Reproducible (Open) Science

Ontwikkelfestival ‘R for staRters’

Marc A.T. Teunis, PhD

2021-11-07 19:11:17

Contents

This is part 1 of a series of three course days

- Part 1; Introducing R
- Part 2; Data Wrangling
- Part 3; Visualizations and a bit statistics

The complete source code for the webinars and all dependent data, and files can be found on Github.com/uashogeschoolutrecht.

In part 3, I will show you how to use this Github resource for your own work.

Introducing Reproducible (Open) Science

1. When things go wrong
2. Why Reproducible (Open) Science?
3. The need for learning programming
4. An example of Reproducible (Open) Science

Reproducible (Open) Science =
Reproducible Research + Open Science

Data, methods and logic

Brown, Kaiser & Allison, PNAS, 2018

"...in science, three things matter:

1. the data,

everything else is a distraction."

Data, methods and logic

Brown, Kaiser & Allison, PNAS, 2018

"...in science, three things matter:

1. the data,
2. the methods used to collect the data [...], and

everything else is a distraction."

Data, methods and logic

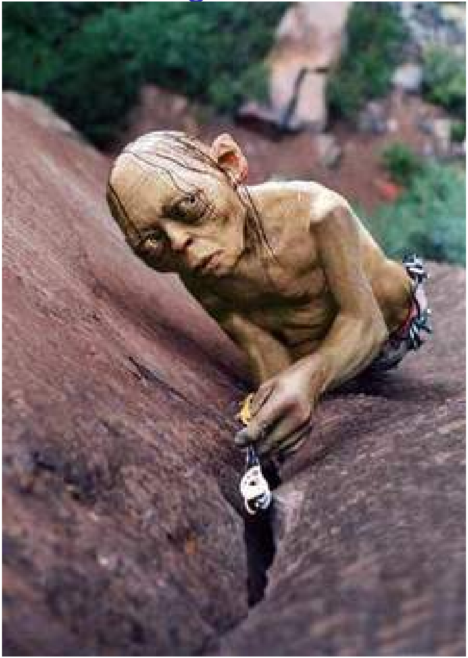
Brown, Kaiser & Allison, PNAS, 2018

"...in science, three things matter:

1. the data,
2. the methods used to collect the data [...], and
3. the logic connecting the data and methods to conclusions,

everything else is a distraction."

Gollums lurking about



Why we need Reproducible (Open) Science?

- To assess validity of science and methods we need access to data, methods and conclusions

Nature Collection on this topic

Why we need Reproducible (Open) Science?

- To assess validity of science and methods we need access to data, methods and conclusions
- To learn from choices other researchers made

Nature Collection on this topic

Why we need Reproducible (Open) Science?

- To assess validity of science and methods we need access to data, methods and conclusions
- To learn from choices other researchers made
- To learn from omissions, mistakes or errors

Nature Collection on this topic

Why we need Reproducible (Open) Science?

- To assess validity of science and methods we need access to data, methods and conclusions
- To learn from choices other researchers made
- To learn from omissions, mistakes or errors
- To prevent publication bias (also negative results will be available in reproducible research)

Nature Collection on this topic

Why we need Reproducible (Open) Science?

- To assess validity of science and methods we need access to data, methods and conclusions
- To learn from choices other researchers made
- To learn from omissions, mistakes or errors
- To prevent publication bias (also negative results will be available in reproducible research)
- To be able to re-use and/or synthesize data (from many and diverse sources)

Nature Collection on this topic

Why we need Reproducible (Open) Science?

