Machine Learning concepts

Siddharth Gehlot

## ***Overview***

Flow –

Data Pre-processing 🡪 Modelling 🡪 Evaluation

Data Pre-processing –

* Import the data
* Clean the data
* Split into training and test sets

Modeling –

* Build the model
* Train the model
* Make predictions

Evaluation

* Calculate performance metrics
* Make a verdict

## ***Data Pre-Processing***

1. Preparing dataset for ML Models –
   1. Import Libraries
   2. Import dataset
   3. Taking care of missing data
   4. Encoding categorical data
   5. Encoding the independent variables
   6. Encoding the dependent variables
   7. Splitting the dataset into training and test sets

Import Datasets –

Dependent Variable – Field to be predicted

Independent Variable – Fields used to make prediction

Taking care of missing data –

1. Impute missing data with mean of values in the field.

Encoding categorical data -

*Convert categories to numbers*

* One hot encoder – independent variable
* Label encoder – dependent variable

Splitting the dataset into training and test set –

Feature scaling is applied after splitting the dataset because test set doesn’t need to be tested on feature scaled data. Only training data needs to be featured scaled for training. Test set is feature scaled separately on same scaler used for training data.

Feature Scaling –

Allows to put all features on same scale. Its done because there are some dominating features causing other features to be ignored.

Feature scaling techniques –

Standardization 🡪 (x- mean(x))/ standard deviation(x)

Put all values of feature between -3 and 3. Standardization should work with all datasets

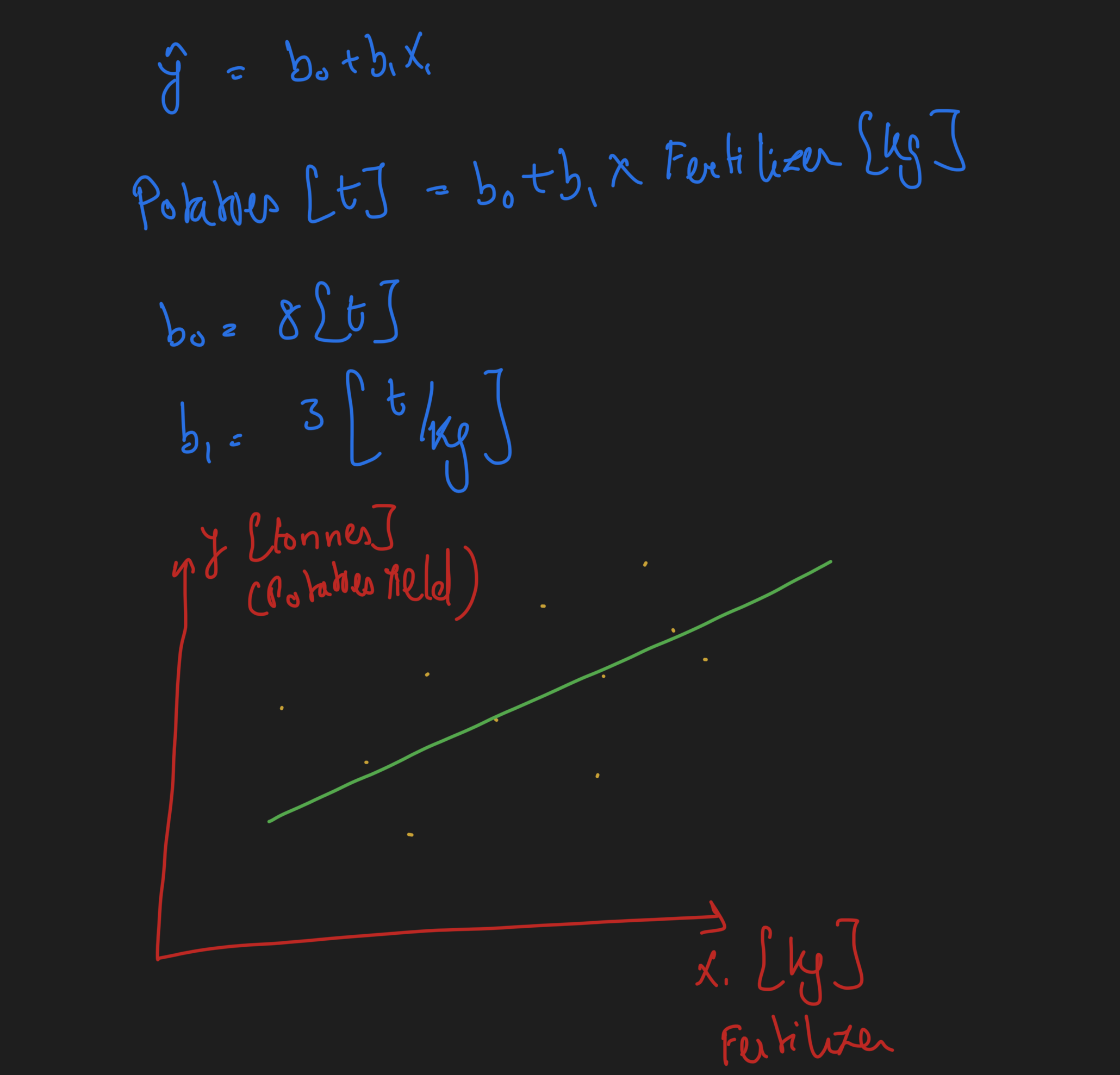
Normalization 🡪 (x – min(x))/(max(x) – min(x))

