

# Statistics for Data Analytics

## Class Assessment 2

Ranu Kishor Parate

X17161452

National College of Ireland

MSc in Data Analytics – 2018/19

# **Multiple Regressions Model**

## **Introduction:**

Multiple regressions are the most common form of linear regression analysis. It refers to a set of techniques for studying the straight-line relationships among two or more variables. It involves a single dependent variable and two or more independent variables. It is used when we want to predict the value of a variable based on the value of two or more other variables.

## **Objectives of the analysis:**

The main purpose of this assessment is to study the selected dataset with the method of multiple regressions.

## **Data Set Source:**

The dataset which is used in this analysis is taken from website namely, European Union data (Link of website: <https://ec.europa.eu/eurostat>). The link from where the data is downloaded is: [http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=hsw\\_hp\\_nuse&lang=en](http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=hsw_hp_nuse&lang=en).

## **About the dataset:**

In this dataset it is given that the percentage value of work related problems by sex that is males and females in the year 1999. For multiple regressions we require one dependent variable here 1 dependent variable is “Values” and for independent variables we are taking 5 independent variables namely; Gender, HProb1, HProb2, HProb3, HProb4.

## **Description of Analysis:**

The analysis is carried out to study the significance of various independent variables on the dependent variable. To analyze the multiple regressions the software which we are using is SPSS.

## **Steps to carry out the analysis in SPSS:**

Follow the steps given below for analyzing multiple regressions in SPSS

Step 1: From the menu bar click on the ‘Analyze’ button. Select ‘Regression’ and under regression click on ‘Linear..’

Step 2: Select dependent variable and move it in Dependent box, in this analysis we are taking dependent variable as ‘Value’ and move independent variables in Independent box, here independent variables are ‘Gender’, ‘HProb1’, ‘HProb2’, ‘HProb3’, ‘HProb4’.

Step 3: Click on ‘Statistics’ button and select ‘Estimates’, ‘Model fit’, ‘R squared change’, ‘Descriptives’, ‘Collinearity diagnostics’ and ‘Durbin-Watson’ and click on ‘Continue’ button.

Step 4: Click on ‘Plots’. Move ‘ZRESID’ in to Y axis and ‘ZPRED’ in to X axis. Under Standardized Residuals Plots box, select ‘Histogram’ and ‘Normal probability plot’ and click on ‘Continue’ button.

Step 5: Click on ‘OK’

**Hypothesis:**

H<sub>0</sub>: The null hypothesis will be that the value of different independent variables affects the dependent variable.

H<sub>1</sub>: The alternate hypothesis will be that the value of different independent variables does not affect the dependent variable.

**Assumptions:**

1. The independent variable must be continuous and dichotomous.
2. The dependent variable must be continuous scaled variable.
3. We assume that there is linear relationship between the dependent and independent variable.

**SPSS interpretation of the output:**

- The predicted variable here is Value that is dependent variable.
- The predictor variables are Gender, HProb1, HProb2, HProb3 and HProb4 that is independent variable.

**Descriptive Statistics:**

Descriptive Statistics			
	Mean	Std. Deviation	N
Value	10.6582	18.54395	122
Gender	.50	.502	122
HProb1	.21	.411	122
HProb2	.20	.399	122
HProb3	.20	.399	122
HProb4	.20	.399	122

A descriptive statistic is a summary statistic that quantitatively describes or summarizes features of a collection of information. The descriptive statistics provide a general overview of the data set. In this dataset there total 122 numbers of variables so the N in the descriptive statistics table represents the total numbers of variables. The average of all the variables according to the dependent and independent variables is represented by the Mean in the descriptive statistics table. This table also calculates the standard deviation of the variables from the calculated mean.

