Poisson Regression with Incidence Rates

Poisson regression is used to model a non-negative integer **count variables** as the outcome. As mentioned earlier, in Public Health and Epidemiology disease incidence is often studied in relation to exposure over 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 when comparing groups

Incidence Rates and Ratios

For incidence rate data

- data come in form of events and pyrs, there is no other 'N'. Data may be aggregated already (table form) or listed as individual cases with follow-up time (in which case there will be an 'N'
- latter form is needed for continuous exposure variables

Summary table for an epidemiological study of some exposure

	Exposure		Total
	yes	no	
case counts	c_e	c_u	$c_e + c_u = C$
person-years	pyr_e	pyr_u	$pyr_e + pyr_u = PYR$
Incidence rate	$IR_e = c_e/pyr_e$	$IR_u = c_u/pyr_u$	$IR = C/PYR$

Incidence Rates and Ratios

The following are cardiovascular disease events in relation to two potential risk factors, age group and obesity (y/n). The data (from SPRM Ch 9):

```
. list, clean noobs

   AgeGroup   Obesit~s   cvd   pyrs   agegrp   obstat
      60{64      Obese    10    245        1        1
      60{64   Notobese    12    640        1        0
      65{69      Obese    34    365        2        1
      65{69   Notobese    45    520        2        0
      70{74      Obese    40    250        3        1
      70{74   Notobese    44    490        3        0
```

If we consider the obesity factor only, and sum the data to fill the table, we have

	Obesity		
	yes	no	Total
CVD cases	84	101	185
person-years	860	1650	2510
Incidence rate	.0976744	.0612121	.0737052

Previously we saw the ratio of count proportions as a natural summary from the Poisson model, showing how counts increase on a multiplicative scale

The incidence rate ratio is defined similarly as

$$\frac{IR_e}{IR_u} = \frac{.0977}{.0612} = 1.5956$$

Indicating about a 1.6 fold excess risk for the obese group

The Incidence Rate Ratio

STATA has modules to perform these types of epidemiologic summaries/analyses functions. The *incidence rate* command:

```
. ir cvd obstat pyrs

                 | obstat                 |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
             cvd |        84         101  |        185
            pyrs |       860        1650  |       2510
-----------------+------------------------+------------
  Incidence rate |  .0976744    .0612121  |   .0737052
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Inc. rate diff. |         .0364623       |    .0124039    .0605207
 Inc. rate ratio |         1.595671       |    1.180226     2.15264 (exact)
 Attr. frac. ex. |         .3733045       |    .1527047     .535454 (exact)
 Attr. frac. pop |         .1695004       |
                 +-------------------------------------------------
```

Poisson Regression Model (and other GLMs) - a note on model fitting

- The Poisson model and other GLMs are fit via *maximum likelihood estimation*. Least squares is an MLE estimator, but only for linear regression

- the *Likelihood Function* $L = f(Y_1, Y_2, \ldots, Y_n|\omega)$ is a key quantity. It is the joint distribution of all the observations (data fixed), expressed as a function of all the parameters (ω). The method finds the values of the parameters (including βs) that most likely gave rise to the data. This is done by taking derivatives of log(L) with respect to parameters, setting equal to zero and solving.

- After the fit, L is a quantity (computed by plugging in those parameters) that can be used to compare models, measure 'fit', etc. We usually work with the *log likelihood*, provided with every model run.

Poisson Regression Model for Incidence Rates

We can fit the model as in the count case. An important difference is the inclusion of the 'exposure' variable to indicate the person-years

```
. poisson cvd obstat, exposure(pyrs) irr
. .
Iteration 1:   log likelihood = -42.981143

Poisson regression                                Number of obs     =          6
                                                  LR chi2(1)        =       9.79
                                                  Prob > chi2       =     0.0018
Log likelihood = -42.981143                       Pseudo R2         =     0.1022
------------------------------------------------------------------------------
         cvd |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      obstat |   1.595671   .2356291     3.16   0.002     1.194671     2.13127
       _cons |   .0612121   .0060908   -28.07   0.000     .0503663    .0743935
   ln(pyrs) |          1  (exposure)
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

The non-obese incidence rate and multiplicative effect of obesity (i.e., the IRR) are produced.

