Tutorial 4 Solution

1. *Matrix Approach to Linear Regression*

Consider the following simple linear relationship between response y and one input feature, x:

$$y \approx \beta_0 + \beta_1 x.$$

Given a data set of n points $(x_1, y_1), ..., (x_n, y_n)$, the model above is then

$$y_i \approx \beta_0 + \beta_1 x_i, \quad i = 1, ..., n. \quad (*)$$

To rewrite (*) in matrix form, we have

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$, then the right-hand side of (*) is $\mathbf{X}\boldsymbol{\beta}$.

The residual sum of squares, $RSS = \sum_{i=1}^{n} [\mathrm{y}_i - (\beta_0 + \beta_1 \mathrm{x}_i)]^2$ is actually equal to

$$\begin{aligned} &(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= [y_1 - (\beta_0 + \beta_1 x_1), y_2 - (\beta_0 + \beta_1 x_2), ..., y_n - (\beta_0 + \beta_1 x_n)] \begin{bmatrix} y_1 - (\beta_0 + \beta_1 x_1) \\ y_2 - (\beta_0 + \beta_1 x_2) \\ \vdots \\ y_n - (\beta_0 + \beta_1 x_n) \end{bmatrix} \end{aligned}$$

Minimizing $RSS = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ w.r.t. $\boldsymbol{\beta}$, we have $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, where A^{-1} is the inverse of the square matrix A.

(a) Consider data set `Colleges.txt`. Write a function in R **using the matrix approach** to perform a simple linear regression of percentage of applicants accepted (Acceptance) on the median combined math and verbal SAT score of students (SAT).

Compare the results with the answers in part (b) of Question 1.

(b) If data set of n points has two input features, x^1, x^2, by matrix approach, the estimate of coefficient is still $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

i. Specify matrix $\mathbf{y}$, $\mathbf{X}$ and $\boldsymbol{\beta}$.

ii. Use your function in part (a) to perform a multivariate linear regression of percentage of applicants accepted (`Acceptance`) on `SAT` and `Top.10p` - percentage of students in the top 10% of their high school graduating class.

Solution:

(a)
```
> dat= read.table("C:/Data/Colleges.txt",header =TRUE,sep= "\t")
> names(dat)
[1] "School"      "School_Type" "SAT"         "Acceptance"  "DPerStudent"
[6] "Top.10p"     "PerPhD"      "GradPer"
> head(dat)
```

```
      School School_Type  SAT Acceptance DPerStudent Top.10p PerPhD GradPer
1    Amherst    Lib Arts 1315         22       26636      85     81      93
2 Swarthmore    Lib Arts 1310         24       27487      78     93      88
3   Williams    Lib Arts 1336         28       23772      86     90      93
4    Bowdoin    Lib Arts 1300         24       25703      78     95      90
5  Wellesley    Lib Arts 1250         49       27879      76     91      86
6     Pomona    Lib Arts 1320         33       26668      79     98      80
```

```
> matrix <- function(x, y) {
+ beta <- solve(t(x )%*% x )%*% t(x )%*% y
+ return( beta )
+ }
> matrix( x = cbind (1,dat$SAT),y = dat$Acceptance )
           [,1]
[1,] 202.2677440
[2,]  -0.1300894
```

Compare the outputs with part (a) of Question 1, they are the same.

```
> lm(Acceptance ~SAT , data =dat )
Call:
lm(formula = Acceptance ~ SAT, data = dat)

Coefficients:
(Intercept)          SAT
   202.2677      -0.1301
```

(b) With two input features x^1 and x^2,

i. Matrix $\mathbf{y}$ doesn't change while $\mathbf{X}$ and $\boldsymbol{\beta}$ now are as below.

$$\mathbf{X} = \begin{bmatrix} 1 & x_1^1 & x_1^2 \\ 1 & x_2^1 & x_2^2 \\ \vdots & \vdots & \\ 1 & x_n^1 & x_n^2 \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix},$$

where the second column of $\mathbf{X}$ is the set of n observation of feature x^1 and the third column is the set of n observations of feature x^2.

ii.

```
> matrix( cbind (1, dat$SAT ,dat$Top.10p), dat$Acceptance )
             [,1]
[1,] 175.54421649
[2,]  -0.08478261
[3,]  -0.41029538
> # Compare outputs with lm()
> lm(Acceptance ~ SAT +Top.10p, data = dat )
Call:
lm(formula = Acceptance ~ SAT + Top.10p, data = dat)

Coefficients:
(Intercept)          SAT      Top.10p
  175.54422     -0.08478     -0.41030
>
```

2. A dataset on house selling price was randomly collected [1], `house_selling_prices_FL.csv`. It's our

[1] *Statistics: The Art and Science of Learning from Data*, 4th, Agresti, Franklin, Klingenberg

interest to model how y = selling price (dollar) is dependent on x = the size of the house (square feet). A simple linear regression model (y regress on x) was fitted, called Model 1.

The given data has another variable, NW, which specifies if a house is in the part of the town considered less desirable (NW = 0).

(a) Derive the correlation between x and y.

(b) Derive a scatter plot of y against x. Give your comments on the association of y and x.

