R-course
regression and lmer()
2017

Justin Sulik

Louis ten Bosch

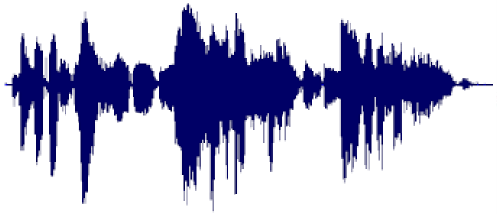
Today’s learning outcomes

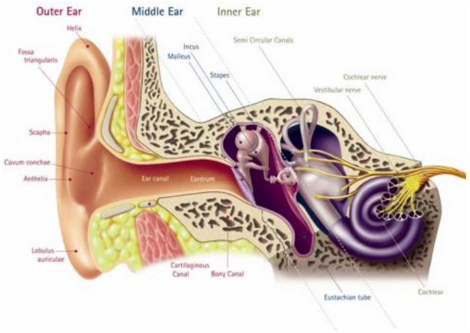
- Understand some of the larger issues involved in doing stats
  - (things to think about before you even start trying to build a model)
- Understand that there is no magic algorithm you can follow to create the perfect model
  - (things to think about while building models)

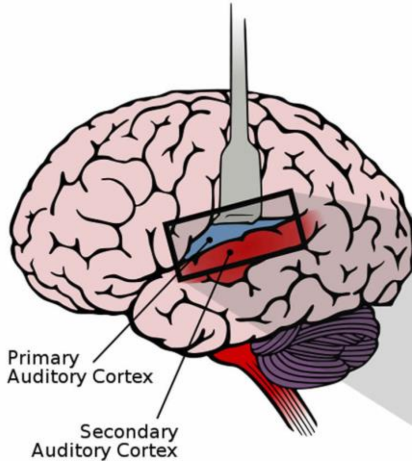
Some abbreviations you'll see

- LM – Linear model
- LMEMs – Linear mixed-effects models
- LMERs – Linear mixed-effects regressions
- G... – generalized ...
- LME4 – an R package for doing these
- lmer() – an R function from the above package

experiment.... example



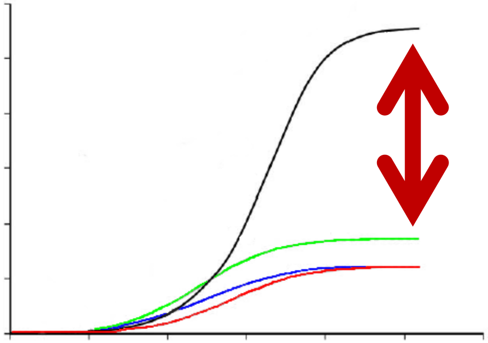




word activation

decision

execution



reaction times,
judgments, cortical
activations
parameters
(stimulus, participants,
experiment-based)

reaction times,
judgments, cortical
activations
parameters
(stimulus, participants,
experiment-based)

Is there any structure in this set?

reaction times,
judgments, cortical
activations
parameters
(stimulus, participants,
experiment-based)

What can we learn from this data set?

How to find structure in this data set?

Options:

- formulate hypotheses, and check them against the data ——— P(data | model)
- find structure in data by deep learning; the resulting model encapsulates the knowledge in the data

reaction times,
judgments, cortical
activations
parameters
(stimulus, participants,
experiment-based)

What can we learn from this data set?

How to find structure in this data set?

Options:

- formulate hypotheses, and check them against the data ——— P(data | model) confirmatory
- find structure in data by deep learning; the resulting model encapsulates the knowledge in the data exploratory

reaction times,
judgments, cortical
activations
parameters
(stimulus, participants,
experiment-based)

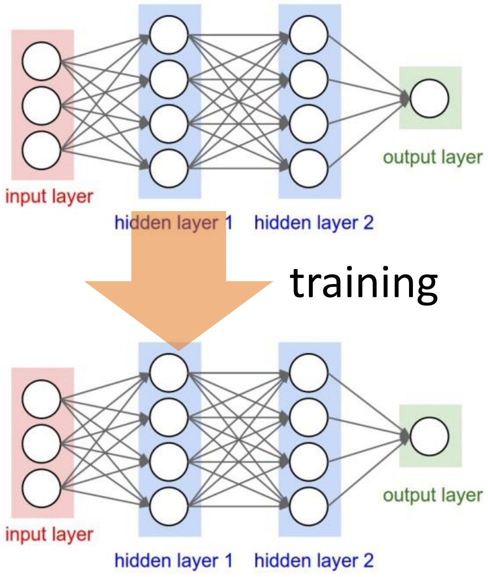
one of the well-known methods: regression modeling

$$\begin{gathered}\log RT = P_{stim,1} + P_{stim,2} + \ldots + P_{stim,N} + \\ P_{subj,1} + P_{subj,2} + \ldots + P_{subj,M} + \\ linear\ interactions + \\ nonlinear\ terms + \\ nonlinear\ interactions + \\ noise + P_{previous\ decisions}\end{gathered}$$

reaction times,
judgments, cortical
activations
parameters
(stimulus, participants,
experiment-based)

another method (fashionable): deep neural networks

input parameters



training

output parameters

knowledge about data structure is in the model parameters

Issues

- What is a good (statistical) model?
- How to find a good model?
- Overfitting, underfitting
- Is my N large enough?
- Structure of the model: fixed versus random effects

The program (2017)

- **Mon Oct 9, 2-4pm**: week 1 (Louis)
  - aim, introduction (this file)
  - p-values, transformations, H0-H1, exploratory versus confirmatory, model selection (R^2, AIC, BIC, ...)
- **Thu Oct 19, 2-4pm**: week 2
  - fixed effects, random effects, slopes (Justin)
  - AIC, BIC, anova(), practice on artificial data set (will be made available) (Louis)
- **Wed Oct 25, 3-5pm**: week 3
  - releveling, centering, contrast coding (Louis)
  - plots and visualization (easy) (Justin)
- **Fri Nov 3, 3-5pm**: week 4 (Louis)
  - prediction, generalization
  - overcoming convergence problems
- **Wed Nov 8, 3-5pm**: week 5 (Justin)
  - glmer
  - bootstrapping
  - visualization, plotting and interaction (more advanced)
  - how to publish (sweave)

Aim of this course

- to provide background in regression and especially lmer(), glmer()
- increase awareness of pitfalls

- there is no absolute truth here
  - statistics must robustly support your narrative in a paper

- “**Lies, damned lies, and statistics**“
  - “phrase describing the persuasive power of numbers, particularly the use of statistics to bolster weak arguments. It is also sometimes colloquially used to doubt statistics used to prove an opponent's point”.

Your background - rondje

- name, group
- experience with R
- experience with lmer, glmer
- what kind of experiments

Statistics is not really simple

- Statistics is often used in a sloppy way
  - "Group A improved significantly during the training ($p < 0.01$), whereas group B did not improve ($p = 0.2$)".
  - frequent fallacy, also occurs in journal articles

- Instead of modelling separate data sets, combine data sets from both groups, and check significance of the interaction
  - group * learning_rate

- See also: Sander Nieuwenhuis, Birte Forstmann & Eric-Jan Wagenmakers (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. Nature Neuroscience.
- Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn (2011) False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant, Psychological Science 22(11) 1359–1366

Causes of statistical misinterpretations

- Domain knowledge but not statistical knowledge
  - Or reverse.
- The questions are ill-posed, or poorly defined
  - what is the size of the lexicon of 4y old infants?
- Poor data quality
  - many outliers
  - missing data (e.g. after combining different data sets)
  - distortions in a sample (e.g. interviewing by telephone)
  - verbal data descriptors (hot - warm - cold)
  - unknown variation in data
  - unknown coverage of variation

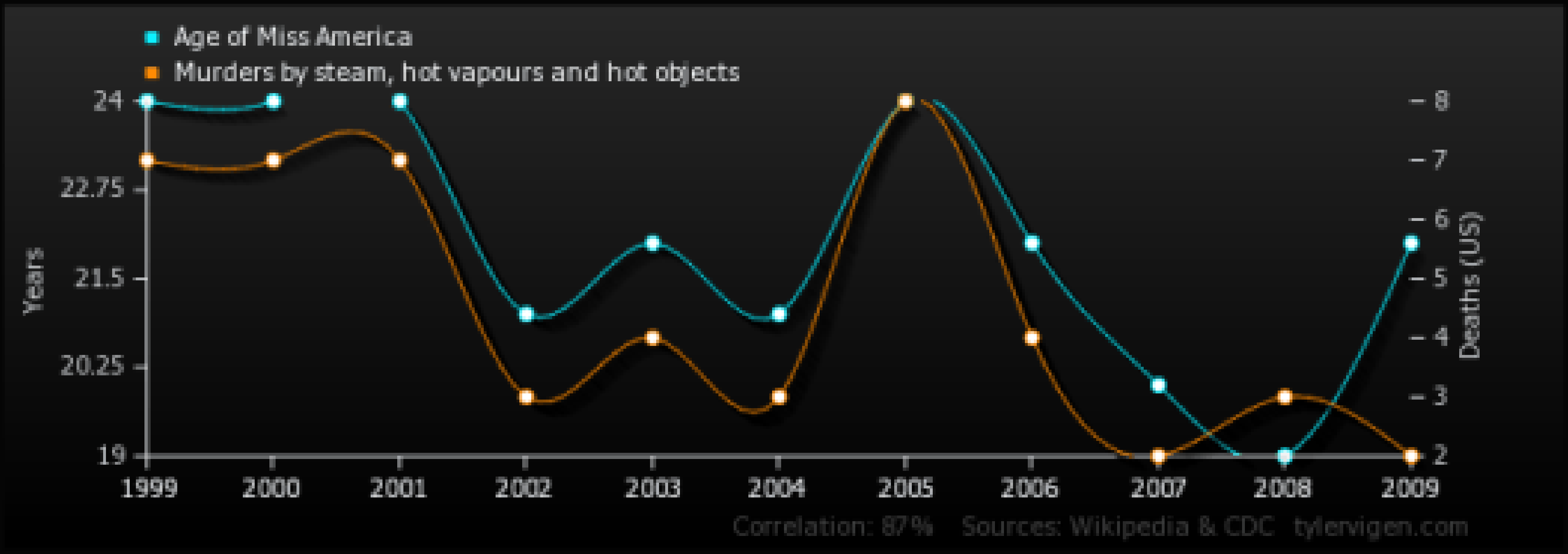
Correlation does not imply causation

When a statistical test shows a correlation between A and B, there are the following possibilities:

- Coincidence (A does not cause B, neither the reverse).
- A (partly) causes B, or reverse.
- A and B both (partly) cause each other.
- A and B are both (partly) caused by a third factor, C.
- B is (partly) caused by C which is correlated to A (or reverse).

