Practical-3

Soumarya Basak

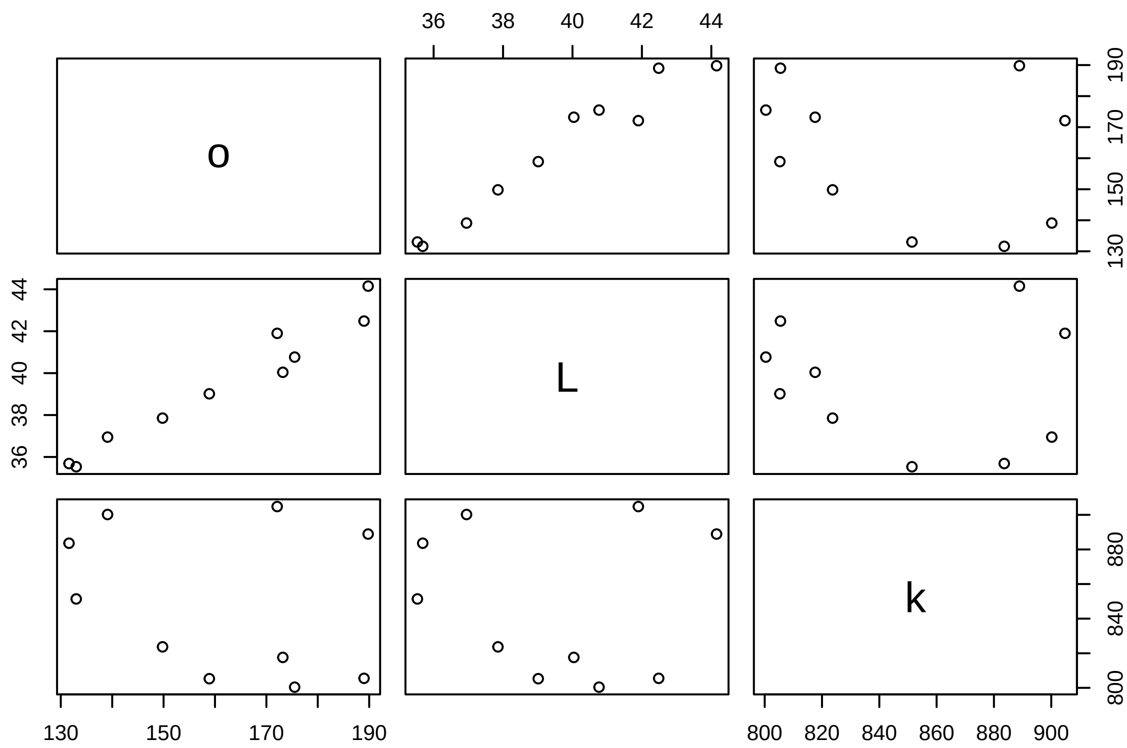
Problem 1

Data Input

```
library(readr)
df<- read.csv("C:\\Users\\souma\\Dropbox\\Mstat_CU\\Sem 2\\Regression_analysis_1\\Data Sets\\cob-doglas
```

