Statistical Analysis Project

Joseph Gonzalez
UC Davis
STA 135 Multivariate Data Analysis
Professor Xiaodong Li

March 19, 2020

1 Data set 1:

1.1 Introduction

For this section, I will perform data analysis on table 7.1(page 372) using multi-linear regression techniques. This data set reflects real estate data, which was gathered from 20 homes in Milwaukee. My goals for analyzing this data set are to generate summary statistics of the variables, calculate the least squares estimate for the coefficients($\hat{\vec{\beta}}$), determine the R^2 statistic, find $\hat{\sigma}^2$, solve for $\widehat{cov(\beta})$, and examine sums of squares. I also plan to conduct simultaneous confidence intervals for the coefficients and test whether the full model(all variables) is statistically better than the reduced model. All tests will be conducted at a 95% confidence level($\alpha = 0.05$) and we will assume that the observations were obtained randomly(independence assumption).

1.2 Summary

The data set contains 20 observations on 3 variables. The three variables are Total Dwelling Size, Assessed Value, and Selling Price. For this analysis, the Total Dwelling Size and Assessed Value are the explanatory variables and Selling Price is the response variable.

Variable Name	Variable Type	Mean	Standard Dev.	Variance
Selling Price(Y)	numeric	76.55	8.071620	65.151053
Total dwelling Size(z_1)	numeric	16.22	2.681084	7.188213
Assessed Value(z_2)	numeric	63.06	7.385639	54.547658

Table 1: Summary Statistics Table

Table 1 shows that the variables are numeric and it provides estimates for the mean, standard deviation, and variance. These values describe the data's spread. The mean will be used for conducting prediction intervals.

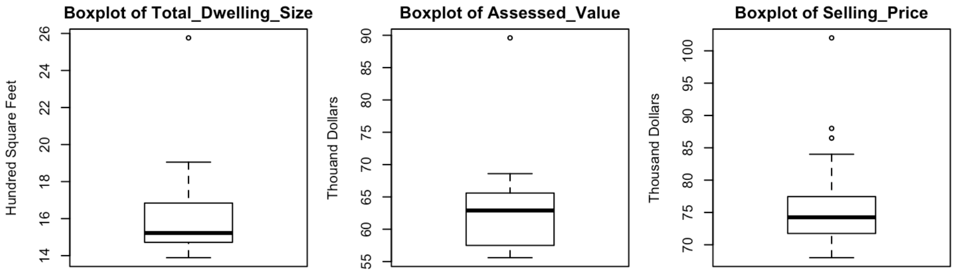

Figure 1: Real Estate Data Set Box plots

In figure 1 provides that quantiles, minimum, maximum, and possible outliers of the data. Selling price has the highest mean value and has three possible outliers. Total dwelling size and assessed value both have one possible outlier.

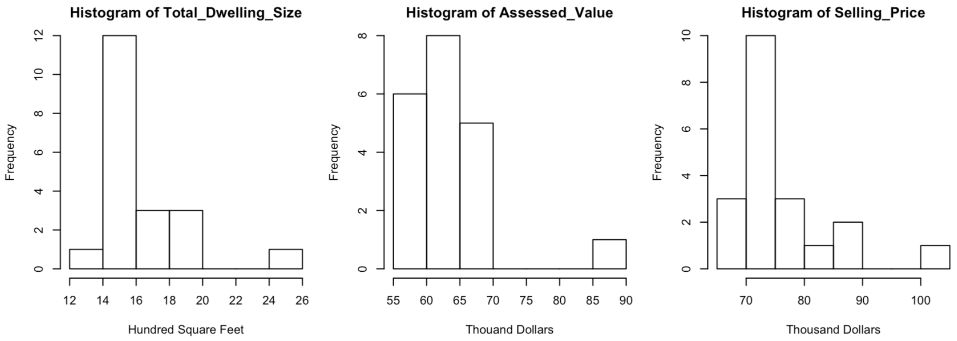

Figure 2: Real Estate Data Histograms

In figure 2, we can see that the distributions for the variables are right-skewed. These plots also show the ranges where most of the data points belong to. For example, we see that most values for total dwelling size are between 14 and 16. Furthermore, these plots also show the possible outliers. These possible outliers are illustrated by the bars that are much farther away from the rest of the data.

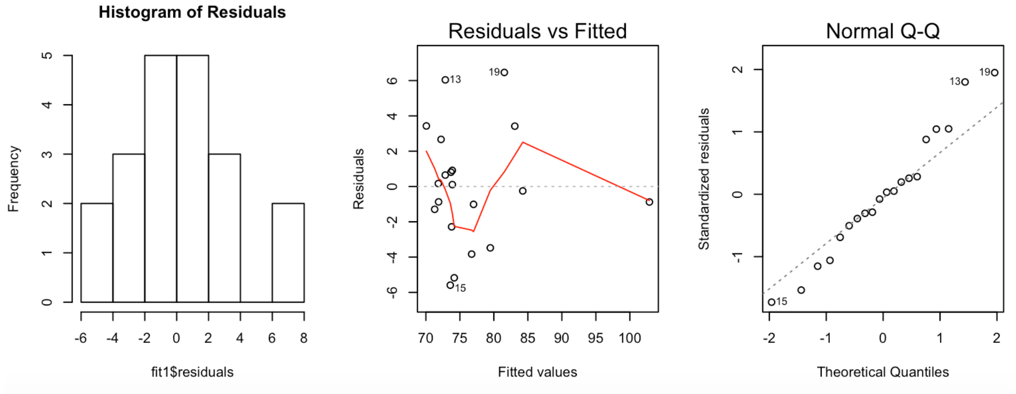

Figure 3: Residual Plots

Figure 3 contains three residual plots that reveal whether the residuals are i.i.d normal and have constant variance. The residuals histogram shows that the residuals are approximately normally distributed with a mean of 0. However, the Q-Q plot shows that the residuals belong to a distribution that has more probability in the tails(heavy-tailed) than the traditional normal distribution. The Residuals vs Fitted plot shows that there is some non-constant variance occurring among the residuals. These setbacks may be caused by some of the outliers in the data. In practice, we would usually try to alleviate these problems but, for the sake of this project's purpose, I will continue with the analysis without making further adjustments.

