ZIP/ZINB - Count Data Models

See Book Chapter 11

- Count data consist of non-negative integer values
- Examples:
  - number of driver route changes per week,
  - the number of trip departure changes per week,
  - drivers' frequency-of-use of ITS technologies over some time period,
  - the number of accidents observed on road segments per year.
- Count data can be properly modeled by using a number of methods, the most popular of which are Poisson and negative binomial regression models.

Poisson Regression Model

- Consider the number of accidents occurring per year at various intersections in a city.
- In a Poisson regression model, the probability of intersection i having y_i accidents per year (where y_i is a non-negative integer) is given by:

$$P(y_i) = \frac{EXP(-\lambda_i)\lambda_i^{y_i}}{y_i!}$$

- Where:

$P(y_i)$ is the probability of intersection i having y_i accidents per year

λ_i is the Poisson parameter for intersection i, which is equal to intersection i's expected number of accidents per year, $E[y_i]$.

- Poisson regression models are estimated by specifying the Poisson parameter λ_i (the expected number of events per period) as a function of explanatory variables.

- The most common relationship between explanatory variables and the Poisson parameter is the log-linear model,

$$\lambda_i = EXP(\beta X_i) \text{ or, equivalently}$$
$$LN(\lambda_i) = \beta X_i,$$

- Where:

X_i is a vector of explanatory variables and

β is a vector of estimable coefficients.

- In this formulation, the expected number of events per period is given by

$$E[y_i] = \lambda_i = EXP(\beta X_i)$$

- For model estimation, note the likelihood function is:

$$L(\beta) = \prod_i P(y_i)$$

- So, with the Poisson equation,

$$L(\beta) = \prod_i \frac{EXP(-\lambda_i)\lambda_i^{y_i}}{y_i!}$$

- Since $\lambda_i = EXP(\beta X_i)$,

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

- Which gives the log-likelihood,

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

Poisson Regression Model Goodness of Fit Measures

- The likelihood ratio test is a common test used to assess two competing models. It provides evidence in support of one model

- The likelihood ratio test statistic is,

$$-2[LL(\beta_R) - LL\ (\beta_U)]$$

  - where

    $LL(\beta_R)$ is the log-likelihood at convergence of the "restricted" model (sometimes considered to have all coefficients in β equal to 0, or just to include the constant term, to test overall fit of the model)

    $LL(\beta_U)$ is the log-likelihood at convergence of the unrestricted model.

  - This statistic is χ^2 distributed with the degrees of freedom equal to the difference in the numbers of coefficients in the restricted an unrestricted model (the difference in the number of coefficients in the β_R and the β_U coefficient vectors).

- Another measure of overall model fit is the ρ^2 statistic. The ρ^2 statistic is,

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

- Where:

  $LL(\beta)$ is the log-likelihood at convergence with coefficient vector β and

  $LL(0)$ is the initial log-likelihood (with all coefficients set to zero).

- The perfect model would have a likelihood function equal to one (all selected alternative outcomes would be predicted by the model with probability one, and the product of these across the observations would also be one) and the log-likelihood would be zero giving a ρ^2 of one

- The ρ^2 statistic will be between zero and one and the closer it is to one, the more variance the estimated model is explaining.

Truncated Poisson Regression Model

- Truncation of data can occur in the routine collection of transportation data.
- Example, if the number of times per week an in-vehicle navigation system is used on the morning commute to work, during weekdays, the data are right truncated at 5, which is the maximum number of uses in any given week.
- Estimating a Poisson regression model without accounting for this truncation will result in biased estimates of the parameter vector β, and erroneous inferences will be drawn.
- Fortunately, the Poisson model is adapted easily to account for such truncation. The right-truncated Poisson model is written as:

$$P(y_i) = \left[\lambda_i^{y_i} / y_i!\right] \Big/ \left[\sum_{m_i=0}^{r} \left(\lambda_i^{m_i} / m_i!\right)\right],$$

- Where:

  $P(y_i)$ is the probability of commuter i using the system y_i times per week,

  λ_i is the Poisson parameter for commuter i;

  m_i is the number of uses per week;

  and r is the right truncation (in this case, 5 times per week).

Negative Binomial Regression Model

- Poisson distribution that restricts the mean and variance to be equal:

$$E[y_i] = VAR[y_i].$$

- If this equality does not hold, the data are said to be under dispersed ($E[y_i] > VAR[y_i]$) or overdispersed ($E[y_i] < VAR[y_i]$), and the coefficient vector will be biased if corrective measures are not taken.

