DATA SCIENCE
and
BUSINESS INTELLIGENT

Author: Egi Safitri

Meeting 10

Regresi

Analisis Regresi

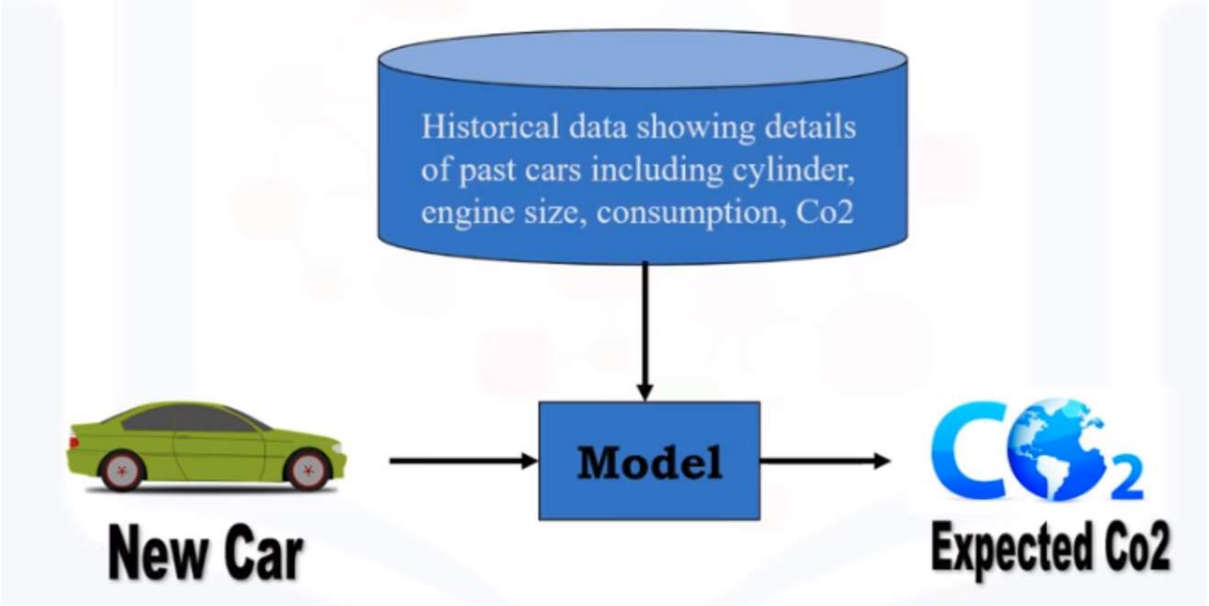
x : variabel bebas

y : variabel tak bebas

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Regresi adalah proses
Memprediksi nilai kontinu

Model Regresi



Analisis Regresi

Regresi Sederhana:

- Regresi sederhana linier
- Regresi sederhana non-linier
- Contoh: memprediksi co2emission vs EngineSize dari semua mobil.

Regresi Variabel Jamak/Berganda:

- Regresi variabel jamak linier
- Regresi variabel jamak non-linier
- Contoh: memprediksi co2emission vs EngineSize dan Cylinders dari semua mobil.

Analisis Regresi

- Prakiraan penjualan produk
- Analisis kepuasan
- Estimasi harga
- Pendapatan pekerjaan
- dst.

Regresi Linier Sederhana

1. Pendahuluan

- Analisis regresi digunakan untuk mempelajari dan mengukur hubungan statistik yang terjadi antara dua atau lebih varibel. Dalam regresi sederhana dikaji dua variabel, sedangkan dalam regresi majemuk dikaji lebih dari dua variabel.
- Dalam analisis regresi suatu persamaan regresi hendak ditentukan dan digunakan untuk menggambarkan pola atau fungsi hubungan yang terdapat antar variabel.
- Variabel yang akan diestimasi nilainya disebut variabel terikat (*dependent variable atau response variable*) dan biasanya diplot pada sumbu tegak (sumbu-y). Sedangkan variabel bebas (*independent variable atau explanatory variable*) adalah variabel yang diasumsikan memberikan pengaruh terhadap variasi variabel terikat dan biasanya diplot pada sumbu datar (sumbu-x).

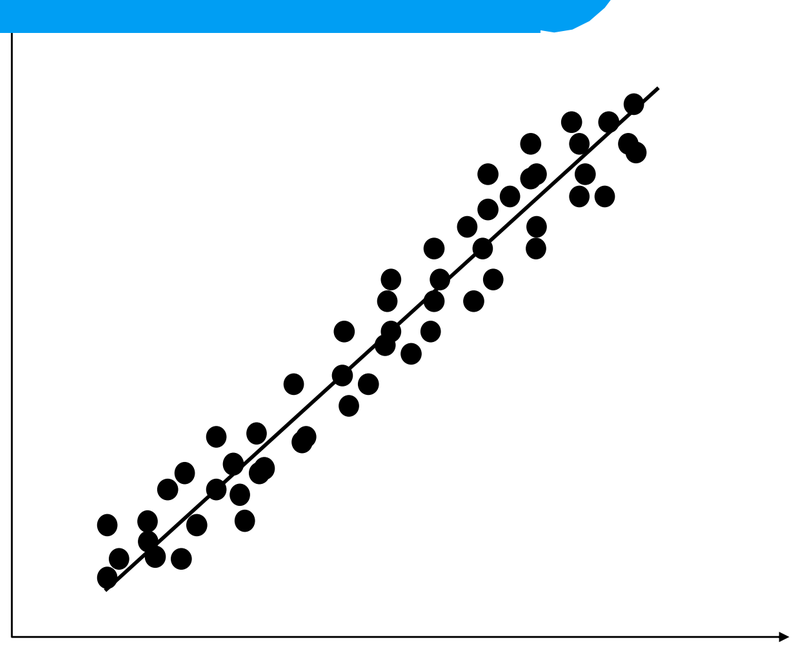
Analisis Regresi

- Analisis korelasi bertujuan untuk mengukur "seberapa kuat" atau "derajat kedekatan" suatu relasi yang terjadi antar variabel.
- Analisis regresi ingin mengetahui pola relasi dalam bentuk persamaan regresi,
- Analisis korelasi ingin mengetahui kekuatan hubungan tersebut dalam koefisien korelasinya. Dengan demikian biasanya analisis regresi dan korelasi sering dilakukan bersama-sama.

