**Stat 401 A – Lab 10**

**Goals:** In this lab, we will learn tools for multiple regression. Specifically, we will see how to:

draw scatterplot matrices and estimate correlation coefficients

fit a multiple regression (more than one X variable), and interpret the results

evaluate diagnostics

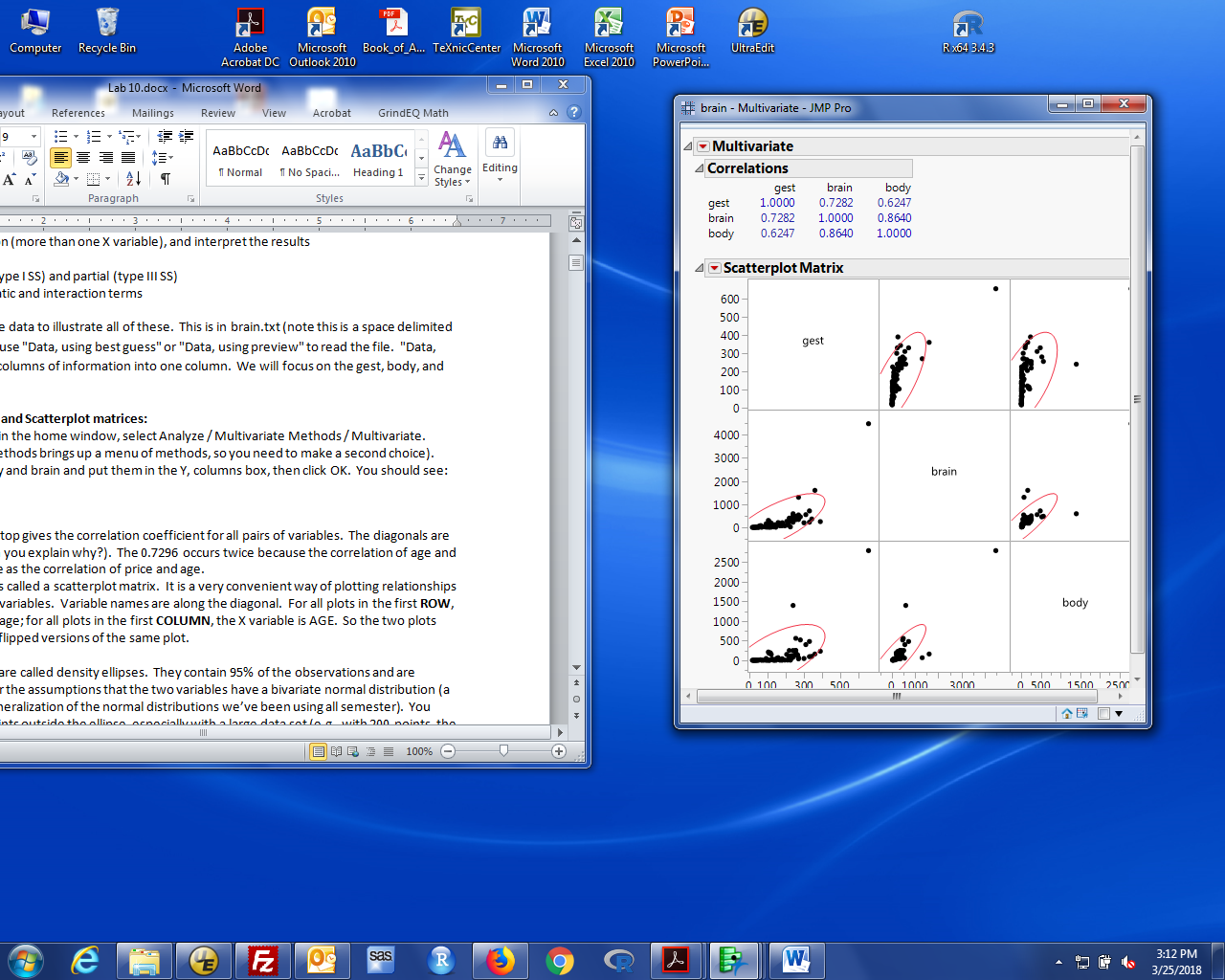
calculate sequential (type I SS) and partial (type III SS)

fit models with quadratic and interaction terms

We will use the brain size data to illustrate all of these. This is in brain.txt (note this is a space delimited file), so you will need to use "Data, using best guess" or "Data, using preview" to read the file. "Data, using default" gloms all columns of information into one column. We will focus on the gest, body, and brain variables.

