Running head: MODELS FOR COUNT DATA

A Model Comparison for Count Data with a Positively Skewed Distribution with an Application to the Number of University Mathematics Courses Completed

Pey-Yan Liou, M.A.
Department of Educational Psychology
University of Minnesota
56 East River Rd
Minneapolis, MN 55455
Email: lioux005@umn.edu
Phone: 612-626-7998

Paper presented at the Annual Meeting of the American Educational Research Association

San Diego, April 16, 2009

Abstract

The current study examines three regression models: OLS (ordinary least square) linear regression, Poisson regression, and negative binomial regression for analyzing count data. Simulation results show that the OLS regression model performed better than the others, since it did not produce more false statistically significant relationships than expected by chance at alpha levels 0.05 and 0.01. The Poisson regression model produced fewer Type I errors than expected at alpha levels 0.05 and 0.01. The negative binomial regression model produced more Type I errors at both 0.05 and 0.01 alpha levels, but it did not produce more incorrect statistically significant relationships than expected by chance as the sample sizes increased.

A Model Comparison for Count Data with a Positively Skewed Distribution with an Application to the Number of University Mathematics Courses Completed

Introduction

Student mathematics achievement has always been an important issue in education. Several reports (e.g., Kuenzi, Matthews, & Mangan, 2006; United States National Academies [USNA], 2007) have stressed that the well being of America and America's competitive edge depend largely on science, technology, engineering and mathematics (STEM) education. USNA (2007) examined the K-12 STEM curriculum and concluded that the key to an innovative and technological society rests in STEM fields. Additionally, growing pressure from globalization has solidified the idea that to maintain a nation's advantage depends not only on how well people educate their children but especially on how well people educate them in mathematics and science (Glenn, 2000).

Data from the 2007 Trends in International Mathematics and Science Study (TIMSS) rank American fourth grade students in 11th place out of 36 nations and eighth grade students in 9th place out of 49 nations on their average mathematics score (Mullis, Martin, & Foy, 2008). In science, the situation is very similar with TIMSS ranking American fourth grade students in 8th place out of 36 nations and eighth grade students in 11th place out of 49 nations on their average science score (Martin, Mullis, & Foy, 2008). In addition, in the 2006 Program for International Student Assessment (PISA), fifteen-year-old American students' mathematics scores ranked 32th out of 52 nations (Baldi, Jin, Skemer, Green, & Herget, 2007). This shows that if the United States wants to maintain its status as a world economic and technological leader and continue to compete with high student achievement Asian and the Organization for Economic Cooperation and Development (OECD) countries, it is imperative that STEM educators, researchers, and the government act to increase the international achievement ranking. However, this trend also reflected how students chose their majors in

higher education. The report, “Science, Technology, Engineering and Mathematics (STEM) Education Issues and Legislative Options” (Kuenzi et al., 2006) mentioned that 16.7 % of American bachelor degrees were conferred in STEM fields during 2002-2003. It also stated that the rate of Americans who have STEM to non-STEM degrees was one of the lowest rates around the world in 1997.

In order to promote STEM education and recruit more students to choose STEM majors, one solution is to increase the number of students taking mathematics courses (Kuenzi, at el., 2006). However, in reality, it seems that many American postsecondary students do not have a solid foundation in mathematics. Research (Parsad & Lewis, 2003) shows that 22% of all U.S. entering freshmen at degree-granting institutions take at least one remedial mathematics class. The definition of remedial mathematics courses was defined as mathematics courses for college-level students lacking those skills necessary to perform college-level work at the level required by the institution (Parsad & Lewis, 2003).

Therefore, the number of mathematics courses completed and the factors that influence course completion should be studied because of the role mathematics plays not only in STEM, but also in the nation’s economics and well-being.

Much educational research has studied the relationship between various factors and student mathematics performance. The most commonly-used statistical models belong to the category of the “general linear model” (Kutner, Nachtsheim, Neter, & Li, 2005), such as simple linear regression analysis, multiple linear regression analysis, and analysis of variance (ANOVA). For instance, Stedman (1997) used simple linear regression to show the relationship between the correct numbers of items students answered in the Second International Mathematics Study exam and the percentage of materials covered in the average U.S. mathematics curriculum. House (2002) performed multiple linear regressions to assess the relationship between various instructional practices and mathematics achievement.

Chuansheng & Stevenson (1995) used ANOVAs to study the difference of Asian-American, Caucasian-American, and East Asian high school students' mathematics achievement as well as students' beliefs about efforts. Moreover, they used multiple regressions to identity the relationship of mathematics scores and variables that account for cultural differences among students. Yee and Eccles (1988) performed 2 (child sex) $\times$ 3 (mathematics ability level) ANOVAs on students' mathematics scores and parents' beliefs and offered explanations.

Path analysis and hierarchical linear modeling are other models used in studying mathematics performance. Parsons, Adler, and Kaczala (1982) used path analysis to draw paths among variables which are related to student mathematics performance. Lee, Croninger, and Smith (1997) used two-level hierarchal linear modeling to differentiate the variability from high schools instead of just analyzing students' information as a single unit.

However, in the examples mentioned above, the outcome variables are continuous (i.e., student's mathematics score). Few articles analyze discrete count variables (i.e., the number of mathematics courses taken). How to use appropriate statistical models to analyze count data is a very substantial research topic in quantitative methodologies. General linear models may not be suitable for analyzing discrete count data.

Yet, despite its violation of fundamental general linear models assumptions, linear regression analysis is still used in educational research for count data. Ayalon and Yogev (1997) used hierarchical linear modeling to estimate the relationship between students' characteristics and course-taking in science and humanity areas separately, and the effects of school characteristics on these relationships. The dependent variable at the student level was the number of course-taking units in either science or humanities. However, the authors did not mention the distribution of the outcome variable, which was the number of course-taking units, and what kind of regression model they used to estimate parameters. According to the equation on pages 344-345 in their article, the models were

The level 1 model:

$$N(units)_{ij} = \beta_{0j} + \beta_{1j}(gender)_{ij} + \beta_{2j}(ethnicity)_{ij} + \beta_{3j}(ability)_{ij} + r_{ij}$$

The level 2 models:

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(sector)_j + \gamma_{12}(size)_j + \gamma_{13}(\%male)_j + v_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}(sector)_j + \gamma_{22}(size)_j + v_{2j}$$

$$\beta_{3j} = \gamma_{30} + \gamma_{31}(sector)_j + \gamma_{32}(size)_j + \gamma_{33}(meanability)_j + v_{3j}$$

It is likely that Ayalon and Yogev assumed that the distribution of the outcome is normal and continuous since they did not mention anything about the distribution of the outcome variable, the number of courses taken, which is likely to be positively-skewed distributed count data. Due to this misuse of the appropriate regression model, the conclusion drawn from their statistical analysis, which influences their estimated standard errors and inferential statistics, may not be valid.

Although some research (Davenport, Davison, Kuang, Ding, Kim, & Kwak, 1998; Davenport, Davison, Wu, Kim, Kuang, Kwak, & Chan, 2004) stated that the more important thing about students' mathematics achievement is course difficulty instead of the number of courses completed, it is still worth studying statistical issues surrounding count data with positively-skewed distributions instead of being normally distributed, because of the issue of statistical conclusion validity which means the "validity of conclusions, or inferences, based on statistical tests of significant" (Pedhazur & Schmelkin, 1991, p.224). The statistical evidence of this research can further assist researchers who study count data related to the number of mathematics courses completed or similar topics.

Literature Review

The regression model is a statistical method which shows the relationship between two or more quantitative variables. The intent is to learn if values of a dependent variable can be predicted from other independent variable(s) in a regression model. In addition, inferential statistics from the model parameters provide a way to evaluate the magnitude of each independent variable which can account for variation in the dependent variable when controlling for other independent variables. The linear regression model utilizing ordinary least squares (OLS) estimation is the most commonly used traditional regression model in the educational research field.

