11-1-2003

Using Zero-inflated Count Regression Models To Estimate The Fertility Of U. S. Women

Dudley L. Poston Jr.
Texas A&M University, dudleyposton@yahoo.com

Sherry L. McKibben
Texas A&M University, sherrymckibben@yahoo.com

Follow this and additional works at: http://digitalcommons.wayne.edu/jmasm

Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Using Zero-inflated Count Regression Models
To Estimate The Fertility Of U. S. Women

Dudley L. Poston, Jr.   Sherry L. McKibben
Department of Sociology
Texas A&M University

In the modeling of count variables there is sometimes a preponderance of zero counts. This article concerns the estimation of Poisson regression models (PRM) and negative binomial regression models (NBRM) to predict the average number of children ever born (CEB) to women in the U.S. The PRM and NBRM will often under-predict zeros because they do not consider zero counts of women who are not trying to have children. The fertility of U.S. white and Mexican-origin women show that zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models perform better in many respects than the Poisson and negative binomial models. Zero-inflated Poisson and negative binomial regression models are statistically appropriate for the modeling of fertility in low fertility populations, especially when there is a preponderance of women in the society with no children.

Key words: Poisson regression, negative binomial regression, demography, fertility, zero counts

Introduction

When analyzing variation in the number of children that women have born to them, demographers frequently use Poisson and negative binomial regression models rather than ordinary least squares models. Poisson and negative binomial regression models are statistically more appropriate for predicting a woman's children ever born (CEB), particularly in societies where mean fertility is low (Poston, 2002). Most women in such populations have children at the lower parities, including zero parity, and few have children at the higher parities. The CEB variable, which by definition

Dudley L. Poston, Jr. is Professor of Sociology, and the George T. and Gladys H. Abell Professor of Liberal Arts. He is co-editing (with Michael Micklin) the *Handbook of Population* (Klewer Plenum, 2004). E-mail him at: dudleyposton@yahoo.com. Sherry L. McKibben is a Lecturer in the Department of Sociology. E-mail: sherrymc. sherrymckibben@yahoo.com.

is a count variable, i.e., a nonnegative integer, is hence heavily skewed with a long right tail.

The statistical modeling of these kinds of CEB data is best based on approaches other than the ordinary least squares (OLS) linear regression model because using it to predict a count outcome, such as CEB, will often "result in inefficient, inconsistent, and biased estimates" (Long, 1997, p. 217) of the regression parameters. Poisson regression models (PRM) and negative binomial regression models (NBRM) have been shown to be statistically more appropriate (Poston, 2002).

However, sometimes there are so many zeros in the count dependent variable that both the PRM and the NBRM under-predict the number of observed zeros; the resulting regression models, therefore, often do not fit the data. Zero-inflated count regression models were introduced by Lambert (1992) and Greene (1994) for those situations when the PRM and the NBRM failed to account for the excess zeros and resulted in poor fit. This paper examines the use and application of zero-inflated count regression models to predict the number of children ever born to U.S. women.

Methodology

The most basic approach for predicting a count variable, such as CEB, is the Poisson regression model (PRM). In the PRM, the dependent variable, namely, the number of events, i.e., in the case of this paper, the number of children ever born (CEB), is a nonnegative integer and has a Poisson distribution with a conditional mean that depends on the characteristics (the independent variables) of the women (Long, 1997; Long & Freese, 2001). The PRM incorporates observed heterogeneity according to the following structural equation:

$$\mu_i = \exp\,(a + X_{1i}\,b_1 + X_{2i}\,b_2 + ... + X_{ki}\,b_k)$$

where μ_i is the expected number of children ever born for the ith woman; X_{1i}, X_{2i} ... X_{ki} are her characteristics; and a, b_1, b_2 ... b_k are the Poisson regression coefficients.

