HW 10: ANOVA for SLR and R^2

Instructions: Work must be shown to receive full credit. You may work with others on the homework, but you must write and turn in your own copy. **This does not mean that you can simply copy someone else's work!!** Also, make sure your homework is neat, stapled, and all answers are written in complete sentences!! Come and see me if you have any questions.

On problems that require the use of R, PLEASE give me the RELEVANT R code and output to for each problem so I can assess partial credit. I may take off for including unnecessary R output. If one problem refers back to output from another problem, make sure to cite that output in your answer. Incorrect one-sentence answers will get little or no credit.

NOTE: If a problem asks you to perform a hypothesis test, make sure to give the hypotheses, test statistic, p-value, and a conclusion in the terms of the problem. Also, if the problem asks you to perform a confidence interval, make sure to interpret the confidence interval.

"By Hand" Problems: For hypothesis tests, you may use R to find the p-value. For confidence intervals, you may use R to find the multiplier.

1. The *olympic.csv* file on Moodle contains 120 years worth of Olympic data with 30,181 observations. We would like to use this data to determine if there is a linear relationship between the height on Olympic athlete and their weight. That is, can height be used to predict weight in a simple linear regression model? Some relevant summaries are included below.

	Variable	Mean	Standard Deviation
$\hat{y}_i = -121.895 + 1.101x_i$	Height(cm)	177.6424	10.92419
$\sum e_i^2 = 2426280$	Weight(kg)	73.75355	15.00499

   (a) Construct a complete ANOVA table for SLR based on the information provided.

   (b) What are the hypotheses represented by the test in the table? That is, state H_O and H_A in both symbols and words.

   (c) Is more of the variability in weight explained by the line or from random error? How do you know this? Give values from the table to justify your answer.

   (d) The p-value for the test is ≈ 0. Interpret this value and state the conclusion of the test in context of the problem.

   (e) Calculate and interpret the value of R^2 using information from elsewhere in the problem.

2. Recall the roller coaster data from earlier in the semester. We were interested in predicting the maximum speed using the maximum drop height. We had a sample of 20 roller coasters. Some relevant summaries follow.

	Variable	Mean	Variance
$\hat{y}_i = 40.66847 + 0.16957x_i$	Drop	140.45	2074.26
$\sum e_i^2 = 374.4$	Speed	64.485	79.35082

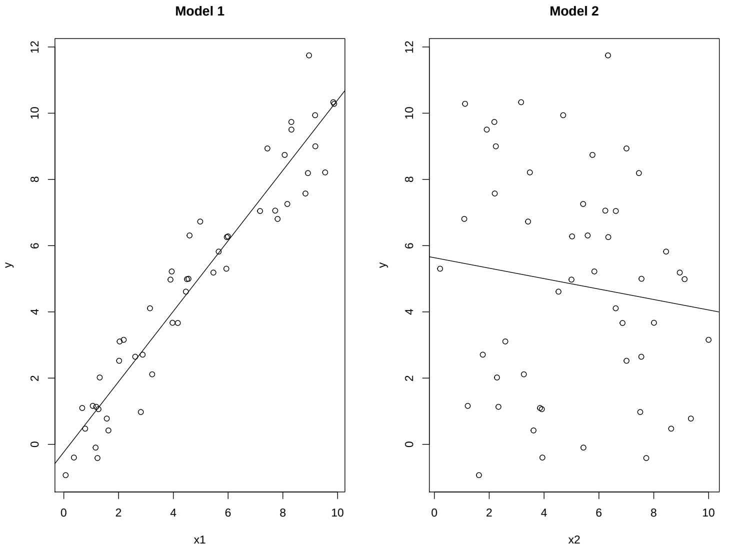
(a) Construct a complete ANOVA table for SLR based on the information provided.

(b) What are the hypotheses represented by the test in the table? That is, state H_O and H_A in both symbols and words.

(c) The p-value for the test is 7.56×10^{-7}. Interpret this value and state the conclusion of the test in context of the problem.

(d) Calculate and interpret the value of R^2 using information from elsewhere in the problem.

3. (From Sheather 2009, pg. 41) Two alternative straight line regression models have been proposed for Y. The first model has Y as a linear function of x_1 while the second model has Y as a linear function of x_2. The plots show the data and least squares regression lines. Recall, SSE stands for residual sum of squares while SSR stands for the regression sum of squares. This is a multiple choice question. Which of the following statements is true? Give a detailed reason to support your choice.



(a) SSE for model 1 is greater than SSE for model 2, while SSR for model 1 is greater than SSR for model 2.

(b) SSE for model 1 is less than SSE for model 2, while SSR for model 1 is less than SSR for model 2.

(c) SSE for model 1 is greater than SSE for model 2, while SSR for model 1 is less than SSR for model 2.

(d) SSE for model 1 is less than SSE for model 2, while SSR for model 1 is greater than SSR for model 2.

"R" Problems:

4. During WWII baseball was still played, though many players served in the military. Suppose we are interested in predicting batting averages towards the the end of the war (`BA44`) using the prior season's batting averages (`BA43`). The data are provided on Moodle in the `ww2baseball.xlxs` file.

   (a) Using `R`, load the data and obtain the regression line for predicting 1944 batting averages using 1943 batting averages.

   (b) Obtain the ANOVA table in `R`.

   (c) What term (give the value) measures the variability explained by the line?

   (d) What term (give the value) measures the variability not explained by the line?

   (e) What term provides an estimate for σ^2? (There are a couple of ways to find this value.)

   (f) Using prior output, provide two different test statistics for the test of $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$.

   (g) State and interpret the value of R^2. What does this tell us about the fit of the line?