Making R accessible for sceptics

Stella Bollmann

June 24, 2020

Questions

1. Why would applied researchers use SPSS?
2. Why should applied researchers use R?
3. How can one make R accessible to the sceptics?

About myself

Me as a social scientist

- Diplom (MSc.) in Psychology (2011)
- PhD (Dr. phil.) in Psychological Methods (2015)

Me as a statistician

- Master in Statistics (2016)
- Postdoc and consultant at Psychological Institute of UZH

About myself

Me as a social scientist

- Diplom (MSc.) in Psychology (2011)
- PhD (Dr. phil.) in Psychological Methods (2015)

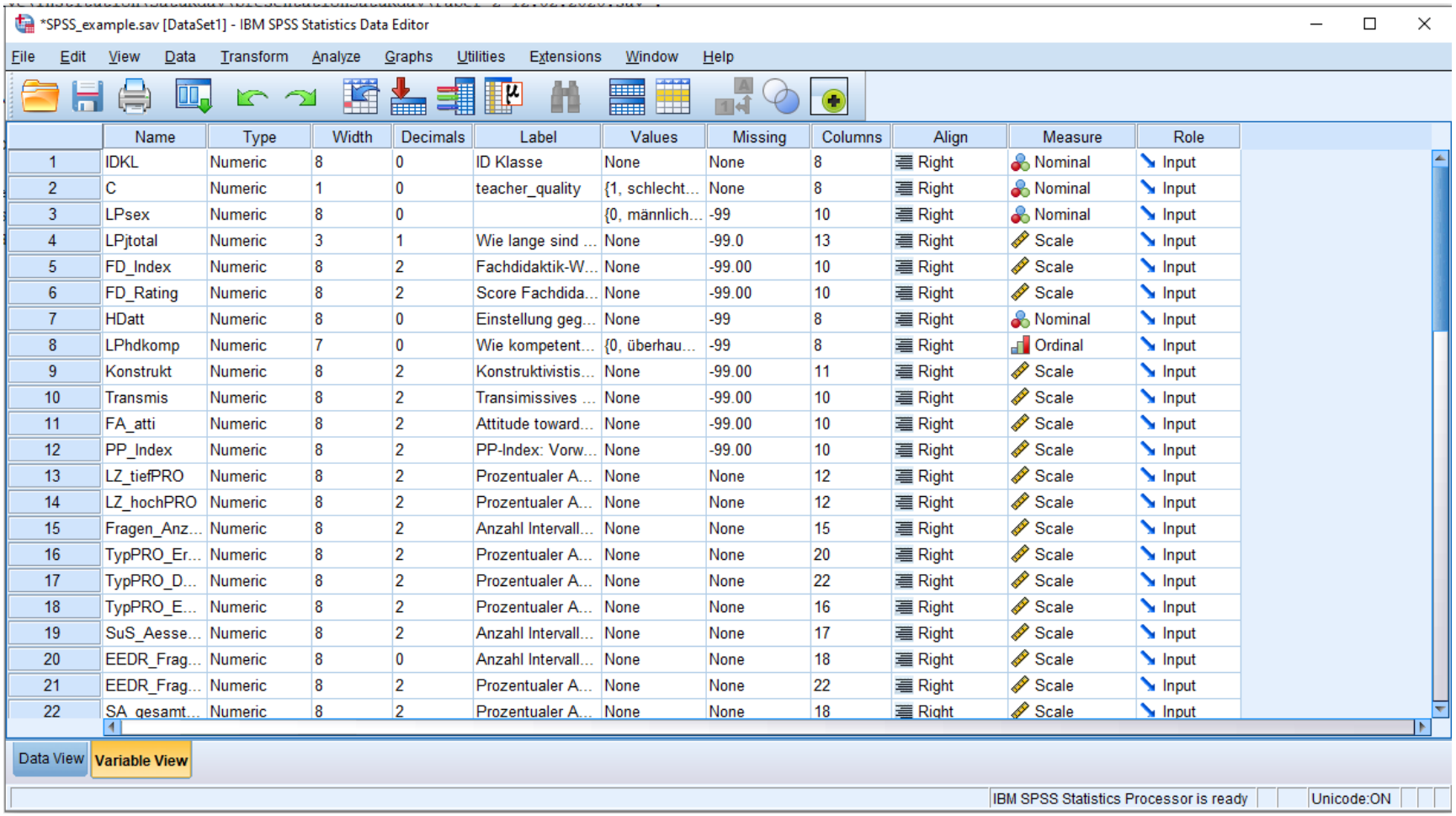
Me as a statistician

- Master in Statistics (2016)
- Postdoc and consultant at Psychological Institute of UZH

Me as an ambassador

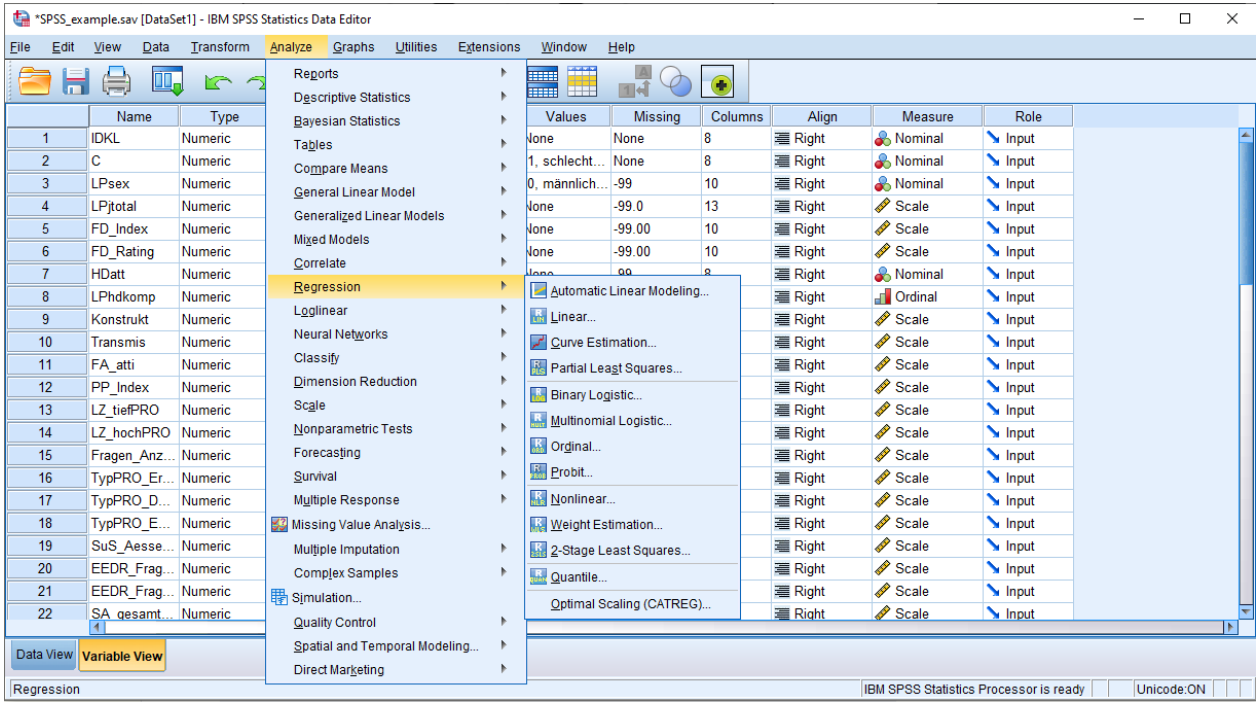
- Statistical consultant at PH Lucerne
- Postdoc researcher in health care science at University Lucerne
- President of the Swiss Statistical Society

