Lecture 20: Count Models

Sravani Vadlamani

QMB 3200: Advanced Quantitative Methods

11/19/2020

For the rest of the semester

Date	Topic
19-Nov	Count Models
24-Nov	Count Models – Stata/ Project Feedback
01-Dec	Project
03-Dec	Final Review
08-Dec	Final Exam

Count Data Models

- Used for modelling the count of things as a function of covariates. Counts are non-negative integers.

Examples

- Count of vehicles in a queue
- Number of defective entities
- Number of failures
- Number of computers/cars/telephones/etc. in a household

Why ‘special’ methodology?

- OLS regression can/will predict values that are negative and will also predict non-integer values
- There are a number of ways to model counts, but Poisson and negative binomial are ‘popular’.
- Also, zero-inflated model can work under certain circumstances

Poisson Models

$$\Pr\left(Y = y_i\right) = \frac{EXP^{-\lambda_i}\lambda_i^{\,y_i}}{y_i\,!};\ y = 0,1,2,\ldots$$

where; $$E\left[y_i\right] = \lambda_i = EXP\left(\beta X_i\right)$$

$$\ln\left(\lambda_i\right) = \beta' X_i$$

Expressions b/w Response and Predictors

- The expression shown is the log-linear form—there are others as well, but log-linear is most common.
- It is the *exp* portion of the expression that constrains the model forecasts to be positive.

Poisson Models

By substituting E[y] = EXP (ßX) in the expression, one easily obtains the likelihood function for all observations:

$$L(\beta) = \prod_i \frac{EXP\left[-EXP\left(\beta X_i\right)\right]\left[EXP\left(\beta X_i\right)\right]^{y_i}}{y_i!}$$

And the log-likelihood is simply:

$$LL(\beta) = \sum_{i=1}^{n}\left[-EXP\left(\beta X_i\right) + y_i\beta X_i - LN\left(y_i!\right)\right]$$

Poisson Model Elasticities

An elasticity is the estimate of the effect of a change in an independent variable on the dependent variable. The elasticity on the count of individual *i* for the k^{th} continuous independent variable is given as:

$$E_{x_{ik}}^{\lambda_i} = \frac{\partial \lambda_i}{\lambda_i} \times \frac{x_{ik}}{\partial x_{ik}} = \beta_k x_{ik}$$

So, if an elasticity was -5.4, then for a 1% increase in the variable would result in a 5.4% decrease in the expected frequency.

Why are elasticities useful? i.e.,

why not just use coefficient estimates?

Pseudo Elasticity

For a discrete variable the previous equation is not suitable.

$$E_{x_{ik}}^{\lambda_i} = \frac{EXP(\beta_k) - 1}{EXP(\beta_k)}$$

So, the pseudo elasticity for an indicator variable is computed as:

Review

1. What data are appropriate for Poisson models?

2. Why can’t regression coefficients be used to reflect ‘effect’ of covariates?

Poisson Models, GOF Measures

Log-likelihood ratio test to compare restricted and unrestricted models

$$-2\left[LL\left(\beta_R\right)-LL\left(\beta_U\right)\right]\approx\chi^2\left(\alpha,df_U-df_R\right)$$

The sum of model deviances, G-square, is equal to zero for a model with perfect fit.

$$G^2=2\sum_{i=1}^{n}y_i LN\left(\frac{y_i}{\hat{\lambda}_i}\right)$$

Poisson Models, GOF Measures

A measure similar to R-square is given as

$$R_p^2 = 1 - \frac{\sum_{i=1}^{n} \left[\frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}} \right]^2}{\sum_{i=1}^{n} \left[\frac{y_i - \bar{y}}{\sqrt{\bar{y}}} \right]^2}$$

Another measure of overall model fit is the ρ-square statistic.

$$\rho^2 = 1 - \frac{LL(\beta)}{LL(0)}$$

Poisson Models, GOF Measures

Because of the non-linearity of the conditional mean and heteroscedasticity in the regression, there is no ‘true’ equivalent of R-square.

Example 1: Intersection Accidents at two-lane rural roads in California and Michigan.

Variable Abbreviation	Variable Description	Maximum / Minimum Values	Mean of Observations	Standard Deviation of Observations
State	Indicator variable for state: 0 = California; 1 = Michigan.	1 / 0	0.29	0.45
Accident	Count of injury accidents over observation period	13 / 0	2.62	3.36
*AADT*1	Average annual daily traffic on major road	33058 / 2367	12870	6798
*AADT*2	Average annual daily traffic on minor road	3001 / 15	596	679
Median	Median width on major road in feet	36 / 0	3.74	6.06
Drive	Number of driveways within 250 feet of intersection center	15 / 0	3.10	3.90

Resultant Model

So, the estimated model has the form:

$$E\left[y_i\right] = \lambda_i = EXP\left(\beta X_i\right)$$
$$= EXP\begin{pmatrix} -0.83 + 0.00008(AADT1_i) + 0.0005(AADT2_i) \\ -0.06(Median_i) + 0.07(Drive_i) \end{pmatrix}$$
$$= EXP^{-0.83} EXP^{0.00008(AADT1)} ....EXP^{0.07(Drive)}$$
$$= (0.436)(AADT1Factor).......(DriveFactor)$$

The model is additive in the exponent or multiplicative on the expected value of y.

