Statistika, Vol. 7 No. 2, 39 – 45
Nopember 2007

Memanfaatkan Model Statistika Dasar: Awal Pembelajaran Statistika Dalam Penelitian

SATWIKO DARMESTO

Sekolah Tinggi Ilmu Statistik
Alamat: Otto Iskandardinata 64 C, Jakarta 13330
e-mail: satwiko@mailhost.bps.go.id
telpon/fax: (021) 851 8787 / (021) 819 7577

ABSTRAK

Pada era globalisasi sekarang ini, salah satu faktor terpenting yang menentukan daya saing sebuah bangsa adalah penguasaan ilmu pengetahuan dan teknologi (IPTEK). Saat ini semua bangsa sedang berpacu dalam penguasaan ilmu pengetahuan dan teknologi (apa pun), tidak hanya sekedar agar bisa survive, tetapi untuk menguasai pasar dengan teknologi yang selalu dikembangkannya. Penguasaan IPTEK dapat diperoleh melalui penelitian, pengembangan, dan penerapan teknologi secara terus menerus dan dilakukan oleh semua pihak yang mempunyai keiinginan untuk maju. Secara khusus, penguasaan IPTEK bangsa Indonesia harus didukung oleh semua pihak, negara dan masyarakat.

Hampir dalam setiap penelitian diperlukan uji satistika di dalam analisa data, sehingga hasil penelitian dapat dipahami oleh semua pihak dan diyakini dapat diterapkan. Statistika sebagai ilmu diterapkan di bidang ekonomi dan sosial untuk mengukur keadaan dan kondisi perekonomian dan pengaruh kondisi tersebut kepada masyarakat. Namun demikian, di bidang teknik pun statistika dipakai untuk mengukur ketahanan/keakuratan pemakaian material pembuatan pesawat terbang, menguji baku mutu produk, atau menguji kemiringan/sudut sayap ekor pesawat terbang apakah cukup reliable untuk diterapkan.

Penguasaan metodologi statistika merupakan kunci keberhasilan di dalam penelitian. Oleh karena itu, penbelajaran ilmu statistika sangat diperlukan oleh peneliti atau siapa pun yang akan melakukan penelitian dan masyarakat luas lainnya.

1. Menikmati Hasil Penelitian

Banyak hasil penelitian sudah dinikmati oleh nasyarakat luas dalam kehidupan sehari-hari. Apabila kita melewati salah satu pintu toll di Jakarta atau Bandung, akan mengingatkan kita pada teori antrian (scheduling). Pengendara biasanya memasuki pintu toll yang disediakan oleh PT. Jasa Marga, sebagai pengelola jalan toll, secara serial. Akan tetapi karena antrian mobil yang terlalu panjang akan mengganggu akses jalan arteri, maka antrian di pintu toll dibuat kombinasi antara serial dan parallel. Dua mobil akan masuk secara serentak dan akan dapat dilayani oleh dua petugas karcis toll. Demikian juga di Statsiun Pengisian Bahan Bakar untuk Umum (SPBU). Desain pompa bensin dibuat secara serial dan parallel oleh pabrikan pompa bensin. Pengendara mobil yang akan mengisi bensin seharusnya mengikuti desain pompa bensin tersebut sehingga (satu) petugas dapat bekerja optimal, antrian tidak kelihatan terlalu panjang, dan pemasukan uang lebih cepat bagi pemilik SPBU.

Statistical Process Control (SPC) merupakan bagian tak terpisahkan dalam menangani proses produksi untuk mencapai keseragaman produk (standard). Dalam proses produksi penyelia akan menghitung rata-rata keseragaman produk dan diletakkan pada center line (CL), kemudian menghitung standar deviasi dari variasi produk, dan menentukan upper control limit (UCL) serta lower control limit (LCL).

Estimasi penduduk Indonesia tahun 2005 dan 2010 dilakukan dengan rumus pertumbuhan penduduk dengan memasukan data kelahiran, kematian, dan perpindahan penduduk. Formula Modified Laspeyers Indices dan data harga komoditi berbagai jenis barang konsumsi dan jasa digunakan untuk menghitung angka inflasi yang nialinya diterbitkan setiap bulan. Angka inflasi selalu diperbandingkan dari satu waktu ke waktu untuk berbagi analisis ekonomi. Estimasi APBN dari besaran-besaran niali per sector beserta asumsi-asumsi harga minyak, nilai dolar, tingginya inflasi dilakukan setiap tahun. Simulasi besaran PDRB

(Product Domestic Regional Bruto) dilakukan dengan melibatkan nilai impor, ekspor, konsumsi pemerintah, investasi, dan konsumsi.

