Teaching of survey statistics: Complex Survey Analysis and Structural Equation Models

Maria Valaste, University of Helsinki

5th Baltic-Nordic Conference on Survey Statistics
16–20 June 2019, Örebro, Sweden

Introduction

- Sample surveys are essential tools in a modern society to provide information of different areas. Surveys are used to produce Official Statistics but also e.g. to gain insight into population attitudes.
- At the moment emerging data sources and developments of big data and open data provides new opportunities (Ridgway 2016) but this data revolution also challenges the data users.
- Often the survey data are collected by a complex sampling design e.g. involving stratification, clustering and unequal inclusion probabilities.
  - Thus, the estimators should be constructed so that the complexities of the sampling design are accounted for.

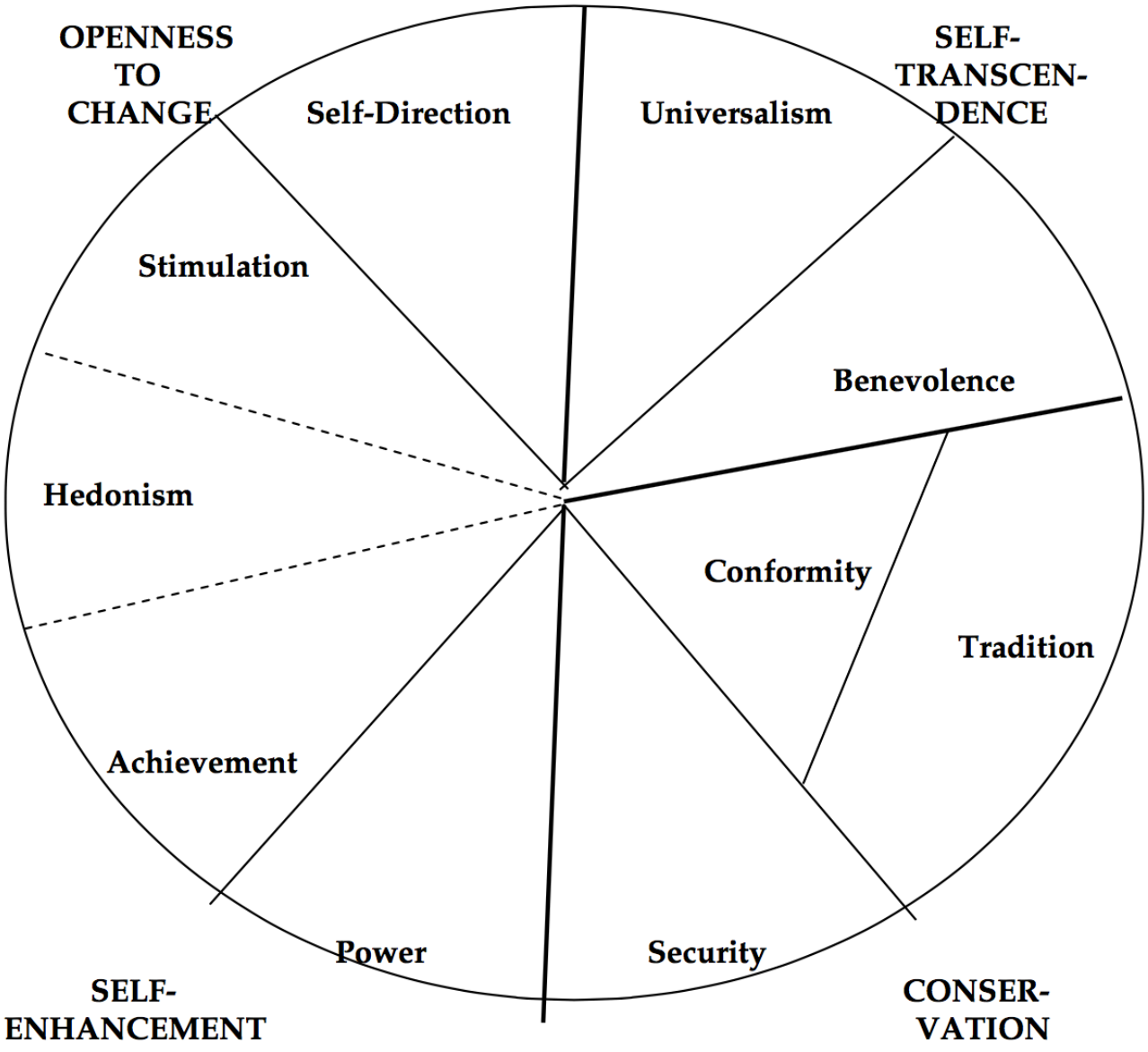
Introduction (Cont.)

