Homework 4 Solution

Due: 3/13/2019 before 11pm. Submit in Canvas (file upload). Rmd file and the html output file (submit both files) are strongly recommended, but not required.

1. [20 points]

Verify the conclusions on the distribution of $\widehat{\beta}_k$ and Cov$(\widehat{\beta}_k, \widehat{\beta}_i)$ in our lecture note `Manova-I`, page 377.

Proof: Since $X_{(k)} = A\beta_k + \epsilon_{(k)}$, then $E(X_{(k)}) = A\beta_k, Var(X_{(k)}) = \sigma_{kk}I_n, Cov(X_{(k)}, X_{(i)}) = \sigma_{ki}I_n$. Since $\hat{\beta}_k = (A'A)^{-1}A'X_{(k)}$,

$$
\begin{aligned}
E(\hat{\beta}_k) &= (A'A)^{-1}A'E(X_{(k)}) = (A'A)^{-1}A'A\beta_k = \beta_k \\
Var(\hat{\beta}_k) &= (A'A)^{-1}A'Var(X_{(k)})A(A'A)^{-1} = (A'A)^{-1}A'\sigma_{kk}I_nA(A'A)^{-1} = \sigma_{kk}(A'A)^{-1} \\
Cov(\hat{\beta}_k, \hat{\beta}_i) &= (A'A)^{-1}A'Cov(X_{(k)}, X_{(i)})A(A'A)^{-1} = (A'A)^{-1}A'\sigma_{ki}I_nA(A'A)^{-1} = \sigma_{ki}(A'A)^{-1} \\
\hat{\beta}_k &\sim N(\beta_k, \sigma_{kk}(A'A)^{-1})
\end{aligned}
$$

2. [45 points]

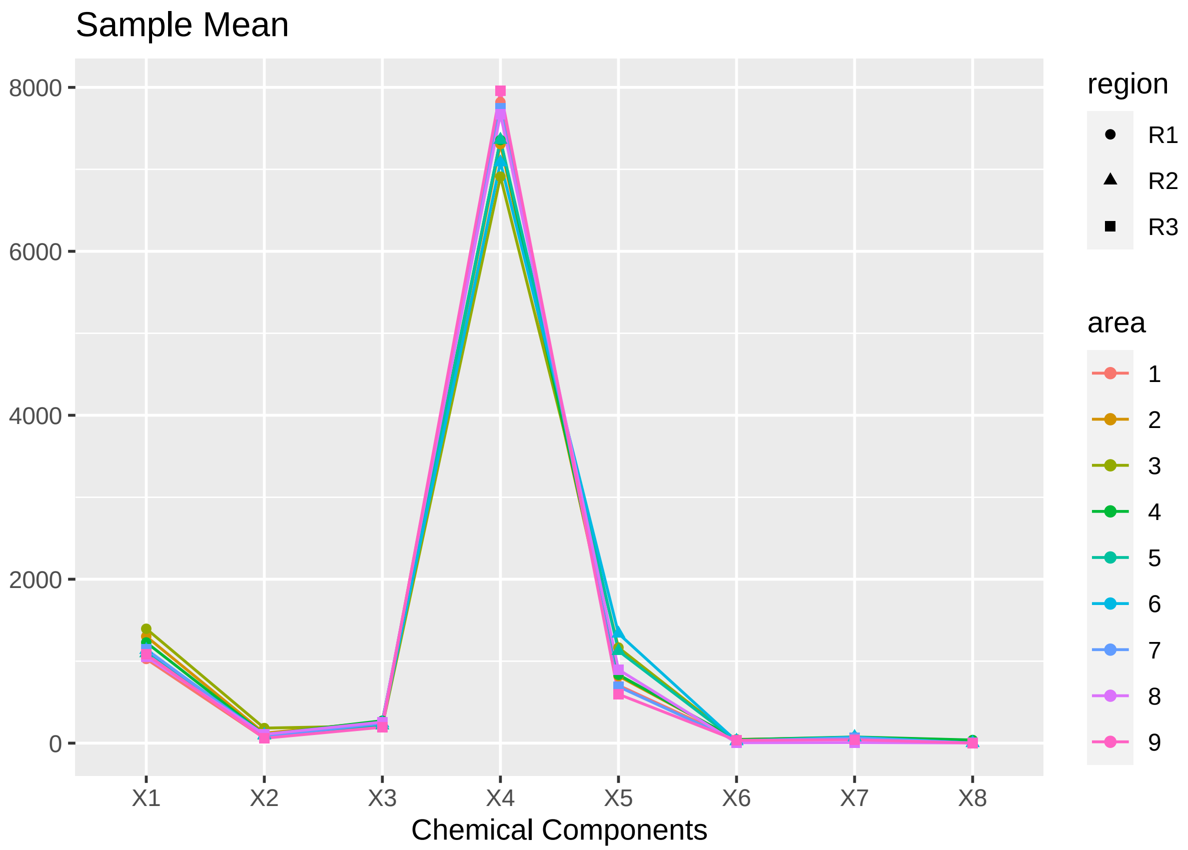
The Olive Oils dataset `olive.dat` (provided in the folder `Data and code` in Canvas) consists of measurements of eight chemical components (columns 2-8) on 572 samples of olive oil. The samples come from three different regions of Italy. The regions are further partitioned into nine areas: areas 1–4 belong to region R1, areas 5 and 6 belong to region R2, and areas 7–9 belong to region R3. The first column of the file provides the indicator for the nine regions. Now, we study the whole data set.

