Data Analysis Project

Srijesh Reddy Yarram

AA-5000-12: Foundations of Analytics

Dr.Srikanth Mudigonda

March 13, 2022

Introduction

The dataset used for analysis in this project is "**debt**" which has been extracted from the package "**faraway**" which has the functions and datasets for books by Julian Faraway. The data arise from a large postal survey on the psychology of debt. The dataset has been obtained by installing the "**faraway**" package and by using '*data(debt)*' function to retrieve the dataset to R-environment.

Numerical summaries of the variables in the dataset:

incomegp	house	children
Min. :1.000	Min. :1.000	Min. :0.0000
1st Qu.:2.000	1st Qu.:2.000	1st Qu.:0.0000
Median :3.000	Median :2.000	Median :1.0000
Mean :3.105	Mean :2.043	Mean :0.9605
3rd Qu.:4.000	3rd Qu.:2.000	3rd Qu.:2.0000
Max. :5.000	Max. :3.000	Max. :4.0000

singpar	agegp	bankacc
Min. :0.00000	Min. :1.000	Min. :0.0000
1st Qu.:0.00000	1st Qu.:2.000	1st Qu.:1.0000
Median :0.0000	Median :2.00	Median :1.000
Mean :0.05592	Mean :2.46	Mean :0.8421
3rd Qu.:0.00000	3rd Qu.:3.00	3rd Qu.:1.0000
Max. :1.00000	Max. :4.000	Max. :1.0000

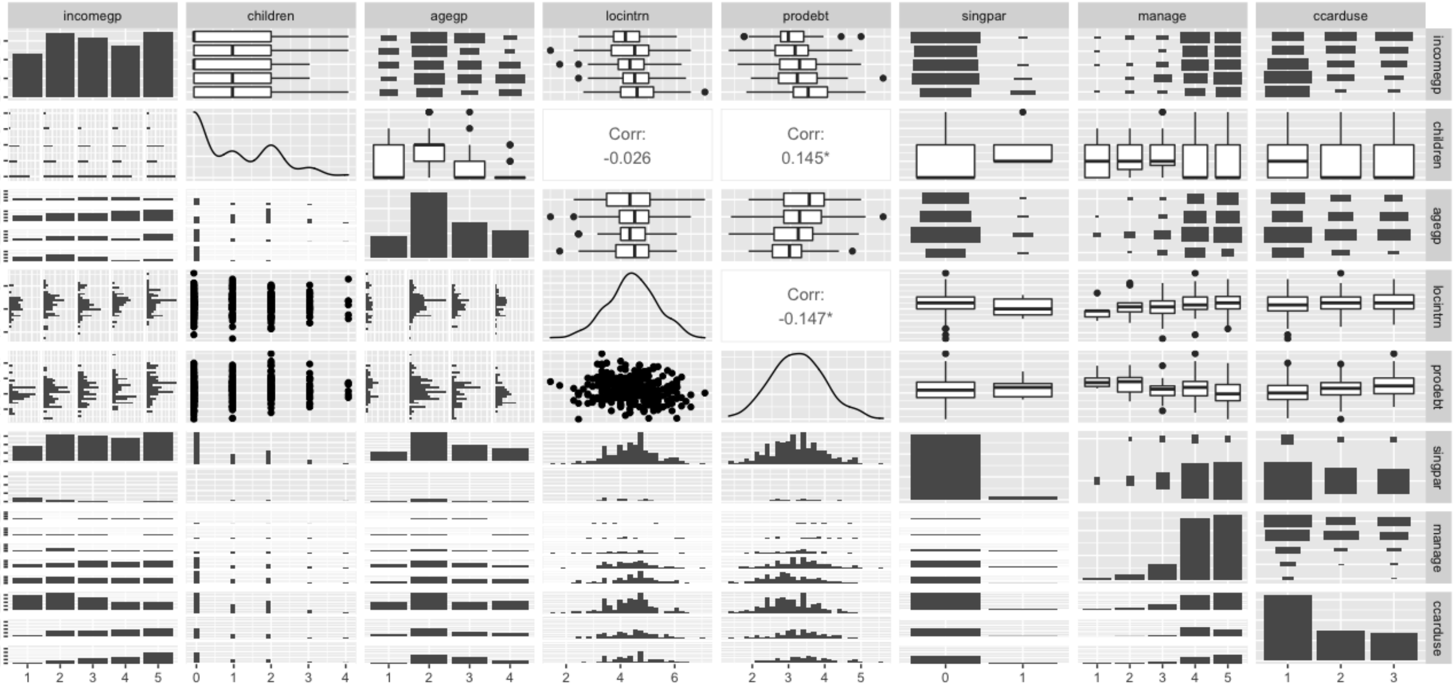
cigbuy	xmasbuy	locintrn	prodebt
Min. :0.000	Min. :0.000	Min. :1.500	Min. :1.350
1st Qu.:0.000	1st Qu.:1.000	1st Qu.:3.830	1st Qu.:2.710
Median :0.000	Median :1.000	Median :4.415	Median :3.180
Mean :0.319	Mean :0.875	Mean :4.413	Mean :3.199
3rd Qu.:1.000	3rd Qu.:1.000	3rd Qu.:5.000	3rd Qu.:3.650
Max. :1.0000	Max. :1.000	Max. :7.000	Max. :5.470

