Lecture 13: GLM for Poisson Data

Dipankar Bandyopadhyay, Ph.D.

BMTRY 711: Analysis of Categorical Data Spring 2011

Division of Biostatistics and Epidemiology

Medical University of South Carolina

GLM for Counts

Situation: The outcome variable Y is a count

GLM for counts have as it's random component the **Poisson Distribution**

1. Number of cargo ships damaged by waves (classic example given by McCullagh & Nelder, 1989)
2. Number of deaths due to AIDs in Australia per quarter
3. Daily homicide counts in California

Poisson Rates

In some cases, the counts are affected by the amount of “exposure”.
For example, the number of homicides may be affected by

1. The underlying population size
2. The local gun laws
3. The size of the police force

When this occurs, we may want to take into account the “denominator” and form a rate such as

$$Y/t = rate$$

where t represents a quantification of exposure.

We will also look that this defined as

$$Y = rate * t$$

Components of the GLM

The components of a GLM for a count response are

1. Random Component: Poisson distribution and model the expected value of Y, denoted by $E(Y) = \mu$.
2. Systematic component: For now we will look at just one explanatory variable x
3. Link: We could use
   (a) **Identity Link** which would give us

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

   But, just as for binomial data, the model can yield $\mu < 0$ (Note $\mu \geq 0$)
   (b) **Log Link** (most common and the canonical link)

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

Poisson Loglinear Model

Our model is

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

Since the log of the expected value of Y is a linear function of explanatory variable(s), and the expected value of Y is a multiplicative function of x:

$$\begin{aligned} \mu &= e^{\alpha+\beta x} \\ &= e^{\alpha} e^{\beta x} \end{aligned}$$

What does this mean for μ? How do we interpret β?

Consider 2 values of x, say $(x_1 \& x_2)$ such that the difference between them equals 1. For example, $x_1 = 10$ and $x_2 = 11$.

Denote, $\mu_1 = E(Y|x = 10)$. Then

$$\mu_1 = e^{\alpha} e^{\beta 10}$$

and the expected value when $x = 11$ is

$$\begin{aligned} \mu_2 &= e^{\alpha} e^{\beta 11} \\ &= e^{\alpha} e^{\beta 10} e^{\beta} \\ &= \mu_1 e^{\beta} \end{aligned}$$

Thus, a 1-unit change in x has a multiplicative effect on the mean of Y.

