Regression Models for Count Data

Jason Brinkley, PhD
Abt Global

Reference

SESUG Paper SD296-2019

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.

https://www.lexjansen.com/sesug/2019/SESUG2019_Paper-296_Final_PDF.pdf

Analysis of Count Data

- Outcomes in the form of counts are becoming an increasingly popular metric in a wide variety of fields.
- Examples: Number of hospitalizations, chronic conditions, medications, etc.
- 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.

Why does this matter?

- Standard methods (regression, t-tests, ANOVA) are ‘ok’ 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

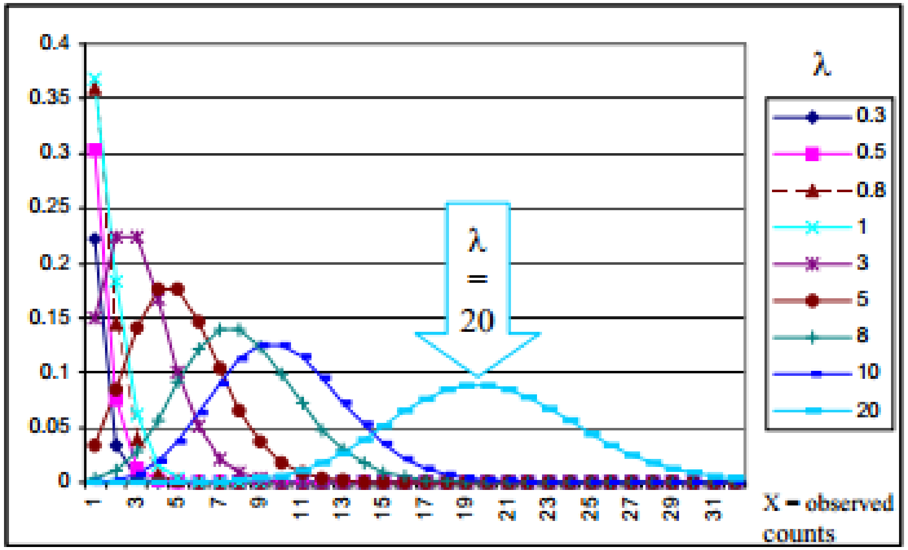
- Nonparametric statistics that rank the data and help researchers look at high counts versus low counts have born fruit.
- But really, we want to look at methods that are designed for and utilize many different facets of count data.

Count Data 101 - 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 variance are the same.**
- If λ is large, then the data 'looks' like normal data and we sometimes approximated it with the normal distribution.

Closer Look – Some Examples

Percent of observations where the random variable X is expected to have the value x, given that the Poisson distribution has a mean of λ= P(X=x, λ) = (e $^{-\lambda}$ * λ ^{X}) / X!



Above: an illustration of how the shape of a Poisson distribution changes as lambda (its mean) changes
X axis is observed counts - Y axis is the percent of total N
FIGURE 1

(from http://www.nesug.org/Proceedings/nesug10/sa/sa04.pdf)

The Secret Sauce – GLM Models

- GLM is the term for the big umbrella framework that encompasses many of the types of regression that we already know about.
- OLS is a type of GLM, but not all GLM are OLS.
- Ordinal regression, logistic regression, probit regression, multinomial regression, tobit regression, and many more follow under the GLM framework.

Example - Affairs

- Fair (1978) did a study on extramarital affairs.
- Suppose we want to examine the impact of children, religiosity, happiness, and time in marriage on the number of admitted marital affairs.
- Let’s start by looking at the variables of interest.