However, count data with a positively-skewed distribution may not fit well in the OLS linear regression model. There are four reasons. First, the OLS linear regression model produces negative values, but count data are always larger than or equal to zero. In other words, OLS linear regression does not account for data being truncated at zero; thus, it could predict negative values which are meaningless (King, 1988; Sturman, 1999). Second, one of the assumptions for validating statistical tests from OLS linear regression is the normality of residuals. Count data with a positively-skewed distribution are unlikely to satisfy this assumption. Third, the validity of hypothesis tests in the OLS linear regression model depends on assumptions about the homogeneity of variance of residuals that are unlikely to be met in count data (Gardner, Mulvey, & Shaw, 1995). Fourth, OLS linear regression is mainly for continuous dependent variables, not discrete variables, like count data. Due to the reasons mentioned above, using OLS regression to analyze count data may lead to conclusions that do not make sense for the data, such as impossible mean predicted values, and incorrect standard errors for significance tests and p-values.

Given these limitations of OLS linear regression for count data with a positively-skewed distribution, some research has suggested using rescaled categories (Gardner et al., 1995) or

transformations (Kutner et al., 2005), since they are commonly applied in education research. Moreover, these methods change the characteristics of count data to more closely match the assumptions of the traditional statistical methods. For instance, count data can be rescaled into a dichotomous variable, and analyzed using logistic regression or a similar technique suitable for binary variables. Another option is to rescale count data to a set of ordered categories, and then use these rescaled categories variables in an analysis, such as ANOVA. However, reducing counts to categories, such as changing a four rating-scale variable into a dichotomous variable, would squander some information, and may lead to diluted statistical power, defeating the purpose of the analysis (Gardner et al., 1995). As for data transformations, one of the disadvantages is that they may obscure the fundamental interconnections between variables. The other is that idiosyncratic transformations make it difficult to compare results across studies.

Another way to analyze this type of count data with its nonnormality of residuals is to use more appropriate statistical methods. One option may be nonparametric (NPAR) statistical methods because nonparametric statistical methods have no distribution assumption (Blum & Fattu, 1954; Harwell, 1988). However, several nonparametric statistical methods rank the data, thus losing some information from the original values. Therefore, if the distribution of count data is known, parametric statistics models should be utilized. The nonlinear regression models are more appropriate statistical methods for the known distribution of count data.

Nonlinear regression models are appropriate for count data because they use probability distributions for the dispersion of the dependent variable scores around the expected value for dependent variables which take on only nonnegative integer values (Kutner et al., 2005).

Therefore, the alternative to rescaled categories, transformations, or NPAN is to adopt an appropriate nonlinear model. The “generalized linear models” (GLMs) are solutions for

modeling count data. GLMs do not assume a normally distributed response variable. The random component in GLMs is the dependent variable, which assumes a certain distribution. For example, the Poisson regression model and the negative binomial regression model could be used to model count data in different conditions.

The Poisson regression model is utilized for count outcome, with large-count outcomes being rare events (Kutner et al., 2005). Moreover, the Poisson regression model is particularly attractive for modeling count data because the model has been extended into a regression framework, it has a simple structure, and it can be easily estimated (Lee, 1986). However, this simplicity is the result of some limiting assumptions: the variance should be equal to the mean of the response count data. Violations of this assumption may have substantial affects on the reliability and efficiency of the model coefficients (Sturman, 1999).

In reality, the mean and the variance of a dependent variable in most educational data are not the same, such as the number of university mathematics courses completed. Instead, the variance of the model often exceeds the value of the mean, a phenomenon called overdispersion (Hilbe, 2007). Moreover, characteristics of count data may yield further violations of assumptions, which may produce flaws in the Possion regression model. For instance, mathematics achievement research may suggest that the number of university mathematics courses completed is a function of several factors, such as high school mathematics scores, the number of high school mathematics courses completed, ACT/SAT mathematics scores, and so on. It suggests that these individual characteristics cause or at least correlate with the number of mathematics courses completed. Therefore, the negative binomial regression may substitute for this situation because the negative binomial regression has an extra parameter which counts for the overdispersion (Hilbe, 2007).

The following subsections in the Literature Review consist of reviewing three regression models: the OLS linear regression model, the Poisson regression model, and the negative

binomial regression model. Further, the use of the Poisson regression model and the negative binomial regression model in academic research will be reviewed. In addition, simulation studies comparing the OLS linear regression model, the Poisson regression model, and the binomial regression model in other disciplines will be discussed. Finally, the research question of this study will be posed.

Ordinary Least Squares Linear Regression

The formulation of linear regression models can be represented in the form of the general linear regression model, which directly follows the derivation from Kutner et al., (2005),

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1} + \varepsilon_i \quad (1)$$

where:

- $\beta_0, \beta_1, ..., \beta_{p-1}$ are parameters
- $X_{i1}, ..., X_{i,p-1}$ are known constants
- ε_i are independent, and follows $N(0, \sigma^2)$
- $i = 1, ..., n$

Since $E\{\varepsilon_i\} = 0$, the response function for regression model (2.1) is:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{p-1} X_{p-1} \quad (2)$$

Least squares is the most commonly-used estimation in traditional linear regression because of its easy computation and the best linear unbiased estimator (BLUE) under the Gauss-Markov assumption (Puntanen & Styan, 1989). Therefore, it is commonly-used in estimation for linear regression models in educational research, and parameters estimated from the least squares estimation are unbiased and have minimum variance among all unbiased linear estimators (Kutner et al., 2005).

The least squares criterion is generalized as follows for the general linear regression model (1):

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - ... - \beta_{p-1} X_{i,p-1})^2 \quad (3)$$

where the least squares estimators are those as a set $(\beta_0, \beta_1, ..., \beta_{p-1})$ minimize Q.

The OLS regression model gives a generally satisfactory approximation for most regression applications. However, because of the positively-skewed distribution of count data, it contradicts the assumptions of commonly employed statistical methods such as OLS (Sturman, 1999). This is "because for count data, the absolute values of the residuals almost always correlate positively with the predictors, the estimated standard errors of the regression coefficients are smaller than their true value. Thus, the t-values associated with the regression coefficients are likely to be inflated (Sturman, 1999, p. 418)." In other words, tests from OLS estimation are likely to be inefficient, and estimates of standard errors inconsistent for count data.

Poisson Regression

Poisson regression is a nonlinear statistical model, and it is the best known model for modeling count data with a Poisson distribution. GLMs have two primary features. First, for some dependent variables μ_i, the probability distribution of y_i given μ_i is a member of the exponential family. For the Poisson regression model, this distribution is the Poisson distribution. Second, there is a "link function" which is a transformation $g(\cdot)$ that linearizes the expected value of y_i. That is, $g(\mu_i) = \sum_j \beta_j x_{ij}$, where $\sum_j \beta_j x_{ij}$ is a linear combination of the predictors.

As Kutner et al. (2005) stated, the Poisson regression model can be expressed as follows:

$$\mu_i = \mu(X_i, \beta) = \exp(X_i' \beta) \quad (4)$$

In Kutner et al. (2005), $X_i^{'}\beta$ is equivalent to the expression of $\mu_i = \sum_j \beta_j x_{ij}$ in Gardner et al. (1995). μ_i are the dependent mean for the i th case, and they are assumed to be a function of the set of independent variables X_i. In other words, $\mu(X_i, \beta)$ is the value of the predictor variables for case i from the function that relates the mean dependent μ_i to X_i. β are the values of the regression coefficients.

The explanation for the formula (4) is that a one-unit change in the predictor variable X_i multiplies the expected values by a factor of $\exp(\beta_j)$, and a one-unit decrease divides the expected incidents by the same amount (Gardner et al., 1995). In other words, "Poisson models are typically used to either summarize predicted counts based on a set of explanatory predictors, or are used for interpretation of exponentiated estimated slopes, indicating the expected change or difference in the incidence rate ratio of the outcome based on changes in one or more explanatory predictors" (Hilbe, 2007, p.43).

The Poisson probability density function below directly follows the derivation DeGroot & Schervish (2002),

$$f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{for} \quad x = 0,1,2,... \qquad (5)$$

$$= 0 \text{ otherwise.}$$

where:

- x is a random variable with a discrete distribution, and it is supposed to be a nonnegative integer.
- λ is a mean under the probability function of X following the Poisson probability function.

