PERBADINGAN BIAS ESTIMATOR KERNEL NADARAYA-WATSON DAN *LOCALLY LINEAR* PADA REGRESI NONPARAMETRIK

BIAS COMPARISON NADARAYA WATSON AND LOCALLY LINEAR KERNEL ESTIMATOR OF NONPARAMETRIC REGRESSION

Zulfikar
Dosen pada Jurusan Sistem Informasi dan Teknik Informatika
STMIK Bahrul Ulum, Jombang
hhtp//www.stmikbu.ac.id Email:stmikbu@telkom.net

Abstrak

Diberikan data (x_i, y_i) serta hubungan antara x_i dan y_i diasumsikan mengikuti model regresi nonparametrik:

$$y_i = m(x_i) + \varepsilon_i, \quad i = 1, 2, ..., n.$$

Kurva regresi *m* diasumsikan bentuknya tidak diketahui dan ε_i sesatan random berdistribusi independen dengan mean nol dan variansi σ^2.

Dalam penelitian ini akan dikaji bentuk estimator untuk *m*(*x*) dengan polinomial lokal derajat *p*. Jika *p* = 0 diperoleh estimator Nadaraya-Watson. Untuk *p* = 1 diperoleh estimator *Locally Linear*. Selanjutnya dibandingkan nilai bias, variansi dan *MSE* dari kedua estimator tersebut. Aplikasi estimator kernel dilakukan pada data Canadian Males dari hasil penelitian Murphy dan Welch (1990).
Kata Kunci: Estimasi Nonparametrik,least square terbobot , Polinomial Lokal

Abstract

Given a data set (x_i, y_i) and connecting between x_i and y_i be assumed to follow nonparametric regression model :

$$y_i = m(x_i) + \varepsilon_i, \quad i = 1, 2, ..., n.$$

Regresssion curve of m be assumed is an unknown form and ε_i, is an error term in the observations are IID with mean 0 and finite variance σ^2.

In this paper propose to exist mean conditional estimators with employ the local polinomial method which polinomial degree p = 0 will be formed the Nadaraya-Watson estimator and p = 1 to exist the Locally Linear estimator. Furthemore, with the same method also be existed the comparison both bias and variance. Kernel estimator will be applied of the Canadian Males Data by Murphy and Welch (1990).
Key words: Nonparametric estimation, weighted least square, Local polinomial

1. Pendahuluan

Dalam proses estimasi nonparametrik, estimator kernel menghasilkan bias yang sangat menggganggu inferensi. Bias memiliki dua sumber , yaitu kurva bias dan *boundary bias* (Racine, 1990). Pertama, kernel order ke-*p*, bias tergantung pada turunan ke-*p*. Kernel order kedua misalnya menghasilkan bias yang berkaitan dengan turunan kedua. Kedua, estimator kernel memiliki kecepatan konvergensian yang lambat dan memiliki bias yang tinggi dekat batas-batas interval. Tipe ini ditunjukkan sebagai *boundary bias.*

Beberapa pendekatan reduksi bias dalam estimasi bertumpu pada sifat-sifat asimtotik, penggunaan kernel order tinggi dan metode resampling. Hardle dan Bowman (1988) menunjukkan bahwa ekspansi asimtotik untuk reduksi bias. Pendekatan reduksi bias hanya menggunakan ekspansi dan tidak menggunakan order lebih rendah. Pendekatan reduksi bias terkoreksi dengan bootstrap untuk estimator kernel yang dilakukan Hardle dan Marron (1991). Pendekatan lain untuk mereduksi bias adalah dengan kernel order tinggi. Bartlett (1963) mempertimbangkan reduksi bias berdasarkan *Mean Square Error* (*MSE*). Gasser dan Muller (1979) dan Rice (1984) menunjukkan bahwa estimator kernel Nadaraya-Watson memiliki kecepatan konvergensi yang lambat dan bias yang tinggi dekat batas-batas pada interval. Gasser dan Muller (1979) menunjukkan perilaku asimtotik secara global dan pengaruhnya menjadi buruk untuk kernel order tinggi.

Dalam penelitian ini digunakan estimator kernel Nadaraya-Watson dan *Locally Linear* untuk mengkaji bias regersi kernel order rendah, seperti kernel order dua.

2. Estimator Kernel pada Fungsi Regresi

Asumsi dasar regresi nonparametrik adalah keberadaan fungsi penghalus *m*(.) dari hubungan respon *y* dan prediktor *x*, yaitu :

$$Y_i = m(x_i) + \varepsilon_i, i = 1,2,...,n \quad (1)$$

dimana $\varepsilon_i \sim N(0,\sigma^2)$ dan $m(x_i)$ kurva regresi yang bentuknya tidak diketahui. Fungsi regresi $m(x_i)$ pada model regresi nonparametrik dapat diestimasi dengan pendekatan kernel yang didasarkan pada fungsi densitas kernel, serta penghalus dengan pendekatan kernel ini selanjutnya dekenal sebagai penghalus kernel (*smoothing kernel*)(Hardle (1991)).

