CS&SS/STAT/SOC 536: Regression for Count Outcomes

Adrian Dobra
adobra@uw.edu

1 Introduction

We assume that the outcome y counts how many times a certain even has occurred. In this situation y can take any positive integer value, i.e.

$$y \in \{0, 1, 2, \ldots\}.$$

It is incorrect to treat y as a continuous outcome and make use of simple linear regression to model it. The Poisson distribution with mean μ is typically use to model the outcome, i.e. $Y \sim Pois(\mu)$. That is, the probability that the outcome takes a value $y \in \{0, 1, 2, \ldots\}$ is

$$P(y|\mu) = \frac{\mu^y}{y!} \exp(-\mu),$$

where $y! = 1 \cdot 2 \cdot \ldots \cdot y$ (this is called "y factorial"). We have $0! = 1$, $1! = 1$, $2! = 2$, $3! = 6$ and, in general, $y! = y \cdot (y-1)!$. In order to understand the assumption we make about our outcome when we use the Poisson distribution, we need to calculate the mean and the variance of y. First we calculate the mean (the first order moment):

$$\begin{aligned} E[Y|\mu] &= \sum_{y \geq 0} yP(y|\mu), \\ &= \sum_{y \geq 1} \frac{\mu^y}{(y-1)!} \exp(-\mu), \\ &= \mu \exp(-\mu) \sum_{y \geq 0} \frac{\mu^y}{y!}, \\ &= \mu. \end{aligned}$$

The second order moment is

$$\begin{aligned} E[Y^2|\mu] &= \sum_{y \geq 0} y^2 P(y|\mu), \\ &= \mu^2 + \mu \cdot \sum_{y \geq 0} \frac{\mu^y}{y!} \exp(-\mu), \\ &= \mu^2 + \mu. \end{aligned}$$

The variance of the outcome is:

$$\begin{aligned} Var[Y|\mu] &= E[Y^2|\mu] - (E[Y|\mu])^2, \\ &= \mu^2 + \mu - \mu, \\ &= \mu. \end{aligned}$$

Therefore, once you assumed that your outcome follows a Poisson distribution, you implicitly assumed that the mean of your outcome is equal with its variance. This is called the assumption of **equidispersion**. *You need to make sure this assumption is actually true for your data. If the variance of the outcome is greater than the mean of the outcome (i.e.,* **overdispersion***), you must use other regression models (e.g., negative binomial regression).*

We assume that we have p explanatory (independent) variables $x = (x_1, x_2, \ldots, x_p)$. The Poisson regression model says that the outcome follows a Poisson distribution whose log-mean is a linear combination of the explanatory variables:

$$Y \sim Pois(\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)) \tag{1}$$

We assume we have observed n samples involving the outcome and the p explanatory variables:

$$\{(y^i, x^i) : i = 1, \ldots, n\},$$

where $x^i = (x_1^i, x_2^i, \ldots, x_p^i)$. For each sample, the Poisson regression model says that the conditional mean of that sample given the corresponding values of the explanatory variables is

$$E[y^i|x^i] = \exp(\beta_0 + \beta_1 x_1^i + \ldots + \beta_p x_p^i).$$

Again, the conditional mean is equal with the conditional variance under the equidispersion assumption:

$$E[y^i|x^i] = Var[y^i|x^i].$$

2 Maximum Likelihood Estimation

We denote $\mu_i = E[y^i|x^i]$. The log-likelihood is:

$$\begin{aligned} l(\beta_0, \beta_1, \ldots, \beta_p|data) &\propto \sum_{i=1}^{n} y^i \log(\mu_i) - \mu_i, \\ &= \sum_{i=1}^{n} y^i(\beta_0 + \beta_1 x_1^i + \ldots + \beta_p x_p^i) - \exp(\beta_0 + \beta_1 x_1^i + \ldots + \beta_p x_p^i). \end{aligned}$$

This log-likelihood is concave, hence it has a unique maximum. The Newton-Raphson algorithm will converge to this maximum and return the MLEs of the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$.