```
------------------------------------------------------------------------------
         cvd |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      obstat |   1.595671   .2356291     3.16   0.002     1.194671     2.13127
       _cons |   .0612121   .0060908   -28.07   0.000     .0503663    .0743935
    ln(pyrs) |          1  (exposure)
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

Note that $.0612$ or $6.1/100$ person-years is the incidence rate in the non-obese and .6125 $\times$ 1.5956 $=$ .09767 or $9.8/100$ person-years is the rate in the obese group, reproducing the numbers in the earlier table.

Poisson Model for Incidence Rates - add an ordinal predictor

We can add the age group variable as an ordinal predictor. Examine CVD by age:

```
. tabstat cvd pyrs, by(AgeGroup) stat(sum)
Summary statistics: sum  by categories of: AgeGroup

AgeGroup |       cvd      pyrs
---------+--------------------
   60{64 |        22       885
   65{69 |        79       885
   70{74 |        84       740
---------+--------------------
   Total |       185      2510
------------------------------
. display 22/885
  .02485876

. display 79/885
  .08926554

. display 84/740
  .11351351
```

Poisson Model for Incidence Rates - add an ordinal predictor

```
. poisson cvd obstat agegrp, exposure(pyrs) irr

Iteration 0:   log likelihood = -21.297411
Iteration 1:   log likelihood = -21.297215
Iteration 2:   log likelihood = -21.297215

Poisson regression                                Number of obs     =          6
                                                  LR chi2(2)        =      53.15
                                                  Prob > chi2       =     0.0000
Log likelihood = -21.297215                       Pseudo R2         =     0.5551

------------------------------------------------------------------------------
         cvd |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      obstat |   1.535171   .2267304     2.90   0.004     1.149323    2.050554
      agegrp |   1.871195   .1845068     6.35   0.000     1.542366     2.27013
       _cons |   .0162154   .0040498   -16.50   0.000      .009939    .0264555
   ln(pyrs) |          1  (exposure)
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

Model here is on IRR scale

Poisson Model for Incidence Rates - Meaning of Predictor Coefficients

exp(coefficient) yield the relative increase/decrease (i.e the IRR) for change in exposure level

a. $cons = .0162$ - estimated mean CVD rate for agegrp = 0 (not particularly useful)

b. $\beta_{agegrp} = 1.87$ - Indicating a 1.87-fold increase in CVD rate per age group increase (1x for group 1, 2x for group 2, 3x for group 3)

c. $\beta_{obstat} = 1.535$ - Indicating a 1.54-fold increase in CVD for obese vs. nonobese- similar to effect seen earlier before adding age

Rates produced for three age groups are:

Nonobese: 0.036, 0.067, 0.127

Obese: 0.055, 0.102, 0.195

Poisson Regression - Categorical Predictor for Age

```
. poisson cvd obstat age2 age3, exposure(pyrs) irr

Iteration 0:   log likelihood = -17.107802
. . .
Iteration 3:   log likelihood = -17.083097
Poisson regression                                Number of obs   =          6
                                                  LR chi2(3)      =      61.58
                                                  Prob > chi2     =     0.0000
Log likelihood = -17.083097                       Pseudo R2       =     0.6432

------------------------------------------------------------------------------
         cvd |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      obstat |   1.468678   .2180134     2.59   0.010     1.097924    1.964629
        age2 |   3.399674   .8229289     5.06   0.000     2.115409    5.463615
        age3 |   4.453633   1.067596     6.23   0.000     2.784004    7.124574
       _cons |   .0220038   .0048363   -17.36   0.000     .0143025    .0338521
   ln(pyrs) |          1  (exposure)
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

Poisson Regression - Categorical Predictor for Age

cvd	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
obstat	1.468678	.2180134	2.59	0.010	1.097924	1.964629
age2	3.399674	.8229289	5.06	0.000	2.115409	5.463615
age3	4.453633	1.067596	6.23	0.000	2.784004	7.124574
_cons	.0220038	.0048363	-17.36	0.000	.0143025	.0338521
ln(pyrs)	1	(exposure)				

Note: _cons estimates baseline incidence rate.

This model may be closer to table-based rates, but different here because separate by obesity

Rates produced for three age groups are:

Nonobese: 0.022, 0.075, 0.098

Obese: 0.032, 0.109, 0.144

Incidence Rates with Individual Case Data

The CVD data was aggregated by groups (age, obesity) with event counts in the groups

- We can also work with individual case data, where each individual either does or does not have the event (coded 0 or 1) over some recorded follow-up time for that individual
- Total person-time and the tally of events produces the overall rate
- Person-time and event counts within covariate groups produces rates according to covariates

Incidence Rates with Individual Case Data

Ex/ Endometrial cancer among women receiving tamoxifen/placebo in a breast cancer clinical trial.

