Poisson versus Negative Binomial Regression

Randall Reese

Utah State University

rreese531@gmail.com

February 29, 2016

Overview

1. Handling Count Data
   - ADEM
   - Overdispersion
2. The Negative Binomial Distribution
   - Foundations of Negative Binomial Distribution
   - Basic Properties of the Negative Binomial Distribution
   - Fitting the Negative Binomial Model
3. Other Applications and Analysis in R
4. References

Count Data

Count Data

- Data whose values come from $\mathbb{Z}_{\geq 0}$, the non-negative integers.
- Classic example is deaths in the Prussian army per year by horse kick (Bortkiewicz)
- Example 2 of Notes 5. (Number of successful "attempts").

Poisson Distribution

- Support is the non-negative integers. (Count data).
- Described by a single parameter $\lambda > 0$.
- When $Y \sim \text{Poisson}(\lambda)$, then

$$\mathbb{E}(Y) = \text{Var}(Y) = \lambda$$

Acute Disseminated Encephalomyelitis

Acute Disseminated Encephalomyelitis (ADEM) is a neurological, immune disorder in which widespread inflammation of the brain and spinal cord damages tissue known as white matter. (National Organization for Rare Disorders)

Individuals who experience ADEM are reported (anecdotally) to be more prone to seizures.

Example Clinical Study

Seizures in individuals with ADEM

A study is performed to test whether ADEM affects an individual's likelihood of having a seizure. An observational study of 274 participants was conducted over 6 months. Every participant had previously had 1+ seizure within the last 2 years. Two main groups were constructed: non-ADEM and ADEM, based upon whether a subject had ADEM or not. The number of seizures experienced by each participant over a 6 month period was recorded.

The blood sodium levels (in milliequivalents per liter [mEq/L]) of each participant were measured three times (0 months, 3 months, 6 months) and averaged. The age and sex of each participant was also recorded.

View Data in SAS

```
data ADEMdata;
  infile "/home/rreese531/my_courses/BioStat ADEM.csv"
    firstobs=2 delimiter=",";
  input seizeNum ADEM sex age bloodNa;
run;

proc print data=ADEMdata;
run;
```

Example Clinical Study

We want to model the average number of seizures, based on predictors ADEM, age, sex, and average BloodNa level.
Let S_i be the number of seizures for participant i. We will assume

$$S_i \sim \text{Pois}(\lambda_i),$$

where λ_i is the expected number of seizures for individual i.

Poisson Regression

$$\log(\lambda_i) = \beta_0 + \beta_{AD}(ADEM)_i + \beta_{ag}(age)_i + \beta_s(sex)_i + \beta_{bNa}(bldNa)_i$$

Poisson Regression in SAS

- Using `proc genmod` and the `log` link function (log-linear regression).

```
proc genmod data=ADEMdata;
model seizeNum = ADEM sex age bloodNa/
       dist=Poisson link=log;
run;
```

Examining the Fitted Model

- The covariates of ADEM and age are significant (as is $\hat{\beta}_0$).
- However, sex and bloodNa level were not determined to be significant.

Model Assumptions on Shaky Ground

- However, what about our assumption that, for a given covariate profile $\mathcal{T}$, the Poisson parameter $\lambda_{\mathcal{T}}$ represents *both* the mean and the variance?
- Heterogeneity may cause us issues here.
- Look at males without ADEM, broken down into 5 year age groups. (BloodNa level not at all significant).
- (For subjects of the same gender and ADEM status, the expected number of seizures is estimated to differ by about 5%).

SAS Code for Mean and Variance Comparisons by Group

```
proc format;
value ageGroup
0 = ’20-24’ 1 = ’25-29’ 2 = "30-34" 3 = "35-39"
4 = "40-44" 5 = "45-49" 6 = "50-54" 7 = "55-59"
8 = "60-64" 9 = "65 and up";

proc sort data=ademdata; by ageGrp;
run;

proc means var mean n data=ademdata;
format ageGrp ageGroup.;
var seizeNum; where ADEM = 0 and sex = 0;
by ageGrp;
run;
```

Overdispersion

- We have some heuristic evidence of overdispersion caused by heterogeneity.
- Also look at Pearson and Deviance statistics (Value/df $\approx$ 1).

Overdispersion

Overdispersion occurs when, for a random variable $Y \sim \text{Pois}(\lambda)$,

$$\mathbb{E}(Y) < \text{Var}(Y).$$

In other words, for a Poisson model, if our variance is larger than our expected value, we have **overdispersion.**

Testing for Overdispersion

Our test for overdispersion is based on an assumption that if $\mathbb{E}(S) = \lambda$, then there is some $\delta > 0$ such that

$$\text{Var}(S) = \lambda + \delta\lambda^2.$$

(More this assumption in a moment. Hint ... Negative Binomial). The Hypotheses:

$$H_0 : \ \delta = 0 \qquad\qquad H_A : \ \delta > 0$$

Employing Lagrange multipliers (see [Cameron & Trivedi, 1998]), we get a test statistic $D \sim \chi^2_1$.

Here $D = 20.8819$, with p-value < 0.0001.

