***STAT 6300 Final Project Report:***

***Multivariate Analysis for the Iris Flower***



*(\*from left to right- Iris Setosa, Iris Virginica and Iris Versicolour)*

Professor name: Dr. Nancy S. Boudreau

Student name: Jing-yi Wu and Yue Zhang

Date: Dec. 7.2014

**1.Introduction:**

The Iris flower data set is a famous multivariate data set which was collected by Ronald Fisher in 1936. It is a good example of a data set that responds well to discriminant analysis and cluster analysis. All data were collected from 3 species of Iris (Iris Setosa, Iris Versicolour and Iris Virginica). After collection, this data set has been analyzed in many different perspectives. We want to investigate whether we can distinguish between the 3 species of Iris. If we can find a way to distinguish between the different lengths and widths of petals and sepals, it will become easier to categorize these three species of flower.

**2. Data collection:**

Our data set is a second-hand data set, which was downloaded from <https://archive.ics.uci.edu/ml/datasets/Iris>. The website is the Machine Learning Repository of University of California. The data set was donated on July 1st, 1988, and does not contain missing values. On each level of Iris flower, we have 50 instances. Four variables were measured on each observation: the length of sepals, the length of petals, the width of sepals and the width of petals, in centimeters. In total we have 150 observations with 4 variables and 3 groups.

**3. Methods**

3.1 Methodology

We will use Multivariate one-way MANOVA, one-way ANOVA and factor analysis to investigate whether the three species of Iris Flower have significant differences in the four variables.

3.2 Model

Where

*:* the response of group at level i and variable at level j

*:* the overall mean

: effect of ith level of group on the overall mean vector

*:* the error of group at level i and variable at level j

3.3 One Way MANOVA Table

|  |  |  |  |
| --- | --- | --- | --- |
| Sum of Variation | Degrees of Freedom | Sum of Squares Error | Wilk’s |
| Treatment(Between) | K | **H** |  |
| Error(Within) | k(n-1) | **E** |  |
| Total | kn-1 | **H**+**E** |  |

**4. Results:**

For this case, the one-way MANOVA has one factor and 4 dependent variables:

Factor ---Species--- 3 levels: 1= *Type 1(Iris Setosa)*; 2= *Type 2 (Iris Versicolour)*; 3= *Type 3 (Iris Virginica).*

The 4 dependent variables:

y1= length of sepals; y2= length of petals;

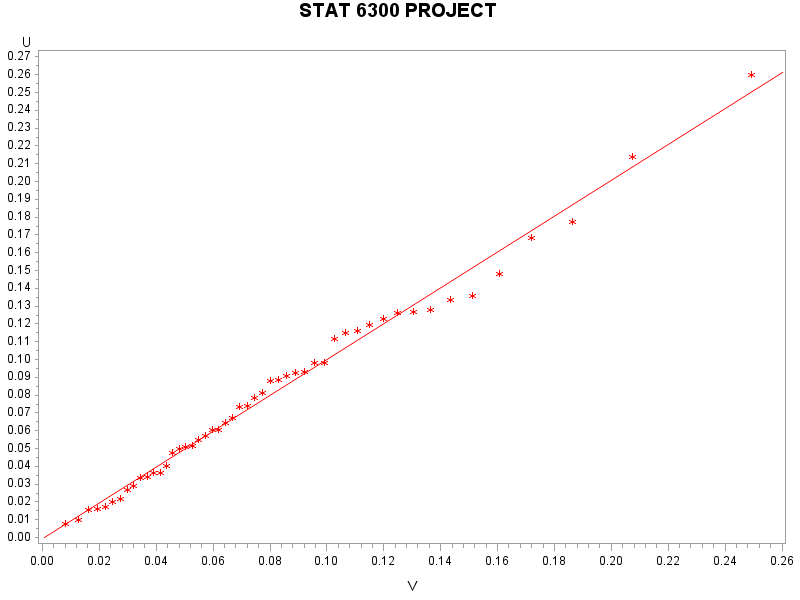
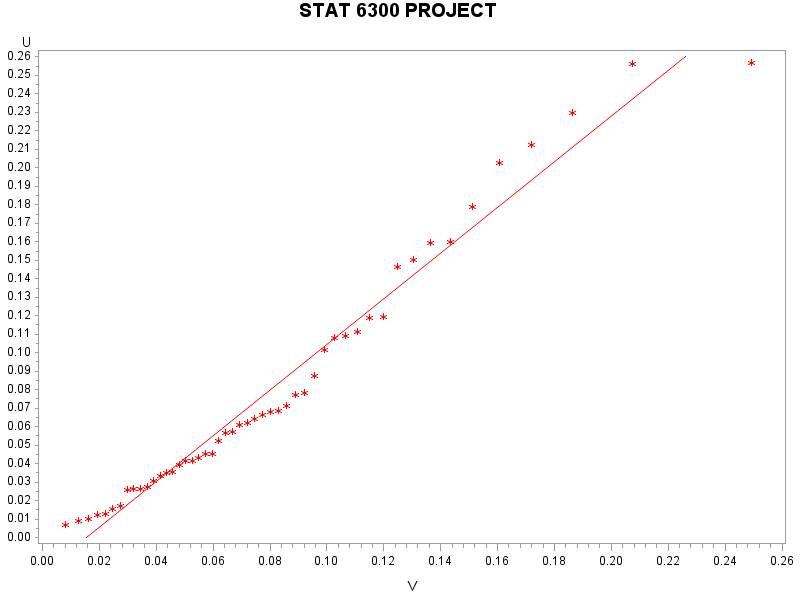
y3= width of sepals; y4= width of petals