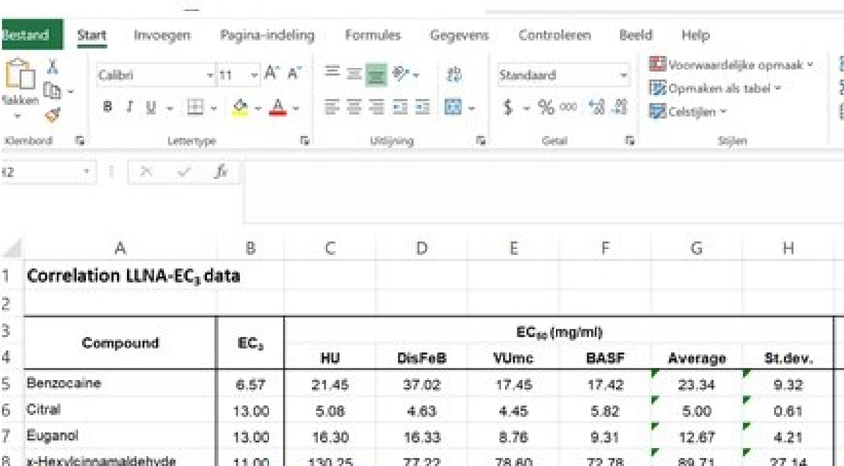
- To assess validity of science and methods we need access to data, methods and conclusions
- To learn from choices other researchers made
- To learn from omissions, mistakes or errors
- To prevent publication bias (also negative results will be available in reproducible research)
- To be able to re-use and/or synthesize data (from many and diverse sources)
- To have access to it all!

Nature Collection on this topic

The *GUI problem*

How would you ‘describe’ the steps of an analysis or creation of a graph when you use GUI* based software?

“You can only do this using code, so it is (basically) impossible in a GUI”

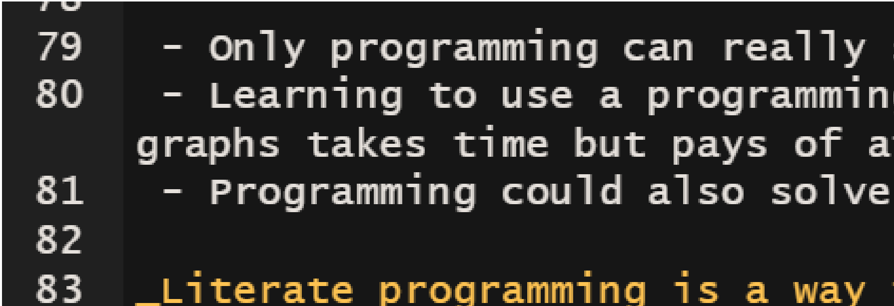


Compound	EC_3	EC_{50} (mg/ml)					
		HU	DisFeB	VUmc	BASF	Average	St.dev.
Benzocaine	6.57	21.45	37.02	17.45	17.42	23.34	9.32
Citral	13.00	5.08	4.63	4.45	5.82	5.00	0.61
Euganol	13.00	16.30	16.33	8.76	9.31	12.67	4.21
α-Hexylcinnamaldehyde	11.00	130.25	77.22	78.60	72.78	89.71	27.14

Programming is essential for Reproducible (Open) Science

- Only programming an analysis (or creation of a graph) records every step

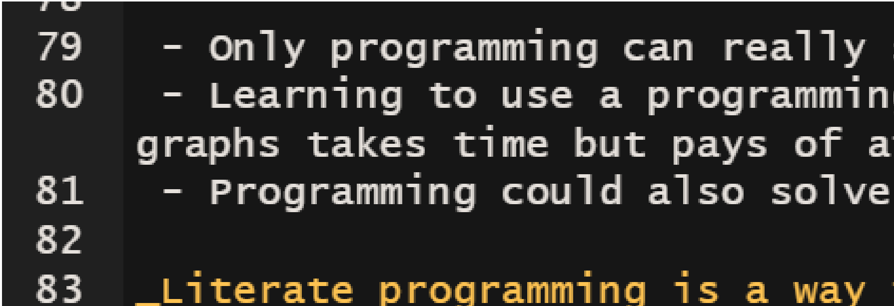
(Literate) programming is a way to connect narratives to data, methods and results



Programming is essential for Reproducible (Open) Science

- Only programming an analysis (or creation of a graph) records every step
- The script(s) function as a (data) analysis journal

