Problem Set 2

Dr. Young

Carson Slater

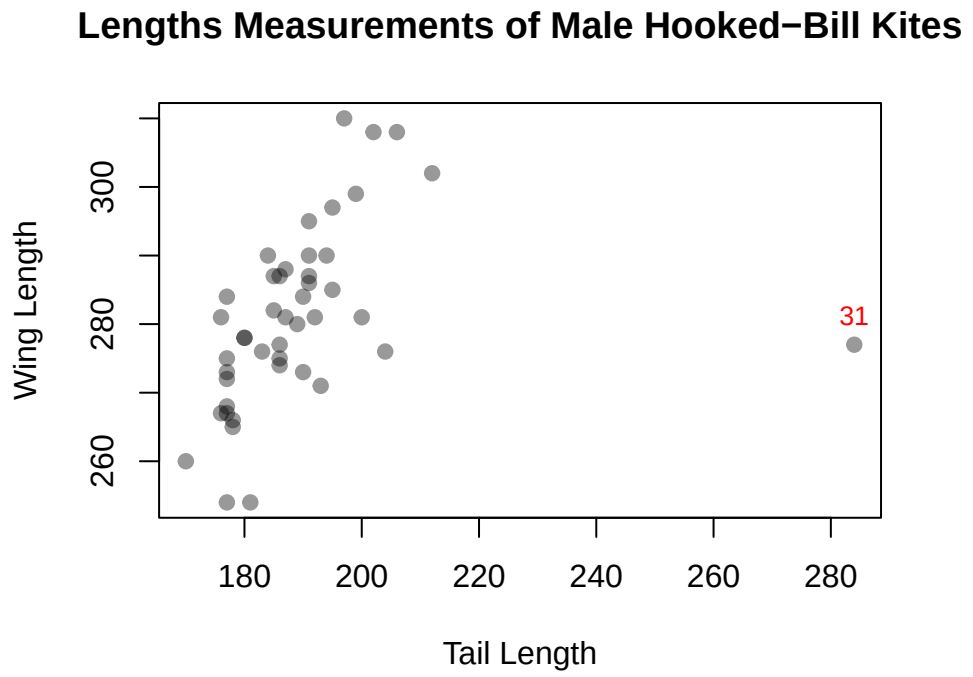
2024-06-28

6.20

```
male <- read.delim(here::here("datasets", "T6-11.dat"), sep = '', header = FALSE)
```

a.

```
colnames(male) <- c("tail", "wing")
plot(male, main = "Lengths Measurements of Male Hooked-Bill Kites",
     xlab = "Tail Length", ylab = "Wing Length",
     pch = 19, col = alpha("black", 0.4))
text(male$tail[31], male$wing[31], labels = c(31), pos = 3, cex = 0.8, col = "red")
```



Aside from observation 31, there appears to be no outliers in this data.

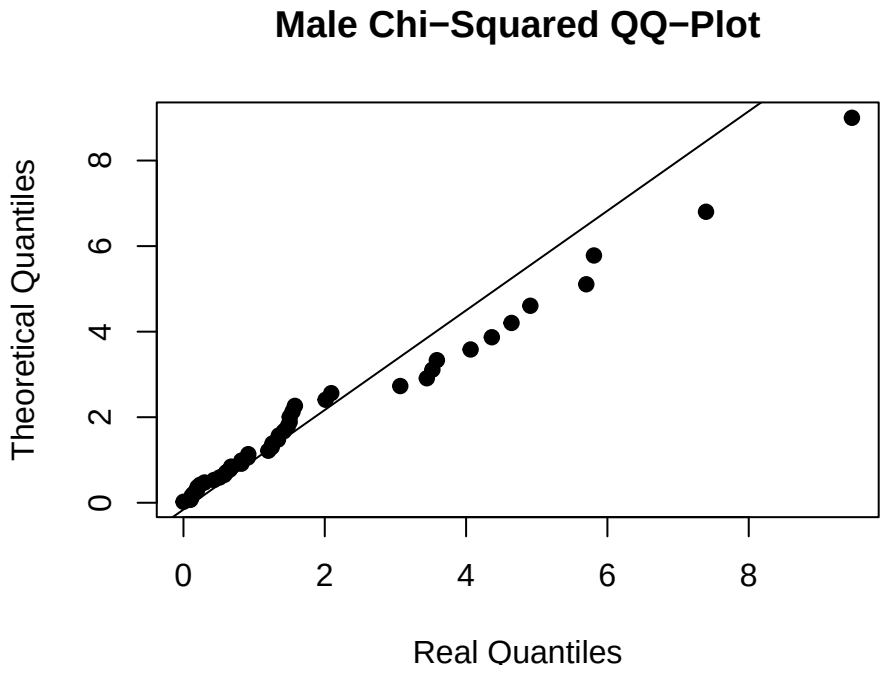
b.

