Poisson Regression

STAT/BIOS 823

Homework 12

Directions

Using `RMarkdown` in `RStudio`, complete the following questions. Launch RStudio and open a new RMarkdown file or use the class RMarkdown template provided and save it on your working directory as a `.Rmd` file. At the end of the activity, save your **pdf** generated from `RMarkdown+Knitr` and submit your homework on the Blackboard.

All questions are mandatory.

Some `R-codes` and **output** have been provided for you. `R-code` and output must be clearly shown.

Homework submitted after the due date will attract a penalty of 10 points per day after the due date.

If you have questions, please post them on the **lesson discussion board**.

Poison distribution can be utilized for analyzing data with counts as the outcome of interest ($Y_i = 0, 1, 2, \cdots$). The main assumption is that the mean and variance of the Poisson distribution are equall. That is, $E\{Y\} = \mu$ and $\sigma^2\{Y\} = \mu$.

1. A study was conducted by Sir Richard Doll and colleagues (Breslow and Day, 1987) based on 1951 data where all British doctors were sent a brief questionnaire about whether they smoked tobacco. Since then information about their deaths has been collected. The data below shows the numbers of deaths from coronary heart disease among male doctors 10 years after the survey. It also shows the total number of person-years of observation at the time of the analysis. Note that this data can be loaded using `data(doctors)` after loading the R package `dobson`. However, you have to convert the strings variables into factors and numerical values where appropriate.

The data is shown on the Table below:

Table 1: Deaths from coronary heart disease after 10 years among British male doctors by age and smoking status in 1951

age	agesq	agecat	smoke	deaths	personyrs
1	1	35-44	smoker	32	52407
2	4	45-54	smoker	104	43248
3	9	55-64	smoker	206	28612
4	16	65-74	smoker	186	12663
5	25	75-84	smoker	102	5317
1	1	35-44	non-smoker	2	18790
2	4	45-54	non-smoker	12	10673
3	9	55-64	non-smoker	28	5710
4	16	65-74	non-smoker	28	2585
5	25	75-84	non-smoker	31	1462

(a) Write an *R*-code to create this dataset or load it from the package `dobson`. Fit a Poisson regression model to the data with the response being the number of `deaths` as follows: $Y_i = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2 + \beta_3 Smoke_i + \beta_4 Smoke_i \times age_i + \epsilon_i$. Use an `offset = log(personyrs)` in your model. From this model, what is the value of the residual deviance? What is the value of the residual degrees of freedom? What is the *pseudo* R^2 value for the model?

(b) Obtain the deviance residuals and present them in an index plot (explore the function *resid*() after fitting your model). Are there any outlying cases?

```
docm1$null.deviance  # NULL DEVIANCE
docm1$deviance  # DEVIANCE RESIDUALS
devresd <- round(resid(docm1, type = "deviance"), 3)
sum(devresd^2)  # Chi-square goodness-of-fit statistic
stddevres <- rstandard(docm1)  # standardized deviance residuals

# standardized Pearson's residuals
stdpearsons <- rstandard(docm1, type = "pearson")
pearson.resid <- round(resid(docm1, type = "pearson"),
    3)  # Pearson residuals
sum(pearson.resid^2)  # Pearson goodness-of-fit statistic
# expected <- fitted.values(docm1, type='response')
expected <- round(predict(docm1, type = "response"),
    2)
```

(c) Create a table with the observed and expected number of deaths together with the age groups and smoking status. What is the value of the deviance residuals goodness-of-fit statistic (G^2)? (**Hint**: sum the squares of the residual deviances) What is the value of the Pearson's Chi-square goodness-of-fit statistic (χ^2)? (**Hint**: sum the squares of the pearson's residuals).

```
kable(cbind(dataDeaths[, c(1, 3, 4, 5)], expected,
    pearson.resid, devresd))
```

age	agecat	smoke	deaths	expected	pearson.resid	devresd
1	35-44	smoker	32	29.58	0.444	0.438
2	45-54	smoker	104	106.81	-0.272	-0.273
3	55-64	smoker	206	208.20	-0.152	-0.153
4	65-74	smoker	186	182.83	0.235	0.234
5	75-84	smoker	102	102.58	-0.057	-0.057
1	35-44	non-smoker	2	3.41	-0.766	-0.830
2	45-54	non-smoker	12	11.54	0.135	0.134
3	55-64	non-smoker	28	24.74	0.655	0.641
4	65-74	non-smoker	28	30.23	-0.405	-0.411
5	75-84	non-smoker	31	31.07	-0.013	-0.013

(d) Perfom a test of the goodness of fit for this model. Are the Poisson's distribution assumptions violated? Based on this model, what conclusions do you reach?