Multiple Regression Analysis

Multiple regression allows to understand a relationship between one dependent variable and more than one independent variables or predictors. Dependent variable in multiple regression should always be continuous while the independent variables can be continuous or dichotomous in nature. Multiple regression equation is represented as:



Where,

Y hat 🡪 Dependent variable

a 🡪 Intercept on the Y-axis

X1, X2…..Xk 🡪 Independent variables

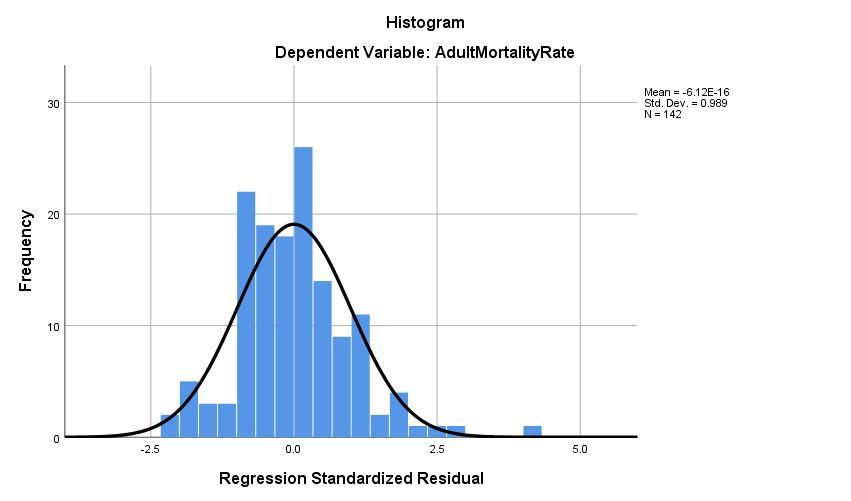
b1, b2…..bk 🡪 Co-efficient of independent variables

**Data:**

For performing multiple regression and build a model on it, data from ‘World Health Organization’ is used. Five different data sets were combined in one for this analysis for the year 2016.