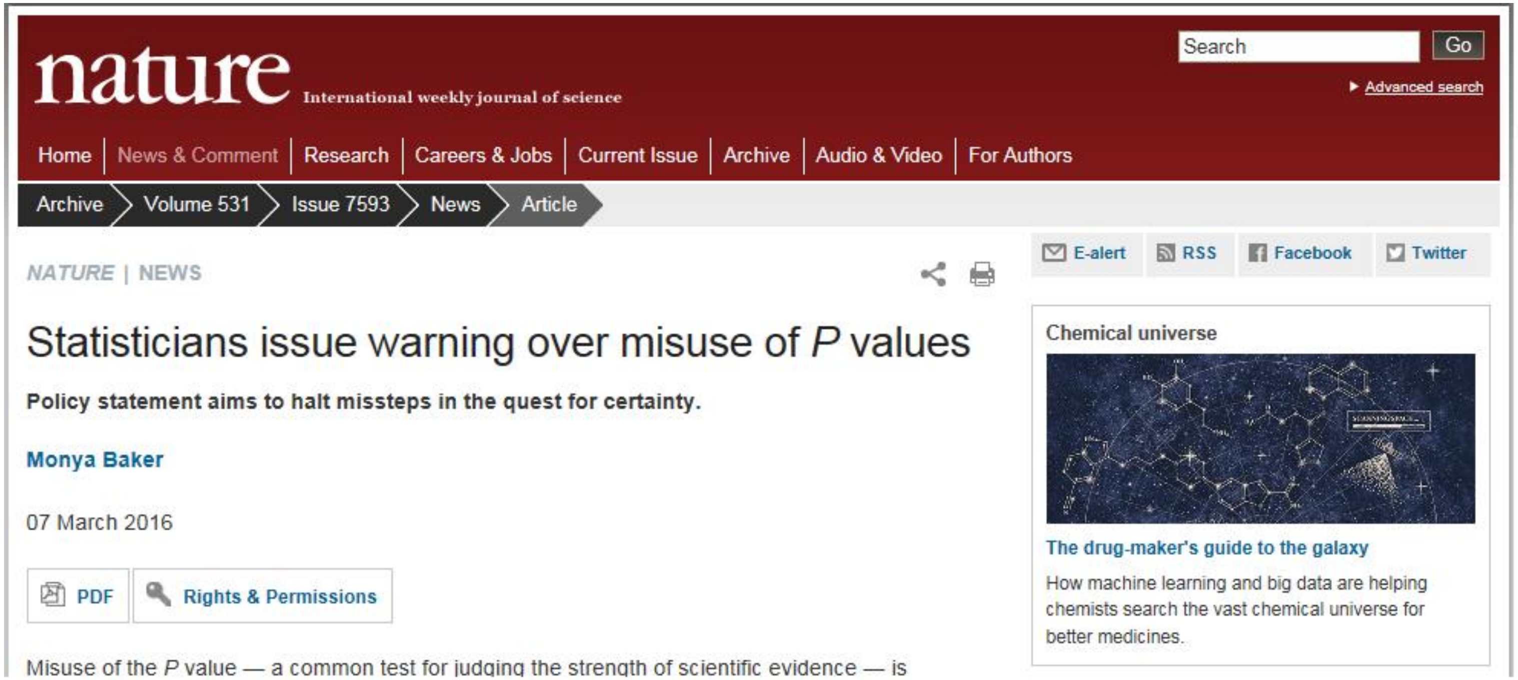
The effect of false causality can be reduced by e.g. splitting groups ("treatment group“, "control group“)

Spurious correlations, Tyler Vigen (2015)



Interpreting p-values

- The p-value does not relate to the probability “an effect being present” in the data
  - but to the probability of ‘rejecting the absence of an effect’ across different data samples, under very specific assumptions about data distributions



NATURE | NEWS

Statisticians issue warning over misuse of *P* values

Policy statement aims to halt missteps in the quest for certainty.

Monya Baker

07 March 2016

Misuse of the *P* value — a common test for judging the strength of scientific evidence — is

Chemical universe

The drug-maker's guide to the galaxy

How machine learning and big data are helping chemists search the vast chemical universe for better medicines.

Many recent papers address the p-value issue

- *p*-value might provide the right answer to the wrong question.
- What we really want to know is *not* the probability of the observations given a hypothesis about the existence of a real effect, but rather the probability that there *is* a real effect given the observations.
- The dichotomy between ‘significant’ and ‘not significant’ is absurd.
  - There’s obviously very little difference between the implication of a *p*-value of 4.7 per cent and of 5.3 per cent, yet the former is regarded as success and the latter as failure.

Interpreting p-values

One of the problems for obtaining good p-values is often the low sample size N.

- However, a low N in itself is not necessarily a problem.
- And: a large N won't solve everything.

Better p-values can be obtained in different ways:

- by better designs:
  - repeated measures (e.g. keep the participant constant across many measurements)
- by using meaningful transformations of your data
  - e.g. in case of continuous data: log(x), -1/x, ...

Interpreting p-values (cont'd)

- or better analysis methods:
  - in case of discrete data (judgements, yes/no): logistic regression
  - by including random factors (week 2)
  - contrast coding (in case a discrete variable has three or more levels)

For tricks and pitfalls with p-values see: Joseph P. Simmons, Leif D. Nelson & Uri Simonsohn (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant (Psychological Science).

- ***Requirements for authors and reviewers (!)***

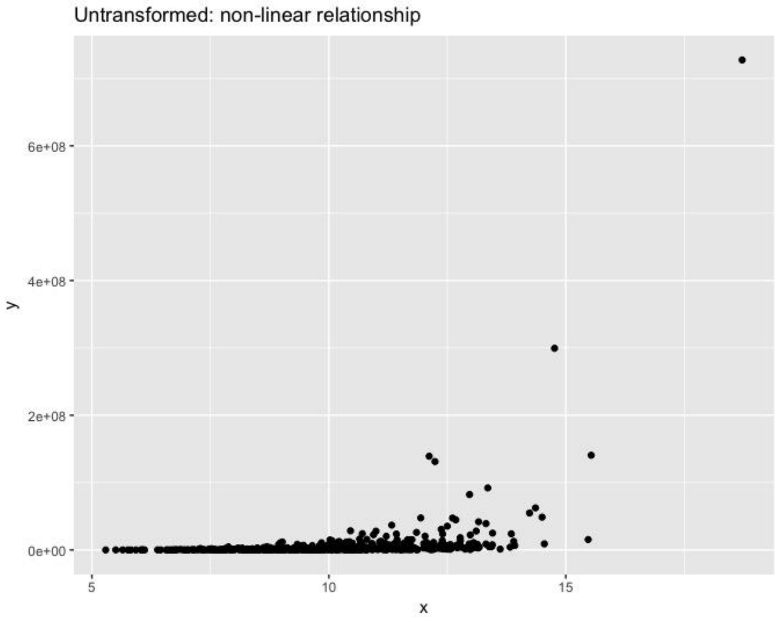
Representation of the problem

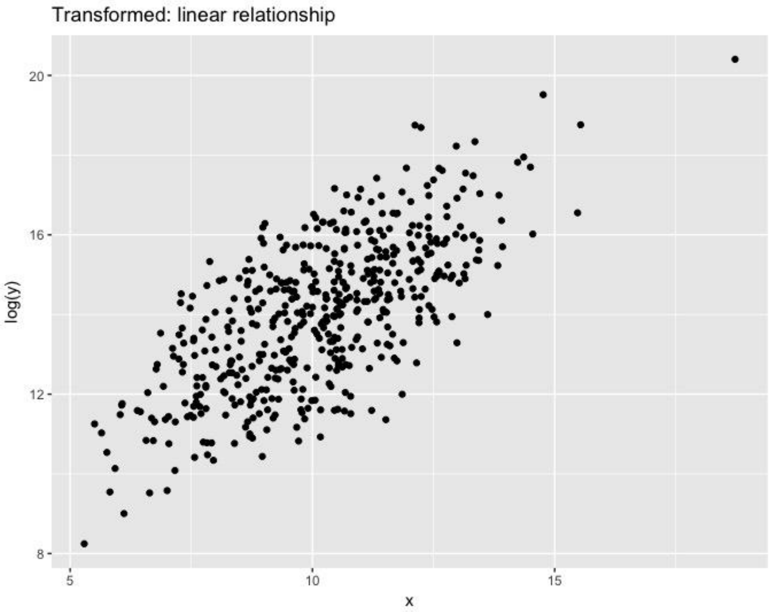
- In general: keep continuous data continuous
  - age -> (young, old)
  - binning may artificially improve p-values
  - binning may help to reduce detrimental effect of outliers

- Transform continuous measurements
  - e.g. log(RT) or 1/RT instead of linear RT
  - Weber’s law: The relationship between stimulus and perception is logarithmic
  - pros and cons, especially if you transform the dependent variable

Representation of the problem

- Transform continuous measurements
  - e.g. log(RT) or 1/RT instead of linear RT
    - Weber's law: The relationship between stimulus and perception is logarithmic
  - pros and cons, especially if you transform the dependent variable
  - In week 5, we'll be looking at alternatives to transformation (GLMERs)





Hypothesis testing

- Hypothesis testing is formulated in terms of two hypotheses:
- H0: the null hypothesis
- H1: the alternate hypothesis

Hypothesis testing

- The hypothesis we want to test is if H1 is true.
- So, there are two possible outcomes:
  - Reject H0 and accept H1 because of sufficient evidence in the sample in favor or H1
  - Do not reject H0 because of insufficient evidence to support H1.

- Failure to reject H0 does not mean the H0 is true.

Significance testing vs Bayesian factors

- This is about **Rejecting a null hypotheses** by a significance test versus **Collecting evidence in favor of** the null hypothesis

- Using p-value testing we can only **accept** the alternative hypothesis (if $p<0.05$) or **not reject** the null hypothesis (if $p>0.05$). You do not **prove** the null hypothesis, or collect evidence for an hypothesis.

- Alternative: Bayesian modelling
  - log likelihoods, Akaike Information Criterion (AIC), BIC,...

- For a proposal, see: Eric-Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, Han van der Maas and Rogier Kievit (2012). An agenda for purely confirmatory research. Perspectives on Psychological Science.

Replication?

