CS-BIGS 5(1): 44-50
http://www.bentley.edu/centers/csbigs/tasneem.pdf

Modeling the Number of Research Papers Produced by Graduate Students Using Zero-Inflated Models

Tasneem Zaihra
McGill University, Canada

The objective of this case study is to discuss a step by step approach in modeling zero inflated over dispersed counts using data on the number of research papers produced by a group of biochemistry students. The model can be further used to study factors associated with differences in productivity of students within the PhD (Biochemistry) stream. We fit a Zero Inflated Negative Binomial (ZINB) regression model in order to predict the number of articles produced during the last three years of PhD from factors indicating the gender of the student, marital status, the number of children aged five or younger and the number of articles produced by a PhD mentor during the last three years. The dispersion parameter is found to be significantly different from zero, suggesting that the counts are over dispersed, and that a Negative Binomial (NB) model is more appropriate than a Poisson model. Vuong's test further suggests that our zero-inflated model is a significant improvement over a standard NB model. Thus, the ZINB model is a clear winner in terms of parsimony and goodness of fit for the data. Based on our model, we find significant disadvantages for females and scientists with children under five and a large positive effect of the number of publications by the mentor. The presentation is accessible to readers with an intermediate level of statistics.

In the analysis of count data, the dependent variable is usually the number of times an event occurs over a fixed period of time or other specified intervals such as distance, area or volume. Some examples of event counts are as follows.

- The number of physician and hospital outpatient visits is often used in modeling the demand for medical care. This number is the dependent variable and it is analyzed using several explanatory variables such as the number of hospital stays, self-perceived health status, the number of chronic conditions as well as socioeconomic variables such as gender, the number of years of schooling, private insurance indicator, etc.
- The number of claims per year on a particular car owner's auto insurance policy.
- The number of workdays missed due to the sickness of a dependent in a four week period.
- The number of papers published per year by a researcher. In this case, the covariates can be gender, the number of years spent as a PhD candidate, ranking of the department, number of publications by a mentor etc.
- The number of days of absence is used by school administrators to study the attendance behavior of high school juniors. Predictors of the number of days of absence include the gender of the student and standardized test scores in math and language arts.
- The number of fish caught by fishermen at a state wildlife park is often used in fisheries study. The regressors can be the length of stay at the wildlife park, the number of people in the group etc.
- The number of times HIV infected men develop a urinary tract infection over a specific time period.

The Poisson (log-linear) regression model for such event counts is the most basic model that explicitly takes into account the nonnegative integer-valued aspect of the dependent count variable. However, real-life data are often characterized by over dispersion (i.e., the variance exceeds the mean) as well as zero inflation (i.e., there are excess zeros). In such situations the Poisson model is not the best fit because of its restrictive property that the conditional variance equals the conditional mean. The negative binomial regression model, which is a generalization of the Poisson regression model, allows for over dispersion but does not take care of excess zeros in the data. In order to take care of excess zeros we resort to zero-inflated counterparts of either Poisson or Negative Binomial models depending upon whether the data are over dispersed or not.

Zero-inflated models

The main motivation for zero-inflated count models is that real-life data frequently display over dispersion and excess zeros (Lambert 1992; Greene 1994). The zero-inflated density is a mixture of a point mass at zero and a count distribution such as Poisson, Geometric or Negative Binomial (NB). There are two sources of zeros: zeros may come from both the point mass and from the count component. For modeling the unobserved state (zero vs. count), a binary model is used: in the simplest case only with an intercept but potentially containing regressors although the vector of regressors in the zero inflation component and the regressors in the count component need not be distinct. In general the response $Y = (Y_{1...} Y_n)$ are independent and

$Y_i \sim 0$ with probability p_i
$\sim f(Y_i, \boldsymbol{\theta}_i | X_i)$ with probability $(1- p_i)$

Where, $f(Y_i, \boldsymbol{\theta}_i | X_i)$ is probability distribution function (pdf) of a Poisson or Negative Binomial Distribution and $\boldsymbol{\theta}_i$ is the vector/ scalar of parameters. It follows that

$Y_i = 0$ with probability $p_i + (1- p_i) f(Y_i=0, \boldsymbol{\theta}_i | X_i)$
$= k$ with probability $(1- p_i) f(Y_i=k, \boldsymbol{\theta}_i | X_i)$, $k=1, 2, 3...$

Also, the parameters $\boldsymbol{\theta} = (\theta_{1...} \theta_n)$ and $\mathbf{p} = (p_{1...} p_n)$ satisfy,

