**Homework from Class 8  
Logistic Regression**

1. Taking multiple vitamin and mineral supplements during pregnancy can reduce the risk of birth defects. A study (based on a sample of several thousand women) was conducted to examine factors associated with not taking vitamin supplements during pregnancy, to better target a public health campaign. The following table presents results of a multiple logistic regression predicting NOT using multiple vitamins during pregnancy.

In this analysis, all variables have been categorized and represented through dummy variables in the model, using reference category coding. A commonly used convention with logistic regression is to indicate the reference group for a categorical predictor with an OR of ‘1.00’ and to dash out the CI or p-value associated with the reference group. So, in the logistic regression below, Whites are the reference group for race/ethnicity, married is the reference group for marital status, etc.

(Yu et.al., Preconceptional and Prenatal Multivitamin-Mineral Supplement Use in the 1988 National Maternal and Infant Health Survey, AJPH 1996)

Multiple logistic regression of risk of not using multivitamin-mineral

supplements during pregnancy

|  |  |  |
| --- | --- | --- |
| Variable | OR | 95% CI |
| Race/ethnicity  Black  Asian  American Indian  White  Marital status  Unmarried  Married  Age, years  <20  20-34  35+  Education, years  < 12  12  > 12  Smoking status  Smoker  Nonsmoker | 1.27  1.14  1.01  1.00  1.08  1.00  1.75  1.28  1.00  1.73  1.59  1.00  0.90  1.00 | 1.12 , 1.36  1.02 , 1.29  0.81 , 1.23  ---  1.01 , 1.15  ---  1.55 , 1.97  1.18 , 1.38  ---  1.60 , 1.87  1.51 , 1.67  ---  0.85 , 0.95  --- |

1A. Which variables are significant predictors of not using vitamins during pregnancy? Explain.

1B. Describe the association between education and not using vitamins during pregnancy.

1C. Describe the association between smoking status and not using vitamins during pregnancy.

2. The YRBS (Youth Risk Behavior Survey) is an annual, national survey of high school students that collects data on health behaviors and conditions. The following are partial results from a logistic regression predicting whether or not a student smokes (reports any tobacco use during the past 30 days) from their grade (9, 10, 11, 12 with grade 9 as the reference category) and gender (coded 1 for females and 0 for males), based on a random sample of n=1,000 students.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Coefficient | Standard  Error | Wald Chi-Square | p-value |
| Intercept  Sex Female  Grade 10  Grade 11  Grade 12 | -1.354  -0.673  0.117  0.240  0.614 | ---  0.163  0.245  0.240  0.235 |  |  |

|  |  |  |
| --- | --- | --- |
| Variable | Odds Ratio | 95% CI for OR |
| Intercept  Sex Female  Grade 10  Grade 11  Grade 12 |  |  |

One note about the above table and R logistic regression output. With logistic regression, which focuses on a categorical outcome variable, most analysts present a ‘Wald chi-square’ statistic for the slope rather than a ‘z-test’ for the slope. I think the reason for this is just to help differentiate analyses of categorical variables (where the chi-square test is the usual analysis) and measurement variables (where t-tests and z-tests are the usual analyses). R gives a z-test for the slope in the output. The formula for the z-statistic is

z = (estimated slope) / (standard error). The formula for the ‘Wald chi-square’ is just the square of the z-value: Wald chi-square = ( (estimated slope) / (standard error) )2. Chi-square statistics with 1 degree of freedom and z-statistics give the same p-values, so the two tests are equivalent. (A ‘Wald’ test is just a test based on a parameter estimate divided by a standard error, and can be presented as either a z-statistic or a chi-square statistic.)

2A. Complete the above tables. The Wald Chi-square statistics in a logistic regression have 1 degree of freedom, and the confidence intervals for a slope from a logistic regression use a critical z-value.

2B. Describe differences in smoking for females vs. males, based on the odds ratios from the above tables. Your description should include both a statement about statistical significance, and if smoking significantly differs between females and males, a description of the direction (who has a odds of smoking, males or females?) and the magnitude (how different are the odds of smoking for females and males?) of the difference.

2C. Describe the differences in smoking across grades, based on the odds ratios from the above tables. Your description should include both a statement about statistical significance and about the direction and magnitude of differences.

2D. What is the predicted probability of smoking (calculated using the slopes from the above logistic regression), for:

- a 9th grade girl

- a 9th grade boy

- a 12th grade boy?

3. A study was conducted to examine factors associated with whether or not people vote. A sample of 500 registered voters was selected and interviewed over the phone a couple of weeks before an election. Demographic data and data on political awareness and affiliation were collected during this interview. After the election, subjects were re-contacted, and asked whether or not they voted.

Hypothetical data are saved in the attached files. Variables in the data set are:

1) an id number, ranging from 1 to 500

2) respondent’s sex, coded 0 for males and 1 for females

3) age in years

4) political party, coded as 1 for democrats, 2 for republicans, and 3 for independents

5) years of education, where 12 corresponds to a high school education, 16 to a 4 year college education

6) household income, in thousands of dollars

7) a political awareness score based on a current events test, ranging from 0 to 10, with higher values indicating higher political awareness

8) whether or not the respondent voted, coded as 0 for did not vote, 1 for voted, and 9 if the respondent could not be reached after the election.

There are no missing data for any variables other than voting.

3A. As a preliminary step in the analysis we need to account for the missing data on voting - re-code the voting variable to account for missing data:

In R, missing data are declared as NA. One way to recode 9 as missing is to use the following:

vote[vote==9] <- NA

(there are two equal signs in the square bracket without a space in between) which re-assigns vote as missing when vote equals 9. As a check, run ‘table(vote)’, which should only give values for vote=0 and vote=1.

3B. What percent of the sample voted (this percent should be based on the number with non-missing data, not on the total sample size)? Give a 95% confidence interval for the percent of registered voters who voted:

Check on the n’s and values to make sure those who didn’t vote are excluded

3C. Does the percent voting differ between males and females? Perform a chi-square test to examine this question, reporting the percent voting for males and for females along with the chi-square statistic and p-value (again, this analysis should ignore those with missing data on voting – the computer packages will automatically exclude those with missing data).

3D. What is the odds ratio describing the odds of voting for females compared to males? Give the 95% confidence interval for this odds ratio.

3E. Perform a multiple logistic regression predicting whether someone votes from their age, sex, and political party. Since political party is a categorical variable, it should be represented in the logistic regression through dummy variables, and we need to choose a reference category. Since Massachusetts is a democratic state, we will use democrat (party = 1) as the reference group.

From the logistic regression output, complete the following summary table, presenting the odds ratio and p-value for each of the independent variables in the model.

Results of a logistic regression predicting voting

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Odds Ratio | p-value | 95% CI |
| Age Sex (F vs. M)  Political Party\*  Republicans  Independents |  |  |  |

\* Democrats are the reference category

Which of the variables in the model are significantly associated with the chance that someone votes? Describe the significant associations.