Regression

Introduction | Types | Practice

Contents

- What is Regression
- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Linear Regression

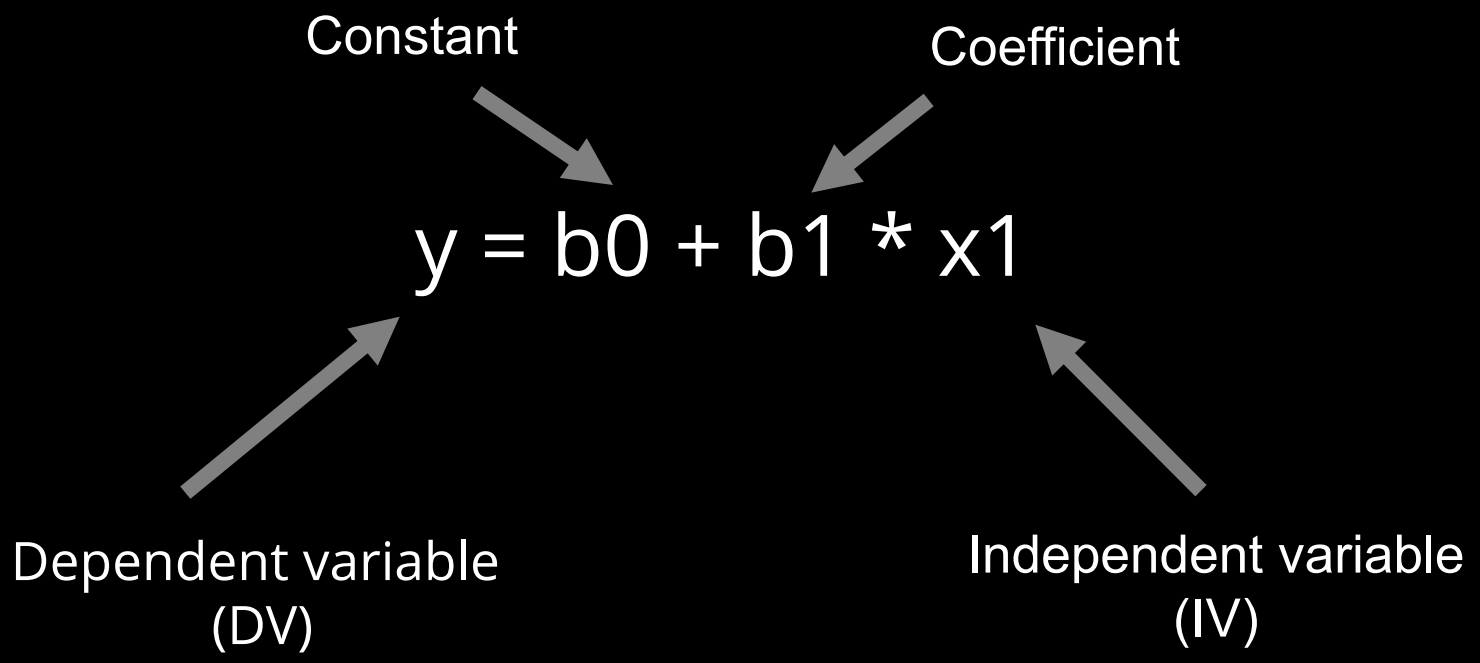
What is Regression?

- Regression is a statistical measure used in finance, investing and other disciplines that attempts to determine the strength of the relationship between one **dependent** variable (usually denoted by Y) and a series of other changing variables (known as **independent** variables).

