**Stat 301 – Lab 11**

**Goals:** In this lab, we will see how to:

Calculate and plot standardized residuals and Cook’s Distance

Identify points that interest us on a plot

Calculate the PRESS statistic

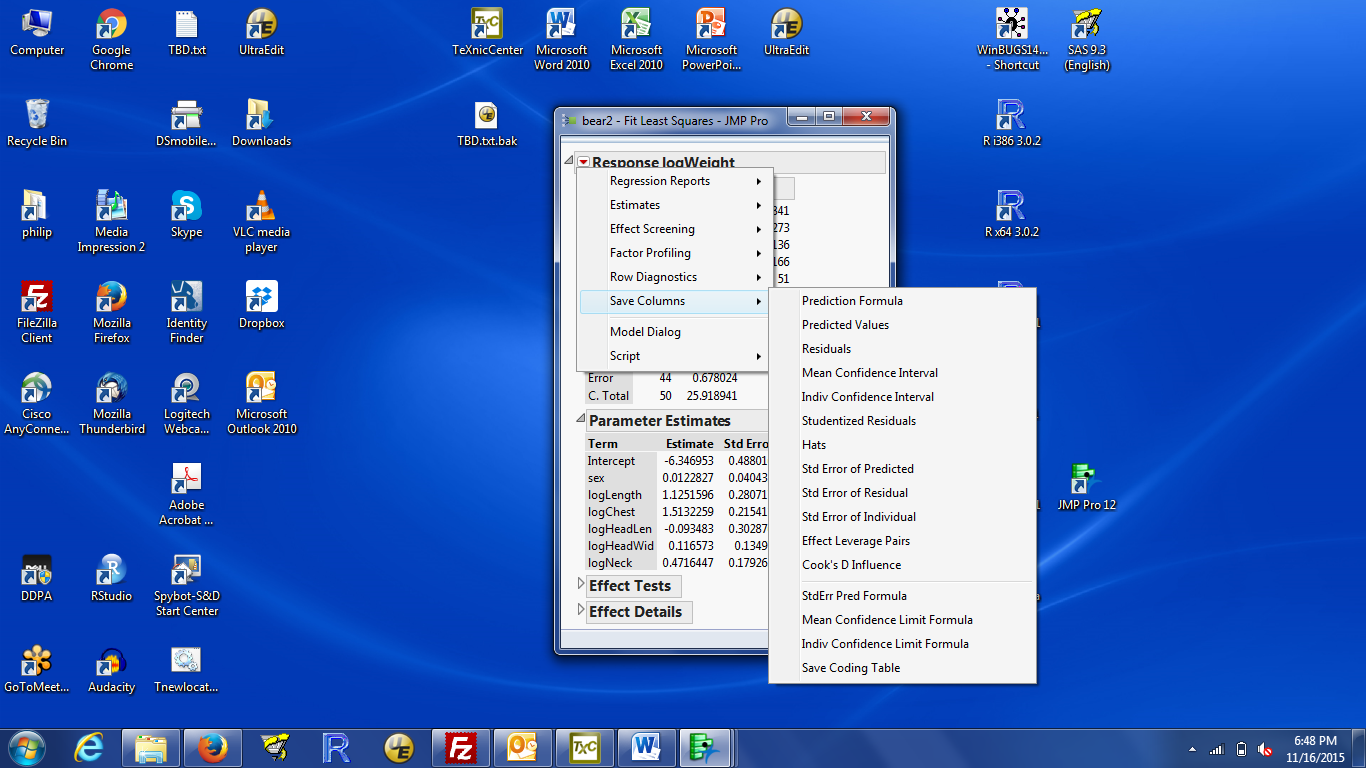
We will use the full bear weight data set to illustrate both of these. Earlier in the semester, we used Chest measurements and Sex to predict the weight of a black bear. In addition to chest, the investigators measured the overall Length, Head Length, Head Width and Neck diameter. Looking at a scatterplot matrix suggests that all variables should be log transformed.

The bear2.txt file on the class web site includes log transformed variables already created for you. Load the data set into JMP. Open the Analyze / Fit Model dialog, enter a model using sex and the five log transformed measurements (log length through log neck) to predict log weight, then run that model.

**Standardized residuals and Cook’s Distance:**

Both are observation-specific values, so they are obtained by saving new columns in the data set.

1. After fitting the model, click the red triangle by Response logWeight and select Save Columns. The list that appears includes all sorts of statistics that could be computed for each observation.



