Menduga dan Menguji Koefisien Regresi Logistik Biner Menggunakan Solver di MS Excel

Bagus Sartono - bagusco4@yahoo.com

(Terinspirasi pertanyaan seorang teman tentang penghitungan koefisien regresi logistik. Beberapa hal dalam tulisan ini overlap dengan tulisan banyak orang)

Pendahuluan

Regresi logistik biner telah banyak digunakan secara luas sebagai salah satu alat analisis pemodelan ketika variabel responnya (Y) bersifat biner. Istilah biner merujuk pada penggunaan dua buah bilangan 0 dan 1 untuk menggantikan dua kategori pada variabel respon. Contoh variabel respon yang dimaksud adalah kesuksesan (sukses – gagal), kesetujuan (setuju – tidak setuju), keinginan membeli (ya – tidak), dan masih banyak lagi.

Pendugaan koefisien model regresi logistik tidak dapat dilakukan dengan metode kuadrat terkecil (ordinary least squares) seperti halnya regresi linear karena pelanggaran asumsi kehogenan ragam. Metode kemungkinan maksimum (*maximum likelihood*) menjadi salah satu alternatif yang dapat digunakan. Tidak sederhananya proses komputasi dari metode ini mengakibatkan pengguna harus berpikir untuk mendapatkan perangkat lunak statistik tertentu. Namun demikian, pengguna yang terbiasa dengan aplikasi MS Excel, dapat memanfaatkan add-ins Solver sebagai solusinya.

Di berbagai artikel di internet, kita dapat menjumpai prosedur penggunaan Solver untuk mendapatkan model regresi logistik. Sayangnya hanya sebatas pendugaan koefisien. Pada tulisan ini akan dipaparkan juga proses menghitung statistik G dan nilai kritis pengujian berdasarkan sebaran chi-square. Kembali solver akan digunakan sebagai salah satu cara mendapatkan nilai tersebut. Pada bagian awal akan dituliskan sekilas tentang regresi logistik biner dan metode *maximum likelihood*, kemudian ditutup dengan ilustrasi penggunaan Solver.

Solver Add-Ins

Solver merupakan salah satu prosedur yang tersedia di MS Excel yang dapat digunakan untuk mencari nilai atau kombinasi beberapa nilai yang menghasilkan output paling optimum. Dengan prosedur ini kita dapat memasukkan fungsi dari satu atau beberapa sel di lembar kerja Excel, kemudian kita dapat memperoleh berapa nilai di sel-sel tadi yang dapat menghasilkan fungsi dengan nilai maksimum, minimim, atau yang paling mendekati nilai target.

Pengguna Excel versi 2007 dapat mengaktifkan add-ins ini melalui menu *Excel Options > Add-Ins*, sedangkan pada versi sebelumnya pada menu *Tools > Add-Ins*.

Regresi Logistik Biner

Jika p_i menyatakan peluang suatu individu ke-i memiliki nilai Y = 1, maka model regresi logistik dengan k buah variabel bebas dapat dituliskan sebagai

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + + \beta_k X_k \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

dengan $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$.

Model regresi logistik adalah model linear antara logit(p) dengan variabel penjelas X. Seperti halnya dalam regresi linear, kita bisa mendapatkan nilai-nilai intersep dan *slope* dari model tersebut.

Namun berbeda halnya dengan di regresi linear yang dapat menggunakan metode kuadrat terkecil (*least squares method*) dalam menentukan dugaan β_0 dan β_i, i = 1, 2, ..., k. Secara statistik, metode tersebut mengasumsikan nilai *variance error* bersifat konstan (homogen). Padahal dalam kasus regresi logistik biner, yang nilai Y mengikuti sebaran bernoulli, yang nilai *variance* merupakan fungsi

dari *p*. Tentu saja pada data yang kita miliki nilai *p* ini bervariasi tergantung pada variabel penjelas X. Karena nilai *p* bervariasi, maka nilai *variance* juga bervariasi sehingga *variance* bersifat heterogen. Pendekatan *weighted least squares* dapat mengatasi masalah ini. Sehingga teknik *iteratively reweighted least squares* (IRLS) dapat dijadikan pilihan metode selain metode *maximum likelihood* (ML) dalam menduga parameter model regresi logistik.

