PBHS 32100 / STAT 22401 Winter 2025 J. Dignam

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

Recall that in linear regression, the (conditional) mean of the response Y is related to covariates directly via the linear function $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p$. The variance on the prediction is from the Gaussian (normal) distribution

A Generalized Approach for Many Model Types

Another approach:

For each model relating Y to predictors$X$, one can specify

- The *link function* $h(\cdot)$, which specifies the relationship between the linear prediction equation ($X\beta$, or the *linear predictor*) and $\mathrm{E}(Y|X)$, the conditional mean of Y
- The probability distribution for the error term ϵ of the model, equivalently, the variance of Y

Then, a unified theory and single estimation approach subsumes a wide variety of models

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
Integer counts	natural log	Poisson	Poisson
Integer counts	natural log	negative binomial	negative binomial
0/1 discrete	logit	binomial	logistic
polychotomous discrete	logit	multinomial	multinomial logistic
real valued, non-negative	inverse	gamma	survival (time to event)

- **Note:** link function addresses "How does the linear predictor $X\beta$ relate to $E(Y)$?"

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 (but can equal 0).

Poisson Distribution:

The probability distribution function of Y is:

$$\Pr(Y = Y) = \frac{e^{-\lambda}\lambda^{y}}{x!}, y = 0, 1, 2, \ldots$$

A single parameter defines the Poisson distribution:

$$\begin{aligned} \mathrm{E}(Y) &= \lambda \quad (> 0) \\ \mathrm{var}(Y) &= \lambda \end{aligned}$$

Poisson Distribution

A **Poisson random variable** is an (integer) count variable over a large population relative to the number of events

Example: Suppose that, on average, there are 3 fatal traffic accidents in Chicago on a holiday weekend

- let random variable Y = the number of fatalities during the holiday
- the parameter λ is the mean of Y, $\rightarrow$ here, $\lambda = 3$

  What is the probability of 5 fatalities during the holiday?

$$\Pr(Y = 5) = \frac{e^{-3} 3^5}{5!} = 0.101$$

3 fatalities?

$$\Pr(Y = 3) = \frac{e^{-3} 3^3}{3!} = 0.224$$

Poisson Regression

A Poisson regression model is sometimes known as a **log-linear model**, and it takes the form:

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

Note that

$$\log\big(\mathrm{E}(Y|X)\big) \neq \mathrm{E}\big(\log(Y|X)\big)$$

- The latter is OLS using log transformation on Y, as we examined earlier.
- The predicted mean of the Poisson model on the count scale is

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

A strong assumption in Poisson regression is that **conditional on the predictors, the mean and variance of the outcome are equal, i.e., following the Poisson dist'n.**

Examples of Poisson observations, random variables

1. The number of persons per year killed by mule or horse kicks, as collected from 20 volumes of Preussischen Statistik on 10 Prussian army corps in the late 1800s (Bortkiewicz, 1898).
2. The number of people in line at the grocery store. Predictors may include the time of day, whether a special event (e.g., holiday, big sporting event) is three or fewer days away, etc. Problems involving queueing theory frequently involve the Poisson distribution
3. The number of awards earned by students in a high school. Predictors include the type of program in which students were enrolled (vocational, general or academic), exam scores, and other factors.

Data are recorded as event counts in some sample size N, and usually N >> events.

Table 2. Frequency distributions for Bortkewitsch's full table

Number of deaths	Observed frequency	Expected frequency, as given by Bortkewitsch and by Keynes	Expected Poisson frequency, as given by Jeffreys
0	144	143·1	139·0
1	91	92·1	97·3
2	32	33·3	34·1
3	11	8·9	8·0
4	2	2·0	1·4
5+	0	0·6	0·2
Total	280	280·0	280·0

If each of the 280 counts of numbers of deaths could reasonably be thought to be independent of all the others, and the number of cavalry officers and their susceptibility to death from horse-kicks could reasonably be thought to be the same for each of the 280 units of observation, then a simple Poisson model for the observed frequencies would be reasonable. The expected frequencies for a Poisson distribution with mean 196/280=0·700 were given by Jeffreys and are reproduced here in Table 2; these show good agreement with the observed frequencies. Table 2 also reproduces the expected frequencies given by Bortkewitsch and quoted by Keynes; these were obtained by fitting a Poisson model to the data for each corps and then summing the expected values across the corps (e.g. Winsor, 1947, p. 158).