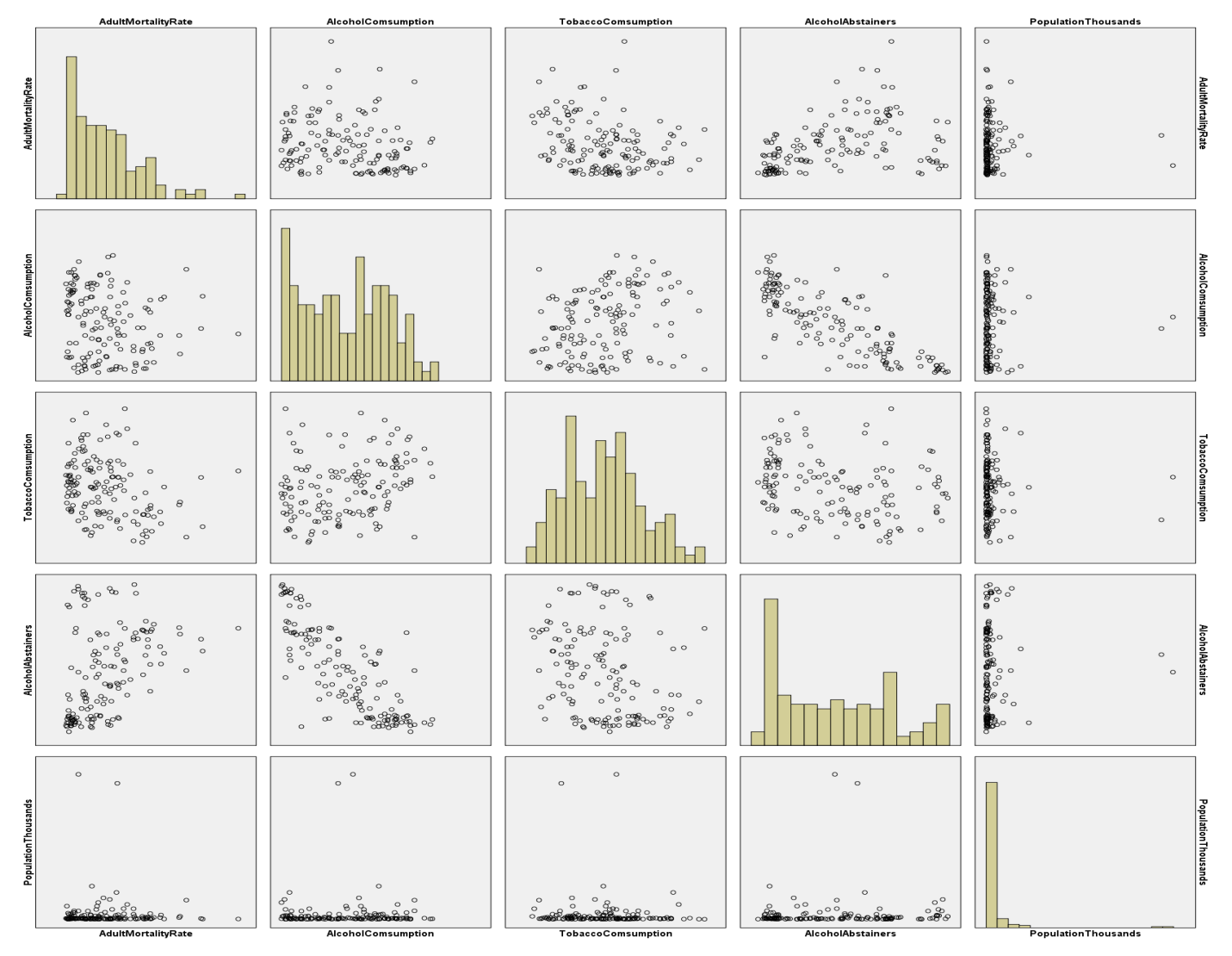
1. First data set is country wise percentage of lifetime alcohol abstainers which by definition means **“Lifetime abstainers are defined as the proportion of adults (15+ years) in a given population who have not consumed any alcohol during their lifetime, assessed at a given point in time. Numerator: The number of lifetime abstainers (15+ years). Denominator: The total number of participants (15+ years) responding to the corresponding question in a given survey.”**
2. Second data set is country wise adult mortality rate which on WHO website is defined as **“Probability that a 15 year old person will die before reaching his/her 60th birthday. The probability of dying between the ages of 15 and 60 years (per 1 000 population) per year among a hypothetical cohort of 100 000 people that would experience the age-specific mortality rate of the reporting year.”**
3. Third data set used in the analysis is country wise alcohol consumption per capita and is defined as **“Total alcohol per capita consumption (APC) is defined as the total (sum of recorded and unrecorded alcohol) amount of alcohol consumed per person (15years of age or older) over a calendar year, in litres of pure alcohol, adjusted for tourist consumption. The estimates for the total alcohol consumption are produced by summing up the 3-year average per capita (15+) recorded alcohol consumption and an estimate of per capita (15+) unrecorded alcohol consumption for a calendar year. Tourist consumption takes into account tourists visiting the country and inhabitants visiting other countries.”**
4. Fourth data set is tobacco consumption by country defined as **“The percentage of the population aged 15 years and over who currently use any tobacco product (smoked and/or smokeless tobacco) on a daily or non-daily basis.”**
5. Fifth data set is total population for countries in thousands defined as **“De facto population in a country, area or region. Figures are presented in thousands.”**

**Assumptions:**

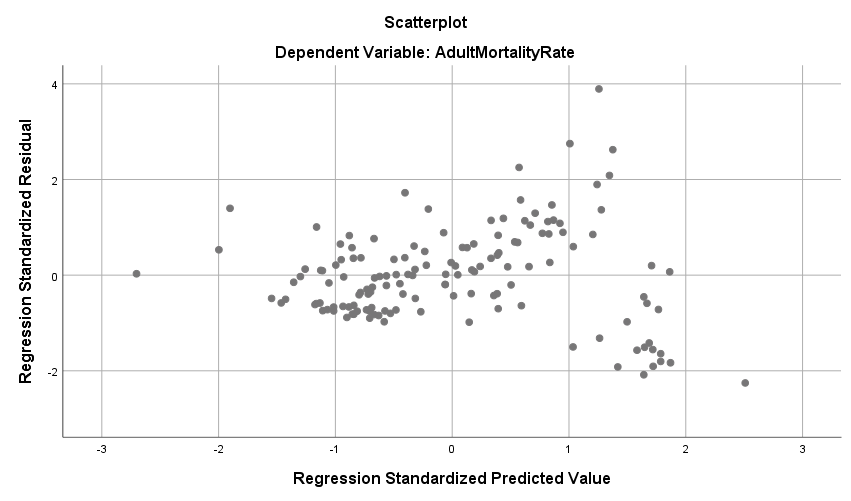
1. Variable Type – Dependent and independent variables both are continuous.
2. Normal Distribution of Regression Residuals – Data used for regression analysis shows that the regression residuals are normally distributed. Refer Fig. 1 below

