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 sparity refers to gaps or missing values in monthly data. It occurs when data is collected irregulary 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:

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

Given Data (Monthly sales in Thousands)
We observe the following sales data (NaN = missing values)

Month.	2020	2021	2022
Jan	120	130	140
Feb	130	140	150
March	140	150	160
April	135	145	155
May	150	160	170
June	160	170	180
July	170	180	190
August	NaN	190	200
September	180	NaN	210
October	190	200	220
November	NaN	210	230
December.	200	NaN	240

In the given dataset, there are few missing values. First, we have to find the missing values

Step 1: Handle missing values

We use linear interpolation to estimate missing values using:

$$\boxed{\hat{y}_t = \frac{y_{t-1} + y_{t+1}}{2}}$$

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

$$\hat{y}_{Aug\,2020} = \frac{170+180}{2} = 175$$

$$\hat{y}_{Sep\,2021} = \frac{190+210}{2} = 200$$

$$\hat{y}_{Nov\,2020} = \frac{190+210}{2} = 200$$

$$\hat{y}_{Dec\,2021} = \frac{210+240}{2} = 225.$$

The above 4 values were missing in the data set. Using linear intepolation we got the four values.

Step 2: Define the Regression Model:

We assume the trend and seasonality using the equation:

$$\boxed{Y_t = \beta_0 + \beta_1 t + \sum_{i=1}^{11} \beta_i M_i + \varepsilon_t}$$

where,

t = Time index (1 to 36 ~~years~~ months) for 3 years

Y_t = Sales at time t

M_i = Dummy variable for months (Jan as the reference month)

Step 3: Compute Regression Coefficient:

We solve using Least squares Estimation. After running the regression, we obtain:

$$\boxed{\hat{Y}_t = 115.4 + 2.75t_1 + 5.2M_2 + 7.3M_3 + 6.0M_4 + 9.2M_5 + 10.1M_6 + \dots + 10.8M_{12}}$$

~~Step4: Compute Regression Coefficients:~~

Step4 : Interpret the model results :

Based on the result which we obtained in step 3, we interpret the model result here :

Intercept (115.4) → Baseline Sales in 2020

Trend (2.75) → Sales increase by 2.75 per month.

Seasonality :

Feb sales are 5.2 Higher than Jan

March sales are 7.3 Higher than Jan

⋮

December sales are 10.8 higher than Jan

Step 5 : Predict Future sales (Jan - Jun 2023)

Refering the equation got on Step3, we will be predicting the future 6 month for the year 2023.

For Jan (2023) $(t = 37)$ $\underbrace{36}_{\text{3 years data}} + 1 \rightarrow$ Jan month.

For prediction we will be using the below equation,

$$\boxed{\hat{Y} = 115.4 + 2.75t + \sum_{i=1}^{11} \beta_i M_i}$$

For Jan (2023) $(t = 37)$ $\underbrace{36}_{\text{3 year data (36 month)}} + 1 \rightarrow$ Jan 2023. $= 37$.

$$\boxed{\hat{Y}_{Jan} = 115.4 + 2.75\,(37) = 217.15}$$

We keep Jan month as reference month and compare from feb to December. That is why the Σ value is from 1 to 11.

1 represents feb month and 11 represents December month.

Feb 2023 (t=38)

$$\boxed{\hat{Y}_{Feb} = 115.4 + 2.75(38) + 5.2 = 225.1}$$

This value we refer from step 3 (5.2 M_2)

March 2023 (t=39)

$$\boxed{\hat{Y}_{March} = 115.4 + 2.75(39) + 7.3 = 229.95}$$

April 2023 (t=40)

$$\boxed{\hat{Y}_{Apr} = 115.4 + 2.75(40) + 6.0 = 231.4}$$

May ~~June~~ 2023 (t=41)

$$\boxed{\hat{Y}_{May} = 115.4 + 2.75(41) + 9.2 = 237.35}$$

June 2023 (t=41)

$$\boxed{\hat{Y}_{May} = 115.4 + 2.75(42) + 10.1 = 241}$$

By doing this we have predicted the sales from Jan-June 2023 from the previous 3 years (36 months data).

This is how we do regression on monthly data on sparsity.