Paper 372-2008

Modeling Zero-Inflated Count Data with Underdispersion and Overdispersion

Adrienne Tin, Research Foundation for Mental Hygiene, New York, NY

ABSTRACT

A common problem in modeling count data is underdispersion or overdispersion. This paper discusses the distinction between overdispersion due to excess zeros and overdispersion due to values that are greater than 0. It shows how to use exploratory data analysis to determine the dispersion patterns and that the dispersion patterns can change depending on the predictors and the subpopulation that are included in the analysis. Further, the paper discusses how to fit zero-inflated models using PROC NLMIXED and compares the model fit. The data is from the National Health and Nutrition Examination Survey (NHANES 2003-2004).

BACKGROUND

In its simplest form, the Poisson distribution models the number of events from a memoryless exponential process where the event rate is constant. The Poisson density function only depends on the mean number of events, *u*.

$$\Pr(y \mid u) = \frac{\exp(-u)u^{y}}{y!}$$

1/u is the event rate.

However, the occurrence rates of many events are not constant. This can be a result of individual or group heterogeneity, contagious effect, or spells (Long, 1997; Eaton, 1978). This background section summarizes the various metamorphism of the Poisson distribution for accommodating heterogeneity. The formulas are largely from Long (1997). Some of the notations are adapted to match what are commonly used in SAS documentation.

If the event rate depends on individual characteristics, then the Poisson model can be modified to let the mean, *u*, be a function of individual characteristics, $\mathbf{x}_i$. Then

$$u_i = \exp(x_i B)$$

$$\Pr(y_i \mid x_i) = \frac{\exp(-u_i)u_i^{yi}}{y_i!}$$

Each case has its expected mean, u_i, but the variance is still constraint to be the mean. If $\mathbf{x}_i$ does not contain all the relevant predictors, then the model may not account for all the overdispersion (Agresti, 2002). In SAS, GENMOD or GLIMMIX can estimate a dispersion parameter, k, of a Poisson model using the deviance or the Pearson statistics, although k is not a parameter in the distribution. With this technique, $\mathrm{Var}(u_i) = ku_i$ where $k > 0$. When $k < 1$, the variance is less than the mean. Therefore, the Poisson model with a dispersion parameter can model both under- and over-dispersion.

The next extension of the poisson model is letting the expected number of events, u_i, be a function of $\mathbf{x}_i$, and some unobserved random variable, e_i:

$$u_i = \exp(x_i B + e_i)$$

Assuming that $\exp(e_i)$ has a gamma distribution with an expected value of 1 and a shape parameter, 1/k, where $k > 0$, then $\Pr(y_i \mid x_i)$ has a negative binomial distribution.

$$\Pr(y_i \mid x_i) = \frac{\Gamma(y_i + 1/k)(ku_i)^{y_i}}{y_i!\Gamma(1/k)(1+ku_i)^{y_i+1/k}}$$

Since $E(\exp(e_i)) = 1, E(u_i) = E(\exp(x_i B + e_i)) = E(\exp(x_i B))$, the expect value of u_i does not change whether we assume a Poisson or a negative binomial distributions. However, the variance functions are different.

With the negative binomial distribution, var(y_i |$\mathbf{x}_i$) = $u_i(1+ku_i) > u_i$, because k > 0. Therefore, the negative binomial assumes that the variance is greater than the mean. It is only appropriate for modeling overdispersion and not for underdispersion.

With some types of data, a high percentage of zero count occurs at all levels of the predictors and cannot be accounted for by the variance functions of the Poisson or the negative binomial distributions. To model this excess of zeros, a zero-inflated model may be appropriate. This kind of models assumes that the observations may belong to two groups. One group, g1, is very likely to have a count of zero. The other group, g2, follows one of the count data distribution, which is either Poisson or negative binomial. The group membership is estimated by a probability, p, which depends on a set of predictors, z_i, that may be different from the count data model.

Assuming that

$$p_i = \Pr(i \in g1 \mid z_i)$$

$$1 - p_i = q_i = \Pr(i \in g2 \mid z_i)$$

then a zero-inflated model has this general form:

$$\Pr(y_i \mid x_i, z_i) = p_i + (1 - p_i)g(u_i) \quad \text{if } y_i = 0 \text{ (Equation 1)}$$

$$\Pr(y_i \mid x_i, z_i) = (1 - p_i)f(u_i) \quad \text{if } y_i > 0 \text{ (Equation 2)}$$

where $g(u_i) = \Pr(y_i=0 \mid x_i)$ in the count data model, and $f(u_i)$ is the density of either the Poisson or the negative binomial distribution. Equation 1 assumes that the zeros are generated from two sources based on the probability of whether a case belongs to group 1 or group 2.

