IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Study Of The Relation Between The Infant Mortality Rate And The Literacy Rate By Using Python

[1]Dr. H. Ravi Sankar, [2]Dr. Y.M. Chenna Reddy, [3]Dr. A.V.S. Raghavaiah, [4]Dr J. Prabhakara Naik&[5]N. Venkata Subbaiah

[1,2 &5]Department of Statistics, Loyola Degree College(YSRR), Pulivendula, AP

[3]Department of Statistics, S.K.S.C. Degree College, Proddatur, AP

[4]Department of Statistics, S.S.N.B. Degree College, Anantapur

ABSTRACT

Infant Mortality Rate (IMR) is one of the important rate to measure the situation of any country whether developed or undeveloped or underdeveloped. IMR depends on so many things like Medical facilities, Economic Conditions, Age fertility rate, Literacy rate etc. It is a study to measure the relation between IMR and Literacy rate through the application of Python.

KEY WORDS

IMR, Literacy Rate, Linear Regression Model, Python

1. INTRODUCTION

Infant Mortality Rate (IMR) is our measure of interest, depends on so many variables throughout the world. It differs from country to country which is given to some extent in the following section. Along with the said causes our interest is extended to another aspect that whether Literacy Rate is showing any significant effect on IMR or not. Among so many people’s view Literacy Rate influence the IMR. To test this opinion, weare planning to collect a reliable data. It is to be tested whether there is relationship exists between our interested variables.For that we are planned to apply Linear Regression Model to establish the relationship and to obtain the estimates of the parameters. And finally we want to test the significance of that estimates and model. The testing is naturally carried out with the traditional methods. Now here we want to apply the analysis through the modern approach Python. It is easy, quick and accurate to apply Python.

2.1. INFANT MORTALITY RATE (IMR)

The death of children of age below 1 year is called Infant Mortality.Infant Mortality Rate (IMR) is the number of deaths of children under one year age per 1000 live births in a given region during the given period. The rate of deaths under age 5 years is Child Mortality Rate (CMR). 9 million infants of age below 1 year were died all over the world in 1990. IMR was recorded as 65 per 1000 live births. In 2015 it was declined to 29 per 1000 live births.

There are so many causes for IMR.Premature birth is main contributor to IMR.Infections, complications during delivery, perinatal asphyxia and birth injuries, environmental and social barriers prevent access to medical facilities are the other contributors for increasing IMR.

Medical

IMR and deaths are related to medical conditions include: low birth weight, sudden infant death syndrome, malnutrition, congenital malformations, and infectious diseases and low income for health care.

Congenital Malformations

Congenital Malformations have had a significant effect on Infant Mortality. These are birth defects that babies are born with, such as cleft lip and palate, Down syndrome, and heart defects. Malnutrition and infectious diseases were the main cause of death in more undeveloped countries. In more developed countries, these birth defects had to do with heart and central nervous system.

Low Birth Weight

If the babies born with weight 3000gm to 3500gm the mortality increases. The mortality rate rapidly increases with decreasing in weight. As compared with normal-birth-weight infants, those with low weight at birth are almost 40 times more likely to die in the neonatal period; for infants with very low weight at birth the relative risk of neonatal death is almost 200 times greater.Infant mortality due to low birth weight is usually a direct cause stemming from other medical complications such as preterm birth, poor maternal nutritional status, lack of prenatal care, maternal sickness during pregnancy, and an unhygienic home environment.

Sudden Infant Death Syndrome

This disease is more common in Western countries. Sudden infant death syndrome(SIDS) is a syndrome where an infant dies in their sleep with no reason behind it. Even with a complete autopsy, no one has been able to figure out what causes this disease.Some researchers have discovered that it is healthier for babies to sleep on their backs instead of their stomachs.This discovery saved many families from the tragedy that this disease causes.

Malnutrition

Inadequate intake of nourishment, such as proteins and vitamins, which adversely affects the growth, energy and development of people all over the world is termed as Malnutrition or undernutrition.Children are most vulnerable as they have yet to fully develop a strong immune system, as well as being dependent upon parents to provide the necessary food and nutritional intake. Factors which contribute to malnutrition are socioeconomic, environmental, gender status, regional location, and breastfeeding cultural practices.

In addition to the above factors there are so many factors that influencing the IMR like Infectious Diseases, Environmental, Socio-economic factors, Medicine and biology, Cultural Influences, Gender Favouritism etc.

2.2. LITERACY RATE

Literacy is the ability to read, write and comprehend information in order to communicate effectively. According to the 2011 Census, any person aged seven and above and has the ability to read and write is considered literate. Literacy and level of education are basic indicators of the level of development achieved by a society. Spread of literacy is generally associated with important traits of modern civilization such as modernization, urbanization, industrialization, communication and commerce. Literacy forms an important input in overall development of individuals enabling them to comprehend their social, political and cultural environment better and respond to it appropriately. Higher levels of education and literacy lead to a greater awareness and also contributes in improvement of economic and social conditions. It acts as a catalyst for social upliftment enhancing the returns on investment made in almost every aspect of development effort, be it population control, health, hygiene, environmental degradation control, employment of weaker sections of the society.

