Discriminant Analysis

**Description:**

Discriminant analysis is a multivariate analysis that uses many dependent variables to predict your independent category variables. The purpose of a discriminant analysis is *classification or prediction* of what category people will go into based on their scores on some other variables.

**Definitions/Abbreviations:**

Dependent variables – here your normal dependent measures are actually used to predict your category measures. These variables are still conceptually dependent variables but are being used as independent variables to categorize.

Independent variables – there aren’t really “groups” anymore, but category membership. You want to use a limited number of categories because with many categories results become difficult to interpret.

Discriminant function – you will get one to many prediction equations (categories – 1) on how to predict what group a person will be in. This function will either be significant (can be used to predict group membership) or non-significant (does not determine the difference between group membership).

Function parts:

D = dz1 + dz2 + dz3 …

D = discriminant score – this is used to predict their category membership.

d = discriminant function coefficient – sort of like slope when you predict something, how strong that predictor is in your equation. These numbers are choosen to maximize group differences.

Z = dependent measures – these are z-scored so that you can see which variable has the strongest prediction due to the larger d numbers (SD = 1, Mean = 0).

D ranges from 0 to 1 = group 1 membership

D ranges from -1 to 0 = group 2 membership

Classification function: Raw score equation used to classify you into groups, you are assigned to the group with the highest classification score.

C = Constant + cx1 + cx2 …

C = classification group

c = classification coefficient – raw score number used to predict your group. You would multiple this number times the score on a dependent variable.

X = a person’s individual score on a dependent measure.

Discriminant versus Classification Function:

* Discriminant function is reported because it’s easy to interpret which dependent measures were the best at predicting group membership. This equation is standardized so that you can compare across scales.
* Classification function is used when you want to calculate scores for a person to figure out what group they belong in (especially with 3 or more groups). This function is easier to use because you do not have to z-score the dependent measures, just multiply by the classification coefficient.

**Types of Discriminant Functions:**

Standard / Direct – each predictor is used in the equation at the same time. The predictors are only assigned unique variance and overlapping variance is assigned by SPSS.

Sequential / Hierarchical – each predictor is used in a certain pre-specified order. Each step shows if the new predictor adds something to the classification equation. Variance is assigned at each step to the predictor added.

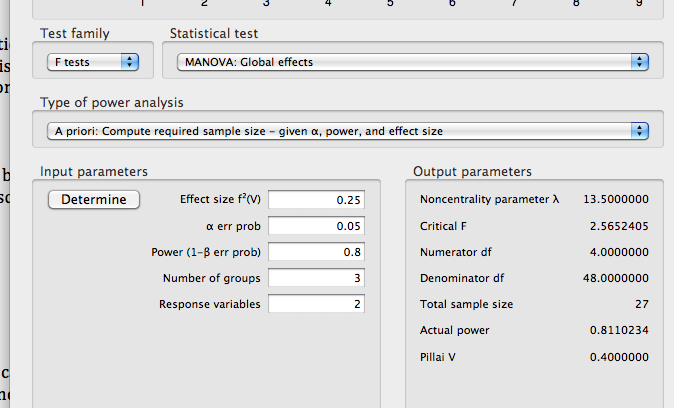
Stepwise / Statistical – similar to a hierarchical discriminant, but predictors are entered in order based on their highest overlap in variance. Variance is assigned at each step to the predictor added.

**Power:**

Power for discriminant analysis is a little tricky, but can be estimated with MANOVA parameters because when MANOVA powers are met, discriminant will also be met.

G\*Power options:

* F-test
* MANOVA: Global Effects
* f = effect size, can convert from eta squared
* alpha = .05
* beta = .80
* Number of groups = number of classifications or categories
* Number of response variables = number of dependent measures.



**Assumptions:**

Small N: Unequal classification/category sizes is will not cause problems in predicting group membership. However, one category with a very small number of people in it will make prediction difficult. The probability of that category is very low, so there’s no function that will distinguish that group – it will lump them into the other groups.

Missing data:

1. Missing dependent measures – these values need to be replaced if possible or eliminate the person. Dependent measures are the most important in predicting membership, so you’ll need something to predict with. There are plenty of replacement options (see data screening notes).
2. Missing category membership – these values are not necessary…but the person will be excluded from the analysis. When discriminant analysis is run, a function is created to predict group membership and then that membership is compared to actual group membership to see how good the function worked. However, you can use significant functions to figure out what category a person should be classified in to fill in this missing data for another analysis.

Normality: This assumption is actually based on linear combinations of the category memberships, which is difficult to test. You do not test normality in the normal residual regression or histogram way since you want the categories to be normal. Therefore, the way to achieve normality is to check for very small categories and to have at least 20 cases in the smallest sample.

Outliers: When trying to predict category membership, outliers are very problematic. You will want to check for univariate and multivariate outliers separated by each category classification.

Homogeneity: Box’s M can be used with p-values > .001 to meet the assumption. You can examine a scatter plot of scores for the discriminant functions for each group.

Linearity: This assumption is a must because a discriminant function *is* a linear regression equation used to predict category membership. Non-linear functions would mean non-significant discriminant functions.

Multicollinearity: When the dependent measures are highly correlated, then using them to predict is using the same scale twice. Only one of the will be important because the other is overshadowed by the first dependent measure. You can average or eliminate variables that are highly correlated.

# Complete Example

In this example, participants were asked if they were working, at home and happy or at home and unhappy. The researchers are trying to predict the difference between the different classifications using marital status, religion, race, and some attitude variables.

