HW 10

1) The file Gesell.MTP concerns a study of whether intelligence can be predicted based on the age at which a child speaks. For each of 21 participants in the study, the variable Age represents the age (in months) at which they spoke their first word, and the variable Score represents the Gesell Adaptive Score. (The Gesell test is an adult intelligence test).

   A) Without looking at the data, how would you expect Score to be related to Age? (Positively or negatively?)
   B) Make a scatterplot of Score versus Age. Does the plot show the relationship you predicted in A)?
   C) Run the simple regression of Score on Age. Get the leverage and Cook’s D values by clicking on Storage in the regression dialog box, and checking the boxes for leverage (Hi) and Cook’s Distance.
   D) Use the regression output to compute the \( p \)-value for the coefficient of Age in the regression. Does this suggest that Score is related to Age, in the direction that was predicted?
   E) What proportion of the variance in Score is explained by Age, based on the regression output?
   F) Are there any data points with high leverage? According to the rule we gave in the course handout, is the Cook’s D corresponding to these points high enough to cause concern?
   G) Delete the data point with the largest value of Cook’s D, by highlighting that case in the Minitab worksheet, and pressing the Del key. Now, re-run the regression. Describe the effects on the \( p \)-value for the slope, and on R-Squared. Is there now strong evidence of a linear relationship between Score and Age?
   H) Do you feel that it is justifiable to have deleted this point from the data set?

2) The file Magazine.MTP contains data on advertising costs and characteristics of magazines. The response variable is PageCost, which represents the cost of a full-page color ad in the magazine. There are 48 magazines listed. Circ is the circulation of the magazine (in thousands), PercMale is the percentage of the readers who are male, and MedIncome is the median income of the readers. (Obviously, a lot of statistical work went into trying to accurately measure these variables!)

   A) Run a multiple regression of PageCost on Circ, PercMale, and MedIncome. Which variables are positively related to PageCost? Which coefficients are statistically significant? Based on the \( p \)-values for the regression coefficients, which variables seem to be useless for predicting PageCost?
B) The *F*-Statistic and corresponding *p*-value in the Analysis of Variance part of the output provides a test of the null hypothesis that the regression is useless for predicting Y, i.e., that all regression parameters besides the intercept are zero. Based on the *p*-value, does the regression seem to be useful for predicting Y? Does this mean that all variables are useful? (Remember your answer to part A).

C) Get a 95% confidence interval for the mean page cost of a magazine with a circulation of 800 thousand, 60% male readership, and median income of $30,000. To do this, click on Options in the regression dialog box. In the box for “Prediction Intervals for New Observations”, enter 800 60 30000.

D) We really shouldn’t have any faith in the regression output until we do some checks to see if the model seems adequate. Make a plot of the residuals versus fitted values, and use this to argue that, in fact, there are problems with the model.

E) To try to find the problem, make a scatterplot of PageCost versus Circ. Identify the magazines corresponding to the two outlier points. These are the two with the highest circulation, so they may be high leverage points. Instead of deleting these points, we are going to try using transformations, in this case by working with the logarithms of these two variables. To create the transformed variables, use Calc → Calculator → Store Result in Variable: LogCost, Expression: log(PageCost). Then repeat the same procedure for Circulation, that is, Calc → Calculator → Store Result in Variable: LogCirc, Expression: log(Circ). Make a scatterplot of LogCost versus LogCirc. Notice that now the linear relationship is stronger, and the variability of Y is much more nearly constant as a function of X.

F) Run the multiple regression of LogCost on LogCirc, PercMale and MedIncome, including a plot of residuals versus fitted values. Note any changes compared with parts A), B) and D). Also, comment on any changes in the R-Squared value.

3) The file MLB92.MTP contains data on the 1992 season of Major League Baseball. Although many variables are given for the 330 players, we will focus on trying to predict the number of homeruns (HR), based on the number of Doubles and the number of Triples.

A) Run the regression of HR on Doubles and Triples. Are the regression coefficients significantly different from zero? What about the intercept?

B) Is it surprising that the coefficient for Triples is negative? (This question is optional, for baseball fans only!)

C) Test the null hypothesis that the true coefficient for Doubles is 1.0, versus a two-tailed alternative, at level .05.

D) Plot the residuals versus fitted values. What problem is indicated?

E) Try taking logs of all three variables. (See Problem 2E). Repeat parts A) C) and D) for the regression of logHR on logDoubles and logTriples.