

## Regression Models for Count Data

### Overview

A good example of the adaptation of the regression model for a variable with a particular distribution (i.e., the generalized linear model) is the modeling of count data. Whenever a measure is a count of something (e.g., number of cars passing, frequency of drug use, number of walking trips), the dependent variable and therefore the residuals tend to be non-normal (often, but not always, there is a high frequency of 0s or low values and a low frequency of higher values). Use of the Poisson link function is designed for this type of count data (Coxe, West, & Aiken, 2009). The Poisson model assumes that the conditional mean and variance of the outcome are approximately equal (i.e., mean and variance taking into account the covariates in the model). When the conditional variance exceeds the conditional mean, which frequently occurs in practice, it is referred to as *overdispersion*. This may bias standard errors and thus statistical tests. The negative binomial model is a related approach but does not require the equal conditional variance and mean, allowing for overdispersion without bias in standard error estimates. When there is no overdispersion, the negative binomial and Poisson are the same. Variants, called zero-inflated models, exist for both types of count models when there are many zero values (see Long, 1997 for additional details).

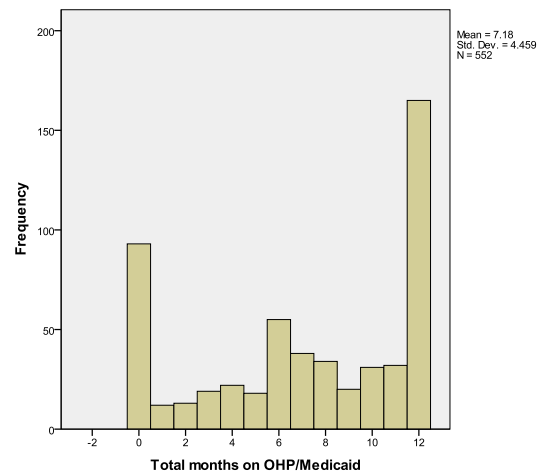
### Poisson Model Example

This example also comes from Karen Seccombe's project focusing on healthcare among welfare recipients in Oregon. The outcome variable is the number of months over a year that respondents were covered by the Oregon Health Plan. Because this is a count of the number of months, a regression model developed to take into account the distributional characteristics of this type of data is most appropriate.

I first did a quick check on the overdispersion issue by examining a frequency histogram and estimating the unconditional variance and mean.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
mosmed Total months on OHP/Medicaid	552	0	12	7.18	4.459	19.882
Valid N (listwise)	552					



These results suggest that the mean and variance are not near equal. This is not the optimal way to investigate overdispersion, since they are unconditional values. One suggested test is to compare the likelihood ratio of the Poisson and the negative binomial models, because they are equivalent when the equal dispersion assumption is met. Just to illustrate its use, I test the Poisson regression model below even though this distribution does not look appropriate for Poisson.

### Poisson Model

```
genlin mosmed with income educat marital depress1
    /model income educat marital depress1 distribution=poisson link=log.
```

Goodness of Fit<sup>a</sup>

	Value	df	Value/df
Deviance	1345.360	353	3.811
Scaled Deviance	1345.360	353	
Pearson Chi-Square	964.522	353	2.732
Scaled Pearson Chi-Square	964.522	353	
Log Likelihood <sup>a</sup>	-1255.992		
Akaike's Information Criterion (AIC)	2521.984		
Finite Sample Corrected AIC (AICC)	2522.154		
Bayesian Information Criterion (BIC)	2541.386		
Consistent AIC (CAIC)	2546.386		

Dependent Variable: Total months on OHP/Medicaid  
Model: (Intercept), income, educat, marital, depress1

a. The full log likelihood function is displayed and used in computing information criteria.  
b. Information criteria are in small-is-better form.

Omnibus Test<sup>a</sup>

Likelihood Ratio Chi-Square	df	Sig.
41.780	4	.000

Dependent Variable: Total months on OHP/Medicaid  
Model: (Intercept), income, educat, marital, depress1

a. Compares the fitted model against the intercept-only model.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	2.059	.0631	1.935	2.182	1064.235	1	.000
income	-5.712E-6	2.5587E-6	-1.073E-5	-6.973E-7	4.984	1	.026
educat	-.016	.0219	-.059	.027	.515	1	.473
marital	-.193	.0498	-.290	-.095	14.969	1	.000
depress1	.028	.0077	.013	.043	13.076	1	.000
(Scale)	1 <sup>a</sup>						

Dependent Variable: Total months on OHP/Medicaid  
Model: (Intercept), income, educat, marital, depress1

a. Fixed at the displayed value.

## Negative Binomial Model

```
genlin mosmed with income educat marital depress1
/model income educat marital depress1 distribution=negbin link=log.
```

Goodness of Fit<sup>a</sup>

	Value	df	Value/df
Deviance	312.822	353	.886
Scaled Deviance	312.822	353	
Pearson Chi-Square	122.929	353	.348
Scaled Pearson Chi-Square	122.929	353	
Log Likelihood <sup>a</sup>	-1086.792		
Akaike's Information Criterion (AIC)	2183.584		
Finite Sample Corrected AIC (AICC)	2183.755		
Bayesian Information Criterion (BIC)	2202.987		
Consistent AIC (CAIC)	2207.987		

Dependent Variable: Total months on OHP/Medicaid  
Model: (Intercept), income, educat, marital, depress1

a. The full log likelihood function is displayed and used in computing information criteria.  
b. Information criteria are in small-is-better form.

Omnibus Test<sup>a</sup>

Likelihood Ratio Chi-Square	df	Sig.
5.348	4	.253

Dependent Variable: Total months on OHP/Medicaid  
Model: (Intercept), income, educat, marital, depress1

a. Compares the fitted model against the intercept-only model.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	2.062	.1918	1.686	2.438	115.498	1	.000
income	-5.328E-6	6.7609E-6	-1.858E-5	7.923E-6	.621	1	.431
educat	-.020	.0656	-.149	.109	.093	1	.761
marital	-.202	.1360	-.469	.064	2.213	1	.137
depress1	.030	.0226	-.014	.074	1.750	1	.186
(Scale)	1 <sup>a</sup>						
(Negative binomial)	1						

Dependent Variable: Total months on OHP/Medicaid  
Model: (Intercept), income, educat, marital, depress1

a. Fixed at the displayed value.

The two approaches lead to different statistical conclusions, but there appears to be an overdispersion problem. A comparison of the two log-likelihoods,  $41.780 - 5.348 = 36.432$  suggests overdispersion using a  $\chi^2$  distribution with 1 df and double the alpha level (i.e., a critical value of 2.71).

## References

- Coxe, S., West, S.G., & Aiken, L.S. (2009). The Analysis of Count Data: A Gentle Introduction to Poisson Regression and Its Alternatives. *Journal of Personality Assessment*, 91, 121–136.
- Cameron, A. C., & Trivedi, P.K. (1998) *Regression Analysis of Count Data*, Cambridge University Press, 1998
- Long, J.S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44, 443-460.