REGRESSION ASSINGMENT LAB 2

ABDUL RAUF

2022-12-03

Contents

Contents

NAME:-*ABDUL RAUF*
ENRL NO:-*GL6092*
ROLL NO:-*22DSMSA116*

LAB 2

1 Assingment 1

1.1 Install `wooldridge` package and use the data.

We have to install the `wooldridge` package by `install.packages("wooldridge")` then use the package by `require("wooldridge")/library("wooldridge")`

```
head(wage1)
```

```
##   wage educ exper tenure nonwhite female married numdep smsa northcen south west construc
## 1 3.10   11     2      0        0      1       0      2    1        0     0    1        0
## 2 3.24   12    22      2        0      1       1      3    1        0     0    1        0
## 3 3.00   11     2      0        0      0       0      2    0        0     0    1        0
## 4 6.00    8    44     28        0      0       1      0    1        0     0    1        0
## 5 5.30   12     7      2        0      0       1      1    0        0     0    1        0
## 6 8.75   16     9      8        0      0       1      0    1        0     0    1        0
##   ndurman trcommpu trade services profserv profocc clerocc servocc    lwage expersq tenursq
## 1       0        0     0        0        0       0       0       0 1.131402       4       0
## 2       0        0     0        1        0       0       0       1 1.175573     484       4
## 3       0        0     1        0        0       0       0       0 1.098612       4       0
## 4       0        0     0        0        0       0       1       0 1.791759    1936     784
## 5       0        0     0        0        0       0       0       0 1.667707      49       4
## 6       0        0     0        0        1       1       0       0 2.169054      81      64
```

```
wage12<-wage1[1:7]
head(wage12)
```

```
##   wage educ exper tenure nonwhite female married
## 1 3.10   11     2      0        0      1       0
## 2 3.24   12    22      2        0      1       1
## 3 3.00   11     2      0        0      0       0
## 4 6.00    8    44     28        0      0       1
## 5 5.30   12     7      2        0      0       1
## 6 8.75   16     9      8        0      0       1
```

We have extracted first seven variables from **wage1** and then stored the data in **wage12**.

1.2 Setting hypothesis

Null Hypothesis: \

$H_0 : \beta_1 = \beta_2 = \beta_3 \quad \cdot\cdot \quad = \beta_n = 0$

VS

Alternative hypothesis: $\backslash\ H_1 : At\ least\ one\ of\ the\ \beta_i\ 's\ are\ not\ zero$

1.2.1 fit the model.

$wage = \beta_0 + \beta_1 \times educ + \beta_2 \times exper + \beta_3 \times tenure + \beta_4 \times nonwhite + \beta_5 \times female + \beta_6 \times married + \epsilon_0$

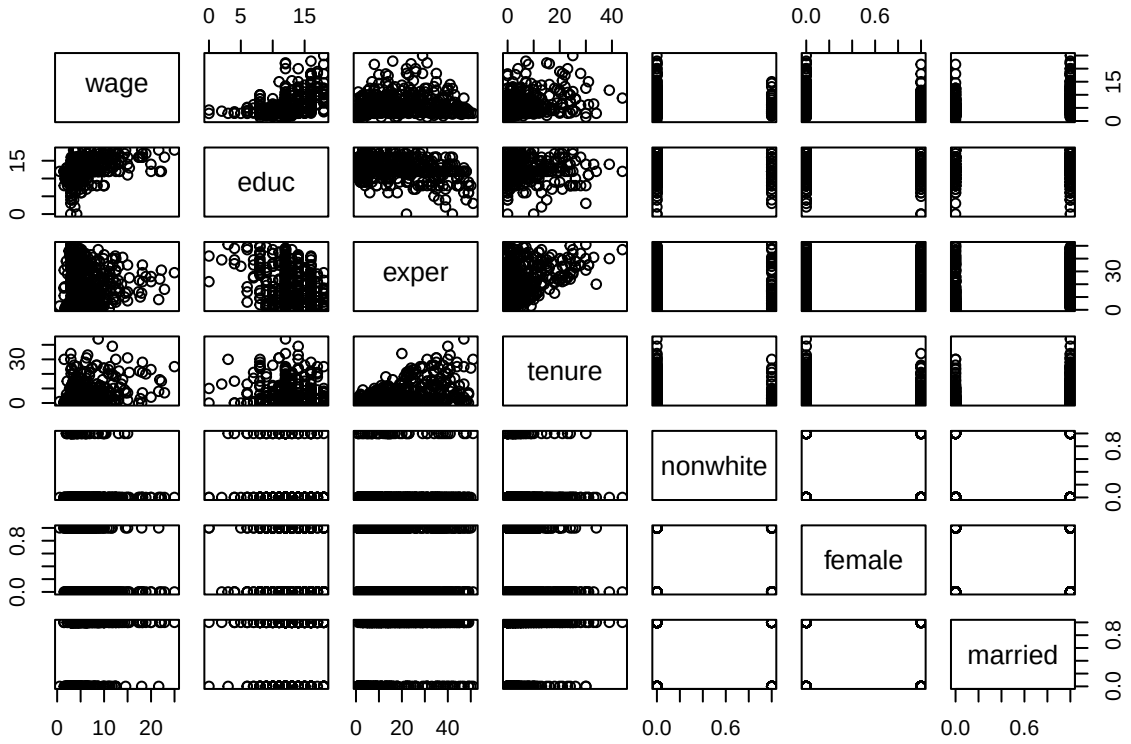
```
m1<-lm(wage~educ+exper+tenure+nonwhite+female+married , data = wage12)
summary(m1)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper + tenure + nonwhite + female +
##     married, data = wage12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6716 -1.8239 -0.4967  1.0403 13.9209
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.60221    0.73107  -2.192   0.0289 *
## educ         0.55510    0.05006  11.090  < 2e-16 ***
## exper        0.01875    0.01204   1.557   0.1201
## tenure       0.13883    0.02116   6.562 1.29e-10 ***
## nonwhite    -0.06581    0.42657  -0.154   0.8775
## female      -1.74241    0.26682  -6.530 1.57e-10 ***
## married      0.55657    0.28674   1.941   0.0528 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.952 on 519 degrees of freedom
## Multiple R-squared:  0.3682, Adjusted R-squared:  0.3609
## F-statistic: 50.41 on 6 and 519 DF,  p-value: < 2.2e-16
```

