Assignment 1 - ECON204(A)

Pakhi Chachra

2023-11-27

Question 1

Ans. (i) A **Simple Linear Regression** models allows to model a relationship between a dependent variable and independent variable. According to the question, the following information is given to us:

Given:

- Average Weekly Expenditure on Advertising $\bar{X}$: \$500
- Average Weekly Sales $\bar{Y}$: \$10,000
- New Expenditure on Advertising (X_i): \$750
- New Predicted Weekly Sales$\hat{y}$: ?

We know that the **Simple Linear Regression Function** to predict y ($\hat{y}$) is given as follows:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{1}$$

However, when the average weekly expenditures and weekly sales are provided to us, then we use $\bar{y}$ instead of $\hat{y}$.

So substituting values for the average weekly expenditure and average weekly sales, we get the Linear Regression Function as:

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \tag{2}$$
$$10000 = \hat{\beta}_0 + \hat{\beta}_1 .500 \tag{3}$$

Similarly, the function for predicted weekly sales when the expenditure is increased to `$750`, would be:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 .750 \tag{4}$$

From the above two equations, we can estimate the $\hat{\beta}_0$ and $\hat{\beta}_1$.

$\hat{\beta}_1$ and $\hat{\beta}_0$ are given as:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{5}$$
$$\hat{\beta}_0 = \bar{Y} - b_1 \bar{X} \tag{6}$$
$$\tag{7}$$

Thus, $\hat{\beta}_1$ and $\hat{\beta}_0$ =

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(750-500)(12000-10000)}{\sum_{i=1}^{n}(750-500)^2} \tag{8}$$

$$\hat{\beta}_1 = 8 \tag{9}$$

$$\hat{\beta}_0 = 10000 - b_1.500 \tag{10}$$

$$\hat{\beta}_0 = 6000 \tag{11}$$

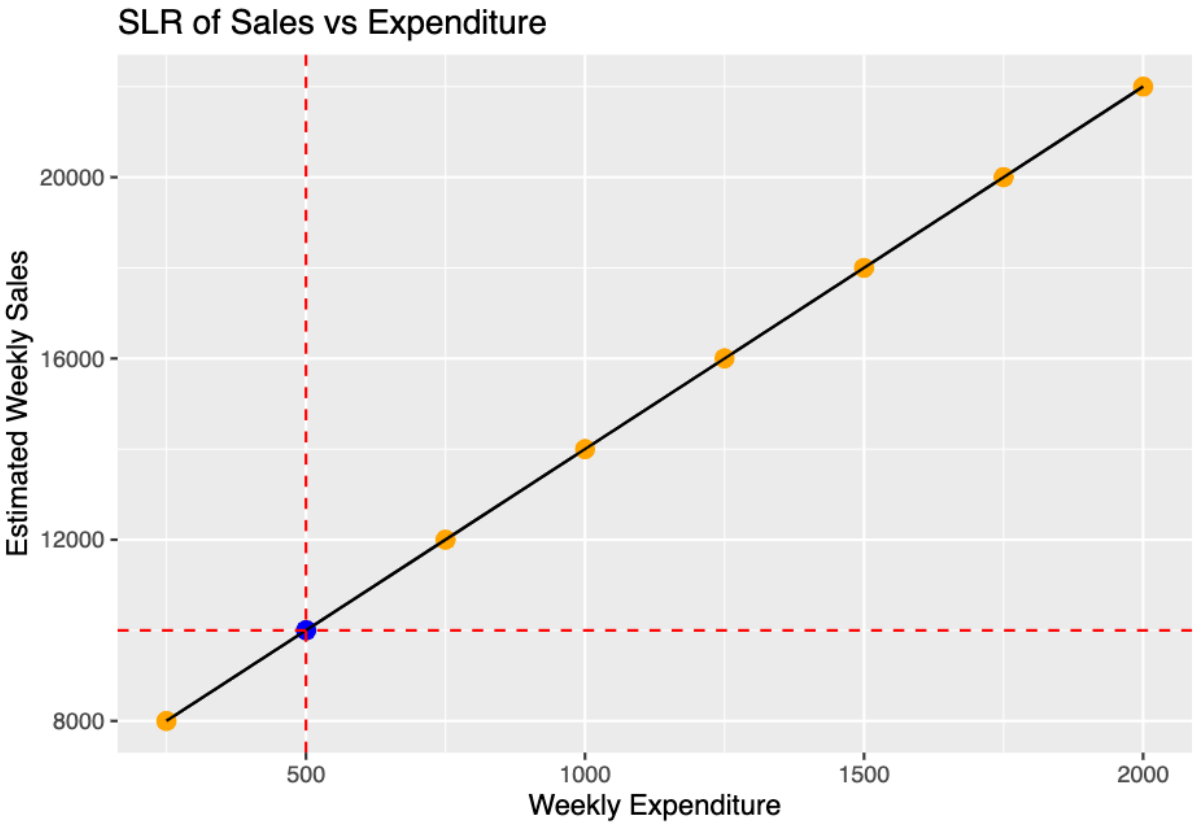
Thus, the estimated linear regression used by the consultant to make this prediction is:

$$\hat{y} = 6000 + 8x \tag{12}$$

(ii) Since we know the estimated linear regression function to use, we can construct a graph for the same. Our linear function is of the form $\hat{y} = 6000 + 8x$.We can begin by constructing a dataframe of x values ranging from 500 to 2000 by using `xlim(500,2000)` and then plot y as a function of x using `lm_eq <- function(x) 8 * x + 6000` and plotting `y=lm_eq`.