## Correlations:

		Correlations					
		Value	Gender	HProb1	HProb2	HProb3	HProb4
Pearson Correlation	Value	1.000	.080	.945	-.183	-.258	-.277
	Gender	.080	1.000	.000	.000	.000	.000
	HProb1	.945	.000	1.000	-.258	-.258	-.258
	HProb2	-.183	.000	-.258	1.000	-.245	-.245
	HProb3	-.258	.000	-.258	-.245	1.000	-.245
	HProb4	-.277	.000	-.258	-.245	-.245	1.000
Sig. (1-tailed)	Value	.	.189	.000	.022	.002	.001
	Gender	.189	.	.500	.500	.500	.500
	HProb1	.000	.500	.	.002	.002	.002
	HProb2	.022	.500	.002	.	.003	.003
	HProb3	.002	.500	.002	.003	.	.003
	HProb4	.001	.500	.002	.003	.003	.
N	Value	122	122	122	122	122	122
	Gender	122	122	122	122	122	122
	HProb1	122	122	122	122	122	122
	HProb2	122	122	122	122	122	122
	HProb3	122	122	122	122	122	122
	HProb4	122	122	122	122	122	122

The relationship or the correlations between the variables is shown in the above correlation table. In this table the relation between independent variable and dependent variables are shown. That is relation between Value which is dependent and Gender, HProb1, HProb2, HProb3, HProb4 which are independent are shown. N is the total number of variables.

## Variables Entered/Removed:

Variables Entered/Removed <sup>a</sup>			
Model	Variables Entered	Variables Removed	Method
1	HProb4, Gender, HProb2, HProb3, HProb1 <sup>b</sup>	.	Enter

a. Dependent Variable: Value

b. All requested variables entered.

Variables entered/ removed table tells us what all variables are fed in the model.

**Model:** It shows multiple Models in a single regression command. Here the model is 1 that means only 1 model is being reported.

**Variables Entered:** It shows what all variables are entered into regression in a block. In this analysis the variables we have entered are Gender, HProb1, HProb2, HProb3 and HProb4.

**Method:** It shows the method which is used by SPSS to analyze the regression. In this analysis the method which is used is “FORCED ENTRY METHOD”. The standard method of entry is simultaneous; all independent variables are entered into the equation at the same time.

### Model Summary:

Model Summary <sup>b</sup>										
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	Change Statistics			Sig. F Change	Durbin-Watson
						F Change	df1	df2		
1	.950 <sup>a</sup>	.903	.899	5.89393	.903	216.358	5	116	.000	3.122

a. Predictors: (Constant), HProb4, Gender, HProb2, HProb3, HProb1

b. Dependent Variable: Value

Model summary summarized and tells us how suitable the model is. The calculated values of all the elements like R, RSquare, Standard deviation error, etc. tells us how suitable the model is.

**R:** The R value in the model summary signifies the simple correlation. It is the correlation between the observed values of the outcome, and the values predicted by the model. It is multiple regression Coefficient value which is (.950) that is 95%. It shows the relationship between the detected and expected values of dependent variable. So the 95% score is above average.

**R Square:** R Square is the proportion of variance accounted for by the model. It calculates how much of the total variation in the dependent variable can be explained by the independent variables. R Square is the coefficient of determination. In this case the R Square is 0.903. In this analysis the independent variable Gender, HProb1, HProb2, HProb3 and HProb4 variability of 90.3% on the dependent variable that is Value.

**Adjusted R Square:** It is the estimation of R Square in the population. It give us idea how well our model is generalized. As the R Square is very large due to the large number of independent variables the adjusted R square is used to calculate this effect. From 0.903 to 0.899

This Model Summary calculates the output of Durbin Watson Test. This test informs us about whether the assumption of independent errors is tenable. The calculated value for Durbin Watson Test is 3.122.

### Coefficients:

Coefficients <sup>a</sup>								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-.330	1.316		-.251	.802		
	Gender	2.969	1.067	.080	2.782	.006	1.000	1.000
	HProb1	43.030	1.668	.954	25.792	.000	.610	1.639
	HProb2	2.671	1.701	.057	1.570	.119	.622	1.607
	HProb3	-.138	1.701	-.003	-.081	.936	.622	1.607
	HProb4	-.838	1.701	-.018	-.492	.623	.622	1.607

a. Dependent Variable: Value

Model column in coefficient table shows that the predictor variables that is Constant, Gender, HProb1, HProb2, HProb3 and HProb4. The first variable in the column is constant also referred as a Y intercept. The b-values tell us to what degree each predictor affects the outcome if the effects of all the predictors are held constant.