- **36.09%** of the variability is explained by this model and this model is significant for 6 & 519 degree of freedom and 5% level of significance but `exper` ,`nonwhite`& `married` are insignificant to this model.

1.3 Plotting the graph

```
pairs(wage12)
```



From the plot we can see that there is some relationship of wage with `educ`,`expr`,`tenure`.

1.4 Names of the variables

```
names(wage12)
```

```
## [1] "wage"     "educ"     "exper"    "tenure"   "nonwhite" "female"   "married"
```

1.5 Correlation matrix for `wage12` data

```
cor(wage12)
```

```
##                wage        educ       exper      tenure    nonwhite      female     married
## wage      1.00000000  0.40590333  0.11290344  0.34688957 -0.03851959 -0.34009786  0.22881718
## educ      0.40590333  1.00000000 -0.29954184 -0.05617257 -0.08465433 -0.08502941  0.06888104
## exper     0.11290344 -0.29954184  1.00000000  0.49929145  0.01435563 -0.04162597  0.31698428
## tenure    0.34688957 -0.05617257  0.49929145  1.00000000  0.01158880 -0.19791027  0.23988874
## nonwhite -0.03851959 -0.08465433  0.01435563  0.01158880  1.00000000 -0.01091747 -0.06225929
## female   -0.34009786 -0.08502941 -0.04162597 -0.19791027 -0.01091747  1.00000000 -0.16612843
## married   0.22881718  0.06888104  0.31698428  0.23988874 -0.06225929 -0.16612843  1.00000000
```

there is no such high correlation among the variables.we can also check from the VIF

1.6 VIF for the given model

```
vif(m1)
```

```
##     educ    exper   tenure nonwhite   female  married
## 1.157103 1.608380 1.407182 1.011547 1.072152 1.182157
```

all the VIF values are not so high which implies that there is no multicolliniarity.

1.7 Female as a dummy variable.

$wage = \beta_0 + \beta 1 \times female + \epsilon_0$

as it is a binary variable so for female model will be

$wage = \beta_0 + \beta_1 \times 1$

for male

$wage = \beta_0 + \beta_1 \times 0$
$i.e$
$wage = \beta_0$

```
m2<-lm(wage~female,data = wage12)
summary(m2)
```

```
##
## Call:
## lm(formula = wage ~ female, data = wage12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5995 -1.8495 -0.9877  1.4260 17.8805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.0995     0.2100  33.806  < 2e-16 ***
## female       -2.5118     0.3034  -8.279 1.04e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.476 on 524 degrees of freedom
## Multiple R-squared:  0.1157, Adjusted R-squared:  0.114
## F-statistic: 68.54 on 1 and 524 DF,  p-value: 1.042e-15
```

```
#female estimated wage =7.009-2.5118=4.4972
```

- As calculated F-statistic is greater than the tabulated F-statistic at 5% level of significance ,1 & 524 degrees of freedom i.e 3.9201 so our model is significant.

