**Andrew Hillard**

**Data 2**

**Homework 7**

Problem 1

1. Two factors are not sufficient. The test of the null hypothesis that 2 factors are sufficient versus the alternative hypothesis that more factors are needed results in a p-value of <.0001.
2. Three factors are sufficient for the MLE Factor Analysis with varimax rotation. The significance test of the null hypothesis that three factors are sufficient versus the alternative hypothesis that more factors are needed results in a p-value of 0.11.

An interpretation of Factor 1 is that it describes the latent variable of the belief that doctor’s control your pain. This matches the factor pattern in SAS. Statements that claim that the person is in control of their health have negative pattern values and statements that claim the doctor is in control of their health have positive pattern values.

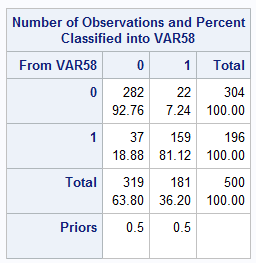
An interpretation of Factor 2 is that it describes the latent variable that you belief that you are in control of your pain. Statements that suggest you have more control have positive values while statements that claim the doctor is in control of your pain have more negative values.

An interpretation of Factor 3 is pretty unclear though may have something to do with a person’s belief that their pain is controlled by more recent actions.

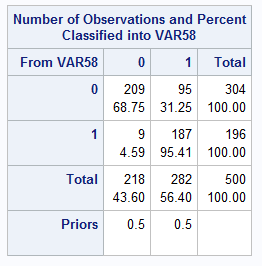
1. The interpretation of Factor 1 and 2 is the same as the interpretation using the MLE / Varimax analysis. Factor 2 is most positively associated with statement two and five. Factor 3 is also different than the MLE / Varimax analysis. Factor 3 also tends to follow a pattern around the belief of whether you believe you can control your pain, but is most positively associated with statement six and seven. Since the three factors have similar interpretations, it’s likely that having less than three factors is appropriate. When we run the principle factor method without the constraint of having three factors, only two factors are selected, which supports our hunch.
2. The interpretation of Factor 1 is the same as in the previous parts, your ratings of statements is positively associated with your belief that doctor’s control your pain. However, the factor pattern is more pronounced than previous parts, likely due to the oblique rotation. Factor 2 is also interpreted similarly, and is correlated with your belief that you, not your doctor, are responsible for your pain. Statements about you being in control of your pain are positively correlated with Factor 2. Factor 3 follows a similar pattern as Factor 2, though is more pronounced for some statements. Factor 2 can be interpreted as being positively correlated with your belief that you are in control. Factor 2 and 3 seem to be complimentary in their correlations to Factor 1.

Problem 2

1. The covariance matrices are not equal for spam and not spam. This conclusion was reached using the Bartlett’s test with PROC DISCRIM in SAS. The p-value was <.0001. The implication of this conclusion is that it’s appropriate to use the quadratic discriminant rule in our discriminant analysis. If the covariance matrices had been equal, the linear discriminant rule would have been more fitting.
2. Below is the classification summary for the test dataset. The discriminant analysis did a good job at classifying not spam and a decent job at classifying spam. For the test data, 92.76% of not spam (0) was classified correctly. Only 81.12% of spam (1) was classified correctly.

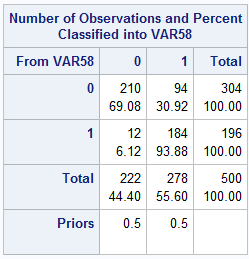


1. Below is the classification summary for the test dataset. The results for the quadratic discriminant analysis were almost opposite what they were for the linear discriminant analysis. The analysis did a good job at classifying “spam,” but not “not spam.” For the test data, 68.75% of not spam (0) was classified correctly. 95.41% of spam was classified correctly.

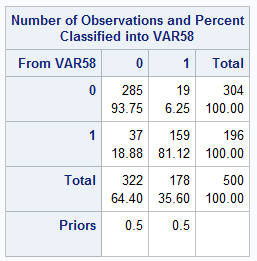


1. I included the following variables based on a forward selection method with entry criteria of significance level .01. This left me with the following 36 attributes: 21, 23, 7, 57, 16, 52, 25, 5, 53, 8, 24, 6, 20, 18, 22, 42, 27, 46, 45, 49, 19, 56, 17, 12, 37, 33, 26, 44, 48, 43, 4, 9, 3, 1, 11, and 35.

The quadratic discriminant analysis produced similar results with the subset of variables. See the cross validation matrix for test data below.

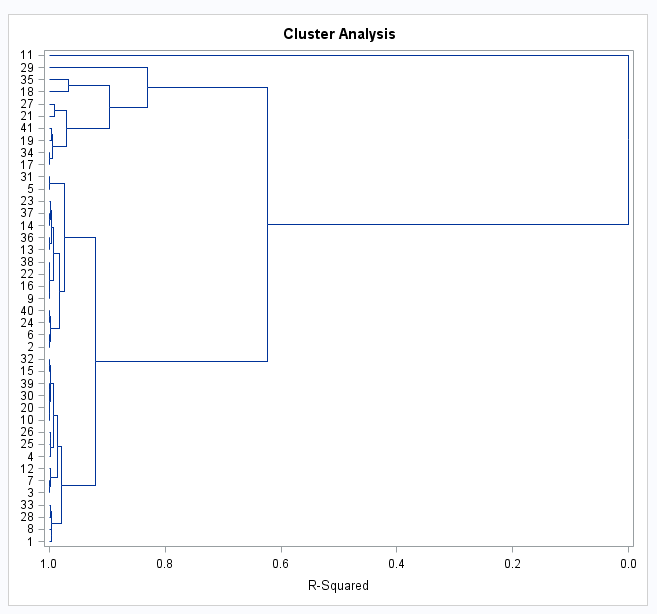


Likewise, the linear discriminant analysis performed about the same with the subset of variables. See cross-validation matrix for test data below.



Problem 3

1. The dendrogram is given below. Based on the plot, and the Pseudo t-squared statistic, I would tentatively say that a cluster size of 5 is appropriate. The dendrogram suggests a clustering of three would be reasonable, as three general groups seem present in the plot. However, the t-squared statistic spikes at 5, which would indicate a cluster group of 6 as reasonable. A cluster size of 5 is somewhere in between. That being said, it is very difficult to determine based on the plot and pseudo statistic.



1. The cluster categories are significant in the 1-way ANOVA and the test produces a p-value of =0.0004. The interpretation is that the cities are clustered into groups that have significant differences in SO2 values. Furthermore, based on the plot below, it look as if clusters 3 and 4 have similar SO2 values, and clusters 1 and 2 each have different SO2 values then the other clusters.

