STAT E-150 - Statistical Methods
Introduction to SPSS

How to access SPSS from Harvard Computer Services

Before you begin to install SPSS from Harvard Computer Services:

- You must have a Harvard ID (HUID) and PIN to activate an FAS account or to access to most online applications. You can find your HUID on your Extension School registration confirmation statement and/or bill. (Note: this is *not* the same as your DCEID which begins with "@")
- If you do not have a PIN, you can request one online at www.pin.harvard.edu. Your PIN will be emailed to the email address you provided when you registered with the Extension School.
- You must email HUIT Support at help@fas.harvard.edu to request an Authorization Code. The Code will be sent to you upon resolution of your support request, and then you can follow the instructions below.

Once you have your HUID, PIN, and authorization code, go to: www.fas.harvard.edu/computing

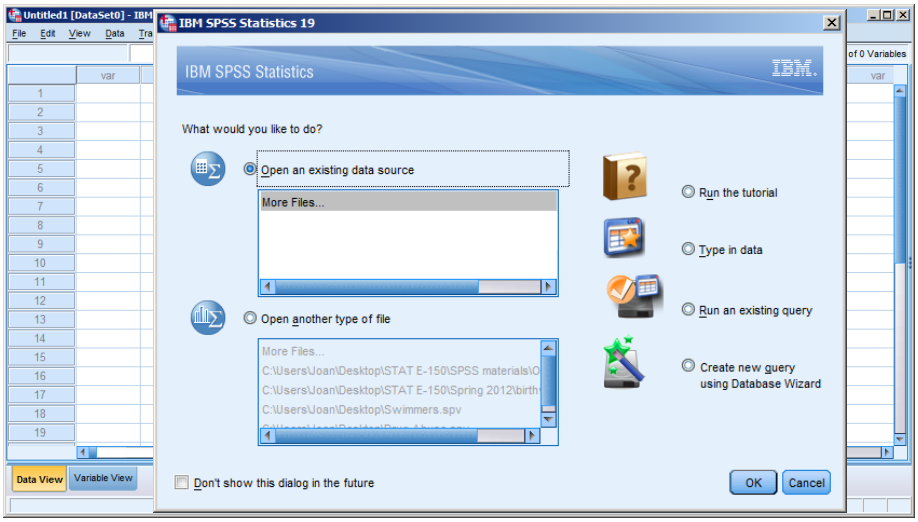
1. Click on "Other Students" under "New to FAS?"
You will use your Harvard ID and PIN in order to create an email account.

2. Once you have an account, click on Software Downloads at the top of the page.
This will lead you to a page where you login using the ID and PIN discussed above.

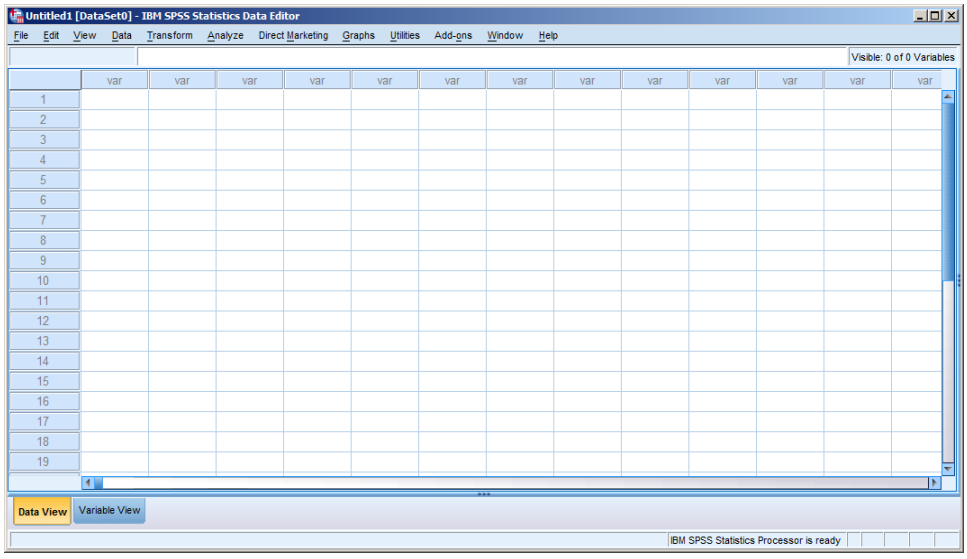
3. After you log in, you will be directed to a list of software downloads and you can download SPSS.

Entering and Saving Data

When you open SPSS your screen should look like this, with a menu dialog box in the front and the Data Editor behind it. The Data Editor looks like a spreadsheet, and that is where you can define variables and enter data.



Click on **Cancel** to access the Data Editor window. Each column represents a variable, and each row represents an observation.

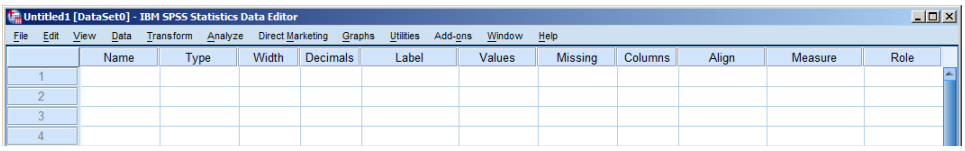


Data can be entered by opening a data file, but we will start by typing data into the Data Editor. First click on the Variable View tab at the bottom of the window.

Here is the data we have been discussing in class:

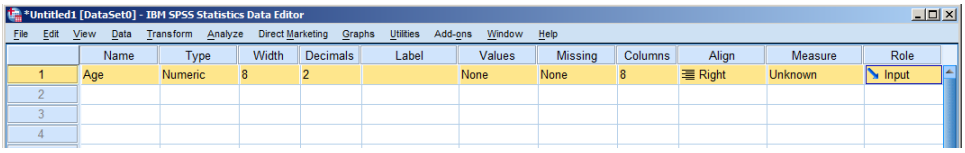
Observation	1	2	3	4	5	6	7	8	9	10
Maternal Age	15	17	18	15	16	19	17	16	18	19
Birthweight (g)	2289	3393	3271	2648	2897	3327	2970	2535	3138	3573

Define the variables by entering a name, the type of data, the number of decimal places, etc., for each variable. You can see the column headings at the top:

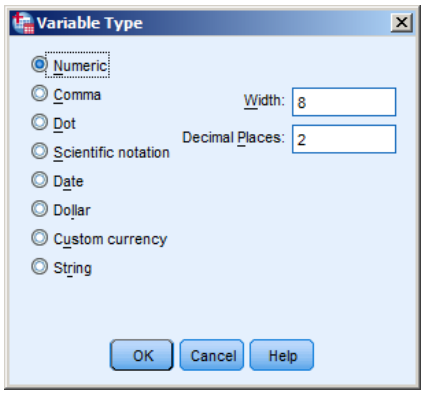


1. Go to the first empty cell in Row 1, under Name, and type in the first variable name. We will use the variable name **Age**.

2. Press **Enter** or **Tab**.

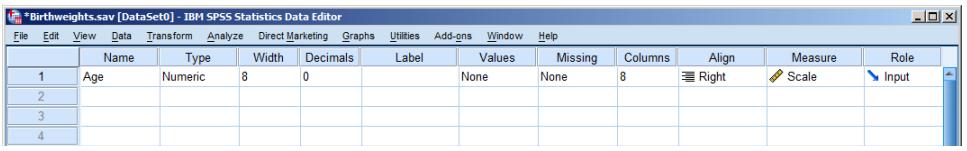


3. Click in the **Type** column, and a small button with three dots will appear. Click on this button to see this dialog box:



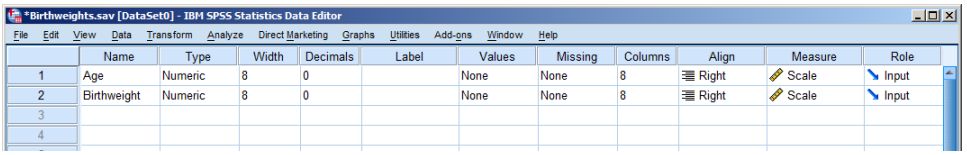
Note that the default variable type is Numeric, which is appropriate for the data we are using. Set the number of decimal places to 0.

4. Set the Measurement Level of these two variables to **Scale.**



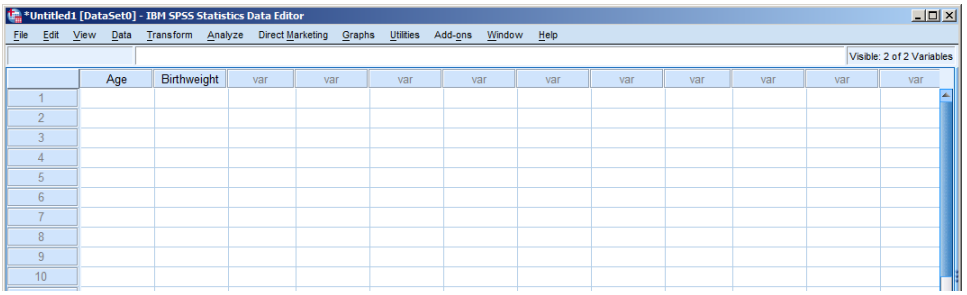
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Age	Numeric	8	0		None	None	8	Right	Scale	Input
2											
3											
4											

5. Repeat these steps to create the **Birthweight** variable.



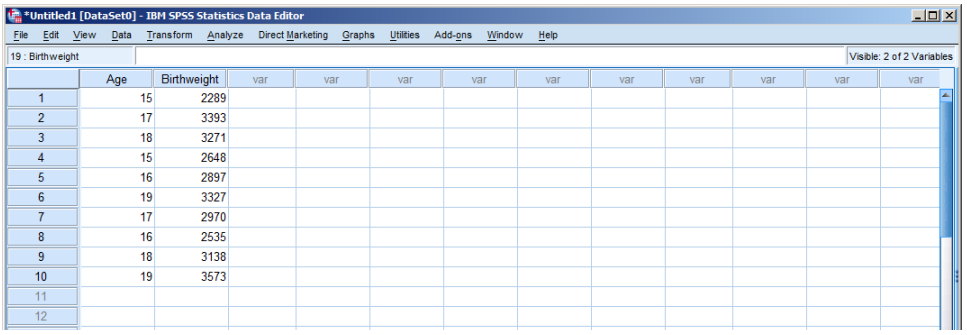
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Age	Numeric	8	0		None	None	8	Right	Scale	Input
2	Birthweight	Numeric	8	0		None	None	8	Right	Scale	Input
3											
4											

6. Switch to the Data View to enter the data:



	Age	Birthweight	var	var	var	var	var	var	var	var	var	var
1												
2												
3												
4												
5												
6												
7												
8												
9												
10												

Your window should look like this:



	Age	Birthweight	var	var	var	var	var	var	var	var	var	var
1	15	2289										
2	17	3393										
3	18	3271										
4	15	2648										
5	16	2897										
6	19	3327										
7	17	2970										
8	16	2535										
9	18	3138										
10	19	3573										
11												
12												

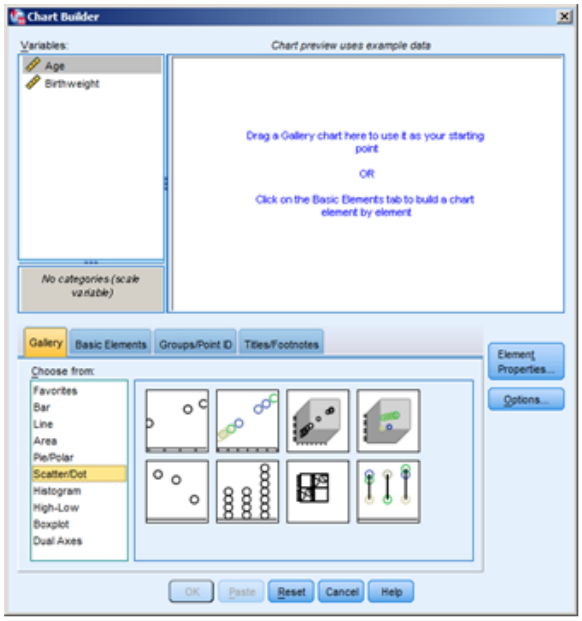
7. To save your file, go to **File > Save As…** and indicate where you would like to save your work. Be sure to use an appropriate filename. Click **Save** when you have entered all of the necessary information.

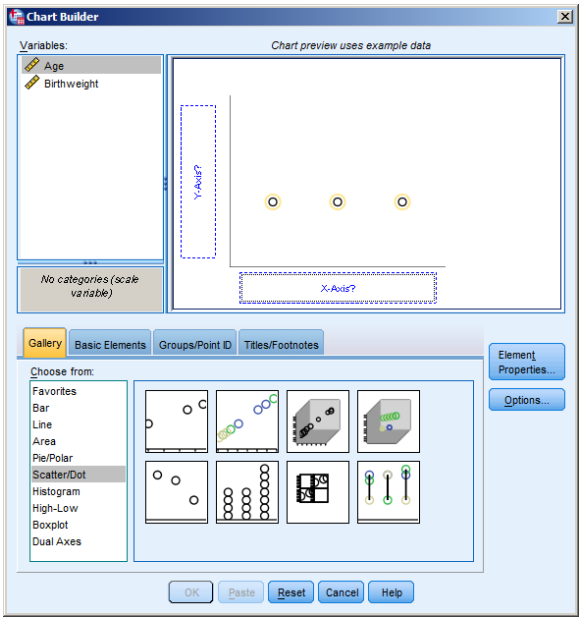
You will see an Output Viewer window that confirms that your file has been saved.

