STAT 425 Homework-1

Sharvi Tomar (stomar2)

01/09/21

Contents

Problem 1:

The data set prostate from the faraway library, is from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. Make a numerical and graphical summary of the data. Comment on any features you find interesting.

```
library(faraway)
prostate_data=data.frame(prostate)
str(prostate_data)
```

```
## 'data.frame':    97 obs. of  9 variables:
##  $ lcavol : num  -0.58 -0.994 -0.511 -1.204 0.751 ...
##  $ lweight: num  2.77 3.32 2.69 3.28 3.43 ...
##  $ age    : int  50 58 74 58 62 50 64 58 47 63 ...
##  $ lbph   : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
##  $ svi    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ lcp    : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
##  $ gleason: int  6 6 7 6 6 6 6 6 6 6 ...
##  $ pgg45  : int  0 0 20 0 0 0 0 0 0 0 ...
##  $ lpsa   : num  -0.431 -0.163 -0.163 -0.163 0.372 ...
```

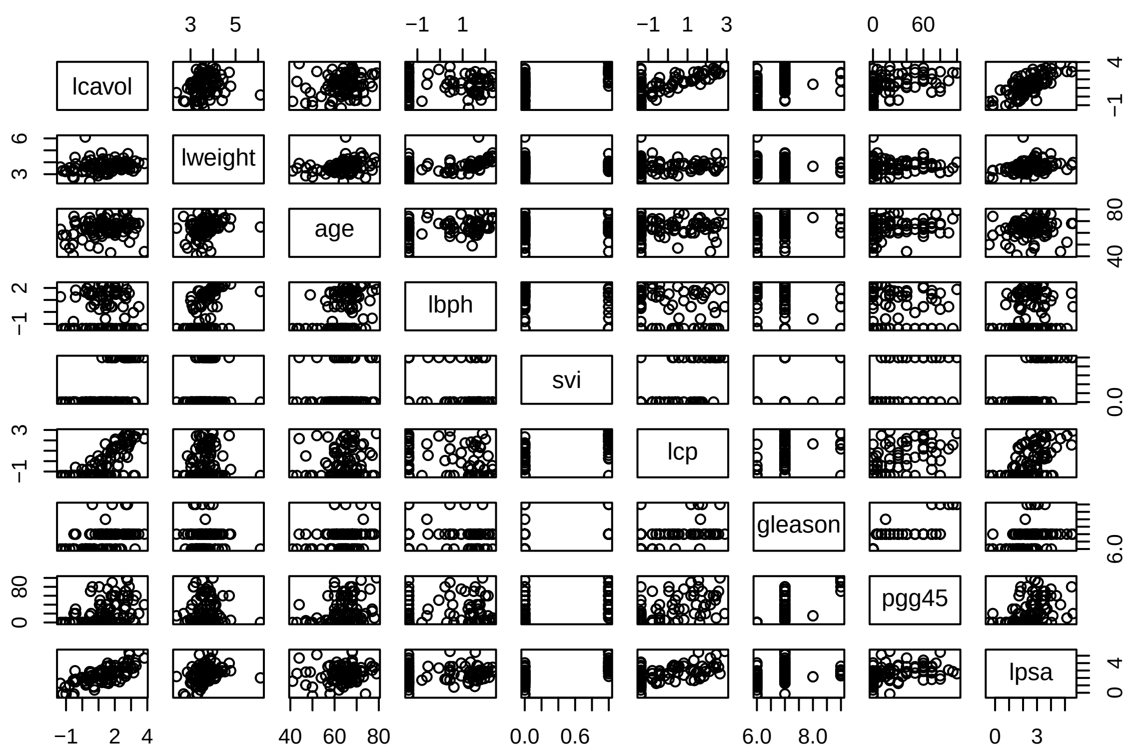
```
# Numerical summary
summary(prostate)
```

```
##      lcavol           lweight          age             lbph
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863
##  Median : 1.4469   Median :3.623   Median :65.00   Median : 0.3001
##  Mean   : 1.3500   Mean   :3.653   Mean   :63.87   Mean   : 0.1004
```

```
##   3rd Qu.: 2.1270   3rd Qu.:3.878   3rd Qu.:68.00   3rd Qu.: 1.5581
##   Max.   : 3.8210   Max.   :6.108   Max.   :79.00   Max.   : 2.3263
##        svi              lcp             gleason          pgg45
##   Min.   :0.0000   Min.   :-1.3863   Min.   :6.000   Min.   :  0.00
##   1st Qu.:0.0000   1st Qu.:-1.3863   1st Qu.:6.000   1st Qu.:  0.00
##   Median :0.0000   Median :-0.7985   Median :7.000   Median : 15.00
##   Mean   :0.2165   Mean   :-0.1794   Mean   :6.753   Mean   : 24.38
##   3rd Qu.:0.0000   3rd Qu.: 1.1786   3rd Qu.:7.000   3rd Qu.: 40.00
##   Max.   :1.0000   Max.   : 2.9042   Max.   :9.000   Max.   :100.00
##        lpsa
##   Min.   :-0.4308
##   1st Qu.: 1.7317
##   Median : 2.5915
##   Mean   : 2.4784
##   3rd Qu.: 3.0564
##   Max.   : 5.5829
```

