Regression Analysis HW3

Tuorui Peng, tuoruipeng2028@u.northwestern.edu

Notation: I use abbr SMW for Sherman-Morrison-Woodbury Formula, which we proved in the first homework.

$$(A+UCV')^{-1} = A^{-1} - A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1}$$

Question 1.1

$\hat{\beta}_{n+1}$ is obtained by

$$\begin{aligned}
0 =& \frac{\partial}{\partial\beta}(Y-X\beta)'(Y-X\beta) + (y_{n+1}-x_{n+1}'\beta)^2\\
=& -2X'(Y-X\beta) + 2(y_{n+1}-x_{n+1}'\beta)(-x_{n+1})\\
\Rightarrow& (X'X + x_{n+1}x_{n+1}')\beta = X'Y + x_{n+1}y_{n+1}\\
\Rightarrow& \beta = (X'X+x_{n+1}x_{n+1}')^{-1}(X'Y+x_{n+1}y_{n+1})\\
\underset{\text{SMW}}{=}& (I - \frac{(X'X)^{-1}x_{n+1}x_{n+1}'}{1+x_{n+1}'(X'X)^{-1}x_{n+1}})(X'X)^{-1}(X'Y+x_{n+1}y_{n+1})\\
=& (I - \frac{(X'X)^{-1}x_{n+1}x_{n+1}'}{1+x_{n+1}'(X'X)^{-1}x_{n+1}})(\hat{\beta}_n - (X'X)^{-1}x_{n+1}y_{n+1}')
\end{aligned}$$

In which the computation cost is estimated as:

- $(X'X)^{-1}x_{n+1}$: $O(d^2)$
- $(X'X)^{-1}x_{n+1}x_{n+1}'$: $O(d)$
- $x_{n+1}'(X'X)^{-1}x_{n+1}$: $O(d)$
- $(I - \frac{(X'X)^{-1}x_{n+1}x_{n+1}'}{1+x_{n+1}'(X'X)^{-1}x_{n+1}})(\hat{\beta}_n - (X'X)^{-1}x_{n+1}y_{n+1}')$: $O(2d^2)$

In total: $\sim O(3d^2+2d)$

Question 1.2

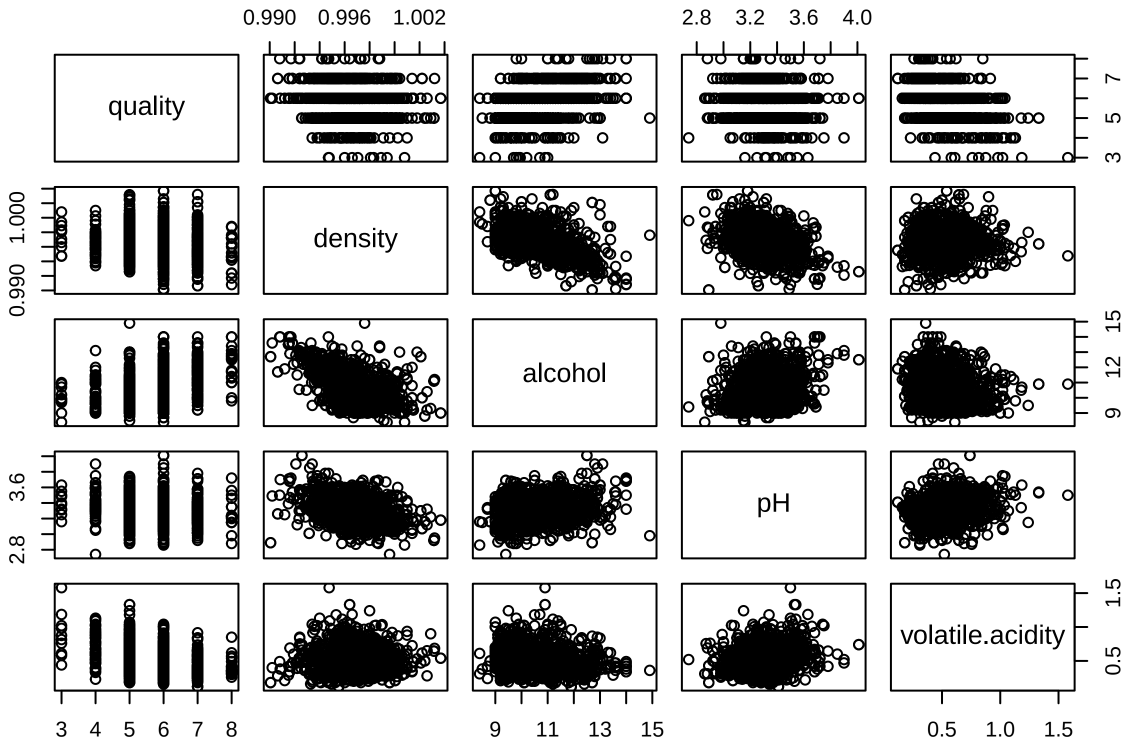
Here I solve the sub-problem (b) directly, in which we just need to check the positive semi-definiteness of matrix $C_\rho \in \mathbb{R}^{n\times n}$, which is equivalent to check the sign of its determinants of $C_\rho \in \mathbb{R}^{m\times m}$, $\forall m \leq n$:

$$
\begin{aligned}
\det \underset{m\times m}{C_\rho} &= \det \begin{bmatrix} 1 & -\rho & \cdots & -\rho \\ -\rho & 1 & \cdots & -\rho \\ \vdots & \vdots & \ddots & \vdots \\ -\rho & -\rho & \cdots & 1 \end{bmatrix}_{m\times m} \\
&= \det \begin{bmatrix} 1 & -\rho & \cdots & -\rho \\ -1-\rho & 1+\rho & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -1-\rho & 0 & \cdots & 1+\rho \end{bmatrix}_{m\times m} \\
&= \det \begin{bmatrix} 1-(m-1)\rho & -\rho & \cdots & -\rho \\ 0 & 1+\rho & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1+\rho \end{bmatrix}_{m\times m} \\
&= (1-(m-1)\rho)\prod_{i=2}^{m}(1+\rho) = (1-(m-1)\rho)(1+\rho)^{m-1}
\end{aligned}
$$

We have positive semi-definiteness of $\underset{n\times n}{C_\rho}$ iff $1-(m-1)\rho \geq 0$, $\forall m \leq n$ iff $\rho \leq \dfrac{1}{n-1}$.

Question 1.3

```
wine <- read.csv("winequality-red.csv", sep = ";")
pairs(wine[, c("quality", "density", "alcohol", "pH", "volatile.acidity")])
```



We have a strong positive relation between `quality` and `alcohol`, and a strong negative relation between `density` and `alcohol`.

Question 1.5

```
sfo <- read.csv("simplified-sfo-weather.csv", sep = ',')
```

(a)

Verify directly we have

$$\begin{aligned} cov(Y-\hat{Y}, Y_{\text{new}}-\hat{Y}_{\text{new}}) =& cov(X\beta+\varepsilon - X(X'X)^{-1}X(X\beta+\varepsilon), Z\beta+\varepsilon_{\text{new}} - Z(X'X)^{-1}X'(X\beta+\varepsilon)) \\ =& cov((I-X(X'X)^{-1}X')\varepsilon, \varepsilon_{\text{new}} - Z(X'X)^{-1}X'\varepsilon) \\ =& cov((I-X(X'X)^{-1}X')\varepsilon, -Z(X'X)^{-1}X'\varepsilon) \\ =& -(I-X(X'X)^{-1}X')\sigma^2 IX(X'X)^{-1}Z' = 0 \end{aligned}$$

(b)

We have

$$\begin{aligned}
Y_{\text{new}} - \hat{Y}_{\text{new}} =& Z\beta + \varepsilon_{\text{new}} - Z(X'X)^{-1}X'(X\beta + \varepsilon) \\
=& \varepsilon_{\text{new}} - Z(X'X)^{-1}X'\varepsilon \\
\sim& N(0, \sigma^2(I - Z(X'X)^{-1}Z'))
\end{aligned}$$