- Tamoxifen is an anti-estrogen that is highly effective in treatment of breast cancer. Tamoxifen binds to estrogen receptors on the tumor and arrests growth.
- Approved for use across a spectrum of the disease from metastatic to early stage and DCIS, and even in cancer prevention among women deemed at high risk
- Tamoxifen is associated with increased risk of endometrial (uterine) cancer. The benefits are generally considered to outweigh the risk for treatment. In prevention, it is a larger concern

Incidence Rates with Individual Case Data

Data from the NSABP B-14 tamoxifen trial (N-, ER+ breast cancer): Endometrial cancer was uncommon (16 cases) relative the the studied group (N > 2800) and amount of follow-up time (10+ years):

```
. list id trt menstat age bmi endo timefree, clean

      id    trt   menstat   age    bmi   endo  timefree
. . . .
. . . .
     359     1         3    59   23.2      0      189.8
     362     1         1    36   27.4      0       41.3
     364     1         3    55   18.6      0      122.4
     366     1         3    55   30.6      0        9.9
     367     2         3    58   23.4      1         95
     371     1         2    49   22.1      0       39.5
     374     2         3    63   25.1      0      199.5
. . . .
. . . .
```

Incidence Rates with Individual Case Data

The data summarized via Incidence Rate analysis (*timefree* is the exposure time)

```
. ir endo trtx timefree
Incidence-rate comparison
                 |         trtx           |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+-----------
            endo |        13           3  |         16
        timefree |   90322.5     79946.1  |   170268.6
-----------------+------------------------+-----------
  Incidence rate |  .0001439    .0000375  |    .000094
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Inc. rate diff. |         .0001064       |    .0000174    .0001954
 Inc. rate ratio |         3.835513       |    1.053989    20.98384 (exact)
 Attr. frac. ex. |         .7392787       |    .0512237    .9523443 (exact)
 Attr. frac. pop |         .6006639       |
                 +-------------------------------------------------
```

Incidence Rates with Individual Case Data

Use Poisson model:

```
. poisson endo trtx, exposure(timefree)
Iteration 0:   log likelihood = -93.429861
Iteration 1:   log likelihood = -93.428424
Iteration 2:   log likelihood = -93.428424

Poisson regression                                Number of obs     =      1,395
                                                  LR chi2(1)        =       5.58
                                                  Prob > chi2       =     0.0182
Log likelihood = -93.428424                       Pseudo R2         =     0.0290

------------------------------------------------------------------------------
        endo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        trtx |   1.344303   .6405126     2.10   0.036     .0889214    2.599685
       _cons |   -10.1905   .5773502   -17.65   0.000    -11.32208    -9.05891
ln(timefree) |          1  (exposure)
------------------------------------------------------------------------------

. display exp(1.344303)
3.8355123
```

Incidence Rates with Individual Case Data

Poisson approach allows incorporation of covariates:

```
.  poisson endo trtx age  bmi, exposure(timefree) ir

Iteration 0:   log likelihood = -92.189405
Iteration 1:   log likelihood = -92.187682
Iteration 2:   log likelihood = -92.187682

Poisson regression                              Number of obs     =      1,395
                                                LR chi2(3)        =       8.06
                                                Prob > chi2       =     0.0448
Log likelihood = -92.187682                     Pseudo R2         =     0.0419
------------------------------------------------------------------------------
        endo |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        trtx |   3.715235   2.381092     2.05   0.041     1.057917    13.04731
         age |   1.032185   .0297447     1.10   0.272     .9755024    1.092161
         bmi |   1.040999   .0452663     0.92   0.355     .9559539    1.133609
       _cons |   2.23e-06   4.38e-06    -6.62   0.000     4.71e-08    .0001052
ln(timefree) |          1  (exposure)
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

Poisson Regression - Testing individual Coefficients

We can test the value of individual covariates as in SLR/MLR. We test

$$H_0 : IRR = e^{\beta} = 1$$

which is is same as

$$H_0 : \log(IRR) = \beta = 0$$

vs.

$$H_A : \log(IRR) \neq \beta = 0$$

Theory from GLMs and the estimation method shows that β is $\approx$ Normal. The test statistic (reported in the `STATA/R` output) for the above hypothesis is:

$$Z = \frac{\hat{\beta}\ (-0)}{\hat{\text{se}}(\hat{\beta})}$$

Poisson Regression - Testing the Whole Model

When we execute the model, the following summary precedes the

