**STATISTICS 201, FALL 2013**
**November 20 Homework**
**Due Wed, November 27**
**CLARIFICATION ADDED ON SUNDAY, NOVEMBER 24**

The file Nov20Hmwk.txt contains a subset of the data for the student and parents height data set that we have used in numerous examples. The subset consists of the male students only, with the case removed that had a clearly erroneous mother's height of 80 inches. There are 75 cases in the data set. The variables in the data file are:

ID = the original ID from the full dataset, which you should use to identify cases
Sex = Male for everyone in this data set (and thus you won't need to use it)
momheight, dadheight and Height = heights in inches for mother, father, and student

The variable names are listed at the top of the data file, so you should use the R option that specifies that they are included.

ASSIGNMENT:

1. Use the variables momheight and dadheight to predict Height.
2. Find case diagnostic values for the four diagnostic measures discussed in class. (These include $t_i$, $h_{ii}$, $(DFFIT)_i$, and Cook's distance.) In this step you are simply supposed to find them all. You don't have to print them all out; you just need to verify that you found them. For instance, you could print out the first few or show the R commands used to find them.
3. For each of the diagnostic measures, identify cases that need to be investigated (if any). Use the variable "ID" to identify them so we know which cases you have identified. *Make sure you use the variable "ID" and not the Row number. They are not the same because females were omitted, so ID numbers are not consecutive.* In this step you should give the ID numbers for all of the cases identified, for each of the 4 measures. If you prefer, you can combine the information in Steps 3 and 4.
4. For each case identified in #3, provide the data values and the diagnostic measure(s) that caused the case to be flagged. In this step you should create a table showing the data values for each of the cases identified in step 3. Step 5 might be more clear if you also include the predicted values (Y-hat) in the table.
5. Choose 4 of the cases flagged and provide an explanation for why that case was flagged as unusual. In some cases, this might include comparing Y to its predicted value.
6. Discuss whether any of the identified cases should be removed from the analysis. For this step you should remember the discussion of outliers, what causes them, and what to do (or not do) about them, from the lecture on October 16.