Creating a Scatterplot

1. Click on **Graphs > Chart Builder**.

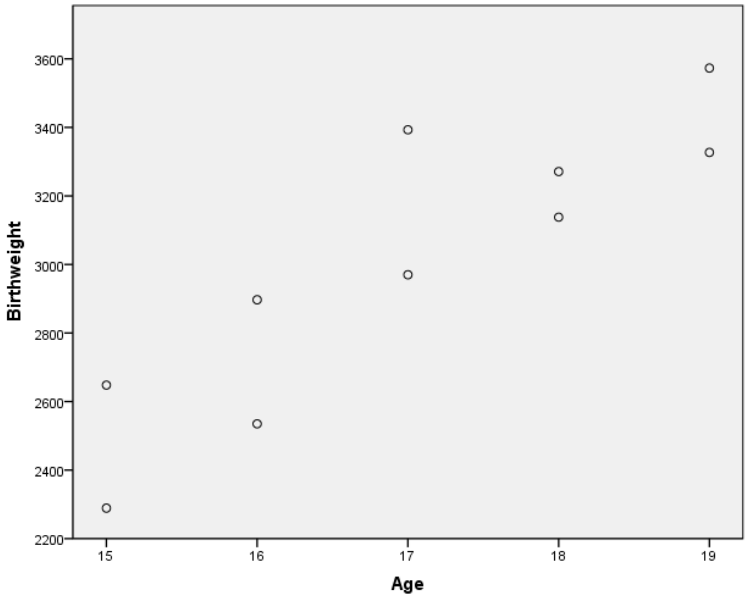
   Select **Scatter/Dot** from the list of charts.
   Drag the first Scatterplot option to the window.





   Drag the explanatory variable, Age, to the x-axis, and drag the response variable, Birthweight, to the y-axis. You will see a sample scatterplot; this is not a display of the data that was entered.

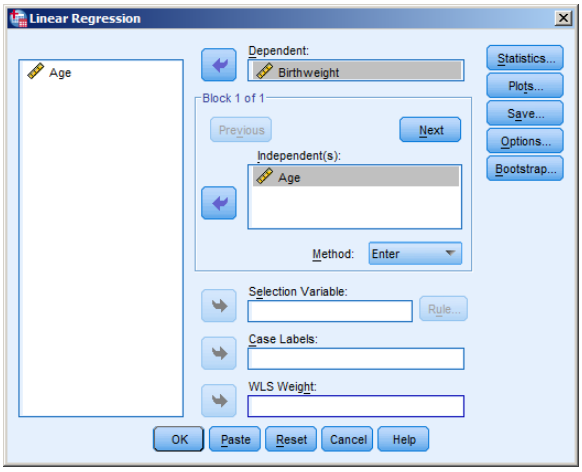
   Click on OK. The scatterplot of the data will appear in the **Output Viewer**.



Finding the Linear Regression Equation

1. Click on **Analyze > Regression > Linear**

   Drag the variables to the appropriate locations and then click on **OK**



To see the output, go to the **Output Viewer** which you can access by using the **Window** menu.

You will see several tables; these are the three we are interested in at this point:

The **Variables** table lists the independent, or explanatory variables in the model:

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Age[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Birthweight

The **Model Summary** table reports the value of the correlation coefficient and the coefficient of determination:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.884[a]	.781	.754	205.308

a. Predictors: (Constant), Age

The **Table of Coefficients** table reports the coefficients in the regression model, as well as t-statistics and significance levels for the regression inference:

Coefficients[a]

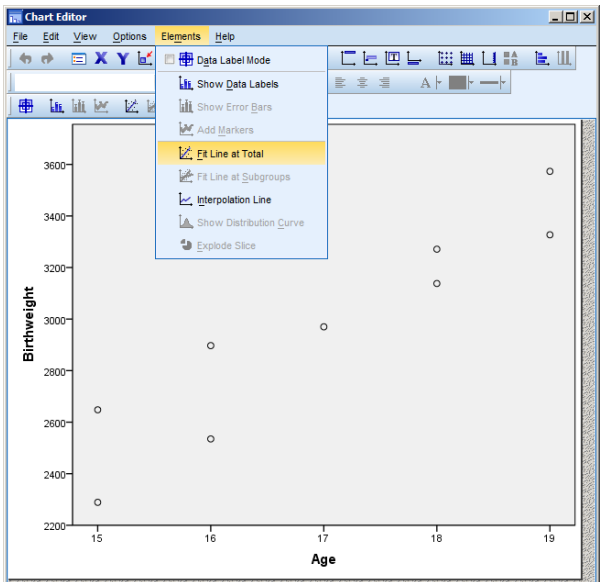
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1163.450	783.138		-1.486	.176
	Age	245.150	45.908	.884	5.340	.001

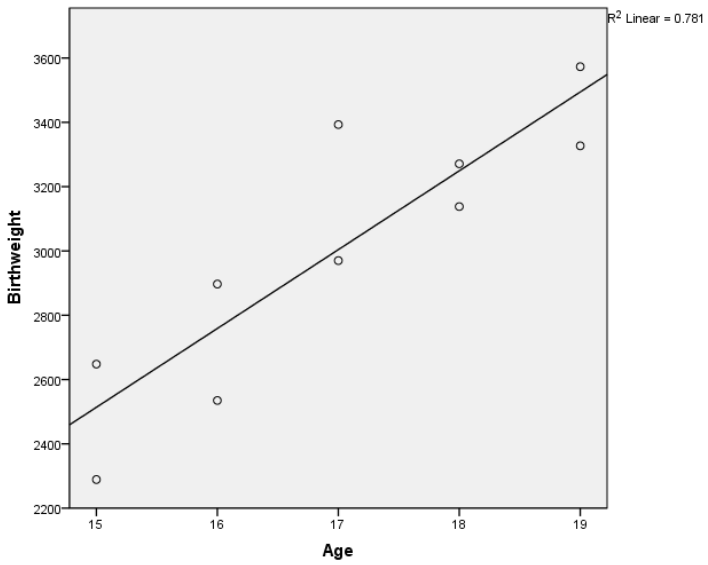
a. Dependent Variable: Birthweight

Displaying the Regression Line

1. Go to the Output viewer and double-click on the scatterplot you created. This will open the Chart Editor.

In the Chart Editor, click on > **Elements** > **Fit Line at Total** and then close the Chart Editor. In the upper right corner, you will see value of R^2



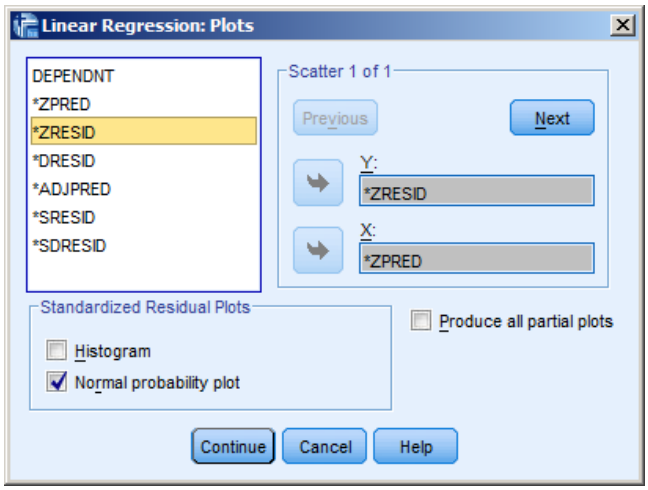


Examining Residuals

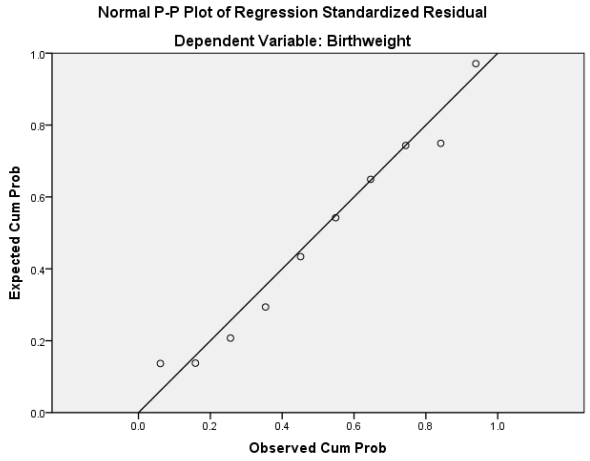
1. Create a scatterplot with the regression line displayed.

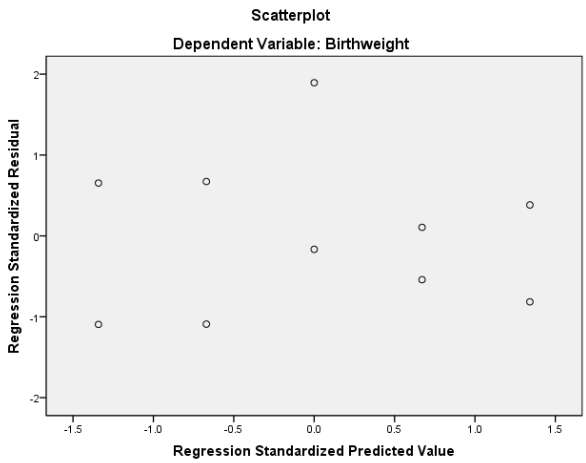
2. Click on **Analyze > Regression > Linear**

3. As before, choose the explanatory and response variables, and click on **Plots**. This will open a new dialog box which will allow you to choose the output you need.



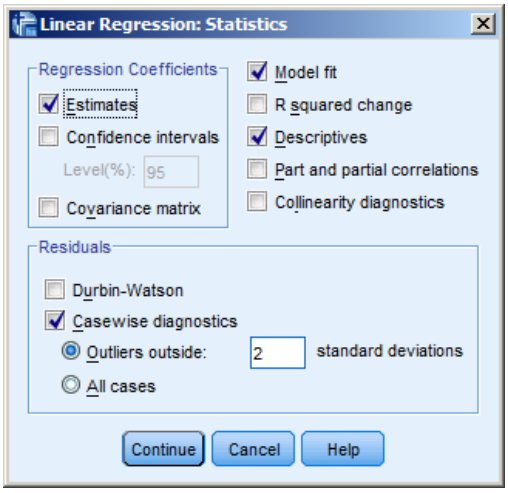
4. In the next dialog box, click on **Plots**, and make the selections shown above in the **Linear Regression Plots** dialog box. This will produce a Normal Probability Plot of the residuals and a scatterplot of the standardized predicted values (*ZPRED) vs. the standardized residuals (*ZRESID).





Getting confidence intervals:

Make the selections shown in the **Linear Regression Statistics** box, and then click on **Continue**:



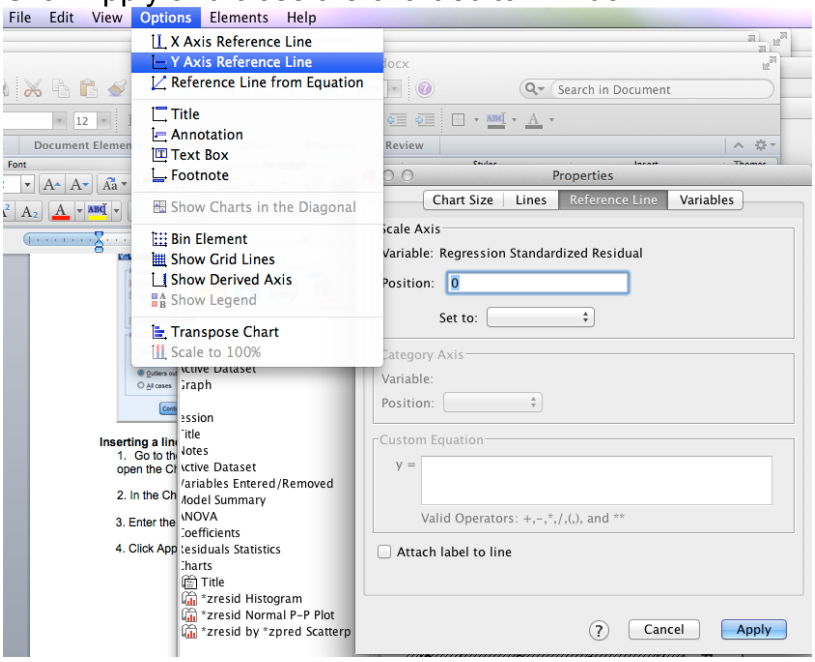
Inserting a line with a constant value, like at 0 for the residual graph.

1. Go to the Output viewer and double-click on the scatterplot you created. This will open the Chart Editor.

2. In the Chart Editor, click on > **Options** > **Y axis reference line**

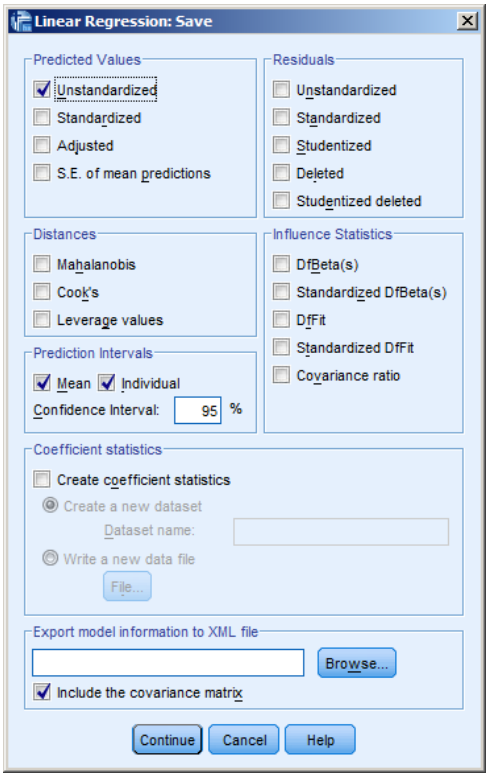
3. Enter the value you want the line at in the Position: box

4. Click Apply and close the chart editor window.

5. 

Making Predictions

Return to your data, and again select **> Analyze > Regression > Linear**
Click on **Save…** and make the selections shown below.

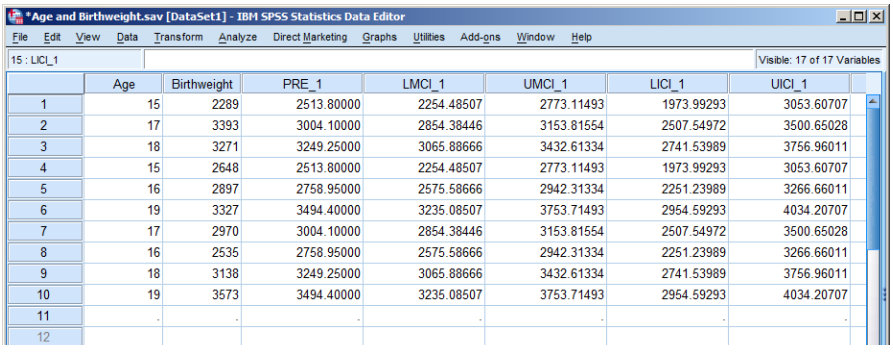


2. Click on **> Window** and select your dataset.
   You will see the table shown below.

   The **PRE_1** column shows the values predicted by the regression equation.

