Linear Models

Tutorial *Corpus Statistics with R*

Andreas Blombach, Philipp Heinrich
and Stefan Evert

Chair of Computational Corpus Linguistics
Friedrich-Alexander-Universität Erlangen-Nürnberg
andreas.blombach@fau.de

Erlangen, 2019.10.08

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Linear regression

- Can random variable Y be predicted from random variable X?

here: focus on linear relationship between variables

- Linear predictor:

$$Y \approx \beta_0 + \beta_1 \cdot X$$

  - $\beta_0 =$ intercept of regression line
  - $\beta_1 =$ slope of regression line

Linear regression

- Can random variable Y be predicted from random variable X?

here: focus on linear relationship between variables

- Linear predictor:

$$Y \approx \beta_0 + \beta_1 \cdot X$$

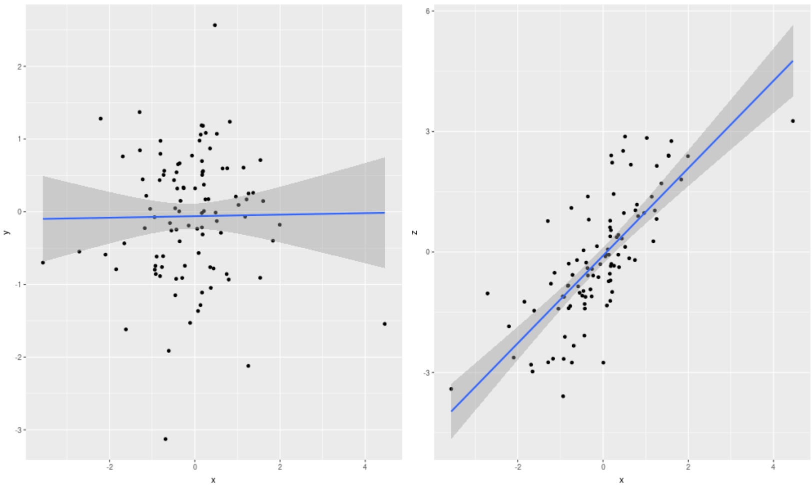
  - $\beta_0 =$ intercept of regression line
  - $\beta_1 =$ slope of regression line

- Least-squares regression minimizes prediction error

$$Q = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i) \right]^2$$

for data points $(x_1, y_1), \ldots, (x_n, y_n)$

Linear relationships



Simple linear regression

- Coefficients of least-squares line

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}_n\bar{y}_n}{\sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2}$$

$$\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1\bar{x}_n$$

Simple linear regression

- Coefficients of least-squares line

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}_n\bar{y}_n}{\sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2}$$

$$\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1\bar{x}_n$$

- Mathematical derivation of regression coefficients
  - minimum of $Q(\beta_0, \beta_1)$ satisfies $\partial Q/\partial\beta_0 = \partial Q/\partial\beta_1 = 0$
  - leads to normal equations (system of 2 linear equations)

$$-2\sum_{i=1}^{n}\left[y_i - (\beta_0 + \beta_1 x_i)\right] = 0 \quad\Rightarrow\quad \beta_0 n + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$-2\sum_{i=1}^{n} x_i\left[y_i - (\beta_0 + \beta_1 x_i)\right] = 0 \quad\Rightarrow\quad \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

  - regression coefficients = unique solution $\hat{\beta}_0, \hat{\beta}_1$

The Pearson correlation coefficient

- Measuring the "goodness of fit" of the linear prediction
  - variation among observed values of Y = sum of squares S_y^2
  - closely related to (sample estimate for) variance of Y

$$S_y^2 = \sum_{i=1}^{n} (y_i - \bar{y}_n)^2$$

  - residual variation wrt. linear prediction: $S_{\text{resid}}^2 = Q$

The Pearson correlation coefficient

- Measuring the "goodness of fit" of the linear prediction
  - variation among observed values of Y = sum of squares S_y^2
  - closely related to (sample estimate for) variance of Y

$$S_y^2 = \sum_{i=1}^{n} (y_i - \bar{y}_n)^2$$

  - residual variation wrt. linear prediction: $S_{\text{resid}}^2 = Q$

- Pearson correlation = amount of variation "explained" by X

$$R^2 = 1 - \frac{S_{\text{resid}}^2}{S_y^2} = 1 - \frac{\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_n)^2}$$

