Statistical Approach To Estimate The Factors That Influence And Useful For The Prediction Of Credit Debt Using Multiple Regression Model

Pramod Ramu,MSc in Data Analytics September 2021,National College of Ireland,x20205759@student.ncirl.ie

***Abstract*—The report showcases that using a sample data a multiple regression model is build which is used to understand the factors that dictate and predict the amount of credit debt through statistical approach.**

***Index Terms*—Descriptive Statistics,Scatter Plots,Multiple Regression Model,Ordinary Least Square,Backward Elimination,Gauss Markov Assumptions.**

I. INTRODUCTION

In recent days most of the people prefer to do cashless transaction.Among that credit card is one of the widely used payment mode.Customer could transact large amount with a short span of time via credit card.The customer should give more importance on its usage,improper use of the card may lead to bankruptcy problem. It works based on the pay later agreement with the user.If the credit card user need to do any payment or buy a product he may use the card as the mode of payment.Once the amount has been paid, it will be added as a debt to the card holder which is need to be repaid later.

To understand the application of the multiple regression concepts , a couple of literature have been referred.Tian jinyu and Zhao xin [1] used multiple regression methods to build a model that predicts audit opinions using 30 companies as samples from Shangai and Shenzen stock markets.Similarly N.J Park, K.M.George and N.Park [2] used multiple regression techniques to build a model that predicts change in the trend. The model uses a dummy variable as the response variable and the predictor variable encompasses qualitative as well as quantitative variables.

The objective of this study is to build a model to understand about the factors which influence and also used for the prediction of the credit debt.So the study incorporates descriptive statistics to do proper analysis of the data before building a model,a linear regression model to estimate the factors,a set of diagnostic tests to achieve the Gauss Markov and other related assumptions and final result which encompasses total summary,performance and fit.

II. METHODOLOGY

The given data set consists of nine variables.Among that 'credit debit in thousands' is the dependent variable 'Y'.'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','Whether the customer has previously defaulted' are the independent variables X1,X2,X3,X4,X5,X6,X7,X8.The independent variables should have a linear relationship as well as statistically significant enough to build the model which depicts the factors that influence the credit debt.To estimate the factors the model uses ordinary least squares(OLS) approach and the focus will be on choosing the ß1,ß2..ßn to minimize the sum of squared residuals.The equation for the model is

$$Y = ß1X1 + ß2X2 + ... + ßnXn \tag{1}$$

The entire process is carried out using R programming language [3].It could be done on other platforms like IBM SPSS [4] or by using python programming language.It depends on the individual choices but more importance should be given to the final outcome of the study.

The process starts by analysis of the data using descriptive statistics and few visualization to understand the relationship and significance of the variables in the given data set which are useful for building the model. The next step would be building the model with the statistically significant as well appropriate predictor variables with the 'Credit debt' as the response variable,parallelly the model need to be evaluated by plotting the results. This will be an iterative process to achieve the best fit line.The process also involves transformations of the variables(log,square or square root) which are involved in the model for the better results as well as to meet the Gauss Markov and other related assumptions and removal of of outliers for the better shape of the model.

On building the model simultaneously a series of diagnostic tests are conducted to meet the Gauss Markov and other related assumptions [5].The assumptions are 1.'correct functional form', which means there should be a straight line relationship between the predictors and response variables.2.'Homoscedasticity',which means the prediction errors should have a constant variance.3.'No auto correlation',which means the errors should be independent.4.'Normal distribution',means the errors should be normally distributed.5.'Absence of multicollinearity',means there should be no predictors which are functions of one other.6.'No influential data points' means there should be no observations in the sample which may have leverage values.

The final step involves plotting the residuals with appropriate visualizations and also provide the detailed summary of the final model and variables involved in it.It encompasses the

	N	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
age	687	20	56	34.87	8.011	64.170	.371	.093	-.603	.186
ed	687	1	5	1.73	.931	.868	1.189	.093	.711	.186
employ	687	0	31	8.36	6.634	44.007	.844	.093	.280	.186
address	687	0	34	8.29	6.850	46.927	.946	.093	.320	.186
income	687	14	446	45.46	36.629	1341.669	3.927	.093	27.130	.186
debtinc	687	.4	41.3	10.225	6.7825	46.002	1.102	.093	1.274	.186
creddebt	687	.011696	20.561310	1.53800213	2.096724771	4.396	3.960	.093	22.876	.186
othdebt	687	.045584	27.033600	3.05196002	3.271136843	10.700	2.724	.093	10.438	.186
default	687	0	1	.26	.440	.194	1.085	.093	-.826	.186
Valid N (listwise)	687									

Fig. 1. Descriptive statistics of the given data set

results of all the diagnostic tests in a visualized format and information regarding the model performance.

