Kajian Metode Estimasi Parameter dalam Regresi Semiparametrik Spline

Wahyu Wibowo[1], Sri Haryatmi[1], I Nyoman Budiantara[2]
wahyu.stk@gmail.com
[1]Jurusan Matematika, Universitas Gadjah Mada Yogyakarta
[2]Jurusan Statistika, Institut Teknologi Sepuluh Nopember Surabaya

Intisari

Pada regresi semiparametrik spline, estimasi kurva regresi dapat diselesaikan dengan metode kuadrat terkecil, kuadrat terkecil terpenalti, dan reproducing kernel Hilbert space. Masing-masing pendekatan memiliki karakteristik yang berbeda. Pada pendekatan kuadrat terkecil, masalah utama adalah memilih banyak knot dan lokasinya. Pada kuadrat terkecil terpenalti dan reproducing kernel mempunyai permasalahan yang sama, yaitu menentukan nilai parameter penghalus yang optimal. Namun, secara matematis, metode reproducing kernel memberi hasil yang lebih umum dibandingkan dengan kuadrat terkecil terpenalti karena berlaku untuk semua derajat polinomial spline yang akan dipergunakan. Makalah ini akan menjelaskan penggunaan metode kuadrat terkecil, kuadrat terkecil terpenalti, dan reproducing kernel hilbert space serta perbedaan masing-masing dalam estimasi kurva regresi semiparametrik spline.

Kata kunci : *Regresi semiparametrik, least square, penalized least square, reproducing kernel Hilbert space, spline*

Abstract

Curve estimation for spline semiparametric regression can be done by using least square, penalized least square and reproducing kernel Hilbert space method. Each methods has different characteristics. Least square method has problem about how to choose the number of knots and their location. Penalized least square and reproducing kernel has the same problem about how to choose the optimal smoothing parameter. However,reproducing kernel approach is more general mathematically than penalized least square due to be valid for any degree of polynomial spline that to be used. This paper will explaine about least square, penalized least square, reproducing kernel Hilbert space as weel as difference of each method in curve estimation for spline semiparametric regression.

Keyword : *semiparametric regression, least square, penalized least square, reproducing kernel Hilbert space, spline*

1. Pendahuluan

Regresi semiparametrik merupakan kombinasi antara regresi parametrik dan regresi nonparametrik. Kombinasi dalam hal ini dimaksudkan bahwa dalam regresi semiparametrik memuat sekaligus model regresi parametrik dan model regresi nonparametrik. Regresi semiparametrik ini muncul karena adanya kasus-kasus pemodelan yang hubungan antar variabelnya selain ada yang linear juga ada yang tidak diketahui bentuknya. Keberadaan dua komponen yang berbeda dalam regresi semiparametrik ini menjadikan pemakaian model ini menjadi luas dan secara teori berkembang sangat pesat. Perkembangan ini selain karena aplikasi juga karena berkembangnya perangkat keras teknologi komputasi yang mempermudah dan mempercepat komputasi.

Estimasi kurva regresi semiparametrik spline dapat diselesaikan dengan metode kuadrat terkecil, kuadrat terkecil terpenalti, dan reproducing kernel hilbert space. Masing-masing pendekatan memiliki karakteristik yang berbeda. Pada pendekatan kuadrat terkecil, masalah utama adalah memilih banyak knot dan lokasinya. Pada kuadrat terkecil terpenalti dan reproducing kernel mempunyai permasalahan sama, yaitu menentukan nilai parameter penghalus yang optimal. Namun, secara matematis, metode reproducing kernel memberi hasil yang lebih umum dibandingkan dengan kuadrat terkecil terpenalti karena berlaku untuk semua derajat polinomial spline yang akan dipergunakan

Makalah ini akan menjelaskan penggunaan metode kuadrat terkecil, kuadrat terkecil terpenalti, dan reproducing kernel hilbert space serta perbedaan masing-masing dalam estimasi kurva regresi semiparametrik spline. Penjelasan akan dimulai dengan konstruksi model regresi semiparametrik spline, metode kuadrat terkecil, kuadrat terkecil terpenalti, dan reproducing kernel Hilbert space. Sebagai penutup akan diberikan kesimpulan berkaitan dengan metode-metode tersebut.

