Statistics for Data Analytics : CA 1

Multiple Regression

and

Time Series Model

Submitted to:

Tony Delaney

Submitted by:

Sobil Dalal

x19148496

Master of Science in Data Analytics

National College of Ireland

Table of Contents

[I. Multiple Regression Model 3](#_Toc36767125)

[A. Introduction 3](#_Toc36767126)

[B. Objective 3](#_Toc36767127)

[C. Data description 3](#_Toc36767128)

[D. Data transformation/cleaning 4](#_Toc36767129)

[E. Tools and Libraries used for analysis 4](#_Toc36767130)

[F. Assumptions 4](#_Toc36767131)

[1. Sample size 4](#_Toc36767132)

[2. Non-zero variance 5](#_Toc36767133)

[3. Outliers and Normality of dependent variable 5](#_Toc36767134)

[4. Correlation & linear relationship 5](#_Toc36767135)

[G. Modeling and verification/diagnostic of model 6](#_Toc36767136)

[1. Model Summary 6](#_Toc36767137)

[2. Coefficients 7](#_Toc36767138)

[3. Anova and F-test 7](#_Toc36767139)

[4. Correct functional form 7](#_Toc36767140)

[5. Homoscedasticity and constant variance of errors 7](#_Toc36767141)

[6. No autocorrelation between errors 8](#_Toc36767142)

[7. Errors are normally distributed 8](#_Toc36767143)

[8. Absence of multicollinearity 8](#_Toc36767144)

[9. No influential data points 8](#_Toc36767145)

[H. Conclusion 9](#_Toc36767146)

[II. Time Series Model 10](#_Toc36767147)

[A. Introduction 10](#_Toc36767148)

[B. Objective 10](#_Toc36767149)

[C. Data description 10](#_Toc36767150)

[D. Data transformation/cleaning 10](#_Toc36767151)

[E. Tools and Libraries used for analysis 10](#_Toc36767152)

[F. Pre-modeling checks 10](#_Toc36767153)

[1. Trend, seasonality, level 10](#_Toc36767154)

[2. Identify auto regressive or moving average model 11](#_Toc36767155)

[3. Differencing and stationary covariance 11](#_Toc36767156)

[G. Modeling and verification/diagnostic of model 11](#_Toc36767157)

[1. Model Summary 11](#_Toc36767158)

[2. Normal distribution of residuals 12](#_Toc36767159)

[3. No autocorrelation for all lags 12](#_Toc36767160)

[H. Conclusion 12](#_Toc36767161)

[References 13](#_Toc36767162)

# Multiple Regression Model

## Introduction

Multiple regression is a statistical modeling technique which helps to establish relationship between a dependent/response variable and more than one independent/predictor variable (Lantz, 2015, p. 172-181). Multiple linear regression technique is a type of multiple regression used in this section to build relation between dependent and other independent variables. The relationship of the independent variables is considered to be linear to the dependent variable (James, 2013, p. 59-71).

Demographic and socio-economic statistics like “cellular subscription” as the dependent variable and “gross national income per capita”, “population median age”, “population living in urban areas”, “total population” as four independent variables, have been used for multi-linear regression modeling.

## Objective

The analysis will help in predicting the percentage of cellular subscribers based on different

demographic and socio-economic factors of a country.

## Data description

The analysis is done based on below variables with the following description for the year 2012:

1. Dependent variable:
   1. Cellular subscription (cellular)
      * + Continuous numerical value ranging from 5% to 186%.
        + Measurement is in terms of percentage with respect to overall population.
        + Source of the data set is extracted from (Anon, 2014).
2. Independent variables:
   1. Gross national income per capita (income)
      * + Continuous numerical value ranging from $390 to $66960.
        + Source of the data set is extracted from (Anon, 2014).
   2. Population median age (popul.med.age)
      * + Continuous numerical value ranging from 15.04 to 45.53 years.
        + Source of the data set is extracted from (Anon, 2014).
   3. Population living in urban areas (popul.urban)
      * + Continuous numerical value ranging from 11% to 100%.
        + Source of the data set is extracted from (Anon, 2014).
   4. Total population (popul.in.thous)
      * + Continuous numerical value ranging from 89 to 1390000 (thousand).
        + Source of the data set is extracted from (Anon, 2014).

