Chapter 9

Poisson Regression Model

Aim: Upon completing the chapter, the learner should be able to estimate the magnitude of the association between disease and exposure, controlling for potential confounders, using a Poisson regression model.

9.1 Model Definition

The Poisson regression model allows us to assess epidemiological studies when the main outcome is an integer number, such as the number of cancer patients, number of immunized children, or number of hospitalized patients. It is assumed that these types of outcomes, Y, are random variables with positive integers and are distributed as the *Poisson probability distribution.* This probability distribution is characterized by one parameter, identified by the Greek letter μ, which is the expected number of events of the outcome of interest and is expressed in the following manner:

$$\Pr\left[Y = y\right] = \frac{e^{-\mu}\mu^{y}}{y!}$$

where the variance of Y is equal to the expected value of Y, $\mathrm{Var}(Y) = E(Y) = \mu$.

The Poisson regression model is used in cohort studies to estimate the expected value of Y among exposure and nonexposure groups, adjusted for the effect of potential confounders. An example of this would be the expected number of lung cancer cases in a 5-year period among smokers and nonsmokers, adjusted for the effects of age and sex (Fox, 2008; Hoffmann, 2004). One advantage of this model is that it can be used to obtain an estimate of the *relative risk (RR),* adjusted for potential confounders. The simplest form of a Poisson regression model, with an exposure variable (E) and one potential confounding variable (C), can be defined with any of the following equivalent expressions:

$$[\text{i}] \qquad \mu_i = T_i * I_i = T_i * e^{\beta_0 + \beta_E * E + \beta_C * C + \beta_{EC} * (EC)}$$

$$[\text{ii}] \qquad I_i = \mu_i / T_i = e^{\beta_0 + \beta_E * E + \beta_C * C + \beta_{EC} * (EC)}$$

$$[\text{iii}] \quad \text{Ln}(\mu_i) = \text{Ln}(T_i) + \beta_0 + \beta_E * E + \beta_C * C + \beta_{EC} * (EC)$$

$$[\text{iv}] \quad \text{Ln}(I_i) = \beta_0 + \beta_E * E + \beta_C * C + \beta_{EC} * (EC)$$

where:

μ_i indicates the expected value for the outcome variable

I_i represents the incidence (the expected cases by time unit or population under the *i*th condition)

T_i represents the sum of the times in the study under the *i*th condition

E denotes the exposure variable

C denotes the effect of the confounder variable

*E***C* denotes the interaction between the exposure and confounder

β_j denotes the coefficients (parameters) associated with the *j*th predictor variables ($j = E$, C, or $E * C$); this value represents the expected changes in the natural logarithm of μ_i in the expression [iii]

β_0 represents the constant term (intercept) in the model

The expression $\text{Ln}(T_i)$ denotes the natural logarithm of T_i under the expression [iii] of the Poisson regression model, which is included as a predictor variable with a coefficient or parameter equal to 1. This type of predictor is identified as an *offset* and has a fixed parameter.

9.2 Relative Risk

In the event that the outcome variable *Y* indicates the occurrence of new cases of a disease in a cohort study, we can determine the incidence of this disease among different groups of exposure through a Poisson regression model, as follows (assuming *E* is dichotomous and *C* is continuous):

Exposure *(E* = 1*)*

$$I_{\text{exp}} = \frac{\mu_{\text{exp}}}{T_{\text{exp}}} = e^{\beta_0 + \beta_E + \beta_C * C + \beta_{EC} * C}$$

Nonexposure *(E* = 0*)*

$$I_{\text{non-exp}} = \frac{\mu_{\text{non-exp}}}{T_{\text{non-exp}}} = e^{\beta_0 + \beta_C * C'}$$

If $C = C'$, then

$$\mathrm{RR} = \frac{I_{\text{exp}}}{I_{\text{non-exp}}} = e^{\beta_E + \beta_{E^*C}{}^*C}$$

In the case of nonsignificant interaction terms (H_0: $\beta_{E^*C} = 0$), we can obtain the adjusted RR using the following:

$$\mathrm{RR}_{\text{adjusted}} = \frac{I_{\text{exp}}}{I_{\text{non-exp}}} = e^{\beta_E^*}$$

where β_E^* is obtained from the model that excludes the interaction term. If the interaction term is significant, it is necessary to estimate the RR in different population subgroups defined by the levels of C.