2. Model

Pandang n sampel random dengan variabel pada masing-masing sampel adalah (y_i, x_i, t_i), $i = 1,2,...,n$. Dalam hal ini diasumsikan y_i sebagai variabel respon, x_i sebagai variabel prediktor yang diketahui berpengaruh linear, dan t_i sebagai variabel prediktor yang tidak diketahui bentuk pengaruhnya terhadap respon. Selanjutnya dibentuk model regresi semiparametrik

$$y_i = \beta_0 + \beta_1 x_i + f(t_i) + \varepsilon_i, \quad i = 1,2,...,n \tag{1}$$

(β_0, β_1) parameter untuk komponen parametrik, dan f komponen nonparametrik, dalam hal ini merupakan fungsi yang tidak diketahui. $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ adalah error random yang saling independen dengan mean nol dan variansi σ^2.

Model (1) dapat dinyatakan dalam notasi matrik menjadi

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\varepsilon} \tag{2}$$

dimana $\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{1n} \end{pmatrix}$, $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$, $\mathbf{f} = \begin{pmatrix} f(t_1) \\ \vdots \\ f(t_n) \end{pmatrix}$, $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$

Masalah estimasi pada regresi semiparametrik muncul karena adanya komponen nonparametrik berupa fungsi *f* yang tidak diketahui bentuknya. Oleh karena itu, hampiran terhadap bentuk fungsi tersebut dapat dilakukan dengan lebih dari satu bentuk fungsi. Beberapa diantaranya adalah spline, kernel, fourier, wavelet, dan polinomial lokal. Secara aplikasi, hampiran-hampiran ini memiliki kelebihan yang berbeda. Sebagai contoh, hampiran spline sangat cocok untuk data yang memiliki perilaku berubah-ubah dalam sub-sub interval tertentu. Hampiran fourier sangat cocok untuk data yang memiliki pola berulang atau musiman. Penjelasan mengenai macam-macam hampiran ini dapat dilihat pada Eggermont dan Lariccia (2009).

3. Metode Kuadrat Terkecil

Metode kuadrat terkecil merupakan metode yang sangat lazim dipergunakan dalam regresi linear. Prinsip metode ini adalah meminimumkan kuadrat residual. Metode ini juga bisa dipergunakan dalam regresi semiparametrik spline. Wibowo, dkk (2009, 2010) telah menggunakan metode ini untuk estimasi parameter pada regresi semiparametrik dan juga

sifat-sifat statistik estimator yang diperoleh. Penggunaan metode ini mensyaratkan bentuk spesifik fungsi *f(t)* dalam model (1). Bentuk fungsi spline yang biasa dipergunakan adalah fungsi basis spline polinomial *truncated*.

Fungsi spline polinomial *truncated* derajat *p* dengan *k* titik knots $\kappa_1, \kappa_2, ..., \kappa_k$ disajikan dalam bentuk,

$$f(t_i) = \alpha_0 + \alpha_1 t_i + \alpha_1 t_i^2 + ... + \alpha_p t_i^{\,p} + \alpha_{p1}(t_i - \kappa_1)_+^p + ... + \alpha_{pk}(t_i - \kappa_k)_+^p \tag{3}$$

dengan $(t_i - \kappa)_+^p = \begin{cases} (t_i - \kappa)^p, & t_i \geq \kappa \\ 0 \quad\quad , & t_i < \kappa \end{cases}$

Sehingga model (1) dapat dinyatakan menjadi

$$y_i = \beta_0 + \beta_1 x_i + \alpha_0 + \alpha_1 t + \alpha_2 t^2 + ... + \alpha_p t^p + \alpha_{p1}(t - \kappa_1)_+^p + ... + \alpha_{pk}(t - \kappa_k)_+^p + \varepsilon_i, \quad i = 1, 2, ..., n \tag{4}$$

Apabila dinyatakan dalam matriks, diperoleh

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{5}$$

dengan

$$\mathbf{Z} = \begin{bmatrix} 1 & x & \cdots & x^p & (t-\kappa_1)_+^p & \cdots & (t-\kappa_k)_+^p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x & \cdots & x^p & (t-\kappa_1)_+^p & \cdots & (t-\kappa_1)_+^p \end{bmatrix}_{n\times(1+p+k)} \qquad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_0 & \alpha_1 & \cdots & \alpha_p & \alpha_{p1} & \cdots & \alpha_{pk} \end{bmatrix}^{\mathrm{T}}_{1\mathrm{x}(1+p+k)}$$