SPSS

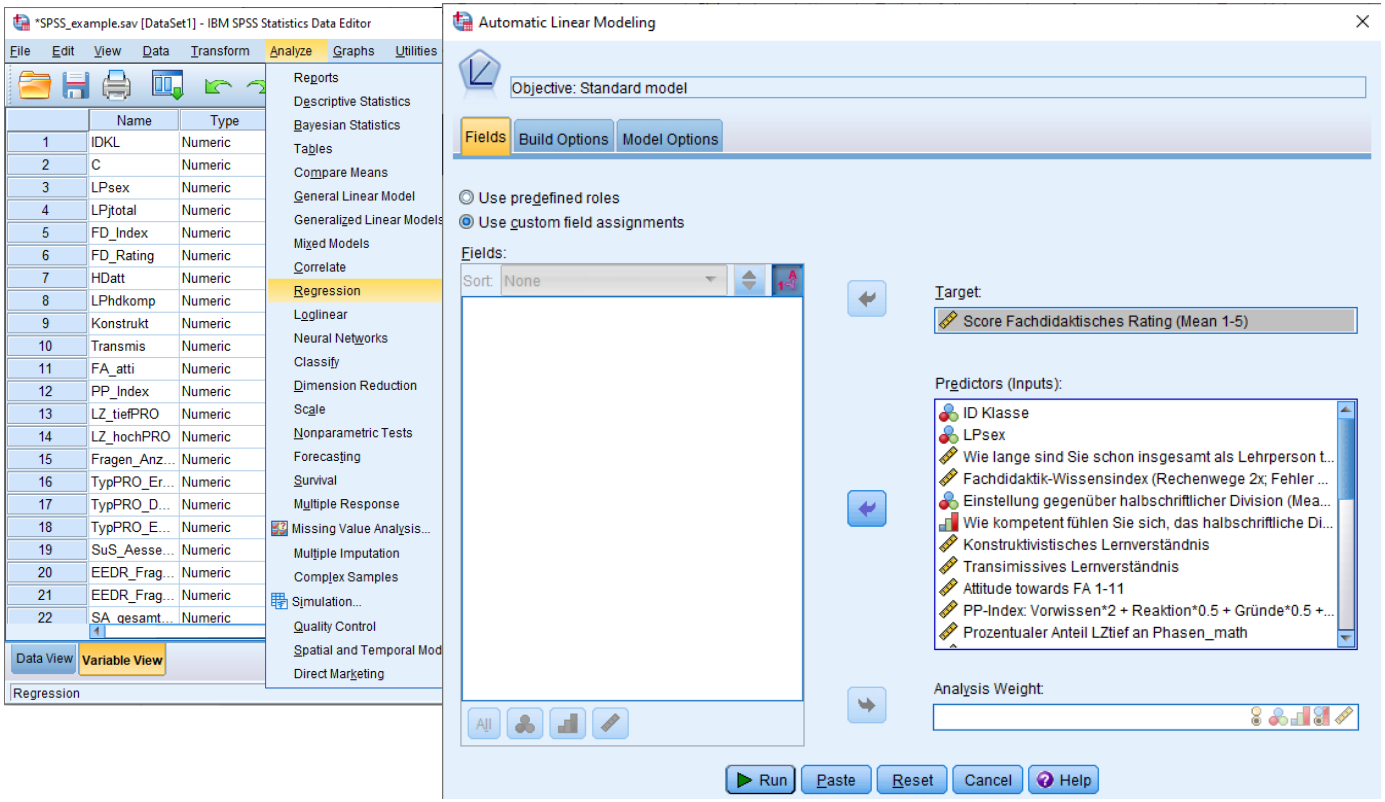


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	IDKL	Numeric	8	0	ID Klasse	None	None	8	Right	Nominal	Input
2	C	Numeric	1	0	teacher_quality	{1, schlecht...	None	8	Right	Nominal	Input
3	LPsex	Numeric	8	0		{0, männlich...	-99	10	Right	Nominal	Input
4	LPjtotal	Numeric	3	1	Wie lange sind ...	None	-99.0	13	Right	Scale	Input
5	FD_Index	Numeric	8	2	Fachdidaktik-W...	None	-99.00	10	Right	Scale	Input
6	FD_Rating	Numeric	8	2	Score Fachdida...	None	-99.00	10	Right	Scale	Input
7	HDatt	Numeric	8	0	Einstellung geg...	None	-99	8	Right	Nominal	Input
8	LPhdkomp	Numeric	7	0	Wie kompetent...	{0, überhau...	-99	8	Right	Ordinal	Input
9	Konstrukt	Numeric	8	2	Konstruktivistis...	None	-99.00	11	Right	Scale	Input
10	Transmis	Numeric	8	2	Transimissives ...	None	-99.00	10	Right	Scale	Input
11	FA_atti	Numeric	8	2	Attitude toward...	None	-99.00	10	Right	Scale	Input
12	PP_Index	Numeric	8	2	PP-Index: Vorw...	None	-99.00	10	Right	Scale	Input
13	LZ_tiefPRO	Numeric	8	2	Prozentualer A...	None	None	12	Right	Scale	Input
14	LZ_hochPRO	Numeric	8	2	Prozentualer A...	None	None	12	Right	Scale	Input
15	Fragen_Anz...	Numeric	8	2	Anzahl Intervall...	None	None	15	Right	Scale	Input
16	TypPRO_Er...	Numeric	8	2	Prozentualer A...	None	None	20	Right	Scale	Input
17	TypPRO_D...	Numeric	8	2	Prozentualer A...	None	None	22	Right	Scale	Input
18	TypPRO_E...	Numeric	8	2	Prozentualer A...	None	None	16	Right	Scale	Input
19	SuS_Aesse...	Numeric	8	2	Anzahl Intervall...	None	None	17	Right	Scale	Input
20	EEDR_Frag...	Numeric	8	0	Anzahl Intervall...	None	None	18	Right	Scale	Input
21	EEDR_Frag...	Numeric	8	2	Prozentualer A...	None	None	22	Right	Scale	Input
22	SA_gesamt...	Numeric	8	2	Prozentualer A...	None	None	18	Right	Scale	Input

