Graduate Admissions

Analyzing and Comparing a Prospective Student's Application Scores with Their Chance of Being Admitted

STA 135 Multivariate Data Analysis

Xiaodong Li

Christopher Ton
SID: 913915130
June 10, 2019

Introduction

The dataset analyzed was retrieved from Kaggle and it introduces various student attributes in terms of exam scores, grade averages and supplemental factors for consideration during the admission review for UCLA's graduate programs. Specifically, the data consists of 500 observations and 8 variates that includes the GRE and TOEFL scores, quality of university, statement of purpose and recommendation letters, grade point averages, whether or not research was conducted during undergrad and finally a probability value for admittance. Analysis involving confidence intervals based on Hotelling's T^2 and Bonferroni correction for the population means regarding both the one sample and two sample cases were conducted. For the two sample tests, groups were divided into two based on the existence of research experience as the categorical factor, a binary value of either 0 or 1. Finally, dimension reduction approaches that are the principal component and linear discriminant analyses were employed in further examination of the dataset.

The assumption of normality was assumed here for each of the variates since the procedure for testing when normality does not exist was not introduced in the course. Relatively, transformations can be done but even so, normally may not be achieved, implying non-parametric methods. Plotting a histogram of each the variates shows distributions that almost identical to the bell curve. Discrepancies in the analysis may be due to the assumed normality of the random variables who are mutually independent.

Inspiration for analyzing this dataset could be answering the if a certain group of students with have had research experience recieved higher chances of being admitted than those without experience in research.

Summary

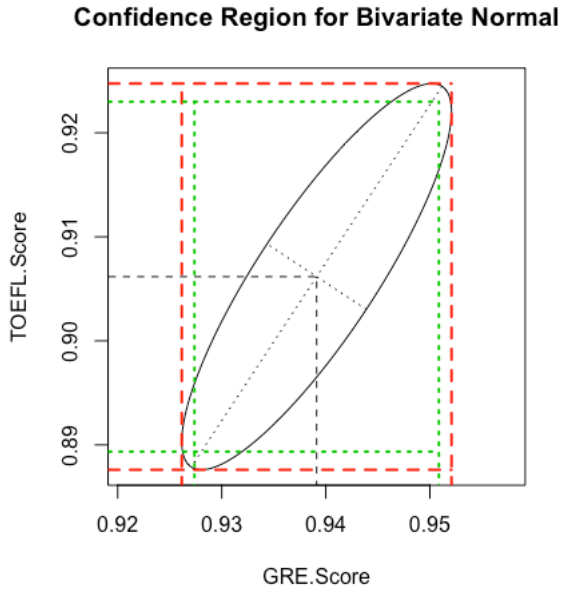
The sample mean vector below are the averages for each of the 9 variables.
$\bar{\vec{x}} = [319.3000, 108.7400, 3.3400, 3.5500, 3.4300, 8.6098, 0.6400, 0.7134]$

The sample covariance matrix:

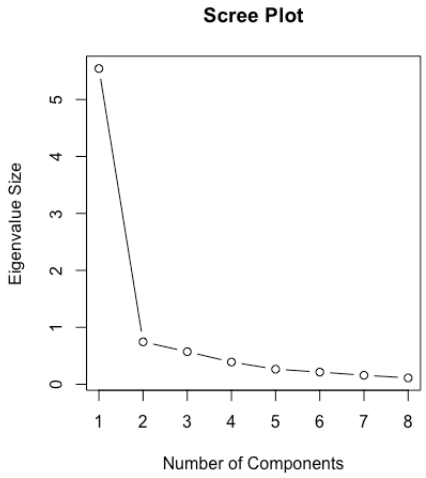
	GRE.Score	TOEFL.Score	University.Rating	SOP	LOR
GRE.Score	149.071429	64.3448980	12.9979592	7.1683673	7.9908163
TOEFL.Score	64.344898	38.0738776	6.2330612	3.9214286	3.5528571
University.Rating	12.997959	6.2330612	1.6167347	0.9520408	0.8916327
SOP	7.168367 3.9214286	0.9520408	0.9107143	0.6413265	
LOR	7.990816	3.5528571	0.8916327	0.6413265	1.0409184
CGPA	7.393939	3.7150490	0.7841510	0.4935816	0.5088633
Research	2.355102	0.9044898	0.2677551	0.1306122	0.2089796
Chance.of.Admit	1.613653	0.7796776	0.1765755	0.1040102	0.1222837

	CGPA	Research	Chance.of.Admit
GRE.Score	7.3939388	2.35510204	1.61365306
TOEFL.Score	3.7150490	0.90448980	0.77967755
University.Rating	0.7841510	0.26775510	0.17657551
SOP	0.4935816	0.13061224	0.10401020
LOR	0.5088633	0.20897959	0.12228367
CGPA	0.5013816	0.12380408	0.10090069
Research	0.1238041	0.23510204	0.03410612
Chance.of.Admit	0.1009007	0.03410612	0.02692494

The values in this table provides the relationship between each of the variates.

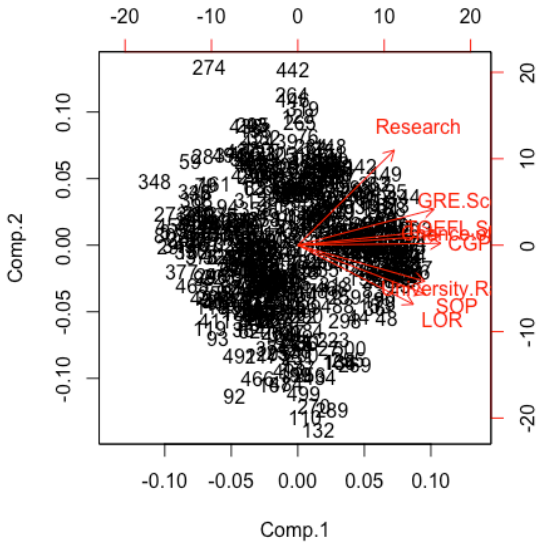


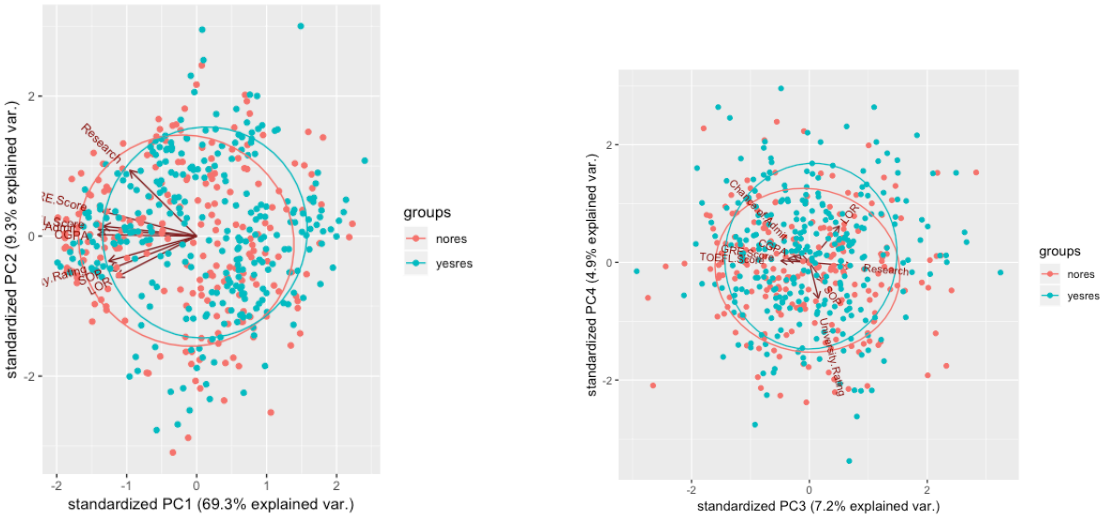
The image above is the graphical representation for the testing of equality of any two particular means. In this case, the test is being done on the GRE.Score and TOEFL.score, whose columns are divided by their maximums, in order to get a percentage score. This confidence ellipse of covers the true parameter μ with probability $1 - \alpha$. Labeled in red is the 95% Simultaneous T^2 Confidence Intervals and in green is the narrower Bonferroni Correction Interval.



