National College of Ireland

Project Submission Sheet – 2018/2019

Sanket Dilip Dayama

Student Name: ……………………………………………………………………………………………………

X18143652

Student ID: ……………………………………………………………………………………………………

Msc Data Analytics 2019

Programme: ……………………………………………………………… **Year:** ………………………

Statistcs for Data Analytics (Cohort B)

Module: ……………………………………………………………………………………………………

Prof. Tony Delaney

Lecturer: ……………………………………………………………………………………………………

Submission Due Date: 7th January 2019

……………………………………………………………………………………………………

CA2 Statistics for Data Analytics

Project Title: ……………………………………………………………………………………………………

1,577

Word Count: ……………………………………………………………………………………………………

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: ………

Date: ………

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):	

Sanket Dilip Dayama

X18143652, Msc Data Analytics Cohort B

CA 2 Statistics

MULTIPLE LINEAR REGRESSION

Analysis on time required to start a business

Link of the data source :

1) http://data.un.org/Data.aspx?d=WDI&f=Indicator_Code%3aIC.REG.DURS
2) http://data.un.org/Data.aspx?d=WDI&f=Indicator_Code%3aIC.REG.PROC
3) http://data.un.org/Data.aspx?d=WDI&f=Indicator_Code%3aIC.REG.COST.PC.ZS

From these three links I have merged three datasets depending on the country. The data in the dataset consists of one year.

Introduction : Multiple regression is an extended version of simple linear regression. It is mostly used to predict the value of the dependent variable from the independent variables present in the data. The independent values are the ones which support to predict the value of dependent variable.

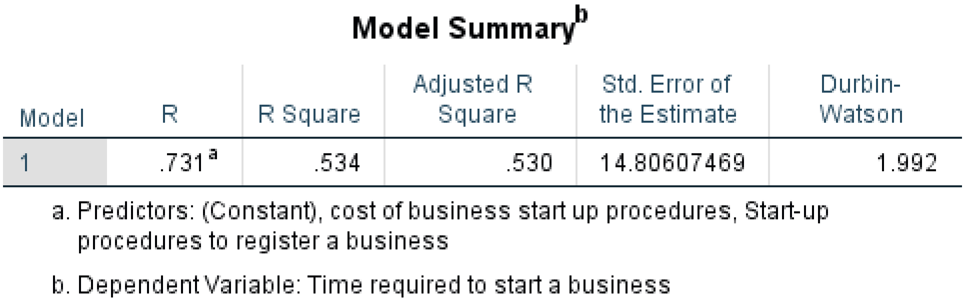
This project consists of three variables in which there is one dependent variable and two independent variable. The variables are listed below,

1) Time required to start a business (dependent)
2) Start up procedures to register a business (independent)
3) Cost of business start up procedures (independent)

Assumptions :

a) The dependent variable should be continuous. The dependent variable used in this project is “Time required to start a business” is the count of days which is continuous in nature.
b) The independent variables used to support the dependent variable should be continuous. Here the independent variables “Start up procedures to register a business” and “Cost of business start up procedures” are continuous in nature.

c) The observations should be independent. This can be checked by the Durbin-Watson test which should be in between 1.5 and 2.5. In this project the value is 1.992, according to the test it shows there is no auto correlation between the observations. The goodness of fit measure that is the R squared value gives the strength of the relationship between the model and the dependent variable. The R squared value for this model is 0.534 which is good level of prediction. The adjusted R square gives us the value if a new variable is taken into consideration in the model and the improvement seen for it. The value for adjusted R square is 0.530.



