Lecture 19: Poisson Regression

Poisson Regression

Recall that a random variable y has Poisson distribution if

$$Pr(y = k) = \frac{\mu^k}{k!} e^{-\mu}, \mu > 0, k = 1, 2, \cdot.s$$

We have $E(y) = var(y) = \mu > 0$. This distribution is used for counts of events occur randomly over time or space, when outcomes in disjoint periods or regions are independent. Examples are plentiful such as traffic accidents, the incidence of rare events or diseases, etc.

African Elephants Data The first of our examples deals with the number of successful matings of 41 male African elephants over a period of eight years, and to examine to what extent these numbers depend on their ages at the onset of the study.

Age	Matings	Age	Matings	Age	Matings
27	0	33	3	39	1
28	1	33	3	41	3
28	1	33	3	42	4
28	1	33	2	43	0
28	3	34	1	43	2
29	0	34	1	43	3
29	0	34	2	43	4
29	0	34	3	43	9
29	2	36	5	44	3
29	2	36	6	45	5
29	2	37	1	47	7
30	1	37	1	48	2

```
32  2  37  6  52  9
33  4  38  2
```

Poisson Regression In view of these count data, we will model the conditional distribution of the response Y given the explanatory variables X using the Poisson distribution whose mean is given by $\mu = \mu(\mathbf{x}) = E(y \mid \mathbf{x}) \geq 0$. To ensure this constraint, we assume

$$\log \mu(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \boldsymbol{\beta}^T \mathbf{x}$$

This is called a **Poisson loglinear model**. The coefficient, say β_1 of $x_1 =$ age, can be interpreted as follows: if we increase age by one unit while keeping the remaining variables fixed. This affects the mean μ of the Poisson response by $100(e^{\beta_1} - 1)\%$.

Estimation

- We obtain the maximum likelihood estimate (MLE) $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ by maximizing the log-likelihood:

$$l(\boldsymbol{\beta}) = \sum_{i}^{n} y_i \log \mu_i - \sum_{i}^{n} \mu_i$$

  Also, the MLE is asymptotically normal:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \hat{\mathbf{V}})$$

  .

- The MLE of the mean is $\mu(\mathbf{x}) = \exp(\hat{\boldsymbol{\beta}}\mathbf{x})$. In particular, $\hat{\mu}_i = \hat{\mu}(\mathbf{x}_i) = \exp(\hat{\boldsymbol{\beta}}\mathbf{x}_i)$, $i = 1, 2, \cdots, n$.
- The covariance of $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{V}}$ is estimated by

$$\widehat{COV}(\hat{\boldsymbol{\beta}}) = (\sum_{i} \hat{\mu}_i \mathbf{x}_i \mathbf{x}_i^T)^{-1}$$

  . The $se(\hat{\beta}_j)$ is the square root of the diagonal elements of the

above matrix.

- Approximate 95% CI for the mean are

$$\exp\left(\hat{\boldsymbol{\beta}}\mathbf{x}_i \pm 1.96\sqrt{\mathbf{x}_i^T(\sum_i \hat{\mu}_i\mathbf{x}_i\mathbf{x}_i^T)^{-1}\mathbf{x}_i}\right),$$

  some other ways to calculate the CI might be better.

- The deviance is

$$D = 2\sum_i y_i \log(y_i/\hat{\mu}_i) \approx \sum \frac{(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i} \sim \chi^2$$

  , which is the Pearson's chi-square statistic with $n-p-1$ df. The model for the mean is doubtful when the chi-square statistic is much greater than $n-p-1$.

- The deviance residuals are defined by

$$d_i = sign(y_i-\hat{\mu}_i)\sqrt{2(y_i\log(y_i/\hat{\mu}_i)-(y_i-\hat{\mu}_i))}.$$

Pearson residuals are given by

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}.$$

Elephant Data

```
data a;
infile elephants.txt firstobs=2;
input age matings;

proc genmod;
 model matings = age / dist  = poisson
                       lrci covb;

proc genmod;
 model matings = age age*age / dist  = poisson
                               lrci covb;
```

```
run;
```

Ceriodaphnia Organisms Data This data set contains the number of Ceriodaphnia organisms that are counted in a controlled environment in which reproduction occurs among the organisms, and there are two strains involved. The data is located at ceriodaphnia.txt. Ceriodaphnia represents the number of reproduction occurred, Concentration is the chemical component in the environment that impairs the reproduction, and Strain is the strain type.

```
Cerio Conc Strain;
datalines;
82  0.00    1
106 0.00    1
63  0.00    1
45  0.50    1
34  0.50    1
26  0.50    1
11  1.50    1
```

```
10  1.50    1
10  1.50    1
14  1.00    2
14  1.00    2
9   1.25    2
12  1.25    2
16  1.25    2
7   1.50    2
...
```

