Predicting the amount of currency in the economy

Srikar

2021-12-01

Aim: To build a regression model to predict the amount of currency in the economy

Data Description:

The data observes the various monetary variables in India over a period of 58 months. The variables in the data set are:

1) M1(the amount of currency in circulation)

2) IIP (the index of inflation),

3) INT (the bank interest rates),

4) UPI (the value of UPI transactions in that month)

5) CC (the value of credit card transactions in that month),

6)DC (the value of debit card transactions in that month).

```
library(readxl)
dat <- read_excel("C:/Users/Srikar/Downloads/Book2.xlsx")
head(dat)
```

```
A tibble: 6 x 6
```

	M1	IIP	INT	UPI	CC	DC
1	9.93	4.97	7.5	-4.71	18.2	20.6
2	9.92	4.86	7.5	-4.71	18.2	20.6
3	9.93	4.91	7.5	-4.71	18.3	20.6
4	9.98	4.92	7.3	-4.61	18.2	20.6
5	9.95	4.84	7.3	-3.54	18.3	20.7
6	9.68	4.86	7.1	-1.61	18.4	20.5

Data Summary:

```
summary(dat)
```

```
       M1              IIP             INT              UPI
 Min.   : 9.651   Min.   :3.989   Min.   :5.500   Min.   :-4.711
 1st Qu.: 9.943   1st Qu.:4.786   1st Qu.:6.450   1st Qu.: 1.358
 Median :10.069   Median :4.837   Median :6.900   Median : 4.044
 Mean   :10.058   Mean   :4.819   Mean   :6.758   Mean   : 2.734
 3rd Qu.:10.158   3rd Qu.:4.879   3rd Qu.:7.287   3rd Qu.: 4.823
 Max.   :10.327   Max.   :4.971   Max.   :7.500   Max.   : 5.610
       CC              DC
 Min.   :18.16   Min.   :20.04
 1st Qu.:18.55   1st Qu.:20.65
 Median :18.75   Median :20.74
 Mean   :18.72   Mean   :20.73
 3rd Qu.:18.93   3rd Qu.:20.87
 Max.   :19.14   Max.   :21.00
```

We observe that the minimum amount of currency circulation in the economy is 9.65 and the maximum is 10.327. The index of inflation ranges between 3.989 to 4.971. The bank interest rate is between 5.5 and 7.5 The range of value of UPI transactions are between -4.711 to 5.610.
The value of credit card transactions lies between 18.16 to 19.14. The value of debit transactions lie between 20 to 21.

Procedure

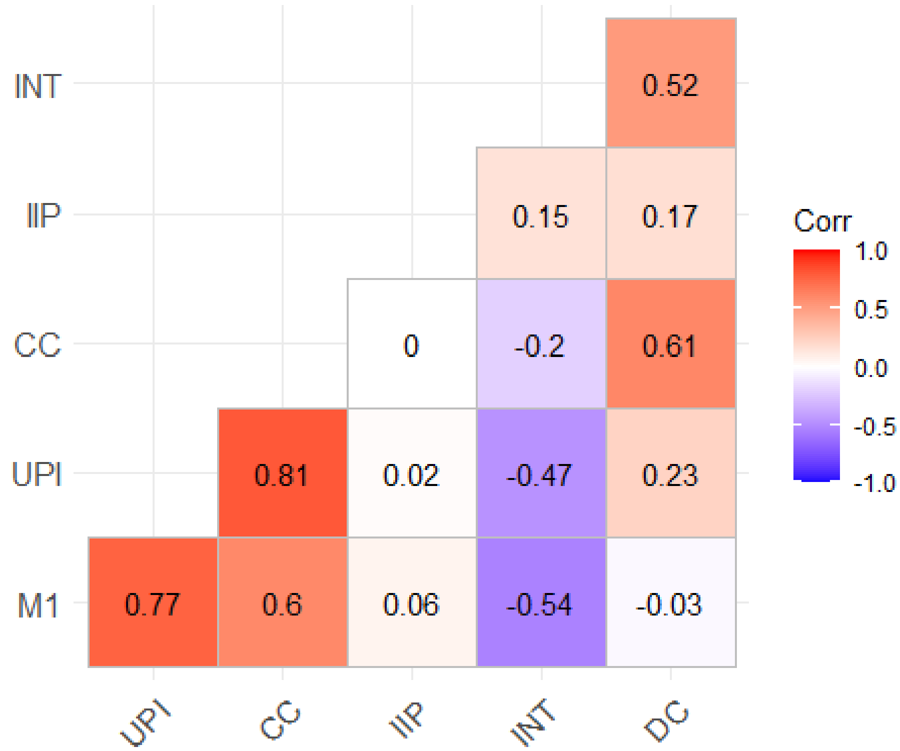
1.Visualizing Correlation plot

```
r=cor(dat)

library(ggplot2)

library(ggcorrplot)

ggcorrplot(r, hc.order = TRUE, type = "lower",lab = TRUE)
```



