Multiple Regression Analysis

Kishore Lakshmanan
StudentID: 20253583
MSc in Data Analytics - B – 2021-22
Statistics for Data Analytics
Continuous Assessment – Semester 1
National College of Ireland, IRELAND
Email:x20253583@student,ncirl.ie

***Abstract*— Nowadays data becoming more powerful weapon for each and every sector whether it would be the historical or predicted one. The most important thing in data should be precise and understandability of the content which is described in it. For that reason, it should be pre-exercised with some sort of tools to get the meaningful data. Out of which, statistics is one of the traditional methodologies to perform various tasks related to the exploration of data. In this project we are going to perform the various statistics techniques like descriptive statistics, regression, diagnostic steps to explore useful information or prediction from the credit dataset.**

Keywords—historical, predicted, statistics, methodologies, exploration.

I. INTRODUCTION

The regression models will be analyzed by the credit dataset which consists of 600+ customer data with attributes like Age in years, Level of education, Years with current employer, Years at current address, Household income in thousands, Debt to income ratio (x100), Other debt in thousands and whether the customer has previously defaulted. In this proposal, we will look into the various statistics methods like descriptive statistics, visualization techniques, regression models, diagnostic steps taken and summary of the performance fit to predict the credit amount of customers. These statistical methods are undertaken by several tools which are available in the market for paid/unpaid versions. Here we are using IBM SPSS and RStudio to visualize and analyze the dataset from the csv format. The following paper will show the above stated concepts in depth with the appropriate examples.

II. DESCRIPTION OF DATA

A. Descriptive Statistics

There are two types of variables exists in this dataset which are continuous variable and categorical variable. Descriptive statistics is useful to describe the dataset with the several information/ideas which is observed through some mathematical notations such as minimum, maximum, mean, standard deviation and skewness etc.

The continuous variable shown in the dataset are Age in years, Years with current employer, Years at current address, Household income in thousands, Debt to income ratio (x100) and other debt in thousands. These continuous variables are also known as scale level data. The categorical variable shown in the dataset are level of education and default variables. This is also known as nominal data.

In Statistics, we use descriptive or summaries to forecast the scale level variables and frequencies or crosstab functions are used for categorical variables. The following tables shows the statistics of continuous and catogorical variables.

```
DESCRIPTIVES VARIABLES=age employ income creddebt debtinc othdebt
  /STATISTICS=MEAN STDDEV VARIANCE MIN MAX SEMEAN KURTOSIS SKEWNESS.
```

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
age	687	20	56	34.87	.306	8.011	64.170	.371	.093	-.603	.186
employ	687	0	31	8.36	.253	6.634	44.007	.844	.093	.280	.186
income	687	14	446	45.46	1.397	36.629	1341.669	3.927	.093	27.130	.186
creddebt	687	.011696	20.561310	1.53800213	.079995040	2.096724771	4.396	3.960	.093	22.876	.186
debtinc	687	.4	41.3	10.225	.2588	6.7825	46.002	1.102	.093	1.274	.186
othdebt	687	.045584	27.033600	3.05196002	.124801655	3.271136843	10.700	2.724	.093	10.438	.186
Valid N (listwise)	687										

Fig 2.1. Descriptive statistics of continuous variables

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
ed * default	687	100.0%	0	0.0%	687	100.0%

ed * default Crosstabulation

			default		
			0	1	Total
ed	1	Count	284	78	362
		Expected Count	267.2	94.8	362.0
	2	Count	138	58	196
		Expected Count	144.6	51.4	196.0
	3	Count	57	29	86
		Expected Count	63.5	22.5	86.0
	4	Count	24	14	38
		Expected Count	28.0	10.0	38.0
	5	Count	4	1	5
		Expected Count	3.7	1.3	5.0
Total		Count	507	180	687
		Expected Count	507.0	180.0	687.0

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Interval by Interval	Pearson's R	.108	.039	2.853	.004[c]
Ordinal by Ordinal	Spearman Correlation	.117	.039	3.075	.002[c]
N of Valid Cases		687			

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.
c. Based on normal approximation.

Fig 2.2. Descriptive statistics of catogorical variables

These categorical variables education and default deft status are more significant in nature. The above crosstabulation describes the relation between the education level and default status.

B. *Visualisation Techniques*

Data visualization is used in almost every aspect in science. Computer approaches are used by scientists from numerous fields to model complicated processes and show phenomena that are difficult to observe directly, such as weather patterns, medical problems, and mathematical relationships.

