Assignment-1

Shraddha Ganesh

2023-11-27

Problem 1

i)

The notation of a simple linear regression function is used by the consultant to make this prediction. While the values of the slope and the intercept are unknown, the independent and the dependent variables are weekly advertisement expenditure and weekly sales respectively. We can, however, solve for the intercept and the slope by solving the two equations given below.

Equation 1: $10,000 = \beta_0 + \beta_1 * 500$

Equation 2: $12,000 = \beta_0 + \beta_1 * 750$

From these two equations, **slope = 8.** By substituting this value in equation in Equation 1, we get **intercept as 6,000.**

Therefore, the estimated simple regression used by the consultant to make this prediction is as follows.

Estimated Equation: $y = 6000 + 8x$

OR, $\text{predicted. weekly. sales} = 6000 + 8 \cdot \text{weekly. advertisement. expenditure}$

ii)

```
install.packages("ggplot2",repos = "http://cran.us.r-project.org")
```

```
## package 'ggplot2' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\shrad\AppData\Local\Temp\Rtmpk3r37R\downloaded_packages
```

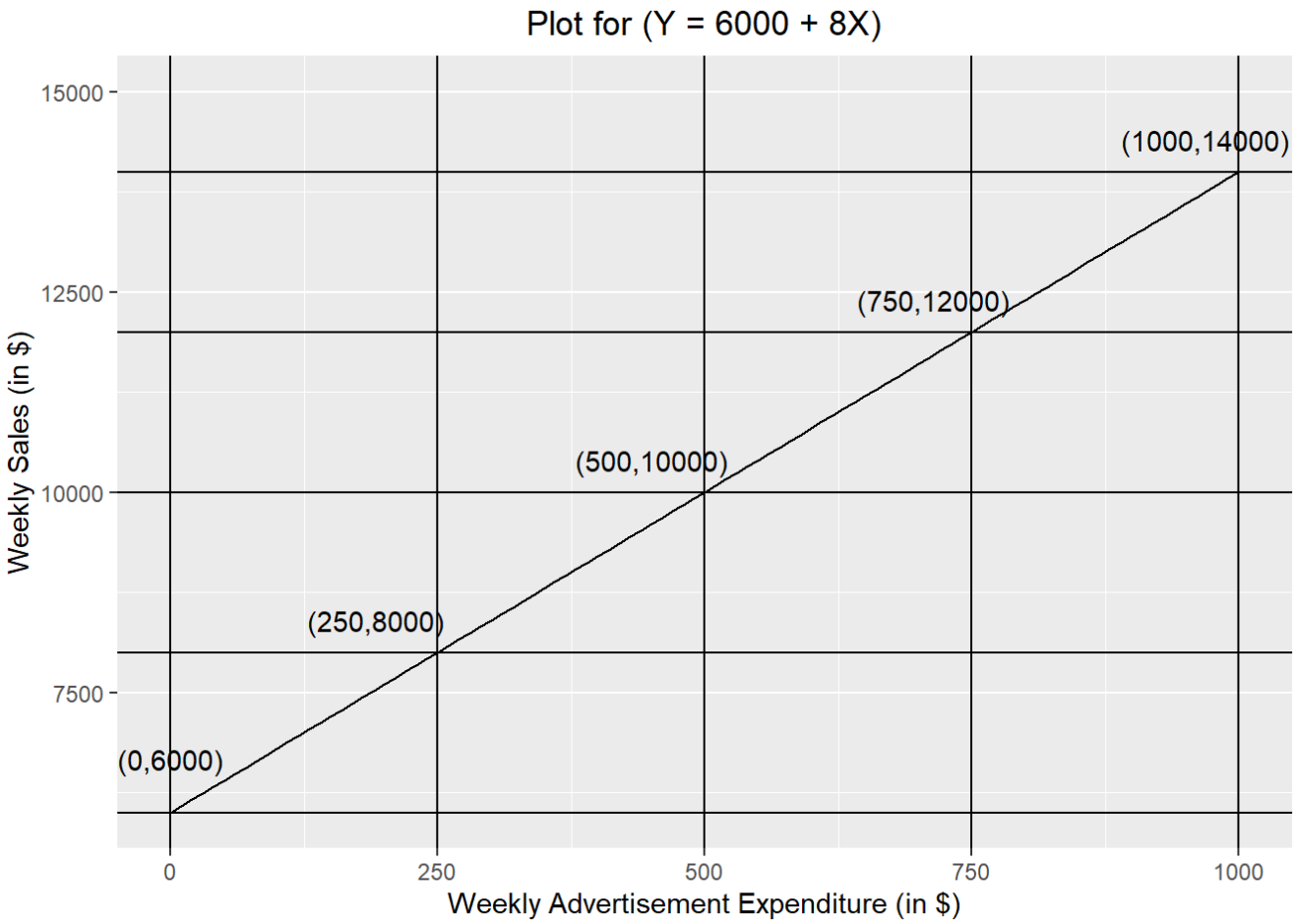
```
library(ggplot2)
```

Graphing the equation:

```
reg_graph = ggplot(data = data.frame(x = 0), mapping = aes(x = x))
lm_eq = function(x){6000 + 8*x}
reg_graph = reg_graph + stat_function(fun = lm_eq) + xlim(0, 1000)+
  ylim(6000,15000)
reg_graph = reg_graph + labs(title = "Plot for (Y = 6000 + 8X)",
                 x = "Weekly Advertisement Expenditure (in $)",
                 y = "Weekly Sales (in $)")
reg_graph = reg_graph + theme(plot.title = element_text(hjust = 0.5))
