Estimating GDP based on Telecommunication

Omkar Ratnoji Tawade  
*MSc. Data Analytics*   
*National College of Ireland*Dublin, Ireland  
x19232136@student.ncirl.ie

# Objective

Telecommunication is an important part of developing and developed nations. It has a great influence on the economic growth of the nation. Telecommunication plays a big role in the business sector by increasing the productivity level. The objective of this analysis is to estimate the GDP of a particular country using telecommunication factors such as the usage of the internet, mobile, and telephone. I have also added the ‘population in urban’ factor in the dataset which I assume will be linear to all the above factors. The basic idea of this analysis is to check whether the countries with high GDP does invest more in telecommunications and vice versa.

# Description of variables

## GDP

The measure category of GDP in SPSS is a scale that suggests that this variable is continuous. In our analysis, it will be a dependent variable.

## Individuals using the Internet

It is also a continuous variable detected by the SPSS. It will be used as an independent variable in this analysis. Value is this variable is calculated as the number of internet users per 100 people.

## Mobile Subscribers

Mobile Subscriber is also a continuous variable and it will be used as an independent variable in this analysis. It is calculated as the number of mobile subscribers per 100 inhabitants.

## Telephone Subscribers

For regression mostly we use the continuous variables, so most of the variables will be continuous. It is also a continuous variable and will be used as an independent variable in the analysis. It is calculated as the number of telephone subscribers per 100 people.

## Population in Urban

This variable is included in the dataset based on the assumption that the population in urban utilizes the above factors more than rural for economic purposes. It is also continuous and will be used as an independent variable while building the model.

All the above data is collected and merged from the UN data portal. In this dataset, I have included the data of 115 nations.

There is no missing value in the dataset.

# Model building and diagnostics

## Examining Data

It is the first step of model building. We need to check the data stored in these variables are credible or not. For this check we take will analyze the descriptive statistic of all variables. If we check the last row of the output table, it shows that valid N (listwise) is 115 which suggests that there no missing value in the data.

Table

Description automatically generated

Now we will explore our dependent variable which is GDP. According to SPSS output, GDP is not normally distributed but it will be not an issue since we check the normality of the prediction error variable to analyze the model efficiency. I have also analyzed the boxplot of the same variable and can observe that there are many outliers. It was obvious because there are few developed nations which have high GDP than developing nations.

Chart, bar chart

Description automatically generated

Similarly, I analyzed all independent variables using the box plot and observed that the population in urban has a similar box plot like GDP. But after comparing the outliers of GDP and population in urban I cannot assume that country that has more population in urban has higher GDP. Other independent variables such as Internet subscribers, mobile subscribers, and telephone subscribers had very few outliers.

After analyzing all independent variables, now we need to check the normality of the dependent variable. In our case, the result of normality was not significant. Below is the output of the normality test.

Table

Description automatically generated

As we can see the significant value of this test is very low so we cannot assume that our dependent variable is normal.

The graph below is the normal Q-Q plot of our dependent variable and by observing that graph we can conclude that values in our dependent variable are highly deviated from normal.

Chart, scatter chart

Description automatically generated

## Assumptions

Before proceeding to actual model building, we need to check a few assumptions[1] on our dataset. This will ensure that our dataset would qualify for multiple regression analysis or not.

1. The relation between IV and DV is Linear:

The first assumption of multiple regression is to check the relation between the independent and dependent variables, and it can be distinguished by a straight line. Below is the scatter plot of our dataset.

A picture containing photo, show, different, many

Description automatically generated

It is evident from the scatter plot that our dependent variable is not that linear with any independent variables. The distance between the points and the line in the above plot is known as residual distance. These residual distances must below to have a good model.

1. No multicollinearity:

To carry out the multiple regression, we must have to check the multicollinearity between our independent variables. To build a good model this should be as low as possible.

Table

Description automatically generated

By analyzing the above output, we can conclude that there is collinearity between the Internet and Telephone. Normally the significant value should not go above 0.7. If the value is above 0.7 then it is sure that there is multicollinearity present. Another way to check multicollinearity is by observing the output of coefficients after model building. Below is the table of coefficient,

Table

Description automatically generated

To meet this assumption, all VIF values must be less than 10 and tolerance value must be greater than 0.2

1. The values of residuals are Independent:

It can be explained that each case should be different from other or individual data points should be independent of others. To verify this assumption, we need to observe the output of the Durbin-Watson statistic.

Table

Description automatically generated

In our study, the Durbin-Watson statistic value is 2.093. A good model has a value very close to 2. If this value goes above 3 or below 1 then it will impact the analysis.

1. The variance of residual is constant:

It can be described as the constant spread of residuals at each point of the predictor variable.

The amount of residual error should be the same at every point in the model. This is also known as homoscedasticity. We can check this assumption by analyzing the scatter plot of the standardized value of the model against the standardized residuals. If the plot is funnel-shaped, then the assumption is violated. If the model produced is good, then we can observe the same spread of residual error as the predictor value across the x-axis increases.

Chart, scatter chart

Description automatically generated

As we can see clearly that this scatters plot is taking the funnel shape, so the assumption is being violated here.

1. The values of the residuals are normally disturbed:

It can be verified by plotting the P-P plot of the model. Residuals will lie along a normal line. The closer the residuals are to that normal line the more efficient is the model. In our study, the plot generated is not the ideal plot we need to have.

Chart, line chart

Description automatically generated

By observing the above graph, we can conclude that residuals are not distributed ideally across the normal line. Data points are hardly near or touched the normal line. We can conclude that our dataset violated this assumption.

