IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# PREDICTION OF DATA THROUGH MULTIPLE LINEAR REGRESSION ANALYSIS

*Dr. Shyamal Kumar Das (Independent researcher)*
*Srabani Complex; Dakbanglow, Barasat*
*Kolkata-700124*

**ABSTRACT**

The main purpose of running a multiple regression analysis is to find out the relationship between the dependent variable and the independent variables, and among and between the independent variables and whether those independent variables are good enough to help in predicting the dependent variable, and those multiple independent variables are chosen to predict the dependent variable. This study is fashioned to create a scalable intervention about the judicious use and application of multiple regression analysis in the field of educational research and for better and greater prediction of students' achievement under different criteria (considered as independent/predictor variables). This study would act as a support system to the curriculum framers /educational evaluation management for elaborate understanding of predicting students' achievement.

Key words: Multiple regression, evaluation, achievement, curriculum.

## I. INTRODUCTION

When more than one independent variable able to predict or explain a dependent variable, generally all those independent variables are put into the 'model' and perform a multiple linear regression analysis. In short Multiple Linear Regressions is an analytical procedure used when more than one independent variable is included in a 'model'.

A MLRM under $k$ independent variables $x_1$, $x_2$, $x_3$, $x_4$, ..., $x_k$ and a dependent variable $Y$, can be written as $\boldsymbol{Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots\cdots\cdots + \beta_k x_k + \varepsilon}$

The parameters $\beta_1$ , $\beta_2$ , $\beta_3$ , $\beta_4$ , ...., $\beta_k$ are the regression coefficients associated with $x_1$, $x_2$, $x_3$, $x_4$, ..., $x_k$ respectively. $\beta_0$ is the constant term i.e. equivalent to the 'y-intercept' particularly for Simple Linear Regression (SLR). $\varepsilon$ is the (residual for the particular observation in the population) random error component reflecting the difference between the observed and fitted linear relationship. There can be various reasons for such difference, e.g., joint effect of those variables not included in the model, random factors which cannot be accounted in the model etc. In general more complex models may include higher powers of one or more independent variables i.e.

$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

Or interaction effects of two or more variables $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$

Multiple Linear Regression (MLR) is one of the most common forms of regression analysis. As predictive analysis it is used to describe data and to analyse the strength and relationship among and between one dependent variable and two or more independent variables. MLR fits a line which clustered around by multi-dimensional cloud of data points and besides that it consists of three stages;

(i) analysing the correlation and directionality of the data,
(ii) estimating the model, i.e., fitting the line,
(iii) evaluating the validity and usefulness of the model.

## II. FORMATION OF REGRESSION EQUATION

Students' scholastic achievement in mathematics (SAM) and attitude towards mathematics (ATM) are related. It is expected that, on an average, higher level of students' attitude towards mathematics (ATM) provides higher Students' scholastic achievement in mathematics (SAM). So a simple linear regression equation can be expressed as;

$SAM = \beta_0 + \beta_1 (ATM)$

$\beta_1$ reflects the change in scholastic achievement in mathematics with respect to change in one unit attitude towards mathematics. $\beta_0$ reflects the students' scholastic achievement in mathematics when their attitude towards mathematics is almost nil as it is expected that a student may achieve minimum without any positive attitude towards mathematics.

Again students' scholastic achievement in mathematics depends upon their level of IQ (empirical evidences suggest that). So $\beta_1$ will over-state the marginal impact of students' level of IQ. Hence if students' level of IQ and attitude towards mathematics are correlated, then the regression model will associate with all the observed influence in students' scholastic achievement in mathematics.

So the better equation is;

$SAM = \beta_0 + \beta_1 (ATM) + \beta_2$ **(students' level of IQ)**

Again students' scholastic achievement in mathematics depends upon their duration of mathematics practice in a day (in hour). Students' scholastic achievement in mathematics would be desirably influenced by duration of mathematics practice in a day (in hour) regardless of students' level of IQ. So $\beta_2$ will over-state the marginal impact of IQ. Hence if duration of study and students' level of IQ are correlated, then the regression model will associate with all the observed influence in students' scholastic achievement in mathematics. So the better equation is;

$SAM = \beta_0 + \beta_1 (ATM) + \beta_2$ **(students' level of IQ)** $+ \beta_3$ ***(Duration of math practice in hour)***

Often it is observed that student' scholastic achievement in mathematics tends to rise more rapidly under the teachers with better mathematics pedagogical content knowledge (MPCK). In order to accommodate such possibility, the equation can be extended as;

$SAM = \beta_0 + \beta_1 (ATM) + \beta_2$ **(students' level of IQ)** $+ \beta_3$ ***(Duration of math practice in hour)*** $+ \beta_4 (MPCK)$

It is the way to proceed for regression modelling of a research. A researcher must consider the experimental conditions and phenomenon in advance to take the decision on how many, way and how to choose the predictor and dependent variable.

## III. ANALYSIS AND INTERPRETATION

Let us try and understand the concept of multiple regressions analysis with the help of this example. Let us try to find out what is the relation among and between students Scholastic Achievement in Mathematics (SAM), students Attitude towards Mathematics (ATM), students' level of IQ, duration of math practice, Teachers' mathematics pedagogical content knowledge (MPCK). All the Class-X students of Kolkata are considered as population and 40 students from three different schools are considered as sample of the study.