The expected value of a zero-inflated Poisson or negative binomial model is:

$$E(y_i \mid x_i, z_i) = u_i - u_i p_i$$

For the zero-inflated Poisson (ZIP) model, the variance is:

$$Var(y_i \mid xi, zi) = u_i(1 - p_i)(1 + u_i p_i)$$

For the zero-inflated negative binomial (ZINB) model, the variance is:

$$Var(y_i \mid xi, zi) = u_i(1 - p_i)[(1 + u_i(p_i + k)]$$

The four models discussed above offer different ways to fit the mean and variance relationship. The Poisson model with a dispersion parameter assumes that the variance is a linear function of the mean. The negative binomial model assumes that the variance is a quadratic function of the mean. Both variance functions are monotone in the first quadrant. The Poisson variance function may be monotone non-increasing or non-decreasing, while the negative binomial variance function is monotone increasing. On the other hand, the variance functions of the ZIP and ZINB models concave downward in the first quadrant, allowing the variance to increase and then decrease with the mean.

The rest of the paper consists of three examples in modeling zero-inflated count data. The outcome is the number of hospitalization in a 12-month period from the National Health and Nutrition Examination Survey (NHANES 2003-2004). The first example presents a simple model to fit the hospitalization data using age as the only predictor. The second example shows that the dispersion pattern in a subgroup differs from the overall disperson pattern. The third example is an attempt to achieve a better fit of the hospitalization data by using age, gender, marital status, and higher order terms as predictors.

EXAMPLE 1 -- EXCESS ZEROS AND UNDERDISPERSION WHEN THE COUNTS ARE GREATER THAN 0

EXPLORATORY DATA ANALYSIS

The National Health and Nutrition Examination Survey (NHANES) of 2003-2004 has a total of 10117 cases with hospitalization information. Around 90% of the cases have no incidence of hospitalization in the 12-month survey period (Table 1). In the substantive sense, a zero-inflated model may be appropriate because we can imagine certain segment of the population tend to be hospitalized because of some chronic illnesses. This part of the data can be fitted by a count data model. For the rest of the population, hospitalization is often the result of an accident, which can be estimated by a logit or probit model.

Figure 1 plots that mean numbers of hospitalization by age group and gender. The x-axis has the midpoint of each age group. For example, 5 represents the age group from birth to 9 years old. The plot shows that the number of hospitalization tends to increase with age, except for female during the child-bearing years. This suggests that age, gender, and marital status are good predictors of the number of hospitalization. In this first example, as a way of illustrating the differences between the Poisson, negative binomial, ZIP, and ZINB models, we use age as the only predictor. In the third example, we will consider age, gender, and marital status as predictors.

Since age is a continuous variable, the cases were grouped by age in decade to analyze the relation between mean and variance. When all cases were included, the data exhibits a pattern of overdispersion. In Figure 2, the variances are always above the means. If we only consider the cases with at least one hospitalization, then the data exhibits a pattern of underdispersion. In Figure 3, the variances are below the means for most age groups, except when ages are between 50 to 59. Since there is overdispersion when all cases are included, if we use an one-part model, then it is very likely that the negative binomial model will have a better fit. If we use a two-part model, then it is very likely that the ZIP model is a better fit, since there is underdispersion when the counts are greater than zero.

Table 1 Frequencies of number of hospitalization

Number of Hospitalization	Freq.	Percent	Cumulative Frequency
0	9118	90.13	9118
1	725	7.17	9843
2	165	1.63	10008
3	57	0.56	10065
4	17	0.17	10082
5	14	0.14	10096
6	21	0.21	10117

Figure 1 Proportion of subjects with hospitalization by age group

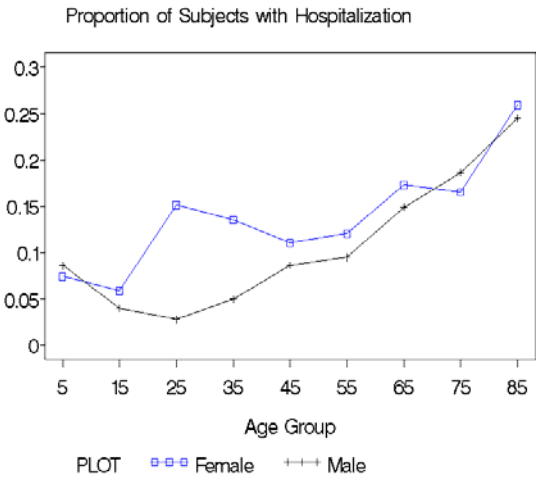


Figure 2 Means and variances by age group (all cases are included.)

