Regressing Analysis for Factors Affecting British Loan Approval

Yan Zhang[1, a], Wanqi Deng[2, b]

[1]Department of Business and Law, Coventry University, Coventry, UK

[2]Department of Business and Law, Coventry University, Coventry, UK

[a]13879038611@163.com, [b]15180454760@163.com

Abstract

This paper focuses on some possible factors: Gender, Level of education, Number of years of employment, Household income in thousands and Total debt in thousands that affect British householders' Loan Approval with multiple linear regression model in SPSS. By analysing the result of regression in SPSS, we can sort out which factors have comparatively stronger impact on the independent variable: Time needed to get the approval from the bank for British and which factor has comparatively less impact. Therefore, the model can be improved by omit the factor that is not significant in the regression model.

Keywords

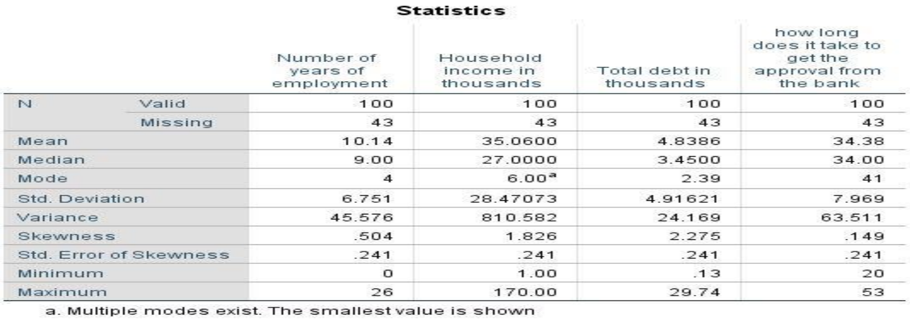
British householders' Loan Approval, Regression Analysis.

1. ITRODUCTION

Loans are a form of debt that can be provided by different establishments such as banks, building societies, credit unions and payday loan companies. Loans can vary from small amounts from one off payments to large purchases that are to be repaid over several years. There are two different types of loans: secured and unsecured. Secured loans are offset against an asset such as property, which is often used for large loans such as mortgages. Whereas, unsecured loans including personal loans, require no collateral. This paper only focus on some impossible factors: Gender, Level of education, Number of years of employment, Household income in thousands and Total debt in thousands that affect British householders' bank Loan Approval with multiple linear regression model in SPSS.

2. DESCRIPTIVE STATISTICS

This paper collect relevant statistics of 100 British householders. The followings are the descriptive statistics.



Statistics

		Number of years of employment	Household income in thousands	Total debt in thousands	how long does it take to get the approval from the bank
N	Valid	100	100	100	100
	Missing	43	43	43	43
Mean		10.14	35.0600	4.8386	34.38
Median		9.00	27.0000	3.4500	34.00
Mode		4	6.00[a]	2.39	41
Std. Deviation		6.751	28.47073	4.91621	7.969
Variance		45.576	810.582	24.169	63.511
Skewness		.504	1.826	2.275	.149
Std. Error of Skewness		.241	.241	.241	.241
Minimum		0	1.00	.13	20
Maximum		26	170.00	29.74	53

a. Multiple modes exist. The smallest value is shown

Fig 1. Descriptive statistics

Descriptive statistics is used to provide a summary of data, highlighting the key aspects and enabling visualisation of the data collected. In a large sample size, exceeding 30 observations, there is a lower possibility of error and outliers are more easily identifiable. It is vital from the data that outliers are taken into consideration, which would highlight an anomaly or an error in collection of the data.

In all of the variables, numbers of years of employment, household income, total debt and the time it takes for approval the distribution of skewness is positive.The average number of years individuals have been employed is 10.14 years. The data shows that there is a range of 26 years of employment in this sample. It is also visible from the data that the most frequent number of years an individual has been working when they are approved the loan is 4 years.