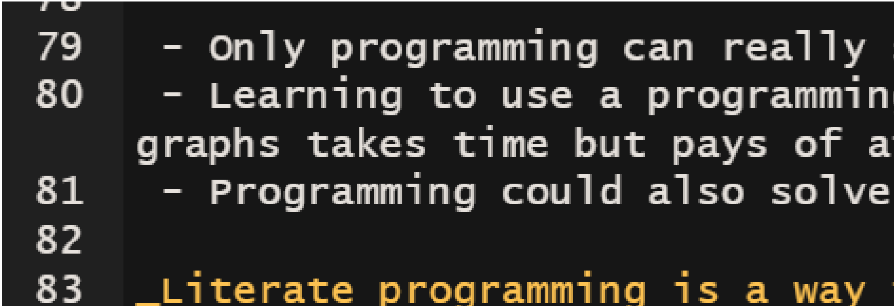
(Literate) programming is a way to connect narratives to data, methods and results



Programming is essential for Reproducible (Open) Science

- Only programming an analysis (or creation of a graph) records every step
- The script(s) function as a (data) analysis journal
- Code is the logic that connects the data and methods to conclusions

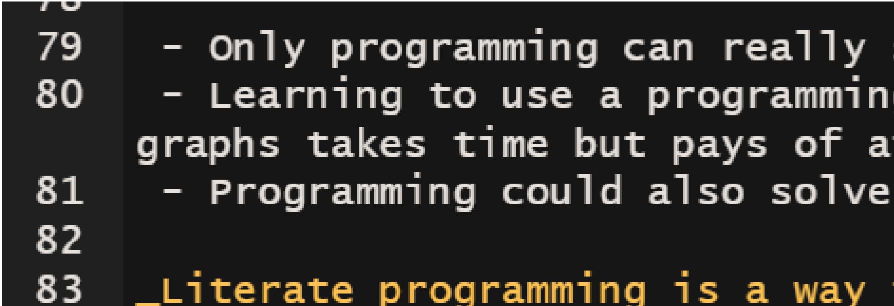
(Literate) programming is a way to connect narratives to data, methods and results



Programming is essential for Reproducible (Open) Science

- Only programming an analysis (or creation of a graph) records every step
- The script(s) function as a (data) analysis journal
- Code is the logic that connects the data and methods to conclusions
- Learning to use a programming language takes time but pays of at the long run (for all of science)

(Literate) programming is a way to connect narratives to data, methods and results



To replicate a scientific study we need at least:

- Scientific context, research questions and state of the art [P]

P = Publication, *D = Data*, *C = Code*, *OAcc = Open Access*, *OSrc = Open Source*

To replicate a scientific study we need at least:

- Scientific context, research questions and state of the art [P]
- (Experimental) model or characteristics of population or matter studied [P]

P = Publication, *D = Data*, *C = Code*, *OAcc = Open Access*, *OSrc = Open Source*

To replicate a scientific study we need at least:

- Scientific context, research questions and state of the art [P]
- (Experimental) model or characteristics of population or matter studied [P]
- Data that was generated and corresponding meta data [D, *C*]

P = Publication, *D = Data*, *C = Code*, *OAcc = Open Access*, *OSrc = Open Source*

To replicate a scientific study we need at least:

- Scientific context, research questions and state of the art [P]
- (Experimental) model or characteristics of population or matter studied [P]
- Data that was generated and corresponding meta data [D, *C*]
- **Exact** (experimental) design of the study [P, *D*, C]

$P = Publication$, $D = Data$, $C = Code$, $OAcc = Open\ Access$, $OSrc = Open\ Source$

To replicate a scientific study we need at least:

- Scientific context, research questions and state of the art [P]
- (Experimental) model or characteristics of population or matter studied [P]
- Data that was generated and corresponding meta data [D, *C*]
- **Exact** (experimental) design of the study [P, *D*, C]
- Exploratory data analysis of the data [*P*, C]

