SESUG Paper SD135-2018

Regression Models for Count Data

Jason Brinkley, Abt Associates

ABSTRACT

Outcomes in the form of counts are becoming an increasingly popular metric in a wide variety of fields. For example, studying the number of hospital, emergency room, or in-patient doctor's office visits has been a major focal point for many recent health studies. Many investigators want to know the impact of many different variables on these counts and help describe ways in which interventions or therapies might bring those numbers down. Traditional least squares regression was the primary mechanism for studying this type of data for decades. However, alternative methods were developed some time ago that are far superior for dealing with this type of data. The focus of this paper is to illustrate how count regression models can outperform traditional methods while utilizing the data in a more appropriate manner. Poisson Regression and Negative Binomial Regression are popular techniques when the data are overdispersed and using Zero-Inflated techniques for data with many more zeroes than is expected under traditional count regression models. These examples are applied to studies with real data.

INTRODUCTION

Outcomes in the form of counts are becoming an increasingly popular metric in a wide variety of fields. In the health sciences examples include number of hospitalizations, chronic conditions, medications, and so on. Many investigators want to know the impact of many different variables on these counts and help describe ways in which interventions or therapies might bring those numbers down. Standard methods (regression, t-tests, ANOVA) are useful for some count data studies. The methods are robust and tend to give valid results in exploring or examining associations. But many of those methods were developed to look at outcomes that run on a 'true' continuum (height, weight) or scores that run across a long range. They are not as good at handling count data where the counts do not go very high. Alternatives to the traditional modeling framework include nonparametric statistics that rank the data and help researchers look at high counts versus low counts are useful. But the methods described in this paper were specifically designed for count data.

COUNT DISTRIBUTIONS

The Poisson distribution expresses the probability of that a set number of events will occur in a fixed time or space interval. Examples include number of hurricanes in a year or location or number of calls in a call center per hour. Whereas the normal distribution is explained through the mean and standard deviation (denoted μ and σ) the Poisson distribution is denoted by one parameter λ. For Poisson data, the mean and standard deviation are the same.

An example of how the probability distributions vary with different values of λ can be given by Lavery (2010). We see $\lambda>0$ and for small values we have a skewed distributions and as the value increases, the distribution becomes more bell shaped and we sometimes approximated it with the normal distribution.

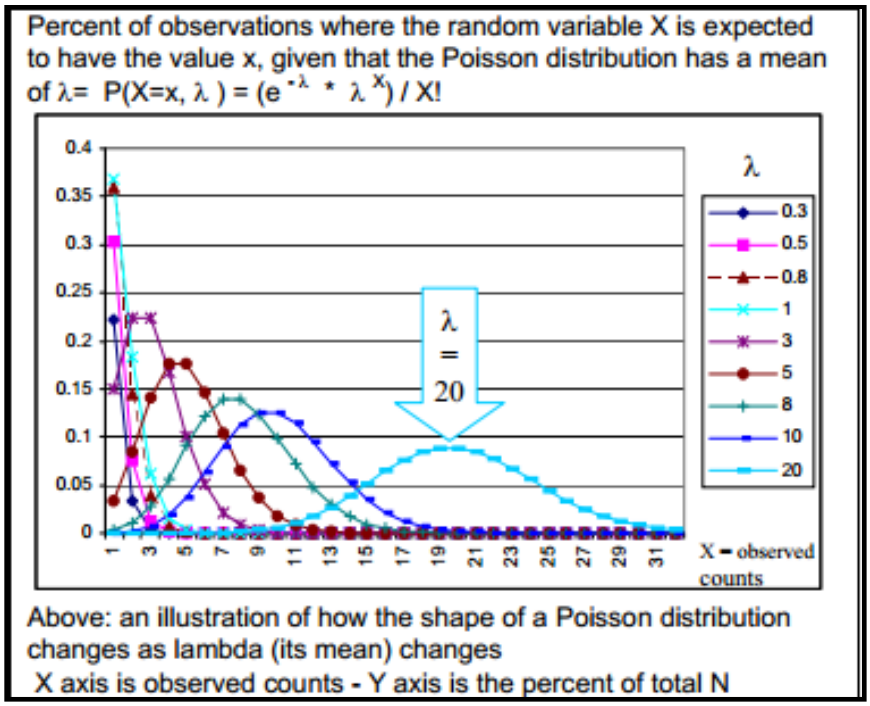

Figure 1 – Examples of Poisson Distributions

The major limitation of the Poisson distribution is that the mean and variance are equal and many practical problems with count data do not have distributions where one can reasonably make this assumption. It is usually the case that the data are 'over-dispersed' in that the variance is much larger than the mean. This happens in many count data scenarios where it is usually the case

