STK 211 Analisis statistika

Materi 8 Analisis Korelasi dan Regresi

Pendahuluan

- Kita umumnya ingin mengetahui hubungan antar peubah
- Analisis Korelasi digunakan untuk melihat keeratan hubungan linier antar dua peubah
- Analisis Regresi digunakan untuk melihat hubungan sebab akibat antar peubah

Analisis Korelasi

Analisis Korelasi

- Mengukur keeratan/kekuatan hubungan antar 2 peubah
- Dinyatakan dalam suatu ukuran nilai → Koefisien korelasi (r_{xy} atau disingkat r)
- Koefisien korelasi pearson (numerik):

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

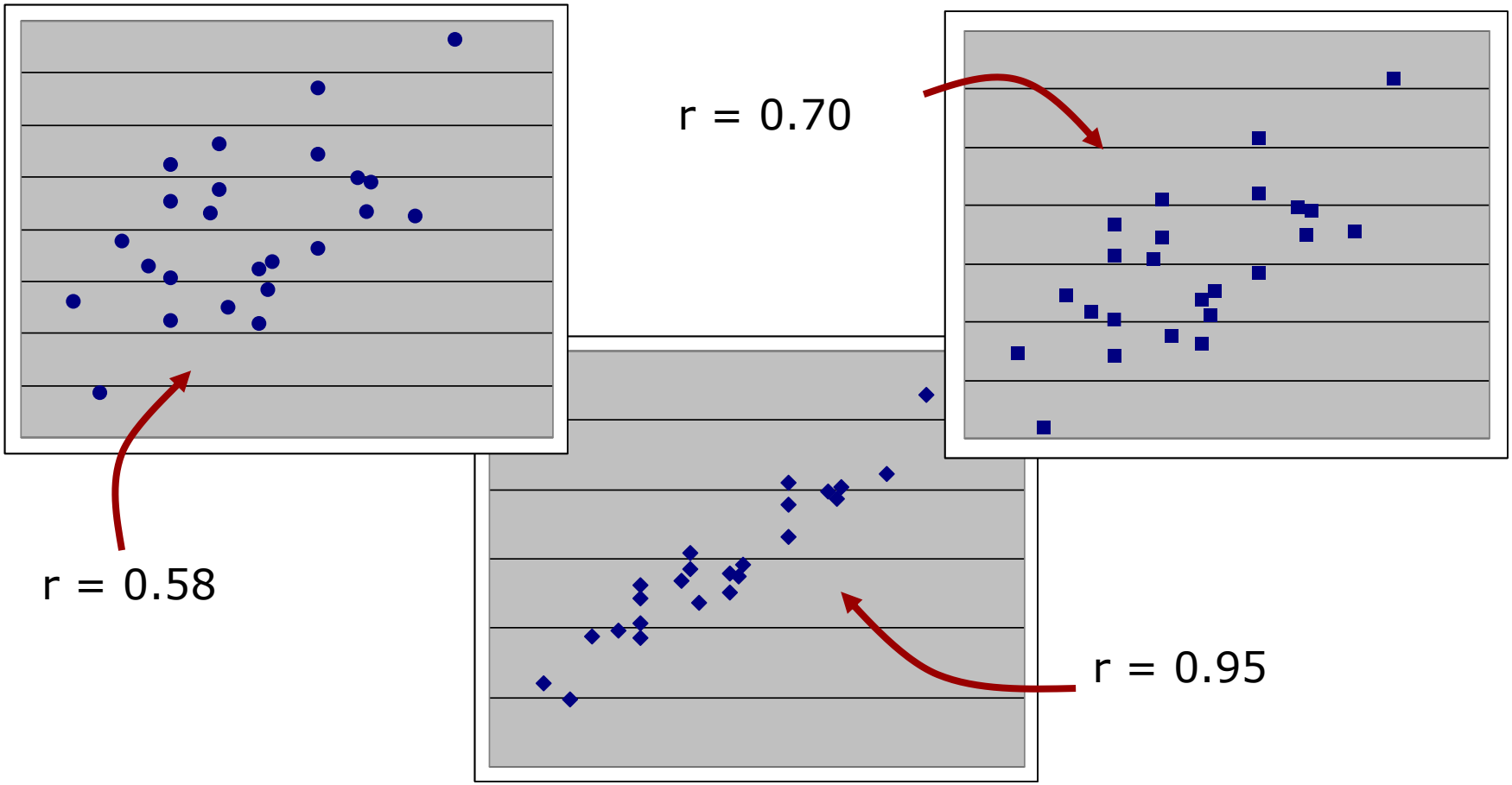
$$S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$S_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}} \quad \text{dan} \quad S_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n - 1}}$$

