STAT 557 **ASSIGNMENT 4** **Fall 2017**

Reading Assignment: Agresti, Chapters 6, 8, 9 and 10.

Written Assignment: Due Tuesday, November 14, in class

1. Between 1977 and 1980 the National Toxicology Program (NTP) of the US Department of Health and Human Services conducted a toxicology and carcinogenesis bioassay of a polybrominated biphenyl mixture (PBB), a flame retardant known as Firemaster F-1. As part of the study, Fisher 334 rats were given 125 oral doses of PBB over a period of six months beginning at about seven weeks after birth. Both male and female rats were used. Occurrence or non-occurrence of a nonlethal lesion, bile duct hyperplasia, was recorded for each of 319 rats for which pathology reports were available. The data are posted in the file

bile.dat

with one line for each rat. There are seven numbers on each line. The first is the rat identification number. This is followed by values for sex, dose level of PBB, initial weight of the rat (in grams), cage tier, age (in weeks) when the rat died or was sacrificed and examined, the response variable. Sex is coded as 1 = female and 0 = male. The response is coded as 1 = hyperplasia present and 0 = hyperplasia absent. The six dose levels were 0, 0.1, 0.3, 1.0, 3.0, and 10.0. The cage tier variable has values 1, 2, 3, 4, 5 corresponding to the tier in which the rat's cage was housed during the six-month experiment, with 1 = top, …, 5 = bottom. SAS code posted in the file **bile.sas** can be used to read the data file and answer some of the following questions. R code is posted in the file **bile.R.**

A. First fit the logistic regression model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Z_{3i} + \beta_4 Z_{4i} + \beta_5 Z_{5i} + \beta_6 Z_{6i} + \beta_7 Z_{7i} + \beta_8 Z_{8i}$$

where Z_1, Z_2, Z_3, Z_4 are dummy variables corresponding to the five cage tier levels, Z_5 denotes sex of the rat, Z_6 is initial weight, Z_7 is age at examination, Z_8 is the dose of PBB, and π denotes the probability of bile duct hyperplasia. According to this model, which of the variables appear to be associated with the incidence of bile duct hyperplasia?

B. How would you interpret the coefficient β_5 in the model in part (A)? Construct a 95 % confidence interval for $\exp(\beta_5)$.

C. How would you interpret $\exp(\beta_8)$ in the model in part (A)? Construct a 95 % confidence interval for $\exp(\beta_8)$.

D. Construct plots of the Pearson residuals versus Z6, Z7 and Z8. Construct similar plots of deviance residuals. Pass smooth curves through these plots (you do not have to submit these plots, just look at them on your computer screen). What do these plots suggest?

E. Try adding interaction terms Z_5Z_6, Z_5Z_7, Z_5Z_8, Z_6Z_7, Z_6Z_8, Z_7Z_8 to the model in part (A). Search for a good model. Did this search produce a model that suggests that increased exposure to PBB increases the risk of bile duct hyperplasia? If so, describe the relationship.

F. For the model you selected in part (E), examine diagnostics on leverage and influence of individual cases on estimates of model parameters to determine if there are highly influential cases or outliers. State your conclusions. (Do not submit plots or lists of diagnostic statistics like DFBETAS, just comment on unusual or highly influential cases if any exist.

G. Apply a clustering procedure to the explanatory variables to form about 20 clusters. Hierarchical clustering can be done with the CLUSTER procedure in SAS, or with the hclus function in R. Compute a local mean deviance plot. Submit this plot. What does this plot suggest? Is there any information in the estimated effects for the clusters?

H. Do whatever additional analyses of these data you think would be useful to select a final model (or models). Report a formula for your model with estimates of parameter values substituted into the formula. Report standard errors for parameter estimates in parentheses beneath the parameter estimates. Write a short paragraph describing your conclusions about possible effects of the level of exposure to PBB on prevalence of bile duct hyperplasia.

2. The table shown below displays data from a study of the relationship of age and marital status in Demark. An individual is classified as divorced if he or she had been divorced at any time prior to the survey, regardless of whether or not that individual is currently remarried. Consequently, an individual is classified as single if the individual has never married. An individual is classified as married if the individual has married once and has never been divorced. The source of the data did not reveal how individuals who are widowed and never divorced are classified; perhaps those people are excluded from the study. Furthermore, the source of the data treats these data as a simple random sample of 185 individuals from the population of Danes who are at least 16 years old. At the time of the survey, the legal age of marriage in Denmark was 16.

