MODUL EKONOMETRIKA
(MA615830)

OLEH:
I WAYAN SUMARJAYA

PROGRAM STUDI MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS UDAYANA
BUKIT JIMBARAN
2017

KATA PENGANTAR

Modul Ekonometrika (MA615830) merupakan mata kuliah pilihan bebas bidang kompetensi. Mata kuliah ini diawali dengan pembahasan tentang konsep ekonometrika dan beberapa istilah dalam bidang ekonometrika. Bab 1 ini juga menitikberatkan perbedaan mendasar antara ekonometrika, matematika ekonomi, dan statistika. Bab 2 membahasa konsep dasar peluang dan statistika yang diperlukan mahasiswa untuk memahami ekonometrika.

Bab 3 membahas konsep regresi linear sederhana sebagai dasar dalam memahami teknik-teknik regresi dalam analisis ekonometrika. Pengembangan pada kasus berganda dibahasa pada Bab 4. Bab 5 membahas salah satu penyimpangan asumsi dalam ekonometrika yaitu heteroskedastisitas. Materi tentang analisis data panel mengakhiri topik mata kuliah ini.

Akhir kata semoga Modul Ekonometrika (MA615830) ini bermanfaat bagi mahasiswa yang mengambil mata kuliah analisis deret waktu. Segala kritik dan saran guna perbaikan modul ini harap dikirim via email ke `sumarjaya@unud.ac.id`.

Bukit Jimbaran, Februari 2017

Penulis

DAFTAR ISI

BAB I

PENGANTAR EKONOMETRIKA

Bab ini membahas konsep dasar ekonometrika, metodologi ekonometrika, dan jenis-jenis data ekonometrika. Materi pada bab ini diadaptasi dari Gujarati (2003), Hill *et al.* (2012), dan Wooldridge (2006).

1.1 Definisi Ekonometrika

Secara harfiah kata ekonometrika (*econometrics*) berarti "pengukuran ekonomi" (*economic measurement*). Namun, meskipun pengukuran merupakan bagian penting dari ekonometrika, ruang lingkup dari ekonometrika lebih luas (Gujarati, 2003). Lebih lanjut, Gujarati (2003) mengatakan bahwa ekonometrika merupakan kombinasi dari teori ekonomi, matematika ekonomi, statistika ekonomi, dan statistika matematika.

Teori ekonomi membuat pernyataan atau hipotesis yang biasanya kualitatif. Sebagai contoh, teori mikroekonomi menyatakan bahwa, hal-hal lain tetaplah sama, reduksi harga suatu komoditas diharapkan meningkatkan kuantitas permintaan untuk komoditas tersebut. Dengan demikian teori ekonomi mempostulatkan hubungan negatif atau terbalik antara harga dan jumlah permintaan dari suatu komoditas. Tetapi, teori itu sendiri tidaklah memberikan ukuran numerik tentang hubungan keduanya; artinya, teori ini tidaklah memberitahu berapa banyak kuantitas akan naik atau turun sebagai akibat dari perubahan harga komoditas. Itu merupakan tugas seorang ekonometrikawan untuk mengestimasi atau memberikan nilai angka numerik.

Matematika ekonomi, di lain pihak, berhubungan dengan bagaimana menyatakan atau mengekspresikan teori ekonomi dalam bentuk matematika (persamaan) tanpa mempertimbangkan keberukuran (*measurability*) atau verifikasi empiris dari teori. Ekonometrika, seperti pada pembahasan sebelumnya, tertarik dengan verifikasi empiris dari teori ekonomi.

Statistika ekonomi berhubungan dengan pengumpulan, pemrosesan, dan penyajian data ekonomi dalam bentuk grafik (*chart*) dan tabel. Sebagai contoh, seorang statistikawan ekonomi mengumpulkan data tentang *gross national product* atau (GNP), tenaga kerja, pengangguran, harga, dan lain-lain. Data yang dikumpulkan ini berupa data mentah untuk diolah lebih lanjut. Seorang statistikawan ekonomi tidak melakukan hal lebih jauh selain proses tersebut. Jika sang statistikawan ekonomi melakukan pengujian teori ekonomi terhadap data yang dikumpulkan maka yang bersangkutan akan menjadi ekonometrikawan.

Statistika matematika memberikan alat kepada ekonometrikawan dalam melaksanakan tugasnya. Seorang ekonometrikawan memerlukan metode khusus dalam menganalisis data ekonomi. Lebih lanjut, seorang ekonometrikawan biasanya berhubungan data *observational*, bukan data *experimental*. Hal ini berarti bahwa seorang ekonometrikawan harus menguasai keahlian yang berbeda-beda jika dibandingkan menganalisis data percobaan (*experimental*). Hal selanjutnya adalah ekonometrikawan harus membiasakan diri dengan sifat alamiah dan struktur data.

1.2 Metodologi Ekonometrika

Pada bagian ini kita akan membahas metodologi ekonometrika yang sering disebut metode tradisional atau klasik (Gujarati, 2003, p.3). Langkah-langkah metode klasik ini adalah sebagai berikut:

1. Pernyataan tentang teori atau hipotesis
2. Spesifikasi model matematika dari teori
3. Spesifikasi model statistika atau ekonometrika
4. Mendapatkan data
5. Mengestimasi parameter model ekonometrika
6. Pengujian hipotesis
7. Prediksi (*prediction*) atau peramalan (*forecasting*)
8. Menggunakan model untuk kendali atau kebijakan

Subbagian berikut mengilustrasikan langkah-langkah metode klasik di atas. Kita akan membahas teori konsumsi Keynes (Gujarati, 2003, p.4-10).

1.2.1 Pernyataan tentang teori atau hipotesis

John Maynard Keynes mengatakan bahwa hukum psikologi dasar menyatakan bahwa pria (wanita) meningkat konsumsinya sebagaimana pendapatannya meningkat, tetapi tidak sebanyak kenaikan pendapatannya. Dengan kata lain, Keynes mempostulatkan bahwa laju perubahan konsumsi untuk satu satuan unit (katakanlah, satu dolar atau rupiah) pada pendapatan, lebih besar dari 0 tetapi lebih kecil dari 1. Keynes mempostulatkan apa yang disebut sebagai *marginal prospensity to consume*, disingkat MPC.

1.2.2 Spesifikasi model matematika dari teori

Meskipun Keynes mempostulatkan hubungan positif antara konsumsi dan pendapatan, namun ia tidak menentukan secara pasti bentuk hubungan fungsional antara keduanya. Seorang matematikawan ekonomi mungkin mengusulkan bentuk fungsi konsumsi berikut:

$$y = \beta_1 + \beta_2 x, \quad 0 < \beta_2 < 1 \tag{1.1}$$

dengan y menyatakan pengeluaran konsumsi, x menyatakan pendapatan, dan β_1 serta β_2 merupakan parameter model, yang selanjutnya disebut sebagai perpotongan (*intercept*) dan lereng (*slope*). Koefisien lereng β_2 mengukur MPC.

Fungsi konsumsi pada persamaan (1.1) hanya memiliki satu persamaan sehingga disebut model persamaan tunggal (*single equation model*). Jika terdapat lebih dari satu persamaan, model ini disebut model persamaan berganda (*multiple-equation model*). Peubah yang muncul pada sisi sebelah kiri persamaan (1.1) disebut peubah takbebas (*dependent variable*) dan peubah di sebelah kanan disebut peubah bebas (*independent* atau *explanatory*). Pada persamaan (1.1) di atas konsumsi y adalah peubah takbebas sedangkan pendapatan x adalah peubah bebas.

1.2.3 Spesifikasi model ekonometrika

Model matematika pada persamaan (1.1) tidaklah terlalu menarik bagi ekonometrikawan, karena model tersebut mengasumsikan hubungan eksak atau deterministik antara konsumsi dan pendapatan. Hubungan antara peubah ekonomi umumnya tidaklah eksak. Sebagai contoh, jika kita mendapatkan data 1000 keluarga di Indonesia, katakanlah, kemudian kita memplot data tersebut dengan konsumsi pada sumbu vertikal dan pendapatan pada sumbu horizontal, maka kita tidak mengharapkan semua amatan akan terletak pada suatu garis lurus. Hal ini disebabkan peubah lain memengaruhi konsumsi seperti ukuran keluarga, usia anggota keluarga, agama keluarga, dan lain-lain.

Agar hubungan antara peubah ekonomi menjadi tidak eksak ekonometrikawan biasanya memodifikasi fungsi konsumsi deterministik pada (1.1) menjadi

$$y = \beta_1 + \beta_2 x + u \tag{1.2}$$

dengan u adalah gangguan (*disturbance* atau *error*). Gangguan ini merupakan peubah acak yang memiliki sifat-sifat peluang yang terdefinisikan dengan baik (*well-defined*). Galat u ini mungkin mewakili semua faktor yang memengaruhi konsumsi tetapi tidak disertakan secara eksplisit.

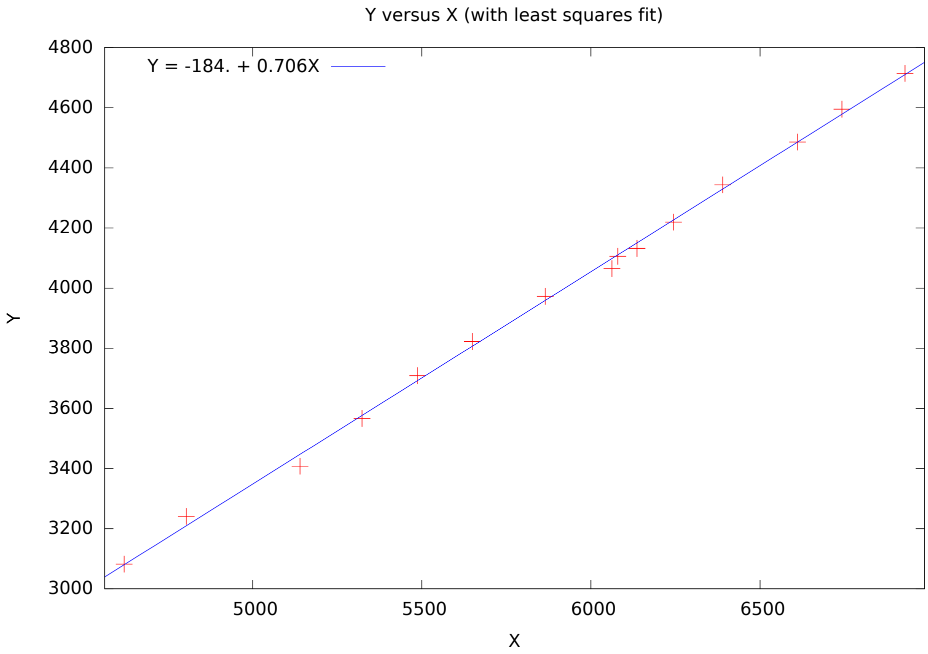
Model (1.2) disebut model ekonometrika, lebih tepatnya, model regresi linear sederhana. Pada model fungsi konsumsi (1.2) dihipotesiskan bahwa konsumsi y berhubungan secara linear dengan peubah penjelas x, tetapi hubungan antara keduanya tidaklah eksak dan tergantung pada variasi individu.

1.2.4 Mendapatkan data

Untuk mengestimasi model ekonometrika pada (1.2) kita memerlukan data. Dengan kata lain, untuk mendapatkan estimasi nilai β_1 dan β_2 kita memerlukan data. Sebagai contoh berikut ini adalah data pengeluaran dan GNP selama tahun 1982–1996. Lihat (Gujarati, 2003, p.6).

Tabel 1.2.1: Data pengeluaran individual dan GNP di Amerika Serikat selama periode 1982–1996 (dalam *billions of dollars*)

Tahun	y	x
1982	3081,5	4620,3
1983	3240,6	4803,7
1984	3407,6	5140,1
1985	3566,5	5323,5
1986	3708,7	5487,7
1987	3822,3	5649,5
1988	3972,7	5865,2
1989	4064,6	6062,0
1990	4132,2	6136,3
1991	4105,8	6079,4
1992	4219,8	6244,4
1993	4343,6	6389,6
1994	4486,0	6610,7
1995	4595,3	6742,1
1996	4714,1	6928,4



Gambar 1.2.1: Plot pengeluaran vs GNP

1.2.5 Mengestimasi parameter model ekonometrika

Langkah selanjutnya adalah mengestimasi parameter model ekonometrika. Dengan teknik analisis regresi sederhana yang nanti akan kita bahas, diperoleh estimasi untuk β_1 dan β_2, berturut-turut adalah $-184{,}08$ dan $0{,}7064$. Dengan demikian esimasi fungsi konsumsi adalah

$$\hat{y} = -184{,}08 + 0{,}7064 x_i \,. \tag{1.3}$$

Tanda topi $\hat{\cdot}$ mengindikasikan bahwa nilai tersebut adalah nilai estimasi. Sebagai konvensi, tanda topi di atas peubah atau parameter mengindikasikan bahwa itu merupakan nilai estimasi. Lalu apa arti dari masing-masing koefisien itu? Nanti, kita akan membahas lebih rinci interpretasi dari koefisien regresi tersebut. Sebagai contoh, angka 0,70 berarti bahwa untuk periode sampel tersebut kenaikan pendapatan sebenarnya sebesar 1 dollar menyebabkan, secara rata-rata, kenaikan sekitar 70 sen pada pengeluaran konsumsi sesungguhnya. Dikatakan rata-rata karena hubungannya tidak eksak.

1.2.6 Pengujian Hipotesis

Setelah model yang disuaikan (*fitted model*) adalah model yang memberikan hampiran atau pendekatan cukup bagus, kita harus membuat kriteria untuk mencari tahu apakah estimasi pada (1.3) mendukung teori yang sedang diuji. Untuk itu kita melakukan statistika inferensial (pengujian hipotesis).

1.2.7 Prediksi atau Peramalan

Jika model terpilih tidak bertentangan dengan hipotesis atau teori yang bersesuaian, kita dapat menggunakannya untuk memprediksi nilai masa depan peubah takbebas (*dependent variable*) y berdasarkan nilai peubah acak prediktor x. Sebagai contoh misalkan nilai GDP adalah 7500 maka nilai prediksi untuk konsumsi adalah

$$\hat{y} = -184{,}08 + 0{,}7064 \times 7500 = 5113{,}922. \tag{1.4}$$

Dengan demikian, diketahui nilai GDP, nilai tengah atau rata-rata prediksi untuk pengeluaran konsumsi adalah 5113,922 *billions of dollars*).

1.2.8 Menggunakan model untuk kendali atau kebijakan

Pada subbagian sebelumnya kita telah mendapatkan estimasi persamaan regresi yang dapat selanjutnya dapat digunakan untuk pengendalian atau kebijakan. Dengan kebijakan fiskal atau moneter yang tepat pemerintah dapat mengendalikan nilai peubah x untuk menghasilkan nilai target y.

1.3 Data dalam Ekonometrika

Pada bagian 1.1 kita telah membahas sekilas bagaimana seorang ekonometrikawan mendapatkan data. Secara garis besar data dapat diperoleh melalui percobaan atau eksperimen (*experimental data*). Data ini lazim dilakukan pada bidang-bidang seperti fisika, pertanian, atau bidang lain yang mana data dapat diperoleh melalui percobaan. Meskipun hal ini dimungkinkan juga dalam bidang ekonomi namun susah dilakukan. Misalnya terdapat beberapa percobaan terencana pada ilmu sosial, namun muncul kendala dalam mengorganisasi atau mendanainya (Hill *et al.*, 2012). Sebaliknya, data-data nonpercobaan (*nonexperimental data*), diperoleh tidak melalui percobaan terkontrol pada individu, perusahaan atau segmen ekonomi (Wooldridge, 2006). Data nonpercobaan ini sering disebut data amatan (*observational data*) atau data restrospektif (*retrospective data*).

1.3.1 Tipe-tipe data ekonomi

Data ekonomi dapat berupa berbagai tipe. Berikut ini tipe-tipe data ekonomi (Hill *et al.*, 2012, p.6–7):

1. Data yang dapat dikumpulkan pada beragam tingkat agregat:

   a) data mikro, yaitu data yang dikumpulkan pada unit-unit pengambilan keputusan ekonomi individu seperti individu, rumah tangga, dan perusahaan;
   b) data makro, yaitu data yang diperoleh dengan menggabungkan semua individu, rumah tangga, perusahaan pada tingkat lokal, daerah, atau nasional.

2. Data yang menyatakan suatu *flow* atau *stock*

   a) *flow*, ukuran *outcome* sepanjang periode waktu, seperti pemakaian bahan bakar bensin sepanjang triwulan terakhir 2016;
   b) *stock*, ukuran *outcome* pada waktu tertentu, seperti kuantitas minyak mentah pada tangki penyimpanan Pertamina pada tanggal 1 September 2016.

3. Data dapat berupa kuantitatif atau kualitatif:

   a) kuantitatif, data dapat dinyatakan dalam angka atau transformasi, seperti harga atau pendapatan per kapita;
   b) kuantitatif, data yang bersifat kategorik, seperti seorang konsumen yang membeli atau tidak memberli suatu produk.

1.3.2 Struktur data ekonomi

Data tampang-lintang

Data tampang-lintang (*cross-section*) adalah data yang terdiri dari suatu sampel individu, rumah tangga, perusahan, kota, negara, atau unit sampel lain yang diambil pada titik waktu tertentu. Sebagai contoh berikut ini adalah data cuplikan tampang-silang dari 526 pekerja pada tahun 1976 Wooldridge (2006).

Tabel 1.3.2: Data tampang-silang upah dan karakteristik lainnya

obsno	*wage*	*educ*	*exper*	*female*	*married*
1	3,10	11	2	1	0
2	3,24	12	22	1	1
3	3,00	11	2	0	0
4	6,00	8	44	0	1
5	5,30	12	7	0	1
⋮	⋮	⋮	⋮	⋮	⋮
525	11,56	16	5	0	1
526	3,50	14	5	1	0

Keterangan: *obsno* adalah nomor amatan, *wage* adalah upah, *educ* adalah pendidikan, *exper* adalah pengalaman, *female* adalah jenis kelamin (wanita), dan *married* adalah status pernikahan.

Data deret waktu

Data deret waktu (*time series*) adalah data yang terdiri dari satu atau lebih peubah yang dikumpulkan sepanjang waktu. Sebagai contoh adalah data GNP pada data Tabel 1.2.1. Data-data ini biasanya dilaporkan dalam harian, bulanan, triwulanan, semesteran, atau tahunan. Contoh lain adalah data saham-saham unggulan dalam IHSG.

Data panel atau longitudinal

Data panel atau longitudinal terdiri data kumpulan data deret waktu untuk setiap data tampang-silang. Lihat Tabel 1.3.3.

Struktur data ekonomi ini berhubungan pula dengan metode yang relevan dengan metode ekonometrika yang relevan.

1.4 Kausalitas dan *Ceteris paribus*

Sebagian besar pengujian dalam teori ekonomi tujuan utama seorang ekonomiwan adalah menyimpulkan bahwa satu peubah, katakanlah pendidikan, memiliki efek kausal atau menyebabkan peubah lain, katakanlah produktivitas karyawan. Hanya dengan menemukan asosiasi antara dua peubah atau lebih mungkin bisa memberikan gambaran, tetapi tanpa hubungan kausalitas hal ini tidaklah menyakinkan. Konsep *ceteris paribus* yang berarti ”faktor lain yang relevan sama” memegang peranan penting dalam analisis kausal Wooldridge (2006). Sebagai contoh dalam menganalisis permintaan konsumen, kita tertarik untuk mengetahui apakah pengaruh mengubah harga barang berpengaruh pada jumlah permintaan, sembari mengganggap bahwa

Tabel 1.3.3: Data panel dua tahun tentang statistika tindak kejahatan di kota

obsno	*city*	*educ*	*exper*	*population*	*married*	*police*
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
⋮	⋮	⋮	⋮	⋮	⋮	⋮
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

faktor lain seperti pendapatan, harga barang lain, dan selera individu tetap. Jika faktor lain ini tidak dianggap tetap, kita tidak bisa mengetahui hubungan kausal perubahan harga pada jumlah permintaan. Mengingat sifat ”nonpercobaan” pada data bidang ilmu-ilmu sosial, mengungkap hubungan kausal adalah tantangan.

1.5 Latihan

1. Carilah contoh data tampang lintang (*cross section*), deret waktu (*time series*), dan data panel. Data dapat diambil dari Internet atau sumber-sumber lain yang relevan. Kemudian klasifikasikan data yang Anda peroleh; apakah termasuk data mikro, makro, *flow*, *stock*, kuantitatif, atau kualitatif. Pengklasifikasian bisa lebih dari satu.

