Jurnal Intersections
Volume 5, No. 2, Agustus 2020
P-ISSN: 2685-7952
Website: https://jurnal.ucy.ac.id/index.php/intersections

ESTIMASI KURVA REGRESI SPLINE PADA DATA LONGITUDINAL DENGAN METODE KUADRAT TERKECIL

Toto Hermawan
Universitas Cokroaminoto Yogyakarta
totohermawan@mail.ugm.acid

Abstrak

Artikel ini mengkaji tentang estimasi regresi spline khususnya penggunaan pada data longitudinal. Data longitudinal adalah data yang diperoleh berdasarkan pangamatan yang dilakukan sebanyak n objek yang saling independen dengan setiap objek diamati secara berulang dalam kurun waktu yang berbeda dan antara pengamatan dalam objek yang sama adalah dependen selain itu data longitudinal adalah data yang mampu membedakan keragaman respon yang disebabkan karena pengukuran yang berulang.Kurva regresi spline diestimasi dengan menggunakan kuadrat terkecil. Terlihat bahwa taksiran kurva regresi spline untuk data longitudinal merupakan kelas pendugaan linear dalam observasi respon $\tilde{y}_i$ dan sangat tergantung pada titik knot $k_1, k_2, \ldots, k_n$

Kata Kunci : data longitudinal, kuadrat terkecil, dan regresi spline

1. PENDAHULUAN

Regresi nonparametrik digunakan apabila bentuk kurva regresi diasumsikan tidak diketahui. Regresi nonparametrik memiliki fleksibilitas yang tinggi dalam mengestimasi kurva regresi. Berbeda dengan regresi parametrik yang mengasumsikan bentuk kurva regresi diketahui seperti linear, kuadratik, kubik, eksponensial atau yang lainnya, pendekatan regresi nonparametrik tidak mengasumsikan bentuk awal dari kurva regresi. Sehingga diperlukan pendekatan dalam mengestimasi kurva regresi nonparametrik,salah satunya adalah metode spline. Eubank [4] menyatakan spline merupakan salah satu model yang mempunyai interpretasi statistik dan interpretasi visual sangat khusus dan sangat baik, melalui pemilihan titik knot optimal Di samping itu, spline mampu menangani karakter data fungsi yang bersifat mulus (smooth) melalui pemilihan parameter penghalus optimal.

Islamiyati (2010) menguraikan penggunaan regresi spline polynomial truncuted pada data cross sectional. Namun dalam perkembangan riset selama ini, telah banyak jenis data pengukuran yang diperoleh di lapangan, bukan hanya dalam bentuk cross sectional, diantaranya data longitudinal. Wu dan Zhang [9] menyatakan data longitudinal adalah data pengamatan yang dilakukan terhadap n obyek yang saling independen, setiap obyek diamati secara berulang dan kontinu dalam kurun waktu tertentu, dimana pengamatan dalam obyek yang sama saling berkorelasi. Perbedaan struktur data tersebut menyebabkan perlu kajian tentang penggunaan regresi spline pada data longitudinal.

Tulisan ini mengkaji tentang estimasi kurva regresi spline pada data longitudinal, dimana metode estimasi yang digunakan adalah metode kuadrat terkecil, dengan memilih titik knot optimal berdasarkan nilai Gross Cross Validation (GCV) terkecil.

2. DATA LONGITUDINAL

Data longitudinal adalah data yang didapatkan dari hasil pengukuran berulang pada beberapa individu (unit cross-sectional) dalam waktu berturut-turut (Diggle, 2002). Salah satu tujuan penelitian menggunakan data longitudinal adalah meneliti apakah ada pengaruh variabel penjelas terhadap variabel respon, termasuk meneliti pengaruh variabel penjelas pada besarnya perubahan variabel respon. Sehingga pada dasarnya analisis data longitudinal adalah analisis regresi pada data longitudinal. Secara umum data longitudinal dapat disajikan pada tabel dibawah ini (Danardono, 2012).