A scree plot, a plot of $\hat{\lambda}_i$ versus i - the ordered magnitude of an eigenvalue from large to small versus its component number- determines an appropriate number of principal components to retain. The elbow area is this number, and the remaining eigenvalues should be insignificant and of the same size. According to the graph, two sample principal components effectively summarize the total sample variance. (pg 445)

First Two Prinicipal Component Scores





The first plot provide the scores for the sample data in the space of the first two principal components while the remaining two present the plots governed by standardization and the less significant components for reference, respectively.

Given that we have classified attributes to separate the population into particular groups, and the goal is to classify a new observation into the respective class. Provided below is the confusion matrix, that counts the instances of these allocations happening.

		Predicted Membership		
		π_1	π_2	
Actual	π_1	$n_{1c} = 155$	$n_{1m} = n_1 - n_{1c} = 65$	n_1
membership	π_2	$n_{2m} = n_2 - n_{2c} = 58$	$n_{2c} = 222$	n_2

Analysis

One Sample Inference

Simultaneous Confidence Intervals Based on T^2 and Bonferroni Correction

Testing the hypothesis $H_0 := \vec{\mu} = \vec{\mu_0}$ and for instance, we can test to see if:

	$\vec{\mu}$	=	$\vec{\mu_0}$
GRE.Score	319.3000		320
TOEFL.Score	108.7400		100
University.Rating	3.3400		4.5
SOP	3.5500		4.6
LOR	3.4300		4.2
CGPA	8.6098		8.7
Research	0.6400		0.7
Chance.of.Admit	0.7134		0.6

Hotelling's T^2 Testing

The procedure used here is a testing of a multivariate mean vector with Hotelling's T^2 where n is the first 50 observations and p is number of variates, 8.
Decision rule: reject H_0 if $T^2 > \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)$

Plugging in and solving the equation for the observed value of $T^2 = n(\vec{\bar{X}} - \vec{\mu_0})^T S^{-1}(\vec{\bar{X}} - \vec{\mu_0})$ which is then compared to the critical value , under the null hypothesis, H_o, $\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)$, we get that:

$T^2 = 1053.814 > 20.23576$, reject the H_0.

P-value based on T^2_{mod}

To further emphasize the decision rule, the p-value can be calculated by $P(Z > T^2_{mod})$ where Z is a random variable that follows the same distribution, $F_{p,n-p}$, and $T^2_{mod} = \frac{(n-1)p}{n-p}T^2 \sim F_{p,n-p} \rightarrow T^2_{mod} = 112.9086$
$P(Z > 112.9086) = 1 - P(Z < 112.9086) = 2.2x10^{-16} < 0.05 = \alpha$, reject H_0

95% T^2 Confidence Intervals

We want to answer the question, " How to find simultaneous intervals for $\mu_1, ..., \mu_p$ with $\vec{\bar{x}}$ and S?"

For each $j = 1, ..., p$, the confidence interval for μ_j is $\bar{x_j} \pm \frac{s_j}{\sqrt{n}}t_{n-1}(\alpha/2)$

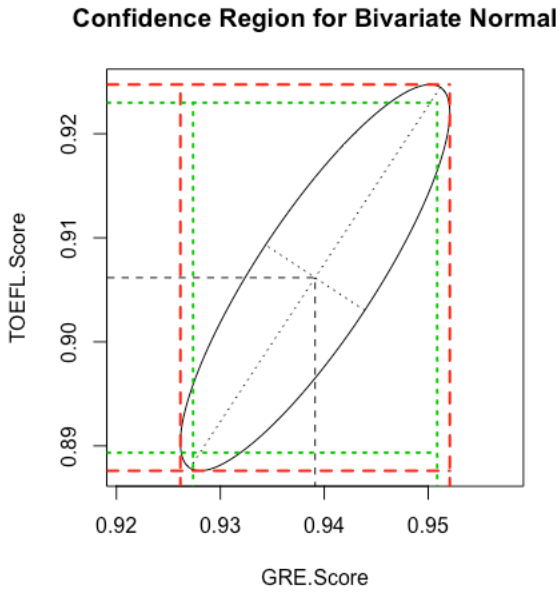
Define T^2-intervals as $\vec{\bar{x}_j} - s_j\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)} \leq \mu_j \leq \vec{\bar{x}_j} + s_j\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)}$ $_{(result 5.3 of page 225)}$