Dua estimator kernel terkenal adalah estimator Nadaraya-Watson dan *Locally Linear*. Metode yang digunakan adalah dengan penetapan polinomial lokal (Wand dan Jones, 1995). Metode ini menetapkan estimasi *least square* terbobot :

$$\beta = (\beta_0, \beta_1, ..., \beta_p)^T,$$

pada dejarat *p* dengan meminimumkan :

$$\sum_{i=1}^{n} \left[Y_i - \beta_0 - \beta_1 (x_i - x) - ... - \beta_p (x_i - x)^p \right]^2 K_h (x_i - x), \quad (2)$$

dimana $K_h(x_i - x)$ sebagai pembobot kernel. Solusi standar adalah estimator (*p* + 1) x1:

$$\hat{\beta} = (X^T{}_x W_x X_x,)^{-1} X^T{}_x W_x Y \quad (3)$$

dalam bentuk matrik singular, dimana Y = $(Y_1,...,Y_n)^T$ adalah vektor respon,

$$X_x = \begin{bmatrix} 1 & x_1 - x & . & . & . & (x_1 - x)^P \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ 1 & x_n - x & . & . & . & (x_n - x)^P \end{bmatrix}$$

adalah sebuah bentuk matrik *n x (p + 1)* dan $W_x = diag\{K_h(x_1 - x),...,K_h(x_n - x)\}$ merupakan matrik diagonal *n x n* pada pembobot. Ketika estimator pada *m(x)* adalah koefisien intercept diperoleh :

$$\hat{m}(x; p, h) = e^T{}_1(X^T{}_x W_x X_x,)^{-1} X^T{}_x W_x Y \quad (4)$$

dimana e_1 adalah vektor *(p + 1) x 1* bernilai 1 dalam masukan pertama dan nol untuk yang lainnya. Schucany (2004) menyatakan bahwa ketika *p* = 0 secara eksplisit akan membentuk formula estimator Nadaraya-Watson dan *p* = 1 akan membentuk estimator *Locally Linear*. Untuk p = 0 diperoleh estimasi :

$$\hat{m}_{NW}(x;0,h) = n^{-1}\sum_{i=1}^{n} K_h(x_i - x)Y_i / \sum_{i=1}^{n} K_h(x_i - x) \quad (5)$$

sebagai bentuk estimator kernel Nadaraya-Watson. Sedangkan untuk *p* = 1 diperoleh estimasi $\hat{m}_{LL}(x;1,h) =$

$$= n^{-1}\sum Y_i \frac{\left[K_h(x_i - x)\sum (x_i - x)^2 K_h(x_i - x) - (x_i - x)K_h(x_i - x)\sum (x_i - x)K_h(x_i - x)\right]}{\sum\left[K_h(x_i - x)\sum (x_i - x)^2 K_h(x_i - x) - (x_i - x)K_h(x_i - x)\sum (x_i - x)K_h(x_i - x)\right]}$$

Jika $\hat{s}_j = \sum (x_i - x)^j K_h(x_i - x), j = 1,2$ maka diperoleh estimator kernel *Locally Linear*

$$\hat{m}_{LL}(x,1,h) = n^{-1}\frac{\sum K_h(x_i - x)\left[\hat{s}_2 - (x_i - x)\hat{s}_1\right]Y_i}{\sum K_h(x_i - x)\left[\hat{s}_2 - (x_i - x)\hat{s}_1\right]} \quad (6)$$

Untuk mendapatkan persamaan bias, dan variansi maka dibuat asumsi sebagai berikut :

(i) Fungsi *m, m'* dan m'' kontinu dalam [0,1].

(ii) Kernel K simetris sekitar nol.

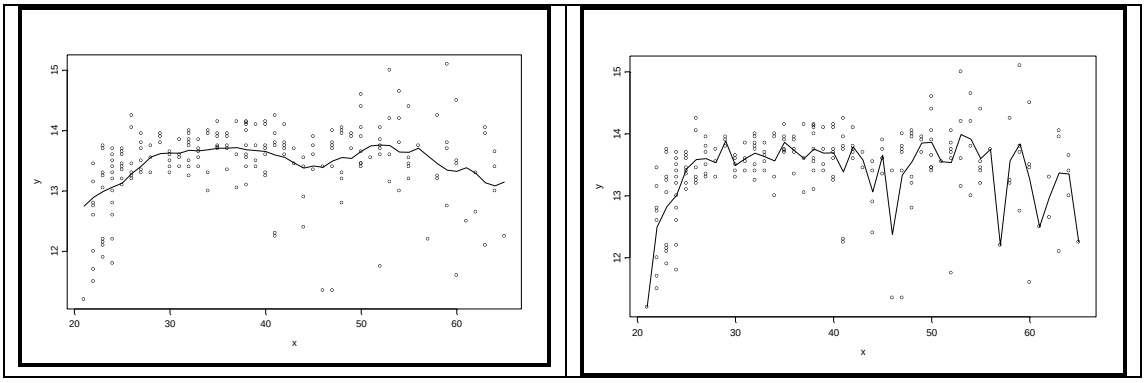
Sehingga diperoleh bias dan variansi sebagai berikut :