Perhatikan bahwa untuk model sederhana dengan satu buah variabel bebas

$\log(p/(1-p)) = \beta_0 + \beta_1 X$

$p/(1-p) = \exp(\beta_0 + \beta_1 X)$

$p = \exp(\beta_0 + \beta_1 X) - p \exp(\beta_0 + \beta_1 X)$

$p\,(1 + \exp(\beta_0 + \beta_1 X)) = \exp(\beta_0 + \beta_1 X)$

Sehingga $p = \exp(\beta_0 + \beta_1 X) / (1 + \exp(\beta_0 + \beta_1 X))$

Dengan kata lain, model regresi logistik biner dapat dituliskan sebagai

$$P(Y=1) = p_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \quad \text{..... (2)}$$

Koefisien β selanjutnya diduga menggunakan metode *maximum likelihood*. Secara sederhana dapat disebutkan bahwa metode ini berusaha mencari nilai koefisien yang memaksimumkan fungsi likelihood. Dengan nilai Y yang bersifat biner, kita dapat menggunakan Bernoulli sebagai sebaran variabel Y sehingga fungsi *likelihood* akan berbentuk

$$L = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \quad \text{dengan } p_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \quad \text{..... (3)}$$

Jelas bahwa nilai β menentukan besarnya nilai fungsi *likelihood* L. Secara komputasi bekerja dengan operator perkalian kurang menyenangkan dibandingkan dengan penjumlahan. Transformasi logaritma dapat digunakan mengubah perkalian menjadi penjumlahan, dan kemudian fungsi *likelihood* diganti dengan fungsi *log-likelihood*. Perhatikan bahwa fungsi logaritma besifat monoton naik, sehingga jika *log-likelihood* mencapai maksimum maka fungsi *likelihood* juga demikian. Bentuk fungsi yang dimaksimumkan dengan demikian adalah

$$LL = \log(L) = \sum_{i=1}^{n} \log p_i^{y_i}(1-p_i)^{1-y_i} = \sum_{i=1}^{n} y_i \log p_i + (1-y_i)\log(1-p_i) \quad \text{..... (4)}$$

Penduga bagi koefisien β diperoleh sebagai solusi bagi permasalahan memaksimumkan LL.

Pengujian peranan variabel bebas, X, dalam model dapat dilakukan menggunakan uji *likelihood ratio* dengan formula

$$G = 2\log\left[\frac{\text{likelihood tanpa peubah bebas}}{\text{likelihood dengan peubah bebas}}\right]$$

Statistik uji-G digunakan untuk menguji peranan variabel penjelas di dalam model secara bersama-sama (Hosmer & Lemeshow, 1989) dengan hipotesis :

