Metode Likelihood Lokal Dengan Pembobot Kernel Pada Regresi Nonparametrik Dengan Respon Normal

Toha Saifudin

Departemen Matematika
Fakultas Sains dan Teknologi
Universitas Airlangga

Abstrak

Model regresi nonparametrik berbentuk $y_i = s(x_i) + \varepsilon_i$, untuk i = 1, 2, ..., n dengan $s(x_i)$ adalah fungsi halus. Bermacam-macam metode pendugaan model regresi nonparametrik telah dikembangkan oleh para peneliti. Kebanyakan metode pendugaan yang digunakan merupakan metode bebas distribusi. Dalam paper ini kami bertujuan untuk mendapatkan penduga model regresi nonparametrik menggunakan metode berbasis distribusi yaitu likelihood lokal dengan pembobot kernel yang diterapkan pada respon berdistribusi normal.

Kata kunci : regresi nonparametrik, fungsi halus, likelihood lokal, pembobot kernel.

PENDAHULUAN

Penelitian tentang analisis regresi nonparametrik mengalami perkembangan yang pesat. Hal tersebut disebabkan metode nonparametrik tidak membutuhkan asumsi mengenai bentuk dari fungsi regresi, dan memberikan fleksibilitas pada data sampel untuk mencari bentuk fungsional yang dapat menggambarkan data dengan baik. Secara umum model regresi nonparametrik antara peubah respon dengan satu prediktor dapat dinyatakan sebagai

$$y_i = s(x_i) + \varepsilon_i,\ i = 1,2,...,n. \qquad (1)$$

dengan y_i adalah nilai peubah respon Y, x_i adalah nilai peubah prediktor X, ε_i adalah galat model dengan mean 0 dan varians σ^2, dan $s(x_i)$ adalah fungsi regresi yang bentuknya tidak diketahui (Hardle, 1990).

Pendugaan $s(x_i)$ secara nonparametrik dilakukan berdasarkan data pengamatan dengan teknik *smoothing*. Teknik ini tidak mengasumsikan distribusi probabilitas dari respon. Ada beberapa teknik *smoothing* dalam regresi nonparametrik antara lain Histogram, Penduga Kernel, Penduga Deret Orthogonal, Penduga Spline, K-NN, Deret Fourier, Wavelet, dan lain – lain (Hardle (1990), Eubank (1988)).

Dalam analisis regresi seringkali dijumpai kenyataan bahwa variabel respon diasumsikan mengikuti distribusi tertentu. Asumsi distribusi dari respon ini seringkali diperlukan dalam pemodelan regresi data uji hidup. Permasalahan dalam regresi tersebut adalah mendapatkan penduga model yang menyatakan hubungan keterkaitan variabel prediktor terhadap respon yang sesuai dengan asumsi distribusi yang diketahui tersebut. Untuk respon yang berdistribusi normal akan mempunyai bentuk model regresi yang berbeda dengan ketika respon berdistribusi eksponensial. Oleh karena itu dalam paper ini kami akan membahas bagaimana metode menduga model regresi apabila respon berdistribusi tertentu, khususnya Normal dengan bentuk fungsional regresinya tidak diketahui. Metode yang kami gunakan adalah likelihood lokal dengan menggunakan pembobot kernel.

PEMBAHASAN

1. Maksimum likelihood lokal

Diketahui *n* data berpasangan $\{(x_1, y_1),...,(x_n, y_n)\}$ yang saling bebas dan diasumsikan bahwa untuk $X = x$, fungsi kepadatan peluang dari *Y* adalah

$$Y \mid x \sim f(y \mid \theta) \tag{2}$$

dengan θ adalah parameter yang merupakan fungsi dari *x* yaitu $\theta = s(x)$, $s(x)$ adalah fungsi penghalus. Berdasarkan sampel berukuran n di atas, fungsi *likelihood* dari $Y \mid x$ adalah

$$L(\theta \mid \boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{n} f(y_i \mid s(x_i)). \tag{3}$$

untuk $\boldsymbol{x} = \{x_1, x_2, ..., x_n\}$ dan $\boldsymbol{y} = \{y_1, y_2, ..., y_n\}$.

