Homework 2

NAME: Your Name
DUE DATE: March 1st, 11:59 pm

Problem 1 [ISL 3.4] (4 points)

I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

b) Answer (a) using test rather than training RSS.

c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression and also the training RSS for the cubic relationship. Would we expect one to be lower than the other, would expect them to be the same, or is there not enough information to tell? Justify your answer.

d) Answer (c) using test rather than training RSS.

Problem 2 [ISL 4.1] (2 points)

Using a little bit of algebra, prove that

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

is equivalent to

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X}.$$

In other words, the logistic function representation and logit representation for the logistic regression model are equivalent.

Problem 3 [ISL 3.14] (8 points)

This problem focuses on the *collinearity* problem.

a) Perform the following commands in **R**:

```r
set.seed(1)
x1 = runif(100)
x2 = 0.5*x1+rnorm(100)/100
y = 2+2*x1+0.3x2+rnorm(100)
```

The last line corresponds to creating a linear model in which y is a function of x_1 and x_2. Write out the form of the linear model. What are the regression coefficients?

b) What is the correlation between x_1 and x_2? Create a scatterplot displaying the relationship between the variables.

c) Using this data, fit a least squares regression to predict y using x_1 and x_2. Describe the results obtained. What are $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$? How do these relate to the true β_0, β_1, and β_2? Can you reject the null hypothesis $H_0 : \beta_1 = 0$? How about the null hypothesis $H_0 : \beta_2 = 0$?

d) Now fit a least squares regression to predict y using only x_1. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

e) Now fit a least squares regression to predict y using only x_2. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

f) Do the results in (c)-(e) contradict each other? Explain your answer.

g) Now suppose we obtain one additional observation, which was unfortunately mismeasured.

```r
x1 = c(x1, 0.1)
x2 = c(x2, 0.8)
y = c(y, 6)
```

Re-fit the linear models from (c)-(e) using this new data. What effect does this new observation have on each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

Problem 4 Linear Regression (10 points)

Use the data in the College_Extra.csv file on U.S. colleges, found in the Data Sets folder on Canvas and taken from the US College Scorecard data set, to supplement the data in College.csv. In other words, for each college in College.csv, add the following predictors about student's home zip codes:

- **PCT_BA**: Average percentage of people over the age of 25 with Bachelors degrees
- **MEDIAN_HH_INC**: Average median household income
- **POVERTY_RATE**: Average poverty rate
- **MD_EARN_WNE_P10**: Median earnings of federally-aided students who are employed 10 years after enrollment

Then, construct a good fitting model for a federally-aided student's median salary 10 years after enrollment. Consider potential transformations of the outcome and variables as well as interaction terms. Examine the fit of your model using regression diagnostics.

Describe your findings in clearly written text, tables, and figures.

Problem 5 Logistic Regression (14 points)

Wells in Bangladesh used for drinking are often contaminated by natural arsenic which can cause diseases as exposure accumulates in someone's body. If someone's well is contaminated, a neighbor's well may be safe and so if they agree, one can switch to sharing their well. After a research team measured arsenic levels in all the wells in a certain area, residents were urged to switch wells if theirs was labelled unsafe (more than 0.5 hundred micrograms per liter, i.e. 50 micrograms). A few years later the researchers returned to see who had switched wells. The data are in the file wells.txt, found in the Data Sets folder. They include 5 variables for 3020 wells

- **switch** is a binary indicator for whether the household switched wells
- **arsenic** is the level of arsenic in the well in hundreds of micrograms per liter

- **dist** is the distance to the nearest safe well in meters
- **assoc** is whether household members are active in community organizations
- **educ** is the number of years of education of the head of household.

First construct a training data set of the first 2520 wells and a test data set of the last 500.

a. Construct a good logistic regression model predicting the decision to switch wells as a function of the 4 predictors (arsenic, distance, association and education) on the training data. Consider potential transformations of continuous variables and possible interactions.

b. Compute the confusion matrix on the test data using p = 0.5 as a cutoff and discuss what this tells you about the predictive model you have constructed (e.g. sensitivity, specificity, error rate, etc.)

c. Construct an ROC plot and compute the area under the ROC curve. What does this curve tell you about choice of threshold that balances sensitivity with specificity (i.e., how would you balance risk of switching and not switching?)