Before plotting, we set the working directory using `setwd()` and then continue with the problem:

```
setwd("/Users/pakhichachra/Desktop/econ support/ECON204")
library(ggplot2)
p <- ggplot(data = data.frame(x = 0), mapping = aes(x = x))
lm_eq <- function(x) 8 * x + 6000
p <- p + stat_function(fun = lm_eq) + xlim(250, 2000) +
  labs(x= "Average Weekly Expenditure", y= "Average Weekly Sales", title= "SLR of WS vs WE")
data_points <- data.frame(x= c(250, 500, 750, 1000, 1250, 1500, 1750, 2000), y= lm_eq(c(250, 500, 750,
p <- p + geom_point(data = data_points, aes(x = x, y = y), color = "orange", size = 3)+ labs(x= "Weekly
p <- p + geom_point(aes(x = 500, y = 10000), color = "blue", size = 3)
p <- p + geom_line(data = data_points, aes(x = x, y = y), color = "black")
p <- p + geom_vline(xintercept = 500, linetype = "dashed", color = "red")
p + geom_hline(yintercept = 10000, linetype = "dashed", color = "red")
```




```
knitr::opts_chunk$set(echo = TRUE)
```

In the graph above, the blue point signifies the weekly average values of sales $10,000 and weekly average value of expenditure $500.

Question 2

Ans. (i) *Given:*

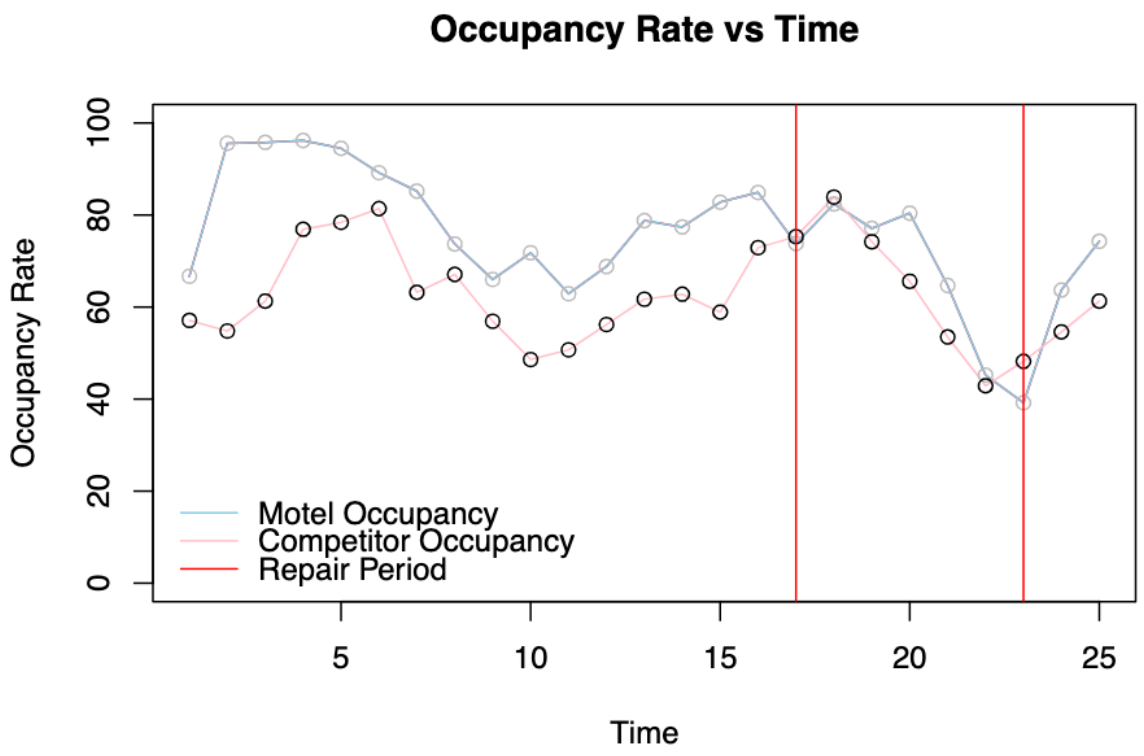
- Dataset `motel`: Variables `motel_pct` (damaged motel occupancy rates) and `comp_pct` (competitior occupancy rates)

