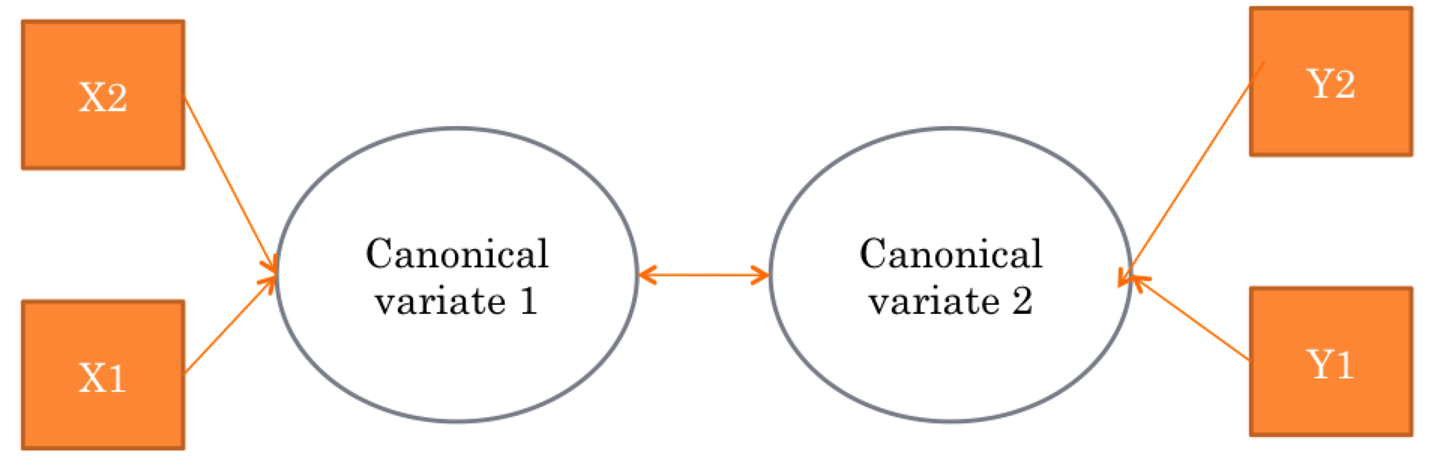
Canonical Correlation

**Description:**

* Canonical correlation is a multivariate correlation, which finds the relationship between several independent variables and dependent variables.
  + Usually you are interested in this type of analysis when you want to compare two concepts that are measured by several variables at once, but instead of averaging or combining the variables, you can use all the variables to create this overall factor of what you are analyzing.
  + Your variables *need* to be continuous.
* Comparison to regression:
  + In multiple regression, you can have several IVs but only one DV.
  + You can consider this analysis a half-way step between multiple linear regression and full structural equation modeling.
* How it works:
  + Regressions are used to create predicted values for both the IVs and DVs, and then these values are correlated. Remember that in multiple linear regression, *R* is the correlation between Yhat (predicted values for each person) and Y. This value indicates how close you are to getting their score correctly. This analysis creates Xhat and Yhat and correlates those two values.
  + In regression – there is only one “best” way to combine the variables to create the closest predicted values usually calculated with least squares. In canonical correlation – there are several ways to combine the variables (similar to MANOVA combinations of the DVs, Wilk’s/Roy’s). Therefore, you get several “solutions” but usually only a couple are understandable.
* Eigenvalues – A mathematical representation of the variance accounted for by that grouping of items. Here’s sort of how eigenvalues work:
  + You have data – that data contains a large pot of variance to work with. This analysis will group that variance together based on the relationships you specify (i.e., these are the IVs, these are the DVs). Eigenvalues are a number that represents how much variance that grouping accounts for out of the total amount of variance in your data set. These numbers are then usually scaled to percentages to make them more interpretable.
* Orthogonality: No overlapping variance across variates. Orthogonal means that once you use the variance once for one pairing (see eigenvalues above), you cannot use that variance again. If the first relationship covers 80% of the variance, not very likely that you will find another significant variate pair because you have used up all the variance.

**Research Questions:**

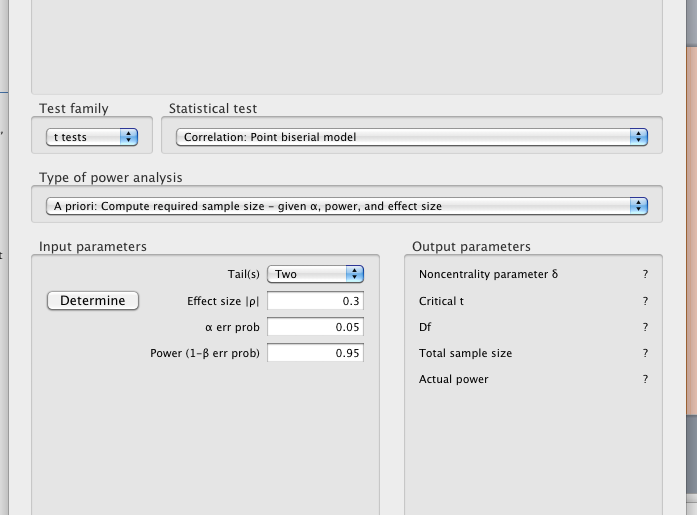
* How many canonical variate pairs did you find?
  + How many significant correlations can you find between the two sets of variables? Basically, this question is how many significant solutions did you find?
  + To assess this research question, you will look at the number of significant pairings, usually only the first couple account for the most variance, and not all are significant.
  + You will get output for the following number of pairs: IV/DV with the smallest number of variables – therefore if you have 4 IVs and 3 DVs, you will get three pairs.
* Importance of canonical variates – interpretation of the solutions that were significant.
  + How strong is the correlation across IV-DV?
    - Usually a very important question – this finding assesses how much the overall variables are related to each other.
  + How strong is the correlation on each side (IV-IV/DV-DV)?
  + How strong is the correlation between each IV-DV combination?



**Power:**

Use G\*Power with correlation options:

* Test family: t-test
* Correlation – Point Biserial model
* Two tails
* Effect size rho: what would you expect the correlation to be between these two concepts? You can hit determine to use their calculator for rho.
* Alpha = .05.
* Power = .80.



**Assumptions:**

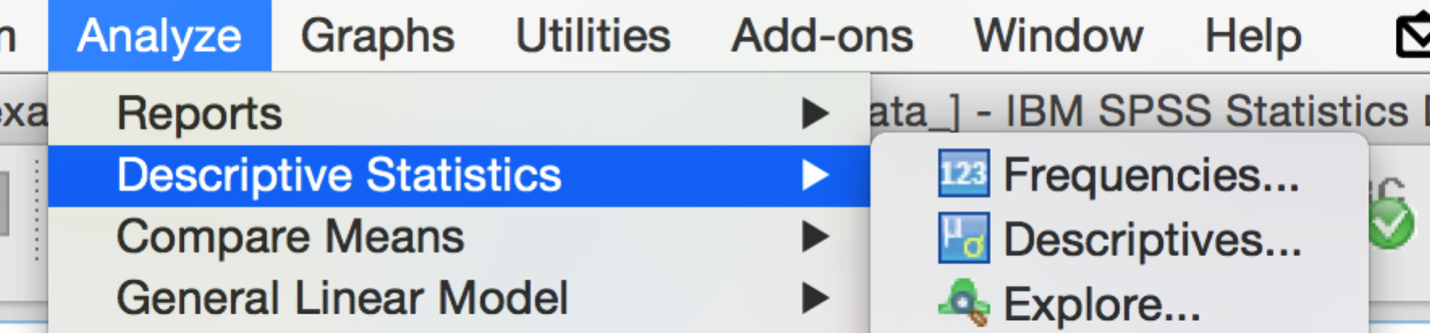
1. Sample size – you need at least 10 people to each IV and DV if your variables are very reliable. Larger sample sizes are helpful.
2. Missing data – very sensitive to the different types of data replacement. You might try various version of data replacement to see which theoretical picture makes the most sense.
3. Outliers – just as with regression, outliers will change the relationship between the IVs and DVs. You will have a hard time combining the IVs and DVs if there are scores that are crazy.
4. Multivariate Normality – you want the combinations of the IVs and DVs to be normally distributed.
5. Mutlivariate Linearity – since this is regression, you want the combinations of the IVs and DVs to be linear so you can create the best combination.
6. Homoscedasticity – you want the variance across all the IV/DV combinations to be the same.

