STAT 32950: Homework 3

Robert Winter

Table of Contents

1 Exercise 3: Practice Canonical Correlation Analysis, Low Dimensions

Download the data from `stiffness.DAT`.

```
stiffness =
 ↪  read.table("C:/Users/rewin/OneDrive/Documents/STAT_32950/Homework/HW3/stiffness.DAT")
 ↪  %>%
  rename(X1 = V1,
         X2 = V2,
         X3 = V3,
```

```
          X4 = V4,
          dj2 = V5)
```

The data were obtained by taking four different measures of stiffness, x_1, x_2, x_3, and x_4 of each of $n = 30$ boards. The first measurement involves sending a shock wave down the board, the second measurement is determined while vibrating the board, and the last two measurements are obtained from static tests. The squared distances $d_j^2 = (x_j - \bar{x})^T S^{-1}(x_j - \bar{x})$ are also included as the last column in the data. (ref. Table 4.3 in J&W)

Let $X = (X_1, X_2)$ be the vector of variables representing the dynamic measures of stiffness, and let $Y = (X_3, X_4)$ be the vector of variables representing the static measures of stiffness.

1.1 Part (a)

Perform a canonical correlation analysis of these data.

We perform a canonical correlation analysis of (X, Y) below.

```
# Create X and Y vectors
X = select(stiffness, c(X1, X2))
Y = select(stiffness, c(X3, X4))

cancor(X,Y)
```

```
$cor
[1] 0.91291935 0.06805556

$xcoef
             [,1]         [,2]
X1 -0.0006687933 -0.001237328
X2  0.0001106253  0.001430402

$ycoef
             [,1]         [,2]
X3 -0.0002497238  0.001573032
X4 -0.0003515941 -0.001453802

$xcenter
      X1       X2
1906.100 1749.533
```

```
$ycenter
      X3       X4
1509.133 1724.967
```

1.2 Part (b)

Write the first canonical variates U_1 and V_1 as linear combinations of the components of X and Y respectively.

The first canonical variates are:

$$U_1 = a_1^T X = \begin{bmatrix} a_{11} & a_{12} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \approx \begin{bmatrix} -0.000669 & 0.000111 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = -0.000669X_1 + 0.000111X_2$$

and

$$V_1 = b_1^T Y = \begin{bmatrix} b_{11} & b_{12} \end{bmatrix} \begin{bmatrix} X_3 \\ X_4 \end{bmatrix} \approx \begin{bmatrix} -0.000250 & -0.000352 \end{bmatrix} \begin{bmatrix} X_3 \\ X_4 \end{bmatrix} = -0.000250X_3 - 0.000352X_4.$$

1.3 Part (c)

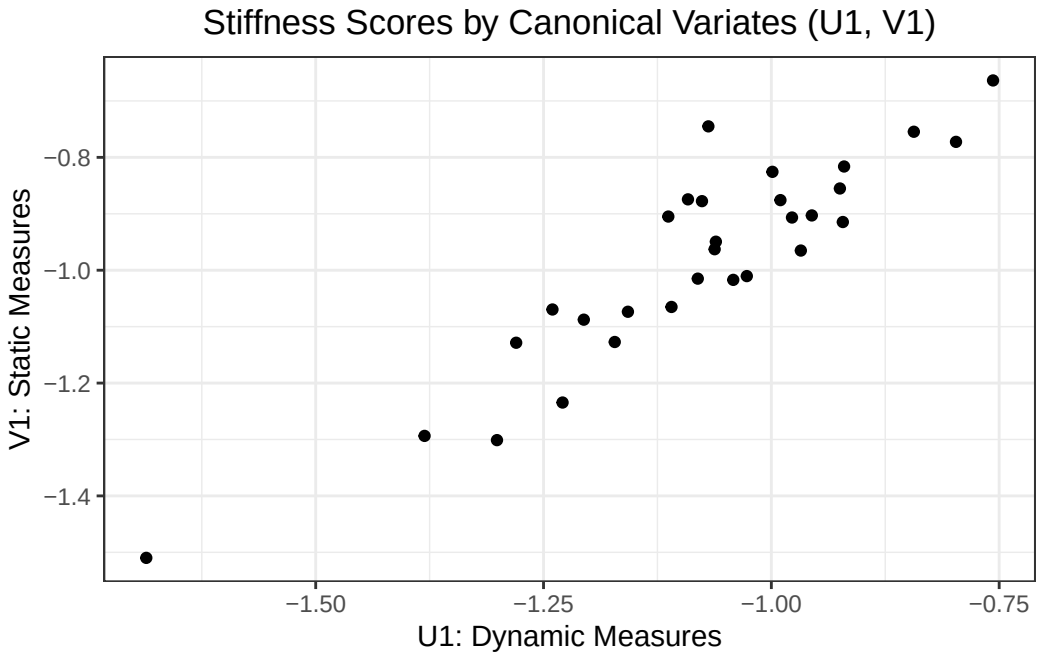
Produce two scatterplots of the data: one in the coordinate plane of the first canonical variate pair (U_1, V_1), and one in the plane of the second pair (U_2, V_2).

```r
# U data
U = as.matrix(X) %*% cancor(X,Y)$xcoef %>%
  as.data.frame() %>%
  rename(U1 = V1,
         U2 = V2)

# V data
V = as.matrix(Y) %*% cancor(X,Y)$ycoef %>%
  as.data.frame()
```

