National College of Ireland

Project Submission Sheet – 2019/2020

School of Computing

Student Name:	RICHARD SALDANHA		
Student ID:	x18183034@student.ncirl.ie		
Programme:	MASTER OF SCIENCE IN DATA ANALYTICS – FULL TIME	**Year:**	2019
Module:	STATISTICS FOR DATA ANALYTICS		
Lecturer:	Prof. TONY DELANEY		
Submission Due Date:	28th November 2019		
Project Title:	Statistical Analysis on the Views of Americans in the field of science and technology development in the distinct future.		
Word Count:	2998		

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 bibliography 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: 23rd November 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **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.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Statistical Analysis on the Views of Americans in the field of science and technology development in the distinct future.

I.OBJECTIVE

The main objective of this project is to carry out a statistical analysis on the thought process of citizens in America in the field of science and technology development in the next five decades by applying two statistic models which are multiple linear regression and binary logistic regression, the interpretation of the results will be done using IBM SPSS.

II. Research Questions

1. Can the respondent's age give good results of the survey depending upon which state they reside, education earned, employment status?
2. Based on the answers of the different survey questions can be predict which type of the respondent would answer such type of questions in the future, considering their age and state in which they reside?

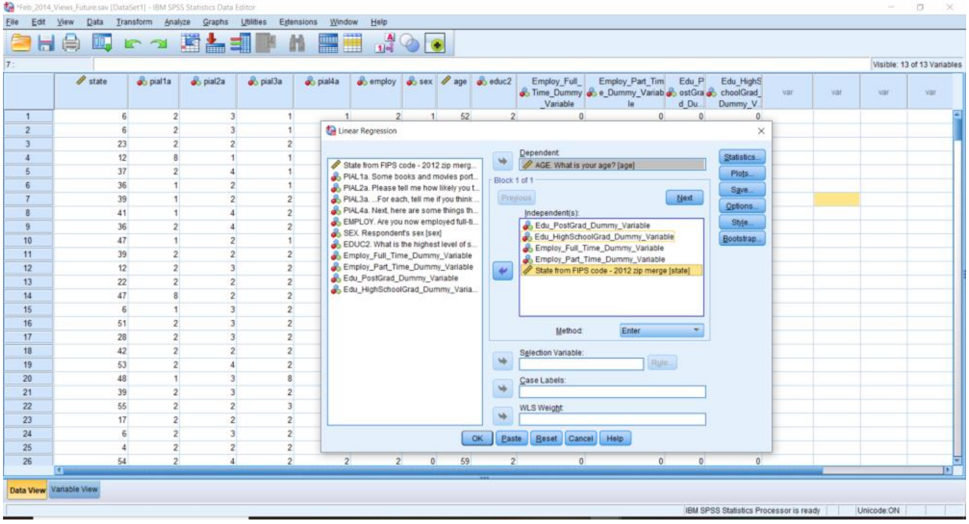
III. DATA

1. The Dataset used in this analysis is from Pew Research Center under the category of Science which is found on the second page. The link to it is mentioned below
   https://www.pewresearch.org/science/dataset/february-2014-smithsonian-science/
2. The downloaded dataset contained many valuable information such as documentation in word and pdf format along with datasets in the form of excel and SPSS.
3. It is about a survey about views of the citizens in America based on the subject lines of Science and Future in the next five decades to come.
4. This dataset contains a lot of information consisting of 51 variables a huge portion of which are questions asked either my cell phone or through landline to Americans of all age groups along with the other information such as Gender, Language in which the interview was conducted and many more.
5. In this project, the dataset has been cleaned i.e. there is no missing values and the volume of the dataset has been reduced to 364 records and 13 variables.
6. We will to apply two statistical models the first one is binary logistic regression and the other one is multiple linear regression.
7. In this project we will be making use of software tool of IBM SPSS for Statistics of Version 26 to interpret each model and to draw valuable insights from them.

IV. MODELS

1. Multiple Linear Regression
   1. In multiple regression,we try to establish a relationship between one dependent variable and two or more independent variables.The dependent variable is of scale measure while the indpendent variable have a certain criteria like for instance the independent variable cannot have multiple categories they need to be dichotomous which means that they can have only two levels which is 0 and 1 or they can be continous i.e. of a scale measure.
   2. In our project, for multiple linear regression analysis we have considered our respondent's age as our dependent variable and the state in which they reside, the education earned and the employment status as our independent vraiables.
   3. As our two variables education earned (educ2) and employment status(employ) are categorical having multiple categories we have scaled down the number of categories to three and created n-1 ie. two dummy variables for each variables.
   4. Let's perform Multiple Linear regression

1. Go to Analyze ➔ Regression ➔ Linear and then enter the dependent and independent varaiables and make sure that the method choosen is **Enter** from the dropdownlist. See below screen shot



Linear Regression

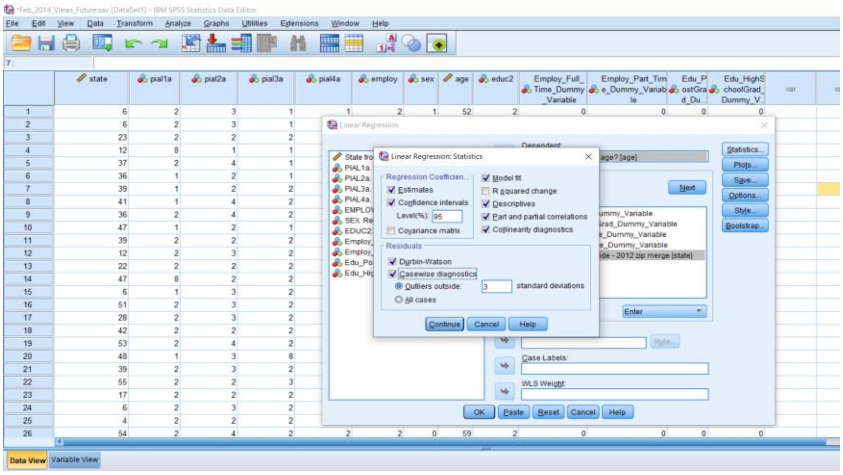
Dependent: AGE: What is your age? [age]

Block 1 of 1

Independent(s): Edu_PostGrad_Dummy_Variable, Edu_HighSchoolGrad_Dummy_Variable, Employ_Full_Time_Dummy_Variable, Employ_Part_Time_Dummy_Variable, State from FIPS code - 2012 zip merge [state]

Method: Enter

2. Next, Click on Statistics button and under regression coefficents section select Estimates,Confidence interval, Model fit, Descriptives, Part and partial correlations and Collineartity diagnostics.For Residuals we choose the Durbin-Watson test and case residuals for upto 3 standard devaiation. See below screen shot.