For getting a qualitative understanding, data visualization provides an important set of tools and strategies. The charts below are the basic techniques:

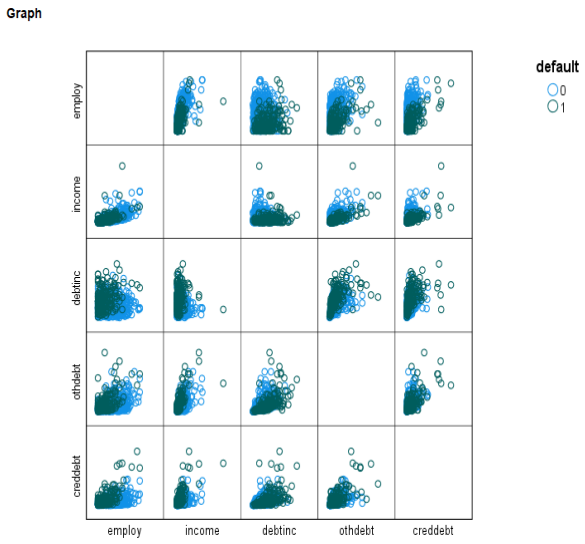


Fig 2.3. Overview of data using Scatter plot

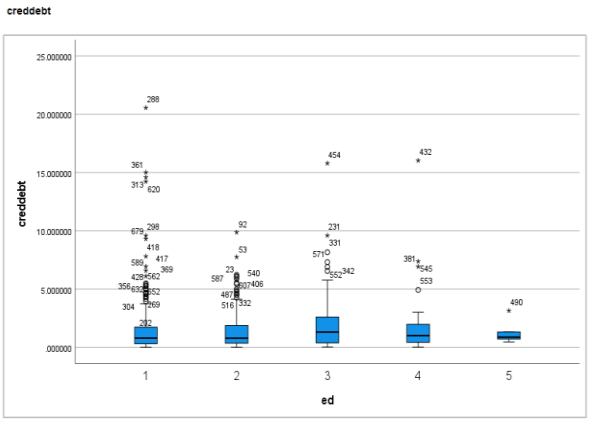


Fig 2.4. Box Plot diagram for education level vs creditdebt

The above boxplot shows that the 4th grade of education is normally distributed based on the whiskers extension on both sides of the box. This is also useful for handling large amount of data.

The histogram chart will shows whether the variables are normally distributed or skewed on its position. The below diagram describes the relation between default and creditdebt. In this both of them are positvely skewed in nature. So, it explains the strong relationship between them.

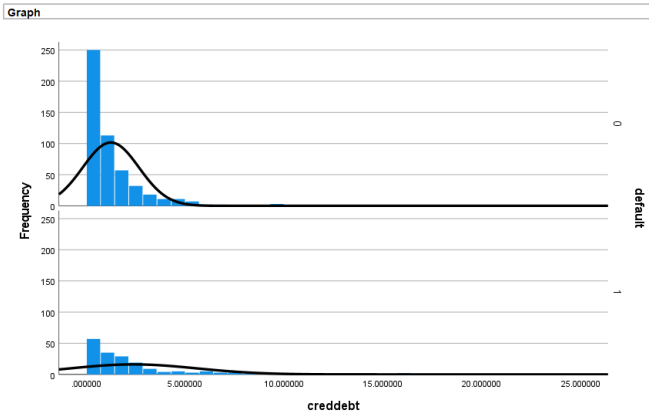


Fig 2.4. Histogram chart for default vs creditdebt

III. MODEL BUILDING PROCESS

In this section, we are going to analyze the various regression models to select the appropriate one which would pass all the diagnostic measures. In this dataset we are having one dependent variable and 8 independent variables. So, there would be a possible of 2^9 models from this dataset. We need to find the proper predictors to create our own model. The following data will show the clear idea about the model selection process.

A. *Steps taken for achieving final regression model*

The following measures should be taken prior to the selection of best model.

1) Variable Selection:

Independent variable selection is the first stage while selecting the model. The correlation concept is the best way to analyze the relationship between the depend and independent variables.