Assume: ~

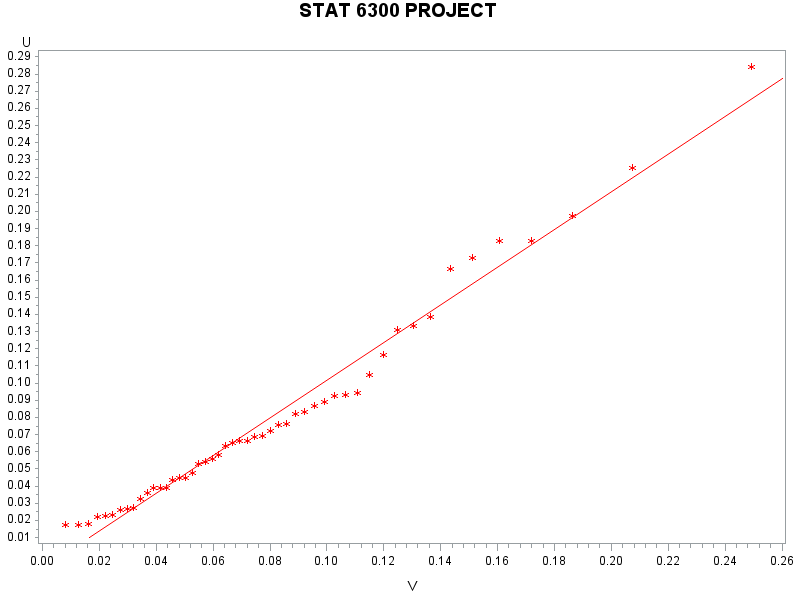
4.1 Normality check and Outlier Detection

We need to check the multivariate normality assumption for type 1 (Iris Setosa), type 2 (Iris Versicolour) and type 3 (Iris Virginica).

Type 1(Iris Setosa) Type 2 (Iris Versicolour)



Type 3 (Iris Virginica)



The three Q-Q plots based on the beta distribution indicate that the data in the three groups are all multivariate Normal, as upon visual inspection, the sample quantiles do not deviate from the theoretical quantiles.

Test for potential outliers:

|  |  |  |
| --- | --- | --- |
| Group | T.S. | R.R. |
| Iris Setosa | 12.3276 | Reject if > 15.89 |
| Iris Versicolour | 12.4895 | Reject if > 15.89 |
| Iris Virginica | 13.6691 | Reject if > 15.89 |

*(\*The SAS Output is in the appendix #1)*

Conclusion:

With and , the critical value is . Since the max , there is insufficient evidence to conclude the existence of an outlier in group 1.

With and , the critical value is . Since the max , there is insufficient evidence to conclude the existence of an outlier in group 2.

With and , the critical value is . Since the max , there is insufficient evidence to conclude the existence of an outlier in group 3.

In conclusion, our data are multivariate normal and there is no trace of outliers.

4.2 One-Way MANOVA :

We used the model to analyze our data.

Based on our SAS Output, we have the following results.

Test for the equal mean vector:

| **Characteristic Roots and Vectors of: E Inverse \* H, where H = Type III SSCP Matrix for GRP E = Error SSCP Matrix** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Characteristic Root** | **Percent** | **Characteristic Vector V'EV=1** | | | |
| **Y1** | **Y2** | **Y3** | **Y4** |
| **32.1919292** | **99.12** | -0.06840592 | -0.12656121 | 0.18155288 | 0.23180286 |
| **0.2853910** | **0.88** | 0.00198791 | 0.17852670 | -0.07686357 | 0.23417227 |
| **0.0000000** | **0.00** | 0.10268742 | -0.19415509 | -0.22502879 | 0.37627520 |
| **0.0000000** | **0.00** | -0.24194505 | 0.10603485 | 0.10535376 | 0.00000000 |

