Metode Statistika

Analisis Korelasi dan Regresi

Dr. Kusman Sadik, M.Si

Dept. Statistika IPB - 2015

Hubungan Dua Peubah atau Lebih

PEUBAH	KASUS	PENGUMPULAN DATA	JENIS HUBUNGANNYA
1. Dosis pupuk 2. Banyaknya padi yg di-hasilkan /ha	Diduga dosis pupuk mempengaruhi banyaknya padi yg dihasilkan/ha	Dosis pupuk ditentukan dahulu, faktor-faktor lain yg mempengaruhi banyaknya padi diken-dalikan sehingga pengaruhnya konstan, ke-mudian diamati banyaknya padi yg dihasilkan	Perubahan banyaknya padi yg dihasilkan/ha dipengaruhi oleh perubahan dosis pupuk → HUB SEBAB AKIBAT
1. Tinggi badan 2. Berat badan	Diduga tinggi badan dan berat badan memiliki hubungan	Dimulai dengan mengamati tinggi badan da-hulu, disusul mengamati peubah yg dianggap relevan (berat badan), atau sebaliknya.	Pengamatan thdp kedua peubah dilakukan secara bersamaan. Sulit untuk mengatakan bahwa peru-bahan satu peubah disebabkan oleh perubahan peubah lainnya → bukan HUB SEBAB AKIBAT Ingin diketahui kekuatan dan arah hubungannya

Hubungan Dua Peubah atau Lebih (2)

PEUBAH	KASUS	PENGUMPULAN DATA	JENIS HUBUNGANNYA
1. Banyaknya barang terju-al/minggu 2. Adanya hari libur/tidak 3. Harga barang	Diduga banyaknya ba-rang terjual/minggu dipe-ngaruhi oleh berbagai peubah, misalnya harga barang, ada/ tidaknya hari libur dlm minggu tsb	Harga barang ditentukan lebih dahulu, faktor-faktor lain yg mempengaruhi banyaknya barang terjual dikendalikan sehingga pengaruhnya konstan, kemudian diamati banyaknya barang yg terjual pada minggu ada hari libur dan minggu tanpa hari libur	Perubahan banyaknya barang yg terjual dipengaruhi oleh perubahan harga dan ada/tidaknya hari libur → Hub SEBAB AKIBAT
1. Bobot badan 2. Bobot jantung	Diduga bobot badan dan bobot jantung memiliki hubungan	Dimulai dengan mengamati bobot badan terlebih dahulu, segera disusul mengamati peubah yg dianggap relevan (dalam hal ini bobot jantung), atau sebaliknya.	Pengamatan thdp kedua peubah dilakukan secara bersamaan. Sulit untuk mengatakan bahwa perubahan satu peubah disebabkan oleh peubah lainnya.→ bukan SEBAB AKIBAT. Ingin diketahui model matematisnya (HUB KUANTITATIF)

Contoh Kasus Lain

- Umur vs tinggi tanaman
- Biaya promosi vs volume penjualan
- Produktivitas pertanian vs (tanaman bahan pangan, tanaman perkebunan rakyat, peternakan dan perikanan)
- Produksi padi vs luas lahan sawah
- Tinggi badan vs berat badan
- Bobot badan vs bobot jantung

Analisis Hubungan

1. Jenis/tipe hubungan

3. Ukuran Keterkaitan

2. Skala pengukuran variabel

4. Pemodelan Keterkaitan

Relationship vs Causal Relationship

- Tidak semua hubungan (*relationship*) berupa hubungan sebab-akibat (*causal relationship*).
- Penentuan suatu hubungan bersifat sebab-akibat memerlukan pendapat/pengetahuan dari bidang ilmu terkait.

Alat Analisis Keterkaitan/Hubungan

- Ditentukan oleh:
  1. Skala pengukuran data/variabel
  2. Jenis hubungan antar variabel

Relationship	Numerik	Kategorik
Numerik	Korelasi Pearson, Spearman	Tabel Ringkasan
Kategorik	Tabel Ringkasan	Spearman (ordinal), Chi Square

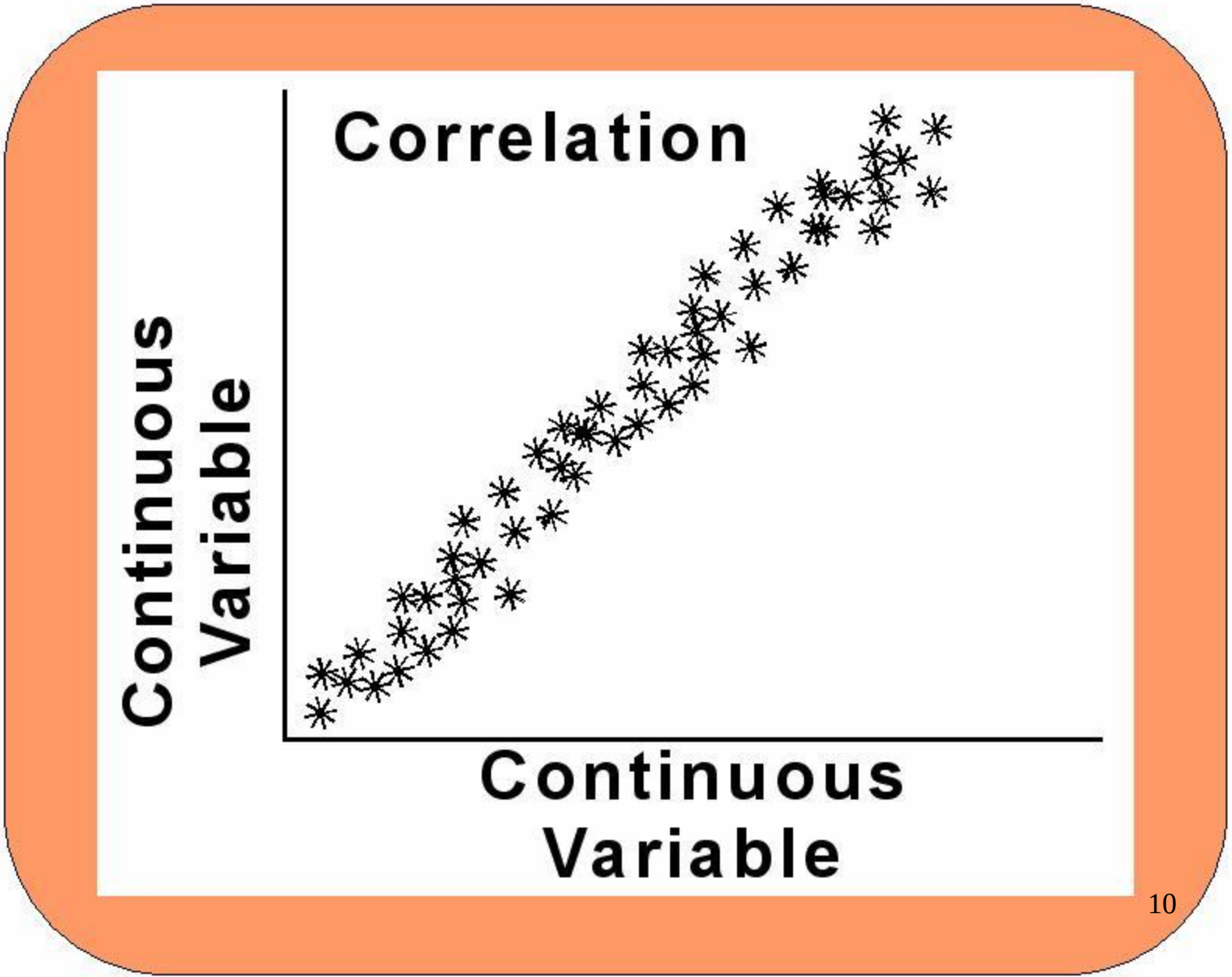
Hubungan Sebab Akibat

(Causal Relationship)

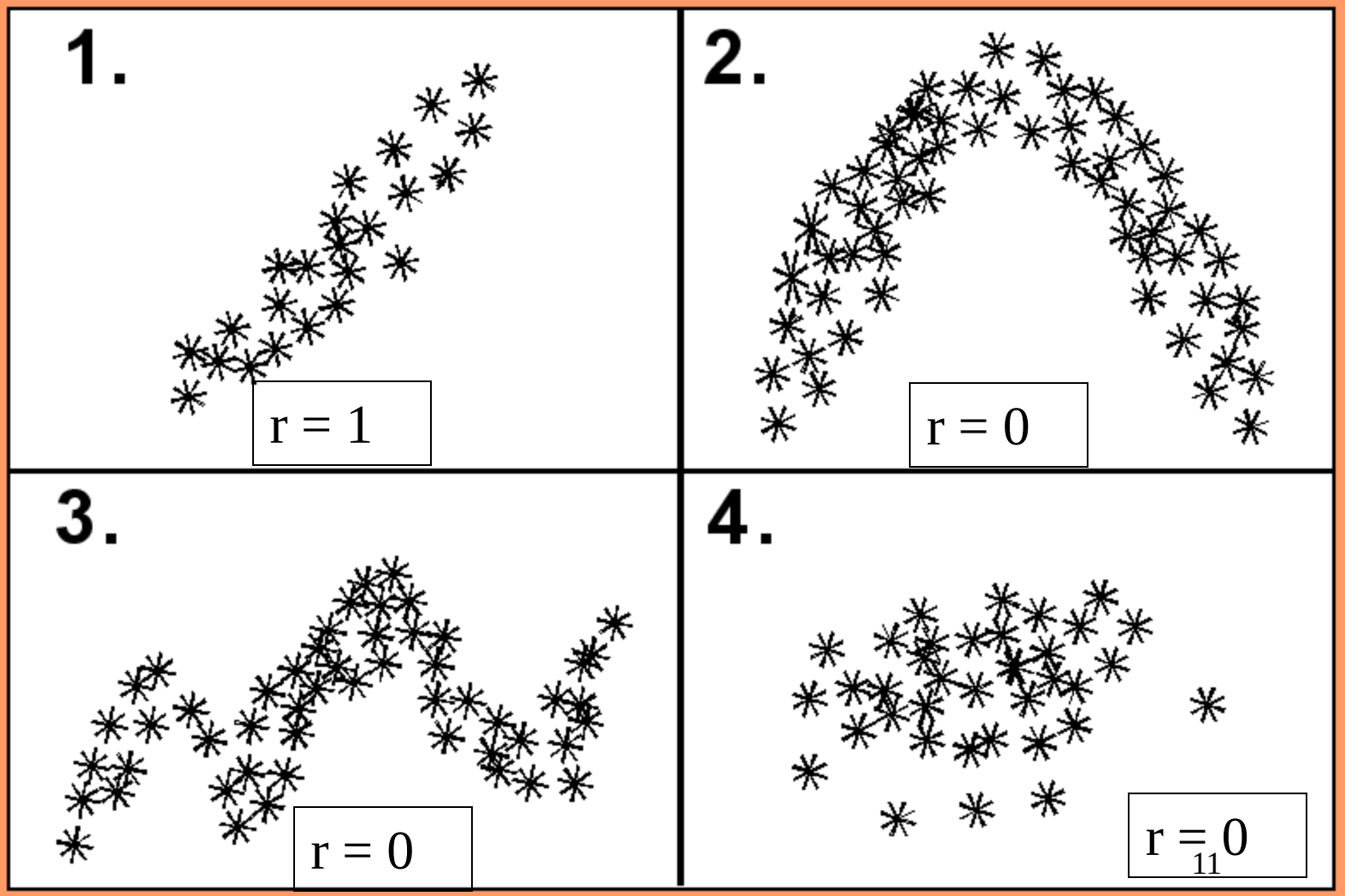
Variabel Y	Variabel X	
	Numerik	Kategorik
Numerik	Regresi Linier	Regresi Linier
Kategorik	Regresi Logistik	Regresi Logistik

Korelasi (r)