(c) Derive R^2 of Model 1. Verify that $\sqrt{R^2} = |cor(y, x)|$. In which situation we can have $\sqrt{R^2} = cor(y, x)$?

(d) Form a model (called Model 2) which has two regressors (x and NW). Write down the equation of Model 2.

(e) Report the coefficient of variable NW in Model 2. Interpret it.

(f) Estimate the price of a house where its size is 4000 square feet and is located at the more desirable part of the town.

(g) Report the R^2 of Model 2. Interpret it.

Solution:

```
> house = read.csv("C:/Data/house_selling_prices_FL.csv")
> names(house) # names of columns
```

```
[1] "House"    "Taxes"    "Bedrooms" "Baths"    "Quadrant" "NW"       "price"
[8] "size"     "lot"
```

```
> dim(house) # 100 observations and 9 columns
```

```
[1] 100   9
```

```
> house$NW = as.factor(house$NW) # to declare that NW is categorical
> attach(house)
```

(a)
```
> cor(price, size)
```

```
[1] 0.7612621
```

(b) The plot is given below. It shows a quite strong positive association and quite linear between price and size of a house. This agrees with the correlation value of 0.76.

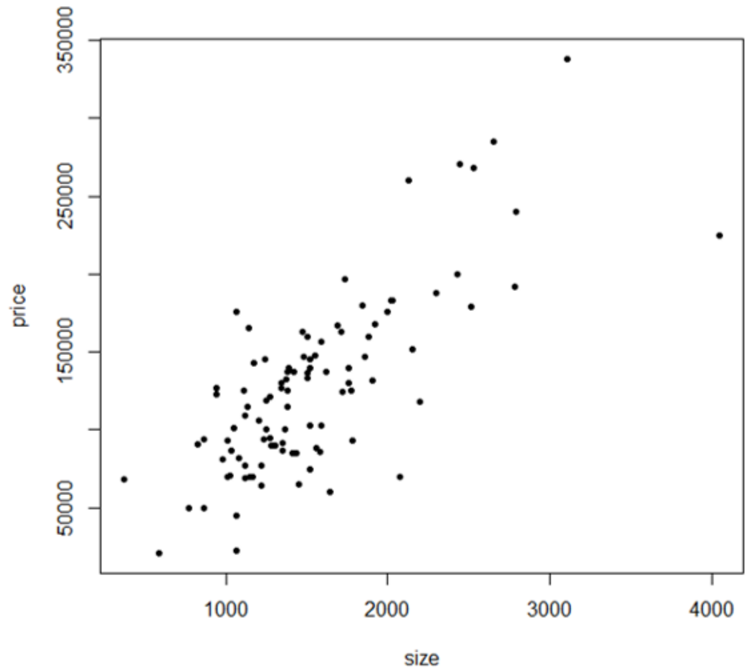
```
> plot(size, price, pch = 20)
```

(c)
```
> M1 = lm(price ~ size, data = house)
> summary(M1)
```

```
Call:
lm(formula = price ~ size, data = house)

Residuals:
   Min     1Q Median     3Q    Max 
-98567 -23582   2404  18843  89345 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9161.159  10759.786   0.851    0.397    
size           77.008      6.626  11.622   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




```
Residual standard error: 36730 on 98 degrees of freedom
Multiple R-squared:  0.5795,	Adjusted R-squared:  0.5752
F-statistic: 135.1 on 1 and 98 DF,  p-value: < 2.2e-16
```

For Model 1, $R^2 = 0.5795$. Indeed, $\sqrt{0.5795} = 0.761 = |cor(y, x)|$.

When $cor(y, x) > 0$ then in a simple model $y \sim x$, we always have $\sqrt{R^2} = cor(y, x)$.

(d) Form a model (called Model 2) which has two regressors (x and NW). Write down the equation of Model 2.

```
> M2 = lm(price ~ size + NW, data = house)
> summary(M2)

Call:
lm(formula = price ~ size + NW, data = house)

Residuals:
   Min     1Q Median     3Q    Max 
-83207 -22968    215  14135 109149 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -15257.514  11908.297  -1.281 0.203160    
size            77.985      6.209  12.560  < 2e-16 ***
NW1          30569.087   7948.742   3.846 0.000215 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34390 on 97 degrees of freedom
```

```
Multiple R-squared:  0.6352,        Adjusted R-squared:  0.6276
F-statistic: 84.43 on 2 and 97 DF,  p-value: < 2.2e-16
```

The fitted equation of Model 2:

$$\hat{y} = -15257.5 + 77.99x + 30569.1I(NW = 1).$$

(e) The estimated coefficient of NW in Model 2 is 30569.1. This value means: for two houses of the same size (fix x), the house in the more desirable part (NW = 1) is \$30569.1 more than the one in the less desirable part (NW = 0).

(f)
```
> predict(M2, newdata=data.frame(size=4000, NW = "1"))
        1
327252.1
```

The mean price of a house with size $x = 4000$ and NW = 1 is \$327252.1.

(g) The fitted Model 2 has $R^2 = 0.6352$. It means, Model 2 can explain 63.52% variance in the observed response.