VOL. 1 TH. 2005 ISSN : 1907 - 2562

PROCEEDING

SNM-2005

SEMINAR NASIONAL MATEMATIKA TAHUN 2005

Jakarta, 30 Juli 2005

Diselenggarakan Oleh :
Departemen Matematika FMIPA Universitas Indonesia

VOL. 1 TH. 2005
ISSN : 1907-2562

PROCEEDING

SNM-2005

SEMINAR NASIONAL MATEMATIKA TAHUN 2005

Jakarta, 30 Juli 2005

Alamat Redaksi :
Gedung D Lantai 2 FMIPA UI, Kampus UI Depok, 16424
Telp. 021 – 7862719 Fax : 021 – 7863439

MODEL REGRESI TERSENSOR KIRI PADA NILAI AMATAN NOL
(Left Censored Regression Model at Zero Observations Value)

Fitria Virgantari[1], fitriav@yahoo.com
Siswadi[2], siswadi@inrr.org

[1]Program Studi Matematika, FMIPA, Universitas Pakuan Bogor
[2]Departemen Matematika, FMIPA, Institut Pertanian Bogor

ABSTRACT

Research aims at estimating censored regression model and ordinary regression model on 500 artificial observations containing zero value of dependent variables have been carried out. Criteria of best model were used according to properties of unbiasedness and minimum variance estimator. Results show that parameter estimate of censored regression model is more accurate than ordinary regression model. Censored regression model also yield smaller MSE (Mean Square Error) than ordinary regression model indicating that censored regression model is better than ordinary regression model.

Keywords: zero observations, censored regression model, unbiasedness, MSE

PENDAHULUAN

Analisis regresi merupakan analisis yang digunakan untuk menjelaskan pola hubungan antara peubah tak bebas dengan satu atau lebih peubah bebas dari segugus pengamatan. Pendugaan parameter dalam persamaan regresi biasa dilakukan melalui metode kuadrat terkecil dengan asumsi kenormalan, kebebasan dan kehomogenan ragam. Namun, fenomena yang terjadi kadang menghasilkan respons yang berstruktur kontinu dengan kisaran yang mungkin sangat besar, sehingga menyebabkan timbulnya masalah heteroskedastisitas (ketidakhomogenan ragam). Hal ini sering dijumpai pada data survey konsumsi/pengeluaran rumah tangga; di mana sebagian rumah tangga tidak mengkonsumsi jenis komoditas tertentu (*zero consumption*); sedangkan rumah tangga yang lain mengkonsumsi dengan jumlah yang sangat bervariasi atau tersensor kiri pada nilai amatan nol. Sebagai contoh adalah pengeluaran untuk tembakau, pada karakter (X) tertentu. Nilai ini akan bernilai nol bila obyek penelitian bukan perokok (*zero consumption*), dan akan bervariasi bila X adalah perokok. Hal ini akan berimplikasi pada metode pendugaan parameter dari model yang digunakan.

Model regresi klasik dengan metode kuadrat terkecil (*Ordinary Least Square/OLS*) akan cukup baik didekati apabila asumsi kenormalan, kebebasan dan kehomogenan ragam dipenuhi. Namun, semakin banyak pengamatan bernilai nol pada data yang diperoleh akan menyebabkan timbulnya masalah heteros-kedastisitas. Penggunaan metode OLS akan menghasilkan penduga yang berbias dan tidak konsisten karena asumsi yang mendasari tidak dipenuhi. Sedangkan penghilangan pengamatan bernilai nol *(zero consumption)* tersebut akan mengurangi ukuran sampel dan tidak mencerminkan keadaan yang sebenarnya, karena rumah tangga dengan *zero consumption* tetap merupakan bagian dari populasi dan *zero consumption* merupakan keputusan rumah tangga yang bersangkutan. Pada keadaan tersebut model alternatif yang dapat digunakan adalah model regresi tersensor. Pada model ini, data dipecah dalam dua bagian (yang bernilai nol dan bukan) sehingga fungsi kepekatannya merupakan fungsi kepekatan bersama, kemudian pendugaan parameternya dilakukan dengan metode kemungkinan maksimum. Tulisan ini bertujuan untuk mengkaji penerapan model regresi tersensor pada data yang mengandung banyak amatan bernilai nol pada peubah tak bebas, dan membandingkannya dengan model regresi biasa.

