MEDIAN
Jurnal Arsitektur dan Planologi

TEHNIK ANALISIS MODEL LOGIT DAN PROBIT UNTUK PENELITIAN DALAM BERBAGAI BIDANG ILMU

Sidik Budiono
Tenaga Pengajar Pada Program Studi Ekonomi Pembangunan – STIE Ottow Geisler
Alamat : Kampus STIE Ottow Geisler Kota Raja Abepura Jayapura

ABSTRACT

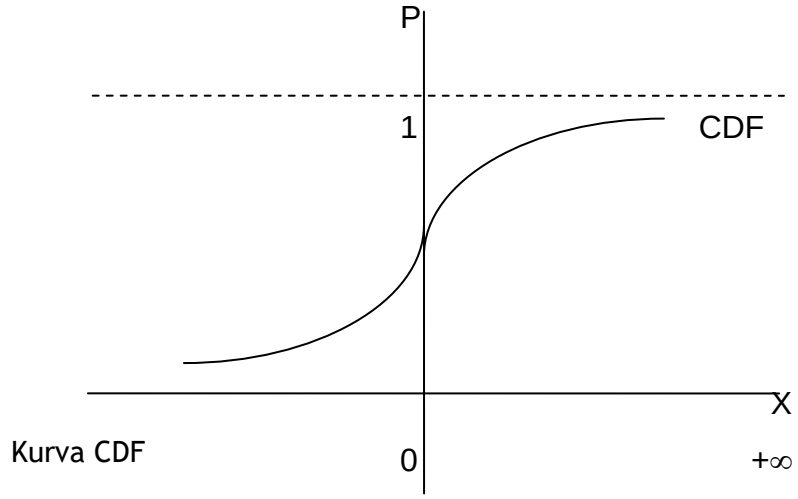
We frequently face natural and social phenomen together. There are quantitative or/and qualitative data. Solving to econometrics analysis is not easy to apply because of some basic assumptions must be required. This paper want to explain how we make solution for quantitative or/and qualitative data for analysis. Logit and Probit model have some advantages and weakness as data type themselves.

Key Words*: Probit, Logit, Ordinary Least Square, Weighted Least Square, Cumulative Density Function*

1. PENGANTAR

Model Logit-Probit memiliki banyak masalah (Gujarati, 1999), seperti (1) Non-normalitas dari u_i, (2) hetero-skedastisitas dari u_i, (3) Kemungkinan dari $\hat{Y}_i$, yang nilainya di luar 0 – 1 dan (4) secara umum nilai R^2 yang rendah. Tetapi masalah-masalah tersebut dapat diatasi. Misalnya, kita dapat menggunakan Weighted Least Square (WLS) untuk mengatasi masalah heteroskedastisitas atau menaikkan ukuran sample untuk meminimalkan masalah non-normalitas. Masalah dasar Model Logit Probit adalah karena model mengasumsikan $P_i = E(Y = 1 \mid X)$ meningkat secara linier seiring X, yaitu efek marjinal dari X tetap konstan.

Suatu model probabilitas yang memiliki dua hal: (1) jika X_i meningkat $P_i = E(Y= 1 \mid X)$ meningkat tetapi nilainya antara 0 – 1, dan (2) hubungan antara P_i dan X_i nonlinier, "mendekati 0 dengan tingkat yang semakin melambat saat X_i yang semakin kecil dan mendekati 1 dengan tingkat yang semakin melambat saat X_i yang semakin besar". Secara geometris, model yang kita ingin lihat akan seperti gambar berikut bahwa probabilitas ada di antara 0 dan 1 dan bervariasi secara non linier dengan X.



Gambar 1. Kurva *Cumulative Density Function* (CDF)

Kurva **sigmoid** (s-shaped) di atas sangat ber-*resembles cumulatif distribution function* (CDF) dari variabel random. Sehingga CDF dapat digunakan untuk regresi model di mana variabel responnya dikotomi, memiliki nilai 0-1. Walaupun semua CDF berbentuk S, tetapi untuk tiap-tiap variabel random ada CDF yang unik. Karena alasan historis dan praktis, CDF biasanya memilih untuk menggambarkan model respon 0-1 adalah (1) logistik dan (2) normal, yang pertama bagi model logit yang kedua untuk model probit (atau normit).

