**Lab 5:** **Transformations, QQ plots**

**Goals:**

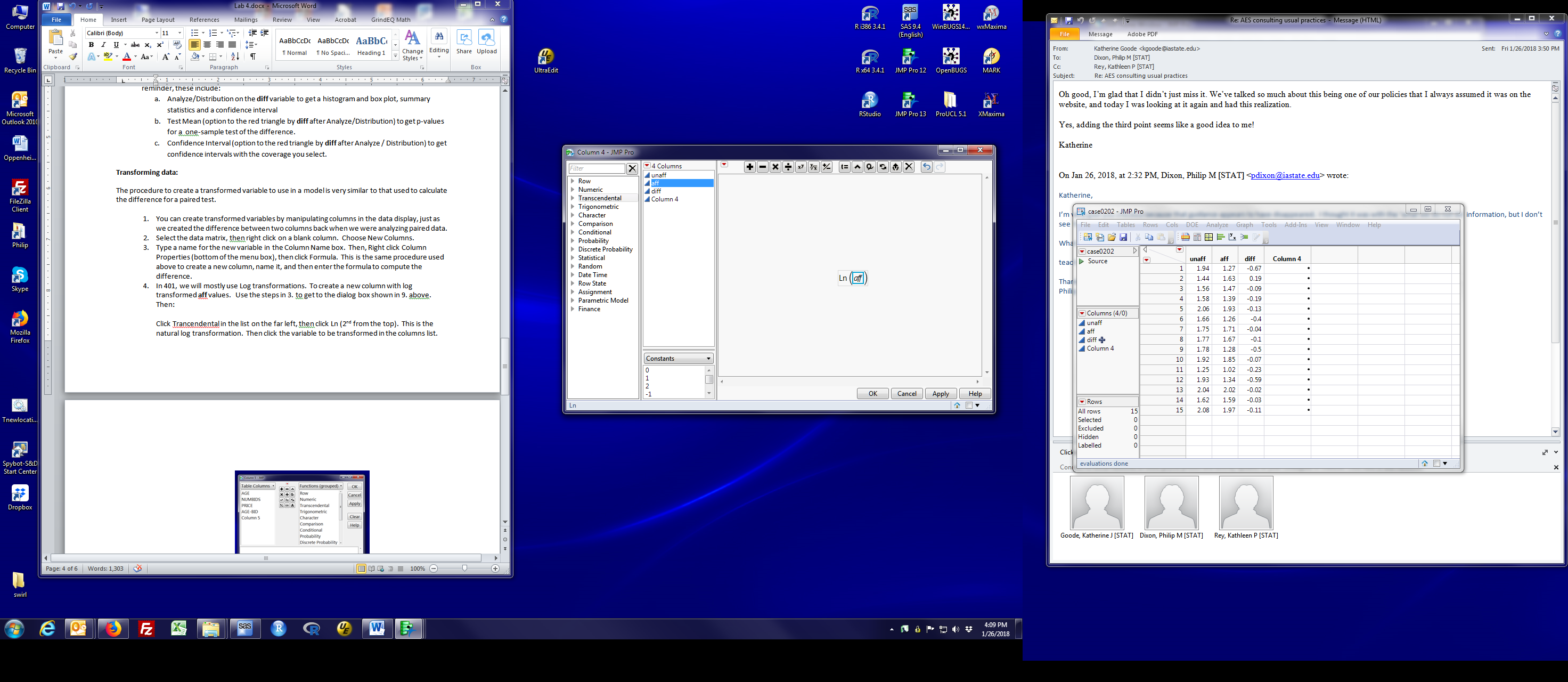
1. How to compute transformations
2. How to draw QQ plots
3. How to store residuals from an analysis, then draw a QQ plot
4. How to plot residuals vs predicted values

**Transforming data:**

The procedure to create a transformed variable is very similar to that used to calculate differences.

