Regresi Nonparametrik Kernel dalam Pemodelan Jumlah Kelahiran Bayi di Jawa Barat Tahun 2017

Safni Chusnaifah Junianingsih*, Yayat Karyana

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Indonesia.

*schusnaifah@gmail.com, yayatkaryana@gmail.com

Abstract. Regression analysis is one of the analytical tools used to determine the effect of multiple predictor variables (X) on response variables (Y). The approach in regression analysis is divided into two, parametric approaches and nonparametric approaches. On nonparametric regression analysis, the shape of the regression curve is unknown, the data arega expected to look for its own estimation form so that it has high flexibility. Estimation of regression functions is performed with the Nadaraya Watson kernel estimator using Gaussian kernel functions. In this method requires bandwidth (h) or finer parameters as a balance controller between the smoothness of the function and the suitability of the function of the data. Optimum bandwidth (h) is obtained by minimizing the Generalized Cross Validation (GCV) value. Based on the analysis, obtained in a simple linear regression model obtained a Mean Square Error (MSE) value of 552976772 and a Standard Error (SE) of 24437,98. While in the kernel nonparametric regression model, the optimum bandwidth (h) is 0,50, Mean Square Error (MSE) is 96832714, and the Standard Error (SE) value is 10226,4. So it can be concluded that the kernel nonparametric regression model is better than a simple linear regression model.

Keywords: *Generalized Cross Validation (GCV), Baby Birth, Nadaraya Watson, Kernel Regression*

Abstrak. Analisis regresi merupakan salah satu alat analisis yang digunakan untuk mengetahui pengaruh dari beberapa variabel prediktor (X) terhadap variabel respon (Y). Pendekatan dalam analisis regresi dibagi menjadi dua, yaitu pendekatan parametrik dan pendekatan nonparametrik. Pada analisis regresi nonparametrik bentuk kurva regresi tidak diketahui, data diharapkan mencari sendiri bentuk estimasinya sehingga memiliki fleksibilitas yang tinggi. Estimasi fungsi regresi dilakukan dengan estimator kernel Nadaraya Watson menggunakan fungsi kernel Gaussian. Metode ini membutuhkan bandwidth (h) atau parameter penghalus sebagai pengontrol keseimbangan antara kemulusan fungsi dan kesesuaian fungsi terhadap data. Bandwidth (h) optimum diperoleh dengan meminimumkan nilai Generalized Cross Validation (GCV). Berdasarkan analisis diperoleh pada model regresi linear sederhana diperoleh nilai Mean Square Error (MSE) sebesar 552976772 dan niai Standard Error (SE) sebesar 24437,98. Sedangkan pada model regresi nonparametrik kernel diperoleh bandwidth (h) optimum sebesar 0,50, Mean Square Error (MSE) sebesar 96832714, dan nilai Standard Error (SE) sebesar 10226,4. Sehingga dapat disimpulkan bahwa model regresi nonparametrik kernel lebih baik daripada model regresi linear sederhana.

Kata Kunci: *Generalized Cross Validation (GCV), Kelahiran Bayi, Nadaraya Watson, Regresi Kernel.*

Corresponding Author
Email: yayatkaryana@gmail.com

A. Pendahuluan

Analisis regresi adalah salah satu alat analisis yang termasuk dalam statistika inferensi. Analisis regresi digunakan menelusuri pola hubungan, atau untuk mengetahui pengaruh dari beberapa variabel prediktor *(X)* terhadap variabel respon *(Y)*. Ada berbagai macam analisis regresi, di antaranya yaitu Regresi Linear Sederhana, Regresi Linear Berganda, Regresi Non-Linear, Regresi Ridge, Regresi Kernel, Regresi Logistik. Pendekatan dalam analisis regresi dibagi menjadi dua, yaitu pendekatan parametrik dan pendekatan nonparametrik.

Pendekatan parametrik memiliki bentuk hubungan yang diketahui antara variabel prediktor dengan variabel respon misalnya membentuk pola linear, kuadratik, eksponensial, dan polinomial dan juga harus memenuhi asumsi-asumsi klasik. Menurut Rifai (2019) adanya asumsi yang tidak terpenuhi menyebabkan analisis menjadi kurang tepat karena akan memberikan hasil yang tidak akurat. Sedangkan dalam regresi nonparametrik bentuk kurva regresi tidak diketahui, data diharapkan mencari sendiri bentuk estimasinya sehingga memiliki fleksibilitas yang tinggi (Sukarsa *et. al.* 2012).