a. Provide appropriate graphical summaries for the dataset. You may plot the sample means of all nine areas (using different color in `geom_line`) in one figure, using x-axis as the indicator for chemical components, and also use different line styles (or different shapes of points in `geom_point`) for the three sub-regions to better understand the dataset. Comment on the distinctiveness between the chemical composition of the olive oils in the three main regions, as well as individually between the sub-regions. Can you see interactions between regions and chemical components? [15 points]

```r
library(tidyverse)

olive <- read.table("olive.dat", header = T) %>%
  rename(area = group.id) %>%
  mutate(region = c(rep("R1", 4), rep("R2", 2), rep("R3", 3))[area]) %>%
  mutate_at(vars(area, region), as.factor) %>%
  select(region, everything())

olive %>% group_by(area, region) %>%
  summarise_all(mean) %>%
  reshape2::melt(c("area", "region")) %>%
  qplot(variable, value, group = area, color = area, shape = region, data = .,
    geom = c("line", "point"), main = "Sample Mean", xlab = "Chemical Components", ylab = "")
```



Weak interaction between regions and chemical components.

b. Perform a one-way multivariate analysis of variance to test for differences in mean chemical composition among the three regions. [10 points]

```
fit.lm <- olive %>% lm(as.matrix(.[, -(1:2)]) ~ region, data = .)
summary(car::Manova(fit.lm)) %>% print(SSP = F)
```

```
##
## Type II MANOVA Tests:
##
## ------------------------------------------
##
## Term: region
##
## Multivariate Tests: region
##                  Df test stat approx F num Df den Df     Pr(>F)
## Pillai            2  1.593690 276.0350     16   1126 < 2.22e-16 ***
## Wilks             2  0.031702 324.3008     16   1124 < 2.22e-16 ***
## Hotelling-Lawley  2 10.816547 379.2552     16   1122 < 2.22e-16 ***
## Roy               2  8.494086 597.7713      8    563 < 2.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Reject H_0 that the chemical components among 3 regions are the same.

c. Test for the interaction between regions and chemical components. What are the C and M matrices being used in this hypothesis? [10 points]

```r
C <- cbind(0, diag(2))
M <- rbind(1, -diag(7))
car::linearHypothesis(fit.lm, C, P = M) %>% print(SSP = F)
```

```
##
## Multivariate Tests:
##                  Df test stat approx F num Df den Df     Pr(>F)
## Pillai            2  1.567412 291.9375     14   1128 < 2.22e-16 ***
## Wilks             2  0.039526 324.1159     14   1126 < 2.22e-16 ***
## Hotelling-Lawley  2  8.944258 359.0481     14   1124 < 2.22e-16 ***
## Roy               2  6.627284 533.9698      7    564 < 2.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Reject H_0 that there is no interaction between regions and chemical components. $C = (\mathbf{0}_2, I_2), M = (\mathbf{1}_7, -I_7)'$.