Since the first eigenvalue is large (99.12%) and others are small, it should be 1 dimensional. Thus, we use Roy's test.

| **MANOVA Tests for the Hypothesis of No Overall GRP Effect H = Type III SSCP Matrix for GRP E = Error SSCP Matrix  S=2 M=0.5 N=71** | | |
| --- | --- | --- |
| **Statistic** | **Value** | **P-Value** |
| **Wilks' Lambda** | 0.02343863 | <.0001 |
| **Pillai's Trace** | 1.19189883 | <.0001 |
| **Hotelling-Lawley Trace** | 32.47732024 | <.0001 |
| **Roy's Greatest Root** | 32.19192920 | <.0001 |

***Hypothesis:***

***Test Statistics:*** , p-value<0.0001,

***Decision rule:*** if p-value< , we reject null hypothesis.

***Conclusion:*** Since p-value <0.0001<0.05, we reject the null hypothesis. We conclude that there is sufficient evidence to indicate a significant difference in mean among 3 Types of Iris at α=0.05.

4.3 Univariate Test

We need to perform the univariate test for 4 variables: length of sepals, the length of petals, the width of sepals and the width of petals.

(1) For Y1 Length of Sepals:

| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **GRP** | 2 | 63.21213333 | 31.60606667 | 119.26 | <.0001 |

***Hypothesis:***

***Test Statistics:*** , p-value<0.0001,

***Decision rule:*** if p-value < , reject null hypothesis.

***Conclusion:*** Since p-value <0.0001<0.05, we reject the null hypothesis. We conclude that sufficient evidence to indicate difference among the mean of Length of Sepals at =0.05.

(2)For Y2 Length of Petals variable:

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **GRP** | 2 | 11.34493333 | 5.67246667 | 49.16 | <.0001 |

***Hypothesis:***

***Test Statistics:*** , p-value<0.0001,

***Decision rule:*** if p-value < , reject null hypothesis.

***Conclusion:*** Since p-value <0.0001<0.05, we reject the null hypothesis. We conclude that sufficient evidence to indicate difference among the mean of Length of Petals at =0.05.

(3) For Y3 Width of Sepals variable:

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **GRP** | 2 | 437.1028000 | 218.5514000 | 1180.16 | <.0001 |
|  |  |  |  |  |  |

***Hypothesis:***

***Test Statistics:*** , p-value<0.0001,

***Decision rule:*** if p-value < , reject null hypothesis.

***Conclusion:*** Since p-value <0.0001<0.05, we reject the null hypothesis. We conclude that sufficient evidence to indicate difference among the mean of Width of Sepals at =0.05.

(4) For Y4 Width of Petals:

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **GRP** | 2 | 80.41333333 | 40.20666667 | 960.01 | <.0001 |
|  |  |  |  |  |  |

***Hypothesis:***

***Test Statistics:*** , p-value<0.0001,

***Decision rule:*** if p-value < , reject null hypothesis.

***Conclusion:*** Since p-value <0.0001<0.05, we reject the null hypothesis. We conclude that sufficient evidence to indicate a difference among the mean of Width of Petals at =0.05.

4.4 Multiple comparison (Tukey’s Method)

We need to do a further analysis to know where the differences are. We will use Tukey’s multiple comparisons.

1. For Y1 (length of sepals):

| **Means with the same letter are not significantly different.** | | | |
| --- | --- | --- | --- |
| **Tukey Grouping** | **Mean** | **N** | **GRP** |
| A | 6.5880 | 50 | 3 |
|  |  |  |  |
| B | 5.9360 | 50 | 2 |
|  |  |  |  |
| C | 5.0060 | 50 | 1 |

The mean length of sepals for the Iris Virginica is significantly higher than the mean length of sepals for the other two types of flower (Iris Versicolour ,Iris Setosa); and the mean length of sepals for Iris Versicolour is significantly higher than the mean of the length of sepals for Iris Setosa. Therefore, we can distinguish the three types of Iris flower by the length of sepals, the longest one is Iris Virginica, the Iris Versicolour is second longest, and the length of Iris Setosa is the shortest.

1. For Y2 (length of petals):

| **Means with the same letter are not significantly different.** | | | |
| --- | --- | --- | --- |
| **Tukey Grouping** | **Mean** | **N** | **GRP** |
| A | 3.42800 | 50 | 1 |
|  |  |  |  |
| B | 2.97400 | 50 | 3 |
|  |  |  |  |
| C | 2.77000 | 50 | 2 |

The mean length of petals for the Iris Setosa is significantly higher than the mean length of petals for the other two types of flower (Iris Versicolour ,Iris Virginica); and the mean length of petals for Iris versicolour is significantly higher than the mean of the length of sepals for Iris Virginica. Therefore, we can distinguish the three types of Iris flower by the length of petals: the longest one is Iris Setosa, the Iris Virginica is followed, the length of Iris Versicolour is the shortest.

