HW4_Question6

Wookyeong Song (most of them from Yan-Yu Chen)

2022/10/27

Multiple regression by matrix algebra in R

You need to submit your codes alongside with the answers, plots, outputs, etc. You are required to use R Markdown: Please submit a .rmd file **and** its corresponding .html file.

Consider the following data set with 5 cases, one response variable Y and two predictor variables X_1, X_2.

case	Y	X1	X2
1	-0.97	-0.63	-0.82
2	2.51	0.18	0.49
3	-0.19	-0.84	0.74
4	6.53	1.60	0.58
5	1.00	0.33	-0.31

Consider the first-order model for the following questions:

(a)

Create the design matrix $\mathbf{X}$ and the response vector $\mathbf{Y}$. Calculate $\mathbf{X}'\mathbf{X}$, $\mathbf{X}'\mathbf{Y}$ and $(\mathbf{X}'\mathbf{X})^{-1}$.

```
X<-cbind(rep(1,5), matrix(c(-0.63,0.18,-0.84,1.60,0.33,
                            -0.82,0.49,0.74,0.58,-0.31), ncol=2))
Y<-c(-0.97,2.51,-0.19,6.53,1.00)
t(X)%*%X
```

```
##      [,1]   [,2]   [,3]
## [1,] 5.00 0.6400 0.6800
## [2,] 0.64 3.8038 0.8089
## [3,] 0.68 0.8089 1.8926
```

```
t(X)%*%Y
```

```
##         [,1]
## [1,]  8.8800
## [2,] 12.0005
## [3,]  5.3621
```

```r
solve(t(X)%*%X)
```

```
##             [,1]        [,2]        [,3]
## [1,]  0.21184719 -0.02140278 -0.06696786
## [2,] -0.02140278  0.29134054 -0.11682948
## [3,] -0.06696786 -0.11682948  0.60236791
```

(b)

Obtain the least-squares estimators $\hat{\beta}$.

```r
solve(t(X)%*%X, t(X)%*%Y)
```

```
##          [,1]
## [1,] 1.265271
## [2,] 2.679724
## [3,] 1.233270
```

(c)

Obtain the hat matrix $\mathbf{H}$. What are rank($\mathbf{H}$) and rank($\mathbf{I} - \mathbf{H}$)?

(Hint: You may use `rankMatrix()` in library `Matrix`)

```r
H<-X%*%solve(t(X)%*%X)%*%t(X) # hat matrix
H
```

```
##             [,1]       [,2]        [,3]       [,4]        [,5]
## [1,]  0.74859901 0.02181768  0.01132102 -0.1770289  0.39529119
## [2,]  0.02181768 0.27197293  0.35049579  0.2534024  0.10231125
## [3,]  0.01132102 0.35049579  0.82936038 -0.1072487 -0.08392853
## [4,] -0.17702890 0.25340235 -0.10724866  0.7973084  0.23356681
## [5,]  0.39529119 0.10231125 -0.08392853  0.2335668  0.35275928
```

```r
Matrix::rankMatrix(H)[1]
```

```
## [1] 3
```

```r
Matrix::rankMatrix(diag(1,5)-H)[1]
```

```
## [1] 2
```

(d)

Calculate the trace of $\mathbf{H}$ and compare it with $rank(\mathbf{H})$ from part (c). What do you find?

```r
sum(diag(H))
```

```
## [1] 3
```

For projection matrix, the rank of matrix equals the trace of matrix.

(e)

Obtain the fitted values, the residuals, SSE and MSE. What should be the degrees of freedom of SSE?

```
Yhat<-as.vector(H%*%Y)
Yhat
```

```
## [1] -1.43423719  2.35192330 -0.07307774  6.26812586  1.76726576
```

```
residuals<-Y-Yhat
residuals
```

```
## [1]  0.4642372  0.1580767 -0.1169223  0.2618741 -0.7672658
```

```
SSE<-sum(residuals^2)
SSE
```

```
## [1] 0.91145
```

The degree of freedom of SSE should be $n-3=2$.

```
MSE<-SSE/3
MSE
```

```
## [1] 0.3038167
```

Consider the nonadditive model with interaction between X_1 and X_2 for the following questions:

(f)

Create the design matrix. Obtain the hat matrix $\mathbf{H}$. Find $rank(\mathbf{H})$ and $rank(\mathbf{I}-\mathbf{H})$. Compare the ranks with those from part (c), what do you observe?

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i, \quad i = 1, \cdots, 5$$

$$\beta = [\beta_0, \beta_1, \beta_2, \beta_3]'$$

```
X<-cbind(X,X[,2]*X[,3])
X
```

```
##      [,1]  [,2]  [,3]    [,4]
## [1,]    1 -0.63 -0.82  0.5166
## [2,]    1  0.18  0.49  0.0882
## [3,]    1 -0.84  0.74 -0.6216
## [4,]    1  1.60  0.58  0.9280
## [5,]    1  0.33 -0.31 -0.1023
```

```
H<-X%*%solve(t(X)%*%X)%*%t(X)
H
```

```
##               [,1]       [,2]        [,3]        [,4]         [,5]
## [1,]  0.995761996 0.05537216 -0.02643351 -0.02101565 -0.003685001
## [2,]  0.055372163 0.27652824  0.34537029  0.27458248  0.048146828
## [3,] -0.026433513 0.34537029  0.83512745 -0.13107993 -0.022984289
## [4,] -0.021015645 0.27458248 -0.13107993  0.89578648 -0.018273381
## [5,] -0.003685001 0.04814683 -0.02298429 -0.01827338  0.996795843
```

```
Matrix::rankMatrix(H)[1]
```

```
## [1] 4
```

```
Matrix::rankMatrix(diag(1,5)-H)[1]
```

```
## [1] 1
```

We observe that rank($\mathbf{H}$) is larger than in (d) by 1.

(g)

Obtain the least-squares estimators $\hat{\boldsymbol{\beta}}$.

```
solve(t(X)%*%X, t(X)%*%Y)
```

```
##          [,1]
## [1,] 1.051738
## [2,] 1.987286
## [3,] 1.804233
## [4,] 1.387774
```

(h)

Obtain the fitted values, the residuals, SSE and MSE. What should be the degrees of freedom of SSE?

```
Yhat<-as.vector(H%*%Y)
Yhat
```

```
## [1] -0.9627998  2.4159250 -0.1450905  6.5657047  1.0062607
```

```
residuals<-Y-Yhat
residuals
```

```
## [1] -0.007200196  0.094075045 -0.044909459 -0.035704724 -0.006260666
```

```
SSE<-sum(residuals^2)
SSE
```

```
## [1] 0.01223284
```

The degree of freedom of SSE should be $n-4=1$. Therefore, $MSE=SSE=0.012$.

(i)

Which of the two models appears to fit the data better?

The second model fits the data better since it has a much smaller SSE therefore much larger R^2 ($R^2 = 1 - SSE/SSTO$ and $SSTO$ is the same).