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

Martin Schweinberger <www.martinschweinberger.de>

contact: m.schweinberger@uq.edu.au

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

Before we start, download the zip-file "AdvancedStatzForLinguists", unzip wherever you please and open it (all you will ever need - for this work shop - is in that folder)!

You can automatically download the zipped folder from https://martinschweinberger.de/docs/materials/AdvancedStatzForLinguists.zip.

All code and more elaborate explainations of what we will cover is available at the LADAL website (Language Technology and Data Analysis Laboratory; https://slcladal.github.io/index.html) 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

What this course is **NOT**

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

What will we **NOT** cover?

- Basic concepts (probability, significance, etc.)
- Yes, everything ww will do will be 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?

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

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