

Homework 4: due March 19

1. Suppose we collect data for a group of students in a statistics class with variables X_1 = hours studied, X_2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = 6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.
 - (a) Estimate the probability that a student who studies for 40h and has an undergrad GPA of 3.5 gets an A in the class.
 - (b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?
2. This question should be answered using the [Weekly](#) data set. This data contains 1,089 weekly S&P stock market returns for 21 years, from the beginning of 1990 to the end of 2010.
 - (a) Produce some numerical and graphical summaries of the [Weekly](#) data. Do there appear to be any patterns?
 - (b) Use the full data set to perform a logistic regression with [Direction](#) as the response and the five lag variables plus [Volume](#) as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?
 - (c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.
 - (d) Now fit the logistic regression model using a training data period from 1990 to 2008, with [Lag2](#) as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).
3. Using the [Boston](#) data set, fit classification models in order to predict whether a given suburb has a crime rate above or below the median. Explore logistic regression models using various sub-sets of the predictors. Describe your findings.