, aes(col=p, fill=p), alpha=1) +
  viridis::scale_color_viridis(discrete = TRUE, option = "D")+
  scale_shape_manual(values = c(15,17,18, 19)) +
  theme_classic() +
  geom_hline(yintercept=0, linetype='dotted') +
  ggtitle('A custom coefficient stability plot!') +
  labs(subtitle="Error bars represent 99% confidence intervals") +
  theme(axis.text.x = element_blank(),
        axis.title = element_blank(),
        axis.ticks.x = element_blank()) +
  coord_cartesian(xlim=c(1-min_space, max(coef_grid_logit$model)+min_space),
                  ylim=c(-0.1, 1.6)) +
  guides(fill=F, shape=F, col=F)

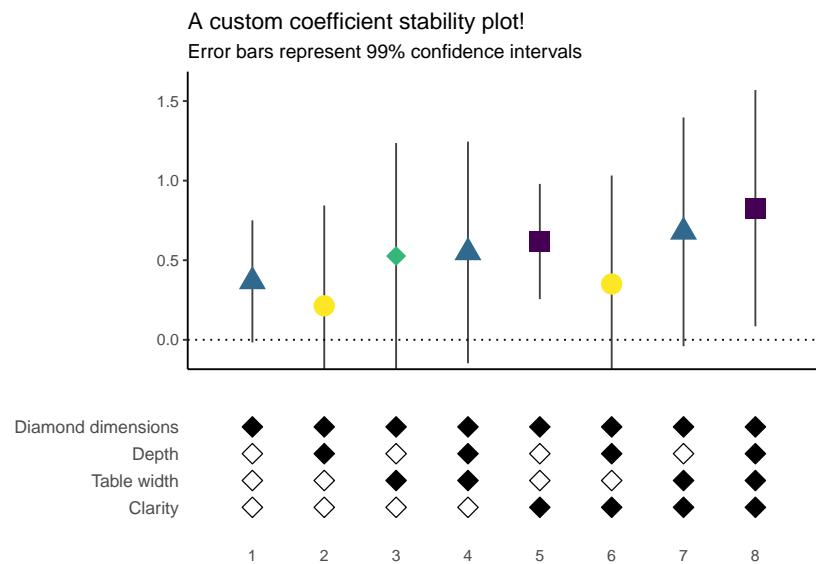
control_plot = ggplot(control_grid_logit) +
  geom_point(aes(x = model, y = y, fill=value), shape=23, size=4) +
  scale_fill_manual(values=c('#FFFFFF', '#000000')) +
  guides(fill=F) +
  scale_y_continuous(breaks = unique(control_grid_logit$y),
                     labels = unique(control_grid_logit$key),
                     limits=c(min(control_grid_logit$y)-1, max(control_grid_logit$y)+1))
  scale_x_continuous(breaks=c(1:max(control_grid_logit$model))) +
  coord_cartesian(xlim=c(1-min_space, max(control_grid_logit$model)+min_space)) +
  theme_classic()
```

```

theme(panel.grid.major.y = element_blank(),
      panel.grid.minor.y = element_blank(),
      axis.title = element_blank(),
      axis.text.y = element_text(size=10),
      axis.ticks = element_blank(),
      axis.line = element_blank())

cowplot::plot_grid(coef_plot, control_plot, rel_heights=c(1,0.5),
                   align='v', ncol=1, axis='b')

```



To get different model specification (e.g., probit vs. logit)

```

starb_probit = function(spec, data, rhs, ...) {
  # Unpack ...
  l = list(...)
  get_mfx = ifelse(is.null(l$get_mfx), F, T) # Set a default to F

  spec = as.formula(spec)
  if (get_mfx) {
    model = glm(
      spec,
      data = data,
      family = binomial(link = 'probit'),
      weights = data$weight
    ) %>%
  }
}

```

```

    margins() %>%
      summary
    row = which(model$factor == rhs)
    coef = model[row, 'AME'] %>% as.numeric()
    se   = model[row, 'SE']  %>% as.numeric()
    p    = model[row, 'p']   %>% as.numeric()
  } else {
    model = glm(
      spec,
      data = data,
      family = binomial(link = 'probit'),
      weights = data$weight
    ) %>%
      broom::tidy()
    row = which(model$term == rhs)
    coef = model[row, 'estimate'] %>% as.numeric()
    se   = model[row, 'std.error'] %>% as.numeric()
    p    = model[row, 'p.value']  %>% as.numeric()
  }
}

z = qnorm(0.995)
return(c(coef, p, coef + z * se, coef - z * se))
}

probit_dfs = stability_plot(
  grid = grid1,
  data = diamonds,
  lhs = lhs_var,
  rhs = rhs_var,
  model = starb_probit,
  get_mfx = T,
  perm = perm_controls,
  base = base_controls,
  fe_always = F,
  run_from = 3,
  run_to = 5
)

# We'll put the probit DFs on the left, so we need to adjust the model numbers according
# so the probit and logit DFs don't plot on top of one another!
coef_grid_probit = probit_dfs[[1]] %>%
  mutate(model = model + max(coef_grid_logit$model))

control_grid_probit = probit_dfs[[2]] %>%
  mutate(model = model + max(control_grid_logit$model))

```

```
coef_grid    = bind_rows(coef_grid_logit, coef_grid_probit)
control_grid = bind_rows(control_grid_logit, control_grid_probit)

panels = stability_plot(
  coef_grid = coef_grid,
  control_grid = control_grid,
  data = diamonds,
  lhs = lhs_var,
  rhs = rhs_var,
  perm = perm_controls,
  base = base_controls,
  fe_always = F,
  run_from = 5,
  run_to = 6
)

coef_plot = panels[[1]] + geom_vline(xintercept = 8.5,
                                      linetype = 'dashed',
                                      alpha = 0.8) +
  annotate(
    geom = 'label',
    x = 4.25,
    y = 1.8,
    label = 'Logit models',
    size = 6,
    fill = '#D3D3D3',
    alpha = 0.7
  ) +
  annotate(
    geom = 'label',
    x = 12.75,
    y = 1.8,
    label = 'Probit models',
    size = 6,
    fill = '#D3D3D3',
    alpha = 0.7
  ) +
  coord_cartesian(ylim = c(-0.5, 1.9))

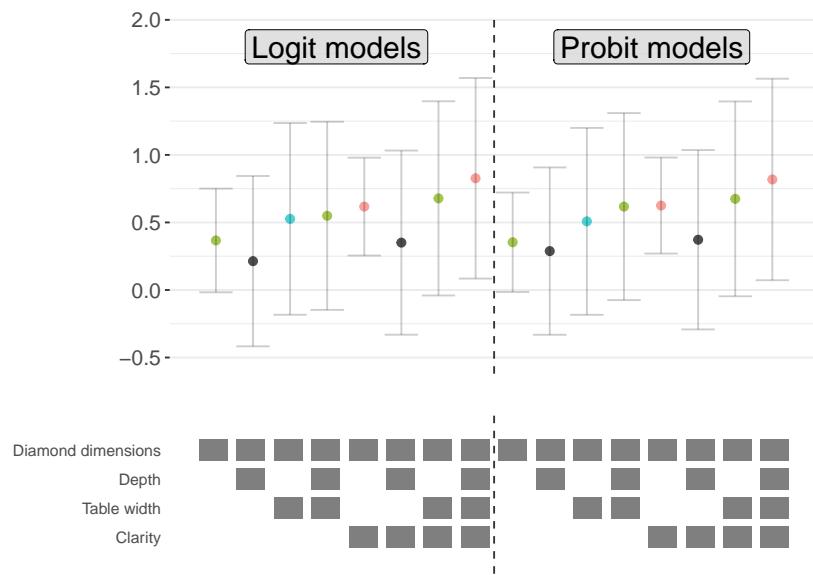
control_plot = panels[[2]] + geom_vline(xintercept = 8.5,
                                         linetype = 'dashed',
                                         alpha = 0.8)