There are 13 variables in the data set. Out of those the variables that I have observed to be converted into Factor form are incomegp, house, singpar, agegp, bankacc, bsocacc, manage, ccarduse, cigbuy, xmasbuy.

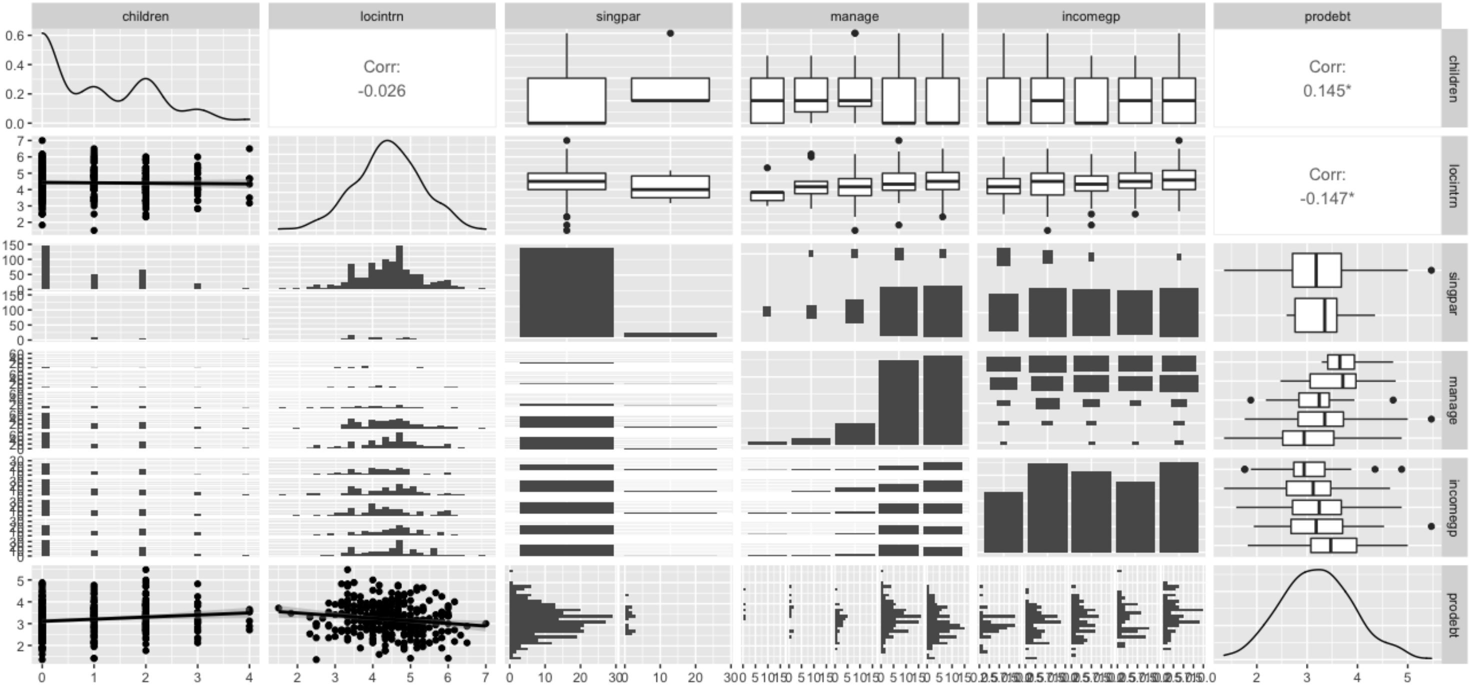
Variable	Description of the variable	Measurement type	role
incomegp	Income group (1=lowest, 5=highest) (ordinal variable)	factor	predictor
house	security of housing tenure (1=rent, 2=mortgage, 3=owned outright) (nominal variable)	factor	predictor
children	number of children in household	Continuous numeric	predictor
singpar	is the respondent a single parent? (nominal variable)	factor	predictor
agegp	age group (1=youngest) (ordinal variable)	factor	predictor
bankacc	does the respondent have a bank account? (nominal variable)	factor	predictor
bsocacc	does the respondent have a building society account? (nominal variable)	factor	predictor
manage	self-rating of money management skill (high values=high skill) (nominal variable)	factor	predictor
ccarduse	how often did s/he use credit cards (1=never… 3=regularly) (ordinal variable)	factor	predictor
cigbuy	does s/he buy cigarettes? (nominal variable)	factor	predictor
xmasbuy	does s/he buy Christmas presents for children? (nominal variable)	factor	predictor
locintrn	score on a locus of control scale (high values=internal)	Continuous numeric	predictor
prodebt	score on a scale of attitudes to debt (high values=favourable to debt	Continuous numeric	outcome