Literacy rate is considered as division of the number of literates of a given age range by the corresponding age group population and multiply the result by 100. Alternatively, apply the same method using the number of illiterates to derive the illiteracy rate; or by subtracting the literacy rate from 100%.The average literacy rate in India stands at 74.04%. While Kerala has the highest literacy rate in India at 93.91%, Bihar has the least literacy rate in India of 63.82%.

One of the main factors contributing to this relatively low literacy rate is usefulness of education and availability of schools in vicinity in rural areas. There is a shortage of classrooms to accommodate all the students. In addition, there is no proper sanitation in most schools. There are so many schools had no drinking water facility and no toilets.Severe caste disparities also exist.Discrimination of lower castes has resulted in high dropout rates and low enrolment rates. The National Sample Survey Organisation and the National Family Health Survey collected data in India on the percentage of children completing primary school which are reported to be only 36.8% and 37.7% respectively.The large proportion of illiterate females is another reason for the low literacy rate in India. Inequality based on gender differences resulted in female literacy rate being lower at 65.46% than that of their male counterparts at 82.14%. Due to strong stereotyping of female and male roles, Sons are thought of to be more useful and hence are educated. Females are pulled to help out on agricultural farms at home as they are increasingly replacing the males on such activities which require no formal education. Fewer than 2% of girls who engaged in agriculture work attended school.

2.3. LINEAR REGRESSION MODEL

Linear regression is a linear approach to modelling the relationship between a scalarresponse (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

- If the goal is prediction, forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.
- If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.
- Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares cost function as in Ridge regressionL^2-norm (penalty) and lasso (L^1-norm penalty). Conversely, the least squares approach can be

used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.

Consider a simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + €_i,\ i=1,2,\ldots\ldots,n$$

WhereYis termed as the dependent or study variable and *X* is termed as independent or explanatory variable. The terms β_0 and $\beta_1$are the parameters of the model. The parameter$\beta_0$is termed as intercept term and the parameter$\beta_1$is termed as slope parameter. These parameters are usually called as regression coefficients. The unobservable error component € accounts for the failure of data to lie on the straight line and represents the difference between the true and observed realization of *y*.This is termed as disturbance or error term. There can be several reasons for such difference, e.g., the effect of all deleted variables in the model, variables may be qualitative, inherit randomness in the observations etc. We assume that ε is observed as independent and identically distributed random variable with mean zero and constant varianceσ^2 . Later, we will additionally assume thatε is normally distributed.

2.4. PYTHON

Python is an interpreted, high-level, general programming language. Created by Guido van Rossum and first released in 1991. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

Python was conceived in the late 1980s as a successor to the ABC language. Python 2.0, released in 2000. Python 3.0, released in 2008, was a major revision of the language that is not completely backward-compatible and much Python 2 code does not run unmodified on Python 3. The Python 2 language, i.e. Python 2.7.x, was officially discontinued on 1 January 2020 (first planned for 2015) after which security patches and other improvements will not be released for it. With Python 2's end-of-life, only Python 3.5.x and later are supported.

Since 2003, Python has consistently ranked in the top ten most popular programming languages in the TIOBE Programming Community Index where, as of February 2020, it is the third most popular language (behind Java, and C). It was selected Programming Language of the Year in 2007, 2010, and 2018. Large organizations that use Python include Wikipedia, Google, Yahoo!, CERN, NASA, Facebook, Amazon, Instagram, Spotify and some smaller entities like ILM and ITA. The social news networking site Reddit is written entirely in Python. Python can serve as a scripting language for web applications and has been successfully embedded in many software products as a scripting language. Python is commonly used in artificial intelligence projects and many operating systems include Python as a standard component. Python is used extensively in the information security industry, including in exploit development. Due to Python's user-friendly conventions and easy-to-understand language, it is commonly used as an intro language into computing sciences with students. This allows students to easily learn computing theories and concepts and then apply them to other programming languages.

3. METHODOLOGY

3.1. SIMPLE LINEAR REGRESSION

Simple Linear Regression Model:

$Y_i = \beta_0 + \beta_1 X_i + €_i$, i = 1,2,,……..,n

Where Y = Dependent Variable

X = Independent Variable

β_0 = Intercept parameter (which captures the influences of all known variables on Y)

β_1 = Slope parameter or Regressor constant of Y on X (which gives the effect of independent variable X on Y)

$€_i$ = Error or Random or Disturbance variable (which captures the influences of all unknown variables)

SOURCES OF ERROR VARIABLE

The main sources of error variable are

(i) Due to omission of some variables which have influence on dependent variable Y

(ii) Due to the misspecification of the linear model

(iii) Due to aggregation of the values on the variables and

(iv) Error involved in the measurement of the variables (Measurement Error) etc.