1.3 Analysis

The first objective is to construct a multiple linear regression model. This model will look like:

$$Y_i = \beta_0 + \beta_1 * Z_{i1} + \beta_2 * Z_{i2} + \mathcal{E}$$

$$i = 1, .., n$$

To find the values for the coefficients in this model, we use the least square estimate formula below.

$$\hat{\beta} = (Z^T Z)^{-1} Z^T \vec{Y}$$

The design matrix(Z) is a 20 by 3 matrix, where the first column is all ones, the second column is the dwelling size data and the third column is the assessed value data.

$$\hat{\beta} = \begin{bmatrix} 30.967 \\ 2.634 \\ 0.0452 \end{bmatrix}$$

After using the formula, the result is the 3 by 1 matrix(above) with three values. The first value is Y-intercept, which is the selling price of the house when dwelling size and assessed values are zero. This has no practical meaning because we cannot have a zero dwelling size or a zero assessed value. The other values, 2.634 and 0.0452, are the coefficients of dwelling size and assessed value. The values for these coefficients suggest that dwelling size and assessed values have a positive association with selling price.

With these values, the fitted regression line is:

$$\hat{Y_i} = 30.967 + 2.634 * Z_{i1} + 0.0452 * Z_{i2}$$

1.3.1 Sums of Squares

Next, we evaluate the multi-linear regression's fit with the data through the sum of squares.

	Sum squares	Degrees of Freedom	Mean Squares
Explained Sum of Squares	1032.875	2	
Residual Sum of Squares	204.9949	17	12.06
Total Sum of Squares	1237.87	19	

Table 2: Sums of Squares

In table 2, we can see that there are more explained sum of squares than unexplained sum of squares. This means that more of the variation in the sales price is explained with its relationship with dwelling size and assessed value. Since there are more explained sum of squares than unexplained sum of squares, we can determine that this regression has a relatively good fit on the data set. We can quantify this fit by the r-squared value:

$$R^2 = 0.834397$$

From table 2, we can also get an estimate for σ^2. This value is located in the mean squares column. Generally, an estimate for σ^2 is $\hat{\sigma^2} = \frac{ESS}{TSS} = \frac{1}{n-r-1}||\hat{\vec{\mathcal{E}}}||$.

$$\hat{\sigma}^2 = 12.05853$$

Later, this value will be used to calculate the confidence intervals for the coefficients. It is also used to calculate the $c\hat{o}v(\beta)$, which is equal to $\hat{\sigma}^2(Z^TZ)^{-1}$.

$$c\hat{o}v(\beta) = \begin{bmatrix} 62.129 & 3.068 & -1.765 \\ 3.068 & 0.617 & -0.207 \\ -1.765 & -0.2074 & 0.0813 \end{bmatrix}$$

1.3.2 Confidence Intervals

Using $\hat{\sigma}^2$, $\hat{\beta}_j$(j=1,2) and a t-quantile($\alpha = 0.05$ and 17 d.f.), we calculate the one-at-a-time confidence intervals:

$$\beta_1 \in (0.9769312, 4.291868)$$

$$\beta_2 \in (-0.5564991, 0.6468668)$$

The C.I for β_1 suggests that its value is significantly different from 0. On the contrary, the C.I. for β_2 suggests that its value is not significantly different from zero. This indicates that the assessed value could be dropped from the model.

Next, we can calculate the C.I. based on simultaneous methods. The first method is the confidence region. This test uses an F distribution quantile as a component to determine the width of the C.I.

Confidence Region

$$\beta_0 \in (6.556739, 55.37639)$$

$$\beta_1 \in (0.2015372, 5.067262)$$

$$\beta_2 \in (-0.8379773, 0.9283451)$$

The confidence intervals(above) show that β_0 and β_1 are significantly different from 0. For β_2, we see that this interval contains 0, which suggests it is not significantly different than 0. Figure 4 shows an illustration of the confidence region for β_1 and β_2. In this figure, we can see that 0 resides in the ellipse for β_2.

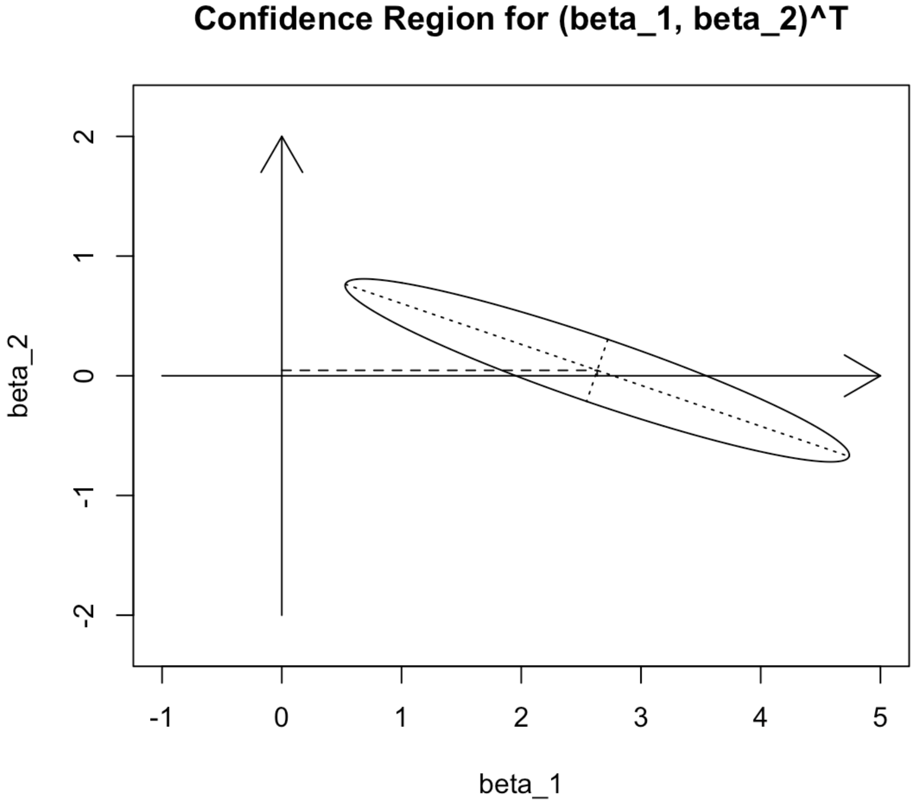

Figure 4: Confidence Region For β_1 and β_2

Bonferroni correction is another method for simultaneous confidence intervals. This method uses the same formula as the one-at-a-time confidence interval, but this method adjusts for the increase in type 1 error rate by dividing α by r+1(r is the number of coefficients).

