Regresi Linier (Linear Regression)

1

PERTEMUAN KE-9: ANALISIS TERAPAN

AGUS S. MUNTOHAR

1

Pengertian Regresi

2

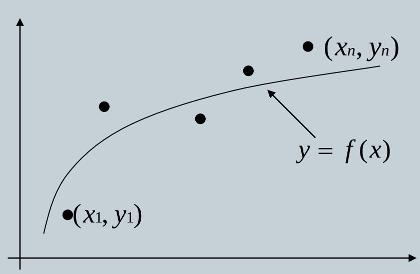
- Diberikan *n* data $(x_1,y_1), (x_2,y_2),\ldots, (x_n,y_n)$ dengan model persamaan $y = f(x)$ untuk setiap data (*best fit*). Best fit dilakukan dengan meminimalkan nilai kuadrat dari kesalahan perkiraan (S_r).

Kesalahan perkiraan pada setiap titik data:

$$\varepsilon_i = y_i - f(x_i)$$

Jumlah kuadrat kesalahan perkiraan :

$$S_r = \sum_{i=1}^{n} (y_i - f(x_i))^2$$



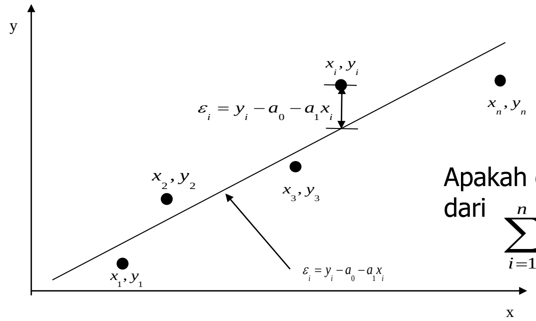
Gambar 1 Dasar model regresi

2

Regresi Linier : Kriteria#1

3

Diberikan *n* titik data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ best fit $y = a_0 + a_1 x$ untuk data tersebut



Gambar 2. Regersi Linier dari data y vs. x dengan nilai sisa (residuals) terhadap setiap data x_i.

Apakah dengan meminimalkan nilai dari $\sum_{i=1}^{n} \varepsilon_i$ dapat dijadikan kriteria, dimana

$$\varepsilon_i = y_i - (a_0 + a_1 x_i)$$

3

Contoh untuk Kriteria#1

4

Example: Given the data points (2,4), (3,6), (2,6) and (3,8), best fit the data to a straight line using Criterion#1

Table. Data Points

x	y
2.0	4.0
3.0	6.0
2.0	6.0
3.0	8.0

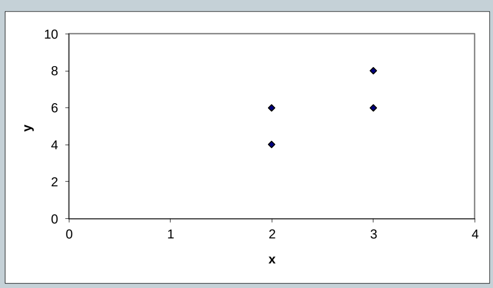


Figure. Data points for y vs. x data.

4

Regresi Linier : Kriteria#1

5

Using $y=4x-4$ as the regression curve

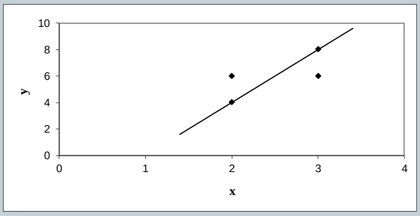


Figure. Regression curve for y=4x-4, y vs. x data

Table. Residuals at each point for regression model y = 4x – 4.

x	y	$y_{predicted}$	$\varepsilon = y - y_{predicted}$
2.0	4.0	4.0	0.0
3.0	6.0	8.0	-2.0
2.0	6.0	4.0	2.0
3.0	8.0	8.0	0.0
			$\sum_{i=1}^{4} \varepsilon_i = 0$

5

Regresi Linier : Kriteria#1

6

Using $y=6$ as a regression curve

Table. Residuals at each point for y=6

x	y	$y_{predicted}$	$\varepsilon = y - y_{predicted}$
2.0	4.0	6.0	-2.0
3.0	6.0	6.0	0.0
2.0	6.0	6.0	0.0
3.0	8.0	6.0	2.0
			$\sum_{i=1}^{4} \varepsilon_i = 0$

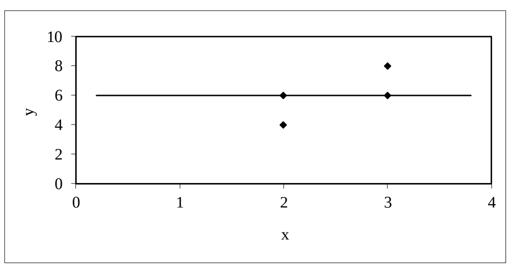


Figure. Regression curve for y=6, y vs. x data

6

Linear Regression – Criterion #1

7

$$\sum_{i=1}^{4} \varepsilon_i = 0$$ for both regression models of y=4x-4 and y=6.