The PRM is appropriate when the mean and the variance of the count distribution are similar, and is less applicable when the variance of the distribution exceeds the mean, that is, when there is over-dispersion in the count data. If there is significant over-dispersion in the distribution of the count, the estimates from the PRM will be consistent, but inefficient. "The standard errors in the Poisson regression model will be biased downward. Resulting in spuriously large z-values and spuriously small p-values" (Long & Freese, 2001; Cameron & Trivedi, 1986), which could lead the investigator to make incorrect statistical inferences about the significance of the independent variables.

This is addressed by adding to the PRM "a parameter that allows the conditional variance of (the count outcome) to exceed the conditional mean" (Long, 1997, 230). This extension of the Poisson regression model is the negative binomial regression model (NBRM). The NBRM adds to the Poisson regression model the error term ε according to the following structural equation:

$$\mu_i = \exp\,(a + X_{1i}\,b_1 + X_{2i}\,b_2 + ... + X_{ki}\,b_k + \varepsilon_i)$$

However, sometimes there are many more zeros in the count dependent variable than are predicted by the PRM or NBRM, resulting in an overall poor fit of the model to the data. Zero-inflated models respond to this problem of excess zeros "by changing the mean structure to allow zeros to be generated by two distinct processes" (Long & Freese, 2001, p. 250).

Consider a few examples of excess zeros. Suppose one wishes to survey visitors to a national park to predict the number of fish they caught. Suppose that some of the visitors did not fish, but data were not available on who fished and who did not fish. The data gathered hence have a preponderance of zeros, some of which apply to persons who fished and caught no fish, and others to persons who did not fish (Stata, 2001; Cameron & Trivedi, 1998).

Or consider the problem of predicting the number of publications written by scientists. Some scientists will never publish either because they have chosen not to do so, or, perhaps, because they are not permitted to do so. But assume that there are no data telling which scientists have a zero probability of ever publishing. As with the example of the number of fish caught, there will be a preponderance of zeros among scientists with regards to the number of articles published. Some of the zeros will apply to scientists who tried to publish but were not successful and others to scientists who did not try to publish (Long & Freese, 2001; Long, 1990).

Finally, consider the example to be addressed in this paper, namely, the number of children born to women. Some women will choose not to have children and are referred to as voluntarily childless women. Other women will try to have children but will not be successful in their attempts and are referred to as involuntarily childless women (Poston, 1976; Poston & Kramer, 1983). But, assume that it is not directly known to which group each woman belongs. Thus among women of the childbearing ages of 15-49, there will be many zeros on the CEB dependent variable; some of the zeros will apply to women who tried to produce children but were not successful, and others to women who voluntarily opted against having children.

Long and Freese (2001) stated that in zero-inflated models it is assumed that "there are two latent (i.e., unobserved) groups. An individual in the Always-0 Group (Group A) has

an outcome of 0 with a probability of 1, while an individual in the Not Always-0 Group (Group ~A) might have a zero count, but there is a nonzero probability that she has a positive count" (p. 251).

In all cases, the investigator does not know into which of the two groups the respondents fall. If it was known into which group each subject was placed, one could subtract the persons belonging to the Always-0 Group from the total sample, and estimate Poisson or negative binomial regression models. But typically one does not have this kind of information, thus requiring the introduction of zero-inflated regression.

The estimation of zero-inflated regression models involves three steps: 1) predicting membership in the two latent groups, Group A and Group ~A; 2) estimating the number of counts for persons in Group ~A; and 3) computing "the observed probabilities as a mixture of the probabilities for the two groups" (Long & Freese, 2001, p. 251).

To analyze the fertility of U.S. women, one would follow these steps (for detail, see Long & Freese, 2001. p. 251-252; Cameron & Trivedi, 1998, p. 125-127, 211-215).

In Step 1, use a logistic regression model to predict the woman's membership in Group A (never have children) or Group ~A (may or may not have children). The independent variables used in the logistic equation may be "referred to as inflation variables since they serve to inflate the number of 0s" (Long & Freese, 2001, p. 251).