```
. poisson cvd obstat age2 age3, exposure(pyrs) irr

Iteration 0:   log likelihood = -17.107802
. . .
Iteration 3:   log likelihood = -17.083097

Poisson regression                                Number of obs     =          6
                                                  LR chi2(3)        =      61.58
                                                  Prob > chi2       =     0.0000
Log likelihood = -17.083097                       Pseudo R2         =     0.6432
```

The likelihood ratio test (LR above, equaling 61.58)) is s function of the difference in likelihoods for this model vs. a *null* model with no predictors. **this is tantamount to the overall F test in linear regression.**

To illustrate, we can run the two models and contrast

```
.* RUN null model
. poisson cvd, exposure(pyrs) irr

Iteration 0:   log likelihood = -47.874227
Iteration 1:   log likelihood = -47.874227

Poisson regression                                Number of obs     =          6
                                                  LR chi2(0)        =      -0.00
                                                  Prob > chi2       =          .
Log likelihood = -47.874227                       Pseudo R2         =    -0.0000

. * SAVE it
. est store nul
.
.* RUN full model
. poisson cvd obstat age2 age3, exposure(pyrs) irr
Iteration 0:   log likelihood = -17.107802
. . .
Iteration 3:   log likelihood = -17.083097

Poisson regression                                Number of obs     =          6
                                                  LR chi2(3)        =      61.58
                                                  Prob > chi2       =     0.0000
Log likelihood = -17.083097                       Pseudo R2         =     0.6432
. . .
```

```
. * SAVE it
. est store big
.
.* CONTRAST these models
. lrtest big nul

Likelihood-ratio test                                 LR chi2(3)  =     61.58
(Assumption: nul nested in big)                       Prob > chi2 =    0.0000
```

The above computation (test statistic is)

$$D = -2\{\log\hat{L_n} - \log\hat{L_b}\}$$

which here is

$$-2\{\ -47.874 - (-17.083)\} = 61.58$$

This is compared to a χ^2 statistic with degrees of freedom equal to the number of parameters (3 here). Result is to reject the null hypothesis that the likelihood ratio is 1. This means (some) predictors matter.

Poisson Regression - Testing Subsets of Parameters

Contrasting different models follows the same strategy - contrasting the likelihood values between models.

- For example, to test the contribution of the categorical age variable (2 indicators), we can run the two models and compute as above.

- Alternatively, we can use a post-estimation test as done earlier in SLR/MLR. This is a different test (Wald test) but inference should be similar to LR test.

Poisson Regression - Testing Subsets of Parameters

```
. poisson cvd obstat age2 age3, exposure(pyrs) irr
. . .
Iteration 3:   log likelihood = -17.083097

Poisson regression                                Number of obs   =          6
                                                  LR chi2(3)      =      61.58
                                                  Prob > chi2     =     0.0000
Log likelihood = -17.083097                       Pseudo R2       =     0.6432
------------------------------------------------------------------------------
         cvd |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      obstat |   1.468678   .2180134     2.59   0.010     1.097924    1.964629
        age2 |   3.399674   .8229289     5.06   0.000     2.115409    5.463615
        age3 |   4.453633   1.067596     6.23   0.000     2.784004    7.124574
       _cons |   .0220038   .0048363   -17.36   0.000     .0143025    .0338521
   ln(pyrs) |          1  (exposure)
------------------------------------------------------------------------------
. test age2 age3
 ( 1)  [cvd]age2 = 0
 ( 2)  [cvd]age3 = 0

           chi2(  2) =   38.89
         Prob > chi2 =    0.0000
```

Poisson Regression - Model and Coefficients

The LR test would use the same steps as earlier:

1. Run larger model (with age group vars and obesity), save log likelihood value (it is -17.083097)
2. Run model omitting age group vars, save likelihood value (it is -42.981143)
3. Compute D = -2(-42.981143 - (-17.083097)) = 51.80
4. Result is larger than χ^2 with 2 df, then age variable(s) are significantly contributing to model fit

Both the Wald and LR test conclude: keep age variables

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

         Deviance goodness-of-fit =   3.49824
         Prob > chi2(2)           =    0.1739

         Pearson goodness-of-fit  =  3.525841
         Prob > chi2(2)           =    0.1715
```

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.

