Diana Rueda

STA 135 – Multivariate Data Analysis

Xiaodong Li

Winter 2021

Final Project

Diana Rueda

3/19/2021

1. Multivariable Data Analysis

Introduction

The goal of this multivariable data analysis is to identify and explore the relationships as well as make inferences between the variables in dataset T1-7 through a multivariable linear regression model.

Dataset T1-7 contains average ratings of different factors over the course of treatment for radiotherapy patients. The columns are as follows:

Col. 1: x1 = number of symptoms Col. 2: x2 = amount of activity (1-5 scale) Col. 3: x3 = amount of sleep (1-5 scale) Col. 4: x4 = amount of food consumed (1-3 scale) Col. 5: x5 = appetite (1-5 scale) Col. 6: x6 = skin reaction (0, 1, 2 or 3)

For this analysis we will focus on the effects of the other variables on the number of symptoms a patient develops. We might be able to identify which behaviors contribute to a higher chance of developing symptoms while on radiotherapy.

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   symptoms = col_double(),
##   activity = col_double(),
##   sleep = col_double(),
##   eat = col_double(),
##   appetite = col_double(),
##   skin_reaction = col_double()
## )
```

Summary

The first six rows of our data look like this:

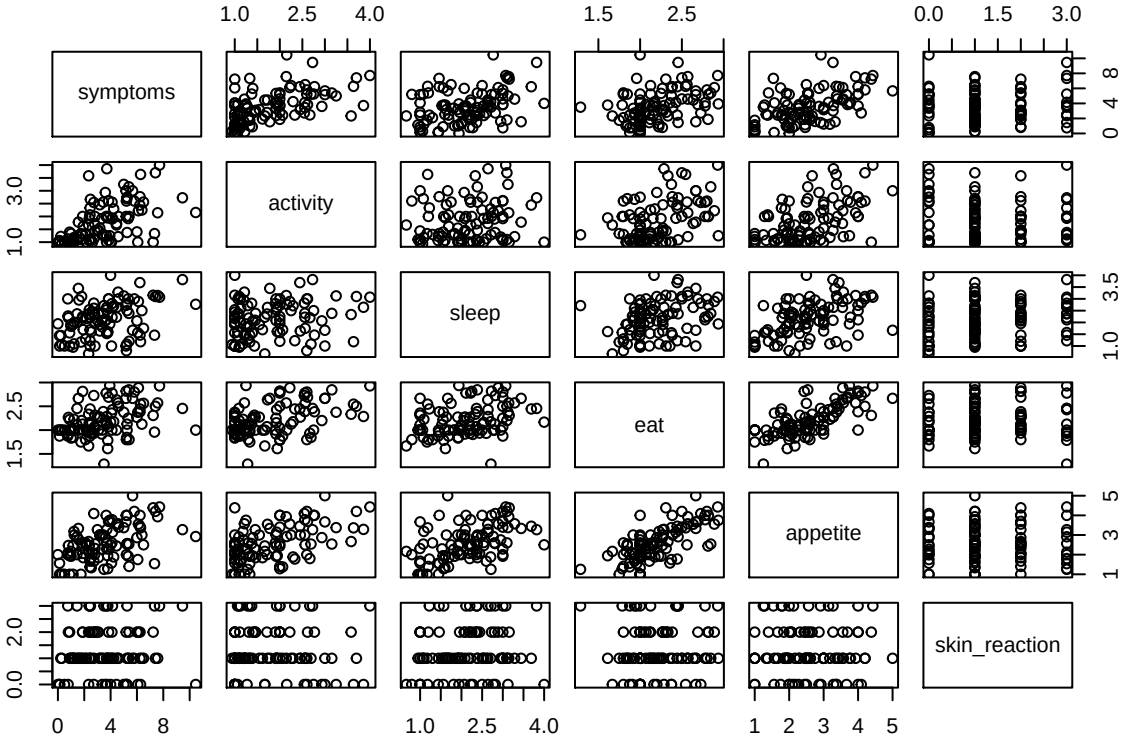
```
## # A tibble: 6 x 6
##   symptoms activity sleep   eat appetite skin_reaction
##      <dbl>    <dbl> <dbl> <dbl>    <dbl>         <dbl>
## 1    0.889    1.39  1.56   2.22     1.94             1
## 2    2.81     1.44  0.999  2.31     2.31             2
## 3    1.45     1.09  2.36   2.46     2.91             3
## 4    0.294    0.941 1.06   2        1                1
## 5    2.73     2.54  2.82   2.73     4.09             0
## 6    3.94     1.25  1.94   2.94     3.75             1
```

Summary statistics:

```
##     symptoms         activity         sleep            eat
```

```
##  Min.   : 0.000   Min.   :0.941   Min.   :0.666   Min.   :1.286
##  1st Qu.: 1.887   1st Qu.:1.111   1st Qu.:1.564   1st Qu.:2.000
##  Median : 3.404   Median :1.641   Median :2.178   Median :2.139
##  Mean   : 3.542   Mean   :1.809   Mean   :2.138   Mean   :2.209
##  3rd Qu.: 5.178   3rd Qu.:2.323   3rd Qu.:2.712   3rd Qu.:2.440
##  Max.   :10.461   Max.   :4.000   Max.   :4.000   Max.   :2.937
##     appetite     skin_reaction
##  Min.   :1.000   Min.   :0.000
##  1st Qu.:1.924   1st Qu.:1.000
##  Median :2.500   Median :1.000
##  Mean   :2.575   Mean   :1.276
##  3rd Qu.:3.272   3rd Qu.:2.000
##  Max.   :5.000   Max.   :3.000
```

Upon a first glance at our data distribution below we can tell that variables eat and appetite are correlated. Thus in our analysis we should start by selecting only one of the two. In addition we can use intition to see what variables can provide interesting insights. For example, it does not make much sense to try to use a type of skin reaction to predict symptoms since that is a symptom itself.



Analysis

Following our intuition from the data summary we start by fitting a linear regression model with symptoms as the response variable and activity, sleep and eat as the explanatory variables.

```
##
## Call:
## lm(formula = Y ~ Z)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6899 -1.1605 -0.0809  0.7767  6.1618
```

```
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.5551     1.1916  -1.305  0.19506
## Z                NA         NA      NA       NA
## Zactivity     1.2807     0.2518   5.087 1.86e-06 ***
## Zsleep        0.6940     0.2474   2.805  0.00612 **
## Zeat          0.5871     0.6211   0.945  0.34694
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.727 on 94 degrees of freedom
## Multiple R-squared:  0.3791, Adjusted R-squared:  0.3593
## F-statistic: 19.13 on 3 and 94 DF,  p-value: 9.144e-10
```

From the estimates and its significance at level alpha = 0.05 we can see that eat does not provide enough information to be significant. Perhaps we can try to see what removing this variable does to the model.

```
##
## Call:
## lm(formula = Y ~ Z)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8097 -1.1995 -0.1087  0.8463  5.9603
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5899     0.6140  -0.961  0.33907
## Z                NA         NA      NA       NA
## Zactivity     1.3815     0.2279   6.062 2.71e-08 ***
## Zsleep        0.7638     0.2360   3.236  0.00167 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.726 on 95 degrees of freedom
## Multiple R-squared:  0.3732, Adjusted R-squared:   0.36
## F-statistic: 28.28 on 2 and 95 DF,  p-value: 2.314e-10
```

As we can see the statistics remain stable while activity and sleep remain equally significant as before. Removing eat did not affect our linear model significantly. Therefore our chosen model for analysis is Symptoms ~ activity + sleep at level alpha = 0.05 .

The least squared estimates corresponding to this model are:

```
##            symptoms
##          -0.5899431
## activity  1.3815076
## sleep     0.7637762
```

with an R-squared statistic of 0.3731798.

The estimated covariance matrix of the estimates in $\hat{\beta}$ is:

```
cov_est
```

```
##                          activity       sleep
##           0.37697753 -0.07242676 -0.10082910
```

```
## activity -0.07242676  0.05194201 -0.01008368
## sleep    -0.10082910 -0.01008368  0.05570451
```

Testing H-null: beta_j = 0

From t-test for $\hat{\beta}_1$ we obtained a t-statistic of 6.0616922 and a 1.985251
Therefore we can reject the null hypothesis that the estimate $\hat{\beta}_1$ is zero at level alpha = 0.05 .

From t-test for $\hat{\beta}_2$ we obtained a t-statistic of 3.2360924 and a 1.985251 Therefore we can reject the null hypothesis that the estimate $\hat{\beta}_2$ is zero at level alpha = 0.05 .

We proceed to find confedence intervals for these estimators.

The following 95% confidence interval was found for $\hat{\beta}_1$

```
## [ 0.9290532 , 1.833962 ]
```

The following 95% confidence interval was found for $\hat{\beta}_2$

```
## [ 0.2952211 , 1.232331 ]
```

The following 95% confidence region based simultaneous confidence intervals were found for $\hat{\beta}_1$ and $\hat{\beta}_2$ respectivly:

```
## [ 0.7328212 , 2.030194 ]
```

```
## [ 0.09200618 , 1.435546 ]
```

The following 95% confidence intervals for $\hat{\beta}_1$ and $\hat{\beta}_1$ were obtained with bonferroni correction

```
## [ 0.8260708 , 1.936944 ]
```

```
## [ 0.1885742 , 1.338978 ]
```

Finally we use an F-test to test for H_0: beta_1 = beta_2 = 0 at level alpha = 0.05

with a level alpha = 0.05 we find that the f-statistic is 168.4947024 while the critical value is 18.4241563.
Therefore we can reject the null hypothesis that the estimate $\hat{\beta}_1 = \hat{\beta}_2 = 0$ at level alpha = 0.05 .

Conclusion

The result above agree that the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ are not zero and therefore the variables in the model proposed have a significant weight in the determination of the response variable “symptoms”.
We can use this model to predict the value of symptoms for a new patient with significant certainty.

Let’s say we have a new patient with a score of 2.5 in activity and 3.0 in sleep. What can they expect?
The following is a 95% confidence interval for a new obsevation