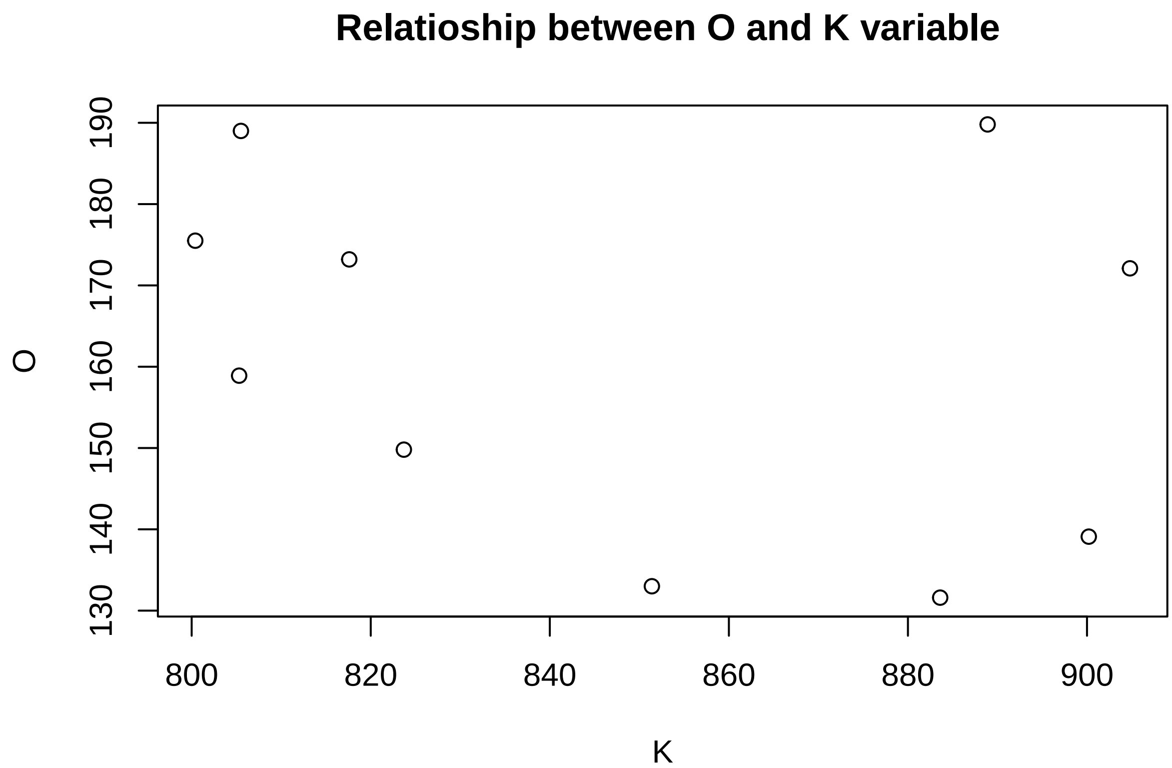
Scatter Plot

```
plot(df[,-1])
```



We can see that `k` and `o` are not linearly related.

```
plot(df$k, df$o, xlab="K", ylab = "O", main = " Relatioship between O and K variable")
```



Lets go for Cobb-Doglas Model

$$O_t = \alpha L_t^{\beta_1} k_t^{\beta_2} u_t$$
$$logO_t = \alpha^* + \beta_1 logL_t + \beta_2 logK_t + v_t$$

Let's Fit the model

```
reg1<- lm(log(o)~log(L)+log(k),data = df)
summary(reg1)
```

```
##
## Call:
## lm(formula = log(o) ~ log(L) + log(k), data = df)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0186868 -0.0056866 -0.0004878  0.0031244  0.0296869
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.22246    0.72109   3.082  0.01776 *
## log(L)       1.77558    0.06505  27.298 2.27e-08 ***
## log(k)      -0.54416    0.09704  -5.607  0.00081 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01447 on 7 degrees of freedom
```

```
## Multiple R-squared:  0.9915, Adjusted R-squared:  0.9891
## F-statistic: 410.5 on 2 and 7 DF,  p-value: 5.557e-08
```

Test for Durbin Watson Test

```r
library(car)
```

```
## Warning: package 'car' was built under R version 4.1.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.1.3
```

```r
durbinWatsonTest(reg1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.2397808     0.8869685   0.002
##  Alternative hypothesis: rho != 0
```

So, under 95% level of significant we accept that there is Autocorrelation in the data.

