BuletinIlmiah Mat. Stat. danTerapannya (Bimaster)
Volume 08, No. 4 (2019), hal 633-638.

ESTIMASI MODEL REGRESI NONPARAMETRIK KERNEL MENGGUNAKAN ESTIMATOR *NADARAYA-WATSON*

Nurul Anisa, Naomi Nessyana Debataraja, Shantika Martha

INTISARI

Pendekatan regresi nonparametrik dilakukan untuk memodelkan data yang tidak diketahui bentuk fungsinya. Salah satu regresi nonparametrik yang sering digunakan adalah regresi kernel. Tujuan penelitian ini adalah untuk mengestimasi model regresi nonparametrik menggunakan regresi kernel dengan estimator Nadaraya-Watson pada data indeks pembangunan manusia di Indonesia. Berdasarkan hasil analisis yang telah dilakukan, dapat disimpulkan bahwa untuk data indeks pembangunan manusia diperoleh bandwitdh optimal dengan estimator Nadaraya-Watson sebesar 1,384884. Hasil estimasi tersebut memperoleh nilai koefisien determinasi sebesar 63,2% dan menghasilkan nilai Mean Absolute Percentage Error (MAPE) sebesar 2,5% yang berarti bahwa kemampuan estimasi menggunakan regresi nonparametrik kernel sangat baik.

Kata Kunci: regresi kernel, *bandwidth*, *gaussian*.

PENDAHULUAN

Analisis regresi merupakan analisis data yang menggambarkan hubungan antara variabel dependen dengan satu atau beberapa variabel independen[1]. Dalam analisis regresi ada dua jenis pendekatan yang dapat digunakan untuk mengestimasi kurva regresi, yaitu pendekatan regresiparametrik dan pendekatan regresi nonparametrik. Pendekatan regresi nonparametrik dilakukan untuk memodelkan data yang tidak diketahui bentuk fungsinya. Jenis data yang fluktuatif dan tidak membentuk suatu pola hubungan tertentu akan sulit didekati dengan regresi parametrik sehingga pendekatan nonparametrik merupakan yang paling tepat digunakan. Meskipun regresi nonparametrik merupakan regresi untuk mengatasi pemodelan data yang tidak membentuk pola hubungan tertentu, akan tetapi model regresi nonparametrik tetap dapat digunakan untuk memodelkan data yang berbentuk apa saja, baik linear maupun nonlineardikarenakan tidak adanya asumsi yang harus dipenuhi[2]. Salah satu regresi nonparametrik yang sering digunakan adalah regresi kernel. Kelebihan kernel adalah dapat mencapai tingkat kekonvergenan yang relatif cepat[3]. Pada regresi kernel dikenal suatu estimator yang biasa digunakan untuk mengestimasi fungsi regresi yaitu estimator *Nadaraya-Watson.* Estimasi pendekatan kernel tergantung pada dua parameter yaitu *bandwidth* dan fungsi kernel[4]. Ada beberapa jenis fungsi kernel antara lain *uniform*, *triangle*, *epanechnikov*, *gaussian*, dan *cosinus*[5]. Fungsi kernel yang biasa digunakan adalah fungsi kernel *gaussian*, karena lebih mudah dalam perhitungan serta fungsi bobot kernel tersebut terdefinisi atau memiliki nilai pada semua bilangan riil[4]. *Bandwidth* yang terlalu kecil akan menyebabkan fungsi yang diestimasi tersebut menjadi sangat kasar sehingga hubungan variansinya tinggi dan memiliki potensi bias yang rendah, sebaliknya jika *bandwidth* yang terlalu besar menyebabkan fungsi yang diestimasi tersebut menjadi sangat mulus sehingga hubungan variansinya rendah dan memiliki potensi bias yang besar[4]. Pemilihan *bandwidth* optimal berdasarkan nilai *cross validation* (CV) minimum.

Tujuan yang dicapai dari penelitian ini adalah menentukan estimasi model regesi nonparametrik menggunakan regresi kernel dengan estimator *Nadaraya-Watson.* Penelitian ini menggunakan data sekunder yang diperoleh dari Badan Pusat Statistika. Variabel yang digunakan adalah indeks pembangunan manusia di Indonesia pada tahun 2017 dan angka harapan hidup di Indonesia pada tahun 2017. Metode yang digunakan adalah metode regresi kernel dengan fungsi kernel *gaussian*.

Estimator yang digunakan adalah estimator *Nadaraya-Watson*. Pemilihan parameter menggunakan *cross validation* (CV).

