Introduction to Generalized Linear Models
Part 2 of 3

Dr Rafael de Andrade Moral
Associate Professor of Statistics, Maynooth University

rafael.deandrademoral@mu.ie
https://rafamoral.github.io

Outline

- ~~The normal model: A recap~~
- ~~Models for binary data~~
- ~~Models for binomial data~~
- Models for multinomial data
- Models for count data
- Extensions: overdispersion models
- Extensions: zero-inflated models

Models for Multinomial Data

The Multinomial GLM

- Example: elephant grass grazing data

The Multinomial GLM

- We now are looking at multivariate extensions of the Bernoulli and binomial distributions
- For the sake of simplicity we will consider the *individual* case (denominator of 1)
- A Bernoulli random variable can assume only two values, success of failure
- Recall the Bernoulli variable Y_i:

$$Y_i = \begin{cases} 1, & \text{if success} \\ 0, & \text{if failure} \end{cases}$$

- We can re-write it as a bivariate response variable $\mathbf{Y}_i$ such that:

$$\mathbf{Y}_i = \begin{cases} (1,0)^\top, & \text{if success} \\ (0,1)^\top, & \text{if failure} \end{cases}$$

The Multinomial GLM

- The vectors are of dimension 2 because there are only 2 possible categories of response
- Extending this to K possible response categories, we have:

$$Y_{ik} = \begin{cases} 1, & \text{if category } k \\ 0, & \text{otherwise} \end{cases}$$

- As an example, for $K = 3$ categories, this translates to

$$\mathbf{Y}_i = \begin{cases} (1,0,0)^\top, & \text{if category 1} \\ (0,1,0)^\top, & \text{if category 2} \\ (0,0,1)^\top, & \text{if category 3} \end{cases}$$

The Multinomial GLM

- We write $Y_i \sim \text{Multinomial}(\boldsymbol{\pi}_i)$, where $\boldsymbol{\pi}_i = (\pi_1, \pi_2, \ldots, \pi_K)^\top$ is now a vector of probabilities
- Recall that probabilities are bounded in the $(0, 1)$ interval
- Therefore, we must make the restriction

$$\sum_{j=1}^{K} \pi_{ij} = 1$$

- This commonly translates as taking one of the categories as the *reference category*
- In the `nnet` package in R which implements this model as the `multinom` function, the first category is taken as reference, and therefore

$$\pi_{i1} = 1 - \sum_{j=2}^{K} \pi_{ij}$$

The Multinomial GLM

- There are many variations and extensions to multinomial GLMs
- We will now see the basic building blocks of the so-called *generalized logits model*
- We have $K-1$ logits for a multinomial model with K categories
- Taking the first category as reference, we have

$$\eta_{ki} = \log\left(\frac{\pi_{ki}}{\pi_{1i}}\right) = \beta_{0k} + \beta_{1k}x_{1i} + \ldots + \beta_{pk}x_{pi}$$

- This means that a generalized logits model estimates $(K-1)\times(p+1)$ parameters

The Multinomial GLM

- For our grazing experiment example, we have $K = 3$ categories, and therefore, 2 logits:

$$\begin{aligned} \log\left(\frac{\pi_{2i}}{\pi_{1i}}\right) &= \beta_{02} + \beta_{12}\text{month}_i \\ \log\left(\frac{\pi_{3i}}{\pi_{1i}}\right) &= \beta_{03} + \beta_{13}\text{month}_i \end{aligned}$$

- $\beta_{02}, \beta_{12}, \beta_{03},$ and β_{13} are the four parameter estimates we see when we fit the model using `multinom`

The Multinomial GLM

- It is easy to obtain fitted probabilities for each category based on predictors:

$$\log\left(\frac{\pi_{ki}}{\pi_{1i}}\right) = \eta_{ki} \Leftrightarrow \frac{\pi_{ki}}{\pi_{1i}} = e^{\eta_{ki}} \Leftrightarrow \pi_{ki} = \pi_{1i}e^{\eta_{ki}}$$