- Replication “crisis” attributed to several factors:
  - publication bias
  - poor use (bad use) of data and of statistical models, poor conclusions
  - too much explorative model optimization

- Suggested solutions:
  - preregistration of the model, methods, variables
  - triple blind set-up

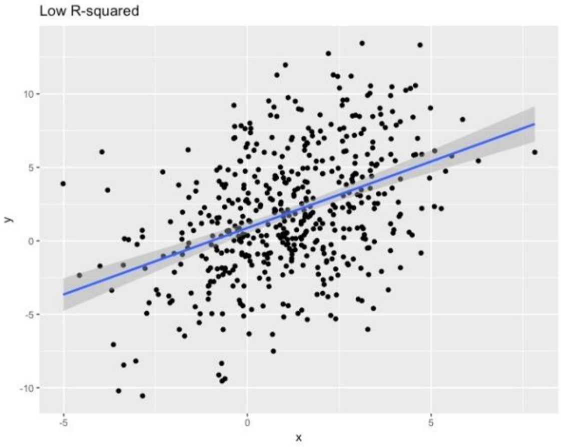
Exploratory versus confirmatory

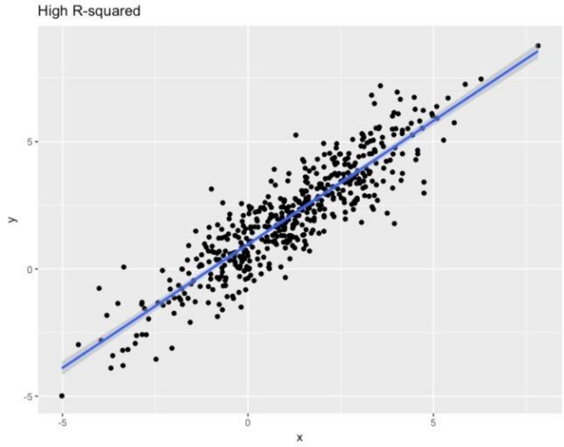
- From “Random effects structure for confirmatory hypothesis testing: Keep it maximal”, Barr et al. 2013, JML, page 277:

> “In this paper we have focused largely on confirmatory analyses. **We hope this emphasis will not be construed as an endorsement of confirmatory over exploratory approaches.**
> Exploratory analysis is an important part of the cycle of research, without which there would be few ideas to confirm in the first place.
> We do not wish to discourage people from exploiting all the many new and exciting opportunities for data analysis that LMEMs offer (see Baayen, 2008 for an excellent overview). ...”

What is a good model?

- Criteria
  - modeling of the data: R^2, log likelihood *log(P(data|model))*
  - ease of interpretation, given a theory; theoretical elegance; plausibility w.r.t. cognitive process
  - parsimony: AIC, BIC
  - generalizability (validity on unseen new data)





What is a good model?

- How to search? in practice:
  - start with database with “enough relevant variation” (not just large N)
  - start with defendable moderately complex model
  - too simple models might incorrectly assign significance
  - too complex models may distribute significance among many predictors & might fail to converge
  - theory or intuition might guide toward model simplification

How to start...

- do an exploratory **data** analysis, e.g.
- look at ranges, frequency distribution
  - outliers
  - systematic measurement errors
- check for normality
  - tests: shapiro wilk, kolmogorov-smirnov, anderson darling
- **use plots**

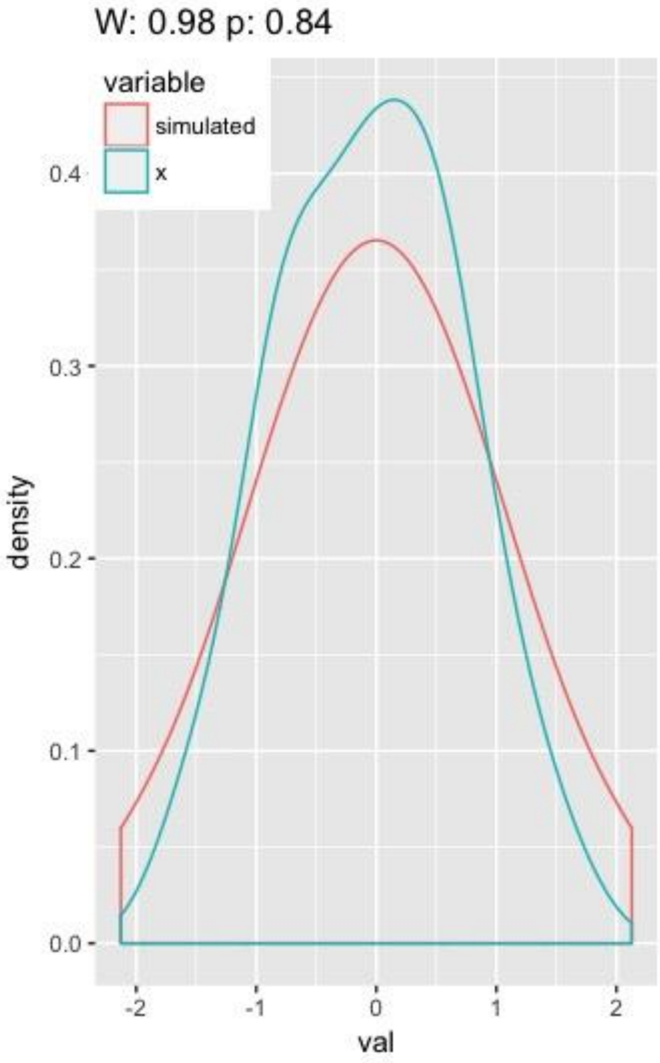
Normality test

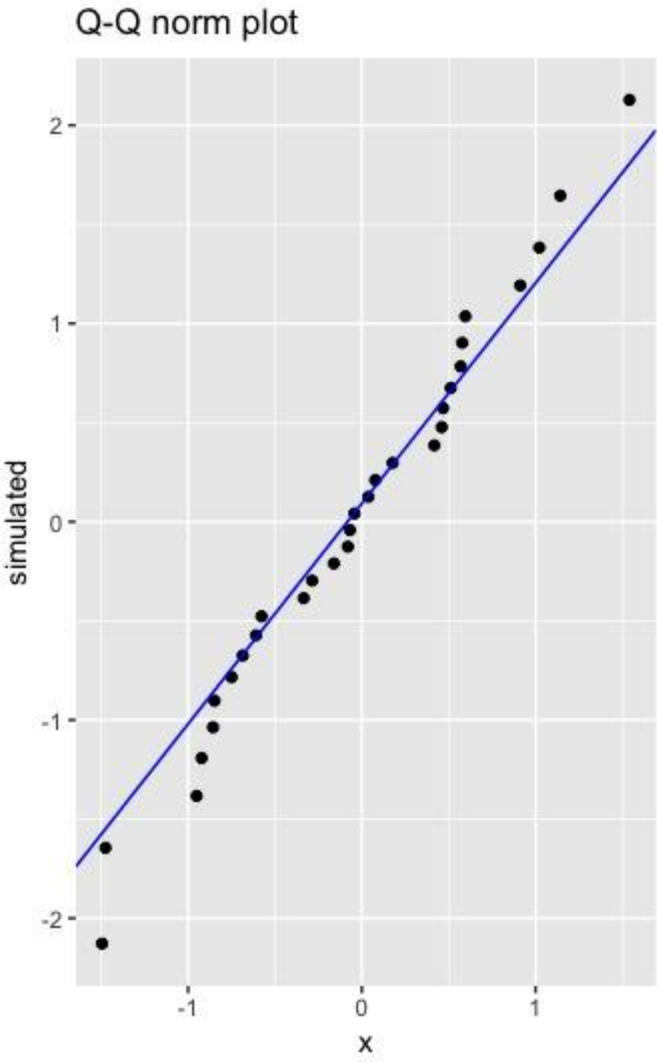
- Normality test is a null hypothesis tests **against** the assumption of normality. It does do not **prove** normality.
  - When N small, even big departures from normality may not be detected
  - When N large, a small deviation from normality may lead to a rejected H_0

```
set.seed(100)
x <- rbinom(15,5,.6)
shapiro.test(x)
W = 0.8816, p-value = 0.0502
x <- rlnorm(20,0,.4)
shapiro.test(x)
W = 0.9405, p-value = 0.2453
```

- So, in both these cases (binomial and lognormal variates) the p-value is causing a **failure to reject** the null hypothesis (that the data are normal).
- Failure to reject is not the same thing as accepting/proving.