The B column in the coefficient table is the values for the regression equation for predicting the dependent variable from the independent variable.

A standardized coefficient value is expressed in standard deviations.

The regression equation is given by:

$$y=b_0+b_1X_1+b_2X_2+.....+b_nX_n$$

The column B under unstandardized coefficient gives the value of  $b_0$ ,  $b_1$ ,  $b_2$ ,  $b_3$ ,  $b_4$ ,  $b_5$ .

The coefficient for Gender is 2.969. So for every unit increase in Gender, a 2.97 unit increase in Value is predicted, holding all other variables constant.

The coefficient for HProb1 is 43.030. So for every unit increase in HProb1, 43 unit increases in Value is predicted, holding all other variables constant.

The coefficient for HProb2 is 2.671. So for every unit increase in HProb2, a 2.67 unit increase in Value is predicted, holding all other variables constant.

For every unit increase in HProb3, we expect a -0.138 unit decrease in the Value, holding all other variables constant.

For every unit increase in HProb4, we expect a -0.838 unit decrease in the Value, holding all other variables constant.

Gender x (2.969) + HProb1 x (43.030) + HProb2 x (2.671) + HProb3 x (-0.138) + HProb4 x (-0.838) will give the predicted value of the dependent variable (Value).

t and Sig. – These are the t-statistics and their associated 2-tailed p-values used in testing whether a given coefficient is significantly different from zero. Using an alpha of 0.05:

The significance of each independent variable is:

The significance value can be check by comparing the p-value and alpha. Here we will take alpha = 0.05. If p-value < 0.05 then the independent variable must be significant.

- For Gender,  $0.006 < 0.05$ , this is significant.
- For HProb1,  $0.000 < 0.05$ , this is significant.
- For HProb2,  $0.119 > 0.05$ , this is not significant.
- For HProb3,  $0.936 > 0.05$ , this is not significant.
- For HProb4,  $0.623 > 0.05$ , this is not significant.

## ANOVA:

ANOVA <sup>a</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	37579.581	5	7515.916	216.358	.000 <sup>b</sup>
	Residual	4029.656	116	34.738		
	Total	41609.237	121			

a. Dependent Variable: Value

b. Predictors: (Constant), HProb4, Gender, HProb2, HProb3, HProb1

The Total variance is partitioned into the variance which can be explained by the independent variables (Model) and the variance which is not explained by the independent variables (Error). These are categories as Regression, Residual and Total.

Sum of Squares which is associated with the three sources of variance: Total, Residual and Regression.

df is the Degree of Freedom associated with the three sources of variance. For total variance the df is calculated as N-1 where N is the total number of variance.

Mean Square this is calculated as the sum of squares divides by its respective degree of freedom.

F value is calculated as Mean Square of Regression divided by Mean Square of Residual. Here Mean Square of Regression = 7515.916 and Mean Square of Residual = 34.738. Therefore,

