REGRESSION ANALYSIS IN STATISTICS

Presented

By

Debanjan Basak

Alumni

of

Sitalkuchi College

Contact: +91-8016126478

Email:

debanjanbasak111@gmail.com

debanjan.geo@nbu.ac.in

OUTLINE

- ✔ ***Introduction***
- ✔ ***Concepts of Regression***
- ✔ ***Assumptions***
- ✔ ***Simple Bivariate Analysis:***
- ✔ ***Fitting of Regression Trend Line by Least Square Method.***
- ✔ ***Residual Mapping.***

INTRODUCTION

What is regression?

- A reliable method of identifying which variables have impact on a topic of interest.
- We can determine which factors matter most
- Which factors to be ignored
- How these factors influence each other

IMPORTANT TERMS IN REGRESSION

- **Dependent Variable:** The main factor that you're trying to understand or predict.
- **Independent Variable:** These are the factors that you hypothesize, show an impact on the dependent variable
- **Coefficients:** They represent the change in the response variable for a unit change in predictor variable
- **Error term:** The difference between the actual and predicted values

GENERAL EQUATION FOR REGRESSION

- The simple linear regression equation is

$$y = \beta_0 + \beta_1 x_1$$

- The Multiple linear regression equation is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

- Coefficient is computed by subtracting mean from the variable and dividing by the standard deviation
- Error sum of squares is calculated by $\varepsilon_i = y_i - y_i\hat{}$

LEAST SQUARES METHOD

- ✔ The Least Squares method is a form of mathematical regression analysis used to determine the line of best fit for a set of data,
- ✔ It aims to create a straight line that minimizes the sum of the squares of the errors
- ✔ It is mostly used for data fitting
- ✔ Our aim is to minimise the sum of squares

Fitting of Regression Trend Line by Least Square Method

Average Monthly Temperature (^{0}C)	38	29	25	21	34	41	43	28
Electricity Bill (Rs.)	85	75	60	55	90	100	120	70

On the basis of the data

i) Draw a Regression Line following the least square method.

ii) Predict the probable ^{0}C in average monthly temperature of rupees in electricity bill is 80.

iii) Predict the Probable rupees in electricity bill of ^{0}C, when average monthly temperature is 40^0 C.

iv) Interpret the relationship.

Average Monthly Temperature (^{0}C) (x)	Electricity Bill (Rs.) (y)	x.y	X^2
38	85	3230	1444
29	75	2175	841
25	60	1500	625
21	55	1155	441
34	90	3060	1156
41	100	4100	1681
43	120	5160	1849
28	70	1960	784
Σ x = 259	Σ y = 655	Σ x.y =22340	$\Sigma X^2 = 8821$

It is known as the method of least square. The general equation of a straight line is given by **y = a+bx , where a = intercept, b= slope.** The equation to solve for a and b are

$$\Sigma y = na + b \Sigma x \text{ …….. (i)}$$

$$\Sigma xy = a \Sigma x + b\Sigma x^2 \text{ …….. (ii)}$$

By substituting the values from the table we

get, 655= 8a + 259b …… (i)

22340 = 259a +8821b …….. (ii)

Multiplying equation (i) by 32.38 we get,

21208.90 = 259a +8386.42b …… (iii)
22340= 259a +8321 b …… (ii)