Descriptive Stats - Predictors

=1 if male				
MALE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	315	52.41	315	52.41
1	286	47.59	601	100.00

=1 if have kids				
KIDS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	171	28.45	171	28.45
1	430	71.55	601	100.00

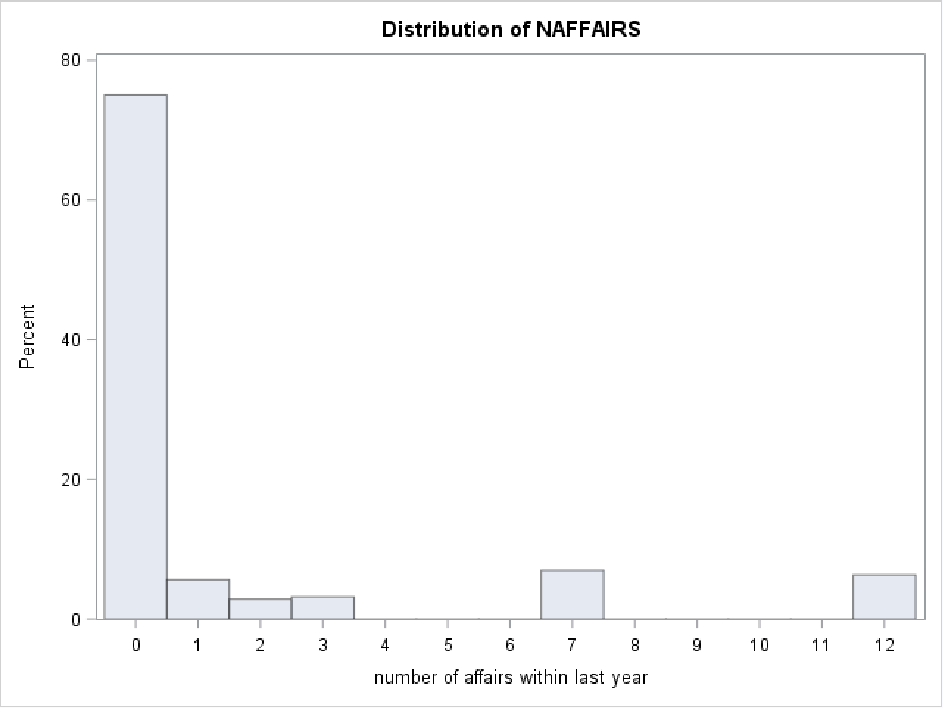
5 = very relig., 4 = somewhat, 3 = slightly, 2 = not at all, 1 = anti				
RELIG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	48	7.99	48	7.99
2	164	27.29	212	35.27
3	129	21.46	341	56.74
4	190	31.61	531	88.35
5	70	11.65	601	100.00

5 = vry hap marr, 4 = hap than avg, 3 = avg, 2 = smewht unhap, 1 = vry unhap				
RATEMARR	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	16	2.66	16	2.66
2	66	10.98	82	13.64
3	93	15.47	175	29.12
4	194	32.28	369	61.40
5	232	38.60	601	100.00

The UNIVARIATE Procedure
Variable: YRSMARR (years married)

Moments			
N	601	Sum Weights	601
Mean	8.17769551	Sum Observations	4914.795
Std Deviation	5.57130315	Variance	31.0394188
Skewness	0.0781888	Kurtosis	-1.5705532
Uncorrected SS	58815.3483	Corrected SS	18623.6513
Coeff Variation	68.1280337	Std Error Mean	0.2272582

Outcome – Number of Affairs



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

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). But the median is 0 (75% of the data is 0 values).

SAS Code

```
*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;
```

Traditional Regression Output

The GLM Procedure

Dependent Variable: NAFFAIRS number of affairs within last year

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	973.382446	88.489313	9.38	<.0001
Error	589	5555.699085	9.432426		
Corrected Total	600	6529.081531			