   The **LMCI_1 and UMCI_1** columns show the upper and lower bounds of the 95% *confidence* interval for the mean values of y.

   The **LICI_1 and UICI_1** columns show the upper and lower bounds of the 95% *prediction* interval for the individual values of y.



	Age	Birthweight	PRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1
1	15	2289	2513.80000	2254.48507	2773.11493	1973.99293	3053.60707
2	17	3393	3004.10000	2854.38446	3153.81554	2507.54972	3500.65028
3	18	3271	3249.25000	3065.88666	3432.61334	2741.53989	3756.96011
4	15	2648	2513.80000	2254.48507	2773.11493	1973.99293	3053.60707
5	16	2897	2758.95000	2575.58666	2942.31334	2251.23989	3266.66011
6	19	3327	3494.40000	3235.08507	3753.71493	2954.59293	4034.20707
7	17	2970	3004.10000	2854.38446	3153.81554	2507.54972	3500.65028
8	16	2535	2758.95000	2575.58666	2942.31334	2251.23989	3266.66011
9	18	3138	3249.25000	3065.88666	3432.61334	2741.53989	3756.96011
10	19	3573	3494.40000	3235.08507	3753.71493	2954.59293	4034.20707
11	.	.	.	.	.	.	.
12							

To make predictions for a value of the explanatory variable which is not in your dataset:

1. Click on **> Window** and select your dataset.

2. Enter the new value in the column for the explanatory variable. Here we will make predictions for a mother who is 20 years of age.

3. Click on **> Analyze > Regression > Linear**
   then click on **Save**

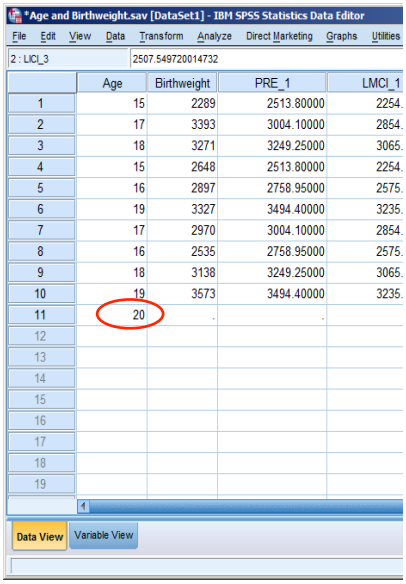
   You will see the same dialog box you used earlier, and will generate the same results for the new variable you entered.

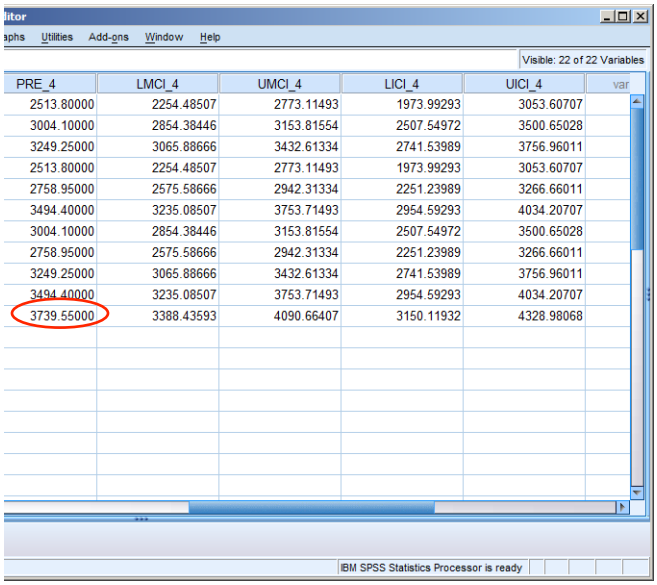
   Click on **Continue**

4. In the next dialog box, click on **OK**

5. Click on **Window** and select your dataset.
   At the left side of the spreadsheet you will see the value you entered.

6. Scroll to the right; in the row where you entered the value of the explanatory variable you will see the predicted value of the response variable as well as the confidence interval and the prediction interval for this value.

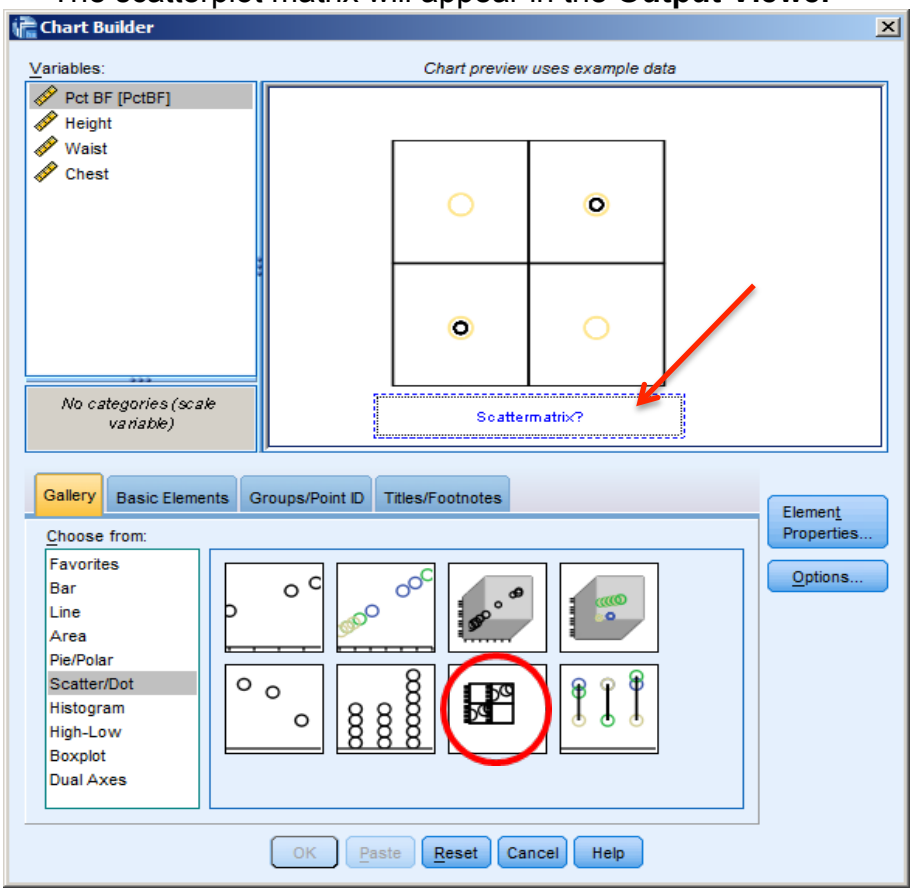




Create a scatter plot matrix to investigate multicolinearity

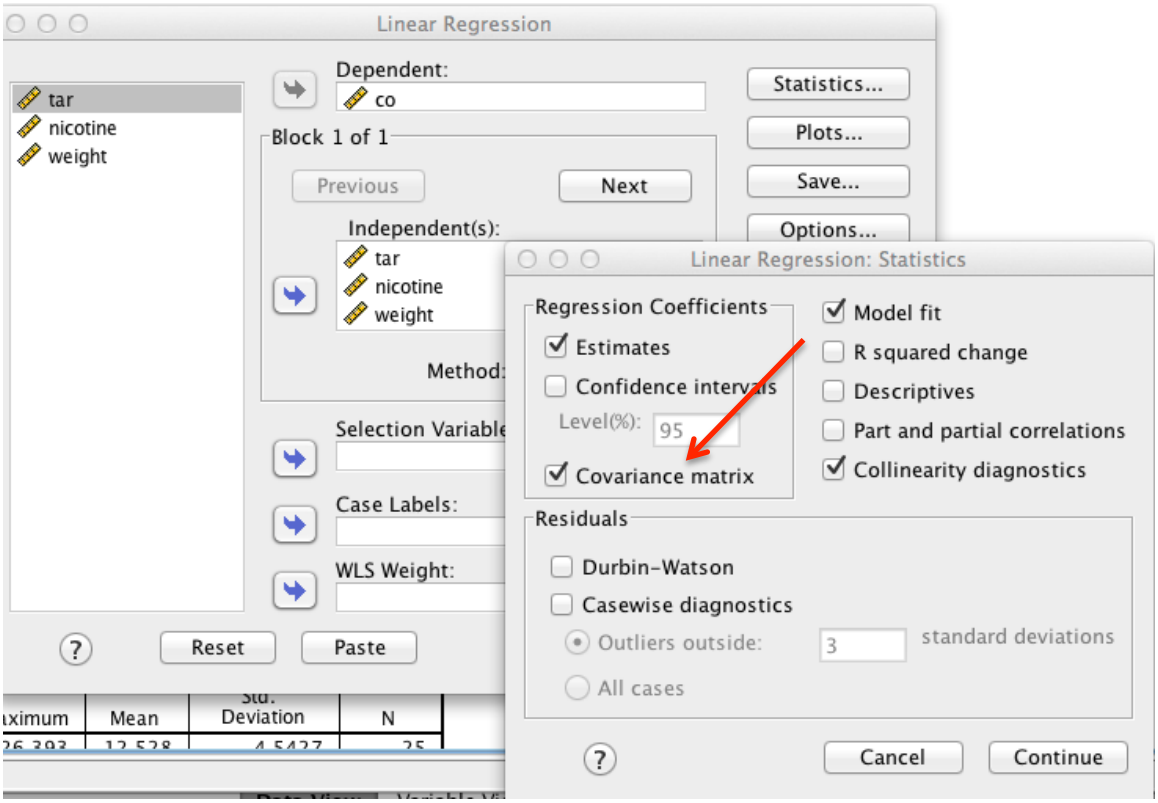
Creating a Scatterplot Matrix

Click on **Graphs > Chart Builder.**
Select **Scatter/Dot** from the list of charts.
Drag the **Scatterplot Matrix** to the window.
Drag the matrix variables to the horizontal axis.
Click on OK.
The scatterplot matrix will appear in the **Output Viewer**



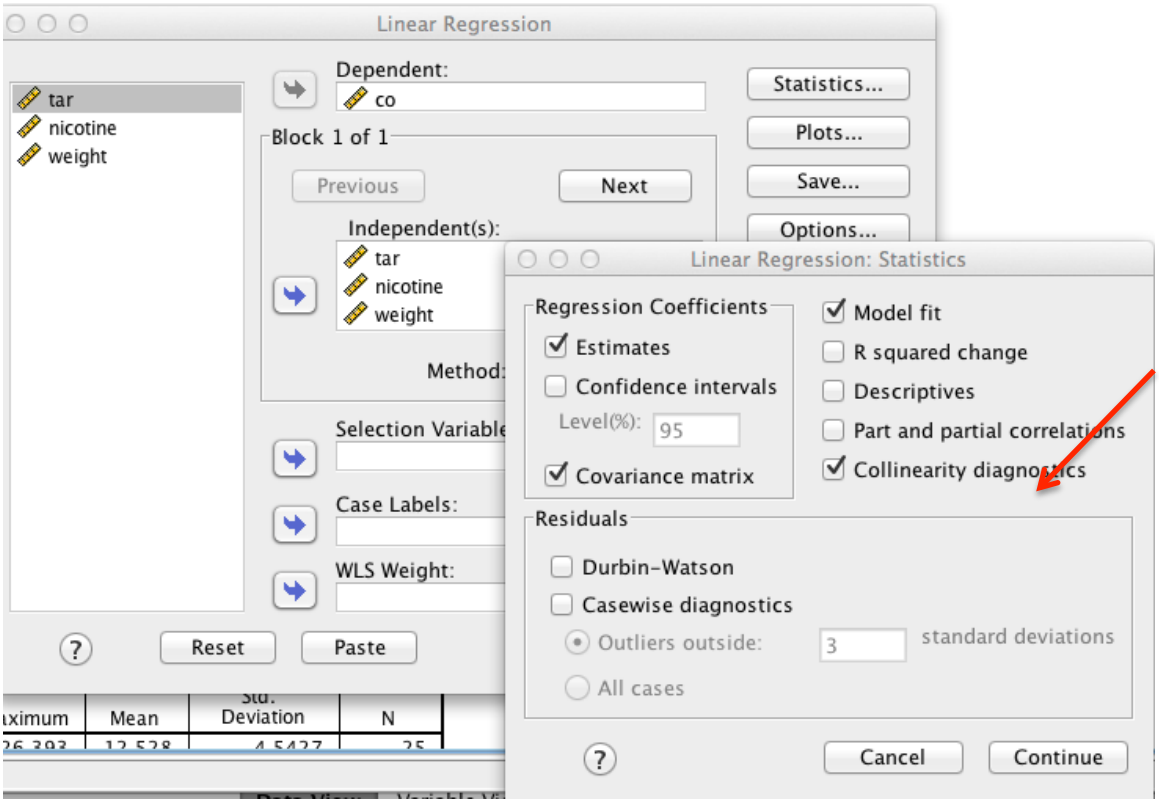
Create a correlation matrix, numbers to go with the graph above:

- Correlation matrix
- Run regression
- click on statistics button.
- Select correlation matrix
- Click continue



