



# Modeling count data

Stefan Hartmann  
HHU Düsseldorf

- Debunking some myths about statistical modeling
- Basics of Poisson and negative binomial regression
- Common problems: Overdispersion and zero-inflation
- Modelling strategies for overdispersed and zero-inflated count data

## Debunking some myths about statistical modelling

### ■ **Myth 1: There is one right model for every dataset.**

→ No: "All models are wrong but some are useful" (George Box)

### ■ **Myth 2: You can find the "right" model using a flowchart**

→ No: Each dataset presents a challenge on its own right, and modelling involves clearly stating your prior assumptions and expectations and "translating" them into a model.

### ■ **Myth 3: Statistical modelling is objective.**

→ No: Modelling always involves the researcher's choices and expectations.

- 1. The goal is to estimate **parameter values** (e.g., the number of speech errors for drunk vs. sober participants).
- 2. These parameters have **true (population) values**, which are approximated by taking a sample.
- 3. The **population** from which the sample is taken is **infinitely large**.
- 4. Samples drawn from the population are **representative** and **random** (e.g., samples are from all American English speakers, drawn randomly).

## Simple linear regression

- We can predict all kinds of data using the following general equation:

$$\text{Outcome}_i \sim (\text{model}) + \text{error}_i$$

- In regression, the model we fit is linear  $\rightarrow$  we summarize the data with a straight line.

## Simple linear regression

- Simple linear regression boils down to fitting a **straight line** to our data.
- In mathematical terms, a straight line can be defined by two parameters:
  - the **intercept** ( $b_0$ )
  - the **slope** ( $b_1$ )

## What is a statistical model?

- "A statistical model describes the relationship between one or more variables on the basis of another variable or variables." (Hilbe 2011)

$$Y = \beta_0 + \beta X + \varepsilon$$

$Y$ : response / outcome / dependent variable

$\beta$ : coefficient for  $X$ , i.e. slope describing the rate of change based on a one-unit change in  $X$ , holding other predictor values constant

$\beta_0$ : intercept, i.e. value of  $Y$  when  $X = 0$

$\varepsilon$ : error term

# Straight lines are (not) enough

## Why do we need Generalized Linear Models?

- Linear models are well-suited for data with a numeric/continuous response variable – e.g. reaction times
- (Caveat: in many cases, straight lines are not the best fit even for those data types; check out e.g. polynomial regression modeling for "curving" straight lines)
- But in many cases, we are dealing with **categorical** response variables – e.g., variant A vs. variant B; error vs. no error etc.



## Models and distributions

- Most statistical models assume that the observed data has been "generated" by a process following a certain distribution
- as such, both theoretical and empirical distributions play a role in modeling:
  - the distribution that we observe in the actual data,
  - a theoretical probability distribution that helps modelling the observed data.

# Straight lines are (not) enough

## The challenge

- We have to 'squeeze'  $[-\infty, +\infty]$  in the interval  $[0,1]$  in order to model probabilities.
- We can do this using the logistic function
- in logistic regression, the logit function serves as the so-called link function



## Basic idea of GLMs

- when fitting a simple linear model, the underlying assumption is that the process that generates the response variable follows a normal distribution
- GLMs generalize the linear model framework to incorporate data-generated processes that follow any distribution.

Simple linear regression:

$$y_i = \text{Normal}(\mu_i, \sigma)$$

Binomial logistic regression:

$$\begin{aligned} y_i &= \text{binomial}(N = 1, p) \\ &= \text{bernoulli}(p) \end{aligned}$$

Poisson logistic regression:

$$y \sim \text{Poisson}(\lambda)$$

Bernoulli  
distribution  
is a special  
case of the  
binomial  
distribution  
with  $N = 1$

## Logistic regression

- In the context of logistic regression, we are usually interested in modeling  $p$  as a function of one or more predictors.

- Remember our regression equation:

$$y_i = \beta_0 + \beta_1 x_i$$

- Ideally, we want different probabilities for different values of  $x$ .
- The regression equation  $\beta_0 + \beta_1 x_i$  can predict any continuous variable – but **probabilities** have to be between 0 and 1!
- As such, we have to 'squeeze' the output in the interval  $[0,1]$  – and that's what the logistic function does!

