MODUL 1

Teori Regresi

Prof. Dr. Sugiarto, M.Sc.

PENDAHULUAN

Dalam kondisi sehari-hari, kita sering menjumpai hubungan kausal suatu variabel dengan satu atau lebih variabel lainnya. Hubungan kausal tersebut memungkinkan kita untuk memprediksi nilai suatu variabel jika nilai variabel yang memengaruhinya diketahui. Bila bentuk dari dua variabel yang berhubungan kausal tersebut dapat ditemukan, dimungkinkan untuk mengetahui seberapa besar pengaruh perubahan satu unit variabel bebas terhadap variabel terikatnya. Modul ini memberikan dasar-dasar pengetahuan kepada mahasiswa untuk mendapatkan garis atau fungsi yang dapat menggambarkan hubungan kausal dari variabel-variabel yang menjadi perhatian, yang dikenal dengan garis regresi. Bila garis yang dimaksud dapat ditemukan, terhadap garis tersebut dapat dibuat suatu model matematis yang sangat berguna untuk memprediksi nilai variabel terikat apabila besarnya nilai variabel yang memengaruhinya diketahui atau ditetapkan.

Setelah mempelajari dan memahami isi modul ini, mahasiswa diharapkan memiliki kompetensi umum dalam mematut garis regresi yang merepresentasikan sebaran data aslinya. Selain itu, diharapkan mahasiswa memiliki kompetensi khusus dalam:

1. menjelaskan hubungan antara variabel,
2. menjelaskan pengertian regresi,
3. menjelaskan cara mematut garis regresi yang baik,
4. menjelaskan konsep populasi dan sampel dalam kaitannya dengan regresi,
5. menjelaskan konsep koefisien regresi,
6. menjelaskan konsep *the least square method*,
7. menjelaskan konsep parameter regresi.

Untuk memudahkan Anda dalam mencerna materi dalam modul ini, materi dalam modul ini dikemas dalam tiga kegiatan belajar, yaitu Kegiatan Belajar 1 mengenai mematut garis regresi, Kegiatan Belajar 2 mengenai analisis regresi, dan Kegiatan Belajar 3 mengenai penggunaan SPSS.

KEGIATAN BELAJAR 1

Mematut Garis Regresi

Fokus dalam Kegiatan Belajar 1 ini adalah mematut garis regresi yang mampu merepresentasikan sebaran data aslinya. Untuk mengantar pembaca memahami paparan di bagian ini secara runtut, ulasan berikut akan dibagi menjadi tiga bagian, yaitu hubungan antarvariabel, tren hubungan, dan mematut garis regresi.

A. HUBUNGAN ANTARVARIABEL

Dalam kondisi sehari-hari, kita sering menjumpai hubungan suatu variabel dengan satu atau lebih variabel lainnya. Contohnya dapat dilihat berikut.

1. Tingkat pendidikan seseorang berhubungan dengan besarnya gaji yang diperolehnya. Lazimnya, dengan semakin tinggi tingkat pendidikan seseorang, gaji yang akan diperoleh makin tinggi pula. Saat seseorang melamar kerja, yang bersangkutan akan memprediksi bahwa pada kondisi normal, secara rata-rata gajinya akan lebih tinggi dibandingkan calon lain yang memiliki tingkat pendidikan lebih rendah (*ceteris paribus*).
2. Biaya iklan suatu produk berhubungan dengan volume penjualannya. Semakin sering perusahaan menayangkan iklan produknya di berbagai media masa, makin besar biaya iklan yang dikeluarkan oleh perusahaan dengan harapan meningkatkan volume penjualan produknya. Direktur pemasaran perusahaan sangat berkepentingan untuk mengetahui hubungan permintaan produknya dengan biaya iklan. Penelitian terkait kedua variabel tersebut akan sangat membantu memperoleh informasi elastisitas pengeluaran iklan dari permintaan produk perusahaan, yang dalam hal ini mencerminkan rata-rata tanggapan permintaan terhadap kenaikan anggaran untuk iklan. Pengetahuan ini sangat berguna dalam penetapan anggaran iklan yang optimum. Pengusaha memerlukan informasi ini untuk bisa memutuskan apakah akan menguntungkan jika melakukan sejumlah pengeluaran untuk iklan tertentu.
3. Semakin mahal harga suatu barang, lazimnya makin bagus kualitas barang tersebut dibandingkan barang padanannya dan sebagainya.

Dalam hal ini, didapati adanya hubungan kualitas barang dengan harganya.

4. Belanja konsumsi perorangan berhubungan dengan pendapatan riil perorangan setelah pajak atau pendapatan yang bisa dibelanjakan. Analisis terhadap kedua variabel tersebut memungkinkan peramalan kecenderungan konsumsi marjinal (*marginal propensity to consume* = *MPC*), yaitu rata-rata perubahan dalam konsumsi untuk setiap perubahan pendapatan riil.
5. Tingkat perubahan upah kerja berhubungan dengan tingkat pengangguran. Dengan mengetahui hubungan kedua variabel tersebut, ahli ekonomi perburuhan dapat meramalkan perubahan rata-rata upah pada tingkat pengangguran tertentu. Pengetahuan tersebut berguna untuk menyatakan sesuatu mengenai proses yang berhubungan dengan inflasi dalam suatu perekonomian karena peningkatan upah tampaknya akan tecermin dalam peningkatan harga barang di masyarakat.
6. Analis investasi dapat meramalkan perubahan harga saham atas dasar pengetahuan perubahan dalam indeks pasar (misalkan indeks harga saham gabungan). Bila hubungan kedua variabel tersebut dapat diramalkan, manfaatnya bagi analis ataupun investor sangatlah berguna.

Berbagai ilustrasi yang dikemukakan tersebut memaparkan hubungan dua variabel (yang untuk selanjutnya dinotasikan dengan variabel *X* dan variabel *Y*) dalam konteks kausal. Dalam arti, bila besarnya variabel *X* berubah, besarnya variabel *Y* akan terpengaruh. Variabel yang memengaruhi variabel lain disebut sebagai variabel bebas (*independent variable*, yang selanjutnya dinotasikan dengan *X*), sedangkan variabel yang dipengaruhi oleh variabel lain disebut variabel terikat (*dependent variable*, yang selanjutnya dinotasikan dengan *Y*). Pada konteks hubungan kausal, variabel terikat adalah variabel yang dipengaruhi oleh variabel bebas. Hubungan kausal tersebut memungkinkan kita untuk memprediksi[1] nilai suatu variabel jika nilai variabel yang memengaruhinya diketahui. Bila bentuk hubungan dari dua variabel yang berhubungan kausal tersebut dapat ditemukan, dimungkinkan untuk mengetahui seberapa besar pengaruh perubahan satu unit variabel bebas terhadap variabel terikatnya. Dengan demikian, untuk

[1] Istilah lain yang lazim digunakan selain kata memprediksi adalah memperkirakan, meramalkan, mengestimasi

keperluan meramalkan, yang harus diketahui adalah *tren* (kecenderungan) hubungan dan persamaan matematis variabel-variabel yang menjadi perhatian. Pada prinsipnya, bila persamaan matematis dari variabel-variabel yang menjadi perhatian dapat diketahui, *tren* hubungan dari variabel-variabel tersebut juga dapat diketahui. Dengan pertimbangan *tren* hubungan dari variabel-variabel yang menjadi perhatian lebih mudah diperoleh daripada persamaan matematisnya, pemaparan dalam bab ini akan dimulai dari tahapan mengetahui *tren* hubungan dari variabel-variabel yang menjadi perhatian tersebut. Selanjutnya, diarahkan ke penetapan persamaan matematisnya. Guna memudahkan pemahaman pembaca, di bab ini kita akan membatasi perhatian pada masalah yang sederhana dengan hanya mengulas hubungan dua variabel *X* dan *Y*, yang mengarah pada teknik meramalkan nilai variabel *Y* sebagai suatu fungsi linear dari suatu variabel tunggal *X*. Meskipun dalam keseharian kita sering menjumpai adanya lebih dari satu variabel bebas yang memengaruhi satu variabel terikat[2], pada saat ini kita akan menunda dulu pembahasan terkait masalah ketergantungan variabel terikat pada beberapa variabel bebas tersebut dan memfokuskan perhatian pada hubungan dua variabel saja.

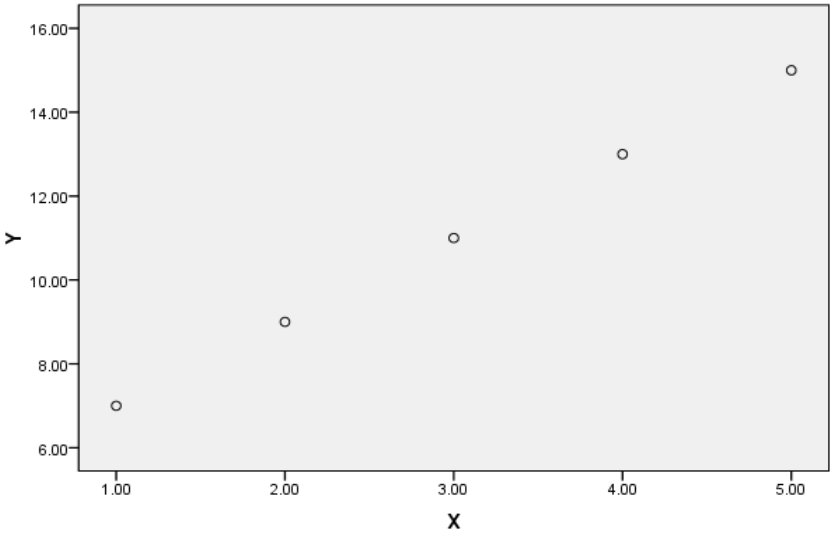
B. TREN HUBUNGAN

Untuk mengetahui *tren* hubungan dua variabel yang menjadi perhatian (selanjutnya dinotasikan dengan *X* dan *Y*), dapat digunakan diagram pencar (*scatter diagram)* yang merepresentasikan pasangan nilai *X* dan *Y* pada sebuah panel sumbu silang variabel-variabel *X* dan *Y*. Dari diagram pencar, dapat diperoleh gambaran berbagai kemungkinan hubungan *X* dan *Y*, yang pada dasarnya dapat diklasifikasikan sebagai hubungan *linear*, *curvilinear*, atau tidak ada hubungan.

Pada Gambar 1.1a hingga Gambar 1.1c ditunjukkan berbagai kemungkinan sederhana sebaran titik yang merepresentasikan pasangan nilai *X* dan *Y* pada sebuah panel sumbu silang variabel-variabel *X* dan *Y*.

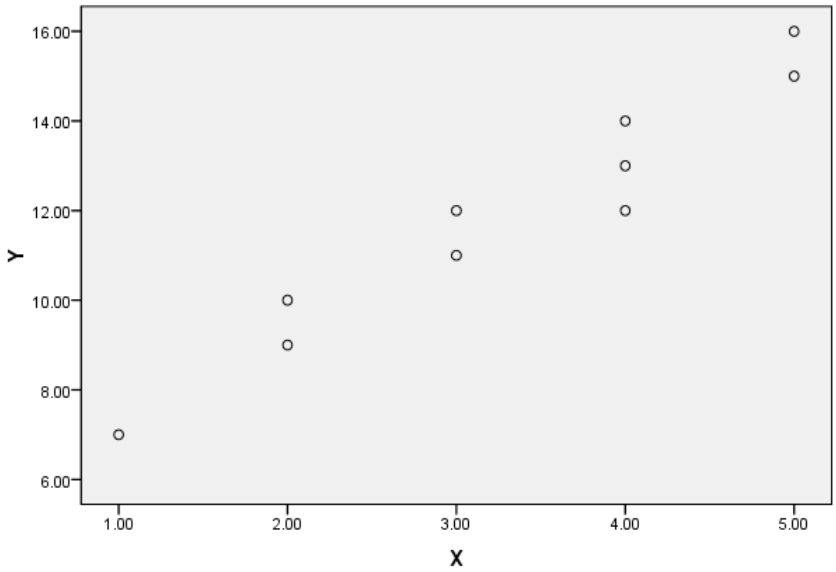
---

[2] Permasalahan adanya beberapa variabel bebas yang memengaruhi satu variabel terikat sering kali dikenal dengan permasalahan variabel majemuk atau *multivariabel*, misalnya besarnya belanja seseorang terhadap produk tertentu dipengaruhi oleh banyak variabel seperti halnya pendapatan per bulan, banyaknya anggota keluarga, umur, selera, warna produk, dan desain produk.



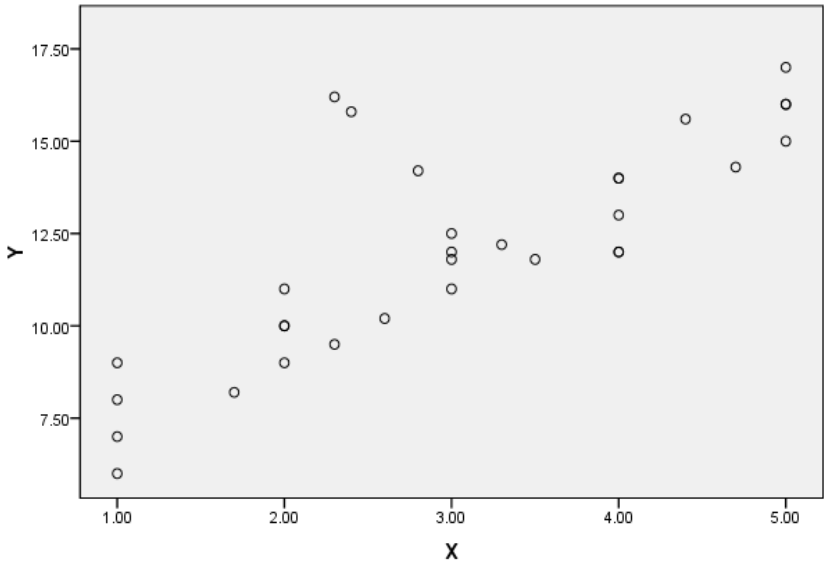
Gambar 1.1a.
Berbagai Kemungkinan Sederhana *Tren* Hubungan Pasangan Nilai *X* dan *Y*

Gambar 1.1a merepresentasikan pasangan nilai *X* dan *Y* yang tersebar secara urut sehingga dapat dengan mudah ditarik garis yang tepat melewati semua titik tersebut. Dari Gambar 1.1a, diketahui *tren* hubungan *X* dan *Y* yang positif karena nilai-nilai *X* dan nilai-nilai *Y* bergerak searah. Bila nilai *X* mengalami kenaikan, nilai *Y* juga mengalami kenaikan. Dengan demikian, dapat diprediksi bahwa bila nilai *X* bertambah besar, nilai *Y* juga akan mengalami peningkatan.



