31535537 עידן
314696972 נעמה

א.

נשים לב. $P_X y = y$ ובנוסף $y \in Im(X) = P_X y = y$ [illegible] שווה לעצמו
וגם כן [illegible] הגדרה זה הוכחנו כי $P_X := X(X^TX)^{-1}X^T$
ולכן מכך שהראינו כי $\hat{y} = X\hat{\beta} = P_X y$

$$\hat{\beta} = \cancel{(X^TX)^{-1}X^T}X(X^TX^{-1})X^Ty = (X^TX^{-1})X^Ty$$

ב.

$$\theta := a^T\beta = \sum_{j=0}^{p} a_j\beta_j \qquad a = (0,-1,1,0\ldots 0)^T \text{ constant vector}$$

$$\Rightarrow \theta = \beta_2 - \beta_1 \Rightarrow \hat{\theta} = a^T\hat{\beta} = \underbrace{a^T(X^TX)^{-1}X^T}_{C^T} y = C^T y$$

אומדן $\hat{\theta} = \hat{\beta}_2 - \hat{\beta}_1$ נחשב את $E(\hat{\theta})$

$$E(\hat{\theta}) = E(\hat{\beta}_2 - \hat{\beta}_1) = E(\hat{\beta}_2) - E(\hat{\beta}_1) = E(\underbrace{(X_2^TX_2)^{-1}X_2^T}_{:A_2}y_2) - E(\underbrace{(X_1^TX_1)^{-1}X_1^T}_{:A_1}y_1)$$

$$= A_2E(y_2) - A_1E(y_1) = A_2E(X_2\beta_2 + \varepsilon_2) - A_1E(X_1\beta_1 + \varepsilon_1)$$

$$= A_2[E(\underbrace{X_2\beta_2}_{\text{קבוע}}) + E(\underbrace{\varepsilon_2}_{\text{הנחת המודל}=0})] - A_1[E(\underbrace{X_1\beta_1}_{\text{קבוע}}) + E(\underbrace{\varepsilon_1}_{=0})] =$$

$$= \cancel{(X_2^TX_2)^{-1}X_2^TX_2}\beta_2 - \cancel{(X_1^TX_1)^{-1}X_1^TX_1}\beta_1 = \beta_2 - \beta_1$$

שונות

$$Cov(\hat{\theta}) = Cov(\hat{\beta}_2 - \hat{\beta}_1) = cov((X_2^TX_2)^{-1}X_2^Ty_2 - (X_1^TX_1)^{-1}X_1^Ty_1)$$

לפי [illegible] [illegible] $\hat{\theta} = \hat{\beta}_2 - \hat{\beta}_1$ [illegible] [illegible] $Cov(\beta_2 - \beta_1) = Cov(\tilde{\theta})$

הצגנו הליניארית של $\hat{\theta}$ $\tilde{A} := (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T$

$$Cov(\tilde{\theta}) = \tilde{A}Cov(\tilde{y})\tilde{A}^T = \sigma^2\tilde{A}\tilde{A}^T = \sigma^2(\tilde{X}^T\tilde{X})^{-1}$$

רלוונטי [illegible] [illegible] (Gauss) בהוכחה [illegible] [illegible]

$$C = \tilde{X}(\tilde{X}^T\tilde{X})^{-1} \qquad \hat{\theta} = C^Ty$$

כלומר [illegible] [illegible] [illegible] [illegible] $d \in R^n$

אם $d = C$ [illegible] [illegible] [illegible] כלומר $d \neq C$ [illegible]

$\tilde{\theta} = d^Ty$ חסר הטיה [illegible] $Var(\hat{\theta}) < Var(\tilde{\theta})$

בהרצאה [illegible] $\theta \in R^n$ כי [illegible] [illegible] [illegible] [illegible]

$$\theta := a^T\beta = \sum^{p} a_j\beta_j \in R$$

ג.

$$\hat{y}_i = x_i\hat{\beta}_i = x_i(x_i^T x_i)^{-1} x_i^T y_i$$

$$h_i = cov(\underbrace{x_i(x_i^T x_i)^{-1} x_i^T}_{=A} y_i , y_i) = A\, cov(y_i, y_i) = x_i(x_i^T x_i)^{-1} x_i^T \sigma^2$$

σ^2 היא השונות של y

כולה נובע מ$(0,\sigma)$ בגלל הנחה הזו

היא ו x_i קבוע.

שאלה 2.

צ"ל: $Y=(Y_1,Y_2,\dots,Y_n)^T\in R^n$ ונניח $E(Y)=\mu$. $\forall v\in R^n$ ונגדיר $Q(v)=v^TAv$

כאשר $A\in R^{n\times n}$ סימטרית. $E(Q(Y))=Q(\mu)+E[Q(Y-\mu)]$

אם נשתמש ברמז הנתון ניתן לרשום את $Y=\mu+\varepsilon$ כאשר $E(\varepsilon)=0$

אזי $E(Y)=E(\mu+\varepsilon)=\mu+E(\varepsilon)=\mu$ ולכן

$$(*)\ E[Q(Y)]=E(Y^TAY)=E((\mu+\varepsilon)^TA(\mu+\varepsilon))=\mu^TA\mu=Q(\mu)$$

$$(**)\ E[Q(Y-\mu)]=E(Q(\varepsilon))=E(\varepsilon^TA\varepsilon)=A\,E(\varepsilon^T\varepsilon)=A\,E(\varepsilon^T)\,E(\varepsilon)=$$
$$=0$$

לפי הנחה ידוע.