Classification/IV(s): Working, Happy at Home, Unhappy at Home

Predictors/DV(s):

Marital – marital status: unmarried, married, divorced

Children – yes there are children, no children

Religion – Protestant, Catholic, Jewish, other

Race – Caucasian or African American

Control – Locus of control, low numbers are internal high numbers are external

Attmar – Attitudes toward current marital status

Attrol – Attitudes toward the role of women

Sel – Socioeconomic status

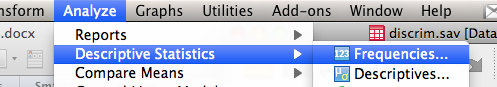
Atthouse – Attitudes towards housework

Educ – Number of years of schooling

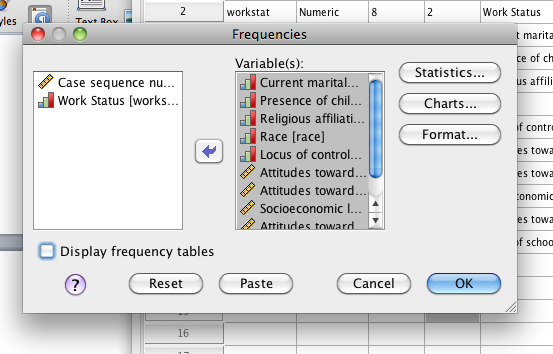
**Research question:** Can we correctly classify people into their classifications using these variables? Which variables are important in classification?

Assumptions Checks:

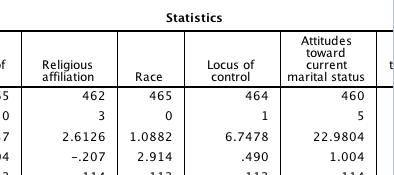
1. Missing data
   1. Analyze > descriptive statistics > frequencies



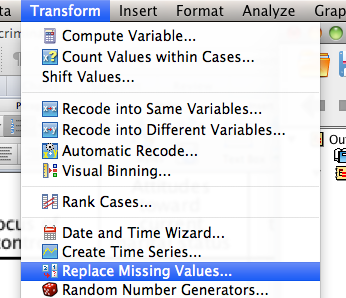
* 1. Move over ALL the dependent variables (your predictor variables).
  2. Pick any options you’d like to see – mean/sd/skew, etc. You can uncheck the frequency table option.



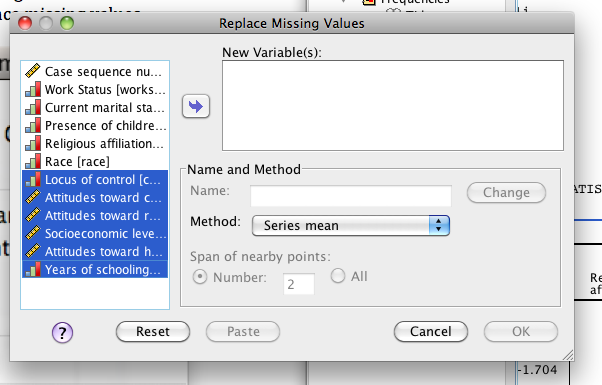
* 1. Look for missing values under missing, and see what type of variable is missing. Religion we probably cannot “mean replace” or guess at to fill in, but locus of control we can fill in as continuous missing data.



* 1. Replace continuous missing data.
     1. Transform > replace missing values.

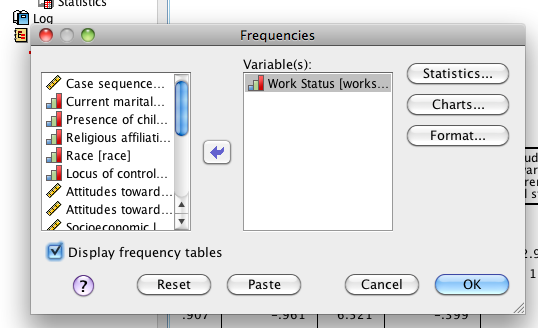


* 1. I’m going to fill in all the continuous data (you have to pick which type of replacement you want before you move them to the right. Hit ok after moving them over.

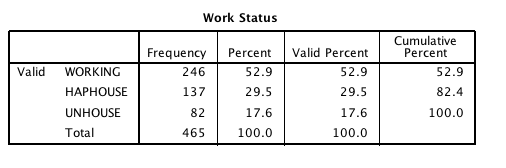


* 1. Now the new variables will be at the end of the data set – be sure to use those in your analysis and not the old ones by accident.
  2. When you have categorical missing data, you usually have to just eliminate the subjects.

1. Small N – this variable you want to check your classification IVs.
   1. Analyze > Descriptive Statistics > Frequencies.
   2. Just move over the classification variable AND ask for the frequency table.

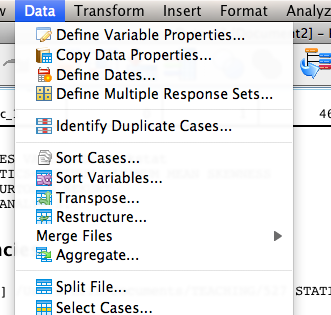


* 1. You want to make sure there’s at least 20 people per classification and no extreme differences in percentages between the classifications (like an 80-20 split).