Selanjutnya untuk menduga $s(x_i)$ berdasarkan likelihood lokal, dalam paper ini terlebih dulu menggunakan pendekatan linier lokal seperti yang dilakukan oleh Tibshirani (1984) dengan bentuk $s(x_i) = \beta_{0i} + \beta_{1i} x_i$. Penduga *likelihood* lokal untuk $s(x_i)$ adalah

$$\hat{s}(x_i) = \hat{\beta}_{0i} + \hat{\beta}_{1i}\, x_i \tag{4}$$

dengan $\hat{\beta}_{0i}$ dan $\hat{\beta}_{1i}$ nilai yang memaksimumkan fungsi *ln likelihood* lokal

$$\ell_i(\beta_{0i}, \beta_{1i} \mid \boldsymbol{x}, \boldsymbol{y}, h) = \sum_{j=1}^{n} \left\{ ln\left(f(y_j \mid \beta_{0i} + \beta_{1i} x_j) \right) \frac{1}{h} K\left(\frac{x_i - x_j}{h} \right) \right\}, \text{ i = 1, 2, ..., n.} \tag{5}$$

K(.) dalam (5) adalah fungsi kernel. Fungsi *ln likelihood* lokal (5) tersebut mengikuti penulisan Santos dan Neves (2007) dengan sedikit penyesuaian.

2. Maksimum likelihood lokal pada model regresi dengan respon berdistribusi Normal

Misalkan diberikan n data pengamatan $\{x_i, y_i\}_{i=1}^{n}$ mengikuti model regresi

$$y_i = s(x_i) + \varepsilon_i \tag{6}$$

dengan $Y_i \sim N(s(x_i), \sigma^2)$, dan $\varepsilon_i \sim N(0, \sigma^2)$.

Fungsi likelihood global berdasarkan model (6) adalah

$$L(s(x) \mid \boldsymbol{x}, \boldsymbol{y}) = \prod_{j=1}^{n} f(y_j \mid \boldsymbol{x}, \boldsymbol{y}) = \prod_{j=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma} exp\left\{ -\frac{1}{2\sigma^2} (y_j - s(x_j))^2 \right\} \right], \tag{7}$$

dan *Logaritma natural* dari persamaan (7) adalah

$$\mathcal{L}(s(x) \mid \boldsymbol{x}, \boldsymbol{y}) = ln\, L(s(x) \mid \boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{n} \left[ln\left(\frac{1}{\sqrt{2\pi}\sigma} \right) - \frac{1}{2\sigma^2} (y_j - s(x_j))^2 \right]. \tag{8}$$

Selanjutnya untuk estimasi parameter $s(x_i)$, i = 1, 2, ..., n, atau yang tidak lain adalah merupakan fungsi regresi nonparametrik, terlebih dahulu mendekati fungsi tersebut dengan pendekatan linier lokal, yaitu

$s(x_i) = \beta_{0i} + \beta_{1i} x_i$, i = 1, 2, ..., n. (9)

Berdasarkan (9), dapat dilihat bahwa menduga $s(x_i)$ adalah identik dengan menduga β_{0i} dan β_{1i}. Untuk mendapatkan penduga parameter lokal β_{0i} dan β_{1i}, i = 1, 2, ..., n, dilakukan dengan mencari β_{0i} dan β_{1i} yang memaksimumkan fungsi *ln likelihood* lokal sebagai berikut :

$$\begin{aligned} \mathcal{L}_i(\beta_{0i}, \beta_{1i} \mid \boldsymbol{x}, \boldsymbol{y}, h) &= \sum_{j=1}^{n} \left[\left(ln\left(\frac{1}{\sqrt{2\pi}\sigma} \right) - \frac{1}{2\sigma^2} (y_j - \beta_{0i} - \beta_{1i} x_j)^2 \right) . \frac{1}{h} K\left(\frac{x_i - x_j}{h} \right) \right] \\ &= ln\left(\frac{1}{\sqrt{2\pi}\sigma} \right) \sum_{j=1}^{n} \frac{1}{h} K\left(\frac{x_i - x_j}{h} \right) - \\ &\quad \frac{1}{2\sigma^2} \sum_{j=1}^{n} (y_j - \beta_{0i} - \beta_{1i} x_j)^2 . \frac{1}{h} K\left(\frac{x_i - x_j}{h} \right) \end{aligned} \tag{10}$$

