Statistics for Data Analytics

Statistical analysis of Multiple and Logistic Regression _ CA 2

MSc Data Analytics B

Sindhujan Dhayalan

ID: X17170265

Statistical analysis of Multiple regression model:

Objective of the analysis:

1) How well do the two predictors tertiary education (%) and self-perceived health as good (%) predict the outcome variable median equivalized net income(Euro)? How much variance in the outcome variable can be explained by these 2 predictors>
2) Which predictor of the 2 predictors has the highest unique contribution in outcome variable?

Data source & details:

Data used for performing the statistical analysis is sourced from Eurostat: https://ec.europa.eu/eurostat/web/gdp-and-beyond/quality-of-life/data.

Three datasets,

- Mean and median income by age and sex - EU-SILC survey (ilc_di03)
- Self-perceived health by sex, age and educational attainment level (hlth_silc_02)
- Population by educational attainment level, sex and age (%) - main indicators (edat_lfse_03)

are extracted, cleaned and merged to form a single dataset for statistical analysis.

Independent variables used are Tertiary education (%) and Self perceived health as good (%) and dependent variable is median equivalized net income(Euro) for 30 different countries over a period of 9 years for males and females. Contributing a total of 540 records of data for statistical analysis of multiple regression model.

Preliminary tests to test that the assumptions of the technique being used are not violated for multiple regression model:

Sample size:

N > 50 + 8m (m = number of independent variables) – Ref: Tabachnick and Fidell(2013, p.123)**.** In this case 2 variables are used; therefore, Sample size must be greater than 66. The size of the data set being used is 540. Hence the sample size used is highly sufficient for multiple regression model.

Multicollinearity:

Correlations

		Median equivalised net income (Euro)	Tertiary education(%)	Good Self-perceived health-Value (%)
Pearson Correlation	Median equivalised net income(Euro)	1.000	.473	.387
	Tertiary education(%)	.473	1.000	.083
	Good Self-perceived health-Value(%)	.387	.083	1.000
Sig. (1-tailed)	Median equivalised net income(Euro)	.	.000	.000
	Tertiary education(%)	.000	.	.027
	Good Self-perceived health-Value(%)	.000	.027	.
N	Median equivalised net income(Euro)	540	540	540
	Tertiary education(%)	540	540	540
	Good Self-perceived health-Value(%)	540	540	540

Fig 1

Using the correlation matrix illustrated in fig 1, it is inferred that the predictors are not highly correlated ($r < 0.9$). More over the correlation between the predictors is significant at level .08. hence Multicollinearity does not exist. The predictors are correlate substantially with the dependent variable 'Median equivalized net income(Euro)' as the correlation value are 0.473 and 0.387 for Tertiary education (%) and Good self-perceived health-value (%) respectively. The values are above 0.3 showing existence of relationship between the independent variables/predictors with the dependent variable.

Collinearity diagnostics performed using IBM SPSS on the variables as part of multiple regression model, can identify problems with multicollinearity that would not be identified by correlation matrix. Coefficients table illustrated in fig 2 displays the absence of multicollinearity as the tolerance value of independent variables are around .90 and are more than .10 value meaning multiple correlation with another variable is low. Also, the Variance inflation factor of the independent variables are less than 10 suggesting very low possibility of multicollinearity.

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)	-33304.955	3639.542		-9.151	.000	-40454.439	-26155.470					
	Tertiary education(%)	563.463	44.469	.444	12.671	.000	476.110	650.817	.473	.480	.442	.993	1.007
	Good Self-perceived health-Value(%)	438.065	43.898	.350	9.979	.000	351.833	524.297	.387	.396	.348	.993	1.007

a. Dependent Variable: Median equivalised net income(Euro)

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	Tertiary education(%)	Good Self-perceived health-Value (%)
1	1	2.927	1.000	.00	.01	.00
	2	.067	6.594	.02	.98	.03
	3	.005	23.308	.98	.01	.97

Fig 2

Normality: Normal probability plot(P-P) of the regression standardized residual of dependent variable (Median equivalized net income(Euro) is illustrated in fig 3. It can be inferred that the points are reasonably along the straight diagonal line, there no major deviations of points from the normality line.