- The p-value for `female` is less than 0.05 so `female` is significant to the model.
- From above calculation table we come to know that the `wage` of the `female` is estimated to be **4.4972** and the `wage` of the `male` is estimated to be **7.0995**.`wage` of the `male` is higher than that of the `female`.
- **11.4%** of the variability of `wage` is explained by the `female`.

1.8 Nonwhite as a dummy variable

$wage = \beta_0 + \beta 1 \times nonwhite + \epsilon_0$

as it is a binary variable so for nonwhite model will be

$wage = \beta_0 + \beta_1 \times 1$

for white

$wage = \beta_0 + \beta_1 \times 0$
$i.e$
$wage = \beta_0$

```
m6<-lm(wage~nonwhite,data = wage12)
summary(m6)
```

```
##
## Call:
## lm(formula = wage ~ nonwhite, data = wage12)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.414 -2.526 -1.259  1.026 19.036
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.9442     0.1700  34.961   <2e-16 ***
## nonwhite     -0.4682     0.5306  -0.882    0.378
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.694 on 524 degrees of freedom
## Multiple R-squared:  0.001484,   Adjusted R-squared:  -0.0004218
## F-statistic: 0.7786 on 1 and 524 DF,  p-value: 0.378
```

```
#estimated nonwhite wage =5.9442-0.4682=5.476
```

- As calculated F-statistic is greater than the tabulated F-statistic at 5% level of significance, 1 & 524 degrees of freedom i.e 3.9201 so our model is insignificant.
- The p-value of nonwhite is greater than 0.05 so nonwhite is insignificant to the model.

- From above calculation table we come to know that the wage of the nonwhite is estimated to be **5.476** and the wage of the white is estimated to be **5.9442** but as it is insignificant so we have to remove nonwhite from our model.wage of the white is higher than that of the nonwhite.
- Negative R squared implies insignificance of explanatory variables i.e `nonwhite` variable.

1.9 Married as a dummy variable

$wage = \beta_0 + \beta 1 \times married + \epsilon_0$

as it is a binary variable so for married model will be

$wage = \beta_0 + \beta_1 \times 1$

for non-married

$wage = \beta_0 + \beta_1 \times 0$
$i.e$
$wage = \beta_0$

```
m7<-lm(wage~married,data = wage12)
summary(m7)
```

```
##
## Call:
## lm(formula = wage ~ married, data = wage12)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.144 -2.181 -1.094  1.406 18.407
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.8439     0.2507  19.320  < 2e-16 ***
## married       1.7296     0.3214   5.381 1.12e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.599 on 524 degrees of freedom
## Multiple R-squared:  0.05236,    Adjusted R-squared:  0.05055
## F-statistic: 28.95 on 1 and 524 DF,  p-value: 1.121e-07
```

```
#total estimated married wage = 4.8439+1.7296=6.5735
```

- As calculated F-statistic is greater than the tabulated F-statistic at 5% level of significance ,1 & 524 degrees of freedom i.e 3.9201 so our model is significant.
- The p-value of married is less than 0.05 so married is significant to the model.
- From above calculation table we come to know that the wage of the married is estimated to be **6.5735** and the wage of the non-married is estimated to be **4.8439**.wage of the non-married is higher than that of the married.
- **5%** of the variability of wage is explained by the married.

1.10 forward selection

1.10.1 1. we add exper to the model

$wage = \beta_0 + \beta_1 \times exper + \epsilon_0$

```
f1<-lm(wage~exper,data = wage12)
summary(f1)
```

```
##
## Call:
## lm(formula = wage ~ exper, data = wage12)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.936 -2.458 -1.112  1.077 18.716
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.37331    0.25699  20.908  < 2e-16 ***
## exper        0.03072    0.01181   2.601  0.00955 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.673 on 524 degrees of freedom
## Multiple R-squared:  0.01275,    Adjusted R-squared:  0.01086
## F-statistic: 6.766 on 1 and 524 DF,  p-value: 0.009555
```

- it is significant to the model as it's p-value is greater than 0.05
- adjusted R-squared is **0.01086** and RSE is **3.673**

1.10.2 2. we add educ in 1

$wage = \beta_0 + \beta_1 \times exper + \beta_2 \times educ + \epsilon_0$

```
f2<-lm(wage~exper+educ,data = wage12)
summary(f2)
```

```
##
## Call:
## lm(formula = wage ~ exper + educ, data = wage12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5532 -1.9801 -0.7071  1.2030 15.8370
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.39054    0.76657  -4.423 1.18e-05 ***
## exper        0.07010    0.01098   6.385 3.78e-10 ***
## educ         0.64427    0.05381  11.974  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.257 on 523 degrees of freedom
## Multiple R-squared:  0.2252, Adjusted R-squared:  0.2222
## F-statistic: 75.99 on 2 and 523 DF,  p-value: < 2.2e-16
```

