MA554 Applied Multivariate Analysis HW3

Aukkawut Ammartayakun

2023-10-29

Problem 1

Prove that $(n-p-1)\mathbf{S} \sim W_m(n-p-1, \mathbf{\Sigma})$ and $\mathbf{S}$ is independent of $\hat{\mathbf{B}}$.

Proof To show that $(n-p-1)\mathbf{S} \sim W_m(n-p-1, \mathbf{\Sigma})$, using the fact that

$$\begin{aligned}\hat{\mathbf{E}}^\top\hat{\mathbf{E}} &= \mathbf{E}^\top(\mathbb{I}-\mathbf{H})^\top(\mathbb{I}-\mathbf{H})\mathbf{E} \\ &= \mathbf{E}^\top(\mathbb{I}-\mathbf{H})\mathbf{E}\end{aligned}$$

and eigendecomposition of $(\mathbb{I}-\mathbf{H})$ which is $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ where $\mathbf{\Lambda} = \text{diag}(1,1,\dots,0)$ snd $\text{rank}(\mathbf{\Lambda}) = n-p-1$. Let $\tilde{\mathbf{E}} = \mathbf{U}^\top\mathbf{E}$, then

$$\hat{\mathbf{E}}^\top\hat{\mathbf{E}} = \tilde{\mathbf{E}}^\top \begin{bmatrix} \mathbb{I}_{n-p-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tilde{\mathbf{E}}$$

Note that the middle matrix is a $m \times m$ matrix with the rank of $n-p-1$. We can see that the matrix in the middle is Wishart distributed with degree of freedom $n-p-1$ and dimension m with

Problem 2

The file `plankton.txt` contains data on the catches of 5 varieties of plankton recorded by a marine research vessel. A total of 25 successive hauls were made. In each haul, two fine mesh nets were simultaneously dragged, one at 29 meter deep and the other at 31 meter deep. For each net in each haul, estimates were made of the "catch" (number of caught) for the 5 varieties A, B, C, D, and E.

```
plankton <- read.table("plankton.txt", skip = 13, nrow = 25)
names(plankton) <- c(paste0(LETTERS[1:5],"_29"), paste0(LETTERS[1:5],"_31"))
head(plankton)
```

```
##    A_29  B_29  C_29  D_29  E_29  A_31  B_31  C_31  D_31  E_31
## 1 2.321 1.537 3.332 2.945 1.907 2.339 1.783 3.093 2.316 1.744
## 2 2.835 2.316 2.745 3.075 1.994 2.862 1.966 2.383 2.991 1.808
## 3 2.748 2.007 2.929 3.107 1.838 2.578 1.605 2.352 3.054 1.435
## 4 2.981 2.142 2.956 3.124 1.762 2.832 1.518 2.179 2.478 1.593
## 5 2.130 1.386 2.586 2.876 1.670 2.186 1.478 2.428 3.096 1.419
## 6 2.867 1.874 3.276 2.843 1.860 2.642 1.717 2.822 2.433 1.589
```

The values of the counts of A, B, C, D and E for depth 29 meters are in columns 1-5; similar counts for depth 31 meters are in columns 6-10. You should view this as multivariate data with $n = 25$ and $p = 10$. Each row corresponds to one haul. You may assume the data consist of a random sample of size 25 from a multivariate normal population. Since numbers in the same row measure the same haul at different depths and for different varieties, it is repeated measurement data, on five measurements, over two levels of a factor (depth = 29 and 31).

a) Test the null hypothesis at the 5% level that, for each variety, the expected catches of plankton were the same at both depths? Do this in two ways, using a Hotelling's T^2 statistic (hence a multivariate test) and simultaneous Bonferronized t-tests (hence a bunch of univariate tests with simultaneous guarantees).