So the matrix M s.t. $M(Y_{\text{new}} - \hat{Y}_{\text{new}}) \sim N(0, \sigma^2 I)$ can be chosen as

$$\begin{aligned}
M =& (I - Z(X'X)^{-1}Z')^{-1/2} \\
\underset{\text{SMW}}{=}& \left(I - Z(X'X + Z'Z)^{-1}Z'\right)^{1/2}
\end{aligned}$$

(c)

Since we have proven the normality and the independence between $Y - \hat{Y}$ and $Y_{\text{new}} - \hat{Y}_{\text{new}}$ (which is equivalent to covariance being zero for normal variables), we have

$$A = \frac{\|M(Y_{\text{new}} - \hat{Y}_{\text{new}})\|_2^2 / n}{\|Y - \hat{Y}\|_2^2 / (m-d)} \sim F_{n,m-d}$$

(d)

```r
library('tidyverse')
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
mat_inverse_sqrt <- function(mat){
    a <- eigen(mat)
    idx <- which(a$value > 1e-8)
    return(a$vector[, idx] %*% diag(1 / sqrt(a$value[idx])) %*% t(a$vector[, idx]))
}

f_stat <- function(X,Z,Y,Y_new){
    X <- as.matrix(X)
```

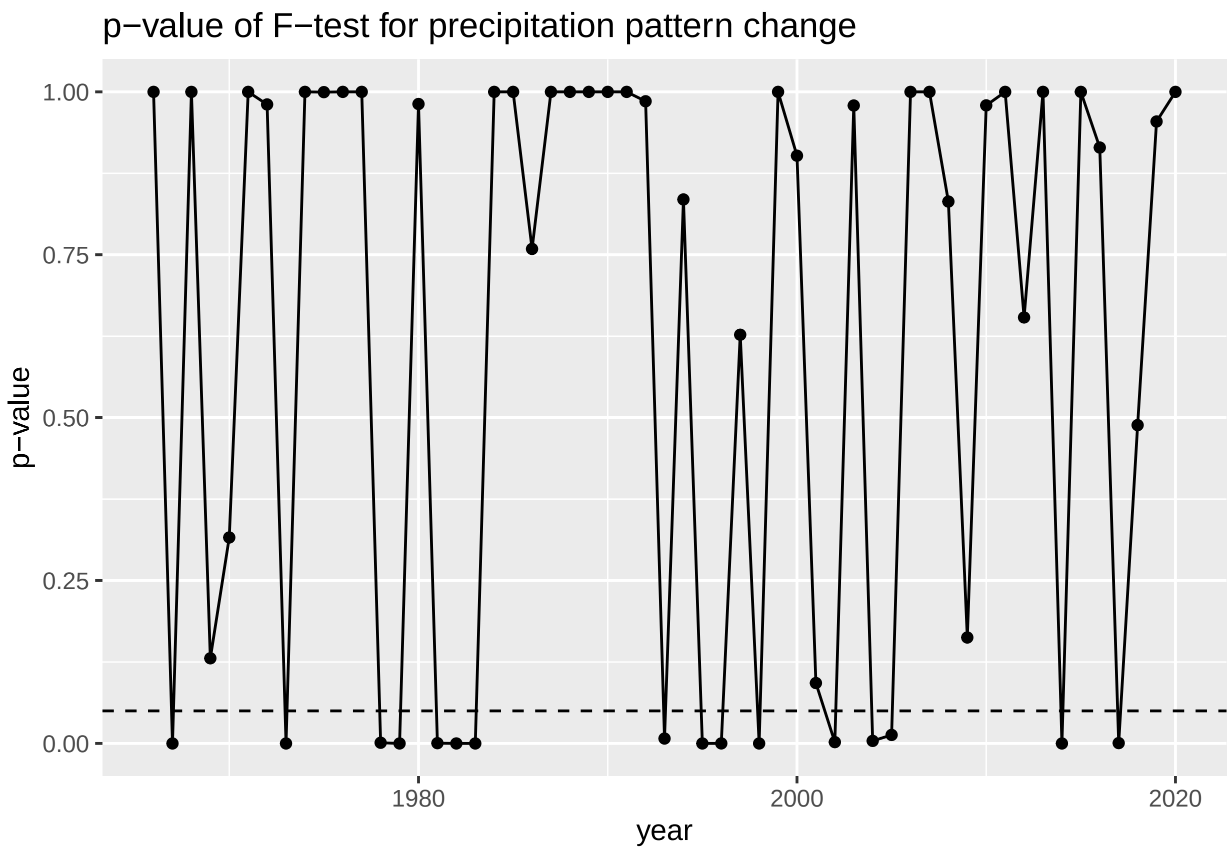
```r
    Z <- as.matrix(Z)
    m <- nrow(X)
    n <- nrow(Z)
    d <- ncol(X)
    beta <- solve(t(X)%*%X)%*%t(X)%*%Y
    Y_hat <- X%*%beta
    Y_new_hat <- Z%*%beta
    eps <- Y-Y_hat
    eps_new <- Y_new-Y_new_hat
    M <- mat_inverse_sqrt(diag(n)-Z%*%solve(t(X)%*%X) %*% t(Z))
    A <- (sum((M%*%eps_new)^2) / n) / (sum(eps^2) / (m-d))
    return(A)
}
time_to_X <- function(date_seq){
    df <- data.frame(interc = rep(1, length(date_seq)), sincomp = sin(2*pi*date_seq/365.25), coscomp
    return(df)
}

years <- c(1966:2020)
p_values <- c()
for(year in years){
    X <- time_to_X(sfo$day[sfo$year < year])
    Z <- time_to_X(sfo$day[sfo$year == year])
    Y <- sfo$precip[sfo$year < year]
    Y_new <- sfo$precip[sfo$year == year]
    dof1 <- nrow(Z)
    dof2 <- nrow(X)-ncol(X)
    p_values <- c(p_values, 1-pf(f_stat(X,Z,Y,Y_new), dof1, dof2))
}
names(p_values) <- years
p_values
```

