Count regression models

B.D. in Business Administration and Economics
Course in Quantitative Methods III

Rosario Barone
University of Rome "Tor Vergata"

rosario.barone@uniroma2.it

GLM routine

More specifically, GLMs are generalization of linear models for situations in which the outcome is not Gaussian, summarized as follows:

- specify distribution for the dependent variable $f(Y|\theta)$;
- specify a link function $g(\cdot)$;
- specify a linear predictor.

Assumption on Y

The distribution of the dependent variable $f(Y|\theta)$ is assumed to belong to the exponential family. Some examples:

- Normal
- binomial (with fixed n)
- multinomial (with fixed n)
- Poisson
- negative binomial (with fixed number of failures).

Model definition

We define the distribution $f(Y|X)$, with mean $\boldsymbol{\mu}$ of the depending on the independent variables, X, through:

$$\mathrm{E}(Y|X) = \boldsymbol{\mu} = g^{-1}(X\boldsymbol{\beta})$$

where:

- $E(Y|X)$ is the expected value of Y conditional on X;
- $X\boldsymbol{\beta}$ is the linear predictor;
- g is the link function.

The variance is typically a function, V, of the mean:

$$\mathrm{var}(Y|X) = \nu(g^{-1}(X\boldsymbol{\beta})).$$

However, by choosing ν as a distribution of the exponential family we get a more flexible model.

Count outcome dependent variable

Let Y denote a count variable ($Y \in \mathbb{N}$). The simplest counts probability distribution is the Poisson distribution, with probability mass function:

$$f(y;\mu) = \frac{e^{-\mu}\mu^y}{y!} = \exp(-\mu)\left(\frac{1}{y!}\right)\exp\left(y \log \mu\right).$$

This has natural exponential form

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)\exp\left\{y_i Q(\theta_i)\right\}$$

with:

- $\theta = \mu$
- $a(\mu) = \exp(-\mu)$
- $b(y) = 1/y!$
- $Q(\mu) = \log \mu$

Count outcome dependent variable

The natural parameter is $\log \mu$, so the canonical link function is $\eta = \log \mu$ (log link).

Let now Y denote a count response variable ($Y \in \mathbb{N}$) and let $\mathbf{x} = (x_1, \ldots, x_k)$ be the vector of observed covariates.

We define a *Poisson loglinear model*, by using a log link to define the relationship between the response variable and the covariates, that is:

$$\log \mu_i = \sum_j \beta_j x_{ij}.$$

Poisson loglinear model

- The Poisson distribution has a positive mean μ;
- although a GLM can model a positive mean using the identity link, it is more common to model $\log \mu$:
  - $\log \mu \in \mathbb{R}$;
  - $\log \mu$ is the natural parameter for the Poisson distribution;
  - the log link is the canonical link for a Poisson GLM.

A Poisson loglinear GLM assumes a Poisson distribution for Y and uses the log link.

Poisson loglinear model

Let consider the simplest case, with a single with explanatory variable X. The Poisson loglinear model is:

$$\log \mu = \alpha + \beta x.$$

The mean satisfies the exponential relationship

$$\mu = \exp(\alpha + \beta x) = e^{\alpha} \left(e^{\beta}\right)^{x}.$$

A 1-unit increase in x has a multiplicative impact of e^{β} on μ: The mean at $x + 1$ equals the mean at x multiplied by e^{β}.

Overdispersion for Poisson GLMs

- overdispersion is not an issue in ordinary regression with normally distributed Y, because that distribution has a separate parameter to describe variability.
- Count data often show greater variability than the Poisson allows: the variances are much larger than the means, whereas Poisson distributions have identical mean and variance.
- A common cause of overdispersion is subject heterogeneity.
- When data does not have good fitting with the Poisson distribution, ML estimates are still consistent but standard errors are incorrect (underestimated).

Overdispersion for Poisson GLMs

How to deal with overdispersion?

- quasi-likelihood approach (as in the binomial case);
- Negative binomial model.

Negative Binomial GLMs

Università di Roma Tor Vergata

The negative Binomial GLMs are an extension of the Poisson GLM that has an extra parameter and accounts better for overdispersion.

Let consider a count variable, $Y \in \mathbb{N}$; the negative binomial distribution has density

$$f(y; k, \mu) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{k}{k+\mu}\right)^k \left(1 - \frac{k}{k+\mu}\right)^y;$$

Negative Binomial GLMs

Università di Roma Tor Vergata

- $E(Y) = \mu$;
- $var(Y) = \mu + \mu^2/k$.
- k^{-1} is a *dispersion parameter*, and as $k \to \infty$, the distribution converges to the Poisson(μ).
- For k fixed, one can express the negative binomial density in natural exponential family form and a model with negative binomial random component is a GLM;
- A variety of link functions are possible: most common is the log link.

Reminder: likelihood equations for GLM

For n independent observations, the likelihood function is:

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log(f(y_i;\theta_i,\psi))$$

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{n} c(y_i,\phi)$$

After some analytics, we get the *likelihood equations*:

$$\frac{\mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\mathrm{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0.$$

Likelihood equations for a Poisson GLM

The general Poisson loglinear model has the matrix form

$$\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$

By assuming the link to be $\eta_i = \log \mu_i$ (log link), we have:

- $\mu_i = \exp(\eta_i)$ and $\partial \mu_i / \partial \eta_i = \exp(\eta_i)$;
- $var(Y_i) = \mu_i$.

Therefore, the likelihood equations for the Poisson GLM are:

$$\frac{\mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\mu_i} \exp(\eta_i) = 0,$$

since $\mu_i = \exp(\eta_i)$,

$$\frac{\mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} (y_i - \mu_i)x_{ij} = 0.$$

Overdispersion for Poisson GLMs and Quasi-likelihood

An alternative way (with respect to the negative Binomial model) for handling overdispersion for counts is the quasi-likelihood approach.

Let $Y_i \sim Pois(\mu_i)$, then:

- $E(Y_i) = \mu_i$,
- $var(Y_i) = \mu_i$.

A simple quasi-likelihood approach uses the alternative variance function

$$\nu(\mu_i) = \phi\mu_i,$$

overdispersion occours when $\phi > 1$.

Estimates are equal to the *ML* case for the Poisson response (ϕ drops out from likelihood equations and it is estimated separately) and the standard errors multiply by $\sqrt{\phi}$.

Poisson GLMs diagnostics

- Deviance of the model
- Likelihood ratio
- Statistics on the residuals (RSS-like statistics):
  - deviance residuals
  - Pearson residuals

Count data regression models in R

- Poisson GLM: `glm(formula, family = poisson, data, ...)`

- Quasi-Likelihood approach for Poisson GLM: `glm(formula, family = quasipoisson, data, ...)`

- Negative Binomial GLM:
  - `library: MASS`
  - `glm.nb(formula, data, ..., link = log)`