3 Interpreting the regression parameters

We assume we keep all the variables in a Poisson regression model fixed with the exception of variable x_k. As we change the values of x_k to $x_k + \delta$, the ratio of expected values of the outcome y is:

$$\begin{aligned}\frac{E[y|\ldots, x_{k-1}, x_k+\delta, x_{k+1}, \ldots]}{E[y|\ldots, x_{k-1}, x_k, x_{k+1}, \ldots]} &= \frac{\exp(\ldots + \beta_{k-1}x_{k-1} + \beta_k(x_k+\delta) + \beta_{k+1}x_{k+1} + \ldots)}{\exp(\ldots + \beta_{k-1}x_{k-1} + \beta_k x_k + \beta_{k+1}x_{k+1} + \ldots)}, \\ &= \exp(\beta_k\delta).\end{aligned}$$

The interpretation of the coefficient β_k is as follows:

For a chance in δ in x_k, the expected count changes by a factor of $\exp(\beta_k\delta)$, holding all other variables constant.

4 Variable Selection

Let $\mathcal{M}$ be a regression model that involves a variable x_{k_0}, $1 \leq k_0 \leq p$. We denote by $\mathcal{M}_{-k_0}$ the regression obtained from $\mathcal{M}$ by deleting variable x_{k_0}. Making a choice between $\mathcal{M}_{-k_0}$ and $\mathcal{M}$ involves testing the null hypothesis:

$$H_0: \quad \beta_{k_0} = 0.$$

The likelihood ratio test statistics is the difference in the deviances of the two models and its asymptotic distribution is Chi-square with 1 degree of freedom. Therefore the p-value associated with the likelihood ratio test is:

$$P\left(\chi_1^2 \geq G^2(\mathcal{M}_{-k_0}|\mathcal{M})\right).$$

The AIC and BIC are obtained with the same formulas given in the handout "Regression for binary outcomes". The dimension of a Poisson regression model that involves p explanatory variables is $p+1$.

5 Residuals and Prediction

Let $(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ be the MLEs corresponding with the Poisson regression model (1). The predicted count at $(x_1, \ldots, x_p)$ is

$$E[y|x_1, \ldots, x_p] = \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots + \widehat{\beta}_p x_p).$$

Again, the variance of the predicted count is equal with the predicted count itself, i.e.

$$E[y|x_1, \ldots, x_p] = Var[y|x_1, \ldots, x_p].$$

The *fitted* value corresponding with the i-th sample is the predicted value at $(x_1^i, \ldots, x_p^i)$. The discrepancy between the observed and the fitted values associated with the i-th sample is quantified using the *Pearson residual*:

$$r_i = \frac{y_i - E[y_i|x_1^i, \ldots, x_p^i]}{\sqrt{Var[y_i|x_1^i, \ldots, x_p^i]}} = \frac{y_i - E[y_i|x_1^i, \ldots, x_p^i]}{\sqrt{E[y_i|x_1^i, \ldots, x_p^i]}}.$$

If the assumption of equidispersion does not hold, i.e.

$$E[y|x_1, \ldots, x_p] \neq Var[y|x_1, \ldots, x_p],$$

your residuals will be large and your model will not fit your data. As before, you can formally test whether the fit of your model is dominated by outliers by calculating the sum of squares of the Pearson residuals:

$$\sum_{i=1}^{n} r_i^2.$$

It can be shown that asymptotically this quantity follows a Chi-square distribution with $n - |\mathcal{M}| = n - p - 1$ degrees of freedom. You decide that your model does not fit the data well if the p-value

$$P\left(\chi^2_{n-p-1} \geq \sum_{i=1}^{n} r_i^2\right).$$

is smaller than 0.05 (or 0.01). When overdispersion occurs, you can either model the counts with respect to a baseline measure (see HW3) or you use negative binomial regression. The negative binomial regression has an extra parameter that allows the variance to become bigger than the mean. Everything else remains the same.