QUALIFYING EXAM SOLUTIONS

Statistical Methods

Fall, 2011

1. (a) The model is

$$Y \sim Poisson(\alpha + \beta X).$$

Suppose (Y_i, X_i) for $i = 1, \cdots, n$ are observed. The loglikelihood function is

$$\ell(\alpha, \beta) = -\sum_{i=1}^{n} Y_i! + \sum_{i=1}^{n} Y_i \log(\alpha + \beta X_i) - \sum_{i=1}^{n} (\alpha + \beta X_i).$$

Then, the likelihood equations are

$$\begin{aligned} \frac{\partial \ell(\alpha, \beta)}{\partial \alpha} &= \sum_{i=1}^{n} \frac{Y_i}{\alpha + \beta X_i} - n \\ \frac{\partial \ell(\alpha, \beta)}{\partial \beta} &= \sum_{i=1}^{n} \frac{Y_i X_i}{\alpha + \beta X_i} - \sum_{i=1}^{n} X_i. \end{aligned}$$

It is clear that those are different from the least square equations. Therefore, the MLE is not equal to the LSE.

(b) Let π_i be the probability. Then the model is

$$\log \frac{\pi_j}{\pi_1} = \beta_{j0} + \beta_{j1} X.$$

Then, the model is equivalent to

$$\pi_j = \pi_1 e^{\beta_{j0} + \beta_{j1} X}$$

and

$$\sum_{j=1}^{J} \pi_j = 1.$$

If we fit $J - 1$ separate logistic regression model for $(1, j)$, we also have

$$\pi_j = \pi_1 e^{\beta_{j0} + \beta_{j1} X}$$

which is the same as the formula from the previous model. However, the constraint is $\pi_j + \pi_1 = 1$, which is different from the previous one. Therefore, the two models are different.

(c) Let π_j be the probability of the j-th category. The model assumption is

$$\log \frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_5} = \theta_j - \beta_2 I_{X=2} - \beta_3 I_{X=3} - \beta_4 I_{X=4},$$

for $j = 1, 2, 3, 4$. Then, we have $\hat{\theta}_1 = -0.9188$, $\hat{\theta}_2 = -0.5183$, $\hat{\theta}_3 = 0.4922$, $\hat{\theta}_4 = 1.8579$, $\hat{\beta}_2 = 0.1176$, $\hat{\beta}_3 = 0.3174$, $\hat{\beta}_4 = 0.5208$. If the last category is used as a baseline, then, we have the model

$$\log \frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_5} = \theta'_j - \beta'_1 I_{X=1} - \beta'_2 I_{X=2} - \beta'_3 I_{X=3}.$$

Compare it with the previous model, we have

$$\theta'_j = \theta_j - \beta_4$$

and

$$\beta'_i = \beta_i + \beta_4.$$

Therefore, we have $\hat{\theta}'_1 = -1.4396$, $\hat{\theta}'_2 = -1.0391$, $\hat{\theta}'_3 = -0.0286$, $\hat{\theta}'_4 = 1.3371$, $\hat{\beta}'_1 = -0.5208$, $\hat{\beta}'_2 = -0.4032$, and $\hat{\beta}'_3 = -0.2034$.

2. (a) Because the model does not contain any interaction effect, it assumes all the interaction effects are zero:

$$\log \lambda_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l,$$

where λ_{ijkl} is the expected value of the response, α_i, β_j, γ_k and δ_l, represents Heart, Comps, Smoke, and BW, respectively. It assumes all the independent variables are independent.

(b) The expected count is

$$e^{5.15385-1.55769-0.04539-0.99056} = 12.94.$$

(c) The residual degree of freedom is $df = 16 - 5 = 11$. Since the residual deviance is $164.7 > \chi^2_{0.05,11} = 19.67$, we reject the null hypothesis and conclude that at least one of the interaction effects are not zero. Therefore, the four independent variables are not independent.

(d) The model is

$$\log \lambda_{ijkl} = \mu + \alpha_i + \beta_j + \delta_l + (\alpha\beta)_{ij} + (\alpha\delta)_{il} + (\beta\delta)_{jl}.$$

Try to interpret it by yourself. Keep in mind that if the interaction between two independent variable is not in the model, then they are independent.

(e) The residual degree of freedom is 9. Since $G^2 = 9.561 < \chi^2_{0.05,9} = 16.92$, we accept the model.

3. (a) Let n be the total number of the subjects. We can choose a random order 1 to n and assign the i-th number to the i-th subject. If the i-th number is less than or equal to $n/2$, then for (D1) we assign one treatment otherwise we assign the second treatment, for (D2) we assign A first then B otherwise B first and A.

