National College of Ireland

Project Submission Sheet – 2021/2022

Student Name: …Polina Prinii……

Student ID: …x21137757……..

Programme: …PGDDA…… **Year:** ……2021……

Module: …Statistics for Data Analytics……

Lecturer: …Hicham Rifai…….

Submission Due Date: …28/11/2021…….

Project Title: … Multiple Linear Regression Analysis: Credit Debt.....

Word Count: …3210…….

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: …Polina Prinii……

Date: …22/11/2021……

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Multiple Linear Regression Analysis: Credit Debt

Polina Prinii
Postgraduate Diploma in Science in Data Analytics
School of Computing, National College of Ireland, Mayor Street Lower, IFSC, Dublin 1, D01 Y300, Ireland
Email: x21137757@student.ncirl.ie

Abstract

This document serves the purpose of a business report which will outline the data used, methods used within Multiple Linear Regression and the findings of analysis on the Credit.csv file.

Keywords

CSV File; Multiple Linear Regression; Independent Variable; Dependent Variable; SPSS; Excel;

Abbrevations and Acronyms

MLR: Multiple Linear Regression

x: Independent Variable

y: Dependent Variable

Introduction

Credit debt is a major factor of everyday life. Millions of people apply for credit cards to assist with financial decisions. Using the Credit.csv file containing 800+ customers who are issued a credit card, a Multiple Linear Regression analysis was undertaken.

The purpose of this analysis was to determine if the present *Independent Variables* denoted as the x variables may have a statistically significant prediction to a person's Credit debt, which in this case is our *Dependent Variable*.

Please note that while all available data can be used as the x variables for the purpose of our MLR as each are to the analyser's interpretation, the above have been chosen based on personal interpretation and selection.

Objective

Using Multiple Linear Regression analysis, the objective of this paper is to determine whether the model derived from the MLR analysis is indeed a reliable predictor source for credit debt for potential future customers.

Multiple Linear Regression is a statistical method that uses multiple explanatory variables also known as independent variables to predict the outcome of a response variable. Multiple Linear Regression is an extension of Simple Linear Regression which uses only one explanatory variable. The MLR equation is as follow:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p,$$

Using the following tools such as Excel for preliminary analysis and SPSS as well as R for in-depth analysis the paper will investigate if there is a significant relationship between the chosen x variables and selected y variable.

The following tests will aim to confirm this are in the following form:

- **Gauss Markov Theorom (Assumptions)**
- **Correlation Coefficient (Pearson's Correlation)**
- **Correlation of Determination (R^2)**
- **Linearity (Relationship between x and y)**
- **Multicollinearity (Variance Inflation Factor)**
- **Residuals vs Fitted Scatterplot (Homoscedasticity)**
- **Significance of Relationship (p value)**

Multiple Regression Assumptions

The following assumptions are taken into consideration to ensure that the data under analyses using MLR is suitable.

The assumptions are as follows:

1. Our y variable is measured on a continuous scale.
2. Our two or more x variables are of either continuous or categorical type.
3. Our data has an independence of observations.
4. A linear relationship is present between our y variable and x variables.
5. Our data needs to show homoscedasticity.
6. Our data must not show multicollinearity.
7. Our data must not contain significant outliers.
8. Our data mut have an approximately normally distributed residual (errors).

Methodology for Analysis

The following are depicted to aid and perform the MLR to derive the results:

A. Prelimenary Analysis using Excel:

In this instance the Excel tool is used to perform a high-level view of the data under analysis. This check is undertaken with the intention to ensure that the data is accurate and does not require any transforming.

Additionally, Excel tool provides a view into the data types under analysis using MLR, together with confirming Assumption #1 to #3...