In Step 2, for women in Group ~A (may or may not have children), depending on whether or not there is over-dispersion in the CEB dependent variable, use either a Poisson regression model or a negative binomial regression model to predict the probabilities of counts 0 to y (where y is the maximum number of children born to a woman). The independent variables used in Step 2 may or may not be the same as those used in Step 1. In the examples shown below, the same independent variables are used in both steps. Using the same variables in both steps is not required. Different variables could be used in each step.

In Step 3, the results from the preceding steps are used to determine the overall probability of 0's, which is "a combination of the probabilities of 0's from each group, weighted by the probability of an individual (woman) being in the group" (Long, 1997, p. 242-243). The probabilities of counts other than 0 are adjusted in a similar way.

Results

Data are available for 1995 for U.S. (non-Hispanic) white and Mexican-origin women, gathered in Cycle 5 of the National Survey of Family Growth (National Center for Health Statistics, 1995). The data are based on personal interviews conducted in the homes of a national sample of 10,847 females between the ages of 14 and 44 in the civilian, non-institutionalized population in the United States. Table 1 reports the descriptive data on children born (CEB) for U.S. white and Mexican-origin women in 1995.

White women have a mean CEB of 1.2 with a variance of 1.6. Mean CEB for Mexican-origin women is 1.9 with a variance of 2.8. For both white and Mexican-origin women, the variance of CEB is greater than the mean of CEB. There are several ways for determining if there is over-dispersion in the CEB data (see Poston, 2002). It turns out that there is not a significant amount of over-dispersion in the CEB data for whites, justifying the use of a Poisson regression model. There is a significant amount of over-dispersion in the CEB data for Mexican-origin women, so that a negative binomial regression model will be appropriate.

Poisson Regression versus Zero-inflated Poisson Regression

A Poisson regression model is thus estimated for the white women that predicts their CEB with socioeconomic and location characteristics that have been shown in the demographic literature to be associated with fertility. The independent variables pertain to education, rural residence, poverty status, age, regional location, and religion. Some are measured as dummy variables and others as interval.

Table 1. Data for Children Ever Born: U.S. White and Mexican-Origin Women, Ages 15-49.

Group	Mean	Standard Dev.	Variance	No. of Cases
White	1.2471	1.2839	1.6486	6,456
Mexican	1.8864	1.6592	2.7531	924

Source of Data: National Center for Health Statistics (1995).

They are the following: X_1 is the woman's education measured in years of school completed; X_2 is a dummy variable indicating whether the woman lives in a rural area; X_3 is a dummy variable indicating whether the woman is classified as being in poverty (poverty status is based on whether the woman's family income is below the national poverty threshold, adjusted for family size).

Continuing, X_4 is the woman's age measured in years; X_5 to X_7 are three dummy variables representing the woman's region of residence, namely, X_5 residence in the Midwest, X_6 residence in the South, and X_7 residence in the West; residence in the Northeast is the reference category; and X_8 to X_{10} are three dummy variables reflecting the woman's religion, as follows: X_8 indicates if the woman's religion is Protestant, X_9 if she is Catholic, and X_{10} if she has no religion, or religion is not specified; Jewish religion is the reference category. The first panel of Table 2 reports the results of the Poisson regression equation predicting CEB for U.S. white women in 1995.

According to the Poisson coefficients shown in the first panel of Table 2, four of the ten independent variables are significantly related with the CEB of white women. The higher the woman's education, the fewer her CEB; the older her age, the higher her CEB. If she is a rural resident or in poverty, she will have more children than urban residents or women not living in poverty. The geographic location and religion variables are not statistically significant.

Using the above Poisson regression results, the predicted probabilities of each white woman may be calculated for each count of CEB from 0 to 10. The mean of the predicted probabilities at each count may then be determined, using this formula (Long & Freese, 2001):

$$\overline{\Pr}(y = m) = \frac{1}{N} \sum_{i=1}^{N} \hat{\Pr}(y_i = m \mid x_i)$$

where y = m = the count of children ever born, and x_i are the above ten independent variables.