d. Test for the equivalence of means of chemical components among the 4 areas within region R1. What are the C and M matrices being used in this hypothesis? [10 points]

```r
fit2.lm <- filter(olive, region == "R1") %>% lm(as.matrix(.[, -(1:2)]) ~ area, data = .)
C <- cbind(0, diag(3))
M <- diag(8)
car::linearHypothesis(fit2.lm, C, P = M) %>% print(SSP = F)  # OR summary(Manova(fit2.lm))
```

```
##
## Multivariate Tests:
##                  Df test stat  approx F num Df   den Df     Pr(>F)
## Pillai            3  1.561753  42.62048     24 942.0000 < 2.22e-16 ***
## Wilks             3  0.066872  58.14938     24 905.4961 < 2.22e-16 ***
## Hotelling-Lawley  3  5.905260  76.44031     24 932.0000 < 2.22e-16 ***
## Roy               3  4.476791 175.71405      8 314.0000 < 2.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Reject H_0 that means of chemical components among the 4 areas within region R1 are equivalent. $C = (\mathbf{0}_3, I_3), M = I_8$.

3. [35 points]

Consider the `fbiwide` data in `classdata` package. Use the code `devtools::install_github("heike/classdata")` to install this package if you haven't done so. Only focus on the states `California`, `Iowa`, `Illinois`, `District of Columbia` and `New York`.

a. Take the log transformation of the data first, and then take the difference between consecutive years (current year minus previous year). Why this transformation provides the change rate of the number of crimes? [5 points]

Solution: $\ln y_{i+1} - \ln y_i = \ln(y_{i+1}/y_i)$.

b. Use the transformed data to compare the change rates of different crimes. Summarize `Year` to three groups: from 1961 to 1980, from 1981 to 2000, from 2001 to 2017. Call this variable as `decade`. Perform a two-way Manova for the interaction between states and decade. [10 points]

Solution: Drop `Rape` and `Legacy.rape` to avoid `NA`.

```r
fbi <- classdata::fbiwide %>%
  filter(State %in% c("California", "Iowa", "Illinois", "District of Columbia", "New York")) %>%
```

```r
  mutate(Decade = case_when(Year %in% 1961:1980 ~ "D1", Year %in% 1981:2000 ~ "D2", Year %in% 2001:
  select(-Abb, -Year, -Population, -Rape, -Legacy.rape) %>%
  group_by(State) %>% mutate_if(is.numeric, ~ c(NA, diff(log(.)))) %>%
  select(Decade, everything()) %>% na.omit()

fit.lm <- fbi %>% lm(as.matrix(.[, -(1:2)]) ~ State * Decade, data = .)
summary(car::Manova(fit.lm))$multivariate.tests$`State:Decade` %>% print(SSP = F)
```

```
##
## Multivariate Tests: State:Decade
##                  Df test stat approx F num Df   den Df    Pr(>F)
## Pillai            8 0.1797433 1.003677     48 1560.000 0.4670361
## Wilks             8 0.8305393 1.008528     48 1258.769 0.4581355
## Hotelling-Lawley  8 0.1919253 1.012939     48 1520.000 0.4494719
## Roy               8 0.0876279 2.847906      8  260.000 0.0047533 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cannot reject H_0 that there is no interaction between states and decade.

c. Report and explain the sum of square matrices from question (b). Verify their sum is equal to the corrected total sum of squares matrix. [10 points]