2. Dengan cara yang sama klasifikasikan data yang saya unggah pada situs e-learning Unud dan Google Drive.

BAB II

TINJUAN KONSEP PELUANG DAN STATISTIKA

Bab ini meninjau kembali konsep peluang dan statistika yang akan digunakan dalam mata kuliah ini. Materi ini diadaptasi dari buku-buku berikut: Gujarati (2003), Wooldridge (2006), dan, Stock and Watson (2012), dan Hill *et al.* (2012).

2.1 Operator Penjumlahan dan Perkalian

Beberapa sifat penting operator penjumlahan diambil dari (Gujarati, 2003, p.869-912):

1. $\sum_{i=1}^{n} k = nk$, dengan k adalah konstanta.
2. $\sum_{i=1}^{n} kx_i = k\sum_{i=1}^{n} x_i$, dengan k adalah konstanta.
3. $\sum_{i=1}^{n}(a + bx_i) = na + b\sum_{i=1}^{n} x_i$, dengan a dan b adalah konstanta.
4. $\sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$.
5. $\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} = \sum_{j=1}^{m}\sum_{i=1}^{n} x_{ij}$.
6. $\sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j = \sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j$
7. $\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} + y_{ij}) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}$.
8. $(\sum_{i=1}^{n} x_i)^2 = \sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} x_i x_j = \sum_{i=1}^{n} x_i^2 + 2\sum_{i<j} x_i x_j$

2.2 Peluang

2.2.1 Fungsi Densitas Peluang

Misalkan X adalah fungsi densitas peluang diskret yang bernilai $x_1, x_2, \ldots, x_n$. Fungsi densitas peluang diskret X didefinisikan sebagai

$$f(x) = \begin{cases} P(X = x_i), & \text{untuk } i = 1, 2, \ldots, n, \\ 0, & \text{untuk } x \neq x_i. \end{cases} \tag{2.1}$$

Misalkan X adalah fungsi densitas peluang kontinu. Fungsi $f(x)$ dikatakan fungsi densitas peluang kontinu X jika kondisi berikut dipenuhi:

1. $f(x) \geq 0$,
2. $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$,
3. $\int_a^b f(x)\,\mathrm{d}x = P(a \leq x \leq b)$, $f(x)\,\mathrm{d}x$ dikatakan elemen peluang (*probability elements*).

2.2.2 Fungsi densitas peluang bersama

Misalkan X dan Y adalah dua peubah acak diskret. Maka fungsi

$$f(x,y) = \begin{cases} P(X = x \text{dan} Y = y) & \\ 0 & \text{jika } X \neq x \text{ dan } Y \neq y \end{cases} \tag{2.2}$$

disebut fungsi densitas peluang bersama diskret. Berhubungan dengan $f(x, y)$ adalah fungsi densitas peluang marginal atau individual $f(x)$ dan $f(y)$. Fungsi marginal ini didefinisikan sebagai berikut

$$f(x) = \sum_y f(x,y) \tag{2.3}$$

dan

$$f(y) = \sum_x f(x,y). \tag{2.4}$$

Selain itu, kita ingat kembali peluang bersyarat yang nanti akan sering digunakan dalam analisis regresi.

$$f(x|y) = P(X = x|Y = y) \tag{2.5}$$

dan

$$f(y|x) = P(Y = y|X = x). \tag{2.6}$$

Fungsi densitas peluang bersyarat ini dapat pula diperoleh dari hubungan

$$f(x|y) = \frac{f(x,y)}{f(y)} \tag{2.7}$$

dan

$$f(y|x) = \frac{f(x,y)}{f(x)}. \tag{2.8}$$

Pada kasus kontinu, fungsi $f(x, y)$ dikatakan fungsi densitas peluang bersama apabila syarat-syarat berikut dipenuhi:

1. $f(x, y) \geq 0$,
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}x\,\mathrm{d}y = 1$,
3. $\int_a^b \int_a^b f(x, y)\,\mathrm{d}x\,\mathrm{d}y = P(a \leq x \leq b, c \leq y \leq d)$.

Fungsi densitas peluang marginal diperoleh dengan mengintegralkan. Dengan demikian, fungsi densitas peluang marginal X dan Y adalah

$$f(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y \tag{2.9}$$

dan

$$f(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x. \tag{2.10}$$

2.3 Saling Bebas secara Statistika

Dua peubah bebas X dan Y dikatakan saling bebas secara statistika apabila

$$f(x,y) = f(x)f(y) \tag{2.11}$$

yaitu peluang bersama dinyatakan sebagai hasil perkalian marginal.

2.4 Nilai Harapan

Nilai harapan didefinisikan sebagai

$$E(X) = \begin{cases} \sum_x x f(x), & \text{jika } x \text{ diskret;} \\ \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}\,x, & \text{jika } x \text{ kontinu.} \end{cases} \tag{2.12}$$

Berikut ini adalah sifat-sifat nilai harapan:

1. Nilai harapan suatu konstanta adalah konstanta itu sendiri. Dengan demikian, jika c adalah suatu konstanta $\mathrm{E}(c) = c$.
2. Jika a dan b adalah konstanta, maka $\mathrm{E}(aX + b) = a\,\mathrm{E}(X) + b$.
3. Jika X dan Y adalah peubah acak yang saling bebas, maka $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$.
4. Jika X adalah peubah acak dengan fungsi densitas peluang $f(x)$ dan $g(x)$ adalah fungsi dari X, maka

$$\mathrm{E}(g(x)) = \begin{cases} \sum_x g(x) f(x), & \text{jika } X \text{ diskret;} \\ \int\limits_{-\infty}^{\infty} g(x) f(x)\,\mathrm{d}x, & \text{jika } X \text{ kontinu.} \end{cases} \tag{2.13}$$

2.4.1 Sifat-sifat varians

Varians didefinisikan sebagai

$$\mathrm{var}(X) = \mathrm{E}(X - \mathrm{E}(X))^2 = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2. \tag{2.14}$$

Berikut ini adalah sifat-sifat varians:

1. Varians dari suatu konstanta adalah nol. Dengan kata lain $\mathrm{var}(c) = 0$, dengan c adalah konstanta.
2. Jika a dan b adalah konstanta, maka $\mathrm{var}(aX + b) = a^2\,\mathrm{var}(X)$.
3. Jika X dan Y adalah peubah acak saling bebas maka
   a) $\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y)$,
   b) $\mathrm{var}(X - Y) = \mathrm{var}(X) + \mathrm{var}(Y)$.
4. Jika X dan Y adalah peubah-peubah acak yang saling bebas dan a dan b adalah konstanta, maka

$$\mathrm{var}(aX + bY) = a^2\,\mathrm{var}(X) + b^2\,\mathrm{var}(Y). \tag{2.15}$$

2.4.2 Sifat-sifat kovarians

Kovarians dua peubah acak X dan Y didefinisikan sebagai

$$\mathrm{cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y). \tag{2.16}$$

Berikut ini sifat-sifat penting kovarians:

1. Jika X dan Y saling bebas, maka kovariansnya nol.
2. Kovarians $\mathrm{cov}(a + bX, c + dY) = bd\,\mathrm{cov}(X, Y)$

2.4.3 Sifat-sifat korelasi

Koefisien korelasi $\mathrm{cor}(X, Y)$, sering ditulis sebagai ρ, didefinisikan sebagai

$$\rho = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}. \tag{2.17}$$

Koefisien korelasi ρ mengukur asosiasi **linear** antara dua peubah acak dan berada di antara -1 dan $+1$. Tanda -1 menunjukkan asosiasi negatif sempurna dan $+1$ mengindikasikan asosiasi positif sempurna.

Berikut ini sifat-sifat peubah yang berkorelasi (*correlated variables*):

1. Misalkan X dan Y dua peubah acak. Maka

   a) $\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y) + 2\,\mathrm{cov}(X, Y)$,

   b) $\mathrm{var}(X - Y) = \mathrm{var}(X) + \mathrm{var}(Y) - 2\,\mathrm{cov}(X, Y)$.

2. Varians dari kombinasi linear

$$\mathrm{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{var}(X_i) + 2 \sum \sum_{i<j} \mathrm{cov}(X_i, X_j). \tag{2.18}$$

2.4.4 Harapan dan varians Bersyarat

Misalkan $f(x, y)$ adalah fungsi densitas peluang bersama X dan Y. Harapan bersyarat X diketahui $Y = y$ didefinisikan sebagai

$$\mathrm{E}(X|Y = y) = \begin{cases} \sum_x x f(x|Y = y), & \text{jika } X \text{ diskret;} \\ \int_{-\infty}^{\infty} x f(x|Y = y)\,\mathrm{d}x, & \text{jika } X \text{ kontinu.} \end{cases} \tag{2.19}$$

Definisi untuk $\mathrm{E}(Y|X = x)$ juga diperoleh dengan cara serupa. Varians bersyarat X diketahui $Y = y$ didefinisikan sebagai

$$\mathrm{var}(X|Y = y) = \begin{cases} \sum_x [X - \mathrm{E}(X|Y = y)]^2 f(x|Y = y), & \text{jika } X \text{ diskret;} \\ \int_{-\infty}^{\infty} [X - \mathrm{E}(X|Y = y)]^2 f(x|Y = y)\,\mathrm{d}x, & \text{jika } X \text{ kontinu.} \end{cases} \tag{2.20}$$

2.4.5 Sifat-sifat harapan dan varians bersyarat

Berikut ini sifat-sifat penting harapan dan varians bersyarat:

1. Jika $f(X)$ adalah fungsi dari X, maka $\mathrm{E}(f(X)|X) = f(X)$, artinya fungsi X sebagai konstanta dalam perhitungan ekspektasinya. Misalnya $\mathrm{E}(X^3|X) = X^3$ karena X diketahui maka X^3 juga diketahui.

2. Jika $f(X)$ dan $g(X)$ adalah fungsi dari X, maka

$$\mathrm{E}[f(X)Y + g(X)|X] = f(X)\,\mathrm{E}(Y|X) + g(X). \tag{2.21}$$

   Contoh $\mathrm{E}[XY + cX^2|X] = X\,\mathrm{E}(Y|X) + cX^2$ dengan c adalah konstanta.

3. Hukum ekspektasi iterasi. Hukum ini menyatakan bahwa

$$\mathrm{E}(Y) = \mathrm{E}[\mathrm{E}(Y|X)]. \tag{2.22}$$

Bentuk ini dapat diperluas menjadi (Wooldridge, 2006)

$$\mathrm{E}(Y|X) = \mathrm{E}[\mathrm{E}(Y|X,Z)|X]. \tag{2.23}$$

4. Jika X dan Y saling bebas maka $\mathrm{var}(Y|X) = \mathrm{var}(Y)$.

5. Varians tak bersyarat

$$\mathrm{var}(Y) = \mathrm{E}(\mathrm{var}(Y|X)) + \mathrm{var}(\mathrm{E}(Y|X)). \tag{2.24}$$

2.4.6 Momen distribusi peluang yang lebih tinggi

Kita telah meninjau nilai tengah (mean) dan varians. Selain ukuran itu ada ukuran momen yang lebih tinggi yaitu momen ke-r yang didefinisikan sebagai

$$\mathrm{E}(X - \mathrm{E}(X))^r. \tag{2.25}$$

Sebagai momen khusus adalah kepencongan (*skewness*) yang didefinisikan oleh

$$\mathrm{Sk} = \frac{\mathrm{E}(X - \mathrm{E}(X))^3}{\sigma^3} \tag{2.26}$$

dan kurtosis yang didefinisikan sebagai

$$\mathrm{Kur} = \frac{\mathrm{E}(X - \mathrm{E}(X))^4}{[\mathrm{E}(X - \mathrm{E}(X))^2]^2}. \tag{2.27}$$

Sebagai catatan apabila nilai $\mathrm{Sk} = 0$ artinya distribusi data simetrik, $\mathrm{Sk} < 0$ berarti data memiliki ekor panjang di sebelah kiri, dan $\mathrm{Sk} > 0$ berarti data memiliki ekor panjang di sebelah kanan. Distribusi dengan kurtosis besar memiliki nilai yang terpusat pada nilai tengah (*mean*) dan memiliki puncak tengah yang relatif tingg, sedangkan untuk distribusi dengan kurtosis lebih kecil memiliki bentuk lebih datar. Sebagai pembanding nilai kurtosis yang digunakan adalah kurtosis distribusi normal yang bernilai 3.

2.4.7 Distribusi-distribusi peluang penting

Berikut ini adalah distribusi peluang penting.

Distribusi normal

Peubah acak X dikatakan berdistribusi normal (atau Gauss) dengan parameter μ dan σ^2, biasa ditulis $N(\mu, \sigma^2)$ jika memiliki fungsi densitas peluang berbentuk

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{\sigma^2}\right], -\infty < x < \infty. \tag{2.28}$$

Selanjutnya didefinisikan transformasi $Z = (X - \mu)/\sigma$ maka

$$f(z) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}z^2\right) \tag{2.29}$$

Sifat-sifat distribusi normal:

1. Simetrik pada sekitar nilai tengah.

2. Sekitar 68% area kurva normal terletak antara nilai $\mu \pm \sigma$, sekitar 95% area berada pada nilai $\mu \pm 2\sigma$, dan sekitar 99,7% area berada pada nilai $\mu \pm 3\sigma$.

3. Jika $X_1 \sim N(\mu_1, \sigma_1^2)$ dan $X_2 \sim N(\mu_2, \sigma_2^2)$ dan diasumsikan keduanya saling bebas. Kombinasi linear $Y = aX_1 + bX_2$ dengan a dan b adalah konstanta, maka

$$Y \sim N[(a\mu_1 + b\mu_2), (a^2\sigma_1^2 + b^2\sigma_2^2)]. \tag{2.30}$$

4. Teorema limit pusat (*central limit theorem*). Misalkan $X_1, X_2, \ldots, X_n$ adalah peubah acak saling bebas masing-masing dengan fungsi densitas peluang dengan nilai tengah (*mean*) μ dan varians σ^2. Misalkan $\bar{X} = \sum_{i=1}^{n} X_i/n$ maka

$$\lim_{n\to\infty} \bar{X} \sim N(\mu, \sigma^2/n). \tag{2.31}$$

   Catatan, ini berlaku tetap berlaku dan tidak tergantung pada fungsi densitas peluang tertentu. Selanjutnya

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1). \tag{2.32}$$

5. Momen ketiga dan keempat distribusi ini adalah

$$\mathrm{E}(X - \mu)^3 = 0 \tag{2.33}$$

   dan

$$\mathrm{E}(X - \mu)^4 = 3\sigma^4. \tag{2.34}$$

6. Untuk distribusi normal kepencongan bernilai 0 dan kurtosis bernilai 3. Dengan demikian uji kenormalan dapat dihitung dengan informasi ini seperti uji kenormalan Jarque-Bera.

Distribusi khi-kuadrat

Misalkan $Z_1, \ldots, Z_k$ adalah distribusi normal baku (yaitu $N(0, 1)$), maka kuantitas $\sum_{i=1}^{k} Z_i^2$ dikatakan berdistribusi khi-kuadrat dengan derajat bebas k. Fungsi densitasnya dinyatakan sebagai

$$f(z, k) = \frac{2^{-k/2}}{\Gamma(k/2)} x^{k/2-1}\, \mathrm{e}^{-x/2}, \quad x > 0. \tag{2.35}$$

Beberapa sifat penting distribusi khi-kuadrat:

1. Distribusi khi-kuadrat bersifat pencong (*skew*), derajat kepencongannya tergantung pada derajat kebebasan. Distribusi ini pencong atau condongke kanan.

2. Nilai tengah distribusi khi-kuadrat adalah k dan variansnya $2k$ dengan k adalah derajat bebas.

3. Jika X_1 dan X_2 adalah peubah acak khi-kuadrat saling bebas dengan derajat bebas k_1 dan k_2 maka $X_1 + X_2$ juga merupakan distribusi khi-kuadrat dengan derajat bebas $k_1 + k_2$.

Distribusi Student *t*

Jika Z_1 berdistribusi normal baku dan Z_2 berdistribusi khi-kuadrat dengan derajat bebas k dan saling bebas dengan Z_1 maka peubah acak

$$\frac{Z_1}{\sqrt{Z_2/k}} \sim t_k \tag{2.36}$$

yaitu berdistribusi t dengan derajat bebas k. Distribusi ini memiliki nilai tengah 0 dan varians $k/(k-2)$.

Distribusi F

Jika Z_1 dan Z_2 peubah acak khi-kuadrat dengan saling bebas dengan derajat bebas masing-masing k_1 dan k_2, maka peubah acak

$$\frac{Z_1/k_1}{Z_2/k_2} \sim F_{k_1,k_2}. \tag{2.37}$$

Beberapa sifat penting:

1. Distribusi F pencong ke kanan.
2. Nilai tengah distribusi F adalah $k_2/(k_2-2)$ dan variansnya

   $$\frac{2k_2^2(k_1+k_2-2)}{k_1(k_2-1)^2(k_2-4)}. \tag{2.38}$$

   Catatan nilai tengah terdefinisikan untuk $k_2 > 2$ dan variansnya untuk $k_2 > 4$.
3. Hubungan distribusi t dan F: $t_k^2 = F_{1,k}$.
4. Jika penyebut k_2 cukup besar, maka hubungan berikut berlaku: $k_1 F \sim \chi^2_{k_1}$.

2.5 Inferensi Statistika

Pada bagian ini kita akan membicarakan konsep penting dalam statistika inferensial.

2.5.1 Estimasi Titik

Misalkan X adalah sampel acak dengan fungsi densitas peluang $f(x;\theta)$, dengan θ adalah parameter dari distribusi. Untuk mempermudah pembicaraan kita anggap bahwa terdapat satu parameter yang tidak diketahui. Asumsikan bahwa kita tahu bentuk fungsional atau teoretis fungsi densitas peluangnya katakanlah t atau F. Namun, kita tidak tahu nilai θ. Oleh karena itu, kita mengambil sampel acak berukuran n dari fungsi densitas peluang ini dan membuat suatu fungsi dari nilai sampel seperti

$$\hat{\theta} = f(x_1, \ldots, x_n) \tag{2.39}$$

yang memberikan estimasi terhadap nilai θ sesungguhnya. Nilai $\hat{\theta}$ disebut statistik (*statistic*) atau penduga (*estimator*). Nilai tertentu dari penduga ini disebut suatu dugaan (*estimate*).

Statistik $\hat{\theta}$ dapat dianggap sebagai peubah acak karena merupakan fungsi dari sampel. Sebagai contoh

$$\hat{\theta} = \frac{x_1 + \cdots + x_n}{n} = \bar{x} \tag{2.40}$$

adalah penduga untuk nilai tengah sesungguhnya, katakanlah μ. Penduga $\hat{\theta}$ yang diperoleh ini disebut sebagai penduga titik (*point estimator*) karena hanya memberikan satu titik tunggal nilai θ.

2.5.2 Estimasi Interval

Daripada mendapatkan nilai tunggal untuk estimasi θ, kita bisa mendapatkan nilai estimasi θ dengan mengonstruksikan dua penduga $\hat{\theta}_1(x_1, \ldots, x_n)$ dan $\hat{\theta}_2(x_1, \ldots, x_n)$ dan katakan dengan kepercayaan (peluang) bahwa $\hat{\theta}_1$ dan $\hat{\theta}_2$ mengandung nilai θ sesungguhnya. Lebih jelasnya, pada estimasi interval ini kita mengonstruksikan dua penduga $\hat{\theta}_1$ dan $\hat{\theta}_2$ sedemikian hingga

$$P(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2) = 1 - \alpha, \quad 0 < \alpha < 1. \tag{2.41}$$

Interval ini disebut dengan interval kepercayaan atau selang kepercayaan $1 - \alpha$ untuk θ. Nilai $1 - \alpha$ disebut koefisien kepercayaan. Jika $\alpha = 0{,}05$; maka $1 - \alpha = 0{,}95$ artinya jika kita mengonstruksi interval kepercayaan dengan koefisien kepercayaan 0,95, maka mengulang konstruksi tersebut dari percobaan berulang kita akan benar 95 dari 100 kasus. Nilai α disebut tingkat signifikansi.

2.6 Metode Estimasi

Secara garis besar terdapat tiga metode pendugaan atau estimasi yaitu: metode kuadrat terkecil (*least squares*), metode kemungkinan maksimum (*maximum likelihood*), dan metode momen (*method of moments*) dan perluasannya. Pada bagian ini kita akan memusatkan perhatian pada metode kemungkinan maksimum (*maximum likelihood*).