Bonferonni C.I.s

$$\beta_0 \in (10.03934, 51.89379)$$

$$\beta_1 \in (0.5486385, 4.720161)$$

$$\beta_2 \in (-0.711975, 0.8023427)$$

Similar to the previous confidence intervals, we see that the confidence interval for the coefficient of assessed values contains 0. Next, we will conduct an F-test to see if the assessed value can be dropped from the model.

1.3.3 F-test For Assessed Value

In this section, the full and reduced model to determine whether the reduced model is statistically better than the full. These models look like:

$$\textit{Full Model}: \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 Z_{i1} + \hat{\beta}_2 Z_{i2}$$

$$Reduced\ Model: \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 Z_{i1}$$

$$i = 1, ..., 20$$

Using R, the following F-statistic and critical value were calculated:

$$F^* = 0.0251$$

$$F_{1,17}(0.05) = 4.45$$

Since the F^* is less than $F_{1,18}$, we fail to reject the null hypothesis. Therefore, there is sufficient evidence to suggest that assessed value can be dropped from the model.

1.4 Conclusion

In this analysis, we first obtained the least square estimates for the dwelling size and assessed value variables. These values were both positive and suggested that the variables have a positive association with selling price. For every 100 feet squared increase in dwelling size, we could expect the selling price to increase by 2.634 thousand dollars on average, holding assessed value constant. For every 1 thousand dollar increase in assessed value, we could expect the selling price to increase by 0.0452 thousand dollars on average, holding dwelling size constant. We also found that there was more explained sum of squares that residual sum of squares. This resulted in a R^2 value of 0.834, which suggests the model has a good fit on the data. Lastly, we formed multiple simultaneous confidence intervals for the coefficients and saw that all the intervals for assessed values contained 0. This lead to the F-test comparison between the full model and reduced model without assessed values. The result suggested that that the reduced model is significantly better than the full model and that assessed value could be dropped from the model.

2 Data set 2:

2.1 Introduction

For this section, I will conduct a two-sample test and LDA on table 6.9(page 344). This data set is based on a study by Jolicoeur and Mosimann, which evaluated the relationship of size and shape for painted turtles. This study's concepts and results are important to the overall analysis of body variation among living organisms. My goals for analyzing this data set are to generate summary statistics of the variables, conduct a multivariate two-sample test using Hotelling's T^2, create a confidence ellipse for the mean difference, and examine the simultaneous confidence intervals using Bonferroni's correction and confidence regions. I also plan to use linear discriminant analysis to set rules for the categorization of new observations($\vec{x}_0$). In addition, I will form diagrams illustrating LDA and calculate the apparent and expected actual error rate.

Variable Name	Variable Type	Mean	Standard Dev.	Variance
Length	Integer	113.4	11.780	138.766
Width	Integer	88.29	7.074	50.042
Height	Integer	40.71	3.355	11.259

Table 3: Summary Statistics For 24 Males

Variable Name	Variable Type	Mean	Standard Dev.	Variance
Length	Integer	136	21.249	451.520
Width	Integer	102.58	13.105	171.732
Height	Integer	52.04	8.046	64.737

Table 4: Summary Statistics For 24 Females

2.2 Summary

First, the full data set was divided into two data sets(one for female turtles and one for male turtles). Tables 3 and 4, contain the summary statistics of length, width, and height for the male and female turtles. The most apparent observation is that the mean values for length, width, and height for the females are larger than the males. This is furthered illustrated in figure 5.

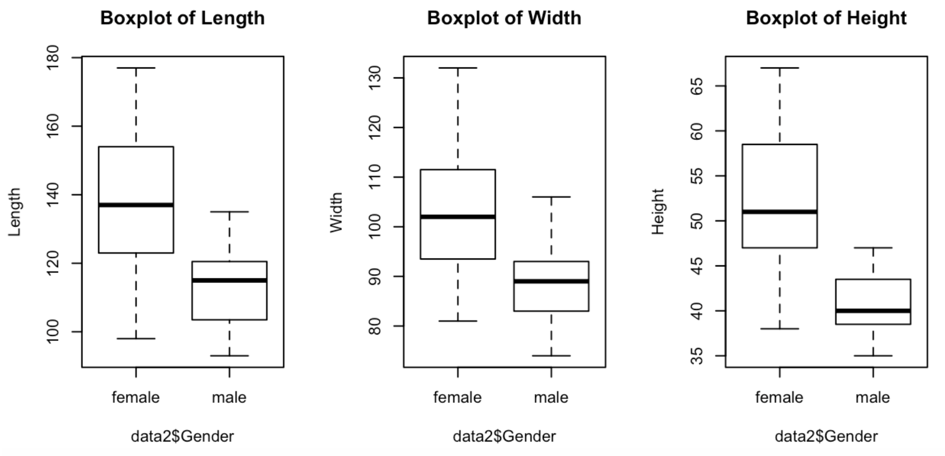

Figure 5: Painted Turtles Data Boxplots

In figure 5, we can see that the means for variables are larger for females than for males. This may suggest that there is a significant difference in the variable means between females and males. To test this, we will conduct several two-sample tests.

2.3 Analysis

2.3.1 Two Sample Hotelling's T^2 Test

To test the difference in means, we first calculate the Hotelling's T^2 and compare it to the F-critical values. Using R, the values are:

$$T^2 = 72.38162$$

$$F_{3,97}(0.05) = 8.833461$$

$$T^2 > F_{3,97}(0.05)$$

Since the T^2 value is more than the F critical values, we can reject the null hypothesis at $\alpha = 0.05$. This suggests that there is a difference between the variables' mean values for male and female turtles.