```

Locating the average weekly values:

```
y1 = 6000
reg_graph = reg_graph+
  geom_vline(xintercept = 0)+
  geom_hline(yintercept = 6000)+
  geom_text(aes(0,y1,label = "(0,6000)",vjust = -2))

y2 = 8000
reg_graph = reg_graph+
  geom_vline(xintercept = 250)+
  geom_hline(yintercept = 8000)+
  geom_text(aes(0,y2,label = "(250,8000)",vjust =-1,hjust=-1))

y3 = 10000
reg_graph = reg_graph+
  geom_vline(xintercept = 500)+
  geom_hline(yintercept = 10000)+
  geom_text(aes(0,y3,label = "(500,10000)",vjust = -1,hjust=-2.65))

y4 = 12000
reg_graph = reg_graph+
  geom_vline(xintercept = 750)+
  geom_hline(yintercept = 12000)+
  geom_text(aes(0,y4,label= "(750,12000)",vjust = -1,hjust=-4.5))

y5 = 14000
reg_graph = reg_graph+
  geom_vline(xintercept = 1000)+
  geom_hline(yintercept = 14000)+
  geom_text(aes(0,y5,label= "(1000,14000)",vjust=-1,hjust=-5.65))

reg_graph
```



Problem 2

```
install.packages("patchwork",repos = "http://cran.us.r-project.org")
```

```
## package 'patchwork' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\shrad\AppData\Local\Temp\Rtmpk3r37R\downloaded_packages
```

```
library(patchwork)
library(readxl)
motel = read_excel("C:/Users/shrad/Downloads/motel.xlsx")
```

i)

```
df1 = motel[,1]
df2 = motel[,5]
df3 = motel[,7]
newdf1 = data.frame(cbind(df1,df2,df3))
final_repa = data.frame(newdf1[17:23, 1:2])
final_not_repa = data.frame(newdf1[1:16,1:2])
comp_not_repa = motel[1:16,6]
comp_repa = data.frame(motel[17:23,6])
during_repair = cbind(final_repa,comp_repa)
not_repair = cbind(final_not_repa,comp_not_repa)
```

Creating A Single Plot:

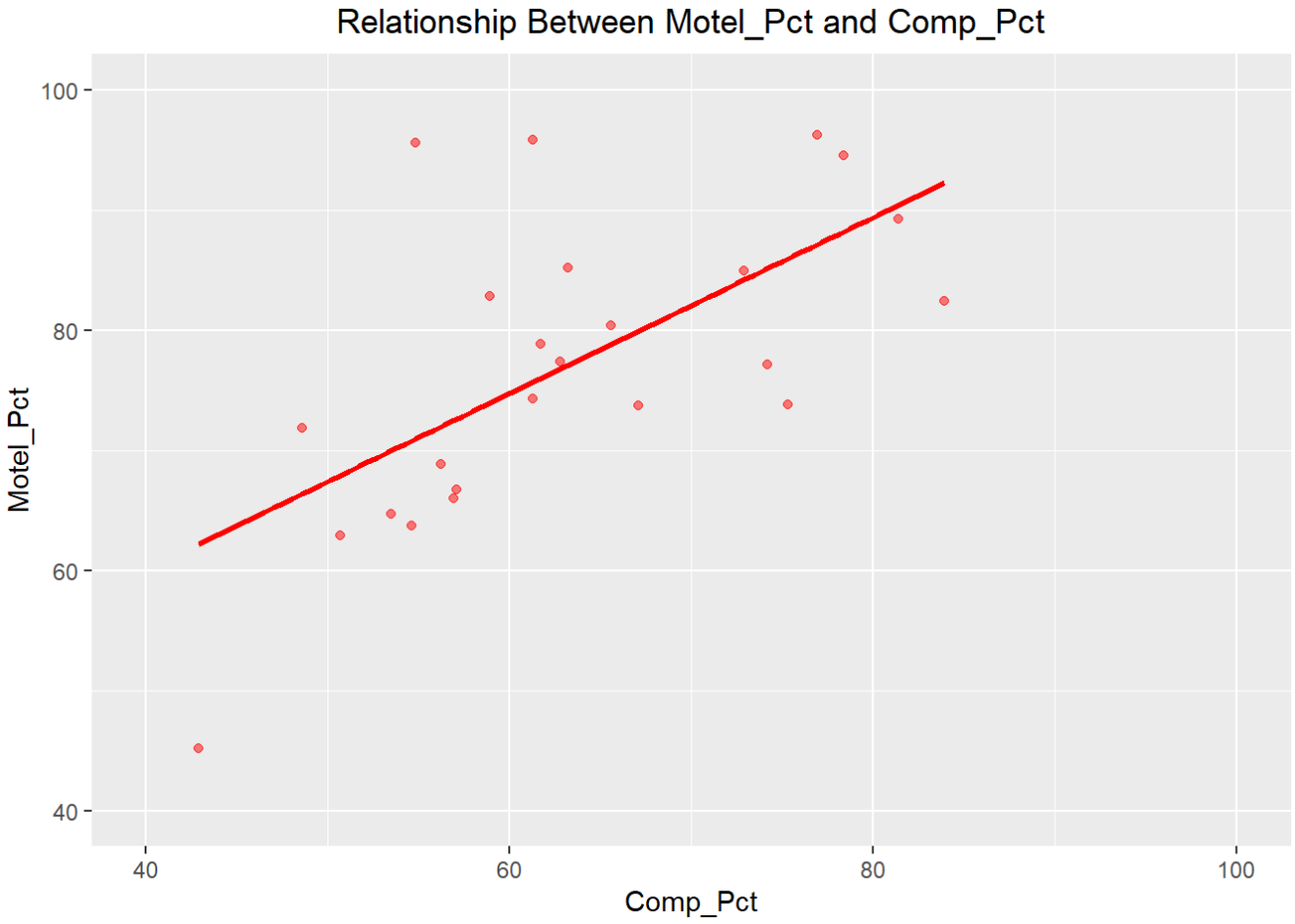
```
plot1 = gg=ggplot(data=not_repair,mapping=aes(x=time))+
  geom_point(mapping=aes(y=motel_pct,color="motel_pct"))+
  geom_line(mapping=aes(y=motel_pct,color="motel_pct"))+
  geom_point(mapping=aes(y=comp_pct,color="comp_pct"))+
  geom_line(mapping=aes(y=comp_pct,color="comp_pct"))+
  xlim(0,25)+
  geom_vline(xintercept = 0.5)+
  geom_vline(xintercept = 16.5)+
  labs(title="Occupany Rates Before Repair",x="Time (month)",y="Occupany Rate")+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_color_discrete(name="Motel")

plot2 = gg=ggplot(data=during_repair,mapping=aes(x=time))+
  geom_point(mapping=aes(y=motel_pct,color="motel_pct"))+
  geom_line(mapping=aes(y=motel_pct,color="motel_pct"))+
  geom_point(mapping=aes(y=comp_pct,color="comp_pct"))+
  geom_line(mapping=aes(y=comp_pct,color="comp_pct"))+
  xlim(0,25)+
  geom_vline(xintercept = 16.5)+
  geom_vline(xintercept = 23.5)+
  labs(title="Occupany Rates During Repair",x="Time (month)",y="Occupany Rate")+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_color_discrete(name="Motel")

final_plot = plot1/plot2
final_plot
```

Interpretation:

For before the repair period - During the pre-repair, the damaged hotel has has consistently higher occupancy rates than its competitor hotel. For during the repair period - During the repair period, for the first two first two months and the last month (17th, 18th, and 23rd), the competitors firm had higher occupancy rates than the damaged hotel. However, for the remaining in between 4 months (19th,20th,21st,and 22nd),the competitor firm has higher occupancy rates than the damaged hotel.