9.3 Parameter Estimation

The procedure for estimating the unknown coefficients in the Poisson regression model is similar to the procedure of logistic regression for obtaining the maximum-likelihood estimates. Under the assumption that the observations are independent, the *likelihood function* for the Poisson regression model is expressed as follows:

$$L(\beta) = \prod_{i=1}^{K} \frac{e^{-\mu_i}\mu_i^y}{y!}$$

where μ_i is the expected value of Y under the ith condition in the Poisson regression model. The coefficients βs that produce the highest value of this likelihood function are the maximum-likelihood estimates (MLEs) for this model. Based on the MLEs estimates, we can also estimate the RRs and test the statistical hypothesis, with the approach similar to that performed for the logistic regression model.

9.4 Example

Suppose we are interested in assessing the difference in the incidence of cardiovascular disease by *sex*, controlling for *age*. Available data for this purpose can be extracted from the epidemiological cohort study of Framingham (Massachusetts), which started in 1948 with a sample of 5,127 subjects, aged 30–62 years old. The following table summarizes the incidence of cardiovascular disease by age and sex:

Age	CVD_M	PY_M	I_M	CVD_F	PY_F	I_F	$RR_{M:F}$
<46	43	7,370	5.83	9	9,205	0.98	5.97
46–55	163	12,649	12.89	71	16,708	4.25	3.03
56–60	155	7,184	21.58	105	10,139	10.36	2.08
61–80	443	15,015	29.50	415	24,338	17.05	1.73
>80	19	470	40.43	50	1,383	36.15	1.12
Total	823	42,688	19.28	650	61,773	10.52	1.83

Note: M, male; F, female; CVD, cardiovascular disease; PY, person-years; and I, incidence per 1,000 persons in 1 year.

The last column of the table shows the RRs between males and females by age group. The observed trend in these RRs indicates that, in the older age groups, the RRs are getting close to 1; therefore, the incidences of cardiovascular disease by sex are quite different for the younger age groups and quite similar for the older age groups. This trend suggests that age has a modifying effect on the relationship between sex and cardiovascular disease (Szklo and Nieto, 2004).

9.5 Programming the Poisson Regression Model

To evaluate the expected number of cases in the Framingham Study by sex and age using the Poisson model, we need to prepare the database as follows:

```
     +-------------------------+
     | age   sex      py   cvd |
     |-------------------------|
  1. |   1     1    7370    43 |
  2. |   1     0    9205     9 |
  3. |   2     1   12649   163 |
  4. |   2     0   16708    71 |
  5. |   3     1    7184   155 |
     |-------------------------|
  6. |   3     0   10139   105 |
  7. |   4     1   15015   443 |
  8. |   4     0   24338   415 |
  9. |   5     1     470    19 |
 10. |   5     0    1383    50 |
     +-------------------------+
```

Note:

- age indicates the code of the age group (1: <46 years; 2: 46–55 years; 3: 56–60 years; 4: 61–80 years; 5: >80 years)
- sex indicates the code of the sex (0 = female, 1 = male)
- py indicates person-years
- cvd indicates the number of cardiovascular disease cases

9.6 Assessing Interaction Terms

According to the trend in the RRs (by age group) that was observed in the previous table, which suggests the presence of an age–sex interaction, we will initially explore the interaction terms in the Poisson regression model using the likelihood ratio test, as follows:

```
. quietly xi: glm cvd i.sex*i.age, fam(poi) lnoff(py)

. estimates store model1

. quietly xi: glm cvd i.sex i.age, fam(poi) lnoff(py)

. lrtest model1 .

Likelihood-ratio test                       LR chi2(4)    =   29.58
(Assumption: . nested in model1)            Prob > chi2   =  0.0000
```

Note: lnoff(py) indicates the inclusion of the natural logarithm of the py variable as an *offset* variable.

The results indicate a significant age–sex interaction term (P-value < .0001), confirming that age has a modifying effect on the relationship that exists between sex and cardiovascular disease. Therefore, it is necessary to estimate the RR (male vs. female) by age group. To carry out this evaluation, we use the Poisson regression model, while also including interaction terms. For example, the resulting models of previous data can be programmed in Stata with the following command:

```
xi: glm cvd i.sex*i.age, fam(poisson) lnoff(py)
```