Gambar 1.1b

Gambar 1.1b memberi penjelasan bahwa pasangan titik *X* dan *Y* juga memiliki kecenderungan bahwa jika *X* meningkat, nilai *Y* besarnya meningkat. Dengan kata lain, didapati *tren* yang positif. Namun, pada Gambar 1.1b, kita tidak bisa membuat satu garis lurus yang secara tepat dapat melewati semua pasangan titik *X* dan *Y* sebagaimana pada Gambar 1.1a.



Gambar 1.1c

Gambar 1.1c menunjukkan sebaran titik pasangan nilai *X* dan *Y* yang lebih tersebar lagi jika dibandingkan Gambar 1.1a ataupun Gambar 1.1b. Pada Gambar 1.1c, titik-titik observasi pasangan *X* dan *Y* juga menunjukkan *tren* hubungan *X* dan *Y* yang positif. Namun, sebagaimana Gambar 1.1b, kita tidak bisa membuat satu garis lurus yang dapat secara tepat melewati semua pasangan titik *X* dan *Y*.

Pada kondisi titik-titik pasangan nilai *X* dan *Y* tersebut tersebar secara tidak urut serta tidak merata, perlu didapatkan suatu garis yang dapat mewakili pola sebaran pasangan nilai *X* dan *Y* tersebut secara baik. Untuk itu, kita perlu mengestimasi suatu garis yang secara geometris diperoleh dengan meletakkan garis secara tepat, yang dapat mewakili pola sebaran pasangan titik *X* dan *Y*. Garis tersebut semestinya melewati sebaran pasangan *X* dan *Y* yang menjadi perhatian. Garis tersebut sering kali disebut garis regresi *Y* pada *X*. Bila garis yang dimaksud dapat ditemukan, terhadap garis tersebut dapat dibuat suatu model matematis yang sangat berguna untuk memprediksi nilai *Y* apabila besarnya nilai variabel *X* diketahui atau ditetapkan. Untuk memperoleh garis regresi, kita harus mematut garis yang mampu merepresentasikan sebaran data aslinya.

C. MEMATUT GARIS REGRESI

Untuk mematut garis regresi yang menggambarkan hubungan kausal variabel *X* dengan variabel *Y*, dapat ditempuh tahapan-tahapan berikut.

1. Mengumpulkan data pasangan nilai variabel *X* dan variabel *Y* yang menjadi perhatian.
2. Membuat diagram pencar dengan melakukan plot titik-titik (X_1,Y_1), $(X_2,Y_2),\ldots,$ (X_n,Y_n) pada sebuah sistem koordinat *rectangular*. Lazimnya variabel bebas (*X*) di plot sumbu horizontal, sedangkan variabel terikat (*Y*) di plot sumbu vertikal.
3. Memvisualisasikan sebuah kurva mulus dari diagram pencar yang terbentuk. Kurva mulus yang dibentuk disebut kurva aproksimasi.

Harapan kita adalah memperoleh sebuah kurva mulus yang representatif. Dalam arti, kurva tersebut memiliki kemampuan untuk mewakili sebaran data yang dihadapi. Dengan kata lain, kurva yang terbentuk memiliki kriteria *good fit* (cocok) dengan data aslinya. Pada kurva yang *good fit,* bila kita mengetahui nilai *X*, kita dapat meramalkan nilai *Y* dengan hasil yang sama dengan *Y* aslinya. Bila nilai Y_i hasil prediksi dilambangkan dengan $\widehat{Y}_i$, untuk setiap nilai *X* akan didapatkan selisih *Y* asli dengan *Y* prediksi sama dengan 0. Untuk itu, kita akan mencermati kembali ilustrasi pada Gambar 1.1a hingga Gambar 1.1c. Dari Gambar 1.1a, dimungkinkan untuk memperoleh kurva yang memiliki kecocokan seratus persen dengan data yang dihadapi. Namun, untuk Gambar 1.1b dan Gambar 1.1c, tidak dimungkinkan mendapatkan kurva yang memiliki kecocokan seratus persen dengan data aslinya karena datanya menyebar. Dalam hal ini, bagaimanapun kita berupaya untuk menetapkan garis yang diharapkan mewakili data yang dihadapi tetap akan muncul *error*[3] (yang dinotasikan dengan e_i). Dalam hal ini, e_i merepresentasikan perbedaan nilai antara nilai *Y* yang sebenarnya (yang sering kali disebut sebagai *Y observed*) dengan nilai *Y* hasil prediksi (yang sering kali dinotasikan dengan $\widehat{Y}_i$).

$$e_i = Y_i - \widehat{Y}_i$$

[3] Sebutan lain bagi *error* adalah galat, *residual*, sisaan, atau kesalahan pengganggu.

Dalam konteks plot pasangan titik-titik *X* dan *Y* pada sistem koordinat *rectangular*, *error* adalah jarak vertikal dari nilai *Y* sebenarnya dengan nilai *Y* hasil prediksi (nilai Y_i hasil estimasi). Komponen *error* muncul karena garis estimasi yang dibentuk tidak mampu mengestimasi secara sempurna nilai *Y* amatan. Apabila semua nilai *error* sama dengan nol (sebagaimana terjadi pada Gambar 1.1a), semua titik pada diagram pencar akan terletak secara sempurna pada garis *tren*. Pada kondisi ini, nilai Y_i (*observed*) sama dengan nilai Y_i hasil prediksi sehingga akan diperoleh *error* sama dengan nol. Dalam praktiknya, kondisi ini jarang sekali terjadi karena lazimnya akan dijumpai *error*.

Dari paparan yang dikemukakan, kita berupaya memperoleh sebuah kurva mulus yang representatif, yang memiliki kemampuan mewakili sebaran data yang dihadapi, dan memenuhi kriteria *good fit* (kecocokan terbaik) terhadap data aslinya. Kurva yang memiliki kriteria *good fit* tentunya kurva yang memiliki keadaan total *erro*r yang kecil dengan penyimpangan yang minimum. Untuk memilih kurva atau garis dengan kecocokan terbaik, ada baiknya terlebih dahulu kita harus mendefinisikan apa yang kita maksud dengan 'kecocokan terbaik', yang secara intuisi layak, obyektif, dan dalam kondisi-kondisi tertentu akan menghasilkan peramalan terbaik bagi Y_i untuk nilai *X* tertentu (X_i). Definisi 'kecocokan terbaik' tercapai jika kita dapat mematut garis atau kurva yang dapat mewakili sebaran data yang dihadapi dengan kecocokan terbaik (*best fitting*). Jika kita dihadapkan pada n pasang observasi pasangan nilai *Y* dan *X*, kita ingin menetapkan persamaan garis dugaan sedemikian rupa agar diperoleh nilai prediksi *Y* yang memiliki nilai sedekat mungkin dengan nilai *Y* yang sebenarnya. Secara umum, mematut garis dapat dilakukan dengan empat cara:

1. dengan pandangan mata,
2. dengan mencari nilai minimum *error*,
3. dengan meminimumkan jumlah dari nilai mutlak *error*,
4. dengan meminimumkan jumlah kuadrat *error*.

1. Mematut Garis dengan Pandangan Mata

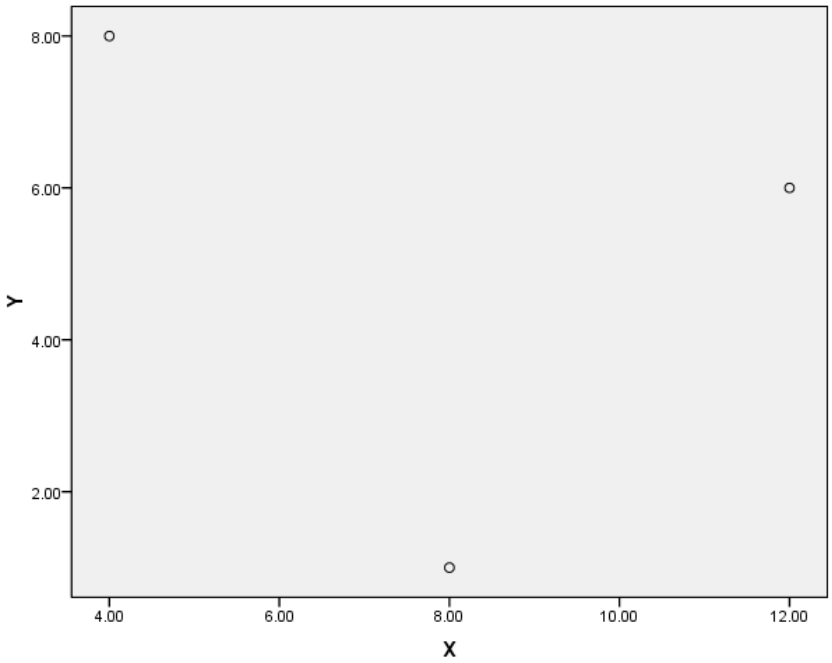
Prosedur memperoleh garis lurus dalam kecocokan terbaik (*best fitting*) dengan pandangan mata dilakukan melalui pencocokan secara visual sebuah garis dan serangkaian data. Dalam hal ini, kita berupaya menggerakkan penggaris sampai kita merasa telah mencapai suatu kondisi yang meminimumkan penyimpangan (deviasi) titik-titik dari calon garis yang akan

kita buat dari titik-titik hasil amatan. Kelemahan dari cara ini adalah dimungkinkannya dibuat lebih dari satu garis yang dapat dipatut dari sebaran titik pasangan *X* dan *Y* yang dimiliki. Pertanyaannya adalah dari berbagai garis yang mungkin dihasilkan tersebut, manakah yang terbaik. Garis manakah yang paling representatif? Dalam hal ini, tidak ada rujukan yang pasti sehingga pada umumnya cara mematut garis dengan pandangan mata tersebut hanya digunakan sebagai tahap awal eksplorasi *tren* hubungan variabel *X* dan *Y* (penjelasan lebih lanjut akan dipaparkan di bab 2).

Pada Gambar 1.2 terpampang tiga pasangan titik *X* dan *Y* yang diperoleh dari plot data pada Tabel 1.1.

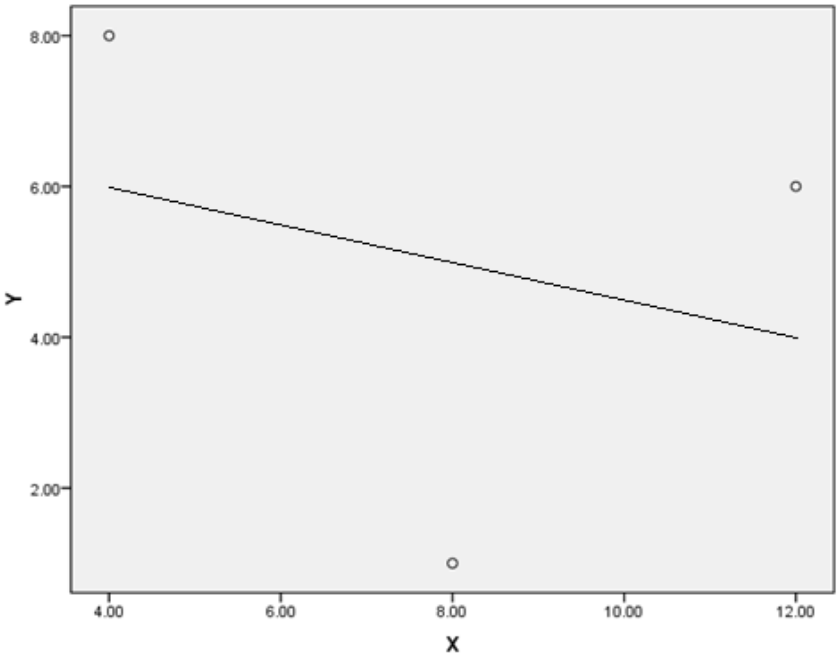
Tabel 1.1.
Pasangan Titik *X* dan *Y*

X	*Y*
4	8
8	1
12	6

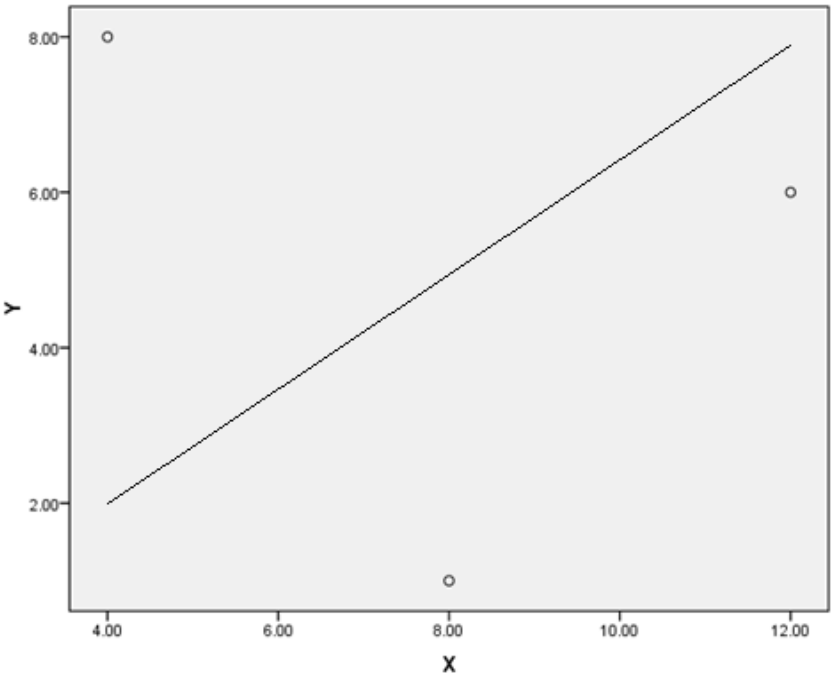


Gambar 1.2.
Plot Pasangan Titik *X* dan *Y* dari Data Tabel 1.1

Atas dasar pasangan titik yang sama tersebut, kita dapat mematut dua garis regresi sebagaimana ditunjukkan pada Gambar 1.2a dan Gambar 1.2b. Pertanyaannya adalah garis manakah yang memiliki kecocokan terbaik terhadap data pada Tabel 1.1. Dalam hal ini, kita membutuhkan kriteria tambahan sebagai rujukan untuk memutuskan garis mana yang lebih baik dalam mematut pasangan data pada Tabel 1.1.