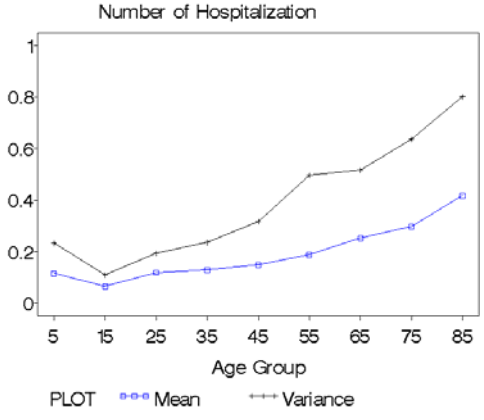
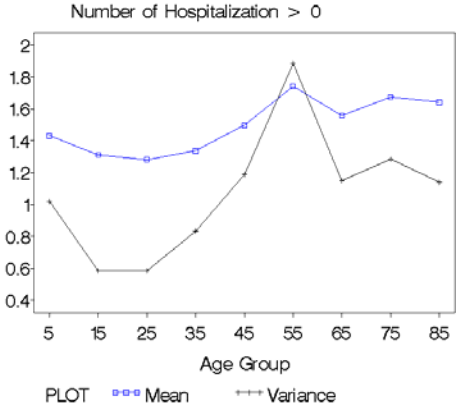


Figure 3 Means and variances by age group (only cases with at least one hospitalization are included.)



MODEL RESULTS AND COMPARISONS

The programs for fitting the ZIP and ZINB models are in appendix I. The Poisson model with dispersion and the negative binomial models are fitted using PROC GLIMMIX. Table 2 lists the results of this simplistic model with age as the only predictor. The Poisson and the negative binomial models are nested models, they can be compared using the log likelihood, likewise with the ZIP and ZINB models. The negative binomial model has one more parameter and a much lower -2 log likelihood than the Poisson model, this means that the negative binomial model is a better fit even without a formal likelihood ratio test. Between the ZIP and the ZINB model, the ZIP model has a much lower -2 log likelihood. The likelihood ratio test with one degree of freedom confirm that the ZIP model is a better fit over the ZINB model with a p-value < 0.0001.

Since the one-part models and the two-part models are not nested models, we cannot compare them using the log likelihood. We can compare them informally by how well they predict. Figure 4 and 5 plot the mean number of hospitalization by age group and compare it with the predictions. There are no big differences among the Poisson, negative binomial, and ZIP. However, ZINB performs extremely poorly. The scale parameter estimate in ZINB model is very close to 0. This is an indication of the lack of overdispersion among cases with at least one hospitalization.

Table 2 Parameter estimates from the one-part models

One-part Models	Poisson w. Dispersion	Negative Binomial
-2 Log Likelihood	9248.66	8190.53
	Estimate (S.E)	Estimate (S.E.)
Intercept	-2.6052*(0.0686)	-2.5632 (0.0644)
Age / 10	0.1883 (0.0133)	0.1767 (0.0139)
Scale	2.0031	5.1960 (0.3565)

Table 3 Parameter estimates from the two-part models

Two-part Models	ZIP	ZINB
-2 Log Likelihood	8312	8598
Logit Part (Pr[case follows a count data distribution])		
Intercept	-2.0094 (0.0917)	-1.1883 (0.0573)
Age / 10	0.1414 (0.0175)	-0.0043 (0.0152)
Poisson / Neg. Bin		
Intercept	-0.4657 (0.0816)	-0.2495 (0.0244)
Age / 10	0.0703 (0.0141)	0.0287 (0.0107)
Scale	NA	5.73E-09

