Statistika, Vol. 8 No. 1, 31 – 36
Mei 2008

Parameter *Quantile-like* dalam Pendugaan Area Kecil Melalui Pendekatan *Penalized-Splines*

KUSMAN SADIK

Tenaga Pengajar di Departemen Statistika IPB, Bogor
Jl. Meranti, Kampus IPB Darmaga, Bogor 16680, Tlp./fax: (0251) 624535

ABSTRAK

Pada beberapa tahun terakhir ini, para statistisi mulai mengembangkan metodologi yang berkaitan dengan pendugaan untuk daerah atau domain survei yang memiliki sampel kecil atau bahkan tidak memiliki sampel satupun. Data yang diperoleh melalui teknik survei yang tepat akan sangat efektif dan memiliki sifat reliabilitas untuk menduga total atau rataan peubah tertentu. Sifat penduga yang demikian dapat dicapai apabila data sampel dari survei mencakup daerah atau domain yang besar. Misalnya, beberapa survei ekonomi yang dilakukan di Indonesia berskala nasional. Pada survei yang demikian banyaknya sampel rumah tangga untuk tiap kecamatan dalam suatu kabupaten sangat kecil (small area). Bahkan bisa terjadi suatu kecamatan tertentu tidak terpilih sebagai daerah survei sehingga sampel rumah tangga dari kecamatan tersebut tidak ada. Persoalannya adalah bagaimana menduga parameter, misalnya tingkat kemiskinan di level kecamatan tersebut sementara sampelnya sangat kecil. Salah satu metode yang banyak dikembangkan untuk pendugaan area kecil (small area estimation / SAE) adalah model yang berbasis pada generalized linear mixed model (GLMM). Beberapa pendekatan lain saat ini mulai didiskusikan oleh para statistisi di dunia. Salah satu metode alternatif tersebut adalah pemodelan yang didasarkan pada kuantil yang dikenal dengan M-quantile P-splines. Aspek penting dari metode ini adalah adanya sifat tegar (robust) terhadap pencilan (outliers) dan bebas sebaran (distribution free).

Kata Kunci: general linear mixed model, empirical best linear unbiased prediction, M-quantile regression, robust estimation, penalized splines

1. Pendahuluan

Survei menjadi salah satu bagian penting dari proses pengambilan keputusan yang berbasis pada data. Sehingga survei sudah dilakukan baik di lembaga penelitian swasta maupun negeri. Bahkan kebijakan publik suatu negara sangat dipengaruhi oleh data-data hasil survei.

Sangat beragam persoalan yang ditemui dalam survei. Namun demikian, ada dua topik utama yang menjadi perhatian para statistisi selama tahun-tahun terakhir ini. Topik tersebut menyangkut persoalan pengembangan teknik penarikan sampel (*sampling technique*) dan pengembangan metodologi pendugaan parameter pupulasi (*estimation methods*).

Biasanya statistik diperoleh dari suatu survei yang didisain untuk memperoleh statistik nasional. Artinya survei semacam ini didisain untuk inferensia bagi daerah (*domain*) yang luas. Persoalan muncul ketika dari survei seperti ini ingin diperoleh informasi untuk area yang lebih kecil, misalnya informasi pada level propinsi, kabupaten, bahkan mungkin level kecamatan.

Statistik area kecil (*small area statistic*) telah menjadi perhatian para statistisi dunia secara sangat serius sejak sepuluh tahun terakhir ini (misalnya Dol, 1991; Ghosh and Rao, 1994; Chand dan Alexander, 1995; Carlin, 1998; dan Rao, 2003). Berbagai metode pendugaan area kecil (*small area estimation*) telah dikembangkan khususnya menyangkut metode yang berbasis model (*model-based area estimation*).

Pendugaan langsung umumnya didasarkan pada teknik penarikan sampelnya (*sampling technique*). Teknik semacam ini telah dikembangkan oleh Cochran (1977), Sewnson dan

Wretman (1992), dan Thompson (1997). Metode yang didasrak pada pemodelan (*model-based*) juga telah dikembangkan, misalnya seperti yang dilakukan oleh Dorfman dan Royall (2001).

