HOMEWORK 5
Due Dec 4, start of class

This homework covers Chapters 12 and 13. 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. In the simple regression model, \( Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \), where \( \epsilon_i \sim N(0, \sigma^2) \).
   (a) Write down the likelihood function for \( \beta_0 \) and \( \beta_1 \) for \( Y_1, \ldots, Y_n \).
   (b) Find the MLEs for \( \beta_0 \) and \( \beta_1 \). To do this, you need to assume two equations with two unknowns.
   (c) Is the least-squares estimator (LSE) different than the MLE for either parameter? (You may use class notes for this - you do not need to re-derive the LSE.)

2. For a sample \( n = 166 \) college students, the following variables were measured:
   (a) Height (cm)
   (b) Mother’s height (cm)
   (c) Father’s height (cm)
   (d) Gender of student (M or F)
   (e) Race (white, black, Asian, Hispanic)

As a researcher, you want to know if parents’ height, gender, and/or race predict height accurately. These are the measurements for the first five students sampled:

- Student 1: 151, 165, 189, Female, White
- Student 2: 155, 169, 183, Female, Hispanic
- Student 3: 175, 157, 191, Male, Hispanic
- Student 4: 176, 181, 177, Male, Asian
- Student 5: 173, 148, 162, Male, Black

(a) Write out the multiple-regression linear model for this data set. (HINT: The race variable is categorical, so there are more than 5 \( \beta \) parameters in this model.) Set Hispanic males as the baseline group. Interpret each parameter in your model.
(b) Write out the design matrix \( X \) for the first 5 students sampled.
(c) Write out the regression model in matrix format for the first five students sampled.

3. Answer the following questions regarding covariance (section 5.2), provide brief justification for your answer:
   (a) Can covariance between two random variables be less than \(-1\)?
   (b) If covariance between two random variables is negative, does their correlation have to be negative?
   (c) If \( \text{Cov}(X, Y) = 0.3 \), what is \( \text{Cov}(100X, Y) \)?
   (d) If \( \text{Corr}(X, Y) = 0.1 \), what is \( \text{Corr}(100X, Y) \)
   (e) What is \( \text{Cov}(X, X) \)?
   (f) What is \( \text{Corr}(X, X) \)?
   (g) What is \( \text{Cov}(100X, 10X) \)?
   (h) What is \( \text{Corr}(100X, 10X) \)?
4. A rock specimen is randomly selected and weighted two different times. Let \( w \) denote the true weight (a number) of the rock, and let \( X_1 \) and \( X_2 \) be the two measured weights. Then, \( X_1 = w + E_1 \) and \( X_2 = w + E_2 \), where \( E_1 \) and \( E_2 \) are the two measurement errors. Suppose that \( E_1 \) and \( E_2 \) are independent and distributed normally with mean 0 and variance 0.1 (i.e., \( E_1, E_2 \sim N(0, 0.1) \)).

(a) What is the mean of \( X_1 \)? What is the mean of \( X_2 \)?
(b) What is \( V(X_1) \)? What is \( V(X_2) \)?
(c) What is \( \text{Corr}(X_1, X_2) \)?

5. **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 \( \beta_1 \).

**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\(^\circ\).
   (f) Calculate the values of the residuals from the least squares line for the four observation for which the temperature is 182\(^\circ\). 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 \( x = \text{tannin concentration by protein precipitation} \) and \( y = \text{perceived astringency as determined by a panel of tasters} \), 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 and do so in a way that conveys information about reliability and precision.

(f) Does it appear that true average astringency for a tannin concentration of .7 is something other than 0? State and test appropriate hypothesis.

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.

<table>
<thead>
<tr>
<th>( x )</th>
<th>1.5</th>
<th>1.5</th>
<th>2.0</th>
<th>2.5</th>
<th>2.5</th>
<th>3.0</th>
<th>3.5</th>
<th>3.5</th>
<th>4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>23.0</td>
<td>24.5</td>
<td>25.0</td>
<td>30.0</td>
<td>33.5</td>
<td>40.0</td>
<td>40.5</td>
<td>47.0</td>
<td>49.0</td>
</tr>
</tbody>
</table>

(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 a 95% CI for \( \mu_{Y|X=3.0} \), the true average etch rate when flow = 3.0. Has this average been precisely estimated?

(e) Calculate a 95% prediction interval (PI) for a single future observation on etch rate to be made when flow = 3.0. Is the predication likely to be accurate?

(f) Would the 95% CI and PI when flow = 2.5 be wider or narrower than the corresponding intervals of parts (d) and (e)? Answer without actually computing the intervals.

(g) Would you recommend calculating a 95% PI for a flow of 6.0? Explain.

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 \texttt{lm()} function in R, and should include:

- The estimates for \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \).
- 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:

\[
\text{my.lm}(X = \text{predictorVar}, Y = \text{responseVar})
\]