Mixed-effects regression models

Martijn Wieling

Center for Language and Cognition Groningen, University of Groningen

Seminar in Statistics and Methodology - March 29, 2011

Overview

- Introduction
- Recap: multiple regression
- Mixed-effects regression analysis
  - Explanation
  - Case-study: Dutch dialect data
- Conclusion

Introduction

- Consider the following situation (taken from Clark, 1973):
  - Mr. A and Mrs. B study reading latencies of verbs and nouns
  - Each randomly selects 20 words and tests 50 participants
  - Mr. A finds (using a sign test) verbs to have faster responses
  - Mrs. B finds nouns to have faster responses
- How is this possible?

Introduction

- Consider the following situation (taken from Clark, 1973):
  - Mr. A and Mrs. B study reading latencies of verbs and nouns
  - Each randomly selects 20 words and tests 50 participants
  - Mr. A finds (using a sign test) verbs to have faster responses
  - Mrs. B finds nouns to have faster responses
- How is this possible?

The language-as-fixed-effect fallacy

- The problem is that Mr. A and Mrs. B disregard the variability in the words (which is huge)
  - Mr. A included a difficult noun, but Mrs. B included a difficult verb
  - Their set of words does not constitute the complete population of nouns and verbs, therefore their results are limited to their words
- This is known as the language-as-fixed-effect fallacy (LAFEF)
  - Fixed-effect factors have repeatable and a small number of levels
  - Word is a random-effect factor (a non-repeatable random sample from a larger population)

Why linguists are not always good statisticians

- LAFEF occurs frequently in linguistic research until the 1970's
  - Many reported significant results are wrong (the method is anti-conservative)!
- Clark (1973) combined a by-subject (F_1) analysis and by-item (F_2) analysis in a measure called *min F'*
  - Results are significant and generalizable across subjects and items when *min F'* is significant
  - Unfortunately many researchers (>50%!) incorrectly interpreted this study and may report wrong results (Raaijmakers et al., 1999)
  - E.g., they only use F_1 and F_2 and not *min F'* or they use F_2 while unneccesary (e.g., counterbalanced design)

Our problems solved...

- Apparently, analyzing this type of data is difficult...
- Fortunately, using mixed-effects regression models solves all our problems!
  - The method is easier than using the approach of Clark (1973)
  - Results can be generalized across subjects and items
  - Mixed-effects models are robust to missing data (Baayen, 2008, p. 266)
  - We can easily test if it is necessary to treat words as a random effect
- As mixed-effects regression models are an extension of multiple regression, a brief recap follows

Our problems solved...

- Apparently, analyzing this type of data is difficult...
- Fortunately, using mixed-effects regression models solves all our problems!
  - The method is easier than using the approach of Clark (1973)
  - Results can be generalized across subjects and items
  - Mixed-effects models are robust to missing data (Baayen, 2008, p. 266)
  - We can easily test if it is necessary to treat words as a random effect
- As mixed-effects regression models are an extension of multiple regression, a brief recap follows

Recap: multiple regression

- Multiple regression: predict one numerical variable on the basis of other independent variables (numerical or categorical)
  - (*Logistic* regression is used to predict a categorical dependent)
- We can write a regression formula as $y = I + ax_1 + bx_2 + ...$
- E.g., predict the reaction time of a participant on the basis of word frequency, word length and subject age: $RT = 200 - 5WF + 3WL + 10SA$

Mixed-effects regression modeling: introduction

- Mixed-effects regression modeling distinguishes fixed-effects and random-effects factors
- Fixed-effects factors:
  - Repeatable levels
  - Small number of levels (e.g., Gender, Word Category)
  - Same treatment as in multiple regression (dummy coding)
- Random-effects factors:
  - Levels are a non-repeatable random sample from a larger population
  - Often large number of levels (e.g., Subject, Item)

What are random-effects factors?

- Random-effects factors are factors which are likely to introduce systematic variation
  - Some participants have a slow response (RT), while others are fast
  - Some words are easy to recognize, others hard
  - The effect of word frequency on RT might be higher for one participant than another (e.g., non-native subjects might have more profit from frequent words than native subjects)
  - The effect of subject age on RT might be different for one word than another (e.g., modern words might be recognized easier by younger subjects)

Specific models for every observation

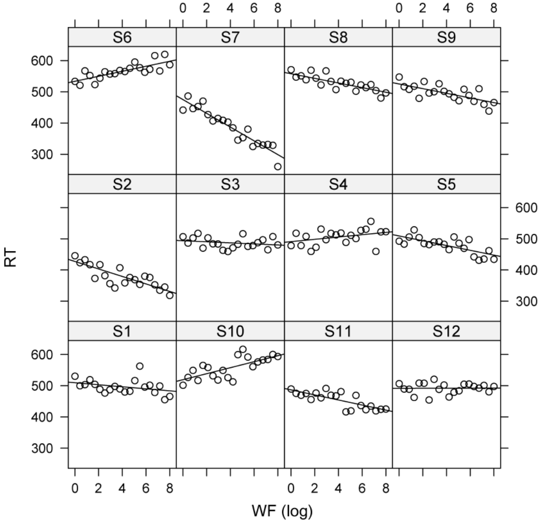
- Mixed-effects regression analysis allow us to use random intercepts and random slopes to make the regression formula as precise as possible for every individual observation in our random effects
  - A single parameter (standard deviation) models this variation for every random slope or intercept
  - The actual random intercepts and slopes are derived from this value
  - Likelihood-ratio tests assess whether the inclusion of random intercepts and slopes is warranted
- Note that multiple observations for each level of a random effect are necessary for mixed-effects analysis to be useful (e.g., participants respond to multiple items)