P = Publication, *D = Data*, *C = Code*, *OAcc = Open Access*, *OSrc = Open Source*

To replicate a scientific study we need at least:

- Scientific context, research questions and state of the art [P]
- (Experimental) model or characteristics of population or matter studied [P]
- Data that was generated and corresponding meta data [D, *C*]
- **Exact** (experimental) design of the study [P, *D*, C]
- Exploratory data analysis of the data [*P*, C]
- **Exact** methods that were used to conduct any formal inference [*P*, C]

P = Publication, *D = Data*, *C = Code*, *OAcc = Open Access*, *OSrc = Open Source*

To replicate a scientific study we need at least:

- Scientific context, research questions and state of the art [P]
- (Experimental) model or characteristics of population or matter studied [P]
- Data that was generated and corresponding meta data [D, *C*]
- **Exact** (experimental) design of the study [P, *D*, C]
- Exploratory data analysis of the data [*P*, C]
- **Exact** methods that were used to conduct any formal inference [*P*, C]
- Model diagnostics [*C*]

$P = Publication$, $D = Data$, $C = Code$, $OAcc = Open\ Access$, $OSrc = Open\ Source$

To replicate a scientific study we need at least:

- Scientific context, research questions and state of the art [P]
- (Experimental) model or characteristics of population or matter studied [P]
- Data that was generated and corresponding meta data [D, *C*]
- **Exact** (experimental) design of the study [P, *D*, C]
- Exploratory data analysis of the data [*P*, C]
- **Exact** methods that were used to conduct any formal inference [*P*, C]
- Model diagnostics [*C*]
- Interpretations of the (statistical) model results/model fitting process [*P*, *C*]

$P = Publication$, $D = Data$, $C = Code$, $OAcc = Open\ Access$, $OSrc = Open\ Source$

To replicate a scientific study we need at least:

- Scientific context, research questions and state of the art [P]
- (Experimental) model or characteristics of population or matter studied [P]
- Data that was generated and corresponding meta data [D, *C*]
- **Exact** (experimental) design of the study [P, *D*, C]
- Exploratory data analysis of the data [*P*, C]
- **Exact** methods that were used to conduct any formal inference [*P*, C]
- Model diagnostics [*C*]
- Interpretations of the (statistical) model results/model fitting process [*P*, *C*]
- Conclusions and academic scoping of the results [P, *C*]

$P = Publication$, $D = Data$, $C = Code$, $OAcc = Open\ Access$, $OSrc = Open\ Source$

To replicate a scientific study we need at least:

- Scientific context, research questions and state of the art [P]
- (Experimental) model or characteristics of population or matter studied [P]
- Data that was generated and corresponding meta data [D, *C*]
- **Exact** (experimental) design of the study [P, *D*, C]
- Exploratory data analysis of the data [*P*, C]
- **Exact** methods that were used to conduct any formal inference [*P*, C]
- Model diagnostics [*C*]
- Interpretations of the (statistical) model results/model fitting process [*P*, *C*]
- Conclusions and academic scoping of the results [P, *C*]
- **Access to all of the above** [OAcc, OSrc]

P = Publication, *D = Data*, *C = Code*, *OAcc = Open Access*, *OSrc = Open Source*

A short example of Reproducible (Open) Science

Assume we have the following question: "Which of 4 types of chairs takes the least effort to arise from when seated in?" We have the following setup:

- 4 different types of chairs
- 9 different subjects (probably somewhat aged)
- Each subject is required to provide a score (from 6 to 20, 6 being very lightly strenuous, 20 being extremely strenuous) when arising from each of the 4 chairs. There is some 'wash-out' time in between the trials. The chair order is randomised.

To analyze this experiment statistically, the model would need to include: the rating score as the **measured (or dependent) variable**, the type of chair as the **experimental factor** and the subject as the **blocking factor**

Mixed effects models

A typical analysis method for this type of randomized block design is a so-called ‘multi-level’ or also called ‘mixed-effects’ or ‘hierarchical’ models. An analysis method much used in clinical or biological scientific practice.

