Bankrupt Data Analysis

Soumya Mukherjee
Susmit Saha
Sujoy Kumar Bhadra

April 2020

Contents

1 Description Of The Data

We are given with an annual financial data on financially sound and bankrupt firms. Data is given in the 5 columns where the last column indicates whether the firm is *financially sound* (by indicator **1**) or *bankrupt* (by indicator **0**). The first four columns are:

- X_1: Ratio of cash flow to total debt,
- X_2: Ratio of net income to total assets,
- X_3: Ratio of current assets to total liability,
- X_4: Ratio of current assets to net sales
- **Bankrupt Firms:** We have $n_1 = 21$ observations on bankrupt firms. Let us define the observation related to ith bankrupt firm as, $\underset{\sim}{X}_{1i} = (X_{i1}, X_{i2}, X_{i3}, X_{i4})$ for $i = 1, \ldots, n_1$, where X_{ij} denotes the observation corresponding to j^{th} variable on i^{th} individual firm.
- **Financially Sound Firms:** There are $n_2 = 25$ observations on financially sound firms. Let us define the observation related to i^{th} financially sound firm as, $\underset{\sim}{X}_{2i} = (X_{i1}, X_{i2}, X_{i3}, X_{i4})$ for $i = 1, \ldots, n_2$.
- **Assumptions:** We assume that $\underset{\sim}{X}_{11}, \ldots, \underset{\sim}{X}_{1n_1}$ and $\underset{\sim}{X}_{21}, \ldots, \underset{\sim}{X}_{2n_2}$ are independent samples from the population with mean vector $\underset{\sim}{\mu}_1$ and $\underset{\sim}{\mu}_2$ respectively and variance-covariance matrix Σ_1 and Σ_2 respectively.
- **Notations:** Also let us define sample mean for these two group as $\underset{\sim}{\bar{x}}_1, \underset{\sim}{\bar{x}}_2$ respectively and the sample co-variance matrices are S_1, S_2 respectively.

2 Exploratory Data Analysis

2.1 Deviation of the variables

We are going to study how our variables are scattered over or behave in 2 different sub-groups using box-plot.

- **Ratio of cash flow to total debt (X_1)**: Our first variable X_1 i.e., ratio of cash flow to total debt mostly takes negative value for the bankrupt firms whereas it takes positive value for most of the financial firms. In both cases we can observe a negatively skewed pattern. Hence financially sound firm group has higher mean than than bankrupt firm group. However there are three outliers in the bankrupt group; two of them have highly negative and the other one has highly positive ratio of cash flow which is very unnatural. Also there is only one outlier in financially sound group which has relatively high negative ratio. Also among financially sound firms the variable X_1 has more dispersion than among the bankrupt firms.

- **Ratio of net income to total assets (X_2)**: Like X_1, X_2 also takes negative values for most of the bankrupt firms whereas it takes positive values for all financially sound firms except two firms which are detected as outliers; having a very high negative ratio of net income for that firm. Among bankrupt firms X_2 is highly dispersed but for financially sound firms this ratio is clustered within a small range. In both the groups the variable X_2 has more-or-less symmetric distribution w.r.t. their means.

- **Ratio of current assets to net sales (X_3)**: The value of ratio of current assets to total liability always takes positive values. However for financially sound firms it takes higher value than for the bankrupt firms. Also in the bankrupt firm group, X_3 possesses a very small dispersion whereas it has a comparatively higher dispersion for financially sound group.

- **Ratio of current assets to total liability (X_4)**: X_4 is slightly more dispersed in bankrupt group than financially sound one. However there is a negatively skewed distribution pattern in the bankrupt group whereas there seems to be a positively skewed pattern in case of financially sound group. According to its corresponding boxplot there is no outlier present in either of the two groups.

BOXPLOT DIAGRAMS:

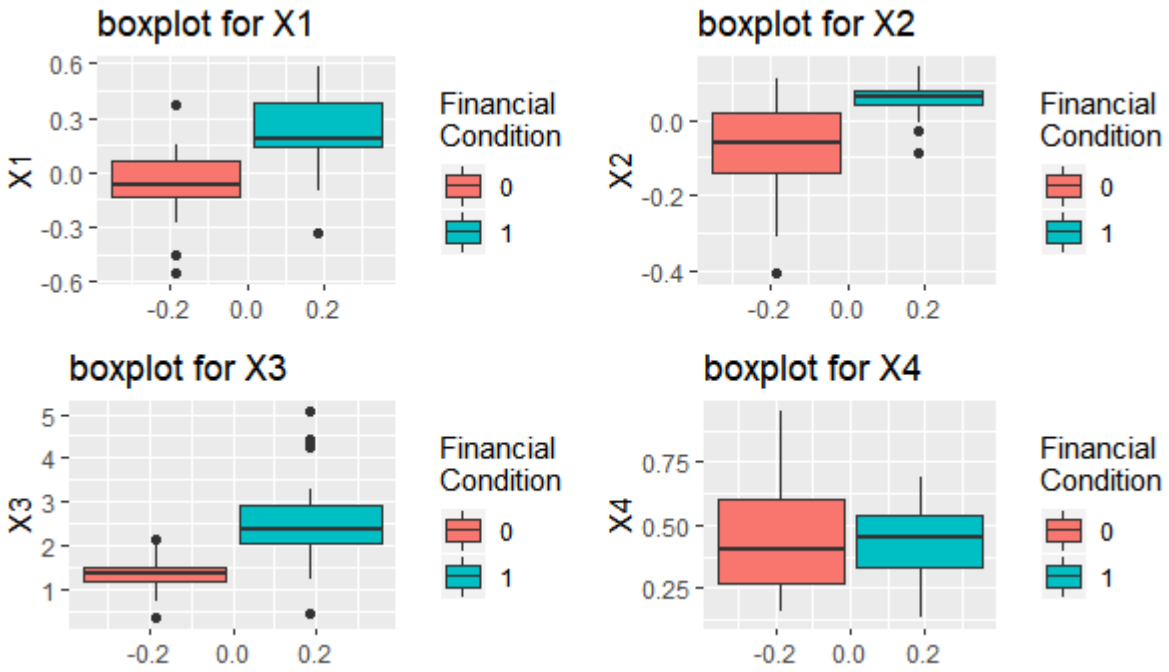


Figure 1: Boxplot Of Every Variable in the two groups

2.2 Correlation Structure

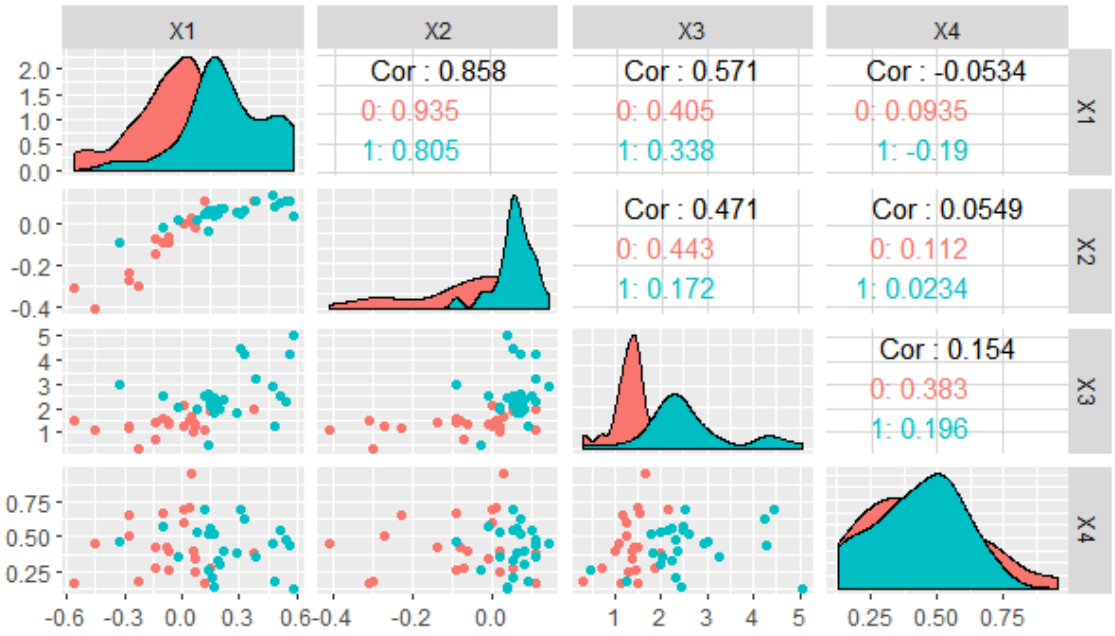


Figure 2: Correlation Plot

We can make following conclusions from the correlation plot :

1. There is a high positive correlation between the variable X_1 and X_2 in both the groups. However association is slightly stronger in the bankrupt group than the financially sound.

2. In both the group there is a weak association between variable X_1 and X_3. Also between variable X_2 and X_3 there is a mild positive correlation for bankrupt firm group but there is very weak association among the financially sound firm group.

3. However for other pairs of variables there seems to be no significant association.

3 Principle Component Analysis

Principle component analysis helps to reduce dimension of a multi-variate data structure. However PCA has nothing to do with the multi-variate normality assumption and hence we can perform PCA before checking the normality.

We have four variables and each variable is a ratio and consequently it is unit free. Hence we can perform PCA separately for the two groups using the corresponding sample co-variance matrices S_1 and S_2 respectively.

3.1 PCA On Bankrupt Group

First we have to decide how many principle components we are going to consider. It can be decided using the *Scree plot*.

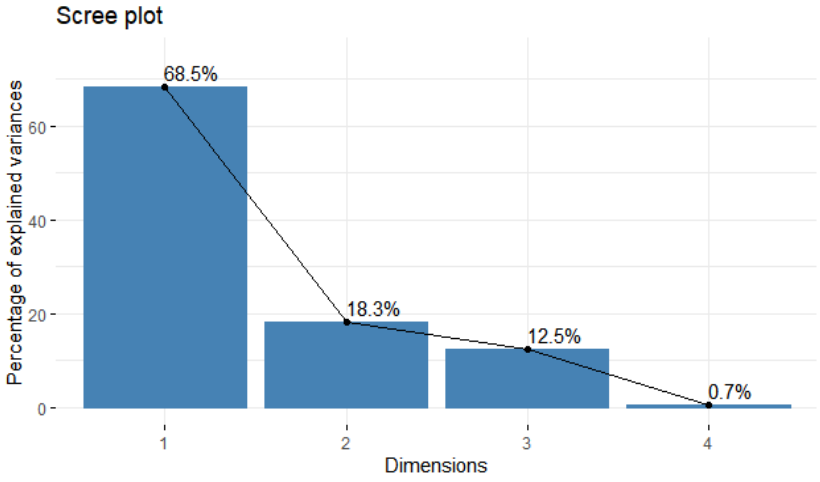


Figure 3: Scree Plot

From the scree-plot, it is pretty clear that a distinct elbow occurs at $i = 2$. Thus, it appears as if only one (the first) PC effectively summarises the total sample variance. However, there is a reverse bend at $i = 3$. Hence we can consider the first three PC's to be our reduced dimension.

Let us look at the sample principle components and the cumulative percentage of total variance explained by the PC's. The four PC's, obtained as the linear combination of these four variables are given as follows:

$$\hat{y}_1 = 0.27x_1 + 0.19x_2 + 0.92x_3 + 0.22x_4$$

$$\hat{y}_2 = 0.74x_1 + 0.48x_2 - 0.22x_3 - 0.41x_4$$

$$\hat{y}_3 = 0.28x_1 + 0.18x_2 - 0.33x_3 + 0.88x_4$$

$$\hat{y}_4 = 0.55x_1 - 0.83x_2 + 0.02x_3 + 0.00x_4$$

The following table gives us the variance and cumulative percentage of total variance explained by each PC:

	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$	$\hat{y}_4$
variance	0.187697793	0.050270167	0.034255667	0.001790659
Cumulative variance	68.49927	86.84509	99.34651	100.00000

We can observe that the first principal component can itself explain 68.5% of the whole variation. Although second and third principle component can only explain 19% and 12% respectively but these three components collectively can explain almost 99% of the total variation. Hence,it is enough to omit the last variable only and can continue to work with a reduced dimension of "3".

3.2 PCA On Financially Sound Group

In this section we are going to perform principle component analysis for the financially sound firm group. Therefore let us first take a look at the *scree plot*.

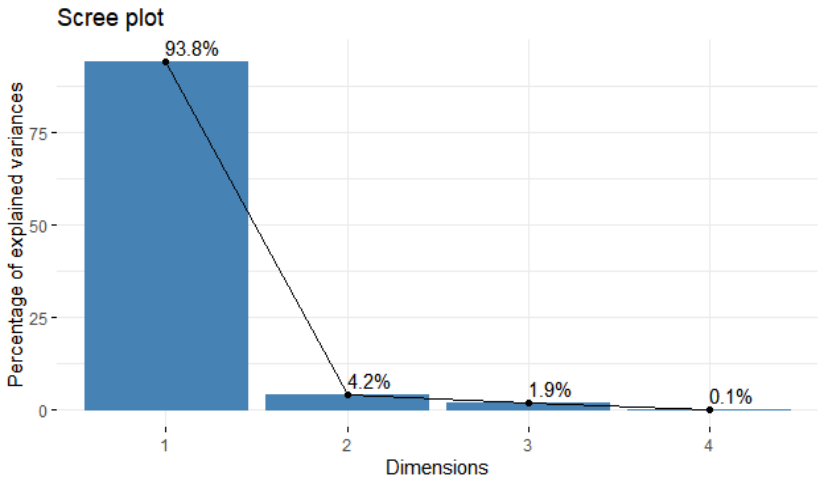


Figure 4: Scree Plot

From the scree-plot, it is pretty clear that a distinct elbow occurs at i = 2. Thus, it appears as if only one (the first) PC can effectively summarise the total sample variance.