Selanjutnya, ditentukan $\mathbf{C} = [\mathbf{x} \;\; \mathbf{z}]$, $\boldsymbol{\omega} = \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix}$, sehingga (5) dapat dinyatakan dengan

$$\mathbf{Y} = \mathbf{C}\boldsymbol{\omega} + \boldsymbol{\varepsilon}$$

Selanjutnya didefinisikan kuadrat residual sebagai berikut,

$$\begin{aligned} \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon} &= (\mathbf{Y} - \mathbf{C}\boldsymbol{\omega})^{\mathrm{T}}(\mathbf{Y} - \mathbf{C}\boldsymbol{\omega}) \\ &= \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - 2\boldsymbol{\omega}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}\mathbf{Y} + \boldsymbol{\omega}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}\mathbf{C}\boldsymbol{\omega} \end{aligned} \tag{6}$$

Estimator $\boldsymbol{\omega}$ diperoleh dengan cara meminimumkan (6) terhadap $\boldsymbol{\omega}$. Selanjutnya (6) diturunkan terhadap $\boldsymbol{\omega}$ dan disamadengankan nol, sehingga diperoleh :

$$\begin{aligned} \frac{\partial \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}}{\partial \boldsymbol{\omega}} &= -2\mathbf{C}^{\mathrm{T}}\mathbf{Y} + 2\mathbf{C}^{\mathrm{T}}\mathbf{C}\boldsymbol{\omega} = \mathbf{0} \\ \mathbf{C}^{\mathrm{T}}\mathbf{C}\boldsymbol{\omega} &= \mathbf{C}^{\mathrm{T}}\mathbf{Y} \end{aligned} \tag{7}$$

yang merupakan persamaan normal. Penyelesaian persamaan ini akan merupakan estimator $\boldsymbol{\omega}$. Sesuai aljabar matrik, karena matrik $\mathbf{C}$ mempunyai rank (3 + *p* + *k*) dan $\mathbf{C}^{\mathrm{T}}\mathbf{C}$ matrik *positive-definite*, maka $\mathbf{C}^{\mathrm{T}}\mathbf{C}$ akan merupakan matrik nonsingular. Sehingga persamaan (6) akan mempunyai penyelesaian tunggal,

$$\hat{\boldsymbol{\omega}} = (\mathbf{C}^{\mathrm{T}}\mathbf{C})^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{Y} \tag{8}$$

Dalam hal ini, $\hat{\boldsymbol{\omega}}$ merupakan estimator kuadrat terkecil $\boldsymbol{\omega}$. Mengingat persamaan (2), estimator (8) berlaku hanya untuk derajat polinomial *p* dan banyak knots *k* yang tertentu. Sehingga lebih tepat kalau estimator ini dinyatakan dengan

$$\hat{\boldsymbol{\omega}}_{p,\kappa} = (\mathbf{C}^{\mathrm{T}}\mathbf{C})^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{Y}$$

$$\hat{\mathbf{Y}} = C\hat{\omega}_{p,\kappa} = \mathbf{C}(\mathbf{C}^{\mathrm{T}}\mathbf{C})^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{Y}$$
$$\hat{\mathbf{Y}} = \mathbf{H}(p;\kappa_1,...,\kappa_k)\mathbf{Y}$$

dengan $\mathbf{H}(p;\kappa_1,...,\kappa_k) = \mathbf{C}(\mathbf{C}^{\mathrm{T}}\mathbf{C})^{-1}\mathbf{C}^{\mathrm{T}}$

Permasalahan selanjutnya adalah bagaimana menentukan derajat polinomial *p* dan banyak knots *k* yang akan digunakan dalam estimator. Untuk keperluan ini akan dipergunakan kriteria *Generalized Cross Validation* (GCV), yang didefinisikan :

$$GCV(p;\kappa_1,\kappa_2,...,\kappa_k) = \frac{n^{-1}\left\|(\mathbf{I}-\mathbf{H}(p;\kappa_1,\kappa_2,...,\kappa_k)\mathbf{y}\right\|^2}{\left[n^{-1}trace(\mathbf{I}-\mathbf{H}(p;\kappa_1,\kappa_2,...,\kappa_k)\right]^2}$$