1. You can create transformed variables by manipulating columns in the data display, just as we created the difference between two columns back when we were analyzing paired data.
2. Select the data matrix, then right click on a blank column. Choose New Columns.
3. Type a name for the new variable in the Column Name box. Then, Right click Column Properties (bottom of the menu box), then click Formula. This is the same procedure used above to create a new column, name it, and then enter the formula to compute the difference.
4. In 587, we will mostly use Log transformations. To create a new column with log transformed **aff** values. Use the steps in 3. to get to the dialog box shown in 9. above. Then:

Click Trancendental in the list on the far left, then click Ln (2nd from the top) or Log (3rd from the top). This is the natural log transformation. This is what I want when I talk about log transformations. Then click the variable to be transformed in the columns list. aff should appear in the formula box and you should see:



1. Click Ok and a new column will appear that contains log aff values.
2. If you do want log base 10, then choose log10 from the transendental list.
3. The square root transformation is obtained by clicking the button on the top row of buttons. Put aff in the x place and leave y blank.
4. If creating a transformed variable in JMP is opaque, you can always read the data set into Excel, create a new spreadsheet column with the transformed variable, save the new worksheet, then read the new data set into JMP.

**QQ plots**

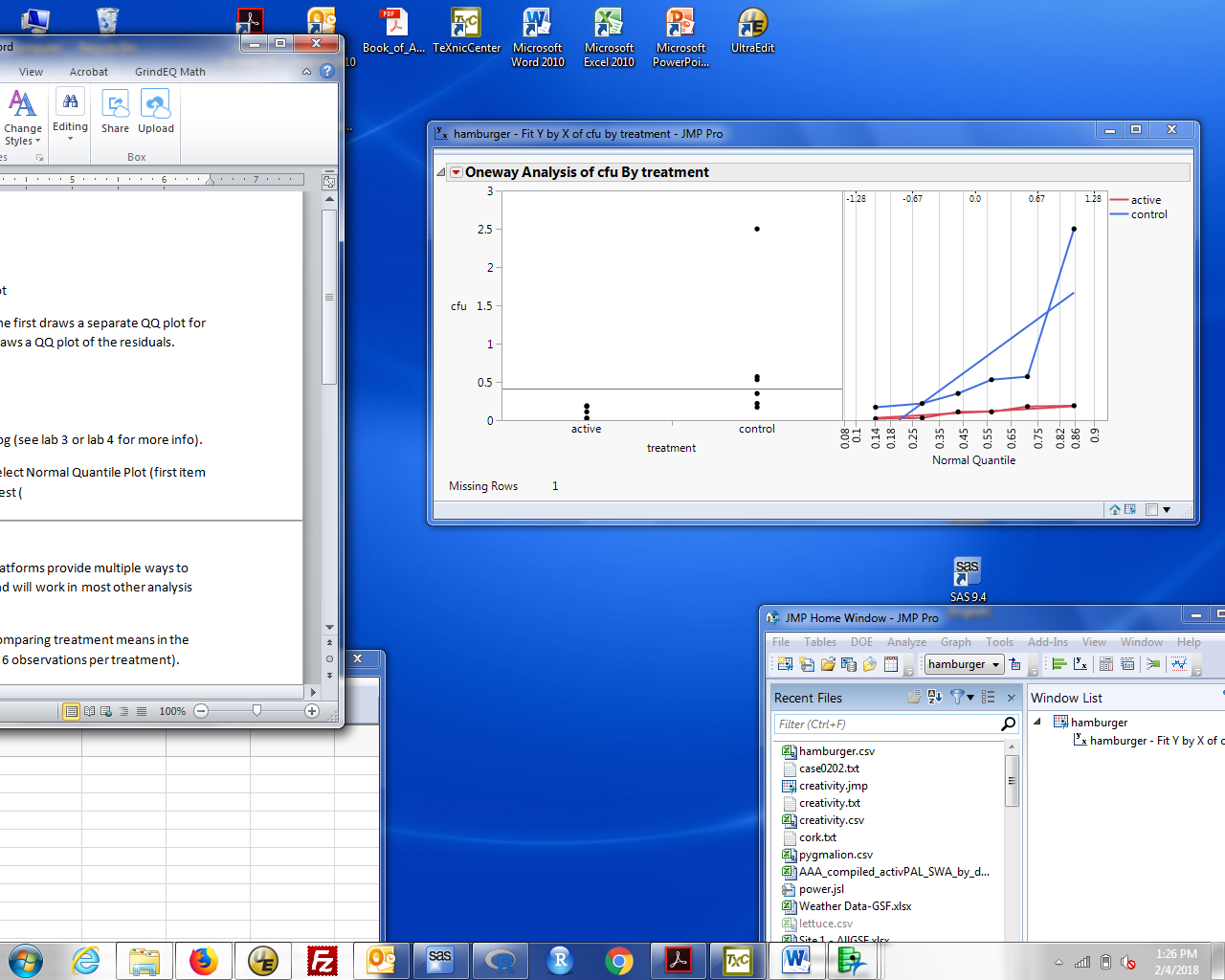
There are two ways to draw QQ plots to assess normality in JMP. The first draws a separate QQ plot for each group of observations. The second saves the residuals then draws a QQ plot of the residuals.