$$F = \frac{7515.916}{34.738} = 216.358$$

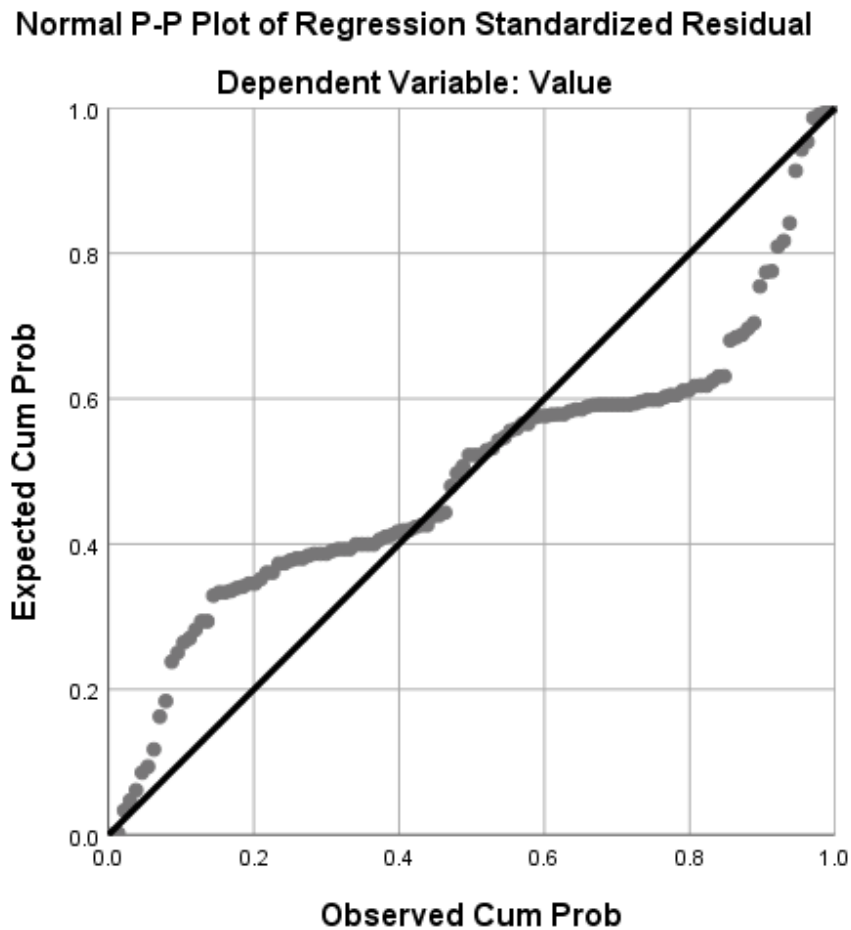
Significance value = 0.000 < 0.05

Hence, we can deduce that predictors that are independent variables are significant in predicting the value which is the dependent variable.

### Charts:

p-p plot:

p-p plot is the Probability-plot.





The p-p plot displays us that the model is not linear and there are many outliers lying outside the linear line. As it is clear from the graph that most of the points are outliers that is most of the points are not in straight line which indicate that the normality assumption is not met.

**Conclusion:**

The multiple linear regression method was conducted to analyze relationship between the Value which is the Dependent variable and “Gender”, “HProb1”, “HProb2”, “HProb3” which are Independent variable. There were many outliers detected and the variables were not linearly distributed.

The analysis was conducted on 5 independent variables out of which 2 independent variable (Gender and HProb1) are significantly contributing towards the prediction of Value.

The other 3 variables (HProb2, Hprob3 and HProb4) are not significant in predicting the value which is dependent variable.

## **Logistic Regression Model (Binary)**

### **Introduction:**

Logistic regression allows testing model to predict categorical outcomes with two or more categories. The independent variable can be either categorical or continuous, or mix of both in one model. It is the form of binomial regression.

### **Objectives of the analysis:**

The main purpose of this assessment is to study the selected dataset with the method of binary logistic regressions.

### **Data Set Source:**

The dataset which is used in this analysis is taken from website namely, European Union data (Link of website: <https://ec.europa.eu/eurostat>). The link from where the data is downloaded is: [http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=hlth\\_egis\\_am7e&lang=en](http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=hlth_egis_am7e&lang=en).

### **About the dataset:**

In this dataset it is given that the percentage value of use of homecare services by sex that is males and females, age and educational attainment level in the year 2014. For binary logistic regression we require dependent variable which should be dichotomous that is it should have two different possible values here dependent variable is “Sex” whose values are Males and Females. And the predictors are required that is independent variables this can be either categorical or continuous or a mix of both in the one model; here independent variables are “EducationAttainmentLevel1”, “EducationAttainmentLevel2” and “Value”, Where “EducationAttainmentLevel1” and “EducationAttainmentLevel2” are categorical covariates and “Value” is covariates.

### **Description of Analysis:**

The analysis is carried out to study the significance of various predictors or independent variables on the dependent variable. To analyze the binary logistic regressions using software name SPSS.