ניתן לראות כי ניתן להגיע מצד אחד לצד השני

$$E[Q(Y-\mu)]+Q(\mu)\underset{(*)(**)}{=}Q(\mu)=\mu^TA\mu=E((\mu+\varepsilon)^TA(\mu+\varepsilon))=$$
$$=E(Y^TAY)=E(Q(Y-\mu))$$

לפי שוויון (*)(**)

השתמשנו ש-$E(\varepsilon)=0$ ו-התוחלת של סכום שווה לסכום התוחלות.

התוחלת של transpose מוגדר כשווה את התוחלות לאיבר עם נעשה

על ה transpose ולקחנו תוחלת.

שאלה 3.

א. מודל רגרסיה הפשוטה. $y_i = \alpha + \beta x_i + \epsilon_i \quad \epsilon_i \overset{iid}{\sim} N(0,\sigma^2)$

simple linear regression $y_i = \alpha + \beta x_i + \varepsilon_i$. Then

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

כאן הנו מודל רגרסיה לינארית פשוטה אשר הוכחנו ש $\hat{\beta}(X^TX)^{-1}X^Ty$

$$Q(b) = \|y - Xb\|^2 = \|y\|^2 + \|Xb\|^2 - 2y^TXb$$

$$= \|y\|^2 + b^TX^TXb - 2(X^Ty)^Tb$$

ומכאן $(X^TX)\hat{\beta} = X^Ty$

$$\frac{\partial Q}{\partial b} = 2b^T(X^TX) - 2(X^Ty)^T = 0 \Rightarrow \beta = (X^TX)^{-1}Xy$$

שווה לאפס

$$Q(\alpha, b_1) = \sum_{i=1}^{n}[y_i - (\alpha + b_1x_i)]^2$$

ולמעשה $(\hat{\beta}_0, \hat{\beta}_1) = \arg\min Q(b_0, b_1)$ ע״י גזירה והשוואה לאפס.

$$\hat{\beta}_1 = \frac{\sum^n (y_i - \bar{y})(x_i - \bar{x})}{\sum^n (x_i - \bar{x})^2} = r_{xy}\frac{s_y}{s_x}$$

$$\hat{\alpha} = \bar{y} - \beta_1\bar{x}$$

עבור סעיף B

Quadratic regression: $Y_i = \alpha + \beta x_i + \gamma x_i^2 + \varepsilon_i$ $(i = 1, \ldots n)$.

$$p = 3, \qquad X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \qquad \theta = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}.$$

הנחנו כי הפיכה X^TX

$$y = X\begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}$$

$$X^Ty \quad X^TX\theta$$

$$(X^TX)^{-1}X^Ty = \cancel{(X^TX)^{-1}(X^TX)}\theta \quad \alpha -$$

$$\theta = (X^TX)^{-1}X^Ty \Rightarrow$$

$$\hat{\alpha} = [(X^TX)^{-1}X^Ty]_1$$

$$\tilde{\beta} = [(X^TX)^{-1}X^Ty]_2$$

$$\gamma = [(X^TX)^{-1}X^Ty]_3$$

ב.

$$\operatorname{cov}(\hat\alpha) = \operatorname{cov}[\underbrace{(X^TX)^{-1}X^T}_{:A} Y]_{1} = A\underbrace{\operatorname{cov}(Y)}_{\sigma^2 I_n}A^T = \sigma^2 AA^T =$$

$$= \sigma^2 (\cancel{X^TX})^{-1}\cancel{X^TX}(X^TX)^{-1} = [\sigma^2 (X^TX)^{-1}]_{11}$$

ניתן לראות כי באותו אופן עבור β נקבל)

$[\sigma^2 (X^TX)^{-1}]_{22}$

$$\hat\beta_1 = \frac{\sum(x_i-\bar x)y_i}{\sum(x_i-\bar x)^2} - \frac{\bar y \overbrace{\sum(x_i-\bar x)}^{=0}}{\sum(x_i-\bar x)^2} = \frac{\sum(x_i-\bar x)y_i}{\sum(x_i-\bar x)^2}$$

$$\operatorname{var}(\beta_1) = \frac{\sum(x_i-\bar x)y_i}{\sum(x_i-\bar x)^2} = \frac{1}{\sum(x_i-\bar x)^2}\operatorname{var}\left(\sum(x_i-\bar x)y_i\right)$$

$$= \frac{1}{(\sum(x_i-\bar x)^2)^2}\operatorname{var}(y_i) + \sum_{i\neq j}^{n}(x_i-\bar x)(x_j-\bar x)\underbrace{\operatorname{cov}(y_i,y_j)}_{=0} =$$

מהנחה כי $e = \sigma^2 I_n$

$$= \frac{\sigma^2}{\sum(x_i-\bar x)^2}$$

$$\operatorname{var}(\hat\alpha) = \operatorname{var}(\bar y - \beta_1\bar x) = v(\bar y) + v(\beta_1\bar x) - 2\operatorname{cov}(\bar y, \beta_1\bar x) =$$

$$= \frac{1}{n^2}\sum\operatorname{var}(y_i) + \bar x^2\frac{\sigma^2}{\sum(x_i-\bar x)^2} - 2\bar x\operatorname{cov}(\bar y, \hat\beta_1) =$$

$$= \sigma^2\left[\frac{1}{n} + \frac{\bar x^2}{\sum(x_i-\bar x)^2}\right]$$

מסקנה: הכלל אונית נכון כי שווה

ואפס. אז

$$\operatorname{Var}(\hat\alpha) = \sigma^2\left[\frac{1}{n} + \frac{\bar x^2}{\sum(x_i-\bar x)^2}\right]$$