Menghadapi data yang bersifat time series, beberapa metoda dapat digunakan untuk melihat pengaruh variable-variable yang saling berpengaruh atau tidak saling berpengaruh terhadap independent mau pun dependent variable. Bahkan mungkin dapat digunakan untuk memprediksi pengaruh satu (atau beberapa) independent variable terhadap dependent variable. Untuk data yang bersifat time series dan cross section (antar beberapa sector, kegiatan, atau perusahaan) biasanya menggunaka analisis data panel.

2. Data Statistik

Statistik secara populer sering diartikan sebagai data atau hasil hitungan berdasarkan data (Djauhari, 2007). Data statistik merupakan data/fakta hasil pengamatan suatu fenomena atau karakteristik tertentu pada suatu lokasi dan kurun waktu tertentu. Data statistik dapat diperoleh dari catatan secara periodik atau pun melalui suatu survey. Berbagai sumber data statistik seperti Dinas di Pemerintah Daerah, Lembaga Riset seperti LIPI dan BPS atau A.C. Nielsen menghasilkan data statistik yang mungkin berguna bagi para periset dalam membuktikan dan memaknai fenomena apa yang sudah, sedang, dan akan terjadi..

Diperlukan ketrampilan dan kesungguhan dalam mencari/memperoleh, menyimpan, dan menggunakan data di unit instansi masing-masing agar data statistik dapat digunakan secara bersama dan bertanggung jawab. Ada kalanya suatu data tentang penduduk berbeda antar instansi atau unit. Oleh karena itu diperlukan kejelian dalam melihat apakah data tersebut relevan dengan penelitian yang sedang kita lakukan. Kejelian melihat apakah kurun waktu data sama antar unit atau instansi tersebut. Demikian pula kejelian dalam melihat definisi atau metoda yang digunakan oleh unit atau instansi tadi. Sedikit saja perbedaan dalam beberapa hal tadi akan menyebabkan data juga berbeda

3. Regresi dan Korelasi

Simple Regresi

Di dalam analisis regresi dan korelasi kita dapat menentukan nature dan kekuatan hubungan dari dua variable, variable independent dan variable dependent. Persamaan simple regression dituliskan sebagai:

$$Y = a + bX$$

dimana Y sebagai dependent variable dan X sebagai independent variable.

Multiple Regresi

Di dalam praktek penelitian dosen, guru, dan mahasiswa, multiple regresi lebih banyak dijumpai dan dipergunakan dalam memecahkan persoalan satu dependent variable dipengaruhi oleh banyak independent variable, sehingga persamaan multiple regression dituliskan sebagai:

$$Y = a + b_1X_1 + b_2X_2 + \ldots\ldots + b_kX_k$$

dimana Y sebagai dependent variable dan X_i sebagai independent variable ke i.

Korelasi

Hubungan (association) antara dependent variable dan independent variable(s) dinyatakan dalam rumus r^2 (r-Squared) yang menyatakan seberapa kuat (besar atau kecil) dependent variable Y dipengaruhi oleh besar kecilnya nilai independent variable(s) X.

Variant dari Regresi: Dummy Variable

Ada kalanya satu atau beberapa variable X merupakan data kualitatif atau dummy (mempunyai nilai 0 atau 1) yang harus diperhitungkan di dalam persamaan. Misal, di dalam persamaan $Y = 3526.4 + 722.5X_1 + 90.02X_2 + 1.2690X_3 + 23.406X_4$. Nilai X_1 MALE, yang merupakan indicator variable diberi kode 1 untuk laki-laki dan diberi kode 0 untuk perempuan, X_2 berupa lamanya pendidikan dalam tahun (EDUC), X_3 adalah lamanya pengalaman dalam bulan (EXPR), dan X_4 jumlah bulan setelah 1 Januari 2004 (TIME). Senua variable ini (termasuk gender MALE tadi) diperhitungkan untuk menghitung upah dasar pegawai Y (SALARY) di suatu perusahaan. Bentuk persamaan garis regresi dengan

memperhitungkan dummy variable (MALE) dalam menentukan upah dasar pegawai adalah sebagai berikut:

SALARY = 3526.4 + 722.5 MALE + 90.02 EDUC + 1.2690 EXPR + 23.406 TIME.