Solution: It is true for Type I MANOVA (not Type II or III) because we have unbalanced dataset.

```r
options(width = 200, digits = 4)
summary(manova(fit.lm))$SS  ## Type I SS
```

```
$State
                   Aggravated.assault Burglary Larceny.theft Motor.vehicle.theft  Murder Robbery
Aggravated.assault            0.11218  0.08121       0.01178             0.07342 0.04100 0.05593
Burglary                      0.08121  0.07734       0.01917             0.07618 0.04651 0.05073
Larceny.theft                 0.01178  0.01917       0.01205             0.02750 0.01351 0.01841
Motor.vehicle.theft           0.07342  0.07618       0.02750             0.09637 0.04910 0.06811
Murder                        0.04100  0.04651       0.01351             0.04910 0.03068 0.03089
Robbery                       0.05593  0.05073       0.01841             0.06811 0.03089 0.05152

$Decade
                   Aggravated.assault Burglary Larceny.theft Motor.vehicle.theft Murder Robbery
Aggravated.assault             0.3275   0.4268        0.3104              0.3249 0.2658  0.4281
Burglary                       0.4268   0.6831        0.4797              0.4147 0.4301  0.6870
Larceny.theft                  0.3104   0.4797        0.3387              0.3027 0.3015  0.4822
Motor.vehicle.theft            0.3249   0.4147        0.3027              0.3230 0.2579  0.4158
Murder                         0.2658   0.4301        0.3015              0.2579 0.2710  0.4326
Robbery                        0.4281   0.6870        0.4822              0.4158 0.4326  0.6908

$`State:Decade`
                   Aggravated.assault   Burglary Larceny.theft Motor.vehicle.theft   Murder    Robbery
Aggravated.assault            0.11269 -0.0160823     0.0232321           0.0385608 -0.02150 -0.0326760
Burglary                     -0.01608  0.0121201    -0.0001228           0.0063452  0.01017  0.0144806
Larceny.theft                 0.02323 -0.0001228     0.0216143          -0.0088273 -0.04090 -0.0061280
Motor.vehicle.theft           0.03856  0.0063452    -0.0088273           0.0968783  0.06908  0.0003171
Murder                       -0.02150  0.0101662    -0.0408957           0.0690814  0.13397  0.0241327
Robbery                      -0.03268  0.0144806    -0.0061280           0.0003171  0.02413  0.0542654

$Residuals
                   Aggravated.assault Burglary Larceny.theft Motor.vehicle.theft Murder Robbery
Aggravated.assault             2.1973   0.3385        0.5800              0.7538 0.5705   1.044
Burglary                       0.3385   1.8377        0.7875              1.3793 1.0278   1.733
```

```
Larceny.theft                  0.5800   0.7875         1.2033              0.8494 0.6829   1.108
Motor.vehicle.theft            0.7538   1.3793         0.8494              3.1008 1.3428   1.910
Murder                         0.5705   1.0278         0.6829              1.3428 5.3683   2.119
Robbery                        1.0445   1.7328         1.1076              1.9101 2.1195   3.960
```

```r
var(fbi[, -(1:2)]) * (nrow(fbi) - 1)  # SSTotal
```

```
                   Aggravated.assault Burglary Larceny.theft Motor.vehicle.theft Murder Robbery
Aggravated.assault             2.7496   0.8305        0.9253               1.191 0.8558   1.496
Burglary                       0.8305   2.6103        1.2862               1.876 1.5146   2.485
Larceny.theft                  0.9253   1.2862        1.5756               1.171 0.9570   1.602
Motor.vehicle.theft            1.1907   1.8765        1.1708               3.617 1.7188   2.394
Murder                         0.8558   1.5146        0.9570               1.719 5.8039   2.607
Robbery                        1.4959   2.4849        1.6020               2.394 2.6071   4.756
```

d. I also want to study the differences of the change rates among `Aggravated.assault`, `Burglary` and `Larceny.theft`. This indicates whether all the crimes decrease or increase at the same rate. Use `Aggravated.assault - Burglary` and `Burglary - Larceny.theft`. Are there interaction effects between states and decades for the difference of the change rates among the three crimes? [10 points]

```r
options(digits = 7)
newfit.lm <- lm(cbind(Aggravated.assault - Burglary, Burglary - Larceny.theft) ~ State * Decade, da
summary(car::Manova(newfit.lm))$multivariate.tests$`State:Decade` %>% print(SSP = F)
```

```
##
## Multivariate Tests: State:Decade
##                  Df test stat approx F num Df den Df  Pr(>F)
## Pillai            8 0.0623951 1.046571     16    520 0.40514
## Wilks             8 0.9383765 1.046137     16    518 0.40561
## Hotelling-Lawley  8 0.0648481 1.045675     16    516 0.40610
## Roy               8 0.0475591 1.545672      8    260 0.14171
```

Cannot reject H_0 that there is no interactions between states and decades.