Figure 4 Observed and predicted from one-part models

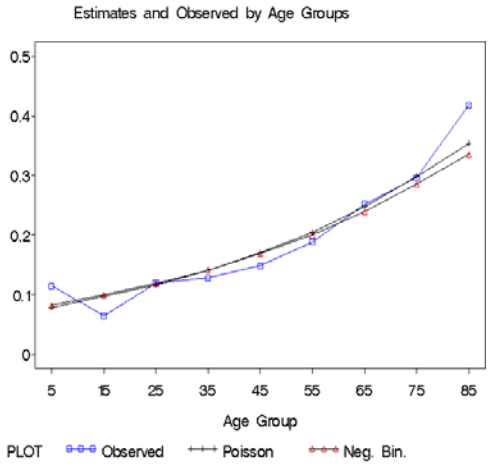
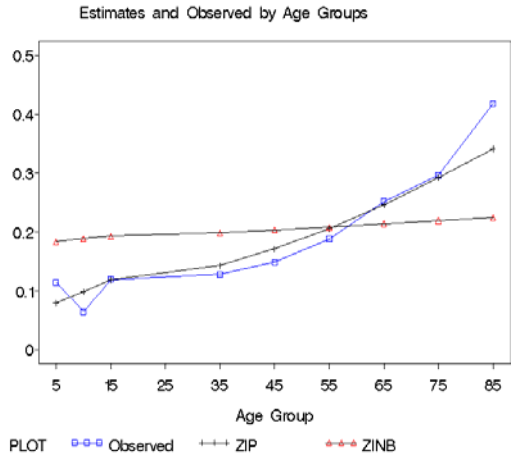


Figure 5 Observed and predicted from two-part models



EXAMPLE 2 - DATA WITH EXCESS ZEROS AND OVERDISPERSION AT ALL LEVELS

Example 1 shows that when the number of hospitalization is greater than zero, the data exhibits underdispersion. However, within some subgroups, the relationship between mean and variance may be different. It turns out that among Black females between the ages of 40 and 89, the data shows an overdispersion pattern when the number of hospitalization is greater than zero. In Figure 6 and 7, the variances are above the mean for all age groups.

Figure 6 Means and variances of the number of hospitalization among Black females.

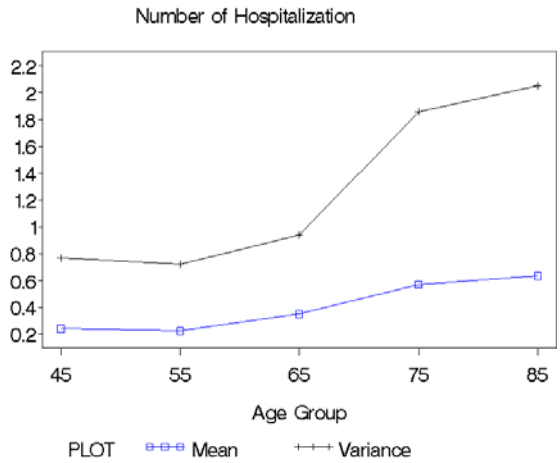
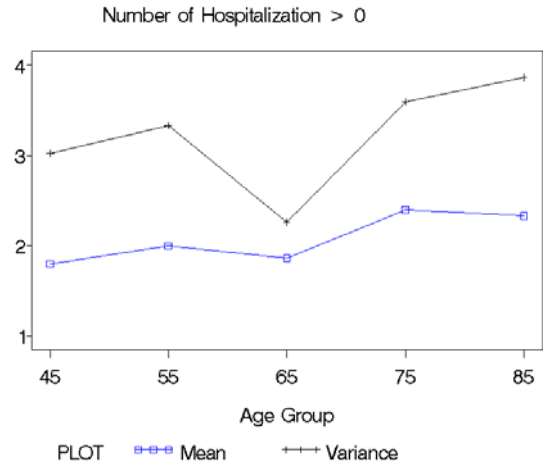


Figure 7 Means and variances of the number of hospitalization among Black females when count > 0



For this subgroup, because of overdispersion at all levels of the data, we can anticipate that the ZINB model fits better than the ZIP model. The ZIP model has a -2 loglikelihood of 447.6. On the other hand, a ZINB model has a -2 loglikelihood of 443.8. Likelihood ratio test indicates that the zero-inflated negative binomial model is more likely at the 0.05 level.

When comparing the standard error estimates between the two models in table 4 and 5, the ZIP model has the common problem of underestimated standard errors when overdispersion is not model properly. For example, the p-value of Age in the ZIP logit model is 0.0586, while that in the ZINB logit model is 0.09.