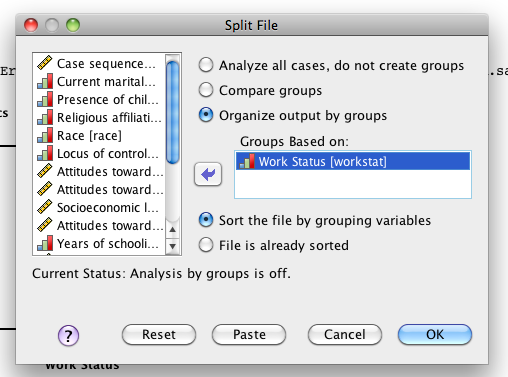


* 1. Here there are is a large category, but it’s not overpoweringly large in comparison to the other categories. They all have at least 20 people.

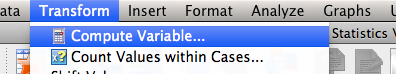
1. Outliers – outliers in this analysis are treated a little differently. You want to run an outlier analysis (the easiest is for multivariate outliers), but separately for each classification.
   1. Split file first. Data > Split file.



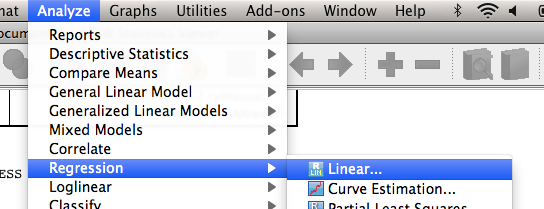
* 1. Organize output by groups > move over your classification IV variable. Hit ok.



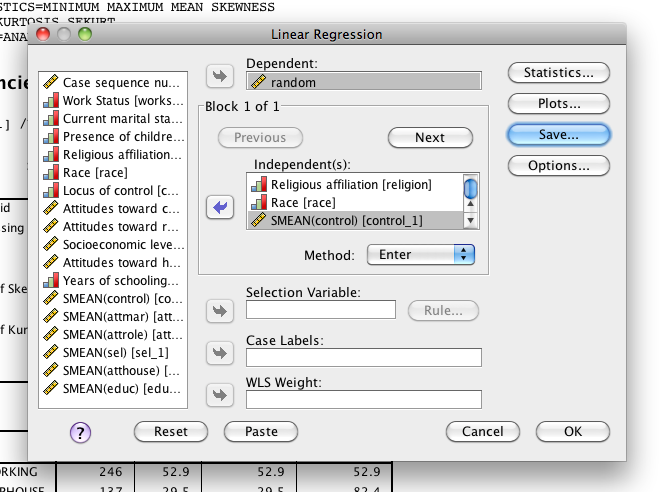
* 1. Create a random variable for a fake regression analysis.
     1. Transform compute variable.

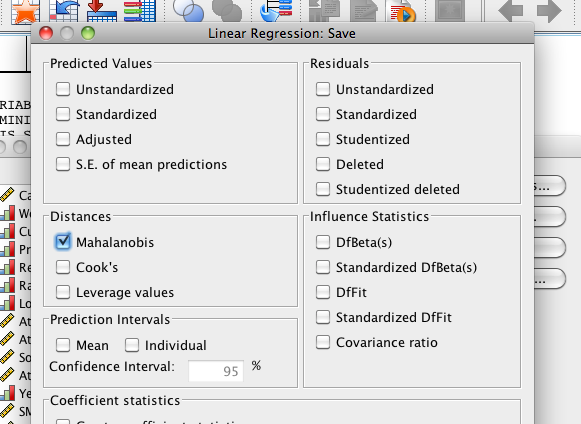


* + 1. Call your variable “random”. Pick random numbers under function group. Use RvChiSq to create a random variable. ? marks need to be filled in with number.
  1. Now run a fake regression and ask for Mahalanobis distance.
  2. Analyze > regression > linear.

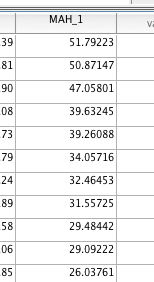


* 1. Random variable in your DV box. DV/Predictor variables in your IV box. Save > Mahalanobis.

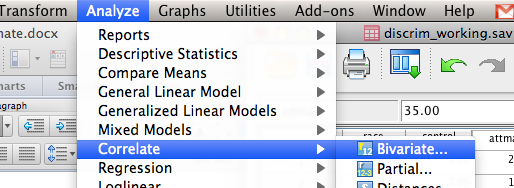




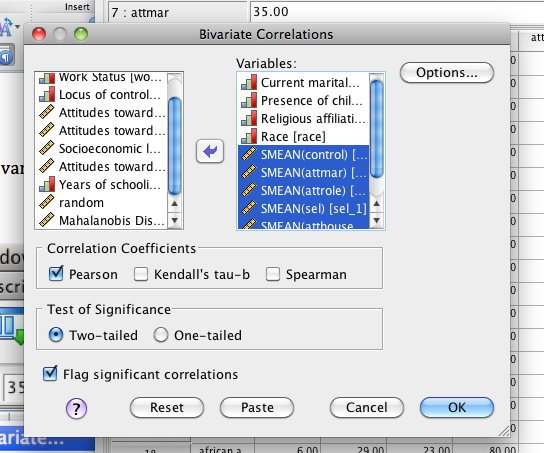
* 1. In the output you’ll see that it ran 3 different regressions, one for each classification. It will on create *one* Mahalanobis variable, but it’s based on which regression they were run on. You check for outliers the same way, but this way they are outliers *within* their classification.
  2. Sort Mahalanobis descending.
  3. Cut off = Chi square with number of variables as the degrees of freedom, p<.001. 10 variables = 29.59.
  4. It appears we have a several outliers, you’ll want to exclude these people.



