ESTIMASI DAN INFERENSI MODEL REGRESI SEMI-PARAMETRIK PROSES PRODUKSI

Tubagus Pamungkas,
Dosen Tetap Pendidikan Matematika FKIP UNRIKA Batam

ABSTRAK

Regresi berganda terdapat kasus khusus dalam sebuah analisa regresi, pada regresi berganda terdapat satu variabel tak bebas yang akan diprediksi, tetapi terdapat dua atau lebih variabel bebas, pemilihan model yang terbaik akan dilakukan dengan 3 metode, yaitu Metode Quadratic Mode Estimator (QME), Metode Symmetrically Trimmed Least Squares (STLS) dan Metode Left Truncated (LT). Data yang digunakan adalah data polusi udara yang disebabkan oleh 7 variabel bebas yang meliputi jumlah kendaraan yang melewati, suhu udara, kecepatan angin, perbedaan temperatur, angin, jam aktif dan hari aktif, Pemotongan atau penyensoran dari suatu variabel respon dalam suatu model regresi adalah salah satu masalah yang sering muncul dalam banyak aplikasi.

Berdasarkan uraian dari pembahasan dan simulasi, terdapat beberapa hal penting yang dapat disimpulkan, dalam pemilihan model terbaik regresi semi-parametrik diperoleh model terbaik adalah dengan menggunakan metode QME, hal tersebut dapat dilihat dari nilai RMSE terkecil. Dalam uji parsial terdapat 2 variabel yang signifikan terhadap variabel dependent yaitu variabel yang berupa cars dan wind.speed dan juga konstanta yang berupa intercept. Dalam diagnostic checking dapat disimpulkan uji kenormalan menggunakan Kolmogorov Smirnov Test ternyata data tidak berdistribusi normal, namun karena data banyak sehingga kenormalan bisa diabaikan, sedangkan untuk uji autokorelasi menggunakan Durbin WatsonTest dapat disimpulkan tidak ada autokorelasi pada residual, untuk uji Homoskesdastisitas menggunakan Breusch Pagan Test dapat disimpulkan residual bersifat homoskedastisitas.

1. PENDAHULUAN

Dalam kehidupan sehari-hari terdapat hal-hal yang dapat diselesaikan menggunakan matematika, statistika adalah salah satu cara dalam mengumpulkan data, mengolah, menganalisa dan menyimpulkan. Analisis Regresi merupakan salah satu teknik untuk melihat hubungan antara 2 variabel atau lebih dan kemudian mengestimasinya menjadi sebuah model yang dapat menjadi sebuah persamaan yang dapat menghubungkan variabel tergantung (*dependent variable*) terhadap variabel-variabel bebas (*independent variabels).* Banyak paper meregresi estimasi non-parametrik untuk efisiensi produksi atas variabel-variabel bebas dalam prosedur-prosedur tertentu untuk menjelaskan faktor yang mungkin mempengaruhi kinerja dari variabel tergantungnya. Model regresi yang menangani situasi tersebut memerlukan satu set persamaan (satu persamaan tunggal saja tidak cukup) yang perlu diselesaikan

secara simultan dan model ini dikenal sebagai *model ekonometrik*. lebih dahulu akan di deskripsikan suatu data yang layak untuk model-model seperti ini. Kita mengajukan prosedur-prosedur bootstrap tunggal dan ganda; keduanya memungkinkan inferensi valid, dan prosedur bootstrap ganda memperbaiki efisiensi statistik dalam regresi. Kita menguji kinerja statistik estimator-estimator kita dengan menggunakan metode Metode Quadratic Mode Estimator (QME), Metode Symmetrically Trimmed Least Squares (STLS) dan Metode Left Truncated (LT) .