From this table, we observe that there is high correlation between UPI value and value of credit card transactions. There is also high correlation between UPI and the money circulated. There is medium correlation between the credit and debit card valuations.

We can observe low correlation with respect to credit card and bank interest rates. There is no correlation between credit card and index of inflation.

2. Building the regression model

```
mod=lm(dat$M1~.,data = dat)
summary(mod)
```

```
   Call:
   lm(formula = dat$M1 ~ ., data = dat)

   Residuals:
         Min        1Q    Median        3Q       Max
   -0.284421 -0.040434  0.008267  0.042592  0.176334

   Coefficients:
                Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 12.134680   1.944621   6.240 8.02e-08 ***
IIP          0.151368   0.091941   1.646  0.10572
INT          0.001321   0.032382   0.041  0.96761
UPI          0.022923   0.008558   2.679  0.00987 **
CC           0.303514   0.130217   2.331  0.02367 *
DC          -0.412858   0.160221  -2.577  0.01285 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08843 on 52 degrees of freedom
Multiple R-squared:  0.6853, Adjusted R-squared:  0.655
F-statistic: 22.65 on 5 and 52 DF,  p-value: 5.495e-12
```

We observe that intercept is significant as its p-value is lesser than the significance value (0.05). This means that if all the values
were 0, then the average circulation of money would be 12.13 thousand crores.

We also observe the variables UPI, CC and DC to be significant as their p-values a re lesser than the significance value (0.05).Variables such as IIP and INT have no significant relationship with the money circulated.

The overall model is significant as the p-value is way below the significance level (0.05). The R-squared value shows 0.6853 which means that the independent variables only contribute to 68.53% of variation of the dependent variable M1, the money circulated in the economy. The adjusted R-Square which is obviously lesser shows the amount of variation that is described by variables that are actually significant.

We can confirm the variables that matter by using the ANOVA model as well.

```
library(stats)
anova(mod)
```

```
Analysis of Variance Table

Response: dat$M1
          Df  Sum Sq Mean Sq F value    Pr(>F)
IIP        1 0.00527 0.00527  0.6738   0.41546
INT        1 0.40030 0.40030 51.1918 2.803e-09 ***
UPI        1 0.42593 0.42593 54.4695 1.225e-09 ***
CC         1 0.00199 0.00199  0.2549   0.61575
DC         1 0.05192 0.05192  6.6399   0.01285 *
Residuals 52 0.40662 0.00782
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the anova model, we can see that even INT shows significance on the model which varies from our intimal summary of our regression model.
We also observe that from the ANOVA model, we can see that CC does not show significance as opposed to our previous model.

The number of reasons could be numerous but to begin with we can check the assumption of multiple regression and checkwhether the model satisfies these are not.

3) Checking the important assumptions

(i) <u>The multicollinearity assumption</u>

```
library(car)

   Warning: package 'car' was built under R version 3.6.3

   Loading required package: carData

vif(mod)

        IIP      INT      UPI       CC       DC
   1.085964 3.350272 4.604383 8.609871 5.735451
```

Since we see that the vif value of the regressors are not above 10, we say that assumption of no multicollinearity is satisfied.
This means that no two variables interact with each other to have an impact on the money circulation in the economy.

(ii) <u>Autocorrelation assumption</u>