Overdispersion is a real problem in working with count data. Most real working examples have mean and variances nowhere near the same. A common method for dealing with overdispersed Poisson data is to fit a negative binomial regression model. The negative binomial distribution is another statistical distribution for count data. The negative binomial distribution looks at the number of failures before 1 or more wins (say X failures until you win one time). The negative binomial distribution can be thought of statistically as a mixture distribution of Poisson and gamma Distributions. The reader should see Pedan (2001) for a review of the theory of negative binomial distributions.

GENERALIZED LINEAR MODELS AND ZERO INFLATION

Regression models are among the best frameworks for exploring the association between multiple variables and an outcome of interest. Generalized linear models (GLM) provide the framework by which regression models can capitalize on the use of alternative distributions such as the Poisson and negative binomial distributions in outcome modeling. They rely on specifying a distribution of interest and a *link* function (usually specified as a log-link for count regression models) to aid in modeling the data. The interested reader should see Ngo (2016) for an overview of generalized linear models.

Finally, generalized linear models can be augmented to account for so-called zero-inflation which can occur in count data where there are more zero values in the data than would occur naturally in count distributed data. Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) models exist as zero inflated analogs for the generalized linear models that use the Poisson and negative binomial distributions for outcomes (respectively). Zero-Inflated models in SAS® take a two stage modeling approach where the data are first modeled for the probability of a zero value to explain the extra zeroes in the data and a second stage model where the data are modeled using a count regression. Erdman (2008) has a soft introduction into the theory of zero-inflated count models.

The task here is to help the reader elucidate which model should be used in which scenario. For that we ask a series of questions based on the distribution of our count outcome. In scenarios where the count

data has a large mean and the data 'look' bell shaped, then Poisson Regression (Poisson distribution based GLM) work well.

In cases where the data are skewed (Poisson data that has not converged to Normal distributions) then Poisson Regression may still be viable. Here, two additional questions are asked; first if the variance is larger than the mean, then the data are likely over-dispersed and Negative Binomial Regression (negative binomial based GLM) may work well. If the data have many zero values then ZIP should be considered. In cases where the data have many zero values and a long tail (high variance) then the ZINB may be your best fit.

