MULYIPLE LINEAR REGRESSION MODEL

Mohammed Swaned M

2022-12-05

Contents

Name : Mohammed Swaned M

Enrollment_no: GN4994

Faculty_no : 22DSMSA104

1 Multiple Linear Regression Model

1.1 Introduction

The Multiple linear regression model is defined as $y = \beta_0 + \beta_1 \times X_1 + \cdots + \beta_p X_p + \epsilon$

Now we consider the wage12 data and the multiple linear regression model for that data is given as

$wage = \beta_0 + \beta_1 \times education + \beta_2 \times tenure + \beta_3 \times nonwhite + \beta_4 \times female + \beta_5 \times married + \epsilon$

Now read data

```
credit=read.csv("credit.csv")
head(credit)
```

```
##    Income Limit Rating Cards Age Education Own Student Married Region Balance
## 1  14.891  3606    283     2  34        11  No      No     Yes  South     333
## 2 106.025  6645    483     3  82        15 Yes     Yes     Yes   West     903
## 3 104.593  7075    514     4  71        11  No      No      No   West     580
## 4 148.924  9504    681     3  36        11 Yes      No      No   West     964
## 5  55.882  4897    357     2  68        16  No      No     Yes  South     331
## 6  80.180  8047    569     4  77        10  No      No      No  South    1151
```

```
credit1=credit[c(-7,-8,-9,-10)]
credit1
```