```
library(lmtest)

   The following objects are masked from 'package:base':

       as.Date, as.Date.numeric

dwtest(dat$IIP~dat$UPI)


    Durbin-Watson test

   data:  dat$IIP ~ dat$UPI
```

```
   DW = 1.0754, p-value = 4.874e-05
   alternative hypothesis: true autocorrelation is greater than 0

dwtest(dat$IIP~dat$INT)


    Durbin-Watson test

   data:  dat$IIP ~ dat$INT
   DW = 1.1187, p-value = 0.0001156
   alternative hypothesis: true autocorrelation is greater than 0

dwtest(dat$IIP~dat$CC)


    Durbin-Watson test

   data:  dat$IIP ~ dat$CC
   DW = 1.0749, p-value = 5.199e-05
   alternative hypothesis: true autocorrelation is greater than 0

dwtest(dat$IIP~dat$DC)


    Durbin-Watson test

   data:  dat$IIP ~ dat$DC
   DW = 1.129, p-value = 0.0001443
   alternative hypothesis: true autocorrelation is greater than 0

dwtest(dat$INT~dat$UPI)


    Durbin-Watson test

   data:  dat$INT ~ dat$UPI
   DW = 0.44828, p-value = 2.936e-14
   alternative hypothesis: true autocorrelation is greater than 0

dwtest(dat$INT~dat$IIP)


    Durbin-Watson test

   data:  dat$INT ~ dat$IIP
   DW = 0.39514, p-value = 2.405e-15
   alternative hypothesis: true autocorrelation is greater than 0

dwtest(dat$INT~dat$CC)
```

```
    Durbin-Watson test

   data:  dat$INT ~ dat$CC
   DW = 0.39253, p-value = 9.467e-16
   alternative hypothesis: true autocorrelation is greater than 0
```

```r
dwtest(dat$INT~dat$DC)
```

```
    Durbin-Watson test

   data:  dat$INT ~ dat$DC
   DW = 0.57223, p-value = 2.301e-11
   alternative hypothesis: true autocorrelation is greater than 0
```

```r
dwtest(dat$UPI~dat$CC)
```

```
    Durbin-Watson test

   data:  dat$UPI ~ dat$CC
   DW = 0.54395, p-value = 5.637e-12
   alternative hypothesis: true autocorrelation is greater than 0
```

```r
dwtest(dat$UPI~dat$DC)
```

```
    Durbin-Watson test

   data:  dat$UPI ~ dat$DC
   DW = 0.040967, p-value < 2.2e-16
   alternative hypothesis: true autocorrelation is greater than 0
```

```r
dwtest(dat$CC~dat$DC)
```

```
    Durbin-Watson test

   data:  dat$CC ~ dat$DC
   DW = 0.10012, p-value < 2.2e-16
   alternative hypothesis: true autocorrelation is greater than 0
```

We can see that the assumption of autocorrelation has been fulfilled as all of them have a p-value less than significance level.

(iii) <u>Heteroscedascity assumption)</u>

```r
library(lmtest)
bptest(mod)
```

```
    studentized Breusch-Pagan test

   data:  mod
   BP = 19.999, df = 5, p-value = 0.00125
```

Since the p-value is lesser than 0.05, the significance level, we say that it is significant and that the it is heteroscedastic.

Since it fulfills the assumptions given above, we can test out by only taking factors that have significant impact on the dependent variable.

4) Performing stepwise procedure

```
step(mod, direction="both")
```

```
   Start:  AIC=-275.7
   dat$M1 ~ IIP + INT + UPI + CC + DC

          Df Sum of Sq     RSS     AIC
   - INT   1  0.000013 0.40663 -277.70
   <none>              0.40662 -275.70
   - IIP   1  0.021195 0.42782 -274.75
   - CC    1  0.042483 0.44910 -271.94
   - DC    1  0.051922 0.45854 -270.73
   - UPI   1  0.056105 0.46273 -270.20