Linear Regression: Statistics

Regression Coefficients

☑ Estimates

☑ Confidence intervals Level(%): 95

☐ Covariance matrix

☑ Model fit

☐ R squared change

☑ Descriptives

☑ Part and partial correlations

☑ Collinearity diagnostics

Residuals

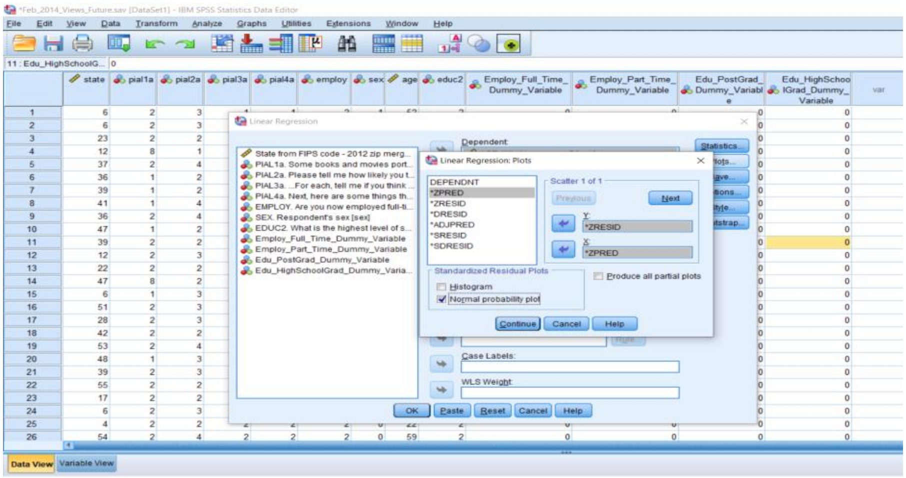
☑ Durbin-Watson

☑ Casewise diagnostics

● Outliers outside: 3 standard deviations

○ All cases

3. Next, click on Plots option and select *ZRESID for Y axis and *ZPRED for X axis and choose Normal probability plot for Standardized Residual Plots. See below screen shot.



Linear Regression: Plots

DEPENDNT, *ZPRED, *ZRESID, *DRESID, *ADJPRED, *SRESID, *SDRESID

Scatter 1 of 1

Y: *ZRESID

X: *ZPRED

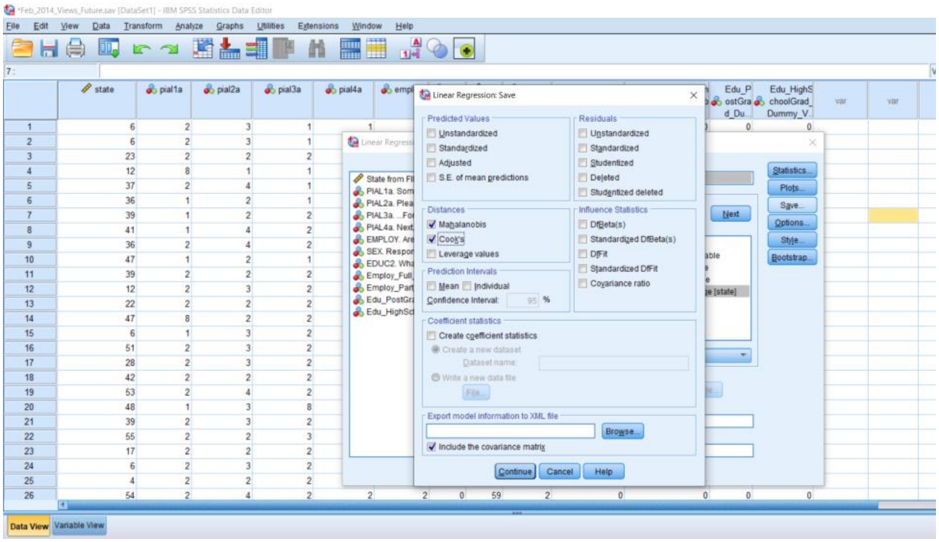
Standardized Residual Plots

☐ Histogram

☑ Normal probability plot

☐ Produce all partial plots

4. Finally, from the save option we select the Mahalanobis and Cook’s option from the Distances section See below screen shot



*Feb_2014_Views_Future.sav [DataSet1] - IBM SPSS Statistics Data Editor

	state	pial1a	pial2a	pial3a
1	6	2	3	1
2	6	2	3	1
3	23	2	2	2
4	12	8	1	1
5	37	2	4	1
6	36	1	2	1
7	39	1	2	2
8	41	1	4	2
9	36	2	4	2
10	47	1	2	1
11	39	2	2	2
12	12	2	3	2
13	22	2	2	2
14	47	8	2	2
15	6	1	3	2
16	51	2	3	2
17	28	2	3	2
18	42	2	2	2
19	53	2	4	2
20	48	1	3	8
21	39	2	3	2
22	55	2	2	3
23	17	2	2	2
24	6	2	3	2
25	4	2	2	2
26	54	2	4	2

Linear Regression: Save

Predicted Values
□ Unstandardized
□ Standardized
□ Adjusted
□ S.E. of mean predictions

Residuals
□ Unstandardized
□ Standardized
□ Studentized
□ Deleted
□ Studentized deleted

Distances
☑ Mahalanobis
☑ Cook's
□ Leverage values

Influence Statistics
□ DfBeta(s)
□ Standardized DfBeta(s)
□ DfFit
□ Standardized DfFit
□ Covariance ratio

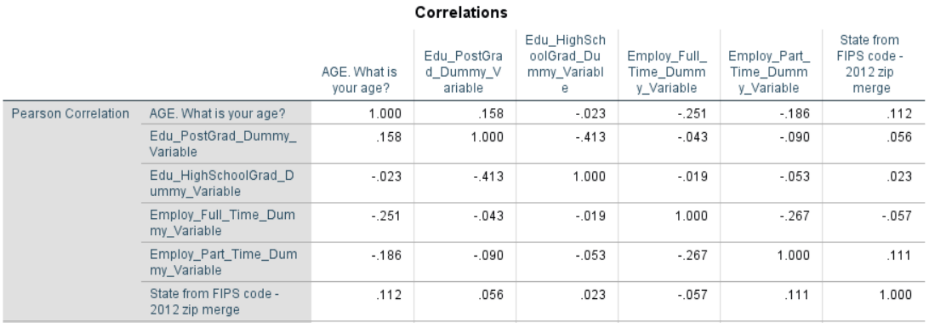
Prediction Intervals
□ Mean □ Individual
Confidence Interval: 95 %

Coefficient statistics
□ Create coefficient statistics
● Create a new dataset
Dataset name:
○ Write a new data file
File...

