Credit Debt Prediction using Multiple Regression

1st Rutuja Dinesh Mehta
MSc in Data Analytics
National College Of Ireland
Dublin, Ireland
x20129751@student.ncirl.ie

I. Introduction

Multiple Linear Regression, which is also known as Multiple Regression, is a statistical approach that uses several exploratory variables for the prediction of a model. It is used to predict the outcome of the single response variable. The main goal of the multiple regression is to build a linear relationship between the two components : two or more exploratory (also known as the independent) variables and the response (also known as the dependent) variable.

Multiple Regression equation :

$$\mathbf{Y = a + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_nX_n} \quad (1)$$

where,
Y : dependent variable
a : constant
X_1, X_2,..., X_n : independent variables
b_1, b_2,...., b_n : slope coefficients of the independent variables

II. Objective and Dataset Description

The dataset consists of 687 rows and 9 columns. The objective of the project is to predict the Credit card debt using several multiple linear regression models based on the various attributes such income, debtinc, creddebt, otherdebt, employer, etc. The detailed description of each of the variable is given below.

Variable	Description of the Variable	Data Type	Unit
age	Age	Numeric	Years
ed	Level of education	Numeric	Level : 1,2,3,4,5
employ	Years at current employer	Numeric	Years
address	Years at current address	Numeric	Years
income	Household income	Numeric	Thousands
debtinc	Debt to income ratio	Numeric	Ratio(x100)
creddebt	Credit card debt	Numeric	Thousands
otherdebt	Other debts	Numeric	Thousands
default	Whether the customer has previously defaulted	Numeric	1 : Yes, 2 : No

Fig. 1. Description of the variables

III. Understanding and Building a Model

A. Descriptive Statistics

Descriptive statistics summarizes all the basic features of the variables about the samples or the measures in the given dataset and identifies the potential relationship amongst the variables. The most commonly used descriptive statistics are a measure of central tendency(mean,median, mode), dispersion(range, variation, standard deviation, skew) and the association(chi-square, correlation). Descriptive statistics for the used dataset is given below.

Descriptive Statistics

		age	ed	employ	address	income	debtinc	creddebt	othdebt	default
N	Valid	837	837	837	837	837	837	837	837	687
	Missing	0	0	0	0	0	0	0	0	150
Mean		35.04	1.72	8.55	8.38	46.58	10.141	1.56440205	3.073980270	.26
Median		34.00	1.00	7.00	7.00	35.00	8.600	.88283300	1.995136000	.00
Mode		29	1	0	2	21	5.4	.085785[a]	3.16608600[a]	0
Std. Deviation		8.053	.931	6.761	6.917	38.427	6.6801	2.109409171	3.387519187	.440
Variance		64.846	.866	45.707	47.841	1476.648	44.623	4.450	11.475	.194
Skewness		.342	1.207	.874	.930	3.748	1.130	3.744	3.217	1.085
Std. Error of Skewness		.085	.085	.085	.085	.085	.085	.085	.085	.093
Kurtosis		-.653	.691	.421	.254	23.064	1.438	20.100	16.915	-.826
Std. Error of Kurtosis		.169	.169	.169	.169	.169	.169	.169	.169	.186
Range		36	4	33	34	433	41.2	20.549614	35.15191600	1
Minimum		20	1	0	0	13	.1	.011696	.0455840000	0
Maximum		56	5	33	34	446	41.3	20.561310	35.19750000	1
Percentiles	25	29.00	1.00	3.00	3.00	24.00	5.100	.38008000	1.043060500	.00
	50	34.00	1.00	7.00	7.00	35.00	8.600	.88283300	1.995136000	.00
	75	41.00	2.00	13.00	12.00	55.50	13.800	1.89689100	3.930448500	1.00

a. Multiple modes exist. The smallest value is shown

Fig. 2. Descriptive Statistics

The spread of the variables value namely frequency distribution, outliers and the skewness can be concluded from the histogram. The below histogram shows us that the age and debtinc have a normal distribution with a small tail in the positive direction. Rest all the variables such as income, creddebt, address, otherdebt, default, ed and employ are positively skewed. Variable income has some outliers, as well.