We plot these variables on the same graph using basic R graphics function `plot()` and add arguments like `lines()` to plot the two graphs together against time and `points()` to enhance the readability and each data point of the graph. We also use the function `abline()` to indicate the repair period of 7 months, i.e., the area between the two lines indicate the ongoing repair period in the damaged motel.

```
library(readxl)
motel <- read_excel("motel.xlsx")
attach(motel)
Occupancy <- plot(motel$time, motel_pct, type="l", col="red", ylim= c(0,100), xlim= c(1,25),
                  xlab= "Time", ylab= "Occupancy Rate", main= "Occupancy Rate vs Time")
motel_occu <- lines(time, motel_pct, col="skyblue")
comp_occu <- lines(time, comp_pct, col="pink")
```

```
motel_occu_points <- points(time, motel_pct, col="gray")
comp_occu_points <- points(time, comp_pct, col="black")

#adding abline to indicate the repair period
abline(v= c(17,23), col= "red")
legend("bottomleft", inset = c(0, 0), legend = c("Motel Occupancy", "Competitor Occupancy", "Repair Per
         c("skyblue", "pink", "red"), lty = 1, bty= "n",x.intersp = 0.75, y.intersp = 0.75, text.font=
```




```
knitr::opts_chunk$set(echo = TRUE)
```

Thus, the space between the ablines at `x=17` and `x=23` indicate the repair period of 7 months between the 17th and 23rd month of the damaged motel.

To determine the higher occupancy rate before the repair period, we will subset the data from the original dataset setting `repair==0` and `time < 17` indicating a period of before repair and then calculate the means of occupancy rate of the damaged motel and the competitor motel.

```
occupancy_before_repair <-subset(motel, repair == 0 & time < 17)
attach(occupancy_before_repair)
mean(occupancy_before_repair$motel_pct)
```

```
## [1] 80.64375
```

```r
mean(occupancy_before_repair$comp_pct)
```

```
## [1] 63.05625
```

```r
knitr::opts_chunk$set(echo = FALSE)
```

Before repair period began, the average occupancy rate of the damaged motel was `80.64375` whereas the average occupancy rate of the competitor motel was `63.05625`. Hence, this suggests that the occupancy rate for the damaged motel was **higher** before the repair period.

To check the higher occupancy rate during the repair period, we will subset the data from the original dataset setting `repair==1` indicating that the repair is undergoing and then calculate the means of occupancy rate of the damaged motel and the competitor motel.

```r
occupancy_during_repair <- subset(motel, repair==1)
attach(occupancy_during_repair)
mean(occupancy_during_repair$motel_pct)
```

```
## [1] 66.11429
```

```r
mean(occupancy_during_repair$comp_pct)
```

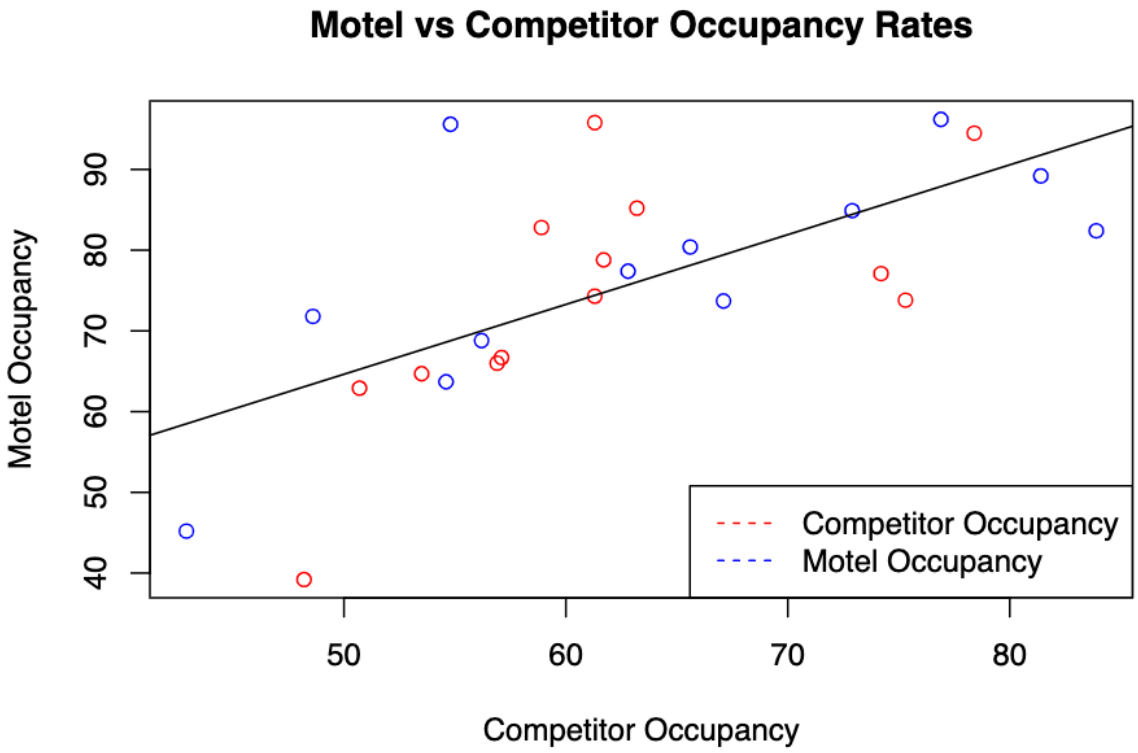
```
## [1] 63.37143
```

```r
knitr::opts_chunk$set(echo = FALSE)
```

From the means above, it is clear that the average occupancy rate of the damaged motel, i.e., `mean(motel_pct) = 66.11429` is higher than that of the competitor motel `mean(comp_pct) = 63.37143` during the repair period.