What I called Standardized residuals are called Studentized residuals by JMP (6th item from the top of the list). Cook’s Distance is the item labelled Cook’s D Influence (5th from the bottom).

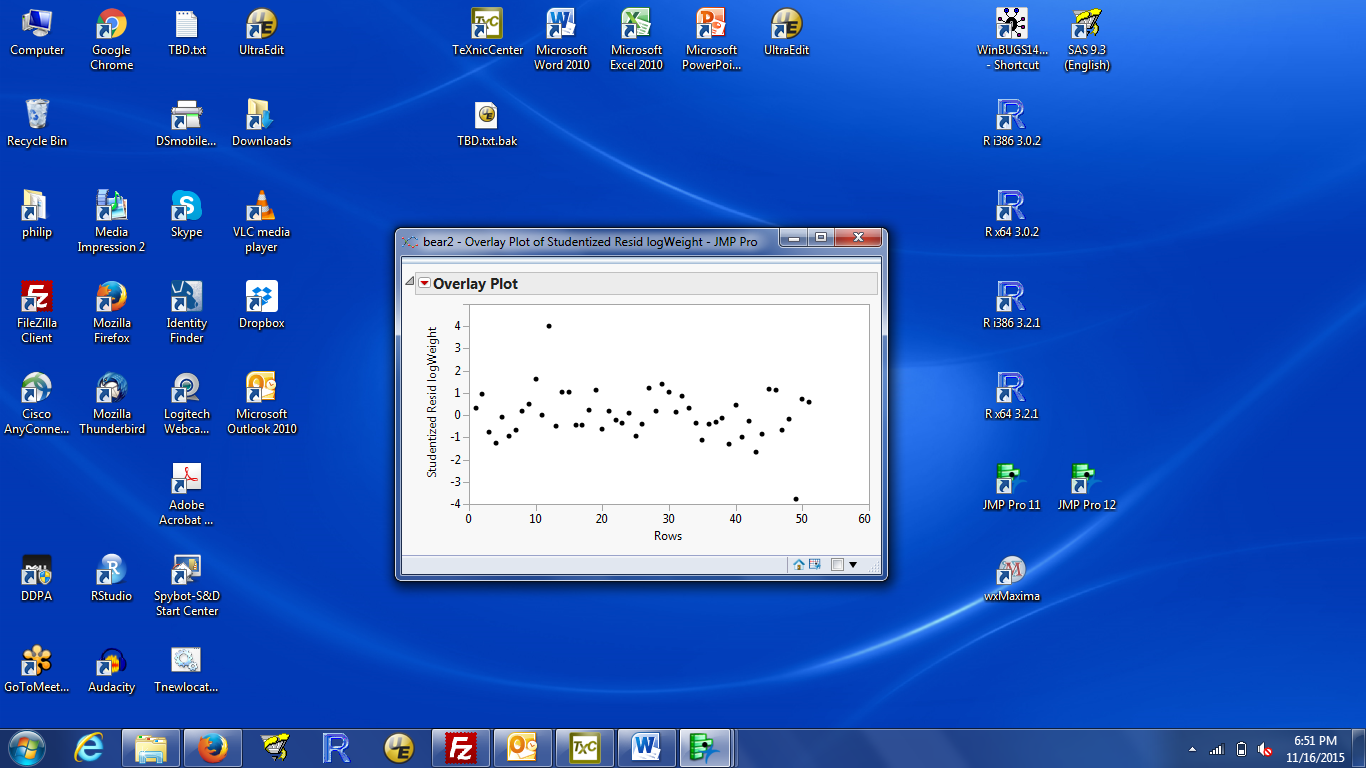
1. After you select your choice, a new column will be added to the data set. The columns are labelled by the quantity you requested and the response variable. That way if you fit models to more than one response variable (or a transformation of the response variable), you can match column to model.
2. Although you could look through the lists of numbers to find interesting values, it is easy to miss a number in a long list. The most common approach to scan a long list is to plot them.

The usual plot is Y = residual or D and X = row number.

Click in the data table to make that window active. Select Graph / Overlay Plot. Select the variable to be plotted (e.g. Studentized Resid logWeight) and put that in the Y box. You have two options for the X variable:

If you have an informative number for each observation (e.g. an id number), you can select that variable and put it in the X box. If you don’t, leave the X box blank. The row number (i.e. observation number) will be used as the X coordinate. There is no id variable for the bears, so I left X blank.