Tabel 1. Bias dan Varisnsi estimator kernel order dua.

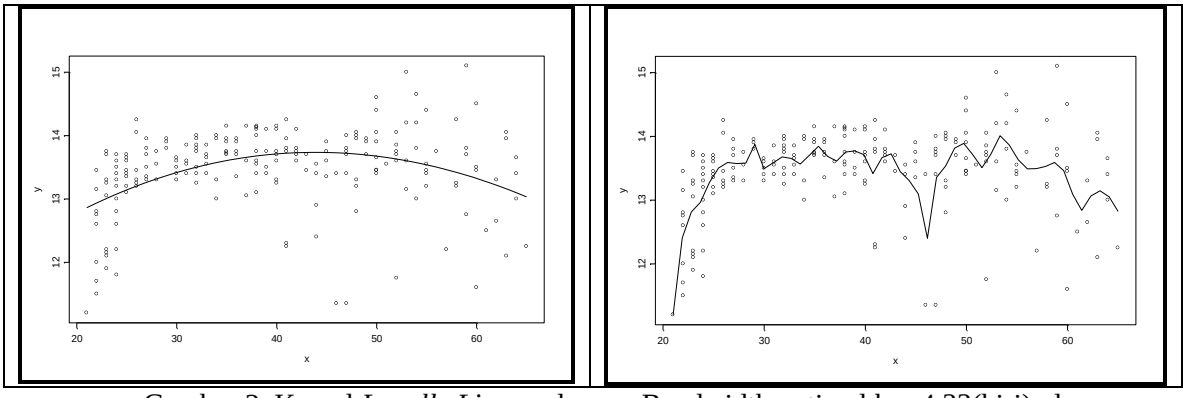
Kernel	Bias	Variansi
Nadaraya-Watson	$Bias\{\hat{m}(x)\} = h_n^2\left(\frac{1}{2}m''(x) + \frac{m'(x)f'(x)}{f(x)}\right)\int z^2K(z)dz$	$Var\{\hat{m}(x)\} = \frac{\sigma^2(x)}{nh_n}\int K^2(z)dz$
Locally Linear	$Bias\{\hat{m}(x)\} = \{\int z^2K(z)dz\}\frac{1}{(2)!}m''h^2$	$Var\{\hat{m}(x)\} = \int K^2(z)dz\frac{\sigma^2(x)}{nh}$

3. Aplikasi Kernel

Untuk mengaplikasikan estimator kernel Nadaraya-Watson dan *Locally Linear* digunakan data penelitian Canadian Males oleh Murphy dan Weclh (1990) dengan melihat hubungan antara pengalaman kerja (*x*) dan besarnya penghasilan (*y*). Bandwidth optimum diperoleh sebesar 7,21 untuk kernel Nadaraya-Watson dan 4,33 untuk kernel *Locally Linear* dengan menggunakan Generalized *Cross Validation* (*GCV*).



Gambar 1. Kernel Nadaraya-Watson dengan Bandwidth optimal h = 7.21(kiri), dan tidak optimal h = 0,5 (kanan)



Gambar 2. Kernel *Locally Linear* dengan Bandwidth optimal h = 4,33(kiri), dan tidak optimal h = 0,1 (kanan)

Pada gambar 1 dan 2 terlihat bahwa kurva regresi dengan bandwidth optimal terlihat smooth dibandingkan dengan bandwidth tidak optimal dan ditunjukkan bahwa nilai bias, variansi dan *MSE* pada bandwidth optimal lebih kecil dibandingkan pada bandwidth tidak

optimal. Nilai variansi dan *MSE* pada *h* optimal lebih kecil untuk kernel Nadaraya-Watson dan *Locally Linear* order dua seperti terlihat pada tabel 2 berikut:

Tabel 2. Perbandingan Nilai, Bias, Variance, *MSE* dan Koreksi Bias antara Dua Estimator

Kernel	Bias	Variance	*MSE*
Nadaraya-Watson			
H = 7,2 (optimal) H = 0,5	0,00647122 0,00595601	0,06750895 0,17100716	0,06755083 0,17010700
Locally Linear			
H = 4,33 (optimal) H = 0,10	0,08190244 0,01783870	0,07979603 0,17838700	0,08650404 0,17875350

Dengan melihat nilai bias, variansi dan *MSE* dapat disimpulkan estimator kernel Nadaraya-Watson lebih baik pada aplikasi data Canadian Males dibandingkan kernel *Locally Linear*.

