MA684 Class 9 Homework

Logistic Regression II

1. From the voting study in last week’s assignment. Last week you ran a multiple logistic regression predicting whether or not a registered voter voted, based on their age, sex, and political party:

|  |  |  |
| --- | --- | --- |
| Variable | OR | p-value |
| Age (yrs)  Sex (1-female 0-male)  Political Party  Democrat  Republican  Independent | 1.02  0.96  1.00  1.26  3.62 | 0.002  0.866  ---  0.295  <0.001 |

The purpose of this question is to take a more complete look at this analysis.

1A. From the output for this logistic regression, the model chi-square is 24.91 (calculated from the difference in -2LogL for the intercept-only model (569.29) and the full model (544.38).

What null hypothesis is tested by the model chi-square?

What are the degrees of freedom for this model chi-square?

What is the p-value corresponding to this model chi-square?

1B. The C-statistic (also called the AUC) for this model is 0.64. Give an interpretation for this C-statistic.

1C. The Hosmer-Lemeshow test was also run for this logistic regression, and gave a chi-square value of 9.33 with 8 df (the degrees of freedom are the number of categories used to create ranked groups in the H-L procedure minus 2, and is given in the output) and a p-value of 0.3156.

What null hypothesis is being tested by the Hosmer-Lemeshow test? What can you conclude from the Hosmer-Lemeshow chi-square value and from the p-value from this test?

1D. We would like to get a single p-value to describe the significance of political party (represented by 2 indicator variables in the above regression) in predicting voting, controlling for age and sex. To do this, I ran a ‘reduced model’ predicting voting from just age and sex. The model chi-square for this reduced model was 7.52, with a p-value of .0247 and a C-statistic of 0.58.

From the full and reduced models, calculate the partial chi-square testing whether political party is associated with voting, controlling for age and sex.

What are the degrees freedom for this chi-square? What is the p-value for this chi-square? What is the chance in the C-statistic due to adding political party to the model?

Question 2 (based on Problem 1 from Chapter 22 of KKM and N). A study was conducted to investigate risk factors for hypertension (high blood pressure) in women. Data from a sample of 680 women are saved under the name ‘Hyperstudy’. Variables in the data set are:

1) id, an id number;

2) age, in years;

3) race, coded as 1 for whites, 2 for blacks, 3 for Hispanics, and 4 for Asians;

4) smoke, coded 1 for smokers and 0 for non-smokers; and

5) hypertension, coded 1 for those with high blood pressure and 0 for those without high blood pressure.

2A. What percent of women in this sample have high blood pressure?

Perform a multiple logistic regression, predicting hypertension from age, race, and smoking. Race should be treated as a categorical variable in this logistic regression. Based on this regression model:

2B. Give a summary table presenting results from this logistic regression. The table should include columns for odds ratios, confidence intervals for the odds ratios, and p-values.

2C. Is the overall model significant? How well does this model predict hypertension - report and interpret the C-statistic for the logistic regression model.

2D. From this multiple logistic regression model, find the odds ratio and confidence interval comparing the odds of hypertension for two subjects who differ in age by 10 years. (We talked about how to translate a change-corresponding-to-one-unit odds ratio to a change-corresponding-to-10-units odds ratio in class, but I’m not sure we talked about how to get a confidence interval for this 10-year odds ratio: first calculate the CI for the one-year odds ratio, and then translate the lower and upper bounds to get the lower and upper bounds of the CI for a 10-year odds ratio).

2E. Based on this model, describe the association between race and hypertension, including a statement about the overall significance of race in the model (based on a multiple-partial test for the 3 dummy variables for race) and a comparison across races based on the odds ratios relating to race.

2F. Based on this model, describe the association between smoking and hypertension.

3. Smoking may have a cumulative effect, so that smoking may have a greater effect in older

smokers (who have generally been smoking longer) than in younger smokers. To investigate,

run a logistic regression model predicting hypertension from age, race, smoking, and the

interaction between age and smoking.

*(To include an interaction term in a logistic regression in R, include the terms for age, race, and smoking in the model and then indicate the two terms for the interaction using an asterisk:*

*log.out <- glm(hypertension ~ age + factor(race) + smoke + age\*smoke,*

*family=binary(link=logit) )*

*To include an interaction term in a logistic regression in SPSS, enter the terms for age, race, and smoking into the model, then highlight both ‘age’ and ‘smoke’ variable in the variable list by holding the Ctrl key down when you click on the second variable and click on the ‘>a\*b>’ button.)*

3A. Based on the test for this interaction term, does the effect of smoking differ for older vs. younger women? Explain.

3B. Regardless of the significance of the interaction, use this model to find the odds ratio comparing the odds of hypertension for:

i) a 30 year old smoker vs. a 30 year old non-smoker, and

ii) a 50 year old smoker vs. a 50 year old non-smoker.

3C. I ran a Hosmer-Lemeshow test on this interaction model, placing subjects into 10 categories based on their predicted risk of hypertension. The Hosmer-Lemeshow chi-square statistic was 7.30 with 8 degrees of freedom (you could check this value). Find the p-value for this test statistic, and give an interpretation of the results of this test.