Regresi berganda terdapat kasus khusus dalam sebuah analisa regresi, pada regresi berganda terdapat satu variabel tak bebas yang akan diprediksi, tetapi terdapat dua atau lebih variabel bebas, dimana bentuk umum dari regresi berganda adalah :

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$, sehingga jika *Y* adalah variabel yang akan diramalkan maka $X_1, X_2, ...X_k$ adalah dapat diuji pengaruhnya terhadap *Y*, dan variabel $X_1, X_2, ...X_k$ tersebut dapat digunakan untuk menduga nilai di masa mendatang. Dimana model regresi secara teori dapat dijelaskan $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$ dengan $\beta_0, \beta_1, \beta_2, ...\beta_k$ adalah parameter tetap, $X_1, X_2, ...X_k$ diukur tanpa galat, sedangkan ε adalah suatu variabel random yang diukur secara menyebar secara normal disekitar nol (nilai tengah ε) dan mempunyai suatu ragam V_ε, sedangkan model regresi secara praktek dapat dijelaskan $Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + ... + b_k X_{ki} + e_i$, untuk i = *1,2,3,….,N* dimana $X_1, X_2, ...X_k$ diasumsikan diukur tanpa galat, $b_0, b_1, b_2, ..., b_k$ adalah penaksir $\beta_0, \beta_1, \beta_2, ...\beta_k$ dan semuanya adalah variabel acak, dengan sebaran bersama yang normal, sedangkan e_i (*i=1,2,3,…,N)* adalah suatu bagian galat taksiran, untuk pengamatan ke-*i* dan diasumsikan merupakan sampel independen dari suatu sebaran normal.

Pemecahan koefisien sendiri dapat dijelaskan sebgai berikut $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$ baik secara eksplisit maupun implisit prakstisi tersebut memuat berbagai asumsi tentang koefisien, ukuran *X* (bahwa *X* diukur tanpa kesalahan) dan ε bagian kesalahan, bentuk pragmatisnya adalah : $Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k + e$ dan untuk setiap vektor pengamatan, dimana pengamatan ke-*i* dinotasikan sebagai : $Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + ... + b_k X_{ki} + e_i$ dimana $b_0 + b_1 X_{1i} + b_2 X_{2i} + ... + b_k X_{ki} = \hat{Y}_i$, sehingga $Y_i = \hat{Y}_i + e_i$, sehingga didapatkan $e_i = Y_i - \hat{Y}_i$ dan metode *Ordinary Least Square* (OLS) atau jumlah kuadrat kecil minimum dari kesalahan tersebut yaitu meminimumkan $\phi = \sum_{i=1}^{n} e_1^{\,2} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$, dimana $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i})^2$.

Pada pemilihan model yang terbaik akan dilakukan dengan 3 metode, yaitu Metode Quadratic Mode Estimator (QME), Metode Symmetrically Trimmed Least

Squares (STLS) dan Metode Left Truncated (LT). data yang digunakan adalah data polusi udara yang disebabkan oleh 7 variabel bebas yang meliputi jumlah kendaraan yang melewati, suhu udara, kecepatan angin, perbedaan temperatur, angin, jam aktif dan hari aktif, kemudian dengan bantuan software program R akan disajikan simulasi pengolahan data menjadi model yang terbaik dengan penggunaan package library(boot), library(misctools), library(maxlik), library(truncreg), library(truncSP). Hasil simulasi model terbaik ditunjukkan dengan RMSE terkecil.

Analisis regresi telah lama dikembangkan untuk mempelajari pola dan mengukur hubungan statistik antara dua atau lebih variabel. Teknik analisis yang mencoba menjelaskan bentuk hubungan antara dua atau lebih variabel atau lebih khususnya hubungan antara peubah-peubah yang mengandung sebab akibat disebut Analisis Regresi. Prosedur analisisnya didasarkan pada distribusi probabilitas bersama variabel-variabelnya. Bila hubungan ini dapat dinyatakan dalam persamaan matematik, maka dapat dimanfaatkan dalam keperluan sehari-hari, misalnya untuk melakukan prediksi, meramal dan sebagainya.

Persamaan matematik yang memungkinkan melakukan peramalan nilai-nilai suatu variabel tak bebas dari satu atau lebih variabel bebas disebut persamaan regresi. Istilah ini berasal dari hasil pengamatan yang dilakukan Sir Francis Galton (1822 – 1911) yang membandingkan antara tinggi badan anak laki-laki dengan tinggi badan bapaknya. Galton menyatakan bahwa tinggi badan anak laki-laki dari bapak yang tinggi pada beberapa generasi kemudian cenderung "mundur" (regressed) mendekati rata-rata populasi.