Table 2. Poisson Regression Model, and Zero-inflated Poisson Regression Model, U.S. White (non-Hispanic) Women, 1995.

Independent Variable	Poisson Model		Zero-inflated Poisson Model			
			Logit		Poisson	
	b	z	b	z	b	z
	Panel 1		Panel 2		Panel 3	
X_1 Education	-.070	-15.45	.905	12.83	-.058	11.84
X_2 Rural Residence	.111	3.87	-.336	-1.55	.097	3.35
X_3 Poverty Status	.377	10.48	-.482	-2.06	.336	8.96
X_4 Age	.076	48.41	-.781	-14.79	.034	16.01
X_5 Midwest	.045	1.37	-.714	-2.63	.007	.22
X_6 South	-.023	-.07	-.289	-1.06	-.048	-1.39
X_7 West	.019	.53	-.187	-0.63	.011	.31
X_8 Protestant	.074	.80	-2.483	-3.02	-.030	-.31
X_9 Catholic	.052	.56	-1.871	-2.28	-.035	-.37
X_{10} No Religion	-.110	-1.14	-2.020	-2.15	-.223	-2.23
Constant	-1.504	-11.98	8.780	8.12	.096	.14
Likelihood Ratio χ^2	2857.59, P = 0.000		458.66, P = 0.000			

Vuong Test of Zip vs. Poisson: 20.16 P = 0.000

Figure 1 is a plot of the mean Poisson predicted probabilities at each count of CEB (the green x symbols), and they may be compared with the observed empirical distribution of CEB (the blue circles). Just over 40 percent (proportion of .4046) of U.S. white women have no children ever born, but the Poisson regression results predict a mean probability at zero count of .361, which is an under-prediction of the observed CEB. The Poisson regression results over-predict the observed CEB data at count one, under-predict at count two, and are more consistent with the observed CEB data at the third and higher counts.

But, a central issue for this paper is the under-prediction by the Poisson regression model of the observed zero counts of CEB for white women. In such a situation, it would be appropriate to estimate a zero-inflated Poisson regression model. The 2nd and 3rd panels of Table 2 present the results of such a model. Recall from the previous section that the first two steps in estimating a zero-inflated model involve 1) using a logistic regression model to predict the woman's group membership in Group A (never have children) or Group ~A (may or may not have children), and 2) for women in Group ~A (may or may not have children), using a Poisson regression model to predict her number of children ever born. Thus there are two panels of zero-inflated Poisson results reported in Table 2. Panel 2, titled "Logit" are the logit coefficients obtained in Step 1, and Panel 3, titled "Poisson" are the Poisson coefficients obtained in Step 2.
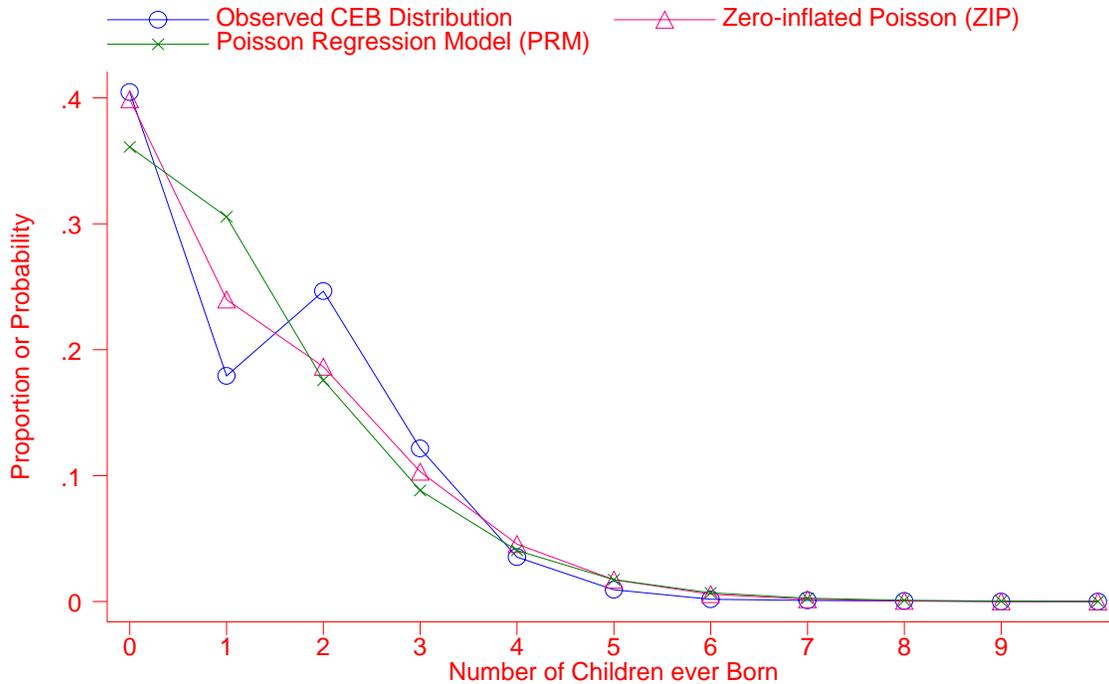
Figure 1. Distributions of CEB, PRM, & ZIP, U.S. White Women