The minimum age of the men with prostate cancer is 41 years and the maximum is 79 years (no young men in the data).

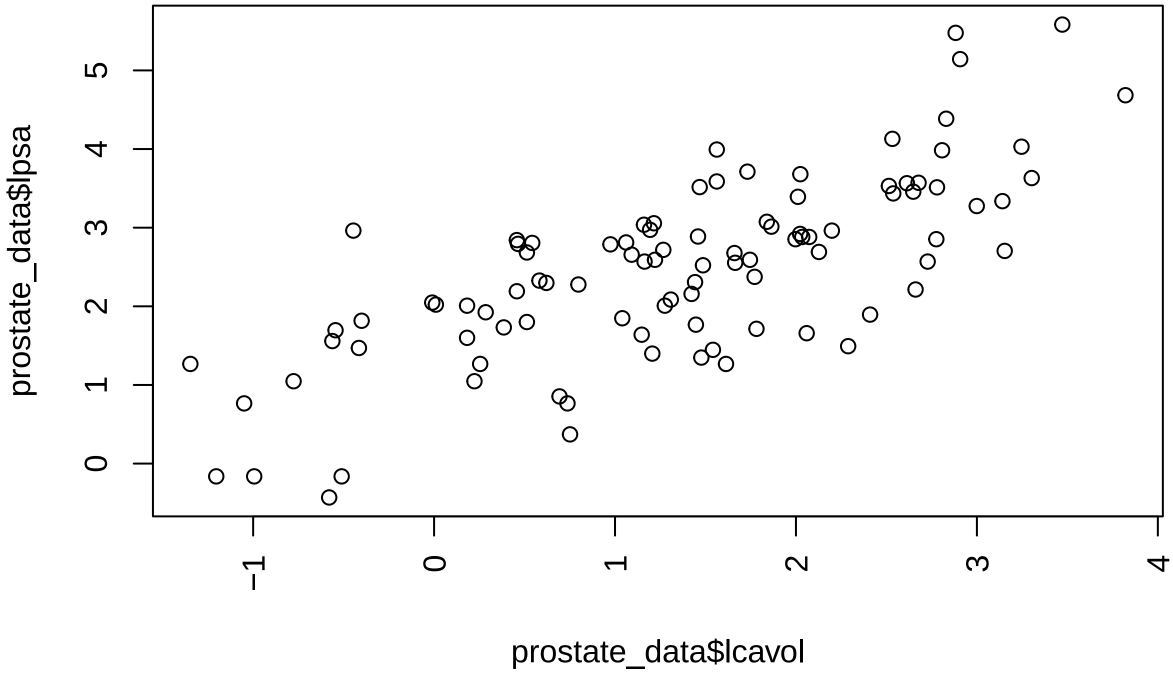
```
# Graphical summary
pairs(prostate)
```



1.From the top-right and botton-left corner, we can see a relationship between 'lcavol and 'lpsa' values.

