Journal of Modern Applied Statistical Methods

Volume 1 | Issue 2 Article 47

11-1-2002

The Statistical Modeling Of The Fertility Of Chinese Women

Dudley L. Poston Jr.
Texas A&M University

Recommended Citation
Poston, Dudley L. Jr. (2002) "The Statistical Modeling Of The Fertility Of Chinese Women," *Journal of Modern Applied Statistical Methods*: Vol. 1 : Iss. 2 , Article 47.
DOI: 10.22237/jmasm/1036109160

Journal of Modern Applied Statistical Methods
Fall 2002, Vol. 1, No 2, 387-396

1538 – 9472/02/$30.00

The Statistical Modeling Of The Fertility Of Chinese Women

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

This article is concerned with the statistical modeling of children ever born (CEB) fertility data. It is shown that in a low fertility population, such as China, the use of linear regression approaches to model CEB is statistically inappropriate because the distribution of the CEB variable is often heavily skewed with a long right tail. For five sub-groups of Chinese women, their fertility is modeled using Poisson, negative binomial, and ordinary least squares (OLS) regression models. It is shown that in almost all instances there would have been major errors of statistical inference had the interpretations of the results been based only on the results of the linear regression models.

Key words: Poisson, negative binomial, OLS, modeling Chinese fertility

Introduction

The national censuses of many countries include a question that asks women about the number of children they have had ever born to them; these are referred to as children ever born (CEB) data. Demographers often use such data in statistical models of fertility. CEB data may be referred to as event count or count data. “An event count refers to the number of times an event occurs... (and) is the realization of a nonnegative integer-valued random variable” (Cameron & Trivedi, 1998, p. 1). For many count variables, such as the CEB variable, its distribution is heavily skewed with a long right tail. This is certainly the case in low-fertility populations, such as China, the population analyzed in this article. This reflects the fact that most women in such populations have children at the lower parities, including zero parity, and few have children at the higher parities. In this paper CEB data from the 1990 census of China are analyzed for five sub-groups of ever-married women. It is shown that the use of linear regression to model CEB for these sub-groups is statistically inappropriate.

Table 1 (all tables and figures are in the appendix) is a compilation of descriptive information on the CEB variable for ever-married Chinese women aged 15-49 from five sub-groups, namely, the Han (the majority nationality group), and four of China’s 55 minority groups (the Korean, Manchu, Hui and Uygur minorities). The Han women have an average of 2.13 children ever born. The Korean and Manchu women have mean CEB values that are less than that of the Han, both with values of 1.8. Hui women have a mean CEB of 2.33, and Uygur women report one of the higher average CEB values of any of the Chinese minority nationalities, a mean of 3.16. Tables and figures appear at the end of this paper.

Figures 1-5 (appendix) show frequency distributions of the observed CEB data (the blue lines with circles as symbols) for these five sub-groups: Han women (Figure 1), Manchu women (Figure 2), Korean women (Figure 3), Uygur women (Figure 4), and Hui women (Figure 5). For Han women (Figure 1), about 8 percent have no children, over 30 percent have one, about 29 percent have two, 19 percent have three, 9 percent have four, 4 percent have five, and progressively

Dudley L. Poston, Jr. is Professor of Sociology, and the George T. and Gladys H. Abell Professor of Liberal Arts, at Texas A&M University. He has co-authored/edited ten books and over 200 refereed journal articles, chapters and reports on various sociological, statistical and demographic topics. He is currently co-editing (with Michael Micklin) the *Handbook of Population*, scheduled to be published in 2004 by Kluwer Academic/ Plenun. Email: dudleyposton@yahoo.com.

smaller percentages of women have children at the higher parities. The Han distribution is heavily skewed with a long right tail. This characterization also applies to the Manchu, Korean and Hui distributions. Only the Uygur women (Figure 4), with one of the highest fertility rates in China, do not show as skewed a CEB distribution as the others, although their distribution too has a long right tail.

A major point is that none of the distributions in Figures 1-5 is normally distributed, and most are heavily skewed, and all have long right tails. Therefore, the statistical modeling of these kinds of CEB data should be based on approaches other than the ordinary least squares (OLS) linear regression model. Using an OLS model to predict a count outcome, such as children ever born, will often "result in inefficient, inconsistent, and biased estimates" (Long, 1997, p. 217) of the regression parameters.

Methodology

There are several alternative models that take into account the characteristics of a count variable such as CEB. The most basic is the Poisson regression model in which "the probability of a count (of CEB) is determined by a Poisson distribution, where the mean of the distribution is a function of the independent variables" (Long, 1997, pp. 217-218), which, in this case would be the characteristics of the individual women. The Poisson regression model, and alternate models such as the negative binomial regression model and some types of zero-inflated regression models, are based on the univariate Poisson distribution, which will now be considered.