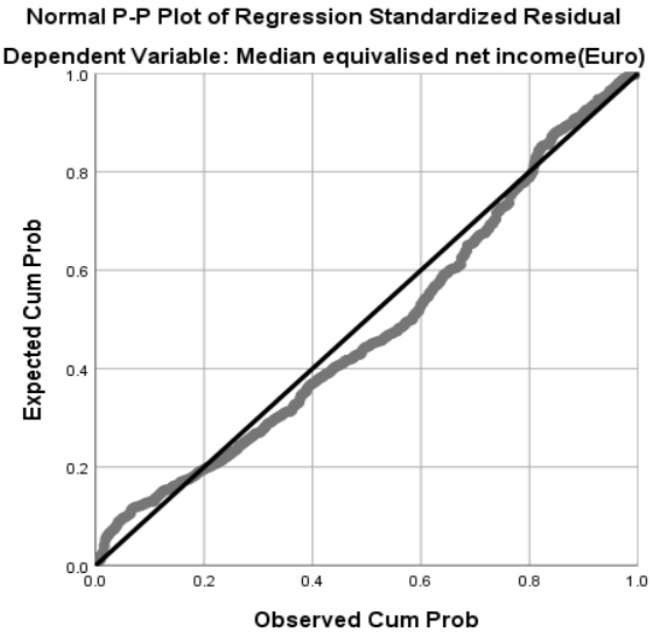

Fig 3

Outliers:

Scatter plot and histogram of dependent variable (Median equivalized net income(Euro)) is illustrated in fig 4:

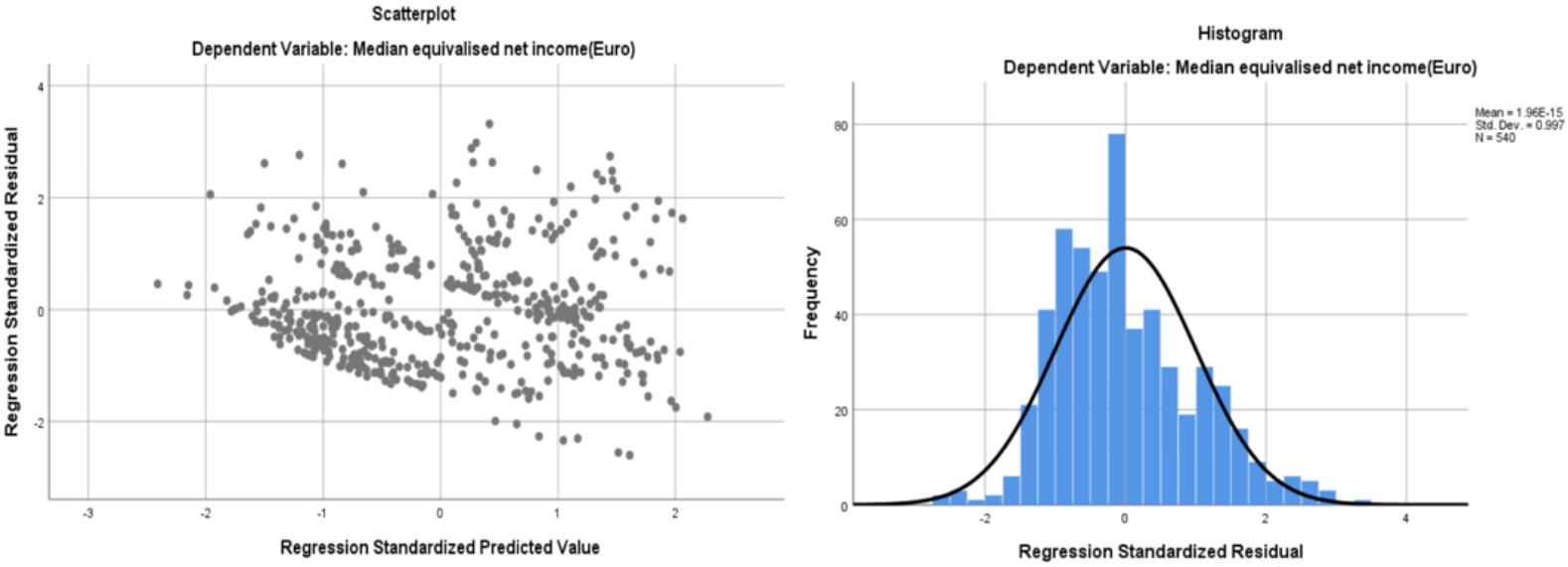

Fig 4

From the standardized residual scatter plot illustrated in Fig 4, it is inferred that the dependent variable does not seem to have any values above 3.3 and below -3.3 As per Tabachnick and Fidell (2013, p. 128), there are no outliers since there are no standardized residual value above 3.3 or

below -3.3. Also, the histogram in fig 4 displays a very well distributed data with no outliers to worry about.