Bortkewitsch (1898, p. 24) noted that the four corps denoted G, I, VI and XI had numerical compositions that were particularly far from the average. He therefore excluded these four corps, to give the observed frequencies in our Table 3, for which the total number of deaths is 122.

Table 3 also contains the expected Poisson frequencies as obtained by Bortkewitsch himself and by Fisher (1925, Section 15, Table 4) for a Poisson distribution with mean 122/200=0·610. The agreement between observed and expected is very good indeed for the smaller data-set.

Table 3. Frequency distribution excluding corps G, I, VI and XI

Number of deaths	Observed frequency	Expected Poisson frequency as obtained by Bortkewitsch and Fisher
0	109	108·7
1	65	66·3
2	22	20·2
3	3	4·1
4	1	0·6
5+	0	0·1
Total	200	200·0

However, a generalised linear model with logarithmic link function and Poisson errors for the observations and with terms for corps and years may be fitted to the corps-by-years table of counts. Goodness-of-fit for these two terms may be summarised by an analysis of deviance (McCullagh & Nelder, 1983, p. 17).

Poisson Regression
Examples of Poisson observations, random variables

A second major use of Poisson regression in Public Health and Epidemiology is in relation to disease incidence over time

- We are interested in disease counts in relation to exposure time. Many deleterious exposures, as well as natural factors such as aging, will have bearing on the event count and must be accounted for when, say, comparing groups.
- Thus, rather than denominator N for a sample, the relevant denominator is the sum of exposure time over all N individuals, known as person-*time* Rather than proportions, we have rates per unit of time at risk.
- This approach also accommodates different lengths of at risk time that may naturally occur.

We will review these types of Poisson models later

Poisson Regression

We illustrate Poisson regression using Example 3 above (school awards):

- `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.

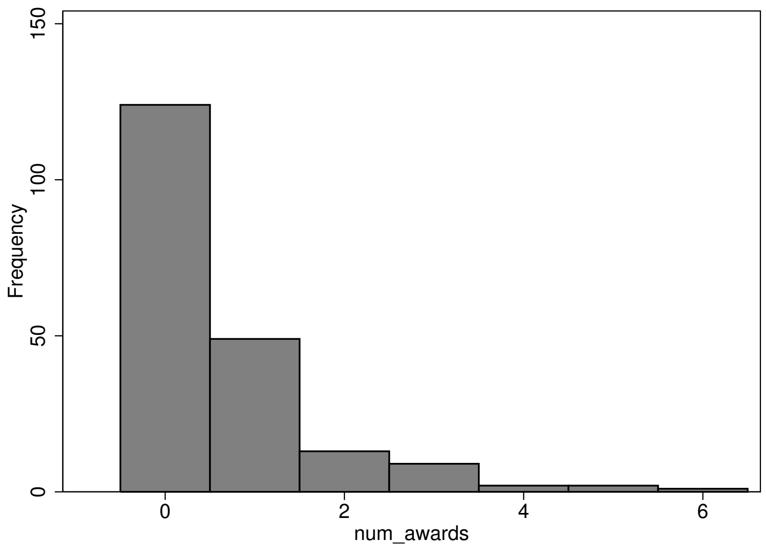
For Poisson regression, we assume that the outcome variable number of awards, conditioned on the predictor variables, will have roughly equal mean and variance.

Poisson Regression - Assumptions

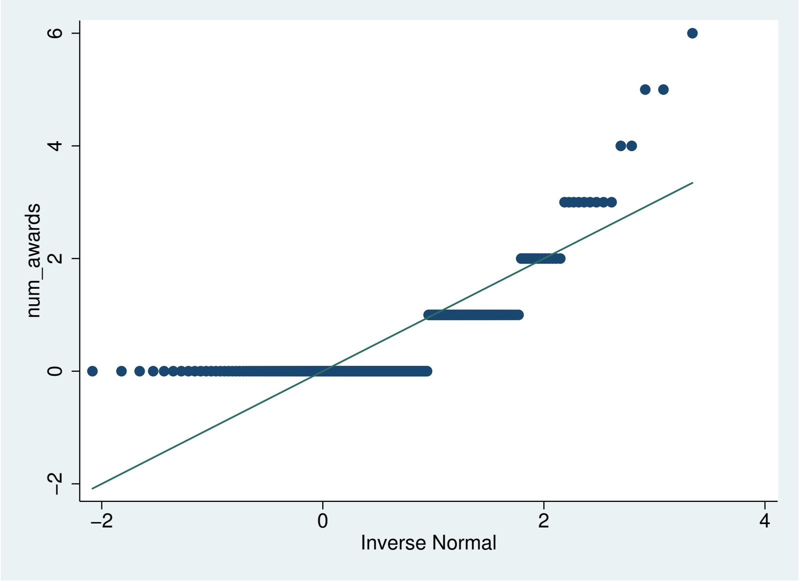
Examining the mean numbers of awards by program type suggests that program type is a good candidate predictor. Additionally, the means and variances are similar within each program (Poisson assumption).

```
. use http://www.ats.ucla.edu/stat/stata/dae/poisson_sim, clear
. sum num_awards

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
  num_awards |       200         .63    1.052921          0          6

. 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 not met. Count outcome variables are sometimes log-transformed and analyzed using OLS regression. However, more than half of the data (124 students) have zero awards

Poisson Regression - Null model