Complete Example

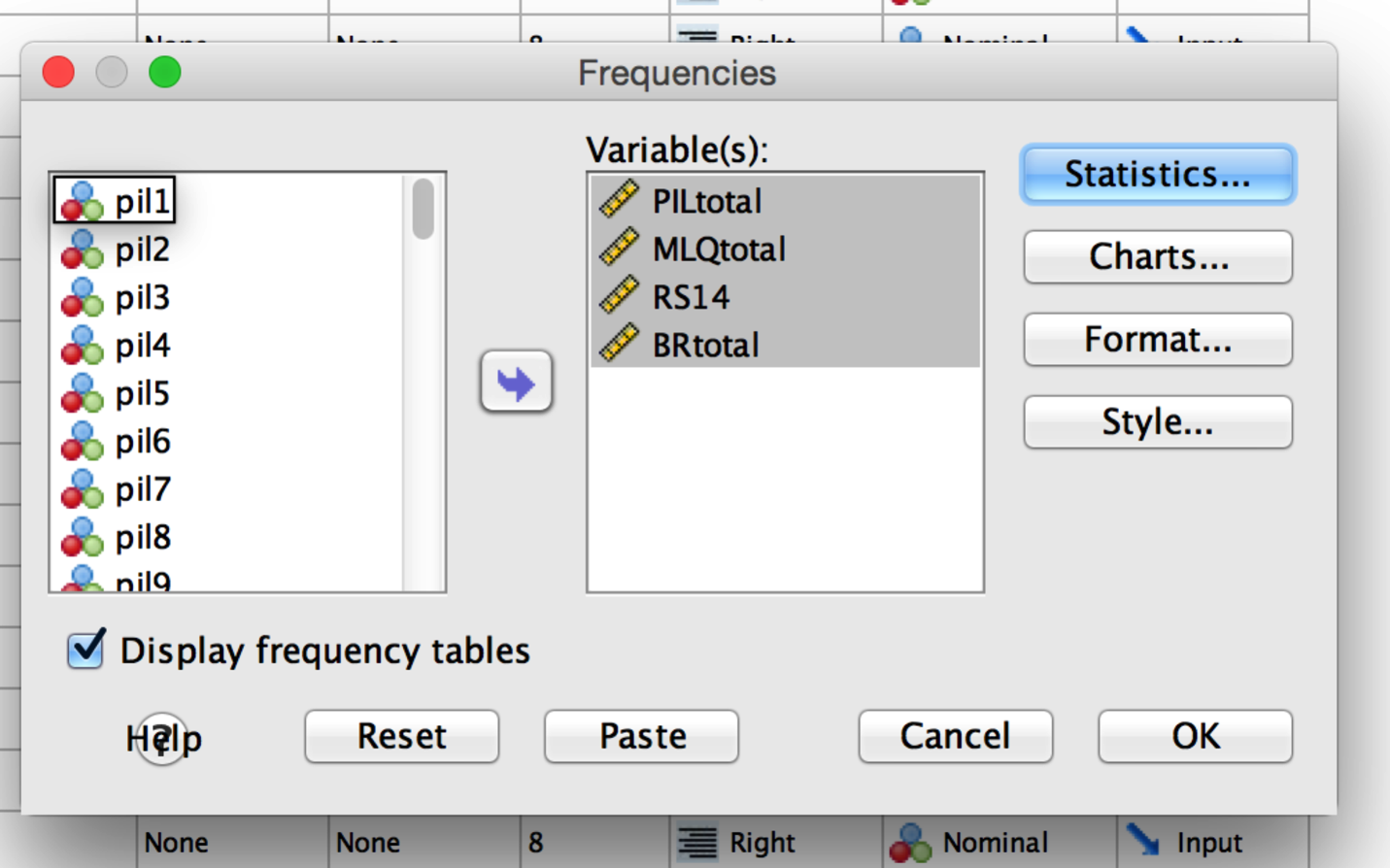
**Research Question:** We used two scales of meaning and purpose in life to assess an underlying meaning factor (PIL and MLQ). Additionally, we used two scales of resiliency (BR, RS14) to assess an underlying resiliency factor. Many researchers argue these are two names for the same idea. Therefore, we wish to use canonical correlation to look at the relationships of these two factors with multiple scales assessing the latent construct.

**Assumption Checks:**

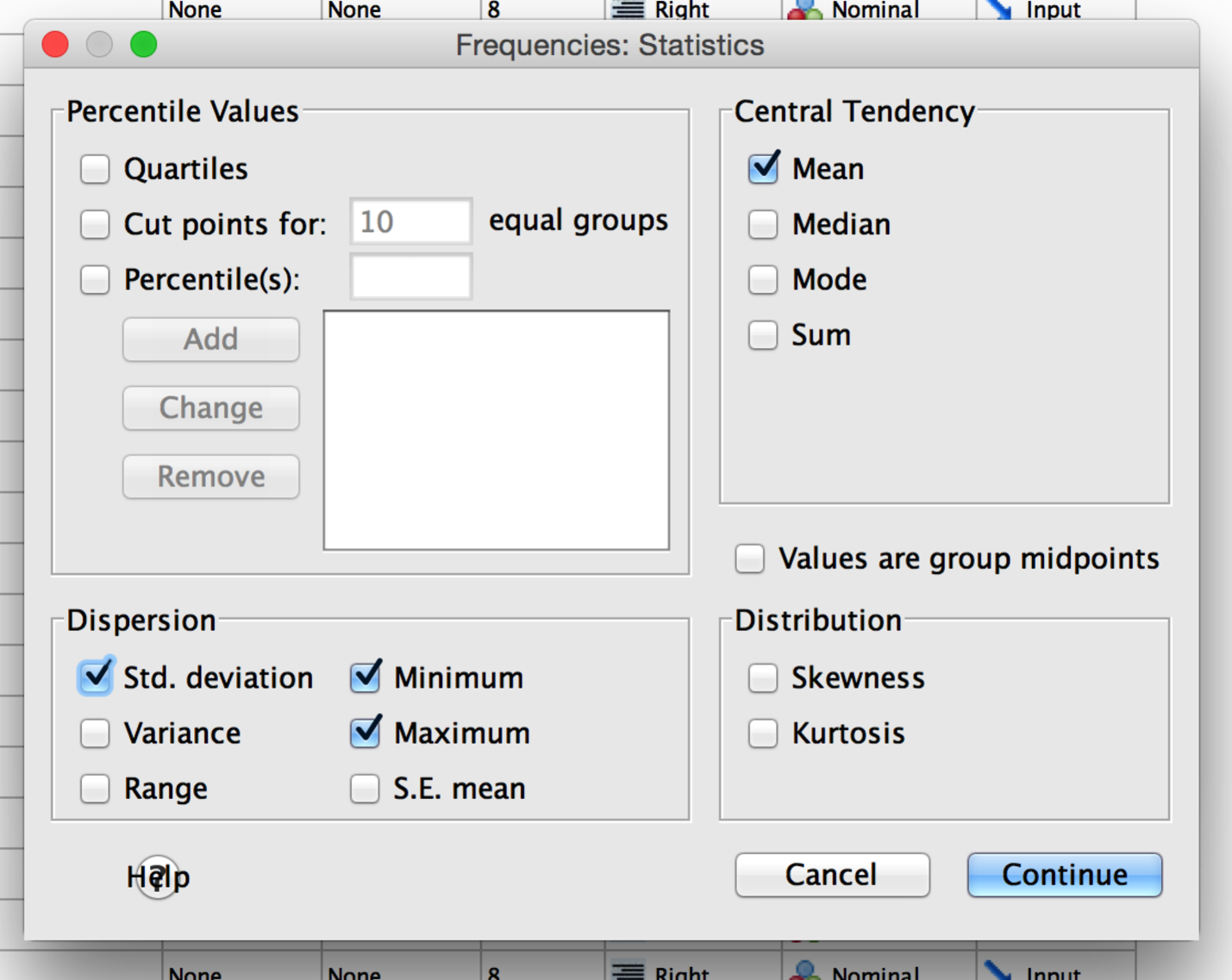
1. Check for missing data and accuracy errors:
   1. Analyze > descriptives > frequencies.

****

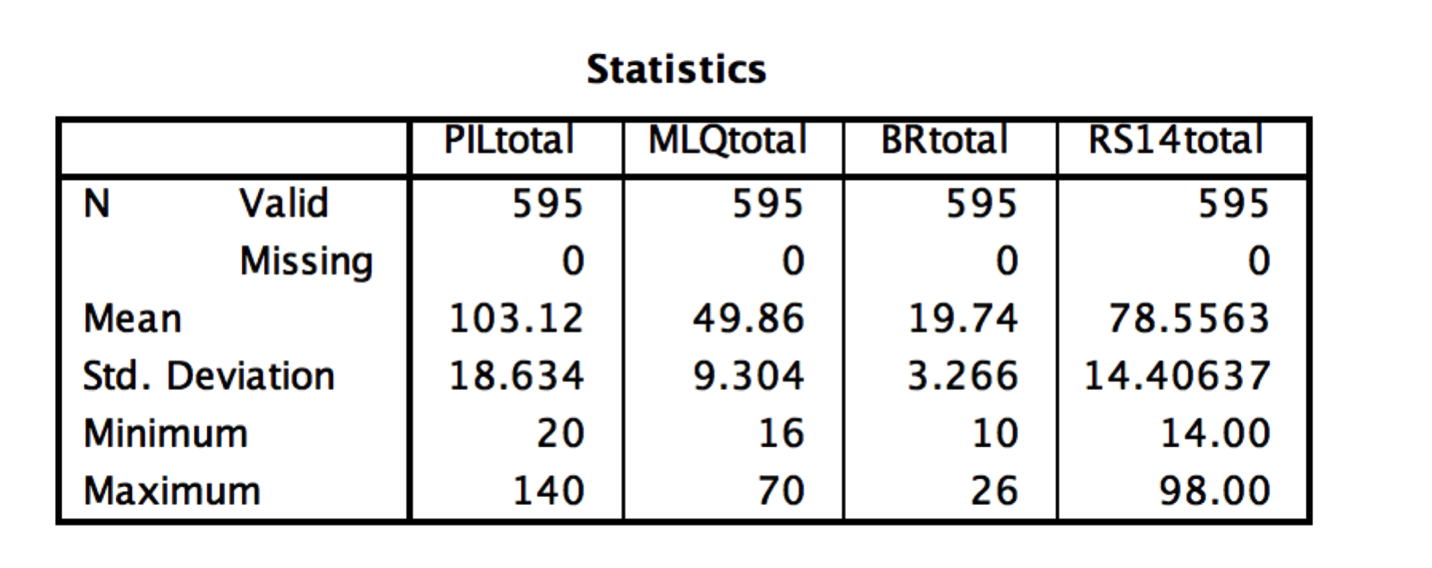
* 1. Move over the variables you are going to use as measures for the IV and DV.

