# Regresi

Bahan Kuliah IF4058 Topik Khusus Informatika I

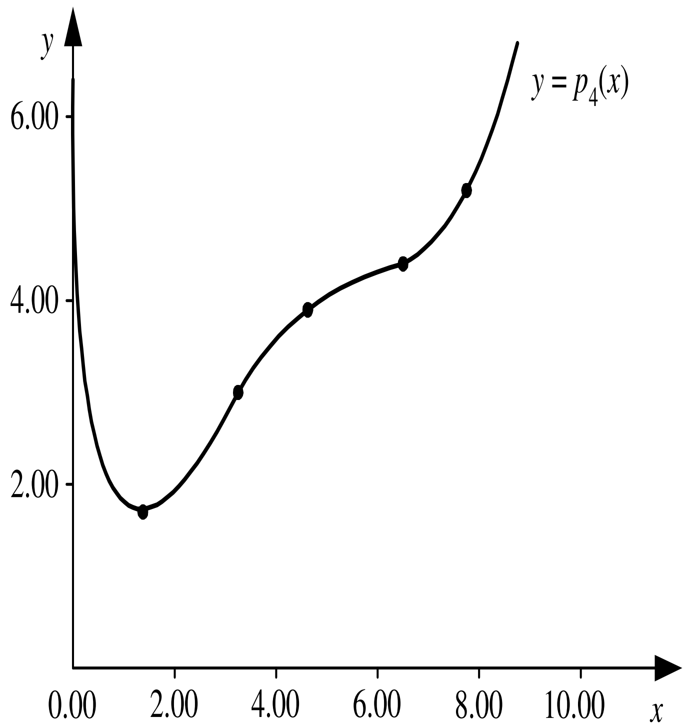
Oleh; Rinaldi Munir (IF-STEI ITB)

# Pendahuluan

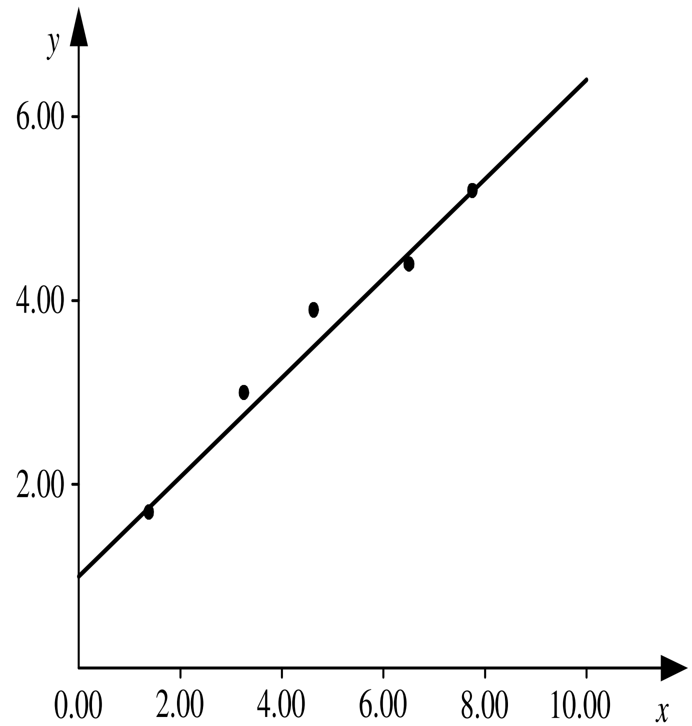
- Regresi adalah teknik pencocokan kurva untuk data yang berketelitian rendah.
- Contoh data yang berketelitian rendah data hasil pengamatan, percobaan di laboratorium, atau data statistik. Data seperti itu kita sebut *data hasil pengukuran.*
- Untuk data hasil pengukuran, pencocokan kurva berarti membuat fungsi mengampiri (*approximate*) titik-titik data.
- Kurva fungsi hampiran tidak perlu melalui semua titik data tetapi dekat dengannya tanpa perlu menggunakan polinom berderajat tinggi.