ASSUMPTIONS ON € OR LINEAR MODEL

The crucial assumptions about error variable or the linear model are given by

(i) E(€$_i$) = 0 $\Longrightarrow$ All the error observations have zero means (By common sense , the average of the effects of all unknown variables is assumed to be zero)
(Unbiased Assumption)

(ii) E(€$_i$€$_j$) = σ^2 , $\forall\, i = j = 1,2, \ldots \ldots \ldots, n$
= 0, $\forall\, i \neq j$
$\Longrightarrow$ All the €$_I$'s are uncorrelated with each other and they have the same unknown varianceσ^2.
(Homoscadastic Assumption)

(iii) €$_i$'s follow Normal distribution with zero mean and the unknown error varianceσ^2. i.e
€$_i$$\sim i.i.d\ N(0,\sigma^2)$

LEAST SQUARES ESTIMATION OR ORDINARY LEAST SQUARES (OLS) ESTIMATORS FOR THE PARAMETERS β$_0$ AND β$_1$

Let $\widehat{\beta_0}$ and$\widehat{\beta_1}$ be the least squares estimators of β$_0$ and β$_1$ respectively. Now the estimated or fitted linear regression model can be written as

$$\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_i, i = 1,2, \ldots \ldots, n$$

Define the residuals from fitted straight line as

$$e_i = Y_i - \widehat{Y_i} \text{ or } e = Y - \hat{Y} = Y - \widehat{\beta_0} - \widehat{\beta_1}\, X$$

By the principle of least squares, we minimise the residual sum of squares w.r.t $\widehat{\beta_0}$ and$\widehat{\beta_1}$ to obtain he set of normal equations. The solutions of the normal equations give the OLS estimators of the parameters β$_0$ and β$_1$.

The first order conditions for the minimization of the residual sum of squares are given by,

(i) $\frac{\partial \sum e_i^2}{\partial \widehat{\beta_0}} = 0 \Rightarrow \frac{\partial}{\partial \widehat{\beta_0}} \left[\sum_{i=1}^{n} \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i\right)^2\right] = 0$

$\Rightarrow -2 \sum_{i=1}^{n} \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i\right) = 0$

$$\Rightarrow n\,\widehat{\beta_0} + \widehat{\beta_1} \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \qquad \longrightarrow (3.1)$$

(ii) $\frac{\partial \sum e_i^2}{\partial \widehat{\beta_1}} = 0 \Rightarrow \frac{\partial}{\partial \widehat{\beta_1}} \left[\sum_{i=1}^{n} \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i\right)^2\right] = 0$

$\Rightarrow -2 \sum_{i=1}^{n} \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i\right) X_i = 0$

$\Rightarrow \widehat{\beta_0} \sum_{i=1}^{n} X_i + \widehat{\beta_1} \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i\, Y_i \longrightarrow (3.2)$

The two normal equations (1) and (2) are given by,

$$\boxed{\begin{aligned} \sum_{i=1}^{n} Y_i &= n\,\widehat{\beta_0} + \widehat{\beta_1} \sum_{i=1}^{n} X_i & \longrightarrow (3.1) \\ \sum_{i=1}^{n} X_i\, Y_i &= \widehat{\beta_0} \sum_{i=1}^{n} X_i + \widehat{\beta_1} \sum_{i=1}^{n} X_i^2 & \longrightarrow (3.2) \end{aligned}}$$

The solutions or the OLS estimators for β$_0$ and β$_1$ are given by

(i) $\widehat{\beta_1} = \dfrac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}$ or $\widehat{\beta_1} = \dfrac{\sum xy}{\sum x^2}$ $\longrightarrow I$

Where $\sum x^2 = \sum X^2 - \frac{(\sum X)^2}{n}$ and $\sum xy = \sum XY - \frac{(\sum X)(\sum Y)}{n}$

Here $\sum x^2$= Corrected sum of squares of X

$\sum xy$ = Corrected sum of products of X and Y.

Similarly we can write $\sum y^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$ = Corrected sum of squares of Y

(ii) $\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}$ $\rightarrow II$

where $\bar{X} = \frac{\sum X}{n}$ and $\bar{Y} = \frac{\sum Y}{n}$

(iii) An unbiased estimate of the unknown variance σ^2 is given by

$\widehat{\sigma^2} = \frac{\sum e^2}{n-2} = \frac{\sum[Y-\hat{Y}]^2}{n-2}$ Or $\widehat{\sigma^2} = \frac{\sum[Y-\widehat{\beta_0}-\widehat{\beta_1}X]^2}{n-2}$

In the simplified form, $\widehat{\sigma^2} = \frac{\sum y^2 - \widehat{\beta_1}\sum xy}{n-2} = \frac{Residual\ Sum\ of\ Squares}{Error\ d.f}$

PROPERTIES OF THE LEAST SQUARES ESTIMATORS $\widehat{\beta_0}$ AND $\widehat{\beta_1}$.

(i) The OLS estimators $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are the linear functions of the y_i's.