   Step:  AIC=-277.7
   dat$M1 ~ IIP + UPI + CC + DC

          Df Sum of Sq     RSS     AIC
   <none>              0.40663 -277.70
   - IIP   1  0.021186 0.42782 -276.75
   + INT   1  0.000013 0.40662 -275.70
   - CC    1  0.051714 0.45835 -272.75
   - UPI   1  0.056493 0.46313 -272.15
   - DC    1  0.113179 0.51981 -265.45


   Call:
   lm(formula = dat$M1 ~ IIP + UPI + CC + DC, data = dat)

   Coefficients:
   (Intercept)          IIP          UPI           CC           DC
      12.08668      0.15133      0.02289      0.30119     -0.40800
```

We observe that after adding and subtracting the variables, we get a model that is built only on significant variables. This model is more reliable and will possess less error. We have removed the variable INT from the model

New model

```r
nmod=lm(dat$M1 ~ IIP + UPI + CC + DC, data = dat)

summary(nmod)
```

```
   Call:
   lm(formula = dat$M1 ~ IIP + UPI + CC + DC, data = dat)

   Residuals:
         Min        1Q    Median        3Q       Max
   -0.283705 -0.040112  0.008579  0.042322  0.176286

   Coefficients:
                Estimate Std. Error t value Pr(>|t|)
   (Intercept) 12.086682   1.533774   7.880 1.74e-10 ***
   IIP          0.151326   0.091065   1.662  0.10247
   UPI          0.022888   0.008435   2.714  0.00896 **
   CC           0.301192   0.116012   2.596  0.01217 *
   DC          -0.408001   0.106229  -3.841  0.00033 ***
   ---
   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

   Residual standard error: 0.08759 on 53 degrees of freedom
   Multiple R-squared:  0.6853, Adjusted R-squared:  0.6615
   F-statistic: 28.85 on 4 and 53 DF,  p-value: 9.495e-13
```

We can see that most variables show significance except IIP which was not removed during the stepwise procedure as it might explain the variation to a minimum extent.