Table 4. ZIP model of older Black females

		Estimate	Std. Error	DF	t Value	Pr > |t|
Logit part	Intercept	-2.772	0.8134	318	-3.41	0.0007
	Age in decade	0.2448	0.129	318	1.9	0.0586
Poisson part	Intercept	-0.1573	0.5806	318	-0.27	0.7866
	Age in decade	0.101	0.08783	318	1.15	0.2508

Table 5. ZINB model of older Black females

		Estimate	Std. Error	DF	t Value	Pr > |t|
Logit part	Intercept	-2.635	0.877	318	-3	0.003
	Age in decade	0.234	0.138	318	1.7	0.090
Negative Binomial	Intercept	-0.367	0.715	318	-0.51	0.608
	Age in decade	0.115	0.107	318	1.08	0.282
	Dispersion	0.211	0.146	318	1.44	0.150

EXAMPLE 3 – A MORE SATISFACTORY MODEL

EXPLORATORY DATA ANALYSIS

Since Figure 1 strongly suggests that in addition to age, gender and marital status may be good predictors of the number of hospitalization, we conduct some exploratory analysis on the relationship between mean and variance when the subjects are divided into subgroups by age, gender, and marital status. Figure 8 and 9 plot the mean versus the variance within the subgroups. The diagonal line represents x=y. The points above the line represent the subgroups with the variance is higher the mean. When all cases are considered, 90% of the subgroups have variance greater than the mean. This implies that if we use a one-part model, then the negative binomial model may be a better fit because of the overdispersion pattern. When only cases with at least one hospitalization are considered, only 21% of the subgroups have variance above the mean. Therefore, underdispersion is still the dominant pattern when the counts are greater than zero.

Figure 8. Mean and Variance by age group, gender, and marital status

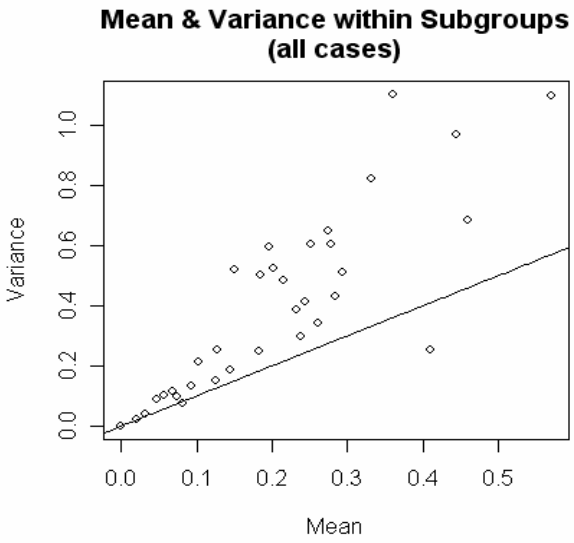
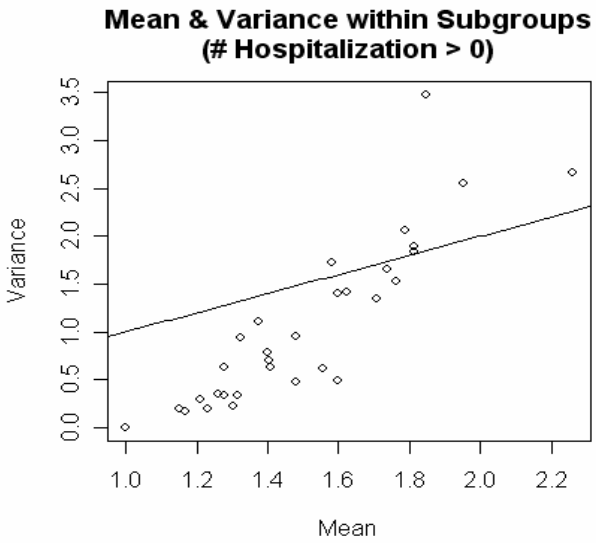


Figure 9. Mean and Variance by age group, gender, and marital status when count is greaten than 0



MODEL RESULTS AND COMPARISON

We tested all four models (Poisson, Negative Binomial, ZIP, and ZINB) up to third order interaction terms and the fourth order term of age since the curve in Figure 1 suggests a fourth order polynomial. The final models include all the significant higher order terms and their lower order terms. The results are in table 6 and 7. Age is measured in decade and is centered to avoid the problem of collinearity when interaction and higher order terms are included. Within the one-part models, the negative binomial has a better fit with a -2 log likelihood of 7978 versus 8912 for the Poisson model. Within the two-part models, NLMIXED converged only with the ZIP model and did not converge with the ZINB model as more second order terms were added. This is probably due to underdispersion, which leads to difficulty in estimating the scale parameter as more predictors are added.

When we compare the predicted means from the negative binomial and the ZIP models, they are not very different by age group (see Figure 10). However, when the subjects are separated by gender and married status, the ZIP model does considerably better among young married females (see Figure 12). This may be attributed to the logit part of the ZIP model, which estimates the probability of hospitalization. If most young married females were hospitalized due to childbirth, then they are hospitalized only once. Estimating their number of hospitalization is like estimating their probability of having a child.

