Regression Analysis for Social Research using MS Excel

Dr Sudhir Kapoor*
Dr. Rashmi Kapoor**
*Associate Professor, Department of Statistics, Hindu College, University of Delhi, Delhi-110007.
** Assistant Professor, Department of African Studies, University of Delhi, Delhi-110007

Data whether qualitative or quantitative are aggregate facts but only when processed and understood in the context that they convey meaning. Once data is collected, next step is cleaning or refining of data to extract the relevant information and make the data meaningful. To mine out useful information, statistical analysis tools are used. Statistical analysis can be used for both qualitative data and quantitative data. The present paper will attempt to provide basic steps regression analysis for social research using MS Excel.

Social research or social science research deals with the study of society, human behaviour and social interactions. Social research employs the similar methods of scientific inquiry as used in natural sciences. As a matter of fact the method of scientific inquiry is same for all sciences. Pearson (1911, 12) said that, "the unity of all sciences consists alone in its methods, not its material." Similarly, Bacon (cited in Thompson, 1911, 81) states that, "the division of sciences are not like different lines that meet in one angle but rather like the branches of trees that join in one trunk." Also, science is not inherently different from the way one thinks in everyday life. Albert Einstein said that "all science is nothing more than a refinement of everyday thinking." In the similar sense, Thomas Huxley said that, "Science is simply commonsense at its best." It is the scientific method that differentiates a science from commonsense.

In social research, a major challenge is objective investigation of social phenomena. The researcher's perception can hinder in objective understanding. Then the indicators under study are not directly measurable. Also, the reliability of obtaining the same results in social research is difficult. The inconsistency of social behaviour may be due to other exogenous and endogenous factors exerting influence. Experimentation that determines the validity and reliability in the scientific research is not possible in social research as in the latter variables cannot be manipulated to identify the definite cause of the phenomenon. The empirical studies in social science can at best determine the dependence of one variable on one or more variables or indicate the magnitude of association between the variables.

Social science students often shy in dealing with numbers and quantitative analyses. Probably they find it too complicated. But with basic understanding of statistical analysis, the different software packages available have empowered researchers to delve in the realm previously circumvented. The software packages have made managing, treating and analyses of data convenient and provide results in the format that are meaningful and interpretable.

There are several data analysis tools available for organization, analysis and graphical representation of data. NVIVO and QDA Miner are good for qualitative analyses whereas Stata, SPSS, R and Excel are some of the packages and tools easily available for organising, computing, analyzing, and predicting quantitative data. Stata and SPSS are paid statistical packages but R is a free programming language that needs to be learnt, though not difficult. MS Excel is most easily accessible software programme that organizes, formulates and analyses data

in order to discover trends and patterns. This spreadsheet program is also a data management tool that also gives enables graphical representation for summary information of data.

Statistical analysis has become an integral part of analysis of both qualitative and quantitative data that has been facilitated by tools and programmes available. Statistical analysis helps to discover trends and underlying patterns in the sample population and on that basis make predictions for the whole population. In the present paper statistical data analyses of one social research topic will be dealt with. First correlation will be determined, then regression and finally ANOVA will be computed. Correlation is a statistical method used to analyse relationship between a dependent variable and one or more independent variable to determine the association between the dependent and independent variable/s if any and also determine the strength of the relationship between the variables. Regression model helps to determine the model for the relationship between the dependent variable and independent variable by determining the function that fit the values of independent variables that vary linearly. Linear regression enables prediction or estimation of the value of dependent variable from the selected value of the independent variable or explanatory variable. ANOVA is analysis of variance.

In this paper, statistical modeling will be used to estimate the relationship between female labour force participation rate (FLFPR) and per capita GDP using MS Excel. The variable FLFPR is the factor to be analyzed and predicted and is called the dependent variable. Then there are other factors called the independent variables that are perceived to impact the dependent variable. There can be more than one independent factor that explains the variation in the dependent variable. In the present example, there is only one independent variable, that is, per capita GDP and hence simple linear regression model will be used. The broad theme is the relationship between the economic growth and women empowerment will be analyzed. Taking GDP Per Capita as a proxy for economic growth and FLFPR as an indicator for women's empowerment, the research question to be studied will be 'does the change in the per capita GDP of a country affect FLFPR?' The hypothesis then is 'FLFPR in the age group 15-64 increases with the increase in per capita GDP'. The FLFPR in the age group 15-64 is the dependent variable (Y) whereas per capita GDP is independent variable (X). The Table 1 gives a hypothetical data of the two variables.

Table 1: The FLFPR and per capita GDP (X) (2010-2019).

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
FLFPR	26	25	24	24	23	23	23	22	22	22
GDP	1300	1400	1450	1550	1750	1750	1870	1990	2080	2150

H_0 (null hypothesis): The FLFPR in the age group 15- 64 (Y) is not affected by the change in the per capita GDP (X).