```
. poisson num_awards
. . .
Poisson regression                                Number of obs     =        200
                                                  LR chi2(0)        =       0.00
                                                  Prob > chi2       =          .
Log likelihood = -231.86356                       Pseudo R2         =     0.0000
------------------------------------------------------------------------------
  num_awards |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -.4620355   .0890871    -5.19   0.000    -.6366429    -.287428
------------------------------------------------------------------------------
.* output on counts scale
. poisson num_awards, irr
. . .
------------------------------------------------------------------------------
  num_awards |  Inc. Rate   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |        .63   .0561249    -5.19   0.000     .5290656    .7501906
------------------------------------------------------------------------------
```

constant term here is just mean overall. First model is on natural log(mean counts) scale, second is on mean counts scale

Poisson Regression - Program type as a Predictor

- categories for program (‘general program’ is baseline/reference group)

```
. poisson num_awards acad voc

Iteration 0:   log likelihood = -205.26518
Iteration 1:   log likelihood = -205.25743
Iteration 2:   log likelihood = -205.25743

Poisson regression                                Number of obs     =        200
                                                  LR chi2(2)        =      53.21
                                                  Prob > chi2       =     0.0000
Log likelihood = -205.25743                       Pseudo R2         =     0.1147
------------------------------------------------------------------------------
  num_awards |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        acad |   1.609438   .3473253     4.63   0.000     .9286925    2.290183
         voc |   .1823213   .4409585     0.41   0.679    -.6819415    1.046584
       _cons |  -1.609438   .3333333    -4.83   0.000    -2.262759   -.9561164
------------------------------------------------------------------------------
```

Poisson Regression - Program type as a Predictor

coefficients are used to predict log of mean counts by group. Note that

a. $\exp(\beta_0) = \exp(-1.6094) = .200$ - mean awards for the general education group (the reference group here)

b. $\exp(\beta_0 + \beta_{voc}) = \exp(-1.6094 + .1823) = .24$ - mean awards for the vocational group

c. $\exp(\beta_0 + \beta_{acad}) = \exp(-1.6094 + 1.6084) = 1.0$ - mean awards for the academic group

These are the same means for the general, vocational, and academic programs as shown in table earlier.

Tests shown are comparisons to reference (general ed.) group

Poisson Regression - Program type as a Predictor

Same model on the mean count scale. The β coefficients here are the incidence rate ratios (IRR)

```
. poisson num_awards acad voc, irr
. . .

Poisson regression                                Number of obs     =        200
                                                  LR chi2(2)        =      53.21
                                                  Prob > chi2       =     0.0000
Log likelihood = -205.25743                       Pseudo R2         =     0.1147
------------------------------------------------------------------------------
  num_awards |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        acad |   4.999999   1.736626     4.63   0.000     2.531197    9.876743
         voc |        1.2   .5291501     0.41   0.679     .5056343    2.847906
       _cons |   .2000001   .0666667    -4.83   0.000      .104063    .3843828
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

Here, the coefficient for voc is the ratio of mean counts for vocational vs general; coefficient for $acad$ is the ratio of means for academic vs general. Coefficients give the *multiplicative* effect

The tests vs reference group are the same as before

Poisson Regression - Adding (continuous) Math Score to Model

```
. poisson num_awards acad voc 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]
-------------+----------------------------------------------------------------
        acad |   1.083859    .358253     3.03   0.002     .3816962    1.786022
         voc |   .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
------------------------------------------------------------------------------
```

Poisson Regression - Model and Coefficients

Results (βs) are increase/decrease in log(counts) on an additive scale. Again, to get relative increase in counts per unit of X on a multiplicative scale, we request the incidence rate ratio:

```
. poisson num_awards acad voc math, irr

. . .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 |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        acad |   2.956065   1.059019     3.03   0.002     1.464767    5.965674
         voc |   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 and Coefficients

```
------------------------------------------------------------------------------
  num_awards |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        acad |   2.956065   1.059019     3.03   0.002     1.464767    5.965674
         voc |   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
------------------------------------------------------------------------------
```

This model indicates:

- academic program has a 2.95 fold greater mean awards than general program. Smaller than before after adjusting for continuous math score
- vocational program has a nonsignificant 1.45 fold greater mean awards than general program
- per point of math score, mean awards goes up by a small but significant amount - 1.07 or about 7%

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, large p-value indicates good fit.