2. LANDASAN TEORI

2.1. Model regresi biasa

Persamaan regresi linear berganda adalah persamaan dari satu peubah tak bebas (Y) dengan satu atau lebih peubah bebas $(X_1, X_2, \ldots, X_p)$. Hubungan antara peubah-peubah tersebut dapat dirumuskan dalam bentuk persamaan :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \varepsilon_i \quad (1)$$

atau dalam bentuk matriks:

$$Y = X\beta + \varepsilon \quad (2)$$

di mana X adalah matriks peubah bebas berukuran nxk, Y adalah vektor peubah tak bebas berukuran nx1, β adalah vektor parameter berukuran kx1, ε adalah vektor galat (sisaan) berukuran nx1; n adalah banyaknya pengamatan, dan k=p+1 adalah banyaknya parameter.
Pendugaan parameter dalam model regresi biasa ini dilakukan dengan meminimumkan jumlah kuadrat sisaan atau

$$\varepsilon'\varepsilon = Y'Y - 2\beta' X'Y + \beta' X'X\beta \tag{3}$$

Sebagai nilai dugaan, maka akan dipilih β sedemikian rupa sehingga nilai $\varepsilon'\varepsilon$ akan minimum. Caranya adalah dengan mendiferensialkan persamaan (3) terhadap β dan kemudian disamakan dengan nol, yaitu:

$$\frac{\partial(\varepsilon'\varepsilon)}{\partial\beta} = -2X'Y + 2X'X\beta = 0 \tag{4}$$

sehingga akan didapatkan:

$$\hat{\beta} = (X'X)^{-1}(X'Y) \tag{5}$$

bila $X'X$ adalah matriks nonsingular (berpangkat penuh). Apabila matriks $X'X$ tidak berpangkat penuh, maka penduga β dicari dengan matriks kebalikan umum. Penduga tersebut bersifat tidak unik, dan solusi umumnya (Kshirsagar, 1983) adalah:

$$\overset{*}{\beta} = \hat{\beta} + (I - H)z \tag{6}$$

di mana $H = S^{-}S$ adalah matriks idempoten berukuran pxp yang mempunyai sifat $H^2 = H$, $SH = S$, pangkat H = pangkat S = pangkat X = tr H; dan z adalah vektor sembarang; sedangkan

$$\hat{\beta} = S^{-}X'Y \tag{7}$$

di mana S^{-} adalah kebalikan umum dari $S = X'X$.

2.2. Model regresi tersensor

Misalkan Y^* adalah peubah yang berdistribusi normal dengan ragam σ^2. Anggaplah $(y_1^*, y_2^*, \ldots, y_n^*)$ adalah sampel berukuran n dan data tercatat hanya pada nilai-nilai Y^* yang lebih besar dari nol. Untuk nilai-nilai $Y^* \leq 0$, dimasukkan nilai 0, atau:

$$Y = \begin{cases} Y^*, untuk Y^* > 0 \\ 0, selainnya \end{cases} \tag{8}$$

di mana $Y_i^* = X_i'\beta + \varepsilon_i$, β adalah vektor parameter berukuran kx1, X_i adalah vektor peubah bebas berukuran kx1 dan ε adalah sisaan yang merupakan peubah acak bebas dan berdistribusi normal dengan nilai tengah nol dan ragam σ^2; n adalah banyaknya pengamatan, dan k=p+1 adalah banyaknya parameter.