First, we plot the data in the (U_1, V_1) (i.e., first canonical variate) plane:

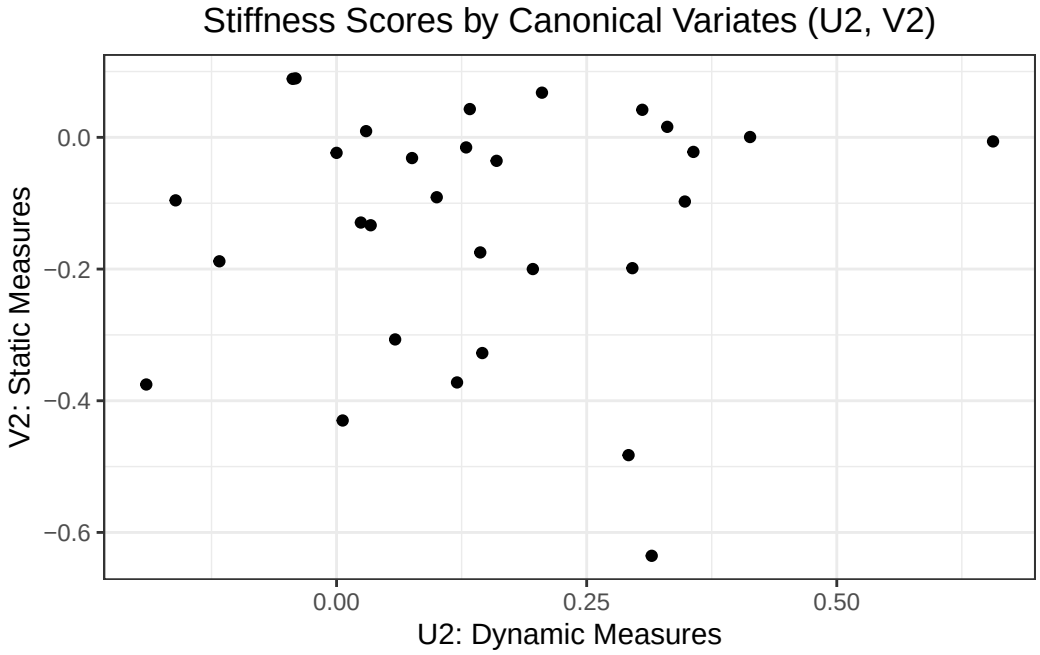
```r
# (U1, V1) scatterplot
ggplot(mapping = aes(x=U$U1, y = V$V1)) +
  theme_bw() +
  geom_point() +
  xlab("U1: Dynamic Measures") +
```

```
  ylab("V1: Static Measures") +
  ggtitle("Stiffness Scores by Canonical Variates (U1, V1)") +
  theme(plot.title = element_text(hjust = 0.5))
```