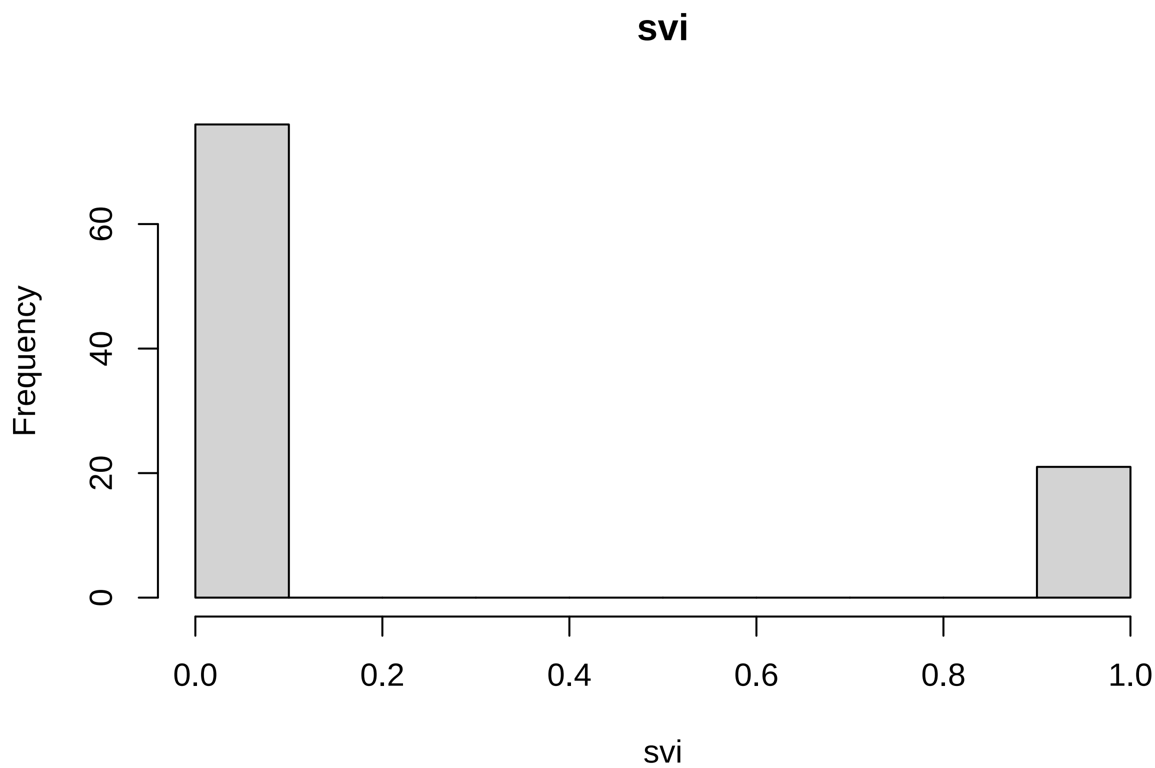
2.The nature of scatter plots of 'svi' and 'gleason' are different and hence, we can check the distributions of these variables to understand better.

```r
# Plotting the variable 'lpsa' with 'lcavol'
plot(prostate_data$lpsa ~ prostate_data$lcavol, las=3)
```



There is an overall positive linear relationship between lspa and lcavol. The log of prostate specific antigen (lspa) seems to increase with the increase in log cancer vol (lcavol).

```r
hist(prostate_data$svi,main="svi",xlab="svi")
```

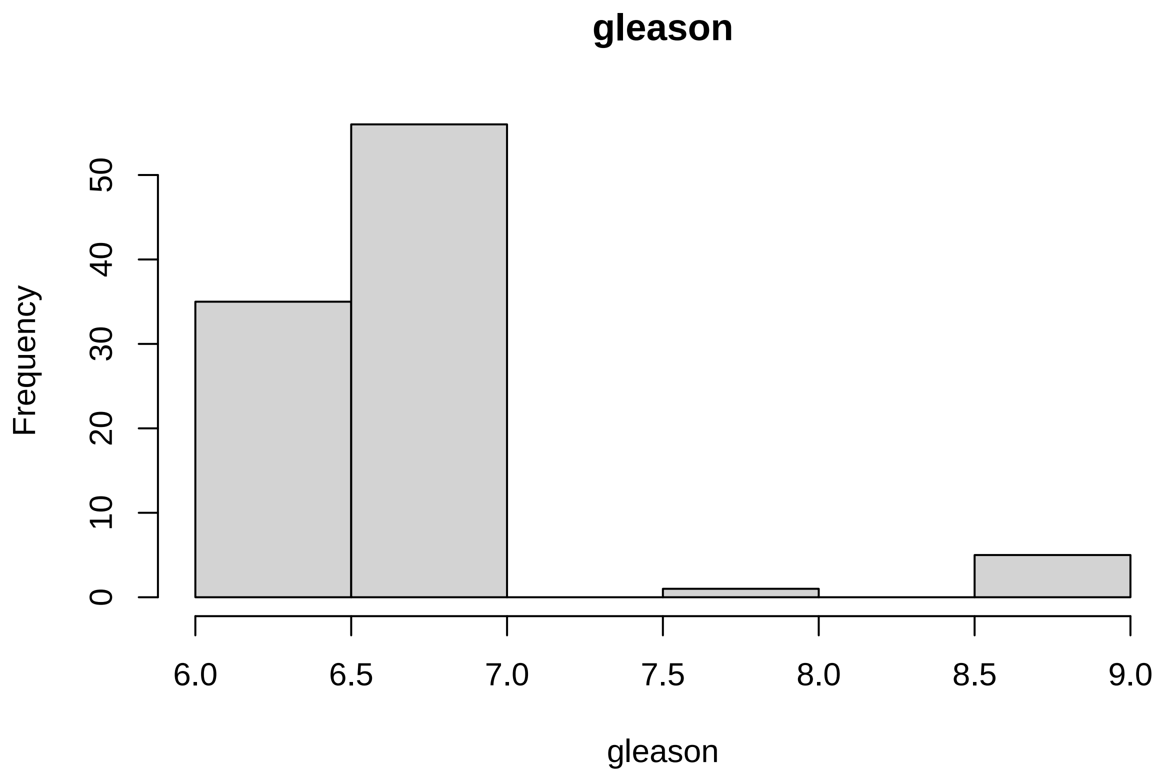



```
table(prostate_data$svi)
```

```
##
##  0  1
## 76 21
```