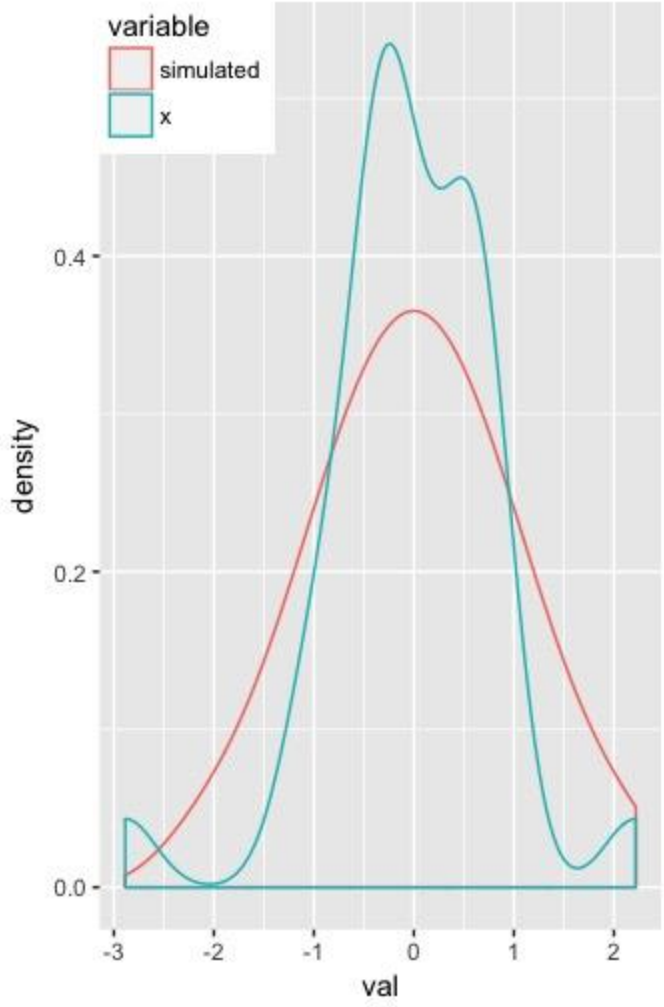
Normality plots



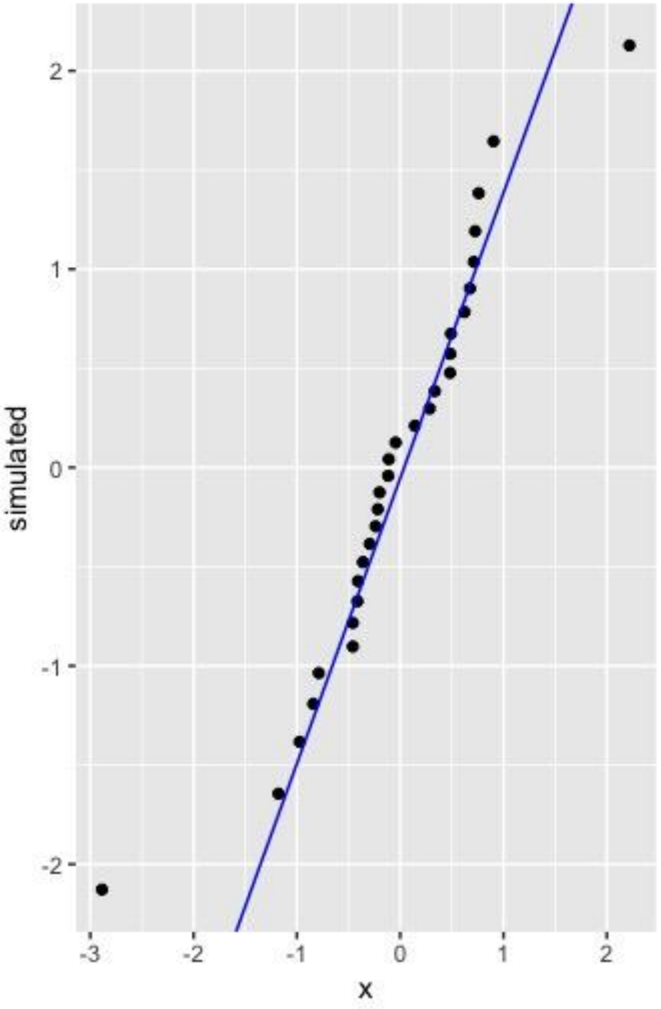


Normality plots

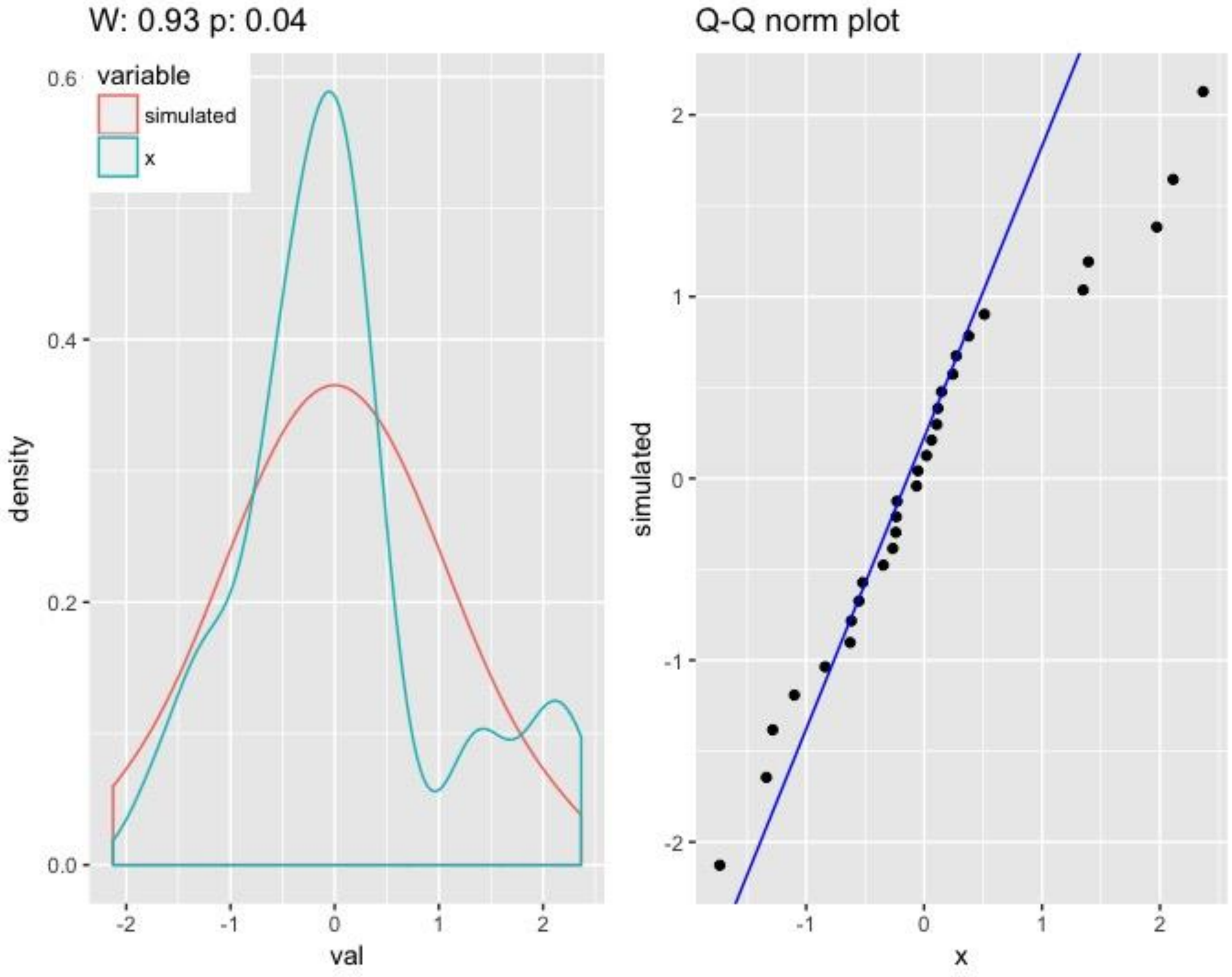
W: 0.92 p: 0.02



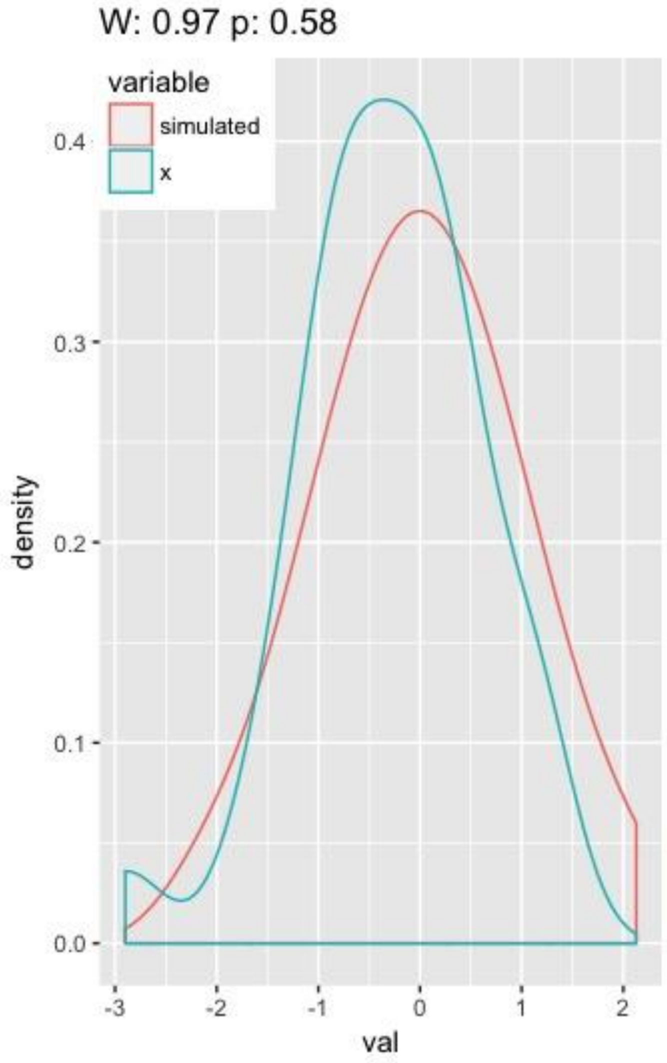
Q-Q norm plot

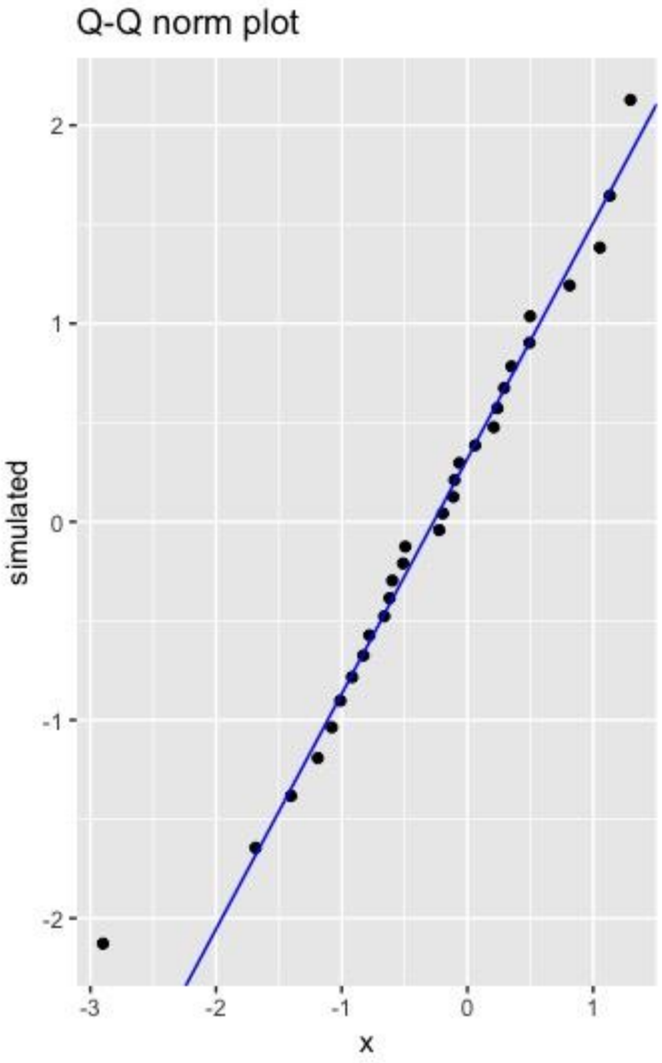


Normality plots



Normality plots





lm()

- lm() fits linear model (linear regression)
- Basic form of a formula in R:

```
Y ~ X1 + X2 + ... X3*X4 + ... + X5:X6 + ...
```

- * denotes all main effects and interactions
- : denotes only the interaction
- ^n higher powers (squares, ...)
- - removes a specified term
- I() brackets the portions of a formula where operators are used mathematically

lm(), no random effects yet

- Example of a regression formula in R

```
Y ~ X1 + X2 + ... X3*X4 + ... + X5:X6
```

- ... mathematically, this is:

$$Y = \alpha + \beta_1 X1 + \beta_2 X2 + \ldots \beta_{..} X3 + \beta_{..} X4 + \beta_{..} X3 * X4 \ldots + \beta_{..} X5\ X6$$

glm()

- example: glm(formula, family = binomial, data)
- The family argument species the error distribution and link function. See ?family for more information
  - binomial(link = "logit")
  - gaussian(link = "identity")
  - poisson(link = "log")