All values of features will be between 0 and 1. Normalization is recommended when dataset is normal.

## ***Machine Learning Models***

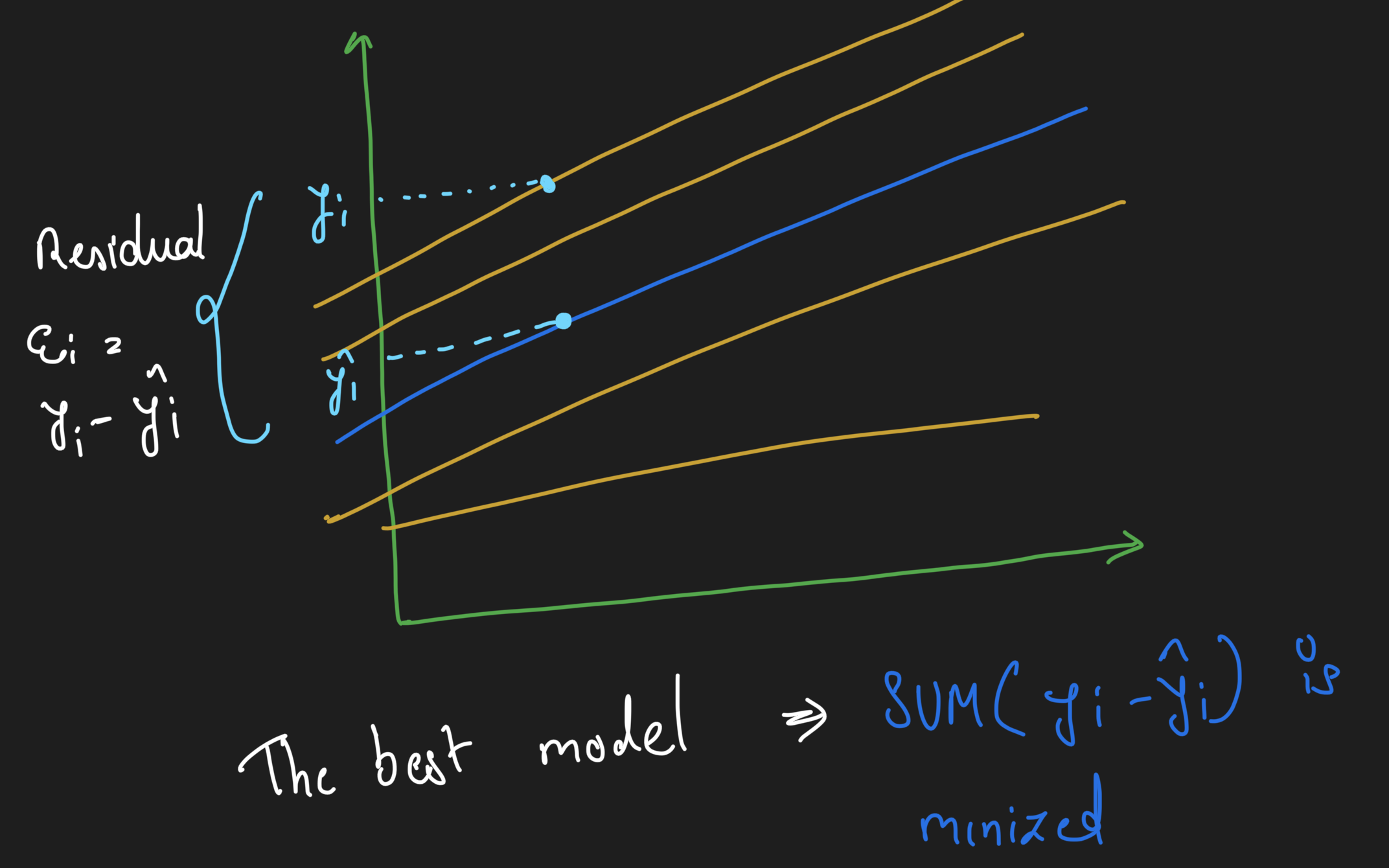
### Simple Linear Regression

Y = b0 + b1X1



Ordinary Least Squares method –

To identify which linear regression model is best.