$$\operatorname{cov}(\bar{Y}, \hat{\beta}_1) = 0$$

$$\Downarrow$$

$$= \operatorname{cov}(\beta_0 + \beta_1 \bar{X} + \bar{\varepsilon}, \hat{\beta}_1) = \operatorname{cov}(\bar{\varepsilon}, \hat{\beta}_1) =$$

$$\operatorname{cov}\left(\bar{\varepsilon}, \frac{\sum^n (X_i - \bar{X})(\beta_0 + \beta_1 X + \varepsilon_i)}{\sum (X_i - \bar{X})^2}\right) = \operatorname{cov}\left(\bar{\varepsilon}, \frac{\sum (X_i - \bar{X})\varepsilon_i}{\sum (X_i - \bar{X})^2}\right)$$

$$= \frac{1}{\sum (X_i - \bar{X})^2} \operatorname{cov}(\bar{\varepsilon}, \sum (X_i - \bar{X})\varepsilon_i) = \frac{1}{\sum (X_i - \bar{X})^2} \frac{1}{n} \sum \operatorname{cov}(\varepsilon_j, \sum (X_i - \bar{X})\varepsilon_i)$$

$$= \frac{1}{\sum (X_i - \bar{X})^2} \cdot \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1} (X_i - \bar{X}) \operatorname{cov}(\varepsilon_j, \varepsilon_i) = 0$$

$$\operatorname{cov}(\hat{\alpha}, \hat{\beta}) = \operatorname{cov}(\bar{Y} - \hat{\beta}_1 \bar{X}, \hat{\beta}) = \underbrace{\operatorname{cov}(\bar{Y}, \hat{\beta})}_{=0} - \bar{X} \operatorname{cov}(\hat{\beta}) =$$

$$= -\bar{X}\left[\frac{\sigma^2}{\operatorname{Var}(X)}\right] = -\bar{X}\,[\sigma^2 (X^T X)^{-1}]_{11}$$

$$:\varepsilon$$

$$\operatorname{cov}(\hat{\gamma}) = \operatorname{cov}[\underbrace{(X^T X)^{-1} X^T}_{:A} (\cancel{\beta})] = A \underbrace{\operatorname{cov}(Y)}_{\sigma^2 I_n} A^T = \sigma^2 A A^T =$$

$$= \sigma^2 \cancel{(X^T X)^{-1}} \cancel{X^T X} (X^T X)^{-1} = [\sigma^2 (X^T X)^{-1}]_{33}$$

.3 דהמשך?

B) רווח סמך ל β תחת מודל A:

אם $\varepsilon \sim N(0,1)$ אז: $\hat{\beta} \sim N(\beta, \sigma^2 (X^TX)^{-1})$.

נחשב באופן הבא:

$$\frac{\hat{\beta}_j-\beta_j}{\sqrt{\hat{var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j-\beta_j}{\sqrt{\hat{\sigma}^2(X^TX)^{-1}_{jj}}} \cdot \frac{\sqrt{\sigma^2(X^TX)^{-1}_{jj}}}{\sqrt{\sigma^2(X^TX)^{-1}_{jj}}} = \frac{(\hat{\beta}_j-\beta_j)/\sqrt{\sigma^2(X^TX)^{-1}_{jj}}}{\sqrt{\hat{\sigma}^2(X^TX)^{-1}_{jj}}/\sqrt{\sigma^2(X^TX)^{-1}_{jj}}} =$$

$$\frac{(\hat{\beta}_j-\beta_j)/\sqrt{\sigma^2(X^TX)^{-1}_{jj}}}{\hat{\sigma}/\sigma} \sim \frac{N(0,1)}{\sqrt{\chi^2_{n-p-1}/(n-p-1)}} \sim t_{n-p-1}$$

$$\Downarrow$$

לכן נבנה סטטיסטי הוא $T_j = \frac{\hat{\beta}_j-\beta_j}{\sqrt{\hat{\sigma}^2(X^TX)^{-1}_{jj}}} \sim t_{n-p-1}$

$$1-\alpha = P\left(t_{n-p-1,\frac{\alpha}{2}} \le T_j \le t_{n-p-1,1-\frac{\alpha}{2}}\right) =$$

$$P\left(\hat{\beta}_j - \sqrt{\hat{\sigma}^2(X^TX)^{-1}_{jj}}\, t_{n-p-1,1-\frac{\alpha}{2}} \le \beta_j \le \hat{\beta}_j + \sqrt{\hat{\sigma}^2(X^TX)^{-1}_{jj}} \right)$$

לכן רס"ל β_j ברמת אמון $1-\alpha$ הוא: $\left[\hat{\beta}_j \pm \sqrt{\hat{\sigma}^2(X^TX)^{-1}_{jj}} \cdot t_{n-p-1,1-\frac{\alpha}{2}}\right]$

סעיף ה

נבדוק $H_0: \gamma=0$

כי תחת H_0 $\gamma_j=0$

$$T_j = \frac{\hat{\gamma}_j - 0}{\sqrt{\hat{\sigma}^2(X^TX)^{-1}_{jj}}} \sim t_{n-p-1}$$

כאשר אנו מקבלים את ההנחות לסעיף קודם.