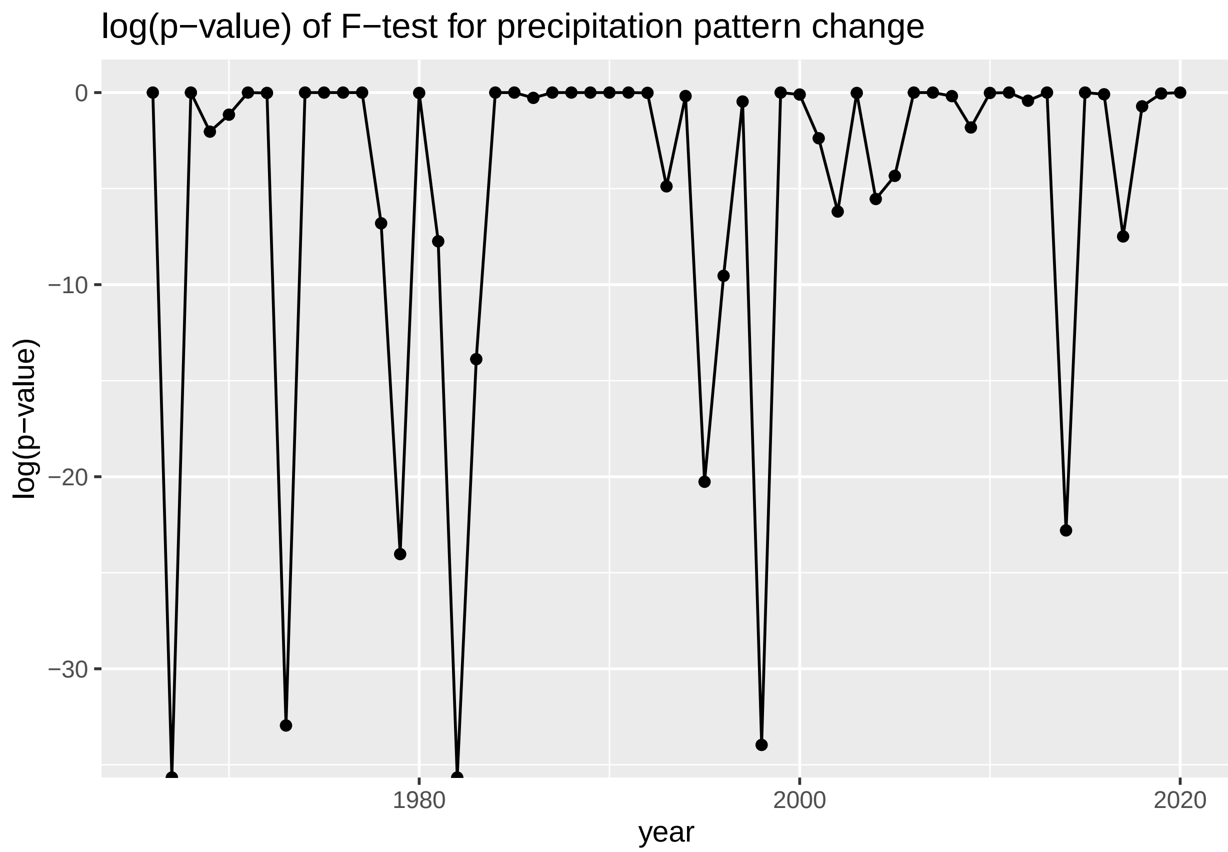
```
##         1966         1967         1968         1969         1970         1971
## 1.000000e+00 0.000000e+00 9.999818e-01 1.307204e-01 3.161205e-01 1.000000e+00
##         1972         1973         1974         1975         1976         1977
## 9.807176e-01 4.884981e-15 1.000000e+00 9.995939e-01 1.000000e+00 9.999999e-01
##         1978         1979         1980         1981         1982         1983
## 1.106125e-03 3.661982e-11 9.813982e-01 4.332939e-04 0.000000e+00 9.375169e-07
##         1984         1985         1986         1987         1988         1989
## 1.000000e+00 1.000000e+00 7.589744e-01 1.000000e+00 1.000000e+00 1.000000e+00
##         1990         1991         1992         1993         1994         1995
```

```
## 1.000000e+00 9.999999e-01 9.855533e-01 7.573062e-03 8.348714e-01 1.583855e-09
##         1996         1997         1998         1999         2000         2001
## 7.187248e-05 6.272902e-01 1.776357e-15 1.000000e+00 9.020412e-01 9.274056e-02
##         2002         2003         2004         2005         2006         2007
## 2.040332e-03 9.790616e-01 3.924335e-03 1.308682e-02 1.000000e+00 1.000000e+00
##         2008         2009         2010         2011         2012         2013
## 8.316644e-01 1.626295e-01 9.792580e-01 1.000000e+00 6.538253e-01 1.000000e+00
##         2014         2015         2016         2017         2018         2019
## 1.257557e-10 1.000000e+00 9.145681e-01 5.582391e-04 4.885333e-01 9.545072e-01
##         2020
## 1.000000e+00
```

```
# plot of p-value v.s. year
ggplot(data.frame(years = years, p_values = (p_values)), aes(x = years, y = p_values)) + geom_line()
```




```
# plot of log(p-value) v.s. year
ggplot(data.frame(years = years, p_values = log(p_values)), aes(x = years, y = p_values)) + geom_line
```



In the above we plotted log(p-value) v.s. year plot and p-value v.s. year plot. Seems in most years we don't reject the null hypothesis that there's no change in the precipitation pattern. But there are some years we observe significant low p-value, suggesting a rejection to null hypothesis.

(e)

I would say that 'changing over time' should be some kind of smooth, structural change, in which sense we should observe a long-range low p-value in the p-value v.s. year plot. But in the above plot we don't observe such a long-range low p-value. Actually we can see that in most of the years the p-value is nearly one, suggesting no change in the precipitation pattern. So I would say these 'outlier' p-values might just due to some occassional, short-term incidents happening in those years, instead of a long-term 'change in the precipitation pattern'.