- adjusted r squared get increased so it is significant as it becomes **0.2222** and RSE get reduced **3.257**

1.10.3 3. we add `tenure` in 2

$wage = \beta_0 + \beta_1 \times exper + \beta_2 \times educ + \beta_3 \times tenure + \epsilon_0$

```
f3<-lm(wage~exper+educ+tenure,data = wage12)
summary(f3)
```

```
##
## Call:
## lm(formula = wage ~ exper + educ + tenure, data = wage12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6068 -1.7747 -0.6279  1.1969 14.6536
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.87273    0.72896  -3.941 9.22e-05 ***
## exper        0.02234    0.01206   1.853   0.0645 .
## educ         0.59897    0.05128  11.679  < 2e-16 ***
## tenure       0.16927    0.02164   7.820 2.93e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.084 on 522 degrees of freedom
## Multiple R-squared:  0.3064, Adjusted R-squared:  0.3024
## F-statistic: 76.87 on 3 and 522 DF,  p-value: < 2.2e-16
```

- this model is significant as p-value is less than 0.05 but by the use of the `tenure` in the model `exper` get insignificant so we removed the tenure from our model.

1.10.4 4. we add `nonwhite` to our model

$wage = \beta_0 + \beta_1 \times exper + \beta_2 \times educ + \beta_4 \times nonwhite + \epsilon_0$

```
f4<-lm(wage~exper+educ+nonwhite,data = wage12)
summary(f4)
```

```
##
## Call:
## lm(formula = wage ~ exper + educ + nonwhite, data = wage12)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -5.538 -1.982 -0.709  1.205 15.835
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.38683    0.77481  -4.371 1.49e-05 ***
## exper        0.07009    0.01099   6.378 3.95e-10 ***
## educ         0.64412    0.05405  11.917  < 2e-16 ***
## nonwhite    -0.01621    0.47006  -0.034    0.972
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.26 on 522 degrees of freedom
## Multiple R-squared:  0.2252, Adjusted R-squared:  0.2207
## F-statistic: 50.56 on 3 and 522 DF,  p-value: < 2.2e-16
```

- From the above table we find that model is significant as p-value is less than 0.05 .
- But p-value of the `nonwhite` is greater than 0.05 so we remove `nonwhite` from our model.

1.10.5 5. we add`married` to our model

$wage = \beta_0 + \beta_1 \times exper + \beta_2 \times educ + \beta_5 \times married + \epsilon_0$

```
f5<-lm(wage~exper+educ+married,data = wage12)
summary(f5)
```

```
##
## Call:
## lm(formula = wage ~ exper + educ + married, data = wage12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7049 -2.0168 -0.5597  1.2077 15.5241
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.37293    0.75990  -4.439 1.11e-05 ***
## exper        0.05688    0.01164   4.888 1.36e-06 ***
## educ         0.61285    0.05423  11.300  < 2e-16 ***
## married      0.98945    0.30920   3.200  0.00146 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.229 on 522 degrees of freedom
## Multiple R-squared:  0.2401, Adjusted R-squared:  0.2357
## F-statistic: 54.97 on 3 and 522 DF,  p-value: < 2.2e-16
```

*From the above table we find that model is significant as p-value is less than 0.05 .

*p-value of the `married` is less than 0.05 so it is significant to our model.

*our adjusted R-squared get increased as it becomes **0.2357** so this model is better than previous model in 2.

1.10.6 6. we add `female` to our model

$wage = \beta_0 + \beta_1 \times exper + \beta_2 \times educ + \beta_5 \times married + \beta_6 \times female + \epsilon_0$

```
f6<-lm(wage~exper+educ+married+female,data = wage12)
summary(f6)
```

```
##
## Call:
## lm(formula = wage ~ exper + educ + married + female, data = wage12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4057 -1.9042 -0.5982  1.1454 14.6545
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.79066    0.75121  -2.384   0.0175 *
## exper        0.05567    0.01106   5.035 6.59e-07 ***
## educ         0.58332    0.05166  11.292  < 2e-16 ***
## married      0.66024    0.29685   2.224   0.0266 *
## female      -2.06710    0.27221  -7.594 1.45e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.066 on 521 degrees of freedom
## Multiple R-squared:  0.3158, Adjusted R-squared:  0.3105
## F-statistic: 60.12 on 4 and 521 DF,  p-value: < 2.2e-16
```

- From the above table we find that model is significant as p-value is less than 0.05 .
- p-value of the `female` is less than 0.05 so it is significant to our model.
- our adjusted R-squared get increased so this model is better than previous model in 5 i.e **0.3105**.
- RSE is also less than that of model in 5 i.e **3.066** so it is better model.