ii)

```
motel_pct = data.frame(motel[,5])
comp_pct = data.frame(motel[,6])
plot=ggplot(motel,aes(x=comp_pct,y=motel_pct))+
  geom_point(color="red",alpha=.5)+
  geom_smooth(method="lm",se=FALSE,color="red")+
  labs(title="Relationship Between Motel_Pct and Comp_Pct", x = "Comp_Pct",
       y = "Motel_Pct")+theme(plot.title = element_text(hjust = 0.5))+
  xlim(40,100)+ylim(40,100)
plot
```



Interpretation:

There is a positive relationship between compt_pct and motel_pct. Therefore, as compt_pct increases, so does motel_pct. Increasing occupany among competitors may indicate general growth in demand for motels. While there is a correlation between the damaged motel's occupancy and its competitors occupancy, we cannot establish causation with certainty. the motel industry in general seems to have higher occupancy rates as demand for motels increases.

iii)

```
rig_line = lm(motel_pct~comp_pct,motel)
coef=coef(rig_line)
coef
```

```
## (Intercept)    comp_pct
##  21.3999883   0.8646393
```

Estimated Linear Regression Equation: $\mathrm{MOTEL.PCT} = 21.4 + 0.86 \cdot \mathrm{COMP.PCT}$

Interpretation:

The slope is 0.86, therefore, if the competitors occupancy increases by 1%,the damaged motel's occupancy increases by 0.86%. The intercept is 21.4,therefore, when the competitor's occupancy is 0, the damaged motel's occupancy will be 21.4.

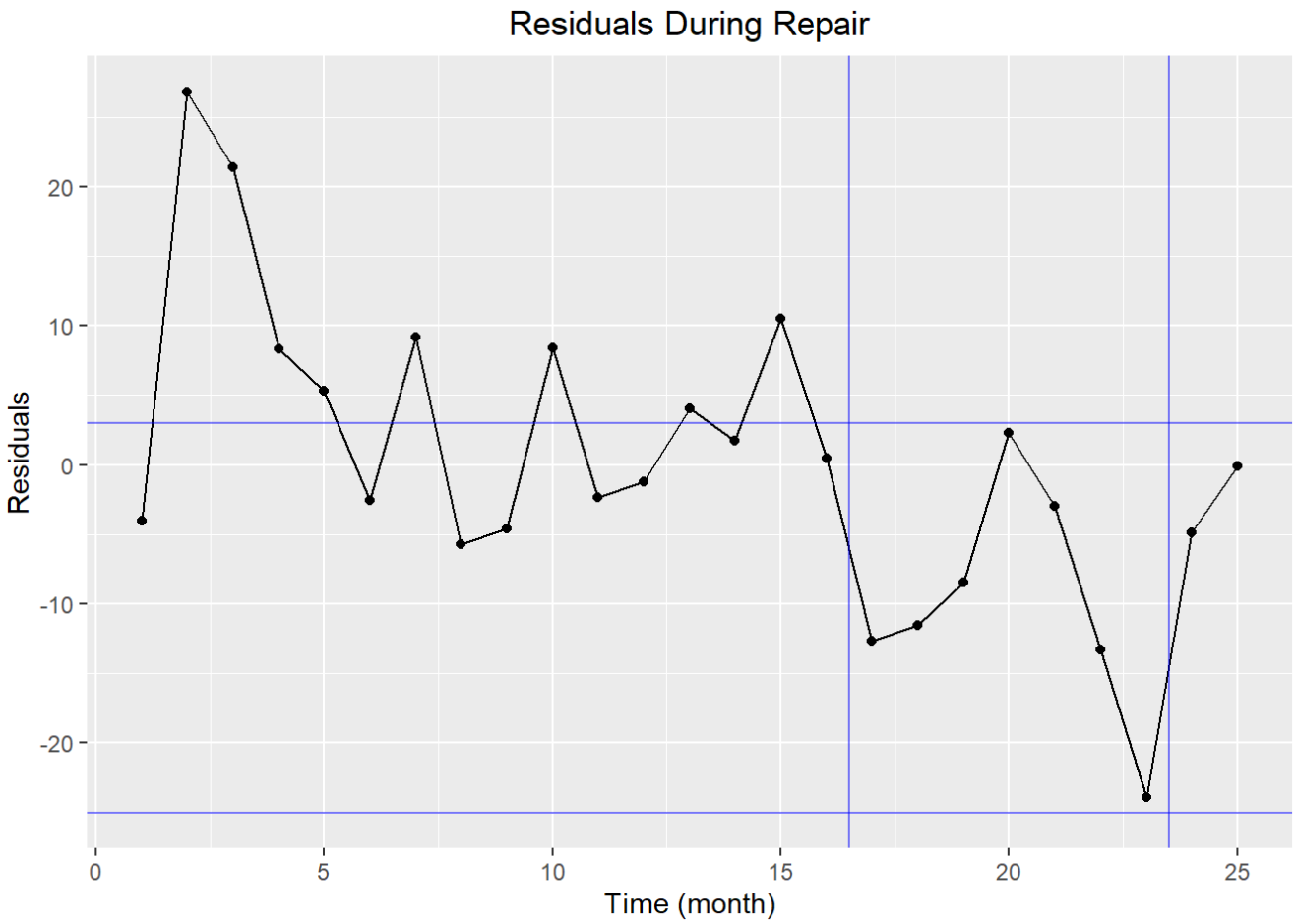
iv)