### **Steps to carry out the analysis in SPSS:**

Follow the steps given below for analyzing binary logistic regressions in SPSS

Step 1: From the menu bar click on the ‘Analyze’ button. Select ‘Regression’ and under regression click on ‘Binary Logistic..’

Step 2: Select dependent variable and move it in Dependent box, in this analysis we are taking dependent variable as ‘Sex’ which is dichotomous and move independent variables that is predictors in Block 1 of 1 box, here independent variables are ‘Value’, ‘EducationAttainmentLevel1’ and ‘EducationAttainmentLevel2’.

Under 'Method', Enter should be selected.

Step 3: Click on 'Categorical' button move the categorical independent variables (EducationAttainmentLevel1 and EducationAttainmentLevel2) in the categorical covariates box. Also select the categorical variables one by one and under change contrast click on radio button name 'First' under Reference Category, then click on 'Change' and click on 'Continue' button.

Step 4: Click on 'Options' button. Under 'Statistics and Plots' check or select 'Classification plots', 'Hosmer-Lemeshow goodness-of-fit', 'Casewise listing of residuals' and 'CI for exp(B)'.

Step 5: Click on 'Continue' and then on 'OK'.

### Assumptions:

1. The dependent variable must be dichotomous. In this analysis the Sex is used which is having 2 variables as Males and Females.
2. The independent variable can be either categorical or continuous or a mix of both in the one model.
3. Multicollinearity should not present between the independent variables.

### Hypothesis:


H<sub>0</sub>: The null hypothesis will be that the dependent variable (Males and Females) gets affected by the predictors or independent variables.

H<sub>1</sub>: The alternate hypothesis will be that dependent variable does not get affected by the predictors or independent variables.

### SPSS interpretation of the output:

- The predicted variable here is Sex that is dependent variable.
- The predictor variables are Value, EducationAttainmentLevel1 and EducationAttainmentLevel2 that is independent variable.

### Dependent Variable Encoding:



Original Value	Internal Value
Males	0
Females	1

Here to analyze the Educational Attainment Level under binary logistic regression the Sex has been encoded to 0 as Males and 1 as Females.

### Variables in the Equation:

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.007	.066	.010	1	.921	1.007

The logistic regression model uses odds ratio, which is defined as the probability of an event of interest compared with the probability of not having an event of interest.

$$\text{Odd Ratio} = \frac{\text{probability of an event of interest}}{1 - \text{probability of an event of interest}}$$

Null model  $B_0 \ln(p/p-1) = 0.007$

### Classification Table:

Classification Table <sup>a,b</sup>					
Observed			Predicted		Percentage Correct
			SEX		
Step 0	SEX	Males	0	455	.0
		Females	0	458	100.0
	Overall Percentage				50.2

a. Constant is included in the model.

b. The cut value is .500

Classification table tells us the overall percentage of the model. Here the overall percentage of the model for the analysis of this data is 50.2%. This means, that the prediction is 50.2% true for the selected model. Here the prediction of Males is 455 and for Female id 458. This means that the predictions of Males are nearly equal to the Females.

## Omnibus Tests of Model of Coefficients:

### Block 1: Method = Enter

**Omnibus Tests of Model Coefficients**

		Chi-square	df	Sig.
Step 1	Step	15.669	2	.000
	Block	15.669	2	.000
	Model	15.669	2	.000

The Omnibus Test of Model gives us an overall indications of how well the model performs, over and Block 0. This is referred as 'goodness of fit' test. In this the significant value should be less than 0.05. Here in this case, the significant value is 0.000 which is less than 0.05 ( $0.000 < 0.05$ ). Therefore, this model with whatever the predictor values are set is better than SPSS's original shown Block 0 model. That is the model is good fit for the binary logistic regression. The value of Chi-square is 15.669 at 2 degree of freedom.