Let us look at the sample PCs and the cumulative percentage of total variation explained by each PCs. The 4 sample principle components of the financially sound firm group are given as follows:

$$\hat{y}_1 = 0.07x_1 + 0.01x_2 + 1.00x_3 + 0.03x_4$$

$$\hat{y}_2 = 0.91x_1 + 0.16x_2 - 0.06x_3 - 0.39x_4$$

$$\hat{y}_3 = 0.36x_1 + 0.14x_2 - 0.06x_3 + 0.92x_4$$

$$\hat{y}_4 = 0.20x_1 - 0.98x_2 - 0.01x_3 + 0.07x_4$$

The following table gives us the variance and cumulative percentage of total variance explained by each PC:

	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$	$\hat{y}_4$
variance	1.0534392509	0.0467476968	0.0217290128	0.0006657062
Cumulative variance	93.84077	98.00507	99.94070	100.00000

Here we can observe that the first PC alone can explain 93% of the total variation. Hence we can consider whole data to be reduced by the first principle component satisfactorily.

4 Normality Checking

4.1 Methodology

We want to check whether our samples come from a p-variate multi-normal distribution or not. In our case $p = 4$. We have proceeded as follows:

- First, we try to investigate our data using some graphical procedures.
- Then we apply *Shapiro-Wilk's* test to check that each variable (i.e. the marginal distribution) are coming from uni-variate normal distribution or not .
- Lastly we apply *Royston's* test to verify presence of multivariate normality and observe whether our previous conclusions are compatible with this test's result or not and then make final conclusion about the normality of the sample.

We have applied the above-mentioned procedure to both the groups separately.

4.1.1 Graphical Procedure

We know that, if $\underset{\sim}{x} \sim N_p(\underset{\sim}{\mu}, \Sigma)$ then $(\underset{\sim}{x} - \underset{\sim}{\mu})'\Sigma^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu}) \sim \chi^2_p$. Suppose $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_n$ be a random sample coming from p-variate normal distribution with mean vector $\underset{\sim}{\mu}$ and variance-covariance matrix Σ. Then *Mahalanobis's distance* is defined as :

$$d_i = (\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})'S^{-1}(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}}) \quad \forall i = 1, \ldots, n$$

where $\bar{\underset{\sim}{x}}$ is the sample mean and S is the sample variance-co variance matrix.

For large n, d_i follows asymptotically χ^2_p for all $i = 0, 1, \ldots, n$. We sort d_i's in ascending order as $d_{(1)} < d_{(2)} < \ldots < d_{(n)}$. Let, q_i (for all $i = 1, 2, \ldots, n$) denote the $\frac{(i-0.5)}{n}$ th quantile of the χ^2_p distribution. We then plot the pairs $(q_i, d_{(i)})$. If the points lie on or close to the 45° line, we conclude that d_i comes from a χ^2_p distribution and hence the original sample might come from a p-variate normal distribution.

4.1.2 Shapiro-Wilk's Test

We can suspect that if the marginals of a multivariate distribution do not follow uni-variate normal distribution then the multivariate random variable may not follow multivariate normal distribution. For checking uni-variate normality we use *Shapiro-Wilk's* test.
In this test we want to verify the null hypothesis that a particular sample $\{x_1, x_2, \ldots, x_n\}$ coming from a normal distribution. Our test statistics is defined as,

$$W = \frac{(\sum_{i=1}^n a_i x_{(i)})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

where,

- $x_{(i)}$ is the i th order statistic $\quad \forall \quad i = 1, 2, ..., n.$
- $\bar{x}$ is the sample mean
- $(a_1, a_2, \ldots, a_n) = \frac{\underset{\sim}{m}' V^{-1}}{C}$
- $\underset{\sim}{m}$ and V is the mean vector and co-variance matrix of order statistics of *i.i.d.* random variables sampled from standard normal distribution respectively.
- C is a vector norm defined as, $C = ||V^{-1}\underset{\sim}{m}|| = (\underset{\sim}{m}' V^{-1} V^{-1} \underset{\sim}{m})^{\frac{1}{2}}$.

We make our conclusions based on the p-value.

4.1.3 Royston's Test

Royston's test is an extension of the uni-variate *Shapiro-Wilk's test* to check multi-normality. In this test procedure we want to test H_0 : sample points are coming from a multi-variate normal distribution against H_1 : not H_0. . Let W, denote the value of the Shapiro-Wilk statistic for the j^{th} variable in a p-variate distribution. Then define,

$$R_j = \left[\Phi^{-1}\Big\{\frac{1}{2}\big(\Phi(-\frac{(1-W_j)^{\lambda} - \mu}{\sigma}\big)\Big\}\right]^2 \qquad \forall \;\; j = 1(1)p$$

Then define the statistic as, $H = \frac{\xi}{p}\sum_{j=1}^{p} R_j$, where $\xi = \frac{p}{1+(p-1)\bar{c}}$ and $\bar{c}$ is an estimate of the average correlation among the R_j's.

Under H_0, H approximately follows $\chi^2_{(\xi)}$. We will use the p-value to interpret the result of this test.

{ **NOTE:** The details of the estimation procedure of "$\bar{c}$" can be found in the paper [1] cited in the References section. We implement the version available in the **MVN** package in R.}

4.2 Findings Regarding Bankrupt Firms

4.2.1 Graphical Method:

We plot the pairs $(q_i, d_{(i)})$ obtained from the observations of bankrupt firms and found that almost every point lies close to that 45° line except for the last three ordered squared Mahalanobis's distances. So, we can suspect that the samples may not come from a multivariate normal distribution.

