Regression Models for Count Data and Examples

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).[1]

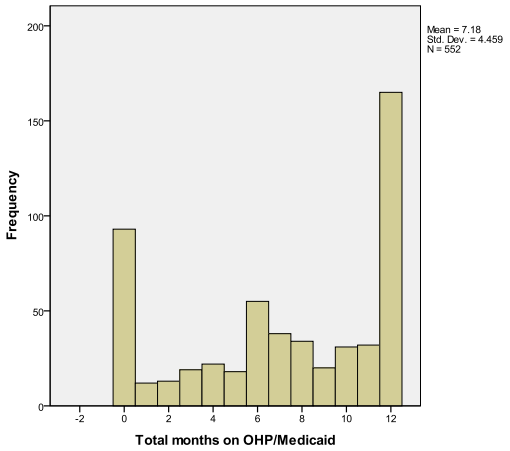
Poisson Model Example

This example also comes from Karen Seccombe's project[2] 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. This is just to illustrate the concept. The assumption is really about conditional variance (residual variance), so these descriptives *are not a test of the assumption*.

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 (at least globally) 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 (Long, 1997), because they are equivalent when the equal dispersion assumption is met. This is not the only test available—there have been many proposed (see Vives, Losilla, Rodrigo, & Portell, 2008, for example). *It is very important to state* that the because Poisson is a special case of negative binomial (if equidispersion is met the two approaches are the same), then *there is no strong need to justify the use of negative binomial* with a test of overdispersion. Just to illustrate the likelihood ratio comparison test Long mentions, I test the Poisson regression model below to compare to the negative binomial for didactic purposes.

---

[1] Although overdispersion seems to be the most common and well-studied problem, underdispersion (variance is less than the mean) can also occur. There are a number of proposed solutions, although not enough evidence to select a clear winner in all situations. The most popular solution seems to be the Conway–Maxwell–Poisson (CMP) method and evidence presented by Huang (2017) suggests that this might work well in many situations. The CMP approach and others are not available in SPSS, but see COMPoissonReg package in R or PROC COUNTREG in SAS.

[2] Seccombe, K., Newsom, J.T., & Hoffman, K. (2006). Access to healthcare after welfare reform. *Inquiry, 43*, 167-179.

SPSS

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

Goodness of Fit[b]

	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[a]	-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[a]

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[a]						

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

a. Fixed at the displayed value.

Note that pseudo-R^2 values can be computed for Poisson but is not available in SPSS. Hand computations (or spreadsheet) would be fairly simple using the equations given in the “Multiple Logistic Regression and Model Fit” handout.

R

```
> rm(d)
>
> library(haven)
> d = read_sav("c:/jason/spsswin/mvclass/count.sav")

> library(MASS)
> model1 <- glm(mosmed ~ income + educat + marital + depress1,family="poisson",data=d)
> summary(model1)

Call:
glm(formula = mosmed ~ income + educat + marital + depress1,
  family = "poisson", data = d)

Deviance Residuals:
  Min    1Q  Median    3Q   Max
-4.3169 -1.5178  0.1631  1.3575  2.6246

Coefficients:
        Estimate  Std. Error z value       Pr(>|z|)
(Intercept) 2.058720360 0.063107111 32.623 < 0.0000000000000002
income   -0.000005712 0.000002559 -2.232       0.025584
educat   -0.015722632 0.021905588 -0.718       0.472914
marital   -0.192553055 0.049769108 -3.869       0.000109
depress1   0.027815115 0.007692180  3.616       0.000299

(Dispersion parameter for poisson family taken to be 1)

  Null deviance: 1387.1 on 357 degrees of freedom
Residual deviance: 1345.4 on 353 degrees of freedom
AIC: 2522

Number of Fisher Scoring iterations: 5
```

Negative Binomial Model Example

SPSS

On the print command, adding solution(exponentiated) gives Exp(B) in the output, which is the number of times increment in the average count (or prevalence) for each increase in X compared to the prior value of X (Coxe et al., 2009, p. 132), also known as the incident rate ratio (Hardin & Hilbe, 2007). For binary X, this is the number of times (multiplicative) difference in the average count for the X=1 group vs. the X=0 group, e.g., number times in the mean number of months covered for the married vs. nonmarried group. Also, pseudo-R^2 values are not available in SPSS for negative binomial either, but could be computed by hand.[3]

```
genlin mosmed with income educat marital depress1
  /model income educat marital depress1 distribution=negbin(MLE) link=log
 /print solution(exponentiated) summary fit history.
```