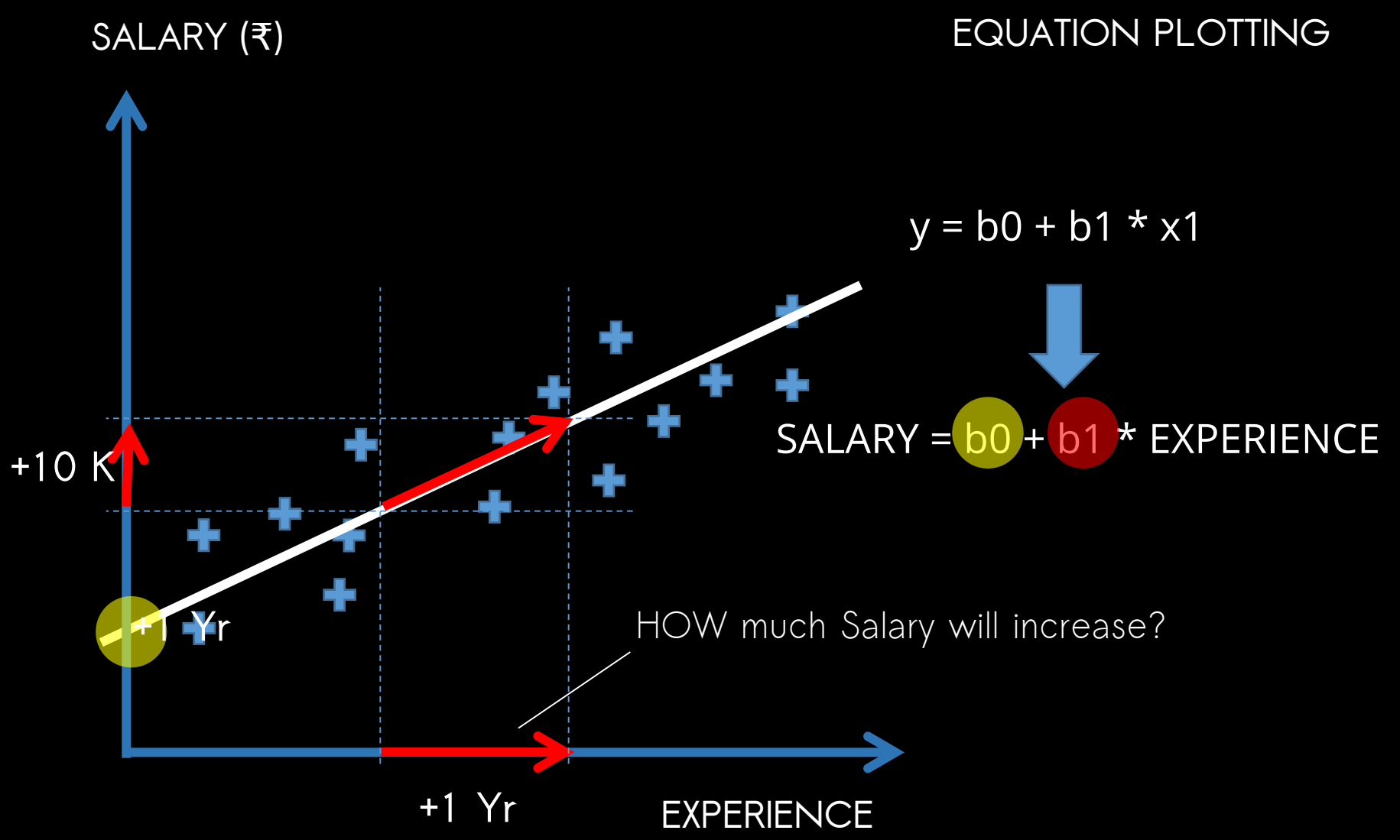
TYPES:

- Simple Linear Regression
- Multi Linear Regression
- Polynomial Linear Regression