The Univariate Poisson Distribution

Figures 1-5 also show for the five sub-groups of Chinese women the univariate Poisson distributions (the purple lines with triangle symbols) that correspond to the mean CEB values for the respective groups. The shape of the univariate Poisson distribution depends entirely on the value of the mean, and is based on the following formula:

$$\mathbf{Pr}(Y = y) = \frac{\mathbf{exp}(-\mu)\mu^{y}}{y!} \quad , \quad y = \mathbf{0, 1, 2, ...}$$

where the parameter μ represents the mean, and **y** is an integer indicating the number of times the count has occurred, ranging from 0 to some higher positive integer.

This purely theoretical distribution was developed by the French mathematician Simeon-Denis Poisson (1781-1840) and is fundamental in the statistical analysis of an assortment of issues involving radioactivity, traffic, and many other count events that occur in time and/or space.

Some properties of the theoretical Poisson distribution are (Long & Freese, 2001, p. 224):

1)With increasing values of the mean, μ, the shape of the distribution moves to the right; this is seen in the above CEB distributions;

2) The variance of the univariate Poisson distribution equals the mean, μ, a property known as equi-dispersion. Empirically, however, the variance of many count variables tends to be greater than the mean. To illustrate, the descriptive CEB data in Table 1 indicate that the variance of CEB for Uygur women is more than twice its mean. The variance of CEB for Hui women is also larger than its mean.

3) As μ increases, the probability of zero counts decreases.

4) As μ increases, the Poisson distribution approximates a normal (Gaussian) distribution.

Consider once again Figures 1-5. Observe their empirical distributions of children ever born, and compare these distributions with the univariate Poisson distributions that correspond to their mean CEB values. For Han women (Figure 1), the fitted Poisson distribution (the purple line with triangle symbols) slightly over-predicts the observed proportion of women with zero children, under-predicts the proportion with one child, slightly under-predicts the proportion with two children, and predicts fairly well the proportions of women at the higher parities. The univariate Poisson distributions for the other four nationality groups of Chinese women also show various patterns of under-prediction and over-prediction of the numbers of women at most of the different counts of children ever born. In some cases these patterns of under- and over-prediction are similar to those of the Han Chinese shown in Figure 1, and in other cases they are not.

One should not expect the univariate Poisson distributions to perfectly predict the proportions of women at each count of CEB because the Poisson distributions do not take into account the heterogeneity of the women. That is, one reason why the Poisson distributions shown in Figures 1-5 do not perfectly fit the observed CEB distributions is that the women in the five samples vary in the numbers of children they produce. It would be unrealistic to expect that all Han women have the same rate of child production, that all Manchu women have the same rate, and similarly for the other groups of women. The researcher needs to introduce heterogeneity into the models by drawing on the observed characteristics of the women. Therefore, the issue of statistical modeling will now be considered and the results of the analyses presented.

Results

Most demographic analyses of CEB have used linear regression models (e.g., see Ritchey, 1975; Johnson, 1979; Janssen and Hauser, 1981; Entwisle and Mason, 1985; Bean and Tienda, 1987). This is an appropriate statistical strategy if the mean CEB count is high because in such a situation the distribution of the dependent variable tends to be approximately normal. But if the mean of the counts is not high, as is the case with children ever born responses of women in low-fertility populations, then the "common regression estimators and models, such as ordinary least squares in the linear regression model, ignore the restricted support for the dependent variable" (Cameron &Trivedi, 1998, p. 2).

There is a host of regression models that may be used in the analysis of count data (see Cameron & Trivedi, 1998). The Poisson regression model is the most basic and the standard model for analyzing count outcomes and is derived from the Poisson distribution. The Poisson regression model is an appropriate strategy 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. In such instances an alternate modeling approach would be negative binomial regression.

The Poisson Regression Model

In a Poisson regression model, the dependent variable, namely, the number of events, i.e., the number of children ever born, is a nonnegative integer and has a Poisson distribution with a conditional mean that depends on the characteristics (the independent variables) of the women. The model thus 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 i^{th} woman; $\mathbf{X}_{1i}$, $\mathbf{X}_{2i}$ ... $\mathbf{X}_{ki}$ are her characteristics; and **a**, b_1, b_2 ... b_k are the Poisson regression coefficients.

The Poisson regression model is a nonlinear model, predicting for each individual woman the number of children she has had ever born to her, μ_i. The **X** variables are related to μ nonlinearly. Some applications of the Poisson regression model will now be illustrated in separate statistical analyses of children ever born for Han, Korean, and Manchu women, using data from the 1% Sample of the 1990 Census of China. The Chinese samples have been restricted to ever-married women between the ages of 15 to 49. Poisson models would appear to be appropriate for estimating CEB for the Han, Manchu and Korean because their mean and variance CEB values are so similar (Table 1).

