Homework 3

Due Monday July 23, 2012 *(before class starts!)*

1. Please download the "PennSexSurvey.txt" dataset and import it into R. Read the description about the Penn Sex Survey that was done a couple years ago on the course website.

   Let $X_i \overset{\text{iid}}{\sim} N(\mu_m, \sigma_m^2)$, $i = 1, ..., n_m$ and $Y_i \overset{\text{iid}}{\sim} N(\mu_f, \sigma_f^2)$, $i = 1, ..., n_f$ where X_i records the male's BMI and Y_i records the female's BMI. We'll assume that X_i's are independent to Y_i's

   (a) Suppose $\sigma_m^2 = \sigma_f^2 = \sigma^2$ where σ^2 is unknown.

      i. Show that the statistic

      $$\sigma_p^2 = \frac{(n_m - 1)\hat{\sigma}_m^2 + (n_f - 1)\hat{\sigma}_f^2}{n_m + n_f - 2} \tag{1}$$

      is an unbiased estimate for σ^2. Find the distribution of

      $$(n_m + n_f - 2)\frac{\sigma_p^2}{\sigma^2} \tag{2}$$

      Hint: Use the following fact. If $A \sim \chi_a^2$ and $B \sim \chi_b^2$ and A and B are independent from each other, $A + B \sim \chi_{a+b}^2$. Also, we're dealing with the case when μ_m and μ_f are unknown.

      ii. Consider the following statistic

      $$\frac{\bar{X} - \bar{Y} - (\mu_m - \mu_f)}{\sigma_p \sqrt{\frac{1}{n_m} + \frac{1}{n_f}}} \tag{3}$$

      What is the distribution of this statistic? Specify the distribution and the parameter(s) related to that distribution.
      Hint: You must provide mathematical justification. This question is very similar to question 4(e) from Homework 2

   (b) Suppose we want to test the assumption that $\sigma_m^2 = \sigma_f^2$. Conduct a hypothesis test, specify the test statistic, and explicitly derive the sampling distribution from the test statistic. Do your reject the null at $\alpha = 0.05$?

2. These are some conceptual questions about simple linear regression. *No math required, seriously* except for the last question.

   (a) (TRUE/FALSE) For any data (X_i, Y_i), the simple regression line goes through $\bar{X}$ and $\bar{Y}$. Briefly explain your answer.

(b) (Bonus question) Depending on your TRUE/FALSE answer above, what does this say about "regressing to the mean"?

(c) What is the difference between CIs and PIs in linear regression. Explain briefly.

(d) If the Goodness of Fit test decides that there is no linear relationship between X and Y, does that imply that there is no association between X and Y? Why or why not?

(e) If we reject the null $H_0 : \beta_1 = 0$ in the hypothesis test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$, does this imply that there is an association between X and Y? Does this imply that there is a linear association between X and Y? Briefly explain your answers.

(f) Consider the two fitted regression lines

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \epsilon_i \tag{4}$$

$$X_i = \hat{\beta'}_0 + \hat{\beta'}_1 Y_i + \epsilon_i \tag{5}$$

If you are given Y_i and are asked to find X_i in a simple linear regression, are the X_is you find from both equations of the regression, (4) and (5), identical? Why or why not?

(g) Fill in the following ANOVA table where $n = 128$ and we're fitting a simple linear regression

SS	MSE	DF
SSE =	MSE = 30	DFE =
SSR = 10000	MSR =	DFR =
SST =	MST =	DFT =

Table 1: ANOVA table for $n = 128$ and simple linear regression as our model for the linear relationship between X and Y.

3. *Fraud Detection* In this question, we'll explore one technique used in criminology to detect fraud in numerical-based data. In particular, we'll take a look at how the average number of the leading digit '1' differs between fraudulent and real data.

Download the cancer gene expression data set from the course website and import it into R. Each row in the data corresonds to one gene while each columun corresponds to a sample (think of it as an individual, for simplicity). We'll consider each row to be a measurement while each columun represents an individual. *Note that the last column contains the labels for whether it is fake (i.e. 1) or not (0).*

Using the following snippets of R-code to extract the frequency of the digit '1' in the leading digit of each measurement. Remember to take out the last column when you run the above function!

```
# R Code
data = read.csv(''filename.csv'')
data.clean = data[,-ncol(data)] #To take the out the last column.

# AND #
extractLeadingDigit = function(measurement,digit=1) {
```

```
  # Function to extract the leading digit
  firstDigit = substring(abs(measurement),1,1)  # Extracts the first digit
  return(sum(firstDigit == digit)) #gives freq. of '1' in the leading digit
}
```

(a) Are there any differences between the frequencies of the digit "1" between the fake data and the real data? Conduct a hypothesis test, choose the appropriate test statistic, and a reasonable sampling distribution. Do you reject the null at the 0.05 significance level?

