Multivariate Data Analysis

Ryan Smith

Multiple Linear Regression

Introduction:

For my first data set analysis, I will perform multiple linear regression on the data from table 8.5 in the texbook. The available data for each of 61 cities is the population in thousands, percent of population with professor degrees, percent of people employed over the age of 16, percent of people employed in Government jobs, and median home value($100,000). My goal is to fit a multiple linear regression model to find the relationship between my response variable(median home value) and 2 variables which will be my x_1(population), x_2(professional degree percent), . I will assume that the responses and explanatory variates satisfy the following model:

$$Y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \ldots + \beta_r z_{ir} + \epsilon_i, \quad i = 1, \ldots, n,$$

where e_i is iid N(0, sigma^2). and use the least squares estimate to find the best fitting regression coefficients for my data. I will then calculate the R^2, sample variance, and various other confidence regions based on the least squares estimate.

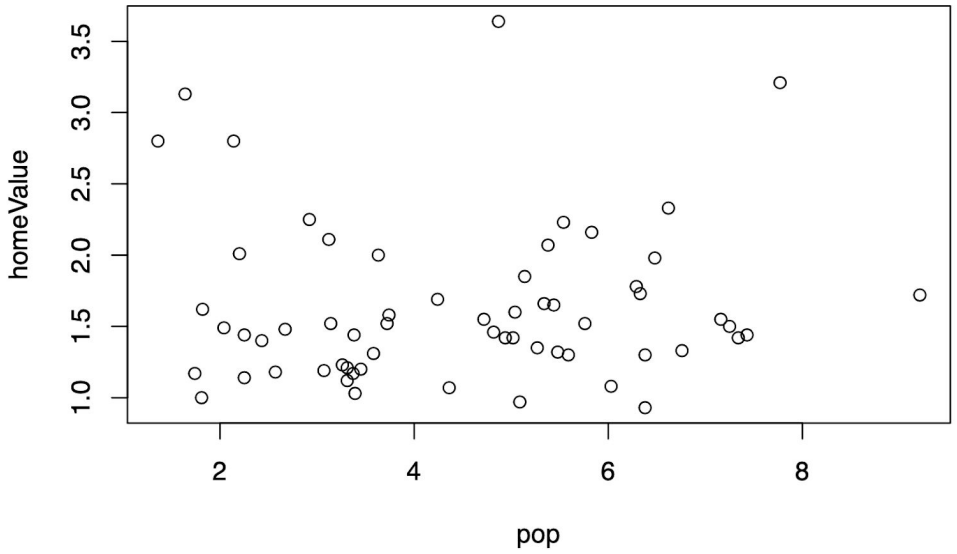
Summary:

A header of my data is:

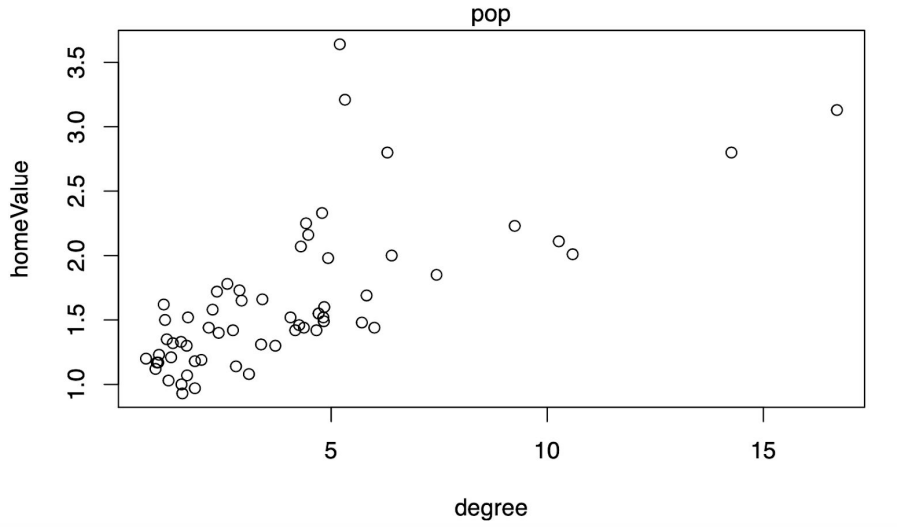
```
##    pop degree homeValue
## 1 2.67   5.71      1.48
## 2 2.25   4.37      1.44
## 3 3.12  10.27      2.11
## 4 5.14   7.44      1.85
## 5 5.54   9.25      2.23
## 6 5.04   4.84      1.60
```

. I use the third column as my response variable(homeValue), the first column(pop) as my first explanatory variate(c_1), and the second column(percent of population with professional degree) as my second explanatory variate(x_2).