B. In-Depth Analysis using R & SPSS:

In this instance the SPSS tool is used to calculate the following: together with addressing Assumption #3 to #8:

- Gauss Markov Theorem tests and tells us that a certain set of assumptions are met. Assumptions being #3 to #8. The Theorem is applied throughout the total period of analysis of this report and is interpreted as the outlined tests are applied.
- Correlation coefficient otherwise known as Pearson's Correlation Coefficient represented as the R value. The following analysis aids in identifying which x variables are the best predictors for Credit debt based on the y variable of Credit Debt. A value closer to 1 indicates a strong indicator.
- Correlation of Determination represented by the R^2 value from the results of a MLR analysis. This will tell us the statistical measurement of how our x variables explain the difference in our y variable. In other words, outline the strength of the linear relationship between x and y. A value closer to 1 indicates a strong indicator.
- Linearity, this analysis aids to determine if there is a linear relationship between our x variables and our y variable. In this analysis two views are illustrated:
    - Linear relationship between each x variable and the one y variable.

- Linear relationship between y variable and the x variables collectively.

- Multicollinearity test aids in determining if the x variables are highly correlated with each other. The desired outcome is in a form of a negative result.
- Homoscedasticity identified as part of our MLR analysis, specifically within the Residuals vs Fitted plot.
- Significance of Relationship determined by the p-value denoted to be significant if less than 0.05. This aids in rejecting the null hypothesis which in other words confirms that changes in each predictor (x variable) are related to changes in the response (y variable).

PRELIMENARY ANALYSIS FINDINGS

On the question of preliminary analysis which demonstrates the data under analysis is not fit for purpose and several transformations are needed.

The data under analysis is outlined below in Table 1 prior to transformation.

TABLE I.

No.	Credit.csv – Data Types			
	Column Name:	***Variable Type:***	***Data Type:***	***Unit of Measurement:***
1.	age	Independent	Continuous	Age (20 - 56)
2.	ed	Independent	Nominal	Grade (1 - 5)
3.	employ	Independent	Continuous	Years (0 – 31)
4.	Address	Independent	Nominal	Code (0 -34)
5.	income	Independent	Continuous	Thousands
6.	debtinc	Independent	Continuous	(x100)
7.	creddebt	Dependent	Continuous	Thousands
8.	othdebt	Independent	Continuous	Thousands
9.	default	Independent	Nominal	Code (0 or 1)

[a.] Credit.csv – Data Types

Fig. 1. Data Types

A snapshot of the first fifteen rows of the data under analysis is attached below for the readers understanding after transforming the data.

	A	B	C	D	E	G	H	J	L
1	age	ed	employ	address	income_tran	debtinc	creddebt_tra	othdebt_tra	default
2	52	1	6	9	29000	16.3	1715.9	3011.1	0
3	48	1	22	15	100000	9.1	3703.7	5396.3	0
4	36	2	9	6	49000	8.6	817.52	3396.48	1
5	36	2	13	6	41000	16.4	2918.22	3805.78	1
6	43	1	23	19	72000	7.6	1181.95	4290.05	0
7	39	1	6	9	61000	5.7	563.27	2913.73	0
8	41	3	0	21	26000	1.7	99.01	342.99	0
9	39	1	22	3	52000	3.2	1154.82	509.18	0
10	47	1	17	21	43000	5.6	587.55	1820.45	0
11	28	1	3	6	26000	10	431.6	2168.4	0
12	29	1	8	6	27000	9.8	402.19	2243.81	0
13	21	2	1	2	16000	18	241.92	2638.08	1
14	25	4	0	2	32000	17.6	2140.16	3491.84	0
15	45	2	9	26	69000	6.7	707.32	3915.68	0
16	43	1	25	21	64000	16.7	951.23	9736.77	0
17	33	2	12	8	58000	18.4	3084.21	7587.79	0
18	26	3	2	1	37000	14.2	204.91	5049.09	0
19	45	1	3	15	20000	2.1	105	315	0
20	30	1	1	10	22000	10.5	1138.83	1171.17	0

The following columns:

- income to income_transformed
- creddebt to creddebt_transformed
- othdebt to othdebt_transformed

underwent transformation in the form of multiplying the records by 10^3 to view the values in thousands as they are presented in a decimal format.

Additionally, the preliminary analysis has shown that the assumptions #1 to #3 are met. Our y variable portrayed as creddebt_transformed in the above image is of continuous type. Our x variables are indeed of continuous or categorical type as shown above in Figure 1, as well as being of independent observation.

