STAT 3690 Homework 2

zhiyanggeezhou.github.io

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca)

Due at Mar 16 11:59 pm (Central Time)

Answers must be submitted electronically via Crowdmark. Please enclose your R source code (if applicable) as well.

1. 10 subjects with bronchus cancer were enrolled in a clinical study. For each of them, two survival times (in week) were recorded: T_1 = survival time from the first hospital admission; T_2 = survival time from the beginning of nontreatability.
   a. The gap between the two survival times is often of interet, because it may reflect the progression of disease as well as the treatment effect. Write down the distribution of time gap $T_1 - T_2$, assuming $[T_1, T_2]^\top \sim MVN_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
   b. Let $\mathbf{T}_i = [T_{i1}, T_{i2}]^\top$, where T_{i1} (resp. T_{i2}) were the observation of T_1 (resp. T_2) for subject i, $i = 1, \ldots, 10$. Suppose $\mathbf{T}_1, \ldots, \mathbf{T}_{10} \overset{\text{iid}}{\sim} MVN_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = [\mu_1, \mu_2]^\top$. Present both the Bonferroni's and Scheffe's simultaneous 95% confidence intervals for μ_1, μ_2 and $\mu_2 - \mu_1$.

Subject	T_1	T_2
1	81	74
2	461	423
3	20	16
4	450	450
5	246	87
6	166	115
7	63	50
8	64	50
9	155	113
10	151	38

Answer to Q1a. $T_1 - T_2 \sim N([1,-1]\boldsymbol{\mu}, [1,-1]\boldsymbol{\Sigma}[1,-1]^\top)$, since $T_1 - T_2 = [1,-1][T_1, T_2]^\top$ and $[T_1, T_2]^\top \sim MVN_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Answer to Q1b.

```
##            [,1]    [,2]
## mu1      40.640 330.760
## mu2      -5.607 288.807
## mu2-mu1 -92.833   4.633
```

```
##            [,1]    [,2]
## mu1      29.06 342.338
## mu2     -17.36 300.556
## mu2-mu1 -96.72   8.522
```

2. Consider the `Wolves` dataset from the package `candisc`. The variable `sex` indicates the sex of wolves (`f=female`, `m=male`), while `location` encodes wolves' habitats (`ar=Arctic`, `rm=Rocky Mountain`). The combination of `location` and `sex` is exactly `group`. Variables `x1` to `x9` correspond to 9 different skull morphological measurements of wolves, respectively. **We will merely focus on six measurements `x4` to `x9`.**
   a. Perform an appropriate test to compare the mean skull measurements of male and female wolves. Is there any statistical evidence to claim that the skull morphology differs between males and females at 5% level? (**Hereafter**, in reporting results of hypothesis testing, don't forget to include your hypotheses, the name of method, the value of test statistic, and the rejection region/p-value, before coming to the conclusion.)
   b. What are the assumptions required to perform the test in part a?
   c. Repeat parts a and b for wolves only from the Arctic.
   d. Provide plausible explanations (both statistical and subject-matter) about any discrepancy between the analysis in parts a and c.
   e. For now we are not sure whether the covariance matrix of the six measurements vary with `sex`. Please confirm it via a hypothesis test at level $\alpha = .05$.

```
##      group location sex  x1  x2  x3   x4   x5   x6   x7   x8   x9
## rmm1  rm:m       rm   m 126 104 141 81.0 31.8 65.7 50.9 44.0 18.2
## rmm2  rm:m       rm   m 128 111 151 80.4 33.8 69.8 52.7 43.2 18.5
## rmm3  rm:m       rm   m 126 108 152 85.7 34.7 69.1 49.3 45.6 17.9
## rmm4  rm:m       rm   m 125 109 141 83.1 34.0 68.0 48.2 43.8 18.4
## rmm5  rm:m       rm   m 126 107 143 81.9 34.0 66.1 49.0 42.4 17.9
## rmm6  rm:m       rm   m 128 110 143 80.6 33.0 65.0 46.4 40.2 18.2
```

Answer to Q2a. Testing hypotheses H_0: identical skull morphology between male and female wolves v.s. H_1: otherwise, we carried on the Wilk's lambda test and obtained 0.3584 as the value of test statistic. The corresponding p-value was .002235 (resp. .002469) for Bartlett's (resp. Rao's) approximation. So, at the .05 level, there was statistical evidence against H_0, i.e., we rejected H_0 and believed that there is difference between males' and females' skull measurements.