In the first case, we want to test

$$H_0 : \boldsymbol{\mu}_{29} - \boldsymbol{\mu}_{31} = \mathbf{0}$$
$$H_1 : \boldsymbol{\mu}_{29} - \boldsymbol{\mu}_{31} \neq \mathbf{0}$$

However, we can define $\boldsymbol{\delta} = \boldsymbol{\mu}_{29} - \boldsymbol{\mu}_{31}$ and make this a one-sample test.

```
# Calculate mean difference vector
differences <- plankton[, 1:5] - plankton[, 6:10]
mean_differences <- colMeans(differences)
mean_differences
```

```
##    A_29    B_29    C_29    D_29    E_29
## 0.14020 0.18664 0.49976 0.43068 0.32592
```

```
# do a T-test
test_result <- HotellingsT2Test(differences)
print(test_result)
```

```
##
##  Hotelling's one sample T2-test
##
## data:  differences
## T.2 = 42.089, df1 = 5, df2 = 20, p-value = 6.081e-10
## alternative hypothesis: true location is not equal to c(0,0,0,0,0)
```

Since p-value is small (6.081×10^{-10}), we can reject the null hypothesis. That is, there is a significant difference between the expected catches of plankton in both depth. This can also be seen from the mean difference calculated earlier although it is not that clear whether that is significant or not at the beginning.

In the second case, we will use the Bonforroni correction, that is the cut-off p-value is now $\frac{5}{m}\% = 1\%$ while we performing $m = 5$ hypothesis tests:

$$H_0 : \mu_{X,29} - \mu_{X,31} = 0$$
$$H_1 : \mu_{X,29} - \mu_{X,31} \neq 0$$

for $X \in \{A, B, C, D, E\}$.

```
bonforroni <- apply(differences, 2, t.test)
bonforroni
```

```
## $A_29
##
##  One Sample t-test
##
## data:  newX[, i]
## t = 2.775, df = 24, p-value = 0.01052
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.03592573 0.24447427
## sample estimates:
## mean of x
##    0.1402
```

```
##
##
## $B_29
##
##  One Sample t-test
##
## data:  newX[, i]
## t = 3.9534, df = 24, p-value = 0.0005928
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.08920235 0.28407765
## sample estimates:
## mean of x
##   0.18664
##
##
## $C_29
##
##  One Sample t-test
##
## data:  newX[, i]
## t = 11.893, df = 24, p-value = 1.504e-11
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.4130316 0.5864884
## sample estimates:
## mean of x
##   0.49976
##
##
## $D_29
##
##  One Sample t-test
##
## data:  newX[, i]
## t = 7.281, df = 24, p-value = 1.603e-07
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.308598 0.552762
## sample estimates:
## mean of x
##   0.43068
##
##
## $E_29
##
##  One Sample t-test
##
## data:  newX[, i]
## t = 6.4252, df = 24, p-value = 1.209e-06
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.2212282 0.4306118
## sample estimates:
```

```
## mean of x
##   0.32592
```

From the test results, we can see that we fail to reject null hypothesis for A but the rest of them, we can say that we reject the null hypothesis that there is no difference in mean for each type of catch in different depth.

b) Following (a), identify all varieties, if any, which differed between depths.