The sum of the residuals is as small as possible, that is zero, but the regression model is not unique.

Hence the above criterion of minimizing the sum of the residuals is a bad criterion.

7

Linear Regression-Criterion#2

8

Will minimizing $\sum_{i=1}^{n} |\varepsilon_i|$ work any better?

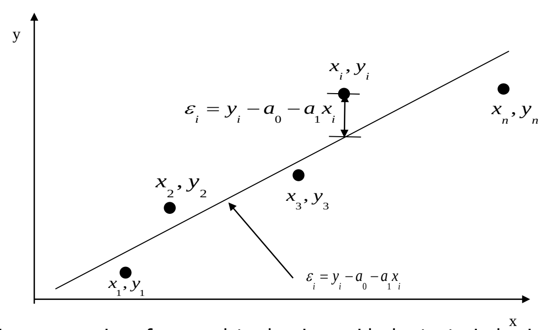


Figure. Linear regression of y vs. x data showing residuals at a typical point, x_i.

8

9

Linear Regression-Criteria 2

Using *y=4x-4* as the regression curve

Table. The absolute residuals employing the y=4x-4 regression model

x	y	$y_{predicted}$	$\lvert\varepsilon\rvert = \lvert y - y_{predicted}\rvert$
2.0	4.0	4.0	0.0
3.0	6.0	8.0	2.0
2.0	6.0	4.0	2.0
3.0	8.0	8.0	0.0
			$\sum_{i=1}^{4}\lvert\varepsilon_i\rvert = 4$

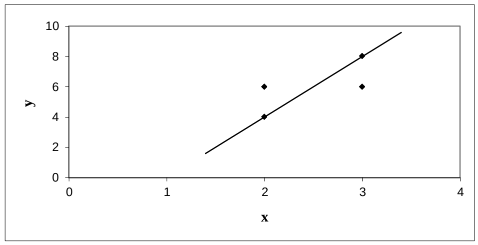


Figure. Regression curve for y=4x-4, y vs. x data

9

10

Linear Regression-Criteria#2

Using *y=6* as a regression curve

Table. Absolute residuals employing the y=6 model

x	y	$y_{predicted}$	$\lvert\varepsilon\rvert = \lvert y - y_{predicted}\rvert$
2.0	4.0	6.0	2.0
3.0	6.0	6.0	0.0
2.0	6.0	6.0	0.0
3.0	8.0	6.0	2.0
			$\sum_{i=1}^{4}\lvert\varepsilon_i\rvert = 4$

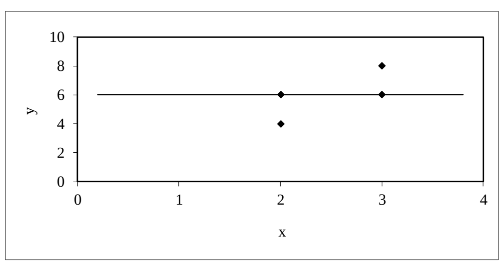


Figure. Regression curve for y=6, y vs. x data

10

Linear Regression-Criterion#2

11

$\sum_{i=1}^{4} |\varepsilon_i| = 4$ for both regression models of y=4x-4 and y=6.

The sum of the errors has been made as small as possible, that is 4, but the regression model is not unique.

Hence the above criterion of minimizing the sum of the absolute value of the residuals is also a bad criterion.

Can you find a regression line for which $\sum_{i=1}^{4} |\varepsilon_i| < 4$ and has unique regression coefficients?

11

Least Squares Criterion

12

The least squares criterion minimizes the sum of the square of the residuals in the model, and also produces a unique line.

$$S_r = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right)^2$$

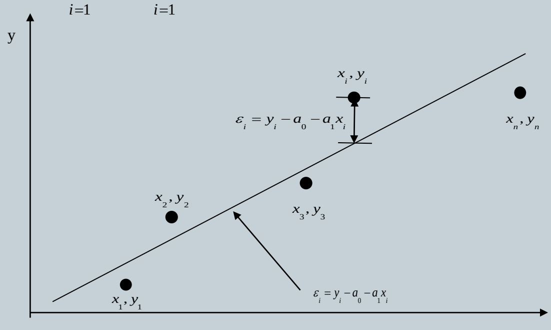


Figure. Linear regression of y vs. x data showing residuals at a typical point, x_i.