After using na.omit() function on the *debt* dataset I have observed that 160 observations got omitted. The original dataset had 464 observations and the cleaned dataset got created with 304 observations.

Analyses



Matrix 1



Matrix2

If we observe the lead diagonal Matrix 2, we see the univariant distribution of the variables the are

-Children seems to have 1 peak value and few number of higher and lower values, where most of the children seem to lie between 0-1 range, and we can see it's a very skewed distribution

-'Locintrn' seems to have one peak value one peak value and rest of the values are dispersed around it.

-'prodebt' has got one peak value and the rest of the values are dispersed around it.

Based on the correlation observations from Matrix 2, I can interpret the strength of association of variables with 'prodebt' are

-'Children' has a correlation coefficient of r= 0.145 which is very weak and positive

-'Locintrn' has a correlation coefficient of r= -0.147 which is very weak and negative

By looking at Matrix 2 I found 'incomegp' and manage to be interesting variables with relationship to outcome variable 'prodebt'. By looking at the graphs in the matrix the 'prodebt' is varying at different layers of the factor variables i.e., 'incomegp' from highest to lowest and manage from low skilled to high skilled in money management. But from the matrix we can only determine the pair-wise associations but not the combined predictive ability.

Now we need to see what the combined predictive ability is, so we fit a linear regression model to the dataset and evaluate its outputs with outcome variable as 'prodebt'.

Model 1:

The first variable that I am going to take here is locintrn,

prodebt (Model 1)= $X_0 + X_1$ (locintrn)

$X_0 = 3.7274$, $X_1 = -0.1197$

prodebt (Model 1)= 3.7274 -0.1197 *(locintrn)

The probability of observing an F-statistic value as large as 6.712 for the 1 predictor and the 302 observations by fluke is 0.01, which means the F- statistic found is credible and we can have confidence in the way in which the 'Model 1' has been estimated.

The adjusted R square value is 0.0185. The amount of variance in 'prodebt' that is shared by the combination is 1.85% variability. This means that 98.15% of the variability is not accounted for. So the majority of the variability in the outcome variable 'prodebt' is not captured by 'locintrn' alone, which means there are other variables that are better predictors which I have not included in my model.

As the p-value of the 'locintrn' is 0.01 which is <0.05 'locintrn' can be said statistically significant predictor variable for 'prodebt'.

The coefficient of the 'locintrn' is negative and for single unit value increase in 'locintrn' the 'prodebt' value drops by 0.1197.

Model 2:

As one variable isn't sufficient enough now we are going to introduce another predictor variable 'manage'

Prodebt(Model 2)= $X_0 + X_1$ (locintrn)+X_2 (manage2)+X_3 (manage3)+X_4(manage4) + X_5 (manage5)

X_0= 4.22082 X_1 = -0.10891

X_3= -0.62750 X_5= -0.70212

X_2= -0.16939 X_4= -0.40193

The probability of observing an F-statistic value as large as 5.113 for the 5 predictors and the 298 observations by fluke is 0.000627 which is nearer to zero, which means the F- statistic found is credible and we can have confidence in the way in which the 'Model 2' has been estimated.

The amount of variance in ‘prodebt’ that is shared by the combination is 6.356% variability. This means that 93.644% of the variability is not accounted for. So the majority of the variability in the outcome variable ‘prodebt’ is not captured by these 2 variables, But the variability got increased by 4.5% than the previous model which tells us that ‘manage’ is a significant predictor. So this is a better-fitted model than before. which means there are other variables that are better predictors which I have not included in my model.