2.6.1 Metode kemungkinan maksimum

Misalkan peubah acak X memiliki fungsi densitas peluang $f(x, \theta)$ yang bergantung pada parameter tunggal θ. Kita tahu fungsi densitas peluangnya, namun tidak tahu nilai parameternya. Misalkan terdapat sampel acak berukuran n. Fungsi densitas peluang bersama dari n sampel ini adalah

$$f(x_1, \ldots, x_n; \theta). \tag{2.42}$$

Karena ini merupakan sampel acak, kita dapat menuliskan (2.42) sebagai hasil kali dari fungsi densitas peluang individual sebagai

$$f(x_1, \ldots, x_n; \theta) = f(x_1; \theta) \cdots f(x_n; \theta). \tag{2.43}$$

Fungsi densitas peluang bersama ini memiliki interpretasi ganda. Jika θ diketahui, kita interpretasikan sebagai peluang bersama mengamati nilai sampel yang diberikan. Di lain pihak, kita bisa mengganggap sebagai fungsi dari θ untuk nilai $x_1, \ldots, x_n$. Nah, interpretasi kedua inilah kita katakan fungsi densitas peluang bersama sebagai fungsi kemungkinan (*likelihood function*) dan menuliskan fungsi ini sebagai

$$L(\theta; x_1, \ldots, x_n) = f(x_1; \theta) \cdots f(x_n; \theta). \tag{2.44}$$

Amati peran θ dalam hal ini, dalam fungsi densitas peluang bersama maupun fungsi likelihood.

Penduga kemungkinan maksimum dari θ adalah nilai θ yang memaksimumkan fungsi kemungkinan sampel $L(\theta; x_1, \ldots, x_n)$. Biasanya diambil log dari kemungkinan ini, kemudian kita turunkan terhadap parameter yang tidak diketahui. Selanjutnya disamakan dengan nol. Nilai penduga yang memenuhi nilai ini disebut penduga kemungkinan maksimum (*maximum likelihood estimator*).

Contoh 2.6.1. Misalkan peubah acak X berdistribusi Poisson dengan nilai tengah (mean) λ. Misalkan $X_1, \ldots, X_n$ adalah peubah acak Poisson dengan nilai tengah λ. Misalkan kita ingin mencari penduga kemungkinan maksimum untuk λ. Fungsi kemungkinannya adalah

$$L(x_1, x_2, \ldots, x_n) = \frac{\mathrm{e}^{-\lambda}\lambda^{x_1} \cdots \mathrm{e}^{-\lambda}\lambda^{x_n}}{x_1!x_2! \cdots x_n!} \tag{2.45}$$

$$= \frac{\mathrm{e}^{-n\lambda}\lambda^{\sum x_i}}{x_1! \cdots x_n!}. \tag{2.46}$$

Kemudian, ambil logaritma dari persamaan (2.46), maka persamaan tersebut menjadi

$$\log L(x_1, \ldots, x_n) = -n\lambda + \sum x_i \log \lambda - \log \prod x_i!. \tag{2.47}$$

Kemudian menurunkan persamaan (2.47) terhadap λ dan menyamakan dengan nol diperoleh

$$-n + \frac{\sum x_i}{\lambda} = 0. \tag{2.48}$$

Selanjutnya menyelesaikan persamaan (2.48) diperoleh $\hat{\lambda} = \sum x_i/n = \bar{x}$. Nilai $\hat{\lambda}$ ini disebut penduga kemungkinan maksimum (*maximum likelihood estimator*).

2.6.2 Sifat-sifat Sampel Kecil

Pada bagian ini kita akan membicarakan sifat-sifat sampel kecil seperti ketakbiasan (*unbiasedness*), varians minimum, penduga takbias terbaik, kelinearan, penduga takbias linear terbaik, penduga MSE minimum.

Ketakbiasan

Suatu penduga $\hat{\theta}$ dikatakan penduga takbias dari θ jika nilai harapan $\hat{\theta}$ sama dengan nilai sesungguhnya θ. Dengan kata lain

$$\mathrm{E}(\hat{\theta}) = \theta \tag{2.49}$$

atau

$$\mathrm{E}(\hat{\theta}) - \theta = 0. \tag{2.50}$$

Jika persamaan ini tidak dipenuhi, maka penduga dikatakan bias dan dihitung sebagai

$$\mathrm{bias}(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta. \tag{2.51}$$

Varians Minimum

Penduga $\hat{\theta}_1$ dikatakan penduga varians minimum dari θ jika varians $\hat{\theta}_1$ lebih kecil atau sama dengan varians $\hat{\theta}_2$ yang merupakan penduga lain dari θ.

Penduga Takbias Terbaik atau Efisien

Jika $\hat{\theta}_1$ dan $\hat{\theta}_2$ dua penduga takbias dari θ dan varians $\hat{\theta}_1$ lebih kecil atau sama dengan varians $\hat{\theta}_2$, maka $\hat{\theta}_1$ dikatakan penduga takbias varians minimum (*minimum-variance unbiased*) atau takbias terbaik (*best unbiased*) atau efisien.

Kelinearan

Suatu penduga $\hat{\theta}$ dikatakan penduga linear dari θ jika penduga tersebut fungsi linear dari amatan sampel. Sebagai contoh

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.52}$$

adalah penduga linear karena merupakan fungsi linear dari nilai x.

Penduga Takbias Linear Terbaik

Jika $\hat{\theta}$ adalah linear, takbias, dan memiliki varians minimum dari semua kelas estimator takbias linear dari θ maka penduga $\hat{\theta}$ dikatakan penduga takbias linear terbaik (*best linear unbiased estimator* disingkat BLUE).

Penduga MSE Minimum

Mean square error (MSE) suatu penduga $\hat{\theta}$ didefinisikan sebagai

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}(\hat{\theta} - \theta)^2. \tag{2.53}$$

Bandingkan dengan varians yang didefinisikan sebagai

$$\mathrm{var}(\hat{\theta}) = \mathrm{E}[\hat{\theta} - \mathrm{E}(\hat{\theta})]^2. \tag{2.54}$$

Lalu apa beda keduanya? Perbedaannya adalah bahwa $\mathrm{var}(\hat{\theta})$ mengukur penyimpangan atau perbedaan dari distribusi $\hat{\theta}$ di sekitar nilai tengah atau nilai harapannya, $\mathrm{E}(\hat{\theta})$. Sedangkan, MSE mengukur penyimpangan sekitar nilai parameter sesungguhnya. Hubungan antara MSE, bias, dan varians adalah sebagai berikut:

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{var}(\hat{\theta}) + [\mathrm{bias}(\hat{\theta})]^2. \tag{2.55}$$

2.6.3 Sifat-sifat Sampel Besar

Seringkali suatu penduga (estimator) tidak memenuhi satu atau lebih sifat-sifat statistika yang diinginkan pada sampel kecil. Namun, sebagaimana ukuran sampel membesar menuju tak berhingga, penduga memiliki beberapa sifat-sifat statistika yang diinginkan. Sifat-sifat ini disebut sifat sampel besar (*large-sampel*) atau asimtotik (*asymptotic*).

Ketakbiasan asimptotik

Suatu penduga $\hat{\theta}$ dikatakan penduga takbias asimtotik dari θ jika

$$\lim_{n\to\infty} \mathrm{E}(\hat{\theta}_n) = \theta. \tag{2.56}$$

Simbol $\hat{\theta}_n$ berarti bahwa penduga berdasarkan sampel berukuran n.

Contoh 2.6.2. Misalkan varians sampel dari peubah acak X adalah $s^2 = \sum(x_i - \bar{x})^2/n$. Dapat ditunjukkan bahwa

$$\mathrm{E}(s^2) = \sigma^2\left(1 - \frac{1}{n}\right) \tag{2.57}$$

Jelas bahwa pada pada sampel kecil s^2 bersifat bias. Namun sebagaimana n membesar menuju takhingga, $\mathrm{E}(s^2)$ mendekati σ^2 yang sesungguhnya.

Konsistensi

Suatu penduga $\hat{\theta}$ dikatakan penduga konsisten jika nilainya mendekati parameter sesungguhnya θ sebagaimana ukuran sampel membesar dan membesar. Jika dalam limit (sebagaimana n membesar menuju takhingga) distribusi $\hat{\theta}$ menuju titik tunggal θ, dan jika distribusi $\hat{\theta}$ memiliki penyebaran atau varians nol, kita katakan $\hat{\theta}$ adalah penduga konsisten dari θ.

Secara formal, suatu penduga $\hat{\theta}$ dikatakan penduga konsisten dari θ jika peluang bahwa nilai mutlak selisih antara $\hat{\theta}$ dan θ kurang dari δ (suatu bilangan positif kecil sebarang) mendekati satu. Dengan kata lain,

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| < \delta) = 1, \quad \delta > 0. \tag{2.58}$$

Persamaan (2.58) biasanya dituliskan sebagai

$$\mathrm{plim}_{n\to\infty}\hat{\theta} = \theta. \tag{2.59}$$

Sebagai catatan, sifat ketakbiasan dan konsistensi secara konseptual adalah hal yang berbeda. Sifat ketakbiasan berlaku untuk sebarang ukuran sampel, namun konsistensi hanya untuk sifat sampel besar.

Syarat cukup (*sufficient condition*) untuk konsistensi adalah bias dan varians menuju nol sebagaimana ukuran sampel membesar sampai tak berhingga. Dengan kata lain,

$$\lim_{n\to\infty} \mathrm{E}(\hat{\theta}_n) = \theta \tag{2.60}$$

dan

$$\lim_{n\to\infty} \mathrm{var}(\hat{\theta}_n) = 0. \tag{2.61}$$

Alternatif lain, kondisi cukup untuk konsistensi adalah $\mathrm{MSE}(\hat{\theta})$ menuju nol sebagaimana n membesar menuju tak berhingga.

Contoh 2.6.3. Misalkan $x_1, x_2, \ldots, x_n$ adalah sampel acak dari distribusi dengan nilai tengah μ dan varians σ^2. Tunjukkan bahwa rata-rata sampel $\bar{x}$ adalah penduga konsisten dari μ. Kita tahu bahwa $\mathrm{E}(\bar{X}) = \mu$ dan $\mathrm{var}(\bar{X}) = \sigma^2/n$. Karena $\mathrm{E}(\bar{X}) = \mu$ berapapun ukuran sampelnya maka jelas bersifat takbias. Sebagaimana n membesar menuju tak berhingga $\mathrm{var}(\bar{X})$ menuju nol. Jadi jelas bahwa $\bar{x}$ adalah penduga konsisten untuk μ.

Berikut ini adalah sifat-sifat peluang limit.

1. Invarian (sifat Slutsky). Jika $\hat{\theta}$ adalah penduga konsisten dari θ dan jika $h(\hat{\theta})$ adalah sebarang fungsi kontinu dari $\hat{\theta}$, maka

$$\mathrm{plim}_{n\to\infty}h(\hat{\theta}) = h(\theta). \tag{2.62}$$

   Ini berarti bahwa jika $\hat{\theta}$ penduga konsisten dari θ, maka $1/\hat{\theta}$ juga merupakan penduga konsisten dari $1/\theta$, demikian pula $\log(\hat{\theta})$ adalah penduga konsisten dari $\log(\theta)$. Ingat bahwa, sifat ini tidak berlaku untuk operator ekspektasi E, yaitu jika $\hat{\theta}$ adalah penduga takbias dari θ, yakni $\mathrm{E}(\hat{\theta}) = \theta$ tidaklah benar bahwa $1/\hat{\theta}$ adalah juga penduga takbias dari θ, yaitu

$$\mathrm{E}\left(\frac{1}{\hat{\theta}}\right) \neq \frac{1}{\mathrm{E}(\hat{\theta})} \neq \frac{1}{\theta}. \tag{2.63}$$

2. Jika c adalah konstanta, maka $\mathrm{plim}_{n\to\infty} c = c$.
3. Jika $\hat{\theta}_1$ dan $\hat{\theta}_2$ adalah penduga-penduga konsisten, maka
   a) $\mathrm{plim}(\hat{\theta}_1 + \hat{\theta}_2) = \mathrm{plim}\,\hat{\theta}_1 + \mathrm{plim}\,\hat{\theta}_2$,
   b) $\mathrm{plim}(\hat{\theta}_1\hat{\theta}_2) = \mathrm{plim}\,\hat{\theta}_1 \mathrm{plim}\,\hat{\theta}_2$,
   c) $\mathrm{plim}(\hat{\theta}_1/\hat{\theta}_2) = \mathrm{plim}\,\hat{\theta}_1/\mathrm{plim}\,\hat{\theta}_2$.

   Ingat secara umum tidaklah berlaku $\mathrm{E}(\hat{\theta}_1/\hat{\theta}_2) \neq \mathrm{E}(\hat{\theta}_1)/\,\mathrm{E}(\hat{\theta}_2)$. Demikian pula, $\mathrm{E}(\hat{\theta}_1\hat{\theta}_2) \neq \mathrm{E}(\hat{\theta}_1)\,\mathrm{E}(\hat{\theta}_2)$. Namun, jika $\hat{\theta}_1$ dan $\hat{\theta}_2$ berdistribusi saling bebas maka tentu saja $\mathrm{E}(\hat{\theta}_1\hat{\theta}_2) = \mathrm{E}(\hat{\theta}_1)\,\mathrm{E}(\hat{\theta}_2)$

Efisiensi Asimptotik

Misalkan $\hat{\theta}$ adalah penduga dari θ. Varians dari distribusi asimptotik $\hat{\theta}$ disebut variansi asimtotik dari $\hat{\theta}$. Jika $\hat{\theta}$ konsisten dan varians asimtotiknya leibh kecil dibandingkan varians asimptotik dari semua penduga konsisten θ yang lain, maka $\hat{\theta}$ disebut efisien asimptotik (*asymptotically efficient*).

Kenormalan asimptotik

Suatu penduga $\hat{\theta}$ dikatakan berdistribusi normal asimtotik jika distribusi pengambilan sampelnya menuju distribusi normal sebagaimana ukuran sampel n membesar menuju takhingga.

Latihan Soal

1. Misalkan terdapat fungsi densitas peluang bersama

$$f(x,y) = 2 - x - y, \quad 0 \leq x \leq 1; 0 \leq y \leq 1.$$

   Hitunglah:
   a) $f(x)$
   b) $f(y)$
   c) $\mathrm{E}(X)$
   d) $\mathrm{E}(Y)$
   e) $\mathrm{E}(XY)$
   f) $\mathrm{E}(Y|X)$
   g) $\mathrm{E}(X|Y)$
   h) $\mathrm{var}(X)$
   i) $\mathrm{var}(Y)$

j) $\mathrm{cov}(X, Y)$

k) $\mathrm{cor}(X, Y)$

2. Misalkan X dan Y adalah peubah acak, a_1, a_2, b_1, dan b_2 adalah konstanta. Buktikan sifat-sifat berikut:

   a) Jika X dan Y saling bebas maka $\mathrm{cov}(X, Y) = 0$.

   b) $\mathrm{cov}(a_1X + b_1, a_2Y + b_2) = a_1a_2 \,\mathrm{cov}(X, Y)$.

   c) $\mathrm{cor}(a_1X + b_1, a_2Y + b_2) = \pm \,\mathrm{cor}(X, Y)$.

   d) Jika X dan Y saling bebas maka $\mathrm{E}(Y|X) = \mathrm{E}(Y)$.

3. Misalkan Y_1, Y_2, dan Y_3 adalah sampel dari suatu populasi $N(\mu, \sigma^2)$, namun Y_1, Y_2, dan Y_3 tidak saling bebas. Bahkan nilai

$$\mathrm{cov}(Y_1, Y_2) = \mathrm{cov}(Y_2, Y_3) = \mathrm{cov}(Y_1, Y_3) = \frac{\sigma^2}{2}.$$

   Misalkan $\bar{Y} = (Y_1 + Y_2 + Y_3)/3$.

   a) Hitung $\mathrm{E}(\bar{Y})$

   b) Hitung $\mathrm{var}(\bar{Y})$

4. Latihan berikut memberikan contoh bahwa pasangan peubah acak X dan Y dengan nilai tengah bersyarat Y diketahui X bergantung pada X, namun $\mathrm{cor}(X, Y) = 0$. Misalkan X dan Z adalah dua peubah acak normal baku, yaitu $N(0, 1)$, yang saling bebas dan misalkan $Y = X^2 + Z$.

   a) Buktikan bahwa $\mathrm{E}(Y|X) = X$.

   b) Buktikan bahwa $\mathrm{E}(Y) = 1$.

   c) Buktikan bahwa $\mathrm{E}(XY) = 0$. Petunjuk: Gunakan fakta bahwa momen-momen ganjil dari suatu distribusi normal baku semuanya adalah nol.

   d) Buktikan bahwa $\mathrm{cov}(X, Y) = 0$ dan dengan demikian $\mathrm{cor}(X, Y) = 0$.

5. Misalkan Y_1, Y_2, Y_3, dan Y_4 adalah peubah acak bebas dan berdistribusi identik dari suatu populasi dengan nilai tengah μ dan varians σ^2. Misalkan $\bar{Y} = (Y_1 + Y_2 + Y_3 + Y_4)/4$ menyatakan rata-rata dari keempat peubah acak tersebut.

   a) Hitunglah nilai harapan dan varians $\bar{Y}$!

   b) Sekarang misalkan terdapat penduga yang lain untuk μ, yaitu

$$W = \frac{1}{8}Y_1 + \frac{1}{8}Y_2 + \frac{1}{4}Y_3 + \frac{1}{2}Y_4.$$

   Buktikan bahwa W adalah penduga takbias untuk μ. Kemudian hitung varians W.

   c) Berdasarkan jawaban (a) dan (b), penduga untuk μ mana yang akan Anda pilih, $\bar{Y}$ atau W?

6. Misalkan $\bar{Y}$ menyatakan rata-rata sampel dari suatu sampel acak dengan nilai tengah μ dan varians σ^2/n. Misalkan terdapat dua penduga alternatif untuk μ:

$$W_1 = \frac{(n-1)}{n}\bar{Y}$$

dan

$$W_2 = \frac{\bar{Y}}{2}.$$

a) Buktikan bahwa W_1 dan W_2 adalah penduga bias untuk μ dan hitunglah biasnya. Apa yang terjadi pada bias saat $n \to \infty$? Berikan komentar Anda pada perbedaan yang penting dalam bias untuk kedua penduga sebagaimana ukuran sampel membesar.

b) Hitunglah limit peluang (*probability limit*) plim(W_1) dan plim(W_2). Penduga mana yang konsisten?

c) Hitung var(W_1) dan var(W_2).

d) Berikan argumentasi bahwa W_1 merupakan penduga yang lebih baik dibandingkan $\bar{Y}$ jika μ "mendekati" nol. Lengkapi penjelasan Anda dengan membandingkan nilai bias dan varians yang telah dihitung.

7. Misalkan $X_1, X_2, \ldots, X_n$ adalah sampel acak berukuran n dari suatu populasi dengan nilai tengah μ dan varians σ^2. Daripada menggunakan semua n amatan, misalkan suatu penduga yang sangat sederhana yakni hanya menggunakan dua sampel pertama

$$X^* = \frac{X_1 + X_2}{2}$$

a) Buktikan bahwa X^* adalah penduga linear.

b) Buktikan bahwa X^* adalah penduga takbias.

c) Hitunglah varians X^*.

8. Misalkan $X_1, X_2, \ldots, X_n$ adalah sampel acak berukuran n dari suatu populasi dengan $N \sim (\mu, \sigma^2)$. Misalkan terdapat penduga terbobot

$$\tilde{X} = \frac{X_1}{2} + \frac{X_2}{3} + \frac{X_3}{6}$$

a) Buktikan bahwa $\tilde{X}$ adalah penduga linear.

b) Buktikan bahwa $\tilde{X}$ adalah penduga takbias.

c) Hitunglah varians $\tilde{X}$.

2.7 Inferensi Statistika

Pendugaan dan pengujian hipotesis merupakan dua cabang inferensi statistika klasik. Pengujian hipotesis dapat dinyatakan sebagai berikut. Misalkan kita mempunyai peubah acak dengan fungsi densitas peluang $f(x;\theta)$, dengan θ adalah parameter distribusi. Setelah mendapatkan sampel acak berukuran n kita mendapatkan pendugaan titik $\hat{\theta}$. Namun, karena parameter sesungguhnya θ biasanya tidak diketahui, maka muncul pertanyaan: apakah mungkin sampel kita berasal dari fungsi densitas peluang $f(x;\theta) = \theta^*$? Dalam pengujian hipotesis ini disebut dengan hipotesis nol (null) dan biasanya dinotasikan dengan H_0. Hipotesis nol ini akan diuji melawan hipotesis alternatif, yang dinotasikan H_a atau lazimnya ditulis H_1. Hipotesis H_1 ini menyatakan bahwa $\theta \neq \theta^*$.

Hipotesis nol dan hipotesis alternatif dapat berupa hipotesis sederhana (*simple*) atau komposit (*composite*). Hipotesis dikatakan sederhana jika hipotesis tersebut menspesifikasikan nilai parameter distribusi. Jika tidak, hipotesis dikatakan hipotesis komposit.