The household income of individuals in this sample has a significant range from £1,000 to £170,000. The sample shows a median value of £27,000 and a most frequently occuring value of £6,000. This suggesting that there is a higher proportion of those with a lower income requesting approval for loans. Which is backed up by the skewness which is positive.

The total debt of the individuals in the sample has a range of £29,610 with an average of £4,838.60, demonstrating the average debt of individuals is relatively low. Whereas the most popular amount of debt is considerable lower at £2,390.

The length of time for bank loans to be approved has an average of 34.38 days which does not vary significantly from the median at 34 days. There is a 7.969 day deviation from the mean and a variance of 63.511 days. The length of time for bank loans to be approved has a positive skewness.

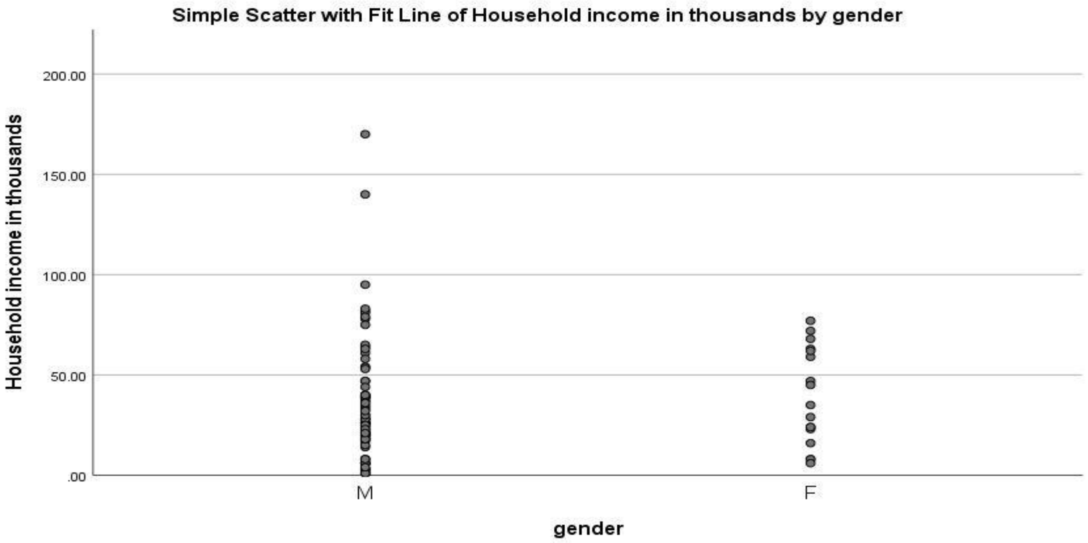


Fig 2. Relationship between gender and Household income (in thousands)

The graph shows that males' highest income is higher compared to females'. The range of income for males is £170,000 per annum and females significantly lower with a range of £80,000. It is visible that the distribution of pay is denser for males, whereas, females have a smaller variance of income.

Before analysing the correlation coefficient between number of years of employment and household income and do the regression, it is essential to make sure that both of these variables are nearly normal distribution.