Specific models for every observation

- $RT = 200 - 5WF + 3WL + 10SA$ (general model)
  - The intercepts and slopes may vary (according to the estimated standard variation for each parameter) and this influences the word- and subject-specific values
- $RT = 400 - 5WF + 3WL - \ 2SA$ (word: scythe)
- $RT = 300 - 5WF + 3WL + 15SA$ (word: twitter)
- $RT = 300 - 7WF + 3WL + 10SA$ (subject: non-native)
- $RT = 150 - 5WF + 3WL + 10SA$ (subject: fast)
- And it is easy to use!

```
> lmer( RT ~ WF + WL + SA + (1+SA|Wrd) + (1+WF|Subj) )
```

Specific models for every subject



Case study: Dutch dialects w.r.t. standard Dutch

- The goal of this study is to investigate which factors predict the dialect distances of 562 words in 424 locations from standard Dutch
- We use a mixed-effects regression model for this purpose
  - Random-effects factors: Location, Word and Transcriber
- Several location-, speaker- and word-related factors are investigated
  - E.g., number of inhabitants, average age of inhabitants, speaker age, speaker gender, word frequency and word category

Geographic distribution of locations

Determining dialect distances

- We use phonetic transcriptions of 562 words in 424 locations in NL
- These are compared to standard Dutch transcriptions using the Levenshtein algorithm (Levenshtein, 1965)
  - The Levenshtein algorithm measures the minimum number of insertions, deletions and substitutions to transform one string into another

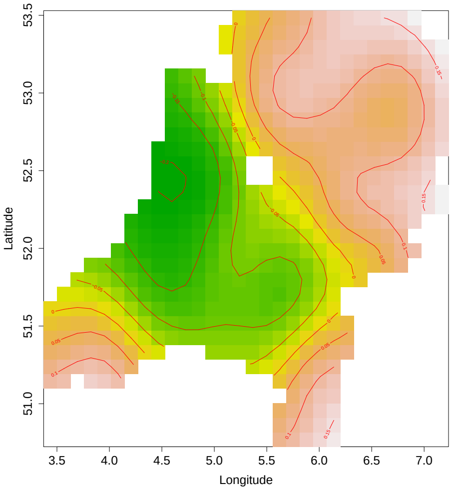
b		ɪ	n	d	ə	n
b	ɛ	i	n	d	ə	
	1	1				1

- The distance between the dialectal and standard Dutch pronunciation is based on the total cost of the operations (above: 3)
- We actually use more sensitive, automatically determined, sound distances: e.g., contrasting [a]:[ɑ] from [a]:[i] (Wieling et al., 2007)

The influence of geography

- An important determinant for dialect variation is geographic location (people in nearby locations have more contact than in distant locations)
- We include geography by predicting dialect distances with a Generalized Additive Model (GAM) which models the interaction between longitude and latitude
  - The fitted values of this GAM are included as a predictor in our model
  - (The details of this procedure are outside the scope of this lecture)

Fitted GAM for dialect distance from standard Dutch



Final model: fixed-effects

	Estimate	Std. Error	*t*-value
Intercept	-0.0153	0.0105	-1.4561
GAM distance (geography)	0.9684	0.0274	35.3239
Population size (log)	-0.0069	0.0026	-2.6386
Population average age	0.0045	0.0025	1.8049
Population average income (log)	-0.0005	0.0026	-0.1988
Noun instead of Verb/Adjective	0.0409	0.0122	3.3437
Word frequency (log)	0.0198	0.0060	3.2838
Vowel-consonant ratio (log)	0.0625	0.0059	10.5415

*t-values indicate significance if $|t| > 2$ (two-tailed) or $|t| > 1.65$ (one-tailed)

Final model: random effects

Factors	Rnd. effects	Std. Dev.	Cor.	
Word	Intercept	0.1394		
	Pop. size (log)	0.0186		
	Pop. avg. age	0.0086	-0.856	
	Pop. avg. income (log)	0.0161	0.867	-0.749
Location	Intercept	0.0613		
	Word freq. (log)	0.0161	-0.084	
	Noun instead of Verb/Adjective	0.0528	-0.595	0.550
Transcriber	Intercept	0.0260		
Residual		0.2233		

*The inclusion of all random slopes and intercepts was warranted by likelihood-ratio tests

*A richer random effect structure is likely possible, but not computationally feasible (now: 24 CPU hours!)