Simple Linear Regression



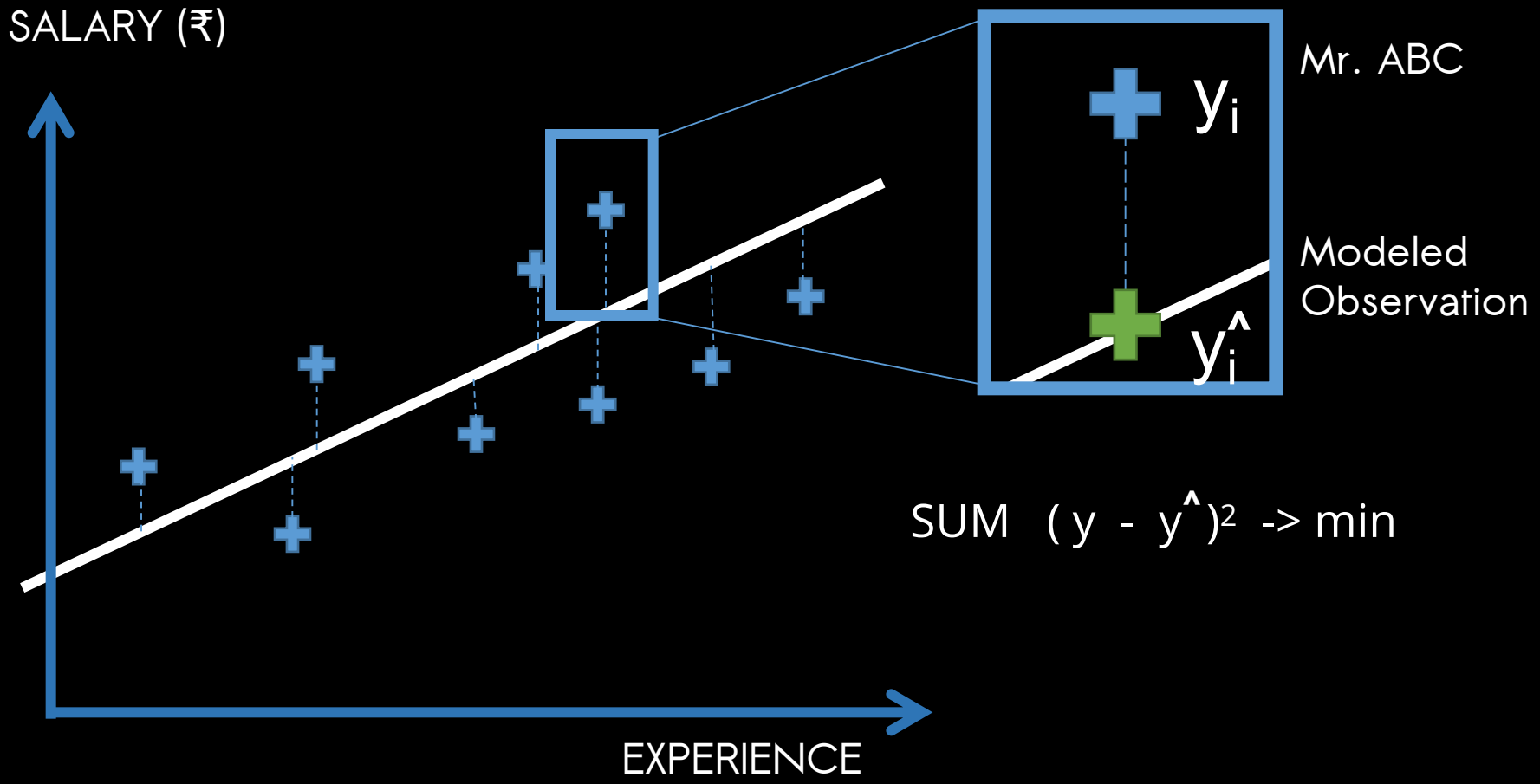
Simple Linear Regression



Simple Linear Regression

ORDINARY LEAST SQUARES

- How SLR finds **Best Fitting Line** from our Data



Simple Linear Regression

ANALYZING DATASET

IV

YearsExperienc	Salary
1.1	39343
1.3	46205
1.5	37731
2	43525
2.2	39891
2.9	56642
3	60150
3.2	54445
3.2	64445
3.7	57189
3.9	63218
4	55794
4	56957
4.1	57081
4.5	61111
4.9	67938

