BIST P8110: Applied Regression II

20. Poisson Regression: Case Study

Qixuan Chen

Department of Biostatistics
Columbia University

Case Study: Poisson Regression

- Suppose the following insurance claims data are classified by two factors: age group (with two levels) and car type (with three levels). The variable N represents the number of insurance policyholders and the variable Y represents the number of insurance claims.

```
data insure;
input N Y car $ age;
log_N = log(N);
datalines;
500 42 small 1
1200 37 medium 1
100 1 large 1
400 101 small 2
500 73 medium 2
300 14 large 2
;
```

Poisson Regression Model

- The Poisson regression model for the insurance claims rate is

$$\log\{E(Y_i|\text{age}_i, \text{car}_i)\} = \log(n_i) + \beta_0 + \beta_1 \text{car}_{1i} + \beta_2 \text{car}_{2i} + \beta_3 \text{age}_i$$

where,

$$\text{car}_1 = \begin{cases} 1 & \text{car=median} \\ 0 & \text{ortherwise} \end{cases}$$

$$\text{car}_2 = \begin{cases} 1 & \text{car=large} \\ 0 & \text{ortherwise} \end{cases}$$

$$\text{age} = \begin{cases} 1 & \text{age group} = 2 \\ 0 & \text{age group} = 1 \end{cases}$$

SAS Codes

- The following SAS statements invoke the GENMOD procedure to fit the Poisson regression model

```
proc genmod data=insure;
 class car(ref=`small') age(ref=`1')/param=ref;
 model Y = car age /link=log dist=poi offset=log_N type3;
run;
```

SAS Outputs

```
                    The GENMOD Procedure

                     Model Information

       Data Set                     WORK.INSURE
       Distribution                     Poisson
       Link Function                        Log
       Dependent Variable                     Y
       Offset Variable                    log_N


    Number of Observations Read              6
    Number of Observations Used              6


                  Class Level Information

                                    Design
         Class     Value          Variables

         car       large          1      0
                   medium         0      1
                   small          0      0

         age       1              0
                   2              1
```

SAS Outputs

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	2	2.8207	1.4103
Scaled Deviance	2	2.8207	1.4103
Pearson Chi-Square	2	2.8416	1.4208
Scaled Pearson X2	2	2.8416	1.4208
Log Likelihood		837.4533	
Full Log Likelihood		-16.4638	
AIC (smaller is better)		40.9276	
AICC (smaller is better)		80.9276	
BIC (smaller is better)		40.0946	

Algorithm converged.

SAS Outputs

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.6367	0.1318	-2.8950	-2.3784	400.20	<.0001
car	large	1	-1.7643	0.2724	-2.2981	-1.2304	41.96	<.0001
car	medium	1	-0.6928	0.1282	-0.9441	-0.4414	29.18	<.0001
age	2	1	1.3199	0.1359	1.0536	1.5863	94.34	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
car	2	72.82	<.0001
age	1	104.64	<.0001

Goodness of Fit Test

- The "Criteria For Assessing Goodness Of Fit" table (page 6) contains statistics that summarize the model fit.
- Test goodness of fit using Deviance (or using Pearson Chi-Square)
  - H_0: there is no lack of fit. vs. H_α: there is lack of fit.
  - Test statstic: $D = 2.8207$ with df $= 2$
  - $Pr(\chi^2_2 \geq 2.8207) = 0.244 > 0.05$
  - Conclusion: Fail to reject H_0, and conclude that the specified model fits the data reasonably well.

Overdispersion

- The "Criteria For Assessing Goodness Of Fit" table can also be used to identify overdispersion
  - If no overdispersion, the ratio of Deviance to DF or the ratio of Pearson Chi-square to DF, **Value/DF**, should be about one.
  - **Value/DF** > 1 may indicate overdispersion given that there is no sufficient evidence of lack of fit of the model.

Overdispersion

- One method to "fix" overdispersion is to correct the standard errors of the estimates using scale option
  - Specify the scale option (scale=d or scale=p) in the model statement. The scaled Deviance or scaled Pearson Chi-Square is forced to be one.
  - The standard errors of the regression coefficients are multiplied by a factor $= \sqrt{\text{Value}/\text{DF}}$.

SAS Code: "Fix" Overdispersion

- The SAS code is as follows:

```
proc genmod data=insure;
 class car(ref='small') age(ref='1')/param=ref;
 model Y = car age /link=log dist=poi
                     offset=log_N type3 scale=p;
run;
```

SAS Outputs: With Scale Option

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	2	2.8207	1.4103
Scaled Deviance	2	1.9853	0.9926
Pearson Chi-Square	2	2.8416	1.4208
Scaled Pearson X2	2	2.0000	1.0000
Log Likelihood		589.4219	
Full Log Likelihood		-16.4638	
AIC (smaller is better)		40.9276	
AICC (smaller is better)		80.9276	
BIC (smaller is better)		40.0946	

Algorithm converged.

