Statistics 500 - Homework # 10, Fall 2020
Due: noon Saturday, 11/7/2020

Reading Assignment: Statistical Sleuth, Chapters 10-12 (Multiple Regression).

1. Data **(CO2_2014Vehicles.csv)** on new vehicles for the 2014 model year are available from the Environmental Protection Agency. A random sample of 200 vehicles was selected. Using these data we wish to predict the CO2 emissions of the vehicles in city driving (cityCO2). The explanatory variables are listed below. (This is one of datasets we analyze for Lab 11).

- Engine – displacement of the engine in liters (Min = 1, Max = 6.8)
- Cylinder – number of cylinders (Min = 3, Max = 12)
- City MPG – Fuel economy in city driving (MPG) (Min = 11, Max = 40)
- Gears – number of gears (Min = 1, Max = 9)
- Intake – Number of intake valves per cylinder (Coded as 1 if 2 and 0 otherwise)
- Exhaust – Number of exhaust valves per cylinder (Coded as 1 if 2 and 0 otherwise)

Use SAS to run the multiple linear regression model with Engine, Cylinder, City MPG, Gears, Intake, and the interaction between Cylinder and Intake as explanatory variables. Use the output to answer the following questions.

a. Give the estimated regression line for predicting the cityCO2 values from the four explanatory variables (Engine, Cylinder, City MPG, Gears) for vehicles with two intake valves per cylinder and for vehicles that do not have two intake valves per cylinder. What is the difference in these two equations?
b. Conduct a t-test for the significance of the interaction between Cylinder and Intake in the model that includes Engine, Cylinder, City MPG and Gears and Intake. Report the null and alternative hypotheses, test statistic, p-value, decision and conclusion.
c. Explain why you should not conduct a t-test for the significance of the variables Cylinder and Intake in this model.

2. In this problem we will examine additional variables collected in the Berkeley Guidance Study. The relationship between weight and height of 18 year-old girls was examined on Homework 9. The variables in the bigger data set include height in centimeters at ages 2, 9 and 18 (HT2, HT9, HT18), weights in kilograms (WT2, WT9, WT18), leg circumference in centimeters at ages 9 and 18 (LG9 and LG18), and strength in kilograms at ages 9 and 18 (ST9 and ST18). Two additional measures of body type are also given, somatotype (SOMA) at age 18, on a scale from 1, very thin, to 7, very obese, and body mass index (BMI) at age 18, computed as weight in kilograms at age 18 divided by the square of height at age 18 in meters. The data are posted in the file **BGSgirls2.txt** and *partial* SAS code is **BGSgirls2.sas**. There is one line of data for each of 70 girls with the variables appearing in the following order:

ID: Girl identification number
WT2: Weight(kg) at two years
HT2: Height(cm) at two years
WT9: Weight(kg) at nine years
HT9: Height(cm) at nine years
LG9: Leg circumference(cm) at nine years
ST9: Strength(kg) at nine years
WT18: Weight(kg) at eighteen years
HT18: Height(cm) at eighteen years
LG18: Leg circumference(cm) at eighteen years
ST18: Strength(kg) at eighteen years
BMI: Body Mass Index at eighteen years

SOMA: Somatotype (SOMA), on a scale from 1, very thin, to 7, very obese

We will examine how well the measurements at ages 2 and 9 can predict BMI at age 18.

a. Compute the sample correlations between BMI at age 18 and each of the following explanatory variables HT2, HT9, WT2, WT9, and ST9. Which of these explanatory variables have significant correlations with BMI at age 18?

b. Compute the sample correlations among the five explanatory variables HT2, HT9, WT2, WT9, and ST9. Which of these explanatory variables have significant correlations with other explanatory variables?

c. Find least squares estimates of the parameters in the regression of BMI at age 18 on strength at age 9,

$$BMI_i = \beta_0 + \beta_1(ST9_i) + \varepsilon_i \qquad i = 1, 2, \ldots, 70.$$

Is the slope significantly different from zero? What do the residual plots reveal?

d. Construct a partial residual plot to display the conditional association between BMI at age 18 and strength at age 9 (ST9) given weight at age 9. First compute the residuals from the regression of BMI on WT9, and call them **rbmiwt9**. Then compute the residuals from regressing ST9 on WT9, and call them **rst9wt9.** Then plot **rbmiwt9** against **rst9wt9.** Describe what this plot reveals about the conditional relationship between BMI at age 18 and strength at age 9, after adjusting out the effects of weight at age 9?

e. Complete the regression of **rbmiwt9** on **rst9wt9**. Report the value of the estimated slope and its standard error.

f. Now compute the multiple regression of the body mass index at age 18 on both weight at age 9 and strength at age 9, i.e. fit the model

$$BMI_i = \beta_0 + \beta_1(ST9_i) + \beta_2(WT9_i) + \varepsilon_i \qquad i = 1, 2, \ldots, 70$$

Is the estimate of β_1 for this model, the coefficient for strength at age 9, the same as the estimate of β_1 for the model in part (c)? Did you expect the estimates to be different? Explain.

g. Compare the least squares estimate of the slope for the regression in part (e) to the least squares estimate of the partial regression coefficient for strength at age 9 (ST9) in part (f). Are the two estimates the same? Are the standard errors of the two estimates the same?

h. Fit the multiple regression model

$$BMI_i = \beta_0 + \beta_1(HW2_i) + \beta_2(WT2_i) + \beta_3(HT9_i) + \beta_4(WT9_i) + \beta_5(ST9_i) + \varepsilon_i$$

And report the R^2 value. What proportion of the variation in BMI values at age 18 can be attributed to changes in the conditional means for BMI as the five explanatory variables vary across girls?

i. Report estimates of the six partial regression coefficients for the model in part (h), their standard errors, and the value of the corresponding t-tests and p-values (for two-sided alternatives to the null hypothesis). For each t-test, explicitly state the null hypothesis that is tested and state you conclusion.

3. A dataset was collected from home sales in Ames, Iowa between 2006 and 2010. The variables collected are:

- Year Built: The year the house was built.
- Basement Area (in sq. ft): The amount of area in the house below ground level.
- Living Area (in sq. ft): The living area in the home. Includes the amount below ground level.
- Total Room: The number of rooms in the house.

- Garage Cars: The number of cars that can be placed in the garage.
- Year Sold: The year the home was sold.
- Sale Price: The sale price of the home (our response variable).
- Garage Size: S = Small (Garage Cars = 0 or 1) or L = Large (Garage Cars = 2 or more)
- Age (in yrs.) = Age of house = Year Sold – Year Built

The data from 2,925 sales can be found in the file **AmesHousing.csv**. For all parts requiring a hypothesis test, make sure to state the null and alternative hypotheses, test statistic, p-value, decision, and conclusion in context.

First, we will look at predicting the Sale Price of the house from the two explanatory variables: Living Area and Age.

a. Give a description of the parameters (β's) for Living Area and Age in the multiple linear regression model.
b. What is the value of R^2 and its interpretation for the model including Living Area and Age?
c. Give the ANOVA Table and F-test for the overall significance of the model.
d. Give the t-test of the significance of Living Area in the model, and the t-test of the significance of Age in the model.
e. In addition to Living Area and Age, we now add two additional explanatory variables Basement Area and Total Room into the multiple linear regression model.
    (i) How much is the reduction to the Sums of Squares for Error for adding Basement Area and Total Room to the model with Living Area and Age?
    (ii) Provide the partial F-test for the significance of Basement Area and Total Room in the model with Living Area and Age.