SUMMARY

In summary after concluding the necessary transformation for the data under analyses the preliminary analysis has shown that the data proposed for analyses using MLR is fit for purpose and has met Assumption #1 to #3.

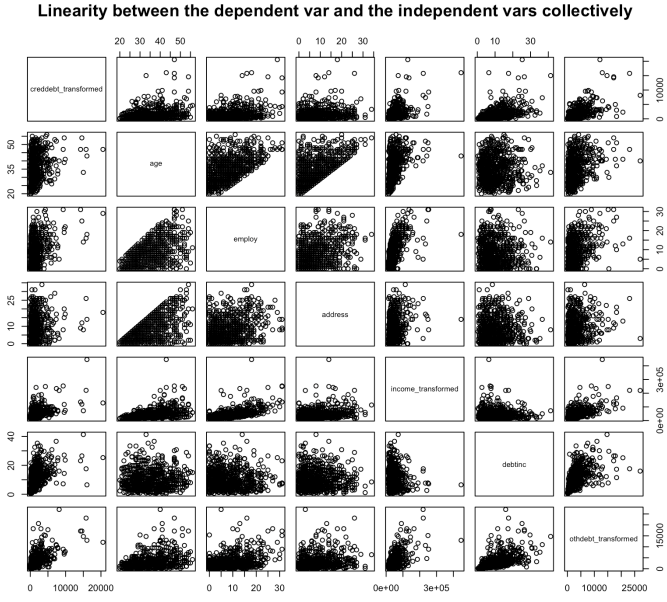
IN-DEPTH ANALYSIS FINDINGS

Let the project now turn to reviewing the findings of the in-depth analysis stage using the SPSS & R tools. Findings are not reviewed in the order of the methods outlined in the Methodology section.

A. Pre MLR Analysis - Linearity Findings:

Linearity tests for a linear relationship between our y variable and the x variables as selected. As defined in the methodology section the analysis will look to identify a linear relationship between each x variable and the one y variable as well as between the y variable and x variables collectively. The analysis excludes the Nominal Type x variables.

With the use of the R the analysis achieves to identify a linear relationship on both grounds simultaneously. Linear relationships are illustrated in the below snapshot.



The results as shown in the above snapshot, indicate that a good linear relationship is presented between the y variable (creddetb_transformed) and the selected x variables.

To further strengthen the analysis of linear relationship between the y and x variables the analysis moves to compare the means between x and y to confirm that the sig value of Deviation (p-value) is less than 0.05. This in essence will state that the deviation from linearity is less than .5%.

The mean comparison is conducted using the SPSS. Further analysis using mean comparison showed that all the x variables within the data under analysis for the exception of the Nominal type variables are of linear relationship. On average the ANOVA results show cased a p-value of .000 apart from the x variable

income_transformed which had a 1.000 Deviation from Linearity thus confirming that the relationship between the y variable is not linear as per the snapshot below.

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
creddebt_transformed * othdebt_transformed	Between Groups	(Combined)	2.977E+9	684	4351722.68	.222	.989
		Linearity	1.246E+9	1	1.246E+9	63.482	.015
		Deviation from Linearity	1.731E+9	683	2533925.19	.129	1.000
	Within Groups		39252464.5	2	19626232.3		
	Total		3.016E+9	686			

Summary:

In summary the linearity test and the results derived indicate that a linear relationship is present in all x variables present within the data under analysis. Together these results suggest that assumption #4 is not violated thus, allowing the analysis to continue to Pearson's Correlation with the aim to identify the best x variables aka predictors for the dependent y variable.

B. *Pre MLR Analysis - Correlation Coefficients Findings:*

Analysis continues with the use of Pearson's Correlation, denoted by the R value. The objective with the following test is to measure the statistical relationship or association between two continuous variables, in this instance between the dependent and independent variables.

The statistical test takes the following approach with the use of the SPSS tool. A test between the x variables collectively against the y variable is conducted.

