

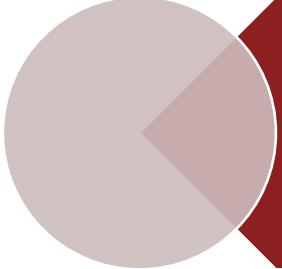
# Modul 8

# Regresi Linear

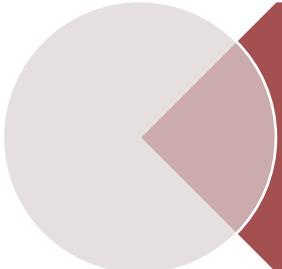
YT

MATH1042 – Peluang dan Statistika

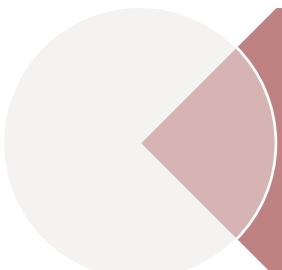
# Outline



## 8.1 Pengertian dasar Regresi Linear



## 8.2 Ukuran Kecocokan Model



## 8.3 Transformasi Data untuk Regresi Linear

# 8.1 Pengertian dasar Regresi Linear

# Capaian Pembelajaran

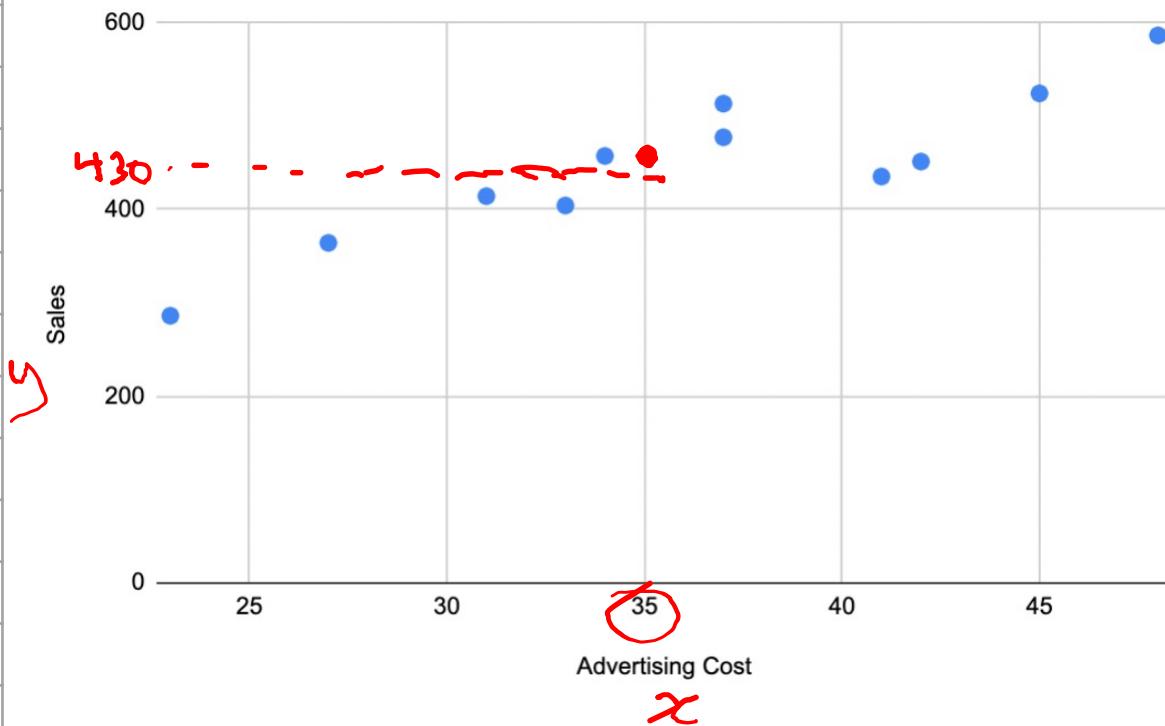
- Mahasiswa mampu menerapkan komputasi model Regresi Linear, baik secara perhitungan **manual** maupun dengan bantuan **Python**.

# Prasyarat Pembelajaran

- Estimasi Interval
- Teori Peluang Dasar

# Kaitan antara Variabel

Adv Cost (x)	Sales (y)
37	477
45	524
37	513
48	586
33	404
23	286
31	414
34	457
27	364
41	435
42	451
48	549
25	249
26	315
45	537
44	521



Jika adv cost = 35, berapakah salesnya?

**Variabel Bebas (Regressor):**

adalah variabel yang nilainya mempengaruhi variabel lain.

Notasi:  $x$

**Variabel Terikat (Response):**

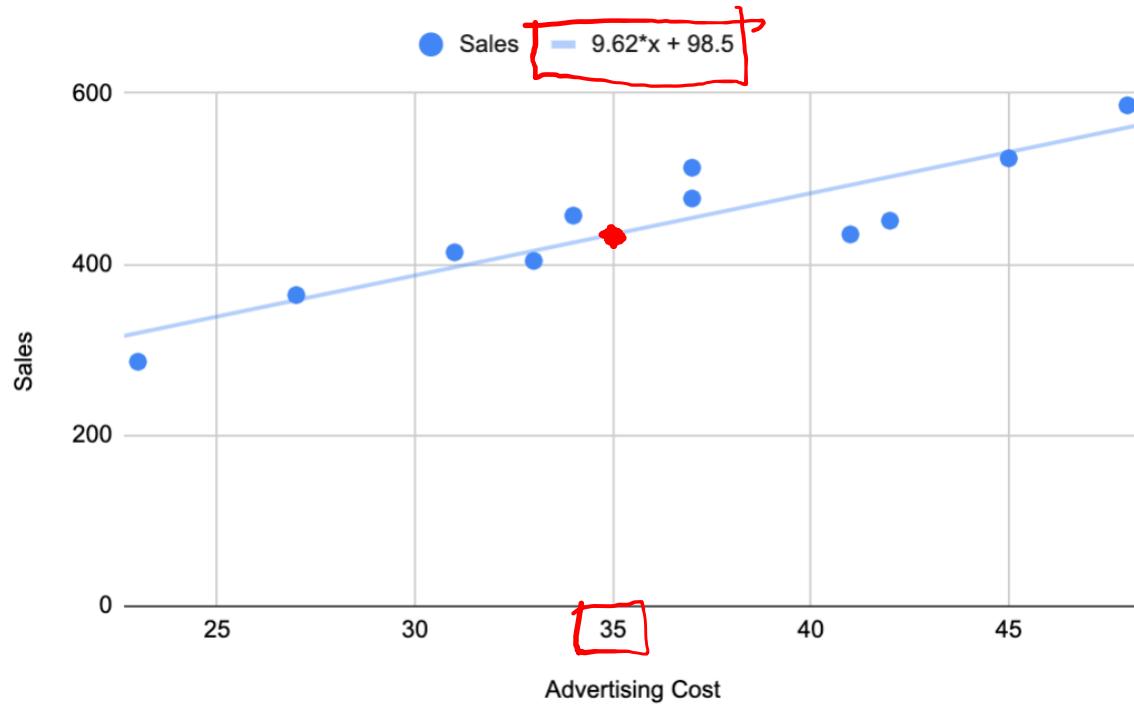
adalah variabel yang nilainya dipengaruhi variabel lain.

Notasi:  $y$

# Data dan Model

Data:

Adv Cost (x)	Sales (y)
37	477
45	524
37	513
48	586
33	404
23	286
31	414
34	457
27	364
41	435
42	451
48	549
25	249
26	315
45	537
44	521



Model:

$$\hat{y} = 9.62x + 98.5$$

Data:

adalah nilai yang secara kenyataannya terjadi;

merupakan realisasi dari unpredictability.