4. Kernel Nadaraya-Watson Order Tinggi

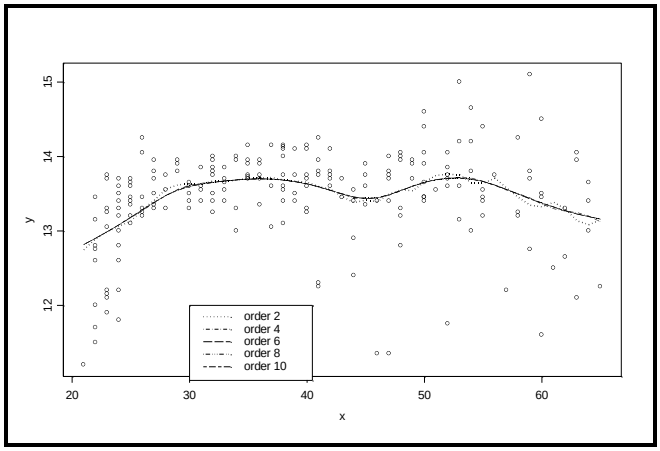
Setelah didapatkan estimator terbaik, maka selanjutnya dilakukan perbandingan dengan kernel order tinggi, yaitu untuk order 4, 6, 8 dan 10. Hasil perhitungan bias, variansi dan *MSE* kernel order tinggi terlihat pada tabel 3 berikut:

Tabel 4.4. Nilai Bias, Variansi dan *MSE* pada estimator kernel Nadaraya - Watson

Kernel	Bias	Variansi	*MSE*
NW Order dua	0,00675854	0,06748859	0,06753422
NW Order empat	0,01112049	0,06064958	0,06061947
NM Order enam	0,01048195	0,06008253	0,06019240
NM Order delapan	0,01040195	0,05925814	0,05925814
NM Order sepuluh	0,01040195	0,05925814	0,05925814

Pada tabel di atas terlihat bahwa semakin tinggi order, nilai variansi dan *MSE* semakin kecil.

Untuk melihat kurva regresi kernel order tinggi disajikan pada gambar 3. Gambar 3 terlihat bahwa pada order 4 permukaan kurva sudah *smooth* dan hampir berimpit dengan order 6, 8 dan 10.



Gambar 3. Kernel Nadaraya-Watson Order Tinggi

Kesimpulan

a. Berdasarkan aplikasi data dari Murphy dan Welch (1990) didapatkan bahwa estimator terbaik dijumpai pada kernel Nadaraya-Watson yang ditunjukkan pada bandwidth optimal diperoleh nilai variansi dan MSE dari estimator kernel Nadaraya-Watson order-2 lebih kecil dari estimator kernel Locally-Linear order-2.
b. Pada estimasi kernel Nadaya-Watson order lebih tinggi (4, 6, 8 dan 10) mempunyai nilai bias, variansi dab MSE yang lebih kecil dari kernel Nadaraya-Watson order-2. Untuk order-4, 6, 8 dan 10 memperlihatkan pola yang hampir berimpit.

Daftar Pustaka

Bartlett, M. S. (1963), Statistical Estimation of Density Functions, *Sankhya* 25, 245-254.

Gasser, T and Muller, H. G(1981), *Kernel estimation of Regression Functions, in Smoothing Techniques for Curve estimation*, Springer-Verlag, Berlin, Heildleberg, New York, pp. 23-68.

Hardle, W. (1991), *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.

Hardle, W. and Bowman, S. (1988), Bootstrapping in Nonparametric Regression Local Adaptive Smoothing and Confidence Bounds, *Journal of the American Statistics Assosiation* 83, 102-110.

Hardle, W. and Maron, S. (1991), Bootstrapping Simultaneous Error Bar for Nonparametric regression. *Annals of Statistics* 19, 778-796.

Murphy, K. M. dan Welch, F. (1990), Empirical Age- Earning Profiles, *Journal of Labour Economics* 8(2), 202-229.

Racine, J. (1998), *Bias-Corrected Kernel Regression*, Department of Economics, University of South Florida, Tampa, FL., USA 33620.

Rice, J. A. (1984), Boundary Modification for Kernel Regression, *Communication in Statistics* 13, 893-900.

Schucany, W. R. (2004), *Kernel Smoothers: An Overview of Curve Estimations for The First Graduate Course in Nonparametric Statistics.* Department of Statistical Science, SMU, Dallas TX.