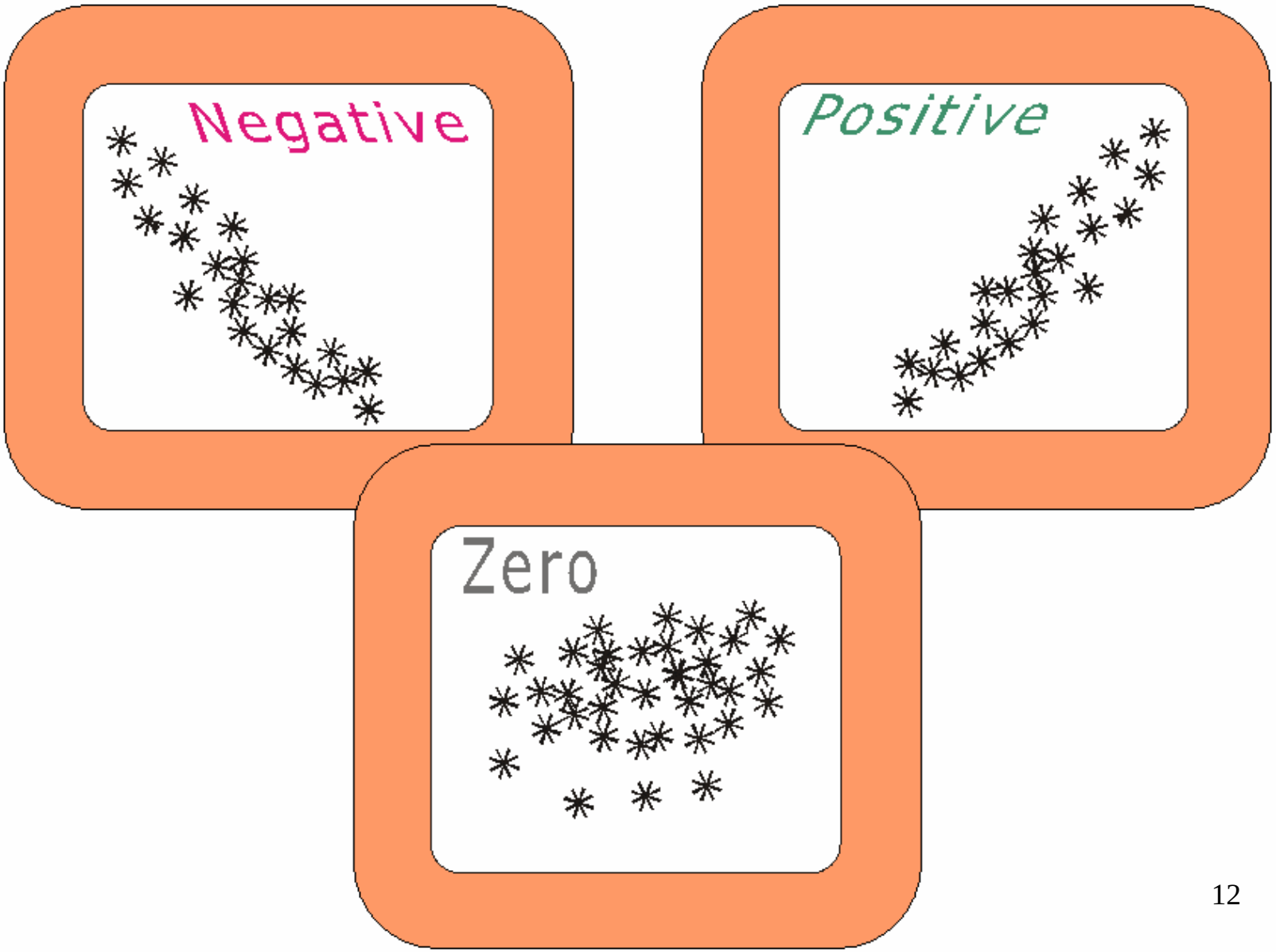
Overview



Relationships between Continuous Variables



Korelasi

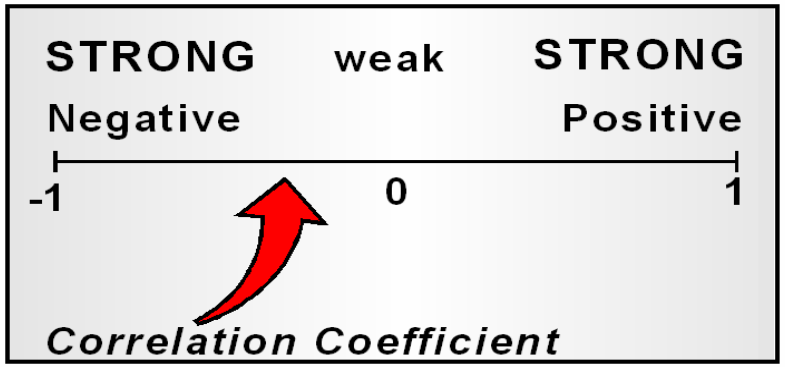


Koefisien Korelasi (r)

- Tidak menggambarkan hubungan sebab akibat
- Nilainya berkisar antara -1 dan 1
- Tanda (+) atau (-) → arah hubungan
  - (+) searah;
  - (-) beralawanan arah
- Koefisien Korelasi Pearson → hubungan linier
- Koefisien Korelasi Spearman (rank correlation) → trend relationship

Koefisien Korelasi Pearson (r)

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



$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$S_{xx} = \sum (x_i - \bar{x})^2 \text{ dan } S_{yy} = \sum (y_i - \bar{y})^2$$

Notasi lain:

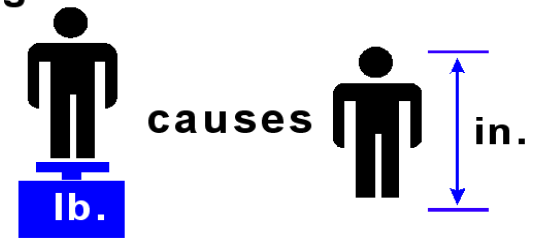
$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}, \quad S_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

$$S_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}.$$

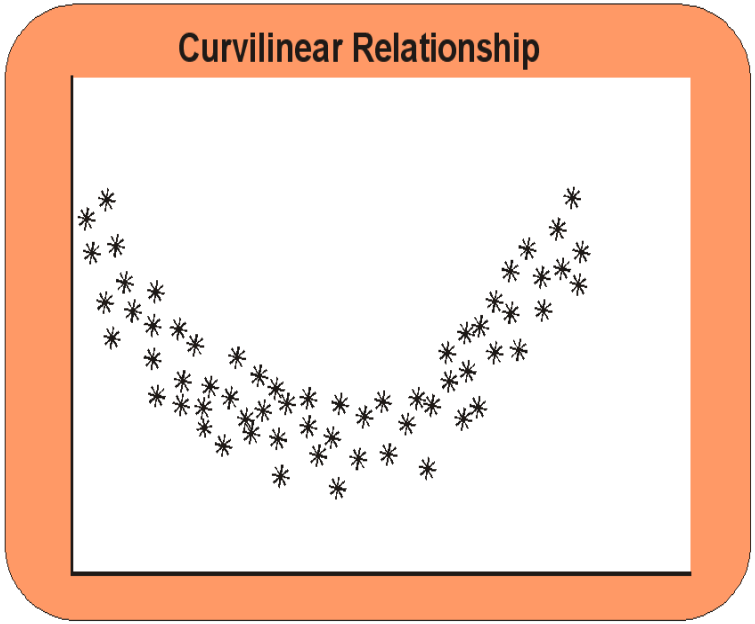
Korelasi !!!

