Modeling count data

Stefan Hartmann
HHU Düsseldorf

Course overview

- 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

Before we start...

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.

Basics of frequentist modeling

- 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).

(quoted from Sonderegger 2023)

The linear model

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 → we summarize the data with a straight line.

(Field et al. 2015: 246)

The linear model

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)**

Basics of statistical modelling

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

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

β_0: intercept, i.e. value of Y when $X = 0$

ε: 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.

Basics of statistical modeling

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' [–∞, +∞] 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

Straight lines are (not) enough

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: $y_i = Normal(\mu_i, \sigma)$

Binomial logistic regression: $y_i = binomial(N = 1, p)$

$= bernoulli(p)$

Poisson logistic regression: $y \sim Poisson(\lambda)$

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

(Winter 2021: 199f.)

Straight lines are (not) enough

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!

(Winter 2021: 199f.)

Straight lines are (not) enough

Link functions

Linear regression: $I(\beta_0 + b_1 * x_i)$

$$\downarrow$$

$$y_i = Normal(\mu_i, \sigma)$$

Logistic regression: $logistic(\beta_0 + b_1 * x_i)$

$$\downarrow$$

$$y_i = Bernoulli(p)$$

Poisson regression: $logistic(\beta_0 + b_1 * x_i)$

$$\downarrow$$

$$y_i = Poisson(\lambda_i)$$

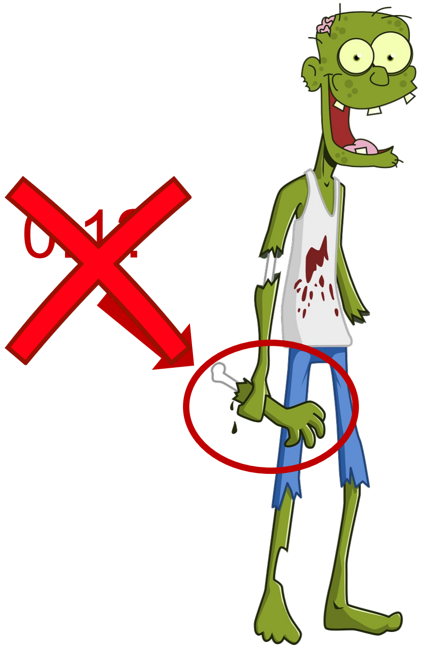
Count data

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

https://en.wikipedia.org/wiki/Count_von_Count#/media/File:Count_von_Count_kneeling.png

What are count variables?

- observations that can only take on nonnegative integer values



Statistical model

What's a statistical model

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

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

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

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

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

Statistical models

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

(Hilbe 2011: 8)

Statistical models

Mean-variance relationship for selected count models

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

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

(Hilbe 2011: 12)

Statistical models

Mean-variance relationship for selected count models

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

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

(Hilbe 2011: 12)

Inner workings of a count model

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.

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

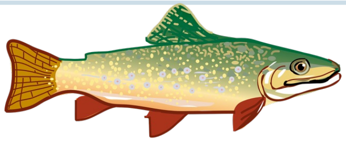
or, inversely:

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

(Hilbe 2014: 16)

Statistical models for count data

Poisson model

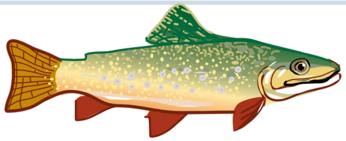


- Poisson distribution:
  - is bound by zero – no negative counts possible!
  - discrete (categorical) distribution, i.e. only positive integers possible.
  - one parameter λ, which specifies the rate of a count process
  - unique feature of Poisson distribution: mean and variance are the same → the 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.

(Hilbe 2014, Winter 2019)

Statistical models for count data

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 wth a given mean)
  - The mean and variance of the model are approximately identical

(Hilbe 2014: 36)

Overdispersion, zero-inflation, zero truncation

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.

(Hilbe 2011: 10)

Overdispersion, zero-inflation, zero truncation

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.

https://aosmith.rbind.io/2019/03/06/lots-of-zeros/

Overdispersion, zero-inflation, zero truncation

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

https://aosmith.rbind.io/2019/03/06/lots-of-zeros/

Statistical models for count data

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.

(Hilbe 2011: 10)

Negative binomial model

Assumptions

- The response, is a count consisting of nonnegative integers.
- As the value of λ 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.

(Hilbe 2014: 133)

Adding exposure variables

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 controle for a country's size.
- For exposure variables, the rate λ is split into two components:
  - the mean number of events μ,
  - per unit of exposure τ 'tau'

$$\log(\mu) = \beta_0 + \beta_0 \times exposure\ variable + \log(\tau)$$

- 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)`

Fitting count models

A recipe (although "cookbook approaches" are problematic ☺)

- Check if the data are count data.
- Try a Poisson model.
- Check for Poisson overdispersion – if there is overdispersion, try negative binomial model
- Check for zero-inflation (and, for small lambdas, zero-truncation)
  - Remember that NB zero-inflation is not the same as Poisson zero-inflation
- If necessary, explore which other count models are around (e.g. in Hilbe's book on modeling count data).

Hands-on task

Fake dataset

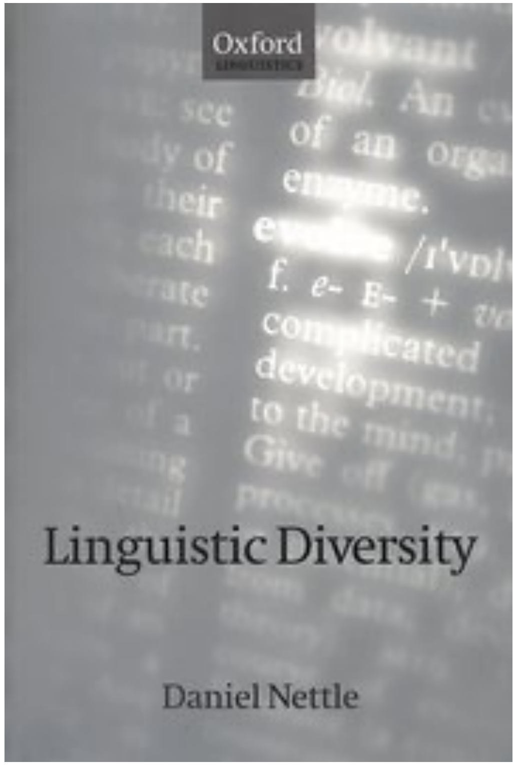
Creating a fake dataset

- To get a feel for the Poisson distribution, we first create a dataset with simulated Poisson-distributed data using the R function *rpois*.

Authentic 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?



(see Winter 2019: 34)

Example study

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
- https://graphvar.uni-bonn.de/

References

- 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.