Lecture 14: Logistic and Poisson Regression

Nancy R. Zhang

Statistics 203, Stanford University

February 24, 2010

Count data

1. Men and women were asked whether they believed in the after life (1991 General Social Survey).
2. Results:

	Y	N or U	
M	435	147	582
F	375	134	509
Total	810	281	1091

3. Question: is belief in afterlife independent of gender?

Contingency Tables

	Y	N or U	
M	435	147	582
F	375	134	509
Total	810	281	1091

1. Model: $Y_{ij} \sim Poisson(\lambda_{ij})$.
2. H_0: Independence. i.e. $\lambda_{ij} = \lambda\alpha_i\beta_j$.
3. H_A: λ_{ij} arbitrary.
4. Pearson's χ^2 Test:

$$\chi^2 = \sum_{i,j} \frac{(Y_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_1 \quad \text{(under } H_0\text{)}$$

5. Why 1 df? Independence model has 5 (λ, 2 α's, 2 β's) parameters, 2 constraints $\Rightarrow$ 3 df. Unrestricted model has 4 parameters.

Under independence:

$$\log E(Y_{ij}) = \log \lambda_{ij} = \log \lambda + \log \alpha_i + \log \beta_j.$$

What about variance? Because the data is Poisson,

$$Var(Y_{ij}) = E(Y_{ij}) = \lambda_{ij}.$$

Thus, the variance scales with the mean.

- Log stabilizes variance.
- But unlike before, we are explicit modeling data as Poisson rather than Gaussian – added power if the data is indeed Poisson.

Why Poisson?

1. Count data is always > 0.
2. Poisson distribution:

$$Poisson(k) = \sum_{i=1}^{k} Poisson(1)$$

By central limit theorem,

$$\frac{Poisson(k) - k}{\sqrt{k}} \rightarrow N(0, 1)$$

Thus "large poissons are like Gaussians". But small Poissons are quite different.

Similarities and differences with Gaussian, Logistic

1. Mean = $g(X\beta)$.
    1. Gaussian: g is identity.
    2. Binomial: g is logit.
    3. Poisson: g is log.

    g is called the "link" function.
2. Distribution of $Y \Rightarrow$ dependence of variance on mean.
    1. Gaussian: Variance constant in mean.
    2. Binomial: $Var(\pi) = \pi(1-\pi)$.
    3. Poisson: $Var(\lambda) = \lambda$.

There are many other models of this type, collectively called "generalized linear models."

Contingency table - regression model

Suppose that we have a k by m table. After life example: $k = m = 2$. We call this a $k \times m$ contingency table.

1. Model:

$$Y_{ij} \sim Poisson(\lambda_{ij})$$

2. Mean function:

$$\log \lambda_{ij} = \log \lambda + \log \alpha_i + \log \beta_j$$

3. Pearson test for independence:

$$\chi^2 = \sum_{ij} \frac{(Y_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{k-1,m-1} \quad \text{(under } H_0\text{)}$$

Poisson Regression

1. Model fitting: Newton Raphson.
2. Confidence intervals: same as for Binomial, use local Gaussianity.
3. Assessment of model fit: Deviance residuals.

Log-linear versus Logit models

1. Loglinear models are of use primarily when at least two variables are response variables. With a single categorical response, it is simpler and more natural to use logit models.
2. When you have two variables (e.g. Gender versus after-life belief), then logit might treat one as explanatory and the other as response, while there is an equivalent loglinear model.
3. Loglinear models view data as N independent cell counts rather than individual classifications of n subjects, $n = \sum_{i=1}^{N} Y_i$, and do not treat the row sums as fixed.

2×2 tables

	Y	N or U	
M	435	147	582
F	375	134	509
Total	810	281	1091

Model: $Y_{ij} \sim Poisson(\lambda_{ij})$.

If you have two Poissons, $Poiss(\lambda_{i1})$ and $Poiss(\lambda_{i2})$, then conditioned on their sum, each count is a binomial.

$$Y_{i,1} | Y_{i,1} + Y_{i,2} \sim Binomial\left(Y_{i,1} + Y_{i,2}, \frac{\lambda_{i1}}{\lambda_{i1} + \lambda_{i2}}\right)$$

Then,

$$\begin{aligned} \text{logit} P(1|\text{ row } = i, \text{row sum} = n_i) &= \log \frac{P(1|\text{ row } = i, \text{row sum} = n_i)}{P(2|\text{ row } = i, \text{row sum} = n_i)} \\ &= \log \frac{\lambda_{i1}}{\lambda i2} \\ &= \log \lambda_{i1} - \log \lambda_{i2}. \end{aligned}$$

2×2 tables

$$\text{logit}P(1|\text{ row } = i, \text{row sum} = n_i) \quad = \quad \log \lambda_{i1} - \log \lambda_{i2}.$$

Under the null hypothesis:

$$H_0: \quad \lambda_{ij} = \lambda * \alpha_i * \beta_j,$$

$$\log \lambda_{ij} = \log \lambda + \log \alpha_i + \log \beta_j.$$

$$\text{logit}P(1|\text{ row } = i, \text{row sum} = n_i) = \log \beta_1 - \log \beta_2 \equiv \delta$$

The key is that the above logit does not depend on i. In binomial regression, we are modeling

$$\text{logit}P(1|X) = \beta_0 + \beta_1 X.$$

So testing H_0 is equivalent to testing $\beta_1 = 0$ in logistic regression.

2×2 tables

Thus...

1. Testing the hypothesis $H_0 : \lambda_{ij} = \lambda * \alpha_i * \beta_j$ in the Poisson model is the same as testing independence in the logistic model.
2. To test this hypothesis, you fit the model with λ_{ij} arbitrary, and then use Chi-square test on the difference of deviances.
3. The difference of deviances will be the same as the logit model, but the absolute deviances will be different.