It is worth noting that in the ZIP model results, the signs of the terms diverge in the logit part and the Poisson part. The logit part models the probability of hospitalization (q=1-p). In the logit part, the signs of the coefficient for Female and Married are positive. This means among people who were around the age of 30, married females had a higher probability of hospitalization. In the Poisson part, the sum of the coefficients of Female, Married, and their interaction is negative. This means that among people who were around the age of 30 and were hospitalized, married females tended to have lower counts. Therefore, a two-part model provides more opportunities for hypothesis testing when the probability of an event and the event counts may be driven by different subject characteristics.

Figure 10 Observed and predicted means

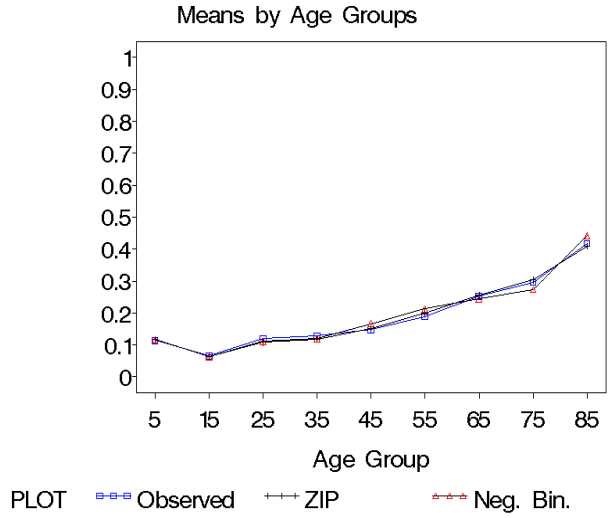


Figure 11. Observed and predicted means of unmarried females

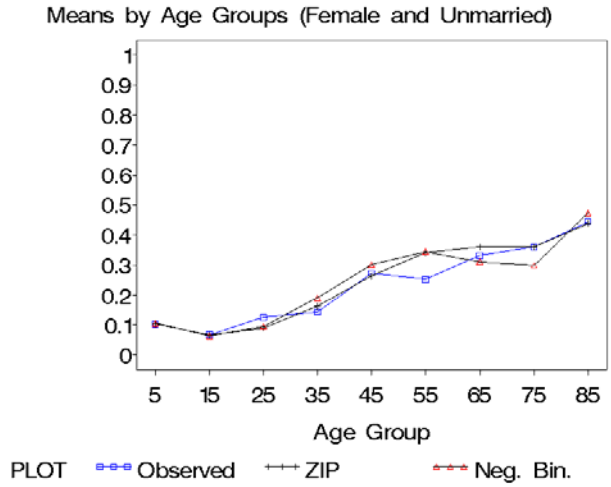


Figure 12. Observed and predicted means of married females

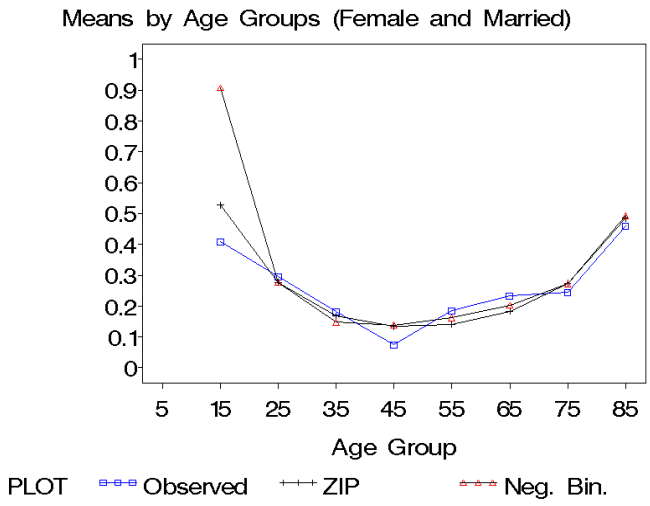


Figure 13. Observed and predicted means of unmarried males

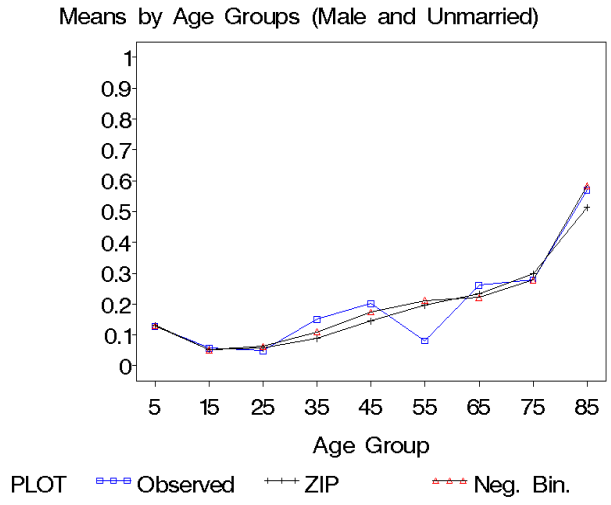