SAS Output: With Scale Option

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.6367	0.1571	-2.9446	-2.3288	281.67	<.0001
car	large	1	-1.7643	0.3247	-2.4006	-1.1280	29.53	<.0001
car	medium	1	-0.6928	0.1529	-0.9924	-0.3932	20.54	<.0001
age	2	1	1.3199	0.1620	1.0024	1.6374	66.40	<.0001
Scale		0	1.1920	0.0000	1.1920	1.1920		

NOTE: The scale parameter was estimated by the square root of Pearson's Chi-Square/DOF.

LR Statistics For Type 3 Analysis

Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
car	2	2	25.63	0.0376	51.25	<.0001
age	1	2	73.65	0.0133	73.65	<.0001

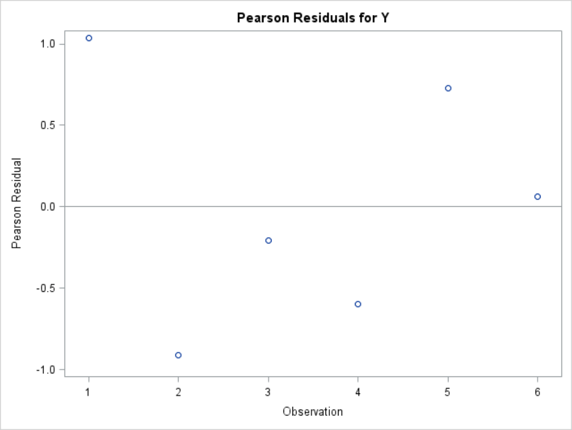
NOTE: The scale estimate is now 1.1920 (=$\sqrt{1.4208}$, Pearson Chi-Square Value/DF) and the standard error is multiplied by 1.1920 compared to the SAS outputs on page 7.

Residuals

We can output the Pearson residuals plot by using the following SAS statements:

```
ods graphics on;
proc genmod data=insure plots=RESCHI;
 class car(ref='small') age(ref='1')/param=ref;
 model Y = car age /link=log dist=poi
                   offset=log_N type3 scale=p;
run;
ods graphics off;
```

SAS Output: Pearson Residuals



NOTE: No observations have absolute value of the Pearson residuals larger than 2.0

Hypothesis Testing

- To check whether the car type is a significant predictor of the insurance claim rate, we can use the output in "LR Statistics For Type 3 Analysis"
  - $H_0 : \beta_1 = \beta_2 = 0$ vs. H_α : at least one β not equal to zero
  - Test statistic: $\frac{(\text{Deviance}_{\text{age}} - \text{Deviance}_{\text{age+car}})}{(n-p)*\hat{\phi}} = 25.63$
  - $Pr(F_{r,n-p} \geq 25.63) = Pr(F_{2,2} \geq 25.63) = 0.0376$
  - Conclusion: Reject H_0 at $\alpha = 0.05$. There is sufficient evidence to conclude that the car type is significantly associated with the insurance claim rate.

Regression Coefficient Interpretation

- Interpretation of β_1
  - $\hat{\beta}_1 = -0.6928$ with a 95% CI: (-0.9924, -0.3932)
  - $e^{\hat{\beta}_1} = e^{-0.6928} = 0.5002$ with a 95% CI: (0.3707, 0.6749)
  - Interpretation: The insurance claim rate among the insurance policyholders with median size cars is 49.9% percent less than that of policyholders with small size cars, and this decrease could be as little of 32.5% or as much as 62.9% with 95 percent confidence.

Relative Risk

- The relative risk of insurance claims among policyholders who have large size cars and are in age group 1 versus the policyholders who have median size cars and are in age group 1:

$$
\begin{aligned}
&\frac{E(Y_i|\text{car}_{1i} = 0, \text{car}_{2i} = 1, \text{age}_i = 0)/n_i}{E(Y_j|\text{car}_{1j} = 1, \text{car}_{2j} = 0, \text{age}_j = 0)/n_j} \\
&= \frac{e^{\hat{\beta}_0 + \hat{\beta}_2}}{e^{\hat{\beta}_0 + \hat{\beta}_1}} \\
&= e^{\hat{\beta}_2 - \hat{\beta}_1} \\
&= e^{-1.7643 - (-0.6928)} \\
&= 0.3425
\end{aligned}
$$

Model Prediction

- Expected number of insurance claims per 1000 insurance policyholders who drive large cars and are in age group 2:

$$
\begin{aligned}
& \hat{E}(Y|\text{car}_1 = 0, \text{car}_2 = 1, \text{age} = 1) \\
&= N \times e^{\hat{\beta}_0 + \hat{\beta}_2 + \hat{\beta}_3} \\
&= 1000 \times e^{-2.6367 - 1.7643 + 1.3199} \\
&= 45.9
\end{aligned}
$$