You could also use one-way ANOVA but I will illustrate why this is not a good idea

What do we minimally need, to replicate the science of this experiment?

I will show:

- the data

In the next few slides, I will hopefully convince you of the power of (literate) programming to communicate such an analysis.

Example reproduced from: Pinheiro and Bates, 2000, *Mixed-Effects Models in S and S-PLUS*, Springer, New York.

What do we minimally need, to replicate the science of this experiment?

I will show:

- the data
- an exploratory graph

In the next few slides, I will hopefully convince you of the power of (literate) programming to communicate such an analysis.

Example reproduced from: Pinheiro and Bates, 2000, *Mixed-Effects Models in S and S-PLUS*, Springer, New York.

What do we minimally need, to replicate the science of this experiment?

I will show:

- the data
- an exploratory graph
- a statistical model

In the next few slides, I will hopefully convince you of the power of (literate) programming to communicate such an analysis.

Example reproduced from: Pinheiro and Bates, 2000, *Mixed-Effects Models in S and S-PLUS*, Springer, New York.

What do we minimally need, to replicate the science of this experiment?

I will show:

- the data
- an exploratory graph
- a statistical model
- the statistical model results

In the next few slides, I will hopefully convince you of the power of (literate) programming to communicate such an analysis.

Example reproduced from: Pinheiro and Bates, 2000, *Mixed-Effects Models in S and S-PLUS*, Springer, New York.

What do we minimally need, to replicate the science of this experiment?

I will show:

- the data
- an exploratory graph
- a statistical model
- the statistical model results
- a model diagnostic

In the next few slides, I will hopefully convince you of the power of (literate) programming to communicate such an analysis.

Example reproduced from: Pinheiro and Bates, 2000, *Mixed-Effects Models in S and S-PLUS*, Springer, New York.

What do we minimally need, to replicate the science of this experiment?

I will show:

- the data
- an exploratory graph
- a statistical model
- the statistical model results
- a model diagnostic
- some conclusions

In the next few slides, I will hopefully convince you of the power of (literate) programming to communicate such an analysis.

Example reproduced from: Pinheiro and Bates, 2000, *Mixed-Effects Models in S and S-PLUS*, Springer, New York.

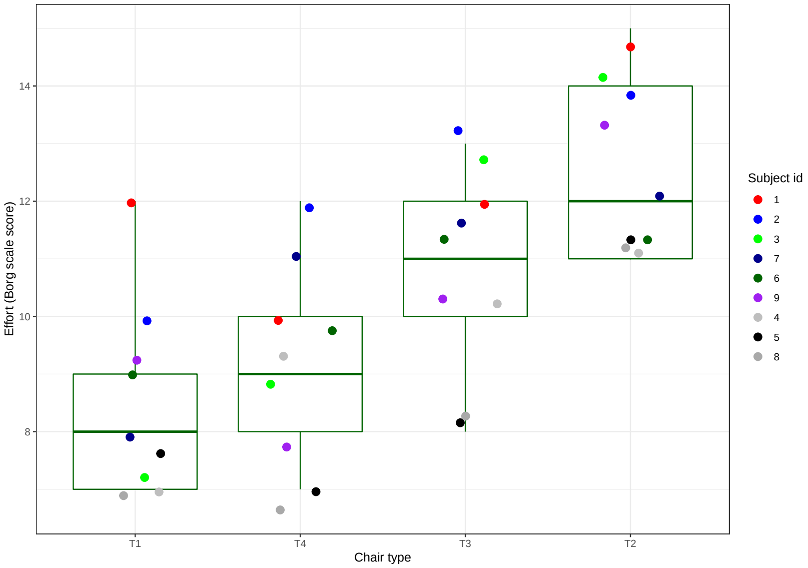
The data of the experiment

Wretenberg, Arborelius & Lindberg, 1993

```
library(nlme)
ergoStool %>% as_tibble()
```

