First name: Susheela

Last name: Polepalli

M-number: M10727836

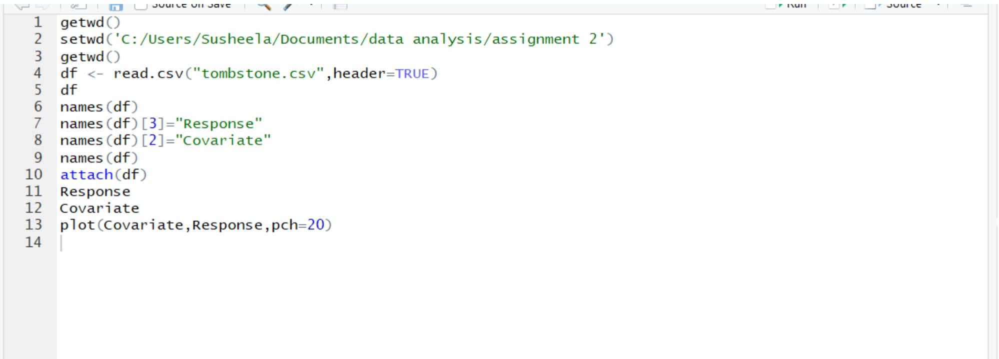
Homework 2

BANA7038

1. Read <tombstone.csv> into R. Use response variable = Marble Tombstone Mean Surface Recession Rate, and covariate = Mean SO2 concentrations over a 100 year period. Description: Marble Tombstone Mean Surface Recession Rates and Mean SO2 concentrations over a 100 year period.

Solution:

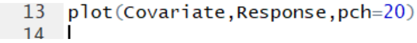
The following code is used to read the data and attach the variables to the workspace.



```
getwd()
setwd('C:/Users/Susheela/Documents/data analysis/assignment 2')
getwd()
df <- read.csv("tombstone.csv",header=TRUE)
df
names(df)
names(df)[3]="Response"
names(df)[2]="Covariate"
names(df)
attach(df)
Response
Covariate
plot(Covariate,Response,pch=20)
```

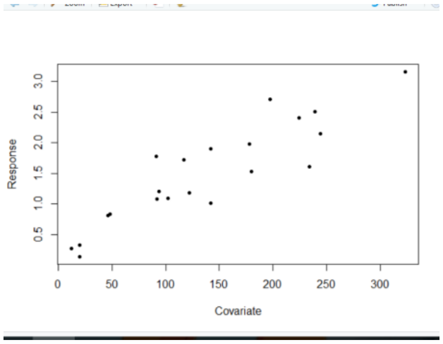
2. Plot data and briefly describe what you observe.

The command “plot” is used to plot the data



```
plot(Covariate,Response,pch=20)
```

The plot is shown as below:



Observations:

The above plot represents the tombstone dataset. The x axis represents the covariate “SO2 concentrations over 100 year period.” And y axis represents the response variable “Marble Tombstone Mean Surface Recession Rates”. The response variable increasing with increasing covariate variable.

3. Perform linear regression using lm() function

3.1. Obtain coefficient estimates $\widehat{\beta_0}, \widehat{\beta_1}$.

3.2. Obtain fitted values and the sum of fitted values.

3.3. Obtain the sum of all values of response variable.

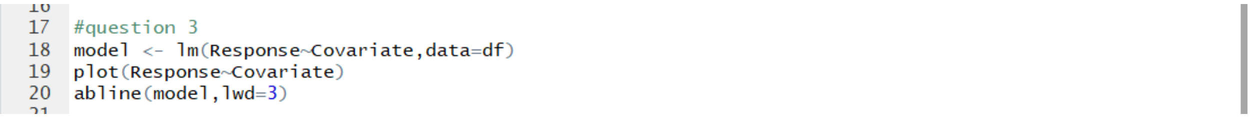
3.4. Obtain residuals and the sum of residuals.

3.5. Obtain the standard errors of $\widehat{\beta_0}, \widehat{\beta_1}$.

Solution:

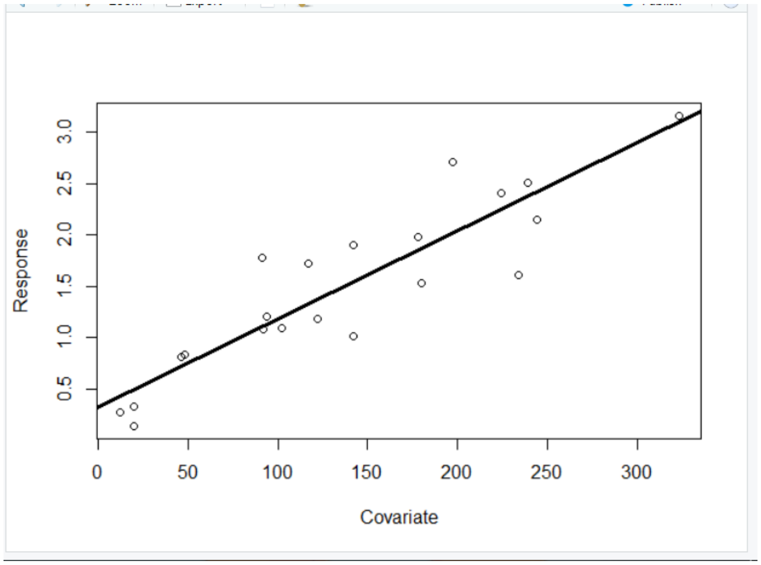
Linear regression:

The following code is used to perform linear regression



```
#question 3
model <- lm(Response~Covariate,data=df)
plot(Response~Covariate)
abline(model,lwd=3)
```

The output is shown as below:

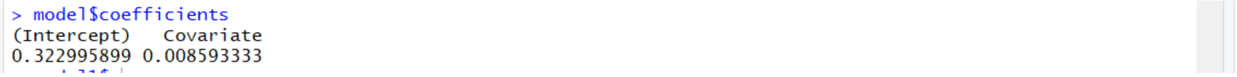


3.1: Coefficient estimates $\widehat{\beta_0}$, $\widehat{\beta_1}$.

Code:

```
model$coefficients
```

Output:

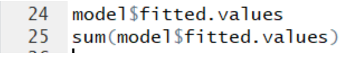


```
> model$coefficients
(Intercept)   Covariate 
0.322995899 0.008593333 
```

The coefficient estimate $\widehat{\beta_0}$=0.322995899, $\widehat{\beta_1}$= 0.008593333.

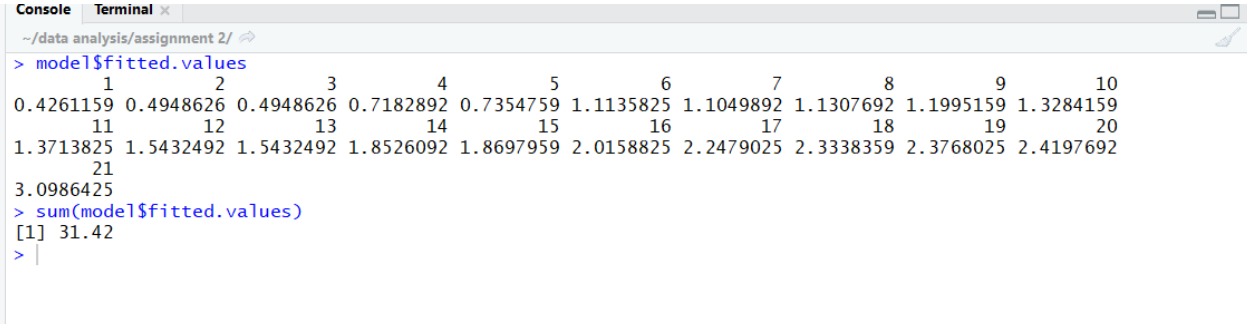
3.2. Fitted values and the sum of fitted values.

