Forum Statistika dan Komputasi, Oktoberl 2005, p: 12 – 16 Vol. 10 No. 2
ISSN : 0853-8115

PENDEKATAN GENERAL LINEAR MIXED MODEL PADA SMALL AREA ESTIMATION

Khairil A. Notodiputro dan Anang Kurnia

Departemen Statistika FMIPA IPB

Abstract

Small area estimation is commonly used to describe smaller domain or sub-population. Small area estimation is an important measuring instrument to estimate parameter of smaller domain borrowing strength of population parameter estimate through statistical models with random influence. In this paper we showed the contribution of statistical methods in small area estimation using general linear mixed models.

Keywords : small area estimation, general linear mixed model

PENDAHULUAN

Dalam suatu survey tingkat nasional seringkali kita dihadapkan pada masalah tingkat akurasi hasil yang berbeda antara level nasional dengan level dibawahnya baik propinsi maupun kabupaten/kota. Tingkat akurasi pada level nasional akan berlipat kali lebih baik daripada tingkat akurasi pada level kabupaten/kota. Suatu metode yang dikembangkan untuk menangani kasus tersebut adalah *small area estimation* yang merupakan himpunan dari berbagai metode statistika yang berupaya untuk memanfaatkan keakuratan/kekuatan penduga parameter pada level nasional untuk menduga parameter pada level kabupaten/kota atau priopinsi.

Small area biasanya diterjemahkan sebagai wilayah yang lebih kecil dari suatu wilayah populasi. Jika Indonesia merupakan wilayah populasi, maka propinsi atau kabupaten/ kota adalah yang dimaksud dengan *small area* karena geografi. Dalam ilmu statistik, perhatian terhadap *small area* merupakan bagian atau partisi dari wilayah populasi baik berdasarkan geografi, sosial-ekonomi, budaya atau yang lainnya.

Dalam makalah ini diperlihatkan kontribusi metode statistika dalam *small area estimation* melalui *composite estimation* dengan menggunakan konsep *general linear mixed model.*

KAJIAN PUSTAKA

Model Linear Campuran

Bentuk umum model linear campuran disajikan sebagai berikut:

$$y = X\beta + Zb + e \qquad (1)$$

X adalah matriks (n x p) dan Z berukuran (n x q), sedangkan β merupakan pengaruh tetap dan b pengaruh acak dimana e ~ N(0, Σ) serta b ~ N(0, D). Σ dan D merupakan komponen ragam yang tidak diketahui dan biasa diduga dari data dimana $\Sigma = \sigma^2 I_n$ dan $D = \sigma_b^2 I_n$.
Nilai harapan y jika b diketahui adalah

$$E(y \mid b) = X\beta + Zb \text{ dengan ragam } \Sigma \qquad (2)$$

Dari (1) *marginal distribution* bagi y adalah normal dengan nilai tengah Xβ dan ragam $V = \Sigma + ZDZ'$ sehingga log-kemungkinan bagi (β,θ) untuk $\theta = (\sigma^2, \sigma_b^2)$ adalah

$$\log L(\beta,\theta) = -\frac{1}{2}\log|V| - \frac{1}{2}(y - x\beta)' V^{-1}(y - x\beta) \qquad (3)$$

Jika θ *fixed* (tetap), penduga bagi β adalah penyelesaian dari

$$(X'V^{-1}X)\beta = X'V^{-1}y \qquad (4)$$

yang tidak lain adalah penyelesaian melalui *generalized* atau *weighted least-square.*

Perhatikan log-kemungkinan untuk seluruh parameter (β,θ,b)

$$L(\beta,\theta,b) = p(y \mid b)\, p(b) \qquad (5)$$

Berdasarkan kondisi (2) dan b ~ N(0, D), maka

$$\log L(\beta,\theta,b) = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(y - X\beta - Zb)' \Sigma^{-1}(y - X\beta - Zb) - \frac{1}{2}\log|D| - \frac{1}{2}b'D^{-1}b \qquad (6)$$

