A Generalized Approach for Many Model Types

- Noting and taking advantage of commonalities among linear models for different response variable types, Nelder and Wedderburn and later McCullagh (UChicago) and Nelder developed **Generalized Linear Models**.
- This approach generalizes many types of models into one framework, unifying theory and estimation methods.
- For each model relating Y to predictors X, one specifies
  - The *link function* $h(\cdot)$, which indicates the relationship between the linear prediction equation and $\mathrm{E}(Y)$;
  - The distribution for the error term ϵ of the model.
- Then, a unified theory and single estimation approach subsumes a wide variety of models.

Note: See Canvas for the examples completed with R

A Generalized Approach for Many Model Types

- A Few of the Several Types of GLMs:

Response	Link Function	Error Term	Model
Continuous ($\approx$ normal)	identity	normal	linear
0/1 discrete	logit	Binomial	logistic
polychotomous discrete	logit	multinomial	multinomial logistic
Integer counts	natural log	Poisson	Poisson regression
real valued, non-negative	inverse	Gamma	survival

Note: See Canvas for the examples completed with R

Regression With Normal Response

- Linear regression
- Response: Continuous, $\approx$ Normal $(E(Y|X), \sigma^2)$
- Link = identity = $h(w) = w$
- Error: Normal (does not depend on predictors)

$$h[E(Y|X)] = E(Y|X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Note: See Canvas for the examples completed with R

Regression With Binary/Bernoulli Response

- Logistic regression
- Response: Binary (0, 1), Bernoulli(p),
  $\mathrm{E}(Y) = p$, $\mathrm{var}(Y) = p(1-p)$
- Link = logit = $h(w) = \log\left(\dfrac{w}{1-w}\right)$
- Error: Bernoulli/binomial (and depends on predictors)

$$h[E(Y|X)] = \log\left[\frac{E(Y|X)}{1-E(Y|X)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

where $E(Y|X) = P(Y = 1|X) = p(X)$
(p is a function of predictors)

Note: See Canvas for the examples completed with R

Regression With Poisson Response

- Poisson regression, log-linear regression
- Response: Poisson(λ), $\mathrm{E}(Y) = \lambda$, $\mathrm{var}(Y) = \lambda$
- Link $= \log_e$ $\quad h(w) = \log_e(w)$
- Error: Poisson (and depends on predictors)

$$h[E(Y|X)] = \log_e(E(Y|X)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

where $E(Y|X) = \lambda(X)$ $\quad$ (λ is a function of the predictors)

Poisson Regression

- **Poisson regression** is used to model **count variables** as outcome. The outcome (i.e., the count variable) in a Poisson regression cannot take on negative values (can equal 0).

- A Poisson regression model is sometimes known as a **log-linear model** (link function is log) and the model takes the form:

$$\log\big(\mathrm{E}(Y|X)\big) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p.$$

CAUTION: This is NOT the same as the OLS transformation

$$\mathrm{E}\big(\log(Y|X)\big) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

In fact, $\log\big(\mathrm{E}(Y|X)\big) \geq \mathrm{E}\big(\log(Y|X)\big)$,
and the latter is OLS using log transformation on Y.

Note: See Canvas for the examples completed with R

Poisson Regression

- In fact, $\mathrm{E}\big(\log(Y|X)\big) \leq \log\big(\mathrm{E}(Y|X)\big)$
- The predicted mean of Poisson model is $\mathrm{E}(Y|X) = \exp\big(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p\big)$.
  - For OLS $\exp\big(\mathrm{E}[\log(Y|X)]\big) \leq E(Y|X)$
    (bias, we tend to underestestimate the mean)
  - For Poisson model (a GLM),
    we model $\log(E(Y|X))$ and hence $E(Y|X)$ directly.

Note: See Canvas for the examples completed with R

- Recall that the mean and variance of Poisson distribution are the same.
- Therefore, a very strong assumption in Poisson regression is – **conditional on the predictors, the conditional mean and variance of outcome are equal.**

Poisson Regression

- **Examples of Poisson observations:**

  1. The number of persons killed by mule or horse kicks in the Prussian army per year. Ladislaus Bortkiewicz collected data from 20 volumes of Preussischen Statistik. These data were collected on 10 corps of the Prussian army in the late 1800s over 20 years.

  2. The number of people in line in front of you at the grocery store. Predictors may include the number of items currently offered at a special discounted price and whether a special event (e.g., a holiday, a big sporting event) is three or fewer days away. Analyses involving queueing frequently involve the Poisson distribution.

  3. The number of awards earned by students at one 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 a mathematics exam.

