Negative Binomial Regression

Shengping Yang PhD ,Gilbert Berdine MD

In the hospital stay study discussed recently, it was mentioned that "in case overdispersion exists, Poisson regression model might not be appropriate." I would like to know more about the appropriate modeling method in that case.

Although Poisson regression modeling is widely used in count data analysis, it does have a limitation, i.e., it assumes that the conditional distribution of the outcome variable is Poisson, which requires that the mean and variance be equal. The data are said to be overdispersed when the variance exceeds the mean. Overdispersion is expected for contagious events where the first occurrence makes a second occurrence more likely, though still random. Poisson regression of overdispersed data leads to a deflated standard error and inflated test statistics. Overdispersion may also result when the success outcome is not rare. To handle such a situation, negative binomial regression is commonly recommended.

The Poisson distribution can be considered to be a special case of the negative binomial distribution. The negative binomial considers the results of a series of trials that can be considered either a success or failure. A parameter ψ is introduced to indicate the number of failures that stops the count. The negative binomial distribution is the discrete probability function that there will be a given number of successes before ψ failures. The negative binomial distribution will converge to a Poisson distribution for large ψ.

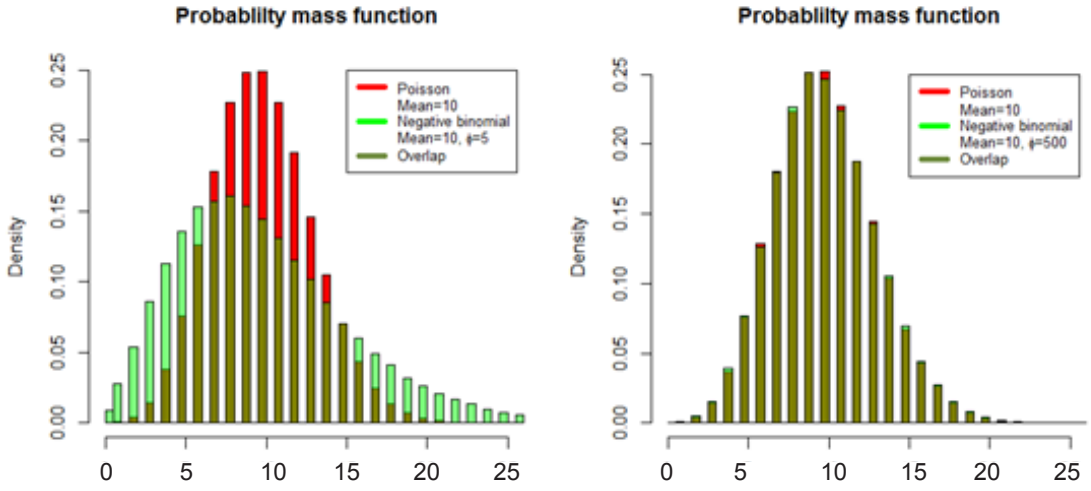

Figure 1. *Comparison of Poisson and negative binomial distributions.*

Corresponding author: Shengping Yang
Contact Information: Shengping.yang@ttuhsc.edu.
DOI: 10.12746/swrccc2015.0310.135

Figure 1 shows that when ψ is small (e.g., ψ=5), a negative binomial distribution is more spread than a Poisson distribution with the same mean. However, when ψ is large (e.g., ψ =500), the two distributions mostly overlap.

1. The negative binomial regression model.

Negative binomial distribution is defined as a discrete distribution of the number of successes in a sequence of independent and identically distributed Bernoulli trials before a specified number of failures are observed. More intuitively, it can be viewed as a Poisson distribution with parameter λ, where λ itself is not fixed but a random variable which follows a Gamma distribution. The Gamma distribution is a continuous probability function with a shape parameter ψ, a rate parameter δ, and the Gamma function Γ. The physical meaning of the Gamma distribution is an average or expected waiting time for a random event. Now, suppose that variable v_i follows a Gamma distribution that

$$k(v_i;\psi;\delta) = \frac{\delta^{\psi}}{\Gamma(\psi)} v_i^{\psi-1} e^{-v_i\delta}, \text{ where } v_i > 0, \delta > 0, \psi > 0,$$

and $E(v_i) = \psi / \delta$, $Var(v_i) = \psi / \delta^2$. By setting $\psi = \delta$, and transforming the one parameter Gamma distribution as a function of the Poisson mean λ_i, we get

$$g(\lambda_i;\psi;\mu_i) = \frac{(\psi/\mu_i)^{\psi}}{\Gamma(\psi)} \lambda_i^{\psi-1} e^{-\lambda_i\psi/\mu_i}.$$

To get the marginal distribution of the outcome variable y_i , we have

$$f(y_i;\psi,\mu_i) = \int_0^{\infty} g(y_i;\lambda_i,\psi,\mu_i) h(\lambda_i) d\lambda_i$$

$$f(y_i;\psi,\mu_i) = \int_0^{\infty} \frac{\left(\psi/\mu_i\right)^{\psi}}{\Gamma(\psi)} \lambda_i^{\psi-1} e^{-\lambda_i\psi/\mu_i} \frac{e^{-\lambda_i}\lambda_i}{y_i!} d\lambda_i$$

$$= \left(\frac{y_i + \psi - 1}{\psi - 1}\right)\left(\frac{\psi}{\mu_i + \psi}\right)^{\psi}\left(\frac{\mu_i}{\mu_i + \psi}\right)^{y_i},$$

which is the probability mass function of a negative binomial distribution (NB2). Note that

$$E(y_i;\psi,\mu_i) = \mu_i, \text{ and}$$

$$Var(y_i;\psi,\mu_i) = \mu_i + \frac{\mu_i^2}{\psi},$$

where the conditional variance is greater than the conditional mean of the outcome variable (overdispersion). In addition, it is clear that when ψ is very large, $\frac{\mu_i^2}{\psi} \to 0$, $Var(y_i;\psi,\mu_i) \sim \mu_i = E(y_i;\psi,\mu_i)$, and the expected value or mean converges to the variance.