Now, we plot the data in the (U_2, V_2) (i.e., second canonical variate) plane:

```
# (U1, V1) scatterplot
ggplot(mapping = aes(x=U$U2, y = V$V2)) +
  theme_bw() +
  geom_point() +
  xlab("U2: Dynamic Measures") +
  ylab("V2: Static Measures") +
  ggtitle("Stiffness Scores by Canonical Variates (U2, V2)") +
  theme(plot.title = element_text(hjust = 0.5))
```



1.4 Part (d)

Based on the two pots and the values of the canonical correlations $\{\rho_1^*, \rho_2^*\}$, comment on the correlation structure "captured" by each canonical pair.

We found in Part (a) that the first canonical correlation, $\rho_1^* \approx 0.913$, is very close to 1. This strong correlation is also depicted in the first scatterplot in Part (c). As such, the first canonical variate pair (U_1, V_1) captures the strong correlation between a board's dynamic and static measures of stiffness. In particular, if a board's shock wave score (X_1) is low and its vibration score (X_2) is high, such that U_1 is high, then it is likely that the board's static stiffness scores (X_3, X_4) are both low, such that V_1 is high — and vice versa.

Meanwhile, we also saw in Part (a) that the second canonical correlation, $\rho_2^* \approx 0.068$, is very close to 0, making it significantly weaker. This weak (nearly nonexistent) correlation is also depicted in the second scatterplot in Part (c). This indicates that there is not much correlation between boards' dynamic scores (X_1, X_2) and static scores (X_3, X_4) beyond what is already explained by the correlation between (U_1, V_1), as described above.

2 Exercise 4: Multivariate vs. Univariate Inference

Input the dataset `fly.dat` of 15 observations on X_1 = antenna length (mm) and X_2 = wing length (mm) of two species of flies.

```
fly =
 ↪  read.table("C:/Users/rewin/OneDrive/Documents/STAT_32950/Homework/HW3/fly.dat")
```

Define two univariate variables Y_1 = {antenna length} + {wing length}, Y_2 = {wing length}. Treat the data as bivariate samples from two populations (species `Af` and `Apf`) with equal covariance.

```
fly = fly %>%
  mutate(Y1 = Ant.Length + Wing.Length,
         Y2 = Wing.Length)

af = fly %>% subset(Species == "Af")
apf = fly %>% subset(Species == "Apf")
```

2.1 Part (a): Property of Hotelling's T^2

2.1.1 Part (i)

Compute a Hotelling's T^2-statistic for the hypothesis of equality of the mean vectors in the two species based on (Y_1, Y_2). Is the hypothesis of equality of the means accepted?

Let μ_{Af} denote the mean vector of Y_1 and Y_2 for flies of species Af, and μ_{Apf} denote the corresponding mean vector for flies of species Apf. Then Hotelling's T^2 statistic corresponding to $H_0 : \mu_{Af} = \mu_{Apf}$ is $T^2_* \approx 55.881$. Under H_0, $T^2 \sim \frac{(n_{Af}+n_{Apf}-2)p}{n_{Af}+n_{Apf}-p-1} F_{p,n_{Af}+n_{Apf}-p-1} = \frac{(9+6-2)(2)}{9+6-2-1} F_{2,9+6-2-1} = \frac{13}{6} F_{2,12}$; equivalently, $\frac{6}{13} T^2 \sim F_{2,12}$. Then under H_0, $p = \mathbb{P}(T^2 > 55.881) = \mathbb{P}(\frac{6}{13} T^2 > \frac{6}{13}(55.881)) \approx \mathbb{P}(\frac{6}{13} T^2 > 25.791) \approx 4.519 \times 10^{-5} < 0.001$. Thus, we reject the null hypothesis that the mean vectors of the two species of flies are equal.

```
# "Scaled" T2 stat and p-value
HotellingsT2Test((select(af, c(Y1, Y2))),
                 (select(apf, c(Y1, Y2))))
```

```
    Hotelling's two sample T2-test

data:  (select(af, c(Y1, Y2))) and (select(apf, c(Y1, Y2)))
T.2 = 25.791, df1 = 2, df2 = 12, p-value = 4.519e-05
alternative hypothesis: true location difference is not equal to c(0,0)
```

```r
# "Unscaled" T2 stat
n_Af = nrow(af)
n_Apf = nrow(apf)
p = 2
(n_Af + n_Apf - 2)*p / (n_Af + n_Apf - 1 - p) *
 ↪  HotellingsT2Test((select(af, c(Y1, Y2))),

 ↪  (select(apf, c(Y1, Y2))))$statistic
```

```
        [,1]
[1,] 55.8807
```