2. PEMBAHASAN

a. Model Logit

Untuk menjelaskan hubungan kepemilikan kategori-1 dengan pendapatan, dalam bentuk Model Logit Probit (Gujarati, 1995):

$$P_i = E(Y = 1|X) = \beta_1 + \beta_2 X_i \quad (1)$$

Di mana X adalah pendapatan dan Y= 1 berarti responden memiliki kategori-1. Persamaan tersebut dapat ditulis menjadi:

$$P_i = E(Y = 1|X) = \frac{1}{\left(1+e^{-(\beta_1+\beta_2 X_i)}\right)} \quad (2)$$

atau menjadi

$$P_i = \frac{1}{(1+e^{-Z_i})} = \frac{e^Z}{1+e^Z} \quad (3)$$

$Z_i = ß_1 + ß_2 X_i$

Persamaan (3) menggambarkan apa yang dikenal sebagai fungsi distribusi logistik (kumulatif). Mudah dipastikan bahwa jika Z_i berkisar antara - ∞ sampai +∞, P_i berkisar antara 0 dan 1 dan bahwa P_i secara nonlinier berhubungan dengan Z_i (yaitu X_i) sehingga memenuhi syarat sebelumnya. Namun dalam memenuhi syarat tersebut menimbulkan masalah pendugaan karena P_i non linier tidak hanya pada X tetapi juga pada ß seperti terlihat pada persamaan (2). Hal ini berarti bahwa kita tidak dapat menggunakan prosedur *Ordinay Least Square* (OLS) untuk menduga parameter. Tetapi persamaan (2) dapat dilinierkan. Jika P_i, kemungkinan memiliki kategori-1, yang ditentukan dari persamaan (3), maka (1-P_i) Kemungkinan memiliki kategori-0, adalah:

$$1 - P_i = \frac{1}{(1+e^{-Z_i})} \quad (4)$$

dapat ditulis menjadi :

$$\frac{P_i}{1-P_i} = \frac{1+e^{Z_i}}{(1+e^{-Z_i})} = e^{Z_i} \quad (5)$$

Sekarang P_i /(1 – P_i) adalah *odds ratio* dalam keinginan memiliki kategori-1 – rasio dari probabilitas bahwa suatu responden akan memiliki suatu kategori-1 terhadap probabilitas memiliki kategori-0. Sehingga, jika P_i = 0.8, artinya *odds* tersebut adalah 4 terhadap 1 dalam kemungkinan memiliki kategori-1.
Sekarang jika kita menggunakan log natural dari persamaan (5) diperoleh hasil yang menarik sebagai berikut :

$L_i = \ln (P_i/(1 - P_i) = Z_i$
$= ß_1 + ß_2 X_i \quad (6)$

L adalah log dari *odd ratio*, tidak hanya linier dalam X, tetapi juga linier dalam parameter. L dinamakan juga logit.

1. Saat P berkisar dari 0-1, *logit* L berkisar dari -∞ sampai +∞. Sehingga walau kemungkinan berada di antara 0 dan 1, logit tidak terlalu terbatas.
2. Walaupun L linier untuk X, tetapi probabilitasnya tidak. Kebalikannya dengan Model Logit Probit dalam persamaan (1) dimana probabilitas meningkat secara linier dengan X.
3. Walaupun kita hanya memasukkan satu variable X, atau regresor pada model tetapi jumlah regresor dapat ditambah sesuai teori yang mendasarinya.
4. Jika L positif artinya jika nilai regresor naik, odds bahwa regresand = 1 meningkat, dan jika L negatif, *odds* bahwa regressand = 1 menurun saat nilai X meningkat. Untuk membedakannya, logit menjadi negatif dan pengaruhnya meningkat saat rasio odds menurun dari 1 menjadi 0 dan menjadi meningkat dan positif saat rasio odds meningkat dari 1 sampai tak terbatas.
5. Secara lebih formal, interpretasi dari model logit untuk persamaan (6) adalah sebagai berikut: $ß_2$, *slope*, pengukur perubahan pada L untuk satu unit perubahan X, yaitu menunjukkan

berapa selisih log dalam kategori-1 berubah saat pendapatan berubah sebesar satu unit. $ß_1$ intersep adalah nilai dari selisih log dalam kategori-1 jika pendapatan nol.