Penelitian ini berupa studi literatur dan studi kasus yang dimulai dengan menentukan fungsi kernel yang akan digunakan. Langkah selanjutnya adalah menentukan *bandwidth* optimal berdasarkan nilai CV minimum. Kemudian dilakukan estimasi model regresi nonparametrik kernel. Setelah itu uji signifikansi parameter untuk mengetahui hubungan parameter di dalam model, serta menentukan nilai *Mean Absolute Percentage Error* (MAPE) untuk mengetahui kebaikan dari suatu model.

ANALISIS KORELASI

Analisis korelasi bertujuan untuk mengukur keeratan dua variabel yaitu dengan kata lain variabel independen (X) dan variabel dependen (Y). Untuk mencari koefisien korelasi dapat digunakan dengan uji hipotesis dan kriteria uji sebagai berikut:

H_0: tidak terdapat korelasi yang signifikan antara variabel dependen dan variabel independen.

H_1: terdapat korelasi yang signifikan antara variabel dependen dan variabel independen.

Statistik uji yang digunakan adalah:

$$r_{XY} = \frac{n\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{\sqrt{\left(n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right)}\sqrt{\left(n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right)}}$$

dimana r_{XY} adalah korelasi antara variabel independen dengan variabel dependen, X_i adalah variabel independen ke-i, Y_i adalah variabel dependen ke-i, dan n adalah jumlah data.

Kriteria pengambilan keputusan yakni tolak H_0 jika nilai $r_{XY} > r_{tabel}$ atau nilai *p-value* $\leq \alpha$.

Nilai koefisien korelasi dapat diinterpretasikan dalam Tabel 1 [6]:

Tabel 1 Tabel Interpretasi Nilai Koefisien Korelasi

Nilai Koefisien Korelasi	Interpretasi
0,00 – 0,29	Korelasi sangat lemah
0,30 – 0,49	Korelasi lemah
0,50 – 0,69	Korelasi cukup
0,70 – 0,79	Korelasi kuat
0,80 – 1,00	Korelasi sangat kuat

REGRESI NONPARAMETRIK

Pada prinsipnya pendekatan nonparametrik dilakukan untuk memodelkan data yang tidak diketahui bentuk fungsinya.Model regresi nonparametrik secara matematis dapat ditulis[2]:

$$Y_i = m(X_i) + \varepsilon_i$$

dimana Y_i adalah variabel dependen ke-i, X_i adalah variabel independen ke-i, $m(X_i)$ adalah fungsi regresi yang tidak diketahui, dan ε_i adalah pengukuran residual yang tidak dapat dijelaskan dengan fungsi regresi $m(X_i)$.

PEMILIHAN *BANDWIDTH* OPTIMAL

Bandwidth adalah parameter pemulus yang berfungsi untuk mengontrol kemulusan dari kurva yang diestimasi dan sebagai ukuran kesesuaian fungsi pada data, sehingga dalam memilih nilai *bandiwdth* diharapkan nilai optimal. *Bandwidth* yang optimal diperoleh dengan menghitung nilai *cross validation* (CV). Pemilihan *bandwidth* yang optimal diperoleh berdasarkan nilai CV minimum. Metode CV dapat dinyatakan sebagai berikut:

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{m}(X_{\neq i})\right)^2$$

dimana CV adalah *cross validation*, h adalah *bandwidth*, Y_i adalah variabel dependen ke-i, dan $\hat{m}(X_{\neq i})$ adalah nilai estimasi.

ESTIMATOR *NADARAYA-WATSON*

Pada regresi kernel dikenal suatu estimator yang biasa digunakan untuk mengestimasi fungsi regresi yaitu estimator *Nadaraya-Watson*. Estimator *Nadaraya-Watson* (N-W) dapat ditulis dalam persamaan sebagai berikut [2]:

$$\widehat{m}(X_i) = \frac{\sum_{j=1}^{n} K\left(\frac{X_i - X_j}{h}\right) Y_j}{\sum_{j=1}^{n} K\left(\frac{X_i - X_j}{h}\right)} \tag{1}$$

dengan

$$W_{ij}(X) = \frac{K\left(\frac{X_i - X_j}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_i - X_j}{h}\right)} \tag{2}$$

dimana $\{W_{ij}(x)\}$ merupakan barisan bobot-bobot positif dan memiliki karakteristik $\sum_{j=1}^{n} W_{ij}(x) = 1$, maka diperoleh:

$$\widehat{m}(X_i) = \sum_{j=1}^{n} W_{ij}(x) Y_j \tag{3}$$

Sehingga estimator *Nadaraya-Watson* (N-W) merupakan rata-rata terboboti dari $\{Y_j\}$.