Gambar 1.2a.
Garis Regresi A



Gambar 1.2b
Garis Regresi B

2. Mematut Garis dengan Mencari Nilai Minimum *Error*

Cara 2 merupakan salah satu alternatif yang diperhitungkan untuk menjawab kesulitan dari cara 1. Untuk tujuan ini, kita akan memilih fungsi regresi dengan cara sedemikian rupa sehingga dicapai jumlah *error* sekecil mungkin, dalam notasi matematis dinyatakan $\sum(Y_i - \widehat{Y}_i)$ minimum. Kalau nilai yang kita ramalkan untuk Y_i dilambangkan dengan $\widehat{Y}_i$, persamaan prediksi Y_i sebagai berikut.

$$\widehat{Y}_i = \widehat{\alpha} + \beta X_i$$

Dalam hal ini, $\hat{\alpha}$ merupakan estimasi untuk α dan $\hat{\beta}$ merupakan estimasi untuk β yang sebenarnya.[4] Besaran α dikenal sebagai *intercept*, yaitu titik potong garis regresi dengan sumbu *Y* pada saat *X* bernilai 0. Besaran β dikenal dengan *slope* atau kemiringan garis regresi yang menyatakan besarnya perubahan nilai *Y* setiap *X* berubah satu satuan. Melanjutkan contoh terdahulu dari pasangan data yang tertera pada Tabel 1.1, kita tampilkan *error* dari kedua garis yang dipatut tersebut dalam Tabel 1.2 dan Tabel 1.3.

Tabel 1.2.
Error dari Garis yang Dipatut pada Gambar 1.2a

X	*Y*	$\hat{Y}$	*Error* = Y- $\hat{Y}$
4	8	6	2
8	1	5	-4
12	6	4	2
			Total *error* = 0

Tabel 1.3.
Error dari Garis yang Dipatut pada Gambar 1.2b

X	*Y*	$\hat{Y}$	*Error* = Y- $\hat{Y}$
4	8	2	6
8	1	5	-4
12	6	8	-2
			Total *error* = 0

Didapati hasil total *error* kedua garis yang dipatut adalah 0. Berdasarkan cara dua, dapat dinyatakan bahwa kedua garis tersebut sama baiknya dalam mematut pasangan data pada Tabel 1.1. Tentunya, pembaca mempertanyakan bagaimana mungkin pasangan data yang sama diwakili oleh dua garis yang memiliki kecenderungan hubungan yang berlawanan, karena Gambar 1.2a

[4] $\hat{\alpha}$ dibaca α topi, α cap, α hat, atau prediksi α yang dalam hal ini merupakan estimasi dari α yang sebenarnya.

$\hat{\beta}$ dibaca β topi, β cap, β hat atau prediksi β yang dalam hal ini merupakan estimasi dari β yang sebenarnya

menunjukkan kecenderungan hubungan *X* dan *Y* negatif sedangkan Gambar 1.2b menunjukkan kecenderungan hubungan *X* dan *Y* positif.

Secara logika, cara mematut garis dengan mencari nilai minimum ***error*** tersebut masuk akal, namun cara tersebut mengandung kelemahan dan tidak menjamin diperolehnya sebuah garis estimasi yang baik. Dengan menggunakan cara meminimumkan jumlah *error*, setiap *error* yang muncul diberi bobot yang sama, tidak peduli apakah *error* yang muncul bernilai besar atau kecil. Padahal, semakin kecil nilai *error* berarti semakin dekat nilai prediksi terhadap nilai amatannya, sedangkan semakin besar *error* yang muncul berarti semakin jauh nilai prediksi dari nilai amatan yang sebenarnya.

Untuk memudahkan pemahaman pembaca, berikut akan diberikan contoh yang ekstrem. Tabel 1.4 mengilustrasikan adanya empat buah *error*, yaitu e_1, e_2, e_3, dan e_4. Bila *error* e_2 dan e_3 serta juga e_1 dan e_4 memperoleh bobot yang sama meskipun e_2 dan e_3 lebih dekat ke fungsi regresi daripada e_1 dan e_4, berarti semua *error* menerima tingkat penting yang sama tidak peduli seberapa dekat atau seberapa jauh terpencarnya observasi individual dari fungsi regresi. Pada kondisi demikian, jumlah dari seluruh *error* tersebut adalah nol meskipun e_1 dan e_4 terpencar lebih jauh di sekitar garis regresi daripada e_2 dan e_3. Kondisi ini terjadi karena *error* yang positif mengompensasi *error* yang negatif.

Tabel 1.4.
Gambaran *Error* yang Ekstrem

Nomor	*Error*	Nilai
1	e_1	15
2	e_2	1
3	e_3	-1
4	e_4	-15
		Total *error* = 0

3. Mematut Garis dengan Meminimumkan Jumlah dari Nilai Mutlak *Error*

Cara ketiga adalah meminimumkan jumlah dari nilai mutlak[5] *error* atau dalam notasi matematisnya meminimumkan $\sum \left| Y_i - \hat{Y}_i \right|$. Dengan cara ini, tanda positif dan negatif dari *error* dapat dihilangkan karena terhadap setiap *error* digunakan harga mutlaknya. Dengan demikian, besaran *error* yang positif tidak lagi mengompensasi *error* yang negatif sehingga pendekatan ini 'dianggap' dapat mengatasi permasalahan pada cara kedua yang telah dibahas sebelumnya. Sebagai ilustrasi, akan digunakan contoh pasangan data yang tertera pada Tabel 1.1. Tampilan nilai mutlak *error* dari kedua garis yang dipatut (Gambar 1.2a dan Gambar 1.2b) terlihat pada Tabel 1.5 dan Tabel 1.6.

Tabel 1.5.
Nilai Mutlak *Error* dari Gambar 1.2a

X	*Y*	$\hat{Y}$	*Error*=Y- $\hat{Y}$	Nilai mutlak *error*, |e|
4	8	6	2	2
8	1	5	-4	4
12	6	4	2	2
			Total *error* = 0	Total nilai mutlak *error* = 8

Tabel 1.6.
Nilai Mutlak *Error* dari Gambar 1.2b

X	*Y*	$\hat{Y}$	*Error* = Y- $\hat{Y}$	Nilai mutlak *error*, |e|
4	8	2	6	6
8	1	5	-4	4
12	6	8	-2	2
			Total *error* = 0	Total nilai mutlak *error* =12

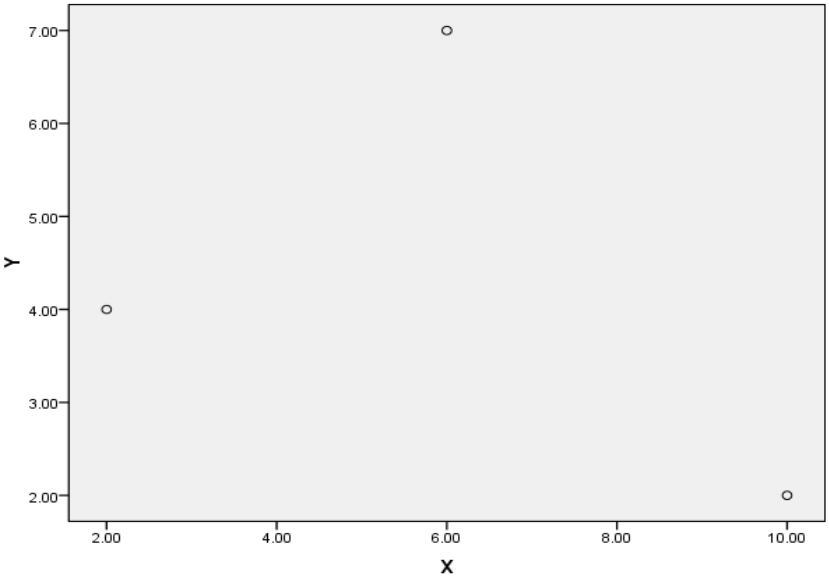
[5] Simbol untuk nilai mutlak adalah dua garis paralel | |.

Karena total nilai mutlak *error* garis regresi dari Gambar 1.2b lebih kecil dari total nilai mutlak *error* garis regresi dari Gambar 1.2b (8 < 12), dinyatakan bahwa garis regresi pada Gambar 1.2a lebih baik dalam mematut pasangan data pada Tabel 1.1 dibandingkan garis regresi pada Gambar 1.2b.

Hingga pada tahap ini, mungkin kita tergoda untuk menyimpulkan bahwa meminimumkan jumlah mutlak *error* merupakan kriteria terbaik untuk mendapatkan kurva dengan kepatutan yang baik. Namun, sebelum kita memutuskan, ada baiknya kita menguji lebih jauh keunggulan dari kriteria ini. Pada Gambar 1.3, ditampilkan dua diagram pencar yang identik dengan dua garis yang dipatut atas dasar tiga pasang data dari Tabel 1.7.

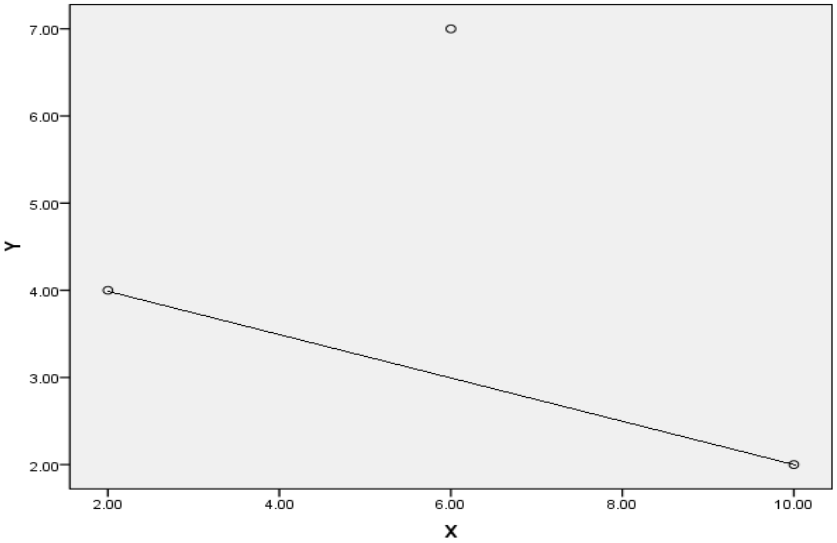
Tabel 1.7.
Ilustrasi Data untuk Metode Kuadrat Terkecil

X	*Y*
2	4
6	7
10	2



Gambar 1.3.
Diagram Pencar Data pada Tabel 1.7

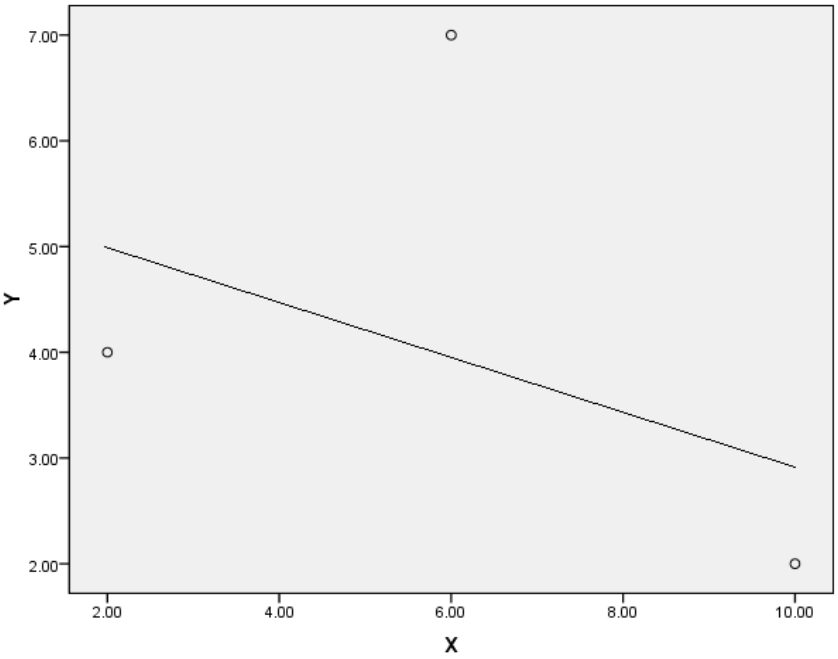
Selanjutnya, kita akan mematut dua garis regresi yang masing-masing tampak pada Gambar 1.3a dan Gambar 1.3b. Tabel 1.8 memperlihatkan *error* dari garis regresi pada Gambar 1.3a. Tabel 1.9 memperlihatkan *error* dari garis regresi pada Gambar 1.3b.



Gambar 1.3a

Tabel 1.8.
Error dari Garis Regresi pada Gambar 1.3a

X	***Y***	$\widehat{Y}$	***Error*=Y-** $\widehat{Y}$	**Nilai mutlak *error*, |e|**
2	4	4	0	0
6	7	3	4	4
10	2	2	0	0
			Total *error*=4	Total nilai mutlak *error*= 4



Gambar 1.3b

Tabel 1.9.
Error dari Garis Regresi pada Gambar 1.3b

X	*Y*	$\widehat{Y}$	*Error*=Y- $\widehat{Y}$	Nilai mutlak *error*, |e|
2	4	5	-1	1
6	7	4	3	3
10	2	3	-1	1
			Total *error*=1	Total nilai mutlak *error*= 5

Atas dasar informasi dari Tabel 1.8 dan Tabel 1.9 tersebut, didapati total nilai mutlak *error* Gambar 1.3a adalah 4, sedangkan total nilai mutlak *error* Gambar 1.3b adalah 5. Dengan demikian, berdasarkan kriteria 3, dinyatakan Gambar 1.3a memiliki kepatutan yang lebih baik dibandingkan Gambar 1.3b. Namun, secara intuisi tampak bahwa Gambar 1.3b sebenarnya lebih mampu mewakili sebaran data dibandingkan Gambar 1.3a karena garis pada Gambar 1.3b melalui bagian tengah dari titik-titik pasangan data. Bila dicermati, tampak bahwa Gambar 1.3a mengabaikan peranan titik observasi yang

terdapat di tengah, sedangkan Gambar 1.3b lebih memperhatikan peranan titik observasi yang terdapat di tengah. Gambar 1.3b dibuat dengan memperhatikan semua titik observasi, sedangkan Gambar 1.3a semata-mata dibuat mengarah pada pemenuhan cara 3, yaitu meminimumkan total nilai mutlak *error*. Pada kondisi ini, Gambar 1.3b tampak lebih mewakili titik-titik observasi dibanding Gambar 1.3a. Atas dasar paparan tersebut, cara 3 dianggap memiliki kelemahan karena cara ini tidak memperhatikan semua titik observasi, khususnya mengabaikan peranan titik observasi yang terdapat di tengah. Dari paparan dengan grafik, pembaca dapat mendeteksi kelemahan tersebut secara lebih jelas. Secara nalar, kita dapat mengatakan bahwa semakin jauh jarak titik amatan dari garis estimasi, semakin serius *error* yang muncul. Dalam hal ini, kita lebih menoleransi beberapa *error* dengan nilai mutlak yang kecil sebagaimana terlihat pada Gambar 1.3b daripada satu *error* dengan nilai mutlak yang besar sebagaimana terlihat pada Gambar 1.3a. Meskipun atas dasar kriteria cara 3, Gambar 1.3a memiliki total nilai mutlak *error* yang lebih kecil dibandingkan Gambar 1.3b. Dalam hal ini, kita cenderung menyatakan memilih Gambar 1.3b untuk mewakili sebaran titik observasi (Levin, Richard I & Rubin, David S, 1998).