cowplot::plot_grid(
  coef_plot,
```

```

control_plot,
  rel_heights = c(1, 0.5),
  align = 'v',
  ncol = 1,
  axis = 'b'
)

```



33.1.2 rdfanalysis

- Not recommend

Installation

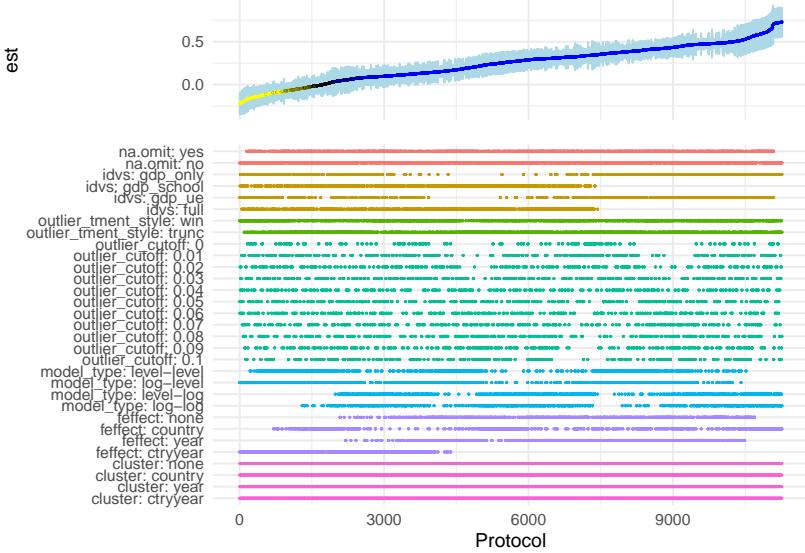
```
devtools::install_github("joachim-gassen/rdfanalysis")
```

Example by the package's author

```

library(rdfanalysis)
load(url("https://joachim-gassen.github.io/data/rdf_est.RData"))
plot_rdf_spec_curve(est, "est", "lb", "ub")

```



Shiny app for readers to explore

```
design <- define_design(steps = c("read_data",
                                    "select_idvs",
                                    "treat_extreme_obs",
                                    "specify_model",
                                    "est_model"),
                        rel_dir = "vignettes/case_study_code")

shiny_rdf_spec_curve(ests, list("est", "lb", "ub"),
                     design, "vignettes/case_study_code",
                     "https://joachim-gassen.github.io/data/wb_new.csv")
```

33.2 Coefficient stability

(Oster, 2017)

- Coefficient stability can be evident against omitted variable bias.
- But coefficient stability alone can be misleading, but combining with R^2 movement, it can become informative.

Packages

- **mplot**: graphical Model stability and Variable Selection

- **robomit**: Robustness checks for omitted variable bias (implementation of

```
library(robomit)

# estimate beta
o_beta(
  y      = "mpg",           # dependent variable
  x      = "wt",            # independent treatment variable
  con   = "hp + qsec",     # related control variables
  delta = 1,               # delta
  R2max = 0.9,             # maximum R-square
  type  = "lm",            # model type
  data   = mtcars          # dataset
)
#> # A tibble: 10 x 2
#>   Name           Value
#>   <chr>          <dbl>
#> 1 beta*          -2.00
#> 2 (beta*-beta controlled)^2    5.56
#> 3 Alternative Solution 1       -7.01
#> 4 (beta[AS1]-beta controlled)^2  7.05
#> 5 Uncontrolled Coefficient    -5.34
#> 6 Controlled Coefficient      -4.36
#> 7 Uncontrolled R-square        0.753
#> 8 Controlled R-square         0.835
#> 9 Max R-square                0.9
#> 10 delta                     1
```

Appendix A

Appendix

A.1 Git

Cheat Sheet

Cheat Sheet in different languages

Learn Git

Interactive Cheat Sheet

Ultimate Guide of Git and GitHub for R user

- Setting up Git: `git config` with `--global` option to configure user name, email, editor, etc.
- Creating a repository: `git init` to initialize a repo. Git stores all of its repo data in the `.git` directory.
- Tracking changes:
 - `git status` shows the status of the repo
 - * File are stored in the project's working directory (which users see)
 - * The staging area (where the next commit is being built)
 - * local repo is where commits are permanently recorded
 - `git add` put files in the staging area
 - `git commit` saves the staged content as a new commit in the local repo.
 - * `git commit -m "your own message"` to give a messages for the purpose of your commit.

- History
 - `git diff` shows differences between commits
 - `git checkout` recovers old version of fields
 - * `git checkout HEAD` to go to the last commit
 - * `git checkout <unique ID of your commit>` to go to such commit
- Ignoring
 - `.gitignore` file tells Git what files to ignore
 - `cat . gitignore *.dat results/` ignore files ending with “dat” and folder “results”.
- Remotes in GitHub
 - A local git repo can be connected to one or more remote repos.
 - Use the HTTPS protocol to connect to remote repos
 - `git push` copies changes from a local repo to a remote repo
 - `git pull` copies changes from a remote repo to a local repo
- Collaborating
 - `git clone` copies remote repo to create a local repo with a remote called `origin` automatically set up
- Branching
 - `git check - b <new-branch-name`
 - `git checkout master` to switch to master branch.
- Conflicts
 - occur when 2 or more people change the same lines of the same file
 - the version control system does not allow to overwrite each other’s changes blindly, but highlights conflicts so that they can be resolved.
- Licensing
 - People who incorporate General Public License (GPL’d) software into their own software must make their software also open under the GPL license; most other open licenses do not require this.
 - The Creative Commons family of licenses allow people to mix and match requirements and restrictions on attribution, creation of derivative works, further sharing and commercialization.

- Citation:
 - Add a CITATION file to a repo to explain how you want others to cite your work.
- Hosting
 - Rules regarding intellectual property and storage of sensitive info apply no matter where code and data are hosted.

A.2 Short-cut

These are shortcuts that you probably remember when working with R. Even though it might take a bit of time to learn and use them as your second nature, but they will save you a lot of time.

Just like learning another language, the more you speak and practice it, the more comfortable you are speaking it.

function	short-cut
navigate folders in console	" " + tab
pull up short-cut cheat sheet	ctrl + shift + k
go to file/function (everything in your project)	ctrl + .
search everything	cmd + shift + f
navigate between tabs	Crtl + shift + .
type function faster	snip + shift + tab
type faster	use tab for fuzzy match
cmd + up	
ctrl + .	

Sometimes you can't stage a folder because it's too large. In such case, use Terminal pane in Rstudio then type `git add -A` to stage all changes then commit and push like usual.

A.3 Function short-cut

apply one function to your data to create a new variable: `mutate(mod=map(data,function))`
 instead of using `i in 1:length(object)`: `for (i in seq_along(object))`
 apply multiple function: `map_dbl`
 apply multiple function to multiple variables:`map2`
`autoplot(data)` plot times series data

`mod_tidy = linear(reg) %>% set_engine('lm') %>% fit(price ~ ., data=data)` fit lm model. It could also fit other models (stan, spark, glmnet, keras)

- Sometimes, data-masking will not be able to recognize whether you're calling from environment or data variables. To bypass this, we use `.data$variable` or `.env$variable`. For example `data %>% mutate(x=.env$variable/.data$variable)`
- Problems with data-masking:
 - Unexpected masking by data-var: Use `.data` and `.env` to disambiguate
 - Data-var cant get through:
 - Tunnel data-var with `{}` + Subset `.data` with `[]`
- Passing Data-variables through arguments

```
library("dplyr")
mean_by <- function(data,by,var){
  data %>%
    group_by({{{by}}}) %>%
    summarise("{var}":=mean({var})) # new name for each var will be created by
}

mean_by <- function(data,by,var){
  data %>%
    group_by({{{by}}}) %>%
    summarise("{var}":=mean({var})) # use single {} to glue the string, but hard
}
```

- Trouble with selection:

```
library("purrr")
name <- c("mass","height")
starwars %>% select(name) # Data-var. Here you are referring to variable named "name"

starwars %>% select(all_of((name))) # use all_of() to disambiguate when

averages <- function(data,vars){ # take character vectors with all_of()
  data %>%
```

```

    select(all_of(vars)) %>%
      map_dbl(mean,na.rm=TRUE)
}

x = c("Sepal.Length","Petal.Length")
iris %>% averages(x)

# Another way
averages <- function(data,vars){ # Tunnel selections with {{}}
  data %>%
    select({{vars}}) %>%
    map_dbl(mean,na.rm=TRUE)
}

x = c("Sepal.Length","Petal.Length")
iris %>% averages(x)

```

A.4 Citation

include a citation by [Farjam_2015]

cite packages used in this session

