STA 223 Homework 4

Bohao Zou (bhzou@ucdavis.edu)

February 17, 2021

1. Set 4, Problem 5

Since for each i only one of the Y_{im} can satisfy $Y_{im} = 1$ and by conditional probability, we can have:

$$\begin{aligned}
&P(Y_{ij} = 1 | Y_{ij} = 1 \quad or \quad Y_{i1} = 1) \\
&= \frac{P(Y_{ij} = 1 | Y_{ij} = 1) + P(Y_{ij} = 1 | Y_{i1} = 1)}{P(Y_{ij} = 1) + P(Y_{i1} = 1)} \\
&= \frac{P(Y_{ij} = 1)}{P(Y_{ij} = 1) + P(Y_{i1} = 1)} \\
&= \frac{\pi_{ij}}{\pi_{ij} + \pi_{i1}} \\
&logit(P(Y_{ij} = 1 | Y_{ij} = 1 \quad or \quad Y_{i1} = 1)) = \eta_{ij}
\end{aligned}$$

We can set $\pi_{i1} = b$, $exp(\eta_{ij}) = a_{ij}$, for each $2 \leq j \leq M$, we have $\frac{\pi_{ij}}{b} = a_{ij}$, so we can have $M - 1$ equations.

$$\begin{aligned}
\pi_{i2} &= a_{i2} b \\
\pi_{i3} &= a_{i3} b \\
&\cdots \\
\pi_{iM} &= a_{iM} b \\
1 - b &= \sum_{j=2}^{M} \pi_{ij}
\end{aligned}$$

By using those $M - 1$ equations we can solve $M - 1$ parameters. We can first solve b,

$$\begin{aligned}
1 - b &= \sum_{j=2}^{M} a_{ij} b \\
1 &= (\sum_{j=2}^{M} a_{ij} + 1) b \\
b = \pi_{i1} &= \frac{1}{1 + \sum_{j=2}^{M} exp(\eta_{ij})}
\end{aligned}$$

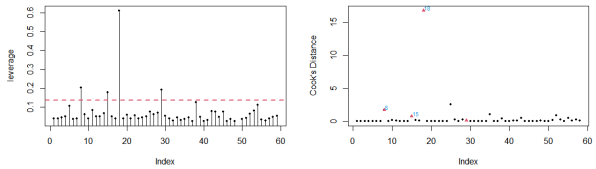
Then, we can have the formula of π_{ij}

$$\pi_{ij} = \frac{exp(\eta_{ij})}{1 + \sum_{j=2}^{M} exp(\eta_{ij})}$$

2. Set 4, Problem 6

(a)

The Cook's distance and leverage plots are showed below:



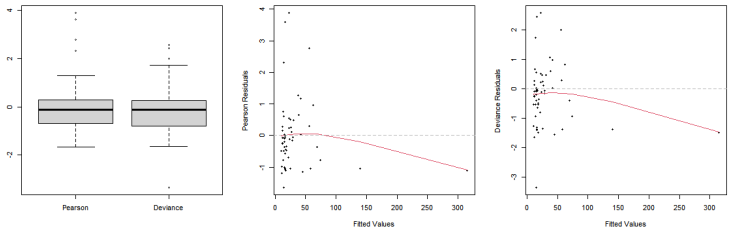
From this plot we can know the index of 8, 15, 18, 29 cases are identified as outliers.

(b)

The overdispersion of this model is 10.359, From this value we can know the overdispersion parameters are significantly bigger than 1. The means the truth variance has big difference with expected variance. This may arise because of the model is lack fitted with this dataset.

(c)

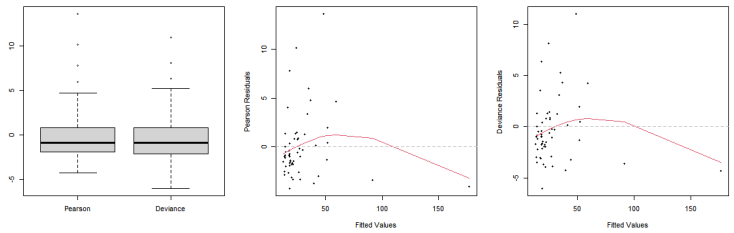
We can use the boxplot and the fitted value vs. deviances and Pearson residuals to check the goodness of fit for negative binomial model.



From this plot we can know there still exist lack-of-fit in this model. but the overdispersion of this model is 1.15. It is much better than Poisson regression.

(d)

We can use the boxplot and the fitted value vs. deviances and Pearson residuals to check the goodness of fit for quasi-likelihood model with log link function.



From this plot we can know there still exist lack-of-fit in this model. The overdispersion is 10.35. So, there is no difference with previous Poisson regression analysis.