Poisson Regression - Model Fit

To see the basis of the fit test, we can look at observed vs, predicted values:

```
. predict cvd_pred
(option n assumed; predicted number of events)
. list AgeGroup ObesityStatus cvd cvd_pred, noobs clean

    AgeGroup   Obesit~s   cvd   cvd_pred
       60{64      Obese    10    7.91759
       60{64   Notobese    12    14.0824
       65{69      Obese    34   40.10115
       65{69   Notobese    45   38.89889
       70{74      Obese    40   35.98157
       70{74   Notobese    44   48.01822

. gen chipart = (cvd - cvd_pred)^2 / (cvd_pred)
. tabstat chipart, stat(sum)
    variable |       sum
-------------+----------
     chipart |   3.52584
```

Note: This is the Pearson χ^2 sum that we compute for χ^2 tests in frequency tables

Fitting as a GLM

An alternative way to fit Poission regression is using the "glm" function (Stata or R), specifying which "family" to use (default is linear regression).

```
. glm cvd obstat age2 age3, family(Poisson) exposure(pyrs)

Iteration 0:   log likelihood = -17.619973
Iteration 1:   log likelihood = -17.083841
Iteration 2:   log likelihood = -17.083097
Iteration 3:   log likelihood = -17.083097

Generalized linear models                          Number of obs   =          6
Optimization     : ML                              Residual df     =          2
                                                   Scale parameter =          1
Deviance         =  3.498240465                    (1/df) Deviance =    1.74912
Pearson          =  3.525841336                    (1/df) Pearson  =   1.762921

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

                                                   AIC             =   7.027699
Log likelihood   = -17.08309669                    BIC             =  -.0852785
------------------------------------------------------------------------------
```

```
             |                 OIM
         cvd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      obstat |   .3843624   .1484419     2.59   0.010     .0934215    .6753032
        age2 |   1.223679   .2420611     5.06   0.000     .7492484    1.698111
        age3 |    1.49372   .2397136     6.23   0.000      1.02389     1.96355
       _cons |  -3.816539    .219792   -17.36   0.000    -4.247323   -3.385754
   ln(pyrs) |          1  (exposure)
------------------------------------------------------------------------------
```

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

- The overall fit statistics as well as other measures are provided.

Fitting as a GLM

The model executed in R:

```
> library(foreign)
> cvd <- read.dta("CVD_factors.dta")
> cvd
  AgeGroup ObesityStatus cvd pyrs agegrp obstat age2 age3
1    60{64         Obese  10  245      1      1    0    0
2    60{64      Notobese  12  640      1      0    0    0
3    65{69         Obese  34  365      2      1    1    0
4    65{69      Notobese  45  520      2      0    1    0
5    70{74         Obese  40  250      3      1    0    1
6    70{74      Notobese  44  490      3      0    0    1
>.
>
> Pois <- glm(cvd ~ offset(log(pyrs)) + obstat + age2 + age3, data=cvd, family=poisson)
>
>
> Pois

Call:  glm(formula = cvd ~ offset(log(pyrs)) + obstat + age2 + age3,
    family = poisson, data = cvd)
```

```
Coefficients:
(Intercept)       obstat         age2         age3
    -3.8165       0.3844       1.2237       1.4937

Degrees of Freedom: 5 Total (i.e. Null);  2 Residual
Null Deviance:     65.08
Residual Deviance: 3.498  AIC: 42.17
```

Poisson Regression - summary measures

In SLR/MLR, we have a partition of variability captured by R^2. In Poisson regression (and other log-linear models), analogues to R^2 have been sought. One simple one is

$$R^2_{pseud} = 1 - \frac{\log L(\hat{\boldsymbol{\beta}})}{\log \hat{L}_0}$$

where $\log L(\hat{\boldsymbol{\beta}})$ is the log likelihood for the current model and $\log \hat{L}_0$ is the log likelihood for the null model.