Pada pendugaan yang berbasis pada rancangan survei (*design-based*), pembobot rancangan $w_j(s)$ memiliki peranan penting dalam membentuk penduga berbasis rancangan $\hat{Y}$ bagi *Y*. Pembobot ini tergantung pada s dan elemen *j* ($j \in s$). Salah satu bentuk pembobot yang penting adalah $w_j(s)=1/\pi_j$, dimana $\pi_j = \Sigma_{\{s:j\in s\}} p(s)$, $j=1, 2, .., N$. Apabila tidak informasi penyerta (auxiliary information), maka penduga langsung dapat diekspresikan sebagai $\hat{Y} = \Sigma_{\{s:j\in s\}} w_j(s) y_j$. Dalam kasus ini, rancangan tidak berbias apabila terpenuhi $\Sigma_{\{s:j\in s\}} p(s) w_j(s) = 1$ untuk $j=1, 2, \ldots, N$. Pembobot ini merupakan bentuk umum dari penduga Horvitz-Thompson (Cochran, 1977).

2. Pendugaan Area Kecil (*Small Area Estimation*)

Pendugaan area kecil merupakan konsep terpenting dalam pendugaan parameter secara tidak langsung di suatu area yang relatif kecil dalam persampelan survei (*survey sampling*). Metode pendugaan area kecil digunakan untuk menduga karakteristik dari subpopulasi (domain yang lebih kecil). Pendugaan langsung (*direct estimation*) pada suppopulasi tidak memiliki presisi yang memadai karena kecilnya jumlah sampel yang digunakan untuk memperoleh dugaan tersebut.

Alternatif metode lain adalah dengan cara menghubungkan area tersebut dengan area lain melalui model yang tepat. Dengan demikian dugaan tersebut merupakan dugaan tidak langsung (*indirect estimation*), dalam arti bahwa dugaan tersebut mencakup data dari domain yang lain. Rao (2003) menyebutkan bahwa prosedur pendugaan area kecil pada dasarnya memanfaatkan kekuatan area sekitarnya (*neighbouring areas*) dan sumber data diluar area yang statistiknya ingin diperoleh.

Pendugaan tidak langsung dapat menggunakan pendekatan model secara umum. Misalkan diasumsikan bahwa $\theta_i = g(\overline{Y}_i)$ untuk beberapa spesifikasi g(.) dihubungkan dengan data penyerta spesifik pada area i, $z_i = (z_{1i}, \ldots, z_{pi})^T$ melalui suatu model linear

$$\theta_i = z_i^T\beta + b_i v_i, \qquad i = 1, \ldots, m$$

dimana b_i adalah konstanta positif yang diketahui dan β adalah vektor berukuran px1. Sedangkan v_i adalah pengaruh acak (random effects) spesifikasi area yang diasumsikan bebas dan menyebar identik (*independent and identically distributed*, *iid*) dengan

$E_m(v_i) = 0$ dan $V_m(v_i) = \sigma_v^2$ (≥ 0), atau $v_i \sim$ iid $(0, \sigma_v^2)$

Pendugaan tidak langsung untuk rataan populasi di area kecil i, ($\overline{Y}_i$), diperlukan informasi mengenai penduga langsungnya yaitu $\hat{\overline{Y}}$. Dengan menggunakan metode James-Stein akan diperoleh:

$$\hat{\theta}_i = g(\hat{\overline{Y}}) = \theta_i + e_i$$

$$\hat{\theta}_i = z_i^T\beta + b_i v_i + e_i, \quad i = 1, \ldots, m$$

dimana galat penarikan sampel (sampling error) ei adalah bebas dengan

$E_p(e_i \mid \theta_i) = 0$ dan $V_p(e_i \mid \theta_i) = \psi_i$, atau $v_i \sim$ iid $(0, \sigma_v^2)$