Dependent variable ($Y$) = students' scholastic achievement in mathematics (SAM)

Independent variables ($x_1$, $x_2$, $x_3$, *and* $x_4$) = Students' Attitude towards Mathematics (ATM) ($x_1$), students' level of IQ ($x_2$), duration of math practice in hour ($x_3$), Teachers' mathematics pedagogical content knowledge (MPCK) ($x_4$). The dataset and multiple linear regression equation are given below.

| S/No | SAM | ATM | IQ | Hour | MPCK |
|---|---|---|---|---|---|
| 1 | 76 | 69 | 100 | 4 | 75 |
| 2 | 56 | 52 | 89 | 4 | 66 |
| 3 | 59 | 60 | 94 | 3 | 68 |
| 4 | 68 | 71 | 100 | 4 | 61 |
| 5 | 74 | 73 | 99 | 3 | 72 |
| 6 | 82 | 74 | 125 | 5 | 77 |
| 7 | 87 | 75 | 125 | 4 | 75 |
| 8 | 62 | 64 | 102 | 3 | 54 |
| 9 | 77 | 64 | 120 | 3 | 58 |
| 10 | 68 | 71 | 109 | 5 | 64 |
| 11 | 43 | 59 | 98 | 5 | 51 |
| 12 | 49 | 58 | 76 | 4 | 66 |
| 13 | 35 | 56 | 76 | 2 | 60 |
| 14 | 68 | 73 | 99 | 4 | 70 |
| 15 | 77 | 67 | 103 | 3 | 68 |
| 16 | 84 | 76 | 123 | 4 | 72 |
| 17 | 40 | 50 | 79 | 3 | 66 |
| 18 | 58 | 64 | 81 | 5 | 63 |
| 19 | 66 | 75 | 89 | 4 | 72 |
| 20 | 71 | 69 | 99 | 3 | 45 |

| S/No | SAM | ATM | IQ | Hour | MPCK |
|---|---|---|---|---|---|
| 21 | 64 | 58 | 97 | 2 | 42 |
| 22 | 55 | 62 | 90 | 3 | 62 |
| 23 | 60 | 69 | 94 | 3 | 72 |
| 24 | 66 | 70 | 103 | 4 | 64 |
| 25 | 73 | 68 | 116 | 2 | 60 |
| 26 | 60 | 58 | 101 | 4 | 62 |
| 27 | 54 | 60 | 97 | 3 | 56 |
| 28 | 68 | 67 | 100 | 3 | 66 |
| 29 | 75 | 84 | 121 | 3 | 58 |
| 30 | 43 | 54 | 76 | 2 | 62 |
| 31 | 55 | 69 | 79 | 3 | 55 |
| 32 | 40 | 62 | 71 | 3 | 50 |
| 33 | 53 | 70 | 94 | 4 | 72 |
| 34 | 78 | 86 | 102 | 3 | 68 |
| 35 | 70 | 80 | 106 | 4 | 70 |
| 36 | 76 | 71 | 120 | 3 | 69 |
| 37 | 36 | 40 | 71 | 3 | 46 |
| 38 | 45 | 53 | 73 | 2 | 68 |
| 39 | 71 | 80 | 108 | 3 | 66 |
| 40 | 50 | 62 | 89 | 3 | 54 |

Accordingly I have fit the regression model using students Scholastic Achievement in Mathematics (SAM) as dependent variable and students Attitude towards Mathematics (ATM), students' level of IQ, duration of math practice in hour, and Teachers' mathematics pedagogical content knowledge (MPCK) as the predictor or independent variables.

| Independent Variable | Coefficients | | SE | t | P | L/ 95% | U/ 95% |
|---|---|---|---|---|---|---|---|
| Intercept | $\beta_0$ | -33.0396 | 8.0643 | -4.0970 | 0.0002 | -49.4112 | -16.6681 |
| ATM | $\beta_1$ | 0.4415 | 0.1355 | 3.2576 | 0.0025 | 0.1663 | 0.7166 |
| IQ | $\beta_2$ | 0.5809 | 0.0811 | 7.1590 | 0 | 0.4161 | 0.7456 |
| Hour | $\beta_3$ | -0.8026 | 1.1870 | -0.6761 | 0.5033 | -3.2125 | 1.6072 |
| MPCK | $\beta_4$ | 0.1952 | 0.1229 | 1.5873 | 0.1214 | -0.0544 | 0.4448 |

The least square regression equation is;

$$\mathbf{SAM = -33.0397 + 0.4415 \cdot ATM + 0.5809 \cdot IQ - 0.8026 \cdot Hour + 0.1952 \cdot MPCK}$$

### 3.1 Interpretation of the Coefficients in the Multiple Linear Regression Equation

Regression coefficients are the estimates of the unknown population parameters and describe the relationship between a predictor variable and the dependent variable. These regression coefficients represent the mean change in the dependent variable for one unit of change in the predictor variable while holding other predictors in the model constant.

The sign of each coefficient indicates the direction of the relationship between a predictor variable and the dependent variable i.e. the mean change in the dependent variable given a one-unit increase in the predictor variable.

- A positive signed coefficient indicates that if the value of predictor variable increases, the mean of the dependent variable also tends to increases (i.e. relation is in direct proportion)
- A negative signed coefficient indicates that if the value of predictor variable increases, the mean of the dependent variable tends to decreases (i.e. relation is in inverse proportion)
- Example, the coefficient ($\boldsymbol{\beta_1}$) is +0.4415, means dependent value **(*Y*)** increases by 0.4415 for every one unit change in the predictor, likewise the coefficient ($\boldsymbol{\beta_3}$) -0.8026, the mean dependent value **(*Y*)** decreases by 0.8026 for every one unit change in the predictor.

It's easy to think that variables with larger coefficients are more important because they represent a larger change in the dependent variable but it does not mean that a predictor variable with larger coefficient identify itself more important (i.e. *larger coefficients don't necessarily identify more important predictor variables*).

### 3.2 Multicollinearity in Regression Model

In regression analysis Multicollinearity occurs when two or more independent variables are highly correlated with each other, hardly those independent variables provides any unique or independent information in the regression model. In short multicollinearity exists when:

- One independent variable is correlated with another independent variable.
- One independent variable is correlated with a linear combination of two or more independent variables.

Its presence can adversely affect the regression results. If the correlation is high enough among and between independent variables, there is every possibility for the problems when fitting and interpreting the regression model.