The value of Durbin Watson Test statistics is , d= 0.8869685

Remedial Measure

```r
library(orcutt)
```

```
## Warning: package 'orcutt' was built under R version 4.1.3
```

```
## Loading required package: lmtest
```

```
## Warning: package 'lmtest' was built under R version 4.1.3
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.1.3
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
cochrane.orcutt(reg1, convergence =5, max.iter=1000)
```

```
## Cochrane-orcutt estimation for first order autocorrelation
##
## Call:
## lm(formula = log(o) ~ log(L) + log(k), data = df)
##
##  number of interaction: 311
##  rho 0.996755
##
## Durbin-Watson statistic
## (original):    0.88697 , p-value: 7.837e-04
## (transformed): 2.34510 , p-value: 4.69e-01
##
##  coefficients:
## (Intercept)      log(L)      log(k)
```

```
##    1.894243    1.759270   -0.136710
```

The Final Model

$$logO_t = 1.894243 + (1.759270 \times logL_t) - (0.136710 \times logK_t)$$

Problem 2

Data

```
library(readr)
df<- read.csv("C:\\Users\\souma\\Dropbox\\Mstat_CU\\Sem 2\\Regression_analysis_1\\Data Sets\\coal.csv")
head(df)
```

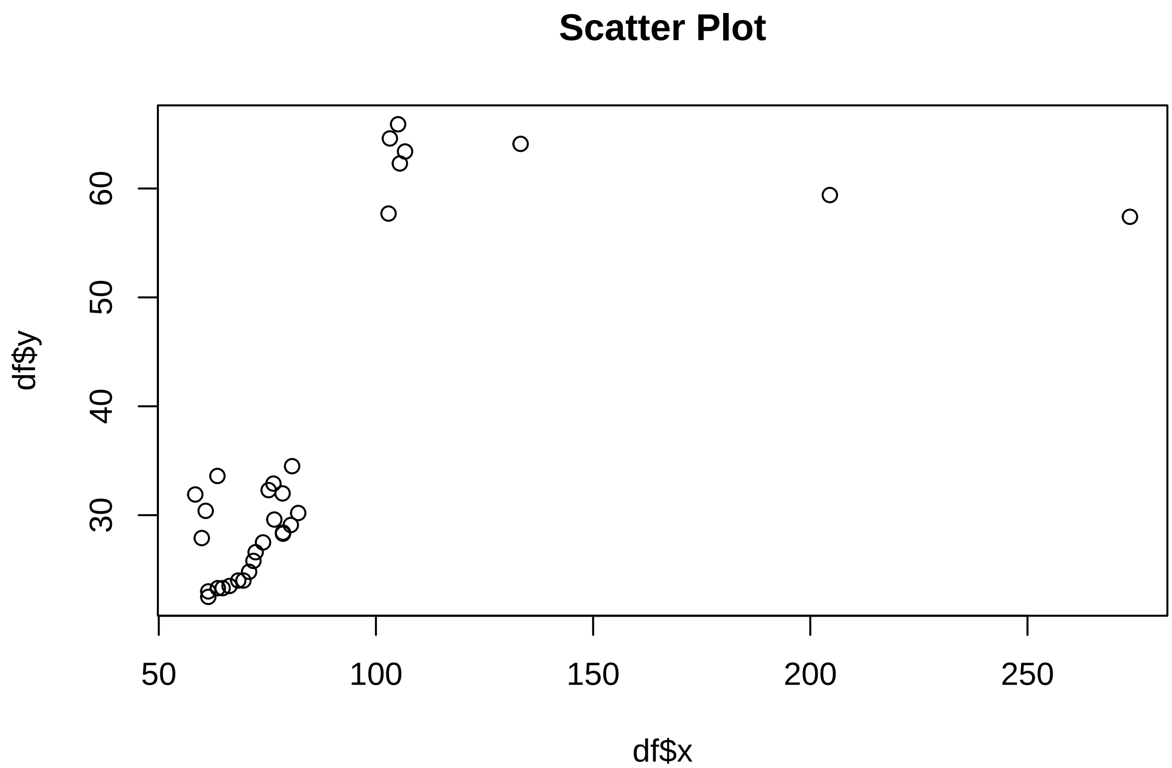
```
##   ï..Year Price.of.oil.cents. Price.of.Bit.Coal
## 1    1950                80.7              34.5
## 2    1951                76.4              32.9
## 3    1952                75.3              32.3
## 4    1953                78.5              32.0
## 5    1954                80.4              29.1
## 6    1955                78.6              28.4
```

```
colnames(df)<- c("sl.no","x","y")
head(df)
```

```
##   sl.no    x    y
## 1  1950 80.7 34.5
## 2  1951 76.4 32.9
## 3  1952 75.3 32.3
## 4  1953 78.5 32.0
## 5  1954 80.4 29.1
## 6  1955 78.6 28.4
```