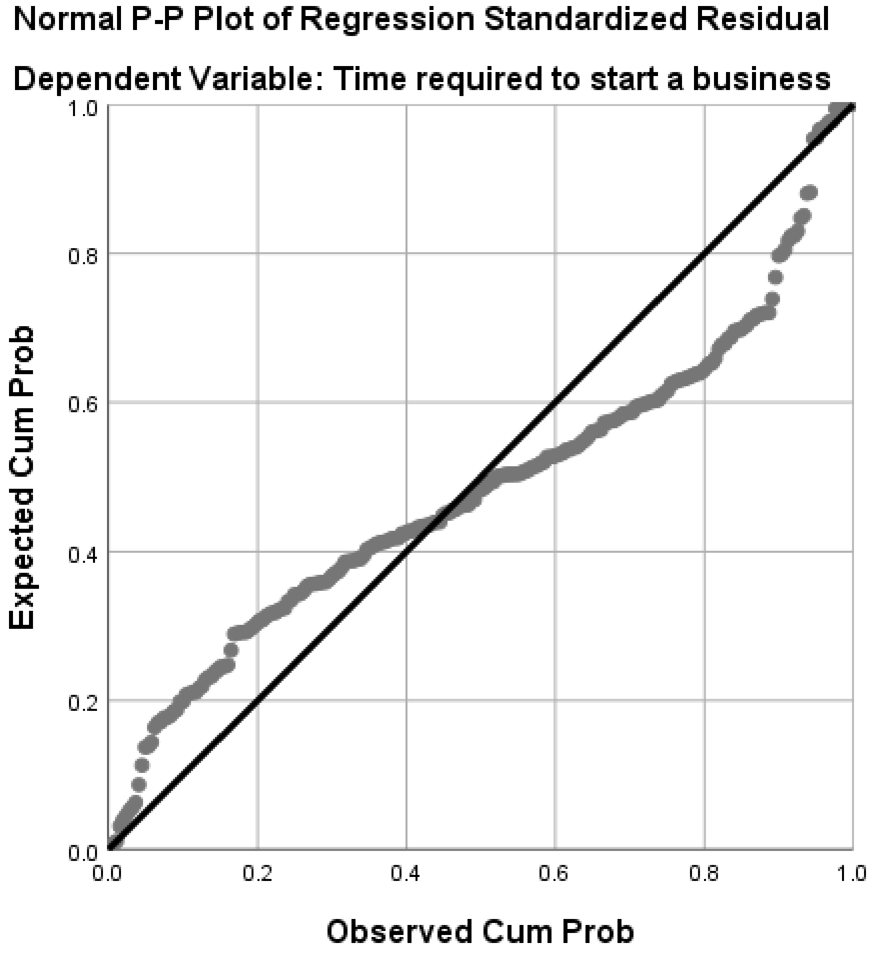
Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.731[a]	.534	.530	14.80607469	1.992

a. Predictors: (Constant), cost of business start up procedures, Start-up procedures to register a business

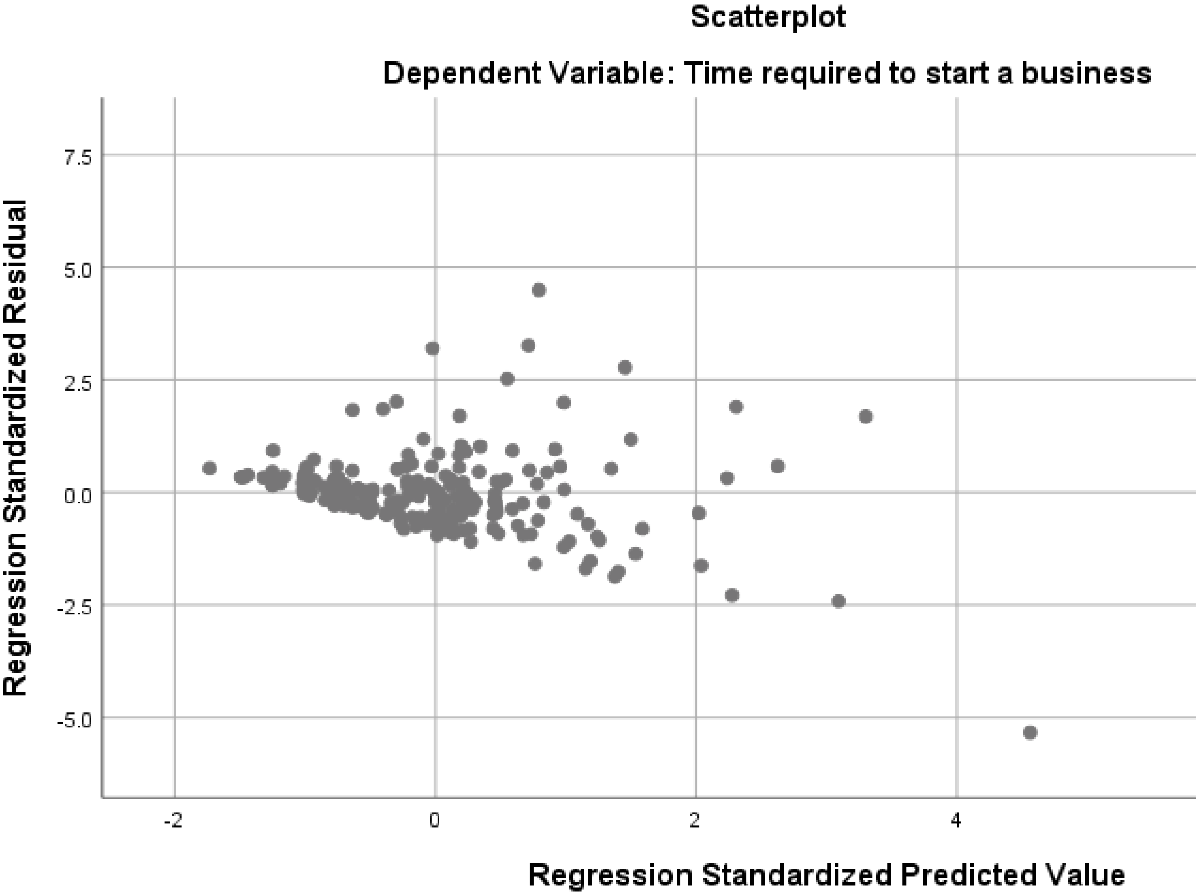
b. Dependent Variable: Time required to start a business