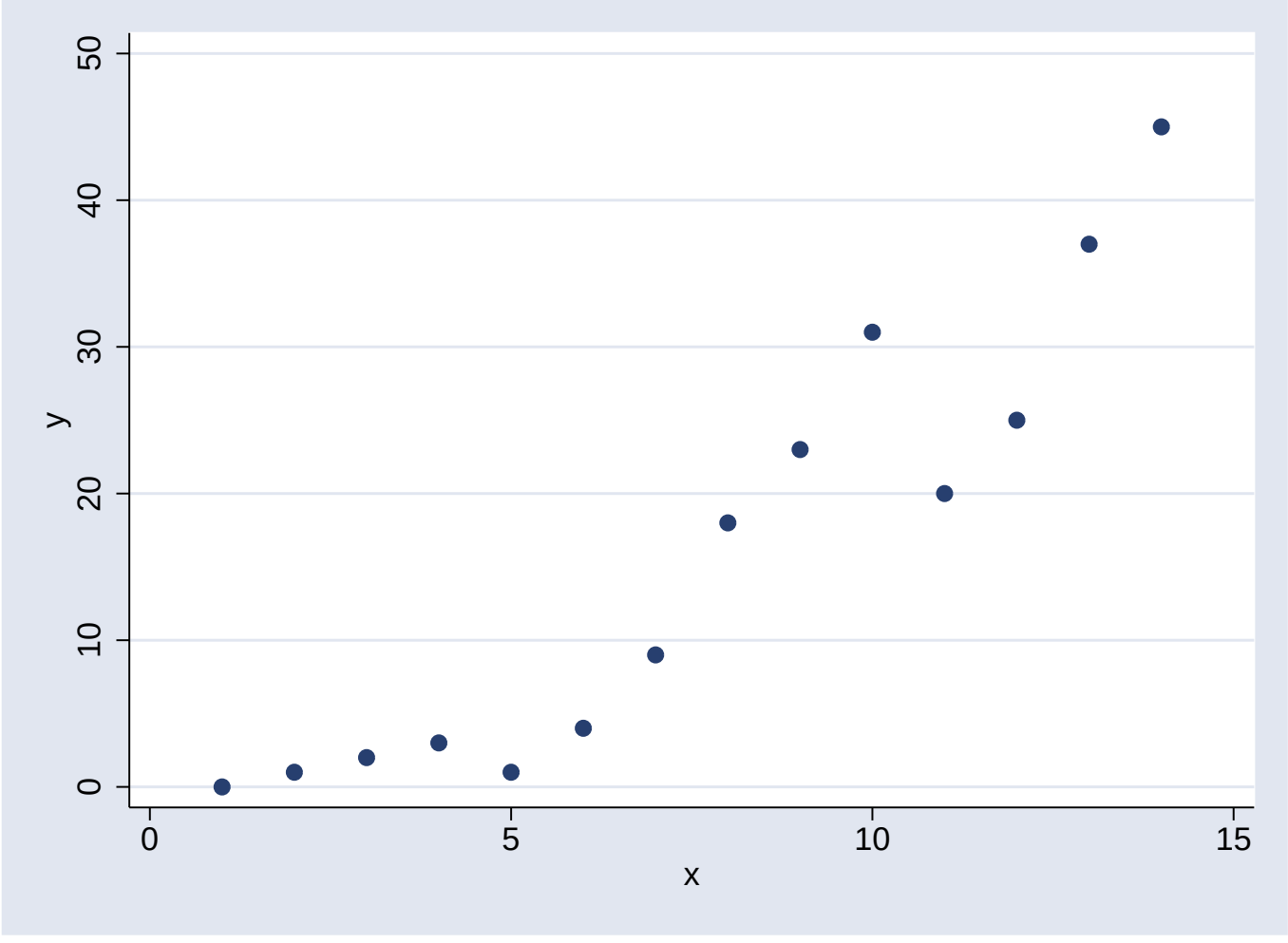
1. If $\beta = 0$, then $e^0 = 1$ and μ_2 is the same as μ_1. That is, $\mu = E(Y)$ is not related to x.
2. if $\beta > 0$, then $e^\beta > 1$ and μ_2 is e^β times **larger** than μ_1.
3. if $\beta < 0$, then $e^\beta < 1$ and μ_2 is e^β times **smaller** than μ_1.

Example

The following data represents the number of deaths from AIDS in Australia per quarter in 1983 - 1986.

Month Period	Deaths	Month Period	Deaths
1	0	8	18
2	1	9	23
3	2	10	31
4	3	11	20
5	1	12	25
6	4	13	37
7	9	14	45

Graphically, the data look like



Poisson Regression Model

We can model the Poisson regression model using GENMOD as

```
proc genmod;
 model y = x /dist=poi link = log;
run;
```

and get the following results:

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	0.3396	0.2512	1.83	0.1763
x	1	0.2565	0.0220	135.48	<.0001

Therefore, our estimated model is

$$log(\mu) = 0.3396 + 0.2565x$$

Interpretation

We see that for a 1-unit increase in month, the expectation (or mean) number of AIDs deaths increases by a factor of $e^{0.2565} = 1.292$ with a 95% C.I. of $(exp(0.2133), exp(0.2997)) = (1.238, 1.349)$.