Scatter Plot

```
plot(df$x,df$y,main = "Scatter Plot")
```



Lets fit the model

```
reg2<- lm(y~x,data=df)
summary(reg2)
```

```
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.7402  -7.2003  -4.3925   0.9037  25.6624
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.97256    4.70852   3.392  0.00196 **
## x            0.23088    0.04783   4.827  3.8e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.72 on 30 degrees of freedom
## Multiple R-squared:  0.4372, Adjusted R-squared:  0.4184
## F-statistic:  23.3 on 1 and 30 DF,  p-value: 3.795e-05
```

Check for influencial Observation

```
influence.measures(reg2)
```

```
## Influence measures of
##   lm(formula = y ~ x, data = df) :
##
##       dfb.1_     dfb.x    dffit cov.r   cook.d    hat inf
## 1  -0.000958  0.000284 -0.00162 1.106 1.36e-06 0.0322
## 2  -0.007519  0.003023 -0.01133 1.107 6.64e-05 0.0336
## 3  -0.011552  0.004908 -0.01696 1.107 1.49e-04 0.0341
## 4  -0.020748  0.007345 -0.03298 1.104 5.62e-04 0.0329
## 5  -0.050639  0.015428 -0.08500 1.090 3.71e-03 0.0323
## 6  -0.056624  0.019908 -0.09026 1.088 4.18e-03 0.0328
## 7  -0.042706  0.016985 -0.06465 1.098 2.15e-03 0.0336
## 8  -0.041430  0.010552 -0.07337 1.093 2.77e-03 0.0319
## 9  -0.057623  0.020259 -0.09186 1.088 4.33e-03 0.0328
## 10 -0.063283  0.028475 -0.09030 1.091 4.18e-03 0.0347
## 11 -0.072476  0.034802 -0.09993 1.089 5.12e-03 0.0356
## 12 -0.081852  0.039989 -0.11179 1.085 6.39e-03 0.0358
## 13 -0.093776  0.047325 -0.12576 1.079 8.06e-03 0.0364
## 14 -0.103547  0.054298 -0.13582 1.076 9.39e-03 0.0372
## 15 -0.102985  0.055765 -0.13253 1.079 8.95e-03 0.0380
## 16 -0.108670  0.061713 -0.13583 1.080 9.40e-03 0.0394
## 17 -0.110333  0.064802 -0.13504 1.083 9.29e-03 0.0406
## 18 -0.109310  0.065577 -0.13202 1.086 8.89e-03 0.0415
## 19 -0.119387  0.074419 -0.14072 1.086 1.01e-02 0.0434
## 20 -0.111476  0.069488 -0.13140 1.090 8.81e-03 0.0434
## 21 -0.030443  0.019427 -0.03535 1.118 6.46e-04 0.0448
## 22  0.006124 -0.003854  0.00717 1.119 2.66e-05 0.0439
## 23  0.040377 -0.026328  0.04625 1.119 1.10e-03 0.0462
## 24  0.043931 -0.026403  0.05299 1.112 1.45e-03 0.0416
## 25  0.040178  0.096395  0.30367 0.936 4.38e-02 0.0348
## 26  0.040584  0.166574  0.46307 0.773 9.25e-02 0.0359   *
## 27  0.022870  0.158991  0.40856 0.839 7.50e-02 0.0368
## 28  0.030878  0.142333  0.38764 0.857 6.83e-02 0.0361
## 29  0.055241  0.141585  0.43791 0.792 8.39e-02 0.0349   *
## 30 -0.140544  0.296946  0.41264 0.973 8.12e-02 0.0648
## 31  0.148652 -0.202602 -0.21625 1.424 2.41e-02 0.2557   *
## 32  3.264561 -4.106430 -4.21730 1.363 6.55e+00 0.6023   *
```