```
## [ 1.680279 , 8.630029 ]
```

---

2. LDA

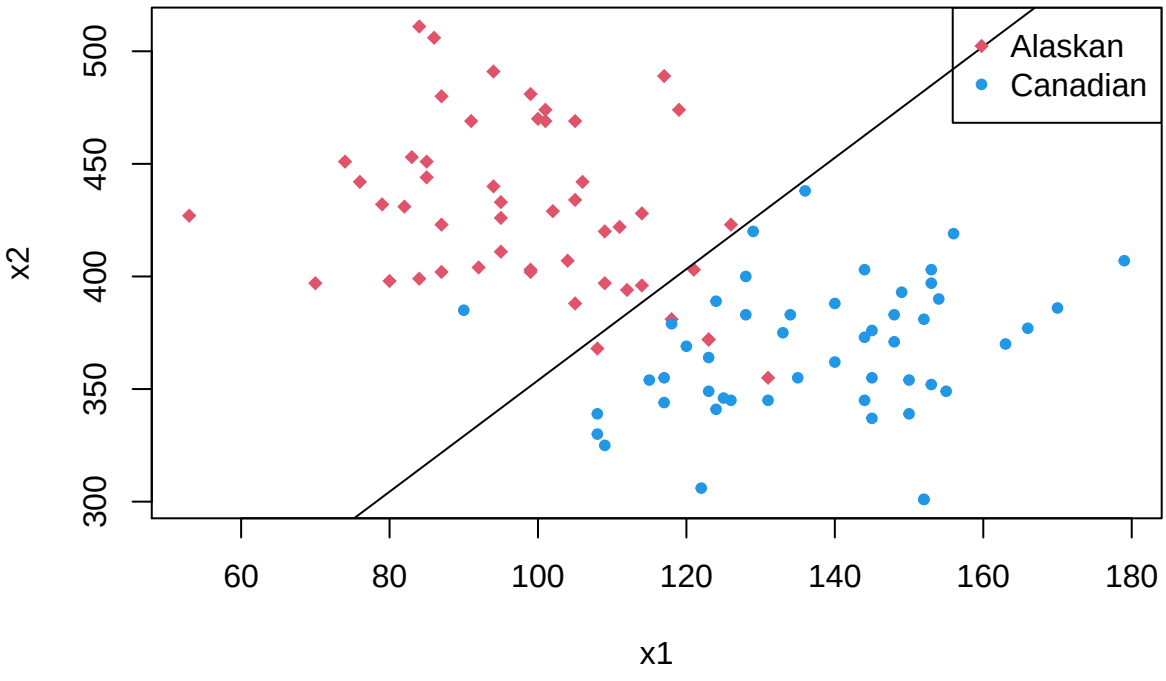
Introduction

T11_02 LDA These data were edited from file T11-2.DAT on disk from book Salmon data (growth-ring diameters) Col. 1: location (1 = Alaskan, 2 = Canadian) Col. 2: gender (1 = female, 2 = male) Col. 3: X1 = diameter of rings for 1st yr freshwater growth (.01 in) Col. 4: X2 = diameter of rings for 1st yr marine growth (.01 in)

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   location = col_double(),
##   gender = col_double(),
##   x1 = col_double(),
##   x2 = col_double()
## )
```

```
## Call:
## lda(location ~ x1 + x2, data = salmon1, prior = c(1, 1)/2)
##
## Prior probabilities of groups:
##   1   2
## 0.5 0.5
##
## Group means:
##       x1     x2
## 1  98.38 429.66
## 2 137.46 366.62
##
## Coefficients of linear discriminants:
##            LD1
## x1  0.04458572
## x2 -0.01803856
```

```
##
## true_class  1  2
##          1 44  6
##          2  1 49
```



3. PCA

Introduction

The goal of this analysis is to find the minimum amount of data we can use to obtain the most information. We will be maximizing the information obtained by using the components that provide more than 90% of the variance.

This analysis is to be carried away on data which contains the carapace measurements in milimeters for painted turtles. The asspects that are meassured are width, length and height. In addition to this it also contains a column with gender information.

T6_9 Carapace measurements in milimeters for painted turtles Col. 1: x1 = length Col. 2: x2 = width Col. 3: x3 = height Col. 4: Gender (1 = female,2 = male)