CI for μ_1 : $[0.9261557, 0.9520796]$
CI for μ_2 : $[0.8876064, 0.9247269]$

Bonferroni Method of Multiple Comparisons

defined as $\bar{x_j} - \frac{s_j}{\sqrt{n}}t_{n-1}(\frac{\alpha}{2p}) \leq \bar{x_j} + \frac{s_j}{\sqrt{n}}t_{n-1}(\frac{\alpha}{2p})$
CI for μ_1 : $[0.9273743, 0.950861]$
CI for μ_2 : $[0.8893513, 0.922982]$

Both Confidence Intervals Together



Two-Sample Inference

Means for each group:

	$\vec{\mu_1}$ (no research)	$\vec{\mu_2}$ (research)
GRE.Score	0.9097059	0.9473739
TOEFL.Score	0.8665909	0.9142262
University.Rating	2.5636364	3.5464286
SOP	2.9181818	3.7321429
LOR	3.0954545	3.7892857
CGPA	8.2347273	8.8449286
Chance.of.Admit	0.6349091	0.7899643

Two independent p-variate random samples with the same population covariance
Test $H_0 : \vec{\mu_1} = \vec{\mu_2}$

Hotelling's T^2 Testing

Hotelling's T^2 sampling distribution implies the equivalent $H_0 : \vec{\mu_1} - \vec{\mu_2} = \vec{\delta_0}$
$T^2 = ((\vec{x_1} - \vec{x_2}) - \vec{\delta_0})^T((\frac{1}{n_1} + \frac{1}{n2})S_{pooled})^{-1}((\vec{x_1} - \vec{x_2}) \sim \frac{(n_1+n_2-2)p}{n_1+n_2-1-p}F_{p,n_1+n_2-1-p}(\alpha)$
Decision rule: reject H_0 if $T^2 > \frac{(n_1+n_2-2)p}{n_1+n_2-1-p}F_{p,n_1+n_2-1-p}(\alpha)$

$T^2 = 55.07828 > 15.73899$, reject H_0

Rejecting the null hypothesis promotes checking significant components by:

95% **Component-Wise Simultaneous Confidence Intervals**

$(\bar{x_{1j}} - \bar{x_{2j}}) - \sqrt{\frac{(n_1+n_2-2)p}{n_1+n_2-1-p} F_{p,n_1+n_2-1-p}(\alpha) \sum(\frac{1}{n}) S_{pooled,j}} \leq \mu_{1j} - \mu_{2j} \leq (\bar{x_{1j}} - \bar{x_{2j}}) + \sqrt{\frac{(n_1+n_2-2)p}{n_1+n_2-1-p} F_{p,n_1+n_2-1-p}(\alpha) \sum(\frac{1}{n}) S_{pooled,j}}$

GRE.Score	[-0.05928446 ,-0.01605167]
TOEFL.Score	[-0.08284477, -0.01242580]
University.Rating	[-1.79789797 ,-0.16768645]
SOP	[-1.53551361 ,-0.09240847]
LOR	[-1.37644171 ,-0.01122063]
CGPA	[-1.02136026 ,-0.19904233]
Chance.of.Admit	[-0.24844278 ,-0.06166760]

95% **Bonferroni Correction Simultaneous Confidence Intervals**

$(\bar{x_{1j}} - \bar{x_{2j}}) - S_{pooled,j} \sqrt{\frac{1}{n1} + \frac{1}{n2}} t_{n_1+n_2-2}(\frac{\alpha}{2p}) \leq \mu_{1j} - \mu_{2j} \leq (\bar{x_{1j}} - \bar{x_{2j}}) + S_{pooled,j} \sqrt{\frac{1}{n1} + \frac{1}{n2}} t_{n_1+n_2-2}(\frac{\alpha}{2p})$