Export model information to XML file
Browse...
☑ Include the covariance matrix

Continue Cancel Help

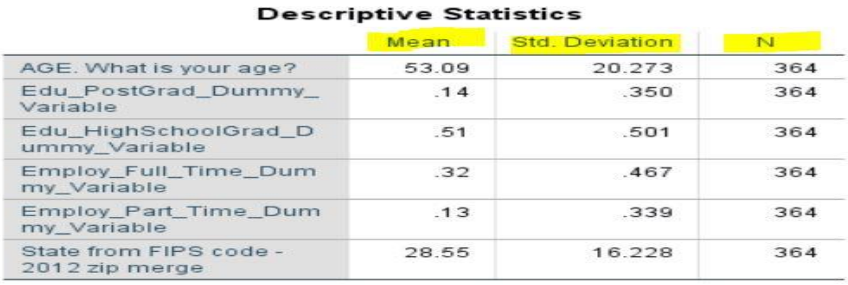
5. Click on ok, and observe the output window
6. Now, Lets interpret the output of multiple linear regression
    1. Correlations
        1. Information about the correlation between the dependent and independent variables is provided by the correlation table.
        2. We look for relationships among dependent and independent variables and also between independent variables.
        3. We observe the Pearson Correlation section to check if there exist a relation between Dependent and indpendent variables value above .3 which is prefered and among independent variables of .7 or higher is not prefered otherwise we will have to discard the variable from our analysis.In our analysis there doesn't seem to be any relation between the dependent variable and the multiple independent variables as the values are below .3 whereas in the case of independent variables there is no values .7 or more so we will retain all of the indpendent variables for our analysis.See below screenshot



Correlations

		AGE. What is your age?	Edu_PostGrad_Dummy_Variable	Edu_HighSchoolGrad_Dummy_Variable	Employ_Full_Time_Dummy_Variable	Employ_Part_Time_Dummy_Variable	State from FIPS code - 2012 zip merge
Pearson Correlation	AGE. What is your age?	1.000	.158	-.023	-.251	-.186	.112
	Edu_PostGrad_Dummy_Variable	.158	1.000	-.413	-.043	-.090	.056
	Edu_HighSchoolGrad_Dummy_Variable	-.023	-.413	1.000	-.019	-.053	.023
	Employ_Full_Time_Dummy_Variable	-.251	-.043	-.019	1.000	-.267	-.057
	Employ_Part_Time_Dummy_Variable	-.186	-.090	-.053	-.267	1.000	.111
	State from FIPS code - 2012 zip merge	.112	.056	.023	-.057	.111	1.000

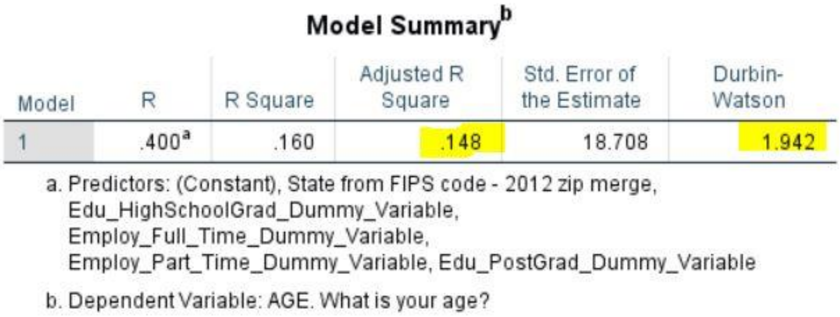
    2. Descriptive Statistics:
        1. The decriptive Statistics table summarizes the Mean, Standard Deviation and total number of records N for the dependent and independent variables.See below Screenshot.



Descriptive Statistics

	Mean	Std. Deviation	N
AGE. What is your age?	53.09	20.273	364
Edu_PostGrad_Dummy_Variable	.14	.350	364
Edu_HighSchoolGrad_Dummy_Variable	.51	.501	364
Employ_Full_Time_Dummy_Variable	.32	.467	364
Employ_Part_Time_Dummy_Variable	.13	.339	364
State from FIPS code - 2012 zip merge	28.55	16.228	364

3. <u>Model Summary</u>
    1. In model summary we have the R square,Adjusted R Square and value for Durban Watson test.
    2. In our analysis of Multiple linear regression we got .148 as our adjusted R Square value indicating 14.8% variance in our dependent variable is explained by our independent variables and Durban Watson value of 1.942 as the value lies between 0 and less than 2 it means that there is a positive autocorrelation.See below screenshot.



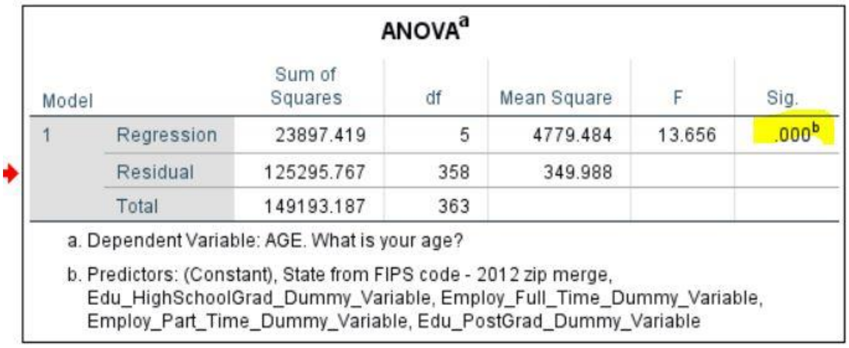
Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.400[a]	.160	.148	18.708	1.942

a. Predictors: (Constant), State from FIPS code - 2012 zip merge, Edu_HighSchoolGrad_Dummy_Variable, Employ_Full_Time_Dummy_Variable, Employ_Part_Time_Dummy_Variable, Edu_PostGrad_Dummy_Variable

b. Dependent Variable: AGE. What is your age?

4. <u>ANOVA:</u>
    1. When we look at our ANOVA table we observe that we have obtained a significant value of .000 which is less than 0.05 so there is support for the model and it is of statistical significance. See below screenshot.