Multiple linear regression

- Linear regression with multiple predictor variables

$$Y \approx \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

minimises

$$Q = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) \right]^2$$

for data points $(x_{11}, \ldots, x_{1k}, y_1),\ \ldots,\ (x_{n1}, \ldots, x_{nk}, y_n)$

- Multiple linear regression fits *n*-dimensional hyperplane instead of regression line

Multiple linear regression: The design matrix

- Matrix notation of linear regression problem

$$\mathbf{y} \approx \mathbf{Z}\beta$$

- "Design matrix" $\mathbf{Z}$ of the regression data

$$\mathbf{Z} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \ldots & y_n \end{bmatrix}'$$

$$\beta = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & \ldots & \beta_k \end{bmatrix}'$$

  - $\mathbf{A}'$ denotes transpose of a matrix; $\mathbf{y}, \beta$ are column vectors

General linear regression

- Matrix notation of linear regression problem

$$\mathbf{y} \approx \mathbf{Z}\beta$$

- Residual error

$$Q = (\mathbf{y} - \mathbf{Z}\beta)'(\mathbf{y} - \mathbf{Z}\beta)$$

General linear regression

- Matrix notation of linear regression problem

$$\mathbf{y} \approx \mathbf{Z}\beta$$

- Residual error

$$Q = (\mathbf{y} - \mathbf{Z}\beta)'(\mathbf{y} - \mathbf{Z}\beta)$$

- System of normal equations satisfying $\nabla_\beta Q = 0$:

$$\mathbf{Z}'\mathbf{Z}\beta = \mathbf{Z}'\mathbf{y}$$

General linear regression

- Matrix notation of linear regression problem

$$\mathbf{y} \approx \mathbf{Z}\beta$$

- Residual error

$$Q = (\mathbf{y} - \mathbf{Z}\beta)'(\mathbf{y} - \mathbf{Z}\beta)$$

- System of normal equations satisfying $\nabla_\beta Q = 0$:

$$\mathbf{Z}'\mathbf{Z}\beta = \mathbf{Z}'\mathbf{y}$$

- Leads to regression coefficients

$$\hat{\beta} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$$

General linear regression

- Predictor variables can also be functions of the observed variables $\rightarrow$ regression only has to be linear in coefficients β
- E.g. polynomial regression with design matrix

$$\mathbf{Z} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^k \\ 1 & x_2 & x_2^2 & \cdots & x_2^k \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^k \end{bmatrix}$$

corresponding to regression model

$$Y \approx \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k$$

Linear statistical models

- Linear statistical model (ϵ = random error)

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

  - $x_1, \ldots, x_k$ are not treated as random variables!
  - $\sim$ = "is distributed as"; $\mathcal{N}(\mu, \sigma^2)$ = normal distribution

Linear statistical models

- Linear statistical model (ϵ = random error)

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

  - $x_1, \ldots, x_k$ are not treated as random variables!
  - $\sim$ = "is distributed as"; $\mathcal{N}(\mu, \sigma^2)$ = normal distribution

- Mathematical notation:

$$Y \mid x_1, \ldots, x_k \sim \mathcal{N}\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \sigma^2\right)$$

Linear statistical models

- Linear statistical model (ϵ = random error)

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

  - $x_1, \ldots, x_k$ are not treated as random variables!
  - $\sim$ = "is distributed as"; $\mathcal{N}(\mu, \sigma^2)$ = normal distribution

- Mathematical notation:

$$Y \mid x_1, \ldots, x_k \sim \mathcal{N}\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \sigma^2\right)$$

- Assumptions
  - error terms ϵ_i are i.i.d. (independent, same distribution)
  - error terms follow normal (Gaussian) distributions
  - equal (but unknown) variance σ^2 = homoscedasticity