There are two popular ways to measure multicollinearity:

- computing a coefficient of multiple determinations for each independent variable,
- computing variances inflation factor for each independent variable.

It is possible to detect multicollinearity using a metric known as the variances inflation factor (VIF), which measures the correlation and strength of correlation among and between the independent variables in a regression model.

To find VIF at first we are to find $\mathbf{R^2}$ for each of the four independent variables by performing individual regression using one independent variable (say ATM) as the dependent variable and the other three (i.e. IQ, Hour, and MPCK) as the independent variables. The value of VIF starts at 1 and has no upper limit. For the interpretation of VIF values there is a general rule; value of 1 indicates there is no correlation among and between a given dependent variable (practically one independent variable act that) and any other independent variables in the model.

- A value between 1 and 5 indicates there is moderate correlation among and between a given dependent variable (practically one independent variable act that) and any other independent variables in the model, but this is often not serve enough to require attention.

- A value greater than 5 indicates potentially severe correlation among and between a given dependent variable (practically one independent variable act that) and any other independent variables in the model. In this case, the coefficient estimates and p-value in the regression output are likely unreliable.

| Independent variable | $R^2$ | VIF | Correlated with other Independent variable | VIF |
|---|---|---|---|---|
| ATM | 0.501759 | 2.007060 | Moderate correlation | $1 / (1 - R^2)$ |
| IQ | 0.454481 | 1.833116 | Moderate correlation | |
| Hour | 0.145164 | 1.169815 | Moderate correlation | |
| MPCK | 0.244491 | 1.323610 | Moderate correlation | |

- It reveals, as each of the VIF values for the independent variables in regression model are moderate multicollinearity is not a problem for this model.

- Multicollinearity makes it hard to assess the relative importance of independent variables, but it does not affect the usefulness of the regression equation for prediction. Even when multicollinearity is great, the least-squares regression equation can be highly predictive. So, if you are only interested in prediction, multicollinearity is not a problem.

### 3.3 Interpretation of P-Value for a Predictor Variable in a Multiple Linear Regression Equation

There is one of the most important reason for performing a multiple linear regression analysis is to determine which (if any) of the independent variables are significant predictors of the dependent variable because independent variables as useful predictor help to "explain" the dependent variable.

The P-value helps in determining whether the relationships observed in sample also exists in the larger population. The P-value against each independent variable test the null hypothesis i.e. there is no significant correlation between dependent variable ($Y$) and the independent variable ($x_i$). If there is no correlation, there is no association between the changes in the independent variable and the shifts in the dependent variable. In other words, there is insufficient evidence to conclude that there is effect at the population level.

This can be explained as follow;

| Independent variables | P-value | α = 0.05 | For the entire population sample data provides... |
|---|---|---|---|
| ATM | 0.0002 < 0.05 | Sig. | enough evidence to reject the null hypothesis. |
| IQ | 0 < 0.05 | Sig. | enough evidence to reject the null hypothesis. |
| Hour | 0.5033 > 0.05 | N/Sig. | insufficient evidence to reject the null hypothesis. |
| MPCK | 0.1214 > 0.05 | N/Sig. | insufficient evidence to reject the null hypothesis. |

There is enough evidence to indicate that the students' ATM is a significant predictor of students' SAM keeping the students' level of IQ, duration of mathematics practice (i.e. in Hour), and teachers' MPCK the same.

There is enough evidence to indicate that the students' level of IQ is a significant predictor of students' SAM keeping the students' ATM, duration of mathematics practice (i.e. in Hour), teachers' MPCK the same.

There is not enough evidence to indicate that the students' duration of mathematics practice (i.e. in Hour) is a significant predictor of students' SAM keeping the students' ATM, level of IQ, and teachers' MPCK the same.

There is not enough evidence to indicate that the teachers' MPCK is a significant predictor of students' SAM keeping the students' ATM, level of IQ, and duration of mathematics practice (i.e. in Hour) the same.

Hardly the value of regression coefficients ($\beta_i$) indicates the importance of an independent variable ($\boldsymbol{x_i}$). P-value calculations incorporate a variety of properties, but a measure of importance is not among them. A very low p-value can reflect properties other than importance, such as a very precise estimate and a large sample size (i.e. *low p-values don't necessarily identify predictor variables that are practically important*).

In some case t-test reveals that none of the independent variable is significant predictor of dependent variable, even though the p-value from the F-test suggests that at least there must be one independent variable which would be a significant predictor of dependent variable. In that case we are to follow the backwards selection process where at the initial stage all independent variables are included in the initial model. Then the "least significant" independent variable is removed from the model and the model is re-fit with the remaining independent variables. Again, the least-significant independent variable is removed and the model is re-fit with the remaining independent variables. This process of removing one independent variable at a time is continued until all remaining independent variables are "significant" predictors of the dependent variable.

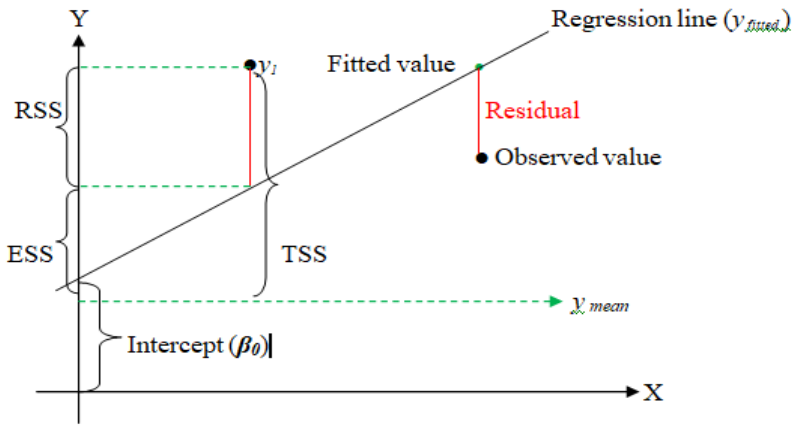
A variable is 'least significant' means its P-value is highest from the t-test, and a variable is 'significant' means its P-value is less than 0.05 (or so).