Variable 'svi' takes only 2 values 0 and 1. Most men(76 out of 97) in the data, do not have seminal vesicle invasion or svi. It makes sense to represent it as a factor variable while performing linear regression.

```
hist(prostate_data$gleason,main="gleason",xlab="gleason")
```

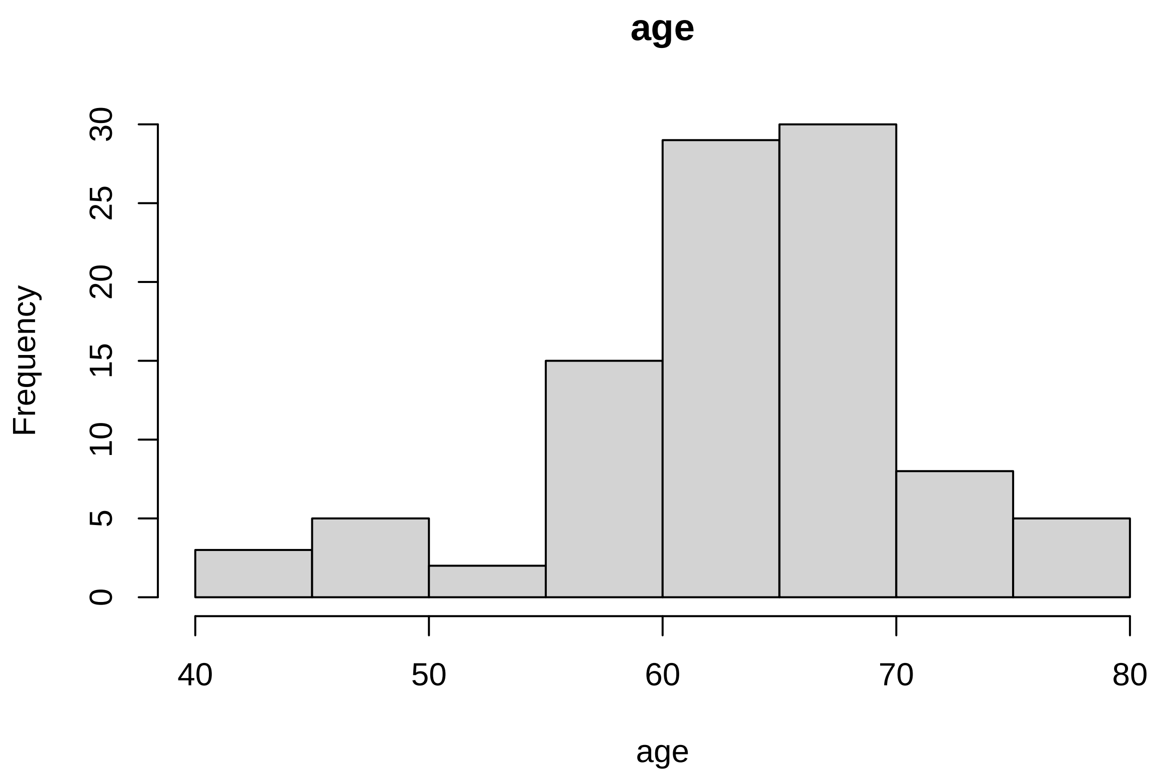



```
table(prostate_data$gleason)
```

```
##
##  6  7  8  9
## 35 56  1  5
```

Variable 'gleason' takes the values 6,7,8,9 and stands for Gleason score. Most people (91 out of 97) in the data have a gleason score of 6-7 and only a few have a higher gleaon score of 8-9.

```
# Plotting distribution of 'age'
hist(prostate_data$age,main="age",xlab="age")
```


age
Frequency
0
5
10
15
20
25
30
40
50
60
70
80
age

Problem 2

Show that for the SLR model, the coefficient of determination R2 is equal to the square of the correlation coef-

