Exploration 8B: How Outliers Influence Regression

For this exploration we are going to investigate the relationship between the appraised values of homes and the actual sale prices of the homes. (All of the homes in the data file “C08 Homes.xls” were sold during a three-month period of time in the Rochester, NY region in 2000.)

First, construct a linear regression model for predicting the price of the home from the appraised value. Be sure that you have the routine construct the diagnostic graphs (Fitted vs. Actual and Residuals vs. Fitted). Also, make sure that you have the routine calculate the fitted values and residuals on the data worksheet. We will need all of these graphs and figures to explore the data.

1. What is the equation of your regression model? Is this model any good?
2. What does this model mean?
3. Now, on the residuals vs. fitted graph, draw horizontal lines to mark one standard error of estimate above and one below the horizontal axis (residuals = 0 along the axis). Draw similar lines to mark two, three and four standard errors. How many of the observations fall within 1, 2, 3, and 4 standard errors of the predicted values? What proportion of the observations fall in these ranges? (There are 275 observations total.)
4. Are there any outliers in the data - observations more than 4 standard errors from zero? How many are there? How do you think these outliers influence the quality of the regression model? How would the regression model change if you removed these outliers and re-ran the regression routine?
5. Next, we are going to identify the outliers and remove them from the data. To do this, we need to look at the actual data and sort it from smallest to largest residuals. Before we can do this, however, we need to delete the empty column between the data and the fitted values (this should appear in column N). To delete the column, place your cursor on the column header (N), right click, and select “Delete”. Now, sort the data from smallest to largest residuals (see the “How To Guide” for section 5B.) Locate any observations with residuals more than 4 standard errors from zero. Delete these observations from the data by deleting the rows the data are in (right click on the row, select “Delete”). Now, create a new regression model to predict the price of a home from its appraised value.
6. How do these two models compare, both the equations and quality of the fit?