Log $(\boldsymbol{\theta}) = \mathbf{B}\boldsymbol{\beta}$ and logit $(\mathbf{p}) = \mathbf{G}\mathbf{v}$

Thus, the design matrices for **G** and **B** contain potentially different sets of experimental factor and covariate effects that pertain to the probability of the zero state and the Poisson/Negative Binomial mean in the nonzero state, respectively (Lambert 1992). Therefore, the β's have interpretations in terms of the effect of a covariate or factor level and the **p**'s have interpretations in terms of the effect on the mean number of zeros. The default link function is the logit link, but other links such as the probit can also be used. The full set of parameters of the model and potentially the dispersion parameter (if a negative binomial count model is used) can be estimated by maximum likelihood. Inference is typically performed for all parameters except the dispersion parameter, which is treated as a nuisance parameter even if a negative binomial model is used.

Analysis of the count of publications produced during the last three years by PhD (Biochemistry) Students

The dataset for this case study is an example of over dispersed and zero-inflated counts. The response is a count of publications produced by a PhD Biochemistry student; and the dataset consists of a sample of 915 biochemistry graduate students from Long (1997) with the following variables:

- **art:** the count of articles produced during the last three years of PhD
- **fem:** the factor indicating the gender of a student, with levels Men and Women
- **mar:** the factor indicating the marital status of a student, with levels Single and Married
- **kid5:** the number of children aged five or younger
- **phd:** the prestige of the PhD department
- **ment:** the count of articles produced by a PhD mentor during the last three years

We demonstrate the use of zero-inflated models in this dataset, and examine the effects of gender differences along with other factors such as mentoring, marriage and family on productivity in terms of producing articles during the period of a PhD candidacy for Biochemistry students.

Summary of the dataset

The number of articles produced during the last three years of a PhD program varies from 0 to 19. The dataset has 494 males and 421 females. The number of articles produced by a PhD mentor during the last three years varies from 0 to 77. The Mean Count of articles produced = 1.693 which is almost one third of the variance of the number/count of articles produced = 3.710. Thus the data are over dispersed even though we have not yet considered covariates. When we look into the matrix of scatter plots (Figure 1) it seems the covariates 'phd' (again, the prestige of a PhD department) and 'ment' (again, the count of articles produced by a PhD mentor during the last three years) are correlated. The

correlation between them is equal to 0.2604. Pearson's product-moment correlation test gives a p-value of $1.110\ 10^{-15}$ indicating that the correlation is significant.