**Correlation coefficients and Scatterplot matrices:**

1. From the menu in the home window, select Analyze / Multivariate Methods / Multivariate. (Multivariate Methods brings up a menu of methods, so you need to make a second choice).
2. Select gest, body and brain and put them in the Y, columns box, then click OK. You should see:



1. The table at the top gives the correlation coefficient for all pairs of variables. The diagonals are always 1.00 (can you explain why?). The 0.7282 occurs twice because the correlation of gest and brain is the same as the correlation of brain and gest (can you explain why?).
2. The plot below is the scatterplot matrix. I find them to be a very convenient way of plotting relationships among multiple variables. We discussed the interpretation of these in lecture. To remind you: Variable names are along the diagonal. For all plots in the first **ROW**, the Y variable is gest; for all plots in the first **COLUMN**, the X variable is gest. So the two plots "across" the diagonal are flipped versions of the same plot.

The red ellipses are called density ellipses. They contain 95% of the observations and are computed under the assumptions that the two variables have a bivariate normal distribution (a two-variable generalization of the normal distributions we’ve been using all semester). You expect some points outside the ellipse, especially with a large data set (e.g., with 200 points, the ellipse is expected to contain 95% of 200 = 190 points, which puts 10 points outside the ellipses). I’m not emphasizing these ellipses, but some folks find them helpful to identify unusual observations. They're not very helpful unless the data are bivariate normal. When not, I find them distracting.

You can eliminate the density ellipses and just show the points by:

clicking the red triangle by scatterplot matrix,

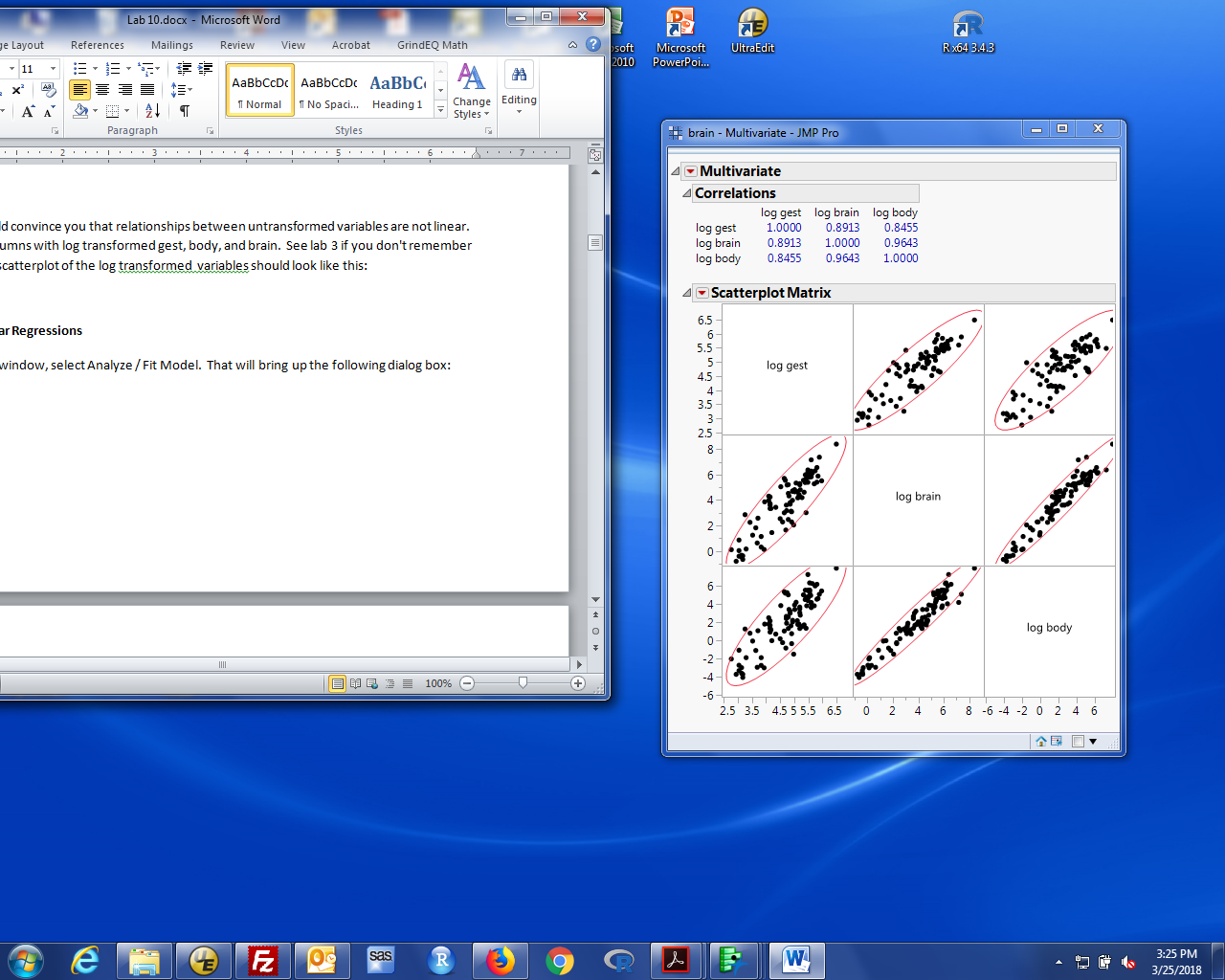
unchecking Density ellipses.

