Continuous Assignment Statistics
MSCDA B

x20225776
MSc in Data Analytics
National College of Ireland
Dublin, Ireland
x20225776@student.ncirl.ie

***Abstract*—A multiple Regression Model is a powerful tool when you want to predict a numerical value by a set of other variables. This report provides insights for an understanding of the criteria that dictate the amount of credit card debt by financial behaviour of 687 customers. The final model is a parsimonious four variables subset that explained 73% of the dependent variable variance and respected all the Ordinary Least Square assumptions.**

***Index Terms*—Statistics, Multiple Regression Model.**

I. Introduction

The use of data is increasing rapidly and is probably one of the most demanded skills nowadays. Therefore, words like Machine Learning, Data Mining, Artificial Intelligence, Business Intelligence are all related to new careers like Data Science or Data Analytics.

However, the foundations of this field are not as younger at all. Although statistics is an older field, the big difference is that we have not had that amount of data, followed by the interest in using the data for continuous improvement in any process.

This document will be a Multiple Regression Analysis and the assumptions behind it. Then, using the data set about Credit Card's debt, it will be tried to understand the criteria that dictate the amount of credit card debt for each customer.

II. Description of the data

The data set is related to the financial characteristics and behaviour of 687 customers. The data frame has nine variables listed in the following list:

- **Age:** Age in years.
- **Education:** Level of education.
- **Years Current Employ:** Years with current employer.
- **Years Current Address:** Years at current address.
- **Household Income:** Household income in thousands.
- **Debt Income ratio:** Debt to income ratio (in percentage).
- **Other Debt:** Other debt in thousands.
- **Previous Defaulted:** Whether the customer has previously defaulted.
- **Credit Debt:** the amount of credit card debt in thousands.

Exploring the data set in Table I is noticed that we have a complete data frame with no missing values in any variable.

The main feature is that customers are people currently labour force with a minimum of 20 years old and a maximum of 56 (35 years old on average).

Furthermore, the people in the sample are mainly on the first two levels of education, and an average of 1.73.

In particular, about the financial behaviour, the average level of debt is around 10%. That means the proportion of all debt (Credit debt plus Other debts) and the household income, so it is not over-indebted. Moreover, just 26.2% of the sample had previously defaulted.

TABLE I
STATISTICAL DESCRIPTIVE

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Age	687	34.90	8.01	20	29	40	56
Education	687	1.73	0.93	1	1	2	5
Years_Current_Employ	687	8.36	6.63	0	3	12	31
Years_Current_Address	687	8.29	6.85	0	3	12	34
Household_Income	687	45.50	36.60	14	24	54	446
Debt_Income_ratio	687	10.20	6.78	0.40	5.00	14.00	41.30
Credit_Debt	687	1.54	2.10	0.01	0.37	1.89	20.60
Other_Debt	687	3.05	3.27	0.05	1.04	3.93	27.00
Previous_Defaulted	687	-	-	0	-	-	1

III. Modelling

A. Dependent Variable: Credit card debt

The dependent variable in the model is Credit Card debt in thousands, which is visible in the Figure 1 the distribution has a positive skew with a long tail. It can bring many problems to respect the assumptions of linear regression. One possible solution is to apply the logarithmic function and use the value in thousands. However, the values closer to zero can generate negative distortions that make problematic later interpretations.

As is shown in Figure 2 long-tail distributions have been fixed considerably, and now it is closer to a normal distribution.

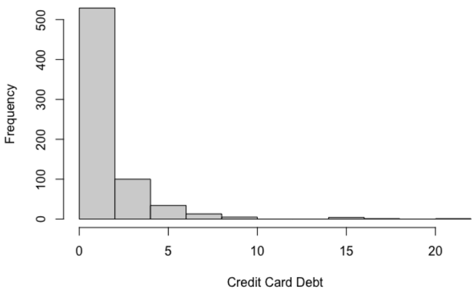


Fig. 1. Histogram of the Amount of Credit Card Debt (in thousands).

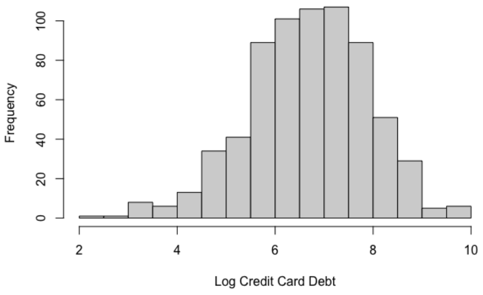


Fig. 2. Histogram of the Amount of Credit Card Debt (log).