Contoh 2.7.1. Misalkan $X \sim N(\mu, \sigma^2)$ dan kita bisa nyatakan hipotesis

$$H_0 : \mu = 18 \quad \text{dan} \quad \sigma = 4 \tag{2.64}$$

yang merupakan hipotesis sederhana, atau

$$H_0 : \mu = 18 \quad \text{dan} \quad \sigma > 4 \tag{2.65}$$

yang merupakan hipotesis komposit karena nilai σ tidak ditentukan.

Untuk menguji hipotesis nol, yaitu menguji validitasnya, kita menggunakan informasi sampel untuk memperoleh apa yang disebut sebagai statistik uji. Seringkali stastistik uji ini merupakan penduga titik dari parameter yang tidak diketahui. Kemudian, kita berusaha mencari distribusi pengambilan sampel dari statistik uji tersebut dan menggunakan selang kepercayaan (*confidence interval*) atau uji signifikansi (*test of significance*) untuk menguji hipotesis nol.

Contoh 2.7.2. Misalkan distribusi tinggi dari pria pada suatu populasi berdistribusi normal dengan nilai tengah μ inci dan simpangan baku $\sigma = 2{,}5$ inci. Suatu sampel acak 100 pria diambil dari populasi ini memiliki tinggi rata-rata 67 inci. Dengan kata lain kita diberitahu bahwa $X_i \sim N(\mu, 2{,}5^2)$, $\bar{X} = 67$, dan $n = 100$. Misalkan diasumsikan kita akan menguji hipotesis berikut:

$$H_0 : \mu = \mu^* = 69, \tag{2.66}$$
$$H_1 : \mu \neq 69. \tag{2.67}$$

Pertanyaannya adalah: dapatkan sampel dengan $\bar{X} = 67$, yakni statistik uji, yang berasal dari populasi ini dengan nilai tengah 69? Secara intuitif kita tidak akan menolak hipotesis nol jika $\bar{X}$ "cukup dekat" dengan μ. Jika tidak demikian, kita mungkin akan menolaknya dan memilih hipotesis alternatif. Namun, bagaimana cara kita memutuskan bahwa $\bar{X}$ cukup dekat dengan μ? Kita dapat menggunakan dua pendekatan: (1) selang kepercayaan dan (2) uji signifikansi.

2.7.1 Pendekatan Selang Kepercayaan

Kita mengetahui bahwa $X_i \sim N(\mu, \sigma^2)$ dan kita juga tahu statistik uji $\bar{X} \sim N(\mu, \sigma^2/n)$. Karena kita tahu distribusi peluang $\bar{X}$, kita dapat membuat selang kepercayaan $100(1-\alpha)$ untuk μ berdasarkan nilai $\bar{X}$ dan memeriksa apakah selang ini menyertakan nilai $\mu = \mu^* = 69$. Jika memang demikian, kita tidak akan menolak hipotesis nol; sebaliknya, jika tidak kita akan menolak hipotesis nol. Dengan demikian, jika $\alpha = 0{,}05$, kita akan mempunyai selang kepercayaan 95% dan selang ini menyertakan μ^*, kita mungkin tidak akan menolak hipotesis nol. Artinya, 95 dari 100 interval akan menyertakan μ^*. Lalu, bagaimana cara kerjanya? Kita tahu bahwa $\bar{X} \sim N(\mu, \sigma^2/n)$, maka

$$Z_i = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \tag{2.68}$$

yaitu peubah acak normal baku. Dengan tabel normal, kita tahu bahwa

$$\Pr(-1{,}96 \leq Z_i \leq 1{,}96) = 0{,}95. \tag{2.69}$$

Dengan kata lain

$$\Pr\left(-1{,}96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1{,}96\right) = 0{,}95 \tag{2.70}$$

atau, dengan pengaturan ulang masing-masing suku kita akan memperoleh,

$$\Pr\left[\bar{X} - 1{,}96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1{,}96\frac{\sigma}{\sqrt{n}}\right] = 0{,}95. \tag{2.71}$$

Ini adalah selang kepercayaan 95% untuk μ. Begitu selang ini kita peroleh, langkah selanjutnya adalah memerika apakah $\mu = \mu^*$ terletak di dalam selang ini. Jika memang demikian, kita tidak akan menolak hipotesis nol; jika tidak, kita akan menolaknya.

Contoh 2.7.3. Lihat kembali Contoh 2.7.2. Kita peroleh selang kepercayaan 95% untuk μ adalah

$$66{,}51 \leq \mu \leq 67{,}49. \tag{2.72}$$

Selang ini jelas tidak mengikutkan $\mu = 69$. Dengan demikian, kita dapat menolak hipotesis nol bahwa nilai μ sesungguhnya adalah 69 dengan koefisien kepercayaan 95%. Dalam istilah uji hipotesis, selang kepercayaan yang telah kita bentuk disebut daerah penerimaan (*acceptance region*) dan area di luar daerah penerimaan disebut daerah kritis (*critical region*) atau daerah penolakan (*region(s) of rejection*) hipotesis nol. Batas bawah dan batas atas dari daerah penerimaan disebut nilai kritis (*critical values*). Jika nilai yang dihipotesiskan berada di dalam daerah penerimaan, maka kita tidak akan menolak hipotesis nol; demikian pula sebaliknya.

Dalam memutuskan untuk menerima atau menolak hipotesis nol H_0, kita akan melakukan dua jenis kesalahan. Pertama, kita menolak H_0, padahal H_0 benar. Kesalahan ini disebut kesalahan tipe I (*type I eror*). Dengan demikian, pada contoh sebelumnya $\bar{X}$ mungkin berasal dari populasi dengan nilai tengah 69. Kesalahan kedua adalah kita tidak menolak H_0 padahal H_0 salah. Kesalahan ini disebut kesalahan tipe II (*type II error*).

Idealnya, kita meminimumkan kesalahan tipe I dan II. Namun, sayangnya untuk sebarang ukuran sampel, tidaklah memungkinkan untuk meminimumkan kedua kesalahan secara simultan. Pendekatan klasik untuk masalah ini adalah sebagai berikut. Dianggap kesalahan tipe I lebih serius dibandingkan kesalahan tipe II. Dengan demikian, kita seharusnya berusaha untuk menjaga agar melakukan kesalahan tipe I pada tingkat yang cukup kecil misalnya 0,01 atau 0,05 dan berusaha untuk meminimumkan peluang memiliki kesalahan tipe II.

Peluang melakukan kesalahan tipe I disebut α dan disebut tingkat signifikansi (*level of significance*) dan peluang melakukan kesalahan tipe II disebut β. Peluang tidak melakukan kesalahan tipe II disebut kekuatan dari uji (*power of the test*). Dengan kata lain, kekuatan uji adalah kemampuan untuk menolak hipotesis nol yang salah. Pendekatan klasik adalah menguji hipotesis pada α tertentu seperti 0,01 atau 0,05 dan memaksimumkan kekuatan, dengan kata lain meminumkan β.

Dengan demikian, kita sekarang tahu bahwa $(1-\alpha)$ adalah satu dikurangi peluang melakukan kesalahan tipe I. Jadi koefisien kepercayaan 95% berarti bahwa kita menerima bahwa paling banyak 5% peluang melakukan kesalahan tipe I–kita tidak ingin menolak hipotesis sebenarnya lebih dari 5 dari 100 kali.

2.7.2 Nilai p atau Tingkat Signifikansi Eksak

Pada bagian sebelumnya kita dapat memilih nilai α pada sebarang tingkat, misalnya 1, 5, atau 10 persen. Daripada memilih nilai α sebarang, kita dapat memperoleh nilai peluang (*probability value*, biasanya disingkat p-value) atau tingkat signifikansi eksak. Nilai p atau p-value didefinisikan sebagai tingkat signifikansi terkecil yang mana hipotesis nol dapat ditolak. Perangkat lunak statistika biasanya menghitung p-value ini secara otomatis.

Contoh 2.7.4. Dalam suatu pengujian hipotesis diperoleh p-value $= 0{,}001$. Ini berarti pada $\alpha = 0{,}01$ dan $\alpha = 0{,}05$ kita akan menolak hipotesis nol. Sebaliknya pada pengujian hipotesis dengan p-value $= 0{,}065$ maka pada $\alpha = 0{,}01$ dan $\alpha = 0{,}05$ kita akan menerima hipotesis nol.

2.7.3 Pendekatan Uji Signifikansi

Ingat bahwa

$$Z_i = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1). \tag{2.73}$$

Biasanya nilai $\bar{X}$ dan n diketahui atau dapat diestimasi, tetapi nilai μ dan σ sesungguhnya tidak diketahui. Namun, jika σ ditentukan dan diasumsikan dalam H_0 bahwa $\mu = \mu^*$, suatu nilai tertentu, maka Z_i dapat dihitung secara langsung menggunakan tabel normal atau perangkat lunak statistika. Jika nilai peluang ini kecil, katakanlah kurang dari 1 persen atau 5 persen, maka kita dapat menolak hipotesis nol.

Contoh 2.7.5. Lihat kembali Contoh 2.7.2. Kita akan melakukan prosedur yang disebut sebagai uji Z. Pada contoh kita, $\mu = \mu^* = 69$, maka diperoleh

$$Z = \frac{67 - 69}{2{,}5/\sqrt{100}} = \frac{-2}{0{,}25} = -8. \tag{2.74}$$

Jika kita lihat tabel distribusi normal kita tahu bahwa peluang Z melebihi 3 atau -3 adalah sekitar 0,001. Peluang melebihi 8 ini bahkan lebih kecil. Dengan demikian kita dapat menolak hipotesis nol.

2.7.4 Langkah-langkah Pengujian Hipotesis

Secara garis besar langkah-langkah pengujian hipotesis adalah sebagai berikut:

1. Nyatakan hipotesis nol H_0 dan hipotesis alternatif H_1. Sebagai contoh $H_0 : \mu = 69$ dan $H_1 : \mu \neq 69$.

2. Pilih statistik uji. Sebagai contoh $\bar{X}$.

3. Tentukan distribusi peluang dari statistik uji, misalnya $\bar{X} \sim N(\mu, \sigma^2/n)$.

4. Pilih tingkat signifikansi, yakni peluang melakukan kesalahan tipe I, α.

5. Hitunglah selang kepercayaan $100(1 - \alpha)\%$ menggunakan distribusi peluang statistik uji. Jika nilai parameter dalam hipotesis nol berada dalam daerah kepercyaan ini, jangan menolak hipotesis nol. Sebaliknya jika berada di luar dareah ini, kita menolak hipotesis nol. Ingat bahwa, menerima atau menolak hipotesis nol artinya kita mengambil risiko salah sebesar α persen.

2.8 Aljabar Vektor dan Matriks untuk Ekonometrika

Berikut ini diberikan beberapa penerapan aljabar vektor dan matriks yang relevan untuk ekonometrika. Bagian ini diadaptasi dari Judge *et al.* (1982), Gujarati (2003), dan Stock and Watson (2012).

2.8.1 Beberapa sifat penting operasi vektor dan matriks

Transpos matriks. Misalkan $\mathbf{A}$ dan $\mathbf{B}$ adalah dua matriks maka $(\mathbf{A}+\mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$ dan $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$.

Balikan matriks. Misalkan matriks $\mathbf{A}$ dan $\mathbf{B}$ adalah terbalikkan (*invertible*), maka $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

Matriks definit positif dan semidefinit positif. Misalkan $\mathbf{V}$ adalah matriks bujur sangkar. Matriks $\mathbf{V}$ dikatakan definit positif jika $\mathbf{c}^\top\mathbf{V}\mathbf{c} > 0$ untuk semua vektor taknol $\mathbf{c}$ yang berdimensi $n \times 1$. Dengan cara serupa $\mathbf{V}$ dikatakan semidefinit positif jika $\mathbf{c}^\top\mathbf{V}\mathbf{c} \geq 0$ untuk semua vektor taknol $\mathbf{c}$ yang berdimensi $n \times 1$.

Matriks idempoten. Suatu matriks $\mathbf{A}$ dikatakan idempoten jika $\mathbf{A}$ adalah matriks bujur sangkar dan $\mathbf{AA} = \mathbf{A}$. Jika $\mathbf{A}$ adalah suatu matriks idempoten yang juga simetrik, maka $\mathbf{A}$ adalah semidefinit positif dan A memiliki r nilai eigen yang sama dengan 1 dan $n-r$ nilai eigen yang sama dengan 0, dengan $r = \text{rank}(\mathbf{A})$.

2.8.2 Turunan vektor dan matriks

Misalkan $f(\mathbf{B}, \mathbf{x})$ adalah fungsi bernilai real dan terdeferensialkan dari matriks $\mathbf{B}$ yang berdimensi $(T \times K)$, vektor kolom $\mathbf{x}$ yang berdimensi $T \times 1$ dan $\mathbf{b}(\mathbf{x})$ adalah fungsi terdeferensialkan dari $\mathbf{x}$ dengan nilai dalam ruang Euclid berdimensi K. Selanjutnya didefinisikan

1. $\partial f/\partial\mathbf{x}$ adalah vektor kolom berdimensi T dengan elemen ke-i $\partial f/\partial x_i$,
2. $\partial f/\partial\mathbf{B}$ adalah matriks berdimensi $T \times K$ dengan elemen ke-ij $\partial f/\partial b_{ij}$,
3. $\partial\mathbf{b}/\partial\mathbf{x}^\top$ adalah matriks berdimensi $K \times T$ dengan elemen ke-ij $\partial b_i/\partial x_j$.

Misalkan $\mathbf{a}$ dan $\mathbf{x}$ adalah vektor-vektor berdimensi $T \times 1$, $\mathbf{y}$ adalah vektor berdimensi $K \times 1$, $\mathbf{A}$ adalah matriks berdimensi $T \times T$, dan $\mathbf{B}$ adalah matriks berdimensi $T \times K$. Selanjutnya

1. $\dfrac{\partial\mathbf{a}^\top\mathbf{x}}{\partial\mathbf{x}} = \mathbf{a}$,
2. $\dfrac{\partial\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\partial\mathbf{x}} = (\mathbf{A}+\mathbf{A}^\top)\mathbf{x}$,
3. $\dfrac{\mathbf{x}^\top\mathbf{B}\mathbf{y}}{\partial\mathbf{B}} = \mathbf{x}\mathbf{y}^\top$.

2.9 Elastisitas

Alat favorit seorang ekonomiwan adalah elastisitas. Elastisitas merupakan persentase perubahan pada satu peubah yang berasosiasi dengan 1% perubahan pada peubah lain untuk pergerakan pada kurva tertentu. Elastisitas y yang bersesuaian dengan perubahan pada x, dinotasikan ε_{yx}, didefinisikan sebagai

$$\varepsilon_{yx} = \text{lereng} \times \frac{x}{y}.$$

Lereng (*slope*) diperoleh dari turunan $\mathrm{d}y/\mathrm{d}x$. Sebagai contoh untuk hubungan linear $y = \beta_0 + \beta_1 x$, lereng untuk model ini adalah $\mathrm{d}y/\mathrm{d}x = \beta_1$, dan elastistasnya adalah $\varepsilon_{yx} = \beta_1 x/y$.

BAB III

REGRESI LINEAR SEDERHANA

3.1 Latar Belakang

Sebelum membicarakan apa itu analisis regresi linear sederhana perhatikan terlebih dahulu contoh berikut yang diadaptasi dari Gujarati (2003). Contoh berikut adalah populasi total dari 60 keluarga dengan pendapatan mingguan X dan pengeluaran untuk konsumsi mingguan Y, keduanya dalam satuan dolar (lihat Tabel 3.1.1). Enam puluh keluarga tersebut dibagi menjadi 10 kelompok pendapatan (dari \$80 sampai dengan \$260) dan pengeluaran mingguan untuk masing-masing keluarga pada tiap kelompok yang berbeda ditunjukkan di dalam Tabel 3.1.1. Dengan demikian kita melihat bahwa terdapat 10 nilai X yang tetap (*fixed*) yang bersesuaian dengan nilai-nilai Y melawan masing-masing nilai X. Artinya, terdapat 10 X subpopulasi. Tabel 3.1.1 memberikan informasi tentang nilai tengah atau rata-rata pengeluaran konsumsi mingguan yang bersesuaian dengan 10 tingkat pendapatan. Sebagai contoh rata-rata konsumsi untuk keluarga yang berpenghasilan mingguan \$100 adalah \$77, sedangkan untuk keluarga dengan penghasilan mingguan \$260 adalah \$173. Jadi kita akan mendapatkan 10 nilai tengah untuk 10 nilai subpopulasi Y. Nilai ini disebut nilai harapan bersyarat (*conditional expected values*) karena nilai ini tergantung pada nilai peubah Y yang diberikan, dan kita tulis ini sebagai $\mathrm{E}(Y|X)$.

Hal penting yang perlu diperhatikan adalah perbedaan antara nilai harapan bersyarat dan nilai harapan tidak bersyarat (*unconditional expected value*) dari pengeluaran untuk keenam puluh keluarga, yang dinotasikan dengan $\mathrm{E}(Y)$. Nilai ini tentu saja bisa diperoleh dengan menjumlahkan pengeluaran konsumsi mingguan untuk keenam puluh keluarga dalam populasi dan membaginya dengan 60 yaitu \$ 7272/60 = \$121,20. Nilai ini diperoleh dengan mengabaikan perbedaan tingkat pendapatan dalam keluarga-keluarga yang berbeda-beda. Kalau kita lihat kembali Tabel 3.1.1 jelas bahwa tiap-tiap nilai harapan bersyarat berbeda dengan nilai harapan tak bersyarat \$121,20.

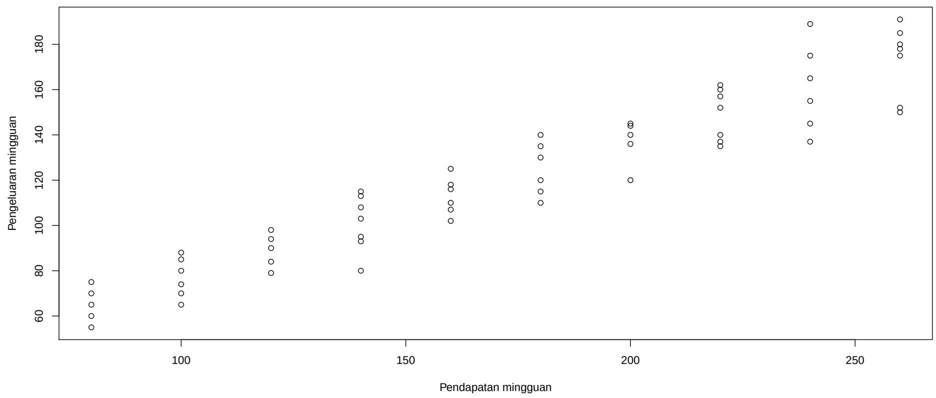
Apa perbedaan kedua nilai tengah tersebut? Jawaban sudah disebutkan di atas. Namun, kita bisa pertegas lagi. Nilai tengah tak bersyarat menjawab pertanyaan ”Berapakah nilai harapan dari pengeluaran konsumsi mingguan dari sebuah keluarga?” Namun jika kita bertanya ”Berapakah nilai harapan dari pengeluaran konsumsi mingguan dari suatu keluarga dengan pendapatan mingguan \$180?” maka kita akan mendapatkan jawaban \$125. Dengan kata lain, apabila kita bertanya ”Berapakah nilai prediksi nilai tengah terbaik untuk pengeluaran mingguan dari suatu keluarga dengan pendapatan mingguan \$180?”, kita akan mendapatkan jawaban \$125. Lalu apa artinya? Pengetahuan tentang tingkat pendapatan memungkinkan kita untuk memprediksi nilai tengah pengeluaran konsumsi dengan lebih baik jika dibandingkan kita tidak memiliki informasi tersebut.

Apabila kita plot data X dan Y dan masing-masing nilai harapan bersyarat dihubungkan, maka kita akan memperoleh garis regresi populasi (*population regression line*) atau kurva regresi populasi (*population regression curve*) atau regresi Y pada X (*regression of Y on X*).

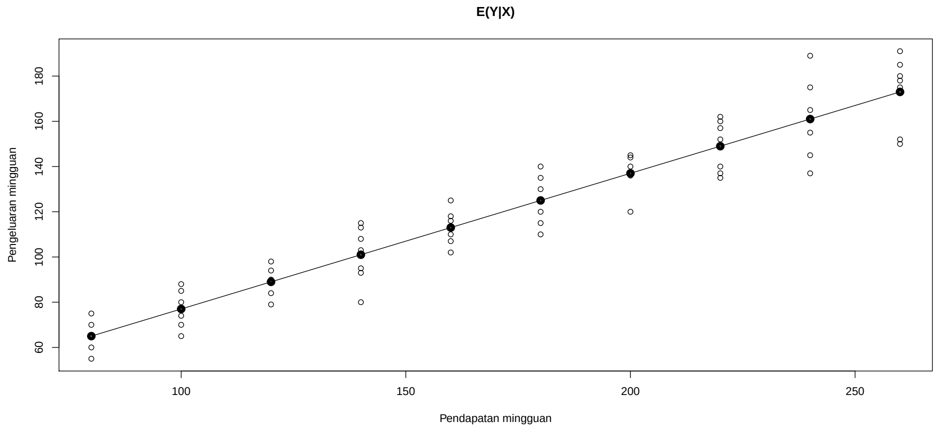
Apabila kita plot data pada Tabel 3.1.1 kita akan memperoleh Gambar 3.1.1. Kemudian apabila kita hubungkan titik-titik pada sepuluh tingkat pengeluaran dengan nilai $\mathrm{E}(Y|X)$, kita akan memperoleh garis regresi populasi seperti pada Gambar 3.1.2.