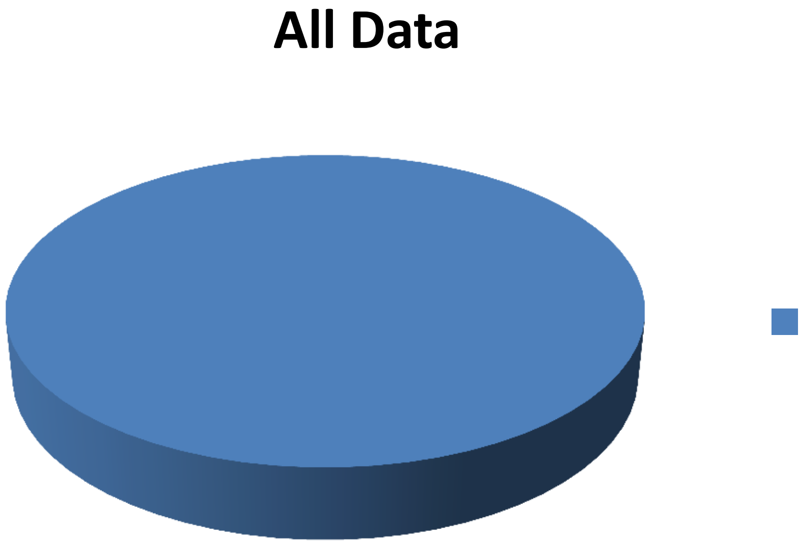
types of variables, scales

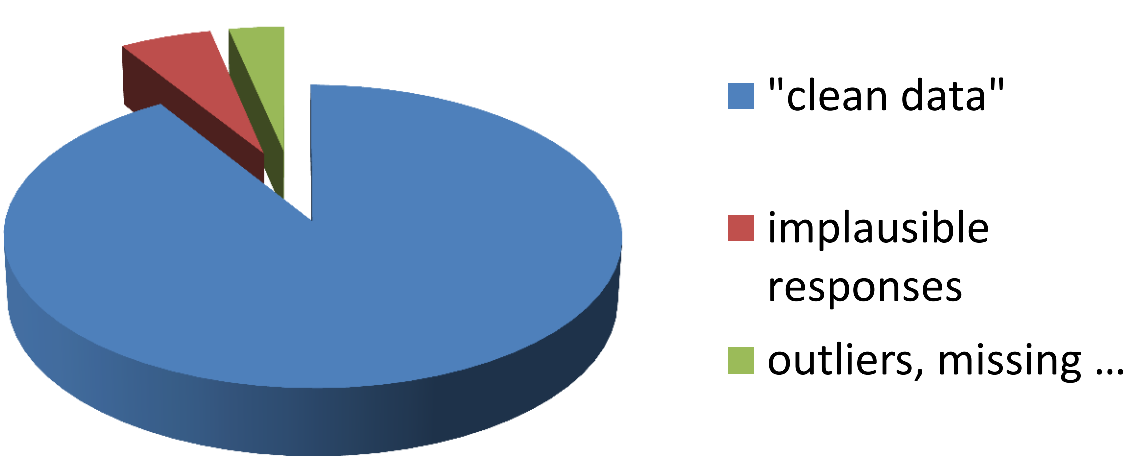
binary	0/1
nominal	amsterdam, rotterdam, nijmegen
ordinal	cold, tepid, warm, hot
interval	degree C
ratio	weight

Further reading

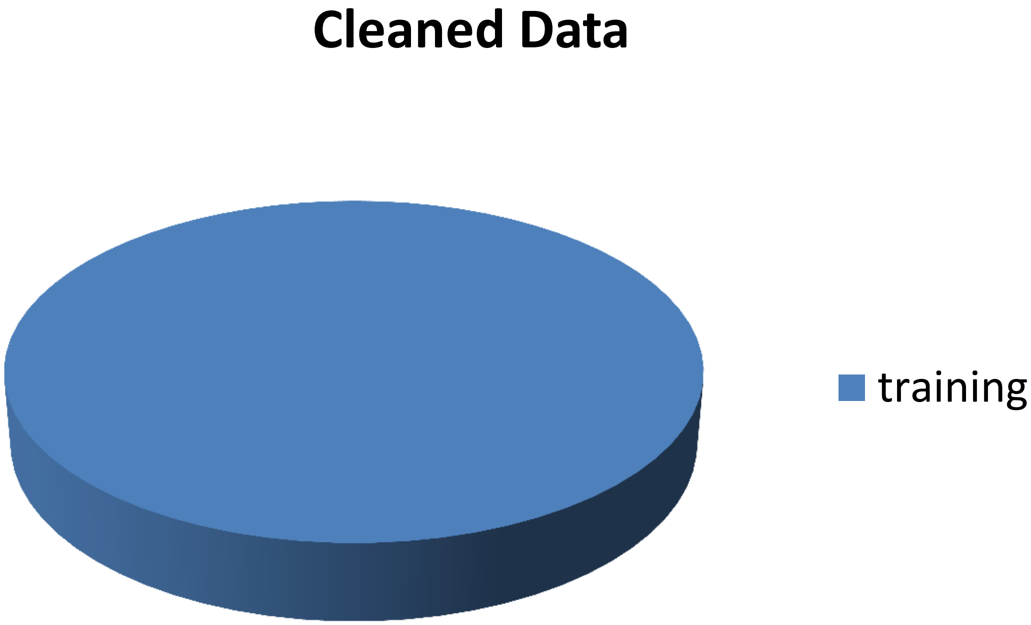
- Allerhand, M. (2012). A Tiny Handbook of R. Berlin: Springer. ISBN 9783642179808.
- Baayen, R. H. (2008). Analyzing Linguistic Data: A Practical Introduction to Statistics Using {R}. Cambridge: Cambridge University Press.
- Faraway, J. J. (2006). Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models. Boca Raton, FL: Chapman and Hall.
- Johnson, K. (2008). Quantitative Methods in Linguistics. Malden, MA: Blackwell. ISBN 978-1-4051-4425-4.
- Menard, S. W. (2009). Logistic Regression: From Introductory to Advanced Concepts and Applications. Thousand Oaks, CA: Sage. ISBN 978-1-4129-7483-7.
- Webpage of the R Study Group at UPenn Linguistics Dept.
- Dale J. Barr, Roger Levy, Christoph Scheepers, Harry J. Tily (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. JML 68 p 255-278.
- Wurm & Fisicaro (2014). What residualizing predictors in regression analyses does (and what it does not do). JML 72, p. 37-48.
- Baker, Monya (2016). Statisticians issue warning about misuse of p-values. Nature.

Data sets


All Data


All Data
"clean data"
implausible responses
outliers, missing ...

Cleaned training set

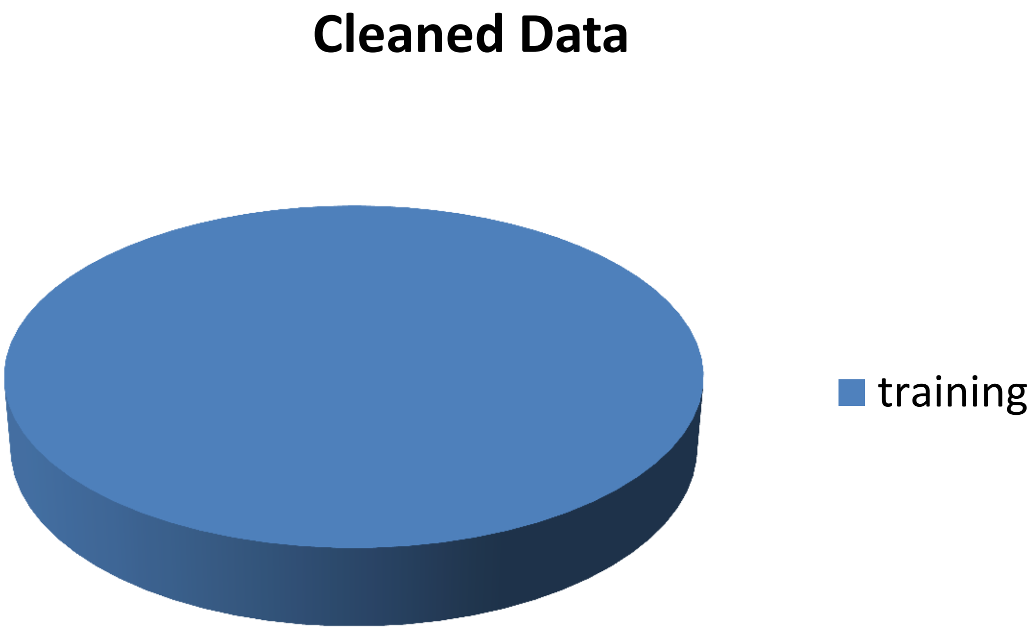


data set you optimize regression models on

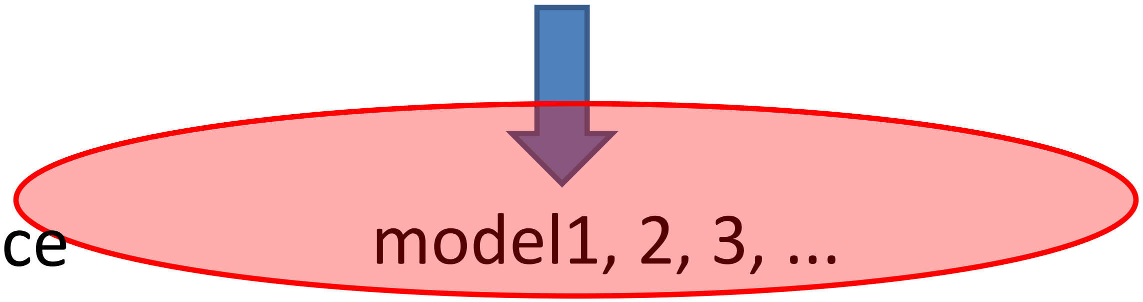
↓

model 1, 2, 3, ...