Since $\widehat{\beta_1} = \frac{\sum x_i y_i}{\sum x_i^2} = \sum w_i y_i$ where $w_i = \frac{x_i}{\sum x_i^2} = the\ weights$

$$\left[\because \sum w_i = 0, \sum w_i^2 = \frac{1}{\sum x_i^2}\ and \sum w_i x_i = \sum w X_i = 1\right]$$

Similarly $\widehat{\beta_0}$ can be written as linear combination of y_i's.

(ii) The OLS estimators $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are unbiased estimators of β_0 and $\beta_1$respectively.

i.e E $[\widehat{\beta_1}] = \beta_1$ and E $[\widehat{\beta_0}] = \beta_0$

(iii) The variances of $\widehat{\beta_1}$ and $\widehat{\beta_0}$ are given by $Var\left(\widehat{\beta_1}\right) = E\left[\widehat{\beta_1} - \beta_1\right]^2$

Or $Var\left(\widehat{\beta_1}\right) = \frac{\sigma^2}{\sum x^2}$ $\rightarrow III$

Where $\sum x^2 = \sum X^2 - \frac{(\sum X)^2}{n}$

$Var\left(\widehat{\beta_0}\right) = E\left[\widehat{\beta_0} - \beta_0\right]^2$

Or $Var\left(\widehat{\beta_0}\right) = \sigma^2 = \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum x^2}\right]$ $\rightarrow IV$

It can be easily shown that $\widehat{\beta_0}$ and $\widehat{\beta_1}$ having the minimum least variances among all the linear unbiased estimators for β_0 and β_1. Hence we say that $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are the "Best Linear Unbiased Estimators" (BLUE's) for the parameters β_0 and β_1 respectively.

Also $\widehat{\sigma^2}$ is an unbiased estimator of unknown variance

$\widehat{\sigma^2} = \left[\frac{\sum y^2 - \widehat{\beta_1}\sum xy}{n-2}\right]$ $\rightarrow V$

Note: The formulae I, II, III, IV and V are important to remember to deal with the inference or tests of significance for the parameters of the simple linear regression model.

TESTS OF THE SIGNIFICANCE OF $\widehat{\beta_0}$ AND $\widehat{\beta_1}$

I TEST OF THE SIGNIFICANCE OF $\widehat{\beta_1}$

H_0: $\beta_1 = 0$ or $\widehat{\beta_1} = 0$ i.e There is no significant effect of X on Y or $\widehat{\beta_1}$ is not significant.

To test the H_0, the Student's t-test statistic is given by,

$$t = \frac{\widehat{\beta_1}}{S.E(\widehat{\beta_1})}\ where\ \widehat{\beta_1} = \frac{\sum xy}{\sum x^2}$$

IJCRT

Or $t = \frac{\widehat{\beta_1}\sqrt{\sum x^2}}{\hat{\sigma}} \sim t_{n-2}\, d.f$

Where $\hat{\sigma} = \sqrt{\left[\frac{\sum y^2 - \widehat{\beta_1}\sum xy}{n-2}\right]}$ and

$\sum x^2 = \sum X^2 - \frac{(\sum X)^2}{n}$

$\sum xy = \sum XY - \frac{(\sum X)(\sum Y)}{n}$

$\sum y^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$

We compare t_{cal} with t critical value for n-2 d.f at either 5% or 1% level of significance and draw the inference accordingly.

II TEST OF THE SIGNIFICANCE OF $\widehat{\beta_0}$

We state the null hypothesis as H_0: $\beta_0 = 0$ or $\widehat{\beta_0} = 0$ i.e the intercept estimator $\widehat{\beta_0}$ is not significant. To test the H_0, the Student's t-test statistic is given by

$t = \frac{\widehat{\beta_0}}{S.E(\widehat{\beta_0})}$ *where* $\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}.$ Here $\bar{X} = \frac{\sum X}{n}$ and $\bar{Y} = \frac{\sum Y}{n}$

$$Or\ \ t = \frac{\widehat{\beta_0}}{\hat{\sigma}\sqrt{\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum x^2}\right]}} \sim t_{n-2}\, d.f$$

We compare t_{cal} value with the t critical value for n-2 d.f at either 5% or 1% level of significance and draw the inference accordingly.

III PREDICTION OF Y FOR GIVEN VALUE X = X_0

The estimated simple linear regression model is given by $\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1}X.$ This linear regression equation of Y on X can be used to predict the value of Y for a given value of X.