****

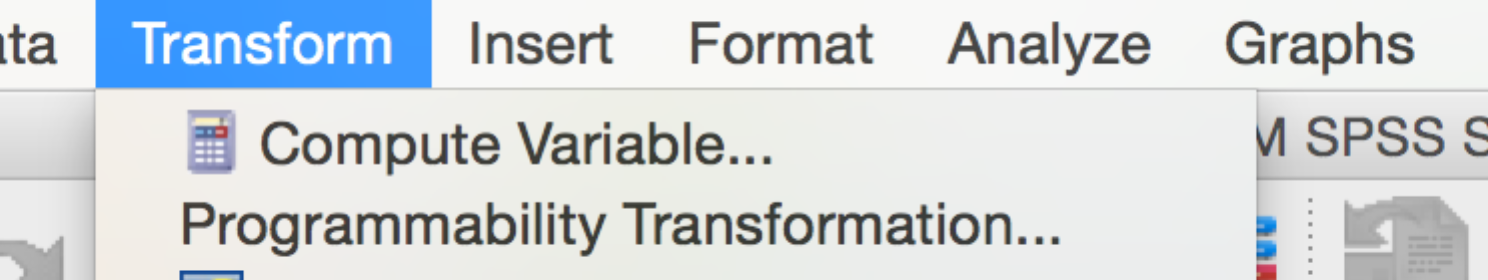
* 1. Click statistics > mean, standard deviation, minimum, and maximum.

****

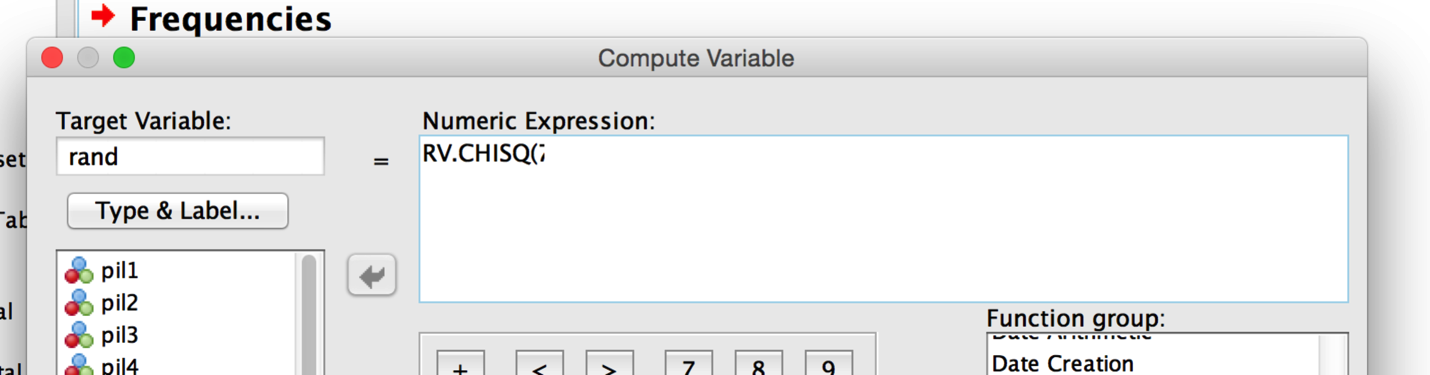
* 1. Click continue and ok.
  2. Look at the minimum and maximum to make sure they are in range (they are for these variables), check out the means/SDs to makes sure they don’t seem crazy for accuracy.
  3. Look at the missing line to make sure there isn’t any missing data (which you can use linear trend at point to replace a small number of points, see data screening notes).



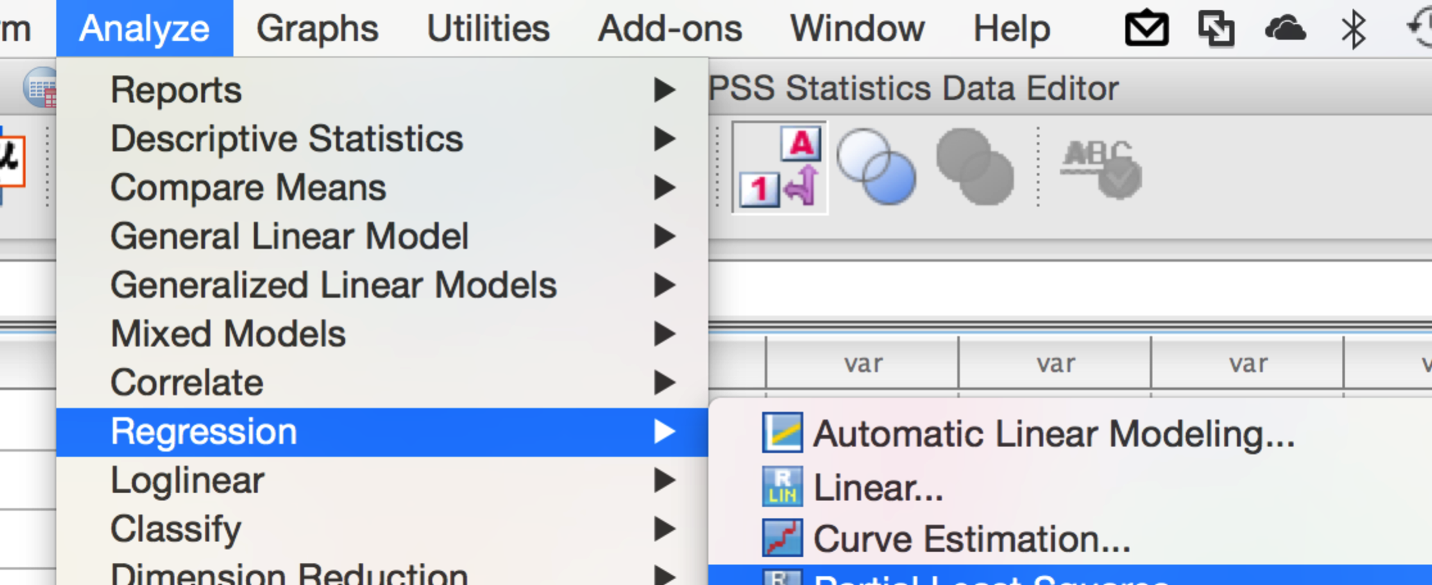
1. Outliers.
   1. This analysis is a multivariate analysis, so we are going to use Mahalanobis distance to check for outliers.
   2. While canonical correlation is a type of regression, you run the analysis through syntax. Therefore, we will use a fake regression to check all the IVs and DVs together at once.
   3. First, make a random variable:
      1. Transform > compute variable.



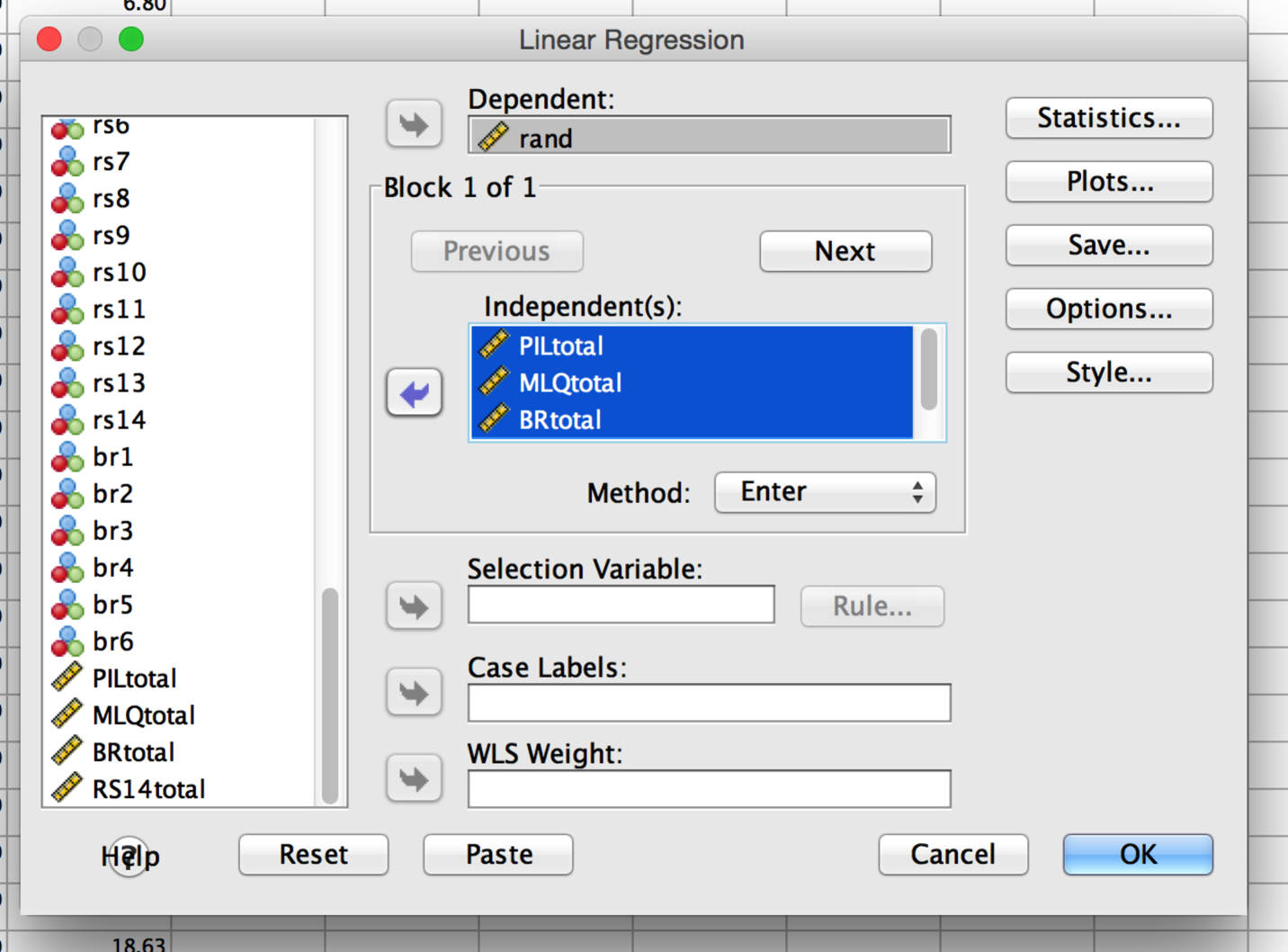
* + 1. Enter a target variable name (rand or random).
    2. Use the function group option to find random number generators.
    3. Use RV.ChiSquare as your random number generator.
    4. The number 7 works well for the ?? to generate from.