Statistical inference for linear models

- Model comparison with ANOVA techniques
  - Is variance reduced significantly by taking a specific explanatory factor into account?
  - intuitive: proportion of variance explained (like R^2)
  - mathematical: F statistic $\rightarrow$ p-value

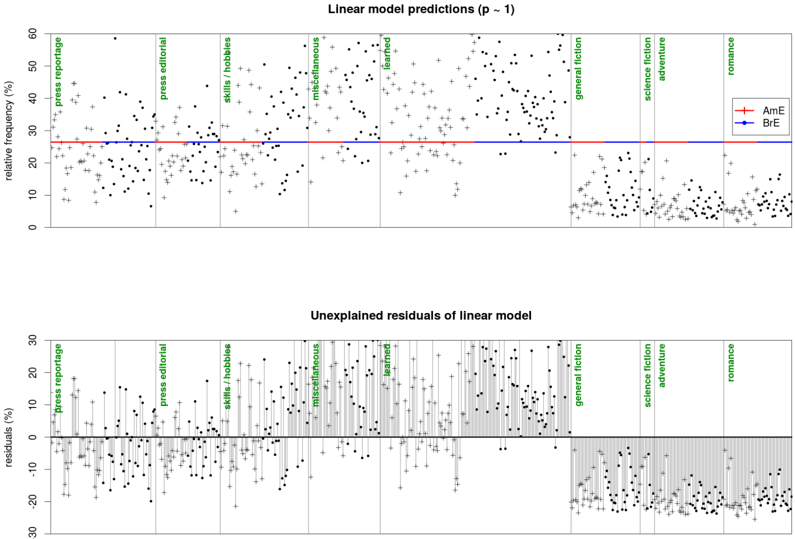
Statistical inference for linear models

- Model comparison with ANOVA techniques
  - Is variance reduced significantly by taking a specific explanatory factor into account?
  - intuitive: proportion of variance explained (like R^2)
  - mathematical: F statistic $\rightarrow$ p-value
- Parameter estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots$ are random variables
  - t-tests ($H_0 : \beta_j = 0$) and confidence intervals for β_j
  - confidence intervals for new predictions

Statistical inference for linear models

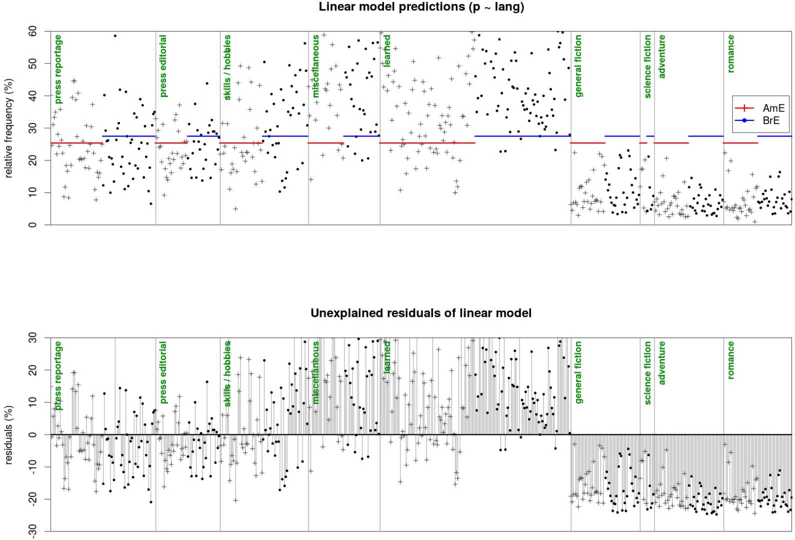
- Model comparison with ANOVA techniques
  - Is variance reduced significantly by taking a specific explanatory factor into account?
  - intuitive: proportion of variance explained (like R^2)
  - mathematical: F statistic $\rightarrow$ p-value
- Parameter estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots$ are random variables
  - t-tests ($H_0 : \beta_j = 0$) and confidence intervals for β_j
  - confidence intervals for new predictions
- Categorical factors: dummy-coding with binary variables
  - e.g. factor x with levels *low*, *med*, *high* is represented by three binary dummy variables $x_{\text{low}}, x_{\text{med}}, x_{\text{high}}$
  - one parameter for each factor level: $\beta_{\text{low}}, \beta_{\text{med}}, \beta_{\text{high}}$
  - NB: β_{low} is "absorbed" into intercept β_0
    model parameters are usually $\beta_{\text{med}} - \beta_{\text{low}}$ and $\beta_{\text{high}} - \beta_{\text{low}}$
  - mathematical basis for standard ANOVA