```r
confint(nmod)
```

```
                      2.5 %      97.5 %
   (Intercept)  9.010321694 15.16304315
   IIP         -0.031327306  0.33397841
   UPI          0.005969897  0.03980636
   CC           0.068501301  0.53388193
   DC          -0.621069955 -0.19493290
```

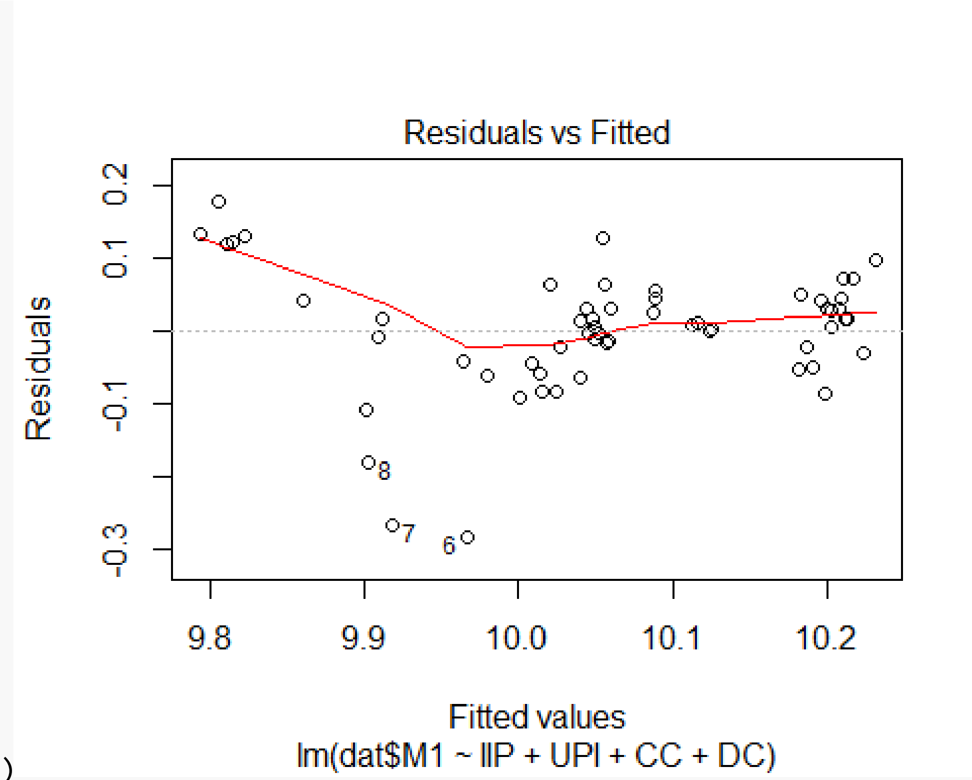
Using a 95 confidence interval, we can say that if we were to keep sampling the data, we will get these ranges of values for our independent variables between the given range.

We still see that the R-squared is more or less the same. Although the model is significant, it might not be perfect due to several factors such as missing variables, missing data or having non-normal distribution of erro

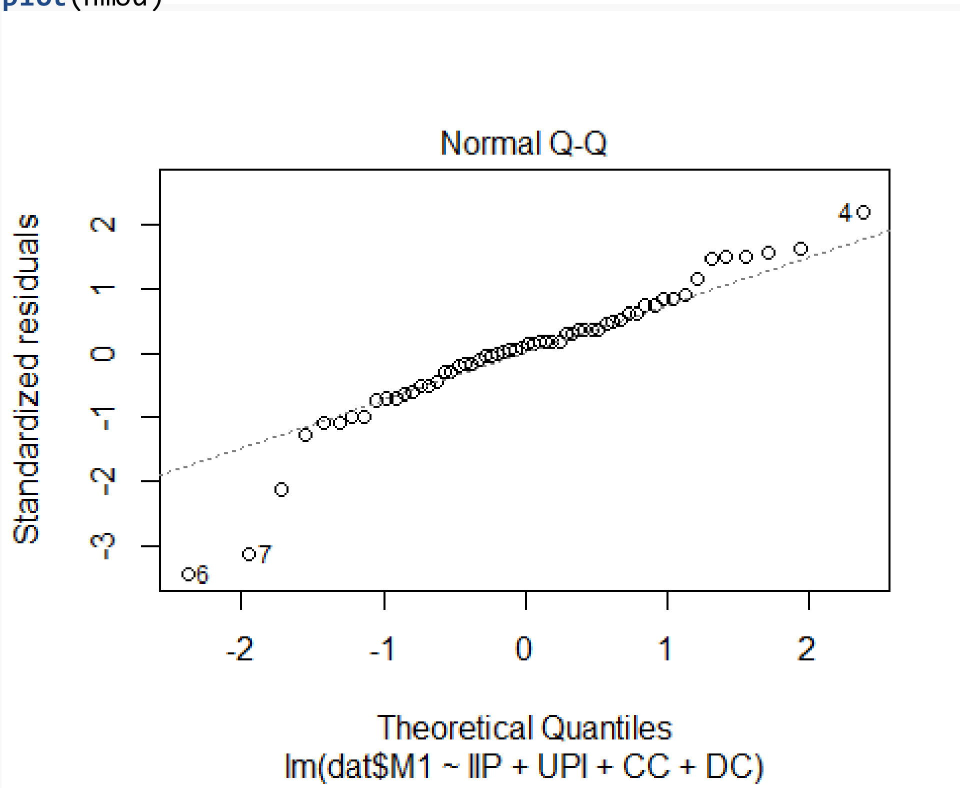
```
rs.

 We can test the assumption of normality of errors to verify the model


5) Checking the normality of error assumption
```




```
plot(nmod)
```



We observe that residual line deviates a lot fro, the dotted line showing that there i s a huge deviation from the observed balies.

We see that as the fitted values become larger, the error becomes smaller.

From the Normal Q-Q plot, we observe that the data is not exacltyu normal and is skewed towards the left. We also see some outliers in
the dataset. We can test the normlaity of errors using the Shapiro-Wilk test to clari fy.