Apakah dari persamaan di atas dapat diketahui bahwa ada diskriminasi antara laki-laki dan perempuan dalam menentukan upah dasar pegawai?

Atau satu contoh lain, banyaknya bensin yang dipakai sebuah kendaraan/mobil dipengaruhi oleh berat kendaraan dan jenis transmisi mobil tersebut, $Y = -2.925 + 70112X_1 + 0.0041\ X_2$ dimana Y adalah CITYMPG, X_1 adalah WEIGHT kendaraan, dan X_2 adalah AUTO (indicator variable, diberi kode 1 untuk mobil automatic dan kode 0 untuk mobil dengan manual transmission). Persamaan regresi yang menyertakan dummy variable untuk mobil automatic atau manual adalah sebagai berikut:

CITYMPG = -2.925 + 70112 WEIGHT + 0.0041 AUTO

Variant dari Regresi: Interaction Variables

Salah satu tipe variable yang dipakai di dalam regresi adalah interaction variable yang dapat dibentuk dengan mengalikan nilai dua variable independent X_1 dan X_2. Dengan terjadinya interaksi ini, akan memberikan efek ke dalam persamaan yang dibuat. Persamaan regresi biasa adalah

$Y = a + b_1X_1 + b_2X_2.$

Namun, apabila kita masukkan interaction variable antara X_1 dan X_2, maka persamaan menjadi

$Y = a + b_1X_1 + b_2X_2 + b_3X_1.X_2.$ Atau secara matematis $Y = a + (b_1 + b_3X_2)X_1 + b_2X_2$

Sebagai contoh adalah kemenangan pemain bulutangkis Susi Susanti dipengaruhi oleh kekuatan lob dan smash dari Susi Susanti, namun juga oleh kombinasi lob dan smash dari Susi Susanti, interaksi keduanya INTERACT = LOB*SMASH). Dengan demikian persamaan regresi menjadi

WINS = -19.3 + 0.00626 LOB + 1.00 SMASH – 0.000210 INTERACT

Model di bawah ini merupakan pengembangan dari model SALARY di atas dengan menambahkan interaksi antara EDUC dan EXPR, sehingga persamaan menjadi

SALARY=3006 + 688 MALE + 138 EDUC + 5.68 EXPR + 22.4 TIME – 0.364 EDUCEXPR

4. Regresi Menggunakan Time-Series Data

Tujuan dari penggunaan time series data adalah dapat melakukan prediksi/ forecast atas dependent variable untuk waktu ke depan. Para peneliti dapat menggunakan salah satu dari dua regression model: causal regression model atau extrapolative regression model. Extrapolative model menggunakan explanatory variables, mendiskripsikan perkembangan masa lalu sehingga dapat diperhitungkan di masa depan.

Lag Variable

Pada waktu menggunakan time-series data, ada kalanya menghubungkan nilai dependent variable pada waktu ini dengan nilai explanatory variable pada waktu yang sama. Misal, Penjualan bulan ini dihubungkan denga biaya iklan bulan ini. Namun, efek iklan yang dipasang bulan lalu mungkin baru dirasakan pada bulan ini, efek pemasangan iklan bulan ini akan dirasakan pada bulan depan, dan seterusnya. Dengan demikian model persamaan dengan time lag (jeda waktu) tersebut dapat dituliskan sebagi berikut:

$Y = a + b_1X_t + b_2X_{t-1} + b_3X_{t-2}$

Dalam model ini sales (Y) sebagai fungsi dari biaya iklan pada bulan ini (X_{t-1}) dan dua bulan sebelumnya (X_{t-2}).

Trend Dalam Time-Series Regression

Trend dala time-series data adalah tendensi untuk bergerak naik atau turun dalam kurun waktu tertentu. Pergerakan ini mungkin membentuk kurve lurus (straight line) atau kurva melengkung (curvelinear pattern) Analisis regresi dapat dipakai untuk memodelkan trend tertentu dan mengextrapolate trend ini untuk estimasi yang akan datang.

