**Stat 401 – Lab 11**

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

construct and use indicator variables

calculate regression diagnostics

fit regression models to all subsets of variables

calculate the PRESS statistic to quantify “out of sample” prediction error

**Note:** There is a lot of material this week. JMP provides multiple ways to accomplish many of this week's tasks**.** I show you the basics. The lab 11 extra document describes options that could be helpful.

**Constructing indicator variables by hand:**

Simplest: edit the data set by hand and add a column with the indicator values. Remember to add a column, either Cols/New Column, or double click on the top row of the data set window (the row with the variable names).

If you create your own indicator variables, you control how they are defined. However, it is a bit harder to get model comparison tests. If you need this, see the lab 11 extra document. Or, let JMP create the indicator variables for you.

If you need to do this for a lot of observations, you can write a formula to do this. See the lab 11 extra document.

**Creating indicator variables automatically in JMP:**

When you use a nominal variable (red bars) in a regression model, JMP automatically converts that to one or more indicator variables. If the nominal variable has two levels (e.g. time with “E” and “L”), JMP creates one indicator variable. If it has *k* levels, JMP creates *k-1* indicators.

In lecture, I said (or will say) that the definition of indicator variables is arbitrary. Class and the text use ‘last level is 0’ indicators. JMP uses +1/-1 indicators. For two levels, the first, e.g., “E”, gets a +1 and the last, e.g. “L”, gets a -1. To create this indicator automatically, Analyze / Fit model, then add the nominal variable, e.g. time, (not the indicator version of time) to the model. Then run the model. My example has time (nominal) and light (continuous) in the model.

JMP can report results for automatic indicator variables the way we have discussed in class. To get these, fit the model with the automatic indicator variables. Then click the red triangle by Response, select Estimates / Indicator Parameterization Estimates. You see a new box of estimates labelled Indicator Parameterization. These estimates follow the "set last category to 0" parameterization. That means L gets the value of 0 and E gets the values of 1. You see this by time[E] as the label for the parameter. The estimated value (12 and a bit) is the negative of what you got when you created your own indicator variable with E = 0 and L=1.

**Using an indicator variable in a regression:**

Just add the variable to the model.

**Creating an interaction variable that involves an indicator variable:** No different from creating an interaction of two continuous variables. See Lab 10 discussion of quadratic and cross-product variables for the instructions.

**Regression diagnostics**:

Load the bear.txt data file (you probably will have to use Data Import Best Guess or Preview). This contains various measurements made on black bears in Pennsylvania. The goal is to predict bear weight without having to actually weigh the bear. Measuring length, chest girth, etc. is easy. Weighing a 400 lb bear in the field is not. The bear.txt file has the original measurements.

We will consider the additive linear model that predicts weight from length, chest, headlen, headwid, and neck.