Goodness of Fit[a]

	Value	df	Value/df
Deviance	465.990	352	1.324
Scaled Deviance	465.990	352	
Pearson Chi-Square	213.284	352	.606
Scaled Pearson Chi-Square	213.284	352	
Log Likelihood[b]	-1069.459		
Akaike's Information Criterion (AIC)	2150.919		
Finite Sample Corrected AIC (AICC)	2151.158		
Bayesian Information Criterion (BIC)	2174.202		
Consistent AIC (CAIC)	2180.202		

Dependent Variable: mosmed Total months on OHP/Medicaid
Model: (Intercept), total annual household income, highest ed level, marital status, ces-d mental health score higher scores more depressed

a. Information criteria are in smaller-is-better form.

b. The full log likelihood function is displayed and used in computing information criteria.

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
9.108	4	.058

Dependent Variable: mosmed Total months on OHP/Medicaid
Model: (Intercept), total annual household income, highest ed level, marital status, ces-d mental health score higher scores more depressed

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

Parameter Estimates

			95% Wald Confidence Interval		Hypothesis Test				95% Wald Confidence Interval for Exp(B)	
Parameter	B	Std. Error	Lower	Upper	Wald Chi-Square	df	Sig.	Exp(B)	Lower	Upper
(Intercept)	2.061	.1440	1.778	2.343	204.800	1	.000	7.850	5.920	10.410
total annual household income	-5.340E-6	5.1591E-6	-1.545E-5	4.772E-6	1.071	1	.301	1.000	1.000	1.000
highest ed level	-.019	.0493	-.116	.077	.154	1	.695	.981	.890	1.080
marital status	-.201	.1035	-.404	.002	3.782	1	.052	.818	.668	1.002
ces-d mental health score higher scores more depressed	.030	.0171	-.004	.063	3.018	1	.082	1.030	.996	1.065
(Scale)	1[a]									
(Negative binomial)	.513	.0598	.409	.645						

Dependent Variable: mosmed Total months on OHP/Medicaid
Model: (Intercept), total annual household income, highest ed level, marital status, ces-d mental health score higher scores more depressed

a. Fixed at the displayed value.

R

```
> rm(d)
>
> library(haven)
> d = read_sav("c:/jason/spsswin/mvclass/count.sav")

> model2 <-glm.nb(mosmed ~ income + educat + marital + depress1,data=d)
> summary(model2,digits = 3)
```

[3] Coxe and colleagues (2009) point out that the pseudo-R^2 values are problematic for negative binomial models because the nondispersion parameter differs in the intercept and full model. Nonetheless, familiar Pseudo-R^2 values, like McFadden, appear to be widely used with negative binomial models. Cameron and Windmeier (1996) discuss several pseudo-R^2 values that could be used with negative binomial models, and conclude the deviance R^2 is the best approach (based on deviance residuals from the null model and the full model, but a simple transformation of the McFadden pseudo-R^2). The deviance R^2 is implemented in Stata (e.g., Hardin & Hilbe, 2007; Hlibe, 2011).

```
Call:
glm.nb(formula = mosmed ~ income + educat + marital + depress1,
  data = d, init.theta = 1.948228735, link = log)

Deviance Residuals:
   Min    1Q  Median    3Q    Max
-2.62360 -0.74107  0.07741  0.58559  1.20398

Coefficients:
        Estimate  Std. Error z value      Pr(>|z|)
(Intercept) 2.060541269 0.136173648 15.132 <0.0000000000000002
income   -0.000005340 0.000005231 -1.021       0.3074
educat   -0.019345224 0.047217810 -0.410       0.6820
marital   -0.201266828 0.103368878 -1.947       0.0515
depress1   0.029654962 0.017205760  1.724       0.0848

(Dispersion parameter for Negative Binomial(1.9482) family taken to be 1)

  Null deviance: 475.27 on 357 degrees of freedom
Residual deviance: 465.99 on 353 degrees of freedom
AIC: 2150.9

Number of Fisher Scoring iterations: 1
       Theta: 1.948
     Std. Err.: 0.227

 2 x log-likelihood: -2138.919
```