1. For Y3 (width of sepals):

| **Means with the same letter are not significantly different.** | | | |
| --- | --- | --- | --- |
| **Tukey Grouping** | **Mean** | **N** | **GRP** |
| A | 5.55200 | 50 | 3 |
|  |  |  |  |
| B | 4.26000 | 50 | 2 |
|  |  |  |  |
| C | 1.46200 | 50 | 1 |

The mean width of sepals for the Iris Virginica is significantly higher than the mean width of sepals for the other two types of flower (Iris Versicolour ,Iris Setosa); and the mean width of sepals for Iris versicolour is significantly higher than the mean of the width of sepals for Iris Setosa. Therefore, we can distinguish the three types of Iris flower by the width of sepals: the widest one is Iris Virginica, the Iris Versicolour is less wider, and Iris Setosa has the narrowest width of sepals.

1. For Y4 (width of petals):

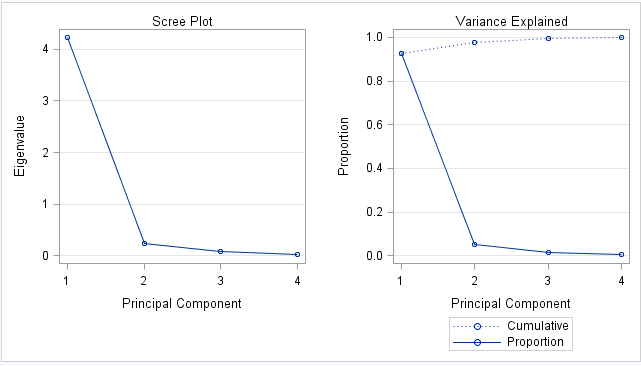
| **Means with the same letter are not significantly different.** | | | |
| --- | --- | --- | --- |
| **Tukey Grouping** | **Mean** | **N** | **GRP** |
| A | 2.02600 | 50 | 3 |
|  |  |  |  |
| B | 1.32600 | 50 | 2 |
|  |  |  |  |
| C | 0.24600 | 50 | 1 |

The mean width of petals for the Iris Virginica is significantly higher than the mean width of petals for the other two types of flower (Iris Versicolour ,Iris Setosa); and the mean width of petals for Iris versicolour is significantly higher than the mean of the width of petals for Iris Setosa. Therefore, we can distinguish the three types of Iris flower by the width of petals, the widest one is Iris Virginica, the Iris Versicolour is followed, Iris Setosa has the narrowest width of petals.

4.5 Principal component analysis:

We use principal component analysis to maximize the variance of a linear combination of variables. We first check the correlation among the 4 variables. *(\*The SAS Output is Appendix 4)* indicates that Y1 has high correlation with Y3 and Y4, and Y3 has high correlation with Y4. Therefore, we can use principal components analyze our data.

| **Eigenvalues of the Covariance Matrix** | | | | |
| --- | --- | --- | --- | --- |
|  | **Eigenvalue** | **Difference** | **Proportion** | **Cumulative** |
| **1** | 4.22824171 | 3.98557096 | 0.9246 | 0.9246 |
| **2** | 0.24267075 | 0.16446125 | 0.0531 | 0.9777 |
| **3** | 0.07820950 | 0.05437441 | 0.0171 | 0.9948 |
| **4** | 0.02383509 |  | 0.0052 | 1.0000 |



We use the first 3 rules to decide the number of principal components to keep.

We can see that the first eigenvalue takes up more than 85% out of total. The average of the eigenvalues is 1.143, and only one eigenvalue greater than the mean. Also, from the Scree plot, we keep all PC's to that left of the elbow. That would be the first one.

| **Eigenvectors** | | | | |
| --- | --- | --- | --- | --- |
|  | **Prin1** | **Prin2** | **Prin3** | **Prin4** |
| **Y1** | 0.361387 | 0.656589 | -.582030 | 0.315487 |
| **Y2** | -.084523 | 0.730161 | 0.597911 | -.319723 |
| **Y3** | 0.856671 | -.173373 | 0.076236 | -.479839 |
| **Y4** | 0.358289 | -.075481 | 0.545831 | 0.753657 |

The coefficient of the principal component of Y3 is 0.856671, which is much larger than others. So Y3 dominates the first principal component. Additionally, we can see Y2 is negative and very close to zero in first component.