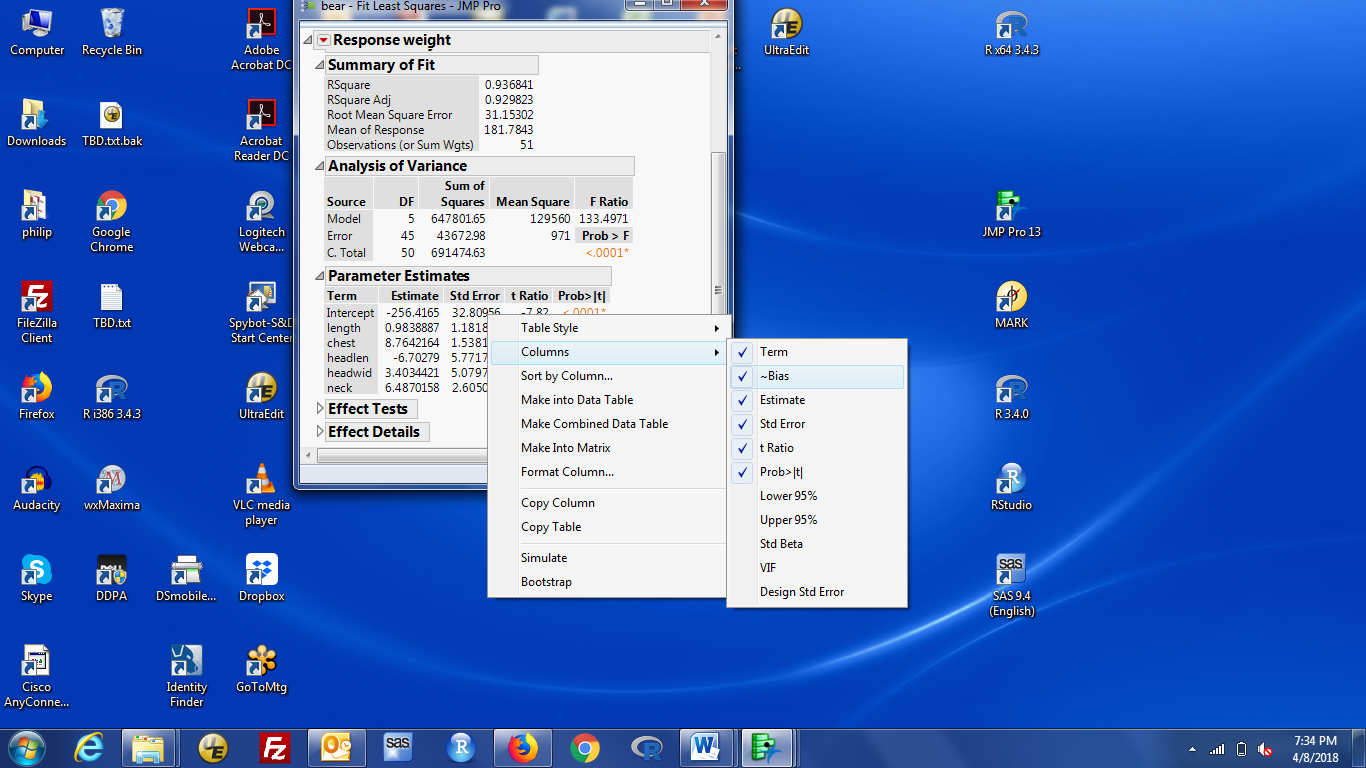
Start the Analyze / Fit Model dialog. Put weight in the Y box and the 5 X variables: length, chest, headlen, headwid, and neck in the Construct Model Effects box, then click Run to fit the model.

We have talked about two types of diagnostic. Some, like residuals and Cook's D, have values for each observation. Some, like VIF, are variable specific. They are obtained in different ways

Observation-specific measures (e.g. Cook’s D, standardized residuals): Fit the model, using Analyze / Fit Model, not Stepwise. Click the red triangle for the model, select Save Columns. Select the columns you want: e.g. Studentized Residuals, Cook’s D, or one of the other variables. New columns are added to the data table.

To plot against observation number, the best plot comes from Overlay Plot. The problem with Graph Builder is that we don't have a variable containing the observation number. Overlay Plot doesn't require this. Graph / Overlay Plot (very bottom of the list). Put Cook's D in the Y box and leave the X box empty. Click done and you get a plot with observation number on the X axis.

Variable-specific measures (e.g. VIF): Fit the model, then find the Parameter Estimates table. Right click inside the table. You should get a menu that includes Columns. Select Columns and you will get a menu with options for numbers to include in the table:



left-click on VIF (near the bottom of the list) to put a check by that statistic. A column with VIF values will be added to the Parameter Estimates table.

**To calculate the PRESS statistic:**

1. Fit a model. Then, click the red triangle for that model, select Row Diagnostics, then select Press.
2. A small output box, labelled Press, with two numbers is included with the output.

