REGRESI NONPARAMETRIK KERNEL *ADJUSTED*

Novita Eka Chandra

Universitas Islam Darul ‘Ulum Lamongan
novitaekachandra@gmail.com

Sri Haryatmi dan Zulaela

Jurusan Matematika FMIPA UGM

ABSTRACT. *Nadaraya Watson's kernel adjusted regression estimator is an estimator whose kernel is taken from the family of scale-location associated with the classical kernel density estimator. Based on these estimator, it can be obtained optimal bandwith and scale parameter. This estimator gives a better estimation results compared with Naradaya Watson's classical kernel regression estimator. This is proven by the small grade MSE which is given by this estimator.*
Keywords*: estimator, Nadaraya Watson kernel, location-scale family.*

ABSTRAK. *Estimator regresi kernel Nadaraya Watson adjusted merupakan estimator yang kernelnya diambil dari keluarga skala-lokasi yang berkaitan dengan estimator densitas kernel klasik. Berdasarkan estimator tersebut dapat diperoleh nilai bandwidth dan parameter skala yang optimal. Estimator ini memberikan hasil estimasi yang lebih baik dibandingkan dengan estimator regresi kernel Nadaraya Watson klasik. Hal ini terlihat berdasarkan nilai MSE yang diberikan estimator ini lebih kecil.*
Kata Kunci*: estimator, kernel Nadaraya Watson, keluarga skala-lokasi.*

1. PENDAHULUAN

Hubungan fungsional antara variabel respon dan variabel prediktor dinyatakan ke dalam bentuk model regresi. Pada model regresi terdapat suatu fungsi yang disebut fungsi regresi. Untuk mengestimasi fungsi regresi yang belum diketahui bentuknya digunakan pendekatan regresi nonparametrik. Estimasi fungsi regresi nonparametrik menggunakan metode *smoothing* tertentu. Salah satunya dengan estimator kernel Nadaraya Watson.

Pada regresi nonparametrik kernel klasik, pemilihan *bandwidth* merupakan masalah utama dari estimasi kernel. Sebaliknya, pemilihan kernel tidak memberikan pengaruh yang nyata dari suatu estimasi. Berdasarkan hal tersebut, muncullah suatu gagasan baru yaitu estimasinya tidak difokuskan pada *bandwidth* saja. Akan tetapi, kernel memiliki peran yang jauh lebih besar dari biasanya. Kernel akan diambil dari keluarga skala-lokasi yang berkaitan dengan estimator densitas kernel klasik. Estimasi ini dikenal dengan estimasi densitas kernel *adjusted* yang telah diper\-kenalkan oleh Srihera dan Stute (2011). Selanjutnya, estimasi tersebut diperluas oleh Eichner dan Stute (2012) dalam kasus regresi, yang kemudian disebut regresi nonparametrik kernel *adjusted*.

Ukuran kebaikan estimator dapat dilihat berdasarkan tingkat kesalahannya. Semakin kecil tingkat kesalahannya semakin baik estimasinya. Ukuran kesalahan tersebut dapat dilihat dari MSE (*Mean Square Error*). Pada penelitian ini akan dibandingkan kinerja dari estimator kernel Nadaraya-Watson *adjusted* dan estimator kernel Nadaraya Watson klasik yang dilihat berdasarkan nilai MSE masing-masing estimator.

2. ESTIMATOR DENSITAS KERNEL *ADJUSTED*

Menurut Srihera dan Stute (2011) dalam tulisannya, diambil fungsi kernel K^* yang berasal dari keluarga skala-lokasi yang berkaitan dengan estimator densitas kernel klasik, yaitu

$$K^*(x) = \sigma \hat{f}_h(\sigma x + \theta) \tag{1}$$

dengan $\theta \in \mathbb{R}$ dan $\sigma > 0$ merupakan parameter lokasi dan skala.

Berdasarkan definisi kernel yang diambil pada Persamaan 1, estimator densitas kernel *adjusted* adalah

$$\hat{f}_{hA}(x) = \frac{\sigma}{n^2 h^2} \sum_{j=1}^{n} \sum_{i=1}^{n} K\left(\frac{\sigma x - \sigma X_i + h\theta - hX_j}{h^2} \right) \tag{2}$$

Teorema 2.1. Diberikan bahwa K merupakan fungsi simetris dan f terdiferensialkan dua kali secara kontinu pada x, serta $E\left(X^2\right)<\infty$. Selanjutnya, jika $n\to\infty$ dan $h\to 0$ sedemikian hingga $nh\to\infty$, maka

$$Bias\left[\hat{f}_{hA}(x)\right]=\frac{f'(x)h}{\sigma}\int f(y)(\theta-y)dy+\frac{f''(x)h^2}{2\sigma^2}\int f(y)(\theta-y)^2\,dy+o\left(h^2\right)$$

dan

$$nhVar\left[\hat{f}_{hA}(x)\right]=\sigma f(x)\int f^2(y)\,dy+o(1).$$

Kedua persamaan tersebut berlaku untuk $\theta\in\mathbb{R}$ dan $\sigma>0$.