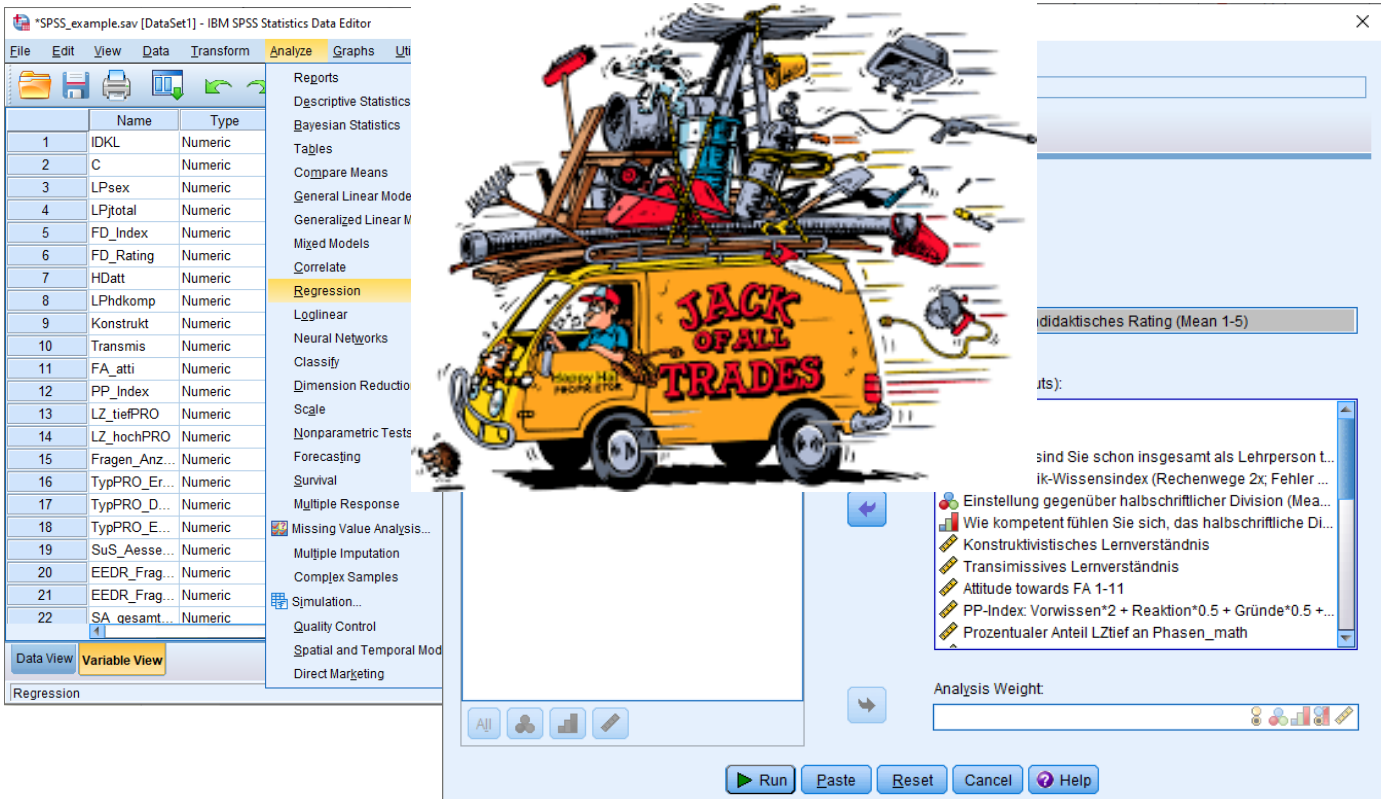
SPSS


*SPSS_example.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help
Reports
Descriptive Statistics
Bayesian Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Simulation...
Quality Control
Spatial and Temporal Modeling...
Direct Marketing
Automatic Linear Modeling...
Linear...
Curve Estimation...
Partial Least Squares...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...
Quantile...
Optimal Scaling (CATREG)...
Name Type Values Missing Columns Align Measure Role
1 IDKL Numeric None None 8 Right Nominal Input
2 C Numeric 1, schlecht... None 8 Right Nominal Input
3 LPsex Numeric 0, männlich... -99 10 Right Nominal Input
4 LPjtotal Numeric None -99.0 13 Right Scale Input
5 FD_Index Numeric None -99.00 10 Right Scale Input
6 FD_Rating Numeric None -99.00 10 Right Scale Input
7 HDatt Numeric Right Nominal Input
8 LPhdkomp Numeric Right Ordinal Input
9 Konstrukt Numeric Right Scale Input
10 Transmis Numeric Right Scale Input
11 FA_atti Numeric Right Scale Input
12 PP_Index Numeric Right Scale Input
13 LZ_tiefPRO Numeric Right Scale Input
14 LZ_hochPRO Numeric Right Scale Input
15 Fragen_Anz... Numeric Right Scale Input
16 TypPRO_Er... Numeric Right Scale Input
17 TypPRO_D... Numeric Right Scale Input
18 TypPRO_E... Numeric Right Scale Input
19 SuS_Aesse... Numeric Right Scale Input
20 EEDR_Frag... Numeric Right Scale Input
21 EEDR_Frag... Numeric Right Scale Input
22 SA_gesamt... Numeric Right Scale Input
Data View
Variable View
Regression
IBM SPSS Statistics Processor is ready
Unicode:ON

SPSS


*SPSS_example.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities
Name Type
1 IDKL Numeric
2 C Numeric
3 LPsex Numeric
4 LPjtotal Numeric
5 FD_Index Numeric
6 FD_Rating Numeric
7 HDatt Numeric
8 LPhdkomp Numeric
9 Konstrukt Numeric
10 Transmis Numeric
11 FA_atti Numeric
12 PP_Index Numeric
13 LZ_tiefPRO Numeric
14 LZ_hochPRO Numeric
15 Fragen_Anz... Numeric
16 TypPRO_Er... Numeric
17 TypPRO_D... Numeric
18 TypPRO_E... Numeric
19 SuS_Aesse... Numeric
20 EEDR_Frag... Numeric
21 EEDR_Frag... Numeric
22 SA_gesamt... Numeric
Data View Variable View
Regression
Reports
Descriptive Statistics
Bayesian Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Simulation...
Quality Control
Spatial and Temporal Mod
Direct Marketing
Automatic Linear Modeling
Objective: Standard model
Fields Build Options Model Options
Use predefined roles
Use custom field assignments
Fields:
Sort: None
All
Target:
Score Fachdidaktisches Rating (Mean 1-5)
Predictors (Inputs):
ID Klasse
LPsex
Wie lange sind Sie schon insgesamt als Lehrperson t...
Fachdidaktik-Wissensindex (Rechenwege 2x; Fehler ...
Einstellung gegenüber halbschriftlicher Division (Mea...
Wie kompetent fühlen Sie sich, das halbschriftliche Di...
Konstruktivistisches Lernverständnis
Transimissives Lernverständnis
Attitude towards FA 1-11
PP-Index: Vorwissen*2 + Reaktion*0.5 + Gründe*0.5 +...
Prozentualer Anteil LZtief an Phasen_math
Analysis Weight:
Run Paste Reset Cancel Help