## Data transformation/cleaning

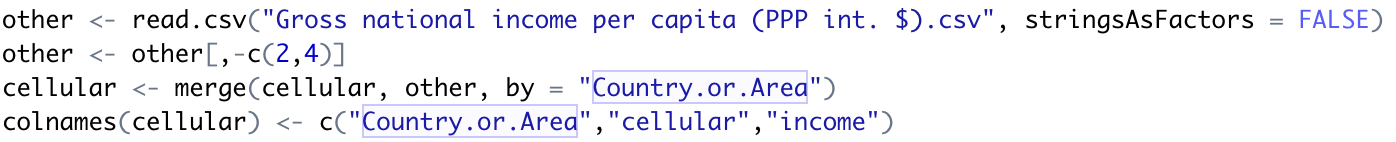
All the data files (.csv) are merged by the common column “Country.or.Area” as per Figure-1 snippet in R language to create a data frame “cellular”.

Figure: 1

Summary of the data frame is in Figure 2.

A screenshot of a cell phone

Description automatically generated

Figure: 2

## Tools and Libraries used for analysis

* + 1. R Language
    2. R libraries:
       1. psych
       2. leaps
       3. car

## Assumptions

### Sample size

According to Tabachnick & Fidell (2013) the sample size should be adequate of multilinear regression and follow

N > 50 + 8M where N = Sample size, M = Number of factors

Therefore, N = 162 > 82 that is derived as (50 + 8\*4) is satisfactory for the model.

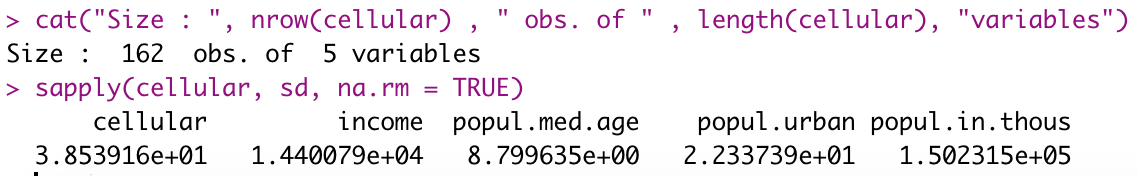


Figure: 3

### Non-zero variance

From the Figure-3, it can be seen that the variance is non-zero for all the variables in the sample.

### Outliers and Normality of dependent variable

|  |  |
| --- | --- |
| From the box plot in Figure-4, it is clearly apparent that there are no outliers in the dependent variable “cellular”.  Moreover, the histogram plot in Figure-4 also suggests that the data is distributed significantly normally. | Figure: 4 |

### Correlation & linear relationship

|  |
| --- |
|  |

Figure: 5

From Figure-5, it can be inferred from Pearson’s product-moment correlation (r) and scatterplot that cellular has positive linear significant relationship with income, popul.med.age, popul.urban and non-significant negative relationship with popul.in.thous.

Moreover, it can also be seen that, there might be some relation between independent variables as-well. Therefore, the optimal model must contain interaction between them.

The preliminary assumptions are met for modeling multilinear regression model. Analysis are further done on the model fitted on the sample data.

## Modeling and verification/diagnostic of model

### Model Summary

For building the best model to fit the sample data, approaches like best predictors using “regsubsets” function, “backward step" and “forward step” using R function step and manual addition or deletion of independent variable was performed.

|  |  |
| --- | --- |
| According to the concept of parsimony, from Figure-5 it is noticeable that number of predictors after 3 provide almost the similar result and thus considering additional parameter to the model will not be beneficial.  On top of that, addition of more factors to the model might result in overfitting to the sample. | Figure : 5 |

From Figure-5, it is clear that among all the independent variables popul.med.age, popul.urban and popul.med.age:popul.urban variables will provide most of the information to predict the cellular dependent variable.

Figure-6, contains the model summary. The accuracy of the model is inferred from adj. R2.

|  |  |
| --- | --- |
| R2 value is 0.5778. This implies that 57.78% of the variation in the value of cellular subscription can be explained by population median age, population living in urban area and interaction between both.  Adjusted R2 value of the model is 0.5696 which provide estimate value for the sample’s population by adjusting the R2 to balance the effect of the number of predictors. | Figure: 6 |

### Coefficients

Coefficients for the model are -83.026, 5.86, 2.28 and -0.066 for Intercept, popul.med.age, popul.urban and popul.med.age:popul.urban respectively. The t-test p value (Pr(>|t|)) for all the coefficients are less than 0.05 suggesting that all the coefficients are significant for predicting the value of cellular (Figure-6).