Misuses of the Correlation Coefficient

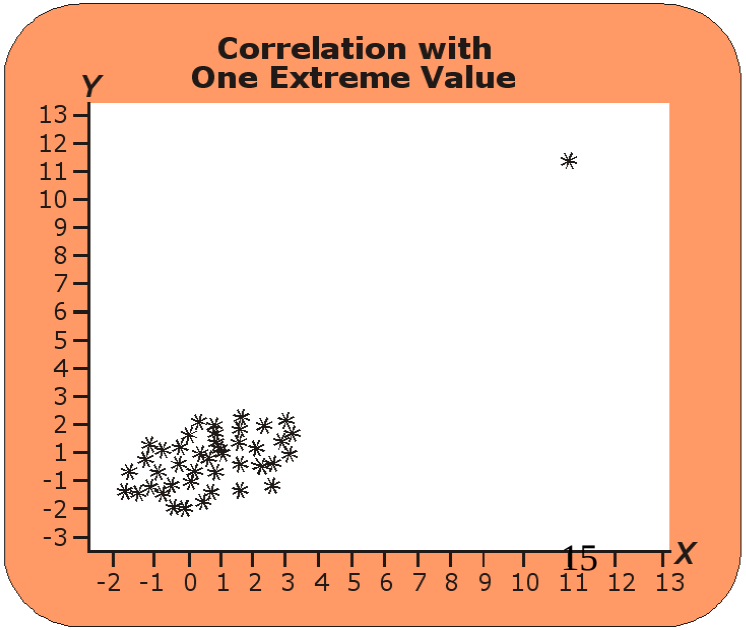
Strong correlation does not mean



Missing Another Type of Relationship



Extreme Data Values



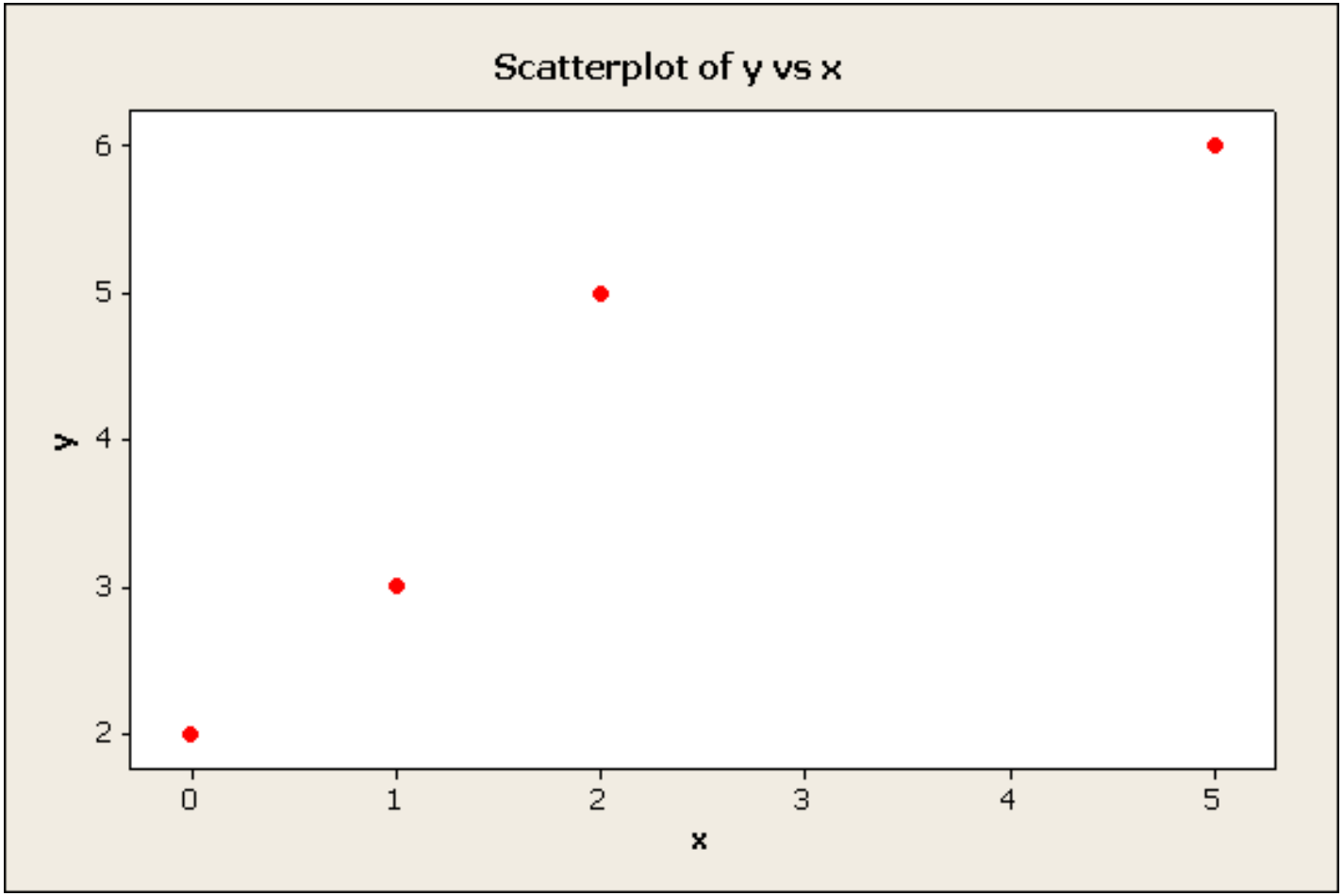
Contoh

Diketahui data pengeluaran iklan (x milyar) dengan total profit penjualan suatu produk komputer perbulan (y milyar) selama 4 bulan sebagai berikut:

x	:	2	1	5	0
y	:	5	3	6	2

a. Buat scatter plot untuk data tersebut.
b. Hitung koefisien korelasinya.

(a) Scatter Plot : x dengan y



(b) Koefien Korelasi (r)

	x	y	x^2	y^2	xy
	2	5	4	25	10
	1	3	1	9	3
	5	6	25	36	30
	0	2	0	4	0
Total	8	16	30	74	43
	Σx	Σy	Σx^2	Σy^2	Σxy

$$r = \frac{43 - \frac{8 \times 16}{4}}{\sqrt{30 - \frac{8^2}{4}}\sqrt{74 - \frac{(16)^2}{4}}} = .930$$

Mendenhall : Example 12.7, hlm. 534

The heights and weights of $n = 10$ offensive backfield football players are randomly selected from a county's football all-stars. Calculate the correlation coefficient for the heights (in inches) and weights (in pounds) given in Table 12.4.

Heights and Weights of $n = 10$ Backfield All-Stars

Player	Height, x	Weight, y
1	73	185
2	71	175
3	75	200
4	72	210
5	72	190
6	75	195
7	67	150
8	69	170
9	71	180
10	69	175

Mendenhall : Example 12.7, hlm. 534

Solution You should use the appropriate data entry method of your scientific calculator to verify the calculations for the sums of squares and cross-products,

$$S_{xy} = 328 \qquad S_{xx} = 60.4 \qquad S_{yy} = 2610$$

using the calculational formulas given earlier in this chapter. Then

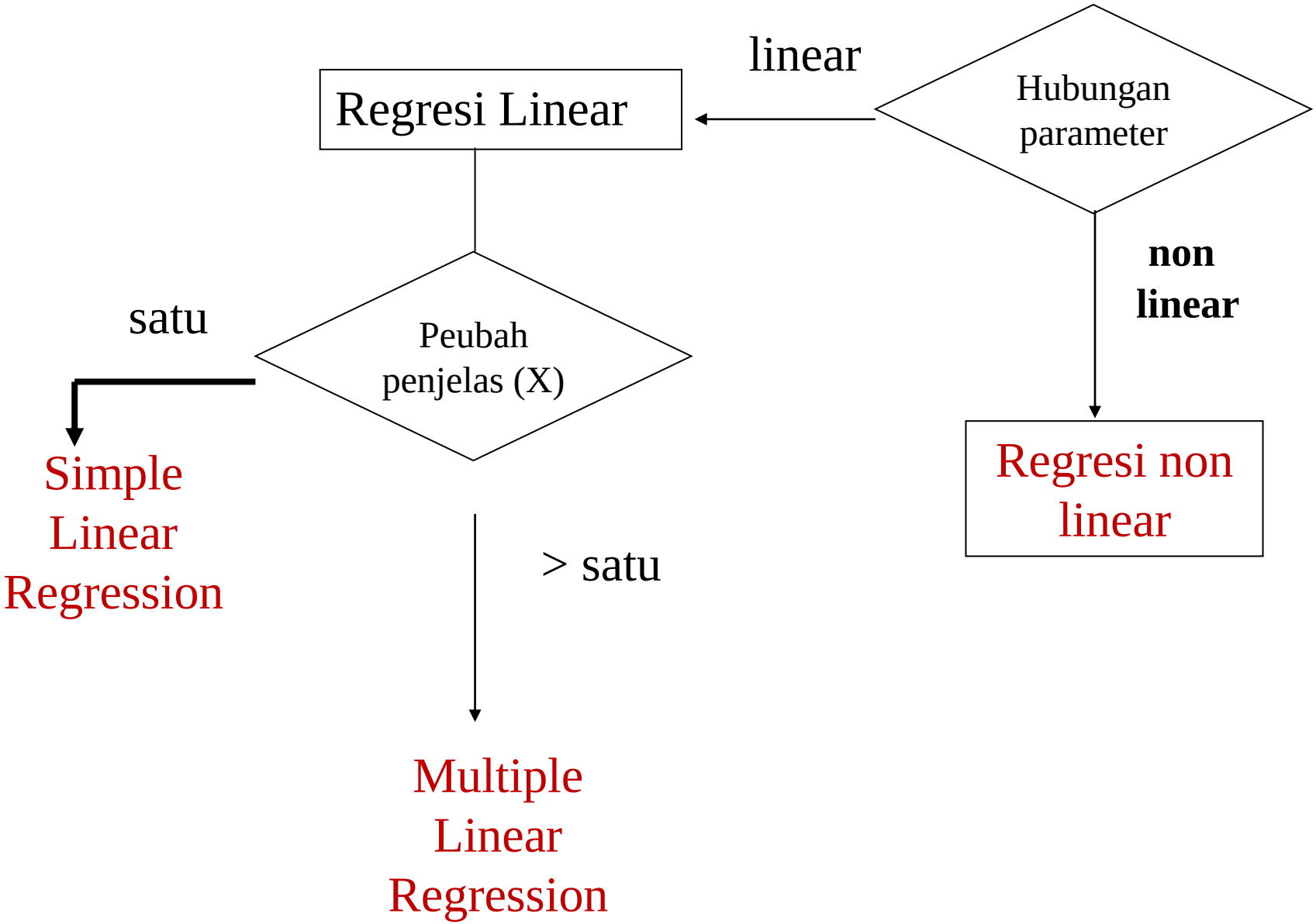
$$r = \frac{328}{\sqrt{(60.4)(2610)}} = .8261$$

