Statistics 500 - Homework # 11, Fall 2020
Due: noon Friday, 11/13/2020

Reading Assignment: PACQ: Chapters 1, 2, 3 and 6. Review Appendices A through D as needed.

1. In this problem we will revisit the Berkeley Guidance Study that you worked on in last homework. The relationship between weight and height of 18 year-old girls was examined on Homework 10. 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

a. Fit the multiple regression model

$$\mathrm{BMI}_i = \beta_0 + \beta_1(\mathrm{HW2}_i) + \beta_2(\mathrm{WT2}_i) + \beta_3(\mathrm{HT9}_i) + \beta_4(\mathrm{WT9}_i) + \beta_5(\mathrm{ST9}_i) + \varepsilon_i$$

and use the following residual plots to assess model assumptions (Do not submit the plots; just examine the plots and briefly describe the insight provided by each plot).

(i) Normal probability plot (q-q plot) of residuals and the related Shapiro-Wilk test.
(ii) Plot of the residuals versus the estimates of the conditional means for BMI.
(iii) Individual plots of the residuals versus each of the five explanatory variables.
(iv) Partial regression plots based on the regression of BMI on each explanatory variable after adjusting out the linear relationships between BMI and the other four explanatory variables and the linear relationships between the particular explanatory variable and the other four explanatory variables. You could construct the partial residual plots in the manner described in homework 10, or you can get all five partial regression plots by including the “partial” option in the model statement of the REG procedure as shown below:

```
proc reg data=set1;
 model bmi = ht2 ht9 wt2 wt9 st9 / ss1 ss2 vif partial;
 output out=set1 r=resid p=yhat;
 run;
```

b. Given that an outlier that should be detected from part (a), refit the model and check the four plots listed in (a) and assess whether model assumptions are violated or not.

c. For the 69 observations (without the outlier that was detected from part (a)), use a backward selection procedure to search for a model using $\alpha_{stay}=0.05$. For this question, just consider the five variables mentioned in part (a): `ht2 ht9 wt2 wt9 st9`. For your final model, report the estimated coefficients and their standard errors.

d. For the 69 observations (without the outlier that was detected from part (a)), check all possible models that could be constructed using at most the five variables `ht2 ht9 wt2 wt9 st9` and then give the best one that you recommend. Justify your choice.

e. Check assumptions of equal variance and normality for the one you selected in part (d).

f. Is there concerns about multicollinearity for the explanatory variables of the model you picked in part (d)?

2. A dataset (analyzed in last homework) 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

We continue the analysis of this dataset with more predictors. 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.

a. The data from 999 sales can be found in the file **TrainingSet.csv**. You will use these 999 observations and SAS to determine a final multiple linear regression model for predicting sale price from the explanatory variables: Basement Area, Living Area, Total Room, Garage Size, and Age. Use a significance level of $\alpha=0.05$ to determine significance of the overall model and of individual explanatory variables. Use stepwise method and all possible models to arrive at a final multiple linear regression model. Then, use this model to answer the following questions:
   (i) Explain your choice of the final model.
   (ii) Give the table of parameter estimates for the final model.
   (iii) Give the ANOVA table for the final model.
   (iv) Give the value of R^2 for the final model. Give an interpretation of this value.
   (v) Check model assumptions and also look for outliers, high leverage points, and potential influential points.

b. Then, fit the regression model with your selected variables to the data contained in the file **EvaluationSet.csv**. This file includes the remaining 1,924 sales. For this model, report the following:
   (i) The table of parameter estimates for the final model.
   (ii) The ANOVA table for the final model.
   (iii) The value of R^2 for the final model. Give an interpretation of this value.
   (iv) The value of MS_{Error}. Compare this value to the MS_{Error} from the model fit with the Training data set above.