2. REGRESI PARAMETRIK

Uji regresi keseluruhan baik yang berparamater (regresi parametrik) dan juga regresi yang kita asumsikan *smooth* (regresi semiparametrik) terlebih dahulu akan dibahas untuk regresi parametrik untuk parameter $\beta_1, \beta_2, \ldots, \beta_k$ yang merupakan elemen β dalam model $\boldsymbol{y} = X\beta + \varepsilon$. dalam hal ini akan mengasumsikan bahwa y berdistribusi $N_n (X\beta, \sigma^2 I)$, dimana X berukuran $n \times (k+1)$ dari rank $k+1 < n$. x tersebut adalah konstanta yang ditetapkan.

METODE KUADRAT TERKECIL

Prosedur penarikan garis regresi yang banyak dikenal adalah metode kuadrat terkecil (ordinary least squares) atau yang lebih dikenal dengan istilah OLS. Metode ini memilih suatu garis regresi yang membuat jumlah kuadrat jarak vertikal dari titik-titik yang dilalui garis lurus tersebut sekecil mungkin, dimana jika model populasi regresi linier ganda adalah $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$ sedangkan model estimasi regresi linier ganda adalah $Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k + e$ maka estimasi OLS pada regresi linier ganda adalah :

Persamaan regresi populasi : $y = X\beta + u$

Residual (estimasi dari galat acak) : $\hat{u} = y - X\hat{\beta}$

Jumlah Kuadrat Galat (JKG) : $\hat{u}'\hat{u} = (y - X\hat{\beta})'(y - X\hat{\beta})$

$$= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}$$

Meminimumkan JKG : $\frac{\partial(\hat{u}'\hat{u})}{\partial\hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0$

Estimator OLS : $\hat{\beta} = (X'X)^{-1}X'y$

ESTIMASI REGRESI LINIER

Perkiraan regresi linier dibuat untuk meningkatkan ketelitian dengan menggunakan variabel tambahan x_i yang berkorelasi dengan y_i. Bila hubungan antara x_i dan y_i diuji, mungkin ditemukan bahwa walaupun hubungannya mendekati linier, garisnya tidak melalui titik origin. Hasil ini menyarankan suatu perkiraan yang didasarkan pada regresi linier dari y_i pada x_i lebih baik daripada rasio dua variabel.

Misalkan bahwa y_i dan x_i masing-masing diperoleh untuk setiap unit dalam sampel rata-rata populasi $\bar{X}$ dan x_i diketahui. Perkiraan regresi linier $\bar{Y}$, rata-rata populasi y_i adalah

$$\bar{y}_{lr} = \bar{y} + \beta(\bar{X} - \bar{x})$$

Dimana notasi lr menyatakan *regresi linier* dan β adalah suatu koefisien perkiraan dari perubahan dalam *y* bila *x* meningkat. Alasan utama dari perkiraan ini adalah jika $\bar{x}$ dibawah rata-rata, harus mengira $\bar{y}$ juga dibawah rata-rata dari suatu jumlah $\beta(\bar{X} - \bar{x})$ karena regresi dari y_i pada x_i .untuk suatu perkiraan jumlah populasi *Y*, kita ambil $\bar{Y}_{lr} = N\hat{y}_{lr}$.

Watson (1937) menggunakan suatu regresi dari luas daun untuk memperkirakan rata-rata luasnya pada suatu pabrik. Prosedurnya adalah dengan menimbang seluruh daun-daun pada pabrik. Untuk sebuah sampel kecil dari daun, luas dan berat masing-masing daun telah ditetapkan. Rata-rata regresi pada berat daun, inti dari aplikasinya adalah bahwa berat daun dapat ditemukan dengan cepat tetapi penentuannya menyita banyak waktu.