B. Independent Variables: subset of variables

The set of independent variables also have positive skewness. This characteristic is common in variables as debts and incomes. Also, because the sample's mean is 35 years old is normal to expect to have more cases in lower values in the variables ”Years in the current address” and ”Years in the current Employ”. In Figure 3 it is possible to see a histogram with the distribution for each variable in the subset.

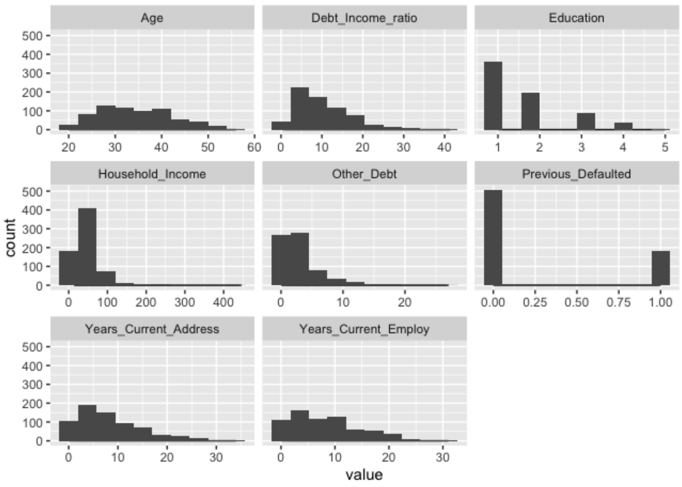
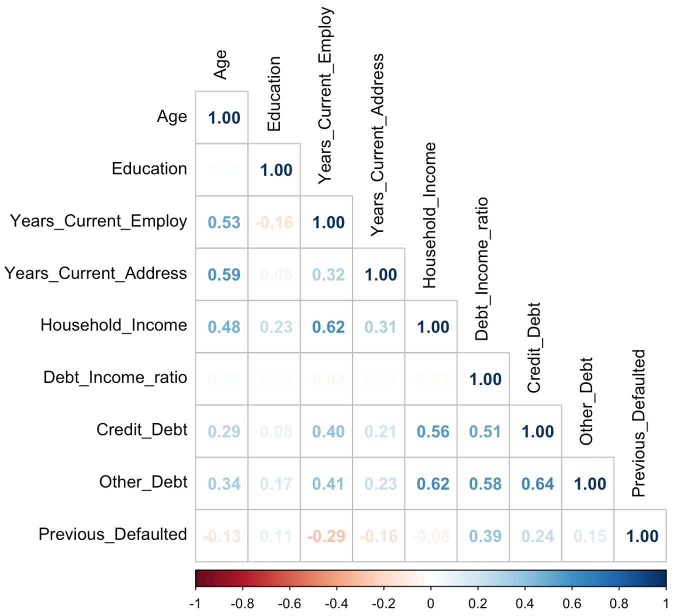


Fig. 3. Partial Correlation between variables.

C. Partial correlation

On the other hand, the statistical relation between all variables is shown in Figure 4. The blue colour represents a positive correlation, and the red colour represents negative relation. The intensity of the colour means how strong the relationship is and the correlation coefficient.

The main feature is the strong correlation between ”Credit debt” and ”Other Debt” (0.64), which will be potentially a good predictor, moreover, between independent variables like ”Household income” and ”Other debt” (0.62), and ”Years in the current Employ” (0.62). However, this kind of higher correlations (over 0.5) could be potentially dangerous with the multicollinearity assumptions; these considerations will be evaluated on the modelling process.



	Age	Education	Years_Current_Employ	Years_Current_Address	Household_Income	Debt_Income_ratio	Credit_Debt	Other_Debt	Previous_Defaulted
Age	1.00								
Education	0.02	1.00							
Years_Current_Employ	0.53	-0.16	1.00						
Years_Current_Address	0.59	0.06	0.32	1.00					
Household_Income	0.48	0.23	0.62	0.31	1.00				
Debt_Income_ratio	0.02	0.01	-0.03	0.01	-0.02	1.00			
Credit_Debt	0.29	0.08	0.40	0.21	0.56	0.51	1.00		
Other_Debt	0.34	0.17	0.41	0.23	0.62	0.58	0.64	1.00	
Previous_Defaulted	-0.13	0.11	-0.29	-0.16	-0.08	0.39	0.24	0.15	1.00

Fig. 4. Partial Correlation between variables.