2.3.2 Bonferroni's Correction for Simultaneous Confidence Intervals

In addition to Hotelling's T^2 two-sample test, we can conduct simultaneous confidence intervals for the difference in means. The Bonferroni method is preferred because it has a smaller confidence width than the confidence region.

$$Length \rightarrow \mu_{11} - \mu_{21} \in (-34.98915, -10.344185)$$

$$Width \rightarrow \mu_{12} - \mu_{23} \in (-21.84471, -6.738628)$$

$$Height \rightarrow \mu_{13} - \mu_{23} \in (-15.75477, -6.911898)$$

We are overall 95 percent confident that $\mu_{11} - \mu_{21}$ is between -34.98915 and -10.344185, $\mu_{12} - \mu_{23}$ is between -21.84471 and -6.738628, and $\mu_{13} - \mu_{23}$ is between -15.75477 and -6.911898. We see that the confidence intervals are in a negative range, which further supports that the variable means for the females are larger than the males.

2.3.3 Linear Discriminant Analysis

Next, we conduct linear discriminant analysis to set boundaries in the data to classify a new observation $\vec{x}_0$ to either the female sample or male sample. Using Fisher's Rule, for a new observation:

$\vec{x}_0$, it is assigned to the males group if:

$$(\vec{x}_0 - \vec{x}_1)^T S_{pooled}^{-1} (\vec{x}_0 - \vec{x}_1) \leq (\vec{x}_0 - \vec{x}_2)^T S_{pooled}^{-1} (\vec{x}_0 - \vec{x}_2)$$

$\vec{x}_0$, it is assigned to the females group if:

$$(\vec{x}_0 - \vec{x}_1)^T S_{pooled}^{-1} (\vec{x}_0 - \vec{x}_1) \geq (\vec{x}_0 - \vec{x}_2)^T S_{pooled}^{-1} (\vec{x}_0 - \vec{x}_2)$$

Figure 6 provides pairwise plots for the male and female samples.

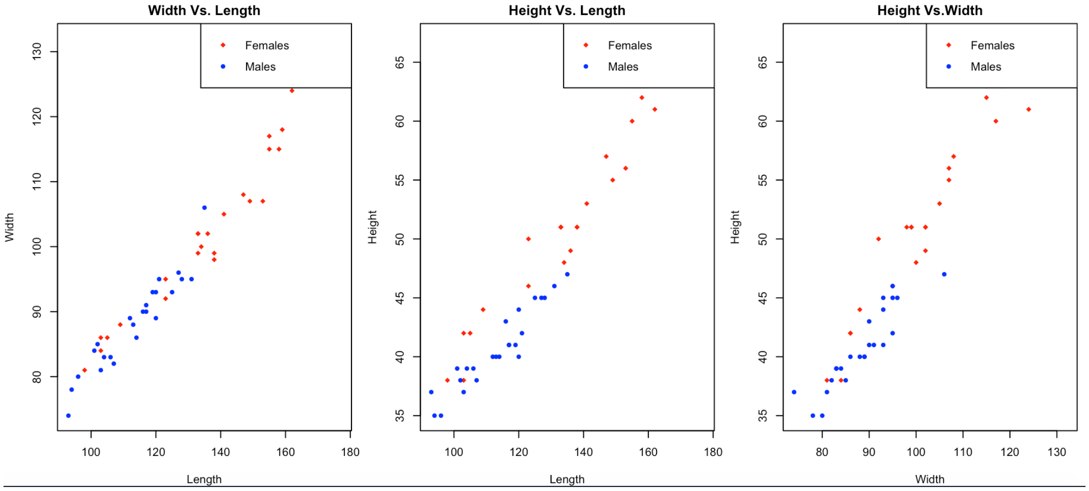

Figure 6: Pairwise Plots

In figure 6, the red data points represent females and the blue represent males. Intuitively, we can see that the data points almost are divided on their own. The female turtles generally have larger values than the males in all three categories. Next, figure 7 illustrates the boundary points for the two classes.

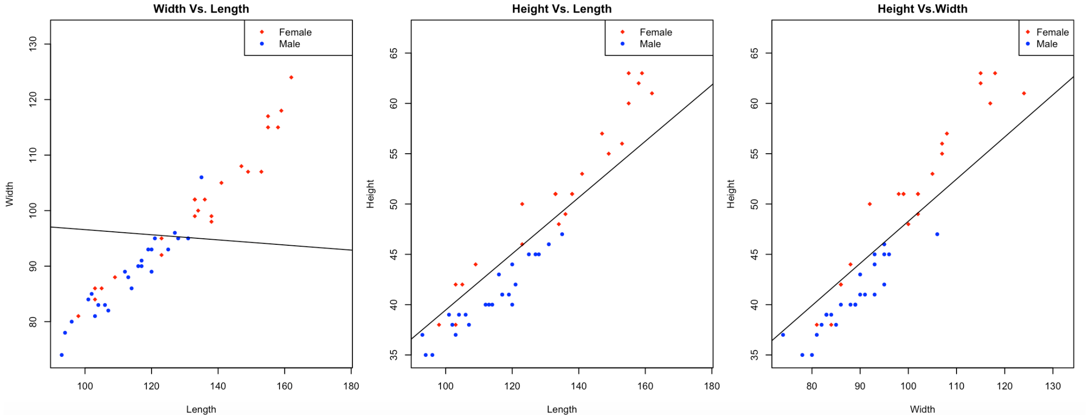

Figure 7: LDA Plots

Generating these boundary lines can help decide which class a new observation belongs to. To best describe the categorization process, one must imagine a new data point added to these plots without a blue or red color. If the point resides on the side of the line that has mostly red points, then it is suggested that this new point belongs to the female class. If the point resides on the side of the line that has mostly blue points, then it is suggested that this new point belongs to the male class. It is also important to point out that not all red and blue data points belong to the side with the majority of female or male turtles. These are misclassifications and we can evaluate these using a confusion matrix and Lachenbruch's holdout.