לכן המבחן לבדיקת ההשערה ש $\gamma=0$ הוא דחיית H_0

אם $$\left|\frac{\hat{\gamma}}{\sqrt{\hat{\sigma}^2(X^TX)^{-1}_{jj}}}\right| \ge t_{n-p-1,1-\frac{\alpha}{2}}$$

Quiz 1

Menachem Sokolik and Idan Keipour

12 5 2021

```
My_lm <- function(X,Y){
  #' @param X is matrix with n rows and p+1 columns
  #' @param Y is vector with length n
  # TODO: Complete the code.
  #       Don't use any external libaries or functions for computing linear models.
  X <- as.matrix(X)
  X <- cbind(1,X) # n*(p+1)
  Y <- as.matrix(Y)
  n <- nrow(X)
  p <- ncol(X) - 1
  dof <-  n-p-1
  # t(X) is the transponse, %*% is matrix multiplication,
  # solve takes the inverse
  A <- solve(t(X) %*% X)
  # this comes from looking at formula for beta hat and
  # multiplying the additional terms needed
  beta_hat <- A %*% t(X) %*% Y
  Y_hat <- X %*% beta_hat
  # just filling in formula
  residuals <- Y - (Y_hat)
  sigma2_hat <- t(residuals) %*% residuals/dof
  # I’ve put as.numeric() at the end to convert it to a single scalar number
  # instead of a 1 by 1 matrix.
  sigma2_hat <- as.numeric(sigma2_hat)
  # covariance matrix of estimated regression coefficients
  # is just the estimated residual variance times solve(t(X) %*% X)
  # we calculuted earlier and stored as A
  beta_cov <- sigma2_hat * A
  # diag takes the diagonal of the matrix, sqrt makes it go
  # from variance to standard deviation
  beta_SE <- sqrt(diag(beta_cov))

  T_stat <- beta_hat/beta_SE
  P_Values <- 2*(1-pt(abs(T_stat),dof))

  # residual sum of squares divided by total sum of squares
  R_squared <- 1- sum((residuals)^2)/sum((Y - mean(Y))^2)
  return(list(beta_hat, beta_cov, T_stat, P_Values, sigma2_hat, R_squared, dof, residuals, X,
Y, Y_hat))
}
```

```
df <- read.csv("startups.csv")
```

i

```
str(df)
```

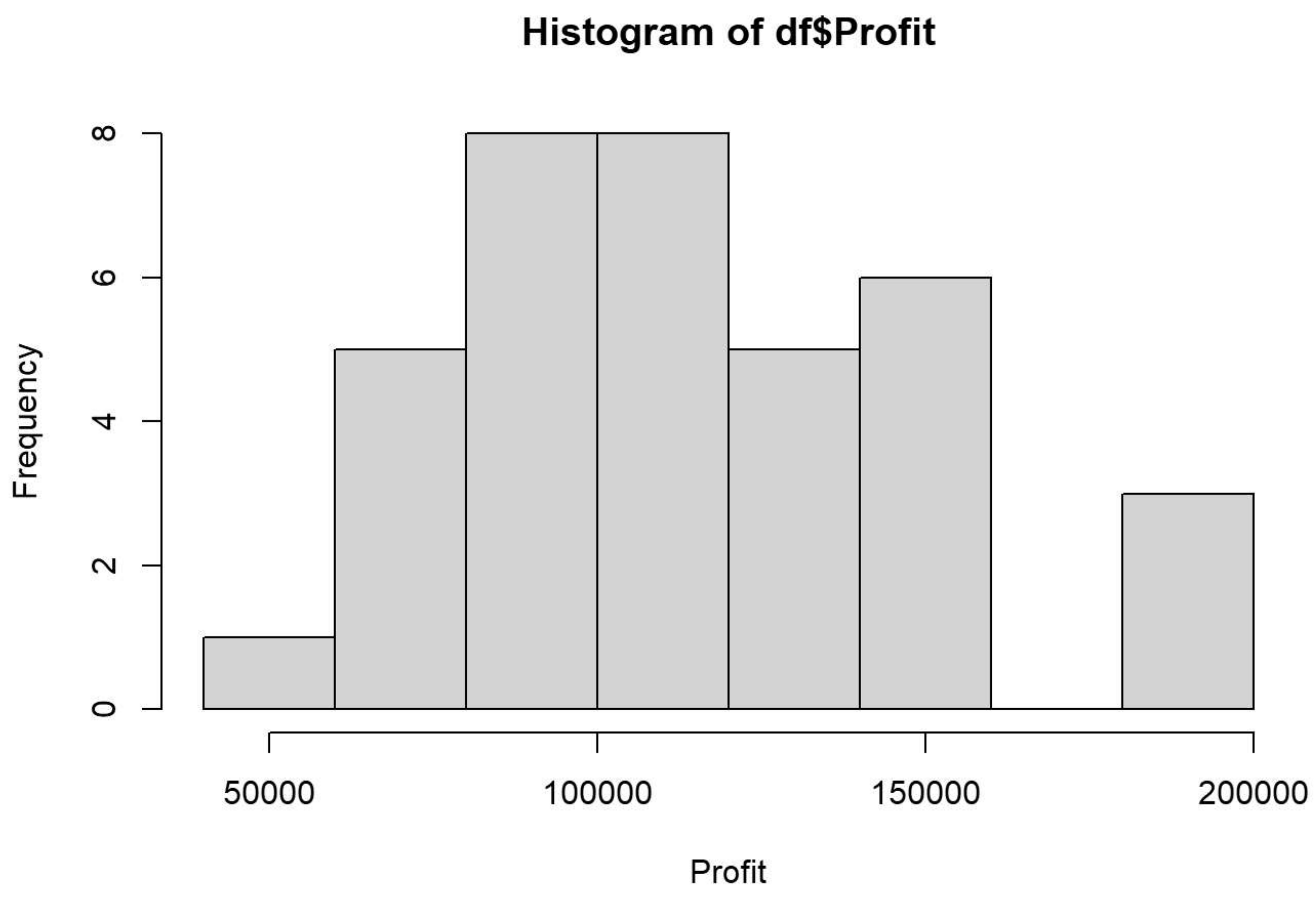
```
## 'data.frame':	36 obs. of  5 variables:
##  $ R.D.Spend      : num  165349 162598 153442 131877 130298 ...
##  $ Administration : num  136898 151378 101146 99815 145530 ...
##  $ Marketing.Spend: num  471784 443899 407935 362861 323877 ...
##  $ State          : chr  "New York" "California" "Florida" "New York" ...
##  $ Profit         : num  192262 191792 191050 156991 155753 ...
```

```
summary(df)
```

```
##    R.D.Spend      Administration   Marketing.Spend     State          
##  Min.   :  1000   Min.   : 51283   Min.   :  1904   Length:36         
##  1st Qu.: 45837   1st Qu.:105078   1st Qu.:139922   Class :character  
##  Median : 74662   Median :122700   Median :239453   Mode  :character  
##  Mean   : 76578   Mean   :123041   Mean   :224241                     
##  3rd Qu.:105066   3rd Qu.:144485   3rd Qu.:300633                     
##  Max.   :165349   Max.   :182646   Max.   :471784                     
##      Profit      
##  Min.   : 49491  
##  1st Qu.: 95212  
##  Median :108643  
##  Mean   :115300  
##  3rd Qu.:136795  
##  Max.   :192262  
```