D. Modelling the Multiple Regression Model

The first step was to consider all the variables in the data set (see Table II. Model 1 gives an R-square of 0.63, which is very good. Model 2 considers parsimony principles, so it was possible to explain the same variance of credit card debt with fewer variables. From Model 2, all the variables with no statistical significance were dropped on Model 2. The second model got an R-square of 0.62, which is closer but not improved.

On the other hand, trying to improve the model's performance, I have been attempting to correct the bias of the variables ”Household income” and ”Debt Income ratio”, applying a logarithmic transformation to both of them in Model 3 4, respectively.

Finally, in Model 4, the variable "Years in current Employ" had lost the statistical significance, and Model 5 is trying to predict with a reducted and parsimony subset of 4 variables:

- Other Debt
- Previously Defaulted
- Log transformation of Household Income, and
- Log transformation Debt Income ratio

TABLE II
MANUAL BACKFORWARD MODELLING

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	-2.75^{***}	-2.46^{***}	-7.15^{***}	-8.16^{***}	-8.30^{***}
	(0.16)	(0.08)	(0.29)	(0.28)	(0.25)
Years_Current_Employ	0.03^{***}	0.03^{***}	0.00	0.01	
	(0.01)	(0.01)	(0.01)	(0.01)	
Household_Income	0.02^{***}	0.02^{***}			
	(0.00)	(0.00)			
Debt_Income_ratio	0.14^{***}	0.14^{***}	0.14^{***}		
	(0.01)	(0.01)	(0.01)		
Other_Debt	-0.13^{***}	-0.13^{***}	-0.15^{***}	-0.13^{***}	-0.13^{***}
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)
as.factor(Previous_Defaulted)1	0.18^{*}	0.17^{*}	0.21^{**}	0.32^{***}	0.30^{***}
	(0.08)	(0.08)	(0.07)	(0.06)	(0.06)
Age	0.01				
	(0.01)				
as.factor(Education)2	0.09				
	(0.07)				
as.factor(Education)3	0.22^{*}				
	(0.09)				
as.factor(Education)4	0.09				
	(0.14)				
as.factor(Education)5	-0.14				
	(0.35)				
Years_Current_Address	0.01				
	(0.01)				
log(Household_Income)			1.62^{***}	1.53^{***}	1.59^{***}
			(0.09)	(0.08)	(0.06)
log(Debt_Income_ratio)				1.27^{***}	1.27^{***}
				(0.04)	(0.04)
R^2	0.63	0.62	0.68	0.73	0.73
Adj. R^2	0.62	0.61	0.68	0.73	0.73
Num. obs.	687	687	687	687	687

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

The Table above is a resume of many combinations tested as the exponential transformation of some variables and different varieties of variables as numeric or categorical (for instance, education). So after this iterative and manual process of selection, the best model is Model 5.

IV. Diagnostics and assumption checking

The Multiple Regression Assumptions can be expressed as:

1) **Linearity:**
Linearity can be defined as "the dependent variable is linearly related to the independent variables. There should be no systematic relationship between the residuals and the predicted (that is, fitted) values". It can be compared in the graphic "Residual vs Fitted" in Figure III. Despite the graph having a bit curve, it can be considered that this assumption is fully completed.

2) **Errors have constant variance (homoscedasticity):**
This definition is "the error terms have a constant variance" in Figure III the graph "Homogeneity of Variance) shows that we have met this assumption.

3) **No Autocorrelation between errors:**
Under this assumption, the dependent variable is normally distributed for a fixed set of predictor values. Then the residual values should be normally distributed with a mean of 0.

The Normal Q-Q plot is a probability plot of the standardized residuals against the values that would be expected under normality.

The graph "Normality of residuals" shows it, even if it can be improved, especially in the first quarter of the line, the 45 degrees of the line suggest that we have met this assumption as well.

TABLE III
DURBIN WATSON TEST MODEL

lag	Autocorrelation D-W	Statistic	p-value
1	0.0198	1.96	0.622

Alternative hypothesis: rho != 0

4) **No multicollinearity:**
Suppose Multicollinearity is when two or more predictors are functions of one other (possibly latently). When Multicollinearity is present, standard errors will be larger than they should be. and one mechanism for testing this is the "Variance Inflation Factor", which all are lower than 3. And also see Graphic of Collinearity[1] in Figure 5.