```

package=ls(sessionInfo()$loadedOnly) for (i in package){print(toBibtex(citation(i)))}

package=ls(sessionInfo()$loadedOnly)
for (i in package){
  print(toBibtex(citation(i)))
}

```

A.5 Install all necessary packages/libaries on your local machine

Get a list of packages you need to install from this book (or your local device)

```

installed <- as.data.frame(installed.packages())

head(installed)
#>                               Package
#> AbnormalReturns  AbnormalReturns
#> addinslist       addinslist

```

```

#> admisc           admisc
#> agridat         agridat
#> akima           akima
#> AlignAssign     AlignAssign
#>
#> AbnormalReturns C:/Users/tn9k4/OneDrive - University of Missouri/Documents/R/win-l...
#> addinslist       C:/Users/tn9k4/OneDrive - University of Missouri/Documents/R/win-l...
#> admisc           C:/Users/tn9k4/OneDrive - University of Missouri/Documents/R/win-l...
#> agridat         C:/Users/tn9k4/OneDrive - University of Missouri/Documents/R/win-l...
#> akima           C:/Users/tn9k4/OneDrive - University of Missouri/Documents/R/win-l...
#> AlignAssign     C:/Users/tn9k4/OneDrive - University of Missouri/Documents/R/win-l...
#>                         Version Priority      Depends
#> AbnormalReturns  0.1.0    <NA> R (>= 2.10)
#> addinslist        0.4.0    <NA> R (>= 3.1.0)
#> admisc            0.24    <NA> R (>= 3.5.0)
#> agridat          1.20    <NA>           <NA>
#> akima            0.6-2.3 <NA> R (>= 2.0.0)
#> AlignAssign      0.5.0    <NA>           <NA>
#>
#> AbnormalReturns
#> addinslist       curl, remotes, DT (>= 0.1), miniUI (>= 0.1), shiny (>=\n0.13.2), sh...
#> admisc
#> agridat
#> akima
#> AlignAssign
#>                         LinkingTo
#> AbnormalReturns   <NA>
#> addinslist        <NA>
#> admisc            <NA>
#> agridat          <NA>
#> akima            <NA>
#> AlignAssign      <NA>
#>
#> AbnormalReturns
#> addinslist
#> admisc
#> agridat          AER, agricolae, betareg, broom, car, coin, corrgram, desplot,\ndply...
#> akima
#> AlignAssign
#>                         Enhances          License License_is_FOSS
#> AbnormalReturns   <NA> MIT + file LICENSE           <NA>
#> addinslist        <NA> MIT + file LICENSE           <NA>
#> admisc            <NA>           GPL (>= 3)           <NA>
#> agridat          <NA>           CC BY-SA 4.0           <NA>
#> akima            <NA> ACM / file LICENSE           <NA>

```

A.5. INSTALL ALL NECESSARY PACKAGES/LIBARIES ON YOUR LOCAL MACHINE919

```
#> AlignAssign      <NA>          GPL-2          <NA>
#> License_restricts_use OS_type MD5sum NeedsCompilation Built
#> AbnormalReturns      <NA>      <NA>      <NA>          no 4.0.4
#> addinslist      <NA>      <NA>      <NA>          no 4.0.5
#> admisc      <NA>      <NA>      <NA>          yes 4.0.5
#> agridat      <NA>      <NA>      <NA>          no 4.0.5
#> akima      yes      <NA>      <NA>          yes 4.0.5
#> AlignAssign      <NA>      <NA>      <NA>          no 4.0.4

write.csv(installed, file.path(getwd(), 'installed.csv'))
```

After having the `installed.csv` file on your new or local machine, you can just install the list of packages

```
# import the list of packages
installed <- read.csv('installed.csv')

# get the list of packages that you have on your device
baseR <- as.data.frame(installed.packages())

# install only those that you don't have
install.packages(setdiff(installed, baseR))
```


Appendix B

Bookdown cheat sheet

```
# to see non-scientific notation a result
format(12e-17, scientific = FALSE)
#> [1] "0.0000000000000012"
```

B.1 Operation

R commands to do derivatives of a defined function Taking derivatives in R involves using the `expression`, `D`, and `eval` functions. You wrap the function you want to take the derivative of in `expression()`, then use `D`, then `eval` as follows.

simple example

```
#define a function
f=expression(sqrt(x))

#take the first derivative
df.dx=D(f, 'x')
df.dx
#> 0.5 * x^-0.5

#take the second derivative
d2f.dx2=D(D(f, 'x'), 'x')
d2f.dx2
#> 0.5 * (-0.5 * x^-1.5)
```

Evaluate

* The first argument passed to `eval` is the expression you want to evaluate *

the second is a list containing the values of all quantities that are not defined elsewhere.

```
#evaluate the function at a given x
eval(f,list(x=3))
#> [1] 1.732051

#evaluate the first derivative at a given x
eval(df.dx,list(x=3))
#> [1] 0.2886751

#evaluate the second derivative at a given x
eval(d2f.dx2,list(x=3))
#> [1] -0.04811252
```

B.2 Math Expression/ Syntax

Full list

Aligning equations

```
\begin{aligned}
a &= b \\
X &\sim \text{Norm}(10, 3) \\
5 &\leq 10
\end{aligned}
```

$$\begin{aligned} a &= b \\ X &\sim \text{Norm}(10, 3) \\ 5 &\leq 10 \end{aligned}$$

Cross-reference equation