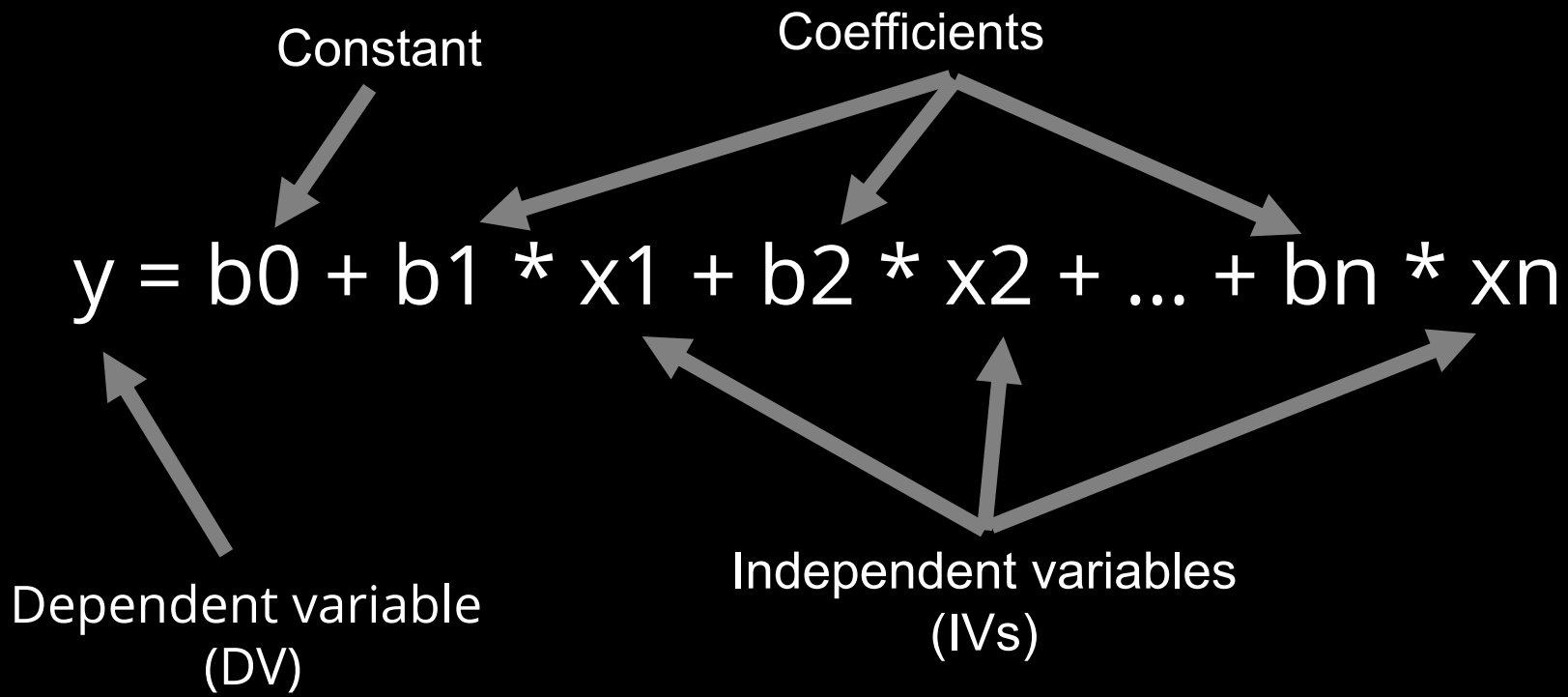
DV

Simple Linear Regression

LET's CODE!

- Prep your Data Preprocessing Template
  - **Import** Dataset
  - **No** need for ***Missing Data***
  - Splitting into **Training** & **Testing** dataset
  - Keep **Feature Scaling** but least preffered here
- Co-relate Salaries with Experience
- Later carry out prediction
- Verify the Values of prediction
- Prediction on TEST SET

Multi Linear Regression



Multi Linear Regression

ASSUMPTIONS OF LINEAR REGRESSION

- Linearity
- Homoscedasticity
- Multivariate normality
- Independence of Errors
- Lack of Multicollinearity

Multi Linear Regression

DUMMY VARIABLES

Categorical Variable

R&D Spend	Administratic	Marketin	State	Profit
165349.2	136897.8	471784.1	New York	192261.8
162597.7	151377.59	443898.5	California	191792.1
153441.51	101145.55	407934.5	Florida	191050.4
144372.41	118671.85	383199.6	New York	182902
142107.34	91391.77	366168.4	Florida	166187.9
131876.9	99814.71	362861.4	New York	156991.1
134615.46	147198.87	127716.8	California	156122.5
130298.13	145530.06	323876.7	Florida	155752.6
120542.52	148718.95	311613.3	New York	152211.8
123334.88	108679.17	304981.6	California	149760
101913.08	110594.11	229161	Florida	146122
100671.96	91790.61	249744.6	California	144259.4

Multi Linear Regression

DUMMY VARIABLES

D

D Spend	Administratic	Marketin	State
165349.2	136897.8	471784.1	New York
162597.7	151377.59	443898.5	California
153441.51	101145.55	407934.5	Florida
144372.41	118671.85	383199.6	New York
142107.34	91391.77	366168.4	Florida
131876.9	99814.71	362861.4	New York
134615.46	147198.87	127716.8	California
130298.13	145530.06	323876.7	Florida
120542.52	148718.95	311613.3	New York
123334.88	108679.17	304981.6	California
101913.08	110594.11	229161	Florida
100671.96	91790.61	249744.6	California

