BIOS663 Homework 6

Investigators are interested in the effect of dermal nicotine exposure in a population of Latino tobacco workers in North Carolina. (Nicotine can be absorbed from tobacco leaves through the skin and can cause nicotine poisoning, which is characterized by nausea, vomiting, headache, and dizziness.) Data were collected on tobacco work tasks and risk factors for exposure to nicotine during a summer tobacco work season. Nicotine exposure was measured by levels of cotinine, a nicotine metabolite, contained in saliva. Other covariates of interest include age, body mass index, education, work conditions (working in wet conditions is believed to increase nicotine absorption), type of tobacco work ("priming" refers to picking or harvesting the tobacco and is expected to result in highest nicotine exposures, "barning" refers to putting the harvested tobacco into a barn for curing, "topping" refers to breaking the flower off the top of the plant, and "other" refers to farm work that does not involve tobacco contact, such as driving a truck), and smoking (smokers would also have nicotine exposure through cigarettes, and it is not known whether exposure to tobacco leaves would increase cotinine levels to a similar extent in both smokers and non-smokers).

The variables are available in the file Nicotine.dat on the BLACKBOARD website. They are listed in the file in the following order.

- COTININE: salivary cotinine concentration (in ng/mL)
- AGE: age (in years)
- BMI: body mass index (in kg/m^2)
- EDUC: years of education
- WET: takes value 1 if work conditions on day of measurement were wet and takes value 0 otherwise
- TASK: takes value 1 for priming, 2 for barning, 3 for topping, and 4 for other work not involving tobacco contact
- LNNSMOKE: natural logarithm of (1 + number of cigarettes smoked per day)

To report a test, provide H_0, the test statistic, the degrees of freedom, the p-value, the decision (accept/reject H_0), and an interpretation of the result in terms of the subject matter. The entire homework should be typed neatly, and no SAS code or output should appear along with the assignment. It should be in the form of a professional report.

1. Dichotomize salivary cotinine concentration based on its median value, so that we can divide these tobacco workers into two groups: those with high or low level of cotinine. Use this dichotomized variable, denoted by cotinineBinary, as response variable.

(a) Report its association with variable 'task', and the odds ratios for task level 1 vs 4, 2 vs. 4 and 3 vs. 4.

Using reference cell coding, and kept level 4 as reference level, the odds ratios are the $\exp(b_j)$, $j = 1, 2, 3$, where b_j is the logistic regression coefficient for task level 1, 2, and 3.

(b) Still consider the logistic regression with task as the only covariate. Provide a table of parameter estimates and standard errors using (a) cell mean coding and (b) reference cell coding of task, and give the interpretations of parameters in both coding schemes.

The SAS code would be similar to the code in slide 24 of lecture 17. The output is skipped here. For cell mean coding, the four covariates are four indicators of each task group, and the coefficients are the log odds of each group. For reference cell coding, the four covariates are intercept plus indicators of three task groups. The intercept is the log odds of the baseline group, and each other coefficient is the log odds ratio of one group vs. the reference group.

(c) Use both task, wet as covariates to fit a logistic model. *Based on this model*, evaluate model fitting by a Chi-square test.

Following the code in lecture notes. The Chi-square test will not be valid if you add interaction into your model. The idea of Chi-square test is to compare a subset model with the full model. However, if task, wet and task*wet are all included in the model, you are estimating means of each cell, which is a full model. Then if you run Chi-square test, it is to compare full model vs. full model.

(d) Fit the Full Model in Every Cell: Use cotinineBinary as the response and task, wet, and lnnsmoke as predictors and explain the results.

Similar to the code for home work 5, but using logistic regression here. A full model in each cell should have 8 parameters. For example, for reference cell coding, you should have degree of freedom 16 = 1 (intercept) + 3 (task) + 1 (wet) + 1 (lnnsmoke) + 3 (task*wet) + 3 (task*lnnsmoke) + 1 (wet*lnnsmoke) + 3 (task*wet*lnnsmoke).

2. Based on their salivary cotinine concentrations, divide the samples into three groups of approximately equal sizes, and then delete the middle

group and only use the two remaining groups in the following analysis. Use the new group indicator as response variable. Repeat the logistic regression analyses (a)-(d) in question 1 using this new data and the new response variable. Compare the significance levels of the results obtained in question 2 and in question 1. Explain why you see smaller or larger p-values in question 2. Note you cannot compare models in question 2 and models in question 1 by any test, since only a subset of samples are used in question 2.

If you see warning message about "complete separation", it indicates the two groups of the response variable can be perfectly separated by the model. If you see warning message about "quasi-complete separation", it indicates the the two groups can be almost completely separated . In these situations, the maximum likelihood estimate is not reliable, and the logistic regression model might be invalid. Usually these situations occur when sample size is small. In this homework, ignore models with such warning messages.

```
The basic idea in tis question is to answer:  How should we dichotomize
a continuous variable?  We can just sort the continuous variable
and divide it into two groups of approximately equal sizes.  However,
for those observations in the middle of the sorted vector, the
dichotomization may be highly in-accurate.  Then an alternative
approach is to drop those observations in the middle, and only
use those observations with more extreme values to define the
dichotomized variable.  The drawback is then the total sample
size is reduced.  Therefore, if we see a smaller p-value in question
2 than in question 1, it means we got better dichotomization,
which beat the disadvantage due to smaller sample size.
```