d) There is a linear relationship between the dependent and independent variables.

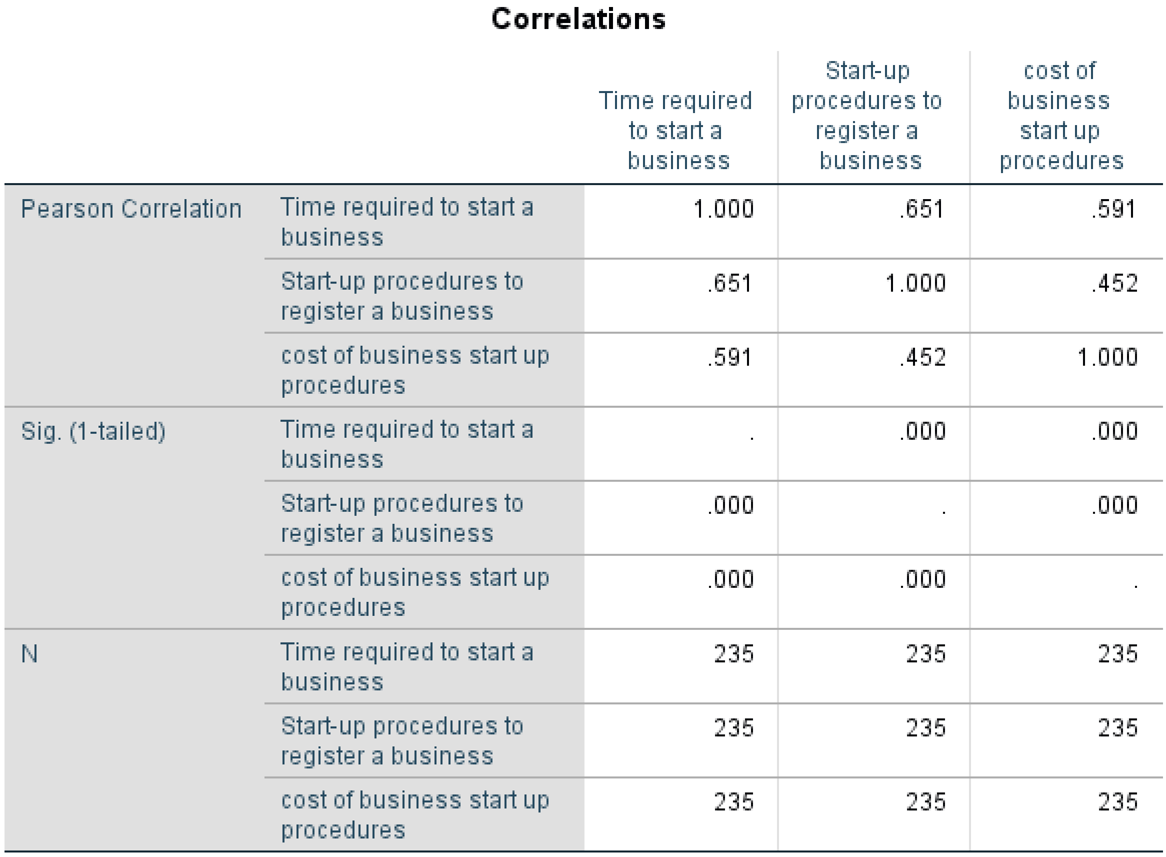


The reason we see this plot is the theoretical distribution has higher deviation than empirical distribution that is it tells us there are differences in the higher density data variable regions.