SPSS


*SPSS_example.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Uti
Name Type
1 IDKL Numeric
2 C Numeric
3 LPsex Numeric
4 LPjtotal Numeric
5 FD_Index Numeric
6 FD_Rating Numeric
7 HDatt Numeric
8 LPhdkomp Numeric
9 Konstrukt Numeric
10 Transmis Numeric
11 FA_atti Numeric
12 PP_Index Numeric
13 LZ_tiefPRO Numeric
14 LZ_hochPRO Numeric
15 Fragen_Anz... Numeric
16 TypPRO_Er... Numeric
17 TypPRO_D... Numeric
18 TypPRO_E... Numeric
19 SuS_Aesse... Numeric
20 EEDR_Frag... Numeric
21 EEDR_Frag... Numeric
22 SA_gesamt... Numeric
Data View Variable View
Regression
Reports
Descriptive Statistics
Bayesian Statistics
Tables
Compare Means
General Linear Mode
Generalized Linear M
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reductio
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Simulation...
Quality Control
Spatial and Temporal Mod
Direct Marketing
JACK OF ALL TRADES
didaktisches Rating (Mean 1-5)
uts):
sind Sie schon insgesamt als Lehrperson t...
ik-Wissensindex (Rechenwege 2x; Fehler ...
Einstellung gegenüber halbschriftlicher Division (Mea...
Wie kompetent fühlen Sie sich, das halbschriftliche Di...
Konstruktivistisches Lernverständnis
Transimissives Lernverständnis
Attitude towards FA 1-11
PP-Index: Vorwissen*2 + Reaktion*0.5 + Gründe*0.5 +...
Prozentualer Anteil LZtief an Phasen_math
Analysis Weight:
All
Run Paste Reset Cancel Help

SPSS

Simple advantages:

- Graphical interface
- Straightforward structure of available methods
- Straightforward structure of results
- (Almost) no estimation errors

SPSS in convenient but...

SPSS is rather...

- Expensive
- Intransparent
- Not reproducible
- Focused on getting results instead of actually analyzing data
- Promotes analyzing data without understanding the process behind it

→ Potentially promotes **bad scientific practice**

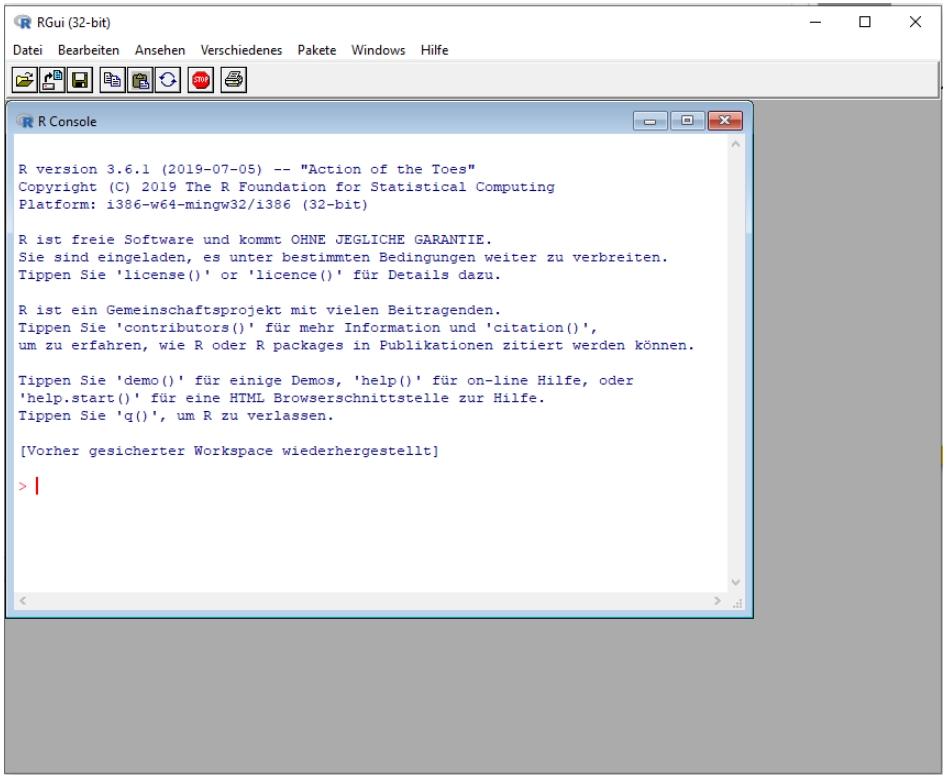
My first contact with R as a social scientist

- 2009/10: As a student tutor for an R course

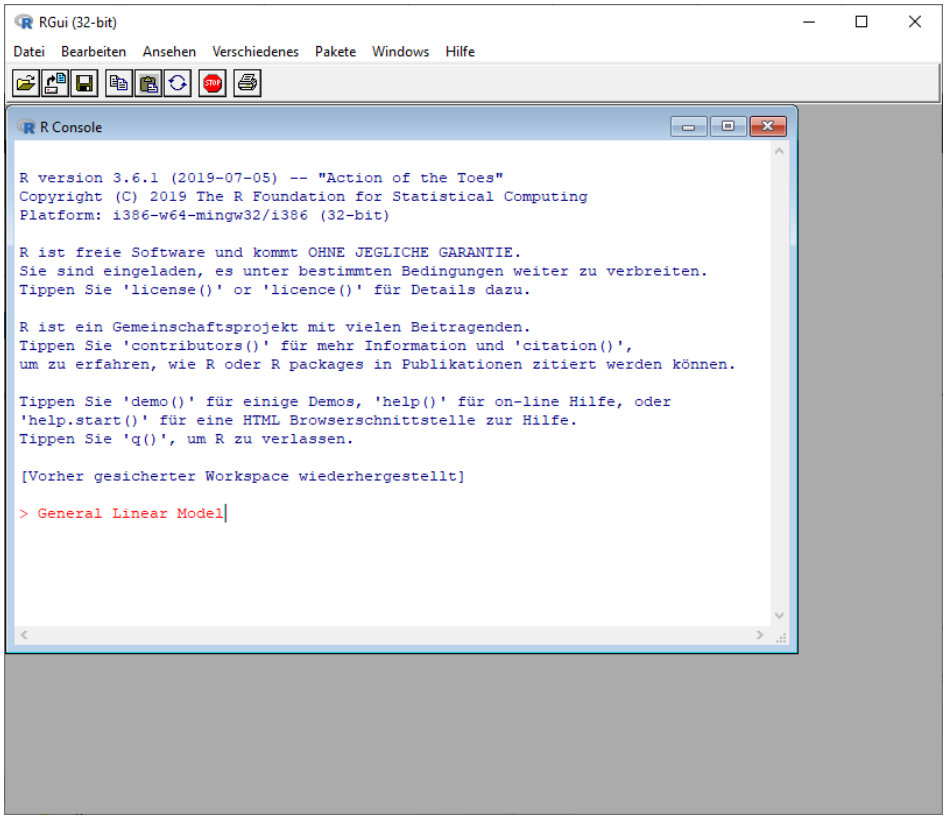
Problem:

- Back then, the main purpose of R was to replace SPSS
- People tried to make R resemble SPSS as closely as possible; yet, the same flaws remained
- For most users, R was just more inconvenient and less intuitive than SPSS

My first contact with R as a social scientist



My first contact with R as a social scientist

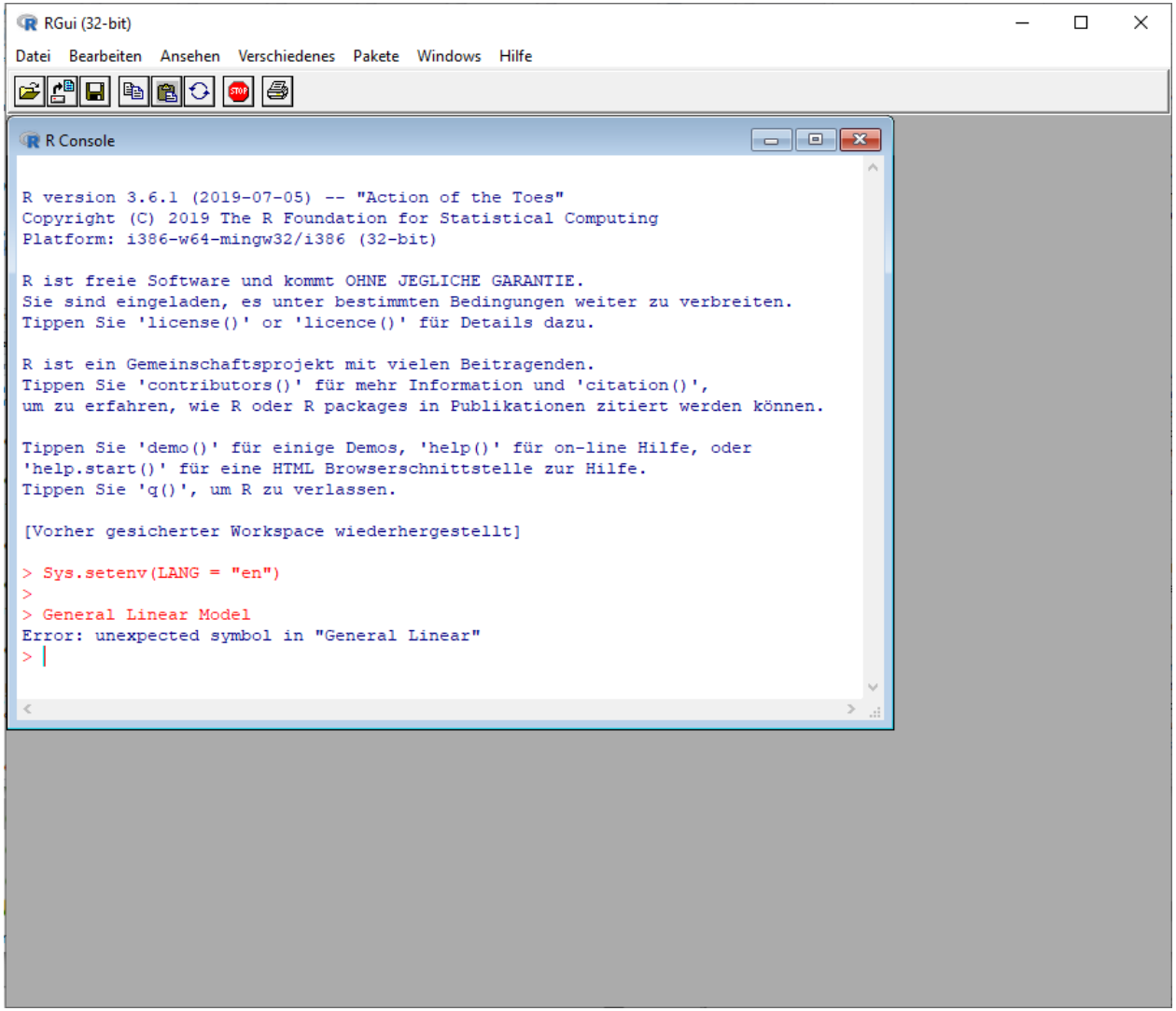


Experience with R as a statistician

- Several R courses as a statistics student (2011 – 2016)
- Discovery of the endless potential of R
  - Data simulation and shiny apps
  - Writing own functions / packages
  - Availability of source code
  - Contactable developers
  - Textmining
  - Analysis of geographical data
  - ...

Advantages of R

Forces you to know what you are doing



Advantages of R

Open Source

```
library(multilevel)
```

```
## Warning: package 'multilevel' was built under R version 4.0.2
```

```
## Loading required package: nlme
```

```
## Loading required package: MASS
```

```
ICC1
```

