First Name: Susheela

Last Name: Polepalli

M-Number: M10727836

Homework 3

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:

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. **2. Obtain R^2, explain what it means.**

Solution:

Code:

```
model2 <- lm(Response~Covariate,data=df)
plot(Response~Covariate)
abline(model2,lwd=3)
summary(model2)
summary(model2)$r.square
```

Output:

```
> summary(model2)$r.square
[1] 0.8115724
>
```

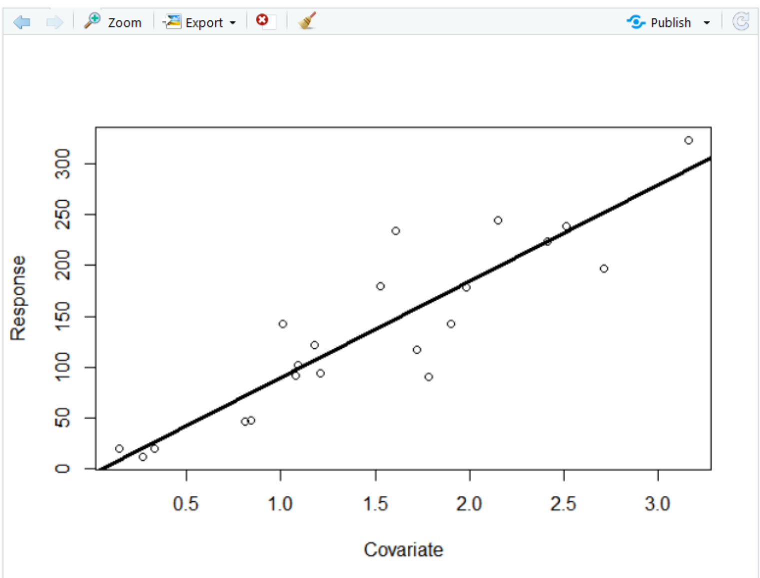
Observations of R^2:

Since R^2 is a proportion, it is always a number between 0 and 1.

If R^2 = 1, all of the data points fall perfectly on the regression line. The predictor x accounts for all of the variation in y.

If R^2 = 0, the estimated regression line is perfectly horizontal. The predictor x accounts for none of the variation in y.

The R^2=0.8115724 means that the data points are close to the fitted regression line.



3. Perform the following hypothesis testing and interval estimation using lm() and other related R functions.

3.1. Perform t tests, obtain t statistics and p values, interpret the results, make a conclusion (i.e. reject or not reject) and explain why. Note: please explain what the null hypothesis is.

Solution:

Code:

```
summary(model2)$coef[,3] # t value
summary(model2)$coef[,4] # p value
```

Output:

```
> summary(model2)$coef[,3] # t value
(Intercept)   Covariate
   2.122239    9.046242
> summary(model2)$coef[,4] # p value
 (Intercept)     Covariate
4.718525e-02 2.578534e-08
```

The coefficient t-value is a measure of how many standard deviations our coefficient estimate is far away from 0. We want it to be far away from zero as this would indicate we could reject the null

We reject the null hypothesis and conclude that there is a relationship between response and covariate as the t value is greater than 0 and corresponding p values are less which leads to rejection of null hypothesis.

3.2. Perform ANOVA test (F test), obtain F statistic and p value, interpret the results, make conclusion (i.e. reject or not reject) and explain why. Note: please explain what the null hypothesis is.

Solution:

Code:

```
summary(model2) # f statistic
```

Output:

```
> summary(model2) #summary

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

>
```

F statistic=81.83 and p value=2.579e-08

Null Hypothesis H0:There is no relationship between response and covariate is rejected as the F statistic is > 1.

3.3. Compute confidence interval for coefficients, fitted values (mean response), interpret the meanings of these quantities.

Solution:

Code:

```
confint(model2,level=0.95)
confint(model2,level=0.90)
confint(model2,level=0.99)

predict.lm(model2, interval="confidence")
```