- **Contoh**: diberikan data jarak tempuh (*y*) sebuah kendaraaan - dalam mil- setelah *x* bulan seperti pada tabel di bawah ini

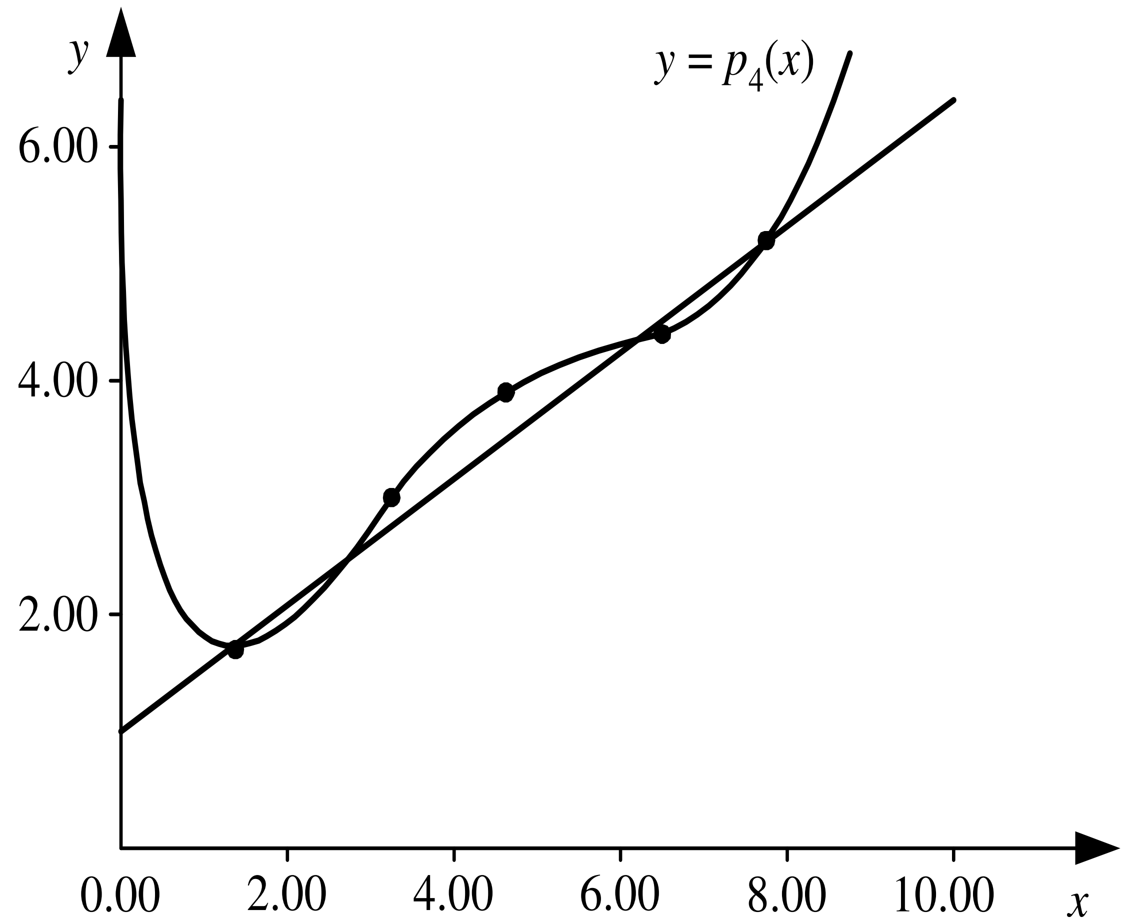
| $x$ | 1.38 | 3.39 | 4.75 | 6.56 | 7.76 |
|---|---|---|---|---|---|
| $y$ | 1.83 | 2.51 | 3.65 | 4.10 | 5.01 |




Interpolasi




Regresi




Dari kedua pencocokan tersebut, terlihat bahwa garis lurus memberikan hampiran yang *bagus*, tetapi belum tentu yang *terbaik*. Pengertian terbaik di sini bergantung pada cara kita mengukur galat hampiran.