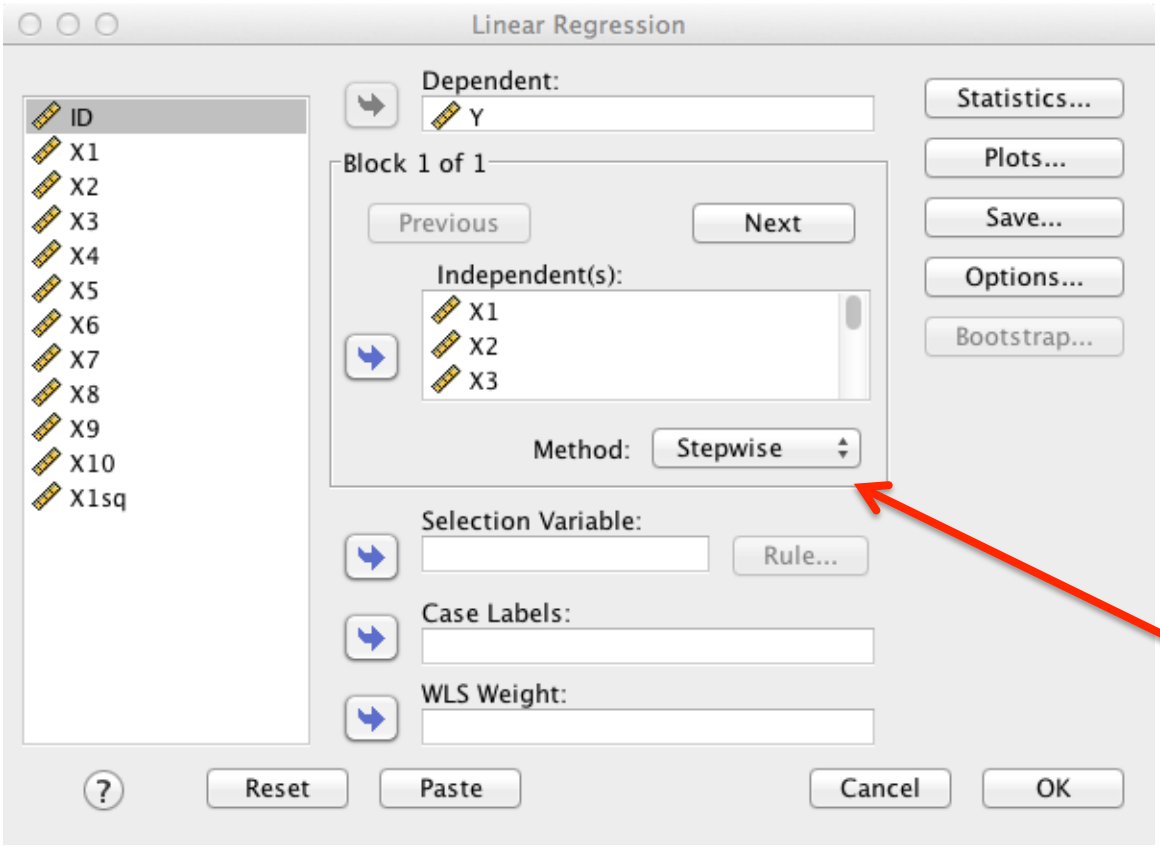
To get the variation inflation factor:

- A statistic to capture multicollinearity
- Under statistics tab, when you are doing your regression analysis



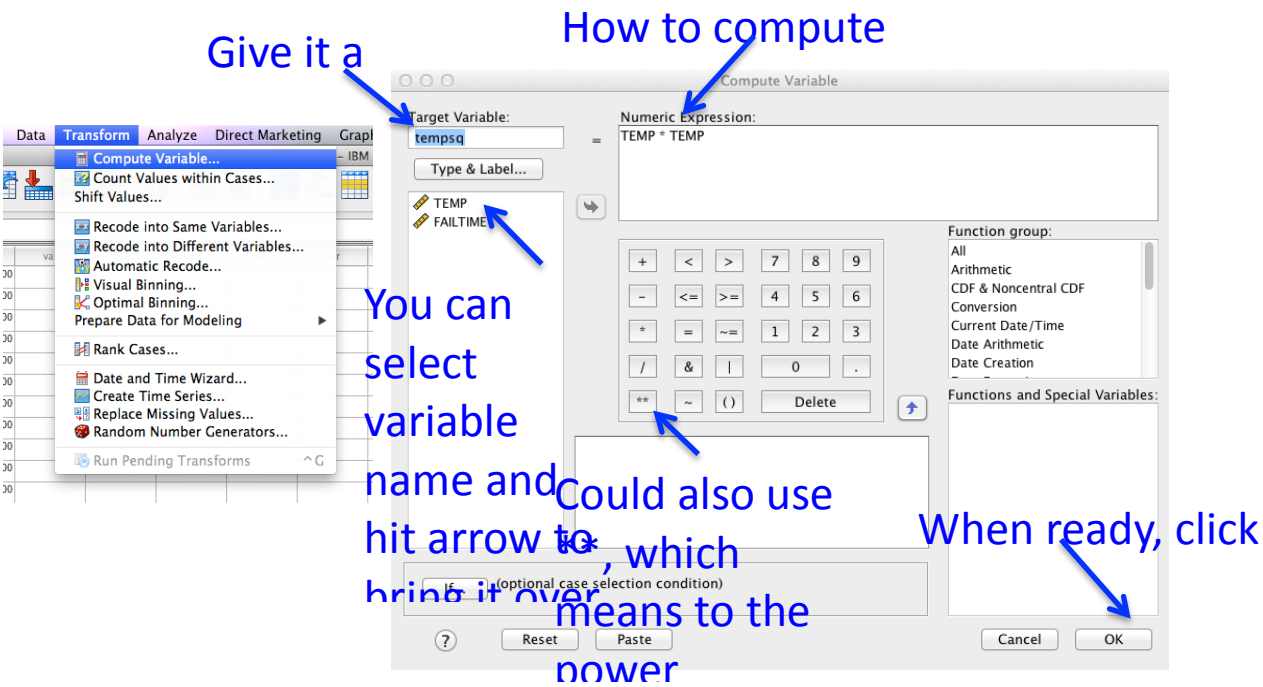
Stepwise regression

- Enter all of your variables, then choose enter method as stepwise



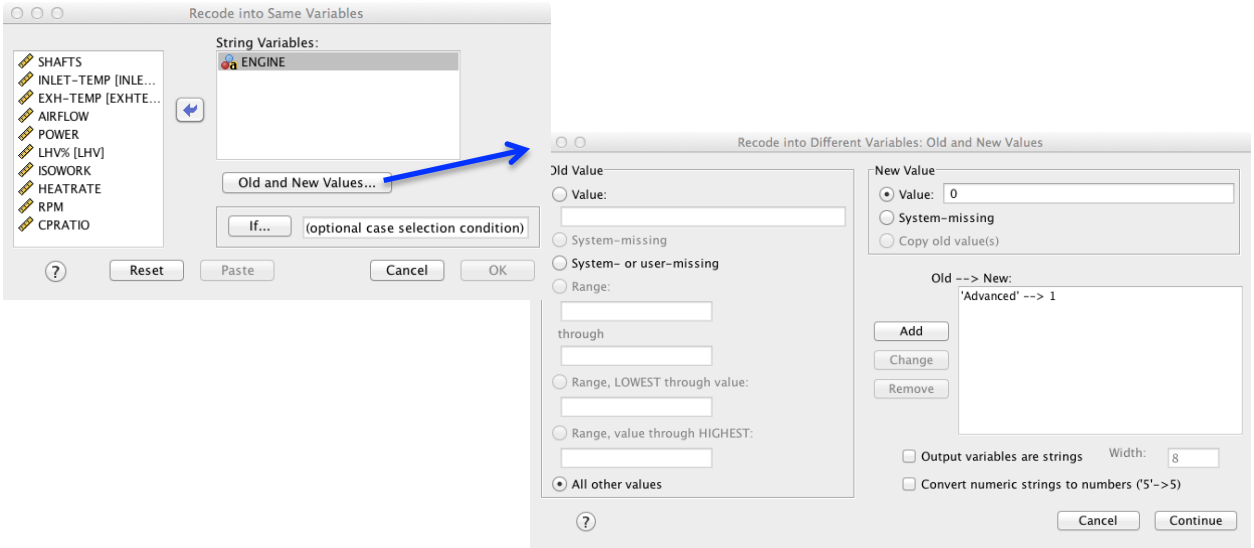
How to create a quadratic term:

Literally create X^2 or X * X:



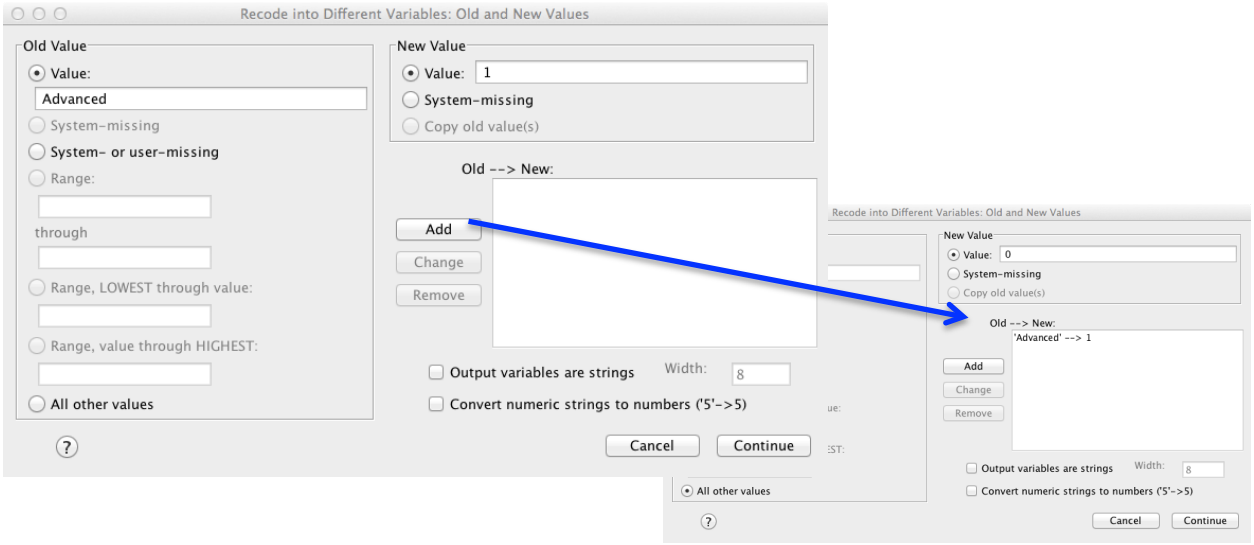
How to create dummy variables

- We have one variable called Engine, that has 3 levels, Traditional, Advanced, and Aeroderiv
- We want to create 2 variables
  - IsAdvan: 1 if advanced, 0 otherwise
  - IsAero: 1 if Aerodynamic, 0 otherwise
- Transform -> Recode into different variables
- Drag, or use arrow to move Engine over to String Variables
- Click on Old and New Values….

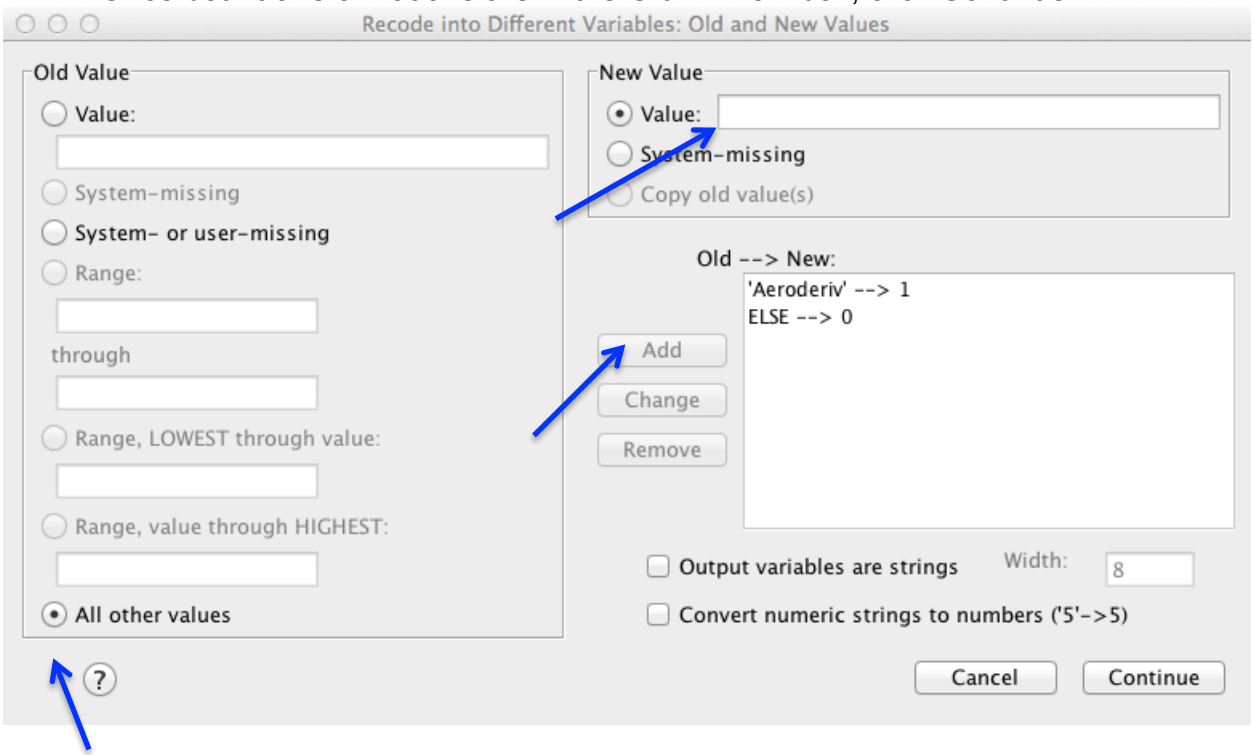


- Enter the value of Engine you want to recode in Old Value: Advanced

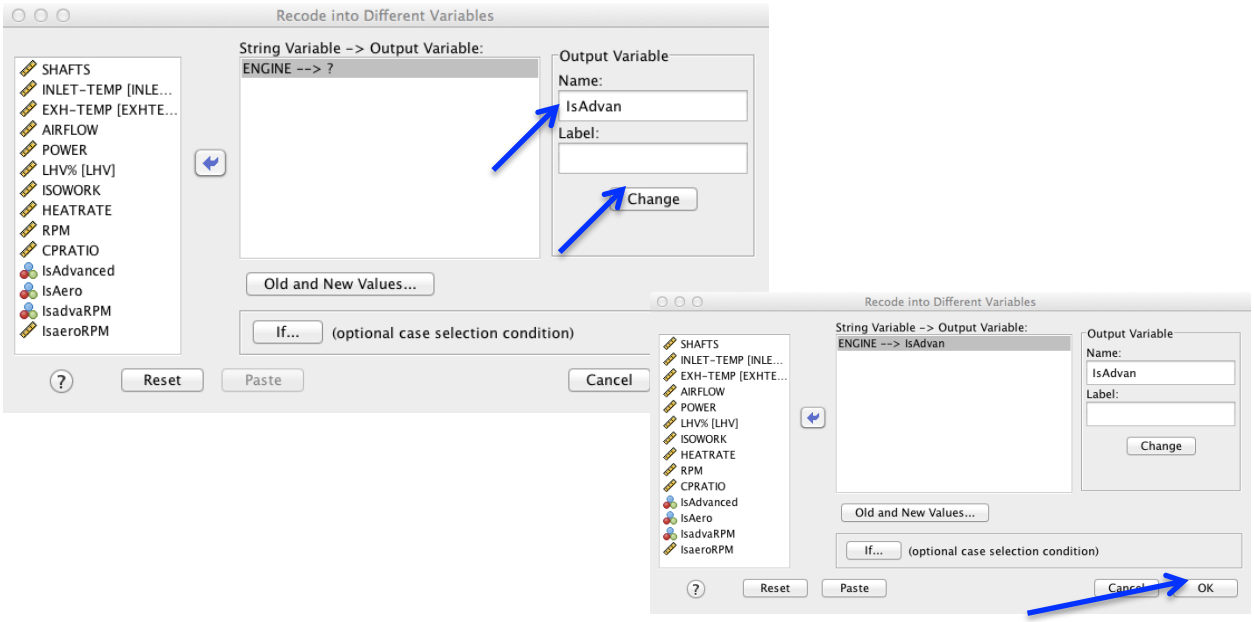
- Enter the New Value: 1
- Then clic Add, it moves to the box, old → new