- We then use the restriction

$$\pi_{1i} = 1 - \sum_{j=2}^{K} \pi_{ji} = 1 - \sum_{j=2}^{K} \pi_{1i}e^{\eta_{ji}}$$

- Solving for π_{1i} yields

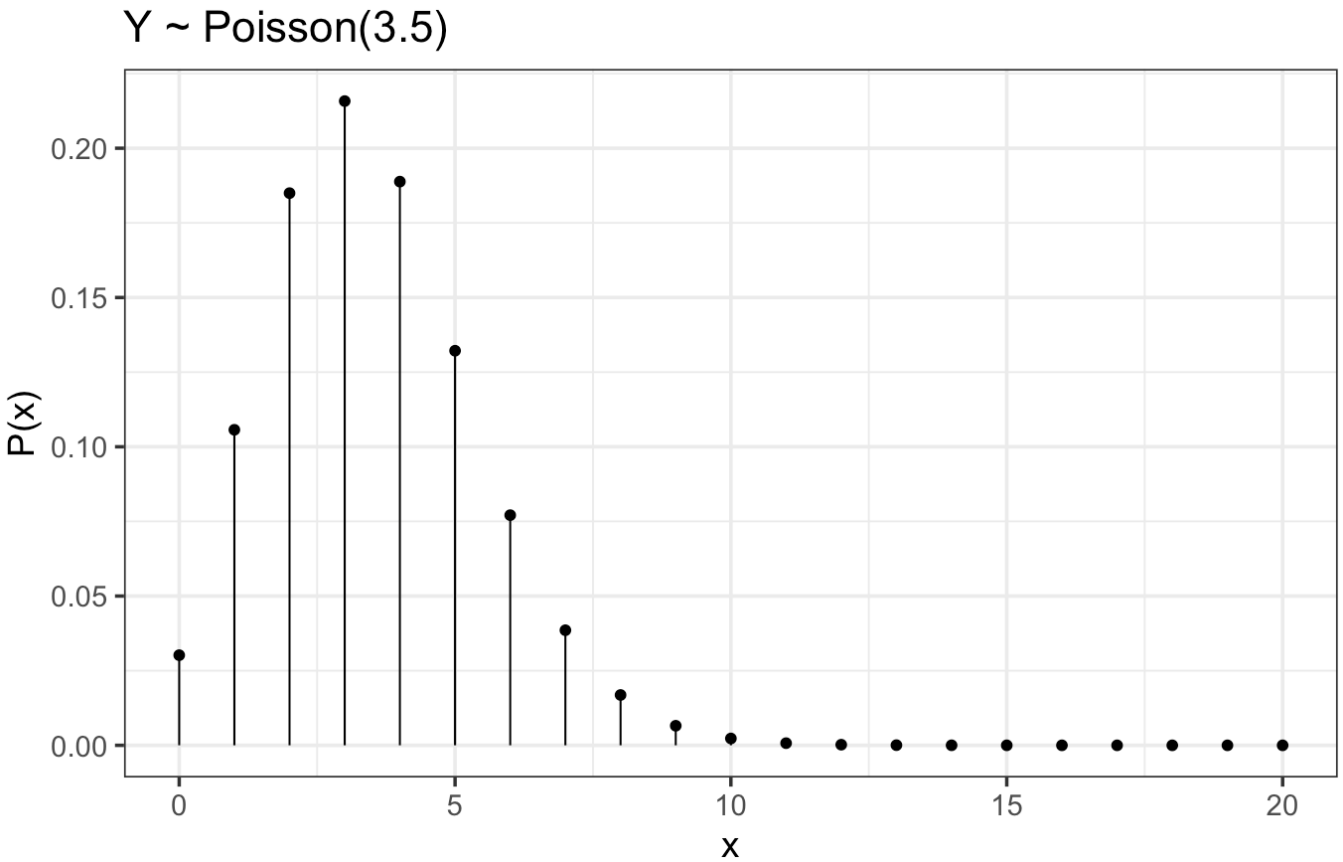
$$\pi_{1i} = \frac{1}{1 + \sum_{j=2}^{K} e^{\eta_{ki}}}$$