(ii) To check whether there is a relationship between two variables, we plot these variables against each other and analyze the trendline. Before doing so, we will plot `motel_pct` (y) against `comp_pct` (x).

```r
library(readxl)
motel <- read_excel("motel.xlsx")
attach(motel)
plot(x= comp_pct, y=motel_pct,col= c("red","blue"),xlab= "Competitor Occupancy", ylab= "Motel Occupancy
     main= "Motel vs Competitor Occupancy Rates")
abline(lm(motel_pct~comp_pct, data = motel), col= "black" )
legend("bottomright", legend= c("Competitor Occupancy", "Motel Occupancy"),col= c("red", "blue"), lty=2
```




```
knitr::opts_chunk$set(echo = FALSE)
```

Since, the linear regression line is upwards sloping, we can conclude that there seems to be a positive relationship between damaged motel occupancy rates vs competitor occupancy rates. That means, if the occupancy rates of competitor motels increase, then occupancy rate of the damaged motel also increase.

To check the strength of the relationship, we us ethe function `cor()` which specifies how strong the relationship between two variables is on a scale of -1 to +1.

```
attach(motel)
cor(comp_pct, motel_pct)
```

```
## [1] 0.6645726
```

```
knitr::opts_chunk$set(echo = FALSE)
```

The strength of the relationship is `0.6645726` which indicates that the positive relationship between motel occupancy rates and competitor occupancy rates is moderately and significantly strong.

To explain why such a relationship might exist, since if in an area, competitor motels and other motels exist, the overall traffic of tourists also increases. That means, there is an increase in the traffic of people coming to that area and hence already in an increase in the usual amount of people coming to stay at motels. Secondly, the other reason behind this can be that if the competitor motels charge a higher price, then the tourists can come to the damaged motel for their stay.

iii) To estimate a linear regression, we use the function `lm()` in R. This result gives the slope and intercept parameters of the linear regression which we then plug into our simple linear regression function

```
attach(motel)
reg_1 <- lm(motel_pct~comp_pct)
reg_1$coefficients[1]
```

```
## (Intercept)
##    21.39999
```

```
reg_1$coefficients[2]
```

```
##  comp_pct
## 0.8646393
```

```
knitr::opts_chunk$set(echo = FALSE)
```

Thus, the estimate will look like as follows:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \hat{x}_i \tag{13}$$

$$\hat{y} = 21.39999 + 0.8646393.\hat{x}_i \tag{14}$$

Interpretation of the slope: This indicates that with a unit increase in the occupancy rate of competitor motels, the occupancy rate in the damaged motel increases by 0.8646393.

Interpretation of the intercept: This indicates that when the occupancy rate of the competitor motels is 0, the occupancy rate of the damaged motel is 21.39999.

iv) Residuals are the difference between Actual Y value and the Predicted Y values in a linear regression model. The equation is given as follows:

$$\hat{u_i} = Y_i - \hat{Y_i} \tag{15}$$

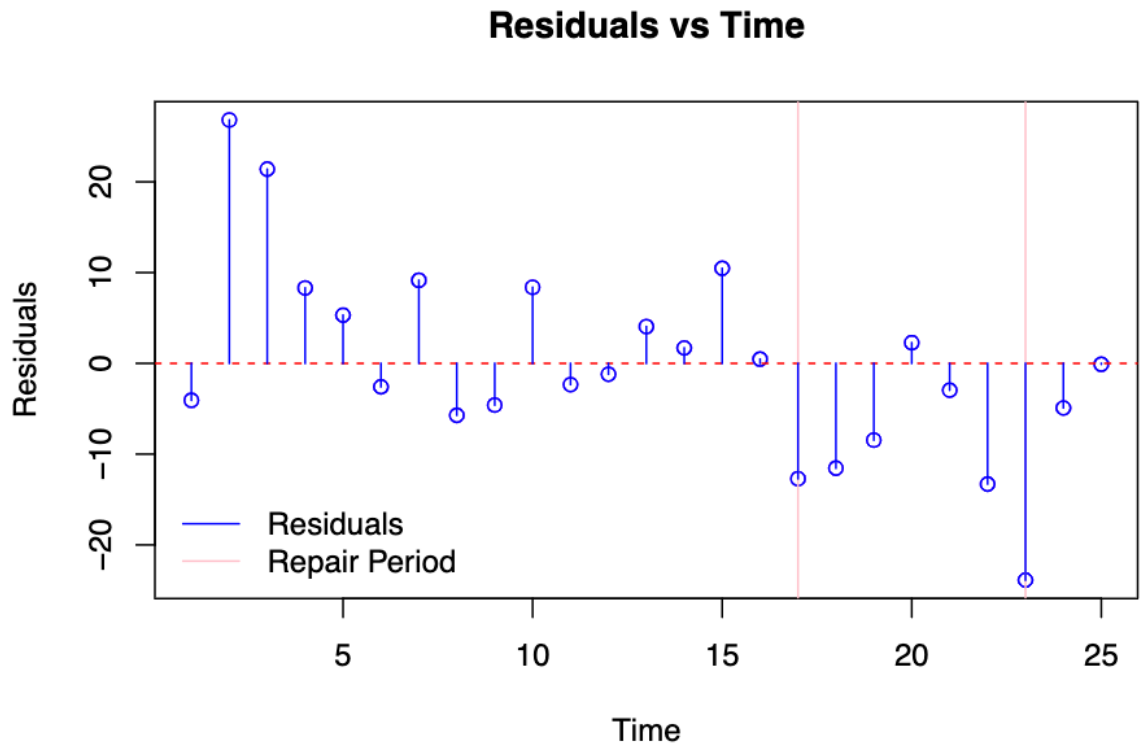
In order to plot residuals against time, we need to calculate the residuals in the linear regression model first. This is achieved by using the function `residuals(lm())` and then plot it against time using the `plot()` function.

```
library(readxl)
motel <- read_excel("motel.xlsx")
attach(motel)
reg_1 <- lm(motel_pct~comp_pct)
Least_Squares_Residuals <- residuals(reg_1)
Least_Squares_Residuals
```

```
##          1          2          3          4          5          6
## -4.0708929 26.8177776 21.3976220  8.3092488  5.3122899 -2.5816281
##          7          8          9         10         11         12
##  9.1548074 -5.7172859 -4.5979650  8.3785413 -2.3372013 -1.1927175
##         13         14         15         16         17         18
```

```
##   4.0517663   1.7006631  10.4727564   0.4678061 -12.7073283 -11.5432263
##          19          20          21          22          23          24
##  -8.4562250   2.2796730  -2.9581913 -13.2930147 -23.8756030  -4.9092946
##          25
##  -0.1023780
```

```
res_time <- plot(motel$time, Least_Squares_Residuals, type = "p", col = "purple", xlab = "Time", ylab =
res_time_points <- points(motel$time, Least_Squares_Residuals, col="blue")
abline(h = 0, col = "red", lty = 2)
repair_per <- abline(v= c(17,23), col= "pink")
segments(motel$time, 0, motel$time, Least_Squares_Residuals, col= "blue", lty=1)
legend("bottomleft", legend= c("Residuals", "Repair Period"),col= c("blue", "pink"), lty=1, bty= "n")
```




```
knitr::opts_chunk$set(echo = FALSE)
```

Result: There is an overprediction of the motel's occupancy rate during the repair period as we see that there are more points below the reference line and resultant residuals will be negative since actual motel occupancy values are smaller than the predicted motel occupancy values.

v) A linear regression with y = `motel_pct` and x = `relprice` demonstrates the relationship of the variables. If **Relprice**, which is the price per room charged by the motel in question relative to its competitor, is higher then the occupancy rate of the damaged motel will be lower since the customers would prefer to get a room which is cheaper. So, intuitively, the sign of the slope coefficient, i.e., $\hat{\beta}_1$ will be **negative.**

As the **relative price goes up**, the **occupancy rate** of the damaged motel **goes down**.

The same can be observed if a graph is plotted:

```
attach(motel)
reg_2 <- lm(motel_pct ~ relprice)
reg_2
```

```
##
## Call:
## lm(formula = motel_pct ~ relprice)
##
## Coefficients:
## (Intercept)     relprice
##       166.7       -122.1
```

```
knitr::opts_chunk$set(echo = FALSE)
```

In the output, the slope given by `reg_2$Coefficients[2]` is `-122.1` and is negative. Thus, the results align with our prediction and we can conclude that the estimated slope will be negative.

Question 3

Ans. (i) *Given*: - Earnings_and_Height Data

The median of any data provides the observation at the 50th percentile. To calculate the median, we use the function `median()`.

```
library(readxl)
Earnings_and_Height <- read_excel("Earnings_and_Height.xlsx")
attach(Earnings_and_Height)
median(height)
```

```
## [1] 67
```

```
knitr::opts_chunk$set(echo = FALSE)
```

Thus, the median height is 67 inches.

(ii) To calculate the average earnings of people who have a height atmost of 67 inches, we can subset the population with height less than or equal to 67 inches from the dataset using the function `subset()`. After doing so, we can calculate the mean of the earnings with this population that only consists of people with a height of **atmost** 67 inches.

```
atmost_67 <- subset(Earnings_and_Height, height <= 67)
mean(atmost_67$earnings)
```

```
## [1] 44488.44
```

```
knitr::opts_chunk$set(echo = FALSE)
```

Thus, the average earnings for workers whose height is at most 67 inches is $44488.44.

(iii) To calculate the average earnings of people who have a height greater 67 inches, we can subset the population with height more than 67 inches from the dataset using the function `subset()`. After doing so, we can calculate the mean of the earnings with this population that only consists of people with a height of **greater than** 67 inches.