Click OK and you get the plot.



Note: The second approach (leaving X vacant to use the row number) only works with Graph / Overlay plot. Using Graph / Graph Builder doesn’t work.

1. See below for how to identify interesting points in this plot
2. Another option is to sort the list of numbers so the interesting values (unusually large or unusually small) are easy to find.

Click on the data window then right click on the column to sort on (e.g. Cook’s D Influence). Select Sort from the menu then select either ascending or descending. Descending puts the largest values first; Ascending puts the smallest values first.

You may get a warning from JMP about ‘Can’t reorder observations’. That’s because there is a plot like the plot of studentized residuals against row number that would be messed up by reordering (and hence renumbering observations). You can ignore this because JMP will open a second data set window with the observations in the sorted order.

**Identifying interesting observations:**

JMP implements linked graphics: when you select a point in one window, it is highlighted in all windows. This makes it easy to identify the interesting observations in a plot.

1. In the plot of Studentized residuals, there is an observation with a studentized residual that’s around 4. Which observation is this?
2. There are two ways to find out:

a) click in the plot window to activate it, then move the pointer over the point of interest. The row number should appear.

b) click in the plot window to activate it, then select the point of interest (left click near the point). Look in the data window and scroll down until you find the selected observation: # 12.

1. You can clear the selection by clicking somewhere far from any observation in the plot window. Or by clicking in the rows box on the far left of the data window. Clicking close to All rows will select all rows. Clicking anywhere else in that box will unselect all rows.
2. You can select multiple points by holding down the left mouse button while selecting or by shift-clicking multiple points.
3. You can work the other direction: if you select rows in the data window, they are highlighted in the plot.
4. If you have multiple plots open, points are highlighted in all windows. To demonstrate this, use Analyze / Multivariate Methods / Multivariate to draw a scatterplot matrix with the five log size variables and log weight. If you’ve closed the plot of studentized residuals vs row number, redraw that plot. Then left-click on the observation with the standardized residual of nearly 4. That observation is highlighted in all plots.

**PRESS:**

The PRESS statistic is available by request after fitting a model using Analyze / Fit Model.

1. Fit the model of interest. We will use the previously fitted bear model, predicting log weight from sex and five log transformed size measurements. The rMSE from this model is 0.124.
2. Click the red triangle by Response LogWeight, hover the mouse over Row Diagnostics to expand that menu and select PRESS.
3. A small results box, labelled Press, should appear below Parameter Estimates and Effect Tests. The Press number is the out-of-sample analog of the error sums-of-squares; the Press RMSE is the out-of-sample analog of the rMSE. In lecture, I have focused on what JMP calls Press RMSE. That is 0.169 for the bear data.

**Self assessment:**

We continue to evaluate the relationships between home sales price and assessed values for 84 homes sold in one neighborhood of Tampa Florida between 2008 and 2009. This is a subset of the data considered in Case Study 2 in the text. The data are in tamsales1.txt on the class web site. You will need to use text import / preview to read the file correctly. All my analyses use untransformed variables for demonstration.

Reminder: each row of data is for the sale of one house. The three variables are the sales price, the assessed value of the land, and the assessed value of the improvements (buildings, e.g. the house, garage, or garden shed, or an in-ground pool). The goal is to develop a model to predict sales price for a home about to be put on the market.

**Questions:**

All questions are based on the model that uses LAND, IMPROVE and their uncentered interaction LAND\*IMPROVE to predict PRICE.

1. Are there any observations that are potential outliers? If so, which observations are they?
2. Are there any observations with unusual influence on the fitted regression? If so, which observations are they?
3. Plot the Land and Improvements value for each sale and find where row 34 is plotted. Can you explain why the Cook’s D value for this observation is so high?
4. Find the rMSE and the Press RMSE. If you wanted to describe the ability of this model to predict new observations, not in the data set, which number is more appropriate?

**Answers:**

1. Yes, row 34 has an unusually small studentized residual (residual < -3) while rows 63 and 76 have unusually large studentized residuals (residual > 3).
2. Yes, rows 34 and 63 have Cook’s D larger than 1, so are considered unusually influential.
3. This sale has extreme values for the two X variables and an unusually low sales price (large negative residuals), so that sale has a disproportionate influence on the prediction equation.
4. rMSE: 131,819.6 PressRMSE: 177,258.4

The PressRMSE is a better description of the ability to predict sales prices of new homes on the market.