III. DATA ANALYSIS

For analyzing the given data,it is good to use the descriptive statistics which gives a quick summary and the characteristics of the given data set.It mainly involves the measures of central tendency and measures of variability.Measures of central tendency encompasses mean,median,mode.It used to describe the center of the data set.Measures of variability consists of variance,standard deviation,minimum and maximum values of the variables,skewness and Kurtosis. It is a better approach to use some visualizations like histograms,scatter plots and correlation matrix of the data for understanding the relationship and the distribution of the variables.

A. Descriptive statistics

The process started by checking the descriptive statistics of the variables of the given data set. From the descriptive statistics figure 1,it can be inferred that the data set does not contain any missing values.Overall there is a variation in mean, variance and standard deviation of the variables. It seems that the mean of 'employ' variable is approximately equal to the mean of 'Income' variable where as there is slight difference in the variance and standard deviation of these two variables..

B. Visualization of the given data set

Scatter plots and other visualizations are used to identify the relationship between the variables.

From the scatter plot figure 2 it seems that there is relationship between the variables and but for the predictor variables with the statistical significance could be deduced by building regression model by using all the predictor variables against the response variable.

IV. BUILDING AND EVALUATION OF THE MODEL

Once the data has been analyzed the next step is to build the model. For selecting the statistically significant variables backward elimination method [5] has been incorporated. In the backward elimination approach all the variables are included in the model and the variables will be removed based on the highest p-value at each run and this process is carried out until the statistically significant variables are obtained.

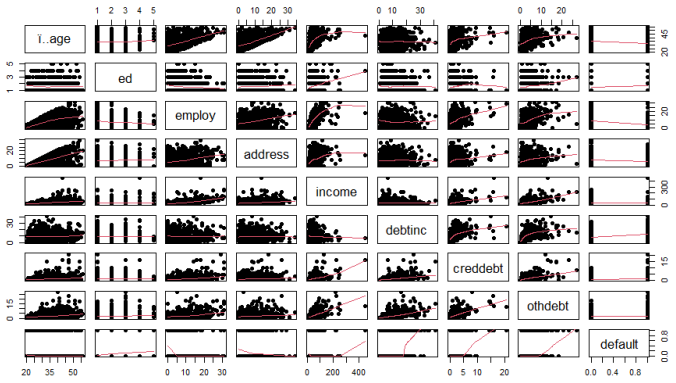

Fig. 2. Visualization of the given data set using scatter plots

A. Designing the model using backward elimination

Linear regression model was built using 'Credit Debt'as the dependent variable and 'level of education','Years with current employer','Years at current address'(address),'Household income in thousands'(income),'Debt to income ratio(X100)'(debtinc),'Other debt in thousands','Whether the customer has previously defaulted' as the independent variables.Initially it was observed that 'level of education had the highest p-value.So it was removed from the model and in the next run 'age' variable was removed,in the following run 'address' was removed.Finally the last element with highest p-value was 'other debt'.After removing it,the statistically significant independent variables in the model were 'Years with current employer','Income','Debt to income ratio' and 'Whether the customer has previously defaulted'.

When the first model was ran it was observed that the predictor errors did have the constant variance.In order to fix this issue log transformation of the response variable ('Credit Debt') was taken into consideration.After doing the transformation and running the model it was observed that the 'Whether the customer has previously defaulted' variable was having high P-value(non- statistically significant).so this variable was removed from the current model.Now in order to check the effect of log transformation of 'credit debit' on the other variables the model with all the independent variables was taken into account .At that point it was observed that 'Other Debt' variable was statistically significant.So the second model had the 'Years with current employer','Income','Debt to income ratio' and 'Other Debt' as the independent variables and the log transformed dependent variable 'Credit Debt'

On testing the second model it was observed that the linearity was not up to the mark.So to attend this issue the thumb rule was to take the log or square or square root transformations of the independent variables.So the focus was on to take the log transformations of each independent variables one by one and run the model.On taking the log transformation of the 'Income' variable and running the model it was found that the 'Years with current employer' was becoming non

```
Call:
lm(formula = creddebt ~ income + debtinc, data = given_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.00069 -0.15669  0.05651  0.21142  0.52508

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.937751   0.059484  -32.58   <2e-16 ***
income       0.150261   0.007361   20.41   <2e-16 ***
debtinc      1.003692   0.038357   26.17   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.284 on 626 degrees of freedom
Multiple R-squared:  0.6296,    Adjusted R-squared:  0.6284
F-statistic:   532 on 2 and 626 DF,  p-value: < 2.2e-16
```

Fig. 3. Final Model

statistically significant.On trouble shooting this issue with help of correlation matrix,it was found that the 'Years with current employer' and 'Income' variable are correlated.The solution was to drop one of the two variables.So it was decided to drop the 'Years with current employer ' variable.On dropping this variable,the third model had the log transformed independent and the dependent variables.