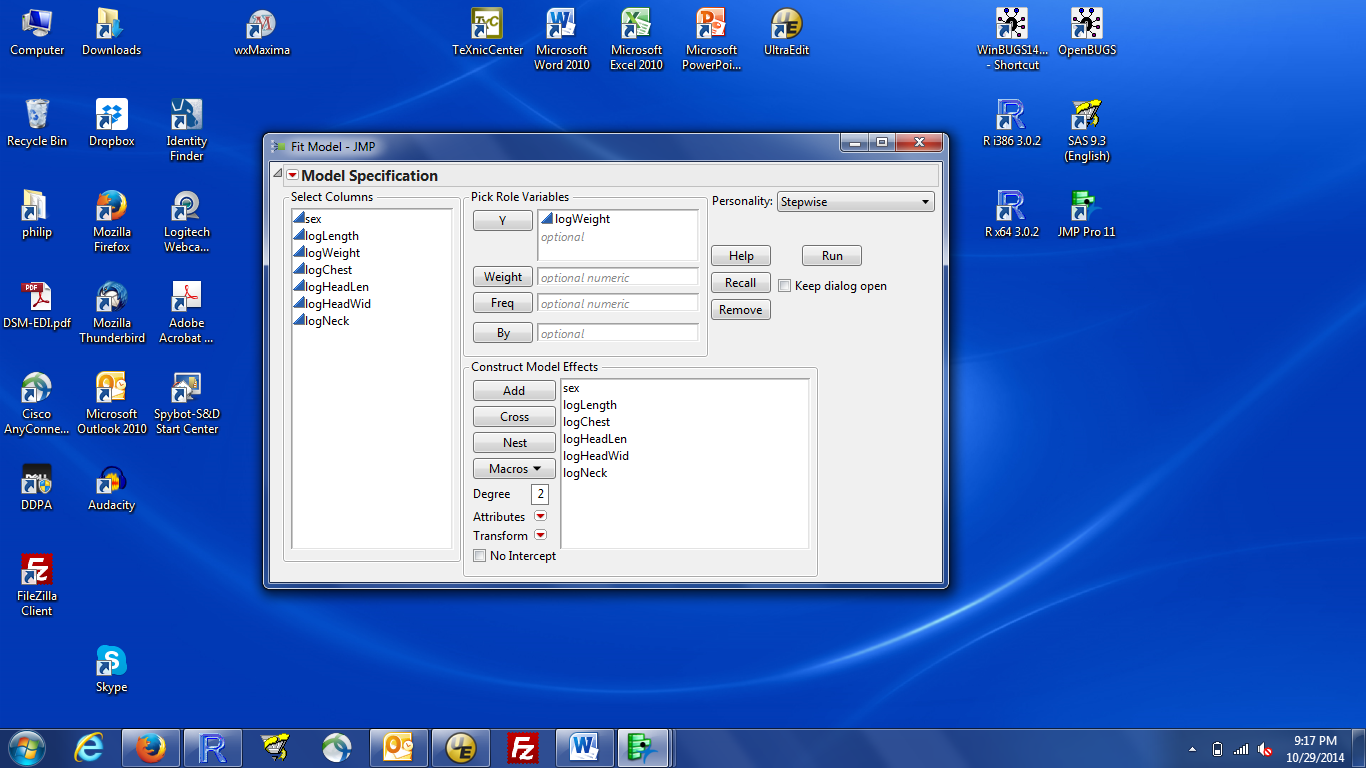
The Press number is a Sum-of-Squares; the Press RMSE is a standard deviation.