2. $$r_{xy}^2 = \left(\frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \right)^2$$

$$= \frac{S_{xy}}{S_{xx}} \cdot \frac{S_{xy}}{S_{yy}}$$

$$= \hat{\beta}_1 \frac{S_{xy}}{S_{yy}}$$

$$= \hat{\beta}_1 \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (y_i - \bar{y})^2}$$

$$= \frac{\sum_i \hat{\beta}_1 (x_i - \bar{x})(y_i - \bar{y})}{\sum (y_i - \bar{y})^2} \qquad [\hat{\beta}_1 (x_i - \bar{x}) = \hat{y}_i - \bar{y}]$$

$$= \frac{\sum_i (\hat{y}_i - \bar{y})(y_i - \bar{y} + \hat{y}_i - \hat{y}_i)}{\sum (y_i - \bar{y})^2}$$

$$= \frac{\sum_{i=1} (\hat{y}_i - \bar{y}) \left[(\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \right]}{\sum_i (y_i - \bar{y})^2}$$

Check below page for proof

$$= \frac{\sum_{i=1} (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} + \frac{\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}{\sum_i (y_i - \bar{y})^2}$$

$$= \frac{FSS}{TSS} + 0 \qquad \left(\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0 \right)$$

$$= R^2$$

ficient r2 XY.

c) $\sum_{i=1} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$ $\quad [\hat{\beta}_1(x_i - \bar{x}) = \hat{y}_i - \bar{y}]$

$= \sum_i \hat{\beta}_1 (x_i - \bar{x})(y_i - \hat{y}_i)$

$= \hat{\beta}_1 \sum_i (x_i - \bar{x}) e_i$ $\quad [y_i - \hat{y}_i = e_i]$

$= \hat{\beta}_1 \left(\sum_i x_i e_i - \bar{x} \sum_i e_i \right)$ $\quad [\sum_i e_i = 0]$

$= \hat{\beta}_1 \sum_i x_i e_i$

$= \hat{\beta}_1 \sum_i x_i \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right)$

$= \hat{\beta}_1 \left(\sum_i x_i y_i - \hat{\beta}_0 \sum_i x_i - \hat{\beta}_1 \sum x_i^2 \right)$

$= \hat{\beta}_1 \left(\sum_i x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_i x_i - \hat{\beta}_1 \sum_i x_i^2 \right)$

$= \hat{\beta}_1 \left(\sum_i x_i y_i - \bar{y} \sum_i x_i + \hat{\beta}_1 \left(\bar{x} \sum_i x_i - \sum_i x_i^2 \right) \right)$

$= \hat{\beta}_1 \left(\sum_i x_i y_i - \frac{\sum_i x_i \sum_i y_i}{n} + \frac{S_{xy}}{S_{xx}} \left(\frac{(\sum_i x_i)^2}{n} - \sum_i x_i^2 \right) \right)$

$= \hat{\beta}_1 \left(\sum_i x_i y_i - \frac{\sum_i x_i \sum_i y_i}{n} + \frac{\sum_i x_i (y_i - \bar{y})}{\sum_i x_i (x_i - \bar{x})} \left(\frac{(\sum_i x_i)^2}{n} - \sum_i x_i^2 \right) \right)$

$= \hat{\beta}_1 \left(\sum_i x_i y_i - \frac{\sum_i x_i \sum y_i}{n} + \frac{\left(\sum_i x_i y_i - \frac{\sum x_i \sum y_i}{n} \right)}{\cancel{\sum x_i^2 - (\sum x_i)^2/n}} \cancel{\left(\sum_i x_i^2 - (\sum_i x_i)^2/n \right)} (-1) \right)$

$= \hat{\beta}_1 \left(\cancel{\sum_i x_i y_i} - \cancel{\frac{\sum_i x_i \sum y_i}{n}} - \cancel{\sum_i x_i y_i} + \cancel{\frac{\sum_i x_i \sum y_i}{n}} \right)$

$= 0$

Figure 1: .

Problem 3

Straight line regression through the originIn this question we shall make the following assumptions: (1)Yis related toxby the simple linear regression modelYi= xi+ei(i= 1,2,...,n), i.e.E(Y|X=xi) = xi (2) The errorse1,e2,...,enare independent from each other. (3) The errorse1,e2,...,enhave a common variance. (4) The errors are normally distributed with a mean 0 and variance 2(especially when the sample size is small), i.e.,e|X N(0, 2). In addition, since the regression model is conditional onXwe can assume that the values of the predictor variablex1,x2,...,xn are known fixed constants.

(a) Show that the least squares estimate of is given by:ˆ = ni=1xiyi ni=1x2i
(b) Under the above assumptions show that: +(i)E[ˆ β] = β (ii)V ar(ˆ β) = 2 ni=1x2i +(iii)ˆ β|X N(

3. a) For regression through origin,

$Y_i = \beta x_i + e_i$

We can estimate β using the Least-squares principle

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta x_i)^2$$

Differentiating and equating to zero (wrt β)

$$2 \sum_{i=1}^{n} (y_i - \beta x_i)(-x_i) = 0$$

$$-\sum_{i=1}^{n} x_i y_i + \beta \sum_{i=1}^{n} x_i^2 = 0$$

$$\beta \sum_i x_i^2 = \sum_i x_i y_i$$

$$\boxed{\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}}$$

b) i) $E[\hat{\beta}] = E\left[\frac{\sum_i x_i y_i}{\sum_i x_i^2}\right]$

$$= \frac{\sum_i x_i E(y_i)}{\sum_i x_i^2} \quad [x_1, x_2 \ldots x_n \to \text{known fixed constants}]$$

$$= \frac{\sum_i x_i (\beta x_i)}{\sum_i x_i^2} \quad [E(Y|X = x_i) = \beta x_i]$$

$$E[\hat{\beta}] = \beta \frac{\cancel{\sum_i x_i^2}}{\cancel{\sum_i x_i^2}}$$

β, 2 ni=1x2i)

ii) $Var(\hat{\beta}) = Var\left(\frac{\sum_i x_i y_i}{\sum_i x_i^2}\right)$

$= \frac{1}{(\sum_i x_i^2)^2} Var(\sum_i x_i y_i)$ [$x_i \to$ known fixed constants]

$= \frac{1}{(\sum_i x_i^2)^{\cancel{2}}} \cancel{(\sum_i x_i^2)} Var(y_i)$ [$Var(Y_i) = \sigma^2 \to$ homogenity of variance]

$Var(\hat{\beta}) = \sigma^2 / \sum_i x_i^2$

Date / /

iii) $\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$

[$x_1, x_2 \ldots x_n \to$ known fixed constants]

- Since $\hat{\beta}$ is a linear combination of the independent normal random variables $y_1, y_2 \ldots y_n$, therefore, the distribution of $\hat{\beta}$ is normal.
- From Q3 b) i) $E[\hat{\beta}] = \beta$
- From Q3 b ii) $Var(\hat{\beta}) = \sigma^2 / \sum_{i=1}^{n} x_i^2$

So, $\hat{\beta} | x \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}\right)$

Problem 4

The web site www.playbill.comprovides weekly reports on the box office ticketsales for plays on Broadway in New York. We shall consider the data for the week October1117, 2004 (referred to below as the current week). The data are in the form of the gross boxoffice results for the current week and the gross box office results for the previous week (i.e.,October 310, 2004). The data are included in the fileplaybill.csv.Fit the following model to the data:Y= 0+ 1x+ewhereYis the gross box office resultsfor the current week (in)$andxisthegrossboxofficeresultsforthepreviousweek(in)$.

```
playbill=read.csv("playbill.csv")
model=lm(CurrentWeek~LastWeek, playbill)
```

Complete the following tasks: (a) Find a 95% confidence interval for the slope of the regression model, $\beta 1$. Is 1 a plausible value for $\beta 1$? Give a reason to support your answer.

```
library(ISwR)
confint(model, 'LastWeek', level=0.95)
```

```
##              2.5 %   97.5 %
## LastWeek 0.9514971 1.012666
```

From above, we can see that 1 lies within the 95% confidence interval and hence, 1 can certainly be a plausible $\beta 1$ value.

(b) Test the null hypothesis H0: $\beta 0$= 10000 against a two-sided alternative. Interpret your result.

```
confint(model, '(Intercept)', level=0.95)
```

```
##                 2.5 %  97.5 %
## (Intercept) -14244.33 27854.1
```

From above, we can see that intercept($\beta 0$) value of 10000 lies within the 95% confidence interval range.Hence, we do NOT reject the null hypothesis i.e. $\beta 0$= 10000.

(c) Use the fitted regression model to estimate the gross box office results for the current week (in \$) for a production with \$400,000 in gross box office the previous week. Find a 95% prediction interval for the gross box office results for the current week (in \$ for a production with \$400,000 in gross box office the previous week. Is \$450,000 a feasible value for the gross box office results in the current week, for a production with \$400,000in gross box office the previous week? Give a reason to support your answer.

```
predict(model,newdata=data.frame(LastWeek = 400000),interval="predict", level=.95)
```

```
##        fit      lwr      upr
## 1 399637.5 359832.8 439442.2
```

From above, we can see that 450,000 does not lie within 95% confidence interval and hence, \$450,000 is not a feasible value.