Bukti:

Dari Persamaan 2, diperoleh

$$E\left[\hat{f}_{hA}(x)\right]=f(x)+\frac{hf'(x)}{\sigma}\int f(y)(\theta-y)\,dy+\frac{h^2 f''(x)}{2\sigma^2}\int f(y)(\theta-y)^2\,dy+o\left(h^2\right).$$

Untuk $h\to 0$, $E\left[\hat{f}_{hA}(x)\right]$ konvergen ke nol. Dengan demikian, estimator $\hat{f}_{hA}(x)$ merupakan estimator tak bias asimtotik bila h konvergen ke nol. Selanjutnya, nilai bias dari Persamaan 2 adalah

$$Bias\left[\hat{f}_{hA}(x)\right]=\frac{hf'(x)}{\sigma}\int f(y)(\theta-y)\,dy+\frac{h^2 f''(x)}{2\sigma^2}\int f(y)(\theta-y)^2\,dy+o\left(h^2\right)$$

dan nilai variansi dari Persamaan 2 sebagai berikut:

$$Var\left[\hat{f}_{hA}(x)\right]=\frac{\sigma}{nh}f(x)\int f^2(y)\,dy+o\left(\frac{1}{nh}\right),h\to 0.$$

$\square$

Berdasarkan Teorema 2.1, diperoleh nilai MSE untuk $\hat{f}_{hA}(x)$ yaitu

$$MSE\left[\hat{f}_{hA}(x)\right]=\frac{\sigma}{nh}f(x)\int f^2(y)\,dy+\left[\frac{hf'(x)}{\sigma}\int f(y)(\theta-y)\,dy+\right.$$

$$\left.\frac{h^2 f''(x)}{2\sigma^2}\int f(y)(\theta-y)^2\,dy\right]^2+o\left(\frac{1}{nh}\right)+o\left(h^4\right),h\to 0,nh\to\infty.$$

Untuk $h\to 0$ dan $nh\to\infty$, nilai MSE dari estimator $\hat{f}_{hA}(x)$ konvergen ke nol. Dengan demikian, estimator $\hat{f}_{hA}(x)$ merupakan estimator yang konsisten, sehingga dapat ditulis $\hat{f}_{hA}(x)\xrightarrow{p} f(x)$.

Selanjutnya, diambil $\theta=E(X)$, dengan begitu nilai MSE dari $\hat{f}_{hA}(x)$ menjadi

$$MSE\left[\hat{f}_{hA}(x)\right]\approx\frac{\sigma}{nh}f(x)\int f^2(y)\,dy+\left[\frac{h^2 f''(x)}{2\sigma^2}Var(X)\right]^2$$

dan diperoleh nilai *bandwidth* optimal dan σ optimal dari $MSE\left[\hat{f}_{hA}(x)\right]$, yaitu

$$h_{opt1}=n^{-1/5}$$

$$\sigma_{opt1}=\left(\frac{\left(f''(x)Var(X)\right)^2}{f(x)\int f^2(y)\,dy}\right)^{-1/5}.$$

3. ESTIMATOR REGRESI KERNEL NADARAYA WATSON *ADJUSTED*

Estimator regresi kernel Nadaraya Watson *adjusted* merupakan perluasan dari estimator regresi kernel Nadaraya Watson klasik. Seperti yang telah dijelaskan sebelumnya, bahwa kernel yang diambil berdasarkan keluarga skala lokasi dari estimator densitas kernel klasik. Dengan demikian, untuk menentukan estimator regresi kernel Nadaraya Watson *adjusted* analog dengan estimator regresi kernel Nadaraya Watson klasik. Berdasarkan estimator Nadaraya Watson, bentuk estimator regresi kernel Nadaraya Watson *adjusted* yaitu

$$\hat{m}_{hA}(x)=\frac{\dfrac{\sigma}{n^2h^2}\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n}K\left(\frac{\sigma x-\sigma X_i+h\theta-hX_j}{h^2}\right)Y_i}{\hat{f}_{hA}(x)} \tag{3}$$

Persamaan 3 dapat ditulis menjadi

$$\hat{m}_{hA}(x) = \frac{1}{n}\sum_{i=1}^{n} W_{ni}(x) Y_i \tag{4}$$

dengan $W_{ni}(x) = \dfrac{\dfrac{\sigma}{nh^2}\displaystyle\sum_{j=1}^{n} K\left(\dfrac{\sigma x - \sigma X_i + h\theta - hX_j}{h^2}\right)}{\hat{f}_{hA}(x)} \geq 0$ dan $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} W_{ni}(x) = 1$ untuk semua $x \in \mathbb{R}$.