Untuk (β,θ) yang diketahui, turunan (6) terhadap b adalah

$$\frac{d\log L}{db} = Z'\Sigma^{-1}(y - X\beta - Zb) - D^{-1}b \qquad (7)$$

dan penduga bagi b adalah penyelesiaan dari

$$(Z'\Sigma^{-1}Z + D^{-1})\, b = Z'\Sigma^{-1}(y - X\beta) \qquad (8)$$

Penduga tersebut dikenal sebagai B*est Linear Unbiased Predictor* (BLUP) dan dalam prakteknya parameter-parameter yang tidak diketahui akan disubstitusi dengan penduganya sehingga kemudian disebut *Empirical Best Linear Unbiased Predictor* (EBLUP).

Perhatikan jika b kita anggap sebagai peubah acak, maka fungsi kemungkinan dapat diinterpretasikan sebagai fungsi kepekatan peluang. Dengan demikian p(b) dapat dipandang sebagai sebaran awal (*prior distribution*) bagi b sehingga *posterior distribution* bagi b akan memiliki nilai tengah $\hat{b}$ dan ragam $(Z'\Sigma^{-1}Z + D^{-1})^{-1}$, dimana $(Z'\Sigma^{-1}Z + D^{-1})^{-1} = (I(\hat{b}))^{-1}$ diperoleh dari turunan kedua (7) terhadap b yang tidak lain adalah Informasi Fisher bagi $\hat{b}$. Penurunan lengkap dapat dilihat pada Pawitan (2001). Interpretasi tersebut tidak lain adalah konsep E*mpirical Bayes* (EB).

Model Small Area Estimation

Suatu wilayah populasi W diasumsikan terdiri dari partisi-partisi wilayah (sub-populasi) W_i yang lebih kecil dan tidak saling beririsan untuk i = 1, 2, 3, ..., k. Secara umum ada tiga pendekatan untuk melakukan pendugaan parameter pada masalah *small area* : (1) *Direct estimation*, (2) *Indirect estimation*, dan (3) *Composite estimation*.

Suatu penduga bagi parameter Υ_i dari suatu sub-populasi W_i secara langsung dapat diperoleh berdasarkan anggota contoh pada sub-populasi tersebut (*direct / design-based estimator*). Menurut Ramsini et all (2001) penduga tersebut merupakan penduga tak bias tetapi memiliki ragam yang besar karena diperoleh dari ukuran contoh yang kecil. Masalah lain akan timbul apabila pada suatu sub-populasi W_i tidak terwakili didalam survey, sehingga yang mungkin dilakukan adalah pendekatan/pendugaan secara tidak langsung dan disebut sebagai *indirect estimator*, Breidt (2001). Penduga tersebut diperoleh dengan memanfaatkan informasi peubah lain yang berhubungan dengan parameter yang diamati, sehingga sering juga disebut *model-based estimator*, Ramsini et al (2001).

Dalam hal ini, dua ide utama digunakan untuk mengembangkan model untuk *small area estimation* (Saei dan Chambers, 2003) yaitu (1) asumsi bahwa keragaman didalam sub-populasi peubah respon dapat diterangkan seluruhnya oleh hubungan keragaman yang bersesuaian pada informasi tambahan, kemudian disebut model pengaruh tetap (*fixed effect*), (2) asumsi keragaman spesifik sub-populasi tidak dapat diterangkan oleh informasi tambahan dan merupakan pengaruh acak sub-populasi (*random effect*). Gabungan dari dua asumsi tersebut membentuk model pengaruh campuran (*mixed models*). Namun demikian kelemahan terjadi jika model yang dibuat tidak merepresentasikan kondisi sebenarnya, Breidt(2001).