NEW YORK	CALIFORNIA
1	0
0	1
0	1
0	1

$$y = b0 + b1 * x1 + b2 * x2 + b3 * x3 + b4 * D1$$

Multi Linear Regression

	Administratic	Marketin	State
349.2	136897.8	471784.1	New York
597.7	151377.59	443898.5	California
41.51	101145.55	407934.5	Florida
72.41	118671.85	38	
07.34	91391.77	3	
376.9	99814.71	3	
15.46	147198.8		
98.13	145530.(		
42.52	148718.95	311613.3	New York
34.88	108679.17	304981.6	California
13.08	110594.11	229161	Florida
71.96	91790.61	249744.6	California

D2 = 1 - D1

Multi Linear Colinearity

DUMMY VARIABLE TRAP

D

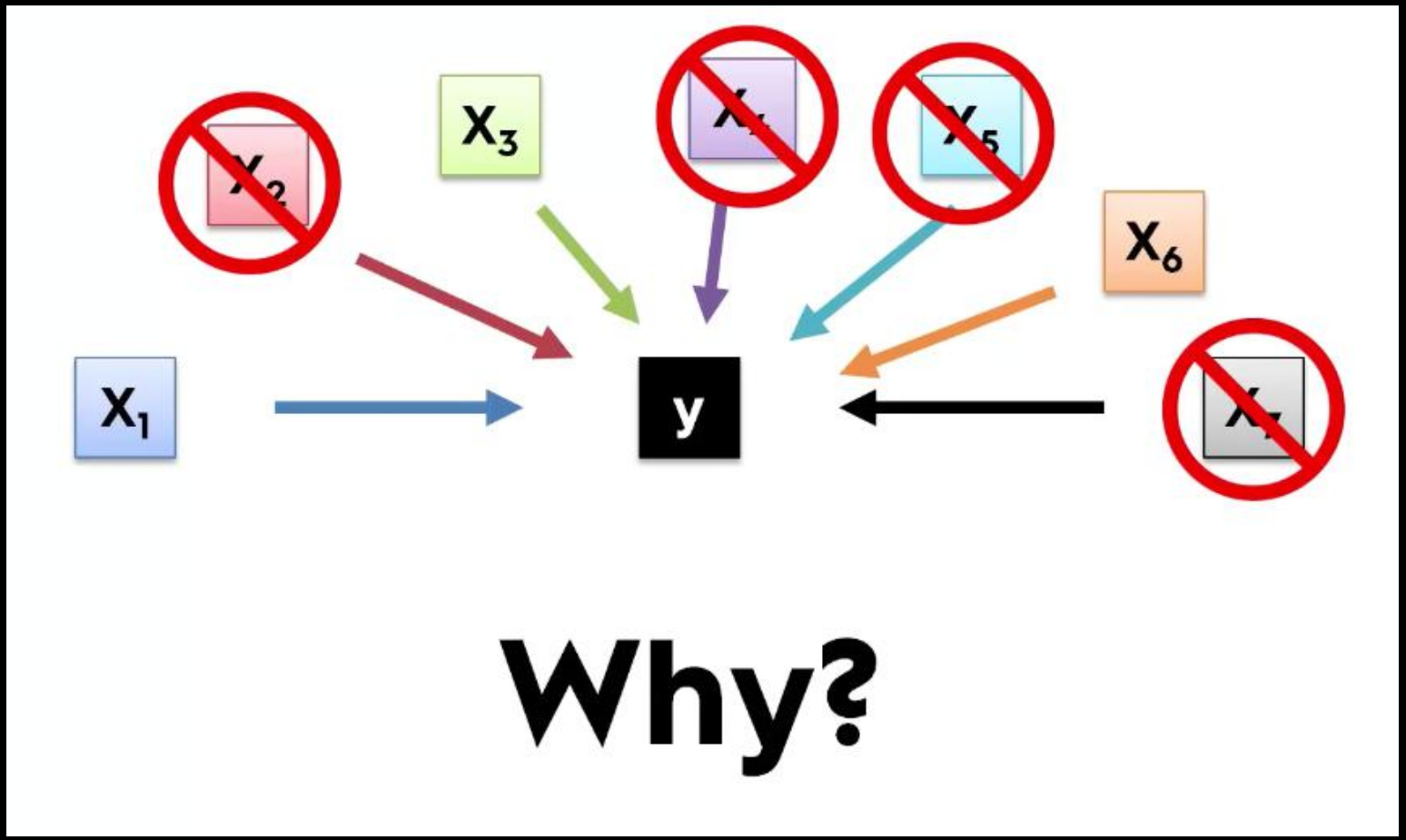
NEW YORK	CALIFORNIA
	0
	1
	1
0	1

$$y = b0 + b1 * x1 + b2 * x2 + b3 * x3 + b4 * D1 + b5 * D2$$

Always OMIT one Dummy Variable

Building A Model

STEP BY STEP

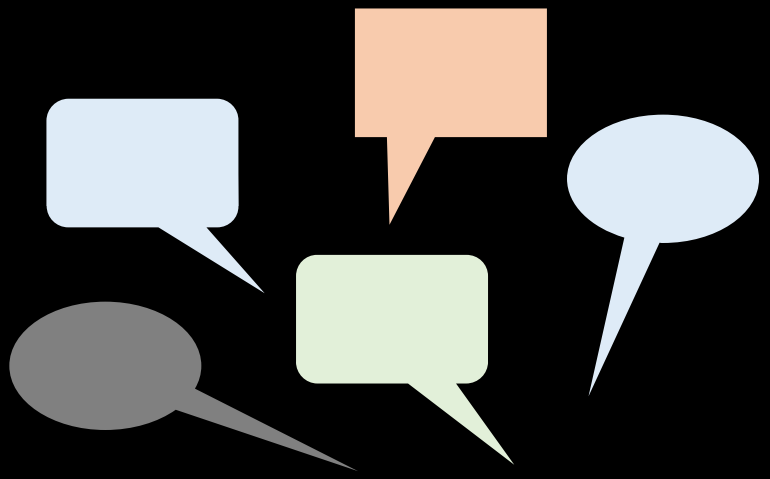


Building A Model

2 REASONS

- GARBAGE IN ⟶ GARBAGE OUT

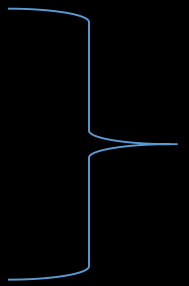
- TOO MUCH EXPLANATION LATER



Building A Model

METHODS OF BUILDING A MODEL

- All - in
- Backward Elimination
- Forward Selection
- Bidirectional Elimination
- Score Comparison



Stepwise regression

Building A Model

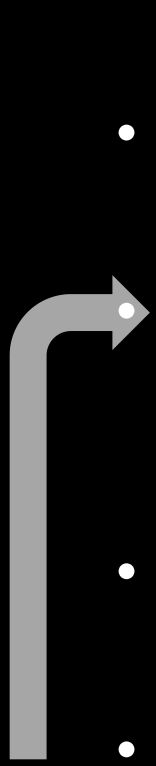
METHODS OF BUILDING A MODEL

- ALL - IN
  - Throw in every variable
  - Prior Knowledge
  - Known Values
  - Preparing Backward elimination

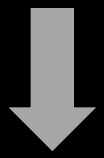
Building A Model

BACKWARD ELIMINATION MODEL