3. General Linear Mixed Model

Rao (2003) mengaitkan model-model di atas sebagai bagian dari *general linear mixed model* (GLMM) yang menggabungkan antara pengaruh tetap (*fixed effects*) dan pengaruh acak (*random effects*) dalam suatu model umum. Datta dan Ghosh (1991) mengemukakan formulasi model GLMM sebagai berikut :

$$\mathbf{y}^P = \mathbf{X}^P\boldsymbol{\beta} + \mathbf{Z}^P\mathbf{v} + \mathbf{e}^P$$

Pada model ini $\mathbf{v}$ dan $\mathbf{e}^P$ bebas dengan $\mathbf{e}^P \sim \mathbf{N}(\mathbf{0}, \sigma^2\boldsymbol{\psi}^P)$ dan $\mathbf{v} \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{D}(\boldsymbol{\lambda}))$, dimana $\boldsymbol{\psi}^P$ adalah matrik definit positif yang diketahui dan $\mathbf{D}(\boldsymbol{\lambda})$ adalah matrik definit positif yang strukturnya diketahui. Sedangkan $\mathbf{X}^P$ dan $\mathbf{Z}^P$ adalah matrik rancangan dan $\mathbf{Y}^P$ adalah vektor *N*

x 1 dari nilai y populasi. Matrik koragam bagi **v** dan **e** masing-masing adalah **G** dan **R**. Persamaan di atas dapat pula ditulis sebagai berikut:

$$\mathbf{y}^{P} = \begin{bmatrix} \mathbf{y} \\ \mathbf{y}* \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \mathbf{X}* \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z} \\ \mathbf{Z}* \end{bmatrix} \mathbf{v} + \begin{bmatrix} \mathbf{e} \\ \mathbf{e}* \end{bmatrix}$$

dimana bagian yang ditandai *asterisk* (*) menunjukkan unit yang tidak tercakup dalam sampel (*nonsampled*). Vektor untuk total (Y_i) pada area kecil adalah berbentuk **Ay** + **Cy*** dengan **A** = $\oplus_{i=1}^{m} \mathbf{1}_{n_i}^{T}$ dan **C** = $\oplus_{i=1}^{m} \mathbf{1}_{N_i - n_i}^{T}$ dimana $\oplus_{i=1}^{m} \mathbf{A}_u$ = blockdiag($\mathbf{A}_1$, …, $\mathbf{A}_m$).

Pada GLMM ini dilakukan pendugaan terhadap kombinasi linear dari parameter yaitu μ = $\mathbf{1}^T\boldsymbol{\beta}$ + $\mathbf{m}^T\mathbf{v}$. Rao (2003) mengemukakan bahwa untuk $\boldsymbol{\delta}$ tertentu yang diketahui maka penduga BLUP (*best linear unbiased prediction*) bagi μ adalah

$$\tilde{\mu}^{H} = t(\boldsymbol{\delta}, \mathbf{y}) = \mathbf{1}^T \tilde{\boldsymbol{\beta}} + \mathbf{m}^T \tilde{\mathbf{v}} = \mathbf{1}^T \tilde{\boldsymbol{\beta}} + \mathbf{m}^T \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$$

dimana

$$\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\boldsymbol{\delta}) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$

$$\tilde{\mathbf{v}} = \tilde{\mathbf{v}}(\boldsymbol{\delta}) = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$$

Model untuk pendugaan tidak langsung, yaitu $\hat{\theta}_i = \mathbf{z}_i^T\boldsymbol{\beta} + b_i v_i + e_i$, $i = 1, \ldots, m$, sebenarnya merupakan kasus khusus dari model GLMM, yaitu

$$\mathbf{y}_i = \hat{\theta}_i, \quad \mathbf{X}_i = \mathbf{z}_i T, \quad \mathbf{Z}_i = b_i$$

dan

$$\mathbf{v}_i = v_i, \quad \mathbf{e}_i = e_i, \quad \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$$

sedangkan

$$\mathbf{G}_i = \sigma_v^2, \quad \mathbf{R}_i = \psi_i$$

sehingga

$$\mathbf{V}_i = \psi_i + \sigma_v^2 b_i^2 \text{ dan } \mu_i = \theta_i = \mathbf{z}_i^T\boldsymbol{\beta} + b_i v_i$$