Fay dan Herriot (1979) secara umum menggunakan persamaan (1) dengan Z hanya mengandung intersep, dengan kata lain model hanya meliputi pengaruh acak area, untuk menduga rata-rata pendapatan sub-populasi (<1000) menggunakan data sensus 1970 di Amerika Serikat.

Model Fay-Herriot tersebut merupakan model dasar bagi pengembangan pemodelan *small area* yaitu $y_i = \omega_i + e_i$; $\omega_i = x_i'\beta + \upsilon_i$, dimana e_i dan υ_i saling bebas dengan $E(e_i) = E(\upsilon_i) = 0$ serta $Var(e_i) = \Sigma_i$ dan $Var(\upsilon_i) = D$ (i = 1, 2, 3, ..., k). Russo et.al (2005 menjabarkan lebih lanjut model *small area* dengan memperjelas pengaruh acak sub-populasi di dalam model.

1. $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$
   vektor data pendukung
2. $\omega_i = \mathbf{x}_i'\beta + z_i\upsilon_i$ untuk i = 1, 2, ..., k
   merupakan parameter yang menjadi perhatian dan diasumsikan memiliki hubungan dengan data pendukung pada (1) sedang υ_i pengaruh acak dengan nilai tengah nol dan ragam σ^2_υ.
3. $\hat{\omega}_i = \omega_i + e_i$
   direct estimate untuk sub-populasi ke-i dengan sampling error
4. $\hat{\omega}_i = \mathbf{x}_i'\beta + z_i\upsilon_i + e_i$ untuk i = 1, 2, ..., k
   model tersebut terdiri dari pengaruh acak dan pengaruh tetap sehingga merupakan bentuk *general linear mixed model* dengan struktur peragam yang diagonal.

Model regresi merupakan upaya untuk membentuk model umum dan memanfatkan kekuatan dan keakuratan pendugaan pada level populasi, sedangkan deviasi sub-populasi untuk menangkap kekhasan yang terjadi pada setiap sub-populasi dan bersifat acak. Dengan demikian jika kita hanya akan memanfaatkan informasi

umum maka $\omega_i = x_i'\beta$, dan jika pengaruh umum dan lokal kita adopsi, diperoleh $\omega_i = x_i'\beta + \upsilon_i$.

Secara statistika model pada point (4) diatas melibatkan pengaruh acak akibat desain sampling (*designed-induced,* e_i) dan pengaruh acak pemodelan sub-populasi (*model-based,* υ_i) serta model tersebut merupakan bentuk khusus *general linear mixed model*. Solusi BLUP merupakan rataan terboboti dari *design-based estimator* ($\hat{\omega}_i$), dan *regression-synthetic estimator* ($x_i'\tilde{\beta}$) serta dinotasikan sebagai berikut :

$$\tilde{Y}_i = \gamma_i \hat{Y}_i + (1-\gamma_i)\, x_i'\tilde{\beta} \qquad (9)$$

dimana $\gamma_i = \dfrac{\sigma^2_v}{\sigma^2_v + \sigma^2_{ei}}$.

Model tersebut dapat dikatakan mengambil keuntungan dengan menggabungkan keragaman (deviasi) sub-populasi dan ketepatan pendugaan berdasarkan *direct estimation* (Rao, 2003).

PENERAPAN PADA DATA BPS

Untuk meningkatkan akurasi perencanaan dan pengendalian pembangunan suatu wilayah sudah barang tentu diperlukan data pendukung yang akurat. Dalam ilustrasi ini disajikan peta kemiskinan di Propinsi Jawa Barat dengan kabupaten/ kota sebagai sub-populasi yang menjadi perhatian.

Peubah yang diamati dan menjadi perhatian dalam ilustrasi ini adalah tingkat kemiskinan yang didefinisikan sebagai rasio jumlah keluarga miskin terhadap total jumlah keluarga. Suatu keluarga dikatakan miskin jika pengeluaran perkapitanya dibawah Rp. 114.000,- (*poverty line* 2002). Sedangkan sumber data yang digunakan adalah SUSENAS 2003 dengan materi informasi berbasis rumah tangga.