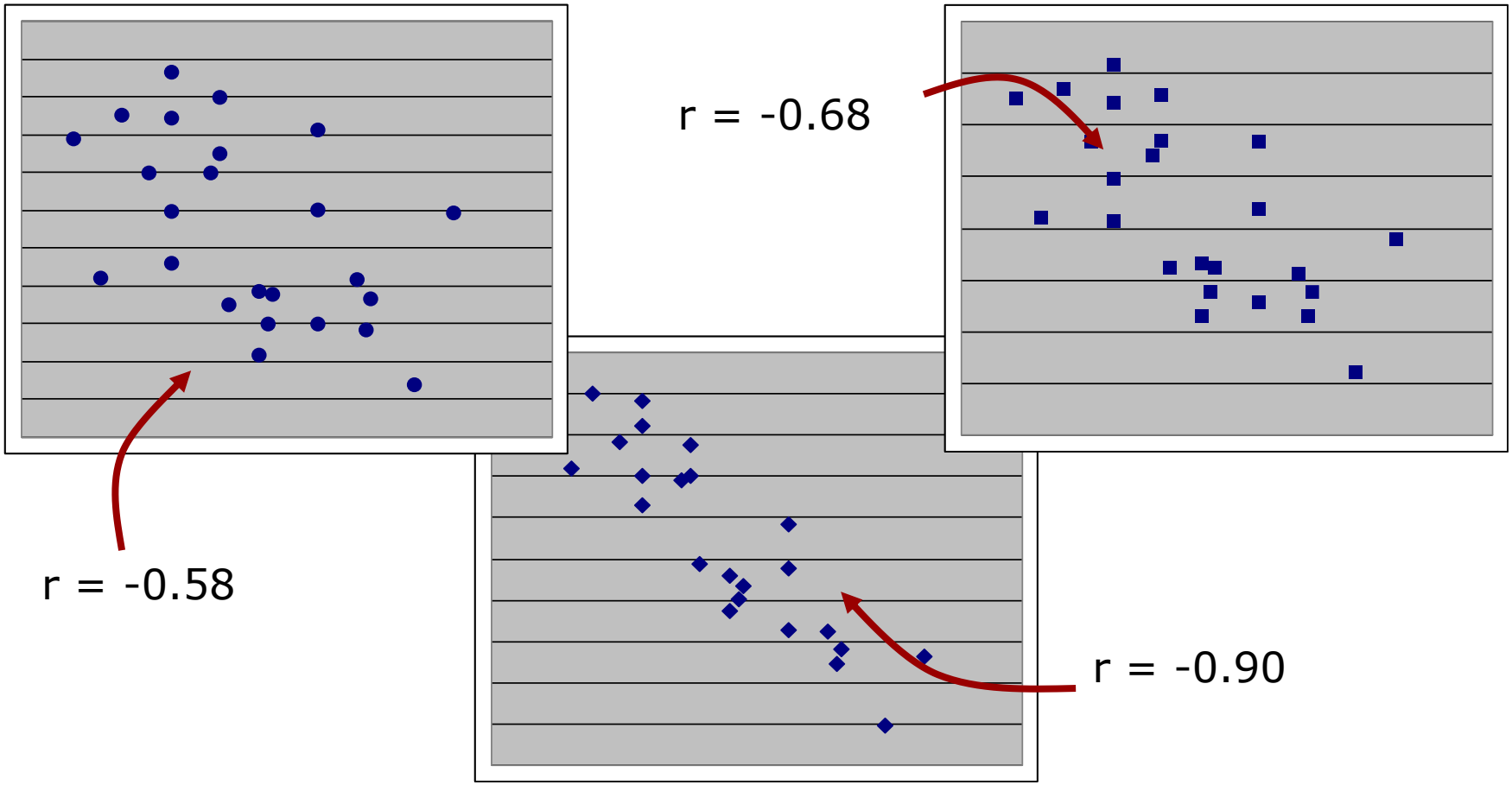
Koefisien Korelasi

- Bernilai antara –1 s/d +1
- Tanda koefisien menunjukkan arah hubungan kedua peubah
- Besarnya koefisien menunjukkan keeratan hubungan kedua peubah

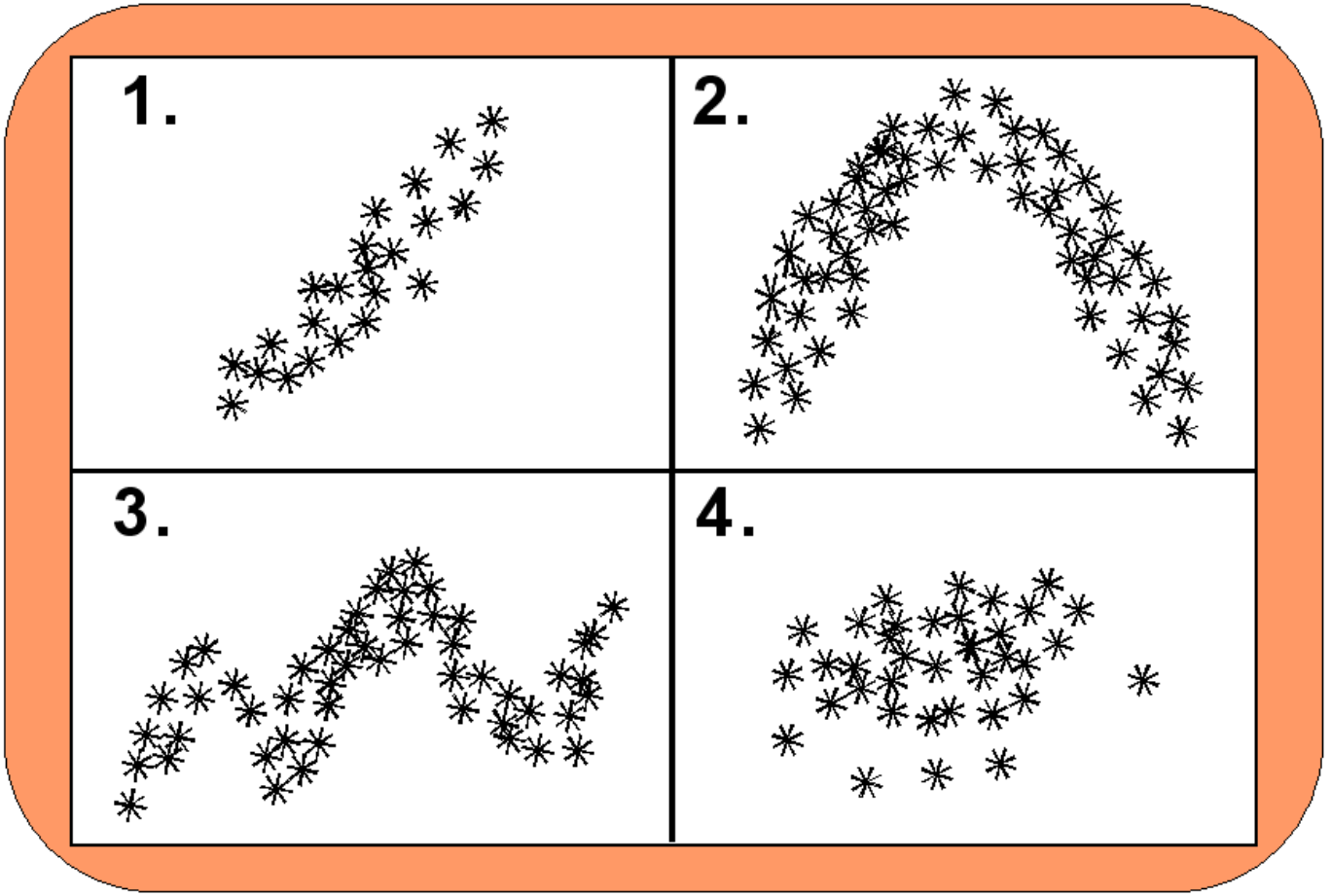
Koefisien Korelasi (+)



