Multiple Linear Regression Analysis for Amount of Credit Card Debt Prediction

MSc in Data Analytics September 2021

Nivedita Vishwanath Hiremath

x21108471@student.ncirl.ie

Abstract: **Credit Debt is dependent on several factors. In this paper, an attempt is made to build a multiple regression model between Credit debt based on a few factors. In this project, Multiple regression model will be applied on Dataset using R.**

I. OBJECTIVES

The objective of this project is to analyze data through descriptive statistics and visualization. Applying different multiple regression models to get satisfactory R square or adjusted R square value for the model. Models are verified by Gauss Markov assumptions. If the assumptions are not satisfied, further transformation of variables and analysis will be performed.

Multiple Linear Regression: It is used to estimate the relationship between two or more independent variables and one dependent variable [1]

Equation:

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

$\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ are the regression Coefficients of the predictor variables– Each coefficient is interpreted as the value that measures a unit change in that predictor variable keeping all other variables as constant

ϵ: model error term(it tells how much variation in our estimate of

R squared: It is the proportion of variation in the response variable explained by predictor variable. R squared value can take between 0 to 1.

Formula,

$$R^2 = 1 - \frac{sum\ squared\ regression\ (SSR)}{total\ sum\ of\ squares\ (SST)}$$

Adjusted R squared: It tells the percentage of variation explained by the predictor variables that affecting the response variable. The Adjusted R-squared value increases when new independent value is added that improves the model fit. It decreases when an insignificant predictor variable is added to the model that decreases model fit.

$$R^{-2} = 1 - \frac{SSres/df_e}{SS_{tot}/df_t}$$

Residual Standard Error: It is the average amount that the response will deviate from the true regression line [2]

Gauss Markov Assumption:

- Correct functional form(Linearity) – Dependent variable should be linearly related to the independent variable. There should not be any kind of curve or grouping factor.
- Errors have constant variance(Homoscedasticity)- This assumption states that the variation of noise should be similar across the values of predictor variables. A graph of std. residuals versus fitted values show the distribution of values of the predictor variables.
- No autocorrelation between errors/Independence of errors – This states that errors or noise should be independent of each other.
- Normal Distribution of errors - This assumption states that the residuals or errors should be normally distributed.
- Absence of Multicollinearity – If there is high multicollinearity between predictor variables this assumption states to avoid multicollinearity between predictor variables.
- No influential Data points – Influential data points are the outliers. This assumption states to avoid influential data points.

II. DESCRIPTION ABOUT THE DATASET

The below Fig. 1 gives information about a dataset.

Number of columns/variables: 9

Number of observations/rows: 687

Dependent variable: creddebt

```
> str(Credit_debt)
'data.frame':   687 obs. of  9 variables:
 $ i..age  : int  52 48 36 36 43 39 41 39 47 28 ...
 $ ed      : int  1 1 2 2 1 1 3 1 1 1 ...
 $ employ  : int  6 22 9 13 23 6 0 22 17 3 ...
 $ address : int  9 15 6 6 19 9 21 3 21 6 ...
 $ income  : int  29 100 49 41 72 61 26 52 43 26 ...
 $ debtinc : num  16.3 9.1 8.6 16.4 7.6 5.7 1.7 3.2 5.6 10
 $ creddebt: num  1.716 3.704 0.818 2.918 1.182 ...
 $ othdebt : num  3.01 5.4 3.4 3.81 4.29 ...
 $ default : int  0 0 1 1 0 0 0 0 0 0 ...
```

Fig. 1. Details of Dataset

Table I. Description of variables in Dataset

Variable	Description
age	age in years
ed	level of education
employ	years with current employer
address	years at current address
income	household income in thousands
debtinc	debt to income ration
creddebt	credit card debt in thousands
othdebt	Other debt in thousands
default	the customer has previously defaulted

The below Fig. 2. shows descriptive statistics of data.

```
> summary(Credit_debt)
    i..age          ed           employ          address          income          debtinc          creddebt          othdebt      
 Min.   :20.00   Min.   : 1.000   Min.   : 0.000   Min.   : 0.000   Min.   : 14.00   Min.   : 0.40   Min.   : 0.0117   Min.   : 0.04558  
 1st Qu.:29.00   1st Qu.: 1.000   1st Qu.: 3.000   1st Qu.: 3.000   1st Qu.: 24.00   1st Qu.: 5.00   1st Qu.: 0.3686   1st Qu.: 1.04306  
 Median :34.00   Median : 1.000   Median : 7.000   Median : 7.000   Median : 34.00   Median : 8.60   Median : 0.8508   Median : 1.98171  
 Mean   :34.87   Mean   : 1.731   Mean   : 8.362   Mean   : 8.285   Mean   : 45.46   Mean   :10.23   Mean   : 1.5380   Mean   : 3.05196  
 3rd Qu.:40.00   3rd Qu.: 2.000   3rd Qu.:12.000   3rd Qu.:12.000   3rd Qu.: 54.00   3rd Qu.:14.05   3rd Qu.: 1.8877   3rd Qu.: 3.93045  
 Max.   :56.00   Max.   : 5.000   Max.   :31.000   Max.   :34.000   Max.   :446.00   Max.   :41.30   Max.   :20.5613   Max.   :27.03360  
    default      
 Min.   :0.000  
 1st Qu.:0.000  
 Median :0.000  
 Mean   :0.262  
 3rd Qu.:1.000  
 Max.   :1.000
```