A plot of x_1(pop) vs Y(homeValue in $100,000) is:



We can see a trend for homevalue to increase as the population increases.
A plot of x_2(percent of people with professional degrees) vs Y(homevalue in $100,000):



Shows a much stronger trend for homevalue to increase as the percent of people with degrees increases.
The mean homeValue is 1.63($100,000). The mean population is 4.469(1,000). The mean percent of population with professional degree is 3.96%.

I add a column of ones to my data to give me the design matrix Z:

```
##          pop degree
## [1,] 1 2.67   5.71
## [2,] 1 2.25   4.37
## [3,] 1 3.12  10.27
## [4,] 1 5.14   7.44
## [5,] 1 5.54   9.25
## [6,] 1 5.04   4.84
```

Analysis:

My least squares estimate , $\hat{\vec{\beta}} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z} \vec{Y}.$, is

```
##                [,1]
##         0.89565387
## pop     0.05021947
## degree 0.13009840
```

, where Beta_0 is 0.896, Beta_1 is.0502, Beta_2 is 0.13.

Using this least squares estimate, I can calculate the R^2 statistics using the formula R^2 = (Explained sum of squares) / (Total sum of squares). The R^2 is 0.4955, which means that 49.5% of the variance of the response variable(Median house price) is explained by the population and percent of people with professional degrees.

The formula for the estimated sample variance is:

$$\hat{\sigma}^2 := \frac{1}{n-r-1} \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \frac{1}{n-r-1} \|\hat{\vec{\epsilon}}\|^2 = \frac{1}{n-r-1} \|\vec{Y} - \boldsymbol{Z}\hat{\vec{\beta}}\|^2.$$

.

This gives us an estimated sample variance of 0.166.

The estimated covariance of beta_hat is

$$\widehat{\mathrm{Cov}}(\hat{\vec{\beta}}) = \hat{\sigma}^2 (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}.$$

which is

```
##                                pop        degree
##         0.027735675 -4.168575e-03 -1.610244e-03
## pop    -0.004168575  8.471869e-04  9.653053e-05
## degree -0.001610244  9.653053e-05  2.975163e-04
```

One at a time confidence intervals for beta_j, alpha = 0.05:

$$\beta_j \in \left[\hat{\beta}_j - \hat{\sigma}\sqrt{\omega_{jj}}t_{n-r-1}(\frac{\alpha}{2}), \hat{\beta}_j + \hat{\sigma}\sqrt{\omega_{jj}}t_{n-r-1}(\frac{\alpha}{2})\right]$$

j = 1: [-0.008043467 , 0.1084824],
j = 2: [0.09557145 , 0.1646253].

Confidence regions based simultaneous confidence intervals for beta_j, alpha = 0.05;

$$\beta_j \in \left[\hat{\beta}_j - \hat{\sigma}\sqrt{\omega_{jj}}\sqrt{(r+1)F_{r+1,n-r-1}(\alpha)}, \hat{\beta}_j + \hat{\sigma}\sqrt{\omega_{jj}}\sqrt{(r+1)F_{r+1,n-r-1}(\alpha)}\right], \quad j = 0, 1, \ldots, r.$$

j = 1: [-0.03358825 , 0.1340272]
j = 2: [0.08043347 , 0.1797633]

Bonferroni correction based simultaneous confidence intervals, alpha = 0.05:

$$\beta_j \in \left[\hat{\beta}_j - \hat{\sigma}\sqrt{\omega_{jj}}t_{n-r-1}\left(\frac{\alpha}{2(r+1)}\right), \hat{\beta}_j + \hat{\sigma}\sqrt{\omega_{jj}}t_{n-r-1}\left(\frac{\alpha}{2(r+1)}\right)\right], \quad j = 0, 1, \ldots, r.$$

j = 1: [-0.02153958 , 0.1219785]
j = 2: [0.08757358 , 0.1726232]