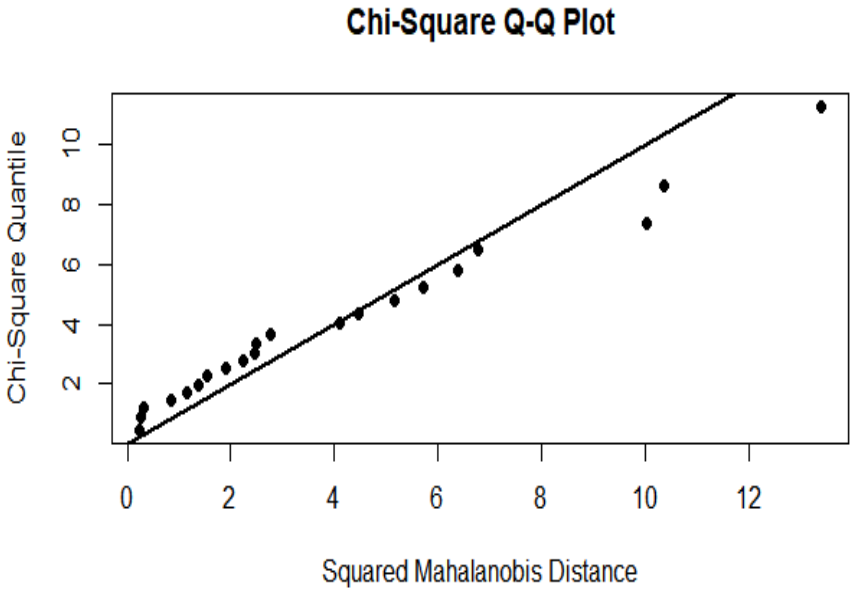


Figure 5: Chi-square Q-Q plot for bankrupt firms

4.2.2 Shapiro-Wilk's Test Results:

Now we try to verify marginal normality of each variable. We summarise our findings in the table below.

Variable	value of test statistic	p-value
X_1	0.9582	0.4800
X_2	0.9108	0.057
X_3	0.9595	0.5057
X_4	0.9372	0.1921

From the above table, we can conclude that each of them are following univariate normal distribution i.e. accept our null hypothesis at 0.05 level of significance.

We may be hopeful about the multi-variate normality of the data. However we verify our assumption through Royston's test.

4.2.3 Royston's Test

Lastly we perform Royston's test and obtain the p.value as **0.13** which is greater than 0.05. Hence we can conclude that the samples are coming from a multi-variate normal distribution.

4.3 Findings Regarding Financially Sound Firms

4.3.1 Graphical Method:

Now we follow the previous procedure for the observations related to financially sound firms. From the chi-square Q-Q plot we can observe that many points are lying far from that 45° diagonal line, specifically the points corresponding to the higher ordered squared distances. Hence it seems that the sample is not coming from a multi-variate normal population.

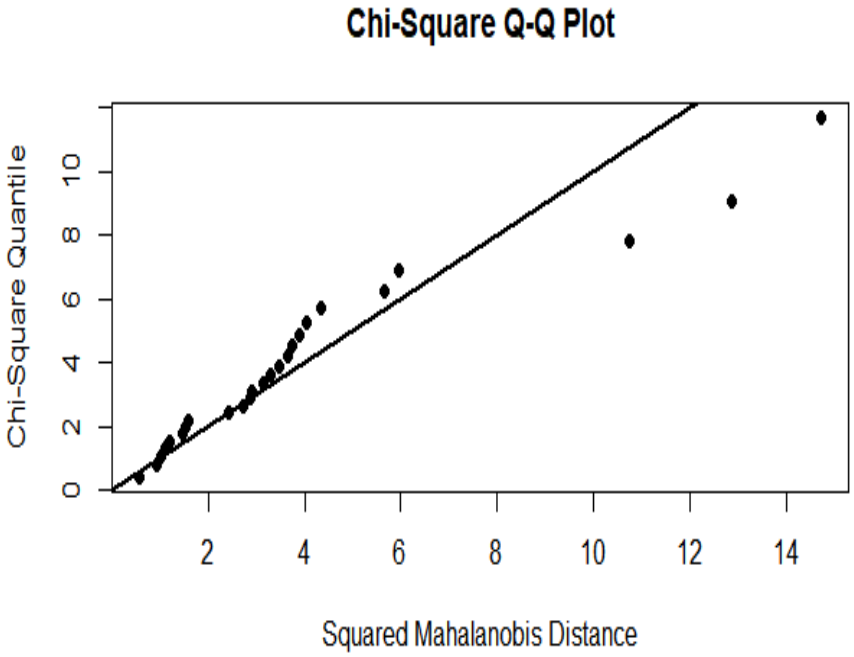


Figure 6: Chi-square Q-Q plot for financially sound firms

4.3.2 Shapiro-Wilk's Test

Results corresponding to Shapiro-Wilk's test are summarized in the table below.

Variable	Value of the statistics	p-value
X_1	0.9417	0.1620
X_2	0.9238	0.0626
X_3	0.9074	0.0267
X_4	0.9614	0.4429

As from the table we can comment that except the variable X_3 i.e., ratio of current assets to total liability all other three variable are coming from a normal population at a 5% level of significance. However we might suspect that the data is coming from a multi-variate normal population and testify our assertion through Royston's test.

4.3.3 Royston's Test

Lastly we perform Royston's test and get the p.value as **0.012** which is less than 0.05 Hence we can conclude that the samples are not coming from a multi-variate normal distribution at 5% level.

5 Power Transformation For Obtaining Normality

Now our objective is to use power transformations to obtain a new data set which holds the multi-variate normality condition.

The Box-Cox transformation is given by,

$$y(\lambda) = \begin{cases} \frac{x^{\lambda}-1}{\lambda}; & \text{if } \lambda \neq 0 \\ log(x); & \lambda = 0 \end{cases}$$

Where, we make a transformation of the given variable x as, $x = U + \gamma$. where γ is set by the user so $U + \gamma$ is strictly positive for these transformations to make sense. The optimal value of λ is the one which results in the best approximation of a normal distribution curve i.e., have the maximum likelihood. However this Box-Cox transformation does not allow negative observation. The Box-Cox family with negatives allowed was proposed by Hawkins and Weisberg in [2] in 2017. It is the Box-Cox power transformation of

$$z = 0.5(U + \sqrt{U^2 + \gamma^2})$$

where, for this family γ is either user selected or is estimated. γ must be positive if U includes negative values and non-negative otherwise, ensuring that z is always positive. We can use **"bcnPower"** function available in **car** package in R to get required estimated values of λ, γ.

Previously we have observed that sample observations coming from financially sound firms are not following multi-normality, whereas in bankrupt group we have multi-variate normality. Hence we try to transform that financially sound group data and then make the same derived transformations in bankrupt group and see still whether each of them follows mutli-variate normality or not, because we have to consider same transformation to keep it comparable.

Using R, we get the following estimated value of λ and γ for financially sound group.