```
. estat gof

         Goodness-of-fit chi2  =  189.4496
         Prob > chi2(196)      =    0.6182
```

A statistically significant (small p-value) here would indicate that the data do not fit the model 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 (i.e., not Poisson data)

Fitting GLMs

An alternative way to fit Poission regression is using the “glm” function (Stata or R), specifying which “family” to use, default is linear regression and “binomial” is logistic regression (for binary outcome).

```
. glm num_awards math acad voc, 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
```

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

- Estimates and tests are same as earlier. Again, βs are are in log(counts) on an additive scale.

- Models fit by separate computer modules for logistic, Poisson, etc can all be fit in a GLM framework and should give same answer.

Poisson Regression with Continuous Predictors

- The previous example, without the continuous math score, could be accommodated by frequency table methods for estimation and testing, although this would get unwieldy with more categorical predictors forming a multidimensional table
- We were able to add a continuous predictor, which cannot be represented by frequencies unless we ‘bin’ the scores into some categories.
- We can have any combination of categorical, ordinal, and continuous predictors in the Poisson model
- **Ex** How does approval of new drugs for chronic diseases relate to disease prevalence and monetary expenditure? New drug approvals are relatively uncommon and in ‘count’ form

Poisson Regression with Continuous Predictors

The data (1990s- mid 2000s, from C &H):

```
. list, noobs clean

             Disease_Area   drugs    prev   expend
    Ischemic Heart Disease      6    8976    198.4
               Lung Cancer      3     874     80.2
                  HIV/AIDS     21    1303   1049.6
               Alcohol Use      2   18092    222.6
  Cerebrovascular Disease       2    9467    108.5
                      COPD      1    4271     48.9
                Depression      7   12785    149.5
                  Diabetes     13   37850    278.4
           Osteoarthritis       5   12345    151.3
               Drug abuse       1    4000    442.1
                  Dementia      9    8931    344.1
                    Asthma      3   15919     41.8
              Colon Cancer      2    1926     70.6
           Prostate Cancer      4    2020     40.1
             Breast Cancer      9    2262    159.5
          Bipolar Disorder      2    2418       35
```

Correlations