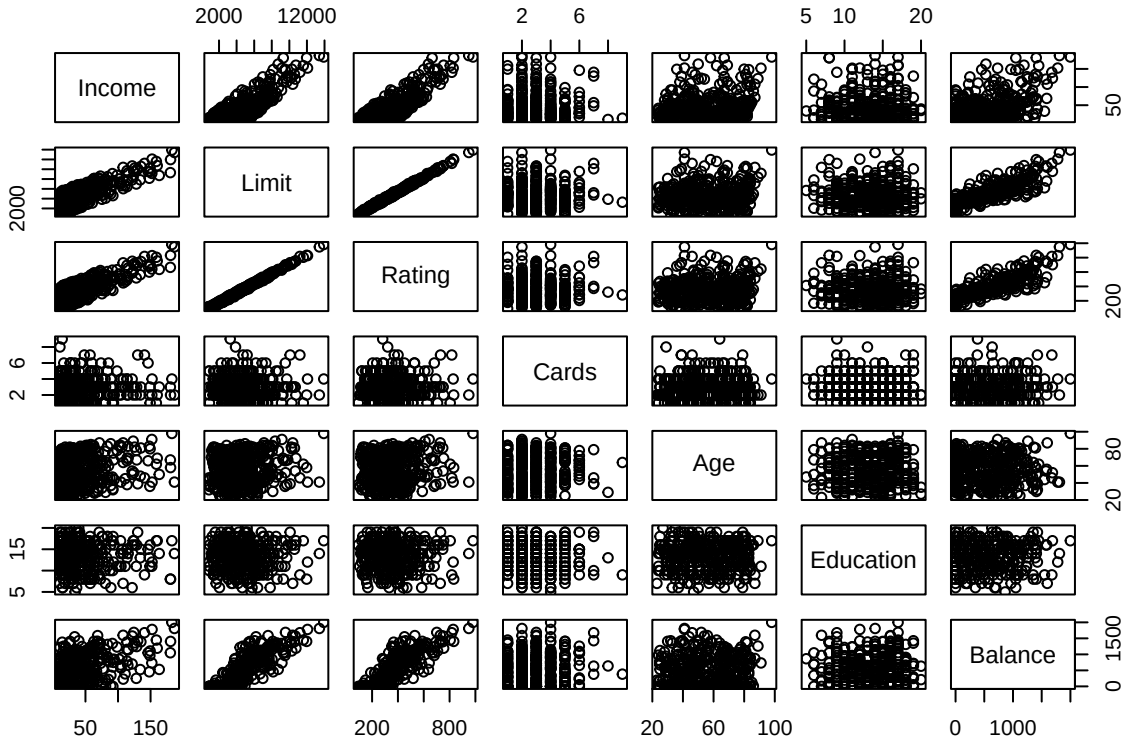
```
##       Income Limit Rating Cards Age Education Balance
## 1     14.891  3606    283     2  34        11     333
## 2    106.025  6645    483     3  82        15     903
## 3    104.593  7075    514     4  71        11     580
## 4    148.924  9504    681     3  36        11     964
## 5     55.882  4897    357     2  68        16     331
## 6     80.180  8047    569     4  77        10    1151
## 7     20.996  3388    259     2  37        12     203
## 8     71.408  7114    512     2  87         9     872
## 9     15.125  3300    266     5  66        13     279
## 10    71.061  6819    491     3  41        19    1350
## 11    63.095  8117    589     4  30        14    1407
## 12    15.045  1311    138     3  64        16       0
## 13    80.616  5308    394     1  57         7     204
## 14    43.682  6922    511     1  49         9    1081
## 15    19.144  3291    269     2  75        13     148
## 16    20.089  2525    200     3  57        15       0
## 17    53.598  3714    286     3  73        17       0
## 18    36.496  4378    339     3  69        15     368
## 19    49.570  6384    448     1  28         9     891
## 20    42.079  6626    479     2  44         9    1048
## 21    17.700  2860    235     4  63        16      89
## 22    37.348  6378    458     1  72        17     968
## 23    20.103  2631    213     3  61        10       0
## 24    64.027  5179    398     5  48         8     411
## 25    10.742  1757    156     3  57        15       0
## 26    14.090  4323    326     5  25        16     671
## 27    42.471  3625    289     6  44        12     654
## 28    32.793  4534    333     2  44        16     467
## 29   186.634 13414    949     2  41        14    1809
## 30    26.813  5611    411     4  55        16     915
## 31    34.142  5666    413     4  47         5     863
## 32    28.941  2733    210     5  43        16       0
## 33   134.181  7838    563     2  48        13     526
## 34    31.367  1829    162     4  30        10       0
## 35    20.150  2646    199     2  25        14       0
## 36    23.350  2558    220     3  49        12     419
## 37    62.413  6457    455     2  71        11     762
## 38    30.007  6481    462     2  69         9    1093
## 39    11.795  3899    300     4  25        10     531
## 40    13.647  3461    264     4  47        14     344
## 41    34.950  3327    253     3  54        14      50
## 42   113.659  7659    538     2  66        15    1155
## 43    44.158  4763    351     2  66        13     385
## 44    36.929  6257    445     1  24        14     976
## 45    31.861  6375    469     3  25        16    1120
## 46    77.380  7569    564     3  50        12     997
## 47    19.531  5043    376     2  64        16    1241
## 48    44.646  4431    320     2  49        15     797
## 49    44.522  2252    205     6  72        15       0
```

```
## 50   43.479  4569    354     4  49        13     902
## 51   36.362  5183    376     3  49        15     654
## 52   39.705  3969    301     2  27        20     211
## 53   44.205  5441    394     1  32        12     607
## 54   16.304  5466    413     4  66        10     957
## 55   15.333  1499    138     2  47         9       0
## 56   32.916  1786    154     2  60         8       0
## 57   57.100  4742    372     7  79        18     379
## 58   76.273  4779    367     4  65        14     133
## 59   10.354  3480    281     2  70        17     333
## 60   51.872  5294    390     4  81        17     531
## 61   35.510  5198    364     2  35        20     631
## 62   21.238  3089    254     3  59        10     108
## 63   30.682  1671    160     2  77         7       0
## 64   14.132  2998    251     4  75        17     133
## 65   32.164  2937    223     2  79        15       0
## 66   12.000  4160    320     4  28        14     602
## 67  113.829  9704    694     4  38        13    1388
## 68   11.187  5099    380     4  69        16     889
## 69   27.847  5619    418     2  78        15     822
## 70   49.502  6819    505     4  55        14    1084
## 71   24.889  