Persamaan (10) diatas dapat ditulis menjadi notasi matrik sebagai berikut:

$$\mathcal{L}_i\left(\boldsymbol{\beta}_i \middle| \boldsymbol{x},\boldsymbol{y},h\right) = M = ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right)tr\left(\boldsymbol{W}_i\right) - \frac{1}{2\sigma^2}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_i\right)'\boldsymbol{W}_i\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_i\right) \tag{11}$$

dengan:

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad \boldsymbol{\beta}_i = \begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix} \quad .$$

$$\boldsymbol{W}_i = \begin{pmatrix} \frac{1}{h}K\left(\frac{x_i - x_1}{h}\right) & 0 & \cdots & 0 \\ 0 & \frac{1}{h}K\left(\frac{x_i - x_2}{h}\right) & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{h}K\left(\frac{x_i - x_n}{h}\right) \end{pmatrix}$$

Bentuk (11) dapat diuraikan sebagai berikut:

$$\begin{aligned} M &= ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right)tr\left(\boldsymbol{W}_i\right) - \frac{1}{2\sigma^2}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_i\right)'\boldsymbol{W}_i\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_i\right) \\ &= ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right)tr\left(\boldsymbol{W}_i\right) - \frac{1}{2\sigma^2}\left(\boldsymbol{Y}' - \boldsymbol{\beta}_i'\boldsymbol{X}'\right)\left(W_i\boldsymbol{Y} - W_i\boldsymbol{X}\boldsymbol{\beta}_i\right) \\ &= ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right)tr\left(\boldsymbol{W}_i\right) - \frac{1}{2\sigma^2}\left(\boldsymbol{Y}'W_i\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{W}_i\boldsymbol{X}\boldsymbol{\beta}_i - \boldsymbol{\beta}_i'\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{Y} + \boldsymbol{\beta}_i'\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X}\boldsymbol{\beta}_i\right) \\ &= ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right)tr\left(\boldsymbol{W}_i\right) - \frac{1}{2\sigma^2}\left(\boldsymbol{Y}'\boldsymbol{W}_i\boldsymbol{Y} - \boldsymbol{\beta}_i'\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{Y} - \boldsymbol{\beta}_i'\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{Y} + \boldsymbol{\beta}_i'\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X}\boldsymbol{\beta}_i\right) \\ &= ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right)tr\left(\boldsymbol{W}_i\right) - \frac{1}{2\sigma^2}\left(\boldsymbol{Y}'\boldsymbol{W}_i\boldsymbol{Y} - 2\boldsymbol{\beta}_i'\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{Y} + \boldsymbol{\beta}_i'\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X}\boldsymbol{\beta}_i\right), \end{aligned} \tag{12}$$

karena $\boldsymbol{\beta}_i'\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{Y}$ adalah matrik 1x1 atau skalar yang transposenya $\left(\boldsymbol{\beta}_i'\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{Y}\right)' = \boldsymbol{Y}'\boldsymbol{W}_i\boldsymbol{X}\boldsymbol{\beta}_i$ mempunyai nilai yang sama.

Nilai dugaan bagi $\boldsymbol{\beta}_i$ adalah $\hat{\boldsymbol{\beta}}_i$ yang memaksimumkan *M*. Nilai maksimum *M* dicapai pada saat $\frac{\partial \ln M}{\partial \boldsymbol{\beta}_i} = 0$, yaitu $-\frac{1}{2\sigma^2}\left(-2\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{Y} + 2\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X}\hat{\boldsymbol{\beta}}_i\right) = 0$, dan setelah diselesaikan, diperoleh penduga

$$\hat{\boldsymbol{\beta}}_i = \left(\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{Y} \; . \tag{13}$$

Selanjutnya penduga model regresi berdasarkan maksimum *likelihood* lokal adalah

$$\hat{Y}_i = \hat{s}\left(x_i\right) = \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}_i = \boldsymbol{X}_i'\left(\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{Y} \tag{14}$$

dengan $\mathbf{X}_i' = \begin{pmatrix} 1 & x_i \end{pmatrix}$.