We elect to change the tail measurement of observation 31 from 284 to 184, as it is believed to be a data entry error.

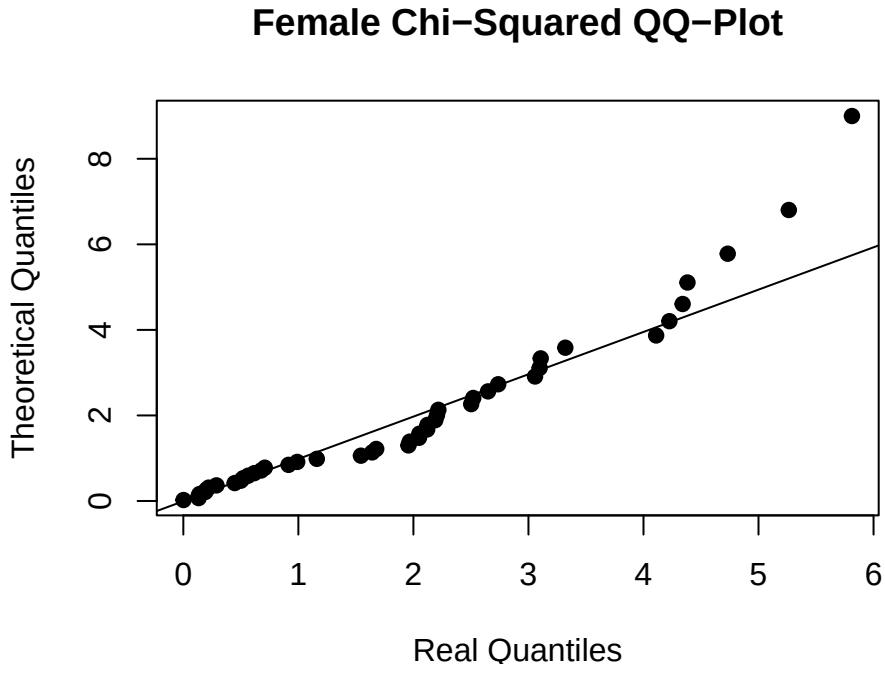
```r
female <- read.delim(here::here("datasets", "T5-12.dat"), sep = '', header = FALSE)

colnames(female) <- c("tail", "wing")

male[31, 1] <- 184

# number of variables and alpha
p <- 2
alpha <- 0.05

# combine two datasets
n_m <- dim(male)[1]
n_f <- dim(female)[1]
bird <- rbind(male, female)
bird$gender <- c(rep('male', n_m),rep('female', n_f))
```

We first check for multivariate normality between the two sets of data.

```r
# test for equality of mean vector
xbar_m <- colMeans(bird[bird$gender == 'male', -3])
names(xbar_m) <- c("tail", "wing")
xbar_f <- colMeans(bird[bird$gender == 'female', -3])
names(xbar_f) <- c("tail", "wing")

Sigma_m <- cov(bird[bird$gender == 'male', -3])
Sigma_f <- cov(bird[bird$gender == 'female', -3])
Sigma_pool <- ((n_m-1)*Sigma_m+(n_f-1)*Sigma_f)/(n_m+n_f-2)
Sigma_poolinv <- solve((1/n_m+1/n_f)*Sigma_pool)


male_dist <- mahalanobis(male, xbar_m, Sigma_m)
female_dist <- mahalanobis(female, xbar_f, Sigma_f)

qqplot(male_dist, qchisq(ppoints(nrow(male)), p),
       main = "Male Chi-Squared QQ-Plot",
       ylab = "Theoretical Quantiles",
       xlab = "Real Quantiles",
       pch = 19)
qqline(male_dist, distribution = \(prob) qchisq(prob, df = p))
```




```
qqplot(female_dist, qchisq(ppoints(nrow(female)), p),
       main = "Female Chi-Squared QQ-Plot",
       ylab = "Theoretical Quantiles",
       xlab = "Real Quantiles",
       pch = 19)
qqline(female_dist, distribution = \(prob) qchisq(prob, df = p))
```