(d) Some promoters of Broadway plays use the prediction rule that next week's gross box office results will be equal to this weeks gross box office results. Comment on the appropriateness of this rule.

```r
summary(model)
```

```
##
## Call:
## lm(formula = CurrentWeek ~ LastWeek, data = playbill)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -36926  -7525  -2581   7782  35443
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.805e+03  9.929e+03   0.685    0.503
## LastWeek    9.821e-01  1.443e-02  68.071   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18010 on 16 degrees of freedom
## Multiple R-squared:  0.9966, Adjusted R-squared:  0.9963
## F-statistic:  4634 on 1 and 16 DF,  p-value: < 2.2e-16
```

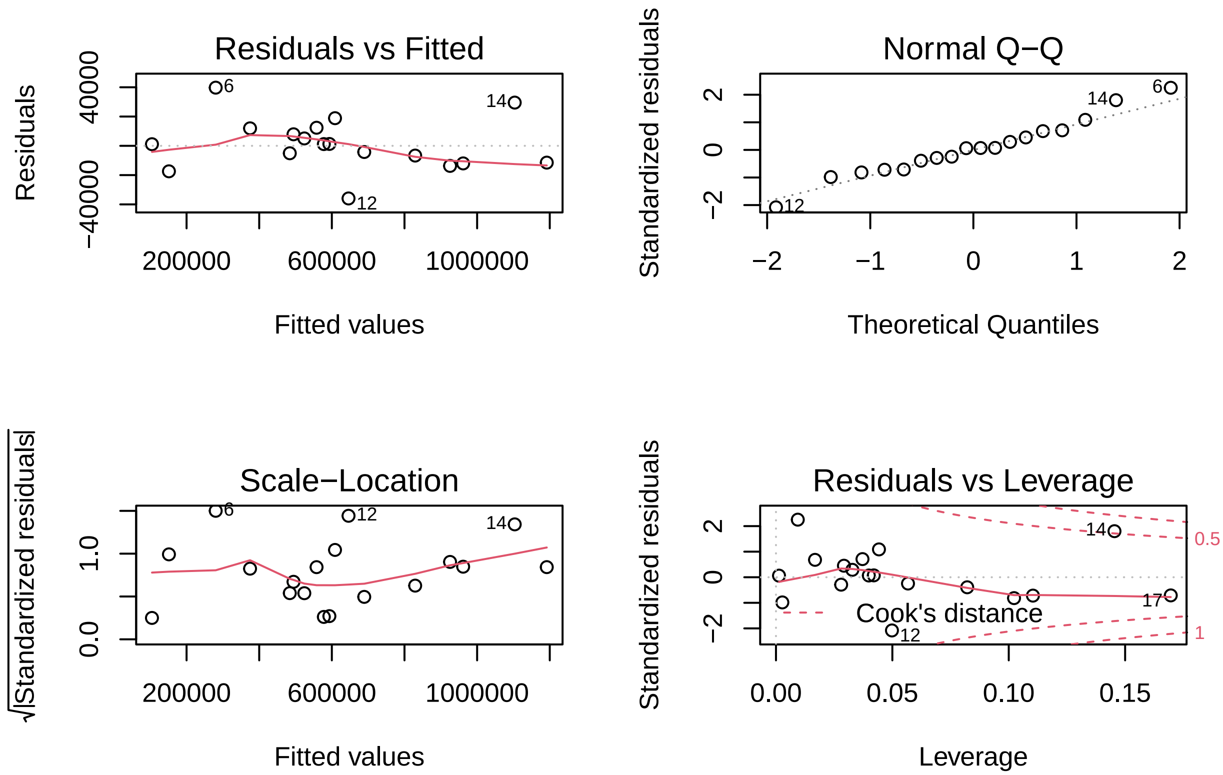
From the model summary, we can see that: a) the coefficient of 'LastWeek' is ~0.982 (close to 1) and the variable statistically significant in the prediction b) the 'Intercept' value is not significant c) High value of R-squared From here, we can say the Current Week value is very close to Last Week value.

To understand the promoters prediction rule better, lets create a linear regression model with no intercept.

```r
# Creating a linear regression model with no intercept
model2=lm(CurrentWeek~ LastWeek-1,data = playbill)
summary(model2)
```

```
##
## Call:
## lm(formula = CurrentWeek ~ LastWeek - 1, data = playbill)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -35948 -10271   1145  10936  39720
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## LastWeek  0.99102    0.00607   163.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17720 on 17 degrees of freedom
## Multiple R-squared:  0.9994, Adjusted R-squared:  0.9993
## F-statistic: 2.665e+04 on 1 and 17 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2, 2))
plot(model2)
```



From the above model we can see that the coefficient of 'LastWeek' is 0.99 (very close to 1). The model's R-squared value is also very high (0.9994). Hence, the promoters prediction rule is reasonably good.