Computing and adding Residuals to the data set motel:

```
rig_line = lm(motel_pct~comp_pct,motel)
residuals = rig_line$residuals
motel$residuals = residuals
```

Plotting residuals against time.

```
res_plot = ggplot()+
  geom_point(motel,mapping = aes(x=time,y=residuals))+
  geom_line(motel,mapping = aes(x=time,y=residuals))+
  geom_vline(xintercept = 16.5,color="blue",alpha=.5)+
  geom_vline(xintercept = 23.5,color="blue",alpha=.5)+
  geom_hline(yintercept = -25,color="blue",alpha=.5)+
  geom_hline(yintercept = 3,color="blue",alpha=.5)+
  labs(title="Residuals During Repair",
       x="Time (month)",
       y="Residuals")+
  theme(plot.title = element_text(hjust = 0.5))
res_plot
```

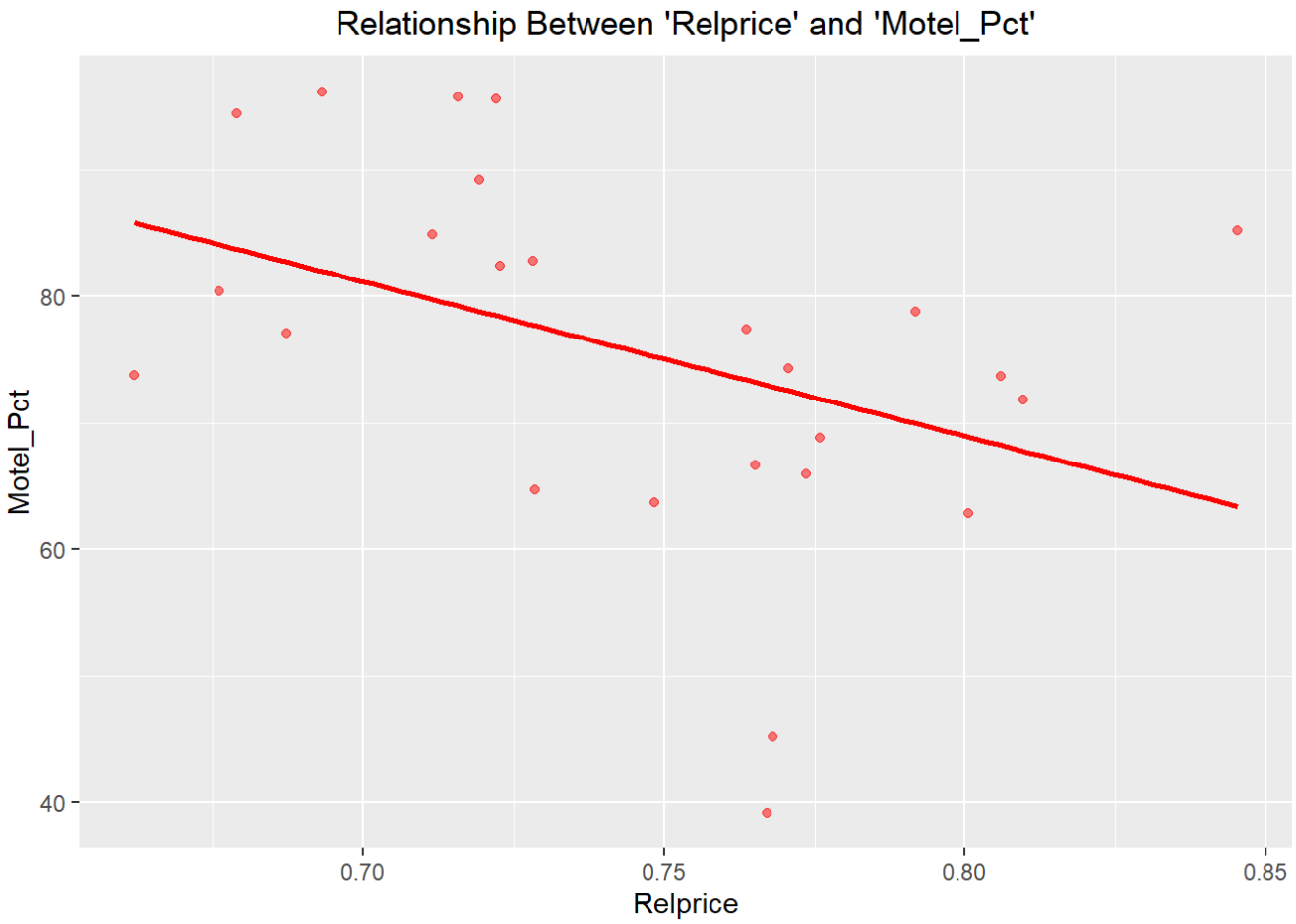


Interpretation:

For all months except the 20th, the residuals have been negative. Therefore, the occupancy rates of the motel was overestimated by the model. The model has, however, under predicted the motel's occupancy rate on the the 20th month by a small margin.

v)

```
p = ggplot(motel,aes(x=relprice,y=motel_pct))+
  geom_point(color="red",alpha=.5)+
  geom_smooth(method="lm",se=FALSE,color="red",alpha=.5)+
  labs(title="Relationship Between 'Relprice' and 'Motel_Pct'",
       x="Relprice",
       y="Motel_Pct")+
  theme(plot.title = element_text(hjust = 0.5))
p
```



Interpretation:

The sign for the slope is negative. Because the estimated regression line is sloping downwards indicating a negative relationship between Relprice and Motel_Pct.

```
new = lm(motel_pct~relprice,motel)
co=coef(new)
co
```

```
## (Intercept)    relprice
##    166.6560   -122.1186
```

$$\text{MOTEL. PCT} = 166.66 + -122.12 \cdot \text{Relprice}$$

Yes, the sign of the estimated slope agrees with our expectation from judging the graph.

Problem 3

```
eh = read_excel("C:/Users/shrad/Downloads/Earnings_and_Height.xlsx")
```