- Next, you want to set all the other levels of Engine to 0.
- Select All other values
- Enter New Value as 0
- Click Add
- Once both transformations are in the Old → New box, click Continue



- Then enter a name into Name: IsAdvan
- Then click Change
- Then click OK



SPSS Instructions for ANOVA
To create side-by-side boxplots of the data:

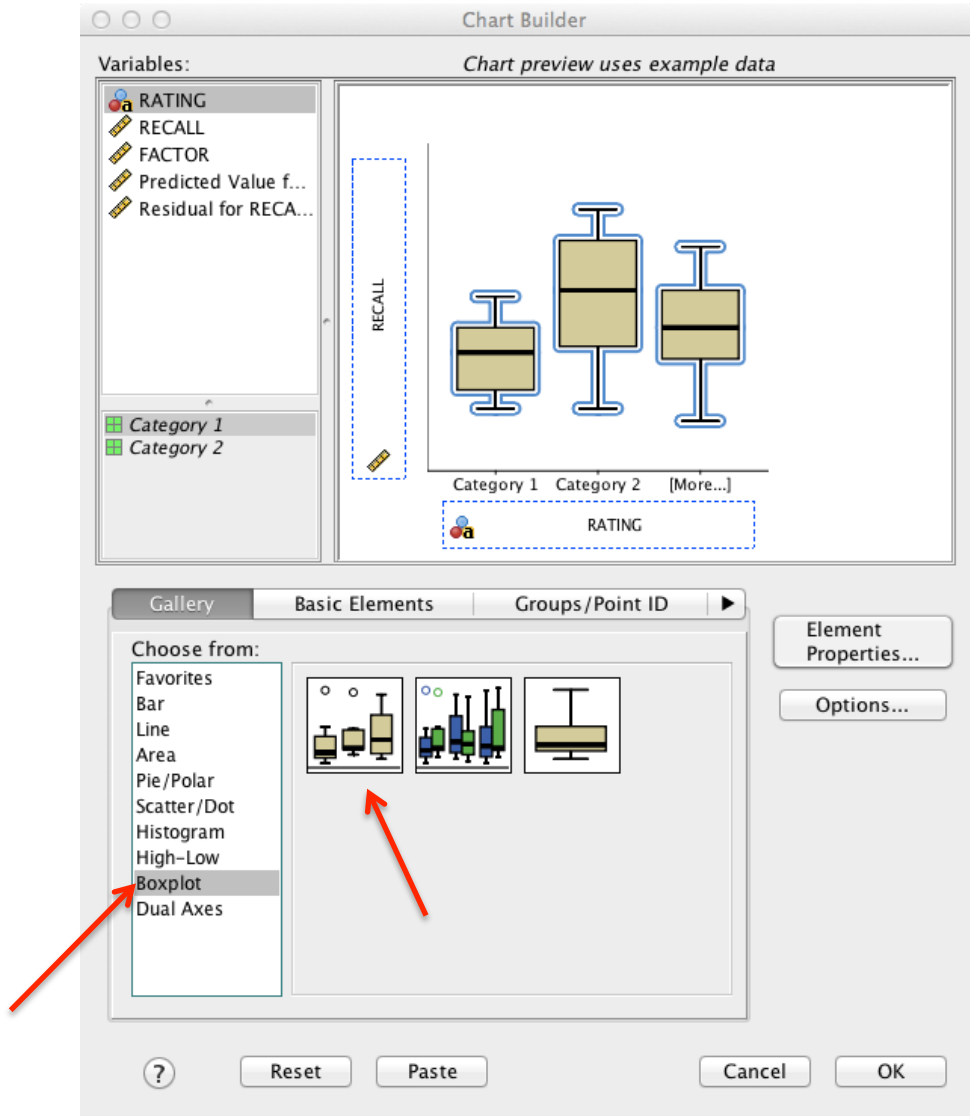
Assume that your file has the groups in one column and the values of the variable in a second column.

Choose **> Graphs > Chart Builder**

Choose **Boxplot** and drag the first boxplot (Simple) to the preview area.

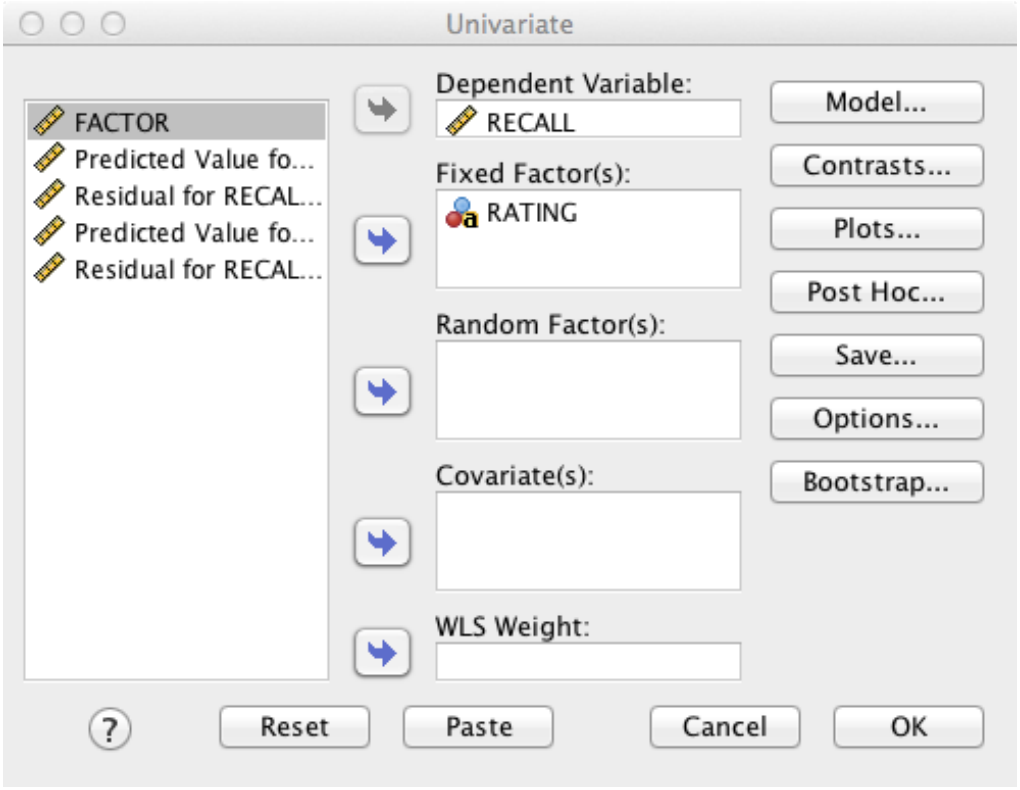
Drag the column with the groups to the x-axis, and the column with the values of the predictor variable to the y-axis.

Click **OK**.



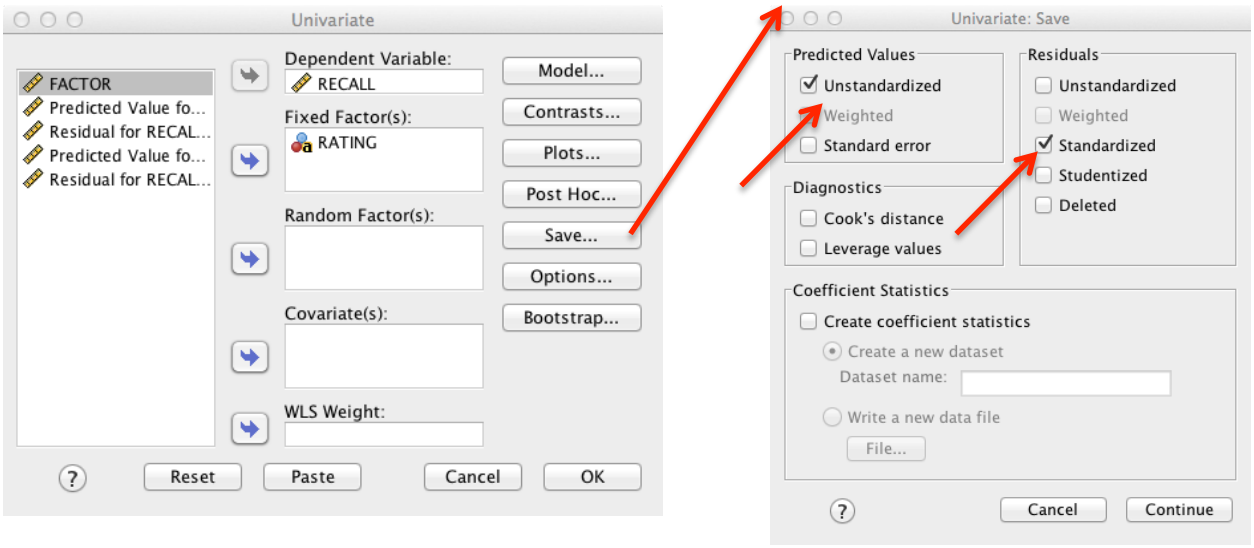
How to run an ANOVA

- **To perform a One-Way Analysis of Variance**
- Choose **> Analyze > General Linear Model > Univariate**
- Identify the response variable and move it to the **Dependent Variable** list. Select the variables that define the groups and move it to the **Fixed Factors** box.

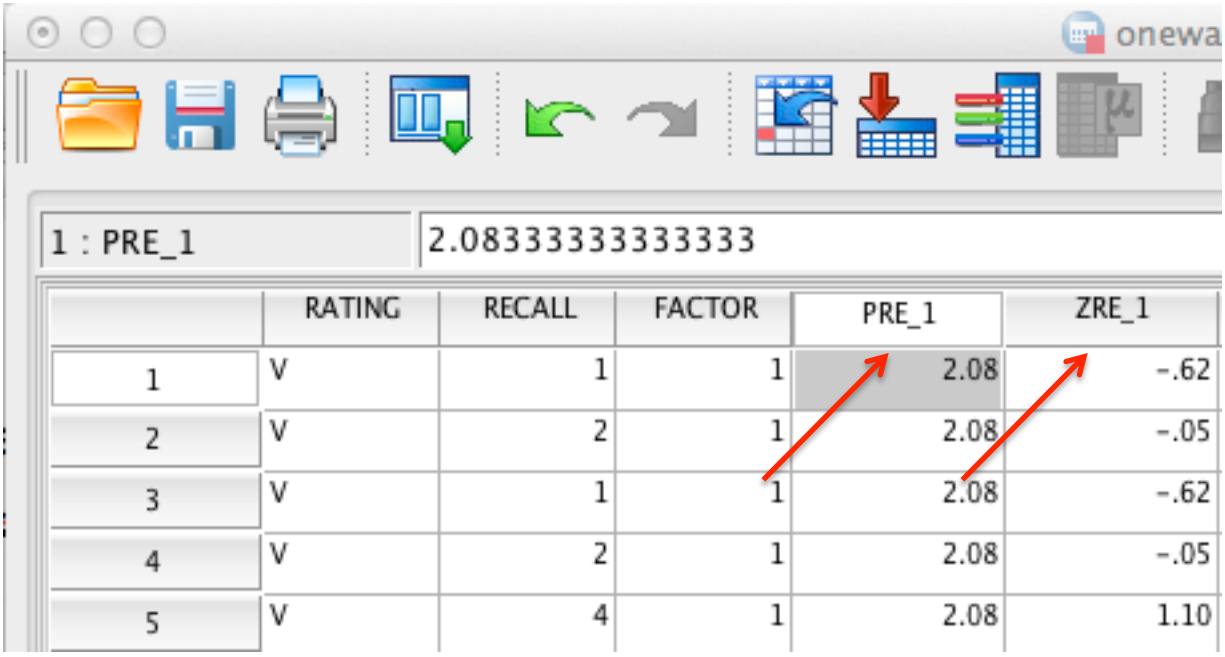


How to create a NPP and residual plot from ANOVA

- You need to save out the residuals and the predicted values.
- Click on save in the univariate window
- Then select standardized residuals and unstandardized predicted values