Apabila persamaan pendugaan tidak langsung disubstitusikan ke dalam pendugaan GLMM akan diperoleh penduga BLUP bagi μ_i atau θ_i yaitu:

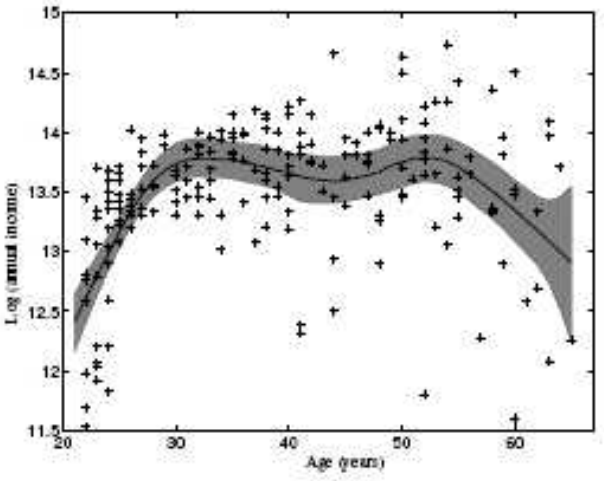
$\tilde{\theta}_i^{H} = \mathbf{z}_i^T\tilde{\boldsymbol{\beta}} + \gamma_i(\hat{\theta}_i - \mathbf{z}_i^T\tilde{\boldsymbol{\beta}})$, dimana $\gamma_i = \sigma_v^2 b_i^2 / (\psi_i + \sigma_v^2 b_i^2)$, dan

$$\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\sigma_v^2) = \left[\sum_{i=1}^{m} \frac{\mathbf{z}_i \mathbf{z}_i^T}{\psi_i + \sigma_v^2 b_i^2}\right]^{-1} \left[\sum_{i=1}^{m} \frac{\mathbf{z}_i \hat{\theta}_i}{\psi_i + \sigma_v^2 b_i^2}\right]$$

4. M-*quantile* P-*Splines* dalam Pendugaan Area Kecil

Chambers dan Tzavidis (2006) mengusulkan suatu pendekatan baru untuk pendugaan area kecil didasarkan pada parameter *quantile-like* pada sebaran bersyarat dari peubah yang menjadi perhatian untuk beberapa kovariat yang diberikan. Model nonparametrik ini dimungkinkan dapat memberikan keuntungan penting apabila bentuk fungsional hubungan antar peubah yang menjadi perhatian dan kovariatnya adalah tidak linier. Spesifikasi yang salah tentang model dapat menyebabkan bias dalam pendugaan parameter area yang kecil.

Pengembangan M-*quantile* dalam pendugaan area kecil merupakan salah satu pendekatan model apabila bentuk fungsional dari hubungan antar peubah dan kovariatnya tidak dapat dispesifikasikan. Sementara regresi *Penalized-spline*, sering dikenal sebagai P-*splines*, adalah suatu metoda nonparametrik yang akhir-akhir ini cukup populer karena fleksibilitasnya (lihat Ruppert, Wand, dan Carrol, 2003). P-*splines* juga mulai didiskusikan sebagai salah satu metode alternatif dalam pendugaan area kecil yang didasarkan pada model pengaruh campuran/*mixed effects models* (Opsomer *et al.*, 2005).