Model:

adalah nilai yang secara idealnya terjadi;

merupakan formulasi dari regularity.

kayak harusnya begini kalau perfect

# Model Regresi Linear Sederhana

$y = mx + c \rightarrow$  yang dicari itu m sama c nya

- Bentuk umum:

$$\hat{y} = \beta_0 + \beta_1 x$$

- Parameter:

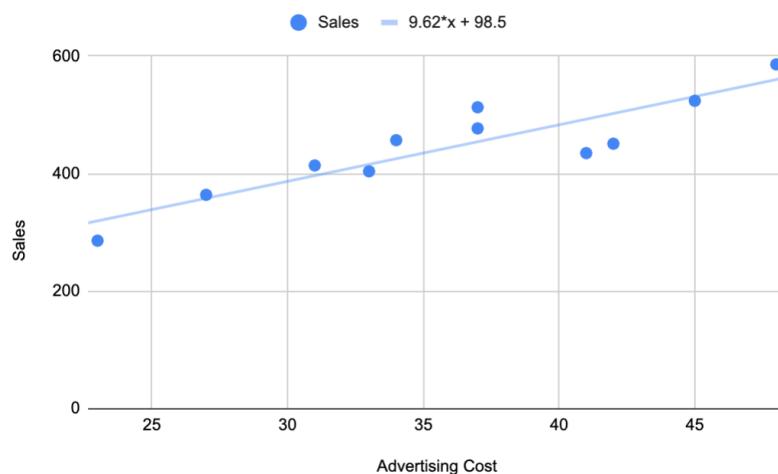
$$\beta_0, \beta_1$$

- Data:

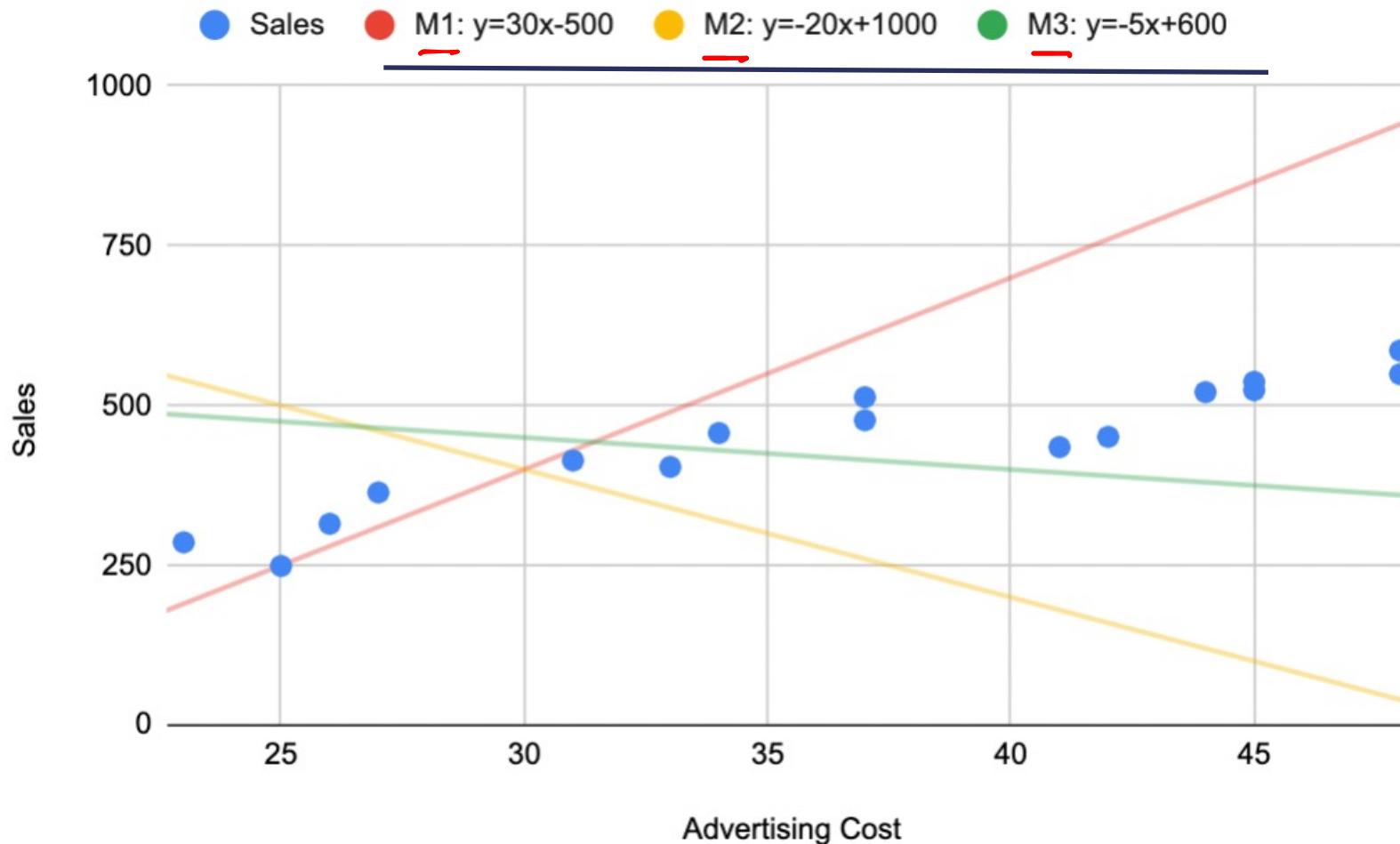
Adv Cost (x)	Sales (y)
37	477
45	524
37	513
48	586
33	404
23	286
31	414
34	457
27	364
41	435
42	451
48	549
25	249
26	315
45	537
44	521

## Visi Misi Regresi Linear

Yang diketahui:	Data hasil <u>realisasi</u> suatu ketidakpastian.
Yang dipercaya: kalo perfect itu jadi kayak model	Hasil data realisasi tersebut haruslah <u>terpancar</u> dari sebuah <u>distribusi/model</u> teoretis, namun dengan tambahan suatu nilai <u>error</u> .
Yang diasumsikan:	Model teoretis diasumsikan dalam bentuk <u>persamaan linear</u> .
Yang dicari:	Nilai <u>parameter</u> model.



# Masalah RL = Masalah memilih parameter



kayak trendline excel, tapi yang dicari itu equationnya yang merupakan persamaan garisnya

# Kaitan antara Model dan Data

Model:

$$\hat{y}^{(i)} = \beta_0 + \beta_1 x^{(i)}$$

y pred ->

n : jumlah data

$x^{(i)}$  : data regressor (var bebas) ke-i

$y^{(i)}$  : data response (var terikat) ke-i

$\hat{y}^{(i)}$  : nilai prediksi model/hasil regresi ke-i

$\varepsilon^{(i)}$  : error dari hasil regresi ke-i

$\beta_0, \beta_1$  : parameter model

Data:

$$y^{(i)} = \beta_0 + \beta_1 x^{(i)} + \varepsilon^{(i)}$$

Error/golat

$$\varepsilon \sim N(0, \sigma)$$

distribusi normal  
jadi secara ekspektasi/rata2  
nilainya adalah 0

Kaitan antara Model & Data:

$$y^{(i)} = \hat{y}^{(i)} + \varepsilon^{(i)}$$

$$\mathbb{E}y^{(i)} = \mathbb{E}(\hat{y}^{(i)} + \varepsilon^{(i)})$$

$$\mathbb{E}y^{(i)} = \mathbb{E}\hat{y}^{(i)} + \mathbb{E}\varepsilon^{(i)}$$

$$\boxed{\mathbb{E}y^{(i)} = \mathbb{E}\hat{y}^{(i)} + 0}$$

Jadi, secara ekspektasi, hasil regresi yang ideal akan bernilai persis seperti data real/kenyataannya,  
kalau perfect ??

# Metode Kuadrat Terkecil

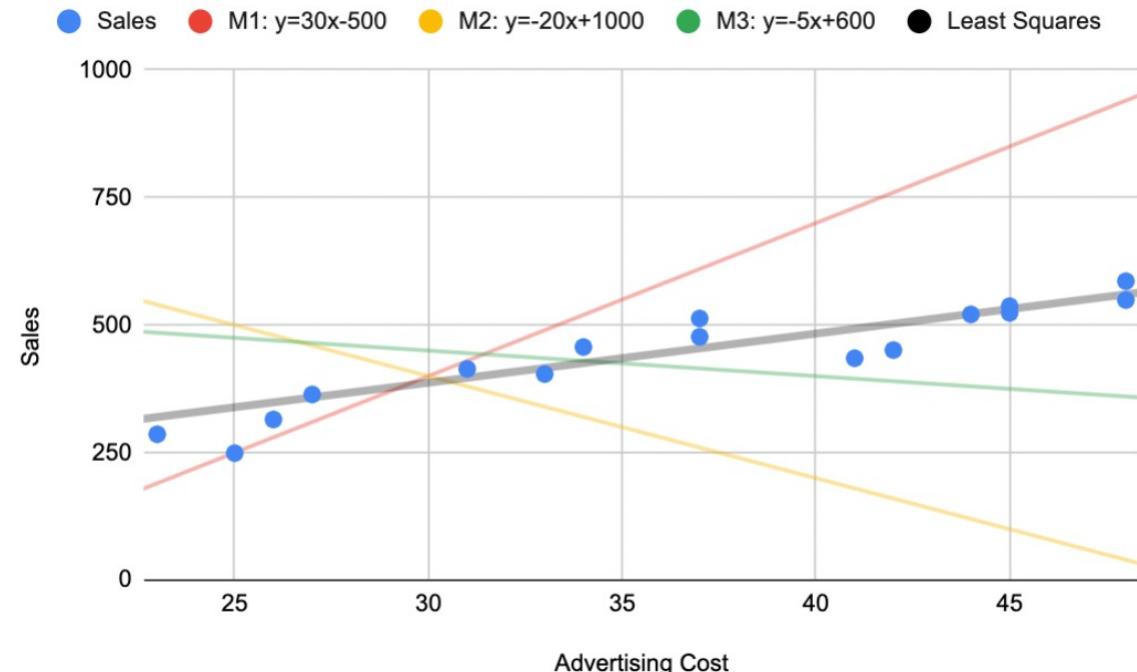
(Least Square)

salah satu cara menentukan parameter terbaik

- Bagaimana membuat model RL terbaik?
  - Dengan menentukan parameter terbaik.
- Bagaimana menentukan parameter terbaik?
  - Dengan menggunakan metode kuadrat terkecil.
- Ide: parameter diatur agar model memberikan hasil yang 'sedekat mungkin dengan data.