Code:



```
model$fitted.values
sum(model$fitted.values)
```

Output:

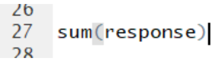


```
> model$fitted.values
        1         2         3         4         5         6         7         8         9        10 
0.4261159 0.4948626 0.4948626 0.7182892 0.7354759 1.1135825 1.1049892 1.1307692 1.1995159 1.3284159 
       11        12        13        14        15        16        17        18        19        20 
1.3713825 1.5432492 1.5432492 1.8526092 1.8697959 2.0158825 2.2479025 2.3338359 2.3768025 2.4197692 
       21 
3.0986425 
> sum(model$fitted.values)
[1] 31.42
> 
```

Sum of fitted values = 31.42

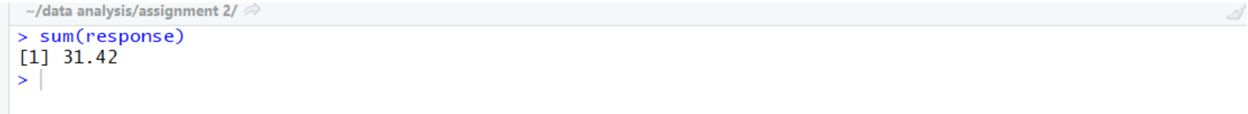
3.3. The sum of all values of response variable.

Sum:



```
sum(response)
```

Output:



```
> sum(response)
[1] 31.42
> 
```

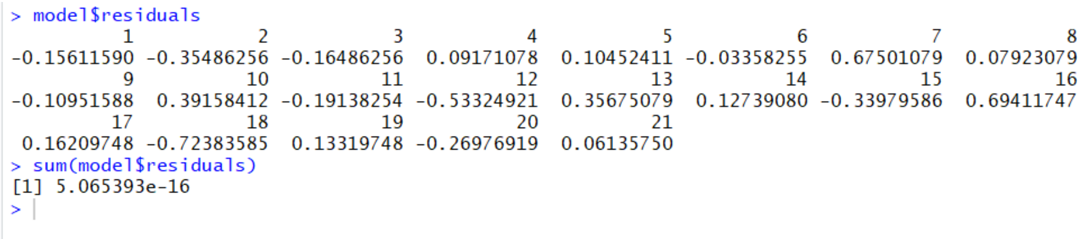
Sum of all values of response variable= 31.42

3.4. Obtain residuals and the sum of residuals.

Code:

```
model$residuals
sum(model$residuals)
```

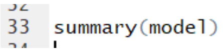
Output:



```
> model$residuals
          1           2           3           4           5           6           7           8 
-0.15611590 -0.35486256 -0.16486256  0.09171078  0.10452411 -0.03358255  0.67501079  0.07923079 
          9          10          11          12          13          14          15          16 
-0.10951588  0.39158412 -0.19138254 -0.53324921  0.35675079  0.12739080 -0.33979586  0.69411747 
         17          18          19          20          21 
 0.16209748 -0.72383585  0.13319748 -0.26976919  0.06135750 
> sum(model$residuals)
[1] 5.065393e-16
> 
```

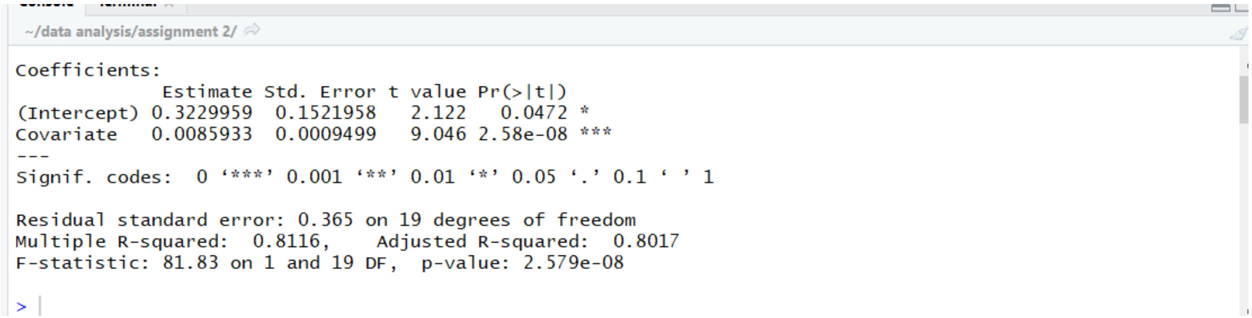
3.5. Obtain the standard errors of $\widehat{\boldsymbol{\beta}_0}$, $\widehat{\boldsymbol{\beta}_1}$.

Code:



```
summary(model)
```

Output:

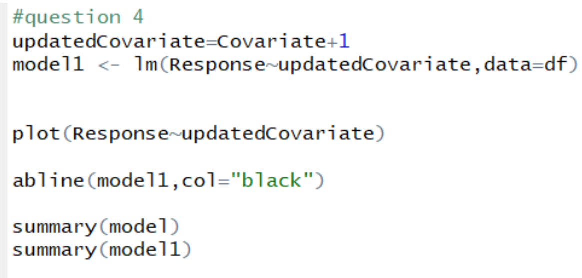


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.3229959  0.1521958   2.122   0.0472 *  
Covariate   0.0085933  0.0009499   9.046 2.58e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.365 on 19 degrees of freedom
Multiple R-squared:  0.8116,	Adjusted R-squared:  0.8017 
F-statistic: 81.83 on 1 and 19 DF,  p-value: 2.579e-08

> 
```

The standard errors of $\widehat{\beta_0}$=0.1521958, $\widehat{\beta_1}$=0.0009499.

4. Suppose we increase SO2 Concentration by one unit, how does such a change influence the Surface Recession Rate?

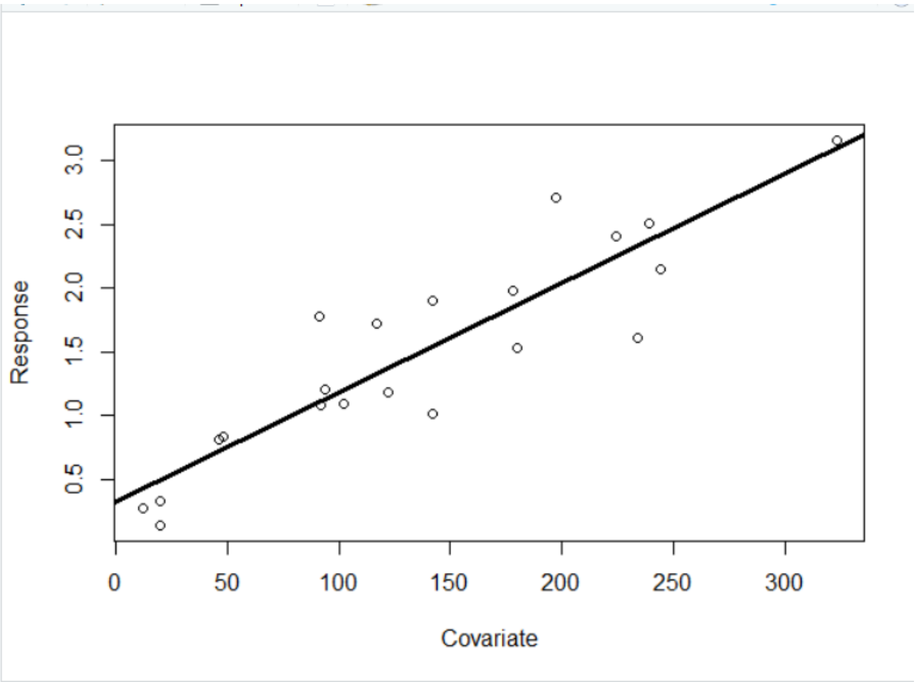
Solution: I have constructed a list “updatedcovariate”with s02 concentration increased by 1 unit , performed linear regression and named it model1.The code for the following is as follows:



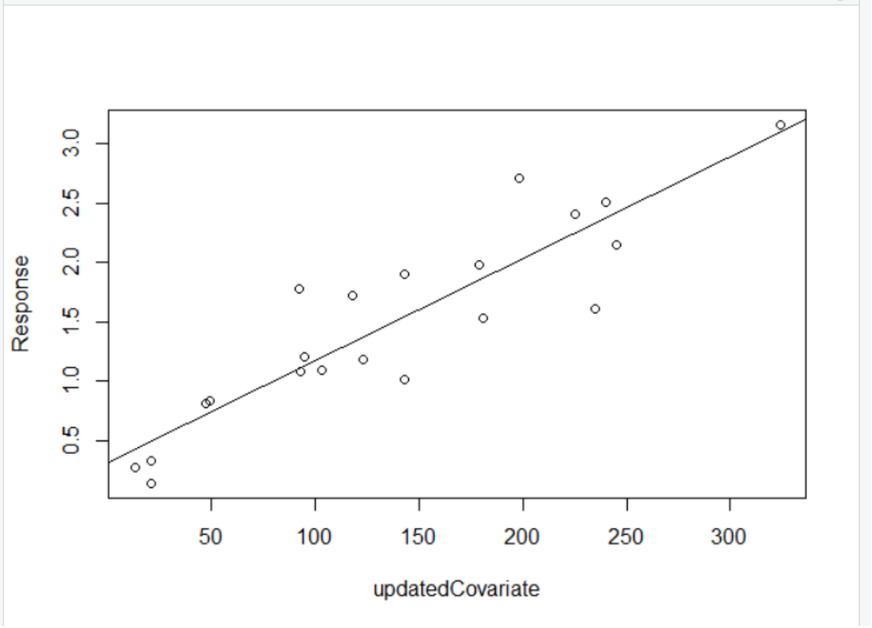
```
#question 4
updatedCovariate=Covariate+1
model1 <- lm(Response~updatedCovariate,data=df)