e) The homoscedasticity here can be explained from the scatter plot where the variances are unable to follow the pattern with the best fit line as we follow the line. The scatter plot is also used to show the effect of variable on one and the other one. Here we can see that the relation of variables show higher difference with each other at high value variables. The scatter plot also shows some outliers which are normally out of the range from -3 to 3. The cook's distance was taken into consideration to remove one of the outlier as the values larger than 1 create the outliers.(Pallant , Julie 2007).



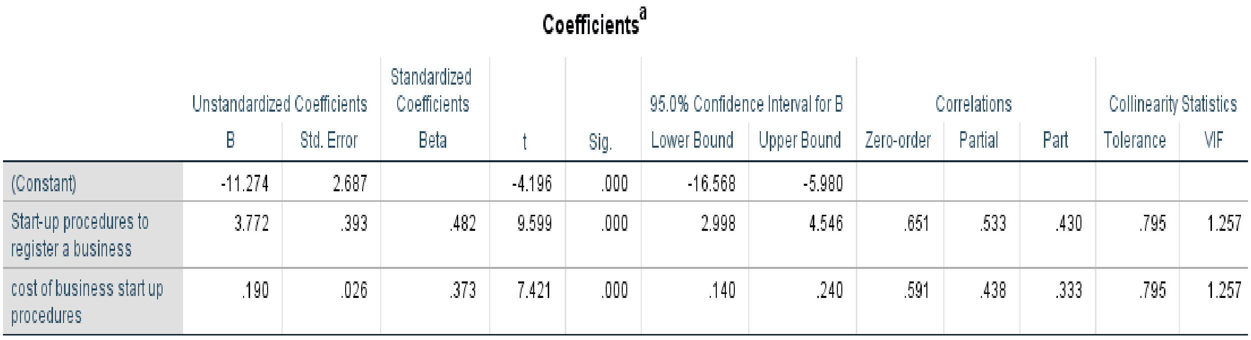
f) The data below does not show multi collinearity as there should be less relation between the independent variables as per the assumption for multiple regression. As seen in the correlations table we can see the independent variables which are "Startup procedures to register a business" and "cost of business start up procedures" show very less relation between the two as the threshold says that it should be less than 0.7. The Sig.(1-tailed) represents the significance level of the correlation which is 0. The N row shows the number of observations taken into consideration in the operation that is 235.



Correlations

		Time required to start a business	Start-up procedures to register a business	cost of business start up procedures
Pearson Correlation	Time required to start a business	1.000	.651	.591
	Start-up procedures to register a business	.651	1.000	.452
	cost of business start up procedures	.591	.452	1.000
Sig. (1-tailed)	Time required to start a business	.	.000	.000
	Start-up procedures to register a business	.000	.	.000
	cost of business start up procedures	.000	.000	.
N	Time required to start a business	235	235	235
	Start-up procedures to register a business	235	235	235
	cost of business start up procedures	235	235	235

In the table shown below we can inspect the tolerance and VIF values for the multi collinearity.

