STATASTICS FOR DATA ANALYTICS

Shubham Siras Katmusare

School of Computing

National College of Ireland

x19195117@student.ncirl.ie

1. **TIME SERIES ANALYSIS**
2. **Objective:**

The objective is to implement the time series analysis for the “Population change - Demographic balance and crude rates at national level [DEMO\_GIND]” data for Netherland as a geopolitical entity over the years 1960-2020. This analysis will focus on the most appropriate tests needs to be considered and implemented to get the optimum level results for the considered dataset. Furthermore, the model built will be used to forecast three periods ahead with the optimal prediction ranges.

1. **Data context and Sources:**

The data used for the analysis is from the geopolitical domain. The data has time series data over the span of 61 years from year 1960 to year 2020. The dataset posses the annual time frequency. The data for respective years seems to be consistent over the span and has 61 records. All the data is numeric and requires no data pre-processing. The dataset is collected form the Eurostat[1].

1. **LOGISTIC REGRESSION**
2. **Objective:**

The objective is to implement the logistic regression model for the “Latino’s employment status in the US” data. As a classification algorithm, classified under the Supervised learning technique, binary logistic regression is popularly used. This regression approach is used to estimate the likelihood of observational values occurring in one of the two types of the dichotomous dependent variable (i.e. two dependent values).

1. **Data context and Sources:**

For logistic regression implementation, the dataset is selected from the Eurostat [2] based on Latino’s employment status in US these days. There are 150 records along with the 7 independent variables namely “Parents”, “marital”, “religion”, “age”, “income”, “sex” and “Primary Language” which will influence the target variable “employment”. As a part of pre-processing, the target variable is converted and labelled as the “0” and “1” which represent the employment status.

1. **Assumptions:**
   1. Binary dependent variable – The target variable is converted to the “0” and “1” signifies that the “Employment” (target) variable is dichotomous.
   2. Sample size – The dataset must contain the ample amount of data in order to avoid overfitting. The dataset contains 150 records states that signifies the minimum sample requirement of 15 records per feature.
   3. Multi Collinearity and Collinearity– From the fig.1, it is concluded that none of the feature has correlation factor more than 0.7 so independent variable are not significantly dependent and affects the target variable individually.

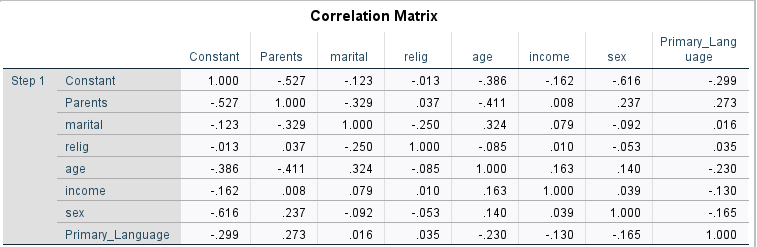


Fig1: Correlations

1. **Diagnostics and assessment:**

The diagnostics is carried out with total 7 independent variables and 1 binary target variable. “Parents”, “marital”, “religion”, “age”, “income”, “sex” and “Primary Language” are provided to predict the “employment” as target in the logistic regression.

The diagnostic carried out using the Omnibus test of model coefficients, model summary, Hosmer-Lemeshow goodness-of-fit, classification table and CI for exp(B) is kept in 95%.

1. Omnibus test:

This Omnibus test conducted to check the model's performance. The degree of significance here is observed as 0.002, which is less than 0.05, demonstrate the importance of statistical measures. Fig.2 depicts the omnibus evaluation of model coefficients.

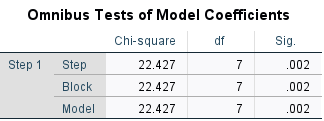


Fig 2 - Omnibus test of model coefficients

2. Model Summary:

The importance of distraction in the target variable can be found in this Model Summary table when forecasting the performance. Using Cox & Snell R Square and Nagelkarke R Square values, this can be achieved, where the values are observed as 0.170 and 0.273 respectively. This performance shown in fig.3, has been used to conclude that the expected value is anywhere between 17% and 28% different from the actual value.

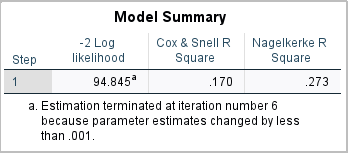


Fig3 – Model Summary

3. Hosmer and Lemeshow Test:

The Hosmer and Lemeshow test determine the importance of meaning that satisfies the principle of the Goodness-of-fit assumption. Fig.4 indicates that the significance value must be more than 0.05, which is 0.752 here, after conducting the analysis. This explains the existence of a correlation between the predictor variable and the dependent variable, and the dataset fits well with the model.

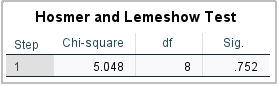


Fig4. Hosmer and Lemeshow Test

4. Classification Table:

The Classification Table refection of the tabular format of the Confusion Matrix, which is used to verify the model's accuracy. The model evaluated on the data set considered has an accuracy of 84.20 percent. The below fig.5 , summaries the model accuracy for target variable "Employment".

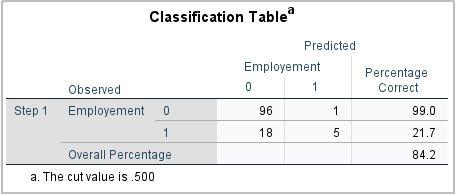


Fig.5 – Classification Table

5. Influence of Independent Variables:

In Fig.5, the value of “B” represents the contribute of the independent feature in deduction of the target variable.