Nilai *p* dan *k* dipilih dengan menyelesaikan optimasi

$GCV(p_{opt};\kappa_{1_{opt}},\kappa_{2_{opt}},...,\kappa_{k_{opt}}) = \min_{\substack{p\in\Re^+ \\ \kappa_1\in\Re^+ \\ \vdots \\ \kappa_k\in\Re^+}} \left(GCV(p;\kappa_1,\kappa_2,...,\kappa_k)\right)$ Metode estimasi kuadrat terkecil secara matematis dapat diselesaikan dengan langkah-langkah yang sederhana dan menghasilkan model statistik yang mudah diinterpretasikan. Kesulitan metode ini adalah menentukan derajat polinomial, banyak knot dan lokasi knot-knot tersebut.

4. Metode Kuadrat Terkecil Terpenalti

Metode kuadrat terkecil terpenalti merupakan perluasan metode kuadrat terkecil dengan menambahkan parameter penghalus dan penalti pada fungsi yang akan dipergunakan. Fungsi yang akan dipergunakan merupakan keluarga fungsi yang terdifferensial pada interval [*a,b*] dan kontinu absolut pada turunan pertama. Bentuknya diberikan sebagai berikut,

$$S(\beta, f) = \frac{1}{n}\sum_{i=1}^{n}\{y_i - (\beta_0 + x_i\beta_1) - f(t_i)\}^2 + \lambda\int_a^b\{f^{(m)}(t)\}^2 dt \qquad (9)$$

Dalam hal ini λ merupakan parameter penghalus, sedangkan penalti diberikan oleh

$$\int_a^b\{f^{(m)}(t)\}^2 dt$$

Jika $\lambda \to 0$, maka hasil estimasi mendekati hasil metode kuadrat terkecil. Sebaliknya, jika $\lambda \to \infty$, maka estimasi akan menginterpolasi titik-titik data. Estimator terbaik merupakan kompromi antara nilai jumlah kuadrat residual dan parameter penghalus λ yang bisa didapatkan dengan meminimumkan nilai GCV.

Dalam Grenn dan Silverman (1994) dinyatakan bahwa fungsi yang meminimumkan (9) merupakan fungsi *natural cubic spline*, yang diberikan sebagai berikut :

$$f(t) = a_i + b_i(t-t_i) + c_i(t-t_i)^2 + d_i(t-t_i)^3,\ t_i \le t \le t_{i+1},\ i = 1,2,...,n$$

$t_1, t_2, \cdots, t_n$ merupakan knot yang memenuhi $a < t_1 < t_1 < \cdots < t_n < b$. Selanjutnya, penalti dalam (9) akan dinyatakan dalam bentuk nilai turunan kedua dengan langkah sebagai berikut.

1. Didefinisikan $f_i = f(t_i)$ dan $\gamma_i = f''(t_i)$, $i = 1,2,\cdots,n$.
2. Pandang $\mathbf{f} = (f_1,\cdots,f_n)^T$, $\boldsymbol{\gamma} = (\gamma_2,\cdots,\gamma_{n-1})^T$, $h_i = t_{i+1} - t_i$, $i = 1,2,\cdots,n-1$
3. Susun matrik, namakan **Q** dan **R,** dengan ketentuan sebagai berikut**.**

$\mathbf{Q}$ adalah matrik berukuran $n\times(n-2)$ dengan element q_{ij}, $i=1,2,\cdots,n$ dan $j=2,3,\cdots,n-1$.

$q_{j-1,j}=h_{j-1}^{-1}$, $q_{ij}=-h_{j-1}^{-1}-h_j^{-1}$, $q_{j+1,j}=h_j^{-1}$ $j=2,3,\cdots,n-1$, $q_{ij}=0$, jika $|i-j|\geq 2$.

Matriks $\mathbf{Q}$ is diindeks mulai $j=2$, sehingga elemen teratas $\mathbf{Q}$ adalah q_{12}.