The coefficients in the "Logit" panel (Panel 2) of Table 2 are the logit coefficients predicting a woman's membership in Group A (never having children). The higher her education the greater the likelihood of her not having children. If she is in poverty, she is likely to not have children. The older her age, the less likely she will not have children. If she lives in the Midwest, she will be less likely than women living in the Northeast to not have children. And if she is a Catholic, or a Protestant, or has no religion, she will be less likely than Jewish women to have no children. The rural, South, and West variables are not significant.

For the purpose of this paper, the more relevant coefficients are shown in the "Poisson" panel (Panel 3) of Table 2; these are the zero-inflated Poisson coefficients predicting the woman's CEB. The higher her education, the less number of children shill will have. If she is a rural resident, or in poverty, she will have more children. The older her age, the more the children. If she has no religion, she will have fewer children than Jewish women. The other variables are not significant.

A relevant comparison is between the zero-inflated Poisson coefficients (Panel 3) and

the Poisson coefficients (Panel 1). Note first that the Poisson coefficients (Panel 1) for most of the independent variables are slightly larger than those for the zero-inflated Poisson coefficients (Panel 3). However, the z-scores for many of the Poisson coefficients are quite a bit larger than the z-scores for the corresponding zero-inflated Poisson coefficients. Thus although the two sets of Poisson coefficients are not too different in magnitude, the standard errors for the zero-inflated coefficients will tend to be larger than they are for the Poisson coefficients.

Regarding issues of interpretation and statistical inference, the results of the two Poisson models allow the investigator to conclude that the effects on a woman's CEB of her education, rural residence, poverty status and age are all statistically significant. However, the zero-inflated Poisson results, but not the basic Poisson results, also allow the investigator to conclude that the "no religion" variable has a statistically significant negative effect on CEB. Women who report no religion have fewer children than women in the reference (Jewish religion) category. This inference would not have been made using the results of the Poisson model (Panel 1).

Is the zero-inflated Poisson regression model (ZIP) statistically preferred over the basic Poisson model regression model (PRM)? There is a formal test statistic, the Vuong test (Vuong, 1989) that determines statistically whether the zero-inflated model is a significant improvement over the Poisson model (for details, see Long, 1997, p. 248; Long & Freese, 2001, p. 261-262). The Vuong statistic is asymptotically normal; if its value is > 1.96, the ZIP model is preferred over the PRM. If Vuong < 1.96, the PRM is preferred. The Vuong test statistic is shown at the base of Table 2, Vuong $= 20.16$. This is clear evidence that the zero-inflated Poisson regression results are preferred over the Poisson regression results.