3954    318     4  75        12     357
## 72   58.781  7402    538     2  81        12    1103
## 73   22.939  4923    355     1  47        18     663
## 74   23.989  4523    338     4  31        15     601
## 75   16.103  5390    418     4  45        10     945
## 76   33.017  3180    224     2  28        16      29
## 77   30.622  3293    251     1  68        16     532
## 78   20.936  3254    253     1  30        15     145
## 79  110.968  6662    468     3  45        11     391
## 80   15.354  2101    171     2  65        14       0
## 81   27.369  3449    288     3  40         9     162
## 82   53.480  4263    317     1  83        15      99
## 83   23.672  4433    344     3  63        11     503
## 84   19.225  1433    122     3  38        14       0
## 85   43.540  2906    232     4  69        11       0
## 86  152.298 12066    828     4  41        12    1779
## 87   55.367  6340    448     1  33        15     815
## 88   11.741  2271    182     4  59        12       0
## 89   15.560  4307    352     4  57         8     579
## 90   59.530  7518    543     3  52         9    1176
## 91   20.191  5767    431     4  42        16    1023
## 92   48.498  6040    456     3  47        16     812
## 93   30.733  2832    249     4  51        13       0
## 94   16.479  5435    388     2  26        16     937
## 95   38.009  3075    245     3  45        15       0
## 96   14.084   855    120     5  46        17       0
## 97   14.312  5382    367     1  59        17    1380
## 98   26.067  3388    266     4  74        17     155
## 99   36.295  2963    241     2  68        14     375
## 100  83.851  8494    607     5  47        18    1311
## 101  21.153  3736    256     1  41        11     298
## 102  17.976  2433    190     3  70        16     431
## 103  68.713  7582    531     2  56        16    1587
```

```
## 104 146.183  9540    682     6  66        15    1050
## 105  15.846  4768    365     4  53        12     745
## 106  12.031  3182    259     2  58        18     210
## 107  16.819  1337    115     2  74        15       0
## 108  39.110  3189    263     3  72        12       0
## 109 107.986  6033    449     4  64        14     227
## 110  13.561  3261    279     5  37        19     297
## 111  34.537  3271    250     3  57        17      47
## 112  28.575  2959    231     2  60        11       0
## 113  46.007  6637    491     4  42        14    1046
## 114  69.251  6386    474     4  30        12     768
## 115  16.482  3326    268     4  41        15     271
## 116  40.442  4828    369     5  81         8     510
## 117  35.177  2117    186     3  62        16       0
## 118  91.362  9113    626     1  47        17    1341
## 119  27.039  2161    173     3  40        17       0
## 120  23.012  1410    137     3  81        16       0
## 121  27.241  1402    128     2  67        15       0
## 122 148.080  8157    599     2  83        13     454
## 123  62.602  7056    481     1  84        11     904
## 124  11.808  1300    117     3  77        14       0
## 125  29.564  2529    192     1  30        12       0
## 126  27.578  2531    195     1  34        15       0
## 127  26.427  5533    433     5  50        15    1404
## 128  57.202  3411    259     3  72        11       0
## 129 123.299  8376    610     2  89        17    1259
## 130  18.145  3461    279     3  56        15     255
## 131  23.793  3821    281     4  56        12     868
## 132  10.726  1568    162     5  46        19       0
## 133  23.283  5443    407     4  49        13     912
## 134  21.455  5829    427     4  80        12    1018
## 135  34.664  5835    452     3  77        15     835
## 136  44.473  3500    257     3  81        16       8
## 137  54.663  4116    314     2  70         8      75
## 138  36.355  3613    278     4  35         9     187
## 139  21.374  2073    175     2  74        11       0
## 140 107.841 10384    728     3  87         7    1597
## 141  39.831  6045    459     3  32        12    1425
## 142  91.876  6754    483     2  33        10     605
##  [ reached 'max' / getOption("max.print") -- omitted 258 rows ]
```