Fig. 2. Descriptive statistics of Dataset

III. Data Visualization

Data Visual Analysis that needs to be done before starting model building steps -

i. Correlation between variables, Scatter plot, and Histogram Analysis of variables:

Correlation explains the significance of relationship between two variables.

Correlation value ranges from -1 to 1, where

-1 = Negative correlation(indirect correlation)

0 = No correlation

1 = Positive correlation(direct correlation)

Correlation formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Scatter plots and correlation matrices are used to recognize the relevancy between dependent and independent variables. If we see a linear relationship between response and predictor variables, then we need to select those variables for our model to predict the response variable and these plots also help in analysing the relationship between independent variables to estimate the Multicollinearity problem. Histogram Analysis of variables helps to apply any form of transformation if it fails to meet the Gauss – Markov Assumptions.

Below is Fig. 3 the snapshot of the Scatter plot and correlation between variables

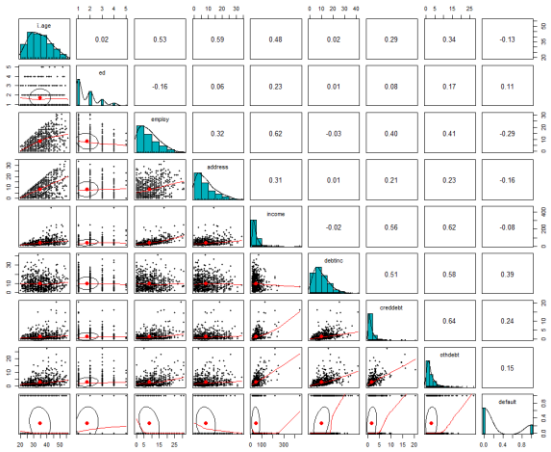

Fig. 3. Scatter plot-Correlation Matrix-Histogram

IV. Models Building Process and Description

By looking at the above Fig. 3 Scatter plot and correlation snapshot. By analysing correlation matrix the variables ‘income’, ‘debtinc’, ‘othdebt’, ‘employ’ correlates with ‘creddebt’.

Model 1: This is built using independent variables ‘employ’, ‘debtinc’, ‘othdebt’, ‘income’ for predicting ‘creddebt’.

```
> model1<-lm(creddebt~employ+debtinc+othdebt+income,data=Credit_debt)
> summary(model1)

Call:
lm(formula = creddebt ~ employ + debtinc + othdebt + income,
    data = Credit_debt)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3309 -0.5792  0.0378  0.5076 13.4381

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.869909   0.141092 -13.253  < 2e-16 ***
employ       0.033763   0.009784   3.451 0.000594 ***
debtinc      0.179174   0.011566  15.492  < 2e-16 ***
othdebt     -0.055379   0.030629  -1.808 0.071033 .
income       0.032168   0.002421  13.288  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.331 on 682 degrees of freedom
Multiple R-squared:  0.5994,    Adjusted R-squared:  0.597
F-statistic: 255.1 on 4 and 682 DF,  p-value: < 2.2e-16
```

Fig. 4. Model 1 Summary

This model is giving F-statistics 255.1and p-value is <0.05 overall model is significant which means independent variables are making a contribution in predicting response variable and Adjusted R-squared value of 0.597. Although the model is significant it is failing because the presence of Heteroscedasticity which means errors does not have the constant variance. This can be verified by looking at the Fig.5 graph of Homogeneity of Variance with fan-out plot and can also verify by ncvTest.

Fig. 5. Homogeneity of Variance(Std residuals vs Fitted values)

```
> ncvTest(model1)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1315.596, Df = 1, p = < 2.22e-16
```

Fig. 6. ncvTest value

We can see above the score test is significant this means we have not met the constant variance assumption ncvTest is failed.