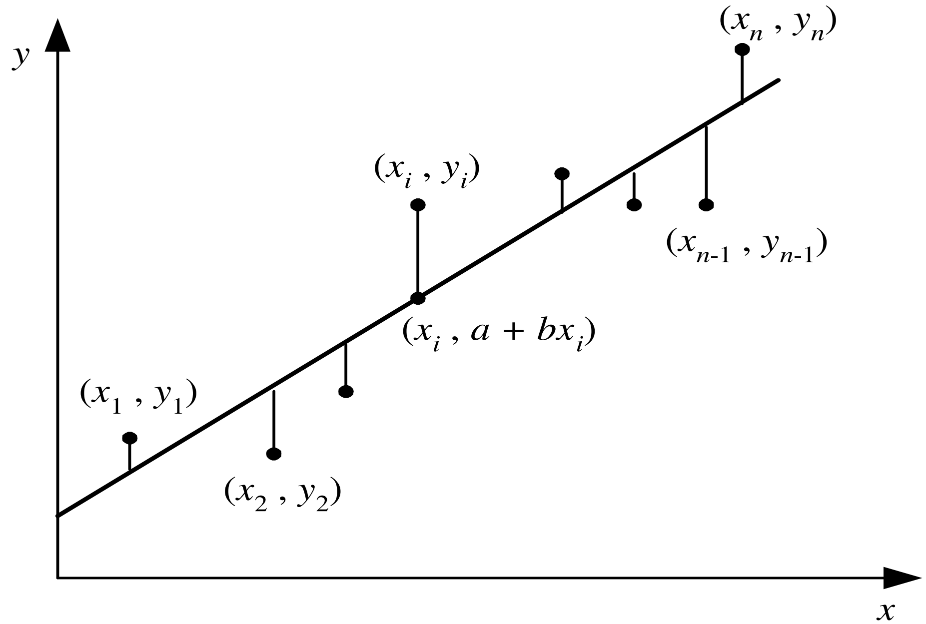
- Prinsip penting yang harus diketahui dalam mencocokkan kurva untuk data hasil pengukuran adalah:
  – Fungsi mengandung sesedikit mungkin parameter bebas
  – Deviasi fungsi dengan titik data dibuat minimum.

- Kedua prinsip di atas mendasari metode **regresi kuadrat terkecil**.

- Manfaat pencocokan kurva untuk data hasil pengukuran:
  1. Bagi ahli sains/rekayasa: mengembangkan formula empirik untuk sistem yang diteliti.
  2. Bagi ahli ekonomi: menentukan kurva kecenderungan ekonomi untuk “meramalkan” kecenderungan masa depan.

# Regresi Lanjar

- Misalkan ($x_i$, $y_i$) adalah data hasil pengukuran. Kita akan menghampiri titik-titik tersebut dengan sebuah garis lurus.
- Garis lurus tersebut dibuat sedemikian sehingga galatnya sekecil mungkin dengan titik-titik data.



- Karena data mengandung galat, maka nilai data sebenarnya, $g(x_i)$, dapat ditulis sebagai

$$g(x_i) = y_i + e_i \qquad\qquad i = 1, 2, ..., n \qquad\qquad (1)$$

- yang dalam hal ini, $e_i$ adalah galat setiap data. Diinginkan fungsi lanjar

$$\mathrm{f}(x) = a + bx \qquad\qquad (2)$$

- yang mencocokkan data sedemikian sehingga deviasinya,

$$r_i = y_i - f(x_i) = y_i - (a + bx_i) \qquad\qquad (3)$$

minimum.

Total kuadrat deviasi persamaan (4) adalah

$$R = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Agar *R* minimum, maka haruslah

$$\frac{\partial R}{\partial a} = -2\sum(y_i - a - bx_i) = 0$$

$$\frac{\partial R}{\partial b} = -2\sum x_i(y_i - a - bx_i) = 0$$

Untuk selanjutnya, notasi ditulis "$\sum$" saja.

***Penyelesaian:***

Masing-masing ruas kedua persamaaan dibagi dengan -2:

$$\sum(y_i - a - bx_i) = 0 \quad \Rightarrow \sum y_i - \sum a - \sum bx_i = 0$$
$$\sum x_i(y_i - a - bx_i) = 0 \quad \Rightarrow \sum x_i y_i - \sum ax_i - \sum bx_i^2 = 0$$

Selanjutnya,

$$\sum a + \sum bx_i = \sum y_i$$
$$\sum ax_i + \sum bx_i^2 = \sum x_i y_i$$

atau

$$na + b\sum x_i = \sum y_i$$
$$a\sum x_i + b\sum x_i^2 = \sum x_i y_i$$

Kedua persamaan terakhir ini dinamakan **persamaan normal**, dan dapat dapat ditulis dalam bentuk persamaan matriks:

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

Solusi (nilai *a* dan *b*) bisa dicari dengan metode eliminasi Gauss

Atau langsung dengan rumus:

$$b = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

$$a = \bar{y} - b\bar{x}$$

Untuk menentukan seberapa bagus fungsi hampiran mencocokkan data, kita dapat mengukurnya dengan **galat RMS** (*Root-mean-square error*):

$$E_{RMS} = \left( \frac{1}{n} \sum_{i=1}^{n} \left| f(x_i) - y_i \right|^2 \right)^2$$

Semakin kecil nilai $E_{RMS}$ semakin bagus fungsi hampiran mencocokkan titik-titik data.