2.1.2 Part (ii)

Should you get the same results if you use the original variables (X_1, X_2)? Why?

We should (and do!) get the same T^2 statistic and p-value if we use the original variables (X_1, X_2) rather than the transformed (Y_1, Y_2). Recall that in Exercise 1(c), we showed that if C is an invertible matrix and we transform data X_j into $Y_j = CX_j$, then Hotelling's T^2 statistic under some null hypothesis with respect to the X_j's is equal to Hotelling's T^2 statistic under the corresponding/transformed null hypothesis with respect to the Y_j's. In this case, let $C = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ and observe that $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = C \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$, since $Y_1 = X_1 + X_2$ and $Y_2 = X_2$. C is invertible; in particular, $C^{-1} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$. Thus, we'd recover the same T^2 statistic if we test the null hypothesis that the difference-in-means vector of the Y variables equals $\mathbf{0}$ as we did when we tested the null hypothesis that the difference-in-means vector of the X variables equals $\mathbf{0}$.

2.2 Part (b)

If you conduct (univariate) two-sample t-tests at a test level $\alpha = 0.05$ performed on each of the variables Y_1 and Y_2 separately (i.e., assuming independence of Y_1 and Y_2), would the hypothesis of equality of species means be accepted? What if the test level $\alpha = 0.01$?

Performing a univariate two-sample t-test on the difference in the means of Y_1 between the two species, we recover a p-value of $0.5202 > 0.05$. Thus, we fail to reject the null hypothesis that the two species have equal means of Y_1 at both the $\alpha = 0.05$ and $\alpha = 0.01$ levels.

```r
t.test(data = fly,
       Y1 ~ Species,
       var.equal = TRUE)
```

```
    Two Sample t-test

data:  Y1 by Species
t = 0.66097, df = 13, p-value = 0.5202
alternative hypothesis: true difference in means between group Af and group Apf is not equal
95 percent confidence interval:
 -0.1461906  0.2750795
sample estimates:
 mean in group Af mean in group Apf
         3.217778          3.153333
```

On the other hand, performing a univariate two-sample t-test on the difference in the means of Y_2 between the two species, we recover a p-value of $0.066 > 0.05$. Thus, we fail to reject the null hypothesis that the two species have equal means of Y_2 at both the $\alpha = 0.05$ and $\alpha = 0.01$ levels.

```
t.test(data = fly,
       Y2 ~ Species,
       var.equal = TRUE)
```

```
    Two Sample t-test

data:  Y2 by Species
t = -2.0047, df = 13, p-value = 0.06628
alternative hypothesis: true difference in means between group Af and group Apf is not equal
95 percent confidence interval:
 -0.253933843  0.009489398
sample estimates:
 mean in group Af mean in group Apf
         1.804444          1.926667
```

2.3 Part (c)

Draw a scatterplot of Y_1 vs. Y_2 for the data in both species groups, marking the data points of the two species groups with different symbols, and explain how it can happen that Parts (a) and (b) have different conclusions.

We plot the flies in our dataset in the (Y_1, Y_2) plane in the scatterplot below.

