CoEDL Summer School 2019
Advanced Statistics for Linguists
(coedlstatzr)

Martin Schweinberger
<www.martinschweinberger.de>
m.schweinberger@uq.edu.au

2019 12 04

CoEDL Summer School 2019 - Advanced Statistics for Linguists (coedlstatzr)

Before we begin

- All you will ever need - for this work shop - is in the folder I sent.
- If you have not received my email type the following into your browser to automatically download that folder https://martinschweinberger.de/docs/materials/AdvancedStatzForLinguists.zip
- Unzip wherever you please and open it!
- All code and more elaborate explainations of what we will cover is available at the website of the *Language Technology and Data Analysis Laboratory* (LADAL) https://slcladal.github.io/index.html

LADAL is hosted by the *School of Languages and Culters* of The University of Queensland, Australia (UQ)

About this Course

What will we cover?

- Simple linear regression
- Fixed-effects regression (linear | logistic)
- Mixed-effects regression (linear | logistic | quasi-poisson)
- Tree-based models (Conditional Inference Trees | Random Forests | Boruta)

Aims

- Understand these methods
- Use these methods
- Being aware of their advantages|disadvantages|problems|issues

About this Course

Why is this course relevant for researchers that already know statistics?

- Best Practices emerge only with time
- Different people know different things (I have never not learned anything when I attended a lecture about sth I already "knew")
- Tips and tricks about model fitting and model diagnostics
- Adding and sharing to this course (*please* let us know if you have tips, tricks, or experience with sth: we are *all* here to learn!)

About this Course

What this course is *not*

- This is not an introduction to statistics
- This is not an introduction to R

What will we *not* cover?

- Basic concepts (probability, significance, etc.)
- Yes, everything is done in R but we cannot go into how R works
- Technical trouble shooting (cry for help and the assitants will come and assits in crying)
- The mathematical underpinning of the models (unless absolutely neccessary)

Timeline

Session 1 (Thursday 10:00 to 11:30)

- Introduction and set up
- Simple linear and multiple fixed-effects regression

Session 2 (Thursday 9:00 to 10:30)

- More multiple fixed-effects regression and start with mixed-effects regression

Session 3 (Friday 11:00 to 12:30)

- Mixed-effects regression

Session 4 (Friday 11:00 to 12:30)

- Tree-based models
- Wrap-up and goodbye

Why R?

Good reasons for using R

- Free open-source software
- Fuly-fledged programming environment
- Enables and enhances full reproducibility | replicability of your research (enables Best Practices)
- Can be used for data science | management | processing | visualization | analytics | presentation
- Massive and friendly support-infrastructure

Recommendations

Things that I wish I had done | known earlier

- Use R projects (Rproj)
- Use tidyverse (yes, i was brought up with base R and still haven't fully adapted)
- Create a GitHub and/or GitLab account and connect R to Git (version control, forking, cloud storage)
- You can use R to create websites (LADAL), apps (Shiny), slides (like these), publications (Rpub)
- You can do NLP, data management, data visualization, data analytics all in R
- R allows geo-spatial visualizations (maps)

What will come next?

Trends that - I believe | predict - will become more frequent in the future

- Mixed-models
- Bayesian mixed-models (problem with frequentist approach: we evaluate the probabilty of H_1 via the H_0 rather than directly)
- Interactive apps (Shiny for public outreach | schools: to allow students to discover language and make things about language more well known)
- Replication, Open Data | Science, collaborative research (hopefully)
- Entering new fields (e.g. History, Cultural and Literary Studies)

Where from here?

Books about statistics that I can recommend (for beginners)

- Field, Miles, and Field (2012), Levshina (2015), Gries (2009), Agresti (1996)

Books about statistics that I can recommend (for advanced)