There are a variety of decorations that you can add to each plot by checking a box for what you want to add.

1. If you want a confidence interval for the population correlation or a test of population correlation = 0, click the red triangle by Multivariate and select Pairwise Correlations. We haven't discussed (and won't) the theory behind inference on a correlation coefficient. Ask if you don’t understand any of the results.

**Data preparation:**

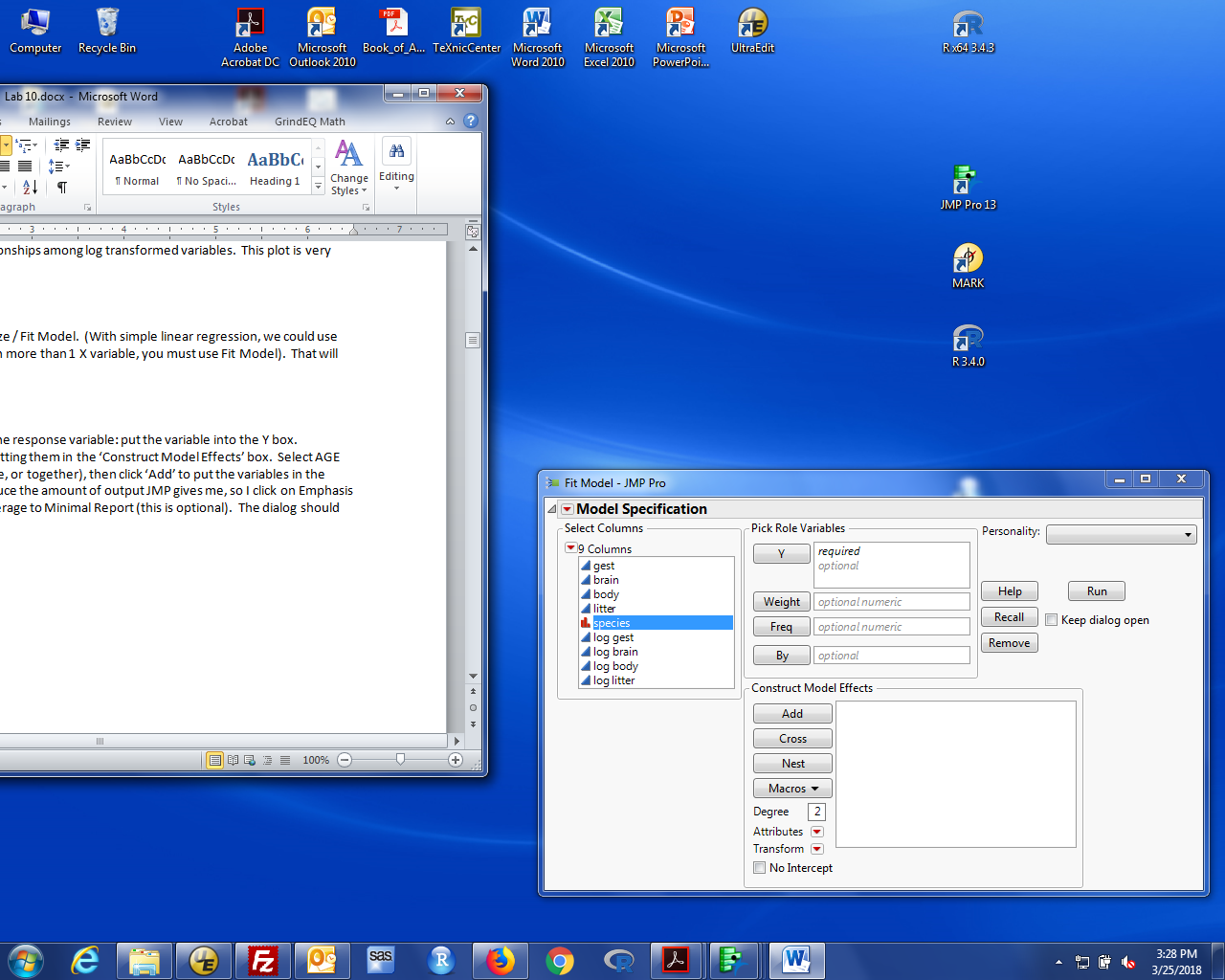
The scatterplot should convince you that relationships between untransformed variables are not linear. Create three new columns with log transformed gest, body, and brain. See lab 3 if you don't remember how to do this. The scatterplot of the log transformed variables should look like this (see next page):



From here on, we will assume linear relationships among log transformed variables. This plot is very much consistent with that assumption.

**Fitting Multiple Linear Regressions**

1. In the home window, select Analyze / Fit Model. (With simple linear regression, we could use either Fit Y by X or Fit Model. With more than 1 X variable, you must use Fit Model). That will bring up the following dialog box:

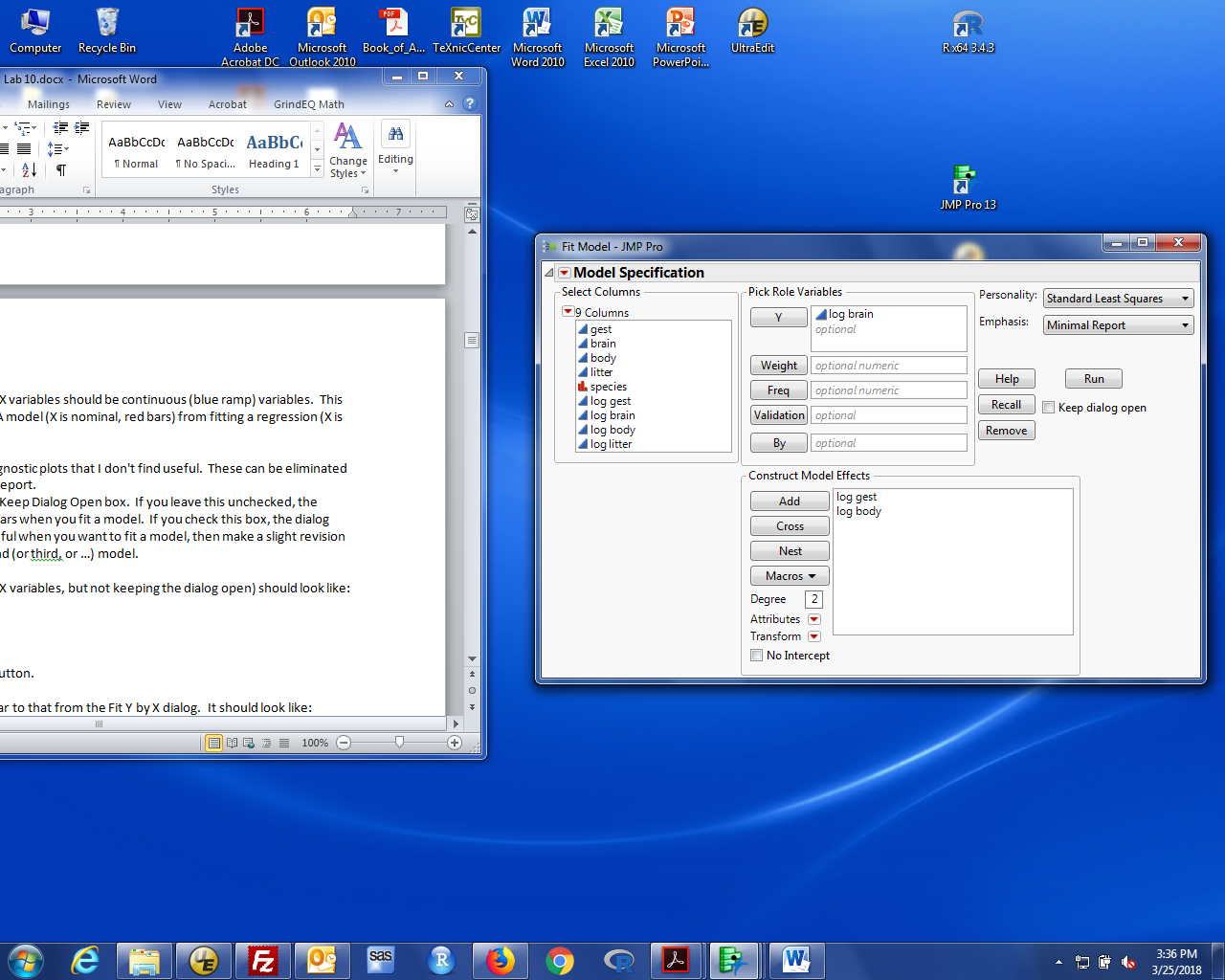


1. We want to use log gest and log body to predict log brain. It should be clear how to specify the response variable: put the log brain into the Y box.
2. The X variables are specified by putting them in the ‘Construct Model Effects’ box. Select log gest and log body (either one at a time, or together), then click ‘Add’ to put the variables in the Model Effects box.
3. If you want to fit a regression, the X variables should be continuous (blue ramp) variables. This is what separates fitting an ANOVA model (X is nominal, red bars) from fitting a regression (X is continuous, blue ramp).
4. Options in the Fit model dialog:

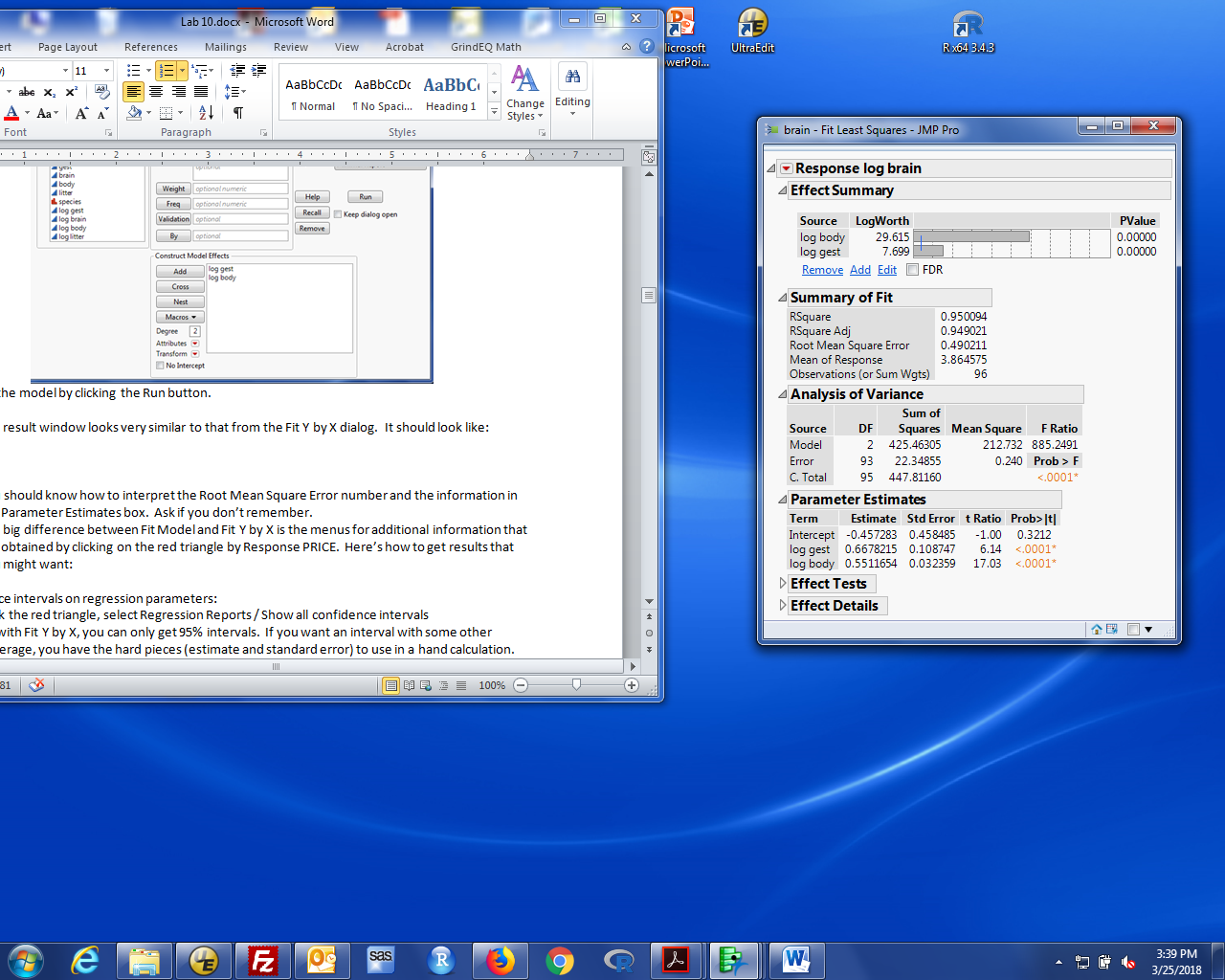
The default is to provide some diagnostic plots that I don't find useful. These can be eliminated by changing Emphasis to Minimal report.

Another useful modification is the Keep Dialog Open box. If you leave this unchecked, the model specification dialog disappears when you fit a model. If you check this box, the dialog stays around. This can be very useful when you want to fit a model, then make a slight revision (e.g. add a variable) and fit a second (or third, or …) model.

The dialog (with minimal report, 2 X variables, but not keeping the dialog open) should look like:



1. Fit the model by clicking the Run button.
2. The result window looks very similar to that from the Fit Y by X dialog. It should look like:



1. You should know how to interpret the Root Mean Square Error number and the information in the Parameter Estimates box. Ask if you don’t remember.
2. The Analysis of Variance box contains the model comparison between an intercept only model (all slopes = 0, reduced model), and the specified model (all slopes potentially not 0, full model).
3. The big difference between Fit Model and Fit Y by X is the menus for additional information that are obtained by clicking on the red triangle by Response log brain. Here’s how to get results that you might want:

Confidence intervals on regression parameters:

Click the red triangle, select Regression Reports / Show all confidence intervals

As with Fit Y by X, you can only get 95% intervals. If you want an interval with some other coverage, you have the hard pieces (estimate and standard error) to use in a hand calculation.