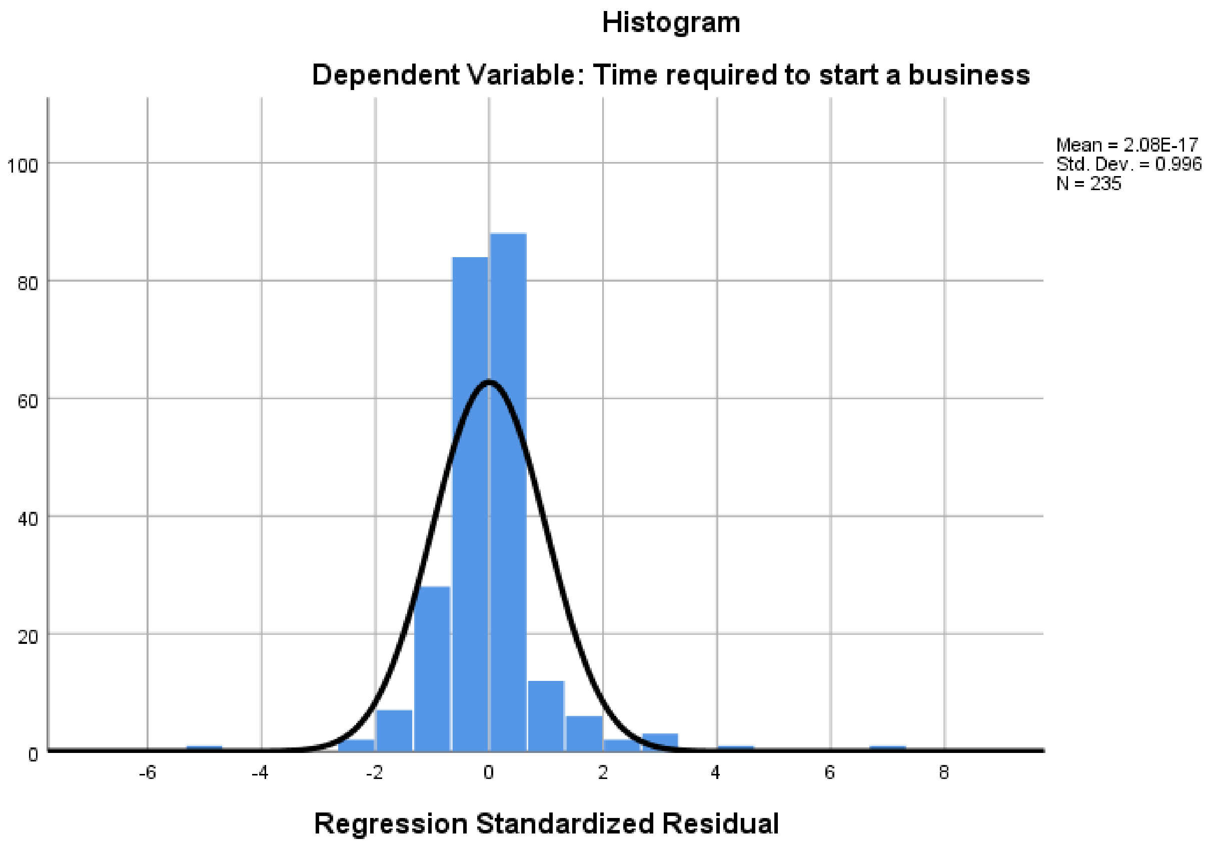


Coefficients[a]

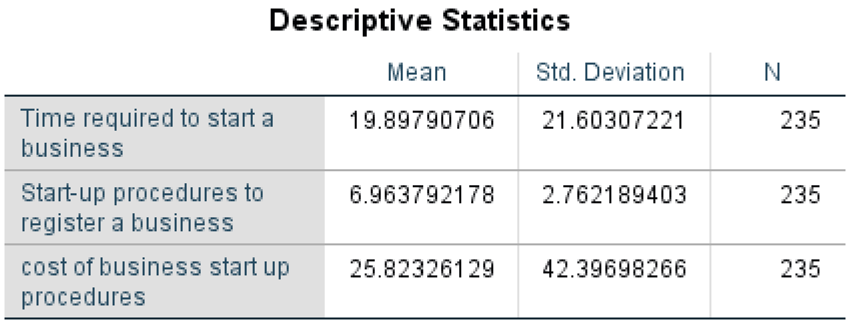
	Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
	B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
(Constant)	-11.274	2.687		-4.196	.000	-16.568	-5.980					
Start-up procedures to register a business	3.772	.393	.482	9.599	.000	2.998	4.546	.651	.533	.430	.795	1.257
cost of business start up procedures	.190	.026	.373	7.421	.000	.140	.240	.591	.438	.333	.795	1.257

The tolerance value of the model is 0.795 which is greater and should be as per the threshold which is greater than 0.1, whereas for VIF values the value should be less than 10 and the model consists of the value of 1.257 which gives us the assurance that there will not be any problem regarding the multi collinearity of the independent variables. It also shows the significance which should be lower than 0.05 to check the impact on dependent variable. The start-up procedures have more impact compared to cost of the business start up procedures on time required to start a business.

g) The histogram below shows that it is normally distributed for the “variable time required to start a business” which is our dependent variable, except for some outliers.