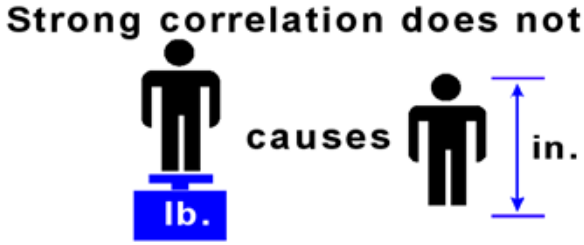
Koefisien Korelasi (-)



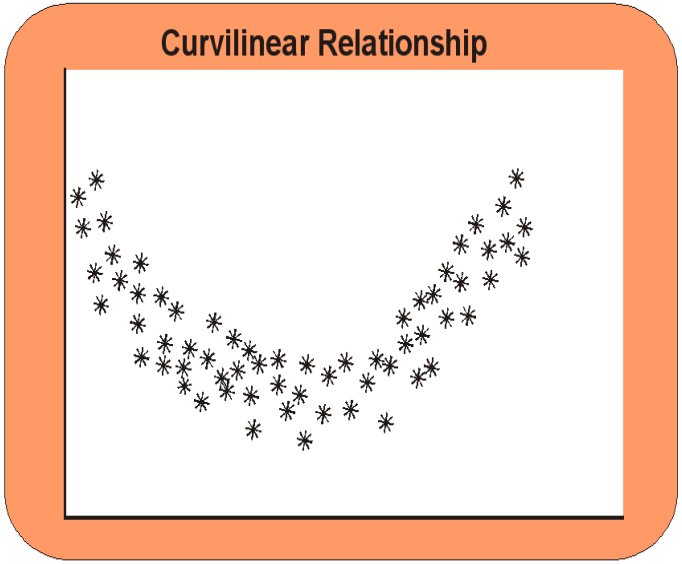
Beberapa Kemungkinan Hubungan Antar 2 Peubah