Untuk mempermudah dalam menganalisis data longitudinal, maka data longitudinal direpresentasikan dalam bentuk formulasi matematis. Berikut adalah notasi yang diperlukan dalam pemodelan data longitudinal (Diggle, 2002).Ada sebanyak i = 1, ..., m individu, yang masing-masing memiliki pengamatan berulang j=1,2,...,n_i. Dengan banyaknya pengamatan berulang untuk individu tidak harus sama. Sehingga secara total ada $N = \sum_{i=1}^{m} n_i$ observasi. Waktu observasi aktual, yaitu saat pengamatan terjadi dinotasikan dengan t_{ij}. Variabel respon dinyatakan sebagai Y_{ij} dengan nilai observasi y_{ij}. Dapat pula dinyatakan sebagai

$$Y_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{bmatrix} \quad (1)$$

Atau $Y_i = (Y_{i1} \quad \dots \quad Y_{in})^t$ dengan nilai observasi $y_i = (y_{i1} \quad \dots \quad y_{in})^t$. Variabel penjelas dinyatakan sebagai

$$X_i = \begin{bmatrix} x_{i11} & \cdots & x_{i1p} \\ x_{i21} & \cdots & x_{i1p} \\ \vdots & \ddots & \vdots \\ x_{in1} & \cdots & x_{inp} \end{bmatrix}$$

Suatu matriks berukuran n_i x p. dengan p adalah banyaknya kovariat. Untuk menunjukkan ke satu observasi untuk semua nilai kovariat, dapat ditulis sebagai vektor $x_i = (x_{ij1} \quad \cdots \quad x_{ijp})^t$ suatu vektor berukuran 1 x p. Variabel penjelas dalam data longitudinal dapat diamati sekali saja dan nilainya sama sampai akhir studi, yang sering dinamakan kovariat awal *(baseline covariat).* Variabel penjelas dapat pula diamati lebih dari satu kali selama waktu studi berjalan, atau sering disebut kovariat bergantung waktu *(time varying covariate).* Sebagai contoh variabel jenis kelamin, usia saat studi dimulai merupakan baseline covariate. Sedangkan usia dan indeks depresi merupakan time varying covariat. Mean variabel respon Y_{ij} adalah $E(Y_{ij}) = \mu_{ij}$ sedangkan bila variabel respon ditulis sebagai, meannya $E(Y_i) = \mu_i$. untuk individu *i*, variansi dari Y_i berupa matriks kovariansi berukuran n_i x n_i

$$\mathrm{Var}(\boldsymbol{Y}_i) = \begin{bmatrix} v_{i11} & \cdots & v_{i1n_i} \\ \cdots & v_{ijk} & \cdots \\ v_{in_i1} & \cdots & v_{in_in_i} \end{bmatrix}$$

dengan $v_{ijk} = Cov(\boldsymbol{Y}_{ij}, \boldsymbol{Y}_{ik})$

Yang membedakan model regresi data longitudinal dengan model regresi biasa *(cross-sectional)* adalah adanya korelasi antar variabel, sehingga pemodelan kovariansi atau dependensi terhadap waktu untuk pengukuran berulang dalam satu individu menjadi hal yang penting. Model untuk kovariansi juga tidak lepas dari model untuk mean nya. Suatu model untuk kovariansi seharusnya dipilih berdasarkan model tertentu mean respon, karena model kovariansi tersebut merepresentasikan kovariansi antara residual sebagai akibat dari pemilihan model tertentu untuk meannya. Dikenal berbagai bentuk kovariansi dalam model longitudinal. Di antaranya sudah diperkenalkan pada pembahasan tentang model sederhana untuk data longgitudinal yaitu

1) Compound Symmetry
2) Banded
3) Autoregressive
4) Heteroscedastic
5) Matrik Kovarian Tidak-terstruktur

6) Matrik Kovarian Diagonal

Menurut Danardono (2012), Ada tiga komponen utama yang membentuk GLM *(Generalized Linear Model)* yaitu

1) Asumsi Distribusi
2) Komponen Sistematik
3) Fungsi Penghubung

Selanjutnya karena dalam GLM mengasumsikan independensi antar observasi. Sementara dalam data longitudinal, pengukuran sering diasumsikan berkorelasi, maka GLM tidak tepat digunakan untuk data longitudinal. Sehingga dibutuhkan model alternatif yaitu dengan *Generalized Estimating Equation* (GEE).