plot(Response~updatedCovariate)

abline(model1,col="black")

summary(model)
summary(model1)
```

The plots are as follows:

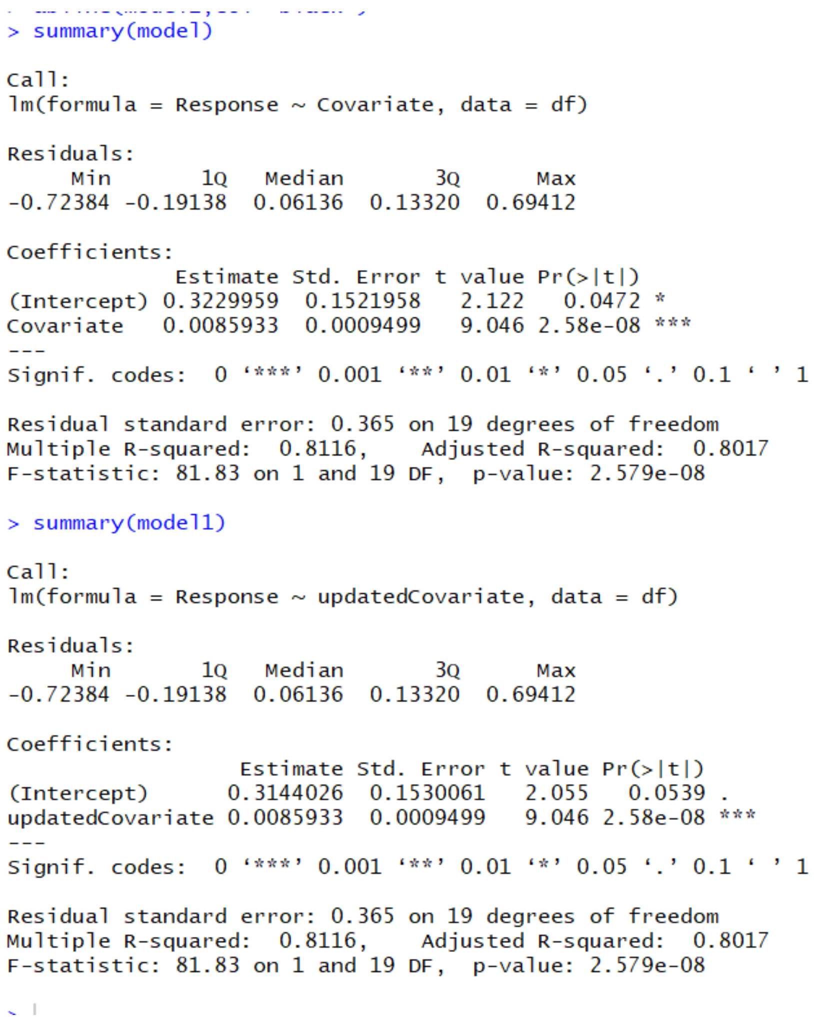
Plot for the original model with original covariate and response values.



Plot for the new model with updated covariate and response values.



The summary of the respective models is used to compare to models.



```
> summary(model)

Call:
lm(formula = Response ~ Covariate, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.72384 -0.19138  0.06136  0.13320  0.69412 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.3229959  0.1521958   2.122   0.0472 *  
Covariate   0.0085933  0.0009499   9.046 2.58e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.365 on 19 degrees of freedom
Multiple R-squared:  0.8116,	Adjusted R-squared:  0.8017 
F-statistic: 81.83 on 1 and 19 DF,  p-value: 2.579e-08

> summary(model1)

Call:
lm(formula = Response ~ updatedCovariate, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.72384 -0.19138  0.06136  0.13320  0.69412 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      0.3144026  0.1530061   2.055   0.0539 .  
updatedCovariate 0.0085933  0.0009499   9.046 2.58e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.365 on 19 degrees of freedom
Multiple R-squared:  0.8116,	Adjusted R-squared:  0.8017 
F-statistic: 81.83 on 1 and 19 DF,  p-value: 2.579e-08
```

There is a change is y intercept estimate whereas the residual values remain the same.

Therefore, there is no much effect on the response variable when the covariate is increased by 1 unit.

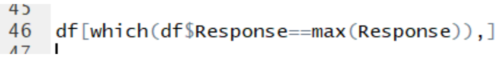
5. Does the intercept of the linear regression have natural interpretation? If so, what does it mean?

Solution:

Yes, the intercept has natural interpretation. The line Y=0.0085933X + 0.3144026 has Y intercept of 0.3144026. This means that when mean SO2 concentrations= 0 ug /m^3 and doesn't affect the tombstone, the surface recession rate of the tombstone is 0.3144026 mm/100 years. This means there are other reasons for surface recession of tombstone apart from SO2 concentration which makes sense.

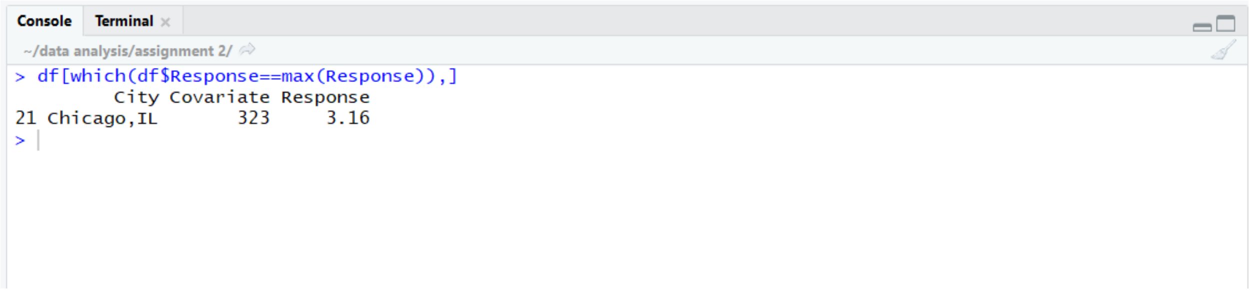
6. Which area (i.e., observation) has the highest Surface Recession Rate?

Code:



```
df[which(df$Response==max(Response)),]
```

Output:



```
> df[which(df$Response==max(Response)),]
         City Covariate Response
21 Chicago,IL       323     3.16
> 
```

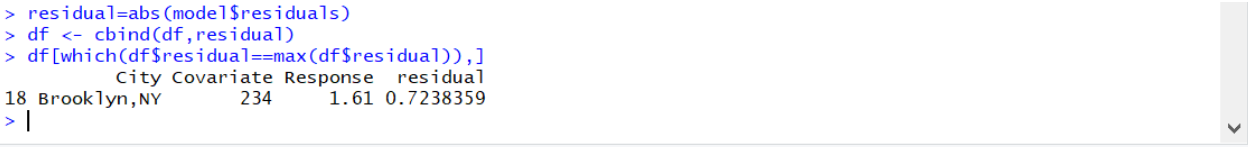
Chicago is having highest surface recession rate.

7. Which area (i.e., observation) has the largest residual (i.e., the largest absolute value) according to the linear regression you just fitted?

Code:

```
residual=abs(model$residuals)
df <- cbind(df,residual)
df[which(df$residual==max(df$residual)),]
```

Output:



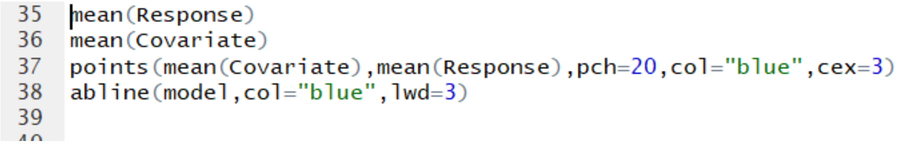
```
> residual=abs(model$residuals)
> df <- cbind(df,residual)
> df[which(df$residual==max(df$residual)),]
          City Covariate Response  residual
18 Brooklyn,NY       234     1.61 0.7238359
> |
```

Brooklyn is having largest residual value.

8. Calculate the mean of covariate and mean of response. Verify the fact that the fitted regression line go through the point $(\bar{x}, \bar{y})$.

Solution:

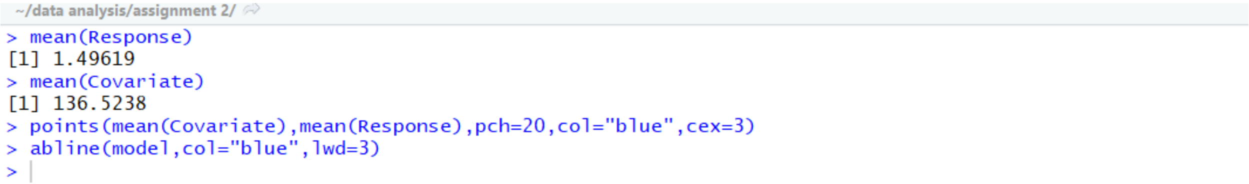
Code:



```
mean(Response)
mean(Covariate)
points(mean(Covariate),mean(Response),pch=20,col="blue",cex=3)
abline(model,col="blue",lwd=3)
```

Output:

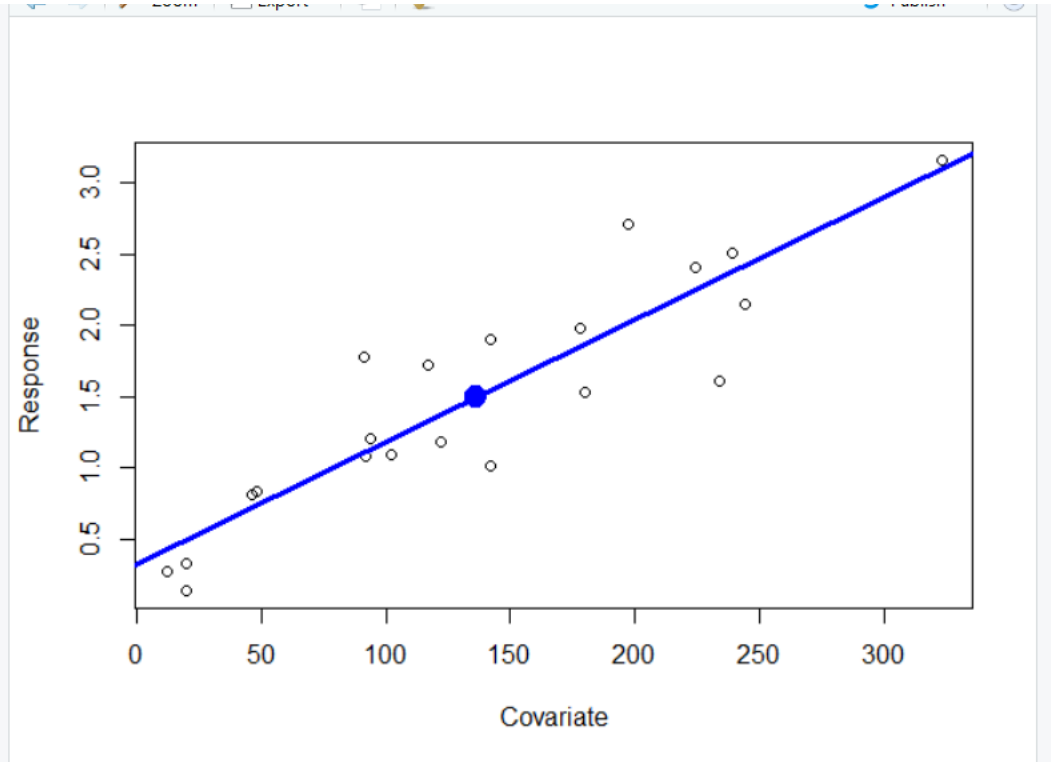
The mean of response and covariate variables are 1.49619 and 136.5238 respectively.



~/data analysis/assignment 2/

```
> mean(Response)
[1] 1.49619
> mean(Covariate)
[1] 136.5238
> points(mean(Covariate),mean(Response),pch=20,col="blue",cex=3)
> abline(model,col="blue",lwd=3)
> 
```

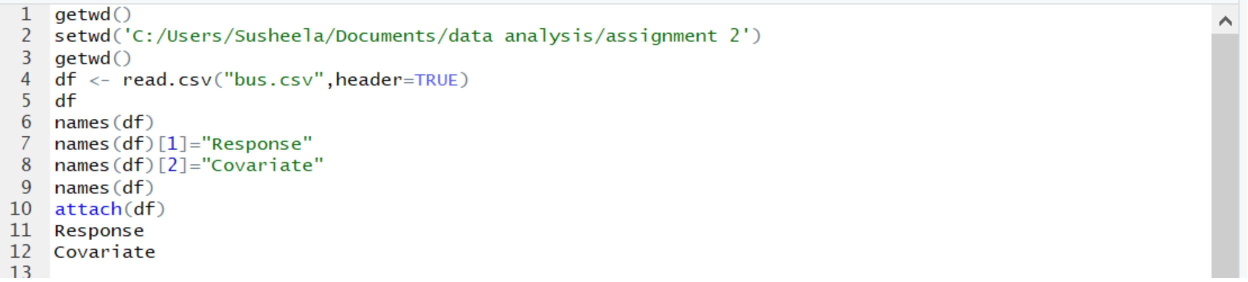
The plot shows that the linear regression model passes through the mean values of covariate and response values respectively.



9. Repeat the same questions (1-8) for the date set <bus.csv>. Description: Cross-sectional analysis of 24 British bus companies (1951). Use response variable = Expenses per car mile (pence), covariate = Car miles per year (1000s).

9.1 Read <bus.csv> into R.

Solution: The following code is used to read the data and attach the variables to the workspace.



```
getwd()
setwd('C:/Users/Susheela/Documents/data analysis/assignment 2')
getwd()
df <- read.csv("bus.csv",header=TRUE)
df
names(df)
names(df)[1]="Response"
names(df)[2]="Covariate"
names(df)
attach(df)
Response
Covariate

```

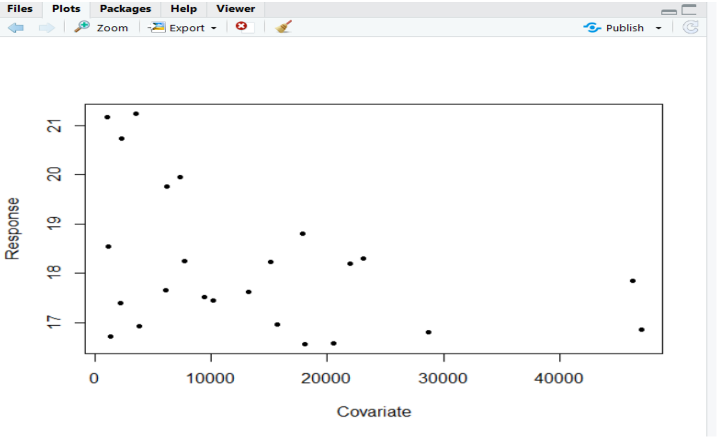
9.2. Plot data and briefly describe what you observe.

Solution:

The command "plot" is used to plot the data:

```
#question2
plot(Covariate,Response,pch=20)
```

The output is as shown below:



Observations:

The above plot represents the bus dataset. The x axis represents the covariate “car miles per year (1000s).” and y axis represents the response variable “Expenses per car mile (pence)”. The response variable is decreasing with increasing covariate variable.

9.3. Perform linear regression using lm() function

9.3.1. Obtain coefficient estimates $\widehat{\beta_0}, \widehat{\beta_1}$.

9.3.2. Obtain fitted values and the sum of fitted values.

9.3.3. Obtain the sum of all values of response variable.

9.3.4. Obtain residuals and the sum of residuals.

9.3.5. Obtain the standard errors of $\widehat{\beta_0}, \widehat{\beta_1}$.

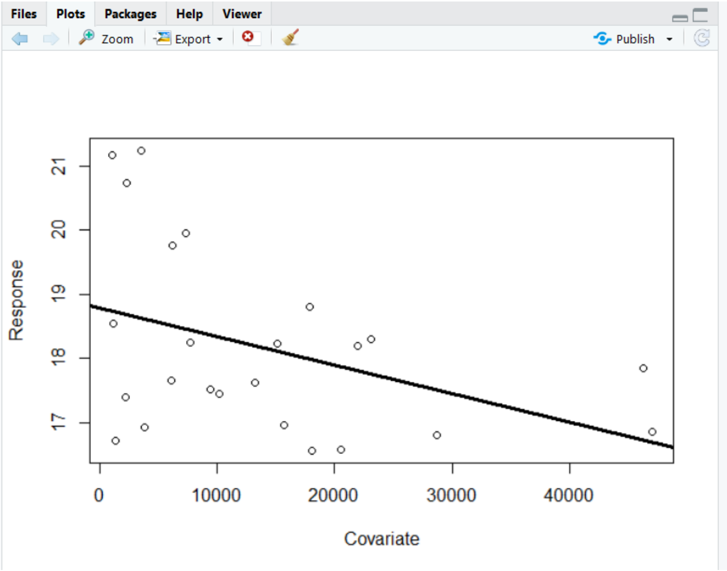
Solution:

Linear regression:

The following code is used to perform linear regression

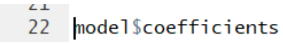
```
#question 3
model <- lm(Response~Covariate,data=df)
plot(Response~Covariate)
abline(model,lwd=3)
```

The output plot is as follows:



9.3.1. Obtain coefficient estimates $\widehat{\beta_0}$, $\widehat{\beta_1}$.

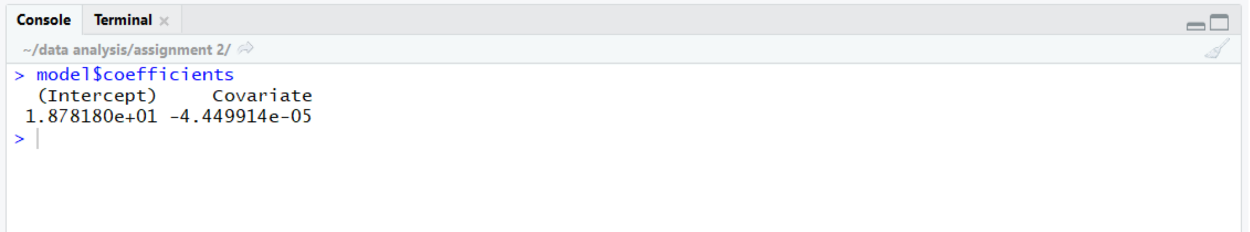
Code:



```
model$coefficients
```

Output:

The coefficients of $\widehat{\beta_0}$ (intercept), $\widehat{\beta_1}$ is as follows:



```
> model$coefficients
  (Intercept)      Covariate 
 1.878180e+01 -4.449914e-05 
> 
```

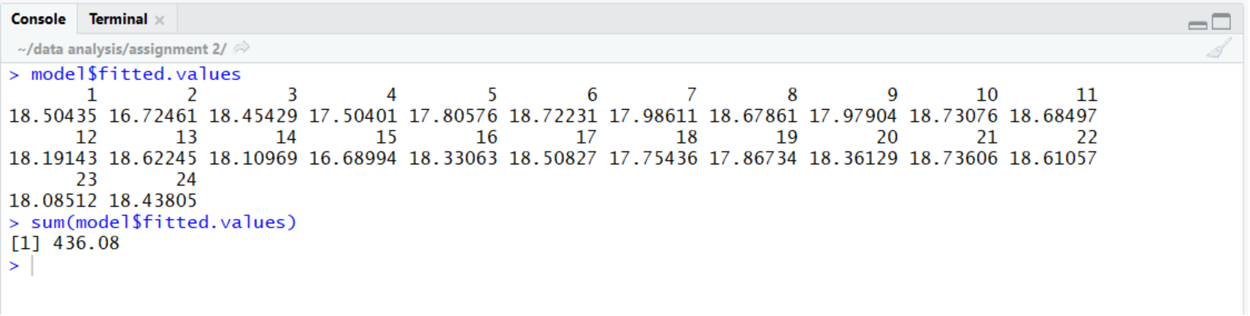
9.3.2. Obtain fitted values and the sum of fitted values.

Code:

```
model$fitted.values
sum(model$fitted.values)
```

Output:

The sum of fitted values =436.08



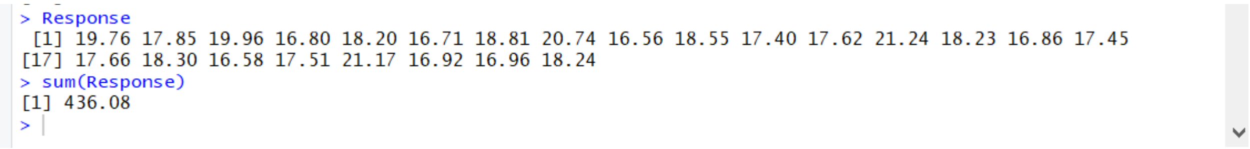
```
> model$fitted.values
       1        2        3        4        5        6        7        8        9       10       11 
18.50435 16.72461 18.45429 17.50401 17.80576 18.72231 17.98611 18.67861 17.97904 18.73076 18.68497 
      12       13       14       15       16       17       18       19       20       21       22 
18.19143 18.62245 18.10969 16.68994 18.33063 18.50827 17.75436 17.86734 18.36129 18.73606 18.61057 
      23       24 
18.08512 18.43805 
> sum(model$fitted.values)
[1] 436.08
> 
```

9.3.3. Obtain the sum of all values of response variable.

Code:

```
Response
sum(Response)
```

Output: Sum of response variable = 436.08



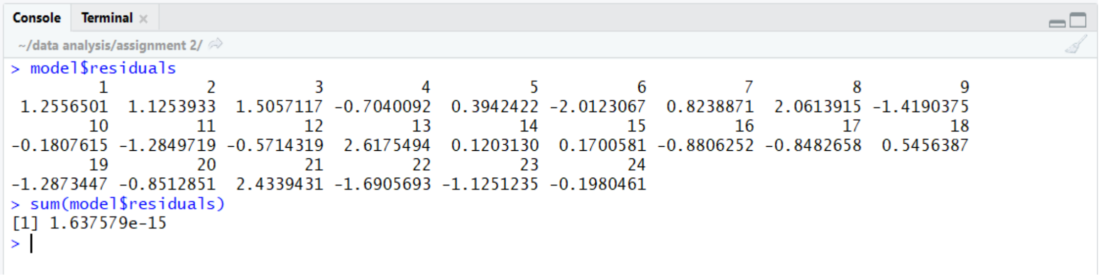
```
> Response
 [1] 19.76 17.85 19.96 16.80 18.20 16.71 18.81 20.74 16.56 18.55 17.40 17.62 21.24 18.23 16.86 17.45
[17] 17.66 18.30 16.58 17.51 21.17 16.92 16.96 18.24
> sum(Response)
[1] 436.08
> 
```

9.3.4. Obtain residuals and the sum of residuals.

Code:

```
model$residuals
sum(model$residuals)
```

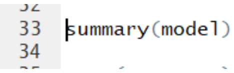
Output:



```
> model$residuals
         1          2          3          4          5          6          7          8          9 
 1.2556501  1.1253933  1.5057117 -0.7040092  0.3942422 -2.0123067  0.8238871  2.0613915 -1.4190375 
        10         11         12         13         14         15         16         17         18 
-0.1807615 -1.2849719 -0.5714319  2.6175494  0.1203130  0.1700581 -0.8806252 -0.8482658  0.5456387 
        19         20         21         22         23         24 
-1.2873447 -0.8512851  2.4339431 -1.6905693 -1.1251235 -0.1980461 
> sum(model$residuals)
[1] 1.637579e-15
> 
```

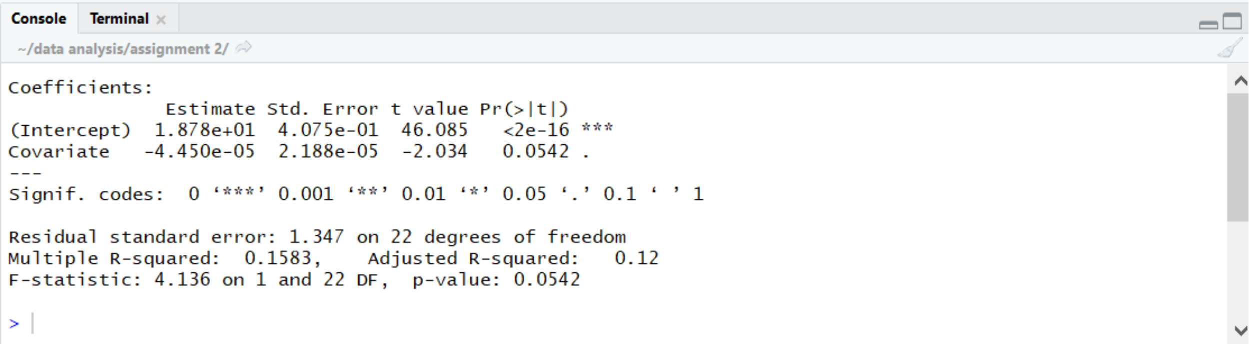
9.3.5. Obtain the standard errors of $\widehat{\beta_0}$, $\widehat{\beta_1}$.

Summary command is used to retrieve the standard errors of $\widehat{\beta_0}$, $\widehat{\beta_1}$ respectively.



```
summary(model)
```

Output:



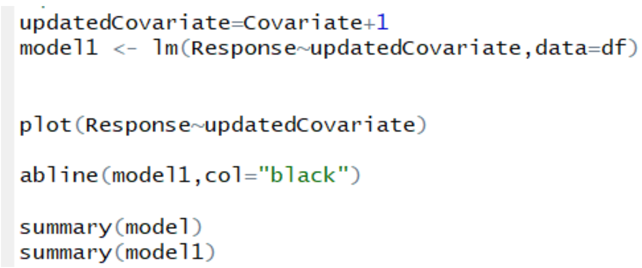
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.878e+01  4.075e-01  46.085   <2e-16 ***
Covariate   -4.450e-05  2.188e-05  -2.034   0.0542 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.347 on 22 degrees of freedom
Multiple R-squared:  0.1583,	Adjusted R-squared:   0.12 
F-statistic: 4.136 on 1 and 22 DF,  p-value: 0.0542

> 
```

Standard errors of $\widehat{\beta_0}$= 4.075e-01 , $\widehat{\beta_1}$= 2.188e-05 .

9.4. Suppose we increase car miles per year by one unit, how does such a change influence the Expenses per car mile?

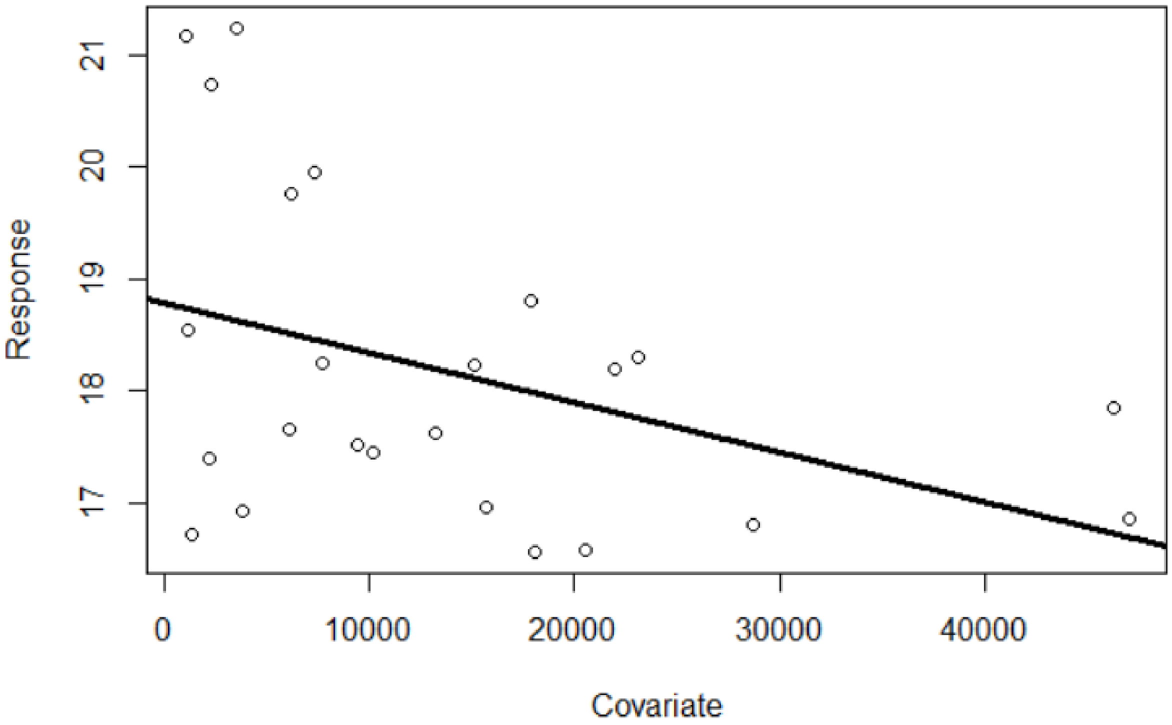
Solution:

I have constructed a list “updatedcovariate” with car miles per year increased by 1 unit (updated covariate), performed linear regression and named it model1.The code for the following is as follows:

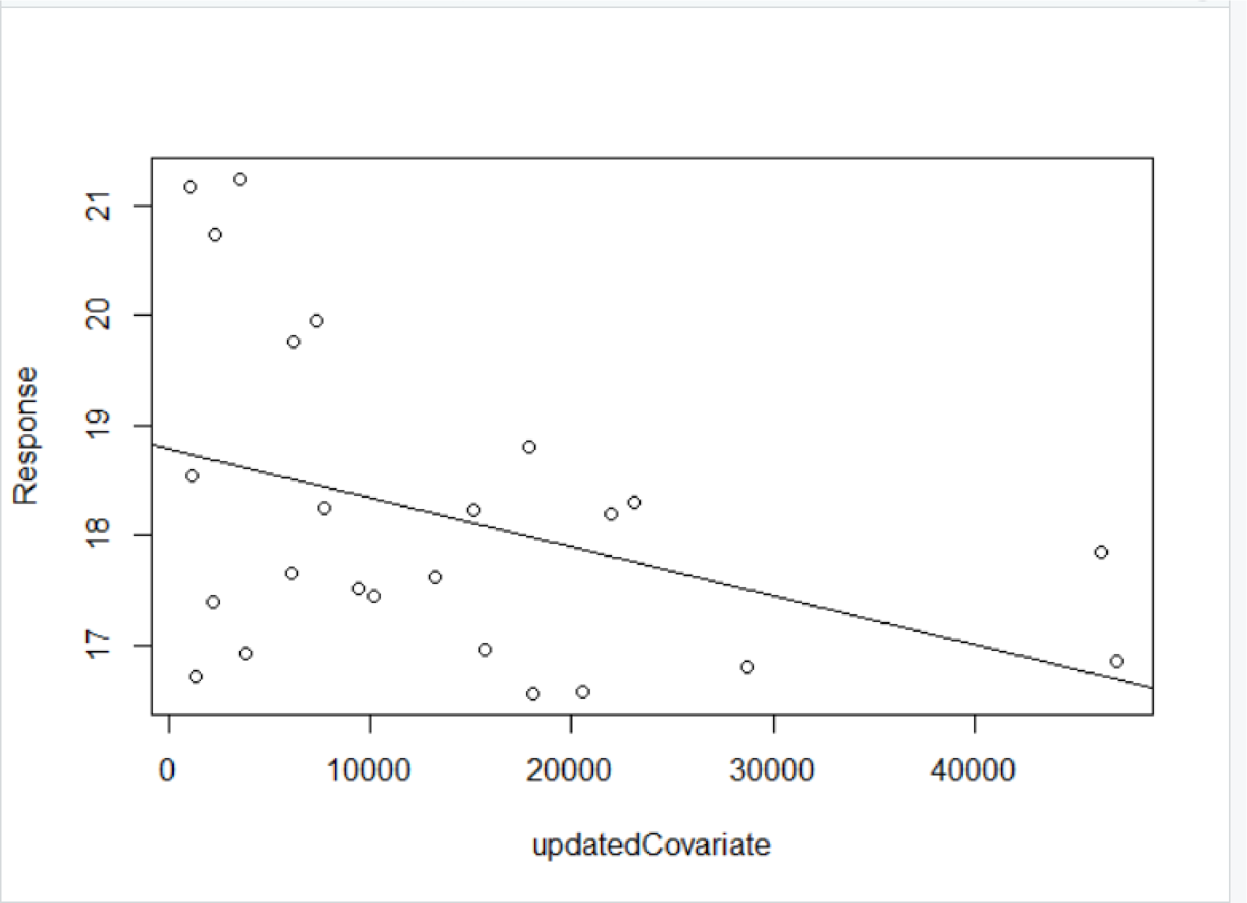


```
updatedCovariate=Covariate+1
model1 <- lm(Response~updatedCovariate,data=df)


plot(Response~updatedCovariate)

abline(model1,col="black")

summary(model)
summary(model1)
```