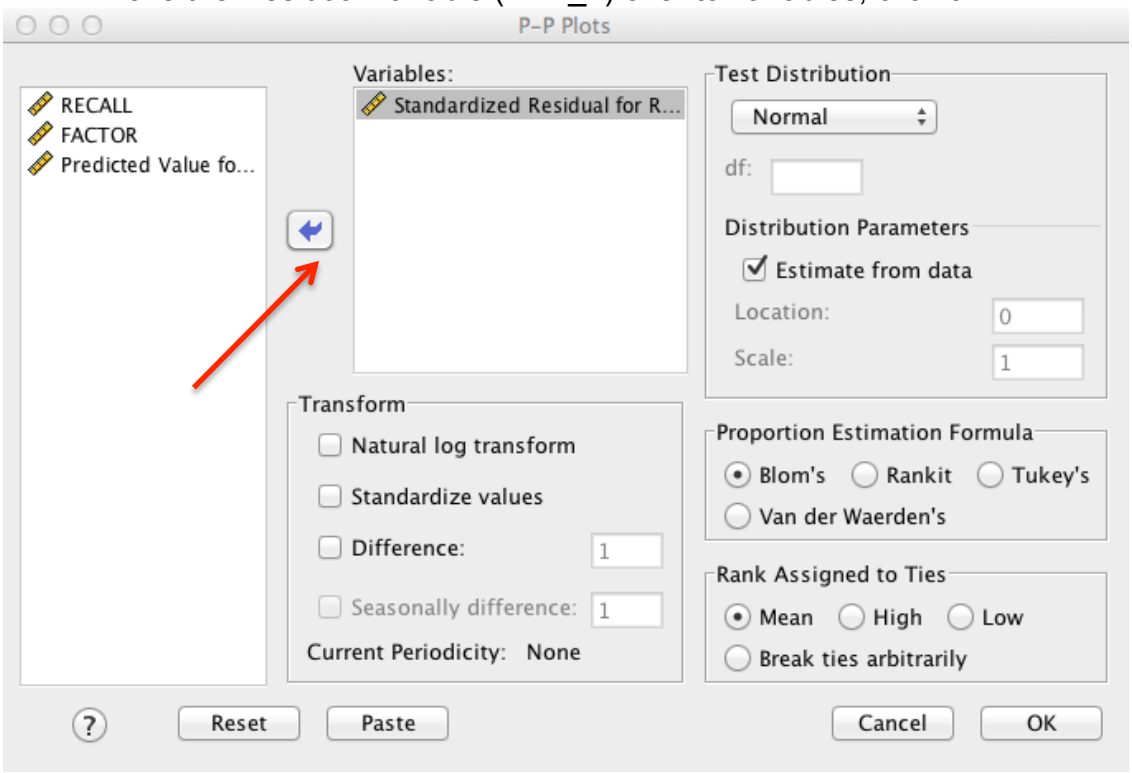


- This will create a new columns in your data
  - PRE_1, the fitted values
  - ZRE_1, the standardized residuals



How to create a NPP plot:

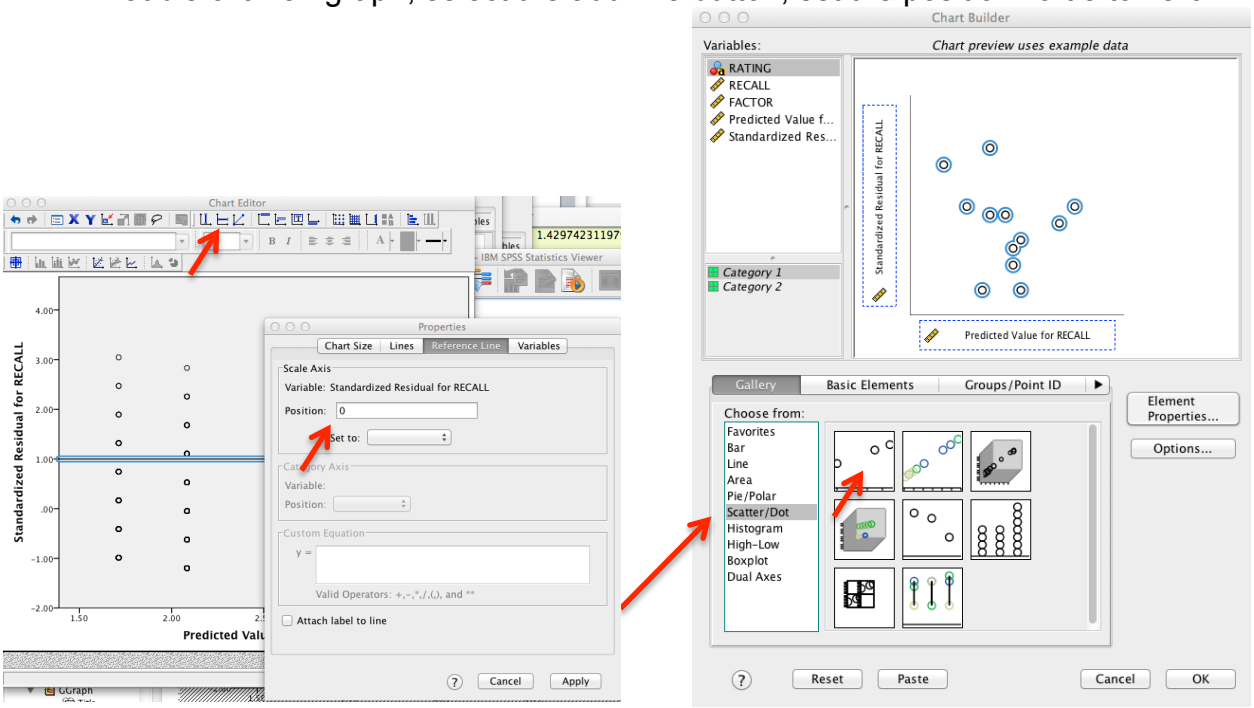
- Go to Analyze -> Descriptive -> P-P plots
- Move the Residual variable (ZRE_1) over to variables, click ok.



How to create a residual plot:

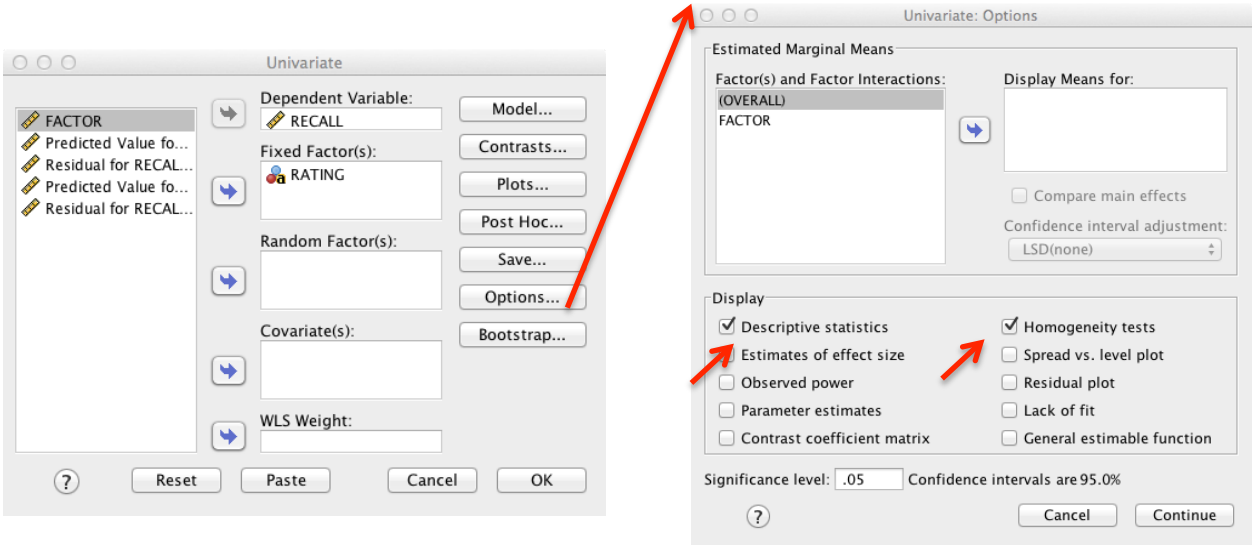
- You are going to make a scatter plot using the predicted values and the residuals
- Goto Graphs -> Chart builder
- Click on scatter plot, drag the first example up to the window

- drag the fitted values (PRE_1) to the x-axis, and the residuals (ZRE_1) to the y axis. click ok
- Then you need to add a line at 0 for ease of interpreting
- Double click on graph, select the add line button, set the position value to zero



How to check equal variance for ANOVA: Levene's test and standard deviations:

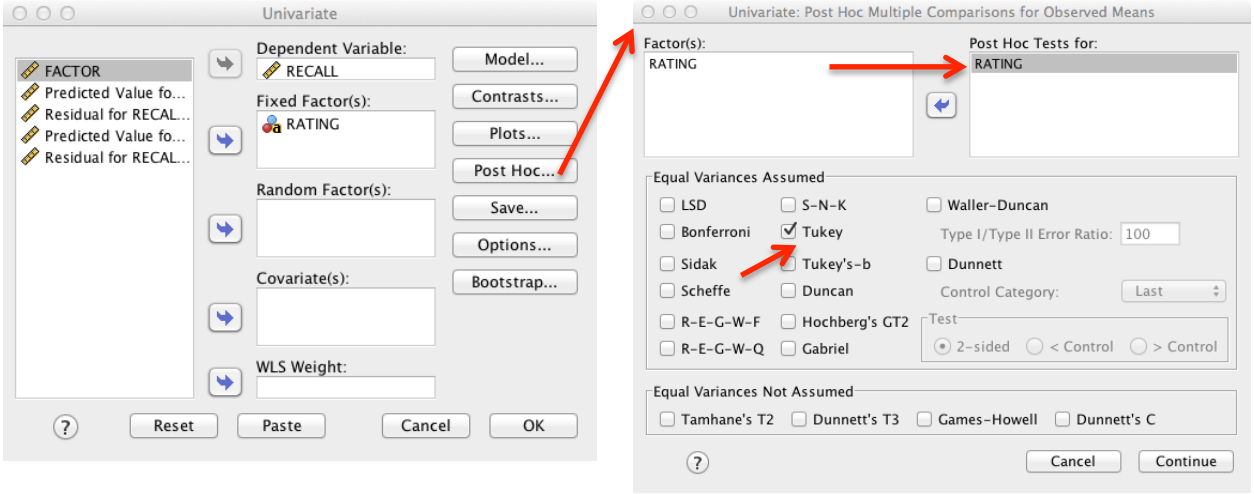
- Both of these can be gotten by running the univariate analysis.
- Click on options
- Then select Descriptive statistics to get the standard deviations
- Select Homogeneity tests to get Levene's test.



How to conduct posthoc tests:

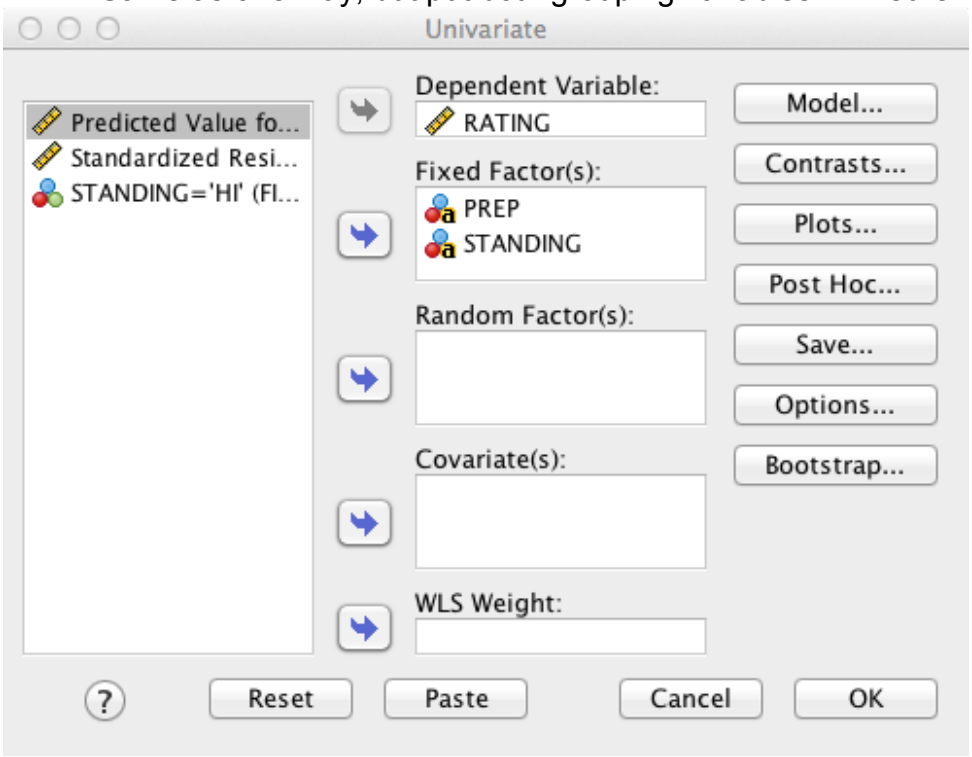
- The option is found by running your univariate analysis

- Then click on Post Hoc…
- Click on the variable you want to test, and use the arrow to move it over into the box entitled ‘Post hoc Tests for’
- Then select your preferred posthoc test, click continue



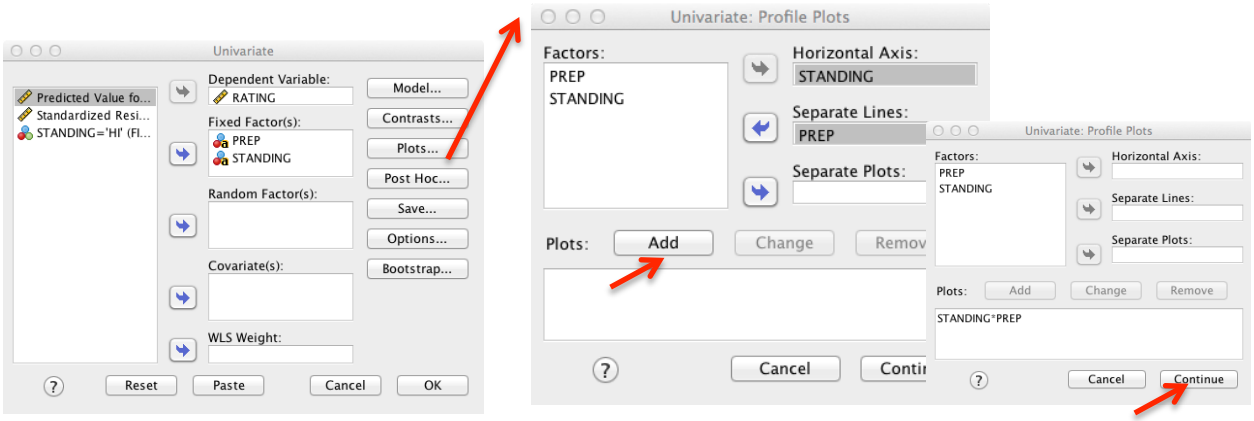
Two-way ANOVA:

- Same as one-way, but put both grouping variables in fixed-effects box



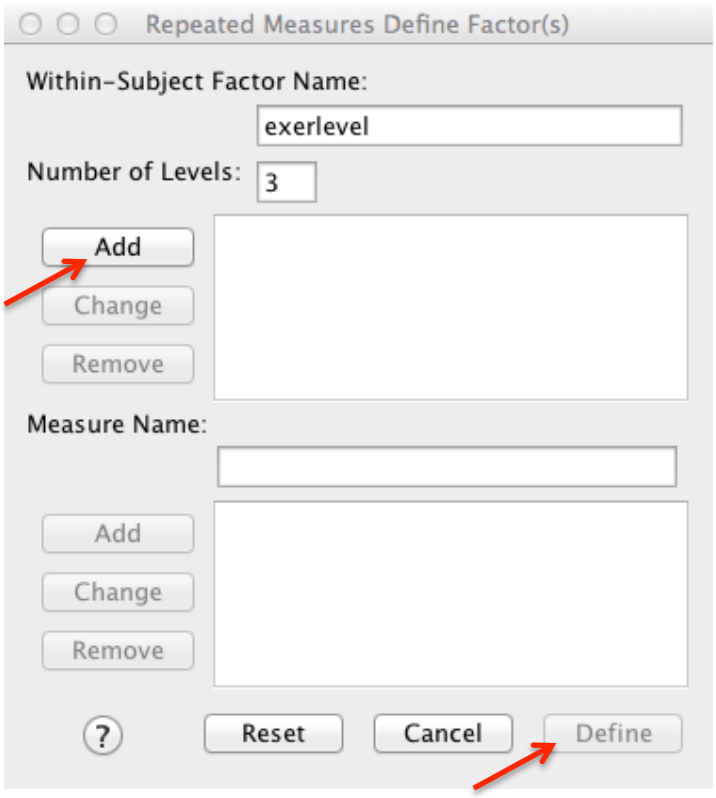
Two-way and repeated measures ANOVA plot

- To look at your data generally, and to look for interactions
- Access from univariate analysis window
- Click on Plots…., Put the variable with the most levels on the horizontal axis, and the other one in the separate lines box.
- Important: Click Add, then click continue



Repeated measures ANOVA:

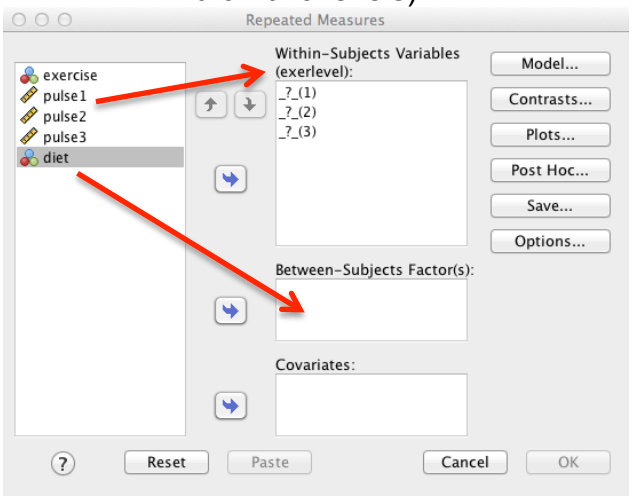
- Analyze – General Linear Model – Repeated measures
- Enter the name of your within subjects factor in the box, put the number of levels
- Click add
- Click define

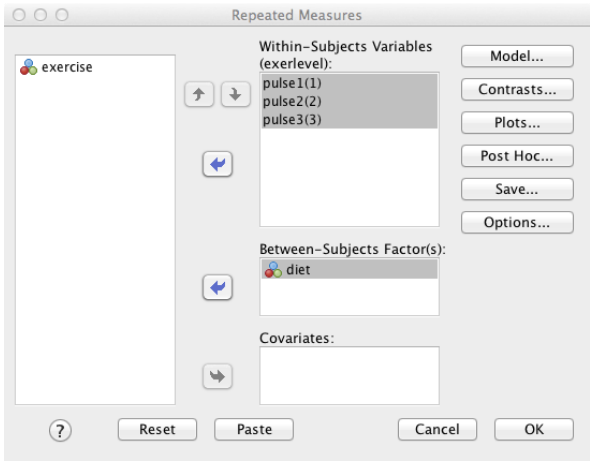


- You need to define your levels of your within subjects variable by dragging them (or using the error) over to the right.
- Then add your between subjects variable to the between subjects box.
- Then everything else is like two-way ANOVA.
  - Plots to make the graph of your data
  - Save to save the standardized residuals (you don't need predicted)

- Remember, you will have to combine the multiple columns it produces into one.

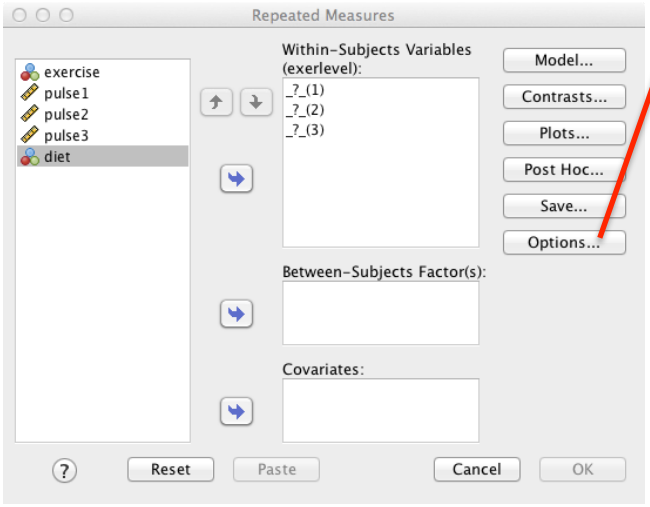
– Posthoc if your between subjects variable will need posthoc tests (more than two levels)

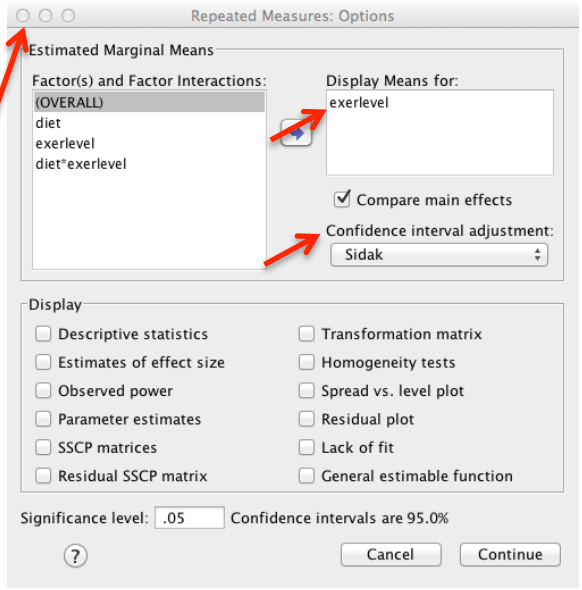




Repeated measures posthoc tests:

- From ANOVA window, click options
- Move your within subject variable over to the 'Display Means For' box
- Click Compare man effects
- For the Confidence Interval Adjustment choose Sidak
- Click continue

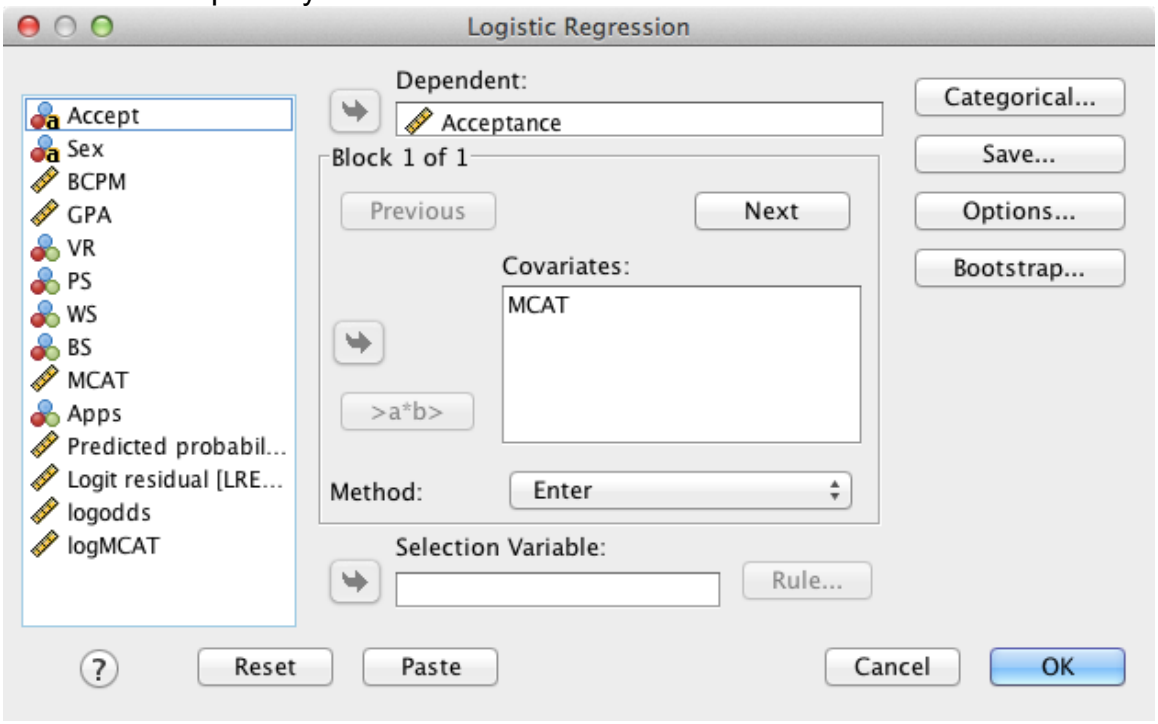




Logistic regression

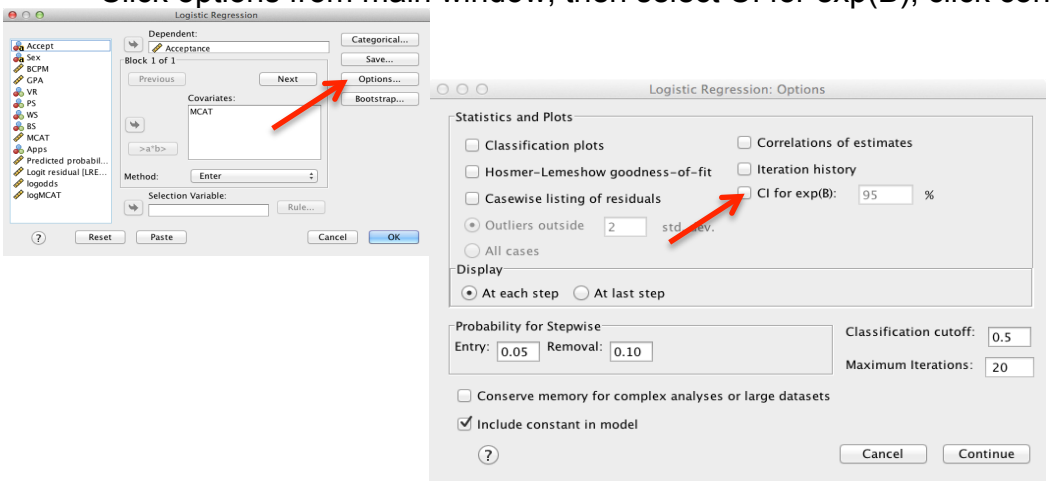
- Analyze -> Regression -> Binary logistic
- Put response in Dependent, predictors in covariates

– If use indicator variables for categorical variables, you don't need to specially define them



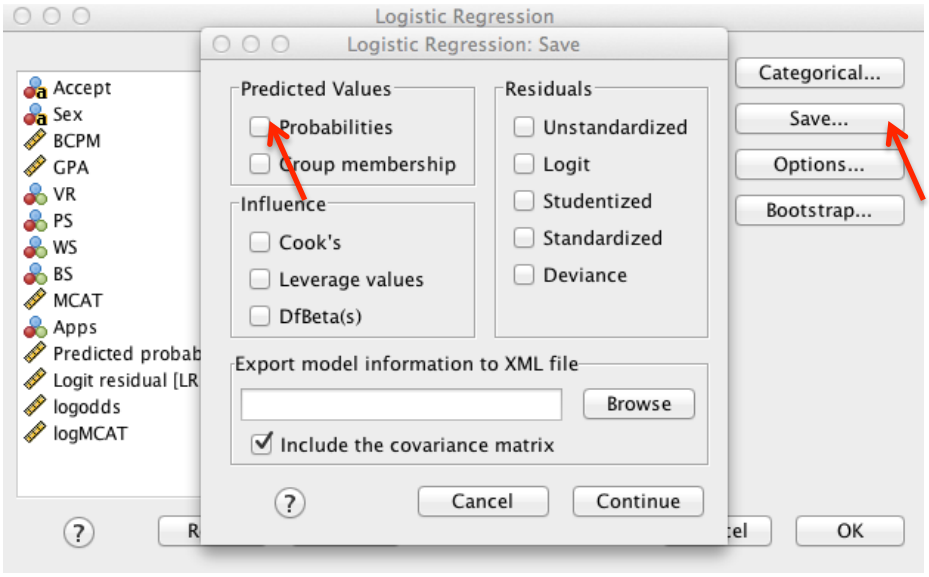
How to get a confidence interval of Exp(B)

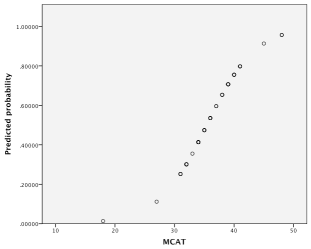
- Click options from main window, then select CI for exp(B), click continue



How to save probabilities for graphs

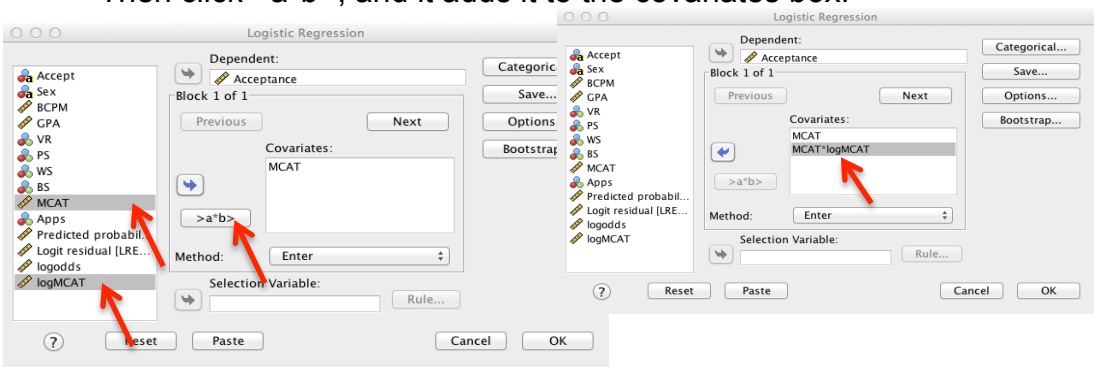
- Click save from main window, click probabilities, then click continue
- Then make a scatter plot of predictor (x axis) and probabilities (y axis)





Test for linearity by adding an interaction term:

- First you need to make the natural log of X, use Transform -> compute variable, choose Arithmetic from the function group on the right, and select ln, hit the blue error next to delete to move it into the expression box.
- Then you can create interaction of X and ln(x) within the logistic dialog box
- Select X and log(X) in the list on the left.
- Then click >a*b>, and it adds it to the covariates box.