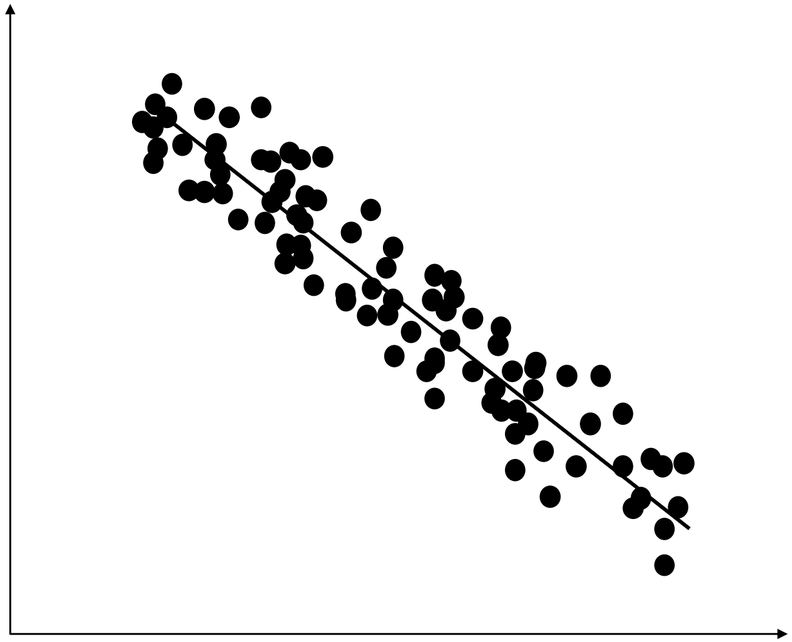
Analisis Regresi

- Langkah pertama dalam menganalisis relasi antar variabel adalah dengan membuat diagram pencar (*scatter diagram*) yang menggambarkan titik-titik plot dari data yang diperoleh. Diagram pencar ini berguna untuk
  - membantu dalam melihat apakah ada relasi yang berguna antar variabel,
  - membantu dalam menentukan jenis persamaan yang akan digunakan untuk menentukan hubungan tersebut.

Scatter Plot

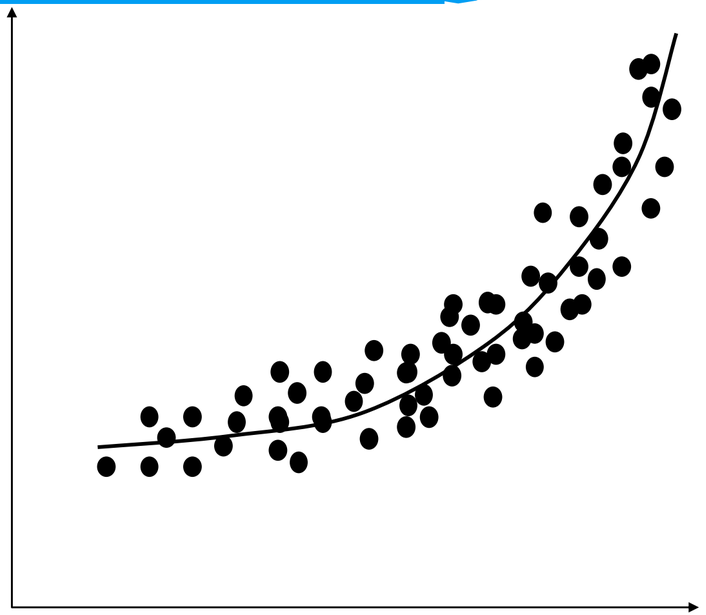


Linier positif

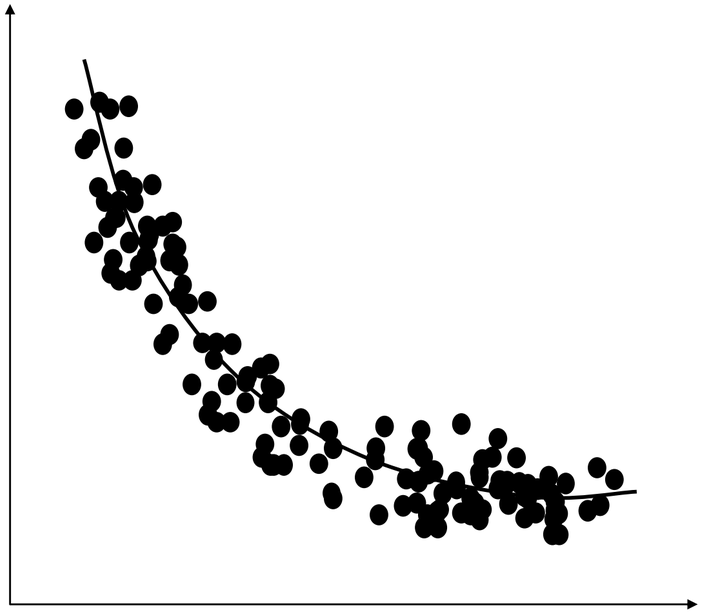


Linier negatif

Scatter Plot



Curvelinier positif



Curvelinier negatif

Analisis Regresi

Asumsi yang digunakan dalam regresi linear adalah sebagai berikut:

a. $E(\in_i) = 0$

b. $E(\in_i^2) = \sigma^2$

c. $E(\in_i \in_j) = \text{cov}(\in_i, \in_j) = 0$

d. X_i konstan

Untuk memperkirakan A dan B dipergunakan metode kuadrat kesalahan terkecil, dimana