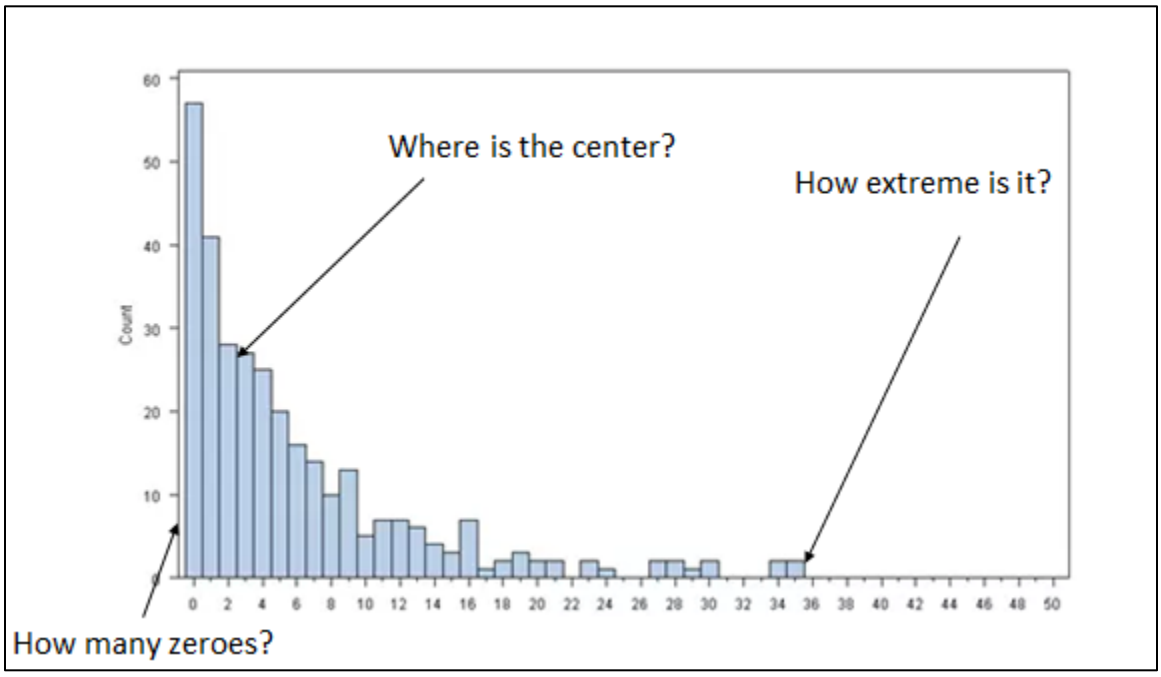

Figure 2 – Determining Count Regression Model

EXAMPLE - AFFAIRS

Fair (1978) did a study on extramarital affairs. Suppose we want to examine the impact of children, religiousity, happiness, and time in marriage on the number of admitted marital affairs. The main outcome (NAFFAIRS) has the following distribution and summary statistics:

Moments			
N	601	Sum Weights	601
Mean	1.45590682	Sum Observations	875
Std Deviation	3.29875773	Variance	10.8818026
Skewness	2.34699789	Kurtosis	4.25688176
Uncorrected SS	7803	Corrected SS	6529.08153
Coeff Variation	226.577531	Std Error Mean	0.13455913

Figure 3 – Univariate Output

The skew in this data illustrates that the data does not run over the typical range of normal type data. There are an unusually high number of observations that are in the seven and 12 count columns. Count

regression models are good at capitalizing on things like this. Note that the mean here is 1.45 and the standard deviation is 3.3 (variance=10.9).

We will focus our analysis on the impact of marriage rating (RATEMARR) on expected number of affairs. We start the analysis by comparing SAS code (here the dataset is called sample) using the same covariates but fitting ordinary least squares (OLS or traditional regression) to Poisson regression GLM.

```
*traditional regression estimates;
proc glm data=sample;
class MALE RATEMARR KIDS RELIG;
model NAffairs = YRSMARR MALE RATEMARR KIDS RELIG/solution;
lsmeans RATEMARR/cl;
run;

*Poisson regression estimates;
proc genmod data=sample;
class MALE RATEMARR KIDS RELIG;
model NAffairs = YRSMARR MALE RATEMARR KIDS RELIG/dist=poisson;
lsmeans RATEMARR/cl;
run;
```

Figure 4 – SAS Code for OLS and Poisson Regression

We use the LS means statement to provide least square means as a way to compare model based estimates for the expected number of affairs for each marriage rating after adjusting for years of marriage, gender, whether there are children, and religiousity. Figure 5 below shows partial GLM output of the OLS model.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
YRSMARR	1	140.0209330	140.0209330	14.84	0.0001
MALE	1	1.2456570	1.2456570	0.13	0.7164
RATEMARR	4	452.6359242	113.1589810	12.00	<.0001
KIDS	1	3.8877359	3.8877359	0.41	0.5211
RELIG	4	241.4478853	60.3619713	6.40	<.0001

RATEMARR	NAFFAIRS LSMEAN	95% Confidence Limits	
1	3.591451	2.056039	5.126862
2	3.927564	3.131669	4.723460
3	1.618354	0.956928	2.279779
4	1.435747	0.948738	1.922755
5	1.097238	0.660390	1.534085

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
YRSMARR	1	101.09	<.0001
MALE	1	1.35	0.2458
RATEMARR	4	231.42	<.0001
KIDS	1	0.26	0.6099
RELIG	4	174.35	<.0001

RATEMARR Least Squares Means

	Log Estimate	Log Lower	Log Upper	Estimate	Lower	Upper
1	1.0275	0.7633	1.2917	2.794072	2.14534	3.63897
2	1.193	1.0427	1.3432	3.296957	2.83687	3.83128
3	0.295	0.1104	0.4795	1.343126	1.11672	1.61527
4	0.1502	0.00418	0.2961	1.162067	1.00419	1.3446
5	-0.244	-0.4018	-0.0862	0.783488	0.66911	0.91746

