Multiple Linear Regression

Introduction

Content

1. Introduction to MLR
2. Statistical Model of MLR
3. Basic Data Checks
4. Model Fitting
5. Parameter Estimation – Ordinary Least Squares Method
6. Interpretation of Partial Regression Coefficients

Multiple Linear Regression

- Multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables.
- The independent variables can be continuous or categorical.
- Multiple Linear Regression is used when we want to predict the value of a variable based on the values of two or more other variables.
- The variable we want to predict is called the dependent variable
- The variables used to predict the value of dependent variable are called independent variables (or explanatory variables/predictors).
- Multiple linear regression requires the model to be linear in the parameters.
- Example: The price house in USD can be dependent variable and area of house, location of house , air quality index in the area, distance from airport etc. can be independent variables.

Statistical Model

$$\boxed{Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p + e}$$

where,

Y	: Dependent Variable
$X_1, X_2 ,\ldots, X_p$	: Independent Variables
$b_0, b_1 ,\ldots, b_p$	: Parameters of Model
e	: Random Error Component

- Independent variables can either be **Continuous or Categorical**
- Multiple linear regression **requires the model to be linear in the parameters**
- Parameters of the model are estimated by Least Square Method.
- The **least squares (LS)** criterion states that the **sum of the squares of errors** (or residuals) **is minimum**.
- Mathematically, following quantity is minimized to estimate parameters using least square method.
- **Error ss=** Σ **(Yi –** $\hat{Yi}$ **)2**

Case Study – Modeling Job Performance Index

Background

- A company conducts different written tests before recruiting employees. The company wishes to see if the scores of these tests have any relation with post-recruitment performance of those employees.

Objective

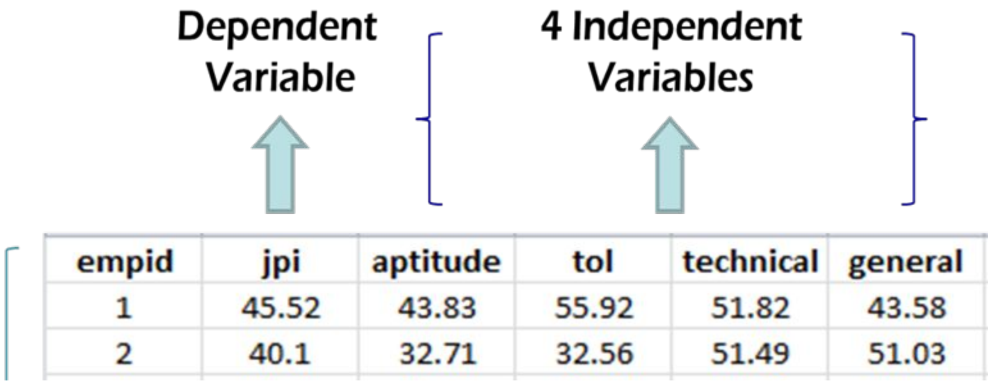
- To predict employees' job performance index after probationary period, based on scores of tests conducted at the time of recruitment

Available Information

- Sample size is 33
- Independent Variables: Scores of tests conducted before recruitment on the basis of four criteria – Aptitude, Test of Language, Technical Knowledge, General Information
- Dependent Variable: Job Performance Index calculated after an employee finishes probationary period (6 months)

Data Snapshot

Performance Index



Dependent Variable

4 Independent Variables

empid	jpi	aptitude	tol	technical	general
1	45.52	43.83	55.92	51.82	43.58
2	40.1	32.71	32.56	51.49	51.03

Columns	Description	Type	Measurement	Possible values
empid	Employee ID	integer	-	-
jpi	Job performance Index	numeric	-	positive values
aptitude	Aptitude score	numeric	-	positive values
tol	Test of Language	numeric	-	positive values
technical	Technical Knowledge	numeric	-	positive values
general	General Information	numeric	-	positive values

Graphical Representation of Data

- It is always recommended to have a general look at your data and behavior of all the variables before moving to modeling.

- This helps you in making intuitive inferences about the data, which can be statistically validated by your final model.