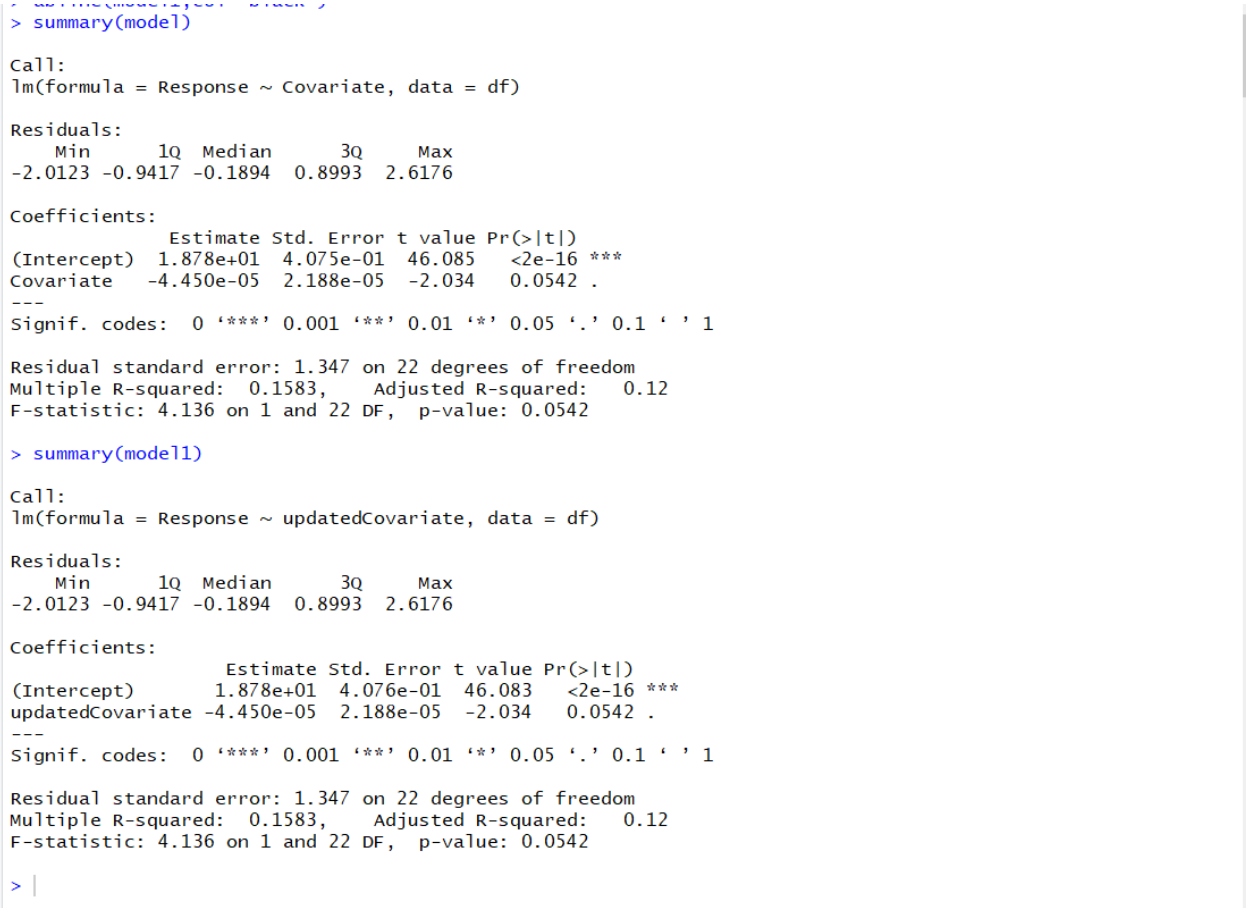
This is the plot which original covariate and response values:



This is the plot and linear regression with response and updated covariate values. :



The summary of respective models for comparison is shown as follows:



```
> abline(model1,col= "black")
> summary(model)

Call:
lm(formula = Response ~ Covariate, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.0123 -0.9417 -0.1894  0.8993  2.6176 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.878e+01  4.075e-01  46.085   <2e-16 ***
Covariate   -4.450e-05  2.188e-05  -2.034   0.0542 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.347 on 22 degrees of freedom
Multiple R-squared:  0.1583,	Adjusted R-squared:   0.12 
F-statistic: 4.136 on 1 and 22 DF,  p-value: 0.0542

> summary(model1)

Call:
lm(formula = Response ~ updatedCovariate, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.0123 -0.9417 -0.1894  0.8993  2.6176 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       1.878e+01  4.076e-01  46.083   <2e-16 ***
updatedCovariate -4.450e-05  2.188e-05  -2.034   0.0542 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.347 on 22 degrees of freedom
Multiple R-squared:  0.1583,	Adjusted R-squared:   0.12 
F-statistic: 4.136 on 1 and 22 DF,  p-value: 0.0542

> 
```

There is no change in y intercept estimate, residual and other corresponding values.

Therefore, there is no effect on the response variable when the covariate is increased by 1 unit.

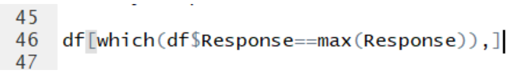
9.5. Does the intercept of the linear regression have natural interpretation? If so, what does it mean?

Solution: No, the intercept of linear regression does not have natural interpretation. This is because in the linear regression, Y intercept is positive i.e. when the car miles per year(1000s) =0, the expenses per car mile is a positive value which is not relevant as the car hasn’t travelled any mile.

9.6. Which area (i.e., observation) has the highest Expenses per car mile (pence)?

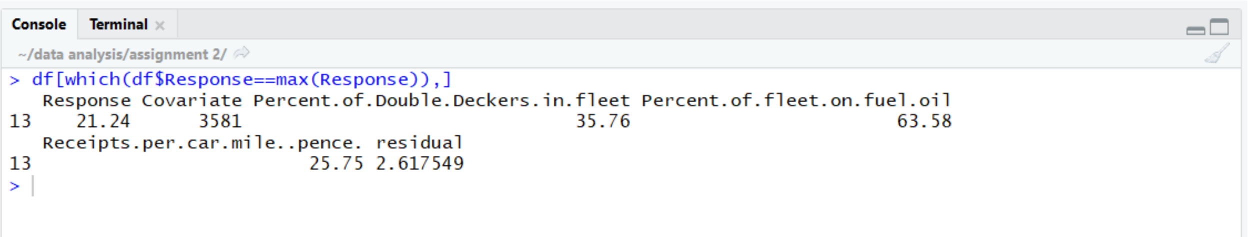
Solution: the 13th observation has the highest expenses per car mile (pence).