6. Dengan suatu nilai variabel bebas tertentu, misalnya X*, jika kita benar-benar ingin menduga bukan selisih dari kategori-1 tetapi probabilitas kepemilikan kategori-1 itu sendiri, hal ini dapat dilakukan secara langsung dari persamaan (3) jika dugaan $ß_1 + ß_2$ diketahui, maka kita harus mengestimasi $ß_1$ dan $ß_2$.
7. Jika Model Logit Probit mengasumsikan bahwa P_i secara linier berhubungan dengan X_i, model logit mengasumsikan bahwa log dari rasio selisih secara linier berhubungan dengan X_i

b. Pendugaan Model Logit

Untuk pendugaan dalam persamaan (6) sebagai berikut:

$$L_i = \ln\left(\frac{P_i}{1-P}\right) = ß_1 + ß_2 X_i \qquad (7)$$

Untuk menduga persamaan (7), kita memerlukan nilai *regressan*d, atau logit, L_i. Hal ini tergantung pada jenis data yang akan dianalisis. Kita membedakannya menjadi dua jenis data: pertama, data pada tingkat individu atau mikro, dan kedua, data terkelompok atau pengulangan.

c. **Data pada Tingkat Individu**

Jika kita memiliki data individu maka pendugaan OLS pada persamaan (7) tidak dapat dilakukan. Berdasar data $P_i = 1$ jika responden memiliki kategori-1 dan $P_i = 0$ jika kategori-0. Tetapi jika kita memasukkan nilai ini langsung pada logit L_i, kita memperoleh (Greene, 2000):

$L_i = \ln\left(\frac{1}{0}\right)$ Jika responden memiliki kategori-1

$L_i = \ln\left(\frac{0}{1}\right)$ Jika responden dengan kategori-0

Hal tersebut tidak memiliki arti. Sehingga, jika kita memiliki data mikro kita tidak dapat menduga (15.61) dengan OLS standar. Dalam keadaan ini kita mungkin harus beralih pada metode maximum likelihood (ML) untuk menduga parameter.

d. Data Terkelompok

Jika X_i tingkat pendapatan, ada N_i responden, n_i responden yang memiliki kategori-1 ($n_i \le N_i$). sehingga kita hitung:

$$\hat{P}_i = \frac{n_i}{N_i} \qquad (8)$$

yaitu frekuensi relatif , kita dapat menggunakannya sebagai dugaan dari P_i sebenarnya untuk setiap X_i. Jika Ni besar, $\hat{P}_i$ akan menjadi penduga yang baik untuk P_i. Sehingga diperoleh logit dugaan sebagai berikut:

$$\hat{L}_i = \ln\left(\frac{\hat{P}_i}{1-\hat{P}_i}\right) = \hat{\beta}_1 + \hat{\beta}_2 X_i \qquad (9)$$

yang akan menjadi pendugaan yang baik dari logit L_i sesungguhnya jika jumlah N_i pada setiap X_i besar.

Sejauh ini, kita belum tentu dapat menggunakan OLS untuk persamaan (9) untuk menduga parameter karena ada *property* dari *stochastic disturbance* yang harus dipenuhi. Terlihat bahwa jika N sangat besar dan tiap observasi dari kelas nilai X didistribusikan secara bebas sebagai variabel binomial, maka :

$$U_i \sim N\left[0, \frac{1}{N_i P_i (1-P_i)}\right] [0,\ 1 / (N_i\ P_i\ (1-\ P_i\ )]$$

(10)

dimana u_i mengikuti distribusi normal dengan *zero mean* dan *variance* sama dengan $1/[N_i\ P_i\ (1 - P_i\ )]$.