(b) For (D1) treatment A and treatment B effect are independent measure, but they are dependent in (D2). However, the treatment effect is nested in subject in (D1) but not in (d2).

(c) For (D1), let $\bar{Y}_1$ be the average of treatment 1 and $\bar{Y}_2$ be the average of the treatment 2. Then

$$V(\bar{Y}_1 - \bar{Y}_2) = 2 \times 10^2/50 = 4.$$

For (D2), let y_{i1} be the measurement of treatment A and y_{i2} be the measurement of treatment B. Let $\delta_i = y_{i1} - y_{i2}$, and $\hat{\delta} = \sum_{i=1}^{100} \delta$.

$$V(\delta_i) = 2 \times 10^2 - 2 \times 0.82 \times 10^2 = 36.$$

Then

$$V(\bar{\delta}) = \frac{36}{100} = 0.36.$$

Therefore, the second measure has lower variance, which is more powerful if the treatment mean is not affected.

(d) We need to include: (i) the nested effect in (D1) and the dependence in (d2); (ii) the power function comparison.

4. (a) An exponential family PMF (or PDF) has the form of

$$f(y, \theta, \phi) = \exp \frac{y\theta - b(\theta)}{a(\phi)} + c(\theta, \phi).$$

The PMF of the negative binomial is

$$f(y; k, \mu) = \exp\{y \log \frac{\mu}{\mu + k} + k \log \frac{k}{\mu + k} - \log \frac{\Gamma(y + k)}{\Gamma(k)\Gamma(y + 1)}.$$

Because k is known, the above is an exponential family if we use $\theta = \log \frac{\mu}{\mu+k}$. In this case, we have $b(\theta) = -k \log(1 - e^{\theta})$ and $\phi = 1$. Thus,

$$E(y) = b'(\theta) = \mu.$$

(b) The canonical link is $g(\mu) = \theta$ which is

$$g(\mu) = \log \frac{\mu}{\mu + k}.$$

(c) We consider the model

$$\log \frac{\mu_i}{\mu_i + k} = \alpha + \beta x_i$$

where $x = 1$ if the i-th person is a patient, and $x = 0$ if not. Let $x_i = 1$ for $i = 1, 2, 3$ and $x_i = 0$ for $i = 4, 5, 6$. Then $\mu_i = k/(1 - e^{\alpha})$ for $i = 1, 2, 3$ and $\mu_i = k/(1 - e^{\alpha+\beta})$ for $i = 4, 5, 6$. The loglikelihood function is

$$\begin{aligned}
\ell(\alpha, \beta) =& \sum_{i=1}^{6} f(y_i; k, \mu_i) \\
=& \sum_{i=1}^{6} \log \frac{\Gamma(y_i + k)}{\Gamma(k)\Gamma(y_i + 1)} + 3k \log(1 - e^{\alpha}) + 3k \log(1 - e^{\alpha+\beta}) \\
&+ \alpha(y_1 + y_2 + y_3) + (\alpha + \beta)(y_4 + y_5 + y_6).
\end{aligned}$$

5. (a) Matching can eliminate the age, gender, location, and SES effects. Otherwise, we do not know the different is caused by WCV or other effects.

(b) For the prior year, the odds ratio is

$$\hat{\theta} = \frac{509 \times 1593}{407 \times 1491} = 1.336.$$

For the post year

$$\hat{\theta} = \frac{672 \times 1625}{375 \times 1328} = 2.193.$$

(c) The main effect of wcv reflects the odds ratio for prior year. When it is combined with the interaction effect, it reflects the odds ratio for the post year.

(d) The interaction effect reflects the change of the odds ratio. We have

$$\log(2.192) - \log(1.336) = 0.4955$$

which is close to the interaction effect.

(e) This model is a saturated model. It is equivalent to the method using original data.

6. (a) State is a fixed effect, Farm (State) and Cow (Farm*State) are random effects.

(b) There is no degree of freedom in the error terms because the estimation of Cow effect used all the observations. Actually, the model is a saturated model.

(c) DF are 4, 120, 1750, respectively.

(d) The F-statistic is

$$F^* = \frac{4289/4}{71463/1750} = 26.3$$

which is $\geq \chi_{0.05,4} = 9.48$. Thus, they are different.

(e) The estimate of the variance is $\hat{\sigma}^2 = 71463/1750 = 40.84$. The variance of the mean is

$$\hat{Y}_{5\cdot\cdot} = \frac{\hat{\sigma}^2}{375} = 0.1089$$

The 95% confidence interval is

$$34.70 \pm 1.96 \times \sqrt{0.1089} = [34.05, 35.35].$$