Figure 14. Observed and predicted means of married males

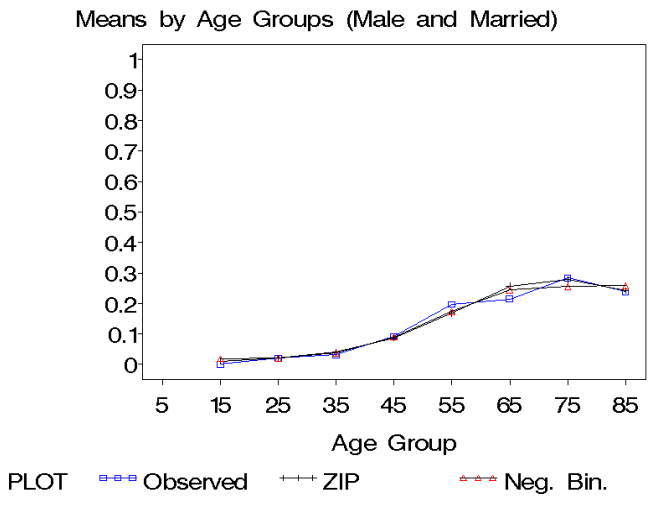


Table 6. Poisson and negative binomial model results from GLIMMIX	Poisson	Negative Binomial
-2 Log Likelihood	8912.42	7977.84
AIC (smaller is better)	8942.42	8009.84
BIC (smaller is better)	9050.75	8125.39
Pearson Chi-Square	19194.91	12363.64
Pearson Chi-Square / DF	1.9	1.22
Predictors	Poisson Estimate	Negative Binomial Estimate
Intercept	-2.5016 (0.1378)	-2.4381 (0.1233)
(Age-30)/10	0.5271 (0.0707)	0.5686 (0.0664)
Female	0.4671 (0.1578)	0.4804 (0.1483)
Married	-1.2106 (0.4490)	-1.2278 (0.3554)
Female * (Age-30)/10	0.1143 (0.0434)	0.1143 (0.0418)
Married* (Age-30)/10	0.3451 (0.3505)	0.1699 (0.3151)
Female*Married	1.494 (0.4864)	1.4186 (0.3931)
(Age-30)/10 * Female*Married	-1.4643 (0.3300)	-1.4311 (0.2861)
[(Age-30)/10]**2	0.03085 (0.0223)	0.0112 (0.0225)
Female * [(Age-30)/10]**2	-0.0447 (0.0143)	-0.0456 (0.0147)
Married * [(Age-30)/10]**2	0.1081 (0.1104)	0.2094 (0.1163)
Female*Married * [(Age-30)/10]**2	0.2541 (0.0558)	0.2510 (0.0529)
[(Age-30)/10]**3	-0.0578 (0.0074)	-0.0623 (0.0070)
Married * [(Age-30)/10]**3	-0.0292 (0.0125)	-0.0423 (0.0142)
[(Age-30)/10]**4	0.00861 (0.0014)	0.0098 (0.0015)
Scale	1.9001	4.3381 (0.3064)