So, we get 4 influential measures— 26,29,31,32 th observation

Lets remove them and fit the data again

```
df1<- df[-c(26,29,31,32),]

reg2<- lm(y~x, data=df1)
summary(reg2)
```

```
##
## Call:
## lm(formula = y ~ x, data = df1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

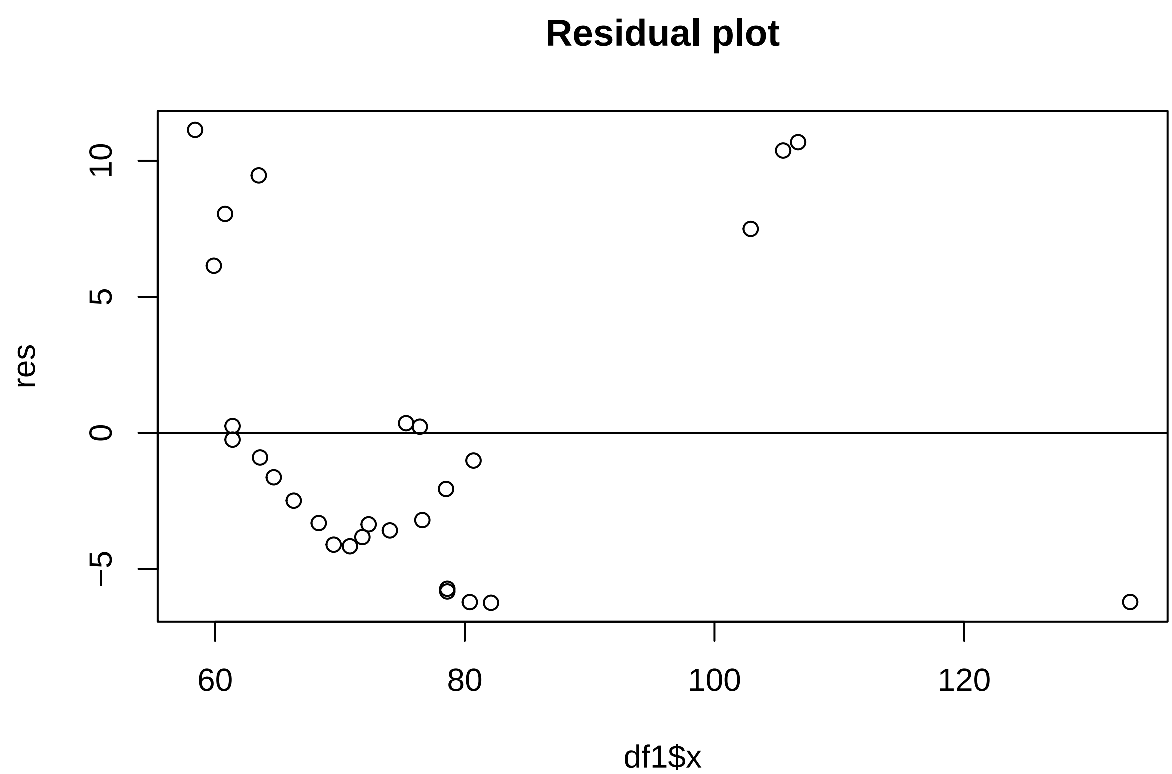
```
## -6.244 -3.900 -1.848  1.802 11.135
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.86915    5.16723  -3.458  0.00189 **
## x             0.66155    0.06598  10.027    2e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.84 on 26 degrees of freedom
## Multiple R-squared:  0.7945, Adjusted R-squared:  0.7866
## F-statistic: 100.5 on 1 and 26 DF,  p-value: 2.005e-10
```

SO, after removing 4 observation, the model fits more well as adjusted R^2 increase.

Lets check for residuals

```
res<- resid(reg2)

# residual plot
plot(df1$x,res,main = "Residual plot")
abline(h=0)
```



Its, unclear from the residual plot that is there heterosecdasticity present or not.

So, lets test it, ##### GQ Test