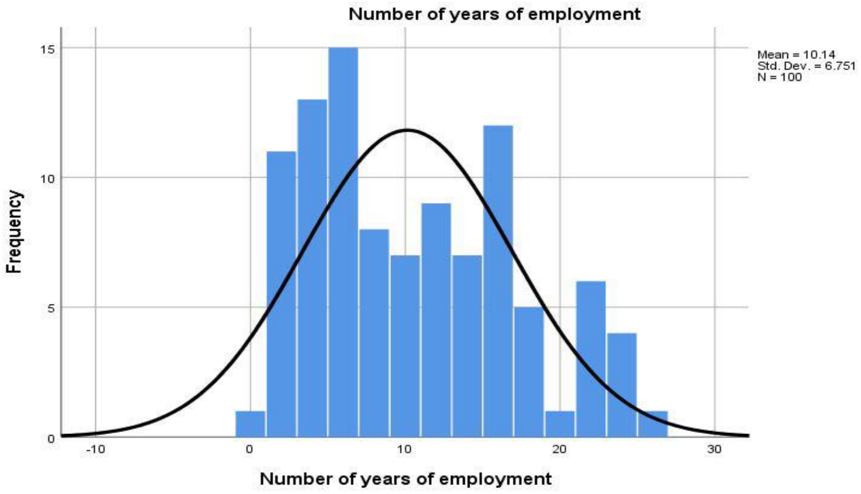


Fig 3. Distribution of number of years of employment

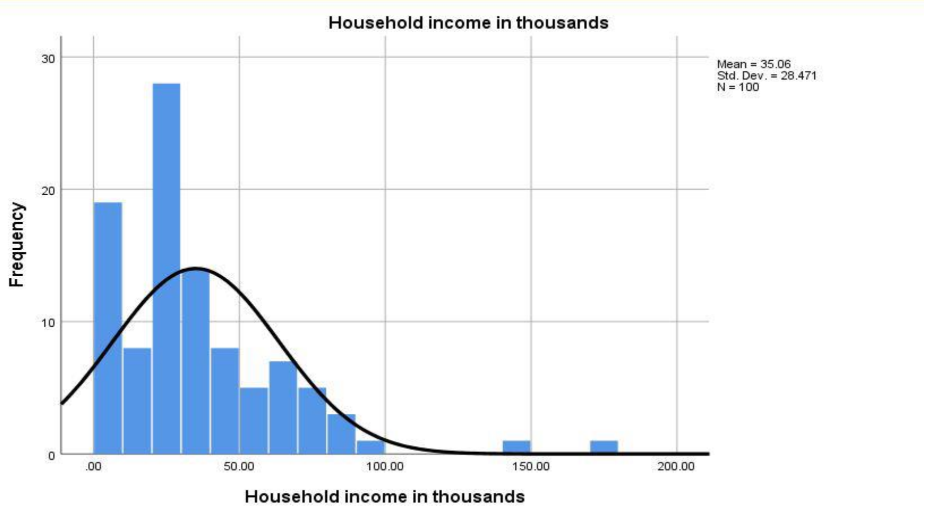
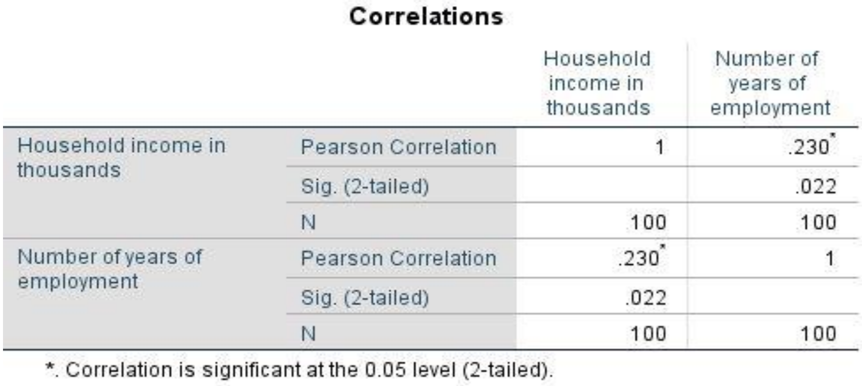


Fig 4. Distribution for household income (in thousands)

These two figures show that both number of years of employment and household income are nearly normal distribution. Therefore, the next step will be to undertake a correlation analysis.



Correlations

		Household income in thousands	Number of years of employment
Household income in thousands	Pearson Correlation	1	.230*
	Sig. (2-tailed)		.022
	N	100	100
Number of years of employment	Pearson Correlation	.230*	1
	Sig. (2-tailed)	.022	
	N	100	100

*. Correlation is significant at the 0.05 level (2-tailed).

Fig 5. Correlation between household income and number of years of employment

The correlation coefficient between household income and number of years of employment is 0.230 or 23%, therefore, an extra year of employment will result in an income increase of £230. The P-value is 0.022, which is smaller than 0.05, showing a positive, although weak, relationship between the number of years of employment and household income.