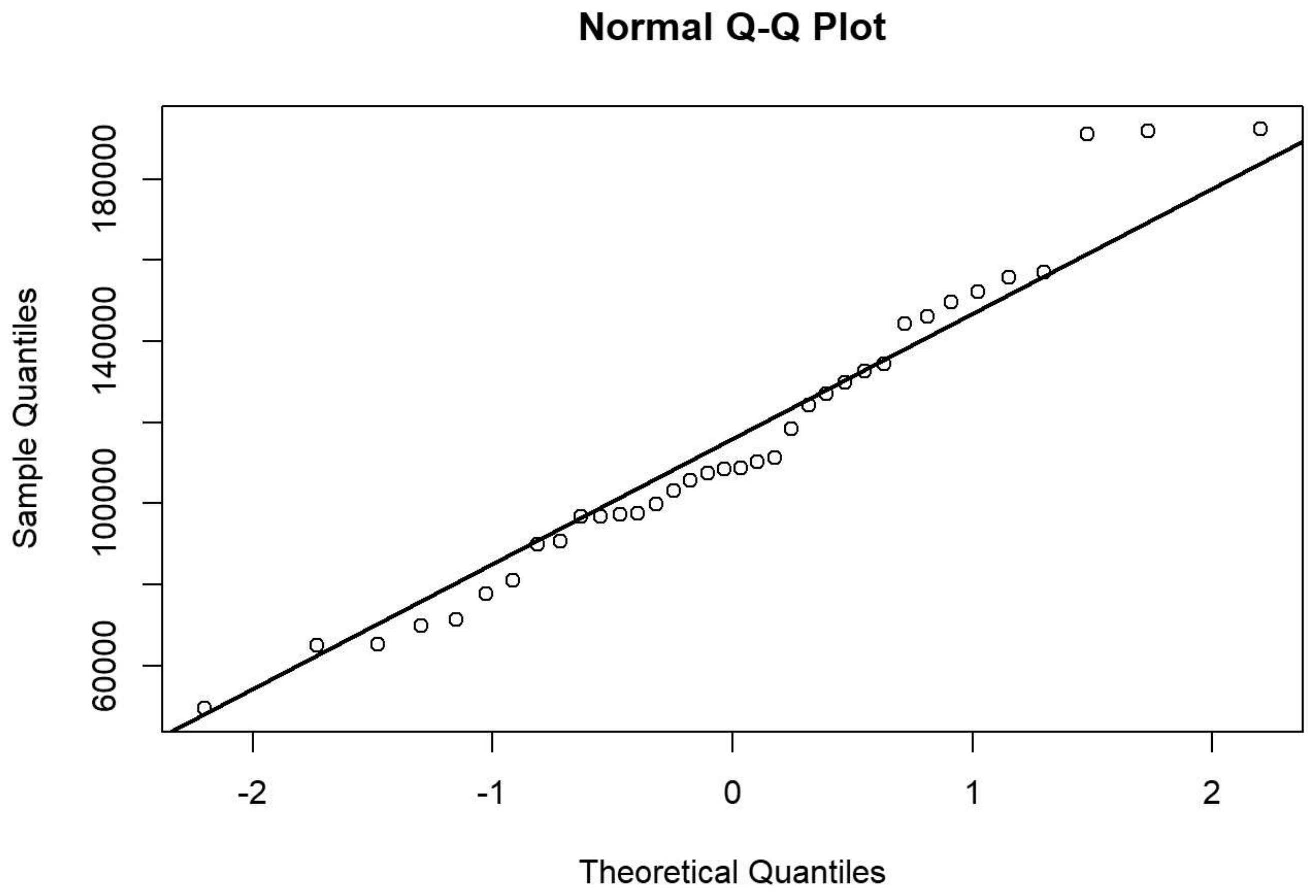
ii

```
hist(df$Profit, xlab = "Profit")
```