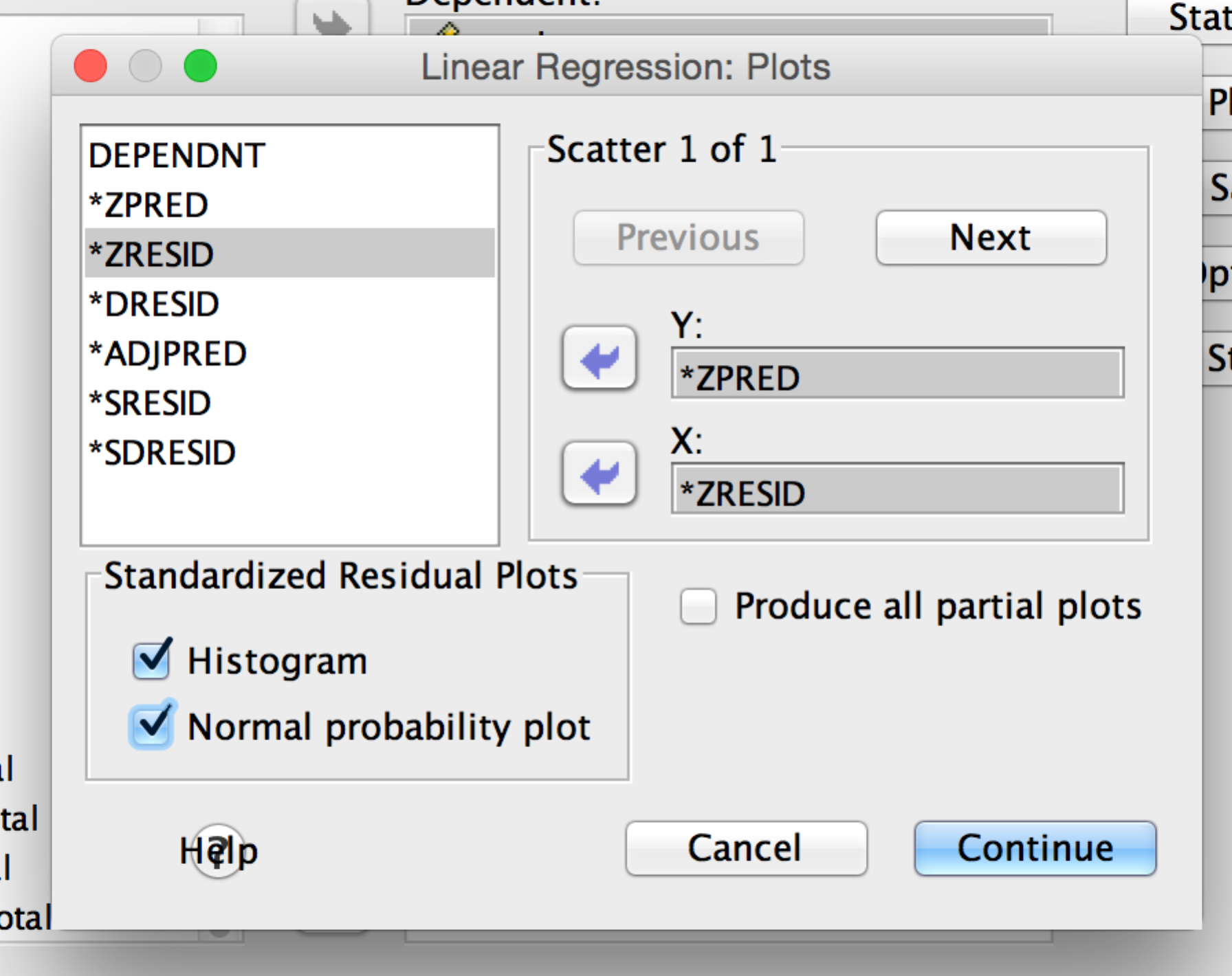
* 1. Now run a fake regression to get outlier scores and assumptions.
     1. Analyze > regression > linear.



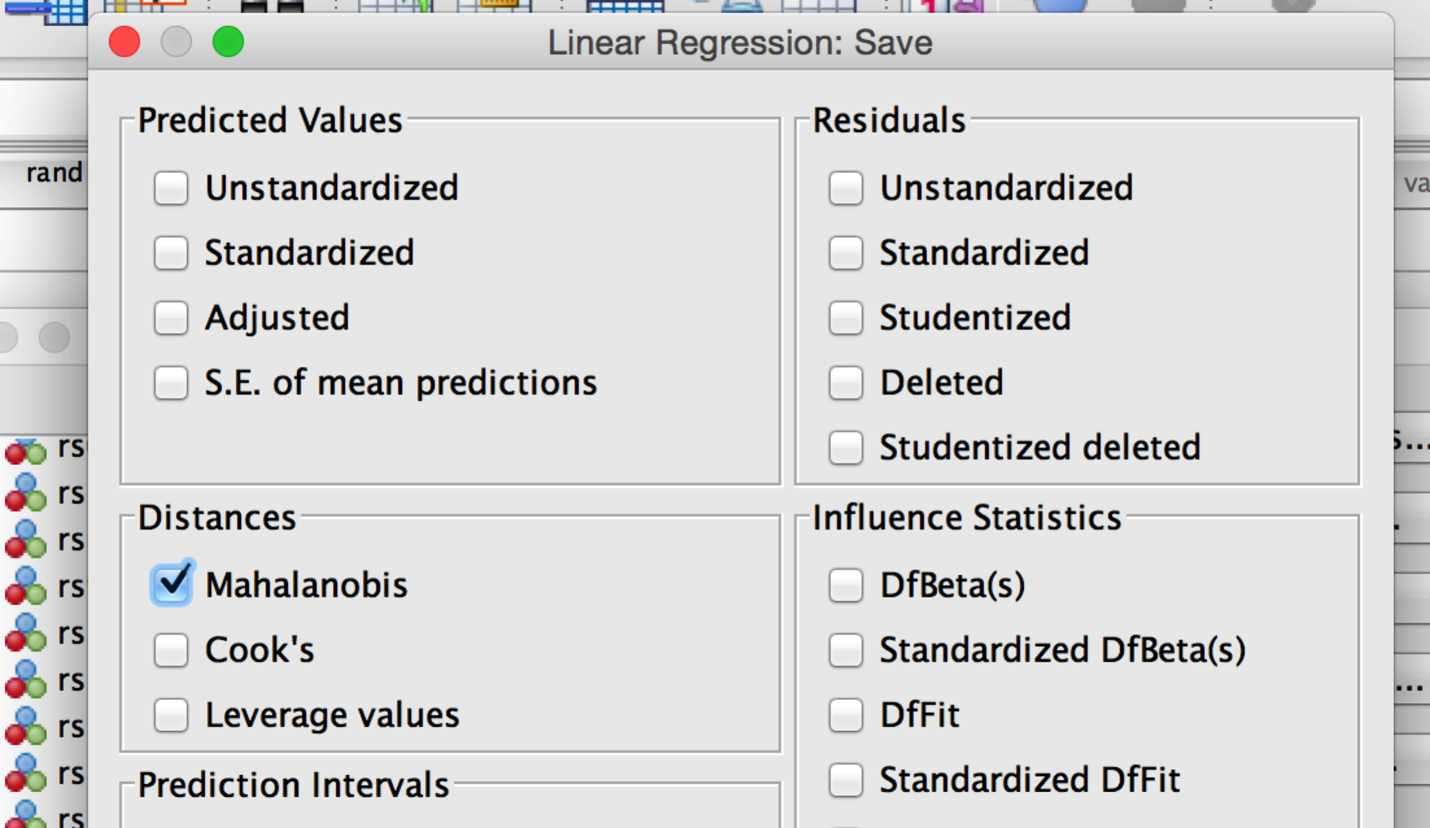
* + 1. Put the random variable in the dependent box, all the IVs and DVs in the independent box.
    2. Click on plots.