Berikut ini beberapa kondisi yang diperlukan, yaitu:

1. $E\left(Y^2\right) < \infty$ dan $E\left(X^2\right) < \infty$.
2. X mempunyai fungsi densitas f yang terdiferensial secara kontinu pada x dan $f(x) > 0$.
3. m terdiferensial secara kontinu dua kali pada x.
4. K merupakan fungsi densitas peluang yang memenuhi $K(-x) = K(x)$ untuk setiap $x \in \mathbb{R}$. Selanjutnya, K mempunyai momen ketiga berhingga, yaitu $\int x^3 K(x)\,dx < \infty$.

Selanjutnya, didefinisikan

$$\begin{aligned}\hat{m}_{hA}(x) - m(x) &= \frac{1}{n}\sum_{i=1}^{n} W_{ni}(x)\{Y_i - m(X_i)\} + \frac{1}{n}\sum_{i=1}^{n} W_{ni}(x)\{m(X_i) - m(x)\} \\ &= I_1 + I_2 \end{aligned} \tag{5}$$

Teorema 3.1. Berdasarkan kondisi (1) sampai (4), untuk $h \to 0$ dan $nh \to \infty$

$$(nh)^{1/2} I_1 \xrightarrow{d} N\left(0, \rho^2(x)\right)$$

dan

$$I_2 = h^2 B(x) + o\left(h^2\right),$$

dengan

$$\rho^2(x) = \frac{\sigma\sigma_1^2(x)\int f^2(u)\,du}{f(x)}$$

merupakan variansi asimtotik dari $(nh)^{1/2} I_1$ dan

$$B(x) = \frac{\{2f'(x)m'(x) + f(x)m''(x)\}Var(X)}{2\sigma^2 f(x)}$$

adalah bias asimtotik dari $h^{-1}I_2$.

Bukti:

Pertama-tama, akan dicari nilai ekspektasi dari I_1, yaitu

$$E[I_1] = 0.$$

Selanjutnya, akan dicari variansi dari I_1, yaitu

$$Var[I_1] = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} W_{ni}(x)\{Y_i - m(X_i)\}\right)^2\right] - \left(E\left[\frac{1}{n}\sum_{i=1}^{n} W_{ni}(x)\{Y_i - m(X_i)\}\right]\right)^2 \tag{6}$$

Perhatikan persamaan pertama pada Persamaan (6)

$$\left(\frac{1}{n}\sum_{i=1}^{n} W_{ni}(x)\{Y_i - m(X_i)\}\right)^2 = \frac{\dfrac{\sigma^2}{n^4h^4}\displaystyle\sum_{i=1}^{n}\sigma_1^2(X_i)\left[\sum_{j=1}^{n} K\left(\frac{\sigma x - \sigma X_i + h\theta - hX_j}{h^2}\right)\right]^2}{\hat{f}_{hA}^2(x)} \tag{7}$$

Karena $\hat{f}_{hA}(x) \xrightarrow{p} f(x)$, maka $\hat{f}_{hA}^2(x) \xrightarrow{p} f^2(x)$ untuk $h \to 0$ dan $nh \to \infty$.

Perhatikan pembilang pada Persamaan (7), diperoleh nilai ekspektasinya yaitu

$$E\left[\frac{\sigma^2}{n^4h^4}\sum_{i=1}^{n}\sigma_1^2(X_i)\left[\sum_{j=1}^{n} K\left(\frac{\sigma x - \sigma X_i + \theta h - hX_j}{h^2}\right)\right]^2\right] = \frac{\sigma\sigma_1^2(x)f(x)\int f^2(u)\,du}{nh}, h \to 0.$$

Dengan demikian, variansi dari I_1 yaitu

$$Var\left[I_1\right] = \frac{\sigma\sigma_1^2\left(x\right)\int f^2\left(u\right)du}{nhf\left(x\right)}.$$

Berdasarkan Teorema Limit Pusat, diperoleh

$$\left(nh\right)^{1/2} I_1 \xrightarrow{d} N\left(0, \rho^2\left(x\right)\right)$$

dengan $\rho^2\left(x\right) = \dfrac{\sigma\sigma_1^2\left(x\right)\int f^2\left(u\right)du}{f\left(x\right)}$.