## Classification Table:

**Classification Table<sup>a</sup>**

			Predicted		Percentage Correct
		Observed	SEX Males	Females	
Step 1	SEX	Males	319	136	70.1
		Females	284	174	38.0
	Overall Percentage				54.0

a. The cut value is .500

The Classification table shows, how well the model is able to predict the correct category for each case. Here the overall percentage of the model for analysis is 54%. When compared with the Block 0, the overall percentage of model in block 0 is 50.2%. The block 1 overall percentage varies with the block 0 overall percentages. The classification table also tells us the individual percentage.

### Hosmer and Lemeshow Test:

#### Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	7.523	8	.481

The poor fit is indicated by Hosmer and Lemeshow test; it indicates the significant value is less than .05. As in this model the significant value is 0.481 in Hosmer Lemeshow test that means, it is greater than 0.05. Therefore the model is good fit.

### Model Summary:

#### Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1250.008 <sup>a</sup>	.017	.023

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

The value of Cox and Snell R Square and Nagelkerke R Square indicates the amount of variation in the dependent variables which is predicted by the predictor variables collectively.

The maximum value is 1 if relationship is perfect and minimum value is 0 if there is no relationship. The values of Cox and Snell R Square and Nagelkerke R Square are pseudo R<sup>2</sup> values. The value of Cox and Snell R Square is 1.7% and the value of Nagelkerke R Square is 2.3%. Nagelkerke R Square value is greater than Cox and Snell R Square.

### Variables in the Equation:

#### Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
Step 1 <sup>a</sup>	Value	.062	.016	14.240	1	.000	1.064	1.030	1.098
	EducationalAttainmentLevel1(1)	-.126	.144	.760	1	.383	.882	.665	1.170
	Constant	-.145	.091	2.528	1	.112	.865		

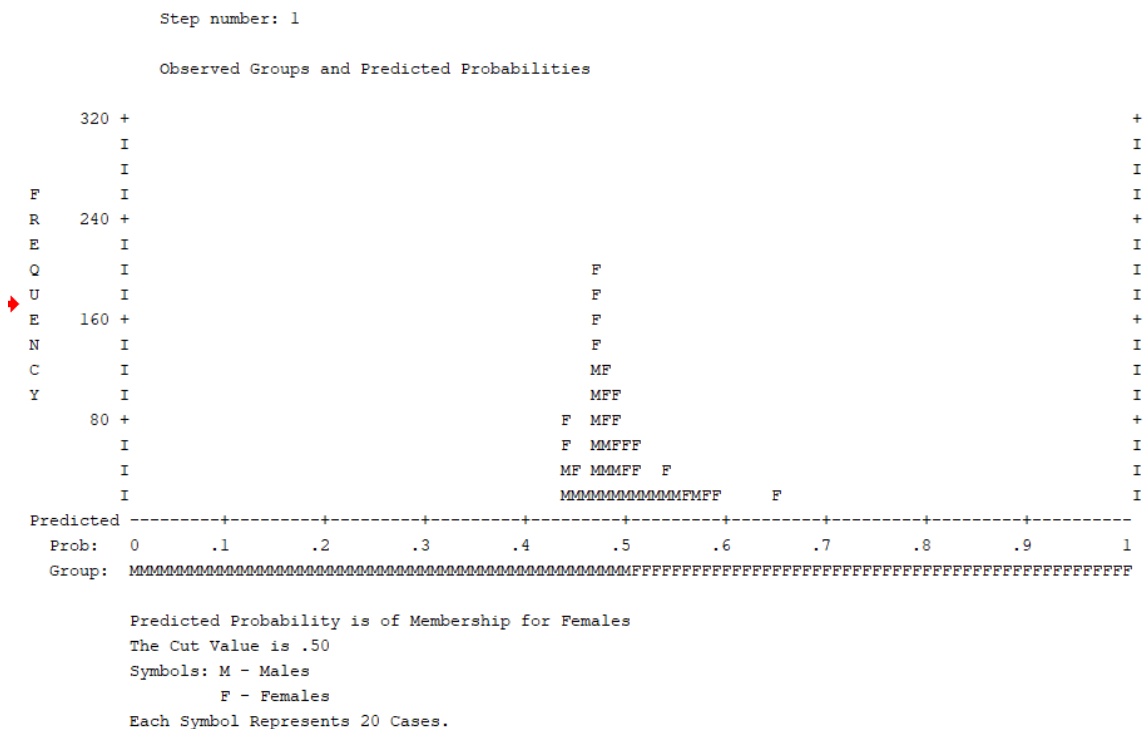
a. Variable(s) entered on step 1: Value, EducationalAttainmentLevel1.