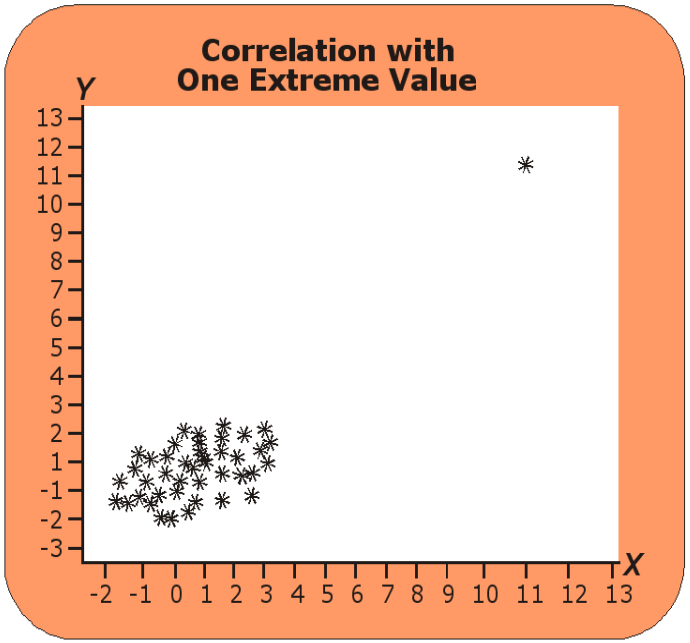
3 Penyalahgunaan Koefisien Korelasi



Missing Another Type of Relationship



Extreme Data Values

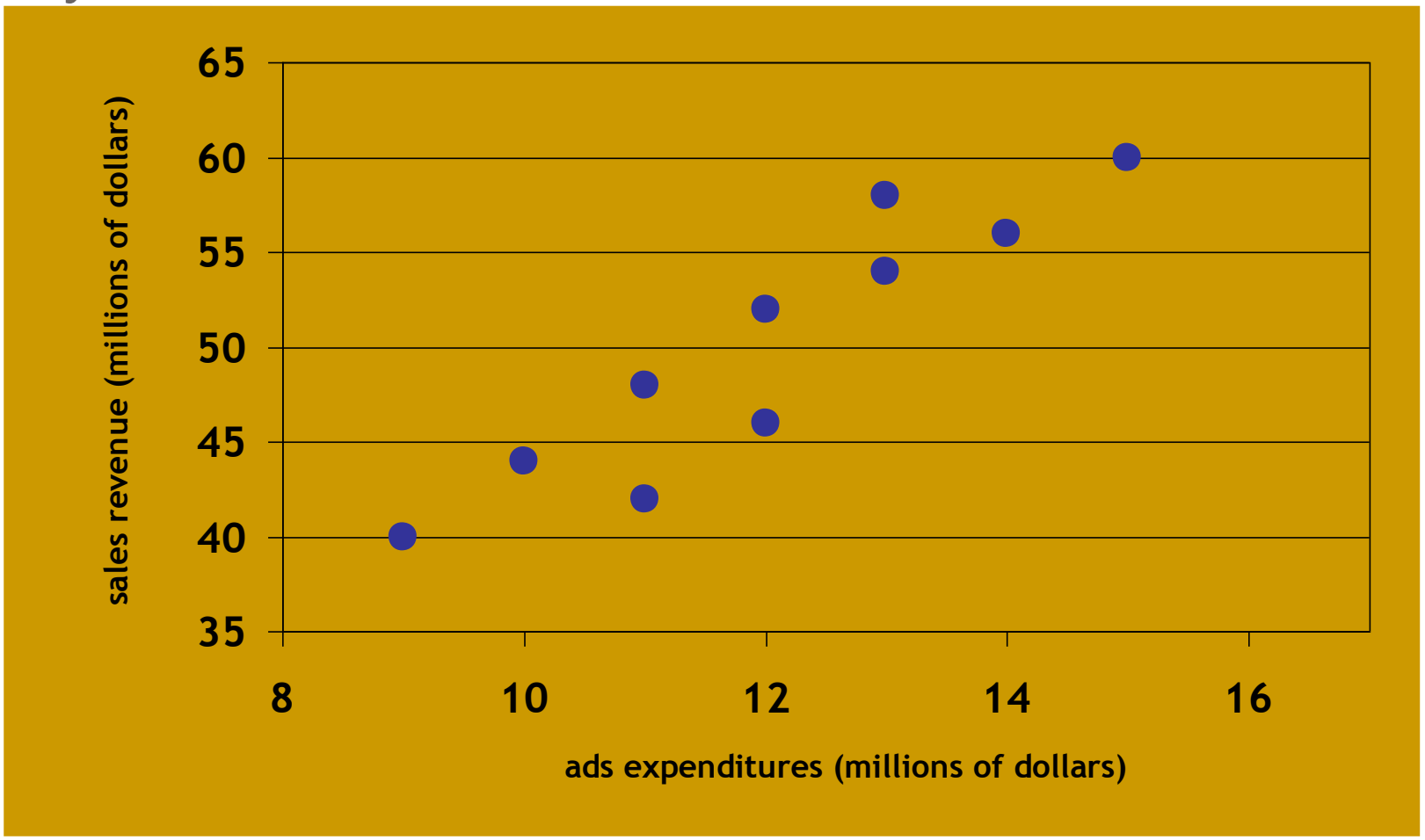


Ilustrasi

Misalnya ingin melihat hubungan antara pengeluaran untuk iklan (ads expenditures, X) dengan penerimaan melalui penjualan (sales revenue, Y)

Waktu	1	2	3	4	5	6	7	8	9	10
X	10	9	11	12	11	12	13	13	14	15
Y	44	40	42	46	48	52	54	58	56	60

$r_{xy} = 0.9226$


65
60
55
50
45
40
35
sales revenue (millions of dollars)
8
10
12
14
16
ads expenditures (millions of dollars)

Analisis Regresi Linear Sederhana

Pengantar

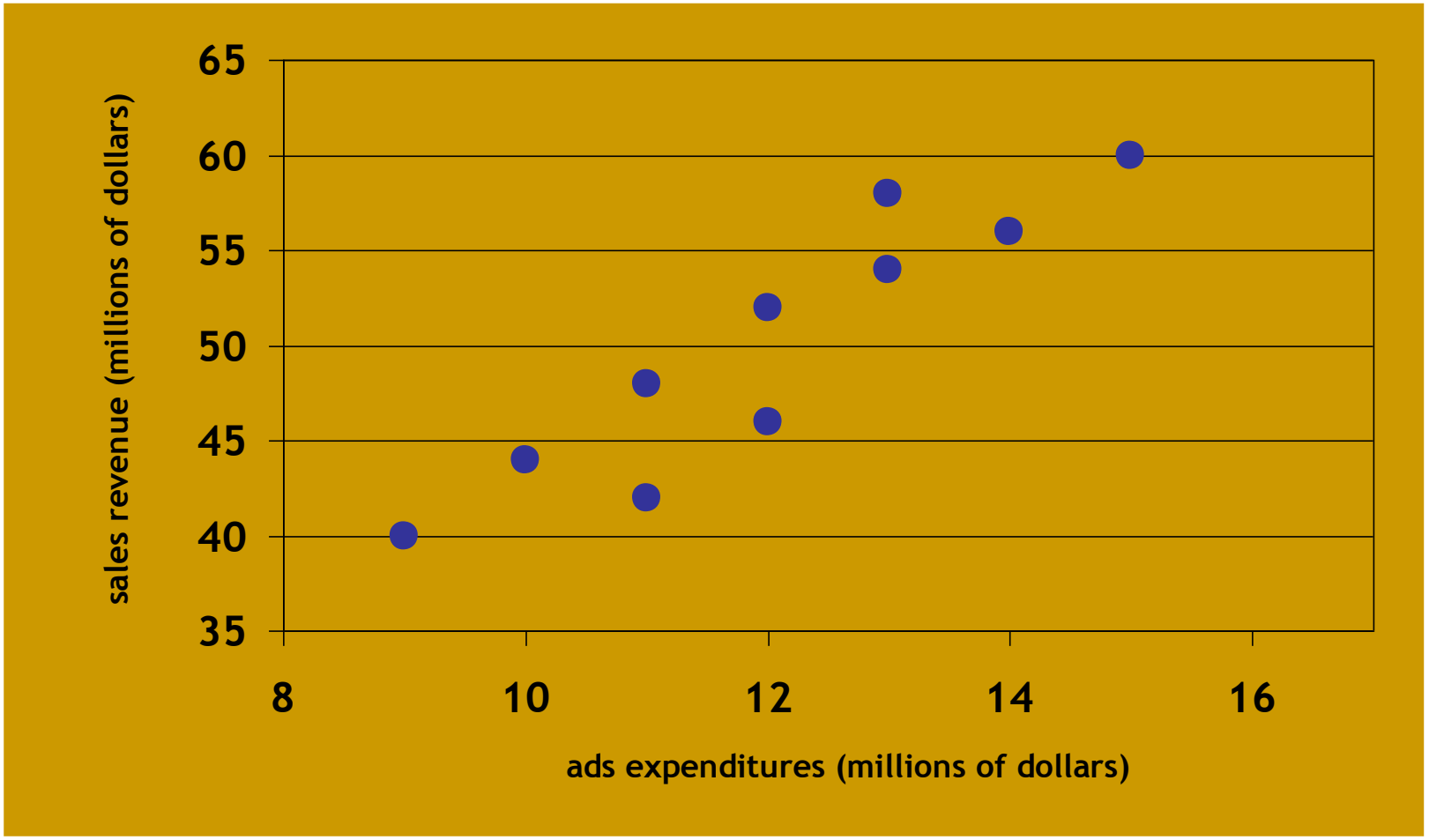
- Terdapat 2 peubah numerik : peubah yang satu mempengaruhi peubah yang lain
- Peubah yang mempengaruhi → X, peubah bebas, peubah penjelas, peubah kovariat
- Peubah yang dipengaruhi → Y, peubah tak bebas, peubah respon

