# Generalized additive modeling and dialectology

## Lecture 3 of advanced regression for linguists

Martijn Wieling and Jacolien van Rij

Seminar für Sprachwissenschaft
University of Tübingen

LOT Summer School 2013, Groningen, June 26

# Today's lecture

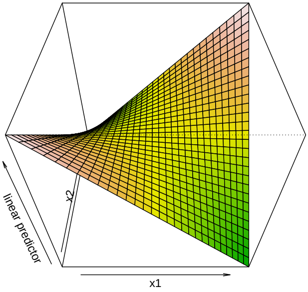
- Introduction
  - Some words about logistic regression
  - Generalized additive mixed-effects regression modeling
  - Standard Italian and Tuscan dialects
- Material: Standard Italian and Tuscan dialects
- Methods: `R` code
- Results
- Discussion

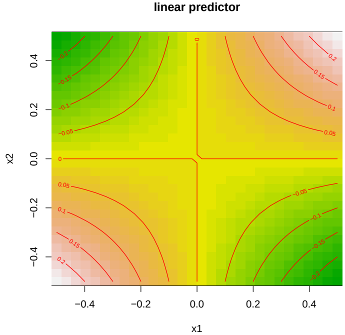
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Netherlands Organisation for Scientific Research

# A linear regression model

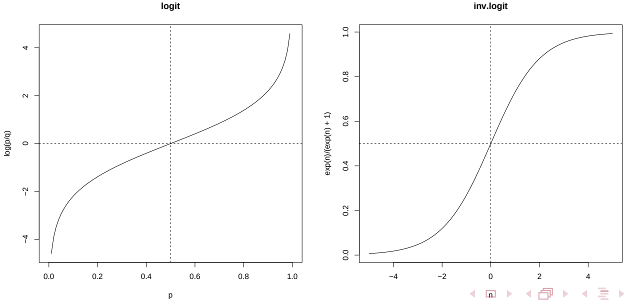
- *linear model*: linear relationship between predictors and dependent variable: $y = a_1 x_1 + ... + a_n x_n$
  - Non-linearities via explicit parametrization: $y = a_1 x_1^2 + a_2 x_1 + ...$
  - Interactions not very flexible



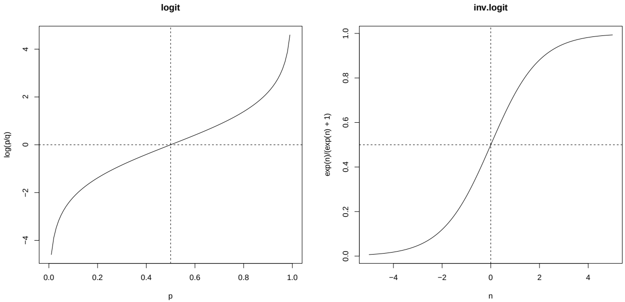
# A generalized linear regression model

- *generalized linear model*: linear relationship between predictors and dependent variable via link function: $g(y) = a_1 x_1 + ... + a_n x_n$
- Examples of link functions:
  - $y^2 = x \Rightarrow y = \sqrt{x}$
  - $\log(y) = x \Rightarrow y = \mathrm{e}^x$
  - $\mathrm{logit}(p) = \log(\frac{p}{1-p}) = x \Rightarrow p = \frac{e^x}{e^x+1}$



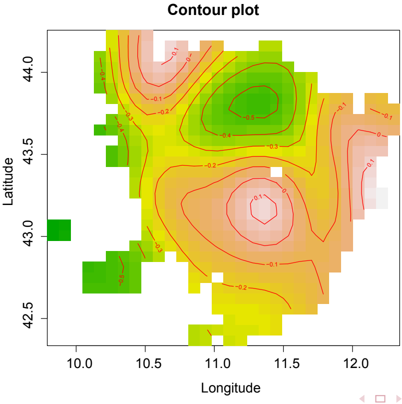
# Logistic regression

- Dependent variable is binary (1: success, 0: failure), not continuous
- Transform to continuous variable via log odds: $\log(\frac{p}{1-p}) = \mathrm{logit}(p)$
- Done automatically in regression by setting `family="binomial"`
- interpret coefficients w.r.t. success as logits: in `R`: `plogis(x)`