Month	Observed Deaths	Fitted
1	0	1.815026
2	1	2.345738
3	2	3.031629
4	3	3.918073
5	1	5.063713
6	4	6.544336
7	9	8.457892
8	18	10.93097
9	23	14.12717
10	31	18.25794
11	20	23.59654
12	25	30.49614
13	37	39.41317
14	45	50.93753

Interpretation of the Poisson Regression Model

1. The marginal effect of x_i (month period i) on μ_i (expected number of deaths in month period i) is for a 1-unit increase in month period the estimated count increases by a factor of $e^{0.2565} = 1.292$
2. That is, the number of deaths is growing at the rate of 29% per year
3. We can look at the predicted probability of number of deaths given a value of x_i. Recall $Y_i \sim$Poisson(μ_i). Thus

$$\widehat{P}(Y_i = y) = \frac{e^{-\hat{\mu}_i}\hat{\mu}_i^y}{y!}$$

where

$$\hat{\mu}_i = 0.3396 + 0.2565x_i$$

For example, consider Month Period 3, the probability of y deaths would be

$$\widehat{P}(Y_i = y) = \frac{e^{-(0.3396+0.2565\cdot 3)}(0.3396 + 0.2565 \cdot 3)^y}{y!}$$

and for $y = 1$

$$\begin{aligned} \widehat{P}(Y_i = 1) &= \frac{e^{-(0.3396+0.2565\cdot 3)}(0.3396+0.2565\cdot 3)^1}{1} \\ &= 0.3658 \end{aligned}$$

Poisson Regression for Rate Data

- Events may occur over time or space (exposure)
- And the amount of exposure may vary from observation to observation

Let

$$Y = \text{count (e.g., number of observed cases)}$$

$$t = \text{days in the community}$$

Then, the sample **rate** of occurrence = Y/t with the expected value of

$$E(Y/t) = \frac{1}{t}E(Y) = \mu/t$$

The Poisson regression model with log link for the expected rate of occurrence is

$$\begin{aligned} log(\mu/t) &= \alpha + \beta x \\ log(\mu) - log(t) &= \alpha + \beta x \\ log(\mu) &= \alpha + \beta x + log(t) \end{aligned}$$

The term "$log(t)$" is an adjustment term.

It is called the **offset**.

In terms of the multiplicative model, the Poisson regression model with a log link for rate data is

$$\mu = te^{\alpha}e^{\beta x}$$

Written in this form, it is clear that

1. The expected value of counts depends on both t and x
2. Both t and x are observed and not parameters of the distribution

Example

- Suppose you observe $Y_1 = 15$ cases of leukemia in Cambridge, and $Y_2 = 30$ cases of leukemia in Boston in 1990.
- The number of cases in Cambridge has distribution

$$Y_1 \sim Poi(\mu_1)$$

  and the number of cases in Boston has distribution

$$Y_2 \sim Poi(\mu_2)$$

- You want to test if the expected number of cases in Cambridge and Boston is the same:

$$H_0 : \mu_1 = \mu_2$$

- We can form a simple WALD (difference in means over non-null standard error) test statistic:
- Since,

$$E(Y_j) = Var(Y_j) = \mu_j,$$

when

$$Y_j \sim Poi(\mu_j),$$

- And the MLE of μ_j is

$$\hat{\mu}_j = y_j \qquad \text{and} \qquad \widehat{Var}(Y_j) = \hat{\mu}_j = y_j$$

- The WALD statistic for

$$H_0 : \mu_1 = \mu_2$$

is

$$\begin{aligned} Z &= \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\widehat{Var}(\hat{\mu}_1 - \hat{\mu}_2)}} \\ &= \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\widehat{Var}(\hat{\mu}_1) + \widehat{Var}(\hat{\mu}_2)}} \\ &= \frac{y_1 - y_2}{\sqrt{y_1 + y_2}} \end{aligned}$$

Wald Test on Computer

- The easiest way to get this test statistic on the computer is using a linear regression model:

$$\mu_j = \beta_0 + \beta_1 x_j,$$

where

$$x_j = \begin{cases} 1 \text{ if group 1 (Cambridge)} \\ 0 \text{ if group 2 (Boston)} \end{cases}.$$