Pearson's Correlation test has revealed that none of the selected x variables have accommodated a perfect correlation (a score close to +1) nonetheless, the analysis identified several x variables of strong correlation (a score of 0.50 to +1).

Correlations

		creddebt_transformed	age	employ	address	income_transformed	debtinc	othdebt_transformed
creddebt_transformed	Pearson Correlation	1	.293**	.401**	.206**	.561**	.513**	.643**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000
	N	687	687	687	687	687	687	687
age	Pearson Correlation	.293**	1	.528**	.593**	.476**	.025	.345**
	Sig. (2-tailed)	.000		.000	.000	.000	.516	.000
	N	687	687	687	687	687	687	687
employ	Pearson Correlation	.401**	.528**	1	.317**	.619**	-.033	.405**
	Sig. (2-tailed)	.000	.000		.000	.000	.389	.000
	N	687	687	687	687	687	687	687
address	Pearson Correlation	.206**	.593**	.317**	1	.313**	.014	.229**
	Sig. (2-tailed)	.000	.000	.000		.000	.722	.000
	N	687	687	687	687	687	687	687
income_transformed	Pearson Correlation	.561**	.476**	.619**	.313**	1	-.024	.622**
	Sig. (2-tailed)	.000	.000	.000	.000		.535	.000
	N	687	687	687	687	687	687	687
debtinc	Pearson Correlation	.513**	.025	-.033	.014	-.024	1	.580**
	Sig. (2-tailed)	.000	.516	.389	.722	.535		.000
	N	687	687	687	687	687	687	687
othdebt_transformed	Pearson Correlation	.643**	.345**	.405**	.229**	.622**	.580**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	
	N	687	687	687	687	687	687	687

**. Correlation is significant at the 0.01 level (2-tailed).

The x variables identified are as follows:

- income_transformed (.561)
- debtinc (.513)
- othdebt_transformed (.643)

Summary:

In summary Pearson's test has identified a total of three strong correlation independent variables, which going forward will be applied to the MLR model against the dependent variable for a final breakdown if the three identified variables are reliable predictors for credit debt.

C. *Multiple Linear Regression Analysis*

The report now moves to perform MLR analysis against the dependent variable identified as creddebt_transformed along the three independent variables as outlined in the *Pre MLR Analysis – Correlation Coefficients Findings.*

The report alongside the MLR analysis aims to confirm that assumptions #5 to #8 are met.

The MLR analysis can be conducted using both R and SPSS. To demonstrate understanding of both tools, findings are pulled from R and SPSS at various stages of analysis. In the below snapshot are the x variables and y variable used throughout the MRL analysis.

Regression

Variables Entered/Removed[a]

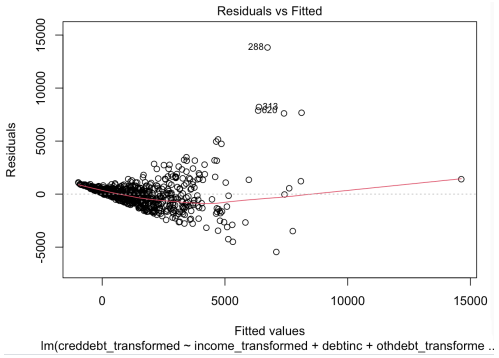
Model	Variables Entered	Variables Removed	Method
1	othdebt_transformed, debtinc, income_transformed[b]	.	Enter

a. Dependent Variable: creddebt_transformed

b. All requested variables entered.

The MLR analysis first looks at meeting assumption #5 which states that the data needs to show homoscedasticity. In the case that the data shows heteroscedasticity the data variables responsible would require to undergo transformation to correct this and meet the assumption.