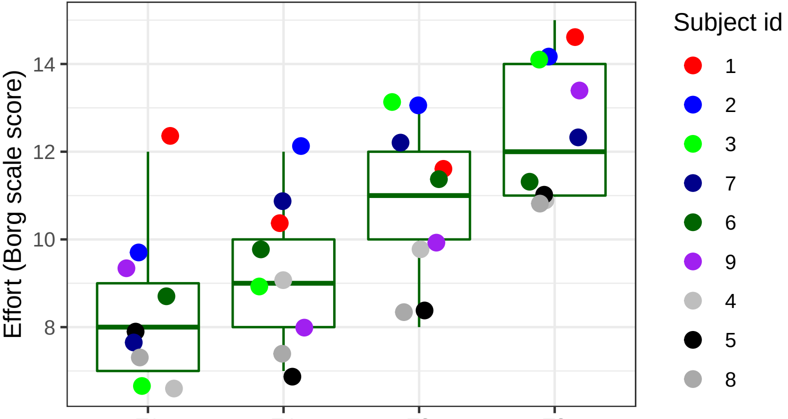
```
## # A tibble: 36 x 3
##    effort Type  Subject
##     <dbl> <fct> <ord>
##  1     12 T1    1
##  2     15 T2    1
##  3     12 T3    1
##  4     10 T4    1
##  5     10 T1    2
##  6     14 T2    2
##  7     13 T3    2
##  8     12 T4    2
##  9      7 T1    3
## 10     14 T2    3
## # ... with 26 more rows
```

An exploratory graph



Mind the variability per subject, what do you see?

- Can you say something about within-subject variability (note 'Minster Blue')?
- Can you say something about between-subject variability (note 'Mister Green', vs 'Mister Black')?
- Which chair type takes, on average the biggest effort to arise from?



The statistical questions

1. Which chair type takes, on average the biggest effort to arise from? (ANOVA / MEM, fixed effects)

- Do individual (within subject) differences play a role in appointing a average score to a chair type? (MEM, random effects)
- Does variability between subjects play a role in determining the 'best' chair type (ANOVA / MEM, confidence intervals)

The statistical model

Statistical models (in R) can be specified by a `model formula`. The left side of the formula is the dependent variable, the right side are the 'predictors'. Here we include a `fixed` and a `random` term to the model (as is common for mixed-effects models)

```
library(nlme)
```

```
ergo_model <- lme(
  data = ergoStool, # the data to be used for the model
  fixed = effort ~ Type, # the dependent and fixed effects
  random = ~1 | Subject # random intercepts for Subject va
)
```

The `lme()` function is part of the `{nlme}` package for mixed effects modelling in R

Example reproduced from: Pinheiro and Bates, 2000, *Mixed-Effects Models in S and S-PLUS*, Springer, New York.

The statistical results

	Value	Std.Error	DF	t-value	p-value
(Intercept)	8.5555556	0.5760123	24	14.853079	0.0000000
TypeT2	3.8888889	0.5186838	24	7.497610	0.0000001
TypeT3	2.2222222	0.5186838	24	4.284348	0.0002563
TypeT4	0.6666667	0.5186838	24	1.285305	0.2109512

Model diagnostics

- Diagnostics of a fitted model is the most important step in a statistical analysis

A residual plot shows the 'residual' error ('unexplained variance') after fitting the model. Under the Normality assumption standardized residuals should:

Model diagnostics

- Diagnostics of a fitted model is the most important step in a statistical analysis
- In most scientific papers the details are lacking

A residual plot shows the 'residual' error ('unexplained variance') after fitting the model. Under the Normality assumption standardized residuals should:

Model diagnostics

- Diagnostics of a fitted model is the most important step in a statistical analysis
- In most scientific papers the details are lacking
- Did the authors omit to perform this step? Or did they not report it?