Answer to Q2b. We assumed the independence across all the wolves. Within each group, the skull measurements were assumed to be identically distributed as a multivariate normal distribution. We also assumed the identical covariance matrix of these measurements between males and females.

Answer to Q2c. Testing hypotheses H_0: identical skull morphology between male and female arctic wolves v.s. H_1: otherwise, we carried on the Wilk's lambda test and obtained .4020 as the value of test statistic. The corresponding p-value was .1236 (resp. .1343) for Bartlett's (resp. Rao's) approximation. So, at the .05 level, there was no statistical evidence against H_0, i.e., we did not reject H_0 and believed that there is no difference in skull morphology between male and female arctic wolves.

The assumptions for this test are similar to those for Q2b, i.e., the independence across arctic wolves, the normality of their skull measurements, and the equal covariance matrix of these measurements between male and female arctic wolves.

Answer to Q2d. Maybe the arctic wolves show less sex difference than Rocky-mountain ones in terms of the six skull measurements, while the sex effect we saw in the full analysis (Q2a) was driven by the Rocky-mountain wolves. It could also be the sample size of the subgroup analysis (Q2c) which was too small to detect a reasonable effect of sex on the six skull measurements.

Answer to Q2e. Testing hypotheses H_0: the covariance matrix of skull measurements does not vary with sex v.s. H_1: otherwise, we carried on the Box's M test and obtained 27.87 as the value of test statistic. The corresponding p-value was .1439. So, at the .05 level, there was no strong statistical evidence against H_0, i.e., we did not reject H_0 and believed that the covariance matrix does not vary with sex.

3. There is a dataset presented by Dean De Cock (2011, *Journal of Statistics Education*, 19(3)). It describes the sale of individual residential property in Ames, Iowa, U.S. from 2006 to 2010, containing

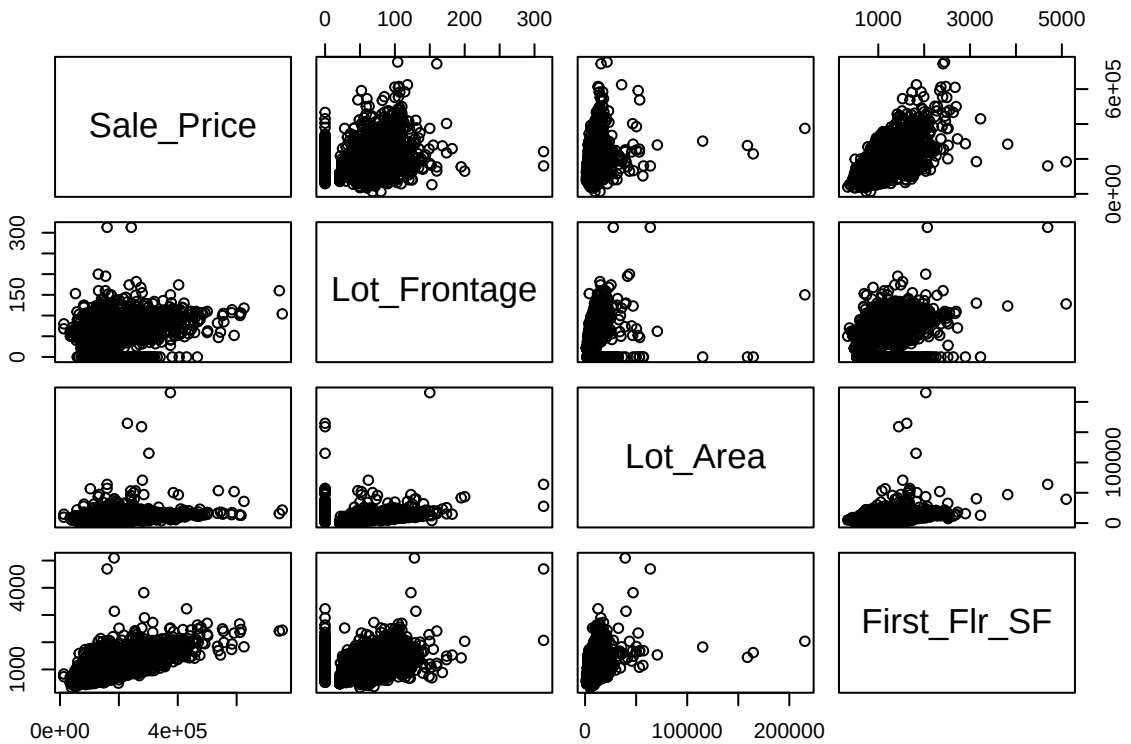
2930 observations and a large number of explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) for the assessment of home values. We will focus on the following five variables:

- `Sale_Price`: sale price (in USD);
- `Lot_Frontage`: linear feet of street connected to property (in feet);
- `Lot_Area`: lot size (in square feet);
- `First_Flr_SF`: first floor square feet;
- `Year_Sold`: year sold.

```
# The code below creates a clean version of the dataset.
# For more information, type ?AmesHousing::make_ames.
install.packages('AmesHousing')
library(AmesHousing)
ames_data <- make_ames()
```

- There are six subquestions.
  a. Create a pairs plot for variables `Sale_Price`, `Lot_Frontage`, `Lot_Area` and `First_Flr_SF`.
  b. Fit a linear regression model with variables `Sale_Price`, `Lot_Frontage`, `Lot_Area` and `First_Flr_SF` as outcomes and `Year_Sold` as the only explanatory variable. Carefully interpret the regression coefficient estimates.
  c. Test the multivariate regression model in part b against the empty model. Then test each of the four univariate regression models against the empty model. By comparing the multivariate result with the four univariate ones, what conclusions can you draw?
  d. Use the Cook's distance to identify the most influential observations for model in part b.
  e. Investigate the distribution of residuals and the overall model fit for model in part b.
  f. Based on your observations above, suggest ways of improving the model fit in part b.

Answer to Q3a.



Answer to Q3b. Since `Year_Sold` is never equal to zero, we can interpret regression coefficients other than intercepts: on average, for the individual residential property sold one year later, the average sale price was about 1855 lower in USD, the average linear feet of street connected to property was about .2986 less, the average lot size was about 138.2 smaller in square feet, and the average area of first floor was about 4.068 smaller in square feet.

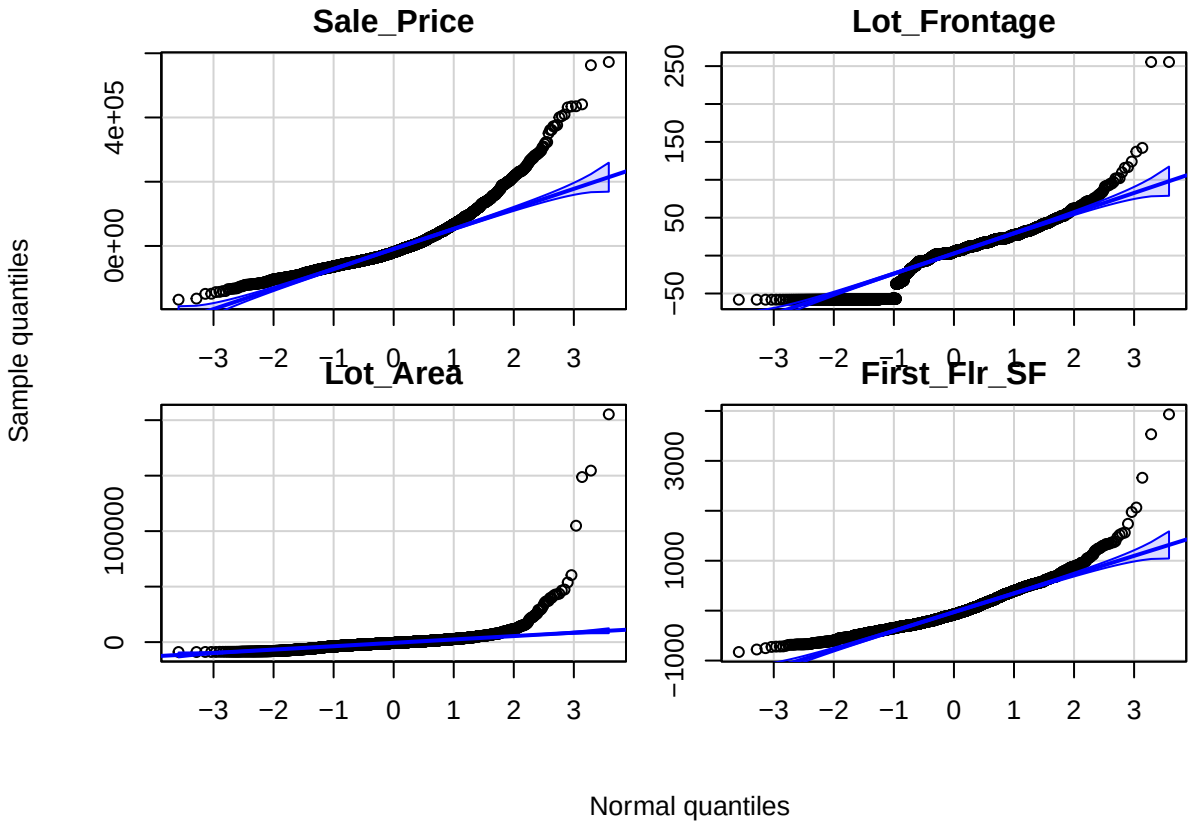
Answer to Q3c. We tested the following five pairs of hypotheses:

1. H0: `cbind(Sale_Price,Lot_Frontage,Lot_Area,First_Flr_SF)~1` vs. H1: `cbind(Sale_Price,Lot_Frontage,Lot_`
2. H0: `Sale_Price~1` vs. H1: `Sale_Price~Year_Sold`
3. H0: `Lot_Frontage~1` vs. H1: `Lot_Frontage~Year_Sold`
4. H0: `Lot_Area~1` vs. H1: `Lot_Area~Year_Sold`
5. H0: `First_Flr_SF~1` vs. H1: `First_Flr_SF~Year_Sold`.

We carried on the Wilk's lambda test for the first pair and LRT for the remaining four. The corresponding p-value were .4252, .09795, .5253, .2115, and .4595 So, at the .05 level, there was no statistical evidence against the null hypothesis in the five tests, i.e., `Year_Sold` is insignificant in all the five linear regression models. (P.S. sometimes, an explanatory variable insignificant for all the univariate models may still be significant for the multivariate model because the multivariate model employs extra information, i.e., the correlation between outcome variables.)

Answer to Q3d. In terms of the Cook's distance, the most influential observation is No. 957.

Answer to Q3e. As seen in the following normal Q-Q plots for residuals, residuals cannot be considered as normally distributed.



Answer to Q3f. Add explanatory variables that explain more variation (since `Year_Sold` is not significant) and transform all the four outcome variables (since residuals are not normally distributed). If the resulting linear model is still not satisfactory, one may remove outliers (including influential observations and outliers

in figures for Q1a and Q3e) and even consider a non-linear model.