Correlations

		age	ed	employ	address	income	debtinc	creddebt	othdebt	default
age	Pearson Correlation	1	.020	.528**	.593**	.476**	.025	.293**	.345**	-.134**
	Sig. (2-tailed)		.601	.000	.000	.000	.516	.000	.000	.000
	N	687	687	687	687	687	687	687	687	687
ed	Pearson Correlation	.020	1	-.155**	.058	.233**	.012	.082*	.169**	.108**
	Sig. (2-tailed)	.601		.000	.130	.000	.754	.031	.000	.004
	N	687	687	687	687	687	687	687	687	687
employ	Pearson Correlation	.528**	-.155**	1	.317**	.619**	-.033	.401**	.405**	-.285**
	Sig. (2-tailed)	.000	.000		.000	.000	.389	.000	.000	.000
	N	687	687	687	687	687	687	687	687	687
address	Pearson Correlation	.593**	.058	.317**	1	.313**	.014	.206**	.229**	-.163**
	Sig. (2-tailed)	.000	.130	.000		.000	.722	.000	.000	.000
	N	687	687	687	687	687	687	687	687	687
income	Pearson Correlation	.476**	.233**	.619**	.313**	1	-.024	.561**	.622**	-.078*
	Sig. (2-tailed)	.000	.000	.000	.000		.535	.000	.000	.040
	N	687	687	687	687	687	687	687	687	687
debtinc	Pearson Correlation	.025	.012	-.033	.014	-.024	1	.513**	.580**	.393**
	Sig. (2-tailed)	.516	.754	.389	.722	.535		.000	.000	.000
	N	687	687	687	687	687	687	687	687	687
creddebt	Pearson Correlation	.293**	.082*	.401**	.206**	.561**	.513**	1	.643**	.240**
	Sig. (2-tailed)	.000	.031	.000	.000	.000	.000		.000	.000
	N	687	687	687	687	687	687	687	687	687
othdebt	Pearson Correlation	.345**	.169**	.405**	.229**	.622**	.580**	.643**	1	.148**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000		.000
	N	687	687	687	687	687	687	687	687	687

Fig 3.1. Correlation between the dependent and independent variables

The above figure 3.1 shows the overall relationship between the variables. In which variable having higher value close to 1 are related to each other. Here correlation between income & othdebt is .622, employ & income is .619, debtinc & othdebt is .580 and so on. From the highly related predictors we need to remove any one of the variables depending upon the p value and also relation between the dependent variable (creditdebt).

After the above steps we need to check the individual p value of all the independent variable to select the best model by running the multiple regression function. This multiple regression method is purely based upon the factor as 2^9 subsets. Out of the possible combination, we would select best fit for our process.

2) Subset method

We need to utilize the best subset method to find the Relevant variable which will be best suited for our model. The below formula is used to find the subset method as shown below and variables are selected through 3 classical approaches.

```
BestModels<-regsubsets(creddebt~income+employ+debtinc+default,data=Creditdebt, nbest = 2, method="exhaustive")
summary(BestModels)
par(mfrow=c(1,1))
plot(BestModels,scale = "adjr2")
plot(BestModels,scale = "bic")

subsets(BestModels, statistic="adjr2")
subsets(BestModels, statistic="bic")
```

The below diagram shows the subset plot using car package in Rstudio. This explains the position of each and every subsets. Hence it will be easy to pick the most suitable one.

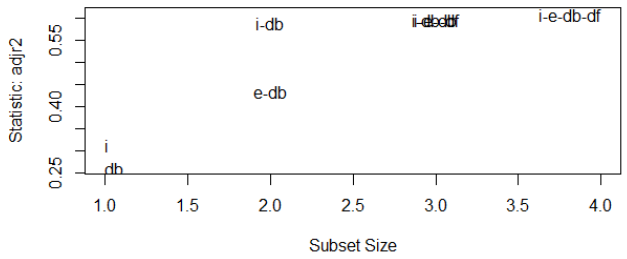


Fig 3.2 Subsets Plot

There are 3 types of automated steps to do for the large p value. They are

Forward selection
Backward selection
Mixed selection

By this method, most of the model gives us the common predictors after the auto removal of the unusual one. Below independent variables are the best one.

Employ, debtinc, income and default

With these variables we are going to implement model building steps.