Selanjutnya R adalah matriks symmetric berukuran $(n-2)\times(n-2)$ dengan elements r_{ij}, i,j=2,3,…,$(n-1)$ sebagai berikut

$$r_{ij}=\tfrac{1}{3}(h_{i-1}+h_i),\ r_{i,j+1}=r_{i+1,i}=\tfrac{1}{6}h_i, \text{ dan } r_{ij}=0 \text{ untuk } |i-j|\geq 2$$

4. Definiskan matrik $\mathbf{K}$ dengan

$$\mathbf{K}=\mathbf{Q}\mathbf{R}^{-1}\mathbf{Q}^{\mathbf{T}}$$

Selanjutnya hubungan antara $\mathbf{Q,R,f,\gamma}$ and $\mathbf{K}$ terdapat dalam pada Green dan Silverman (1994), yang dinyatakan dalam teorema berikut.

Theorem 1:

Vektor $\mathbf{f}$ dan $\boldsymbol{\gamma}$ menentukan *natural cubic spline* f jika dan hanya jika memenuhi

$$\mathbf{Q}^T\mathbf{f}=\mathbf{R}\boldsymbol{\gamma}$$

Selanjutnya penalti akan dapat dinyatakan menjadi

$$\int_a^b\{f''(t)\}^2dt=\boldsymbol{\gamma}^T\mathbf{R}\boldsymbol{\gamma}=\mathbf{f}^T\mathbf{K}\mathbf{f}$$

Bukti dapat dilihat pada Green dan Silverman (1994). Sehingga (9) dapat dinyatakan dalam notasi matrik menjadi ; $S(\boldsymbol{\beta},\mathbf{f})=(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{f})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{f})+\lambda\mathbf{f}^TK\mathbf{f}$

Sebagai hasil teorema ini, jumlah kuadrat terpenalti (9) dapat dinyatakan

$$S(\boldsymbol{\beta},\mathbf{f})=n^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{f})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{f})+\lambda\mathbf{f}^T\mathbf{K}\mathbf{f} \qquad (10)$$

Meminimumkan (9) equivalent dengan meminimumkan,

$$S(\boldsymbol{\beta},\mathbf{f})=(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{f})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{f})+\lambda\mathbf{f}^TK\mathbf{f} \qquad (11)$$

Selanjutnya, dengan menurunkan (10) terhadap $\boldsymbol{\beta}$ and $\mathbf{f}$, dan menyamadengankan nol, diperoleh

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{Y}-\hat{\mathbf{f}})$$
$$\hat{\mathbf{f}}=(\mathbf{I}+\lambda\mathbf{K})^{-1}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}) \qquad (12)$$

Dalam hal ini terhadap dua smoother, $\mathbf{S}_\beta=\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ menghasilkan dugaan $\mathbf{X}\hat{\boldsymbol{\beta}}$ dan smoother lainnya adalah $\mathbf{S}_f=(\mathbf{I}+\lambda\mathbf{K})^{-1}$ menghasilkan dugaan $\hat{\mathbf{f}}$. Selanjutnya dugaan parametrik diberikan oleh $\mathbf{S}_\beta(\mathbf{Y}-\mathbf{f})=\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{Y}-\mathbf{f})\equiv\mathbf{X}\hat{\boldsymbol{\beta}}$ sedangkan dugaan untuk nonparametrik adalah $\hat{\mathbf{f}}=\mathbf{S}_f(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})$. Substitusi dugaan nonparametrik ke bagian pertama (11), menghasilkan

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}=\mathbf{X}^T(\mathbf{Y}-\mathbf{S}_f(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}))\ \Leftrightarrow\ \mathbf{X}^T(\mathbf{I}-\mathbf{S}_f)\mathbf{X}\boldsymbol{\beta}=\mathbf{X}^T(\mathbf{I}-\mathbf{S}_f)\mathbf{Y}$$

Persamaan ini adalah persamaan normal untuk *generalized least square* normal, dengan elemen non-diagonal berupa matrik pembobot $(\mathbf{I}-\mathbf{S}_f)$. Sehingga, parameter $\boldsymbol{\beta}$ dan $\mathbf{f}$ dapat diselesaikan dengan,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T(\mathbf{I}-\mathbf{S}_f)\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I}-\mathbf{S}_f)\mathbf{Y}$$
$$\hat{\mathbf{f}} = \mathbf{S}_f(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}) = -\mathbf{S}_f\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{S}_f\mathbf{Y}$$

Estimator penalized least square masih tergantung pada parameter smoothing λ, oleh karena itu harus dipilih yang optimum dengan meminimumkan GCV. GCV didefinisikan dalam metode ini diformulakan dengan

$$GCV(\lambda) = \frac{\left\|\mathbf{Y}-\hat{\mathbf{Y}}\right\|^2}{(1-n^{-1}tr(\mathbf{A}))^2}$$

A merupakan *hat* matrik yang memenuhi

$$\mathbf{AY} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{f}} = \begin{bmatrix}\mathbf{X} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{f}}\end{bmatrix} = \hat{\mathbf{Y}}$$

dengan

$$\mathbf{A} = \begin{bmatrix}\mathbf{X} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{X}^T\mathbf{X} & \mathbf{X}^T\\ \mathbf{X} & (\mathbf{I}+\lambda\mathbf{K})\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}^T\\ \mathbf{I}\end{bmatrix}$$

Bentuk kurva regresi hasil metode kuadrat terkecil terpenalti kuadrat terkecil bergantung pada parameter penghalus λ. Namun metode ini menjadi lebih komplek jika terdapat lebih dari satu prediktor nonparametrik. Selain itu, interpretasi pengaruh prediktor nonparametrik terhadap respon tidak diberikan secara eskplisit melalui model statistik, akan tetapi melalui kurva regresi.

5. Reproducing Kernel Hilbert Space

Penggunaan metode reproducing kernel Hilbert space dalam regresi semiparametrik pada dasarnya merupakan perluasan metode kuadrat terkecil terpenalti dengan menggunakan fungsi yang terdifferensial pada interval $[a,b]$ dan turunannya yang ke-m kontinu absolut pada interval tersebut. Dengan kata lain, fungsi tersebut termuat di dalam ruang *Sobolev* $W_2^m[a,b]$ dengan $W_2^m([a,b]) = \{f : f, \cdots, f^{(m-1)} \text{ kont. abs.}, \int_a^b \left(f^{(m)}(t)\right)^2 dt < \infty\}$ *Reproducing Kernel Hilbert Space* (RKHS) H_R adalah suatu ruang Hilbert dari fungsi bernilai real pada [0,1] dengan sifat bahwa untuk setiap $t \in [0,1]$, fungsional $L_t(f) = f(t)$ merupakan fungsional linear terbatas, dalam arti bahwa terdapat M sedemikian hingga berlaku

$$|L_t f| = |f(t)| \le M\|f\|$$

H_R dapat didekomposisi menjadi $H_R = H_0 \oplus H_1$ dengan H_0 ruang *Null*, dan H_1 adalah ruang yang tegak lurus dengan ruang Null.

Reproducing kernel dari H_R adalah fungsi R yang didefinisikan pada $[0,1]\times[0,1]$, sedemikian hingga untuk setiap titik $t \in [0,1]$ berlaku $R_t \in H_R$ dengan $R_t(s) = R(t,s)$ dan $L_t f = \langle R_t, f\rangle$, $f \in H_R$

Untuk menggunakan metode reproducing kernel dalam estimasi parameter regresi semiparametrik, model (1) dinyatakan menjadi

$$y_i = \beta_0 + \beta_1 x_i + L_{t_i} f + \varepsilon_i, \quad i = 1,2,...,n \tag{13}$$

dengan $L_{t_i} f = \langle R_{t_i}, f\rangle = f(t_i)$

Selanjutnya, estimasi parameter diperoleh dengan meminimumkan penalized least square

$$S(\beta, f) = \frac{1}{n}\sum_{i=1}^{n}\{y_i - (\beta_0 + x_i\beta_1) - L_{t_i} f\}^2 + \lambda\|P_1 f\|_R^2 \quad (14)$$

dengan $\int_a^b \{f^{(m)}(t)\}^2 dt = \|P_1 f\|_{W_m}^2$, P_1 proyeksi orthogonal f pada H_1 dalam H_R.

Estimator yang akan diperoleh terdiri dari estimator spline parsial sebagai komponen nonparametrik (**f**), dan estimator parametrik (**β**). Estimator spline parsial diperoleh berdasarkan teorema berikut.