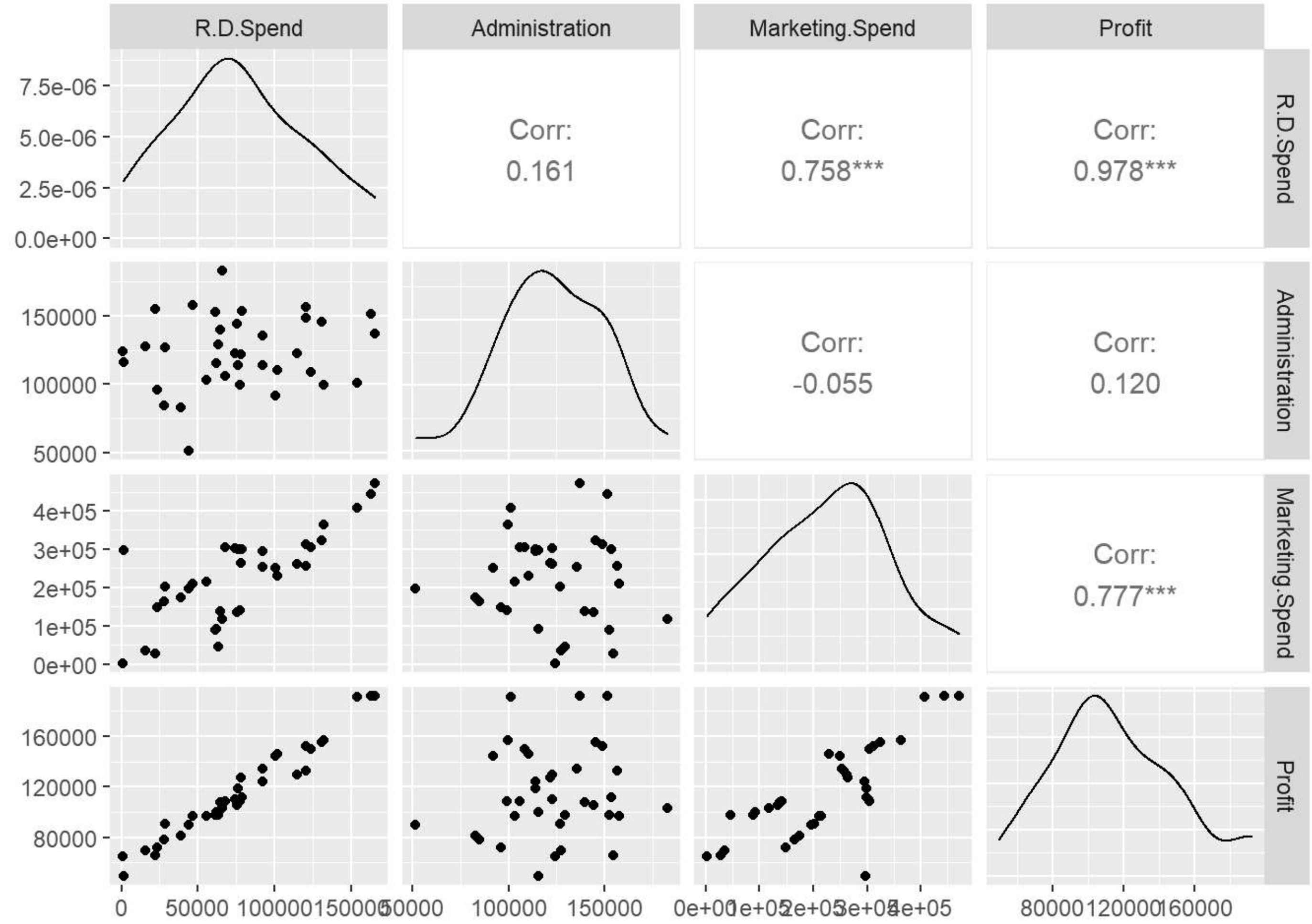
looks like a normal distribution but with long right tail. also can see that with the QQPlot to identify normality.

```
qqnorm(df$Profit)
qqline(df$Profit, lwd = 2)
```