*Note: the "or so" is included because 0.05 is an arbitrary set cut-off point. We may decide to keep an independent variable with a p-value of 0.06, for example.*

### 3.4 Interpretation of R-Squared and the Adjusted R-Squared in a Multiple Linear Regression Equation

In regression analysis a residual is the vertical distance between a data point and the regression line means the difference between an observed value of the dependent variable (y) and the predicted/ fitted value of y i.e. (ŷ). They are positive if they are above the regression line and negative if they are below the regression line. Each data point has one residual. If the regression line actually passes through the point, the residual at that point is zero.

Residual = Observed value – Predicted/ Fitted value = ($\boldsymbol{\varepsilon}$) = y – ŷ.




RSS = Residual sum of square i.e. RSS = $\Sigma (y_1 - y_{fitted})$

TSS = Total sum of square i.e. TSS = $\Sigma (y_1 - y_{mean})$

ESS = Explained sum of square i.e. ESS = $\Sigma (y_{fitted} - y_{mean})$

$$R_{sq} = 1 - (RSS / TSS)$$

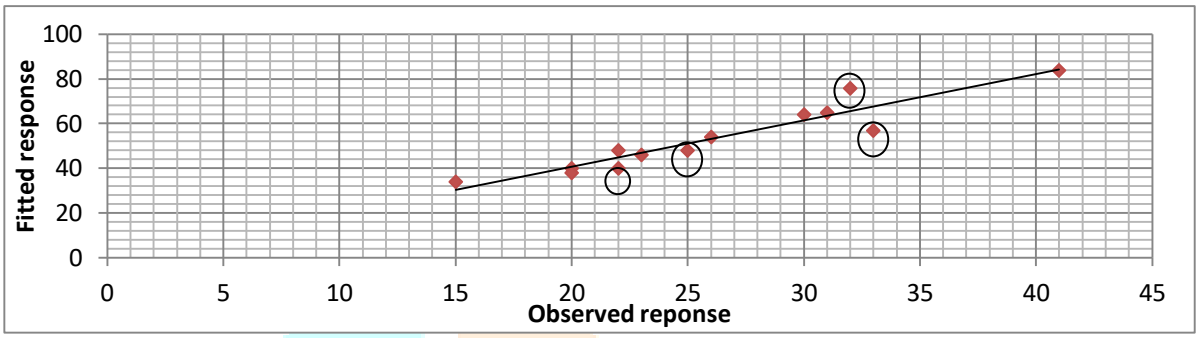
$$R_{sq}^{adj} = 1 - \{(1 - R_{sq}) * (n - 1) / (n - p - 1)$$

R-squared is a statistical measure of how close the data are to the fitted regression line. And it explains the degree to which independent variables ($x_1$, $x_2$, $x_3$, *and* $x_4$) explain the dependent variable ($Y$). But adjusted R-squared adjusts the statistic based on the number of independent variables in the model. The adjusted R-squared is a modified version of $R^2$ for the number of predictors in a model. The adjusted R-squared can be negative only when Residual sum of squares approaches to the total sum of squares, that means the explanation towards response is very low or negligible but not all the time. So, Negative Adjusted $R^2$ means insignificance of explanatory variables. The results may be improved with the increase in sample size or avoiding correlated independent variables. R-squared always remain between 0 and 1 (If $R^2 = 0$ then the dependent variable cannot be predicted and $R^2 = 1$ then the dependent variable can be predicted without error) and shows the linear relationship in the sample of data even when there is no basic relationship, but the adjusted R-squared gives the best estimate of the degree of relationship in the basic population.

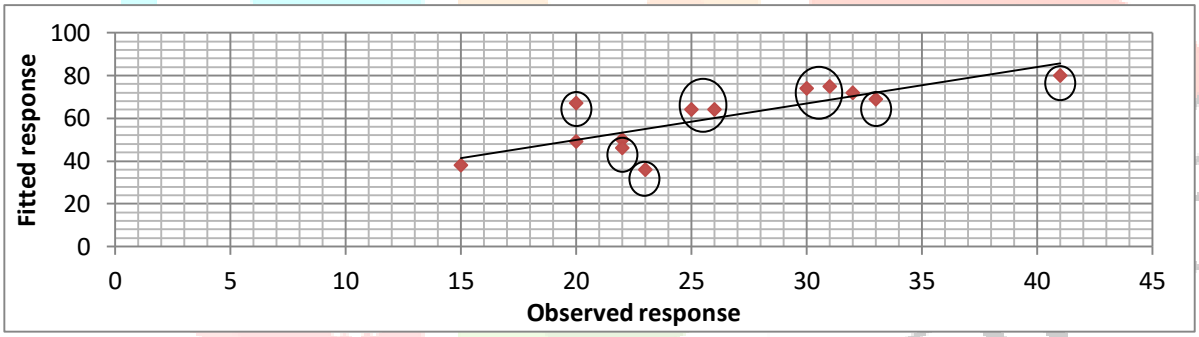
The main problem with R-squared is that it will either stay the same or increase but would not decrease when more independent variables are added to the model, even if they do not have any proper relationship with the dependent variables. Whereas the Adjusted R-squared will only increase when a significant independent variable is added otherwise its value would decrease if the added independent variable is insignificant. Therefore it is always suggested to use Adjusted R-squared to judge goodness of model as it is more reliable and accurate in determining the efficiency of the model. It is to be noted that, (i) for one independent variable, R-square and Adjusted R squared would be exactly same, (ii) for the addition of more non-significant independent variables the gap between R-squared and Adjusted $R^2$ would increase, (iii) the Adjusted R-squared includes the number of independent variables in its formula where as the R-squared does not, (iv) R-squared cannot verify whether the coefficient ballpark figure and its predictions are prejudiced. It also does not show if a regression model is satisfactory; it can show an R-squared figure for a good model, or a high R-squared figure for a model that doesn't fit.