Training set

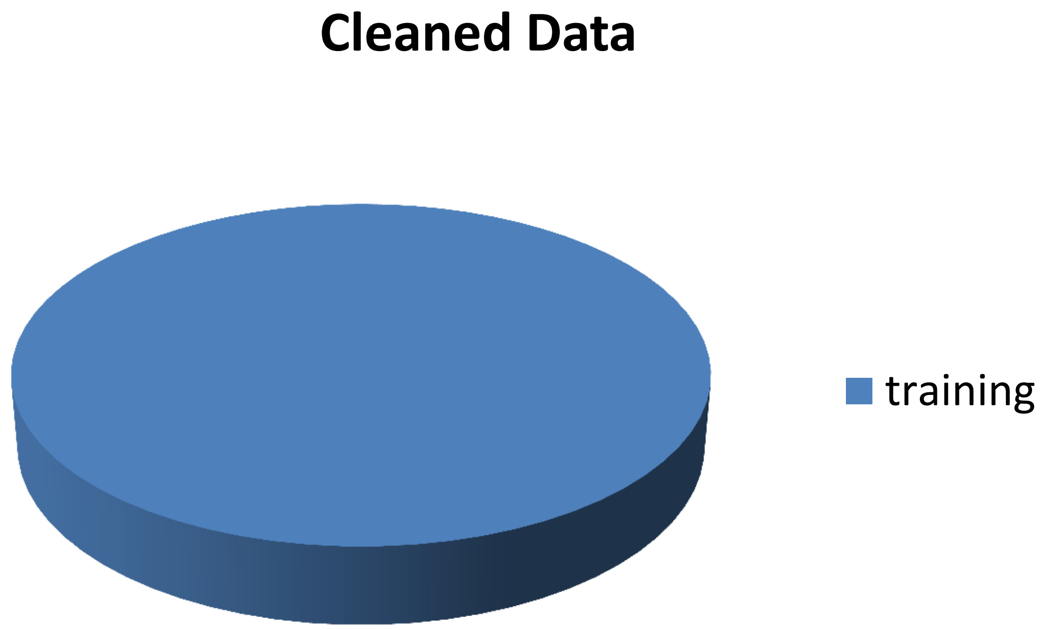


data set you optimize regression models on

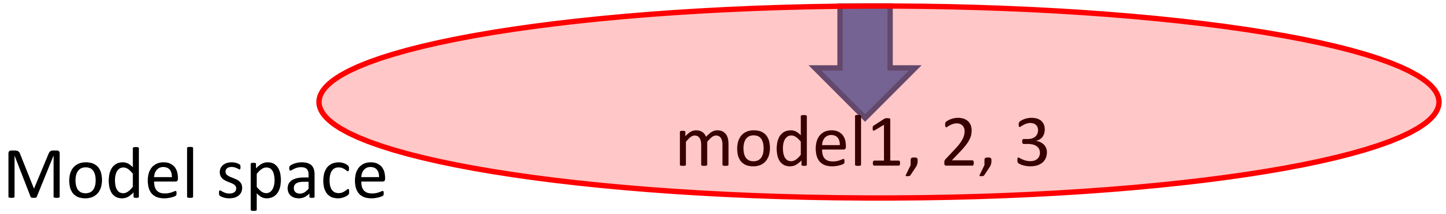


Model space

model1, 2, 3, ...



data set you optimizing regression models on


model1, 2, 3

Model space

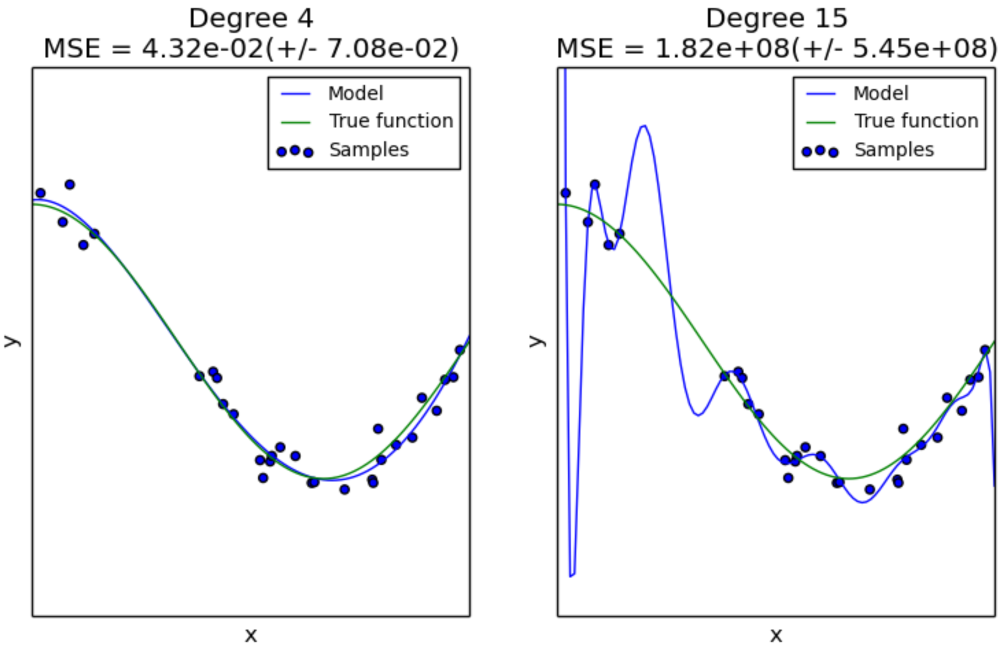
Log likelihood, R2
AIC, BIC (the smaller the better)
anova (if nested)
theoretical elegance
generalizability

Large models

- If you start at the rich side, final model is biased to become (too) large
  - good match on data
  - poor match on unseen data
  - AIC and BIC help to avoid constructing too large models
- Large model may fail to converge
  - model may be OK, but you cannot be sure

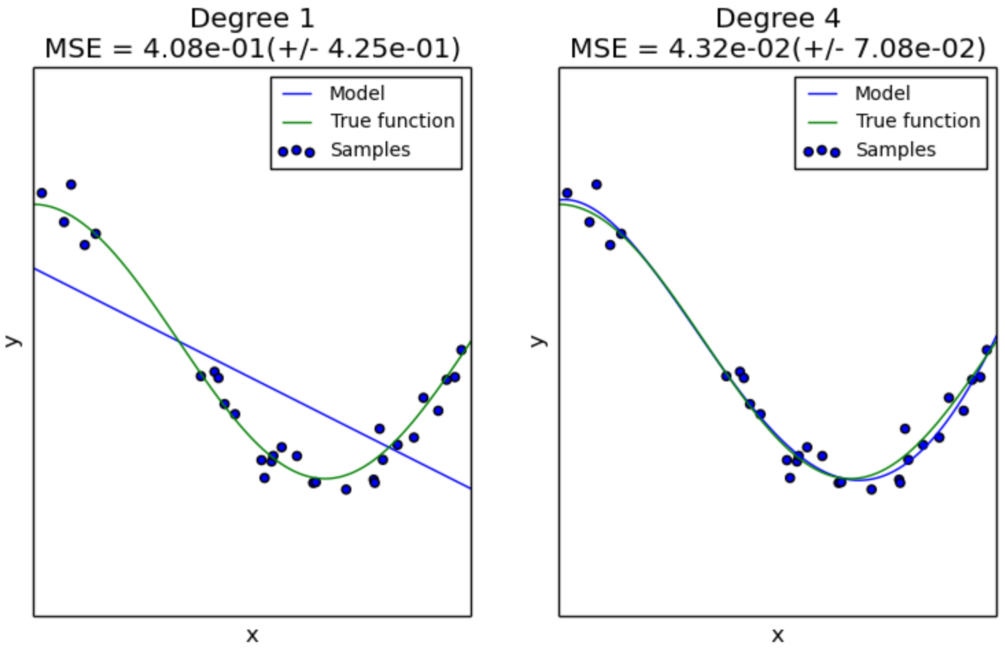
Too large …

- A large model may be *overfitting*
  - Excellent match on training data, poor fit on unseen new data

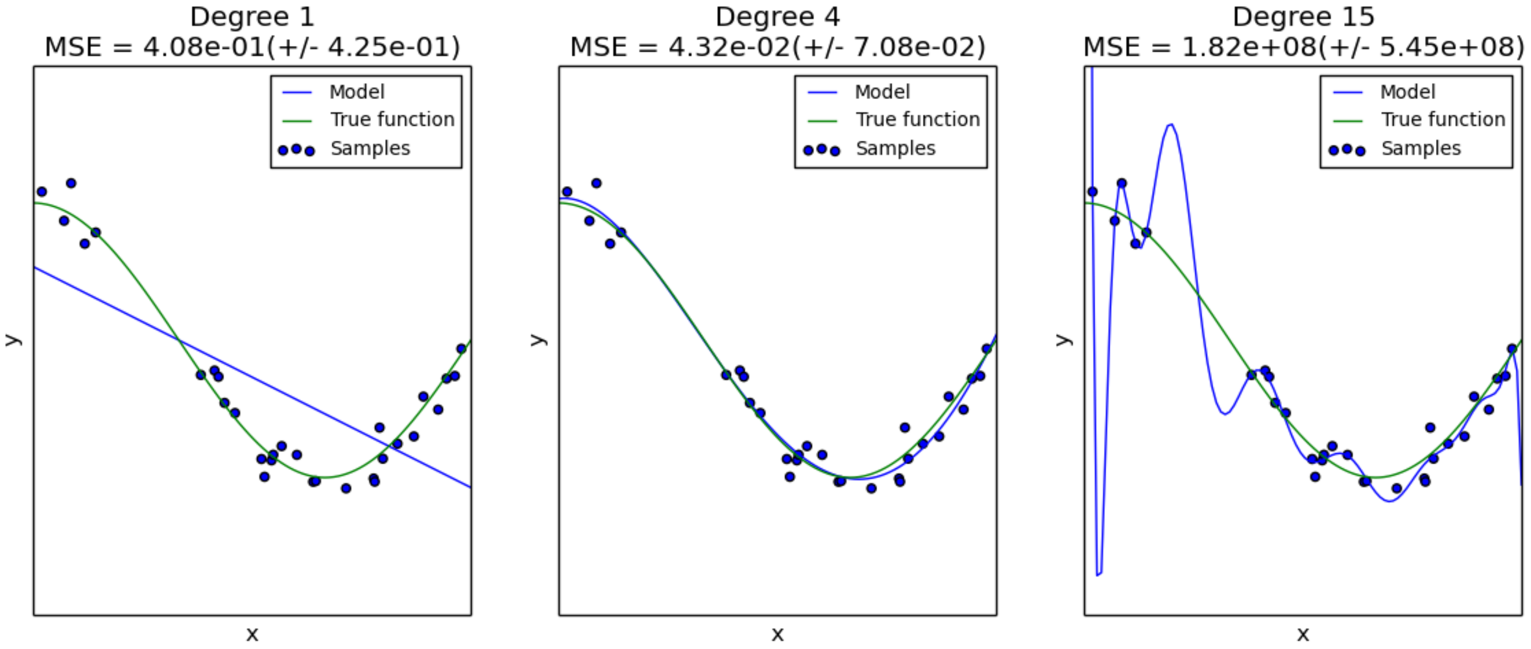


Too small …

- A too small model underfits the data
  - i.e. may not grasp the details that are in the data



“Model Bias” versus “Model Variance”


Degree 1
MSE = 4.08e-01(+/- 4.25e-01)

Degree 4
MSE = 4.32e-02(+/- 7.08e-02)

Degree 15
MSE = 1.82e+08(+/- 5.45e+08)

Model too simple -> underfitting

Model too complex -> overfitting

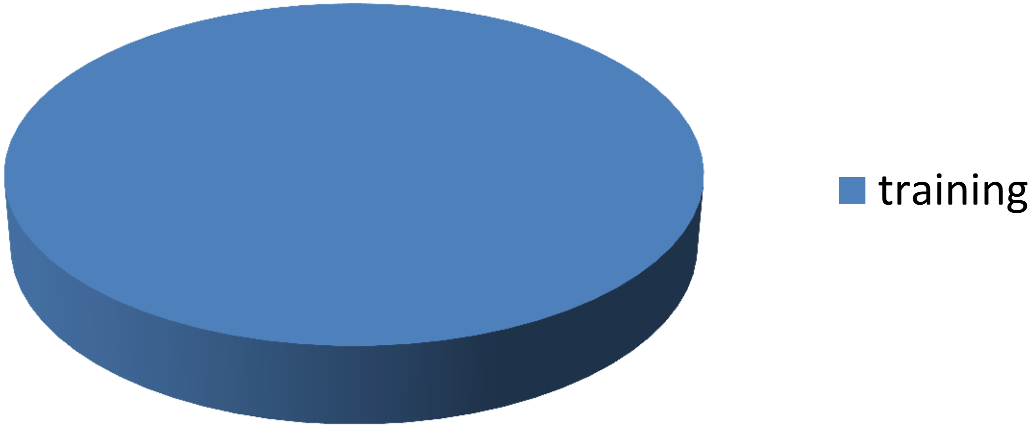
Too large ...