Model sebenarnya : $Y = A + BX + \in$

Model perkiraan : $Y = a + bX + e$

a, b, dan e adalah penduga untuk A, B, dan $\in$

$Y_i = a + bX_i + e_i$ atau $e_i = Y_i - (a + bX_i)$ dan $\sum_i e_i^2 = \sum (Y_i - (a + bX_i))^2$.

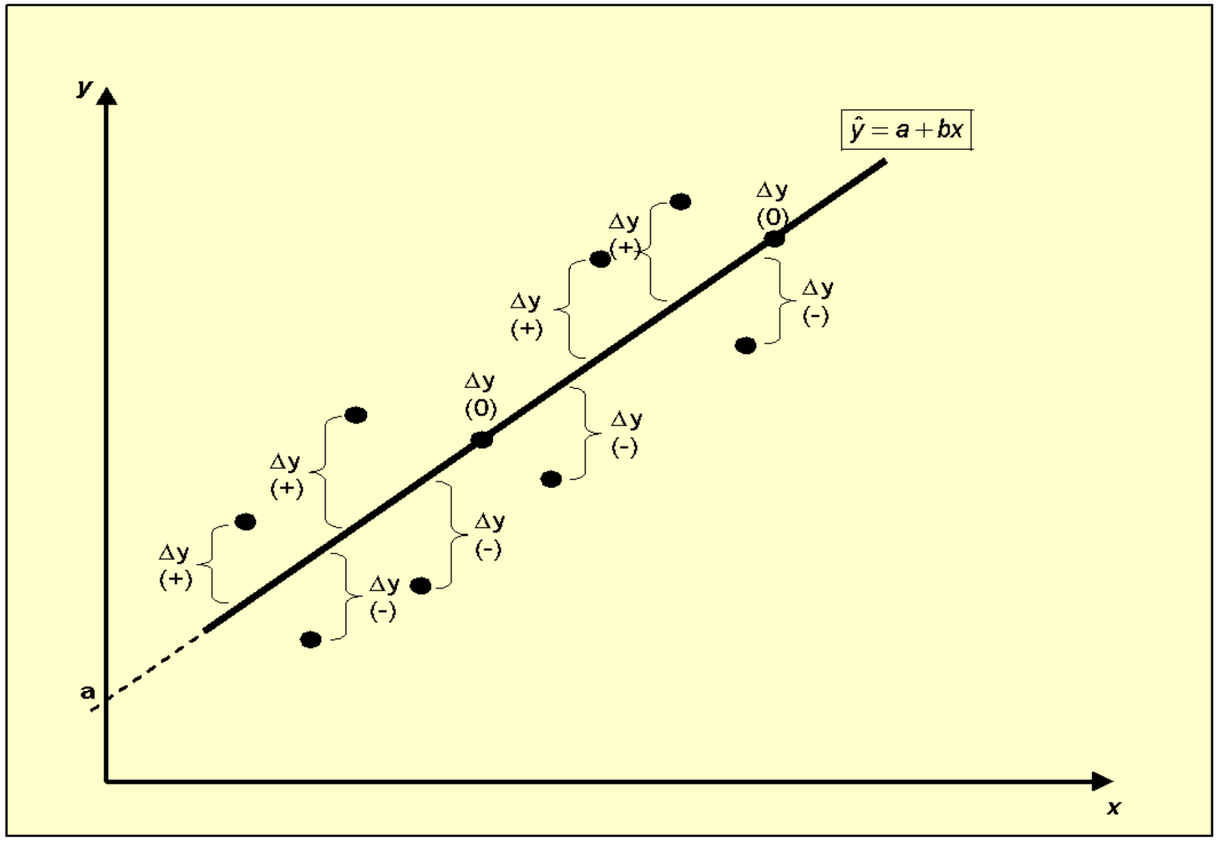
Analisis Regresi

penurunan parsial terhadap a dan b yang sederhana diperoleh

$$a = \overline{Y} - b\overline{X} = \frac{\sum_i Y_i \sum_i X_i^2 - \sum_i X_i \sum_i X_i Y_i}{n\sum_i X_i^2 - \left(\sum_i X_i\right)^2}$$ dan

$$b = \frac{n\sum_i X_i Y_i - \sum_i X_i \sum_i Y_i}{n\sum_i X_i^2 - \left(\sum_i X_i\right)^2}$$

2. Analisis Regresi Linear



Gambar 2 Garis regresi linier pada diagram pencar

Regresi Linier Untuk Memprediksi Nilai Kontinu

x : variabel bebas

y : variabel tak bebas

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Nilai kontinyu / numerik

Topologi Regresi Linier

Regresi Linier Sederhana:

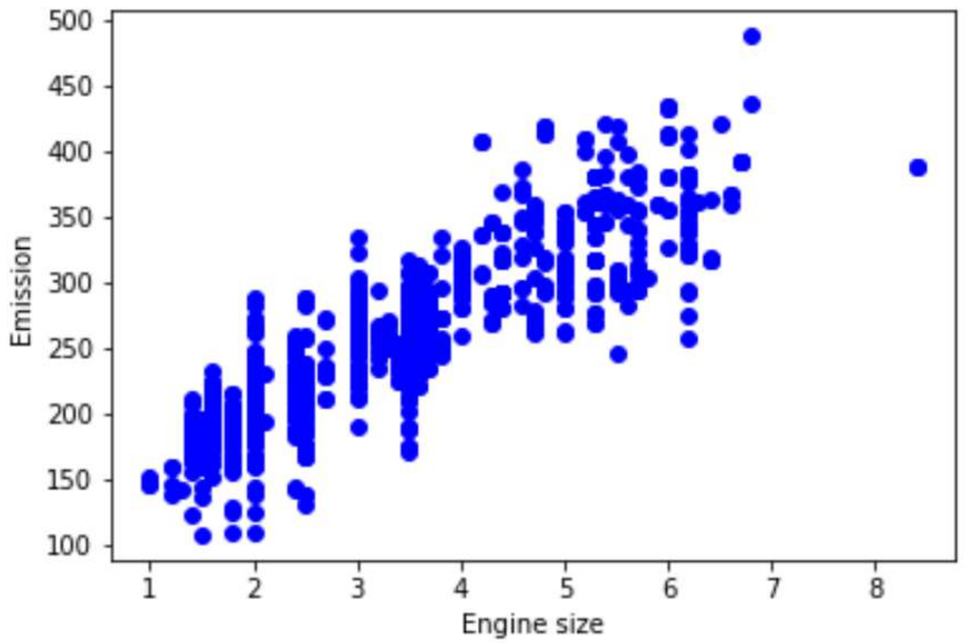
- Memprediksi co2emission vs EngineSize dari semua mobil
  - variabel bebas (x): EngineSize
  - variabel tak bebas (y): co2emission

Regresi Linier Variabel Jamak:

- Memprediksi co2emission vs EngineSize dan Cylinders dari semua mobil
  - variabel bebas (x): EngineSize, Cylinders, dst.
  - variabel tak bebas (y): co2emission

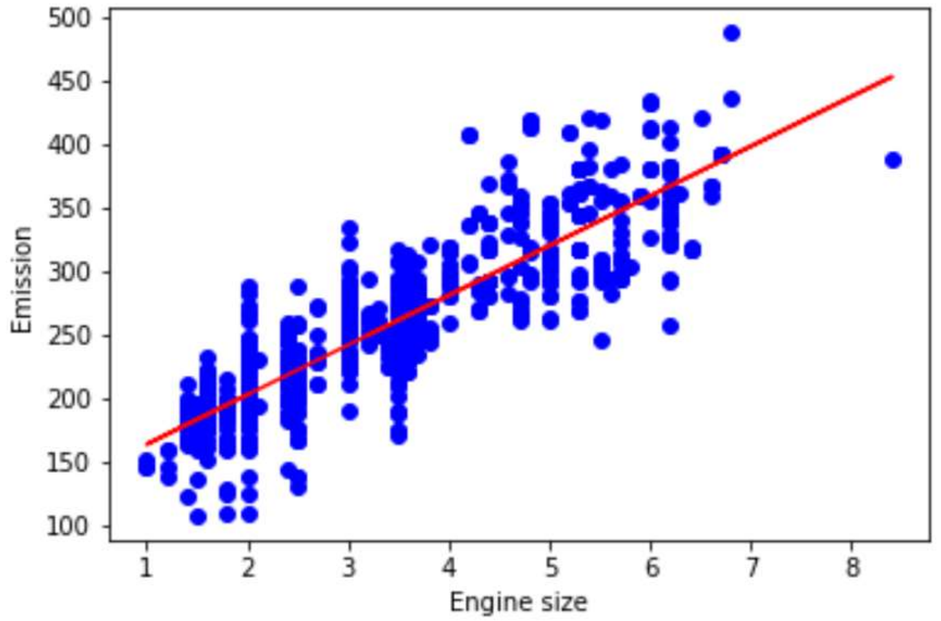
Diagram Pencar Regresi Linier

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?



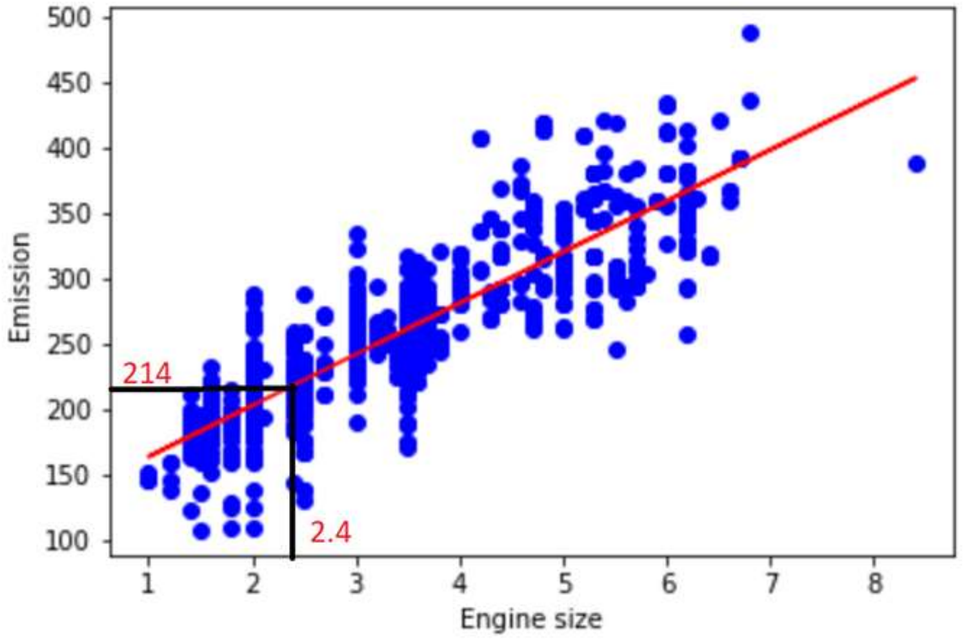
Garis Regresi Linier

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?