i)

```
median = median(eh$height)
median
```

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

ii)

```
d1 = data.frame(eh[eh$height<=67,8:9])
mean1 = mean(d1[,1])
mean1
```

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

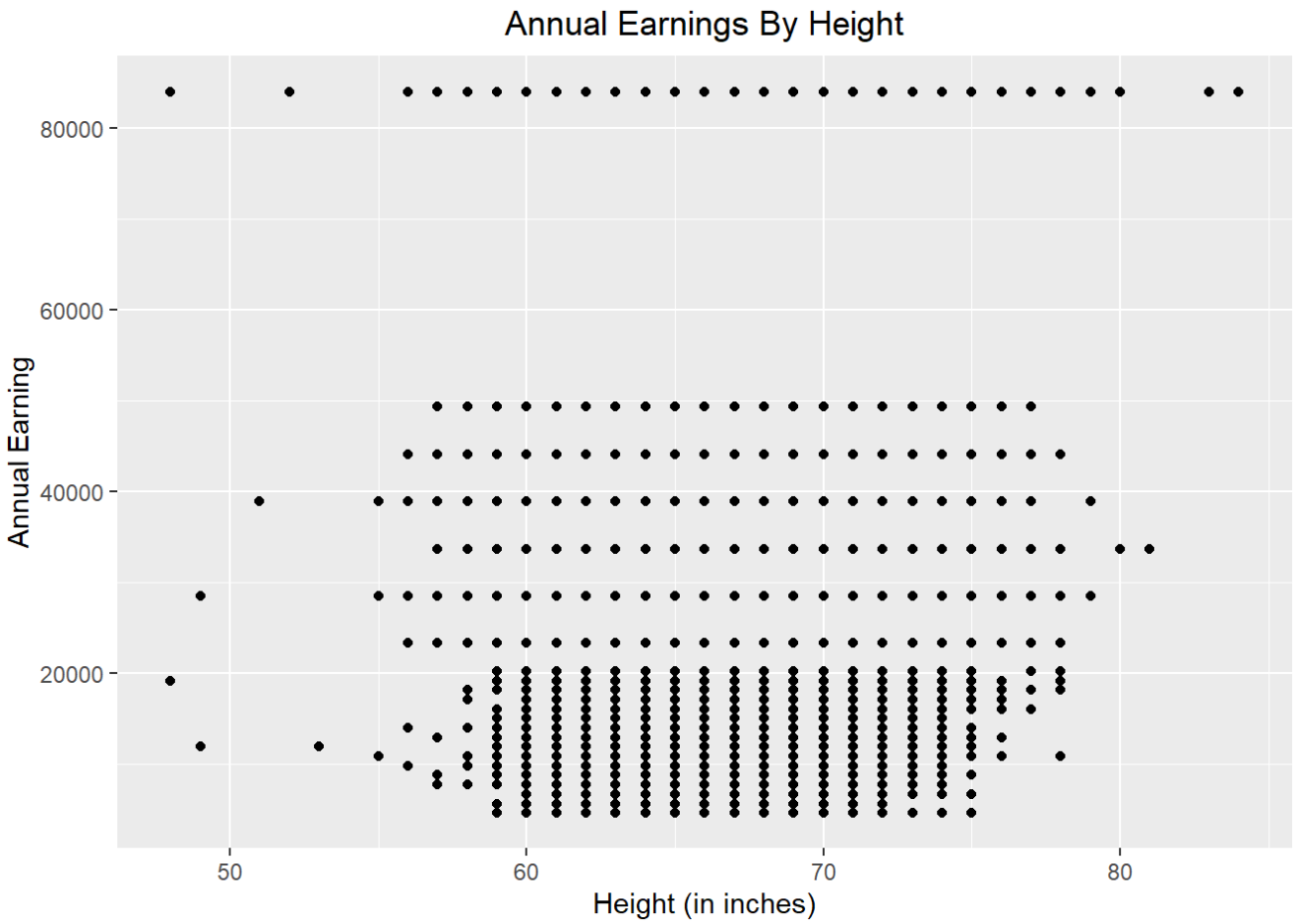
iii)

```
d2 = data.frame(eh[eh$height>67,8:9])
mean2 = mean(d2[,1])
mean2
```

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

iv)

```
ann_plot = ggplot(eh,aes(x=height,y=earnings))+
  geom_point()+
  labs(title="Annual Earnings By Height",
       x="Height (in inches)",
       y="Annual Earning")+
  theme(plot.title = element_text(hjust = 0.5))
ann_plot
```



v)

```
reg = lm(earnings~height,eh)
reg_co = coef(reg)
reg_co
```

```
## (Intercept)      height
##   -512.7336    707.6716
```

SLR: $\text{Earnings} = -512.7 + 707.7 \cdot \text{Height}; \text{Residualstandarderror} : 26780$

a)

Estimated Slope = 707.7. Therefore, as the hight increases by 1 inch, the earnings increase by $707.7.

b)

Predicted earnings for a 67 inch worker.

```
earning = -512.7 + 707.7 * 67
earning
```

```
## [1] 46903.2
```

Predicted earnings for a 70 inch worker.

```
earning1 = -512.7 + 707.7 * 70
earning1
```

```
## [1] 49026.3
```

Predicted earnings for a 65 inch worker.

```
earning2 = -512.7 + 707.7 * 65
earning2
```

```
## [1] 45487.8
```

C)

```
r_squared = summary(reg)$r.squared
r_squared
```

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

vi)

Height can be correlated to other factors like age, which can determine earnings. Similarly, height can also be correlated to gender as the average heights of the genders differ. If gender discrimination exists in this data set, men are likely to earn higher than women. Therefore, it is very unlikely that the error term u_i has a zero conditional mean, given Height (x_i).

Problem 4

i)

Given, Estimated Error Variance = $\hat{\sigma}^2 = 2.04672$

To Find - RSS.

$\hat{\sigma}^2 = RSS/n - 2 = 2.04672$

$2.04672 = RSS/51 - 2$

$2.04672 = RSS/49$

$2.04672 \cdot 49 = RSS$

$RSS = 100.2893$

ii)

Given, estimated variance of β_2 = 0.00098

To Find, standard error of β_2 and $\sum(x_i - \bar{x})^2$.