Code:



```
df[which(df$Response==max(Response)),]
```

Output:

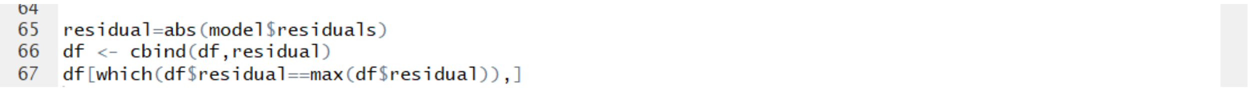


```
> df[which(df$Response==max(Response)),]
   Response Covariate Percent.of.Double.Deckers.in.fleet Percent.of.fleet.on.fuel.oil
13    21.24      3581                              35.76                        63.58
   Receipts.per.car.mile..pence. residual
13                         25.75 2.617549
> 
```

9.7. Which area (i.e., observation) has the largest residual (i.e., the largest absolute value) according to the linear regression you just fitted?

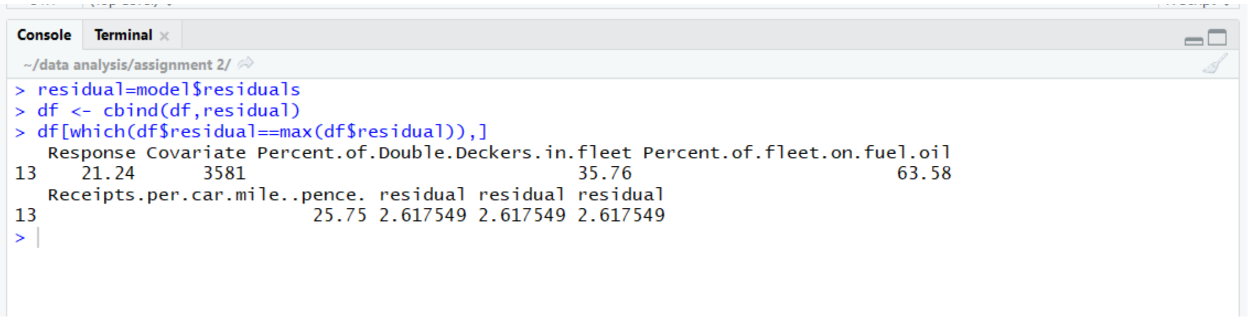
Solution: The 13th observation is having the largest residual value.

Code:



```
residual=abs(model$residuals)
df <- cbind(df,residual)
df[which(df$residual==max(df$residual)),]
```

Output:

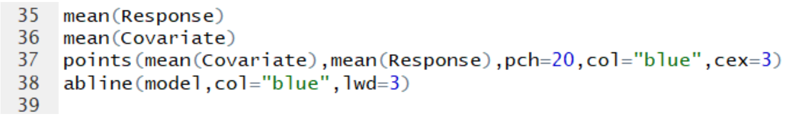


```
> residual=model$residuals
> df <- cbind(df,residual)
> df[which(df$residual==max(df$residual)),]
   Response Covariate Percent.of.Double.Deckers.in.fleet Percent.of.fleet.on.fuel.oil
13    21.24      3581                              35.76                        63.58
   Receipts.per.car.mile..pence. residual residual residual
13                         25.75 2.617549 2.617549 2.617549
> 
```

9.8. Calculate the mean of covariate and mean of response. Verify the fact that the fitted regression line go through the point $(\bar{x}, \bar{y})$.

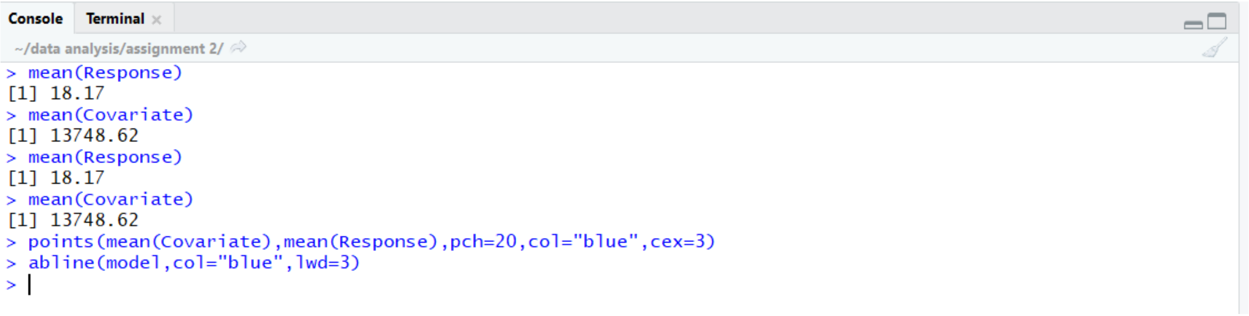
Solution:

Code:



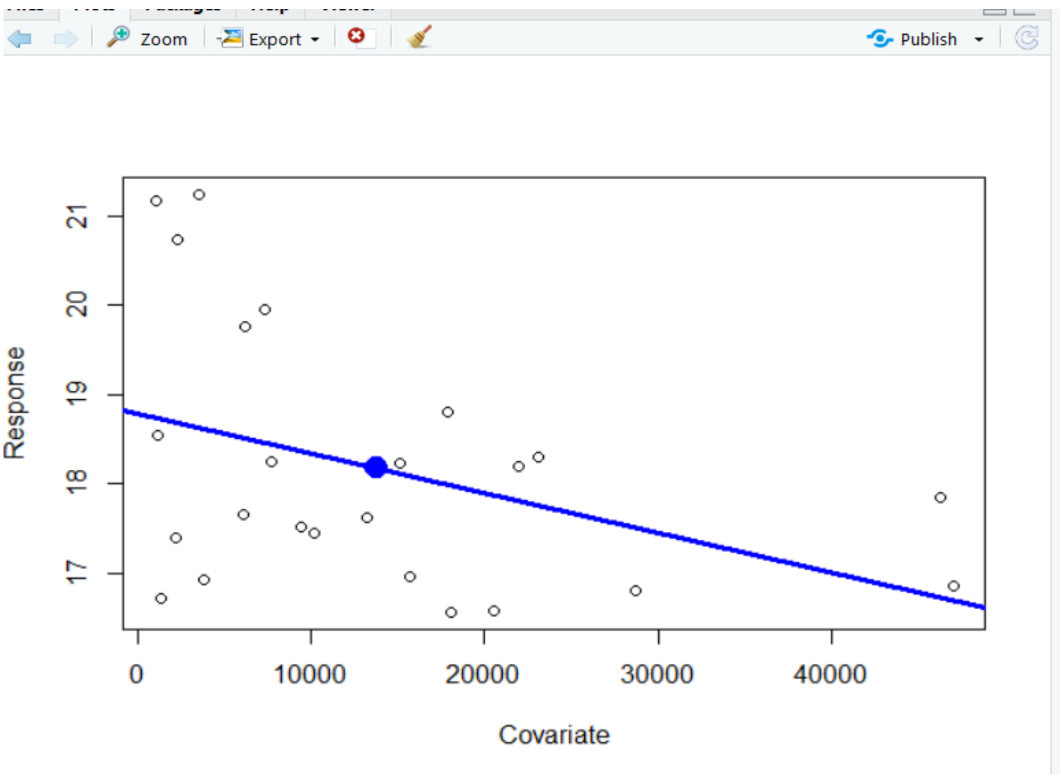
```
mean(Response)
mean(Covariate)
points(mean(Covariate),mean(Response),pch=20,col="blue",cex=3)
abline(model,col="blue",lwd=3)

```

Output: the mean values of response and covariate variables are 18.17 and 13748.62 respectively.



```
> mean(Response)
[1] 18.17
> mean(Covariate)
[1] 13748.62
> mean(Response)
[1] 18.17
> mean(Covariate)
[1] 13748.62
> points(mean(Covariate),mean(Response),pch=20,col="blue",cex=3)
> abline(model,col="blue",lwd=3)
> |
```

Plot:



As you can see, the linear regression line passes through the means of response and covariate variables.