Tabel 3.1.1: Pendapatan mingguan keluarga

	X									
	80	100	120	140	160	180	200	220	240	260
	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
Y	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
		88		113	125	140		160	189	185
				115				162		191
Total	325	462	445	707	678	750	685	1043	966	1211
$\mathrm{E}(y\|x)$	65	77	89	101	113	125	137	149	161	173



Gambar 3.1.1: Plot pendapatan mingguan versus pengeluaran mingguan.



Gambar 3.1.2: Garis regresi populasi pada data plot pendapatan mingguan versus pengeluaran mingguan.

3.2 Konsep Fungsi Regresi Populasi

Pada bagian sebelumnya kita telah melihat bahwa $\mathrm{E}(Y|X_i)$ adalah fungsi dari X_i, dengan X_i adalah nilai-nilai dari X. Ini kita bisa nyatakan secara simbolis sebagai

$$\mathrm{E}(Y|X_i) = f(X_i) \tag{3.1}$$

dengan $f(X_i)$ adalah fungsi dari peubah penjelas dari peubah acak X. Pada contoh kita sebelumnya $\mathrm{E}(Y|X_i)$ adalah fungsi linear dari X_i. Persamaan (3.1) disebut fungsi harapan bersyarat (*conditional expectation function*) atau fungsi regresi populasi (*population regression function*) atau regresi populasi (*population regression*). Secara singkat, fungsi ini berarti nilai harapan distribusi bersyarat Y diketahui X_i secara fungsional berhubungan dengan X_i. Dengan istilah yang lebih sederhana, dapat dikatakan bahwa fungsi ini memberikan informasi bagaimana nilai tengah atau rata-rata peubah respon Y bervariasi dengan X.

Kemudian muncul pertanyaan bagaimana bentuk fungsi $f(X_i)$ ini? Dalam kondisi nyata kita tidak memiliki keseluruhan populasi untuk diuji. Dengan demikian bentuk fungsional fungsi regresi populasi adalah pertanyaan empiris. Biasanya sebagai langkah awal kita mengasumsikan bentuk fungsi regresi populasi adalah fungsi linear dari X, katakanlah

$$\mathrm{E}(Y|X_i) = \beta_0 + \beta_1 X_i \tag{3.2}$$

dengan β_0 dan β_1 adalah parameter yang tidak diketahui, namun nilainya tetap yang disebut dengan koefisien regresi; β_0 disebut juga koefisien intersep dan β_1 disebut koefisien lereng (*slope*). Persamaan (3.2) disebut fungsi regresi populasi, model regresi populasi linear, atau regresi populasi linear. Istilah regresi, persamaan regresi, dan model regresi adalah sinonim. Tujuan analisis regresi adalah mengestimasi β_0 dan β_1.

3.2.1 Makna Istilah Linear

Kita telah membicarakan konsep dari regresi linear sederhana. Tentu saja muncul pertanyaan tentang maksud kata linear. Secara garis besar kata linear dapat diinterpretasikan dalam dua cara yang berbeda. Pertama, linear dalam peubah. Dalam interpretasi ini nilai harapan bersyarat $\mathrm{E}(Y|X)$ adalah fungsi linear dari X. Dalam interpretasi ini fungsi regresi seperti

$$\mathrm{E}(Y|X) = \beta_0 + \beta_1 X_i^2 \tag{3.3}$$

bukanlah fungsi linear karena peubah X muncul sebagai pangkat. Demikian pula, dengan fungsi regresi

$$\mathrm{E}(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2. \tag{3.4}$$

Interpretasi kedua kelinearan adalah linear dalam parameter, artinaya $\mathrm{E}(Y|X)$ adalah fungsi linear dari parameter. Hal ini berarti, peubah X bisa linear atau tidak. Dalam interpretasi ini fungsi regresi pada (3.3) dan (3.4) adalah linear dalam parameter regresi β. Tentu saja, pada interpretasi kedua ini model dengan bentuk

$$\mathrm{E}(Y|X) = \beta_0 + \beta_1^2 X_i \tag{3.5}$$

tidaklah linear dalam parameter (atau nonlinear). Lalu interpretasi mana yang kita gunakan? Pembicaraan tentang analisis regresi linear yang dimaksud dengan linear adalah linear dalam parameter.

3.2.2 Spesifikasi stokastik fungsi regresi populasi

Kita telah melihat bagaimana konsep fungsi regresi populasi diterapkan pada data pendapatan dan pengeluaran mingguan. Pada contoh kasus tersebut terlihat bahwa konsumsi meningkat seiring dengan pendapatan meningkat. Namun, terlihat pula bahwa pengeluaran masing-masing keluarga tidaklah selalu meningkat sebagaimana pendapatan meningkat. Sebgai contoh, untuk keluarga yang berpenghasilan \$100 terdapat satu keluarga yang pengeluaran konsumsinya \$65 yang lebih kecil daripada pengeluaran dua keluarga dengan pendapatan hanya \$80. Apa yang bisa kita amati? Melihat kembali Tabel 3.1.1 kita lihat bahwa pengeluaran terkelompok pada rata-rata konsumsi semua keluarga pada X_i yang bersangkutan, atau sekitar nilai harapan bersyaratnya. Kita bisa menyatakan penyimpangan masing-masing individu Y_i disekitar nilai harapannya sebagai berikut

$$u_i = Y_i - \mathrm{E}(Y|X_i) \tag{3.6}$$

atau

$$Y_i = \mathrm{E}(Y|X_i) + u_i \tag{3.7}$$

dengan simpangan u_i adalah peubah acak yang tidak teramati yang bernilai positif atau negatif dan disebut sebagai gangguan stokastik (*stochastic disturbance*) atau suku galat stokastik (*stochastic error term*). Kemudian, $\mathrm{E}(Y|X_i)$ adalah komponen sistematik atau deterministik. Komponen U_i adalah surrogate atau *proxy* dari semua peubah yang diabaikan atau dibuang yang memengaruhi Y namun tidak disertakan dalam model regresi. Sebagai contoh apabila $\mathrm{E}(Y|X_i)$ diasumsikan linear dalam X_i maka kita bisa nyatakan

$$Y_i = \mathrm{E}(Y|X_i) + u_i = \beta_0 + \beta_1 X_i + u_i. \tag{3.8}$$

Kalau kita lihat kembali data pendapatan dan pengeluaran konsumsi di atas, misalkan sekarang kita ambil pengeluaran konsumsi untuk individu dengan penghasilan $X_i = \$120$ (lihat kembali Tabel 3.1.1) maka kita akan peroleh

$$Y_1 = 79 = \beta_0 + \beta_1 \times 120 + u_1, \tag{3.9}$$
$$Y_2 = 84 = \beta_0 + \beta_1 \times 120 + u_2, \tag{3.10}$$
$$\vdots \tag{3.11}$$
$$Y_5 = 98 = \beta_0 + \beta_1 \times 120 + u_5. \tag{3.12}$$

Sekarang, jika ambil nilai harapan pada kedua sisi pada (3.7) kita akan memperoleh

$$\mathrm{E}(Y_i|X_i) = \mathrm{E}(\mathrm{E}(Y|X_i)) + \mathrm{E}(u_i|X_i) \tag{3.13}$$
$$= \mathrm{E}(Y|X_i) + \mathrm{E}(u_i|X_i). \tag{3.14}$$

Namun, mengingat $\mathrm{E}(Y_i|X_i)$ sama dengan $\mathrm{E}(Y|X_i)$ maka berarti $\mathrm{E}(u_i|X_i) = 0$.

3.2.3 Signifikansi suku galat stokastik

Pada bagian sebelumnya telah dijelaskan bahwa suku galat (*disturbance term*) u_i adalah pengganti (*surrogate*) untuk semua peubah yang dihilangkan dari model namun secara bersama-sama memengaruhi Y. Tentu saja muncul pertanyaan: ”Kenapa tidak memasukkan semua peubah ini ke dalam model secara tersurat” atau ”Kenapa tidak mengembangkan model regresi dengan peubah sebanyak mungkin?”. Ada beberapa alasan kenapa hal itu tidak dilakukan (Gujarati, 2003). Berikut alasan-alasan tersebut.

1. **Ketidakjelasan atau kekaburan teori (*vagueness of theory*)**. Dalam hal ini teori, jika ada, yang menentukan tingkah laku Y mungkin tidak lengkap. Artinya, kita mungkin tahu bahwa, misalnya pengeluaran migguan X memengaruhi pengeluaran mingguan Y, namun kita tidak yakin tentang peubah lain yang memengaruhi Y. Dengan demikian u_i dapat digunakan sebagai pengganti untuk semua peubah yang dikeluarkan atau dihilangkan dari model.

2. **Ketidaktersediaan data.** Bahkan jika kita tahu beberapa peubah yang dikeluarkan dari model sehingga kita akan membuat model regresi linear berganda, bukan regresi linear sederhana, kita mungkin tidak memperoleh informasi kuantitatif tentang peubah ini. Sebagai contoh data tentang harta keluarga sebagai salah satu peubah bisa digunakan. Namun, data ini bisa saja tidak tersedia sehingga kita harus menghilangkannya.

3. **Peubah inti versus peubah periferal (*peripheral variables*).** Selain pendapatan X_1, banyak anak X_2, jenis kelamin X_3, agama X_4, pendidikan X_4, dan daerah geografis X_5 juga memengahui pengeluaran konsumsi. Namun, cukup mungkin bahwa pengaruh bersama dari semua atau beberapa peubah ini kemungkinan kecil sekali atau nonsistematik atau acak. Artinya, memasukkan peubah-peubah ini bisa jadi tidak terlalu bermanfaat. Kesulitan lainnya adalah peubah seperti jenis kelamin, pendidikan, dan agama sulit dikuantifikasi.

4. **Keacakan intrinsik dalam tingkah laku manusia**. Bahkan jika telah memasukkan semua peubah ke dalam model, terdapat beberapa keacakan intrinsik dalam masing-masing individu Y yang tidak dapat dijelaskan bagaimanapun kita mencoba.

5. **Peubah proksi/kuasa yang tidak bagus.** Meskipun model regresi klasik mengasumsikan bahwa peubah X dan Y diukur secara akurat, namun dalam praktiknya data bisa "terganggu" akibat kesalahan pengukuran. Sebagai contoh teori konsumsi Milton Friedman mengatakan bahwa konsumsi permanen $Y^{(p)}$ sebagai fungsi pendapatan permanen $X^{(p)}$. Namun, dalam praktiknya peubah ini tidak bisa diamati secara langsung, sehingga digunakan peubah proksi seperti konsumsi saat ini Y dan pendapatan saat ini X. Kedua peubah ini tentu bisa diamati. Namun, terjadi masalah kesalahan pengukuran karena nilai teramati belum tentu sama dengan $Y^{(p)}$ dan $X^{(p)}$. Jadi dalam hal ini, u_i dapat mewaliki galat pengukuran ini.

6. **Prinsip kehematan (*parsimony*)**. Prinsip berarti apabila kita bisa menjelaskan tingkah laku Y dengan dua atau tiga peubah penjelas X dan jika teori kita tidak cukup kuat untuk mengatakan bahwa peubah lain mungkin bisa dimasukkan, kenapa harus memasukkan peubah lain? Nah, kalau u_i menyatakan peubah lainnya tentu saja kita tidak bisa mengeluarkan peubah yang relevan dan penting hanya untuk menjaga agar model regresi tetap sederhana.

7. **Bentuk fungsional yang salah.** Bahkan jika kita secara teoretis memperoleh peubah yang menjelaskan suatu fenomen atau bahkan jika kita bisa mendapatkan data pada peubah-peubah tersebu, sering kali kita tidak mengetahui bentuk fungsional antara peubah bebas dan takbebas. Sebagai contoh andaikan model pendapatan dan pengeluaran adalah

   $$Y = \beta_0 + \beta_1 X_i + u_i. \tag{3.15}$$

   Apakah model ini model yang cocok untuk menjelaskan hubungan antara Y dan X? Bagaimana jika model kedua dengan bentuk

   $$Y = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i. \tag{3.16}$$

Tabel 3.3.2: Suatu sampel acak dari Tabel 3.1.1

Y	X
70	80
65	100
90	120
95	140
110	160
115	180
120	200
140	220
155	240
150	260

Tabel 3.3.3: Suatu sampel acak dari Tabel 3.1.1

Y	X
55	80
88	100
90	120
80	140
118	160
120	180
145	200
135	220
145	240
175	260

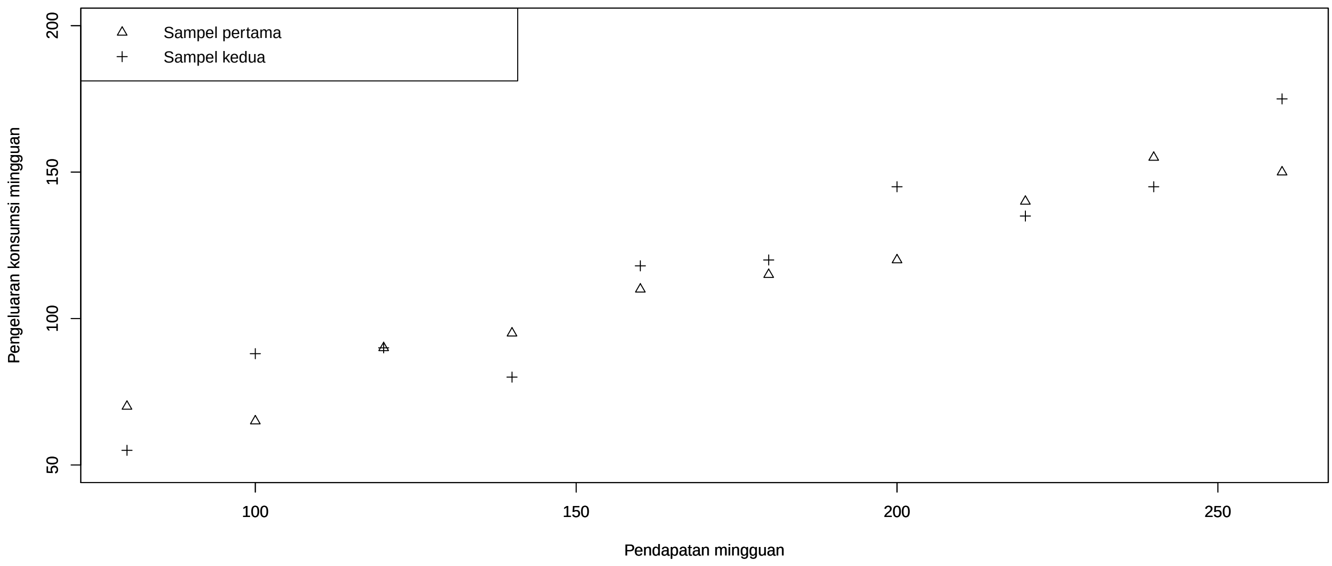
lebih tepat? Pada model dengan dua peubah, yakni Y dan X, plot pencar (*scatter plot*) dapat digunakan untuk melihat bentuk hubungan antara kedua peubah. Tetapi dalam model linear berganda, tidaklah mudah untuk menentukan bentuk fungsional yang tepat karena kita tidak bisa secara grafis memvisualisaikan peubah-peubah ini dalam banyak dimensi.

3.3 Fungsi Regresi Sampel

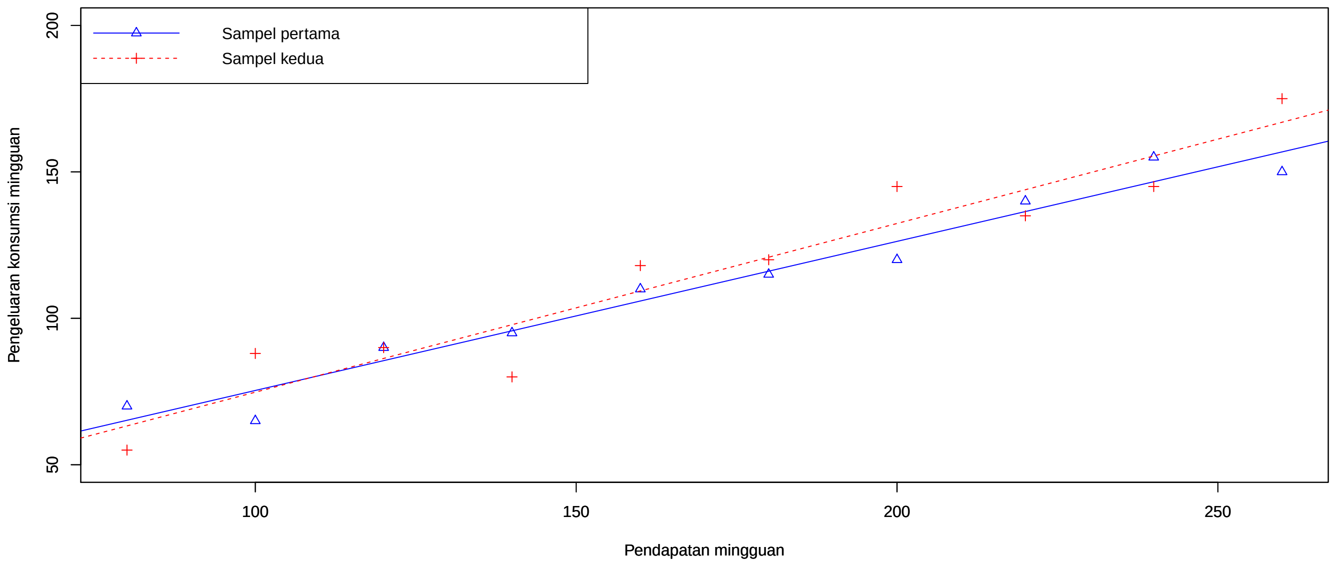
Pada praktiknya kita menggunakan sampel untuk mengestimasi fungsi regresi populasi. Sebagai contoh Tabel 3.3.2 adalah sampel acak dari populasi pada Tabel 3.1.1. Kemudian, Tabel 3.3.3 adalah sampel acak lain dari populasi pada Tabel 3.1.1. Sampel pada Tabel 3.3.2 dan 3.3.3 kemudian diplot. Hasil plot ini dapat dilihat pada Gambar 3.3.3. Kemudian masing-masing garis regresi sampel diplot pada Gambar 3.3.4. Bagaimana cara menghitung garis regresi ini akan dibicarakan pada subbab berikutnya. Secara umum, kita akan memperoleh n persamaan regresi sampel yang berbeda untuk n sampel yang berbeda. Ingat bahwa, tidak ada cara yang pasti untuk mengetahui fungsi regresi populasi karena memang kita tidak pernah tahu. Kalau dilihat lagi Gambar 3.3.4, kita tidak dapat dengan pasti menentukan mana yang menyatakan persamaan regresi populasi karena adanya fluktuasi dalam pengambilan sampel.

Bagian sebelumnya kita telah membahas fungsi regresi populasi berbentuk

$$Y_i = \beta_0 + \beta_1 X_i + u_i. \tag{3.17}$$



Gambar 3.3.3: Plot kedua sampel pengeluaran versus pendapatan dari Tabel 3.3.2 dan Tabel 3.3.3.



Gambar 3.3.4: Plot kedua sampel pengeluaran versus pendapatan dari Tabel 3.3.2 dan Tabel 3.3.3 disertai garis regresi.

Bentuk fungsi regresi populasi pada (3.17) pada praktiknya diestimasi dengan fungsi regresi sampel dengan bentuk

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i \tag{3.18}$$

dalam hal ini $\hat{u}_i$ adalah sisaan atau residual.

