7.3 Poisson Regression for Incidence Rate

Definition. Suppose the data are aggregated by certain levels of predictors and the response variable is the number of cases n per T person-time. We assume that the observations n are realizations of N, a Poisson random variable with rate λT. The Poisson model has the form

$$\mathbb{E}(N) = \lambda T \, = \, \exp\{\beta_0 + \beta_1 \, x_1 + \cdots + \beta_k \, x_k + \ln(T)\}.$$

The $\ln(T)$ term in the exponent is called the **offset**. In this case, the Poisson regression models the incidence rate. In practice, we regress n/T on the predictor variables, and write the fitted model in terms of the estimated incidence rate, $\widehat{\lambda} = \exp\{\widehat{\beta}_0 + \widehat{\beta}_1 \, x_1 + \cdots + \widehat{\beta}_k \, x_k\}$.

Fitted regression coefficients are interpreted in terms of estimated incidence rate, and a predicted value of the incidence rate is computed according to the formula $\lambda^0 = \exp\{\widehat{\beta}_0 + \widehat{\beta}_1 \, x_1^0 + \cdots + \widehat{\beta}_k \, x_k^0\}$. If it is desired to predict the number of cases n^0 for a given duration T^0, it can be achieved by writing $n^0 = \lambda^0 \cdot T^0 = T^0 \cdot \exp\{\widehat{\beta}_0 + \widehat{\beta}_1 \, x_1^0 + \cdots + \widehat{\beta}_k \, x_k^0\}$.

Example. A longitudinal epidemiological observational study was conducted in patients who underwent mitral heart valve replacement surgery. The primary endpoint of the study was symptomatic arterial thrombosis (a blood clot in the artery). The data file "thrombosis_data.csv" contains age category (mid-decade values, 45/55/65/75/85 years), gender(M/F), New York Heart Association (NYHA) classification (class I=no limitation in ordinary physical activity, class II=slight limitation in activity, class III=significant limitation in activity); number of cases of arterial thrombosis (AT) in each combination of age, gender, and NYHA classification; and duration, the total number of patient-years for that combination. The following SAS and R codes fit the Poisson regression model.

In SAS:

```
proc import out=thrombosis_data datafile="./thrombosis_data.csv"
dbms=csv replace;

proc genmod;
class gender(ref="F") NYHA(ref="I");
   model AT/duration=age gender NYHA/dist=poisson link=log;
run;
```

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-7.4282	0.2225	-7.8643	-6.9921	1114.49	<.0001
age		1	0.0117	0.0031	0.0057	0.0177	14.67	0.0001
gender	M	1	0.2902	0.0924	0.1092	0.4712	9.87	0.0017
gender	F	0	0.0000	0.0000	0.0000	0.0000	.	.
NYHA	II	1	0.1732	0.1179	-0.0578	0.4042	2.16	0.1417
NYHA	III	1	0.4387	0.1129	0.2174	0.6599	15.10	0.0001
NYHA	I	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

We can see that age, gender, and NYHA class III are significant predictors of the incidence rate of thrombosis. The fitted model is

$$\hat{\lambda} = \exp\{-7.4282 + 0.0117 \cdot age + 0.2902 \cdot male + 0.1732 \cdot NYHAII + 0.4387 \cdot NYHAIII\}.$$

As age increases by one year, the estimated incidence rate of thrombosis increases by $(\exp(0.0117) - 1) \cdot 100\% = 1.176871\%$. For males, the estimated incidence rate is $\exp(0.2902) \cdot 100\% = 133.6695\%$ of that for females. For patients with NYHA class III, the estimated incidence rate is $\exp(0.4387) \cdot 100\% = 155.069\%$ of that for NYHA class I patients. Thus, older males who have more severe symptoms of a heart disease and have more limitations in physical activities have higher incidence rate of thrombosis.

In R:

```
thrombosis.data<-read.csv(file="./thrombosis_data.csv", header=TRUE, sep=",")

gender.rel<- relevel(as.factor(thrombosis.data$gender), ref="F")
NYHA.rel<- relevel(as.factor(thrombosis.data$NYHA), ref="I")
```

#fitting Poisson model
summary(fitted.model<- glm(AT ~ age + gender.rel + NYHA.rel
+ offset(log(duration)), data=thrombosis.data, family=poisson(link=log)))

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.428238   0.222509 -33.384  < 2e-16
age          0.011728   0.003062   3.831 0.000128
gender.relM  0.290210   0.092365   3.142 0.001678
NYHA.relII   0.173202   0.117872   1.469 0.141723
NYHA.relIII  0.438667   0.112899   3.885 0.000102
```

Suppose now we would like to predict the number of thrombosis cases per 10,000 patient-years for males in their 60s with NYHA class II. We obtain

$$n^0 = 10000 \cdot \exp\{-7.4282 + 0.0117 \cdot 65 + 0.2902 + 0.1732\} = 20.20723.$$

In SAS:

```
/*using fitted model for prediction*/
data prediction;
input age gender$ NYHA$;
cards;
65 M II
;

data thrombosis_data;
 set thrombosis_data prediction;
run;

proc genmod;
class gender NYHA;
   model AT/duration=age gender NYHA/dist=poisson link=log;
 output out=outdata p=pred_IR;
run;

data outdata;
 set outdata;
  pred_AT=10000*pred_IR;
run;
```

```
proc print data=outdata (firstobs=31) noobs;
var pred_AT;
run;
```

pred_AT
20.2438

In R:

#using fitted model for prediction
print(predict(fitted.model, data.frame(age=65, gender.rel="M", NYHA.rel="II", duration=10000), type="response"))

```
20.24329
```

□