Testing for Overdispersion

We can test for overdispersion in SAS.
A few small changes are made in the previous `proc genmod` sequence:

- Change `dist` to `negbin`.
- Add `scale = 0 noscale` options.

```
proc genmod data=ADEMdata;
model seizeNum = ADEM sex age bloodNa/
       dist=negbin scale=0 noscale link=log;
run;
```

The Negative Binomial Distribution

In the presence of Poisson overdispersion for count data, an alternative distribution called the

Negative Binomial Distribution

may avail a better model.

The Negative Binomial Distribution

First Definition: Bernoulli Trials

The number of successes in a sequence of independent and identically distributed Bernoulli trials before a specified (and fixed) number of failures occurs.

Denote the fixed number of failures as $r > 0$ and the probability of success in each Bernoulli trial as $p \in (0, 1)$.

$$\text{NegBinom}(r, p).$$

The pmf is then given by

$$f(k) = \binom{k+r-1}{k} \cdot (1-p)^r p^k \qquad k \in \{0, 1, 2, 3, \dots\}$$

The Negative Binomial Distribution

Second Definition: Gamma-Poisson Mixture

If we let the Poisson means follow a gamma distribution with shape parameter r and rate parameter $\beta = \frac{1-p}{p}$ (so Pois(λ) mixed with Gamma(r, β)), then the resulting distribution is the negative binomial distribution.

The pmf is then given by

$$f(k) = \frac{\Gamma(k+r)}{k!\Gamma(r)} \cdot (1-p)^r p^k \qquad k \in \{0, 1, 2, 3, \ldots\}$$

This is an important extension in that it allows for r to be any positive real number.

$$
\begin{aligned}
f(k; r, p) &= \int_0^\infty f_{\mathrm{Poisson}(\lambda)}(k) \cdot f_{\mathrm{Gamma}\left(r, \frac{1-p}{p}\right)}(\lambda)\, d\lambda && (1)\\
&= \int_0^\infty \frac{\lambda^k}{k!} e^{-\lambda} \cdot \lambda^{r-1} \frac{e^{-\lambda(1-p)/p}}{\left(\frac{p}{1-p}\right)^r \Gamma(r)}\, d\lambda && (2)\\
&= \frac{(1-p)^r p^{-r}}{k!\,\Gamma(r)} \int_0^\infty \lambda^{r+k-1} e^{-\lambda/p}\, d\lambda && (3)\\
&= \frac{(1-p)^r p^{-r}}{k!\,\Gamma(r)}\, p^{r+k}\, \Gamma(r+k) && (4)\\
&= \frac{\Gamma(r+k)}{k!\,\Gamma(r)}\, p^k (1-p)^r. && (5)
\end{aligned}
$$

Basic Properties of the Negative Binomial Dist.

Let $Y \sim \text{NegBinom}(r, p)$. Then

$$\mathbb{E}(Y) = \frac{pr}{(1-p)} \equiv \mu$$

$$\text{Var}(Y) = \frac{pr}{(1-p)^2} = \mu + \frac{1}{r}\mu^2$$

Hence our assumption on the variance in the test for overdispersion. Note that as $r \to \infty$, we get the Poisson distribution.

Basic Properties of the Negative Binomial Dist.

$$\mathrm{Var}(Y) = \frac{pr}{(1-p)^2} = \mu + \frac{1}{r}\mu^2$$

This extra parameter in the variance expression allows us to construct a more accurate model for certain count data, since now the mean and the variance do not need to be equal.

Fitting the Negative Binomial Model in SAS

To fit a log-linear model assuming the Negative Binomial distribution in SAS, we do

```
proc genmod data=ADEMdata;
model seizeNum = ADEM sex age bloodNa/
       dist=negbin link=log;
run;
```

Also finds an estimate of $\delta = \frac{1}{r}$, our **dispersion parameter**. See [SAS Help 9.3] for further information.

Examining Goodness of Fit

- Examine the Pearson Statistic/df. Should be close to 1.
- Also can look at AIC. (Lower is better).

Doing this in R

```
adem = read.csv("BioStat ADEM.csv")

ademPoisson = glm(numSeize˜.,fam = poisson, d = adem)
ademPoisson

ademNegBinom = glm.nb(numSeize ˜., data = adem )
ademNegBinom
```

Doing this in R

Results in R:

```
Call:  glm(form = numSeize~., fam = poisson, d = adem)

Coefficients:
(Intercept)   ADEM          Sex           Age     BloodNa
  0.3688593   0.43787    -0.02158      0.01265   -0.0005014

Degrees of Freedom: 273 Total (i.e. Null);  269 Residual
Null Deviance:      503.4
Residual Deviance: 454.5   AIC: 1120
```

Doing this in R

Results in R:

```
Call:  glm.nb(for = numSeize ~ .,
 d = adem, init.theta = 4.130387732, link = log)

Coefficients:
(Intercept)  ADEM      Sex          Age         BloodNa
   0.368726  0.419239  -0.023199    0.012360   -0.000384

Degrees of Freedom: 273 Total (i.e. Null);  269 Residual
Null Deviance:     323.5
Residual Deviance: 294.7  AIC: 1083
```

Some things in R

- In the `R` output, `init.theta` refers to r, where r is as above.

References

Cameron, A. C. and Trivedi, P. K. (1998)
Regression Analysis of Count Data
Cambridge: Cambridge University Press.

SAS/STAT(R) User Guide 9.3
The GENMOD Procedure
SAS Institute Inc.