However, serious issues in using Poisson regression when modeling count data arise. The Poisson regression model uses a one-parameter model to describe the distribution of the dependent variable because it assumes that the variance is a function of the mean. This may

be too strict for most data, particularly in designs where observations may not be drawn in strictly independent trials, such as spatial or time autocorrelation (Hilbe, 2007). Furthermore, "overdispersion" frequently happens when Poisson regression is used for modeling count data. The definition of the overdispersion is that the variance of the model exceeds the value of the mean (Hilbe, 2007). Overdispersion is caused by positive correlation between responses or by an excess variation between response probabilities or counts. According to Hilbe (2007), overdispersion occurs when:

> (a) the model omits important explanatory predictors;
>
> (b) the data include outliers;
>
> (c) the model fails to include a sufficient number of interaction terms;
>
> (d) a predictor needs to be transformed to another scale; or when
>
> (e) the assumed linear relationship between the response and the link function and predictors is mistaken, i.e., the link is misspecified (p.52).

In contrast to overdispersion, if the response variance is smaller than the mean, it is called "underdispersion." And if the response variance is equal to the mean, it is called "equidispersion" (Hoffman, 2004). Some researchers (Gardner et al., 1995; Hoffman, 2004) stated that if the response variance is not equal to the mean, the estimates in the Poisson regression model are inefficient. Moreover, the Poisson regression model may produce biased estimates of its variance terms and lead to inappropriate inferences about the regression, since overdispersion may cause standard errors of the estimates to be underestimated (Hilbe, 2007). Sturman's (1999) simulation study showed that when count response data were in the overdispersion condition, the estimates of standard errors from the Poisson regression model seemed to be less than their true value, which leads to inflated t coefficient and Type I errors. If overdispersion happens, its consequences for parameter estimates in the Poisson regression models are like the problem of heteroscedasticity in linear models (Gardner et al., 1995;

Sturman, 1999).

Obviously, when the response count variables do not have equal mean and variance, modeling these kinds of positively-skewed variables has limitations as described above. Unless extremely restrictive assumptions are met, the Poisson model produces incorrect estimates of its variance terms and misleading inferences about the regression. Although regression parameters are consistently estimated, standard errors are biased downwards leading to the rejection of too many false null hypotheses (Caudill & Mixon, 1995). Therefore, it is important to consider alternative regression models.

Negative Binomial Regression

The negative binomial regression model is more flexible than the Poisson model and is frequently used to study count data with overdispersion (Hilbe, 2007; Hoffman, 2004). In fact, the negative binomial regression model is in many ways equivalent to the Poisson regression model because the negative binomial model could be viewed as a Poisson-gamma mixture model (Hilbe, 2007). However, the difference is that the negative binomial regression model has a free dispersion parameter. In other words, the Poisson regression model can be considered as a negative binomial regression model with an ancillary or heterogeneity parameter value of zero (Hilbe, 2007). In the negative binomial regression model, a random term reflecting unexplained between-subject differences is included (Gardner et al., 1995), that is, the negative binomial regression adds an overdispersion parameter to estimate the possible deviation of the variance from the expected value under Poisson regression. Therefore, using the negative binomial regression to model count data with a Poisson distribution has the consequence of generating more conservative estimates of standard errors and may modify parameter estimates (Hilbe, 2007).

The negative binomial probability density function below directly follows the derivation from Hilbe (2007),

$$\frac{\Gamma(y+v)}{\Gamma(y+1)\Gamma(v)}\left(\frac{1}{1+\lambda/v}\right)^{v}\left(1-\frac{1}{1+\lambda/v}\right)^{y} \quad (6)$$

where:

- Γ is the gamma function.
- λ is the mean of the negative binomial distribution.
- v is the dispersion parameter.
- y is the dependent variable.

The Use of Poisson Regression and Negative Binomial Regression in Other Disciplines and Educational Research

The Poisson regression model and the negative binomial regression model have been introduced to many academic disciplines where they have been utilized to analyze count data with positive skew. For instance, in the field of history, researchers have examined how to use these more appropriate regression models in analyzing factors that influenced number of white mob violent acts against African Americans in the American South (Beck & Tolnay, 1995). In the field of politics, researchers (King, 1989) have demonstrated the power of these regression models for observed international relations event count data. In the field of economics, researchers have studied the relationship between the number of patents applied for and received by companies, as well as research and development expenditures (Hausman, Hall, & Griliches, 1984). Poisson regression and the negative binomial regression have been used in the field of psychology as well where researchers studied the number of violent incidents happening in a community (Gardner, et al., 1995). And in business, researchers have used Poisson regression and the negative binomial regression to study factors that affect absenteeism (Sturman, 1999).

Compared with other disciplines, however, Poisson regression and the negative binomial regression seem to be rarely used in educational research even when outcome variables are

count data with a positive skewed distribution. This may be because it seems little attention is paid to examining count data. None of the educational research articles found by the author used Poisson regression, and only two educational research articles found by the author used the negative binomial regression model: Goyette's study (1999) and Cole's study (2006). Yet, even while it is a positive sign to see the negative binomial regression model being used, the statistical analyses approaches may be challenged.

In Goyette's study, the negative binomial regression was utilized to model the number of applications of Asian American and White high school students for college. One of Goyette's reasons for using the binomial regression model in this study was that the negative binomial regression model is appropriate for the specification of the count dependent variables which were defined as a number of repeatable events within a certain, fixed interval. The second reason, as Goyette described, "I favor the negative binomial model over a simple Poisson model because the Poisson model is based on the assumption that the mean of college application equals its variance (p.26)." It is valuable that Goyette mentioned the difference between the Poisson model and the negative binomial model, and the reason why Goyette used the negative binomial regression model. However, Goyette did not provide the values of the mean and the variance of the number of the student applications in this study clearly, For instance, whether the variance is larger than the mean, or the mean is larger than the variance. Therefore, the results from the way Goyette used the negative binomial regression may not be valid.

On the other hand, Cole's study (2006) utilized negative binomial regression to analyze the number of ethnocentric courses provided in tribal, black, and mainstream colleges and universities. This article did an excellent job to explain the factors, such as the total number of courses, undergraduate enrollment, year, and school locations, which affect the number of ethnocentric courses offered, as well as interpreting the coefficient of variables from the

negative binomial regression model. However, the researcher did not mention any reason for using the negative binomial regression model; for instance, there are no descriptive statistics for the number of ethnocentric courses offered, such as the mean, variance, skewness, and kurtosis. Without this basic statistical information, it is hard to decide whether the negative binomial regression works better than other regressions in this situation.

Although these two articles did not provide enough statistical information for readers to decide if the negative binomial regression was more appropriate, the way they utilized the negative binomial regression for analyzing count data provides a valuable example for educational researchers who analyze count data. Therefore, there is a need for researchers to examine when the OLS regression model, the Poisson regression model, and the negative binomial regression model could be better used in different conditions, such as with different means, variances, and sample sizes.

The Simulation Studies of OLS Regression, Poisson Regression, and Negative Binomial Regression

Two simulation studies in other disciplines have investigated the behavior of OLS regression, Poisson regression, and negative binomial regression for analyzing count data in different conditions. One is King's article (1988); he used sample size as the condition for comparing the OLS regression model and the Poisson regression model. The other is Sturman's research (1999); he used different distributions of the dependent variable, sample size, and distributions of the independent variables as the different conditions.

King (1988) compared the differences of the OLS regression model and the Poisson regression model for political science event count data. He commented that the OLS regression model, which is the most common model applied in political science, produces misspecification, inefficiency, bias, inconsistency, and insufficiency for modeling the count data. Therefore, he proposed the Poisson regression for modeling political event count data.

King conducted a Monte Carlo study with 12 different sample sizes (n=10, 20, 30, 40, 50, 100, 150, 200, 500, 750, 1000, and 2000), and simulated data from a Poisson distribution with parameter $\theta_i = X\beta$. β and X were arbitrarily chosen. A (3×1) β vector was chosen to include $\beta_1 = 2.0$, $\beta_2 = 0.4$, and $\beta_3 = -3.0$. His results showed that OLS regression produced more errors than the Poisson regression for all sample sizes. Most importantly, when the sample size increased, the number of OLS errors increased. Some researchers may still prefer to transform count data with a positively-skewed distribution using a logarithmic transformation to make the distribution of this count data approximately normal. King also commented that "if one must use the logged OLS (LOLS) model, collecting fewer observations might yield better results. (p.853)" In order words, the LOLS model may work for small sample size count data with a Poisson distribution. However, King mentioned that increases in sample size of this kind of count data produce increases in the variance of the count data. This would cause overdispersion for the transformed count data, so the LOLS model provided more Type I errors than the Poisson regression model. Therefore, the Poisson regression model still appears to be a better choice for analyzing count data with a Poisson distribution.