ANOVA[a]

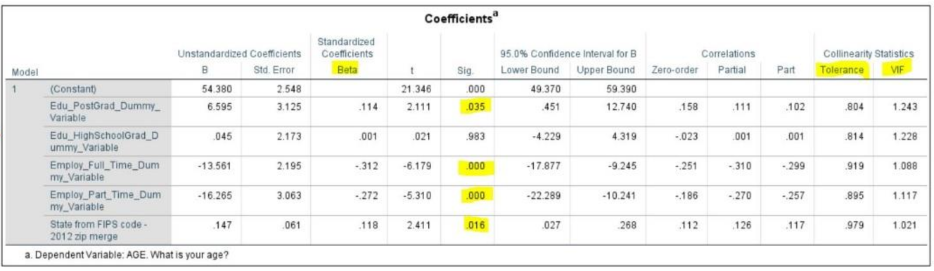
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	23897.419	5	4779.484	13.656	.000[b]
	Residual	125295.767	358	349.988		
	Total	149193.187	363			

a. Dependent Variable: AGE. What is your age?

b. Predictors: (Constant), State from FIPS code - 2012 zip merge, Edu_HighSchoolGrad_Dummy_Variable, Employ_Full_Time_Dummy_Variable, Employ_Part_Time_Dummy_Variable, Edu_PostGrad_Dummy_Variable

5. <u>Coefficients</u>
    1. To know about the contribution of each indepent variables in predicting the dependent variables we observe the Coefficients table.
    2. For comparing individual independnt variables contribution to predict the dependent variable we look at the Standardized Coeffients Beta column and look for a high value without considering(-ve)sign in our case, the variable **Employee_Full_Time_Dummy_Variable** has the highest value of -0.312 which is an indication that it is the strongest contributor to predict the dependent variable.
    3. We also look at the Sig. column for variables with values less than 0.05 to check which variables are “making a statistically significant unique contribution to the equation” (Pallant,2013,p.167).In our observation we found that there are 4 independent variables which are having significance value less than 0.05 and they are **State from FIPS code-2012-zip merge,Employ_Part_Time_Dummy_Variable,Employ_Full_Time_Dummy_Varaiable,Edu_PostGrad_Dummy_Variable**.

4. Edu_HighSchool_Dummy_Variable doesn’t make a contribution of unique significance to the prediction of dependent variable.
5. We also observe the two columns Tolerance and VIF (Variance Inflation Factor)under the Collinearity Statistics to check for multicollinearity in the model.Tolerance indicates “how much of the variability of the specified independent is not explained by the other independent variables in the model and is calculated using the formula of 1-R squared for each variable”(Pallant,2013,p.164). In our observation, the tolerance value for each individual independent variables are greater than 0.1 which means that there is no issue of multicollinearity also the VIF values are less then 10, so we can say that the independent variables are not highly correlated with each other.See below screen shot

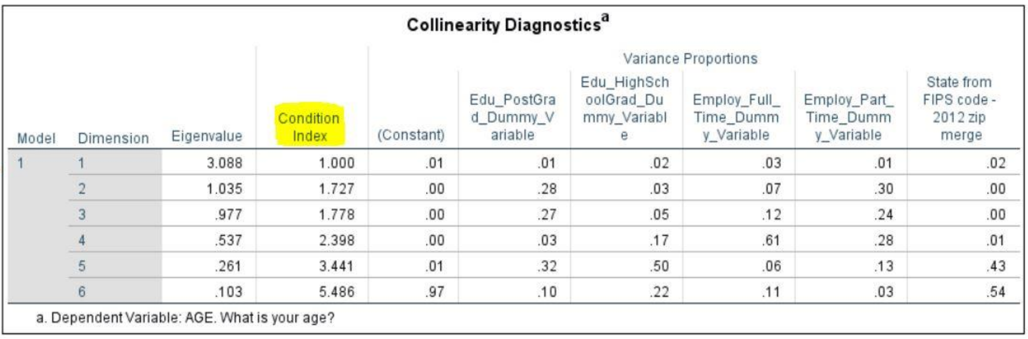


Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	54.380	2.548		21.346	.000	49.370	59.390					
	Edu_PostGrad_Dummy_Variable	6.595	3.125	.114	2.111	.035	.451	12.740	.158	.111	.102	.804	1.243
	Edu_HighSchoolGrad_Dummy_Variable	.045	2.173	.001	.021	.983	-4.229	4.319	-.023	.001	.001	.814	1.228
	Employ_Full_Time_Dummy_Variable	-13.561	2.195	-.312	-6.179	.000	-17.877	-9.245	-.251	-.310	-.299	.919	1.088
	Employ_Part_Time_Dummy_Variable	-16.265	3.063	-.272	-5.310	.000	-22.289	-10.241	-.186	-.270	-.257	.895	1.117
	State from FIPS code - 2012 zip merge	.147	.061	.118	2.411	.016	.027	.268	.112	.126	.117	.979	1.021

a. Dependent Variable: AGE. What is your age?

6. Collinearity Diagonostic
    1. Another, way to check the assumptions of multicollinearity is by observing the Collinearity Diagnostic in which we observe the condition index column.
    2. According to the study conducted by Ibm.com(case2019), If the values in the condition indexes are greater than 15, then there is collinearity among the independent variables and if the values detected are greater than 30 which means there is a serious matter of concern about the independent variables.
    3. In our analysis we found that the values in Condition Matrix are neither greater than 15 nor it is more than 30. See below screen shot.



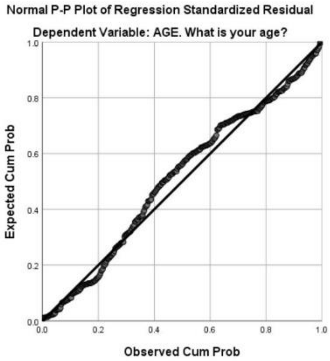
Collinearity Diagnostics[a]

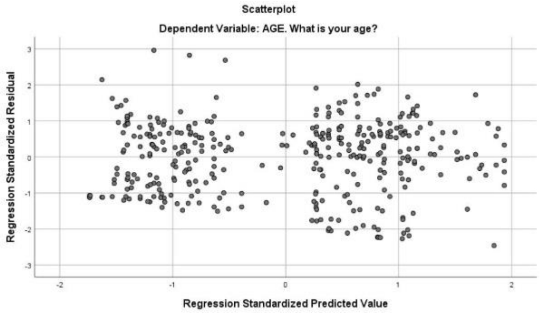
Model	Dimension	Eigenvalue	Condition Index	Variance Proportions					
				(Constant)	Edu_PostGrad_Dummy_Variable	Edu_HighSchoolGrad_Dummy_Variable	Employ_Full_Time_Dummy_Variable	Employ_Part_Time_Dummy_Variable	State from FIPS code - 2012 zip merge
1	1	3.088	1.000	.01	.01	.02	.03	.01	.02
	2	1.035	1.727	.00	.28	.03	.07	.30	.00
	3	.977	1.778	.00	.27	.05	.12	.24	.00
	4	.537	2.398	.00	.03	.17	.61	.28	.01
	5	.261	3.441	.01	.32	.50	.06	.13	.43
	6	.103	5.486	.97	.10	.22	.11	.03	.54

a. Dependent Variable: AGE. What is your age?