		Observed Counts			
Age Group	X	Single	Married	Divorced	Total
17 – 21	19	17	1	0	18
21 – 25	23	16	8	0	24
25 – 30	27.5	8	17	1	26
30 – 40	35	6	22	4	32
40 – 50	45	5	21	6	32
50 – 60	55	3	17	8	28
60 – 70	65	2	8	6	16
70+	75	1	3	5	9

A variable, X, was created to model trends across age. We will do the same, although it would be better to have the actual age of each respondent. These data have been posted as the file **dmstatus.dat** in the data folder of the course Blackboard page. Code for part (a) has been posted at **dmstatus.sas** and **dmstatus.R** on the course Blackboard page.

A. Using marriage as the baseline category, fit the following polychotomous logistic regression model to the data (call it model A). Report a table of estimated coefficients and their standard errors

$$\log\left(\frac{\pi_{\text{sin gle},i}}{\pi_{\text{married},i}}\right) = \beta_0 + \beta_1(X_i - 16)$$

$$\log\left(\frac{\pi_{\text{divorced},i}}{\pi_{\text{married},i}}\right) = \alpha_0 + \alpha_1(X_i - 16)$$

i) Report a value of a log-likelihood ratio G^2 test, and its degrees of freedom and p-value, for testing the fit of this model against the general alternate of eight different and independent multinomial distributions for the eight age groups. State your conclusion. Consider if you can trust the reported p-value.

ii) Give interpretations of α_0 and α_1 for this model.

iii) Create a plot of the estimated proportions against age. The plot should have one curve for each marital status category. This plot should also display the observed proportions of respondents in the three response categories at each age level. What does this plot reveal?

B. Fit the following model to the data (call it model B).

$$\log\left(\frac{\pi_{\text{sin gle},i}}{\pi_{\text{married},i}}\right) = \beta_0 + \beta_1(X_i - 16) + \beta_2(X_i - 16)^2$$

$$\log\left(\frac{\pi_{\text{divorced},i}}{\pi_{\text{married},i}}\right) = \alpha_0 + \alpha_1(X_i - 16) + \alpha_2(X_i - 16)^2$$

i) Report a value of a log-likelihood ratio G^2 test, and its degrees of freedom and p-value, for testing the fit of this model against the general alternate of eight different and independent multinomial distributions for the eight age groups. State your conclusion.

ii) Create a plot of the estimated proportions against age, with one curve for each marital status category. This plot should also display the observed proportions of respondents in the three response categories at each age level. Comparing this plot to the plot in part A(iii), describe how model B provides improvement over model A, if at all.

C. Fit the following model (call it model C) to the data.

$$\log\left(\frac{\pi_{\text{sin gle, i}}}{\pi_{\text{married, i}}}\right) = \alpha_1 + \beta_1 \log(X_i - 16)$$

$$\log\left(\frac{\pi_{\text{divorced, i}}}{\pi_{\text{married, i}}}\right) = \alpha_2 + \beta_2 \log(X_i - 16)$$

i) Report a value of a log-likelihood ratio G^2 test, and its degrees of freedom and p-value, for testing the fit of this model against the general alternate of eight different and independent multinomial distributions for the eight age groups. State your conclusion.

ii) Create a plot of the estimated proportions against age, with one curve for each marital status category. This plot should also display the observed proportions of respondents in the three response categories at each age level. Does this model appear to fit the data as well as model B?

D. Fit the following model (call it model D) to the data:

$$\log\left(\frac{\pi_{\text{sin gle,i}}}{\pi_{\text{married,i}}}\right) = \alpha_1 + \beta_1 \log(X_i - 16) + \gamma_1 [\log(X_i - 16)]^2$$

$$\log\left(\frac{\pi_{\text{divorced, i}}}{\pi_{\text{married, i}}}\right) = \alpha_2 + \beta_2 \log(X_i - 16) + \gamma_2 [\log(X_i - 16)]^2$$

E. For models A, B C, and D, create a table with one row for each model and columns for the number of parameters in the model, the -2(log-likelihood) or deviance value, AIC, and SBC (or SC). Which of the four models is "best" for these data? Write a few sentences to support your decision. You may define what you mean by "best".