R-Square	Coeff Var	Root MSE	NAFFAIRS Mean
0.149084	210.9493	3.071226	1.455907

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

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-0.744126863	B	0.52518740	-1.42	0.1570
YRSMARR	0.110057104		0.02856494	3.85	0.0001
MALE 0	-0.092117114	B	0.25348531	-0.36	0.7164
MALE 1	0.000000000	B	.	.	.
RATEMARR 1	2.494212836	B	0.80973589	3.08	0.0022
RATEMARR 2	2.830326689	B	0.44277919	6.39	<.0001
RATEMARR 3	0.521116136	B	0.38224923	1.36	0.1733
RATEMARR 4	0.338508882	B	0.30765518	1.10	0.2717
RATEMARR 5	0.000000000	B	.	.	.
KIDS 0	0.220142951	B	0.34290050	0.64	0.5211
KIDS 1	0.000000000	B	.	.	.
RELIG 1	2.125430699	B	0.58212824	3.65	0.0003
RELIG 2	0.949478200	B	0.44908542	2.11	0.0349
RELIG 3	1.231680995	B	0.46175214	2.67	0.0079
RELIG 4	0.080100782	B	0.43057534	0.19	0.8525
RELIG 5	0.000000000	B	.	.	.

Traditional Regression Output

The GLM Procedure

Dependent Variable: NAFFAIRS number of affairs within last year

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	973.382446	88.489313	9.38	<.0001
Error	589	5555.699085	9.432426		
Corrected Total	600	6529.081531			

R-Square	Coeff Var	Root MSE	NAFFAIRS Mean
0.149084	210.9493	3.071226	1.455907

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

Estimated Effect of Rate Marriage?

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

SAS outputs 'LS Means' as model adjusted means. Adjusted for all the other variables in the model. Here we are saying that this model suggests that everything else being held constant, people who rate their marriage as 'very unhappy' had an average of 3.6 affairs in the last 12 months.

Poisson Regression Output

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.4885	0.1762	-1.8339	-1.1431	71.34	<.0001
YRSMARR		1	0.0771	0.0078	0.0618	0.0924	97.33	<.0001
MALE	0	1	-0.0791	0.0681	-0.2125	0.0544	1.35	0.2456
MALE	1	0	0.0000	0.0000	0.0000	0.0000		.
RATEMARR	1	1	1.2715	0.1537	0.9702	1.5727	68.44	<.0001
RATEMARR	2	1	1.4369	0.1037	1.2336	1.6402	191.91	<.0001
RATEMARR	3	1	0.5389	0.1182	0.3073	0.7705	20.80	<.0001
RATEMARR	4	1	0.3941	0.1024	0.1934	0.5948	14.81	0.0001
RATEMARR	5	0	0.0000	0.0000	0.0000	0.0000		.
KIDS	0	1	0.0530	0.1035	-0.1499	0.2559	0.26	0.6089
KIDS	1	0	0.0000	0.0000	0.0000	0.0000		.
RELIG	1	1	1.4343	0.1578	1.1251	1.7435	82.65	<.0001
RELIG	2	1	0.7734	0.1433	0.4926	1.0542	29.13	<.0001
RELIG	3	1	0.8856	0.1428	0.6058	1.1654	38.47	<.0001
RELIG	4	1	0.0424	0.1494	-0.2505	0.3352	0.08	0.7768
RELIG	5	0	0.0000	0.0000	0.0000	0.0000		.
Scale		0	1.0000	0.0000	1.0000	1.0000		

Least Square Means

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

Fitting a count regression model to count data takes advantage of distribution ideas and their impact on variance calculations. In essence, we tell SAS that this is count data and that extra insight helps us get smaller standard errors than traditional OLS models (even with a log transformation).

Mousetraps

- The whole point of this early discussion is to illustrate that Poisson regression modeling is a better mousetrap for this kind of count data than traditional OLS models.
- But Poisson regression utilizes the Poisson distribution. How do we know if/when that is appropriate?
- More importantly, when do we know if something is wrong?

Variation is the key!

- Poisson distribution assumes mean and variance (st dev squared) are the same.
- If variance and mean are different than Poisson may not be the best fit.
- **Affairs data: mean is 1.44, var is 10.88!**
- That doesn't mean we abandon Count Regression.