Dalam model *small area estimation*, tingkat kemiskinan merupakan peubah respon atau yang akan diduga dan peubah bebasnya adalah peubah-peubah yang diasumsikan mempengaruhi dan atau menggambarkan tingkat kemiskinan, seperti:

1. proporsi rumah bukan milik sendiri
2. proporsi atap terluas rumah bukan beton/genteng
3. proporsi jenis dinding rumah terluas bukan tembok
4. proporsi jenis lantai rumah terluas tanah
5. luas lantai per jumlah anggota rumah tangga
6. proporsi sumber air minum terbuka (sumur tak terlindung, mata air, air sungai, air hujan dan lainnya)
7. proporsi penggunaan fasilitas air minum tidak sendiri
8. proporsi penggunaan fasilitas buang air besar tidak sendiri
9. proporsi daya PLN terpasang 450 VA atau tanpa meteran
10. rata-rata pengeluaran non makanan sebulan per kapita.

Langkah-langkah pendugaan parameter dapat disederhanakan sebagai berikut:

1. definiskan model $y_i = X_i\beta + Z_ib + e_i$
2. $\hat{y}_i = y_i + e_i$, dimana $\hat{y}_i$ adalah *design-based estimator* yang diperoleh secara langsung berdasarkan data dan desain survey.
3. perbaiki model pada (1) menjadi $\hat{y}_i = X_i\beta + Z_ib + e_i$
4. tentukan besaran penduga tingkat kemiskinan

$$\tilde{y}_i = \gamma_i \hat{y}_i + (1-\gamma_i)\, x_i'\tilde{\beta}$$

dengan $\gamma_i = \dfrac{\sigma^2_v}{\sigma^2_v + \sigma^2_{ei}}$

Hasil analisis diperoleh tingkat kemiskinan di Propinsi Jawa Barat berdasarkan *design-based estimator* adalah 10.30% dengan RSE 2.21%. RSE yang relatif lebih kecil jika dibadingkan dengan RSE untuk kabupaten/ kota menggambarkan tingkat keakuratan yang lebih baik pada level propinsi dibandingkan dengan level kabupaten/ kota. Tabel 1 menyajikan pendugaan tingkat kemiskinan pada kabupaten/kota di Provinsi Jawa Barat berdasarkan data SUSENAS tahun 2003.

Secara umum penduga berdasarkan *composite estimation* menghasilkan RSE yang lebih kecil kecuali untuk Kota Depok, Kota Bekasi, Kabupaten Bekasi, dan Kabupaten Karawang, hasil sejalan juga diperlihatrkan oleh penduga berbasis model. Hal tersebut mungkin disebabkan oleh model yang kurang bagus dalam menggambarkan kondisi kabupaten/kota tersebut karena keheterogenan dan keterbatasan data. Selain itu, yang juga menjadi perhatian adalah RSE Kota Depok dan Kota Bekasi diatas 70%, suatu statistik yang sangat besar dalam suatu pendugaan parameter, yang secara aljabar diduga karena kecilnya penduga tingkat kemiskinan di dua wilayah tersebut. Interpretasi lain yang bisa diambil, menjelaskan bahwa wilayah Jawa Barat bagian selatan relatif masih tinggi tingkat kemiskinannya.

Tabel 1. Pendugaan tingkat kemiskinan (dalam persen) berdasarkan *design-based*, *model-based* dan *composite estimator*