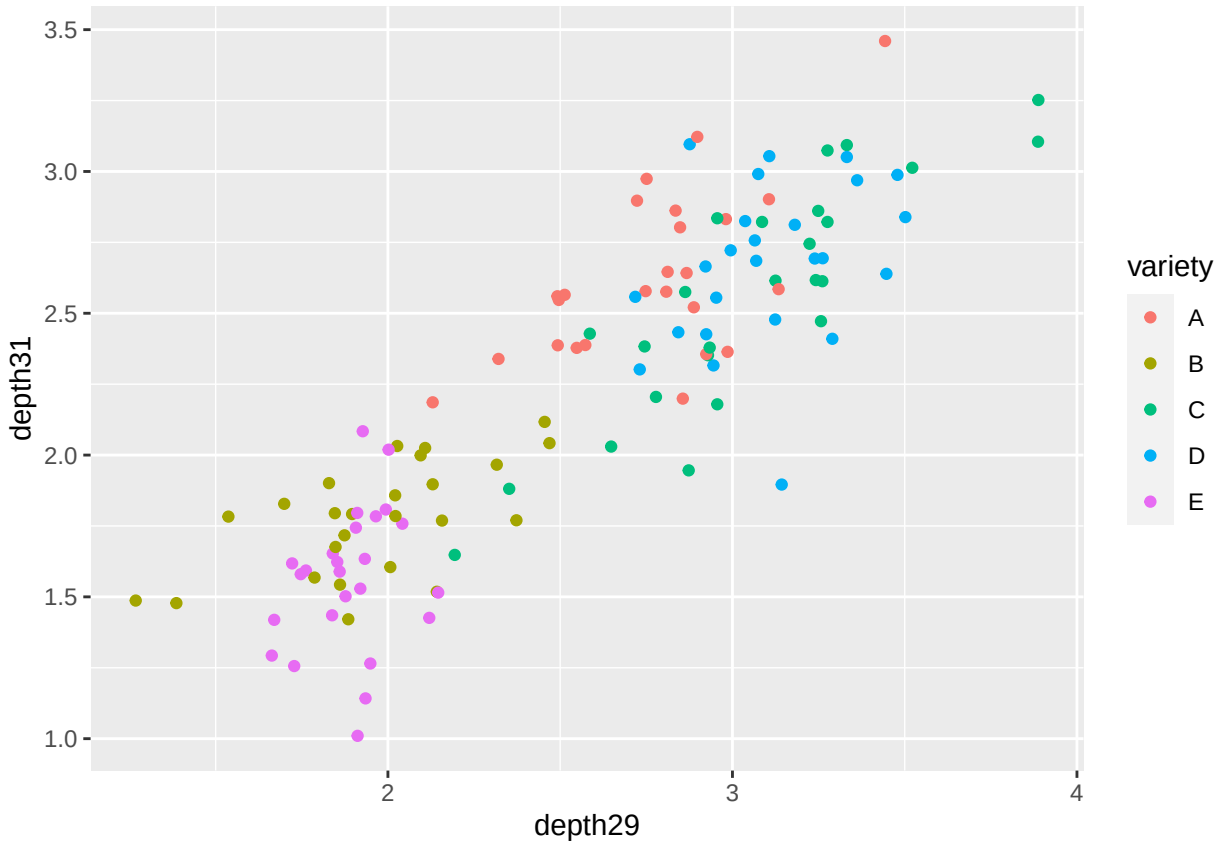
As said in (a), From the test results, we can see that we fail to reject null hypothesis for A but the rest of them, we can say that we reject the null hypothesis that there is no difference in mean for each type of catch in different depth. In another word, B, C, D, and E catch are differs between depths.

c) Here is a different way to think of this data set. Each haul corresponds to 5 groups (the 5 varieties), and the data point is 2-dimensional, representing the catches from the two different depths. See the data frame `plankton2` below. Think of 5 groups of 2-dimensional vectors (with 25 vectors in each group). Conduct a Multivariate Analysis of Variance (MANOVA) to see if there was any statistical evidence to suggest that the five varieties are different in terms of their mean catch, measured by 2-dimension data at two depths. In addition, please draw a scatter plot of these 2-dimensional data points with colors indicating the five varieties.

```
plankton2 <- plankton %>% mutate(haul = 1:n()) %>%
        pivot_longer(A_29:E_31,
                     names_to = c("variety", ".value"),
                     names_pattern = "([A-Z])_(.*)") %>%
        rename(depth29 = '29', depth31 = '31')
head(plankton2,10)
```

```
## # A tibble: 10 x 4
##     haul variety depth29 depth31
##    <int> <chr>     <dbl>   <dbl>
##  1     1 A          2.32    2.34
##  2     1 B          1.54    1.78
##  3     1 C          3.33    3.09
##  4     1 D          2.94    2.32
##  5     1 E          1.91    1.74
##  6     2 A          2.84    2.86
##  7     2 B          2.32    1.97
##  8     2 C          2.74    2.38
##  9     2 D          3.08    2.99
## 10     2 E          1.99    1.81
```

```
#scatterplot
ggplot(plankton2, aes(x=depth29, y=depth31, colour = variety)) + geom_point()
```




```
fit <- manova(cbind(depth29,depth31) ~ variety, data = plankton2)
summary(fit, test = "Wilks")
```

```
##            Df   Wilks approx F num Df den Df    Pr(>F)
## variety     4 0.16267   44.012      8    238 < 2.2e-16 ***
## Residuals 120
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From MANOVA test, we can see that p-value is 2.2×10^{-16} and that means there is statistical evidence against null hypothesis, that is the treatment effect is not zero. In another word, there is a difference in mean between number of the different plankton caught in different depth.

Problem 3

Amitriptyline is prescribed by some physicians as an antidepressant. However, there are also conjectured side effects that seem to be related to the use of the drug: irregular heartbeat, abnormal blood pressures, and irregular waves on the electrocardiogram among other things. Data gathered on 17 patients who were admitted to the hospital after an amitriptyline overdose are given in the data set `JW_ami_data.txt`. The two response variables are

- Y_1 Total TCAD plasma level (TOT)
- Y_2 Amount of amitriptyline present in TCAD plasma level (AMI)

The five predictor variables are

- Z_1 Gender: 1 if female, 0 if male (GEN)
- Z_2 Amount of antidepressants taken at time of overdose (AMT)
- Z_3 PR wave measurement (PR)

- Z_4 Diastolic blood pressure (DIAP)
- Z_5 QRS wave measurement (QRS)