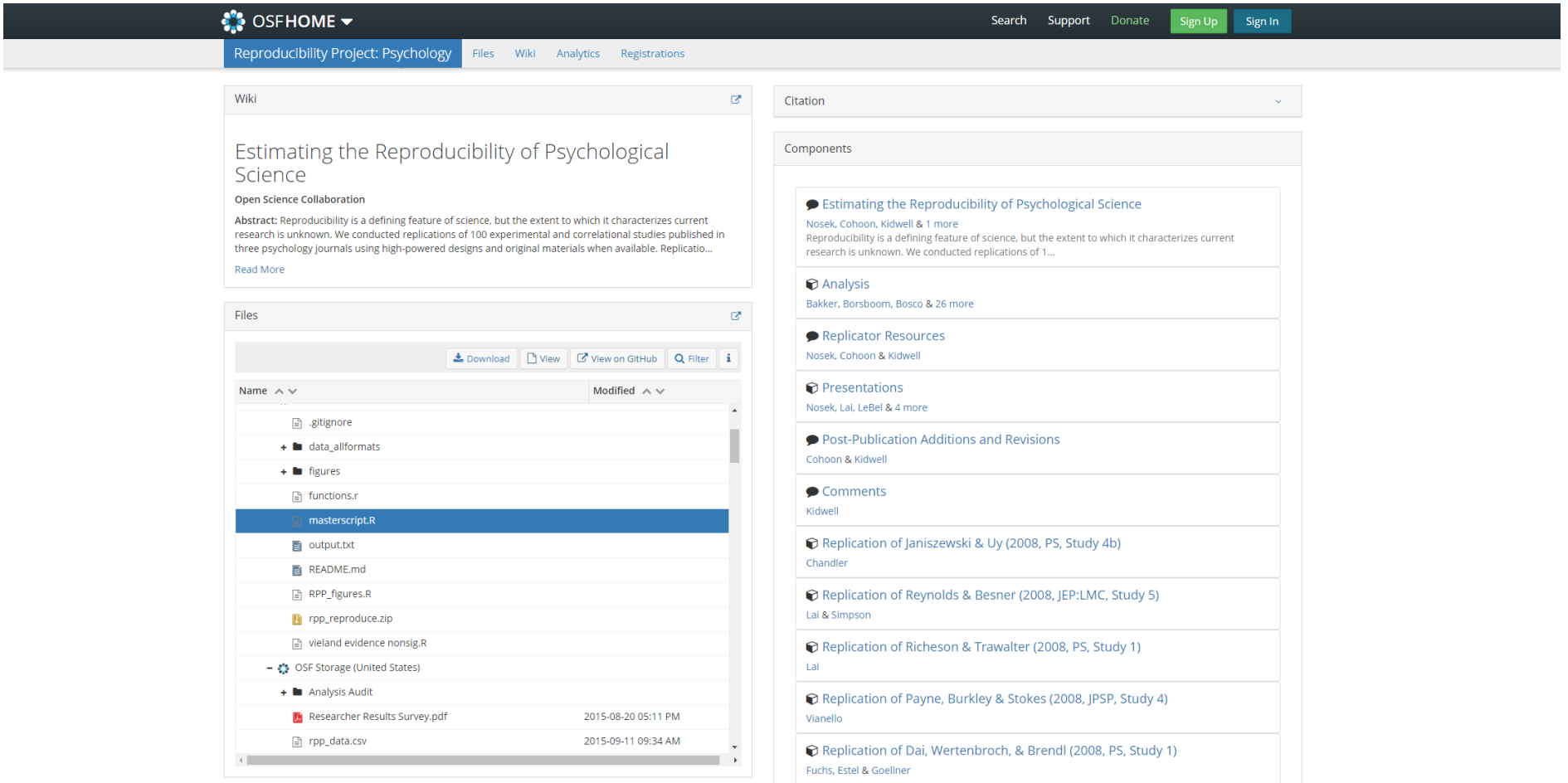
```
## function (object)
## {
##     MOD <- summary(object)
##     MSB <- MOD[[1]][1, 3]
##     MSW <- MOD[[1]][2, 3]
##     GSIZE <- (MOD[[1]][2, 1] + (MOD[[1]][1, 1] + 1))/(MOD[[1]][1,
##         1] + 1)
##     OUT <- (MSB - MSW)/(MSB + ((GSIZE - 1) * MSW))
##     return(OUT)
## }
## <bytecode: 0x0000000016016340>
## <environment: namespace:multilevel>
```

```
ICC2
```

```
## function (object)
## {
##     MOD <- summary(object)
##     MSB <- MOD[[1]][1, 3]
##     MSW <- MOD[[1]][2, 3]
##     OUT <- (MSB - MSW)/MSB
##     return(OUT)
## }
```