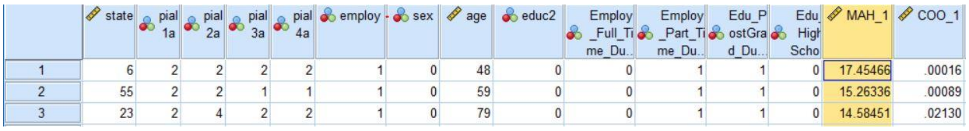
7. Interpreting the Charts
    1. In multiple linear regression, another way to check for the assumptions is by observing the normality Probability Plot of the regression standardised Residual and the Scatter plot.
    2. From the screen shot below it is evident and as Pallant(2013) states that the normal P-P plot has points probably lying in a straight diagonal line commencing from bottom left to top right which gives us a suggestion that there is no major deviations from normality.
    3. The scatter plot obtained showed us that there is a rectangular distribution of residuals and most of the score lies in the center.In

our case the standardised residual is not more than 3.3 or less than -3.3 so no suitable outliers are found.See below Screenshots



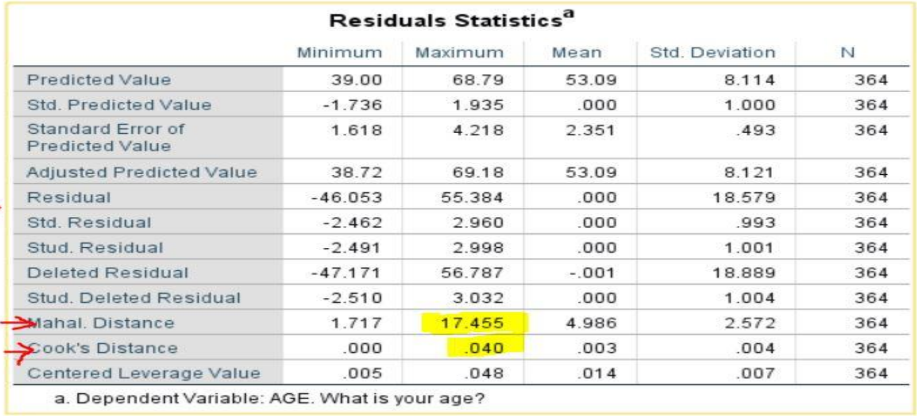


8. <u>Residual Statistics</u>
   1. Pallant(2013) suggest that by observing the Mahalanobis distances and Cook's Distance from the Residual statistics table we can check for the outliers in our system.
   2. The critical chi-square value obtained is 11.070 at an alpha value of 0.05 with 5 degrees of freedom as in our case there are 5 independent variables and the value obtained from the table is 17.455 which is more than our critical value obtained and now we can sort the MAH_1 column in descending order and find out the cases which are having higher values than critical value and terminate those cases. We found three cases which can be seen from below screenshot.



	state	pial 1a	pial 2a	pial 3a	pial 4a	employ	sex	age	educ2	Employ _Full_Ti me_Du..	Employ _Part_Ti me_Du..	Edu_P ostGra d_Du...	Edu_ High Scho	MAH_1	COO_1
1	6	2	2	2	2	1	0	48	0	0	1	1	0	17.45466	.00016
2	55	2	2	1	1	1	0	59	0	0	1	1	0	15.26336	.00089
3	23	2	4	2	2	1	0	79	0	0	1	1	0	14.58451	.02130

   3. From, the residual table the Maximum value for Cook's Distance obtained was .040 and Pallant(2013) states that if the values are less than 1 there are no dominant outliers in our predicted variables that will have a major problem in our analysis. See below screenshot



Residuals Statistics[a]

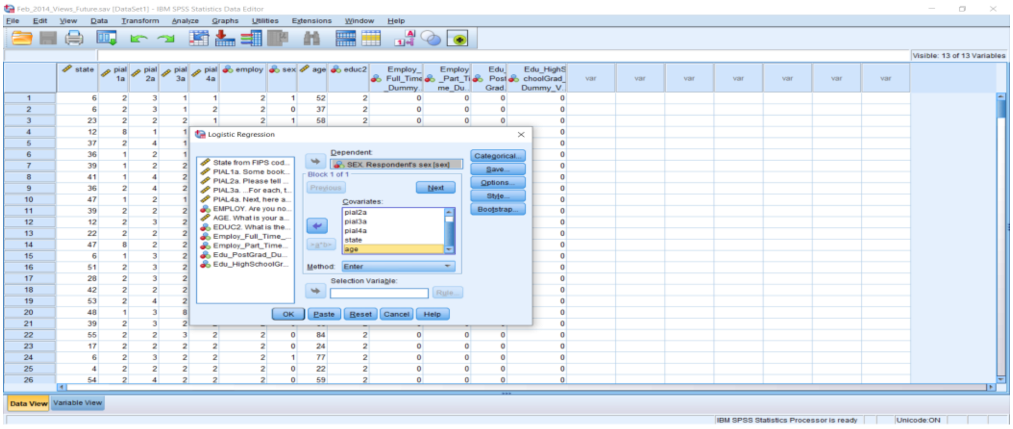
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	39.00	68.79	53.09	8.114	364
Std. Predicted Value	-1.736	1.935	.000	1.000	364
Standard Error of Predicted Value	1.618	4.218	2.351	.493	364
Adjusted Predicted Value	38.72	69.18	53.09	8.121	364
Residual	-46.053	55.384	.000	18.579	364
Std. Residual	-2.462	2.960	.000	.993	364
Stud. Residual	-2.491	2.998	.000	1.001	364
Deleted Residual	-47.171	56.787	-.001	18.889	364
Stud. Deleted Residual	-2.510	3.032	.000	1.004	364
Mahal. Distance	1.717	17.455	4.986	2.572	364
Cook's Distance	.000	.040	.003	.004	364
Centered Leverage Value	.005	.048	.014	.007	364

a. Dependent Variable: AGE. What is your age?

<u>Conclusion:</u>

From this we can conclude that the respondent's age will give us decent results of the survey as none of the assumptions were violated and the model proved to be statistically significant from the interpretations of different outputs.