4. Mematut Garis dengan Meminimumkan Jumlah Kuadrat *Error*

Tiga cara mematut garis regresi yang dikemukakan sebelumnya mengandung kelemahan. Berikut adalah cara yang merupakan penyempurnaan dari tiga cara sebelumnya, yaitu menggunakan metode kuadrat terkecil biasa (*method of ordinary least squares, OLS*)[6]. Metode kuadrat terkecil biasa (yang selanjutnya akan disebut sebagai metode kuadrat terkecil) dikemukakan oleh Carl Friedrich Gauss, seorang ahli matematika bangsa Jerman. Dengan asumsi-asumsi tertentu, metode kuadrat terkecil mempunyai beberapa sifat statistik yang sangat menarik yang membuatnya menjadi satu metode analisis regresi yang paling kuat (*powerful*) dan populer (Gujarati, 2009). Metode ini dianggap paling baik karena, di samping dapat mengatasi perbedaan tanda pada *error*, juga dapat menggambarkan semua peranan titik observasi yang tersebar. Dengan metode ini, masalah perbedaan tanda pada *error* dapat diatasi (mengatasi masalah pada cara 2), di samping juga metode ini memperhatikan semua titik observasi yang tersebar (mengatasi masalah pada cara 3). Dengan metode kuadrat terkecil, pemilihan garis dengan kecocokan terbaik dilakukan dengan meminimumkan jumlah

[6] Pada umumnya metode kuadrat terkecil biasa lazim dituliskan sebagai metode kuadrat terkecil.

kuadrat penyimpangan nilai-nilai *Y* yang diamati dari nilai-nilai *Y* yang diramalkan (*Y* prediksi). Fungsi regresi ditetapkan dengan cara sedemikian rupa sehingga jumlah kuadrat *error* adalah sekecil mungkin yang dalam notasi matematis dituliskan $\sum e_i^2 = \sum \left(Y_i - \hat{Y}_i\right)^2$ minimum. Dalam hal ini, e_i^2 adalah kuadrat *error*.

Dengan menguadratkan setiap e_i, metode kuadrat terkecil memberikan bobot yang lebih besar kepada *error* seperti e_1 dan e_4 daripada *error* e_2 dan e_3 pada contoh di Tabel 1.4. Dengan menggunakan kriteria kuadrat terkecil, persamaan prediksi yang menghasilkan nilai jumlah *error* terkecil (pada cara 2) belum tentu menghasilkan jumlah kuadrat *error* yang terkecil. Kalau setiap nilai *error* tersebut kita kuadratkan dan kemudian kita jumlahkan, hasilnya disebut jumlah kuadrat *error*. Makin kecil nilai jumlah kuadrat *error* yang diperoleh berarti garis *trend* yang terbentuk makin mendekati diagram pencarnya.

Guna memudahkan pemahaman pembaca, kita akan melanjutkan contoh terdahulu dari pasangan data yang tertera pada Tabel 1.7. Kita akan menampilkan nilai kuadrat *error* dari data pada Tabel 1.8 dan Tabel 1.9 dalam Tabel 1.10 dan Tabel 1.11.

Tabel 1.10.
Kuadrat *Error* dari Garis Regresi pada Gambar 1.3a

X	*Y*	$\hat{Y}$	*Error*=*Y*-$\hat{Y}$	Nilai kuadrat *error*, $(e)^2$
2	4	4	0	0
6	7	3	4	16
10	2	2	0	0
			Total *error*=4	Total nilai kuadrat *error*= 16

Tabel 1.11.
Kuadrat *Error* dari Garis Regresi pada Gambar 1.3b

X	*Y*	$\hat{Y}$	*Error*=*Y*-$\hat{Y}$	Nilai kuadrat *error*, $(e)^2$
2	4	5	-1	1
6	7	4	3	9
10	2	3	-1	1
			Total *error*=1	Total nilai kuadrat *error*= 11

Dengan cara 4, terbukti bahwa kecenderungan kita sewaktu menyatakan memilih Gambar 1.3b untuk mewakili sebaran titik observasi dibandingkan Gambar 1.3a ternyata berdasar. Total nilai kuadrat *error* dari Gambar 1.3a (16) > Total nilai mutlak *error* Gambar 1.3b (11).

Dari ilustrasi yang dikemukakan, diketahui bahwa penggunaan metode kuadrat terkecil dalam mematut garis regresi adalah yang terbaik dibandingkan tiga cara yang dikemukakan sebelumnya. Dengan demikian, dalam mematut persamaan garis regresi yang *good fit* akan digunakan metode kuadrat terkecil. Persamaan regresi yang didasarkan metode kuadrat terkecil biasa (*method of ordinary least squares, OLS)* lazim disebut regresi OLS.

Garis regresi yang dibentuk atas dasar metode kuadrat terkecil yang mengaproksimasi himpunan titik-titik (X_i,Y_i) memiliki persamaan $Y_i = \alpha + \beta X_i + \varepsilon_i$. Dalam hal ini, kita dapat menyatakan bahwa komponen *error* merupakan kesalahan statistik (*statistical error*). Komponen *error* ini merupakan suatu variabel acak (*random variabel*) yang merepresentasikan kegagalan model yang dibentuk dalam mencocokkan diri sepenuhnya terhadap data yang dihadapi. Dalam hal ini, ε_i adalah suatu variabel acak yang independen dan memiliki nilai mean = 0 dan variance = σ^2. Pada dasarnya, distribusi Y_i dan ε_i adalah identik, hanya bedanya nilai mean ε_i terletak pada 0.

LATIHAN

Untuk memperdalam pemahaman Anda mengenai materi di atas, kerjakanlah latihan berikut!

1) Pada tabel berikut, telah diteliti 10 observasi terkait empat variabel, yaitu piutang, biaya operasi, laba, dan arus kas.

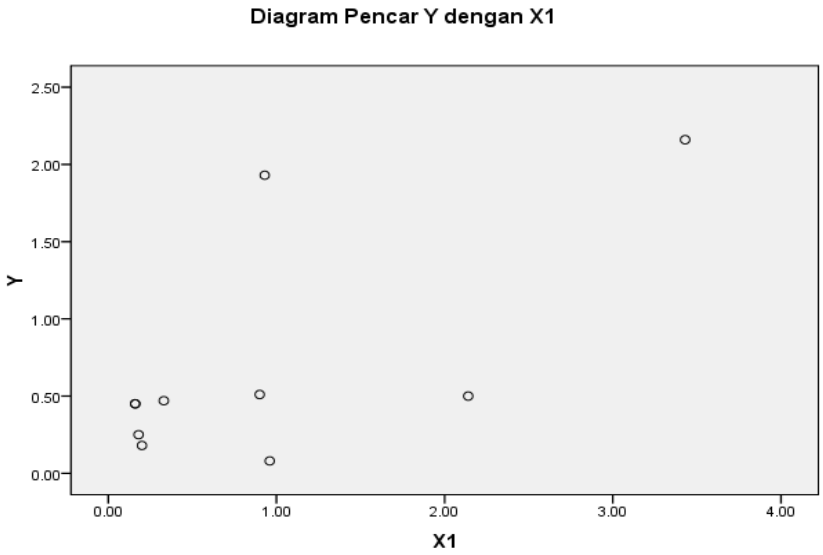
Piutang (*Y*)	Biaya operasi (X_1)	Laba (X_2)	Arus kas (X_3)
0.47	0.33	1.51	1.58
0.18	0.20	0.04	0.02
2.16	3.43	0.38	0.08
1.93	0.93	2.04	3.29
0.08	0.96	2.30	0.88
0.45	0.16	0.10	0.89
0.50	2.14	1.59	0.46
0.51	0.90	0.89	0.80
0.45	0.16	0.12	0.88
0.25	0.18	2.30	1.68

Bila dibuat diagram pencar *Y* dengan masing-masing *X*, tentukan indikasi *trend* hubungan variabel *Y* dengan masing-masing *X*!

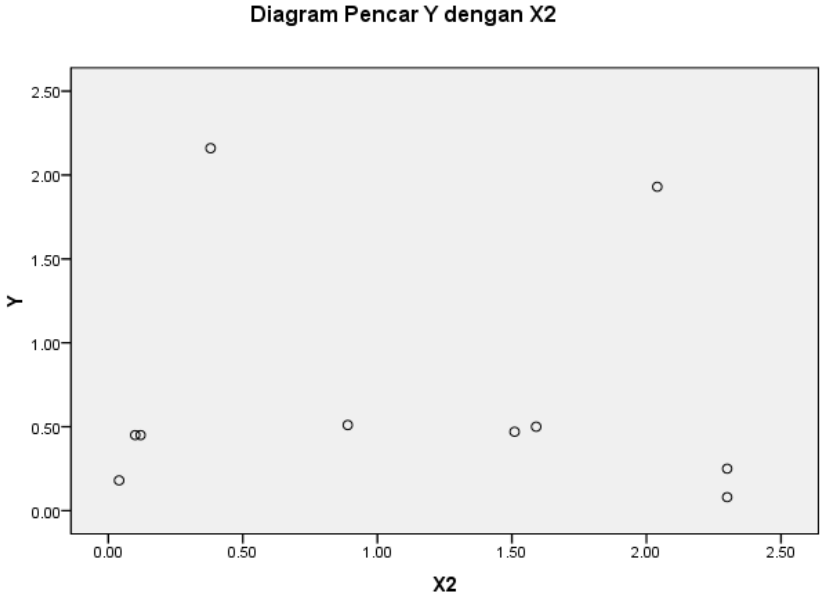
2) Pada waktu melakukan estimasi garis regresi dari pasangan pengamatan *X* dan *Y*, muncul komponen *error*. Jelaskan penyebab munculnya komponen *error* tersebut!
3) Jelaskan pengertian estimasi yang *underestimate* dan yang *overestimate*!
4) Jelaskan kelemahan dari cara mematut garis regresi dengan pandangan mata!
5) Jelaskan kelemahan dari cara mematut garis regresi dengan mencari nilai minimum *error*!
6) Jelaskan kelemahan dari cara mematut garis regresi dengan meminimumkan jumlah dari nilai mutlak *error*!

Petunjuk Jawaban Latihan

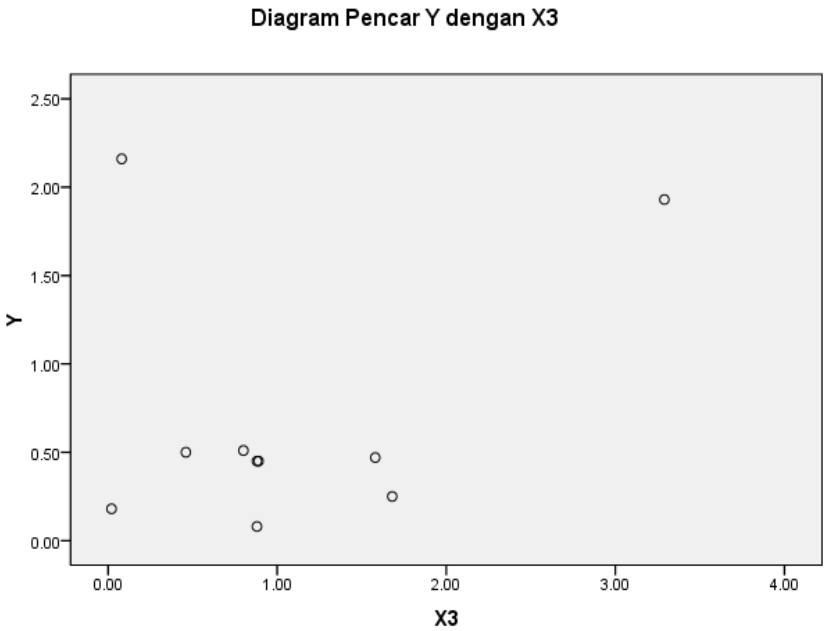
1) *Trend* hubungan variabel *Y* dengan masing-masing *X*.



Indikasi *trend* hubungan Y dengan X_1 adalah positif.



Indikasi *trend* hubungan Y dengan X_2 adalah negatif.



Indikasi *trend* hubungan Y dengan X_3 adalah positif.

Dari ketiga diagram pencar yang terbentuk, tampak bahwa *trend* hubungan Y dengan X_1 lebih jelas menunjukkan pola fungsi linier dibandingkan *trend* hubungan Y dengan X_2 ataupun Y dengan X_3.

2) Komponen *error* muncul karena garis estimasi yang dibentuk tidak mampu mengestimasi secara sempurna nilai Y amatan.
3) Pada kondisi nilai Y_i (*observed*) lebih besar dari nilai Y_i hasil prediksi, akan diperoleh *error* positif. Dalam hal ini, berarti kita menaksir nilai Y_i terlalu rendah (*underestimate*). Sebaliknya, bila nilai Y_i (*observed*) lebih kecil dari nilai Y_i hasil prediksi, akan diperoleh *error* negatif. Dalam hal ini, berarti kita menaksir nilai Y_i terlalu tinggi (*overestimate*).
4) Kelemahan dari cara mematut garis regresi dengan pandangan mata adalah dimungkinkannya dibuat lebih dari satu garis yang dapat dipatut dari sebaran titik pasangan X dan Y yang dimiliki. Dalam hal ini, tidak ada rujukan yang pasti sehingga pada umumnya cara mematut garis dengan pandangan mata tersebut hanya digunakan sebagai tahap awal eksplorasi *trend* hubungan variabel X dan Y.
5) Kelemahan dari cara mematut garis regresi dengan mencari nilai minimum *error* adalah setiap *error* yang muncul diberi bobot yang sama dan tidak peduli apakah *error* yang muncul bernilai besar atau kecil.

