Introduction to Multiple Linear Regression - I

Content

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 a dependent variable and sze of house, location of house , air quality index in the area, distance from airport etc. can be independent variables.

Statistical Model

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p + e$$

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, the following quantity is minimized to estimate parameters using the least square method.
- Error ss= $\Sigma (Yi - \hat{Yi})2$

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 a probationary period, based on test scores conducted at the time of recruitment

Available Information

- Sample size is 33
- Independent Variables: Test scores 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 a probationary period (6 months)

Data Snapshot

Performance Index

Dependent Variable

4 Independent Variables

Observations

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
3	50.61	56.64	54.84	52.29	52.47

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 variables before moving to modelling.
- This helps you to make intuitive inferences about the data, which can be statistically validated by your final model.
- The simplest way of doing this is to create a scatter plot matrix, which will give bivariate relationships between variables.

```
#Importing the Data
```

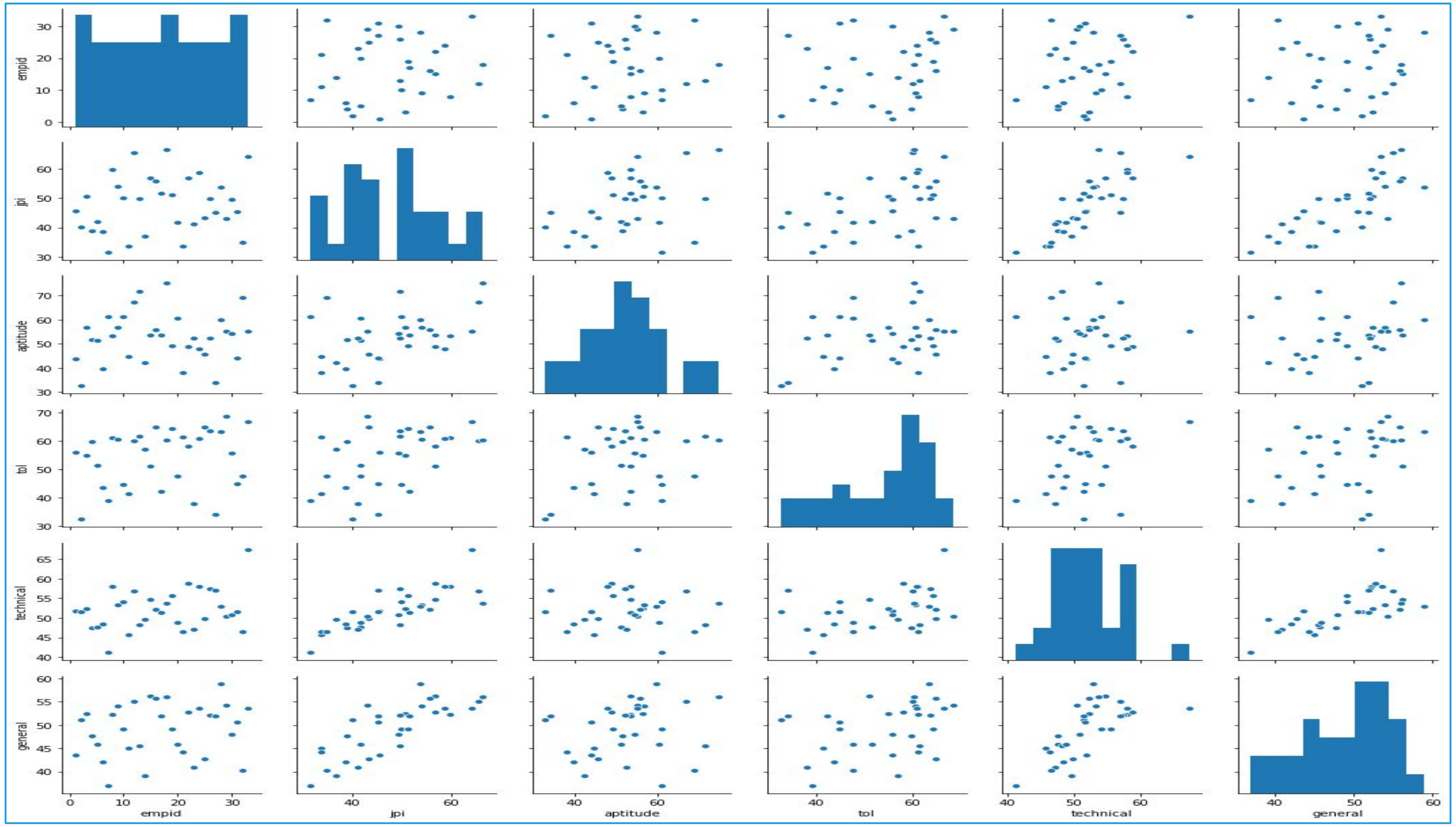
```
import pandas as pd
perindex = pd.read_csv("Performance Index.csv")
```

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

```
import seaborn as sns
sns.pairplot(perindex)
```

Scatter Plot Matrix

The `pairplot()` function in the **`seaborn library`** gives a scatter plot matrix and distribution of all variables using histograms.



Model for the Case Study

Dependent Variable

Independent Variables

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

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

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

Parameter Estimation Using ols() function in Python

#Model Fit

```
import statsmodels.formula.api as smf

jpimodel=smf.ols('jpi ~ tol + aptitude + technical +general',
data=perindex).fit()

jpimodel.params
```

- *ols() 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

```
Intercept    -54.282247
tol            0.033372
aptitude       0.323562
technical      1.095467
general        0.536834
dtype: float64
```

Interpretation :

- *jpimodel.params gives the model parameters.*
- *Signs of each parameter represent their relationship with the dependent variable.*

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 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

In this session we have covered the **basics of multiple linear regression using Python**.

Follow these simple steps to carry out your first analysis:

Understand the Data	• Ensure the data is complete and consistent • Identify dependent and independent variables
Data Visualization	• `pairplot()` function from **`seaborn`** library gives scatter plot matrix
Fit a Model	• **ols()** function from library **statsmodels** fits a linear regression model