Perkiraan regresi, $\bar{y} + \beta(\bar{X} - \bar{x})$ memperbaiki rata-rata sampel dari pengukuran sebenarnya dengan perkiraan regresi yang cepat dari pengukuran sebenarnya. Perkiraan yang cepat ini tidak bebas dari bias. Bila $x_i - y_i$ = *D*, sehingga perkiraan cepat adalah sempurna kecuali untu satu bias konstan *D*, kemudian dengan β = 1 perkiraan regresi menjadi $\bar{y} + (\bar{X} - \bar{x}) = \bar{X} + (\bar{y} - \bar{x})$ = (rata-rata populasi dari perkiraan cepat) + (penyesuaian untuk bias)

Jika tidak ada model regresi linier yang diumpamakan, pengetahuan tentang sifat perkiraan regresi adalah dari cakupan yang sama seperti pengetahuan tentang perkiraan rasio. Perkiraan regresi adalah konsisten, dalam pengertian sederhana bila sampel terdiri dari seluruh unit populasi, $\bar{x} = \bar{X}$ dan perkiraan regresi mengurangi $\bar{Y}$

. Sebagaimana akan diperlihatkan, perkiraan regresi secara umum adalah bias, tetapi rasio biasnya untuk kesalahan baku menjadi kecil bila sampel besar.

Dengan suatu pemilihan β yang sesuai, perkiraan regresi termasuk seperti kasus-kasus khusus dari rata-rata per unit maupun perkiraan rasio, bila β diambil sama dengan nol, $\bar{y}_{lr}$ mengurangi $\bar{y}$. bila $\beta = \bar{y}/\bar{x}$,

$$\begin{aligned} \bar{y}_{lr} &= \bar{y} + \beta(\bar{X} - \bar{x}) \\ &= \bar{y} + \bar{y}/\bar{x}(\bar{X} - \bar{x}) \\ &= \bar{y}/\bar{x}(\bar{X}) \\ &= \hat{\bar{Y}}_R \end{aligned}$$

3. REGRESI SEMI-PARAMETRIK

Pemotongan atau penyensoran dari suatu variabel respon dalam suatu model regresi adalah salah satu masalah yang sering muncul dalam banyak aplikasi. Pemotongan terjadi, sebagai contoh, pada saat mengamati nilai dari kerusakan infrastruktur yang diasuransikan, dalam suatu kebakaran, pencurian, atau kejadian-kejadian lainnya yang serupa, karena kehilangan-kehilangan yang nilainya lebih kecil dari yang bisa dikurangi tidak akan bisa dilaporkan pada perusahaan asuransi. Proses penyensoran sering terjadi pada saat meneliti durasi, misalnya pengangguran dalam ekonomi perburuhan, waktu bertahan dalam percobaan bidang kedokteran, dan waktu kegagalan komponen dalam proses-proses industri.

Dalam hal ini, digunakan model regresi seperti berikut:

$y_i = m(x_i) + \varepsilon_i$, $\quad i=1,2,...n$

dengan y merupakan variabel respon laten, x merupakan variabel penjelas, m(x) merupakan nilai yang tidak diketahui yang nilainya p+1 ($p \geq 1$) kali serta merupakan fungsi yang bisa diturunkan (di-diferensialkan), dan ε adalah kesalahan acak yang terdistribusi secara independen dan merata dengan rata-rata nol dan variansi terbatas pada nilai-nilai tertentu.

Metode Symmetrically Trimmed Least Squares (STLS) .

Estimator kuadrat terkecil yang terpotong secara simetris (Powell, 1986) bisa digunakan untuk menangani pemotongan atau penyensoran dalam pengaturan model regresi (semi)-parametris, yakni pada saat $m(x_i)$ dalam (1) bisa dideskripsikan secara parametris, sebagai contoh dengan polinomial $m(x_i) = \beta_0 + \beta_1 x_i + ... + \beta_p x_{i^p}$. Pemotongan (atau penyensoran) dari variabel respon mengenalkan suatu ketidaksimetrisan dalam suatu distribusi. Estimator STLS dan SCLS secara simetris memotong dan menyensor, secara berurutan, variabel respon dalam rangka untuk mengembalikan kesimetrisan distribusi pada $\beta_0 + \beta_1 x_i + ... + \beta_p x_{i^p}$. Pada cara ini, estimator kuadrat terkecil bersifat konsisten dan tegak lurus secara asimptotis, dalam

kondisi-kondisi aturan tertentu, termasuk asumsi kesalahan yang terdistribusi secara simetris.