3.4 Metode kuadrat terkecil

Sebelum membicarakan metode kuadrat terkecil perhatikan masalah optimisasi berikut. Misalkan kita ingin mencari penduga m yang meminimumkan

$$\sum_{i=1}^{n}(Y_i - m)^2 \tag{3.19}$$

yang mengukur jarak antara penduga m dan titik-titik sampel. Bagaimana cara menghitung m yang meminimumkan ini? Tentu saja dengan mengambil turunan dan menyamakan dengan nol maka kita akan memperoleh

$$\frac{\mathrm{d}}{\mathrm{d}m}\sum_{i=1}^{n}(Y_i - m)^2 = -2\sum_{i=1}^{n}(Y_i - m) = -2\sum_{i=1}^{n} Y_i + 2nm = 0. \tag{3.20}$$

Menyelesaikan persamaan ini dalam m menghasilkan $m = \bar{Y}$. Ingat $m = \bar{Y}$ ini adalah penduga dari $\mathrm{E}(Y)$. Dengan konsep yang sama kita bisa menghitung pendugaan kuadrat terkecil untuk model regresi linear sederhana dengan bentuk

$$Y_i = \beta_0 + \beta_1 X_i + u_i. \tag{3.21}$$

Ingat bahwa garis regresi berdasarkan estimasi ini adalah $\hat{\beta}_0 + \hat{\beta}_1 X_i$. Dengan demikian penduga kuadrat terkecil akan meminimumkan

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X)^2. \tag{3.22}$$

Selanjutnya menurunkan 3.22 terhadap β_0 dan β_1 diperoleh

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) \tag{3.23}$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)X_i \tag{3.24}$$

Kemudian menyamakan dengan nol akan diperoleh persamaan kuadrat terkecil

$$\bar{Y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{X} = 0, \tag{3.25}$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \hat{\beta}_0 \bar{X} - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} X_i^2 = 0. \tag{3.26}$$

Menyelesaikan persamaan di atas untuk $\hat{\beta}_0$ dan $\hat{\beta}_1$ menghasilkan

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \tag{3.27}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \tag{3.28}$$

Pada bagian sebelumnya telah dihitung nilai penduga kuadrat terkecil (*ordinary least squares*) untuk β_0 dan β_1. Untuk sebarang $\hat{\beta}_0$ dan $\hat{\beta}_1$ didefinisikan nilai suaian (*fitted*) value untuk Y sebagai

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i. \tag{3.29}$$

Kemudian sisaan (*residual*) untuk amatan ke-i didefinisikan sebagai selisih nilai aktual Y_i dan nilai suaian $\hat{Y}_i$ yaitu

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i. \tag{3.30}$$

3.4.1 Sifat-sifat aljabar statistik kuadrat terkecil

Berikut ini adalah beberapa sifat aljabar penting dari statistik kuadrat terkecil (Stock and Watson, 2012) dan Wooldridge (2006).

1. Jumlah, demikian pula rata-rata sampel, dari sisaan kuadrat terkecil adalah nol. Dengan kata lain,

$$\sum_{i=1}^{n} \hat{u}_i = 0. \tag{3.31}$$

   Ini adalah akibat langsung dari kondisi pada kuadrat terkecil. Interpretasi lain adalah rata-rata dari nilai suaian $\hat{Y}_i$ sama dengan rata-rata sampel. Artinya

$$\frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i = \bar{Y}. \tag{3.32}$$

2. Kovarians sampel antara regresor atau peubah bebas dan sisaan adalah nol. Dengan demikian

$$\sum_{i=1}^{n} \hat{u}_i X_i = 0. \tag{3.33}$$

3. Titik-titik $(\bar{X}, \bar{Y})$ akan selalu berada pada garis regresi.

Berdasarkan sifat-sifat penting diatas dapat kita lihat bahwa Y_i dapat didekomposisi menjadi dua yaitu nilai suaian $\hat{Y}_i$ dan residual $\hat{u}_i$. Nilai suaian dan sisaan tidak berkorelasi dalam sampel. Apabila kita defisikan jumlah kuadrat total (*total sum of squares*), disingkat JKT; jumlah kuadrat regresi atau jumlah kuadrat yang dijelaskan (*explained sum of squares*), disingkat JKR; dan jumlah kuadrat galat/sisaan (*residual sum of squares*), disingkat disingkat JKG. Jadi kita definisikan,

$$\text{JKT} = \sum_{i=1}^{n} (Y_i - \bar{Y}), \tag{3.34}$$

$$\text{JKR} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}), \tag{3.35}$$

$$\text{JKG} = \sum_{i=1}^{n} \hat{u}_i^2. \tag{3.36}$$

Sehingga diperoleh hubungan

$$\text{JKT} = \text{JKR} + \text{JKG}. \tag{3.37}$$

3.4.2 Kecocokan suai (*goodness of fit*)

Kita telah melihat bagaimana menggunakan metode kuadrat terkecil untuk mengestimasi parameter β_0 dan β_1. Asumsikan jumlah kuadrat total JKT tidak sama dengan nol (Catatan: JKT akan sama dengan nol apabila semua nilai Y_i sama. Dan hal ini tentu saja tidaklah mungkin.). Dengan demikian apabila kita bagi JKT dengan JKT sendiri maka diperoleh

$$\frac{\text{JKT}}{\text{JKT}} = \frac{\text{JKR}}{\text{JKT}} + \frac{\text{JKG}}{\text{JKT}} \tag{3.38}$$

atau

$$1 = \frac{\text{JKR}}{\text{JKT}} + \frac{\text{JKG}}{\text{JKT}}. \tag{3.39}$$

Kuantitas R-kuadrat regresi, disebut pula koefisien determinasi, didefinisikan sebagai

$$R^2 = \frac{\text{JKR}}{\text{JKT}} = 1 - \frac{JKG}{\text{JKT}}. \tag{3.40}$$

Nilai R^2 ini adalah rasio dari variasi yang dijelaskan dengan variasi total. Dengan demikian, ini dapat diinterpretasikan sebagai fraksi atau pecahan variasi sampel pada Y yang dijelaskan oleh X. Nilai $0 \leq R^2 \leq 1$. Ketika kita menginterpretasikan R^2 biasanya dikalikan dengan nilai 100 untuk mengubahnya menjadi persen: ”$100 \cdot R^2$ adalah persentase dari variasi sampel pada Y yang dijelaskan oleh X.”

Jika semua titik data terletak pada garis yang sama maka, metode kuadrat terkecil memberikan suaian yang sempurna pada data. Dalam hal ini $R^2 = 1$. Nilai R^2 yang dekat dengan nol mengindikasikan suaian yang tidak bagus.

3.4.3 Asumsi-asumsi dalam regresi linear sederhana

Sebelum membicarakan lebih lanjut sifat-sifat pendugaan metode kuadrat terkecil, terlebih dahulu kita lihat kembali apa saja asumsi yang digunakan pada saat kita menerapkan analisis regresi linear sederhana, yang dalam hal ini disebut pula sebagai analisis regresi linear klasik. Asumsi-asumsi berikut diambil dari Gujarati (2003) dan Wooldridge (2006).

1. Model regesi linear. Model regresi linear yang dimaksud adalah linear dalam parameter. Sebagai contoh $Y_i = \beta_0 + \beta_1 X_i + u_i$.

2. Nilai X tidak berubah dalam pengambilan sampel secara berulang. Nilai regresor X tidak berubah dalam pengambilan sampel secara berulang atau X dianggap nonstokastik.

3. Nilai tengah gangguan atau galat u_i adalah nol. Diberikan nilai X_i, nilai tengah atau nilai harapan peubah acak gangguan u_i adalah nol. Secara teknis nilai tengah bersyarat u_i adalah nol, yaitu $\text{E}(u_i|X_i) = 0$.

4. Homoskedastisitias varians atau varians yang sama untuk u_i. Dengan kata lain $\text{var}(u_i|X_i) = \sigma^2$.

5. Tidak terdapat autokorelasi di antara gangguan. Artinya untuk sebarang dua nilai X_i dan X_j, korelasi antara u_i dan u_j, untuk $i \neq j$ adalah nol. Dengan kata lain $\text{cov}(u_i, u_j|X_i, X_j) = 0$.

6. Kovarians antara u_i dan X_i adalah nol. Dengan kata lain $\text{E}(u_i X_i) = 0$.

7. Banyak amatan n harus lebih besar dari jumlah parameter yang akan diestimasi. Dengan kata lain, banyak akamatan n harus lebih besar dari jumlah peubah bebas atau regresor.
8. Variabilitas pada nilai X. Nilai X pada suatu sampel tidaklah boleh sama semuanya. Secara teknis, $\text{var}(X) > 0$.
9. Model regresi dispesifikasikan dengan benar. Dengan kata lain, tidak terdapat bias spesifikasi atau galat dalam model yang digunakan dalam analisis empiris.
10. Tidak terjadi multikolinearitas yang sempurna. Artinya, tidak terdapat hubungan linear sempurna di antara peubah-peubah prediktor.

3.4.4 Nilai harapan dan varians penduga kuadrat terkecil

Pada subabagian sebelumnya kita telah memperoleh

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}, \tag{3.41}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \tag{3.42}$$

Selnjutnya kita ingin mengetahui apakah $\hat{\beta}_0$ dan $\hat{\beta}_1$ bersifat takbias. Terlebih dahulu kita nyatakan

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \tag{3.43}$$

$$= \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \tag{3.44}$$

$$= \frac{\sum_{i=1}^n (X_i - \bar{X})(\beta_0 + \beta_1 X_i + u_i)}{\sum_{i=1}^n (X_i - \bar{X})^2}. \tag{3.45}$$

Kemudian sederhanakan

$$\sum_{i=1}^n (X_i - \bar{X})(\beta_0 + \beta_1 X_i + u_i) = \beta_0 \sum_{i=1}^n (X_i - \bar{X}) + \beta_1 \sum_{i=1}^n (X_i - \bar{X})X_i + \sum_{i=1}^n (X_i - \bar{X})u_i. \tag{3.46}$$

Selanjutnya substitusikan (3.46) pada pembilan pada persamaan (3.45). Hasilnya adalah

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n (X_i - \bar{X})u_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \tag{3.47}$$

Sekarang

$$\mathrm{E}(\hat{\beta}_1) = \mathrm{E}\left(\beta_1 + \frac{\sum_{i=1}^n (X_i - \bar{X})u_i}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) \tag{3.48}$$

$$= \beta_1 + \mathrm{E}\left(\frac{\sum_{i=1}^n (X_i - \bar{X})u_i}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) \tag{3.49}$$

$$= \beta_1 + \frac{\sum_{i=1}^n (X_i - \bar{X})\,\mathrm{E}(u_i)}{\sum_{i=1}^n (X_i - \bar{X})^2} \tag{3.50}$$

$$= \beta_1 + 0 = \beta_1. \tag{3.51}$$

Jadi $\hat{\beta}_1$ bersifat takbias. Bagaimana dengan $\hat{\beta}_0$? Kita tuliskan

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{3.52}$$

$$= \beta_0 + \beta_1 \bar{X} + \bar{u} - \hat{\beta}_1 \bar{X} \tag{3.53}$$

$$= \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{X} + \bar{u}. \tag{3.54}$$

Sehingga

$$\mathrm{E}(\hat{\beta}_0) = \mathrm{E}(\beta_0 + (\beta_1 - \hat{\beta}_1)\bar{X} + \bar{u}) = \beta_0 + \mathrm{E}(\beta_1 - \hat{\beta}_1)\bar{X} + \mathrm{E}(\bar{u}) = \beta_0 \tag{3.55}$$

karena $\mathrm{E}(\hat{u}) = 0$ dan $\mathrm{E}(\hat{\beta}_1) = \beta_1$. Sekarang bagaimana dengan varians $\hat{\beta}_1$ dan $\hat{\beta}_0$?

$$\operatorname{var}(\hat{\beta}_1) = \left(\frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)^2 \operatorname{var}\left(\sum_{i=1}^n (X_i - \bar{X})u_i\right) \tag{3.56}$$

$$= \left(\frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)^2 \left(\sum_{i=1}^n (X_i - \bar{X})^2\right) \operatorname{var}(u_i) \tag{3.57}$$

$$= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}. \tag{3.58}$$

Dengan cara yang sama dapat ditunjukkan

$$\operatorname{var}(\hat{\beta}_0) = \frac{\sigma^2 n^{-1}}{\sum_{i=1}^n (X_i - \bar{X})^2} \sum_{i=1}^n X_i^2. \tag{3.59}$$

Coba Anda tunjukkan!

3.4.5 Mengestimasi varians galat

Ingat kembali sisaan yang didefinisikan sebagai

$$\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = (\beta_0 + \beta_1 X_i + u_i) - \hat{\beta}_0 - \hat{\beta}_1 X_i \tag{3.60}$$

atau dapat pula dinyatakan sebagai

$$\hat{u}_i = u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)X_i. \tag{3.61}$$

Penduga takbias untuk σ^2 adalah

$$\hat{\sigma}^2 = \frac{1}{(n-2)} \sum_{i=1}^n \hat{u}_i^2 \tag{3.62}$$

dan bersifat takbias. Coba Anda tunjukkan!

Teorema 3.4.1 (Wooldridge (2006))**.** *Berdasarkan asumsi kelinearan dalam parameter, sampel acak, variasi sampel dalam peubah prediktor, nilai tengah bersyarat, dan homoskedastisitas berlaku*

$$\mathrm{E}(\hat{\theta}^2) = \sigma^2. \tag{3.63}$$

Sebagai catatan penduga alamiah dari θ adalah

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} \tag{3.64}$$

yang disebut galat baku regresi (*standard error of the regression*, disingkat SER). Nama lain dari SER adalah galat baku estimasi (*standar error of the estimate*) dan galat kuadrat nilai tengah akar (*root mean squared error*).

3.4.6 Sifat-sifat penduga kuadrat terkecil

Penduga kuadrat terkecil memenuhi sifat BLUE. Hal ini dijamin oleh Teorema Gauss-Markov yang menyatakan bahwa diberikan asumsi-asumsi model liner regresi klasik, penduga kuadraat terkecil, dalam keluarga penduga linear takbias, memiliki varians minimum dan memenuhi sifat BLUE.

3.5 Model Regresi Linear Normal Klasik

Pada bagian sebelumnya kita telah membahas bagaimana menggunakan metode kuadrat terkecil untuk mendapatkan estimasi parameter. Namun, metode kuadrat terkecil tidak membuat asumsi probabilistik terhadap u_i. Oleh karena itu, untuk melakukan inferensi terhadap fungsi regresi populasi dari fungsi regresi sampel, Teorema Gauss-Markov tidak dapat digunakan. Oleh karena itu diasumsikan galat u_i menyebar normal. Penambahan asumsi ini pada model regresi klasik yang telah kita bahas dikenal sebagai model regresi linear normal klasik.

Asumsi model regresi linear normal klasik ini mengasumsikan masing-masing u_i berdistribusi normal dengan nilai tengah $\mathrm{E}(u_i) = 0$, varians $\mathrm{E}(u_i - \mathrm{E}(u_i))^2 = \mathrm{E}(u_i^2) = \sigma^2$ dan kovarians $\mathrm{cov}(u_i, u_j) = \mathrm{E}(u_i u_j) = 0$ untuk $i \neq j$. Asumsi ini secara singkat dapat ditulis sebagai $u_i \sim N(0, \sigma^2)$. Lebih jelasnya diasumsikan saling bebas dan berdistribusi identik, atau IID normal. Pertanyaanya kenapa harus mengasumsikan kenormalan?

Menurut Gujarati (2003) ada beberapa alasan:

1. Kita telah membicarakan bahwa galat u_i menyatakan pengaruh kombinasi pengaruh banyak variabel independen pada variabel dependen yang tidak secara eksplisit dimasukkan pada model regresi. Harapan kita tentu saja pengaruh variabel yang dihilangkan ini atau diabaikan ini kecil dan acak. Teorema limit pusat menyatakan jika terdapat sejumlah besar variabel acak yang berdistribusi bebas dan identik, dengan sedikit pengecualian, distribusi dari jumlahnya menuju distribusi normal.

2. Varian lain dari Teorema Limit Pusat mengatakan bahwa bahkan jika jumlah variabel tidak sangat besar atau jika variabel-variabel ini tidak saling bebas tegas (*strictly independent*), jumlahnya mungkin tetap normal.

3. Dengan asumsi kenormalan, distribusi peluang dari penduga kuadrat terkecil dapat diperoleh dengan mudah karena sifat kenormalan yaitu sebarang fungsi linear dari variabel yang menyebar normal juga menyebar normal. Apabila kita lihat pada bagian sebelumnya $\hat{\beta}_0$ dan $\hat{\beta}_1$ adalah fungsi linear dari u_i. Dengan demikian, jika u_i menyebar normal, maka begitu pula dengan $\hat{\beta}_0$ dan $\hat{\beta}_1$.

4. Distribusi normal merupakan distribusi yang cukup sederhana yang hanya melibatkan dua parameter, yaitu nilai tengah dan varians. Sifat-sifat distribusi ini telah banyak dipelajari secara ekstensif.

5. Apabila kita menangani data dengan ukuran kecil atau berhingga, katakanlah 100, asumsi kenormalan memegang peranan penting. Asumsi ini membantu kita tunuk mendapatkan distribusi peluang eksak dari penduga kuadra terkecil.

3.5.1 Sifat-sifat Penduga Kuadrat Terkecil di bawah Asumsi Kenormalan

Berdasarkan asumsi bahwa u_i berdistribusi normal, maka penduga kuadrat terkecil memiliki sifat-sifat berikut (Gujarati, 2003):

1. Bersifat takbias.
2. Memiliki varians minimum. Lebih jelasnya, penduga takbias minimum atau penduga efisien.
3. Konsisten. Artinya, seiring meningkatnya ukuran sampel menuju tak berhingga, penduga-penduga tersebut konvergen menuju nilai parameter populasi sesungguhnya.
4. Intersep $\hat{\beta}_0 \sim N(\beta_0, \sigma^2_{\hat{\beta}_0})$ dengan
$$\sigma^2_{\hat{\beta}_0} = \frac{\sigma^2 \sum X_i^2}{n \sum (X_i - \bar{X})^2} \tag{3.65}$$
5. Lereng $\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$ dengan
$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{n \sum (X_i - \bar{X})^2} \tag{3.66}$$
6. Kuantitas $(n-2)(\hat{\sigma}^2/\sigma^2)$ berdistribusi χ^2 dengan derajat bebas $(n-2)$. Informasi ini dapat digunakan untuk melakukan inferensi terhadap parameter sesungguhnya σ^2 dari nilai estimasi $\hat{\sigma}$.
7. Estimasi $(\hat{\beta}_0, \hat{\beta}_2)$ menyebar dan saling bebas dari $\hat{\sigma}^2$. Keduanya memiliki varians minimum dari semua keluarga penduga tak bias, apakah linear atau tidak.

Berdasarkan asumsi kenormalan galat, kita dapat menuliskan model regresi linear sederhana sebagai
$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2). \tag{3.67}$$
Fungsi likelihood untuk (3.67) adalah
$$L(\beta_0, \beta_1, \sigma^2) = \frac{1}{(\sqrt{2\pi})^n \sigma^n} \exp\left[-\frac{1}{2} \sum_{i=1}^{n} \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{\sigma^2}\right]. \tag{3.68}$$

Selanjutnya menghitung loglikelihood dan menurunkan terhadap β_0 dan β_1 akan menghasilkan nilai $\tilde{\beta}_0$ dan $\tilde{\beta}_1$ seperti pada metode kuadrat terkecil. Namun untuk
$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2 \tag{3.69}$$
yang berbeda dengan $\hat{\sigma}^2 = [1/(n-2)] \sum \hat{u}_i^2$ yang diperoleh dari metode kuadrat terkecil. Catatan nilai $\hat{\sigma}^2$ ini telah ditunjukkan sebagai penduga takbias dari σ^2.

3.6 Pendugaan Selang dan Uji Hipotesis

3.6.1 Pendugaan Selang

Pada bagian sebelumnya kita telah membahas distribusi dari $\widehat{beta}_0$ dan $\hat{\beta}_1$. Karena keduanya menyebar normal, maka kita dapat membentuk statistik

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\mathrm{SE}(\hat{\beta}_1)} \tag{3.70}$$

Dalam praktiknya σ tidak diketahui, sehingga biasanya diestimasi dengan penduga takbias $\hat{\sigma}^2$. Dengan kata lain

$$t = \frac{\hat{\beta}_1 - \beta_1}{\mathrm{SE}(\hat{\beta}_1)} = \frac{(\beta_1 - \beta_1)\sqrt{\sum(X_i - \bar{X})^2}}{\hat{\sigma}}. \tag{3.71}$$

Statistik t ini memiliki derajat bebas $n-2$. Selanjutnya dengan mengatur masing-masing suku diperoleh

$$P(-t_{\alpha/2} \leq t \leq t_{\alpha/2}) = 1 - \alpha. \tag{3.72}$$

Sebagai contoh untuk β_0 kita akan memperoleh

$$P[\hat{\beta}_0 - t_{\alpha/2}\mathrm{SE}(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2}\mathrm{SE}(\hat{\beta}_0)] = 1 - \alpha \tag{3.73}$$

atau bisa juga ditulis sebagai

$$\hat{\beta}_0 \pm t_{\alpha/2}\mathrm{SE}(\hat{\beta}_0). \tag{3.74}$$

Interpretasi selang kepercayaan ini adalah sebagai berikut. Diberikan selang kepercayaan, misalnya 95%, dalam jangka panjang, 95 dari 100 kasus interval, berisi nilai β sesungguhnya. Kita tidak bisa menginterpretasikan sebagai peluang bahwa 95% bahwa selang berisi nilai β sesungguhnya karena interval ini sekarang tetap dan tidak lagi acak. Dengan demikian, kita hanya bisa mengatakan apakah β berada di dalam selang atau tidak. Peluang bahwa selang kepercayaan berisi nilai β adalah 0 atau 1.