- **Contoh**: Tentukan persamaan garis lurus yang mencocokkan data pada tabel di bawah ini. Kemudian, perkirakan nilai *y* untuk *x* = 1.0.
- **Penyelesaian:**

| $i$ | $x_i$ | $y_i$ | $x_i^2$ | $x_i\, y_i$ |
|---|---|---|---|---|
| 1 | 0.1 | 0.61 | 0.01 | 0.061 |
| 2 | 0.4 | 0.92 | 0.16 | 0.368 |
| 3 | 0.5 | 0.99 | 0.25 | 0.495 |
| 4 | 0.7 | 1.52 | 0.49 | 1.064 |
| 5 | 0.7 | 1.47 | 0.49 | 1.029 |
| 6 | 0.9 | 2.03 | 0.81 | 1.827 |
| | $\sum x_i = 3.3$ | $\sum y_i = 7.54$ | $\sum x_i^2 = 2.21$ | $\sum x_i\, y_i = 4.844$ |

Diperoleh sistem persamaan lanjar:

$$\begin{bmatrix} 6 & 3.3 \\ 3.3 & 2.21 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 7.54 \\ 4.844 \end{bmatrix}$$

Solusi SPL di atas adalah: $a = 0.2862$
$b = 1.7645$

Persamaan garis regresinya adalah: $f(x) = 0.2862 + 1.7645x$.

Perbandingan antara nilai $y_i$ dan $f(x_i)$:

| $i$ | $x_i$ | $y_i$ | $f(x_i) = a + bx_i$ | deviasi | (deviasi)$^2$ |
|---|---|---|---|---|---|
| 1 | 0.1 | 0.61 | 0.46261 | 0.147389 | 0.02172 |
| 2 | 0.4 | 0.92 | 0.99198 | -0.07198 | 0.00518 |
| 3 | 0.5 | 0.99 | 1.16843 | -0.17844 | 0.03184 |
| 4 | 0.7 | 1.52 | 1.52135 | -0.00135 | 0.00000 |
| 5 | 0.7 | 1.47 | 1.52135 | -0.05135 | 0.00264 |
| 6 | 0.9 | 2.03 | 1.87426 | 0.15574 | 0.02425 |
| | | | | | $\sum = 0.08563$ |

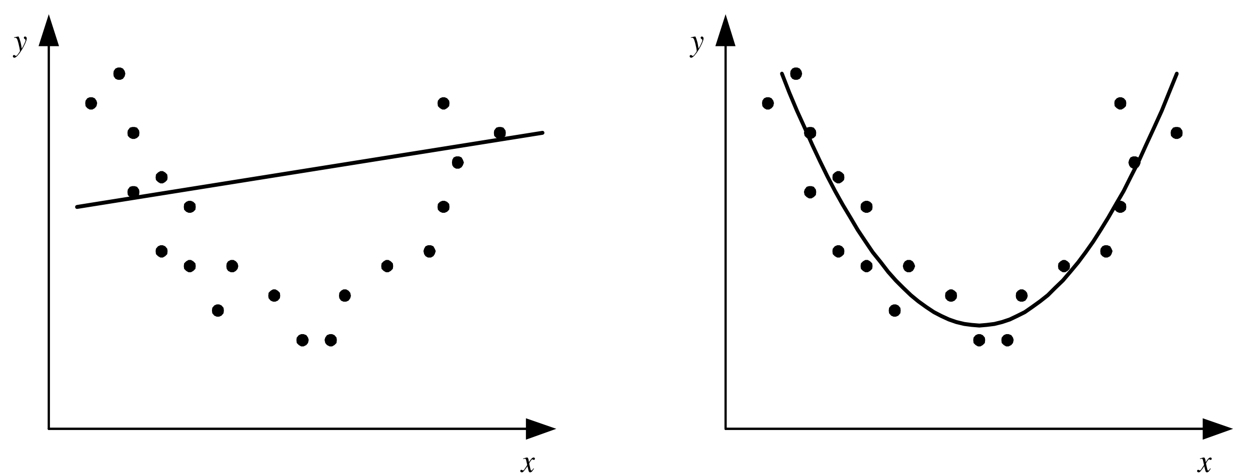
Taksiran nilai $y$ untuk $x = 1.0$ adalah