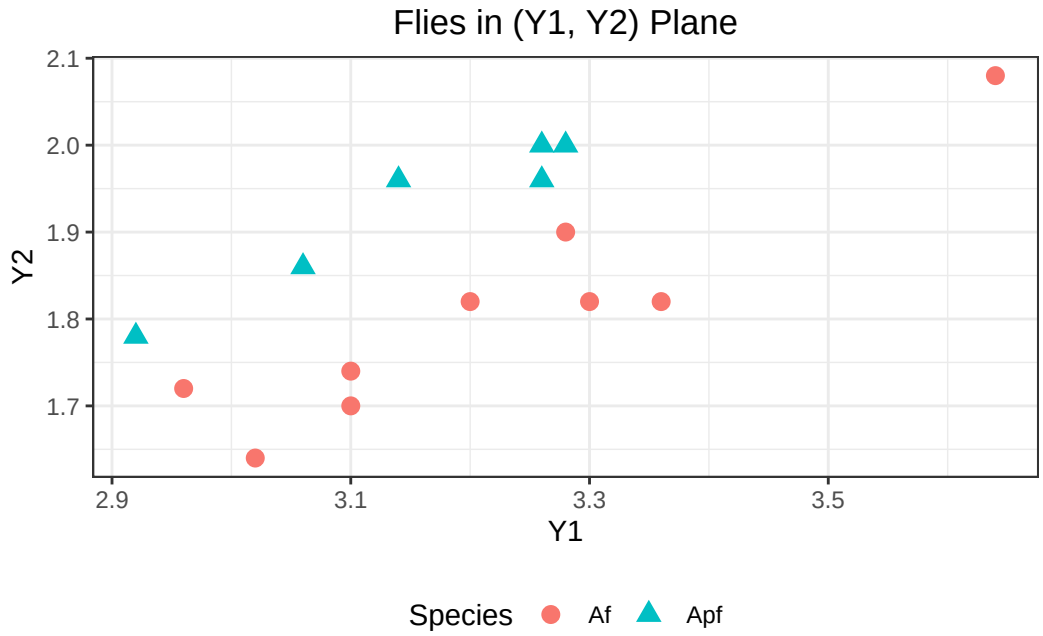
When we look just at the Y_1 coordinates of the points (imagine, for example, projecting all the points onto the Y_1-axis), notice that each species' flies are distributed rather similarly, with a great deal of overlap between the two species and both species' means around 3.2.

This explains our first result in Part (b): when we look just at the flies' Y_1 coordinates, there does not appear to be a significant difference between the two species' means.

However, when we look just at the Y_2 coordinates of the points (imagine, again, projecting all the points onto the Y_2-axis), the two species are more differentiated. In particular, flies of the Af species tend to have smaller values of Y_2 than flies of the Apf species. On the Y_2-axis, the two species still have some overlap (e.g., some Af flies have larger Y_2 values than some Apf flies, and one Af fly has a larger Y_2 value than *every* Apf fly!), but much less than in the previous case. This helps explain our second result in Part (b): when we look just at the flies' Y_2 coordinates, there is a nearly-significant difference between the two species means at the $\alpha = 0.05$ level, where this lack of significance is likely hampered by the overlap between the two groups.

Finally, when we look at the (Y_1, Y_2) coordinates of the points together, the two species are highly distinguishable. In particular, each species' points form a "diagonal," upward-sloping cloud, with the Apf cloud shifted slightly above the Af cloud. Now, visualized in two full dimensions, the two species' clouds have no overlap (at least in this particular dataset): if we imagine drawing a tight shape around all the Af points and another tight shape around all the Apf points, the two shapes do not intersect. From this perspective, the two groups have clearly discernible means, which is why we recovered such a significant p-value in Part (a).

```
ggplot(fly, aes(x = Y1, y = Y2, col = Species, shape = Species)) +
  theme_bw() +
  geom_point(size = 3) +
  theme(legend.position = "bottom") +
  ggtitle("Flies in (Y1, Y2) Plane") +
  theme(plot.title = element_text(hjust = 0.5))
```



2.4 Part (d): Comparison: Confidence Region vs. Simultaneous Confidence Intervals

2.4.1 Part (i)

Draw a 98% confidence ellipse for the species mean differences of Y_1 and Y_2 based on Hotelling's T^2.