# Straight lines are (not) enough

## Link functions

**Linear regression:**  $I(\beta_0 + b_1 * x_i)$   
↓  
 $y_i = \text{Normal}(\mu_i, \sigma)$

**Logistic regression:**  $\text{logistic}(\beta_0 + b_1 * x_i)$   
↓  
 $y_i = \text{Bernoulli}(p)$

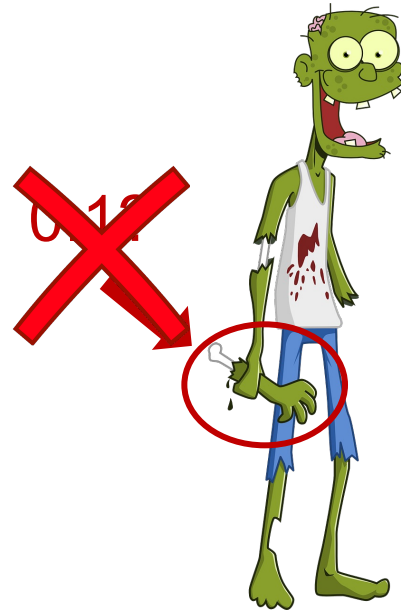
**Poisson regression:**  $\text{logistic}(\beta_0 + b_1 * x_i)$   
↓  
 $y_i = \text{Poisson}(\lambda_i)$

- Especially in corpus linguistics, we often work with count data
  - Token frequencies
  - Type frequencies
  - in general, all kinds of frequencies



# What are count variables?

- observations that can only take on nonnegative integer values



## What's a statistical model

- When there's more than one predictor, we have a separate beta and X value for each predictor:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon$$

- the predicted / expected mean value of the response is usually denoted as  $\hat{y}$ :

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

(Note that the error term is gone here because  $\hat{y}$  is the *estimated* value)



## Relationship to probability distributions

- All parametric statistical models are based on probability distributions
- in frequentist approaches, the probability density function (PDF) whose parameters we are trying to estimate is assumed to describe the population data

## Mean-variance relationship for selected count models

Model	Mean	Variance
Poisson	$\mu$	$\mu$
Linear Negative Binomial (NB1)	$\mu$	$\mu(1+\alpha)=\mu+\alpha\mu$
Negative binomial (NB2)	$\mu$	$\mu(1+\alpha\mu)=\mu+\alpha\mu^2$
Poisson inverse Gaussian	$\mu$	$\mu(1+\alpha\mu^2)=\mu+\alpha\mu^3$
Negative binomial-P	$\mu$	$\mu(1+\alpha\mu^p)=\mu+\alpha\mu^p$
Generalized Poisson	$\mu$	$\mu(1+\alpha\mu)^2=\mu+2\alpha\mu^3+\alpha^2\mu^3$

(=  $\lambda$  – there are notational differences, we will stick to  $\lambda$ /lambda in the rest of this tutorial because that's what R's inbuilt Poisson functions use)

## Mean-variance relationship for selected count models

Model	Mean	Variance
<b>Poisson</b>	$\mu$	$\mu$
Linear Negative Binomial (NB1)	$\mu$	$\mu(1+\alpha)=\mu+\alpha\mu$
<b>Negative binomial (NB2)</b>	$\mu$	$\mu(1+\alpha\mu)=\mu+\alpha\mu^2$
Poisson inverse Gaussian	$\mu$	$\mu(1+\alpha\mu^2)=\mu+\alpha\mu^3$
Negative binomial-P	$\mu$	$\mu(1+\alpha\mu^p)=\mu+\alpha\mu^p$
Generalized Poisson	$\mu$	$\mu(1+\alpha\mu)^2=\mu+2\alpha\mu^3+\alpha^2\mu^3$

(=  $\lambda$  – there are notational differences, we will stick to  $\lambda$ /lambda in the rest of this tutorial because that's what R's inbuilt Poisson functions use)

## Structure of a count model

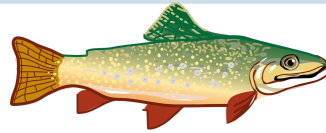
- Count models have the basic structure of a linear model, but the left-hand side of the equation is in log form.

$$\ln(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

or, inversely:

$$\lambda = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)$$

## Poisson model

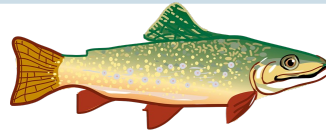


### ■ Poisson distribution:

- is bound by zero – no negative counts possible!
- discrete (categorical) distribution, i.e. only positive integers possible.
- one parameter  $\lambda$ , which specifies the rate of a count process
- unique feature of Poisson distribution: mean and variance are the same → the higher the mean, the greater the variance (equidispersion)
- Problem: equidispersion criterion hardly ever satisfied! Usually we face under- or, more commonly, overdispersion.

### ■ Poisson regression models $\lambda$ as a function of some predictors.

## Poisson model



### ■ Assumptions of Poisson models

- The response variable must be a count – this criterion is violated when a Poisson model is employed on a count response that does not include the possibility of having zero counts! (e.g., age in days, length of stay at the hospital)
- Observations are independent of one another
- No cell has substantially more or less than what is expected based on the mean of the empirical distribution (e.g., data should not have more zero counts than is expected based on a Poisson distribution with a given mean)
- The mean and variance of the model are approximately identical

## Overdispersion

- Poisson **overdispersion** occurs in data where the variability of the data is greater than the mean.
- The standard errors of an overdispersed model are biased and cannot be trusted.