rumus dasarnya

$$\hat{y} = \beta_0 + \beta_1 x$$



# Residu/Error Kuadrat

R square

Definisi. (Residu suatu titik)

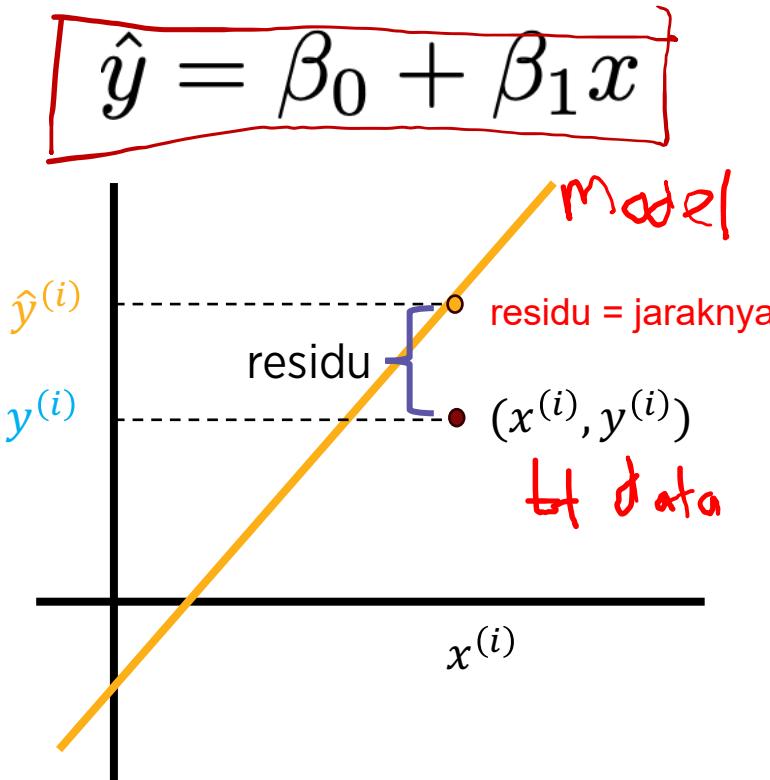
Misalkan  $(x^{(i)}, y^{(i)})$  adalah sebuah titik data.

Misalkan pula  $\hat{y}^{(i)}$  adalah nilai prediksi dari model RL.  
 $y_{\text{pred}}$

Maka, nilai residu/error kuadrat dari titik ke-i adalah:

$$\begin{aligned}\varepsilon^{(i)} &= \text{data} - \text{model} \\ &= (y^{(i)} - (\beta_0 + \beta_1 x^{(i)}))^2\end{aligned}$$

rumusnya ini pake aja



# Residu/Error Kuadrat

Definisi. (Residu keseluruhan dataset)

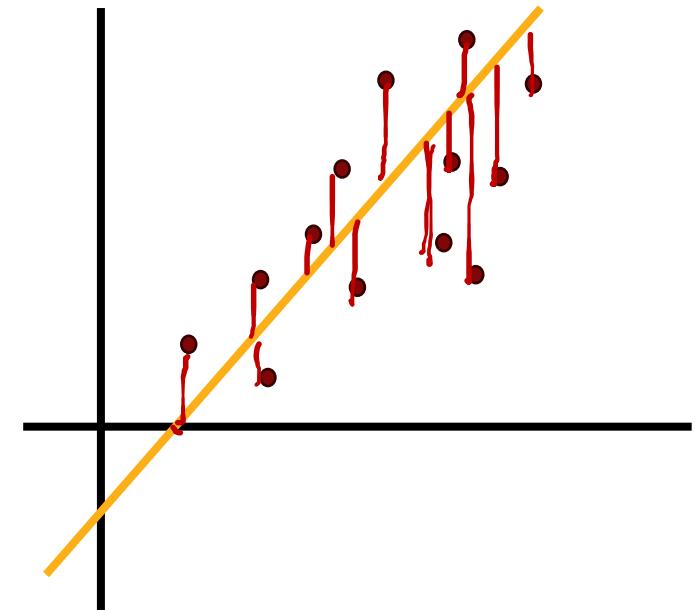
Misalkan  $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$  adalah keseluruhan titik data.

Maka, nilai residu/error kuadrat (*SSE-Sum of Squared Error*) dari model terhadap keseluruhan data adalah:

$$SSE = \sum_{i=1}^n \underbrace{(y^{(i)} - (\hat{y}^{(i)}))}_\text{residu untuk titik-i}^2$$

sigma artinya di jumlahkan semua

$$\hat{y} = \beta_0 + \beta_1 x$$



## Metode kuadrat terkecil

Prinsip: model yang baik adalah yang SSEnya sekecil-kecilnya.

# Kuadrat Terkecil: Meminimalkan Residu

- Diberikan data latih:  $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$  x\_train & y\_train nya
  - Dicari nilai:  $\beta_0$  dan  $\beta_1$
  - Agar nilai residu sekecil-kecilnya:  
$$\sum_{i=1}^n (y^{(i)} - \beta_0 + \beta_1 x^{(i)})^2$$
  - Masalah optimisasi multivariabel tanpa kendala:
    - Tentukan nilai  $\beta_0$  dan  $\beta_1$  agar fungsi biaya berikut bernilai sekecil-kecilnya:
- $\min \beta_0, \beta_1$  SSE( $\beta_0, \beta_1$ ) =  $\sum_{i=1}^n (y^{(i)} - \beta_0 + \beta_1 x^{(i)})^2$
- generalisasi cost funct
- cost function  
loss function

# Kuadrat Terkecil: Meminimalkan Residu

- Tentukan nilai  $\beta_0$  dan  $\beta_1$  agar SSE bernilai sekecil-kecilnya:

$$\text{SSE}(\beta_0, \beta_1) = \sum_{i=1}^n (y^{(i)} - \beta_0 + \beta_1 x^{(i)})^2$$

- Agar optimal,  $\beta_0$  &  $\beta_1$  harus memenuhi

$$\frac{\partial \text{SSE}(\beta_0, \beta_1)}{\partial \beta_0} = 0, \quad \frac{\partial \text{SSE}(\beta_0, \beta_1)}{\partial \beta_1} = 0$$

- Solusi:

$$\beta_1 = \frac{\sum_{i=1}^n (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sum_{i=1}^n (x^{(i)} - \bar{x})^2}$$

## Teorema Min/Max Fungsi Multivariabel

--> turunannya = 0

Titik min/max dari fungsi  $F(x_1, x_2, \dots, x_n)$  haruslah merupakan titik stasioner:

$$\frac{\partial F}{\partial x_1} = 0, \quad \frac{\partial F}{\partial x_2} = 0, \quad \dots \quad \frac{\partial F}{\partial x_n} = 0$$

PR: Buktikan!

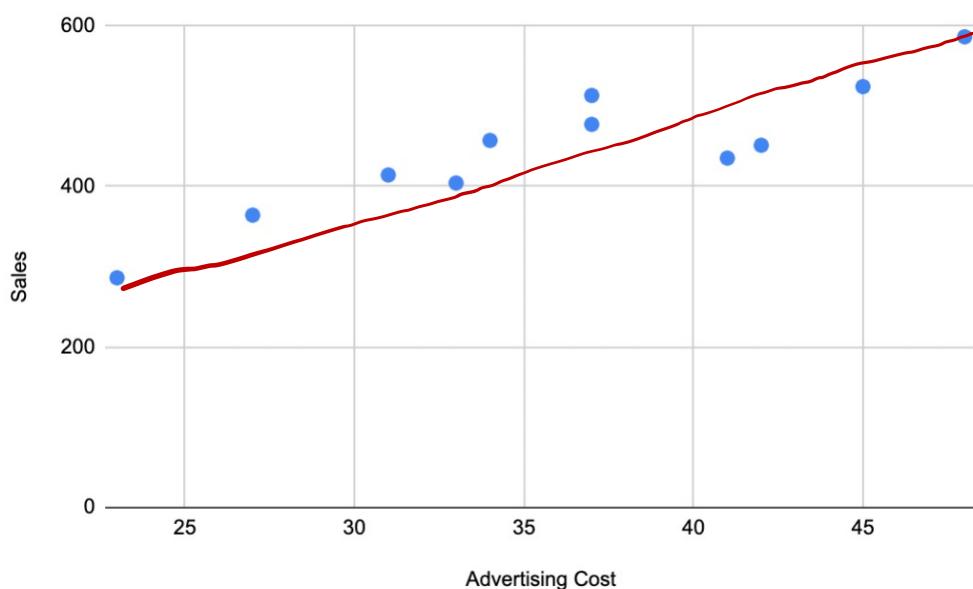
$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

rumus b0 dan b1 nya

# Contoh Perhitungan

Adv Cost (x)	Sales (y)
37	477
45	524
37	513
48	586
33	404
23	286
31	414
34	457
27	364
41	435
42	451
48	549
25	249
26	315
45	537
44	521

$$\bar{x} = 36.625; \quad \bar{y} = 442.625$$



cara yang sama dipake buat equation trendline  
excel kayaknya (harusnya)

$$\beta_1 = \frac{\sum_{i=1}^n (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sum_{i=1}^n (x^{(i)} - \bar{x})^2}$$
$$= 10.839$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 45.669$$