-1131.10 = -434.58b

Or 434.58b = 1131.10

Or, **b = 2.60**

Substituting the value of b in **eq (i)** we get,

655 = 8a + 259b

Or, 655 = 8a +259* 2.60
Or, 655 = 8a + 665.63

Or, - 8a = 665.40 – 655

Or, -8a = 10.63

Or, a = 10.63/-8

Or, **a = -1.33**

y= a+ bx

y= -1.33 + 2.60x

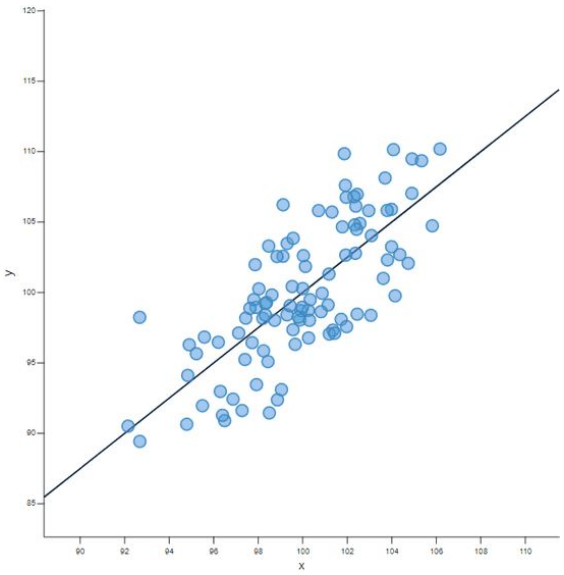
Therefore, Regression Equation:

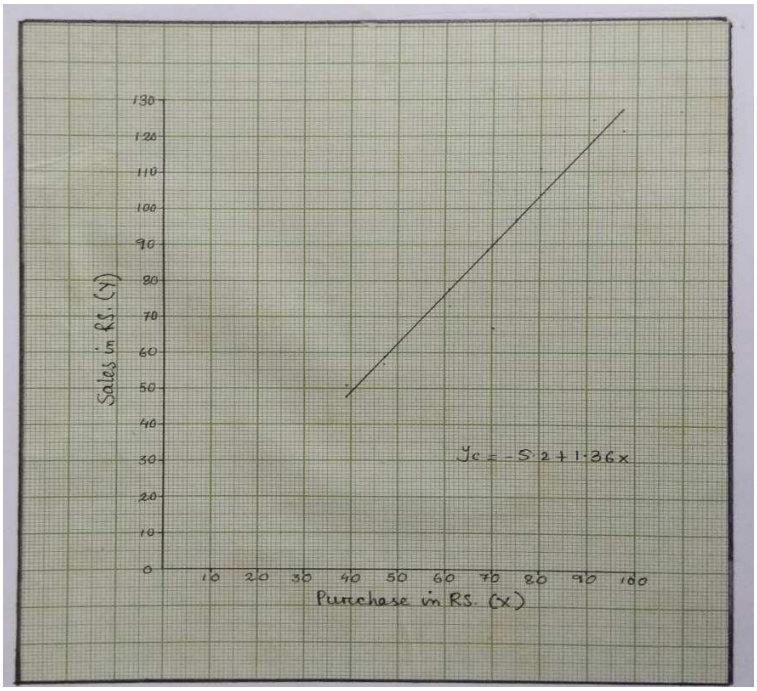
y= -1.33 + 2.60x

x	**y_c**
38	47.12
29	36.83
25	
21	
34	
41	
43	
28	

The Slope of the Regression line ‘b’, which is also known as regression co-efficient, shows the estimated average change in y with respect to x. Thus from the above regression equation, we can say that the relationship between **Average Monthly Temperature** and **Electricity Bill (Rs.)** is such that a ^{0}C increase in temperature by and large cause an increase of 2.60 Rs. in electricity bill.

SCATTER DIAGRAM AND REGRESSION LINE OF Y ON X





MEAN SQUARE ERROR

- ✔ The mean squared error tells you how close a regression line is to a set of points.
- ✔ It does this by taking the distances from the points to the regression line
- ✔ The squaring is necessary to remove any negative signs.
- ✔ Formula for Mean square error is E = Sum(Y – Y)^2

Residual Mapping

Sl. No.	Name of the District	Percentage of Urban Population (y)	Percentage of total worker (x)
1	Patna	22.54	3.16
2	Gaya	7.63	1.40
3	Sahabad	8.22	2.63
4	Serdan	4.20	2.24
5	Champaran	5.21	1.47
6	Muzaffarpur	5.25	1.33
7	Darbhanga	4.43	1.45
8	Monghyr	11.87	3.16
9	Bhagalpur	10.61	2.00
10	Saharsa	4.53	0.58
11	Purnea	6.34	1.25
12	Santal Pargana	5.76	1.55
13	Palamau	4.96	1.20
14	Hazaribagh	12.87	1.93
15	Ranchi	13.67	4.13
16	Dhanbad	43.51	8.95
17	Sighbhum	26.24	11.84

(i) Find out the regression equation of y on x using the values given above (1971).