F-test, H_0: beta_1 = beta_2 = 0:

Compare $\frac{1}{\hat{\sigma}^2}(\boldsymbol{C}\hat{\vec{\beta}})^\top \left(\boldsymbol{C}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{C}^\top\right)^{-1}(\boldsymbol{C}\hat{\vec{\beta}})$ = 28.48, and $(r-q)F_{r-q,n-r-1}(\alpha).$ = 6.311

Confidence interval for z_0^T*Beta, alpha = 0.05:

$$\vec{z}_0^\top\vec{\beta} \in \left[\vec{z}_0^\top\hat{\vec{\beta}} - \hat{\sigma}t_{n-r-1}(\frac{\alpha}{2})\sqrt{\vec{z}_0^\top(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\vec{z}_0},\ \vec{z}_0^\top\hat{\vec{\beta}} + \hat{\sigma}t_{n-r-1}(\frac{\alpha}{2})\sqrt{\vec{z}_0^\top(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\vec{z}_0}\right]$$

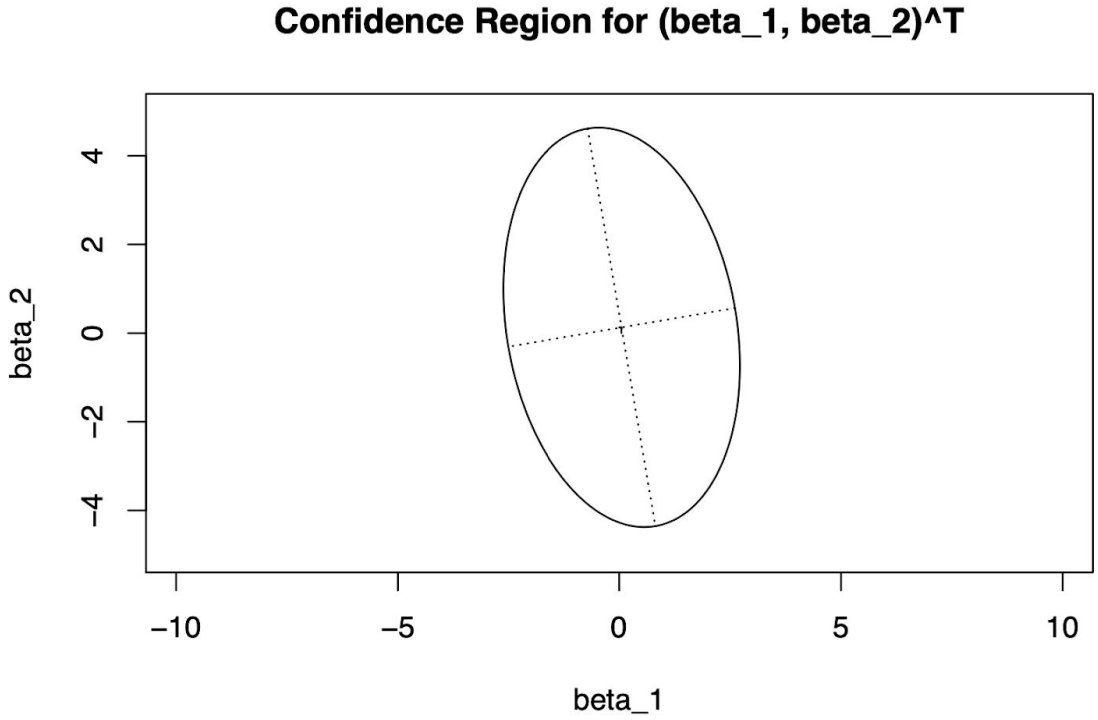
= [7.743934 , 16.04615].

Prediction interval for Y_0 = z_0^T beta + epsilon_0, alpha = 0.05:

$$Y_0 \in \left[\vec{z}_0^\top \hat{\vec{\beta}} - \hat{\sigma} t_{n-r-1}(\frac{\alpha}{2})\sqrt{1+\vec{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\vec{z}_0},\ \vec{z}_0^\top \hat{\vec{\beta}} + \hat{\sigma} t_{n-r-1}(\frac{\alpha}{2})\sqrt{1+\vec{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\vec{z}_0}\right]$$

= [7.664441 , 16.12565].