Jadi, model RL terbaik adalah:

$$\hat{y} = 45.669 + 10.839x$$

# Perspektif Statistika dalam RL

- Besarnya  $\beta_0$  dan  $\beta_1$  pada model regresi yang terbaik sebetulnya bersifat terselubung/unknown;
- Seperti halnya kita tidak mengetahui nilai statistik **populasi**, namun dapat membuat inferensi melalui statistik **sampel**, kita juga dapat:
  - menghitung nilai  $b_0$  dan  $b_1$  (**BUKAN** nilai yang sesungguhnya); hanya approximasi yang mendekati
  - dengan menggunakan data yang ada;
  - sebagai estimasi dari nilai  $\beta_0$  dan  $\beta_1$  (nilai yang sesungguhnya).
- Konsekuensi:
  - Nilai  $b_0$  dan  $b_1$  sebagai estimator dapat dilakukan prosedur: cara berpikir statistika bukan machine learning
    - Estimasi titik
    - Estimasi interval
    - Uji Hipotesis

# Estimator standar deviasi error

data

$$y^{(i)} = \underbrace{\beta_0 + \beta_1 x^{(i)}}_{\text{deterministik}} + \underbrace{\varepsilon^{(i)}}_{\text{random}}$$
$$\varepsilon \sim N(0, \sigma)$$

nilai parameter yang tidak diketahui  
→ perlu diestimasi

Estimator untuk  $\sigma^2$  (variansi residu model RL):

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

MSE (Mean Squared Error)

Estimator untuk  $\sigma$  (stdev residu model RL):

$$RMSE = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (\hat{y}^{(i)} - y^{(i)})^2}$$

RMSE (Root Mean Squared Error)

rmse untuk menghitung std dev di distribusi normal errornya

salah satu contoh penerapannya, gunakan untuk estimasi yang lain juga

# Penerapan RMSE untuk Estimasi Interval RL

- Ingat kembali Rule of Thumb Distribusi Normal: Jika  $\varepsilon \sim N(0, \sigma)$  maka:

$$\mathbb{P}(-2\sigma < \varepsilon < 2\sigma) = 95\%$$

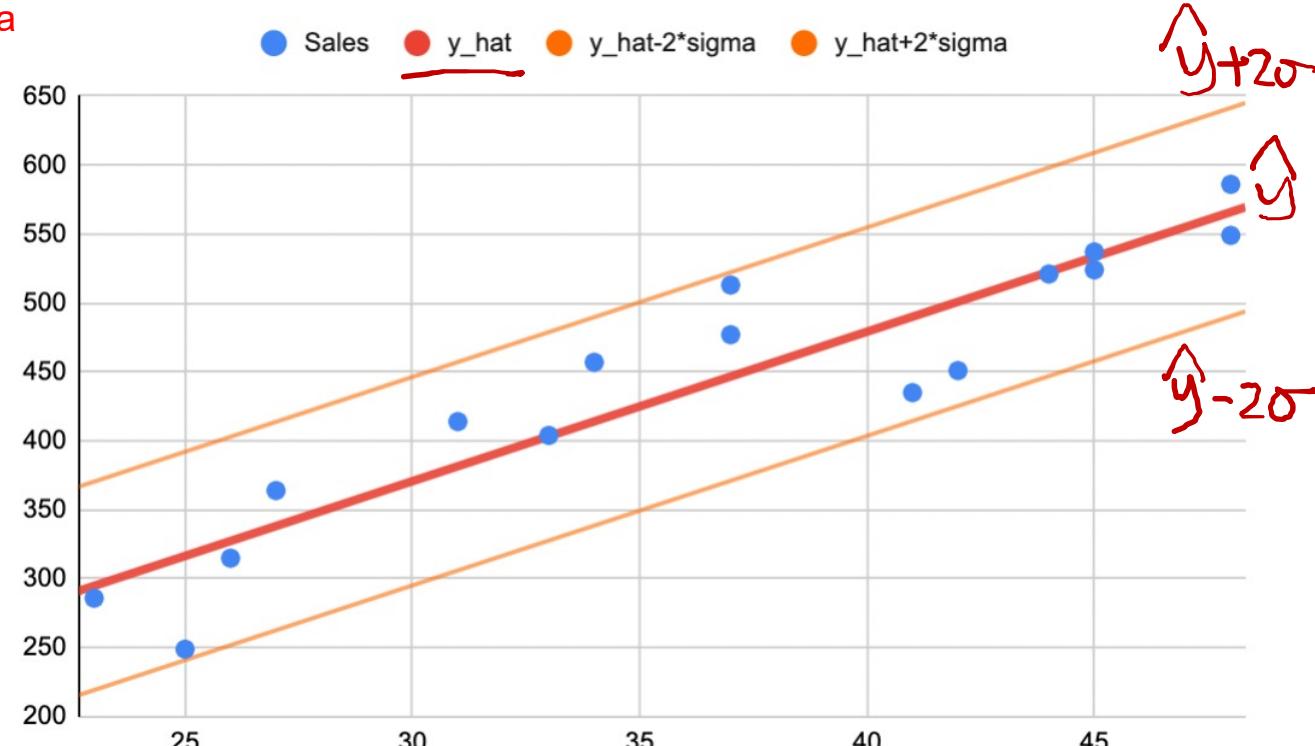
-> kalo misalkan di realisasiinya susah ga bisa di rumusin gini  
pake teori coba coba juga gpp harusnya

- Interpretasi dalam RL:

Dengan tingkat kepercayaan 95%:

- Error model akan selalu terjaga di interval  $(-2\sigma, 2\sigma)$
- Nilai variabel respons terhadap  $x$  akan berada dalam interval  $(\hat{y} - 2\sigma, \hat{y} + 2\sigma)$

$\sigma \leftarrow \text{RMSE}$



## **8.2 Ukuran Kecocokan Model**

# Capaian Pembelajaran

- Mahasiswa mampu mengevaluasi suatu model regresi linear dengan menggunakan ukuran-ukuran yang sesuai & relevan.

# Prasyarat Pembelajaran

- Model dasar Regresi Linear

# Kamus Notasi

Notasi/Istilah (Style 1)	Notasi/Istilah (Style 2)	Interpretasi	Formula
	$S_{xx}$	Jumlah simpangan $x$	$S_{xx} = \sum_{i=1}^n (x^{(i)} - \bar{x})^2$
	$S_{yy}$	Jumlah simpangan $x$	$S_{yy} = \sum_{i=1}^n (y^{(i)} - \bar{y})^2$
	$S_{xy}$	Korelasi	$S_{xy} = \sum_{i=1}^n (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})$
	MSE (Mean Squared Error)	Rata-rata dari ukuran kuadrat jarak	$MSE = \frac{1}{n-2} \sum_{i=1}^n (\hat{y}^{(i)} - y^{(i)})^2$
	RMSE (Root Mean Squared Error)	Akar dari rata-rata dari ukuran kuadrat jarak	$RMSE = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (\hat{y}^{(i)} - y^{(i)})^2}$

# Kamus Notasi

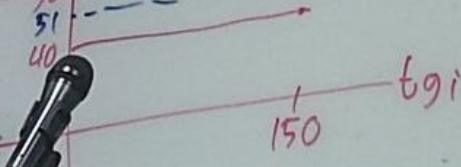
Notasi/Istilah (Style 1)	Notasi/Istilah (Style 2)	Interpretasi	Formula
SSE (Error Sum of Squares)	RSS (Residual Sum of Squares)	<ul style="list-style-type: none"> <li>Jumlah dari semua error kuadrat</li> <li>Error yang <u>diakibatkan oleh kesalahan prediksi.</u> salah kita wkwk</li> </ul>	$SSE = RSS = \sum_{i=1}^n (\hat{y}^{(i)} - y^{(i)})^2$
SSR (Regression Sum of Squares)	ESS (Explained Sum of Squares)	<p>Error yang <u>diakibatkan oleh teknik regresi</u> (yang tak terhindarkan).</p> <p>ini emang salah dari datanya, ga bisa diapa apain</p>	$SSR = ESS = \sum_{i=1}^n (\hat{y}^{(i)} - \bar{y})^2$
SST (Total Sum of Squares)	TSS (Total Sum of Squares)	Error total secara keseluruhan.	$SST = TSS = \sum_{i=1}^n (y^{(i)} - \bar{y})^2$
	R Square ( $R^2$ )	<u>Ukuran kecocokan model regresi.</u>	$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$ $R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}$

$$\beta_1 = \frac{\sum_{x=1}^n (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sum_{x=1}^n (x^{(i)} - \bar{x})^2}$$