3) Model Building

In the model selection, we are going to analyze from simple steps to complex based on the p value. If the acquired P value for the variable is not seems to be under threshold, then we would require to remove that variable from our model and then if needed we can add new variable to check again. Then the above process keeps on continuing until we find a best fit with the appropriate p value.

```
model1<-lm(creddebt~age+ed+employ+address+income+debtinc+othdebt+default,data = Creditdebt)
summary(model1)

model2<-update(model1,~.-age-ed-address-othdebt)
summary (model2)
```

```
Call:
lm(formula = creddebt ~ employ + income + debtinc + default,
    data = Creditdebt)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0024 -0.6196  0.0131  0.5263 12.9984

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.809635   0.116857 -15.486  < 2e-16 ***
employ       0.048338   0.010138   4.768 2.28e-06 ***
income       0.027923   0.001758  15.882  < 2e-16 ***
debtinc      0.147173   0.008072  18.232  < 2e-16 ***
default      0.645131   0.130860   4.930 1.03e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.311 on 682 degrees of freedom
Multiple R-squared:  0.6113,    Adjusted R-squared:  0.609
F-statistic: 268.2 on 4 and 682 DF,  p-value: < 2.2e-16
```

Fig 3.3 Sample modelbuilding process

With the selected model we will plot the graph to look at the simple diagnostic methods like normality, collinearity etc., to check if all passes the criteria or not. If the model fails to do, then we need to do some transformation process (see later) and stepwise analysis to rectify it.

B. Intermediate Models

The intermediate models are based upon the number of predictors we have chosen to build our model. If we have more predictors, intermediate model also more and if it is less, models also be less. Hence intermediate models are directly proportional to the selected number of variables.

```
model1<-lm(creddebt~age+ed+employ+address+income+debtinc+othdebt+default,data = Creditdebt)
summary(model1)
```

Firstly, all the predictors were gone through the model selection and after the correlation some of the variables are removed based upon the value. Any of the highly related variables must be out from the model to get the better result.

```
model2<-update(model1,~.-age-employ-othdebt)
summary (model2)
```

Secondly, we have analyzed the possible models through the selection methods where we got almost similar combinations of variables. Out of all the possible combinations, we need to select the suitable ones through the various testing methods like linearity, heteroscedacity, normality etc., After all the test there will be small number of models to analyze.

```
model3<-update(model1,~.-age-ed-address-othdebt)
summary (model)
```

Finally, we need to manually add and remove the variables to boost the strength of our model. Even though the variable is not under the selected variable it is possible to improve the model. So manually entering method is also the best one. Here address variable is added to the model to give its best. So, the following predictors are the one which is included in our final model.

Employ, debtinc, income, default and address

After gone through the summary of the variables, we have noticed some problem is there in our model. So we need to undergo some treatment and transformation from our model.

C. *Treatment of outliers and transformation undertaken*

Our model has linearity and heteroscedacity problems. So, we need to groom our model using some of the transformation methods like transforming our variables into new predictors by using various mathematical expressions like log, cos, sqrt and multiply etc.., which gives the interaction effect into our model.

We can see the below model consists of various interactions which will produce the model with removal of heteroscedacity issue and outliers are carried out by adding address variables into it.

```
model7<-lm((log(creddebt))~employ*debtinc*default*income+address,data = Creditdebt)
summary(model7)
```

```
Call:
lm(formula = (log(creddebt)) ~ employ * debtinc * default * income +
    address, data = Creditdebt)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9612 -0.4155  0.1071  0.5440  1.5925

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -2.812e+00  2.049e-01 -13.726  < 2e-16 ***
employ                        6.418e-02  1.719e-02   3.733 0.000205 ***
debtinc                       1.353e-01  2.207e-02   6.133 1.48e-09 ***
default                       7.760e-01  3.914e-01   1.983 0.047800 *
income                        2.368e-02  5.485e-03   4.317 1.82e-05 ***
address                       8.937e-03  4.549e-03   1.965 0.049854 *
employ:debtinc               -2.498e-03  1.838e-03  -1.359 0.174693
employ:default                5.333e-02  3.748e-02   1.423 0.155301
debtinc:default              -4.797e-02  2.978e-02  -1.611 0.107738
employ:income                -7.098e-04  2.931e-04  -2.422 0.015716 *
debtinc:income               -1.098e-04  6.269e-04  -0.175 0.861051
default:income               -3.728e-03  1.203e-02  -0.310 0.756791
employ:debtinc:default       -9.732e-05  2.712e-03  -0.036 0.971390
employ:debtinc:income         2.614e-05  3.497e-05   0.747 0.455141
employ:default:income        -2.198e-04  6.560e-04  -0.335 0.737624
debtinc:default:income       -4.707e-05  8.724e-04  -0.054 0.956990
employ:debtinc:default:income -9.640e-07  4.727e-05  -0.020 0.983733
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.751 on 670 degrees of freedom
Multiple R-squared:  0.6237,    Adjusted R-squared:  0.6147
F-statistic: 69.41 on 16 and 670 DF,  p-value: < 2.2e-16
```