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$

H_1 : minimal ada satu β yang tidak sama dengan 0

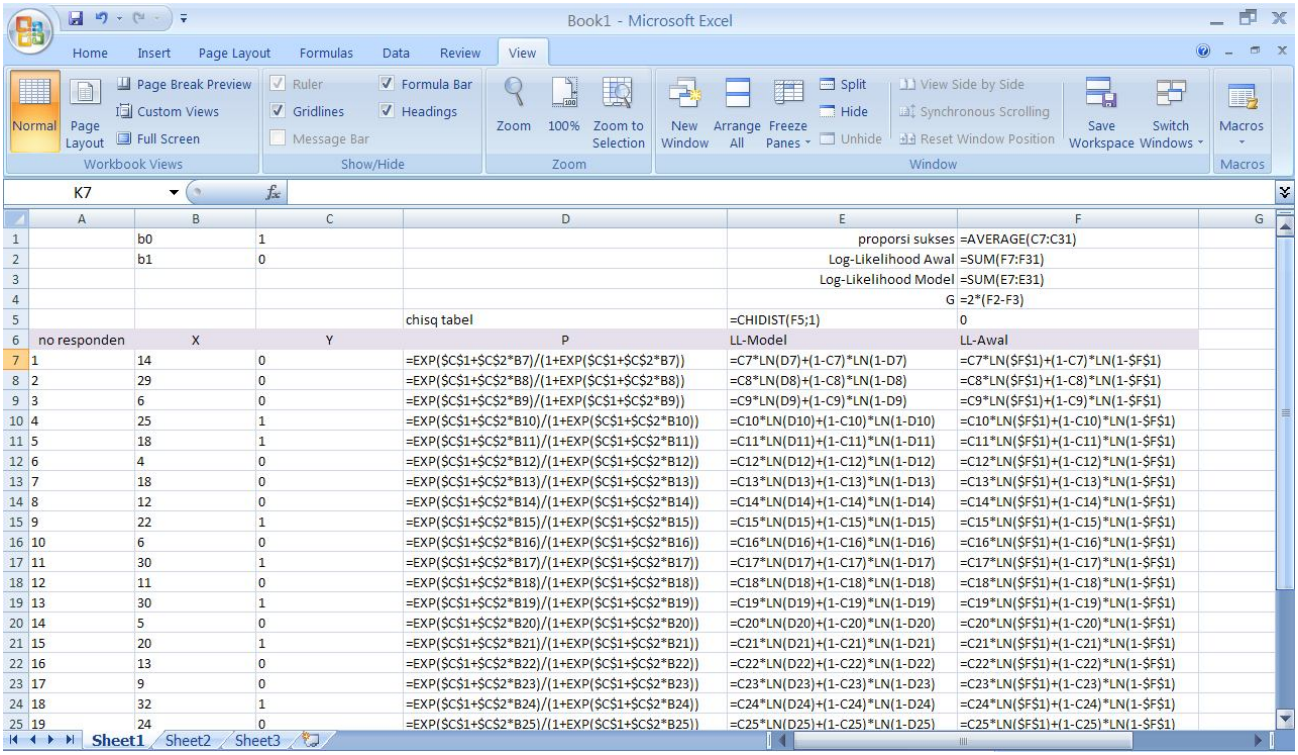
Jika H_0 benar, statistik G ini mengikuti sebaran χ^2 dengan derajat bebas k.

Ilustrasi Penggunaan Solver

Sebagai ilustrasi penggunaan Solver untuk menduga koefisien regresi logistik sekaligus melakukan pengujiannya akan digunakan data dari suatu studi yang menilai kemampuan programmer dalam menyelesaikan program yang rumit, dan menghubungkan kemampuan ini dengan tingkat pengalaman

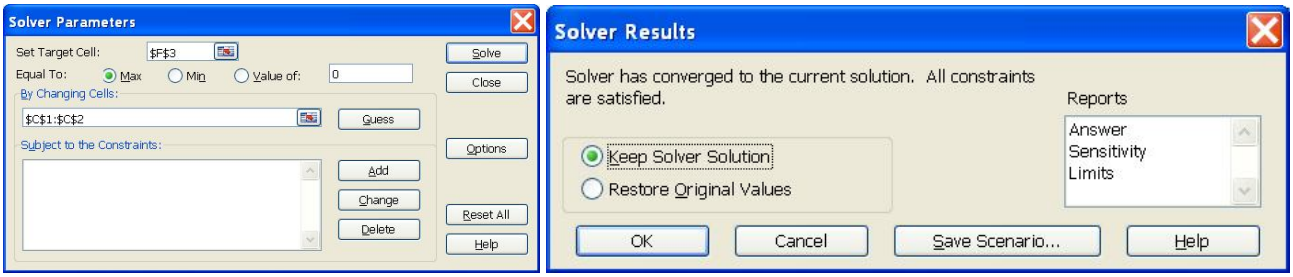
mereka. Dua puluh lima orang programmer terlibat dalam studi ini. Data yang dikumpulkan meliputi dua variable yaitu X=lamanya pengalaman membuat program (dalam bulan) dan Y=keberhasilan membuat program (1=berhasil, 0=tidak).

Pertama yang harus dilakukan adalah menyiapkan lembar kerja berisi data dan formula-formula yang diperlukan. Nilai dugaan awal untuk b_0 dan b_1 dapat diisi secara acak. Kemudian pada kolom p, dihitung nilai dugaan peluang Y=1 dengan memasukkan nilai b_0, b_1, dan X pada persamaan (2). Nilai LLmodel didapat menggunakan fungsi (4) dengan nilai p berasal dari kolom P. Sedangkan LLawal menggunakan nilai proporsi yang berhasil (Y = 1) pada data. Nilai total LLawal dan LLmodel tinggal menjumlahkan masing-masing kolom tersebut. Pada ilustrasi Gambar 1, fungsi LLmodel pada sel F3 yang akan dimaksimumkan dengan mengubah-ubah nilai b_0 dan b_1 pada sel C1 dan C2. Gambar 2 menunjukkan tampilan Solver yang meminta pengguna mengisi sel mana yang dioptimumkan, bentuk fungsi tujuan, dan sel mana yang akan dicari supaya hasilnya optimum. Hasil dari dugaan koefisien regresi disajikan pada Gambar 4.