**Fig. 1**

1. Linear Relationship between dependent and independent variables –

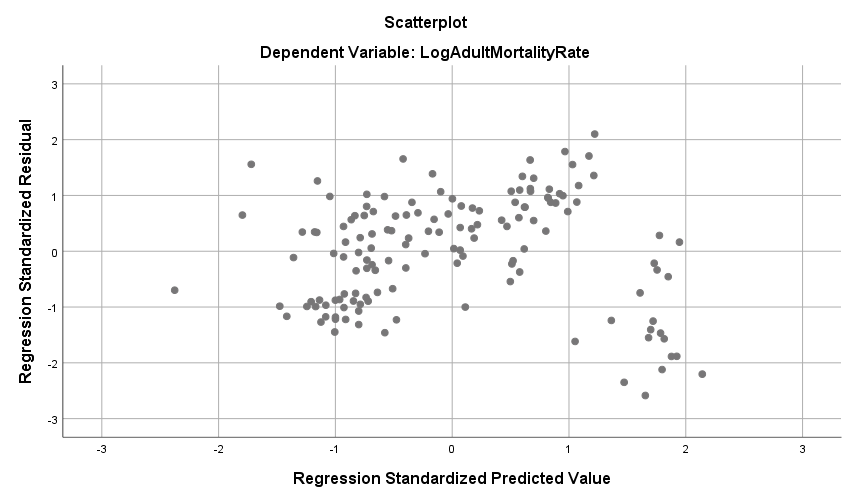


1. From above scatter plot matrix, we can see that linear relationship is present for dependent variable ‘Adult Mortality Rate’ and independent variables ‘Alcohol Consumption’ & ‘Alcohol Abstainers’. However, a satisfying linear relationship is not found for ‘Tobacco Consumption’ as compared with other independent variables and not at all for ‘Population in Thousands’ if we ignore the two outliers. So, for regression analysis we’ll only take dependent variable ‘Adult Mortality Rate’ and independent variables ‘Alcohol Consumption’, ‘Alcohol Abstainers’.
2. Homoscedasticity – Fig. 2 shows that Homoscedasticity is not met. From the residual plot present below, we can see that the plot is heterocedastic.

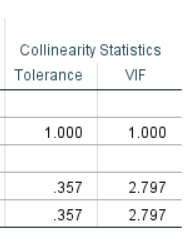


**Fig. 2**

So, taking log of dependent variable ‘Adult Mortality Rate’. After taking log of the dependent variable, the plot is as follows:



1. Collinearity / Multicollinearity – The VIF values for data are less than 10. Fig. 3



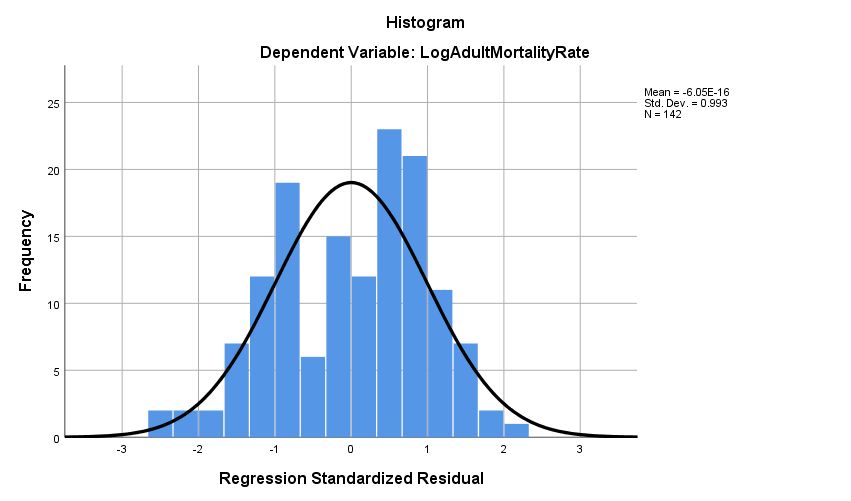
**Fig.3**

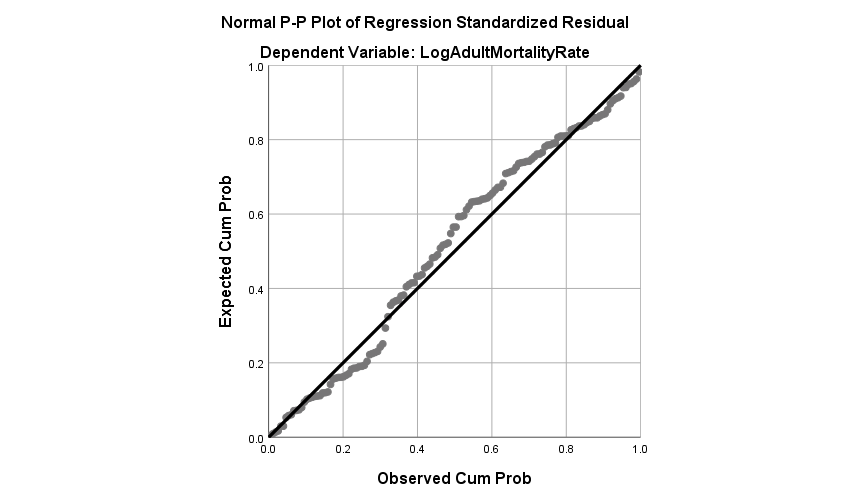
1. Independence of Errors Durbin-Watson Statistic: From Fig. 4 we can see that the Durbin Watson Statistic is near to 2.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model Summaryc** | | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Durbin-Watson |
| 1 | .413a | .170 | .164 | .21685 |  |
| 2 | .441b | .194 | .183 | .21445 | 1.931 |
| a. Predictors: (Constant), AlcoholAbstainers | | | | | |
| b. Predictors: (Constant), AlcoholAbstainers, AlcoholComsumption | | | | | |
| c. Dependent Variable: LogAdultMortalityRate | | | | | |

**Fig. 4**

1. Normality of Errors: Histogram of standardized residuals is normally distributed which can also be seen from the Normal Probability Plot below





**Regression Analysis:**

As the assumption of homoscedasticity was not met, so log of Y was taken and then the regression analysis was carried out.

1. Correlation Matrix: From Fig. 5 we can see that there is a strong inverse relation between Alcohol Consumption and Alcohol Abstainers with Pearson correlation coefficient = -0.802. Strong positive correlation can also be found between Alcohol Abstainers and LogAdultMortalityRate.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Correlations** | | | | |
|  | | LogAdultMortalityRate | AlcoholComsumption | AlcoholAbstainers |
| Pearson Correlation | LogAdultMortalityRate | 1.000 | -.238 | .413 |
| AlcoholComsumption | -.238 | 1.000 | -.802 |
| AlcoholAbstainers | .413 | -.802 | 1.000 |
| Sig. (1-tailed) | LogAdultMortalityRate | . | .002 | .000 |
| AlcoholComsumption | .002 | . | .000 |
| AlcoholAbstainers | .000 | .000 | . |
| N | LogAdultMortalityRate | 142 | 142 | 142 |
| AlcoholComsumption | 142 | 142 | 142 |
| AlcoholAbstainers | 142 | 142 | 142 |