The base value here is manage1. Except for manage5 with a p-value of 0.0286(<0.05) which is statistically significant no other category in manage is statistically significant because the p-values are >0.05. Moving from manage1 to manage5 the average value of ‘prodebt’ decreases by 0.70212 i.e. the mean values of manage1 and manage5 differs by -0.70212. The p-value of 'locintrn' is 0.0179 which is statistically significant as it is <0.05.

By keeping the remaining coefficients constant and for a unit increase of ‘locintrn’ the ‘prodebt’ decreases by 0.108. and by keeping all the other variables constant and for every increase of the unit value of manage5 the ‘prodebt’ decreases by 0.702.

The coefficient increased by the value of 0.0117 for ‘locintrn’ after the introduction of the 'manage' variable. That means that the slope has increased from -0.1197 to -0.108. which means for a unit change in ‘locintrn’ keeping manage as constant the ‘prodebt’ value increases by 0.0117 than the previous model.

Model 3:

As 2 variables aren’t sufficient enough now we are going to introduce another predictor variable ‘children’

Prodebt(Model 3)= $X_0 + X_1$ (locintrn)+X_2 (manage2)+X_3 (manage3)+X_4(manage4) + X_5 (manage5)+X_6(children)

X_0= 4.10936 X_1 = -0.10867

X_2= -0.16782 X_3= -0.64655

X_4= -0.37547 X_5= -0.67198

X_6= 0.09213

The probability of observing an F-statistic value as large as 5.374 for the 6 predictors and the 297 observations by fluke is 2.75e-05 which is nearer to zero, which means the F- statistic found is credible and we can have confidence in the way in which the 'Model 3' has been estimated.

The amount of variance in 'prodebt' that is shared by the combination is 7.971% variability. This means that 92.029% of the variability is not accounted for. So the majority of the variability in the outcome variable 'prodebt' is not captured by these 3 variables, But the variability got increased by 1.615% than the previous model which tells us that 'children' is a significant predictor. So this is a better-fitted model than before. which means there are other variables that are better predictors which I have not included in my model.

The base value here is manage1. Except for manage5 with a p-value of 0.0347(<0.05) which is statistically significant no other category in manage is statistically significant because the p-values are >0.05. Moving from manage1 to manage5 the average value of 'prodebt' decreases by 0.67198 i.e. the mean values of manage1 and manage5 differs by -0.70212. The p-value of 'locintrn' and 'children' are 0.0131 which is statistically significant as it is <0.05.

The changes in the coefficient of 'locintrn' and 'manage' from model 2 to model 3 are not that significant as the signs didn't change and the values of coefficients are pretty close to the previous model. It implies that the slope of the new model didn't change much compared to the previous model.

The coefficient of children implies that keeping the 'locintrn' and 'manage' constant for every unit increase in the value of children the 'prodebt' value increases by 0.09213.

Model 4:

As 3 variables aren't sufficient enough now we are going to introduce another predictor variable 'singpar'

Prodebt(Model 4)= X_0 + X_1 (locintrn)+X_2 (manage2)+X_3 (manage3)+X_4(manage4) + X_5 (manage5)+X_6(children)+X_7(singpar1)

X_0= 4.11192 X_1 = -0.10994

X_2= -0.15357 X_3= -0.63730

X_4= -0.36994 X_5= -0.66825

X_6= 0.09409 X_7 = -0.07511

The probability of observing an F-statistic value as large as 4.619 for the 7 predictors and the 296 observations by fluke is 6.368e-05 which is nearer to zero, which means the F- statistic found is credible and we can have confidence in the way in which the 'Model 4' has been estimated.

The amount of variance in 'prodebt' that is shared by the combination is 7.716% variability. This means that 92.284% of the variability is not accounted for. So the majority of the variability in the outcome variable 'prodebt' is not captured by these 4 variables, But the variability got decreased by 0.255% than the previous model which tells us that 'singpar1' is not a significant predictor. So this is not a better-fitted model than before. which means there are other variables that are better predictors which I have not included in my model.

The base value here is manage1. Except for manage5 with a p-value of 0.0360(<0.05) which is statistically significant no other category in manage is statistically significant because the p-values are >0.05.Moving from manage1 to manage5 the average value of 'prodebt' decreases by 0.66825 i.e. the mean values of manage1 and manage5 differs by -0.70212. The p-value of 'locintrn' and children are 0.0131 which is statistically significant as it is <0.05. the p-value of singpar1 is 0.6716 (>0.05) which means that it is not statistically significant variable.