Predicted values for new observations:

Click the red triangle, select Save Columns / Prediction Formula

The data table will have a new column, labelled Pred Formula. You get predictions for new observations by entering the X values in the data table. (Remember, if log brain is predicted from log gest and log body, you need to enter values for both variables).

Note: If you choose Predicted Values instead of Predicted Formula, you get the values for all observations in the data table, but the predicted values are not computed when you enter new observations.

CI’s or PI’s for new observations:

Click the red triangle, select Save Columns / Mean Confidence Interval Formula for CI’s, or

Save Columns / Indiv Confidence Interval Formula for PI’s.

Note: Again, there are two related options. The Formula versions update the computed values when you add new observations to the data table. The non-Formula versions compute values for what is in the table and nothing more.

**Diagnostics:**

All are obtained using the menu under the red triangle by the Y variable (i.e., continuing the above list of additional information).

Plot of residuals vs predicted values:

Click the red triangle, select Row Diagnostics / Plot Residual by Predicted. You can also save predicted values and save residuals, then plot those two variables. Remember the new variables are provided as new columns in the data set.

Normal Quantile Plots: Not shown automatically (unlike Fit Y by X). Here’s how to get a plot:

Click the red triangle, select Save Columns / Residuals, then Analyze / Distribution, looking at the residual column. Click the red triangle by the variable name (Residual PRICE, not Distributions) and click Normal Quantile plot.

The dots are the data, the straight line is what is expected if the errors have a normal distribution. The curved lines are approximate prediction intervals if the errors are normal. I haven’t talked about computing these (and won’t). You should be more concerned about lack of normality if the data fall outside the prediction intervals.

**Partial (Type III) SS:**

The minimal report hides the type III SS. To see them, click the grey triangle by Effect Tests. These are the type III SS and associated F tests. These F statistics are the square of the coefficient-specific T tests.

**Sequential SS**:

Analyze / Fit Model reports the partial SS (type III) for each term in the model by default. If you really want the sequential SS (de-emphasized in this year's 401), fit the model, click the red triangle by price, and select the Estimates tab. Select Sequential Tests. The sequential F tests are added to the output for the Fit Model; the box of output is labelled with the words sequential (Type I).

**Interaction or quadratic terms in a multiple regression model:**

There are two ways to fit a model with an interaction or quadratic term:

create a variable with the product or square of the appropriate X variable(s) by hand