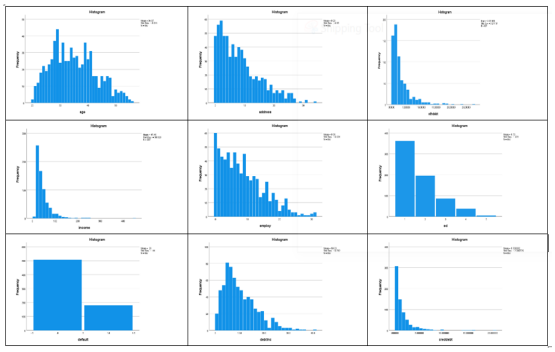

Fig. 3. Histogram

The correlation matrix shown below depicts us the strength of the relationship between the numeric coefficients. This can be used as an input or as a diagnostic for the future advanced analysis. The widely used Pearson Correlation is used for this

dataset which typically ranges from -1 to +1. The relation is said to be positive if the coefficient lies between 0 and 1 and a negative if the coefficient lies between -1 and 0. The strong relation can be seen between income and otherdebt when related to creddebt . Age and default are moderately related to creddebt. And the weakness can be seen when education and employment are in relation to creddebt.

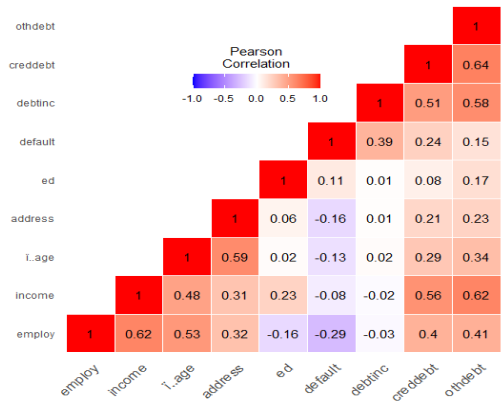

Fig. 4. Pearson Correlation

B. Model Building

a) First Model :

Variables used for building the first Model are :

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	default, income, ed, address, debtinc, age, employ, othdebt[b]	.	Enter

a. Dependent Variable: creddebt
b. All requested variables entered.

Fig. 5. Variable Model 1

All the independent variables such as income, ed, address, debtinc, age, employ, otherdebt and default are used to build a model have linearity with the variable creddebt. It can be concluded from the below model that the four non-significant variables age, ed, address and otherdebt doesn't fulfill the criteria of p<0.05 which has a p-value of 0.194, 0.203, 0.123 and 0.139 respectively.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	.784[a]	.615	.611	1.308447810	.615	135.443	8	678	.000	1.992

a. Predictors: (Constant), default, income, ed, address, debtinc, age, employ, othdebt
b. Dependent Variable: creddebt

Fig. 6. Model 1 Summary

The output of the Model 1 can be represented by the regression equation as :

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1855.071	8	231.884	135.443	.000[b]
	Residual	1160.760	678	1.712		
	Total	3015.831	686			

a. Dependent Variable: creddebt
b. Predictors: (Constant), default, income, ed, address, debtinc, age, employ, othdebt

Fig. 7. Anova Table Model 1

Coefficients[a]

Model		Unstandardized Coefficients: B	Unstandardized Coefficients: Std. Error	Standardized Coefficients: Beta	t	Sig.	95.0% Confidence Interval for B: Lower Bound	95.0% Confidence Interval for B: Upper Bound	Collinearity Statistics: Tolerance	Collinearity Statistics: VIF
1	(Constant)	-1.543	.288		-5.356	.000	-2.108	-.977		
	age	-.011	.009	-.044	-1.299	.194	-.029	.006	.503	1.988
	ed	-.077	.060	-.034	-1.274	.203	-.195	.041	.794	1.259
	employ	.047	.011	.148	4.086	.000	.024	.069	.431	2.322
	address	.014	.009	.046	1.544	.123	-.004	.032	.636	1.573
	income	.031	.003	.550	12.438	.000	.026	.036	.291	3.439
	debtinc	.160	.012	.518	13.350	.000	.136	.184	.378	2.648
	othdebt	-.045	.030	-.070	-1.480	.139	-.104	.015	.255	3.919
	default	.659	.132	.138	4.996	.000	.400	.918	.741	1.350

a. Dependent Variable: creddebt

Fig. 8. Coefficient Summary Model 1

$$Y = -1.543 - 0.011(age) - 0.077(ed) + 0.047(employ) + 0.014(address) + 0.031(income) + 0.160(debtinc) - 0.045(otherdebt) + 0.659(default)$$

Next, the global hypothesis test is conducted to check if any of the regression coefficients are other than 0. The significance level of 0.05 can be used here.

$H_0 : b_1 = b_2 = b_3 = b_4 = b_5 = b_6 = b_7 = b_8$
H_1 : Hardly any b's are 0

We reject null hypotheses since the p value in the ANOVA table is 0.000 which appears to be less than the significance level 0.05 and so arriving at the conclusion that atleast one of the regression coefficients is unequal to 0.

$$H_0 : b_n = 0$$
$$H_1 : b_n \neq 0$$

Null hypothesis is rejected for the variables age, ed, address and otherdebt as the p-value is greater than the significance level 0.05.