1.2 Make the scatter plot

```
pairs(credit1)
```



From the above plot, we may see that there is a linear relationship between TV and sales, radio“andsales’ and newspaper and sales. To figure out more we obtain the correlation coefficient among the variables.

1.3 Calculate the correlation coefficient

```
cor(credit1)
```

```
##                Income       Limit      Rating       Cards         Age    Education      Balance
## Income     1.00000000  0.79208834  0.79137763 -0.01827261 0.175338403 -0.027691982  0.463656457
## Limit      0.79208834  1.00000000  0.99687974  0.01023133 0.100887922 -0.023548534  0.861697267
## Rating     0.79137763  0.99687974  1.00000000  0.05323903 0.103164996 -0.030135627  0.863625161
## Cards     -0.01827261  0.01023133  0.05323903  1.00000000 0.042948288 -0.051084217  0.086456347
## Age        0.17533840  0.10088792  0.10316500  0.04294829 1.000000000  0.003619285  0.001835119
## Education -0.02769198 -0.02354853 -0.03013563 -0.05108422 0.003619285  1.000000000 -0.008061576
## Balance    0.46365646  0.86169727  0.86362516  0.08645635 0.001835119 -0.008061576  1.000000000
```

1.4 Fitting of multiple Linear Regression Model

To estimate the coefficients of the variables TV, radio and newspaper, we fit the following model

```
M1=lm(Balance~., data = credit)
summary(M1)
```

```
##
## Call:
## lm(formula = Balance ~ ., data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -161.64  -77.70  -13.49   53.98  318.20
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -479.20787   35.77394 -13.395  < 2e-16 ***
## Income        -7.80310    0.23423 -33.314  < 2e-16 ***
## Limit          0.19091    0.03278   5.824 1.21e-08 ***
## Rating         1.13653    0.49089   2.315   0.0211 *
## Cards         17.72448    4.34103   4.083 5.40e-05 ***
## Age           -0.61391    0.29399  -2.088   0.0374 *
## Education     -1.09886    1.59795  -0.688   0.4921
## OwnYes       -10.65325    9.91400  -1.075   0.2832
## StudentYes   425.74736   16.72258  25.459  < 2e-16 ***
## MarriedYes    -8.53390   10.36287  -0.824   0.4107
## RegionSouth   10.10703   12.20992   0.828   0.4083
## RegionWest    16.80418   14.11906   1.190   0.2347
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.79 on 388 degrees of freedom
## Multiple R-squared:  0.9551, Adjusted R-squared:  0.9538
## F-statistic: 750.3 on 11 and 388 DF,  p-value: < 2.2e-16
```

1.5 Checking Significance of Model

To check the significance of the model, we check F statistic and for that we set the hypotheses as follows:

Null Hypothesis:$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

Alternative Hypothesis:$H_1 :$ *Atleast one* $\beta_i \neq 0, i = 1, 2, \cdots, k.$

F-statistic: 750.3 on 11 and 388 DF, p-value: $< 2.2e-16$ Since the F statistic is 750.3 on 11 and 388 df with p-value $<$ 2.2e-16 i.e. almost zero. Hence we *reject the null hypothesis* that means there is at least oneβ_i that is not equal to zero. Therefore, our model is significant.

1.6 Checking Significance of Variables

To check the significance of variable(s), we check the $t - ratios$ and its corresponding $p - values$. We set the hypotheses as follows: $H_0 : \beta_i = 0\ Vs\ H_1 : \beta_i \neq 0\, where\, i = 0, 1, 2, 3$ Now we check the t-statistics and p value of the corresponding variables one by one and decide that which one is significant. SO the p-value for *intercept* term is 2×10^{-16} that is almost zero, hence we reject the null hypothesis and accept that $\beta_0 \neq 0$. The p-value of *Income* is 2×10^{-16}that is almost zero, hence we reject the null hypothesis and accept that $\beta_1 \neq 0$ The p-value of *Limit* is 1.21×10^{-8} that is almost zero, hence we reject the null hypothesis and accept that $\beta_3 \neq 0$. The p-value of *Rating* is 0.0211 that is less than 0.05, hence we reject the null hypothesis and accept that $\beta_5 \neq 0$. The p-value of *Cards* is 5.40×10^{-5} that is almost zero, hence we reject the null hypothesis and accept that $\beta_5 \neq 0$. The p-value of *Age* is 0.034 that is less than 0.05, hence we reject the null hypothesis and accept that $\beta_5 \neq 0$. The p-value of *Student* is 2×10^{-16} that is almost zero, hence we reject the null hypothesis and accept that $\beta_5 \neq 0$. The p-value of *Education* is 0.4921 that is greater than