Estimated parameter	X_1	X_2	X_3	X_4
$\hat{\lambda}$	0.43351945	1.01352799	0.04555263	1.99833949
$\hat{\gamma}$	0.2359694	0.1664129	3.7527159	3.0851024

To verify normality, we apply Royston's test on the transformed data-set get the p.value as 0.204 which is much greater than 0.05. Hence we can conclude that the transformed data regarding financially sound firm group follow multi-variate normal distribution at 5% level.

Now we make that obtained transformation on the bankrupt dataset and perform Royston's test to check normality. In this case we get the p.value as 0.1 which is greater than 0.05. Hence the transformed bankrupt data have multi-variate normality at 5% level.

Hence we shall proceed with these transformed dataset for the rest part of our analysis.

6 Exploratory Analysis Of The Transformed Data

We are going to analyse these transformed dataset graphically to see the amount of difference between the present dataset and the original on and to check how our drawn inferences from these new datasets will be relevant to the original ones.

- **Deviation of the variables:** We can use boxplot to visualize the spread of the each variable under the transformed dataset over the entire real line. From those boxplots we can observe that:

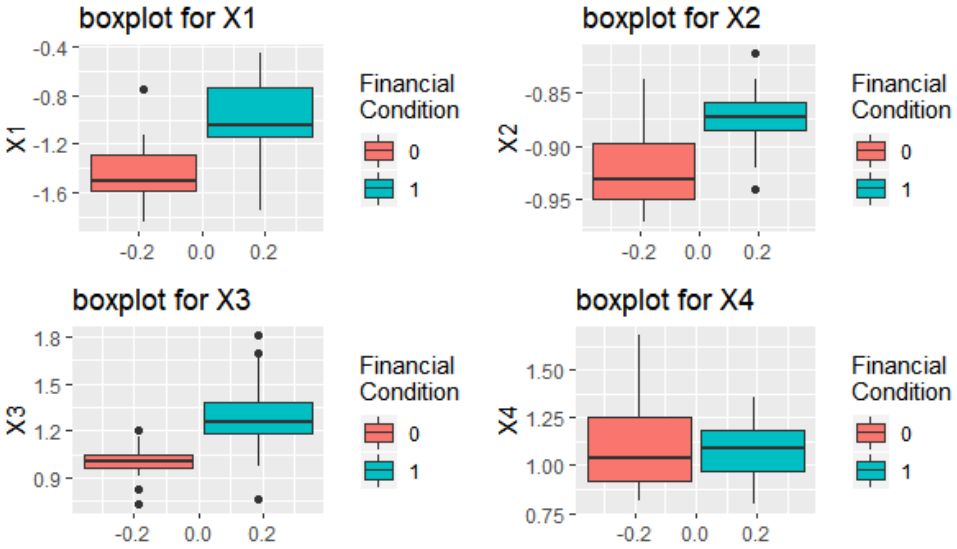


Figure 7: Boxplot for The Transformed Dataset

1. In the original dataset, variables X_1 and X_2 take positive value for financially sound firms and negative value for bankrupt firms. However in the transformed data, these two random variables take negative value for both the cases but they tend to take highly negative values for bankrupt firm compared to the financially sound group. There is no significant change in the deviation structure of the variable X_1 in each group; whereas for variable X_2 in financially sound group have larger range compared to the original one.
2. For variable X_3 and X_4 the only change is that they tend to take lesser value than the original scenario.

- **Correlation Structure**: First we are going to check the correlation structure among the variables in two groups.

Figure 8: Correlation Plot

From the correlation plot we can say that correlation structure between variables are almost same from the original dataset.

Hence we can safely conclude that inference about "difference in mean vectors" of the two groups using the transformed dataset can give relevant inference about the mean difference of the original sample. But the inferences drawn about the "mean vector" for each group using the transformed data set does not have any relevant interpretation for the mean vectors of the original dataset.

7 Test for equality of Dispersion Matrices

We want to test $H_0 : \Sigma_1 = \Sigma_2$ against $H_1 : \Sigma_1 \neq \Sigma_2$. where Σ_1 and Σ_2 are the dispersion matrices for bankrupt firms and financially sound firms respectively. *Box's M test* is commonly used in testing such problem. But is highly sensitive towards multivariate normality assumption. As our data-set is not holding the multivariate normality assumption, decision made from this test may not be consistent. Henceforth we can apply this test procedure on the transformed dataset. The **Box's M** statistics for 2 groups is given by,

Define,

$$M = (n_1 + n_2 - 2)ln|S_{pooled}| - \sum_{l=1}^{2}(n_l - 1)ln|S_l|$$

Where $S_{pooled} = \frac{1}{n_1+n_2-2}[(n_1 - 1)S_1 + (n_2 - 1)S_2]$.

Also define,

$$C = 1 - \left(\frac{2p^2 + 3p - 1}{6(p + 1)}\right)\left[\sum_{l=1}^{2}\frac{1}{(n_l - 1)} - \frac{1}{\sum_{l=1}^{2}(n_l - 1)}\right]$$

Then our test statistic is the product of M and C.

Under H_0, $M.C$ **follows** χ^2 with distribution $df = \frac{p(p+1)}{2}$. Hence we reject H_0 against H_1 at level α if $M.C > \chi^2_{df}$.

Box's χ^2 approximation works well if each n_l exceeds 20 and if the number of variables p and the number of groups g do not exceed 5.

Our problem satisfies the above conditions implying it is appropriate to use Box's M test. Here $df = 10$.

- **Result And Conclusion:** We perform this test procedure on our transformed dataset. From R, we get $MC = 36.587$ and $\chi^2_{10;0.95} = 18.30704$. Hence, the test rejects H_0 at 5% level of significance.
- **Consequence:**
  1. As a consequence of this unequal variance matrices we **can not perform profile analysis**.
  2. Also in discriminant analysis **Fisher's linear discriminant function may not be much efficient**.

8 Test For The Equality Of Mean Vectors

Previously from the boxplots, we have observed that each variable has completely different range of values for the two different groups. Hence we can intuitively draw our inference that their theoretical mean vectors are not equal. In this section we are going to justify our intuition using a proper testing procedure.

We are want to test $H_0 : \underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2 = \underset{\sim}{0}$ against $H_1 : \underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2 \neq \underset{\sim}{0}$. However our original sample do not hold multi-variate normality assumption. Hence we are going to perform this testing problem on the transformed data-set. We are going to check whether the mean difference of the transformed vector is zero or not.