Cara Kerja Regresi Linier

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?



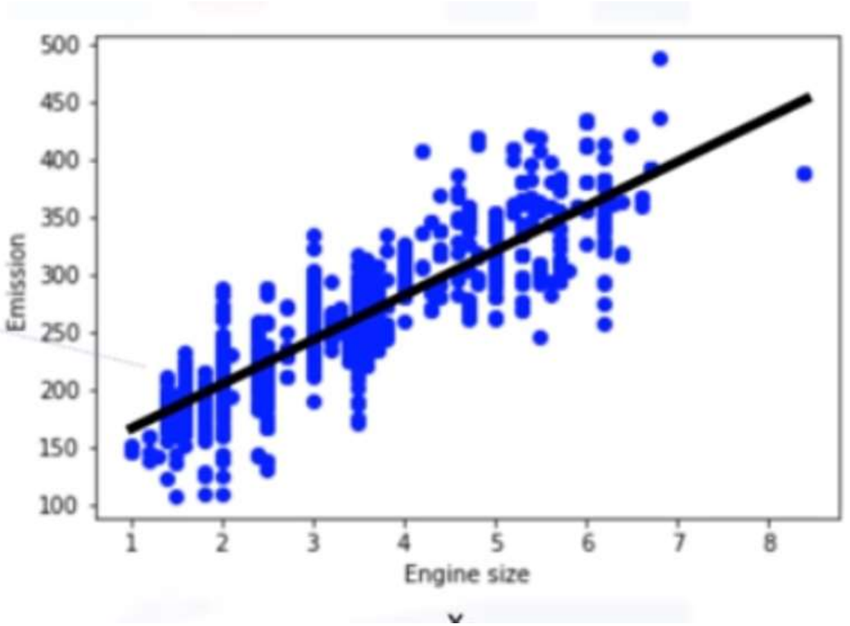
Cara Kerja Regresi Linier

$$\hat{y} = \theta_0 + \theta_1 x_1$$

response variable

a single predictor

y



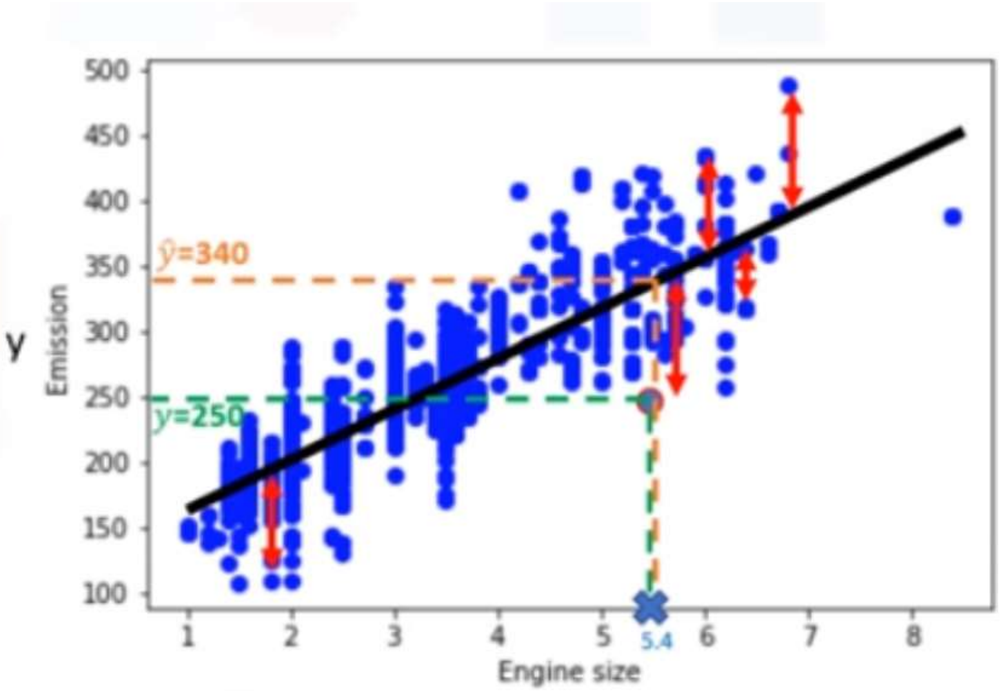
Cara Mencari Parameter Model Terbaik

x_1 = **5.4** independent variable
y= 250 actual Co2 emission of x1

$\hat{y} = \theta_0 + \theta_1 x_1$
$\hat{y}$ = 340 the predicted emission of x1

Error = y- $\hat{y}$
= 250 – 340
= -90

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



Estimasi Parameter

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

X_1 (ENGINESIZE) — y (CO2EMISSIONS)