Fig.3.4 Transformation of predictors

With these interactive models, we need to pick the suitable one because will all these variables collinearity problem arises. So, here we use stepwise method to shortlist the above patterns to become the best one. The below shown picture will explain the stepwise method.

```
model7<-lm((log(creddebt))~employ*debtinc*default*income+address,data = Creditdebt)
summary(model7)
model8<-step(model7)
model8<-update(model8,~.-default-employ-income-debtinc:default-default:income)
summary(model8)
plot(model8)
```

```
Step:  AIC=-385.68
(log(creddebt)) ~ employ + debtinc + default + income + address +
    employ:debtinc + employ:default + debtinc:default + employ:income +
    default:income

                 Df Sum of Sq    RSS     AIC
- employ:debtinc  1    0.5499 380.07 -386.69
<none>                       379.52 -385.68
- address         1    2.4979 382.02 -383.17
- employ:default  1    3.9023 383.42 -380.65
- default:income  1    6.3794 385.90 -376.23
- debtinc:default 1   12.8989 392.42 -364.72
- employ:income   1   15.3764 394.90 -360.40

Step:  AIC=-386.69
(log(creddebt)) ~ employ + debtinc + default + income + address +
    employ:default + debtinc:default + employ:income + default:income

                 Df Sum of Sq    RSS     AIC
<none>                       380.07 -386.69
- address         1    2.5547 382.63 -384.08
- employ:default  1    3.4147 383.49 -382.54
- default:income  1    6.2594 386.33 -377.46
- debtinc:default 1   12.5637 392.64 -366.34
- employ:income   1   15.4194 395.49 -361.37
>
```

Fig. 3.4 Stepwise Method

```
Call:
lm(formula = (log(creddebt)) ~ debtinc + address + employ:default +
    employ:income, data = Creditdebt)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4776 -0.4698  0.1067  0.5759  1.7463

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.726e+00  6.948e-02 -24.843  < 2e-16 ***
debtinc         1.018e-01  5.052e-03  20.152  < 2e-16 ***
address         1.866e-02  4.824e-03   3.869  0.00012 ***
employ:default  2.664e-02  9.632e-03   2.766  0.00583 **
employ:income   5.181e-04  3.936e-05  13.161  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8246 on 682 degrees of freedom
Multiple R-squared:  0.5382,    Adjusted R-squared:  0.5355
F-statistic: 198.7 on 4 and 682 DF,  p-value: < 2.2e-16
```

Fig 3.5 Final Model

The above shown method is our final model with the removal of some predictors and interactive predictors to solve collinearity issues.

IV. DIAGNOSTICS AND ASSUMPTION CHECKING

A. *Heteroscedacity*

We have seen the heteroscedacity problem when the original predictors are gone through the model. So, we have assigned log to dependent variable and some interactions within the predictors to rectify it. The following diagram illustrates the corrected model of heteroscedacity.

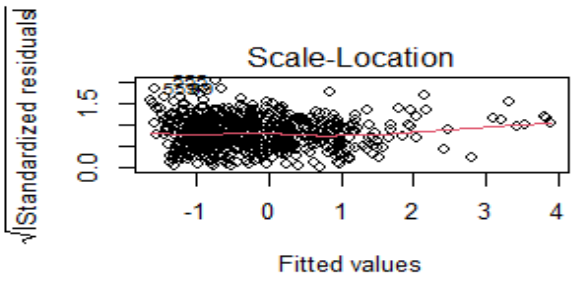


Fig 4.1 Corrected model from heteroscedacity

Heteroscedacity can be tested into two method.They are

ncvTest

bpTest

The value of ncv test should not be less than 0.05 and value of bp test should be grater than 0.05

```
> ncvTest(model8)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.854068, Df = 1, p = 0.17331
>
```

Fig 4.2 ncvTest

```
> bptest(model)  #defaults to to testing against explanatory variables

        studentized Breusch-Pagan test

data:  model
BP = 115.73, df = 9, p-value < 2.2e-16
```

Fig 4.3 BP Test

B. *Normal Distribution*

The below graph shows the normality of our model. This can be achieved by analyzing various models.