- The simplest way of doing this is creating a scatter plot matrix, which will give bivariate relationships between variables.

```
#Importing the Data
```

```r
perindex<-read.csv("Performance Index.csv",header=TRUE)
```

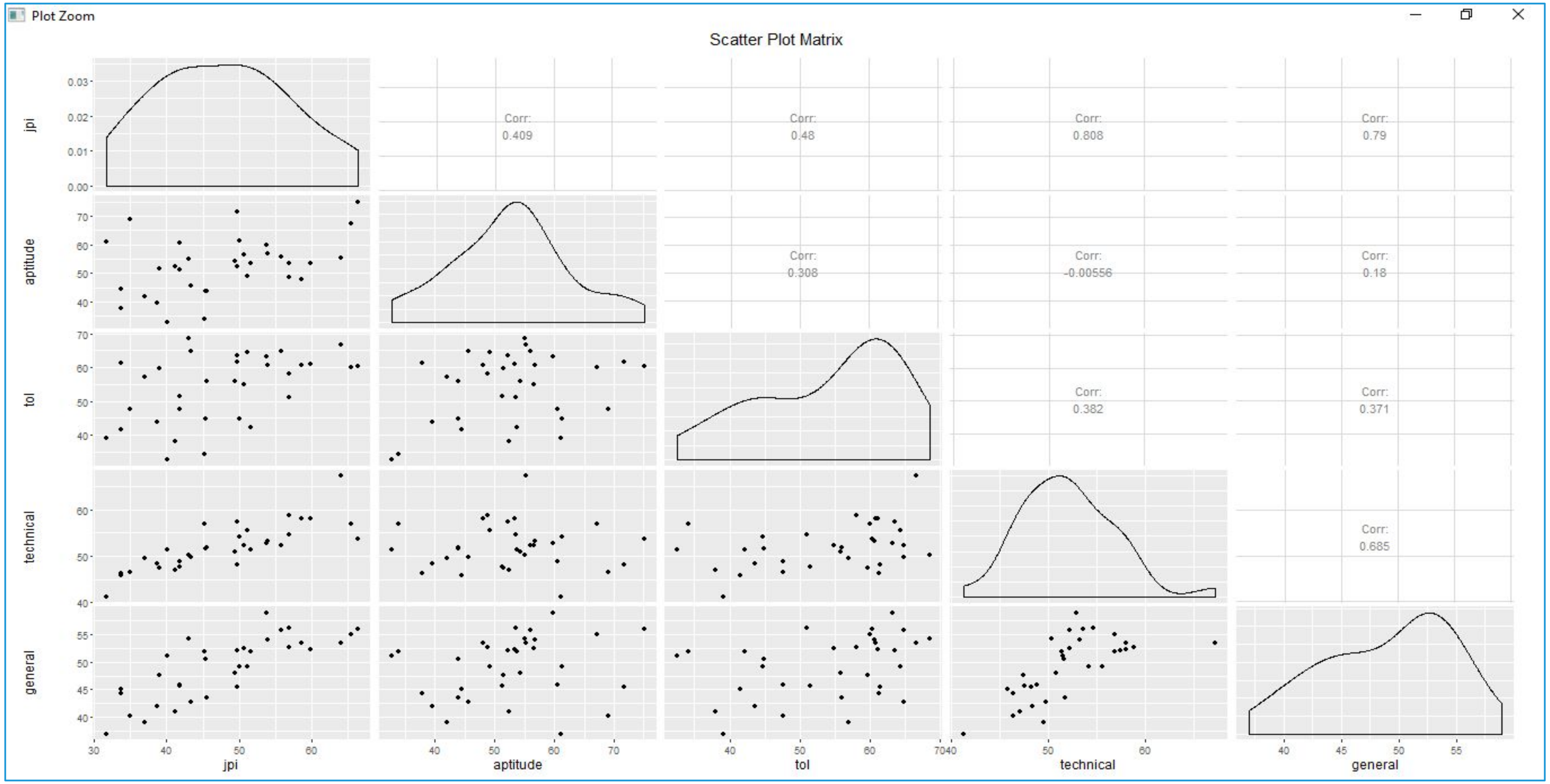
```
#Graphical Representation of the Data
```

```r
library(GGally)

ggpairs(perindex[,c("jpi","aptitude","tol","technical","general")],
        title="Scatter Plot Matrix",
        columnLabels=c("jpi","aptitude","tol","technical","general"))
```

Scatter Plot Matrix

> `ggpairs()` function in package `GGally` gives a Generalised Pairs Plot which not only visualises bivariate scatter relationships, but also gives their quantified representation, in the form of Correlation Coefficients, along with Distribution for each variable



Simple v/s Multiple Linear Regression

- Using simple linear regression to solve such a problem is not wrong. For instance, we can study the impact of aptitude on job performance, then see the impact of technical expertise on job performance and so on. But is this approach efficient? Certainly not!