A selection of independent variables is used that reflect socioeconomic and locational characteristics that have been shown to be associated with fertility. The independent variables pertain to age, education, residence, regional location, and marital status. Some are measured as dummy variables, and others as interval. They are the following:$\mathbf{X}_1$ is the woman's age measured in years (age); $\mathbf{X}_2$ to $\mathbf{X}_5$ are four dummy variables representing the levels of education of the women, namely, $\mathbf{X}_2$, completed at least some elementary school; $\mathbf{X}_3$, completed at least some middle school; $\mathbf{X}_4$, completed at least some high school; and $\mathbf{X}_5$, completed at least some college; illiterate women are treated as the reference group; $\mathbf{X}_6$ is the woman's employment status, a dummy variable coded 1 if she is employed; $\mathbf{X_7}$ and $\mathbf{X_8}$ are dummy variables representing the woman's residence in a city (yes/no) and her residence in a town (yes/no); the reference category is residing in a rural area; $\mathbf{X_9}$ to $\mathbf{X_{13}}$ are five dummy variables representing the woman's region of residence, namely, $\mathbf{X_9}$ residence in the North, $\mathbf{X_{10}}$ residence in the East, $\mathbf{X_{11}}$ residence in the South Central, $\mathbf{X_{12}}$ residence in the Southwest, and $\mathbf{X_{13}}$ residence in the Northwest; residence in the Northeast region is treated as the reference category; and $\mathbf{X_{14}}$ and $\mathbf{X_{15}}$ are two dummy variables reflecting the woman's marital status as follows: $\mathbf{X_{14}}$ indicates if the woman is widowed (yes or no), and $\mathbf{X_{15}}$ if she is divorced (yes or no); currently married is the reference category.

The Poisson regression model is estimated with maximum likelihood procedures. Table 2 reports the results of the above Poisson regression model for Han women, Manchu women and Korean women. All three models converged after three iterations. The overall structure of the models may be appraised with the Likelihood Ratio χ^2 statistic, which tests the null hypothesis ($\mathbf{H_0}$) that all the Poisson coefficients are not significantly different from zero. In all three models the null hypothesis may be rejected, indicating that there is some predictive utility in the three models. This conclusion is reinforced by the significant values of the three Pseudo $\mathbf{R}^2$ statistics.

The decision to use a Poisson regression approach to model CEB for the Han, Manchu and Korean women may be formally and directly appraised with the Poisson Goodness of Fit χ^2 test statistic (bottom of Table 2); it compares the observed empirical distribution with the distribution predicted by the Poisson regression model. The null hypothesis ($\mathbf{H_0}$) is that there is no difference between the observed data and the modeled data, indicating that the Poisson model fits the data. A small χ^2 value, with a probability > 0.05, indicates that one cannot reject the null hypothesis that the observed CEB data are Poisson distributed. In all three models, the values of the Poisson Goodness of Fit χ^2 statistic indicate that using Poisson regression to model the CEB data was appropriate.

The Poisson regression coefficients for the fifteen independent variables will now be examined. Table 2 reports for each independent variable the value of the Poisson coefficient (**b**) and its standard error (**s.e.**). Coefficients that are not significant have been asterisked. The Poisson coefficients indicate the degree of nonlinear association of the independent variable with the dependent variable of CEB, controlling for the effects of the other independent variables.

Looking first at the model for Han women, age is positively associated with CEB. And the four education dummy variables are negatively associated with CEB (the reference variable here is illiterate status). If the woman is employed ($\mathbf{X_6}$), she has fewer children than if she is not employed. Women who live in cities ($\mathbf{X_7}$), or in towns ($\mathbf{X_8}$), have fewer children than women who live in rural areas. Women who live in the North ($\mathbf{X_9}$), or in the South Central ($\mathbf{X_{11}}$), or in the Southwest ($\mathbf{X_{12}}$), or in the Northwest ($\mathbf{X_{13}}$) have more children than women living in the Northeast region (the reference region). The CEB of women living in the East ($\mathbf{X_{10}}$) is not significantly different from the CEB of women living in the Northeast. The CEB of widowed women ($\mathbf{X_{14}}$) is not significantly different from the CEB of married women, but the CEB of divorced women ($\mathbf{X_{15}}$) is significantly less than that of married women. None of the signs of the Poisson coefficients are surprising. They are what one would expect.

The effects of the Poisson coefficients for the independent variables in the other two regression models, those for Manchu women and for Korean women, are quite similar in sign, and in magnitude as those for Han women. However, more of the coefficients in the Manchu and Korean models are not statistically significant compared to the number of insignificant coefficients in the Han model. Five of the fifteen coefficients in the Manchu regression model are not significant (four of the region variables, and the widowed variable). And eleven of the fifteen coefficients in the Korean model are not statistically significant; only the age, college, city residence, and divorced variables are statistically significant.

It was noted earlier in the review of the demographic literature on the statistical modeling of children ever born that many CEB analyses have used linear regression approaches. It was also noted that such a strategy is not appropriate in low fertility populations owing to the heavily skewed distribution of CEB. One thus might ask how similar, or different, would the regression results reported in Table 2 be if linear regression models had been used instead of Poisson regression models.