- In other words

$$\mu_1 = \beta_0 + \beta_1$$

$$\mu_2 = \beta_0$$

- Note that

$$\mu_1 - \mu_2 = \beta_1$$

- If

$$H_o : \beta_1 = \mu_1 - \mu_2 = 0,$$

then

$$H_o : \mu_1 = \mu_2$$

SAS PROC GENMOD

```
data one;
input city cases;
cards;
1 15
0 30
;

proc genmod data=one;
 model cases = city  / link=id /* identity or linear model */
                       dist = poi;
run;
```

```
/* SELECTED OUTPUT */

Analysis Of Parameter Estimates

                                Standard      Wald 95%          Chi-
Parameter   DF  Estimate           Error   Confidence Limits   Square  Pr > ChiSq

Intercept    1   30.0000          5.4772   19.2648   40.7352    30.00      <.0001
city         1  -15.0000          6.7082  -28.1478   -1.8522     5.00      0.0253
```

The Wald Statistic from this output is

$$Z^2 = \left(\frac{y_1 - y_2}{\sqrt{y_1 + y_2}} \right)^2 = 5$$

with $p = .0253$ so we reject the null that the expected number of cases are the same.

Log-linear model

- Since the number of cases has to be non-negative, as before, you may want to use a log-linear model

$$log(\mu_j) = \beta_0 + \beta_1 x_j$$

or, equivalently,

$$\mu_j = exp(\beta_0 + \beta_1 x_j)$$

where

$$x_j = \begin{cases} 1 \text{ if group 1 (Cambridge)} \\ 0 \text{ if group 2 (Boston)} \end{cases} .$$

- In other words

$$log(\mu_1) = \beta_0 + \beta_1$$

$$log(\mu_2) = \beta_0$$

- Note that

$$log(\mu_1) - log(\mu_2) = \beta_1$$

- If

$$H_o : \beta_1 = log(\mu_1) - log(\mu_2) = 0,$$

  then

$$H_o : \mu_1 = \mu_2 = exp(\beta_0)$$

- Then, we can use the Wald Statistic

$$\begin{aligned} Z &= \frac{\widehat{\beta}_1}{\sqrt{\widehat{Var}(\widehat{\beta}_1)}} \\ &= \frac{log(\hat{\mu}_1) - log(\hat{\mu}_2)}{\sqrt{\widehat{Var}[log(\hat{\mu}_1)] + \widehat{Var}[log(\hat{\mu}_2)]}} \end{aligned}$$

SAS PROC GENMOD

```
data one;
input city cases;
cards;
1 15
0 30
;

proc genmod data=one;
 model cases = city  / link=log dist = poi;
run;
```

```
/* SELECTED OUTPUT */

Analysis Of Parameter Estimates

                                Standard      Wald 95%            Chi-
Parameter    DF   Estimate       Error   Confidence Limits      Square  Pr > ChiSq

Intercept     1     3.4012      0.1826     3.0434    3.7590     347.04      <.0001
city          1    -0.6931      0.3162    -1.3129   -0.0734       4.80      0.0284
```

The Wald Statistic from this output is

$$Z^2 = \left(\frac{\widehat{\beta}_1}{\sqrt{\widehat{Var}(\widehat{\beta}_1)}} \right)^2 = 4.8$$

with $p = .0284$ so we reject the null that the expected number of cases are the same.

Main Interest: Rates

- However, what you are probably really more interested in, is if the “rate” of leukemia in the two cities are the same.
- Here, we define the **rate** as the number of events per number at risk in a given time.
- In particular, the number of cases could be higher in Boston because there are many more people who live in Boston than Cambridge.
- Because we want to look at the rate, we rewrite

$$\begin{aligned} \mu_j &= \text{Expected number of events in population } j \\ &= r_j T_j \end{aligned}$$

where

$$
\begin{aligned}
r_j &= \text{rate of events in population } j \\
&= \#\text{ leukemia cases per 10000 at risk} \\
&\quad \text{per year in Cambridge} \\
&= \#\text{ leukemia cases per person-years at risk} \\
&\quad \text{in Cambridge}
\end{aligned}
$$

and

$$T_j = \text{persons-years at risk in city } j \text{ in 1990}$$

Intuitively, we can think of T_j as

$$
\begin{bmatrix} \text{number of people living} \\ \text{in City } j \text{ at any time} \\ \text{in 1990} \end{bmatrix} \times \begin{bmatrix} \text{average amount of time a} \\ \text{person spent in City } j \\ \text{in 1990} \end{bmatrix}
$$

Usually T_j is provided by an investigator.