- Structural equation models are often used to assess unobservable latent constructs (Bollen 1989; Vehkalahti & Everitt 2019). Structural equation modelling (SEM) is a powerful technique that is used to analyse structural relationships. SEM includes e.g.
  - confirmatory factor analysis, path analysis, and latent growth modelling.
- I'll consider some implications of the data revolutions for teaching survey statistics and I present a practical example where the design variables are included in SEM with R package lavaan.survey (Oberski 2019; Lumley 2010, 2004).

Data

- ESS 2016 data (ESS Round 8)
- All 23 countries are included
- The ESS questionnaire includes a well-established 21-item measure of human values, which was developed by the Israeli psychologist, Professor Shalom Schwartz. The 'Human Values Scale' is designed to classify respondents according to their basic value orientations. The Human Values Scale has been included in every ESS round.

Human values (documentation)



Variables

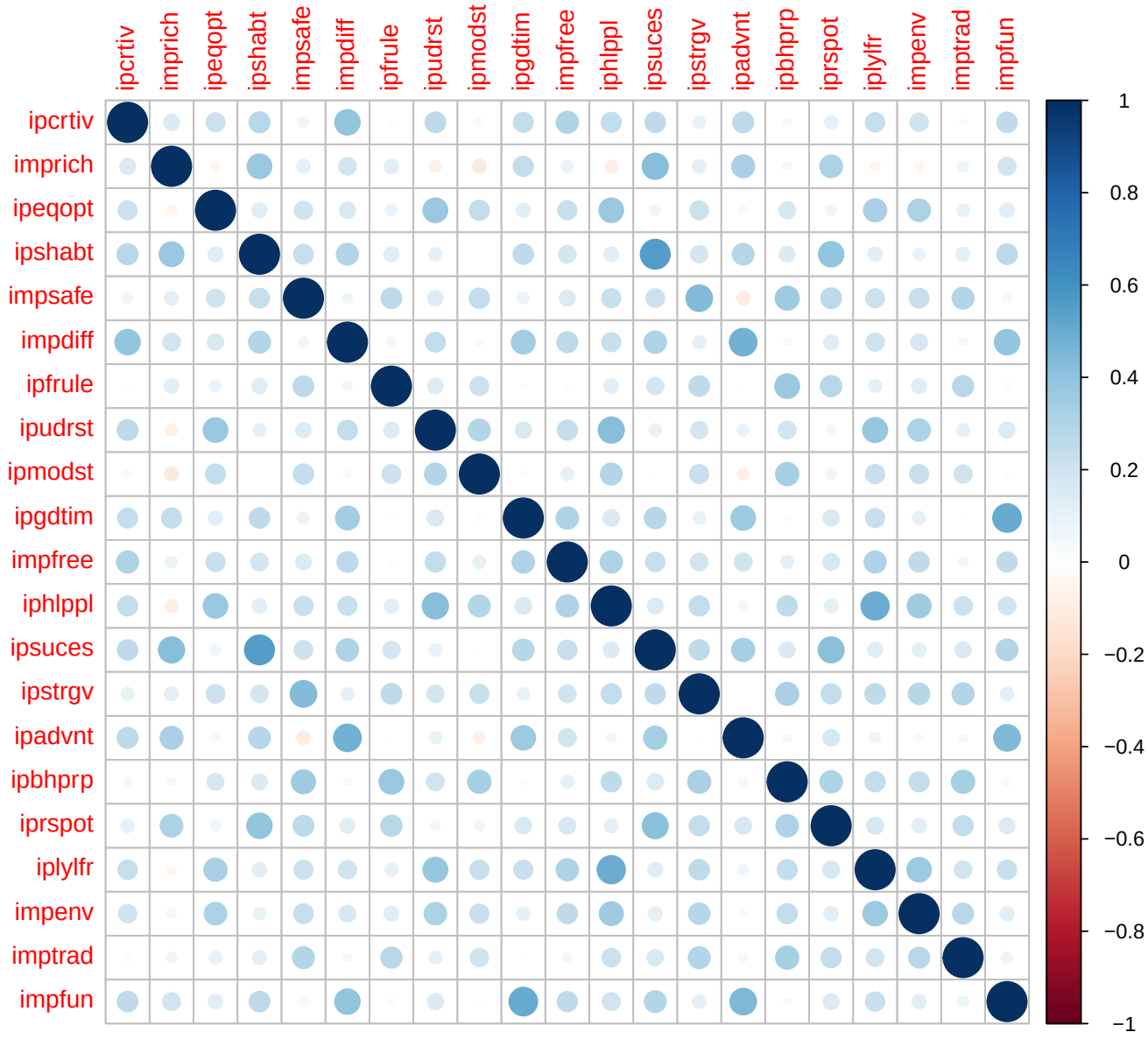
- ipcrtiv - Important to think new ideas and being creative
- imprich - Important to be rich, have money and expensive things
- ipeqopt - Important that people are treated equally and have equal opportunities
- ipshabt - Important to show abilities and be admired
- impsafe - Important to live in secure and safe surroundings
- impdiff - Important to try new and different things in life
- ipfrule - Important to do what is told and follow rules
- ipudrst - Important to understand different people
- ipmodst - Important to be humble and modest, not draw attention
- ipgdtim - Important to have a good time
- impfree - Important to make own decisions and be free
- iphlppl - Important to help people and care for others well-being
- ipsuces - Important to be successful and that people recognize achievements
- ipstrgv - Important that government is strong and ensures safety
- ipadvnt - Important to seek adventures and have an exciting life
- ipbhprp - Important to behave properly
- iprspot - Important to get respect from others
- iplylfr - Important to be loyal to friends and devote to people close
- impenv - Important to care for nature and environment
- imptrad - Important to follow traditions and customs
- impfun - Important to seek fun and things that give pleasure