Pada penelitian ini akan dilakukan pemodelan terhadap jumlah kelahiran bayi di Jawa Barat tahun 2017 menggunakan regresi nonparametrik kernel dengan estimator Nadaraya Watson dan fungsi kernel Gaussian.

B. Metodologi Penelitian

Regresi Linear Sederhana

Menurut Suyono (2018), model regresi linear sederhana adalah model probabilistik yang menyatakan hubungan linear antara dua variabel dimana salah satu variabel memengaruhi variabel yang lain. Variabel yang memengaruhi dinamakan variabel prediktor dan variabel yang dipengaruhi dinamakan variabel respon. Model probabilistik regresi linear sederhana adalah:

$$Y = \beta_O + \beta_1 X + \varepsilon \quad \ldots(2.1)$$

Dan model taksirannya adalah:

$$\hat{Y} = b_O + b_1 X \quad \ldots(2.2)$$

Estimator untuk β_O dan β_1 selanjutnya masing-masing dinotasikan dengan b_O dan b_1. Estimator ini disebut estimator kuadrat terkecil (*least squared estimator*). Rumus untuk b_1 adalah:

$$b_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} {X_i}^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \quad \ldots(2.3)$$

Sedangkan rumus untuk b_0 adalah:

$$b_0 = \frac{\sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} X_i}{n} \quad \ldots(2.4)$$

Asumsi Normalitas

Uji normalitas bertujuan untuk mengetahui apakah residual dari model regresi berdistribusi normal atau tidak. Cara mudah untuk menentukan apakah galat mengikuti distribusi normal adalah dengan menilai plot probabilitas normal (Normal PP-Plot). Jika galat mendekati garis diagonal pada grafik, maka terdistribusi secara normal. Apabila residual tidak mengikuti distribusi normal maka statistik uji tidak berlaku.

Asumsi Homogenitas

Asumsi homogenitas/homoskedastisitas menyatakan bahwa residual memiliki variansi yang konstan. Jika varians tidak homogen, maka akan menyulitkan dalam mengukur standar deviasi yang benar dari prediksi residual. Pengujian dilakukan dengan membuat *scatter plot* antara nilai prediksi pada sumbu horizontal dan residual pada sumbu vertikal.

Regresi Nonparametrik Kernel

Regresi nonparametrik kernel merupakan metode untuk memperkirakan ekspektasi bersyarat dari variabel acak dengan menggunakan fungsi kernel. Model regresi nonparametrik secara umum adalah sebagai berikut (Eubank, 1988):

$$Y_i = m(X_i) + \varepsilon_i \quad \ldots(2.5)$$

Tujuan dari regresi kernel yaitu untuk menemukan hubungan nonlinear antara sepasang

variabel acak *X* dan *Y*. Menurut Hardle *et. al*. (2004) ekspektasi bersyarat dari variabel *y* terhadap *x* yaitu:

$$E(Y|X) = m(X_i) \quad \ldots(2.6)$$

atau

$$E(Y|X = x) = \int \frac{yf(x,y)}{f(x)} dy \quad \ldots(2.7)$$

dimana:

$f(x,y)$ = fungsi densitas bersama *(X,Y)*

$f(x)$ = fungsi densitas marginal *X*

$m(X_i)$ = fungsi yang tidak diketahui untuk mendapatkan dan menggunakan bobot kernel yang sesuai

Estimator Densitas Kernel

Estimator kernel disebut juga estimator densitas kernel Rosenblatt-Parzen karena dikenalkan pertama kali oleh Parzen (1962) dan Rosenblatt (1956) (Hardle, 1990). Secara umum kernel *K* dengan *bandwidth (h)* didefinisikan sebagai berikut:

$K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right)$ untuk -∞ < u < ∞ dan h > 0 …(2.8)

dengan

$$u = \frac{x - X_i}{h}$$

Fungsi kernel *K* harus memenuhi beberapa syarat yaitu:

$$K(u) \geq 0, \text{untuk semua } x$$

$$\int_{-\infty}^{\infty} K(u)du = 1$$

$$\int_{-\infty}^{\infty} x^2 K(u)du = \sigma^2 > 0$$

$$\int_{-\infty}^{\infty} uK(u)\, dx = 0$$

Estimator densitas kernel dengan fungsi $\hat{f}_h(x)$ adalah sebagai berikut:

$\widehat{f_h}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)$ …(2.9)

$\widehat{f_h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)$ …(2.10)