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$\theta_1 = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$$

$$\bar{x} = (2.0 + 2.4 + 1.5 + \dots)/9 = 3.34$$

$$\bar{y} = (196 + 221 + 136 + \dots)/9 = 256$$

$$\theta_1 = \frac{(2.0 - 3.34)(196 - 256) + (2.4 - 3.34)(221 - 256) + \dots}{(2.0 - 3.34)^2 + (2.4 - 3.34)^2 + \dots}$$

$$\theta_1 = 39$$

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

$$\theta_0 = 256 - 39 * 3.34$$

$$\theta_0 = 125.74$$

$$\hat{y} = 125.74 + 39 x_1$$

Prediksi dengan Model Regresi Linier

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$Co2Emission = \theta_0 + \theta_1 \, EngineSize$$

$$Co2Emission = 125 + 39 \, EngineSize$$

$$Co2Emission = 125 + 39 \times 2.4$$

$$Co2Emission = 218.6$$

Kelebihan Regresi Linier

- Sederhana
- Tidak perlu tuning parameter
- Mudah dipahami dan diinterpretasikan

Latihan

Buatlah Model Regresi Linier Sederhana menggunakan Rapid Miner

Contoh Regresi Linier Variabel Jamak

Efektivitas variabel-variabel bebas terhadap prediksi

- Apakah kegelisahan, kehadiran dosen, dan jenis kelamin mempunyai efek pada kinerja ujian mahasiswa?

Prediksi dampak perubahan

- Seberapa besar kenaikan/penurunan tekanan darah terhadap kenaikan/penurunan BMI dari pasien?

Prediksi Nilai Kontinu pada Regresi Linier Variabel Jamak

X: Independent variable

Y: Dependent variable

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

$$Co2\ Em = \theta_0 + \theta_1 Engine\ size + \theta_2 Cylinders + \ ...$$

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ ... + \theta_n x_n$$

$$\hat{y} = \theta^T X$$

$$\theta^T = [\theta_0, \theta_1, \theta_2, ...] \qquad X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ ... \end{bmatrix}$$

MSE Untuk Menunjukkan Error Pada Model

$$\hat{y} = \theta^T X$$

$\hat{y}_i = 140$ the predicted emission of x_i

$y_i = 196$ actual value of x_i

$y_i - \hat{y}_i = 196 - 140 = 56$ residual error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

Estimasi Parameter Regresi Linier Variabel Jamak

Cara-cara mengestimasi parameter θ

Least Squares
- Operasi aljabar linier
- Perlu waktu yang lama untuk dataset yang besar (lebih dari 10000 baris)

Algoritma optimisasi
- Gradient Descent
- Metode yang sesuai apabila dataset sangat besar