We have two population from normal distribution but their co-variance matrices are unequal. Hence here we are going to use the test statistics as,

$$T^2 = (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)' (\frac{S_1}{n_1} + \frac{S_2}{n_2})^{-1} (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)$$

As our sample sizes are large enough, we are going to use an approximate distribution of T^2 as, under H_0, $T^2 \sim \frac{vp}{v-p+1} F_{p,v-p+1}$, where

$$v = \frac{p + p^2}{\sum_{i=1}^{2} \frac{1}{n_i} \left(tr[(\frac{S_i}{n_i}(\frac{S_1}{n_1} + \frac{S_2}{n_2})^{-1})^2] + (tr[\frac{S_i}{n_i}(\frac{S_1}{n_1} + \frac{S_2}{n_2})^{-1}])^2 \right)}$$

However, here we consider v to be the integral part of v.

We reject H_0 against H_1 at level 0.05 if $T^2 > \frac{vp}{v-p+1} F_{0.05;p,v-p+1}$.

Using R, we get the value of the test statistic as, $T^2 = 46.12$ and the critical value is 11.20569. Hence we **reject** our null hypothesis at 5% level.

Thus we have obtained a strong evidence for the presence of difference in mean vectors. Hence we may look for obtaining an interval estimate for this difference in mean vectors.

9 Interval Estimation Of Mean Differences

- **Separate Interval**: From the previous section we have find that the difference of the two mean vectors is significantly differ from zero. Hence interval estimation of mean differences of each variable can have a significant interpretation.

  The 100(1-α)% confidence intervals for differences of means of two normal populations is given by,

$$\left[(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2) - t_{\frac{\alpha}{2};n_1+n_2-2} \sqrt{(\frac{1}{n_1} + \frac{1}{n_2}) S_{pooled}}, (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2) + t_{\frac{\alpha}{2};n_1+n_2-2} \sqrt{(\frac{1}{n_1} + \frac{1}{n_2}) S_{pooled}} \right].$$

  We are going to summarize our results in the following table.

Mean difference	lower bound	upper bound
$\mu_{11}^{'} - \mu_{12}^{'}$	-0.64352322	-0.28113824
$\mu_{21}^{'} - \mu_{22}^{'}$	-0.06757147	-0.02756421
$\mu_{31}^{'} - \mu_{32}^{'}$	-0.40219790	-0.18328102
$\mu_{41}^{'} - \mu_{42}^{'}$	-0.09845179	0.13119652

- **Simultaneous Interval:** An 100(1-α)% simultaneous confidence interval for all linear combinations $\underset{\sim}{a}^{'}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$ are provided by,

$$\left[\underset{\sim}{a}^{'}(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2) - \sqrt{\chi^2_{\alpha;p}}\sqrt{\underset{\sim}{a}^{'}(\frac{S_1}{n_1} + \frac{S_2}{n_2})\underset{\sim}{a}}, \underset{\sim}{a}^{'}(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2) + \sqrt{\chi^2_{\alpha;p}}\sqrt{\underset{\sim}{a}^{'}(\frac{S_1}{n_1} + \frac{S_2}{n_2})\underset{\sim}{a}}\right]$$

  In our case, $\underset{\sim}{a} = \underset{\sim}{e}_i$; where $\underset{\sim}{e}_i$ is a p-dimensional vector whiose each component is zero except the i$^{\text{th}}$ component which is 1.

  Summarizing the results in the following table.

Mean difference	lower bound	upper bound
$\mu_{11}^{'} - \mu_{12}^{'}$	-0.73315589	-0.19150557
$\mu_{21}^{'} - \mu_{22}^{'}$	-0.07904231	-0.01609337
$\mu_{31}^{'} - \mu_{32}^{'}$	-0.45071795	-0.13476097
$\mu_{41}^{'} - \mu_{42}^{'}$	-0.16409548	0.19684021

- **Comments**: From the above interval estimates of the mean differences, we can say that:

  1. In both the cases (i.e., separate and simultaneous), we can observe that mean differences for the transformed variables X_1, X_3 significantly greater than zero compare to the other two variables. Hence we can say that for X_1, X_3 bankrupt firm has lower average than financially sound group. Clearly, these inferences have satisfactory agreement with our comments from those boxplots. Also for each variable bankrupt firm group have lower mean than the financially good firm population.
  2. We are going to summarize interval lengths for each mean differences for the two cases separate and simultaneous both in the below:

Transformed variable	Length of the interval	
	Separate	Simultaneous
X_1	0.36	0.54
X_2	0.04	0.06
X_3	0.22	0.32
X_4	0.23	0.36

  We can observe that simultaneous intervals have greater length than the separate intervals which is expected from the theoretical point of view.

10 Factor Analysis:

In factor analysis, our primary motive is to draw inferences on latent factors which are unobservable but largely influence the behaviour of the variables at hand. Here we try to describe the covariance/ correlation structure between the given variables in terms of the underlying and unobservable random quantities called "factors" using an Orthogonal m-factor Model.

10.1 Orthogonal Factor Model

An Orthogonal m-factor model assumes that $\underset{\sim}{X}$ can be written as linear combination of a set of "m" common factors $F_1, \ldots, F_m$ and "p" additional unique factors $\epsilon_1, \ldots, \epsilon_p$. Thus the model can be expressed in matrix form as follows:

$$\underset{\sim}{X} - \underset{\sim}{\mu} = L\underset{\sim}{F} + \underset{\sim}{\epsilon}$$

where, L is the $m \times p$ loading matrix, $\underset{\sim}{F}$ is the vector of common factors which are unobservable and $\underset{\sim}{\epsilon}$ is the vector of specific factors.

Model Assumptions:

$We\ assume\ that,$

$$E(\underset{\sim}{F}) = \underset{\sim}{0}$$
$$V(\underset{\sim}{F}) = I$$
$$E(\underset{\sim}{\epsilon}) = \underset{\sim}{0}$$
$$V(\underset{\sim}{\epsilon}) = \Psi = diag(\Psi_1, \Psi_2, ..., \Psi_p)$$
$$Cov(\underset{\sim}{F}, \underset{\sim}{\epsilon}) = 0$$