(b) Formulate a simple linear regression that explains the relationship between the number of '1's in the leading digit and whether the data is fradulent or not. To be more specific, let

$$X_i = \begin{cases} 1 & \text{if the data is fake} \\ 0 & \text{if the data is real} \end{cases}$$

and Y_i be the frequency of "1"s in the leading digit. What is (numerically) $\hat{\beta}_0$ and $\hat{\beta}_1$? What are the interpretation of these estimates?

(c) What is the test for linear relationship using the slope β_1 in the context of this setup? Verbally explain what the hypothesis test

$$H_0 : \beta_1 = 0 \text{ vs } H_a : \beta_1 \neq 0$$

is testing in the $X_i = 1$ or 0 set up.

(d) What is the p-value of the above test? How does this p-value compare to the one obtained in (a)

(e) Predict the mean number of the leading digit "1" in the new data for fradulent data and non-fradulent data. What are the 95% confidence intervals? What are the 95% prediction intervals?

(f) Mathematiclaly prove that $\hat{\beta}_0 = \bar{X}_{fraud}$ and $\hat{\beta}_1 = \bar{X}_{truth} - \bar{X}_{fraud}$ where the subscript denotes the sample mean of the group specified in the subscript.

(g) Show that the MSE for this regression is the same as the estimate for the pooled variance, $\hat{\sigma}_p^2$, in the case of two-sample tests.

(h) Show that the p-value of the test in (c) is identical to the p-value you obatin by doing a two-sample test with equal variance assumption. What does this say about the two-sample test and simple linear regression?

(i) Show that the confidence intervals obtained in (d) are identical to those confidence intervals testing the difference in means between two groups, under an equivariance assumption.

4. *Baseball* This question is semi-inspired by the awesome movie "Moneyball' with Brad Pitt in 2011. See the dramatic ESPN Youtube Video: `http://www.youtube.com/watch?v=UuXwYZ3AQUO` and the small scene from the movie `http://www.youtube.com/watch?v=GdCOWspWfWY`

   Download the baseball data set from the course website and import it into R. Each row in the data corresonds to an MLB player in a particular year and each column corresponds to his performancce based on several metrics. Here, we'll only look at the metric yearID(the baseball season), R (runs), RBI (run batted in) and HR (number of home runs). For more information on what these metrics are measuring, see the Wikipedia article on RBI, Home runs, and Sabermetrics.

(a) What is the word that the video link above says between 0:43 and 0:45 in the ESPN video?

(b) It is believed that a higher RBI generally means higher HR. Test this hypothesis using simple linear regression by setting up H_0 and H_a, stating the test statistic, and computing the p-value. Make sure you check the assumptions of the regression (e.g. homoscedasticity, linearity, and normality) with relevant plots. Do you rejec the null at $\alpha = 0.05$?

(c) Provide 95% confidence intervals for the slope.

(d) Provide the estimate for a baseball player with RBI = 60. Provide 99% CIs and PIs for this estimate.

(e) Moneyball claims that good baseball players that make decent number of runs, R, are overlooked because of their age (see 0:16-0:25 of the Moneyball clip). Here, we're going to test this hypothesis by using the number of years each player has been in season as our indicator for the player's age.

Conduct a hypothesis test to prove the Moneyball's statician that most recruiters think there is a relationship between age and the number of runs. Make sure you check the assumptions of the regression (e.g. homoscedasticity, linearity, and normality) with relevant plots. Do you reject the null at $\alpha = 0.05$?
Hint: You have to come up with an R code to reformat the yearID data into the number of years a player has been in season. For example, for the first playe, "aaronha01", he has been in season from 1970 to 1976. Your R code should reformat 1970 to 1, 1971 to 2, 1972 to 3, and so on and so forth. This is probably the hardest thing about this question. But, I do hope it will give you some practice into how to deal with real data (i.e. how to clean up real data

5. All the questions below are bonus questions. Solving them *honestly* will get you a decent amount of bonus points.

(a) Show that $\bar{Y}$ is indepednent of $\hat{\beta}_1$.

(b) Derive the sampling distribution for $\hat{\beta}_1$. In particular, mathematically show that

$$\frac{\hat{\beta}_1 - \beta_0}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{n-2}$$

Hint: To show this, you must show that r_i is independent of $\hat{\beta}_1$. I would suggest computing the covariance between r_i and $\hat{\beta}_1$ to justify your mathematical argument.