Pengantar

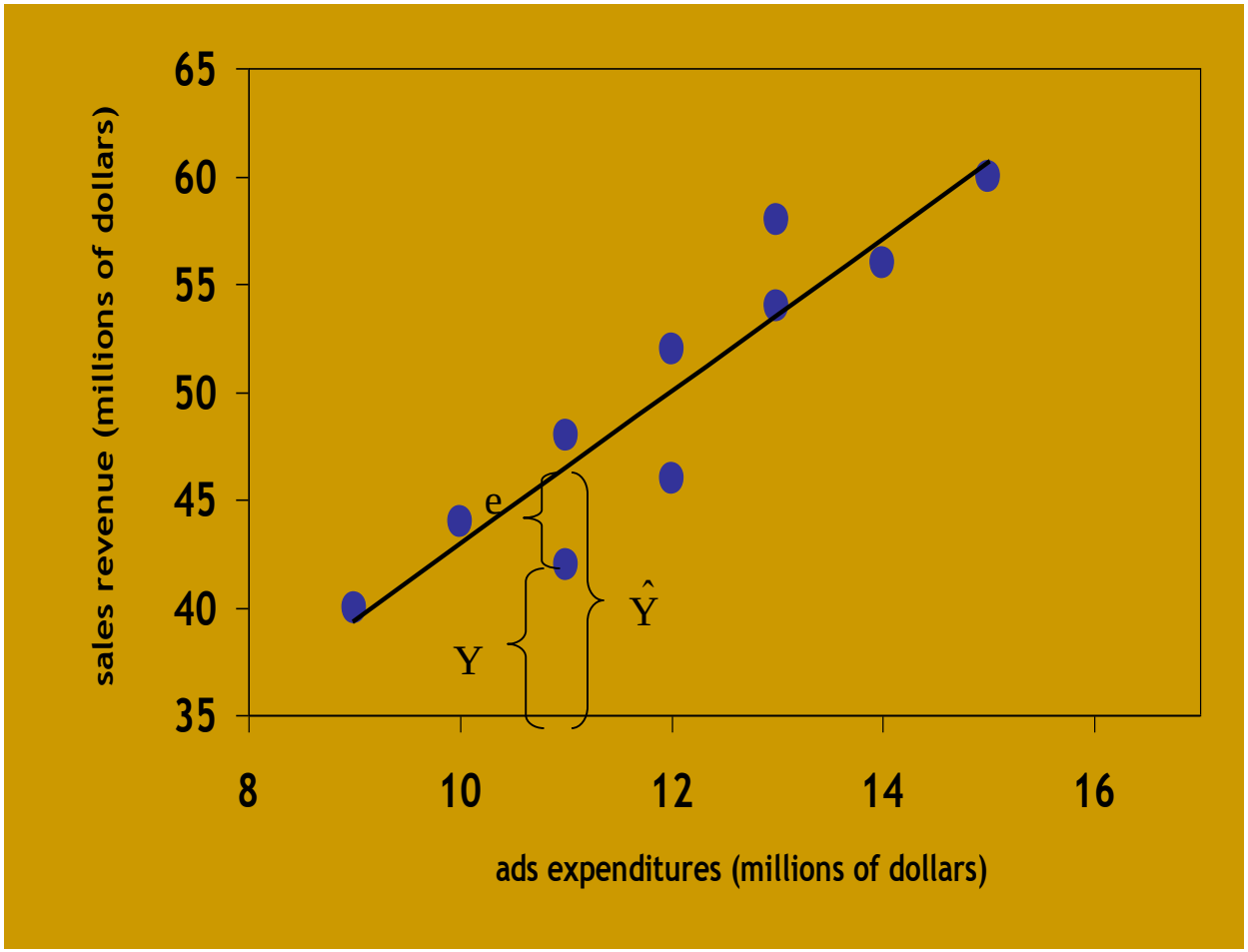
Misalnya ingin melihat hubungan antara pengeluaran untuk iklan (ads expenditures, X) dengan penerimaan melalui penjualan (sales revenue, Y)

Waktu	1	2	3	4	5	6	7	8	9	10
X	10	9	11	12	11	12	13	13	14	15
Y	44	40	42	46	48	52	54	58	56	60

Pengantar


65
60
55
50
45
40
35
sales revenue (millions of dollars)
8
10
12
14
16
ads expenditures (millions of dollars)

Pengantar



Ingin dibuat model

$$Y = \alpha + \beta X$$

Model memuat error, selisih nilai sebenarnya dengan dugaan berdasar model

$$e = Y - \hat{Y}$$

Bagaimana mendapatkan α dan β?

Metode yang digunakan : OLS (ordinary least squares/kuadrat terkecil), mencari α dan β sehingga jumlah kuadrat error paling kecil

Cari penduga α dan β sehingga

minimum $$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[Y_i - \alpha - \beta X_i \right]^2$$

Bagaimana mendapatkan α dan β?

$$b = \frac{\sum_{i=1}^{n} [X_i - \overline{X}][Y_i - \overline{Y}]}{\sum_{i=1}^{n} [X_i - \overline{X}]^2}$$

$$a = \overline{Y} - b\overline{X}$$

Rata-rata Y

Rata-rata X

X	Y	X-rata	Y-rata	(X-rata)(Y-rata)	$(X\text{-rata})^2$
10	44	-2	-6	12	4
9	40	-3	-10	30	9
11	42	-1	-8	8	1
12	46	0	-4	0	0
11	48	-1	-2	2	1
12	52	0	2	0	0
13	54	1	4	4	1
13	58	1	8	8	1
14	56	2	6	12	4
15	60	3	10	30	9