**To fit regression models with all subsets of variables:**

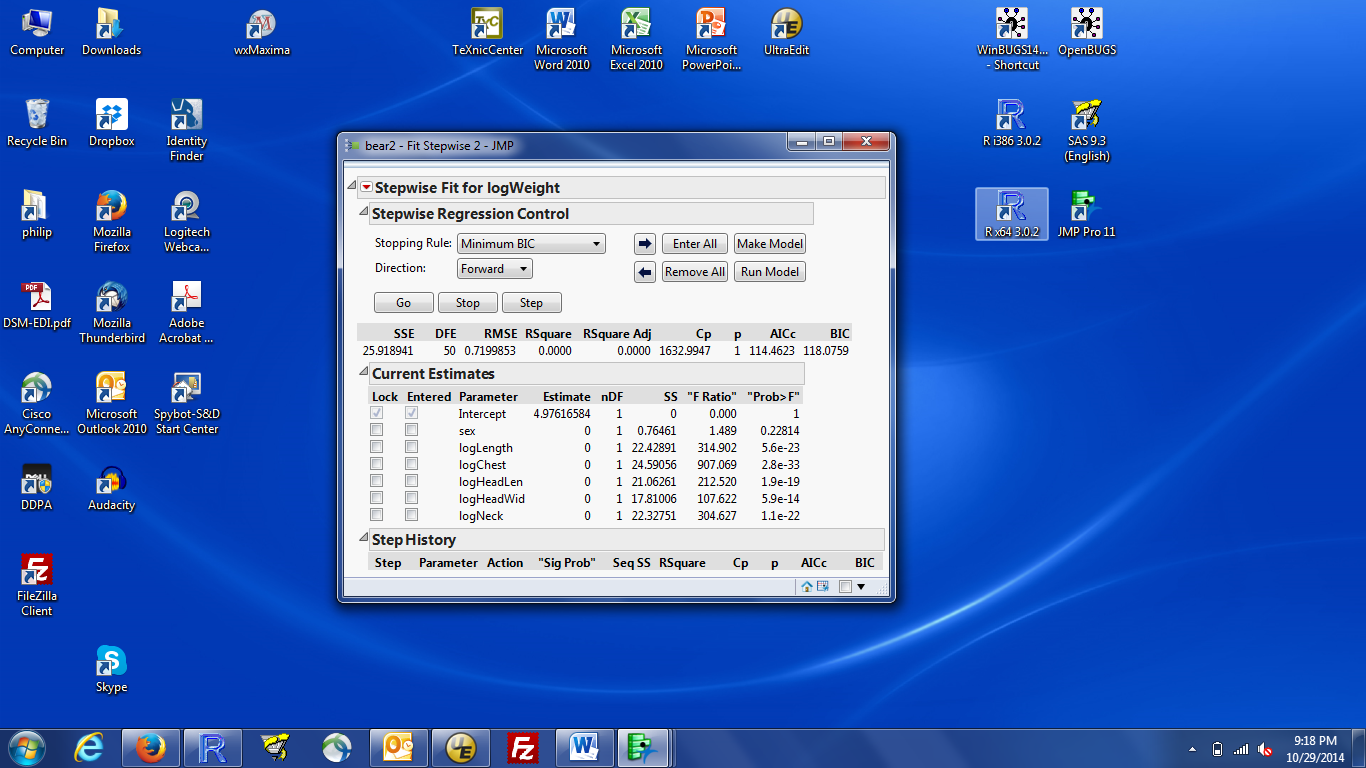
Load the bearlog.txt data file (you will probably have to use Data Import Best Guess or Preview). This is the log transformed versions of the data in bear.txt. If you look at the scatterplot matrices for the original data and for the log transformed data, you see why we should work with the log transformed data. Biological theory (allometry) also suggests linear relationships between log transformed size measures.

The question is ‘which variables?’. To answer this:

1. Analyze / Fit Model. Select logWeight as the Y variable and all other variables as X variables.
2. Look for the Personality box (top right of the Fit Model dialog). It shows Standard Least Squares by default. Left click on that and change the Personality to Stepwise. The Fit Model dialog should look like:



1. Then run the analysis. You should get a results box that looks like:



1. Click on the red triangle at the top left (by Stepwise Fit), and left click on All Possible Models
2. You will get a dialog box that helps you limit the number of models that see results from. JMP will fit all possible models; you don’t want to see results from all of them. You can limit the number of variables that can go in the model. Here JMP supplies 6, the number of X variables. That’s reasonable. You can also see how many models you want to see. This is the number of one-variable models, the number of two-variable models, … up to the number of 6 variable models. The default depends on the number of X variables. For this data set, JMP suggests 15 results per number of variables, which here would give you results for 90 models. I think that’s more than I want to see, so I suggest you change the 15 down to 4 or 5. If you decide later that you want to see more models, you can always rerun with a larger number. Then click OK.