Dalam Model Logit Probit, *disturbance term* dalam model logit adalah heteroskedastik. Sehingga dari pada menggunakan OLS kita lebih baik menggunakan *weighted least squares* (WLS). Untuk tujuan empiris kita akan mengganti P_i yang tidak diketahui dengan $\hat{P}_i$ dan menggunakan :

$$\hat{\sigma}^2 = \frac{1}{N_i \hat{P}_i (1-\hat{P}_i)} \qquad (11)$$

sebagai estimator dari σ^2 .
Sekarang kita mendeskripsikan beberapa langkah dalam pendugaan regresi logit pada persamaan (7):

1. Untuk setiap tingkat pendapatan X, probabilitas kepemilikan kategori-1 adalah
   $\hat{P}_i = n_i / N_i$
2. Untuk setiap X_i , logit adalah
   $$\hat{L}_i = \ln\left[\frac{\hat{P}_i}{1-\hat{P}_i}\right]$$
3. Untuk memecahkan masalah heteroskedastisitas, maka persamaan (7) harus diubah sebagai berikut:
   Dan ditulis sebagai:

$$\sqrt{w_i}\ L_i = \beta_1 \sqrt{w_i} + \beta_2 \sqrt{w_i}\ X_i + \sqrt{w_i}\ u_i \qquad (12)$$

Ditulis sebagai: $L_i^* = ß_1\sqrt{w_i} + ß_2 X_i^* + v_i$
Di mana bobot $w_i = N_i\hat{P}_i(1-\hat{P}_i)$ $L_i^* = L_i$ yang diubah atau diberi bobot; $X_i^* = X_i$ yang diubah atau diberi bobot; dan v_i = *error term* yang diubah. Dengan demikian bahwa *error term* v_i adalah homoskedastik.

4. Pendugaan persamaan (12) dengan OLS – bahwa WLS adalah OLS dengan data yang diubah. Perhatikan bahwa dalam (12) tidak ada intersep yang tampak secara eksplisit. Sehingga harus digunakan regresi dengan cara biasa untuk melakukan pendugaan (12)

e. Model Probit

Untuk menjelaskan sifat dari suatu variabel dependen dikotomi kita harus memilih CDF yang tepat. Greene (2000) menjelaskan bahwa Model logit menggunakan fungsi logistik kumulatif, seperti pada persamaan (2). Tetapi bukan CDF ini saja yang dapat digunakan. Karena CDF normal dalam beberapa penerapan dapat digunakan. Model pendugaan yang timbul dari CDF normal biasa disebut model probit, walaupun sering disebut model normit. Pada prinsipnya CDF normal dapat menggantikan CDF logistik pada persamaan (2).
Untuk mendukung model probit, diasumsikan bahwa dalam contoh kepemilikan kategori-1 keputusan dari responden i untuk memiliki kategori-1 atau tidak, tergantung pada suatu unobservable utility index I_i (juga dikenal sebagai *variable latent*) yaitu yang ditentukan oleh satu atau lebih variabel eksplanator, seperti nilai X_i, dengan suatu cara sehingga nilai indeks I_i yang semakin besar, semakin besar kemungkinan suatu responden memiliki kategori-1. Indeks I_i adalah:

$$I_i = ß_1 + ß_2 X_i \qquad (13)$$

Di mana Xi adalah nilai variabel X dari responden ke-i.
Indeks berhubungan dengan keputusan aktual responden. Dengan Y = 1 jika responden memiliki kategori-1 dan Y = 0 jika tidak. Oleh karena diasumsikan bahwa ada suatu tingkat kritis atau *threshold* dari indeks, I_i^* sebagaimana jika I_i melebihi I_i^*, responden dengan kategori-1, dan sebaliknya. Threshold I^*_i, seperti I_i, tidak terobservasi, tetapi jika kita mengasumsikan bahwa ia terdistribusi secara normal dengan *mean* dan varian yang sama, maka dimungkinkan tidak hanya untuk menduga parameter dari indeks yang ada pada persamaan (13) tetapi juga untuk mendapat informasi mengenai indeks yang tidak terobservasi itu sendiri.