```
shapiro.test(nmod$residuals)
```

```
    Shapiro-Wilk normality test

   data:  nmod$residuals
   W = 0.92403, p-value = 0.001384
```

As seen from the diagram and observation from the test, we observe that the data i s not normal although most of the data plots
are very close the mean. There are a few outliers as well which makes the data not exaclty normal.

6) Testing the model

Although the model might not be perfect and not fulfil assumptions of normlaity o f error due to outliers and some skewed datapoints,
we can test the models strength as for now by substituing the dependanct varibales from the dataset into the model.

Comparison of Observed and Expected value

```
library(broom)
```

```
prediction=augment(nmod)
prediction
```

```
     A tibble: 58 x 11
      `dat$M1`   IIP    UPI    CC    DC .fitted  .resid   .hat .sigma  .cooks
d
         <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>   <dbl>  <dbl>  <dbl>    <dbl
>
    1     9.93  4.97 -4.71   18.2  20.6    9.81  0.117  0.166  0.0866 0.0857
    2     9.92  4.86 -4.71   18.2  20.6    9.79  0.132  0.140  0.0862 0.0855
    3     9.93  4.91 -4.71   18.3  20.6    9.81  0.120  0.165  0.0865 0.0897
    4     9.98  4.92 -4.61   18.2  20.6    9.80  0.176  0.147  0.0844 0.164
    5     9.95  4.84 -3.54   18.3  20.7    9.82  0.130  0.101  0.0864 0.0552
```

```
    6      9.68  4.86 -1.61   18.4  20.5    9.97 -0.284  0.111  0.0780 0.295
    7      9.65  4.82 -0.807  18.6  20.8    9.92 -0.268  0.0519 0.0798 0.108
    8      9.72  4.76 -0.826  18.5  20.8    9.90 -0.181  0.0512 0.0846 0.0484
    9      9.79  4.77 -0.451  18.4  20.7    9.90 -0.108  0.0600 0.0871 0.0208
   10      9.93  4.68 -0.329  18.5  20.7    9.91  0.0159 0.0571 0.0884 0.00042
2
     ... with 48 more rows, and 1 more variable: .std.resid <dbl>
```

We see that the model predicts close to the observed values but does not exactly give us values very near to the observed values itself.
We see that our model with an R-Sqauyred of 65% is able to gives values close to the observed values itself.

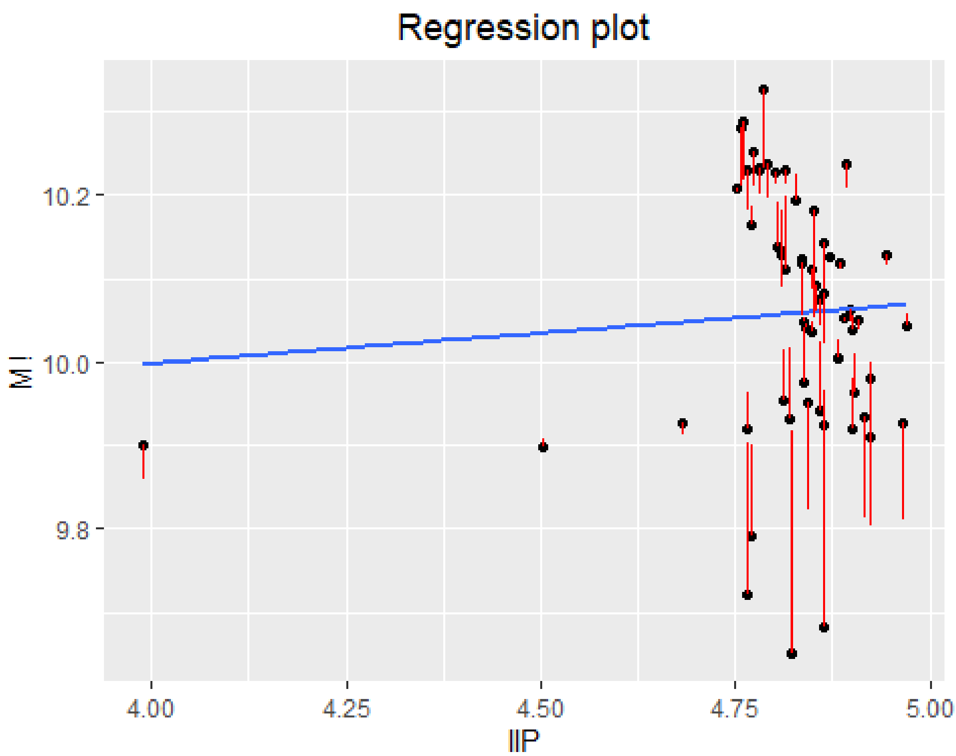
We can see an indepth comparion with the Repsonse variable (M1) and Regressors from the new model.