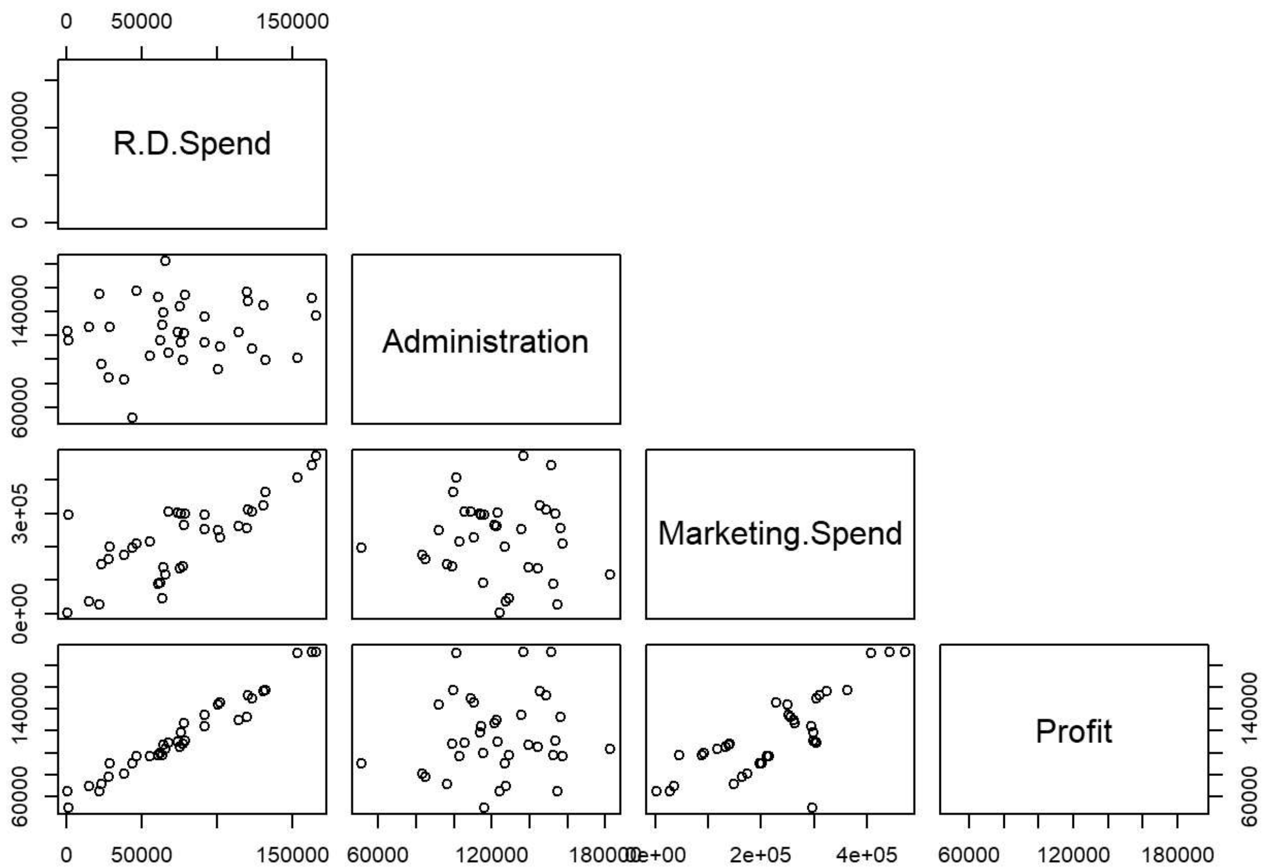


iii

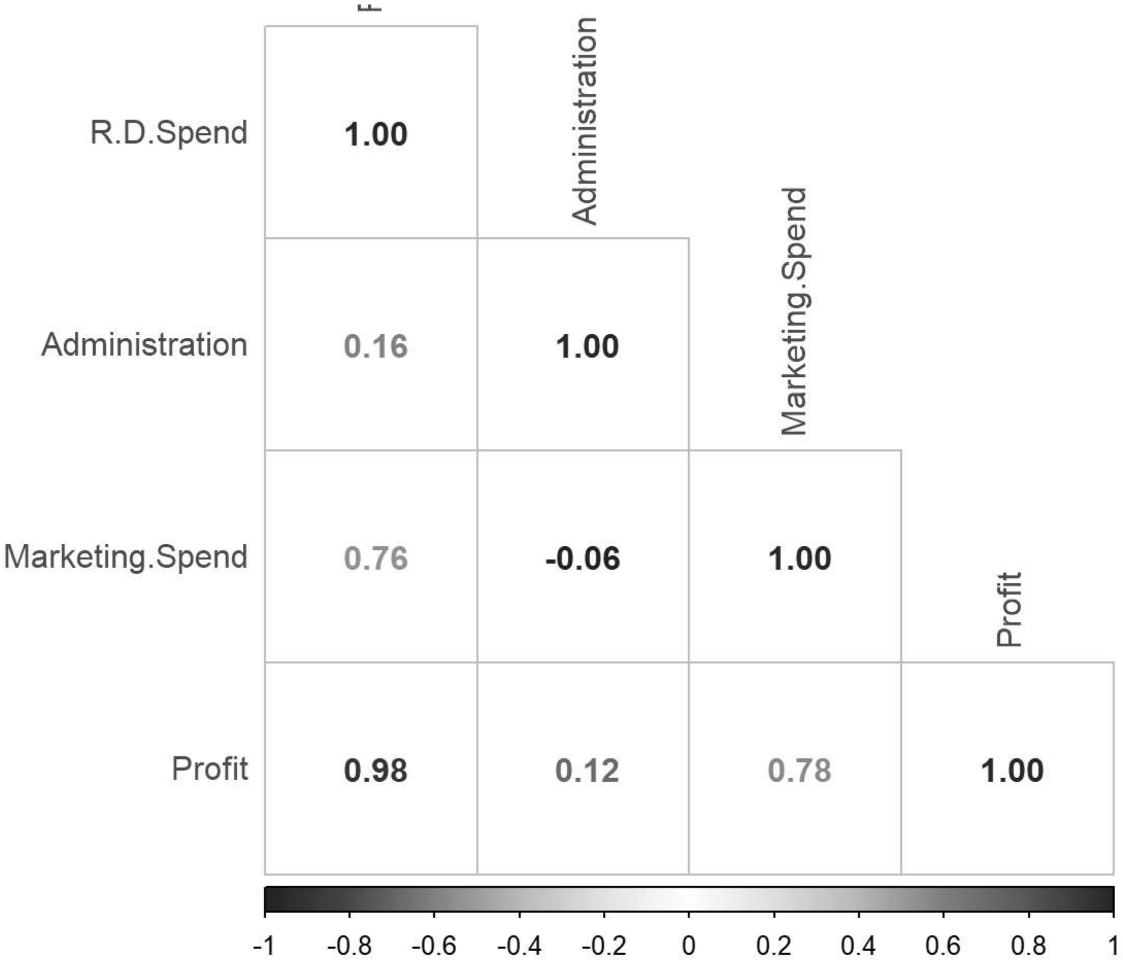
```
# automatic for the correlation coefficients.
# The diagonal consists of the densities of the three variables and the upper panels consist
 of the correlation coefficients between the variables.
ggpairs(df[c(-4)])
```




```
pairs(~ R.D.Spend + Administration + Marketing.Spend + Profit, data = df, upper.panel = NULL)
```



```
# by hand
cor_df <- cor(df[c(-4)])
corrplot(cor_df, method="number",type = "lower" )
```




```
cor_df
```

```
##                 R.D.Spend Administration Marketing.Spend    Profit
## R.D.Spend       1.0000000     0.16054446      0.75826916 0.9782919
## Administration  0.1605445     1.00000000     -0.05504912 0.1196753
## Marketing.Spend 0.7582692    -0.05504912      1.00000000 0.7770678
## Profit          0.9782919     0.11967533      0.77706784 1.0000000
```