Figure 5 – Partial OLS and Poisson Regression Output

We see that the models have very different output in terms of the p-values for type 3 (last variable in the model) and we see different estimates for least square means. There are also smaller confidence intervals for the Poisson model. However, figure 3 suggested that the Poisson Regression output may not be the best fit given the difference in the mean (1.46) and variance (10.88). We compare fit using AIC below. First, we explore a negative binomial fit with code listed in Figure 6.

```
*NB regression estimates;
proc genmod data=sample;
class MALE RATEMARR KIDS RELIG;
model NAffairs = YRSMARR MALE RATEMARR KIDS RELIG/dist=nb;
lsmeans RATEMARR/cl;
run;
```

Figure 6 – Negative Binomial Regression SAS Code

We transform the output into what we see in Figure 7 below. Note that the big difference in confidence limits

RATEMARR Least Squares Means

	Estimate	Lower	Upper
1	2.88752577	0.73956	11.274
2	3.01078246	1.48795	6.09214
3	1.06932756	0.57764	1.97941
4	1.07071859	0.66074	1.73499
5	0.71720039	0.47783	1.07654

Figure 7 – Transformed Least Square Means

So how do we ascertain the 'best' option for modeling? Goodness of fit measures such as Akaike's Information Criterion (AIC) provide a reasonable bell-weather for which model would be preferable. In scenarios where the same model is fit with different assumptions (here OLS versus Poisson versus NB) one can look at AIC values to determine 'best' fit. Figure 8 below has AIC values for all three models, it is clear that the OLS model has better fit (AIC = 1963) versus Poisson Regression (AIC = 2852). However, the negative binomial model was fit and we see a reduction of AIC around 25% (from 1963 to 1478).

OLS

Root MSE	3.07123
Dependent Mean	1.45591
R-Square	0.1491
Adj R-Sq	0.1332
AIC	1963.61475
AICC	1964.23485
SBC	1413.39789

Poisson

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	589	2334.9475	3.9643
Scaled Deviance	589	2334.9475	3.9643
Pearson Chi-Square	589	3950.2486	6.7067
Scaled Pearson X2	589	3950.2486	6.7067
Log Likelihood		-251.0708	
Full Log Likelihood		-1414.4687	
AIC (smaller is better)		2852.9374	
AICC (smaller is better)		2853.4680	
BIC (smaller is better)		2905.7206	

NB

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	589	340.3148	0.5778
Scaled Deviance	589	340.3148	0.5778
Pearson Chi-Square	589	574.4757	0.9753
Scaled Pearson X2	589	574.4757	0.9753
Log Likelihood		436.9640	
Full Log Likelihood		-726.4339	
AIC (smaller is better)		1478.8678	
AICC (smaller is better)		1479.4879	
BIC (smaller is better)		1536.0495	

Exhibit 8 – Affair Model Fit Statistics

In addition, ZINB models were fit to this data and the corresponding AIC was about 1432. So while ZINB does have the 'best' fit, the gains in AIC may not be enough to actually deploy a zero-inflated model in practice. Figure 9 does show code and corresponding output that illustrates a different interpretation from such a model. However, there should be a little hesitation in direct comparison of ZINB and negative binomial models given the additional parameters introduced in a two stage model.

```
*Zero Inflated NB regression estimates;
proc genmod data=sample;
class MALE RATEMARR KIDS RELIG;
model NAffairs = YRSMARR MALE RATEMARR KIDS RELIG/dist=zinb;
zeromodel kids relig ratemarr;
lsmeans RATEMARR/cl;
run;
```

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
YRSMARR	1	12.39	0.0004
MALE	1	0.97	0.3237
RATEMARR	4	7.20	0.1258
KIDS	1	1.06	0.3041
RELIG	4	8.34	0.0800

LR Statistics For Type 3 Analysis of Zero Inflation Model			
Source	DF	Chi-Square	Pr > ChiSq
KIDS	1	5.42	0.0199
RELIG	4	18.49	0.0010
RATEMARR	4	26.35	<.0001

Figure 9 – Zero-Inflated Negative Binomial Code and Select Output

EXAMPLE - NEEDLESTICKS

Mann, Larsen, and Brinkley (2014) looked at negative binomial regression as a way to model pediatric IV stick attempts. The process is actually a negative binomial distribution (count attempts to start an IV until a success). Finding such a process is rare in the medical literature and figure 10 below shows that the data match up well with a theoretical negative binomial counts.