- To account for cases when $E[y_i] \neq VAR[y_i]$, a negative binomial model is used.
- The negative binomial model is derived by rewriting the λ_i equation such that,

$$\lambda_i = EXP(\beta X_i + \varepsilon_i)$$

- where $EXP(\varepsilon_i)$ is a Gamma-distributed error term with mean 1 and variance α^2.
- The addition of this term allows the variance to differ from the mean as below,

$$VAR[y_i] = E[y_i][1+ \alpha E[y_i]] = E[y_i]+ \alpha E[y_i]^2$$

- The Poisson regression model is regarded as a limiting model of the negative binomial regression model as α approaches zero, which means that the selection between these two models is dependent upon the value of α.

- The parameter α is referred to as the overdispersion parameter.
- The negative binomial distribution has the form,

$$P(y_i) = \frac{\Gamma((1/\alpha) + y_i)}{\Gamma(1/\alpha) y_i!} \left(\frac{1/\alpha}{(1/\alpha) + \lambda_i} \right)^{1/\alpha} \left(\frac{\lambda_i}{(1/\alpha) + \lambda_i} \right)^{y_i}$$

where $\Gamma(.)$ is a gamma function. This results in the likelihood function,

$$L(\lambda_i) = \prod_i \frac{\Gamma((1/\alpha) + y_i)}{\Gamma(1/\alpha) y_i!} \left(\frac{1/\alpha}{(1/\alpha) + \lambda_i} \right)^{1/\alpha} \left(\frac{\lambda_i}{(1/\alpha) + \lambda_i} \right)^{y_i}$$

Zero-Inflated Poisson and Negative Binomial Regression Models

- Zero events can arise from two qualitatively different conditions.
    1. One condition may result from simply failing to observe an event during the observation period.
    2. Another qualitatively different condition may result from an inability to ever experience an event.
- Two states can be present, one being a normal count-process state and the other being a zero-count state.
- A zero-count state may refer to situations where the likelihood of an event occurring is extremely rare in comparison to the normal-count state where event occurrence is inevitable and follows some know count process

- Two aspects of this non qualitative distinction of the zero state are noteworthy:

    1. There is a preponderance of zeroes in the data—more than would be expected under a Poisson process.

    2. A sampling unit is not required to be in the zero or near zero state into perpetuity, and can move from the zero or near zero state to the normal count state with positive probability.

- Data obtained from two-state regimes (normal-count and zero-count states) often suffer from overdispersion if considered as part of a single, normal-count state because the number of zeroes is inflated by the zero-count state.

Zero-inflated Poisson (ZIP)

- Assumes that the events, $Y = (y_1, y_2, \ldots\ldots, y_n)$, are independent and the model is

$$y_i = 0 \text{ with probability } p_i + (1 - p_i)\, EXP(-\lambda_i)$$

$$y_i = y \text{ with probability } \frac{(1 - p_i)\, EXP(-\lambda_i)\, \lambda_i^y}{y!} \cdot$$

where y is the number of events per period.

Zero-inflated negative binomial (ZINB)

- regression model follows a similar formulation with events, $Y = (y_1, y_2, \ldots\ldots, y_n)$, being independent and,

$$y_i = 0 \text{ with probability } p_i + (1 - p_i)\left[\frac{\frac{1}{\alpha}}{\left(\frac{1}{\alpha}\right) + \lambda_i}\right]^{1/\alpha}$$

$$y_i = y \text{ with probability } (1 - p_i)\left[\frac{\Gamma\left(\left(\frac{1}{\alpha}\right) + y\right) u_i^{1/\alpha}(1 - u_i)^y}{\Gamma\left(\frac{1}{\alpha}\right) y!}\right], y=1, 2, 3...$$

- where $u_i = (1/\alpha)/\left[(1/\alpha) + \lambda_i\right]$.
  - Zero-inflated models imply that the underlying data-generating process has a splitting regime that provides for two types of zeros.
  - The splitting process can be assumed to follow a logit (logistic) or probit (normal) probability process, or other probability processes.

- A point to remember is that there must be underlying justification to believe the splitting process exists (resulting in two distinct states) prior to fitting this type of statistical model. There should be a basis for believing that part of the process is in a zero-count state.

- To test the appropriateness of using a zero-inflated model rather than a traditional model, Vuong (1989) proposed a test statistic for non-nested models that is well suited for situations where the distributions (Poisson or negative binomial) are specified. The statistic is calculated as (for each observation *i*),

$$m_i = LN\left(\frac{f_1\left(y_i \mid X_i\right)}{f_2\left(y_i \mid X_i\right)}\right)$$

- where:

$f_1(y_i|X_i)$ is the probability density function of model 1, and

$f_2(y_i|X_i)$ is the probability density function of model 2.