Casewise diagnostics:

Casewise Diagnostics[a]

Case Number	Std. Residual	Median equivalised net income (Euro)	Predicted Value	Residual
495	3.316	46739	18830.07	27908.926

a. Dependent Variable: Median equivalised net income(Euro)

Fig 5

The fig 5 depicts one case with residual value of 3.316. It is displayed that the case number 495 had a median equivalized net income of 46739 but the predicted value is 18830.07. Clearly the model predicted the value of this case wrongly, the income is more than predicted value. To test if the case in the above diagnostics is impacting on other results of our model, we check the maximum value for Cook's distance in residual statistics table illustrated in fig 6. The maximum value is .020 lesser than 1, suggesting no problems as per the Tabachnick and Fidell (2013, p.75).

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-465.14	31513.73	16156.64	6209.115	540
Std. Predicted Value	-2.677	2.473	.000	1.000	540
Standard Error of Predicted Value	369.297	1306.936	612.407	180.955	540
Adjusted Predicted Value	-570.29	31740.25	16161.89	6212.685	540
Residual	-22709.471	26567.457	.000	8550.929	540
Std. Residual	-2.651	3.101	.000	.998	540
Stud. Residual	-2.662	3.106	.000	1.001	540
Deleted Residual	-22900.682	26647.758	-5.246	8596.511	540
Stud. Deleted Residual	-2.677	3.131	.000	1.003	540
Mahal. Distance	.003	11.546	1.996	1.971	540
Cook's Distance	.000	.020	.002	.003	540
Centered Leverage Value	.000	.021	.004	.004	540

a. Dependent Variable: Median equivalised net income(Euro)

Fig 6

1) Evaluating the model:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics			
					R Square Change	F Change	df1	df2	Sig. F Change
1	.588[a]	.345	.343	8566.838	.345	141.572	2	537	.000

a. Predictors: (Constant), Good Self-perceived health-Value(%), Tertiary education(%)

b. Dependent Variable: Median equivalised net income(Euro)

Fig 7

The Model summary table is illustrated in fig 7. The R square value in the table is 0.345. This value is the variance in the dependent variable explained by the model which includes the 2 predictors. Illustrating that about 35 percent of variance in median equivalized net income is explained by this model including predictors/independent variables (Good self-perceived health-value (%) and Tertiary education (%). The ANOVA table illustrated in fig 8, contains Sig. value = .000 meaning p < .0005, hence rejecting the null hypothesis.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.222E+10	3	7408012278	104.583	.000[b]
	Residual	3.797E+10	536	70833765.77		
	Total	6.019E+10	539			

a. Dependent Variable: Median equivalised net income(Euro)

b. Predictors: (Constant), Good Self-perceived health-Value(%), SEX, Tertiary education (%)

Fig 8

2) Evaluating each of independent variables:

Coefficients[a]

Model		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
1	(Constant)	-33304.955	3639.542		-9.151	.000	-40454.439	-26155.470					
	Tertiary education(%)	563.463	44.469	.444	12.671	.000	476.110	650.817	.473	.480	.442	.993	1.007
	Good Self-perceived health-Value(%)	438.065	43.898	.350	9.979	.000	351.833	524.297	.387	.396	.348	.993	1.007

a. Dependent Variable: Median equivalised net income(Euro)

Fig 9

Evaluating the contribution of each independent variable, Standardized Coefficients Beta of independent variable in the Coefficients table illustrated in Fig 9 are compared with each other. We find that the Beta value of Tertiary education (%): .444 is higher than Good self-perceived health-value (%): .350. The values in standardized coefficients beta depicts that variable tertiary education (%) makes the strongest unique contribution to the prediction of dependent variable: median equivalized net income(Euro). The Sig. value for tertiary education (%) and Good self-perceived health-value (%) is .000 indicating they are making a significant unique contribution to the prediction median equivalized net income(Euro). Using the Part values in correlations for Tertiary education(%): .442 and Good self-perceived health-

value (%): .348 and squaring the value results that tertiary education(%) explains 19 percent of variance in median equivalized net income(Euro) and Good self-perceived health-value(%) explains 12 percent of variance in median equivalized net income(Euro).The total R Square value in model summary table illustrated in fig 7 is 35 percentage and is close to the all squared part correlation values summed up 31 percentage .