$SSE = RSS \rightarrow$  error kkn regresi

$SSR = ESS \rightarrow$  error bkn kkn regresi

$$SST = TSS \text{ brt}$$



$$R^2 = 1 - \frac{RSS}{TSS}$$

model baik:  
 $\rightarrow ESS$  tinggi  
 $\rightarrow RSS$  rendah

$\rightarrow R^2$  tinggi

$$= \frac{TSS - RSS}{TSS}$$

$$= \frac{ESS}{TSS}$$



# SSE = RSS

= Cost Function

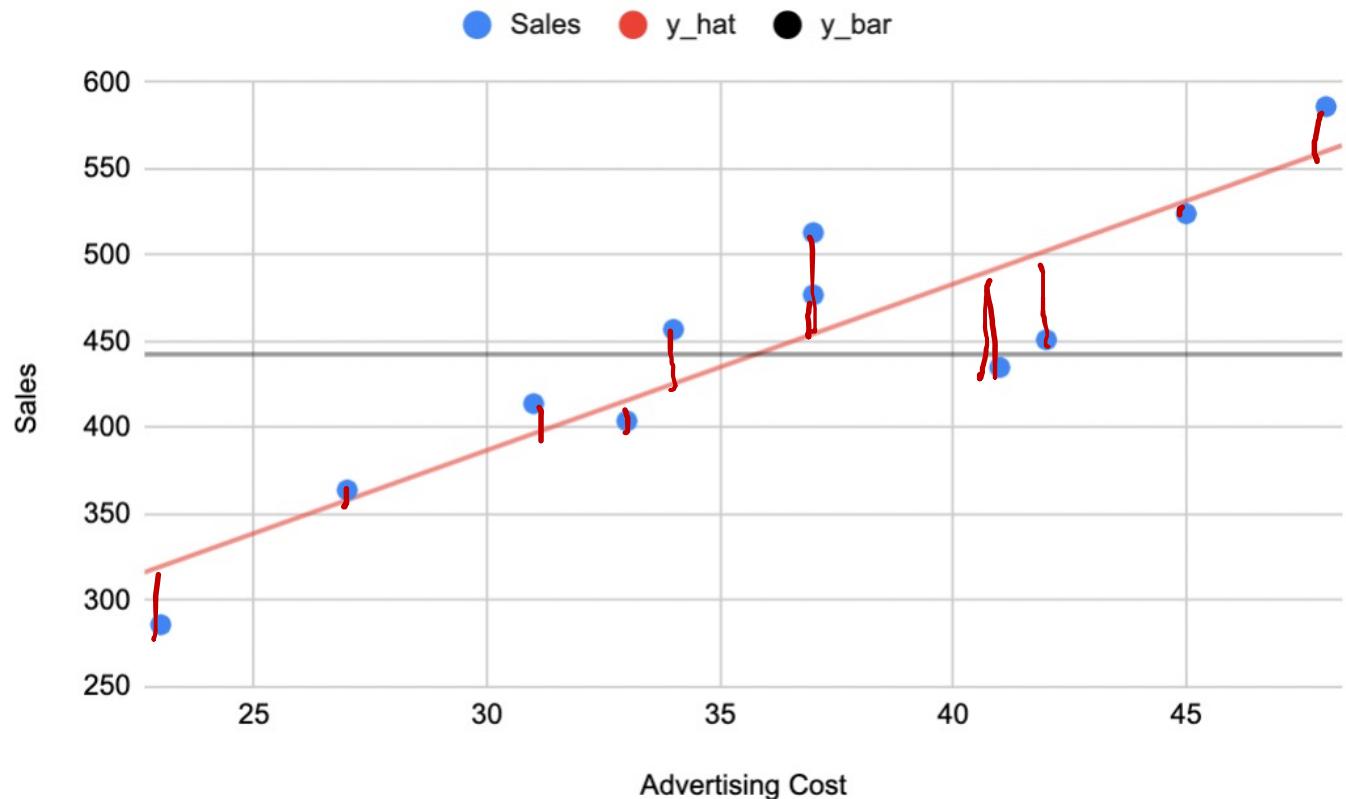
-> yang diusahakan dibenerin dari regresion kita

- Sum of Square Error = Regression Sum of Square
- Merupakan simpangan antara model dan data.

cost function

$$SSE = RSS = \sum_{i=1}^n (\hat{y}^{(i)} - y^{(i)})^2$$

*model    data*



# **SSR = ESS**

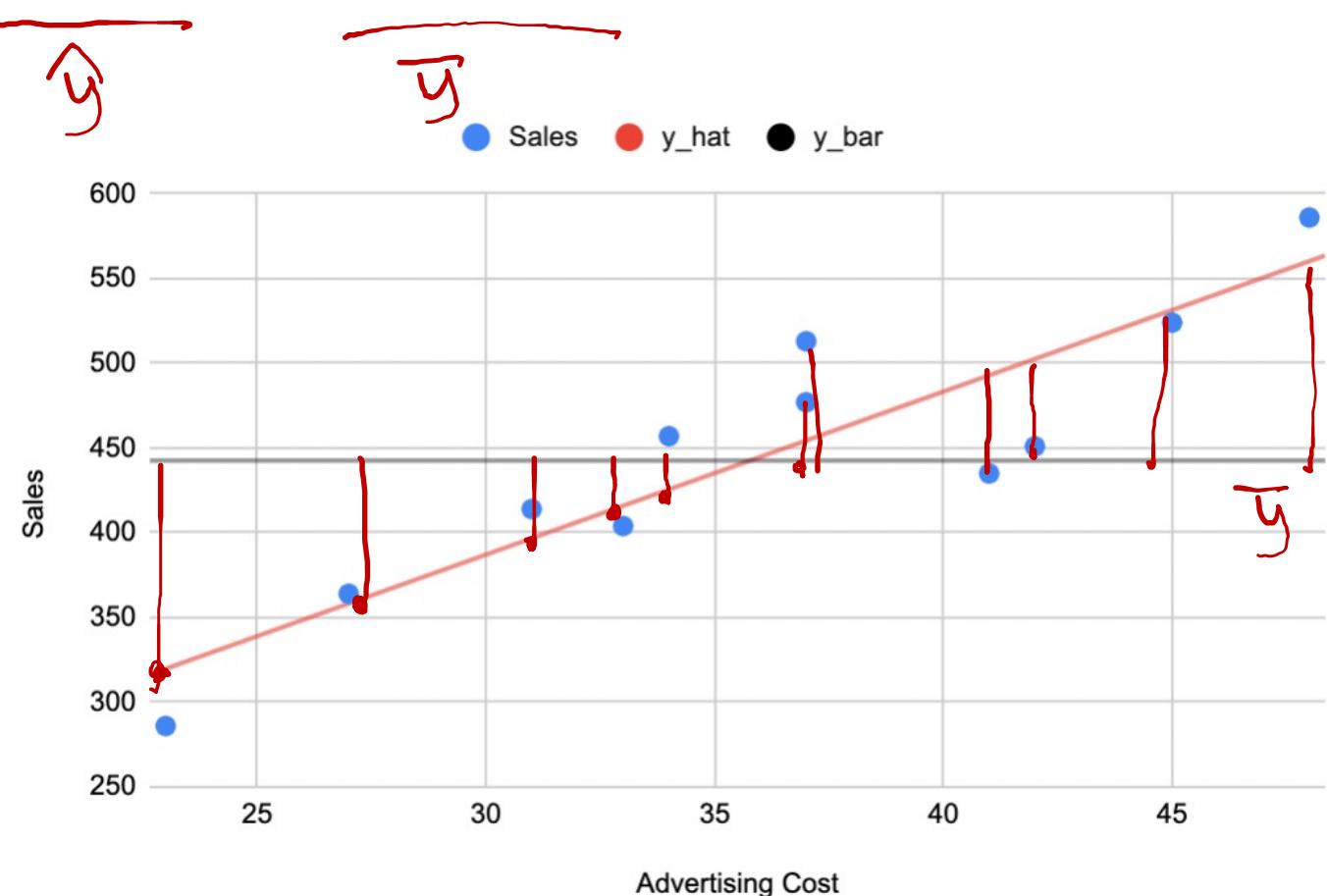
= salah dari sananya

-> there's nothing we can do about it kalo ini tjd aowkawok

- Sum of Square Regression = Explained Sum of Square
- Merupakan simpangan antara model dan rata-rata.