We will use the hamburger data to demonstrate this.

**Separate plots for each group:**

Load the hamburger data file (see lab 4) and start the Fit Y by X dialog (see lab 3 or lab 4 for more info).

Click the red triangle by Oneway Analysis of cfu by treatment and select Normal Quantile Plot (first item in the second box of options). You can do this either before or after running the t-test. You will see something like:



The plot on the right is added to the dot plot of the data. This shows a QQ plot for each group separately. The jagged lines with dots are the QQ plot of the data. The straight lines are what is expected if the observations followed a normal distribution. The red data (active) fits a normal; the blue data (control) do not.

This is a QQ plot of the observations. When draw separately for each group, it is the same as a QQ plot of the residuals. That is because within each group, the residuals are just a shifted version of the data values. As discussed in lecture, this is only true when you look at each group separately. When you look at both groups together, the properties of the residuals are no longer the same as those of the observations.

To look at residuals for all observations as one QQ plot, you need to store the residuals. The Fit Y by X dialog allows you to do this.

**Calculating and storing residuals:**

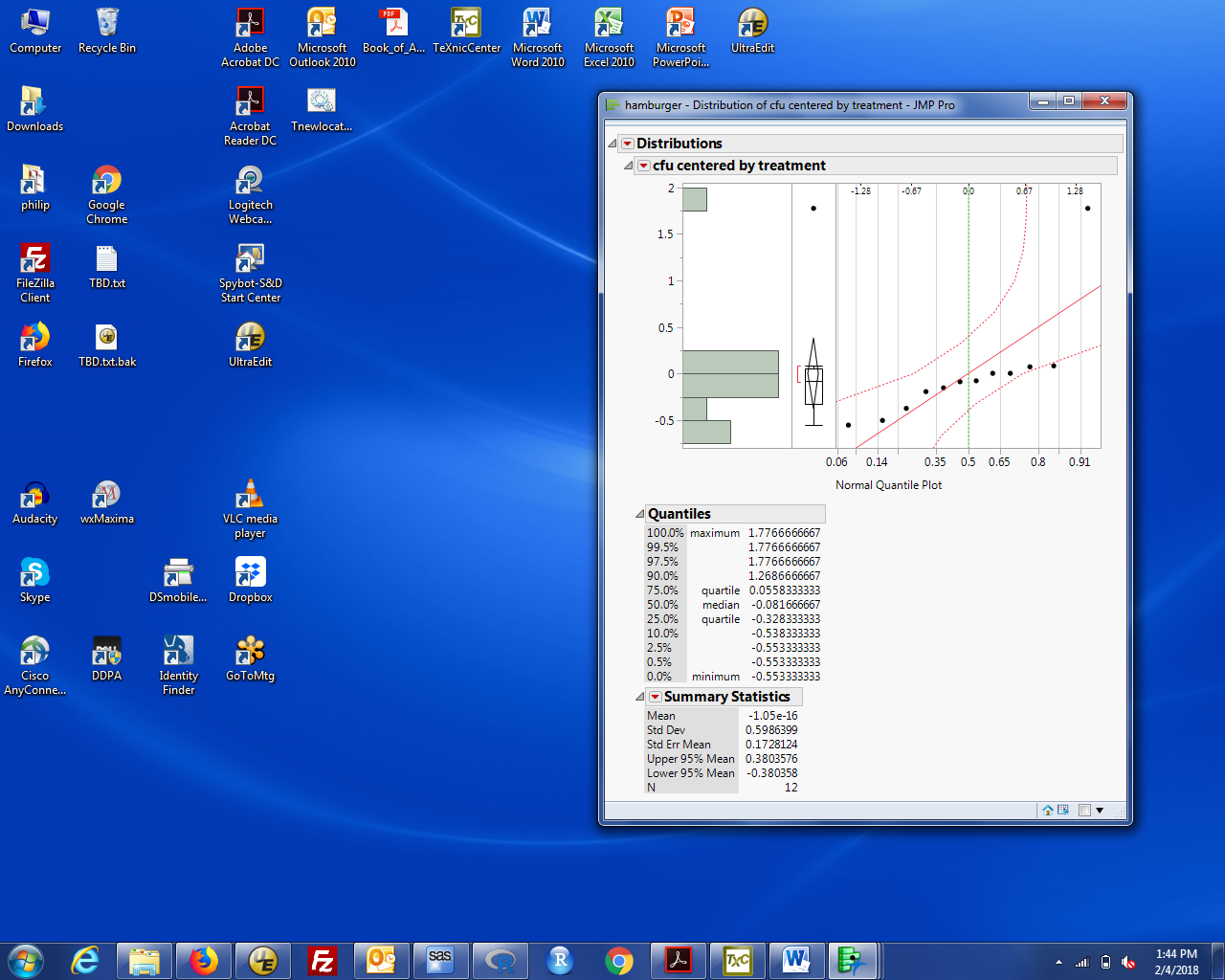
Some details depend on the analysis you’re doing. Some analysis platforms provide multiple ways to extract and manipulate residuals. Here’s what works in Fit Y by X and will work in most other analysis procedures.