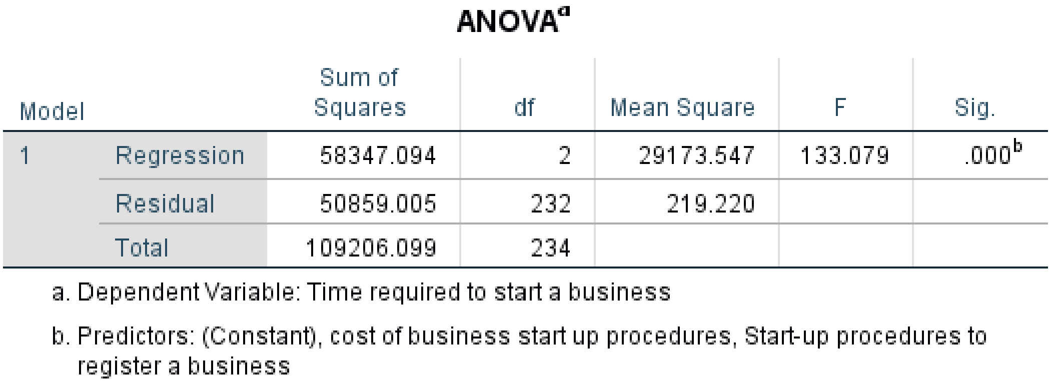
SPSS results :



Descriptive Statistics

	Mean	Std. Deviation	N
Time required to start a business	19.89790706	21.60307221	235
Start-up procedures to register a business	6.963792178	2.762189403	235
cost of business start up procedures	25.82326129	42.39698266	235

The table above shows that the model consists of 235 samples and the mean and standard deviation are calculated.



ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	58347.094	2	29173.547	133.079	.000^b
	Residual	50859.005	232	219.220		
	Total	109206.099	234			

a. Dependent Variable: Time required to start a business

b. Predictors: (Constant), cost of business start up procedures, Start-up procedures to register a business

The observations predicted were 58347.094 out of the total 109206.099 as the sum of squares columns depicts. This model predicts 2 out of 234 degrees of freedom.

Conclusion : The performed analysis gives us the regression equation as follows,

Time required to start a business = -11.274 + 3.772(Startup procedures to register a business) + 0.190(cost of business start up procedures)

Both the constants contribute in predicting the values for time required to start a business which is a dependent variable and we can see that the constant affects negatively on dependent variable.

LOGISTIC REGRESSION

Introduction : A logistic regression is performed when a prediction is to be done of dichotomous dependent variable based on the independent variables. The independent variables can be continuous or categorical. The data used for logistic regression is the same one which was used for multiple regression. The dependent variable was converted to a dichotomous variable.

The data conists of 3 columns :

- Startup success(Yes/No) (dependent)
  Description : It consists of a data that if the start up was successful or not. Here 0 represents the start up has failed to sustain and 1 represents the start up is a success.
- cost of business start up procedures. (independent)
  Description : The cost associated with the start up and its procedures.
- start up procedures to register a business. (independent)

Description : The time required to register the business.

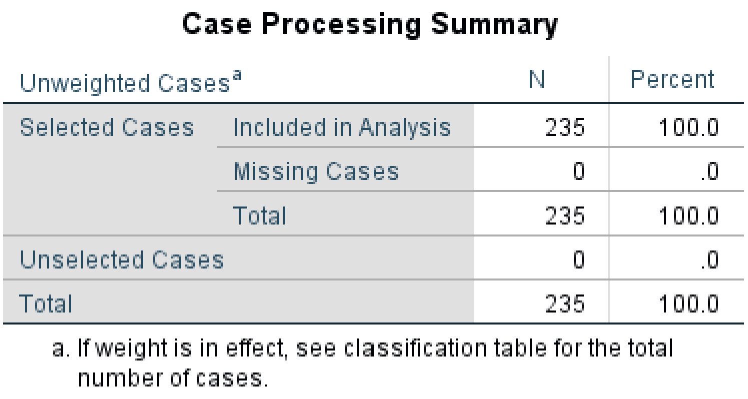
Assumptions : 1) The variable to be taken as dependent should be dichotomous, Startup Success is dichotomous.

2) There must be more than one independent variable. There are 2 independent variables in this model startup procedures to register a business and cost of business startupprocedures.

3) The model should consist of large sample size here its is of 235.

4) The multi collinearity should be less between the independent variables, that is the observations should not be dependent of each other.