Forecast sederhana menggunakan simple regressi dengan mengganti notasi X dengan notasi T (Time period)

$$Y_{T+1} = a + b_1 (T+1)$$

Persamaan Kuadrat (quadratic) trend

$$Y_t = a + b_1 t + b_2 t^2$$

Atau persamaan trend dengan kurva S

$$Y_t = \exp (a + b_1 (1/t)) \text{ dimana exp adalah nilai 2.7}$$

Model ini dapat digunakan untuk membuat model demand suatu produk sepanjang hidup produk tersebut. Pada awalnya demang sedikit sampai produk tadi dikenal. Kemudian demand menanjak sampai puncaknya dan kemudian menurun. Namun persamaan ini tidak dapat digunakan untuk melakukan estimasi. Perlu diubah dengan membuat logaritma di kedua sisi persamaan sehingga menjadi

$$\ln(Y_t) = a + b_1(1/t) \text{ : Bila } Y'_t = \ln(y_t) \text{ dan } t' = 1/t \text{ maka } Y'_t = a + b_1 t'$$

mengganti t = T +1, maka persamaan untuk forecasting adalah

$$Y'_{t+1} = a + b_1 (1/T+1)$$

Ini adalah forecast dari Y'_{T+1} atau logaritma natural dari Y_{T+1}

5. Multivariate Analysis

Dalam multivariate, seluruh variable harus random dan berhubungan sehingga efek dari seluruh variable tidak dapat diinterpretasikan sendiri-sendiri, tetapi secara bersama-sama. Multivariate digunakan untuk mengukur, menjelaskan, dan mem-predict derajad hubungan antar variates (weighted combinations of variables). Multivariate analysis adalah pengembangan beberapa teknik dalam menganalisis data. Multivariate teknik dibagi menjadi beberapa tipe:

Multiple Regression

Metoda ini sangat cocok untuk analisis bila masalah riset melibatkan single metric dependent variable dihubungkan dengan satu atau lebih metric independent variables. Tujuan dari penggunaan analisis multiple regresi adalah mem-predict perubahan di dalam dependent variable sesuai dengan perubahan di dalam beberapa independent variables. Bila periset ingin melakukan prediksi nilai dari dependent variable, maka metoda ini sangat berguna. Misal: biaya makan di luar rumah (dependent) sangat dipengaruhi oleh informasi mengenai family income, family size, umur kepala keluarga (independent).

Beberapa contoh lain: Company sales di prediksi dengan informasi pengeluaran untuk iklan, jumlah salespeople, dan jumlah toko yang membawa produk tersebut. Jumlah penumpang pesawat dipengaruhi oleh harga tiket pesawat, jenis pesawat, jadwal penerbangan. Penggunaan produk PT. ABX dipengaruhi oleh: delivery speed, price level, price flexibility, manufacture's image, service, sales force image, product quality.

Stepwise estimation dari PT. ABX (missal dengan 7 variable independent) mendapatkan persamaan multiple regresi (hanya dengan variable $X_3 + X_5 + X_6$):

$$Y = -6.520 + 3.376X_3 + 7.621X_5 + 1.406X_6$$

Multiple Discriminat Analysis

Bila single dependent variable dikotomus (misal: male-female) atau multikotomus (misal: tinggi-sedang-rendah) dan nonmetric, maka metoda multivariate yang cocok untuk persoalan tersebut adalah Multiple Discriminant Analysis (MDA). Dalam hal ini independent variable adalah metric.

MDA sangat berguna pada situasi dimana seluruh sample dapat dibagi menjadi beberapa group berdasarkan kelas yang sudah diketahui dalam dependent variables. Tujuan utama dari MDA adalah mengerti perbedaan-perbedaan group dan mem-predict kecenderungan anggota (individual atau object) akan menjadi anggota group atau kelas berdasar beberapa metric dari independent variables. Sebagai contoh: MDA dapat digunakan untuk membedakan innovators dari noninnovators sesuai dengan demographic dan psychographic profile. Atau membedakan perokok berat dari perokok ringan, laki-laki dari

perempuan, pembeli merk terkenal dari pembeli merk local, Frequent flyers dari nonfrequent flyers, good credit risk dari poor credit risk.

$Y = X_1 + X_2 + X_3 + \ldots\ldots\ldots + X_n$

(nonmetric) (metric)
(categorical dependent variable)

$Z = W_1X_1 + W_2X_2 + \ldots\ldots\ldots\ldots + W_nX_n$

dimana Z = discriminant score
Wi = discriminant weight for variable i
Xi = independent variable i

Factor Analysis

Factor analysis termasuk variasinya seperti component analysis dan common factor analysis, adalah pendekatan secara statistic yang dapat digunakan untuk menganalisa interrelationship (saling keterhubungan) di antara sejumlah besar variables dan untuk menjelaskan variables tadi dalam faktor yang umum. Tujuan FA adalah memperkecil jumlah informasi dari original faktor menjadi lebih sedikit factor, tetapi dengan sesedikit mungkin kehilangan informasi.