Run the desired analysis, for example use Fit Y by X followed by Means/ANOVA/Pooled t to get a t-test. After you get the analysis, click the red triangle by ‘Oneway analysis …’. Select Save / Save Residuals. A column is added to the data table. That column has the residuals. If you use Save / Save Predicted, you get a column with the predicted values. These can be used for any subsequent plot or analysis.

To get a QQ plot from a column of residuals:

Select the data frame. You should have created a new column with the residuals. For the hamburger analysis, this will be labelled **cfu centered by treatment**. Select Analyze / Distribution and put the residual column into the Y box, and click OK. You should recognize this results box from lab 1 when we calculated descriptive statistics.

Click the red triangle by cfu centered on treatment and select normal quantile plot. A new plot will appear next to the vertical histogram and box plot. Here's what it should look like:



The dots are the observations and the solid straight line is what you would expect if the data were normal. The dashed curved lines are uncertainty bounds that depend on the sample size. If the observations are normal, you expect almost all the points to fall inside these bounds. I don't use these bounds because they are very conservative. Notice that these data, with a clear outlier, still appear normal (by the all points within the bounds criterion).

If you want to get rid of the box plot and vertical histogram (so you only have the QQ plot), click the red triangle by cfu centered by treatment and:

click Outlier Box Plot to uncheck it

select Histogram Options and click Histogram to uncheck it.