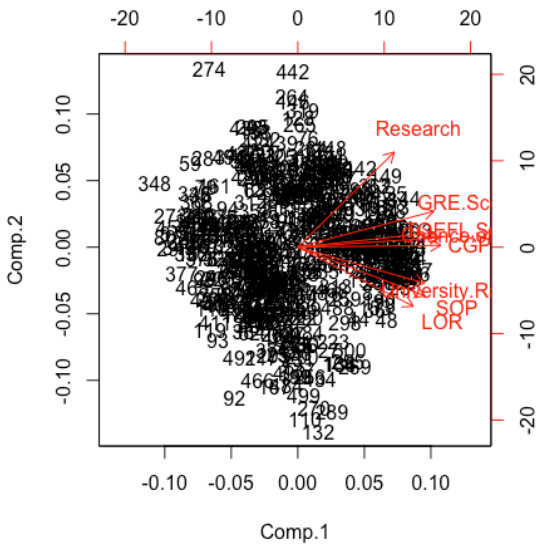
GRE.Score	[-0.05263986 ,-0.02269627]
TOEFL.Score	[-0.07202183 ,-0.02324874]
University.Rating	[-1.54734498 ,-0.41823944]
SOP	[-1.31371766 ,-0.31420442]
LOR	[-1.16661604 ,-0.22104630]
CGPA	[-0.89497530, -0.32542729]
Chance.of.Admit	[-0.21973665 ,-0.09037374]

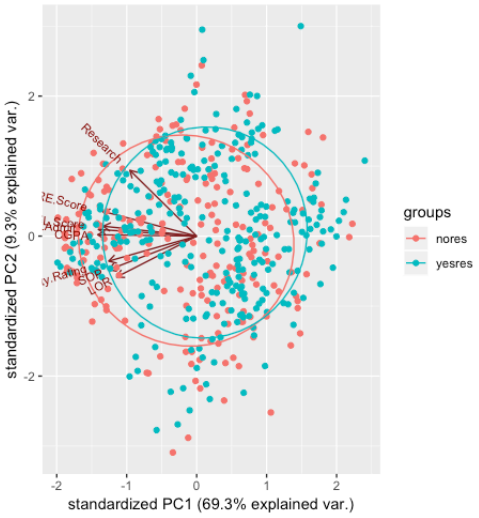
Alternatively, Bonferroni correction also rejects if:

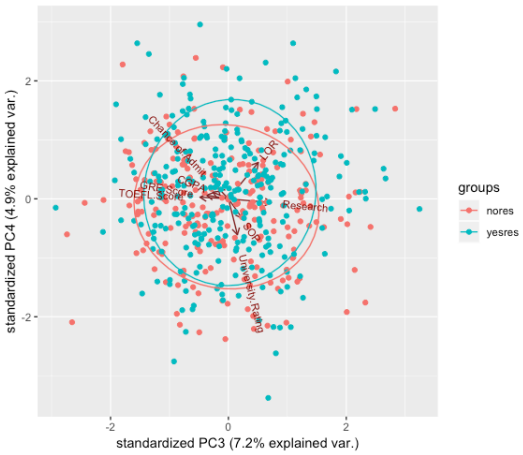
$MAX_{1 \leq j \leq p} |\frac{(\bar{x_{1j}} - \bar{x_{2j}})}{S_{pooled,j} \sqrt{\frac{1}{n1} + \frac{1}{n2}}}| \geq t_{n_1+n_2-2}(\frac{\alpha}{2p}) = 2.747761$

GRE.Score	6.913191
TOEFL.Score	5.367320
University.Rating	4.783394
SOP	4.475320
LOR	4.032453
CGPA	5.887783
Chance.of.Admit	6.586969

Principal Component Analysis







Contribution to First Principal Component

On the topic of the first principal component specifically, we can find out which variate contributes more to its determination, based on loadings and correlations.