$$SSR = ESS = \sum_{i=1}^n (\hat{y}^{(i)} - \bar{y})^2$$

*model*      *rata-rata*

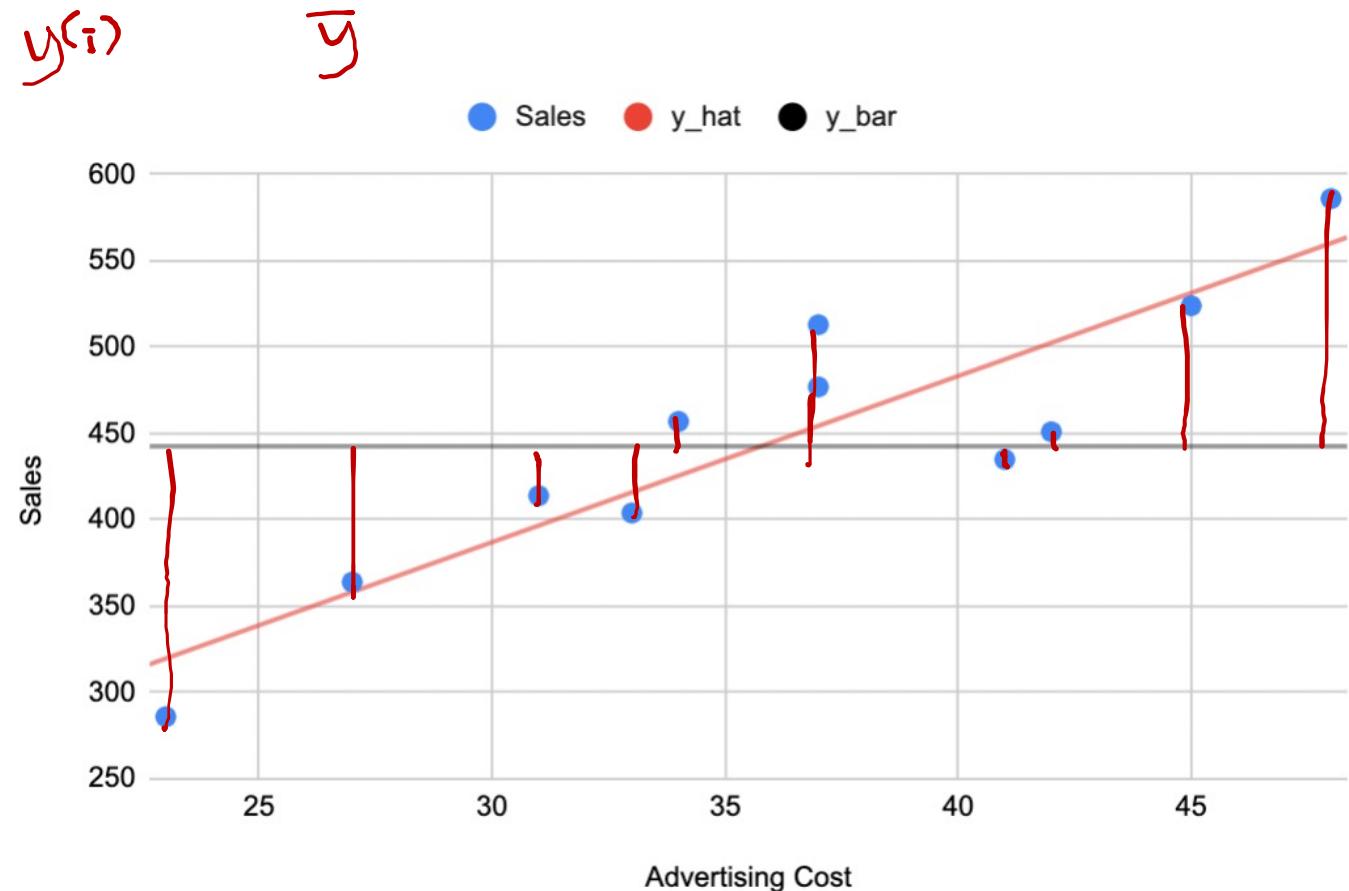


# **SST = TSS**

- Sum of Square Total = Total Sum of Square
- Merupakan simpangan antara **data** dan rata-rata.

$$SST = TSS = \sum_{i=1}^n (y^{(i)} - \bar{y})^2$$

→ Varians dari  $y$



# Teorema Ukuran Kecocokan Linear

- Untuk sembarang dataset  $\{(x^{(i)}, y^{(i)})\}$  dan model regresi linear  $\hat{y} = \beta_0 + \beta_1 x$ , berlaku:
  - $SSE + SSR = SST$
  - $RSS + ESS = TSS$

$$\sum_{i=1}^n (\hat{y}^{(i)} - y^{(i)})^2 + \sum_{i=1}^n (\hat{y}^{(i)} - \bar{y})^2 = \sum_{i=1}^n (y^{(i)} - \bar{y})^2$$

# Koefisien Determinasi R<sup>2</sup>

metrik performa untuk model Linear Regression

- Adalah suatu metrik/ukuran performa suatu model RL.

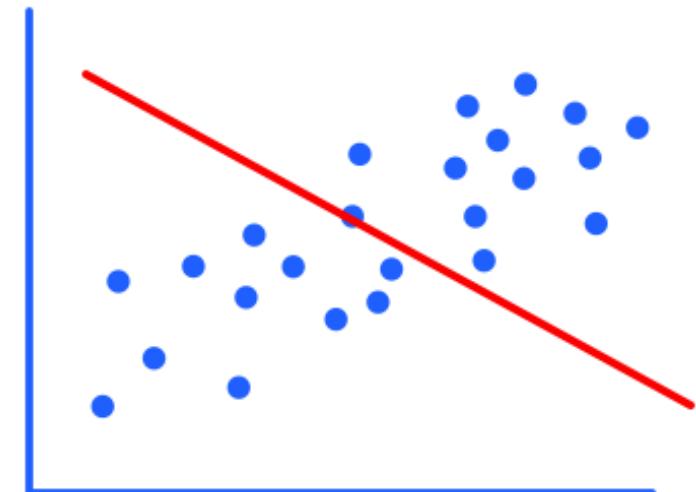
- Definisi:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

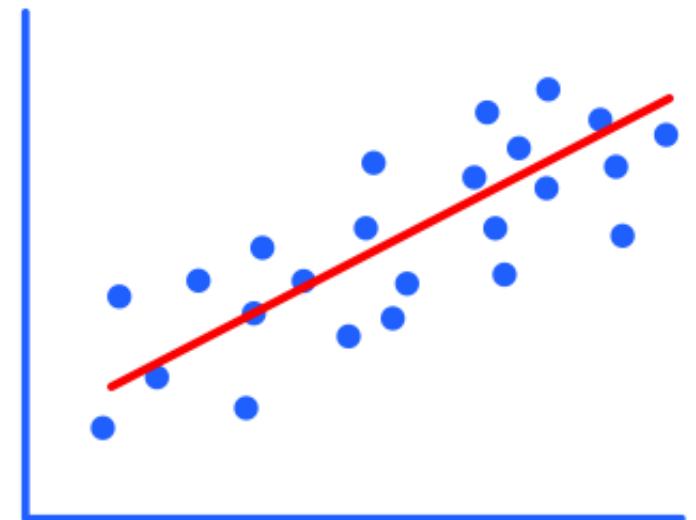
- Nilai R<sup>2</sup> selalu berkisar pada interval [0,1]

- Interpretasi:

- R<sup>2</sup> makin dekat dengan 0, model semakin buruk.
- R<sup>2</sup> makin dekat dengan 1, model semakin baik.



R<sup>2</sup> rendah



R<sup>2</sup> tinggi

# Koefisien Korelasi $\rho$

koef korelasi ini yang bisa dikasih tingkat kepercayaan juga

- Adalah suatu metrik/ukuran performa suatu model RL.