Model 2: This model is built to rectify the presence of Heteroscedasticity by transforming the response variable with log or sqrt. and also we can see in Fig. 3 correlation value there is a high correlation between 'employ' and 'income' of 0.62 which may lead to Multicollinearity problem removing anyone predictor variable will help to resolve it. Independent variable 'income' is removed.

```
> model2<-lm(log(creddebt)~debtinc+othdebt+employ,data=Credit_debt)
> summary(model2)

Call:
lm(formula = log(creddebt) ~ debtinc + othdebt + employ, data = Credit_debt)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3603 -0.4730  0.1132  0.5946  2.3666

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.832800   0.076802  -23.86  < 2e-16 ***
debtinc      0.096272   0.006352   15.16  < 2e-16 ***
othdebt      0.040887   0.014397    2.84  0.00465 **
employ       0.060284   0.005785   10.42  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8575 on 683 degrees of freedom
Multiple R-squared:  0.4999,    Adjusted R-squared:  0.4977
F-statistic: 227.5 on 3 and 683 DF,  p-value: < 2.2e-16
```

Fig. 7 . Model 2 Summary

This model is giving an Adjusted R-squared value of 0.4977. But it is failing of linearity assumption, this does not satisfactorily describe association between dependant and independent variables. This can be confirmed by looking at a Fig. 8. residuals vs fitted graph it has a curved-like structure.

Fig. 8 . Linearity (Residuals vs Fitted values)

Model 3: This model is built to rectify linearity assumption failure from Model 2. This can be done by non-linear transformation of predictors, such as $\log(X)$,$\sqrt{X}$ or X^2 . From the above Fig. 3 snapshot of histogram it is seen that 'debtinc' is left-skewed applying log transformation it will be resolved.

```
> model3<-lm(log(creddebt)~log(debtinc)+othdebt+employ,data=Credit_debt)
> summary(model3)

Call:
lm(formula = log(creddebt) ~ log(debtinc) + othdebt + employ,
    data = Credit_debt)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3531 -0.4378  0.1083  0.5386  2.2176

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.830337   0.110600 -25.591  < 2e-16 ***
log(debtinc)  0.947966   0.051128  18.541  < 2e-16 ***
othdebt       0.037007   0.012969   2.854  0.00445 **
employ        0.062963   0.005414  11.630  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8085 on 683 degrees of freedom
Multiple R-squared:  0.5554,    Adjusted R-squared:  0.5535
F-statistic: 284.4 on 3 and 683 DF,  p-value: < 2.2e-16
```

Fig. 9. Model 3 Summary

This model is giving F-statistic of 284.4 and p-value<0.0.5. So this model is significant and Adjusted R-squared values as 0.5535 and it has least Residual standard error of 0.8085 compared to Model 1 and Model 2 .

V. Diagnostic plots for Model 3:

To satisfy Gauss-Markov for Model3 , Diagnostic plots are built and validated in R.

i. Linearity : In Linearity graph Fig. 10 all points are randomly scattered across the line and does not follow any systemic pattern, it is just random noise. Proving the linearity assumption.

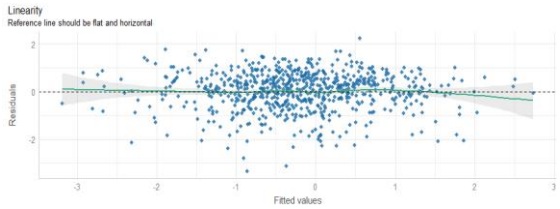

Fig. 10 . Linearity(Residuals vs Fitted values)

ii. Homoscedasticity: In the Homogeneity of Variance graph,all points are randomly scattered and systemic pattern of fan in or fan out is not found.

This can also verify by conducting ncvTest. The ncvTest Fig. 12. shows p-value =0.15742 this means we have met the non-significant variance assumption.

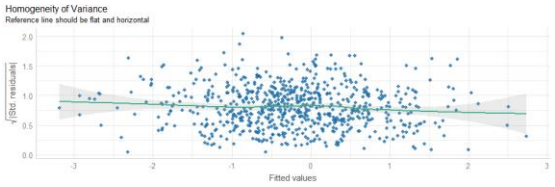

Fig. 11. Homogeneity of Variance(Std residuals vs Fitted values)

```
> ncvTest(model3)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.998799, Df = 1, p = 0.15742
```

Fig. 12. ncvTest

iii. No autocorrelation between errors: This can be verified by the Durbin-Watson test it informs whether the assumption of independent errors is tenable. The D-W statistics value should lie near 2 not less than 1 or greater than 3.

```
> durbinWatsonTest(model3)
 lag Autocorrelation D-W Statistic p-value
   1      0.02840422       1.942519   0.448
 Alternative hypothesis: rho != 0
```

Fig. 13. Durbin – Watson Test

The formula used to calculate Durbin-Watson(D-W) Statistic is:

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

Where, e_t is the residual figure and T is the number of observations

For the current dataset, the D-W Statistic value is 1.942519 which is close to 2. This satisfies the model is not violating the assumption.

iv. Normal Distribution of errors: This assumption has been validated by looking at the below normality of residuals Fig. 14. they appear close to normal distribution.