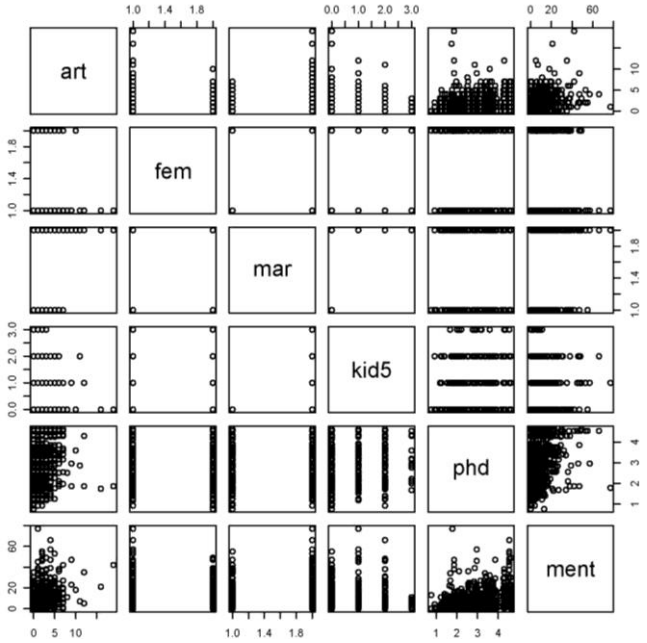

Figure 1. Matrix of scatter plots for the variables

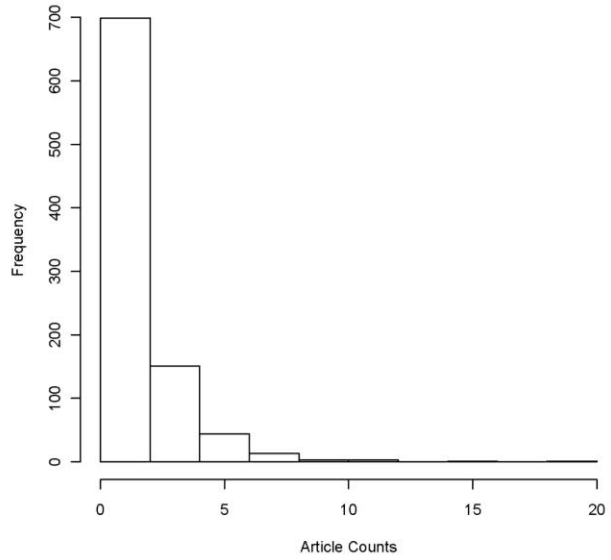

Figure 2. Histogram for the number of publications

This is understandable since the prestige of a department and the number of articles its faculty members are producing are obviously highly correlated. Furthermore, performing a best subsets regression analysis on the full data as well as the subset of data excluding zero counts yields models inclusive of the 'ment' variable but leaving out the 'phd' variable, further indicating that only one of the two variables is good enough for modeling counts. In addition, the count of articles produced during the last three years of a PhD program are clumped at zero (total zeros = 275 out of 915 counts). The histogram for the count of articles (Figure 2) further indicates that this dataset has an abundance of zeros.

To account for the extra zeros in the number of publications by PhD biochemists (about 30% in the sample didn't publish any article), we can imagine that there were two groups of PhD candidates as suggested by Germán Rodríguez (the article can be found at the following link http://data.princeton.edu/wws509/stata/overdispersion.html). For one group, the publications would not be important, while for the other group, a large number of publications would be important. The members of the first group would publish no articles, whereas the members of the second group would publish 0, 1, 2…, articles that may be assumed to have a Poisson or a Negative Binomial distribution. Therefore, this dataset is a classic example of zero-inflated and over dispersed count data.

As we know, count data are highly non-normal and are not well-estimated by ordinary least square regression. Poisson or Negative Binomial (NB) models might be more appropriate if there are no excess zeros; and a Zero-inflated Poisson (ZIP) regression does better when there is no over dispersion in the data. Therefore, we are suspecting a zero-inflated Negative Binomial (ZINB) model will provide a better fit to these data. Either a NB model or a ZIP or a ZINB could account for this over dispersion. An advantage of the NB model is that the Poisson model is nested within it. When the estimated over dispersion parameter is zero, the conditional mean is then equal to the conditional variance and the NB model reduces to the Poisson model (see Long 1997 and Cameron and Trivedi 1998 for details on nesting). Both Long (1997) and Cameron and Trivedi (1998) note that the unobserved heterogeneity that can cause over dispersion can also cause excess zeros.

The ZIP model does not allow for between-subject heterogeneity; however, the over dispersion in the raw data could be the result of a process that gave rise to the zero inflation. On the other hand, the NB model will model the between-subject heterogeneity, but it will enforce the same process for the zero and nonzero counts. As we suspect that there is a separate process for the zero and nonzero counts and for between-subject heterogeneity, we instead try modeling the dataset using ZINB. To begin with, we fit Poisson, NB, and their zero-inflated analog models to the dataset. The following table enumerates the estimates of coefficients, their standard errors and p-values for the test for significance of model coefficients for the model which includes all independent variables.

Table 1. Estimates of model coefficients and their standard errors in brackets ( ) along with the p-values for their test of significance in square brackets [ ]

Count model coefficients (Poisson/ NB with log link)	Poisson	Zero-inflated Poisson(ZIP)	Negative Binomial (NB)	Zero- Inflated Negative Binomial (ZINB)
Intercept	.305 (0.103) [0.00310]	0.641 (0.121) [1.27e-07]	0.256 (0.137) [0.0621]	0.417 (0.143) [0.003]
Gender (Female)	-0.225 (0.055) [3.92e-05]	-0.209 (0.063) [0.001]	-0.216 (0.073) [0.002887]	-0.195 (0.076) [0.010]
Marital Status (Married)	0.155 (0.061) [0.01142]	0.104 (0.071) [0.144573]	0.150 (0.082) [0.067]	0.097 (0.084) [0.248]
Child below 5 years of age	-0.185 (0.0401) [4.08e-06]	-0.143 (0.047) [0.002]	-0.176 (0.053) [0.001]	-0.152 (0.054) [0.005]
PhD	0.0128 (0.026) [0.62714]	-0.006 (0.031) [0.842]	0.0153 (0.036) [0.670]	-0.001 (0.036) [0.984]
Mentor	0.025(0.002) [<2e-16]	0.0181 (0.002) [3.07e-15]	0.029 (0.003) [<2e-16]	0.025 (0.003) [1.28e-12]
Log (theta) Dispersion Parameter			.817 (.271)	0.976 (0.135) [5.70e-13]
Zero-inflation coefficients (binomial with logit link)				
Intercept		-0.577 (0.509) [0.257]		-0.192 (1.322) [0.884]
Gender (Female)		0.110 (0.280) [0.695]		0.636 (0.848) [0.453]
Marital Status (Married)		-0.354 (0.318) [0.265]		-1.498 (0.938) [0.110]
Children below 5 years of age		0.217 (0.196) [0.269]		0.628 (0.443) [0.156]
PhD		0.001 (0.145) [0.993]		-0.037 (0.308) [0.903]
Mentor		-0.134 (0.045) [0.003]		-0.881 (0.316) [0.005]