```
\begin{equation}
(\#eq:1)
a=b
\end{equation}
```

$$a = b \tag{B.1}$$

to refer in a sentence (B.1) (\@ref(eq:1))

Syntax	Notation
Math	
\pm	\pm
\geq	\geq
\leq	\leq
\neq	\neq
\equiv	\equiv
\circ	\circ
\times	\times
\cdot	\cdot
\trianglelefteq	\trianglelefteq
\triangleeq	\triangleeq
\propto	\propto
\subset	\subset
\subseteq	\subseteq
\leftarrow	\leftarrow
\rightarrow	\rightarrow
\Leftarrow	\Leftarrow
\Rightarrow	\Rightarrow
\approx	\approx
\mathbb{R}	\mathbb{R}
$\sum_{n=1}^{10} n^2$	$\sum_{n=1}^{10} n^2$
$\$ \$ \sum_{n=1}^{10} n^2 \$ \$$	$\sum_{n=1}^{10} n^2$
x^n	x^n
x_n	x_n
\overline{x}	\overline{x}
\hat{x}	\hat{x}
\tilde{x}	\tilde{x}
\checkmark	\checkmark
$\underset{\gamma}{\underbrace{\operatorname{argmax}}}$	
$\frac{a}{b}$	$\frac{a}{b}$
$\frac{a}{\bar{b}}$	$\frac{a}{\bar{b}}$
$\frac{a}{\bar{b}}$	$\frac{a}{\bar{b}}$
$\binom{n}{k}$	$\binom{n}{k}$
$x_1 + x_2 + \dots + x_n$	$x_1 + x_2 + \dots + x_n$
x_1, x_2, \dots, x_n	x_1, x_2, \dots, x_n
$\mathbf{x} = \langle x_1, x_2, \dots, x_n \rangle$	$\mathbf{x} = \langle x_1, x_2, \dots, x_n \rangle$
$x \in A$	$x \in A$
$ A $	$ A $
$x \in A$	$x \in A$
$x \subset B$	$x \subset B$

Syntax	Notation
$\$x \ \backslash subseteq B\$$	$x \subseteq B$
$\$A \ \backslash cup B\$$	$A \cup B$
$\$A \ \backslash cap B\$$	$A \cap B$
$\$X \ \backslash sim \{\\sf Binom\}(n, \ \backslash pi)\$$	$X \sim \text{Binom}(n, \pi)$
$\$\\mathrm{P}\{X \ \\leq x\} = \\{\\tt pbinom\}(x, n, \ \backslash pi)\$$	$P(X \leq x) = \text{pbinom}(x, n, \pi)$
$\$P(A \ \\mid B)\$$	$P(A B)$
$\$\\mathrm{P}\{A \ \\mid B\}\$$	$P(A B)$
$\$\\{1, 2, 3\}\$$	$\{1, 2, 3\}$
$\$\\sin(x)\$$	$\sin(x)$
$\$\\log(x)\$$	$\log(x)$
$\$\\int_{a}^{b} f(x) \\; dx\$$	$\int_a^b f(x) \, dx$
$\$\\left(\\int_{a}^{b} f(x) \\; dx\\right)\$$	$\left(\int_a^b f(x) \, dx\right)$
$\$\\left[\\int_{-\\infty}^{\\infty} f(x) \\; dx\\right]\$$	$\left[\int_{-\infty}^{\infty} f(x) \, dx\right]$
$\$\\left. F(x) \\right _{a}^{b}\$$	$F(x) _a^b$
$\$\\sum_{x=a}^{b} f(x)\$$	$\sum_{x=a}^b f(x)$
$\$\\prod_{x=a}^{b} f(x)\$$	$\prod_{x=a}^b f(x)$
$\$\\lim_{x \\rightarrow \\infty} f(x)\$$	$\lim_{x \rightarrow \infty} f(x)$
$\$\\displaystyle \\lim_{x \\rightarrow \\infty} f(x)\$$	$\lim_{x \rightarrow \infty} f(x)$
Greek Letters	
$\$\\alpha A\$$	αA
$\$\\beta B\$$	βB
$\$\\gamma \\Gamma\$$	$\gamma \Gamma$
$\$\\delta \\Delta\$$	$\delta \Delta$
$\$\\epsilon E\$$	ϵE
$\$\\zeta Z \\sigma\$$	$\zeta Z \sigma$
$\$\\eta H\$$	ηH
$\$\\theta \\vartheta \\Theta\$$	$\theta \vartheta \Theta$
$\$\\iota I\$$	ιI
$\$\\kappa K\$$	κK
$\$\\lambda \\Lambda\$$	$\lambda \Lambda$
$\$\\mu M\$$	μM
$\$\\nu N\$$	νN
$\$\\xi \\Xi\$$	$\xi \Xi$
$\$\\o O\$$	$\o O$
$\$\\pi \\Pi\$$	$\pi \Pi$
$\$\\rho \\varrho P\$$	$\rho \varrho P$
$\$\\sigma \\Sigma\$$	$\sigma \Sigma$
$\$\\tau T\$$	τT
$\$\\upsilon \\Upsilon\$$	$\upsilon \Upsilon$

Syntax	Notation
<code>\phi \varphi \Phi</code>	$\phi\varphi\Phi$
<code>\chi X</code>	χX
<code>\psi \Psi</code>	$\psi\Psi$
<code>\omega \Omega</code>	$\omega\Omega$
<code>\cdot</code>	\cdot
<code>\dots</code>	\dots
<code>\ddots</code>	\ddots
<code>\ldots</code>	\ldots

Limit $P(\lim_{n \rightarrow \infty} \bar{X}_n = \mu) = 1$

$$P(\lim_{n \rightarrow \infty} \bar{X}_n = \mu) = 1$$

Matrices

```
$$\begin{array}{rrr}
1 & 2 & 3 \\
4 & 5 & 6 \\
7 & 8 & 9
\end{array}
$$
```

$$\begin{matrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{matrix}$$

```
$$\mathbf{X} = \left[ \begin{array}{rrr}
1 & 2 & 3 \\
4 & 5 & 6 \\
7 & 8 & 9
\end{array} \right]
$$
```

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

Aligning Equations

Aligning Equations with Comments

```
\begin{aligned}
 3+x &= 4 \quad \& \text{(Solve for } x \text{.)} \\
 x &= 4-3 \quad \& \text{(Subtract 3 from both sides.)} \\
 x &= 1 \quad \& \text{(Yielding the solution.)}
\end{aligned}
```

$$\begin{aligned} 3 + x &= 4 && \text{(Solve for } x \text{.)} \\ x &= 4 - 3 && \text{(Subtract 3 from both sides.)} \\ x &= 1 && \text{(Yielding the solution.)} \end{aligned}$$

B.2.1 Statistics Notation

```
$$
f(y|N,p) = \frac{N!}{y!(N-y)!} \cdot p^y \cdot (1-p)^{N-y} = \binom{N}{y} \cdot p^y \cdot (1-p)^{N-y}
$$
```

$$f(y|N,p) = \frac{N!}{y!(N-y)!} \cdot p^y \cdot (1-p)^{N-y} = \binom{N}{y} \cdot p^y \cdot (1-p)^{N-y}$$

