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.

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.
Negative Binomial Model

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

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 – 9.08 = 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).
R Examples

### Poisson

```r
> 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
-4.3169  -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

```r
> 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      -0.000005340  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

Note: the anova function in R could be used to obtain a likelihood ratio test from comparing the intercept-only (null) model and the full model for either the Poisson or negative binomial test (see the multiple logistic example handout for illustration).

### References

