Homework 6, Due Monday, October 24, 2016

You are welcome to use R but attach the code.

1. The data are described in Problems 9.1 on pages 144-145 and the problem continues as Problem 10.3 on page 163. Refer to the specific questions below. R code to load these data into R is attached. Use logistic regression to answer the questions below.

   (a) Calculate the ratio of the odds of TB given exposure to biomass fuel over the odds of TB given no exposure for the pooled data. Also give an approximate 95% confidence interval and interpret in terms of the problem. The data you need are in `TB.pool`. Fit a logistic regression model with just biomass as the explanatory variable. Compare the results to those you got on the midterm exam.

   (b) Provide me with an estimate of the common odds ratio adjusted for income along with an approximate 95% interval. Fit a logistic regression model with biomass and income as explanatory variables. Compare these results to the those you got using the CMH method on the exam.

   (c) Fit a model assuming an interaction between biomass and income. Use a drop-in-deviance test to assess the evidence for the existence of an interaction. Compare the results to the results you got from the Breslow-Day method on the exam.

   (d) Summarize the results as you did on the exam.

2. Extra Credit for STAT 425 and required for STAT 525: Above you were able to get results when you fit the model with an interaction in it. Note however that the model fit perfectly (residual deviance of 0 on 0 df). We were even able to compare this model to the simpler additive model with a drop-in-deviance test. I computed the sample proportions, transformed them to logits, and attempted to fit an ordinary least squares (OLS) model involving the interaction. I got the following.

```
p<-cases/(cases+controls)
logit<-log(p/(1-p))
summary(lm(logit~biomass*income))
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              -0.3238         NA      NA       NA
biomassYes                1.4765         NA      NA       NA
income>1000              -0.7116         NA      NA       NA
biomassYes:income>1000   -0.1534         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:    NaN
F-statistic:   NaN on 3 and 0 DF,  p-value: NA
```

This model did not fit. Why can we get valid results fitting a logistic regression model but not the OLS model? (The answer has nothing to do with the OLS fit being invalid because it is applied to proportions. For example, compare the coefficient estimates from the OLS fit to the logistic regression model fit.).

3. We looked at using `prop.trend.test` to test for a linear trend in risk. Recall the data were the presence or absence of coronary heart disease with body weight being the risk factor. Weight was binned into 5 ordinal categories. The data were.

```
chd<-array(c(32,31,50,66,78,558,505,594,501,739),dim=c(5,2),
dimnames=list(c("<150","151-160","161-170","171-180",">180"),
 c("CHD","no CHD")) )
chd
        CHD no CHD
<150     32    558
151-160  31    505
161-170  50    594
171-180  66    501
>180     78    739
# get the number of trials by row
n<-rowSums(chd)
# get the counts
d<-chd[,1]
prop.trend.test(d,n)
     Chi-squared Test for Trend in Proportions

data:  d out of n ,
 using scores: 1 2 3 4 5
X-squared = 15.361, df = 1, p-value = 8.88e-05
```

Recall that the test is based on fitting a weighted least squares model with the proportions as responses and the weight categories treated as a quantitative variable with values 1 through 5. We will fit a logistic regression model to these data to test for a linear trend.

```
score<-1:5
fit<-glm(cbind(chd[,1],chd[,2])~score,family=binomial)
summary(fit)
```

We talked about the `Residual Deviance` but not about the `Null Deviance` This is the deviance associated with the reduced model that arises if all the regression coefficients are 0. In this case it is the deviance associated with the null hypothesis $H_0 : \beta_1 = 0$. To test for a trend we compare the full model

$$logit(\pi(x)) = \beta_0 + \beta_1 x$$

to the reduced model

$$logit(\pi(x)) = \beta_0$$

(a) Using the residual and null deviances carry out a drop in deviance test to compare these two models. Compare the results (value of test statistic and p-value) to those from `prop.trend.test`.

(b) What advantages do the logistic regression model have over the results from `prop.trend.test`?

Agresti notes that the test statistic from the Cochran-Armitage trend test is related to a score test statistic associated with the test for a linear trend in the logistic regression model. With large sample sizes the drop-in-deviance test statistic and the score statistic will be fairly close. Also, this answers in an indirect way Kyle's question about why the trend test statistic of Chapter 11 *only has one df.*

4. Suppose you have a single categorical risk factor measured at 3 levels (L = low, M = medium, and H = high). Assuming that H is chosen to be the reference category define two dummy variables X_1 and X_2 for the L and M groups. A logistic regression model

$$logit\pi(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

will be fit to the data.

(a) Interpret the parameters in terms of relevant odds ratios.

(b) Are there any structural relationships imposed on the odds ratios by this model? (Hint: This question may make more sense after you read the next problem).

5. The labeling of the categorical variable in the previous question implies a natural ordering. Suppose you code the risk factor as a single quantitative variable X with values 0 for L, 1 for M, and 2 for H.

(a) Describe an appropriate logistic regression model for this situation, providing an interpretation of the parameters.

(b) Are there any structural relationships imposed on the odds ratios by this model?