# A generalized additive model (1)

- *generalized additive model (GAM)*: relationship between individual predictors and (possibly transformed) dependent variable is estimated by a non-linear smooth function: $g(y) = s(x_1) + s(x_2, x_3) + a_4x_4 + ...$
  - multiple predictors can be combined in a (hyper)surface smooth



# A generalized additive model (2)

- Advantage of GAM over manual specification of non-linearities: the optimal shape of the non-linearity is determined automatically
  - appropriate degree of smoothness is automatically determined on the basis of cross validation to prevent overfitting
- Choosing a smoothing basis
  - Single predictor or isotropic predictors: thin plate regression spline
    - Efficient approximation of the optimal (thin plate) spline
  - Combining non-isotropic predictors: tensor product spline
- Generalized Additive Mixed Modeling:
  - Random effects can be treated as smooths as well (Wood, 2008)
  - `R`: `gam` and `bam` (package `mgcv`)
- For more (mathematical) details, see Wood (2006)

# Standard Italian and Tuscan dialects

- Standard Italian originated in the 14th century as a written language
- It originated from the prestigious Florentine variety
- The *spoken* standard Italian language was adopted in the 20th century
  - People used to speak in their local dialect

- In this study, we investigate the relationship between standard Italian and Tuscan dialects
  - We focus on lexical variation
  - We attempt to identify which social, geographical and lexical variables influence this relationship

# Material: lexical data

- We used lexical data from the *Atlante Lessicale Toscano* (ALT)
  - We focus on 2060 speakers from 213 locations and 170 concepts
  - Total number of cases: 384,454
  - For every case, we identified if the lexical form was different from standard Italian (1) or the same (0)

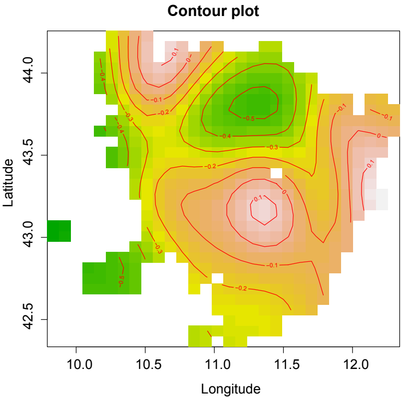
# Geographic distribution of locations

# Material: additional data

- In addition, we obtained the following information:
  - Speaker age
  - Speaker gender
  - Speaker education level
  - Speaker employment history
  - Number of inhabitants in each location
  - Average income in each location
  - Average age in each location
  - Frequency of each concept

# Modeling geography's influence with a GAM

```
# logistic regression: family="binomial"
> geo = gam(NotStd ~ s(Lon,Lat), data=tusc, family="binomial")
> vis.gam(geo,view=c("Lon","Lat"),plot.type="contour",color="terrain",...)
```



# Adding a random intercept to a GAM

```
> model = bam(NotStd ~ s(Lon,Lat) + s(Concept,bs="re"),
              data=tusc, family="binomial")
> summary(model)

Family: binomial
Link function: logit

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3620     0.1152  -3.142  0.00168 **

Approximate significance of smooth terms:
              edf Ref.df Chi.sq p-value
s(Lon,Lat)  27.85  28.77   2265  <2e-16 ***
s(Concept) 168.63 169.00  66792  <2e-16 ***

R-sq.(adj) =  0.253   Deviance explained = 20.9%
fREML score = 5.4512e+05  Scale est. = 1         n = 384454
```

# Adding a random slope to a GAM

```
> model2 = bam(NotStd ~ s(Lon,Lat) + CommSize.log.z + s(Concept,bs="re")
                        + s(Concept,CommSize.log.z,bs="re"),
                data=tusc, family="binomial")
> summary(model2)

Family: binomial
Link function: logit

Parametric coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.3625     0.1161  -3.123   0.002 **
CommSize.log.z      -0.0587     0.0224  -2.621   0.009 **

Approximate significance of smooth terms:
                               edf Ref.df Chi.sq p-value
s(Lon,Lat)                    27.7  28.71   1984  <2e-16 ***
s(Concept)                   168.6 169.00  82474  <2e-16 ***
s(Concept,CommSize.log.z)    154.2 170.00  33956  <2e-16 ***

R-sq.(adj) =  0.257   Deviance explained = 21.3%
fREML score = 5.4476e+05  Scale est. = 1         n = 384454
```