Suppose X = X_0(given), the predicted value of Y for given X = X_0 is given by

$$\widehat{Y_0} = \widehat{\beta_0} + \widehat{\beta_1}X_0.$$

ANOVA FOR TWO VARIABLE LINEAR MODEL

Consider the residuals for two variable linear model as

$e_i = (y_i - \widehat{y_i}) = (y_i - \widehat{\beta_1}x_i)$or $e = (y - \widehat{\beta_1}x)$

Where y_i's and x_i's are in the deviation form i.e $y = [Y - \bar{Y}]$ and x= $[X - \bar{X}]$

i.e e = y- $\widehat{\beta_1}$x

or y = $\widehat{\beta_1}$ x + e

$$\Rightarrow \sum y^2 = \sum [\widehat{\beta_1}x + e]^2$$

$$= \sum \left[\widehat{\beta_1}^2 x^2 + e^2 + 2\widehat{\beta_1}xe\right]$$

$$\Rightarrow \sum y^2 = \widehat{\beta_1}^2 \sum x^2 + \sum e^2 + 2\,\widehat{\beta_1}\sum xe$$

$\Rightarrow \sum y^2 = \widehat{\beta_1}(\widehat{\beta_1}\sum x^2) + \sum e^2 + 2\ \widehat{\beta_1}\ (0)\ \ [\because \sum xe = 0]$

$\Rightarrow \sum y^2 = \widehat{\beta_1}\sum xy + \sum e^2 \rightarrow$ ANOVA Model for two variable linear model.

$$[\because \widehat{\beta_1} = \frac{\sum xy}{\sum x^2}]$$

Or Total Sum of Squares = Regression Sum of Squares + Residual Sum of Squares

TSS = Reg SS + Res SS

Where Total SS = $\sum y^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$

Regression SS = $\widehat{\beta_1}\sum xy = \widehat{\beta_1}\ [\sum xy = \sum XY - \frac{(\sum X)(\sum Y)}{n}]$

Residual SS = [Total SS – Regression SS]

IJCRT

ANOVA TABLE FOR TWO VARIABLE LINEAR MODEL

S.V	d.f	SS	MSS	F_{cal}
Due to Regression (X)	2-1 = 1	$\widehat{\beta_1}\sum xy$	$\frac{\widehat{\beta_1}\sum xy}{2-1}$	$\frac{r^2/2-1}{1-r^2/n-2}$
Residual	n-2	$\sum y^2 - \widehat{\beta_1}\sum xy = \sum e^2$	$\frac{\sum e^2}{n-2}$	
Total	n-1	$\sum y^2$		

SIMPLE COEFFICIENT OF DETERMINATION (r^2)

For a simple linear regression model, the square of the correlation coefficient r^2 determines the validity of the linear regression model and r^2 is called the 'coefficient of simple determination' of linear regression of Y on X.

We have, TSS = Regression SS + Residual SS, r^2 is defined as the ratio of the Regression SS to the Total SS.

i.e $r^2 = \frac{Regression\ SS}{Total\ SS}$ or $r^2 = \frac{\widehat{\beta_1}\sum xy}{\sum y^2}$

or $r^2 = 1 - \frac{Residual\ SS}{Total\ SS} = 1 - \frac{\sum e^2}{\sum y^2}$

The value of r^2 always lies between 0 and 1.

i.e $0 \leq r^2 \leq 1$

The quantity $(r^2)100$ gives the percentage of the variation in dependent variable Y explained by linear influence of the given independent variable X.

$r^2 = \frac{\widehat{\beta_1}\sum xy}{\sum y^2}$, $\sum y^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$\therefore\ t^2 = \frac{r^2(n-2)}{(1-r^2)}$$

$$\therefore\ t^2 = F = \frac{r^2/1}{1-r^2/n-2} \sim F_{(1,n-2)}$$

3.2. JUPITOR

In 2014, Fernando Perez, software developer announced a spin-off project from IPython called Project Jupyter. IPython continued to exist as a Python shell and a kernel for Jupyter, while the notebook and other language-agnostic parts of IPython moved under the Jupyter name. Jupyter is language agnostic and it supports execution environments (aka kernels) in several dozen languages among which are Julia, R, Haskell, Ruby, and of course Python (via the IPython kernel). In 2015, GitHub and the Jupyter Project announced native rendering of Jupyter notebooks file format (.ipynb files) on the GitHub platform.

Project Jupyter's operating philosophy is to support interactive data science and scientific computing across all programming languages via the development of open-source software. According to the Project Jupyter website, "Jupyter will always be 100% open-source software, free for all to use and released under the liberal terms of the modified BSD license."