2 Assingment 2

2.1 Read the data.

```
credit<-read.csv("credit.csv")
head(credit)
```

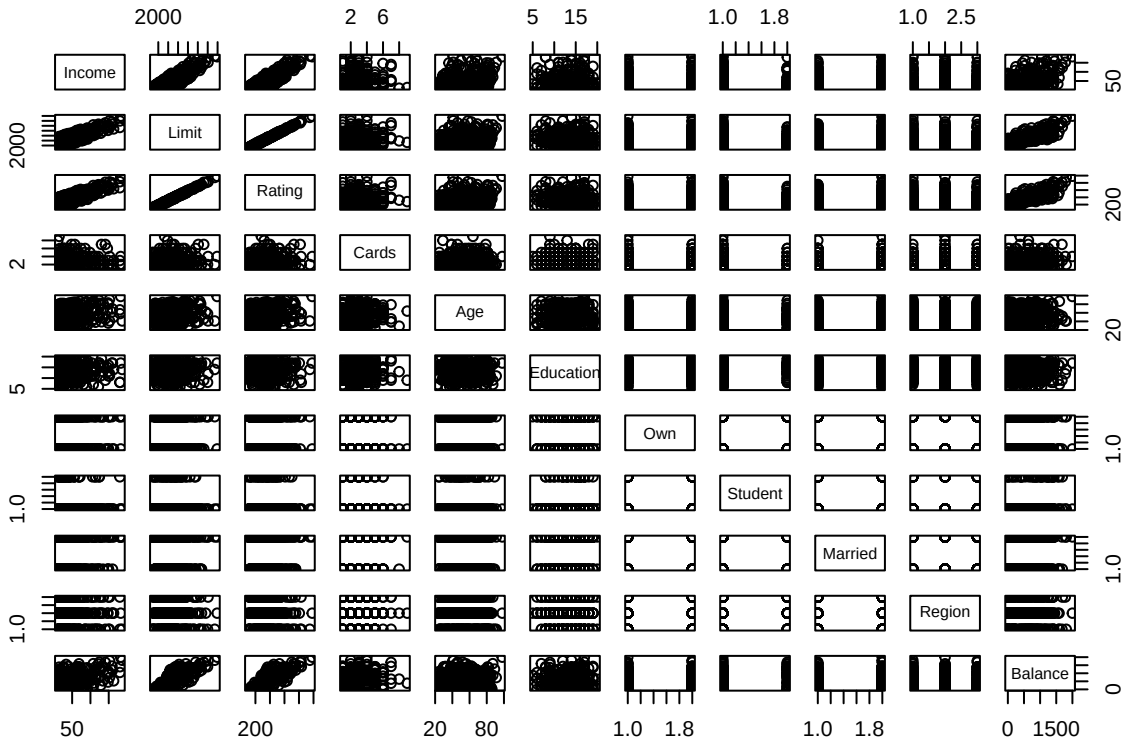
```
##    Income Limit Rating Cards Age Education Own Student Married Region Balance
## 1  14.891  3606    283     2  34        11  No      No     Yes  South     333
## 2 106.025  6645    483     3  82        15 Yes     Yes     Yes   West     903
## 3 104.593  7075    514     4  71        11  No      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
```

```
own<-factor(credit$Own,labels=c(0,1)) #transforming the (yes,no) values in (0 & 1) i.e integer
student<-factor(credit$Student,labels=c(0,1)) #transforming the (yes,no) values in (0 & 1) i.e integer
married<-factor(credit$Married,labels=c(0,1)) #transforming the (yes,no) values in (0 & 1) i.e integer
own
```

```
##   [1] 0 1 0 1 0 0 1 0 1 1 0 0 1 0 1 1 1 1 1 0 1 1 0 0 1 1 1 0 1 1 1 0 1 0 1 1 1 1 1 0 1 0 1 1 1
##  [46] 1 1 0 0 0 0 0 0 0 0 1 1 1 1 0 1 1 1 1 0 1 1 1 1 1 0 0 1 1 0 1 0 0 1 1 0 1 0 0 1 0 1 0 1 0 1
##  [91] 0 0 0 0 1 1 0 1 1 0 0 1 0 0 1 1 0 0 0 0 1 1 0 1 0 1 1 0 1 0 1 0 1 1 1 1 1 1 0 0 1 0 0 1 1
## [136] 1 1 0 1 0 1 0 0 1 0 0 1 0 1 0 1 1 1 0 0 1 1 1 1 0 1 0 0 1 0 0 0 1 1 1 0 1 0 0 0 0 0 1 1 1
## [181] 0 0 1 1 1 1 0 0 0 1 1 1 0 1 0 1 0 1 1 0 0 1 1 1 1 0 0 1 0 1 1 0 0 0 1 1 0 0 1 1 1 1 0 0 0
## [226] 1 1 1 1 0 0 0 0 0 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 1 1 0 1 0 1 0 1 0 0 1 0 1 1 1 0 1 1 1 0 1
## [271] 0 1 1 0 1 0 1 0 1 1 1 1 1 1 0 1 1 0 1 0 0 1 0 0 0 1 1 0 1 1 0 1 1 0 1 1 0 1 0 1 0 1 0 1 0
## [316] 0 0 0 1 0 0 1 0 0 0 1 0 1 1 1 0 0 0 0 1 1 0 0 0 0 1 1 0 1 0 0 0 0 0 1 1 0 0 0 1 1 0 0 0 0
## [361] 1 1 1 0 0 0 0 0 1 0 0 0 1 1 1 1 1 0 0 0 1 0 0 0 0 1 0 0 1 0 1 1 0 0 0 0 0 1 0 1
## Levels: 0 1
```