Similarly, Sturman (1999) compared eight models for analyzing positively skewed count data, the number of incidents of absenteeism in the field of business. The eight models were OLS, OLS with a transformed dependent variable, Tobit, Poisson, overdispersed Poisson, negative binomial, ordinal logistic, and ordinal probit. Sturman calculated the frequency with which each model incorrectly identified a statistically significant relationship from each simulated data set. In order to ensure generalizability, the author simulated five distributions for predicting absenteeism count data from previous articles about absenteeism. These distributions had varying degrees of skewness and kurtosis. Sturman's simulation estimated the likelihood of producing Type I errors under the influences of regression model type,

distribution of the dependent variable, sample size, and distribution of the independent variables. Sturman's results showed that OLS regression does not produce more false positives than expected by chance. The Poisson regression model yielded too many false statistically significant relationships. Sturman's study also showed that the negative binomial model produced fewer false positives. In other words, the negative binomial can serve as a conservative check of the results because it is likely to better match the characteristics of the data. Sturman also used MANOVA to examine whether the characteristics of the simulation study influence the number of Type I errors. He concluded that four factors, the regression model, distribution of the dependent variable, sample size, and distribution of the independent variables, all impacted the likelihood of Type I errors.

Both simulation studies provided empirical statistical results that showed differences between OLS regression and Poisson regression. However, their conclusions are not consistent. King (1988) stated that Poisson regression produced lower Type I error rates for all kinds of sample sizes for Poisson count data. On the other hand, Sturman (1999) concluded that Poisson regression tended to produce higher Type I error rates in different conditions, including distribution of the dependent variable, sample size, and distribution of the independent variables. Since Sturman did not provide detailed Type I error rates for individual conditions, the contradictions between the two articles may come from two major differences. One is that the count data in King's article followed a Poisson distribution, but the count data in Sturman's article were positively-skewed instead of following an entirely Poisson distribution. The other is that there were many variables which caused different conditions in Sturman's article, but the different conditions in King's article varied only by sample size. Beyond the OLS regression and the Poisson regression, Sturman also examined the negative binomial regression model, and concluded that the negative binomial was a better model for analyzing positively-skewed distributed count data.

After examining these two simulation studies, it is still not clear which regression model is more appropriate for analyzing count data like the number of university mathematics courses completed. Therefore, this study will investigate the conditions under which model could be better used for count data with a positively-skewed distribution under different sample size conditions.

Research Question

This study seeks to examine the Type I error rates for three models: OLS regression, Poisson regression, and negative binomial regression models for different sample sizes.

Methodology

Research Design

This study used a Monte Carlo simulation to evaluate which of the regression models was the most appropriate for different sample size conditions based on the Type I error rate. The Monte Carlo study employed a fully-crossed, factorial design with two independent variables: sample size and regression. The sample size contained seven conditions: 20, 40, 80, 100, 250, 500, and 2500. The three regression models, OLS, Poisson, and negative binomial, served as a within-subjects factor.

For Type I error calculation, the parameter values for β_1 and β_2 were set as zero. In order to create a good model of a sampling distribution, the number of replications was set at 10,000 for each sample size condition. The dependent variable was the Type I error rate over the two parameters. The descriptive statistics for the estimated correlations were computed.

In general, this Monte Carlo simulation explored the extent to which regression model is more likely to provide expected Type I error rates for simulated count data under different sample sizes.

The parameters and correlation relationships among variables for this study were from

Minnesota Mathematics Assessment Project (MNMAP) supported by the National Science Foundation under NSF0627986. This dataset included 4414 students who graduated from Minnesota high schools and enrolled at the University of Minnesota in Fall, 2002 or Fall, 2003 (Harwell, Post, Cutler, Anderson, Maeda, Wu, & Hyesook, 2006). The data consisted of student background variables as well as cumulative mathematics courses completed. The descriptive statistics of the number of university mathematics courses completed is shown in Table 1, and the distribution is positively-skewed (see Figure 1). The reason may be that students who complete more mathematics courses major in STEM fields; on the other hand, students who are in the humanity, business, or social science fields take fewer STEM courses. The descriptive statistics of the number of university mathematics courses completed is listed in Table 1.

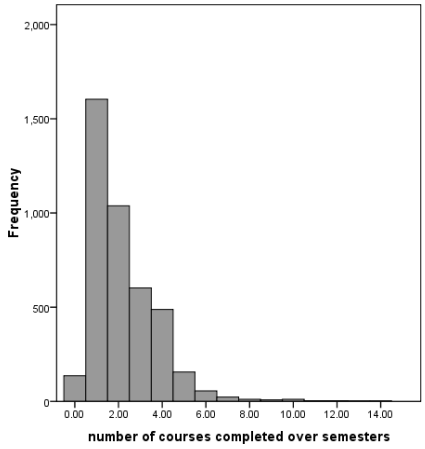


Figure 1 The distribution of the number of university mathematics courses completed in the MNMAP dataset

	Statistic	Std. Error
Mean	2.2008	0.02422
Variance	2.432	
Std. Deviation	1.55942	
Median	2	
Skewness	1.875	0.038
Kurtosis	6.661	0.076
Minimum	0	
Maximum	14	

Table 1
The Descriptive Statistics of the Number of University Mathematics Courses Completed in the MNMAP dataset

Moreover, the histogram for the residuals of the number of university mathematics courses completed from the regression model (Figure 2) with ACT mathematics and high school mathematics GPA as predictors, the P-P plot (Figure 3) and the Scatter Plot (Figure 4)

suggest that the assumption of normality and homoscedasticity are violated. In the histogram of residuals, it is positively-skewed instead of normally distributed. In the P-P plot of the regression standardized residual, most dots do not fall on the straight line. In the scatter plot, the residual points are spread unevenly throughout the range of the predicted values, and there appears to be a pattern to the residual plot. Based on the residual plot, the straight line does not appear to be a reasonable model. It can be said that there is more scatter above the 0-line than below, and an increasing, positive trend can be seen.

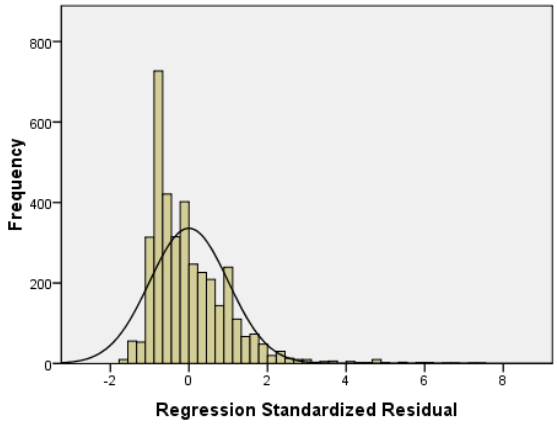


Figure 2 The histogram of residuals of the number of university mathematics courses completed in the MNMAP dataset

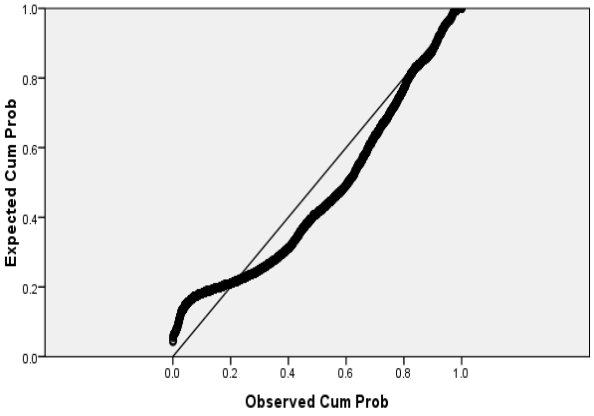


Figure 3 The normal P-P Plot of the regression standardized residual of number of university mathematics courses completed in the MNMAP dataset