Kabupaten/Kota		Design-Based Estimator		Model-Based Estimator		Composite Estimator	
		Statistik	RSE	Statistik	RSE	Statistik	RSE
3201	Kab. Bogor	10.63	8.49	15.53	5.82	15.09	5.00
3202	Kab. Sukabumi	16.52	7.46	15.80	5.87	15.87	4.90
3203	Kab. Cianjur	20.35	6.51	19.61	6.25	19.68	5.20
3204	Kab. Bandung	10.77	7.84	11.43	7.05	11.37	5.92
3205	Kab. Garut	26.90	5.51	20.01	6.18	20.63	5.01
3206	Kab. Tasikmalaya	18.97	6.79	22.95	6.57	22.59	5.57
3207	Kab. Ciamis	19.56	6.90	17.42	5.94	17.62	4.92
3208	Kab. Kuningan	10.80	11.37	9.11	8.74	9.27	7.22
3209	Kab. Cirebon	17.13	7.48	12.44	7.48	12.87	6.06
3210	Kab. Majalengka	14.39	9.21	13.27	5.80	13.37	4.84
3211	Kab. Sumedang	6.27	15.12	9.80	8.38	9.48	7.25
3212	Kab. Indramayu	3.61	17.92	6.31	19.86	6.07	17.17
3213	Kab. Subang	8.98	11.49	7.98	10.75	8.07	8.89
3214	Kab. Purwakarta	4.77	16.50	7.08	11.60	6.87	9.98
3215	Kab. Karawang	6.12	13.70	8.66	26.56	8.43	22.64
3216	Kab. Bekasi	2.63	21.54	3.64	43.28	3.54	36.84
3271	Kota Bogor	2.47	25.50	3.85	27.41	3.73	23.56
3272	Kota Sukabumi	2.71	27.36	4.00	25.82	3.88	22.14
3273	Kota Bandung	0.74	37.66	4.49	21.00	4.15	18.84
3274	Kota Cirebon	5.63	18.70	5.06	18.84	5.11	15.59
3275	Kota Bekasi	1.10	33.16	1.55	88.70	1.51	75.54
3276	Kota Depok	1.22	35.13	1.65	86.17	1.61	73.18

PENUTUP

Metode berbasis model dan *composite* yang digunakan untuk pendugaan tingkat kemiskinan dalam makalah ini memiliki potensi untuk dapat dipercaya dengan dukungan RSE yang lebih kecil jika dibandingkan dengan pendugaan yang dilakukan secara langsung. Namun demikian, pengembangan dan perbaikan metodologi serta pemodelan untuk meningkatkan keakuratan pendugaan masih harus dikembangkan.

Ilustrasi dalam makalah ini mengasumsikan kelinearan model serta pembatasan peubah karakteristik rumah tangga berdasarkan data SUSENAS. Pengembangan model serta memperluas cakupan peubah bisa dilakukan untuk memperbaiki pendugaan. Namun demikian, ketersediaan, kelengkapan dan kualitas data juga perlu menjadi perhatian.

DAFTAR PUSTAKA

Breidt, F.J., (2004), "Small Area Estimation for Natural Resource Surveys", <http://www.stat.colostate.edu/~nsu/starmap/pps/breidt.msts.pdf>, [28 April 2005]

Fay, R.E. and Herriot, R.A., (1979), "Estimates of income for small places: an application of James-Stein procedures to Census data". *Journal of the American Statistical Association*, Vol. 74, p.269-277.

Pawitan, Y., (2001), *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford : Clarendon Press.

Ramsini, B. et.all, (2001), "Uninsured Estimates by County: A Review of Options and Issues",

<www.odh.ohio.gov/Data/OFHSurv/ofhsrfq7.pdf>, [25 Mei 2005]

Rao, J.N.K., (2003), *Small Area Estimation*, New York : John Wiley and Sons.

Russo, C., M. Sabbatini dan R. Salvatore, (2005), "General Linear Models in Small Area Estimation : an assessment in agricultural surveys", Paper presented in The Mexsai Conference. <www.siap.sagarpa.gob.mx/mexsai/trabajos/t44.pdf>, [29 April 2005]

Saei, A. dan R. Chambers, (2003), "Small Area Estimation: A Review of Methods Based on the Application of Mixed Models", S^3RI Methodologi Working Paper M03/16, University of Southampton, UK.

⊗