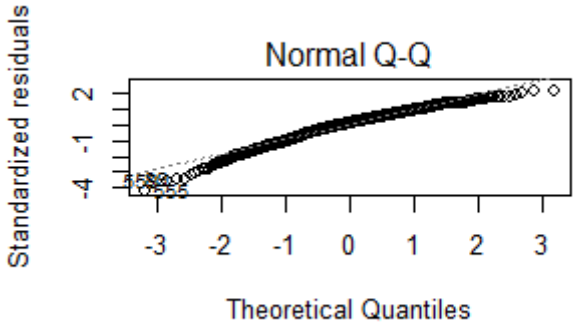


Fig 4.4 Normal Q-Q plot

C. *Linearity*

Linearity will be measured by the residual's vs fitted model like how well the depended and independent variables are related to each other. The red line in the graph should be flat and horizontal for the well-defined model.

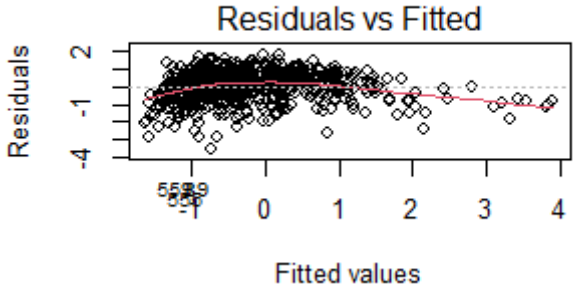


Fig 4.5 Residuals vs fitted plot

D. *Multicollinearity*

First step is to calculate correlation between two predictors. If the correlation is closer to 1 then the assumption of multicollinearity exists. But we need to analyze further to get the exact solution by the below variation inflation factor test. If the VIF is greater than 5 then it is possible to have multicollinearity problem.

```
vif(model8)
     debtinc        address employ:default  employ:income
    1.184583       1.101722       1.230015       1.154807
```

Fig 4.4 variation inflation factor test

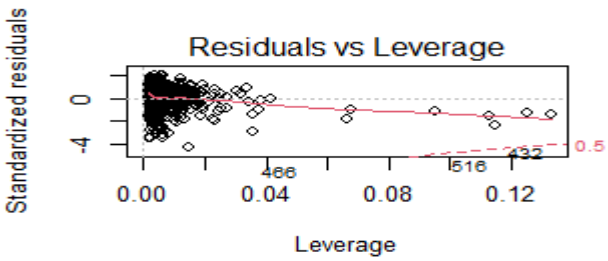
```
> durbinWatsonTest(model8)
 lag Autocorrelation D-W Statistic p-value
   1       0.04314552      1.911111   0.214
 Alternative hypothesis: rho != 0
```

Fig4.5 Test for independence of error

E. *Influential Data Points*

Influential data points are checked by the cooks Distance method. It measures the coefficient on each observation and it has value of 1 and higher are marked as influential. We have 2 or 3 observations in our model. So, we need to remove that to get the better model. It will work in large datasets but in small datasets it creates large impact.

```
> cooks.distance(model8)
            1             2             3             4             5             6             7
1.301153e-04 1.142274e-03 1.478662e-05 8.136064e-04 1.721445e-05 3.018443e-05 4.562769e-03
            8             9            10            11            12            13            14
1.250193e-03 4.731300e-05 4.888227e-05 9.225517e-05 3.928451e-03 7.125367e-04 4.456425e-05
           15            16            17            18            19            20            21
4.651582e-03 3.321097e-04 2.336488e-03 2.094453e-03 2.424216e-04 2.956751e-04 8.089728e-06
           22            23            24            25            26            27            28
2.371614e-03 1.353534e-03 6.911348e-05 1.674686e-03 8.418862e-06 8.998161e-07 3.060264e-05
           29            30            31            32            33            34            35
9.519174e-04 1.235389e-03 2.383855e-04 6.415587e-05 8.371409e-04 9.636687e-06 2.057331e-03
           36            37            38            39            40            41            42
7.032836e-05 1.139331e-03 1.229010e-04 2.198552e-05 1.082686e-03 8.055737e-05 6.462414e-06
           43            44            45            46            47            48            49
4.746162e-04 1.634789e-05 6.921727e-04 1.501708e-03 6.774250e-04 4.017094e-03 3.491616e-03
           50            51            52            53            54            55            56
9.482841e-05 1.390132e-03 3.346804e-04 1.316475e-03 5.378711e-04 7.599726e-04 2.253240e-03
           57            58            59            60            61            62            63
1.069437e-03 5.842466e-04 9.068266e-05 1.772143e-04 6.161888e-04 2.860187e-03 1.954506e-04
           64            65            66            67            68            69            70
7.862477e-03 7.237289e-04 2.843117e-04 2.641547e-04 1.869965e-03 3.470059e-04 5.929405e-04
           71            72            73            74            75            76            77
9.815089e-04 3.991553e-06 2.296550e-09 2.247943e-03 2.712465e-03 1.457783e-03 7.709609e-04
           78            79            80            81            82            83            84
3.418035e-03 1.995129e-04 4.105632e-04 2.937174e-05 4.646542e-04 1.297116e-04 2.422103e-03
           85            86            87            88            89            90            91
8.243206e-05 1.795888e-04 5.816482e-04 1.897696e-07 4.853727e-03 2.790481e-04 1.368747e-04
           92            93            94            95            96            97            98
```