Prediksi Menggunakan Regresi Linier Variabel Jamak

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

$$\hat{y} = \theta^T X$$

$$\theta^T = [125, 6.2, 14, \dots]$$

$$\hat{y} = 125 + 6.2x_1 + 14x_2 +$$

$$Co2Em = 125 + 6.2EngSize + 14\ \text{Cylinders} + \dots$$

$$Co2Em = 125 + 6.2 \times 2.4 + 14 \times 4 + \dots$$

$$Co2Em = 214.1$$

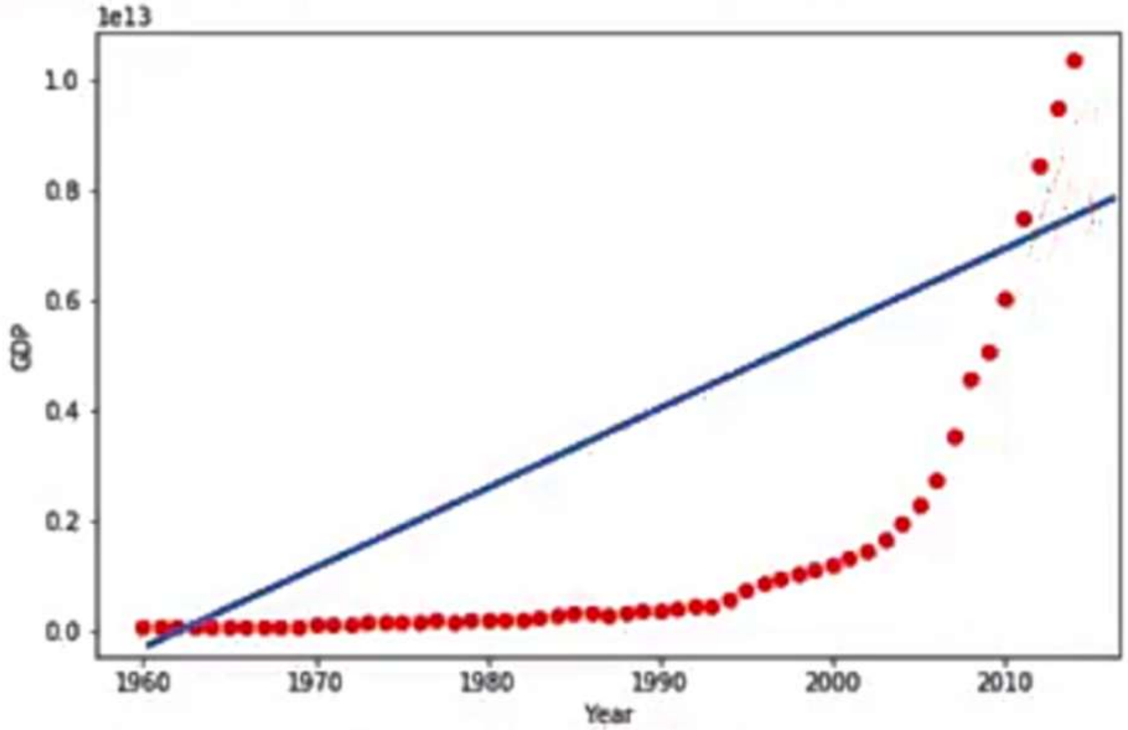
Latihan

- Buatlah Model Regresi Linier Variabel Jamak menggunakan Rapid Miner

Regresi Non Linier

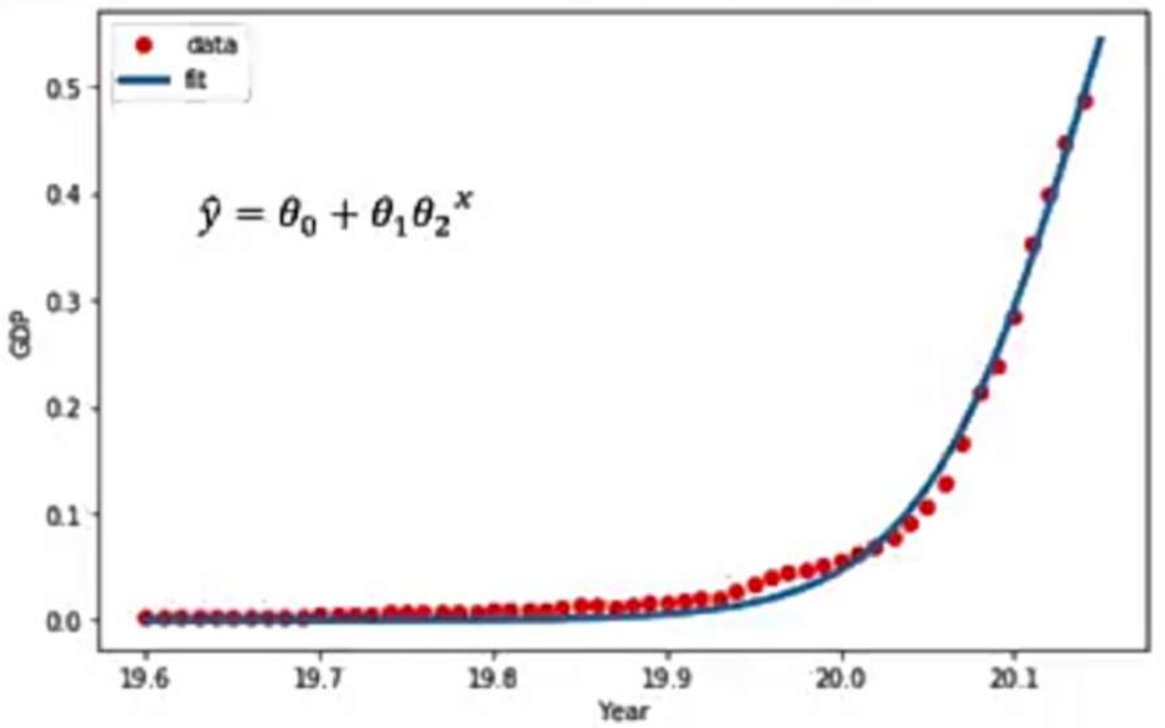
Mengapa Regresi Non-Linier Diperlukan?

	Year	Value
0	1960	5.918412e+10
1	1961	4.955705e+10
2	1962	4.668518e+10
3	1963	5.009730e+10
4	1964	5.906225e+10
5	1965	6.970915e+10
6	1966	7.587943e+10
7	1967	7.205703e+10
8	1968	6.999350e+10
9	1969	7.871882e+10
...	.....	.....