Bentuk umum fungsi kernel *gaussian* adalah sebagai berikut:

$$K(X) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{X_i - X_j}{h}\right)^2\right) \tag{4}$$

kemudian fungsi tersebut disubstitusikan ke Persamaan (1) menjadi:

$$\widehat{m}(X_i) = \frac{\sum_{j=1}^{n} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{X_i - X_j}{h}\right)^2\right) Y_j}{\sum_{j=1}^{n} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{X_i - X_j}{h}\right)^2\right)} \tag{5}$$

dimana X_i adalah variabel independen ke-i, X_j adalah variabel independen ke-j, Y_j adalah variabel dependen ke-j, dan h adalah *bandwidth*.

MEAN ABSOLUTE PERCENTAGE ERROR (MAPE)

Mean Absolute Percentage Error atau biasa dikenal dengan MAPE adalah cara untuk mengukur keakuratan pendugaan. Nilai MAPE dapat dicari menggunakan rumus:

$$MAPE = \frac{\sum_{i=1}^{n} \left|\frac{\hat{Y}_i - Y_i}{Y_i}\right|}{n} \times 100\%$$

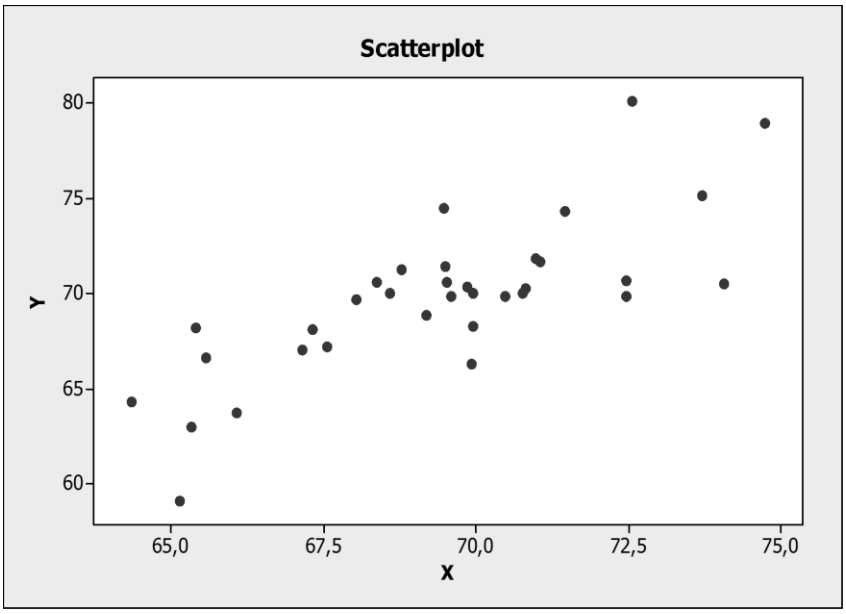
dimana n adalah jumlah data, Y_i adalah variabel dependen ke-i, dan $\hat{Y}_i$ adalah nilai estimasi ke-i.
Kriteria keakuratan MAPE adalah sebagai berikut[7]:

Tabel 2 Tabel Kriteria MAPE

MAPE	Hasil Keputusan
< 10%	Kemampuan prediksi sangat baik
10% - 20%	Kemampuan prediksi baik
20% -50%	Kemampuan prediksi cukup
>50%	Kemampuan prediksi buruk

METODE PENELITIAN

Data pada penelitian ini menggunakan data sekunder yang diambil dari Badan Pusat Statistik (BPS) tahun 2017. Penelitian ini terdiri dari 34 provinsi di Indonesia. Variabel-variabel yang digunakan dalam penelitian ini adalah variabel dependen (Y) yaitu indeks pembangunan manusia di Indonesia dan variabel independen (X) yaitu angka harapan hidup di Indonesia. Langkah awal dalam penentuan model hubungan antara indeks pembangunan manusia dengan masing-masing variabel yaitu membuat *scatterplot* antara variabel dependen dan variabel independen.



Gambar 1 *Scatterplot*

Pada Gambar 1 dapat diketahui bahwa variabel tersebut tidak membentuk pola data yang spesifik. Hal ini menunjukkan bahwa dengan sebaran data yang cenderung acak, bahwa variabel saling bebas.