Linear model for passive frequencies



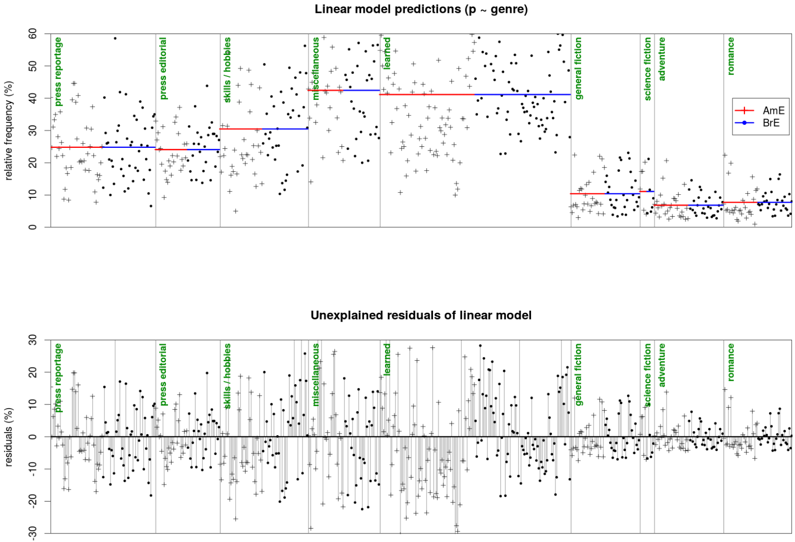
R2:
0

Linear model for passive frequencies



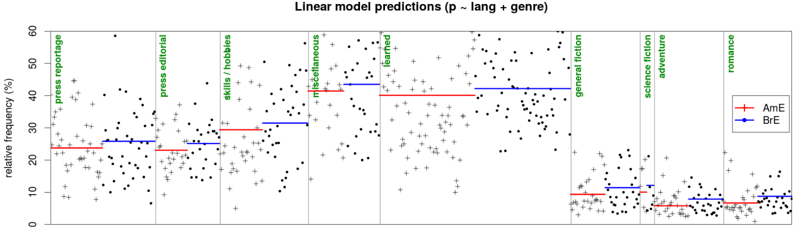
FQS: 189173.8 (R2: 0.36%)

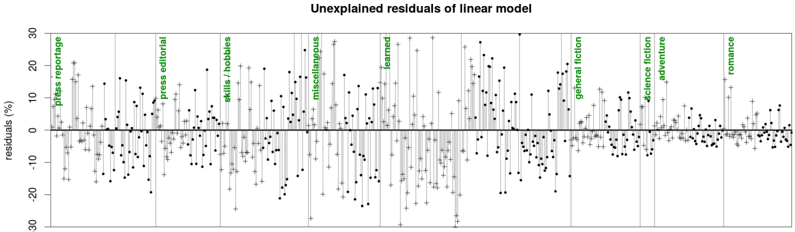
Linear model for passive frequencies



FQS: 77731 (R2: 59.06%)