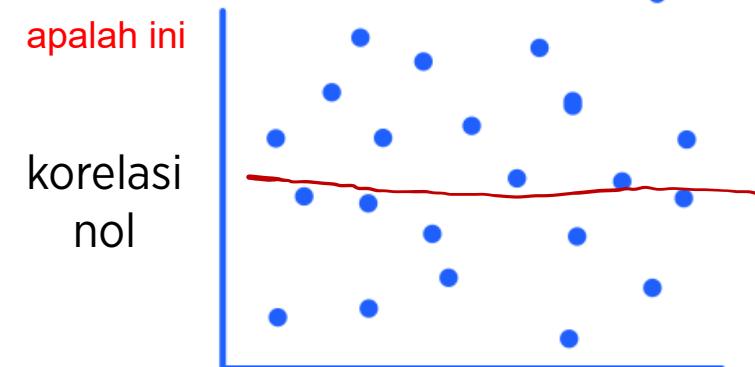
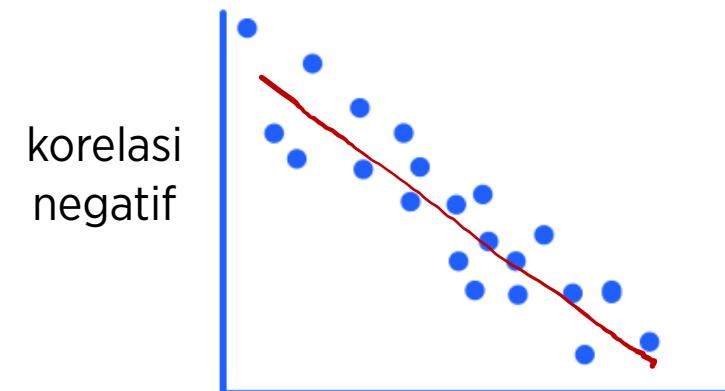
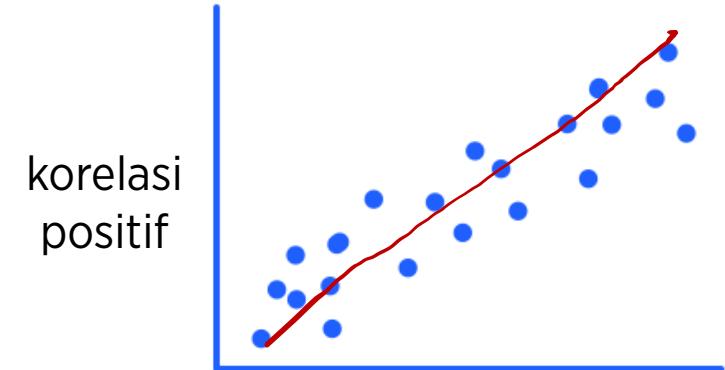
- Definisi:

$$\rho = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

- Nilai  $\rho$  selalu berkisar pada interval  $[-1,1]$

- Interpretasi:

- $\rho$  makin dekat dengan 1, terjadi korelasi positif.
- $\rho$  makin dekat dengan -1, terjadi korelasi negatif.
- $\rho$  makin dekat dengan 0, tidak terjadi korelasi.



# **8.3 Transformasi Data untuk Regresi Linear**

# **Capaian Pembelajaran**

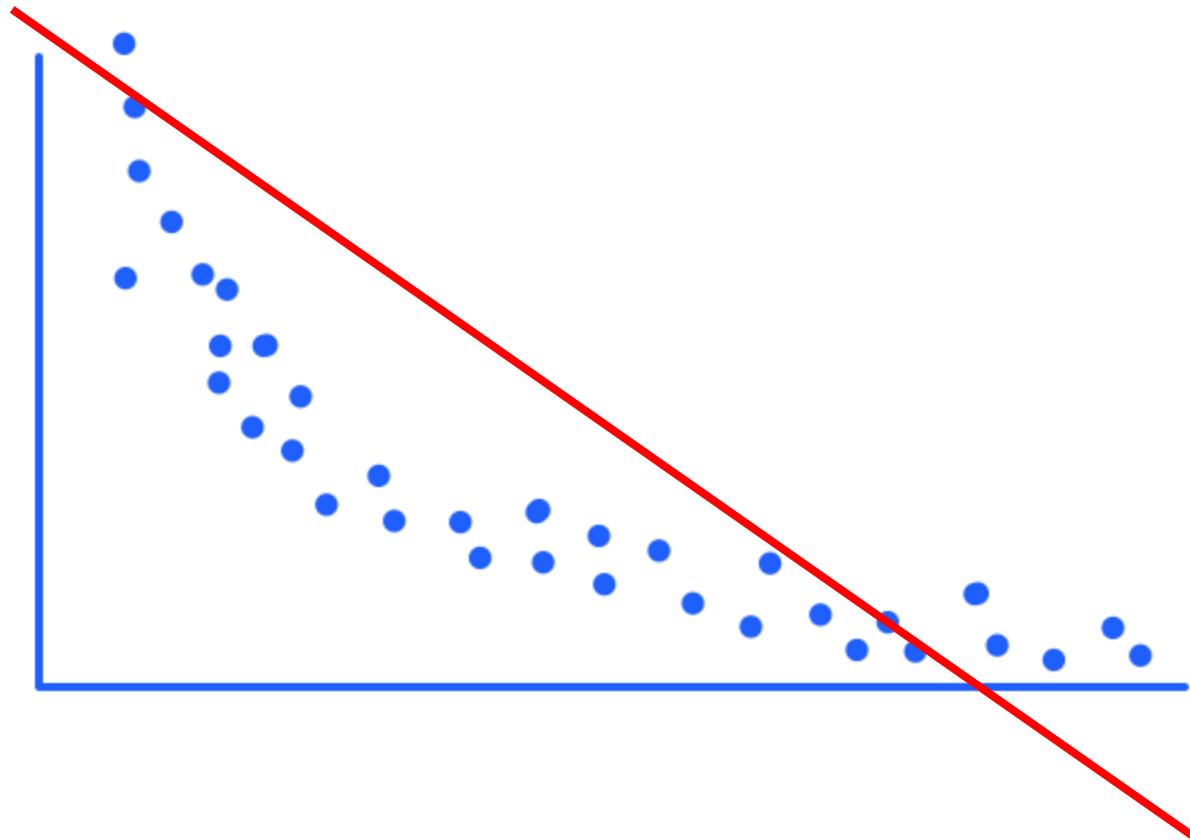
- Mahasiswa mampu mengevaluasi suatu model regresi linear dengan menggunakan ukuran-ukuran yang sesuai & relevan.

# **Prasyarat Pembelajaran**

- Model dasar Regresi Linear

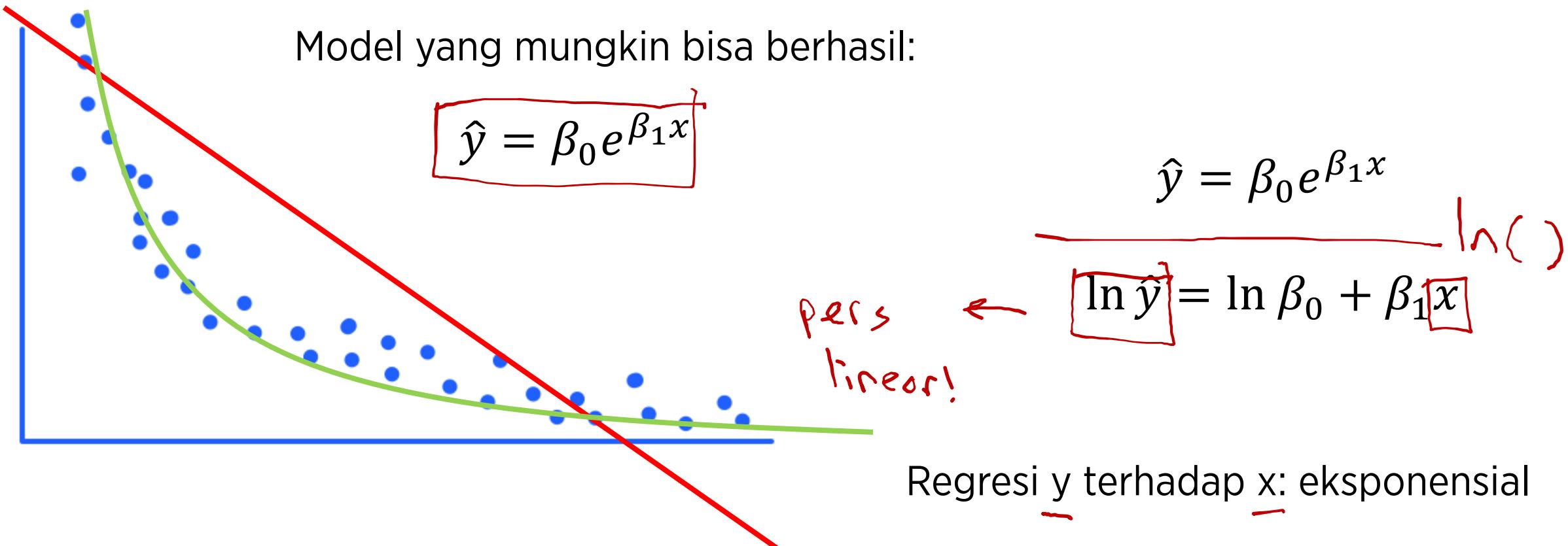
# Data Tidak Linear

kalo ga boleh pake regression lain selain linear kita bisa transform dikit datanya biar bisa di lin reg