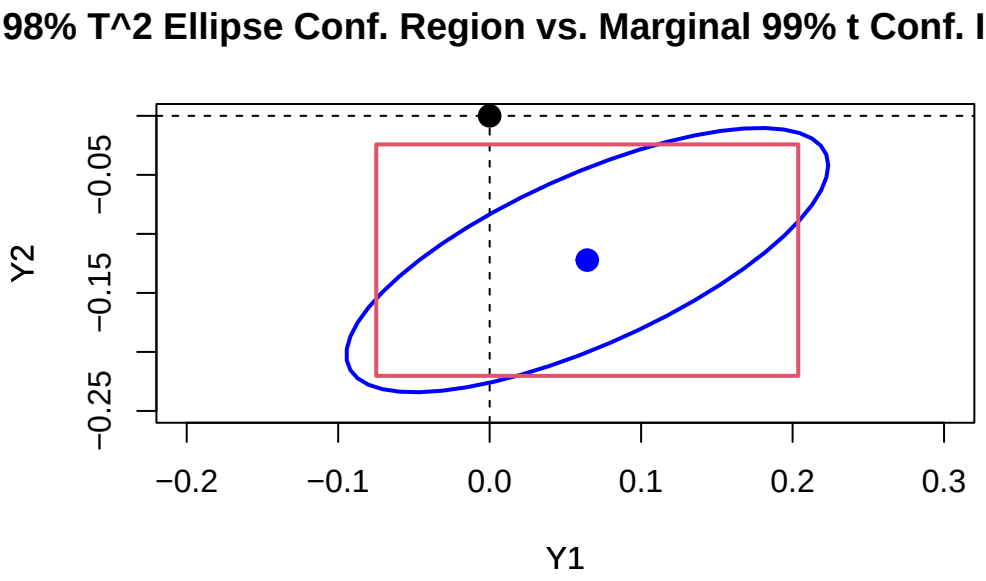
See the plot under Part (ii) below.

2.4.2 Part (ii)

In the same graph, draw a rectangle corresponding to univariate (marginal) 99% confidence intervals for the mean differences of Y_1 and Y_2.

```
Ydata = cbind(fly$Y1, fly$Y2)

dy1bar = mean(af$Y1) - mean(apf$Y1)
dy2bar = mean(af$Y2) - mean(apf$Y2)

n = nrow(fly)
p = 2

# For Ellipse:
```

```r
F = qf(0.98, df1 = p, df2 = n-p)
r = sqrt(F*(n-1)*p/(n*(n-p)))

# For rectangle:
alpha = 0.01
se = sqrt(diag(cov(Ydata)))/sqrt(n)
q = 1-(alpha/2)
cr = qt(q,n-1)
L = dy1bar - cr*se[1]
R = dy1bar + cr*se[1]
D = dy2bar - cr*se[2]
U = dy2bar + cr*se[2]

# Plot
# plot(dy1bar, dy2bar, xlim = c(-0.5,0.5), ylim = c(-0.25, 0), xlab = "Y1",
 ↪  ylab = "Y2") # center of regions
plot(0, 0, pch = 19, cex = 1.5, col = "black", xlim = c(-0.2,0.3), ylim =
 ↪  c(-0.25, 0), xlab = "Y1", ylab = "Y2") # origin (0,0)
abline(v=0, col = "black", lty = 2) # dotted y-axis
abline(h=0, col = "black", lty = 2) # dotted x-axis
ellipse(c(dy1bar, dy2bar), shape = cov(Ydata), radius = r) # ellipse CR
rect(L, U, R, D, border = 2, lwd = 2) # marginal rectangular CR
title("98% T^2 Ellipse Conf. Region vs. Marginal 99% t Conf. Intervals",
 ↪  xlab = "Y1", ylab = "Y2")
```



2.4.3 Part (iii)

Explain that the rectangle is a 98% confidence region by the Bonferroni method.