Dengan asumsi normalitas, probabilitas bahwa I_i^* kurang dari atau sama dengan I_i dapat diperoleh dari Kurva CDF normal terstandarisasi yaitu:

$$P_i = P(Y=1 \mid X) = P(I^*_i \le I_i) = P(Z \le ß_1 + ß_2 X_i) = F(ß_1 + ß_2 X_i) \qquad (14)$$

Di mana $P(Y=1 \mid X)$ berarti probabilitas bahwa suatu kejadian terjadi dengan nilai tertentu dari variable X, atau eksplanatori, dan dimana Z_i adalah variabel normal

standar, seperti, Z~N(0,σ²). F adalah CDF normal standar, yang ditulis secara eksplisit dalam konteks ini sebagai:

$$F(I_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{I_i} e^{-Z^2/2dz} \quad (15)$$

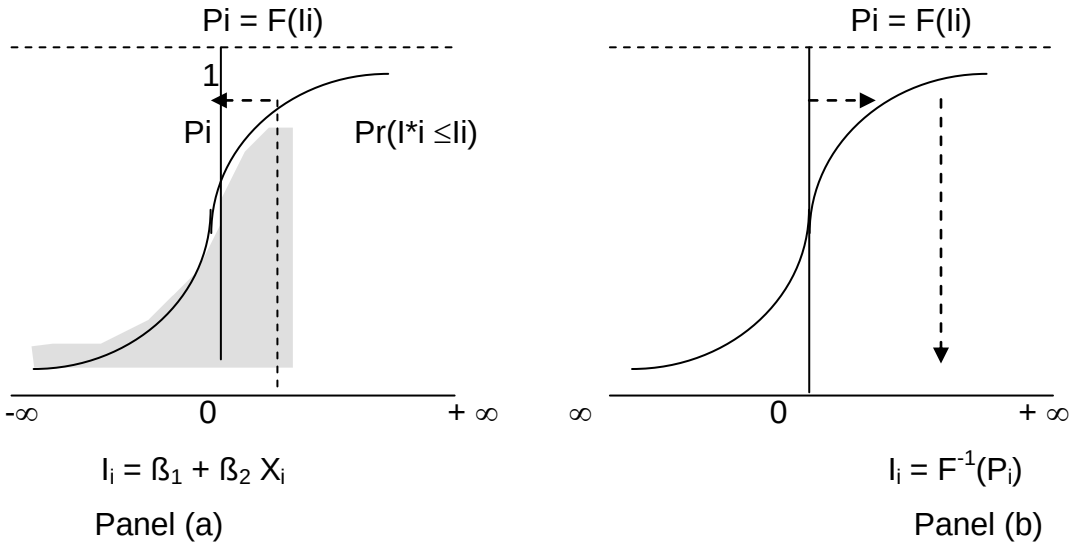
$$F(I_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{ß1 + ß2\ Xi} e^{-Z^2/2dz}$$

karena P menggambarkan probabilitas bahwa suatu kejadian akan terjadi, yaitu probabilitas memiliki kategori-1, dihitung dengan area kurva normal standar dari $-\infty$ sampai I_i.

Untuk mendapat informasi mengenai I_i, indeks utilitas, seperti juga pada $ß_1$ dan $ß_2$ dan kita mengambil invers dari persamaan (13) untuk mendapat:

$I_i = F^{-1}(I_i) = F^{-1}(P_i)$ (16)
$= ß_1 + ß_2\ X_i$

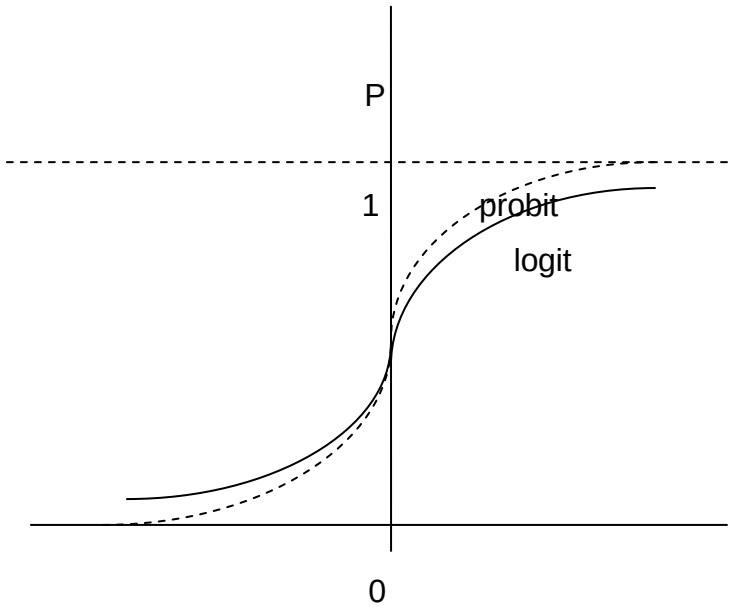
Di mana F^{-1} adalah invers dari CDF normal. Dari panel (a) kita memperoleh dari ordinat probabilitas kumulatif kategori-1 dengan $I_i^* \leq I_i$ di mana di panel (b) kita mendapat dari absis nilai dari I_i dengan nilai P_i tertentu yang secara sederhana berlawanan dengan sebelumnya.



Gambar 2. Fungsi Probabilitas dan Fungsi Invers

f. Model Logit dan Probit

Dalam Model Logit-Probit, logit dan probit memberi hasil yang seragam secara kualitatif, kita akan menfokuskan perhatian pada model logit dan probit karena masalah dengan Model Logit-Probit yang sudah kita sebut sebelumnya. Antara model probit dan logit mana yang lebih baik? pada hampir semua penerapan model hampir sama, perbedaan utamanya adalah distribusi logistik yang memiliki *fatter tail*, dapat dilihat pada Gambar 3. Artinya probabilitas Pi mendekati nol atau 1 untuk logit lebih lambat dari pada probit. Tetapi tidak ada alasan tepat untuk memilih yang satu dibanding yang lain. Agar praktis banyak peneliti memiliki model logit karena *comparative mathematical simplicity*-nya.



Gambar 3. Kurva Probit dan Logit

Walaupun seragam, tetapi harus hati-hati untuk menginterpretasikan koefisien dugaan dari dua model. Karena walaupun logistik standar (basis dari logit) dan standar distribusi normal (basis probit) keduanya memiliki nilai *mean* nol, variannya berbeda; 1 untuk normal standar dan $\pi^2/3$ untuk distribusi logistik, di mana $\pi \approx 22/7$. Sehingga, jika mengalikan koefisien probit dengan sekitar 1.81 (yang mendekati $=\pi/\sqrt{3}$) akan diperoleh pendekatan dari koefisien logit.

3. PENUTUP

Secara intuisi jelas bahwa jika kita melakukan pendugaan garis regresi hanya berdasar observasi, hasil intersep dan koefisien slope akan berbeda dengan jika semua observasi digunakan. Jadi kedua metode memiliki tujuan analisis yang sama untuk data kualitatif dan kuantitatif sekaligus. Kedua metode mengikuti distribusi logistik yang memiliki *fatter tail* yang berbeda sehingga angka dasar (basis) juga akan berbeda.

DAFTAR PUSTAKA

Gujarati, Damodar, 1999, *Essential of Econometrics,* Irwin McGraw-Hill, Singapore

Gujarati, Damodar, 1995, *Basic Econometrics,* third edition, Mc. Graw Hill, New York.

Greene W.H, 2000, *Econometric Analysis,* fourth edition, Prentice – Hall.Inc, New Jersey.

Pindyck R.S dan Rubinfield D.L, 1991, *Econometric Model & Economic Forecast,* Edisi Internasional, third edition, Mc. Graw Hill, New York.