BIOS 6110
Applied Categorical Data Analysis

Instructor: Yuan Huang, Ph.D.

Department of Biostatistics

Fall 2017

Part IV

Poisson Regression

Objectives

In this part, you will learn the following models

- Poisson regression for count data
- Poisson regression for rate data
- Models for handling overdispersion

Key

- Understand count data an rate data
- Interpretation of prediction model and estimated coefficients
- Testing and CI for the coefficients and means
- Understand different link functions: log link and identity link
- Understand different random component for the count data: Poisson and Negative Binomial
- Detection for overdispersion

Reading: Agresti (2002), Section 3.3

Count Data

Poisson regression is perhaps the second most common GLM, after logistic regression. It applies when the response is a count, such as the number of events occurring in time or space.

For example,

- $Y =$ number of parties attended in the past month
- $Y =$ number of imperfections on each of a sample of silicon wafers used in manufacturing computer chips

In epi studies, count data is a common data type that naturally arises from studies investigating the incidence or mortality of disease.

Example used in this lecture: Horseshoe Crab Data. See description in SAS code *PoissonModel*.

Poisson Distribution

- The density function for a Poisson distribution is

$$
\begin{aligned}
\Pr(Y = y) &= \frac{\mu^y}{y!}\exp\{-\mu\}, \quad \mu > 0, y = 0, 1, 2, \ldots \\
&= \exp\{-\mu + y\log(\mu) - \log(y!)\}
\end{aligned}
$$

- The mean and variance of Y are

$$E(Y) = Var(Y) = \mu$$

That is, for Poisson distributions, the variance equals the mean.

Poisson Regression with Log Link

For a single explanatory variable x, the Poisson loglinear model is

$$\log \mu = \alpha + \beta x,$$

which implies

$$\mu = \exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x}$$

- The mean of Y at $x+1$ equals the mean of Y at x multiplied by e^{β}:

  $$\frac{\mu(x+1)}{\mu(x)} = \frac{e^{\alpha}(e^{\beta})^{x+1}}{e^{\alpha}(e^{\beta})^{x}} = e^{\beta} \to \mu(x+1) = e^{\beta} \cdot \mu(x).$$

  Hence, a 1-unit increase in x has a multiplicative impact of e^{β} on μ.
- If $\beta = 0$, then $e^{\beta} = 1$: the mean of Y does not change as x changes
- If $\beta > 0$, then $e^{\beta} > 1$: the mean of Y increases as x increases
- If $\beta < 0$, then $e^{\beta} < 1$: the mean of Y decreases as x increases

Poisson Regression with Identity Link

For a single explanatory variable x, the Poisson model with identity link is

$$\mu = \alpha + \beta x$$

- The estimated μ can be negative.
- The mean of Y at $x + 1$ equals the mean of Y at x plus β:

  $$\mu(x+1) - \mu(x) = \alpha + \beta(x+1) - (\alpha + \beta x) = \beta \rightarrow \mu(x+1) = \mu(x) + \beta.$$

  Hence, a 1-unit increase in x has an additive effect of β on μ.
- If $\beta = 0$: the mean of Y does not change as x changes
- If $\beta > 0$: the mean of Y increases as x increases
- If $\beta < 0$: the mean of Y decreases as x increases

Overdispersion: Greater Variability than Expected

Count data often vary more than we would expect if the response distribution truly were Poisson.

Example: Female Horseshoe Crabs Data. In the following table, the variances are much larger than the means, whereas Poisson distributions have identical mean and variance

Table 3.3. Sample Mean and Variance of Number of Satellites

Width	No. Cases	No. Satellites	Sample Mean	Sample Variance
$<$23.25	14	14	1.00	2.77
23.25–24.25	14	20	1.43	8.88
24.25–25.25	28	67	2.39	6.54
25.25–26.25	39	105	2.69	11.38
26.25–27.25	22	63	2.86	6.88
27.25–28.25	24	93	3.87	8.81
28.25–29.25	18	71	3.94	16.88
$>$29.25	14	72	5.14	8.29