$$y = f(1.0) = 0.2862 + 1.7645(1.0) = 2.0507$$

Galat RMS adalah $E_{RMS} = \left(\frac{0.08563}{6}\right)^{1/2} = 0.119464$

# Pelanjaran

- Regresi lanjar hanya tepat bila data memiliki hubungan lanjar antara peubah bebas dan peubah terikatnya.
- Gambar berikut memperlihatkan bahwa garis lurus tidak tepat mewakili kecenderungan titi-titik data. Fungsi kuadratik lebih tepat menghampiri titik-titik tersebut.




($a$) ($b$)

- Langkah pertama dalam analisis regresi seharusnya berupa penggambaran titik-titik data pada diagram kartesian

- Kemudian secara visual memeriksa data untuk memastikan apakah berlaku suatu model lanjar atau model nirlanjar.

- Penggambaran titik-titik ini sekaligus juga sangat membantu dalam mengetahui fungsi yang tepat untuk mencocokkan data.

- Meskipun fungsi hampiran berbentuk nirlanjar, namun pencocokan kurva dengan fungsi nirlanjar tersebut dapat juga diselesaikan dengan cara regresi lanjar.

Tiga macam fungsi nirlanjar di bawah ini:

1. Persamaan pangkat sederhana

   $y = Cx^b$ , $C$ dan $b$ konstanta.

2. Model eksponensial

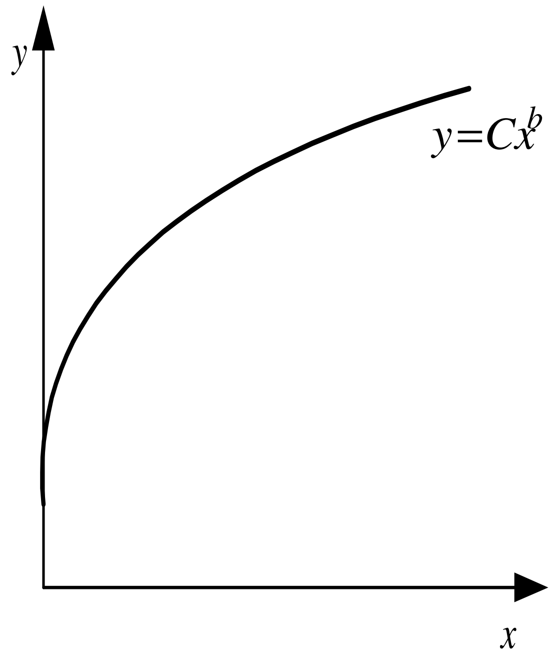
   $y = Ce^{bx}$, $C$ dan $b$ konstanta.

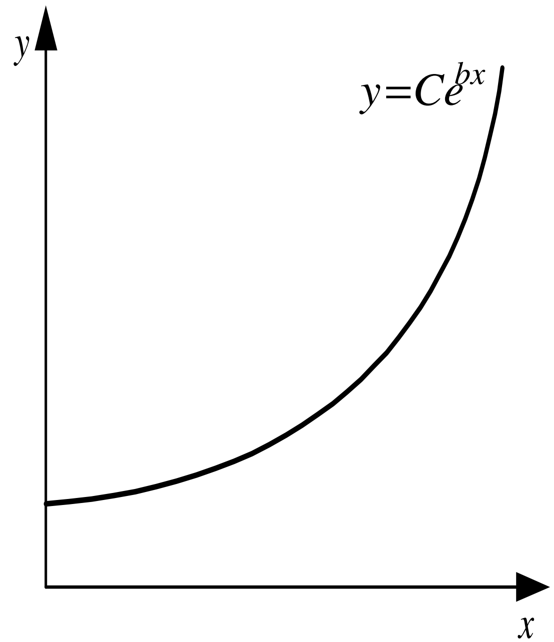
   Contoh: - model pertumbuhan populasi
   - model peluruhan zat radioaktif