	A	B	C	D	E	F	G
1		b0	1		proporsi sukses	=AVERAGE(C7:C31)	
2		b1	0		Log-Likelihood Awal	=SUM(F7:F31)	
3					Log-Likelihood Model	=SUM(E7:E31)	
4					G	=2*(F2-F3)	
5				chisq tabel	=CHIDIST(F5;1)	0	
6	no responden	X	Y	P	LL-Model	LL-Awal	
7	1	14	0	=EXP($C$1+$C$2*B7)/(1+EXP($C$1+$C$2*B7))	=C7*LN(D7)+(1-C7)*LN(1-D7)	=C7*LN($F$1)+(1-C7)*LN(1-$F$1)	
8	2	29	0	=EXP($C$1+$C$2*B8)/(1+EXP($C$1+$C$2*B8))	=C8*LN(D8)+(1-C8)*LN(1-D8)	=C8*LN($F$1)+(1-C8)*LN(1-$F$1)	
9	3	6	0	=EXP($C$1+$C$2*B9)/(1+EXP($C$1+$C$2*B9))	=C9*LN(D9)+(1-C9)*LN(1-D9)	=C9*LN($F$1)+(1-C9)*LN(1-$F$1)	
10	4	25	1	=EXP($C$1+$C$2*B10)/(1+EXP($C$1+$C$2*B10))	=C10*LN(D10)+(1-C10)*LN(1-D10)	=C10*LN($F$1)+(1-C10)*LN(1-$F$1)	
11	5	18	1	=EXP($C$1+$C$2*B11)/(1+EXP($C$1+$C$2*B11))	=C11*LN(D11)+(1-C11)*LN(1-D11)	=C11*LN($F$1)+(1-C11)*LN(1-$F$1)	
12	6	4	0	=EXP($C$1+$C$2*B12)/(1+EXP($C$1+$C$2*B12))	=C12*LN(D12)+(1-C12)*LN(1-D12)	=C12*LN($F$1)+(1-C12)*LN(1-$F$1)	
13	7	18	0	=EXP($C$1+$C$2*B13)/(1+EXP($C$1+$C$2*B13))	=C13*LN(D13)+(1-C13)*LN(1-D13)	=C13*LN($F$1)+(1-C13)*LN(1-$F$1)	
14	8	12	0	=EXP($C$1+$C$2*B14)/(1+EXP($C$1+$C$2*B14))	=C14*LN(D14)+(1-C14)*LN(1-D14)	=C14*LN($F$1)+(1-C14)*LN(1-$F$1)	
15	9	22	1	=EXP($C$1+$C$2*B15)/(1+EXP($C$1+$C$2*B15))	=C15*LN(D15)+(1-C15)*LN(1-D15)	=C15*LN($F$1)+(1-C15)*LN(1-$F$1)	
16	10	6	0	=EXP($C$1+$C$2*B16)/(1+EXP($C$1+$C$2*B16))	=C16*LN(D16)+(1-C16)*LN(1-D16)	=C16*LN($F$1)+(1-C16)*LN(1-$F$1)	
17	11	30	1	=EXP($C$1+$C$2*B17)/(1+EXP($C$1+$C$2*B17))	=C17*LN(D17)+(1-C17)*LN(1-D17)	=C17*LN($F$1)+(1-C17)*LN(1-$F$1)	
18	12	11	0	=EXP($C$1+$C$2*B18)/(1+EXP($C$1+$C$2*B18))	=C18*LN(D18)+(1-C18)*LN(1-D18)	=C18*LN($F$1)+(1-C18)*LN(1-$F$1)	
19	13	30	1	=EXP($C$1+$C$2*B19)/(1+EXP($C$1+$C$2*B19))	=C19*LN(D19)+(1-C19)*LN(1-D19)	=C19*LN($F$1)+(1-C19)*LN(1-$F$1)	
20	14	5	0	=EXP($C$1+$C$2*B20)/(1+EXP($C$1+$C$2*B20))	=C20*LN(D20)+(1-C20)*LN(1-D20)	=C20*LN($F$1)+(1-C20)*LN(1-$F$1)	
21	15	20	1	=EXP($C$1+$C$2*B21)/(1+EXP($C$1+$C$2*B21))	=C21*LN(D21)+(1-C21)*LN(1-D21)	=C21*LN($F$1)+(1-C21)*LN(1-$F$1)	
22	16	13	0	=EXP($C$1+$C$2*B22)/(1+EXP($C$1+$C$2*B22))	=C22*LN(D22)+(1-C22)*LN(1-D22)	=C22*LN($F$1)+(1-C22)*LN(1-$F$1)	
23	17	9	0	=EXP($C$1+$C$2*B23)/(1+EXP($C$1+$C$2*B23))	=C23*LN(D23)+(1-C23)*LN(1-D23)	=C23*LN($F$1)+(1-C23)*LN(1-$F$1)	
24	18	32	1	=EXP($C$1+$C$2*B24)/(1+EXP($C$1+$C$2*B24))	=C24*LN(D24)+(1-C24)*LN(1-D24)	=C24*LN($F$1)+(1-C24)*LN(1-$F$1)	
25	19	24	0	=EXP($C$1+$C$2*B25)/(1+EXP($C$1+$C$2*B25))	=C25*LN(D25)+(1-C25)*LN(1-D25)	=C25*LN($F$1)+(1-C25)*LN(1-$F$1)	