```
ami_data <- read.table("JW_ami_data.txt")
names(ami_data) <- c("TOT", "AMI", "GEN", "AMT", "PR", "DIAP", "QRS")
summary(ami_data)
```

```
##       TOT             AMI              GEN              AMT
##  Min.   : 500   Min.   : 384.0   Min.   :0.0000   Min.   : 350
##  1st Qu.: 652   1st Qu.: 458.0   1st Qu.:0.0000   1st Qu.: 750
##  Median : 896   Median : 653.0   Median :1.0000   Median :1750
##  Mean   :1120   Mean   : 882.4   Mean   :0.7059   Mean   :2146
##  3rd Qu.:1131   3rd Qu.: 941.0   3rd Qu.:1.0000   3rd Qu.:3000
##  Max.   :3389   Max.   :3149.0   Max.   :1.0000   Max.   :7500
##        PR             DIAP          QRS
##  Min.   :135.0   Min.   : 0   Min.   : 60.00
##  1st Qu.:160.0   1st Qu.:60   1st Qu.: 80.00
##  Median :180.0   Median :60   Median : 98.00
##  Mean   :174.9   Mean   :52   Mean   : 97.18
##  3rd Qu.:185.0   3rd Qu.:70   3rd Qu.:111.00
##  Max.   :220.0   Max.   :90   Max.   :140.00
```

```
m1 <- lm(TOT ~ GEN+AMT+PR+DIAP+QRS, data  = ami_data)
summary(m1)
```

a) Perform a regression analysis using only the first response Y_1 (TOT). Suggest and fit an appropriate linear regression model. Identify all independent variables (not including the intercept, since one is typically not interested in the intercept) that have non-zero coefficients for predicting Y_1 at the 5% level, using the Bonferroni correction. Construct a 95% prediction interval for Total TCAD for $z_1 = 1, z_2 = 1200, z_3 = 140, z_4 = 70$, and $z_5 = 85$.

```
##
## Call:
## lm(formula = TOT ~ GEN + AMT + PR + DIAP + QRS, data = ami_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -399.2 -180.1    4.5  164.1  366.8
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.879e+03  8.933e+02  -3.224 0.008108 **
## GEN          6.757e+02  1.621e+02   4.169 0.001565 **
## AMT          2.848e-01  6.091e-02   4.677 0.000675 ***
## PR           1.027e+01  4.255e+00   2.414 0.034358 *
## DIAP         7.251e+00  3.225e+00   2.248 0.046026 *
## QRS          7.598e+00  3.849e+00   1.974 0.074006 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 281.2 on 11 degrees of freedom
## Multiple R-squared:  0.8871, Adjusted R-squared:  0.8358
## F-statistic: 17.29 on 5 and 11 DF,  p-value: 6.983e-05
```

```
p1 <- data.frame(
        GEN = 1,
        AMT = 1200,
        PR = 140,
        DIAP = 70,
        QRS = 85
)
predict(m1,p1, interval = 'prediction', level = 0.99)
```

```
##        fit       lwr     upr
## 1 729.5248 -241.5605 1700.61
```

```
m2 <- lm(AMI ~ GEN+AMT+PR+DIAP+QRS, data  = ami_data)
summary(m2)
```

b) Repeat Part (a) using the second response Y_2 (AMI).