	Female	Male
Female	17	7
Male	2	22

Confusion Matrix for Width Vs. Length

The first confusion matrix identifies the number of misclassifications for categorization of width and length for female and male turtles. The off-diagonal values indicate the misclassifications and we can see that 4 female turtles were misclassified. In this case, the proportion of cases that are incorrectly predicted is 0.1875(apparent error rate). Using Lachenbruch's holdout, we calculate the expected actual error rate to be 0.2293.

	Female	Male
Female	20	4
Male	0	24

Confusion Matrix For Height Vs. Length

For height vs. length, the apparent error rate is 0.08333 and the expected actual error rate is also 0.08333.

	Female	Male
Female	17	7
Male	0	24

Confusion Matrix For Height Vs. Width

For height vs. width, the apparent error rate is 0.1458 and the expected actual error rate is 0.1667.

2.4 Conclusion

From the beginning, we saw that the length, width, and height mean values were larger in female turtles than in male turtles. We later tested this idea using Hotelling's T^2 and Bonferroni's confidence intervals. These tests provided enough evidence to support that the mean values for length, width, and height of females are generally larger than the mean values for length, width, and height of males. Then, we conducted linear discriminant analysis to categorize new observations into these sample groups. Using a pairwise comparison of the variables with respect to gender, we created a boundary line that provided a visual of how we would categorize a new observation. Using confusion matrices and Lachenbruch's holdout, we calculate the error rates of our classification scheme. The rates are roughly low with the lowest being 0.0833(apparent and expected actual error rates) from the height vs. length confusion matrix. In addition, this classification rarely misclassified males(only 2 misclassified for width vs. length). With the total number of observations in mind, the error rates suggest that the classification design performs well at categorizing new observations.

3 Data set 3:

3.1 Introduction

For this section, I will conduct principle component analysis on table 8.5(page 474) to explain the variance-covariance structure of these variables by their linear combinations.

This data is based on census-tract data. My goals for analyzing this data set are to summarize the data, generate the first and second principal components, determine the proportion of total sample variance due to the sample principal components, and compare the variates' contributions to the determination of the sample principal component based on loadings. I also intend to interpret the principle components and explain the number of dimensions the data can be summarized in.

3.2 Summary

Variable Name	Variable Type	Mean	Standard Dev.	Variance
Total Population	Numeric	4.469	1.843	3.397
Professional degree	Numeric	3.962	3.110	9.673
Pct. Employed(16+)	Numeric	71.42	7.458	55.626
Gov. Employment	Numeric	26.91	9.438	89.067
Med. Home Value	Numeric	1.636	0.564	0.3286

Table 5: Summary Statistics Table

Table 5 provides estimates for the variables in the census-tract data set(61 observations). Professional degree represents the percentage of individuals that have a degree, Pct. Employed(16+) represents the percentage of individuals that are over 16 and employed, gov. employment represents the percentage of government employed individuals, and med. home values represents the median home value($100,000s). From this table, we can see that the pct. employed(16+) has the largest mean and gov. employment has the largest variance.

3.3 Analysis

In this section, the goal is to find the number of components that can accurately summarize the original data set. In this case, the principal components will have the following form:

$$Y_i = \vec{v}_i\vec{X} = v_{i1}X_1 + v_{i2}X_2 + v_{i3}X_3 + v_{i4}X_4 + v_{i5}X_5$$
$$i = 1, ..., 5$$

Using R, we can calculate the proportion of variance, eigen vectors, and eigen values. Table 6 reveals the eigen values for each component. It can be seen that the property $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_5 > 0$. These eigen values can also be used to calculate the total proportion of sample variance due to its corresponding principal component. Table 7 shows the proportion of variance for each principal component and the cumulative proportion.

	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
Eigen Values	1.9919183	1.3675266	0.8641573	0.5350610	0.2413367

Table 6: Eigen Values

Importance of components:					
	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
Standard Deviation	1.411	1.1694	0.9296	0.7315	0.4913
Proportion of Variance	0.3984	0.2735	0.1728	0.1070	0.04827
Cumulative Proportion	0.3984	0.6719	0.8447	0.9517	1

Table 7: Principal Components

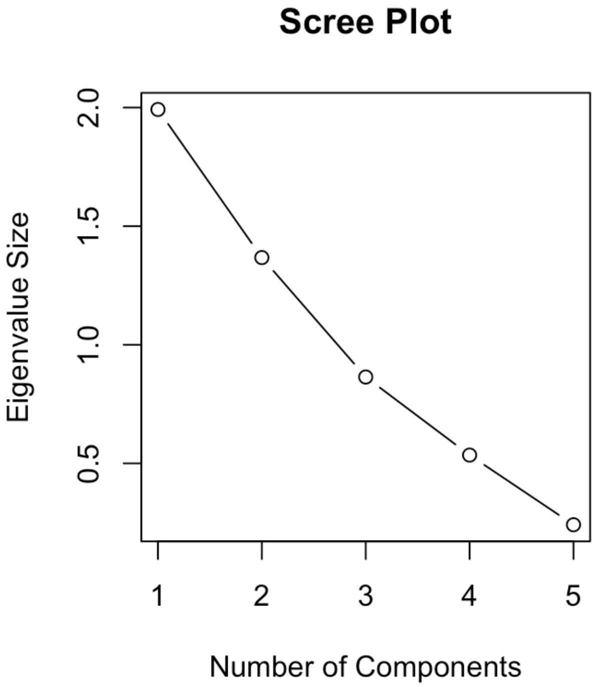

Figure 8: Scree Plot

The scree plot, figure 8, plots the eigen values(y) against the number of components(x). In this plot, we look for a bend in the line that suggests the number of principal components that should be used. The slight bend appears to occur at 3 or 4. While this plot may suggest 3 principal components, a more conservative approach would use 4 principal components(>90% cumulative variance).

In table 7, we see that 3 principle components summarize 84.47% of the information from the original data set and 4 principle components summarize 95.17% of the information from the original data set. It is a general rule to use the number of principal components that summarize over 90% of the data. Therefore, this table suggests 4 principal components are sufficient.