```
             |  drugs    prev  expend
-------------+--------------------------
       drugs |  1.0000
        prev |  0.1961  1.0000
      expend |  0.7850 -0.0474  1.0000
```

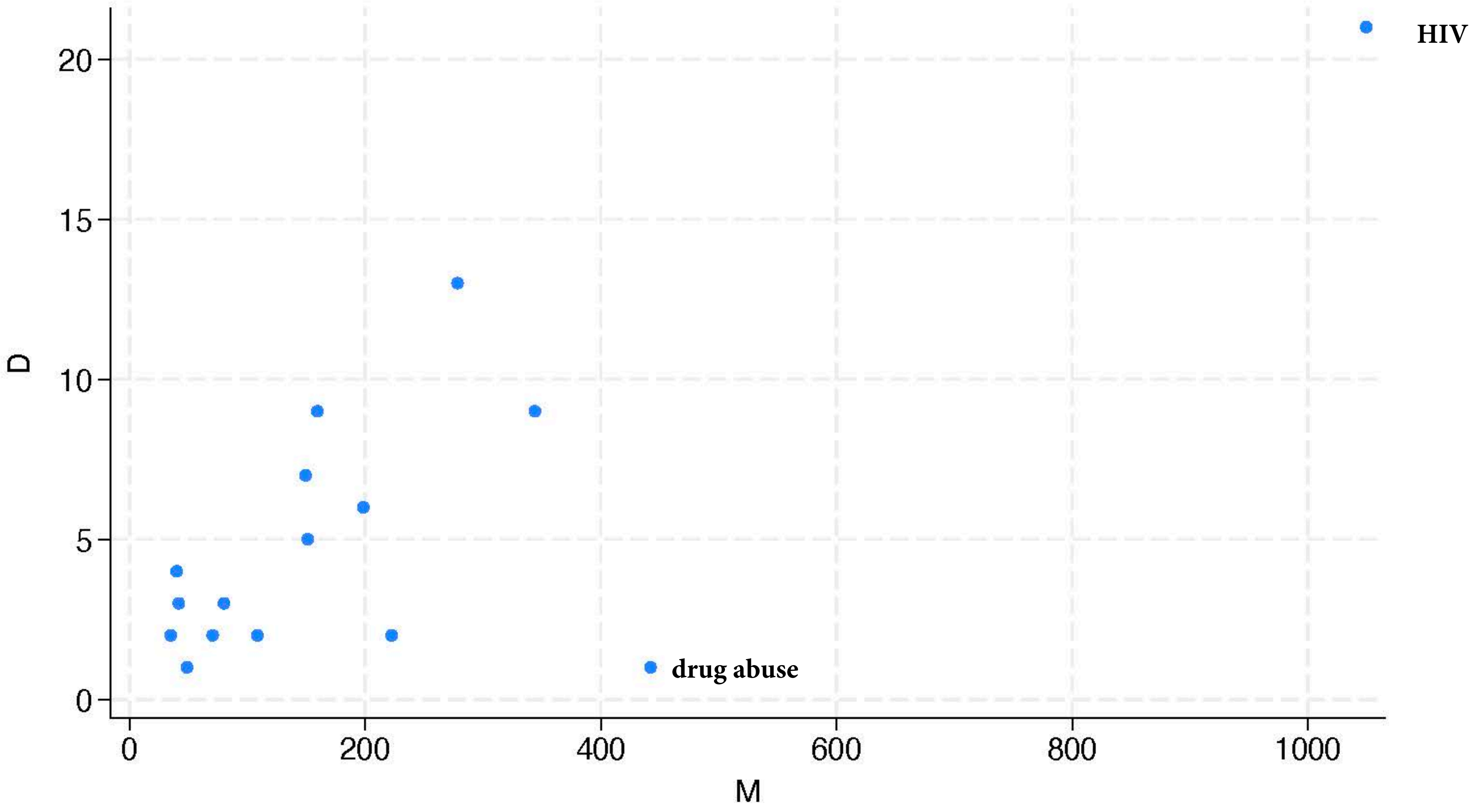
prev: prevalence per 100,000

expend: dollars in millions

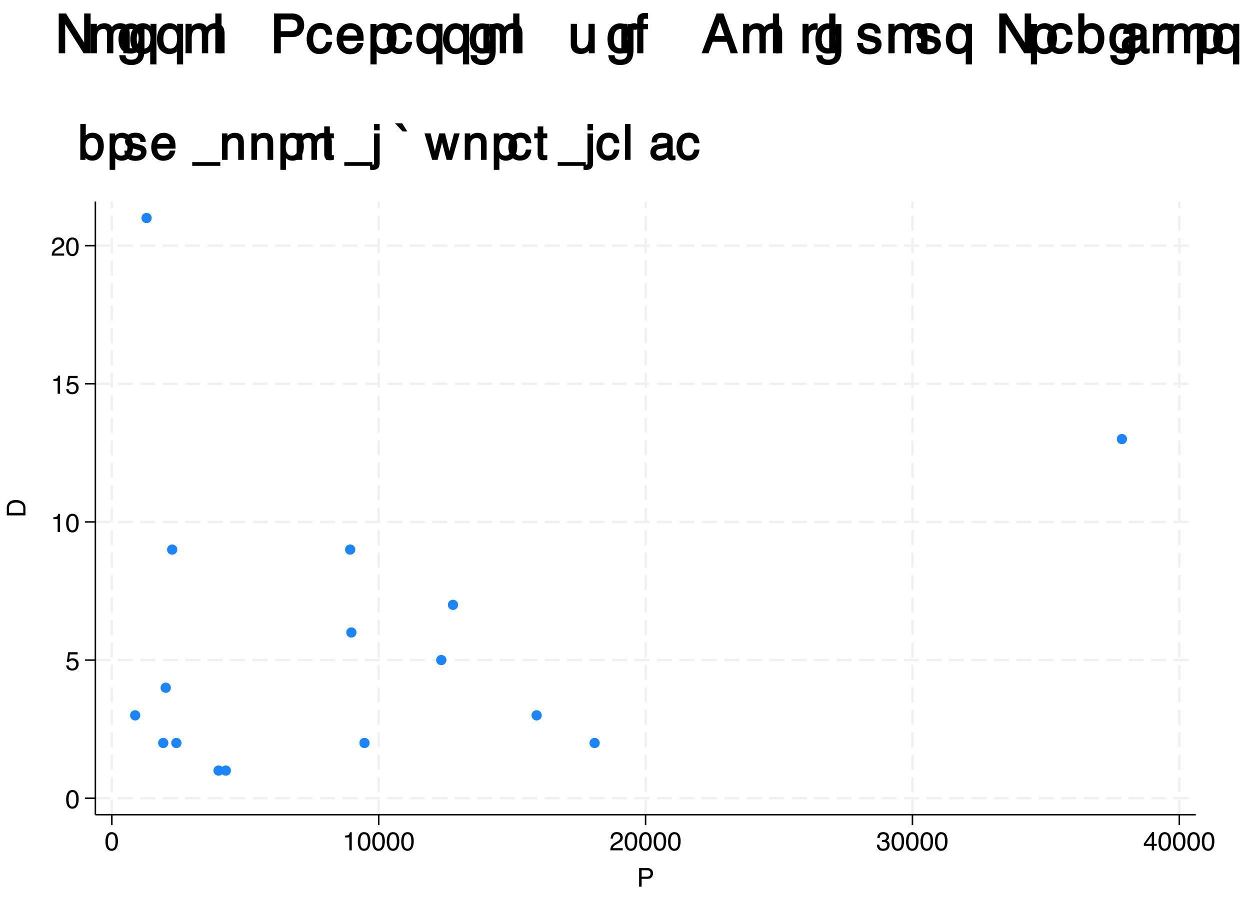
Poisson Regression with Continuous Predictors

The data (1990s- mid 2000s, from C &H):

. . scatter drugs expend



Plot shows trend towards increasing approvals with expenditure


Nmgqqml Pcepcqqgml ugf Aml rgsmsq Npcbgampq
bpse _nnpmt _j ` wnpct _jcl ac
D
0
5
10
15
20
0
10000
20000
30000
40000
P

Poisson Regression - drug approvals

Model on log(counts) scale:

```
. poisson drugs prev expend

Iteration 0:   log likelihood = -38.407767
Iteration 1:   log likelihood =  -38.07115
Iteration 2:   log likelihood = -38.070036
Iteration 3:   log likelihood = -38.070036

Poisson regression                                Number of obs   =         16
                                                  LR chi2(2)      =      38.88
                                                  Prob > chi2     =     0.0000
Log likelihood = -38.070036                       Pseudo R2       =     0.3381

------------------------------------------------------------------------------
       drugs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        prev |    .000027   9.51e-06     2.84   0.005     8.37e-06    .0000456
      expend |    .001998   .0003008     6.64   0.000     .0014084    .0025877
       _cons |   .8777568   .2073627     4.23   0.000     .4713334     1.28418
------------------------------------------------------------------------------
```

Poisson Regression

counts scale:

```
. poisson drugs prev expend, irr

Iteration 0:   log likelihood = -38.407767
Iteration 1:   log likelihood =  -38.07115
Iteration 2:   log likelihood = -38.070036
Iteration 3:   log likelihood = -38.070036

Poisson regression                                Number of obs   =         16
                                                  LR chi2(2)      =      38.88
                                                  Prob > chi2     =     0.0000
Log likelihood = -38.070036                       Pseudo R2       =     0.3381

------------------------------------------------------------------------------
       drugs |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        prev |   1.000027   9.51e-06     2.84   0.005     1.000008    1.000046
      expend |      1.002   .0003014     6.64   0.000     1.001409    1.002591
       _cons |   2.405498   .4988105     4.23   0.000     1.602129    3.611706
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

Poisson Regression

Predictions from the model:

```
. predict dhat


. list disease drugs dhat, clean

                 disease    drugs       dhat
  1.  Ischemic Heart Disease    6    4.55644
  2.             Lung Cancer    3   2.890991
  3.                HIV/AIDS   21   20.28902
  4.             Alcohol Use    2    6.11686
  5. Cerebrovascular Disease    2   3.858106
  6.                    COPD    1    2.97662
  7.              Depression    7   4.579959
  8.                Diabetes   13   11.65872
  9.         Osteoarthritis    5   4.542173
 10.              Drug abuse    1    6.48245
 11.                Dementia    9   6.088735
 12.                  Asthma    3   4.019379
 13.            Colon Cancer    2   2.917786
 14.         Prostate Cancer    4   2.752262
 15.           Breast Cancer    9   3.516701
 16.        Bipolar Disorder    2   2.753795
```

Fitting GLMs - considering alternate models

We have have circumstances where Poisson model is not correct for count data, for example:

- When the variance exceeds the mean, we have an *overdispersed* Poisson random variable - which may be better modeled by the *negative binomial* distribution
- When we have more than the expected number of cases with count of zero, we have a *zero-inflated* Poisson, a hybrid model that Stata or R can fit.