Linear model for passive frequencies



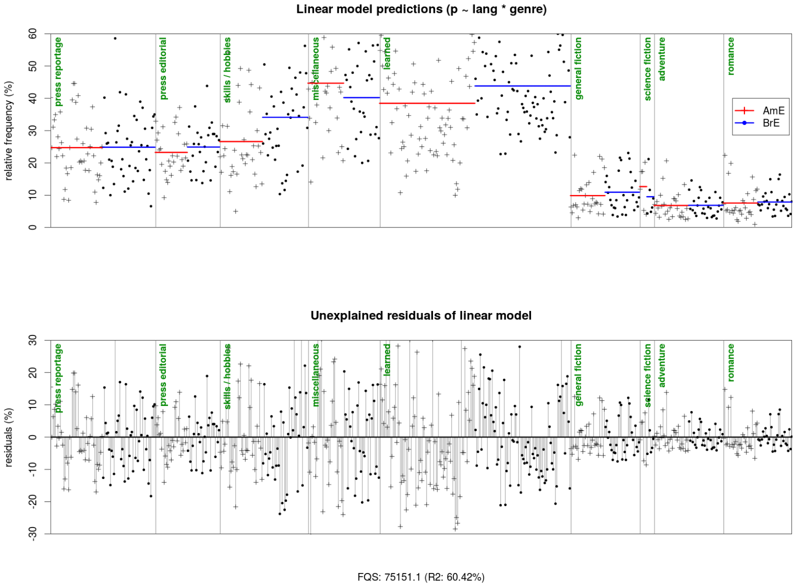


FQS: 77060.6 (R2: 59.41%)

Interaction terms

- Standard linear models assume independent, additive contribution from each predictor variable x_j $(j = 1, \ldots, k)$
- Joint effects of variables can be modelled by adding interaction terms to the design matrix (+ parameters)
- Interaction of numerical variables (interval scale)
  - interaction term for variables x_i and x_j = product $x_i \cdot x_j$
  - e.g. in multivariate polynomial regression:
    $Y = p(x_1, \ldots, x_k) + \epsilon$ with polynomial p over k variables
- Interaction of categorical factor variables (nominal scale)
  - interaction of x_i and x_j coded by one dummy variable for each combination of a level of x_i with a level of x_j
  - alternative codings e.g. to have separate parameters for independent additive effects of x_i and x_j
- Interaction of categorical factor with numerical variable

Linear model for passive frequencies



Generalised linear models

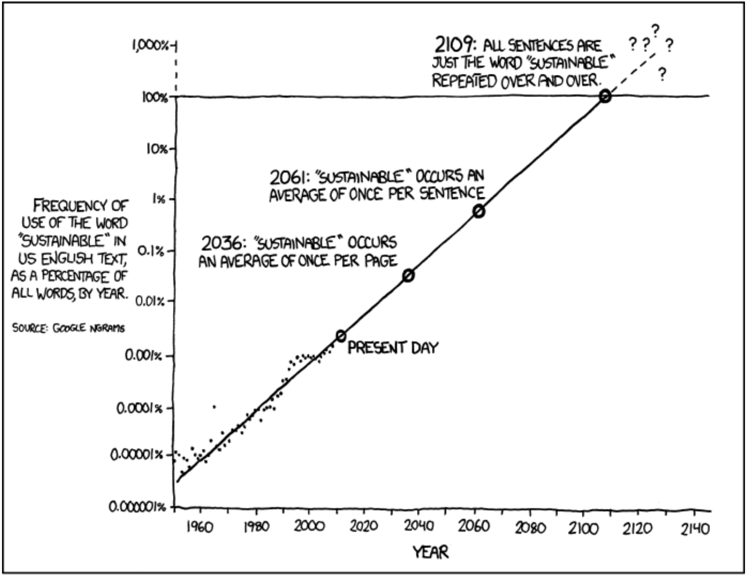
- Linear models are flexible analysis tool, but they ...
  1. only work for a numerical response variable (interval scale)
  2. assume independent (i.i.d.) Gaussian error terms
  3. assume equal variance of errors (homoscedasticity)
  4. cannot limit the range of predicted values