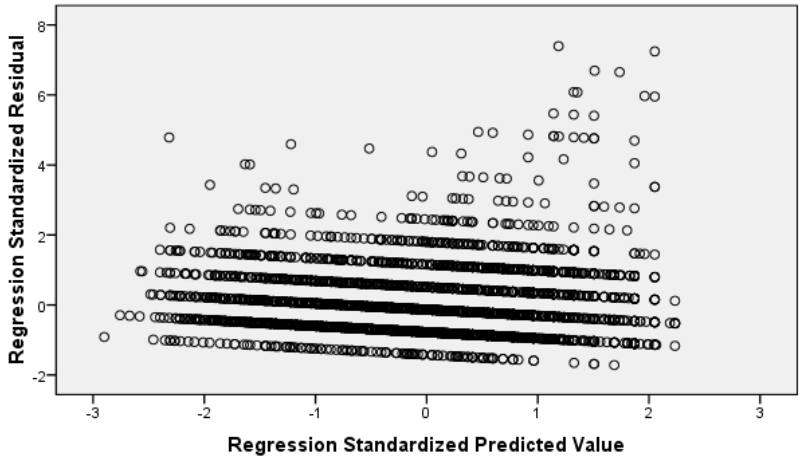


Figure 4 The scatterplot of the number of college mathematics courses completed in the MNMAP dataset

Data Simulation & Analyses

Three variables in this study were simulated based on MNMAP data: ACT mathematics score, high school (HS) mathematics GPA, and error term for each case. The ACT mathematics and high school mathematics GPA variables were created based on specific skewness and kurtosis. On the other hand, the distribution for the error term was tend to simulate as a Poisson distribution (more details are described later).

Data in different conditions were generated with the SAS System ("The SAS System for Windows 9.0," 2003). First, the two independent variables, ACT mathematics and high school mathematics GPA, were not normally distributed according to the real values in the MNMAP dataset. In order to simulate multivariate non-normal data, a method from Fleishman (1978) was applied in this study. The Fleishman transformation function is as follows:

$$Y = a + bX + cX^2 + dX^3 \tag{7}$$

Vale & Maurelli (1983) stated that Fleishman's method provides "an advantage over the other procedures because it can easily be extended to generate multivariate random numbers with specified intercorrelations and univariate means, variances, skews, and kurtoses" (p.465). The intercorrelation coefficients and descriptive statistics for the three variables were from the MNMAP data listed in Table 2 and Table 3.

Table 2

Correlation Coefficients Matrix Between Three Variables

		Total number of university mathematics courses	ACT mathematics	HS mathematics GPA
Total number of university mathematics courses	Pearson Correlation	1	0.178*	0.108*
	N	4144	4105	3827
ACT mathematics	Pearson Correlation	0.178*	1	0.492*
	N	4105	4105	3790

HS mathematics GPA	Pearson Correlation	0.108*	0.492*	1
	N	3827	3790	3827

* indicates $p < .05$

Table 3

Descriptive Statistics for the Three Variables

	Mean	Std.	Skewness	Kurtosis
Total number of university mathematics courses	2.200	1.559	1.875	6.661
ACT mathematics	24.650	4.959	-0.150	-0.624
HS mathematics GPA	3.357	0.643	-1.056	0.648

The correlation between the independent variables representing ACT mathematics and the HS mathematics GPA in this simulation study was set to 0.492, which comes from the MNMAP data. However, since the distributions of the ACT mathematics and the HS mathematics GPA were not normal, the intermediate correlation (ρ) between the two variables was calculated by using Vale and Maurelli's (1983) method:

$$R_{12} = \rho(b_1 b_2 + 3b_1 d_2 + 3d_1 b_2 + 9d_1 d_2) + \rho^2(2c_1 c_2) + \rho^3(6d_1 d_2) \quad (8)$$

Using SAS syntax from Fan & Fan (2005), the intermediate correlation coefficient was calculated as 0.53947.

Second, the error terms were also simulated with Fleishman's (1978) transformation function (7), described above. The mean of the number of mathematics courses completed over semesters in the MNMAP dataset was 2.2. Therefore, the mean and variance of the error term were set to 2.2. According to the skewness and kurtosis of a Poisson distribution described above in the equations 2.16 and 2.17, the skewness and kurtosis were set as 0.674 and 0.455. Further, in order to calculate the Type I error rate, the correlations between the error term and the ACT mathematics and between the error term and the HS mathematics GPA were each set to zero. Therefore, in this study, the known parameters of the distributions

are shown in Table 4. In addition, the desired correlation matrix is shown in Table 5.

Table 4

The Desired Parameters for the Three Distributions Simulated

	Mean	Std.	Skewness	Kurtosis
Error term	2.200	1.483	0.674	0.455
ACT mathematics	24.650	4.959	-0.150	-0.624
HS mathematics GPA	3.357	0.643	-1.056	0.648

Table 5

The Desired Correlation between Three Simulated Variables

	Error term	ACT mathematics	HS mathematics GPA
Error term	1	0	0
ACT mathematics	0	1	0.492
HS mathematics GPA	0	0.492	1

The estimated parameters and correlations based on 1,000,000 samples are shown in Table 6 and Table 7 below. The values in Table 6 indicate the estimated parameters were close to the known parameters. In addition, the correlations in Table 7 were also very close to the known correlations. Based on the correlation test formula (9) below, the simulated correlation between X1 and X2 is not statistically significant, since the t-value is -0.346 (p =0.492, and r =0.49117).

$$t = \frac{r - p}{\sqrt{[(1 - r^2)(1 - p^2)]/(n - 2)}} \tag{9}$$

Figure 7 shows the distribution of the simulated error term. However, based on these 1,000,000 simulated samples, there were 39,759 error terms lower than zero. The percentage is 0.0398. Due to the characteristics of the Poisson and the negative binomial regression model, if the value of the dependent variable is below zero, then the regression cannot be

defined. Therefore, in this study, two ways were used to solve this situation (some dependent variable values lower than zero). One is to add 1 to every value of the dependent variable since the minimum original simulated error term is -0.616. Therefore, only the mean of the error term would be changed, and the variance, skewness, and kurtosis were still the same. Figure 8 shows the distribution of the simulated adjusted error term. The second way is that when the error term was below zero, it would be truncated as zero. The descriptive estimated parameters for the adjusted error term and truncated error term are also listed in Table 6, and correlations between other variables are listed in Table 7. Figure 9 shows the distribution of the simulated truncated error term.

Table 6

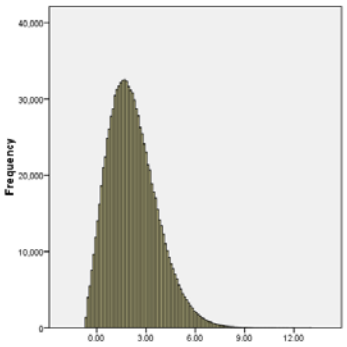
The Estimated Parameters for the Three Distributions Based on 1,000,000 Samples

	Mean	Std.	Skewness	Kurtosis	Minimum	Maximum
ACT mathematics	24.650	4.962	-0.150	-0.624	11.707	34.838
HS mathematics GPA	3.356	0.644	-1.055	0.652	-2.676	4.064
Error term	2.200	1.484	0.674	0.455	-0.616	12.100
Error term plus one	3.200	1.484	0.674	0.455	0.384	13.100
Truncated error term	2.210	1.468	0.728	0.471	0	12.100

Table 7

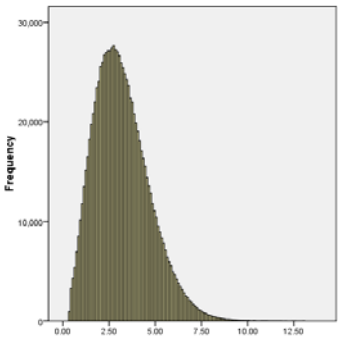
The Estimated Correlation between Three Simulated Variables

	ACT mathematics	HS mathematics GPA	Error term	Error term plus one	Truncated error term
ACT mathematics	1.00000	0.49117	0.00078	0.00078	0.00076
HS mathematics GPA	0.49117	1.00000	0.00015	0.00015	0.00012
Error term	0.00078	0.00015	1.00000	1.00000	0.99927
Error term plus one	0.00078	0.00015	1.00000	1.00000	0.99927
Truncated error term	0.00076	0.00012	0.99927	0.99927	1.00000