Values and categories

- There are six alternatives in the questionnaire to answer.

1 Very much like me
2 Like me
3 Somewhat like me
4 A little like me
5 Not like me
6 Not like me at all
7 Refusal
8 Don't know
9 No answer

Correlations



Descriptive Statistics for Observed Variables

Variable	Mean	SD	min	max	Missing (%)
impdiff	3.01	1.36	1	6	1.86
impenv	2.18	1.05	1	6	1.71
impfree	2.18	1.10	1	6	1.79
impfun	3.00	1.34	1	6	1.86
imprich	4.11	1.34	1	6	1.79
impsafe	2.37	1.23	1	6	1.68
imptrad	2.76	1.36	1	6	1.70
ipadvnt	3.85	1.45	1	6	1.89
ipbhprp	2.71	1.25	1	6	1.97
ipcrtiv	2.59	1.26	1	6	1.91
ipeqopt	2.18	1.08	1	6	1.85
ipfrule	3.23	1.40	1	6	2.30
ipgdtim	2.92	1.33	1	6	1.84
iphlppl	2.20	1.01	1	6	1.80
iplylfr	1.96	0.95	1	6	1.75
ipmodst	2.65	1.22	1	6	1.93
iprspot	3.17	1.37	1	6	2.26
ipshabt	3.23	1.40	1	6	2.02
ipstrgv	2.34	1.20	1	6	2.26
ipsuces	3.19	1.37	1	6	2.09
ipudrst	2.38	1.08	1	6	1.99

Model and tools

- Idea is to perform factor analysis and take into account the survey design.
- Possible tools:
  - for exploratory factor analysis (EFA): stats::factanal(), psych::fa() and
  - for confirmatory factor analysis (CFA) lavaan::cfa()
  - lavaan.survey::lavaan.survey() to fit factor analysis model while taking the survey design into account
  - survey::svyfactanal() fit factor analysis model in complex surveys (experimental).
- The book by Seppo Laaksonen (2018, pp. 22–26) inspired the model: fit the 4-factor model and take the survey design into account.