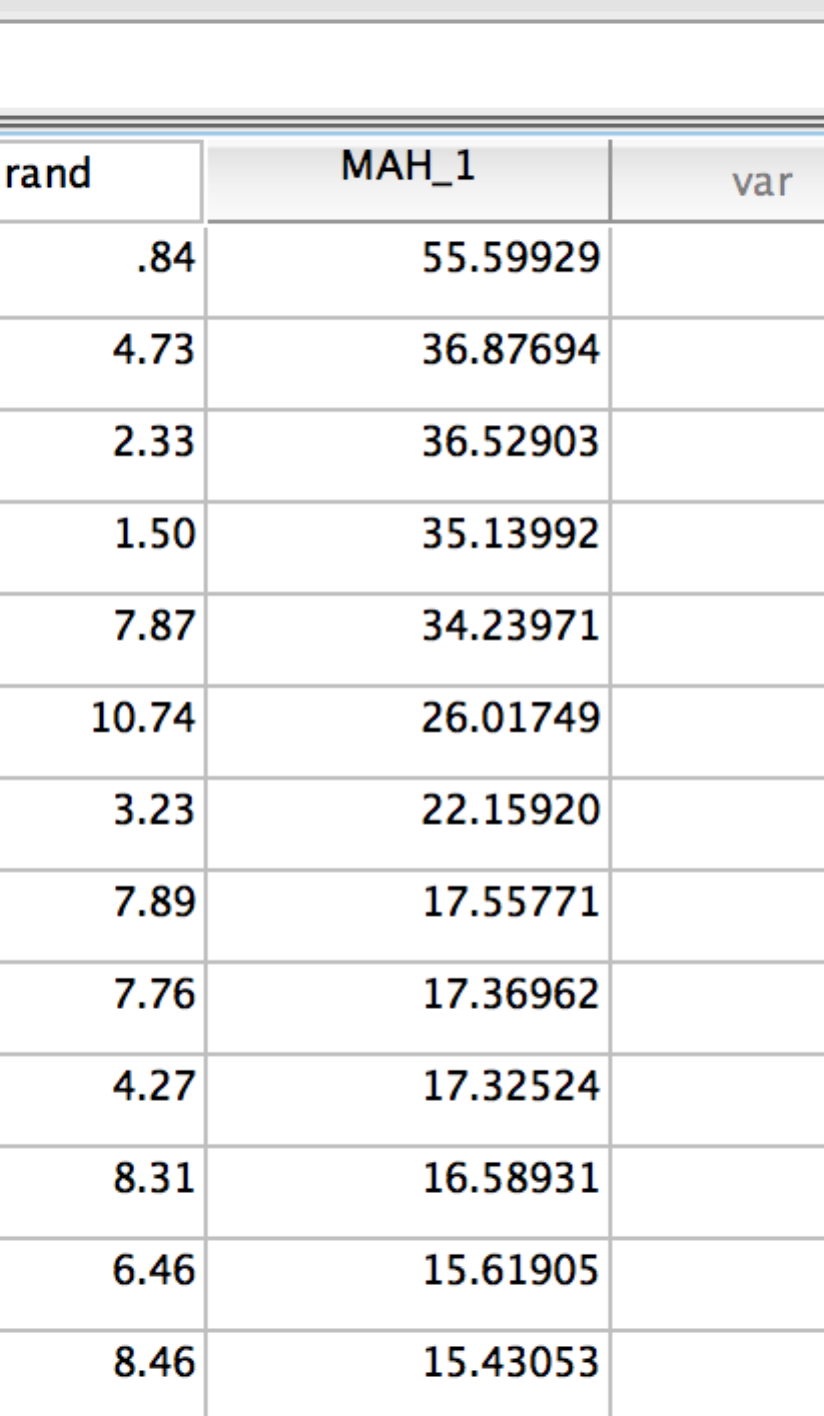
* + 1. Zpred goes in Y, Zresidual goes in X.
    2. Click histogram and normal probability plots for normality and linearity assumption checks.

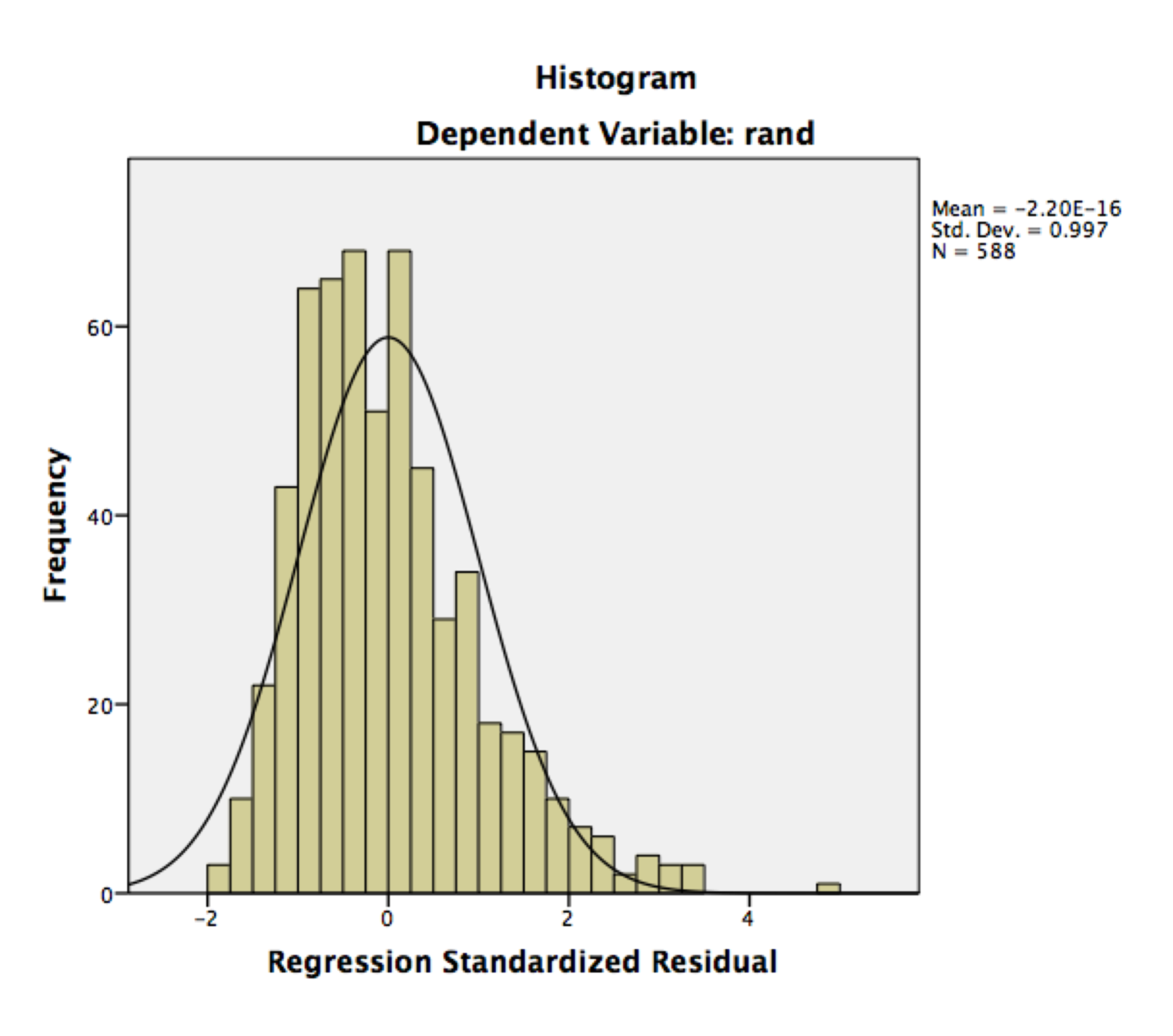


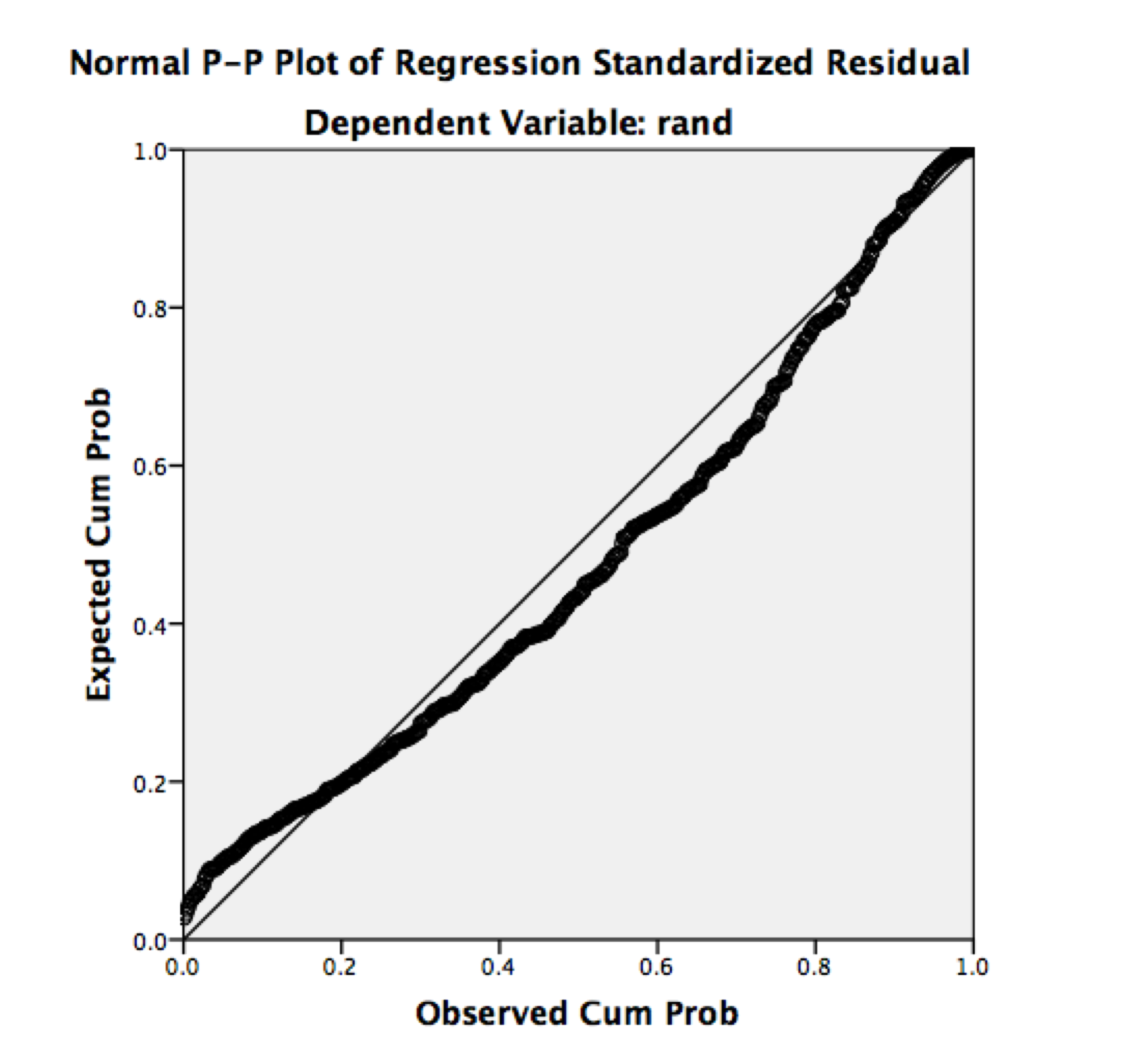
* + 1. Click options and pick Mahalanobis.