$\text{Standard. error} = \sqrt{\overline{0.00098}}$

Standard. error = 0.03130495

$\sum(x_i - \bar{x})^2 = (n-1) \cdot \text{Estimated. variance of} \beta_2$

$\sum(x_i - \bar{x})^2 = (51-1) \cdot 0.00098$

$\sum(x_i - \bar{x})^2 = 49 \cdot 0.00098$

$\sum(x_i - \bar{x})^2 = 0.049$

iii)

Given slope value = 0.18.

Interpretation:

When the percentage of males 18 years or older in high school graduates increases by 1%, the state's mean income of males who are 18 years of age or older increases by 0.18 thousand dollars.

iv)

$\hat{\beta}_0 = \bar{y} - \beta_1 \cdot \bar{x}$

slope = 0.18 (from iii)

$\hat{\beta}_0 = 15.187 - 0.18 \cdot 69.139$

$\hat{\beta}_0 = 2.74198$

v)

$\sum(x_i - \bar{x})^2 = 0.049; \bar{x} = 69.139$

$\sum(x_i - 69.139)^2 = 0.049 \quad ; (a-b)^2 = a^2 + b^2 - 2ab$

$\sum(x_i^2 + 69.139^2 - 2(x_i)(69.139)) = 0.049$

$\sum x_i^2 + \sum(69.139)^2 - 2 \cdot 69.139 \cdot \sum x_i = 0.049$

$\sum x_i^2 + \sum(69.139)^2 - 2 \cdot 69.139 \cdot n\bar{x} = 0.049$

$\sum x_i^2 + \sum(69.139)^2 - 2 \cdot 69.139 \cdot 51 \cdot 69.13 = 0.049$

$\sum x_i^2 + \sum(69.139)^2 - 2 \cdot 69.139 \cdot 51 \cdot 69.13 = 0.049$

$\sum x_i^2 + n(69.139)^2 - 487517.1 = 0.049$

$\sum x_i^2 + 51 \cdot (69.139)^2 - 487517.1 = 0.049$

$\sum x_i^2 + 51 \cdot 4780.201 - 487517.1 = 0.049$

$\sum x_i^2 + 243790.3 - 487517.1 = 0.049$

$\sum x_i^2 - 243726.8 = 0.049$

$\sum x_i^2 = 0.049 + 243726.8$

$\sum x_i^2 = 243726.8$

vi)

Known,

β_0 = 2.74198, β_1 = 0.18, y_i = 12.274, x_i = 58.3.

Regression Equation:

$$y_i = 2.74198 + 0.18 \cdot xi \text{ "}$$

Calculating predicted y_i value with the regression equation for x_i = 58.3.

$$y_i = 2.74198 + 0.18 \cdot 58.3$$

$$y_i = 13.23598$$

$$\text{Least. Squares. Residual} = \text{Actual. V alue} - \text{P redicted. V alue}$$

$$\text{Least. Squares. Residual} = 12.274 - 13.23598$$

$$\text{Least. Squares. Residual} = -0.96198$$

Problem 5

i)

We Know that,

$$\hat{\beta_1} = \hat{p}_{xy} \cdot (\hat{\sigma}_y / \hat{\sigma}_x)$$

From this we have,

$$\hat{\beta}_{xy} = \hat{r}_{xy} \cdot (\hat{\sigma}_x / \hat{\sigma}_y)$$

And,

$$\hat{\beta}_{yx} = \hat{r}_{xy} \cdot (\hat{\sigma}_y / \hat{\sigma}_x)$$

Therefore,

$$\hat{\beta}_{yx} \cdot \hat{\beta}_{xy} = \hat{r}_{xy} \cdot (\hat{\sigma}_x / \hat{\sigma}_y) \cdot \hat{r}_{xy} \cdot (\hat{\sigma}_y / \hat{\sigma}_x)$$

$$\hat{\beta}_{yx} \cdot \hat{\beta}_{xy} = \hat{r}_{xy}^{\,2}$$

Or,

$$\hat{\beta}_{yx} \cdot \hat{\beta}_{xy} = r^2$$

ii)

If $\hat{\beta}_{yx} \cdot \hat{\beta}_{xy} = r^2 = 1$, then there is a perfect linear relationship between X and Y. That is, the explained sum of squares = 1 and the residual sum of squares = 0. Therefore, all points lie on the regression line or on the straight line. Hence, it does not matter if we regress Y on X or X on Y because the two regression lines would coincide. It wouldn't matter which variable is dependent and which is independent.