# Poisson Regression

*Shengping Yang PhD ,Gilbert Berdine MD*

*In a recent study, we were evaluating how risk factors, such as the timing of corticosteroid treatment, are associated with hospital length of stay for pediatric patients who were admitted due to acute asthma exacerbations. I noticed that the length of stay was recorded in the database as integers, such as 0 (if less than 24 hours), 1, 2, 3…, and the distribution seems to be quite skewed. I would like to know more about how to analyze data from such a study.*

**T** his type of data, like hospital length of stay (LOS), is often called "count data" since the observations of the outcome variable can take only non-negative integers, such as 0, 1, 2... In general, methods developed for data with normal or binomial distributed outcome variables are not appropriate for analyzing such data.

One methodology for dealing with count data is the Poisson distribution. The Poisson distribution has many real world applications, including predicting the rates of mutations, distribution of traffic flow, and radioactive decay.

Consider an event that during any time increment can have either a success or a failure. Radioactive decay is a truly random example. One can look at hospital discharge, extubation, or other medical issues and test whether or not the results are random. For any time increment, there is a probability for success. The probability for each unit being examined is assumed to be the same, so a parameter λ is defined to be the average number of successes during each time increment. If there are $N$ units in the system, then:

$$\lambda = Np ,$$

where $p$ is the probability that any given unit will have a success.

***Corresponding author:*** Shengping Yang
***Contact Information:*** Shengping.yang@ttuhsc.edu.
**DOI**: 10.12746/swrccc2015.0309.125

Rearranging gives:

$$p = \lambda/N .$$

The probability that exactly $K$ successes will be observed during a time increment is a simple problem of combinatorial probability given by a variation on the binomial distribution:

$$P(K) = p^K (1-p)^{N-K} = (\lambda/N)^K (1-\lambda/N)^{N-K} .$$

The limit of this expression as $N$ becomes very large is the Poisson distribution:

$$P(K) = \lambda^K e^{-\lambda}/K! .$$

The Poisson distribution has a mean, or an average expected value, of λ. The variance of a Poisson distribution is also equal to λ. The Poisson distributions with different λ can be seen in Figure 1. This distribution can be used as a null hypothesis to see if the observed distribution is random or not.

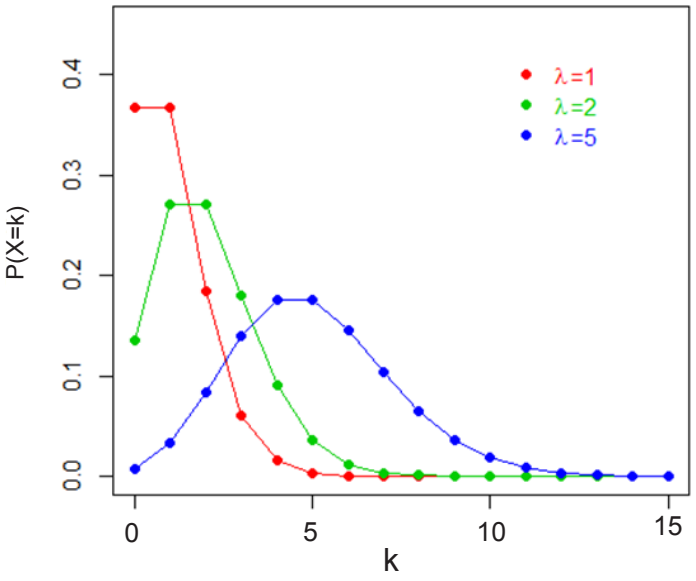
Recall that when the outcome variable had a normal distribution, ordinary linear regression can be used to evaluate the relationship between the outcome and the explanatory variables. The linear model used is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} . \qquad (1)$$

The count data collected in your study is not suitable for this linear model. Compared to normally distributed data, which allows negative values, count data are all non-negative integers and thus the mean value of

counts is always greater than 0. (Note that normal distribution cannot impose such a restriction.) Also, the distribution of count data is skewed toward the right, and the variance of count data tends to increase with the mean; thus the usual assumption of homoscedasticity would not be appropriate for count data.

**Figure 1**. Poisson distributions with different λ




We have previously (Yang and Berdine, 2014) discussed the application of logistic regression in situations where the outcome variable is binary, *i.e.*, 0 represents control and 1 represents case. Instead of directly modeling the binary outcome, the logit of the expected value of the outcome variable being a case is used as a transformed version of the dependent variable, and the logistic regression model is

$$\log_e\left(p_1 \big/ (1-p_i)\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}\,, \quad (2)$$

where $p_i$ is the expected value of the outcome being a case for subject $i$.

For count data, Poisson regression is commonly used.

**1.The Poisson regression model.**

The Poisson regression model assumes that the observed outcome variable follows a Poisson distribution and is characterized by a mean expected value ($\lambda$ in the above discussion) which is also its variance. The Poisson Regression attempts to 'fit' this parameter (which we will call $\mu$) to a linear model of input or explanatory variables. The simple linear model (Eq. 1) cannot be used as $\mu$ can take on only positive values. A log transformation of $\mu$ solves this problem. The Poisson Regression model, therefore, is:

$$\log_e(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}\,, \quad (3)$$

Take the exponential of both side of Eq. (3), we have,

$$\mu_i = \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik})\,. \quad (4)$$

Further, the likelihood function (Poisson Distribution) is,

$$L = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} \frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i}\,.$$

Take the logarithm of the likelihood function; we obtain the log-likelihood function,

$$l = log(L) = \sum\left(y_i \beta' X_i - e^{\beta' X_i} - log(y_i!)\right)\,. \quad (5)$$

One can see the relationship between the log of the Poisson distribution and our regression model (3).