By keeping other variables constant and by increasing the ‘singpar’ value by 1 unit the ‘prodebt’ value decreases by 0.07511 units. And this change is not significant as the variable ‘singpar’ is failing the model.

By running anova test on all the four models it is clear that Model 2 is a better fit model compared to all the 4 models put together. Because Model 2 has got the least p-value i.e. 0.001082 which is highly statistically significant and the Residual sum of squares (RSS) is 145 where as the Model 3 has got p-value 0.013 which is statistically significant but not as much as Model 2and RSS of 142 and the Model 4 has got p-value of 0.6715 which is not statistically significant.

Model 5:

As 4 variables aren’t sufficient enough to make an efficient model we are going to introduce another predictor variable ‘incomegp’. This variable is chosen because there are 5 layers to it from lowest to the highest where chance of association is more in one of the 5 layers. The mode of measurement is currency for both debt and the income group which I felt is a sign of strong association. so I felt the strength of association can be more when it comes to ‘prodebt’ and ‘incomegp’.

Prodebt(Model 5)= X_0 + X_1 (locintrn)+X_2 (manage2)+X_3 (manage3)+X_4(manage4) + X_5 (manage5)+X_6(children)+X_7(singpar1)+ X_8(incomegp2)+ X_9 (incomegp3)+X_{10} (incomegp4)+X_{11} (incomegp5)

X_0= 4.11192 X_1 = -0.10994 X_2= -0.15357 X_3= -0.63730

X_4= -0.36994 X_5= -0.66825 X_6= 0.09409 X_7 = -0.07511

X_8 = 0.06027 X_9= 0.21261 X_{10}= 0.23197 X_{11}= 0.50315

The probability of observing an F-statistic value as large as 4.886 for the 11 predictors and the 292 observations by fluke is 6.059e-07 which is nearer to zero, which means the F- statistic found is credible and we can have confidence in the way in which the ‘Model 4’ has been estimated.

The amount of variance in ‘prodebt’ that is shared by the combination is 12.36% variability. This means that 87.64% of the variability is not accounted for. So the majority of the variability in the outcome variable ‘prodebt’ is not captured by these 5 variables, But the variability got increased by 4.644% than the previous model which tells us that ‘incomegp’ is a significant predictor. So this is a better-fitted model than before. which means there are other variables that are better predictors which I have not included in my model.

The p-value’s of “locintrn” and children are 0.003937 and 0.023933 which are statistically significant as they are <0.05. The p-value of “singpar”, manage are >0.05 so they are not statistically significant and The base value for “incomegp” is incomegp1. Except for incomegp5 with a p-value of 0.000189(<0.05) which is statistically significant no other category in “incomegp” is statistically significant because the p-values are >0.05.Moving from incomegp1 to incomegp5 the average value of “prodebt” increases by 0.50315 i.e. the mean values of incomegp1 and incomegp5 differs by 0.50315.

By keeping other variables constant and increasing the “incomegp” by unit value the “prodebt” value increases by 0.50315. for every 2 person increase in incomegp5 there is a unit increase in the “prodebt” value.

By running anova test on all the five models it is clear that Model 2 and Model5 are both better fit models when all the 5 models put together. Because Model 2 and Model 5 has got the least p-value i.e. 0.0006671 and 0.0007063 which shows that they are statistically significant and their Residual sum of squares (RSS) is 145 and 131. Model 3 is significant but the p-value is higher than Model 2 and 5. Model 4 is not statistically significant as p-value > 0.05. hence we can take Model 5 as the best fir among all the above models as the p-value is nearer to zero and the RSS value is least compared to every other model.

Conclusions

One should have a high money management skill in order to be in the significant category. So learning money management is necessary.

Being in the high income group increases the chances of being favourable to debt.

Being single or married is has got no much affect on the favourable attitude towards debt.

The higher the frequency of the credit card usage the more the individual is inclined towards the favourable attitude towards debt.

The frequency of the usage of credit cards that is “ccarduse” is found to be one of the variables to predict an individual’s attitude towards debt. When I added the “ccarduse” in the multiple linear regression model the it has given a significant increase of 3.71% compared to the last model that is Model 5 taking the variability to 16.07%. Having a 3.71% rise in the variability means it is a useful predictor for the model. And also the more the usage of credit card the more the person is inclined towards having a debt. That is also one of the reasons I felt that it can be a useful predictor.