1. No influential cases biasing the model:

It can be tested by analyzing the cook distance of the independent variable. Influential data points and significant outliers can have unnecessary influence over the model. No data point should have a cook distance greater than 1, otherwise, it would be problematic. Below is the plot of cook distance of our model

Chart

Description automatically generated

In our study, we can analyze that only one country (China) has a cook distance greater than 1. So, we can neglect this data point to make our model more efficient.

After all these assumptions are satisfied then OLS estimators (ß1OLS and ß2OLS) are the best linear unbiased estimators. These estimators are then used to determine the regression coefficients.

## Model Parameters:

There are many parameters we need to check to conclude that model produced is good. Below is the example of multiple regression output in the SPSS.

Table

Description automatically generated

Table

Description automatically generated

Model: This indicates the number of models produced. SPSS allows producing multiple models in single regression command.

R: It is the correlation value between the observed and predicted value of the dependent variable. It is the square-root of R-squared.

R-square: It is the ratio or amount of variance in the dependent variable which can be predicted from the independent variables. It is a measure of the strength of association. R-square is also called a ‘coefficient of determination’.

Adjusted R-square: When more and more independent variables are added into the model then the R-square value of the model will keep on increasing. It will temp to add more independent variables but adding more independent variables will introduce overfitting. Adjusted R-square is used to determine how significant the correlation is after adding the independent variables.

Std. Error of the Estimate: It is the standard deviation of the error term.

Coefficient: These are the values required for the regression equation to predict dependent variables using the independent variables. There is a column of ‘significance’ beside the coefficients which suggests that the value of coefficients is significant or not. If coefficients have significance below 0.05 then that coefficient should not be considered in the equation.

# Model Building

The dataset contains 5 variables of which I am considering ‘GDP’ as the dependent variable and the rest of the other variables are independent variables. I will start the model building process by observing the descriptive statistics of all variables. This step can be used as validating the data to check whether any garbage value is present in the data. Normally it can be detected by observing the min-max value of the variables.

We can also visualize the outliers using the boxplot. Below is the box plot of each variable.

Chart, bar chart

Description automatically generatedA picture containing chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Since it is population data so it will be not normally distributed data. From the plot, we can understand that there are too many outliers in population and GDP data. But we cannot exclude this outlier from our data since they are real data. For example, in the population data, China and India will be always an outlier since their population is much higher than the mean population of countries. So, I am not excluding any case from the data.

Now I will observe the frequency plot of each variable to examine how they are distributed, following are the plot of each variable.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generated

The frequency distribution of variables should be normal. By observing the frequency plots, we can conclude that only mobile\_subscribers have a normal distribution. Telephone and Internet subscribers also have normal distributions but not as clear as mobile subscribers. Whereas GDP and Population do not have a normal distribution, and these are the variables that will affect the efficiency of the model. We can take the log transformation of GDP and Population which will make the distribution of these variables normal distributions. But I will be proceeding with all variables without any transformations.

Now I will check whether my independent variables are correlated or not. In SPSS there is an option for correlate under the analyze tab. The output of correlation is as follows,

Table

Description automatically generated

From the above table, it is evident that telephone and Internet are highly correlated since the value of correlation is above 0.7. To avoid correlation, we can exclude one of the variables. In this case, we can exclude the telephone\_subscribers assuming that Internet\_subscribers will be a more important factor than telephone\_subscribers while estimating the GDP.

But for the current analysis, I am considering both variables for model building.

Now we can move for actual model building and it’s quite simple in SPSS just we have to use the linear regression function under the analyze tab. We have to declare which is our dependent and independent variable in that dialog box and also, we can check and uncheck the parameters that we can observe in the output. So below is the output of the multiple linear regression,

Table

Description automatically generated

Now by observing the Adjusted R-squared value we can assume that model produced is not a good model. So, we need to check the assumptions now to confirm whether we can produce a better model or not. We had discussed all six assumptions earlier and concluded that only five out of six assumptions were violated. Following are the violated assumptions,

1. The relation between IV and DV is Linear (In our study it was not linear).
2. No multicollinearity (In our study, Telephone and Internet variables were correlated).
3. The variance of residuals is constant (In our study the graph of residual and predicted value was fan-shaped which suggested that variance was not constant.)
4. The values of residuals are normally distributed (In our study residual points hardly touched the normal line which suggests that the value of residuals was not normally distributed.)
5. No influential cases biasing the model (In our study, we observed that population and GDP is high for a country like China which act like as an outlier, and cook distance for the same case was higher than 1 which suggest that this particular case is responsible for biasing the model)

So, to produce a better model we will need to perform the transformation on our variables like log transformation on GDP and population variable to get that normal distribution. We will need to neglect the telephone variable as it is highly correlated to the Internet variable this will solve our multicollinearity problem. For a better model, we can neglect ‘China’ from the list as we observed that it is highly influential for biasing the model. I believe after performing all the above steps we can have a better model. But for now, our model will be bad at predicting the GDP value based on independent variables.

In multiple linear regression, we get the coefficient of independent variables as the output. Using these coefficients, we can predict the value of the dependent variable. Below is the output of coefficients,

Table

Description automatically generated

By observing the above table, we can conclude that population and telephone are two variables that have coefficients that are significant enough to use in a linear equation. We cannot use the coefficient of Internet and Mobile in the equation since the coefficient value of these variables is not significant. We can represent the linear equation of our model as follows.

GDP = ß1(Population\_Urban) + ß2(Internet\_Subscribers) + ß3(Mobile\_Subscribers) + ß4(Telephone\_Subscribers)

References

1. Statistics Solutions. (2018). Assumptions of Linear Regression.at *www.statisticssolutions.com/assumptions-of-linear-regression/*