3. RERRESSION ANALYSIS

Multiple variables that can impact the approval of a loan from a bank requires the use of the multiple linear regression model, which analyses the factors significance.

The equation of multiple linear regression model is:

$Y= \beta0+ \beta1* + \beta2*X2 + \beta3*X3 + \beta4*X4+ \beta5*X5 + e1$

Y: How long does it take to get the approval from the bank

X1:Gender

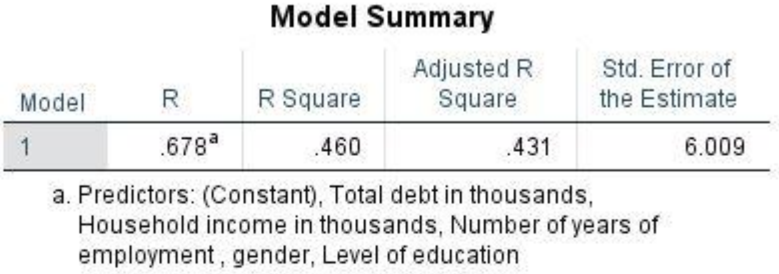
X2: Level of education

X3: Number of years of employment

X4: Household income in thousands

X5: Total debt in thousands

Regressing those variables in SPSS, outcomes are shown as following:



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.678[a]	.460	.431	6.009

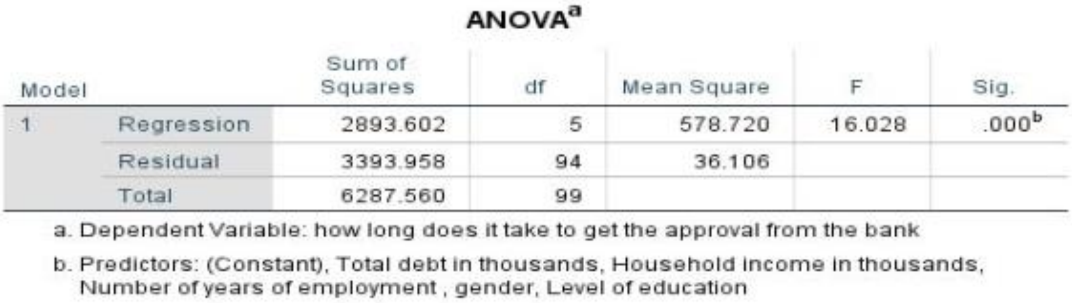
a. Predictors: (Constant), Total debt in thousands, Household income in thousands, Number of years of employment , gender, Level of education

Fig 6. Model Summary table

The second column of the output is R value: 0.678; the third column is R square: 0.460; the fourth column is adjusted R Square: 0.431. Adjusted R square equals to 1-((n-1)*(1-R^2))/ (n-k-1), n is the sample size, k is the number of independent variable. Adjusted R square will usually fall when adding new independent variables, while R square will not, which means R square will increase with the increase number of independent variables (Jeffrey M.Wooldridge, 2016).The last column is standard error of the estimate.

The most significant information in Model Summary Table is R square, which provides a measure of the goodness of fit for the whole regression model. This measures the proportion of the variation in the dependent variable that can be by the independent variables. R square lies between 0 and 1. The greater the R square is, the more powerful of the model.

In the regression model, R square is 0.460, which means that 46% of the variation in the dependent variable: how long does it take to get the approval from the bank can be explained by the variation of the independent variables: Gender, Level of education, number of years of employment, household income and total debt. This result indicates that this model is not significantly powerful but still able to explain some variation of independent variable with these selected independent variables.



ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2893.602	5	578.720	16.028	.000[b]
	Residual	3393.958	94	36.106		
	Total	6287.560	99			

a. Dependent Variable: how long does it take to get the approval from the bank

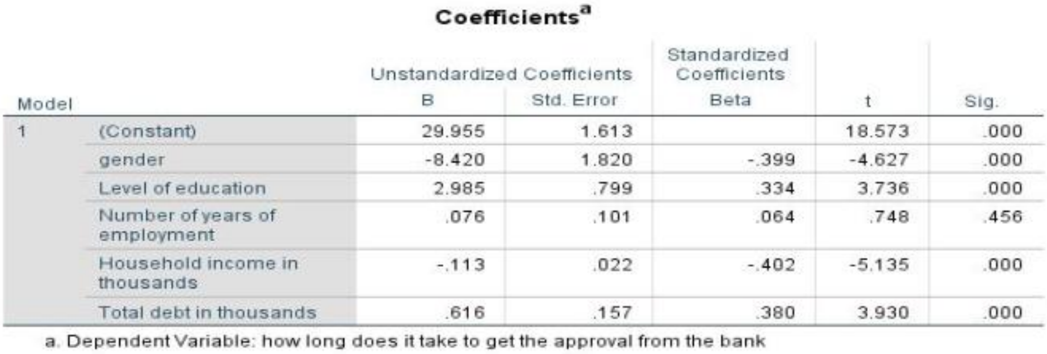
b. Predictors: (Constant), Total debt in thousands, Household income in thousands, Number of years of employment , gender, Level of education

Fig 7. ANOVA Table

This table shows the result of F-Test. F-test tests the overall significance of the regression equation based on analysis of variance.

The row 1, column 2 is SSE (Explained Sum of Squares); Row 2, column 2 is SSR (Residual Sum of Squares); Row 3, column 2 is SST (Total Sum of Squares). SSE+SSR=SST. The third column shows the degree of freedom of SSE, SSR, SST respectively. The fourth column indicates mean Square of SSE and SSR, but there is no statistic for SST.

Row 1, column 5 shows the F value. F = Mean Square of Regression/ Mean Square of Residual= 578.720/36.106=16.028. While the critical value is 2.2899 under the 5% significance level. 16.028>2.2899, therefore, the null hypothesis is rejected: $R^2 = 0$ and this model is significant and can explain the variation of dependent variable with selected independent variables. The same result can be obtained by comparing the P-value in the last column and significance level: 0.5% >0, which means this model is significant-the same result gained before.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	29.955	1.613		18.573	.000
	gender	-8.420	1.820	-.399	-4.627	.000
	Level of education	2.985	.799	.334	3.736	.000
	Number of years of employment	.076	.101	.064	.748	.456
	Household income in thousands	-.113	.022	-.402	-5.135	.000
	Total debt in thousands	.616	.157	.380	3.930	.000

a. Dependent Variable: how long does it take to get the approval from the bank

Fig 8. Table for coefficients

This table shows specific coefficients of multiple linearity regression model.

If the unstandardised coefficients are used, this model can be written as following:

Time taking to get the approval from the bank = 29.955 - 8.420*(Gender) + 2.985*(Level of education) + 0.076*(Number of years of employment)- 0.113*(Household income) + 0.616*(Total debt) + e1.

If the standardised coefficients are used, this model can be written as following:

Time taking to get the approval from the bank = -0.399*(Gender) - 0.334*(Level of education) + 0.064*(Number of years of employment) - 0.402*(Household income) + 0.380(Total debt) + e1.

Gender is a dummy variable, Gender=1: male; Gender=0: female.

The only difference between unstandardised coefficients and standardised coefficients is that standardised coefficients are in normalised units while unstandardised coefficients are in original units. In practice, unstandardised coefficients are used to interpret the regression model. The next step is interpreting the specific means of unstandardised coefficients.

β0 is 29.955, which means it need take 29.99 days to get the approval from the bank when do not consider all of these factors.

β1is -8.420, which means if other factors remain fixed, females on average will take another 8.42 days to get approval from the bank.

β2 is 2.985, which means if other factors remain fixed, when the education level of a people increase by one level, for example qualifying from an undergraduate to postgraduate. The length of time of approval from the bank will increase by 2.895 days.

β3 is 0.076, which means if other factors remain fixed, an extra year of employment will increase the approval time by another 0.076 days.