Pada kasus pemotongan sebelah kiri (pata t=0), dan untuk model polinom, estimator parametris STLS bisa didefinisikan sebagai:

$$h(x_0) = \arg\min_\theta \sum_{i=1}^{n} (y_i - \max(\frac{1}{2} y_i, x_{iT}\beta))^2, \theta$$

dengan $x_1 = (1, x_1, ..., x_{i^p})^T$ dan $\beta = (\beta_0, ..., \beta_p)^T$

Oleh karena itu, perlu didefinisikan estimator STLS lokal untuk $m(x_0)$ dalam (1) dengan variabel respon yang terpotong di sebelah kiri pada t=0 oleh $m(x_0) = e^T \theta_h(x_0)$ dengan e=(1,0,....,0)T dan

$$\hat{\theta}_h(x_0) = \arg\min_\theta \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right)\left(y_i - \max\left(\frac{1}{2} y_i, z_i^T, \theta\right)\right)^2$$

dengan $z_i = (1, (x_i - x_0), ..., (x_i - x_0)^p)^T, \theta = (\theta_0, \theta_1, ..., \theta_p)^T$ dan K merupakan fungsi kernel dengan ordo r, yang memenuhi $\int K(u)du = 1, \int u^k K(u)du = 0$ untuk k=1, …, r-1 dan $\int u^T K(u)du \neq 0$. Pilihan-pilihan alternatif yang umum untuk K(.) adalah fungsi densitas probabilitas (pdf) Gaussian atau fungsi tricube sebagaimana dijelaskan pada persamaan berikut ini.

Quadratic Mode Estimator (QME) dan STLS

Metode QME, metode STLS dan metode Left Truncated akan diaplikasikan pada data produksi polusi udara yang dilakukan penarikan sampel sebanyak 460 responden, dimana informasi data disajikan dalam tabel 4.1.

Tabel 4.1

	Part.polusi	**Mobil**	**suhu**	**Kec.udara**	**Perb.suhu**	**Sudut.angin**	**jam**	**hari**
1	3.66356	7.74414	-4.4	4.2	0.0	18.0	19	116
2	3.04452	8.03398	-5.7	4.8	-0.3	69.1	9	506
3	3.71357	4.70048	-13.5	4.3	0.2	80.0	3	95

4	2.94444	7.52510	1.4	3.0	0.1	177.0	22	161
5	4.06044	7.76260	4.1	5.6	1.1	287.0	7	80
6	3.68888	7.88683	5.8	2.3	-0.1	200.0	9	33
7	3.33220	7.81521	2.7	1.9	0.4	228.0	7	129
8	3.36730	7.77779	7.1	8.9	0.2	220.0	15	155
9	2.07944	6.89163	4.1	2.0	0.1	183.0	9	132
10	3.33220	5.71373	-9.0	3.4	0.0	80.0	6	98
. .	. .	. .	. .	. .	. .	. .	. .	. .

Data selengkapnya bisa di lihat di lampiran 4.1

Didapatkan model terbaik sebagai berikut :
Dari nilai residual setiap model di dapatkan
"Root Mean Square Error"

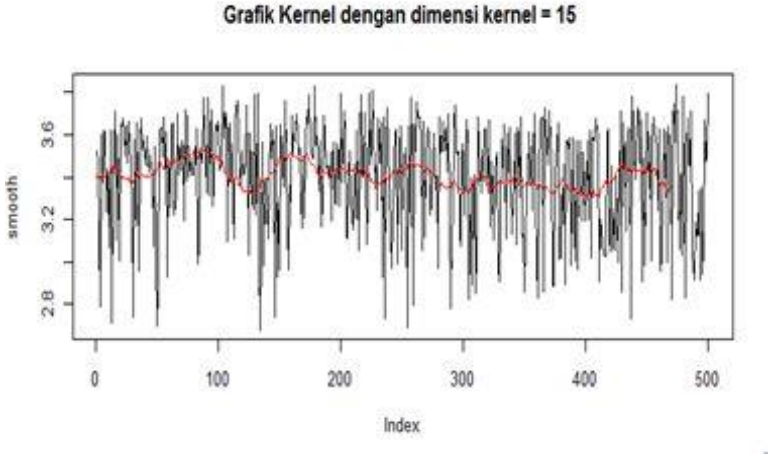
QME	STLS	Lt
0.4759233	4277.689	0.4759233