Terdapat beberapa jenis fungsi kernel yang umum digunakan untuk estimasi data diantaranya yaitu:

1. Kernel Epanechnikov : $K(u) = \frac{3}{4}(1 - u^2)$;|u| ≤ 1, 0 selainnya
2. Kernel Kuartik : $K(u) = \frac{15}{16}(1 - u^2)^2$;|u| ≤ 1, 0 selainnya
3. Kernel Gaussian : $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$;$-\infty < u < \infty$

Estimator Nadaraya Watson

Estimasi fungsi regresi $m(X)$ yang belum diketahui dilakukan menggunakan estimator Nadaraya Watson. Penurunan dari estimator yang diekspresikan dalam bentuk $m(X)$ dari fungsi densitas peluang bersama $f(x,y)$ adalah sebagai berikut:

$$m(X_i) = E(Y|X) = E(Y|X = x) = \int \frac{yf(x,y)}{f(x)} dy$$

Dari persamaan di atas, estimasi Nadaraya Watson untuk m(X) yang belum diketahui nilainya dari fungsi regresi adalah:

$$\widehat{m}(X) = \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)} \quad \ldots(2.11)$$

$$\hat{m}(X) = \sum_{i=1}^{n} W_i(X) Y_i \quad \ldots(2.12)$$

dengan

$$W_i(X) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)} \quad \ldots(2.13)$$

dan

$$\sum_{i=1}^{n} W_i(X) = \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)} = 1 \quad \ldots(2.14)$$

Sehingga dapat dikatakan bahwa estimator Nadaraya Watson merupakan rata-rata terboboti dari (Y_i).

Pemilihan *Bandwidth* Optimum

Bandwidth atau parameter penghalus yang dinotasikan dengan *h*, merupakan pengontrol keseimbangan antara kemulusan fungsi dan kesesuaian fungsi terhadap data. Dalam memilih *h* optimal sangatlah penting agar estimator yang diperoleh juga optimal (Eubank, 1988).

Salah satu metode untuk mendapatkan *h* optimal adalah dengan menggunakan kriteria *Generalized Cross Validation* (GCV), yang didefinisikan sebagai berikut:

$$GCV = \frac{MSE}{\left(\frac{1}{n} tr(1-H(h))\right)^2} \quad \ldots(2.15)$$

Dengan $H(h) = X(X'X + nhI)^{-1}X'$ dan $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - m_h(x_i))^2$. Kebaikan suatu estimator dapat dilihat dari tingkat kesalahannya. Semakin kecil tingkat kesalahannya, semakin baik estimasinya. Berdasarkan penelitian Handayani (2019), Yuniarti dan Hartati (2017) ukuran ketepatan estimator:

1. Mean Square Error (MSE):
   $$MSE = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \ldots(2.16)$$
2. Standard Error (SE)
   $$SE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} \quad \ldots(2.17)$$

C. Hasil Penelitian dan Pembahasan

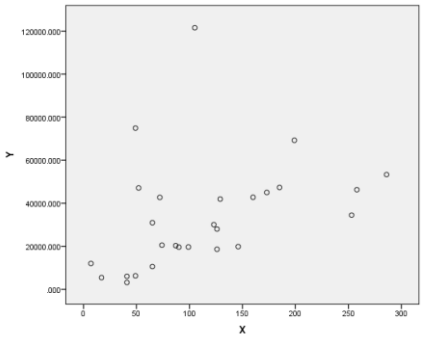
Deskripsi Data

Tabel 1. Hasil Statistik Deskriptif

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Y	27	3170.000	121566.000	33983.55556	25892.349772
X	27	7	286	113.96	74.086
Valid N (listwise)	27				

Berdasarkan Tabel 2 rata-rata jumlah kelahiran bayi di Jawa Barat sebesar 33983,55556≈33984, sedangkan rata-rata jumlah kematian bayi di Jawa Barat sebesar 113,96≈114.



Gambar 1. *Scatter* Plot X terhadap Y

Pada Gambar 1 menunjukkan bentuk kurva yang menggambarkan hubungan antara variabel prediktor terhadap variabel respon. Kurva tidak membentuk pola linear maupun nonlinear, sehingga dapat disimpulkan bahwa data variabel ini merupakan komponen nonparametrik.

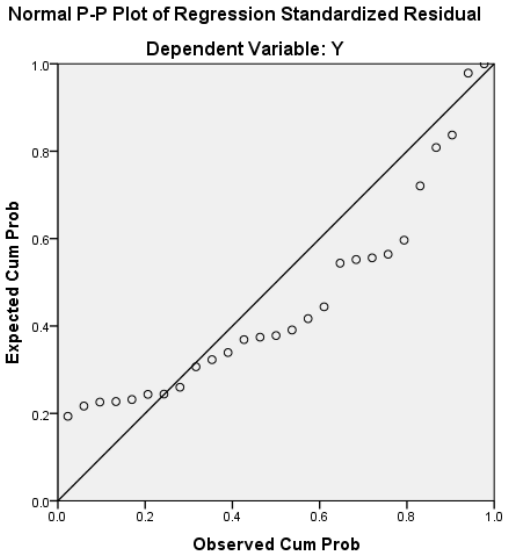
Menaksir Koefisien Regresi

Berdasarkan penaksiran parameter yang telah dilakukan diperoleh nilai taksiran $\beta_0 = 18898{,}69$ dan nilai $\beta_1 = 132{,}37$ sehingga model regresi linear sederhana yang diperoleh adalah:

$$\hat{Y} = 18898{,}69 + 132{,}37(X)$$

Pengujian Asumsi Normalitas

Hasil uji normalitas dapat dilihat dari gambar di bawah ini. Asumsi normalitas adalah sisaan yang dibentuk model regresi linear terdistribusi normal, bukan variabel respon ataupun variabel prediktor.



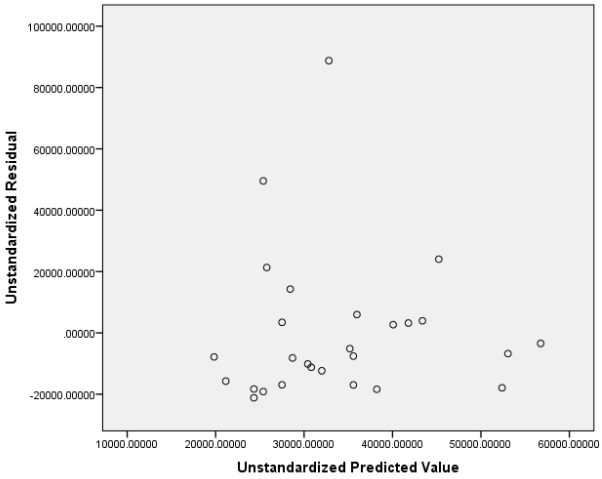
Gambar 2. PP-Plot Residual

Berdasarkan Gambar 2, terlihat bahwa sebaran titik menjauh dari garis diagonal, sehingga dapat disimpulkan bahwa residual pada model tidak berdistribusi normal.

Pengujian Asumsi Homogenitas

Pada bagian ini, pengujian homogenitas dilakukan dengan membuat *scatterplot* (alur sebaran) antara nilai prediksi dan residual. Hasil pengujian dapat dilihat pada gambar di bawah

ini.



Gambar 3. Plot Nilai Prediksi dan Residual

Berdasarkan Gambar 3 terlihat bahwa sebaran titik tidak membentuk suatu pola tertentu atau dapat dikatakan bahwa sebaran titik acak, sehingga dapat disimpulkan bahwa asumsi homogenitas terpenuhi.

Pemilihan *Bandwidth* Optimum

Pemilihan *bandwidth* (h) optimum diperoleh dengan kriteria *Generalized Cross Validation* (GCV) yang minimum. Pada penelitian ini menggunakan 5 nilai *bandwidth* yang diperoleh dengan bantuan *software R 3.6.1* dan menghasilkan nilai GCV yang disajikan pada Tabel 3.

Tabel 3. Nilai *Bandwidth* Optimum

No	*Bandwidth (h)*	GCV
1	0,50	4410461034
2	0,45	4411713738
3	0,40	4411923898
4	0,35	4411939981
5	0,30	4411940338

Dari hasil di atas, diperoleh *bandwidth* optimum dengan nilai GCV yang minimum sebesar 4410461034 adalah 0,50.

Estimasi Model Regresi Nonparametrik Kernel

Estimasi kernel Nadaraya Watson merupakan metode estimasi pengembangan dari metode histogram. Berikut disajikan rumus persamaan untuk estimasi Nadaraya Watson dalam mengestimasi $m(X)$ atau $\hat{m}(X)$.