β4 is -0.133, which means if other factors remain fixed, when a household's income increases by one thousand pounds, the time taken for approval from the bank will increase by 0.133 days.

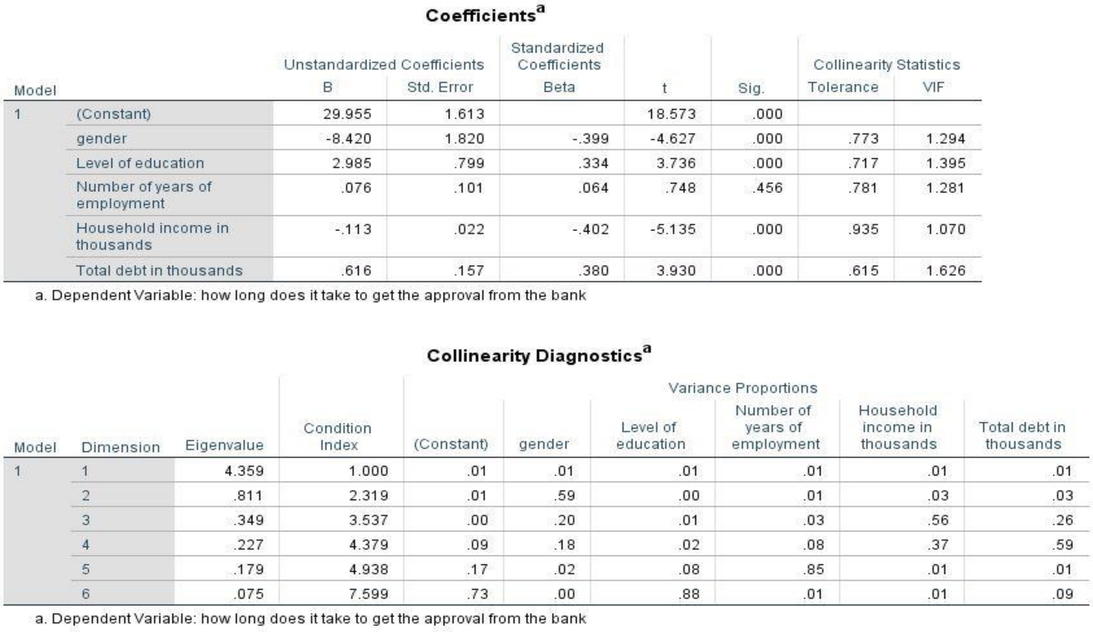
β5 is -0.616, which means if other factors remain fixed, when a people's total debt increases by one thousand pounds, it will take another 0.616 days to get get approval from the bank.

The fifth column show the T-value of these five variables respectively. T-value equals to β/Se(β) in this model. The last column shows the P-value of those five independent variables respectively. According to the hypothesis tests analysis above, except the variable: Number of years of employment, other independent variables are all significance, which means all of them can explain the variation of the dependent variable: Time taken to get the approval from the bank.

In conclusion, when comparing the critical T-value (F-value) with the calculated T-value(F-value) or significance level with P- value, it is identifiable whether an independent whether the model is significant (Anderson,Sweeney, Williams, Freeman, Shoesmith, 2017). According to regression analysis above, this model is significant, which means it can explain some variation of dependent variable with some selected independent variables, but it is a little unreliable because R square is only 0.460. According to the T-tests analysis above, all the independent variables are significant, however the regression analysis identifies that not all the selected independent variables can explain the variation of dependent variable. The independent variable: Number of years of employment cannot explain the variation of dependent variable: The time taking to get the approval from the bank. This model omits a significant that can explain the independent variable: The time taking to get the approval from the bank.