A confidence region for (beta_1, beta_2)^T, for alpha = 0.05:



Conclusion:

The least squares estimate for beta_hat is (0.896, 0.050, 0.130). This means that the intercept for this line, beta_hat_0 is 0.896(or $89,600). Beta_hat_1 is 0.05, which means that for every increase of 1 in x_1(population in thousands), we expect the median home value (Y) to increase by 0.05(or $5,000). Beta_hat_2 is 0.130, which means that for every increase of 1 in x_2(percent of people with professional degrees), we expect the median home value(Y) to increase by 0.130(or $13,000).

The R^2 is 0.4955, which means that 49.55% of the change in the response variable(median home price) is explained by the two explanatory variables.

The estimated sample variance is 0.166.

I calculated the one at a time confidence intervals for beta_j, j = 1,2 using three methods: one-at-a-time, confidence regions based simultaneous confidence intervals, and Bonferroni correction based simultaneous confidence intervals. For beta_1, the

width of the interval for one at a time is 0.116, and for beta_2 it is .0691. For confidence region based simultaneous confidence intervals, the width of the interval for beta_1 is 0.1675, and for beta_2 it is 0.099. For Bonferroni correction based simultaneous confidence intervals, for beta_1 the width is 0.1004, and for beta_2 it is 0.085. This means that for both beta_1 and beta_2, the width of the confidence intervals at level (1-alpha) is smaller using the Bonferroni correction based confidence interval, so it performs better than confidence region based simultaneous confidence intervals.

For the F-test: null hypothesis(h0): beta_1 = beta_2. Reject if the critical value is greater than the F-statistic. The F-statistic is 28.48, and the critical value is 6.3, so do not reject the null hypothesis at level alpha.

The (1- alpha) confidence interval for a new observation z_0^T * beta is [7.743934 , 16.04615]. This corresponds to a confidence interval for the expectation of Y_0 given z_0.

The (1- alpha) prediction interval for Y_0 = z_0^T beta + epsilon_0 is [7.664441 , 16.12565].

Two Sample Test & Linear Discriminant Analysis:

Introduction:

For these analysis, I will use the data from table 11.6 in the textbook. This data set contains data on 85 applicants for a Graduate School of Business. It contains three set of data, one for class 1; students that should be admitted, one for class 2: students that should not be admitted, and one for class 3: students that are borderline. I removed the third class because I am trying to do classification into two classes for the LDA. For each of these classes it contains data on the applicant's GPA(x_1), and their GMAT score(x_2). For the two sample test, I will test whether or not the null hypothesis, h0 : mu1 = m2, which is equivalent to finding a confidence interval for mu1 - mu2. For LDA, I will classify a new observation x_0 to either class 1 or class 2, and find the error rate and plot the decision boundary.

Summary:

The pooled sample variance is used to estimate the population variance:

$$S^2_{pooled} = \frac{n_1 - 1}{n_1 + n_2 - 2} S_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2^2. \quad =$$

```
##                GPA         GMAT
## GPA    0.03504253 -0.01094352
## GMAT -0.01094352  0.41822196
```

The sample means for admit are:

```
##         GPA        GMAT Acceptance
##    3.382857    5.630000    1.000000
```

```
##         GPA        GMAT Acceptance
##    2.482500    4.470714    2.000000
```

Where acceptance = 1 means the applicant should be admitted, and acceptance = 2 means that the applicant should not be admitted.