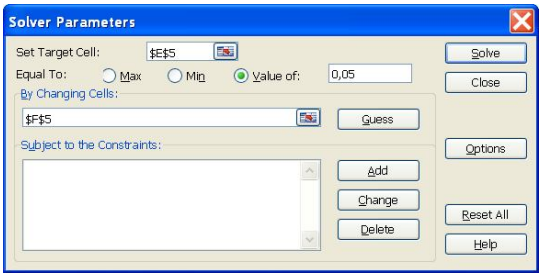
Gambar 1. Layout data dan formula yang digunakan untuk menduga koefisien regresi logistik



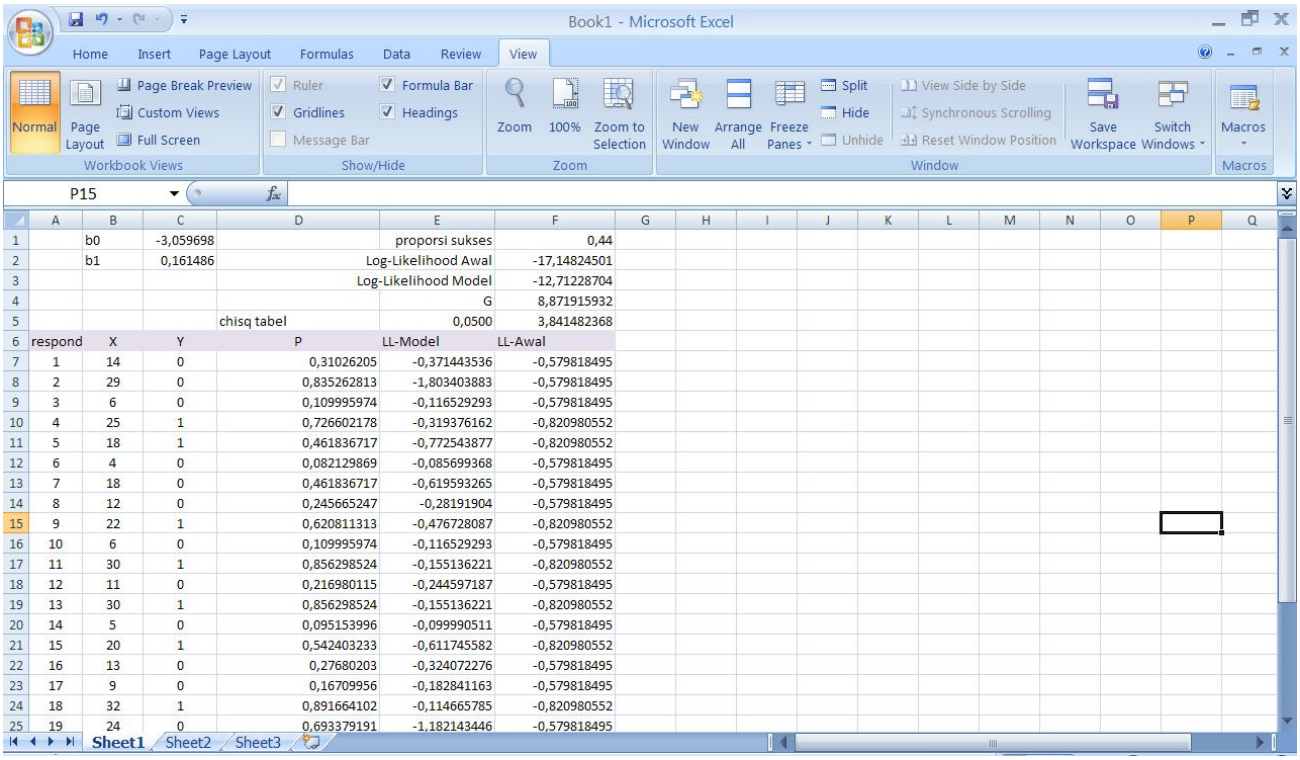
Gambar 2. Kotak dialog Solver yang harus diisi oleh pengguna untuk mendapatkan dugaan b_0 dan b_1