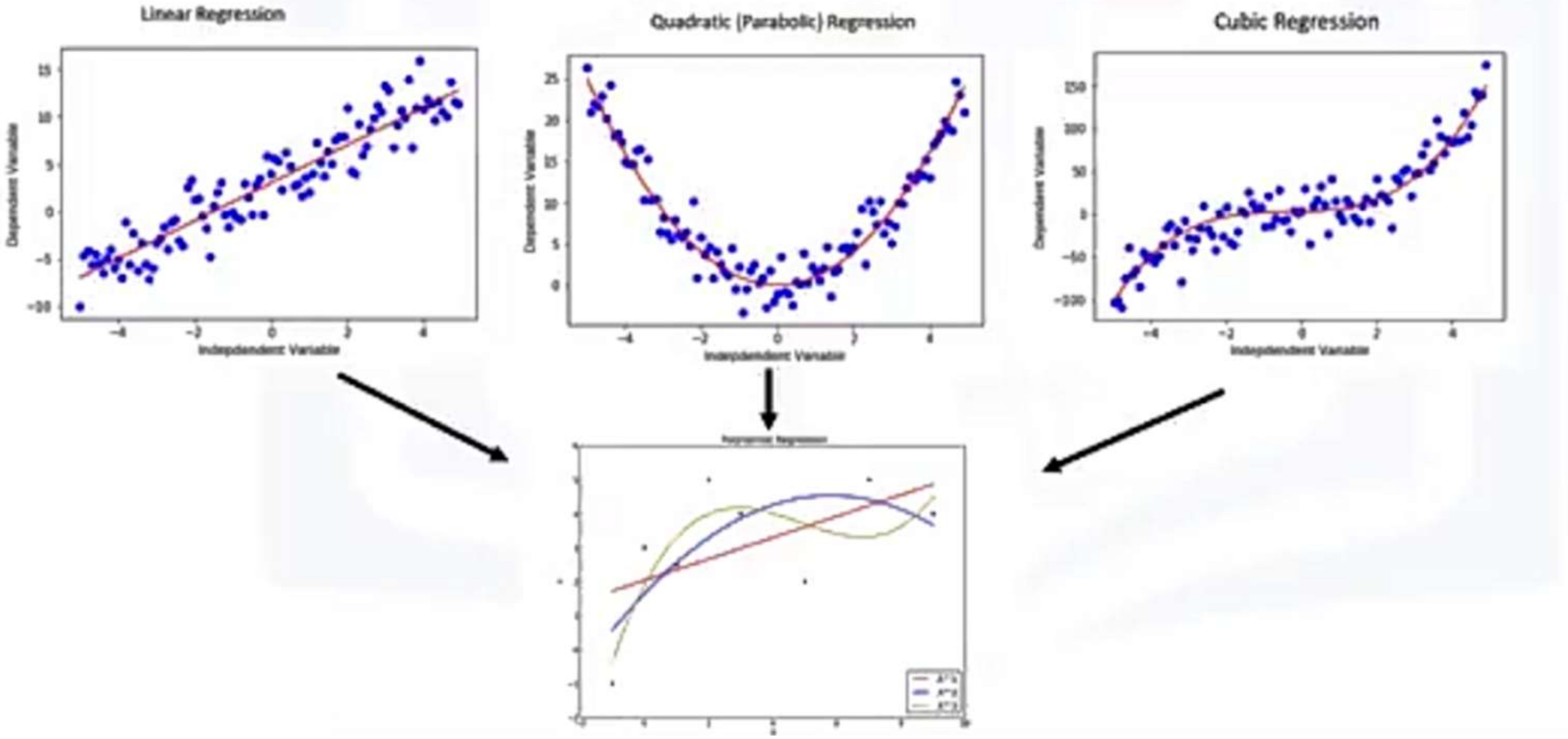


Mengapa Regresi Non-Linier Diperlukan?

	Year	Value
0	1960	5.918412e+10
1	1961	4.955705e+10
2	1962	4.668518e+10
3	1963	5.009730e+10
4	1964	5.906225e+10
5	1965	6.970915e+10
6	1966	7.587943e+10
7	1967	7.205703e+10
8	1968	6.999350e+10
9	1969	7.871882e+10
...	.....	.....



Tipe Regresi



Regresi Polinomial

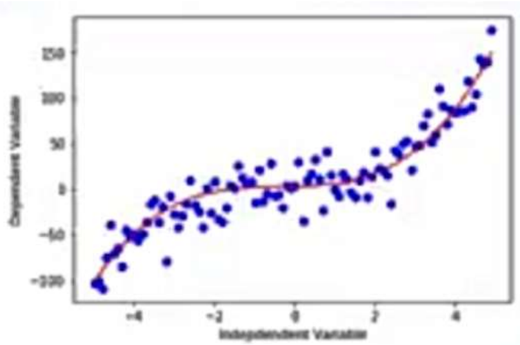
Beberapa data yang berbentuk kurva dapat dimodelkan dengan regresi linier

Contoh: $\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$

Model regresi polinomial dapat ditransformasikan menjadi model regresi linier.

$x_1 = x$
$x_2 = x^2$
$x_3 = x^3$

$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$



Regresi Non-Linier

Memodelkan hubungan tidak linier antara variabel tak bebas dengan himpunan variabel bebas

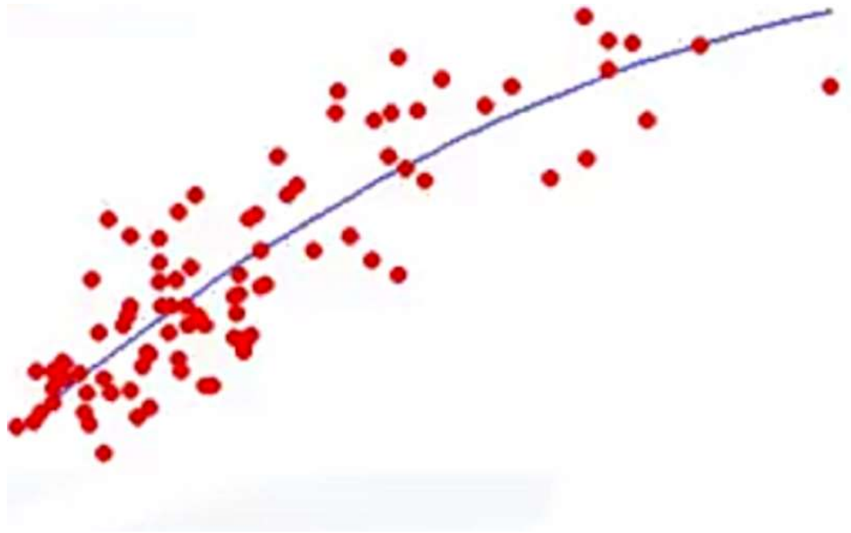
ŷ berupa fungsi non-linier dari parameter θ dan fitur x.

$$\hat{y} = \theta_0 + \theta_2{}^2 x$$

$$\hat{y} = \theta_0 + \theta_1 \theta_2{}^x$$

$$\hat{y} = \log(\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3)$$

$$\hat{y} = \frac{\theta_0}{1 + \theta_1{}^{(x-\theta_2)}}$$



Regresi Linier atau Non-Linier?

Cara untuk mengetahui apakah permasalahan cocok diselesaikan dengan regresi linier atau non linier

- Pengamatan visual atas data (visualisasi)
- Pengamatan akurasi hasil pemodelan

Cara untuk memodelkan data apabila visualisasi mengindikasikan non-linier

- Regresi polinomial
- Regresi non-linier
- Transformasi data non-linier menjadi linier


THANK
YOU