or $r = .83$. This value of r is fairly close to 1, the largest possible value of r, which indicates a fairly strong positive linear relationship between height and weight.

Analisis Regresi

Definisi

- Linier (*linear*) : linier dalam parameter
- Sederhana (*simple*) : hanya satu peubah penjelas (x)
- Berganda (*multiple*) : lebih dari satu peubah penjelas (x)


Hubungan parameter
linear
Regresi Linear
non linear
Regresi non linear
Peubah penjelas (X)
satu
Simple Linear Regression
> satu
Multiple Linear Regression

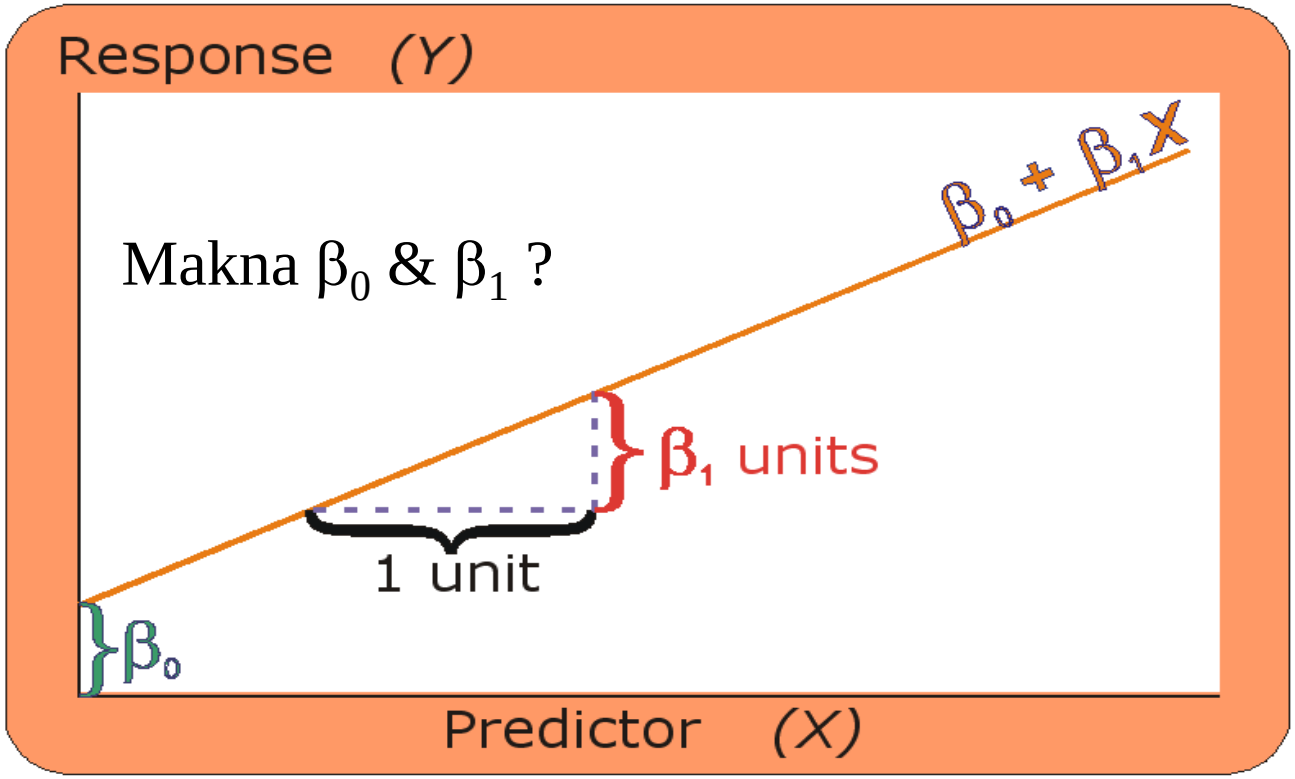
ANALISIS REGRESI

- **Hubungan Antar Peubah:**
  - Fungsional (deterministik) → Y=f(X) ; misalnya: Y=10X
  - Statistik (stokastik) → amatan tidak jatuh pas pada kurva
  - Misal: IQ vs Prestasi, Berat vs Tinggi, Dosis Pupuk vs Produksi, Profit vs Biaya Iklan