Table 7 presents the loadings for the random variables $X_1, .., X_5$. These loadings or

Loadings:					
	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
Total Population	0.263	0.465	0.784	0.217	0.235
Professional Degree	-0.593	0.326	-0.164	-0.145	0.703
Employed(16+)	0.326	0.605	-0.255	-0.663	-0.194
Government Employed	-0.479	-0.252	0.551	- 0.572	-0.277
Median Home Value	-0.493	0.500		0.407	-0.580

Table 8: Loadings

their absolute values allow us to compare the variables contribution to the determination of the principal components. For the first principal component, professional degree contributes the most(0.593). For the second principal component, employed(16+) contributes the most(0.605). For the third, fourth, and fifth principal components, total population(0.784), Employed(16+)(0.663), and professional degree(0.703) contribute the most.

3.4 Conclusion

For principal component analysis, the goal is to explain the variance-covariance structure of a set of variables by linear combinations of these variables. In other words, we hope to reduce the number of columns in the data set without losing too much information. For this specific situation, we conducted PCA on a census-tract data set with 5 variables and 61 observations. Using R, we generated the eigen values and the importance of components. This table revealed that 4 principle components are sufficient to summarize the data. While the scree plot suggests the 3 principal components are sufficient, it is safer to use four because the cumulative proportion of variance is above 90%. In addition, the loadings provided details for which variables contributed more to the determination of the principle components.

