Poisson Regression

- Poisson regression is used to model count variables.
- Poisson regression has a number of extensions useful for count models.

Examples of Poisson regression

- The number of awards earned by students at a secondary or high school.

- Predictors of the number of awards earned include the type of program in which the student was enrolled (e.g., vocational, general or academic) and the score on their final exam in math.

Conventional OLS regression

- Count outcome variables are sometimes log-transformed and analyzed using OLS regression.
- Many issues arise with this approach, including loss of data due to undefined values generated by taking the log of zero (which is undefined) and biased estimates.

Poisson Regression : Crabs Example

The crabs data set The crabs data set is derived from Agresti *(2007, Table 3.2, pp.76-77)*. It gives 4 variables for each of 173 female horseshoe crabs.

- **Satellites** number of male partners in addition to the female's primary partner
- **Width** width of the female in centimeters
- **Dark** a binary factor indicating whether the female has dark coloring (yes or no)
- **GoodSpine** a binary factor indicating whether the female has good spine condition (yes or no)

Poisson Regression : Crabs Example

```
require(glm2)
data(crabs)
head(crabs)
summary(crabs[,1:4])
```

Poisson Regression : Crabs Example

```
> head(crabs)
Satellites Width Dark GoodSpine Rep1 Rep2
1          8 28.3   no        no    2    2
2          0 22.5  yes        no    4    5
3          9 26.0   no       yes    5    6
4          0 24.8  yes        no    6    6
5          4 26.0  yes        no    6    8
....
```

Poisson Regression : Crabs Example

```
> summary(crabs[,1:4])
 Satellites         Width         Dark      GoodSpine
 Min.   : 0.000   Min.   :21.0   no :107   no :121
 1st Qu.: 0.000   1st Qu.:24.9   yes: 66   yes: 52
 Median : 2.000   Median :26.1
 Mean   : 2.919   Mean   :26.3
 3rd Qu.: 5.000   3rd Qu.:27.7
 Max.   :15.000   Max.   :33.5
```

Poisson Regression : Crabs Example

Fit a Poisson regression model with the number of Satellites as the outcome and the width of the female as the covariate. What is the multiplicative change in the expected number of crabs for each additional centimeter of width?

```
crabs.pois <- glm2(Satellites ~ Width,
data=crabs, family="poisson")
summary(crabs.pois)
exp(0.164)
```

Poisson Regression : Crabs Example

```
> summary(crabs.pois)

Call:
glm2(formula = Satellites ~ Width,
family = "poisson", data = crabs)

......
......
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.30476    0.54224  -6.095  1.1e-09 ***
Width        0.16405    0.01997   8.216  < 2e-16 ***
---
......
```
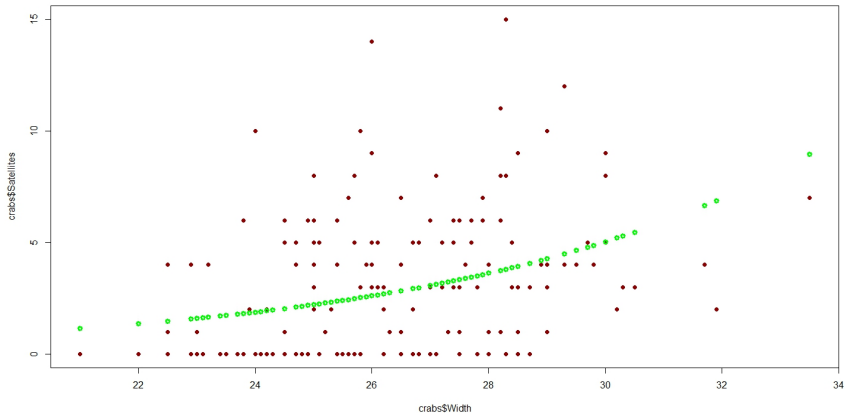
Poisson Regression : Crabs Example

Poisson Regression : Crabs Example

```
plot(crabs$Width, crabs$Satellites,
pch=16, col="darkred")
points(crabs$Width, crabs.pois$fitted.values,
col="green", lwd=3)
```

Poisson Regression with R

Description of the data

- ▶ For the purpose of illustration, we have simulated a data set for the last example.
- ▶ The data set is called *poissonreg.csv*
- ▶ In this example, **num_awards** is the outcome variable and indicates the number of awards earned by students at a high school in a year

Predictor Variables

- **math** is a continuous predictor variable and represents students' scores on their math final exam,

- **prog** is a categorical predictor variable with three levels indicating the type of program in which the students were enrolled.

- **prog** is coded as $1 = "General"$, $2 = "Academic"$ and $3 = "Vocational"$.

Poisson Regression with R

```
         id         num_awards              prog             math
 1     : 1    Min.   :0.00    General   : 45    Min.   :33.0
 2     : 1    1st Qu.:0.00    Academic  :105    1st Qu.:45.0
 3     : 1    Median :0.00    Vocational: 50    Median :52.0
 4     : 1    Mean   :0.63                      Mean   :52.6
 5     : 1    3rd Qu.:1.00                      3rd Qu.:59.0
 6     : 1    Max.   :6.00                      Max.   :75.0
   (Other):194
```
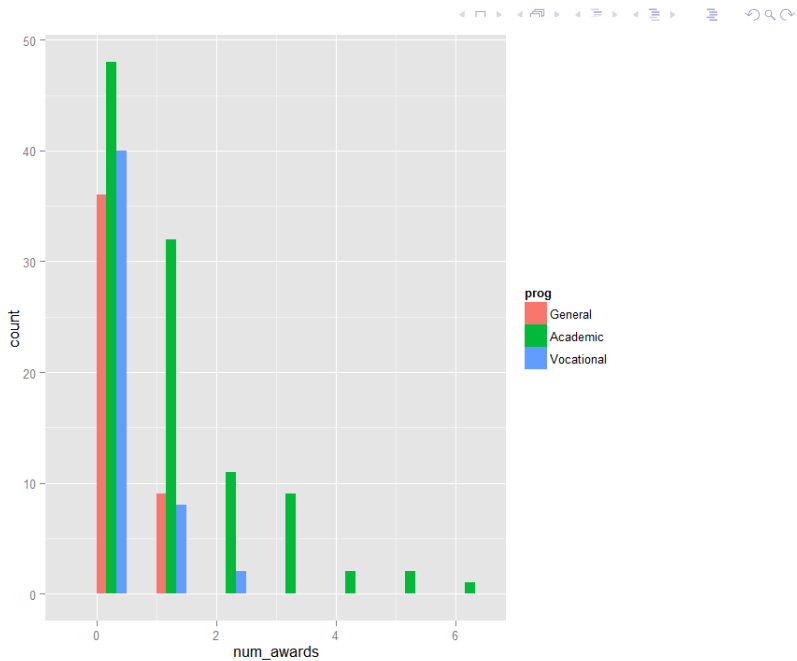
Poisson Regression with R

- ▶ Each variable has 200 valid observations and their distributions seem quite reasonable.
- ▶ The mean and variance of our outcome variable are more or less the same.
- ▶ Our model assumes that these values, conditioned on the predictor variables, will be equal (or at least roughly so).

- Additionally, the means and variances within each level of prog–*the conditional means and variances*–are similar.

- A conditional histogram separated out by program type is plotted to show the distribution.