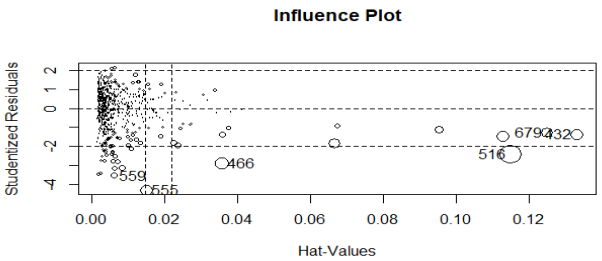


Fig 4.6 Influential data points

V. MODEL SUMMARY

In this summary section, we are going to see the performance and fit of the model. Also, we overlook into the null hypothesis and equation of multiple regression which is suitable for our model.

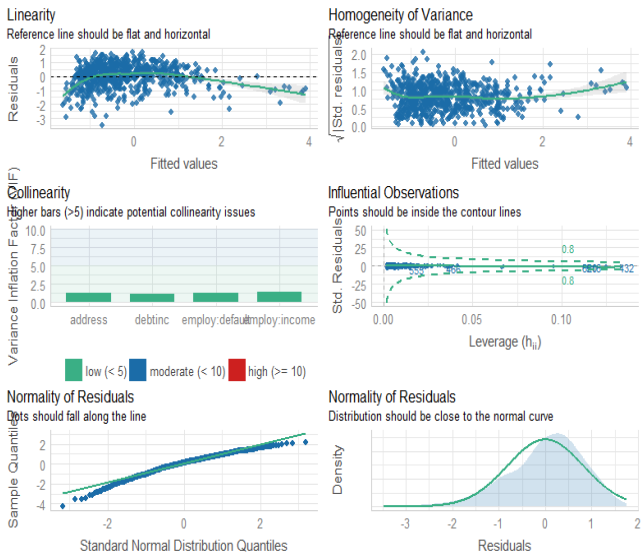


Fig 5.1 Overall performance of the model

Multiple regression equation syntax for all model is shown below.

Multiple Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

The below shown equation is used for our regression model.

$$Y = \beta_0 + \beta_1 X_6 + \beta_2 X_4 + \beta_3 X_3 X_9 + \beta_4 X_3 X_5 + €$$

By this equation we can predict how much credit the person could get in future. Based upon the test analysis of our model, accuracy of the prediction also good. Employ, debtinc, income, default and address are the best dependent variable combination which gave the best fit for our analysis. These variables are named as X1 to X9. Y is the independent variable and here we are predicting its value by the certain model that is selected.

Our model does not reject the null hypothesis and says there will be a relationship between the dependent variable Y and independent variable X.

A multiple regression's major null hypothesis is that there is no association between the X variables and the Y variables. Y variables that is, the observed data fits the model. Y values to those predicted by the multiple regression equation is no better than what you would expect by chance.

In this Model, we have got 62% of adjusted R square value before sorting the collinearity issue and after reducing some observations we have got 54%. By this model, all the diagnostic steps were passed and achieved the best performance level. Therefore, I have selected this model for as final one for multiple regression process.

By adding the othdebt predictor we are receiving more adjusted R square, but some of the diagnostic steps are failed to satisfy the model.

Reference links:

https://uc-r.github.io/descriptives_categorical
https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php