2.2 Plot the graph

```
plot(credit)
```



* From the above plot we found that balance has a linear relationship with income,limit,rating and age.

2.3 correlation matrix for credit data

```
cor(credit[,-7:-10])
```

```
##                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
```

- There is high correlation among the variables i.e income & limit,income & rating ,limit & rating.
- As balance is a response variable so there must be a correlation of the variables with them

2.4 Setting hypothesis

Null Hypothesis: \

$H_0 : \beta_1 = \beta_2 = \beta_3 \quad \cdots \quad = \beta_n = 0$

VS

Alternative hypothesis: \ $H_1 : At\ least\ one\ of\ the\ \beta_i\ 's\ are\ not\ zero$

2.4.1 fit the model.

$Balance = \beta_0 + \beta_1 \times income + \beta_2 \times limit + \beta_3 \times rating + \beta_4 \times cards + \beta_5 \times age + \beta_6 \times education + \beta_7 \times balance + \epsilon_0$

```
a2<-lm(Balance~Income+Limit+Rating+Cards+Age+Education+own+student+married, data = credit)
summary(a2)
```

```
##
## Call:
## lm(formula = Balance ~ Income + Limit + Rating + Cards + Age +
##     Education + own + student + married, data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -171.66  -75.32  -11.29   54.42  309.98
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -468.40374   34.35512 -13.634  < 2e-16 ***
## Income        -7.80200    0.23395 -33.349  < 2e-16 ***
## Limit          0.19308    0.03268   5.909 7.52e-09 ***
## Rating         1.10227    0.48923   2.253   0.0248 *
## Cards         17.92327    4.33228   4.137 4.31e-05 ***
## Age           -0.63468    0.29325  -2.164   0.0310 *
```

```
## Education     -1.11503    1.59592  -0.699   0.4852
## own1         -10.40665    9.90410  -1.051   0.2940
## student1     426.46919   16.67770  25.571  < 2e-16 ***
## married1      -7.01910   10.27803  -0.683   0.4951
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.72 on 390 degrees of freedom
## Multiple R-squared:  0.9549, Adjusted R-squared:  0.9539
## F-statistic: 918.2 on 9 and 390 DF,  p-value: < 2.2e-16
```

- From the above calculated table we came to know that our model is significant as our p-value is less than 0.05
- own and married are insignificant as their p-value is greater than 0.05 except them all the variables are significant.

```
a3<-lm(Balance~Income+Limit+Rating+Cards+Age+student, data = credit)
summary(a2)
```

```
##
## Call:
## lm(formula = Balance ~ Income + Limit + Rating + Cards + Age +
##     Education + own + student + married, data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -171.66  -75.32  -11.29   54.42  309.98
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -468.40374   34.35512 -13.634  < 2e-16 ***
## Income        -7.80200    0.23395 -33.349  < 2e-16 ***
## Limit          0.19308    0.03268   5.909 7.52e-09 ***
## Rating         1.10227    0.48923   2.253   0.0248 *
## Cards         17.92327    4.33228   4.137 4.31e-05 ***
## Age           -0.63468    0.29325  -2.164   0.0310 *
## Education     -1.11503    1.59592  -0.699   0.4852
## own1         -10.40665    9.90410  -1.051   0.2940
## student1     426.46919   16.67770  25.571  < 2e-16 ***
## married1      -7.01910   10.27803  -0.683   0.4951
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.72 on 390 degrees of freedom
## Multiple R-squared:  0.9549, Adjusted R-squared:  0.9539
## F-statistic: 918.2 on 9 and 390 DF,  p-value: < 2.2e-16
```

- This model is better than the previous model as in this case adjusted R-squared is greater than the previous one and RSE is minimum in this case.