Formatted Model Output

Independent Variable	Estimated Parameter	*t*-statistic
Constant	-0.826	-3.57
Average Annual Daily Traffic on Major Road	0.0000812	6.90
Average Annual Daily Traffic on Minor Road	0.000550	7.38
Median width in feet	- 0.0600	- 2.73
Number of driveways with 250 feet of intersection	0.0748	4.54
Number of observations	84	
Restricted Log likelihood (constant term only)	-246.18	
Log likelihood at convergence	-169.26	
Chi-squared and associated *p*-value	153.85	<0.0000001
R_p-Squared	0.4792	
G^2	176.5	

Model Elasticities

Independent Variable	Elasticity
Average Annual Daily Traffic on Major Road	1.045
Average Annual Daily Traffic on Minor Road	0.327
Median width in feet	-0.228
Number of driveways with 250 feet of intersection	0.232

Major road traffic has approximately 3 x the effect on crashes than does minor road traffic. Increasing the median width 1% decreases expected crash count by .22. Driveways nearby increase crashes.

Poisson Model Restriction

- The Poisson distribution has one parameter, λ, which represent the distribution mean and variance.

- Often in real data the variance is not equal to the mean (e.g. statistically), and the Poisson model is not appropriate for the count process.

- We say that there is ‘extra’ heterogeneity across the Poisson means.

Over-dispersion

- Over-dispersion (VAR[Y] > E[Y]) occurs in the following conditions:
- A Poisson process over an interval whose length is random rather than fixed
- Inter-subject variability (each individual has Poisson process with mean Z as a random variable. In this case we assume $E(Z) = \lambda$ and $VAR(Z) = \lambda / \phi$, where ϕ is larger or smaller than one.

Over-dispersion (cntd.)

- When over-dispersion occurs we should change the model to accommodate it.
- We can:
  - assume the overdispersion is gamma distributed across means—resulting in a negative binomial model (or Poisson-gamma model)
  - assume the overdispersion is normally distributed (Poisson-normal model)

Poisson & Negative Binomial Models

$$VAR(Y_i) = \sigma^2 E(Y_i); \text{ where}$$

$$\sigma^2 = \text{ dispersion parameter}$$

$$\sigma^2 > 1 \text{ (overdispersion)}$$

$$\sigma^2 < 1 \text{ (underdispersion)}$$

Negative Binomial Models

$$\lambda_i = EXP\left(\beta' x_i + \varepsilon_i\right)$$

where;

EXP^{ε_i} is gamma distributed with mean = 1

and variance α

Negative Binomial Model

The model has an additional parameter alpha, such that:

$$VAR(y_i) = E[y_i]\{1 + \alpha E(Y_i)\}$$

When α= 0, the model “collapses” to the Poisson model. The over-dispersion rate is given by:

$$\frac{VAR(y_i)}{E[y_i]} = \{1 + \alpha E(Y_i)\}$$

Test for Overdispersion

A test by Cameron and Trevedi (1990). It is based on the assumption that under the Poisson model $(y_i - E[y_i])^2 - E[y_i]$ has mean zero.

$$H_0: \quad VAR[y_i] = E[y_i]$$

$$H_A: \quad VAR[y_i] = E[y_i] + \alpha g(E[y_i])$$

Test for Overdispersion

To conduct this test, a simple linear regression is estimated where Z_i is regressed on W_i, where,

$$Z_i = \frac{\left(y_i - E\left(y_i\right)\right)^2 - y_i}{E\left(y_i\right)\sqrt{2}}$$

$$W_i = \frac{g\left(E\left(y_i\right)\right)}{\sqrt{2}}$$

After running the regression $Z_i = b_i W$ with $g(E[y_i]) =$ and $g(E[y_i]) =$, if b is statistically significant in both cases, then H_0 is rejected for the particular function g.

Negative Binomial Model

Independent Variable	Estimated Parameter	*t*-statistic
Constant	-0.931	-2.37
Average Annual Daily Traffic on Major Road	0.0000900	3.47
Average Annual Daily Traffic on Minor Road	0.000610	3.09
Median width in feet	- 0.0670	- 1.99
Number of driveways with 250 feet of intersection	0.0632	2.24
Overdispersion parameter, α	0.516	3.09
Number of observations	84	
Restricted Log likelihood (constant term only)	-169.26	
Log likelihood at convergence	-153.28	
Chi-squared and associated *p*-value	31.95	<0.0000001

Poisson Model

Independent Variable	Estimated Parameter	*t*-statistic
Constant	-0.826	-3.57
Average Annual Daily Traffic on Major Road	0.0000812	6.90
Average Annual Daily Traffic on Minor Road	0.000550	7.38
Median width in feet	- 0.0600	- 2.73
Number of driveways with 250 feet of intersection	0.0748	4.54
Number of observations	84	
Restricted Log likelihood (constant term only)	-246.18	
Log likelihood at convergence	-169.26	
Chi-squared and associated *p*-value	153.85	<0.0000001
R_p-Squared	0.4792	
G^2	176.5	

Review

- Describe ‘over-dispersion’
- How does the negative binomial model arise?
- Are there other models for over-dispersion that could be used?