Poisson regression

Statistics III - CdL SSE

Tommaso Rigon
Università degli Studi di Milano-Bicocca

Homepage

Alan Agresti

Distinguished Professor Emeritus, University of Florida

- GLMs for **count data** are very common and have theoretical connections with binary and binomial models.
- This unit focuses on **Poisson regression models**.
- I will not cover the analysis of **contingency tables**.
- Such a topic is nonetheless discussed in the textbook but is not part of the exam.
- The most important aspects have been already covered in Unit B.

The content of this Unit is covered in **Chapter 5** of Salvan et al. (2020). Alternatively, see **Chapter 7** of Agresti (2015).

Notation and recap

- In a **Poisson regression** model, we observe Y_i independent Poisson random variables, so that

$$Y_i \stackrel{\text{ind}}{\sim} \text{Poisson}(\mu_i), \qquad g(\mu_i) = \eta_i = \boldsymbol{x}_i^T\beta, \qquad i = 1, \dots, n.$$

- The **canonical link** is $g(\cdot) = \log(\cdot)$, which implies a **multiplicative structure**

$$\mu_i = \exp(\boldsymbol{x}_i^T\beta) = \exp(\beta_1)^{x_{i1}} \times \cdots \times \exp(\beta_p)^{x_{ip}} = \prod_{j=1}^{p} \alpha_j^{x_{ij}}, \qquad \alpha_j = \exp(\beta_j).$$

- Under the canonical link, the **likelihood equations** are

$$\sum_{i=1}^{n} (y_i - \mu_i)x_{ir} = 0, \qquad r = 1, \dots, p.$$

  The solution therefore has a nice interpretation as a **method of moments** estimator, in that

$$\sum_{i=1}^{n} y_i x_{ir} = \sum_{i=1}^{n} \mathbb{E}(Y_i) x_{ir}, \qquad r = 1, \dots, p.$$

Interpretation of the regression coefficients

- Under the **logarithmic link**, the mean has a multiplicative structure, namely

$$\mu_i = \exp(\boldsymbol{x}_i^T \beta) = \exp(\beta_1)^{x_{i1}} \times \cdots \times \exp(\beta_p)^{x_{ip}} = \prod_{j=1}^{p} \alpha_j^{x_{ij}}, \qquad \alpha_j = \exp(\beta_j).$$

- As a result, a **unitary increase** of the jth covariate from x_{ij} to $x_{ij} + 1$ has the following impact on the new mean, say μ_{new}

$$\mu_{\text{new}} = \alpha_1^{x_{i1}} \times \cdots \times \alpha_j^{x_{ij}+1} \times \cdots \times \alpha_p^{x_{ip}} = \alpha_j \left(\alpha_1^{x_{i1}} \times \cdots \times \alpha_p^{x_{ip}}\right) = \alpha_j \mu_i.$$

  In other words, the regression parameters, once exponentiated, can be interpreted as **relative changes** of the mean, namely

$$\alpha_j - 1 = \exp(\beta_j) - 1 = \frac{\mu_{\text{new}} - \mu_i}{\mu_i}.$$

> The interpretation in terms of relative changes is a consequence of the **logarithmic link** function. Therefore, the same interpretation applies whenever this link is used, including the Gamma GLM.

Exposure rate

- Often the expected value of a response count Y_i is proportional to an index t_i, the **exposure**.
- For instance, t_i might be an amount of time and/or a population size, such as in modeling crime counts for various cities. Or, it might be a spatial area, such as in modeling counts of plant species.
- In these case, the **sample rate** is Y_i/t_i, with expected value μ_i/t_i. With explanatory variables, a model for the expected rate under a **logarithmic link** has the form

$$\log\left(\frac{\mu_i}{t_i}\right) = \boldsymbol{x}_i^T\beta, \qquad \Longrightarrow \qquad \log\mu_i = \boldsymbol{x}_i^T\beta + \log t_i,$$

  Because $\log(\mu_i/t_i) = \log\mu_i - \log t_i$, the model makes the adjustment $\log t_i$ to the linear predictor. This adjustment term is called an **offset**, implemented in **R** using the `offset` option.
- The fit corresponds to using $\log t_i$ as an **explanatory variable** in the linear predictor for $\log(\mu_i)$ and **forcing its coefficient** to equal 1.
- Summarising, for this model, the response counts $Y_i \sim \text{Poisson}(\mu_i)$ satisfy

$$\mu_i = t_i\exp(\boldsymbol{x}_i^T\beta).$$

  The mean has a proportionality constant for t_i that depends on the values of the covariates.

Overdispersion

- In Poisson regression the main **assumption** is that $Y_i \sim \text{Poisson}(\mu_i)$, implying that

$$\text{var}(Y_i) = \mu_i,$$

  where implicitly we have set $\phi = 1$.

- However, from the analysis of the residuals or by computing the X^2 statistic we may realize that the data present **overdispersion**, namely the **correct model** is such that

$$\text{var}(Y_i) = \phi\mu_i,$$

  with $\phi > 1$. This implies that the Poisson regression model is **misspecified**.

- The two most common solutions to overdispersion are the following:
  i. the usage of **quasi-likelihoods**;
  ii. using another parametric distribution; a typical choice is the **negative-binomial**.

Zero-inflation

- In practice, the frequency of **zero outcomes** is often **larger than expected** under a Poisson regression.
- Because the **mode** of a Poisson distribution is the integer part of its mean, a Poisson GLM can be inadequate when the mean is relatively large but the modal response is 0.
- Such data are called **zero-inflated**. This often occurs when:
  - many subjects have a true zero response (structural zeros), and
  - many others have positive counts, so the overall mean is not near zero.
- Example: the number of times individuals report exercising (e.g., going to a gym) in the past week:
  - some people exercise frequently,
  - some exercise occasionally but not in the past week (a random zero),
  - others never exercise (a structural zero),
- The two most common solutions to zero-inflation are the following:
  i. Zero-inflated Poisson (ZIP) model, a **mixture model**;
  ii. Hurdle models (model zero vs nonzero first, then model the remaining data).

References

Agresti, A. (2015), *Foundations of Linear and Generalized Linear Models*, Wiley.

Salvan, A., Sartori, N., and Pace, L. (2020), *Modelli lineari generalizzati*, Springer.