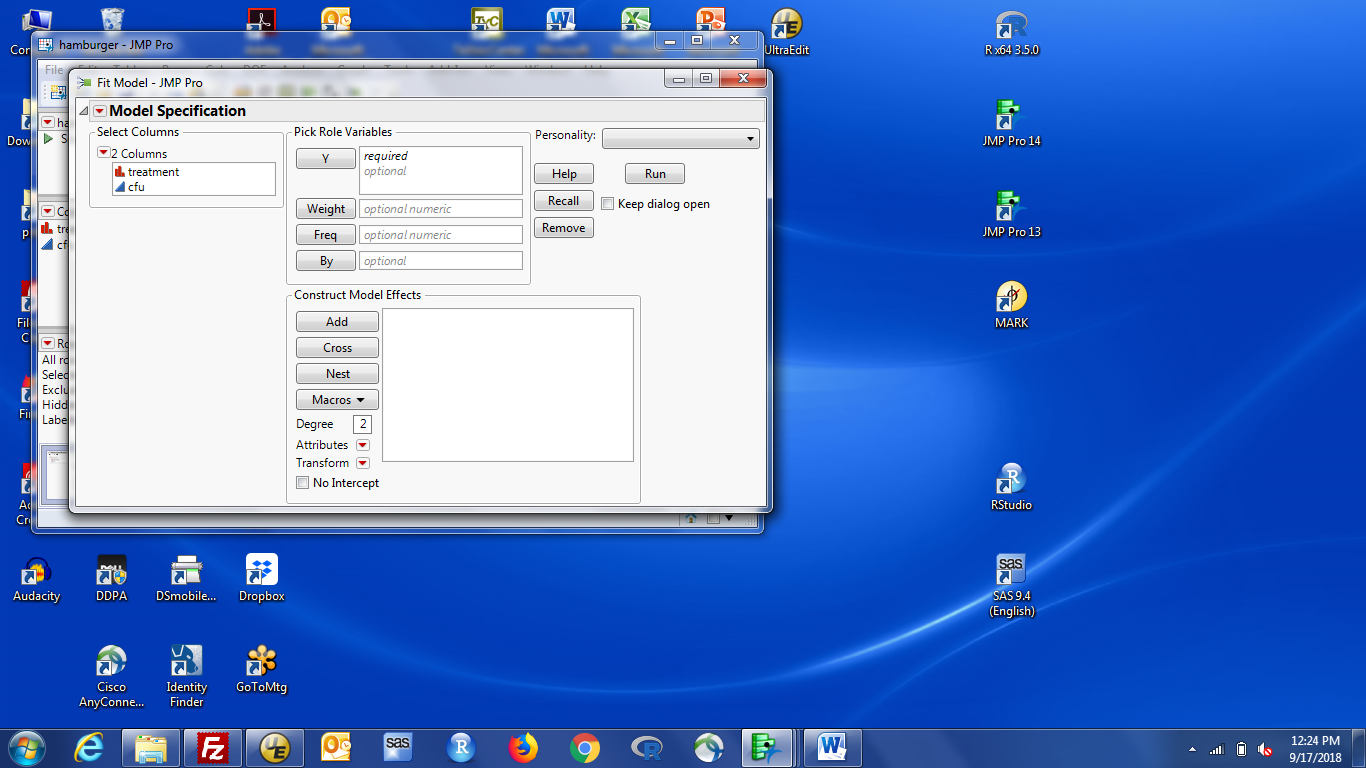
Now you just have the QQ plot.

Remember, you can copy the plot to your HW answers either by copying (ctrl-C) the entire Distributions window, pasting into a Word document then deleting unwanted text, or by saving the results in the Journal (ctrl-J) then saving the Journal as a .rtf or .doc file.