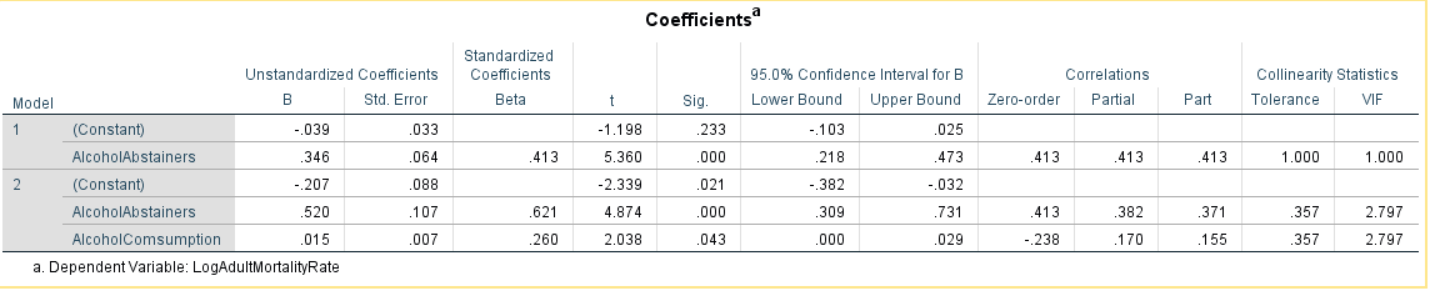
**Fig. 5**

1. Model Summary: Fig. 6 below shows R2 for model 2 is 0.194 which means that 19.4% of the variations in dependent variable is accounted for by dependent variables present in model 2. Adjusted R2 is also near the R2, which represent the same thing but for population from where the sample has been taken.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model Summaryc** | | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Durbin-Watson |
| 1 | .413a | .170 | .164 | .21685 |  |
| 2 | .441b | .194 | .183 | .21445 | 1.931 |
| a. Predictors: (Constant), AlcoholAbstainers | | | | | |
| b. Predictors: (Constant), AlcoholAbstainers, AlcoholComsumption | | | | | |
| c. Dependent Variable: LogAdultMortalityRate | | | | | |

**Fig. 6**

1. Coefficients:



From above table,

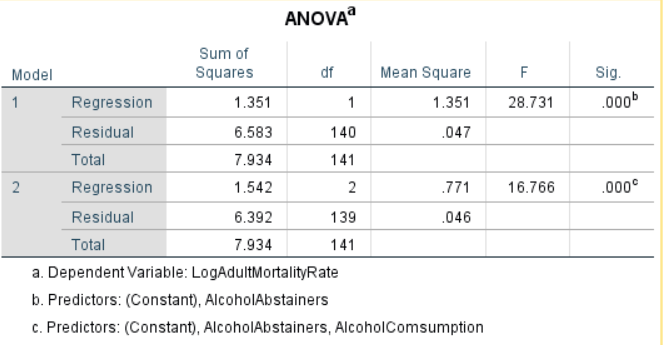
The regression equation is: LogAdultMortalityRate = -0.207 + 0.520(AlcoholAbstainers) + 0.015(AlcoholConsumption)

This equation means that for every unit increase in Alcohol Abstainers, LogAdultMortality rate increases by 0.520. Similarly, for every unit increase in Alcohol Consumption, LogAdultMortality rate increases by 0.015.

As the significance for both independent variables is less than 0.05, this means that both independent variables are making significant contribution to the model.

Standardized Beta values present in the table are measured in terms of standard deviation. As alcohol abstainers increases by 1 standard deviation, LogAdultMortality increase by 0.621 of a standard deviation. Similarly, When the alcohol consumption increases by 1 SD, LogAdultMortality increase by 0.260 standard deviations.

ANOVA: Analysis of variance is significant



**Conculsion:**

A multiple linear regression was calculated to predict Adult Mortality Rate based on Alcohol Consumption and Alcohol Abstainers. A significant regression equation was found (F(2, 139) = 16.766, p < .000), with an R2 of .194. Log of Adult Mortality Rate is equal to -0.207 + 0.520(AlcoholAbstainers) + 0.015(AlcoholConsumption), where, Alcohol Abstainers is measured in percentage and Alcohol Consumption is measured in litres per capita. Log of Adult Mortality rate increased by 0.520 for every unit increase in Alcohol Abstainers and Log of Adult Mortality rate increased by 0.015 for every unit increase in Alcohol Consumption. Both Alcohol Abstainers and Alcohol Consumption were significant predictors of weight.