The variable in the equation table gives the information about the contribution of each of the predictor variables. Here Wald test is used, in this value in the column Sig having value less than

0.05 is 'Value' with Significance of 0.000 from this we can say that Value contribute significantly to the model. And EducationalAttainmentLevel1 does not contribute to the model. The value in B column can be used to calculate the probability. The negative (-) value in the B column tells us the direction of influence for each variable. It indicates that an increase in the independent variable score will result in a decreased probability in the dependent variable. In this the EducationalAttainmentLevel1 value is -.126

The values in column Exp (B) are the odd ratios for each independent variable. The odd ratio of Value is 1.064 this is greater than 1 and for EducationalAttainmentLevel1 it is .882 which is less than 1. This means that the probability of outcome of value is more compared to EducationalAttainmentLevel1.

## Frequency Graph:



The frequency graph above shows that most of the cases are covered in the middle of the graph. This indicates that the probability of Males and Females for Educational Attainment Level is equal that is probability is 50:50.

## Conclusion:

From this analysis we can conclude that however multicollinearity exists between the independent variable than the variable Value has the highest significance for defining the Sex that is Males or Females. The other independent variables are not significant to bring the result.

## References:

- [1] What is Multiple Linear Regression? Available at: <https://www.statisticssolutions.com/what-is-multiple-linear-regression/> (Accessed 19 November 2018)
- [2] Multiple Regression Available at: <https://www.statisticssolutions.com/multiple-regression/> (Accessed 19 November 2018)
- [3] Statistical, Graphics, and Sample Size Software Available at: [https://ncsswpengine.netdnssl.com/wpcontent/themes/ncss/pdf/Procedures/NCSS/Multiple\\_Regression.pdf](https://ncsswpengine.netdnssl.com/wpcontent/themes/ncss/pdf/Procedures/NCSS/Multiple_Regression.pdf) (Accessed 19 November 2018)
- [4] Multiple Regression Analysis using SPSS Statistics Available at: <https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php> (Accessed 19 November 2018)
- [5] 'Descriptive Statistics' (2018) *Wikipedia* Available at: [https://en.wikipedia.org/wiki/Descriptive\\_statistics](https://en.wikipedia.org/wiki/Descriptive_statistics) (Accessed 19 November 2018)
- [6] Selection Process for Multiple Regression - Statistics Solutions Available at: <https://www.statisticssolutions.com/selection-process-for-multiple-regression/> (Accessed 19 November 2018)
- [9] SPSS Annotated Output Regression Analysis Available at: <https://stats.idre.ucla.edu/spss/output/regression-analysis/> (Accessed 19 November 2018)
- [10] Linear Regression Analysis in SPSS Statistics - Procedure, assumptions and reporting the output. Available at: <https://statistics.laerd.com/spss-tutorials/linear-regression-using-spss-statistics.php> (Accessed 19 November 2018)
- [11] Jonathan Lambert, 2017. *SPSS: How to generate a Multiple Linear Regression model - Part 2* [Online Video] Available at: <https://www.youtube.com/watch?v=oMmbUTZKFcM> (Accessed 19 November 2018)
- [12] Mark L. Berenson, David M. Levine and Timothy C. Krehbiel (2012) *Basic Business Statistics*. New Jersey: Pearson Education, Inc.
- [13] Pallant, J (2017) *SPSS Survival Manual*. England: McGraw-Hill Education
- [14] 'Logistic regression' (2018) *Wikipedia* Available at: [https://en.wikipedia.org/wiki/Logistic\\_regression](https://en.wikipedia.org/wiki/Logistic_regression) (Accessed 29 November 2018)