$\overline{X} = 12$

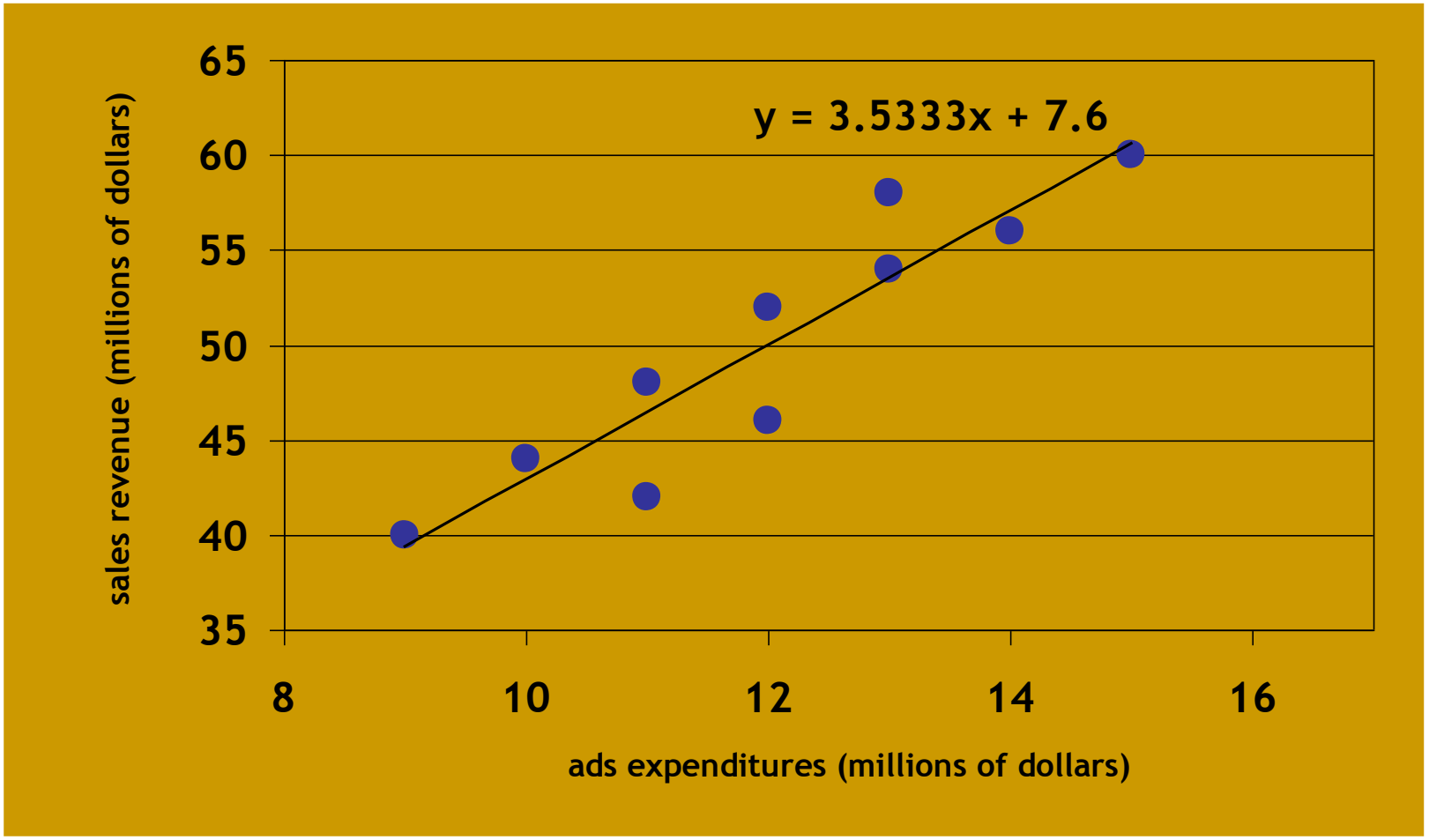
$\overline{Y} = 50$

$\sum [X - \overline{X}][Y - \overline{Y}] = 106$

$\sum [X - \overline{X}]^2 = 30$

b = 106 / 30 = 3.533

a = 50 – 3.533 (12) = 7.60


y = 3.5333x + 7.6
sales revenue (millions of dollars)
65
60
55
50
45
40
35
8
10
12
14
16
ads expenditures (millions of dollars)

Interpretasi a dan b

- a = besarnya nilai Y ketika X sebesar 0
  - pada bbrp kasus nilai a adalah penyesuaian
- b = besarnya perubahan nilai Y ketika X berubah satu satuan. Tanda koefisien b menunjukkan arah hubungan X dan Y

Pada kasus ilustrasi

- a = 7.6 → besarnya sales revenue jika tidak ada belanja iklan adalah 7.6 mlo
- b = 3.533 → jika belanja iklan dinaikkan 1 juta dolar maka sales revenue naik 3.533 juta dolar

Uji Terhadap koefisien β

$H_0 : \beta = 0$ (artinya X tidak mempengaruhi Y)

$H_1 : \beta \neq 0$ (artinya X mempengaruhi Y)

$$stat \quad \text{uji} = t = \frac{b - 0}{s_b} \qquad s_b = \sqrt{\frac{\sum_{i=1}^{n} \left[Y_i - \hat{Y}_i \right]^2}{(n-k) \sum_{i=1}^{n} \left[X_i - \overline{X} \right]^2}}$$

Tolak H_0 jika nilai t melebihi nilai t pada tabel dengan derajat bebas (n-2) dengan tingkat kesalahan $\alpha/2$

Uji Terhadap koefisien β

- Nilai $s_b = 0.52$
- Nilai $t = 6.79$
- Nilai t pada tabel (db = 8, $\alpha = 5\%$) = 2.306
- Kesimpulan : Tolak H_0, data mendukung kesimpulan adanya pengaruh ads expenditure terhadap sales revenue.

Ukuran Kebaikan Model

- Menggunakan koefisien determinasi (R^2, R-squared)
- R-squared bernilai antara 0 s/d 1
- R-squared adalah persentase keragaman data yang mampu diterangkan oleh model
- R-squared tinggi adalah indikasi model yang baik

Ukuran Kebaikan Model

$$R^2 = \frac{\sum \left[\hat{Y}_i - \overline{Y}\right]^2}{\sum \left[Y_i - \overline{Y}\right]^2}$$

- Model dalam ilustrasi bisa ditunjukkan memiliki R-squared 0.85 atau 85%

ANALISIS REGRESI LINIER BERGANDA

Pengantar

- Pada sesi sebelumnya kita hanya menggunakan satu buah X, dengan model $Y = \alpha + \beta X$
- Dalam banyak hal, yang mempengaruhi Y bisa lebih dari satu. Model umum regresi linear berganda adalah