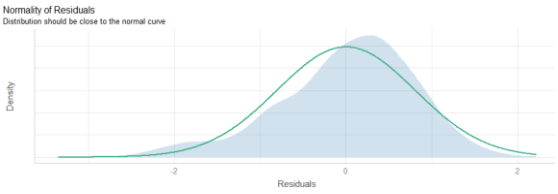

Fig. 14 . Normality of Residuals

v. Absence of Multicollinearity: This can be validated by a test known as the Variance Inflation Factor(VIF) test. The VIF provides a measure of multicollinearity among the independent variables. The VIF value should be less than <5 between independent variables to signify no multicollinearity between independent variables [3]

The formula to calculate VIF.

$$VIF_j = \frac{1}{1-R_j^2}$$

Where R_j^2 is the coefficient of determination when predictor j is regresses against all other predictors.

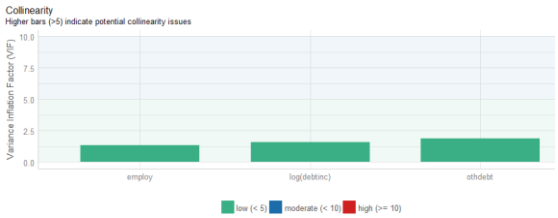

Fig. 15. Variance Inflation Factor plot

By the above graph plot, all variables' VIF value is <5. Hence no problem with Multicollinearity.

vi. No influential Data points: To check if there are any influential observations in the model, we look for the Residuals vs Leverage plot Fig. 16. As all the points lie inside cooks distance or contour lines, we can say that there are no influential data points.

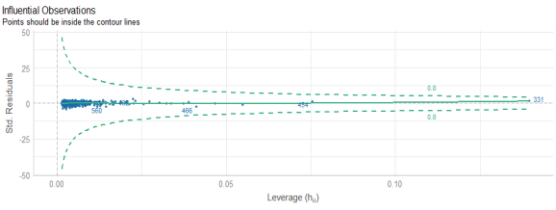

Fig. 16 . Influential Observations(Std.Residuals vs Leverage)

VI. SUMMARY OF THE FINAL MODEL3.

To get a final Model certain steps were followed -.

- Analysis on Data – Initially scatter plot and correlation matrix analysis were done. It is always good to consider those independent variables which are highly correlated with dependent variables but not with each other.
- Diagnostic plot analysis – This is done to each model to ensure our model is BLUE(best linear unbiased estimator)this is verified by Gauss-Markov assumptions. Failing to satisfy the assumption transformation steps were taken.

For Model1 by analyzing the data independent variables ‘employ’, ‘debtinc’, ‘othdebt’, ‘income’ is taken to predict ‘creddebt’ because these have high correlation values with ‘creddebt’, but this failed to meet the Homoscedasticity assumption. Thus Model1 was rejected.

For Model2 to rectify the Homoscedasticity issue with Model1 log transformation was applied on the dependent variable and also independent variable ‘income’ was also removed because it has a high correlation with the ‘employ’ independent variable in the model. But this model also failed because of the Linearity issue this does not help us to express the relationship between dependent and independent variables. Hence Model2 was rejected.

For Model 3 is built by taking minimal variables ‘debtinc’, ‘othdebt’, ‘employ’ which are significant in predicting response variable and also log transformation was applied on independent variable ‘debtinc’ . It was chosen because from above Fig. 3 histogram plot we could see ‘debtinc’ was left-skewed.

Further, Model 3 satisfies all Gauss-Markov assumptions which means that our regression model is BLUE(best linear unbiased estimator).It has F-statistic as 284.4 and its p-value < 0.05 which is significant that means our selected variables are making significant contribution in predicting response variable and it has a R-squared value of 0.5554 and Adjusted R squared(goodness of fit to model) value of 0.5535. This indicates that 55.3% of the variance for ‘creddebt’ is being explained by the features ‘debtinc’, ‘othdebt’, ‘employ’. Model 3 also showed least Residual standard error(badness of fit) value of 0.8085 compared to Model1 and Model 2.

REFERENCES:

[1] R. Bevans, “An introduction to multiple linear regression,” *Scribbr*, Feb. 20, 2020. https://www.scribbr.com/statistics/multiple-linear-regression/#:~:text=Multiple%20linear%20regression%20is%20a (accessed Nov. 25, 2021).

[2] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, "An Introduction to Statistical Learning With Applications in R", 8 th ed, pp. 71-81, 2017

[3] The Investopedia Team, “Variance Inflation Factor Definition,” *Investopedia*, Oct. 31, 2021. https://www.investopedia.com/terms/v/variance-inflation-factor.asp (accessed Nov. 25, 2021).