Analysis:

The equation for the Hotelling's T^2 for two sample test h0: mu1 - mu2 = delta0:

$$T^2 = \left((\bar{\vec{x}}_1 - \bar{\vec{x}}_2) - \vec{\delta}_0\right)^\top \left(\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \boldsymbol{S}_{pooled}\right)^{-1} \left((\bar{\vec{x}}_1 - \bar{\vec{x}}_2) - \vec{\delta}_0\right)$$

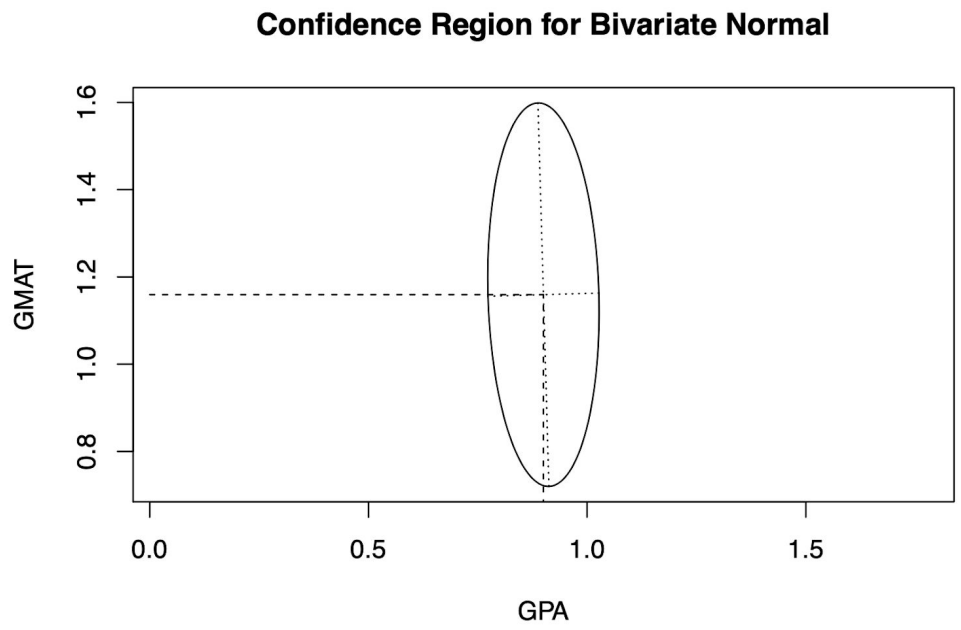
Delta0 is zero in this test because we are testing for the different of means being zero; ie mu1 = mu2. T^2 = 393.8942. The decision rule for this test is:

$$\text{We reject } H_0 : \vec{\mu}_1 - \vec{\mu}_2 = \vec{\delta}_0 \text{ at the level of } \alpha \text{ if}$$

$$T^2 > \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - 1 - p} F_{p, n_1 + n_2 - 1 - p}(\alpha),$$

The critical value is 6.462.

A confidence region for mu1 - mu2 is:



This confidence ellipse doesn't contain zero, so we can interpret it as mu1 =! Mu2.

Using the simultaneous confidence intervals and Bonferroni Simultaneous confidence intervals to check the significant components:

```
## 95% simultaneous confidence interval

##           [,1]     [,2]
## GPA  0.7731684 1.027546
## GMAT 0.7198915 1.598680

## 95% Bonferroni simultaneous confidence interval

##           [,1]     [,2]
## GPA  0.7850059 1.015708
## GMAT 0.7607860 1.557785
```

Using LDA on the data set with prior probabilities ½,½, gives us:

```
## Coefficients of linear discriminants:
##              LD1
## GPA  -5.0483359
## GMAT -0.6546848
```