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   length = col_double(),
##   width = col_double(),
##   height = col_double(),
##   gender = col_character()
## )
```

Summary

The first six rows of our data look like this:

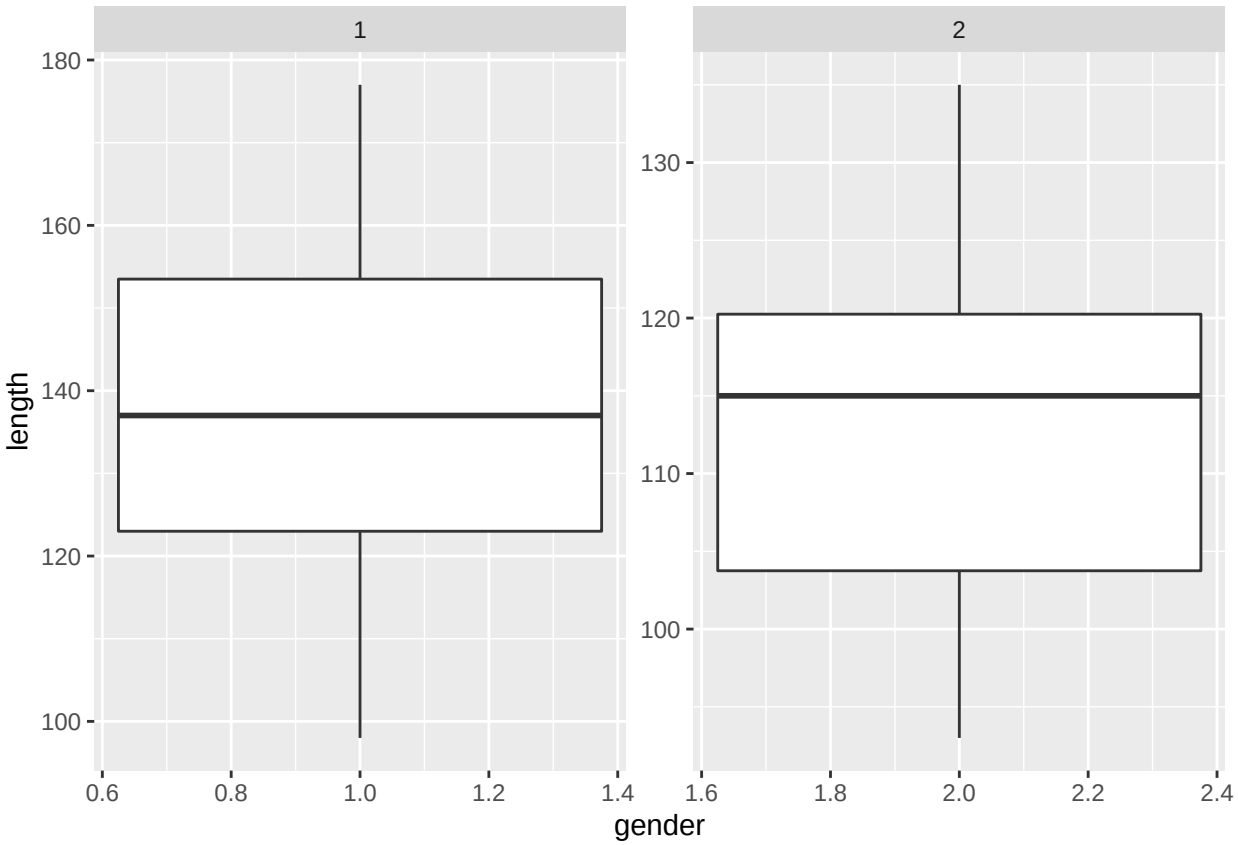
```
## # A tibble: 6 x 4
##   length width height gender
##    <dbl> <dbl>  <dbl>  <int>
## 1     98    81     38      1
## 2    103    84     38      1
## 3    103    86     42      1
## 4    105    86     42      1
## 5    109    88     44      1
## 6    123    92     50      1
```

Summary statistics:

```
##      length          width            height          gender
##  Min.   : 93.0   Min.   : 74.00   Min.   :35.00   Min.   :1.0
##  1st Qu.:106.8   1st Qu.: 86.00   1st Qu.:40.00   1st Qu.:1.0
##  Median :122.0   Median : 93.00   Median :44.50   Median :1.5
##  Mean   :124.7   Mean   : 95.44   Mean   :46.38   Mean   :1.5
##  3rd Qu.:136.5   3rd Qu.:102.00   3rd Qu.:51.00   3rd Qu.:2.0
##  Max.   :177.0   Max.   :132.00   Max.   :67.00   Max.   :2.0
```

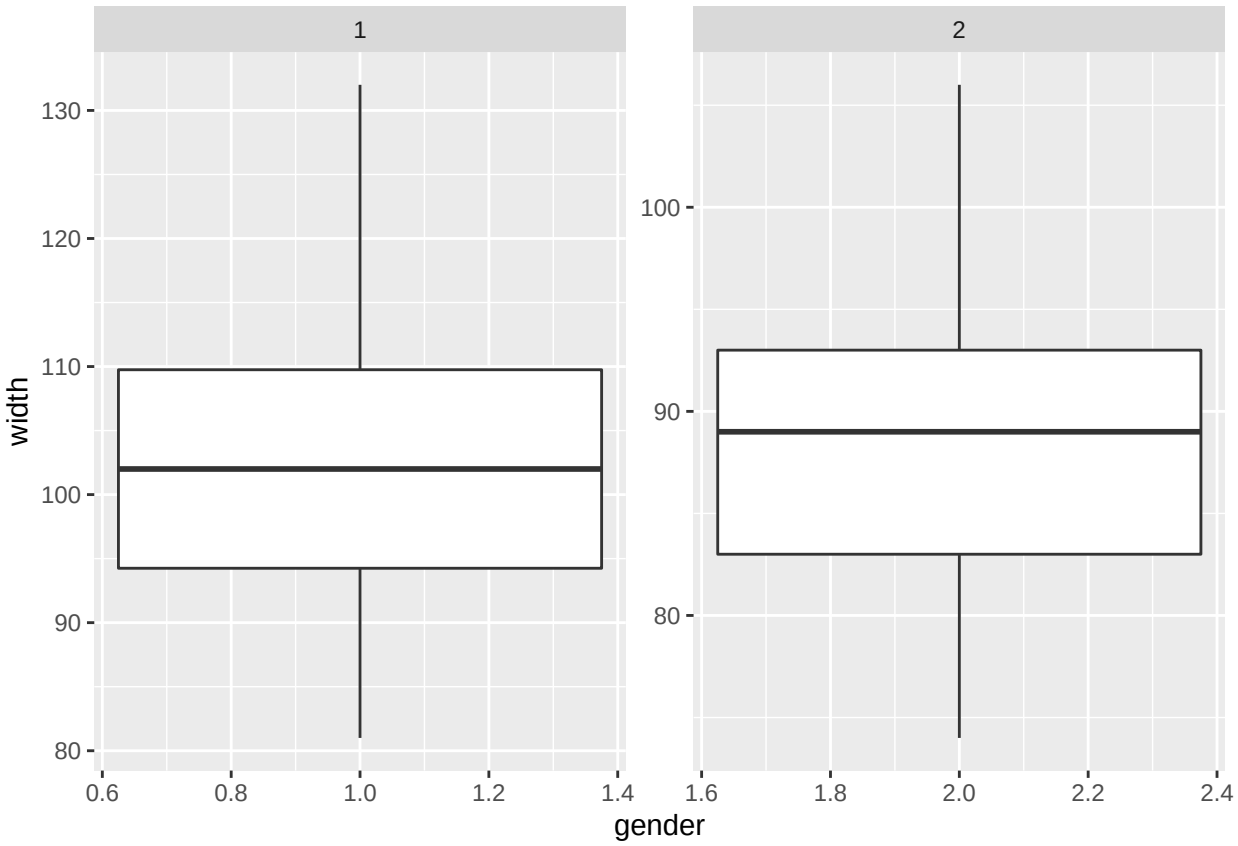
The following is a boxplot of length by gender

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



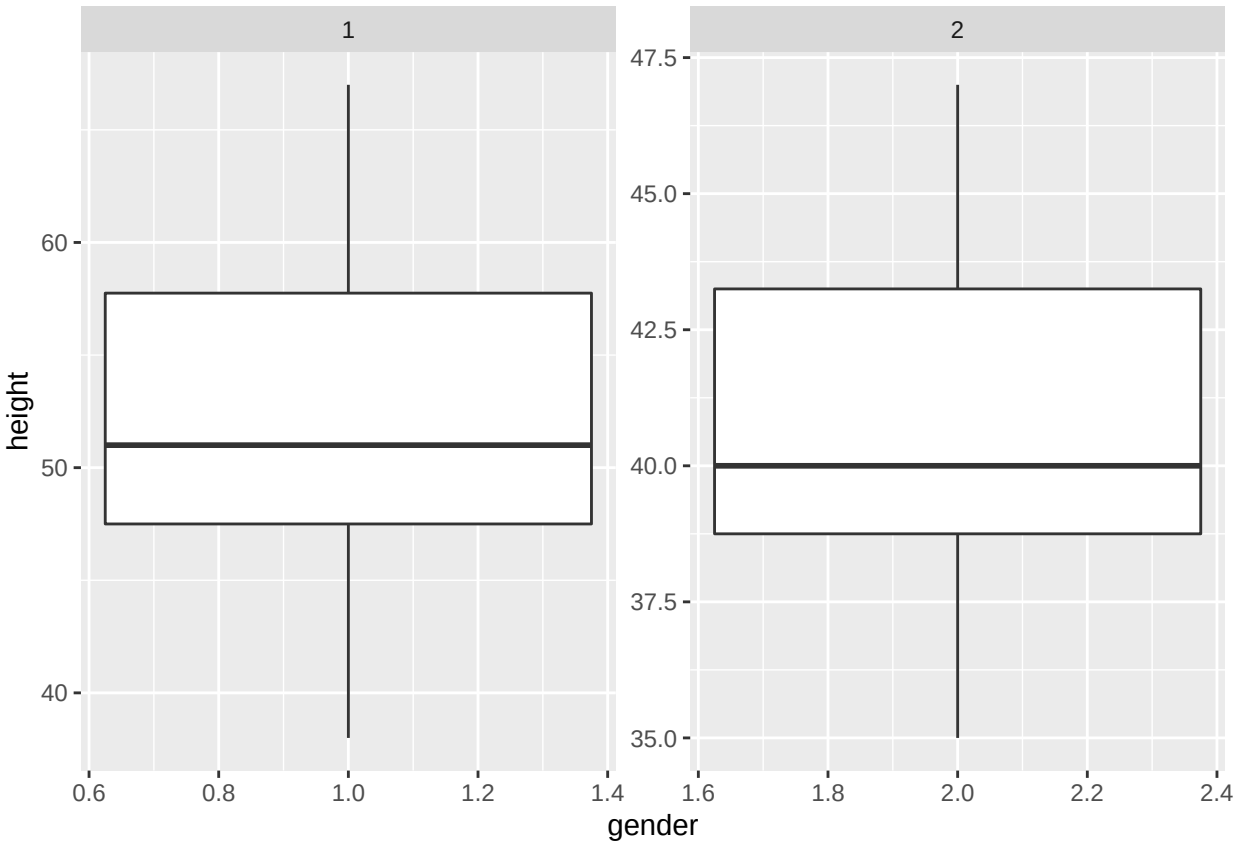
The following is a boxplot of width by gender

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



The following is a boxplot of height by gender

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



From these plots and from knowledge we can tell that there might be a relationship between size of the turtle and gender.

Analysis

The principal components of this data are as follow

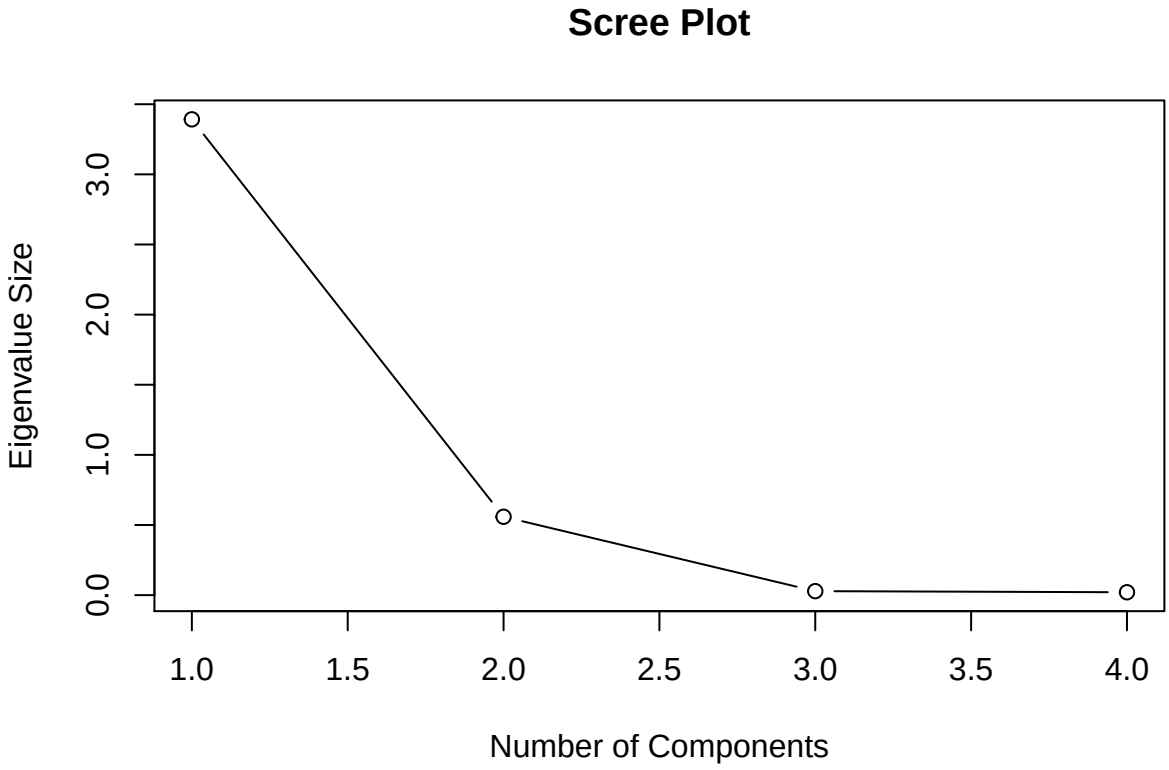
```
## Importance of components:
##                           Comp.1    Comp.2      Comp.3      Comp.4
## Standard deviation     1.8417953 0.7475347 0.168581315 0.143395486
## Proportion of Variance 0.8480525 0.1397020 0.007104915 0.005140566
## Cumulative Proportion  0.8480525 0.9877545 0.994859434 