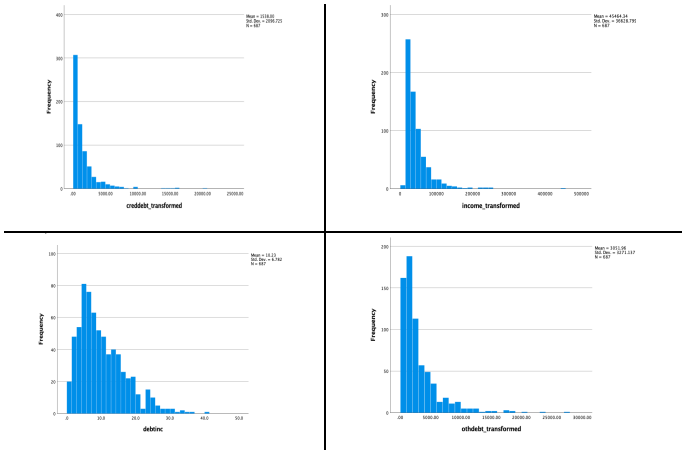
Homoscedasticity can be determined through the aid of a Residual vs Fitted plot. This is achieved by running the plot() function in R and reviewing the results. The report will review the below snapshot to determine if the data shows homoscedasticity.



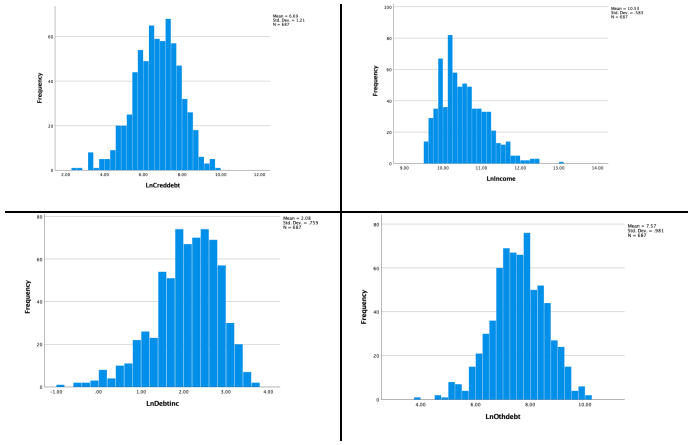
Based on the above snapshot there is a clear suggestion that the data under analysis is showing heteroscedasticity (meaning an unequal scatter) which indicates that there is a systematic change in the spread of residuals over the range of the measured values.

This presents an issue as assumption #5 of homoscedasticity is violated. To correct this the data under analysis must undergo additional transformations in the form which is identified at a later stage. The report will first look to identify which variables are a good fit for transformation. This is achieved by plotting each variable using a Histogram.

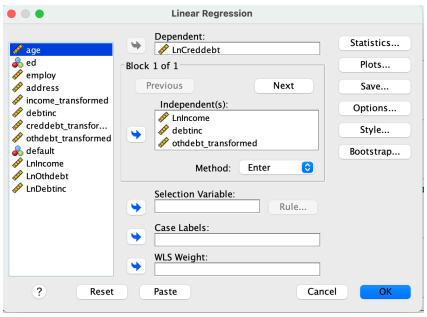
Review of each variable under analysis has shown that all four variables (one dependent and three independent) are showing an intense skew in the distribution, visualized by a Histogram as shown in the below snapshot.



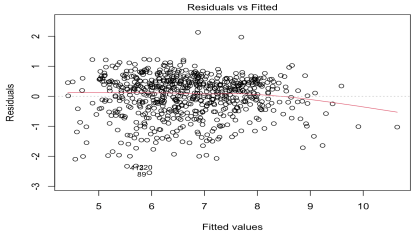
Hence, each variable is a good candidate for transformation. Transformation is applied in the form of computing the variables to derive the natural log of each, denoted as ln(). New distributions for the variables under work is portrayed in the below snapshot.



Concluding transformations, the report is presented with new data for analysis. The analysis continuous by producing a new MLR model based on the four new variables to re-evaluate the Residuals vs Fitted plot to determine if homoscedasticity has been met. Having applied different combinations between the original and transformed variables the analysis concluded the below combination as the best fit for the model which meets assumptions #5.



The new Residual vs Fitted Plot shown below allows the analysis in confidence to confirm that the data has met assumption #5 of homoscedasticity.