First solution

```r
# Factor analysis: 4 factors, varimax-rotation and save factor scores
factors <- factanal(values_nomiss, 4, rotation="varimax",
                    scores="regression")
print(factors, digits=2, cutoff=.3, sort=TRUE) # print solution
```

```
##
## Call:
## factanal(x = values_nomiss, factors = 4, scores = "regression",     rotation
##
## Uniquenesses:
## ipcrtiv imprich ipeqopt ipshabt impsafe impdiff ipfrule ipudrst ipmodst
##    0.68    0.59    0.69    0.49    0.64    0.60    0.73    0.61    0.73
## ipgdtim impfree iphlppl ipsuces ipstrgv ipadvnt ipbhprp iprspot iplylfr
##    0.58    0.74    0.52    0.44    0.67    0.54    0.58    0.61    0.58
##  impenv imptrad  impfun
##    0.69    0.72    0.45
##
## Loadings:
##         Factor1 Factor2 Factor3 Factor4
## ipeqopt  0.54
## ipudrst  0.60
## iphlppl  0.64
## iplylfr  0.57
## impsafe          0.54
```

First solution (Cont.)

```
## ipfrule          0.50
## ipstrgv          0.51
## ipbhprp          0.62
## imptrad          0.52
## ipgdtim                  0.60
## ipadvnt                  0.58    0.32
## impfun                   0.72
## imprich                          0.53
## ipshabt                          0.66
## ipsuces                          0.66
## ipcrtiv  0.41
## impdiff  0.30            0.48
## ipmodst  0.33    0.38
## impfree  0.41
## iprspot          0.42            0.44
## impenv   0.48
##
##                Factor1 Factor2 Factor3 Factor4
## SS loadings       2.41    2.11    1.83    1.78
## Proportion Var    0.11    0.10    0.09    0.08
## Cumulative Var    0.11    0.22    0.30    0.39
##
## Test of the hypothesis that 4 factors are sufficient.
## The chi square statistic is 12747.3 on 132 degrees of freedom.
## The p-value is 0
```