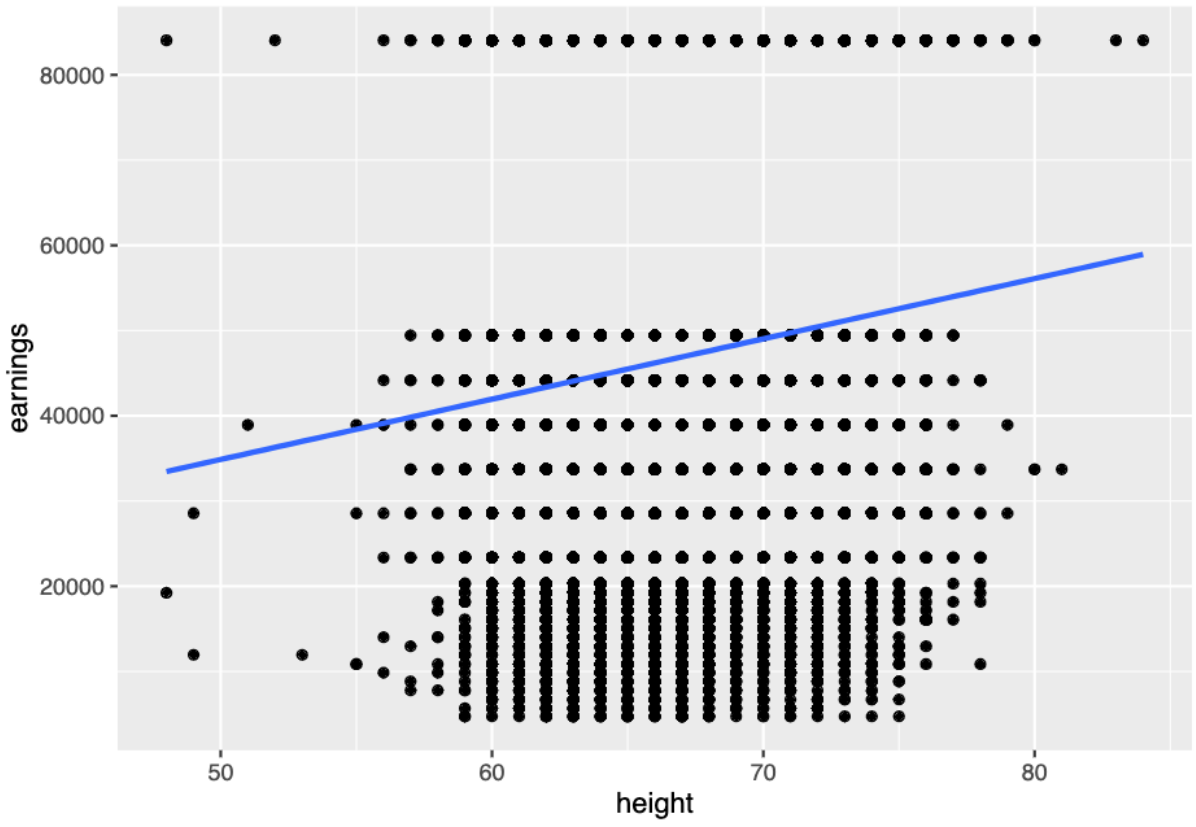
```
greater_67 <- subset(Earnings_and_Height, height > 67)
mean(greater_67$earnings)
```

```
## [1] 49987.88
```

Thus, the average earnings for workers whose height is greater than 67 inches is $49987.88.

(iv) To create a scatterplot of annual earnings on height, we take earnings on the y-axis and height on the x-axis. To create the same, we use the function `geom_point()`under the library `ggplot2`.

```
attach(Earnings_and_Height)
library(ggplot2)
ggplot(Earnings_and_Height, aes(x= height, y= earnings)) + geom_point() + geom_smooth(method = "lm", se
```



From the above scatterplot, we notice that the points are very far apart from the best fit line. This indicates a weak relationship between Earnings and Height.

(v) To calculate the SLR and the standard errors, we use the function of `lm()` of earnings on height (with earnings on the y-axis and height on the x-axis) to execute the same.

```
reg_earn_height <- lm(earnings~height)
summary(reg_earn_height)$coefficients
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -512.7336 3386.85615 -0.1513892 8.796704e-01
## height       707.6716   50.48922 14.0162889 2.129867e-44
```

Thus, since we can see from the output that $\hat{\beta_0}$ = -512.73 and $\hat{\beta_1}$ = 707.67. Thus, SLR can be given as:

$$\hat{y} = -512.73 + 707.67x + \hat{u} \tag{16}$$

The **standard error** associated with $\hat{\beta_0}$ is 3386.85615 and height ($\hat{\beta_1}$) is 50.48922.

a) The estimated slope is equal to $\hat{\beta_1}$ and is equal to **707.6716.**

b) 1. *Predicting earnings someone who is 67 inches tall*

$$\hat{y} = \hat{\beta_0} + \hat{\beta_1}x \tag{17}$$
$$\hat{y} = -512.7336 + 707.6716X67 \tag{18}$$
$$\hat{y} = 46,901.2636 \tag{19}$$

The predicted earnings for someone who is 67 inches tall is `$46,901.2636`.

2. *Predicting earnings someone who is 70 inches tall*

$$\hat{y} = \hat{\beta_0} + \hat{\beta_1}x \tag{20}$$
$$\hat{y} = -512.7336 + 707.6716X70 \tag{21}$$
$$\hat{y} = 49,024.2784 \tag{22}$$

The predicted earnings for someone who is 70 inches tall is `$49,024.2784`.