1.000000000
##
## Loadings:
##        Comp.1 Comp.2 Comp.3 Comp.4
## length  0.524  0.311         0.792
## width   0.525  0.292  0.636 -0.484
## height  0.537        -0.758 -0.362
## gender -0.401  0.901 -0.145
```

We can se that the first two components provide more than 90% of the variance.

The eigenvalues of the correlation matrix are:

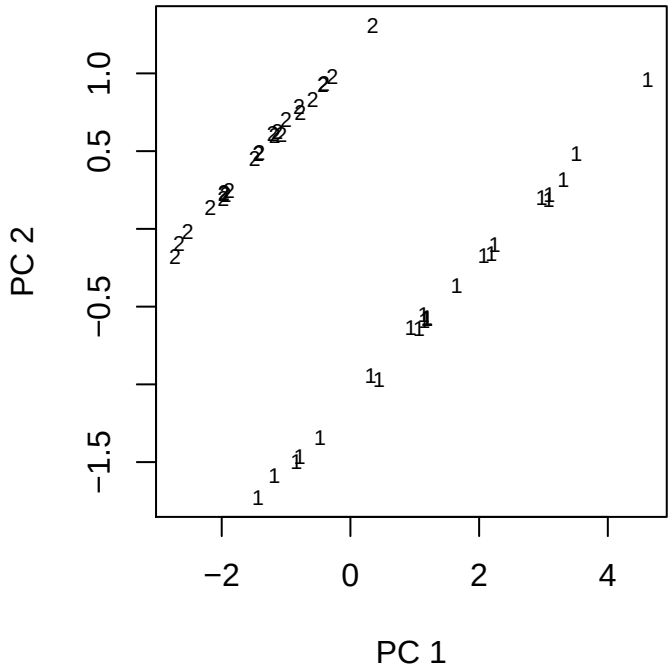
```
##     Comp.1     Comp.2     Comp.3     Comp.4
## 3.39220996 0.55880811 0.02841966 0.02056227
```

From the following plot we can see the component contribution to the distribution.

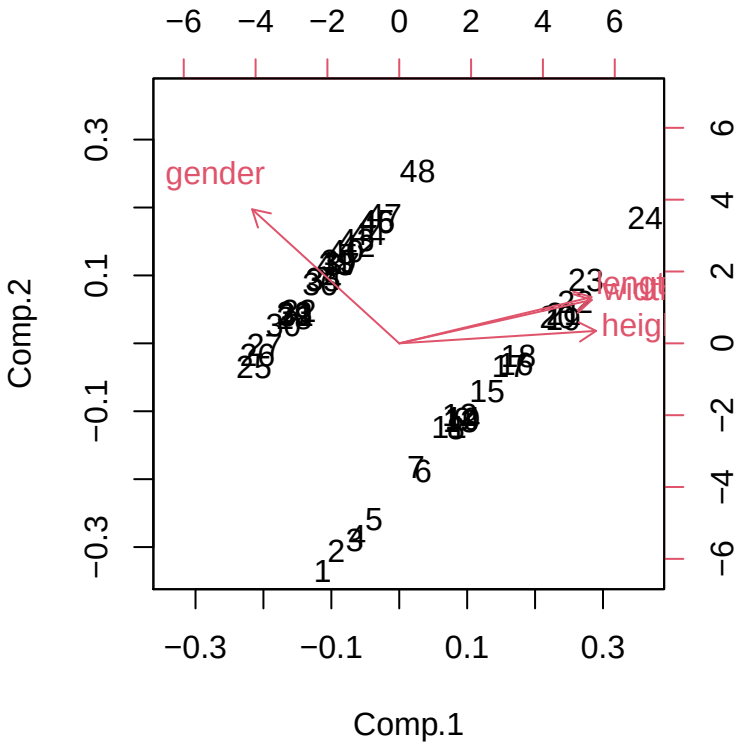


Notice the elbow occurs at 2 components. Thus lets continue our analysis with this in mind.

Plotting the PC scores for the sample data in the space of the first two principal components:



Notice there is a clear distinction between the genders. Now we use a biplot to see more detail.



From this point we can observe that the three variables containing measurements point to the same direction. Two of them in particular seem to be highly correlated. That is widht and lenght.

Conclusion

Since the first two components provide more than 90% of the variance, they are enough to provide the necesary information about this data. In addition we can see that there is correlation between the variables but ultimatly a distinction between female and male turtles.

Code Appendix