Setelah itu dilakukan analisis korelasi yang bertujuan untuk mengukur keeratan dua variabel yaitu dengan kata lain variabel X dan Y. Hasil analisis korelasi antara indeks pembangunan manusia (Y) dengan angka harapan hidup (X) menghasilkan nilai korelasi 0,780 lebih besar dari nilai r tabel sebesar 0,339. Maka dapat disimpulkan bahwa terdapat korelasi antara indeks pembangunan manusia dan angka harapan hidup. Korelasi bernilai positif (+) menunjukkan hubungan antara kedua variabel tersebut bersifat positif atau dengan kata lain semakin meningkatnya angka harapan hidup maka akan meningkat pula indeks pembangunan manusia.

Dalam menentukan model regresi nonparametrik kernel, nilai yang ditentukan terlebih dahulu adalah *bandwidth*. Pemilihan *bandwidth* (h) berpengaruh terhadap kemulusan grafik yang akan diperoleh. Oleh karena itu diperlukan *bandwidth* yang optimal dengan kesalahan estimasi yang tidak terlalu besar menggunakan *cross validation* (CV). Dengan menggunakan *software* R diperoleh nilai *bandwidth* yang optimal adalah 1,384884.

Pada regresi nonparametrik kernel untuk mengetahui variabel independen yang berpengaruh terhadap variabel dependen dilakukan dengan pengujian parsial terhadap parameter. Uji parsial dilakukan menggunakan bantuan *software* R, sehingga diperoleh nilai Sig. sebesar 0,00. Nilai Sig. kurang dari nilai alpha (0,05), sehingga dapat disimpulkan bahwa parameter berpengaruh signifikan terhadap model regresi nonparametrik kernel.

Berdasarkan nilai parameter signifikan dengan *bandwidth* yang optimal maka dapat dilakukan estimasi untuk indeks pembangunan manusia di Indonesia. Model regresi nonparametrik kernel dengan estimator *Nadaraya-Watson* berdasarkan fungsi kernel *gaussian* adalah sebagai berikut:

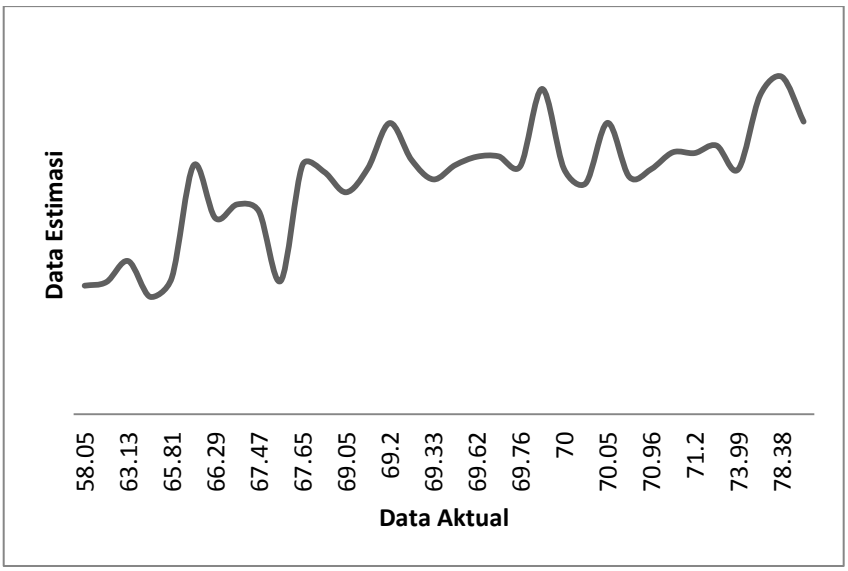
$$\widehat{m}(X_i) = \frac{\sum_{j=1}^{n} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{X_i - X_j}{1{,}384884}\right)^2\right) Y_j}{\sum_{j=1}^{n} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{X_i - X_j}{1{,}384884}\right)^2\right)}$$

Hasil estimasi dari indeks pembangunan manusia di Indonesia menggunakan regresi nonparametrik kernel dengan variabel yang mempengaruhi yaitu angka harapan hidup di Indonesia dapat dilihat pada Tabel 4:

Tabel 4 Hasil Estimasi Indeks Pembangunan Manusia di Indonesia

Provinsi	Hasil Estimasi	Provinsi	Hasil Estimasi
Aceh	70,20245	Nusa Tenggara Barat	65,24685
Sumatera Utara	69,55747	Nusa Tenggara Timur	65,97018
Sumatera Barat	69,84993	Kalimantan Barat	70,36395
Riau	70,94715	Kalimantan Tengah	70,23061
Jambi	70,79035	Kalimantan Selatan	69,21562
Sumatera Selatan	70,05793	Kalimantan Timur	73,7113
Bengkulu	69,72678	Kalimantan Utara	72,41398
Lampung	70,37647	Sulawesi Utara	70,98406
Kep, Bangka Belitung	70,37647	Sulawesi Tengah	68,22386
Kep, Riau	70,18622	Sulawesi Selatan	70,33105
DKI Jakarta	72,50555	Sulawesi Tenggara	70,62076
Jawa Barat	72,41398	Gorontalo	67,90906
Jawa Tengah	74,04572	Sulawesi Barat	64,31629
DI Yogyakarta	74,57499	Maluku	65,07894
Jawa Timur	70,81613	Maluku Utara	68,57918
Banten	70,19029	Papua Barat	64,99655
Bali	71,33506	Papua	64,82841

Berikut adalah kurva dari data aktual dan data estimasi dari indeks pembangunan manusia di Indonesia menggunakan regresi nonparametrik kernel:



Gambar 2 Grafik Data Aktual dan Data Estimasi dengan Regresi Nonparametrik Kernel

Untuk menentukan besar variabel angka harapan hidup dapat menjelaskan variabel indeks pembangunan manusia adalah dengan menghitung koefisien determinasi. Perhitungan koefisien determinasi dilakukan menggunakan bantuan *software* R. Hasil dari perhitungan koefisien determinasi sebesar 0,632. Nilai ini mengandung arti bahwa pengaruh angka harapan hidup terhadap indeks pembangunan manusia adalah sebesar 63,2% sedangkan 36,8% indeks pembangunan manusia dipengaruhi oleh variabel lain yang tidak diteliti. Untuk mengukur keakuratan estimasi dari model regresi nonparametrik kernel adalah dengan menghitung nilai MAPE. Hasil dari perhitungan *Mean Absolute Percentage Error* (MAPE) dari model regresi nonparametrik kernel diperoleh nilai MAPE

0,025 atau 2,5%, maka dapat disimpulkan bahwa kemampuan estimasi menggunakan regresi kernel sangat baik.

KESIMPULAN

Berdasarkan hasil analisis yang telah dilakukan untuk menentukan estimasi model regresi nonparametrik kernel dengan estimator *Nadaraya-Watson*, maka dapat disimpulkan bahwa untuk data indeks pembangunan manusia diperoleh *bandwidth* optimal untuk estimator *Nadaraya-Watson* dengan fungsi kernel *gaussian* sebesar 1,384884. Hasil estimasi tersebut memperoleh nilai koefisien determinasi sebesar 63,2% dan menghasilkan nilai MAPE sebesar 2,5% yang berarti bahwa kemampuan estimasi sangat baik.

DAFTAR PUSTAKA

[1]. Hosmer, D.W, and Lemeshow, S., *Applied Logistic Regression*. New York: John Wiley & Sons. 2000.
[2]. Suparti, Santoso, R., Prahutma, A., Devi, A.R., *Regresi Nonparametrik*. Jawa Timur: Wade Group. 2017.
[3]. Kurniasih, D., Efisiensi Relatif Estimator Fungsi Kernel Gaussian Terhadap Estimator Polonomial DalamPeramalan USD Terhadap JPY. *UNNES Journal of Mathematics*, 2013. **2**(2), 79-84.
[4]. Saputra, J.A, Pemilihan Bandwidth Pada Estimator Nadaraya-Watson Dengan Tipe Kernel Gaussian Pada Data Time Series. *Jurnal Matematika*, 2016, 2-7.
[5]. Hardle, W., *Applied Nonparametric Regression*. New York: Cambridge Univercity Press. 1994.
[6]. Suliyanto., *Ekonometri Terapan: Teori & Aplikasi dengan SPSS*. Yogyakarta. ANDI. 2011.
[7]. Halimi, R., Pembuatan Aplikasi Peramalan Jumlah Permintaan Produk Dengan Metode Time Series Exponential Smoothing Holts Winter di PT. Telekomunikasi Indonesia Tbk. *Jurnal Teknik Pomits*, 2013, 1:1-6.

NURUL ANISA : JurusanMatematika FMIPA Untan, Pontianak, nrl.anisa219@gmail.com
NAOMI NESSYANA DEBATARAJA : JurusanMatematika FMIPA Untan, Pontianak, naominessyana@math.untan.ac.id
SHANTIKA MARTHA : JurusanMatematika FMIPA Untan, Pontianak, Shantika.martha@math.untan.ac.id