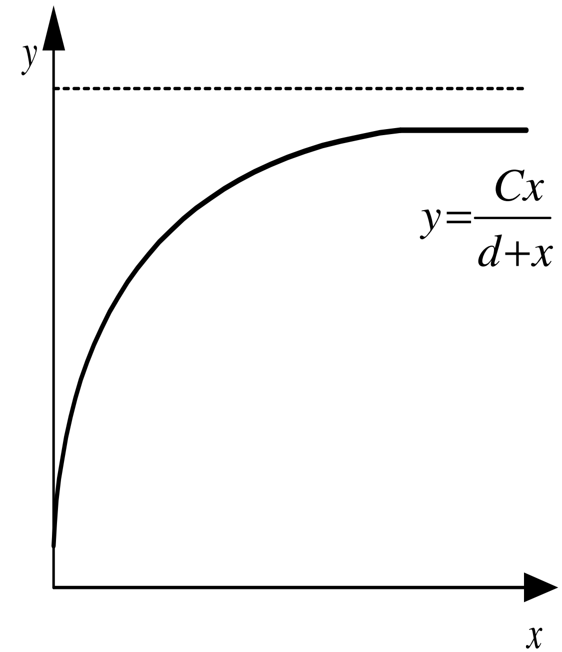
3. Persamaan laju pertumbuhan jenuh (*saturation growth-rate*)

   $$y = \frac{Cx}{d + x}$$ , $C$ dan $d$ konstanta.

Contoh: model pertumbuhan bakteri kondisi pembatas
(misalnya dibatasi oleh jumlah makanan)


y
$y=Cx^b$
x


y
$y=Ce^{bx}$
x


y
$y=\frac{Cx}{d+x}$
x

**Pelanjaran Persamaan Pangkat Sederhana**

Misalkan kita akan mencocokkan data dengan fungsi

$y = Cx^b$

Lakukan pelanjaran sebagai berikut:

$y = Cx^b \quad \Leftrightarrow \quad ln(y) = ln(C) + b\, ln(x)$

Definisikan

$Y = ln(y); \quad a = ln(C); \quad X = ln(x)$

Persamaan regresi lanjarnya adalah:

$Y = a + bX$

Lakukan pengubahan dari $(x_i,y_i)$ menjadi $(ln(x_i), ln(y_i))$, lalu hitung $a$ dan $b$ dengan cara regresi lanjar. Dari persamaan $a = ln(C)$, kita dapat menghitung nilai

$C = e^a$

Sulihkan nilai $b$ dan $C$ ke dalam persamaan pangkat $y = Cx^b$.

- **Contoh**: Cocokkan data berikut dengan fungsi $y = Cx^b$.

**Penyelesaian:**

| $i$ | $x_i$ | $y_i$ | $X_i = ln(x_i)$ | $Y_i = ln(y_i)$ | $X_i^2$ | $X_i Y_i$ |
|---|---|---|---|---|---|---|
| 1 | 0.1500 | 4.4964 | -1.8971 | 1.5033 | 3.5990 | -2.8519 |
| 2 | 0.4000 | 5.1284 | -0.9163 | 1.6348 | 0.8396 | -1.4980 |
| 3 | 0.6000 | 5.6931 | -0.5108 | 1.7393 | 0.2609 | -0.8884 |
| 4 | 1.0100 | 6.2884 | 0.0100 | 1.8387 | 0.0001 | 0.0184 |
| 5 | 1.5000 | 7.0989 | 0.4055 | 1.9599 | 0.1644 | 0.7947 |
| 6 | 2.2000 | 7.5507 | 0..7885 | 2.0216 | 0.6217 | 1.5940 |
| 7 | 2.4000 | 7.5106 | 0.8755 | 2.0163 | 0.7665 | 1.7653 |
| | | | $\sum X_i$ = -1.2447 | $\sum Y_i$ = 12.7139 | $\sum X_i^2$ = 6.2522 | $\sum X_i Y_i$ = -1.0659 |

Diperoleh sistem persamaan lanjar

$$\begin{bmatrix} 7 & -1.2447 \\ -1.2447 & 6.2522 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 12.7139 \\ -1.0659 \end{bmatrix}$$

Solusi SPL di atas: $a = 1.8515$ dan $b = 0.1981$.