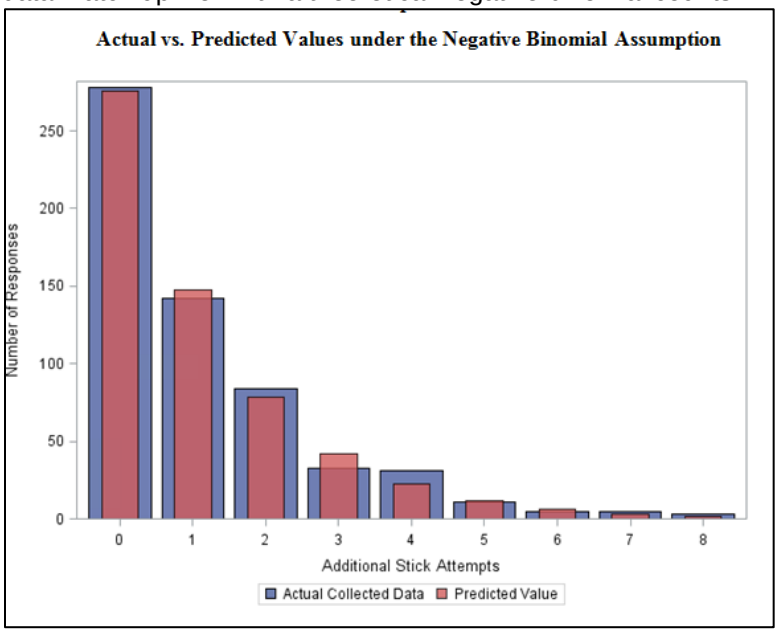

Figure 10 – Negative Binomial Actual Versus Predicted Values

Figure 11 below compares type 3 tests from an OLS model (AIC = 1932) and a Negative Binomial Regression model (AIC = 1533). Clearly, the p-values and interpretation of potential model effects differ

between the use of OLS and NB fits. It is important to note that ZINB were also fit with little change in the results.

OLS Output

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
SHIFT	1	547	13.30	0.0003
DIFF1	2	547	13.45	<.0001
Dehydrated	1	547	28.69	<.0001
COOPCH1	1	547	12.37	0.0005
Nurse1Exp	1	547	11.82	0.0006
OSBDM	1	547	3.08	0.0798

Neg Binomial Output

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
SHIFT	1	15.01	0.0001
DIFF1	2	24.09	<.0001
Dehydrated	1	20.75	<.0001
COOPCH1	1	12.73	0.0004
Nurse1Exp	1	10.16	0.0014
OSBDM	1	1.36	0.2429

Figure 11 – Comparing OLS and Negative Binomial Regression Output

CONCLUSION

There are many different models for count regression data and it is not always clear exactly which model is best for each scenario. It may be that a combination of univariate distribution analysis combined with multiple model fits (and comparing goodness of fit measures) is a viable avenue in exploring which count regression model is best equipped for exploring each specific data set. Analysts should be aware of the differences in these models and that one should explore multiple fits before deciding on which one is the most appropriate for interpretation.

REFERENCES

Lavery, R. 2010 “An Animated Guide: An Introduction to Poisson Regression”. Northeast SAS User Group Meetings. Available at https://www.lexjansen.com/nesug/nesug10/sa/sa04.pdf.

Hilbe, J. M. (2011). Negative binomial regression. 2nd edition. Cambridge, UK: Cambridge University Press.

Pedan A. (2001). Analysis of Count Data Using the SAS® System. SUGI 26 Meetings. Available at http://www2.sas.com/proceedings/sugi26/p247-26.pdf.

Ngo T.H.D. (2016). Generalized Linear Models for Non-Normal Data. SAS Global Forum 2016 Proceedings. Available at https://support.sas.com/resources/papers/proceedings16/8380-2016.pdf.

Erdman D., Jackson L., Sinko A. (2008). Zero-Inflated Poisson and Zero-Inflated Negative Binomial Models Using the COUNTREG Procedure. SAS Global Forum 2008 Meetings. Available at http://www2.sas.com/proceedings/forum2008/322-2008.pdf.

Fair, R.C. (1978). A Theory of Extramarital Affairs. Journal of Political Economy, 86, 45–61.

Mann J, Larsen P , Brinkley J . (2014) Exploring the use of negative binomial regression modeling for pediatric peripheral intravenous catheterization. J Med Stat Inform 2:6.doi:10.7243/2053-7662-2-6

CONTACT INFORMATION

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

Jason Brinkley
Senior Associate
919-294-7745
Jason_Brinkley@abtassoc.com