The analysis must note that having undertaken the natural logarithm transformation to all variables under analysis the variables entered to the MLR model have changed going forwards. In the below snapshot the report can view the x and y variables used throughout the MLR analysis.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	othdebt_transformed, debtinc, LnIncome[b]	.	Enter

a. Dependent Variable: LnCreddebt

b. All requested variables entered.

The MLR analysis then looks at Multicollinearity with the aim of determining if the x variables aka the independent variables are highly correlated with one another. For a model to be of good fit assumption #6 must be met. The analysis can either meet or reject this assumption by reviewing the results outputted from the MLR model, specifically

Multicollinearity is depicted by running a MLR analysis and selecting the collinearity diagnostics as part of the overall analysis. A Coefficients[a] table is outputted as a result. The desired information which will outline Multicollinearity lies within the VIF column.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	-11.226	.733		-15.317	.000		
	LnIncome	1.600	.070	.771	22.985	.000	.426	2.349
	debtinc	.148	.006	.832	25.499	.000	.450	2.224
	othdebt_transformed	.000	.000	-.395	-9.598	.000	.282	3.541

a. Dependent Variable: LnCreddebt

VIF or otherwise known as Variance Inflation Factor tells the measure of how much the behavior (variance) of an independent variable is influenced by other independent variables. The results of VIF look to be below 10 which demonstrates a high correlation.

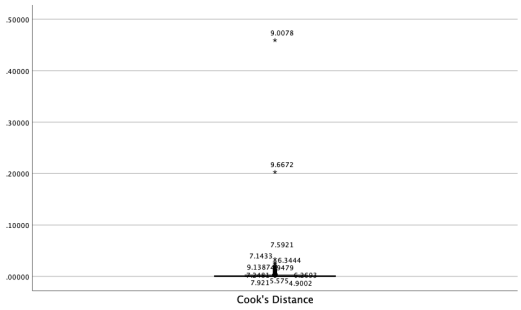
The above snapshot shows that the three x variables under analysis have a low correlation between one another thus, meeting assumption #6 which state that the data under analyses must not show multicollinearity.

Having met the assumption of the data set not showing Multicollinearity the analysis moves to expand on Pearson's test by undertaking the *Correlation of Determination* test. Denoted by the R^2 value, the report looks to identify the statistical measurement of how our x variables explain the difference in our y variable.

With the aid of the below snapshot the analysis can state that the R^2 value suggests a strong relationship between the y variable and the x variables in work as the total variation in the y variable can be explained by 67%, which in this case is relatively large.

```
Residual standard error: 0.6932 on 683 degrees of freedom
Multiple R-squared:  0.6732,    Adjusted R-squared:  0.6717
F-statistic: 468.9 on 3 and 683 DF,  p-value: < 2.2e-16
```

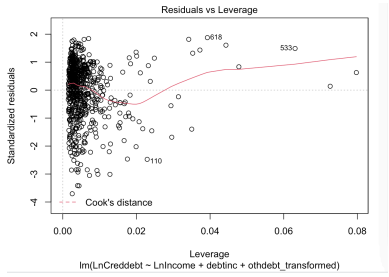
The report prior to continuing to the final evaluation of the fit to the model aims to meet assumptions #7 and #8. To do this the analysis first looks at Cook's Distance to meet assumption #7 which aims to identify any significant outliners to which need be removed from the dataset as a mitigation method to not violate the assumption. This is achieved through the MLR analysis by selecting Cook's Distance and drafting a boxplot as shown in the below snapshot.



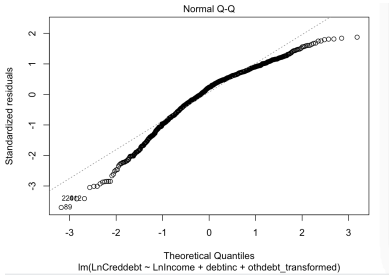
Cook's distance has identified two significant outliers which are to be removed. Them being: 9.0078 and 9.6672. This is achieved by removing the entries at the base file using Excel.

Having removed the significant outliers, the analysis has additionally removed two more outliers which after the initial removal posed significant importance.

The report can now confirm that is has met assumption #7 as the significant outlier have been dealt with. This is demonstrated in the below snapshot where it is observed that no variables are outside dotted lines.



