Jurnal UJMC, Volume 2, Nomor 1, Hal. 1 - 7
pISSN : 2460 - 3333 eISSN : 2579-907X

ESTIMASI MODEL REGRESI SEMIPARAMETRIK DISKRIT

Baiq Diah Fitasari[1], Sri Haryatmi[2], dan Zulaela[2]

[1] Universitas Islam Al-Azhar Mataram, baiq_diah_fitasari@yahoo.com,
[2] Universitas Gadjah Mada, s_kartiko@yahoo.com,
[3] Universitas Gadjah Mada, zulaela@ugm.ac.id

Abstract. Approaches that are used to estimate the regression function are parametric regression model approaching and nonparametric regression model approaching. Semiparametric regression is association of parametric regression and nonparametric regression. Semiparametric regression is used if the relation pattern between independent variables and dependent variables has the known pattern and also has the unknown pattern. Estimating the unknown regression approaching and in this case is using the Nadaraya Watso estimator. Semiparametric estimator is better than nonparametric estimator for the data which have the unknown relationpattern between dependent and independent variable and also has the unknown relation pattern between dependent and independent variable by using the MSE value.

Keywords: *estimator, parametric, Nadaraya-Watson estimator, semiparametric estimator.*

Abstrak. Pendekatan yang digunakan untuk mengestimasi fungsi regresi ada dua jenis yaitu pendekatan model regresi parametrik dan pendekatan model regresi nonparametrik. Regresi semiparametrik merupakan gabungan antara regresi parametrik dan nonparametrik. Regresi semiparametrik digunakan jika pola hubungan antara sekumpulan variabel independen terhadap variabel dependen ada yang polanya diketahui dan ada pula yang polanya tidak dapat diketahui. Mengestimasi fungsi regresi yang tidak diketahui dapat menggunakan pendekatan estimator kernel dalam hal ini mengunakan estimator Nadaraya-Watson. Estimator semiparametrik lebih baik dibandingkan dengan estimator nonparametrik untuk data yang sebagian pola hubungan variabel dependen dan independennya diketahui dan sebagian polanya tidak diketahui dapat dilihat dari nilai MSE yang lebih kecil.

Kata Kunci: *estimator, parametrik, estimator Nadaraya-Watson, estimator semiparametrik.*

1 Pendahuluan

Analisis regresi merupakan alat statistik yang banyak digunakan dalam berbagai bidang, yang bertujuan untuk mengetahui hubungan antara variabel dependen dan variabel independen [3]. Pendekatan model regresi semiparametrik merupakan pendekatan model baru dalam regresi yang menggabungkan antara regresi parametrik dan nonparametrik, dalam artian sebagian variabel independennya bersifat parametrik dan sebagian lain bersifat nonparametrik. Regresi semiparametrik digunakan jika pola hubungan antara sekumpulan variabel independen terhadap variabel dependen ada yang polanya diketahui dan ada pula yang polanya tidak dapat diketahui [8].

Mengestimasi fungsi regresi yang tidak diketahui dapat menggunakan pendekatan estimator kernel [9]. Kernel merupakan suatu fungsi yang mewakili

variabel-variabel independen yang bersifat nonparametrik. Estimator kernel memiliki bentuk yang lebih fleksibel dan perhitungan matematisnya mudah disesuaikan. Estimasi model regresi semiparametrik dapat dilakukan dengan berbagai metode yang ada misalnya metode kuadrat terkecil, metode *likelihood*, metode *Mean Square Error* (MSE), *Root Mean Squared Error* (RMSE) dan lain-lain [5], [6].

Variabel diskrit merupakan variabel yang hasil pengukurannya (kodomain) berupa bilangan bulat. Variabel diskrit sering juga dinyatakan sebagai variabel kategori. Contoh variabel diskrit dikotomi adalah jenis kelamin, status perkawinan, sedangkan variabel diskrit polikotomi contohnya yaitu tingkat pendidikan. Untuk beberapa model regresi semiparametrik kontinu telah dibahas oleh [4] dan [7]. Berdasarkan uraian di atas penulis tertarik untuk mengkaji tentang estimasi model regresi semiparametrik diskrit dan simulasinya menggunakan program R.

2 MODEL REGRESI SEMIPARAMETRIK

Bentuk dari model regresi semiparametrik didefinisikan sebagai berikut [8]:

$$Y_i = m(X_i) + \varepsilon_i, i = 1, 2, .., n \tag{1}$$

dengan Y_i adalah nilai variabel dependen ke-i, X_i adalah nilai variabel independen ke-i, m adalah fungsi regresi yang tidak diketahui untuk diestimasi dan ε_i adalah *error* dengan mean $E(\varepsilon_i) = 0$ dan variansi $Var(\varepsilon_i) = \sigma^2 < \infty$, serta

$$\begin{aligned} m(x) &= r(x;\beta)\omega(x) \\ &=: m_\omega(x;\beta), untuk \quad x \in N^d \end{aligned} \tag{2}$$

dengan $r(x;\beta)$ adalah fungsi parametrik yang tergantung pada parameter yang tidak diketahui $\beta = (\beta_1, \cdots, \beta_p)^T$ dan $\omega(.)$ fungsi koreksi perkalian nonparametrik.

2.1 Komponen Parametrik

Bentuk dari model linier didefinisikan sebagai berikut [3]:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, 2, \cdots, n \tag{3}$$

dengan Y_i adalah nilai variabel dependen ke-i, X_i adalah nilai variabel independen ke-i dan ε_i adalah *error* dengan mean $E(\varepsilon_i) = 0$ dan variansi $Var(\varepsilon_i) = \sigma^2 < \infty$. Berdasarkan Persamaan 3 dan dengan menggunakan metode kuadrat terkecil akan diperoleh estimator untuk β_0 dan β_1 sebagai berikut:

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X} \tag{4}$$

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \tag{5}$$

2.2 Komponen Nonparametrik

2.2.1 Gabungan Kernel Diskrit Dalam metode gabungan kernel diskrit fungsi kernel $K_{x,h}(\cdot)$ merupakan fungsi massa probabilitas (f.m.p) dengan *support* S_x yang tidak tergantung pada h dan $x \in S_x$. Selain itu, diberlakukan dua asumsi sebagai berikut [2]:

$$\lim_{h\to 0} E(K_{x,h}) = x \tag{6}$$

$$\lim_{h\to 0} Var(K_{x,h}) = 0 \tag{7}$$

dimana $K_{x,h}$ adalah variabel acak diskrit dengan f.m.p adalah $K_{x,h}(\cdot)$.

2.2.2 Estimasi Regresi Nonparametrik Diketahui estimator Nadaraya-Watson sebagai berikut [5]:

$$\tilde{m}_n(x) = \frac{\sum_{i=1}^{n} Y_i K_{x,h}(X_i)}{\sum_{j=1}^{n} K_{x,h}(X_j)}, \quad x \in \mathbf{N} \tag{8}$$

dengan $h = h(n) > 0$ adalah urutan sebarang parameter *smoothing* yang memenuhi $\lim_{n\to\infty} h(n) = 0$ dan $K_{x,h}(\cdot)$ adalah gabungan fungsi kernel diskrit.

Teorema 1. *Misalkan f merupakan f.m.p dari variabel acak diskrit X yang didefinisikan pada $\mathbf{N}$. Diasumsikan bahwa $f(x) = P(X = x) > 0$ untuk $x \in \mathbf{N}$. Selanjutnya, andaikan bandwidth $h = h(n) > 0$ memenuhi $\lim_{n\to\infty} h = 0$ dan kernel diskrit $K_{x,h}(\cdot)$ memenuhi asumsi 6 dan 7, maka bias dan variansi dari $\tilde{m}_n(x)$ yaitu:*

$$Bias[\tilde{m}_n(x)] = \left[m^{(2)}(x) + 2m^{(1)}(x)\left(\frac{f^{(1)}}{f}\right)(x)\right]\frac{Var(K_{x,h})}{2} + O\left(\frac{1}{n}\right) + o(h) \tag{9}$$

$$Var[\tilde{m}_n(x)] = \frac{Var(Y|X=x)}{nf(x)}[P(K_{x,h} = x)]^2 + o(\frac{1}{n}) \tag{10}$$

dengan $f^{(1)}, m^{(1)}$ dan $m^{(2)}$ merupakan turunan hingga.

Bukti. Estimator regresi nonparametrik pada Persamaan 8, dapat ditulis sebagai:

$$\tilde{m}_n(x) = \frac{N_n(x;h)}{D_n(x;h)} \tag{11}$$

dengan $D_n(x;h) = n^{-1}\sum_{j=1}^{n} K_{x,h}(X_j)$ dan $N_n(x;h) = n^{-1}\sum_{i=1}^{n} Y_i K_{x,h}(X_i)$. Konvergensi dari $D_n(x;h)$ ke $f(x)$ ditentukan menurut [1] dan dengan cara yang sama konvergensi dari $N_n(x;h)$ ke $mf(x)$ dapat diperoleh. Kemudian, berdasarkan deret Taylor dapat ditulis sebagai berikut:

$$\begin{aligned}\tilde{m}_n(x) = m(x) &+ \frac{1}{f(x)}[N_n(x;h) - (mf)(x)] - \frac{(mf)(x)}{f^2(x)}[D_n(x;h) - f(x)] \\ &- \frac{1}{f^2(x)}[N_n(x;h) - (mf)(x)][D_n(x;h) - f(x)] \\ &+ \frac{N_n(x;h)}{f^3(x)}[D_n(x;h) - f(x)]^2[1 + o(1)]a.s\end{aligned} \tag{12}$$

Ekspektasi dari $D_n(x;h)$ dapat didekati dengan

$$E[D_n(x;h)] = f[E(K_{x,h})] + \frac{1}{2}Var(K_{x,h})f^{(2)}(x) + o(h) \quad , h \to 0$$

Dengan cara yang sama, untuk ekspektasi dari $N_n(x;h)$ diperoleh:

$$E[N_n(x;h)] = (mf)[E(K_{x,h})] + \frac{1}{2}Var(K_{x,h})(mf)^{(2)}(x) + o(h) \qquad (13)$$

Dengan demikian, berdasarkan asumsi 6 dari gabungan kernel diskrit, diperoleh:

$$Bias[D_n(x;h)] = E[D_n(x;h)] - f(x) = \frac{1}{2}Var(K_{x,h})f^{(2)}(x) + o(h) \qquad (14)$$

dan

$$Bias[N_n(x;h)] = E[N_n(x;h)] - (mf)(x) = \frac{1}{2}Var(K_{x,h})(mf)^{(2)}(x) + o(h) \quad (15)$$

Selanjutnya,
$E\Big[N_n(x;h)[D_n(x;h) - f(x)]^2\Big]$

$$= O(1/n)^2 + O(1/n) + E[D_n(x;h) - f(x)]^2 E[N_n(x;h)] \qquad (16)$$

sehingga diperoleh

$$E[\tilde{m}_n(x)] - m(x) = \Big[\frac{(mf)^{(2)}(x)}{f(x)} - \frac{mf^{(2)}(x)}{f(x)}\Big]\frac{Var(K_{x,h})}{2} + O(1/n) + o(h)$$

dan

$$\begin{aligned} MSE(x) &= Var[\tilde{m}_n(x)] + Bias^2[\tilde{m}_n(x)] \\ &= \Big[m^{(2)}(x) + 2m^{(1)}(x)\Big(\frac{f^{(1)}}{f}\Big)(x)\Big]^2 \frac{Var^2(K_{x,h})}{4} \\ &= \frac{E(Y_1^2|X_1 = x) - f(x)E^2(Y_1|X_1 = x)}{nf(x)}[P(K_{x,h} = x)]^2 \\ &\quad + o\Big(h^2 + \frac{1}{n}\Big) \end{aligned} \qquad (17)$$

3 ESTIMASI REGRESI SEMIPARAMETRIK

Bentuk estimator semiparametrik dari m dapat ditulis sebagai:

$$\hat{m}_n(x) = r_0(x)\tilde{\omega}_n(x) = \sum_{i=1}^{n} \frac{Y_i K_{x,h}(X_i)}{\sum_{j=1}^{n} K_{x,h}(X_j)} \times \frac{r_0(x)}{r_0(X_i)}, \quad x \in \mathbf{N} \qquad (18)$$

Teorema 2. *Misalkan diberikan x titik di* $\mathbf{N}$ *yang memenuhi* $f(x) = P(X = x) > 0$. *Diasumsikan bahwa fungsi regresi memenuhi* $m(x) = r_0(x)\omega(x)$ *dengan* $r_0(x) = r(x;\beta_0)$ *sebagai awal tetap. Maka, dengan syarat* $h = h(n) \to 0$ *untuk* $n \to \infty$, *estimator* $\hat{m}_n(x)$ *membuktikan*

$$Bias[\hat{m}_n(x)] = \left[r_0(x)\omega^{(2)}(x) + 2r_0(x)\omega^{(1)}(x)\left(\frac{f^{(1)}}{f}\right)(x)\right]\frac{Var(K_{x,h})}{2} + O\left(\frac{1}{n}\right) + o(h) \tag{19}$$

$$Var[\hat{m}_n(x)] = \frac{Var(Y|X=x)}{nf(x)}[P(K_{x,h} = x)]^2 + o\left(\frac{1}{n}\right) \tag{20}$$

dengan $f^{(1)}, \omega^{(1)}$ *dan* $\omega^{(2)}$ *merupakan beda hingga.*

Bukti. Bukti dari teorema ini diperoleh dengan cara yang sama pada Teorema 1. Estimator semiparametrik $\hat{m}_n(x)$ dapat ditulis sebagai

$$\hat{m}_n(x) = \frac{H_n(x;h)}{F_n(x;h)}$$

dan dengan deret Taylor diperoleh persamaan yang sama seperti Persamaan 12 dengan $H_n(x;h) = n^{-1}\sum_{i=1}^{n}[r_0(x)/r_0(X_i)]Y_iK_{x,h}(X_i)$ dan $F_n(x;h) = n^{-1}\sum_{j=1}^{n}K_{x,h}(X_j) = D_n(x;h)$. Nilai ekspektasi dari $\hat{m}_n(x)$ yaitu

$$E[H_n(x;h)] = (mf)(x) + \frac{1}{2}Var(K_{x,h})r_0(x)(\omega f)^2(x) + o(h) \tag{21}$$

dan nilai biasnya adalah sebagai berikut

$$\begin{aligned} Bias[H_n(x;h)] &= E[H_n(x;h)] - (mf)(x) \\ &= \frac{1}{2}Var(K_{x,h})r_0(x)(\omega f)^2(x) + o(h) \end{aligned}$$

Kemudian untuk variansi,

$$Var[H_n(x;h)] = n^{-1}r_0(x)\left[(E(\omega f)^2(K_{x,h}^2)) - (E(\omega f)(K_{x,h}))^2\right]$$

Untuk mendapatkan bias $\hat{m}_n(x)$ dan Var $\hat{m}_n(x)$, menggunakan argumen yang sama seperti dalam bukti Teorema 1.

4 STUDI KASUS

Dalam penelitian ini akan digunakan data sekunder. Data sekunder yang digunakan merupakan informasi data yang diperoleh dari SMA Muhammadiyah Kecamatan Masbagik Kabupaten Lombok Timur Provinsi Nusa Tenggara Barat. Data tersebut berupa data pembelajaran kooperatif tipe jigsaw dan dan aktivitas belajar siswa terhadap prestasi belajar siswa kelas X tahun pelajaran 2011/2012. Data ini digunakan untuk studi kasus dalam mengestimasi model

regresi semiparametrik diskrit dan untuk melakukan analisis data menggunakan program R. Pada penelitian ini yang menjadi variabel independen adalah pembelajaran kooperatif tipe jigsaw dan aktivitas belajar siswa dengan variabel independen adalah prestasi belajar siswa, serta yang menjadi sampel pada penelitian ini adalah kelas X.3 yaitu sebanyak 28 orang siswa.

Analisis data menggunakan program R. Dari hasil analisis data diperoleh hasil sebagai berikut:

1. Komponen Parametrik
   Dengan menggunakan regresi linier diperoleh nilai $\hat{\beta}_0 = 19,8626915$ dan $\hat{\beta}_1 = 0,9539114$.
2. Komponen Nonparametrik
   Dengan menggunakan estimator Nadaraya-Watson diperoleh nilai $MSE = 4605,595$
3. Komponen Semiparametrik
   Dengan menggunakan hasil estimator yang pada sub bab 3 diperoleh nilai $MSE = 88,74602$

5 KESIMPULAN

Berdasar atas hasil dan pembahasan dapat diambil kesimpulan sebagai berikut:

1. Estimator model regresi semiparametrik diskrit diperoleh sebagai berikut:

$$\hat{m}_n(x) = r_0(x)\tilde{\omega}_n(x) = \sum_{i=1}^{n} \frac{Y_i K_{x,h}(X_i)}{\sum_{j=1}^{n} K_{x,h}(X_j)} \times \frac{r_0(x)}{r_0(X_i)}, \quad x \in \mathbf{N}$$

2. Bias asimtotik dan variansi asimtotik dari estimator model regresi semiparametrik diskrit diperoleh sebagai berikut:

$$Bias[\hat{m}_n(x)] = \left[r_0(x)\omega^{(2)}(x) + 2r_0(x)\omega^{(1)}(x)\left(\frac{f^{(1)}}{f}\right)(x)\right]\frac{Var(K_{x,h})}{2} + O\left(\frac{1}{n}\right) + o(h)$$

$$Var[\hat{m}_n(x)] = \frac{Var(Y|X=x)}{nf(x)}[P(K_{x,h}=x)]^2 + o\left(\frac{1}{n}\right)$$

3. Nilai MSE dari model regresi semiparametrik diskrit adalah MSE=88,74602 lebih kecil dibandingkan dengan nilai MSE dari model regresi nonparametrik MSE=4605,595. Jadi, dapat dikatakan bahwa dalam hal ini estimator regresi semiparametrik diskrit lebih baik dibandingkan dengan estimator regresi nonparametrik.

Daftar Pustaka

[1] Abdous, B., Kokonendji, C. C., dan Senga Kiesse, T. 2010. On Semiparametric Regression for Count Explanatory Variables *Journal of Statistical Planning dan Inference.* 6:1537-1548.

[2] Kokonendji, C. C., dan Senga Kiesse, T. 2011. Discrete Associated Kernels Method and Extensions. *Statistical Methodology*. 8:497-516.
[3] Draper, R. N., dan Smith, H. 1996. *Applied Regression Analysis*. Johan Wiley dan Sons, Inc.
[4] Fan, J., Wu, Y., dan Feng, Y. 2009. Local Quasi-Likelihood with a Parametric Guide. *The Annals of Statistics*. 37:4153-4283.
[5] Hardle, W. 1991. *Smoothing Techniques with Implementation in S*. SpringerVerlag. New York.
[6] Hastie, T. J., dan Tibshirani. 1990. *Generalized Additive Model 4th ed.* Chapman dan Hall. London.
[7] Martins-Filho, C., Mishra, S., dan Ullah, A. 2008. A Class of Improved Parametrically Guided Nonparametric Regression Estimators. *Econometrics Reviews*. 27:542-573.
[8] Ruppert, D., Wand, M. P., dan Carrol, R. J. 2003. *Semiparametric Regression*. Cambridge University. United Kingdom.
[9] Wand, M. P., dan Jones, M. C. 1995 *Kernel Smoothing*. Chapman dan Hall. London.