Gambar 1. Hubungan antar Peubah yang tidak Linier

Pada *generalized linear mixed model* mengasumsikan bahwa keragaman yang berhubungan dengan sebaran bersyarat bagi *y* jika diketahui vektor kovariat ***x*** bisa dinyatakan sebagai struktur hirarki yang sebelumnya telah dispesifikasikan. Suatu pendekatan alternatif ke pemodelan keragaman dari sebaran bersyarat ini adalah melalui regresi linier M-*quantile* yang tidak tergantung pada struktur hirarki data (Chambers dan Tzavidis, 2006). Regresi M-*quantile* mengintegrasikan konsep regresi *quantile* dan regresi *expectile* di dalam suatu kerangka umum yang didefinisikan oleh suatu generalisasi "*quantile-like*" dari regresi berdasarkan pada fungsi pengaruh. Pratesi *et al.* (2006) mengembangkan regresi M-*quantile* untuk suatu kasus dimana hubungan antara peubah yang diminati (*y*) dengan peubah kovariatnya (*x*) tidak linier melalui P-*splines* dan digunakan untuk pendugaan area kecil.

Misalkan digunakan kovariat tunggal yaitu *x*. Suatu model P-*spline* untuk kuantil bersyarat ke-q bagi y apabila diberikan x adalah:

$$Q_q(x,\psi) = \beta_{0\psi}(q) + \beta_{1\psi}(q)x + ... + \beta_{p\psi}(q)x^p + \sum_{k=1}^{K}\beta_{(p+k)\psi}(q)(x-\kappa_k)_+^p$$

dimana ψ adalah fungsi pengaruh yang dispesifikasi, $(t)_+^p = t^p$ jika t > 0 dan 0 untuk selainnya, *p* adalah derajat dari *spline*, dan κ_k untuk k = 1, ..., K adalah gugus simpul (*knots*), nilai simpul K dipilih besar dan pengaruh simpul diletakkan sebagai pembatas bagi ukuran koefisien *spline*. Suatu versi *penalized* dari M-penduga general bisa digunakan untuk memperoleh penduga:

$$\sum_{i=1}^{n}\rho(y_i - Q_q(x_i,\psi)) + (\frac{\lambda}{2})\sum_{k=1}^{K}\beta_{(p+k)\psi}^2$$

dimana fungsi ρ memberikan kontribusi pada masing-masing sisaan pada fungsi tujuan, λ adalah pengganda Lagrange yang mengendalikan taraf pemulusan pada hasil fitting, dan n adalah banyaknya unit sampel. Penduga bagi parameter regresi nonparametrik model M-*quantile* bisa diperoleh melalui pemecahan persamaan:

$$\sum_{i=1}^{n}\psi_q(y_i - \mathbf{x}_i\boldsymbol{\beta})\mathbf{x}_i + \lambda\sum_{k=1}^{K}\beta_{(p+k)\psi} = \mathbf{0}$$

menggunakan *iteratively reweighted penalized least squares* (IRPLS). Penduga rataan untuk area kecil j dapat dinyatakan sebagai berikut:

$$\hat{\bar{y}}_j = \int t d\hat{F}_{CD,j}(t) = \frac{1}{N_j}\left(\sum_{i\in s_{n_j}} y_i + \sum_{i\in r_j}\mathbf{x}_i\hat{\boldsymbol{\beta}}_\psi(\hat{\bar{q}}_j) + \frac{N_j - n_j}{n_j}\sum_{i\in s_{n_j}}(y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}_\psi(\hat{\bar{q}}_j))\right)$$

dimana $\hat{F}_{CD,j}(t)$ adalah dugaan fungsi sebaran kumulatif untuk masing-masing area kecil, $\hat{\bar{q}}_j$ adalah nilai rataan koefisien sampel M-*quantile* untuk semua unit dalam area j, s_{nj} dan r_j dinotasikan sebagai sampel dan non-sampel dalam area j, dan N_j adalah ukuran populasi dalam area j. Nilai yi yang tidak teramati untuk unit populasi $i \in r_j$ diprediksi menggunakan $\mathbf{x}_i\hat{\boldsymbol{\beta}}_\psi(\hat{\bar{q}}_j)$.