Apabila ada 7 attribute X pada multiple regression, apakah 7 attribute tadi dapat dikelompokkan (group) dan dengan demikian jumlah factor akan berkurang. Misal X_1, X_2, X_5 menjadi kelompok 1 dan X_3, X_4, X_6, dan X_7 sebagai kelompok 2.

Cluster Analysis

Cluster analysis adalah nama dari multivariate teknik yang bertujuan membentuk kelompok (group) berdasar karakteristik yang dipunyainya. Cluster analysis mengklasifikasi object: respondent, produk, entitas lain, sehingga setiap object sama atau hamper sama dengan yang lain di dalam satu cluster sesuai dengan criteria yang telah ditetapkan. Dengan demikian secara internal ada homogenitas dan secara eksternal terjadi heterogenitas, sehingga apabila kita melakukan plotting secara geometric maka objek yang sama akan saling berdekatan dan yang tidak sama akan menjauh.

6. Data Panel

Data dalam analisis ekonometrika mungkin berupa data time series, data cross section, atau data panel. Data panel merupakan gabungan data time series dan data cross section. Data panel merupakan data yang memuat unit-unit individu yang sama yang diamati dalam jangka waktu tertentu. Data panel ditengarai oleh T, periode waktu (t = 1,2,3,...., T) yang kecil/sedikit dan N, jumlah individu (I = 1,2,3,.....,N) yang besar/banyak. Dapat pula terjadi sebaliknya, data panel terdiri dari periode waktu yang besar/banyak dan jumlah individu yang kecil/sedikit. Regresi menggunakan data panel disebut sebagai model regresi data panel.

Asumsi model regresi klasik tidak dapat digunakan dalam model data panel karena bertambahnya gangguan menjadi: gangguan antar waktu (time series related disturbance), gangguan antar individu (cross section disturbance), dan gangguan antar waktu dan antar individu.

Dengan analisis data panel dapat diungkap perilaku individu yang berbeda selama jangka waktu tertentu untuk memperoleh parameter estimasi. Model regresi data panel yang memuat efek spesifik individu dapat dituliskan sebagai berikut:

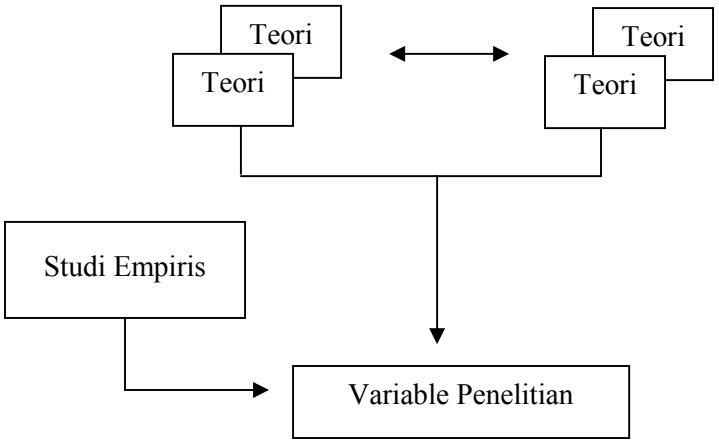
$y_{it} = \alpha_i + \beta' x_{it} + \varepsilon_{it}$

Pada persamaan di atas, y_{it} merupakan nilai variable tak bebas dan x_{it} adalah variable bebas untuk setiap individu i pada periode t (i = 1,2,3,..,N dan t = 1,2,3,....,T. α_i merupakan efek individu yang dapat bernilai constant selama periode t atau mungkin berbeda untuk setiap individu ke i. Pada x_{it} terdapat K slope yang menunjukkan jumlah variable bebas yang digunakan dalam model. Balanced panel merupakan data panel yang memiliki jumlah observasi yang sama untuk setiap unit individunya, sehingga total observasi adalah sebesar N x T.

7. Variable Penelitian Dalam Model

Dari beberapa uraian di atas tentu harus disesuaikan dengan data dan kondisi riset yang akan dijalankan di bidang tertentu. Permasalahan harus dicari, data harus dikumpulkan, dan pisau (alat) yang tepat harus digunakan. Teori dipelajari dan studi empiris diobservasi. Periset harus terampil dalam mengumpulkan variable-variable yang "dicurigai" akan terkena pengaruh dari variable lain atau memberikan pengaruh kepada variable lain.

