Count Data Models

Count Data Models

Count data examples

- Consumer demand: the number of products that a consumer buys on Amazon
- Recreational data: the number of trips taken per year
- Family economics: the number of children a couple has
- Health demand: the number of doctors visits

Count data dependent variable

- The dependent variable is counts (a non-negative integer): y = 0, 1, 2, 3, 4, …
- The sample is concentrated on a few small discrete values.
- We study the factors affecting the average number of the dependent variable.

Poisson model

Poisson model

- The Poisson model predicts the number of occurrences of an event.
- The Poisson model states that the probability that the dependent variable *Y* will be equal to a certain number *y* is:

$$p(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}$$

- For the Poisson model, μ is the intensity or rate parameter.

$$\mu = \exp\left(\mathbf{x}_i'\beta\right)$$

- Interpretation of the coefficients: one unit increase in x will increase/decrease the average number of the dependent variable by the coefficient expressed as a percentage.

Properties of the Poisson distribution

- *Equidispersion property* of the Poisson distribution: the equality of mean and variance.

$$E(y|x) = var(y|x) = \mu$$

  This is a restrictive property and often fails to hold in practice, i.e., there is "overdispersion" in the data. In this case, use the negative binomial model.

- *Excess zeros problem* of the Poisson distribution: there are usually more zeros in the data than a Poisson model predicts. In this case, use the zero-inflated Poisson model.

Marginal effects for the Poisson model

- The marginal effect of a variable on the average number of events is:

$$\partial E(y|x)/\partial x_j = \beta_j \exp(\mathbf{x}_i'\beta)$$

- Interpretation of the marginal effects: one unit increase in x will increase/decrease the average number of the dependent variable by the marginal effect.

Negative binomial model

- The negative binomial model is used with count data instead of the Poisson model if there is overdispersion in the data.
- Unlike the Poisson model, the negative binomial model has a less restrictive property that the variance is not equal to the mean (μ).

$$var(y|x) = \mu + \alpha\mu^2$$

- Another functional form is $var(y|x) = \mu + \alpha\mu$, but this form is less used.
- The negative binomial model also estimates the overdispersion parameter α.

Test for overdispersion

- Estimate the negative binomial model which includes the overdispersion parameter α and test if α is significantly different than zero.
- H_0: $\alpha = 0$ or H_a: $\alpha \neq 0$
- We have three cases:
  - When $\alpha = 0$, the Poisson model.
  - When $\alpha > 0$, overdispersion (frequently holds with real data).
  - When $\alpha < 0$, underdispersion (not very common).

Incidence rate ratios (irr)

- For the Poisson and negative binomial models, in addition to reporting the coefficients and marginal effects, we can also report the incidence rate ratios.
- The incidence rate ratios report exp(b) rather than b.
- Interpretation of the incidence rate ratios: irr=2 means that for each unit increase in x, the expected number of y will double.

Hurdle or two-part models

- The two-part model relaxes the assumption that the zeros (whether or not there are events) and positives (how many events) come from the same data generating processes.
- Example: different factors may affect whether or not you practice a particular sport and how many times you practice your sport in a month.
- We can estimate two-part models similar to the truncated regression models.
- If the process generating the zeros is $f_1(.)$ and the process generating the positive responses is $f_2(.)$ then the two-part hurdle model is defined by the following probabilities:

$$g(y) = \begin{cases} f_1(0) \ \ if \ y = 0 \\ \dfrac{1 - f_1(0)}{1 - f_2(0)} f_2(y) \ \ if \ y \geq 1 \end{cases}$$

- If the two processes are the same, this is the standard count data model.
- The model for the zero versus positive responses is a binary model with the specified distribution, but we usually estimate it with the probit/logit model.

Zero-inflated models

- The zero-inflated model is used with count data when there is an excess zeros problem.
- The zero-inflated model lets the zeros occur in two different ways: as a realization of the binary process (z=0) and as a realization of the count process when the binary variable z=1.
- Example: you either like hiking or you do not. If you like hiking, the number of hiking trips you can take is 0, 1, 2, 3, etc. So you may like hiking, but may not take a trip this year. We are able to generate more zeros in the data.
- If the process generating the zeros is $f_1(.)$ and the process generating the positive responses is $f_2(.)$ then the zero-inflated model is:

$$g(y) = \begin{cases} f_1(0) + (1 - f_1(0))f_2(0) \;\; if \; y = 0 \\ (1 - f_1(0))f_2(y) \;\; if \; y \geq 1 \end{cases}$$

- The zero-inflated model is less frequently used than the hurdle model.
- The zero-inflated models can handle the excess zeros problem.