### Anova and F-test

F-statistic is 70.71 on 3 variables, 155 degree of freedom and the p value is also less than 0.05. This helps in rejecting the null hypothesis that the coefficients are 0 of independent variables. Therefore, the independent variables better fit the dependent variable than intercept only model (Figure-6).

### Correct functional form

Residual vs fitted plot in Figure-7 explains that the model is following linear form and doesn’t have a pattern, thus meets the assumption of multi-linear regression.

|  |
| --- |
| Figure: 7 |

### Homoscedasticity and constant variance of errors

|  |  |
| --- | --- |
| Scale-location plot in Figure-7 shows most of the residual data points in scatterplot are concentrated at 0 and have fairly rectangular shape. | Figure: 8 |

Additionally, non-constant variance score test is also non-significant (> 0.5), thus suggests residuals have fairly constant variance (Figure-8).

### No autocorrelation between errors

|  |  |
| --- | --- |
| The p-value of Durbin Watson test is non-significant, thus according to alternate hypothesis, the autocorrelation is absent in the residuals (Figure-9). | Figure: 9 |

### Errors are normally distributed

|  |  |
| --- | --- |
| As per Normal Q-Q plot in Figure-7 and histogram plot of residuals in Figure-10, the residuals are normally distributed. | Figure: 10 |

### Absence of multicollinearity

|  |
| --- |
| Figure: 11 |

The VIF (Variance Inflation Factor) values are 11.43, 15.34, 38.99 of popul.med.age, popul.urban and popul.med.age:popul.urban respectively. Interaction term is present to handle collinearity of popul.med.age and popul.urban (Figure-11).

The VIF is greater than 10, which ideally should be less. Even after 5 iteration of removing the influential data point VIF was still greater than 10 and also affected other assumptions. Therefore, kept the VIF same as above.

### No influential data points

|  |
| --- |
| Figure: 12 |

From residual vs leverage plot (Figure-7) and summary of cooks.distance model (Figure-12), it is evident that the cook’s distance of all the data points is less than 0.1 and thus there are no influential data points to create disproportionate impact on the model.

## Conclusion

Multiple-linear regression model was applied to study the relationship of factors to predict “cellular subscription”. Among four independent variables, “population median age”, and “population living in urban areas”, along with interaction variable of both are significant enough to determine the dependent variable. All the above-mentioned assumptions of dependent, independent and residuals have been satisfied, except absence of multicollinearity for the multilinear regression model.

The adjusted R2 of the model is 0.5697 which suggests including other independent variables might help to better predict the cellular subscription of the country.

# Time Series Model

## Introduction

Time Series is a sequence of value of a feature taken at a specific frequency such as annually, monthly, weekly. Gross Domestic Product (GDP) is a monetary value used to evaluate country’s economic growth. GDP is predicted based on previous year’s GDP value for the country Ireland. Exponential smoothing model and autoregressive integrated moving average (ARIMA) model have been used to estimate the GDP.

## Objective

The analysis will help in predicting the GDP of Ireland based upon previous year values.

## Data description

The analysis is done based on GDP variable with the following description:

* Source of the data set is extracted from (Anon, 2020).
* Numerical values ranging from 4.402e+09 to 3.827e+11 dollars.
* The frequency of the data is 1 that is annually from 1970 to 2018.
* Total observations are 49.

## Data transformation/cleaning

The data is transformed and cleaned with following steps:

1. The data file (.csv) is imported with two columns that is year and GDP.
2. GDP column is divided by billion to convert to billion term and sorted in ascending order by year.
3. A new time series variable is created for the analysis. Below is the summary (Figure-13).

|  |
| --- |
| Figure : 13 |

## Tools and Libraries used for analysis

* + 1. R Language
    2. R libraries:
       1. fpp2
       2. tseries

## Pre-modeling checks

### Trend, seasonality, level

As per the Figure-14, the plot of time-series of GDP is following a trend and no seasonality (frequency is 1). It is furthermore clear from smoothing plot with different q values (3,5,7).