Problem 5

In this problem we want to test that the identity:TSS=FSS+RSS. In order to do that, test the following identities:

(a) Show that (yi−ˆyi) = (yi−¯y)−ˆ 1(xi−¯xx)
(b) Show that (ˆyi−¯y) =ˆβ1(xi−¯x)
(c) Utilizing the fact thatˆ 1=SXYSXX, show that ni=1(yi−ˆyi)(ˆyi−¯y) = 0
(d) Finally test TSS=FSS+RSS

5.a) $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$\hat{y}_i - \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y}$

$= \cancel{\bar{y}} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \cancel{\bar{y}}$ [$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$]

$(\hat{y}_i - \bar{y}) = \hat{\beta}_1 (x_i - \bar{x})$ — (1)

$(y_i - \hat{y}_i) = (y_i - \hat{y}_i + \bar{y} - \bar{y})$

$= (y_i - \bar{y}) - (\hat{y}_i - \bar{y})$ [Replacing $(\hat{y}_i - \bar{y})$ from (1)]

$(y_i - \hat{y}_i) = (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})$

b) $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$\hat{y}_i - \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y}$

$= \cancel{\bar{y}} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \cancel{\bar{y}}$ [$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$]

$\hat{y}_i - \bar{y} = \hat{\beta}_1 (x_i - \bar{x})$

c) $\sum_{i=1} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$ $\quad [\hat{\beta}_1(x_i - \bar{x}) = \hat{y}_i - \bar{y}]$

$= \sum_i \hat{\beta}_1 (x_i - \bar{x})(y_i - \hat{y}_i)$

$= \hat{\beta}_1 \sum_i (x_i - \bar{x}) e_i$ $\quad [y_i - \hat{y}_i = e_i]$

$= \hat{\beta}_1 \left(\sum_i x_i e_i - \bar{x} \sum_i e_i \right)$ $\quad [\sum_i e_i = 0]$

$= \hat{\beta}_1 \sum_i x_i e_i$

$= \hat{\beta}_1 \sum_i x_i \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right)$

$= \hat{\beta}_1 \left(\sum_i x_i y_i - \hat{\beta}_0 \sum_i x_i - \hat{\beta}_1 \sum x_i^2 \right)$

$= \hat{\beta}_1 \left(\sum_i x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_i x_i - \hat{\beta}_1 \sum_i x_i^2 \right)$

$= \hat{\beta}_1 \left(\sum_i x_i y_i - \bar{y} \sum_i x_i + \hat{\beta}_1 \left(\bar{x} \sum_i x_i - \sum_i x_i^2 \right) \right)$

$= \hat{\beta}_1 \left(\sum_i x_i y_i - \frac{\sum_i x_i \sum_i y_i}{n} + \frac{S_{xy}}{S_{xx}} \left(\frac{(\sum_i x_i)^2}{n} - \sum_i x_i^2 \right) \right)$

$= \hat{\beta}_1 \left(\sum_i x_i y_i - \frac{\sum_i x_i \sum_i y_i}{n} + \frac{\sum_i x_i (y_i - \bar{y})}{\sum_i x_i (x_i - \bar{x})} \left(\frac{(\sum_i x_i)^2}{n} - \sum_i x_i^2 \right) \right)$

$= \hat{\beta}_1 \left(\sum_i x_i y_i - \frac{\sum_i x_i \sum y_i}{n} + \frac{\left(\sum_i x_i y_i - \frac{\sum_i x_i y_i}{n} \right)}{\cancel{\sum_i x_i^2 - (\sum x_i)^2/n}} \cancel{\left(\sum_i x_i^2 - (\sum_i x_i)^2/n \right)} (-1) \right)$

$= \hat{\beta}_1 \left(\cancel{\sum_i x_i y_i} - \cancel{\frac{\sum_i x_i \sum y_i}{n}} - \cancel{\sum_i x_i y_i} + \cancel{\frac{\sum_i x_i \sum y_i}{n}} \right)$

$= 0$

d) $TSS = \sum_i (y_i - \bar{y})^2$

$$= \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

$$= \sum_i \left[(y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2(y_i - \hat{y})(\hat{y}_i - \bar{y}) \right]$$

$$= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2 \cancelto{0}{\sum_i (y_i - \hat{y})(\hat{y}_i - \bar{y})}$$

$$= \sum_i r_i^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 0$$

$TSS = RSS + FSS$