Pada umumnya dari banyak teori akan muncul variable yang mungkin relevan dan akan digunakan dalam penelitian. Tidak tertutup kemungkinan variable juga muncul dari studi empiris yang pernah dilakukan para peneliti lain. Setelah melalui kajian-kajian terhadap variable tersebut maka dipilihlah variable yang akan masuk ke dalam penelitian. Dalam hal ini dari kedua sumber inilah yang mungkin akan menghasilkan variable penelitian seperti yang kita kehendaki.



Beberapa contoh penggunaan variable pembentuk model (bidang yang menyangkut masalah transportasi): perkembangan jumlah penduduk ke usia siap kerja di Bodetabek, menyebabkan penambahan penumpang KA dari Bodetabek ke Jakarta (karena Jakarta menawarkan 1001 jenis pekerjaan). Apabila tidak ada penambahan gerbong KA pengangkut penumpang Bodetabek, maka penumpang akan naik di atas gerbong dan akan mudah menimbulkan kecelakaan.

Variable harga tiket, mudah mendapatkan tiket, banyaknya rute penerbangan, jumlah penerbangan, kemudahan/keinginan mobilitas masyarakat saat ini akan mempengaruhi jumlah penumpang yang akan terbang dari satu tempat ke tempat yang lain. Saat ini terjadi OpenAir antara Amerika Serikat dan European Union. Pesawat-pesawat Amerika akan bebas masuk ke negara-negara Eropa, demikian juga pesawat-pesawat Eropa akan bebas msuk daratan Amerika dengan membawa penumpang masing-masing. Variable apa yang akan terpengaruh dan variable apa yang mempengaruhi?

Variable mana yang mendorong munculnya variable lain? Pelabuhan atau fasilitas dibangun dengan harapan makin banyak masyarakat yang akan menggunakan fasilitas tadi (pull demand). TransJakarta dibangun dengan harapan semua masyarakat menggunakan bus TransJakarta sehingga jalan di Jakarta tidak macet. Ataukah masyarakat membutuhkan banyak rute dan bus TransJakarta dan mau meninggalkan kendaraan pribadinya (push supply) supaya Jakarta tidak macet?

Pendapatan masyarakat Jakarta (atau kota lain), biaya transportasi, kemudahan memperoleh layanan, perpencaran anggota keluarga, budaya arisan/bertemu/kumpul keluarga akan menentukan moda transportasi (sarana) apa yang dibutuhkan masyarakat di Jakarta. Data demografis dan psikografis akan memberikan gambaran yang lebih luas dalam memandang persoalan penentuan sarana transportasi.

Faktor apa saja yang mempengaruhi "kesemrawutan" pengendara mobil, motor,dan alat transportasi lain di jalan raya? Faktor pendidikan berlalu lintas, jumlah pengendara berumur muda, mudah memperoleh SIM, ujian SIM tanpa praktek, budaya semau gue, tilang tidak pernah dilakukan/ditegakkan, polisi menerima uang tilang, banyak kendaraan dan jenis

kendaraan, kurang/tidak ada angkutan umum missal, jalan terlalu sempit, belum ada pengaturan jalan secara optimal, lampu lalu lintas sering mati atau sudah tidak sesuai lagi, rambu lalu lintas tidak jelas, kondisi jalan banyak lubang, dll. Semua variable ini mungkin harus dikurangi sehingga membuat simple permasalahan untuk kemudian segera dapat dipecahkan.

Daftar Pustaka

Dielman, Terry E., Applied Regression Analysis for Business and Economics, Thomson Information/Publishing Group, Boston, 1991

Djauhari, A., Maman, Statistik: Salah satu Indikator Utama Peradaban, Institut Teknologi Bandung, Bandung, 2007

Eppen, Garry D., Gould, F.J., Schmidt, Charles P., Introductory Management Science, Prentice-Hall, New Jersey, 1993

Hair, Joseph F., Anderson, Rolph E., Tatham, Ronald L., Black, William C., Multivariate Data Analysis, Prentie Hall, New Jersey, 1995

Pyndick, Robert S. and Daniel L. Rubindfield, Econometric Models and Economic Forecast. McGraw-Hill International, New York, 1998