Douglas Bates, Reinhold Kliegl, Shravan Vasishth, and Harald Baayen, Parsimonious mixed models, arXiv:1506.04967v1 [stat.ME] 16 Jun 2015

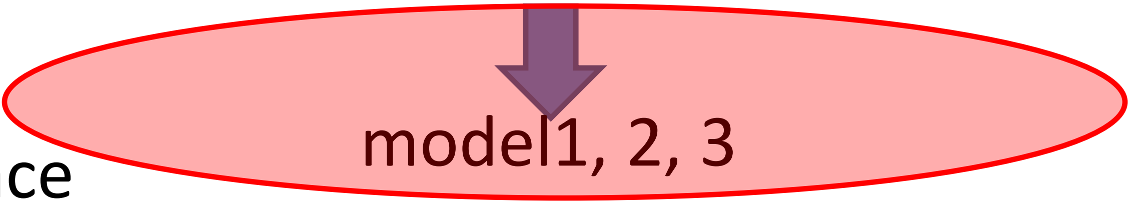
Recently, Barr et al. (2013) recommended fitting ‘maximal’ models with all possible random effect components included. Estimation of maximal models, however, may not converge.

We show that failure to converge typically is not due to a suboptimal estimation algorithm, but is a consequence of attempting to fit a model that is too complex to be properly supported by the data, irrespective of whether estimation is based on maximum likelihood or on Bayesian hierarchical modeling [...]

Cleaned Data



data set you train regression models on

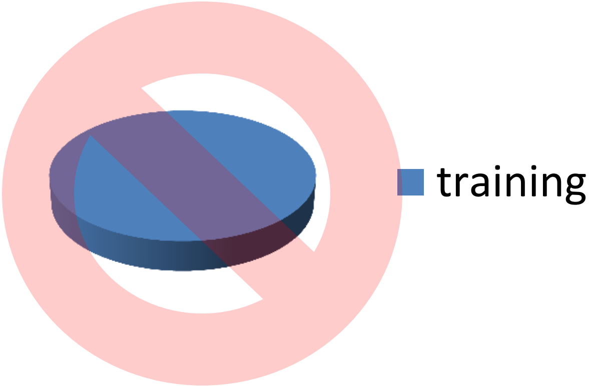


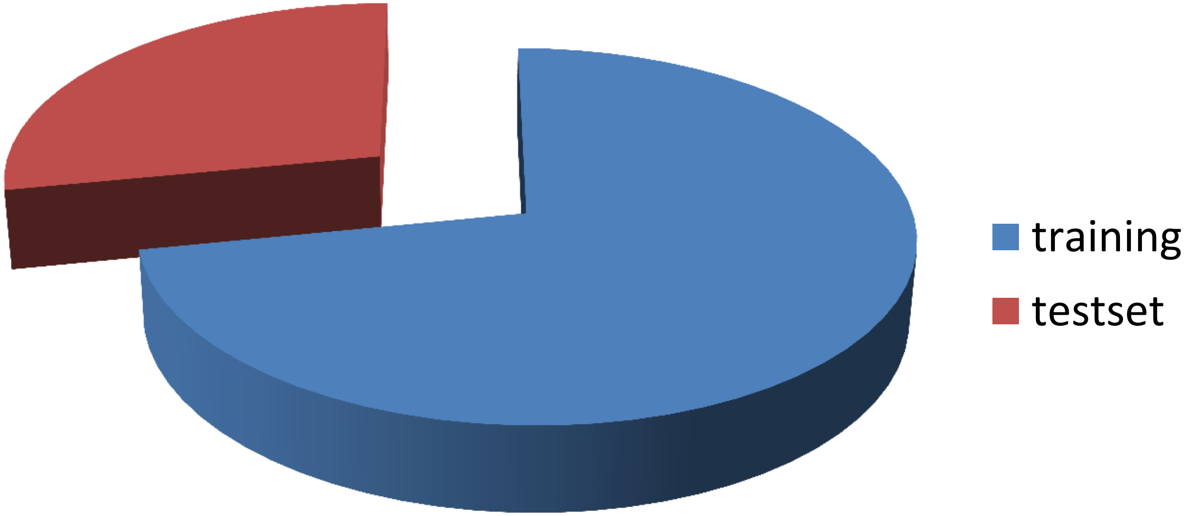
model1, 2, 3

Model space

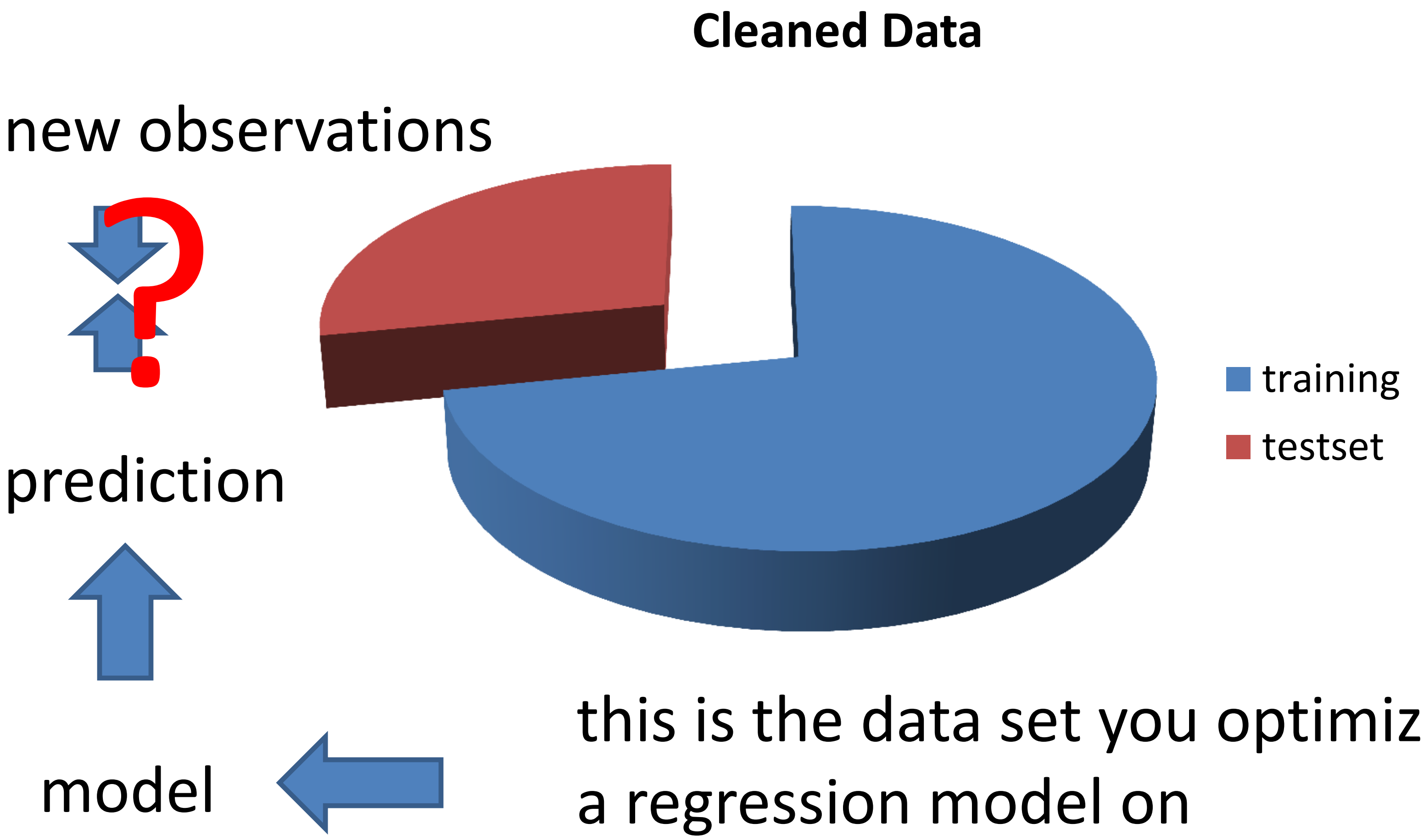
Log likelihood, R2	
AIC, BIC	(the smaller the better)
anova	(if nested)
theoretical elegance	
generalizability	