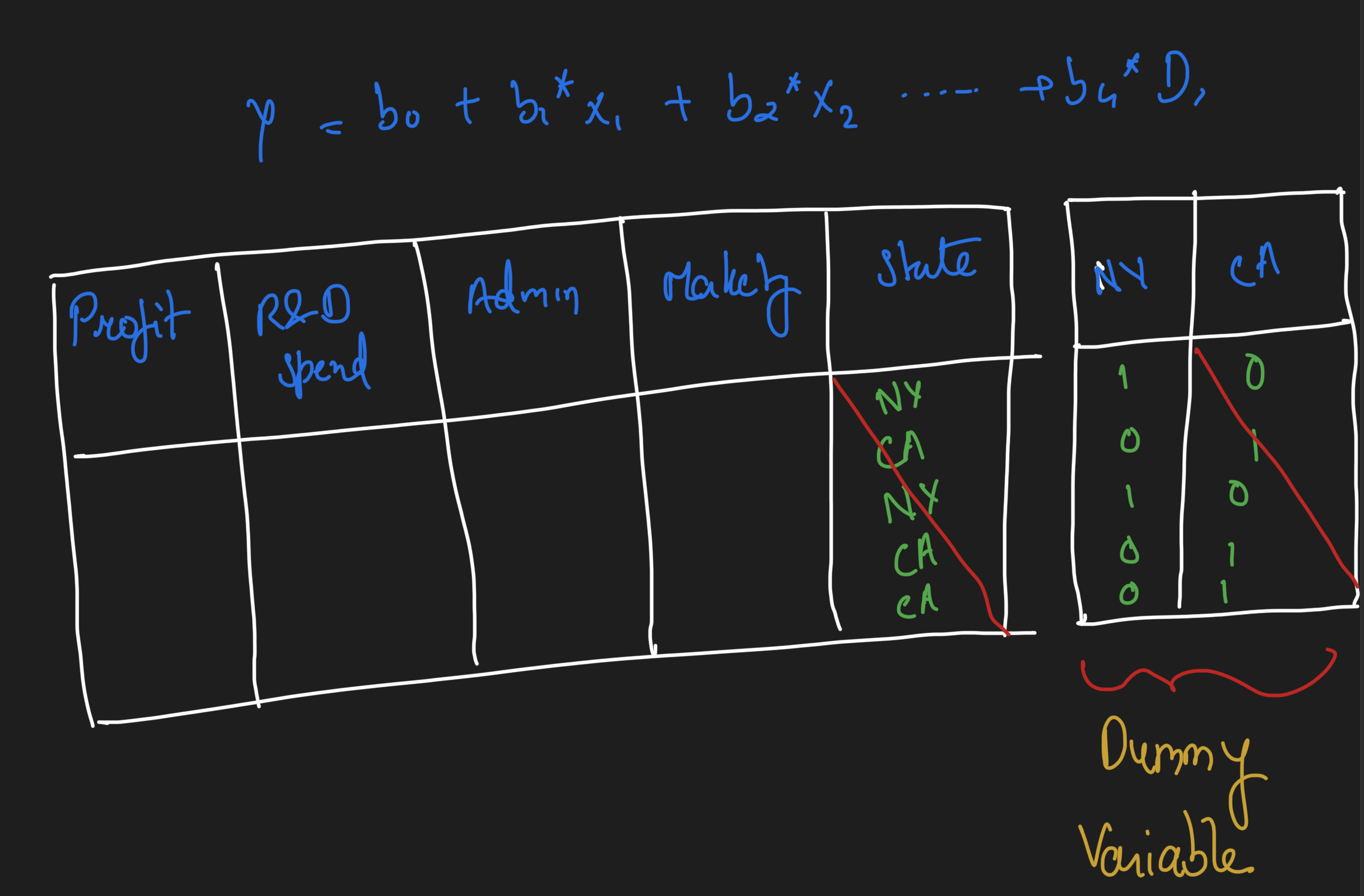


### Multiple Linear Regression

Assumptions –

1. Linearity – There is linear relationship between x and y
2. Homoscedasticity – Equal variance
3. Multivariate Normality – Normality error distribution
4. Independence – Of observations includes ‘no autocorrelation’
5. Lack of multicollinearity – Predictors are not correlated with each other
6. The outlier check – extra steps to remove outlier

Dummy Variables –



Only one dummy variable should be used in this case.

Dummy variable trap -

What will happen if second dummy variable is included?

D2 = 1-D1

So its basically duplicating a variable. Effect of one dummy variable is impacted by another dummy variable. This is called Multicollinearity. Because of this model can’t distinguish between the effect of D-1 from effects of D2. Therefore it won’t work properly.

Always omit one dummy variable i.e. if there are 10 dummy variable consider only 9.

P-value -

Statistical Significance

Example –

H0 – This is a fair coin

H1 – This is not a fair coin

We will start with assumption that H0 is correct hypothesis and we will check if assumption is correct

Toss1 - Tail 0.5

Toss2 - Tail 0.25

Toss3 - Tail 0.125

Toss4 - Tail 0.06

Toss5 - Tail 0.03

………

To reject the null hypothesis, a p-value needs to be less than the chosen significance level (usually 0.05)

### Stepwise Regression

Backward Elimination –

Remove fields based on p-value starting with keeping all fields and then get rid of it with each iteration of calculating p-value

1. Select significance level to stay in the model
2. Fit the full model with all possible predictors
3. Consider the predictors with highest p-value
4. Remove the predictor
5. Fit model with this variable
6. Iterate through step 3 to 5 until the model is ready

All -in –

* Use all variables. Can do with this approach when you have prior knowledge
* In case there are no other option
* Preparing for Backward eliminition

Forward Elimination –

1. Select a significance level to enter the model
2. Fit all simple regression model. Select one with the lowest p-value
3. Keep the variable and fit all possible models with one extra predictor added to the one you already have
4. Consider the predictor with the lowest p-value
5. Move back to step 3 i.e. add variable with each iteration till model is ready

Bidirectional Elimination –

1. Select significance level to enter and to stay in the model
2. Perform the next step of forward selection
3. Perform all steps of backward elimination
4. Back to step 2 until no new variable can enter and no old variable can exit

Score Comparison –

1. All possible models
2. Select the best model