Semua *error* menerima tingkat penting yang sama dan tidak peduli seberapa dekat atau seberapa jauh terpencarnya observasi individual dari fungsi regresi. Padahal, semakin kecil nilai *error* berarti semakin dekat nilai prediksi terhadap nilai amatannya, sedangkan semakin besar *error* yang muncul berarti semakin jauh nilai prediksi dari nilai amatan yang sebenarnya. Kelemahan ini memungkinkan besaran *error* yang positif mengompensasi *error* yang negatif.

6) Cara ini tidak memperhatikan semua titik observasi, khususnya mengabaikan peranan titik observasi yang terdapat di tengah.

RANGKUMAN

1. Dalam kondisi sehari-hari, kita sering menjumpai hubungan kausal suatu variabel dengan satu atau lebih variabel lainnya. Hubungan kausal tersebut memungkinkan kita untuk memprediksi (meramalkan atau mengestimasi) nilai suatu variabel jika nilai variabel yang memengaruhinya diketahui. Bila bentuk hubungan dari dua variabel yang berhubungan kausal tersebut dapat ditemukan, dimungkinkan untuk mengetahui seberapa besar pengaruh perubahan satu unit variabel bebas terhadap variabel terikatnya.
2. Upaya untuk mendapatkan bentuk hubungan dari dua variabel yang berhubungan kausal dapat ditempuh dengan menemukan sebuah kurva mulus yang representatif, yang memiliki kemampuan mewakili sebaran data yang dihadapi, dan memenuhi kriteria *good fit* (kecocokan terbaik) terhadap data aslinya. Kurva yang memiliki kriteria *good fit* tentunya kurva yang memiliki keadaan total *erro*r yang kecil dengan penyimpangan yang minimum. Kurva tersebut disebut garis regresi.
3. Mematut garis dapat dilakukan dengan empat cara:
   a. dengan pandangan mata,
   b. dengan mencari nilai minimum *error*,
   c. dengan meminimumkan jumlah dari nilai mutlak *error* $\sum \left| Y_i - \hat{Y}_i \right|$,
   d. dengan meminimumkan jumlah kuadrat *error.*
4. Dari empat cara tersebut, mematut garis dengan meminimumkan jumlah kuadrat *error* dianggap sebagai yang terbaik sehingga garis regresi yang dicari didasarkan pada metode ini yang sering kali dikenal dengan metode OLS.

TES FORMATIF 1

Pilihlah satu jawaban yang paling tepat!

1) Upaya untuk mendapatkan bentuk hubungan dari dua variabel yang berhubungan kausal dapat ditempuh dengan menemukan sebuah kurva mulus yang representatif, yang memiliki kemampuan mewakili sebaran data yang dihadapi, dan memenuhi kriteria *good fit* (kecocokan terbaik) terhadap data aslinya. Kurva yang memiliki kriteria *good fit* tentunya kurva yang memiliki keadaan ….
   A. total *erro*r yang kecil dengan penyimpangan yang minimum
   B. kurva tersebut berbentuk lurus
   C. kurva tersebut memiliki pijakan nalar yang kuat
   D. menghasilkan total *error* yang wajar

2) Mematut garis dapat dilakukan dengan cara berikut, *kecuali* ....
   A. dengan pertimbangan bisnis
   B. dengan mencari nilai minimum *error*
   C. dengan meminimumkan jumlah dari nilai mutlak *error*
   D. dengan meminimumkan jumlah kuadrat *error*

3) Dari berbagai kemungkinan mematut garis regresi yang *best fitting*, yang dianggap paling baik adalah ….
   A. dengan pandangan mata
   B. dengan mencari nilai minimum *error*
   C. dengan meminimumkan jumlah dari nilai mutlak *error* $\sum \left| Y_i - \hat{Y}_i \right|$
   D. dengan meminimumkan jumlah kuadrat *error*

4) Hubungan kausal memungkinkan kita untuk memprediksi nilai suatu variabel jika nilai variabel yang memengaruhinya ….
   A. tidak diketahui
   B. tidak bisa dikontrol
   C. diketahui
   D. memiliki *error*

5) Komponen *error* muncul karena garis estimasi yang dibentuk ….
   A. tidak mampu mengestimasi secara sempurna nilai *Y* amatan
   B. mampu mengestimasi secara sempurna nilai *Y* amatan
   C. tidak mampu mengestimasi secara sempurna nilai *Y* prediksi
   D. mampu mengestimasi secara sempurna nilai *Y* prediksi

6) Secara logika, cara mematut garis dengan mencari nilai minimum *error* tersebut masuk akal, tetapi cara tersebut mengandung kelemahan dan tidak menjamin diperolehnya sebuah garis estimasi yang baik karena ….
   A. dengan menggunakan cara meminimumkan jumlah *error*, setiap *error* yang muncul diberi bobot yang tidak sama dan tidak peduli apakah *error* yang muncul bernilai besar atau kecil
   B. dengan menggunakan cara meminimumkan jumlah *error*, setiap *error* yang muncul diberi bobot sesuai dengan besarnya *error* yang muncul
   C. dengan menggunakan cara meminimumkan jumlah *error*, setiap *error* yang muncul diberi bobot yang sama dan tidak peduli apakah *error* yang muncul bernilai besar atau kecil
   D. dengan menggunakan cara meminimumkan jumlah *error*, *error* yang kecil diberi bobot besar, sedangkan *error* yang besar diberi bobot kecil

7) Secara nalar, kita dapat mengatakan bahwa semakin jauh jarak titik amatan dari garis estimasi, ….
   A. semakin serius *error* yang muncul
   B. semakin bagus *good fit* yang diperoleh
   C. semakin kecil *error* yang muncul
   D. semakin dekat nilai estimasi terhadap nilai amatan sebenarnya

8) Pada PT Gembira Sentosa, biaya iklan produknya berhubungan dengan volume penjualannya. Semakin sering PT Gembira Sentosa menayangkan iklan produknya di berbagai media masa, makin besar biaya iklan yang dikeluarkan oleh perusahaan, tetapi makin meningkat pula volume penjualan produknya. Dalam hal ini, sifat hubungan variabel biaya iklan dan volume penjualan PT Gembira Sentosa adalah ….
   A. tidak berhubungan
   B. positif
   C. negatif
   D. tidak mungkin diketahui

9) Variabel *X* dan variabel *Y* dinyatakan memiliki konteks hubungan kausal, bila ….
   A. besarnya variabel *X* berubah, besarnya variabel *Y* akan terpengaruh
   B. besarnya variabel *X* berubah, besarnya variabel *Y* tidak akan terpengaruh

C. besarnya variabel *X* berubah, besarnya variabel *Y* terkadang akan terpengaruh terkadang tidak terpengaruh
D. besarnya variabel *X* berubah, besarnya variabel *Y* akan terpengaruh hanya pada nilai-nilai tertentu saja dari variabel *X*

10) Diagram pencar (*scatter diagram)* adalah diagram yang merepresentasikan pasangan nilai *X* dan *Y* pada sebuah panel sumbu silang variabel-variabel *X* dan *Y*. Dari diagram pencar dapat diperoleh gambaran berbagai kemungkinan hubungan *X* dan *Y*, yang pada dasarnya dapat diklasifikasikan sebagai hubungan, *kecuali* ....
A. *linear*
B. *curvilinear*
C. tidak ada hubungan
D. volatil

Cocokkanlah jawaban Anda dengan Kunci Jawaban Tes Formatif 1 yang terdapat di bagian akhir modul ini. Hitunglah jawaban yang benar. Kemudian, gunakan rumus berikut untuk mengetahui tingkat penguasaan Anda terhadap materi Kegiatan Belajar 1.

$$\text{Tingkat penguasaan} = \frac{\text{Jumlah Jawaban yang Benar}}{\text{Jumlah Soal}} \times 100\%$$

Arti tingkat penguasaan: 90 - 100% = baik sekali
80 - 89% = baik
70 - 79% = cukup
< 70% = kurang

Apabila mencapai tingkat penguasaan 80% atau lebih, Anda dapat meneruskan dengan Kegiatan Belajar 2. **Bagus!** Jika masih di bawah 80%, Anda harus mengulangi materi Kegiatan Belajar 1, terutama bagian yang belum dikuasai.

KEGIATAN BELAJAR 2

Konsep dalam Analisis Regresi

Fokus dalam Kegiatan Belajar 2 adalah konsep dalam analisis regresi. Pada bagian ini, akan dipaparkan konsep-konsep terkait seperti halnya konsep populasi dan sampel dalam kaitannya dengan garis regresi serta peran komponen *error*.

A. KONSEP YANG DIGUNAKAN DALAM ANALISIS REGRESI

Pada ulasan sebelumnya, berulang kali disinggung tentang garis regresi sebagai garis yang mewakili sebaran data aslinya. Garis regresi yang diperoleh merupakan bahan dasar dalam melakukan analisis regresi. Dewasa ini, analisis regresi merupakan alat analisis yang sangat populer yang sering kali digunakan untuk mengetahui pengaruh satu atau lebih variabel terhadap variabel lain. Sebagaimana disebutkan sebelumnya, variabel yang memengaruhi variabel lain disebut sebagai variabel bebas (yang sering kali dinotasikan sebagai variabel *X*), sedangkan variabel yang dipengaruhi oleh variabel lain disebut variabel terikat (yang sering kali dinotasikan sebagai variabel *Y*). Di samping sebutan variabel bebas untuk variabel *X* dan variabel terikat untuk variabel *Y*, didapati variasi penyebutan yang lazim digunakan sebagaimana ditayangkan pada Tabel 1.12.

Tabel 1.12.
Variasi Penyebutan Variabel *X* dan *Y* pada Analisis Regresi

Sebutan Variabel *X*	Sebutan Variabel *Y*
Variabel independen (*independent variable*)	Variabel dependen (*dependent variable*)
Variabel tak gayut	Variabel gayut
Variabel bebas	Variabel terikat
Variabel bebas	Variabel tak bebas
Variabel yang bisa dikontrol	Variabel respons
Variabel peramal	Variabel yang diramalkan
Variabel yang menjelaskan	Variabel yang dijelaskan
EXplanatory variable	*EXplained variable*
Predictor	*Predictand*
Variabel yang meregresi (*regressor*)	Variabel yang diregresi (*regressand*)
Perangsang atau variabel kendali (*stimulus or control variable*)	Variabel tanggapan (*response variable*)

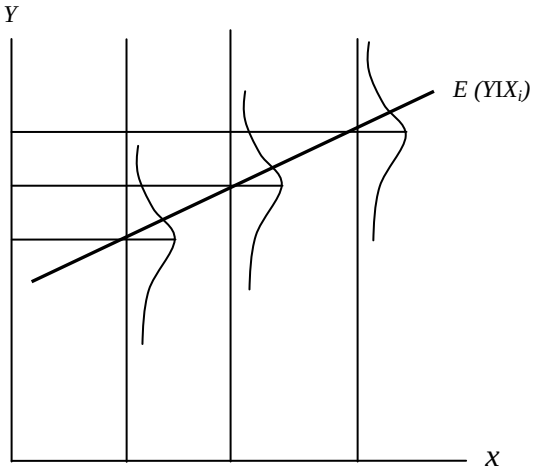
Istilah regresi diperkenalkan oleh Francis Galton berdasarkan telaahannya tentang sifat-sifat keturunan. Galton mendapati fenomena bahwa meskipun ada kecenderungan orang tua yang berbadan tinggi dan mempunyai anak yang berbadan tinggi, lalu orang tua yang berbadan pendek mempunyai anak yang berbadan pendek, distribusi tinggi suatu populasi tidak berubah secara mencolok dari generasi ke generasi. Ia mengungkapkan bahwa ayah yang jangkung akan cenderung memiliki anak yang jangkung pula, tetapi secara rata-rata tidak sejangkung ayahnya. Begitu pula ayah yang pendek akan cenderung memiliki anak yang pendek, tetapi secara rata-rata tidak sependek orang tuanya. Dengan demikian, ada kecenderungan bagi rata-rata tinggi anak-anak dengan orang tua yang memiliki tinggi tertentu untuk bergerak atau mundur **(*regress*)** ke arah ketinggian rata-rata seluruh populasi, sedangkan distribusi tinggi suatu populasi tidak berubah secara mencolok dari generasi ke generasi. Temuan Galton tersebut diperkuat oleh Karl Pearson yang mengumpulkan lebih dari seribu catatan tinggi badan anggota kelompok keluarga. Karl Pearson mendapati fenomena yang mirip dengan temuan Galton bahwa rata-rata tinggi badan anak laki-laki dari kelompok ayah yang tinggi adalah kurang dari tinggi badan ayah mereka. Sementara itu, rata-rata tinggi badan anak lelaki dari kelompok ayah yang pendek adalah lebih tinggi dari tinggi badan ayah mereka. Dengan demikian, temuan mengarah pada mundurnya tinggi badan anak laki-laki yang tinggi ataupun yang pendek ke arah rata-rata tinggi badan semua laki-laki. Galton menyebut kondisi ini sebagai kemunduran ke arah sedang.

Secara umum dapat dikemukakan bahwa penafsiran regresi dewasa ini berbeda dari penafsiran regresi menurut Galton. Dewasa ini, analisis regresi berguna dalam menelaah hubungan dua variabel atau lebih, terutama untuk menelusuri pola hubungan yang modelnya belum diketahui dengan sempurna sehingga dalam terapannya lebih bersifat eksploratif. Saat ini, analisis regresi lebih berkenaan dengan studi ketergantungan suatu variabel (variabel tak bebas) pada variabel lainnya (variabel bebas) dengan tujuan memprediksi atau meramalkan nilai rata-rata hitung (mean) atau rata-rata populasi variabel tak bebas. Bila Galton tertarik untuk memikirkan mengapa didapati kestabilan pada distribusi tinggi suatu populasi, pandangan modern tentang regresi lebih memikirkan bagaimana rata-rata tinggi anak laki-laki berubah dengan melihat tinggi ayah mereka. Dalam pandangan modern, perhatian diberikan pada meramalkan tinggi badan rata-rata anak laki-laki dengan mengetahui tinggi badan ayah mereka.