Table 3 reports ordinary least squares regression results for the same Han, Manchu and Korean populations using the same independent and dependent variables. There are many differences between the OLS regression results shown in Table 3 and the Poisson regression results shown in Table 2. The most important differences have to do with the statistical significance of many of the coefficients. For instance, in the equations for the Han women, and in the equations for the Korean women, more OLS coefficients are statistically significant than are the corresponding Poisson coefficients. In the two Manchu equations, the same five coefficients do not achieve statistical significance.

Among Han women all the OLS coefficients are significant, whereas two of their corresponding Poisson coefficients are not significant. Among the Korean women, six of their fifteen OLS coefficients are not significant, but eleven of their Poisson coefficients are not significant.

Had an OLS model, instead of a Poisson model, been used to predict the number of children ever born among Korean women, incorrect statistical inferences would have been made for the effects of five of the fifteen variables. The results of the OLS model would have allowed the inferences that Korean women who have completed middle school ($\mathbf{X_4}$), and high school ($\mathbf{X_5}$), have fewer children than Korean women who are illiterate. In the Poisson regression these coefficients are not significant. Also, the OLS regression results permit the inferences that employed Korean women ($\mathbf{X_5}$) have fewer children ever born than unemployed Korean women, and women living in towns ($\mathbf{X_8}$) have a lower CEB than women living in rural areas; these are two more erroneous statistical inferences. And, according to the OLS results, it would have been concluded that women living in the South Central region ($\mathbf{X_{11}}$) have more children ever born than women living in the Northeast region, another incorrect inference.

Poisson regression models were estimated for Han, Manchu and Korean women because their mean and variance values for CEB were similar (Table 1). However, Poisson regression models were not estimated for the Hui and Uygur women because their respective variance CEB values were larger than their corresponding mean CEB values (Table 1) indicating the apparent presence for each group of over-dispersion in their CEB distributions.

If there is significant over-dispersion in the distribution of the count (CEB) variable for a population, the estimates from the Poisson regression model will be consistent, but inefficient. "Further the standard errors from the (Poisson regression model) will be biased downward, resulting in spuriously large z-values" (Long, 1997, p. 230), which could lead the investigator to make incorrect statistical inferences about the significance of the variables. This situation is addressed by extending the Poisson regression model by adding "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, which is now considered.

The Negative Binomial Regression Model

It was noted earlier that the Poisson regression model "accounts for observed heterogeneity (i.e., observed differences among sample members) by specifying the (predicted count, $\boldsymbol{\mu}$) as a function of the observed" independent variables (Long & Freese, 2001, p. 243). Often, however, the Poisson regression model does not fit the observed data because of over-dispersion. "That is, the model underestimates the amount of dispersion in the outcome" (Long & Freese, 2001, p. 243). In the negative binomial regression model, variation in $\boldsymbol{\mu}$ "is due both to variation in (the independent variables) among the individuals (in the sample population) and to unobserved heterogeneity introduced by $\boldsymbol{\varepsilon}$" (Long, 1997, p. 231). The term $\boldsymbol{\varepsilon}$ is a "random error that is assumed to be uncorrelated with (the independent variables) ... ($\boldsymbol{\varepsilon}$ may be thought of) "either as the combined effects of unobserved variables that have been omitted from the model or as another source of pure randomness" (Long, 1997, p. 231).

The negative binomial regression model thus adds to the Poisson regression model the error term **ε** according to the following structural equation:

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

It may be shown that the distribution of the observations in the negative binomial regression model is still Poisson. In the negative binomial regression model, the mean structure is the same as in the Poisson regression model, but the distribution about the mean is not the same (Long, 1997, p. 233: Long & Freese, 2001, p. 243). If there is not a statistically significant amount of dispersion in the count outcome data, then the negative binomial regression model will reduce to the Poisson regression model.

One way, therefore, to test for dispersion in the count outcome it to estimate a negative binomial regression model along with a Poisson regression model, and to compare the results of the two models. Like the Poisson regression model, the negative binomial regression model is estimated by maximum likelihood procedures.

As already noted, given a data-set with over-dispersion, if one were to estimate both Poisson and negative binomial regression models, both will have the same mean structure. But the Poisson model will tend to under-estimate the dispersion in the dependent variable. Hence, "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, p. 243; Cameron & Trivedi, 1986, p. 31). Also, in the negative binomial model, compared to the Poisson regression model, there will be an increased probability of both low and high counts.

The left panel of Table 4 contains the results of a negative binomial regression model using the fifteen independent variables to estimate the number of children ever born for ever-married Hui women. For comparison purposes, the middle panel of the table contains the results of a Poisson regression estimating Hui CEB using the same independent variables. And in the right panel are presented the results from an OLS regression.

Comparing the values of the negative binomial regression coefficients (left panel of Table 4) with the values of the Poisson regression coefficients (middle panel), it may be seen that the two sets of coefficients are virtually identical. This suggests that there is not a significant amount of dispersion in the CEB data for the Hui women.