0.05, hence we fail to reject the null hypothesis and accept that $\beta_3 = 0$. The p-value of *Own* is 0.2832 that is greater than 0.05, hence we fail to reject the null hypothesis and accept that $\beta_2 = 0$. The p-value of *Married* is 0.4107 that is greater than 0.05, hence we fail to reject the null hypothesis and accept that $\beta_6 = 0$. The p-value of *Region* is 0.4083 that is greater than 0.05, hence we fail to reject the null hypothesis and accept that $\beta_6 = 0$. This means that the variables *Education,Own,Married and Region*is not significant in this model. Since these variables is not significant, so we remove these variable from the model. Hence our final model is: $Balance = \beta_0 + \beta_1 \times Income + \beta_2 \times Limit + \beta_3 \times Ratings + \beta_4 \times Cards + \beta_5 \times Age + \beta_8 \times Student + \epsilon$

```
M2=lm(Balance~Income+Limit+Rating+Cards+Age+Student,data=credit)
summary(M2)
```

```
##
## Call:
## lm(formula = Balance ~ Income + Limit + Rating + Cards + Age +
##     Student, data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -170.00  -77.85  -11.84   56.87  313.52
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -493.73419   24.82476 -19.889  < 2e-16 ***
## Income        -7.79508    0.23342 -33.395  < 2e-16 ***
## Limit          0.19369    0.03238   5.981 4.98e-09 ***
## Rating         1.09119    0.48480   2.251   0.0250 *
## Cards         18.21190    4.31865   4.217 3.08e-05 ***
## Age           -0.62406    0.29182  -2.139   0.0331 *
## StudentYes   425.60994   16.50956  25.780  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.61 on 393 degrees of freedom
## Multiple R-squared:  0.9547, Adjusted R-squared:  0.954
## F-statistic:  1380 on 6 and 393 DF,  p-value: < 2.2e-16
```

1.7 Adequacy of Model

1.7.1 Residual Standerd Error

$RSE = \sqrt{(\frac{RSS}{n-p-1})}$

where $RSS = \sqrt{(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ and is called residual sum of squares(RSS)

```
Balance=credit$Balance
Income=credit$Income
Limit=credit$Limit
Rating=credit$Rating
Cards=credit$Cards
Age=credit$Age
Balancehat=M2$coefficients[1]*Income+M2$coefficients[2]*Limit+M2$coefficients[3]*Rating+M2$coefficients
#alternaively
resid=M2$residuals
```

```
rss=sum(resid^2)
rss
```

```
## [1] 3821620
```

```
n=length(credit$Income)
p=5
rse=sqrt(rss/n-p-1)
rse
```

```
## [1] 97.71412
```

Now Calculate RSE from the rectified model

```
resid2=M2$residuals
rss2=sum(resid2^2)
rss2
```

```
## [1] 3821620
```

```
p=5
rse2=sqrt(rss/n-p-1)
rse2
```

```
## [1] 97.71412
```

1.7.2 R Squared

Multiple R-squared: 0.9551, Adjusted R-squared: 0.9538. The reported R squared is 0.9551 that is approximately 0.95. So we see that 95% variability of Balance is explained by Income, Limit, Rating, Cards, Age,Education,Own,Student,Married and Region and the adjusted R- squared is 0.9538 when insignificant variables(Education,Own,Married and Region) is attached. After omitting these insignificant variables from the model and we examine the R-squared and adjusted R squared. They are as follows: Multiple R-squared: 0.9547, Adjusted R-squared: 0.954.We find that there is no difference in R squared but adjusted R-squared has increased a little bit. That is theevidence that if we remove any insignificant variable from the model, them adjusted R-squared increased.

