MATH 160 Introduction to Applied Statistics Spring 2010

Notes on Assignment #1

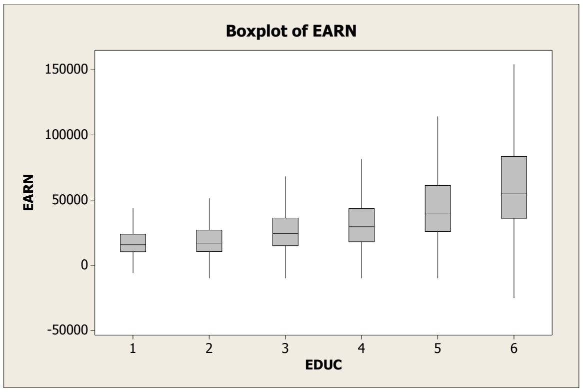
1. The description in the text states that data was collected on 71076 subjects. The data file itself has only 55899 records. An explanation for this discrepancy is not clear. One possibility is that 71076 refers all ages while 55899 refers to those ages 25 to 64 that were selected by the text authors (for reasons that are not explained).

2. EDUC is a categorical variable rather than a quantitative variable. The values (1,2,3,4,5,6) are numbers but adding these numbers is not meaningful. For example, adding 2 for a person with some high school and a 5 for a person with bachelor's degree gives 7 which has no meaning on this scale. So, box plots, histograms and measures such as mean & median are not appropriate here. Counts and proportions for each category would be appropriate. These could be displayed in a bar chart (counts or proportions) or a pie chart (proportions only).

   Note that bar charts and histograms differ in several important ways. In a bar chart, the width of the bars has no meaning and the size of the gap between the bars has not meaning. In a histogram, the width of the bars tells us the size of the bin range and there are no gaps between bars since each possible value falls into exactly one bin. (Note that some bins might be empty giving the appearance of gaps.)

3. The category values (1,2,3,4,5,6) for the EDUC variable have no obvious inherent meaning. A description of each level is required here. The coding is described in the text's description of the data set in the Data Appendix starting on page D-8.

4. It is reasonable to conjecture a connection between education level and earnings. To explore this conjecture, you could examine side-by-side boxplots of earning distribution for each education level (see below). Looking for a possible connection goes beyond what you are asked for in this assignment. Exploring this potential connection is a reasonable way to address the interpretation you were asked to provide.



Note: Outliers are not included in these boxplots.

5. A scatterplot is not appropriate here since EDUC is a categorical variable rather than a quantitative variable (see Note 2). Side-by-side box plots of the EARN distribution for individual EDUC levels would appropriate for this type of comparison.