4. MULTICOLLINEARITY TEST

Multicollinearity describes the correlation among independent variables. The collinearity diagnostics measures the association between the independent variables. When there is a strong association existing between variables there will be implications on the model, as the β estimates will be difficult to distinguish from one another. There are two forms of multicollinearity: perfect and imperfect. Imperfect multicollinearity will be highly correlated where as perfect multicollinearity a movement of one or more variables will directly explain another independent variable.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	29.955	1.613		18.573	.000		
	gender	-8.420	1.820	-.399	-4.627	.000	.773	1.294
	Level of education	2.985	.799	.334	3.736	.000	.717	1.395
	Number of years of employment	.076	.101	.064	.748	.456	.781	1.281
	Household income in thousands	-.113	.022	-.402	-5.135	.000	.935	1.070
	Total debt in thousands	.616	.157	.380	3.930	.000	.615	1.626

a. Dependent Variable: how long does it take to get the approval from the bank

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions					
				(Constant)	gender	Level of education	Number of years of employment	Household income in thousands	Total debt in thousands
1	1	4.359	1.000	.01	.01	.01	.01	.01	.01
	2	.811	2.319	.01	.59	.00	.01	.03	.03
	3	.349	3.537	.00	.20	.01	.03	.56	.26
	4	.227	4.379	.09	.18	.02	.08	.37	.59
	5	.179	4.938	.17	.02	.08	.85	.01	.01
	6	.075	7.599	.73	.00	.88	.01	.01	.09

a. Dependent Variable: how long does it take to get the approval from the bank

Fig 9. Collinearity Diagnostics Table

Multicollinearity can be detected by high variance inflation factors (VIF), shown on the above coefficient table. In each case of the variables the variance inflation factors are below 5 indicating that there is no evidence if multicollinearity, therefore, the variables are not closely related.

Multicollinearity of data has consequences such as the estimates will be sensitive to changes, therefore, there may be a notable difference in the coefficients of the data if a seemingly insignificant variable is intentionally omitted. The estimates will also remain unbiased with an increase in the variables variances and standard errors; "The presence of multicollinearity alone does not lead to bias in estimating the parameters, and this is perhaps a major factor leading to complacency on the parts of investigators", (Willis and Perlack 2017). Another consequence that may be faced by multicollinearity of data is the T values will decrease, and the overall fit of the equation of non correlating variables will remain mainly uninfluenced.

5. CONCLUSION

In the process of bank loan approval, the bank requires a certain specification of information to assess an individual's eligibility for a loan. From analysing the data it can be concluded that the independent variables play a significant part within the amount of days it will take to receive an approved loan.

The descriptive statistics of the data from the sample highlights the mean approval days for a loan is 34.38 days, which can be affected by the variables within the model. The standard deviation has been calculated to identify the closeness to the mean, the data depicts that the standard deviation for each variable is relatively close to the mean value. The data also concludes that those with a lower income are more likely to be applying for a loan. Although, the sample is defined as large; as the sample size is greater than 30, the sample will be significantly more reliable given a considerably larger sample size. The hypothesis test identifies that the whole model is significant and all the independent variables are significant, but exclude the independent variable (the number of years of employment) in regression analysis. This suggests that the regression model may not be accurate. And it can be proved by analysis the R squared (0.460), representing a relatively poor line of best fit. Multicollinearity test of the data is concludes that there is no evidence of direct correlation between the variables, shown by the variance inflation factor. This is a advantageous result for the sample of data as it avoids potential consequences.

Approval for a bank loan is affected by the independent variables. It highlights that the individuals applying for a loan will experience marginally different approval waiting times depending on their length of employment, level of education, gender and current household income.

REFERENCES

[1] Wooldridge, J. (2016). Introductory econometrics. 6th ed. Singapore: Cengage learning.

[2] ANDERSON, D., Sweeney, D., Williams, T., Freeman, N. and Shoesmith, E. (2017). STATISTICS FOR BUSINESS & ECONOMICS. 4th ed. Australia: CENGAGE LEARNING.

[3] Willis, C.E. and Perlack, R.D. (2017) ‘Multicollinearity: Effects, Symptoms, and Remedies’. Journal of the Northeastern Agricultural Economics Council [online] 7 (01), 55–61. available from <https://ageconsearch.umn.edu/record/159045/files/Multicollinerarity.pdf> [2 April 2019].