2.5 `student` as a dummy variable

$Balance = \beta_0 + \beta_1 \times student + \epsilon_0$

as it is a binary variable so for student, model will be

$Balance = \beta_0 + \beta_1 \times 1$

for non-student

$Balance = \beta_0 + \beta_1 \times 0$
i.e
$Balance = \beta_0$

```
d1<-lm(Balance~student, data = credit)
summary(d1)
```

```
##
## 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 ***
## student1      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
```

- From the above table we found that our model is significant as our p-value is less than 0.05
- balance of the student is 480.37+396.46 = **876.83** and balance for non-student is **480.37**

2.6 `married` as a dummy variable

$Balance = \beta_0 + \beta_1 \times married + \epsilon_0$

```
d2<-lm(Balance~married, data = credit)
summary(d2)
```

```
##
## 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 ***
## married1      -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
```

- From the above calculated table we found that our model is insignificant as our p-value is greater than 0.05.

2.7 `own` as a dummy variable

$Balance = \beta_0 + \beta_1 \times own + \epsilon_0$

as it is a binary variable so for own, model will be

$Balance = \beta_0 + \beta_1 \times 1$

for non-own

$Balance = \beta_0 + \beta_1 \times 0$
$i.e$
$Balance = \beta_0$

```
d3<-lm(Balance~own, data = credit)
summary(d3)
```

```
##
## 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 ***
## own1           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
```

- From the above calculated we found that our model is in significant as it is greater than 0.05.

3 Assingment 3

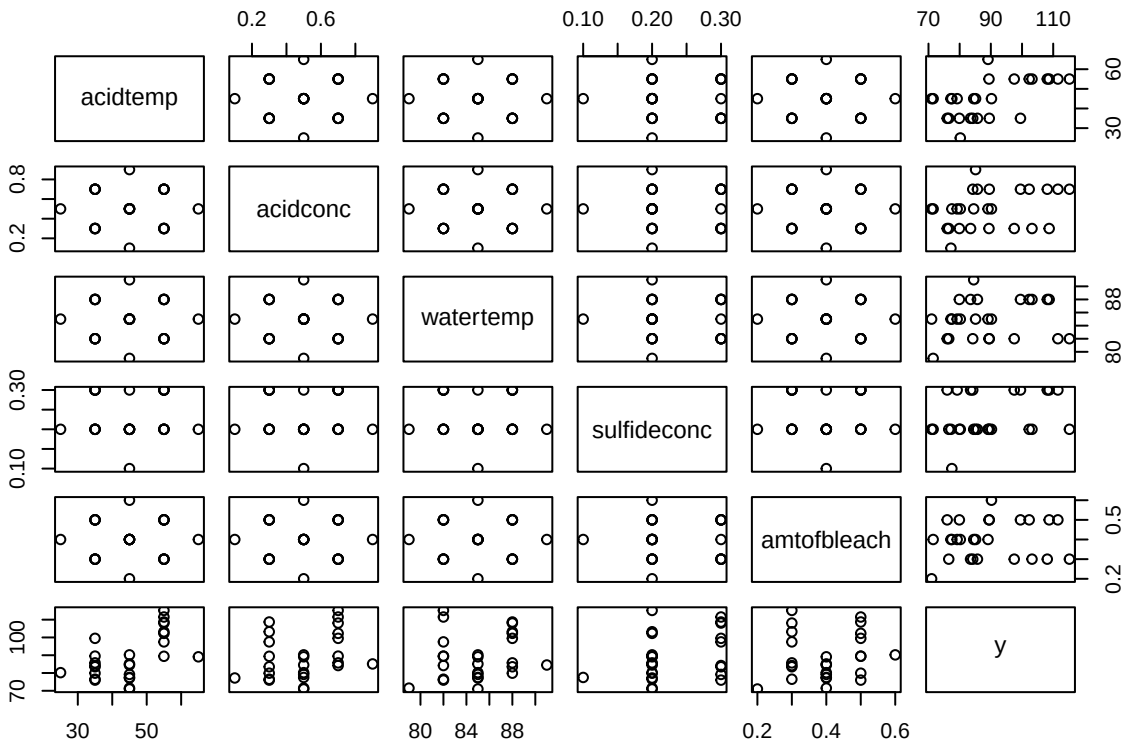
3.1 read the data from excel to r

```
chemical<-read.csv("ChemicalData.csv")
head(chemical)
```

```
##   acidtemp acidconc watertemp sulfideconc amtofbleach    y
## 1       35      0.3        82         0.2         0.3 76.5
## 2       35      0.3        82         0.3         0.5 76.0
## 3       35      0.3        88         0.2         0.5 79.9
## 4       35      0.3        88         0.3         0.3 83.5
## 5       35      0.7        82         0.2         0.5 89.5
## 6       35      0.7        82         0.3         0.3 84.2
```