With reference to the model trained in this analysis concludes that the feature “Relig” (Religion) is more inclined towards the prediction of the “Employment” and “sex” is less prone for the same.

So, the final logistic regression equation comprises as following which can be used to predict the further more results on same dataset.

log(p/1-p) = 6.071 + (-9.70) \* Parents + (-0.123) \* marital + (-0.002) \* relig + (-0.050) \* age + (-0.005) \* income + (-1.484) \* sex + (-0.038) \* primary\_language

1. **CONCLUSION:**

This research was conducted to evaluate the efficacy of various features "Latino employment status in the US" . The binary logistic regression model review concludes this with 84.2 percent accuracy and "Religion" as an efficient measure to predict the status of employment.

**V. ASSUMPTIONS:**

The following list of assumptions were considered during the multiple regression analysis.

*A.* Homoscedasticity test :

Check for each value of the predictors variance of the error term should be constant.

*B.* Independence of errors

*C.* Independence of each data point for the dataset

*D.* Predictor variables must be independent of the error

*E.* Regression residuals follow Normally distribution curve.

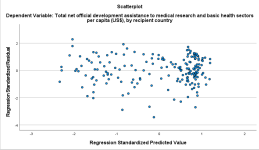
*F.* Absence of multicollinearity between independent variables.

*G.* Absence of influential data points.

*H.* There is linear relationship between predictors and target variable.

**VI. ANALYSYS AND RESULTS:**

Post data transformation, the multiple regression Model analysis is generated on the SPSS utility software tool.

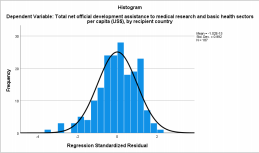


*Fig2 : Scatter plot of standardized residual Vs standardized predicted value*

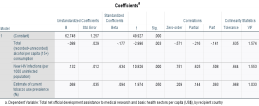
From the figure 2, the scatterplot between the standardized residual Vs standardized predicted value do not follow any relationship and looks noisy so it can be concluded that the assumption of homoscedasticity is followed.

*Fig3 : Model Summary*

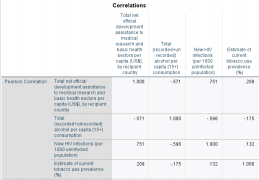
The Durbin-Watson value in our analysis is 2.246 from the Figure 3, which is closer to 2 and hence it can be concluded

that there is no autocorrelation between the errors. *Fig 4 : Frequency Vs Regression residual*

The above figure 4, it represents the relation between the frequency of the data points and regression standardized residual which follow the normal distribution curve. This is deduction of assumption of regression residuals following the normal distribution curve.

*Fig 5 : Coefficients*

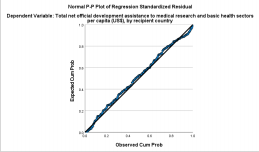
The figure 5, the statistical VIF value sustains around 1, precisely is do not exceed 10 which represents that there is absence of multicollinearity between predictor variables.

*Fig 6 : Correlations*

From the figure 6, the correlations between all the features can depicted.

Furthermore, the hook’s distance values do not exceed 0.01 for any of the data entry points for the given dataset so it can

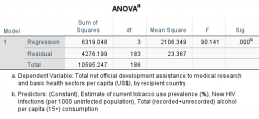
be concluded as there are no outliers in the dataset which may affect the prediction.

 *Fig 7 : P-P plot of regression residual*

From the figure 7, it is demonstrated that there is linear relationship between the target variable and predictors and both are linearly dependent.

Above analysis deduce that the dataset follows all the assumptions to consider the multiple linear regression analysis to apply on the population data.

In figure 3, the R value is 0.772 which shows a good correlation between the predicted and the present values of the dependent or target variable. The co-efficient of determination ( R Square) conclude that the 59.60 percent of the variance in target variable can be deduce using predictors.

*Fig7 : ANOVA table*

From the ANOVA table, p value is 0.000 which is less than .05. further, it can be depicted that all the predictors are significantly influencing the target variable. Figure 5, coefficients also confirms that for “Total Alcohol Consumption Per Capita” , “Communication Disease By Country” and “Estimate Of Tobacco Use Prevalence” with significance value less than .05 is better fit and statistically significant to predict the target variable “Total Net Official Development Assistance”. Interestingly, the predictor “Total Alcohol Consumption Per Capita ” has the negative co-efficient of 0.088 which is inversely affecting the target variable.

**VII. CONCLUSION :**

The Multiple Regression Analysis carried out on the above dataset can be used to predict the “Total Net Official Development Assistance” for the different countries for the year

2018-19. The analysis confirms that “Total Alcohol Consumption Per Capita” , “Communication Disease By Country” affects slight significantly than “Estimate Of Tobacco Use Prevalence” to target variable “Total Net Official Development Assistance”.

This model is affirmatively inclined towards all the assumptions considered for the dataset so this model can be implemented on the population data of this dataset.

**VIII. REFERENCES :**

1. [1] https://apps.who.int/gho/data/node.main 2. [2]https://apps.who.int/gho/data/node.main.SDG 3B?lang=en

3. [3]https://apps.who.int/gho/data/node.main.SDG 35?lang=en

4. [4]https://apps.who.int/gho/data/view.main.SDG 33v?lang=en

5. [5]https://apps.who.int/gho/data/node.main.SDG 3A?lang=en