Selanjutnya, perhatikan pembilang untuk I_2 dalam Persamaan (5), nilai ekspektasinya adalah

$$E\left[\frac{\sigma}{n^2h^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K\left(\frac{\sigma x - \sigma X_i + h\theta - hX_j}{h^2}\right)\left\{m\left(X_i\right) - m\left(x\right)\right\}\right]$$

$$= \frac{f'\left(x\right)m'\left(x\right)h^2 Var\left(X\right)}{\sigma^2} + \frac{f\left(x\right)m''\left(x\right)h^2 Var\left(X\right)}{2\sigma^2} + o\left(h^2\right)$$

dengan $\theta = E\left(X\right)$. Karena $\hat{f}_{hA}\left(x\right) \xrightarrow{p} f\left(x\right)$, maka diperoleh

$$I_2 = \frac{\left\{2f'\left(x\right)m'\left(x\right) + f\left(x\right)m''\left(x\right)\right\}Var\left(X\right)}{2\sigma^2 f\left(x\right)} + o\left(h^2\right).$$

□

Berdasarkan Teorema 3.1, didapatkan nilai MSE dari $\hat{m}_{hA}\left(x\right)$, yaitu

$$MSE\left[\hat{m}_{hA}\left(x\right)\right] \approx \frac{\sigma\sigma_1^2\left(x\right)\int f^2\left(u\right)du}{nhf\left(x\right)} + \left(\frac{h^2\left\{2f'\left(x\right)m'\left(x\right) + f\left(x\right)m''\left(x\right)\right\}^2 Var\left(X\right)}{2\sigma^2 f\left(x\right)}\right)^2$$

dan didapatkan nilai *bandwidth* optimal dan σ optimal dari $MSE\left[\hat{m}_{hA}\left(x\right)\right]$ yaitu

$$h_{opt2} = n^{-1/5},$$

$$\sigma_{opt2} == \left(\frac{\left\{2f'\left(x\right)m'\left(x\right) + f\left(x\right)m''\left(x\right)\right\}^2 Var\left(X\right)}{f\left(x\right)\sigma_1^2\left(x\right)\int f^2\left(u\right)du}\right)^{-1/5}.$$

4. STUDI KASUS

Padi merupakan salah satu tanaman budidaya yang digunakan sebagai bahan makanan pokok penduduk Indonesia. Salah satu penyakit pada tanaman padi adalah penyakit tungro. Virus tungro ini disebarkan oleh hama wereng hijau. Semakin banyak populasi wereng hijau, maka luas lahan serangannya pun bertambah. Selain itu, penyakit tungro dapat menyebabkan gagal panen dan kerugian secara ekonomi.

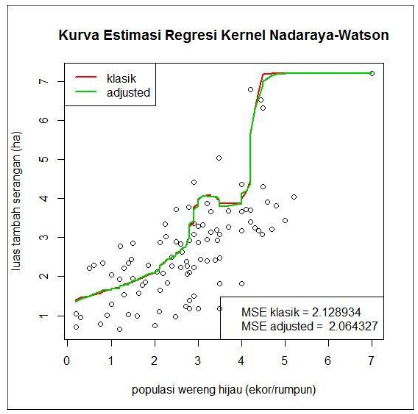
Data yang digunakan dalam penelitian ini merupakan data sekunder yaitu data serangan virus tungro tanaman padi di provinsi Nusa Tenggara Barat pada tahun 2003 - 2010. Data ini bersumber dari Balai Proteksi Tanaman Pangan dan Holtikultura (BPTPH) provinsi Nusa Tenggara Barat dan diambil dari skripsi yang ditulis oleh Fitasari (2011). Dalam studi kasus ini, variabel prediktornya X adalah populasi wereng hijau (ekor per rumpun), sedangkan variabel responnya Y adalah luas tambah serangan (ha). Selanjutnya, data ini digunakan untuk membandingkan nilai MSE dari estimator regresi kernel Nadaraya Watson klasik dengan estimator regresi kernel Nadaraya Watson *adjusted*.

Pengolahan data menggunakan program R.3.0.3, digunakan pula fungsi kernel *Gaussian* dan nilai parameter skala-lokasi yang telah ditetapkan yaitu nilai θ adalah mean dari populasi wereng hijau dan σ adalah standar deviasi dari populasi wereng hijau. Dari hasil pengolahan data, diperoleh beberapa hasil estimasi pada tabel berikut:

Tabel 1. Hasil estimasi.

x	$\hat{m}_h(x)$	$\hat{m}_{hA}(x)$
0,38	1,48	1,46
4,25	3,90	3,86
6,53	7,17	6,96

Berdasarkan Tabel 1, baik dengan estimator regresi kernel klasik maupun *adjusted* diperoleh hasil estimasi yang tidak jauh berbeda. Disimpulkan bahwa semakin banyak populasi wereng hijau, maka semakin besar luas tambah serangannya. Selain itu, diperoleh pula kurva estimasi regresi kernel Nadaraya Watson sebagai berikut:



Gambar 1. Kurva estimasi regresi kernel Nadaraya Watson

Terlihat bahwa kurva kedua estimator tersebut memiliki bentuk kurva yang hampir berhimpitan sehingga sulit untuk menentukan estimator mmana yang memiliki kinerja yang lebih baik. Dengan demikian, digunakan nilai MSE untuk membandingkan kinerja kedua estimator. Berdasarkan nilai output yang didapatkan, disimpulkan bahwa estimator regresi kernel Nadaraya Watson *adjusted* lebih baik. Hal ini terlihat berdasarkan nilai MSE yang lebih kecil.

5. KESIMPULAN DAN SARAN

Berdasarkan pembahasan yang telah dilakukan, diperoleh bentuk estimator regresi kernel Nadaraya Watson *adjusted* pada Persamaan (3). Dalam hasil studi kasus dengan data serangan virus tungro pada tanaman padi di provinsi Nusa Tenggara Barat diperoleh bahwa hasil estimasi dengan kedua estimator yaitu klasik maupun *adjusted* tidak jauh berbeda. Untuk membandingkan kedua estimator tersebet dilihat berdasarkan nilai MSE terkecil. Karena nilai MSE estimator regresi kernel Nadaraya Watson *adjusted* lebih kecil, maka estimator regresi kernel Nadaraya Watson *adjusted* memberikan hasil estimasi yang lebih baik dibandingkan estimator regresi kernel Nadaraya Watson klasik.

DAFTAR PUSTAKA

Azizah, Vita, 2008, Aplikasi Regresi Nonparametrik Kernel dalam Data Finansial, *Skripsi*, Universitas Gadjah Mada, Yogyakarta.

Bain, L. J. dan Engelhardt, Max, 1992, *Introduction to Probability and Mathematcal Statistics*, Duxbury Press, California.

Eichner, Gerrit dan Stute, Winfried, 2012, Kernel Adjusted Nonparametric Regression, *Journal of Statistical Planning and Inference*, 142, pp. 2537-2544.

Fitasari, B. D., 2011, Metode Dekomposisi untuk Peramalan Serangan Virus Tungro pada Tanaman Padi di Provinsi Nusa Tenggara Barat, *Skripsi*, Universitas Mataram, Mataram.

Hardle, Wolfgang, 1991, *Smoothing Techniques with Implementation in S*, Springer-Verlag, New York.

Hardle, Wolfgang, 1994, *Applied Nonparametric Regression*, Berlin.

Hayati, Laila, 2007, Regresi Nonparametrik untuk Mengestimasi Total Populasi, *Tesis*, Universitas Gadjah Mada, Yogyakarta.

Holmes, M. H., 2013, *Introduction to Pertubation Methods*, Springer, New York.

Lael, O. A. F., 2013, Estimator Regresi Kernel Nadaraya Watson Adaptif, *Tesis*, Universitas Gadjah Mada, Yogyakarta.

Pujiastuti, C. E., 1996, Deret Fourier dan Kernel dalam Regresi Nonparametrik, *Tesis*, Universitas Gadjah Mada, Yogyakarta.

Purnamasari, Y. K., 2013, Perbandingan Model Regresi Nonparametrik Spline dan Regresi Nonparametrik Kernel, *Skripsi*, Universitas Gadjah Mada, Yogyakarta.

Roussas, G.G., 1997, *A Course in Mathematical Statistics Second Edition*, Academic Press, United States of America.

Srihera, Ramidha dan Stute, Winfried, 2011, Kernel Adjusted Density Estimation, *Statistics and Probability Letters*, 81, pp. 571-579.

Wand, M. P. dan Jones, M. C., 1995, *Kernel Smoothing*, Chapman and Hall, London.