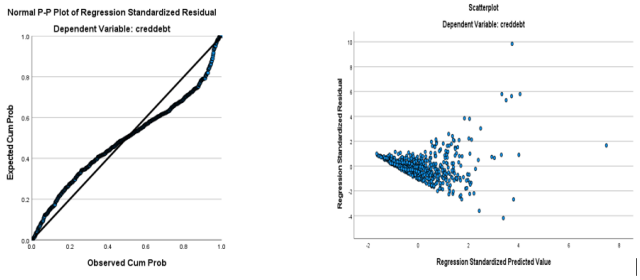

Fig. 9. P-P plot and the Scatter Plot for Model 1

The better result can be obtained from the model where these four variables are removed (value with greatest p-value or with smallest t-statistic) one by one with some applied transformations.

b) Second Model:
In this model, the independent variables age, ed, address and otherdebt, the ones with the p-value greater than 0.05 are removed and the income, debtinc and the otherdebt are transformed into their logarithmic values. The findings for the Model 2 are shown below :

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	.951[a]	.905	.905	.53819	.905	2177.444	3	683	.000

a. Predictors: (Constant), othdebt_log, income_log, debtinc_log
b. Dependent Variable: creddebt_log

Fig. 10. Model 2 Summary

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1892.094	3	630.698	2177.444	.000[b]
	Residual	197.831	683	.290		
	Total	2089.925	686			

a. Dependent Variable: creddebt_log
b. Predictors: (Constant), othdebt_log, income_log, debtinc_log

Fig. 11. Anova Table Model 2

Coefficients[a]

Model		Unstandardized Coefficients: B	Unstandardized Coefficients: Std. Error	Standardized Coefficients: Beta	t	Sig.	95.0% Confidence Interval for B: Lower Bound	95.0% Confidence Interval for B: Upper Bound	Collinearity Statistics: Tolerance	Collinearity Statistics: VIF
1	(Constant)	-22.619	.365		-62.012	.000	-23.335	-21.903		
	income_log	2.940	.052	1.416	56.624	.000	2.838	3.042	.222	4.514
	debtinc_log	2.923	.051	1.835	57.335	.000	2.823	3.023	.135	7.391
	othdebt_log	-1.903	.048	-1.543	-40.050	.000	-1.996	-1.809	.093	10.709

a. Dependent Variable: creddebt_log

Fig. 12. Coefficients Summary Model 2

Again, both the hypothesis; the individual and the global, were checked. It was found that the R and the R Square value got changed with the p-value as 0.000 for the three independent coefficients. Though the R and the R Square values, 95% and 90%, respectively stands good, it can be observed from the Fig.13 that the P-P plot has datapoints far away from the fitted distribution line along with the scatter plot with a lot of noise. Thus, giving a heteroscedasity reason for rejecting this model.
Few more transformations on the independent coefficients are applied in the next model to get the best fit regression model.

c) Third Model :
In this model, the independent variables age, ed, address and otherdebt, the ones with the p-value greater than 0.05 and the employ variable is removed and the income and debtinc are transformed into their squareroot values. Default variable

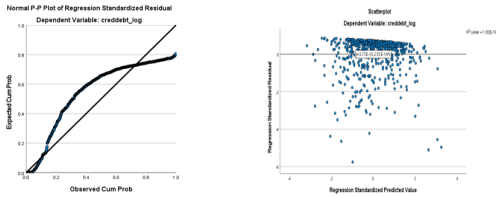

Fig. 13. P-P plot and the Scatter Plot for Model 2

with 0.002 p-value is also considered in this model. The findings for the Model 3 are shown below :

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	.820[a]	.672	.670	.33132	.672	461.962	3	677	.000	1.898

a. Predictors: (Constant), debtinc_sqrt, income_sqrt, default
b. Dependent Variable: creddebt_sqrt

Fig. 14. Model 3 Summary

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	152.135	3	50.712	461.962	.000[b]
	Residual	74.317	677	.110		
	Total	226.452	680			

a. Dependent Variable: creddebt_sqrt
b. Predictors: (Constant), debtinc_sqrt, income_sqrt, default

Fig. 15. Anova Table Model 3

Coefficients[a]