- **Model regresi linear sederhana:**

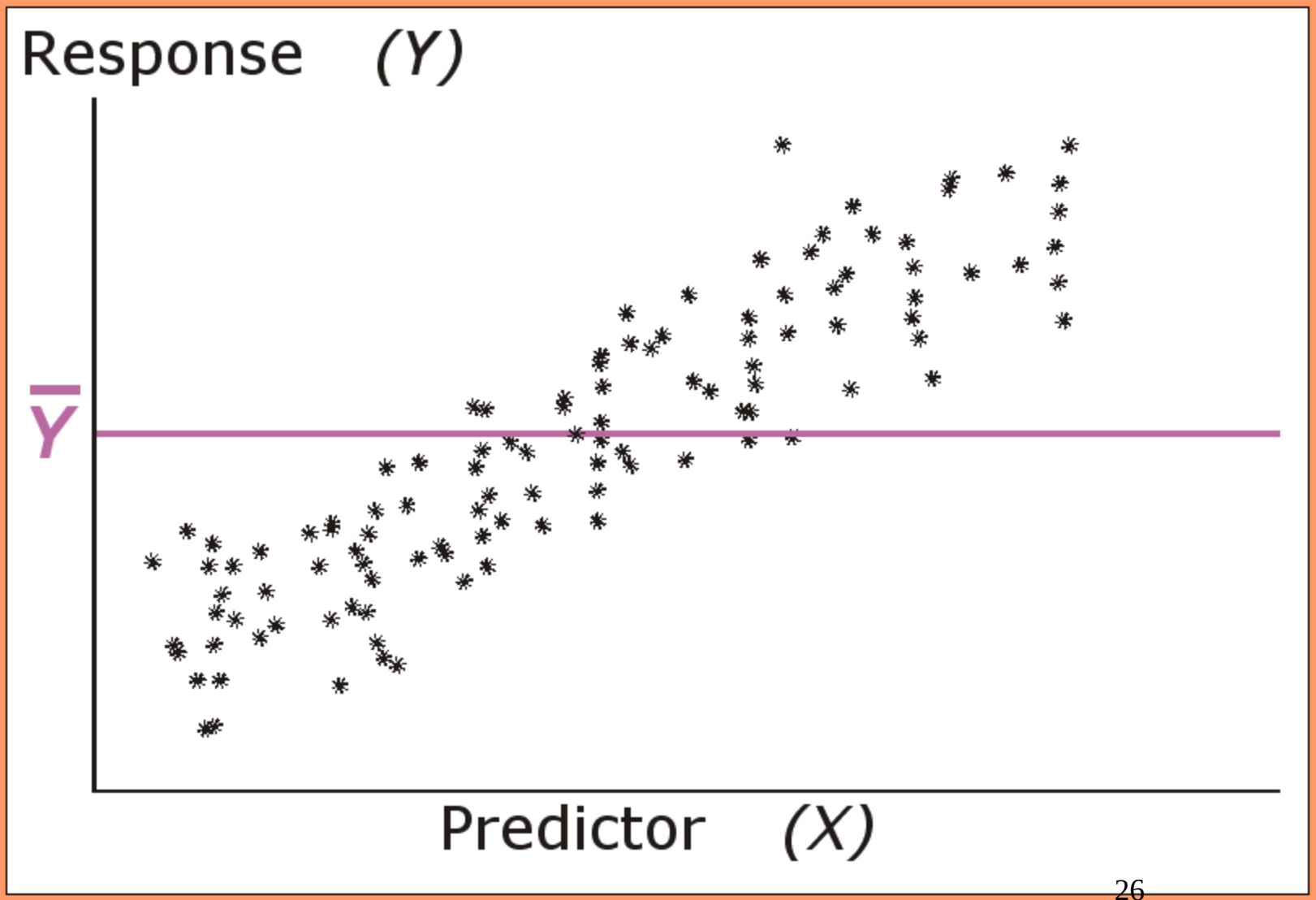
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \;\; ; i = 1,2,...,n$$

Simple Linear Regression Model



Interpretasi : β_0 adalah nilai Y ketika X = 0, sedangkan β_1 adalah perubahan nilai Y untuk setiap perubahan X sebesar satu satuan unit.

The Baseline Model



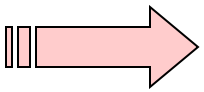
Analisis Regresi

- Pendugaan terhadap koefisien regresi:

  → b_0 penduga bagi β_0 dan b_1 penduga bagi β_1

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$b_0 = \bar{y} - b_1\bar{x}$$



Metode Kuadrat Terkecil (Least Square)

Analisis Regresi

Bagaimana Pengujian terhadap model regresi ??

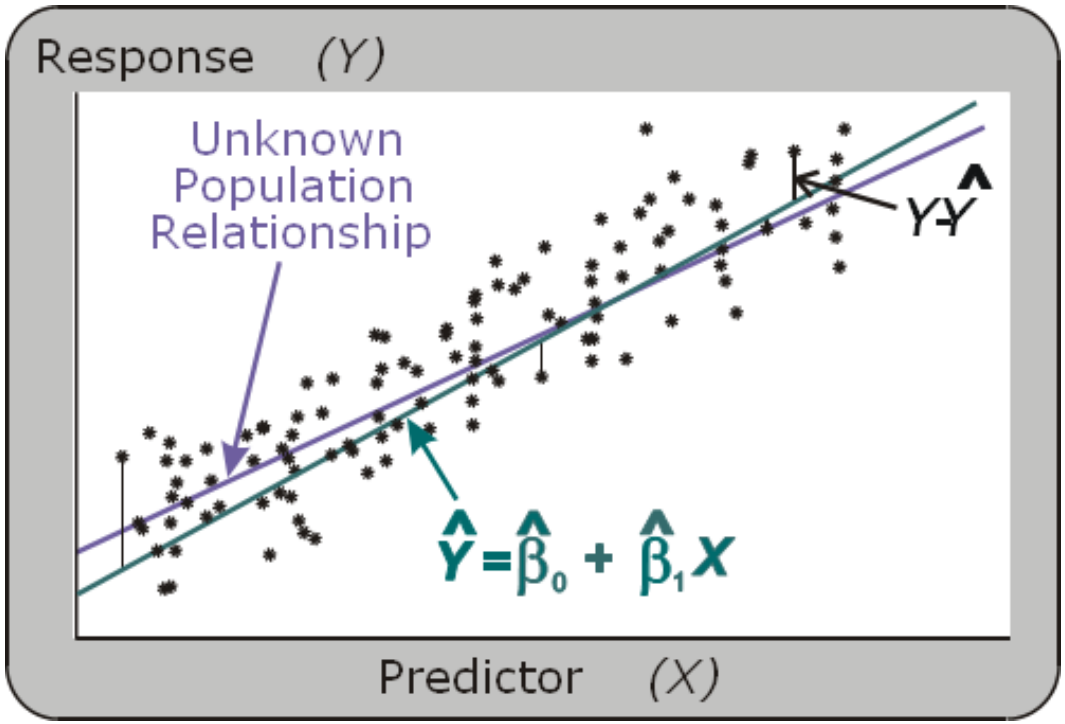
- parsial (per koefisien) → uji-t
- bersama → uji-F (Anova)

Bagaimana menilai kesesuaian model ??

- R^2 (Koefisien Determinasi: persentase keragaman Y yang mampu dijelaskan oleh X)

Metoda Kuadrat Terkecil

- Pendugaan parameter pada regresi didapat dengan meminimumkan jumlah kuadrat galat (*error*).