H_1 (alternative hypothesis): The FLFPR in the age group 15-64 (Y) is affected by the change in the per capita GDP (X).

Correlation Coefficient (r)

Correlation between two variables indicates that the change in the value of one variable tends to effect the variation in the other variable. In other words the two variables tend to change together. Correlation coefficient indicates the strength of the linear relationship between the two variables and whether it is statistically significant or not. The range of correlation coefficient is between -1 to 1 but closer to zero implies a weak linear relationship. The point to be emphasized is that correlation coefficient is a constant not affected by the units of variable meaning that the units of variables in no way influence the determination of correlation. It by no way suggests any cause-effect relation.

Determination of correlation coefficient using 'Analysis Toolpak' in MS Excel. This 'Toolpak' is inbuilt in Excel but has to be enabled manually. For that follow the steps given below:

1. Open a Excel file, click 'Option', Select 'Add-ins'
2. Click 'Add-ins' on the left sidebar
3. In the dialogue box that appears at the bottom in the Manage drop-down box, select 'Excel Add-ins' and then click 'Go'
4. From the 'Add-ins available' Checkbox 'Analysis ToolPak' , then click 'OK'
5. Data Analysis tool will appear on the extreme right in the Data Tab of Excel.

To determine correlation analysis between the dependent variable and independent variable in Excel the steps given below were undertaken.

1. Go to 'Data' tab on Excel sheet, click 'Data Analysis'.
2. Select 'Correlation' from 'Analysis Tools', click 'OK'
3. Correlation dialogue box will open, for 'Input Range' select both dependent variable and independent variable with their labels'
4. Check box 'Labels'
5. Give 'Output Range' if you want analysis on the same Excel sheet, click 'OK'.
6. The following Table 2 appears

Table 2: Correlation Analysis output

	GDP	LFPR
GDP	1	
FLFPR	-0.95369	1

The absolute value of the correlation coefficient indicates the strength of correlation. Table 2 shows that correlation coefficient is -0.95369 indicating a strong negative correlation between FLFPR and per capita GDP. In this case, since it is negative, therefore, correlation is inverse correlation but the absolute value being high shows strong negative correlation between the two variables. Thus for this data if the FLFPR decreases then per capita GDP will increase.

Regression Analysis:

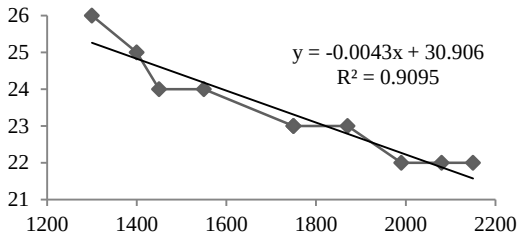
Regression analysis is set of methods that inform whether the independent variable is significant in influencing the dependent variable. It also estimates the relationship between the dependent variable and independent variable and determines the magnitude of change in the dependent variable when the independent variable experiences a variation. It also enables making predictions, projections and forecasting. There are several ways of conducting Regression analysis using Excel. It can be determined by:

a. Using linear regression formula: this requires sound understanding of theoretical statistical knowledge. This method will not be discussed here.

b. Steps for making a line graph using scatter plot and then analyzing the trend line

i. Insert data on the Excel sheet with independent variable data in the first column and dependent variable data in the second column so that independent variable is on x axis and dependent variable on y axis
ii. Select the data of the two columns
iii. Go to Insert tab, click the icon 'Scatter', select 'Scatter with straight lines and markers', click 'OK'
iv. A scatter plot with line will appear on the worksheet
v. Right click anywhere on the line obtained, dialogue box appears, select 'Add trendline'
vi. In the dialogue box that appears, check on 'Linear', then check on 'Display Equation on Chart', check 'Display R-squared value on the chart', click Close
vii. Regression equation and R-squared value appear on the chart with trendline

Graph1: Line graph for the Linear Regression



The equation obtained is Y= -0.0043 x + 30.906 (1)

From the line of regression the information regarding the estimated mean value of the dependent variable can be obtained for any selected value of the independent variable. Also, the graph also indicate the strength of the correlation. More closer are the point to the regression line the stronger is the relationship between the two variables.89

c. Using 'Analysis Toolpak' in Excel.

Steps given below were followed for using 'Data Analysis tool' to conduct regression analysis.

i. Click 'Data' tab in the toolbar of the Excel sheet, select 'Data Analysis'
ii. From the 'Analysis tools' select 'Regression', click 'OK'.
iii. Regression dialogue box will appear, put dependent variable data in the 'Input Y Range' and independent variable data in 'Input X Range'
iv. Checkbox the 'Labels' if labels are there on the top of dependent variable and independent variable ranges
v. By default 'Confidence Level' is 95% which is most commonly accepted value.
vi. In the 'Output option' select the option 'Output Range' for output to appear on the same page. Other options for output are also available.
vii. Click 'OK'
viii. Regression analysis output appears.