Why is Multiple Linear Regression a better method than Simple Linear Regression?

Single predictor provides inadequate information about the response variable

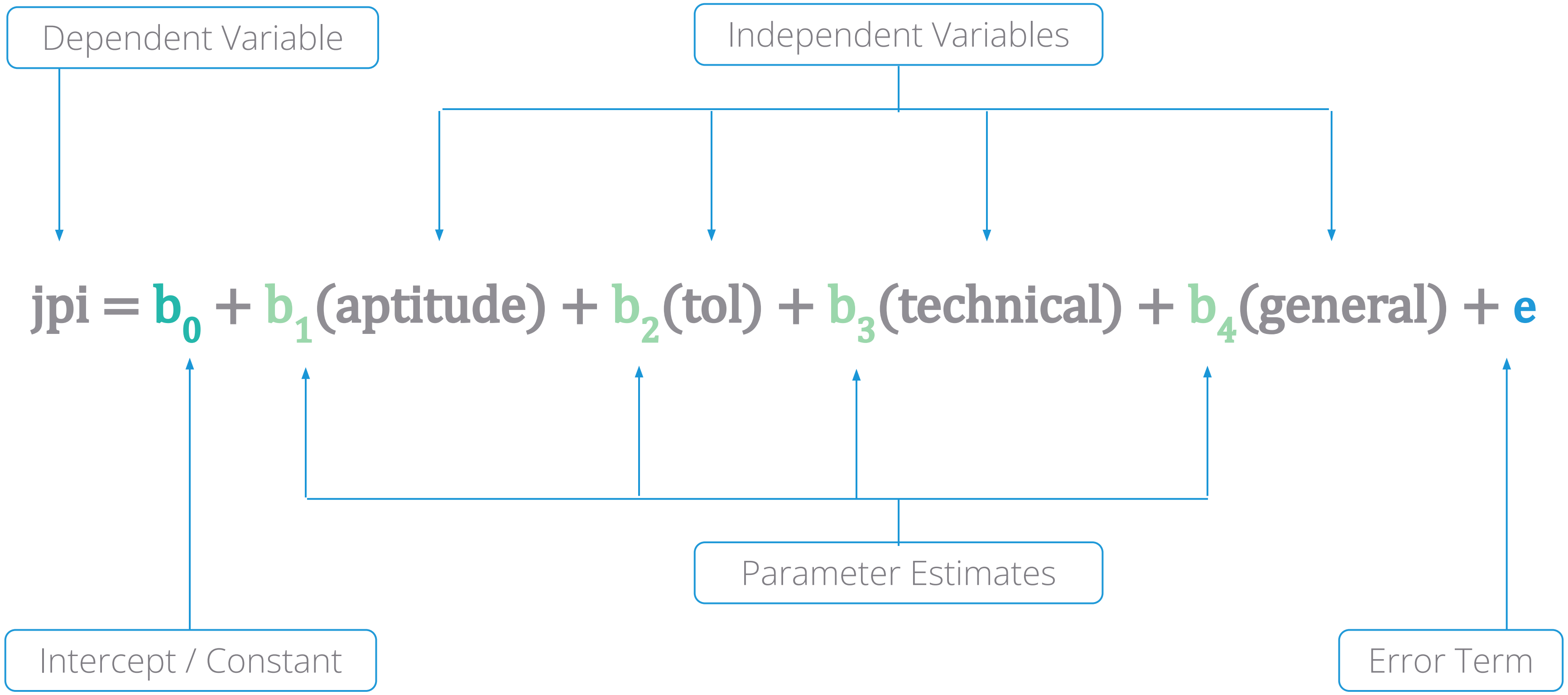
Simultaneous study of multiple variables is essential as the response is always influenced by more than one variables

Questions that MLR Answers

Multiple linear regression analysis of employee performance indices answers the following questions:

1. Do tests conducted at the time of recruitment determine a candidate's performance in the first six months in that job?
2. Which of the four test scores is most significant in determining job performance?
3. Can any of the tests be discontinued?
4. Can performance of newly recruited candidates be estimated based on the scores of tests conducted at the time of their recruitment?