12

Finding Constants of Linear Model

13

Minimize the sum of the square of the residuals: $S_r = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$

To find a_0 and a_1 we minimize S_r with respect to a_1 and a_0.

$$\frac{\partial S_r}{\partial a_0} = -2\sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)(-1) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)(-x_i) = 0$$

giving

$$\sum_{i=1}^{n} a_0 + \sum_{i=1}^{n} a_1 x_i = \sum_{i=1}^{n} y_i$$

$$(a_0 = \bar{y} - a_1 \bar{x})$$

$$\sum_{i=1}^{n} a_0 x_i + \sum_{i=1}^{n} a_1 x_i^2 = \sum_{i=1}^{n} y_i x_i$$

13

Finding Constants of Linear Model

14

Solving for a_0 and a_1 directly yields,

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

and

$$a_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad (a_0 = \bar{y} - a_1 \bar{x})$$

14

Example 1

15

The torque, T needed to turn the torsion spring of a mousetrap through an angle, is given below. Find the constants for the model given by

$$T = k_1 + k_2\theta$$

Table: Torque vs Angle for a torsional spring

Angle, θ	Torque, T
Radians	*N-m*
0.698132	0.188224
0.959931	0.209138
1.134464	0.230052
1.570796	0.250965
1.919862	0.313707

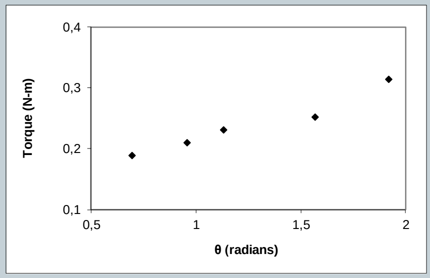


Figure. Data points for Angle vs. Torque data

15

Example 1 cont.

16

The following table shows the summations needed for the calculations of the constants in the regression model.

Table. Tabulation of data for calculation of important summations

	θ	T	θ^2	$T\theta$
	Radians	*N-m*	*Radians²*	*N-m-Radians*
	0.698132	0.188224	0.487388	0.131405
	0.959931	0.209138	0.921468	0.200758
	1.134464	0.230052	1.2870	0.260986
	1.570796	0.250965	2.4674	0.394215
	1.919862	0.313707	3.6859	0.602274
$\sum_{i=1}^{5} =$	6.2831	1.1921	8.8491	1.5896

Using equations described for a_0 and with $n = 5$

$$k_2 = \frac{n\sum_{i=1}^{5}\theta_i T_i - \sum_{i=1}^{5}\theta_i \sum_{i=1}^{5} T_i}{n\sum_{i=1}^{5}\theta_i^2 - \left(\sum_{i=1}^{5}\theta_i\right)^2}$$

$$= \frac{5(1.5896) - (6.2831)(1.1921)}{5(8.8491) - (6.2831)^2}$$

$$= 9.6091\times 10^{-2} \text{ N-m/rad}$$

16

Example 1 cont.

17

Use the average torque and average angle to calculate k_1

$$\bar{T} = \frac{\sum_{i=1}^{5} T_i}{n} \qquad \bar{\theta} = \frac{\sum_{i=1}^{5} \theta_i}{n}$$

$$= \frac{1.1921}{5} \qquad = \frac{6.2831}{5}$$

$$= 2.3842 \times 10^{-1} \qquad = 1.2566$$

Using,

$$k_1 = \bar{T} - k_2 \bar{\theta}$$

$$= 2.3842 \times 10^{-1} - (9.6091 \times 10^{-2})(1.2566)$$

$$= 1.1767 \times 10^{-1} \text{ } N\text{-}m$$

17

Example 1 Results

18

Using linear regression, a trend line is found from the data

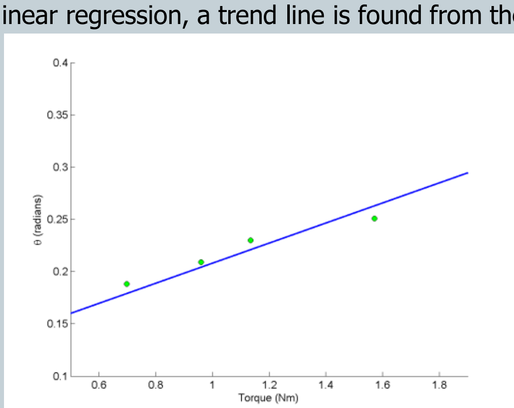


Figure. Linear regression of Torque versus Angle data

Can you find the energy in the spring if it is twisted from 0 to 180 degrees?