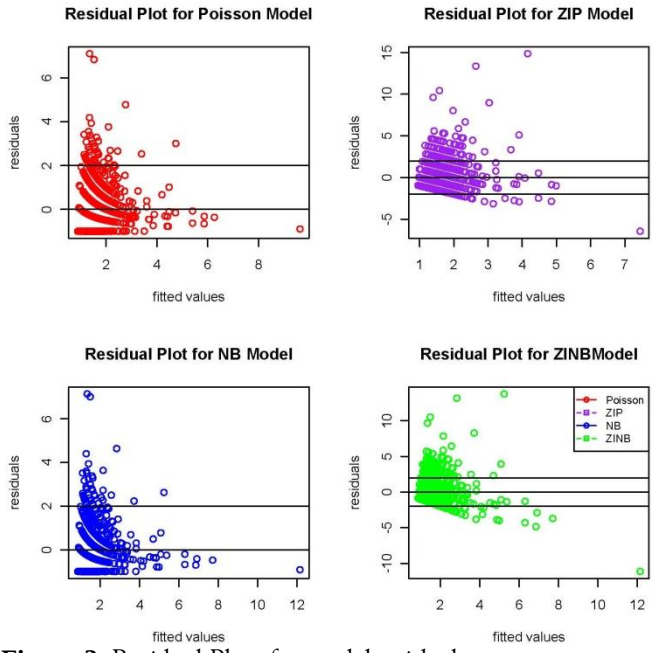

Figure 3. Residual Plots for models with the same regressors in both the zero and count portions of the model

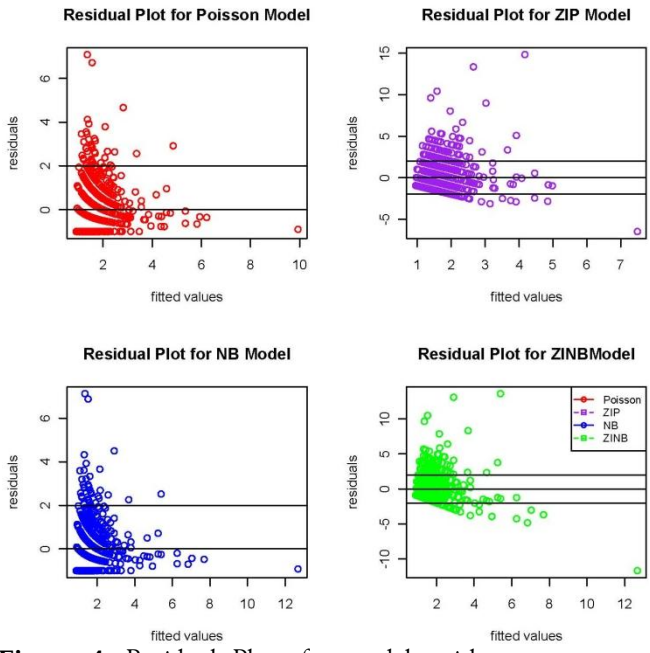

Figure 4. Residual Plots for models with no regressors (except the intercept) in the zero portion of the model

Looking at the estimates of regression parameters in Table 1 it is quite clear that all the approaches will lead to the same kind of conclusions. For instance, the negative coefficient for 'gender' and 'children' below five years of age' in all the model fits indicates that females and individuals with children under five years of age have fewer publications as compared to others. The results for the model without the independent variable 'phd' are not included here for brevity but can be obtained from the author.