Final regression equation:

Y = bo + b1X1 + b2X2 + ……+ bnXn

Median equalized net income = -33304.955 + 563.463 Tertiary education + 438.065 Good self-perceived health value

Result: Multiple regression was carried out to assess the unique contribution of two independent variable on a dependent variable. The model contained 2 independent variable/predictors namely tertiary education (%) and Good self-perceived health-value (%) and one outcome variable: Median equivalized net income(Euro). preliminary tests are carried out to ensure no violation of assumptions of the technique being used for multiple regression model. 35 percent of variance in median equivalized net income is explained by this model including predictors/independent variables.
The analysis answers the two queries,
Query 1: the model including tertiary education (%) and Good self-perceived health-value (%) explains 35 percent of median equivalized net income(Euro).
Query 2: Two independent variables, Tertiary education (%) produces the largest unique contribution (beta = .444) and Good self-perceived health-value (%) also produces a significant contribution (beta = .350).

Statistical analysis of Logistic regression model:

Objective of the analysis:

1) How well do the predictors predict the casualty by Gender? What is the positive and negative predictive values?
2) Which predictor of all the predictors has the highest unique contribution in outcome variable?

Data source & details:

Data used for performing the statistical analysis is sourced from Data.gov.uk: https://data.gov.uk/dataset/6efe5505-941f-45bf-b576-4c1e09b579a1/road-traffic-accidents

A single dataset containing data on accidents across Leeds. Data includes number of vehicles, number of people and vehicles involved, road surface, weather conditions, severity of any casualties, sex of causality and Causality class. For the statistical analysis of this model, Road surface, Casualty class, Casualty severity categorical variables and number of vehicles numeric variable is considered as dependent variables and Sex of the Casualty is taken as the dependent variable.

Road surface is categorized into Dry, Frost/Ice, Snow, Wet/Damp.

Casualty class is categorized into Driver, Passenger and pedestrian.

Casualty severity is categorized into severe and slight.

Sex of casualty is categorized into male and female

Sample size:

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	2686	100.0
	Missing Cases	0	.0
	Total	2686	100.0
Unselected Cases		0	.0
Total		2686	100.0

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

Fig10

The case processing summary table illustrated in fig 10, displays the sample size included in analysis (2686) and 100% of the sample is used for this statistical analysis as expected.

Dependent variable encoding & Categorical variables coding:

Dependent Variable Encoding

Original Value	Internal Value
Female	0
Male	1

Categorical Variables Codings

			Parameter coding		
		Frequency	(1)	(2)	(3)
Road Surface	Dry	2052	.000	.000	.000
	Frost /	34	1.000	.000	.000
	Snow	6	.000	1.000	.000
	Wet / Da	594	.000	.000	1.000
Casualty Class	Driver	1599	.000	.000	
	Passenge	738	1.000	.000	
	Pedestri	349	.000	1.000	
Casualty Severity	Serious	297	.000		
	Slight	2389	1.000		

Fig11

As illustrated in the fig 11, Using IBM SPSS the dependent variable (casualty sex) of male and female is encoded into 0 and 1. Categorical variables are coded as illustrated in fig 11 with the help of IBM SPSS.

Result of analysis without the existence of any independent variables:

Block 0: Beginning Block

Classification Table[a,b]

			Predicted		
			Sex of Casualty		Percentage Correct
	Observed		Female	Male	
Step 0	Sex of Casualty	Female	0	1087	.0
		Male	0	1599	100.0
	Overall Percentage				59.5

a. Constant is included in the model.

b. The cut value is .500

Fig12

As per the fig 12, Block 0 classification table, the overall percentage of correctly classified cases is 59.5 and IBM SPSS has guessed that males are the causalities by accidents. By further analysis, It can be verified if the prediction from IBM SPSS on the dependent variable without any independent variable is valid.

Results of analysis with the existence of independent variables:

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	161.427	7	.000
	Block	161.427	7	.000
	Model	161.427	7	.000

Fig13

The **goodness of fit test** also called as Omnibus test of model coefficients indicated how well the model performs with the inclusion of the independent variables compared to Block 0 without any variables. To prove a that this model is better than IBM SPSS guess, the Sig. value from this model must be lesser than .05. The value in current case is .000 as illustrated in Fig13.hence indicating that this model is better than IBM SPSS guess. The Chi square value is 161.427 with 7 degrees of freedom.