A residual plot shows the ‘residual’ error (‘unexplained variance’) after fitting the model. Under the Normality assumption standardized residuals should:

Model diagnostics

- Diagnostics of a fitted model is the most important step in a statistical analysis
- In most scientific papers the details are lacking
- Did the authors omit to perform this step? Or did they not report it?
- If you do not want to include it in your paper, put it in an appendix!

A residual plot shows the 'residual' error ('unexplained variance') after fitting the model. Under the Normality assumption standardized residuals should:

Model diagnostics

- Diagnostics of a fitted model is the most important step in a statistical analysis
- In most scientific papers the details are lacking
- Did the authors omit to perform this step? Or did they not report it?
- If you do not want to include it in your paper, put it in an appendix!

A residual plot shows the 'residual' error ('unexplained variance') after fitting the model. Under the Normality assumption standardized residuals should:

1. Be normally distributed around 0

Model diagnostics

- Diagnostics of a fitted model is the most important step in a statistical analysis
- In most scientific papers the details are lacking
- Did the authors omit to perform this step? Or did they not report it?
- If you do not want to include it in your paper, put it in an appendix!

A residual plot shows the 'residual' error ('unexplained variance') after fitting the model. Under the Normality assumption standardized residuals should:

1. Be normally distributed around 0
2. Display no obvious 'patters'

Model diagnostics

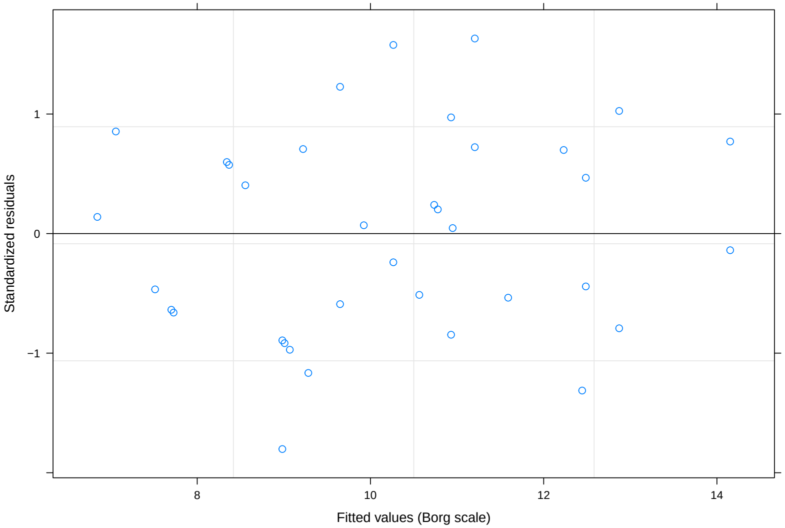
- Diagnostics of a fitted model is the most important step in a statistical analysis
- In most scientific papers the details are lacking
- Did the authors omit to perform this step? Or did they not report it?
- If you do not want to include it in your paper, put it in an appendix!

A residual plot shows the ‘residual’ error (‘unexplained variance’) after fitting the model. Under the Normality assumption standardized residuals should:

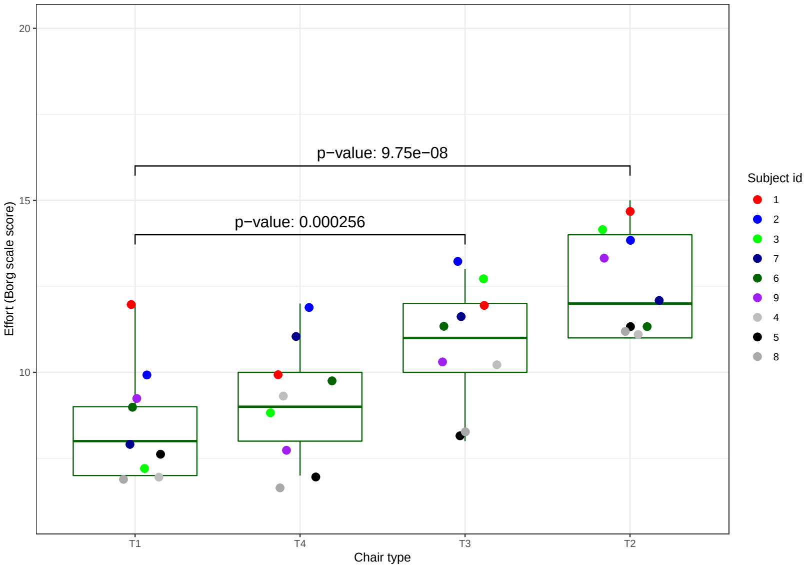
1. Be normally distributed around 0
2. Display no obvious ‘patters’
3. Should display overall equal ‘spread’ above and below 0 (‘assumption of equal variance’)