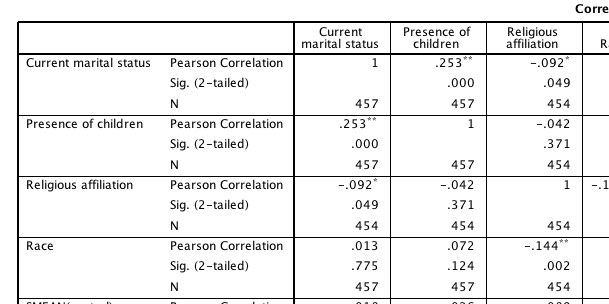
1. Multicollinearity – you’ll want to check your DV/predictor variables to make sure you aren’t using the same ones twice.
   1. Analyze > correlate > bivariate.



* 1. Move over all the DV/predictor variables.

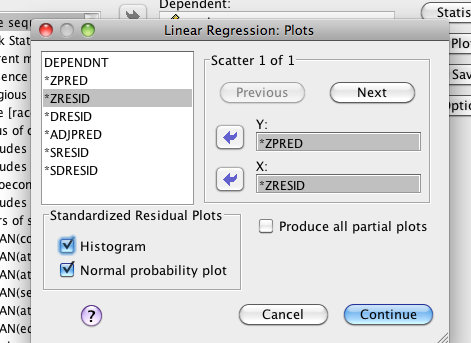


* 1. Look for any correlations above >.9.

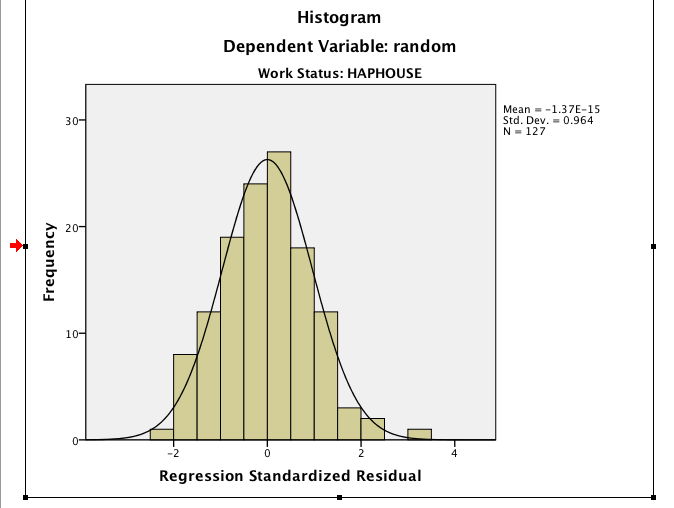
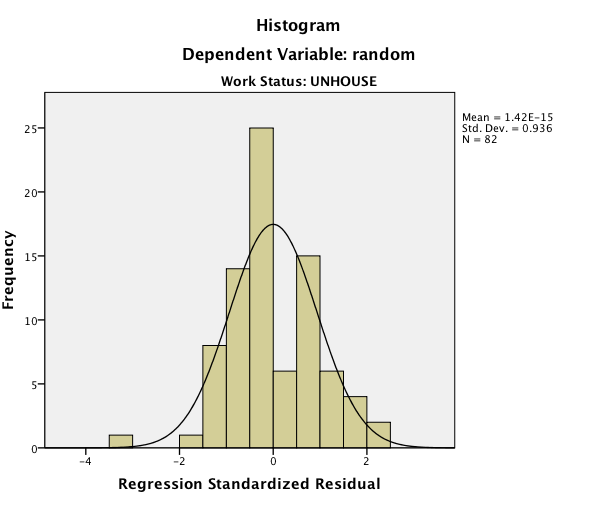
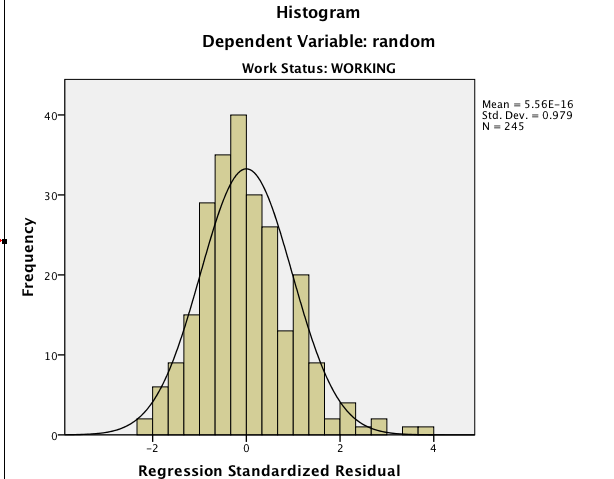


* 1. Combine or only use one of the variables when they are highly correlated. These data seem to be fine.

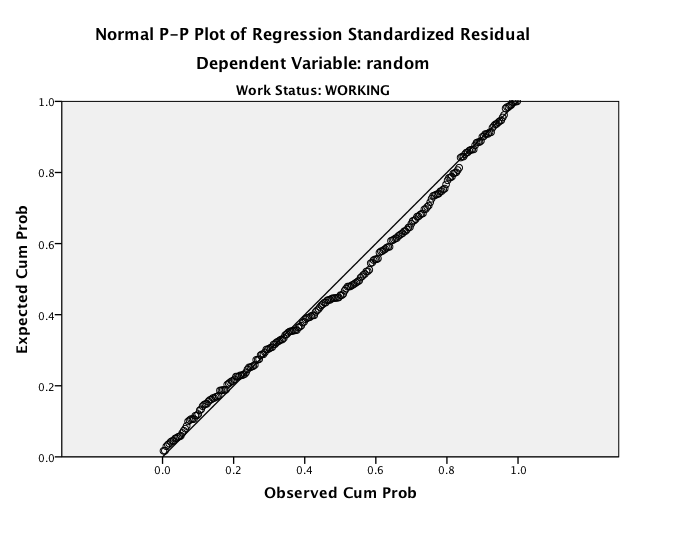
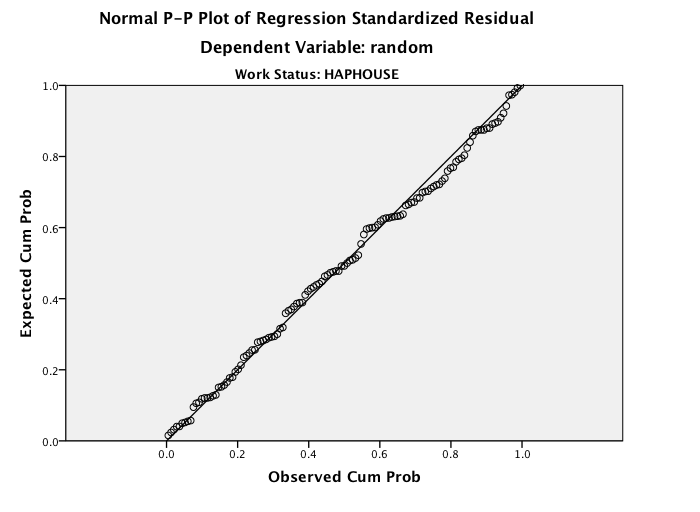
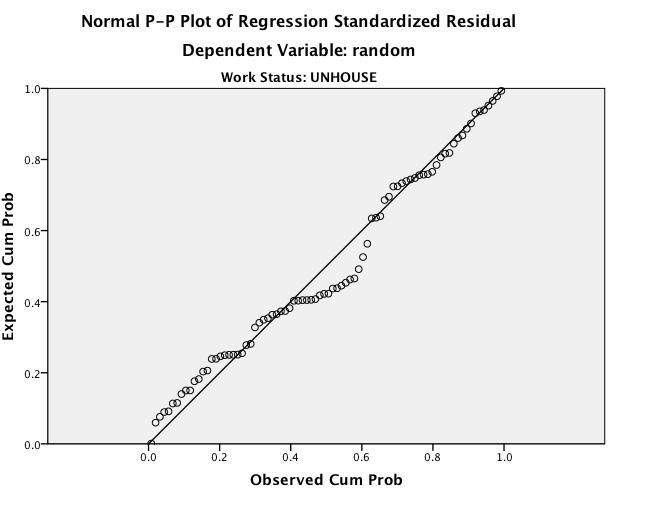
1. Normality/Homogeneity/Linearity – there are not very good ways to check these variables since technically you want the assumptions to be met for the classification variable. However, here’s what people normally do:
   1. Run a fake regression *separated* by classification. If the assumptions are met as you would run a MANOVA, then the assumptions are met for discriminant.
   2. Run the same regression you did earlier for outliers. You will get *three* outputs, one for each classification. You want to look at the histograms, residual and PP plots.
   3. When you run the regression, ask for those plots.
      1. Analyze > regression > linear.
      2. Random into the DV box. DVs/Predictors into the IV box.
      3. Hit plots.
      4. Zpred in Y, Zresid in X. Check the boxes for histogram and normality probability plot.



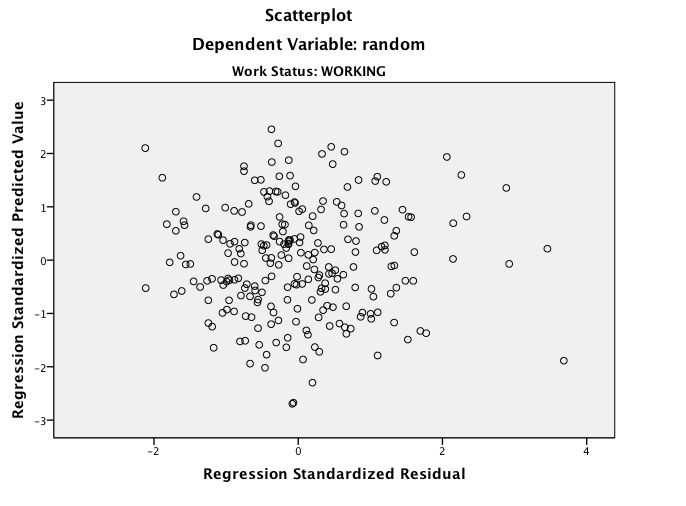
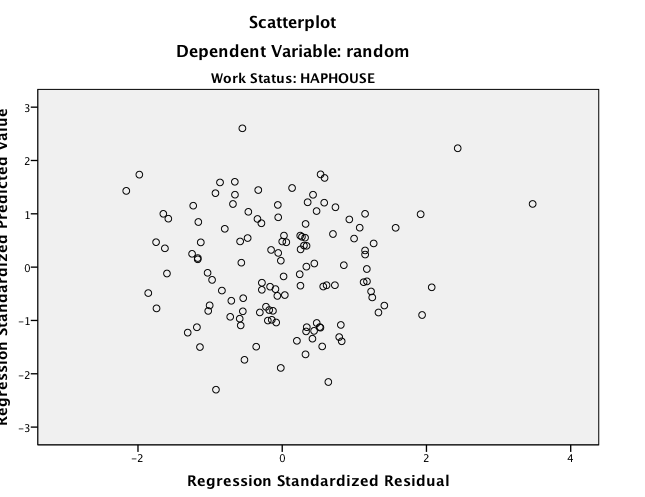
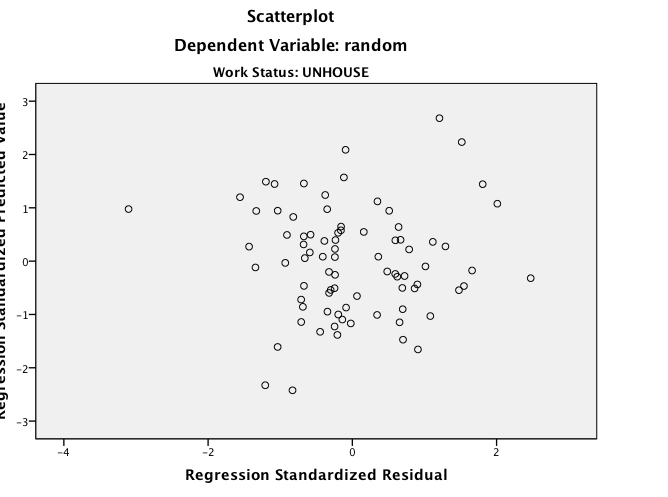
* 1. You will want to look at all three regressions.
  2. Normality:

* 1. Linearity:

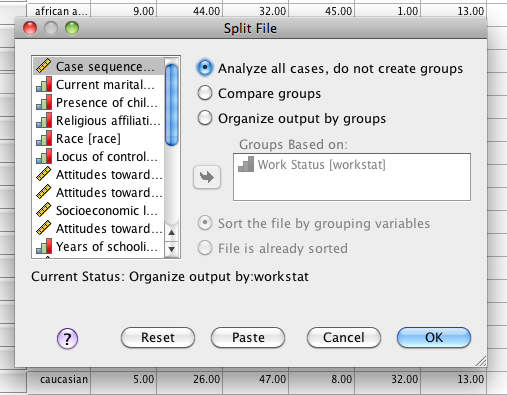
* 1. Homogeneity:

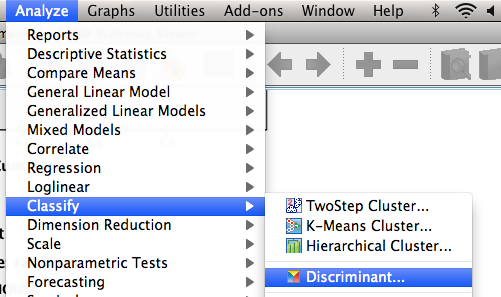
Running the test:

**Important!** Before you run this analysis, be sure to turn OFF the split file.

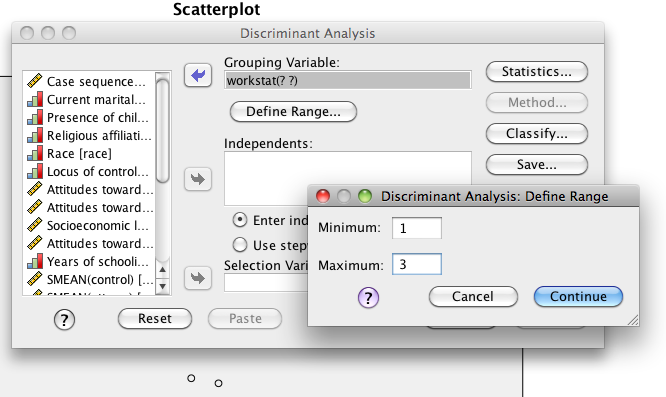
1. Data > split file. Hit analyze all cases, do not create groups.



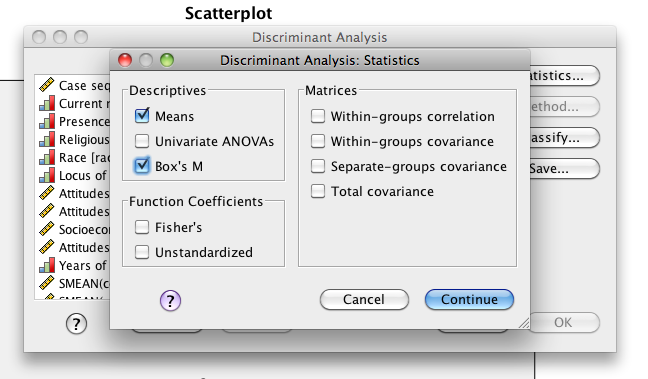
1. Analyze > classify > discriminant.