**Difference between R-squared and the adjusted R-squared**

| | R-squared | Adjusted R-squared |
|---|---|---|
| 1 | Gives the percentage of explained variation as if all the independent variables in the model affect the dependent variable. | Gives the percentage of variation explained by only those independent variables that in reality affect the dependent variable. |
| 2 | When added more significant or insignificant independent variables the value of R-squared either stay the same or increase. | When added more significant or insignificant independent variables the value of adjusted R-squared increase or decrease respectively. |




The graph represents 89% ($r^2$ = **0.893853**) explained and 11% unexplained independent variables (within the circle) in the model affect the dependent variable.




The graph represents the 65% (**$r^2$=0.649945)** explained and 35% unexplained independent variables (within the circle) in the model affect the dependent variable.

The first model fits the data better than the second one because the data points are closer to the regression line means the more variance that is accounted for by the regression model the closer the data points will fall to the fitted regression line.

But it should be always remembered that higher R- Squared values are not always inherently good (means the model fits the data better) and vis-à-vis.

When attempts are made to predict on Behavioural Science (i.e. attitude, creativity, IQ level, anxiety, effectiveness of certain human behaviour etc) in those cases generally R-Squared value typically become lower than 50%, as comparing with the physical process human behaviours are hardly predictable and reveals the model has a good fit. Again, if R-Squared value is low against statistically significant predictors this can be strongly concluded that changes in predictor values are associated with changes in the dependent values and reveals the model has a good fit. As per the below mentioned table obtained from statistical data of 40 students mentioned before:

| | |
|---|---|
| **R-Squared** | **$r^2$ = 0.8451** |
| **Adjusted R-Squared** | **$r^2_{adj}$ = 0.8274** |
| **Residual Standard error** | 5.7426 on 35 degrees of freedom. |
| **Overall F-statistics** | 47.7469 on 4 and 35 degrees of freedom. |
| **Overall P-value** | 0 |

- **$r^2$ = 0.8451** (approximately 0.85) means 85% of the variance in *Y* is predictable as if all the independent variables in the model affect the dependent variable. So 15% of the variance remains unexplained. In general, the higher the R-squared, the better the model fits the data. But it should be remembered that this might be an over-estimation, and it's often better to look at Adjusted R-Squared (which penalizes the amount of predictors you used).
- **$r^2_{adj}$ = 0.8274** (approximately 0.83) means 83% of the variation explained by only those independent variables that in reality affect the dependent variable.
- **$r^2 - r^2_{adj}$ = 0.8451 − 0.8274 = 0.0177** (approximately 0.02) means only 2% variations fluctuate in the explanation of independent variables in the model affect the dependent variable.

### 3.5 Hypotheses for F-Test in Multiple Linear Regressions

R-squared reveals how well the model fits the data, and the F-test is related to it. F-tests can evaluate multiple model terms simultaneously, which allows them to compare the fits of different linear models. In contrast, t-tests can evaluate just one term at a time.

The overall F-test compares the model that we specify to the model with no explanatory variables. This type of model is also known as an intercept-only model.

The F-test for overall significance has the following two hypotheses:

Null hypothesis:

**$H_0$:** All the coefficients are zero (0) or $H_0$: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \cdots\cdots = \beta_k = 0$

This implies that none of the independent variables are significant predictors of the dependent variable.

Alternative hypothesis:

**$H_a$:** At least one coefficient is not Zero (0) or $H_a$: At least one $\beta_k \neq 0$

This implies that at least one of the independent variables is a significant predictor of the dependent variable.

In statistical output, we can find the overall F-test **(**F-test does not indicate which of the parameters $\beta_k \neq 0$ is not equal to zero, only that at least one of them is linearly related to the response variable) in the ANOVA table. As per the below mentioned table obtained from statistical data of 40 students mentioned before:

| Source | df | ∑of Square | Mean Square | F | P |
|---|---|---|---|---|---|
| Regression | 4 | 6298.2046 | 1574.5511 | 47.7469 | **0.00001** |
| Residual error | 35 | 1154.1954 | 32.9770 | | |
| Total | 39 | 7452.4 | | | |

The ANOVA table reveals;

⟰ How much "Explaining" the model doing? i.e.

The ratio SSM/SST = $\mathbf{R^2}$ is known as the squared multiple correlation coefficients. This value is the proportion of the variation in the response variable that is explained by the response variables. The square root of $R^2$ is called the multiple correlation coefficients, the correlation between the observation $y_i$ and the fitted values $\hat{y}$.

$\mathbf{R^2}$ = (Regression of ∑of Square / Total of ∑of Square) = (6298.2046 / 7452.4) = 0.845124

It means the data provide sufficient evidence to conclude that the regression model fits the data better than the model with no independent variables means the independent variables in the model improve the fit.

⟰ **As** P-value is less than 0.01 we should continue the analysis.

Generally speaking, if none of our independent variables are statistically significant, the overall F-test is also not statistically significant. Occasionally, the tests can produce conflicting results. This disagreement can occur because the F-test of overall significance assesses all of the coefficients jointly whereas the t-test for each coefficient examines them individually. For example, the overall F-test can find that the coefficients are significant jointly while the t-tests can fail to find significance individually.

These conflicting test results can be hard to understand, but think about it this way. The F-test sums the predictive power of all independent variables and determines that it is unlikely that all of the coefficients equal zero. However, it's possible that each variable isn't predictive enough on its own to be statistically significant. In other words, our sample provides sufficient evidence to conclude that the model is significant, but not enough to conclude that any individual variable is significant.