- Generally, these measures are not as reliable as fits measures as in the linear regression setting (although R^2 can be misleading there too)

Poisson Regression - summary measures

Ex/ for the full model (age groups and obesity vs. null model we have

$$R^2_{pseud} = 1 - \frac{-17.083096}{-47.874227}$$

$$= 1 - 0.3568 = 0.6432$$

As given in the output for the model of interest:

```
. poisson cvd obstat age2 age3, exposure(pyrs) irr
. . .

Poisson regression                                Number of obs   =          6
                                                  LR chi2(3)      =      61.58
                                                  Prob > chi2     =     0.0000
Log likelihood = -17.083097                       Pseudo R2       =     0.6432
. . . .
```

Alternate Models when Poisson does not appear to be correct model

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

```
. nbreg cvd obstat age2 age3, exposure(pyrs)

Fitting Poisson model:

Iteration 0:   log likelihood = -17.107802
Iteration 1:   log likelihood = -17.083109
Iteration 2:   log likelihood = -17.083097
Iteration 3:   log likelihood = -17.083097

Fitting constant-only model:

Iteration 0:   log likelihood = -26.909909
Iteration 1:   log likelihood = -25.989194
Iteration 2:   log likelihood = -25.570977
Iteration 3:   log likelihood = -25.567453
Iteration 4:   log likelihood = -25.567453

Fitting full model:
```

```
Iteration 0:   log likelihood = -22.857883
Iteration 1:   log likelihood = -19.909726
Iteration 2:   log likelihood =  -19.73484  (not concave)
Iteration 3:   log likelihood = -19.347292  (not concave)
Iteration 4:   log likelihood = -17.635909
Iteration 5:   log likelihood = -17.197221
Iteration 6:   log likelihood = -17.112039
Iteration 7:   log likelihood = -17.089736
Iteration 8:   log likelihood = -17.084648
Iteration 9:   log likelihood = -17.083448
Iteration 10:  log likelihood = -17.083167
Iteration 11:  log likelihood = -17.083109
Iteration 12:  log likelihood = -17.083098
Iteration 13:  log likelihood = -17.083096  (not concave)
Iteration 14:  log likelihood = -17.083096

Negative binomial regression                      Number of obs     =          6
                                                  LR chi2(3)        =      16.97
Dispersion     = mean                             Prob > chi2       =     0.0007
Log likelihood = -17.083096                       Pseudo R2         =     0.3318

------------------------------------------------------------------------------
         cvd |      Coef.   Std. Err.      z    P>|z|     [95 Conf. Interval]
-------------+----------------------------------------------------------------
```

```
      obstat |   .3843709    .148442     2.59   0.010     .0934299    .6753118
        age2 |   1.223679   .2420612     5.06   0.000     .7492478     1.69811
        age3 |   1.493717   .2397137     6.23   0.000     1.023887    1.963547
       _cons |  -3.816542   .2197921   -17.36   0.000    -4.247327   -3.385758
    ln(pyrs) |          1  (exposure)
-------------+----------------------------------------------------------------
    /lnalpha |  -18.60187   1339.439                     -2643.855    2606.651
-------------+----------------------------------------------------------------
       alpha |   8.34e-09   .0000112                             0           .
------------------------------------------------------------------------------
LR test of alpha=0: chibar2(01) = 6.8e-07              Prob >= chibar2 = 0.500
```

Note: Poisson is a special case when $\alpha = 0$ - test above is for $H_0 : \alpha = 0$, which is NOT rejected. This means Poisson model is adequate. Parameters are very similar to Poisson model fit earlier.

Summary – Poisson Regression and GLMs

The Poisson data-based model provided a strong inferential and data exploratory tool for events in relation to person-time of exposure

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 form, and then consider variations on the model that may fit better

The Poisson provides a bridge between linear regression, discrete event counts, and rates of failure, which relates to time to event (survival) data

Summary – Courses Beyond Linear Regression

This course introduces a few common variations on linear regression. Further development of these methods and others are extensively covered in other University of Chicago courses such as:

- *Categorical (Discrete) Data Analysis* - all types of discrete outcome variables - binomial, multinomial, ordinal, counts
- *Generalized Linear Models* - all GLMs
- *Biostatistical Methods* - logistic, Poisson, hazard (failure rate, time to event) in health science context
- *Applied Survival Analysis* - analysis of time to event data, including regression methods for these data
- *Applied Longitudinal Data Analysis* - extending regression to repeated measurements on Y