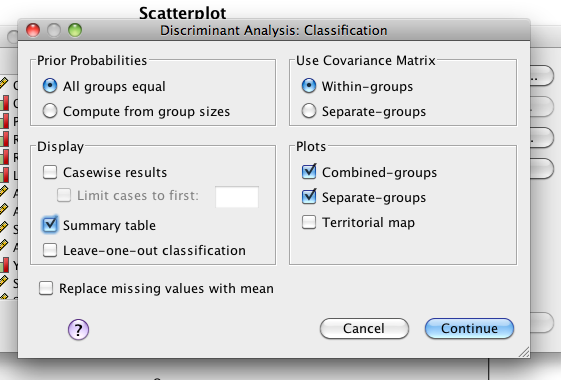
1. Move over the classification/IV into the group variable box. When the ? marks show up hit define range. Type in the highest and lowest numbers that you want to classify (i.e. you can do 1 and 2 only or do 1 2 3 by entering 1 and 3.



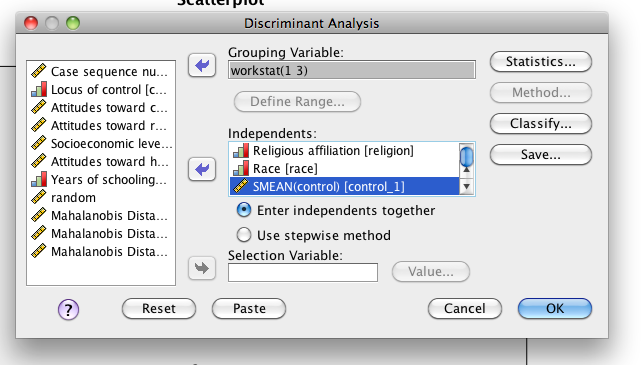
1. Hit statistics. Click “means” for descriptives and Box’s M for homogeneity (just a double check).



1. Hit classify. Pick the summary table, separate groups and combined groups options.

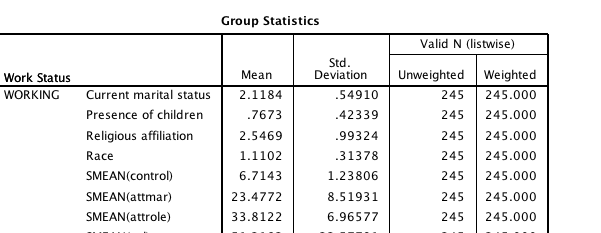


1. Move over your DV/Predictors into the Independents box. Hit ok.

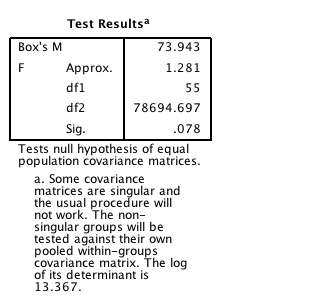


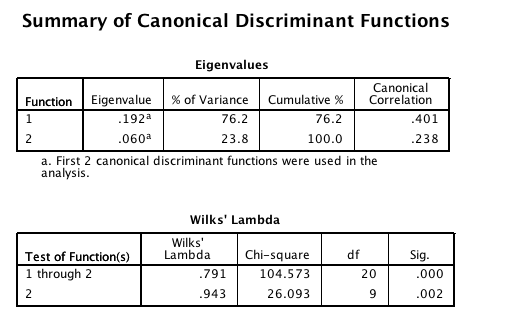
The output:

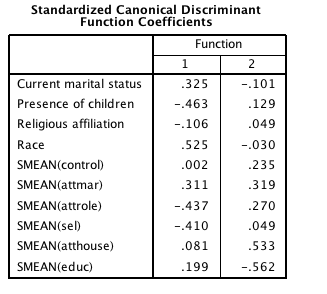
1. The first large box you’ll see is the descriptive statistics separated by group and then a total for the entire dataset regardless of group.



1. The next box is for homogeneity, Box’s M. You want p > .001. Here we are ok because p = .078.

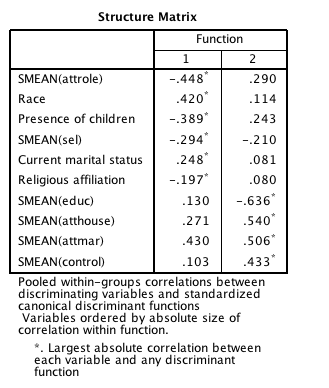


1. Summary of the canonical discriminant functions.
   1. First box covers the canonical correlation (basically a correlation with lots of IVs and lots of DVs instead of one IV and one DV). This box will tell you how much variance is accounted for in your classifications by each discriminant function. There will be classfications – 1 number of functions and the variance will always add to 100.
   2. Wilk’s lambda box – this box tells you which functions were able to significantly predict group membership. There may be several functions, but only one of them may be significant. Here both of our functions can discriminate between classifications.
      1. Function 1 through 2 distinguishes the difference between 1 and 2. Function 2 distinguishes the difference between 2 and 3. The only time classification gets tricky is if function 1 predicts 1 and function 2 predicts 3…that’s when the classification function is used to predict if 1 or 3 is higher.
2. **Standardized** Canonical Discriminant function coefficients – since this label says standardized, these are the little d values for your prediction equation.
   1. **D**  = *d*Z1 + *d*Z2 + *d*Z3…
   2. **D** = .325 (z-current marital status) + -.463 (z-children) + -.106 (z-religion).



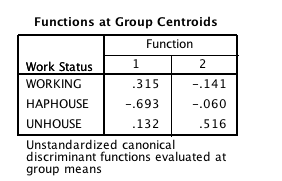
* 1. There are two columns because you have two equations.
     1. Since these are standardized, you can think of them as the higher, the better the predictor (absolute value).
     2. So in the first function (the separation between working and happy house – go back to variable view to figure that out) marital status, children, religion, race, attitude marriage, attitude role of woman, and ses are all important predictors of the difference between working and happy houses.
     3. In the second function, locus of control, attitude marriage, attitude role of women, attitude housework, and education levels predict the difference between happy and unhappy people at home.
     4. Anything positive will put people into the first group, anything negative will put people into the second group.