According to me for the assessment of statistical significance of any estimate, Confidence Interval can be used as hypothesis testing. If the Confidence Interval capture the value of "no effect" this represents a statistically non-significant result. At the same time if the Confidence Interval does not capture the value of "no effect" this represents a different that is statistically non-significant result. The value of "no effect" is the absolute measure, i.e. absolute risk, absolute risk reduction and the number needed to treat. Means a specific intervention leads to zero (0) risk reduction [Risk in control group – Risk in intervention group = 0], it has no effect compared with the control. Thus in situations dealing with absolute measure the value of "no effect" is zero (0).

| Coefficient | Estimate | SE | 95% CI |
|---|---|---|---|
| ATM ($\beta_1$) | 0.4415 | 0.1355 | (0.1664 , 0.7166) Containing no '0' hence Sig. |
| IQ ($\beta_2$) | 0.5809 | 0.0811 | (0.4163 , 0.7455) Containing no '0' hence Sig. |
| Hour ($\beta_3$) | -0.8026 | 1.1870 | (-3.2123 , 1.6071) Containing '0' hence N/Sig. |
| MPCK ($\beta_4$) | 0.1952 | 0.1229 | (-0.0543 , 0.4447) Containing '0' hence N/Sig. |

The degrees of freedom for the t* critical value is the 'df' in the Analysis of Variance table and that is; df = n – p – 1= 40 – 4 – 1 = 35 (for this problem, n = 40, p = Number of independent variables = 4).

t* critical value for a 95% Confidence Interval and df, 35 is **2.030108**

Lower limit = (Estimate) – (t* critical value x SE) = 0.4415 – (2.030108x 0.1355) = 0.1664

Upper limit = (Estimate) + (t* critical value x SE) = 0.4415 + (2.030108x 0.1355) = 0.7166

Interpretation of the 95% confidence interval for the student's ATM ($\beta_1$): we are 95% sure that the rate of student's SAM will be between 0.1664 lower to 0.7166 higher for a student with 1 more quantitative value of ATM, when that student's quantitative value of level of IQ, duration of mathematics practice in hour, and his perception of teacher's MPCK are constant.

### 3.6 The Use of Multiple Linear Regression Equation for Prediction

- Suppose we are to predict the SAM of a student who's ATM, level of IQ, duration of math practice in hour, and his teacher's MPCK are 78, 105, 4, and 67 respectively. In this particular situation,

$Y_{(SAM)} = \beta_0 + \beta_1 (ATM) + \beta_2$ (students' level of IQ) $+ \beta_3$ (*Duration of math practice in hour*) $+ \beta_4 (MPCK)$

SAM = − 33.0397 + 0.4415·ATM + 0.5809·IQ − 0.8026·Hour + 0.1952·MPCK

= − 33.0397 + 0.4415 x 78 + 0.5809 x 105 – 0.8026 x 4 + 0.1952 x 67

= 72.2598 (That student is expected to secure 72 marks in mathematics)

- Suppose we are to predict the IQ level of a student who's SAM, ATM, duration of math practice in hour, and his teacher's MPCK are 43, 54, 2, and 62 respectively. In this particular situation,

$Y_{(SAM)} = \beta_0 + \beta_1 (ATM) + \beta_2$ (students' level of IQ) $+ \beta_3$ (*Duration of math practice in hour*) $+ \beta_4 (MPCK)$

SAM = − 33.0397 + 0.4415·ATM + 0.5809·IQ − 0.8026·Hour + 0.1952·MPCK

43 = − 33.0397 + 0.4415 x 43 + 0.5809 x IQ − 0.8026 x 2 + 0.1952 x 62

43 = − 33.0397 + 18.9845 + 0.5809 x IQ – 1.6052 + 12.1024

43 = 0.5809 x IQ – 3.558

0.5809 x IQ = 43 + 3.558

IQ = 46.558 / 0.5809

IQ = 74.9836 (It indicates the students level of IQ is approximately 75)

**** From dataset check Sl. No. 30***

## IV. CONCLUSION

Due to various reasons it becomes necessary to predict students' scholastic achievement. In that case basing on previously collected data of individual student, authority predicts the achievement. But it is evident those choose data under different independent / predictor variables fall in multicollinearity. Its presence can adversely affect the regression results i.e. bias result of students' scholastic achievement. In order to protect that an idea is shared. There is no hard and fast rule to decide the predictor variables rather it may be flexible and entirely depend upon the curriculum framers /educational evaluation management, avoiding the multicollinearity.

## V. ACKNOWLEDGEMENT

***This article is dedicated to my high school teachers, Late Madhu Sudan Saha (Headmaster & Mathematics Teacher), Late Ranjit Kumar Mitra (English Teacher), Chayan Kanti Dharchoudhury (Physical Science Teacher), & Dr. Parimalendu Chakraborty (Biology Teacher) of R. E. College Model School Durgapur, for whom today I am someone in the society.***

## BIBLIOGRAPHY


- DOI:10.5923/j.ajms.20170704.05
- DOI: 10.33564/IJEAST.2019.v04i06.045
- Guilden Kaya Uyanik, & Nese Guiler (2013): "A Study on Multiple Linear Regression Analysis", *Procedia-Social and Behavioural Sciences 106*; pp: 234-240. DOI: 10.1016/j.sbspro.2013.12.027
- Jim Frost, MS: "Regression Analysis; An Intuitive Guide for using and interpreting Linear Models".
- Jim Frost, MS: "Introduction to Statistics; An Intuitive Guide for Data Driven Decision".
- Jim Frost, MS: "Hypothesis Testing; An Intuitive Guide for Data Driven Decision".
- Serlin, Ronald C., Harwell, & Michael R. (2004): "More Powerful Tests of Predictor Subsets in Regression Analysis under Nonnormality", *Psychological Methods*, Vol-9(4), pp: 492-509. DOI: 10.1037/1082-989X.9.4.492
- STAT TREK: stattrek.com/multiple-regression
- The SAGE Encyclopaedia of Educational Research, Measurement, and Evaluation (2018): DOI: https://dx.doi.org/10.4135/9781506326139.n453