3. Set 4, Problem 7

(a)

We can use Poisson regression with log link, the linear predictors are showed below:

$$\eta_i = \beta_0 + x_{SnailCon\_1}\beta_1 + x_{hygieneI\_1}\beta_2 + x_{hygieneI\_2}\beta_3 + x_{medicalI\_1}\beta_4 + x_{No.ofIn}\beta_5 \tag{1}$$

The assumptions are: Each y_i follows the Poisson distribution with mean λ_i, The responses y_i are independent of one another, and each y_i is a non-negative integer. The hygiene index coded as two indicator variables. The hygiene index 0 would be the natural baseline level.

(b)

We can give a consideration that the occurrence of schistosomiasis can be related with the number of inhabitants of the village. The role of the predictor is used for testing if the consideration is true or false. I do not expect this to be a relevant predictor. If I find it to be significant, I would design another data which randomly selects the same number of people from each village.

(c)

For the case of MLEs, $\Sigma = I^{-1}(\hat{\beta})$, where $I(\hat{\beta})$ is the finitely estimated information matrix. where A is a $2 \times p$ matrix that for the elements a_{ii} and a_{jj} are 1, the others elements are all 0. The a_{ii} means the i-th row and i-th col. The i, j are the indices of snail control and medical access respectively. The output we need is the estimated coefficients, the estimated information matrix at last iteration. Then we have:

$$[A(\hat{\beta} - \beta)]^T (A\Sigma A^T)^{-1} [A(\hat{\beta} - \beta)] \sim \chi_2^2$$

(d)

From the linear predictors (1) we can know the coefficients of snail control index and medical access are β_1 and β_4. The null hypothesis is $\beta_1 - \beta_4 > 0$, the null hypothesis is $\beta_1 - \beta_4 \leq 0$. We can use the theory from (c) to build the test statistic under null hypothesie:

$$[A(\hat{\beta} - \beta)]^T (A\Sigma A^T)^{-1} [A(\hat{\beta} - \beta)] \sim \chi_1^2$$

where A is a $1 \times p$ matrix that [0, 1, 0, 0 -1, 0].

(e)

The components can be this:

$$\eta_i = \beta_0 + x_{SnailCon\_1}\beta_1 + x_{hygieneI\_1}\beta_2 + x_{hygieneI\_2}\beta_3 + x_{medicalI\_1}\beta_4 + x_{SnailCon\_1} \times x_{No.ofIn}\beta_5$$

4. Set 4, Problem 8

(a)

The AIC criteria of proportional odd model is 3165.398. The AIC criteria of baseline odd model is 2477.033. Because those models all have the same number of predictors, we can know the baseline odd model fitted better than proportional odd model.

In the proportional odd model, the P-Values of "volatile.acidity","fixed.acidity","citric.acid", "residual.sugar", "free.sulfur.dioxide", "total.sulfur.dioxide", "density" are lower than significant level alpha = 0.05. Those predictors determine the quality of the wine.

In the baseline odd model, the P-values of "volatile.acidity", "residual.sugar", "chlorides", "total.sulfur.dioxide", "density", "sulphates" and "alcohol" are all significant under level 0.05. In this model, those predictors determine the quality of wine.

(b)

In this logistic regression model, The P-Values of predictors "volatile.acidity", "citric.acid", "chlorides", "free.sulfur.dioxide", "total.sulfur.dioxide", "sulphates" and "alcohol" are all lower than 0.05. Those predictors are related with the quality of wine.

(c)

Based on the fits and inference of the binomial regression with those of the multinomial regression models, the model of binomial regression I would prefer for this application. This is because the predicted accuracy of binomial regression is much better than the multinomial regression.

5. Set 4, Problem 11

(a)

We can use Poisson regression with log link, the linear predictors are showed below:

$$\eta_i = v_i\beta_1 + a_i\beta_2 + e_i\beta_3$$

(b)

We can add an interception item into this model. This interception item can represent the effect that number of cells are in a inactive state where they can not respond to the stimulus.

$$\eta_i = \beta_0 + \upsilon_i\beta_1 + a_i\beta_2 + e_i\beta_3$$

(c)

I would extend the linear predictor as this :

$$\eta_i = \beta_0 + \upsilon_i\beta_1 + a_i\beta_2 + e_i\beta_3 + e_i a_i\beta_4 + e_i\upsilon_i\beta_5$$

(d)

I would use one of zero inflated Poisson model or zero inflated negative binomial model to analyze this data. If the overdispersion of zero inflated Poisson model is roughly equal with 1, I would use zero inflated Poisson model. Otherwise, I would use zero inflated negative binomial model. This is because for the zero count in the dataset, if the count data is zero, there are two possible cases, the first is some cells are active but do not respond for the stimulate. The second is the cells are all inactive so that they can not respond to the stimulus.

6. Set 4,Problem 12

(a)

I would use Poisson or negative binomial regression model for this data set.

(b)

I would build negative binomial regression model first and use leverage values and Cook's distance to remove the ourliers. Then use the clean dataset to fit a poisson regression and negative binomial model respectively. If the overdispersion of Poisson model is roughly equal with 1, I would use Poisson model. Otherwise, I would use zero inflated negative binomial model.