Sampel $y_1, y_2, \ldots, y_n$ disebut dengan sampel tersensor, sedangkan Y disebut dengan peubah tersensor. Pada pengamatan $Y = 0$ yang berarti $Y^* \leq 0$, maka;

$$P(Y = 0) = P(Y^* \leq 0) \tag{9}$$

Secara teoritis, pendugaan parameter pada model regresi tersensor menurut Maddala (1983) dilakukan dengan memisahkan peng-amatan y_i yang sama dengan nol dan y_i yang lebih besar dari nol. Misalkan N_0 adalah banyaknya pengamatan di mana $y_i=0$ dan N_1 adalah banyaknya pengamatan di mana $y_i>0$, dan didefinisikan:

$$F_i = F(x_i'\beta, \sigma^2) = \int_{-\infty}^{x_i'\beta} \frac{1}{\sigma\sqrt{2\pi}} e^{-t^2/2\sigma^2} dt \tag{10}$$

$$f_i = f(x_i'\beta, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i'\beta}{\sigma}\right)^2} \tag{11}$$

$$\Phi_i = \int_{-\infty}^{x_i'\beta/\sigma} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

$$\phi_i = \sigma f_i = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i'\beta}{\sigma}\right)^2}$$

Φ_i dan ϕ_i masing-masing adalah fungsi distribusi dan fungsi kepekatan normal baku yang dievaluasi pada $x_i'\beta/\sigma$. Misalkan pula

$\gamma_i = \dfrac{\phi_i}{1 - \Phi_i}$ dan

$Y_1' = (y_1, y_2, \ldots, y_{N_1})$ adalah vektor berukuran $1xN_1$ pada pengamatan y_i yang lebih besar dari nol

$X_1' = (x_1, x_2, \ldots, x_{N_1})$ adalah matriks berukuran kxN_1 bagi nilai x_i untuk y_i yang lebih besar dari nol

$X_0' = (x_{N_1+1}, \ldots, x_N)$ adalah matriks berukuran kxN_0 bagi nilai x_i untuk y_i sama dengan nol

$\gamma_0' = (\gamma_{N_1+1}, \ldots, \gamma_N)$ adalah vektor $1xN_0$ bagi nilai γ_i untuk nilai y_i sama dengan nol,

maka untuk pengamatan $Y=0$:

$$P(Y=0) = P(Y^* \leq 0) = P(\varepsilon \leq -x_i'\beta)$$
$$= \int_{-\infty}^{-x_i'\beta} f(u)du = \int_{x_i'\beta}^{\infty} f(u)du$$

$$= 1 - F(x_i'\beta, \sigma^2) = (1-F_i) \qquad (14)$$

Berdasarkan (9), maka $P(Y=0)$ + $P(Y>0)=1$, sehingga fungsi kepekatan peluang dari y_i adalah:

$$f(y_i) = (1-F_i)I_{(0)}(y_i) + \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(y_i - x_i'\beta)^2\right\} I_{(0,\infty)}(y_i) \qquad (15)$$

dengan I adalah fungsi indikator yang didefinisikan $I_A(y_i) = \begin{cases} 1, y_i \in A \\ 0, y_i \notin A \end{cases}$ (16)

Dalam model regresi linear, fungsi *likelihood* merupakan fungsi kepekatan peluang bersama dari peubah tak bebas Y_i. Karena faktor sisaan merupakan peubah acak yang saling bebas, maka Y_i juga merupakan peubah acak bebas, sehingga fungsi kemungkinan bagi persamaan (15) adalah:

$$L = \prod_{i=1}^{n} f(y_i) = \prod_0 (1-F_i) \prod_1 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y_i - x_i\beta}{\sigma}\right)^2} \qquad (17)$$

di mana suku pertama meliputi N_0 pengamatan untuk $y_i=0$ dan suku kedua untuk N_1 pengamatan pada $y_i>0$, dan log fungsi kemungkinannya adalah:

$$log\, L = \sum_0 log(1-F_i) + \sum_1 log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \sum_1 \frac{1}{2\sigma^2}(y_i - x_i'\beta)^2 \qquad (18)$$

di mana $\sum\limits_0$ meliputi N_0 pengamatan untuk $y_i=0$ dan $\sum\limits_1$ untuk N_1 pengamatan pada $y_i>0$.