Overdispersion

- Overdispersion is a real problem in working with count data.
- Most real working examples have mean and variances nowhere near the same. Which means that while Poisson regression is doing better than OLS, in some cases we could be doing better.
- However, we do have an answer!

Negative Binomial Regression

- 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 (only important for the mathematics).

Neg Binomial Output

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.8362	0.5394	-2.8934	-0.7790	11.59	0.0007
YRSMARR		1	0.0848	0.0269	0.0321	0.1376	9.94	0.0016
MALE	0	1	-0.0115	0.2429	-0.4876	0.4645	0.00	0.9622
MALE	1	0	0.0000	0.0000	0.0000	0.0000	.	.
RATEMARR	1	1	1.3928	0.7235	-0.0253	2.8108	3.71	0.0542
RATEMARR	2	1	1.4346	0.4040	0.6428	2.2264	12.61	0.0004
RATEMARR	3	1	0.3994	0.3669	-0.3196	1.1185	1.19	0.2763
RATEMARR	4	1	0.4007	0.3096	-0.2061	1.0076	1.67	0.1956
RATEMARR	5	0	0.0000	0.0000	0.0000	0.0000	.	.
KIDS	0	1	-0.1005	0.3085	-0.7051	0.5041	0.11	0.7446
KIDS	1	0	0.0000	0.0000	0.0000	0.0000	.	.
RELIG	1	1	1.6502	0.5422	0.5874	2.7129	9.26	0.0023
RELIG	2	1	1.2506	0.4446	0.3792	2.1220	7.91	0.0049
RELIG	3	1	1.1606	0.4554	0.2681	2.0531	6.50	0.0108
RELIG	4	1	0.2694	0.4202	-0.5542	1.0930	0.41	0.5215
RELIG	5	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion		1	6.8821	0.7737	5.5211	8.5787		

Neg Binomial Code

```
*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;
```

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

Comparing and Assessing Fit

- The measures for assessing model fit are not the same in these models as in traditional OLS models.
- There is no ANOVA table and we aren't doing 'sum of squares' so a measure like R-Square doesn't really apply. We have multiple measures of model fit, most of them centered on the notion 'does this model fit the data well?'
- One popular measure for that is called Akaike's Information Criterion (AIC) which assess how well the regression model is using the information provided in the data. Smaller values are better.

Comparing OLS, Poisson, and Neg Binomial AIC

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	

Zero-Inflation

- The question does make sense but let's reframe it in the context of the traditional problem with count models.
- Count data has a lower bound (you don't get negative counts), so the data could have a lower than expected spread if the data comes with an unusually high amount of zeroes.

Zero-Inflated Poisson Models (ZIP)

- The zero-inflated Poisson (ZIP) model is a recent innovation which adjusted Poisson regression models to account for instances where one has more zeroes than would fit under the traditional Poisson Regression Model.

From the Institute for Digital Research and Education

- “Zero-inflated Poisson regression is used to model count data that has an excess of zero counts. Further, theory suggests that the excess zeroes are generated by a separate process from the count values and that the excess zeroes can be modeled independently. Thus, the **ZIP** model has two parts, a Poisson count model and the logit model for predicting excess zeroes.”

Zero-Inflated Negative Binomial

- I can put the idea of zero-inflation and overdispersion together to get the Zero-Inflated Negative Binomial Model!
- It is the best of both worlds when I have a long tail AND a lot of zeroes.

Zero Inflated Negative Binomial

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

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance		1385.7973	
Scaled Deviance		1385.7973	
Pearson Chi-Square	579	632.0970	1.0917
Scaled Pearson X2	579	632.0970	1.0917
Log Likelihood		-692.8986	
Full Log Likelihood		-692.8986	
AIC (smaller is better)		1431.7973	
AICC (smaller is better)		1433.7106	
BIC (smaller is better)		1532.9649	

```
*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;
```

Diminishing Returns

- Think of using AIC the same way you use R-Square.
- Lower AIC is better fit of model to data, but you want to balance complexity with reduction.
- NB model shows that overdispersion is a real problem. But ZIP shows that zero counts are a real problem.
- Sometimes it is about illustrating the right conclusions. ZINB suggests that some previously significant factors are not significant. This causes adjustments in clinical conclusions.