Teorema 2 :

Apabila $H_R = H_0 \oplus H_1$ dan $\phi_1, \phi_2, \ldots, \phi_m$ merupakan basis di ruang H_0, serta $\mathbf{T}_{n\times m}$ merupakan matrik *full rank* berorde $n \times m$ yang diberikan oleh

$\mathbf{T}_{n\times m} = \{L_i\phi_v\}$, $i = 1, 2, \ldots, n$; $v = 1, 2, \ldots, m$

maka fungsi f yang meminumkan

$$S(\beta, f) = \frac{1}{n}\sum_{i=1}^{n}\{y_i - (\beta_0 + x_i\beta_1) - L_i f\}^2 + \lambda\|P_1 f\|_R^2$$

adalah

$$\hat{\mathbf{f}}_\lambda = \sum_{v=1}^{m} d_v\phi_v + \sum_{i=1}^{n} c_i\varphi_i \quad (15)$$

dengan

$\varphi_i = P_1\eta_i$

$\mathbf{d} = (d_1, d_2, \ldots, d_m)' = \mathbf{(T'M^{-1}T)^{-1}T'M^{-1}(Y - X\beta)}$ $\mathbf{c} = (c_1, c_2, \ldots, c_n)' = \mathbf{M^{-1}(I - T(T'M^{-1}T))^{-1}T'M^{-1}(Y - X\beta)}$

$\mathbf{M} = \Sigma + n\lambda$, $\Sigma = \{\langle \varphi_i, \varphi_j \rangle\}$; $i, j = 1, 2, \ldots, n$

Sedangkan estimator parametrik yang bersesuaian adalah

$$\hat{\boldsymbol{\beta}}_\lambda = \mathbf{(X'(I - A(\lambda))X)^{-1}X'(I - A(\lambda))'Y} \quad (16)$$

Bukti diberikan pada akhir makalah. Sampai di sini, telah diperoleh estimator untuk regresi semiparametrik spline dengan metode (RKHS). Nilai taksiran untuk variabel respon dapat dinyatakan menjadi

$$\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta}_\lambda + \mathbf{f}_\lambda \quad (17)$$

Kelebihan estimator yang diperoleh dengan metode RKHS adalah keumumannya yang berlaku untuk sebarang polinomial spline sampai derajat-*m*. Interpretasi pengaruh prediktor nonparametrik terhadap respon dapat dengan mudah dilakukan melalui kurva regresi. Jika terdapat lebih dari satu prediktor nonparametrik, maka akan muncul kesulitan secara matematis dan grafis.

6. Kesimpulan dan Saran

Masalah utama pada estimasi parameter regresi semiparametrik adalah adanya komponen nonparametrik berupa fungsi yang tidak diketahui bentuknya. Penggunaan metode kuadrat terkecil mengasumsikan bentuk fungsi spline polinomial truncated dan memberikan kemudahan interpretasi melalui model statistik. Penggunaan kuadrat terkecil terpenalti menghasilkan fungsi *natural cubic spline* sebagai komponen nonparametrik dan bentuk kurva regresinya tergantung pada parameter penghalus. Estimator yang diperoleh dengan reproducing kernel Hilbert space juga tergantung pda parameter penghalus, namun

estimatornya berlaku untuk sebarang derajat polinomial spline. Sehingga estimator yang diperoleh dengan pendekatan reproducing kernel Hilbert space bersifat lebih umum dibandingkan dengan estimator yang diperoleh dengan metode kuadrat terkecil maupun kuadrat terkecil terpenalti.

Daftar Pustaka

Green, P.J. and Silverman, B.W., 1994, *Nonparametric Regression and Generalized Linear Model*, Chapman & Hall, London

Eggermont, P.P.B., and Lariccia, V.N., 2009, *Maximum Penalized Likelihood Estimation, Volume II : Regression*, Springer Series in Statistics

Wahba, G., 1990, *Spline Model for Observational Data*, SIAM, XII, Philadelphia

Wibowo, W., Haryatmi, S., Budiantara, I.N., 2009, Least Square Methods for Estimating Curve of Spline Semiparametric Regression, *Proceeding of National Seminar on Mathematic and Mathematic Education*, Yogyakarta State University, December 5th 2009, p. 633-645, ISBN : 978-979-16353-3-2

Wibowo, W., Haryatmi, S., Budiantara, I.N., 2010, Inference And Confidence Interval For Regression Curve In Spline Semiparametric Model, *Proceeding of National Seminar on Mathematic and Mathematic Education*, University of Muhammadiyah Malang, January 30th , 2010

Bukti Teorema 2 :

Untuk bukti estimator nonparametrik, dapat dilihat pada Wahba (1990), sedangkan untuk estimator parametrik diberikan secara singkat.