3. *Predicting earnings someone who is 65 inches tall*

$$\hat{y} = \hat{\beta_0} + \hat{\beta_1}x \tag{23}$$
$$\hat{y} = -512.7336 + 707.6716X65 \tag{24}$$
$$\hat{y} = 45,485.9204 \tag{25}$$

The predicted earnings for someone who is 65 inches tall is `$45,485.9204`.

c) To calculate the value R^2 we use the regression stored in the variable `reg_earn_height`. After doing so, we take the summary of this variable and call upon the value of R^2 which is already a part of the output using a dollar sign like: `summary(reg_earn_height)$r.squared`.

```r
attach(Earnings_and_Height)
reg_earn_height <- lm(earnings~height)
summary(reg_earn_height)$r.squared
```

```
## [1] 0.0108753
```

Thus, the value of R^2 is 0.01087538.

vi) In order to check whether height is uncorrelated with other factors that cause earning, we can build correlation matrix of the dataset "Earnings_and_Height".

```r
cor(Earnings_and_Height)
```

```
##                    sex          age          mrd         educ      cworker
## sex         1.000000000 -0.009223857  0.007184516  0.006921959  0.02467021
## age        -0.009223857  1.000000000 -0.135604032 -0.054165810  0.13413142
## mrd         0.007184516 -0.135604032  1.000000000  0.041262942 -0.06279348
## educ        0.006921959 -0.054165810  0.041262942  1.000000000  0.12863400
## cworker     0.024670211  0.134131424 -0.062793482  0.128633997  1.00000000
## region      0.032760781 -0.021519056  0.017787332 -0.017504684  0.02553214
## race        0.018051224 -0.060432769  0.051250735 -0.122529491 -0.03847701
## earnings    0.052363075  0.100367178 -0.349826028  0.387966598  0.08116608
## height      0.699878410 -0.044797860  0.017139693  0.115682722  0.03516452
## weight      0.285128007  0.075667012 -0.021433702 -0.016816916  0.03424116
## occupation  0.246388673 -0.011743443 -0.010664353 -0.494219124 -0.09744557
##                  region         race     earnings       height       weight
## sex         0.0327607814  0.01805122  0.05236308  0.69987841  0.2851280072
## age        -0.0215190559 -0.06043277  0.10036718 -0.04479786  0.0756670118
## mrd         0.0177873322  0.05125073 -0.34982603  0.01713969 -0.0214337018
## educ       -0.0175046842 -0.12252949  0.38796660  0.11568272 -0.0168169162
## cworker     0.0255321408 -0.03847701  0.08116608  0.03516452  0.0342411569
## region      1.0000000000  0.17171918 -0.04009433  0.01943432  0.0001038323
## race        0.1717191820  1.00000000 -0.12847725 -0.15806364 -0.0573196714
## earnings   -0.0400943263 -0.12847725  1.00000000  0.10428470  0.0191760144
## height      0.0194343231 -0.15806364  0.10428470  1.00000000  0.3742334978
## weight      0.0001038323 -0.05731967  0.01917601  0.37423350  1.0000000000
## occupation -0.0069050725  0.06740642 -0.30055328  0.11828985  0.0848495990
##              occupation
## sex         0.246388673
## age        -0.011743443
## mrd        -0.010664353
## educ       -0.494219124
## cworker    -0.097445566
## region     -0.006905072
## race        0.067406424
## earnings   -0.300553278
## height      0.118289850
## weight      0.084849599
## occupation  1.000000000
```

After building the correlation matrix, we observe that **height** is correlated with other variables in the data set to different degrees with the most being with **sex** that of `0.69987841` and with **mrd** that of `0.01713969`.

Similarly, these factors are also correlated with **earnings** to some degree like a correlation of `0.05236308` with **sex** and `-0.34982603` with **mrd**. This means that there are external factors that impact both earnings and height and thus, there is a bias in the residuals that is not explained by the independent variable in the model (**omitted variable bias**). Hence, the regression error term does not have a conditional mean of zero given height (X_i).

Question 4

(i) Given to us, is the information on Districts in Columbia with 51 observations and the estimated error variance

$$\hat{\sigma} = 2.04672$$

.

Mathematically, the estimated error variance is given by the following equation:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} \tag{26}$$

Substituting values of the error variance:

$$2.04672 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{51-2} \tag{27}$$

$$2.04672 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{49} \tag{28}$$

$$100.28928 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{29}$$

The expression $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the **sum of squared residuals** and is equal to `100.28928`.

(ii) Given:

- Estimated variance of $\beta_2 = 0.00098$
- Estimated error of $\beta_2 =$ 'sqrt(var(β_2))

Thus,the estimated standard error of β_2 is:

$$\hat{se} = \sqrt{0.00098} \tag{30}$$

Thus, the estimated standard error of β_2 is **0.03130495.**

Mathematically, the standard error of β_2 is given by:

$$\hat{se}(\hat{\beta}_2) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{31}$$

Substituting values to find the value of $\sum_{i=1}^{n}(x_i - \bar{x})^2$

Rearranging and solving the equation:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \hat{\sigma}^2 \times \frac{1}{SE(\hat{\beta}_1)^2} \tag{32}$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 2.04672 \times \frac{1}{0.00098} \tag{33}$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 2088.4898 \tag{34}$$

Thus, $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 2088.4898$.

