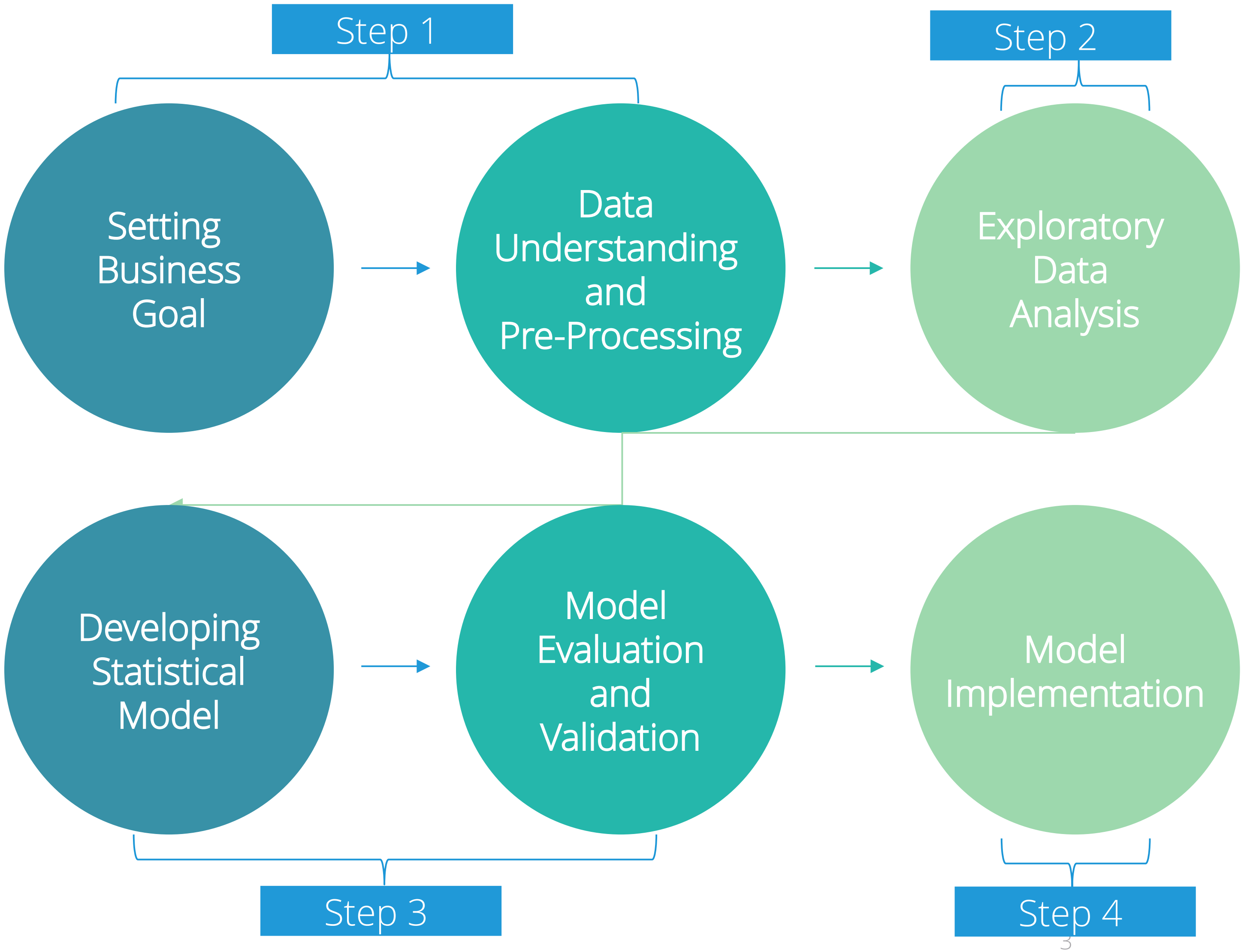
Introduction to Predictive Modelling

What is Predictive modelling?

- Statistical model created **to best predict** → the outcome or probability of an outcome
- Models **developed using** → Historical data or Purposely collected data
- Predictive analytics is used in **financial services, insurance, telecommunications, retail, travel, healthcare, pharmaceuticals, sports and several other fields**

Predictive modelling – General Approach



Multiple Linear Regression

Introduction

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.
- The variable we want to model/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

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p + e$$

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 the Least Squares 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 squares method.
- **Error ss= $\Sigma (Yi - \hat{Y}i)2$**

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

Model for the Case Study

Dependent Variable

Independent Variables

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

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

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

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
- 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() func
helpful when

Interpretation of Partial Regression Coefficients

- For every unit increase in the independent variable (X), 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 the 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

Individual Testing – Using t Test

Testing **which variable is significant**

Objective	To test the **null hypothesis** that **parameters of individual variables are equal to zero**

Null Hypothesis (H_0): $b_i = 0$

Alternate Hypothesis (H_1): $b_i \neq 0$

where $i = 1,2,\ldots,p$

Test Statistic	$t = \dfrac{Estimated\ b_i}{Standard\ Error\ of\ Estimated\ b_i}$
Decision Criteria	Reject the null hypothesis **if p-value < 0.05**