In Part (ii), we constructed 99% marginal simultaneous confidence intervals along each axis using the following formula. Since the confidence intervals were at the 99% level, we had $\alpha = 0.01$, and we considered two confidence intervals, corresponding to $k \in \{1, 2\}$, with one on each axis. $\Delta\bar{Y}_k$ denotes the between-species difference in the means of Y_k for $k \in \{1, 2\}$, and s_{kk} denotes the (k, k) entry in the 2×2 sample covariance matrix $S = \text{Cov}(Y_1, Y_2)$.

$$\begin{aligned} CI_k^{marg} &= \left(\Delta\bar{Y}_k - t_{n-1,\alpha/2}\sqrt{\frac{s_{kk}}{n}}, \Delta\bar{Y}_k + t_{n-1,\alpha/2}\sqrt{\frac{s_{kk}}{n}} \right) \\ &= \left(\Delta\bar{Y}_k - t_{15-1,0.01/2}\sqrt{\frac{s_{kk}}{15}}, \Delta\bar{Y}_k + t_{15-1,0.01/2}\sqrt{\frac{s_{kk}}{15}} \right) \\ &= \left(\Delta\bar{Y}_k - t_{14,0.005}\sqrt{\frac{s_{kk}}{15}}, \Delta\bar{Y}_k + t_{14,0.005}\sqrt{\frac{s_{kk}}{15}} \right) \end{aligned}$$

If we were to construct 98% Bonferonni simultaneous confidence intervals along each axis, we would use the following formula. Since these confidence intervals would be at the 98% level, we'd have $\alpha = 0.02$. Again, we consider two confidence intervals, corresponding to $k \in \{1, 2\}$, with one interval on each axis. And since we are considering two intervals (corresponding to two variables) in total, we have $p = 2$.

$$\begin{aligned} CI_k^{Bonf} &= \left(\Delta\bar{Y}_k - t_{n-1,\alpha/(2p)}\sqrt{\frac{s_{kk}}{n}}, \Delta\bar{Y}_k + t_{n-1,\alpha/(2p)}\sqrt{\frac{s_{kk}}{n}} \right) \\ &= \left(\Delta\bar{Y}_k - t_{15-1,0.02/(2\cdot 2)}\sqrt{\frac{s_{kk}}{15}}, \Delta\bar{Y}_k + t_{15-1,0.02/(2\cdot 2)}\sqrt{\frac{s_{kk}}{15}} \right) \\ &= \left(\Delta\bar{Y}_k - t_{14,0.005}\sqrt{\frac{s_{kk}}{15}}, \Delta\bar{Y}_k + t_{14,0.005}\sqrt{\frac{s_{kk}}{15}} \right) \\ &= CI_k^{marg} \end{aligned}$$

That is, the Bonferroni 98% confidence intervals are precisely equal to the marginal 99% confidence intervals. In short, this is because we have doubled the α level from 0.01 to 0.02, but we have also scaled the α level down by a factor of 2 to account for the two simultaneous intervals being considered. These two factors essentially "cancel each other out," leaving the confidence intervals unchanged from Part (ii).

2.4.4 Part (iv)

Is the zero vector $\mathbf{0} = (0, 0)$ in any of the two regions (ellipse by Part (i) or rectangle by Part (ii))? Compare and comment on the goodness of the two regions. Which one is better?

As shown above, neither of the two confidence regions contains the zero vector. That is, using either T^2 confidence ellipses or simultaneous Bonferroni confidence intervals, we find that the two fly species have statistically significantly different average (Y_1, Y_2) profiles at the $\alpha = 0.02$ level. While both region constructions are useful, the T^2-based elliptic region is more informative. Indeed, this region's area is clearly smaller than that of the confidence rectangle, and as such, it gives a more precise set of possible $(\Delta\bar{Y}_1, \Delta\bar{Y}2)$ pairs than the confidence rectangle. Unlike the confidence rectangle, the elliptic region also captures the covariance structure of Y_1 and Y_2, giving the reader a clearer sense of the relationship between the two variables. On the other hand, the rectangular region is arguably more interpretable than the ellipse, as it allows the reader to easily pick out the confidence interval along one axis without needing to condition on a covariate value along the other axis.