(c)

If the leverage values are bigger than $\frac{2p}{n}$ then treated those cases as potential outliers. If the Cook's distances are bigger than $\frac{4}{n-p}$ then treated those cases as potential outliers. The estimator for the fraction of recordings that is affected is the number of intersection of potential leverage outliers and Cook's distance outliers.

7. Set 5, Problem 1

For the $g(\mu_i) = log(\mu_i)$ we can have $E[Y_i] = \mu_i = exp(\eta_i)$. For the $g(\mu_i) = \frac{1}{\mu_i}$ we can have $E[Y_i] = \mu_i = \frac{1}{\eta_i}$. From the property of Gamma distribution we can know the mean of Gamma distribution must great than 0. This means $E[Y_i] = \mu_i = \frac{1}{\eta_i}$ is not suitable for negative η_i but $exp(\eta_i)$ is fitted for every real number. This is the advantage for *log* link function and the disadvantage for inverse function.

8. Set 5, Problem 8

(a)

We can use binomial regression with logit link, the linear predictors are showed below:

$$\eta_i = \beta_0 + age\beta_1 + National_{US}\beta_2 + FAT\beta_3 + FAT \times National_{US}\beta_4$$

(b)

We would use log-log link function. $\eta = log(-log(1-\mu))$. This is because this link function is suitable for $0 < \mu < 1$ and it is a asymmetric link function.

(c)

The null hypothesis is that $\beta_4 = 0$ and the alternative hypothesis is $\beta_4 \neq 0$. I would plan to discuss the set-up of the hypothesis with the investigator. As the Japan as the baseline, the difference between the effect of FAT on breast cancer of U.S. and Japan is $effect = \beta_2 + FAT\beta_4$. The change of FAT will change this $effect$ by β_4. So, if we want to demonstrate the effect of FAT is the same in U.S. and Japan, we can demonstrate that $\beta_4 = 0$. So, We can conduct this null hypothesis.

(d)

$$[A(\hat{\beta} - \beta)]^T (A\Sigma A^T)^{-1} [A(\hat{\beta} - \beta)] \sim \chi_1^2$$

where A is a $1 \times p$ matrix that [0, 0, 0, 0, 1], $\Sigma = I^{-1}(\hat{\beta})$, where $I(\hat{\beta})$ is the finitely estimated information matrix. From this we can get p-value.

(e)

If we want to find out whether the effect of age and fat intake is linear or not, we can build this model:

$$\begin{aligned}
&\eta_i = g(\pi_i, \alpha) = \beta_0 + age\beta_1 + FAT\beta_2 + age^{\alpha}\beta_3 + FAT^{\alpha}\beta_4 \\
&\eta_i = g(\pi_i, \alpha) \approx g(\pi_i, \alpha_0) + (\alpha - \alpha_0)\gamma_i = \beta_0 + age\beta_{13} + FAT\beta_{24} + (\alpha - \alpha_0)\gamma_i \\
&= \beta_0 + age\beta_{13} + FAT\beta_{24} + \gamma_i\beta_5
\end{aligned}$$

where $\alpha_0 = 1$ and $\gamma_i = \frac{\partial g}{\partial \alpha}|_{\alpha=\alpha_0}$. For a given link function g, we can add γ_i to the linear predictor η_i and give the null hypothesis $\beta_5 = 0$. If the testing is significant, we can derive the age and FAT are all nonlinear effect.

(f)

For continuous predictors age and Fat, we can fit $logit(E(Y|X = x)) = \alpha + \sum_{j=1}^{p-1} g_j(x_j)$ with component function $g_1, \ldots, g_{p-1}$. Following the result of (e) and the relationships between Y and age, Fat are all

complex. We can know there exist non-linear between age, Fat with Y. We can use Smoothing Methods to build this model.

In this question, we can set $\alpha = 0$ and $g_i = g_j = g, i \neq j$. By those assumptions, we can get $logit(E(Y|X = x)) = g(x_{age}) + g(x_{fat})$. In this question, I will use Quotient type as the function and set kernel K as Gaussian kernel.

$$logit(E(Y|X = x)) = \frac{\sum_{i=1}^{n} \frac{y_i}{h} K(x_{age,i} - X_{age,i})}{\sum_{i=1}^{n} \frac{K(x_{age,i} - X_{age,i})}{h}} + \frac{\sum_{i=1}^{n} \frac{y_i}{h} K(x_{FAT,i} - X_{FAT,i})}{\sum_{i=1}^{n} \frac{K(x_{FAT,i} - X_{FAT,i})}{h}}$$

$$K(x) = \frac{1}{\sqrt{2\pi} exp(-x^2/2)}$$

By using this model, we can build a smoothing line and give the probability that $P(Y = 1|X = x)$.