Residual plot

```
plot(ergo_model) ## type = 'pearson' (standardized residual
```



The conclusions in a plot

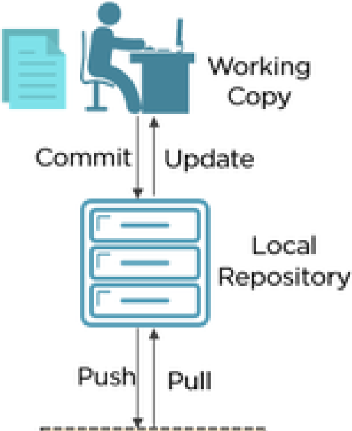


And the most important part. . .

odz: *Practice what you preach*

If you want to reproduce, add-on, falsify or apply your own ideas to this example, you can find the code (and data) in Github.com

In webinar 3, I will show you how to actually run, use and organize code like this!



Thank you for your attention!

Example; The Open Science Framework OSF

OSF - Reproducible Project: Psychology

- 100 publications in Psychology journals

$$RP : Psychology = P + D + C + OSrc\ (+OAcc)$$

$P = Publication$, $D = Data$, $C = Code$, $OAcc = Open\ Access$, $OSrc = Open\ Source$

OSF - Reproducible Project: Psychology

- 100 publications in Psychology journals
- Results from half of these publications could be reproduced

$$RP : Psychology = P + D + C + OSrc \ (+OAcc)$$

$P = Publication$, $D = Data$, $C = Code$, $OAcc = Open\ Access$, $OSrc = Open\ Source$

OSF - Reproducible Project: Psychology

- 100 publications in Psychology journals
- Results from half of these publications could be reproduced
- Full access to P, D and C in OSF

$$RP : Psychology = P + D + C + OSrc\ (+OAcc)$$

$P = Publication$, $D = Data$, $C = Code$, $OAcc = Open\ Access$, $OSrc = Open\ Source$

OSF - Reproducible Project: Psychology

- 100 publications in Psychology journals
- Results from half of these publications could be reproduced
- Full access to P, D and C in OSF
- The publication is not published in an OAcc journal but:

$$RP : Psychology = P + D + C + OSrc \ (+OAcc)$$

$P = Publication$, $D = Data$, $C = Code$, $OAcc = Open\ Access$, $OSrc = Open\ Source$

OSF - Reproducible Project: Psychology

- 100 publications in Psychology journals
- Results from half of these publications could be reproduced
- Full access to P, D and C in OSF
- The publication is not published in an OAcc journal but:
- The submitted manuscript is available in OSF

$$RP : Psychology = P + D + C + OSrc\ (+OAcc)$$

$P = Publication$, $D = Data$, $C = Code$, $OAcc = Open\ Access$, $OSrc = Open\ Source$

OSF - Reproducible Project: Psychology

- 100 publications in Psychology journals
- Results from half of these publications could be reproduced
- Full access to P, D and C in OSF
- The publication is not published in an OAcc journal but:
- The submitted manuscript is available in OSF
- The R code used is available in OSF

$RP : Psychology = P + D + C + OSrc\ (+OAcc)$

$P = Publication$, $D = Data$, $C = Code$, $OAcc = Open\ Access$, $OSrc = Open\ Source$