Output:

```
~/data analysis/assignment 3/
> confint(model2,level=0.95)
                  2.5 %     97.5 %
(Intercept) 0.004446349 0.64154545
Covariate   0.006605098 0.01058157
> confint(model2,level=0.90)
                    5 %       95 %
(Intercept) 0.059829082 0.5861627
Covariate   0.006950771 0.0102359
> confint(model2,level=0.99)
                   0.5 %     99.5 %
(Intercept) -0.112426440 0.75841824
Covariate    0.005875634 0.01131103
> predict.lm(model2, interval="confidence")
         fit       lwr       upr
1  0.4261159 0.1276356 0.7245962
2  0.4948626 0.2094375 0.7802876
3  0.4948626 0.2094375 0.7802876
4  0.7182892 0.4729586 0.9636199
5  0.7354759 0.4930475 0.9779043
6  1.1135825 0.9248239 1.3023412
7  1.1049892 0.9152900 1.2946885
8  1.1307692 0.9438424 1.3176960
9  1.1995159 1.0192244 1.3798074
10 1.3284159 1.1572428 1.4995889
11 1.3713825 1.2021867 1.5405784
12 1.5432492 1.3761806 1.7103178
13 1.5432492 1.3761806 1.7103178
14 1.8526092 1.6666152 2.0386032
15 1.8697959 1.6820050 2.0575867
16 2.0158825 1.8103314 2.2214336
17 2.2479025 2.0069821 2.4888230
18 2.3338359 2.0781916 2.5894801
19 2.3768025 2.1135420 2.6400631
20 2.4197692 2.1487417 2.6907967
21 3.0986425 2.6921265 3.5051585
> 
```

3.4. Plot data points, the regression line, the confidence interval for fitted values (to show that the interval is wider on both sides and narrow in the center).

Solution:

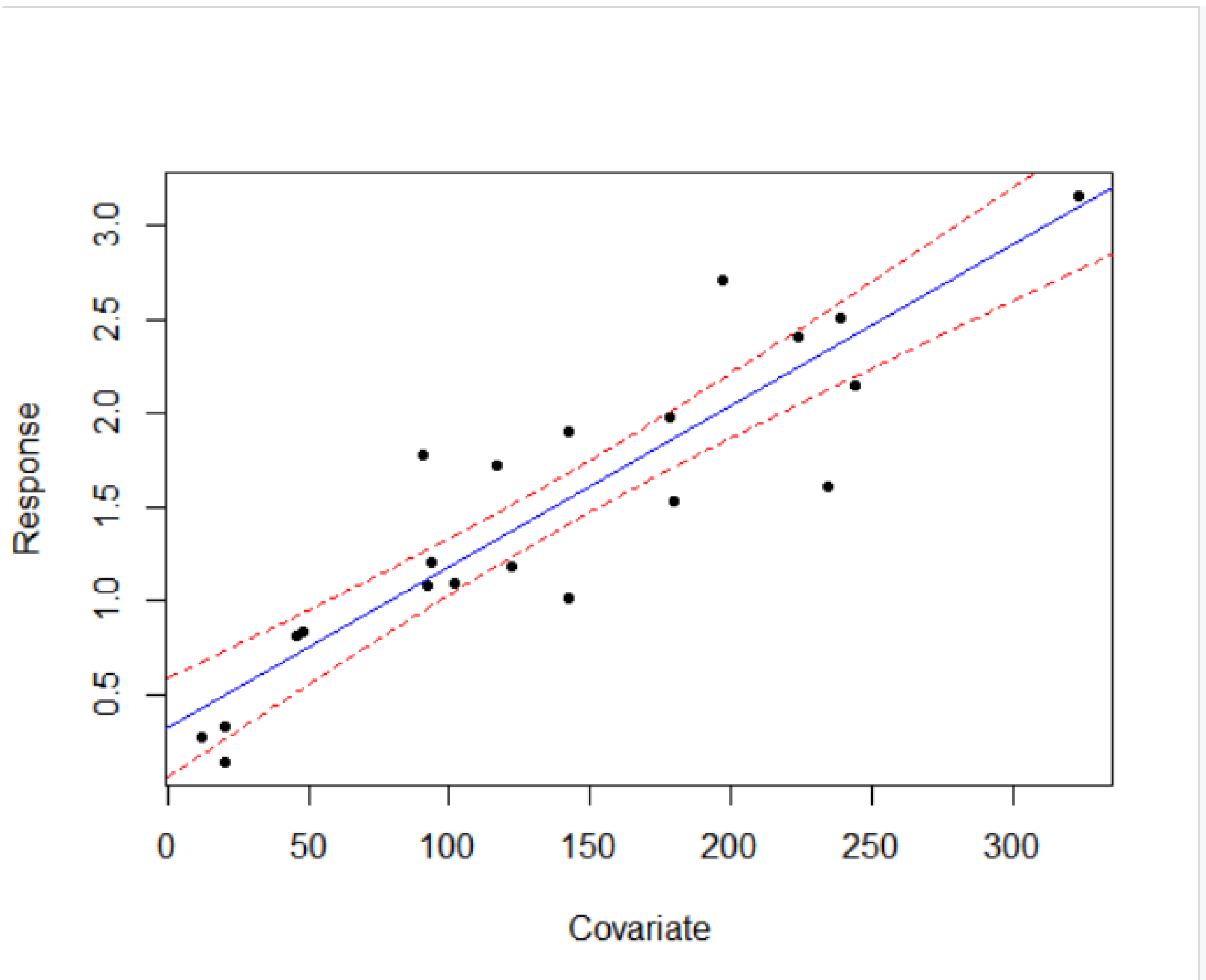
Code:

```
predict.lm(model2, interval="confidence")

newx<-seq(0,500)
conf<-predict(model2,newdata=data.frame(Covariate=newx),interval = c("confidence"),level = 0.90,type="response")
plot(Covariate,Response,pch=20)
model2 <- lm(Response ~ Covariate, data=df)
abline(model2,col="blue")
lines(newx,conf[,3],col="red",lty=2)
lines(newx,conf[,2],col="red",lty=2)