Table 7. ZIP model results	Parameter Estimate
Logit Part (Pr[case follow a Poisson distribution])	
Intercept	-2.6534 (0.1612)
(Age-30)/10	0.2507 (0.0751)
Female	1.3302 (0.2288)
Married	1.0138 (0.3079)
Female * (Age-30)/10	-0.1911 (0.0885)
Married* (Age-30)/10	-1.1297 (0.2411)
[(Age-30)/10]**2	0.1654 (0.0237)
Female * [(Age-30)/10]**2	-0.1545 (0.0332)
Married * [(Age-30)/10]**2	0.2095 (0.0465)
[(Age-30)/10]**3	-0.0249 (0.0061)
Female * [(Age-30)/10]**3	0.0244 (0.0077)
Poisson Part	
Intercept	0.1000 (0.1485)
(Age-30)/10	0.2324 (0.0866)
Female	-0.6203 (0.2043)
Married	-1.8847 (0.4148)
Female * (Age-30)/10	0.3121 (0.0855)
Married* (Age-30)/10	1.2576 (0.3292)
Female*Married	1.6594 (0.4021)
(Age-30)/10 * Female*Married	-1.7129 (0.3084)
[(Age-30)/10]**2	-0.0826 (0.0251)
Female * [(Age-30)/10]**2	0.0841 (0.0286)
Married * [(Age-30)/10]**2	-0.1028 (0.0987)
Female*Married * [(Age-30)/10]**2	0.2918 (0.0533)
[(Age-30)/10]**3	-0.0263 (0.0082)
Female * [(Age-30)/10]**3	-0.0220 (0.0067)
Married * [(Age-30)/10]**3	-0.0203 (0.0111)
[(Age-30)/10]**4	0.0064 (0.0013)

CONCLUSION

When we analyze count data with over- or under-dispersion, it is useful to investigate the patterns of dispersion using exploratory data analysis. The pattern of dispersion may differ within subgroups. The four models discussed in this paper offer different ways to fit the mean and variance relation. A zero-inflated model provides additional opportunities for testing hypotheses on group heterogeneity as well as subject heterogeneity.

REFERENCE

Agresti, A. (2002), *Categorical Data Analysis*, NJ: John Wiley & Sons, Inc..

Eaton, W., Fortin, A. (1978), A Third Interpretation for the Generating Process of the Negative Binomial Distribution, American Sociological Review, 43:264-267.

Flynn, M. (2005), Fitting Extended Count Data Models to Insurance Claims, www.casact.org/education/ratesem/2005/handouts/flynn2.ppt

Long, J. Scott, (1997), *Regression Models for Categorical and Limited Dependent Variables*, CA: Sage Publications.

APPENDIX I

The NLMIXED program to fit a zero-inflated Poisson or negative bionomial model has four parts. The first part specifies the starting values. For this particular model and dataset, the convergence is not very sensitive to starting values. The default values are set to 0. The second and third parts specify the logit and the Poisson models respectively. The last part states the loglikelihood function that combines the two parts. In the program, q0 = (1-p) presents the probability of the case following the Poisson distribution. In this way, greater estimates in both parts mean higher probability or larger count in hospitalization.

```
*** ZIP Model
proc nlmixed data=nhanes.data;
    /* starting values *;
      parms  b0 0 b1 0 bp0 0 bp1 0;

    /* the logit model. Agedec is age/10 */
    eta_p = bp0 + bp1*agedec;
    q0 = exp(eta_p)/(1+exp(eta_p));

    /* the Poisson model */
    eta = b0 + b1*agedec;
      mu = exp(eta);

/* the ZIP loglikelihood function that combines two parts of the model */
    /* q0 represents the probability of the case following
a Poisson distribution*/
    if numhosp=0 then loglike = log(1-q0 + (q0)*exp(-mu));
else loglike = log(q0) + numhosp*log(mu) - mu - lgamma(numhosp+1);

 model numhosp ~ general(loglike);

/* ask NLMIXED to output the predictions for both models */
predict mu out=work.zip_eta;
predict q0 out=work.zip_etap;
run;

*** ZINB Model
proc nlmixed data=nhanes.data;

    /* the starting values */
parms  b0 0 b1 0 bp0 0 bp1 0 k 1;

    /* the logit model. Agedec = age/10 */
eta_p = bp0 + bp1*agedec;
    q0 = exp(eta_p)/(1+exp(eta_p));

    /* the negative binomial model */
    eta = b0 + b1*agedec;
      mu = exp(eta);

/* the ZINB loglikelihood function that combines the two parts of the mdoel */
    /* q0 = 1-p = the probability of a case
having the negative binomial distribution */
if numhosp=0 then loglike = log(1-q0 + q0*exp(-mu));
else  loglike = log(q0) + lgamma(numhosp + (1/k) - lgamma(numhosp+1) -
lgamma(1/k) + numhosp*log(k*mu) - (numhosp+(1/k))*log(1+k*mu));

model numhosp ~ general(loglike);
run;
```

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Adrienne Tin
Research Foundation for Mental Hygiene
105Riverside Drive
New York, NY 10032
Work Phone: 212-543-1152
E-mail: tinad@pi.cpmc.columbia.edu