Generalized Estimating Equation (GEE) yang merupakan perluasan dari GLM dalam hal spesifikasi korelasi antara dua respon yang berbeda $y_i$dan y_k. Sama halnya GLM, GEE mempunyai spesifikasi linear prediktor sebagai berikut;

$$g(\mu_{ij}) = \eta_{ij} = x_i\boldsymbol{\beta}$$

Jenis dan pemilihan fungsi penghubung sama seperti pada GLM. Variansi dari respon *y* ditentukan oleh

$$V(y_{ij}) = \phi v(\mu_{ij})$$

dengan $v(\mu_{ij})$ adalah fungsi variansi dan ϕ parameter skala yang diketahui atau diestimasi. Tambahan spesifikasi GEE yang tidak terdapat pada GLM adalah spesifikasi korelasi antara dua respon yang berbeda, atau sering disebut sebagai *working correlation matrix* R_i berukuran $n_i \; x \; n_i$ yang bergantung pada parameter α, sehingga matriks korelasi ini sering juga ditulis $R_i(\alpha)$. Jika didefinisikan A_i matriks diagonal berukuran $n \; x \; n$ dengan V (μ_{ij} adalah elemen diagonalnya) dan matriks variansi-kovariansi untuk y_i adalah

$$V(\alpha) = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2}$$

Maka estimasi GEE untuk β adalah solusi dari

$$\sum_{i=1}^{m} D_i^T [V(\hat{\alpha})]^{-1} (y_i - \mu_i) = 0$$

dengan $\hat{\alpha}$ merupakan estimator konsisten α dan $D_i = \partial\mu_i/\partial\beta$. Matriks D_i disebut sebagai matriks derivatif, yaitu matriks yang berisi turunan dari μ_i terhadap komponen β. Matriks ini mentransformasikan unit asal berupa Y_i menjadi unit pada skala $g(\mu_i)$. Skala pada unit fungsi penghubung $g(\mu_i)$ ini dapat digunakan untuk memberikan interpretasi pada nilai β

3. MODEL REGRESI SPLINE

Spline adalah potongan polinomial order p dengan titik bersama dari potonganpotongan tersebut disebut dengan knot. Titik knot merupakan perpaduan dua kurva yang menunjukkan pola perubahan perilaku kurva pada selang yang berbeda. Penggunaan titik knot banyak digunakan dalam regresi nonparametrik, karena secara visual dapat menunjukkan setiap perubahan pola perilaku yang terjadi dalam interval waktu tertentu (Islamiyati, 2009). Misalkan pola perubahan yang terjadi pada data sebanyak lima pola perubahan, dimana titik terjadinya pola perubahan tersebut disebut titik knot. Pola perubahan yang terjadi, yaitu pola pertama cenderung naik, kemudian menurun pada pola kedua. Selanjutnya pola ketiga juga menunjukkan kecenderungan turun tetapi penurunannya berbeda dengan pola kedua. Pola keempat mengalami kenaikan kembali dan terus naik pada pola kelima tetapi dengan kecenderungan naik yang berbeda pula. Contoh ini menunjukkan bahwa dengan penggunaan regresi spline, sangat memungkinkan dalam satu data terdapat beberapa pola perubahan dalam setiap interval waktu berbeda.