4.6 Factor Analysis

In Factor analysis, we represent our variables as linear combinations of a few random factors f1, f2, f3. The assumptions are:

1.  for all *j* and for all *i* ≠ *j*.
2. for all *i* and for all *i* ≠ *j*.
3. for all *i* and *j*.
4. 

| **Rotated Factor Pattern** | | |
| --- | --- | --- |
|  | **Factor1** | **Factor2** |
| **Y1** | 0.95940 | 0.04635 |
| **Y2** | -0.14254 | 0.98519 |
| **Y3** | 0.94357 | -0.30562 |
| **Y4** | 0.93190 | -0.25853 |

Therefore, we can use factor analysis to reanalyze our data and check our answer.

| **Eigenvalues of the Correlation Matrix: Total = 4 Average = 1** | | | | |
| --- | --- | --- | --- | --- |
|  | **Eigenvalue** | **Difference** | **Proportion** | **Cumulative** |
| **1** | 2.91849782 | 2.00446735 | 0.7296 | 0.7296 |
| **2** | 0.91403047 | 0.76727360 | 0.2285 | 0.9581 |
| **3** | 0.14675688 | 0.12604204 | 0.0367 | 0.9948 |
| **4** | 0.02071484 |  | 0.0052 | 1.0000 |

From the SAS output, we can find that there are only 1 eigenvalue greater than 1. But the second eigenvalue 0.914 is very close to 1. SAS automatically chose two factors for us. From rotated factor pattern, Y1, Y3, and Y4 are highly correlated with factor 1 with correlation 0.9594, 0.94357, and 0.93190. Y2 and factor 2 are highly correlated with correlation 0.98519. So we put y1(length of sepals), y3(width of sepals) and y4(width of petals) into Factor 1, and we put y2(length of petals) into Factor2. The Scree plot also indicates that the elbow is at 3, we should keep 2 factors.



For Factor 1:y1=length of sepals, y3= width of sepals and y4= width of petals

Factor 2: y2=length of petals

The correlation of factor 1 and 2 is 0 *(\*appendix # 4)*, which indicates the two factors are independent. Now we analyze the data with these two factors.

| **MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall GRP Effect H = Type III SSCP Matrix for GRP E = Error SSCP Matrix  S=1 M=0 N=72.5** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Statistic** | **Value** | **F Value** | **Num DF** | **Den DF** | **Pr > F** |
| **Wilks' Lambda** | 0.10746795 | 610.42 | 2 | 147 | <.0001 |
| **Pillai's Trace** | 0.89253205 | 610.42 | 2 | 147 | <.0001 |
| **Hotelling-Lawley Trace** | 8.30509968 | 610.42 | 2 | 147 | <.0001 |
| **Roy's Greatest Root** | 8.30509968 | 610.42 | 2 | 147 | <.0001 |

It is two dimensional so we use Roy’s test.

***Hypothesis:***

***Test Statistics:***F=610.42 , p-value<0.0001,

***Decision rule:*** if p-value< , reject null hypothesis.

***Conclusion:*** Since p-value <0.0001<0.05, we reject the null hypothesis. We conclude that there is sufficient evidence to indicate a difference among 3 types of Iris at α=0.05.

The univariate tests of the two factors both reject the null hypothesis, which indicates we need perform multiple comparisons to distinguish the three types of Iris flowers. *(\*The SAS output for the univariate test is in the appendix #5)*

Factor 1 Factor 2

| **Means with the same letter are not significantly different.** | | | |
| --- | --- | --- | --- |
| **Tukey Grouping** | **Mean** | **N** | **GRP** |
| A | 1.04164 | 50 | 3 |
|  |  |  |  |
| B | 0.08379 | 50 | 2 |
|  |  |  |  |
| C | -1.12544 | 50 | 1 |

| **Means with the same letter are not significantly different.** | | | |
| --- | --- | --- | --- |
| **Tukey Grouping** | **Mean** | **N** | **GRP** |
| A | 0.7132 | 50 | 1 |
|  |  |  |  |
| B | -0.0762 | 50 | 3 |
|  |  |  |  |
| C | -0.6370 | 50 | 2 |

Factor 1 :y1=length of sepals, y3= width of sepals and y4= width of petals

We can conclude for these three variables Iris Virginca has the highest mean, Iris Versicolour is in the second place, and Iris Setosa has the smallest mean.

Factor 2: y2=length of petals

Iris Setosa has the highest mean Iris Virginca is in the second place, Iris Versicolour has the smallest mean in length of petals.

The result from factor analysis is the same as our previous result based on the real variables, which makes sense.

4.7 Discriminant Analysis

Our goal is to find group separation and discriminate functions of variables since those are useful to describe the difference between the 3 groups.

Now we standardize the data to get the standardized discriminate function. This allows us to determine which variable(s) caused the rejection of the null hypothesis in the previous test. *(\*The SAS output is in the appendix #2)*

The standardized discriminant function is

Since 0.32049 and 0.1767 are significantly larger than the other two values, we can conclude that width of sepals and width of petals contribute most to the mean of the among three types of Iris flower.