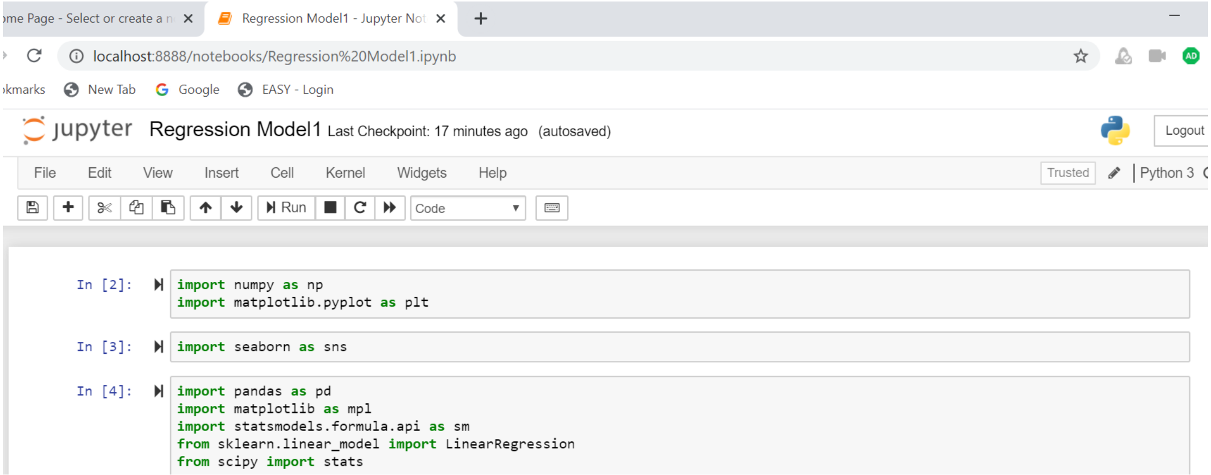
The Jupyter Notebook has become a popular user interface for cloud computing, and major cloud providers have adopted the Jupyter Notebook or derivative tools as a frontend interface for cloud users. Examples include Amazon's SageMaker Notebooks, Google's Colaboratory and Microsoft's Azure Notebook.

4. EMPERICAL INVESTIGATION

To test the relation between IMR and Literacy Rate we have collected the data of different states/union territories of India from Reserve Bank Publications. The data is recorded as

States/Union Territories	Literacy Rate in 2011 (Per Cent)	Infant Mortality Rate in 2016 (Per Thousand)
Andaman and Nicobar Islands	86.63	16
Andhra Pradesh	67.02	34
Arunachal Pradesh	65.39	36
Assam	72.19	44
Bihar	61.8	38
Chandigarh	86.05	14
Chhattisgarh	70.28	39
Dadra and Nagar Haveli	76.24	17
Daman and Diu	87.1	19
Delhi	86.21	18
Goa	88.7	8
Gujarat	78.03	30
Haryana	75.55	33
Himachal Pradesh	82.8	25
Jammu and Kashmir	67.16	24
Jharkhand	66.41	29
Karnataka	75.37	24
Kerala	94	10
Lakshadweep	91.85	19
Madhya Pradesh	69.32	47
Maharashtra	82.34	19
Manipur	76.9	11
Meghalaya	74.43	39
Mizoram	91.33	27
Nagaland	79.6	12
Odisha	72.89	44
Puducherry	85.85	10
Punjab	75.84	21
Rajasthan	66.11	41
Sikkim	81.42	16
Tamil Nadu	80.09	17
Tripura	87.22	24
Uttar Pradesh	67.68	43
Uttarakhand	78.82	38
West Bengal	76.26	25

First we will study the relationship between IMR and Literacy Rate through the plot graph using Jupitor platform in Python. The libraries required for the application and the plot are taken in the screen shot of Jupitor which shown below.

