**Homework 2: Problems To Turn In**

**(4) Drink is a Downer (ACM 12.9):** For this problem we use drinking status as the outcome and are interested in sex and depression (either categorical or continuous) as predictors.

1. Fit a simple logistic regression of drink on sex and use it to find the odds of being a regular drinker for women and men and the corresponding odds ratio. Does there appear to be a sex difference?
2. Repeat part (a) but do the calculations separately for people who are depressed and people who are not depressed and compare the odds ratios for the two groups. (Note: for this part you do not need to give the individual odds by sex; just focus on the odds ratios.)
3. What do the results from part (b) suggest about whether there is an interaction between sex and depression? Explain carefully what such an interaction would mean. Then fit an appropriate model and test whether the interaction is significant. (The depbysex interaction variable has been provided for your convenience.) Do the results confirm your theory? If not, what do you think might have happened?
4. Instead of using the case indicator one could instead use the continuous depression rating, CESD. Fit a logistic regression of drink on sex, CESD and their interaction. (the cesdbysex interaction has been provided for your convenience.) Give as precise an interpretation as you can of this interaction. In particular you should carefully explain the meanings of the regression coefficients b1, b2 and b3 and their corresponding odds ratios and create a rough sketch of the log odds as a function of sex and depression score. Based on your model does it seem as if there is a significant interaction? Are the results more or less significant than those in part (c)? Explain what the results tell you in real-world terms and give a possible explanation for why the test has come out the way it has.

**(5) Homework Is Depressing (ACM):** For this problem we will model depression status (yes or no) as a function of sex, age, income and general health (but not how much statistics homework one has!)

1. Fit the logistic regression model with all 4 predictors, treating health as a continuous variable. Which of these variables appear to be associated with depression status? Are all the relationships in the expected direction? Discuss briefly.
2. Perform an appropriate graphical check to determine whether a linear term adequately describes the relationship between health status and depression. Note: Because the health status variable only has four values you do not need to go through the tedious process of calculating quartiles--you simply need to rerun the model using an appropriate set of indicator variables (which I have provided for you!) In addition to creating the appropriate plot, discuss whether the model using the indicators appears superior to the original model with the linear term. Is it legitimate to formally test this? If the relationship appears non-linear suggest and test an appropriate transformation of the health variable.

**Note: For the rest of the problem, use the model from part (a).**

1. Briefly describe the Hosmer-Lemeshow test for overall goodness of fit and use it to evaluate this model. Does the model appear to fit adequately?
2. Obtain one or more pseudo-R2 values for this data set and explain as carefully as you can what they tell you about the model performance.
3. Compute (i) the true positive rate (sensitivity) and (ii) the true negative rate (specificity)--or if you prefer the false negative rate for this data set for a range of threshold values and plot them to identify a good cutoff for predicting that a person will be depressed. What are your sensitivity, specificity and overall error rate at this cutoff? Do you think the model performs well in this sense?
4. Obtain the ROC curve to go with your calculations from part (e) and the corresponding AUC value. Does the model perform well by this standard?