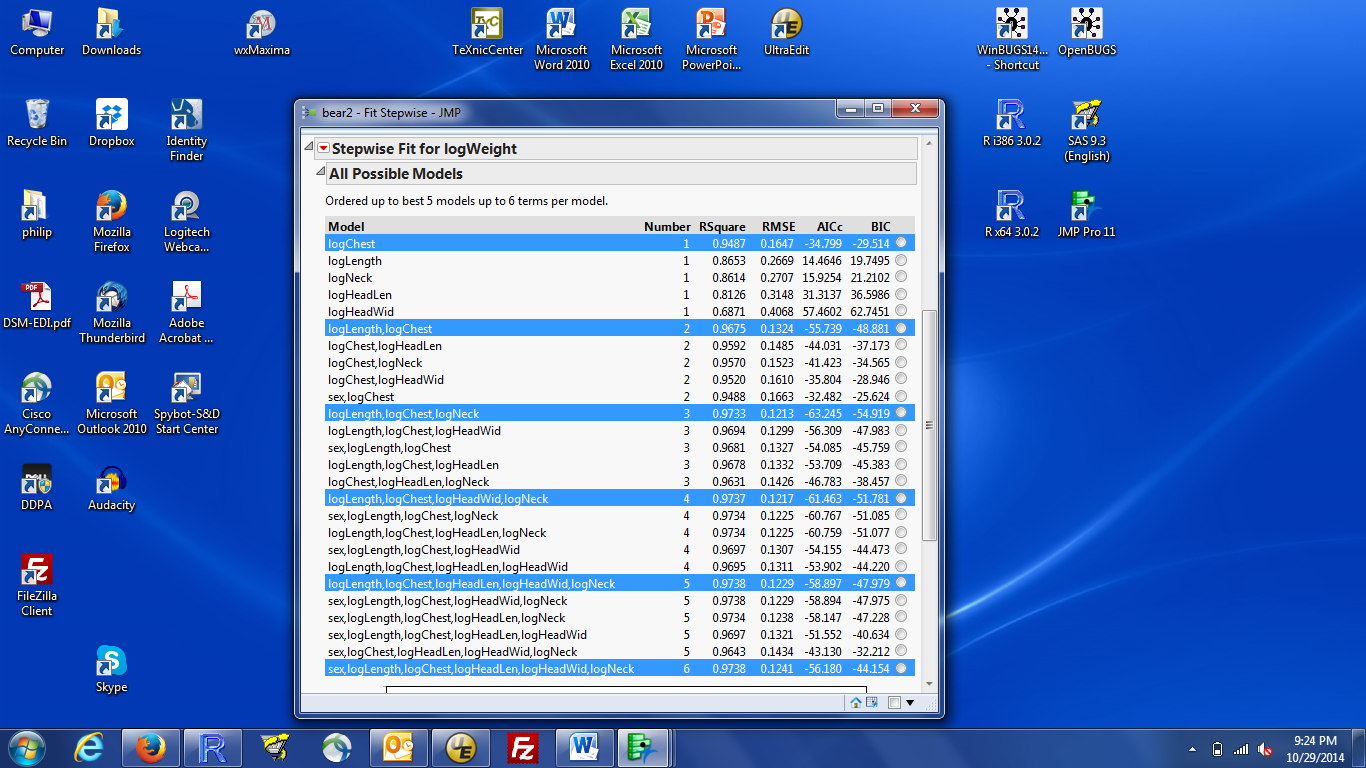
Notes:

1) The choice of 15 or 4 or 5 is a *display* option. JMP will fit all models. For 6 variables, that is 63 models.