The formal statistical test for appraising the presence of dispersion in the negative binomial distribution is the parameter, alpha (in the Poisson regression model, thus, alpha = 0). (See StataCorp, 2001, volume 2, p. 386-387, 390-391; Long & Freese, 2001, p. 243-245 for more discussion.) At the bottom of Table 4 (left panel) is the value of alpha, and immediately below it, the likelihood-ratio χ^2 test of alpha. The value of alpha is .000, indicating that there is not a statistically significant amount of dispersion in the distribution of CEB for the Hui women. The likelihood ratio χ^2 test of alpha has a value of .000, with a probability of .5.

This χ^2 test is based on a comparison of the value of the final log likelihood from the negative binomial regression model and the corresponding value from the Poisson model. There is no difference in the values, indicating that the CEB data for the Hui women are Poisson distributed. This conclusion is reinforced by the results of the Poisson Goodness of Fit of Fit χ^2 (bottom of the middle panel of the table), which has a probability of 1.0. This means that the Poisson model fits the data; the Poisson goodness of fit χ^2 test indicates that given the Poisson regression model one cannot reject the null hypothesis that the observed data are Poisson distributed.

Before leaving the CEB regressions for the Hui women, the Poisson results will be compared with the OLS regression results. What kinds of inference errors would have been made had the Hui CEB been estimated with a linear regression model? The results of the OLS regression model would have allowed the conclusion that among the Hui women employment status (**X_6**) has a significant negative effect on CEB. Thus it would have been inferred that employed women have fewer children ever born than women who are not employed. This turns out to be an incorrect inference. The Poisson regression model results indicate no statistical relationship between employment status and CEB.

Similar errors of inference would have been regarding the effects on CEB of the woman's location in the East region (**X_{10}**), the South Central region (**X_{11}**), and the Northwest region (**X_{13}**). For all three of these regional location variables the

OLS regression results indicate that the effects are significant, but the Poisson regression results show they are not. The Poisson regression model is the more statistically appropriate approach for modeling CEB among the Hui women.

Finally, the estimation of children ever born among the Uygur women may be considered. For Uygur women the variance of their CEB is more than twice the magnitude of the mean of their CEB, values of 6.99 and 3.16, respectively. Table 5 presents in the left panel the results of a negative binomial regression model estimating Uygur CEB, along with the results of a Poisson regression model in the center panel, and the results of an OLS regression model in the right panel. The first question is whether there is enough over-dispersion in Uygur CEB to justify the use of a negative binomial regression model.

The first indication that the negative binomial model is appropriate is the fact that the coefficients from the model are very different from the corresponding coefficients from the Poisson model. A second and more formal indication is that alpha, the over-dispersion parameter (bottom of the table, left panel), has a value of .113, with a probability of .005. And the likelihood-ratio χ^2 test of alpha has a high value of 776.0, with a probability of .000, indicating that the probability that one would observe these data if the process was Poisson, i.e., if alpha = 0, is virtually zero. The Uygur data are clearly not Poisson. A final and related indication is that the Poisson Goodness of Fit χ^2 test statistic performed on a Poisson regression of the Uygur CEB data (bottom of the middle panel of the table) has a probability of .000. This means that the Poisson model does not fit the data; according to the Poisson goodness of fit χ^2 test, the null hypothesis that the observed data are Poisson distributed must be rejected.

The negative binomial and Poisson coefficients (Table 5) may now be compared. First, the signs of the effects of the independent variables on CEB are all the same. Also, the six predictors that are not statistically significant in one model are not significant in the other model. However, for thirteen of the independent variables, the standard errors in the Poisson model are smaller than those in the negative binomial model (the standard errors for the age variable ($\mathbf{X_1}$) are the same in both models). This means that for thirteen of the fourteen independent variables, in the Poisson model the z-values will be spuriously high and the p-values spuriously low. Although there would have been no errors of statistical inference had these Poisson regression results, rather than the negative binomial regression results, been used to predict Uygur CEB, the potential for error is much greater using the Poisson results. For all the above reasons, the negative binomial model is the preferred regression model for predicting children ever born among Uygur women.

Finally, the results of the negative binomial regression predicting Uygur CEB may be compared with the OLS results (left and right panels of Table 5). Would any inference errors been committed had the OLS results been used? The major error that would have occurred is with regard to the effect on CEB of employment status. The results of the OLS regression model indicate that among Uygur women employment status ($\mathbf{X_6}$) has a statistically significant negative effect on CEB. Thus one would have inferred that employed Uygur women have fewer children ever born than Uygur women who are not employed, controlling for the effects of the other independent variables. This turns out to be an incorrect inference. The negative binomial regression results show no statistical relationship between employment status and CEB. Some of the implications of the research reported in this paper will now be addressed.