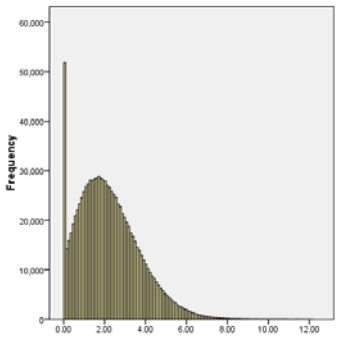
Mean =2.20, SD=1.484, N=1,000,000

Figure 7 The histogram of the simulated error term



Mean =3.20, SD=1.484, N=1,000,000

Figure 8 The histogram of the simulated adjusted error term



Mean =2.21, SD=1.464, N=1,000,000

Figure 9 The histogram of the simulated truncated error term

After simulating datasets under different sample size conditions, OLS regression, Poisson regression, and negative binomial regression, were used to estimate parameters for each regression result. Further, Type I error rates were calculated. In this study, Type I error rates at two nominal levels were estimated: 0.05 and 0.01. Since the null hypothesis is true, the estimated Type I error rate should equal 0.05 and 0.01 within sampling error, respectively. Type I error rate inflation was said to exist when the proportion of the rejection exceeded the upper and the lower limit of the criterion interval of

$$\alpha \pm 1.96\sqrt{\frac{\alpha(1-\alpha)}{N}} \tag{10}$$

Therefore, due to the 10,000 sample size, the criterion interval for the Type I error rate at 0.05 level is (0.045728 to 0.054271), and at 0.01 level is (0.008049 to 0.01195).

In this paper, the OLS regression model is estimated with PROC REG in the SAS system. PROC REG uses ordinary least squares, and the default link function is the identity link because it assumes that the relationship between the dependent variable and the independent variables is linear. On the other hand, PROC GENMOD was used for the Poisson regression model and the negative binomial regression model. The estimation in SAS GENMOD for the two models uses maximum likelihood estimation (MLE), and the link

function is log link. MLE is a preferred method of parameter estimation in statistics, particularly in non-linear modeling with non-normal data. MLE has four optimal properties in estimation: sufficiency (complete information about the parameter of interest contained in its MLE estimator), consistency (true parameter value that generated the data recovered asymptotically), efficiency (lowest-possible variance of parameter estimates achieved asymptotically), and parameterization invariance (same MLE solution obtained independent of the parameterization used) (Myung, 2003). Moreover, the log link function keeps the outcome variable from the fitted model positive. In the SAS output, a scale parameter has been included, even though the degree of freedom is zero for the scale parameter in Poisson regression. The difference in PROC GENMOD for Poisson regression and negative binomial regression is that negative binomial regression relaxes the assumption about equality of the mean and the variance in Poisson regression. This is because the scale parameter for all GLM count models is defined as one, and does not enter into the estimation process. PROC GENMOD allows the specification of a scale parameter to fit overdispersed Poisson distribution. SCALE indicates the value of the overdisperison scale parameter used in adjusting output statistics. PROC GENMOD allows the specification of a dispersion parameter to fit binomial distributions. Dispersion is the estimate of the log of the dispersion parameter. If dispersion is greater than zero the response variable is over-dispersed. If dispersion is less than zero the response variable is under-dispersed. And if the log of the dispersion parameter equals zero, the model reduces to the simpler Poisson model. In contrast to Hoffman's (2004) view that the negative binomial is not a true generalized linear model, Hilbe (2007) argued that negative binomial regression does have GLM status when it is a member of the single-parameter exponential family of distributions. Therefore, SAS GENMOD is appropriate for use in negative binomial regression.

Results

The purpose of this study was to assess whether these three models: OLS, Poisson, and negative binomial regression, preserved the appropriate Type I error rate, which is the probability of rejecting the null hypothesis when it is true. The results from the two methods used to modify less-than-zero dependent variable values will be discussed below separately.

Adjusted Error Term Results (Adding 1 to Each Value)

Tables 8 and 9 show the Type I error rates for the three regression models at $\alpha =$ 0.05 and 0.01 for the ACT mathematics scores (X1), and Tables 10 and 11 show the Type I error rate at alpha level equal to 0.05 and 0.01 for high school mathematics GPA (X2). To identify the estimated Type I error rates that differed from the nominal alpha level, the criterion interval given in equation 10 was used. If the estimated Type I error rate was beyond the criterion interval, then an asterisk was included with the results.

At $\alpha = 0.05$, OLS regression only produced more statistically significant relationships than expected by chance in the 500 sample size condition for X1 estimation. Poisson regression produced fewer Type I errors than expected by chance in all seven sample sizes in both two independent variables. On the other hand, negative binomial regression yielded more Type I errors when the sample sizes were 20, 40, and 2500 for X1, and when the sample sizes were 20 and 40 for X2.

For $\alpha = 0.01$, OLS regression produced more Type I errors when the sample sizes were 100 and 500 for X2. Poisson regression still produced much fewer Type I errors in all seven sample sizes for both X1 and X2. Negative binomial regression yielded more Type I errors when the sample sizes were 20, 40, and 80 for X1, and when the sample sizes were 20, 40, 80, and 100 for X2.

Table 12 and 13 show the average Type I error rates of X1 and X2 for the three regressions at alpha are equal to 0.05 and 0.01. OLS regression seemed to have stable Type I

error rates across different sizes. Poisson regression produced fewer Type I errors than expected by chance. Negative binomial regression yielded more Type I errors than expected by chance when the sample size was small, but it seemed to perform well in controlling Type I error rates when the sample size became larger.

Table 8

Type I Error Rates for the Three Regressions at Alpha = 0.05 Level for X1 (ACT mathematics)

	20	40	80	100	250	500	2500
OLS regression	0.0518	0.0513	0.0486	0.0495	0.0498	0.0550**	0.0474
Poisson regression	0.0166**	0.0178**	0.0186**	0.0191**	0.0192**	0.0193**	0.0166**
Negative Binomial regression	0.0727**	0.0656**	0.0507	0.0508	0.0464	0.0505	0.0414**

** The estimated Type I error rate is beyond the criterion interval of Type I error rate

Table 9

Type I Error Rates for the Three Regressions at Alpha = 0.01 level for X1 (ACT mathematics)

	20	40	80	100	250	500	2500
OLS regression	0.0095	0.0093	0.0108	0.0105	0.0116	0.0104	0.0102
Poisson regression	0.0020**	0.0013**	0.0022**	0.0023**	0.0016**	0.0027**	0.0019**
Negative Binomial regression	0.0193**	0.0160**	0.0123**	0.0113	0.0098	0.0095	0.0082

** The estimated Type I error rate is beyond the criterion interval of Type I error rate

Table 10

Type I Error Rates for the Three Regressions at Alpha = 0.05 level for X2 (HS mathematics GPA)

	20	40	80	100	250	500	2500
OLS regression	0.0512	0.0527	0.0502	0.0505	0.0501	0.0531	0.0534
Poisson regression	0.0177**	0.0183**	0.0181**	0.0190**	0.0207**	0.0192**	0.0197**
Negative Binomial regression	0.0701**	0.0632**	0.0527	0.0512	0.0458	0.0476	0.0473

** The estimated Type I error rate is beyond the criterion interval of Type I error rate

Table 11

Type I Error Rates for the Three Regressions at Alpha = 0.01 level for X2 (HS mathematics GPA)

	20	40	80	100	250	500	2500
OLS regression	0.0108	0.0099	0.0102	0.0127**	0.0112	0.0123**	0.0103
Poisson regression	0.0015**	0.0017**	0.0023**	0.0025**	0.0017**	0.0021**	0.0016**
Negative Binomial regression	0.0202**	0.0147**	0.0112**	0.0132**	0.0102	0.0110	0.0081

** The estimated Type I error rate is beyond the criterion interval of Type I error rate

Table 12

The Average Type I Error Rates for the Three Regressions at Alpha = 0.05 level

	20	40	80	100	250	500	2500
OLS regression	0.0515	0.0520	0.494	0.0500	0.0500	0.0541	0.0504
Poisson regression	0.0172	0.0181	0.0184	0.0191	0.0200	0.0193	0.0182
Negative Binomial regression	0.0714	0.0644	0.0517	0.0510	0.0461	0.0491	0.0444