2. Application of negative binomial regression in count data analysis.

Applying a negative binomial regression to model count data with overdispersion is straightforward using available statistical software. For example, the SAS code (see below) for negative binomial regression modeling is very similar to that for Poisson regression. In the hospital length of stay study, the outcome variable is *LOS* (length of stay), and the risk factors of interest are corticosteroid (whether or not a patient was treated with corticosteroids within 60 minutes), *race*, and *gender.* Comparing with Poisson regression, the only change is to replace the *dist* (distribution) option "*poisson*" with "*nb*", which is the abbreviation for "negative binomial".

```
proc genmod;
    class corticosteriods race gender;
    model LOS = corticosteriods race gender
<other risk factors>/dist=nb;
run;
```

As previously described, the class statement is used to define that *corticosteriods*, *race*, and *gender* are all categorical variables. And the *dist=nb* option is used to indicate that the conditional distribution of *LOS* is assumed to be negative binomial.

Applying a negative binomial regression using *R* software is also straightforward. Instead of using the glm function previously used for applying Poisson regression, the *glm.nb* function, which is modified version of the *glm* function, can be readily used. Since the *glm.nb* function is specifically developed for negative binomial regression modeling, we do not need to include the "*family*=" option.

glm.nb (LOS ~ corticosteroid + race + gender + <other risk factors>, data=data).

3. Assumptions of Negative binomial regression.

Negative binomial regression shares many common assumptions with Poisson regression, such as linearity in model parameters, independence of individual observations, and the multiplicative effects of independent variables. However, comparing with Poisson regression, negative binomial regression allows the conditional variance of the outcome variable to be greater than its conditional mean, which offers greater flexibility in model fitting. Note that negative binomial regression does not handle the underdispersion situation, where the conditional variance is smaller than the conditional mean. Fortunately, underdispersion is rare in practice.

4. Interpretation of the result of a Poisson regression.

A statistical test is highly recommended after running a Poisson regression to determine whether overdispersion exists. One such test is to apply an ordinary least square regression without the intercept term (Cameron and Trivedi, 1996). The dependent variable is defined as $\frac{(y_i - \hat{y}_i)^2 - y_i}{\hat{y}_i}$, where $\hat{y}_i = \exp(X_i\beta)$ is the predicted value obtained using the Poisson regression. By using the $\hat{y}_i$ as the independent variable in the ordinary least square regression, a *t* test can be performed, and a small p value ($p<0.05$) implies that overdispersion exists.

5. Zero-inflated and zero-truncated Poisson/negative binomial regressions.

There are many potential causes of overdispersion; one is the excessive zeros in the data. For example, suppose that the outcome of interest is the number of hospital re-admissions, and then it is quite obvious that there might be a large number of patients who do not have any re-admission (the number of re-admission is 0 for these patients). If this is the cause of overdispersion, then zero-inflated regression is appropriate to model these data (Lambert, 1992).

Considering that zero-inflated count data are generated by a mixture of two statistical processes, i.e., the first one always generates zero counts and the second one generates both zero and nonzero counts. Correspondingly, a logit model can be used to determine if the outcome of an individual observation is from the always-zero or not-always-zero groups, and then a Poisson or negative binomial model can be used to model the counts for the not-always-zero group. We will not discuss the details of zero-inflated regression here; however, the implementation of zero-inflated regression is straightforward in SAS, and the example code is

```
proc genmod;
   class <corticosteriods risk factors>;
   model count = <risk factors> /dist=zip;
   zeromodel = <risk factors>/link=logit;
   run;
```

The *dist* (distribution) option is defined as "*zip*" (zero-inflated Poisson), and an addition statement, *zeromodel* is used to model which group (either the always-zero or the not-always-zero) an individual observation comes from.

In contrast, sometimes, the data-generating process does not allow 0s; for example, if we are interested in modeling the number of wounds for patients at a trauma center, we might find that all the patients have at least one wound. Under this situation, a zero-truncated negative binomial regression can be used.

Author affiliation : Shengping Yang is a biostatistician in the Department of Pathology at TTHUSC. Gilbert Berdine is a pulmonary physician in the Department of Internal Medicine at TTUHSC.
Submitted: 3/26/2015
Accepted: 4/7/2015
Published electronically: 4/15/2015

References

1. Cameron AC, Trivedi P K. Count data models for financial data. Handbook of Statistics 1996, 14, Statistical Methods in Finance, 363-392, Amsterdam
2. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 1992; 34(1): 1-14.
3. North-Holland C, Greenwood M, Yule GU. An enquiry into the nature of frequency distributions representative of multiple happenings with particular reference of multiple attacks of disease or of repeated accidents. J R Statist Soc 1920; 83: 25–279