Other types of Underdispersion

- There are certainly instances of other types of underdispersion then high zero counts.
- And if you have true Poisson data with mean and variance 1 then you will have a lot of zero counts which means the data is not underdispersed.
- I have not seen any practical examples of underdispersed data in the same way we see overdispersed data.

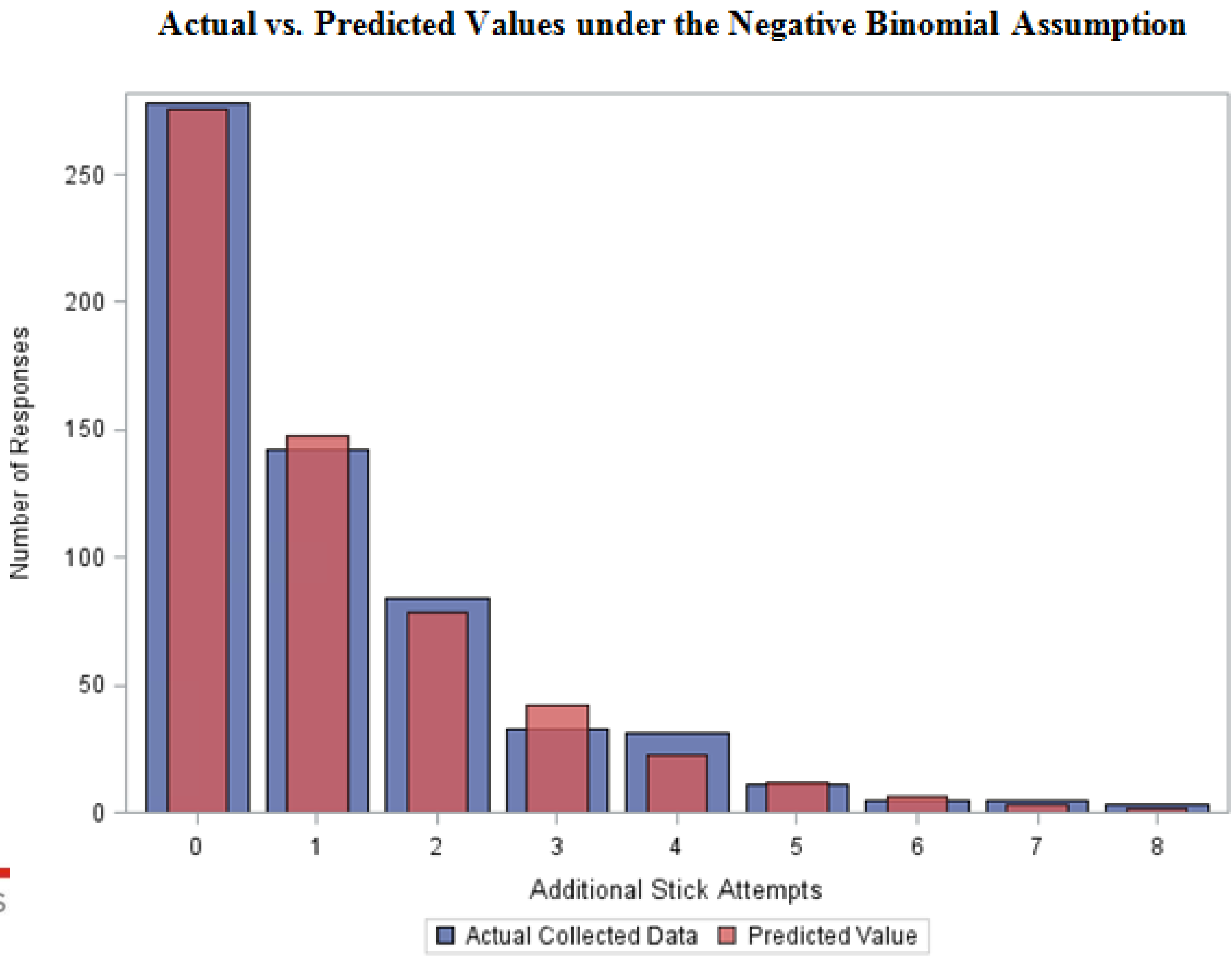
Another Example

- 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).
- Rare to find that in medical literature.

Variables of interest

- Number of additional stick attempts – ranges from 0 to 8 and represents the number of attempts beyond the first.
- Shift – Day or Night Shift
- Diff1 – Nurse assessment if child will be difficult stick on first attempt (yes/no)
- Dehydrated – Child is dehydrated
- Coop1 – Nurse assessment if child is cooperative on first attempt (yes/no)
- Nurse1Exp – Is the nurse who tries to start the IV first a novice or junior (1 year experience or less)
- OSBDM – Mean score of patients reaction to painful procedure across all attempts. (Operational Scale of Behavior Distress)

The Data


