A pipeline for twin studies

An application of the **knitr** package

Stéphanie van den Berg

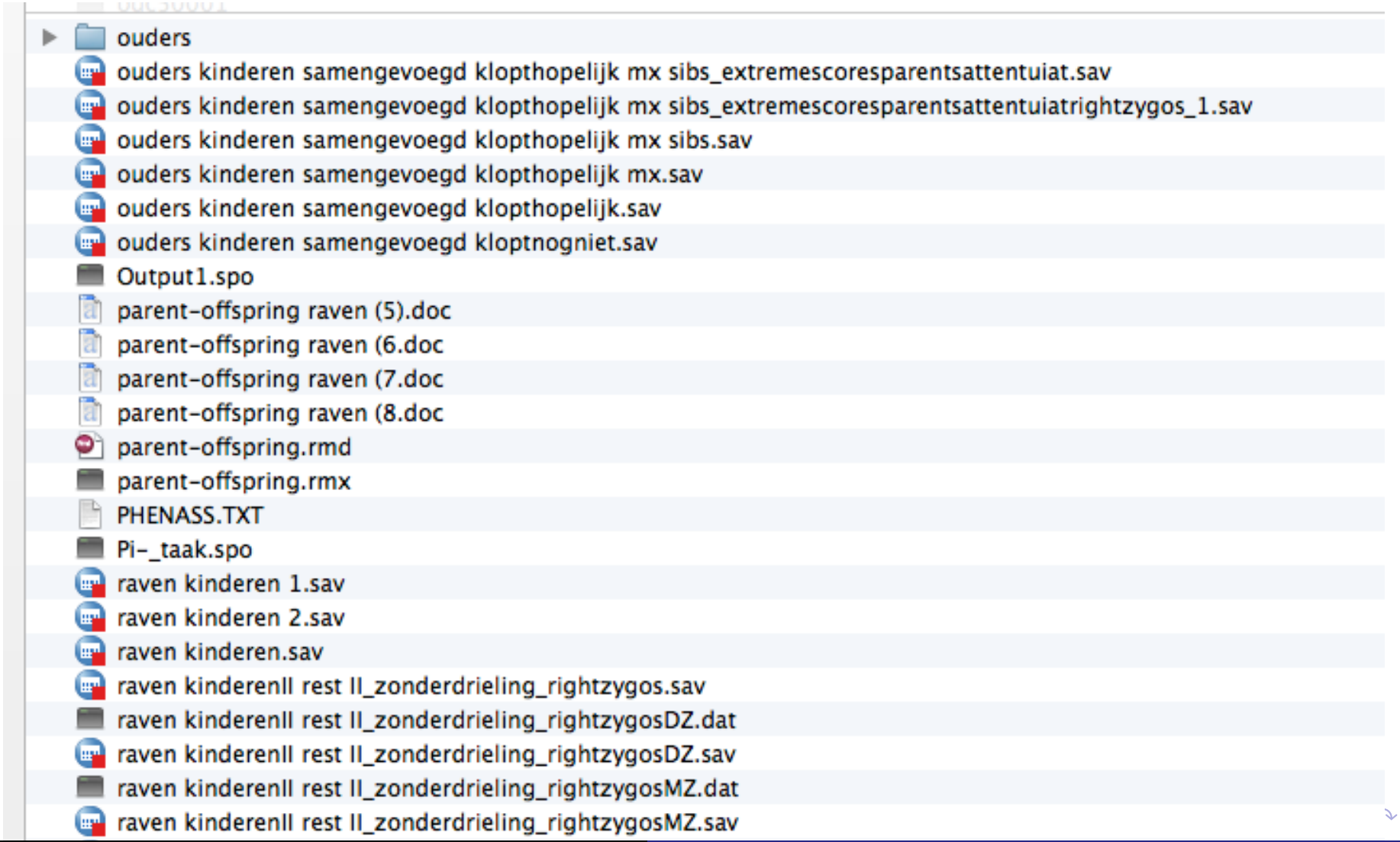
December 15, 2014

Reproducible Research

For yourself and for your peers.
Two types of reproducibility:

- Methodological replication
  - Data collection
  - Data selection
  - Data treatment
  - Statistical model and estimation
- Replication of non-exact statistical results
  - Bootstrap methods
  - Permutation tests
  - Sampling methods, e.g. MCMC

Data files


ouders
ouders kinderen samengevoegd klopthopelijk mx sibs_extremescoresparentsattentuiat.sav
ouders kinderen samengevoegd klopthopelijk mx sibs_extremescoresparentsattentuiatrightzygos_1.sav
ouders kinderen samengevoegd klopthopelijk mx sibs.sav
ouders kinderen samengevoegd klopthopelijk mx.sav
ouders kinderen samengevoegd klopthopelijk.sav
ouders kinderen samengevoegd kloptnogniet.sav
Output1.spo
parent-offspring raven (5).doc
parent-offspring raven (6.doc
parent-offspring raven (7.doc
parent-offspring raven (8.doc
parent-offspring.rmd
parent-offspring.rmx
PHENASS.TXT
Pi-_taak.spo
raven kinderen 1.sav
raven kinderen 2.sav
raven kinderen.sav
raven kinderenII rest II_zonderdrieling_rightzygos.sav
raven kinderenII rest II_zonderdrieling_rightzygosDZ.dat
raven kinderenII rest II_zonderdrieling_rightzygosDZ.sav
raven kinderenII rest II_zonderdrieling_rightzygosMZ.dat
raven kinderenII rest II_zonderdrieling_rightzygosMZ.sav