Count Data Models Example

- Poisson and negative binomial models, truncated models, and zero-inflated models.
- We want to study the factors influencing the number of doctor visits.
- Data are from the U.S. Medical Expenditure Panel Survey for 2003.
- Dependent variable: number of doctor visits.
- Independent variables: has private insurance, has Medicaid insurance, age, education.

Number of doctor visits (y)	Percent frequency
0	11%
1	9%
2	9%
3	9%
4	9%
5	7%
6	6%
7	5%
8	5%
>=9	…

Poisson and negative binomial model coefficients and marginal effects

Number of doctor visits	Poisson coefficients	Poisson marginal effects	Negative binomial coefficients	Negative binomial marginal effects
Has private insurance	0.15*	1.04*	0.16*	1.08*
Has Medicaid insurance	0.29*	1.96*	0.28*	1.96*
Age	0.01*	0.07*	0.01*	0.07*
Education	0.02*	0.17*	0.02*	0.16*
Intercept	0.74*		0.75*	
Alpha (overdispersion parameter)			0.81*	

- Interpretation of the coefficients: individuals who have private insurance are expected to have a 15% (16%) increase in the number of doctor visits for the Poisson (negative binomial model).
- Interpretation of the marginal effects: individuals who have private insurance are expected to have 1.04 (1.08) additional doctor visits for the Poisson (negative binomial) model.
- The overdispersion parameter α is significantly different from zero, therefore we should use the negative binomial model instead of the Poisson model.

Truncated Poisson and negative binomial model coefficients and marginal effects

Number of doctor's visits	Logit coefficients	Zero truncated Poisson coefficients	Zero trunc Poisson marginal effects	Zero trunc negative binomial coefficients	Zero trunc neg binomial marginal effects
Has private insurance	0.64*	0.09*	0.67*	0.10*	0.71*
Has Medicaid insurance	0.46*	0.24*	1.83*	0.26*	1.83*
Age	0.04*	0.01*	0.05*	0.01*	0.05*
Education	0.04*	0.02*	0.16*	0.02*	0.15*
Intercept	-1.71*	1.25*		1.10*	
Alpha (dispersion parameter)				0.77*	

- Interpretation of the coefficients for the logit model: individuals who have private insurance, have Medicaid insurance, who are older, or have higher education are more likely to have a positive number of doctor visits.
- Interpretation of the coefficients and marginal effects for the truncated Poisson (truncated negative binomial) model: for individuals with positive number of doctor visits, those who have private insurance have a 9% (10%) higher number of doctor visits or 0.67 (0.71) additional number of doctor visits.
- Note similar results of the two-step to one-step models but that doesn't have to be the case.

Zero-inflated Poisson and negative binomial model coefficients and marginal effects

Number of doctor's visits	Inflated model coefficients	Zero-inflated Poisson coefficients	Zero-inflated Poisson marginal effects	Inflated model coefficients	Zero-inflated negative binomial coefficients	Zero-inflated negative binomial marginal effects
Has private insurance	-0.64*	0.09*	1.05*	-2.56*	0.10*	0.95*
Has Medicaid insurance	-0.46*	0.24*	1.95*	-0.88*	0.26*	1.85*
Age	-0.04*	0.01*	0.07*	-0.13*	0.01*	0.06*
Education	-0.04*	0.02*	0.17*	0.05*	0.02*	0.16*
Intercept	1.70*	1.25*		7.42*	1.05*	
Alpha (dispersion parameter)					0.75*	

- Interpretation of the coefficients and marginal effects for the zero-inflated Poisson (truncated negative binomial) model: for individuals who are inclined to go the doctor, those who have private insurance have a 9% (10%) higher number of doctor visits or 1.05 (0.95) additional number of doctor visits.

- Inflated model is similar to a logit model (binary model) but the coefficients are different.
- Note similar results of the truncated and zero-inflated models. In this case, there are a higher number of additional doctor visits if we use inflated models than if we use truncated models.
- Note that we do not have to use the same variables in the binary models (zero vs. positive outcome) and the truncated or zero-inflated second-step models (for the positive outcomes).