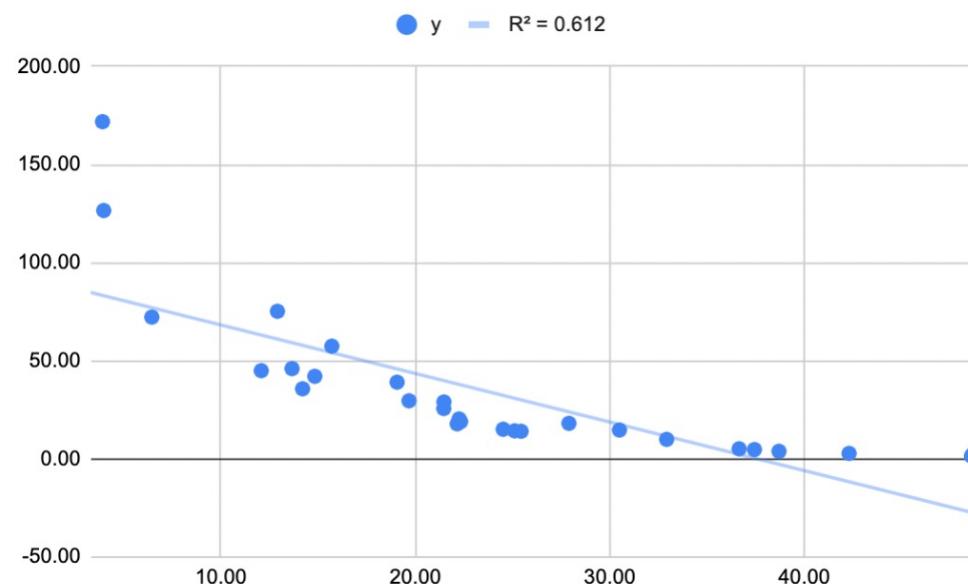
- Gambar di samping merupakan ‘hasil terbaik’ dari suatu model Regresi Linear.
- Tentunya, model tersebut tidak cukup baik untuk melakukan prediksi.
- Solusi: Lakukan transformasi!

# Transformasi yang Melinearkan

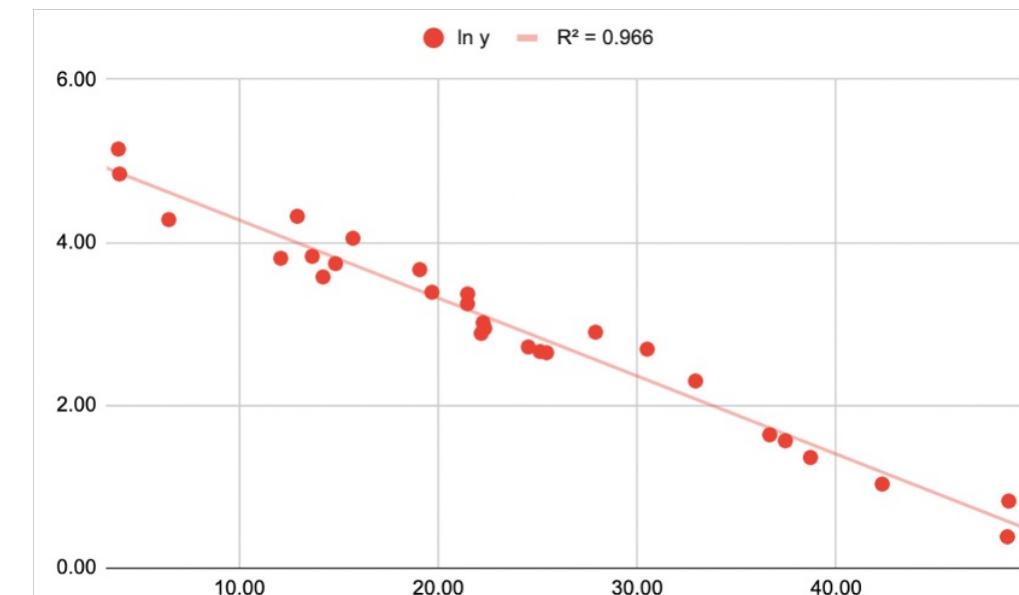


# Contoh Penerapan

x	y	ln y
48.67	2.28	0.82
36.66	5.15	1.64
21.47	25.70	3.25
19.68	29.63	3.39
30.50	14.73	2.69
37.45	4.78	1.57
6.45	72.27	4.28
24.52	15.13	2.72
27.90	18.14	2.90
13.66	46.03	3.83
15.71	57.46	4.05
38.71	3.90	1.36
12.91	75.24	4.32
48.61	1.47	0.39
14.84	42.11	3.74
42.31	2.81	1.03
21.48	29.03	3.37
3.92	171.88	5.15
22.26	20.38	3.01
22.16	17.88	2.88
32.93	9.96	2.30
25.45	14.09	2.65
25.12	14.31	2.66
14.21	35.75	3.58
19.07	39.11	3.67
3.98	126.60	4.84
22.33	19.03	2.95
12.08	44.98	3.81



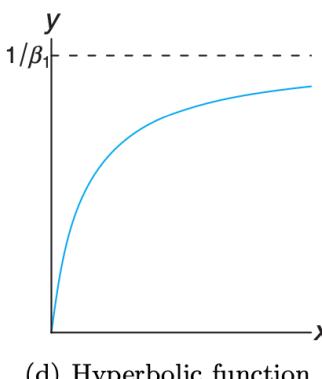
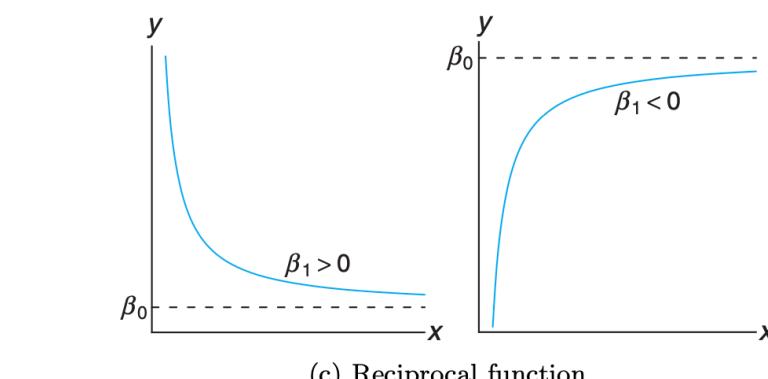
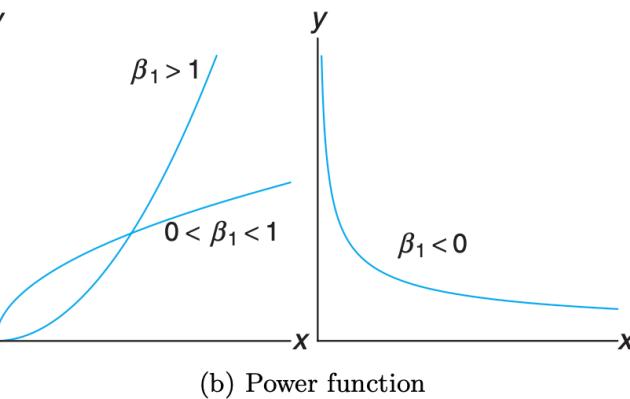
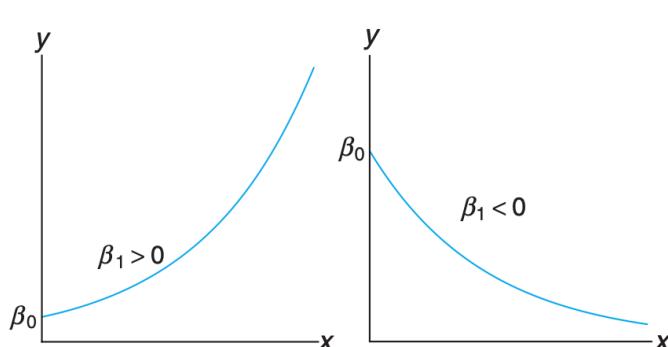
Regrasi linear y thd x



Regrasi linear  $\ln y$  thd x

# Kumpulan transformasi dasar GLM

Functional Form Relating $y$ to $x$	Proper Transformation	Form of Simple Linear Regression
Exponential: $y = \beta_0 e^{\beta_1 x}$	$y^* = \ln y$	Regress $y^*$ against $x$
Power: $y = \beta_0 x^{\beta_1}$	$y^* = \log y; \quad x^* = \log x$	Regress $y^*$ against $x^*$
Reciprocal: $y = \beta_0 + \beta_1 \left(\frac{1}{x}\right)$	$x^* = \frac{1}{x}$	Regress $y$ against $x^*$
Hyperbolic: $y = \frac{x}{\beta_0 + \beta_1 x}$	$y^* = \frac{1}{y}; \quad x^* = \frac{1}{x}$	Regress $y^*$ against $x^*$



Langkah-langkah melakukan transformasi **General Linear Model**:

1. Buat scatterplot dari data.
2. Dari Scatter Plot, buat tebakan mengenai relasi antara  $x$  dan  $y$ .
3. Lakukan transformasi yang sesuai.
4. Buat regresi linear sesuai variabel hasil transformasi.

# Tuhan Memberkati