## END NOTES

**1.** "Mathematics Pedagogical Content Knowledge (MPCK) is an amalgam of Subject Matter Knowledge (SMK) in mathematics, Pedagogical Content Knowledge (PCK), knowledge of student characteristics and students' acceptance as per their biographical and demographical aspects and knowledge of creating congenial learning environment"

**2.**

Multiple Regression Model: $\mathbf{Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots\cdots\cdots + \beta_k x_k + \varepsilon}$

Multiple Regression Equation: $\mathbf{E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots\cdots\cdots + \beta_k x_k}$

Estimated Multiple Regression Equation: $\mathbf{\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots\cdots + b_k x_k}$, Where $\mathbf{b_0, b_1, b_2, b_3, b_4, \cdots\cdots, b_k}$ are the estimates of $\mathbf{\beta_0, \beta_1, \beta_2, \beta_3, \ldots\ldots, \beta_k}$

**3.** In reality it is difficult to conduct multiple regression analysis manually by the researchers. They always prefer to use computer software (SAS, SPSS, Minitab, Excel, etc.). Excel is a widely used, researcher friendly available software application that supports multiple regressions.

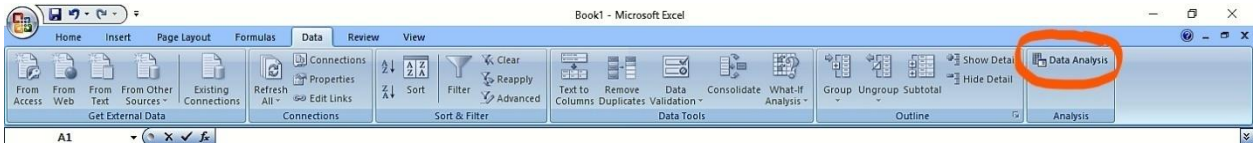
**Enabling Excel:**

At first before starting the analytical work we need to determine whether excel is enabling or not, to do so we are to follow the steps;

- Open Excel.
- Click the data tab.
- If Data Analysis button is available in the upper right corner, we are ready to go ahead. If not we are to set to Analysis Tool Pak in data tab, to do so we are to follow the steps;

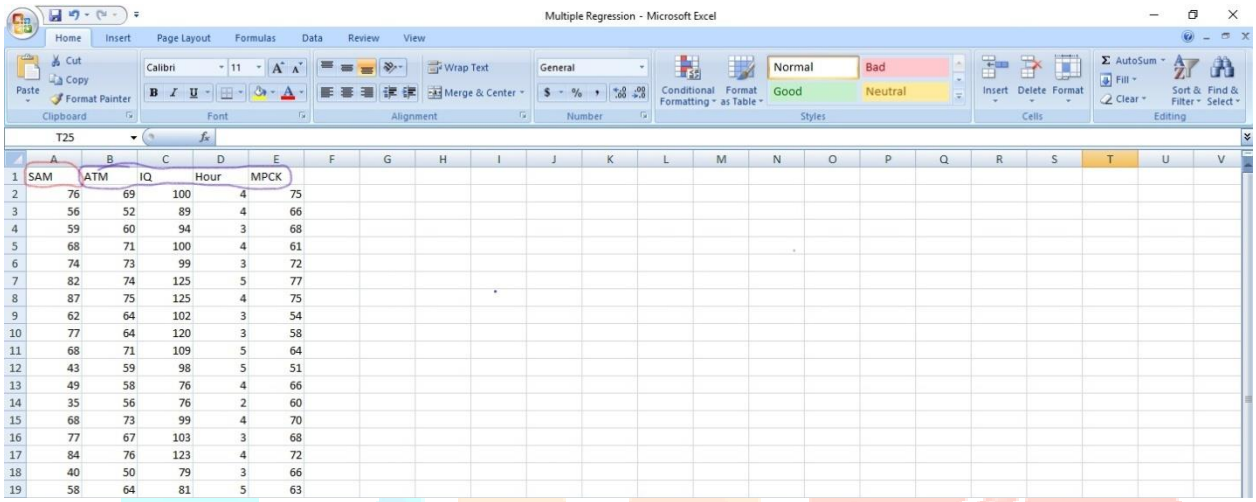
Open Excel → File → Options → Add-ins → Analysis Tool Pack → Go

Ultimately we would get Data Analysis Tool Pak in the upper right corner under the Data tab.




**Data Entry with Excel:**

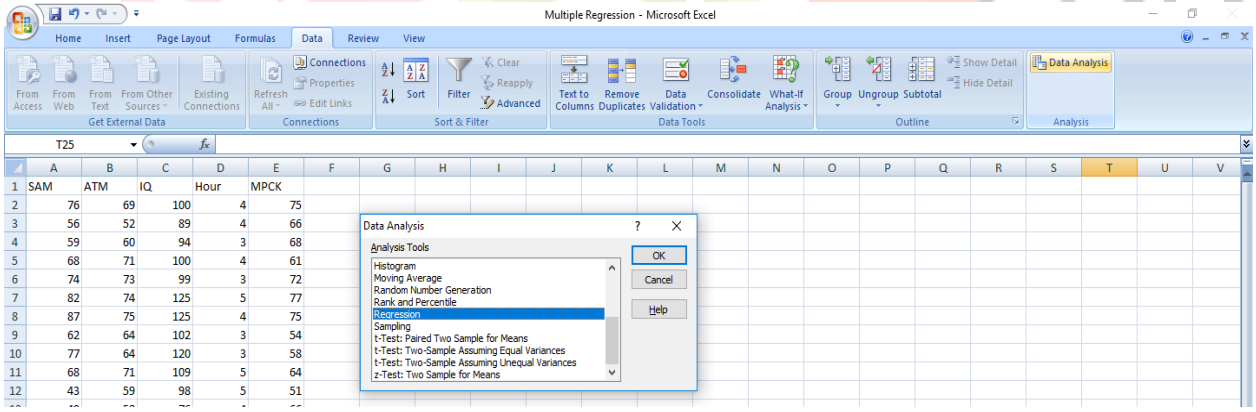
We are to enter the data on Excel spread sheet logically, like dependent variable at the first column and independent variables in the other subsequent adjacent columns and vis-à-vis.