Model for the Case Study



Dependent Variable

Independent Variables

$$\text{jpi} = b_0 + b_1(\text{aptitude}) + b_2(\text{tol}) + b_3(\text{technical}) + b_4(\text{general}) + e$$

Parameter Estimates

Intercept / Constant

Error Term

Parameter Estimation using Least Square Method

Parameters	Coefficients
Intercept	-54.2822
aptitude	0.3236
tol	0.0334
technical	1.0955
general	0.5368

E(jpi)= -54.2822 + 0.3236 (aptitude) + 0.0334 (tol) + 1.0955 (technical) + 0.5368 (general) + e

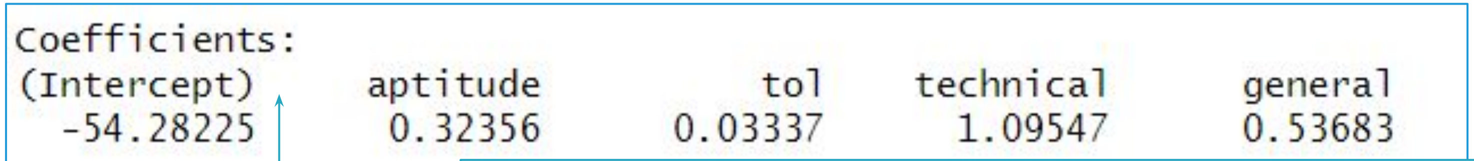
Parameter Estimation Using lm function in R

```
#Model Fit
```

```
jpimodel<-lm(jpi~aptitude+tol+technical+general, data=perindex)
jpimodel
```

- ❑ *lm() fits a linear regression.*
- ❑ *~ separates dependent and independent variables*
- ❑ *Left hand side of tilde(~) represents the dependent variable and right-hand side shows independent variables*
- ❑ *+ separates multiple independent variables.*

```
#Output
```



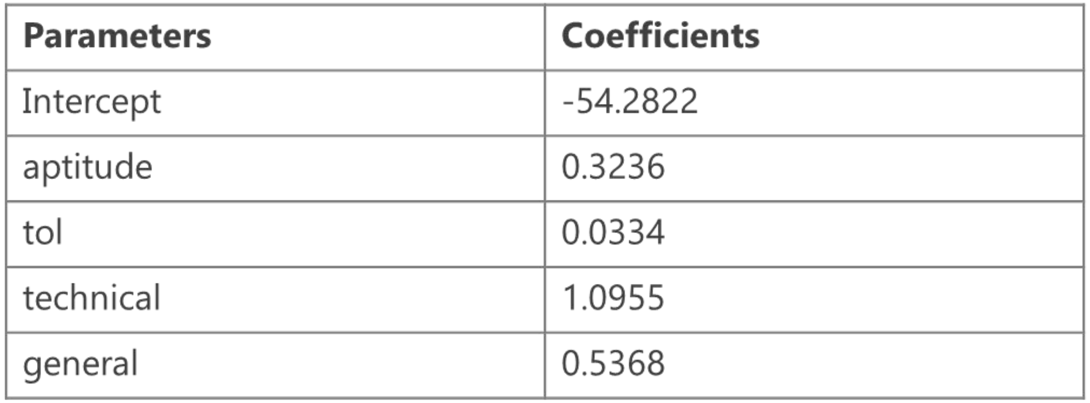
```
Coefficients:
(Intercept)     aptitude          tol    technical      general
  -54.28225      0.32356      0.03337      1.09547      0.53683
```

- □ *Coefficients are the model parameters.*
- □ *Signs of each parameter represent their relationship with the dependent variable.*

* **~. in lm() function uses all variables except the dependent variable. This is helpful when the data has a large number of predictors.**

Interpretation of Partial Regression Coefficients

- For every unit increase in the independent variable (X), the expected value of the dependent variable (Y) will change by the corresponding parameter estimate (b), keeping all the other variables constant



Parameters	Coefficients
Intercept	-54.2822
aptitude	0.3236
tol	0.0334
technical	1.0955
general	0.5368

- From the parameter estimates table, we observe that the parameter estimate for Aptitude Test is 0.3236

  We can infer that for one unit increase in aptitude test score, the expected value of job performance index will increase by 0.3236 units

Quick Recap

Understand the Data

- Ensure the data is complete and consistent
- Identify dependent and independent variables

Simple Data Check

- Use **`pairs()`** function to yield a simple scatter plot
- Use **`ggpairs()`** function from **`GGally`** package for a more nuanced plot (Recommended)

Fit a Model

- Run a regression and obtain ordinary least square estimates of parameters
- lm() function fits a linear regression model