Diagnostics

Residual Analysis

We will start our diagnostics for modeling this dataset by examining the residuals from the four models with and without the independent variable 'phd' in the model (since the independent variable 'phd' is significantly correlated to 'ment' as discussed earlier in the article). We plot residuals for the four models with the same regressors in the zero and count portions of the model (Figure 3) as well as the residuals for models with no regressors (except the intercept) in the zero portion of the model (Figure 4). These residual plots have several deviance residuals larger than two in absolute value with few real outliers. Looking at the residual plots, it is evident that fitting a model that takes into account over dispersion and zero inflation is a good idea. Also, looking at the residual plot of zero-inflated negative binomial

model without including independent variable 'phd' gives smaller residuals. This suggests it is better to fit a zero-inflated negative binomial model *art ~ fem + mar + kid5 + ment | 1*, that is the model which excludes the independent variable 'phd' from the count portion of the model and with no regressors for the zero portion of the model (since none of them seem to be significant).

Comparing the Current Model to a Null Model

To show that the model with regressors fits the data significantly better than the null model, i.e., the intercept-only model, we can compare the current model to a null model without predictors using chi-squared test on the difference of log likelihoods. Table 2 enumerates the chi-squared test statistic for comparing zero-inflated models with regressors to the intercept-only model. As can be seen from the table the likelihood ratio test statistic indicates that the overall zero-inflated models with regressors are statistically significant.

Table 2. Likelihood ratio test statistics

Models being Compared	Chi-Square test statistic for the likelihood ratio test (p-value)
Zero- Inflated Negative Binomial model with regressors versus the intercept-only model	$\chi^2_{(10)} = 119.89$ (<.0001)* $\chi^2_{(4)}$ 97.78 (<2.2e-16)**
Zero- Inflated Poisson model with regressors versus the intercept-only model	$\chi^2_{(10)} = 149.24$ (<.0001)* $\chi^2_{(4)} = 117.21$ (<2.2e-16)**

* Indicates model with same covariates in zero portion of the model
** Indicates model with intercept only in zero portion of the model

Akaike's Information Criterion (AIC) for each Model

We will also compute the AIC for each model. It is a very simple way to compare models with different numbers of parameters. AIC is defined as

$$AIC = -2\log L + 2p$$

where p is the number of parameters in the model. The first term is essentially the deviance and the second a penalty for the number of parameters. When comparing models, the smaller the AIC, the better the fit. Further details on the justification and properties of AIC can be found in Akaike (1974).

Table 3. AIC for the fitted models

Model	AIC
Poisson	3312.349
Negative Binomial (NB)	3134.096
Zero- Inflated Negative Binomial (ZINB)	3136.096
Zero-inflated Poisson	3253.576

The AIC values provide further support in favour of the models which take into consideration the over dispersion in the data.

Vuong`s Test to compare zero-inflated models to their counterparts

Note that the likelihood ratio test does not indicate in any way if our zero-inflated model is an improvement over its counterpart count model. We can determine this by running the corresponding standard NB model and then performing a Vuong test of the two models, see Vuong 1989; model 1 is in our case a zero-inflated model and model 2 its non zero-inflated analog). The test-statistic for the Vuong non-nested test is asymptotically distributed as a standard normal distribution under the null hypothesis that the models are indistinguishable. A large, positive test statistic provides evidence of the superiority of a zero inflated model over its non-inflated counterpart. We can see from Table 4 that the zero-inflated models are better than their non zero-inflated analogs in both versions of the zero-inflated models (with and without the independent variable **'phd'**).

Table 4. Vuong's Test Statistic

Models being Compared	Test Statistic (p-value)
ZINB versus Negative Binomial	2.242 (0.0125)* 2.123 (0.0169)**
ZIP versus Poisson	4.180 (1.454e-05)* 3.268 (0.00054)**

*Indicates model with same covariates in zero portion of the model
**Indicates model with intercept only in zero portion of the model

Comparing the observed number of zeros to the expected number of zeros

For completeness, the observed numbers of zeros in the count of articles produced along with the expected number of counts from various models considered are also enumerated below. The percentage of observed zeros in the sample is 30.05%, while the predicted percentage of zeros by Poisson model is only 20.87% which is an underestimation. However, the ZINB model with an intercept only in the zero inflation portion of the model predicts 30.38% of zeros which is quite close to the observed percentage of zeros.