Appendix

```
options(digits = 4)
## Q1b
dataset = data.frame(
  T1 = c(81, 461, 20, 450, 246, 166, 63, 64, 155, 151),
  T2 = c(74, 423, 16, 450, 87, 115, 50, 50, 113, 38)
)
n = nrow(dataset); p = ncol(dataset)
mu_hat <- colMeans(dataset)
sample_cov <- cov(dataset)

alpha <- .05
a1 = c(1,0); a2 = c(0,1); a3 = c(-1,1)
A = rbind(a1, a2, a3)
row.names(A)=c('mu1','mu2','mu2-mu1')

# Bonferroni's simultaneous CIs
m = nrow(A)
c = qt(1-alpha/2/m, n-1)
(Bonferroni <- cbind(
    A %*% mu_hat - c * sqrt(diag(A %*% sample_cov %*% t(A))/n),
    A %*% mu_hat + c * sqrt(diag(A %*% sample_cov %*% t(A))/n)
))

# Scheffe's simultaneous CIs
c = sqrt(p*(n-1)/(n-p) * qf(1-alpha, p, n-p))
(Scheffe <- cbind(
    A %*% mu_hat - c * sqrt(diag(A %*% sample_cov %*% t(A))/n),
    A %*% mu_hat + c * sqrt(diag(A %*% sample_cov %*% t(A))/n)
))

## Q2a
X <- as.matrix(candisc::Wolves[, c('x4', 'x5', 'x6', 'x7', 'x8', 'x9')])
sex = factor(candisc::Wolves$sex, levels = c('m', 'f'))

# Wilk's lambda test (Bartlett's approximation to the null distribution)
X_m <- X[sex == 'm',]
X_f <- X[sex == 'f',]
n <- nrow(X); p <- ncol(X); m <- 2
SSPcor = (n-1)*cov(X)
SSPw <- (nrow(X_m) - 1)*cov(X_m) + (nrow(X_f) - 1)*cov(X_f)
(Lambda <- det(SSPw)/det(SSPcor))
(cri.point = exp(qchisq(0.95, p*(m-1))/((p+m)/2-n+1)))
Lambda <= cri.point
(p.val = 1-pchisq(((p+m)/2-n+1)*log(Lambda), p*(m-1)))

# Wilk's lambda test (Rao's approximation to the null distribution)
summary(manova(X ~ sex), test = 'Wilks')
summary(car::Manova(lm(X ~ sex)), test.statistic='Wilks')

## Q2c
X <- as.matrix(candisc::Wolves[candisc::Wolves$location=='ar', c('x4', 'x5', 'x6', 'x7', 'x8', 'x9')])
```

```r
sex = factor(candisc::Wolves$sex[candisc::Wolves$location=='ar'], levels = c('m', 'f'))

# Wilk's lambda test (Bartlett's approximation to the null distribution)
X_m <- X[sex == 'm',]
X_f <- X[sex == 'f',]
n <- nrow(X); p <- ncol(X); m <- 2
SSPcor = (n-1)*cov(X)
SSPw <- (nrow(X_m) - 1)*cov(X_m) + (nrow(X_f) - 1)*cov(X_f)
(Lambda <- det(SSPw)/det(SSPcor))
(cri.point = exp(qchisq(0.95, p*(m-1))/((p+m)/2-n+1)))
Lambda <= cri.point
(p.val = 1-pchisq(((p+m)/2-n+1)*log(Lambda), p*(m-1)))

# Wilk's lambda test (Rao's approximation to the null distribution)
summary(manova(X ~ sex), test = 'Wilks')
summary(car::Manova(lm(X ~ sex)), test.statistic='Wilks')

## Q2e
X <- as.matrix(candisc::Wolves[, c('x4', 'x5', 'x6', 'x7', 'x8', 'x9')])
sex = factor(candisc::Wolves$sex, levels = c('m', 'f'))
heplots::boxM(lm(X ~ sex))

## Q3a
ames_data <- AmesHousing::make_ames()
Y = ames_data[,c('Sale_Price', 'Lot_Frontage', 'Lot_Area', 'First_Flr_SF')]
pairs(Y)

## Q3b
fit1 = lm(cbind(Sale_Price, Lot_Frontage, Lot_Area, First_Flr_SF)~Year_Sold, data=ames_data)
coef(fit1)

## Q3c
fit0 = lm(cbind(Sale_Price, Lot_Frontage, Lot_Area, First_Flr_SF)~1, data=ames_data)
anova(fit1, fit0, test='Wilks')
fit10 = lm(Sale_Price~1, data=ames_data); fit11 = lm(Sale_Price~Year_Sold, data=ames_data);
anova(fit11, fit10, test='LRT')
fit20 = lm(Lot_Frontage~1, data=ames_data); fit21 = lm(Lot_Frontage~Year_Sold, data=ames_data);
anova(fit21, fit20, test='LRT')
fit30 = lm(Lot_Area~1, data=ames_data); fit31 = lm(Lot_Area~Year_Sold, data=ames_data);
anova(fit31, fit30, test='LRT')
fit40 = lm(First_Flr_SF~1, data=ames_data); fit41 = lm(First_Flr_SF~Year_Sold, data=ames_data);
anova(fit41, fit40, test='LRT')

## Q3d
resids <- residuals(fit1)
X <- model.matrix(fit1)
H <- X %*% solve(crossprod(X)) %*% t(X)
Hii = diag(H)
SigmaHatLS <- crossprod(resids)/(n - ncol(X))
cook_values <- Hii/((1 - Hii)^2*ncol(X)) * diag(resids %*% solve(SigmaHatLS) %*% t(resids))
which(cook_values==max(cook_values))

## Q3e
```

```r
name = colnames(resids)
op <- par(mfrow = c(2,2),
          oma = c(5,4,0,0),
          mar = c(1,1,2,2))
for (i in 1:ncol(resids)){
  car::qqPlot(resids[,i], main = name[i], id = F)
}
title(xlab = "Normal quantiles",
      ylab = "Sample quantiles",
      outer = TRUE, line = 3)
par(op)
```