- Using the two equations in blue above we are able to recover all probabilities; we compute the reference probability, and then use it to compute the probabilities for all other categories

Models for Count Data

The Poisson GLM

- The Poisson distribution is a discrete probability distribution over the non-negative integers $0, 1, 2, \ldots$

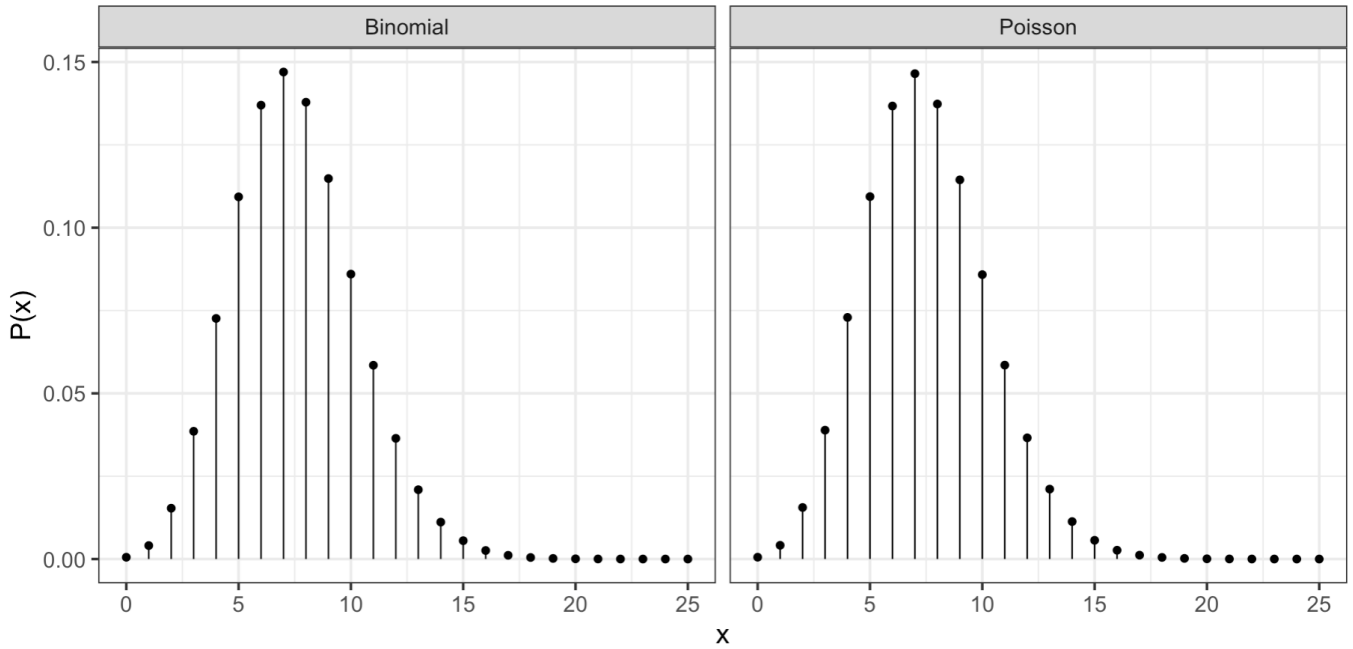


The Poisson GLM

- The number of times an event occurs is a common form of data, e.g.
  - number of identified animal species in quadrats at different locations
  - number of eggs laid per female of an insect species
  - number of customers purchasing at a particular shop per day
  - number of people in each cell of a contingency table summarising survey responses
- Simplest assumption of the underlying process: counts arise at some average *rate* λ, occurring in a fixed interval of time or space
- The Poisson model offers a starting point for analysis

The Poisson GLM

- The Poisson distribution can be seen as the limit of a binomial distribution with a small probability of success very large m
- We have that under these conditions Binomial$(m, \pi) \approx$ Poisson$(m\pi)$
- Example: $m = 1000$ and $\pi = 0.0075$:



The Poisson GLM

- Consider events over time
- Write λ for the average rate per unit time
- The distribution for the number of events Y in an interval of length t is Poisson(λt), with probability mass function

$$P(Y = y) = \frac{e^{-\lambda t}(\lambda t)^y}{y!}, \qquad y = 0, 1, 2, \ldots$$

- We have that

$$\mathrm{E}(Y) = \mathrm{Var}(Y) = \lambda t = \mu$$

- This is often referred to as *equidispersion*
- In many simple applications, counts will be observed over identical time periods, areas, etc.