Spline orde p dengan knot $k_1, k_{2,} K, k_m$ diberikan dalam fungsi f dengan bentuk:

$$f(t_i) = \sum_{j}^{p} \beta_j t_i^j + \sum_{j=1}^{m} \beta_{j+p} (t_i - k_j)_+^p \qquad (2)$$

Dengan $\beta_0, \beta_1, K, \beta_{p+j}$ adalah parameter regresi, dan

$$(t_i - k_j)_+^p = \begin{cases} (t_i - k_j)_+^p, t_i - k_j \geq 0 \\ 0, t_i - k_j \prec 0 \end{cases} \qquad (3)$$

Salah satu cara pemilihan titik knot optimal adalah menggunakan metode generalized cross validation (GCV). Kriteria GCV didefenisikan:

$$GCV(k) = \frac{MSE(k)}{[n^{-1}trace(1-A(k))]^2}$$

Dengan :

$$MSE(k) = n^{-1}\tilde{y}^T[(I-A(k))^T(I-A(k))]\tilde{y}$$

K = adalah titik knots,

$A(k) = X(X^TX)^{-1}X^T$, $A(k)$ adalah matrik berukuran nxn

$$\tilde{y} = (\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_r)^T \quad (4)$$

4. ESTIMASI KURVA REGRESI SPLINE PADA DATA DATA LONGITUDINAL

Data longitudinal yang diukur berulangkali berdasarkan waktu diberikan oleh $(t_{ij}, y_{ij}), j = 1,2,...,n; i = 1,2,...,n$, dimana n_i menyatakan banyaknya pengukuran berulang dari obyek ke-i.Jika diberikan model regresi nonparametrik untuk data longitudinal maka diperoleh suatu bentuk seperti pada (1). Spline pada data longitudinal diberikan dengan bentuk persamaan :

$$f_i(t_{ij}) = \sum_{l=0}^{p} \beta_{li} t_{ij}^1 + \sum_{m=1}^{r} \beta_{(p=m)i}(t_{ij} - k_m)_+^p$$

Dimana :

$k_1, k_2, ..., k_r$ = titik knot

P = jumlah orde

t_{ij} = pengaruh variabel waktu pada objek ke-i dengan pengulangan ke-j

β = parameter

Spline orde p , dapat dimodelkan sebagai berikut:

$$f_i(t_{ij}) = \beta_{0i} + \beta_{1i}t_{1j}^1 + \beta_{2i}t_{2j}^2 + ... + \beta_{pi}t_{pj}^p + \sum_{m=1}^{r} \beta_{(p=m)i}(t_{ij} - k_m)_+^p$$

Menurut model spline pada (5), maka model regresi nonparametrik berdasarkan (1) dapat ditulis :

$$f_i(t_{ij}) = \beta_{0i} + \beta_{1i}t_{1j}^1 + \beta_{2i}t_{2j}^2 + ... + \beta_{pi}t_{pj}^p + \beta_{(p+1)i}(t_{ij} - k_1)^p + \beta_{(p+2)i}(t_{ij} - k_1)^p + ... + \beta_{(p+r)i}(t_{ij} - k_r)^p + \varepsilon_{ij}$$

yang dapat disajikan dengan bentuk matriks, yaitu:

$$\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_i} \\ \hline y_{21} \\ y_{22} \\ \vdots \\ y_{2n_i} \\ \hline \vdots \\ \\ \\ \overline{y_{n1}} \\ y_{n1} \\ \vdots \\ y_{nn_i} \end{bmatrix} = \begin{bmatrix} 1 & t_{11} & {t_{11}}^2 & \cdots & {t_{11}}^p & (t_{11}-k_1)_+^1 & \cdots & (t_{11}-k_r)_+^p \\ 1 & t_{12} & {t_{12}}^2 & \cdots & {t_{12}}^p & (t_{12}-k_1)_+^1 & \cdots & (t_{12}-k_r)_+^p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{1n_i} & {t_{1n_i}}^2 & \cdots & {t_{1n_i}}^p & (t_{1n_i}-k_1)_+^1 & \cdots & (t_{1n_i}-k_r)_+^p \\ \hline 1 & t_{21} & {t_{21}}^2 & \cdots & {t_{21}}^p & (t_{21}-k_1)_+^1 & \cdots & (t_{21}-k_r)_+^p \\ 1 & t_{22} & {t_{22}}^2 & \cdots & {t_{22}}^p & (t_{22}-k_1)_+^1 & \cdots & (t_{22}-k_r)_+^p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{2n_i} & {t_{2n_i}}^2 & \cdots & {t_{2n_i}}^p & (t_{2n_i}-k_1)_+^1 & \cdots & (t_{2n_i}-k_r)_+^p \\ \hline \vdots & & & & & & & \\ \\ \\ \overline{1} & & & & & & & \\ 1 & t_{n1} & {t_{n1}}^2 & \cdots & {t_{n1}}^p & (t_{n1}-k_1)_+^1 & \cdots & (t_{n1}-k_m)_+^p \\ 1 & t_{n2} & {t_{n2}}^2 & \cdots & {t_{n2}}^p & (t_{n2}-k_1)_+^1 & \cdots & (t_{n2}-k_m)_+^p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{nn_i} & {t_{nn_i}}^2 & \cdots & {t_{nn_i}}^p & (t_{nn_i}-k_1)_+^1 & \cdots & (t_{nn_i}-k\ )_+^p \end{bmatrix} \begin{bmatrix} \beta_{01} \\ \beta_{11} \\ \beta_{21} \\ \vdots \\ \beta_{p1} \\ \beta_{(p+1)1} \\ \vdots \\ \beta_{(p+r)1} \\ \hline \beta_{02} \\ \beta_{12} \\ \beta_{22} \\ \vdots \\ \beta_{p2} \\ \beta_{(p+1)2} \\ \vdots \\ \beta_{(p+r)2} \\ \hline \vdots \\ \\ \\ \overline{\beta_{0n}} \\ \beta_{1n} \\ \beta_{2n} \\ \vdots \\ \beta_{pn} \\ \beta_{(p+1)n} \\ \vdots \\ \beta_{(p+r)n} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1n_i} \\ \hline \varepsilon_{21} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{2n_i} \\ \hline \vdots \\ \\ \\ \overline{\varepsilon_{n1}} \\ \varepsilon_{n2} \\ \vdots \\ \varepsilon_{nn_i} \end{bmatrix}$$

Model matriks pada (7) dapat disederhanakan dalam bentuk:

$$\begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \vdots \\ \tilde{y}_n \end{bmatrix}_{nx1} = \begin{bmatrix} \tilde{1} & \tilde{t}_1\tilde{t}_1^{\,2} & \cdots & \tilde{t}_1^{\,p} & (\tilde{t}_1-k_1)^p & \cdots & (\tilde{t}_1-k_m)_+^p \\ \tilde{1} & \tilde{t}_2\tilde{t}_2^{\,2} & \cdots & \tilde{t}_2^{\,p} & (\tilde{t}_2-k_1)^p & \cdots & (\tilde{t}_2-k_m)_+^p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \tilde{1} & \tilde{t}_n\tilde{t}_n^{\,2} & \cdots & \tilde{t}_n^{\,p} & (\tilde{t}_n-k_1)^p & \cdots & (\tilde{t}_n-k_m)_+^p \end{bmatrix} \begin{bmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \\ \vdots \\ \tilde{\beta}_n \end{bmatrix} + \begin{bmatrix} \tilde{\varepsilon}_1 \\ \tilde{\varepsilon}_2 \\ \vdots \\ \tilde{\varepsilon}_n \end{bmatrix}$$

Jika dituliskan dalam notasi matriks, dapat ditulis menjadi

$$\tilde{y} = X\tilde{\beta} + \tilde{\varepsilon}$$

Untuk memperoleh bentuk pendugaan $\tilde{\beta}$ dilakukan melalui metode kuadrat terkecil dengan cara meminimumkan Jumlah Kuadrat Galat (JKG):

$$\varepsilon^T\varepsilon = (\tilde{y} - X\tilde{\beta})^T(\tilde{y} - X\tilde{\beta})$$

$$M = \varepsilon^T\varepsilon = (\tilde{y} - X\tilde{\beta})^T(\tilde{y} - X\tilde{\beta})$$

$$= (\tilde{y}^T\tilde{y} - 2\tilde{\beta}^T X^T\tilde{y} - \tilde{\beta}^T X^T X\tilde{\beta})$$