1.8 Dummy Variable

1.8.1 Studnt as dummy variable

Here we model the data as follows: $Balance = \beta_0 + \beta_1 \times Student$

```
M3=lm(Balance~Student,data = credit)
summary(M3)
```

```
##
## Call:
## lm(formula = Balance ~ Student, data = credit)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -876.82 -458.82  -40.87  341.88 1518.63
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   480.37      23.43   20.50  < 2e-16 ***
## StudentYes    396.46      74.10    5.35 1.49e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 444.6 on 398 degrees of freedom
## Multiple R-squared:  0.06709,    Adjusted R-squared:  0.06475
## F-statistic: 28.62 on 1 and 398 DF,  p-value: 1.488e-07
```

Table displays the coefficient estimates and other information associated with the model. So the Model of the data as follows: $Balance = 480.37 + 396.46 \times Student$

However, we notice that the p-value for the dummy variable that is 1.488×10^{-7}. This means that the variable *Student* is significant.

1.8.2 Married as dummy variable

Here we model the data as follows: $Balance = \beta_0 + \beta_1 \times Married$

```
M4=lm(Balance~Married,data = credit)
summary(M4)
```

```
##
## Call:
## lm(formula = Balance ~ Married, data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -523.29 -451.03  -60.12  345.06 1481.06
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  523.290     36.974  14.153   <2e-16 ***
## MarriedYes    -5.347     47.244  -0.113     0.91
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 460.3 on 398 degrees of freedom
## Multiple R-squared:  3.219e-05,  Adjusted R-squared:  -0.00248
## F-statistic: 0.01281 on 1 and 398 DF,  p-value: 0.9099
```

Table displays the coefficient estimates and other information associated with the model. So the Model of the data as follows: $Balance = 523.290 - 5.347 \times Married$ However, we notice that the p-value for the dummy variable is high that is 0.9099 This means that the variable *Married* is insignificant. ### Own as dummy variable Here we model the data as follows: $Balance = \beta_0 + \beta_1 \times Own$

```
M5=lm(Balance~Own,data = credit)
summary(M5)
```

```
##
## Call:
## lm(formula = Balance ~ Own, data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -529.54 -455.35  -60.17  334.71 1489.20
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   509.80      33.13  15.389   <2e-16 ***
## OwnYes         19.73      46.05   0.429    0.669
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 460.2 on 398 degrees of freedom
## Multiple R-squared:  0.0004611,  Adjusted R-squared:  -0.00205
## F-statistic: 0.1836 on 1 and 398 DF,  p-value: 0.6685
```

Table displays the coefficient estimates and other information associated with the model. So the Model of the data as follows: $Balance = 509.80 + 19.73 \times Own$ However, we notice that the p-value for the dummy variable is high that is0.6685 This means that the variable Own is insignificant.

1.8.3 Region as dummy variable

Here we model the data as follows: $Balance = \beta_0 + \beta_1 \times Region$

```
M6=lm(Balance~Region,data = credit)
summary(M6)
```

```
##
## Call:
## lm(formula = Balance ~ Region, data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -531.00 -457.08  -63.25  339.25 1480.50
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   531.00      46.32  11.464   <2e-16 ***
## RegionSouth   -12.50      56.68  -0.221    0.826
## RegionWest    -18.69      65.02  -0.287    0.774
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 460.9 on 397 degrees of freedom
## Multiple R-squared:  0.0002188,  Adjusted R-squared:  -0.004818
## F-statistic: 0.04344 on 2 and 397 DF,  p-value: 0.9575
```

Table displays the coefficient estimates and other information associated with the model. So the Model of the data as follows: $Balance = 531 - 12.50 \times RegionSouth - 18.69 \times RegionWest$ However, we notice that the p-value for the dummy variable is high that is0.9575 This means that the variable *Region* is insignificant.