BMTRY 711 Analysis of Categorical Data
Spring 2011
Exam 2
Total: 100 points
Due: Apr 15, 2011 at 5.00PM

NAME:

INSTRUCTIONS:

1. Work independently. You may freely use your texts, notes and any other printed resources. You cannot discuss the assignment with individuals (i.e., faculty, students, colleagues, etc.) other than me. Deviations from this policy will result in a failing grade for the class.
2. You should prepare a written summary of the data using a combination of written text, figures and tables. No SAS code should be a part of your write up (you may include it as an appendix, though). You may assume the audience will be familiar with the methods used to summarize the data, so you do not have to elaborate on methodology; however, you should cite the method used so that the results would be reproducible.
3. There is no minimum or maximum number of pages required for the write up. You should exercise your own judgement on how much summary is required to answer the proposed questions. In general, if the question asks if there is an association, a response of "yes" or "no" is NOT sufficient.

- Problem 1:(60 points) DATASET DESCRIPTION:
  The data is based on a randomized-placebo controlled clinical trial involving 326 participants across 5 clinical centers. The clinical area is generalized anxiety disorder, and the primary outcome is whether or not a participant responded to treatment. "Response" is defined as a 50% reduction from baseline in the Hamilton Anxiety Scale (HAM-A) total score. Participants are followed for up to 8 weeks. Data set is located in the course webpage as exam2prob1.txt

PROBLEMS:

1. Estimate the percentage of responders in each treatment arm and conduct an appropriate test. Summarize the association.
2. The randomization plan stratified the randomization by clinical site. Therefore, clinical site should be considered a stratification variable in the analysis regardless of statistical significance.
   (a) Test for a common odds ratio across site.
   (b) If appropriate, estimate the common odds ratio, with confidence interval, by two methods.
3. The 50% reduction in the total score can be affected by the length of time an individual takes the medication. Derive a new binary indicator variable that represents completers (MAXWEEK = 8), and test for a common odds ratio when the response and treatment group is stratified by this variable.
4. Using appropriate combinations of treatment, baseline score, clinical site, and trial completion status, develop a multiple logistic regression model.

(a) Estimate the goodness of fit using the deviance. Hint: I have 12 *df* for my deviance. Yours may differ.

(b) Fit a higher order model and calculate the change in deviance. Comment on model fit.

(c) Estimate the odds ratio, with 95% CI, that an individual on the active drug who completes the trial will respond when compared to an individual on placebo that completes the trial. (Hint: Create your own dummy codes for the variables. Write down the design matrix for both groups and subtract. Use this result and the SAS help file to write the appropriate ESTIMATE statement.)

(d) Estimate the same comparison this time using non-completers.

(e) Estimate the odds ratio of response to treatment for subjects that are on active compound who complete the trial versus the subjects on active compound that do not complete the trial.

5. Does the active compound work? Describe the circumstances of the highest likelihood of response.

- Problem 2. (60 points) DATASET DESCRIPTION:
Five groups of animals were exposed to a dangerous substance in varying concentrations. Let z_i be the $\log_{10}$(concentration), n_i be the number of animals, y_i be the number died and p_i be the proportion died in a particular group. This dose-response data is given below:

Concentration	z_i	n_i	y_i	p_i
0.00001	-5	6	0	0
0.0001	-4	6	1	0.167
0.001	-3	6	4	0.667
0.01	-2	6	6	1
0.1	-1	6	6	1

PROBLEMS:

1. Plot the empirical logits versus $\log_{10}$(concentration). Could the relationship be linear ?

$$\text{Hint: Empirical logit} = \log\Big(\frac{y_i + 0.5}{n_i - y_i + 0.5}\Big)$$

2. Using SAS PROC LOGISTIC, or whatever software you like, fit the model

$$\log\Big(\frac{\pi}{1-\pi}\Big) = \beta_0 + \beta_1 * \log_{10}(\text{concentration})$$

where π_i is the probability of death for the ith group. Then test the null hypothesis $H_0 : \beta_1 = 0$ using (a) Wald test and (b) a likelihood ratio test. What do you find ?

3. Find an estimate and approximate 95% confidence interval for LD_{50}. (Hint: LD_{50} is the dose level for which about 50% of the individuals will be dead.)

4. Consider π_2, the probability of death at concentration level 0.0001. Calculate approximate 95% confidence intervals for π_2 in three different ways: (i) by applying a normal approximation to the raw proportion $p_2 = \frac{y_2}{n_2}$, without using the logit model; (ii) by applying a normal approximation to the fitted value of logit(π_2) under the model and taking the expits $\big(\text{expit(x)} = \frac{e^x}{1+e^x}\big)$ of the endpoints; (iii) by applying a normal approximation to the fitted value $\hat{\pi}_2$ under the model. Compare the intervals. Assuming that the true model is true, which procedure is probably the best ? Which is probably the worst ? Why ?