Second solution

```
# fit model, package lavaan
l <- ' equality  =~ ipeqopt+ipudrst+iphlppl+iplylfr+ipcrtiv+impfree+iphlppl+iplylfr+impenv
       tradition =~ impsafe+ipfrule+ipmodst+ipstrgv+ipbhprp+imptrad
       success   =~ imprich+ipshabt+ipsuces+iprspot
       enjoy     =~ ipadvnt+impfun +impdiff+ipgdtim
     '
fit_lavaan <- cfa(l, data=ESS_part, orthogonal=TRUE)

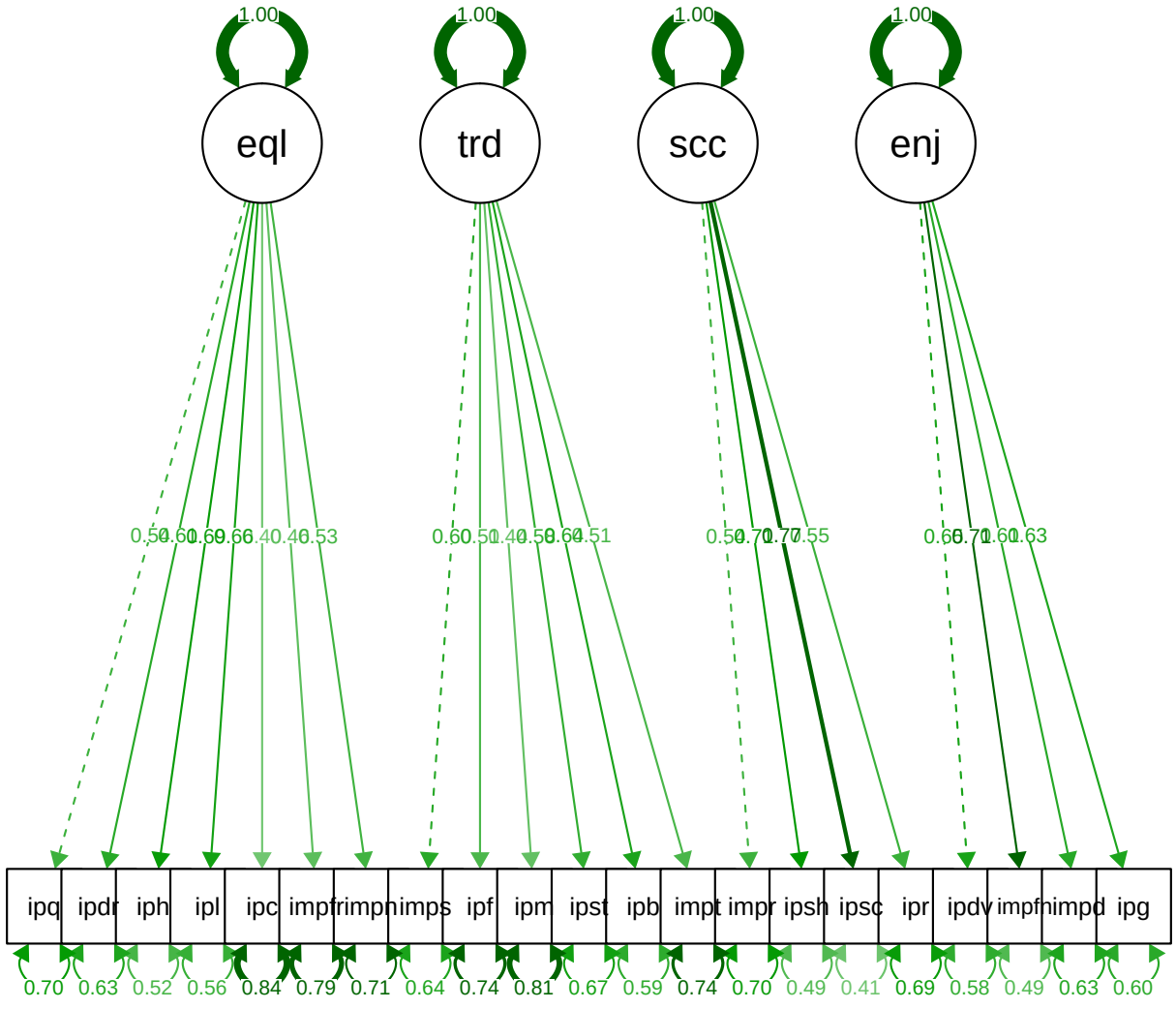
# take out the loadings from the results
inspect(fit_lavaan,what="std")$lambda
```

```
##         equlty tradtn succss enjoy
## ipeqopt  0.545  0.000  0.000 0.000
## ipudrst  0.606  0.000  0.000 0.000
## iphlppl  0.693  0.000  0.000 0.000
## iplylfr  0.662  0.000  0.000 0.000
## ipcrtiv  0.400  0.000  0.000 0.000
## impfree  0.457  0.000  0.000 0.000
## impenv   0.534  0.000  0.000 0.000
## impsafe  0.000  0.602  0.000 0.000
## ipfrule  0.000  0.505  0.000 0.000
## ipmodst  0.000  0.438  0.000 0.000
## ipstrgv  0.000  0.577  0.000 0.000
## ipbhprp  0.000  0.641  0.000 0.000
## imptrad  0.000  0.509  0.000 0.000
## imprich  0.000  0.000  0.544 0.000
## ipshabt  0.000  0.000  0.714 0.000
## ipsuces  0.000  0.000  0.768 0.000
## iprspot  0.000  0.000  0.554 0.000
## ipadvnt  0.000  0.000  0.000 0.647
## impfun   0.000  0.000  0.000 0.715
## impdiff  0.000  0.000  0.000 0.607
## ipgdtim  0.000  0.000  0.000 0.633
```

Second solution (Cont.)

```
# a plot, std display the standardized parameter estimates
# (print standardized estimates: standardizedSolution(fit_lavaan))
semPaths(fit_lavaan, title = FALSE, curvePivot = TRUE, what="std")
```

Second solution (Cont.)