Lastly, the report can close out on meeting all the outlined assumptions by confirming that assumption #8 is met and state that the residuals (errors) are of normal distribution approximately. This is outline by drafting a Normal Q-Q plot, below snapshot illustrates such plot.



The above snapshot clearly suggests that an approximate normality to the distribution of residuals exists in the data under analysis. Interestingly, the Q-Q plot observes Kurtosis which refers to the measure of "Taildness". In the above instance the Q-Q plot is observing a fat tail having both ends of the plot stray from the straight line.

The idea behind meeting the Gauss-Markov Theorem and ensuring all assumptions are met or if violated dealt with is for the purpose of constructing a dataset of good fit for the MLR model.

Overall, the report has succeeded in meeting all the assumption throughout both the preliminary analysis and the MLR analysis.

Throughout the overall analysis a few violations have been met and addressed to better the model fit to the MLR analysis. The report now moves to the final section of interpreting and reporting the output of the MRL analysis. At a high level the report can state that the chosen independent variables are good predictors for the dependent variable under analysis.

Multiple Linear Regression Analysis findings

The results in this report indicate that the chosen independent variables (referred by of:

- LnIncome
- debtinc
- othdebt_transformed

are good predictors for Credit Debt.

A multiple linear regression model was ran using tools such as SPSS and R to predict Credit Debt from Income, Debt to Ratio and Other Debt.

The above outlined variables have demonstrated statistical significance in predicting Credit Debt. The statistical significance is concluded based on the ANOVA results of $F(3, 678) = 485.457$ with the p value being less than .0005 (p value resulting in .000). Additionally, the R^2 value of 0.6732 tells that 67% of the variance in the dependent variable can be explained by the independent variables, a percentage above 50% is a strong indicator as it accounts for more than half.

Lastly, all three independent variables added statistical significance to the prediction. This is explained by the p value derived from the Coefficients results. The p value represents the Significance of Relationship, which if less than 0.05 which rejects the null hypothesis, in other words confirms that changes in each predictor (x variable) are related to changes in the response (y variable).

Thus, the report can conclude that the dataset applied, and transformation undertaken under analysis are a good model fit and safely state that Income, Debt to Income Ratio and Other Debt predict Credit Debt to future customers.

Acknowledgments

I would like to express my appreciation to National College of Ireland who has provided me the possibility to complete this report. A special thanks I give to our lecturer Hicham Rifai who teaches Statistics for Data Analytics, as well as Tony Delaney for producing the course material online. Both have helped me to coordinate my project, especially in understanding the objective and in writing this report.

References

The following online resources were exercised throughout this report.

[1] phrasebank.manchester.ac.uk, 'Reporting Results' [Online]. Available: https://www.phrasebank.manchester.ac.uk/reporting-results/

[2] mymoodle.ncirl.ie, 'Statistics for Data Analytics' [Online]. Available: https://mymoodle.ncirl.ie/course/view.php?id=225

[3] statisticssolutions.com, 'Testing Assumptions of Linear Regression in SPSS' [Online]. Available: https://www.statisticssolutions.com/testing-assumptions-of-linear-regression-in-spss/

[4] statistics.laerd.com, 'Linear Regression Analysis using SPSS Statistics' [Online]. Available: https://statistics.laerd.com/spss-tutorials/linear-regression-using-spss-statistics.php

[5] support.minitab.com, 'Natural log (log base e) function' [Online]. Available: https://support.minitab.com/en-us/minitab/19/help-and-how-to/calculations-data-generation-and-matrices/calculator/calculator-functions/logarithm-calculator-functions/natural-log-log-base-e-function/

[6] towardsdatascience.com, 'Q-Q Plots Explained' [Online]. Available: https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0

[7] statistics.laerd.com, 'Multiple Regression Analysis using SPSS Statistics' [Online]. Available: https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php

[8] sphweb.bumc.bu.edu, 'The Multiple Linear Regression Equation' [Online]. Available: https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704-ep713_multivariablemethods/bs704-ep713_multivariablemethods2.html