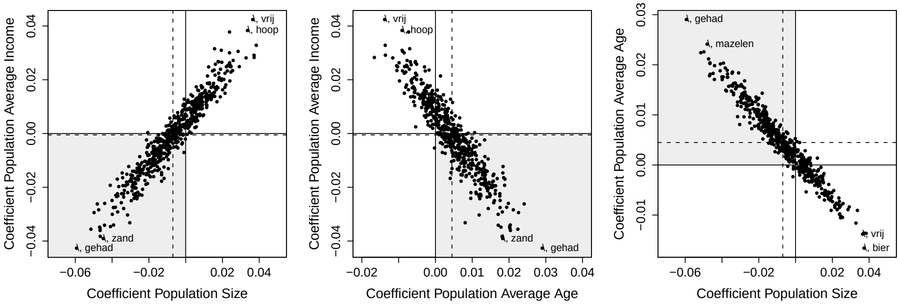
Final model: interesting correlational structure

Factors	Rnd. effects	Std. Dev.	Cor.	
Word	Intercept	0.1394		
	Pop. size (log)	0.0186		
	Pop. avg. age	0.0086	-0.856	
	Pop. avg. income (log)	0.0161	0.867	-0.749
Location	Intercept	0.0613		
	Word freq. (log)	0.0161	-0.084	
	Noun instead of Verb/Adjective	0.0528	-0.595	0.550
Transcriber	Intercept	0.0260		
Residual		0.2233		

*The inclusion of all random slopes and intercepts was warranted by likelihood-ratio tests

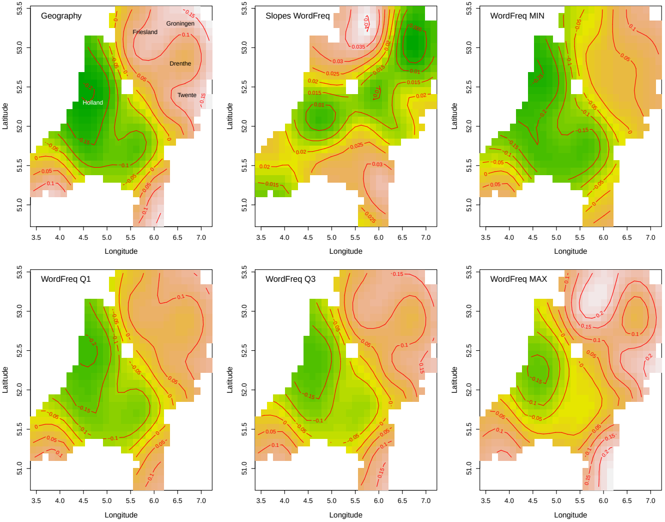
*A richer random effect structure is likely possible, but not computationally feasible (now 24 CPU hours!)

Correlation structure of by-word random slopes

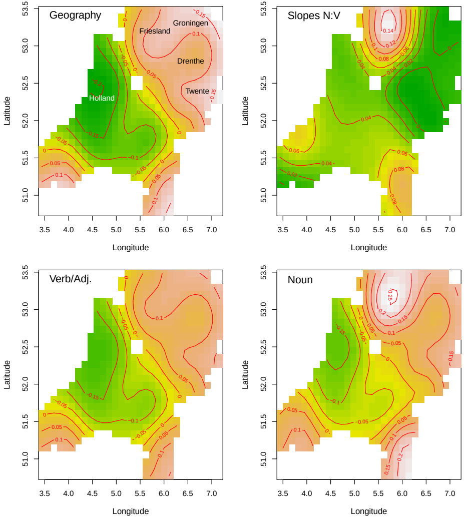


- $LD = -0.0069 PS - 0.0005 PI + 0.0045 PA + \ldots$ (general model)
- $LD = -0.0600 PS - 0.0420 PI + 0.0290 PA + \ldots$ (*gehad*: extreme pattern)
- $LD = \quad 0.0380 PS + 0.0420 PI - 0.0110 PA + \ldots$ (*vrij*: inverted pattern)

By-location random slopes for word frequency



By-location random slopes for Noun-Verb contrast



Case study conclusions

- Our model explained about 45% of the variation in the data with respect to the distance from standard Dutch
- We identified a number of location- and word-related variables playing an important role in predicting the dialect distance from standard Dutch
  - Geography (i.e. social contact between locations)
  - Location-related factors: population size and average age
  - Word-related factors: word category, word frequency and vowel-cons. ratio
- Using a mixed-effects regression approach ensures our results are generalizable and enabled us to quantify and study the variation of individual words and speakers

What you should remember...

- Mixed-effects regression models offer an easy-to-use approach to obtain generalizable results when there are multiple random-effect factors
- Mixed-effects regression models allow a fine-grained inspection of the variability of the random effects, which may provide additional insight in your data
- Mixed-effects regression models are easy in R
    - Lab session: Thursday March 31, 9:00 - 11:00

Thank you for your attention!