STA206_ASS3

Zhen Zhang

October 14, 2015

(a) Model Equations: $y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \epsilon$

Coefficient Vector: $\beta = [\beta_0, \beta_1, \beta_2]$

Design Matrix: $X = [\mathbf{1}, X_1, X_2]$

Response Vector: y

(b) First read data into R

```
X1 <- c(-0.63, 0.18, -0.84, 1.6, 0.33)
X2 <- c(-0.82, 0.49, 0.74, 0.58, -0.31)
Y <- c(-0.97, 2.51, -0.19, 6.53, 1)
```

Then create design matrix **X** and the response vector **Y**

```
(X <- matrix(c(rep(1, 5), X1, X2), 5, 3))
```

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

Now calculate $\mathbf{X'X}$, $\mathbf{X'Y}$ and $(\mathbf{X'X})^{-1}$

```
# XX value
(XX <- 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
```

```
# XY value
(XY <- t(X) %*% Y)
```

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

```r
# XX Inverse
(XXInv <- solve(XX))
```

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

(c) The least square estimator $\hat{\beta}$ is

```r
(beta <- XXInv %*% XY)
```

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

(d) The hat matrix $\mathbf{H}$ is

```r
(H <- X %*% XXInv %*% t(X))
```

```
##             [,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
```

And the rank is

```r
rankMatrix(H)
```

```
## [1] 3
## attr(,"method")
## [1] "tolNorm2"
## attr(,"useGrad")
## [1] FALSE
## attr(,"tol")
## [1] 1.110223e-15
```

```r
rankMatrix(diag(5) - H)
```

```
## [1] 2
## attr(,"method")
## [1] "tolNorm2"
## attr(,"useGrad")
## [1] FALSE
## attr(,"tol")
## [1] 1.110223e-15
```

(e)

```r
(Y_fit <- X %*% beta)
```

```
##             [,1]
## [1,] -1.43423719
## [2,]  2.35192330
## [3,] -0.07307774
## [4,]  6.26812586
## [5,]  1.76726576
```

```r
(e <- Y - Y_fit)
```

```
##            [,1]
## [1,]  0.4642372
## [2,]  0.1580767
## [3,] -0.1169223
## [4,]  0.2618741
## [5,] -0.7672658
```

```r
(SSE <- t(e) %*% e)
```

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

Since SSE has degree of $5 - 2 - 1 = 2$, so

```r
(MSE <- SSE/2)
```

```
##          [,1]
## [1,] 0.455725
```

(f) Model Equations: $y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_1 X_2 + \epsilon$

Coefficient Vector: $\beta = [\beta_0, \beta_1, \beta_2, \beta_3]$

(g) First create the new design matrix X_{Int}

```r
(X_Int <- matrix(c(rep(1, 5), X1, X2, X1 * X2), 5, 4))
```

```
##      [,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
```

The response vector, Y_{Int}, is unchanged:

```r
(Y_Int <- Y)
```

```
## [1] -0.97  2.51 -0.19  6.53  1.00
```

Hat Matrix H_{Int}

```r
(H_Int <- X_Int %*% solve(t(X_Int) %*% X_Int) %*% t(X_Int))
```

```
##              [,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
```

Now calculate the rank:

```r
rankMatrix(H_Int)
```

```
## [1] 4
## attr(,"method")
## [1] "tolNorm2"
## attr(,"useGrad")
## [1] FALSE
## attr(,"tol")
## [1] 1.110223e-15
```

```r
rankMatrix(diag(5) - H_Int)
```

```
## [1] 1
## attr(,"method")
## [1] "tolNorm2"
## attr(,"useGrad")
## [1] FALSE
## attr(,"tol")
## [1] 1.110223e-15
```

The rank of H is now 4, which is the number of β. Also, it is one larger than the H in the previous part, since the number of variables adds one, which is the interaction term.

(h) The β_{Int} here is

```r
beta_Int <- solve((t(X_Int) %*% X_Int)) %*% (t(X_Int) %*% Y_Int)
```

(i) Fitted values:

```r
(Y_fit_Int <- X_Int %*% beta_Int)
```

```
##             [,1]
## [1,] -0.9627998
## [2,]  2.4159250
## [3,] -0.1450905
## [4,]  6.5657047
## [5,]  1.0062607
```

```r
(e_Int <- Y_Int - Y_fit_Int)
```

```
##               [,1]
## [1,] -0.007200196
## [2,]  0.094075045
## [3,] -0.044909459
## [4,] -0.035704724
## [5,] -0.006260666
```

```r
(SSE_Int <- t(e_Int) %*% e_Int)
```

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

The degree of freedom for SSE here is $5 - 3 - 1 = 1$

```r
(MSE_Int <- SSE_Int/1)
```

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

(j) The second model (the model with interaction) fits the data better, since its MSE and SSE is smaller than the previous one.