Actual vs. Predicted Values under the Negative Binomial Assumption
Number of Responses
250
200
150
100
50
0
0
1
2
3
4
5
6
7
8
Additional Stick Attempts
Actual Collected Data
Predicted Value

OLS Model Output

Fit Statistics	
-2 Res Log Likelihood	1929.7
AIC (Smaller is Better)	1931.7
AICC (Smaller is Better)	1931.7
BIC (Smaller is Better)	1936.0

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

NB Output

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	547	575.3659	1.0519
Scaled Deviance	547	575.3659	1.0519
Pearson Chi-Square	547	581.7581	1.0635
Scaled Pearson X2	547	581.7581	1.0635
Log Likelihood		-436.2602	
Full Log Likelihood		-757.2393	
AIC (smaller is better)		1532.4785	
AICC (smaller is better)		1532.8088	
BIC (smaller is better)		1571.3492	

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

Zero-Inflated NB Output

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance		1504.9988	
Scaled Deviance		1504.9988	
Pearson Chi-Square	539	579.6854	1.0755
Scaled Pearson X2	539	579.6854	1.0755
Log Likelihood		-752.4994	
Full Log Likelihood		-752.4994	
AIC (smaller is better)		1538.9988	
AICC (smaller is better)		1540.1385	
BIC (smaller is better)		1612.4213	

