Overview on statistical modelling

Michele Gubian

Graz, October 2020

Outline

- Scientists vs. Engineers
  - or: Why linear models?
- Correlation vs. Causation
- Why only one response variable? (in linear models)
- Ordinary linear models
- Linear Mixed-Effects Models
- Generalised Additive Mixed Models (GAMMs)

Scientists vs. Engineers

- Example scientific scenario: How is anger expressed through prosody (in a certain language)?
- Collect angry and neutral speech
  - e.g. from call center recordings
- Label manually a small subset of the data
- Build one or more models on the data (after data pre-processing)

Scientists vs. Engineers

- Why a model?
  - To isolate and understand the components of anger expressed by prosody
  - Isolate, because prosody (speech in general) varies for tons of other unrelated reasons
  - Understand, because it’s science :)

Scientists vs. Engineers

- The type of model you build is simple, usually a linear (ME) model. Why?
  - Wrong answer: because you don't have enough data, otherwise you would use a DNN
  - Better answer: because you need to isolate and understand, and a LM decomposes sources of change and adds them together

Scientists vs. Engineers

- Example LM (oversimplified)

  f0 range ~ anger + other factors (e.g. age, gender, position in sentence, etc.)

- Each factor (RHS) may or may not contribute in determining (or “explaining”) a variation (with sign) in the response variable (LHS)

Scientists vs. Engineers

- What about a DNN instead?

  f0 range ~ DNN(anger, other factors)

- Prediction accuracy would increase
  - Because reality is way more complex than linear
- Interpretability would decrease (or vanish)
  - Because a DNN usually contains millions of parameters connected by non-linear relationships

Scientists vs. Engineers

- Example engineering scenario: detect anger automatically from speech recordings
- Collect (huge amounts of) angry and neutral speech, e.g. from call center recordings
- Label manually a small subset of the data
- Build a classifier = a magic box that tells you anger yes/no given a stretch of speech

  anger ~ complicated_function(lots of acoustic features)

Scientists vs. Engineers

- Why a classifier?
  - Purpose is to predict (guess) anger yes/no from speech (prosody), not to understand and isolate factors contributing to anger in prosody
  - No need to interpret (the millions of) parameters in the classifier
  - The more data, the higher the support for complexity, the better predictions

    “There is no better data than more data”

Correlation vs. Causation

- A model like this

  f0 range ~ anger + other factors

  provides support for correlation, not causation

  e.g. anger co-occurs with greater f0 range

  not anger implies greater f0 range

Correlation vs. Causation

- A model like this

  f0 range ~ anger + other factors

  usually has the structure:
  - RHS: causes
  - LHS: effect (only one)
- But this is prior knowledge, not a result

Correlation vs. Causation

- A model like this

  f0 range ~ anger + other factors

  is interpreted like this

  “the presence of anger contribute for an increase of 50Hz in f0 range”

  which implies the direction of causation

  i.e. anger causes a change in prosody, not prosody causes anger in the speaker

  but the model does not know that

Correlation vs. Causation

- We can swap roles

  prob(anger) ~ f0 range + other prosodic features

  and maintain the direction of causality

  “an increase of 50Hz in f0 range is associated to an increase of 0.3 in prob(anger)”

- Useful to examine the efficacy of different prosodic features in revealing (not causing!) anger

Why only one response variable?

- Why not?

  f0 range + other features ~ anger + other factors

  or

  (F1, F2) ~ factors affecting a vowel

  i.e. we would like to predict many (correlated) effects on a number of (correlated) acoustic features

Why only one response variable?

- It is legal :)
  - i.e. no conflict with causation or anything logical
- But estimation is hard
- Most (R) stats packages have trouble with multiple-response models
  - lme4 does not support them
  - mgcv (GAMMs) does but with (heavy) limitations
  - brms (Stan, Bayesian) does

A look ahead

- we will do stats on curves and trajectories, e.g. f0 or formants
- we will play the scientists, e.g. what aspect of f0 track shape reflects the presence of anger?
- instead of a numeric feature extracted 'by hand', like f0 range, we will have other ‘shape features’ computed by FPCA
- we will put one of these features (at the time) on the LHS of a LMER model
- we will also look at what GAMMs do, which is similar, but technically quite different
- so we need to use LMER (and LM) as building block