Conclusion

This article considered distributions of CEB data for five sub-groups of Chinese women. It was shown that they were not normal (Gaussian), but, rather, heavily skewed with long right tails. Such distributions are characteristic of low-fertility populations. Given such distributions, a linear regression model is inappropriate for the statistical modeling of children ever born. Fifteen socioeconomic and locational variables drawn from the 1990 Census of China were then used as independent variables to model CEB for the Han and minority group women.

For the Han and Manchu and Korean women, both Poisson regression and ordinary least squares (OLS) regression models were estimated. And for the Hui and Uygur women, these same two approaches along with negative binomial regression were used. It was shown that in almost all instances there would have been major errors of statistical inference had the interpretations been based only on the results of linear regression models.

The literature on the statistical modeling of CEB data indicates that in many instances, linear regression models have been used. The decision to use a linear model, however, is only appropriate if the average CEB value is high. When the mean of a count outcome is high, say, at least above 4 or 5, but certainly around 8 or 9, then the distribution of the outcome will often tend to be approximately normal. However, few populations these days, except mainly those in sub-Saharan Africa, have fertility this high. It would appear thus that the use of a linear model for modeling a fertility variable such as children ever born is becoming more and more inappropriate. And in low fertility populations, such as China, using a linear model would clearly be inappropriate statistically.

References

Bean, F. D., Tienda, M. (1987). *The Hispanic population of the United States*. New York: Russell Sage Foundation.

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.

Entwisle, B., & Mason, W. M. (1985). Multilevel effects of socioeconomic development and family planning programs on children ever born. *American Journal of Sociology*, *91*, 616-649.

Janssen, S. G., & Hauser, R. M. (1981). Religion, socialization, and fertility. *Demography*, *18*, 511-528.

Johnson, N. E. (1979). Minority-group status and the fertility of Black Americans, 1970: A new look. *American Journal of Sociology*, *84*, 1386-1400.

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.

Ritchey, P. N. (1975). The effect of minority group status on fertility: A re-examination of concepts. *Population Studies*, *29*, 249-257.

StataCorp. (2001). *Stata statistical software: release 7.0.* College Station, Texas: Stata Corporation.

Appendix: Tables and Figures

Table 1: Descriptive Data for Children Ever Born: Ever-Married Han, Manchu, Korean, Uygur, and Hui Women, Ages 15-49, China, 1990

Group	Mean	Standard Dev.	Variance	No. of Cases
Han	2.1326	1.4202	2.0170	216,312
Manchu	1.8047	1.1745	1.3795	20,210
Korean	1.7959	1.0478	1.0978	3,837
Uygur	3.1577	2.6443	6.9921	14,553
Hui	2.3289	1.7662	3.1194	17,976

Source of Data: 1% Sample of the 1990 Census of China. The sample of Han women is a 1/10 sample of the 1% sample.

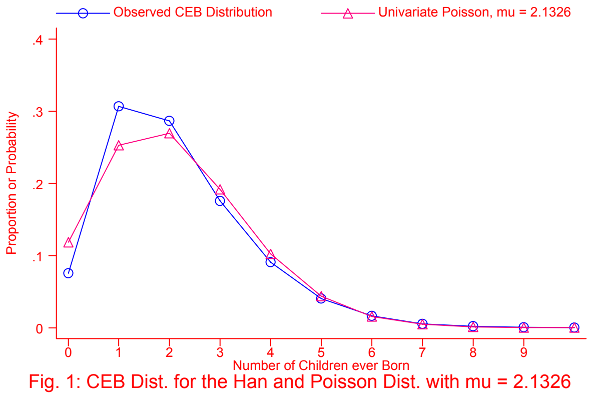


Fig. 1: CEB Dist. for the Han and Poisson Dist. with mu = 2.1326

Table 2: Poisson Regression Models Predicting Number of Children Ever-born for Ever-Married Han, Manchu and Korean Women, Ages 15-49, China, 1990

	Han		Manchu		Korean	
Sample Size	216,312		20,210		3,837	
Independent Variable	**b**	**s.e.**	**b**	**s.e.**	**b**	**s.e.**
X_1 Age	.050	.000	.055	.001	.052	.002
X_2 Elem. Sch	-.092	.004	-.076	.019	.005	.085*
X_3 Middle Sch	-.239	.005	-.189	.020	-.058	.085*
X_4 High School	-.353	.007	-.248	.025	-.117	.089*
X_5 College	-.583	.020	-.466	.054	-.301	.123
X_6 Employ Status	-.063	.005	-.095	.012	-.013	.031*
X_7 City Residence	-.398	.006	-.335	.022	-.234	.038
X_8 Town Residence	-.096	.004	-.055	.013	-.029	.028*
X_9 North Region	.018	.007	.050	.013	-.099	.086*
X_{10} East Region	-.003	.006*	-.055	.069*	-.045	.181*
X_{11} S. Central Reg.	.120	.006	.014	.054*	.152	.134*
X_{12} SW Region	.034	.007	.060	.097*	-.182	.290*
X_{13} NW Region	.089	.008	-.029	.071*	-.032	.236*
X_{14} Widowed	-.022	.012*	-.040	.048*	-.023	.071*
X_{15} Divorced	-.285	.028	-.261	.081	-.341	.129
Constant	-.809	.010	-1.057	.034	-1.145	.111
Pseudo R^2	.145	.000	.136	.000	.112	.000
Likelihood Ratio χ^2	106740.4	0.00	8456.5	0.00	1283.5	0.00
Poisson Goodness of Fit χ^2	106486.4	1.00	7527.8	1.00	1322.9	1.00

*Coefficient not significant at p <.05.

