Bios 312: Modern Regression Analysis
February 15, 2012
Lab 4: Logistic regression

In this lab, we will use logistic regression to analyze the salary dataset. We will consider the covariates gender and year of degree.

Part 1: Initial dataset manipulation

1. Read in the salary dataset
2. Remove all observations that are not from 1995
3. Create an indicator variable for male gender
4. Create a new outcome variable (salhigh) to indicate if salary is above $7600 per month.
   4.1. What proportion of subjects have a salary above $7600

Part 2: High salary and gender (unadjusted model)

In part 2, we will estimate the association between salary and gender. First, we will conduct this analysis using the familiar method, a 2x2 table. We will then conduct the same analysis using logistic regression and compare the two approaches

1. Create a 2x2 contingency table of salhigh by male. From this table, calculate
   1.1. Pearson's Chi-squared test statistic (without continuity correction)
   1.2. The likelihood ratio Chi-squared test statistic
   1.3. The odds of high salary for males compared to the odds of a high salary for females (the odds ratio). Interpret
   1.4. The standard error of the log odds ratio (hint: the variance of the log odds ratio is given by $1/a + 1/b + 1/c + 1/d$ where a, b, c, and d are the cell totals)
   1.5. Pr(high salary | male)
   1.6. Pr(high salary | female)
2. Conduct a logistic regression of male on salhigh. From the regression output, calculate or identify
   2.1. The Wald Chi-squared test statistic (hint: Remeber if Z is Normal, the Z^2 follow a Chi-squared distribution)
   2.2. The likelihood ratio Chi-squared statistic
   2.3. The odds of high salary for males compared to the odds of a high salary for females (the odds ratio). Interpret the odds ratio.
   2.4. The standard error of the log odds ratio
   2.5. Pr(high salary | male)
   2.6. Pr(high salary | female)
3. Obtain estimates for the predicted probabilities 2.5 and 2.6 (use Stata or R commands to get these numbers)
4. Create an indicator variable for female gender

4.1. Fit a logistic regression model using female gender as the covariate. Compare the output to model using male as the covariate.

Part 3: High salary and year of degree (unadjusted model)

1. Create suitable summary statistics for the probability of high salary by year of degree
2. Fit a logistic regression model with salhigh and yrdeg
   2.1. Calculate the odds ratio and interpret the results
3. Suppose the investigators would rather model the number of years it has been since each professor has achieved his or her degree
   3.1. Create a new variable yrssince95 = 95 - yrdeg
   3.2. Fit a logistic regression model with salhigh and yrssince95
   3.3. Compare the output from model 3.2 with 2.1
4. From model 3.2, create a variable for the predicted probability of having a high salary.
   4.1. Calculate the odds and log odds of having a high salary
   4.2. Plot yrssince95 versus the log odds of a having a high salary
   4.3. Plot yrssince95 versus the odds of having a high salary
   4.4. Plot yrssince95 versus the probability of having a high salary
5. Conduct a likelihood ratio test to test the statistical significance of yrssince95
   5.1. Fit a logistic regression model with just the intercept. Save the deviance from this model
   5.2. Fit logistic regression model 3.2. Save the deviance from this model.
   5.3. Subtract the deviance obtained from 5.2 from the deviance obtained from 5.1. Compare this result to the likelihood ratio test statistic given in the output from model 3.2 (and 2.1)