Assessing Overdispersion

Given that the negative binomial model corrects well for overdispersion and its results will be equal to the Poisson if the data are equidispersed, there generally should be no reason to demonstrate that there was an overdispersion violation. But this computation illustrates the difference between the two approaches with this example. The Poisson and the negative binomial approaches lead to different statistical conclusions, but there appears to be an overdispersion problem. A comparison of the two likelihood ratio chi-squares (Long, 1997), 41.780 – 9.108 = 32.672 suggests overdispersion using a χ^2 distribution with 1 df and double the alpha level (i.e., a critical value of 2.71). Because the negative binomial model does not assume equidispersion, it would be the preferred modeling approach in this case, and given the equivalence to Poisson when equidispersion is met, there does not seem to be any reason to use Poisson if the negative binomial approach is available.

Pseudo-R^2 Computation

For the negative binomial model, for convenience, I computed the McFadden pseudo-R^2. I tested the intercept only (or constant only) model by leaving off the predictors (keep the same variables on the `genlin` command to make sure the N is the same as with the full model). Then use the negative log likelihood from the full model above and the intercept only model (see Coxe et al., 2009 for the general strategies and other options). The equation below should use 2 × the log likelihood value printed by SPSS, for the full model (model k), the loglikelihood value printed below is -1074.013, so the -2LL should be 2 × -1074.013 = -2138.918. The obtained log likelihood value from the null model with no predictors (intercept only model) for this example (not shown) was -1069.460, so the -2LL value is -1069.460 × 2 = -2148.026.

$$
\begin{aligned}
R^2_{McFadden} &= 1 - \left[\frac{-2LL_k}{-2LL_{null}} \right] \\
&= 1 - \left[\frac{\text{-2138.918}}{\text{-2148.026}} \right] \\
&= .0042
\end{aligned}
$$

This is one possible pseudo-R-squared value, and there is not extensive evidence on the performance of these measures with negative binomial models (but see Cameron & Windmeijer, 1996).

Goodness of Fit [a]

	Value	df	Value/df
Deviance	462.862	356	1.300
Scaled Deviance	462.862	356	
Pearson Chi-Square	199.173	356	.559
Scaled Pearson Chi-Square	199.173	356	
Log Likelihood [b]	-1074.013		
Akaike's Information Criterion (AIC)	2152.027		
Finite Sample Corrected AIC (AICC)	2152.061		
Bayesian Information Criterion (BIC)	2159.788		
Consistent AIC (CAIC)	2161.788		

Dependent Variable: mcamed Total months on OHP/Medicaid

Write-up

(I report only the negative binomial model here, because it should be generally preferred)

A negative binomial model was used to examine the relation of income, education, marital status, and depression to the number of months covered by the Oregon Health Plan. Together the predictors accounted for a marginally significant amount of variance in the outcome, likelihood ratio $\chi^2(4) = 9.108$, p = .058, McFadden pseudo-R^2 = .042, representing approximately 4.2% of the variance. Income and education were not significant predictors of months covered, B = .000, SE_B = .000, p = .31, 95% CI[-.000,.000] and B = -.019, SE_B = .049, p = .68, 95% CI[-.116,.077], respectively. Marital status and depression were marginally significant predictors of months covered, B = -.201, SE_B = .104, p = .052, 95% CI[-.404,.002] and B = .030, SE_B = .017, p = .082, 95% CI[-.004,.063], respectively.

Given the results in this case were not significant, I did not get into a detailed interpretation of the coefficients, but one could describe the odds increase in the number of months covered for each unit increase in the predictor by using the exponential transformation of the slope. For example, for marital status, $e^{-.201} = .818$*, in which the corresponding statement along the lines of the following could be added:*

Being unmarried was associated with an approximately 1.22 times more months covered compared with being married ($1/e^{-.201} = 1/.818 = 1.22$).

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

Cameron, A. C., & Windmeijer, F. A. (1996). R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics, 14*(2), 209-220.

Hardin, J. W., & Hilbe, J. M. (2007). *Generalized linear models and extensions*. Stata press.

Hilbe, J. M. (2011). Negative binomial regression. Cambridge University Press.

Long, J.S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

Vives, J., Losilla, J. M., Rodrigo, M. F., & Portell, M. (2008). Overdispersion tests in count-data analysis. *Psychological Reports, 103*, 145-160.