A R Code

```
###data set 1 bone marrow
# data
data1= read.table("T7-1.DAT", header=FALSE)
colnames(data1)=c("Total_Dwelling_Size", "Assessed_Value", "Selling_Price")
attach(data1)

#Summary of data
sapply(data1,class)
summary(data1)
sapply(data1,var)
par(mfrow=c(1,3))
xlabels=c("Hundred Square Feet", "Thouand Dollars", "Thousand Dollars")
#histogram
for (i in 1:3){
  hist(data1[,i], main=paste("Histogram of", names(data1[i])),xlab=xlabels[i])
}
par(mfrow=c(2,2))
#boxplot
for (i in 1:3){
  boxplot(data1[,i], main=paste("Boxplot of", names(data1[i])),ylab=xlabels[i])
}
#histogram for residuals
fit1=lm(Selling_Price˜ Total_Dwelling_Size+Assessed_Value, data = data1)
anova(fit1)
par(mfrow=c(1,3))
hist(fit1$residuals, main="Histogram of Residuals")
plot(fit1, which=1)
plot(fit1, which=2)

------------------------------------------------------
#Tests
n <- length(Selling_Price)
Z <- cbind(rep(1,n),as.matrix(data1[,1:2]))
Z
sapply(Z,class)
r <- dim(Z)[2]-1

# least square estimates
beta_hat <- solve(t(Z)%*%Z)%*%t(Z)%*%Selling_Price
beta_hat

#Sums of squares
#SSTO
```

```
Yadjust=data1$Selling_Price-mean(data1$Selling_Price)*rep(1,n)
SSTO=sum(Yadjust^2)

#ESS
Yadjusted2=fit1$fitted.values-mean(data1$Selling_Price)*rep(1,n)
SSRtotal=sum(Yadjusted2^2)
SSRtotal
#RSS
Rres1=0
sum(fit1$residuals^2)

#R squared value
R_square <- 1 - sum((Selling_Price - Z%*%beta_hat)^2)/
sum((Selling_Price-mean(Selling_Price))^2)
R_square

# sigma_hat_square
sigma_hat_square <- sum((Selling_Price - Z%*%beta_hat)^2)/(n-r-1)
sigma_hat_square

# estimated covariance of hat{beta}
cov_B = sigma_hat_square * solve(t(Z)%*%Z)
cov_B

#confidence interval for beta_1
alpha=0.05
j <- 2
cat('[',
    beta_hat[j+1] - qt(1-alpha/2, n-r-1)*sqrt(sigma_hat_square *
    solve(t(Z)%*%Z)[j+1,j+1]),',', beta_hat[j+1] + qt(1-alpha/2, n-r-1)*
    sqrt(sigma_hat_square * solve(t(Z)%*%Z)[j+1,j+1]),']')

# confidence region based simultaneous confidence intervals

j <- 0
cat('[',
    beta_hat[j+1] - sqrt((r+1)*qf(1-alpha, r+1, n-r-1))*
    sqrt(sigma_hat_square * solve(t(Z)%*%Z)[j+1,j+1]),
    ',',
    beta_hat[j+1] + sqrt((r+1)*qf(1-alpha, r+1, n-r-1))*
    sqrt(sigma_hat_square * solve(t(Z)%*%Z)[j+1,j+1]),
    ']')

j <- 1
cat('[',
    beta_hat[j+1] - sqrt((r+1)*qf(1-alpha, r+1, n-r-1))*
    sqrt(sigma_hat_square * solve(t(Z)%*%Z)[j+1,j+1]),
    ’,’,
    beta_hat[j+1] + sqrt((r+1)*qf(1-alpha, r+1, n-r-1))*
    sqrt(sigma_hat_square * solve(t(Z)%*%Z)[j+1,j+1]),
    ’]’)

j <- 2
cat(’[’,
    beta_hat[j+1] - sqrt((r+1)*qf(1-alpha, r+1, n-r-1))*
    sqrt(sigma_hat_square * solve(t(Z)%*%Z)[j+1,j+1]),
    ’,’,
    beta_hat[j+1] + sqrt((r+1)*qf(1-alpha, r+1, n-r-1))*
    sqrt(sigma_hat_square * solve(t(Z)%*%Z)[j+1,j+1]),
    ’]’)

# Bonferroni correction based simultaneous confidence intervals

j <- 0
cat(’[’,
    beta_hat[j+1] - qt(1-alpha/(2*(r+1)), n-r-1)*
    sqrt(sigma_hat_square * solve(t(Z)%*%Z)[j+1,j+1]),
    ’,’,
    beta_hat[j+1] + qt(1-alpha/(2*(r+1)), n-r-1)*
    sqrt(sigma_hat_square * solve(t(Z)%*%Z)[j+1,j+1]),
    ’]’)

j <- 1
cat(’[’,
    beta_hat[j+1] - qt(1-alpha/(2*(r+1)), n-r-1)*
    sqrt(sigma_hat_square * solve(t(Z)%*%Z)[j+1,j+1]),
    ’,’,
    beta_hat[j+1] + qt(1-alpha/(2*(r+1)), n-r-1)*
    sqrt(sigma_hat_square * solve(t(Z)%*%Z)[j+1,j+1]),
    ’]’)


j <- 2
cat(’[’,
    beta_hat[j+1] - qt(1-alpha/(2*(r+1)), n-r-1)*
    sqrt(sigma_hat_square * solve(t(Z)%*%Z)[j+1,j+1]),
    ’,’,
    beta_hat[j+1] + qt(1-alpha/(2*(r+1)), n-r-1)*
    sqrt(sigma_hat_square * solve(t(Z)%*%Z)[j+1,j+1]),
    ’]’)
```

```
# F-test
# H_0: beta_1 = beta_2 = 0

C <- matrix(c(0,0,1,0,0,1),2,3)

df_1 <- qr(C)$rank # df_1: rank of matrix R

f_stat <- (t(C%*%beta_hat)%*%
solve(C%*%solve(t(Z)%*%Z)%*%t(C))%*%(C%*%beta_hat)/df_1)/sigma_hat_square
f_stat

cval_f <- qf(1-alpha, 2, n-r-1)
cval_f


# t-test for single coefficient
# H_0: beta_j = 0, H_a: beta_j != 0

j <- 1
t_stat <- (beta_hat[j+1] - 0)/sqrt(sigma_hat_square * solve(t(Z)%*%Z)[j+1,j+1])
t_stat

alpha <- 0.05
cval_t <- qt(1-alpha/2, n-r-1)
cval_t

# confidence interval for z_0^T beta

z_0 <- c(1, mean(Z[,2]), mean(Z[,3]))

cat('[',
    z_0%*%beta_hat - sqrt(sigma_hat_square)*
    sqrt(t(z_0)%*%solve(t(Z)%*%Z)%*%z_0)*qt(1-alpha/2, n-r-1),
    ',',
    z_0%*%beta_hat + sqrt(sigma_hat_square)*
    sqrt(t(z_0)%*%solve(t(Z)%*%Z)%*%z_0)*qt(1-alpha/2, n-r-1),
    ']')

# prediction interval for Y_0 = z_0^T beta + epsilon_0

cat('[',
    z_0%*%beta_hat - sqrt(sigma_hat_square)*
    sqrt(1+t(z_0)%*%solve(t(Z)%*%Z)%*%z_0)*qt(1-alpha/2, n-r-1),
    ',',
    z_0%*%beta_hat + sqrt(sigma_hat_square)*
    sqrt(1+t(z_0)%*%solve(t(Z)%*%Z)%*%z_0)*qt(1-alpha/2, n-r-1),
    ’]’)

# Confidence Region for (beta_1, beta_2)ˆT

center <- beta_hat[2:3]
es<-eigen(C%*%solve(t(Z)%*%Z)%*%t(C))
e1<-es$vec %*% diag(sqrt(es$val))
r1<-sqrt(df_1*cval_f*sigma_hat_square)
theta<-seq(0,2*pi,len=250)
v1<-cbind(r1*cos(theta), r1*sin(theta))
pts<-t(center - (e1%*%t(v1)))
plot(pts,type="l",main="Confidence Region for
(beta_1, beta_2)ˆT",xlab="beta_1",ylab="beta_2",asp=1,
     xlim = c(-1,5),ylim=c(-1,1))
segments(0,center[2],center[1],center[2],lty=2) # highlight the center
segments(center[1],0,center[1],center[2],lty=2)
arrows(-1,0,5,0)
arrows(0,-2,0,2)

th2<-c(0,pi/2,pi,3*pi/2,2*pi)   #adding the axis
v2<-cbind(r1*cos(th2), r1*sin(th2))
pts2<-t(center-(e1%*%t(v2)))
segments(pts2[3,1],pts2[3,2],pts2[1,1],pts2[1,2],lty=3)
segments(pts2[2,1],pts2[2,2],pts2[4,1],pts2[4,2],lty=3)

#############################################
-------------------------------------------------
#dataset 2
# data
data2= read.table("T6-9.DAT", header=FALSE)
colnames(data2)=c("Length", "Width", "Height", "Gender")
attach(data2)

#boxplot
for (i in 1:3){
  boxplot(data2[,i]˜data2$Gender,
  main=paste("Boxplot of", names(data2[i])),ylab=xlabels[i])