(ii) What is the average rate of change y in with respect to x

(iii) Further calculate residual and plot the residual over map.

Name of the District	Percentage of Urban Population (y)	Percentage of total worker (x)	xy	x^2	Y_c
Patna	22.54	3.16	71.23	9.99	
Gaya	7.63	1.40	10.68	1.96	
Sahabad	8.22	2.63	21.62	6.92	
Serdan	4.20	2.24	9.41	5.02	
Champaran	5.21	1.47	7.66	2.16	
Muzaffarpur	5.25	1.33	6.98	1.77	
Darbhanga	4.43	1.45	6.42	2.10	
Monghyr	11.87	3.16	37.51	9.99	
Bhagalpur	10.61	2.00	21.22	4	
Saharsa	4.53	0.58	2.63	0.34	
Purnea	6.34	1.25	7.93	1.56	
Santal Pargana	5.76	1.55	8.93	2.40	
Palamau	4.96	1.20	5.95	1.44	
Hazaribagh	12.87	1.93	24.84	3.72	
Ranchi	13.67	4.13	56.46	17.06	
Dhanbad	43.51	8.95	389.41	80.10	

Sighbhum	26.24	11.84	310.68	140.19	
	Σ y = 197.66	Σ x = 50.27	Σ x.y =999.56	ΣX^2 = 290.72	

Σ y = na + b Σx …….. (i)
Σxy = a Σx + bΣx2 …….. (ii)

197.66= 17a + 50.27 b …… (i)

22340 = 50.27a +290.72 b …….. (ii)

Multiplying equation (i) by 2.96 we get,

585.07 = 50.27 a +148.80 b …… (iii)

999.24 = 50.27 a +290.72 b ……. (ii)

-414.17 -141.92 b
=

Or 141.92 b = 414.17

Or, b = 2.92

Substituting the value of b in eq (i) we get,

197.66 = 17a + 50.27 b

Or, 197.66 = 17 a +50.27* 2.92

Or, a = 2.99

(ii) Same as previous one

Name of the District	Percentage of Urban Population (y)	Y hat	Residual (Y-Y hat)	(Y-Y hat)2	Categories
Patna	22.54	12.22	10.32	106.50	High Positive
Gaya	7.63	7.08	0.55	0.30	High Positive
Sahabad	8.22	10.67	-2.45	6.00	Medium Negative
Serdan	4.20	9.53	-5.33	28.41	
Champaran	5.21	7.28	-2.07	4.28	
Muzaffarpur	5.25	6.87	-1.62	2.62	
Darbhanga	4.43	7.22	-2.79	7.78	
Monghyr	11.87	12.22	-0.35	0.12	
Bhagalpur	10.61	8.83	1.78	3.17	
Saharsa	4.53	4.68	-0.15	0.02	
Purnea	6.34	6.64	-0.3	0.09	
Santal Pargana	5.76	7.52	-1.76	3.10	
Palamau	4.96	6.49	-1.8	3.24	
Hazaribagh	12.87	8.63	4.25	18.06	
Ranchi	13.67	15.05	-1.38	1.90	
Dhanbad	43.51	29.12	14.39	207.07	
Sighbhum	26.24	37.56	-11.32	128.14	
			Σ= 0.03	Σ = 520.8	