- For example, suppose the true rate of leukemia in Cambridge is

$$r_1 = \text{10 leukemia cases per 100,000 person-years}$$

and

$$T_1 = \text{500,000 person-years}$$

- Then, the expected number of cases is

$$\begin{aligned} \mu_j &= r_j T_j \\ &= \left(\frac{\text{10 cases}}{\text{100,000 person-year}} \right) 500,000 \text{person-years} \\ &= 50 \text{ cases} \end{aligned}$$

- Note, also, that a 'rate' always involves time

$$r_j = \frac{\#\text{ events}}{\text{(number at risk)} \times \text{(time at risk)}}$$

Hypotheses of Interest

- Thus, if we were really interested in testing whether the rates were the same; the hypothesis

  $$H_0 : \mu_1 = \mu_2,$$

  or, in terms of the r_j's

  $$H_0 : r_1 T_1 = r_2 T_2$$

  is not a valid test of

  $$H_0 : r_1 = r_2$$

  unless $T_1 = T_2$.
- Since Cambridge and Boston have very different population sizes, we would think that the WALD test statistic for

  $$H_0 : \mu_1 = \mu_2,$$

  is not answering our question of interest.

WALD Test for rates

- The WALD statistic for

$$H_0 : r_1 = r_2$$

  is

$$Z = \frac{\widehat{r}_1 - \widehat{r}_2}{\sqrt{\widehat{Var}(\widehat{r}_1) + \widehat{Var}(\widehat{r}_2)}},$$

  where the $\widehat{r}_j$ are estimated under the alternative.

- Throughout, we assume that T_j is known
- Then, with

$$Y_1 \sim Poi(\mu_1) = Poi(r_1 T_1)$$

  and

$$Y_2 \sim Poi(\mu_2) = Poi(r_2 T_2)$$

The MLE's (under the alternative that the rates are not equal) are

$$Y_j = \hat{\mu}_j = \widehat{r}_j T_j,$$

or, equivalently,

$$\begin{aligned} \widehat{r}_j &= \frac{Y_j}{T_j} \\ &= \frac{\text{observed events in group } j}{\text{person-years exposure}} \end{aligned}$$

It's more common to deal with log's of rates.

Log-linear model

- As with the cases, since rates always have to be positive (although not constrained to be in [0,1]):

$$r_j > 0$$

  so you often see r_j modelled as a log-linear model

- Since the number of cases has to be non-negative, as before, you may want to use a log-linear model

$$log(r_j) = \beta_0 + \beta_1 x_j$$

  or, equivalently,

$$r_j = exp(\beta_0 + \beta_1 x_j)$$

  where

$$x_j = \begin{cases} \text{1 if group 1 (Cambridge)} \\ \text{0 if group 2 (Boston)} \end{cases} .$$

- In other words

$$log(r_1) = \beta_0 + \beta_1$$

$$log(r_2) = \beta_0$$

- Note that

$$log(r_1) - log(r_2) = log\left(\frac{r_1}{r_2}\right) = \beta_1,$$

where the ratio of the two rates,

$$RR = \frac{r_1}{r_2} = exp(\beta_1)$$

is often called the rate ratio.

- If

$$RR = \frac{r_1}{r_2} = exp(\beta_1) = 1$$

or, equivalently, if

$$H_o : \beta_1 = log(r_1) - log(r_2) = 0,$$

then the rates in the two cities are the same:

$$H_o : r_1 = r_2 = exp(\beta_0)$$

- We can use maximum likelihood on the computer to get a Wald test that

$$H_o : \beta_1 = log(r_1) - log(r_2) = 0,$$

- With $\widehat{r}_j = \frac{Y_j}{T_j}$, the Wald Statistic is

$$\begin{aligned} Z &= \frac{\widehat{\beta}_1}{\sqrt{\widehat{Var}(\widehat{\beta}_1)}} \\ &= \frac{log(\widehat{r}_1) - log(\widehat{r}_2)}{\sqrt{\widehat{Var}[log(\widehat{r}_1)] + \widehat{Var}[log(\widehat{r}_2)]}} \end{aligned}$$

Estimating Rate Ratio

- We can estimate the rate ratio on the computer as well,

$$\widehat{RR} = \frac{\widehat{r}_1}{\widehat{r}_2} = exp(\widehat{\beta}_1)$$

- One can get a 95% confidence interval on the rate ratio by first getting a confidence interval on the $log(RR)$ scale

$$\beta_1 = \log(RR) = \log\left(\frac{r_1}{r_2}\right)$$

  and then exponentiating the endpoints

- In particular, the 95% CI for the RR is

$$\exp\{[\widehat{\beta}_1] \pm 1.96\sqrt{\widehat{Var}[\widehat{\beta}_1]}\}.$$

Estimation on Computer

- We want to estimate the regression parameters for the model

$$log(r_j) = \beta_0 + \beta_1 x_j,$$

- For city j, we have the number of cases Y_j is Poisson with mean

$$\mu_j = r_j T_j$$

- In particular, μ_j has the log-linear model

$$log(\mu_j) = log(r_j) + log(T_j) = (\beta_0 + \beta_1 x_j) + log(T_j)$$

- We can think of $log(T_j)$ as being a covariate in the model for $log(\mu_j)$, but the coefficient of $log(T_j)$ in the model is 1.
- A covariate with coefficient 1 is called an 'offset'.

- For Cambridge, the model is

$$log(\mu_1) = log(r_1) + log(T_1) = \beta_0 + \beta_1 + log(T_1)$$

- For Boston, the model is

$$log(\mu_2) = log(r_2) + log(T_2) = \beta_0 + log(T_2)$$

- Then, in SAS Proc Genmod, you would use a log-linear model for the number of cases Y_1 and Y_2, with the above log-linear model.
- In Proc Genmod, you specify $log(T_j)$ as an offset.

Example

- Suppose you observe

$$Y_1 = 15$$

cases of leukemia in Cambridge, and

$$Y_2 = 30$$

cases of leukemia in Boston in 1990.

- Suppose the average time than an individual lived in one of the cities in 1990 was the same, and was $.75$ year.
- Suppose there were $T_1 = 500,000$ people living in Cambridge at sometime in 1990 and $T_2 = 3,000,000$ people who lived in Boston at sometime in 1990.
- Then, the person-years for Cambridge are

$$T_1 = (500,000)(.75) = 375,000$$

and for Boston are

$$T_2 = (3,000,000)(.75) = 2,225,000$$

The rates in the two cities are:

```
 Cambridge        Boston
  15/375,000    30/2,225,000
 .0000400        .0000135
```

The Estimated Rate Ratio is:

$$\widehat{RR} = \frac{\widehat{r}_1}{\widehat{r}_2} = \frac{15/375,000}{30/2,225,000} = 2.97$$

Thus, the rate is almost 3 times higher in Cambridge.

SAS PROC GENMOD

```
data one;
input city cases exposure;
 log_T = log(exposure);
cards;
1 15 375000
0 30 2225000
;

proc genmod data=one;
 model cases = city  / link=log  dist = poi offset=log_T ;
 estimate 'logrr' city 1 /exp;
run;
```

/* SELECTED OUTPUT */

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-11.2141	0.1826	-11.5719	-10.8562	3772.66	<.0001
city	1	1.0874	0.3162	0.4676	1.7072	11.83	0.0006

Contrast Estimate Results

Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > Ch
logrr	1.0874	0.3162	0.05	0.4676	1.7072	11.83	0.0
Exp(logrr)	2.9667	0.9381	0.05	1.5962	5.5137		

The Wald Statistic from this output is

$$Z^2 = \left(\frac{\widehat{\beta}_1}{\sqrt{\widehat{Var}(\widehat{\beta}_1)}} \right)^2 = 11.83$$

with $p = 0.0006$ so we reject the null that the expected number of cases are the same.