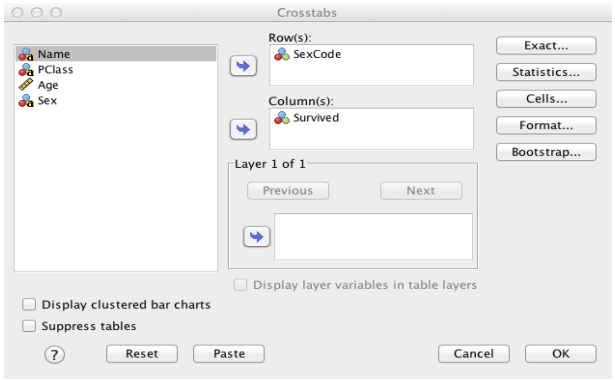
how to make a table of the counts:

- Analyze -> Descriptive Statistics -> crosstabs
- Put one categorical variable in the rows, one in the columns

Accept * Sex Crosstabulation

Count

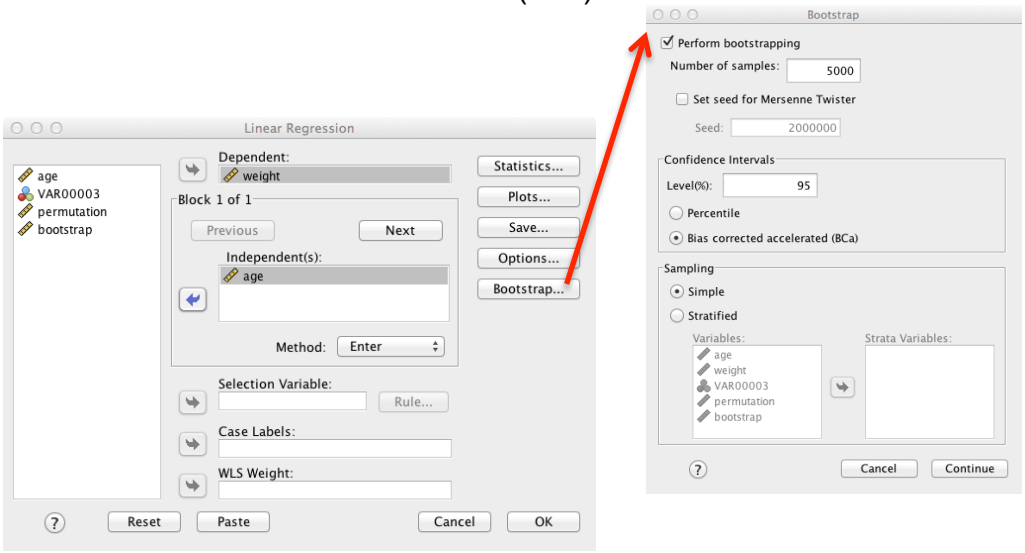
		Sex		Total
		F	M	
Accept	A	18	12	30
	D	10	15	25
Total		28	27	55



Nonparametric tests

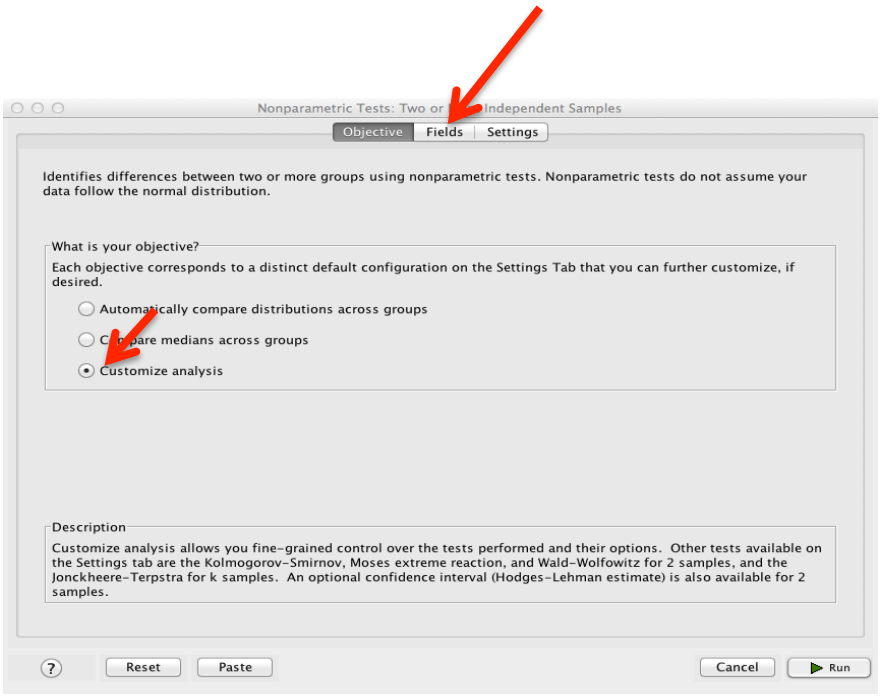
Bootstrap in SPSS

- Just a little box to check!
- Click perform bootstrapping
- Choose number of samples
  - book recommends 5,000
- Confidence intervals
  - choose bias corrected (Bca)

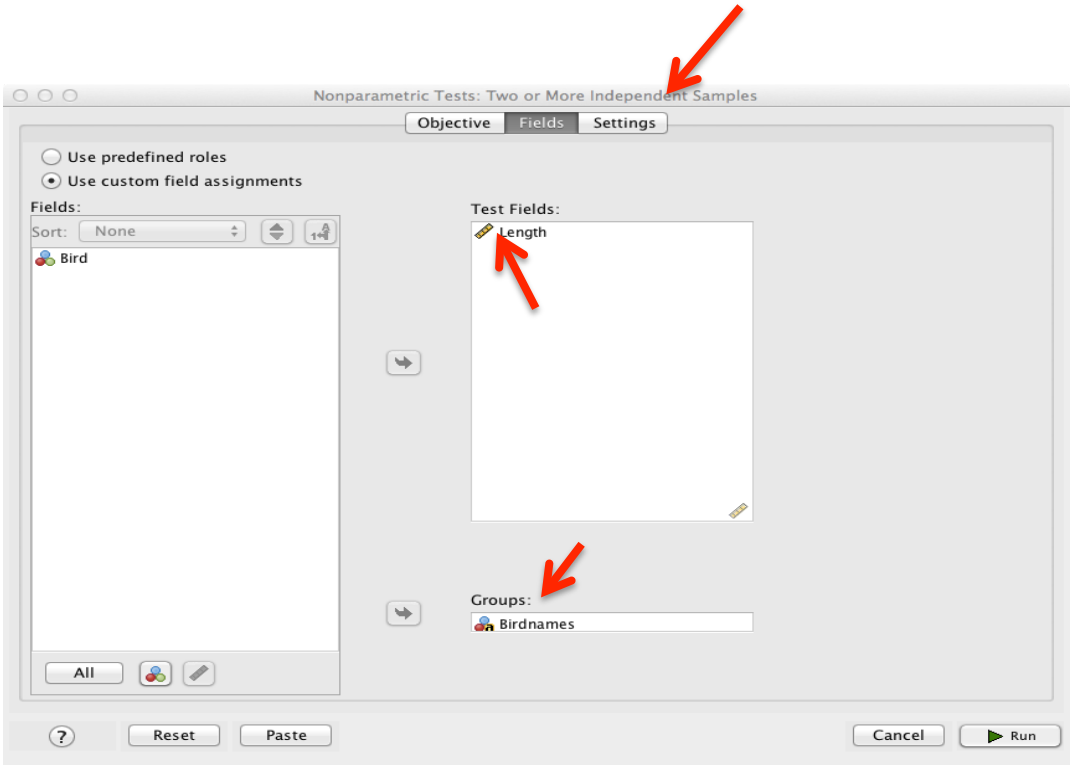


How to run Wilcoxon-Mann-Whitney and Kruskal-Wallis

- need your grouping variable in a single column, it can be numbers or words, but no dummy variables needed
- Analyze -> Nonparametric Test-> independent samples
- Click Customize analysis
- Click Fields at the top



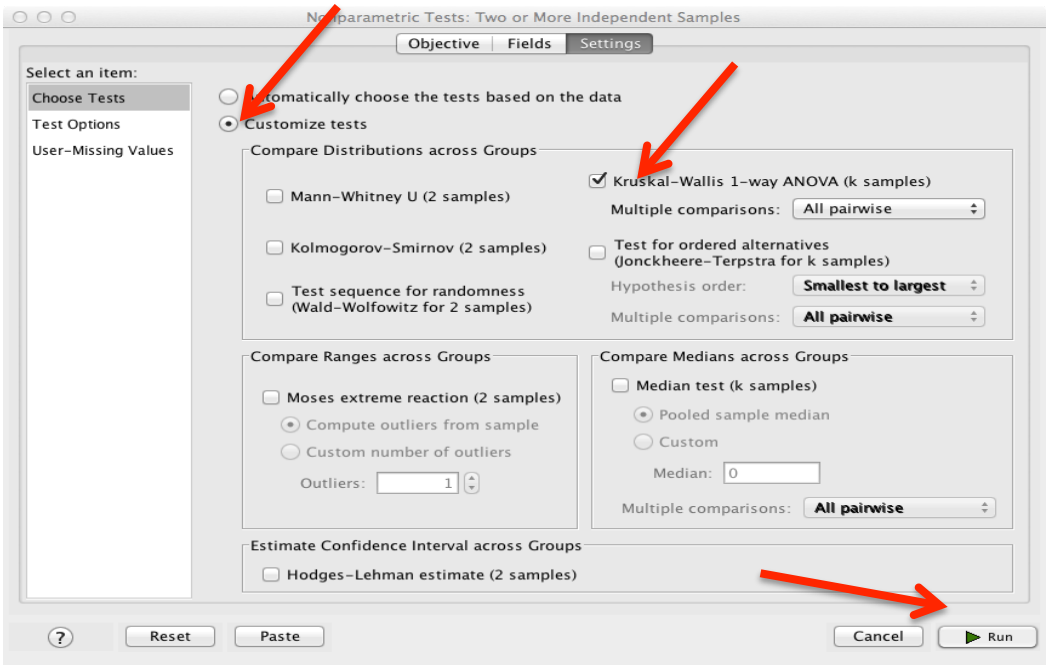
- Move your dependent/response variable to the Test Fields
- Move your grouping variable to Groups
- Click the Settings tab at the top



On the settings tab:

- Click customize tests

- Select either the Mann-Whitney or Kurskal-Wallis
- Click Run



For posthoc tests:

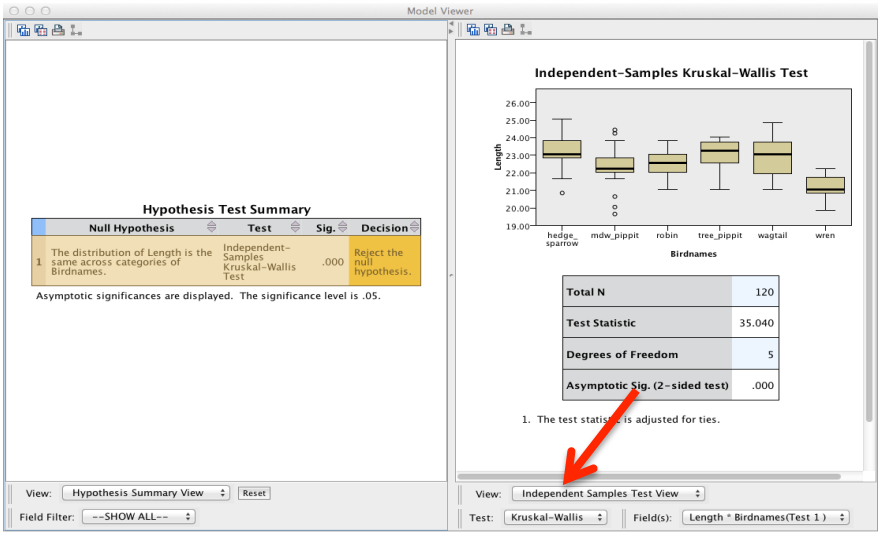
- Double click on the results

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of RECALL is the same across categories of FACTOR.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

- This brings up the window below.
- From the pull down menu by View, choose pairwise comparisons



- The graph at the top gives you the ranks for each group, the table below gives the pvalues for each comparison, with correction for multiple comparisons.