Importance of limited number of files

Preferably:

- One master data file
- One syntax file
- One output file
- One manuscript

Importance of one syntax file

- read master data file
- select cases and variables
- recoding, restructuring
- treatment of missing data
- full analysis
- save one final output file

From SPSS to R

R forces you to use syntax (no click and play).

R allows you to save workspaces.

```
# perform a computationally demanding analysis
data <- read.table('datafile.dat')
output <- do.analysis(data=data, NIterations=1000000, NChains=36)
# save your workspace, including all your results so far
save.image('mcmc_results.RData')
# next month, you can load this workspace and continue working
load('mcmc_results.RData')
summary(output); plot(output)
```

Weaving and Knitting

Even better:
Integrate analysis and word processing
One file where it's always clear:

- how the numbers and your tables ended up in your manuscript
- what the figures actually represent

Dynamic Report Generation

The idea is to weave or knit both statistical language and manuscript with text, tables and figures.
"dynamic report generation"

Dynamic Report Generation

- Results are automatically changed with new or extra data
- Results are automatically changed with alternate treatment of data

Change results in text, tables, figures, title, anything you like!
Even interpretation . . .

knitr

Knitting R language with different text formats, eg. LaTeX, HTML, and Markdown.
I use it with RStudio. There you can make .Rnw documents that contain both R syntax and LaTeX syntax.

IQ example

Simulate data, compute mean and standard deviation, and plot histogram.

Twin concordance rates

Suppose 1% of the population has autism, so prevalence π is 0.01.
Q: Is there a genetic component?
Is it more likely to be autistic if you have a twin sibling that is autistic?

- $P(\text{Affected}) = \pi$
- $P(\text{Affected}|\text{co-twin is affected}) > \pi$
- $P(\text{Affected}|\text{identical co-twin is affected}) >$
  $P(\text{Affected}|\text{fraternal co-twin is affected}) > \pi$

Twin tables

Identical twins

	Affected	Healthy
Affected	20	6
Healthy	8	9000

Fraternal twins

	Affected	Healthy
Affected	15	8
Healthy	9	9000

Twin tables

Identical twins

Both Affected	Discordant	Both Healthy
20	14	9000

Fraternal twins

Both Affected	Discordant	Both Healthy
15	17	9000

Bayesian estimation of twin concordance rates

Van den Berg & Hjelmborg (2012). *Behavior Genetics*, *42*(5), 857-65.

$$
\begin{aligned}
g(\lambda, \mu^{\mathrm{MZ}}, \mu^{\mathrm{DZ}} | \mathbf{y}^{\mathrm{MZ}}, \mathbf{y}^{\mathrm{DZ}}) \quad &= \quad p\left(\frac{\exp(\lambda)}{1+\exp(\lambda)}, \frac{\exp\left(\mu^{\mathrm{MZ}}\right)-1}{1+\exp\left(\mu^{\mathrm{MZ}}\right)}, \frac{\exp\left(\mu^{\mathrm{MZ}}\right)-1}{1+\exp\left(\mu^{\mathrm{MZ}}\right)} | \mathbf{y}^{\mathrm{MZ}}, \mathbf{y}^{\mathrm{DZ}} \right) \\
&\times \quad \frac{\exp(\lambda)}{(1+\exp(\lambda))^2} \frac{2\exp(\mu^{\mathrm{MZ}})}{(1+\exp(\mu^{\mathrm{MZ}}))^2} \frac{2\exp(\mu^{\mathrm{DZ}})}{(1+\exp(\mu^{\mathrm{DZ}}))^2}
\end{aligned}
$$

Too complicated for the average twin researcher . . .

A pipeline for concordance rates

One can develop an .Rnw file that only requires plugging in your data on twins, and *schwoop*, you immediately get results: text, tables and figures. Thus, from raw data to manuscript in one click.

A pipeline for concordance rates

We're in the process of developing a general pipeline for analysing twin data:

- qualitative and quantitative
- univariate and multivariate
- with or without psychometric models
- continuous and ordinal data
- with or without modelling genotype-environment interaction

Discussion

Several ethical issues involved:

- how much do you help with interpretation?
- how much of the method and results sections do you prepare?
- database fishing expeditions
- method fishing expeditions

Discussion

Pros:

- standardization of presenting results
- complete tractability of analysis
- not unlike current practice of using scripts
- not unlike current practice of following standard procedures
- supports the use of newly developed methods (BayesTwin `R` package)