SPLIT: Training set and test set





Training set and test set



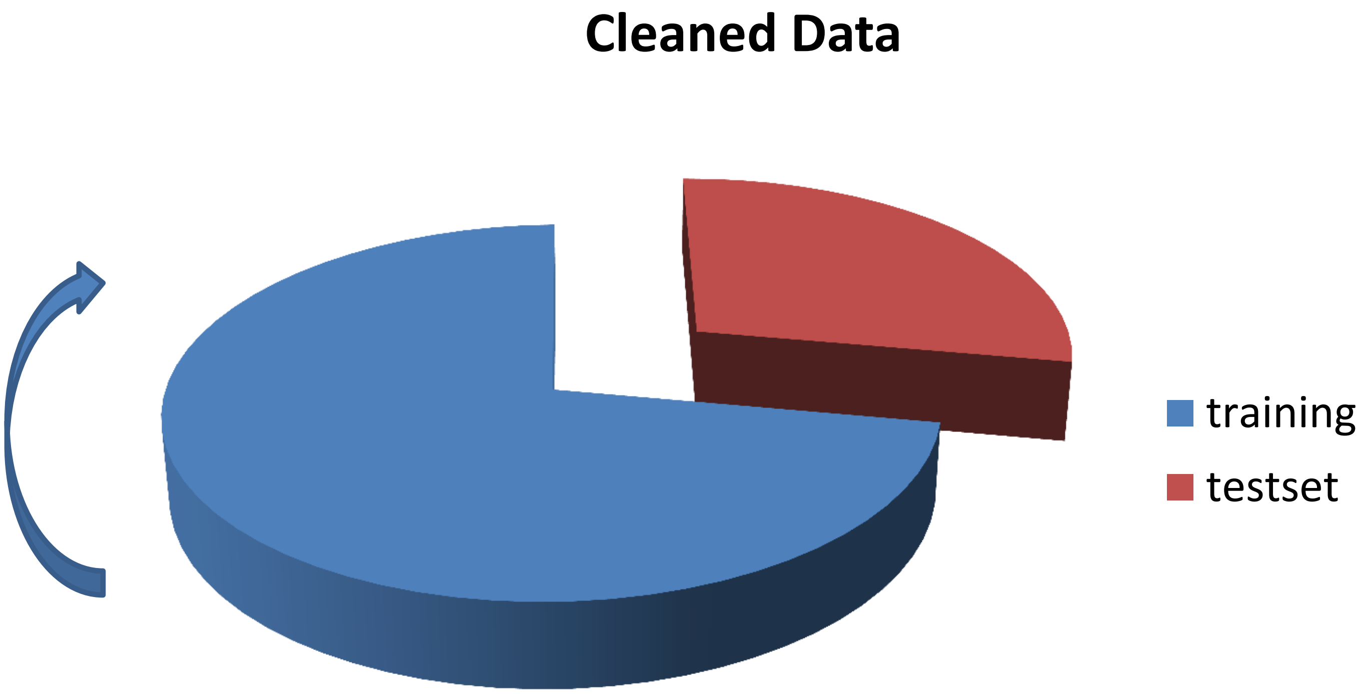
new observations

prediction

model

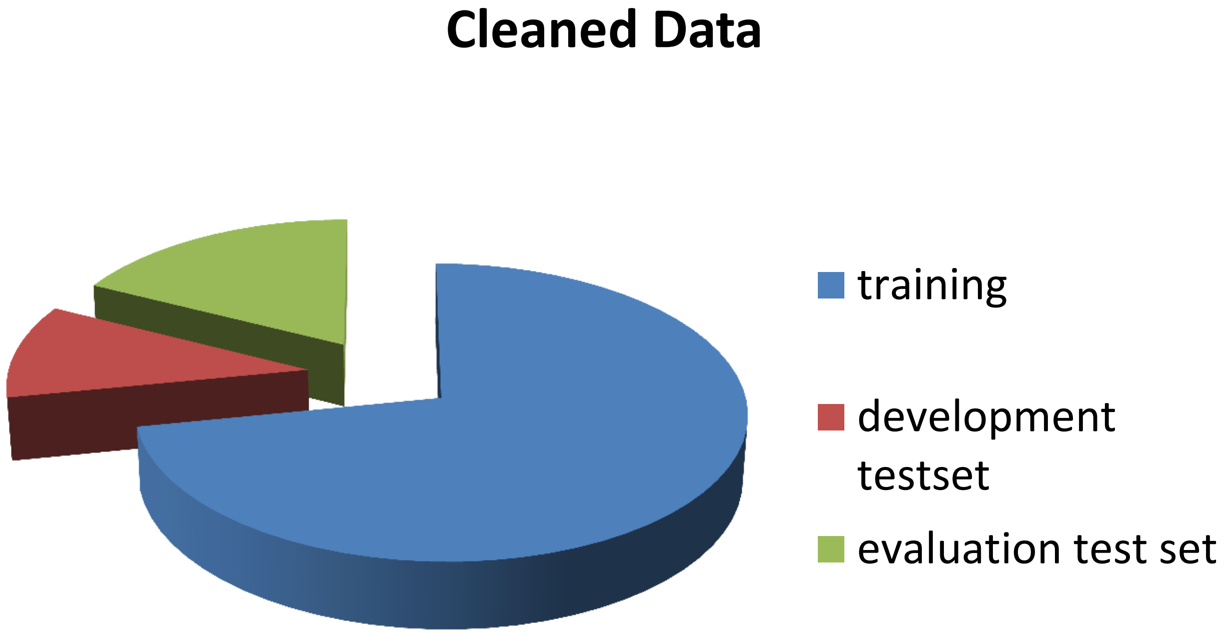
this is the data set you optimize a regression model on

Training set and test set



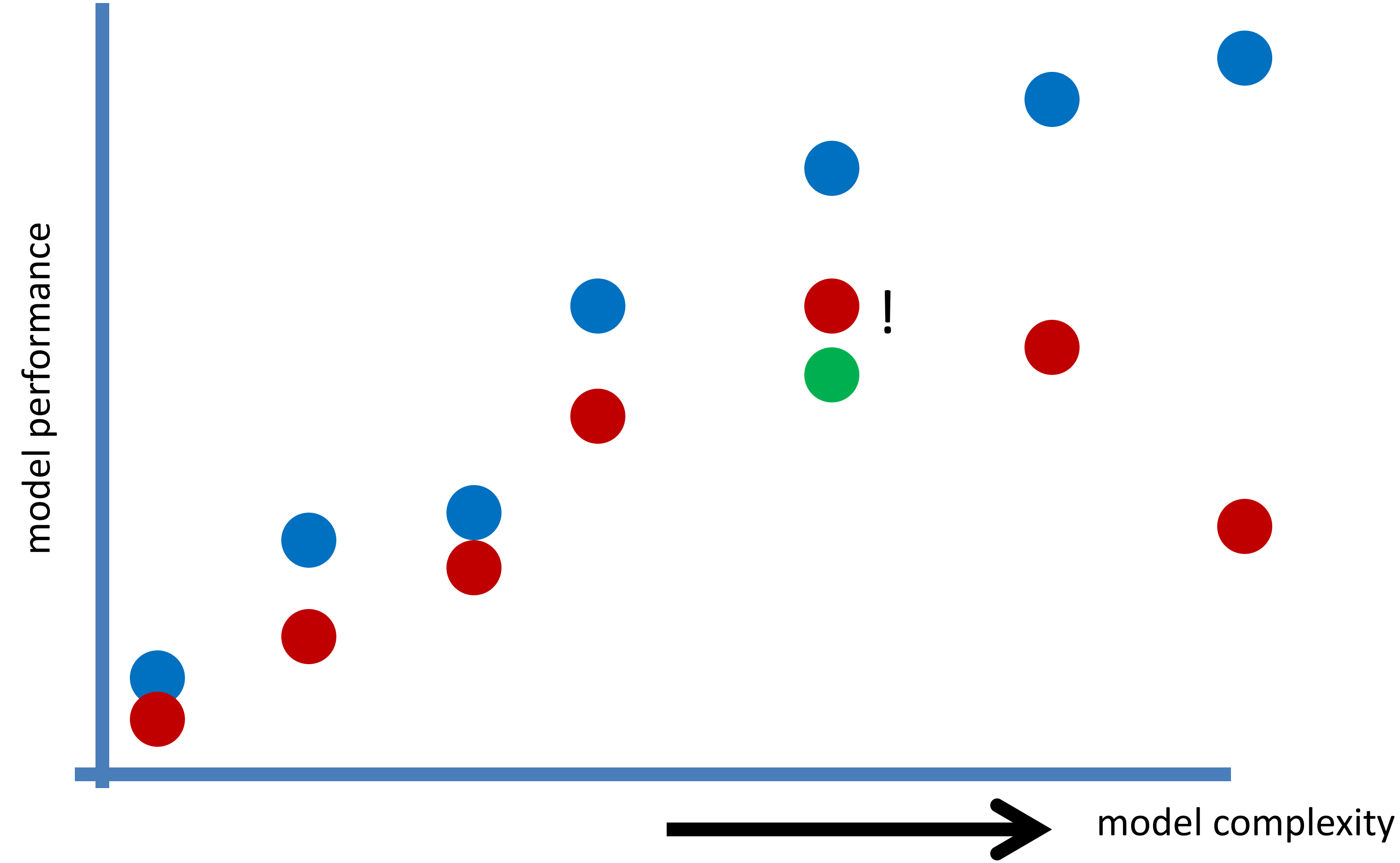
- Rotate test partition
- Report average over all N partitions (N-fold cross validation)

Training set and 2 test sets



- During optimization on training set, inspect performance on development test set
- If performance on development test set starts to decrease, stop optimization
- Report performance of resulting model on new evaluation test set

Performance on training set, development test set, evaluation test set



Generalization

- In the search of a good model, two types of error are particularly relevant
  - Bias
  - Variance
- The *bias* is the error from erroneous assumptions in the learning algorithm.
  - High bias can cause to miss the relevant relations in the data (underfitting).
- The *variance* is the error from sensitivity to small fluctuations in the training set.
  - High variance can cause overfitting: modeling the random noise in the training data, rather than the intended outputs.

Generalization

- On a new data set, the *expected error* is a sum of three terms

bias + variance + irreducible error

(*irreducible error* comes from the noise in the problem itself)

- Too simple models will have a large bias
- Too complex models have a large variance
  - See e.g. https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff

Take-home messages

- We’re **not** learning about LMEMs for the sake of reducing our p-values
- We’re doing it because it’s (quite often) a better representation of the data

Take-home messages

- A too-rich model has a risk of overfitting
  - very good account of data used for training, but poor generalization to unseen data
- Failure to converge is often an indication of overparametrization
- A too poor model is biased
  - does not model the relevant details in data

Take-home messages

- AIC, BIC help avoiding overfitting
- There is no unique recipe for defining a good model
  - AIC, anova, log likelihood,…
  - Elegance
  - Theoretical foundation
  - Generalization power
- A principled way (not yet common in social sciences) is the smart usage of different held-out test sets

The program

- **Mon Oct 9, 2-4pm**: week 1 (Louis)
  - aim, short introduction (this file)
  - p-values, transformations, H0-H1, exploratory versus confirmatory, model selection (R^2, AIC, BIC, ...)

- **Thu Oct 19, 2-4pm**: week 2
  - fixed effects, random effects, slopes (Justin)
  - AIC, BIC, anova(), practice on artificial data set (will be made available) (Louis)

- **Wed Oct 25, 3-5pm**: week 3
  - releveling, centering, contrast coding (Louis)
  - plots and visualization (easy) (Justin)

- **Fri Nov 3, 3-5pm**: week 4 (Louis)
  - prediction, generalization
  - overcoming convergence problems

- **Wed Nov 8, 3-5pm**: week 5 (Justin)
  - glmer
  - bootstrapping
  - visualization, plotting and interaction (more advanced)
  - how to publish (sweave)

Next time

- You will get homework (github or email)
- We will start with artificial database
  - good to start with [small] dataset with known structure
- Download R
  - see Justin's mail
- See you next week
- Questions?