Ceriodaphnia Organisms Data

```
proc import datafile = ceriodaphnia.csv
 out = Ceriodaphnia
 dbms = CSV
 REPLACE
 ;
```

```
proc genmod;
model Cerio = Conc Strain / dist  = poisson
                            lrci covb;
run; quit;
```

Skin Cancer Example:

- Skin cancer for women in Dallas and Minn
- 8 age group and two cities (0 = Minn, 1 = Dallas)
- y_{ij} = count (number of cases) in age i and city j
- n_{ij} = population size in age i and city j

Objective: Determine whether the risk for skin cancer adjusted for age is higher in one area than in the other.

$risk$ = probability of developing skin cancer
λ_{ij} = probability of developing skin cancer in (i, j)th group.

y	city	age	pop
1	0	1	172675
16	0	2	123065
30	0	3	96216
71	0	4	92051
102	0	5	72159
130	0	6	54722
133	0	7	32185
40	0	8	8328
4	1	1	181343
38	1	2	146207
119	1	3	121374
221	1	4	111353
259	1	5	83004
310	1	6	55932
226	1	7	29007
65	1	8	7538

Poisson Regression y_{ij} has a poisson distribution with mean

$$E(y_{ij}) = n_{ij}\lambda_{ij}.$$

Poisson regression concerns the model

$$\log(y_{ij}) = \log(n_{ij}) + \beta_0 + \sum_{i=1}^{7} \beta_i age_i + \beta_8 city,$$

which is equivalent to

$$\log(\lambda_{ij}) = \beta_0 + \sum_{i=1}^{7} \beta_i age_i + \beta_8 city.$$

Interpretations:

- $\beta_0 = \log(\lambda_{81}) = \log$ of the risk in age 8 group in Dallas.
- $\beta_i = \log(\lambda_{i1}) = \log$ of the risk in age i group in Dallas (for $i = 1, 2, ..., 7$).
- $\beta_0 + \beta_8 = \log(\lambda_{80}) = \log$ of the risk in age 8 group in Minn.

- $\beta_i + \beta_8 = \log(\lambda_{i0}) = \log$ of the risk in age i in Minn (for $i = 1, 2, ..., 7$).
- $\beta_8 = \log(\lambda_{i1}) - \log(\lambda_{i0}) = \log(\frac{\lambda_{i1}}{\lambda_{i0}})$
- $RR_i = \frac{\lambda_{i0}}{\lambda_{i1}} = \exp(\beta_8), i = 1, ..., 8$ where RR_i is the relative risk in age i between two cities.

SAS Code

```
*-------------------------------------------------
---*
| Comparison of incidence of nonmelanoma skin cancer among  |
| women in Minneapolis St. Paul and Dallas Ft Worth.        |
| KKMN, 3e Table 24-1, p688                                 |
*-------------------------------------------------
---*;
```

```
title LOGISTIC AND POISSON REGRESSION;

proc import datafile = skincancer.csv
 out = skin
 dbms = CSV
 REPLACE
 ;


data skin;
  set skin;
   lpop = log(pop);
run;

proc genmod;
 class age city;
```

```
 model y = age city/ dist  = poisson
 offset= lpop lrci;
 run;


proc genmod;
 class age city;
 model y/pop = age city/ dist  = poisson
                         lrci;
 run;


proc genmod;
 class age city;
 model y/pop = age city/ dist  = bin
                         lrci;
run;
```

Note: For those situations in which n_{ij} is large and λ_{ij} is very small, the Poisson distribution can be used to approximate the binomial distribution. The larger the n_{ij} and the smaller the λ_{ij}, the better is the approximation. For the skin cancer data, we expect λ_{ij} to be small, therefore, the two models (Binomial and Poisson) should give very similar conclusions.

When studying linear regression, our models were of the form

$$E[\mathbf{y}] = \beta_0 + \beta_1 \mathbf{x}_1 + \ldots + \beta_{p-1}\mathbf{x}_{p-1}.$$

The response $\mathbf{y}$ was continuous, not discrete, and we wanted to predict the mean response and explain the variability among the observed outcomes.