1. Konsep Populasi dan Sampel dalam Kaitannya dengan Garis Regresi

Bila kita melakukan telaah berulang kali pada lingkup populasi yang melibatkan lebih dari satu responden dengan tingkat penghasilan bulanan yang sama (X), didapati besarnya belanja per bulan (Y) dari responden-responden yang diamati tidak selalu sama. Fenomena menunjukkan nilai Y berfluktuasi dan berkelompok di sekitar nilai sentral tertentu. Dalam hal ini, didapati banyak kemungkinan nilai Y yang membentuk suatu populasi dengan suatu pola distribusi Y tertentu. Distribusi peluang Y dengan nilai X tertentu disimbolkan sebagai $p[Y|X]$. Distribusi peluang yang serupa juga terjadi pada nilai X tertentu yang lain. Untuk setiap nilai X tertentu, didapati berbagai nilai Y yang membentuk suatu populasi atau untuk setiap nilai X tertentu akan terdapat suatu distribusi peluang $p[Y|X]$ tertentu.

Bila nilai X disusun secara urut pada sumbu horizontal dan pada kondisi untuk masing-masing nilai X terdapat distribusi peluang $p[Y|X]$, akan didapati berbagai kemungkinan susunan distribusi tersebut. Pada kondisi nilai Y, ditampilkan pada sumbu vertikal. Salah satu kemungkinan bentuk susunan distribusi populasi Y adalah sebagaimana tampak pada Gambar 1.4.



Gambar 1.4.
Salah satu kemungkinan distribusi peluang $p[Y|X]$

Dengan adanya kemungkinan-kemungkinan tersebut, dapat diperkirakan sulit menganalisis populasi yang demikian. Agar masalah tersebut dapat diatasi, diperlukan beberapa anggapan (asumsi) mengenai aturan-aturan dari populasi. Asumsi-asumsi tersebut sebagai berikut.

a. Distribusi peluang p[$Y_i|X_i$] memiliki varian yang sama, yang besarnya adalah σ^2 pada setiap X_i .

b. Nilai mean (rata-rata) *Y* yang diharapkan dan yang dinotasikan dengan E(Y_i)[7] terletak pada sebuah garis lurus yang disebut sebagai garis regresi yang sebenarnya (garis regresi populasi).

$$E(Y_i) = \mu_i = \alpha + \beta X_i$$

Sebagaimana dinyatakan sebelumnya, parameter populasi α menggambarkan *intercept* dan β menggambarkan arah garis (*slope*). Parameter-parameter tersebut akan diestimasi dengan informasi-informasi dari sampel (yaitu a sebagai penduga[8] α dan b sebagai penduga β).

c. Variabel-variabel acak Y_i adalah bebas secara statistik satu sama lain. Artinya, bila salah satu nilai pengamatan di dalam Y_i berubah, perubahan Y_i tersebut tidak akan memengaruhi Y_i yang lainnya. Sebagai gambaran, jika hasil nilai Y_1 kecil, nilai Y_1 yang kecil tersebut tidak akan memengaruhi nilai Y_2 untuk menjadi kecil atau sebaliknya. Dengan kata lain, Y_2 tidak terpengaruh oleh Y_1. Demikian pula untuk pengamatan-pengamatan yang lain juga tidak saling terpengaruh dan memengaruhi satu sama lain.

Dari pembahasan sebelumnya, diperoleh gambaran bahwa tiap rata-rata bersyarat E($Y|X_i$) merupakan fungsi dari X_i atau dalam notasi matematis dituliskan E($Y|X_i$)= f(X_i)[9]. Dalam konteks ini, f(X_i) menggambarkan suatu fungsi dari variabel yang menjelaskan, yaitu X_i, sedangkan E($Y|X_i$) merupakan fungsi linear dari X_i. Persamaan E($Y|X_i$)= f(X_i) dikenal sebagai fungsi regresi populasi (*population regression function=PRF*) atau singkatnya disebut sebagai regresi populasi. Fungsi tersebut hanya

[7] Notasi E(Y_i) dibaca nilai harapan dari Y yang ke i.

[8] Penduga merupakan istilah lain dari *estimator*.

[9] Catatan: rata-rata bersyarat tidak harus selalu terletak pada garis lurus. Rata-rata tersebut bisa saja terletak pada suatu kurva. Pembahasan tentang hal ini akan dikemukakan di bab selanjutnya.

menyatakan bahwa rata-rata (populasi) dari distribusi Y untuk X_i tertentu berhubungan secara fungsional dengan X_i. Dengan kata lain, nilai rata-rata populasi bervariasi bersama dengan X.

Dalam keadaan nyata, kita tidak mempunyai seluruh populasi yang tersedia untuk pengujian. Oleh karena itu, bentuk fungsional fungsi regresi populasi merupakan pertanyaan empiris meskipun dalam kasus khusus teorinya bisa menyatakan sesuatu. Sebagai contoh, seorang ekonom mungkin menduga bahwa belanja konsumsi berhubungan secara linear dengan pendapatan. Karena itu, sebagai pendekatan awal, dapat diasumsikan bahwa fungsi regresi populasi $E(Y|X_i)$ merupakan fungsi linear dari X_i dalam persamaan berikut.

$$E(Y|X_i) = \alpha + \beta X_i \qquad (1)$$

Besaran α dan β merupakan parameter yang tidak diketahui besarnya, tetapi tetap (*fiXed*) dan dikenal sebagai koefisien regresi. Dalam hal ini, α adalah *intercept* dan β merupakan koefisien kemiringan regresi (*slope coefficient*). Persamaan (1) dikenal sebagai fungsi regresi populasi linier atau regresi populasi linier. Dalam pembahasan selanjutnya, istilah regresi, persamaan regresi, dan model regresi akan digunakan dengan arti yang sama. Dalam analisis regresi, fokus kita adalah untuk menaksir fungsi regresi populasi pada persamaan (1), yaitu menaksir nilai α dan β yang tidak diketahui berdasarkan pengamatan atas X dan Y.

Dengan membatasi pembahasan kita pada populasi nilai Y yang berhubungan dengan X yang tetap (*fiXed*), kita telah dengan sengaja menghindari permasalahan teknik penarikan sampel. Karena hampir dalam semua situasi praktis lebih memungkinkan mengobservasi sampel daripada populasi, lazimnya yang kita miliki adalah suatu sampel nilai-nilai Y yang berhubungan dengan beberapa X yang tetap. Pada kondisi ini, kita bisa mengembangkan konsep fungsi regresi sampel (*sample regression function* = *SRF*) untuk menyatakan garis regresi sampel yang ditujukan untuk mewakili garis regresi populasi dengan menuliskan kembali persamaan (1) menjadi berikut ini.

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i . \qquad (2)$$

Keterangan:

$\hat{Y}_i$ = penaksir (estimator) $E(Y|X_i)$ dibaca Y 'hat' atau Y 'cap' atau Y 'topi'

$\hat{\alpha}$ = penaksir α

$\hat{\beta}$ = penaksir β

Persamaan (1) dan persamaan (2) menyatakan fungsi regresi yang *deterministic*. Kalau sebuah nilai *X* disubstitusikan ke dalam persamaan, nilai *Y* menjadi tertentu dan tidak ada ruang yang disediakan untuk kesalahan (*error*). Pada kondisi ini, nilai taksiran *Y* akan sama persis dengan nilai *Y* aslinya. Dalam kenyataannya, meskipun berhadapan dengan data populasi, garis regresi yang dibentuk lazimnya tidak mampu mewakili sebaran data aslinya secara persis sama. Pembaca dapat melihat bahwa titik-titik yang dipetakan ada yang menyimpang dari garis regresi yang dipatut, baik menyimpang secara positif maupun negatif. Penyimpangan yang terjadi tampaknya berpola acak (*random*). Dengan demikian, sebagaimana dipaparkan sebelumnya akan muncul komponen *error*. Secara nalar lebih masuk akal bila fungsi regresi yang dipergunakan mempertimbangkan komponen *error* dalam persamaannya yang mengarah pada fungsi regresi dengan bentuk stokastik. Model stokastik yang kita gunakan untuk menghubungkan *X* dengan *Y* merupakan modifikasi sederhana dari model *deterministic*. Jika pada model *deterministic*, dinyatakan berikut ini.

Y = nilai rata-rata *Y* untuk nilai *X* tertentu

Sebelumnya kita menotasikan nilai rata-rata *Y* untuk nilai *X* tertentu sebagai

$$E(Y|X_i) = \alpha + \beta X_i .$$

Sekarang, kita dapat menyatakan fungsi regresi dalam bentuk stokastik sebagai berikut.

Y = nilai rata-rata *Y* untuk nilai *X* tertentu + kesalahan acak

$Y = \mathrm{E}(Y|X_i)$ + kesalahan acak

$$Y = \alpha + \beta X_i + kesalahan\,acak$$

Lazimnya kesalahan acak dalam konteks populasi dinotasikan dengan ε, sedangkan kesalahan dalam konteks sampel dinotasikan dengan *e*. Dengan demikian, fungsi regresi populasi dapat dinyatakan sebagai berikut.

$$Y_i = E(Y|X_i) + \varepsilon_i$$

Sementara itu, fungsi regresi sampel, Y_i yang teramati dapat dinyatakan sebagai berikut.

$$Y_i = E(Y|X_i) + \varepsilon_i = \widehat{Y}_i + e_i$$

$$Y_i = \widehat{\alpha} + \widehat{\beta} X_i + e_i \qquad (3)$$

Untuk $X = X_i$, kita mempunyai pengamatan sampel $Y = Y_i$ dengan prediksi Y_i yang bernilai $\widehat{Y_i}$ dan e_i sebagai *error* yang menyatakan kesalahan acak. Menimbang dalam kondisi realita, sering kali kita lebih banyak melakukan observasi berkenaan dengan sampel daripada dengan populasi. Maka itu, tujuan utama kita dalam analisis regresi mengarah pada menaksir fungsi regresi populasi atas dasar fungsi regresi sampel.

2. Peran Komponen *Error*

Dari paparan sebelumnya, diperoleh gambaran bahwa nilai Y_i yang diamati pada umumnya tidak selalu tepat sama dengan $\widehat{Y_i}$ (nilai prediksi Y_i). Hal ini dapat dimengerti karena nilai-nilai suatu variabel acak memang biasanya tidak sama dengan nilai harapan atau nilai rata-rata hitungnya dan biasanya menyimpang dari nilai harapannya. Oleh karena itu, secara umum dapat dituliskan seperti berikut ini.

$$\varepsilon_i = Y_i - E(Y_i) = Y_i - \mu_i = Y_i - \alpha - \beta X_i$$

Dalam hal ini, ε_i mencerminkan perbedaan nilai Y_i dengan nilai harapan atau rata-rata hitung dari Y_i [yang dalam bahasa matematisnya dinotasikan dengan $E(Y_i)$]. Sebagaimana dipaparkan sebelumnya, Y_i merupakan variabel acak dan $E(Y_i) = \mu_i$ adalah sebuah parameter. Maka itu, ε_i adalah variabel acak. Pada umumnya ε_i muncul karena terjadinya kesalahan pengukuran dan kesalahan stokastik.

Kesalahan pengukuran muncul ketika kita tidak tepat dalam melakukan penimbangan. Sebagai gambaran, saat seseorang ingin mengetahui pengaruh pendapatan terhadap konsumsi, kesalahan pengukuran sering kali muncul dalam pengukuran tingkat pendapatan responden. Hal ini lazim terjadi karena orang sering kali tidak mau jujur sewaktu menginformasikan pendapatannya.

Kesalahan stokastik adalah kesalahan yang disebabkan oleh tidak mungkinnya sesuatu hal untuk secara tepat diduplikasi. Meskipun suatu eksperimen dijalankan dengan sangat teliti, kesalahan stokastik dimungkinkan selalu terjadi. Kesalahan stokastik dapat diperkecil dengan melakukan pengawasan yang lebih ketat pada eksperimen yang dilakukan, tetapi kesalahan stokastik tidak dapat dihilangkan sama sekali. Kesalahan stokastik biasanya tidak bisa diduga besarnya. Misalnya, saat kita melakukan suatu telaah tentang pengaruh variabel pendapatan terhadap tingkat konsumsi seseorang, meskipun secara ketat kita telah mengukur tingkat pendapatan,

kita tidak akan dapat memperoleh tingkat konsumsi yang benar-benar sama antarsatu responden dengan responden lainnya. Hal ini disebabkan masih terdapat faktor ekstern yang tidak bisa diawasi dengan teliti, misalnya warisan atau jumlah anggota keluarga. Karena kita telah menganggap hal-hal di luar pengamatan sebagai sesuatu yang konstan, lebih baik kita memasukkan hal tersebut sebagai faktor ekstern yang ikut memengaruhi regresi yang dalam hal ini akan masuk ke komponen *error*. Dengan demikian, dalam konteks regresi, *error* dapat berperan sebagai pengganti semua variabel yang tidak diperhitungkan dalam model walaupun dalam kenyataannya secara bersama-sama variabel-variabel yang tidak dimasukkan dalam model tersebut memiliki pengaruh terhadap *Y*. Pertanyaan yang sering kali dikemukakan adalah jika kita tahu bahwa variabel-variabel tersebut memiliki pengaruh pada *Y*, mengapa tidak semua variabel tersebut dimasukkan secara eksplisit dalam model regresi. Dalam hal ini, ada beberapa alasan yang dapat dikemukakan seperti berikut.

1. Kita ingin berupaya agar model regresi kita sesederhana mungkin (prinsip *parsimony*). Dalam hal ini, kita berupaya menjelaskan perilaku *Y* dengan menggunakan sesedikit mungkin variabel *X*. Dengan demikian, variabel-variabel *X* lain yang kurang penting ataupun kurang relevan akan masuk komponen *error*.
2. Merujuk pada teori yang mendasari hubungan variabel *X* dengan variabel *Y*. Dalam teorinya, dinyatakan bahwa pendapatan memengaruhi konsumsi seseorang, tetapi kita mungkin tidak mengetahui atau tidak yakin mengenai variabel lain yang memengaruhi konsumsi. Oleh karena itu, komponen *error* bisa digunakan sebagai pengganti untuk semua variabel yang tidak dimasukkan atau dihilangkan dari model. Pada konteks ini, sering kali komponen *error* juga disebut sebagai komponen *residual* yang menampung variabel-variabel lain yang sebenarnya berpengaruh pada *Y*, tetapi tidak secara eksplisit dipertimbangkan masuk model.
3. Bahkan, bila kita tahu variabel apa saja yang tidak dimasukkan dalam model dan karena itu kita berkeinginan mempertimbangkan penggunaan regresi majemuk (regresi berganda) dalam menjelaskan perilaku variabel *Y*, kita mungkin tidak memiliki informasi kuantitatif mengenai variabel-variabel tersebut. Sebagai contoh, secara prinsip kita dapat memasukkan kekayaan seseorang atau bahkan kekayaan keluarga sebagai variabel yang menjelaskan konsumsi seseorang di samping variabel pendapatan yang telah kita masukkan dalam model. Sayangnya, informasi mengenai kekayaan seseorang atau kekayaan keluarga biasanya tidak tersedia.