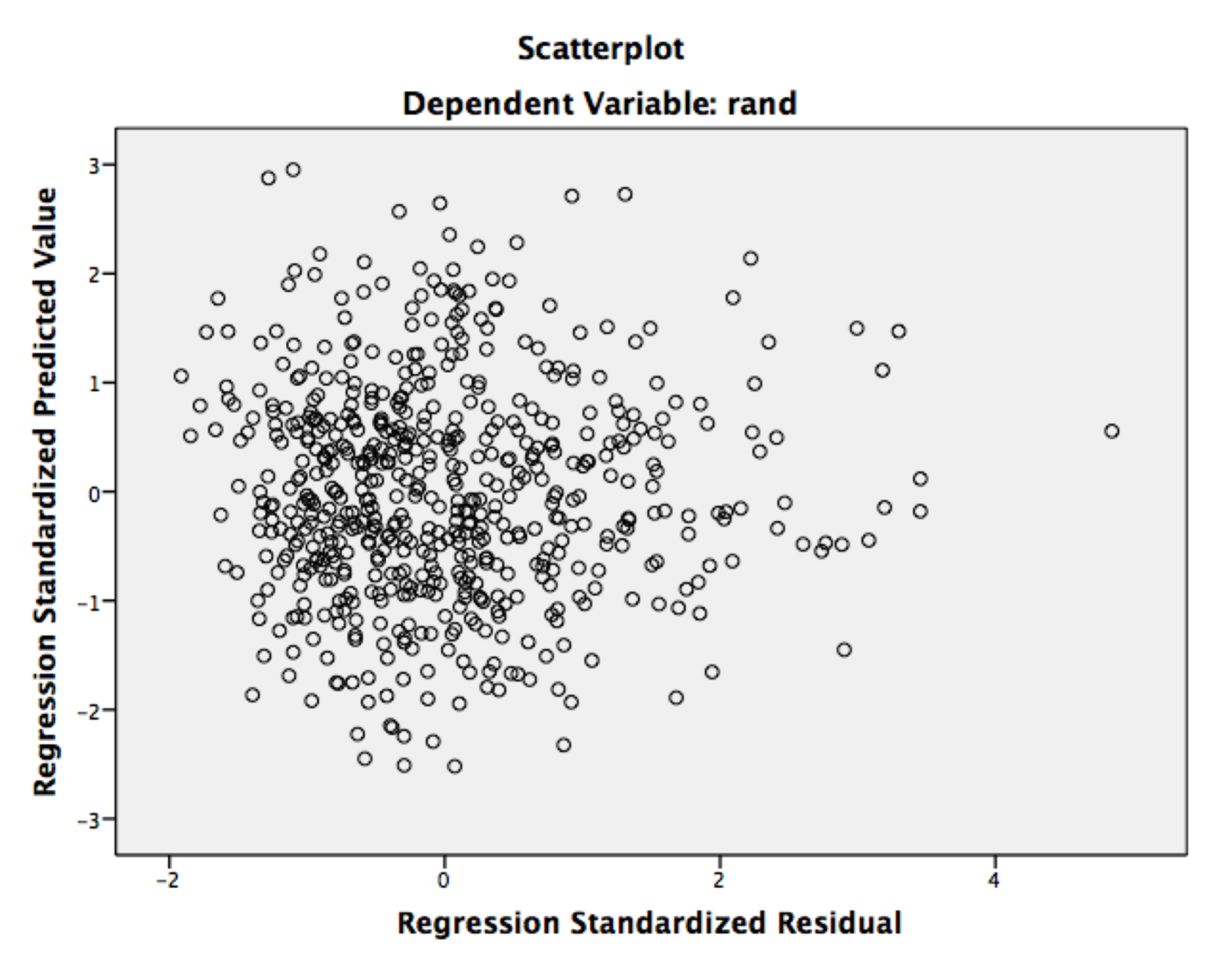
* + 1. Click continue and ok.
  1. Let’s check out the Mahalanobis scores.
     1. Our cut off will be df = 4 for four variables, p<.001, 18.47 using a chi-square table.
     2. Therefore, I had 7 outliers. I deleted them for the rest of the analyses.



1. Normality – let’s check out if the multivariate histogram is roughly normal. Remember you have to rerun the fake regression if you deleted people.
   1. I would say this data is roughly normal, with one outlier-ish person. We do not want to double delete people though, so I’m leaving them in. 
2. Linearity – this analysis creates multiple regression equations to be able to figure out how to combine your variables, so you would need the variables to be linearly related.
   1. This plot looks pretty good.



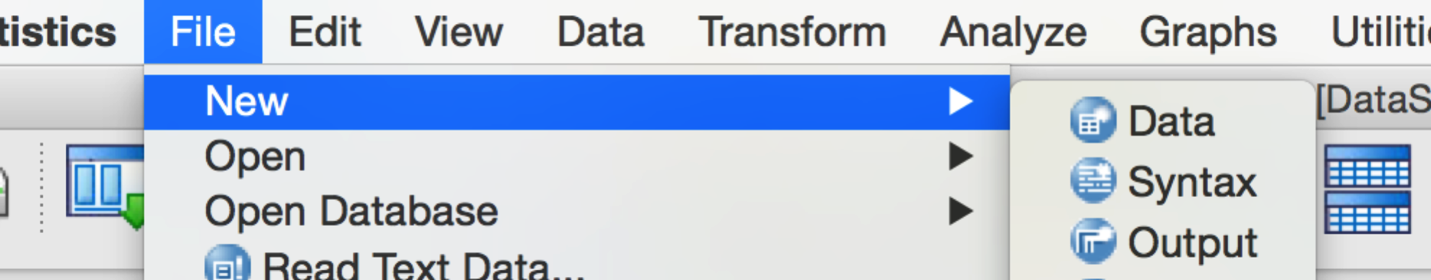
1. Homoscedasticity – again, since canonical correlation is a fancy form of regression, you would want to make sure that that spread of the residuals is the same across all the variables.



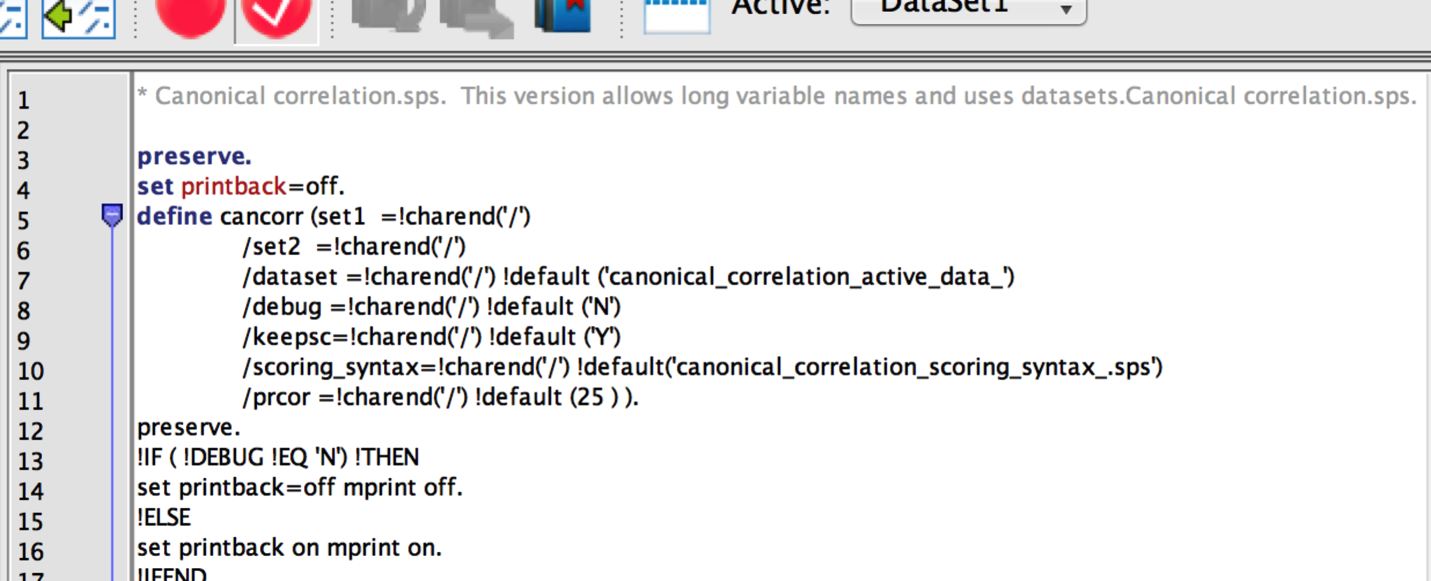
* 1. We see that we have more data clumped at the beginning, so from 2 to 2 it’s ok, but after that we are starting to see unequal spread.