2. Binary Logistic Regression
    1. Pallant(2013) states that to predict categorical oucomes with two or more categories Logistic Regression is used.In binary logistic regression the dependent variable must be a dichotomous variable and the multiple independent variables can be nominal, ordinal or of scale measure.
    2. In this project, the respondent's sex is considered as the dependent variable and it is dichotomous i.e. it has only two levels (0 and 1 indicating 0 for Female and 1 for Male) while the survey questions answered like pial1a, pial2a, pial3a, pial4a are having multiple categories and are transformed to scale measures while the variables age, and state are scale variables.
    3. Now, Let's perform Binary Logistic Regression
    1. Go to Analyze➔ Regression➔Binary Logistic and enter the dependent variable in dependent section and independent variables in the Covariates section. See below screen shot.



Logistic Regression

Dependent: SEX. Respondent's sex [sex]

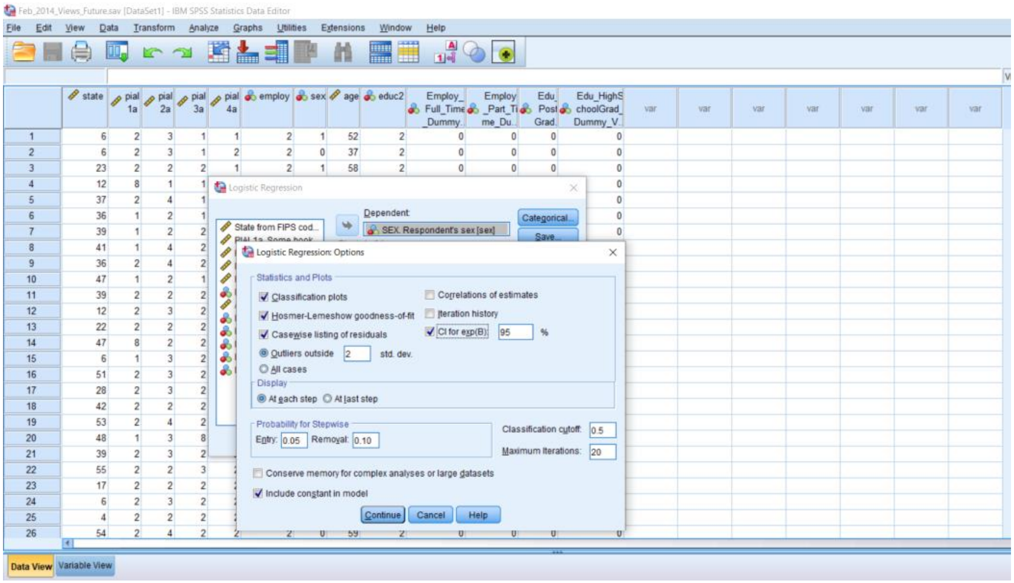
Block 1 of 1

Covariates: pial2a, pial3a, pial4a, state, age

Method: Enter

Selection Variable:

    2. Now, click on the options button we select Classification plots, Hosmer-Lemeshow goodness-of-fit, Confidence Interval CI for exp(B) of 95% interval along with Case wise listing of residuals up to 2 standard deviation. See below screen shot.



Logistic Regression: Options

Statistics and Plots

☑ Classification plots
☑ Hosmer-Lemeshow goodness-of-fit
☑ Casewise listing of residuals
● Outliers outside 2 std. dev.
○ All cases
□ Correlations of estimates
□ Iteration history
☑ CI for exp(B): 95 %

Display

● At each step ○ At last step

Probability for Stepwise

Entry: 0.05 Removal: 0.10

Classification cutoff: 0.5

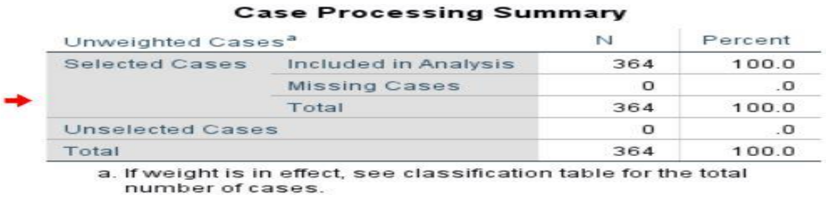
Maximum Iterations: 20

□ Conserve memory for complex analyses or large datasets

☑ Include constant in model

Feb_2014_Views_Future.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help
Logistic Regression
Dependent: SEX. Respondent's sex [sex]
Categorical...
Save...
Continue
Cancel
Help
Data View
Variable View

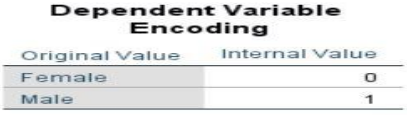
    3. Click on continue and ok and observe the output window
    4. Now, Let's interpret the output of Binary Logistic Regression
        1. We first observe the case processing summary and the dependent variable encoding and found that 364 total number of records are taken for analysis, no missing cases are observed, dependent variables are correctly coded to 0 and 1, 0 indicating Female and 1 indicating Male. See below screen shot



Case Processing Summary

Unweighted Cases^{a}		N	Percent
Selected Cases	Included in Analysis	364	100.0
	Missing Cases	0	.0
	Total	364	100.0
Unselected Cases		0	.0
Total		364	100.0

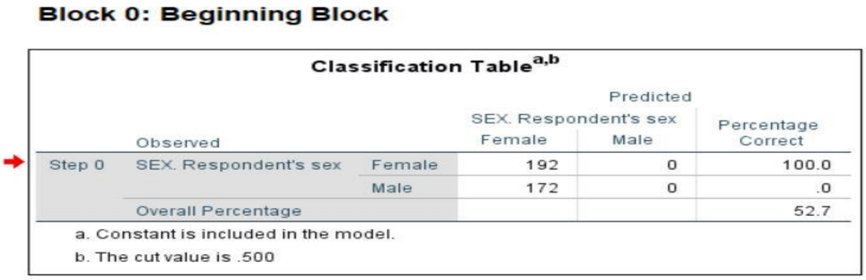
a. If weight is in effect, see classification table for the total number of cases.



Dependent Variable Encoding

Original Value	Internal Value
Female	0
Male	1

5. Block 0: Beginning Block
   1. Pallant(2013) suggest that from the Block 0 we interpret the results of our analysis without the use of any of the independent variables.
   2. On observation of the classification table we can see that the percentage of the cases correctly classified is 52.7 percent, which means that IBM SPSS assumed that the female respondents would give such type of answers on the basis of the fact that Female respondent's are more than that of Male respondent's population.See below screen shot.