Finally, as the final testing of improving, I removed the "influential data Point" identified with Cook's Distance. But after removing these 5 cases, the performance of the Best Model was similar, and all assumptions checking remained similar as well. See Figure 6.

V. Model Summary

In conclusion, the Best Model can explain the 73% of the Credit card debt variance, which is very high considering that it uses only four variables. The mathematical expression is defined like:

(i) $Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \beta_4 * X_4 + \mu$

(ii) $log(Creditcard\_Debt) = \beta_0 + \beta_1 * Other\_debt + \beta_2 * Previous\_Defaulted + \beta_3 * log(Household\_income) + \beta_4 * log(Debt\_Income\_ratio)$

(iii) $log(\hat{Y}) = -8.30 + -0.13 * Other\_debt + 0.30 * Previous\_Defaulted + 1.59 * log(Household\_income) + 1.27 * log(Debt\_Income\_ratio)$

Just as a final reminder, because the Y predicted is the log of $\hat{Y}$, it is essential to apply the exponential to predicted values to get the amount of credit card debt in thousands. And always the real value of Y is equal to value predicted of Y ($\hat{Y}$ plus the error term (ϵ).

(iv) $Y = \hat{Y} + \epsilon$

[1] Note: during the modelling process was tried to use log(Other_Debt), but it transformation generate bigger issues of multicollinearity in the model.