**How to Run Analysis:**

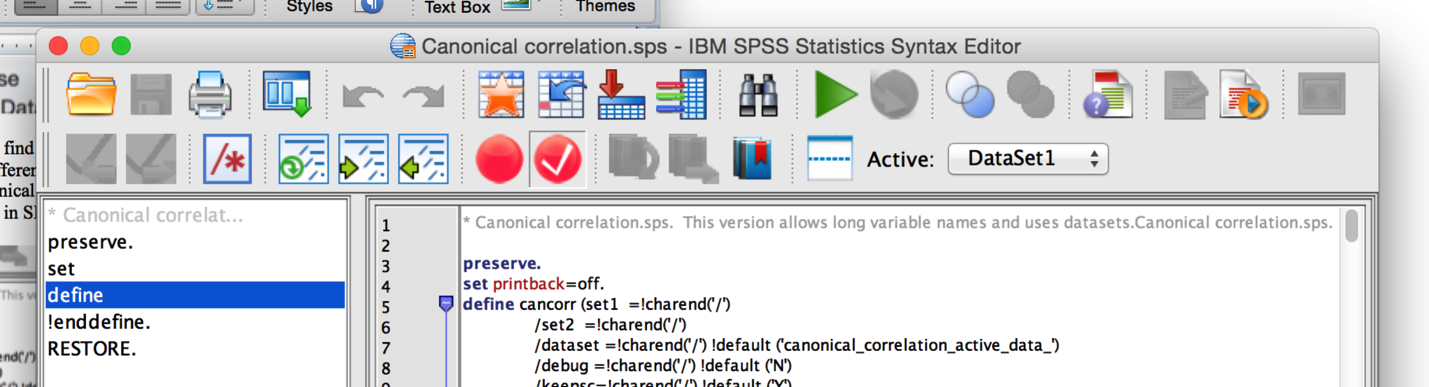
1. First we have to create syntax to run this analysis.
   1. File > New >Syntax.



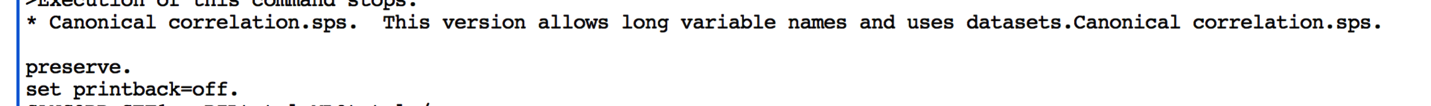
* 1. Now we have to find the canonical correlation file that comes with SPSS. (note: this will look different on your computer).
  2. Search for Canonical correlation.sps. Double click to open it. (You can use the include function in SPSS, but this method seems easier to me.).



* 1. Highlight the entire file (control + A).
  2. Click the big green > button to run the syntax.



* 1. You will get one line of information in your output.

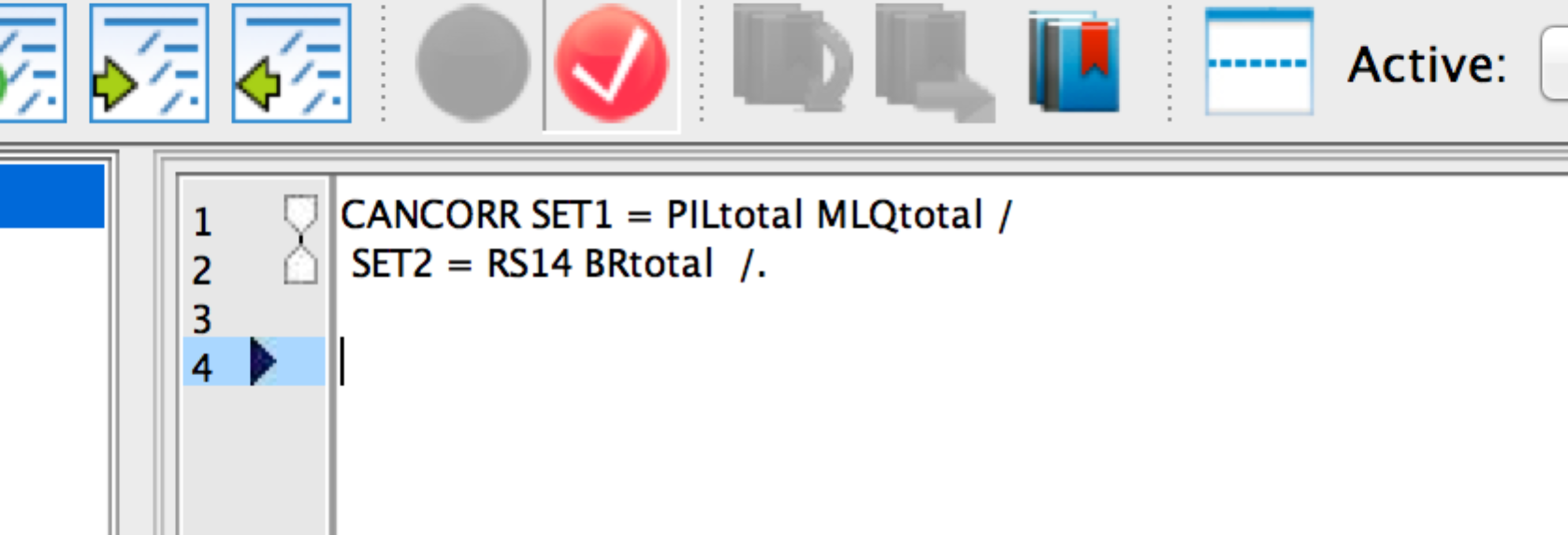


* 1. In your separate syntax window use the following code:

CANCORR SET1 = PILtotal MLQtotal /

SET2 = RS14total BRtotal /.

* + 1. Note that you will have to change out the variable names here for your data. They must be spelled exactly the same for this analysis to work.
  1. Highlight both rows and click the big green arrow to run.



**Reading the Output:**

1. Correlations between IVs only – this section tells you how correlated the IVs are (really either side could be IV or DV, there’s no causal implication here, but it’s easier to call them IVs and DVs as set 1 and set 2). You actually do *want* them to be highly correlated or why are you trying to combine them together? Think about how you wanted the variables to be correlated in MANOVA.
   1. How to report: *r* = .52

Run MATRIX procedure:  
  
  
Correlations for Set-1  
         PILtotal MLQtotal  
PILtotal   1.0000    .5187  
MLQtotal    .5187   1.0000

1. Correlation for your DV set – same idea as above.
   1. How to report: *r* = .50

Correlations for Set-2  
         RS14tota  BRtotal  
RS14tota   1.0000    .4955  
BRtotal     .4955   1.0000

1. Correlations between set 1 and set 2. These are the correlations between your IVs and DVs as individual variables. Basically the first three parts of the output is a correlation matrix, broken down into parts.
   1. PIL and RS *r* = .71
   2. PIL and BR *r* = .51
   3. MLQ and RS *r* = .56
   4. MLQ and BR *r* = .17
   5. In a sense you want these to be correlated because you are saying that you think that the combined set 1 (IVs) and the combined set 2 (DVs) are correlated to each other overall. Therefore, the individual variables are probably correlated too.

Correlations Between Set-1 and Set-2  
         RS14tota  BRtotal  
PILtotal    .7142    .5077  
MLQtotal    .5606    .1688

1. Canonical correlates: these values represent the double headed arrow on the picture shown at the beginning. It’s the correlation between the combined IV score and the combined DV score. You will get more than one – remember there are multiple ways to combine these values. You will get whichever is smaller number of IVs or number of DVs.
   1. First canonical variate: *r* = .76
      1. That means that 58% of the variance between the two combined variables overlaps in the first equation. (I squared .76).
   2. Second canonical cariate: *r* = .29
      1. Another 9% of the variance can be explained by the second equation.
   3. Remember that these are orthogonal, the variances DO NOT overlap.

Canonical Correlations  
1       .755  
2       .293

1. Tests of those variates: this section tells you if the *r* values listed above are significantly greater than zero. Basically, it’s testing if the amount of variance accounted for is better than nothing.
   1. First canonical variate: X2(4) = 545.45, *p* < .001
   2. Second canonical variate: X2(4) = 545.45, *p* < .001
   3. Both are significant!

Test that remaining correlations are zero:  
      Wilk's   Chi-SQ       DF     Sig.  
1       .393  545.449    4.000     .000  
2       .914   52.358    1.000     .000

1. Standardized canonical coefficients – for the IV set only. These values are the loadings for the variables onto the overall latent variable (the combined variable). This section tells you which variables were more strongly related to the overall score. You can think about these loadings as beta values in a regular regression.
   1. You will also get the raw coefficients – these values are the b values in regression and are generally ignored in this type of analysis.
   2. The PIL was more strongly related to meaning than the MLQ in the first variate, but in the second variate, the MLQ was more strongly related to meaning.

