Semester: VIII Subject: AIFB Academic Year: 2024-25

REGRESSION ON YEARLY SPARSITY:

Yearly sparsity refers to gaps or missing values in yearly data. It occurs when data is collected irregularly or some years have incomplete or missing observations:

Problem setup:

Given a dataset with yearly observations, some years may have missing values for key metrics.

Example:

Year.	Metric (Y)
2015	50
2016	—
2017	55
2018	60
2019	—
2020	70
2021	75

Our goal is to fit a regression model to estimate missing values and predict future trends.

To solve this we will use linear regression approach.

The linear regression equation:

$$Y = mX + b$$

Y = Dependent Variable, X = Independent Variable (year)

Semester: VIII Subject: AIFB AcademicYear: 2024-25

m = Slope , b = Intercept.

Step 1: Consider only the known data from the table:

Step 2: Fit a linear model:

Calculate:

$$m = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$b = \bar{Y} - m\bar{X}$$

$$\bar{X} = \frac{2015+2017+2018+2020+2021}{5} = 2018.2$$

$$\bar{Y} = \frac{50+55+60+70+75}{5} = 62.$$

X	Y	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})^2$	
2015	50	-3.2	-12	10.24	38.4
2017	55	-1.2	-7	1.44	8.4
2018	60	-0.2	-2	0.04	0.4
2020	70	1.8	8	3.24	14.4
2021	75	2.8	13	7.84	36.4

$\sum (X_i - \bar{X}) = 0$ $\sum (X_i - \bar{X})^2 = 22.8$ $\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 98$

$$m = \frac{98}{22.8} = 4.29 \simeq 4$$

$m = 4.29$

$b = 62 - 4.29(2018.2)$

$\boxed{b = -8596.07}$

Step3: Estimate the missing years.

The regression equation is $\boxed{Y = 4.2X - 8596.07}$

2016 : $Y = 4.29(2016) - 8596.07 = 52.57$ $\boxed{=56}$.

2019 : $Y = 4.29(2019) - 8596.07 = 65.44$ $\boxed{=65}$

Step4: Predict the future values: (2025)

$Y = 4.29(2025) - 8596.07 = 91.18$ $\boxed{=91}$.

By using Linear Regression the missing values and the future forecast value is found.

Linear Interpolation (Best for small gaps)

If two years have known values, missing years between them can be estimated using linear interpolation:

$$Y_{missing} = Y_{prev} + \frac{(Y_{next} - Y_{prev})}{(Year_{next} - Year_{prev})} \times (Year_{missing} - Year_{prev})$$

Example:

Year	Revenue($M)
2015	50
2016	—
2017	55

Semester: VIII Subject: AIFB Academic Year: 2024-25

$$Y_{2016} = 50 + \frac{(55-50)}{(2017-2015)} \times (2016-2015)$$

$$= 50 + \frac{5}{2} \times 1 = \boxed{52.5}$$

So, 2016 revenue is estimated as $52.5M.

This is how the yearly sparsity is estimated using regression.

REGRESSION ON MONTHLY DATA ON SPARSITY:

Monthly sparsity refers to gaps or missing values in monthly data. It occurs when data is collected irregularly or some years have incomplete or missing observations.

The following steps has to be followed:

Step 1: Handle sparsity (estimate missing values)

Step 2: Fit a regression model with trend and seasonality.

Step 3: Predict future sales for the next 6 months.

Let's consider the below example for regression on monthly data on sparsity:

Problem statement:-

A company tracks its monthly sales for 3 years (36 months). Some months have missing sales data.

Consider the below data set: