

Analisis Regresi I: Regresi dengan Satu Prediktor Metode Kuadrat Terkecil dan Inferensi Parameter Regresi

Prof. Dr. Zanzawi Soejoeti



PENDAHULUAN

Dalam banyak bidang penelitian ilmiah, variasi dalam pengukuran pengukuran percobaan suatu variabel yang berhubungan satu sama lainnya besarnya berubah selama percobaan itu. Dengan memasukkan secara eksplisit data variabel-variabel yang berpengaruh ini ke dalam analisis statistik, sering kali dapat menilai sifat hubungannya, kemudian memanfaatkan informasi ini guna meningkatkan deskripsi dan inferensi tentang variabel yang menjadi perhatian utamanya. Menelusuri hubungan antarvariabel juga penting dalam hal nilai satu variabel dapat diperkirakan dari observasi pada variabel-variabel yang lain, bahkan dapat mengendalikan dan mengoptimalkan dengan memanipulasi faktor-faktor yang berpengaruh.

Analisis regresi adalah kumpulan metode statistik yang mempelajari perumusan model matematik yang menggambarkan hubungan antara beberapa variabel, dan menggunakan hubungan yang dimodelkan ini untuk tujuan memperkirakan dan inferensi statistik yang lain. Bagaimana alat statistik modern yang sangat kuat seperti diberi nama “regresi” perlu penjelasan. Menurut sejarah, kata “regresi” dalam teknis, seperti sekarang ini pertama kali digunakan oleh Francis Galton, yang menganalisis tinggi anak dan tinggi rata-rata orang tua mereka. Dari observasinya, Galton menyimpulkan bahwa anak dari orang tua yang sangat tinggi (pendek) umumnya lebih tinggi (pendek) dari rata-rata tetapi tidak setinggi (sependek) orang tua mereka. Hasil ini dipublikasikan dalam tahun 1885 dengan judul *Regresi Toward Mediocrity in Hereditary Stature* (Regresi terhadap sifat sedang dalam tinggi turunan). Dalam hubungan ini istilah “regresi” berarti tinggi anak cenderung ke arah rata-rata daripada ke nilai-nilai yang lebih ekstrem.

KEGIATAN BELAJAR 1**Analisis Regresi dengan Satu Prediktor**

Stilah analisis regresi dikembangkan ke arah titik yang sekarang mencakup analisis data yang melibatkan dua variabel atau lebih dengan tujuan guna menyingkap sifat hubungannya, kemudian menelusurinya untuk maksud memperkirakan (prediksi). Studi tentang hubungan antara beberapa variabel lazim dalam banyak bidang penyelidikan ilmiah. Manajer periklanan suatu perusahaan tertarik akan hubungan antara biaya yang dikeluarkan untuk iklan dan kenaikan dalam penjualan. Perhatian utama dalam pengobatan dengan radiasi adalah seberapa jauh kerusakan sel yang disebabkan karena lama dan intensitas radiasi. Guna meramalkan banjir, ahli hidrologi harus mempelajari tingkat penyaluran sungai yang diukur pada tempat tertentu dalam hubungannya dengan curah hujan dan tingkat penyaluran pada suatu tempat di hulu sungai dalam tenggang waktu yang sesuai. Dalam studi kesadaran politik, seorang ahli sosiologi mungkin ingin menghubungkan persentase orang yang mempunyai hak pilih dengan faktor sosial ekonomi, seperti struktur umur, tingkat pendidikan, dan penghasilan rata-rata. Bagian personalia suatu perusahaan sering menelusuri hubungan antara nilai pekerjaan karyawan dan evaluasi awal berdasarkan skor interview pada waktu masuk bekerja. Dengan bermacam ragamnya sifat hubungan antara beberapa variabel yang mungkin, klasifikasi yang luas dan beberapa contoh akan menolong dalam membeberkan ruang lingkup analisis regresi.

1. Hubungan Deterministik

Beberapa variabel sering dihubungkan oleh suatu hukum yang dapat dinyatakan dengan suatu fungsi matematik yang tepat. Beberapa dasar teoretis yang dikenal secara luas membenarkan bentuk fungsional itu dan setiap penyimpangan data observasi dari hubungan ini dipandang sebagai kesalahan (sesatan) eksperimental.

Misalnya, x rupiah didepositokan dengan bunga $100r\%$ bersusun tahunan dan y menunjukkan banyak uang dalam deposito n tahun maka y berhubungan dengan x , r , dan n dengan rumus eksak.

$$y = x(1+r)^n$$

yang dikenal sebagai hukum bunga bersusun. Sebagai contoh kedua, waktu t yang diperlukan sebuah bola besi untuk mencapai permukaan bumi jika dijatuhkan dari ketinggian h , berhubungan dengan h menurut hukum gravitasi

$$\text{fisika} \quad t = \sqrt{\frac{2h}{g}}$$

dengan g = konstanta gravitasi. Galileo mula-mula mempostulasikan bahwa h proporsional dengan t^2 dan meninggalkan hal ini bagi pembuat percobaan, kemudian untuk mendapatkan taksiran yang akurat tentang g . Nilai sekarang untuk banyak maksud, adalah sedemikian sehingga percobaan dan analisis data lebih lanjut tidak diperlukan. Hal-hal seperti ini dikeluarkan dari ruang lingkup analisis regresi.

2. Hubungan Semideterministik

Dalam banyak hal yang lain, teori yang telah mapan menentukan suatu bentuk untuk hukum hubungan beberapa variabel, tetapi bukan nilai-nilai parameter tertentu yang tampak dalam hubungan itu. Untuk mempelajari tentang parameter-parameter ini, kita harus melakukan percobaan. Ketepatan alat pengukur yang terbatas, gangguan keadaan percobaan yang tidak dapat dikendalikan, dan faktor-faktor yang lain menyebabkan timbulnya kesalahan percobaan dalam data yang biasanya menyebabkan suatu variasi tentang hubungan yang sebenarnya.

Contoh 2.1

Tiap gas mempunyai perbandingan panas tertentu γ . Jika kita melakukan percobaan tanpa ada variasi di dalam panas jika kita mengubah volume v dan mengukur tekanan P , hukum gas yang ideal menyatakan:

$$PV^\gamma = \text{konstan}$$

dengan perbandingan panas tertentu γ harus ditaksir dari data hasil percobaan pada P dan v .

Dua contoh jenis lainnya adalah menaksir konstan dalam bentuk persamaan yang diketahui untuk suatu reaksi kimia, dan menaksir ekspansi panas dalam campuran logam yang baru.

Dalam berbagai keadaan latar belakang teoretis menganjurkan bentuk hubungan yang masuk akal, tetapi dasar teoretis itu tidak tepat atau dapat diterima secara umum. Lagi pula, fluktuasi tambahan sering dihasilkan oleh variabel yang tak terkendali yang tidak termasuk dalam hubungan itu.

Contoh 2.2

Misalkan, suatu pabrik memproduksi benda dalam kelompok dan manajer produksi ingin menghubungkan biaya produksi satu kelompok (y) dengan ukuran kelompok (x). Bagian tertentu dari biaya itu praktis konstan, seberapa pun ukuran kelompok x , setidaknya dalam rentang variasi x yang realistik. Biaya gedung dan administrasi serta gaji pengawas termasuk dalam kategori ini, dan kita tulis biaya tetap ini bersama-sama dengan F . Bagian kedua berbanding langsung dengan banyak unit yang diproduksi. Misalnya, bahan baku dan tenaga kerja yang diperlukan untuk menghasilkan produksi itu termasuk dalam kategori ini. Misalkan, c menunjukkan biaya variabel untuk memproduksi satu unit benda. Tanpa adanya faktor-faktor lain maka kita dapat mengharapkan akan mempunyai hubungan biaya ukuran deterministik dalam bentuk

$$y = F + cx$$

Tetapi, kita harus memperhatikan bagian ketiga dari biaya yang besarnya sedikit banyak bersifat tak terduga. Kadang-kadang mesin produksi macet sewaktu memproduksi yang mengakibatkan sejumlah waktu menganggur dan biaya perbaikan yang berbeda-beda. Fluktuasi dalam kualitas bahan baku dapat juga terjadi yang berakibat perlambatan proses produksi. Jadi, hubungan deterministik dapat ditutup dengan bagian berpeluang yang diakibatkan oleh faktor-faktor itu dan faktor-faktor yang tak dapat ditemukan lainnya. Oleh karena itu, hubungan antara y dan x harus diselidiki dengan analisis statistik data biaya ukuran kelompok.

3. Hubungan Empiris

Berbeda dengan keadaan di atas banyak fenomena alam yang terdiri dari variabel-variabel yang saling berhubungan, atau satu variabel yang bergantung pada sejumlah variabel yang berpengaruh atau penyebab, di mana hubungan itu tidak dikendalikan oleh hukum alam yang tepat. Gambar beberapa nilai pengamatan variabel-variabel ini pada kertas grafik melukiskan dalam bentuk yang agak kasar, hubungan jalin-menjalin dengan fluktuasi peluang. Berikut adalah beberapa contoh keadaan dengan bentuk hubungan yang melandasinya benar-benar tidak diketahui. Setelah memperoleh pengetahuan yang cukup tentang hubungan empiris ini memungkinkan bagi peneliti untuk merumuskan suatu teori yang mengarah pada rumus matematik sehingga menjadi semideterministik.

Contoh 2.3

Untuk memerangi polusi mobil, penelitian sedang dilakukan untuk menentukan komposisi kimia bahan tambahan bensin yang akan meningkatkan kualitas pemancaran. Satu aspek penelitian adalah mempelajari hubungan antara banyak bahan tambahan tertentu dan pengurangan pemancaran nitrogen oksida. Bahan ramuan yang lain dapat juga menghasilkan beberapa pengaruh, tetapi banyaknya tetap konstan selama studi. Beberapa mobil merek A dipilih sebagai unit percobaan. Banyak nitrogen oksida yang ke luar di ukur untuk tiap mobil pertama tanpa bahan tambahan, kemudian dengan bahan tambahan sebanyak x yang ditentukan. Pengurangan nitrogen oksida yang diperoleh diambil sebagai respons y karena rumitnya reaksi kimia dan keadaan internal mesin mobil, rumus deterministik hubungan antara y dan x di luar jangkauan pengetahuan sekarang.

Contoh 2.4

Misalkan, hasil tanaman tomat y dalam suatu percobaan pertanian akan dipelajari dalam hubungannya dengan dosis x suatu pupuk tertentu, sedangkan faktor pendukung yang lain, seperti irigasi dan jenis tanah sedapat mungkin tetap konstan. Percobaan itu terdiri dari penggunaan berbagai dosis pupuk yang berbeda-beda yang meliputi rentang yang diinginkan, dalam petak yang berbeda, dan selanjutnya mencatat hasil tomat dari petak-petak ini. Dosis pupuk yang berbeda biasanya akan mengakibatkan hasil yang berbeda, tetapi hubungan itu tidak diharapkan mengikuti rumus matematik yang tepat. Di samping variasi peluang yang tak dapat diperkirakan, bentuk yang mendasari hubungan dalam hal ini tidak dapat ditentukan dari dasar teoretis yang mana pun.

Contoh 2.5

Ketangkasan operator yang baru dilatih untuk melakukan pekerjaan keterampilan bergantung pada lamanya periode latihan dan sifat program latihan. Guna menilai efektivitas program latihan, kita harus melakukan studi percobaan tentang hubungan antara pertumbuhan dalam keterampilan atau pengetahuan y dan lamanya latihan x . Akan tetapi, tidak mungkin bahwa hubungan ini deterministik karena kenyataan yang sederhana saja bahwa tidak ada dua individu yang tepat serupa. Namun demikian, analisis data 2 variabel itu dapat menolong kita menilai sifat hubungan dan

menggunakannya dalam menilai dan merancang program latihan semacam itu.

Contoh-contoh itu melukiskan daerah penerapan analisis regresi dalam konteks yang sederhana untuk menentukan bagaimana satu variabel dihubungkan dengan variabel yang lain. Dalam keadaan yang lebih rumit beberapa variabel mungkin saling berhubungan atau satu variabel yang menjadi perhatian utama mungkin bergantung pada beberapa variabel yang mempengaruhi dan studi tentang hubungan semacam itu memerlukan observasi dan analisis semua variabel-variabel ini. Dalam Contoh 2.5 perkembangan pengetahuan dapat dipelajari dalam hubungannya dengan IQ, skor pada ujian ketangkasan awal, banyak latihan yang diterima dalam kelas dan laboratorium, dan sebagainya. Demikian juga hasil proses kimia dapat dipelajari dalam hubungannya dengan beberapa variabel, seperti temperatur dalam sistem, konsentrasi awal ramuan, atau tingkat pendinginan. Kegunaan analisis regresi meluas pada masalah-masalah multivariat ini. Analisis ini memberikan metode untuk pembentukan model untuk hubungan-hubungan seperti tersebut, yaitu menaksir parameter yang tidak diketahui, menentukan variabel mana yang penting dan mana yang tidak penting, dan akhirnya menggunakan model ini untuk perkiraan dan pengendalian.

A. DIAGRAM TITIK

Dalam membeberkan konsep-konsep dasar, kita mulai dengan suatu percobaan untuk menentukan hubungan antara dua variabel x dan y . x bertindak sebagai variabel independen yang nilai-nilainya dikendalikan oleh pembuat percobaan, sedangkan y dependen (bergantung) pada x dan terkena sumber sesatan yang tak dapat dikendalikan.

Variabel *independen* atau *terkendali* adalah juga disebut variabel *prediktor* dan ditulis dengan x . Variabel *respons* atau *akibat* ditulis dengan y .

Ketergantungan y pada x adalah satu arah sehingga kita terutama tertarik dengan situasi yang nilai-nilai x ditetapkan tanpa ada sesatan yang mengganggu (keadaan dengan x dan y keduanya di luar pengendalian pembuat percobaan dan hanya dapat diamati dengan mengambil sampel random dibicarakan kemudian di belakang). Untuk pembicaraan yang lebih

konkret, kita misalkan n mobil merek A digunakan dalam percobaan, seperti dilukiskan dalam contoh 2.3. Banyak nitrogen oksida yang dipancarkan (dikeluarkan) oleh tiap mobil diukur pertama-tama tanpa bahan tambahan digunakan dalam bensin penuh dan pemancaran nitrogen oksida diukur lagi. Pengurangan dalam banyak nitrogen oksida, kemudian dicatat sebagai variabel respons y . Data itu dapat disusun, seperti dalam Tabel 2.1.

Tabel 2.1
Bentuk Data

Banyak bahan tambahan (x)	x_1, x_2, \dots, x_n
Pengurangan nitrogen oksida (y)	y_1, y_2, \dots, y_n

Guna memberikan gambaran contoh dalam angka-angka, kita pandang data yang diberikan dalam Tabel 2.2 sebagai observasi yang diperoleh dalam suatu percobaan dengan $n = 10$ mobil. Banyak bahan tambahan x dan pengurangan dalam nitrogen oksida y diukur dalam satuan tertentu. Tujuan nilai x yang berbeda digunakan dalam percobaan itu, dan beberapa diantaranya digunakan replikasi untuk lebih dari satu mobil. Pandangan sepintas tabel di atas menunjukkan bahwa y umumnya naik jika y juga naik, tetapi sukar untuk mengatakan tentang bentuk hubungan itu hanya dengan melihat data dalam tabel.

Tabel 2.2
Banyak Bahan Tambahan dan Pengurangan
dalam Nitrogen Oksida dari 10 Mobil

Bahan tambahan (x)	1	1	2	3	4	4	5	6	6	7
Pengurangan dalam nitrogen oksida (y)	2,1	2,5	3,1	3,0	3,8	3,2	4,3	3,9	4,4	4,8

Dalam mempelajari hubungan antara dua variabel, langkah pertama yang kita tempuh adalah menggambarkan data sebagai titik-titik dalam kertas grafik. Gambar hasilnya yang dinamakan diagram titik menunjukkan apakah

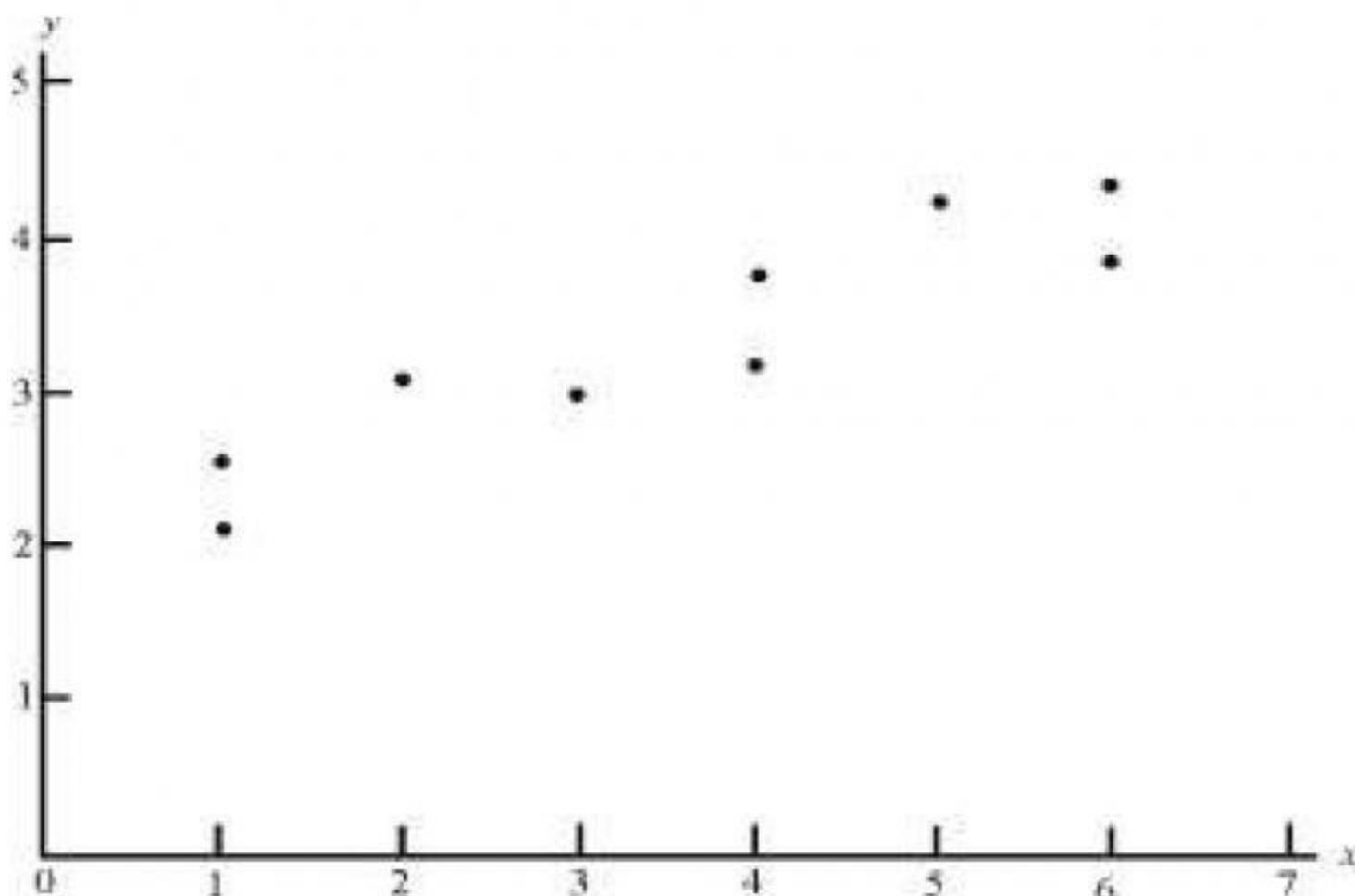
titik-titik itu bergerombol di sekitar garis lurus atau kurva dan juga memberikan kesan visual tentang besar variasi di sekitar garis atau kurva itu. Dalam banyak hal tidak ada hubungan teoretis yang diketahui sebelumnya untuk digunakan sehingga informasi itu tergambar dalam diagram titik yang berguna dalam mencari model matematik yang sesuai.

Diagram titik observasi dalam Tabel 2.2 tampak dalam Gambar 2.1. Diagram titik ini mengungkap hubungannya kira-kira bersifat linear; yakni titik-titik itu terlihat bergerombol di sekitar garis lurus. Oleh karena hubungan linear merupakan hubungan yang paling sederhana penanganannya matematiknya maka kita sajikan secara rinci analisis regresi statistik untuk hal ini. Keadaan yang lain sering kali dapat diubah menjadi linear dengan

Langkah pertama dalam analisis

Dalam meneliti hubungan antara dua variabel menggambar diagram titik adalah langkah awal yang penting yang harus dilakukan sebelum melakukan analisis statistik yang formal. Diagram titik memberikan pandangan tentang sifat hubungan yang ditunjukkan data itu.

menggunakan transformasi yang sesuai untuk satu variabel atau keduanya.



Gambar 2.1
Diagram Titik Data dalam Tabel 2.2

B. REGRESI GARIS LURUS

Jika hubungan antara y dan x tepat merupakan garis lurus maka kedua variabel itu dihubungkan dengan rumus:

$$y = \alpha + \beta x$$

dengan α menunjukkan tinggi titik potong garis dengan sumbu y dan β merupakan lerengan garis itu, atau perubahan dalam y per unit perubahan dalam x (lihat Gambar 2.2).

Dalam keadaan yang bukan deterministik cukup beralasan untuk mempostulasikan bahwa hubungan linear yang melandasinya ditutupi gangguan ini di mana kita dapat merumuskan model regresi linear berikut sebagai representasi sementara model hubungan antara y dan x . Selanjutnya kita ikuti dengan analisis statistik kita.

Model statistik

Kita anggap bahwa respons Y_i berhubungan dengan nilai variabel terkendali x_i melalui

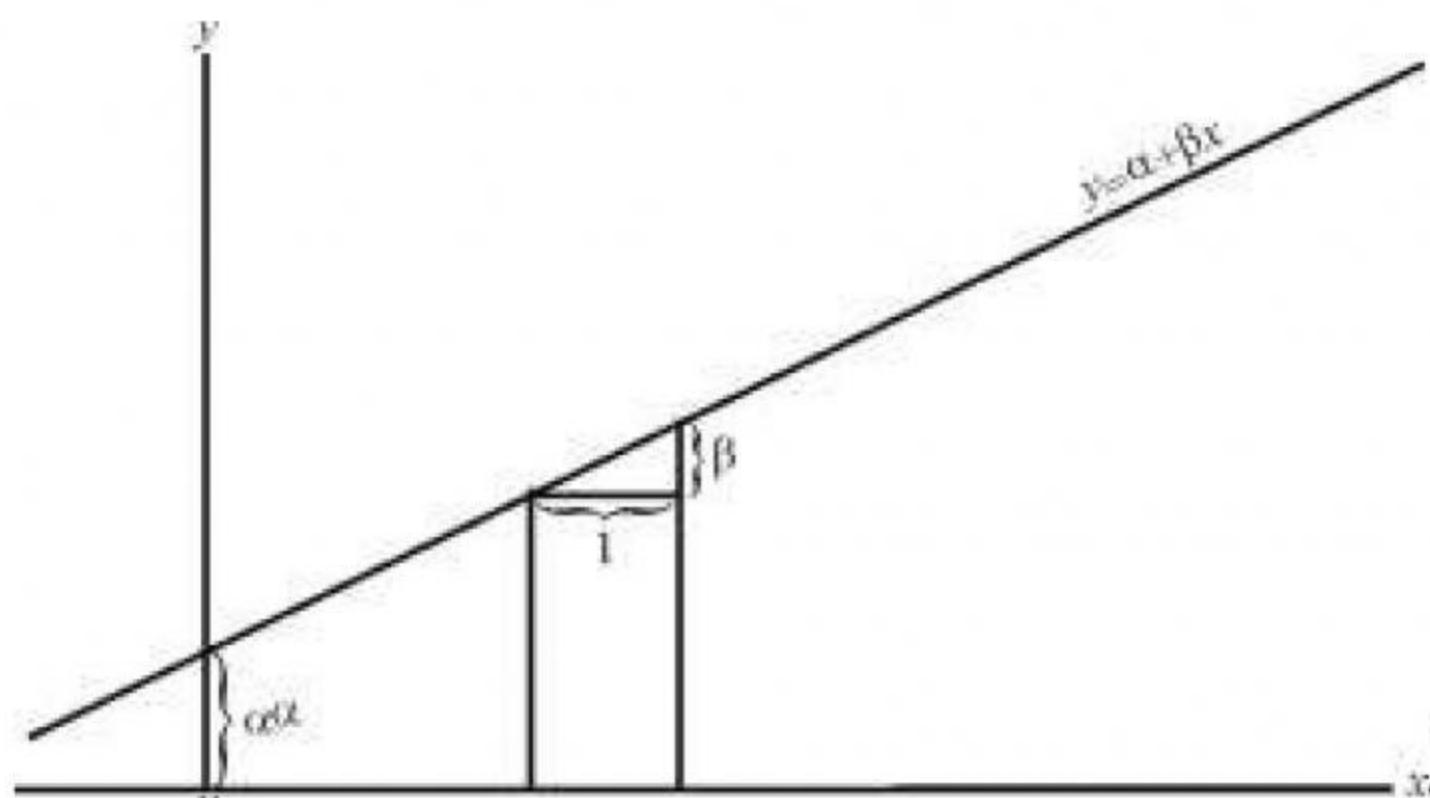
$$Y_i = \alpha + \beta x_i + e_i; i = 1, 2, \dots, n$$

dengan

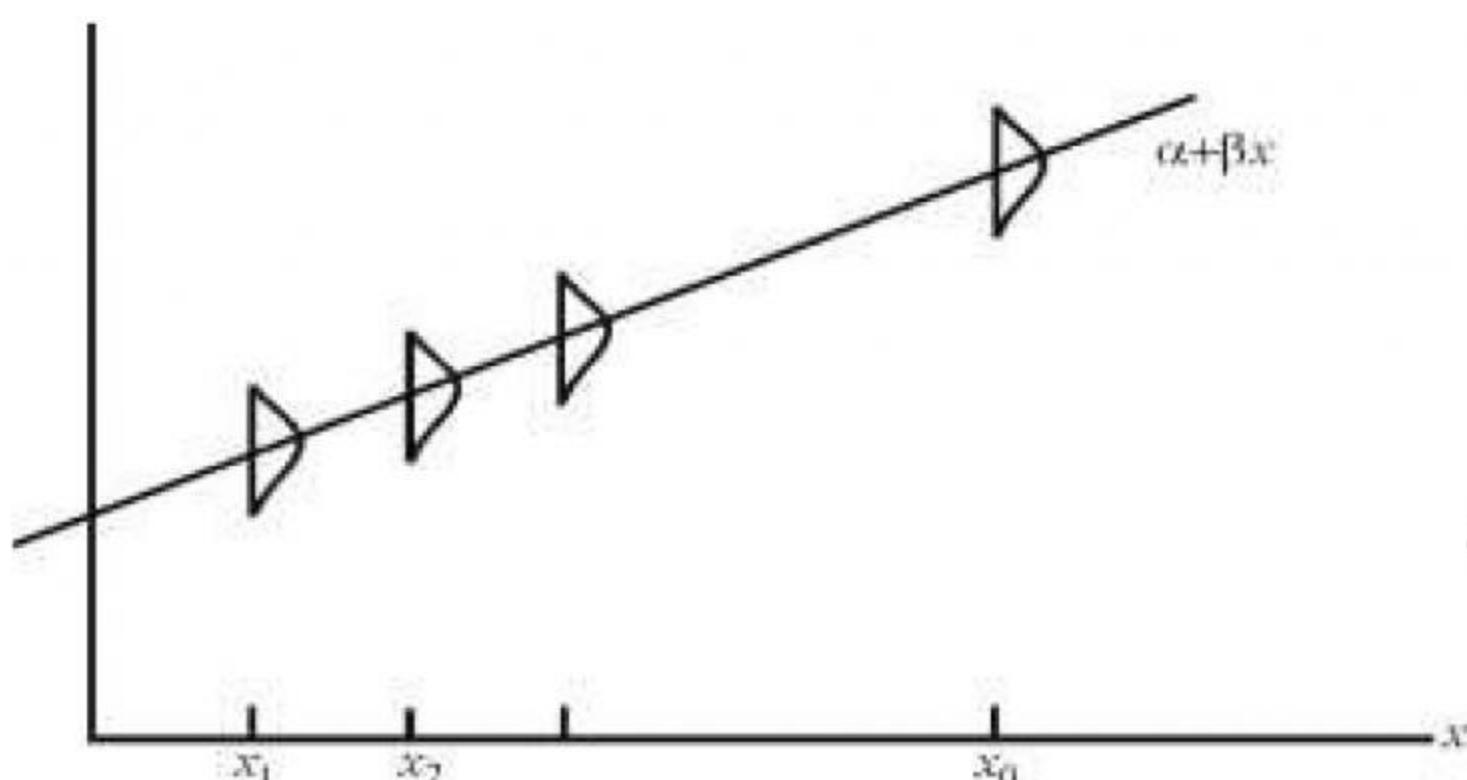
- a. x_1, x_2, \dots, x_n himpunan nilai variabel terkendali x yang dipilih pembuat percobaan untuk studi itu.
- b. e_1, e_2, \dots, e_n adalah komponen sesatan yang tidak diketahui yang ditambahkan dalam hubungan linear yang sebenarnya. Ini adalah variabel random tak teramat yang kita menganggapnya independen dan berdistribusi normal dengan mean nol dan variansi tak diketahui σ^2 .
- c. Parameter α dan β , yang bersama-sama menentukan garis lurus adalah tidak diketahui.

Menurut model ini tiap observasi Y_i yang berkaitan dengan tingkat x_i variabel terkendali adalah sampel random satu observasi dari distribusi normal dengan mean $= \alpha + \beta x_i$ dan deviasi standar $= \sigma$. Satu interpretasi dari ini jika kita mencoba mengamati nilai yang sebenarnya pada garis itu, alam menambah sesatan random e pada kuantitas ini. Struktur sesatan ini dilukiskan dalam Gambar 2.3, yang menunjukkan beberapa distribusi normal

variabel Y . Masing-masing distribusi itu mempunyai variansi sama dan mean- $mean$ -nya terletak pada garis lurus yang sebenarnya yang tidak diketahui $\alpha + \beta x$. Di samping kenyataan bahwa σ tidak diketahui, garis yang mean-mean distribusi normal ini terletak juga tidak diketahui lokasinya. Memang, satu tujuan penting analisis statistik adalah menaksir garis ini.



Gambar 2.2
Grafik Garis Lurus $y = \alpha + \beta x$

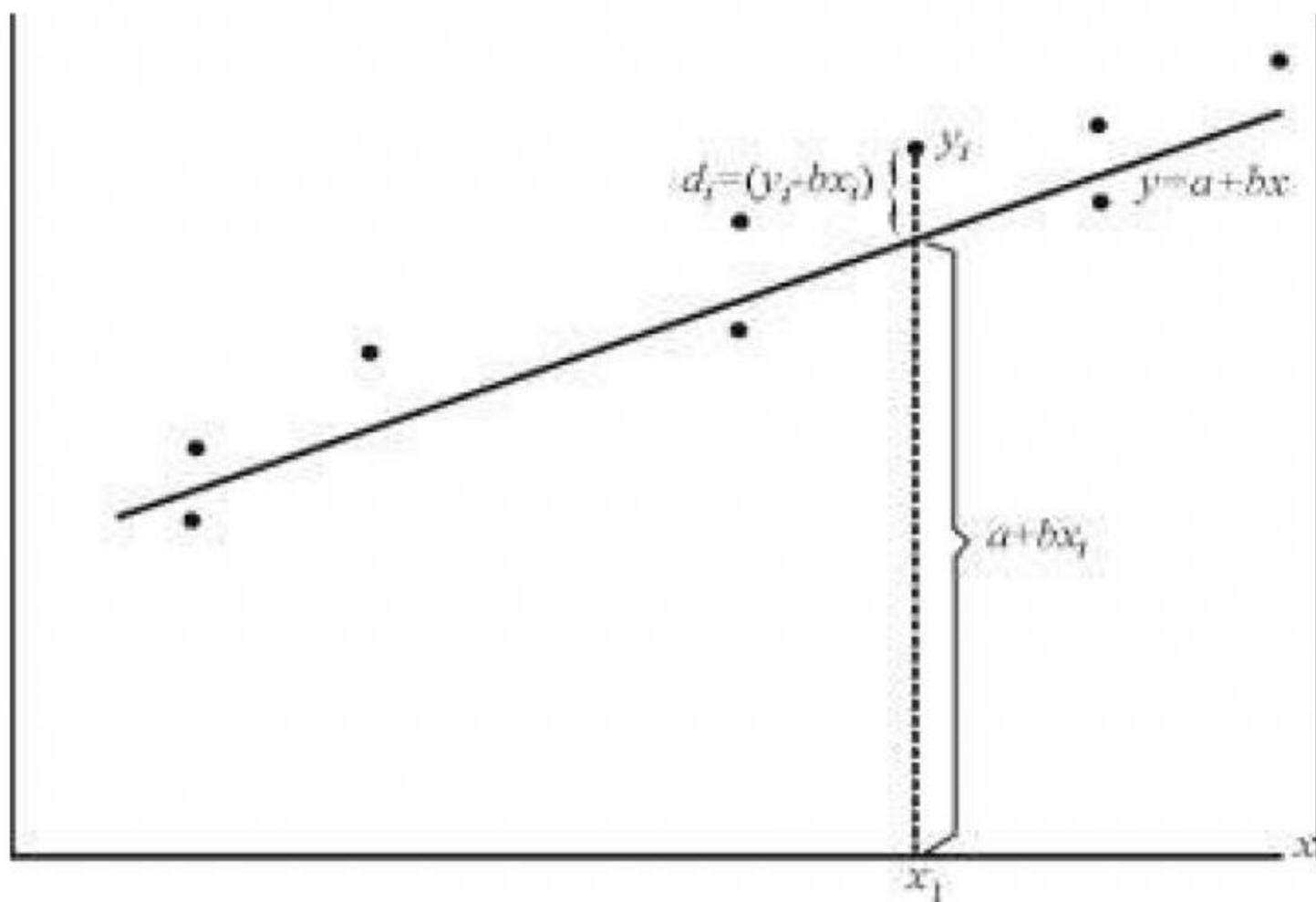


Gambar 2.3
Distribusi Normal Y dengan Mean pada Garis Lurus

C. METODE KUADRAT TERKECIL (*LEAST SQUARES*)

Jika kita anggap sementara bahwa perumusan model di atas itu benar, kita dapat melangkah selanjutnya untuk menaksir garis regresi dan menyelesaikan beberapa masalah inferensi yang berkaitan. Masalah menaksir parameter regresi α dan β dapat dipandang sebagai menaksir garis lurus terbaik pada diagram titik. Satu cara yang sederhana adalah menggeser-geser penggaris yang transparan pada diagram itu untuk menentukan secara visual garis lurus yang cukup cocok dengan data. Meskipun cara ini sederhana untuk digunakan, metode “penaksiran mata” ini mempunyai beberapa kelemahan yang serius. Sebagai prosedur yang subjektif, tidak ada ruang untuk inferensi statistik, seperti menghitung interval kepercayaan atau uji hipotesis. Kedua, metode ini tidak dapat digunakan dalam studi dengan lebih dari dua variabel karena dalam hal ini diagram titik tidak dapat digambarkan. Meskipun hanya dua variabel metode ini tidak mudah digunakan jika hubungan itu tampaknya berbentuk kurva. Metode kuadrat terkecil yang kita pelajari di sini, adalah metode yang objektif dan efisien untuk menaksir parameter regresi dan aplikasinya tidak terbatas pada model garis lurus saja.

Misalkan, garis sebarang $y = a + bx$ digambarkan pada diagram titik, seperti terlihat pada Gambar 2.4. Pada nilai variabel terkendali x_i , nilai y yang diperkirakan oleh garis ini adalah $a + bx$, sedangkan nilai y pengamatan adalah y_i . Selisih antara 2 nilai ini adalah $(y_i - a - bx) = d_i$ yang merupakan jarak vertikal titik itu dari garisnya.



Gambar 2.4
Penyimpangan Observasi dari Garis $y = a + bx$

Dengan memandang selisih-selisih semacam itu pada semua n titik, kita ambil:

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

Sebagai ukuran keseluruhan selisih titik-titik pengamatan dari garis taksirannya. Besar D jelas bergantung pada garis yang digambar itu. Dengan perkataan lain bergantung pada a dan b dan dua kuantitas yang menentukan garis itu. Sangat cocoknya garis terhadap diagram titik ditandai dengan nilai D yang sekecil mungkin. Sekarang kita nyatakan prinsip dasar yang kita ikuti di sini.

1. Prinsip Kuadrat Terkecil

Prinsip kuadrat terkecil terdiri dari menentukan nilai parameter-parameter yang tidak diketahui yang meminimumkan keseluruhan penyimpangan (selisih). Keseluruhan selisih D didefinisikan sebagai:

$D = \sum$ (respons pernyataan-respons perkiraan)² dengan respons perkiraan memuat parameter-parameter yang tidak diketahui. Nilai-nilai parameter yang ditentukan dinamakan *taksiran kuadrat terkecil*.

Untuk model garis lurus, respons perkiraan adalah $a + bx_i$ berkaitan dengan respons pengamatan y_i dan prinsip kuadrat terkecil, meliputi

penentuan a dan b yang meminimumkan $D = \sum_1^n (y_i - a - bx_i)^2$. Kuantitas a

dan b yang ditentukan dengan prinsip ini masing-masing ditulis dengan $\hat{\alpha}$ dan $\hat{\beta}$ dan dinamakan *taksiran kuadrat terkecil* parameter regresi α dan β .

Garis lurus taksiran yang paling cocok diberikan dengan rumus:

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

Guna melukiskan jalan pikiran kuadrat terkecil ini, kita gunakan data dalam Tabel 2.2 untuk menghitung D yang kita lakukan dalam Tabel 2.2. Untuk dua pilihan nilai a dan b ; $a = 0$, $b = 1$ dan $a = 2$, $b = 0,5$. Dari Tabel 2.3 kita dapat melihat bahwa lebih baik memilih $a = 2$, $b = 0,5$ daripada $a = 0$, $b = 1$ karena nilai D lebih kecil. Untungnya kita tidak harus melakukan metode coba-coba seperti ditunjukkan dalam Tabel 2.3 untuk memperoleh taksiran kuadrat terkecil $\hat{\alpha}$ dan $\hat{\beta}$.

Hasil analitis tersedia untuk penaksiran kuadrat terkecil dalam model regresi garis lurus. Untuk menyederhanakan penyajian hasil ini kita kenalkan beberapa notasi dasar:

2. Notasi Dasar

$$\bar{x} = \frac{1}{n} \sum x_i ; \quad \bar{y} = \frac{1}{n} \sum y_i$$

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2$$

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

Kuantitas \bar{x} dan \bar{y} adalah mean sampel nilai-nilai x dan y ; S_{xx} dan S_{yy} adalah jumlah kuadrat deviasi dari mean. Kita telah kenal dengan ungkapan ini dalam kaitannya dengan definisi variansi sampel, tetapi di sini

nilai-nilai x ditetapkan oleh pembuat percobaan. S_{xy} adalah jumlah hasil kali silang deviasi.

Tabel 2.3
Menghitung Nilai D untuk Dua Pilihan a dan b .

X	Y	$a = 0, b = 1$			$a = 2, b = 0,5$		
		$a + bx$	$d = y - a - bx$	d^2	$a + bx$	d	d^2
1	2,1	1	1,1	1,21	2,5	-0,4	0,16
1	2,5	1	1,5	2,25	2,5	0	0
2	3,1	2	1,1	1,21	3,0	0,1	0,01
3	3,0	3	0	0	3,5	-0,5	0,25
4	3,8	4	-0,2	0,04	4,0	-0,2	0,04
4	3,2	4	-0,8	0,64	4,0	-0,8	0,64
5	4,3	5	-0,7	0,49	4,5	-0,2	0,04
6	3,9	6	-2,1	4,41	5,0	-1,1	1,21
6	4,4	6	-1,2	1,44	5,0	-0,6	0,36
7	4,8	7	-2,2	4,48	5,5	-0,7	0,49
			D = 17,65			D = 3,20	

3. Penjabaran Taksiran Kuadrat Terkecil

Menurut prinsip kuadrat kecil, kita harus menentukan kuantitas a dan b sehingga $D = \sum (y_i - a - bx_i)^2$ minimum. Pertama-tama kita tulis:

$$y_i - a - bx_i = (y_i - \bar{y}) - b(x_i - \bar{x}) + (\bar{y} - a - b\bar{x})$$

Dengan mengkuadratkan kedua ruas kita peroleh

$$\begin{aligned} (y_i - a - bx_i)^2 &= (y_i - \bar{y})^2 + b^2(x_i - \bar{x})^2 + (\bar{y} - a - b\bar{x})^2 - 2b(x_i - \bar{x})(y_i - \bar{y}) \\ &\quad - 2b(x_i - \bar{x})(\bar{y} - a - b\bar{x}) + 2(y_i - \bar{y})(\bar{y} - a - b\bar{x}) \end{aligned}$$

Sekarang kedua ruas kita jumlahkan untuk $i = 1, 2, \dots, n$, dan mencatat bahwa kedua suku terakhir ruas kanan rumus di atas hilang setelah penjumlahan karena $\sum(x_i - \bar{x}) = 0$ dan $\sum(y_i - \bar{y}) = 0$. Maka, kita punya:

$$D = S_{yy} + b^2 S_{xx} + n(\bar{y} - a - b\bar{x}) - 2b \cdot S_{xy}$$

Sekarang kita dapat mengatur suku-suku itu dan menyelesaikan pengkuadratannya dan kita peroleh:

$$\begin{aligned}
 D &= n(\bar{y} - a - b\bar{x})^2 + b^2 S_{xx} - 2bS_{xy} + S_{yy} \\
 &= n(\bar{y} - a - b\bar{x})^2 + \left(b^2 S_{xx} - 2bS_{xy} + \frac{S_{xy}^2}{S_{xx}} \right) + S_{yy} - \frac{S_{xy}^2}{S_{xx}} \\
 &= n(\bar{y} - a - b\bar{x})^2 + \left(b\sqrt{S_{xx}} - \frac{S_{xy}}{\sqrt{S_{xx}}} \right)^2 + \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right)
 \end{aligned}$$

Suku terakhir tidak memuat a dan b . Dua suku pertama adalah bentuk kuadrat dan nilai minimumnya adalah nol. Maka, D akan minimum jika:

$$\begin{aligned}
 b\sqrt{S_{xx}} - \frac{S_{xy}}{\sqrt{S_{xx}}} &= 0 & \text{atau} & \quad b = \frac{S_{xy}}{S_{xx}} \quad \text{dan} \\
 \bar{y} - a - b\bar{x} &= 0 & \text{atau} & \quad a = \bar{y} - b\bar{x}
 \end{aligned}$$

4. Garis Regresi Kuadrat Terkecil

Dari penjabaran yang kita lakukan di atas kita peroleh rumus-rumus taksiran kuadrat terkecil:

Taksiran kuadrat terkecil untuk α : $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

Taksiran kuadrat terkecil untuk β : $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$

Taksiran $\hat{\alpha}$ dan $\hat{\beta}$, kemudian dapat digunakan untuk menentukan garis taksiran:

Garis regresi kuadrat terkecil: $\hat{y} = \hat{\alpha} + \hat{\beta}x$

Dengan menggunakan taksiran kuadrat terkecil nilai D yang minimum itu adalah:

$$S_{yy} - \frac{S_{xy}^2}{S_{xx}} = S_{yy} - \hat{\beta}^2 S_{xx}$$

Kuantitas ini dinamakan jumlah kuadrat residu atau jumlah kuadrat sesatan dan ditulis JKS.

Jadi,

Jumlah kuadrat residu atau jumlah kuadrat sesatan adalah

$$JKS = S_{yy} - \beta S_{xx} = \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

Dengan bentuk kedua mengikuti kenyataan bahwa JKS adalah keseluruhan deviasi semua observasi di sekitar garis regresi taksiran $\hat{y} = \hat{\alpha} + \hat{\beta}x$. Deviasi individual observasi y_i dari garis taksiran dinamakan *residu*. Di samping memberikan alternatif hitungan untuk JKS, residu memainkan peranan penting dalam pemeriksaan anggapan modelnya.

$$\text{Residu} = y_i - \hat{\alpha} - \hat{\beta}x_i ; i = 1, 2, \dots, n$$

Dalam menerapkan metode kuadrat terkecil pada himpunan data, akan memudahkan jika pertama-tama menghitung kuantitas dasar $\bar{x}, \bar{y}, S_{xx}, S_{yy}$ dan S_{xy} yang kita kenalkan di muka maka rumus-rumus berikutnya dapat digunakan untuk mendapatkan garis regresi kuadrat terkecil, residu, dan nilai JKS. Penghitungan untuk data yang disajikan dalam Tabel 2.2 ditunjukkan dalam Tabel 2.4.

Tabel 2.4
Penghitungan Garis Kuadrat Terkecil dan JKS

x	y	x^2	y^2	xy	$\hat{\alpha} + \hat{\beta}x$	Residu
1	2,1	1	4,41	2,1	2,387	-0,287
1	2,5	1	6,25	2,5	2,387	0,113
2	3,1	4	9,61	6,2	2,774	0,326
3	3,0	9	9,00	9,0	3,161	-0,161
4	3,8	16	14,44	15,2	3,548	0,252
4	3,2	16	10,24	12,8	3,548	-0,348
5	4,3	25	18,49	21,5	3,935	0,365
6	3,9	36	15,21	23,4	4,322	-0,422
6	4,4	36	19,36	26,4	4,322	0,078
7	4,49	49	23,04	33,6	4,709	0,091
Jumlah=39	35,1	193	130,05	152,7		0,007

$$\bar{x} = 3,9 ; \bar{y} = 3,51$$

$$S_{xx} = 193 - 10(3,9)^2 = 40,9$$

$$S_{yy} = 130,05 - 10(3,51)^2 = 6,85$$

$$S_{xy} = 152,7 - 10(3,9)(3,51) = 15,81$$

$$\hat{\beta} = \frac{15,81}{40,9} = 0,387$$

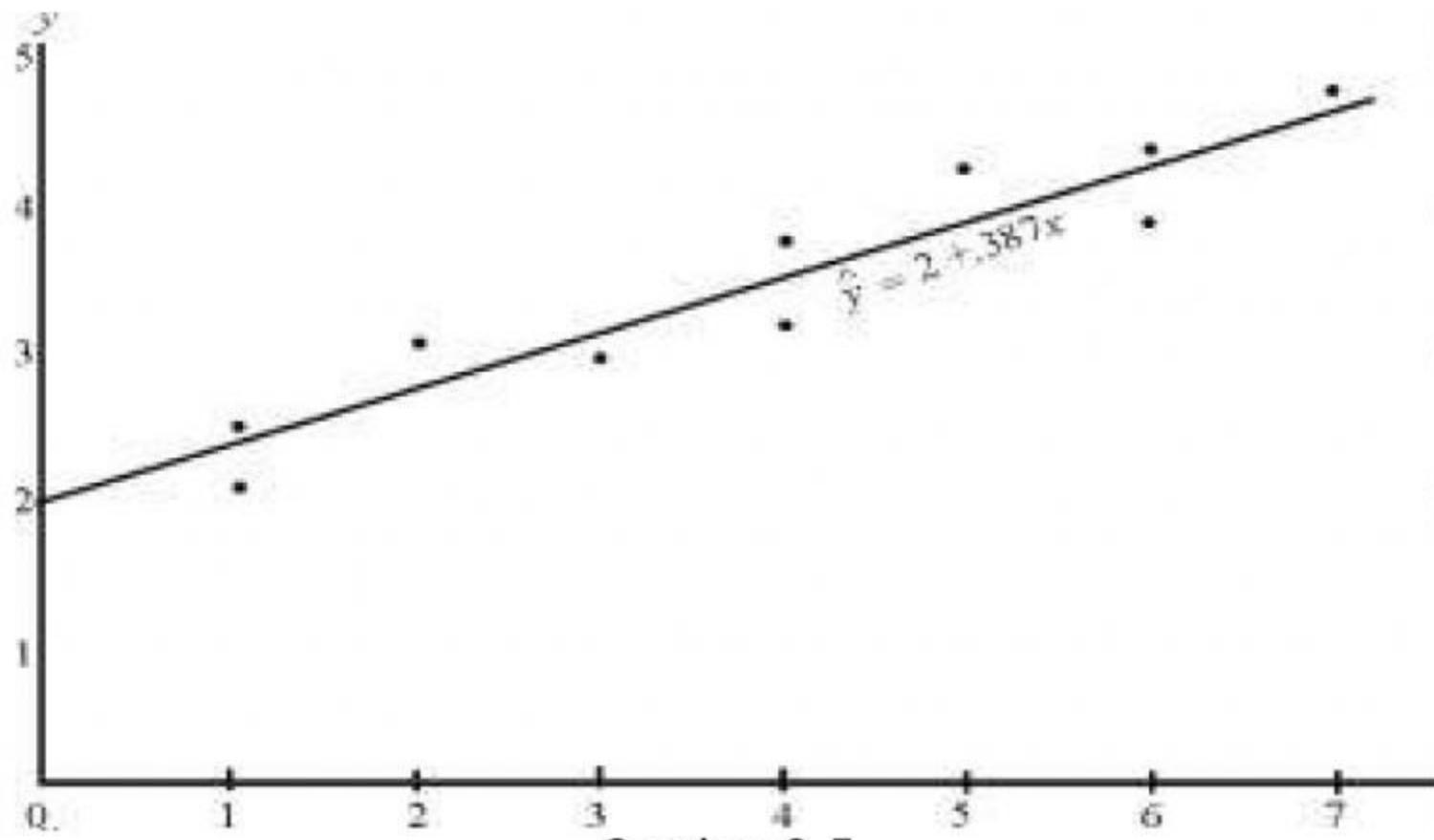
$$\hat{\alpha} = 3,51 - 0,387(3,9) = 2,00$$

$$JKS = 6,85 - \frac{(15,81)^2}{40,9} = 0,74$$

Persamaan garis taksiran dengan metode kuadrat terkecil adalah:

$$\hat{y} = 2,00 + 0,387x$$

Gambar 2.5 menunjukkan gambar data bersama dengan garis regresi kuadrat terkecil.



Gambar 2.5
Garis Regresi Kuadrat Terkecil Data Tabel 2.2

Penyimpangan $y_i - 2,00 - 0,387x_i$ nilai y_i pengamatan dari garis taksiran dihitung dalam kolom terakhir Tabel 2.4. Perlu membiarkan banyak angka di belakang koma dalam kolom residu ini untuk menunjukkan hitungan alternatif JKS sebagai jumlah kuadrat residu. Dari bentuk kedua ini kita peroleh:

$$JKS = (-0,287)^2 + (0,113)^2 + (0,326)^2 + \dots + (0,091)^2 = 0,7376$$

Perbedaan antara 0,7376 dan 0,74 disebabkan karena pembulatan. Teoretis, jumlah residu harus sama dengan nol, dan perbedaan antara jumlah 0,007 dan nol juga karena pembulatan. Bentuk penghitungan ini akan lebih sesuai dilakukan dengan komputer.



LATIHAN

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

- 1) Identifikasi nilai-nilai parameter α , β , dan σ dalam model statistik $Y = 8 - 6x + e$
Dengan e variabel random normal dengan mean = 0 dan deviasi standar 4.
- 2) Pandang model regresi linear $Y = \alpha + \beta x + e$ dengan $\alpha = 3$, $\beta = 5$, dan variabel random normal e mempunyai deviasi standar 3.
 - a. Berapa mean respons y jika $x = 4$? Jika $x = 5$?
 - b. Apakah respons pada $x = 5$ selalu lebih besar dari respons pada $x = 4$? Jelaskan!
- 3) Dipunyai enam pasang nilai (x, y)

x	1	2	3	3	4	5
y	9	5	6	3	3	1

- a. Gambarkan diagram titiknya!
 - b. Hitunglah \bar{x} , \bar{y} , S_{xx} , S_{xy} , S_{yy} !
 - c. Hitunglah taksiran kuadrat terkecil $\hat{\alpha}$ dan $\hat{\beta}$!
 - d. Tentukan garis taksiran dan gambarkan garis itu pada diagram titik!
 - e. Hitunglah nilai-nilai residu dan tunjukkan bahwa jumlahnya nol!
 - f. Hitunglah jumlah kuadrat residu (JKS) dengan:
 - (i) menjumlahkan kuadrat nilai-nilai residu
 - (ii) menggunakan rumus $JKS = S_{yy} - \hat{\beta}^2 S_{xx}$
- 4) Dihitung dari himpunan data (x, y) , dicatat nilai-nilai statistik sebagai berikut.
 $n = 14$; $\bar{x} = 3,5$; $\bar{y} = 2,32$; $S_{xx} = 10,82$; $S_{yy} = 1,035$;
 dan $S_{xy} = 2,677$
 - a) Hitunglah persamaan garis regresi taksiran!
 - b) Hitunglah jumlah kuadrat sesatan!



Dalam bentuknya yang paling sederhana, analisis regresi mempelajari bagaimana variabel respons (y) bergantung pada variabel prediktor (x).

Langkah pertama yang penting dalam mempelajari hubungan antara variabel y dan x adalah menggambar diagram titik data (x_i, y_i) , $i = 1, 2, \dots, n$. Jika gambar ini menunjukkan (mengesankan) hubungan garis lurus maka dirumuskan *model regresi garis lurus*:

Respons $=$ (garis lurus dalam x) + (sesatan random)

$$Y_i = \alpha + \beta x_i + e_i$$

Sesatan random dianggap independen, berdistribusi normal dengan mean 0 dan deviasi standar sama σ .

Parameter regresi α dan β ditaksir dengan *metode kuadrat terkecil* yang meminimumkan jumlah kuadrat deviasi $\sum(y_i - \alpha - \beta x_i)^2$. Taksiran kuadrat terkecil $\hat{\alpha}$ dan $\hat{\beta}$ menentukan garis regresi taksiran $\hat{y} = \hat{\alpha} + \hat{\beta}x$ yang dapat digunakan untuk memperkirakan nilai y dari x .

Selisih $(y_i - \hat{y}_i)$ = (nilai respons pengamatan) – (nilai respons perkiraan) dinamakan *residu*.

Rumus-rumus:

Taksiran kuadrat terkecil: $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$
 $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

Garis regresi taksiran: $\hat{y} = \hat{\alpha} + \hat{\beta}x$

Residu: $\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta}x_i$

Jumlah kuadrat sesatan: $JKS = \sum e_i^2$
 $= S_{yy} - \frac{S_{xy}^2}{S_{xx}}$

TES FORMATIF 1

Pilihlah satu jawaban yang paling tepat!

Dipunyai sembilan pasang nilai (x, y) sebagai berikut.

x	1	1	1	2	3	3	4	5	5
y	9	7	8	10	15	12	19	24	21

- 1) Kita hitung \bar{x} sama dengan
 - A. 2,43
 - B. 2,78
 - C. 2,91
 - D. 2,99

- 2) Kita hitung \bar{y} sama dengan
 - A. 10,68
 - B. 11,72
 - C. 13,89
 - D. 14,74

- 3) Kita hitung S_{xx} sama dengan
 - A. 16,62
 - B. 17,21
 - C. 19,21
 - D. 21,56

- 4) Kita hitung S_{yy} sama dengan
 - A. 304,89
 - B. 318,67
 - C. 329,92
 - D. 339,29

- 5) Kita hitung S_{xy} sama dengan
- 78,78
 - 87,87
 - 92,92
 - 99,99
- 6) Kita hitung $\hat{\beta}$ sama dengan
- 2,162
 - 2,982
 - 3,654
 - 4,276
- 7) Kita hitung $\hat{\alpha}$ sama dengan
- 2,610
 - 3,740
 - 4,120
 - 4,760
- 8) Kita hitung nilai perkiraan \hat{y} untuk $x = 3$ sama dengan
- 14,702
 - 16,706
 - 17,972
 - 18,276
- 9) Kita hitung nilai-nilai residu maka $\sum e_i$ sama dengan
- 0
 - 1
 - 2
 - 3
- 10) Kita hitung JKS sama dengan
- 15,99
 - 17,03
 - 18,27
 - 19,98

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.

$$\text{Tingkat penguasaan} = \frac{\text{Jumlah Jawaban yang Benar}}{10} \times 100\%$$

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**Inferensi Parameter Regresi****A. BEBERAPA SIFAT TAKSIRAN KUADRAT TERKECIL**

Penting untuk diingat bahwa garis regresi $\hat{y} = \hat{\alpha} + \hat{\beta}x$ yang diperoleh dengan prinsip kuadrat terkecil adalah suatu *taksiran*, berdasarkan data sampel, untuk garis regresi yang sebenarnya yang tidak diketahui $y = \alpha + \beta x$. Dalam masalah pemancaran nitrogen oksida dari mobil kita (Contoh 2.3), garis taksirannya $2 + 0,387x$. Ini menegaskan bahwa tiap unit bahan tambahan meningkatkan mean pengurangan nitrogen oksida dengan 0,387. Jika misalnya, $x = 3,2$ unit bahan tambahan dicoba kita dapat juga menggunakan garis regresi taksiran untuk menghitung banyak pengurangan taksiran sebagai $2 + (0,387)(3,2) = 3,24$. Dua pertanyaan tentang nilai taksiran ini segera timbul di sini:

1. Meskipun nilai $\hat{\beta} = 0,387$, dapatkah $\beta = 0$ sehingga y tidak bergantung pada x ? Berapakah nilai yang pantas untuk β ?
2. Seberapa besar ketidakpastian harus disertakan pada taksiran $2 + 0,387(3,2) = 3,24$ untuk titik $\alpha + \beta(3,2)$ pada garis regresi yang sebenarnya?

Untuk menjawab pertanyaan ini dan yang berkaitan, kita harus mengetahui sesuatu tentang distribusi sampling penaksir kuadrat terkecil $\hat{\alpha}$ dan $\hat{\beta}$. Untuk menghindarkan hitungan-hitungan aljabar yang panjang kita sebutkan saja distribusi ini dan sifat-sifatnya tanpa pembuktian.

1. Penaksir kuadrat terkecil adalah penaksir tak bias, yakin:

$$E(\hat{\alpha}) = \alpha \quad \text{dan} \quad E(\hat{\beta}) = \beta$$

2. $\text{Var}(\hat{\alpha}) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]$, dan $\text{Var}(\hat{\beta}) = \frac{\sigma^2}{S_{xx}}$

3. Distribusi $\hat{\alpha}$ dan $\hat{\beta}$ masing-masing adalah normal dengan mean α dan β , deviasi standarnya adalah akar dari variansi yang diberikan dalam (b).
4. $s^2 = \text{JKS}/(n-2)$ adalah penaksir tak bias untuk σ^2 . Juga, $(n-2)s^2/\sigma^2$ berdistribusi khi-kuadrat dengan derajat bebas $(n - 2)$, dan ini independen dengan $\hat{\alpha}$ dan $\hat{\beta}$.
5. Dengan mengganti σ^2 dalam (b) dengan taksiran sampel s^2 dan kita pandang akar dari variansi itu, kita peroleh sesatan standar taksiran untuk $\hat{\alpha}$ dan $\hat{\beta}$ sebagai berikut.

$$\text{SST}(\hat{\alpha}) = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

$$\text{SST}(\hat{\beta}) = \frac{s}{\sqrt{S_{xx}}}$$

6.

$$\frac{\sqrt{S_{xx}} (\hat{\beta} - \beta)}{s} \quad \text{berdistribusi t dengan db} = (n - 2)$$

$$\frac{(\hat{\alpha} - \alpha)}{s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \quad \text{berdistribusi t dengan db} = (n - 2)$$

Dapat kita sebutkan juga di sini sifat baik yang kuat taksiran kuadrat terkecil di samping sifat-sifat yang telah kita sebutkan di atas, yakni penaksir $\hat{\alpha}$ dan $\hat{\beta}$ tidak hanya tak bias, tetapi juga mempunyai variansi yang terkecil di antara semua penaksir tak bias untuk α dan β . Dengan perkataan lain $\hat{\alpha}$ dan $\hat{\beta}$ adalah penaksir tak bias yang terbaik untuk α dan β .

B. BEBERAPA MASALAH INFERENSI YANG PENTING

Sekarang kita akan mempelajari bagaimana menguji hipotesis, menghitung interval kepercayaan untuk parameter regresi α dan β .

1. Inferensi untuk Lerengan β

Dalam analisis regresi mungkin menjadi masalah yang menarik untuk menentukan apakah respons yang diharapkan bergantung atau tidak dengan besar variabel terkendali. Menurut model itu, respons yang diharapkan berhubungan dengan tingkat variabel terkendali melalui:

$$E(Y|x) = \alpha + \beta x$$

dengan $(Y|x)$ berarti respons Y berkaitan dengan tingkat variabel terkendali x tertentu. Respons yang diharapkan $\alpha + \beta x$ tidak berubah dengan berubahnya x jika dan hanya jika $\beta = 0$. Oleh karena itu, kita dapat menguji hipotesis $H_0 : \beta = 0$ terhadap alternatif satu atau dua sisi, bergantung pada sifat hubungan yang dipelajari. Menurut sifat (*f*) di atas, hipotesis nol H_0 harus diuji dengan $H_0 : \beta = 0$, statistik pengujian

$$t = \frac{\hat{\beta} \sqrt{S_{xx}}}{s} ; \text{db} = n - 2$$

Contoh 2.6

Apakah data yang tertuang dalam Tabel 2.2 mendukung dugaan bahwa bahan tambahan mengurangi banyak nitrogen oksida dalam rentang x yang dipelajari?

Untuk menjawab pertanyaan ini, kita pandang uji hipotesis $H_0 : \beta = 0$ terhadap $H_1 : \beta > 0$. Oleh karena $s^2 = \text{JKS}/(n-2) = 0,74/8 = 0,0925$ maka statistik pengujinya adalah

$$t = \frac{\hat{\beta} \sqrt{S_{xx}}}{s} = \left[\sqrt{\frac{40,9}{0,0925}} \right] (0,387) = 8,14 ; \text{db} = 8$$

Dengan derajat bebas 8, nilai tabel 5% atas adalah 1,860. Maka nilai t hitungan sangat signifikan dan H_0 ditolak. Ini berarti adanya pengurangan yang signifikan dalam nitrogen oksida jika digunakan bahan tambahan.

Perlu diingatkan di sini tentang interpretasi uji $H_0 : \beta = 0$. Jika H_0 tidak ditolak kita mungkin akan mengumpulkan bahwa y tidak bergantung pada x . Pernyataan semacam itu mungkin salah. Pertama, kita baru

menunjukkan tidak adanya hubungan linear pada rentang nilai-nilai x dalam percobaan. Di sana tidak ada dasar untuk menarik kesimpulan tentang hubungan untuk nilai-nilai x di luar rentang observasi. *Kedua*, interpretasi tidak ada ketergantungan hanya berlaku jika perumusan model kita benar. Jika diagram titik menunjukkan hubungan berbentuk kurva, tetapi dengan kurang hati-hati kita merumuskan model linear dan menguji $H_0 : \beta = 0$, kesimpulan menerima H_0 harus diinterpretasikan sebagai “tidak ada hubungan linear”, bukan “tidak ada hubungan”. Dalam modul di belakang nanti kita akan mempelajari hal ini lebih jauh. Pandangan kita saat ini adalah menganggap bahwa model itu dirumuskan dengan benar dan membicarakan berbagai masalah inferensi yang berkaitan dengan itu.

Secara lebih umum, kita dapat menguji apakah β sama atau tidak sama dengan suatu nilai tertentu, tidak harus nol.

Untuk menguji $H_0 : \beta = \beta_0$

$$\text{statistik pengujinya adalah } t = \frac{(\hat{\beta} - \beta_0) \sqrt{S_{xx}}}{s}; \text{ db} = n - 2$$

Kecuali uji hipotesis, dapat juga kita menghitung interval kepercayaan untuk parameter β menggunakan distribusi t . Misalnya,

Interval kepercayaan 95% untuk β adalah:

$\hat{\beta} \pm t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{S_{xx}}}$ dengan $t_{\frac{\alpha}{2}}$ adalah titik 2,5% atas distribusi t dengan

$\text{db} = n - 2$.

Contoh 2.7

Untuk data pengurangan dalam nitrogen oksida yang tertuang dalam Tabel 2.2, interval kepercayaan 95% untuk β adalah

$$0,387 \pm (2,306) \frac{0,304}{6,4} = 0,387 \pm 0,110 \text{ atau } (0,277; 0,497)$$

Ini berarti bahwa kita 95% yakin bahwa dengan menambahkan ekstra unit bahan tambahan kita akan mencapai mean pengurangan dalam nitrogen oksida antara 0,277 dan 0,497.

2. Inferensi tentang α

Meskipun dalam praktik kurang begitu penting, prosedur inferensi, seperti yang kita bicarakan di atas dapat kita gunakan untuk parameter α . Prosedur ini menggunakan distribusi t dengan $db = n - 2$, dinyatakan untuk $\hat{\alpha}$ dalam sifat (f) di atas; misalnya:

Interval kepercayaan 95% untuk α adalah:

$$\hat{\alpha} \pm t_{\frac{\alpha}{2}} \cdot s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

untuk uji $H_0 : \alpha = \alpha_0$ statistik pengujinya adalah

$$t = \frac{(\hat{\alpha} - \alpha_0)}{s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \quad \text{dengan } db = n - 2$$

Contoh 2.8

Untuk data pengurangan dalam nitrogen oksida dalam Tabel 2.2, interval kepercayaan 95% untuk α adalah:

$$2,00 \pm (2,306)(0,304) \sqrt{\frac{1}{10} + \frac{3,9^2}{40,9}} = 2,00 \pm 0,48 \text{ atau } (1,52 ; 2,48)$$

Untuk menguji $H_0 : \alpha = 3,00$ kita hitung statistik penguji:

$$t = \frac{2,00 - 3,00}{(0,304) \sqrt{\frac{1}{10} + \frac{3,9^2}{40,9}}} = -4,79$$

Dengan tingkat signifikansi 5% maka $H_0 : \alpha = 3,00$ ditolak.

Catatan:

1. Variabel Prediktor tidak Dapat Dikendalikan oleh Pembuat Percobaan

Analisis model regresi linear yang kita sajikan di atas berdasarkan anggapan bahwa variabel independen x tidak random. Pembuat percobaan memilih nilai-nilai x untuk diikutkan dalam percobaan, kemudian mengamati

variabel random y yang berkaitan dengan nilai-nilai x yang dipilih ini. Jenis prosedur ini cocok dalam kebanyakan percobaan yang terkendali, seperti studi tentang hubungan antara penambahan berat y dan banyak karbohidrat x yang dimakan, hasil tanaman y dan dosis pupuk x , waktu reaksi y dan besar perangsangan x . Tingkat variabel penyebab x dipilih oleh pembuat percobaan pada rentang nilai yang realistik, dan respons y pada tingkat x tertentu dipandang sebagai variabel random berdistribusi normal, seperti dilukiskan dalam Gambar 2.3 di atas.

Dalam studi percobaan yang lain yang melibatkan dua variabel x dan y , meskipun x mungkin dipandang sebagai variabel penyebab yang mempengaruhi respons y , pembuat percobaan mungkin tidak dapat memilih nilai-nilai x yang dapat dikendalikan. Dalam populasi unit-unit percobaan, x dan y keduanya dipandang sebagai variabel random yang berdistribusi peluang bersama tertentu. Pembuat percobaan memilih sampel random n unit percobaan dan mengamati pasangan nilai $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Tujuan percobaan itu masih untuk menggunakan jenis data ini guna merumuskan model untuk memperkirakan nilai harapan y jika nilai x diketahui, misalnya administrator suatu program pendidikan mungkin ingin mempelajari hubungan antara skor x yang dicapai peserta pada ujian masuk dan kinerja mereka selanjutnya y dalam program. Model perkiraan y dari x akan bermanfaat dalam perencanaan ujian masuk standar yang konsisten dengan tujuan program. Demikian juga bermanfaat dalam meramalkan kinerja dalam program. Himpunan data akan terdiri dari skor x dan skor y bagi sampel random dengan n peserta; jelas dalam keadaan ini pembuat percobaan tidak akan dapat mengendalikan nilai-nilai x .

Jika x dan y keduanya variabel random yang diamati dengan sampling random, perumusan matematik suatu model perkiraan dan syarat-syarat yang berkaitan dengan itu didasarkan atas distribusi peluang *normal bivariat*, tetapi selama x dipandang sebagai variabel penyebab yang mempengaruhi y dan tujuan pengambilan sampel adalah untuk membuat perkiraan tentang y dari nilai x , langkah-langkah operasional analisis sama seperti yang telah kita bicarakan di atas. Dengan perkataan lain, nilai pengamatan x dipandang sebagai tingkat tertentu yang percobaan itu dilakukan, dan hasilnya diinterpretasikan sebagai bersyarat pada himpunan nilai-nilai x pengamatan. Berbeda dengan itu, teknik-teknik khusus harus digunakan dalam keadaan yang kesalahan pengukuran cukup besar terlibat baik dalam mencatat nilai-nilai x maupun nilai-nilai y .