We get the discriminant function from the procedure PROC DISCRIM.

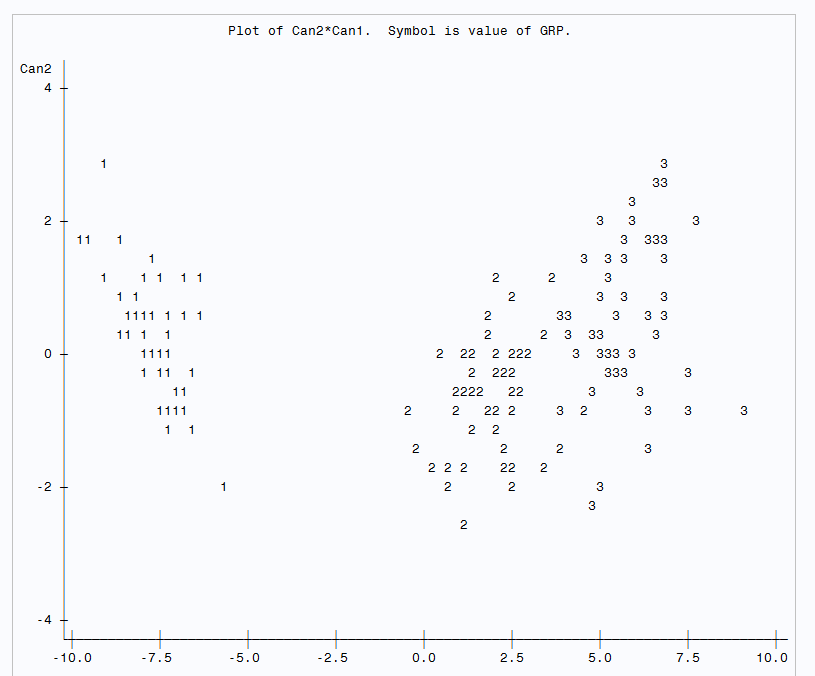
| **Linear Discriminant Function for GRP** | | | |
| --- | --- | --- | --- |
| **Variable** | **1** | **2** | **3** |
| **Constant** | -85.20986 | -71.75400 | -103.26971 |
| **Y1** | 23.54417 | 15.69821 | 12.44585 |
| **Y2** | 23.58787 | 7.07251 | 3.68528 |
| **Y3** | -16.43064 | 5.21145 | 12.76654 |
| **Y4** | -17.39841 | 6.43423 | 21.07911 |

Y1(the length of sepals) and Y2(the length of petals) contribute more than Y3( the width of sepals) and Y4 (the width of petals) in separate group 1(Iris Setosa). Y1 contribute more than other variables in separate group 2 (Iris Versicolour). Y4 contributes more than other variables in separate group 3 (Iris Virginca).

These 3 discriminant functions are used to distinguish new observations. The new observation will be classified into the group of which obtain the largest value when it is plugged into the 3 functions.

The SAS gives us the results of these classification.

| **Number of Observations and Percent Classified into GRP** | | | | |
| --- | --- | --- | --- | --- |
| **From GRP** | **1** | **2** | **3** | **Total** |
| **1** | |  | | --- | | 50 | | 100.00 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 50 | | 100.00 | |
| **2** | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 48 | | 96.00 | | |  | | --- | | 2 | | 4.00 | | |  | | --- | | 50 | | 100.00 | |
| **3** | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 1 | | 2.00 | | |  | | --- | | 49 | | 98.00 | | |  | | --- | | 50 | | 100.00 | |
| **Total** | |  | | --- | | 50 | | 33.33 | | |  | | --- | | 49 | | 32.67 | | |  | | --- | | 51 | | 34.00 | | |  | | --- | | 150 | | 100.00 | |
| **Priors** | |  | | --- | | 0.33333 | |  | | |  | | --- | | 0.33333 | |  | | |  | | --- | | 0.33333 | |  | |  |



We can see that group 1 (Iris Setosa) separates with the other 2 groups (Iris Versicolour and Iris Virginca). Iris Versicolour overlaps with Iris Virginca a little bit.

Then we test for significance.

***Hypothesis:***

***Test Statistics:*** F=199.15 , p-value<0.0001,

***Decision rule:*** if p-value< , reject null hypothesis.

***Conclusion:*** Since p-value <0.0001<0.05, we reject the null hypothesis. We conclude that there is sufficient evidence to indicate is significant at α=0.05.