```

Output:



4. Using the output from summary(), suppose we want to test for null hypothesis of H_0: $\beta_1 = 0.01$ against the alternative hypothesis H_1: $\beta_1 \neq 0.01$, what do you conclude? Reject or not reject? Explain why.

Solution:

The code:

```
summary(model2)
```

```
> summary(model2) #summary

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(lm(Response~Covariate, offset=0.01*Covariate))

```

```
Call:
lm(formula = Response ~ Covariate, offset = 0.01 * Covariate)

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   -0.0014067  0.0009499  -1.481   0.1551  
---
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 t value is < 0. Thus we reject the hypothesis H0: $\boldsymbol{\beta_1 = 0.01}$.

5. Using the output from summary(), suppose we want to test for null hypothesis of $H_0: \beta_1 = 0.02$ against the alternative hypothesis $H_1: \beta_1 \neq 0.02$, what do you conclude? Reject or not reject? Explain why.

```
> summary(model2)

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(lm(Response~Covariate, offset=0.02*Covariate))

Call:
lm(formula = Response ~ Covariate, offset = 0.02 * Covariate)

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   -0.0114067  0.0009499 -12.008 2.56e-10 ***
---
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

> 
```

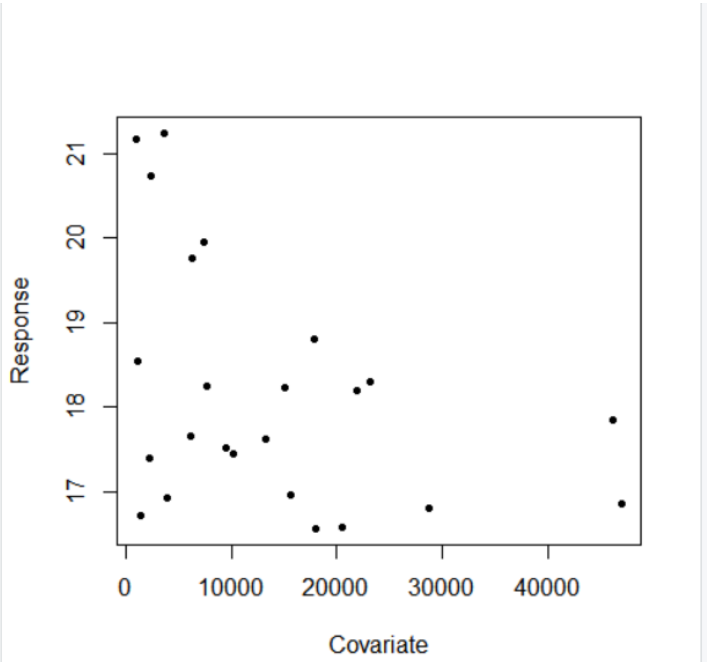
We reject the null hypothesis $H_0: \beta_1 = 0.02$.

6. Repeat the same questions (1-3) 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).

6.1. Read <bus.csv> into R. Use response variable = Expenses per car mile (pence), covariate = Car miles per year (1000s).

Solution:

```
getwd()
setwd("C:/Users/Susheela/Documents/data analysis/assignment 3")
getwd()
df <- read.csv("bus.csv", header=TRUE)
names(df)[1]="Response"
names(df)[2]="Covariate"
names(df)
attach(df)
Response
Covariate
plot(Covariate,Response,pch=20)
```



6.2. Obtain R^2, explain what it means.

Solution:

Code:

```
model2 <- lm(Response~Covariate,data=df)
plot(Response~Covariate)
abline(model2,lwd=3)
summary(model2)$r.square  # r square value
```

Output:

```
> abline(model2,lwd=3)
> summary(model2)$r.square  # r square value
[1] 0.1582641
>
```

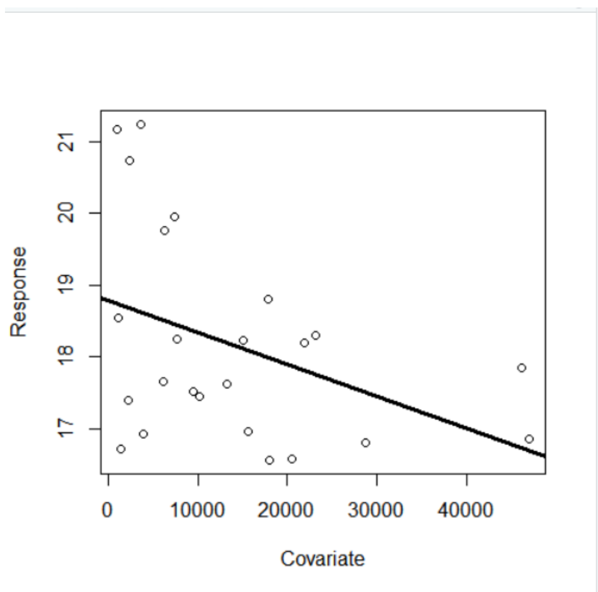
Observations of R^2 value:

Since R^2 is a proportion, it is always a number between 0 and 1.

If R^2 = 1, all of the data points fall perfectly on the regression line. The predictor x accounts for all of the variation in y.

If R^2 = 0, the estimated regression line is perfectly horizontal. The predictor x accounts for none of the variation in y.

The R^2=0.1582641 means that the data points are far from the fitted regression line.



6.3. Perform the following hypothesis testing and interval estimation using lm() and other related R functions.

6.3.1. Perform t tests, obtain t statistics and p values, interpret the results, make a conclusion (i.e. reject or not reject) and explain why. Note: please explain what the null hypothesis is.

Solution:

Code:

```
summary(model2)$coef[,3] # t value
summary(model2)$coef[,4] # p value
```

Output:

```
> summary(model2)$coef[,3] # t value
(Intercept)    Covariate 
   46.08506    -2.03383 
> summary(model2)$coef[,4] # p value
 (Intercept)     Covariate 
2.223005e-23 5.420264e-02 
> 
```

The Null hypothesis H0: β_0=0 is rejected.

The coefficient t-value is a measure of how many standard deviations our coefficient estimate is far away from 0. We want it to be far away from zero as this would indicate we could reject the null

We reject the null hypothesis and conclude that there is a relationship between response and covariate as the t value is greater than 0 and corresponding p values are less which leads to rejection of null hypothesis.

6.3.2. Perform ANOVA test (F test), obtain F statistic and p value, interpret the results, make conclusion (i.e. reject or not reject) and explain why. Note: please explain what the null hypothesis is.

Solution:

Code:

```
summary(model2) # f statistic
```

Output:

```
Console  Terminal
~/data analysis/assignment 3/
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

> 
```

The Null hypothesis H0: β_0=0 is rejected.

6.3.3. Compute confidence interval for coefficients, fitted values (mean response), interpret the meanings of these quantities.

Solution:

Code:

```
confint(model2,level=0.95)
confint(model2,level=0.90)
confint(model2,level=0.99)

predict.lm(model2, interval="confidence")
```

Output:

```
Console  Terminal ×
~/data analysis/assignment 3/
> confint(model2,level=0.95)
                    2.5 %       97.5 %
(Intercept)  1.793660e+01 1.962700e+01
Covariate   -8.987441e-05 8.761294e-07
> confint(model2,level=0.90)
                      5 %          95 %
(Intercept)  1.808199e+01  1.948162e+01
Covariate   -8.206937e-05 -6.928910e-06
> confint(model2,level=0.99)
                    0.5 %       99.5 %
(Intercept) 17.6330280571 1.993058e+01
Covariate   -0.0001061721 1.717378e-05
> predict.lm(model2, interval="confidence")
        fit      lwr      upr
1  18.50435 17.83996 19.16874
2  16.72461 15.14429 18.30493
3  18.45429 17.81459 19.09398
4  17.50401 16.61724 18.39078
5  17.80576 17.12523 18.48629
6  18.72231 17.92084 19.52377
7  17.98611 17.38583 18.58639
8  18.67861 17.90780 19.44942
9  17.97904 17.37647 18.58161
10 18.73076 17.92321 19.53831
11 18.68497 17.90978 19.46016
12 18.19143 17.62077 18.76209
13 18.62245 17.88895 19.35595
14 18.10969 17.53614 18.68324
15 16.68994 15.07660 18.30328
16 18.33063 17.73733 18.92392
17 18.50827 17.84182 19.17471
18 17.75436 17.04387 18.46486
19 17.86734 17.21895 18.51574
20 18.36129 17.75861 18.96396
21 18.73606 17.92468 19.54744
22 18.61057 17.88462 19.33651
23 18.08512 17.50835 18.66190
24 18.43805 17.80568 19.07041
> 
```

6.3.4. Plot data points, the regression line, the confidence interval for fitted values (to show that the interval is wider on both sides and narrow in the center).

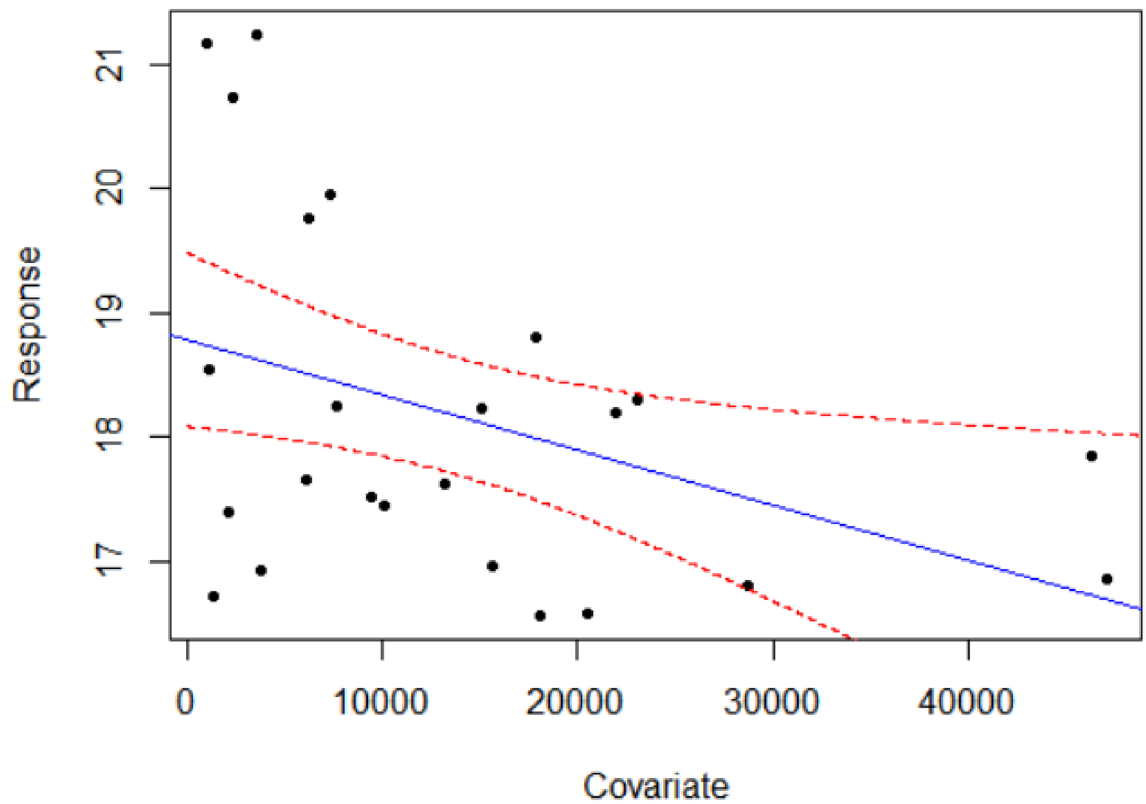
Solution:

Code:

```
newx<-seq(0,49000)
conf<-predict(model2,newdata=data.frame(Covariate=newx),interval = c("confidence"),level = 0.90,type="response")
plot(Covariate,Response,pch=20)
model2 <- lm(Response ~ Covariate, data=df)
abline(model2,col="blue")
lines(newx,conf[,3],col="red",lty=2)
lines(newx,conf[,2],col="red",lty=2)

```

Output:



21
20
19
18
17
Response
0
10000
20000
30000
40000
Covariate