Another way to judge whether ZIP is preferred over PRM is to ascertain if the results from a ZIP regression improve the prediction of the mean probability at count zero. Recall from the discussion of Figure 1 (above) that the Poisson regression results under-predicted the mean probability at count zero. Figure 1 also contains mean predicted probabilities at each count that are based on the results of the zero-inflated Poisson model.

The ZIP predictions are shown in Figure 1 as maroon diamonds. The ZIP model predicts a probability at count zero of .3988, which is very close to the observed proportion of CEB at count zero of .4046. The expected probabilities from the ZIP model at counts 1, 2 and 3 are also closer to the corresponding observed CEB counts than are those predicted by the PRM. The ZIP results seem to do a much better job predicting the observed CEB at counts 0, 1, 2, and 3 than do the results from the PRM.

Negative Binomial Regression versus Zero-inflated Negative Binomial Regression

In the above example predicting CEB for U.S. white women, it was first determined that there was not a significant amount of over-dispersion in CEB, thus justifying modeling CEB with the PRM. But recall that for the U.S. Mexican-origin women, the variance of CEB was significantly greater than the mean of CEB. In such a case, the PRM is not appropriate. Instead a negative binomial regression model (NBRM) is preferred.

The first panel of Table 3 reports the results of a negative binomial regression (NBRM). The same independent variables are used in this regression, as were used in the regressions shown in Table 2, except that age is excluded and "no religion" is used as the reference religion category. Excluding age resulted in a better fit of the negative binomial model with the data. The "no religion" variable is removed from the equation and used as the reference category because the Jewish variable was removed altogether from the equation; only 2 of the 924 Mexican-origin women were Jewish, so there was insufficient variation in this variable. Thus, regression results are shown in Table 3 for eight independent variables.

The negative binomial regression coefficients in the first panel of Table 3 indicate that only two of the eight independent variables are significantly related with the CEB of Mexican-origin women. The higher the woman's education, the lower her fertility; and if she is living in poverty, she will have more children than women not in poverty. The other independent variables are not statistically significant.

There is a large number of zeros for the CEB of Mexican-origin women. Almost 27 percent of them have zero children. Although this is not quite as high as the level of zero parity among white women (40.4 percent of the white women have no children), a zero-inflated negative binomial regression model (ZINB) was estimated to see if model fit would be improved over that of the NBRM. Its results may be compared with those of the NBRM shown in the first panel of table 3.

Recall that zero-inflated models produce two sets of coefficients (see the discussion above). Thus, the coefficients in the "Logit" panel (Panel 2) of Table 3 are the logit coefficients predicting a Mexican-origin woman's membership in Group A (never having children). The higher her education the greater her likelihood of not having children. And if she is Catholic, she is less likely than women with no religion to have no children. The other independent variables are not significant.

Table 3. Negative Binomial Regression Model, and Zero-inflated Negative Binomial Regression Model, U.S. Mexican-origin Women, 1995

Independent Variable	Negative Binomial Model		Zero-inflated Negative Binomial Model			
			Logit		Negative Binomial	
	b	z	b	z	b	z
	Panel 1		Panel 2		Panel 3	
X_1 Education	-.076	-9.13	.120	3.04	-.058	7.35
X_2 Rural Residence	.156	1.36	.150	.33	.177	1.57
X_3 Poverty Status	.231	3.84	-.089	-.29	.213	3.51
X_4 Midwest	.351	.85	9.288	.04	.698	1.76
X_5 South	.442	1.09	8.922	.04	.697	1.82
X_6 West	.478	1.18	8.951	.04	.736	1.93
X_7 Protestant	.202	1.72	-.491	-1.18	.084	.69
X_8 Catholic	.180	1.71	-.903	-2.41	-.004	-.03
Constant	.694	1.63	-11.176	-.05	.589	1.46
Likelihood Ratio χ^2	126.36, P = 0.000			98.05, P = 0.000		