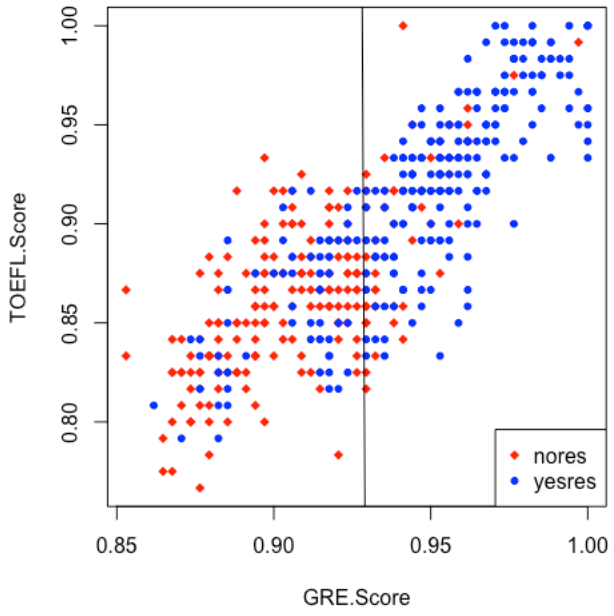
Correlation coefficients between first principal component and its variables, based on population covariance matrix where $\lambda_1 = 2.7789425219$
$Corr(Y_k, X_j) = \frac{Cov(Y_k, X_j)}{\sqrt{Var(Y_k)Var(X_j)}} = v_{kj}\sqrt{\frac{\lambda_k}{\sigma_{jj}}}$

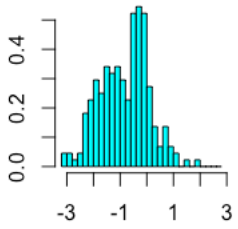
	Comp.1 (loadings)	Comp. 2(loadings)	$\sqrt{\sigma_{jj}}$	$v_{kj}\sqrt{\frac{\lambda_k}{\sigma_{jj}}}$
GRE.Score	0.374	0.269	0.03322102	18.76715155
TOEFL.Score	0.371	0.107	0.05068223	4.63936352
University.Rating	0.349	-0.271	1.14351180	0.15991720
SOP	0.350	-0.365	0.99100362	0.16034346
LOR	0.318	0.443	0.92544957	0.11716590
CGPA	0.393		0.60481280	0.04021808
Research	0.265	0.712	0.49688408	0.01547562
Chance.of.Admit	0.390		0.14114040	0.04095256

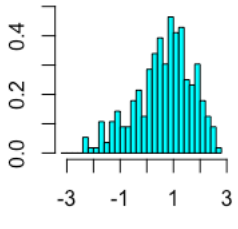
The proportion of total variance due to the first principal component:

$$\frac{Var(Y_1}{Var(Y_1)+...+Var(Y_p)} = \frac{\lambda_1}{\lambda_1+...+\lambda_p} = \frac{\lambda_1}{\sigma_{11}+...+\sigma_{pp}} = 69.3\%$$

Linear Discriminant Analysis







The equal costs and equal priors discriminant function $\hat{y} = \hat{a}^T x = (\bar{x_1} - \bar{x_2})^T S_{pooled}^{-1} x$ such that:
$\bar{y_1} = \hat{a}^T \bar{x_1}$, $\bar{y_2} = \hat{a}^T \bar{x_2}$

Group means:

	GRE.Score	TOEFL.Score	University.Rating	SOP	LOR	CGPA	Chance.of.Adm
0	0.9097059	0.8665909	2.563636	2.918182	3.095455	8.234727	0.6349091
1	0.9473739	0.9142262	3.546429	3.732143	3.789286	8.844929	0.7899643

Coefficient of Linear Discriminants:

	LD1 Coefficients	$\bar{y_1}$	$\bar{y_2}$
GRE.Score	27.40858991	21.12437	22.59511
TOEFL.Score	-4.65896494		
University.Rating	0.10320122		
SOP	0.04983113		
LOR	0.06374563		
CGPA	-0.40450929		
Chance.of.Admit	4.64907822		

Define $\hat{m} = \frac{1}{2}(\bar{x_1} - \bar{x_2})^T S_{pooled}^{-1}(\bar{x_1} + \bar{x_2}) = \frac{1}{2}(\bar{y_1} + \bar{y_2}) = 21.85974$

The decision rule or Fisher's rule becomes:
allocate x_0 to π_1 if $\hat{y_0} \geq \hat{m} = 21.85974$
allocate x_0 to π_2 if $\hat{y_0} < \hat{m} = 21.85974$

Confusion matrix

	0	1
0	155	65
1	58	222

Error rates or misclassification probabilities provide judgments on the quality of classification procedures. Without any constraint on the specefic type of procedure, there is an apparent error rate calculated from the confusion matrix. (pg 598)

The apparent error rate is $\frac{n_{1m}+n_{2m}}{n_1+n_2} = 24.6\%$, the proportion of items in the training set that are misclassified.