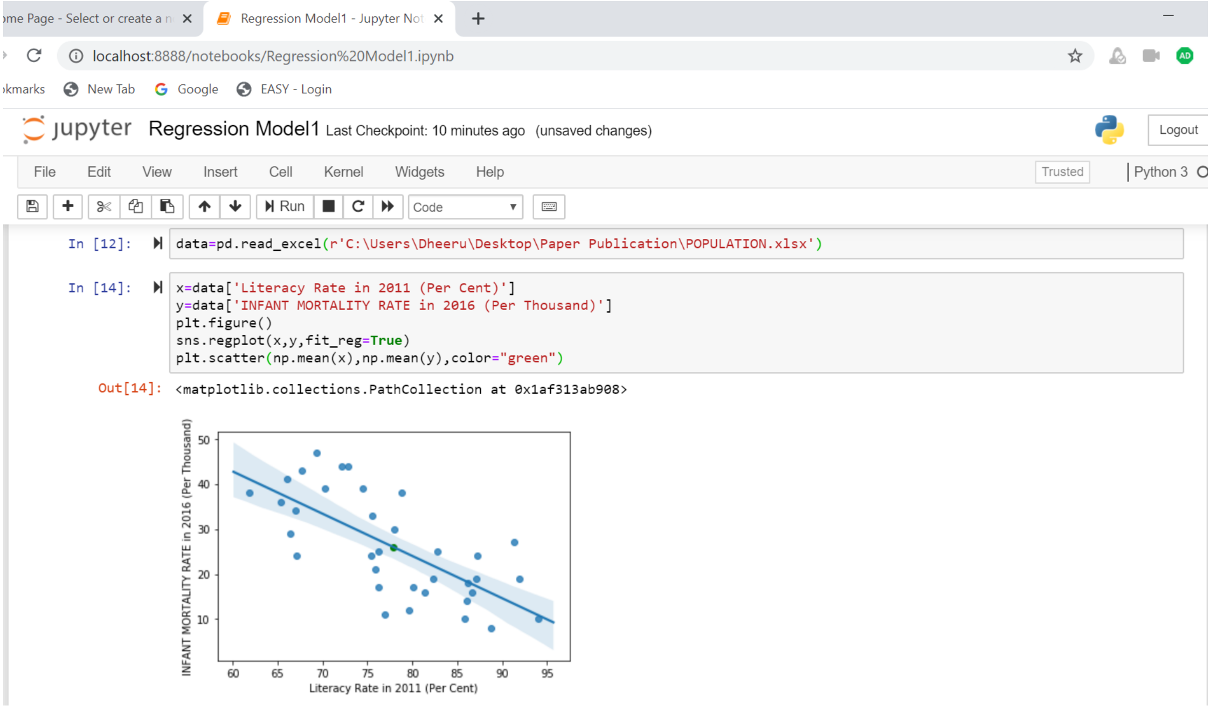


ome Page - Select or create a n
Regression Model1 - Jupyter Not
localhost:8888/notebooks/Regression%20Model1.ipynb
kmarks
New Tab
Google
EASY - Login
jupyter
Regression Model1 Last Checkpoint: 17 minutes ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code

```
In [2]:  import numpy as np
         import matplotlib.pyplot as plt

In [3]:  import seaborn as sns

In [4]:  import pandas as pd
         import matplotlib as mpl
         import statsmodels.formula.api as sm
         from sklearn.linear_model import LinearRegression
         from scipy import stats
```


Regression Model1 - Jupyter Not
localhost:8888/notebooks/Regression%20Model1.ipynb
jupyter Regression Model1 Last Checkpoint: 10 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
INFANT MORTALITY RATE in 2016 (Per Thousand)
Literacy Rate in 2011 (Per Cent)
10
20
30
40
50
60
65
70
75
80
85
90
95

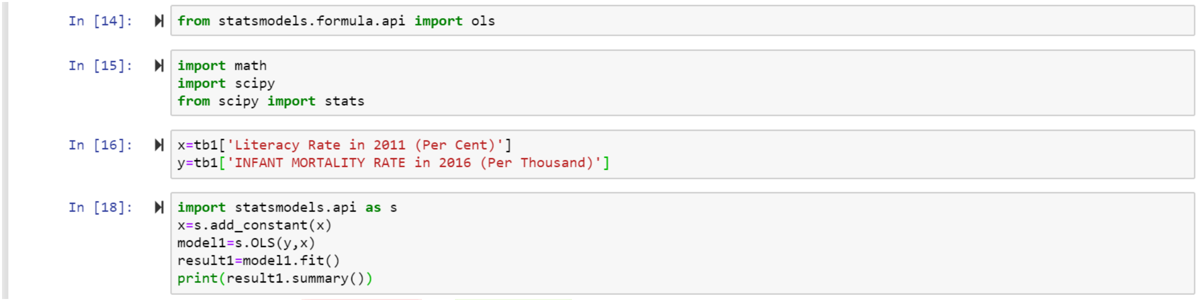
```
In [12]:  data=pd.read_excel(r'C:\Users\Dheeru\Desktop\Paper Publication\POPULATION.xlsx')

In [14]:  x=data['Literacy Rate in 2011 (Per Cent)']
          y=data['INFANT MORTALITY RATE in 2016 (Per Thousand)']
          plt.figure()
          sns.regplot(x,y,fit_reg=True)
          plt.scatter(np.mean(x),np.mean(y),color="green")

Out[14]: <matplotlib.collections.PathCollection at 0x1af313ab908>
```

From the plot it is clear that there exists linear relationship between IMR and Literacy Rate. We can observe the states with lower Literacy Rate witnessed high Infant Mortality Rate and vice-versa.

Now linear relationship can be studied through Simple Linear regression Model through Jupitor platform of Python. The screenshot of Regression Analysis is given below.