```r
library(tidyverse)



ggplot(prediction, aes(dat$IIP, dat$M1)) +
  geom_point() +
  stat_smooth(method = lm, se = FALSE) +
  geom_segment(aes(xend = dat$IIP, yend = .fitted), color = "red", size =
                0.3)+xlab("IIP")+ylab("M!")+ggtitle("Regression plot")+theme
(plot.title = element_text(hjust = 0.5))
```

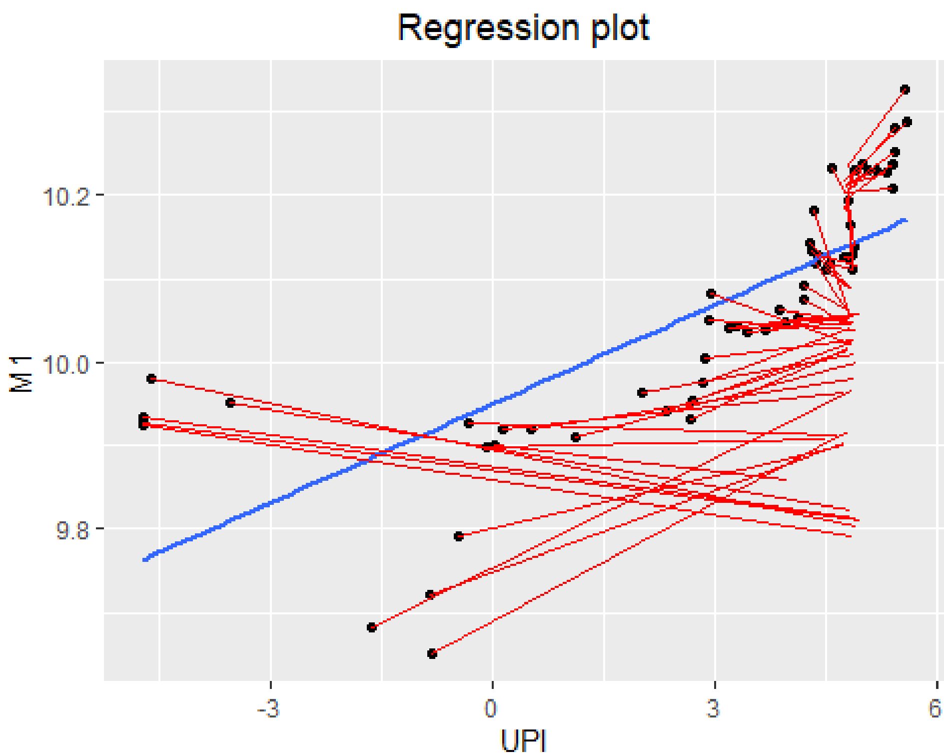
```
   `geom_smooth()` using formula 'y ~ x'
```




```
 We obsevre that with respect to IIP, the residual lines are faw away from th
e regression line itself.

ggplot(prediction, aes(dat$UPI, dat$M1)) +
  geom_point() +
  stat_smooth(method = lm, se = FALSE) +
  geom_segment(aes(xend = dat$IIP, yend = .fitted), color = "red", size =
                   0.3)+xlab("UPI")+ylab("M1")+ggtitle("Regression plot")+theme
(plot.title = element_text(hjust = 0.5))