```
library(lmtest)
gqtest(reg2, fraction = 8 ,order.by = ~x,data = df1)
```

```
##
##  Goldfeld-Quandt test
##
## data:  reg2
## GQ = 3.1441, df1 = 8, df2 = 8, p-value = 0.06281
## alternative hypothesis: variance increases from segment 1 to 2
```

The p-value is =0.06281 So, we fail to reject the null, SO the data is homoscedastic

Check for Autocorrelation

Durbin Watson Test

```
library(car)
durbinWatsonTest(reg2)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.7580202     0.4392214       0
##  Alternative hypothesis: rho != 0
```

Since the p value is less than 0.05

SO, we reject the null that autocorrelation is absent

So, there is Autocorrelation in the data

Remedial Measure

```
library(orcutt)
cochrane.orcutt(reg2, convergence = 8, max.iter = 1000)
```

```
## Cochrane-orcutt estimation for first order autocorrelation
##
## Call:
## lm(formula = y ~ x, data = df1)
##
##  number of interaction: 16
##  rho 0.925707
##
## Durbin-Watson statistic
## (original):    0.43922 , p-value: 4.543e-08
## (transformed): 1.13436 , p-value: 6.381e-03
##
##  coefficients:
## (Intercept)           x
##    3.235024    0.428367
```