Logistic Regression Analysis

**Data:**

For performing logistic regression and build a model on it, data from ‘Eurostat’ is used. Only a single data set was used here.

Data Title 🡪 Total population living in a dwelling with a leaking roof, damp walls, floors or foundation, or rot in window frames of floor

Objective 🡪 The model will predict whether percentage of population living in a particular housing type and belonging to a certain age group is male or a female. Data from the referenced link (Europe.eu (2018) Total population living in a dwelling with a leaking roof, damp walls, floors or foundation, or rot in window frames of floor for Year 2017) is taken only for year 2017.

**Assumptions:**

1. Dependent variable outcomes are mutually exclusive and is dichotomous in nature having levels: Male and Female.
2. Sample size is 961 rows which is enough to carry out the analysis.

**Regression Analysis:**

**Block 0 (Beginning Block):** In this block, none of the independent variables are entered. This block is only based on independent variable.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classification Tablea,b** | | | | | |
|  | Observed | | Predicted | | |
|  | Gender | | Percentage Correct |
|  | Females | Males |
| Step 0 | Gender | Females | 481 | 0 | 100.0 |
| Males | 480 | 0 | .0 |
| Overall Percentage | |  |  | 50.1 |
| a. Constant is included in the model. | | | | | |
| b. The cut value is .500 | | | | | |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables in the Equation** | | | | | | | |
|  | | B | S.E. | Wald | df | Sig. | Exp(B) |
| Step 0 | Constant | -.002 | .065 | .001 | 1 | .974 | .998 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables not in the Equation** | | | | | |
|  | | | Score | df | Sig. |
| Step 0 | Variables | HousingType(1) | .001 | 1 | .974 |
| AgeGroup | .002 | 2 | .999 |
| AgeGroup(1) | .002 | 1 | .964 |
| AgeGroup(2) | .001 | 1 | .982 |
| PercentageOfPopulation | .068 | 1 | .794 |
| Overall Statistics | | .079 | 4 | .999 |

**Block 1:** In block 1 all the independent variables are entered into the model in one go using method ‘Enter’.

1. Omnibus Tests of Model Coefficients:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Omnibus Tests of Model Coefficients** | | | | |
|  | | Chi-square | df | Sig. |
| Step 1 | Step | .079 | 4 | .999 |
| Block | .079 | 4 | .999 |
| Model | .079 | 4 | .999 |

As per omnibus test, we can see that coefficients are not significant.

1. Model Summary: -2LL is too high in the model summary which shows that the model is not fit. Moreover, Cox & Snell R square and Nagelkerke R square is also .000 and reflects that the independent variables actually don’t account for any variations in dependent variable which is Gender. On the other hand, Hosmer and Lemeshow test shows an opposite scenario.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model Summary** | | | |
| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
| 1 | 1332.148a | .000 | .000 |
| a. Estimation terminated at iteration number 2 because parameter estimates changed by less than .001. | | | |

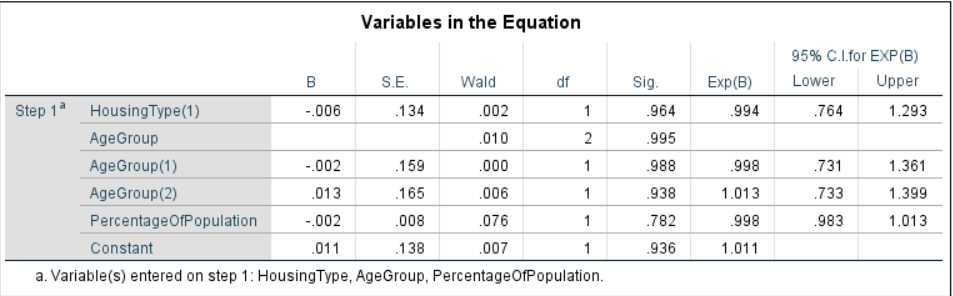
1. Hosmer and Lemeshow Test: While the model summary shows that the model is not fit, Hosmer and Lemeshow Test shows a significance value of 0.937, which says model is a good fit. This test is contradicting with model summary somehow which is confusing.

|  |  |  |  |
| --- | --- | --- | --- |
| **Hosmer and Lemeshow Test** | | | |
| Step | Chi-square | df | Sig. |
| 1 | 2.366 | 7 | .937 |