}

##################Analysis

#### two-sample Hotelling’s T2 test  -------
```

```
males=which(data2$Gender=="male")
females=which(data2$Gender=="female")
Maledata <- data2[males,1:3]
Femaledata <- data2[females,1:3]
sapply(data2,class)
summary(Maledata)
summary(Femaledata)
sapply(Maledata[,1:3],var)
sapply(Maledata[,1:3],sd)
sapply(Femaledata[,1:3],var)
sapply(Femaledata[,1:3],sd)
par(mfrow=c(1,3))
xlabels=c("Length", "Width", "Height")

# now we perform the two-sample Hotelling T^2-test
n<-c(24,24)
p<-3
xmean1<-colMeans(Maledata)
xmean2<-colMeans(Femaledata)
d<-xmean1-xmean2
S1<-var(Maledata)
S2<-var(Femaledata)
Sp<-((n[1]-1)*S1+(n[2]-1)*S2)/(sum(n)-2)
t2 <- t(d)%*%solve(sum(1/n)*Sp)%*%d
t2

alpha<-0.05
cval <- (sum(n)-2)*p/(sum(n)-p-1)*qf(1-alpha,p,sum(n)-p-1)
cval


# to check the significant components
# simultaneous confidence intervals
wd<-sqrt(cval*diag(Sp)*sum(1/n))
Cis<-cbind(d-wd,d+wd)

cat("95% simultaneous confidence interval","\n")
Cis

#Bonferroni simultaneous confidence intervals
wd.b<- qt(1-alpha/(2*p),n[1]+n[2]-2) *sqrt(diag(Sp)*sum(1/n))
Cis.b<-cbind(d-wd.b,d+wd.b)
cat("95% Bonferroni simultaneous confidence interval","\n")
Cis.b
```

```
-----------------------------------------------------------------------------------
#LDA Analysis on Dataset 2
library(rrcov)

par(mfrow=c(1,3), mar=c(4,4,2,1))
plot(data2$Length,data2$Width,xlab="Length",ylab="Width",
     pch=rep(c(18,20),each=24),col=rep(c(2,4),each=24),main="Width Vs. Length")
legend("topright",legend=c("Females","Males"),pch=c(18,20),col=c(2,4),cex=1)

plot(data2$Length,data2$Height,xlab="Length",ylab="Height",
     pch=rep(c(18,20),each=24),col=rep(c(2,4),each=24),main="Height Vs. Length")
legend("topright",legend=c("Females","Males"),pch=c(18,20),col=c(2,4),cex=1)

plot(data2$Width,data2$Height,xlab="Width",ylab="Height",
     pch=rep(c(18,20),each=24),col=rep(c(2,4),each=24),main="Height Vs.Width")
legend("topright",legend=c("Females","Males"),pch=c(18,20),col=c(2,4),cex=1)


## Method 2: use function LDA in MASS package:

#Width Vs Length----------------------------------------------------------
library(MASS)
lda.obj <- lda(Gender~Length+ Width,data=data2,prior=c(1,1)/2)
plda <- predict(object=lda.obj,newdata=data2)

# Confusion matrix
table(data2[,4],plda$class)

#Expected actual error rate
n <- dim(data2)[1]
n_M <- 0
for (i in 1:n){
  lda.obj <- lda(Gender~ Length + Width,data=data2[-c(i),],prior=c(1,1)/2)
  plda <- predict(object=lda.obj,data2[c(i),])
  n_M <- n_M + (plda$class != data2[c(i),]$Gender)
}
n_M/n

#plot the decision line
gmean <- lda.obj$prior %*% lda.obj$means
const <- as.numeric(gmean %*%lda.obj$scaling)
slope <- - lda.obj$scaling[1] / lda.obj$scaling[2]
intercept <- const / lda.obj$scaling[2]

#Plot decision boundary
```

```
plot(data2[,1:2],pch=rep(c(18,20),each=24),
col=rep(c(2,4),each=24),main="Width Vs. Length")
abline(intercept, slope)
legend("topright",legend=c("Female","Male"),pch=c(18,20),col=c(2,4))

#Height V Length----------------------------------------------------------
lda.obj <- lda(Gender˜Length+ Height,data=data2,prior=c(1,1)/2)
plda <- predict(object=lda.obj,newdata=data2)

# Confusion matrix
table(data2[,4],plda$class)

#Expected Actual Error Rate
n <- dim(data2)[1]
n_M <- 0
for (i in 1:n){
  lda.obj <- lda(Gender˜ Length + Height,data=data2[-c(i),],prior=c(1,1)/2)
  plda <- predict(object=lda.obj,data2[c(i),])
  n_M <- n_M + (plda$class != data2[c(i),]$Gender)
}
n_M/n


#plot the decision line
gmean <- lda.obj$prior %*% lda.obj$means
const <- as.numeric(gmean %*%lda.obj$scaling)
slope <- - lda.obj$scaling[1] / lda.obj$scaling[2]
intercept <- const / lda.obj$scaling[2]

#Plot decision boundary
plot(data2[,c(1,3)],pch=rep(c(18,20),
each=24),col=rep(c(2,4),each=24),main="Height Vs. Length")
abline(intercept, slope)
legend("topright",legend=c("Female","Male"),pch=c(18,20),col=c(2,4))


#Height V. Width-------------------------------------------------------
?lda
lda.obj <- lda(Gender˜Width+ Height,data=data2,prior=c(1,1)/2)
plda <- predict(object=lda.obj,newdata=data2)

# Confusion matrix
table(data2[,4],plda$class)

#Expected Actual Error Rate
```

```
n <- dim(data2)[1]
n_M <- 0
for (i in 1:n){
  lda.obj <- lda(Gender~ Width + Height,data=data2[-c(i),],prior=c(1,1)/2)
  plda <- predict(object=lda.obj,data2[c(i),])
  n_M <- n_M + (plda$class != data2[c(i),]$Gender)
}
n_M/n

#plot the decision line
gmean <- lda.obj$prior %*% lda.obj$means
const <- as.numeric(gmean %*%lda.obj$scaling)
slope <- - lda.obj$scaling[1] / lda.obj$scaling[2]
intercept <- const / lda.obj$scaling[2]

#Plot decision boundary
plot(data2[,2:3],pch=rep(c(18,20),each=24),
col=rep(c(2,4),each=24),main="Height Vs.Width")
abline(intercept, slope)
legend("topright",legend=c("Female","Male"),
pch=c(18,20),col=c(2,4))


-----------------------------------------------
  #################################################
----------------------------------------------
#dataset 3 PCA
data3= read.table("T8-5.DAT", header=FALSE)
colnames(data3)=c("Tot_population",
"Profess_deg", "employed_16", "Gov_employ", "Med_home_value")
attach(data3)
sapply(data3,class)
summary(data3)
sapply(data3,var)
sapply(data3,sd)

# correlation matrix
census.pc <- princomp(data3, cor=TRUE)

summary(census.pc, loadings = TRUE)

# Showing the eigenvalues of the correlation matrix:
(census.pc$sdev)^2

# A scree plot:
```

```
plot(1:(length(census.pc$sdev)),  (census.pc$sdev)^2, type='b',
     main="Scree Plot", xlab="Number of Components", ylab="Eigenvalue Size")

# Plotting the PC scores for the sample data
#in the space of the first two principal components:
par(pty="s")
plot(census.pc$scores[,1], census.pc$scores[,2],
     xlab="PC 1", ylab="PC 2", type ='n',
     lwd=2, main="Principle Component Scores")
```