Individual Testing – Using t Test

Parameters	Coefficients	Standard Error	t statistic	p-value
Intercept	-54.2822	7.3945	-7.3409	0.0000
aptitude	0.3236	0.0678	4.7737	0.0001
tol	0.0334	0.0712	0.4684	0.6431
technical	1.0955	0.1814	6.0395	0.0000
general	0.5368	0.1584	3.3890	0.0021

p-values for aptitude, technical and general are < 0.05

p-value for test of language (tol) is > 0.05

Therefore, tol is the only insignificant variable

Measure of Goodness of Fit – R Squared

R^2 is the proportion of variation in the dependent variable which is explained by the independent variables. Note that R^2 always increases if variable is added in the model

$$R^2 = \frac{Explained\ Variation}{Total\ Variation} \quad ¿ \quad \frac{\sum_{i=1}^{n} \left(\hat{Y}i - \bar{Y} \right) 2}{\sum_{i=1}^{n} \left(Yi - \bar{Y} \right) 2}$$

The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the modelling.

$$R_a^{\ 2} = 1 - \frac{n - 1}{n - p - 1}(1 - R^2)$$

The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model.

Normally, greater than 0.7 is considered as a benchmark for accepting goodness of fit of a model.

Understanding Summary Output

```
#Model Summary
summary(jpimodel)
```

summary() generates a detailed

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -54.28225    7.39453  -7.341 5.41e-08 ***
aptitude      0.32356    0.06778   4.774 5.15e-05 ***
tol           0.03337    0.07124   0.468   0.6431
technical     1.09547    0.18138   6.039 1.65e-06 ***
general       0.53683    0.15840   3.389   0.0021 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.549 on 28 degrees of freedom
Multiple R-squared:  0.8768,    Adjusted R-squared:  0.8592
F-statistic: 49.81 on 4 and 28 DF,  p-value: 2.467e-12
```

Interpretation :

- Reject null hypothesis that no variables are significant as p-value is<0.05
- aptitude, technical, general are significant variables (p-values<0.05)

Summary of Findings

Significant variables → **Aptitude**, **Technical knowledge**, **General information**

Out of four dependent variables, **three affect job performance index positively**

R^2 → **0.88**

88% of the variation in job performance index is explained by the model & 12% is unexplained variation

Fitted Values and Residuals

- Fitted values (also called 'Predicted Values') are calculated using estimated model parameters and by substituting values of independent variables. The model now will include only the significant variables.

Estimated Model:

E(jpi)= **-54.40644** + **0.33335***aptitude + **1.11663***technical + **0.54316***general

Values of Independent Variables for First Employee	
aptitude	43.83
tol	55.92
technical	51.82
general	43.58

Fitted Values and Residuals

Values of Independent Variables for First Employee	
aptitude	43.83
technical	51.82
general	43.58

aptitude technical general

jpi= **-54.40644** + **0.33335***43.83 + **1.11663***51.82 + **0.54316***43.58

Predicted jpi= 41.73850

Residual= Observed jpi – Predicted jpi = 45.52 - 41.73850 = 3.781497

Fitted Values and Residuals

```
#Model Fitting after eliminating the insignificant variable
```

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

The insignificant variable **tol** is not included in the new model

```
#Output
```

```
Coefficients:
(Intercept)     aptitude    technical      general
   -54.4064       0.3333       1.1166       0.5432
```

Estimated values of the model parameters using the new model

* To get the fitted values and the residuals values, the model should include only the significant variables

Fitted Values and Residuals

```
#Adding Fitted Values and Residuals to the Original Dataset
perindex$pred<-fitted(jpimodel_new)

perindex$resi<-residuals(jpimodel_new)
```

fitted() and **residuals()** fetch fitted values and residuals respectively.

#Output

```
  empid   jpi aptitude   tol technical general     pred      resi
1     1 45.52    43.83 55.92     51.82   43.58 41.73850  3.781497
2     2 40.10    32.71 32.56     51.49   51.03 41.70973 -1.609731
3     3 50.61    56.64 54.84     52.29   52.47 51.36215 -0.752151
4     4 38.97    51.53 59.69     47.48   47.69 41.69149 -2.721486
5     5 41.87    51.35 51.50     47.59   45.77 40.71145  1.158549
6     6 38.71    39.60 43.63     48.34   42.06 35.61699  3.093010
```

Interpretation :

- **pred** values are calculated based on the values of the model parameters
- **resi** is the difference between the actual **jpi** values and the **pred** values.

Predictions for New Dataset

- New data set should have all the independent variables used in the model
- Column names of all common variables in the new and old datasets should be identical
- Note that missing values will be taken as 0 (which can be incorrect)

```
#Importing New Dataset
perindex_new<-read.csv("Performance Index new.csv", header=TRUE)

perindex_new$pred<-predict(jpimodel_new,perindex_new)

head(perindex_n
```

predict() returns predicted values. Fitted model is the first argument and new dataset object is the second argument. This ensures R uses parameters from the fitted model for predictions on new data.

	empid	jpi	tol	technical	general	aptitude	pred
1	34	66.35	59.20	57.18	54.98	66.74	61.55258
2	35	56.10	64.92	52.51	55.78	55.45	53.00898
3	36	48.95	63.59	57.76	52.08	51.73	55.62154
4	37	43.25	64.90	50.13	42.75	45.09	39.82060
5	38	41.20	51.50	47.89	45.77	50.85	40.87977
6	39	50.24	55.77	51.13	47.98	53.86	46.70139