```
library(readr)
radio <- read_table2("~/Documents/winter 2021/sta 135/T1-7.DAT", col_names = c("symptoms", "activity","s

head(radio)
summary(radio)
plot(radio)
Y <- as.matrix(radio[,1])
n <- length(Y)
Z <- cbind(rep(1,n),as.matrix(radio[,2:4]))
r <- dim(Z)[2]-1
model = lm(Y~Z)
summary_first = summary(model)

summary_first
Y <- as.matrix(radio[,1])
n <- length(Y)
Z <- cbind(rep(1,n),as.matrix(radio[,2:3]))
r <- dim(Z)[2]-1
model = lm(Y~Z)
```

```r
summary = summary(model)

summary
# least square estimates
beta_hat <- solve(t(Z)%*%Z)%*%t(Z)%*%Y
beta_hat
# R^2 statistic
R_square <- 1 - sum((Y - Z%*%beta_hat)^2)/sum((Y-mean(Y))^2)

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

# estimated covariance of hat{beta}
cov_est = sigma_hat_square * solve(t(Z)%*%Z)
cov_est
Omega <- solve(t(Z)%*%Z)
# 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])

alpha <- 0.05
cval_t <- qt(1-alpha/2, n-r-1)
# t-test for single coefficient
# H_0: beta_j = 0, H_a: beta_j != 0

j <- 2
t_stat <- (beta_hat[j+1] - 0)/sqrt(sigma_hat_square * solve(t(Z)%*%Z)[j+1,j+1])

alpha <- 0.05
cval_t <- qt(1-alpha/2, n-r-1)
# One-at-a-time confidence interval for beta_j
j <- 1
cat('[',
    beta_hat[j+1] - qt(1-alpha/2, n-r-1)*sqrt(sigma_hat_square * Omega[j+1,j+1]),
    ',',
    beta_hat[j+1] + qt(1-alpha/2, n-r-1)*sqrt(sigma_hat_square * Omega[j+1,j+1]),
    ']')
# One-at-a-time confidence interval for beta_j
j <- 2
cat('[',
    beta_hat[j+1] - qt(1-alpha/2, n-r-1)*sqrt(sigma_hat_square * Omega[j+1,j+1]),
    ',',
    beta_hat[j+1] + qt(1-alpha/2, n-r-1)*sqrt(sigma_hat_square * Omega[j+1,j+1]),
    ']')
# Confidence region based simultaneous confidence intervals
j <- 1
cat('[',
    beta_hat[j+1] - sqrt((r+1)*qf(1-alpha,r+1,n-r-1))*sqrt(sigma_hat_square * Omega[j+1,j+1]),
    ',',
    beta_hat[j+1] + sqrt((r+1)*qf(1-alpha,r+1,n-r-1))*sqrt(sigma_hat_square * Omega[j+1,j+1]),
    ']')
```