Vuong Test of Zero-inflated Negative Binomial versus Negative Binomial 70.67, P = 0.000

A comparison may be made between the zero-inflated negative binomial coefficients (Panel 3) and the negative binomial coefficients (Panel 1). Note first that the NBRM coefficients (Panel 1) for four of the independent variables are slightly larger than those for the ZINB coefficients (Panel 3). And the z-scores for four of the NBRM coefficients are larger than those for the corresponding ZINB coefficients. The results of the two models in Panels 1 and 3 allow the investigator to conclude that the effects on a woman's CEB of her education and poverty status are statistically significant. However, the zero-inflated negative binomial regression results, but not the negative binomial results, also allow the investigator to conclude that Mexican-origin women living in the West have more children than those living in the Northeast. The NBRM results did not allow this inference to be made.

One may compare the ZINB regression results with the NBRM results to determine if one is statistically preferred over the other. The Vuong test statistic provided at the base of Table 3 has a value of 70.67. Clearly the zero-inflated negative binomial regression results are preferred over the basic negative binomial regression results.

Conclusion

This article considered a situation that frequently occurs when modeling count variables, namely, that there is a preponderance of zero counts. The application addressed in this paper involved the estimation of Poisson regression models (PRM) and negative binomial regression models

(NBRM) to predict the average number of children ever born (CEB) to women in the U.S. This is a count variable, and in a low fertility society such as the U.S., it is skewed with a long right tail.

It was noted in this article that many U.S. women have no children, resulting in a very large percentage of zero counts. But two groups of women have no children; one group will have zero CEB because they have chosen to never have children; another group will have no children even though they are trying to do so. PRM and NBRM are best suited to predict CEB counts among women who are having, or trying to have, children. Thus these models end up under-predicting zero counts because strictly speaking they are not able to consider the zero counts of women who are not trying to have children. Zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models have been proposed to handle such situations.

Analyses conducted in this paper of the fertility of U.S. white and Mexican-origin women in 1995 demonstrated that the zero-inflated models performed better in many respects than the straightforward Poisson and negative binomial models. Not only were the coefficients in the ZIP and ZINB models different from those in the PRM and NBRM, it was also shown that errors of statistical inference, in terms of failing to include significant effects, would have been made had the investigator only relied on the results of the PRM and NBRM.

It would appear that the use of zero-inflated Poisson and negative binomial regression models are statistically appropriate for the modeling of fertility in low fertility populations. This is especially the case when there is a preponderance of women in the society with no children.

References

Cameron, A. C., & Trivedi, P. K. (1986). Econometric models based on count data: comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*, *1*, 29-53.

Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data*. Cambridge, U.K.: Cambridge University Press.

Greene, W. H. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. *Stern School of Business, Department of Economics, Working Paper Number 94-10*.

Lambert, D. (1992). Zero-inflated Poisson regression with an application to defects in manufacturing. *Technometrics*, *34*, 1-14.

Long, J. S. (1990). The origins of sex differences in science. *Social Forces*, *68*, 1297-1315.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, California: Sage Publications.

Long, J. S., & Freese, J. (2001). *Regression models for categorical dependent variables using Stata*. College Station, Texas: Stata Press.

National Center for Health Statistics (NCHS). (1995). *National Survey of Family Growth, Cycle V*. Hyattsville, Maryland: Department of Health and Human Services.

Poston, D. L., Jr. (1976). Characteristics of voluntarily and involuntarily childless wives. *Social Biology*, *23*, 198-209

Poston, D. L., Jr. (2002). The statistical modeling of the fertility of Chinese women. *Journal of Applied Statistical Methods*, *1*(2), 387-396.

Poston, D. L., Jr., & Kramer, K. B. (1983). Voluntary and involuntary childlessness in the United States, 1955-1973. *Social Biology*, *30*, 290-306.

StataCorp. (2001). *Stata statistical software: Release 7.0*. Vol. IV, College Station, Texas: Stata Corporation, 481.

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