Mean of Residual = 0.03/17= 0.00
SD of Estimate = √520.80/17-2 = 5.89

Categories

0 to 5.89 = Medium Positive

5.89 to 11.78 = High Positive

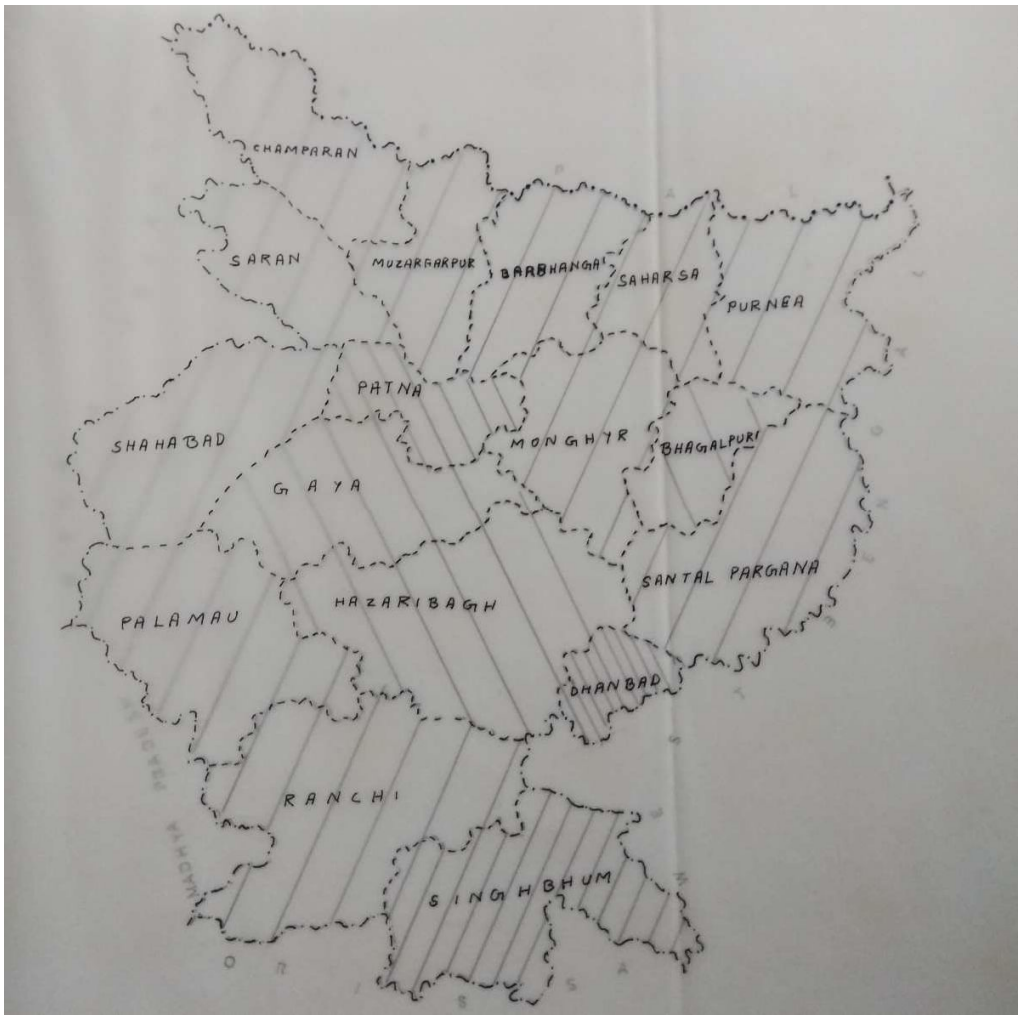
> 11.78 = Very high Positive

0 to - 5.89 = Medium Negative

-5.89 to -11.78 = High Negative

> -11.78 = Very high Negative

Residuals from Regression of Urbanisation on Percentage of employment in Manufacturing Industries in Bihar (1971)



THANK YOU

For you’re attention

Please don’t hesitate to ask, I will be always ready to help you