3. Pemilihan bandwidth

Pemilihan *bandwidth* (dinotasikan h) sangat penting dalam mendapatkan penduga maksimum *likelihood* lokal**.** *Bandwidth* yang optimal diperoleh dengan cara meminimumkan GCV. Kriteria GCV pada penduga maksimum *likelihood* lokal didefinisikan sebagai berikut:

$$GCV\left(h\right) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\left(h\right)\right)^2}{\left[n^{-1}tr\left(1 - A\left(h\right)\right)\right]^2} \tag{15}$$

dengan ***A****(h)* adalah matrik yang diperoleh berdasarkan bentuk (Eubank,1988)

$$\hat{\boldsymbol{Y}} = A\left(h\right)\boldsymbol{Y} \; . \tag{16}$$

Berdasarkan pesamaan (14), maka diperoleh bahwa baris ke-i, i = 1, 2, …, n dari ***A****(h)* adalah $\boldsymbol{X}_i'\left(\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}_i$.

4. Algoritma pendugaan model regresi

Algoritma untuk menduga model adalah sebagai berikut :

i. Memasukkan sampel berpasangan (x_i , y_i) berukuran *n*.
ii. Memilih *bandwidth* optimal berdasarkan kriteria GCV pada persamaan (15).
iii. Menghitung dugaan parameter linier lokal β_{0i} dan β_{1i} berdasarkan persamaan (13).
iv. Menghitung dugaan model regresi, $\hat{y}_i$ berdasarkan persamaan (14).

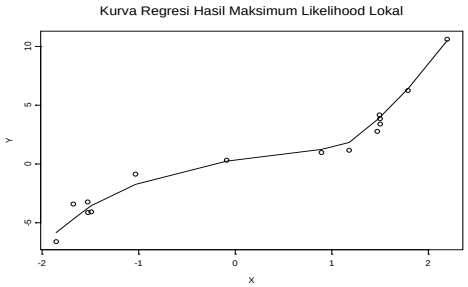
5. Studi kasus

Dalam bagian ini, kami berikan sebuah contoh ilustrasi menggunakan data bangkitan. Sebuah sampel berpasangan berukuran 15 observasi dibangkitkan berdasarkan hubungan $y = x^3 + \varepsilon$ dengan $y \sim N(0,5)$ dan $\varepsilon \sim N(0, 0.6)$ ditunjukkan dalam Tabel 1.

Tabel 1. Data bangkitan

No	Y	X	No	Y	X
1	10.61	2.19	11	-4.07	-1.49
2	-4.13	-1.52	12	6.25	1.79
3	-0.86	-1.03	13	-3.23	-1.53
4	3.39	1.50	14	-3.41	-1.68
5	4.17	1.49	15	0.97	0.89
6	-6.61	-1.85			
7	0.32	-0.09			
8	3.85	1.50			
9	2.77	1.47			
10	1.17	1.18			

Untuk sampel di atas, dengan fungsi kernel Epanichnikov untuk pembobotan, diperoleh bandwidth optimal h = 1.07, GCV = 0.8245638, MSE = 0.3812685, R^2 = 0.9807584 dengan plot dugaan regresi seperti dalam Gambar 1.



Gambar 1. Plot dugaan regresi dengan bandwidth optimal 1,07

KESIMPULAN

Pendugaan model regresi nonparametrik menggunakan metode maksimum likelihood lokal dapat dilakukan dengan terlebih dulu melakukan pendekatan parametrik lokal terhadap fungsi halus *s(x)*. Untuk respon berdistribusi Normal dan pendekatan fungsi halus menggunakan linier lokal yaitu $s(x_i) = \beta_{0i} + \beta_{1i} x_i$, diperoleh penduga model regresi

$$\hat{Y}_i = \hat{s}(x_i) = \boldsymbol{X}_i'(\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{Y}$$

dengan $\mathbf{X}_i' = (1 \quad x_i)$.

DAFTAR PUSTAKA

Eubank, R.M., 1988. *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York.
Hardle, W, 1990, *Applied Nonparametrik Regression*, Cambridge University Press, New York.
Santos, J.A., dan Neves, M.M., 2007. A local maximum likelihood estimator for Poisson regression, *Metrika*, DOI 10.1007/s00184-007-0156-1, © Springer-Verlag.
Tibshirani, R.J., 1984. *Local Likelihood Estimation*, Stanford University, California.