IDS 702: Module 3.1

Poisson regression

Dr. Olanrewaju Michael Akande

Generalized linear models

- As we've seen over the last few modules, we may often need to work with outcome variables that are not continuous.
- Clearly, the standard linear regression will not suffice in those situations.
- Specifically, we saw how to use logistic and probit regression to handle binary response variables.
- In other scenarios however, our outcome variable will not be binary either.
- How should we handle that?

GENERALIZED LINEAR MODELS

- For example, we may want to predict
  - Whether someone prefers product A, B, or C (nominal)
  - Political ideology on an ordered 3 scale outcome, such as "very liberal", "moderate", "very conservative" (ordinal)
  - The number of times an event happens (counts)
- The classes of models we will use to handle these types of responses are referred to as generalized linear models (GLMs).
- Note that GLMs includes the linear, logistic and probit regressions we already covered.

Components of GLMs

Generally, GLMs have three major components:

1. The random component describes the randomness of the outcome variable Y through a pdf or pmf f, with parameter θ_i. That is,

$$y_i|\boldsymbol{x}_i \sim f(y_i|\theta_i) \quad \text{OR} \quad y_i|\boldsymbol{x}_i \sim f(y_i;\theta_i) \quad \text{OR} \quad y_i|\boldsymbol{x}_i \sim f(\theta_i)$$

2. The systematic component defines a linear component of the predictors. That is, for each observation i,

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

3. The link function g connects the random and systematic components through $\mu_i = \mathbb{E}[Y_i]$, that is

$$\eta_i = g(u_i)$$

where g is a monotonic and differentiable function (for those with some math background).

In standard linear regression, g is the identity link $n_i = g(u_i) = u_i$, whereas in logistic regression, g is the logit function.

Poisson regression

- Suppose you have count data (non-negative integers) as your response variable.
- For example, we may want to explain the number of c-sections carried out in hospitals using potential predictors such as
  - hospital type, that is, private vs public
  - location
  - size of the hospital
- The models we have covered so far are not adequate for count data.
- While this is generally the case, there are instances where linear regression, with some transformations (especially taking logs) on the response variable, might still work reasonably well for count data.
- Thus, one can attempt to fit a linear regression model first, check to see if the assumptions of the model are violated, and then move on to a more appropriate model if needed.

Poisson regression

- A good distribution for modeling count data with no limit on the total number of counts is the Poisson distribution.
- Why would the Binomial distribution be inappropriate when there is no limit on the total number of counts?
- The Poisson distribution is parameterized by λ and the pmf is given by

$$\Pr[Y = y] = \frac{\lambda^y e^{-\lambda}}{y!}; \quad y = 0, 1, 2, \ldots; \quad \lambda > 0.$$

- An interesting feature of the Poisson distribution is.

$$\mathbb{E}[Y = y] = \mathbb{V}[Y = y] = \lambda.$$

- When our data fails this assumption, we may have what is known as over-dispersion and may want to consider the Negative Binomial distribution instead, or try a Bayesian specification (STA 602!).
- With no predictors, the best guess for λ is the sample mean, that is, $\hat{\lambda} = \sum_{i=1}^{n} \frac{y_i}{n}$.

Poisson regression

- With predictors, we want to index λ with i, where each λ_i is a function of $\boldsymbol{x}_i$. We can therefore write the random component of this glm as

$$y_i|\boldsymbol{x}_i \sim \text{Poisson}(\lambda_i); \quad i = 1, \ldots, n.$$

- We must ensure that $\lambda_i > 0$ at any value of $\boldsymbol{x}_i$, therefore, we need a link function that enforces this. A natural choice is the natural logarithm, so that we have

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}.$$

- Combining these pieces give us our full mathematical representation for the Poisson regression.

- In R, use the `glm` command but set the option `family = "poisson"`.

- Clearly, λ_i has a natural interpretation as the "expected count", and

$$\lambda_i = e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}}$$

means that we can interpret the e^{β_j}'s as multiplicative effects on the expected counts.

Poisson regression

- For predictions, we can look at the expected counts, that is,

$$\hat{\lambda}_i = e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip}}$$

- Interpretation of e^{β_j}:
  - For continuous x_j: the expected count of Y increases by a multiplicative factor of e^{β_j} when increasing x_j by one unit.
  - For binary x_j: the expected count of Y increases by a multiplicative factor of e^{β_j} for the group with $x_j = 1$ in comparison to the group with $x_j = 0$.

Poisson regression

- For example, suppose
  - Suppose the response variable is the number of mating for elephants, and let x_1 represent the age of the elephants
  - Also suppose $\hat{\beta}_j = 0.069$, so that $e^{\hat{\beta}_j} = e^{0.069} = 1.0714$.
  - Then, an increase in age of one year increases the expected number of mating for elephants by 7 percent.

Poisson regression

- The raw residuals $e_i = y_i - \hat{\lambda}_i$ are difficult to interpret since variance is equal to the mean in Poisson distributions.
- Use the Pearson's residuals instead:

$$r_i = \frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$$

- Plot the r_i's versus the predicted $\hat{\lambda}_i$'s, as well as the x_j values for each predictor j, to look for trends suggesting model misspecification.
- We can also use those to identify potential outliers.
- We can still check for multicollinearity, do model validation using RMSE, and do model selection via forward, backward and stepwise selection for Poisson regression
- We can also perform a change in deviance test to compare nested models.
- We will look at an example soon.

Poisson regression in terms of rate

- Recall that for aggregated data, the logistic regression model is

$$y_i|x_i \sim \text{Bin}(n_i, \pi_i); \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip},$$

  where n_i represents the population size for each "count" y_i.

- Here, we are really interested in learning about, explaining or estimating the probability/proportion/rate, $\pi_i \in (0, 1)$.

- The maximum likelihood estimate of each $\hat{\pi}_i = \frac{y_i}{n_i}$.

- When dealing with very rare events, the rates π_i will be very small, and sometimes really close to 0. Using the logistic regression model in these applications may not be ideal.

- Generally, estimation under the logistic regression model often fails for rates or probabilities close to 0 or 1.

- When dealing with these rare events, it turns out that we often will be able to take advantage of the Poisson approximation to the binomial.

Poisson regression in terms of rate

- To take advantage of this relationship, one way to rewrite the Poisson regression model is

$$y_i|\boldsymbol{x}_i \sim \text{Poisson}\left(\lambda_i = n_i\pi_i\right); \quad i = 1, \ldots, n.$$

- However, since we are really interested in the original rate π_i, we want to model that instead of the "expected counts" λ_i.

- That is, we can write

$$\log\left(\pi_i = \frac{\lambda_i}{n_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}.$$

- Since each n_i is then known, we can write

$$\Rightarrow \log\left(\lambda_i\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \log\left(n_i\right).$$

- Thus, rate data can be modeled by including the $\log(n)$ term with coefficient of 1 (called an offset). This offset is modeled with `offset()` option in R.

```
Model <- glm(successes ~ predictor, data=Data_agg, offset=log(n), family=poisson)
```

What's next?

Move on to the readings for the next module!