Now the third model was tested.Now all the variables were statistically significant.Decided to check the impact of changing the transformation type of the independent variables in order to improve the model.So the square root transformation was taken into the account.Initially changed the transformations of all the variables.On testing there was no much positive impact.Considering this result,square root transformation was only implemented 'Income' variable ,keeping the other independent variables log transformed.It was found that 'Other Debt' variable had the high leverage value.So this variable was dropped and the model was tested again.This model gave the satisfactory results and meets all the assumptions.The third model encompasses log transformed dependent variable 'Cred Debt' and square root transformed 'Income' variable and log transformed Debt to income ratio'.The adjacent R-Squared value for this model was 0.6284 and the prediction accuracy was 63 percent figure 3 .

In order to check the alternate option,another model was developed by dropping 'income' variable and including 'years with current employer' variable and 'Other Debt variable'.The results were not up to the mark.Initially log transformation was applied on all the independent variables.There was no much changes.Later applied square root transformation on 'years with current employer' and log transformation for 'Debt to income ratio' variable.Found satisfactory results with meeting all the assumptions.The adjacent R-Squared value for this model was 0.5389 and prediction accuracy is 54 percent.

B. Transformations of the variables

So in order to meet the Gauss and other relevant assumptions,it is good to transform the independent variables.In the model,log transformation was implemented on the dependent variable 'Credit Debt' to achieve the homoscedasticity.Square root and log transformations was implemented on the independent variables,to achieve linearity.

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.101329, Df = 1, p = 0.75024
```

Fig. 4. Non Constant Variance Test