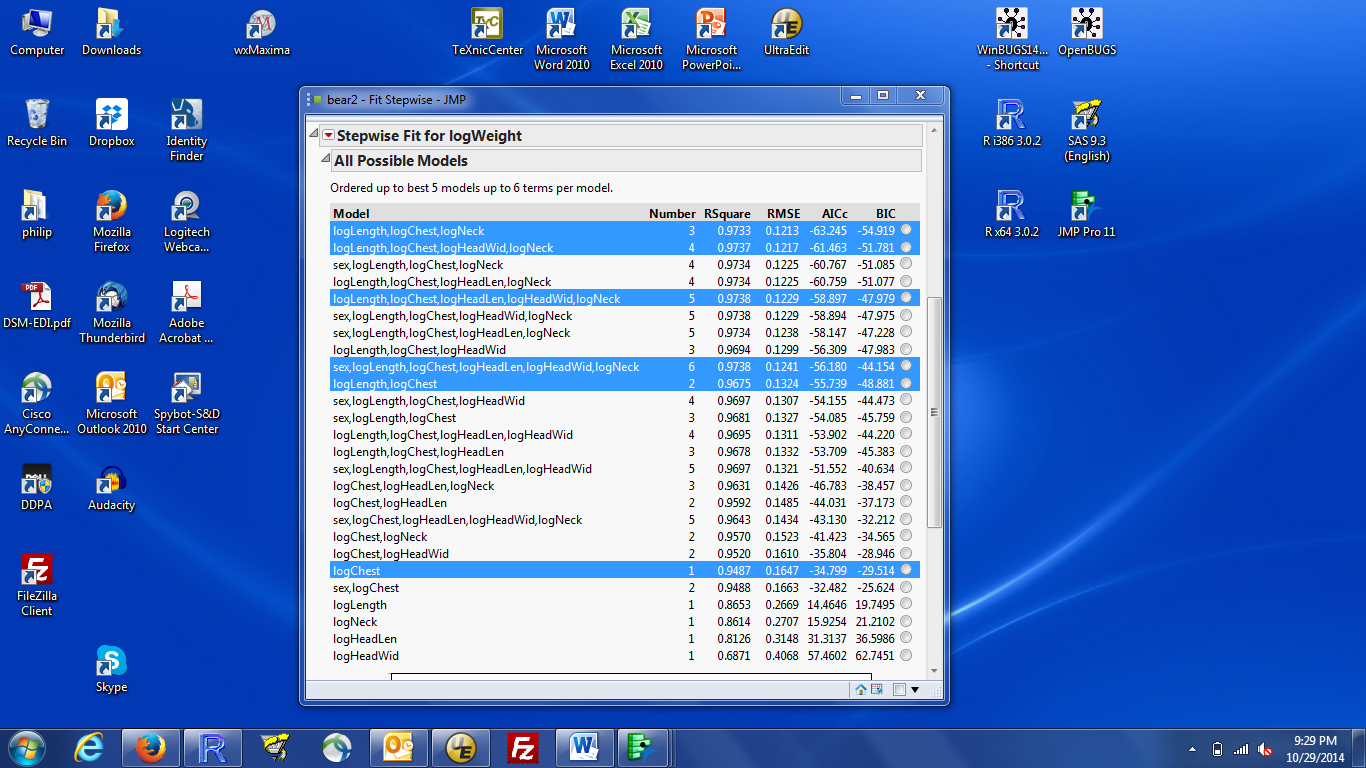
The choices in the menu box only control how much you see. You don’t want to see results for all models!

2) We will talk in lecture about ‘obeying the hierarchy’. If you have interactions or polynomial terms, you probably want to obey the hierarchy, so you want to check the box. If you don’t have any interactions or polynomial terms (the situation here), this option has no effect.