Selang kepercayaan untuk σ^2 dapat dihitung berdasarkan asumsi kenormalan. Kita tahu bahwa $(n-2)\hat{\sigma}^2/\sigma^2$ berdistribusi χ^2_{n-2}. Dengan cara serupa kita akan memeperoleh selang kepercayaan berbentuk

$$P\left[(n-2)\frac{\hat{\sigma}^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq (n-2)\frac{\hat{\sigma}^2}{\chi^2_{1-\alpha/2}}\right] = 1 - \alpha. \tag{3.75}$$

Interpretasi selang kepercayaan ini serupa dengan interpretasi pada kasus sebelumnya. Jika kita mendapatkan selang kepercayaan 95%, kita akan mengharapkan 95 dari 100 selang ini akan berisis nilai σ^2 sesungguhnya.

3.6.2 Uji Hipotesis

Pada pengujian hipotesis terhadap parameter, kita dapat menggunakan pendugaan selang seperti pada bagian sebelumnya. Buat selang kepercayaan $100(1-\alpha)\%$ untuk β. Apabila β, di bawah H_0, berada pada selang ini, jangan tolak H_0, namun jika di luar selang ini tolak H_0.

Pengujian hipotesis menggunakan statistik uji dapat diringkas sebagai berikut:

1. Dua ekor: $H_0 : \beta = \beta^*$ versus $H_1 : \beta \neq \beta^*$ dengan daerah kritis $|t| > t_{\alpha/2,\mathrm{db}}$,
2. Ekor kanan: $H_0 : \beta \leq \beta^*$ versus $H_1 : \beta > \beta^*$ dengan daerah kritis $t > t_{\alpha,\mathrm{db}}$,
3. Ekor kiri: $H_0 : \beta \geq \beta^*$ versus $H_1 : \beta < \beta^*$ dengan daerah kritis $t < -t_{\alpha,\mathrm{db}}$.

Dengan cara serupa kita juga dapat melakukan uji hipotesis untuk σ^2.

3.7 Pemeriksaan Diagnostik

Salah satu syarat penting yang harus dipenuhi adalah asumsi kenormalan. Untuk melihat ini dapat dilakukan dengan beberapa cara. Pertama, memplot histogram sisaan. Jika kita melihat bentuk menyerupai lonceng, maka ada indikasi sisaan berdistribusi normal. Cara kedua adalah dengan melihat plot kuantil-kuantil normal. Jika sisaan berada mendekati garis lurus, maka ada indikasi sisaan berdistribusi normal. Ketiga, uji kenormalan menggunakan uji Anderson-Darling atau Jarque-Bera.

BAB IV

REGRESI LINEAR BERGANDA

Pada bab sebelumnya kita telah mempelajari model regresi linear dengan dua peubah yaitu satu peubah prediktor dan satu peubah respons. Kita juga telah mempelajari bagaimana mendapatkan persamaan kuadrat terkecil (*least squares*) dan sifat-sifat penduga yang diperoleh dengan metode kuadrat terkecil.

Kita juga telah melihat pengembangan model regresi berganda (*multiple regression*) tersebut untuk tiga peubah, yakni dua peubah prediktor dan satu peubah respons. Kita juga telah bagaimana menghitung penduga dengan metode kuadrat terkecil untuk model regresi berganda dan proses penghitungan pendugaan cenderung melibatkan banyak manipulasi aljabar yang ”menjemukan”.

Bab ini akan membahas penggunaan notasi aljabar matriks untuk model regresi linear berganda. Keuntungan menggunakan aljabar matriks adalah memudahkan dalam memodelkan regresi dengan berapapun banyak peubah. Dengan kata lain, begitu model dengan k peubah dirumuskan dan diselesaikan dengan notasi matriks, solusinya berlaku untuk satu, dua, tiga, atau sebarang jumlah peubah.

4.1 Model Regresi Linear dengan k Peubah

Misalkan model regresi populasi dengan peubah respons Y dan peubah prediktor X_1, X_2, X_3,..., X_k yang dapat ditulis sebagai

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, \quad i = 1, 2, \ldots, n \tag{4.1}$$

dengan β_0 adalah intersep, $\beta_1, \ldots, \beta_k$ adalah koefisien lereng (*slope*) partial, u adalah galat, i adalah amatan ke-i, dan n adalah ukuran populasi. Notasi pada persamaan (4.1) pada dasarnya merupakan pernyataan dari persamaan simultan berbentuk

$$\begin{aligned}
Y_1 &= \beta_0 + \beta_1 X_{11} + \cdots + \beta_k X_{k1} + u_1 \\
Y_2 &= \beta_0 + \beta_1 X_{12} + \cdots + \beta_k X_{k2} + u_2 \\
&\vdots \\
Y_n &= \beta_0 + \beta_1 X_{1n} + \cdots + \beta_k X_{kn} + u_n
\end{aligned} \tag{4.2}$$

Sistem persamaan (4.2) dapat dituliskan dalam bentuk yang lebih ”elegan” sebagai berikut

$$\begin{array}{ccccccc}
\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} & = & \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix} & \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} & + & \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \\
\mathbf{Y} & = & \mathbf{X} & \boldsymbol{\beta} & + & \mathbf{u} \\
n \times 1 & & n \times (k+1) & (k+1) \times 1 & + & n \times 1
\end{array} \tag{4.3}$$

dengan $\mathbf{Y} = n \times 1$ adalah vektor kolom amatan pada peubah respons Y,$\mathbf{X}$ adalah matriks berdimensi $n \times (k+1)$ (kolom pertama yang berisi angka 1menyatakan intersep), $\boldsymbol{\beta}$ adalah vektor kolom berdimensi $(k+1) \times 1$ dari parameter yang tidak diketahui dan $\mathbf{u}$ adalah vektor galat berdimensi $n \times 1$. Dengan demikian model regresi dengan k peubah dapat dinotasikan menggunakan representasi matriks sebagai berikut:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \tag{4.4}$$

4.2 Asumsi-asumsi Model Regresi Linear dalam Notasi Matriks

Seperti halnya asumsi pada model regresi linear sederhana dengan dua peubah, asumsi model regresi linear berganda pun sama dengan beberapa penyesuaian.

1. Diasumsikan $\mathrm{E}(\mathbf{u}) = \mathbf{0}$ dengan $\mathbf{u}$ dan $\mathbf{0}$ adalah vektor kolom berdimensi $n \times 1$; $\mathbf{0}$ adalah vektor nol. Dengan kata lain

$$\mathrm{E}\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} \mathrm{E}(u_1) \\ \mathrm{E}(u_2) \\ \vdots \\ \mathrm{E}(u_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}. \tag{4.5}$$

2. Nilai $\mathrm{E}(\mathbf{u}\mathbf{u}') = \sigma^2\mathbf{I}$ dengan $\mathbf{I}$ adalah matriks identitas berukuran $n \times n$. Dengan kata lain,

$$\mathrm{E}(\mathbf{u}\mathbf{u}') = \begin{bmatrix} \sigma^2 & 0 & 0\cdots & 0 \\ 0 & \sigma^2 & 0\cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}. \tag{4.6}$$

   Persamaan (4.6) disebut matriks varians-kovarians dari galat u.

3. Matriks $\mathbf{X}$ yang berdimensi $n \times (k+1)$ adalah nonstokastik, yaitu matriks yang terdiri dari sejumlah bilangan yang tetap (*fixed numbers*) atau dianggap tidak berubah. Artinya, analisis regresi yang kita lakukan adalah analisis regresi bersyarat pada nilai $\mathbf{X}$ yang tetap.

4. Rank matriks $\mathbf{X}$ adalah $p(\mathbf{X}) = k$, dengan k adalah banyak kolom dalam matriks $\mathbf{X}$ dan k kurang dari banyak amatan n. Asumsi ini berarti kolom-kolom $\mathbf{X}$ saling bebas linear (*linearly independent*), artinya tidak terdapat hubungan linear pasti (*exact linear relationship*) di antara peubah x. Dengan kata lain, tidak terdapat multikolinearitas.

5. Vektor $\mathbf{u}$ berdistribusi normal, artinya $\mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

4.3 Pendugaan Kuadrat Terkecil

Untuk mendapatkan pendugaan kuadrat terkecil terlebih dahulu kita tuliskan regresi sampel dengan k peubah yaitu

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki} + \hat{u}_i, \quad i = 1, 2, \ldots, n \tag{4.7}$$

atau bisa dinyatakan dengan cara serupa

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}. \tag{4.8}$$

Pada regresi linear berganda dengan k peubah, tujuan pendugaan kuadrat terkecil adalah meminimumkan

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} \tag{4.9}$$

dengan $\hat{\mathbf{u}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. Dengan demikian bentuk kuadrat

$$\begin{aligned} \hat{\mathbf{u}}'\hat{\mathbf{u}} &= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{Y}'\mathbf{Y} - 2\hat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{Y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}. \end{aligned} \tag{4.10}$$

Selanjutnya untuk mencari penduga dengan metode kuadrat terkecil dengan menurunkan persamaan (4.10) terhadap $\boldsymbol{\beta}$ diperoleh

$$\frac{\partial \hat{\mathbf{u}}'\hat{\mathbf{u}}}{\partial \hat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}. \tag{4.11}$$

Selanjutnya menyamakan dengan nol dan mengatur suku-suku persamaan pada (4.11) diperoleh

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{4.12}$$

Catatan: Jika $\mathbf{a}' = [a_1 \quad a_2 \quad \cdots \quad a_n]$ adalah vektor baris dari suatu bilangan, $\mathbf{x}$ adalah vektor kolom dari peubah $x_1, x_2, \ldots, x_n$ yaitu

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{4.13}$$

dan matriks simetris

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \tag{4.14}$$

maka turunan dari bentuk

$$\frac{\partial \mathbf{a}'\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}. \tag{4.15}$$

dan

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x} \tag{4.16}$$

yang merupakan vektor kolom dari n elemen, atau

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}'\mathbf{A} \tag{4.17}$$

yang merupakan vektor baris dengan n elemen.

4.3.1 Matriks varians-kovarians

Penggunaan matriks memungkinkan kita untuk tidak saja menghitung varians $\hat{\beta}_i$ untuk sebarang elemen $\hat{\boldsymbol{\beta}}$, namun juga kovarians dari dua elemen pada $\hat{\boldsymbol{\beta}}$. Varians dan kovarians ini diperlukan untuk inferensi statistika. Matriks varians-kovarians didefinisikan sebagai

$$\text{var-cov}(\hat{\boldsymbol{\beta}}) = \mathrm{E}[(\hat{\boldsymbol{\beta}} - \mathrm{E}(\hat{\boldsymbol{\beta}}))(\hat{\boldsymbol{\beta}} - \mathrm{E}(\hat{\boldsymbol{\beta}}))']. \tag{4.18}$$

Bentuk umum matriks varians-kovarians $\hat{\boldsymbol{\beta}}$ adalah sebagai berikut

$$\text{var-cov}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} \text{var}\,(\hat{\beta}_1) & \text{cov}(\hat{\beta}_1, \hat{\beta}_2) & \cdots & \text{cov}(\hat{\beta}_1, \hat{\beta}_k) \\ \text{cov}\,(\hat{\beta}_2, \hat{\beta}_1) & \text{var}(\hat{\beta}_2) & \cdots & \text{cov}(\hat{\beta}_2, \hat{\beta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}\,(\hat{\beta}_k, \hat{\beta}_1) & \text{cov}(\hat{\beta}_k, \hat{\beta}_2) & \cdots & \text{var}(\hat{\beta}_k) \end{bmatrix}. \tag{4.19}$$

Pada bagian ini kita akan menghitung matriks varians-kovarians $\hat{\boldsymbol{\beta}}$. Kita telah mendapatkan

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{4.20}$$

Mensubstitusikan $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ ke persamaan (4.20) diperoleh

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}. \end{aligned} \tag{4.21}$$

Selanjutnya kita peroleh

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}. \tag{4.22}$$

Selanjutnya kita bisa menghitung varians-kovarians $\hat{\boldsymbol{\beta}}$ menggunakan definisi (4.18)

$$\begin{aligned} \text{var-cov}(\hat{\boldsymbol{\beta}}) &= \mathrm{E}[((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u})((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u})'] \\ &= \mathrm{E}[((\mathbf{X}'\mathbf{X})^{-1})\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1})] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{E}(\mathbf{u}\mathbf{u}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \tag{4.23}$$

Pada model regresi dengan dua dan tiga peubah kita peroleh penduga takbias untuk σ^2 adalah $\hat{\sigma}^2 = \sum \hat{u}_i^2/(n-2)$ dan $\hat{\sigma}^2 = \sum \hat{u}_i^2/(n-3)$. Untuk k peubah kita akan peroleh

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-k} = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k}. \tag{4.24}$$

Nilai $\hat{\mathbf{u}}'\hat{\mathbf{u}}$ dapat dihitung menggunakan

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{Y}. \tag{4.25}$$

4.3.2 Sifat-sifat Penduga Kuadrat Terkecil

Menggunakan persamaan (4.21) kita dapat melihat bahwa $\hat{\boldsymbol{\beta}}$ bersifat takbias. Kenapa? Penduga ini juga bersifat BLUE. Misalkan $\hat{\boldsymbol{\beta}}^*$ adalah penduga linear dari $\hat{\boldsymbol{\beta}}$ yang dapat ditulis sebagai

$$\hat{\boldsymbol{\beta}}^* = [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{C}]\mathbf{Y} \tag{4.26}$$

dengan $\mathbf{C}$ adalah matriks konstanta. Selanjutnya

$$\begin{aligned}\hat{\boldsymbol{\beta}}^* &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{C}](\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= \boldsymbol{\beta} + \mathbf{CX}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{u} + \mathbf{Cu}.\end{aligned} \tag{4.27}$$

Agar $\hat{\boldsymbol{\beta}}^*$ takbias maka $\mathbf{CX} = \mathbf{0}$. Dengan cara yang sama seperti pada saat menghitung varians-kovarians $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}} = \mathrm{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} + \mathbf{Cu}][(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} + \mathbf{Cu}]' \tag{4.28}$$

sehingga kita peroleh

$$\text{var-cov}(\hat{\boldsymbol{\beta}}^*) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \sigma^2\mathbf{CC}' \tag{4.29}$$

dengan matriks $\mathbf{CC}'$ adalah matriks semidefinit positif.

4.3.3 Koefisien Determinasi

Koefisien determinasi R^2 didefinisikan sebagai rasio jumlah kuadrat yang dijelaskan regresi (*explained sum of squares*, disingkat ESS) dan jumlah kuadrat total (*total sum of squares*, disingkat TSS). (Lihat kembali foto kopi materi tentang analisis regresi sederhana). Koefisien determinasi R^2 didefinisikan sebagai

$$R^2 = \frac{\text{ESS}}{\text{TSS}} \tag{4.30}$$

Pada kasus dua peubah nilai

$$R^2 = \frac{\hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \tag{4.31}$$

dan pada kasus tiga peubah nilai

$$R^2 = \frac{\hat{\beta}_1 \sum_{i=1}^n y_i(x_{2i} - \bar{x}_2) + \hat{\beta}_2 \sum_{i=1}^n y_i(x_{3i} - \bar{x}_3)}{\sum_{i=1}^n (y_i - \bar{y})^2}. \tag{4.32}$$

Merampatkan (*generalizing*) untuk k peubah diperoleh

$$R^2 = \frac{\hat{\beta}_1 \sum_{i=1}^n y_i(x_{2i} - \bar{x}_2) + \hat{\beta}_2 \sum_{i=1}^n y_i(x_{3i} - \bar{x}_3) + \cdots + \hat{\beta}_k \sum_{i=1}^n y_i(x_{ki} - \bar{x}_k)}{\sum_{i=1}^n (y_i - \bar{y})^2}. \tag{4.33}$$

Dalam notasi matriks

$$\text{TSS} = \sum (y_i - \bar{y})^2 = \mathbf{Y}'\mathbf{Y} - n\bar{y}^2 \tag{4.34}$$

dan

$$\text{ESS} = \hat{\beta}_1 \sum y_i(x_{2i} - \bar{x}_2) + \cdots + \hat{\beta}_k \sum y_k(x_{ki} - \bar{x}_k) = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - n\bar{y}^2. \tag{4.35}$$

Dengan demikian R^2 dalam notasi matriks adalah

$$R^2 = \frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - n\bar{y}^2}{\mathbf{Y}'\mathbf{Y} - n\bar{y}^2}. \tag{4.36}$$

4.4 Uji Hipotesis Koefisien Regresi Individu

Pada analisis regresi berganda diasumsikan masing-masing galat u_i berdistribusi normal dengan nilai tengah 0 dan varians konstan σ^2. Dalam notasi matriks ini dinyatakan oleh

$$\mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \tag{4.37}$$

dengan $\mathbf{u}$ dan $\mathbf{0}$ adalah vektor kolom berukuran $n \times 1$ dan $\mathbf{I}$ adalah matriks identitas dan $\mathbf{0}$ adalah matriks vektor nol. Untuk k peubah dapat ditunjukkan bahwa

$$\hat{\boldsymbol{\beta}} \sim N[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]. \tag{4.38}$$

Catatan: Suatu peubah acak $\mathbf{x} = (x_1, x_2, \ldots, x_n)'$ dikatakan berdistribusi normal multivariat dengan vektor nilai tengah $\boldsymbol{\mu}$ dan matriks kovarians $\boldsymbol{\Sigma}$ apabila memiliki densitas peluang berbentuk

$$f(\mathbf{x}) = (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\big[(-1/2)(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma} (\mathbf{x} - \boldsymbol{\mu})\big]. \tag{4.39}$$

Dalam praktiknya σ^2 tidak diketahui, dan biasanya diestimasi dengan $\hat{\sigma}^2$. Artinya, distribusi yang digunakan adalah distribusi t karena masing-masing elemen $\hat{\boldsymbol{\beta}}$ akan berdistribusi t dengan derajat bebas $n - k$. Ini biasanya dinotasikan oleh

$$t = \frac{\hat{\beta}_i - \beta_i}{\mathrm{se}(\hat{\beta}_i)} \tag{4.40}$$

dengan derajat bebas $n - k$ dan $\hat{\beta}_i$ adalah sebarang elemen dari $\hat{\boldsymbol{\beta}}$.

4.5 Uji Hipotesis Signifikansi Regresi Secara Keseluruhan

Untuk menguji koefisien regresi secara keseluruhan digunakan teknik analisis varians (*analysis of variance*, disingkat ANOVA). Diasumsikan galat u_i berdistribusi normal dan hipotesi nol dinyatakan sebagai $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ dan statistik uji

$$F = \frac{(\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - n\bar{y}^2)/(k-1)}{(\mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y})/(n-k)} \tag{4.41}$$

yang berdistribusi F dengan derajat bebas $k - 1$ dan $n - k$.

Tabel 4.5.1: Formulasi Matriks Tabel Anova untuk Model Regresi Linear untuk k Peubah

Sumber variasi	SS	df	MSS
Regresi	$\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - n\bar{y}^2$	$k-1$	$\frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - n\bar{y}^2}{k-1}$
Sisaan	$\mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$	$n-k$	$\frac{\mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}}{n-k}$
Total	$\mathbf{Y}'\mathbf{Y} - n\bar{y}^2$	$n-1$	

4.6 Prediksi

Misalkan

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix} \tag{4.42}$$

adalah vektor nilai peubah X yang akan digunakan untuk memprediksi $\hat{y}_0$, yaitu nilai tengah prediksi y. Nilai estimasi regresi berganda dapat dinyatakan sebagai

$$\hat{y}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}} \tag{4.43}$$

dengan $\mathbf{x}_i' = [1 \quad x_{1i} \quad x_{ki}]$ dan

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}. \tag{4.44}$$

Apabila amatan $\mathbf{x}_0'$, maka

$$(\hat{y}|\mathbf{x}_0') = \mathbf{x}_0'\hat{\boldsymbol{\beta}}. \tag{4.45}$$

Varians prediksi ini adalah

$$\mathrm{var}(\hat{y}_0|\mathbf{x}_0') = \sigma^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0. \tag{4.46}$$

BAB V

HETEROSKEDASTISITAS

Salah satu asumsi penting dalam model regresi linear klasik adalah galat u_i dalam model regresi populasi adalah homoskedastik, yaitu memiliki varians yang sama. Bab ini membahas bagaimana validitas dari asumsi ini dan mengetahui apa yang terjadi apabila asumsi ini tidak dipenuhi. Materi pada bab ini diadaptasi dari Gujarati (2003).