### Identify auto regressive or moving average model

In Figure-15, the Auto Correlation Function (ACF) slowly decreases to 0 and Partial Auto Correlation Function (PACF) have spikes at 1st and 4th lag. This suggests that the model might be auto regressive. It is further evaluated in next section.

|  |  |
| --- | --- |
| Figure: 14 | Figure: 15 |

### Differencing and stationary covariance

|  |  |
| --- | --- |
| For ARIMA modeling the time series should be covariance stationary. Differencing the series by 2, generates the same. It is verified by Dickey-Fuller test (Figure-16) by rejecting the null hypothesis as the p-value is 0.46 which is insignificant. Figure-15 plot also verifies the same. | Figure: 16 |

## Modeling and verification/diagnostic of model

### Model Summary

For building the best model to fit the sample data, modeling techniques like exponential smoothing and ARIMA using R function holt (damped = True/False), auto ets, arima, auto.arima were performed.

The best model obtained using exponential smoothing is holt damped model with RMSE 12.77, whereas overall the best model is auto regressive ARIMA with accuracy of 11.55 RMSE. Further summary is below:

|  |  |
| --- | --- |
| ARIMA model has p value of 4 and d value of 2. It has AIC of 376.17.  This suggests that the future GDP value will have lingering effects of 4-year preceding GDP value with coefficients -0.2562, -0.2959, -0.0196, 0.3131 respectively (Figure-17). | Figure : 17 |

### Normal distribution of residuals

Residual plot in Figure-19 demonstrates that the residuals are significantly normally distributed.

### No autocorrelation for all lags

|  |  |
| --- | --- |
| All the values are near 0 in ACF vs lag plot in Figure-19 which suggests that autocorrelation is zero for all the lags.  Further, this is verified with Box-Ljung test with null hypothesis that there is no autocorrelation and the model is fit. P-value is 0.689 which is insignificant, thus null hypothesis can’t be rejected (Figure-18).    Figure: 18 | Figure : 19 |

## Conclusion

Exponential smoothing model and ARIMA model was applied to estimate the GDP. Among all the models, auto regressive ARIMA with order 4,2,0 was the best model to predict GDP with accuracy of 376.17 AIC and 11.55 RMSE. All the assumptions and verification of model were performed successfully.

# References

Lantz, L. (2015) Machine learning with R. 2nd edn. Birmingham - Mumbai : Packt.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013) An Introduction to Statistical Learning. Birmingham - New York : Springer.

Anon, 2014. Undata | Record View | Cellular Subscribers (Per 100 Population). [online] Data.un.org. Available at: <http://data.un.org/Data.aspx?d=WHO&f=MEASURE\_CODE%3aWHS9\_CS> [Accessed 23 March 2020].

Anon, 2014. Undata | Record View | Gross National Income Per Capita (PPP Int. $). [online] Data.un.org. Available at: <http://data.un.org/Data.aspx?d=WHO&f=MEASURE\_CODE%3aWHS9\_93> [Accessed 23 March 2020].

Anon, 2014. Undata | Record View | Population (In Thousands) Total. [online] Data.un.org. Available at: <http://data.un.org/Data.aspx?d=WHO&f=MEASURE\_CODE%3aWHS9\_86> [Accessed 23 March 2020].

Anon, 2014. Undata | Record View | Population Living In Urban Areas (%). [online] Data.un.org. Available at: <http://data.un.org/Data.aspx?d=WHO&f=MEASURE\_CODE%3aWHS9\_96> [Accessed 23 March 2020].

Anon, 2014. Undata | Record View | Population Median Age (Years). [online] Data.un.org. Available at: <http://data.un.org/Data.aspx?d=WHO&f=MEASURE\_CODE%3aWHS9\_88> [Accessed 23 March 2020].

Anon, 2020. Undata | Record View | Per Capita GDP At Current Prices - US Dollars. [online] Data.un.org. Available at: <http://data.un.org/Data.aspx?d=SNAAMA&f=grID%3a101%3bcurrID%3aUSD%3bpcFlag%3a1%3bitID%3a9> [Accessed 23 March 2020].

Tabachnick, B., & Fidell, L. (2013). Using Multivariate Statistics. Boston, MA: Pearson Education Inc.