According to the QQ-plots, the male data structure seems to show small deviations from multivariate normality, but the female data structre seems to be mostly consistent with multivariate normality.

Next, we test the equality of the covariance structures.

```
(box <- boxM(bird[, -3], bird[, 3]))
```

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  bird[, -3]
## Chi-Sq (approx.) = 1.2223, df = 3, p-value = 0.7477
```

It appears the Box's M test for homogeneity of covariance structures yields a p-value of 0.7477, which lends to the idea of failing to reject the null hypothesis that the covariance structures are equal between the two groups. We elect to pool the covariance structures in our test.

We denote the pooled variance as,

$$\mathbf{S}_{\text{Pooled}} = \frac{(n_m - 1)\mathbf{S}_m (n_f - 1)\mathbf{S}_f}{n_m + n_f - 2},$$

and calculate the test statistic as

$$T^2 = (\bar{\mathbf{x}}_m - \bar{\mathbf{x}}_f)' \left[\left(\frac{1}{n_m} + \frac{1}{n_f} \right) \mathbf{S}_{\text{Pooled}} \right]^{-1} (\bar{\mathbf{x}}_m - \bar{\mathbf{x}}_f)$$

```
# test statistic
(T2 <- t(xbar_m-xbar_f)%*%Sigma_poolinv%*%(xbar_m-xbar_f))
```

```
##          [,1]
## [1,] 25.66253
```

```
# critical value
(c2 <- (n_m+n_f-2)*p/(n_m+n_f-p-1)*qf(1-alpha,df1=p,df2=n_m+n_f-p-1))
```

```
## [1] 6.273886
```

After conducting a test for the equality of mean vectors, we have that the test statistic exceeds the critical value. Hence, we conclude that the male and female hook-billed kite population mean vectors are not equal.

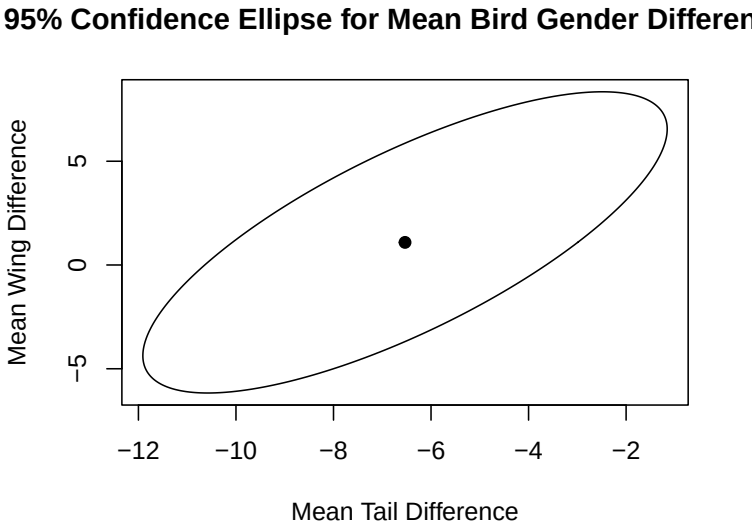
c

We can generate a 95% confidence ellipse with the following equation,

$$\left\{ ((\boldsymbol{\mu}_m - \boldsymbol{\mu}_f) - (\bar{\mathbf{x}}_m - \bar{\mathbf{x}}_f))' \left[\left(\frac{1}{n_m} + \frac{1}{n_f} \right) \mathbf{S}_{\text{Pooled}} \right]^{-1} ((\boldsymbol{\mu}_m - \boldsymbol{\mu}_f) - (\bar{\mathbf{x}}_m - \bar{\mathbf{x}}_f)) \right\} \leq \frac{p(n_m + n_f - 2)}{(n_m + n_f - p - 1)} F_{p, n_m + n_f - p - 1}(\alpha).$$