| SAM | ATM | IQ | Hour | MPCK |
|---|---|---|---|---|
| 76 | 69 | 100 | 4 | 75 |
| 56 | 52 | 89 | 4 | 66 |
| 59 | 60 | 94 | 3 | 68 |
| 68 | 71 | 100 | 4 | 61 |
| 74 | 73 | 99 | 3 | 72 |
| 82 | 74 | 125 | 5 | 77 |
| 87 | 75 | 125 | 4 | 75 |
| 62 | 64 | 102 | 3 | 54 |
| 77 | 64 | 120 | 3 | 58 |
| 68 | 71 | 109 | 5 | 64 |
| 43 | 59 | 98 | 5 | 51 |
| 49 | 58 | 76 | 4 | 66 |
| 35 | 56 | 76 | 2 | 60 |
| 68 | 73 | 99 | 4 | 70 |
| 77 | 67 | 103 | 3 | 68 |
| 84 | 76 | 123 | 4 | 72 |
| 40 | 50 | 79 | 3 | 66 |
| 58 | 64 | 81 | 5 | 63 |

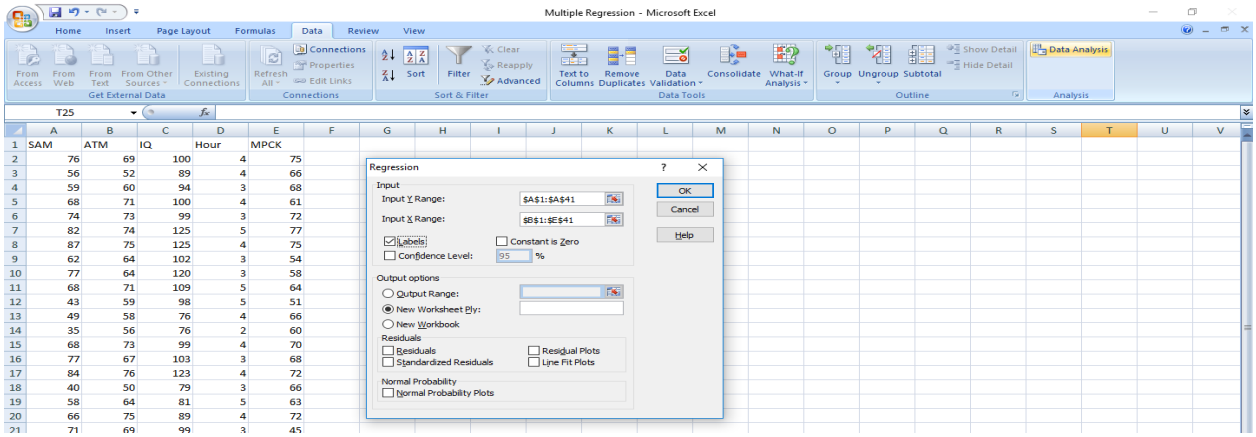
SAM is the dependent variable and ATM, IQ, Hour, and MPCK are independent variables.

Next we are to click on Data Analysis Tool Pak in the upper right corner under the Data tab, this would open Data Analysis dialog box. From the drop-down list we are to select '**Regression**' and click on it.




Excel would display the Regression dialog box, where in the Input Y Range (i.e. Dependent variables, SAM), and Input X Range (i.e. Independent variables, ATM, IQ, MPCK) we have to enter the respective coordinates respectively.

Then click on Label box → click on New Worksheet ply: → click OK to get regression outputs.




If necessity arises we may get additional outputs like Residual plots, and Normal probability plots, to do so we are to select the appropriate box(es) under output options. And Excel spread sheet would be displayed like below.

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | SUMMARY OUTPUT | | | | | | |
| 2 | | | | | | | |
| 3 | *Regression Statistics* | | | | | | |
| 4 | Multiple F | 0.919306439 | | | | | |
| 5 | R Square | 0.845124329 | | | | | |
| 6 | Adjusted I | 0.827424252 | | | | | |
| 7 | Standard E | 5.742561525 | | | | | |
| 8 | Observati | 40 | | | | | |
| 9 | | | | | | | |
| 10 | ANOVA | | | | | | |
| 11 | | *df* | *SS* | *MS* | *F* | *Significance F* | |
| 12 | Regressio | 4 | 6298.20455 | 1574.551137 | 47.74693 | 1.05458E-13 | |
| 13 | Residual | 35 | 1154.19545 | 32.97701287 | | | |
| 14 | Total | 39 | 7452.4 | | | | |
| 15 | | | | | | | |
| 16 | | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| 17 | Intercept | -33.0396859 | 8.064357547 | -4.097001621 | 0.0002356 | -49.41120203 | -16.6681699 |
| 18 | ATM | 0.441538306 | 0.135540068 | 3.257621982 | 0.0025003 | 0.16637734 | 0.716699271 |
| 19 | Hour | -0.8026461 | 1.187099791 | -0.676140377 | 0.5033953 | -3.212586784 | 1.607294582 |
| 20 | MPCK | 0.195202303 | 0.122972875 | 1.587360648 | 0.1214253 | -0.054445904 | 0.44485051 |
| 21 | IQ | 0.580912456 | 0.081143433 | 7.159081534 | 2.384E-08 | 0.41618253 | 0.745642383 |