18

Example 2

19

To find the longitudinal modulus of composite, the following data is collected. Find the longitudinal modulus, E using the regression model $\sigma = E\varepsilon$ and the sum of the square of the residuals.

Table. Stress vs. Strain data

Strain	Stress
(%)	*(MPa)*
0	0
0.183	306
0.36	612
0.5324	917
0.702	1223
0.867	1529
1.0244	1835
1.1774	2140
1.329	2446
1.479	2752
1.5	2767
1.56	2896

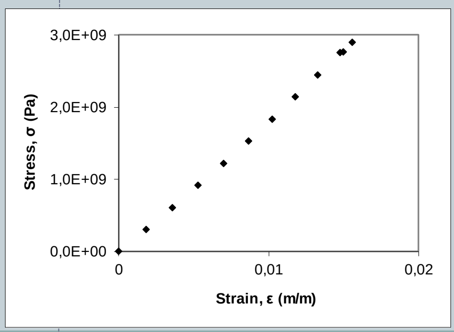


Figure. Data points for Stress vs. Strain data

19

Example 2 cont.

20

Residual at each point is given by

$$\gamma_i = \sigma_i - E\varepsilon_i$$

The sum of the square of the residuals then is

$$S_r = \sum_{i=1}^{n} \gamma_i^2$$

$$= \sum_{i=1}^{n} (\sigma_i - E\varepsilon_i)^2$$

Differentiate with respect to E

$$\frac{\partial S_r}{\partial E} = \sum_{i=1}^{n} 2(\sigma_i - E\varepsilon_i)(-\varepsilon_i) = 0$$

Therefore $$E = \frac{\sum_{i=1}^{n} \sigma_i \varepsilon_i}{\sum_{i=1}^{n} \varepsilon_i^2}$$

20

Example 2 cont.

21

Table. Summation data for regression model

i	ε	σ	ε^2	εσ
1	0.0000	0.0000	0.0000	0.0000
2	1.8300×10^{-3}	3.0600×10^{8}	3.3489×10^{-6}	5.5998×10^{5}
3	3.6000×10^{-3}	6.1200×10^{8}	1.2960×10^{-5}	2.2032×10^{6}
4	5.3240×10^{-3}	9.1700×10^{8}	2.8345×10^{-5}	4.8821×10^{6}
5	7.0200×10^{-3}	1.2230×10^{9}	4.9280×10^{-5}	8.5855×10^{6}
6	8.6700×10^{-3}	1.5290×10^{9}	7.5169×10^{-5}	1.3256×10^{7}
7	1.0244×10^{-2}	1.8350×10^{9}	1.0494×10^{-4}	1.8798×10^{7}
8	1.1774×10^{-2}	2.1400×10^{9}	1.3863×10^{-4}	2.5196×10^{7}
9	1.3290×10^{-2}	2.4460×10^{9}	1.7662×10^{-4}	3.2507×10^{7}
10	1.4790×10^{-2}	2.7520×10^{9}	2.1874×10^{-4}	4.0702×10^{7}
11	1.5000×10^{-2}	2.7670×10^{9}	2.2500×10^{-4}	4.1505×10^{7}
12	1.5600×10^{-2}	2.8960×10^{9}	2.4336×10^{-4}	4.5178×10^{7}
$\sum_{i=1}^{12}$			1.2764×10^{-3}	2.3337×10^{8}

With

$$\sum_{i=1}^{12}\varepsilon_i^2 = 1.2764\times10^{-3}$$

and

$$\sum_{i=1}^{12}\sigma_i\varepsilon_i = 2.3337\times10^{8}$$

Using

$$E = \frac{\sum_{i=1}^{12}\sigma_i\varepsilon_i}{\sum_{i=1}^{12}\varepsilon_i^{2}}$$

$$= \frac{2.3337\times10^{8}}{1.2764\times10^{-3}}$$

$$= 182.84\ GPa$$

21

Example 2 Results

22

The equation $\sigma = 182.84\varepsilon$ describes the data.

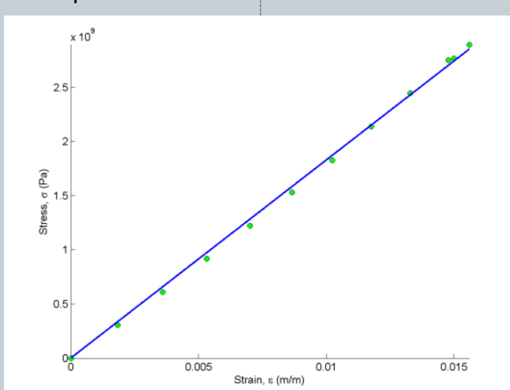


Figure. Linear regression for Stress vs. Strain data

22