```
# simultaneous confidence intervals for the components of mu1-mu2
center <- xbar_m - xbar_f
npoints <- 1e3
theta <- seq(0, 2*pi, length = npoints)
qlevel <- qf(1-alpha,df1=p,df2=n_m+n_f-p-1)

r <- sqrt((1/n_m+1/n_f)*(n_m+n_f-2)*p/(n_m+n_f-p-1)*qlevel)
v <- rbind(r*cos(theta), r*sin(theta))
z <- backsolve(chol(solve(Sigma_pool)),v)+center

# calculate the 95% simultaneous confidence interval
lower_tail <- center[1]-r*sqrt(Sigma_pool[1,1])
upper_tail <- center[1]+r*sqrt(Sigma_pool[1,1])
```

```r
lower_wing <- center[2]-r*sqrt(Sigma_pool[2,2])
upper_wing <- center[2]+r*sqrt(Sigma_pool[2,2])

# plot the ellipse for Scheff's
plot(t(z),type='l',
     main = "95% Confidence Ellipse for Mean Bird Gender Differences",
     xlab = "Mean Tail Difference",
     ylab = "Mean Wing Difference"
     )
points(center[1], center[2],pch=19)
```




```r
table <- rbind(c(lower_tail, upper_tail), c(lower_wing, upper_wing))
rownames(table) <- c("Tail", "Wings")
table |> knitr::kable(col.names = c("Lower", "Upper"))
```

	Lower	Upper
Tail	-11.909075	-1.157591
Wings	-6.168662	8.346440

6.21

We do not have the data so we cannot check for multivariate normality, as only the sample mean vectors and covariance matrices are given.

We now consider the use of pooled variance. The book remarks that if $(n_1 = n_2 = n)$ then $((n-1)/(n+n-2) = 1/2)$ so

$$\begin{aligned}\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2 &= \frac{1}{n}(S_1 + S_2) \\ &= \frac{(n-1)S_1 + (n-1)S_2}{n+n-2}\left(\frac{1}{n} + \frac{1}{n}\right) \\ &= S_{\text{Pooled}}\left(\frac{1}{n} + \frac{1}{n}\right).\end{aligned}$$

With equal sample sizes, the large sample procedure is essentially the same as the procedure based on the pooled covariance matrix. In one dimen- sion, it is well known that the effect of unequal variances is least when $(n_1 = n_2)$ and greatest when (n_1) is much less than (n_2) or vice versa. This is similar in the multivariate case. Hence, we concede that the use of the pooled variance is not unreasonable.

```r
Sigma1 <- matrix(c(0.459,  0.254,  -0.26,  -0.244,
                   0.254,  27.465, -0.589, -0.267,
                  -0.26,  -0.589,   0.030,  0.102,
                  -0.244, -0.267,   0.102,  6.854),
                 nrow = 4, byrow = TRUE)
Sigma2 <- matrix(c(0.944,  -0.89,   0.002,  -0.719,
                  -0.89,   16.432, -0.589,  19.044,
                   0.002,  -0.589,  0.024,  -0.94,
                  -0.719,  19.044, -0.94,   61.854),
                 nrow = 4, byrow = TRUE)

Sigma_pool <- matrix(c(0.701,  0.83,  -0.012, -0.481,
                       0.83,  21.949, -0.494,  9.388,
                      -0.012, -0.494,  0.027,  0.04,
                      -0.481,  9.388,  0.04,  34.354),
                     nrow = 4, byrow = TRUE)

n1 <- 20
n2 <- 20
xbar1 <- c(2.287, 12.6, 0.347, 14.83)
xbar2 <- c(2.404, 7.155, 0.524, 12.840)