Dengan nilai u adalah:

$$u = \frac{x - X_i}{h}, \quad i = 1,2,3, \dots, n$$

Diperoleh fungsi kernel Gaussian untuk penelitian ini adalah:

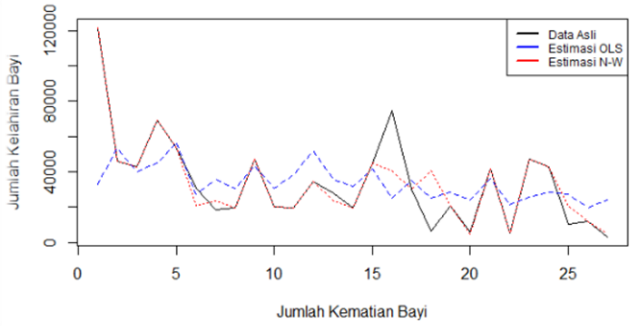
$$K_h(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-X_i}{h}\right)^2}$$

Penelitian ini terdapat 27 pengamatan dan memperoleh *bandwidth* $(h) = 0{,}50$ dari pembahasan sebelumnya pada pemilihan *bandwidth* optimum. Berikut persamaan model jumlah kelahiran bayi dengan pendekatan regresi nonparametrik kernel Nadaraya Watson dengan pemilihan *bandwidth* optimal GCV adalah:

$$Y_i = \frac{\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-X}{0{,}50}\right)^2} Y_i}{\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-X}{0{,}50}\right)^2}} + \varepsilon_i$$

Hasil Estimasi dan Perbandingan Metode Analisis

Dengan hasil estimasi yang diperoleh, berikut disajikan kurva hasil estimasi jumlah kelahiran bayi di Jawa Barat.



Gambar 4. Perbandingan Kurva Hasil Estimasi

Berdasarkan Gambar 4 dapat dilihat bahwa estimasi dengan metode regresi nonparametrik pada kurva berwarna merah menghasilkan estimasi kurva regresi yang lebih baik karena mendekati kurva data yang sebenarnya. Adapun perbandingan ukuran ketepatan estimator disajikan pada Tabel 4.

Tabel 4. Hasil Perbandingan Mean Square Error (MSE) dan Standard Error (SE)

No	Metode	MSE	SE
1	Regresi Linear Sederhana dengan metode OLS	552976772	24437,98
2	Regresi Nonparametrik Kernel dengan estimator Nadaraya Watson	96832714	10226,4

Berdasarkan Tabel 4 dapat dilihat bahwa model regresi nonparametrik kernel Nadaraya Watson memiliki nilai MSE dan SE yang lebih kecil daripada model regresi linear sederhana. Sehingga dapat disimpulkan bahwa model yang terbaik untuk mengestimasi jumlah kelahiran bayi di Jawa Barat adalah menggunakan model regresi nonparametrik kernel.

D. Kesimpulan

Berdasarkan hasil penelitian diperoleh model regresi nonparametrik kernel pada kasus jumlah kelahiran bayi di Jawa Barat sebagai berikut:

$$Y_i = \frac{\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-X}{0,50}\right)^2} Y_i}{\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-X}{0,50}\right)^2}} + \varepsilon_i$$

Apabila dilihat dari ukuran ketepatan estimator, nilai *Mean Square Error* (MSE) dan nilai *Standard Error* (SE) pada model regresi nonparametrik kernel lebih baik untuk data jumlah kelahiran bayi daripada model regresi linear sederhana karena memiliki nilai MSE dan SE yang lebih kecil.

Daftar Pustaka

[1] Dinas Kesehatan Provinsi Jawa Barat. (2017). *Profil Kesehatan Jawa Barat 2017.*

[2] Eubank. (1988). *Spline Smoothing and Nonparametric Regression.* New York: Marcel Dekker.

[3] Handayani, R. (2019). *Analisis Regresi Kernel dengan Estimator Nadaraya Watson.* Yogyakarta.

[4] Hardle, W. (1990). *Applied Nonparametric Regression.* New York.

[5] Hardle, W., Muller, M., Sperlich, S., & Werwatz, A. (2004). Nonparametric and Semiparametric Models. *Springer*.

[6] Rifai, N. K. (2019). Pendekatan Regresi Nonparametrik Dengan Fungsi Kernel Untuk Indeks Harga Saham Gabungan. *Statistika*, 53-61.

[7] Suyono. (2018). *Analisis Regresi untuk Penelitian.* Yogyakarta: Deepublish.

[8] Yuniarti, R., & Hartati, W. (2017). Regresi Nonparametrik Menggunakan Metode Robust dan Cross Validation. *UJMC*, 9-16.