1. The results look like:



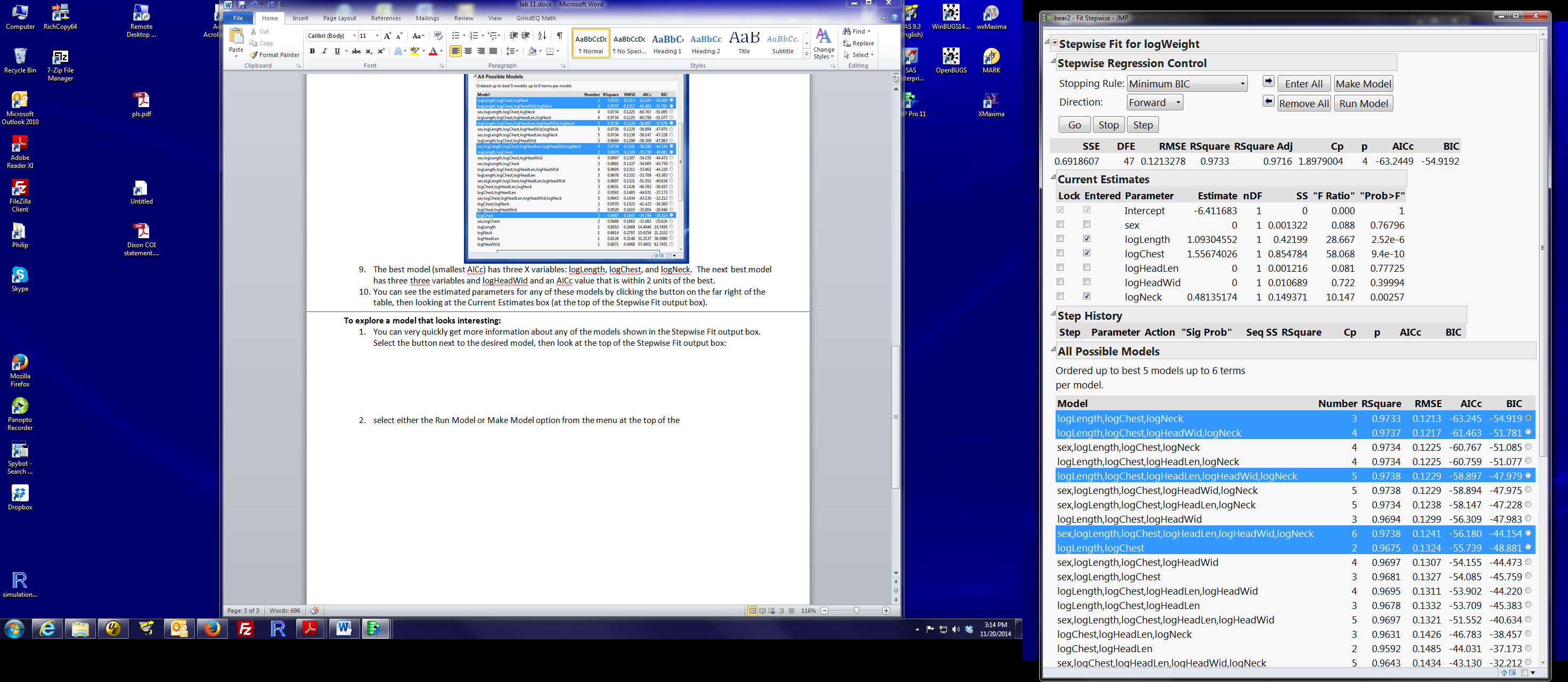
1. You see that JMP is giving you model summary statistics for 5 one-variable models, then 5 two-variable models, and so on. If you provided 4 models per model size in step 5, you would only get the best 4 one-variable models, then the best 4 two-variable models, etc.
2. It is useful to sort this table by increasing value of the AICc statistic (so the model with the smallest AICc is at the top, the second smallest is next, and so on). To sort it, right-click in the table, select Sort by Column, select the AICc column, check Ascending, then OK. The table will now look like:



1. The best model (smallest AICc) has three X variables: logLength, logChest, and logNeck. The next best model has three three variables and logHeadWid and an AICc value that is within 2 units of the best.
2. You can see the estimated parameters for any of these models by clicking the button on the far right of the table, then looking at the Current Estimates box (at the top of the Stepwise Fit output box).