Penduga parameter didapatkan dengan menggunakan turunan pertama dari *log L* terhadap β dan σ^2 yaitu:

$$\frac{\partial logL}{\partial\beta} = -\sum_0 \frac{1}{(1-F_i)}\frac{\partial F_i}{\partial\beta} - \frac{1}{\sigma^2}\sum_1 (y_i - x_i'\beta)(-x_i) = 0$$

(19)

$$-\sum_0 \frac{1}{(1-F_i)} f_i x_i + \frac{1}{\sigma^2}\sum_1 (y_i - x_i'\beta)(x_i) = 0$$

$$-\sum_0 \frac{\sigma f_i x_i}{(1-F_i)} + \frac{1}{\sigma}\sum_1 (y_i - x_i'\beta)(x_i) = 0 \qquad (20)$$

Misalkan $\phi_i = \sigma f_i = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i'\beta}{\sigma}\right)^2}$

$$\Phi_i = \int_{-\infty}^{x_i'\beta/\sigma} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = F_i$$

$$\gamma_i = \frac{\phi_i}{1-\Phi_i}$$

$\gamma_0' = (\gamma_{N_1+1}, \ldots, \gamma_N)$ adalah vektor $1xN_0$ bagi nilai γ_i untuk nilai $y_i=0$, maka persamaan (20) dapat dituliskan:

$$-\sum_0 \gamma_0 x_i + \frac{1}{\sigma}\sum_1 (y_i - x_i'\beta)(x_i) = 0$$

$$-X_0'\gamma_0 + \frac{1}{\sigma}X_1'(Y - X_1\beta) = 0$$

di mana suku pertama meliputi pengamatan untuk $y_i=0$ dan suku kedua meliputi pengamatan untuk $y_i>0$. Selanjutnya,

$$X_1'(Y - X_1\beta) = \sigma X_0'\gamma_0$$

$$(X_1'X_1)\beta = X_1'Y - \sigma X_0'\gamma_0$$

$$(X_1'X_1)^{-1}(X_1'X_1)\beta = (X_1'X_1)^{-1}X_1'Y - \sigma(X_1'X_1)^{-1}X_0'\gamma_0$$

dan didapatkan :

$$\hat{\beta} = (X_1'X_1)^{-1}X_1'Y - \sigma(X_1'X_1)^{-1}X_0'\gamma_0$$

$$= \hat{\beta}_{LS} - \sigma(X_1'X_1)^{-1}X_0'\gamma_0 \qquad (21)$$

di mana $\hat{\beta}_{LS}$ adalah penduga kuadrat terkecil (*least square*) dari β yang didapatkan dari N_1 pengamatan bukan nol.

Sedangkan penurunan terhadap σ^2 adalah:

$$\frac{\partial logL}{\partial\sigma^2} = -\sum_0 \frac{1}{(1-F_i)}\frac{\partial F_i}{\partial\sigma^2} - \sum_1 \frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_1 (y_i - x_i'\beta)^2 = 0 \qquad (22)$$

karena suku pertama sama dengan nol, maka persamaan tersebut akan menjadi:

$$\frac{1}{\sigma^2}\left[\sum_1 y_i^2 - 2\beta'\sum_1 x_i y_i + \beta'\sum_1 x_i y_i\right] = N_1$$

(karena $y_i = x_i'\beta$)

$$\frac{1}{\sigma^2}\left[\sum_1 y_i^2 - \beta'\sum_1 x_i y_i\right] = N_1$$

$$\frac{1}{\sigma^2}\left[\sum_1 (y_i - x_i'\beta)y_i\right] = N_1$$

dan didapatkan:

$$\hat{\sigma}^2 = \frac{\sum (y_i - x_i'\beta)y_i}{N_1} = \frac{Y_1'(Y_1 - X_1\beta)}{N_1} \qquad (23)$$

3. CONTOH KASUS DAN PEMBAHASAN

Untuk mengkaji penerapan model regresi tersensor tersebut, berikut ini akan disajikan hasil analisis terhadap data bangkitan sebanyak 500 pengamatan dengan pengamatan bernilai nol sebanyak 244 atau sekitar 48.8%; dan kemudian akan dibandingkan dengan model regresi biasa (metode OLS).