Output

```
Generalized linear models              No. of obs      =         10
Optimization      : ML                 Residual df     =          0
                                       Scale parameter =          1
Deviance          =  5.31308e-13       (1/df) Deviance =          .
Pearson           =  4.76108e-13       (1/df) Pearson  =          .

Variance function: V(u) = u            [Poisson]
```

```
Link function    : g(u) = ln(u)     [Log]

                                     AIC              =  8.241062
Log likelihood   = -31.20531135      BIC              =  5.31e-13

------------------------------------------------------------------------------
             |                 OIM
         cvd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Isex_1 |  1.786304    .3665609     4.87   0.000     1.067858    2.504751
     _Iage_2 |  1.469314    .3538299     4.15   0.000     .7758204    2.162808
     _Iage_3 |  2.360093    .3473253     6.80   0.000     1.679348    3.040838
     _Iage_4 |  2.858762    .3369284     8.48   0.000     2.198394    3.519129
     _Iage_5 |   3.61029    .3620926     9.97   0.000     2.900601    4.319979
_IsexXage_1_2 |-.6769246    .3931747    -1.72   0.085    -1.447533    .0936837
_IsexXage_1_3 |-1.052307      .38774    -2.71   0.007    -1.812263   -.2923501
_IsexXage_1_4 |-1.238024    .3728725    -3.32   0.001    -1.968841   -.5072074
_IsexXage_1_5 |-1.674611    .4549709    -3.68   0.000    -2.566337   -.7828843
       _cons |-6.930277    .3333333   -20.79   0.000    -7.583599   -6.276956
      ln(py) |         1  (exposure)
------------------------------------------------------------------------------
```

The resulting equation of the Poisson regression model, using only one decimal approximation of the estimated coefficient of this output, is as follows:

$$\mathrm{Ln}\left(\hat{\mu}_i / PY_i\right) = -\mathbf{6.9} + \mathbf{1.8} * \_Isex\_1 + \mathbf{1.5} * \_Iage\_2 + \mathbf{2.4} * \_Iage\_3 + \mathbf{2.9} * \_Iage\_4$$
$$+ \mathbf{3.6} * \_Iage\_5 - \mathbf{0.7} * \_IsexXage\_1\_2 - \mathbf{1.1} * \_IsexXage\_1\_3$$
$$- \mathbf{1.2} * \_IsexXage\_1\_4 - \mathbf{1.7} * \_IsexXage\_1\_5$$

where:

- PY_i = person-years
- *_Isex_1* = 1, if sex = M; *_Isex_1* = 0 if sex = F
- *_Iage_2* = 1, if group of age "2"; *_Iage_2* = 0 other groups of age
- *_Iage_3* = 1, if group of age "3"; *_Iage_3* = 0 other groups of age
- *_Iage_4* = 1, if group of age "4"; *_Iage_4* = 0 other groups of age
- *_Iage_5* = 1, if group of age "5"; *_Iage_5* = 0 other groups of age
- *_IsexXage_1_2* = 1, if *_Isex_1* = 1 and *_Iage_2* = 1, otherwise 0
- *_IsexXage_1_3* = 1, if *_Isex_1* = 1 and *_Iage_3* = 1, otherwise 0
- *_IsexXage_1_4* = 1, if *_Isex_1* = 1 and *_Iage_4* = 1, otherwise 0
- *_IsexXage_1_5* = 1, if *_Isex_1* = 1 and *_Iage_5* = 1, otherwise 0

Considering the previous estimated coefficients, and using the expression [ii] of the Poisson model, we can determine the age-specific incidences as follows:

1. Incidence for the <46 years age group (age2 = 0, age3 = 0, age4 = 0, age5 = 0):

$$I_{<46} = e^{(-\mathbf{6.9} + \mathbf{1.8} * \_Isex\_1)}$$

2. Incidence for the 46–55 years age group (age2 = 1, age3 = 0, age4 = 0, age5 = 0):

$$I_{46-55} = e^{(\mathbf{-6.9+1.8}*\_Isex\_1+\mathbf{1.5}*\_Iage\_2-\mathbf{0.7}*\_IsexXage\_1\_2)}$$

3. Incidence for the 56–60 years age group (age2 = 0, age3 = 1, age4 = 0, age5 = 0):

$$I_{56-60} = e^{(\mathbf{-6.9+1.8}*\_Isex\_1+\mathbf{2.4}*\_Iage\_3-\mathbf{1.1}*\_IsexXage\_1\_3)}$$

4. Incidence for the 61–80 years age group (age2 = 0, age3 = 0, age4 = 1, age5 = 0):

$$I_{61-80} = e^{(\mathbf{-6.9+1.8}*\_Isex\_1+\mathbf{2.9}*\_Iage\_4-\mathbf{1.2}*\_IsexXage\_1\_4)}$$

5. Incidence for the >80 years age group (age2 = 0, age3 = 0, age4 = 0, age5 = 1):

$$I_{>80} = e^{(\mathbf{-6.9+1.8}*\_Isex\_1+\mathbf{3.6}*\_Iage\_5-\mathbf{1.7}*\_IsexXage\_1\_5)}$$

Therefore, to estimate the relative risk (males vs. females) for the first two age groups. We estimate the incidence by sex and then divide these incidences in each age group as follows:

1. The <46 years age group:
   When *_Isex_1* = 1,

$$I_{\text{male}} = e^{(\mathbf{-6.9+1.8})}$$

   And when *_Isex_1* = 0,

$$I_{\text{female}} = e^{(\mathbf{-6.9})}$$

   Then,

$$\text{RR} = \frac{I_{\text{male}}}{I_{\text{female}}} = e^{(1.8)} = 5.96$$

2. The 46–55 years age group:
   When *_Isex_1* = 1 and *_IsexXage_1_2* = 1,

$$I_{\text{male}} = I_{46-55} = e^{(\mathbf{-6.9\ +\ 1.8\ +\ 1.5-0.7})}$$

And when *_Isex_1* = 0 and *_IsexXage_1_2* = 1,

$$I_{\text{female}} = e^{(-6.9+1.5)}$$

Then

$$\text{RR} = \frac{I_{\text{male}}}{I_{\text{female}}} = \exp(1.8 - 0.7) = 3.03$$

To facilitate the estimation of these RRs with 95% confidence intervals in Stata, we can use the *lincom* command in the model, with interaction terms, instead of having one model for each age group. To use this command, after running the model with interaction terms, we enter the name of the predictor with the name Stata assigned to it; then we add the plus sign (+) followed by the corresponding interaction terms and the option *irr*. The syntaxes for the first two age groups will be as follows:

For the <46 years age group:

```
xi: glm cvd i.sex*i.age, fam(poisson) lnoff(py)
          lincom _Isex_1, irr
```

Output

```
(1)  [cvd]_Isex_1 = 0
------------------------------------------------------------------------------
     cvd |        IRR   Std. Err.     z    P>|z|  [95% Conf. Interval]
---------+--------------------------------------------------------------------
     (1) | 5.967359     2.1874    4.87   0.000   2.909142    12.24051
------------------------------------------------------------------------------
```

The incidence of cardiovascular disease in males younger than 46 years old is 5.97 (95% CI: 2.91, 12.24) times the incidence of cardiovascular disease in females younger than 46 years. This greater level of risk is highly significant (*P*-value < .001).

For the 46–55 years age group:

```
lincom _Isex_1 + _IsexXage_1_2, irr
```

Output

```
(1) [cvd]_Isex_1 + [cvd]_IsexXage_1_2 = 0
------------------------------------------------------------------------------
     cvd |      IRR   Std. Err.     z    P>|z|   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     (1) |3.032477   .4312037   7.80   0.000    2.294884    4.007138
------------------------------------------------------------------------------
```

The incidence of cardiovascular disease in males aged 46–55 years old is 3.03 (95% CI: 2.29, 4.01) times the incidence of cardiovascular disease in females aged 46–55 years old. This greater level of risk is highly significant (*P*-value < .001).

9.7 Overdispersion

Using the Poisson regression model, we can also assess the goodness of fit of the model, as was shown in the previous chapter. For example, if we run the model using only the predictor *sex* (in the previous database) as follows:

```
.  xi: glm cvd i.sex, fam(poi) lnoff(py)

Generalized linear models                 No. of obs      =          10
Optimization     : ML                     Residual df     =           8
                                          Scale parameter =           1
Deviance         =  555.0977773           (1/df) Deviance =    69.38722
Pearson          =  517.3470186           (1/df) Pearson  =    64.66838

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

                                           AIC            =    62.15084
Log likelihood   =    -308.7542            BIC            =    536.6771

------------------------------------------------------------------------------
         |                 OIM
     cvd |     Coef.   Std. Err.       z    P>|z|    [95% Conf.   Interval]
---------+--------------------------------------------------------------------
 _Isex_1 |  .6055324   .0524741    11.54   0.000     .5026851    .7083797
    cons | -4.554249   .0392232  -116.11   0.000    -4.631125   -4.477373
  ln(py) |         1  (exposure)
------------------------------------------------------------------------------
```

We discover that the deviance is 555.10, with 8 degrees of freedom; therefore, overdispersion is observed. To assess the departure between the deviance and the degrees of freedom, we obtain the *P*-value in Stata, in the following manner:

```
dis chi2tail(8,555.09)

1.05e-114
```

The results show a very highly significant difference between the deviance and its degrees of freedom (*P*-value < .001). Therefore, the Poisson regression model using only the predictor *sex* is not adequate; including more predictors, exploring another type of model, or assessing the potential correlation between adjacent age groups is called for. For more discussion on this topic, we recommend checking out the books by Cameron and Trivedi (1998), Hilbe (2007), Hoffmann (2004), and Kleinbaum et al. (2008).