The estimate RR is obtained as:
`Exp(logrr)= 2.9667`
We also get a 95% CI fr the RR as:

$$[1.5962, 5.5137]$$

Overdispersion

- For data modelled as counts (i.e., Poisson), we have an inherent limitation
- One “feature” of a Poisson R.V. is that the mean = variance = μ
- When the data observed have a variance greater than predicted under the GLM, we have **overdispersion**
- A common cause is subject heterogeneity

Horseshoe crab example:
Suppose, crab width, weight, color and spine condition are the four predictors that affect a female crab’s number of satellites residing nearby (additional male crab partners) (Y). Suppose that Y has a Poisson distribution at each of the fixed combination of those predictors.

If we model Y as a function of only one of the predictors, we would underestimate the variance of Y since the variance if Y is comprised of all predictor combinations.

Consistency of Parameter Estimates

- Even in the presence of overdispersion, the parameter estimates β's are consistent.
- However, the estimated standard errors will be too small (recall the actual variance is greater than the modelled variance)
- Testing for overdispersion in Poisson data is simple.
- The relationship of the Deviance to the model df is the key
- If Deviance / df > 1, then overdispersion may be present
- If Deviance / df < 1, then underdispersion may be present
- Recall a χ^2 with $df = g$ has a mean of g.
- We will develop a test to see what is a statistically significant amount of over/under dispersion (usually overdispersion)

Negative Binomial Distribution

The negative binomial distribution is used when the number of successes is fixed and we're interested in the number of failures before reaching the fixed number of successes.

The **negative binomial distribution** has the PDF

$$f(y;k,\mu) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\left(\frac{k}{\mu+k}\right)^k\left(1-\frac{k}{\mu+k}\right)^y$$

The negative binomial distribution has $E(Y)=\mu$ and $var(Y)=\mu+\mu^2/k$

The index k^{-1} is called the dispersion parameter.

If $k=0$, then we have the Poisson distribution.

Fitting a Negative Binomial Distribution

All that is required to fit a Negative Binomial Model in GENMOD is to specify “dist = nb”

Recall our AIDs death example:

Month Period	Deaths	Month Period	Deaths
1	0	8	18
2	1	9	23
3	2	10	31
4	3	11	20
5	1	12	25
6	4	13	37
7	9	14	45

Selected Results

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	12	29.6535	2.4711
Scaled Deviance	12	29.6535	2.4711
Pearson Chi-Square	12	28.8473	2.4039
Scaled Pearson X2	12	28.8473	2.4039
Log Likelihood		472.0625	

Here, we see a value of 2.4711 for the Deviance / df ratio. There is the potential for overdispersion since D/$df > 1$.

Fitting NB Model

```
proc genmod;
 model y = x /dist=nb link = log;
run;
```

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	12	14.5844	1.2154
Scaled Deviance	12	14.5844	1.2154
Pearson Chi-Square	12	14.0275	1.1690
Scaled Pearson X2	12	14.0275	1.1690
Log Likelihood		474.3380	

Algorithm converged.

Testing $k = 0$

To test

$$H_0 : k = 0$$

vs.

$$H_A : k \neq 0$$

We can use a LRT.

$$\begin{aligned} LRT &= -2(\text{ LL (Poisson) - LL (negative binomial)}) \\ &= -2(472.0625 - 474.3380) \\ &= 4.551 \end{aligned}$$

on 1 df. Thus, we would reject H_0 and conclude that our data is overdispersed using the Poisson model.

The interpretation of results from a NB regression are the same as the Poisson regression.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-0.0378	0.4050	0.01	0.9257
x	1	0.2963	0.0410	52.14	<.0001
Dispersion	1	0.0934	0.0763		

Where the estimated variance of Y would be

$$\widehat{Var(Y)} = \widehat{\mu} + 0.0934\widehat{\mu^2}$$

where $\log \widehat{\mu} = \widehat{\alpha} + \widehat{\beta}x$