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).

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. 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.

<table>
<thead>
<tr>
<th>Descriptive Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>mosmed Total months on OHP/Medicaid</td>
<td>N</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>552</td>
</tr>
</tbody>
</table>

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 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 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.
SPSS

**Goodness of Fit**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>df</th>
<th>Value/df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviance</td>
<td>1345.360</td>
<td>353</td>
<td>3.811</td>
</tr>
<tr>
<td>Scaled Deviance</td>
<td>1345.360</td>
<td>353</td>
<td></td>
</tr>
<tr>
<td>Pearson Chi-Square</td>
<td>964.622</td>
<td>353</td>
<td>2.732</td>
</tr>
<tr>
<td>Scaled Pearson Chi-Square</td>
<td>964.622</td>
<td>353</td>
<td></td>
</tr>
<tr>
<td>LogLikelihood*</td>
<td>-1255.992</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Akaike's Information Criterion (AIC)</td>
<td>2521.584</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finite Sample Corrected AC (NCC)</td>
<td>2522.154</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bayesian Information Criterion (BIC)</td>
<td>2541.386</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistent AC (CACC)</td>
<td>2548.388</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dependent Variable: Total months on CHIP/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.

**Parameter Estimates**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>R</th>
<th>Std. Error</th>
<th>Lower</th>
<th>Upper</th>
<th>Wald Chi-Square</th>
<th>df</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.059</td>
<td>0.0631</td>
<td>1.935</td>
<td>2.182</td>
<td>1064.235</td>
<td>1</td>
<td>.000</td>
</tr>
<tr>
<td>income</td>
<td>-5.71E-6</td>
<td>2.55E-6</td>
<td>-1.07E-5</td>
<td>-6.97E-7</td>
<td>4.984</td>
<td>1</td>
<td>.026</td>
</tr>
<tr>
<td>educat</td>
<td>-.016</td>
<td>.0219</td>
<td>-.059</td>
<td>.027</td>
<td>.515</td>
<td>1</td>
<td>.473</td>
</tr>
<tr>
<td>marital</td>
<td>-0.193</td>
<td>0.0498</td>
<td>-.290</td>
<td>-.096</td>
<td>14.969</td>
<td>1</td>
<td>.000</td>
</tr>
<tr>
<td>depress1</td>
<td>0.028</td>
<td>0.0077</td>
<td>0.013</td>
<td>0.043</td>
<td>15.076</td>
<td>1</td>
<td>.000</td>
</tr>
</tbody>
</table>

(Dispersion parameter for poisson family taken to be 1)

Dependent Variable: Total months on CHIP/Medicaid
Model: (Intercept), income, educat, marital, depress1
a. Fixed at the displayed value.

R

```r
> rm(mydata)
> library(haven)
> mydata = read_sav("c:/jason/spsswin/mvclass/count.sav")
> library(MASS)
> model1 <- glm(mosmed ~ income + educat + marital + depress1,family="poisson",data=mydata)
> summary(model1)
```

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

Deviance Residuals:
```
     Min      1Q  Median      3Q     Max
-4.3170 -1.5178  0.1614  1.3575  2.6246
```

Coefficients:
```
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.058720360 0.063107111 32.623 < 0.0000000000000002
income -0.000005712 0.000025599 -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
```

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**

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

### Goodness of Fit

<table>
<thead>
<tr>
<th>Value</th>
<th>df</th>
<th>Value of</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviance</td>
<td>465.599</td>
<td>352</td>
</tr>
<tr>
<td>Scaled Deviance</td>
<td>465.599</td>
<td>352</td>
</tr>
<tr>
<td>Pearson Chi-Square</td>
<td>213.284</td>
<td>352</td>
</tr>
<tr>
<td>Scaled Pearson Chi-Square</td>
<td>213.284</td>
<td>352</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-1069.469</td>
<td></td>
</tr>
<tr>
<td>Akaike's Information Criterion (AIC)</td>
<td>2150.919</td>
<td></td>
</tr>
<tr>
<td>Finite Sample Corrected AIC (AICC)</td>
<td>2151.198</td>
<td></td>
</tr>
<tr>
<td>Bayesian Information Criterion (BIC)</td>
<td>2174.302</td>
<td></td>
</tr>
<tr>
<td>Consistent AIC (CAIC)</td>
<td>2190.282</td>
<td></td>
</tr>
</tbody>
</table>

**Omnibus Test**

<table>
<thead>
<tr>
<th>Likelihood Ratio Chi-Square</th>
<th>df</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.108</td>
<td>4</td>
<td>.058</td>
</tr>
</tbody>
</table>

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

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

### Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>B</th>
<th>Std. Error</th>
<th>95% Wald Confidence Interval</th>
<th>Wald Chi-Square</th>
<th>df</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.061</td>
<td>.1440</td>
<td>1.778 - 2.343</td>
<td>204.860</td>
<td>1</td>
<td>.000</td>
</tr>
<tr>
<td>income</td>
<td>-5.340E-006</td>
<td>5.159E-006</td>
<td>-1.545E-005 - 4.727E-006</td>
<td>1.071</td>
<td>1</td>
<td>.301</td>
</tr>
<tr>
<td>educat</td>
<td>-0.019</td>
<td>.0493</td>
<td>-1.107 - .074</td>
<td>.077</td>
<td>1</td>
<td>.730</td>
</tr>
<tr>
<td>marital</td>
<td>-2.01</td>
<td>.035</td>
<td>-3.24 - -0.78</td>
<td>.254</td>
<td>1</td>
<td>.616</td>
</tr>
<tr>
<td>depress1</td>
<td>.030</td>
<td>.0117</td>
<td>.004 - .056</td>
<td>.063</td>
<td>1</td>
<td>.802</td>
</tr>
<tr>
<td>(Scale)</td>
<td>.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Negative binomial)</td>
<td>.513</td>
<td>.0598</td>
<td>.401 - .645</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

- Fixed at the displayed value.

**R**

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

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

Call:
```
glm.nb(formula = mosmed ~ income + educat + marital + depress1, 
   data = mydata, 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 -0.0000000340 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 \( \chi^2 \) 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, I tested the intercept only (or constant only) model by leaving off the predictors (keep the same variables on the \texttt{genlin} command to make sure the N is the same as with the full model). Then use the negative log likelihood from the full mode above and the intercept only model (see Coxe et al., 2009 for the general strategies and other options). The below equation should use \( 2 \times \) 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 \( 2 \times -1074.013 = -2148.918 \). 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 \( \times 2 = -2148.026 \).

\[
R^2_{\text{Nagelkerke}} = \frac{1 - \left(\frac{-2LL_{\text{null}}}{-2LL_k}\right)^{2/n}}{1 - \left(\frac{-2LL_{\text{null}}}{-2LL_{\text{null}}}ight)^{2/n}}
\]

\[
= \frac{1 - \left(\frac{-2148.026}{-2138.918}\right)^{2/358}}{1 - \left(\frac{-2148.026}{-2138.918}\right)^{2/358}}
\]

\[= .00001\]

This is one possible pseudo-\( R^2 \)-squared value, and there is no clear consensus on which pseudo-\( R^2 \) the best to use, particularly with Poisson and negative binomial models.

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 \( \chi^2(4) = 9.108, p = .058 \). Nagelkerke pseudo-\( R^2 = .000 \), representing approximately 0.0% 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 \(\frac{1}{e^{-.201}} = \frac{1}{.818} = 1.22\).

References