Final model

- Create weight variable (Guide to weighting of ESS data)
- Define design

```
# add weight
ESS_part$WEIGHT <- ESS_part$pspwght*ESS_part$pweight

# define design
design_values <- svydesign(id =~ 1, weights=~WEIGHT,
                           data=ESS_part)
```

Final model (Cont.)

```
# define model, package lavaan
mls <- ' equality  =~ ipeqopt+ipudrst+iphlppl+iplylfr+ipcrtiv+impfree+iphlppl+iplylfr+impenv
         tradition =~ impsafe+ipfrule+ipmodst+ipstrgv+ipbhprp+imptrad
         success   =~ imprich+ipshabt+ipsuces+iprspot
         enjoy     =~ ipadvnt+impfun +impdiff+ipgdtim
       '
# Fit the model using lavaan
fit_lavaan_survey <- cfa(mls, data=ESS_part, orthogonal=TRUE)
#summary(fit_lavaan_survey, fit.measures=TRUE)
inspect(fit_lavaan_survey,what="std")$lambda
```

```
##          equlty tradtn succss enjoy
## ipeqopt   0.545  0.000  0.000 0.000
## ipudrst   0.606  0.000  0.000 0.000
## iphlppl   0.693  0.000  0.000 0.000
## iplylfr   0.662  0.000  0.000 0.000
## ipcrtiv   0.400  0.000  0.000 0.000
## impfree   0.457  0.000  0.000 0.000
## impenv    0.534  0.000  0.000 0.000
## impsafe   0.000  0.602  0.000 0.000
## ipfrule   0.000  0.505  0.000 0.000
## ipmodst   0.000  0.438  0.000 0.000
## ipstrgv   0.000  0.577  0.000 0.000
## ipbhprp   0.000  0.641  0.000 0.000
## imptrad   0.000  0.509  0.000 0.000
## imprich   0.000  0.000  0.544 0.000
## ipshabt   0.000  0.000  0.714 0.000
## ipsuces   0.000  0.000  0.768 0.000
## iprspot   0.000  0.000  0.554 0.000
## ipadvnt   0.000  0.000  0.000 0.647
## impfun    0.000  0.000  0.000 0.715
## impdiff   0.000  0.000  0.000 0.607
## ipgdtim   0.000  0.000  0.000 0.633
```

Final model (Cont.)

```
# Fit the 4-factor model while taking the survey design into account.
fit.cfa.surv <- lavaan.survey(fit_lavaan_survey, survey.design = design_values)
#fit.cfa.surv
#summary(fit.cfa.surv, fit.measures=TRUE)
inspect(fit.cfa.surv,what="std")$lambda
```

```
##         equlty tradtn succss enjoy
## ipeqopt  0.531  0.000  0.000 0.000
## ipudrst  0.601  0.000  0.000 0.000
## iphlppl  0.689  0.000  0.000 0.000
## iplylfr  0.645  0.000  0.000 0.000
## ipcrtiv  0.404  0.000  0.000 0.000
## impfree  0.446  0.000  0.000 0.000
## impenv   0.492  0.000  0.000 0.000
## impsafe  0.000  0.611  0.000 0.000
## ipfrule  0.000  0.514  0.000 0.000
## ipmodst  0.000  0.442  0.000 0.000
## ipstrgv  0.000  0.570  0.000 0.000
## ipbhprp  0.000  0.641  0.000 0.000
## imptrad  0.000  0.506  0.000 0.000
## imprich  0.000  0.000  0.558 0.000
## ipshabt  0.000  0.000  0.699 0.000
## ipsuces  0.000  0.000  0.765 0.000
## iprspot  0.000  0.000  0.521 0.000
## ipadvnt  0.000  0.000  0.000 0.626
## impfun   0.000  0.000  0.000 0.750
## impdiff  0.000  0.000  0.000 0.565
## ipgdtim  0.000  0.000  0.000 0.645
```

References

Bollen, K. (1989). *Structural Equations with Latent Variables.* John Wiley & Sons, New York.