We examine the dataset relating school absence days to various factors including mathematics exam scores (Notes on transformations) to contrast some alternate models, all of which can be fit by a GLM procedure

Alternate Models

Looking at mean & variance of the response - not Poisson?

```
. sum daysabs
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     daysabs |       314    5.955414    7.036958          0         35
. by prog: sum daysabs

----------------------------------------------------------------------------
-> prog = 1
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     daysabs |        40       10.65    8.201157          3         34
----------------------------------------------------------------------------
-> prog = 2
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     daysabs |       167    6.934132    7.446304          0         35
----------------------------------------------------------------------------
-> prog = 3
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     daysabs |       107    2.672897    3.733519          0         19
```

Variances appear larger than means here

Fit the Poisson model

```
. poisson daysabs math prog2 prog3

Iteration 0:   log likelihood = -1328.6751
Iteration 1:   log likelihood = -1328.6425
Iteration 2:   log likelihood = -1328.6425

Poisson regression                                Number of obs     =        314
                                                  LR chi2(3)        =     443.73
                                                  Prob > chi2       =     0.0000
Log likelihood = -1328.6425                       Pseudo R2         =     0.1431

------------------------------------------------------------------------------
     daysabs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        math |  -.0068084   .0009311    -7.31   0.000    -.0086332   -.0049835
       prog2 |  -.4398975    .056672    -7.76   0.000    -.5509725   -.3288224
       prog3 |  -1.281364   .0778898   -16.45   0.000    -1.434025   -1.128703
       _cons |   2.651974   .0607367    43.66   0.000     2.532932    2.771015
------------------------------------------------------------------------------
```

Alternate Models

Fit the Negative Binomial model (note: for this dist'n, variance increases as mean increases)

```
. nbreg daysabs math prog2 prog3

Fitting Poisson model:
Iteration 0:   log likelihood = -1328.6751
Iteration 1:   log likelihood = -1328.6425
Iteration 2:   log likelihood = -1328.6425

Fitting constant-only model:
Iteration 0:   log likelihood = -899.27009
Iteration 1:   log likelihood = -896.47264
Iteration 2:   log likelihood = -896.47237
Iteration 3:   log likelihood = -896.47237

Fitting full model:
Iteration 0:   log likelihood = -870.49809
Iteration 1:   log likelihood = -865.90381
Iteration 2:   log likelihood = -865.62942
Iteration 3:   log likelihood =  -865.6289
Iteration 4:   log likelihood =  -865.6289
```

```
Negative binomial regression                      Number of obs     =        314
                                                  LR chi2(3)        =      61.69
Dispersion     = mean                             Prob > chi2       =     0.0000
Log likelihood =  -865.6289                       Pseudo R2         =     0.0344
------------------------------------------------------------------------------
     daysabs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        math |   -.005993   .0025072    -2.39   0.017     -.010907    -.001079
       prog2 |    -.44076    .182576    -2.41   0.016    -.7986025   -.0829175
       prog3 |  -1.278651   .2019811    -6.33   0.000    -1.674526    -.882775
       _cons |   2.615265   .1963519    13.32   0.000     2.230423    3.000108
-------------+----------------------------------------------------------------
    /lnalpha |  -.0321895   .1027882                     -.2336506    .1692717
-------------+----------------------------------------------------------------
       alpha |   .9683231   .0995322                      .7916384    1.184442
------------------------------------------------------------------------------
LR test of alpha=0: chibar2(01) = 926.03              Prob >= chibar2 = 0.000
```

Note: Poisson is a special case when $\alpha = 0$ **- test above is for** $H_0 : \alpha = 0$**, which is rejected** (BTW: test stat. is 2(difference in log likelihoods between models) or 2(-865.6289 - -1328.6425) = 926.03).