```
##
## Call:
## lm(formula = AMI ~ GEN + AMT + PR + DIAP + QRS, data = ami_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -373.85 -247.29  -83.74  217.13  462.72
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.729e+03  9.288e+02  -2.938 0.013502 *
## GEN          7.630e+02  1.685e+02   4.528 0.000861 ***
## AMT          3.064e-01  6.334e-02   4.837 0.000521 ***
## PR           8.896e+00  4.424e+00   2.011 0.069515 .
## DIAP         7.206e+00  3.354e+00   2.149 0.054782 .
## QRS          4.987e+00  4.002e+00   1.246 0.238622
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 292.4 on 11 degrees of freedom
## Multiple R-squared:  0.8764, Adjusted R-squared:  0.8202
## F-statistic:  15.6 on 5 and 11 DF,  p-value: 0.0001132
```

```
p1 <- data.frame(
        GEN = 1,
        AMT = 1200,
        PR = 140,
        DIAP = 70,
        QRS = 85
)
predict(m2,p1, interval = 'prediction', level = 0.99)
```

```
##        fit       lwr      upr
## 1 575.7255 -434.0465 1585.497
```

```
m3 <- lm(cbind(TOT,AMI) ~ GEN+AMT+PR+DIAP+QRS, data  = ami_data)
```

```
summary(m3)
```

c) Perform a multivariate multiple regression analysis using both responses Y_1 and Y_2. Suggest and fit an appropriate linear regression model. Construct a 95% prediction ellipse for both Total TCAD and Amount of amitriptyline for $z_1 = 1, z_2 = 1200, z_3 = 140, z_4 = 70$, and $z_5 = 85$.

```
## Response TOT :
##
## Call:
## lm(formula = TOT ~ GEN + AMT + PR + DIAP + QRS, data = ami_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -399.2 -180.1    4.5  164.1  366.8
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.879e+03  8.933e+02  -3.224 0.008108 **
## GEN          6.757e+02  1.621e+02   4.169 0.001565 **
## AMT          2.848e-01  6.091e-02   4.677 0.000675 ***
## PR           1.027e+01  4.255e+00   2.414 0.034358 *
## DIAP         7.251e+00  3.225e+00   2.248 0.046026 *
## QRS          7.598e+00  3.849e+00   1.974 0.074006 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 281.2 on 11 degrees of freedom
## Multiple R-squared:  0.8871, Adjusted R-squared:  0.8358
## F-statistic: 17.29 on 5 and 11 DF,  p-value: 6.983e-05
##
##
## Response AMI :
##
## Call:
## lm(formula = AMI ~ GEN + AMT + PR + DIAP + QRS, data = ami_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -373.85 -247.29  -83.74  217.13  462.72
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.729e+03  9.288e+02  -2.938 0.013502 *
## GEN          7.630e+02  1.685e+02   4.528 0.000861 ***
## AMT          3.064e-01  6.334e-02   4.837 0.000521 ***
## PR           8.896e+00  4.424e+00   2.011 0.069515 .
## DIAP         7.206e+00  3.354e+00   2.149 0.054782 .
## QRS          4.987e+00  4.002e+00   1.246 0.238622
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 292.4 on 11 degrees of freedom
## Multiple R-squared:  0.8764, Adjusted R-squared:  0.8202
## F-statistic:  15.6 on 5 and 11 DF,  p-value: 0.0001132
```

d) Compare this ellipse with the prediction intervals in Parts (a) and (b). Comment on their differences. Note: the prediction intervals in Parts (a) and (b) would collectively provide a prediction rectangle here.

e) Identify all independent variables (not including the intercept) that have non-zero coefficients for predicting Y_1 or Y_2 at the 5% level, using the Bonferroni correction. In other words, we are testing which of the 10 β values ($2 \times 5 = 10$) is non-zero.