# Varying geography's influence based on concept freq.

- Wieling, Nerbonne and Baayen (2011, *PLOS ONE*) showed that the effect of word frequency varied depending on geography
- Here we explicitly include this in the GAM with `te()`

```
> m = bam(NotStd ~ te(Lon, Lat, Freq, d=c(2,1)) + ...,
                    data=tusc, family="binomial")
```

- As this pattern may be presumed to differ depending on speaker age, we can integrate this in the model as well

```
> m = bam(NotStd ~ te(Lon, Lat, Freq, Age, d=c(2,1,1)) + ...,
                    data=tusc, family="binomial")
```

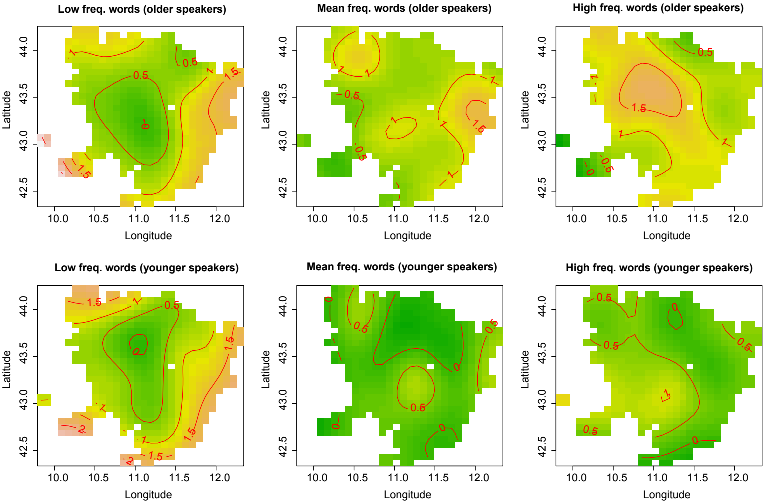
- The results will be discussed next... (Wieling et al., submitted)

# Results: fixed effects and smooths

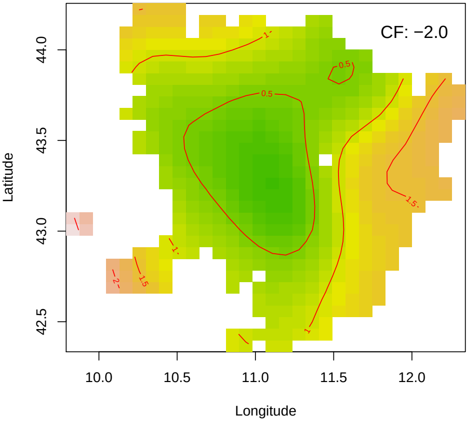
| | Estimate | Std. Error | $z$-value | $p$-value |
|---|---|---|---|---|
| Intercept | -0.4188 | 0.1266 | -3.31 | $< 0.001$ |
| Community size (log) | -0.0584 | 0.0224 | -2.60 | 0.009 |
| Male gender | 0.0379 | 0.0128 | 2.96 | 0.003 |
| Farmer profession | 0.0460 | 0.0169 | 2.72 | 0.006 |
| Education level (log) | -0.0686 | 0.0126 | -5.44 | $< 0.001$ |

| | Est. d.o.f. | Chi. sq. | $p$-value |
|---|---|---|---|
| Geo $\times$ frequency $\times$ speaker age | 225.9 | 3295 | $< 0.001$ |