- The phenomenon of the data having greater variability than expected for a GLM is called overdispersion
- Common causes of overdispersion:
  - Heterogeneity among subjects: some important variables are not included. For example, crabs having a certain fixed width are a mixture of crabs of various weights, colors, and spine conditions.
  - Data are not identically distributed
  - Data are clustered/correlated (will discuss this later)
- Overdispersion is not an issue in ordinary regression models assuming normally distributed Y, because the normal has a separate parameter from the mean (i.e., the variance, σ^2) to describe variability
- For Poisson distributions, the variance equals the mean. Overdispersion is common in applying Poisson GLMs to counts
- Overdispersion is also a concern for logistic regression

Dealing with Overdispersion - Quasilikelihood

When overdispersion is evident, it's usually handled in one of the two ways.

1. By assuming $\text{Var}(y) = \phi\mu$ and estimating the scale parameter ϕ. This This approach, where you modify the variance function directly and do not actually specify a distribution, is then a quasilikelihood model.

- ϕ is usually estimated by the method of moment estimator $\hat{\phi} = X^2/(n-p)$, where X^2 is the Pearson's fit statistics. This χ^2-based estimator is a consistent estimator of ϕ.

- It's also possible to estimate ϕ by a deviance-based estimator $G^2/(n-p)$. However, this estimator is not consistent.

- $$\hat{\beta}_{Quasi} = \hat{\beta}_{Poisson}, \quad SE(\hat{\beta}_{Quasi}) = \sqrt{\hat{\phi}} \times SE(\hat{\beta}_{Poisson}).$$

Disadvantage: : Lacks a log-likelihood, and prevent you from using any of the likelihood-based tools: likelihood ratio tests, AIC/BIC, deviance explained, deviance residuals.

Dealing with Overdispersion - Negative Binomial

2. By changing the response distribution to negative binomial (NB), which is more dispersed than the Poisson.

Let p be the probability of "success" in a Bernoulli trial. In a sequence of Bernoulli trials, the number of failures in a sequence of Bernoulli trials before k successes follows a negative binomial distribution, $NB(k, p)$.

- The mean is $E(Y) = (1-p)k/p = \mu$
- The variance is $Var(Y) = (1-p)k/p^2 = \mu + D\mu^2$, where $D = 1/k$ is the dispersion parameter.

NB distribution can be obtained by a two-stage hierarchical process:

$Z \sim$ Gamma(k,k),

$Y|Z \sim$ Poisson$(((1-p)k/p) \times Z)$,

then, $Y \sim NB(k, p)$

Rate Data

Previously, we focus on the count data Y. We may also be interested in the rate data, Y/t, where t is an interval representing time, space, or other grouping.

Rate data is common under the case of *varying exposure*. In epi study, a commonly used rate is per person-years.

For example, the following data show the survival of patients after heart-valve replacement surgery. The exposure is the total number of patient-follow up months. In this example, it makes more sense to model the mean death rate per patient-month.

Age	Type	Exposure	Death
Under 55	Aortic	1259	4
	Mitral	2082	1
Above 55	Aortic	1417	7
	Mitral	1647	9

Example used in this lecture: British Train Accidents O Data. See description in SAS code *PoissonModel*.

Count Regression for Rate Data

Rate data can be modelled with Poisson regression by using an offset.

The sample rate is Y/t. The expected value of the rate is μ/t with $\mu = E(Y)$. A loglinear model for the expected rate has form

$$\log(\mu/t) = \alpha + \beta x$$

which implies

$$\log(\mu) = \log(t) + \alpha + \beta x$$

- This model looks like a regular Poisson regression but has an *offset* term $\log(t)$ whose coefficient is known, which is 1
- The interpretation of coefficients will stay the same except you can talk about the change in rate, or interpret for the counts but you also need to multiple counts by t.
- The expected number of outcomes satisfies

  $$\mu = t\exp(\alpha + \beta x)$$

  The mean μ is proportional to t, with proportionality constant depending on the value of the explanatory variable. For a fixed value of x, doubling the t also doubles the expected number μ