1. Structure Matrix – this box gives you the correlation between the predictor variables and the discriminant function. It’s a little weird to think about, but this box tells you which variable was *used the most* when predicting.

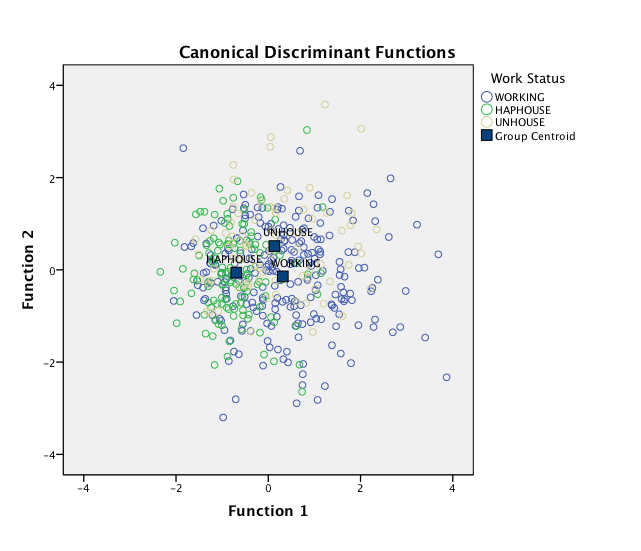


* 1. So even though race had the highest discriminant function for the first equation attitude role of women was the predictor that predicted the most variance. Usually these two boxes will give you the same thing. This box will give you the effect size for each individual variable, while the standardized box gives you a comparison across scales.

1. Functions at Group Centroids – this box is the group mean of means, which is graphed later.

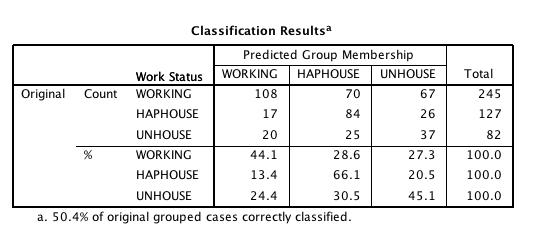


1. The separate graphs are useful and neat looking, but the best graph is the combined groups graph. This graph shows you how the discriminant functions work.



* 1. The first function predicts the difference between working and happy home people.
  2. The second function predicts the difference between happy home and unhappy home people.

1. Classification results – the important box. This box tells you how well you did (and in our case, it’s not very good…only 50%!).



1. The top half the box is the raw frequencies on how close you got. If you were perfect the center diagonal would be numbers and everything else would be zeroes.
2. It appears we did best in the happy house category because 66.1% of them were classified correctly, while only 44 and 45% were classified correctly for the rest.
3. This box is the most common table to make for publication.

Example write up:

Results

A discriminant function analysis was performed using demographic variables and attitude variables to predict working, happy in the home or unhappy in the home work classifications. Demographic variables were race, ethnicity, religion, education level, socioeconomic status, marital status, and the presence of children (yes/no). Attitude variables included attitude towards the role of women, housework, and toward current marital status. Three cases were dropped do to missing data, and 3 multivariate outliers were found using Mahalanobis distance (X2 (10) = 29.59, *p*<.001 as a cut off). Multicollinearity, homogeneity (Box’s M *p*=.078), normality, and linearity all showed to be non-problematic.

Both functions in the discriminant function analysis were significant using Wilks’ Lambda (X2(20) = 104.57, *p*<.001 for function 1 and X2(9) = 26.09, *p*=.002 for function 2). Function 1 predicted the distinction between working participants and participants who were happy being at home. This function showed that participants with low attitudes toward the role of women were more likely to be classified into the working category (*d*=-.44), while positive attitudes toward the role of woman were more likely to be classified in happy homes. Caucasian participants were more likely to be classified as working (*d*=.53). The presence of children actually was predictive for working participants, and individuals who were happy at home did not have children (*d*=-.46). Current marital status showed that divorced participants were more likely to be predicted as working (*d=*.33). Lastly, religion was a predictor of this distinction in that Protestant and Catholic families were more likely to be in happy homes, while Jewish and other religions were more likely to be working (*d*=-.11).

The second function distinguished the difference between participants who were happy at home and participants were who unhappy at home. Education was the strongest predictor of happiness in the home (*d*=-.56), so that participants with more education were more likely to be happy. Increasing attitudes toward housework was more likely to predict unhappy homes (*d*=.53). As the attitude toward current marital status increased, participants were more likely to be classified as in a happy home (*d*=.32). Lastly, locus of control predicted happier homes, in that external focused individuals were more likely to be in a happy home (*d*=.24). Combined, both functions predicted 50.4% of participants into their correct categories. Happy homes were most correctly classified at 66.1%, followed by unhappy homes 45.1% and working homes 44.1%. Please see Table 1 for percentages.

Table 1. *Classification Results for Discriminant Functions*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Working | Happy House | Unhappy House |
| Working | **44.1** | 28.6 | 27.3 |
| Happy House | 13.4 | **66.1** | 20.5 |
| Unhappy House | 24.4 | 30.5 | **45.1** |

*Note*. Correct classifications have been bolded.