```
 lag Autocorrelation D-W Statistic p-value
   1      0.04430307      1.909773   0.252
Alternative hypothesis: rho != 0
```

Fig. 5. Durbin Watson Test Test

C. Removal of outliers

The presence of outliers would cause distortion in the model.In order to remove the outliers interquartile range method will be incorporated.So in the given data set it is the difference between the 75th percentile(Q3) and 25th percentile(q1).So Inter Quartile Range(IQR) is defined as

$$IQR = Q3 - Q1 \tag{2}$$

So it considers the values that lie in the middle range.So any data points whose value is 1.5 times the IQR and more than Q3 that is third quartile OR 1.5 times the IQR and less than the Q1 that is the first quartile would be considered as an outlier.More importance will be given to remove that observation.

V. GAUSS MARKOV AND OTHER ASSUMPTIONS

For the best linear unbiased estimator,Gauss Markov and other relevant assumptions was taken into consideration.

A. Correct Functional Form

Initially there was an issue with the linearity in the model.Later log and square root transformations were used to fix this problem.It is observed from the residual vs fitted plot figure 6 there is a straight line relationship between the predictors and response variable.

B. Errors Have Constatnt Variance/Homoscedasticity

Non constant variance score test (NCV) was implemented to check this assumption.Initially there was a significant score.Log transformation of the dependent variable was used fix this issue.So here the p-value should be non significant to meet the assumption.so the model had P-value was 0.75024 figure 4 and also from the scale-location plot figure 7 it was observed that the model meet the homoscedasticity.

C. No Auto Correlation /Independence of Errors

Durbin-Watson statistic was incorporated to verify there is no autocorrelation between predective errors.Durbin-Watson static value should be near to 2.For the final model the value was approximately 1.91 figure 5.So the model meets the no auto correlation assumption.

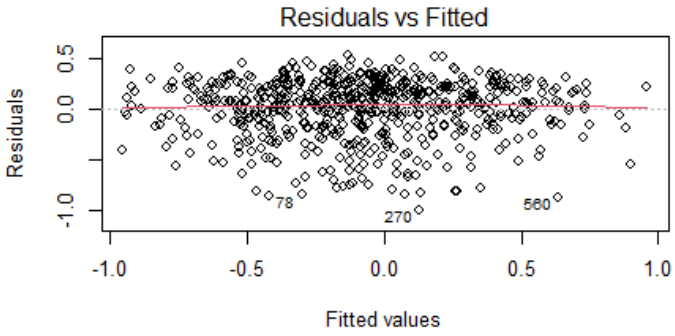

Fig. 6. Linearity

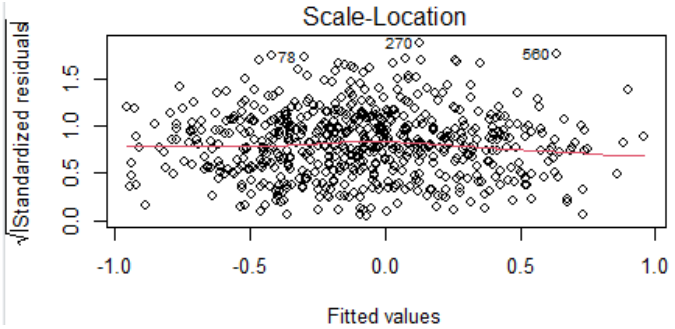

Fig. 7. Homoscedasticity

D. Errors are Normally Distributed

From the Q-Q plot figure 8 it was observed that the predictive errors are normally distributed.

E. Absence of Multicollinearity

VIF test was implemented to test multicollinearity.The VIF values for 'Income' vraiable and 'Debit Income' variabe were 1.001317.

F. No Influential Data Points

To verify that there are no influential data points in the model Cook's distance and leverage values were considered.The CookD values figure 10 was less than 1 and from the residuals vs leverage plot figure 9 it was observed that the meets this assumption.

VI. FINAL SUMMARY OF THE MODEL

In this study two models were developed.One model with adjacent R-squared value of 0.6284 figure 3 which is the final model and the prediction accuracy was 63 percent.Another model with the adjacent R-squared value of 0.5389 and with the prediction accuracy of 54 percent.Even though both the models meets all the assumptions and provide the satisfactory results but on the basis of predictive accuracy, the model with an adjacent R-Squared value of 0.6284 and 63 percent of prediction accuracy is the best among the two models.The scatter plot of the residuals of the final model vs fitted values have shown in figure 11.

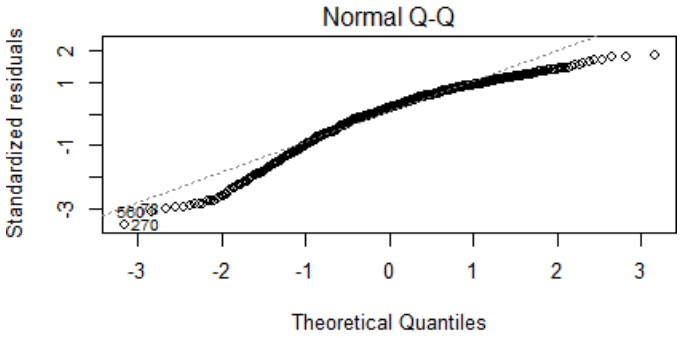

Fig. 8. Q-Q Plot

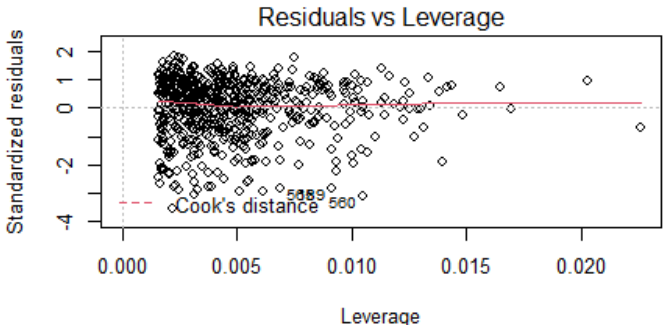

Fig. 9. Residuals vs Leverage

	StudRes	Hat	CookD
189	-2.8674509	0.009105713	0.024898632
240	0.9607217	0.020266305	0.006364934
270	-3.5603507	0.002179957	0.009062214
362	-0.6641504	0.022577444	0.003399319
560	-3.1299732	0.010489515	0.034137729

Fig. 10. No Influential Data Points

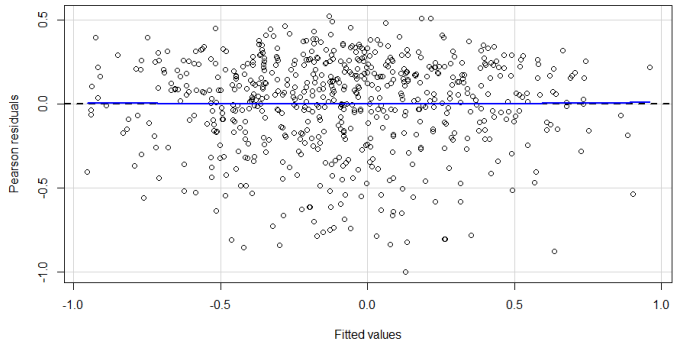

Fig. 11. Final Residual Scatter Plot

REFERENCES

[1] T. Jinyu and Z. Xin, "Apply multiple linear regression model to predict the audit opinion," in *2009 ISECS International Colloquium on Computing, Communication, Control, and Management*, vol. 4. IEEE, 2009, pp. 303–306.

[2] N. Park, K. George, and N. Park, "A multiple regression model for trend change prediction," in *2010 International Conference on Financial Theory and Engineering*. IEEE, 2010, pp. 22–26.

[3] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.

[4] A. Field, *Discovering statistics using IBM SPSS statistics*. sage, 2013.

[5] M. J. Crawley, "An introduction using r," *Á Wiley*, 2005.