- Step 1
  - Select significance level to stay in model (0.05)
- Step 2
  - Fit in full model with all possible predictors
- Step 3
  - Consider the predictor with highest P value
  - If P > SL, go to Step 4, otherwise **go to FIN**
- Step 4
  - Remove the Predictor
- Step 5
  - Fit the model w/o this variable*



MODEL BUILT



Building A Model

FORWARD SELECTION MODEL

1. Select a SL to enter the model
2. Fit all possible simple regression $y \sim x_n$ Select one with lowest P value
3. Keep this variable and fit all possible models with one extra predictor added to the ones you already have
4. Consider the predictor with the lowest P value. go to Step 3, else go to FIN

KEEP PREV MODEL

Building A Model

BIDIRECTIONAL ELIMINATION

1. Select a SL to ENTER and to STAY in the model
   - e.x. SLENTER = 0.05. SLSTAY = 0.05
2. Perform the next of Forward Selection
3. Perform all step of Backward Selection

Building A Model

ALL PROBABLE MODELS

1. Select a criterion of goodness of fit
2. Construct all possible regression model
   - $2^N - 1$
3. Select the one with Best Criterion

BACKWARD ELIMINATION is the Fastest Model, Hence majorly used

LET's CODE!

Congrats for the DAY!
You may rest your Machines ;)

@mitu_skillologies

/mITuSkillologies

@mitu_group

For Queries & Suggestions
CONTACT:

Tejas Rawal 97 65 838775
email: **tejasprawal**@gmail.com