Statistics 191: Introduction to Applied Statistics

Poisson regression

Jonathan Taylor
Department of Statistics
Stanford University

March 8, 2010

Poisson regression

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University

Topics

- Contingency tables.
- Poisson regression.
- Generalized linear model.

Count data

Afterlife

- Men and women were asked whether they believed in the after life (1991 General Social Survey).
-

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

- Question: is belief in the afterlife independent of gender?

Poisson counts

Definition

- A random variable Y is a Poisson random variable with parameter λ if

$$P(Y=j) = e^{-\lambda}\frac{\lambda^j}{j!}, \qquad \forall j \geq 0.$$

- Some simple calculations show that

$$E(Y) = \text{Var}(Y) = \lambda.$$

- Poisson models for counts are analogous to Gaussian for continuous outcomes.

Count data

Jonathan Taylor
Department of Statistics
Stanford University

Contingency table

- Model: $Y_{ij} \sim Poisson(\lambda_{ij})$.
- **Null:** H_0 : independence, $\lambda_{ij} = \lambda \alpha_i \cdot \beta_j, \sum_i \alpha_i = 1, \sum_j \beta_j = 1$.
- **Alternative:** $H_a : \lambda_{ij}$'s are unrestricted
- **Test statistic:** Pearson's X^2 :

$$X^2 = \sum_{ij} \frac{(Y_{ij} - E_{ij})^2}{E_{ij}} \approx \chi^2_1 \text{ under } H_0$$

- Why 1 df ? Independence model has 5 parameters, two constraints = 3 df. Unrestricted has 4 parameters.
- This is actually a *regression model* for the count data.

Count data

Contingency table as regression model

- Under independence

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

- OR, the model has a *log link*.
- What about the variance? Because of Poisson assumption

$$Var(Y_{ij}) = E(Y_{ij})$$

- OR, the *variance function* is

$$V(\mu) = \mu.$$

Count data

Jonathan Taylor
Department of Statistics
Stanford University

Contingency table ($k \times m$)

- Suppose we had k categories on one axis, m on the other (i.e. previous example $k = m = 2$). We call this as $k \times m$ contingency table.
- Independence model:

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

Count data

Contingency tables

- Test for independence: Pearson's

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

- Alternative test statistic

$$G = 2 \sum_{ij} Y_{ij} \log \left(\frac{Y_{ij}}{E_{ij}} \right)$$

Count data

Independence tests

- Unlike in other cases, in this case the *full model* has as many parameters as observations (i.e. it's saturated).
- This test is known as a *goodness of fit* test.
- *How well does the independence model fit this data*?

Count data

Lumber company example

- Y : number of customers visting store from region;
- X_1 : number of housing units in region;
- X_2 : average household income;
- X_3 : average housing unit age in region;
- X_4 : distance to nearest competitor;
- X_5 : distance to store in miles.

Poisson (log-linear) regression model

- Given observations and covariates $Y_i, X_{ij}, 1 \leq i \leq n, 1 \leq j \leq p$.
- **Model:**

$$Y_i \sim Poisson\left(\exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}\right)\right)$$

- Poisson assumption implies the variance function is

$$V(\mu) = \mu.$$

Poisson regression

Interpretation of coefficients

- The log-linear model means covariates have *multiplicative* effect.
- Logistic model:

$$\frac{E(Y|\dots,X_j = x_j + 1,\dots)}{E(Y|\dots,X_j = x_j,\dots)} = e^{\beta_j}$$

- So, one unit increase in variable j results in e^{β_j} (multiplicative) increase the expected count, all other parameters being equal.

Count data

Generalized linear models

- Logistic model:

$$\text{logit}(\pi) = \beta_0 + \sum_j \beta_j X_j \qquad V(\pi) = \pi(1-\pi)$$

- Poisson log-linear model:

$$\log(\mu) = \beta_0 + \sum_j \beta_j X_j, \qquad V(\mu) = \mu$$

- These are the ingredients to a GLM ...

Generalized linear models

Specifying a model

- Given $(Y, X_1, \dots, X_p)$, a GLM is specified by the (link, variance function) pair (V, g).
- Fit using IRLS like logistic.
- Inference in terms of deviance or Pearson's X^2:

$$X^2(\mathcal{M}) = \sum_{i=1}^n \frac{(Y_i - \widehat{\mu}_{\mathcal{M},i})^2}{V(\widehat{\mu}_{\mathcal{M},i})}$$

Statistics 191: Introduction to Applied Statistics
Jonathan Taylor
Department of Statistics
Stanford University

Deviance

- Replaces *SSE* in least squares
- Definition

$$DEV(\mathcal{M}) = -2\left(\log L(\widehat{\mu}(\mathcal{M})|Y,X) - \log(Y|Y,X)\right)$$

- Difference between fitted values of $\mathcal{M}$ and "saturated model" with $\widehat{\mu} = Y$.
- Poisson deviance

$$DEV(\mathcal{M}|Y) = 2\sum_{i=1}^{n}\left(Y_i \log\left(Y_i/\widehat{\mu}_{\mathcal{M},i}\right) + (Y_i - \widehat{\mu}_{\mathcal{M},i})\right)$$

Generalized linear models

Deviance tests

- To test $H_0 : \mathcal{M} = \mathcal{M}_R$ vs. $H_a : \mathcal{M} = \mathcal{M}_F$, we use

$$DEV(\mathcal{M}_R) - DEV(\mathcal{M}_F) \sim \chi^2_{df_R - df_F}$$

- In contingency example $\mathcal{M}_R$ is the independence model

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

with $\mathcal{M}_F$ being the "saturated model."