3.2 plot the data

```
plot(chemical)
```



3.3 correlation matrix of the chemical data

```
cor(chemical)
```

```
##              acidtemp      acidconc watertemp   sulfideconc   amtofbleach         y
## acidtemp    1.0000000  0.000000e+00 0.0000000  0.000000e+00  0.000000e+00 0.5709244
## acidconc    0.0000000  1.000000e+00 0.0000000 -3.491215e-17  0.000000e+00 0.3098757
## watertemp   0.0000000  0.000000e+00 1.0000000  0.000000e+00  0.000000e+00 0.1822235
## sulfideconc 0.0000000 -3.491215e-17 0.0000000  1.000000e+00 -5.079390e-17 0.3303949
## amtofbleach 0.0000000  0.000000e+00 0.0000000 -5.079390e-17  1.000000e+00 0.1318009
## y           0.5709244  3.098757e-01 0.1822235  3.303949e-01  1.318009e-01 1.0000000
```

There is a correlation between `acidtemp` and `y` i.e **0.5709244**(57% correlation between acidtemp and y) ,`sulfideconc` and `y` i.e **0.3303949** (33% correlation between sulfideconc and y) & `acidtemp` and `y` i.e **0.3098757** (30% correlation between acidconc and y)

3.4 Setting hypothesis

Null Hypothesis: \

$H_0 : \beta_1 = \beta_2 = \beta_3 \quad \cdots \quad = \beta_n = 0$

VS

Alternative hypothesis: \ $H_1 : At\ least\ one\ of\ the\ \beta_i\ 's\ are\ not\ zero$

3.4.1 fit the model.

$y = \beta_0 + \beta_1 \times acidtemp + \beta_2 \times acidtemp + \beta_3 \times watertemp + \beta_4 \times sulfideconc + \beta_5 \times amtofbleach + \epsilon_0$

```
chem<-lm(y~.,data = chemical)
summary(chem)
```

```
##
## Call:
## lm(formula = y ~ ., data = chemical)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2133  -5.3674   0.0128   5.1365  21.0837
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -46.6289    55.4735  -0.841 0.410530
## acidtemp      0.7454     0.1888   3.948 0.000795 ***
## acidconc     20.2292     9.4409   2.143 0.044620 *
## watertemp     0.7931     0.6294   1.260 0.222161
## sulfideconc  76.9694    33.6904   2.285 0.033394 *
## amtofbleach  17.2083    18.8817   0.911 0.372952
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 9.25 on 20 degrees of freedom
## Multiple R-squared:  0.5817, Adjusted R-squared:  0.4771
## F-statistic: 5.563 on 5 and 20 DF,  p-value: 0.002272
```

- from the above calculated table we can find that the model is significant as p-value is less than 0.05
- p-value of intercept, watertemp,amtofbleach is greater than 0.05 so it is insignificant but rest of the variables acidtemp,acidconc,sulfideconc are significant.

3.5 removing amtofbleach

```
chem1<-lm(y~acidtemp+acidconc+watertemp+sulfideconc, data = chemical)
summary(chem1)
```

```
##
## Call:
## lm(formula = y ~ acidtemp + acidconc + watertemp + sulfideconc,
##     data = chemical)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7163  -4.5070   0.8337   5.5026  19.3628
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -39.7456    54.7349  -0.726 0.475765
## acidtemp      0.7454     0.1881   3.964 0.000708 ***
## acidconc     20.2292     9.4027   2.151 0.043235 *
## watertemp     0.7931     0.6268   1.265 0.219676
## sulfideconc  76.9694    33.5542   2.294 0.032212 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.213 on 21 degrees of freedom
## Multiple R-squared:  0.5643, Adjusted R-squared:  0.4814
## F-statistic: 6.801 on 4 and 21 DF,  p-value: 0.001126
```

- From the above table we find that our model is significant but intercept and watertemp is insignificant to the model as their p-value is greater than 0.05.so we remove watertemp from our model.

3.6 removing watertemp

```
chem2<-lm(y~acidtemp+acidconc+sulfideconc, data = chemical)
summary(chem2)
```

```
##
## Call:
## lm(formula = y ~ acidtemp + acidconc + sulfideconc, data = chemical)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7163  -5.6578   0.9352   5.3610  16.9837
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.6641    12.6974   2.179 0.040346 *
## acidtemp      0.7454     0.1906   3.911 0.000749 ***
## acidconc     20.2292     9.5302   2.123 0.045278 *
## sulfideconc  76.9694    34.0091   2.263 0.033834 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.338 on 22 degrees of freedom
## Multiple R-squared:  0.5311, Adjusted R-squared:  0.4672
## F-statistic: 8.307 on 3 and 22 DF,  p-value: 0.0007028
```

- From the calculated table we find that our model is significant and now our all the variables as well as intercept get significant as p-value is less than 0.05 .
- Now our model is perfect fit.