Block 0: Beginning Block

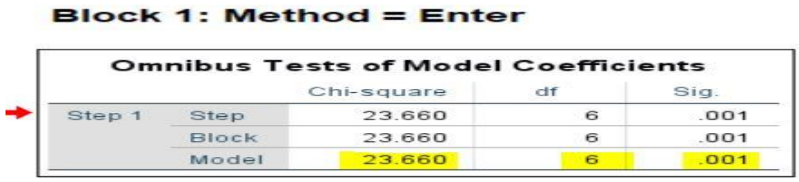
Classification Table^{a,b}

Observed			Predicted		
			SEX. Respondent's sex		Percentage Correct
			Female	Male	
Step 0	SEX. Respondent's sex	Female	192	0	100.0
		Male	172	0	.0
	Overall Percentage				52.7

a. Constant is included in the model.

b. The cut value is .500

6. Block 1: Omnibus Tests of Model Coefficients
   1. "The Omnibus Test of Model Coefficients gives us an overall indication of how well the model performs, over and above the results obtained for Block 0, with none of the predictors entered into the model.This is referred to as a 'goodness of fit'test."(Pallant,2013,p.182)
   2. We observe that the Chi-square obtained is 23.660 with 6 degrees of freedom and the level of significance is 0.001 which is less than 0.005 which means we have got a high significant value which indicates that the set of variables used as our independent variables is better than the assumption made in Block 0 that female respondents would give such answers to the questions. See below screen shot

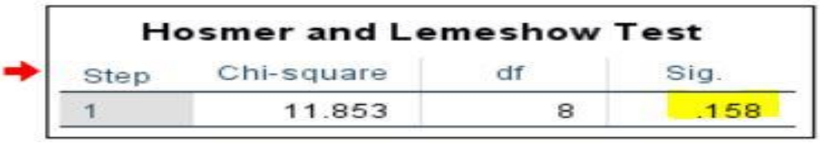


Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	23.660	6	.001
	Block	23.660	6	.001
	Model	23.660	6	.001

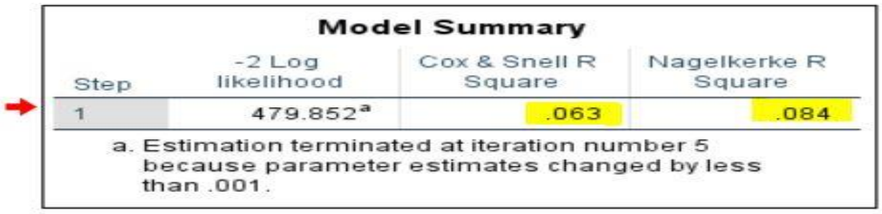
7. Hosmer and Lemeshow Test
   1. In Pallant(2013) we see that the Hosmer and Lemeshow Test is the most reliable test of model as stated by IBM SPSS, but the interpretation of the results is quite different compared to omnibus test.
   2. In this test, Pallant (2013)to support our model we need a significance value greater than 0.05, so in our observations the Chi-square is 11.853 with 8 degrees of freedom holding a significance value of .158 which is greater than 0.05 which indicates that there is a support for the model. See below screen shot



Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	11.853	8	.158

8. Model Summary
    1. Pallant (2013) suggest that observing the Cox & Snell R Square, Nagelkerke R Square values are useful as we get to know about the usefulness of the model. In Pallant (2013) the indication in the amount of variation in the dependent variable is provided by the Cox & Snell R square and the Nagelkerke R square which are pseudo R square statistics as compared to true R square value generated by output in multiple linear regression.
    2. We observe a value of .063 and .084 which gives us a suggestion that between 6.3% and 8.4% the variability is explained by this set of variables. See below screen shot.

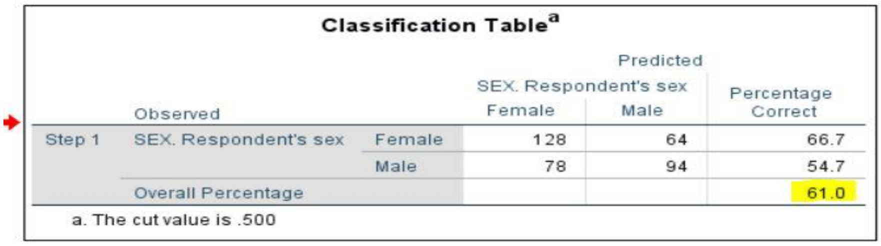


Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	479.852[a]	.063	.084

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

9. Classification Table
    1. From the classification table we can understand the nature of the model in predicting the correct category (Female /Male) respondents for each case.
    2. We do the comparison with Block 0 table to determine how much the model is improved after the consideration of predictors in our model.
    3. The model has classified correctly for each case generating a PAC(Percentage Accuracy in Classification)value of 61.0%, which is considered as an improvement over the 52.7 % in Block 0.See below screen shot



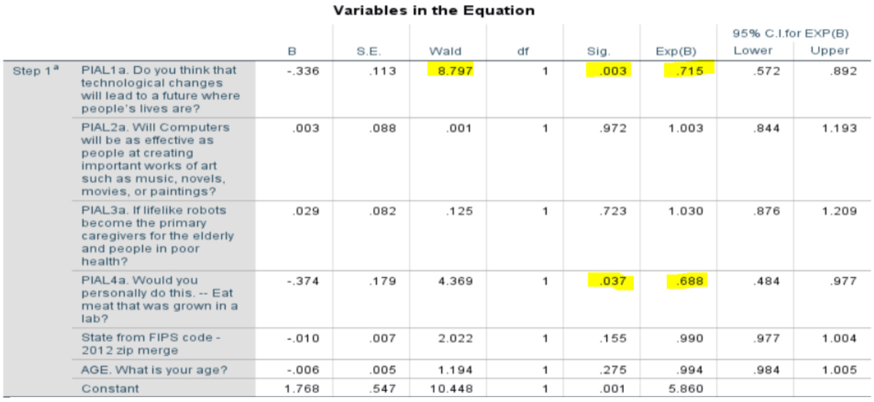
Classification Table[a]

	Observed		Predicted		
			SEX. Respondent's sex		Percentage Correct
			Female	Male	
Step 1	SEX. Respondent's sex	Female	128	64	66.7
		Male	78	94	54.7
	Overall Percentage				61.0

a. The cut value is .500

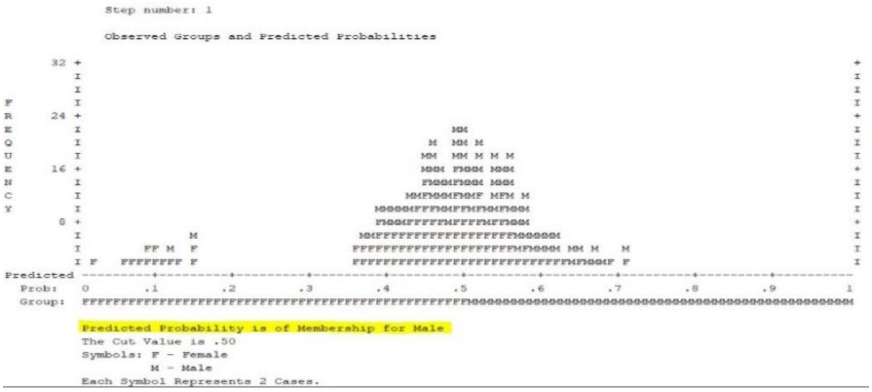
10. Variables in the equation
    1. Pallant(2013)suggest that, Variables in the Equation table plays a significant role as it tells us which independent variables or the predict variables under observation has a significant impact on the dependent variable.
    2. We observe that variables PIAL1a and PIAL4a are statistically significant predictors which are having a significance value less than 0.05 among which the Wald value is the highest for the PIAL1a variable making it the most predicted variable compared to PIAL4a.The remaining variables can be discarded from our analysis.