Fitting same models using GLMs

```
. glm daysabs math prog2 prog3, family(nbinomial)

Iteration 0:   log likelihood = -873.19828
. . .
Iteration 3:   log likelihood = -865.67793

Generalized linear models                          Number of obs   =        314
Optimization     : ML                              Residual df     =        310
                                                   Scale parameter =          1
Deviance         =  350.9751541                    (1/df) Deviance =   1.132178
Pearson          =  331.1757302                    (1/df) Pearson  =   1.068309
Variance function: V(u) = u+(1)u^2                 [Neg. Binomial]
Link function    : g(u) = ln(u)                    [Log]
                                                   AIC             =    5.53935
Log likelihood   = -865.6779288                    BIC             =  -1431.337
------------------------------------------------------------------------------
     daysabs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        math |  -.0059875   .0025416    -2.36   0.018    -.0109689   -.0010061
       prog2 |  -.4407535   .1852477    -2.38   0.017    -.8038322   -.0776747
       prog3 |  -1.278633   .2047766    -6.24   0.000    -1.679988   -.8772782
       _cons |   2.615011   .1991968    13.13   0.000     2.224593     3.00543
------------------------------------------------------------------------------
```

```
. glm daysabs math prog2 prog3, family(Poisson)

Iteration 0:   log likelihood = -1349.4476
. . .
Iteration 3:   log likelihood = -1328.6425

Generalized linear models                          Number of obs   =        314
Optimization     : ML                              Residual df     =        310
                                                   Scale parameter =          1
Deviance         =  1773.953438                    (1/df) Deviance =    5.72243
Pearson          =   2045.65589                    (1/df) Pearson  =    6.59889
Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]
                                                   AIC             =   8.488169
Log likelihood   = -1328.642493                    BIC             =  -8.358387
------------------------------------------------------------------------------
             |                 OIM
     daysabs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        math |  -.0068084   .0009311    -7.31   0.000    -.0086332   -.0049835
       prog2 |  -.4398975    .056672    -7.76   0.000    -.5509725   -.3288224
       prog3 |  -1.281364   .0778898   -16.45   0.000    -1.434025   -1.128703
       _cons |   2.651974   .0607367    43.66   0.000     2.532932    2.771015
------------------------------------------------------------------------------
```

Fitting GLMs

linear models on two response scales - raw and square root

```
. glm daysabs math prog2 prog3, family(Gaussian)

Iteration 0:   log likelihood = -1029.5558
Generalized linear models                          Number of obs   =        314
Optimization     : ML                              Residual df     =        310
                                                   Scale parameter =   41.78862
Deviance         =  12954.47351                    (1/df) Deviance =   41.78862
Pearson          =  12954.47351                    (1/df) Pearson  =   41.78862
Variance function: V(u) = 1                        [Gaussian]
Link function    : g(u) = u                        [Identity]
                                                   AIC             =   6.583158
Log likelihood   = -1029.555844                    BIC             =   11172.16
------------------------------------------------------------------------------
             |                 OIM
     daysabs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        math |  -.0435858   .0150508    -2.90   0.004    -.0730848   -.0140868
       prog2 |   -3.81316   1.138453    -3.35   0.001    -6.044487   -1.581833
       prog3 |  -7.384937   1.215351    -6.08   0.000     -9.76698   -5.002893
       _cons |   12.60373   1.224692    10.29   0.000     10.20338    15.00409
------------------------------------------------------------------------------
```

```
. glm sq_daysabs math prog2 prog3, family(Gaussian)

Iteration 0:   log likelihood = -517.57805
Generalized linear models                          Number of obs   =        314
Optimization     : ML                              Residual df     =        310
                                                   Scale parameter =   1.602587
Deviance         =  496.8019054                    (1/df) Deviance =   1.602587
Pearson          =  496.8019054                    (1/df) Pearson  =   1.602587
Variance function: V(u) = 1                        [Gaussian]
Link function    : g(u) = u                        [Identity]
                                                   AIC             =   3.322153
Log likelihood   = -517.5780451                    BIC             =   -1285.51
------------------------------------------------------------------------------
             |                 OIM
  sq_daysabs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        math |  -.0086047   .0029474    -2.92   0.004    -.0143815   -.0028279
       prog2 |  -.8568447   .2229446    -3.84   0.000    -1.293808   -.4198813
       prog3 |  -1.726006   .2380035    -7.25   0.000    -2.192484   -1.259528
       _cons |   3.447068   .2398328    14.37   0.000     2.977004    3.917131
------------------------------------------------------------------------------
```

- **Note:** This is identical to ordinary MLR model mentioned in last lecture

Summary – Poisson Regression and GLMs

A Poisson data-based model is useful 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. There are alternative variations on Poisson regression that may work. Inference and Interpretation are the same - predicting counts and count ratios by covariates