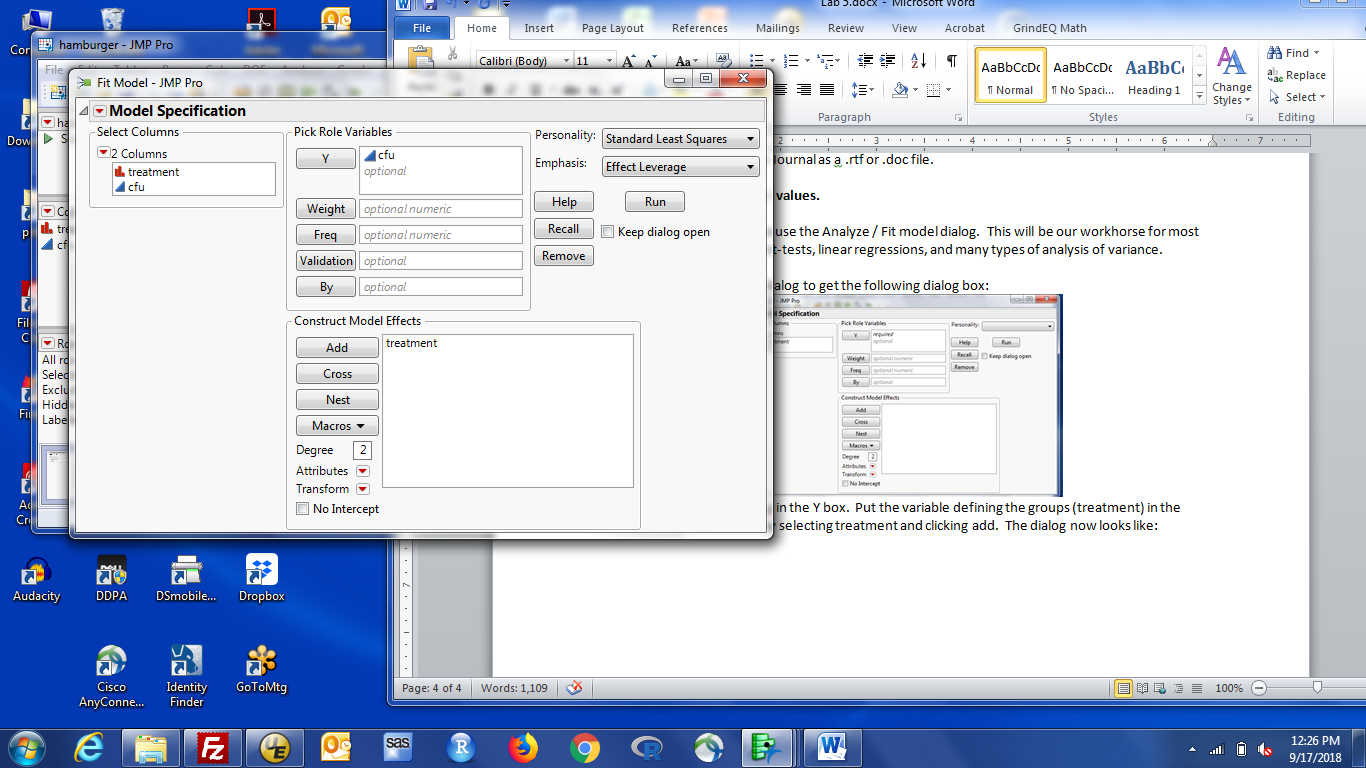
**Plotting residuals vs predicted values.**

The easiest way to do this is to use the Analyze / Fit model dialog. This will be our workhorse for most of the semester because it fits t-tests, linear regressions, and many types of analysis of variance.