3. In our analysis we observe that for both the questions PIAL 1a and PIAL 4a negative values are obtained for B value (-0.336, -0.374) indicating that more answers to this question less likely are the predicted respondent Male. For variables PIAL1a, PIAL4a the odd ratios are (0.715,0.688) which is less than 1 indicating that the way the questions are answered decreases by a factor of (0.715 and 0.688) that the respondent is a Male. See below screen shots.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	PIAL1a. Do you think that technological changes will lead to a future where people's lives are?	-.336	.113	8.797	1	.003	.715	.572	.892
	PIAL2a. Will Computers will be as effective as people at creating important works of art such as music, novels, movies, or paintings?	.003	.088	.001	1	.972	1.003	.844	1.193
	PIAL3a. If lifelike robots become the primary caregivers for the elderly and people in poor health?	.029	.082	.125	1	.723	1.030	.876	1.209
	PIAL4a. Would you personally do this. -- Eat meat that was grown in a lab?	-.374	.179	4.369	1	.037	.688	.484	.977
	State from FIPS code - 2012 zip merge	-.010	.007	2.022	1	.155	.990	.977	1.004
	AGE. What is your age?	-.006	.005	1.194	1	.275	.994	.984	1.005
	Constant	1.768	.547	10.448	1	.001	5.860		



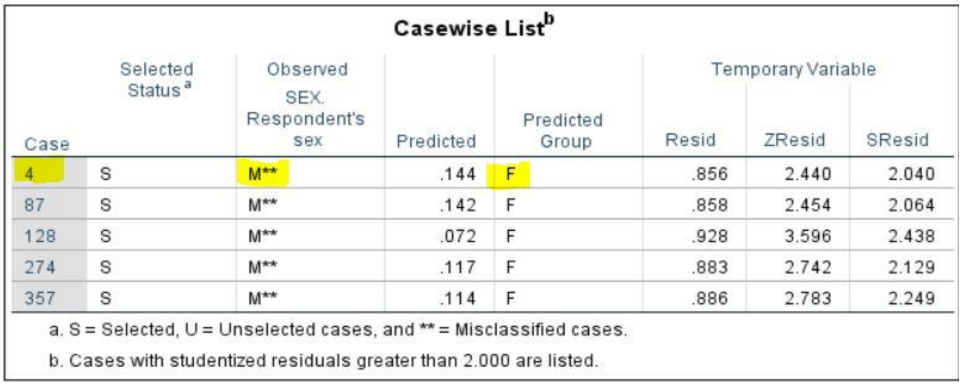
```
              Step number: 1

              Observed Groups and Predicted Probabilities

       32 +                                                                                                    +
          I                                                                                                    I
          I                                                                                                    I
F         I                                                                                                    I
R      24 +                                                                                                    +
E         I                                                MM                                                  I
Q         I                                             M  MM M                                                I
U         I                                            MM  MM M M M                                            I
E      16 +                                            MMM FMMM MMM                                            +
N         I                                            FMMMFMMM MMM                                            I
C         I                                          MMFMMMFMMF MFM M                                          I
Y         I                                      MMMMMFFFMMFFMFMMFMMM                                          I
        8 +                                      FMMMFFFFFMFFFFMFFMMM                                          +
          I              M                     MMFFFFFFFFFFFFFFFFFFFMMMMMM                                     I
          I        FF M  F                    FFFFFFFFFFFFFFFFFFFFFMFMMMM MM M   M                             I
          I F   FFFFFFFF F                    FFFFFFFFFFFFFFFFFFFFFFFFFFFFFMFMMMF F                            I
Predicted ---------+---------+---------+---------+---------+---------+---------+---------+---------+----------
  Prob:   0        .1        .2        .3        .4        .5        .6        .7        .8        .9         1
  Group:  FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMM

          Predicted Probability is of Membership for Male
          The Cut Value is .50
          Symbols: F - Female
                   M - Male
          Each Symbol Represents 2 Cases.
```

11. <u>Casewise List:</u>
    1. From the casewise List we can interpret that for cases 4,87,128,274,357 with ZResid value greater than 2 the model does not fit well and there are cases like 128,274 and 357 which are having higher ZResid value greater than 2.5 indicating a clear sign of outliers and thus they need to be removed from our dataset and redo our analysis.
    2. On observing case 4 it is evident that male respondent will answer such type of questions also considering their age and state in which they reside but Female respondent was predicted.See below shot



Casewise List[b]

Case	Selected Status[a]	Observed SEX. Respondent's sex	Predicted	Predicted Group	Temporary Variable		
					Resid	ZResid	SResid
4	S	M**	.144	F	.856	2.440	2.040
87	S	M**	.142	F	.858	2.454	2.064
128	S	M**	.072	F	.928	3.596	2.438
274	S	M**	.117	F	.883	2.742	2.129
357	S	M**	.114	F	.886	2.783	2.249

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2.000 are listed.

Conclusion:

From this we can conclude that the model proved to be statistically significant and there is support for the model as interpreted by Hosmer and Lemeshow Test. Out of the 6 predictors 2 predictors PIAL1a and PIAL4a are good predictors of which PIAL1a is the strongest predictor with high value of Wald test at 0.003 level of significance. The strongest predictor had an odd ratio of 0.715 which is less than 1 meaning the way questions are answered decreases by a factor of 0.715 that less likely the respondent is a Male.

V. BIBLIOGRAPHY

Pallant, J.(2013)*SPSS survival manual*.5th edn. McGraw-Hill Education (UK).

Ibm.com. (2019). *IBM Knowledge Center*. [online] Available at: https://www.ibm.com/support/knowledgecenter/en/SSLVMB_23.0.0/spss/tutorials/reg_cars_collin_01.html [Accessed 20 Nov. 2019].