(iii) Given: -y_i = the state's mean income of males who are 18 and older

- x_i = percentage of males 18 years or older who are high school graduates
- $\beta_2 = 0.18$

Mathematically,

$$\hat{y} = \beta_0 + \beta_2 x \tag{35}$$

But here $\beta_0 = 0$ since when there are no men earning, there will no income generated. So,

$$\hat{y} = \beta_2 x \tag{36}$$

$$\hat{y} = \beta_2 x \tag{37}$$

$$\hat{y} = 0.18x \tag{38}$$

$$\tag{39}$$

Interpretation: This result indicates that when there is a unit increase in the percentage of males of 18 years and above who are high school graduates, the state's mean income of males who are 18 and older increases by $180.

iv. Given:

$$\bar{x} = 69.139 \tag{40}$$

$$\bar{y} = 15.187 \tag{41}$$

$$\beta_2 = 0.18 \tag{42}$$

$$\tag{43}$$

Mathematically the equation is represented as:

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_2 \bar{x} \tag{44}$$

Substituting the values:

$$15.187 = \hat{\beta}_0 + 0.18 * 69.139 \tag{45}$$

$$2.74198 = \hat{\beta}_0 \tag{46}$$

Thus, the estimate of the intercept parameter, $\hat{\beta}_0$ is 2.74198.

v. Given:

$$\bar{x} = 69.139 \tag{47}$$

$$\sum_{i=1}^{n}(x_i - \bar{x}^2 = 2088.4898 \tag{48}$$

Solution:

Mathematically expanding upon $\sum_{i=1}^{n}(x_i - \bar{x})^2$:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i^2 + \bar{x}^2 - 2x_i\bar{x}) \tag{49}$$

$$= \sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_{i=1}^{n} x_i \tag{50}$$

$$= \sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2\bar{x}n\bar{x}^2 \tag{51}$$

$$= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \tag{52}$$

$$= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \tag{53}$$

$$2088.4898 = \sum_{i=1}^{n} x_i^2 - (51 * 69.139^2) \tag{54}$$

$$245878.7572 = \sum_{i=1}^{n} x_i^2 \tag{55}$$

$$\tag{56}$$

Thus, $\sum_{i=1}^{n} x_i^2 = 245878.7572$

(vi) Given

- $x_i = 58.3$
- $y_i = 12.274$
- $\hat{\beta}_0 = 2.74198$
- $\hat{\beta}_2 = 0.18$

In order to compute the least square residuals, we need to compute the predicted y, i.e., $\hat{y}_i$.

Mathematically, $\hat{y}_i$ is computed by Simple Linear Regression Function:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{57}$$

Substituting the values given:

$$\hat{y} = 2.74198 + 0.18 * 58.3 \tag{58}$$

$$\hat{y} = 13.23598 \tag{59}$$

$$\tag{60}$$

Thus, the predicted value for y ($\hat{y_i}$) is 13.23598.

Now, SSR is computed in the following manner

$$\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 \tag{61}$$

$$\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 \tag{62}$$

$$\sum_{i=1}^{n} u_i^2 = (12.274 - 13.23598)^2 \tag{63}$$

$$\sum_{i=1}^{n} u_i^2 = 0.9254055 \tag{64}$$

$$\tag{65}$$

Thus, the SSR or $\sum_{i=1}^{n} u_i^2$ is equal to **0.9254055**.

Question 5

(i) Given:

- $\hat{\beta}_{YX}$ = Y regression on X
- $\hat{\beta}_{YX}$ = X regression on Y

First we represent $\hat{\beta}_{YX}$ in a slope regression function as follows:

$$\hat{\beta}_{YX} = r\frac{s_y}{s_x} \tag{66}$$

Then we represent $\hat{\beta}_{XY}$ in a slope regression function as follows:

$$\hat{\beta}_{XY} = r\frac{s_x}{s_y} \tag{67}$$

$\hat{\beta}_{YX}$ is calculated by the formula:

$$\hat{\beta}_{YX} = r\frac{s_y}{s_x} \tag{68}$$

Similarly, $\hat{\beta}_{XY}$ is calculated by the formula:

$$\hat{\beta}_{XY} = r\frac{s_x}{s_y} \tag{69}$$

After this, we take the product of equations 58 and 59 and we get:

$$\hat{\beta}_{YX}\hat{\beta}_{XY} = r\frac{s_y}{s_x} \cdot r\frac{s_x}{s_y} \tag{70}$$

s_x and s_y cancel out

Hence, we have proved that:

$$\hat{\beta}_{YX}\hat{\beta}_{XY} = r^2 \tag{71}$$

(ii) If the product of the slopes $\hat{\beta}_{YX}\hat{\beta}_{XY} = 1$, it implies a perfect linear relationship, and the choice of which variable is considered independent or dependent does not matter.The interpretation of the intercepts will differ between the two regressions, but the overall relationship and the prediction of changes in one variable based on changes in the other will be the same.