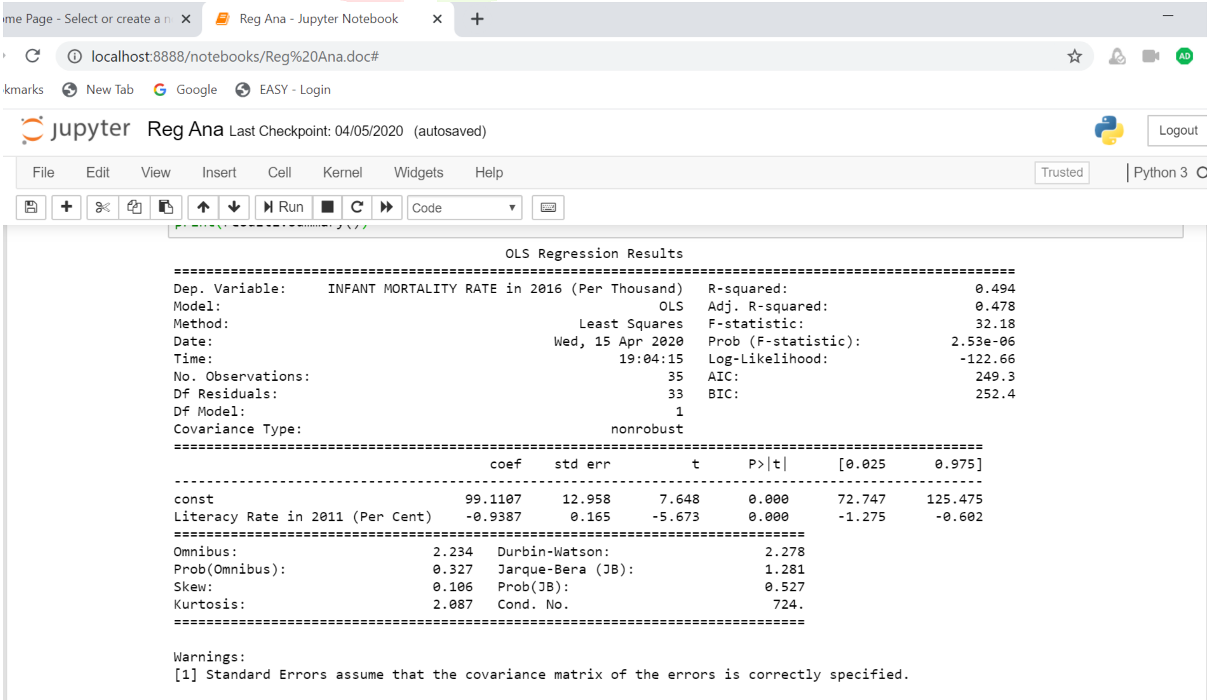


```
In [14]:  from statsmodels.formula.api import ols

In [15]:  import math
          import scipy
          from scipy import stats

In [16]:  x=tb1['Literacy Rate in 2011 (Per Cent)']
          y=tb1['INFANT MORTALITY RATE in 2016 (Per Thousand)']

In [18]:  import statsmodels.api as s
          x=s.add_constant(x)
          model1=s.OLS(y,x)
          result1=model1.fit()
          print(result1.summary())
```


Reg Ana - Jupyter Notebook
localhost:8888/notebooks/Reg%20Ana.doc#
jupyter Reg Ana Last Checkpoint: 04/05/2020 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code

```
                                      OLS Regression Results
=======================================================================================================
Dep. Variable:     INFANT MORTALITY RATE in 2016 (Per Thousand)   R-squared:                       0.494
Model:                                                      OLS   Adj. R-squared:                  0.478
Method:                                           Least Squares   F-statistic:                     32.18
Date:                                          Wed, 15 Apr 2020   Prob (F-statistic):           2.53e-06
Time:                                                  19:04:15   Log-Likelihood:                -122.66
No. Observations:                                            35   AIC:                             249.3
Df Residuals:                                                33   BIC:                             252.4
Df Model:                                                     1
Covariance Type:                                      nonrobust
====================================================================================================
                                       coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------
const                               99.1107     12.958      7.648      0.000      72.747     125.475
Literacy Rate in 2011 (Per Cent)    -0.9387      0.165     -5.673      0.000      -1.275      -0.602
==============================================================================
Omnibus:                        2.234   Durbin-Watson:                   2.278
Prob(Omnibus):                  0.327   Jarque-Bera (JB):                1.281
Skew:                           0.106   Prob(JB):                        0.527
Kurtosis:                       2.087   Cond. No.                         724.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

5. INTERPRETATION

The observation shows that there is significant relation between IMR and Literacy Rate. The Simple Linear Regression Model between IMR and Literacy Rate from the given data is given by

$$IMR = 99.1107 - 0.9387(Literacy\ Rate)$$

The Intercept parameter and Slope parameters are showing significant effect. The overall model is also showing significant effect. The coefficient of simple determination ‘r^2’ gives 49% variation in IMR can be explained by the linear influence of Literacy Rate.

So finally we interpret that even though the parameters and overall model are significant, the Literacy Rate influences IMR moderately i.e nearly 50% only for this data. Those who are interested can analyse through residual analysis, whether the assumptions on residual are how for satisfied with the Python.

Reference:

1. Handbook of Statistics on Indian States, Reserve Bank of India, 2018-19
2. Wikipedia, The New England Journal of Medicine
3. Wikipedia, The Economist
4. Ranking of states and union territories by literacy rate: 2011 Census of India Report (2013)
5. “Primary Education in India: Key Problems” (PDF). Source: Dise. Retrived 15 September 2011.
6. “Educating India”. Source: Scribd. Retrived 15 September 2011.
7. “India’s Literacy Panorama”. Source: Education for all in India. Retrived 15 September 2011.
8. “Gender Inequalities and Demographic Behaviour” (PDF). Source: Snap3. Retrived 15 September 2011.
9. David A. Freedman (2009). Statistical Models: Theory and Practice. Cambridge University Press.
10. Guttag, John V. (12 August 2016). Introduction to Computation and Programming Using Python: With Application to Understanding Data.
11. “Project Jupitor – about Us”. 2018-04-20. Retrieved 2018-05-03.
12. “Project Jupytor // Speaker Deck”.