```
\begin{cases}
\frac{1}{b-a} & \text{for } x \in [a,b] \\
0 & \text{otherwise}
\end{cases}
```

$$\begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

B.3 Table

Fruit	Price	Advantages
Bananas	\$1.34	- built-in wrapper - bright color
Oranges	\$2.10	- cures scurvy - **tasty**

Fruit	Price	Advantages
Bananas	\$1.34	<ul style="list-style-type: none">• built-in wrapper• bright color
Oranges	\$2.10	<ul style="list-style-type: none">• cures scurvy• tasty

(\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{y}

$$(\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{y}$$

Bibliography

- (1976). Corrections: Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, 32(4):954.
- Aaronson, D., Barrow, L., and Sander, W. (2007). Teachers and student achievement in the chicago public high schools. *Journal of Labor Economics*, 25(1):95–135.
- Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425.
- Abadie, A., Angrist, J., and Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70(1):91–117.
- Abadie, A., Athey, S., Imbens, G., and Wooldridge, J. (2017). When should you adjust standard errors for clustering?
- Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california’s tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.
- Abadie, A. and L’Hour, J. (2021). A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association*, 116(536):1817–1834.
- Acemoglu, D., Naidu, S., Restrepo, P., and Robinson, J. A. (2014). Democracy does cause growth. *SSRN Electronic Journal*.
- Agrawal, J. and Kamakura, W. A. (1995). The economic worth of celebrity endorsers: An event study analysis. *Journal of Marketing*, 59(3):56.
- Ahrens, H. and Pincus, R. (1981). On two measures of unbalancedness in a one-way model and their relation to efficiency. *Biometrical Journal*, 23(3):227–235.
- Aiken, L. S. and West, S. G. (2005). Interaction effects. In *Encyclopedia of Statistics in Behavioral Science*.

- Akca, S. and Rao, A. (2020). Value of aggregators. *Marketing Science*, 39(5):893–922.
- Altonji, J., Elder, T., and Taber, C. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy*, 113(1):151–184.
- Amemiya, T. (1984). Tobit models: A survey. *Journal of Econometrics*, 24(1–2):3–61.
- Amemiya, T. and MaCurdy, T. E. (1986). Instrumental-variable estimation of an error-components model. *Econometrica*, 54(4):869.
- Anderson, M. L. (2014). Subways, strikes, and slowdowns: The impacts of public transit on traffic congestion. *American Economic Review*, 104(9):2763–2796.
- Anderson, S. T., Kellogg, R., Langer, A., and Sallee, J. M. (2015). The intergenerational transmission of automobile brand preferences. *The Journal of Industrial Economics*, 63(4):763–793.
- Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2):193–212.
- Angrist, J. D. and Krueger, A. B. (1999). *Empirical Strategies in Labor Economics*, pages 1277–1366. Elsevier.
- Arellano, M. et al. (1987). Computing robust standard errors for within-groups estimators. *Oxford bulletin of Economics and Statistics*, 49(4):431–434.
- Arkhangelsky, D., Athey, S., Hirshberg, D., Imbens, G., and Wager, S. (2019). Synthetic difference in differences. Technical report.
- Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2021). Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118.
- Athey, S. and Imbens, G. W. (2022). Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*, 226(1):62–79.
- Auffhammer, M. and Kellogg, R. (2011). Clearing the air? the effects of gasoline content regulation on air quality. *American Economic Review*, 101(6):2687–2722.
- Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424.
- BABCOCK, P. (2010). Real costs of nominal grade inflation? new evidence from student course evaluations. *Economic Inquiry*, 48(4):983–996.

- Bajari, P., Hong, H., Park, M., and Town, R. (2011). Regression discontinuity designs with an endogenous forcing variable and an application to contracting in health care.
- Balestra, P. and Varadharajan-Krishnakumar, J. (1987). Full information estimations of a system of simultaneous equations with error component structure. *Econometric Theory*, 3(2):223–246.
- Baltagi, B. H. (1981). Simultaneous equations with error components. *Journal of Econometrics*, 17(2):189–200.
- Baltagi, B. H. and Li, Q. (1991). A joint test for serial correlation and random individual effects. *Statistics & Probability Letters*, 11(3):277–280.
- Baltagi, B. H. and Li, Q. (1995). Testing AR(1) against MA(1) disturbances in an error component model. *Journal of Econometrics*, 68(1):133–151.
- Barber, B. M. and Lyon, J. D. (1997). Detecting long-run abnormal stock returns: The empirical power and specification of test statistics. *Journal of Financial Economics*, 43(3):341–372.
- Bareinboim, E., Tian, J., and Pearl, J. (2014). Proceedings of the twenty-eighth aaaa conference on artificial intelligence.
- Baron, R. M. and Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173–1182.
- Barreca, A. I., Guldi, M., Lindo, J. M., and Waddell, G. R. (2011). Saving babies? revisiting the effect of very low birth weight classification. *The Quarterly Journal of Economics*, 126(4):2117–2123.
- Bates, D. M. and Watts, D. G. (1980). Relative curvature measures of nonlinearity. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(1):1–16.
- Bates, D. M. and Watts, D. G. (1981). A relative offset orthogonality convergence criterion for nonlinear least squares. *Technometrics*, 23(2):179.
- Bauer, D. J. and Curran, P. J. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research*, 40(3):373–400.
- Beale, E. M. L. and Little, R. J. A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 37(1):129–145.
- Ben-Michael, E., Feller, A., and Rothstein, J. (2020). Varying impacts of letters of recommendation on college admissions: Approximate balancing weights for subgroup effects in observational studies. *arXiv preprint arXiv:2008.04394*.

- Bendel, R. B. and Afifi, A. A. (1977). Comparison of stopping rules in forward “stepwise” regression. *Journal of the American Statistical Association*, 72(357):46–53.
- Benston, G. J. (1985). The validity of profits-structure studies with particular reference to the ftc’s line of business data. *The American Economic Review*, 75(1):37–67.
- Bento, A., Kaffine, D., Roth, K., and Zaragoza-Watkins, M. (2014). The effects of regulation in the presence of multiple unpriced externalities: Evidence from the transportation sector. *American Economic Journal: Economic Policy*, 6(3):1–29.
- Bera, A. K. and Jarque, C. M. (1981). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 7(4):313–318.
- Bera, A. K., Sosa-Escudero, W., and Yoon, M. (2001). Tests for the error component model in the presence of local misspecification. *Journal of Econometrics*, 101(1):1–23.
- Bertanha, M. and Imbens, G. (2014). External validity in fuzzy regression discontinuity designs.
- Blau, D. (1999). The effect of income on child development. *Review of Economics and Statistics*, 81(2):261–276.
- Blomquist, S., Newey, W., Kumar, A., and Liang, C. (2017). On bunching and identification of the taxable income elasticity.
- Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87(1):115–143.
- Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal*, 15(4):651–675.
- Borusyak, K., Hull, P., and Jaravel, X. (2021). Quasi-experimental shift-share research designs. *The Review of Economic Studies*, 89(1):181–213.
- Bosch, N., Dekker, V., and Strohmaier, K. (2020). A data-driven procedure to determine the bunching window: an application to the netherlands. *International Tax and Public Finance*, 27(4):951–979.
- Bound, J., Brown, C., Duncan, G., and Rodgers, W. (1994). Evidence on the validity of cross-sectional and longitudinal labor market data. *Journal of Labor Economics*, 12(3):345–368.
- Bound, J. and Krueger, A. (1989). The extent of measurement error in longitudinal earnings data: Do two wrongs make a right?

- Bowblis, J. and Smith, A. (2019). Occupational licensing of social services and nursing home quality: A regression discontinuity approach. *ILR Review*, 74(1):199–223.
- BRAV, A. and GOMPERS, P. A. (1997). Myth or reality? the long-run under-performance of initial public offerings: Evidence from venture and nonventure capital-backed companies. *The Journal of Finance*, 52(5):1791–1821.
- BREUSCH, T. S. (1978). TESTING FOR AUTOCORRELATION IN DYNAMIC LINEAR MODELS*. *Australian Economic Papers*, 17(31):334–355.
- Breusch, T. S., Mizon, G. E., and Schmidt, P. (1989). Efficient estimation using panel data. *Econometrica*, 57(3):695.
- Breusch, T. S. and Pagan, A. R. (1980). The lagrange multiplier test and its applications to model specification in econometrics. *The Review of Economic Studies*, 47(1):239.
- Brodeur, A., Clark, A. E., Fleche, S., and Powdthavee, N. (2021). Covid-19, lockdowns and well-being: Evidence from google trends. *Journal of Public Economics*, 193:104346.
- Brown, S. J. and Warner, J. B. (1985). Using daily stock returns. *Journal of Financial Economics*, 14(1):3–31.
- Bruin, J. (2011). newtest: command to compute new test @ONLINE.
- Bullock, J. G., Green, D. P., and Ha, S. E. (2010). Yes, but what's the mechanism? (don't expect an easy answer). *Journal of Personality and Social Psychology*, 98(4):550–558.
- Burger, N. E., Kaffine, D. T., and Yu, B. (2014). Did california's hand-held cell phone ban reduce accidents? *Transportation Research Part A: Policy and Practice*, 66:162–172.
- Busse, M., Silva-Risso, J., and Zettelmeyer, F. (2006). \$1,000 cash back: The pass-through of auto manufacturer promotions. *American Economic Review*, 96(4):1253–1270.