Start the Analyze / Fit Model dialog to get the following dialog box:

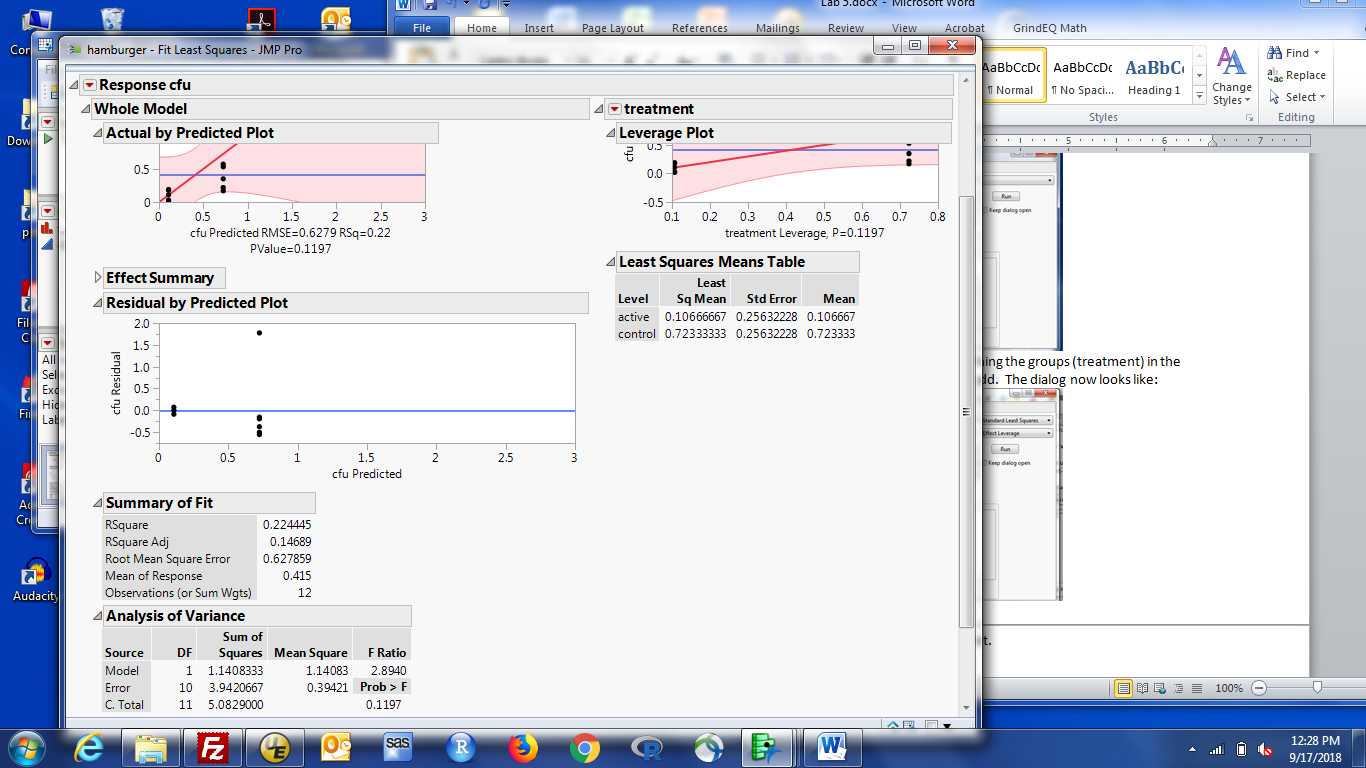


Put the response variable (cfu) in the Y box. Put the variable defining the groups (treatment) in the Construct Model Effects box by selecting treatment and clicking add. The dialog now looks like:



Click the Run box.

The default output includes a small residual vs predicted value plot.



If you want a nicer plot, the strategy is to save the predicted values and residuals, then use graph builder to plot the data.

To save the predicted values and residuals: click the red triangle by Response CFU, select save columns, then prediction formula. Repeat choosing residuals at the last step. The data window now includes two additional columns.

Note: prediction formula and predicted values are very similar. prediction formula is slightly more general. We'll talk more about the difference when we talk about regression.