Note: See Canvas for the examples completed with R

Poisson Regression

We illustrate Poisson regression using Example 3 above:

- `num_awards` is the outcome variable and indicates the number of awards earned by students at a high school in a given year,
- `math` is a continuous predictor variable and represents students' scores on their math final exam, and
- `prog` is a categorical predictor variable with three levels indicating the type of program in which the students were enrolled.

For Poisson regression, we assume that the outcome variable number of awards, conditioned on the predictor variables, will have roughly equal mean and variance.

Poisson Regression - Assumptions

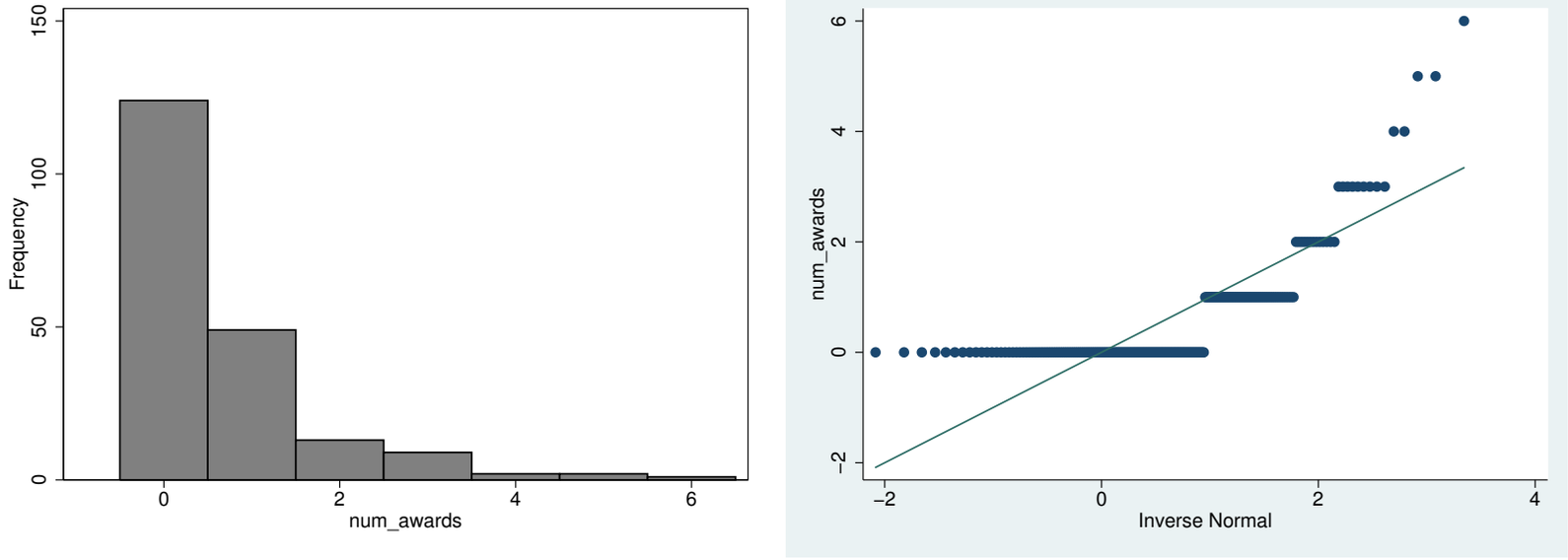
Examining the mean numbers of awards by program type and seems to suggest that program type is a good candidate for predicting the number of awards. Additionally, the means and variances within each level of program – the conditional means and variances – are similar.

```
. use http://statistics.uchicago.edu/~collins/data/STAT224other/poisson_sim, clear
. tabstat num_awards, by(prog) stats(mean sd n)

Summary for variables: num_awards
     by categories of: prog (type of program)

    prog |      mean        sd         N
---------+----------------------------
 general |        .2  .4045199        45
academic |         1  1.278521       105
vocation |       .24  .5174506        50
---------+----------------------------
   Total |       .63  1.052921       200
------------------------------------------
. histogram num_awards, discrete freq
```

```
. qnorm num_awards
```



(a) Histogram of num awards (b) QQ plot

Can we use OLS here? Normality assumption is violated. Count outcome variables are sometimes log-transformed and analyzed using OLS regression. Some issues arise with this approach, for example, more than half of the data (124 students) have **zero** awards (log is undefined)