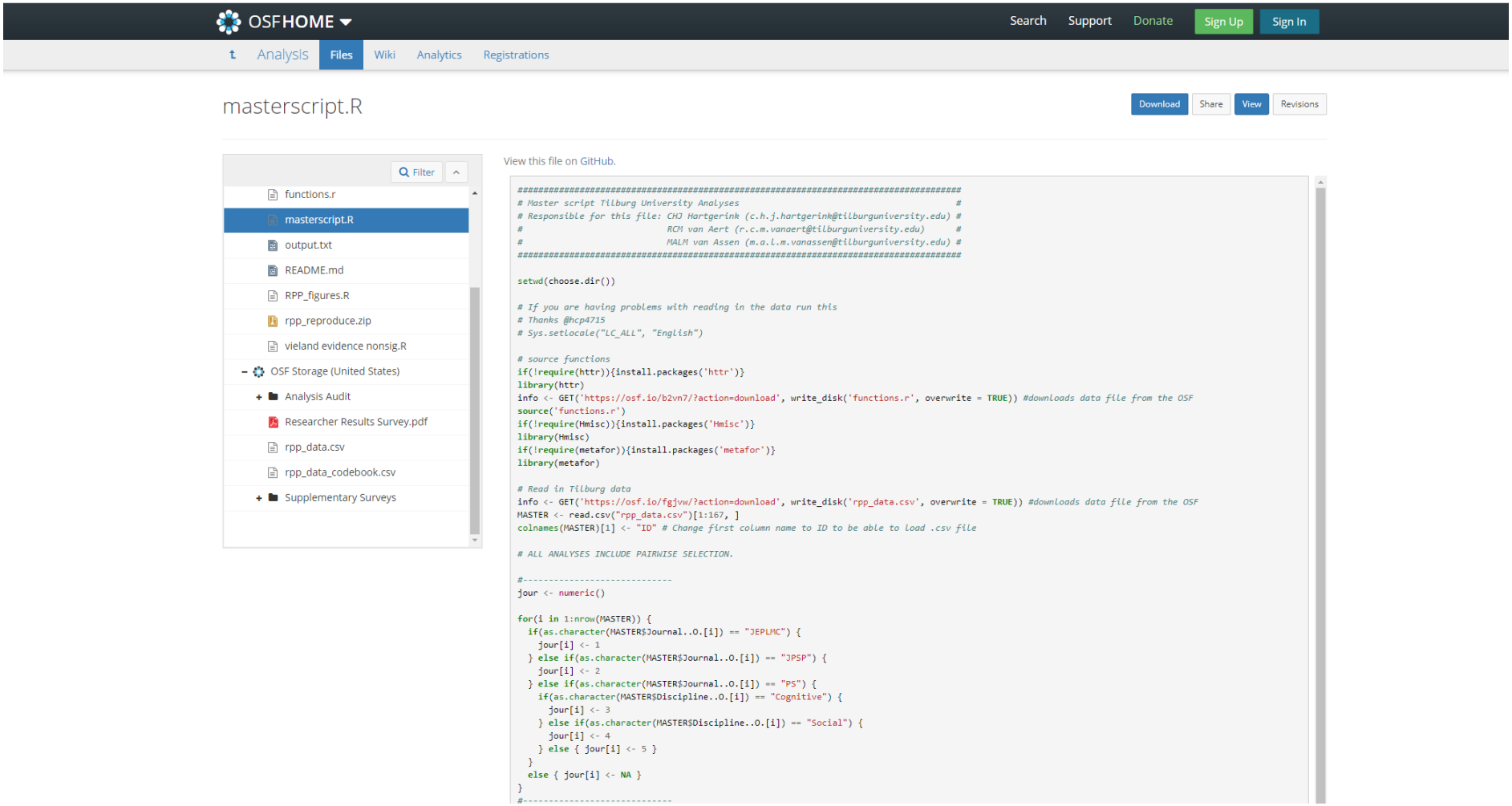
Advantages of R

Open Code / Open Data



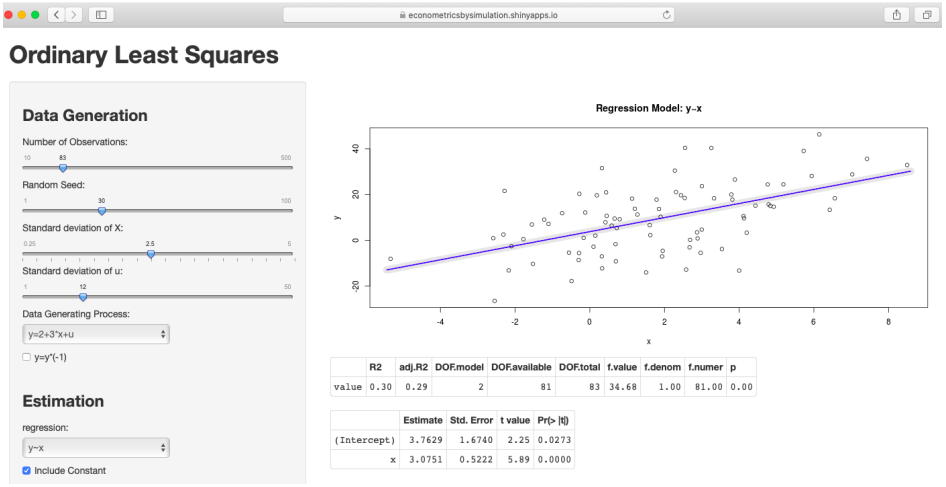
Advantages of R

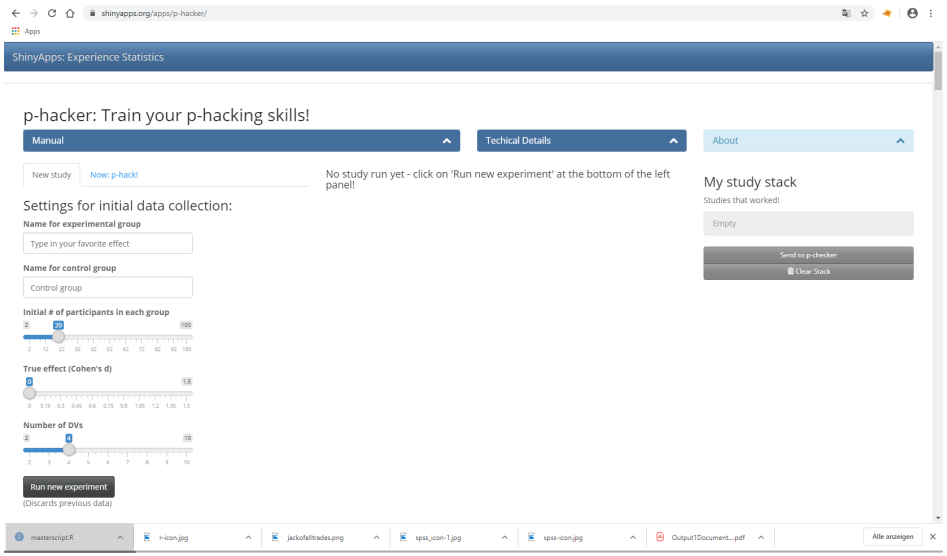
Open Code / Open Data

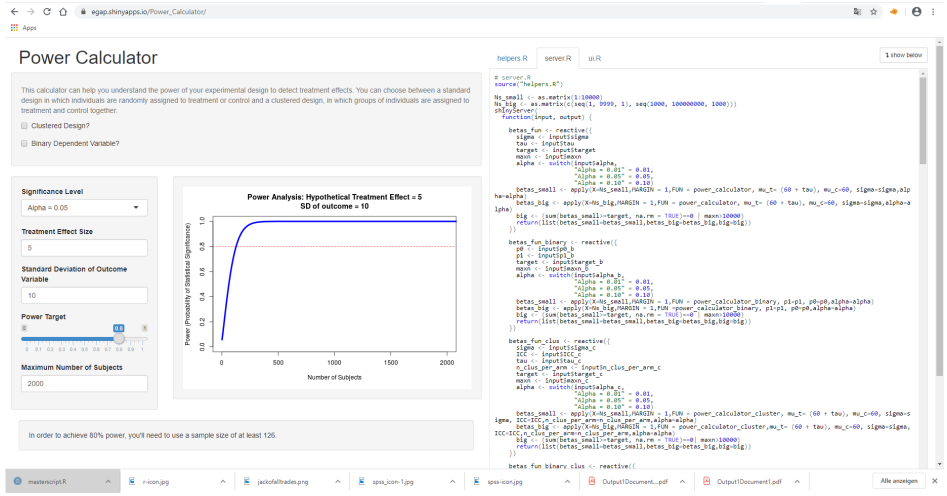


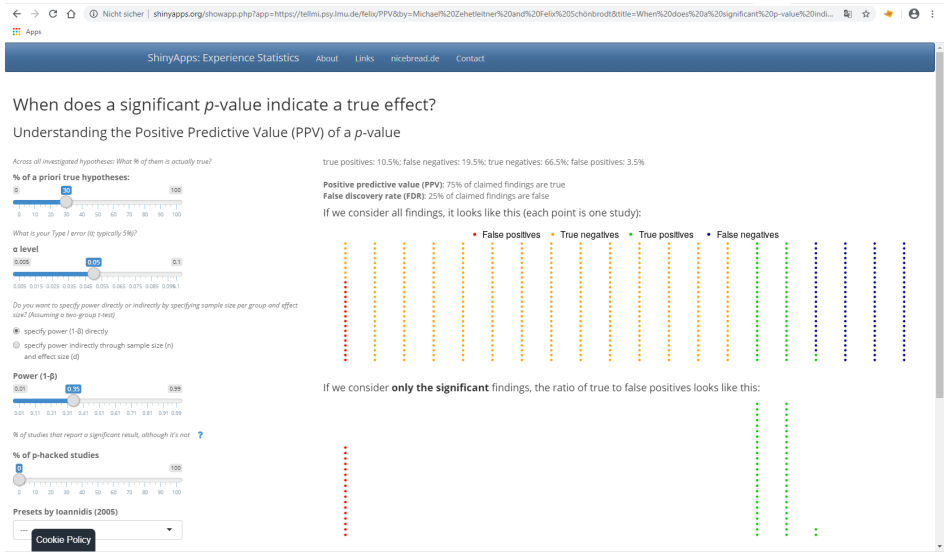
Advantages of R

Shiny Apps









How to make R accessible

In teaching

- First R, not SPSS
- Constantly use R while teaching statistical methods
- Make introduction (vectors, objects, mode, ...) as long as needed but as short as possible
- Practice A LOT: R is a language, languages need to be practiced
- Teach loops and functions only when needed (doctoral students)
- Specific R courses for what is currently needed rather than general R courses

Available packages that can help

For transferring from SPSS:

- **foreign**
- **haven**
- **expss**
- **sjlabelled**
- **labelled**

For teaching R:

- **swirl**
- graphical packages (e.g., **lattice**, **ggplot**)

For analyzing data in the social sciences:

- **psych**
- **gvlma**
- **Hmisc**
- **Multcomp**
- **lavaan**
- **lme4**
- **afex**
- **MASS**

Examples

foreign vs **haven**

```
dat.hv <- read_sav(file="Masterfile CODINGS_Basis Phasen Strategien_LONG_05.08.2019.sav")
dat.fr <- read.spss(file="Masterfile CODINGS_Basis Phasen Strategien_LONG_05.08.2019.sav", to.data.frame=T)
```

```
## re-encoding from UTF-8
```

```
str(dat.hv$IDKL)
```

```
##  num [1:28391] 100 100 100 100 100 100 100 100 100 100 ...
##  - attr(*, "label")= chr "ID Klasse"
##  - attr(*, "format.spss")= chr "F6.0"
##  - attr(*, "display_width")= int 10
```

```
str(dat.fr$IDKL)
```

```
##  num [1:28391] 100 100 100 100 100 100 100 100 100 100 ...
```

```
str(dat.hv$Basis_Form)
```

```
##  dbl+lbl [1:28391] 10, 10, 10, 10, 10, 10,  2,  2,  2,  7,  7,  7,  7,  7, ...
##  @ label        : chr "Basis_Arbeitsformen"
##  @ format.spss  : chr "F3.0"
##  @ display_width: int 11
##  @ labels       : Named num [1:11] 0 1 2 3 4 5 6 7 8 9 ...
##   ..- attr(*, "names")= chr [1:11] "keine Sozialform" "Sonstiges" "Übergänge" "SA Statement LP" ...
```

```
str(dat.fr$Basis_Form)
```

```
##  Factor w/ 11 levels "keine Sozialform",..: 11 11 11 11 11 11 3 3 3 8 ...
```

Examples

expss

```
table(dat.fr$Basis_Form)
```

```
##
## keine Sozialform        Sonstiges        Übergänge  SA Statement LP
##             1445              657             1009              120
##        SA Förder            SA GA            SA PA            SA EA
##             1165             2073             1478            10085
##        KU Förder         KU Kreis       KU frontal
##              802             4315             5242
```

```
table(dat.hv$Basis_Form)
```

```
##
##     0     1     2     3     4     5     6     7     8     9    10
##  1445   657  1009   120  1165  2073  1478 10085   802  4315  5242
```

```
cro(dat.hv$Basis_Form)
```

		#Total
Basis_Arbeitsformen	keine Sozialform	1445
	Sonstiges	657
	Übergänge	1009
	SA Statement LP	120
	SA Förder	1165
	SA GA	2073
	SA PA	1478
	SA EA	10085
	KU Förder	802
	KU Kreis	4315
	KU frontal	5242
	#Total cases	28391

Examples

psych

```
IQdata <- read.table(file="data/IQdata.txt",header=T,dec=",")
```

```
psych::describe(IQdata)
```

```
##           vars   n  mean   sd median trimmed   mad min max range  skew kurtosis
## schule       1 500  3.15 0.57      3    3.17  0.00   2   5     3  0.13     0.17
## alter        2 500 18.95 2.09     19   18.69  1.48  16  25     9  0.96     0.47
## sex          3 499  0.45 0.50      0    0.44  0.00   0   1     1  0.20    -1.96
## figural1     4 500 10.56 3.42     10   10.55  2.97   1  20    19  0.00    -0.33
## figural2     5 500 10.79 3.86     11   10.89  4.45   0  20    20 -0.24    -0.37
## figural3     6 500  8.02 3.33      8    7.75  2.97   0  19    19  0.67     0.24
## verbal       7 500 32.75 8.48     33   33.02  8.90   4  54    50 -0.28    -0.04
## numerisch    8 500 34.40 9.62     35   34.66 10.38   7  58    51 -0.25    -0.43
##              se
## schule     0.03
## alter      0.09
## sex        0.02
## figural1   0.15
## figural2   0.17
## figural3   0.15
## verbal     0.38
## numerisch  0.43
```

Examples

psych

```
library(psych)
load("data/finalOCB.RData")
```

```
psych::alpha(finalOCB[c("SE02_01", "SE02_02", "SE02_03", "SE02_04", "SE02_05", "SE02_06")])
```

```
##
## Reliability analysis
## Call: psych::alpha(x = finalOCB[c("SE02_01", "SE02_02", "SE02_03",
##     "SE02_04", "SE02_05", "SE02_06")])
##
##   raw_alpha std.alpha G6(smc) average_r S/N   ase mean sd median_r
##       0.87      0.87    0.86      0.53 6.7 0.015   51 23     0.52
##
##  lower alpha upper     95% confidence boundaries
## 0.84 0.87 0.9
##
##  Reliability if an item is dropped:
##         raw_alpha std.alpha G6(smc) average_r S/N alpha se  var.r med.r
## SE02_01      0.83      0.84    0.81      0.50 5.1    0.020 0.0067  0.50
## SE02_02      0.85      0.85    0.84      0.54 5.8    0.018 0.0091  0.53
## SE02_03      0.86      0.86    0.84      0.55 6.0    0.017 0.0077  0.53
## SE02_04      0.84      0.85    0.82      0.52 5.5    0.019 0.0070  0.52
## SE02_05      0.83      0.83    0.81      0.50 4.9    0.020 0.0047  0.50
## SE02_06      0.86      0.86    0.84      0.56 6.3    0.017 0.0059  0.54
##
##  Item statistics
##            n raw.r std.r r.cor r.drop mean sd
## SE02_01 178  0.83  0.83  0.80   0.74   42 29
## SE02_02 178  0.75  0.76  0.68   0.64   49 29
## SE02_03 178  0.74  0.74  0.66   0.61   62 32
## SE02_04 178  0.78  0.79  0.74   0.68   57 29
## SE02_05 178  0.84  0.84  0.82   0.76   51 28
```

Examples

Hmisc

```
cormat <- as.data.frame(finalOCB[c("SD01", "SD02_01", "OI", "OCB")])
```

```
mycor <- rcorr(as.matrix(cormat))
mycor
```

```
##           SD01 SD02_01    OI   OCB
## SD01      1.00   -0.11 -0.31 -0.08
## SD02_01 -0.11    1.00  0.06  0.21
## OI      -0.31    0.06  1.00  0.34
## OCB     -0.08    0.21  0.34  1.00
##
## n= 178
##
##
## P
##         SD01   SD02_01 OI     OCB
## SD01           0.1467  0.0000 0.3100
## SD02_01 0.1467         0.4347 0.0052
## OI      0.0000 0.4347         0.0000
## OCB     0.3100 0.0052  0.0000
```

How to make R accessible

In developing

- Create more inclusive environment
- Make documentation / help functions easier understandable
- Better manners on platforms, such as Stack Overflow, etc.

Conclusion

- SPSS is covenient but may enhance bad scientific practice
- R is inconvenient but may help improving scientific practice
- Tools for making R more accessible are available and should be used
- Developers could and should help making R more accessible