Untuk mendapatkan estimator parametrik, persamaan (15) dinyatakan dalam bentuk matrik menjadi

$$\mathbf{f}_\lambda = \mathbf{Td} + \sum \mathbf{c}$$

Selanjutnya, dengan mengingat hubungannya dengan model (2), maka (15) dapat dinyatakan juga menjadi

$$\mathbf{f}_\lambda = \mathbf{Y - X\beta - Mc} + \sum \mathbf{c}$$

Dalam hal ini, $\mathbf{f}_\lambda$ dapat dilihat sebagai hasil smoothing terhadap $\mathbf{Y} - \mathbf{X\beta}$, sehingga secara umum menjadi

$$\mathbf{f}_\lambda = \mathbf{A}(\lambda)(\mathbf{Y - X\beta})$$

$$\begin{aligned}\mathbf{f}_\lambda &= \mathbf{Y - X\beta} - (\sum + n\lambda - \sum)\mathbf{c} \\ &= \mathbf{Y - X\beta} - n\lambda\mathbf{c}\end{aligned}$$

$$\begin{aligned}n\lambda\mathbf{c} &= \mathbf{Y - X\beta} - \mathbf{f}_\lambda \\ &= \mathbf{Y - X\beta} - \mathbf{A}(\lambda)(\mathbf{Y - X\beta}) \\ &= (\mathbf{I - A}(\lambda))(\mathbf{Y - X\beta})\end{aligned}$$

Di lain pihak,

$$\begin{aligned}\sum \mathbf{c} &= \mathbf{f}_\lambda - \mathbf{Td} \\ &= \mathbf{A}(\lambda)(\mathbf{Y - X\beta}) - \mathbf{Td}\end{aligned}$$

$$\begin{aligned}n\lambda\mathbf{c}'\sum \mathbf{c} &= n\lambda\mathbf{c}'\mathbf{A}(\lambda)(\mathbf{Y - X\beta}) \\ &= (\mathbf{I - A}(\lambda))(\mathbf{Y - X\beta})\mathbf{A}(\lambda)(\mathbf{Y - X\beta})\end{aligned}$$

$$S(\boldsymbol{\beta}, \mathbf{f}_\lambda) = (\mathbf{Y} - \mathbf{X\beta} - \mathbf{f}_\lambda)^T(\mathbf{Y} - \mathbf{X\beta} - \mathbf{f}_\lambda) + n\lambda\mathbf{c}'\sum \mathbf{c}$$

$$= (\mathbf{Y} - \mathbf{X\beta})^{\mathrm{T}} (\mathbf{I} - \mathbf{A}(\lambda))^{\mathrm{T}} (\mathbf{I} - \mathbf{A}(\lambda))(\mathbf{Y} - \mathbf{X\beta}) + n\lambda \mathbf{c}' \sum \mathbf{c} \quad = (\mathbf{Y} - \mathbf{X\beta})^{\mathrm{T}} (\mathbf{I} - \mathbf{A}(\lambda))^{\mathrm{T}} (\mathbf{I} - \mathbf{A}(\lambda))(\mathbf{Y} - \mathbf{X\beta}) + (\mathbf{I} - \mathbf{A}(\lambda))(\mathbf{Y} - \mathbf{X\beta})\mathbf{A}(\lambda)(\mathbf{Y} - \mathbf{X\beta})$$

$$\frac{S(\boldsymbol{\beta}, \mathbf{f}_\lambda)}{\partial \boldsymbol{\beta}} = 0$$

$$\boldsymbol{\beta}_\lambda = (\mathbf{X}'(\mathbf{I} - \mathbf{A}(\lambda))\mathbf{X})^{-1} \mathbf{X}'(\mathbf{I} - \mathbf{A}(\lambda))'\mathbf{Y}$$