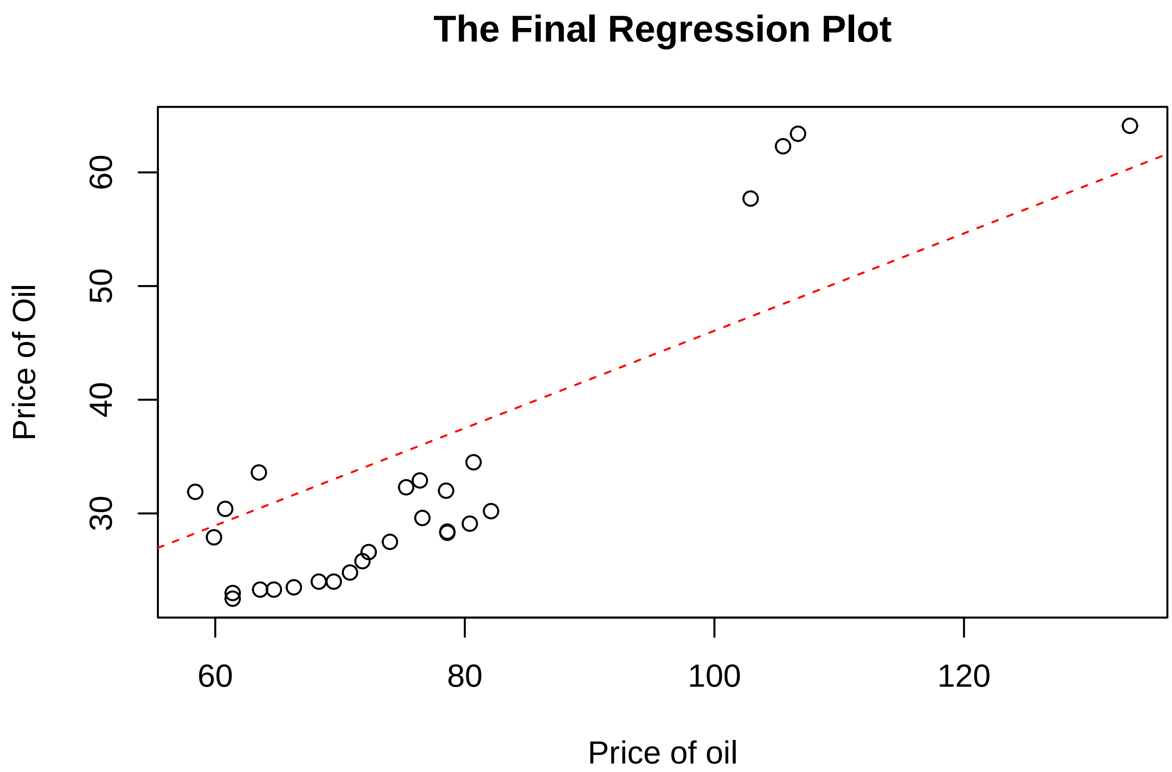
The Final MOdel

$$y = 3.235024 + 0.428367 \times x$$

where y = Price of Bit.coal, and x = Price of oil

Scatter Plot

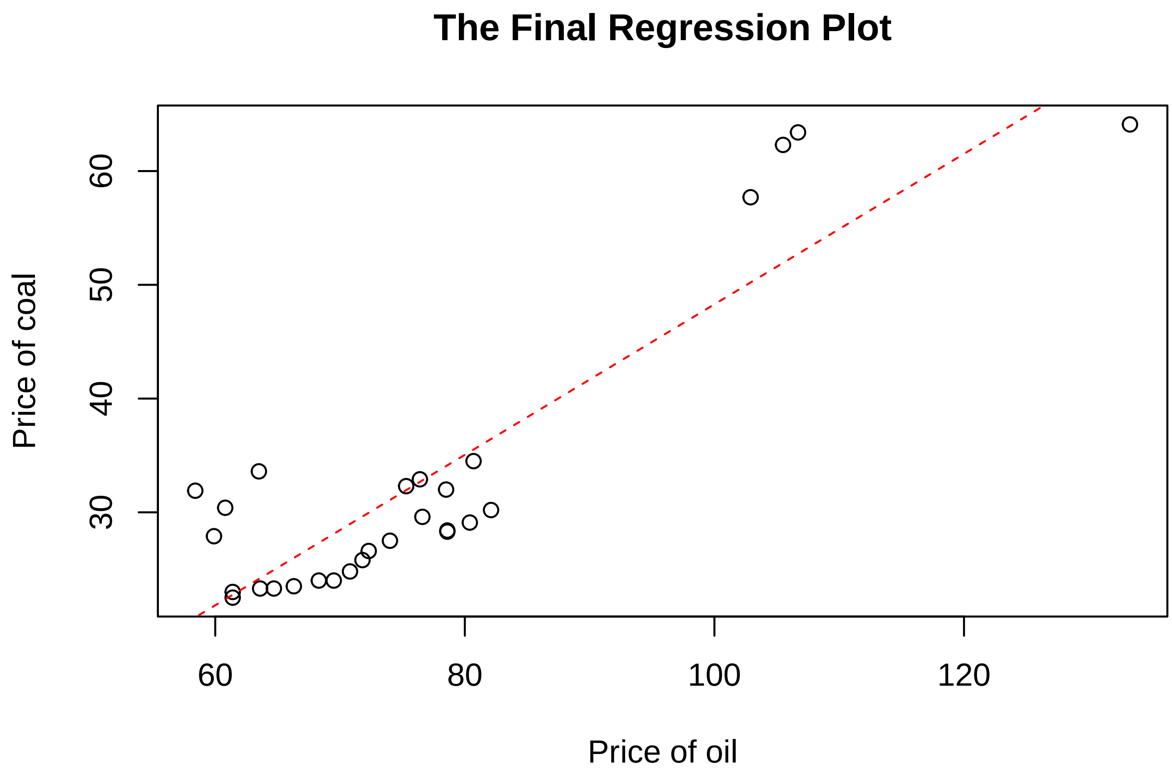
```
plot(df1$x,df1$y,main = "The Final Regression Plot",xlab="Price of oil",ylab = "Price of Oil")
abline(3.235024,0.428367,col='red',lty=2)
```



But as here is only one regressor w can't find the auto correlation among the regressor, and the GQ test also significes the absence of heteroscedasticity so we will stick to the model after removing the influential observation and the model is

$$y = -17.86915 + 0.66155 \times X$$

```
plot(df1$x,df1$y,main = "The Final Regression Plot",xlab="Price of oil",ylab = "Price of coal")
abline(-17.86915,0.66155  ,col='red',lty=2)
```


The Final Regression Plot
Price of coal
60
50
40
30
60
80
100
120
Price of oil