STATISTICS 201, FALL 2013
November 20 Homework
Due Wed, November 27

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}, $(\text{DFFIT})_i$, and Cook's distance.)
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.*
4. For each case identified in #3, provide the data values and the diagnostic measure(s) that caused the case to be flagged.
5. Choose 4 of the cases flagged and provide an explanation for why that case was flagged as unusual.
6. Discuss whether any of the identified cases should be removed from the analysis.