Name: ____________________ SID: ____________________

Collaborators: ____

Week 8 Problem Set

Correlation and Regression
PHW142

You must put your name and SID at the top of the page.

Please include:

- explanations of your reasoning (even if we forget to ask over and over)
- the formulas and major steps when the question asks you to do this
- relevant R code and results
- interpretations in the context of the problem scenario

This problem set is worth 16 points.

Context:

We're going to use a version of the body fat in men dataset Baldi and Moore discuss in several of the exercises in our textbook. (See exercises for chapters 1 and 2.)

We will explore the Quetelet body mass index BMI as a potential predictor of percent body fat. Body mass index is a ratio, defined as mass in kg divided by the square of height in meters. This measure is now widely used in public discourse and conversations about overweight and obesity.

BMI is not a perfect measure of percent body fat; there are football linebackers with high BMI but low percent body fat, and sedentary individuals who appear thin but carry a risky amount of visceral fat.

Until recently, accurately determining percent body fat was complex.

Could BMI serve as a screening tool or even as a predictor?

Precise prediction is the greatest challenge for a regression model. Compared to providing evidence of association and to getting precise estimates of a population coefficient, prediction for individuals is a very tall order.

I have removed two observations from my version of the dataset. There is an individual whose height is clearly a mistake, and an individual who weighs over 350 pounds; the next largest weight is about 100 pounds less.

In the Excel dataset `bodyfat_250_regr_ps.xlsx`, there is a variable for the outcome called `siri`, which is one of the two complex methods of determining percent body fat. The explanatory variable is called `bmi`.

The R Markdown files for the examples in the Reader are the models for the R functions you will need to complete this problem set.

Copy and paste your graphs and results into your report in the places indicated.

Take a look at the solution template before you start working in R.

Paste your scatterplot of the siri and bmi values here:

1. Use the scatterplot to discuss the overall pattern of association and any unusual points. **2 points**

Averages and standard deviations for the outcome siri and explanatory variable bmi for all 250 observations:

Pearson correlation for all 250 observations:

Fitted regression from the `lm` function for all 250 observations.

2. Write out the equation for the fitted regression line.

   State the values of the slope and intercept, including their units. **2 points**

3. **Interpreting the slope, part 2**

   (4 points) Using the standard error for the slope in your `lm` results, and the critical value t^* from the t distribution, demonstrate the calculation of the confidence interval for the slope.

   3.1 What are the degrees of freedom for the t^* value? **$^1\!/_2$ point**

3.2 Calculate the margin of error. **1 point**

If you used OOMPH Stat, paste a screenshot here:

3.3 What are the units for the slope? **½ point**

3.4 What are the lower and upper limits of the confidence interval? **1 point**

3.5 Write a sentence summarizing the confidence interval for the slope, including the units. **1 point**

4. **Interpreting the slope, part 2**

   (4 points) Explain all the details in the t test for the slope with a two-sided alternative hypothesis:

   4.1 State the null and alternative hypotheses. **1 point**

   4.2 Using the slope estimate and the standard error estimate from the output, show the calculation of the test statistic. **1 point**

   4.3 Carefully interpret the P value that the `lm` function gives for this test. **1 point**

4.4 Write a sentence explaining your conclusion about the association between the siri bodyfat value and bmi. **1 point**

5. **(2 points)** Interpret the value of R^2 from this regression and verify that it is the square of the Pearson correlation coefficient.

6. **(2 points)** Interpret the residual standard error of the regression, including the units. Explain why it also tells us that using BMI to predict an individual's percent body fat will not lead to predictions that are useful in practice.

   (Statistics-based answer; not clinical one)