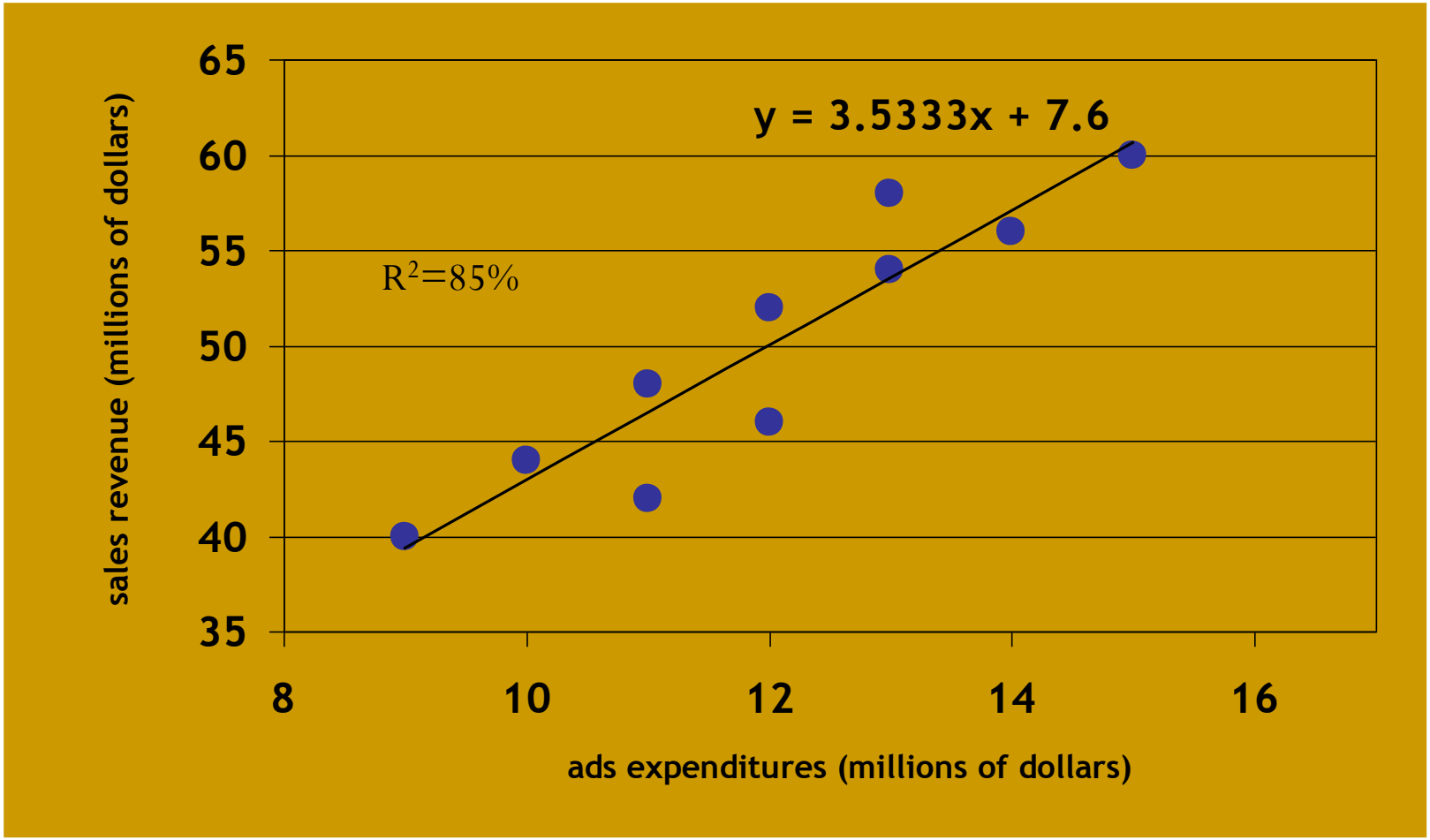
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Ilustrasi

Misalnya dalam satu perusahaan ingin melihat hubungan antara pengeluaran untuk iklan (ads expenditures, X_1) dengan penerimaan melalui penjualan (sales revenue, Y)

Waktu	1	2	3	4	5	6	7	8	9	10
X_1	10	9	11	12	11	12	13	13	14	15
Y	44	40	42	46	48	52	54	58	56	60

Dengan Regresi Linier Sederhana



Ilustrasi

Kemudian terdapat informasi lain mengenai pengeluaran untuk quality control (X_2).

Waktu	1	2	3	4	5	6	7	8	9	10
X_1	10	9	11	12	11	12	13	13	14	15
X_2	3	4	3	3	4	5	6	7	7	8
Y	44	40	42	46	48	52	54	58	56	60

Ilustrasi

- Proses mencari penduga bagi koefisien β_0, β_1, dan β_2 memiliki konsep yang sama dengan model regresi sederhana, namun lebih kompleks. Karena alasan tersebut digunakan bantuan komputer.
- Output standar komputer:
  - ANOVA → pengujian simultan
  - Pengujian Parsial
  - Nilai dugaan koefisien
  - Ukuran kebaikan model

Ilustrasi Output SAS System

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	409.26761	204.63380	46.61	<.0001
Error	7	30.73239	4.39034		
Corrected Total	9	440.00000			

Root MSE	2.09531	R-Square	0.9302
Dependent Mean	50.00000	Adj R-Sq	0.9102
Coeff Var	4.19063		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	17.94366	5.91914	3.03	0.0191
X1	1	1.87324	0.70334	2.66	0.0323
X2	1	1.91549	0.68101	2.81	0.0260

ANOVA

- Digunakan untuk menguji secara simultan pengaruh seluruh X
- H_0: semua $\beta_i = 0$ (tidak ada X yang berpengaruh terhadap Y)
- H_1: ada $\beta_i \neq 0$ (ada X yang berpengaruh terhadap Y)
- Konsep dasar : ANOVA membandingkan besarnya keragaman yang terkandung dalam model dengan keragaman yang tersisa pada error model. Jika rasio keduanya besar, maka X mempengaruhi Y. Rasio itu dilambangkan dengan nilai F. Semakin besar nilai F, semakin kecil nilai-p, cenderung menolak H_0.

ANOVA: ilustrasi

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	409.26761	204.63380	46.61	<.0001
Error	7	30.73239	4.39034		
Corrected Total	9	440.00000			

Seandainya kita gunakan α = 5%, maka nilai-p ini lebih kecil daripada 5%, sehingga kita putuskan TOLAK H_0, artinya ada X yang mempengaruhi sales revenue

Ilustrasi

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	17.94366	5.91914	3.03	0.0191
X1	1	1.87324	0.70334	2.66	0.0323
X2	1	1.91549	0.68101	2.81	0.0260

Modelnya

$Y = 17.94 + 1.87\ X1 + 1.91\ X2$

Uji parsial. Menguji masing-masing X. Karena p-value X1 dan X2 kecil, maka disimpulkan bahwa pengaruh keduanya terhadap sales revenue signifikan secara statistik

Ilustrasi

Root MSE	2.09531	R-Square	0.9302
Dependent Mean	50.00000	Adj R-Sq	0.9102
Coeff Var	4.19063		

Ukuran kebaikan model : R^2. Penambahan X lain dalam model akan selalu meningkatkan nilai R^2, namun menurunkan derajat bebas error. Agar evaluasi terhadap kebaikan model tidak terganggu, nilai R^2 dikoreksi menjadi Adjusted R^2.

