HOMEWORK 5

Due Monday, May 3rd, start of class

This homework covers Chapter 12. You should be working on your homework throughout these two weeks. If you can't solve some of the problems, please come to office hours. Email is fine only for very short questions.

THEORETICAL PORTION

The theoretical problems should be **neatly** numbered, written out, and solved. Do not turn in messy work.

1. A doctor has collected data on cholesterol and eating habits of 200 subjects, and she wants to investigate the relationship between the two. Specifically, the following variables were measured:

   (a) Cholesterol

   (b) Gender (M or F)

   (c) Amount of alcohol per week ("None", "1-2 drinks", "3-5 drinks", "6+ drinks")

   (d) Amount of fish consumed per week (average, in oz)

   (e) Amount of red mean consumed per week (average, in oz)

   (f) Weight (kg)

   These are the actual measurements for the first six subjects sampled:

   - Subject 1: 102, M, "1-2 drinks", 6 oz, 0 oz, 93 kg
   - Subject 2: 95, M, "None", 12 oz, 9 oz, 70 kg
   - Subject 3: 92, F, "3-5 drinks", 8 oz, 3 oz, 83 kg
   - Subject 4: 113, M, "None", 0 oz, 7 oz, 77 kg
   - Subject 5: 132, M, "6+ drinks", 3 oz, 1 oz, 71 kg
   - Subject 6: 148, F, "1-2 drinks", 4 oz, 4 oz, 84 kg

   (a) Write out the multiple-regression linear model for this data set. Please be careful to identify which covariates are continuous (and only require one parameter for their effect) and which ones are categorical, requiring 2 or more parameters for their effect.

   (b) Interpret the parameters of your model.

   (c) Write out the design matrix X for the first 6 subjects sampled.

   (d) Write out the regression model in matrix format for the first six subjects sampled.

2. The following regression model is given:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i1} * x_{i2} + \epsilon_i.$$

   In this model, the response variable, Y, is the amount of time (in minutes) it takes an individual to perform a task. The predictor variable x_1 is the age in years of the participant (centered around the average age), and x_2 is an indicator variable that is equal to '1' if the subject has performed the task before.

   (a) Why would we add the effect β_3 in this model?

   (b) Why would we add the effect β_4 in this model?

   (c) Interpret each of of the covariate effects parameters $\beta_0, \beta_1, \beta_2, \beta_3$, and β_4.

3. **APPM 5570 only**: In the simple regression equation $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$,

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}.$$

   Show that $\hat{\beta}_1$ is an unbiased estimate for β_1.

4. (25 points) You are examining the relationship between weight and height, with the hypothesis that an individual's height (x) can somewhat accurately predict the individual's weight (y). In your dataset, the height values range from 62 inches to 74 inches (mean = 68 inches, SD = 3.5 inches), and the weight values range from 114 to 200 pounds (mean = 151 pounds, SD = 27 pounds).

   (a) In R, you calculate estimates for β_0 and β_1. The fitted model is:

   $$\hat{y}_i = -222.48 + 5.48x_i,$$

   where x_i is the height of the i^{th} individual. Interpret both parameters in the model in the context of the original problem.

   (b) You calculate a second model, using $x_i' = x_i - 62$, and get the following fitted model:

   $$\hat{y}_i = 117.79 + 5.48x_i'.$$

   Interpret the intercept of this model in the context of the original problem.

   (c) You calculate a third model, using $x_i^* = x_i - 68$, and get the following fitted model:

   $$\hat{y}_i = 150.57 + 5.48x_i^*.$$

   Interpret the intercept of this model in the context of the original problem.

   (d) Which parameterization (i.e. model) do you prefer best and why?

COMPUTATIONAL PORTION

The computational portion of your homework should be neatly done and include all graphs, code, and comments, labeled and in order based on the problem you are addressing. Do *not* put graphs in at the end, stick code in random locations, or do anything else that will make this homework difficult to read and grade. **LABELS ARE YOUR FRIEND, USE THEM.** If you turn in something that is messy or out of order, it will be returned to you with a zero. All computations should be done using R, which can be downloaded for free at https://cran.r-project.org/.