**To explore a model that looks interesting:**

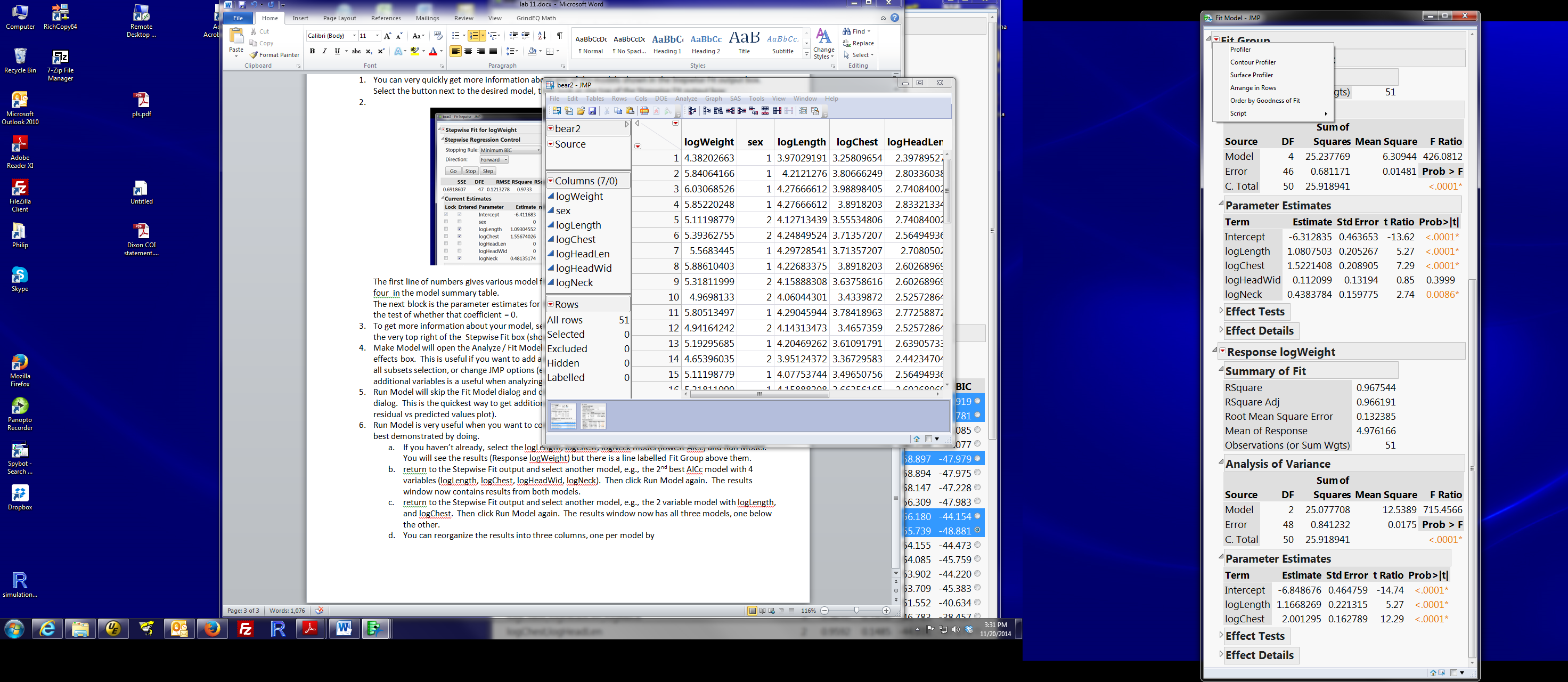
1. You can very quickly get more information about any of the models shown in the Stepwise Fit output box. Select the button next to the desired model, then look at the top of the Stepwise Fit output box:



The first line of numbers gives various model fit statistics, in case you wanted something other than one of the four in the model summary table.

The next block is the parameter estimates for the variables in the desired model. The Prob>F is the p-value for the test of whether that coefficient = 0.

1. To get more information about your model, select either the Run Model or Make Model box from the menu at the very top right of the Stepwise Fit box (shown in the display above)
2. Make Model will open the Analyze / Fit Model dialog, with the specified variables already added to the model effects box. This is useful if you want to add additional variables, even some that were not considered in the all subsets selection, or change JMP options (e.g. centering polynomials). In lecture, we talked (or will talk) about why adding additional variables is useful when analyzing observational data.
3. Run Model will skip the Fit Model dialog and directly give you the model results as if you had run the Fit Model dialog. This is the quickest way to get additional information about the model (e.g. the PRESS statistic or the residual vs predicted values plot).
4. Run Model is very useful when you want to compare detailed results from two up to a few models. This is best demonstrated by doing.
   1. If you haven’t already, select the logLength, logChest, logNeck model (lowest AICc) and Run Model. You will see the results (Response logWeight) but there is a line labelled Fit Group above them.
   2. return to the Stepwise Fit output and select another model, e.g., the 2nd best AICc model with 4 variables (logLength, logChest, logHeadWid, logNeck). Then click Run Model again. The results window now contains results from both models.
   3. return to the Stepwise Fit output and select another model, e.g., the 2 variable model with logLength, and logChest. Then click Run Model again. The results window now has all three models, one below the other.
   4. You can reorganize the results into three columns, one per model by clicking on the red triangle by Fit Group. (That controls actions that work on all three sets of results; if you click the triangle for one model, you only affect that model).



* 1. Select the Arrange in Rows, fill in how many results in one row. I usually choose the number of result sets I have, but you don’t have to.
  2. You now have results for each model displayed side by side. Much easier to compare among models.

1. (optional): JMP provides graphical tools to explore the collection of models in a Fit Group. For example, the Profiler (red triangle by Fit Group / Profiler) brings up plots that portray how the predicted response changes across the range of each X variable for each model. Ask me in lab if you want an explanation of these plots.