Karena itu, kita mungkin terpaksa menghilangkan variabel kekayaan dari model kita walaupun secara logika variabel tersebut sangat relevan dalam menjelaskan konsumsi.

4. Pengaruh bersama variabel-variabel yang tidak dimasukkan dalam model tersebut sedemikian kecil dan tidak sistematis sehingga dari segi kepraktisan dan pertimbangan biaya tidak bermanfaat untuk memasukkan variabel-variabel tersebut dalam model secara eksplisit.
5. Bahkan, bila kita berhasil memasukkan semua variabel yang relevan ke dalam model, ada suatu kerandoman hakiki dalam individual Y yang tidak dapat dijelaskan, tidak peduli betapa keras pun kita berusaha. Dalam hal ini, gangguan *error* sangat baik mencerminkan kerandoman hakiki tersebut.

LATIHAN

Untuk memperdalam pemahaman Anda mengenai materi di atas, kerjakanlah latihan berikut!

1) Jelaskan fokus dalam analisis regresi!
2) Jelaskan anggapan (asumsi) terkait regresi populasi bahwa variabel-variabel acak Y_i adalah bebas secara statistik satu sama lain!
3) Jelaskan yang dimaksud dengan fungsi regresi yang *deterministic*!
4) Jelaskan apa yang dimaksud dengan kesalahan stokastik!
5) Jelaskan, mungkinkah suatu garis regresi tidak memiliki intersep!

Petunjuk Jawaban Latihan

1) Dalam analisis regresi, fokus kita adalah menaksir fungsi regresi populasi, yaitu menaksir nilai α dan β yang tidak diketahui berdasarkan pengamatan atas X dan Y.
2) Artinya, bila salah satu nilai pengamatan di dalam Y_i berubah, perubahan Y_i tersebut tidak akan memengaruhi Y_i yang lainnya. Sebagai gambaran, jika hasil nilai Y_1 kecil, nilai Y_1 yang kecil tersebut tidak akan memengaruhi nilai Y_2 untuk menjadi kecil atau sebaliknya. Dengan kata lain, Y_2 tidak terpengaruh oleh Y_1. Demikian pula untuk pengamatan-pengamatan yang lain juga tidak saling terpengaruh dan memengaruhi satu sama lain.

3) Fungsi regresi yang *deterministic* adalah fungsi regresi yang memenuhi kondisi bila sebuah nilai *X* disubstitusikan ke dalam persamaan. Nilai *Y* menjadi tertentu dan tidak ada ruang yang disediakan untuk kesalahan (*error*). Pada kondisi ini, nilai taksiran *Y* akan sama persis dengan nilai *Y* aslinya.
4) Kesalahan stokastik adalah kesalahan yang disebabkan oleh tidak mungkinnya suatu hal untuk secara tepat diduplikasi.
5) Mungkin bila garis regresi memotong sumbu *Y* pada titik $Y = 0$.

RANGKUMAN

1. Penafsiran regresi dewasa ini berbeda dari penafsiran regresi menurut Galton. Dewasa ini, analisis regresi berguna dalam menelaah hubungan dua variabel atau lebih dan terutama untuk menelusuri pola hubungan yang modelnya belum diketahui dengan sempurna sehingga dalam terapannya lebih bersifat eksploratif.
2. Dikenal fungsi regresi populasi dan fungsi regresi sampel. Secara umum, tujuan utama kita dalam analisis regresi adalah menaksir fungsi regresi populasi atas dasar fungsi regresi sampel karena sangat sering analisis didasarkan pada sampel daripada didasarkan pada populasi.

TES FORMATIF 2

Pilihlah satu jawaban yang paling tepat!

1) Yang dimaksud dengan prinsip *parsimony* dalam pembuatan model adalah ….
   A. memasukkan semua variabel yang secara logika memengaruhi variabel terikat
   B. memasukkan semua variabel yang menurut teori memengaruhi variabel terikat
   C. memasukkan sebanyak mungkin variabel yang dianggap paling memiliki pengaruh terbesar pada variabel terikat
   D. menjaga agar model yang dibuat sesederhana mungkin dengan tetap berupaya untuk mencapai kemampuan penjelasan yang dapat dipertanggungjawabkan

2) Dalam konteks hubungan kausal, variabel yang memengaruhi variabel lain disebut sebagai variabel bebas (*independent variable*), sedangkan variabel yang dipengaruhi oleh variabel lain disebut ….
   A. variabel terikat (*dependent variable*)
   B. variabel tak gayut
   C. variabel yang bisa dikontrol
   D. variabel peramal

3) Pada konteks hubungan kausal, variabel terikat adalah variabel yang dipengaruhi oleh variabel bebas. Berikut adalah sebutan lain bagi variabel terikat, *kecuali* ....
   A. variabel tak bebas
   B. variabel respons
   C. *eXplanatory variable*
   D. variabel dependen

4) Pada konteks hubungan kausal, bila variabel *X* adalah variabel yang meregresi (*regressor*), variabel *Y* adalah ….
   A. *eXplanatory variable*
   B. variabel yang diregresi (*regressand*)
   C. variabel yang bisa dikontrol
   D. variabel tak gayut

5) Dalam konteks analisis regresi, komponen *error* merepresentasikan perbedaan nilai antara ….
   A. nilai *Y* yang sebenarnya dengan nilai *Y* hasil prediksi
   B. nilai *Y* yang sebenarnya dengan nilai *X* yang sebenarnya
   C. nilai *Y* yang sebenarnya dengan nilai *X* hasil prediksi
   D. nilai *Y* hasil prediksi dengan nilai *X* yang sebenarnya

6) Dalam konteks plot pasangan titik-titik *X* dan *Y* pada sistem koordinat *rectangular*, *error* adalah ....
   A. jarak diagonal dari nilai *Y* sebenarnya dengan nilai *Y* hasil prediksi
   B. jarak horizontal dari nilai *Y* sebenarnya dengan nilai *Y* hasil prediksi
   C. jarak vertikal dari nilai *Y* sebenarnya dengan nilai *Y* hasil prediksi
   D. jarak terjauh dari nilai *Y* sebenarnya dengan nilai *Y* hasil prediksi

7) Metode kuadrat terkecil mempunyai beberapa sifat statistik yang sangat menarik yang membuatnya menjadi satu metode analisis regresi yang paling kuat (*powerful*) dan populer. Metode ini dianggap paling baik karena, *kecuali* ....

A. di samping dapat mengatasi perbedaan tanda pada *error* juga dapat menggambarkan semua peranan titik observasi yang tersebar
B. dengan metode ini masalah perbedaan tanda pada *error* dapat diatasi
C. metode ini memperhatikan semua titik observasi yang tersebar
D. semakin serius *error* yang muncul

8) Komponen *error* merupakan suatu variabel acak (*random variabel*) yang merepresentasikan kegagalan model yang dibentuk dalam mencocokkan diri sepenuhnya terhadap data yang dihadapi. Dalam hal ini, ε_i adalah suatu variabel acak yang memenuhi karakteristik berikut, *kecuali* ....
A. independen
B. pada dasarnya distribusi Y_i dan ε_i adalah tidak identik
C. memiliki nilai mean = 0
D. memiliki variance = σ^2

9) Pada umumnya, ε_i muncul karena terjadinya kesalahan pengukuran dan kesalahan stokastik. Kesalahan stokastik adalah kesalahan yang disebabkan oleh ….
A. tidak mungkinnya sesuatu hal untuk secara tepat diduplikasi
B. *error* yang positif mengompensasi *error* yang negatif
C. tidak adanya bias dugaan
D. efisiensi dan efektivitas estimasi yang baik

10) Dalam konteks regresi, *error* dapat berperan sebagai ….
A. pelengkap persamaan regresi
B. pengganti semua variabel yang tidak diperhitungkan dalam model walaupun dalam kenyataannya secara bersama-sama variabel-variabel yang tidak dimasukkan dalam model tersebut memiliki pengaruh terhadap *Y*
C. agar ketepatan garis regresi dugaan dalam mewakili garis regresi populasi tidak dapat dihitung
D. sarana untuk mengkompensasi *underestimate* atau *overestrimate*

Cocokkanlah jawaban Anda dengan Kunci Jawaban Tes Formatif 2 yang terdapat di bagian akhir modul ini. Hitunglah jawaban yang benar. Kemudian, gunakan rumus berikut untuk mengetahui tingkat penguasaan Anda terhadap materi Kegiatan Belajar 2.

$$\text{Tingkat penguasaan} = \frac{\text{Jumlah Jawaban yang Benar}}{\text{Jumlah Soal}} \times 100\%$$

Arti tingkat penguasaan: 90 - 100% = baik sekali
80 - 89% = baik
70 - 79% = cukup
< 70% = kurang

Apabila mencapai tingkat penguasaan 80% atau lebih, Anda dapat meneruskan dengan Kegiatan Belajar 3. **Bagus!** Jika masih di bawah 80%, Anda harus mengulangi materi Kegiatan Belajar 2, terutama bagian yang belum dikuasai.

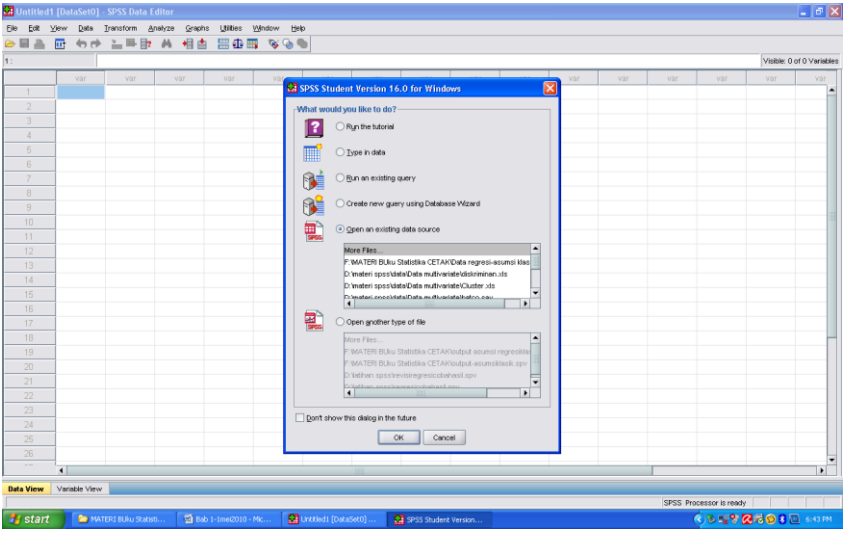
KEGIATAN BELAJAR 3

Penggunaan SPSS

Fokus Kegiatan Belajar 3 adalah penggunaan SPSS. Pada Kegiatan Belajar 1, telah diilustrasikan cara memperoleh diagram pencar (*scatter plot*) secara manual. Pada Kegiatan Belajar 3, dipaparkan tahapan penggunaan SPSS dalam memperoleh diagram pencar. Pemahaman terhadap tahapan yang dikemukakan akan memudahkan mahasiswa dalam memperoleh diagram pencar dari hubungan variabel-variabel yang menjadi perhatian pada tampilan dua dimensi.

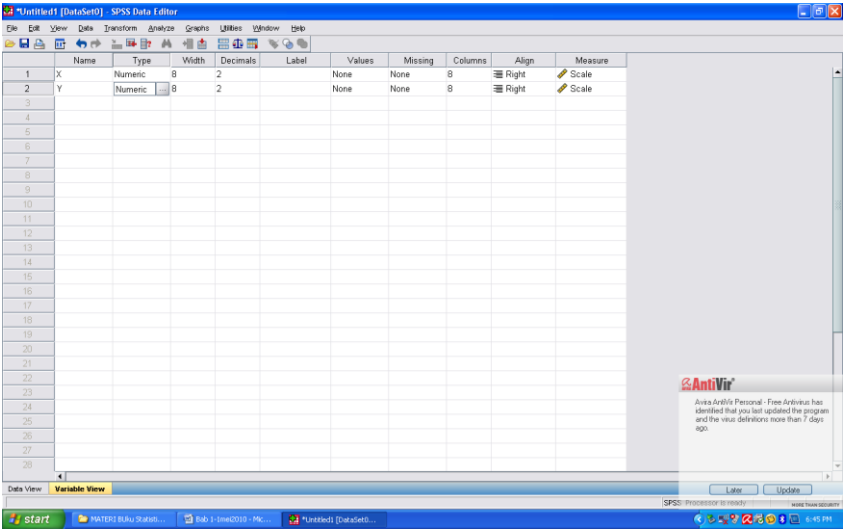
Berikut adalah paparan tahapan tersebut.

1. Buka SPSS. Pada saat SPSS pertama kali dibuka selalu tampak tampilan pertama pada monitor sebagai berikut.

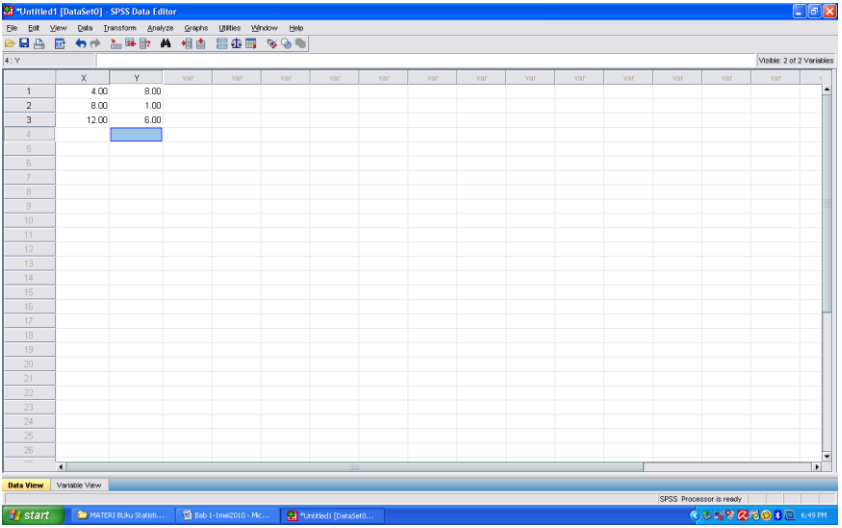


Tampilan layar tersebut disebut SPSS DATA EDITOR yang merupakan Windows utama pada SPSS. DATA EDITOR mempunyai dua fungsi utama, yaitu *input* data yang diolah oleh SPSS atau proses data yang telah di-*input* dengan prosedur statistik tertentu. Berikut ini akan dijelaskan tahapan mulai dari *input* data.