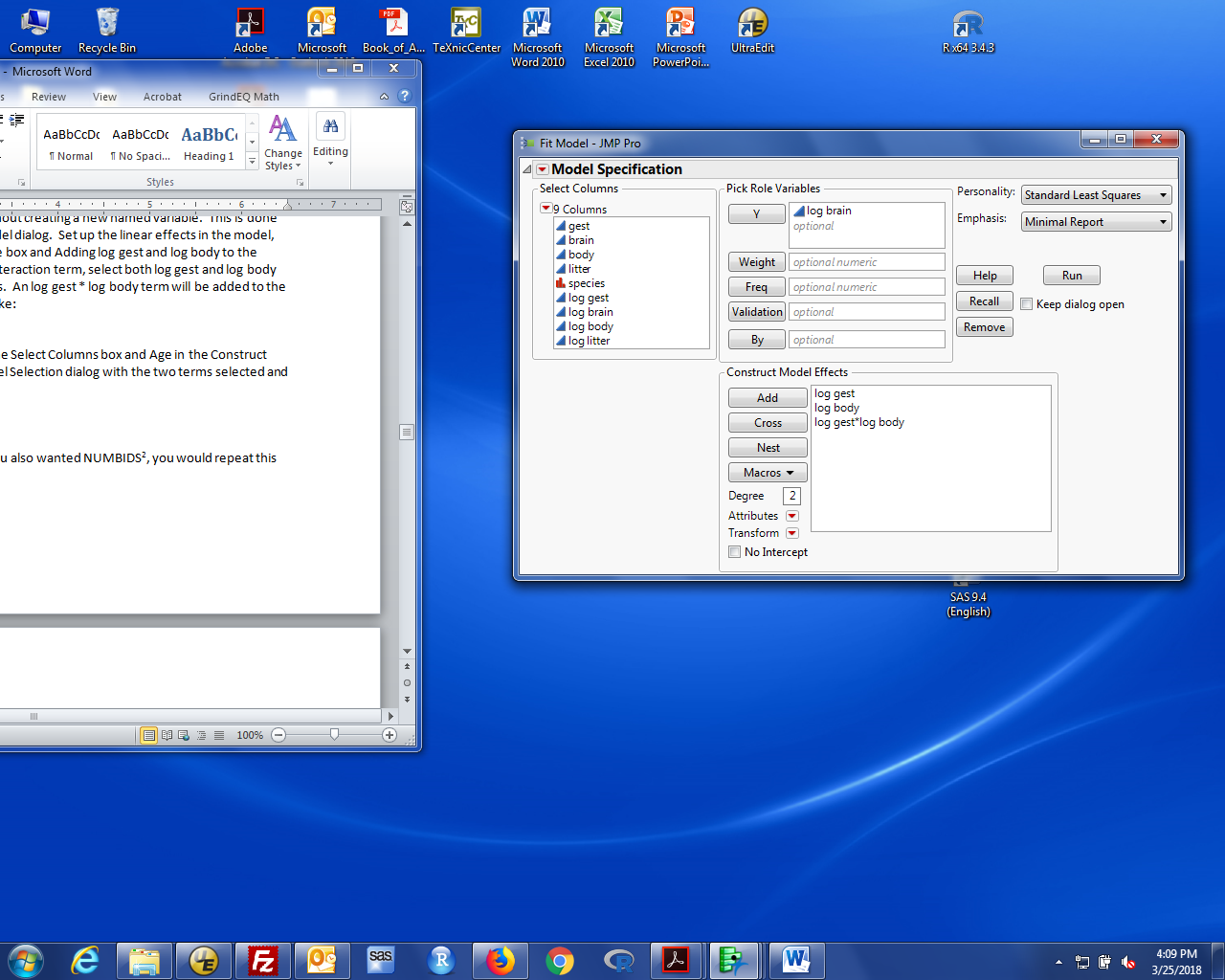
have JMP create that product or square “on the fly”.

Both give the same results, so long as “on the fly” is done with some care. Here’s how to do each:

1. **Creating a new X variable**: This should be familiar by now. Select the data set to make it the active window. Right click on a blank column and select column info. Type in an appropriate column name, then select column properties / edit formula. Use the buttons to write the desired formula, log gest x log body. Then click the OK buttons. You will get a new column with the product. Use that product variable in a multiple regression model in the obvious way: just add it to the Construct Model Effects box in the Analyze/Fit Model dialog.

Quadratic variables (e.g. log gest2) are constructed by log gest x log gest or using the power button (xy), which gives you squares (y=2) by default.

1. **Creating an interaction or quadratic term “on the fly”:** The Fit Model dialog allows you to use “created” variables in a regression model without creating a new named variable. This is done with the **Cross** button in the Analyze / Fit Model dialog. Set up the linear effects in the model, as before by putting log brain in the Y variable box and Adding log gest and log body to the Construct Model Effects box. To include an interaction term, select both log gest and log body in the Select Columns box, then left-click Cross. An log gest \* log body term will be added to the Construct Model Effects box, which will now look like:



To construct a quadratic term, select log gest in the Select Columns box and log gest in the Construct Model effects box, then click Cross.

NOTE: yes, you have to click log gest in **both** places to generate the quadratic term.

If you also wanted log body2, you would repeat this with log body.

**IMPORTANT NOTE:** To recreate class results (and to match results from adding variables), you need to do one more thing anytime you create variables “on the fly”: **Click the red triangle by Model Specification and unclick Center Polynomials.**

We will talk in lecture later about centering polynomials. For now, it is a distraction. You don't have to worry about this if you create the product (or square) term(s) yourself.

If, after fitting a model, the parameter estimates box includes terms like: (log gest-4.71324)\*(log body-2.1305), you didn’t turn off polynomial centering. Your results will not match mine. Recreate and rerun the model with Center Polynomials unclicked.