Table 3: Ordinary Least Squares Regression Models Predicting Number of Children Ever-born for Ever-Married Han, Manchu and Korean Women, Ages 15-49, China, 1990

	Han		Manchu		Korean	
Sample Size	216,312		20,210		3,837	
Independent Variable	**b**	**s.e.**	**b**	**s.e.**	**b**	**s.e.**
X_1 Age	.110	.000	.106	.001	.095	.002
X_2 Elem Sch	-.311	.006	-.277	.023	-.028	.097*
X_3 Middle Sch	-.569	.007	-.478	.024	-.220	.096
X_4 High Sch	-.727	.009	-.591	.027	-.333	.098
X_5 College	-.975	.021	-.858	.048	-.521	.117
X_6 Employ Stat	-.192	.007	-.224	.012	-.062	.030
X_7 City Resid	-.762	.007	-.557	.020	-.383	.034
X_8 Town Resid	-.223	.006	-.110	.013	-.067	.027
X_9 North Region	.023	.009	.091	.013	-.144	.077*
X_{10} East Region	-.019	.008	-.119	.063*	-.100	.175*
X_{11} S. Cent Reg	.262	.008	.025	.052*	.293	.143
X_{12} SW Region	.082	.009	.106	100*	-.370	.234*
X_{13} NW Region	.189	.011	.013	.067*	-.001	.233*
X_{14} Widowed	.096	.021	.074	.062*	-.060	.081*
X_{15} Divorced	-.482	.032	-.354	.067	-.492	.095
Constant	-.951	.014	-1.012	.036	-1.066	.115
R^2 (adj.)	.531	.000	.577	.000	.559	.000
F-test	16293.0	.000	1839.1	.000	1283.5	.000

*Coefficient not significant at p <.05.

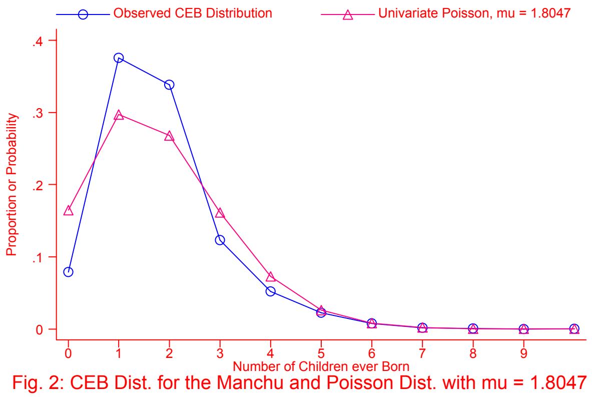
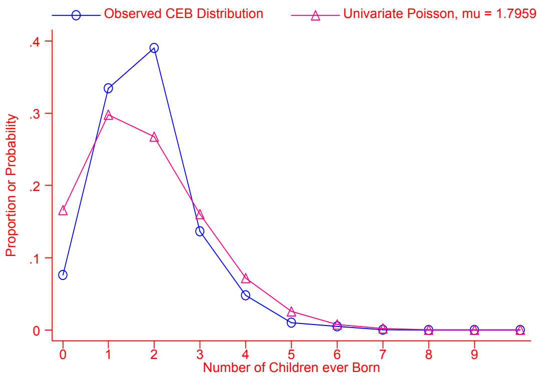


Fig. 2: CEB Dist. for the Manchu and Poisson Dist. with mu = 1.8047



g. 3: CEB Dist. for the Koreans and Poisson Dist. with mu = 1.79

Table 4:Negative Binomial Regression Model (NBR), Poisson Regression Model (PR), and Ordinary Least Squares Regression Model (OLS) Predicting Number of Children Ever-born for 17,976 Ever-Married Hui Women, Ages 15-49, China, 1990