   `geom_smooth()` using formula 'y ~ x'
```



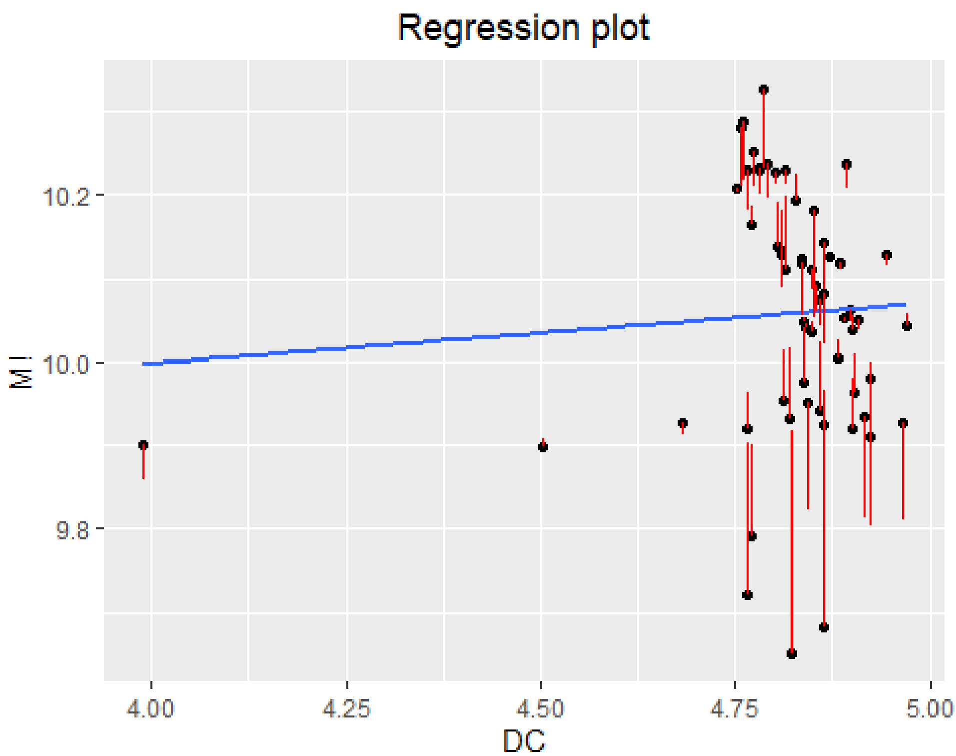
We observe that with respect to UPI, the residual lines are faw away from the regression line itself.

```
ggplot(prediction, aes(dat$CC, dat$M1)) +
  geom_point() +
  stat_smooth(method = lm, se = FALSE) +
  geom_segment(aes(xend = dat$IIP, yend = .fitted), color = "red", size =
                 0.3)+xlab("CC")+ylab("M!")+ggtitle("Regression plot")+theme(
plot.title = element_text(hjust = 0.5))

   `geom_smooth()` using formula 'y ~ x'
```


```
ggplot(prediction, aes(dat$IIP, dat$M1)) +
  geom_point() +
  stat_smooth(method = lm, se = FALSE) +
  geom_segment(aes(xend = dat$IIP, yend = .fitted), color = "red", size =
                 0.3)+xlab("DC")+ylab("M!")+ggtitle("Regression plot")+theme(
plot.title = element_text(hjust = 0.5))

   `geom_smooth()` using formula 'y ~ x'
```



Similarly for CC and DC the residuals are closer to the line as the values increase since the values are clustered amongst the higher values of the observation.

Conclusion:

We observe that we got a model that only explain 65% of the variaion. Although this isn't an ideal R-squared value, it still fulfills the assumptions of the multiple regression. A higher R-square need not mean that our model isn't fit but usually it is better to have more independent variables that truly explain the model. Our model shows that it is not the perfect fit but can be used to predict to some extent.

The drawbacks of the model also include the fact that there are some outliers in the dataset and the dataset isnt exaclty normal. This can be tackled by transforming the data into a perfectly normal fit. We might also require some more independent variables to explain the variation in the data.

Nevertheless, we see that the model's expected values are closed to the bserved values. Since, the values are smaller and are in decimals, the precision really matters in such cases.

The model proves that its useful in the domain of economics as it helps to predict how much currency will be flowing during a given period of time.

This is extremely important as the government must have enough for its citizens so that people can afford and recieve income acordingly. The model will help the government to know when to pump in more money and take away when it is needed.