Table 13

The Average Type I Error Rates for the Three Regressions at Alpha = 0.01 level

	20	40	80	100	250	500	2500
OLS regression	0.0102	0.0096	0.0105	0.0116	0.0114	0.0114	0.0103
Poisson regression	0.0018	0.0015	0.0023	0.0024	0.0017	0.0024	0.0018
Negative Binomial regression	0.0198	0.0154	0.0118	0.0123	0.0100	0.0103	0.0082

Truncated Error Term Results

Tables 14 and 15 show the Type I error rates for the three regressions at $\alpha =$ 0.05 and 0.01 for X1, and Tables 16 and 17 show the Type I error rates for $\alpha = 0.05$ and 0.01 for X2. For $\alpha = 0.05$, OLS regression produced Type I error rates consistent with chance for all seven sample size conditions. In contrast, for X1 using Poisson regression, three out of seven estimated Type I error rates were outside of the criterion interval (sample sizes of 20, 40, and

2500). For X2 using Poisson regression, two out of the seven Type I error rates were outside of the criterion interval (40 and 2500). Poisson regression did not yield statistically significant relationships in the less extreme sample sizes (e.g., 80~500). On the other hand, for X1 in the negative binomial regression, five out of seven Type I error rates were outside of the criterion interval (sample sizes of 20, 40, 80, 100 and 500). For X2 using the negative binomial regression, six out of seven Type I error rates were out of the criterion interval (20, 40, 80, 100, 250 and 500). Only with the largest sample size of 2500 did negative binomial regression produce a Type I error rate consistent with chance. However, Poisson and negative binomial regression produced different kinds of statistically significant relationships. Poisson regression produced lower Type I error rates, but negative binomial regression produced higher.

For $\alpha = 0.01$, OLS and Poisson regression did not produce more statistically significant relationships than expected by chance under the seven sample size conditions. On the other hand, negative binomial regression still produced more Type I error rates in six out of the seven sample sizes for both independent variables.

Table 18 and 19 show the average Type I error rates of X1 and X2 for the three regressions for $\alpha = 0.05$ and 0.01. OLS regression seemed to have stable Type I error rates across different sample sizes. Poisson regression sometimes produced fewer Type I errors than expected, and negative binomial regression yielded more Type I errors than expected by chance.

Table 14

Type I Error Rates for the Three Regressions at Alpha = 0.05 Level for X1 (ACT mathematics)

	20	40	80	100	250	500	2500
OLS regression	0.0505	0.0485	0.0537	0.0508	0.0497	0.0485	0.0472
Poisson regression	0.0438**	0.0457**	0.0508	0.0478	0.0478	0.0438**	0.0437**

Negative Binomial regression	0.0704**	0.0664**	0.0630**	0.0559**	0.0513	0.0704**	0.0473

** The estimated Type I error rate is beyond the criterion interval of Type I error rate

Table 15

Type I Error Rates for the Three Regressions at Alpha = 0.01 level for X2 (HS mathematics GPA)

	20	40	80	100	250	500	2500
OLS regression	0.0116	0.0095	0.0116	0.0098	0.0111	0.0095	0.0110
Poisson regression	0.0085	0.0084	0.0097	0.0091	0.0091	0.0085	0.0106
Negative Binomial regression	0.0240**	0.0180**	0.0138**	0.0129**	0.0130**	0.0240**	0.0111

** The estimated Type I error rate is beyond the criterion interval of Type I error rate

Table 16

Type I Error Rates for the Three Regressions at Alpha = 0.01 level for X1 (ACT mathematics)

	20	40	80	100	250	500	2500
OLS regression	0.0118	0.0091	0.0110	0.0098	0.0111	0.0091	0.0093
Poisson regression	0.0112	0.0092	0.0113	0.0098	0.0098	0.0112	0.0086
Negative Binomial regression	0.0233**	0.0169**	0.0149**	0.0131**	0.0120**	0.0233**	0.0093

** The estimated Type I error rate is beyond the criterion interval of Type I error rate

Table 17

Type I Error Rates for the Three Regressions at Alpha = 0.05 level for X2 (HS mathematics GPA)

	20	40	80	100	250	500	2500
OLS regression	0.0532	0.0463	0.0536	0.0490	0.0537	0.0463	0.0478
Poisson regression	0.0465	0.0421**	0.0509	0.0470	0.0470	0.0465	0.0449**
Negative Binomial regression	0.0786**	0.0661**	0.0624**	0.0563**	0.0568**	0.0786**	0.0480

** The estimated Type I error rate is beyond the criterion interval of Type I error rate

Table 18

The Average Type I Error Rates for the Three Regressions at Alpha = 0.05 level

	20	40	80	100	250	500	2500

OLS regression	0.0519	0.0474	0.0537	0.0499	0.0517	0.0474	0.0475
Poisson regression	0.0452	0.0439	0.0509	0.0474	0.0474	0.0452	0.0443
Negative Binomial regression	0.0745	0.0663	0.0627	0.0561	0.0541	0.0745	0.0477

Table 19

The Average Type I Error Rates for the Three Regressions at Alpha = 0.01 level

	20	40	80	100	250	500	2500
OLS regression	0.0117	0.0093	0.0113	0.0098	0.0111	0.0093	0.0097
Poisson regression	0.0099	0.0088	0.0105	0.0095	0.0095	0.0099	0.0096
Negative Binomial regression	0.0237	0.0175	0.0144	0.0100	0.0125	0.0137	0.0102

In summary, in the adjusted error term results, OLS regression generally seemed to perform to control Type I error rates well. Negative binomial regression controlled the Type I error rate well when the sample size was large, but did not perform well when the sample size was small. Poisson regression seemed to be too conservative because it produces less statistically significant relationships than expected by chance in all sample size conditions. In the truncated error term results, OLS regression performed better than both Poisson and negative binomial regression. Poisson regression tended to produce less Type I errors than expected at the 0.05 level, whereas the negative binomial regression tended to yield more Type I errors than expected at both 0.05 and 0.01.

Discussion

Three regression models have been proposed for the analysis of count data. These methods have been described in the context of modeling the number of university mathematics courses completed, where a large proportion of students took few mathematics

courses, and a small number of students completed a high number of mathematics courses.

The results indicate that OLS regression performs better than Poisson and the negative binomial regression because OLS regression does not produce more or less false statistically significant relationships than expected theoretically. On the other hand, Poisson regression produced too few false statistically significant relationships, and negative binomial regression only controlled the Type I error rate when the sample size was quite large. However, if there are too many zeros in the dependent variable, negative binomial regression may produce too many false statistically significant relationships. In other words, the difference between Poisson regression and the negative binomial regression is that Poisson regression is more conservative because it produces fewer statistically significant relationships than expected by chance. On the other hand, the negative binomial regression is more liberal because it produces more statistically significant relationships than expected by chance.

The results in this study are partially consistent with Struman's (1999) results that OLS regression does not produce more false positives than expected by chance. However, the differences between the current study and Struman's study are that in this study, the Poisson regression model produced fewer false statistically significant relationships, and the negative binomial model produced more false statistically significant relationships than expected by chance, which are the opposite of Sturman's results. There are many differences which may influence results between the two studies, such as the distributions of independent variables, distributions of dependent variables, sample size, and so on, and it is difficult to pinpoint the exact reasons for the discrepancies.

Limitations & Future Research

The present study has several limitations. First, this study used Fleishman' (1978) transformation function to simulate the distribution of the error terms as a Poisson distribution. However, some values were negative in this case, and these were truncated to

zero. Although the skewness and kurtosis of the truncated simulated error terms were close to the simulated error terms according to the characteristics of the Poisson distribution, approximately 4% of the error terms were set to zero. Therefore, the left side of the truncated simulation error term distribution was not like a Poisson distribution.

Second, in this study, the mean and the variance of the error term were set the same. However, in real situations, this assumption is rarely met. Therefore, the distribution of the error term could be featured as over-dispersion or under-dispersion in future simulation studies.