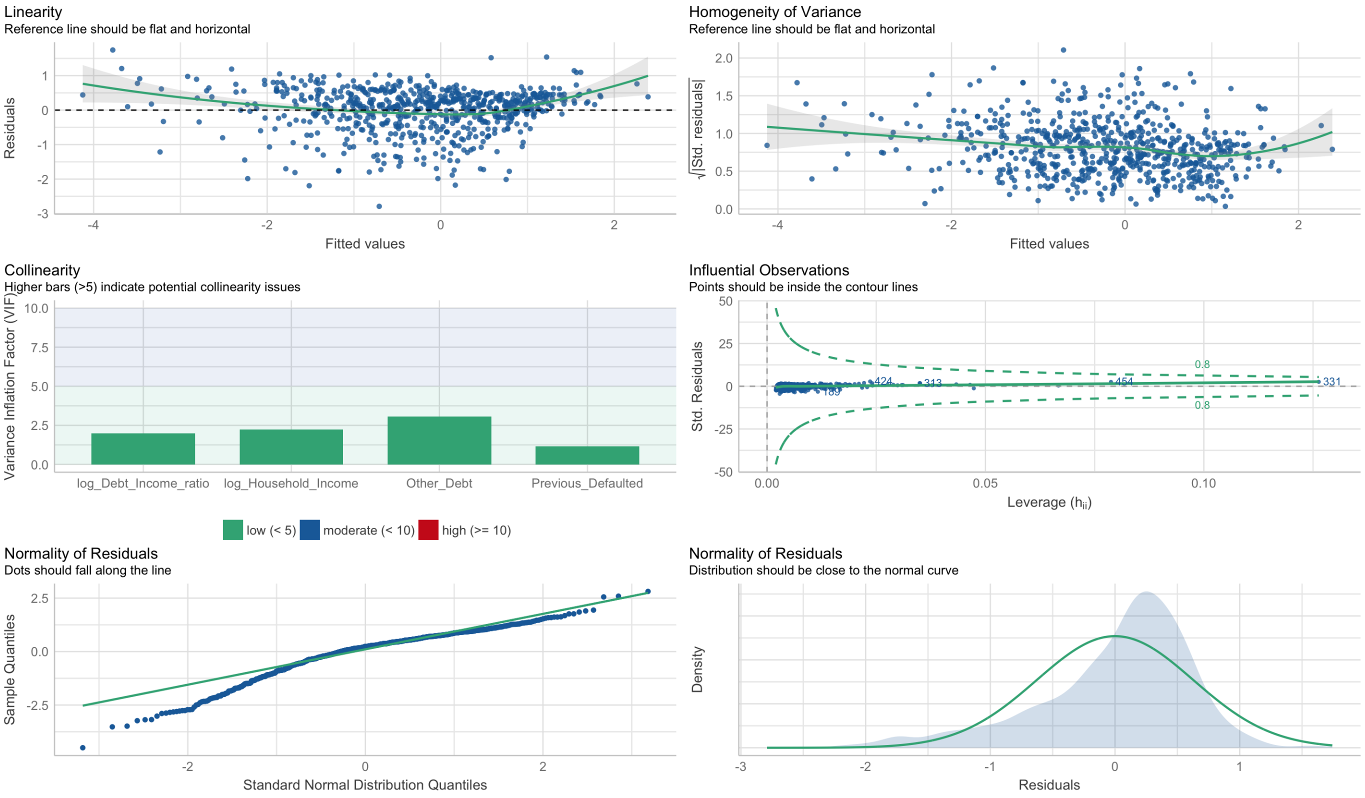


Fig. 5. Partial Correlation between variables.

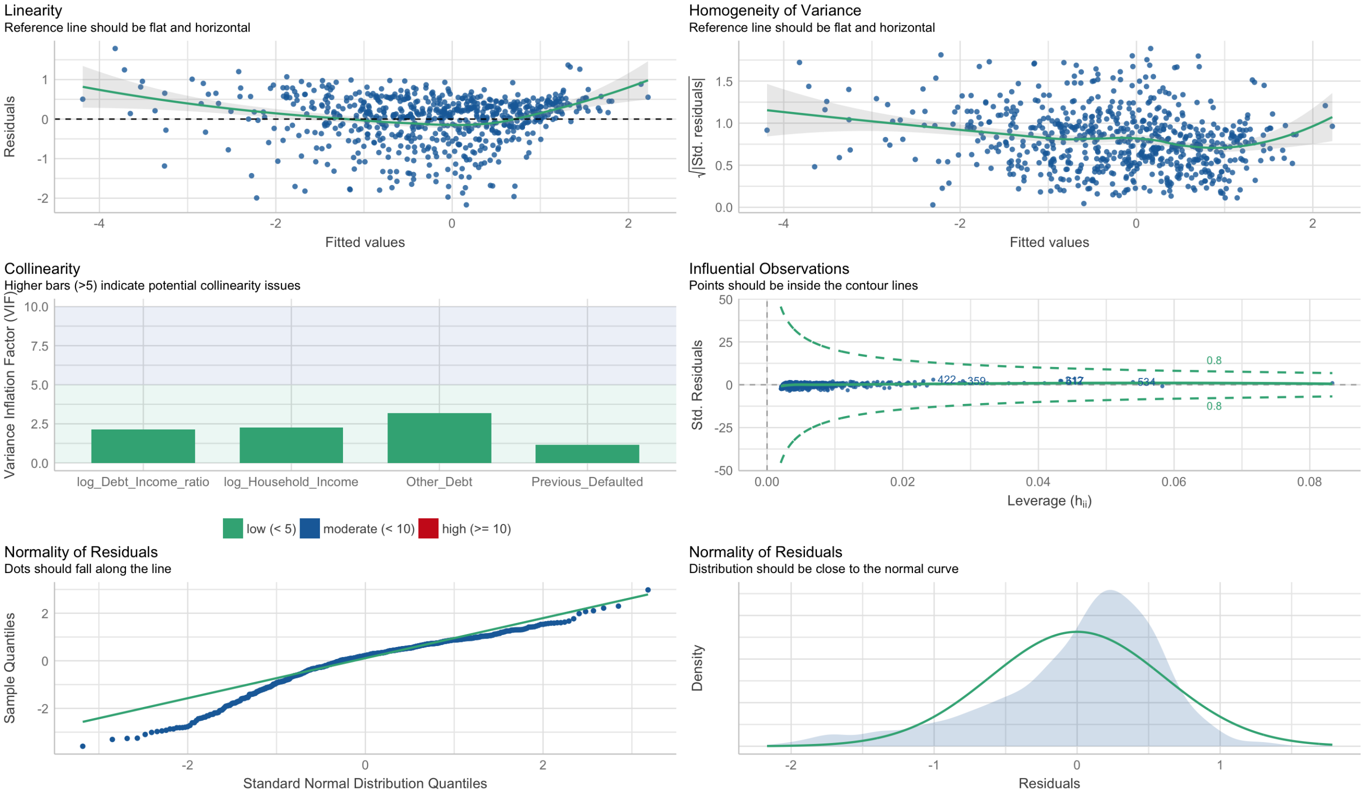


Fig. 6. Partial Correlation between variables.