2. Komentar tentang Model

Prosedur inferensi statistik dasar yang berkaitan dengan model regresi linear dengan satu variabel independen telah kita bicarakan di atas, tetapi penting untuk diingat bahwa keabsahan prosedur ini dan kegunaan kesimpulan-kesimpulannya dibatasi oleh anggapan yang dibuat dalam perumusan model itu. Studi regresi tidak dilakukan dengan mengadakan beberapa uji hipotesis rutin dan menghitung interval kepercayaan untuk parameter-parameternya berdasarkan rumus-rumus yang telah kita pelajari. Kesimpulan-kesimpulan semacam itu dapat sangat menyesatkan jika anggapan yang dibuat dalam perumusan model terlalu tidak sesuai dengan data. Dengan demikian, perlu memeriksa data dengan cermat untuk mendapatkan petunjuk setiap penyimpangan dari anggapan. Kita ulangi, bahwa anggapan-anggapan yang terlibat dalam perumusan model garis lurus adalah:

- a. hubungannya linear;
- b. sesatan-sesatan independen;
- c. variansi konstan;
- d. distribusi normal.

Tentu saja, jika sifat umum hubungan antara y dan x berbentuk kurva, bukan garis lurus, perkiraan yang diperoleh dari menaksir model garis lurus untuk data, pasti akan merupakan hasil yang tidak berarti. Sering kali transformasi data yang cocok akan mengubah hubungan tak linear menjadi hubungan yang kira-kira berbentuk linear. Penyimpangan anggapan independen mungkin kesalahan yang paling serius karena ini secara drastis dapat mengubah kesimpulan yang ditarik dari uji t dan interval kepercayaan. Jika diagram titik menunjukkan besar fluktuasi yang berbeda-beda dalam nilai y pada tingkat x yang berbeda maka anggapan variansi konstan telah dilanggar. Di sini lagi, transformasi data yang sesuai sering kali menolong menstabilkan variansi. Akhirnya, penggunaan distribusi t dalam uji hipotesis dan taksiran interval kepercayaan berlaku selama suku sesatan berdistribusi normal. Sedikit menyimpang dari normal tidak merusak kesimpulan, khususnya jika himpunan data besar. Dengan perkataan lain, pelanggaran anggapan (d) saja tidak seserius pelanggaran anggapan yang lain yang mana saja.



LATIHAN

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

- 1) Dipunyai 5 pasang nilai (x, y) :

x	1	2	3	4	5
y	0,9	2,1	2,4	3,3	3,8

- a. Hitunglah taksiran kuadrat terkecil $\hat{\alpha}$ dan $\hat{\beta}$!
Taksirlah juga variansi sesatan σ^2 !
 - b. Ujilah $H_0 : \beta = 1$ terhadap $\beta \neq 1$ dengan tingkat signifikansi 5%!
 - c. Taksirlah nilai y harapan yang berkaitan dengan nilai $x = 3,5$!
 - d. Hitunglah interval kepercayaan 90% untuk α !
- 2) Sampel random dengan tujuh rumah yang baru-baru ini dijual di suatu kota, nilai taksiran (x) dan harga jual (y) adalah sebagai berikut.

x (ribuan dolar)	83,5	90,0	70,5	100,8	110,2	94,6	120,0
y (ribuan dolar)	88,0	91,2	76,2	107,0	111,0	99,0	118,0

- a. Gambarkan diagram titiknya!
 - b. Hitunglah persamaan regresi taksiran $\hat{y} = \hat{\alpha} + \hat{\beta}x$!
 - c. Hitunglah interval kepercayaan 95% untuk α !
 - d. Hitunglah interval kepercayaan 95% untuk β !
 - e. Lakukan uji $H_0 : \beta = 0$ dengan tingkat signifikansi 5%!
- 3) Menggunakan rumus $\hat{\beta}$ dan JKS, tunjukkan bahwa JKS dapat juga ditulis sebagai:
- a. $JKS = S_{yy} - \hat{\beta}S_{xy}$
 - b. $JKS = S_{yy} - \hat{\beta}^2 S_{xx}$

- 4) Dengan mengingat rumus-rumus $\hat{\alpha}$ dan $\hat{\beta}$, tunjukkan bahwa titik (\bar{x}, \bar{y}) terletak pada garis regresi taksiran.
- 5) Untuk melihat mengapa residu selalu berjumlah nol, kita pandang rumus-rumus $\hat{\alpha}$ dan $\hat{\beta}$, dan tunjukkan bahwa
- Nilai perkiraan \hat{y} adalah $\hat{Y}_i = \bar{y} + \hat{\beta}(x_i - \bar{x})$!
 - Residu adalah $\hat{e}_i = y_i - \hat{y}_i = (y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})$!
Selanjutnya tunjukkan bahwa $\sum \hat{e}_i = 0$!
 - Tunjukkan bahwa $\sum \hat{e}_i^2 = S_{yy} + \hat{\beta}^2 S_{xx} - 2\hat{\beta} S_{xy} = S_{yy} - S_{xy}^2 / S_{xx}$!



RANGKUMAN

Inferensi

1. Inferensi tentang lerengan β didasarkan atas penaksir $\hat{\beta}$.

Sesatan standar taksiran $= \frac{s}{\sqrt{S_{xx}}}$ dan distribusi sampling

$$t = \frac{\hat{\beta} - \beta}{s/\sqrt{S_{xx}}} ; \quad db = n - 2$$

Interval kepercayaan $100(1 - \gamma)\%$ untuk β adalah $\beta \pm t_{\frac{\gamma}{2}} \frac{s}{\sqrt{S_{xx}}}$

Untuk menguji $H_0 : \beta = \beta_0$ digunakan statistik pengujian

$$t = \frac{\hat{\beta} - \beta_0}{s/\sqrt{S_{xx}}} ; \quad db = n - 2$$

2. Inferensi tentang α berdasarkan penaksir $\hat{\alpha}$ sesatan standar

taksiran $s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$ dan distribusi sampling

$$t = \frac{\hat{\alpha} - \alpha}{s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} ; \quad db = (n - 2)$$

Interval kepercayaan $100(1-\gamma)\%$ untuk α adalah:

$$\hat{\alpha} \pm t_{\frac{\gamma}{2}} \cdot s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$



TES FORMATIF 2

Pilihlah satu jawaban yang paling tepat!

- I. Dihitung dari himpunan data nilai-nilai (x, y) , kita catat beberapa statistik sebagai berikut.

$$n = 14$$

$$\bar{x} = 3,5$$

$$\bar{y} = 2,32$$

$$S_{xx} = 10,82$$

$$S_{yy} = 1,035$$

$$S_{xy} = 2,677$$

- 1) Kita hitung persamaan regresi taksiran $\hat{y} = \hat{\alpha} + \hat{\beta}x$ dengan $\hat{\beta}$ sama dengan
 - A. 0,25
 - B. 0,24
 - C. 0,32
 - D. 0,39

- 2) Seperti soal nomor 1, tetapi $\hat{\alpha}$ sama dengan
 - A. 0,96
 - B. 1,01
 - C. 1,45
 - D. 2,21

- 3) Kita hitung taksiran untuk σ^2 , yakni s^2 sama dengan
 - A. 0,004
 - B. 0,009
 - C. 0,011
 - D. 0,031

- 4) Kita hitung interval kepercayaan 95% untuk β
 - A. (0,229 ; 0,271)
 - B. (0,231 ; 0,311)

- C. (0,241 ; 0,322)
 D. (0,244 ; 0,341)
- 5) Untuk menguji $H_0 : \beta = 0$ kita hitung statistik pengujian t , kita peroleh
 A. 9,2
 B. 11,9
 C. 13,1
 D. 15,6
- 6) Kita hitung interval kepercayaan 95% untuk α , kita peroleh
 A. (0,70 ; 2,29)
 B. (0,71 ; 2,99)
 C. (-0,86 ; 3,62)
 D. (-0,94 ; 3,84)
- II Dalam suatu eksperimen untuk menentukan hubungan antara x dan y diperoleh data yang tidak kita sajikan di sini, tetapi telah kita hitung beberapa statistik:
 $n = 25$ $\sum x = 1315$ $\sum y = 235,6$ $\sum xy = 11821,432$
- 7) Kita hitung persamaan regresi taksiran $\hat{y} = \hat{\alpha} + \hat{\beta}x$ dengan $\hat{\beta}$ sama dengan
 A. -0,0798
 B. -0,0911
 C. 0,1011
 D. 0,1101
- 8) Seperti soal nomor 7, tetapi $\hat{\alpha}$ sama dengan
 A. 9,998
 B. 10,225
 C. 11,216
 D. 13,623
- 9) Kita hitung taksiran untuk s^2 , nilai taksiran untuk σ^2 sama dengan
 A. 0,2676
 B. 0,7923

- C. 0,8112
- D. 0,9827

- 10) Kita hitung interval kepercayaan 95% untuk β
- A. $-0,1015 < \beta < -0,0581$
 - B. $-0,2011 < \beta < -0,0921$
 - C. $-0,3011 < \beta < -0,1021$
 - D. $-0,3162 < \beta < -0,1101$
- 11) Untuk menguji $H_0 : \beta = 0$ kita hitung statistik penguji t , kita peroleh
- A. 5,10
 - B. -6,20
 - C. -7,60
 - D. -8,70

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.

$$\text{Tingkat penguasaan} = \frac{\text{Jumlah Jawaban yang Benar}}{11} \times 100\%$$

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.

Kunci Jawaban Tes Formatif

Tes Formatif 1

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

Tes Formatif 2

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

Daftar Pustaka

Battacharyya, G.K. and R.A Johnson (1977). *Statistics Concepts and Methods*. New York: John Wiley.

Freud, J. (1979). *Modern Elementary Statistics*. Prentice Hall.

Kooros, A. (1965). *Elements of Mathematical Economics*. Boston: Houghton Mifflin Company.

Pfeffenberger, R. C. And J. H. Peterson. (1977). *Statistical Methods for Business and Economics*. Richard D. Irwin, Illinois.

Robbins, H. And J. V. Ryzin. (1975). *Introduction to Statistics*. Science Research Associates, Inc.

Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.