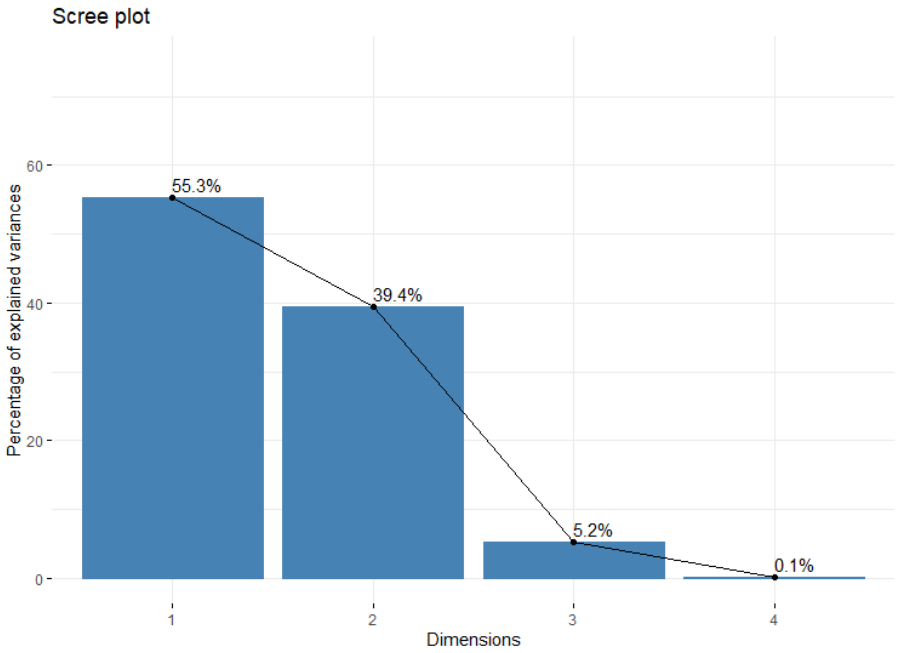
Considering the above assumptions, we can express covariance matrix Σ as follows:

$$\Sigma = LL' + \Psi$$

10.2 Bankrupt firms:

10.2.1 Choosing of "m":

It is very important to decide upon the number of latent variables to be included in our model beforehand. We perform a **principle component analysis** to get an idea about how many components are going to be significant in explaining the variability in this population.



From the scree plot we can observe that almost **96%** of the total variation can be explained by the first two prinicpal components.

Thus the optimal choice of "m" can be taken to be 2.

10.2.2 Estimation of Loading Matrix:

In factor analysis, our primary motive is to estimate L and Ψ. Here we shall use **"Iterative Principal Component Method"** to estimate the loading matrix.

For estimation purpose we replace Σ_1 by S_1 or R_1 and rewrite the model as follows:

$$S_1 = LL' + \Psi \ \ or \ \ R_1 = LL' + \Psi$$

Here we shall work with the sample correlation matrix R_1 for the purpose of our estimation.

The estimated loading matrix is given by:

$$\hat{L}= \begin{pmatrix} 0.98 & -0.18 \\ 0.93 & -0.21 \\ 0.56 & 0.51 \\ 0.14 & 0.60 \end{pmatrix}$$

Interpretation: From the loading matrix, it can be observed that the variables X_1,X_2 are highly loaded on first factor F_1 whereas the variables X_3 and X_4 are moderately loaded on second factor F_2.

Since, we have successfully loaded a particular variable on a particular factor without using any kind of rotation, hence we will not implement any further rotation and will continue to work with the above unrotated estimated loading matrix.

10.2.3 Estimation of Factor Scores and Checking Validity of the Assumptions of OFM :

Here we have estimated the factor scores using the loading matrix estimated above using both "Weighted-Least-Squares" method and "Regression" method. Also we have checked if the estimated factor scores conform to assumptions of **"normality"** and **"pairwise independence"**.

1. **Normality Test:** We have performed multivariate normality test using "Royston's Test".

Method	Bartlett (Weighted Least Squares)	Thompson (Regression)
Obtained p.values	0.5824807	0.5345086

   Thus p.values being much greater than 0.05 we can accept presence of multivariate normality.

2. **Pairwise Independence Test:** Since we have considered a 2-factor model and presence of normality between factor scores has already been accepted above, hence we can use "Pearson test" for independence which is used for checking independence in case of a bivariate normal population.

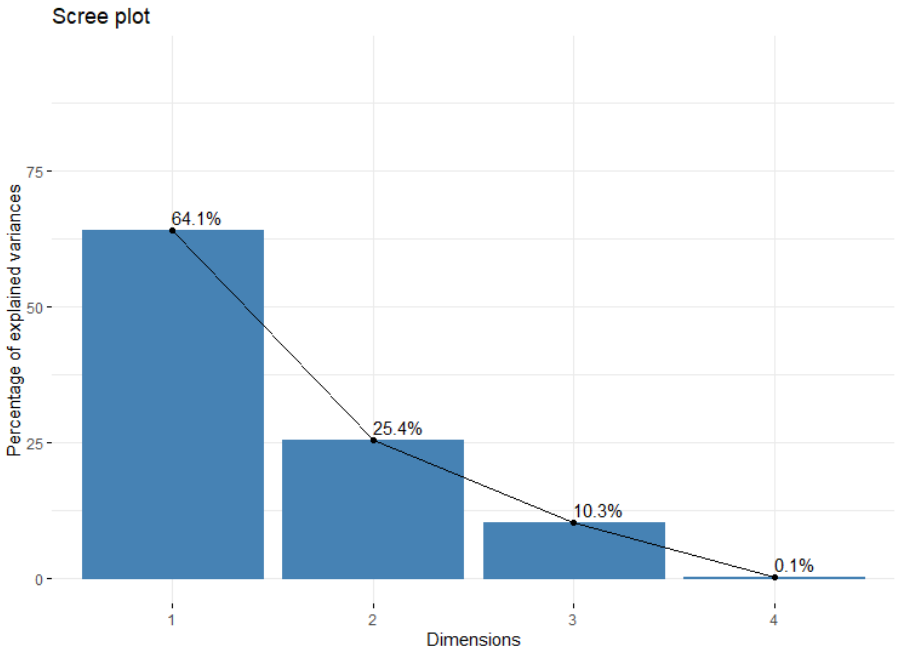
Method	Bartlett (Weighted Least Squares)	Thompson (Regression)
Obtained p.values	0.6894	0.6893

   Thus we conclude that the factors F_1 and F_2 are independent.

10.3 Financially Sound Firms:

10.3.1 Choosing of "m":

Here also we perform a **principle component analysis** to get an idea about how many components are going to be significant in explaining the variability in this group of firms.



From the scree plot we can observe that almost **90%** of the total variation can be explained by the first two prinicpal components.

Thus the optimal choice of "m" can be taken to be 2.

10.3.2 Estimation of Loading Matrix:

Here also we estimate L and Ψ using **"Iterative Principal Component Method"**.

For estimation purpose we replace Σ_2 by S_2 or R_2 and rewrite the model as follows:

$$S_2 = LL' + \Psi \ \ or \ \ R_2 = LL' + \Psi$$

Here also, we shall work with the sample correlation matrix R_2 for the purpose of our estimation.

The estimated loading matrix is then given by:

$$\hat{L}= \begin{pmatrix} 1.23 & -0.06 \\ 0.65 & 0.10 \\ 0.28 & 0.33 \\ -0.11 & 0.81 \end{pmatrix}$$

Interpretation: It can be observed from the estimated loading matrix,that the variables X_1,X_2 are highly loaded on first factor F_1 whereas the variables X_4 is highy loaded on second factor F_2.

