CPH 931 — Fall 2008 — Dr. Charnigo

Written Assignment 3

Written Assignment 3 is due on Wednesday 05 November at 2:30 p.m. under my office door (CPH 203-B) or via e-mail (Word 2003 or PDF) to {RJCharn2@aol.com}.

Acquire the cancer data set from {www.sph.emory.edu/∼dkleinb/logreg2.htm}, which contains information on 288 women diagnosed with endometrial cancer. The variable GRADE is an ordinal response (0 = well differentiated, 1 = moderately differentiated, 2 = poorly differentiated), the variable SUBTYPE is a nominal response (0 = adenocarcinoma, 1 = adenosquamous, 2 = other), and the other variables are dichotomous predictors (AGE: 1 = 65 to 79, 0 = 50 to 64; SMK: 1 = current smoker, 0 = not a current smoker; RACE: 1 = black, 0 = white; ESTROGEN: 1 = used estrogen, 0 = did not use estrogen).

[50] 1. Consider the cancer data set from {www.sph.emory.edu/∼dkleinb/logreg2.htm}. For convenience put X_1 = AGE, X_2 = SMK, X_3 = RACE, and X_4 = ESTROGEN.

[10] a. Treating SUBTYPE as the response variable, fit the polytomous regression model

$$\log\left[\frac{p_{1|\mathbf{x}}}{p_{0|\mathbf{x}}}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4, \quad \log\left[\frac{p_{2|\mathbf{x}}}{p_{0|\mathbf{x}}}\right] = \alpha^* + \beta_1^* x_1 + \beta_2^* x_2 + \beta_3^* x_3 + \beta_4^* x_4,$$

where $p_{1|\mathbf{x}}$, $p_{0|\mathbf{x}}$, etc., have the meanings provided in Lecture 6. Report parameter estimates, standard errors, and p-values.

[10] b. Can SMK be removed from the polytomous regression model in part a? Specify an appropriate null hypothesis in terms of model parameters and then state whether the null hypothesis is accepted or rejected.

[10] c. Suppose that a non-smoking white woman aged 60 who took estrogen has endometrial cancer. Use the polytomous regression model in part a to estimate the probability that her SUBTYPE is adenosquamous.

Hint: First estimate $p_{1|\mathbf{x}}/p_{0|\mathbf{x}}$ and $p_{2|\mathbf{x}}/p_{0|\mathbf{x}}$. Then use

$$p_{0|\mathbf{x}} + p_{1|\mathbf{x}} + p_{2|\mathbf{x}} = 1 \iff 1 + \frac{p_{1|\mathbf{x}}}{p_{0|\mathbf{x}}} + \frac{p_{2|\mathbf{x}}}{p_{0|\mathbf{x}}} = \frac{1}{p_{0|\mathbf{x}}}$$

to estimate $p_{0|\mathbf{x}}$, $p_{1|\mathbf{x}}$, and $p_{2|\mathbf{x}}$.

[10] d. Treating GRADE as the response variable, fit the ordinal logistic regression model

$$\log[O_{2|\mathbf{x}}] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4, \quad \log[O_{12|\mathbf{x}}] = \alpha^* + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4,$$

where $O_{2|\mathbf{x}}$, $O_{12|\mathbf{x}}$, etc., have the meanings provided in Lecture 6. Report parameter estimates, standard errors, and p-values.

[10] e. Suppose that a non-smoking white woman aged 60 who took estrogen has endometrial cancer. Use the ordinal logistic regression model in part d to estimate the probability that her GRADE is well differentiated.

Hint: Begin by estimating $O_{2|\mathbf{x}}$ and $O_{12|\mathbf{x}}$ for such an individual. Noting that $O_{2|\mathbf{x}} = p_{2|\mathbf{x}}/(1-p_{2|\mathbf{x}})$, you can use your estimate of $O_{2|\mathbf{x}}$ to solve for an estimate of $p_{2|\mathbf{x}}$. Then, noting that $O_{12|\mathbf{x}} = (p_{1|\mathbf{x}}+p_{2|\mathbf{x}})/(1-p_{1|\mathbf{x}}-p_{2|\mathbf{x}})$, you can use your estimates of $O_{12|\mathbf{x}}$ and $p_{2|\mathbf{x}}$ to solve for an estimate of $p_{1|\mathbf{x}}$. Finally, an estimate of $p_{0|\mathbf{x}}$ is determined by the estimates of $p_{1|\mathbf{x}}$ and $p_{2|\mathbf{x}}$.

[50] 2. Consider the SARS data set discussed in Lecture 7 and available at
{http://www.richardcharnigo.net/CPH931F08/SARS.xls}.

Note: In all that follows please confine your attention to the data from TAIWAN.

[10] a. Fit a Poisson regression model with DAILYINF as the response variable and TIME as the sole explanatory variable. What is the model-based estimate of the incidence rate ratio comparing day 55 to day 45? What about day 85 to day 75?

Hint: For a generic Poisson regression model, we have

$$\frac{\mu_{\mathbf{x}_{new}}}{\mu_{\mathbf{x}_{old}}} = \frac{\exp[\alpha + \beta_1(x_1+10) + \beta_2 x_2 + \cdots + \beta_k x_k]}{\exp[\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k]}$$

when $\mathbf{x}_{new}$ and $\mathbf{x}_{old}$ differ only by 10 units on X_1. The above expression can, of course, be simplified.

[10] b. Make a plot of DAILYINF versus TIME. Do your answers to part a <u>seem</u> reasonable?

[10] c. Refit the Poisson regression model by including the quadratic term TIME2 along with the linear term TIME. Now what is the model-based estimate of the incidence rate ratio comparing day 55 to day 45? What about day 85 to day 75?

Hint: Modify the hint for part a by changing the numerator x_2 to $(x_1+10)^2$ and the denominator x_2 to x_1^2.

[10] d. Are the two numbers in your answer to part c significantly different from each other? Specify an appropriate null hypothesis and then state whether the null hypothesis is accepted or rejected.

[10] e. Does your answer to part d change if you correct for overdispersion? Investigate with both a multiplicative adjustment to the standard errors in Poisson regression and negative binomial regression.