Hitung $C = e^a = e^{1.8515} = 6.369366$

Jadi,titik-titik $(x, y)$ pada tabel di atas dihampiri dengan fungsi pangkat sederhana:

$$y = 6.369366\, x^{0.1981}$$

**Pelanjaran Model Eksponensial $y = Ce^{bx}$**

Misalkan kita akan mencocokkan data dengan fungsi

$$y = Ce^{bx}$$

Lakukan pelanjaran sebagai berikut:

$$y = Ce^{bx} \quad \Leftrightarrow \ln(y) = \ln(C) + bx \ln(e)$$

$$\Leftrightarrow \ln(y) = \ln(C) + bx \quad \text{(ingat: } \ln(e) = 1\text{)}$$

Definisikan

$$Y = \ln(y); \quad a = \ln(C); \quad X = x$$

Persamaan regresi lanjarnya:

$$Y = a + bX$$

Lakukan pengubahan dari $(x_i, y_i)$ menjadi $(x_i, \ln(y_i))$, lalu hitung $a$ dan $b$ dengan cara regresi lanjar.

Dari persamaan $a = \ln(C)$, kita dapat menghitung nilai $C = e^a$. Sulihkan nilai $b$ dan $C$ ke dalam persamaan eksponensial $y = Ce^{bx}$.

**Pelanjaran Model Laju Pertumbuhan Jenuh** $y = \dfrac{Cx}{d + x}$

Misalkan kita akan mencocokkan data dengan fungsi

$$y = \frac{Cx}{d + x}$$

Lakukan pelanjaran sebagai berikut:

$$y = \frac{Cx}{d + x}$$

$$\frac{1}{y} = \frac{d}{C}\frac{1}{x} + \frac{1}{C}$$

Definisikan

$$Y = 1/y$$
$$a = 1/C$$
$$b = d/C$$
$$X = 1/x$$

Persamaan regresi lanjarnya:

$$Y = a + bX$$

Lakukan pengubahan dari $(x_i,y_i)$ menjadi $(1/x_i, 1/y_i)$, lalu hitung $a$ dan $b$ dengan cara regresi lanjar.

Dari persamaan $a = 1/C$, kita dapat menghitung nilai $C = 1/a$.
Dari persamaan $b = d/C$, kita dapat menghitung $d = bC$ .
Sulihkan $d$ dan $C$ ke dalam persamaan laju pertumbuhan jenuh $y = Cx/(d+x)$.

| Fungsi $y = f(x)$ | Bentuk lanjar $y = a + bX$ | Perubahan peubah dan kontanta |
| --- | --- | --- |
| $y = Cx^b$ | $ln(y) = ln(C) + b\ ln(x)$ | $Y = ln(y)$, $X = ln(x)$, $C = e^a$ |
| $y = Ce^{bx}$ | $ln(y) = ln(C) + bx$ | $Y = ln(y)$, $X = x$, $C = e^a$ |
| $y = \dfrac{Cx}{d + x}$ | $\dfrac{1}{y} = \dfrac{d}{C}\dfrac{1}{x} + \dfrac{1}{C}$ | $Y = 1/y$, $X = 1/x$, $C = 1/a$, $d = bC$ |
| $y = a + \dfrac{b}{x}$ | $y = a + b\dfrac{1}{x}$ | $Y = y$, $X = 1/x$ |
| $y = \dfrac{D}{x + C}$ | $y = \dfrac{D}{C} + \dfrac{-1}{C}(xy)$ | $Y = y$, $X = xy$, $C = \dfrac{-1}{b}$, $D = \dfrac{-a}{b}$ |
| $y = \dfrac{1}{a + bx}$ | $\dfrac{1}{y} = a + bX$ | $Y = \dfrac{1}{y}$, $X = x$ |
| $y = (a + bx)^{-2}$ | $y^{-1/2} = a + bX$ | $Y = y^{-1/2}$, $X = x$ |
| $y = Cxe^{-Dx}$ | $\ln(\dfrac{y}{x}) = \ln(C) + (-Dx)$ | $Y = \ln(\dfrac{y}{x})$, $X = x$, $C = e^a$, $D = -b$ |