5. Contoh Kasus

Sebagai salah satu contoh kasus digunakan data simulasi untuk melihat beberapa karakteristik dari hasil pendugaannya. Pada simulasi ini digunakan 20 area (i = 1, 2, ..., 20), dan masing-masing diulang sebanyak 100 kali. Didalamnya menggunakan 3 (tiga) kovariat. Hasil pendugaan area kecil melalui *Generalized Linear Mixed Model* (GLMM) dan M-*quantile* P-*spline* terdapat pada Tabel 1.

Tabel 1. Hasil Pendugaan Melalui GLMM dan M-*quantile* P-*spline*

Area	Nilai Sebenarnya (θ_i)	GLMM		M-q P-s	
		$\hat{\theta}_i$	StdErr	$\hat{\theta}_i$	StdErr
1	30.514	30.431	0.850	30.621	0.812
2	30.025	29.999	0.850	30.160	0.811
3	25.356	25.999	0.838	26.081	0.804
4	27.127	27.450	0.839	27.441	0.821
5	31.382	30.916	0.844	31.041	0.810
6	26.218	26.745	0.840	26.752	0.801
7	29.241	29.090	0.844	29.082	0.802
8	27.637	27.840	0.852	27.799	0.828
9	30.865	30.645	0.837	30.617	0.800
10	30.846	30.362	0.855	30.305	0.804
11	28.660	28.686	0.836	28.476	0.807
12	31.029	30.793	0.846	31.016	0.827
13	32.272	31.790	0.847	31.706	0.808
14	27.129	27.583	0.845	27.878	0.810
15	29.918	29.779	0.835	29.494	0.813
16	32.612	32.222	0.861	32.147	0.832
17	30.732	30.461	0.837	30.540	0.811
18	31.501	31.175	0.868	31.178	0.828
19	26.243	26.691	0.848	26.376	0.825
20	25.464	26.117	0.838	26.062	0.806
			0.846		0.813

StdErr : *Standard Errror*

Berdasarkan hasil pada Tabel 1, pendugaan melalui M-*quantile* P-*spline* menghasilkan *standard error* yang lebih kecil dibandingkan dengan pendugaan melalui GLMM. Hal ini mungkin disebabkan karena adanya pencilan (*outlier*) pada kovariatnya. Artinya, hasil pendugaan melalui M-*quantile* P-*spline* bersifat lebih tegar (*robust*) dibandingkan GLMM.

6. Kesimpulan

Metode pendugaan area kecil (*small area estimation*) dapat digunakan untuk meningkatkan keakuratan pendugaan dengan cara meningkatkan efisiensi penggunaan sampel melalui fungsi hubung (*link function*) antara penduga langsung dengan pengaruh tetap dan pengaruh acak pada suatu area tertentu.

Metode M-*quantile* P-*spline* dapat dijadikan sebagai salah satu metode alternatif disamping GLMM, khususnya apabila bentuk fungsional dari hubungan antar peubah dan kovariatnya tidak dapat dispesifikasikan. Atau hubungannya tidak bersifat linier, sementara dalam GLMM hubungan tersebut diasumsikan linier.

Daftar Pustaka

[1]. Chambers, R., Tzavidis, N. 2006. *M-quantile Models for Small Area Estimation*, forthcoming in *Biometrika*.
[2]. Rao, J. N. K. 2003. *Small Area Estimation*. John Wiley & Sons, Inc. New Jersey.
[3]. Rao, J. N. K., dan Yu, M. 1994. *Small Area Estimation by Combining Time Series and Cross-Sectional Data*. Proceedings of the Section on Survey Research Method. American Statistical Association.
[4]. Ruppert, R., M. Wand, dan R. Caroll 2003. *Semiparametric Regression*. Cambridge University Press.
[5]. Swenson, B., dan Wretman., J. H. 1989. *The Weighted Regression Technique foe Estimating the Variance of Generalized Regression Estimator*. Biometrika, **76**, 527-537.
[6]. Thompson, M. E. 1997. *Theory of Sample Surveys*. London: Chapman and Hall.