- Busse, M. R., Simester, D. I., and Zettelmeyer, F. (2010). “the best price you'll ever get”: The 2005 employee discount pricing promotions in the u.s. automobile industry. *Marketing Science*, 29(2):268–290.
- Butcher, K., McEwan, P., and Weerapana, A. (2014). The effects of an anti-grade inflation policy at wellesley college. *Journal of Economic Perspectives*, 28(3):189–204.
- Böckerman, P., Kanninen, O., and Suoniemi, I. (2018). A kink that makes you sick: The effect of sick pay on absence. *Journal of Applied Econometrics*, 33(4):568–579.

- Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230.
- Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2019a). **nprobust**: Nonparametric kernel-based estimation and robust bias-corrected inference. *Journal of Statistical Software*, 91(8).
- Calonico, S., Cattaneo, M. D., Farrell, M. H., and Titiunik, R. (2019b). Regression discontinuity designs using covariates. *The Review of Economics and Statistics*, 101(3):442–451.
- Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.
- Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3):414–427.
- Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*. Princeton University Press.
- Campbell, J. Y., Lo, A. W., MacKinlay, A. C., and Whitelaw, R. F. (1998). The econometrics of financial markets. *Macroeconomic Dynamics*, 2(4):559–562.
- Canay, I. A., Santos, A., and Shaikh, A. M. (2021). The wild bootstrap with a “small” number of “large” clusters. *The Review of Economics and Statistics*, 103(2):346–363.
- Card, D. and Krueger, A. (1993). Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania.
- Card, D., Lee, D., Pei, Z., and Weber, A. (2012). Nonlinear policy rules and the identification and estimation of causal effects in a generalized regression kink design.
- Card, D., Lee, D., Pei, Z., and Weber, A. (2015). Inference on causal effects in a generalized regression kink design. *Econometrica*, 83(6):2453–2483.
- Carpenter, C. and Dobkin, C. (2009). The effect of alcohol consumption on mortality: Regression discontinuity evidence from the minimum drinking age. *American Economic Journal: Applied Economics*, 1(1):164–182.
- Cattaneo, M., Jansson, M., and Ma, X. (2019). Simple local polynomial density estimators. *Journal of the American Statistical Association*, 115(531):1449–1455.
- Cattaneo, M. D., Titiunik, R., and Vazquez-Bare, G. (2020). *The Regression Discontinuity Design*, pages 835–857. SAGE Publications Ltd.

- Chaney, P. K., Devinney, T. M., and Winer, R. S. (1991). The impact of new product introductions on the market value of firms. *The Journal of Business*, 64(4):573.
- Chaplin, D., Cook, T., Zurovac, J., Coopersmith, J., Finucane, M., Vollmer, L., and Morris, R. (2018). The internal and external validity of the regression discontinuity design: A meta-analysis of 15 within-study comparisons. *Journal of Policy Analysis and Management*, 37(2):403–429.
- Chen, X., John, G., Hays, J. M., Hill, A. V., and Geurs, S. E. (2009). Learning from a service guarantee quasi experiment. *Journal of Marketing Research*, 46(5):584–596.
- Chen, Y. and Whalley, A. (2012). Green infrastructure: The effects of urban rail transit on air quality. *American Economic Journal: Economic Policy*, 4(1):58–97.
- Chen, Z. and Chan, L. (2012). *Causal Discovery for Linear Non-Gaussian Acyclic Models in the Presence of Latent Gaussian Confounders*, pages 17–24. Springer Berlin Heidelberg.
- Chernozhukov, V. and Hansen, C. (2005). An iv model of quantile treatment effects. *Econometrica*, 73(1):245–261.
- Chetty, R., Friedman, J., Olsen, T., and Pistaferri, L. (2011). Adjustment costs, firm responses, and micro vs. macro labor supply elasticities: Evidence from danish tax records. *The Quarterly Journal of Economics*, 126(2):749–804.
- Chetty, R., Friedman, J., and Rockoff, J. (2013). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates.
- Chung, D. J., Steenburgh, T., and Sudhir, K. (2014). Do bonuses enhance sales productivity? a dynamic structural analysis of bonus-based compensation plans. *Marketing Science*, 33(2):165–187.
- Cullen, J., Jacob, B., and Levitt, S. (2005). The impact of school choice on student outcomes: an analysis of the chicago public schools. *Journal of Public Economics*, 89(5-6):729–760.
- Datta, H., Knox, G., and Bronnenberg, B. J. (2018). Changing their tune: How consumers' adoption of online streaming affects music consumption and discovery. *Marketing Science*, 37(1):5–21.
- Davis, L. (2008). The effect of driving restrictions on air quality in mexico city. *Journal of Political Economy*, 116(1):38–81.
- Davis, L. W. and Kahn, M. E. (2010). International trade in used vehicles: The environmental consequences of nafta. *American Economic Journal: Economic Policy*, 2(4):58–82.

- de Chaisemartin, C. and D'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–2996.
- De Paola, M., Scoppa, V., and Falcone, M. (2012). The deterrent effects of the penalty points system for driving offences: a regression discontinuity approach. *Empirical Economics*, 45(2):965–985.
- Desai, H. and Jain, P. C. (1999). Firm performance and focus: long-run stock market performance following spinoffs. *Journal of Financial Economics*, 54(1):75–101.
- Diamond, A. and Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3):932–945.
- Doleac, J. and Hansen, B. (2020). The unintended consequences of “ban the box”: Statistical discrimination and employment outcomes when criminal histories are hidden. *Journal of Labor Economics*, 38(2):321–374.
- Dolley, J. C. (1933). Characteristics and procedure of common stock split-ups. *Harvard business review*, 11(3):316–326.
- Doornik, J. A. and Hansen, H. (2008). An omnibus test for univariate and multivariate normality*. *Oxford Bulletin of Economics and Statistics*, 70:927–939.
- Dutta, A., Knif, J., Kolari, J. W., and Pynnonen, S. (2018). A robust and powerful test of abnormal stock returns in long-horizon event studies. *Journal of Empirical Finance*, 47:1–24.
- Ebbes, P., Papies, D., and van Heerde, H. J. (2011). The sense and non-sense of holdout sample validation in the presence of endogeneity. *Marketing Science*, 30(6):1115–1122.
- Ebbes, P., Wedel, M., B?ckenholt, U., and Steerneman, T. (2005). Solving and testing for regressor-error (in)dependence when no instrumental variables are available: With new evidence for the effect of education on income. *Quantitative Marketing and Economics*, 3(4):365–392.
- Einav, L., Finkelstein, A., and Cullen, M. R. (2010). Estimating welfare in insurance markets using variation in prices*. *Quarterly Journal of Economics*, 125(3):877–921.
- EJD, Agresti, A., and Finlay, B. (1998). Statistical methods for the social sciences. *Journal of the American Statistical Association*, 93(442):844.
- Esarey, J. and Sumner, J. L. (2017). Marginal effects in interaction models: Determining and controlling the false positive rate. *Comparative Political Studies*, 51(9):1144–1176.

- Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383.
- Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.
- Faraway, J. J. (2016). *Extending the Linear Model with R*. Chapman and Hall/CRC.
- Firpo, S. and Possebom, V. (2018). Synthetic control method: Inference, sensitivity analysis and confidence sets. *Journal of Causal Inference*, 6(2).
- Fornell, C., Mithas, S., Morgeson, F. V., and Krishnan, M. (2006). Customer satisfaction and stock prices: High returns, low risk. *Journal of Marketing*, 70(1):3–14.
- Fox, J. (1991). *Maximum-Likelihood Methods, Score Tests, and Constructed Variables*.
- Freedman, D. A. (2008). Randomization does not justify logistic regression. *Statistical Science*, 23(2).
- Fuller, W. A. and Battese, G. E. (1974). Estimation of linear models with crossed-error structure. *Journal of Econometrics*, 2(1):67–78.
- Furnival, G. M. and Wilson, R. W. (1974). Regressions by leaps and bounds. *Technometrics*, 16(4):499–511.
- Gallego, F., Montero, J.-P., and Salas, C. (2013). The effect of transport policies on car use: Evidence from latin american cities. *Journal of Public Economics*, 107:47–62.
- Gerber, A. S., Green, D. P., Kaplan, E. H., and Kern, H. L. (2010). Baseline, placebo, and treatment: Efficient estimation for three-group experiments. *Political Analysis*, 18(3):297–315.
- Geyskens, I., Gielens, K., and Dekimpe, M. G. (2002). The market valuation of internet channel additions. *Journal of Marketing*, 66(2):102–119.
- Gibbons, C. E., Suárez Serrato, J. C., and Urbancic, M. B. (2018). Broken or fixed effects? *Journal of Econometric Methods*, 8(1).
- Glasser, M. (1964). Linear regression analysis with missing observations among the independent variables. *Journal of the American Statistical Association*, 59(307):834–844.
- Godfrey, L. G. (1978). Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica*, 46(6):1293.

- Goldfarb, A., Tucker, C., and Wang, Y. (2022). Conducting research in marketing with quasi-experiments. *Journal of Marketing*, 86(3):1–20.
- Goldfarb, A. and Tucker, C. E. (2011). Privacy regulation and online advertising. *Management Science*, 57(1):57–71.
- Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.
- Gordon, B. R., Zettelmeyer, F., Bhargava, N., and Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science*, 38(2):193–225.
- Gourieroux, C., Holly, A., and Monfort, A. (1982). Likelihood ratio test, wald test, and kuhn-tucker test in linear models with inequality constraints on the regression parameters. *Econometrica*, 50(1):63.