Data yang dibangkitkan meliputi peubah X_1 dan X_2 serta ε yang masing-masing bebas dan dibangkitkan dari distribusi normal baku (dengan nilai tengah nol dan ragam 1). Sedangkan Y dibangkitkan sesuai dengan model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \tag{24}$$

dengan nilai $\beta_0 = 0$, $\beta_1 = 1$ dan $\beta_2 = 1$. Tahap berikutnya adalah mengambil contoh berukuran 250 dari 500 data yang ada secara sistematis untuk menduga parameter model regresi tersensor dan model regresi biasa. Berdasarkan model yang didapat dari data tersebut, dilakukan validasi model terhadap 250 data sisanya (*cross validation*). Model terbaik ditentukan berdasarkan sifat ketakbiasan masing-masing penduga serta nilai kuadrat tengah galat (MSE) yang dihitung dengan cara:

$$\text{MSE} = \frac{\sum \left(y_i - \hat{y}_i \right)^2}{n} \tag{25}$$

Pendugaan dan analisis dilakukan dengan bantuan *software* SAS versi 8.2e dan Microsoft Excell. Hasil pendugaannya tercantum pada Tabel 1 berikut ini.

TABEL DUGAAN PARAMETER MODEL REGRESI TERSENSOR DAN MODEL REGRESI BIASA

Parameter	Model regresi tersensor		Model regresi biasa	
	Koefisien	P-value	Koefisien	P-value
β_0	-0.00139 (0.0870)	0.9885	0.73103 (0.0459)	0.0001
β_1	1.08199 (0.0860)	0.0001	0.57876 (0.0468)	0.0001
β_2	1.01758 (0.0836)	0.0001	0.4959 (0.0443)	0.0001
MSE	0.03		0.60	

Ket. Angka dalam kurung adalah nilai standard error (galat baku)

Berdasarkan Tabel 1 di atas, terlihat bahwa pada $\hat{\beta}_0$ pada model regresi tersensor adalah -0.001 sedangkan pada model regresi biasa adalah 0.73. Apabila dihubungkan dengan nilai awal ($\beta_0 = 0$) pada data bangkitan, hal ini menunjukkan bahwa $\hat{\beta}_0$ pada model Tobit berbias ke bawah; sedangkan pada model regresi biasa berbias ke atas. Hal tersebut dapat diartikan pula bahwa penduga model Tobit relatif jauh lebih tepat daripada model regresi biasa.

Penduga $\hat{\beta}_1$ dan $\hat{\beta}_2$ pada model Tobit masing-masing adalah 1.08 dan 1.02 sedangkan pada model regresi biasa adalah 0.58 dan 0.50. Apabila dihubungkan dengan nilai awal ($\beta_1 = \beta_2 = 1$) pada data bangkitan, hal tersebut menunjukkan bahwa $\hat{\beta}_2$ dari model Tobit sedikit berbias ke atas, sedangkan penduga model regresi biasa berbias ke bawah. Berdasarkan sifat ketakbiasan tersebut, maka dapat dikatakan bahwa model regresi tersensor lebih baik daripada model regresi biasa.

Berdasarkan Tabel 1 di atas, terlihat pula bahwa nilai MSE pada model Tobit adalah 0.03; sedangkan pada model regresi biasa adalah 0.60; sehingga dapat dikatakan bahwa model regresi tersensor lebih baik daripada model regresi biasa karena menghasilkan nilai MSE yang jauh lebih kecil.

4. KESIMPULAN DAN SARAN

Berdasarkan hasil analisis yang telah dilakukan maka dapat disimpulkan bahwa model regresi tersensor lebih baik daripada model regresi biasa karena menghasilkan nilai MSE yang lebih kecil serta menghasilkan nilai penduga parameter yang relatif lebih tepat. Penelitian lanjutan dapat dilakukan dengan membandingkan berbagai sebaran proporsi pengamatan nol dari data keseluruhan.

5. REFERENSI

[1] G. S. Maddala. *Limited dependent and qualitative variables in econometrics.* New York:Cambridge University Press, 1983, ch. 6, pp. 149-155

[2] D. Heien and C. R. Wessells, "Demand system estimation with microdata: A censored regression approach," *Journal of*

Business & Economic Statistics V ol. 8 (3) pp. 365-371, July 1990.

[3] A. M. Kshirsagar. *A course in linear models.* New York: Marcel Dekker, Inc, 1983, ch. 2, pp. 13-43.

[4] J. Tobin,"Estimation of relationships for limited dependent variables," *Econometrica* Vol. 26, pp. 24-36, 1958.