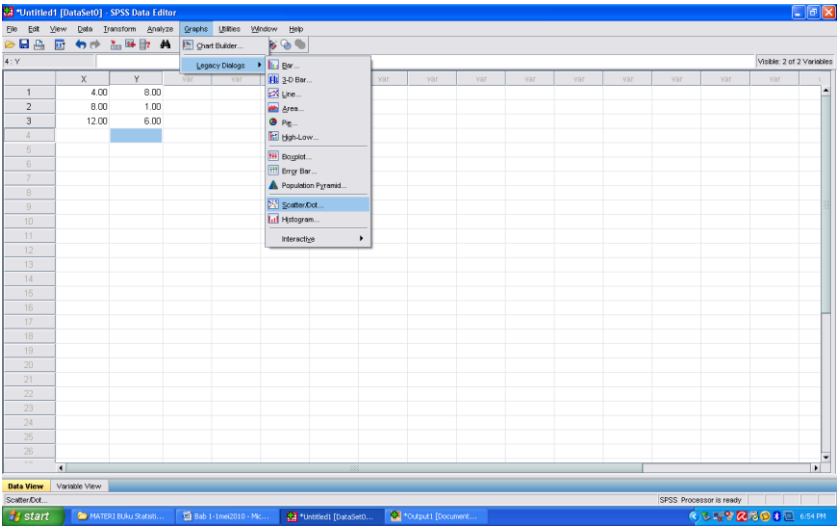
2. Klik *Variabel View*.
3. Beri nama variabel *X* dan variabel *Y*.



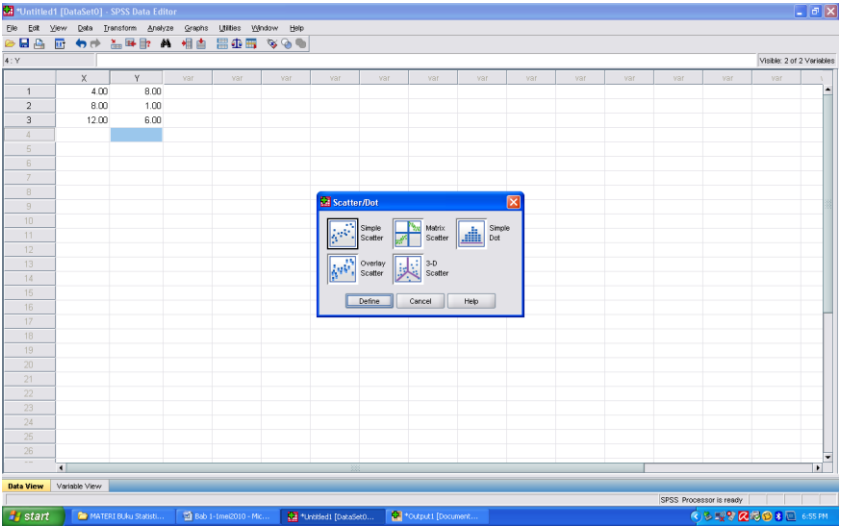
4. Klik *Data View*.
5. Ketikkan data untuk variabel *X* pada kolom 1 ke arah bawah dan juga data untuk variabel *Y* pada kolom 2 ke arah bawah.



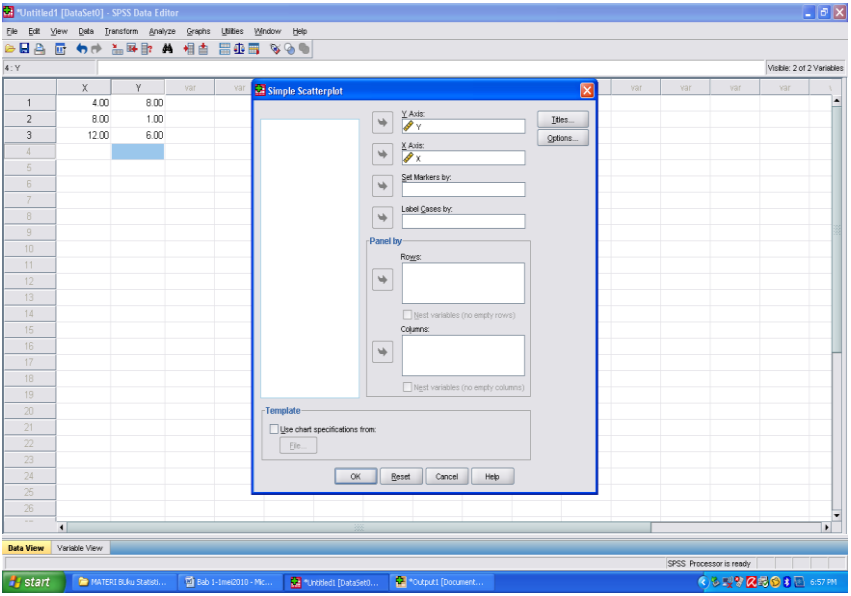
6. Klik *Graph*, kemudian *Legacy Dialogs*, lalu pilih *Scatter*/*Dot*.



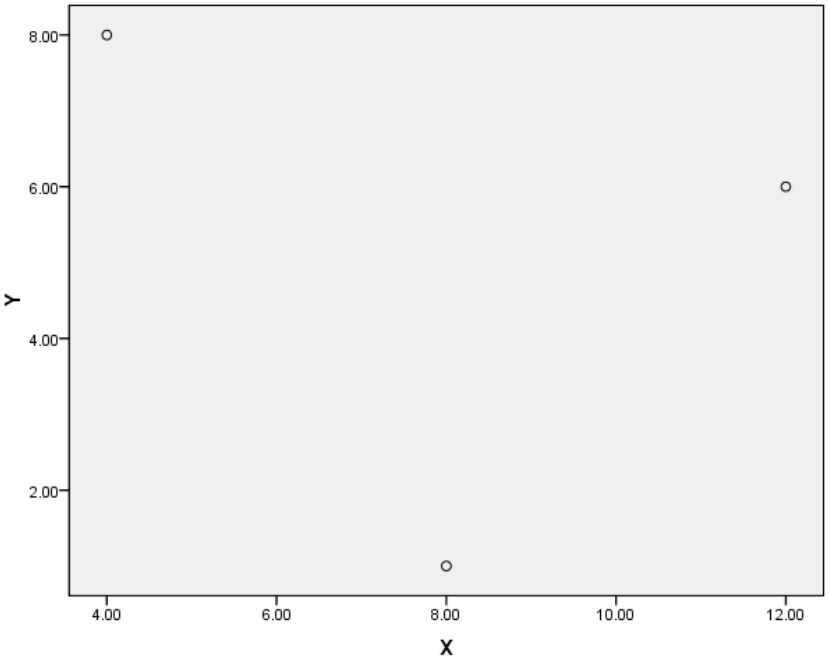
7. Bila Anda klik *Scatter*/*Dot*, akan diperoleh tampilan berikut.



8. Pilih *Simple Scatter*, klik *Define*.
9. Klik *X*, lalu klik anak panah untuk memindahkan *X* ke isian *X* a*X*is. Klik *Y* lalu klik anak panah untuk memindahkan *Y* ke isian *Y* a*X*is.



10. Setelah itu, klik OK akan diperoleh tampilan berikut.



Tentu saja pembaca dapat mengisikan label yang memberi keterangan lebih perinci untuk variabel *X* dan *Y*. Namun, dalam bagian ini, penulis hanya memberikan contoh terkait uraian yang relevan sehingga improvisasi tampilan diserahkan pada pembaca.

LATIHAN

Untuk memperdalam pemahaman Anda mengenai materi di atas, kerjakanlah latihan berikut!

Untuk memperjelas Kegiatan Belajar 3, Anda diharapkan praktik langsung di komputer dengan mengikuti tahap-tahap yang telah diberikan.

Kunci Jawaban Tes Formatif

Tes Formatif 1

1) A
2.) A
3) D
4) C
5) A
6) C
7) A
8) B
9) A
10) D

Tes Formatif 2

1) D
2) A
3) C
4) B
5) A
6) C
7) D
8) B
9) A
10) B

Glosarium

Best fitting	:	kecocokan terbaik.
Diagram pencar (*scatter diagram*)	:	diagram yang merepresentasikan plot pasangan nilai *X* dan *Y* pada sebuah panel sumbu silang variabel-variabel *X* dan *Y*.
Error	:	besaran yang merepresentasikan perbedaan nilai antara nilai *Y* yang sebenarnya (yang sering kali disebut sebagai *Y observed*) dengan nilai *Y* hasil prediksi (yang sering kali dinotasikan dengan $\hat{Y}_i$). Semakin kecil nilai *error* berarti semakin dekat nilai prediksi terhadap nilai amatannya, sedangkan semakin besar *error* yang muncul berarti semakin jauh nilai prediksi dari nilai amatan yang sebenarnya.
Hubungan kausal	:	hubungan yang menyatakan adanya pengaruh satu atau lebih variabel terhadap variabel lain. Hubungan kausal memungkinkan kita untuk memprediksi nilai suatu variabel jika nilai variabel yang memengaruhinya diketahui.
Kurva yang *good fit*	:	kurva yang dibentuk dengan kriteria kecocokan tinggi terhadap data aslinya. Pada kurva yang *good fit*, bila kita mengetahui nilai *X*, kita dapat meramalkan nilai *Y* dengan hasil yang menyerupai *Y* aslinya. Kurva yang memiliki kriteria *good fit* tentunya kurva yang memiliki keadaan *total error* yang kecil dengan penyimpangan yang minimum.
Kurva yang representatif	:	kurva yang memiliki kemampuan mewakili sebaran data yang dihadapi dan memenuhi kriteria *good fit* (kecocokan terbaik) terhadap data aslinya.
Method of ordinary least squares	:	metode kuadrat terkecil, suatu metode mematut garis regresi dengan meminimumkan jumlah kuadrat penyimpangan nilai-nilai *Y* yang diamati dari nilai-nilai *Y* yang diramalkan (*Y* prediksi).

Fungsi regresi ditetapkan dengan cara sedemikian rupa sehingga jumlah kuadrat *error* adalah sekecil mungkin, yang dalam notasi matematis dituliskan $\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$ minimum. Dalam hal ini, e_i^2 adalah kuadrat *error*.

Persamaan garis dugaan : persamaan garis yang digunakan untuk mewakili sebaran data aslinya.

Prediksi yang *overestimate* : kondisi yang menunjukkan nilai Y_i (*observed*) lebih kecil dari nilai Y_i hasil prediksi akan diperoleh *error* negatif. Dalam hal ini, berarti kita menaksir nilai Y_i terlalu tinggi.

Prediksi yang *underestimate* : kondisi yang menunjukkan nilai Y_i (*observed*) lebih besar dari nilai Y_i hasil prediksi sehingga diperoleh *error* positif. Dalam hal ini, berarti kita menaksir nilai Y_i terlalu rendah.

Regresi : berdasarkan telaah Francis Galton tentang sifat-sifat keturunan. Istilah regresi merujuk pada suatu fenomena bergerak atau mundur **(*regress*)** ke arah rata-rata seluruh populasi. Dewasa ini, analisis regresi berguna dalam menelaah hubungan dua variabel atau lebih dan terutama untuk menelusuri pola hubungan yang modelnya belum diketahui dengan sempurna sehingga dalam terapannya lebih bersifat eksploratif. Saat ini, analisis regresi lebih berkenaan dengan studi ketergantungan suatu variabel (variabel tak bebas) pada variabel lainnya (variabel bebas), dengan tujuan memprediksi atau meramalkan nilai rata-rata hitung (mean) atau rata-rata populasi variabel tak bebas.

***Trend* hubungan** : kecenderungan hubungan antara dua atau lebih variabel. Pada umumnya, didapati dua penggolongan *trend*, yaitu *trend* positif atau *trend* negatif. *Trend* *X* dan *Y* yang positif ditandai dengan pergerakan *X* dan *Y* yang searah

(bila *X* bergerak naik, a *Y* bergerak naik; bila *X* bergerak turun, *Y* juga bergerak turun). *Trend X* dan *Y* yang negatif ditandai dengan pergerakan *X* dan *Y* yang tidak searah (bila *X* bergerak naik, *Y* bergerak turun; bila *X* bergerak turun, *Y* juga bergerak naik).

Variabel : variabel merupakan suatu atribut dari sekelompok objek yang diteliti yang memiliki variasi antara satu objek dan objek yang lain dalam kelompok tersebut.

Variabel bebas : variabel yang memengaruhi variabel lain.

Variabel terikat : variabel yang dipengaruhi oleh variabel lain.

Daftar Pustaka

Berenson, Mark L., David M Levine , & Timothy C. Krehbiel. (2005). *Basic Business Statistics: Concepts and Aplications 10ed*. Upper Saddle River. Prentice Hall.

Gujarati. (2009). *Basic Econometrics 5ed*. Singapore: Mc-Grawhill.

Heizer, Jay & Barry Render. (2008). *Operations Management 9ed*. New Jersey: Pearson Education Inc.

Karseno, A.R. (2008). *Statistika Ekonomi II*. Cetakan keempat. Jakarta: Penerbit Universitas Terbuka.

Levin, Richard I & David S. Rubin. (1998). *Statistics for Management 7ed*. New Jersey: Prentice Hall.

McClave, James T., Benson, P.George & Sincich, Terry. (2008). *Statistics for Economics and Business 10 ed*. Upper Saddle River: Pearson Prentice Hall.

Siagian, Dergibson & Sugiarto. (2006). *Metode Statistika untuk Bisnis dan Ekonomi*. Jakarta: Gramedia Pustaka Utama.

SPSS Inc. (2007). *SPSS 16.0 Brief Guide*. USA: Prentice Hall.