Selanjutnya diperoleh:

$$\frac{\partial M}{\partial\tilde{\beta}} = \frac{\partial(\tilde{y}^T\tilde{y} - 2\tilde{\beta}^T X^T\tilde{y} - \tilde{\beta}^T X^T X\tilde{\beta})}{\partial\tilde{\beta}}$$

$$= -2X^T\tilde{y} + 2X^T X\tilde{\beta}$$

Kemudian :

$$-2X^T\tilde{y} + 2X^T X\tilde{\beta} = 0$$
$$2X^T\tilde{y} = 2X^T X\tilde{\beta}$$
$$X^T\tilde{y} = X^T X\tilde{\beta}$$
$$\hat{\tilde{\beta}} = (X^T X)^{-1}X^T\tilde{y} \text{ atau } \beta = (X^T X)^{-1}X^T\tilde{y}$$

Akibatnya pendugaan kurva regresi $f_i(t_{ij})$ diberikan oleh:

$$\hat{\tilde{f}}(k_1, k_2, ..., k_r) = XB$$
$$= X(X^T X)^{-1}X^T\tilde{y} = A(k_1, k_2, ..., k_r)\tilde{y}$$

dengan matriks $A(k_1, k_2, ..., k_r) = X(X^T X)^{-1}X^T$

Terlihat bahwa pendugaan untuk kurva regresi spline untuk data longitudinal merupakan kelas pendugaan linear dalam observasi respon $\tilde{y}_i$ dan sangat tergantung pada titik knot $k_1, k_2, ..., k_r$

5. KESIMPULAN

Estimasi kurva regresi spline untuk data longitudinal dapat disajikan dalam bentuk:

$$\hat{\tilde{f}}(k_1, k_2, ..., k_r) = XB = X(X^T X)^{-1}X^T\tilde{y} = A(k_1, k_2, ..., k_r)\tilde{y}$$

Dengan $\tilde{y}$ adalah variabel respon berorde N ×1 diberikan oleh:

$$\tilde{y} = (\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_r)^T$$

6. DAFTAR PUSTAKA

Besse, P.C., Cardot, H., dan Ferraty, F. (1997), "Simultaneous Nonparametric Regression of Unbalanced Longitudinal Data", Computational Statistics and Data Analysis24, 255 – 270.

Carroll, R.J., Hall, P., Apanasovich, T.V., dan Lin, X. (2004), "Histospline Method in Nonparametric Regression Models with Application to Clustered/Longitudinal Data", Statistica Sinica14, 649 – 674.

Danardono. 2012. Analisis Data Longitudi nal,Yogyakarta:UGM

Diggle, P.J.K., Liang, K.Y., dan Zeger, S.L. (1995), Analysis of Longitudinal Data, Clarendon Press, Oxford.

Eubank, R. L. (1988), Spline smoothing and Nonparametrik Regression, Marcel Dekker, New York.

Green, P.J., dan Silverman, B.W. (1994), Nonparametrik Regression and Generalized Linear Models (a Roughness Penalty Approach), Chapman & Hall, New York.

Islamiyati, Anna. (2009). Estimasi Spline untuk Data Longitudinal dengan Penalized Likelihood. Seminar Nasional Matematika IV ITS, Surabaya.

Islamiyati, Anna (2012). "Regresi Nonparametrik untuk Data Longitudinal". Jurnal Matematika Statistika & Komputasi, Unhas, Makassar.

Wahba, G. (1990), Spline Model for Observational Data, Society For Industrial and Applied Mathematics, Philadelphia.

Wu, H., dan Zhang, J.T. (2006), Nonparametric Regression Methods for Longitudinal Data Analysis, John Wiley & Sons, New Jersey.