7.2 Poisson Regression for Count Data

Definition. If a variable assumes only non-negative integer values (0, 1, 2, ...), it is called a **count variable**. Suppose the response variable y is a count variable, and large values are very unlikely. It means that we can model y as a Poisson random variable, and regress it on a set of predictors $x_1, \ldots, x_k$ through a Poisson regression model defined as:

$$\mathbb{P}(Y = y) = \frac{\lambda^y \exp\{-\lambda\}}{y!}, \quad y = 0, 1, 2, \ldots,$$

where the rate

$$\lambda = \mathbb{E}(y) = \exp\big\{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\big\}.$$

Note that in a Poisson regression model, $\ln(\lambda) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$, and thus it is a generalized linear regression model with the **log-link** function.

Definition. The **fitted Poisson regression model** has the form

$$\widehat{\lambda} = \exp\big\{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_k x_k\big\}.$$

Definition. In the Poisson regression model, the estimates of the regression coefficients are **interpreted** as follows:

• If a predictor variable x_1 is numeric, then the quantity $(\exp\{\widehat{\beta}_1\} - 1) \cdot 100\%$ $(\exp\{\widehat{\beta}_1\} - 1) \cdot 100\%$ represents the estimated percent change in rate when x_1 increases by one unit, while all the other predictors are held fixed. Indeed,

$$\frac{\widehat{\lambda}|_{x_1+1} - \widehat{\lambda}|_{x_1}}{\widehat{\lambda}|_{x_1}} \cdot 100\% = \Big(\exp\big\{\widehat{\beta}_0 + \widehat{\beta}_1 (x_1 + 1) + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_k x_k\big\}$$

$$- \exp\big\{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + 1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_k x_k\big\}\Big) / \exp\big\{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_k x_k\big\} \cdot 100\%$$

$$= \big(\exp\big\{\widehat{\beta}_1\big\} - 1\big) \cdot 100\%.$$

• If a predictor variable x_1 is 0-1 variable, then the quantity $\exp\{\widehat{\beta}_1\} \cdot 100\%$ represents the estimated percent ratio of rates when $x_1 = 1$ and when $x_1 = 0$, while the other predictors are held constant. To see that, we write

$$\frac{\widehat{\lambda}|_{x_1=1}}{\widehat{\lambda}|_{x_1=0}} \cdot 100\% = \frac{\exp\big\{\widehat{\beta}_0 + \widehat{\beta}_1 \cdot 1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_k x_k\big\}}{\exp\big\{\widehat{\beta}_0 + \widehat{\beta}_1 \cdot 0 + \cdots + \widehat{\beta}_k x_k\big\}} \cdot 100\% = \exp\big\{\widehat{\beta}_1\big\} \cdot 100\%.$$

Definition. In a Poisson regression, for a specified set of predictors $x_1^0, \ldots, x_k^0$, the **predicted response** y^0 is computed as

$$y^0 = \exp\left\{\widehat{\beta}_0 + \widehat{\beta}_1\, x_1^0 + \cdots + \widehat{\beta}_k\, x_k^0\right\}.$$

Example. Number of days of hospital stay were recorded for 45 patients with chest pain, along with their gender, age, and history of chronic cardiac illness. The data are given in the file "hospital_stay.csv". We regress the number of days on the other variables via a Poisson regression model. To this end, we run the following SAS code:

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

proc genmod;
class gender(ref="F") illness(ref="no");
   model days=gender age illness/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	-0.8263	0.4702	-1.7479	0.0953	3.09	0.0789
gender	M	1	0.2264	0.2331	-0.2305	0.6834	0.94	0.3315
gender	F	0	0.0000	0.0000	0.0000	0.0000	.	.
age		1	0.0205	0.0079	0.0050	0.0359	6.76	0.0093
illness	yes	1	0.4477	0.2223	0.0119	0.8834	4.05	0.0440
illness	no	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

In the fitted model, the estimated rate is $\widehat{\lambda} = \exp\left\{-0.8263 + 0.2264 \cdot male + 0.0205 \cdot age + 0.4477 \cdot illness\right\}$. Patient's age and the indicator of a chronic

cardiac illness are significant predictors of the average length of stay at the 5% significance level. For a one-year increase in patient's age, the estimated average number of days of hospital stay increases by $(\exp(0.0205)-1)\cdot 100\% = 2.07\%$. Also, the estimated average number of days of hospital stay for patients with a chronic cardiac illness is $\exp(0.4477)\cdot 100\% = 156.47\%$ of that for patients without it.

In R:

```
hospital.stay<-read.csv(file="./hospital_stay.csv", header=TRUE, sep=",")

#specifying reference categories
gender.rel<- relevel(as.factor(hospital.stay$gender), ref="F")
illness.rel<- relevel(as.factor(hospital.stay$illness), ref="no")

#fitting Poisson model
summary(fitted.model<- glm(days ~ gender.rel + age + illness.rel,
data=hospital.stay, family=poisson(link=log)))
```

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.826269   0.470206  -1.757  0.07888
gender.relM      0.226425   0.233142   0.971  0.33145
age              0.020469   0.007871   2.600  0.00931
illness.relyes   0.447653   0.222305   2.014  0.04404
```

Further, the predicted length of stay for a 55-year old male with no chronic cardiac illness is computed as $y^0 = \exp\big\{-0.8263 + 0.2264 + 0.0205\cdot 55\big\} = 1.6949$.

In SAS:

```
data prediction;
input gender$ age illness$;
cards;
M 55 no
;

data hospital_stay;
 set hospital_stay prediction;
```

```
run;

proc genmod;
class gender illness;
model days=gender age illness/dist=poissonlink=log;
 output out=outdata p=pred_days;
run;

proc print data=outdata(firstobs=46) noobs;
var pred_days;
run;
```

pred_days
1.69207

In R:

#using fitted model for prediction
print(predict(fitted.model, data.frame(gender.rel="M", age=55, illness.rel="no"), type="response"))

```
1.692066
```

□