1. The efficiency ratio for a steel specimen immersed in a phosphating tank is the weight of the phosphate coating divided by the metal loss (both in mg/ft^2). The article "Statistical Process Control of a Phosphate Coating Line" (*Wire J. Intl.*, May 1997: 78–81) provides data on tank temperature (x) and efficiency ratio (y), which can be found in the "HW5SteelData.txt" file.

   (a) Plot a histogram of both temperature and efficiency ratio and comment on any interesting features.

   (b) Is the value of the efficiency ratio completely and uniquely determined by tank temperature? Justify.

   (c) Construct a scatter plot of the data Does it appear that efficiency ratio could be well predicted by the value of temperature?

   (d) i. Determine the equation of the estimated regressions line, plot the regression line and plot the data.
   ii. Plot a histogram of the standardized residuals, does the error appear to be normally distributed?
   iii. Create a plot of the fitted values *vs.* the standardized residuals, does the homoscedastic assumption seem reasonable?

   (e) Calculate a point estimate for true average efficiency ratio when tank temperature is 182°.

   (f) Calculate the values of the residuals from the least squares line for the four observation for which the temperature is 182°. Why do they not all have the same sign?

   (g) What proportion of the observed variation in efficiency ratio can be attributed to the simple linear regression relationship between the two variables?

2. Astringency is the quality in a wine that makes the wine drinker's mouth feel slightly rough, dry and puckery. The paper "Analysis of Tannins in Red Wine Using Multiple Methods: Correlations with Perceived Astringency" (*Amer. J. of Enol. and Vitic.*, 2006: 481–485) reported on an investigation to assess the relationship between perceived astringency and tannin concentration using various analytic methods. Data is provided by the authors on x = tannin concentration by protein precipitation and y = perceived astringency as determined by a panel of tasters, and can be found in "SW5WineData.txt" file.

(a) Fit the simple linear regression model to this data (Plot a scatter plot and a plot of the regression line). Then determine the proportion of observed variation in the astringency that can be attributed to the model relationship between astringency and tannin concentration.

(b) i. Construct a scatter plot. Does the simple linear regression model appear to be reasonable in this situation?
ii. Plot a histogram of the standardized residuals, does the error appear to be normally distributed?
iii. Create a plot of the fitted values *vs.* the standardized residuals, does the homoscedastic assumption seem reasonable?

(c) Calculate and interpret a confidence interval for the slope of the true regression line.

(d) Estimate true average astringency when tannin concentration is 0.6 and do so in a way that conveys information about reliability and precision.

(e) Predict astringency for a single wine sample whose tannin concentration is 0.6.

3. Plasma etching is essential to the fine-line pattern transfer in current semiconductor processes. The article "Ion Beam-Assisted Etching of Aluminum with Chlorine" (*J. of the Electrochem. Soc.*, 1985: 2010-2012) gives the accompanying data (read from a graph) on chlorine flow, (x, in units of SCCM) through a nozzle used in the etching mechanism and etch rate (y, in 100 A/min). Fit a linear model predicting etch rate with chlorine flow.

x	1.5	1.5	2.0	2.5	2.5	3.0	3.5	3.5	4.0
y	23.0	24.5	25.0	30.0	33.5	40.0	40.5	47.0	49.0

(a) i. Construct a scatter plot. Does the simple linear regression model appear to be reasonable in this situation?
ii. Plot a histogram of the standardized residuals, does the error appear to be normally distributed?
iii. Create a plot of the fitted values *vs.* the standardized residuals, does the homoscedastic assumption seem reasonable?
iv. What proportion of observed variation in etch rate an be explained by the approximate linear relationship between the two variables?

(b) Does the simple linear regression model specify a useful relationship between chlorine flow and etch rate? Plot the regression line and the data.

(c) Estimate the true average change in etch rate associated with a 1-SCCM increase in flow rate using a 95% confidence interval (CI) and interpret the interval.

(d) Calculate an estimate for the average etch rate when the flow = 3.0.

4. **APPM 5570 only**: Write a function titled 'my.lm' that accepts parameters X and Y, and that performs simple linear regression on the two variables. Your output should look similar to that produced by the `lm()` function in R, and should include:

- The estimates for β_0 and β_1.
- The standard errors for $\hat{\beta}_0$ and $\hat{\beta}_1$.
- R^2, the correlation coefficient.
- The SSE, SSR, and SST.

Email me the function by the due date. Example test code that I will enter is:

```
my.lm(X = predictorVar, Y = responseVar)
```