## Zero-inflation

- Poisson and negative binomial models often underpredict the number of zeros → "excess zeros" problem
- Zero-inflation is when there are more zeros in the data than the distribution allows for.



## Zero truncation

- In some cases, zeros are theoretically impossible
  - e.g. age in days, languages per country, ...
- Just as zero-inflated data have more zeros than the model would predict, zero-truncated data have less

## Negative binomial model (NB2)

- Negative binomial distribution
  - Poisson-gamma mixture model
  - has an extra dispersion parameter  $d$  which is gamma-distributed
  - Negative binomial model adjusts for Poisson overdispersion, it cannot be used to model underdispersed count data.

## Assumptions

- The response, is a count consisting of nonnegative integers.
- As the value of  $\lambda$  increases, the probability of 0 counts decreases.
- $y$  must allow for the possibility of 0 counts.
- The fitted or predicted variable is the expected mean of the distribution of  $y$ .
- The variance is closely approximated as  $\lambda(1+\alpha\lambda) = \lambda + \alpha\lambda^2$
- A foremost goal of NB regression is to model data in which the value of the variance exceeds the mean, or the observed variance exceeds the expected variance.
- A well-fitted NB model has a dispersion statistic approximating 1.0 and an AIC/BIC and log-likelihood statistic less than alternative count models.
- The number of predicted counts is approximately the same as the number of observed counts across the distribution of  $y$ .

## What are exposure variables?

- Rates can be adjusted by exposure variables: for example, when counting the number of languages in a country, one has to control for a country's size.
- For exposure variables, the rate  $\lambda$  is split into two components:
  - the mean number of events  $\mu$ ,
  - per unit of exposure  $\tau$  'tau'
$$\log(\mu) = \beta_0 + \beta_1 \times \text{exposure variable} + \log(\tau)$$
- Caution: adding country size as an exposure variable is effectively saying that the average number of languages occurring in a country is directly proportional to the size of a country!
- in R: `offset(exposure variable)`



## Hands-on task

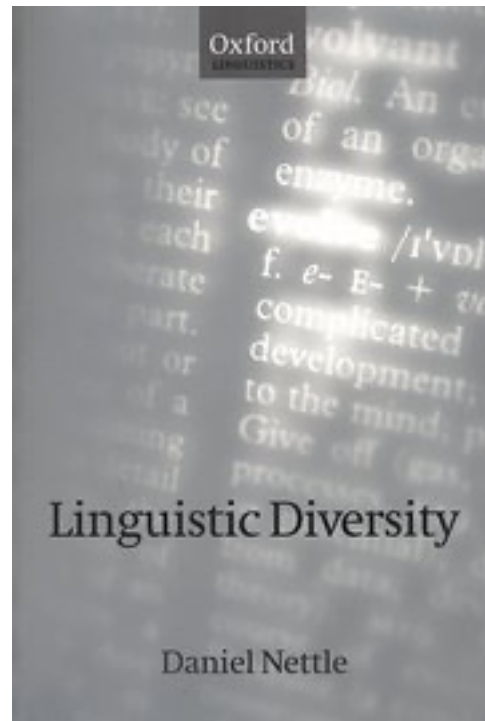
# Fake dataset

## Creating a fake dataset

■ ...

## Nettle (1999)

- hypothesis: linguistic diversity is correlated with climate factors
  - fertile environments → less reason to travel → less language contact → less linguistic diversity, and vice versa
- measured ecolocial risk using a country's Mean Growing Season (MGS) → how many months per year can you grow crops in the country?



## Winter, Perlman & Majid (2018)

- Frequency counts of sensory words from a variety of corpora (COHA, COCA, among others)



# Example study

---

## GraphVar corpus

- A-level exams annotated for spelling errors

- Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.
- Field, Andy, Jeremy Miles, & Zoë Field. 2012. *Discovering Statistics Using R*. Los Angeles: Sage.
- Hilbe, Joseph M. 2011. *Negative binomial regression*. 2nd ed. Cambridge, UK ; New York: Cambridge University Press.
- Levshina, Natalia. 2015. *How to do linguistics with R. Data exploration and statistical analysis*. Amsterdam, Philadelphia: John Benjamins.
- Nettle, Daniel. 1999. *Linguistic diversity*. Oxford ; New York: Oxford University Press.
- Winter, Bodo. 2019. *Statistics for linguists: an introduction using R*. New York: Routledge.