STAT 8004, Homework 7

Group # ... (Replace this)
Members: ... (Replace this)

Apr. 10, 2014

This homework is due Thu., 2014/04/17, 5:30pm.

Instructions: Generate a PDF file from it and submit the PDF file to blackboard. Each group should submit one file with file names **hw[number]-[groupnumber].pdf**. For exmaple, "hw01-1.pdf" for homeowrk 1 and group 1. Please also include your R code in the appendix.

Problem 1. (40 points) Consider the model $Y_i \sim \text{Bernoulli}(\pi_i)$, with

$$\log(\pi_i) = \alpha + \beta x_i.$$

Here $x_i = 0, 1$ is the indicator of the exposure group.

a) (10 points) Consider the model in the 2×2 contingency table setting. What type of contigency table would you like to consider, conditional or unconditional? Suppose we are going to test H_0 whether the odds ratio is 1, how could you formulate the test with α and β?

You can use your results for the previous homework.

b) (10 points) Derive a Wald test for H_0.

c) (10 points) Derive a likelihood ratio test for H_0.

d) (10 points) Derive a score test for H_0.

Problem 2. (50 points) Read Chapter 7.2.3 in the Lachine's. Now we introduce another type of notation to discuss the same problem. For the bivariate model we assume that

$$E(y \mid x_1, x_2) = g^{-1}(\alpha + x_1\beta_1 + x_2\beta_2)$$

for some smooth link $g(\cdot)$. For simplicity we assume that X_1 and X_2 are statistically independent and that X_2 has probability $\mathbb{P}(X_2 = 1) = \phi$. Now consider the values of the coefficients when X_2 is dropped from the model to yeild $\mathbb{E}(y \mid x_1) = g^{-1}(\tilde{\alpha} + x_1\tilde{\beta}_1)$. Note that this notation is simpler than that used in Section 7.2.3.

a) (10 points) Show that the coefficient in the reduced model can be obtained as

$$\mathbb{E}\,(y \mid x_1) = g^{-1}(\tilde{\alpha} + x_1\tilde{\beta}_1) = \mathbb{E}_{x_2}[\,\mathbb{E}\,(y \mid x_1, x_2)] = (1-\phi)g^{-1}(\alpha x_1\beta_1) + \phi g^{-1}(\alpha + x_1\beta_1 + \beta_2).$$

b) (10 points) Let $g(\cdot)$ be the indentity function. Show that $\tilde{\alpha} = \alpha + \phi\beta_2$ and $\tilde{\beta}_1 = \beta_1$.

c) (10 points) Let $g(\cdot)$ be the log function, $g^{-1}(\cdot) = \exp(\cdot)$. Show that $\tilde{\alpha} = \alpha + \log(1-\phi+\phi e^{\beta_2})$ and $\tilde{\beta}_1 = \beta_1$.

d) (10 points) Let $g(\cdot)$ be the logit function and $g^{-1}(\cdot)$ the inverse logit. Assume that X_1 is a binary variable. Show that

$$\tilde{\alpha} = \log\left(\frac{\pi_0}{1-\pi_0}\right), \quad \tilde{\beta}_1 = \log\left(\frac{\pi_1}{1-\pi_1}\right) - \tilde{\alpha},$$

where

$$\begin{aligned} \pi_0 &= \mathbb{E}\,(y \mid x_1 = 0) = \frac{1-\phi}{1+e^{-\alpha}} + \frac{\phi}{1+e^{-(\alpha+\beta_2)}} \\ \pi_1 &= \mathbb{E}\,(y \mid x_1 = 1) = \frac{1-\phi}{1+e^{-(\alpha+\beta_1)}} + \frac{\phi}{1+e^{-(\alpha+\beta_1+\beta_2)}}. \end{aligned}$$

Here $\tilde{\beta}_1$, in general, does not equal to β_1. Note that $e^a/(1+e^a) = (1+e^{-a})^{-1}$.

e) (10 points) Assume that a logit model with $\alpha = 0.2$, $\beta_1 = 0.5$ and $\beta_2 = 1.0$, where $\phi = 0.4$. If X_2 is dropped from the model, show that $\tilde{\alpha} = 0.5637$ and $\tilde{\beta}_1 = 0.4777$.

Problem 3. (50 points) The data “alchyp.txt” comes from a small study in Western Australia of hypertension, alcohol, and obesity. This study was partly designed to mimic a previously reported U.S. study based on a larger sample. A log-linear interaction model is a convenient and effective way of investigating associations among the three variables. A prior-posterior analysis of this 3 x 2 x 4 contingency table using prior information from the previous study (Klatsky et al., 1977) may be appropriate. The previous study reported the general conclusion that alcohol intake and obesity were significantly and independently associated with hypertension (blood pressure). Although a few summary statistics were reported, the full data were not published. One difference between the two studies was in the definition of obesity categories.

The data is listed as follows: the first column (Obesity) contains a numerical value representing the level of obesity (1=low, 2=average, 3=high), the second column (BP) contains a numerical indicator of the presence of hypertension (0=no, 1 =yes). The next five columns are labelled with the levels of alcoholic intake of the subjects, in drinks per day. These columns contain the frequency of observations that have this level of intake, for each group of obesity level and hypertension presence.

a) (10 points) Transform the data to the format that can be handled by logistic model.

b) (10 points) Build up a logistic regression model with two main effects (obesity and alchohol intake) and no interactions. Interprete each parameter in your model.

c) (10 points) Estimate the parameters in your model. What is the risk of having hypertension in each obesity cross alchohol group? List your results into a table.

d) (10 points) Build up a logistic regression model with two main effects (obesity and alchohol intake) and also interaction terms. Interprete each parameter in your model.

e) (10 points) Estimate the parameters in your model. Is the interaction term significant? Assume we would like to keep the interaction term in the model. What is the risk of having hypertension in each obesity cross alchohol group? List your results into a table.