Generalised linear models

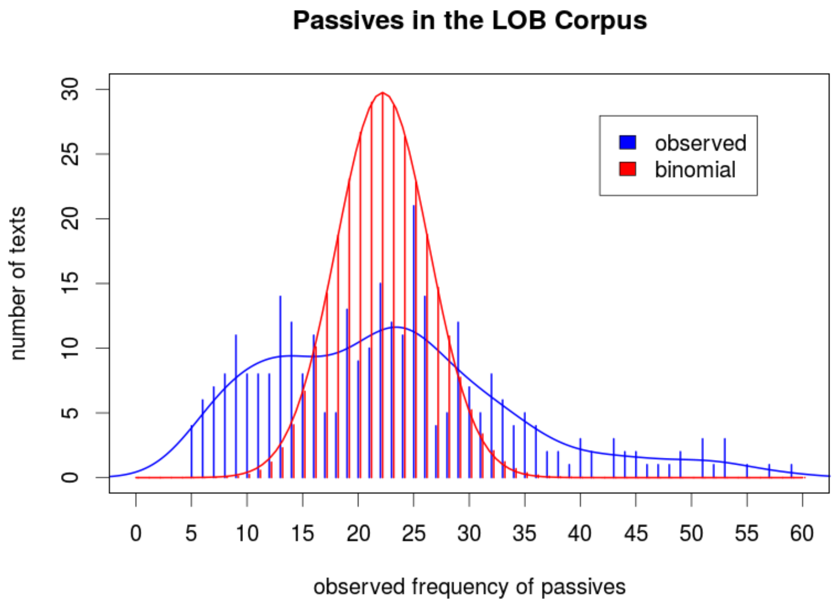
- Linear models are flexible analysis tool, but they ...
  1. only work for a numerical response variable (interval scale)
  2. assume independent (i.i.d.) Gaussian error terms
  3. assume equal variance of errors (homoscedasticity)
  4. cannot limit the range of predicted values
- Linguistic frequency data problematic in all four respects
  - each data point y_i = frequency f_i in one text sample
  - f_i are discrete variables with binomial distribution (or more complex distribution if there are non-randomness effects)
  - linear model uses relative frequencies $p_i = f_i / n_i$
  - Gaussian approximation not valid for small text size n_i
  - sampling variance depends on text size n_i and "success probability" π_i (= relative frequency predicted by model)
  - model predictions must be restricted to range $0 \leq p_i \leq 1$

$\Rightarrow$ General*ised* linear models (GLM)

Sustainability



THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.



Passives in the LOB Corpus
observed
binomial
number of texts
0
5
10
15
20
25
30
0
5
10
15
20
25
30
35
40
45
50
55
60
observed frequency of passives

Generalised linear model for corpus frequency data

- Sampling family (binomial)

$$f_i \sim B(n_i, \pi_i)$$

Generalised linear model for corpus frequency data

- Sampling family (binomial)

$$f_i \sim B(n_i, \pi_i)$$

- Link function (success probability $\pi \leftrightarrow$ odds θ)

$$\pi_i = \frac{1}{1 + e^{-\theta_i}}$$

Generalised linear model for corpus frequency data

- Sampling family (binomial)

$$f_i \sim B(n_i, \pi_i)$$

- Link function (success probability $\pi \leftrightarrow$ odds θ)

$$\pi_i = \frac{1}{1 + e^{-\theta_i}}$$

- Linear predictor

$$\theta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

Mixed Effects Models

a.k.a. hierarchical or multilevel modelling

- useful for situations where observations are clustered or grouped, e.g. by
  - speakers / writers
  - genres / registers
  - items
  - ...

Always include random effects for speaker and genre!

- purpose of mixed effects models: explain variance between groups of observations
- the group variables are so-called *random* effects
- coefficients for each level of random effects are not estimated (as this is done for *fixed* effects), but assumed to be random (and thus *predicted*)

Mixed Effects Models

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{u} + \boldsymbol{\epsilon}$$

$\boldsymbol{y}$	observations
X and Z	design matrices
$\boldsymbol{\beta}$	fixed effects
$\boldsymbol{u}$	random effects
$\boldsymbol{\epsilon}$	random errors

assumptions:

- $\mathbb{E}[\boldsymbol{y}] = X\boldsymbol{\beta}$
- $\mathbb{E}[\boldsymbol{u}] = 0 = \mathbb{E}[\boldsymbol{\epsilon}]$

solutions (assuming normality of $\boldsymbol{u}$ and $\boldsymbol{\epsilon}$)

- BLUE for $\boldsymbol{\beta}$
- BLUP for $\boldsymbol{u}$