Hosmer and Lemeshow test: Most reliable test of model fit available in SPSS.

This test indicates that the model is poor if the significant value is less than .05. The test ran for this model is illustrated in fig 14.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	8.196	6	.224

Fig14

The Sig. value from the Hosmer and Lemeshow test is above .05, thus providing further support for the goodness of the model being used. The Chi-square value for this test is 8.196 with significance value of 0.224.

Model summary:

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	3463.963[a]	.058	.079

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

Fig15

The Cox & Snell R square and Nagelkerke R Square values indicates the amount of variance in the dependent variable explained by the model. From the values illustrated in fig 15, It is inferred that 5.8 percent and 7.9 percent of variability is explained by this set of variables.

Classification table:

Classification Table[a]

Observed			Predicted: Sex of Casualty: Female	Predicted: Sex of Casualty: Male	Percentage Correct
Step 1	Sex of Casualty	Female	440	647	40.5
		Male	303	1296	81.1
	Overall Percentage				64.6

a. The cut value is .500

Fig16

The model correctly classified 64.6 percent of cases overall, improving from 59.5 percent of cases overall depicted from the table in Block 0. There is a 5 % overall increase when all the independent variables are included.

Sensitivity of the model have been accurately identified by the model as 81.1 percent (True positive) and the specificity of the model have been accurately identified by the model as 40.5(True negative).

From these values, it is inferred that model has correctly classified male who have been casualties of accidents and 40.5 percentage of females are correctly classified as casualties of accidents.

The positive predictive value is 67 percent {(1296/ (1296+647)) * 100}, This indicates the of the males predicted to have been the casualties of accident, the model has picked 67 percent accurately.

The negative predictive value is 59 percent {(440/ (440+303)) * 100}, This indicates the of the females predicted to have been the casualties of accident, the model has picked 59 percent accurately.

Variables in the equation table: Depicts the importance of each of the predictor variables:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
Step 1[a]	Casualty Severity(1)	-.131	.134	.965	1	.326	.877	.675	1.140
	Number of Vehicles	-.067	.051	1.760	1	.185	.935	.846	1.033
	Road Surface			4.644	3	.200			
	Road Surface(1)	-.065	.362	.032	1	.858	.937	.461	1.904
	Road Surface(2)	-1.214	.878	1.912	1	.167	.297	.053	1.660
	Road Surface(3)	.158	.099	2.549	1	.110	1.171	.965	1.422
	Casualty Class			147.316	2	.000			
	Casualty Class(1)	-1.135	.094	147.195	1	.000	.321	.268	.386
	Casualty Class(2)	-.417	.135	9.499	1	.002	.659	.505	.859
	Constant	.987	.168	34.303	1	.000	2.683		

a. Variable(s) entered on step 1: Casualty Severity, Number of Vehicles, Road Surface, Casualty Class.

Fig17

From the fig17,

The variables that contribute significantly to the predictive ability of the model will have Sig. value < 0.5, In this model Casualty class is the most significant predictor for the model in explaining the variance of dependent variable. Road surface, number of vehicles and Casualty severity have been less significant.

From Beta values, it is inferred that, Casualty severity, number of vehicles, Causality all are more focused on depicting female casualties. Predicting these independent variables most cause female casualties and Road surface (3) is predicting more male casualties.

Casewise list table was not produced in the analysis as no outliers were found.

Result: Logistic regression was performed to access the impact of a number of independent variables on predicting the sex of the causality. The model contained 4 independent variables and full model including all the independent variables is statistically significant. Casualty class is the most significant predictor of the model. The positive predictive value of the model is 67% and negative predictive value of the model is 59%.

References:

1)TABACHNICK, B. G., & FIDELL, L. S. (2013). *Using multivariate statistics*. Boston, Pearson Education.

2)PALLANT, J. (2016). *SPSS survival manual: a step by step guide to data analysis using SPSS*. Maidenhead, Open University Press.

3) UDEMY COURSE: https://www.udemy.com/spss-statistics-foundation-course-from-scratch-to-advanced/learn/v4/t/lecture/7827104?start=315

4) NCI statistics for data analytics slides.