iv

```
# matrix of profit
# dummy <- for k variables making k-1 categorical.
df$New_York <- ifelse(df$State == "New York", 1,0)
df$Florida <- ifelse(df$State == "Florida", 1,0)
Y <- df["Profit"]
X <- df %>% select(-c("Profit","State"))
My_reg <- My_lm(X,Y)
coeff <- My_reg[[1]]
rownames(coeff)[1] = "Intercept"
coeff
```

```
##                        Profit
## Intercept        5.633027e+04
## R.D.Spend        7.660746e-01
## Administration  -2.616449e-02
## Marketing.Spend  2.384049e-02
## New_York        -2.939637e+03
## Florida         -2.792330e+03
```

v

```r
t_stat <- My_reg[3]
t_stat
```

```
## [[1]]
##                       Profit
##                    8.1873071
## R.D.Spend         15.8224079
## Administration    -0.5077559
## Marketing.Spend    1.3142079
## New_York          -0.8746709
## Florida           -0.9405045
```

vi

```r
CI <- function(alpha, coeff, beta_cov, dof){
  CI_betas <- data.frame(beta=NA, lower_beta=NA, upper_beta=NA)
  for(i in 1:length(coeff)){
  beta <- rownames(coeff)[i]
  lower_beta <- coeff[i] - sqrt(beta_cov[i])*qt(1-alpha/2,dof)
  upper_beta <- coeff[i] + sqrt(beta_cov[i])*qt(1-alpha/2,dof)
  CI_betas <-rbind(CI_betas, c(beta, lower_beta, upper_beta))}
  return(CI_betas)}
```

vii

```r
alpha= 0.05
beta_cov <- diag(My_reg[[2]])
dof <- My_reg[[7]]
CI(alpha, coeff, beta_cov, dof)
```

```
##              beta          lower_beta         upper_beta
## 1            <NA>                <NA>               <NA>
## 2       Intercept     42279.037676838   70381.5040675557
## 3       R.D.Spend   0.667193797647681  0.864955498705348
## 4  Administration  -0.131402078859794 0.0790731063466653
## 5 Marketing.Spend -0.0132075109226804  0.060888492906907
## 6        New_York    -9803.4083213011   3924.13344461858
## 7         Florida   -8855.77633934312   3271.11644734578
```

We will be rejected H_0 when T_j in is absolute value is greater than qt(dof). So when truth comes out we reject the null hypothesis otherwise we do not reject. Also, can not see by p.value test when p.value greater than alpha does not reject the null hypothesis, otherwise rejected. rejecting null hypothesis of intercept and for R.D spend.

```
abs(t_stat[[1]])>qt(0.975, dof)
```

```
##                   Profit
##                     TRUE
## R.D.Spend           TRUE
## Administration     FALSE
## Marketing.Spend    FALSE
## New_York           FALSE
## Florida            FALSE
```

```
P.value <- My_reg[[4]]
P.value
```

```
##                         Profit
##                   3.866198e-09
## R.D.Spend         4.440892e-16
## Administration    6.153375e-01
## Marketing.Spend   1.987366e-01
## New_York          3.887019e-01
## Florida           3.544707e-01
```

question 4.c

```
new_df <- read.csv("Startups_test.csv")
new_df <- cbind("Intercept"=1,new_df)
new_df$New_York <- ifelse(new_df$State == "New York", 1,0)
new_df$Florida <- ifelse(new_df$State == "Florida", 1,0)
real_profit <- new_df$Profit
new_df<- new_df %>% select(-c("Profit","State"))
new_df <- as.matrix(mutate(new_df,"exp_profit" = as.matrix(new_df)%*%coeff))
new_df
```

```
##       Intercept R.D.Spend Administration Marketing.Spend New_York Florida
##  [1,]         1  46014.02       85047.44        205517.6        1       0
##  [2,]         1 134615.46      147198.87        127716.8        0       0
##  [3,]         1  72107.60      127864.55        353183.8        1       0
##  [4,]         1  94657.16      145077.58        282574.3        1       0
##  [5,]         1  20229.59       65947.93        185265.1        1       0
##  [6,]         1 144372.41      118671.85        383199.6        1       0
##  [7,]         1  28754.33      118546.05        172795.7        0       0
##  [8,]         1  65605.48      153032.06        107138.4        1       0
##  [9,]         1 142107.34       91391.77        366168.4        0       1
## [10,]         1  93863.75      127320.38        249839.4        0       1
##       exp_profit
##  [1,]   91315.23
##  [2,]  158649.21
##  [3,]  113705.00
##  [4,]  128845.91
##  [5,]   71579.33
##  [6,]  170021.36
##  [7,]   79376.07
##  [8,]  102199.55
##  [9,]  168741.19
## [10,]  128069.60
```

```
n <- length(real_profit)
RMSE <- sqrt((1/n)*sum((new_df[,7] - real_profit)^2))
RMSE
```

```
## [1] 7561.162
```