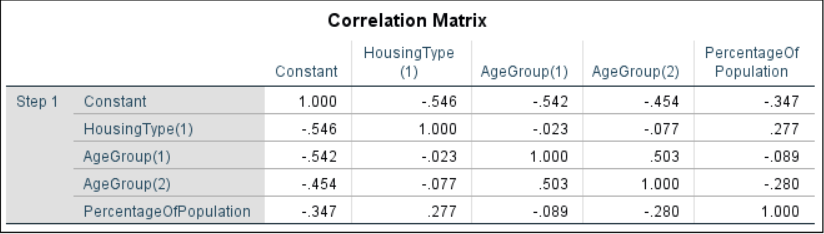
1. Classification Table: The table is showing results with 49.8% accuracy.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classification Tablea** | | | | | |
|  | Observed | | Predicted | | |
|  | Gender | | Percentage Correct |
|  | Females | Males |
| Step 1 | Gender | Females | 176 | 305 | 36.6 |
| Males | 177 | 303 | 63.1 |
| Overall Percentage | |  |  | 49.8 |
| a. The cut value is .500 | | | | | |

1. Variables in the Equation: Variable in below table are not significant. Odds ratio in the table is also very high.



1. Correlation Matrix: We can see below that multi-collinearity exists in the data



Conclusion:

Although few assumptions has been met, this model has many limitations in it and is not fit. If we include another independent variable, it might be a case that model will perform better.

**References:**

World Health Organization (2018) Abstainers, lifetime by country (view.main.52400). Available at: http://apps.who.int/gho/data/view.main.52400. [Accessed 27 November 2018].

World Health Organization (2018) Adult Mortality Data by Country. Available at: http://apps.who.int/gho/data/view.main.1360. [Accessed 27 November 2018].

World Health Organization (2018) Alcohol per capita Consumption by Country. Available at: http://apps.who.int/gho/data/view.main.1800. [Accessed 27 November 2018].

World Health Organization (2018) Age-standardized prevalence of current tobacco smoking among persons aged 15 years and older. Available at:http://apps.who.int/gho/data/view.main.GSWCAH20v. [Accessed 27 November 2018].

World Health Organization (2018) Population Data by Country (all years). Available at: http://apps.who.int/gho/data/view.main.POP2040ALL. [Accessed 27 November 2018].

Slideshare.net Reporting a Multiple Linear Regression in APA (2018). Available at: https://www.slideshare.net/plummer48/reporting-a-multiple-linear-regression-in-apa. [Accessed 28 November 2018].

Europe.eu (2018) Total population living in a dwelling with a leaking roof, damp walls, floors or foundation, or rot in window frames of floor for Year 2017. Available at: http://appsso.eurostat.ec.europa.eu/nui/show.do?query=BOOKMARK\_DS-056148\_QID\_456CBD0C\_UID\_-3F171EB0&layout=TIME,C,X,0;GEO,L,Y,0;HHTYP,L,Z,0;INCGRP,L,Z,1;UNIT,L,Z,2;SEX,L,Z,3;AGE,L,Z,4;INDICATORS,C,Z,5;&zSelection=DS-056148UNIT,PC;DS-056148AGE,TOTAL;DS-056148INCGRP,TOTAL;DS-056148INDICATORS,OBS\_FLAG;DS-056148HHTYP,TOTAL;DS-056148SEX,T;&rankName1=HHTYP\_1\_2\_-1\_2&rankName2=UNIT\_1\_2\_-1\_2&rankName3=AGE\_1\_2\_-1\_2&rankName4=INDICATORS\_1\_2\_-1\_2&rankName5=SEX\_1\_2\_-1\_2&rankName6=INCGRP\_1\_2\_-1\_2&rankName7=TIME\_1\_0\_0\_0&rankName8=GEO\_1\_2\_0\_1&sortC=ASC\_-1\_FIRST&rStp=&cStp=&rDCh=&cDCh=&rDM=true&cDM=true&footnes=false&empty=false&wai=false&time\_mode=ROLLING&time\_most\_recent=false&lang=EN&cfo=. [Accessed 28 November 2018].