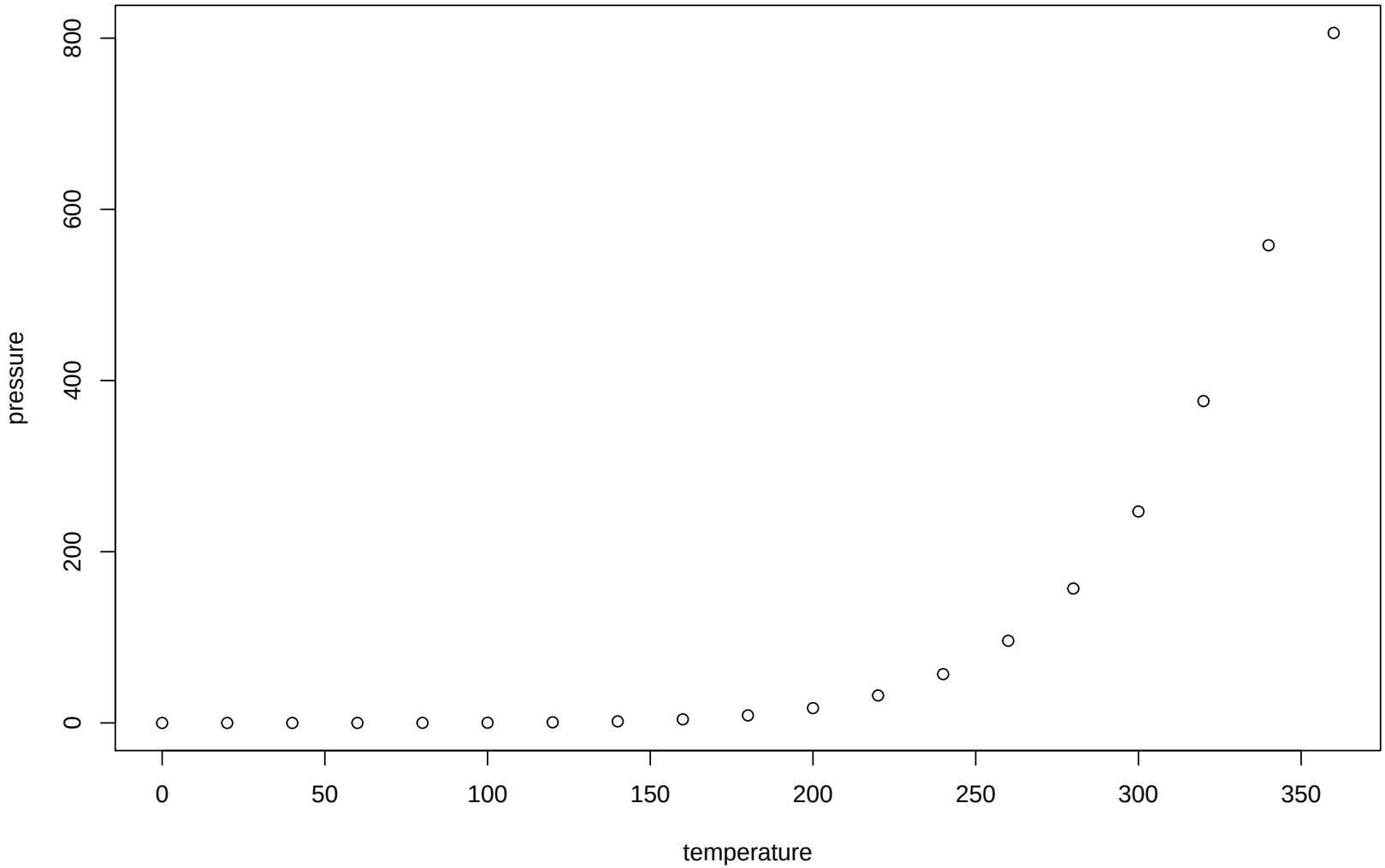
- Baayen (2008), Agresti and Kateri (2011), Pinheiro and Bates (2000), Zuur et al. (2009)

Slide with R Output

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Slide with Plot



References

Agresti, Alan. 1996. *An Introduction to Categorical Data Analysis*. Hoboken, NJ: JohnWiley & Sons.

Agresti, Alan, and Maria Kateri. 2011. *Categorical Data Analysis*. Springer.

Baayen, R Harald. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University press.

Field, Andy, Jeremy Miles, and Zoe Field. 2012. *Discovering Statistics Using R*. Sage.

Gries, Stefan Th. 2009. *Statistics for Linguistics Using R: A Practical Introduction*. Berlin & New York: Mouton de Gruyter.

Levshina, Natalia. 2015. *How to Do Linguistics with R: Data Exploration and Statistical Analysis*. Amsterdam: John Benjamins Publishing Company.

Pinheiro, Jose C., and Douglas M. Bates. 2000. *Mixed-Effects*