***Hypothesis:***

***Test Statistics:*** F=13.79 , p-value<0.0001

***Decision rule:*** if p-value< , we reject null hypothesis.

***Conclusion:*** Since p-value <0.0001<0.05, we reject the null hypothesis. We conclude that there is sufficient evidence to indicate is significant at α=0.05.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test of H0: The canonical correlations in the current row and all that follow are zero** | | | | |
| **Likelihood** | **Approximate** | **Num DF** | **Den DF** | **Pr > F** |
| **Ratio** | **F Value** |
| 0.02344 | 199.15 | 8 | 288 | <.0001 |
| 0.77797 | 13.79 | 3 | 145 | <.0001 |

| **Total Canonical Structure** | | |
| --- | --- | --- |
| **Variable** | **Can1** | **Can2** |
| **Y1** | 0.791888 | 0.217593 |
| **Y2** | -0.530759 | 0.757989 |
| **Y3** | 0.984951 | 0.046037 |
| **Y4** | 0.972812 | 0.222902 |

From the SAS Output above, we can see Y1, Y3, Y4 is highly correlated with Can1, and Y2 is highly correlated with Can2.

**5. Conclusions and Recommendations**:

Conclusions:

(1) In the One-Way Manova analysis, we find that the group factor is significant, which indicates group has significant impact on the four variables. Then univariate analysis indicates the group factor is significant among the four variables which imply group has a significant impact on the four variables (length of sepals, length of petals, width of sepals and width of petals).

(2) In the Tukey Multiple Comparisons, we can see or compare the difference among three groups in each variable. We noticed the order of groups is the same among 3 variables. In second variable(length of petals), the order of groups is different for three groups.

For three variables(length of sepals, width of sepals and width of petals), we find out that Iris Virginca has the highest mean, Iris Versicolour is in the second place, and Iris Setosa has the smallest mean.

For the remaining variable(length of petals), we find out that Iris Setosa has the highest mean, Iris Virginca is in the second place, and Iris Versicolour has the smallest.

If we want to maximize the length of sepals, the width of sepals and the width of petals, we should choose Iris Virginca. If the goal is to maximize the length of petals, the species of Iris Setosa is recommended.

(3) In Principal Component Analysis, we chose the first principle component as an overall measurement. based on the eigenvalues and sree plot. We noticed that the Y3( the width of sepals) dominate this principle component and Y2(the length of petals) is not significant in this case. The variance of the width of sepals is larger than other variables in first component.

(4) In the Factor analysis, we chose 2 factors. The following Manova Analysis and Turkey Multiple Comparison indicate same conclusion with real variable analysis. In other words, we got the same conclusion as (1) and (2) . If the correlation among four variables is high, then either method is fine. However, factor analysis is better if we have many variables.

(5) In the discriminant analysis, new observations can be classified into the 3 groups with few mistakes. If separating Iris Setosa is the goal, the length of sepals and petals should be used; if separating Versicolour is the goal, the length of sepals should be used; if separating Iris Virginca is the goal, the width of petals should be used.

Future Research & Limitation of Study:

(1)If we could redo this project, we would analyze with other more effective methods. For example, we might use discriminant analysis or cluster analysis. This is because discriminant analysis is the nice method for Iris flower data. The dataset will be suitable for Cluster analysis since one type of Iris flower can be easily identified and the rest of two types of Iris flower sharing similar values. Furthermore, we'd like to read more materials on discriminant analysis to obtain a systematic knowledge on it. Hope that we can repeat and surpass the previous analysis from Iris Flower data.

(2)Since both of us are not familiar with botany and evolution of Iris, we analyzed the data by our knowledge from class and experience. We don't know how much those three species flowers close to each other genetically. However, we believe that if we could work with people having knowledge of botany, we can discover and develop more meaningful analysis from the dataset.

(3)We don't have deeper knowledge to use other methods to analyze the data. There are many articles related to this dataset on the Internet. In most articles, authors use this dataset to introduce K-mean and Y-mean algorithm to do the classification. With these techniques, we believe that we can dig out more things from the data set.

Appendix

1.