- Using this, Vuongs' statistic for testing the non-nested hypothesis of model 1 versus model 2 is (Greene, 2000; Shankar et al., 1997),

$$V = \frac{\sqrt{n}\left[\left(\frac{1}{n}\right)\sum_{i=1}^{n} m_i\right]}{\sqrt{\left(\frac{1}{n}\right)\sum_{i=1}^{n}\left(m_i - \overline{m}\right)^2}} = \frac{\sqrt{n}\left(\overline{m}\right)}{S_m}$$

- Where: $\overline{m}$ is the mean ($(1/n)\sum_{i=1}^{n} m_i$), S_m is standard deviation,

- Vuongs' value is asymptotically standard normal distributed (to be compared to *z*-values), and

- if $|V|$ is less than $V_{critical}$ (1.96 for a 95% confidence level), the test does not support the selection of one model over another.

- Large positive values of V greater than $V_{critical}$ favor model 1 over model 2, whereas large negative values support model 2.

		t-statistic of the NB overdispersion parameter α	
		< |1.96|	> |1.96|
Vuong statistic for ZINB(f_1(.)) and NB(f_2(.)) comparison	< - 1.96	ZIP or Poisson as alternative to NB	NB
	> 1.96	ZIP	ZINB

- Because overdispersion will almost always include excess zeros, it is not always easy to determine whether excess zeros arise from true overdispersion or from an underlying splitting regime.
- This could lead one to erroneously choose a negative binomial model when the correct model may be a zero-inflated Poisson.
- The use of a zero-inflated model may be simply capturing model mispecification that could result from factors such as unobserved effects (heterogeneity) in the data.

```
--> poisson;lhs=x5
    ;rhs=one,sr520,x14,x15,x17
    ;zip
    ;limit=6;truncation;upper$

Normal exit from iterations. Exit status=0.

 +---------------------------------------------------------------------+
 | Zero Altered Poisson      Regression Model                          |
 | Logistic distribution used for splitting model.                     |
 | ZAP term in probability is F[tau x ln LAMBDA]                       |
 | Comparison of estimated models                                      |
 |             Pr[0|means]       Number of zeros        Log-likelihood |
 | Poisson          .56606   Act.=   126 Prd.=   115.5      -226.14085 |
 | Z.I.Poisson      .46033   Act.=   126 Prd.=    93.9      -199.33319 |
 | Note, the ZIP log-likelihood is not directly comparable.            |
 | ZIP model with nonzero Q does not encompass the others.             |
 | Vuong statistic for testing ZIP vs. unaltered model is      8.2473  |
 | Distributed as standard normal. A value greater than                |
 | +1.96 favors the zero altered Z.I.Poisson model.                    |
 | A value less than -1.96 rejects the ZIP model.                      |
 +---------------------------------------------------------------------+
```

```
poisson;lhs=x5
    ;rhs=one,sr520,x14,x15,x17
    ;rh2=one,x17
    ;zip=normal
    ;rpm;pts=200;halton
    ;fcn=x14(n),sr520(n)
    ;limit=6;truncation;upper$
```

Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
Nonrandom parameters					
Constant	.71491800	.69105613	1.035	.3009	
X15	-.01954401	.02227875	-.877	.3804	7.15196078
X17	.11296038	.31458872	.359	.7195	1.61078431
Means for random parameters					
X14	-.26808326	.17372280	-1.543	.1228	.62254902
SR520	-.96019265	.31028701	-3.095	.0020	.13725490
Diagonal elements of Cholesky matrix					
X14	.12610375	.11457337	1.101	.2711	
SR520	.51165140	.24727997	2.069	.0385	
Below diagonal elements of Cholesky matrix					
lSR5_X14	-.55832649	.26574401	-2.101	.0356	
Variables in ZERO regime logit probability					
Constant	12.2193479	2.72660999	4.482	.0000	
X17	-7.82774650	1.84263489	-4.248	.0000	.000000

Implied covariance matrix of random parameters

Matrix Var_Beta has 2 rows and 2 columns.

	1	2
1	.01590	-.07041
2	-.07041	.57352

```
Implied standard deviations of random parameters
Matrix S.D_Beta has  2 rows and  1 columns.
                1
         +--------------
        1|      .12610
        2|      .75731

Matrix Cor_Beta has  2 rows and  2 columns.
                1              2
         +-----------------------------
        1|    1.00000       -.73725
        2|    -.73725       1.00000
```