Metoda Kuadrat Terkecil

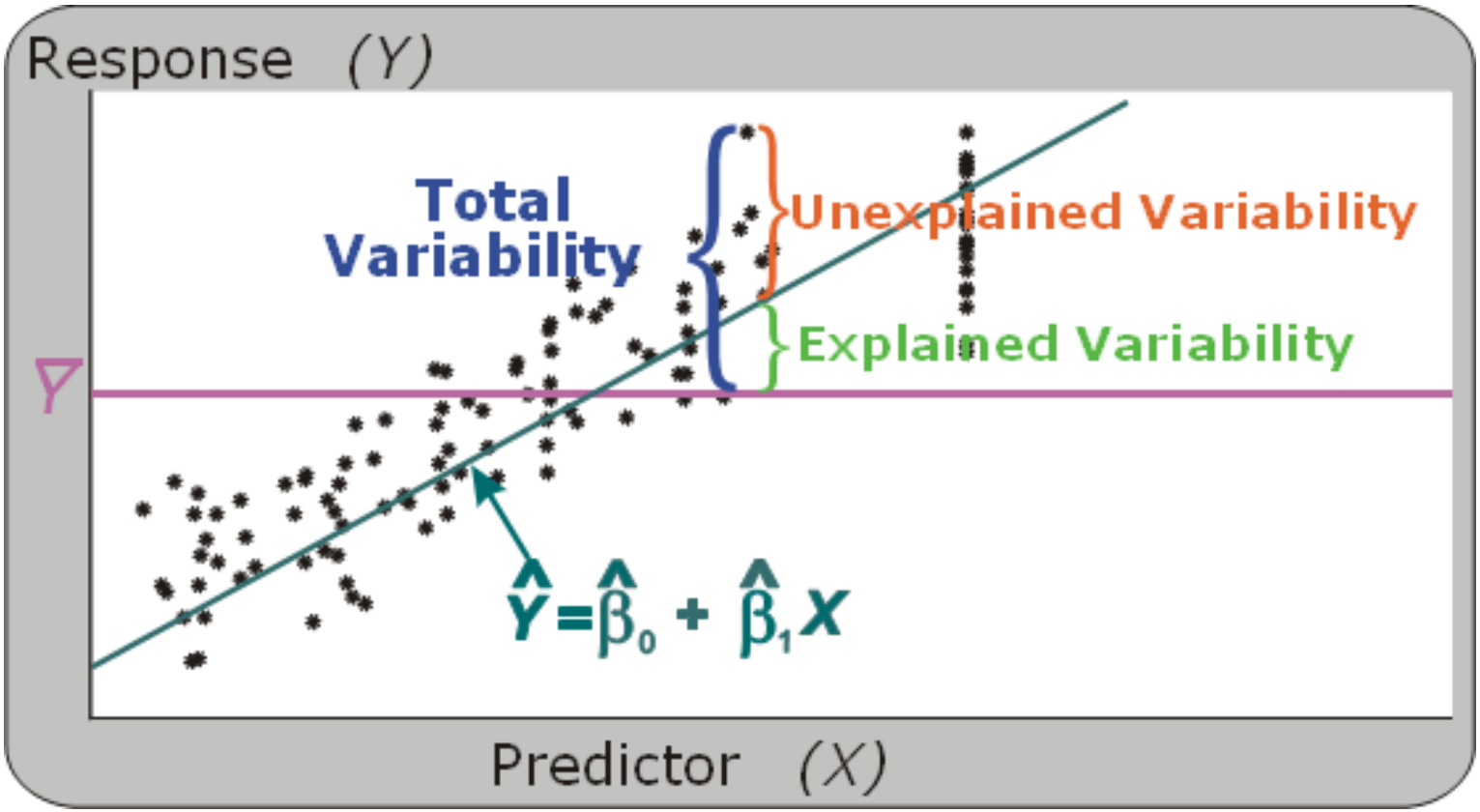
PRINCIPLE OF LEAST SQUARES

The line that minimizes the sum of squares of the deviations of the observed values of y from those predicted is the **best-fitting line.** The sum of squared deviations is commonly called the **sum of squares for error** (SSE) and defined as

$$\text{SSE} = \Sigma(y_i - \hat{y}_i)^2 = \Sigma(y_i - a - bx_i)^2$$

SSE (*Sum of Squares for Error*) = JKG (Jumlah Kuadrat Galat)

Keragaman yang Dapat Dijelaskan dan yang Tidak Dapat Dijelaskan oleh Model

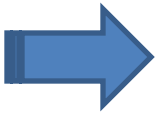


Contoh Data

Apakah semakin besar biaya iklan yang dikeluarkan akan semakin besar pula profit yang diperoleh?

Diamati contoh acak 10 perusahaan yang memproduksi *Laptop*, kemudian dicatat pengeluaran iklan (dalam milyar) dan profit (dalam milyar) selama tahun 2015.

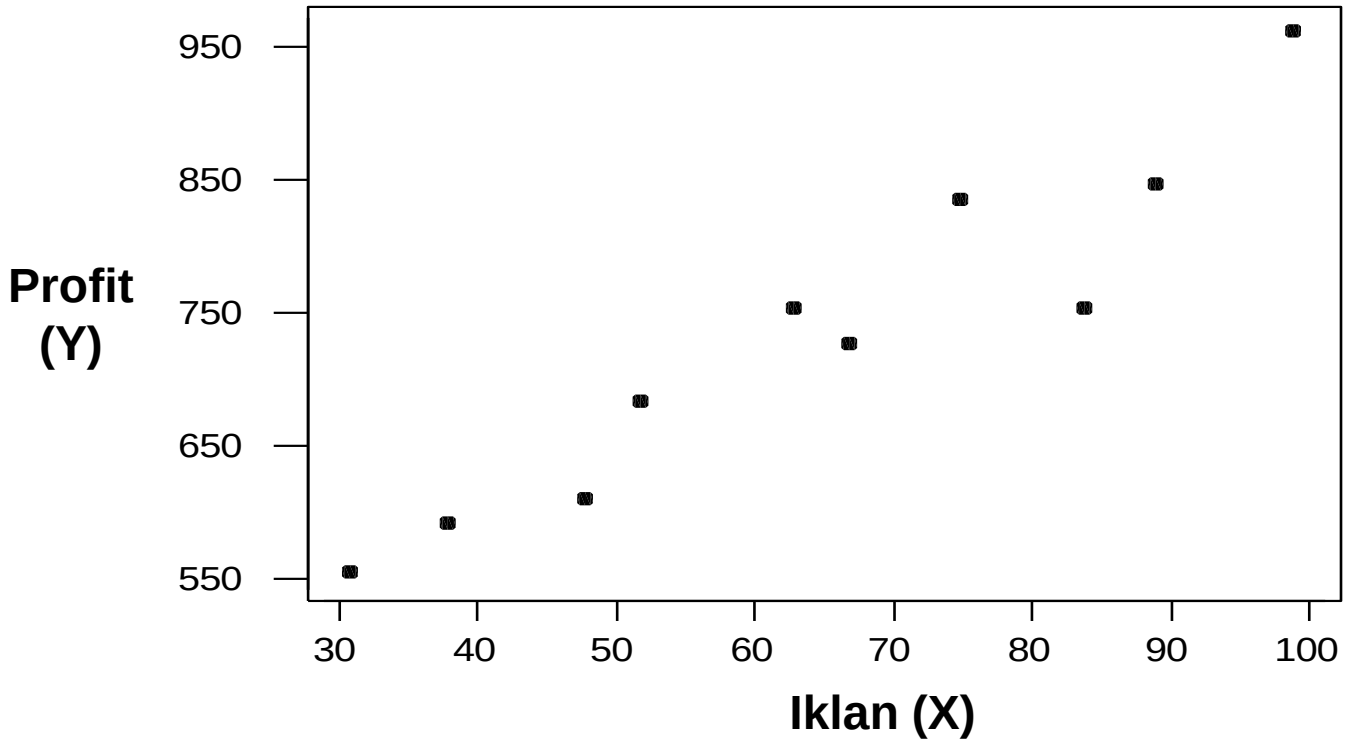
a. Buat scatter plotnya dan jelaskan.
b. Tentukan persamaan model regresinya.
c. Tentukan penduga bagi parameter model regresi tersebut.