Begitu nilai dugaan b_0 dan b_1 diperoleh, maka kita akan memperoleh nilai LLmodel serta statistik uji G. Pada ilustrasi ini diperoleh nilai b_0 = -3.06, b_1 = 0.1615, dan G=8.872.

Proses berikutnya adalah mendapatkan nilai kritis chi-square dengan tingkat kesalahan tertentu. Dalam ilustrasi ini nilai kritis diletakkan pada sel F5 dan tingkat kesalahan pengujian pada sel E5, dengan E5 adalah fungsi dari F5. Fungsi yang digunakan adalah CHIDIST, yaitu fungsi yang menghasilkan peluang suatu titik jatuh di sebelah kanan nilai tertentu pada sebaran chi-square. Kembali Solver digunakan untuk mengoptimumkan E5. Optimum disini adalah yang mendekati nilai target tertentu (dalam ilustrasi dituliskan 0.05, lihat Gambar 3). Hasil akhir dari proses ini ditampilkan pada Gambar 4, dimana nilai kritis dengan α = 5% adalah 3.841 sehingga kita simpulkan bahwa X (pengalaman bekerja sebagai programmer) memiliki pengaruh signifikan terhadap keberhasilan membuat program.



Gambar 3. Kotak dialog untuk mendapatkan nilai kritis



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		b0	-3,059698		proporsi sukses	0,44											
2		b1	0,161486		Log-Likelihood Awal	-17,14824501											
3					Log-Likelihood Model	-12,71228704											
4					G	8,871915932											
5				chisq tabel	0,0500	3,841482368											
6	respond	X	Y	P	LL-Model	LL-Awal											
7	1	14	0	0,31026205	-0,371443536	-0,579818495											
8	2	29	0	0,835262813	-1,803403883	-0,579818495											
9	3	6	0	0,109995974	-0,116529293	-0,579818495											
10	4	25	1	0,726602178	-0,319376162	-0,820980552											
11	5	18	1	0,461836717	-0,772543877	-0,820980552											
12	6	4	0	0,082129869	-0,085699368	-0,579818495											
13	7	18	0	0,461836717	-0,619593265	-0,579818495											
14	8	12	0	0,245665247	-0,28191904	-0,579818495											
15	9	22	1	0,620811313	-0,476728087	-0,820980552											
16	10	6	0	0,109995974	-0,116529293	-0,579818495											
17	11	30	1	0,856298524	-0,155136221	-0,820980552											
18	12	11	0	0,216980115	-0,244597187	-0,579818495											
19	13	30	1	0,856298524	-0,155136221	-0,820980552											
20	14	5	0	0,095153996	-0,099990511	-0,579818495											
21	15	20	1	0,542403233	-0,611745582	-0,820980552											
22	16	13	0	0,27680203	-0,324072276	-0,579818495											
23	17	9	0	0,16709956	-0,182841163	-0,579818495											
24	18	32	1	0,891664102	-0,114665785	-0,820980552											
25	19	24	0	0,693379191	-1,182143446	-0,579818495											

Gambar 4. Hasil akhir pendugaan koefisien dan pengujiannya