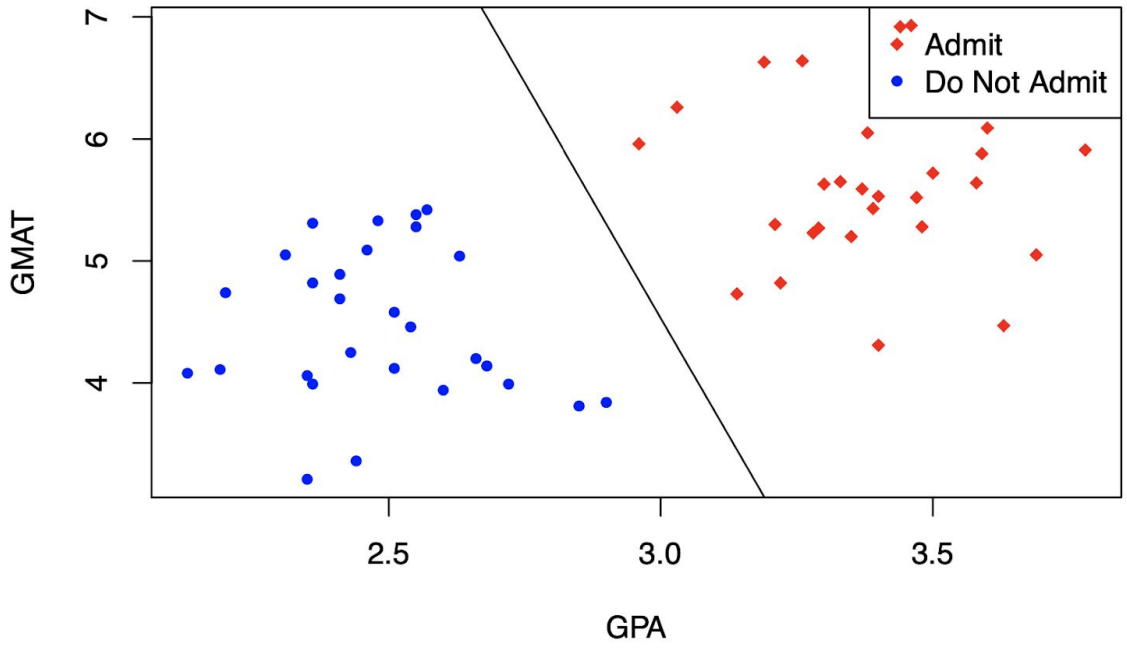
Using the predict function calculated the predicted value for each applicant, and if it's greater than zero assigns it to class 1(admit), and if lower than zero assigns it to class 2(do not admit).

```
##           LD1
## 1  -0.7334571
## 2  -0.8368953
## 3  -1.2996838
## 4  -1.9476754
## 5  -3.8229791
## 6  -3.8926692
```

The confusion matrix for the LDA is:

```
##        1  2
##    1  28  0
##    2   0 28
```

A graph of the observations and decision(classification) boundary is:



Conclusion:

For the two sample test, we performed the Hotellings T^2 test for equal population means. If the T^2 statistic is greater than the critical value, we reject H0: mu1 - mu2 = 0 at the level alpha. The T^2 was 393.9, and the critical value was 6.46, so we reject h0 and conclude at level alpha that the population means , mu1 and mu2, are not equal.

Then I graphed the confidence region for mu1-mu2(mean difference), which is an ellipse with axis x_1 and x_2. This also supports our conclusion that mu1 =! Mu2, because zero is not contained within the confidence ellipse. I then used simultaneous confidence intervals, and Bonferroni simultaneous confidence intervals to check the significant components. The resulting simultaneous confidence intervals are almost the same for each method. Neither component-wise simultaneous confidence intervals contain zero, so they have significant differences.

For the LDA, the fitted linear discriminant analysis line has coefficients GPA(x_1) = -5.048, and GMAT(x_2) = -0.6546. Then I predicted the classes of the original data using the predict function, and it assigns a class to each observation based on whether or not it is above the decision line. The confusion matrix assigns all 56 observations correctly, which is an unfortunate consequence of the data being pre-separated into well defined admit and do-not-admit clases. The plot of the observations with axis x_1 and x_2 highlights this, with the decision boundary perfectly separating the classes with no misclassified objects,

Principal Component Analysis:

Introduction:

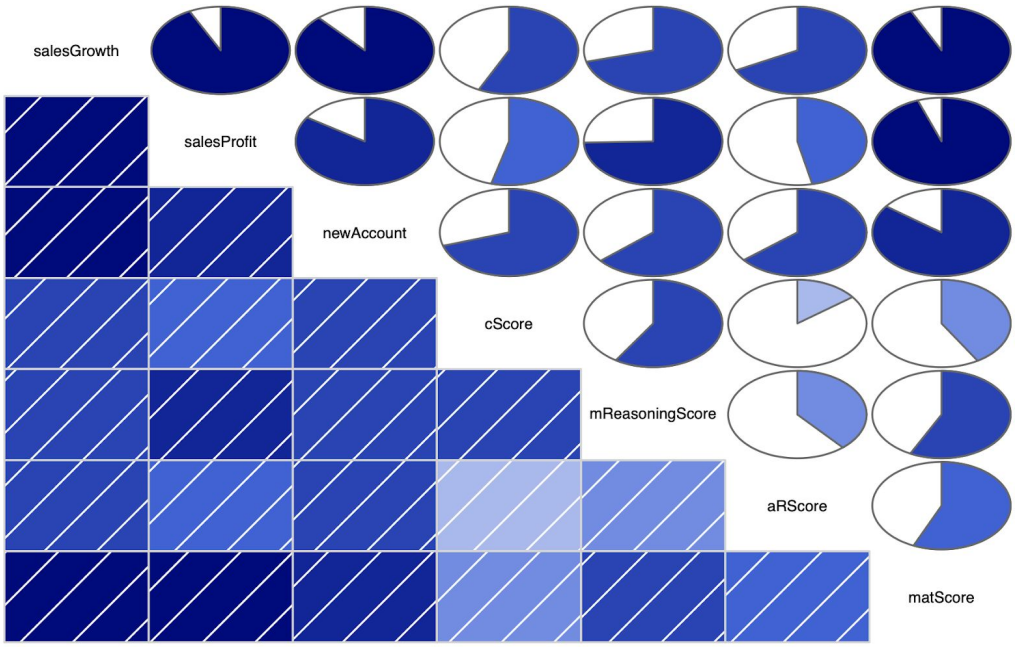
For this analysis, I will perform principal component analysis on the data from table 9.12 in the textbook. I aim to reduce the number of dimensions of the data set while still maintaining a significant portion of the original variance. I will do this by analysing the amount of variance each principal component contributes, and finding a good balance between capturing a large amount of the total variance while not having too many dimensions.

Summary:

The data set I am using is table 9.12 from the textbook. It features 7 columns on job applicants, x_1(sales growth), x_2(sales profits), x_3 (number of new accounts), x_4(creativity score), x_5(reasoning score), x_6(abstract reasoning score), (math score). A header of this data set is:

```
##   salesGrowth salesProfit newAccount cScore mReasoningScore aRScore matScore
## 1        93.0        96.0       97.8      9              12       9       20
## 2        88.8        91.8       96.8      7              10      10       15
## 3        95.0       100.3       99.0      8              12       9       26
## 4       101.3       103.8      106.8     13              14      12       29
## 5       102.0       107.8      103.0     10              15      12       32
## 6        95.8        97.5       99.3     10              14      11       21
```

To visualize the correlation of the variable prior to doing any analysis, a correlogram is:



Analysis:

First I used the princomp function in R to find the principal components using the correlation matrix. I chose to use the correlation matrix because it is equivalent to first scaling the data to standard and then using the covariance matrix. If I had just used the covariance matrix, variables which are in the 100's like salesProfit would always overpower variables like cScore which only go up to 10's in terms of variance. The importance of each principle component is given:

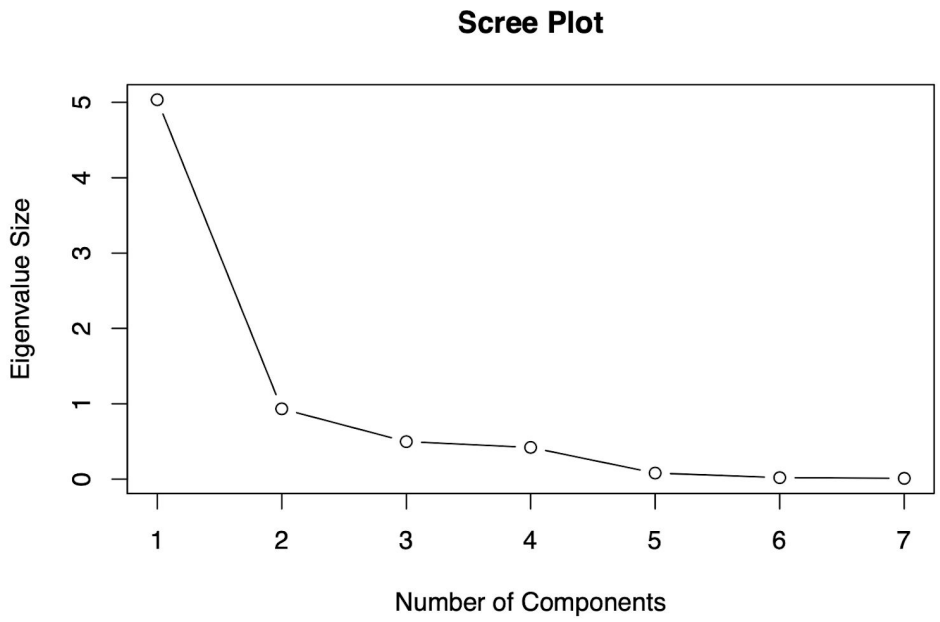
```
## Importance of components:
##                            Comp.1    Comp.2     Comp.3     Comp.4    Comp.5
## Standard deviation     2.2437909 0.9661864 0.70563429 0.64903427 0.2846760
## Proportion of Variance 0.7192283 0.1333594 0.07113139 0.06017793 0.0115772
## Cumulative Proportion  0.7192283 0.8525877 0.92371910 0.98389702 0.9954742
##                             Comp.6      Comp.7
## Standard deviation     0.142620593 0.106488339
## Proportion of Variance 0.002905805 0.001619967
## Cumulative Proportion  0.998380033 1.000000000
```