Pemilihan Model regresi Terpotong terbaik berdasarkan nilai RMSE terkecil RMSE terkecil = 0.4759233 maka
Model regresi terpotong terbaik untuk data adalah "Quadratic Mode Estimator (QME)" dengan ringkasan model regresi terpotong.
Call:
qme(formula = PM10 ~ cars + temp + wind.speed + temp.diff + wind.dir + hour + day, data = data, point = point, direction = arah, beta = metode, covar = TRUE, Cval = metode, level = 1 – alpha

Berdasarkan uraian dari pembahasan dan simulasi, terdapat beberapa hal penting yang dapat disimpulkan, dalam pemilihan model terbaik regresi semi-parametrik diperoleh model terbaik adalah dengan menggunakan metode QME,
Dengan model sebagai berikut :

$$Y(x) = 20851 + 2.014.10^{-1} X_1 - 6.238.10^{-2} X_2 + \hat{m}(t) + \varepsilon$$

hal tersebut dapat dilihat dari nilai RMSE terkecil. Dalam uji parsial terdapat 2 variabel yang signifikan terhadap variabel dependent yaitu Dimana X_1 untuk variabel independen *Cars* dan X_2 untuk variabel independen *Wind Speed*. Sedangkan $\hat{m}(t)$ adalah berupa *plot smoothing fungtion* (gambar 4.2.2). Dalam diagnostic checking dapat disimpulkan uji kenormalan menggunakan *Kolmogorov Smirnov Test* ternyata data tidak berdistribusi normal, namun karena data banyak sehingga kenormalan bisa diabaikan, sedangkan untuk uji autokorelasi menggunakan *Durbin WatsonTest* dapat disimpulkan tidak ada autokorelasi pada residual, untuk uji Homoskesdastisitas menggunakan *Breusch Pagan Test* dapat disimpulkan residual bersifat homoskedastisitas.



Gambar 4.2.2

DAFTAR PUSTAKA

Bain,L.J. and Engelhardt, M., 1992, *Introduction to probability and mathematical statistics*, 2 ed., Duxbury Press, California

Blaxter, L. Hughes, C. and Thight, M., 2001, *How To Research*, Indeks Gramedia., Jakarta.

Cochran, W. G., 1977, *Sampling techniques*, 3 ed., John Wiley and Sons, Inc., New York.

Efron, B. and Tibshirani, R.J.,,1993, *An introduction to the bootstrap*, Chapman and Hall, New York.

Everitt, Brian., 2004, An R and S-Plus Companion to multivariate analysis, Springer., Amerika.

Hardle, W., 1990, *Smoothing techniques with implementation in S*, Springer Verlag,

Hardle, W., Liang, H and Gao, J, 2000, *Partially linear models*, Springer Verlag, Berlin

Haryatmi, Sri., 1988, *Metode Statistika Multivariat*, Universitas Terbuka, Karunika, Jakarta.

Gibbons, J., 1971, *Nonparametric Statistical Inference*, McGraw Hill.

Jhonson, Richard. and Wichern, Dean., 2002, Applied *Multivariat Statistical Analysis*, Pearson Educational International, Amerika

Rencher, Alvin., 2000, *Linear Model in Statistics*, Wiley series in probability and Statistics, Canada.

Rosadi, Dedi., 2011, *Analisis Ekonometrika dan Runtun Waktu Terapan*, Penerbit Andi, Yogyakarta.

Rorres, Anton., 2004, Aljabar Linear Elementer, penerbit Erlangga. Jakarta.

Royden, H.L., 1989, *Real Analysis*, Macmilan Publishing, New York.

Searle, S.R., 1971, *Linear Models*, Wiley Publishers, New York.

Sembiring, R.K., 1995, *Analisis Regresi*, Penerbit ITB, Bandung.

Sumodiningrat, Gunawan., 2007, Ekonometrika Pengantar, BPFE UGM, Yogyakarta.