- Gourieroux, C. and Monfort, A. (1981). On the problem of missing data in linear models. *The Review of Economic Studies*, 48(4):579.
- Greene, W. H. (1990). *Econometric Analysis*.
- Greevy, R. (2004). Optimal multivariate matching before randomization. *Biostatistics*, 5(2):263–275.
- Guo, T., Sriram, S., and Manchanda, P. (2020). “let the sunshine in”: The impact of industry payment disclosure on physician prescription behavior. *Marketing Science*, 39(3):516–539.
- Habel, J., Alavi, S., and Linsenmayer, K. (2021). Variable compensation and salesperson health. *Journal of Marketing*, 85(3):130–149.
- Haitovsky, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(1):67–82.
- Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488.
- Harper, S. and Bruckner, T. A. (2017). Did the great recession increase suicides in the usa? evidence from an interrupted time-series analysis. *Annals of Epidemiology*, 27(7):409–414.e6.
- Hartmann, W., Nair, H. S., and Narayanan, S. (2011). Identifying causal marketing mix effects using a regression discontinuity design. *Marketing Science*, 30(6):1079–1097.
- Hausman, C. and Rapson, D. S. (2018). Regression discontinuity in time: Considerations for empirical applications. *Annual Review of Resource Economics*, 10(1):533–552.

- Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46(6):1251.
- He, S., Hollenbeck, B., and Proserpio, D. (2022). The market for fake reviews. *Marketing Science*.
- Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5(4):475–492.
- Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4):605–654.
- Heckman, J. J., Lalonde, R. J., and Smith, J. A. (1999). *The Economics and Econometrics of Active Labor Market Programs*, pages 1865–2097. Elsevier.
- Heller, R., Rosenbaum, P. R., and Small, D. S. (2009). Split samples and design sensitivity in observational studies. *Journal of the American Statistical Association*, 104(487):1090–1101.
- Henderson, C. R. (1950). Estimation of genetic parameters. *Annals of Mathematical Statistics*, 21:309–310.
- Honda, Y. (1985). Testing the error components model with non-normal disturbances. *The Review of Economic Studies*, 52(4):681.
- Horsky, D. and Swyngedouw, P. (1987). Does it pay to change your company's name? a stock market perspective. *Marketing Science*, 6(4):320–335.
- Hosmer, D. W. and Lemeshow, S. (1992). Confidence interval estimation of interaction. *Epidemiology*, 3(5):452–456.
- Hoyer, P., Shimizu, S., Kerminen, A., and Palviainen, M. (2008). Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378.
- HURVICH, C. M. and TSAI, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307.
- Iacus, S. M., King, G., and Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1):1–24.
- Imai, K., Keele, L., and Tingley, D. (2010a). A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334.
- Imai, K., Keele, L., and Yamamoto, T. (2010b). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1).
- Imai, K. and Kim, I. S. (2020). On the use of two-way fixed effects regression models for causal inference with panel data. *Political Analysis*, 29(3):405–415.

- Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(2):481–502.
- Imbens, G. and Kalyanaraman, K. (2009). Optimal bandwidth choice for the regression discontinuity estimator. Technical report.
- Imbens, G. and Kalyanaraman, K. (2010). Optimal bandwidth choice for the regression discontinuity estimator. Technical report.
- Imbens, G. and Lemieux, T. (2007). Regression discontinuity designs: A guide to practice.
- Israeli, A. (2018). Online map enforcement: Evidence from a quasi-experiment. *Marketing Science*, 37(5):710–732.
- Jaffe, J. F. (1974). Special information and insider trading. *The Journal of Business*, 47(3):410.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *Statistical Learning*, pages 15–57. Springer New York.
- Janakiraman, R., Lim, J. H., and Rishika, R. (2018). The effect of a data breach announcement on customer behavior: Evidence from a multichannel retailer. *Journal of Marketing*, 82(2):85–105.
- Jin, H. and Rubin, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association*, 103(481):101–111.
- Jin, H. and Rubin, D. B. (2009). Public schools versus private schools: Causal inference with partial compliance. *Journal of Educational and Behavioral Statistics*, 34(1):24–45.
- Johansson, P. and Scholtzberg, M. (2022). Rerandomization: A complement or substitute for stratification in randomized experiments? *Journal of Statistical Planning and Inference*, 218:43–58.
- Johnson, P. O. and Neyman, J. (1936). Tests of certain linear hypotheses and their application to some educational problems. *Statistical research memoirs*.
- Johnson, S. A. and Houston, M. B. (2000). A reexamination of the motives and gains in joint ventures. *The Journal of Financial and Quantitative Analysis*, 35(1):67.
- Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, 91(433):222–230.
- Kapelner, A. and Krieger, A. (2014). Matching on-the-fly: Sequential allocation with higher power and efficiency. *Biometrics*, 70(2):378–388.

- Kim, J.-S. and Frees, E. W. (2007). Multilevel modeling with correlated effects. *Psychometrika*, 72(4):505–533.
- Kim, S., Lee, C., and Gupta, S. (2020). Bayesian synthetic control methods. *Journal of Marketing Research*, 57(5):831–852.
- King, G., Honaker, J., Joseph, A., and Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, 95(1):49–69.
- King, G., Lucas, C., and Nielsen, R. A. (2016). The balance-sample size frontier in matching methods for causal inference. *American Journal of Political Science*, 61(2):473–489.
- King, G. and Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, 27(4):435–454.
- King, M. L. and Wu, P. X. (1997). Locally optimal one-sided tests for multiparameter hypotheses. *Econometric Reviews*, 16(2):131–156.
- Kleven, H. (2016). Bunching. *Annual Review of Economics*, 8(1):435–464.
- Knol, M. J. and VanderWeele, T. J. (2012). Recommendations for presenting analyses of effect modification and interaction. *International Journal of Epidemiology*, 41(2):514–520.
- Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963.
- Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, 95(2):391–413.
- Lane, V. and Jacobson, R. (1995). Stock market reactions to brand extension announcements: The effects of brand attitude and familiarity. *Journal of Marketing*, 59(1):63.
- Laurent, R. T. S. and Cook, R. D. (1992). Leverage and superleverage in nonlinear regression. *Journal of the American Statistical Association*, 87(420):985.
- Lee, D. and Lemieux, T. (2009). Regression discontinuity designs in economics.
- Lee, D. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355.
- Lee, D. S., McCrary, J., Moreira, M. J., and Porter, J. R. (2021). Valid t-ratio inference for iv.
- Lewbel, A. (1997). Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and r&d. *Econometrica*, 65(5):1201.

- Lewbel, A. (2012). Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *Journal of Business & Economic Statistics*, 30(1):67–80.
- Lewis, R. A. and Rao, J. M. (2015). The unfavorable economics of measuring the returns to advertising *. *The Quarterly Journal of Economics*, 130(4):1941–1973.
- Li, Y. P., Propert, K. J., and Rosenbaum, P. R. (2001). Balanced risk set matching. *Journal of the American Statistical Association*, 96(455):870–882.
- LIANG, K. and ZEGER, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.
- Liaukonyte, J., Teixeira, T., and Wilbur, K. C. (2015). Television advertising and online shopping. *Marketing Science*, 34(3):311–330.
- Lim, J. H., Rishika, R., Janakiraman, R., and Kannan, P. (2020). Competitive effects of front-of-package nutrition labeling adoption on nutritional quality: Evidence from facts up front-style labels. *Journal of Marketing*, 84(6):3–21.
- Little, R. J. A. (1992). Regression with missing x's: A review. *Journal of the American Statistical Association*, 87(420):1227.
- Little, R. J. A. and Smith, P. J. (1987). Editing and imputation for quantitative survey data. *Journal of the American Statistical Association*, 82(397):58–68.
- Long, Q., Little, R. J. A., and Lin, X. (2010). Estimating causal effects in trials involving multitreatment arms subject to non-compliance: a bayesian framework. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(3):513–531.
- Longnecker, M. P., Chen, M.-J., Probst-Hensch, N. M., Harper, J. M., Lee, E. R., Frankl, H. D., and Haile, R. W. (1996). Alcohol and smoking in relation to the prevalence of adenomatous colorectal polyps detected at sigmoidoscopy. *Epidemiology*, 7(3):275–280.
- Looney, S. W. and Gullledge, T. R. (1985). Use of the correlation coefficient with normal probability plots. *The American Statistician*, 39(1):75.
- LOUGHHRAN, T. and RITTER, J. R. (1995). The new issues puzzle. *The Journal of Finance*, 50(1):23–51.
- Ludwig, J. and Miller, D. L. (2007). Does head start improve children's life chances? evidence from a regression discontinuity design. *The Quarterly Journal of Economics*, 122(1):159–208.
- Lyon, J. D., Barber, B. M., and Tsai, C.-L. (1999). Improved methods for tests of long-run abnormal stock returns. *The Journal of Finance*, 54(1):165–201.

- Mackenzie, D. and Pearl, J. (2018). *The Book of Why: The New Science of Cause and Effect*. ISBN 978-1541698963.
- MacKinlay, A. C. (1997). Event studies in economics and finance. *Journal of economic literature*, 35(1):13–39.