Table 5. Expected number of zeros from various models

Model	Observed Counts	ZINB	NB	Poisson Model	ZIP
Models with same covariates in zero portion of the model	275*	285*	278*	191*	273*
Final model with intercept only in the zero portion of the model	275**	278**	278**	191**	264**

*Indicates model with same covariates in zero portion of the model
**Indicates model with intercept only in zero portion of the model

Final Model

All the predictors of excess zeros in the zero-inflation portions of the model (except '**ment**') were statistically non-significant; '**ment**' was significant at $\alpha = 0.01$ – see

Table 6. Since the only significant predictor of being in the 'always zero' class for ZIP and ZINB model was the number of articles published by the mentor (**'ment'**), we fitted another ZINB model using only intercept and **'ment'** as covariates in the zero portion of the model. However, this did not provide any significant improvement; if the student is not interested in producing any articles the mentor effect may not have much significance. As a result, the final model that we fit in to our Bio Chemists dataset is ZINB with the intercept only in the zero portion of the model, excluding the independent variable **'phd'**. So with that in mind, our final model is parsimonious and does not use any regressors except the intercept term in the zero portion of the model. The Binomial model is used to model the unobserved state (zero vs. count) with only an intercept term, and the count portion contains covariates for gender, marital status, the number of children under the age of five and the number of publications by a mentor.

Table 6. Final Model Coefficients

Count model coefficients (NB with log link)	Zero-inflated Negative Binomial
Intercept	0.303328 (0.082337) [0.000230]
Gender (Female)	-0.216673 (0.072672) [0.002868]
Marital Status (Married)	0.146944 (0.081675) [0.071996]
Kid5	-0.176797 (0.053054) [0.000861]
Articles by Mentor	0.029430 (0.003377) [<2e-16]
Log(theta)	0.817185 (0.119949) [9.57e-12]
Zero-inflation model coefficients (binomial with logit link)	
Intercept	-16.19 (470.21) [0.973]

Interpretation of the Final Model

In order to study the factors associated with differences in productivity (in terms of the number of publication) within the PhD (Biochemistry) stream based on gender differences we fitted the ZINB regression model to predict the count of articles produced during the last three years of PhD **(art)** from factors indicating the gender of the student **(fem),** marital status **(mar),** the number of children aged five or younger **(kid5)** and the count of articles produced by a PhD mentor during the last three years **(ment).** All the predictors except for marital status were statistically significant in the non-zero portion. Looking at the equation for the mean number of articles among those not in the always zero class, we find significant disadvantages for females and scientists with children under five, with a large positive effect of the number of publications by the mentor. As all of the predictors of count in the count portion (except **'mar'**) of the model are statistically significant. Furthermore, the collaboration with the mentor is found to be the most vital factor affecting productivity. For females, opportunities for productivity are significantly decreased by having young children as can be seen from the negative coefficient (-0.177) for the indicator function for children as well as the negative co-efficient associated with the indicator variable for gender (-0.217).

Conclusions

I fitted several models using the commonly used and open sourced statistical package R. However several other statistical packages can perform a similar analysis. For instance, another widely used statistical package is SAS where we can use "proc countreg" to achieve similar analysis. For further details on using SAS for such an analysis, interested readers can refer to the following link:

http://www2.sas.com/proceedings/forum2008/322-2008.pdf

We can see in our model that the dispersion parameter **Log(theta)**=0.817 is significantly different from zero. This suggests that the counts are overdispersed, and that a NB model is more appropriate than a Poisson model. Vuong's test further suggests that our zero-inflated model is a significant improvement over a standard NB model. Thus, for our data the ZINB is a clear winner in terms of parsimony and goodness of fit.

References

Akaike, H. 1974. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* 19(6), 716-723.

Cameron, A. C. and Trivedi, P. K. 1986. Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators. *Journal of Applied Econometrics*, 1, 29-53.

Cameron, A. C. and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

Everitt, B. S. and Hothorn, T. 2006. A Handbook of Statistical Analyses Using R. Boca Raton, FL: Chapman and Hall.

Greene, W. H. 1994. Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. Technical report. Department of Economics. New York University.

Introduction to SAS 2011. UCLA: Academic Technology Services, Statistical Consulting Group. Available at http://www.ats.ucla.edu/stat/r/dae/zinbreg.htm.

Lambert, D. 1992. Zero-Inflated Poisson Regression Models with an Application to Defects in Manufacturing. *Technometrics*, 34, 1-14.

Long, J. S. 1997. *Regression Models for Categorical and Limited Dependent Variables*, Thousand Oaks, CA: Sage Publications.

Long, J. Scott. 1990. The origins of sex differences in science. *Social Forces*. 68(3), 1297-1316.

Vuong, Q. H. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57, 307-333.

Correspondence : rizvitz@gmail.com