The Poisson GLM

- For a random sample of n counts Y_i, $i = 1, \ldots, n$, with possibly different underlying rates μ_i, we have

$$Y_i \sim \text{Poisson}(\mu_i)$$

- We typically model the rates μ_i in terms of observed explanatory variables
- Since $\mu_i > 0$, the most used link function is the *log link*:

$$\log \mu_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

The Poisson GLM

- Example: number of visits to the doctor

The Poisson GLM

- Interpreting the coefficients from a Poisson log-linear model is intuitive
- Since we use the log-link, the inverse link function is the exponential, i.e.

$$\mu_i = e^{\beta_0+\beta_1 x_{1i}+\cdots+\beta_p x_{pi}}$$

- Take a model with a single continuous predictor, i.e. $\mu_i = e^{\beta_0+\beta_1 x_i}$
- For a one-unit increase in the predictor we have

$$\begin{aligned} \mu_i^+ &= e^{\beta_0+\beta_1(x_i+1)} \\ &= e^{\beta_0+\beta_1 x_i \beta_1} \\ &= e^{\beta_0+\beta_1 x_i} e^{\beta_1} \end{aligned}$$

- Therefore, $\mu_i^+ = \mu_i e^{\beta_1}$, i.e. e^{β_1} is the multiplicative effect of an increase in one unit to the predictor variable

Offset variables

- In many situations the length of time during which events are measured varies across individuals
  - e.g.[1] in the doctor visits example, one patient visited the GP 6 times in one year, another 15 times in five years
  - e.g.[2] suppose we monitor people's drinking at social occasions; we find that three people drink 12, 7 and 3 drinks over the course of 7, 5 and 2 hours, respectively
- In such cases, the notation $Y_i \sim \text{Poisson}(\lambda_i t_i)$ is helpful
- We have:

$$\log \mu_i = \log(\lambda_i t_i) = \log \lambda_i + \log t_i$$

- Here, $\log \lambda_i$ is modelled with covariates, and the term $\log t_i$ is referred to as *offset*

Offset variables

- The model is then

$$\begin{aligned} Y_i &\sim \text{Poisson}(\mu_i) \\ \log \mu_i &= \underbrace{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}}_{\log \lambda_i} + \underbrace{\log t_i}_{\text{offset}} \end{aligned}$$

- Note that the offset term can be seen as adding a covariate to the linear predictor associated to a β coefficient equal to 1
- Example: the insurance dataset

Goodness-of-fit Based on Residual Deviance

- For the Poisson and binomial GLMs, one way of checking whether the model fits the data well is to look at the residual deviance
- For a well fitted model, the residual deviance should be approximately equal to the $n - p$, the number of residual d.f.
- This is because, asymptotically, the residual deviance has a χ^2 distribution with $n - p$ d.f.
- If the residual deviance is much larger than expected, this indicates lack-of-fit, which may be due to omission of important predictors, or simply extra variability not accounted for by the model
- Example: *Sitophilus zeamais* progeny data

Half-Normal Plots with a Simulated Envelope

- Another way of empirically checking whether a GLM (or any[1] statistical model, in fact), is to use a graphical technique called *half-normal plot with a simulated envelope*
- It consists in plotting ordered residuals in absolute value versus expected order statistics of the half-normal distribution, and adding a simulated envelope based on the fitted model
- The envelope serves as a guide to what expect if the observed data are a plausible realisation of the fitted model
- In R it is implemented as the `hnp` package
- If the majority of points lie within the simulated envelope, the model can be considered to be well-fitted to the data

[1] as long as response variables can be simulated from it