To see each of the original variates' contribution to each principal component, we compare the loadings. The higher the absolute value of a loading, the more the variate contributes to the corresponding principal component.

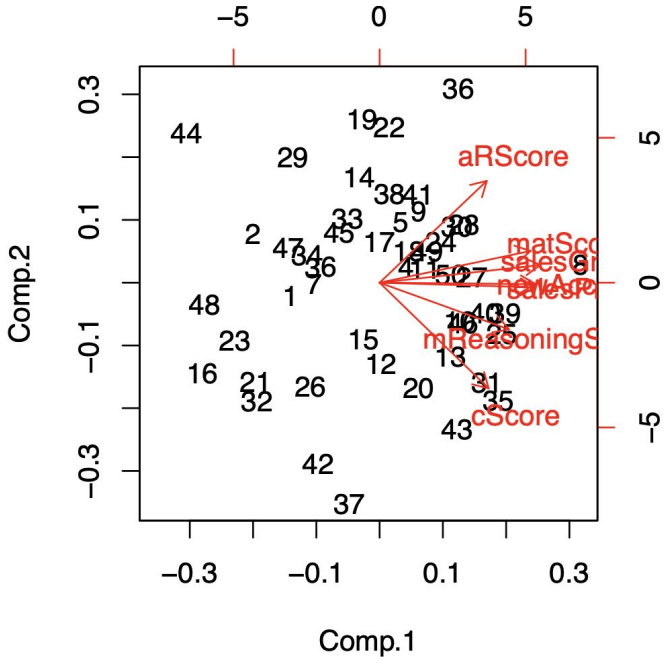
```
## Loadings:
##                  Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## salesGrowth       0.434  0.112                0.632  0.337  0.528
## salesProfit       0.420         0.442               -0.785
## newAccount        0.421        -0.204  0.325 -0.701  0.157  0.399
## cScore            0.294 -0.668 -0.451  0.303  0.261 -0.114 -0.300
## mReasoningScore   0.349 -0.295        -0.847 -0.174  0.197
## aRScore           0.289  0.642 -0.604 -0.154        -0.236 -0.228
## matScore          0.407  0.200  0.434  0.246         0.371 -0.636
```

We can see that all of the original variates contribute a good amount to principal component one, whereas cScore and aRScore contribute much more to principal component 2 than the others.

Next I do a scree plot of eigenvalue size versus number of components, to help me decide how many principal components are necessary when reducing the data set and attempting to not lose too much data.



A plot of the principal component scores for the sample data in the space of the first two principal components :



Because the data labels are just numbers that do not have any intrinsic meaning, this data set does not provide useful information.

Conclusion:

From the “Importance of components chart”, we see that principal component 1 captures 71.9% of the original variance, principal component 2 captures 13.3% of the original variance, and principal component 3 captures 7.11%.

From the scree plot, we see that the elbow occurs at two principal components. So, two principal components is a reasonable number of principal components to use in the data reduction. It reduces the dimensions of the data from 7 to 2, while still capturing 85.26% of the variance contained in the original data.