|  |  |  |
| --- | --- | --- |
| D2 | | |
| Group 1 | Group 2 | Group 3 |
| 12.3276 | 12.4895 | 13.6691 |
| 12.3101 | 10.2908 | 10.8264 |
| 11.0444 | 8.51461 | 9.48003 |
| 10.2221 | 8.08893 | 8.80209 |
| 9.74797 | 7.13371 | 8.78446 |
| 8.60116 | 6.52626 | 8.31315 |
| 7.69928 | 6.43085 | 7.99431 |
| 7.6538 | 6.16105 | 6.65567 |
| 7.23038 | 6.09166 | 6.41876 |
| 7.04021 | 6.06718 | 6.30421 |
| 5.74237 | 5.90277 | 5.6123 |
| 5.72127 | 5.75784 | 5.03188 |
| 5.34906 | 5.58936 | 4.54549 |
| 5.24802 | 5.54112 | 4.47264 |
| 5.18577 | 5.38239 | 4.44874 |
| 4.88911 | 4.73611 | 4.2847 |
| 4.20104 | 4.73258 | 4.16518 |
| 3.77051 | 4.47435 | 4.0077 |
| 3.71265 | 4.45106 | 3.94189 |
| 3.4242 | 4.3762 | 3.66819 |
| 3.30182 | 4.27475 | 3.64513 |
| 3.27025 | 4.23026 | 3.48512 |
| 3.20009 | 3.91786 | 3.32929 |
| 3.08565 | 3.77364 | 3.30504 |
| 2.99648 | 3.56318 | 3.18866 |
| 2.94733 | 3.52787 | 3.18665 |
| 2.75574 | 3.25278 | 3.1402 |
| 2.72236 | 3.11798 | 3.06747 |
| 2.52569 | 2.91999 | 2.80493 |
| 2.19465 | 2.91122 | 2.69108 |
| 2.1744 | 2.76318 | 2.61394 |
| 2.08109 | 2.65531 | 2.55945 |
| 2.01488 | 2.48053 | 2.31141 |
| 1.99451 | 2.45585 | 2.17213 |
| 1.89095 | 2.39731 | 2.16929 |
| 1.70621 | 2.2875 | 2.10935 |
| 1.68485 | 1.94912 | 1.89443 |
| 1.61241 | 1.76776 | 1.89443 |
| 1.48887 | 1.75977 | 1.87651 |
| 1.32301 | 1.63663 | 1.75005 |
| 1.28434 | 1.6243 | 1.58553 |
| 1.26698 | 1.40825 | 1.33526 |
| 1.25273 | 1.29444 | 1.29834 |
| 0.82913 | 1.06762 | 1.26505 |
| 0.76169 | 0.97253 | 1.13958 |
| 0.63625 | 0.84928 | 1.11181 |
| 0.58935 | 0.77256 | 1.07119 |
| 0.49476 | 0.75431 | 0.88826 |
| 0.44911 | 0.49762 | 0.85019 |
| 0.34344 | 0.37832 | 0.83338 |

2. Standardized discriminate function

| **Characteristic Roots and Vectors of: E Inverse \* H, where H = Anova SSCP Matrix for GRP E = Error SSCP Matrix** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Characteristic Root** | **Percent** | **Characteristic Vector V'EV=1** | | | |
| **NY1** | **NY2** | **NY3** | **NY4** |
| **32.1919292** | **99.12** | -0.05664462 | -0.05516376 | 0.32049497 | 0.17668887 |
| **0.2853910** | **0.88** | 0.00164612 | 0.07781377 | -0.13568712 | 0.17849492 |
| **0.0000000** | **0.00** | 0.08503198 | -0.08462566 | -0.39724293 | 0.28681113 |
| **0.0000000** | **0.00** | -0.20034650 | 0.04621702 | 0.18598081 | 0.00000000 |

3.

| **Pearson Correlation Coefficients, N = 150  Prob > |r| under H0: Rho=0** | | | | |
| --- | --- | --- | --- | --- |
|  | **Y1** | **Y2** | **Y3** | **Y4** |
| **Y1** | |  | | --- | | 1.00000 | |  | | |  | | --- | | -0.11757 | | 0.1519 | | |  | | --- | | 0.87175 | | <.0001 | | |  | | --- | | 0.81794 | | <.0001 | |
| **Y2** | |  | | --- | | -0.11757 | | 0.1519 | | |  | | --- | | 1.00000 | |  | | |  | | --- | | -0.42844 | | <.0001 | | |  | | --- | | -0.36613 | | <.0001 | |
| **Y3** | |  | | --- | | 0.87175 | | <.0001 | | |  | | --- | | -0.42844 | | <.0001 | | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.96287 | | <.0001 | |
| **Y4** | |  | | --- | | 0.81794 | | <.0001 | | |  | | --- | | -0.36613 | | <.0001 | | |  | | --- | | 0.96287 | | <.0001 | | |  | | --- | | 1.00000 | |  | |

4.

| **Pearson Correlation Coefficients, N = 150  Prob > |r| under H0: Rho=0** | | |
| --- | --- | --- |
|  | **Factor1** | **Factor2** |
| **Factor1** | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.00000 | | 1.0000 | |
| **Factor2** | |  | | --- | | 0.00000 | | 1.0000 | | |  | | --- | | 1.00000 | |  | |

5.

For factor 1

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **GRP** | 2 | 117.9326734 | 58.9663367 | 279.01 | <.0001 |

For factor 2

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **GRP** | 2 | 46.01715598 | 23.00857799 | 32.84 | <.0001 |