Interpretation

Making inferences about the population mean, we consider a sample with "8" variates, 8 numerical and 1 categoral variable if it was considered as the factor in the two sample classification. The statisitcal model considered here involves a random sample of 50 observations from the multivariate normal distribution. $\vec{X_1}, ..., \vec{X_n} \sim N_p(\vec{\mu}, \sigma)$

One page 5, we test to see if the sample mean vector is equivalent to another observed vector of different mean values. We reject this null hypothesis on the basis of the decision rule, that the Hotelling's T^2 statistic is larger than it distributed critical value. Another result to support this decision is by calculating the p-value or the probability of rejecting the null whens it is true.

Page 7 provides the analysis for testing for differences of two mean vectors or the equality of attributes for all variables from two groups. Since both set of 95% confidence intervals do not contain 0 and the maximum statistic $6.913191 > 2.544644$, we can conclude that the differences between population means for the two groups are significant.

Page 8 interprets mainly first principal component. In general, we can reduce the dimensions of the data without losing much information by accounting for total system variability by considering a small number k of the principal components, explained by the variance-covariance structure of a set of variables through a few linear combinations of these variables (pg 430). The scree plot explains that only the first two are the greatest contributors of information.

Since the proportion of $\frac{\lambda_1}{\sigma_{11}+...+\sigma_{pp}} = 69.3\%$ is not greater than 90%, we cannot say with certainty that the first component alone can replace the original variates without loss of information. Perhaps including the sum of additional components can fulfill the replacement, and thus dimension reduction. (pg 435)

Both groups, students with and without research experience are mainly characterized by CGPA, Chance.of.Admit, TOEFL and perhaps GRE scores.

It is also worthwhile to consider the plot for the less significant principal components to ensure that they should not be informative in their explanation of the total variance.

According to the coefficients, the variables all share equal weight in the first component. The correlation of GRE.score with the first component is significantly larger than that for the other variables, suggesting its utmost importance. Although the relative sizes of the coefficients are roughly the same, it can concluded that CGPA, Chance.of.Admit, TOEFL. Score and perhaps GRE. score contribute more to the determination of the first principal component.

Linear Discriminant Analysis is discussed on page 10. A scatter plot of the two groups, one with research and the other without, shows the each point as students and where they stand in terms of their GRE versus TOEFL scores. The decision line is one that separates the two groups. Pictorially, the figure describes Fisher's procedure for two populations. Points on the scatter plot are projected onto a line in the direction $\hat{a}$. (page 592)
The histogram provides the probability density functions for each group. Assuming that the pdf's $f_1(x), f_2(x)$ are multivariate normal and since Y is the linear combination of normal random variables, $f_1(y)$ and $f_2(y)$ are univariate normal, meaning $f_1(y) \sim N(\mu_{1y}, \sigma_y^2)$ and $f_2(y) \sim N(\mu_{2y}, \sigma_y^2)$ (page 596)

The overlap of the density functions are the misclassifications probabilities based on Y. Where group 0's density visits group 1 is P(2|1), P(misclassifying a π_1 as π_2) and vice versa.

Conclusion

Concerning the Hotelling's test for equality of means, it is evident that the means are not the same for the respective groups with and without research; the confidence interval for "chance.of.admit" does not contain zero, meaning that the differences are not the same. Examining both intervals, it is seen that none of the intervals contain zero, indicating significant differences between both groups. If our initial hypothesis was students with research had a better outlook on being admitted to the graduate program, the testing supports that claim. Besides research, there are variables that contribute much more information in comparison to other various variables. The variable of topic here is the one with maximum variance. According to the plots for the principal component scores, variables that contribute significantly include the gpa, and examination scores among the few. Separately, they cannot account for the total variability but as whole, the addition of respective eigenvalues may exceed the accepted .90 proportionality. If a new student is wondering where they may stand on being accepted to graduate program this fall, it may be reassuring to know the new student's allocation by linear discriminants.