ESS Round 8: European Social Survey Round 8 Data (2016). Data file edition 2.1. NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC. doi:10.21338/NSD-ESS8-2016.

Laaksonen, S. (2018). *Survey Methodology and Missing Data. Tools and Techniques for Practitioners.* Cham, Switzerland: Springer.

Lumley, T. (2004). Analysis of Complex Survey Samples. *Journal of Statistical Software*, 9(8).

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R.* John Wiley & Sons, New York.

Oberski, D. (2019) lavaan.survey. Complex Survey Structural Equation Modeling (SEM). R package version 1.1.3.1, URL: https://cran.r-project.org/web/packages/lavaan.survey/index.html

Ridgway, J. (2016). Implications of the Data Revolution for Statistics Education. *International Statistical Review*, 84(3), pp. 528–549. doi:10.1111/insr.12110.

Vehkalahti, K. & Everitt, B. (2019). *Multivariate Analysis for the Behavioral Sciences*, Second Edition. Boca Raton, Florida: CRC Press.

Extra: Working environment

MV_BaNoCoSS2019_2.Rmd

Knit | Insert | Run

```
```{r, echo = TRUE}
# define model, package lavaan
mls <- ' equality  =~ ipeqopt+ipudrst+iphlppl+iplylfr+ipcrtiv+impfree+iphlppl+iplylfr+impenv
         tradition =~ impsafe+ipfrule+ipmodst+ipstrgv+ipbhprp+imptrad
         success   =~ imprich+ipshabt+ipsuces+iprspot
         enjoy     =~ ipadvnt+impfun +impdiff+ipgdtim
       '

# Fit the model using lavaan
fit_lavaan_survey <- cfa(mls, data=ESS_part, orthogonal=TRUE)
#summary(fit_lavaan_survey, fit.measures=TRUE)
inspect(fit_lavaan_survey,what="std")$lambda

# Fit the 4-factor model while taking the survey design into account.
fit.cfa.surv <- lavaan.survey(fit_lavaan_survey, survey.design = design_values)
#fit.cfa.surv
#summary(fit.cfa.surv, fit.measures=TRUE)
inspect(fit.cfa.surv,what="std")$lambda
```
<!--
a plot, std display the standardized parameter estimates
#semPaths(fit.cfa.surv, title = FALSE, curvePivot = TRUE, what="std")
-->

References

Bollen, K. (1989). *Structural Equations with Latent Variables.* John Wiley \& Sons, New York.

Laaksonen, S. (2018). *Survey Methodology and Missing Data. Tools and Techniques for Practitioners.* Cham, Switzerland: Springer.

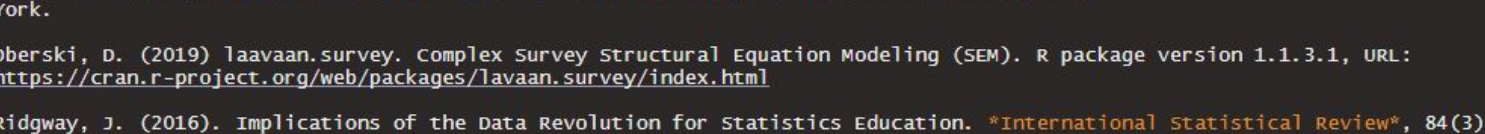
Lumley, T. (2004). Analysis of Complex Survey Samples. *Journal of Statistical Software*, 9(8).

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R.* John Wiley \& Sons, New York.

Oberski, D. (2019) laavaan.survey. Complex Survey Structural Equation Modeling (SEM). R package version 1.1.3.1, URL: https://cran.r-project.org/web/packages/lavaan.survey/index.html

Ridgway, J. (2016). Implications of the Data Revolution for Statistics Education. *International Statistical Review*, 84(3), pp. 528--549. DOI 10.1111/insr.12110.

Extra: Working environment



280:71 | Extra: Working environment | R Markdown
```