Third, the present study attempted to model a set of variables that are typically used in mathematics educational research to understand the performance of OLS, Poisson, and negative binomial regression under different sample size conditions. Reseachers (Hutchinson & Bandalos, 1997) commented that results of Monte Carlo studies would be very useful if the conditions studied were encountered in practice. Therefore, in this study, the values for different between-factor conditions and the intercorrelation coefficients matrix were generated by using the variables in the MNMAP data to make the generated data realistic. However, only two independent variables were used in this study. In order to increase the generalizability, more parameters and more correlations should be studied.

Finally, there are other regression models for modeling count outcomes, such as over-dispersed Poisson, zero-inflated Poisson, zero-inflated negative binomial, and so on. For instance, for the second analysis in this study, about 4% of the data were zero, so they are called zero inflated data. The zero-dispersed Poisson or zero-inflated negative binomial model may be alternative ways to model the data. Therefore, when educational researchers look for a statistical model to fit their count data such as the number of university mathematics courses completed, they should use a more appropriate model in their situations.

Acknowledgement

I would like to express my sincere gratitude to my academic advisor, Dr. Ernest Davenport and Dr. Michael Harwell at the University of Minnesota, for their inspiration, encouragement, and guide for this paper.

References

Ayalon, H., & Yogev, A. (1997). Students, schools, and enrollment in science and humanity courses in Israeli secondary education. *Educational Evaluation and Policy Analysis, 19*(4), 339-353.

Baldi, S., Jin, Y., Skemer, M., Green, P. J., & Herget, D. (2007). Highlights From PISA 2006: Performance of U.S. 15-Year-Old Students in Science and Mathematics Literacy in an International Context (NCES 2008-016). National Center for Education Statistics, Institution of Education Sciences, U.S. Department of Education. Washington, DC. Retrieved January 19, 2008, from http://nces.ed.gov/PUBSEARCH/pubsinfo.asp?pubid=2008016

Beck, E. M., & Tolnay, S. E. (1995). Analyzing historical count data. *Historical Methods, 28*(3), 125-132.

Blum, J. R., & Fattu, M. A. (1954). Nonparametric Methods. *Review of Educational Research,24*(5), 467-487.

Caudill, S. B., & Mixon, F. G. (1995). Modeling household fertility decisions: Estimation and testing of censored regression models for count data. *Empirical Economics, 20,* 183-196.

Chuansheng, C., & Stevenson, H. (1995). Motivation and mathematics achievement: A comparative study of Asian-American, Caucasian-American, and East Asian high school students. *Child Development, 66,* 1215-1234

Cole, W. M. (2006). Accrediting culture: An analysis of tribal and historically black college curriculum. *Sociology of Education, 79,* 355-388.

Davenport, E. C., Davison, M. L., Kuang, H., Ding, S., Kim, S.-K., & Kwak, N. (1998). High school mathematics course-taking by gender and ethnicity. *American Educational Research Journal, 35*(3), 497-514.

Davenport, E. C., Davison, M. L., Wu, Y.-C., Kim, S-K., Kuang, H., Kwak, N., & Chan,C.-K. (2004). The influence of amount and content of high school mathematics coursework on student achievement and ethnic achievement gap. Unpublished manuscript, University of Minnesota.

DeGroot, M. H., & Schervish, M. J. (2002). *Probability and statistics*. Boston: Addison Wesley.

Fang, X., & Fang, X. (2005). Using SAS for Monte Carlo simulation research in SEM. *Structural Equation Modeling. 12*(2), 299-333.

Fleishman, A. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*(4), 521-532.

Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin, 118*(3), 392-404.

Glenn, J. (2000). *Before it's too late: A report to the nation.* National Commission on Mathematics and Science Teaching for the 21st Century. Washington, DC: U.S. Department of Education. Retrieved May, 20, 2008, from http://www.ed.gov/inits/Math/glenn/report.pdf

Goyette, K. (1999). Application to college: A comparison of Asian American and White high school students. The Annual Meeting of the American Educational Research Association. Montreal, Quebec, Canada.

Harwell, M. (1988). Choosing between parametric and nonparametric tests. *Journal of Counseling and Development, 67*, 35-38.

Harwell, M., Post. T.R., Cutler, A., Anderson, E., Maeda, Y., Wu, K., & Hyesook, S. (April, 2006). A longitudinal study of student achievement, course-taking patterns, and persistence in post-secondary mathematics and economics classes after participating in a traditional, University of Chicago School mathematics project, or standards-based mathematics curriculum in high school.

Hausman, J., Hall, B. H., & Griliches, Z. (1984). Econometric models for count data with an application to the patents-R & D relationship. *Econometrica, 52*(4), 909-938.

Hilbe, J. M. (2007). *Negative Binomial Regression*. Cambridge: Cambridge University Press.

King, G. (1988). Statistical models for political science event counts: Bias in conventional procedures and evidence for the Exponential Poisson regression model. *American Journal of Political Science, 32*(3), 838-864.

King, G. (1989). Event count models for international relations: Generalizations and application. *International Studies Quarterly, 33*, 123-147.

Hoffman, J. P. (2004). *Generalized linear models: an applied approach* (2004 ed.). Boston: Pearson Education, Inc.

House, J. D. (2002). Instructional practices and mathematics achievement of adolescent students in Chinese Taipei: Results from the TIMSS 1999 assessment. *Child Study Journal. 32*(3), 157-178.

Kuenzi, J. J., Matthews, C. M., & Mangan, B. F. (2006). Science, technology, engineering and mathematics (STEM) education issues and legislative options. Report of the Congressional Research Service, Library of Congress No. RL33434. Retrieved January 19, 2008, from http://media.umassp.edu/massedu/stem/CRS%20Report%20to20%20Congress.pdf

Hutchinson, S., & Bandalos, D. (1997). A guide to Monte Carlo simulation research for applied researchers. *Journal of Vocational Education Research, 22*(4), 233-245.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models*. New York: McGraw.

Lee, L.F. (1986). Specification test for Poisson regression models. *International Economic Review*, *27*, 689–706.

Lee, V. E., Croninger, R. G., & Smith, J. B. (1997). Course-Taking, equity, and mathematics learning: Testing the constrained curriculum hypothesis in U.S. secondary schools. E*ducational Evaluation and Policy Analysis, 19*(2), 99-121.

Martin, M.O., Mullis, I.V.S., & Foy, P. (with Olson, J.F., Erberber, E., Preuschoff, C., & Galia, J.). (2008). *TIMSS 2007 International Science Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I.V.S., Martin, M.O., & Foy, P. (with Olson, J.F., Preuschoff, C., Erberber, E., Arora, A., & Galia, J.). (2008). *TIMSS 2007 international mathematics report: Findings from IEA's trends in international mathematics and science study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematicsal Psychology, 47,* 90-100.

Parsad, B., & Lewis, L. (2003). *Remedial Education at Degree-granting Postsecondary Institutions in Fall 2000.* (Report No. NCES 2004-010). Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved February 7, 2008, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2004010

Parsons, J. E., Adler, T. F., & Kaczala, C. M. (1982). Socialization of achievement attitudes and beliefs: Parental influences. *Child Development.53,* 310-321.

Pedhazur, E. J., & Schmelkin, L. P. (2007). *Measurement, design, and analysis: An integrated approach.* Lawrence Erlbaum Associates Inc., NJ.

Puntanen, S., & Styan, G. P. H. (1989). The equality of the ordinary least squares estimator and the best linear unbiased estimator. American Statistician. 43(3), 153-161

Stedman, L. C. (1997). International achievement differences: An assessment of a new perspective. *Educational Researcher. 26*(3), 4-15.

Sturman, M. C. (1999). Multiple approaches to analyzing count data in studies of individual differences: The propensity for type 1 errors, illustrated with the case of absenteeism prediction. *Educational and Psychological Measurement, 59*(3), 414-430.

The SAS System for Windows 9.0. (2003). Cary, NC, USA: SAS Institute.

United States National Academies. (2007). *Rising above the gathering storm: Energizing and employing America for a brighter economic future.* Washington, DC: The National Academies Press.

Yee, D. K., & Eccles, J. S. (1988). Parent perceptions and attributions for children's math achievement. *Sex Roles, 19,* 317-333.

Vale, D., & Maurelli, V. (1983). Simulating multivariate nonnormal distributions. *Psychometrika, 48*(3), 465-471.