Standardized Canonical Coefficients for Set-1  
                1        2  
PILtotal    -.823    -.831  
MLQtotal    -.283    1.135  
  
Raw Canonical Coefficients for Set-1  
                1        2  
PILtotal    -.045    -.045  
MLQtotal    -.031     .124

1. Standardized canonical coefficients – for the DV set only.
   1. In this part, we see that the RS was more strongly related than the BR, but they flipped for the second variate.

Standardized Canonical Coefficients for Set-2  
                1        2  
RS14tota    -.906     .710  
BRtotal     -.168   -1.139  
  
  
Raw Canonical Coefficients for Set-2  
                1        2  
RS14tota    -.068     .054  
BRtotal     -.052    -.349

1. Canonical Loadings - correlations between individual variables and their canonical variate. This value is similar to *r*, remember that the standardized loadings are beta, so they are controlling for other variables. This loading is just the relationship between the variable and variate.
   1. These values are very similar to factor analysis – the rules for variables that are useful is that the loading is > .300.

Canonical Loadings for Set-1  
                1        2  
PILtotal    -.970    -.242  
MLQtotal    -.710     .704

Canonical Loadings for Set-2  
                1        2  
RS14tota    -.989     .146  
BRtotal     -.617    -.787

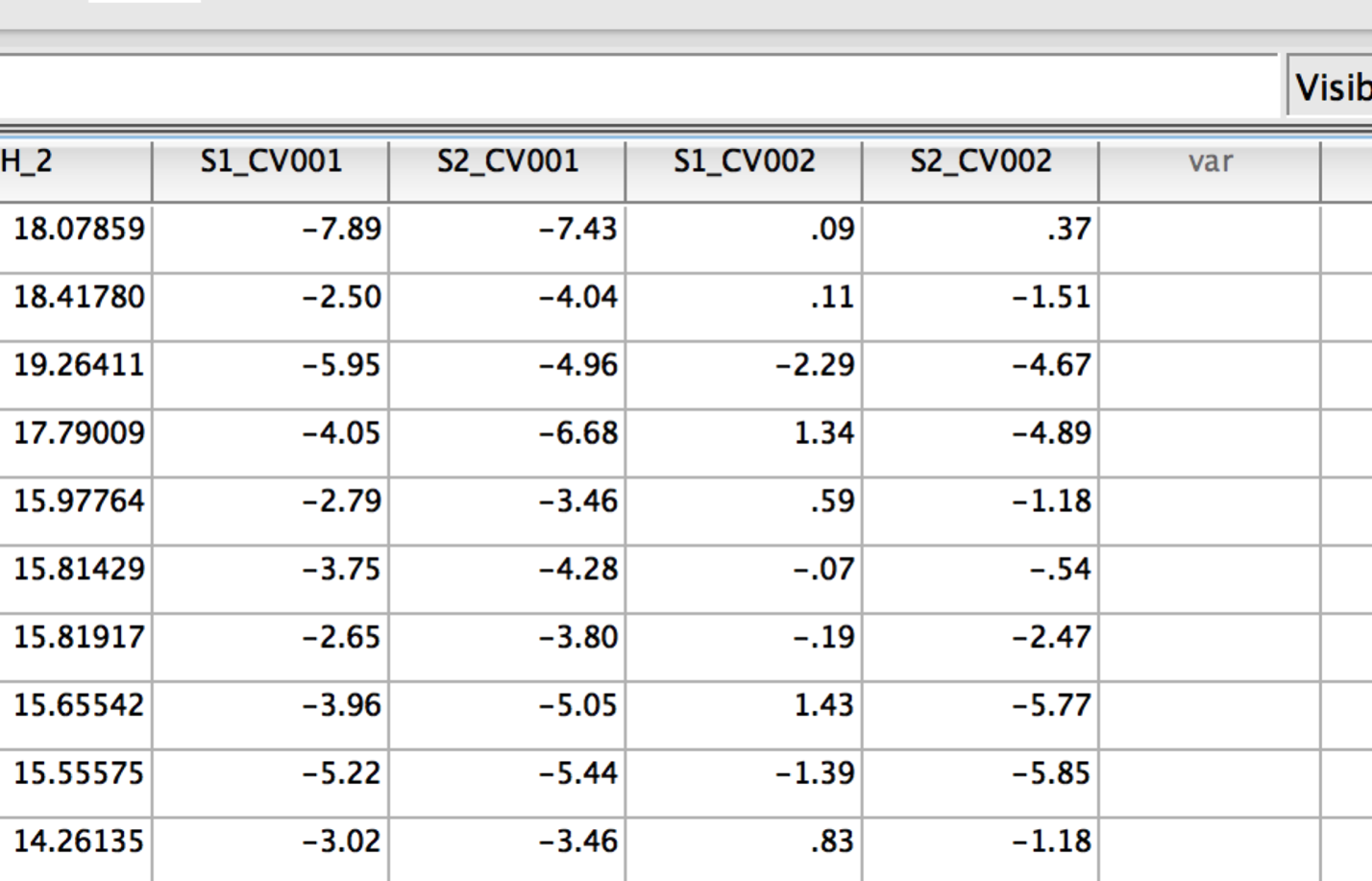
1. Cross loadings – the correlation between the IVs and the other canonical variate (i.e. the overall DV).

Cross Loadings for Set-1  
                1        2  
PILtotal    -.732    -.071  
MLQtotal    -.536     .206  
  
Cross Loadings for Set-2  
                1        2  
RS14tota    -.747     .043  
BRtotal     -.466    -.230

1. Redundancy Analysis – this analysis tells you how much variance the variables account for in the variate. You want the variables to measure their own variate pretty well, and to account for less variance across the other variable (otherwise you are measuring the other latent variable better with your IVs).
   1. Here, the meaning varaibles account for 72% of variance in the first variate, and 28% in the second variate for meaning.
   2. For the meaning variables, they account for 41% of variance in resiliency, and 2% of the variance in resiliency in the second set.

Redundancy Analysis:  
  
Proportion of Variance of Set-1 Explained by Its Own Can. Var.  
               Prop Var  
CV1-1              .723  
CV1-2              .277  
  
Proportion of Variance of Set-1 Explained by Opposite Can.Var.  
               Prop Var  
CV2-1              .412  
CV2-2              .024  
  
Proportion of Variance of Set-2 Explained by Its Own Can. Var.  
               Prop Var  
CV2-1              .680  
CV2-2              .320  
  
Proportion of Variance of Set-2 Explained by Opposite Can. Var.  
               Prop Var  
CV1-1              .387  
CV1-2              .027  
  
------ END MATRIX -----

1. You will also get the scores for participants – these values are the latent scores (the yhat values) for each latent variable.
   1. S1\_CV001 = canonical variate 1, the first set of variables (IV).
   2. S2\_CV001 = canonical variate 1, the second set of variables (DV).
   3. S1\_CV002 = canonical variate 2, the first set of variables (IV).
   4. S2\_CV002 = canonical variate 2, the second set of variables (DV).



**Example Write Up:**

**Results**

This study examined the relationship between the latent variables of meaning and resiliency, as measured by the Purpose in Life, Meaning in Life Questionnaire (meaning), Resiliency Scale 14, and the Brief Resiliency scale (resiliency). The overall scores for these variables were screened for missing data, accuracy, outliers, and assumptions. Seven outliers were removed using Mahalanobis distance as a criterion. Other assumptions were found to be met (linearity, normality, and homoscedasticity).

Canonical correlation was used to examine the overlap between meaning and resiliency. The PIL and MLQ were used to measure meaning, while the RS14 and BRS were used to measure resiliency. Table 1 shows the correlations between all variables. The first canonical variate was significant, *X2*(4) = 545.45, *p* < .001, *r* = .76, indicating that meaning and resiliency were strongly related. The second canonical variate was also significant, *X2*(4) = 545.45, *p* < .001, *r =* .29, explaining another 9% of variance in meaning and resiliency.

Standardized loadings for the variates are in Table 2. Both the PIL and MLQ measured meaning in the first variate, with a stronger loading for the PIL. However, in the second variate, only the MLQ significantly measured meaning. The resiliency variate was measured by both the RS14 and BRS, with a stronger loading for the RS14 in variate one. In the second variate, only the BRS measured resiliency.

Cross loadings and a redundancy analysis indicated that each latent variate was best measured by its own variables, but that there was overlap between the variables. The cross loadings were fairly large for the first variate (Table 2), but the variates were measured better by their own variables (meaning 72%, resiliency 68%) than the other variables (meaning 41% resiliency 39%). The second variate showed the same pattern with stronger overlap for the proposed variables (meaning 28% resiliency 32%) than the opposite variables (meaning 2% resiliency 3%). Overall, meaning and resiliency were significantly related, but distinct items best measured by the PIL and RS14 in the first variate, and the MLQ and BRS in the second variate.

Table 1

*Correlations between measured variables*

|  |  |  |  |
| --- | --- | --- | --- |
|  | MLQ | RS14 | BRS |
| PIL | .52 | .71 | .51 |
| MLQ |  | .56 | .17 |
| RS14 |  |  | .50 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Set 1 | | Set 2 | |
|  | Meaning | Resiliency | Meaning | Resiliency |
| PIL | -0.82\* | (-0.73) | -0.83 | (-0.07) |
| MLQ | -0.28\* | (-0.54) | 1.14\* | (0.21) |
| RS14 | (-0.75) | -0.91\* | (0.04) | 0.71 |
| BRS | (-0.47) | -0.17\* | (-0.23) | -1.40\* |

*Note*. Starred variables indicate significant loadings when examining canonical loadings. Cross loadings are included in parentheses.