Iklan	Profit
31	553
38	590
48	608
52	682
63	752
67	725
75	834
84	752
89	845
99	960

Penyelesaian :

Plot antara pengeluaran Iklan(milyar) dg Profit (milyar)



$$\text{Model: } Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \;\; ; i = 1,2,\ldots,n$$

Penyelesaian :

	x	y	x^2	y^2	xy
	31	553	961	305,809	17,143
	38	590	1,444	348,100	22,420
	48	608	2,304	369,664	29,184
	52	682	2,704	465,124	35,464
	63	752	3,969	565,504	47,376
	67	725	4,489	525,625	48,575
	75	834	5,625	695,556	62,550
	84	752	7,056	565,504	63,168
	89	845	7,921	714,025	75,205
	99	960	9,801	921,600	95,040
Jumlah	**646**	**7,301**	**46,274**	**5,476,511**	**496,125**

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Sxy	24,480.4
Sxx	4,542.4
Syy	146,050.9
b_1	**5.39**
b_0	**381.95**

Analisis Regresi

Contoh output regresi dengan **Minitab** (1)

Regression Analysis (Iklan vs Profit)

The regression equation is Profit = 381.95 + 5.39*Iklan

Predictor	Coef	StDev	T	P
Constant	381.95	42.40	9.01	0.000
Iklan	5.3893	0.6233	8.65	0.000

S = 42.01 R-Sq = 90.3% R-Sq(adj) = 89.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	131932	131932	74.76	0.000
Error	8	14118	1765		
Total	9	146051			

Uji Hipotesis

Bagaimana Pengujian terhadap model regresi ??

- parsial (per koefisien) → uji-t
- bersama → uji-F (Anova)

Bagaimana menilai kesesuaian model ??

- R^2 (Koefisien Determinasi: persentase keragaman Y yang mampu dijelaskan oleh X)

Uji Hipotesis : $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

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

$$S_{yy} = \frac{S_{xy}^2}{S_{xx}} + \text{SSE}$$

Total variability of y	Variability explained by linear relation	Residual or unexplained variability

JK Total = JK Regresi + JK Galat → **JK : Jumlah Kuadrat**

Keragaman total = keragaman yang dapat dijelaskan oleh model + keragaman yang tidak dapat dijelaskan oleh model

R^2 = (JK Regresi)/(JK Total) = JKR/JKT

Uji Hipotesis (1) $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

$$t - hitung = \frac{b_1 - \beta_1}{S_{b_1}}$$

S_{b1} disebut galat baku (*standard of error*) bagi b_1 → SE(b_1)

$$S_{b_1} = \sqrt{\frac{s^2}{\sum (x_i - \bar{x})^2}}$$

$$s^2 = \frac{SSE}{n-2}$$

Tolak Ho jika:
|t-hit| > t(α/2; db=n-2)

$$SSE = \Sigma(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

Uji Hipotesis (2) $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 > 0$

$$t - hitung = \frac{b_1 - \beta_1}{S_{b_1}}$$

$$S_{b_1} = \sqrt{\frac{s^2}{\sum (x_i - \bar{x})^2}}$$

$$s^2 = \frac{SSE}{n-2}$$

Tolak Ho jika:
t-hit > t(α; db=n-2)

$$SSE = \Sigma(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

Uji Hipotesis (3) $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 < 0$

$$t - hitung = \frac{b_1 - \beta_1}{S_{b_1}}$$

$$S_{b_1} = \sqrt{\frac{s^2}{\sum (x_i - \bar{x})^2}}$$

$$s^2 = \frac{SSE}{n-2}$$

Tolak Ho jika:
t-hit $< -$ t(α; db=n-2)

$$SSE = \Sigma(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

Latihan (1)

- Apakah iklan berpengaruh pada profit perusahaan? Uji hipotesis Anda pada taraf nyata $\alpha = 0.05$

Jawaban Ringkas $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

(4). $t - hitung = \dfrac{b_1 - \beta_1}{S_{b_1}} = \dfrac{b_1 - 0}{S_{b_1}}$ = 5.39/0.623 = **8.64**

(3). $S_{b_1} = \sqrt{\dfrac{s^2}{\sum (x_i - \bar{x})^2}}$ = $\sqrt{(S^2/S_{xx})}$ = $\sqrt{(1{,}764.81/4{,}542.4)}$ = 0.623

(2). $s^2 = \dfrac{SSE}{n-2}$ = 1,764.81

Sxy	24,480.4
Sxx	4,542.4
Syy	146,050.9

(1). $SSE = \Sigma(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = S_{yy} - \dfrac{S_{xy}^2}{S_{xx}}$ = 14,118.45

---

t(α/2; db=n-2) = t(0.025; 8) = **2.306**

Karena (t-hit = 8.64) > 2.306 maka **TOLAK H_0**, artinya iklan berpengaruh pada profit perusahaan untuk taraf uji α = 0.05

Latihan (2)

- Apakah semakin besar iklan akan mengakibatkan semakin besar profit? Uji pada taraf nyata $\alpha = 0.05$

Jawaban Ringkas $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 > 0$

(4). $t - hitung = \dfrac{b_1 - \beta_1}{S_{b_1}} = \dfrac{b_1 - 0}{S_{b_1}}$ = 5.39/0.623 = **8.64**

(3). $S_{b_1} = \sqrt{\dfrac{s^2}{\sum (x_i - \bar{x})^2}}$ = $\sqrt{(S^2/S_{xx})}$ = $\sqrt{(1{,}764.81/4{,}542.4)}$ = 0.623

(2). $s^2 = \dfrac{SSE}{n-2}$ = 1,764.81

Sxy	24,480.4
Sxx	4,542.4
Syy	146,050.9

(1). $SSE = \Sigma(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = S_{yy} - \dfrac{S_{xy}^2}{S_{xx}}$ = 14,118.45

---

t(α; db=n-2) = t(0.05; 8) = **1.860**

Karena (t-hit = 8.64) > 1.860 maka **TOLAK H_0**, artinya semakin besar iklan akan mengakibatkan semakin besar profit untuk taraf uji α = 0.05

Latihan (3)

- Berapa profit yang dihasilkan jika iklan yang dikeluarkan 76 milyar? Apakah hasil dugaan ini valid? Kenapa?

Latihan (4)

- Tentukan koefisien determinasinya? Apa maknanya?

Koefisien Determinasi (R^2)

R^2 = (JK Regresi)/(JK Total)

$$= \frac{\frac{(S_{xy})^2}{S_{xx}}}{S_{yy}} = \frac{(S_{xy})^2}{S_{xx} S_{yy}} = 0.903 = \mathbf{90.3\%}$$

Artinya, **90.3** persen keragaman pada *Y* (profit) dapat diterangkan oleh keragaman pada *X* (iklan)

Keterbatasan Korelasi dan Regresi Linear

- Korelasi dan Regresi Linear hanya menggambarkan hubungan yang **linear**
- Korelasi dan metode kuadrat terkecil pada regresi linear tidak resisten terhadap **pencilan**
- Prediksi **di luar selang nilai X** tidak diperkenankan karena kurang akurat
- Hubungan antara dua variabel bisa dipengaruhi oleh **variabel lain** di luar model

Catatan

- Apa itu analisis regresi?
- Apa bedanya dengan korelasi?

Analisis Regresi → Analisis statistika yang memanfaatkan hubungan antara dua atau lebih peubah kuantitatif sehingga salah satu peubah dapat diramalkan dari peubah lainnya.

Korelasi → mengukur keeratan HUBUNGAN LINEAR dari dua variabel

PR/Tugas

Dikumpulkan di TU Dept Statistika, pada hari Senin minggu depan sebelum jam 12.00 (via Ibu Mar)

Catatan : **m** = (digit ke-8) + (digit ke-9) dari NIM

Misal NIM : H24130075 → **m** = 7 + 5 = 12

---

1. Mendenhall (Exercise 12.7 a-c), hal. 511 → y : (data + 0.m)
2. Mendenhall (Exercise 12.20 a-c), hal. 520 → y : (data + 0.m)

Terima Kasih

Materi ini bisa di-download di:

kusmans.staff.ipb.ac.id