Since we have successfully loaded a particular variable on a particular factor without using any kind of rotation, hence we will not implement any further rotation and will continue to work with the above unrotated estimated loading matrix.

10.3.3 Estimation of Factor Scores and Checking Validity of the Assumptions of OFM :

Here we have estimated the factor scores using the loading matrix estimated above using both "Weighted-Least-Squares" method and "Regression" method. Also we have checked if the estimated factor scores conform to assumptions of **"normality"** and **"pairwise independence"**.

1. **Normality Test:** We have performed multivariate normality test using "Royston's Test".

Method	Bartlett (Weighted Least Squares)	Thompson (Regression)
Obtained p.values	0.4876858	0.4619858

   Thus p.values being much greater than 0.05 we can accept presence of multivariate normality.

2. **Pairwise Independence Test:** Since we have considered a 2-factor model and presence of normality between factor scores has already been accepted above, hence we can use "Pearson test" for independence which is used for checking independence in case of a bivariate normal population.

Method	Bartlett (Weighted Least Squares)	Thompson (Regression)
Obtained p.values	0.9025	0.9032

   Thus we conclude that the factors F_1 and F_2 are independent.

11 Discriminant Analysis

Discriminant analysis provides us two disjoint classification regions based on values of the concerned variables obtained from 2 different populations or groups and can predict to which group a new unit (with given values of the variables) will go into.

Although our transformed data has multi-variate normality but we do not have $\Sigma_1 = \Sigma_2$; hence instead of *linear discriminant analysis* we go for *quadratic discriminant analysis*.

Let bankrupt firm population be defined by π_1 and financially sound group be defined as π_2.

- **Classification Region**: We define *quadratic discriminant score function* as follows:

$$d_1^Q(\underset{\sim}{x}) = -\frac{ln|S_1|}{2} - \frac{1}{2}(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_1)' S_1^{-1}(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_1) + ln(p_1)$$

$$d_2^Q(\underset{\sim}{x}) = -\frac{ln|S_2|}{2} - \frac{1}{2}(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_2)' S_2^{-1}(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_2) + ln(p_2)$$

  We allocate a new observation $\underset{\sim}{x_0}$ in the population π_i if

$$d_i^Q(\underset{\sim}{x_0}) = \max\{d_1^Q(\underset{\sim}{x_0}), d_2^Q(\underset{\sim}{x_0})\}.$$

- **Evaluation of the discriminant model**: Since we do not have a large number of data-points hence we do not split the whole data into training set and validation set. However we will execute this discriminant model over the whole dataset and obtain the confusion matrix as given below:

	Predicted	
True	Bankrupt	Financially sound
Bankrupt	19	2
Financially sound	2	23

  Hence, the *apparent error rate*(APER) is $\frac{2+2}{46} \times 100 = 8.7\%$, which is quite low indicating that quadratic discriminant model is working good.

  For further verification we apply *Lachenbruch's holdout procedure*, which is described as below:

  1. "Hold" out the first observation.
  2. Use the remaining $(n-1)$ observations as the training set.
  3. Obtain a classification rule based on the training set.
  4. Predict the observation which we had "held" out using the classification rule.
  5. Repeat 1-4 for all the observations, one at a time.

6. Calculate APER as the proportion of misclassified observations.

In this case the confusion matrix is:

	Predicted	
True	Bankrupt	Financially sound
Bankrupt	17	4
Financially sound	4	21

Hence we can calculate the value of APER as, $\frac{4+4}{46} \times 100 = 17.8\%$ which is quite low.

However as a more sensitive measure Lachenbruch's holdout method indicate slightly greater error rate than the previously mentioned APER.

12 Summary

We have briefly summarized our findings in the following points:

1. From the boxplots we have found that variable X_1, X_2 take mostly negative values for bankrupt firms whereas strictly positive value for financially sound firms. However the other two variables always take positive value in each group. Expect variable X_4, other variables have different deviation structure for these two groups.

2. Among all the variables, X_1 and X_2 have strong positive association; also the variable pairs X_1, X_3 and X_2, X_3 have moderate positive association among themselves. However for all other variable pairs there are no significant associations among them.

3. We have performed principle component analysis and found that for bankrupt population the "first three principle components" explains almost 99% of the total variation whereas in the financially sound group, the "first two components" are able to explain about 98% of the total variation.

4. We then checked the assumption of multivariate normality in both the groups and found that for the bankrupt group the normality condition holds but for the financially sound group assumption of multivariate normality fails to hold.

5. As a consequence of the previous point, we have made a power transformation to the financially sound group to make it a sample from a 4-variate normal distribution and made the same transformation to the bankrupt group. Also we have checked the multivariate normality of the data through Royston test and in each case, on the basis of p-value, we can conclude that the transformed data follows the assumption of multivariate normality.

6. We then perform an exploratory data analysis of these transformed dataset and found that the range of the variables differ from the original one but they possess the same deviation structure and hence inference on the difference of means is relevant with respect to the original dataset. Also the correlation structure has remained more-or-less same w.r.t the original one.

7. We checked the equality of the dispersion matrices of the transformed dataset using the Box-M test and found that they are not equal. As a consequence of this we are not going to perform profile analysis and do not use linear discriminant function in discriminant analysis.

8. Also we checked the equality of the mean vectors and found that the mean difference significantly differs from zero.

9. We have obtained interval estimate of the mean differences and found that for variable X_1 and X_3 they are significantly different from zero. Also the simultaneous intervals have greater length than the separate intervals.

10. We have also performed factor analysis in each of the groups. We have found from the principle component analysis of the transformed dataset that in each group the first two principle component explain the major part of the total variation. Hence we have performed a 2-factor orthogonal model in each group. For bankrupt firm group, variables X_1, X_2 highly loaded on the first factor and X_3 and X_4 are moderately loaded on the factor two. In the financially sound group, X_1, X_2 have high load on the factor one whereas X_4 is highly loaded on the factor two.

11. Lastly we performed discriminant analysis using quadratic discriminant score function. We have found the APER using the whole dataset as the training set as 8.7% which is lesser than the APER obtained using the Lachenbruch's holdout procedure (17.8%). Both the APER indicates that the discriminant analysis using quadratic score is worthy.

13 References

[1] J. P. Royston. Some techniques for assessing multivarate normality based on the shapiro- wilk w. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 32(2):121–133, 1983.

[2] D.M. Hawkins and S. Weisberg. Combining the box-cox power and generalised log transformations to accommodate nonpositive responses in linear and mixed-effects linear models. *South African Statistical Journal*, 51:317–328, 12 2017.

[3] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Applied Multivariate Statistical Analysis. Pearson Prentice Hall, 2007.