Sigma_poolinv <- solve((1/n1+1/n2)*Sigma_pool)

# test statistic
(T2 <- t(xbar1-xbar2)%*%Sigma_poolinv%*%(xbar1-xbar2))
```

```
##          [,1]
## [1,] 16.72817
```

```r
# critical value
(c2 <- (n1+n2-2)*p/(n1+n2-p-1)*qf(1-alpha,df1=p,df2=n1+n2-p-1))
```

```
## [1] 6.679627
```

```r
T2 > c2
```

```
##      [,1]
## [1,] TRUE
```

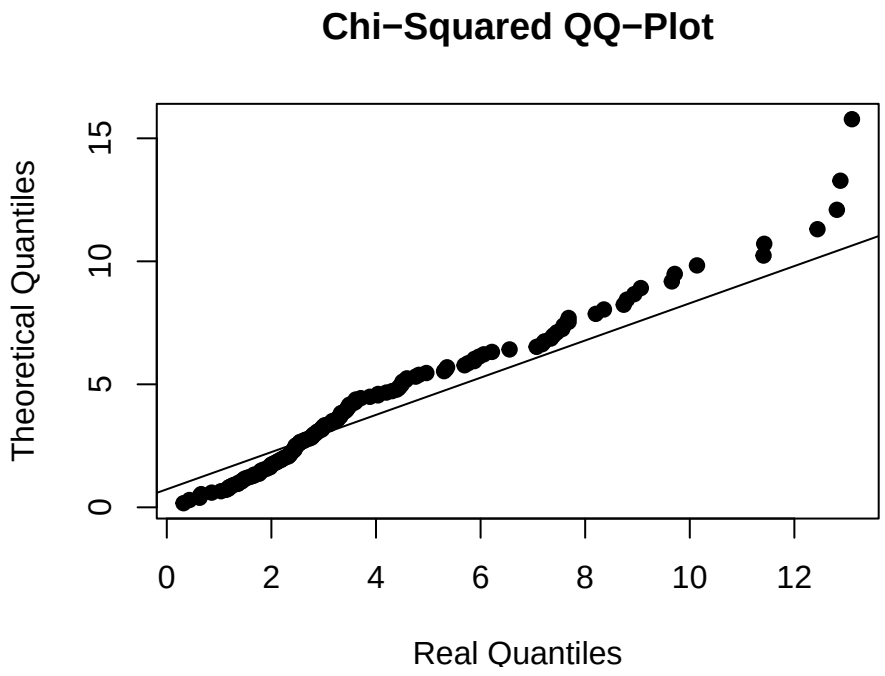
From using the pooled covariance matrix, we find that the test statistic from the sample is equal to 16.728, and the critical value is $T^2 = 6.679$. Therefore we reject

MANOVA Using the iris Data Set

We first check for multivariate normality.

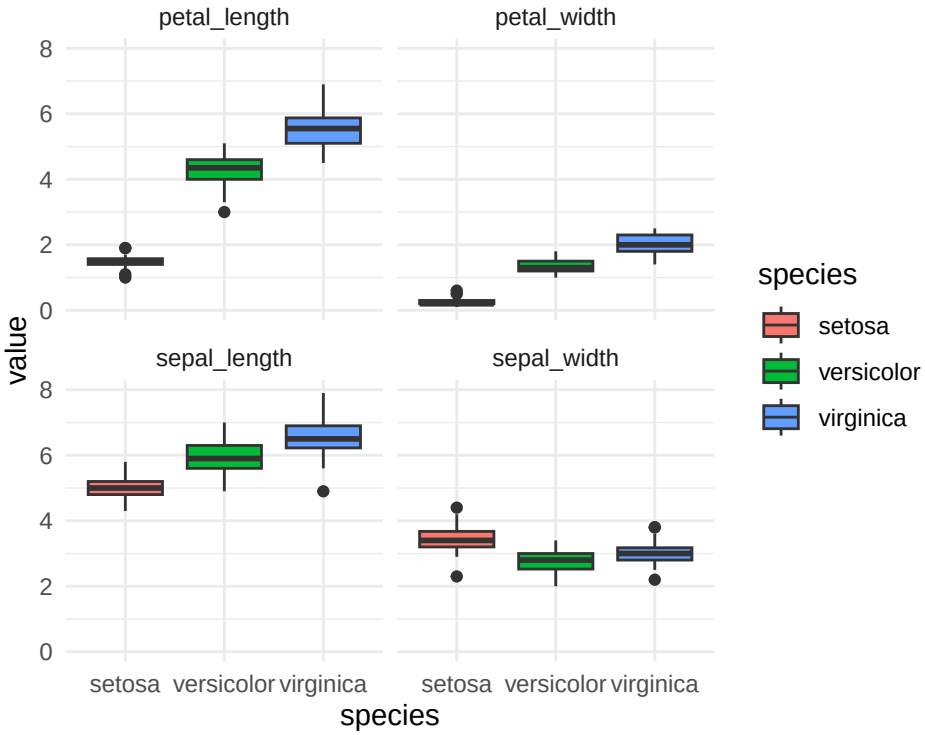
```r
Sigma_iris <- cov(iris[,-5])
xbar_iris <- colMeans(iris[,-5])