Poisson Regression - Model and Coefficients

- In Stata, with categories for program (general is baseline or reference group)

```
. poisson num_awards i.prog math

Iteration 0:   log likelihood = -182.75759
Iteration 1:   log likelihood = -182.75225
Iteration 2:   log likelihood = -182.75225

Poisson regression                              Number of obs   =        200
                                                LR chi2(3)      =      98.22
                                                Prob > chi2     =     0.0000
Log likelihood = -182.75225                     Pseudo R2       =     0.2118
------------------------------------------------------------------------------
  num_awards |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        prog |
   academic  |   1.083859    .358253     3.03   0.002     .3816962    1.786022
   vocation  |   .3698092   .4410703     0.84   0.402    -.4946727    1.234291
             |
        math |   .0701524   .0105992     6.62   0.000     .0493783    .0909265
       _cons |  -5.247124   .6584531    -7.97   0.000    -6.537669    -3.95658
------------------------------------------------------------------------------
```

Note: See Canvas for the examples completed with R

Poisson Regression - Model and Coefficients

- Results (βs) are increase/decrease in log(E(counts)) on an additive scale.
- To interpret the coefficients, one needs to take $\exp(\beta)$'s and interpret them as the expected relative change in counts per unit of X change.
- To get relative increase in counts per unit of X on a multiplicative scale, use "irr" (stands for incidence-rate ratio, similar to risk ratio or relative risk):

```
. poisson num_awards i.prog math, irr

Iteration 0:   log likelihood = -182.75759
Iteration 1:   log likelihood = -182.75225
Iteration 2:   log likelihood = -182.75225

Poisson regression                              Number of obs     =        200
```

Note: See Canvas for the examples completed with R

```
                                                LR chi2(3)        =      98.22
                                                Prob > chi2       =     0.0000
Log likelihood = -182.75225                     Pseudo R2         =     0.2118

------------------------------------------------------------------------------
  num_awards |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        prog |
   academic  |   2.956065   1.059019     3.03   0.002     1.464767    5.965674
   vocation  |   1.447458   .6384309     0.84   0.402     .6097705    3.435942
             |
        math |   1.072672   .0113695     6.62   0.000     1.050618    1.095188
       _cons |   .0052626   .0034652    -7.97   0.000     .0014479    .0191284
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

Poisson Regression - Model Fit

To help assess the fit of the model, the "`estat gof`" command can be used to obtain the **goodness-of-fit** χ^2 test. This is **not** a test of the model coefficients, but rather a test of the model form: Does the Poisson model form fit our data? Thus, a large goodness-of-fit p-value indicates the observed and the predicted data are not too different from each other, i.e., a good fit.

```
. estat gof

Deviance goodness-of-fit =  189.4496
Prob > chi2(196)         =    0.6182

Pearson goodness-of-fit  =  212.1437
Prob > chi2(196)         =    0.2040
```

A statistically significant (small p-value) here would indicate that the model does not fit the data well. In that situation, we may try to determine if there are omitted predictor variables, if our linearity assumption holds and/or if the conditional mean and variance of outcome are very different.

Fitting GLMs

An alternative way to fit Poission regression is using the "glm" function (Stata or R), specifying which "family" to use. The default is linear regression and "binomial" is logistic regression (for binary outcome – Bernoulli is a special case of binomial).

```
. glm num_awards math i.prog, family(poisson)

Iteration 0:   log likelihood = -187.46951
Iteration 1:   log likelihood = -182.75816
Iteration 2:   log likelihood = -182.75225
Iteration 3:   log likelihood = -182.75225

Generalized linear models                         No. of obs      =        200
Optimization     : ML                             Residual df     =        196
                                                  Scale parameter =          1
Deviance         =  189.4496199                   (1/df) Deviance =   .9665797
Pearson          =  212.1437315                   (1/df) Pearson  =   1.082366

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   1.867523
Log likelihood   = -182.7522516                   BIC             =  -849.0206
```

Note: See Canvas for the examples completed with R

```
------------------------------------------------------------------------------
             |                 OIM
  num_awards |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        math |   .0701524   .0105992     6.62   0.000     .0493783    .0909265
             |
        prog |
   academic  |   1.083859    .358253     3.03   0.002     .3816961    1.786022
   vocation  |   .3698092   .4410703     0.84   0.402    -.4946727    1.234291
             |
       _cons |  -5.247124   .6584531    -7.97   0.000    -6.537669    -3.95658
------------------------------------------------------------------------------
```

- Estimates are same as earlier. Again, the β's are in log(counts) on an additive scale.
- In a GLM framework, separate computer modules for logistic, Poisson, etc. would not be needed.

Summary – Poisson Regression and GLMs

- Poisson is a useful model for many phenomena, but has a strong theoretical assumption, that conditional mean and variance of the outcome variable are equal.
- When there seems to be an issue of bad fit, we should first check if our model is appropriately specified, such as omitted variables and functional forms.
- The assumption that the conditional variance is equal to the conditional mean should be checked.

  If not reasonably equal, there are alternative variations on Poisson regression that may work.