Table 3 gives the summary output obtained from the Excel. This summary output of regression analysis provides the information of fitting of the linear regression equation to the present data. Multiple R indicates the correlation which is 0.9536. Multiple R is an absolute value and does not indicate the positive or negative relationship as in the correlation computed in Table 2 using the inbuilt formula for correlation. In this summary output R Square is 0.90 that implies that about 90% of the given data fits the regression model obtained. Standard Error is 0.43, which is the average distance within which data points lay from the regression line. Since the Standard Error is low, therefore, most data points are on the regression line. Smaller the value of Standard Error better is the 'regression analysis model'. Both R Square and Standard Error show goodness of fit. Observations are the number of data entries considered.

Table 3: Summary Output

Regression Statistics	
Multiple R	0.953692824
R Square	0.909530003
Adjusted R Square	0.898221254
Standard Error	0.430654726
Observations	10

Table 4: ANOVA Table

ANOVA					
	Df	*SS*	*MS*	*F*	*Significance F*
Regression	1	14.91629	14.91629	80.42710629	1.90209E-05
Residual	8	1.483708	0.185463		
Total	9	16.4			

Table 4 is the ANOVA table or the analysis of the variance table that shows sources of 'variance' in the model. This tool splits the variance data into two components. One is variation due to Regression that explains the proportion of 'variation between the means of groups' and the other is variation due to Residual or standard error that indicates the 'variation that is random or by chance'.

df is 'degrees of freedom' of the variables. Under the column sum of squares' (SS), the first row shows sum of squares due to regression that is 'variability between the 'means of the group'. Second row shows sum of squares due to residuals displaying 'variability in the residuals'. The third row is total sum of squares that is the total variability obtained by adding the two above variabilities. The lesser the Residual 'Sum of Squares' better is the model obtained. Mean Sum of Square (MS) is the 'mean of sum of squares'. F-value is the calculated value of F statistics that indicates the significance of the regression model obtained. Significance F (the p-value of F determined from the F-Distribution tables) indicates the statistical significance of the result. If it is less than the significance level of 0.05, the model is accepted. Here the 'significance F' is 1.90209E-05 (where E-05 implies 10^{-5}), therefore the model is good.

Table 5: Table of Coefficients

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	30.905899	0.847961	36.44731	3.51971E-10
GDP	-0.004341179	0.000484	-8.96812	1.90209E-05

To form a 'linear regression line', the coefficients given in the Table 5 will be used. It may be noted that the coefficient of per capita GDP (b) is -0.0043 and intercept (a) is 30.906(rounded off).
The linear equation is of the form "y= bx + a",
∴ the equation of the regression line obtained is y= - 0.0043x + 30.906. (2)

Hypothesis testing:

The statistics obtained in the Table 5 provide information to test hypothesis. Hypothesis testing is a method by which the sample data is used to provide the inferences for the whole population. For that p-value is used as it helps to determine which of the two hypothesis, H_0 or H_1 is to be accepted. The Significane F that is the p-value given in the Table of Coefficients (Table 5) is 1.90209E-05. Since the $p < 0.05$, H_0 (null hypothesis) is rejected at 5% level of significance and hence H_1 is accepted.

Interpretations:

Correlation calculated is -0.95369 indicating high correlation between the two variables. The regression equations (both 1&2) obtained using 'Excel ToolPak' and through the linear regression graph is exactly same. The estimated mean value of the dependent variable can be obtained for any selected value of the independent variable. For example, to determine the FLFPR for any value of per Capita GDP, say 1500, put the value of x as 1500 in the equation

Y= -0.0043 * 1500+ 30.906 = 24.456

Thus it can be said that for per capita GDP of 1500, the FLFPR will be 24.456. For any selected value of per capita GDP, the FLFPR can be calculated. This is very useful for predictions and projections. For this data as per capita GDP increases by 1 unit, the estimated decrease in the FLFPR will be about 0.0043 units. Since the $p < 0.05$, null hypothesis is rejected at 5% level of significance.

Conclusions:

The correlation between the FLFPR and the per capita GDP is negative strong correlation. Thus there is inverse relationship between FLFPR and per capita GDP. The regression analysis shows that when per capita GDP increases then FLFPR decreases and vice versa. Since the p-value is less than the significance level there is sufficient evidence that the informs that null hypothesis is to be rejected at 5% level of significance and the linear model obtained above fits the data well. Thus it may be concluded that the FLFPR in the age group 15-64 is affected by the change in the per capita GDP.

References:

Pearson, K. (1911). *The Grammar of Science, 3rd ed*. Adam and Charles Black.

Thompson, J.A. (1911). *Introduction to Science*. The University Press.

Hilal, A.H. & Alabri, S.S. (2013). Using Nvivo for data analysis in qualitative research. In International Interdisciplinary Journal of Education, 2 (2), 181-186. http://www.iijoe.org/v2/IIJOE_06_02_02_2013.pdf