Model		Unstandardized Coefficients: B	Unstandardized Coefficients: Std. Error	Standardized Coefficients: Beta	t	Sig.	95.0% Confidence Interval for B: Lower Bound	95.0% Confidence Interval for B: Upper Bound	Collinearity Statistics: Tolerance	Collinearity Statistics: VIF
1	(Constant)	-1.031	.057		-17.935	.000	-1.144	-.918		
	income_sqrt	.162	.006	.562	25.135	.000	.149	.174	.970	1.031
	default	.097	.032	.073	3.058	.002	.035	.159	.843	1.186
	debtinc_sqrt	.339	.013	.605	25.607	.000	.313	.365	.867	1.153

a. Dependent Variable: creddebt_sqrt

Fig. 16. Coefficients Summary Model 3

It can be clearly seen from the P-P plot below that though the R and R Square values 0.820 and 0.672 respectively are reduced from that of the previous model, the datapoints have almost touched the fitted distribution line. The scatter plot also got noise reduced drastically.
Thus, looking at the transformations in this model and the comparison with the previous two, it can be concluded that the Model 3 stands better in giving us the best fit multiple regression model.

The Final Regression Equation :

Y = -1.O31 + 0.162(income_sqrt) + 0.097(default) + 0.339(debtinc_sqrt)

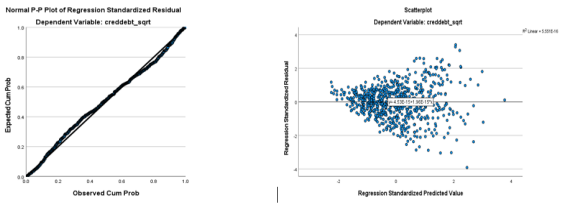

Fig. 17. P-P plot and the Scatter Plot for Model 3

IV. Assumptions

A. Gauss Markov

The Gauss Markov theorem states that if the OLS (Ordinary Least Squares) estimate for the regression coefficients gives you the best linear unbiased estimate possible if the certain set of assumptions in the theorem are met. You will be able to build your own model which resembles closely to the Gauss Markov's if you understand the violated points and correct them. For that, it's crucial to check how the data well fits these assumptions while calculating the regression coefficients.

1) Linear relationship between the dependent variable and the independent variable:

The Gauss Markov theorem states that there should be a linearity amongst dependent and the independent variables. This can be concluded by plotting a scatter plot for all the variables all at once.

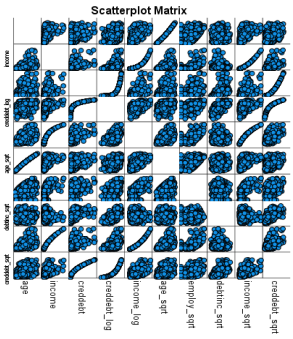

Fig. 18. Linearity between dependent and independent variables

Income_sqrt, debtince_sqrt and default has much better linearity as compared to the other independent variables with the creddebt_sqrt variable. Thus, satisfying the assumption.

2) Multicollinearity :

When two or more variables are strongly related to each other, multicollinearity occurs. This can be problematic while estimating a multiple regression model and we can't easily determine to variance impact on the independent variable then. The correlation coefficient value should not lie between -0.70 and 0.70 or the VIF value should be less than 10.

In the table above, all the independent variables have the VIF

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-1.543	.288		-5.356	.000	-2.108	-.977		
	age	-.011	.009	-.044	-1.299	.194	-.029	.006	.503	1.988
	ed	-.077	.060	-.034	-1.274	.203	-.195	.041	.794	1.259
	employ	.047	.011	.148	4.086	.000	.024	.069	.431	2.322
	address	.014	.009	.046	1.544	.123	-.004	.032	.636	1.573
	income	.031	.003	.550	12.438	.000	.026	.036	.291	3.439
	debtinc	.160	.012	.518	13.350	.000	.136	.184	.378	2.648
	othdebt	-.045	.030	-.070	-1.480	.139	-.104	.015	.255	3.919
	default	.659	.132	.138	4.996	.000	.400	.918	.741	1.350

a. Dependent Variable: creddebt

Fig. 19. Multicollinearity

value less than 10 and hence there is no multicollinearity given to the regression model.

3) Independence of Residuals :

Durbin-Watson statistics helps in testing for the independent residual values i.e no relationship between the observations. It is used for testing the auto-correlation in between the residual values in the range of 0 and 3, value when close to 2 is considered as the good independent residual value.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.784[a]	.615	.611	1.308447810	.615	135.443	8	678	.000	1.992

a. Predictors: (Constant), default, income, ed, address, debtinc, age, employ, othdebt

b. Dependent Variable: creddebt

Fig. 20. Independence of Residuals

Here, the Durbin-Watson value stands out to be 1.992, which has a better auto-correlation.

4) Homoscedasity :

Homoscedasity refers to the scattering of the residuals, for example, if the residuals are evenly distributed or they are bunched together.

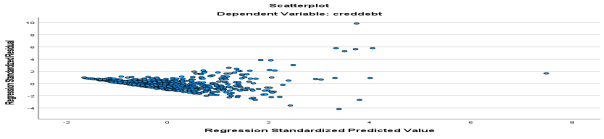

Fig. 21. Homoscedasity

The noise appears very less in the above scatter plot, thus having a minor homoscedasity.

5) No significant outliers :

Cook's distance tests the influential datapoints by measuring the effect of each observation on the regression coefficient. So, if there are any high influential points or the leverage points, they are treated as outliers as it can affect our model and bias the predicted results. The best way to identify those is by checking whether the Cook's distance falls below 1. It

can be inferred from the below table that we don't have any significant outliers as the maximum Cook's distance comes out to be 0.411.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-1.17969716	13.85119152	1.53800213	1.644440490	687
Std. Predicted Value	-1.653	7.488	.000	1.000	687
Standard Error of Predicted Value	.075	.744	.140	.052	687
Adjusted Predicted Value	-1.19280326	12.80848122	1.53669796	1.630108129	687
Residual	-5.472508907	12.87841988	.000000000	1.300796002	687
Std. Residual	-4.182	9.843	.000	.994	687
Stud. Residual	-4.374	9.996	.000	1.010	687
Deleted Residual	-5.985161781	13.28428650	.001304172	1.344184564	687
Stud. Deleted Residual	-4.434	10.818	.002	1.029	687
Mahal. Distance	1.249	220.938	7.988	11.224	687
Cook's Distance	.000	.411	.004	.026	687
Centered Leverage Value	.002	.322	.012	.016	687

a. Dependent Variable: creddebt

Fig. 22. No significant outliers

6) Normal P-P plot of Regression :

The residuals or the errors must be approximately distributed evenly across the normal distributed fitted line with a normal mean of 0. The points should hug a diagonal line for the best normal distribution. In the P-P plot below, the datapoints almost touch the fitted distribution line.

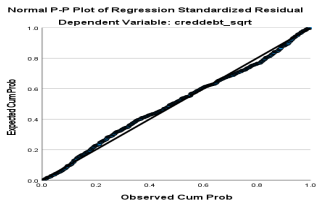

Fig. 23. Normal P-P plot of Regression

V. Model Summary

The three different models were compared with each other depending on the various factors such as R and R Square value, p-value, t-statistics, VIF statistics, P-P plot, scatter plot, etc and many more. And we arrived at our conclusion that the Model 3 bests suits for the prediction.

The reasons for the above said statement are given below :

1) R, R Square and Adjusted R Square helps in determining the coefficients in the model summary. The R Square, proportional to the variance in the dependent coefficient has a value of 0.672. Adjusted R^2 stands out to be 67%. From these values, it is pretty much clear that these variables are useful in predicting the credit card debt.

2) The value of f-statistics is compared with the critical value to test the significance of model, which is 461.96. The degrees of freedom in the numerator is k and that in the denominator is N-(k+1).

$$\textbf{Critical Value = k / (N-(k+1))} \quad (2)$$

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.820[a]	.672	.670	.33132	.672	461.962	3	677	.000	1.898

a. Predictors: (Constant), debtinc_sqrt, income_sqrt, default

b. Dependent Variable: creddebt_sqrt

Fig. 24. Final Model Summary

The critical value comes out to be 5.841 which is lesser than the F computed value. So, we reject the null hypothesis.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	152.135	3	50.712	461.962	.000[b]
	Residual	74.317	677	.110		
	Total	226.452	680			

a. Dependent Variable: creddebt_sqrt

b. Predictors: (Constant), debtinc_sqrt, income_sqrt, default

Fig. 25. Final Anova Table

3) The casewise diagnostic above provides with the actual and predicted values for the dependent variable. This gives us a picture of how our cases stand out even after having a control over all the independent variables.

Casewise Diagnostics[a]

Case Number	Std. Residual	creddebt_sqrt	Predicted Value	Residual
92	3.403	3.14	2.0152	1.12750
231	3.280	3.10	2.0117	1.08674
298	3.065	3.10	2.0819	1.01542
444	-3.025	.94	1.9445	-1.00234
466	-3.912	.91	2.2022	-1.29623
560	-3.044	.76	1.7659	-1.00864

a. Dependent Variable: creddebt_sqrt

Fig. 26. Casewise Diagnostic of the final model

References

[1] Wiley-IEEE Press, "Multiple Regression and Model Building," 10.1002/0471756482.ch3

[2] https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php