SPSS results :



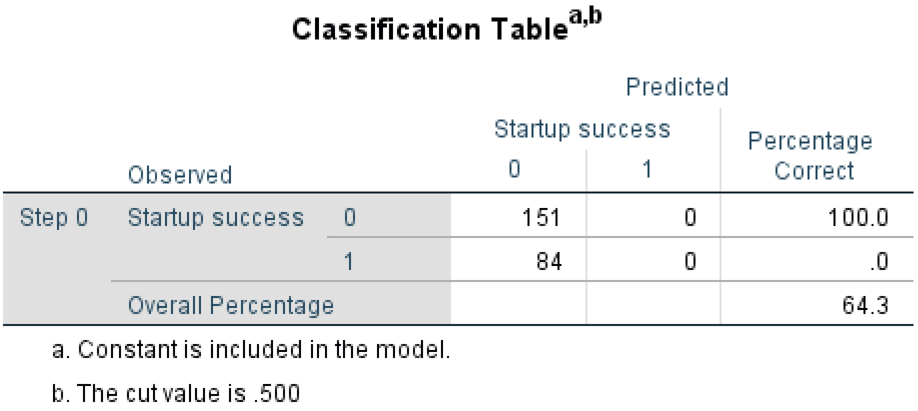
Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	235	100.0
	Missing Cases	0	.0
	Total	235	100.0
Unselected Cases		0	.0
Total		235	100.0

a. If weight is in effect, see classification table for the total number of cases.

The case processing Summary shows that there are total 235 samples and all of them have been processed.

The results show that there are two blocks in which the Block 0 runs the model without considering the independent variables.



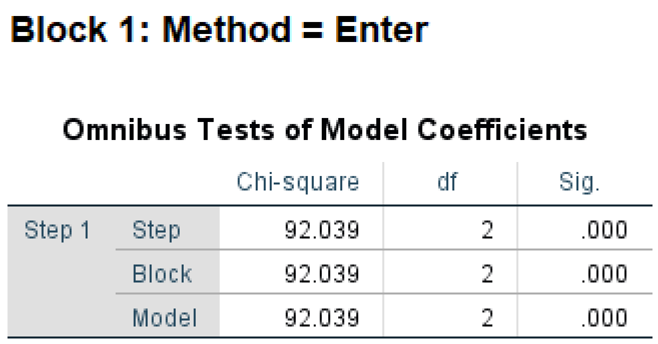
Classification Table[a,b]

			Predicted		
			Startup success		Percentage Correct
	Observed		0	1	
Step 0	Startup success	0	151	0	100.0
		1	84	0	.0
	Overall Percentage				64.3

a. Constant is included in the model.

b. The cut value is .500

The above table in block 0 depicts that the start ups fail to sustain in the market. It is because the independent variable is not taken into the consideration. Here the model predicts that 64.3 % of values are correct.

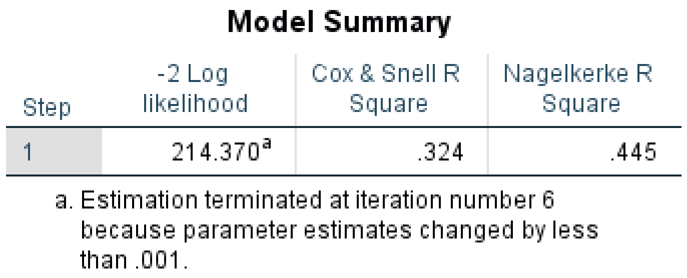


Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	92.039	2	.000
	Block	92.039	2	.000
	Model	92.039	2	.000

The above table in block 1 consists of Omnibus tests of model coefficients, the prediction variables are taken into consideration in the model. The outcomes of the block 0 are compared with block 1 for the goodness of fit test to inspect if there is an impact on dependent variable due to independent variables. The threshold for the significance value should be less than 0.05, here we can see that our table shows 0 where it can be said there is an impact on the dependent variable.

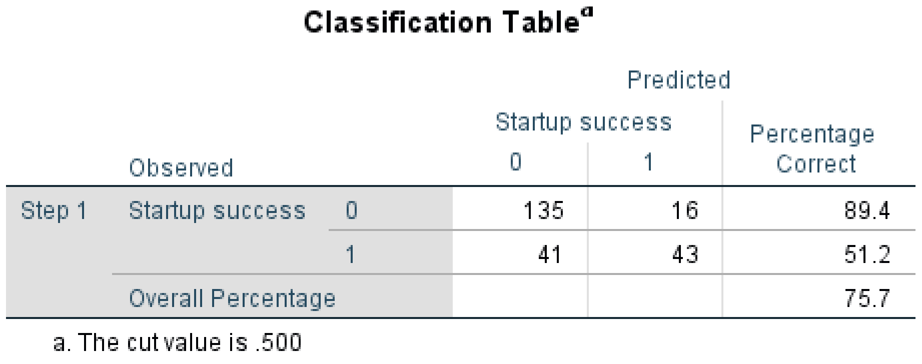


Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	214.370[a]	.324	.445

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

The above table shows us the variation in the dependent variable because of the model. It is revealed by two tests which are the Cox and Snell R Square and Nagelkerke R Square. There is a variation of 32.4 % to 44.5 % in the dependent variable under the influence of this model.