	NBR Model		PR Model		OLS Model	
Independent Variable	**b**	**s.e.**	**b**	**s.e.**	**b**	**s.e.**
X_1 Age	.054	.001	.054	.001	.133	.001
X_2 Elem Sch	-.108	.014	-.108	.014	-.412	.026
X_3 Middle Sch	-.234	.017	-.234	.017	-.559	.029
X_4 High Sch	-.341	.024	-.341	.024	-.674	.037
X_5 College	-.575	.062	-.575	.062	-.998	.079
X_6 Employ	-.013	.017*	-.013	.017*	-.081	.031
X_7 City Res.	-.354	.017	-.354	.017	-.828	.027
X_8 Town Res.	-.072	.017	-.072	.017	-.225	.029
X_9 North Reg.	-.024	.026*	-.024	.026*	-.061	.040*
X_{10} East Reg.	.047	.029*	.047	.029*	.117	.045
X_{11} S. Cent Reg.	.048	.029*	.048	.029*	.109	.045
X_{12} SW Reg.	-.008	.030*	-.008	.030*	-.039	.049*
X_{13} NW Reg.	.287	.026	.286	.026	.689	.042
X_{14} Widowed	-.029	.037*	-.029	.037*	.084	.083*
X_{15} Divorced	-.490	.058	-.490	.058	-.913	.081
Constant	-.959	.039	-.959	.039	-1.742	.066
Pseudo R^2 / R^2 (adj.)	.181	.000	.189	.000	.550	.000
Likelihood Ratio χ^2 or F-test	12072.2	.000	12763.3	.000	1462.7	.000
Alpha	000	.000				
L-Ratio χ^2 test of alpha			.000	.500		
Poisson Goodness of Fit χ^2			11049.0	1.000		

*Coefficient not significant at p <.05.

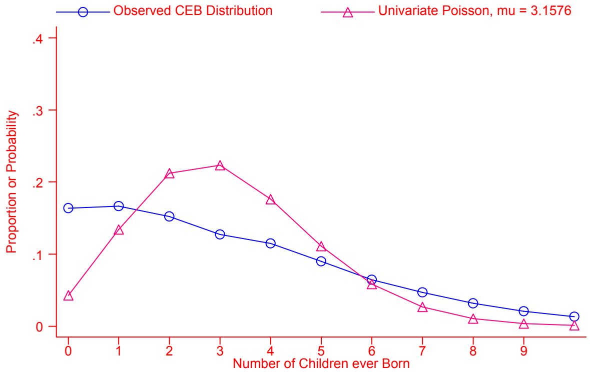


Fig. 4: CEB Dist. for the Uygur & Poisson Dist. with mu = 3.1576

Table 5: Negative Binomial Regression Model(NBR), Poisson Regression Model (PR), and Ordinary Least Squares Regression Model (OLS) Predicting Number of Children Ever-born for14,553 Ever-Married Uygur Women, Ages 15-49, China, 1990

	NBR Model		PR Model		OLS Model	
Independent Variable	**b**	**s.e.**	**b**	**s.e.**	**b**	**s.e.**
X_1 Age	.060	.001	.057	.001	.184	.002
X_2 Elem Sch	.059	.014	.071	.012	.213	.045
X_3 Middle Sch	.071	.018	.085	.015	.202	.055
X_4 High Sch	-.074	.026	-.060	.022	-.196	.075
X_5 College	-.259	.070	-.234	.061	-.608	.183
X_6 Employ	-.019	.016*	-.025	.013*	-.121	.048
X_7 City Res.	-.247	.021	-.248	.018	-.817	.061
X_8 Town Res.	-.052	.019	-.060	.016	-.232	.056
X_9 N Region	-.076	.949*	-.103	.867*	.289	2.326*
X_{10} E Region	.147	.899*	.117	.817*	.798	2.253*
X_{11} S. Cent Reg.	.218	.830*	.195	.750*	1.091	2.113*
X_{12} SW Region	variable not included					
X_{13} NW Region	.649	.783*	.629	.707*	2.116	2.014*
X_{14} Widowed	-.202	.035	-.183	.028	-.608	.117
X_{15} Divorced	-.800	.032	-.802	.029	-1.426	.066
Constant	-1.480	.784*	-1.348	.708*	-4.518	2.017
Pseudo R^2 / R^2 (adj.)	.190	.000	.123	.000	.421	.000
Likelihood Ratio χ^2 or F-test	8042.6	.000	13645.7	.000	755.2	.000
Alpha	.113	.005				
L-Ratio χ^2 test of alpha	776.0	.000				
Poisson Goodness of Fit χ^2			21413.4	.000		

*Coefficient not significant at p <.05.

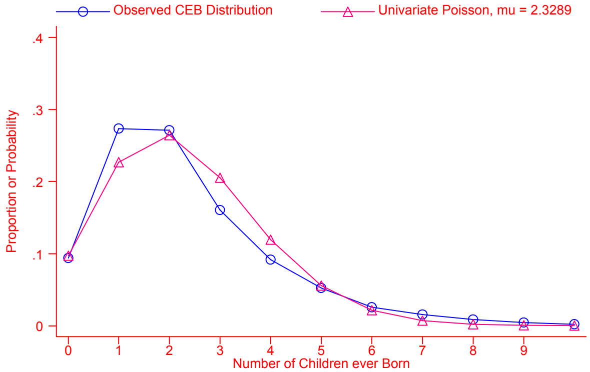


Fig. 5: CEB Dist. for the Hui & Poisson Dist. with mu = 2.3289