Statistics 500 - Homework # 9, Fall 2020
Due by noon Friday, 10/30/2020

Reading Assignment: Statistical Sleuth, Chapters 10 and 11 (Multiple Regression). PACQ, Review Appendices A and B and read Chapter 1

1. The table below shows the height in centimeters and the weight in kilograms at eighteen years of age for ten of the seventy girls in the Berkeley Guidance Study data set.

Height (cm) X	Weight (kg) Y
169.6	71.2
166.8	58.2
157.1	56.0
181.1	64.5
158.4	53.0
165.6	52.4
166.7	56.8
156.5	49.2
168.1	55.6
165.3	77.8

The Berkeley Guidance Study enrolled children born in Berkeley, California, between January 1928 and June 1929, and then measured each child periodically until age 18. The data for all of the girls in the study who were measured at age 18 are posted in the file **BGSgirls.dat**. There is on line for each girl on this data file with three numbers on each line corresponding to a subject identification number, weight, and height, in that order from left to right.

a. Compute least square estimates of the intercept (β_0) and slope (β_1) of a simple linear regression model for predicting weight (Y) from height (X). Report the parameter estimates and their standard errors.

b. Plot weight versus height and insert the estimated regression line on the plot. What does this plot suggest?

c. Construct a plot of the studentized residuals versus $\hat{Y}_i$, where $\hat{Y}_i = b_0 + b_1 X_i$. What does this plot indicate?

d. The diagnostic plots indicate that there is one 18 year-old girl who is extremely heavy given her height. This observation may involve a value for either height or weight that was not properly recorded, or it may just correspond to an unusually heavy girl. You can delete this observation by replacing the value of the weight with a period. Because with is the only girl with weight exceeding 90 kg, you can delete this case in a data step by inserting the code:

```
if(weight > 90) then weight=. ;
```

Or you can use only the subset of data by "where weight le 90;". Re-fit the simple linear regression model. Do the diagnostic plots now appear to show that the data conform to the assumptions of the proposed regression model? If not, what problems remain?

e. Plot the estimated regression lines with the extreme observation included and the extreme observation removed on the same plot. Did deleting the observation in part (d) have a large effect on any of the parameter estimates? Where?

f. Using the data with the outlier deleted, evaluate least squares estimates for the parameters in the quadratic polynomial model

$$Y_i = \alpha_0 + \alpha_1 X_i + \alpha_2 X_i^2 + \varepsilon_i \ .$$

Without any adjustment for simultaneously performing three t-tests, report the values of the least squares estimates of the regression parameters, their standard errors, and p-values for t-tests for testing the null hypotheses that each regression parameters is zero:

a_0 = _____ S_{a_0} = _____ t = _____ p-value = _____

a_1 = _____ S_{a_1} = _____ t = _____ p-value = _____

a_2 = _____ S_{a_2} = _____ t = _____ p-value = _____

Do any of these tests support putting the quadratic term into the model?

2. Suppose that six observations of the yield (Y) of a chemical process were taken at each of four temperature levels (X) for running the process, but you are only given information on the sample means and standard deviations for the observed yields at each temperature. The summary data are

Temperature (^oC)	150	200	250	300
sample mean	66	81	89	92
sample variance	1.15	1.00	1.35	0.90
sample size	6	6	6	6

a. Use this information to compute the least squares estimates of β_0 and β_1 for the simple linear regression model $Y_{ij} = \beta_0 + \beta_1 X_i + \varepsilon_{ij}$. Report values for the estimated coefficient and their standard errors.

b. Complete the following ANOVA table:

Source of variation	df	Sum of Squares	Mean Square
Regression on X			
Residuals			
Lack-of-fit			
Pure error			
Corrected total			

c. Compute a p-value for the lack-of-fit test and state your conclusion.

d. If the test in part (c) rejects the proposed straight line model shown in part (a), find an alternative model that is consistent with the data.