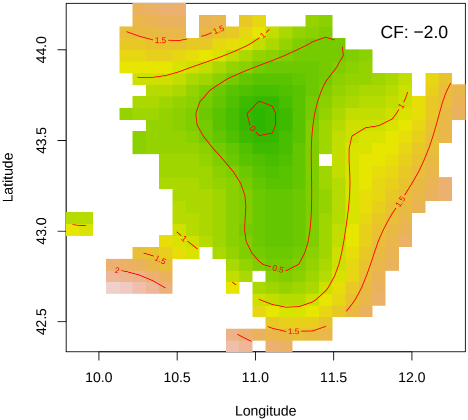
# A complex geographical pattern



# Animation: increasing frequency for older speakers



# Animation: increasing frequency for younger speakers

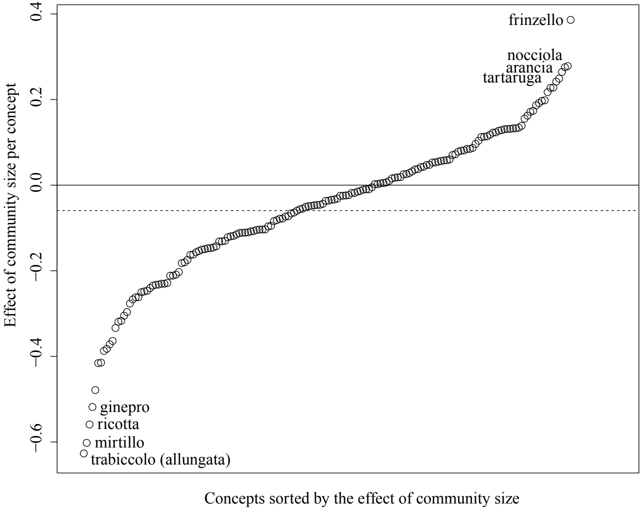


# Results: random effects

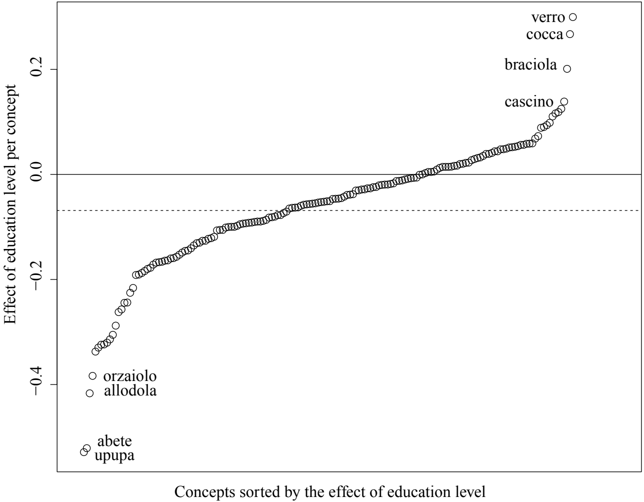
| Factors | Random effects | Std. dev. | $p$-value |
|---|---|---|---|
| Speaker | Intercept | 0.0100 | 0.006 |
| Location | Intercept | 0.1874 | $< 0.001$ |
| Concept | Intercept | 1.6205 | $< 0.001$ |
| | Year of recording | 0.2828 | $< 0.001$ |
| | Community size (log) | 0.1769 | $< 0.001$ |
| | Average community income (log) | 0.2657 | $< 0.001$ |
| | Average community age (log) | 0.2400 | $< 0.001$ |
| | Farmer profession | 0.1033 | $< 0.001$ |
| | Executive or auxiliary worker prof. | 0.0650 | 0.002 |
| | Education level (log) | 0.1255 | $< 0.001$ |
| | Male gender | 0.0797 | $< 0.001$ |

- Complex structure, logistic regression and large dataset: 23 hours of CPU time

# By-concept random slopes for community size



# By-concept random slopes for speaker education level



# Discussion

- Using a generalized additive mixed-effects regression model (GAMM) to investigate lexical differences between standard Italian and Tuscan dialects revealed interesting dialectal patterns
  - GAMs are very suitable to model the non-linear influence of geography
  - The regression approach allowed for the simultaneous identification of important social, geographical and lexical predictors
  - By including many concepts, results are less subjective than traditional analyses focusing on only a few pre-selected concepts
  - The mixed-effects regression approach still allows a focus on individual concepts

- There are some drawbacks to GAMMs, however...
  - `gam` and `bam` are computationally somewhat more expensive than linear mixed-effects modeling using `lmer` (`lme4` package)
  - Model comparison is problematic when including random-effect smooths (i.e. using `anova(gam1,gam2)` is useless)

# Conclusion

- Generalized additive modeling is useful to study non-linear effects
- Use `bam` if your dataset is large
- Use `s()` for predictors which are on the same scale
- Use `te()` when predictors are on a different scale
- (there is also a third option, `ti()`, which should be used when testing main effects and interactions)

- We will experiment with these issues in the lab session after the break!
  - We use a subset of *Dutch* dialect data (faster: no logistic regression)
  - Similar underlying idea: investigate the effect of geography, word frequency, and location characteristics on pronunciation distances from standard Dutch

# Thank you for your attention!