Beberapa Permasalahan

- Multikolinearitas, korelasi antar X menyebabkan variasi dugaan koefisien meningkat
- Heteroskedastisitas, ketidakhomogenan variasi dugaan Y di setiap nilai X
- Autokorelasi, error masih berpola

Multikolinearitas

- Dalam analisis regresi berganda, antar X tidak boleh saling berkorelasi.
- Korelasi antar X menyebabkan dugaan koefisien tidak stabil (memiliki variasi yang besar).
- Hal ini menyebabkan kesimpulan cenderung menyatakan terima H_0 atau pengaruh X tidak signifikan meskipun nilai R^2 sangat tinggi.

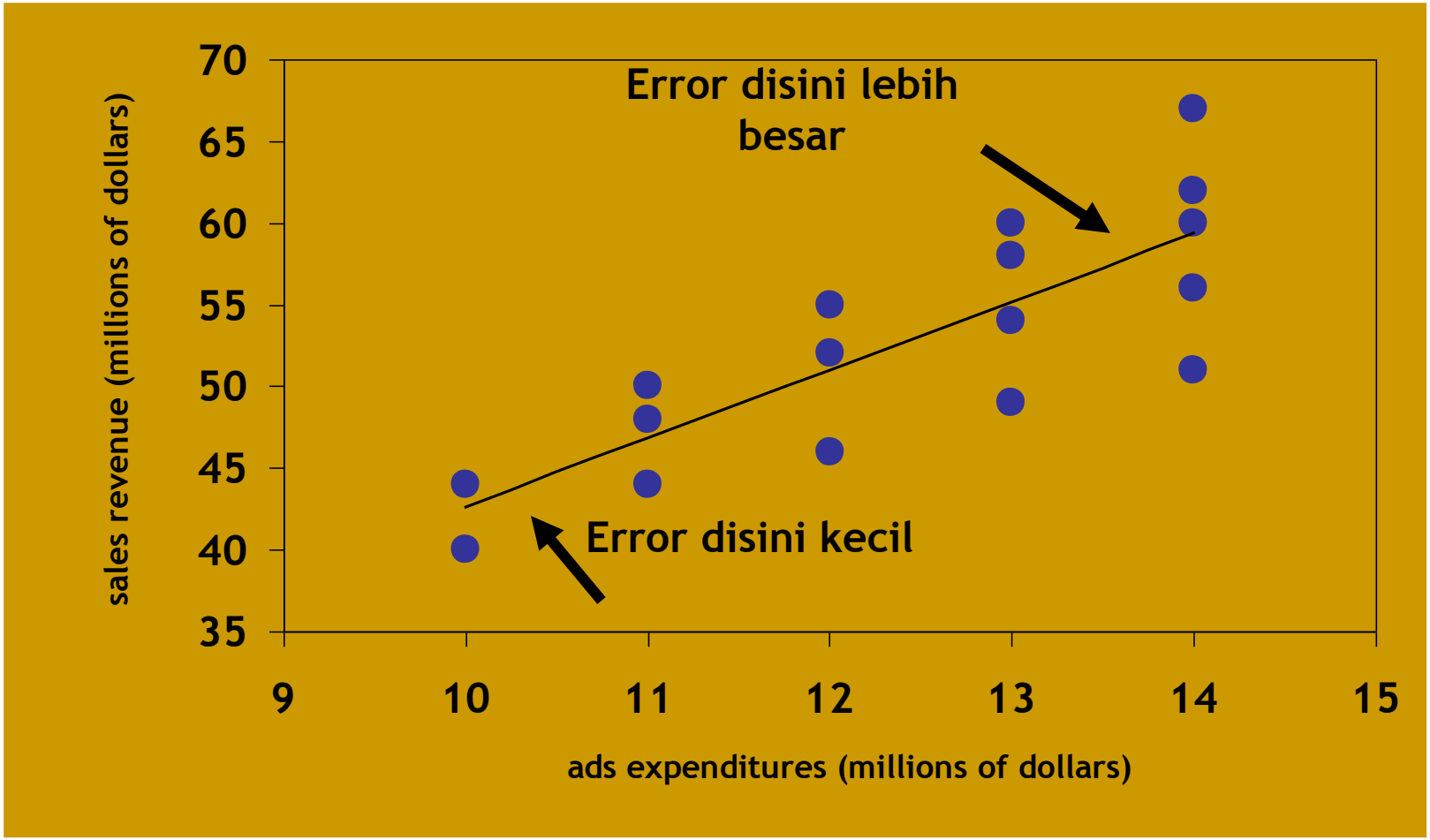
Multikolinearitas

- Dideteksi dengan melihat korelasi antar X.
- Dideteksi dengan nilai VIF (variance inflation factor). Nilai VIF yang lebih dari 10 merupakan **INDIKASI** multikolinearitas mempengaruhi pendugaan.

Multikolinearitas : Penanganan

- Tambah banyaknya data
- Buat restriksi terhadap koefisien, berdasarkan informasi terdahulu
- Buang salah satu variabel yang saling berkorelasi
- Gunakan metode regresi lain (ridge regression, principal component regression, partial least squares, etc)

Heteroskedastisitas



Autokorelasi

- Autokorelasi: korelasi antar error
- Model yang baik → menghasilkan error yang acak, tidak lagi berpola
- Diukur menggunakan statistik D-W (Durbin-Watson)

Autokorelasi: Penanganan

- Masukkan ke dalam model, lag dari variabel Y. Jadi yang mempengaruhi Y selain X adalah Y waktu sebelumnya. Misalkan, harga saat ini ada hubungannya dengan harga kemarin, atau dua hari yang lalu, dst. → lag distributed model

Hal-Hal Lain

- Lakukan terlebih dahulu eksplorasi melalui plot XY:
  - Mungkin ada data pencilan
  - Mungkin perlu transformasi data (misal: model kuadratik)
  - Mungkin perlu pemisahan model (misal: model untuk perusahaan swasta dalam negeri dan swasta asing tidak sama)

Hal-Hal Lain

- Pada kasus regresi berganda, terdapat teknik penyeleksian variabel bebas dalam model:
  - Forward method
  - Backward method
  - Stepwise

- *'All models are wrong, but some are useful' (G. E. P. Box)*

Selesai