```r
# Confidence region based simultaneous confidence intervals
j <- 2
cat('[',
    beta_hat[j+1] - sqrt((r+1)*qf(1-alpha,r+1,n-r-1))*sqrt(sigma_hat_square * Omega[j+1,j+1]),
    ',',
    beta_hat[j+1] + sqrt((r+1)*qf(1-alpha,r+1,n-r-1))*sqrt(sigma_hat_square * Omega[j+1,j+1]),
    ']')
# Bonferroni correction based confidence intervals
j <- 1
cat('[',
    beta_hat[j+1] - qt(1-alpha/(2*(r+1)), n-r-1)*sqrt(sigma_hat_square * Omega[j+1,j+1]),
    ',',
    beta_hat[j+1] + qt(1-alpha/(2*(r+1)), n-r-1)*sqrt(sigma_hat_square * Omega[j+1,j+1]),
    ']')
# Bonferroni correction based confidence intervals
j <- 2
cat('[',
    beta_hat[j+1] - qt(1-alpha/(2*(r+1)), n-r-1)*sqrt(sigma_hat_square * Omega[j+1,j+1]),
    ',',
    beta_hat[j+1] + qt(1-alpha/(2*(r+1)), n-r-1)*sqrt(sigma_hat_square * Omega[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 C
q = 0

Omega_22 = C%*% solve(t(Z)%*%Z) %*%t(C)
f_stat <- t(C%*%beta_hat)%*%solve(Omega_22)%*%(C%*%beta_hat)

cval_f <- qf(1-alpha, 2, n-r-1)
critical = cval_f * df_1 * sigma_hat_square
# prediction interval for Y_0 = z_0^T beta + epsilon_0
z_0 <- c(1, 2.5, 3.0)

cat('[',
    z_0%*%beta_hat - sqrt(sigma_hat_square)*qt(1-alpha/2, n-r-1)*sqrt(1+t(z_0)%*%solve(t(Z)%*%Z)%*%z_0)
    ',',
    z_0%*%beta_hat + sqrt(sigma_hat_square)*qt(1-alpha/2, n-r-1)*sqrt(1+t(z_0)%*%solve(t(Z)%*%Z)%*%z_0)
    ']')
salmon1 <- read_table2("~/Documents/winter 2021/sta 135/T11-2.DAT", col_names = c("location", "gender",

library(MASS)
lda.obj<-lda(location~x1+x2,data=salmon1, prior=c(1,1)/2)
lda.obj
plda<-predict(object=lda.obj, newdata=salmon1)
#determine how well the model fits
true_class <- as.matrix(data.frame(lapply(salmon1[,1], as.character)))
table(true_class, as.matrix(plda$class))

#plot the decision line
gmean <- lda.obj$prior %*% lda.obj$means
```

```r
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(salmon1[,c(3,4)],pch=rep(c(18,20),each=50),col=rep(c(2,4),each=50))
abline(intercept, slope)
legend("topright",legend=c("Alaskan","Canadian"),pch=c(18,20),col=c(2,4))

turtles <- read_table2("~/Documents/winter 2021/sta 135/T6-9.DAT", col_names = c("length", "width","hei

n = length(turtles$gender)
for (i in 1:n) {
  if(turtles$gender[i]=="female"){
    turtles$gender[i] = 1
  }else if(turtles$gender[i]=="male"){
    turtles$gender[i] = 2
  }
}
turtles$gender = as.integer(turtles$gender)
head(turtles)
summary(turtles)
library(ggplot2)
 ggplot(turtles, aes(x=gender, y=length)) +
    geom_boxplot() +
    facet_wrap(~gender, scale="free")
 ggplot(turtles, aes(x=gender, y=width)) +
    geom_boxplot() +
    facet_wrap(~gender, scale="free")
 ggplot(turtles, aes(x=gender, y=height)) +
    geom_boxplot() +
    facet_wrap(~gender, scale="free")
turtles_df <- data.frame(turtles)
turtles_pc <- princomp(turtles_df, cor=T)

# Showing the coefficients of the components:
summary(turtles_pc,loadings=T)

# Showing the eigenvalues of the correlation matrix:
eigenval <- (turtles_pc$sdev)^2
eigenval
# A scree plot:
plot(1:(length(turtles_pc$sdev)), (turtles_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(turtles_pc$scores[,1], turtles_pc$scores[,2],
     xlab="PC 1", ylab="PC 2", type ='n', lwd=2)
# labeling points with state abbreviations:
text(turtles_pc$scores[,1], turtles_pc$scores[,2], labels = turtles$gender, cex=0.7, lwd=2)

# We see the Southeastern states grouped in the bottom left
```

```r
# and several New England states together in the bottom right.
# The biplot can add information about the variables to the plot of the first two PC scores:

biplot(turtles_pc)
```