5.1 Latar Belakang Heteroskedastisitas

Seperti yang diungkapkan pada bagian sebelumnya homoskedastisitas atau varians sama pada model regresi. Dengan kata lain asumsi homoskedastisitas menyatakan bahwa

$$\mathrm{E}(u_i^2) = \sigma^2, \quad i = 1, 2, \ldots, n. \tag{5.1}$$

Apabila terjadi heteroskedastisitas, asumsi model menjadi

$$\mathrm{E}(u_i^2) = \sigma_i^2 \tag{5.2}$$

yang secara garis besar menggambarkan bahwa varians tidaklah konstan lagi. Apakah heteroskedastisitas adalah hal yang wajar? Menurut Gujarati (2003) ada beberapa alasan kenapa varians u_i berubah-ubah.

1. Mengikuti *error-learning models*. Artinya, seiring orang belajar, kesalahan orang tersebut menjadi semakin kecil. Dalam hal ini σ_i^2 diharapkan menurun atau mengecil. Sebagai contoh kalau Anda belajar membaca, pada saat Anda masih kecil tentu berbeda dengan Anda sekarang. Kemampuan membaca yang dahulu "terbata-bata" sekarang sudah lancar.

2. Pertumbuhan. Misalnya, suatu perusahaan dengan profit atau keuntungan besar cenderung memiliki variabilitas yang lebih besar pada kebijakan tentang devidennya dibandingkan perusahaan dengan keuntungan yang lebih rendah.

3. Perbaikan dalam hal pengumpulan data. Seiring dengan semakin canggihnya peralatan pengumpulan data, melakukan kesalahan pun semakin kecil.

4. Pencilan. Heteroskedastisitas dapat terjadi karena adanya pencilan.

5. Mispesifikasi. Kesalahan dalam spesifikasi model dapat menyebabkan munculnya heteroskedastisitas. Hal ini dimungkikankan karena adanya penghilangan beberapa variabel.

6. Kepencongan. Salah satu sumber heteroskedastisitas adalah kepencongan (*skewness*) pada distribusi dari satu atau lebih peubah regresor.

7. Transformasi data yang tidak tepat dan bentuk fungsional data yang tidak tepat.

Gujarati (2003) menambahkan bahwa masalah heteroskedastisitas lebih sering terjadi pada data tampang lintang (*cross-section*) karena biasanya kita lebih sering berhubungan dengna anggota populasi pada titik tertentu, misalnya pengeluaran individu, perusahaan, subdivisi geografis, dan lain-lain. Biasanya data ini berukuran yang berbeda-beda, misalnya kecil, menengah, besar atau pendapatan rendah, medium, atau tinggi. Tetapi, tidak demikian halnya dengan data deret waktu.

5.2 Estimasi Kuadrat Terkecil pada Kasus Heteroskedastisitas

Perhatikan regresi linear sederhana berbentuk

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{5.3}$$

dan kita telah memperoleh bahwa

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n \sum_{i=1}^n Y_i}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2}. \tag{5.4}$$

Namun, variansnya menjadi

$$\operatorname{var}(\hat{\beta}_1) = \frac{\sum_{i=1}^n X_i^2 \sigma_i^2}{(\sum_{i=1}^n X_i^2)^2} \tag{5.5}$$

yang tentu saja berbeda dengan varians dari asumsi homoskedastisitas yaitu

$$\operatorname{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n X_i^2}. \tag{5.6}$$

Catatan: pada kasus heteroskedastisitas estimasi terhadap β_1 tetap bersifat tak bias, tetapi variansnya tidak lagi minimum dan tidak lagi terbaik.

5.3 Metode Kuadrat Terkecil Rampat

Penduga kuadrat terkecil tidak mampu menggunakan informasi tentang variabilitas pada variabel acak Y. Oleh karena itu, digunakan metode lain yang disebut metode kuadrat terkecil rampat (*generalized least squares*, disingkat GLS). Lihat kembali model regresi linear berbentuk

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{5.7}$$

dan apabila dimisalkan $X_{0i} = 1$ untuk setiap i, maka model ini dapat ditulis sebagai

$$Y_i = \beta_0 X_{0i} + \beta_1 X_i + u_i. \tag{5.8}$$

Apabila varians heteroskedastik σ_i^2 diketahui. Apabila persamaan (5.8) dibagi dengan σ_i akan diperoleh

$$\frac{Y_i}{\sigma_i} = \beta_0 \frac{X_{0i}}{\sigma_i} + \beta_1 \frac{X_i}{\sigma_i} + \frac{u_i}{\sigma_i} \tag{5.9}$$

yang selanjutnya dapat dituliskan dalam bentuk

$$Y_i^* = \beta_0^* X_{0i}^* + \beta_1^* X_i^* + u_i^*. \tag{5.10}$$

Kemudian kita peroleh $\mathrm{var}(u_i^*) = 1$ (Cek!). Ini artinya, varians dari galat yang ditransformasi u_i^* menjadi homoskedastik. Estimasi β_0^* dan β_1^* sekarang bersifat BLUE. Dengan kata penduga kuadrat terkecil rampat (*generalized least squares*, disingkat GLS) adalah penduga kuadrat terkecil pada variabel yang tertransformasikan yang memenuhi sifat-sifat kuadrat terkecil.

Untuk mendapatkan penduga GLS kita meminimumkan

$$\sum_{i=1}^{n} \hat{u}_i^{2*} = \sum (Y_i^* - \hat{\beta}_0^* X_{0i}^* - \hat{\beta}_1^* X_i^*)^2 \tag{5.11}$$

atau

$$\sum_{i=1}^{n} \left(\frac{\hat{u}}{\sigma_i}\right)^2 = \sum_{i=1}^{n} \left[\left(\frac{Y_i}{\sigma_i} - \hat{\beta}_0^* \left(\frac{X_{0i}}{\sigma_i}\right) - \hat{\beta}_1^* \frac{X_i}{\sigma_i}\right)\right]^2. \tag{5.12}$$

Menggunakan teknik yang sudah dipelajari dalam penurunan kuadrat terkecil diperoleh

$$\hat{\beta}_1^* = \frac{\sum_{i=1}^n w_i \sum_{i=1}^n w_i X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i \sum_{i=1}^n X_i^2 - (\sum_{i=1}^n w_i X_i)^2} \tag{5.13}$$

dan

$$\hat{\beta}_0^* = \bar{Y}^* - \hat{\beta}_1^* \bar{X}^*. \tag{5.14}$$

Selanjutnya diperoleh

$$\mathrm{var}(\hat{\beta}_1^*) = \frac{\sum_{i=1}^n w_i}{\sum_{i=1}^n w_i \sum_{i=1}^n w_i X_i^2 - (\sum_{i=1}^n w_i X_i)^2}. \tag{5.15}$$

5.4 Mendeteksi Heteroskedastisitas

Secara garis besar metode untuk mendeteksi heteroskedastisitas yaitu metode informal melalui metode grafis dan metode formal melalui prosedur pengujian statistika.

5.4.1 Metode grafis

Menurut Gujarati (2003) jika tidak ada informasi *a priori* atau empiris tentang heteroskedastisitas dalam praktik kita dapa melakukan analisis regresi pada asumsi bahwa tidak terdapat heteroskedastisitas dan melakukan pemeriksaan terhadap sisaan kuadrat $\hat{u}_i^2$ untuk melihat lebih lanjut apakah ada pola sistematis. Ingat bahwa $\hat{u}_i^2$ tidak sama dengan u_i^2, namun ini dapat digunakan jika ukuran sampel cukup besar. Plot $\hat{u}_i^2$ melawan $\hat{Y}_i$, pada kasus dua peubah akan memberikan informasi yang sama dengan plot pada $\hat{u}_i^2$ melawan X. Pada kasus regresi lebih dari dua peubah, $\hat{u}_i^2$ bisa diplot melawan sebarang peubah X yang diikutkan dalam model (Gujarati, 2003).

5.4.2 Metode formal

Uji Park

Bentuk fungsional uji Park adalah

$$\sigma_i^2 = \sigma^2 X_i^{\beta} \mathrm{e}^{v_i} \tag{5.16}$$

atau

$$\ln \sigma_i^2 = \ln \sigma^2 + \beta \ln X_i + v_i \tag{5.17}$$

dengan v_i adalah galat stokastik. Karena σ_i^2 biasanya tidak diketahui, Park menyarankan $\hat{u}_i^2$ sebagai *proxy* dengan meregresikan

$$\ln \hat{u}_i^2 = \ln \sigma^2 + \beta \ln X_i + v_i = \alpha + \beta \ln X_i + v_i. \tag{5.18}$$

Jika β signifikan, ini berarti heteroskedastisitas ada dalam data. Goldfeld dan Quant mengatakan bahwa uji Park memiliki masalah yaitu galat v_i yang masuk ke dalam model mungkin tidak memenuhi asumsi kuadrat terkecil.

Uji Glejser

Ide uji Glejser hampir sama dengan uji Park. Perbedaannya adalah Glejser mengusulkan untuk meregresikan nilai mutlak $\hat{u}_i$ dengan peubah X dianggap berasosiasi dengan σ^2. Beberapa bentuk fungsional uji Glejser adalah sebagai berikut

1. $|\hat{u}_i| = \beta_0 + \beta_1 X_i + v_i$,
2. $|\hat{u}_i| = \beta_0 + \beta_1 \sqrt{X_i} + v_i$,
3. $|\hat{u}_i| = \beta_0 + \beta_1 / X_i + v_i$,
4. $|\hat{u}_i| = \beta_0 + \beta_1 / \sqrt{X_i} + v_i$,
5. $|\hat{u}_i| = \sqrt{\beta_0 + \beta_1 X_i} + v_i$,
6. $|\hat{u}_i| = \sqrt{\beta_0 + \beta_1 X_i^2} + v_i$.

Goldfeld dan Quandt menekankan bahwa galat v_i memiliki nilai harapan taknol, dan memiliki korelasi serial, dan sayangnya heteroskedastik. Kemudian, model 5 dan 6 adalah mondel nonlinear dalam parameter, sehingga tidak bisa diestimasi dari prosedur kuadrat terkecil biasa.

Uji korelasi rank Spearman

Koefisien korelasi rank Spearman didefinisiksan sebagai

$$r_s = 1 - 6\left[\frac{\sum d_i^2}{n(n^2-1)}\right] \tag{5.19}$$

dengan d_i menyatakan selisih pada rank yang diberikan atau ditugaskan pada dua karakteristik yang berbeda pada individu ke-i dan n adalah banyak individu yang diranking. Langkah-langkah pengujian ini adalah sebagai berikut:

1. Lakukan regresi, misalnya $Y_i = \beta_0 + \beta_1 X_i + u_i$ dan hitung sisaan $\hat{u}_i$.
2. Abaikan tanda $\hat{u}_i$, yaitu ambil nilai mutlak $|\hat{u}_i|$. Kemudian ranking $|\hat{u}_i|$ dan X_i (atau $\hat{Y}_i$) baik diurut naik maupun turun. Kemudian hitung korelasi rank Spearman r_s.
3. Asumsikan koefisien korelasi rank populasi ρ_s adalah nol dan $n > 8$, signifikansi sampel r_s dapat diuji menggunakan uji t sebagai berikut:
$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \tag{5.20}$$
dengan derajat bebas $n-2$.
4. Jika nilai t hitung melebihi nilai kritis t, kita tidak akan menerima hipotesis heteroskedastisitas. Jika model regresi memiliki lebih dari satu variabel X, r_s dapat dihitung antara $|\hat{u}_i|$ dan masing-masing peubah X secara terpisah dan diuji signifikansinya.

Uji Goldfeld-Quandt

Uji ini mengasumsikan bahwa varians heteroskedastik σ_i^2 berhubungan positif dengan slah satu dari peubah regresor dalam model. Sebagai contoh σ_i^2 berhubungan secara positif dengan X_i dalam bentuk

$$\sigma_i^2 = \sigma^2 X_i^2 \tag{5.21}$$

dengan σ^2 konstan. Prosedur pengujian Goldfeld-Quandt adalah sebagai berikut:

1. Urut atau ranking amatan berdasarkan nilai X_i, mulai dari nilai X_i terkecil.

2. Hilangkan c amatan di tengah-tengah. Nilai c ditentukan terlebih dahulu, dan bagi sisa amatan $(n - c)$ menjadi dua kelompok $(n - c)/2$ amatan.

3. Lakukan regresi terhadap $(n - c)/2$ amatan pertama dan sisa $(n - c)/2$ amatan sisanya. Kemudian hitung jumlah kuadrat sisaan, katakanlah JKR_1 untuk amatan X_i yang lebih kecil dan JKR_2 untuk amatan X_i yang lebih besar. Masing-masing amatan akan memiliki derajat bebas db $= (n-c)/2-k$ dengan k adalah banyak parameter yang akan diestimasi, termasuk intersep.

4. Hitung rasio
$$\lambda = \frac{\text{JKR}_2/\text{db}}{\text{JKR}_1/\text{db}}. \tag{5.22}$$

Salah satu pedoman pemilihan nilai c adalah sebagai berikut: jika $n = 30$, $c = 4$; jika $n = 60$, $c = 10$. Jika nilai λ lebih besar dari nilai kritis F pada tingkat signifikansi tertentu yang dipilih, kita dapat menolak hipotesis homoskedastisitas.

Uji Breusch-Pagan-Godfrey

Keberhasilan uji Goldfeld-Quandt bergantung tidak hanya pada nilai c, namun pada mengidentifikasi nilai X yang tepat. Untuk mengatasi hal ini uji Breusch-Pagan-Godfrey (BPG) dapat digunakan.

Misalkan model regresi linear berganda

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + u_i. \tag{5.23}$$

Asumsikan varians galat σ_i^2 dinyatakan sebagai

$$\sigma_i^2 = f(\alpha_0 + \alpha_1 Z_{1i} + \cdots + \alpha_m Z_{mi}). \tag{5.24}$$

Dengan kata lain, σ_i^2 adalah fungsi dari variabel nonstokastik Z; satu atau beberapa X dapat menjadi Z. Diasumsikan bahwa

$$\sigma_i^2 = \alpha_0 + \alpha_1 Z_{1i} + \cdots + \alpha_m Z_{mi} \tag{5.25}$$

yang merupakan fungsi linear dari Z. Jika σ_i^2 adalah homoskedastik, kita dapat menguji bahwa $\alpha_0 = \alpha_1 = \cdots = \alpha_m = 0$.

Langkah-langkah pengujian BPG adalah sebagai berikut:

1. Lakukan regresi dan hitung sisaan.

2. Hitung $\tilde{\sigma}^2 = \sum \hat{u}_i^2/n$.

3. Buat variabel $p_i = \hat{u}_i^2/\tilde{\sigma}^2$.

4. Regresikan
$$p_i = \alpha_0 + \alpha_1 X_{1i} + \cdots + \alpha_m Z_{mi} + v_i \tag{5.26}$$

5. Hitung jumlah kuadrat regresi dan definisikan $\Theta = \text{JKR}/2$. Jika Θ melebihi nilai kritis χ^2_{m-1}, kita dapat menolak hipotesis homoskedastisitas.

Uji White

Tidak seperti uji Goldfeld-Quandt yang memerlukan pengurutan amatan yang bersesuaian dengan variabel X yang diduga menyebabkan heteroskedastisitas atau uji BPG yang memerlukan asumsi kenormalan. Uji White tidak memerlukan asumsi kenormalan. Misalkan model regresi berbentuk
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i. \tag{5.27}$$
Prosedur uji White adalah sebagai berikut:

1. Hitung sisaan dari model (5.27).

2. Lakukan regresi berikut
$$\hat{u}_i^2 = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{1i}^2 + \alpha_4 X_{2i}^2 + \alpha_5 X_{1i} X_{2i} + v_i. \tag{5.28}$$
Catatan kuadrat sisaan pada regresi awal diregresikan dengan X, kuadrat, dan perkalian silang dari regresor. Regresor dengan tingkat yang lebih tinggi juga dapat digunakan. Intersep dapat dimasukkan ke dalam model meskipun, misalnya, regresi awal tidak berisi intersep. Hitung R^2 dari regresi ini.

3. Hipotesis nol uji in adalah tidak terdapat heteroskedastisitas. Secara asimtotik
$$nR^2 \sim \chi^2 \tag{5.29}$$
dengan derajat bebas sejumlah banyak regresor (termasuk intersep). Jika nilai khi-kuadrat yang diperoleh melewati nilai kritis kita simpulkan terdapat heteroskedastik.

Uji Koenker-Bassett

Diasumsikan model awal
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} u_i. \tag{5.30}$$
Kemudian lakukan regresi
$$\hat{u}_i^2 = \alpha_0 + \alpha_1 \hat{Y}_i^2 + v_i. \tag{5.31}$$
Hipotesis nolnya adalah $\alpha_1 = 0$. Jika ini tidak ditolak disimpulkan tidak terdapat heteroskedastik.

5.5 Ukuran-ukuran Pemulihan

Ukuran-ukuran pemulihan (*remedial measures*) meliputi kasus pada saat σ_i^2 diketahui dan σ_i^2 tidak diketahui.

5.5.1 Kasus I: σ_i^2 diketahui

Pada saat σ_i^2 diketahui, metode yang paling jelas bisa digunakan adalah metode GLS.

5.5.2 Kasus II: σ_i^2 tidak diketahui

Biasanya σ_i^2 tidak diketahui ada beberapa cara. Misalnya menggunakan galat baku dan varians heteroskedastisitas-konsisten White. Sebagai contoh pada kasus regresi linear biasa White menyarankan penduga

$$\text{var}(\hat{\beta}_1) = \frac{\sum (X_i - \bar{X})^2 \hat{u}_i^2}{(\sum (X_i - \bar{X})^2)^2} \tag{5.32}$$

Beberapa asumsi pola heteroskedastisitas yang mungkin:

1. varians galat proporsional dengan X_i^2, artinya $\mathrm{E}(u_i^2) = \sigma^2 X_i^2$,
2. varians galat proporsionald dengan X_i, artinya $\mathrm{E}(u_i^2) = \sigma^2 X_i$,
3. varians galat proporsional dengan nilai tengah kuadrat Y, atau $\mathrm{E}(u_i^2) = \sigma^2 (\mathrm{E}(Y_i))^2$,
4. transformasi log seperti $\ln Y_i = \beta_0 + \beta_1 X_i + u_i$ sering digunakan untuk mengurangi heteroskedastisitas saat dibandingkan regresi linear sederhana.

BAB VI

Model Regresi Data Panel

Pada bab-bab sebelumnya kita telah membicarakan model regresi untuk data tampang lintang (*cross section*). Pada bab ini kita akan membahas data panel, yang merupakan gabungan dari data tampang lintang dan data deret waktu. Nama lain data panel antara lain data gabungan (*pooled data*, yaitu gabungan amatan deret waktu dan tampang lintang), kombinasi data deret waktu dan tampang lintang (*combination of time series and cross-section data*), data longitudinal, data mikropanel (*micropanel data*), analisis sejarah kejadian (*event history analysis*), dan analisis kohor (*cohort analysis*) (Gujarati, 2003).

Peter Kennedy dalam Hill *et al.* (2012) membagi data panel menjadi tiga jenis berbeda yaitu

1. ”panjang dan sempit” (*long and narrow*), dengan panjang mendeskripsikan dimensi waktu dan sempit mengimplikasikan unit tampang lintang yang relatif cukup kecil;

2. ”pendek dan lebar” (*short and wide*), mengindikasikan banyak individu yang diamati pada waktu yang relatif singkat;

3. ”panjang dan lebar” (*long and wide*), mengindikasikan jumlah individu yang banyak dan waktu yang panjang.

6.1 Materi

Detail materi tentang analisis regresi data panel dapat dibaca pada materi yang diambil dari Gujarati.

Daftar Pustaka

Gujarati, D. N. 2003. *Basic Econometrics*. Fourth edition. McGraw-Hill, Boston.

Hill, R. C., Griffiths, W. E., and Lim, G. C. 2012. *Principles of Econometrics*. Fourth edition. John Wiley and Sons, Asia.

Judge, G. G., Hill, R. C., Griffiths, W., Lütkepohl, H., and Lee, T.-C. 1982. *Introduction to the Theory and Practice of Econometrics*. John Wiley and Sons, New York.

Stock, J. H. and Watson, M. M. 2012. *Introduction to Econometrics*. Pearson Education Limited, England.

Wooldridge, J. M. 2006. *Introductory Econometrics*. Third edition. Thomson South-Western, Australia.