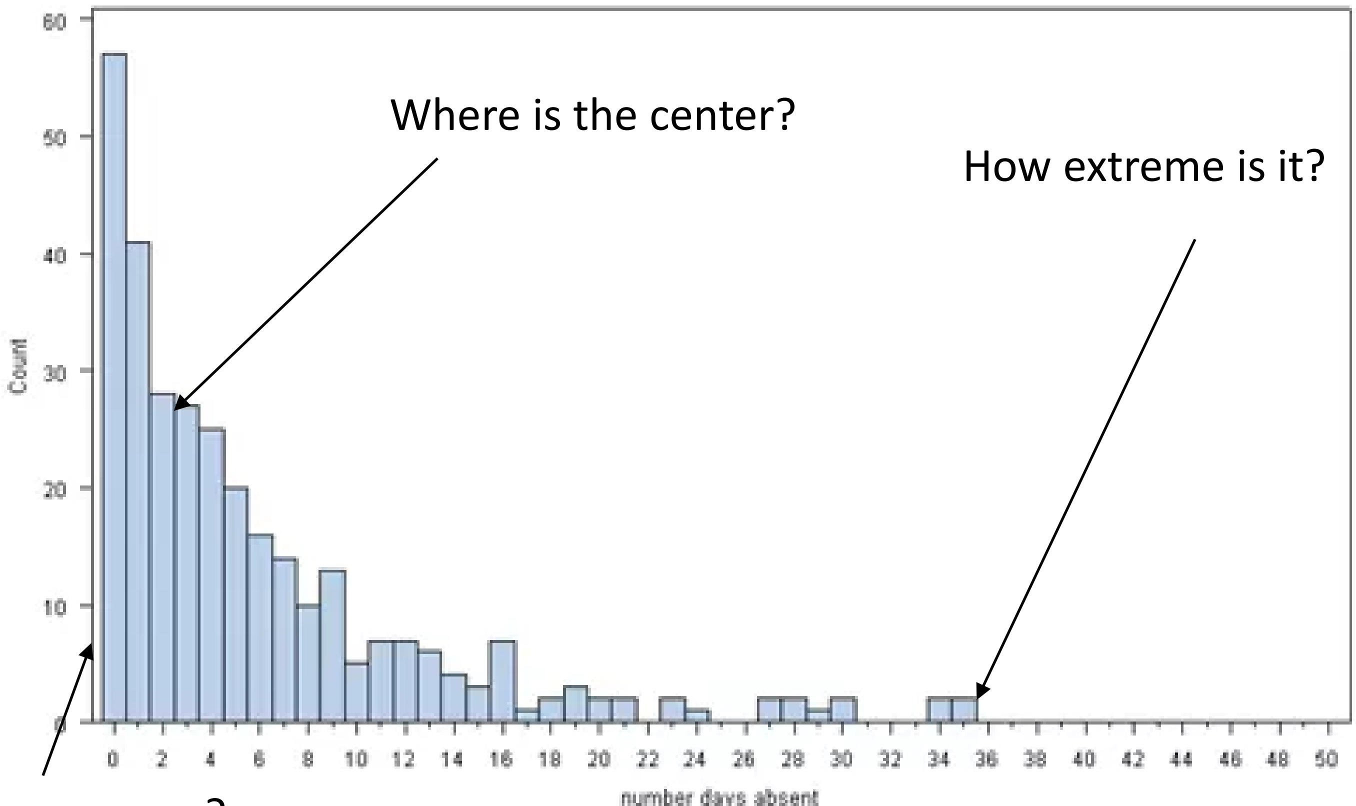
LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
SHIFT	1	3.14	0.0762
DIFF1	2	11.87	0.0026
Dehydrated	1	14.33	0.0002
COOPCH1	1	3.55	0.0594
Nurse1Exp	1	4.87	0.0274
OSBDM	1	2.14	0.1431

LR Statistics For Type 3 Analysis of Zero Inflation Model			
Source	DF	Chi-Square	Pr > ChiSq
SHIFT	1	1.22	0.2695
DIFF1	2	0.49	0.7826
Dehydrated	1	0.07	0.7890
COOPCH1	1	0.00	1.0000
Nurse1Exp	1	0.06	0.8061
OSBDM	1	0.30	0.5827

Dispersion		1	0.2420	0.1189	0.0924	0.6341	

In Summary: What am I looking for?

https://stats.idre.ucla.edu/sas/dae/negative-binomial-regression/



How many zeroes?

Questions?

- Email – Jason.Brinkley@AbtGlobal.com
- X/Twitter - @DrJasonBrinkley

SESUG 2024!

SESUG 2024 Conference

We are so excited to have our in person conference in Bethesda, Maryland in Sepember of 2024!
SESUG registration for the 2024 Conference is now open!

Register Now! **<-- Don't delay, early registration closes July 19!**

The Call for Abstracts for SESUG 2024 is currently open! You can submit a proposal here:

Paper Submissions (START)

What's New for SESUG 2024

- Featured Paper Competition
- New and Revised Academic Sections
- Revised Grant Application and Perks
- Quick Hit Presentations–which do not require a paper submission
- Expanded Networking Opportunities

SESUG 2024 will be held at the North Bethesda Marriott!

September 22-24, 2024