- Magel, R. C. and Hertsgaard, D. (1987). A collinearity diagnostic for nonlinear regression. *Communications in Statistics - Simulation and Computation*, 16(1):85–97.
- Manchanda, P., Packard, G., and Pattabhiramaiah, A. (2015). Social dollars: The economic impact of customer participation in a firm-sponsored online customer community. *Marketing Science*, 34(3):367–387.
- MARDIA, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530.
- Marsaglia, G. and Marsaglia, J. (2004). Evaluating the anderson-darling distribution. *Journal of Statistical Software*, 9(2).
- McCrory, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714.
- McCullagh, P. and Nelder, J. (2019). An outline of generalized linear models. pages 21–47.
- McWilliams, A. and Siegel, D. (1997). Event studies in management research: Theoretical and empirical issues. *Academy of Management Journal*, 40(3):626–657.
- Mei, B. and Sun, C. (2008). Event analysis of the impact of mergers and acquisitions on the financial performance of the u.s. forest products industry. *Forest Policy and Economics*, 10(5):286–294.
- Moorman, C. and Lehmann, D. R. (2004). Assessing marketing strategy. *Marketing Science Institute, Cambridge MA*.
- Morgan, K. L. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2).
- Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica*, 55(4):765.
- Narayanan, S. and Kalyanam, K. (2015). Position effects in search advertising and their moderators: A regression discontinuity approach. *Marketing Science*, 34(3):388–407.
- Narayanan, S. and Nair, H. S. (2013). Estimating causal installed-base effects: A bias-correction approach. *Journal of Marketing Research*, 50(1):70–94.

- Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370.
- Nerlove, M. (1971). A note on error components models. *Econometrica*, 39(2):383.
- Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica: Journal of the econometric society*, pages 1417–1426.
- Nielsen, H., Sørensen, T., and Taber, C. (2010). Estimating the effect of student aid on college enrollment: Evidence from a government grant policy reform. *American Economic Journal: Economic Policy*, 2(2):185–215.
- Oster, E. (2017). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business and Economic Statistics*, 37(2):187–204.
- OWEN, J. and RABINOVITCH, R. (1983). On the class of elliptical distributions and their applications to the theory of portfolio choice. *The Journal of Finance*, 38(3):745–752.
- Park, S. and Gupta, S. (2012). Handling endogenous regressors by joint estimation using copulas. *Marketing Science*, 31(4):567–586.
- Pattabhiramaiah, A., Sriram, S., and Manchanda, P. (2018). Paywalls: Monetizing online content. *Journal of Marketing*, 83(2):19–36.
- Pearl, J. (2013). *Causal Diagrams and the Identification of Causal Effects*, pages 65–106. Cambridge University Press.
- Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60.
- Penfold, R. B. and Zhang, F. (2013). Use of interrupted time series analysis in evaluating health care quality improvements. *Academic Pediatrics*, 13(6):S38–S44.
- Peukert, C., Bechtold, S., Batikas, M., and Kretschmer, T. (2022). Regulatory spillovers and data governance: Evidence from the gdpr. *Marketing Science*.
- Pocock, S. J. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, 31(1):103.
- Preacher, K. J. and Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, 36(4):717–731.
- Qu, Y. and Lipkovich, I. (2009). Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Statistics in Medicine*, 28(9):1402–1414.

- Ramani, N. and Srinivasan, R. (2019). Effects of liberalization on incumbent firms' marketing-mix responses and performance: Evidence from a quasi-experiment. *Journal of Marketing*, 83(5):97–114.
- Reiss, P. C. (2011). Structural workshop paper—descriptive, structural, and experimental empirical methods in marketing research. *Marketing Science*, 30(6):950–964.
- Robins, J. M., Hernán, M. Á., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.
- Rosenbaum, P. R. (2002). Attributing effects to treatment in matched observational studies. *Journal of the American Statistical Association*, 97(457):183–192.
- Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33.
- Rossi, P. E. (2014). Even the rich can make themselves poor: A critical examination of iv methods in marketing applications. *Marketing Science*, 33(5):655–672.
- Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, 29(1):159.
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489.
- Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2:169–188.
- Rubin, D. B. and Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95(450):573–585.
- Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6):110.
- Schabenberger, O. and Pierce, F. J. (2001). *Contemporary Statistical Models for the Plant and Soil Sciences*. CRC Press.
- SCHEVE, K. and STASAVAGE, D. (2012). Democracy, war, and wealth: Lessons from two centuries of inheritance taxation. *American Political Science Review*, 106(1):81–102.
- Schisterman, E. F., Moysich, K. B., England, L. J., and Rao, M. (2003). Estimation of the correlation coefficient using the bayesian approach and its applications for epidemiologic research. *BMC medical research methodology*, 3(1):1–4.

- Shapiro, S. S. and Francia, R. S. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 67(337):215–216.
- Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591.
- Shimizu, S., Hoyer, P., and Hyvärinen, A. (2009). Estimation of linear non-gaussian acyclic models for latent factors. *Neurocomputing*, 72(7-9):2024–2027.
- Shin, J., Sudhir, K., and Yoon, D.-H. (2012). When to “fire” customers: Customer cost-based pricing. *Management Science*, 58(5):932–947.
- Silverman, B. (1969). *Density Estimation for Statistics and Data Analysis*.
- Simonsohn, U., Simmons, J. P., and Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11):1208–1214.
- Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, 13:290.
- Somaini, P. and Wolak, F. A. (2016). An algorithm to estimate the two-way fixed effects model. *Journal of Econometric Methods*, 5(1).
- Somaini, P. and Wolak, F. A. (2021). Twfem: Stata module to efficiently estimate a two-way fixed effects model based on somaini and wolak (2015).
- Sood, A. and Tellis, G. J. (2009). Do innovations really pay off? total stock market returns to innovation. *Marketing Science*, 28(3):442–456.
- Spiller, S. A., Fitzsimons, G. J., Lynch, J. G., and McClelland, G. H. (2013). Spotlights, floodlights, and the magic number zero: Simple effects tests in moderated regression. *Journal of Marketing Research*, 50(2):277–288.
- Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347):730–737.
- Stevenson, B. and Wolfers, J. (2006). Bargaining in the shadow of the law: Divorce laws and family distress. *The Quarterly Journal of Economics*, 121(1):267–288.
- Stock, J. H., Wright, J. H., and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4):518–529.
- Stock, J. H. and Yogo, M. (2005). Testing for weak instruments in linear IV regression. pages 80–108.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1).

- Sun, C. and Liao, X. (2011). Effects of litigation under the endangered species act on forest firm values. *Journal of Forest Economics*, 17(4):388–398.
- Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.
- Swamy, P. A. V. B. and Arora, S. S. (1972). The exact finite sample properties of the estimators of coefficients in the error components regression models. *Econometrica*, 40(2):261.
- Thistletonwaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6):309–317.
- Thoemmes, F., Liao, W., and Jin, Z. (2016). The analysis of the regression-discontinuity design in r. *Journal of Educational and Behavioral Statistics*, 42(3):341–360.
- Tirunillai, S. and Tellis, G. J. (2017). Does offline tv advertising affect online chatter? quasi-experimental analysis using synthetic control. *Marketing Science*, 36(6):862–878.
- Tukey, J. W. (1993). Tightening the clinical trial. *Controlled Clinical Trials*, 14(4):266–285.
- Vach, W. (1994). *Logistic Regression with Missing Values in the Covariates*. Springer New York.
- VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology*, 34(3):211–219.
- von Hippel, P. T. (2009). 8. how to impute interactions, squares, and other transformed variables. *Sociological Methodology*, 39(1):265–291.
- Wallace, T. D. and Hussain, A. (1969). The use of error components models in combining cross section with time series data. *Econometrica*, 37(1):55.
- Wang, Y., Lewis, M., and Schweidel, D. A. (2018). A border strategy analysis of ad source and message tone in senatorial campaigns. *Marketing Science*, 37(3):333–355.
- Wang, Y., Wu, C., and Zhu, T. (2019). Mobile hailing technology and taxi driving behaviors. *Marketing Science*, 38(5):734–755.
- Webb, A. (2002). Statistical pattern recognition.
- Wiles, M. A., Jain, S. P., Mishra, S., and Lindsey, C. (2010). Stock market response to regulatory reports of deceptive advertising: The moderating effect of omission bias and firm reputation. *Marketing Science*, 29(5):828–845.

- Wolfers, J. (2003). Did unilateral divorce laws raise divorce rates? a reconciliation and new results.
- Yule, G. (1899). An investigation into the causes of changes in pauperism in england, chiefly during the last two intercensal decades (part i.). *Journal of the Royal Statistical Society*, 62(2):249.
- Zou, G. Y. (2008). On the estimation of additive interaction by use of the four-by-two table and beyond. *American Journal of Epidemiology*, 168(2):212–224.