iris_dist <- mahalanobis(as.matrix(iris[,-5]), xbar_iris, Sigma_iris)

qqplot(iris_dist, qchisq(ppoints(nrow(iris)), ncol(iris[,-5])),
       main = "Chi-Squared QQ-Plot",
       ylab = "Theoretical Quantiles",
       xlab = "Real Quantiles",
```

```
       pch = 19)
qqline(iris_dist, distribution = \(prob) qchisq(prob, df = ncol(iris[,-5])))
```



These data do not appear to show significant deviations from multivariate normality. We can proceed under the assumption of multivariate normality.

```
# Exploratory Data Analysis
iris |>
  as_tibble() |>
  janitor::clean_names() |>
  pivot_longer(-c(species)) |>
  ggplot(aes(x = species, y = value)) +
  geom_boxplot(aes(fill = species)) +
  facet_wrap(. ~ name) +
  theme_minimal()
```




```
(box <- boxM(iris[, -5], iris[, 5]))
```

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  iris[, -5]
## Chi-Sq (approx.) = 140.94, df = 20, p-value < 2.2e-16
```

Performing Box's M test, we reject the null hypothesis that the covariance structures are equal, but since our assignment is to proceed with MANOVA for the `iris` data, we proceed with caution acknowledging this limitation on our findings.

```
iris_mod <- manova(cbind(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width) ~ Species, iris)

summary.aov(iris_mod, test = "Wilks")
```

```
##  Response Sepal.Length :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## Species       2 63.212  31.606  119.26 < 2.2e-16 ***
## Residuals   147 38.956   0.265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Sepal.Width :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## Species       2 11.345  5.6725   49.16 < 2.2e-16 ***
## Residuals   147 16.962  0.1154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Petal.Length :
##              Df Sum Sq Mean Sq F value    Pr(>F)
```

```
## Species       2 437.10 218.551  1180.2 < 2.2e-16 ***
## Residuals   147  27.22   0.185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Petal.Width :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## Species       2 80.413  40.207  960.01 < 2.2e-16 ***
## Residuals   147  6.157   0.042
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary.aov(iris_mod, test = "Roy")
```

```
##  Response Sepal.Length :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## Species       2 63.212  31.606  119.26 < 2.2e-16 ***
## Residuals   147 38.956   0.265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Sepal.Width :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## Species       2 11.345  5.6725   49.16 < 2.2e-16 ***
## Residuals   147 16.962  0.1154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Petal.Length :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## Species       2 437.10 218.551  1180.2 < 2.2e-16 ***
## Residuals   147  27.22   0.185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Petal.Width :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## Species       2 80.413  40.207  960.01 < 2.2e-16 ***
## Residuals   147  6.157   0.042
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In analyzing the MANOVA results, we observe that all tests yield a significant p-value. We have a violated assumption; however there are $n = 150$ observations, which is greater than $\frac{p(p+1)}{2} = 30$ observations. We this sufficiently large enough sample, we would have reasonable evidence to conclude that there does exists significant differences between mean vectors.