The goal of applying a Poisson regression model is to find the association between the outcome and the potential risk factors. Thus it is of interest to estimate the regression coefficients. However, there is no closed-form solution to find the maximum likelihood estimator of $\beta$. In fact, in order to find such estimators, numeric approaches are exclusively used. Since the focus of this article is to demonstrate the application

of Poisson regression model in count data analysis, we will not discuss the details of the computational issues here.

**2.Application of Poisson regression in count data analysis.**

The objective of the hospital *LOS* study is to evaluate the risk factors associated with *LOS* for pediatric asthma patients. Now, we will illustrate how to apply a Poisson regression to model such data using available statistical software.

Since SAS is one of the widely used software in statistics, we will first provide the SAS code for this study.

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

The SAS *proc genmod* procedure is used for modeling the data. The *class* statement is used to define that *corticosteriod* (whether or not treated with corticosteroid within 60 minutes), *race* and *gender* are all categorical variables. The outcome variable *LOS* takes only non-negative integers, and thus the *dist=poisson* option is used to indicate that the *LOS* is assumed to have a Poisson distribution.

Poisson regression modeling can also be performed using other software, e.g., R, and the code for applying Poisson regression using *R* is,

***glm*** *(LOS ~ corticosteroid + race + gender + <other risk factors>, family="poisson", data=data).*

**3. Assumptions of Poisson regression.**

Like many other regression models, several assumptions have to be made to use a Poisson regression. Most of these assumptions are quite common. For example,

1) the logarithm of the outcome variable is linearly associated with the levels of the risk factors - see Eq.(3);
2) the effects of different risk factors on the outcome are multiplicative – note that because of Eq. (4), we have

$$\mu_i = \exp(\beta_0) \text{x } \exp(\beta_1 X_{i1}) \text{ x } \exp(\beta_2 X_{i2}) \text{x ... x } \exp(\beta_k X_{ik}) .$$

Comparing with the baseline, for any one unit change in $X_{ik}$, its effect on the outcome is $\exp(\beta_k)$ ; and,

3) the outcomes for individual observations are independent given all the risk factors.

In the meantime, there is another quite strong assumption underlying the Poisson regression, *i.e.,* the mean of the outcome variable is equal to its variance. In fact, many times this assumption cannot be satisfied in real data analysis. At the end of this article, we will briefly discuss how violation of this assumption affects the performance of a Poisson regression, and possible alternatives if this assumption is apparently violated.

**4. Interpretation of the result of a Poisson regression.**

Regardless of the software used, the output of a Poisson regression model usually includes regression coefficient estimates and their standard errors, the *Wald* 95% confidence limits, the *Chi-square* test statistics, and the corresponding *p* values. In the hospital LOS study, supposing that the regression coefficient estimate for *corticosteroid* is -0.20 and the corresponding standard error is 0.01, then based on the *Chi-square* test, we conclude that the timing of corticosteroid use is significantly associated with hospital LOS due to a very small *p* value.

More specifically, hospital LOS for patients who received *corticosteroid* treatment within 60 min-

utes was exp (-0.20) =0.82 times that for those who did not receive treatment within 60 minutes.

**5. Overdispersion issue with a Poisson regression.**

It is known that Poisson distribution has only one parameter and that the mean of a Poisson distribution is equal to its variance. By using a Poisson regression model, we're implicitly making a strong assumption that the mean of the count data we are modeling is equal to its variance. However, there are many situations in which this assumption is not valid; for example, the sample variance is greater than the sample mean, which is called overdispersion. A score test or likelihood ratio test can be used to detect overdispersion. For example, under the null hypothesis, the score statistic follows a chi-squared with one degree of freedom. A *p* value less than 0.05 indicates that overdispersion exists.

There are unwanted consequences if overdispersion exists, including

1) the model standard errors will not be correct and may be substantially underestimated. Thus the significance of individual regression coefficients might also be incorrect;

2) model selection might favor overly complex models due to incorrectly calculated deviance; and

3) prediction intervals will be smaller than they should be.

There are many potential causes of overdispersion, including

1) the study cohort might be very heterogeneous, and thus impose additional variability;

2) there might be correlation between individual responses;

3) excessive 0 counts in the data; and

4) some important factors are not included in the regression model.

Should overdispersion exist, Poisson regression modeling is not appropriate. In this case, a negative binomial regression or quasi-likelihood estimation could be applied to handle the excess variation of the data.

**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:** 1/4/2015
**Accepted:** 1/13/2015
**Published electronically:** 1/15/2015

**References**

**1.** Agresti A. An introduction to categorical data analysis, second edition. John Wiley & Sons, Inc. 2007; 74-83. Print.
**2**. Dean, CB. Testing for overdispersion in Poisson and binomial regression models. J Amer Statist Assoc 1992; 87: 451–457
**3.** Hinde J, Demétrio CGB. Overdispersion: models and estimation. Comput Stat Data Anal 1998; 27:151–170.
**4.** Yang S, Berdine G. Categorical Data Analysis - Logistic Regression. The Southwest Respiratory and Critical Care Chronicles 2014; 2(7):51-54.