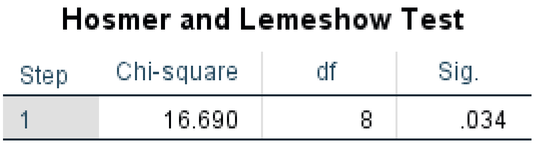


Classification Table[a]

Observed			Predicted		
			Startup success		Percentage Correct
			0	1	
Step 1	Startup success	0	135	16	89.4
		1	41	43	51.2
	Overall Percentage				75.7

a. The cut value is .500

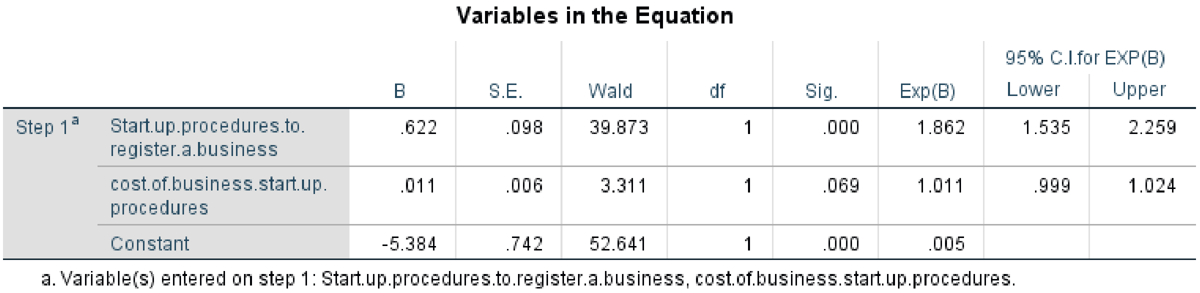
The classification table from block 1 shows us an improved prediction rate that is 75.7% than the block 0 prediction which was 64.3% . This is due to the influence of independent variables which are included in the model for processing.



Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	16.690	8	.034

The goodness of fit is determined by Hosmer and Lemeshow test. Referred from (Pallant , Julie) the significance value should be greater than 0.05 and our model has the significance value of 0.034. The value does not satisfy the criteria where it can be said it is poorly fit.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Start.up.procedures.to.register.a.business	.622	.098	39.873	1	.000	1.862	1.535	2.259
	cost.of.business.start.up.procedures	.011	.006	3.311	1	.069	1.011	.999	1.024
	Constant	-5.384	.742	52.641	1	.000	.005		

a. Variable(s) entered on step 1: Start.up.procedures.to.register.a.business, cost.of.business.start.up.procedures.

The above table depicts that what is the influence of independent variables and their contribution towards model prediction. The significance value should be less than 0.05 (referred from Pallant , Julie). The significance value for an independent variable startup procedures to register a business is 0 which shows that is a major contributor for dependent variable to predict the outcome than the other independent variable cost of business start up procedures which has the value 0.069. The Exp(B) gives us the values for the odds ratio for the variables present in the equation. The odds for startup procedures to register a business 1.862 is greater than the cost of business startup procedures for establishing a successful start up.

Result :

The performed analysis by using Binomial logistic regression gives us the equation,

Startup success = -5.384 + 0.622(Startup procedures to register a business) + 0.11(Cost of business start up procedures)

The probability can be achieved after substitution of independent variables in the equation. If the calculated probability is greater than 0.5 then we can say the startup was successful and established and if less than 0.5 then the start up was failed and could not sustain in the market.

References :

- Pallant, Julie. SPSS Survival Manual : a Step by Step Guide to Data Analysis Using SPSS.
- https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php
- https://statistics.laerd.com/spss-tutorials/binomial-logistic-regression-using-spss-statistics.php
- https://www.knowledgette.com/courses/144851/lectures/2148258