Jurnal Matematika Thales (JMT): 2021 ISSN: 2715-1891
Vol. 03 No. 01

ESTIMASI MODEL REGRESI SEMIPARAMETRIK SPLINEPADA DATA DENGAN OUTLIER MENGGUNAKANMETODE ESTIMASI M ROBUST (SPLINE SEMIPARAMETRIC REGRESSION MODELESTIMATION ON DATA WITH OUTLIERS USING ROBUST M-ESTIMATION METHOD)

PUTRI NILAM CAYO*, SRI HARYATMI

Abstract. Spline semiparametric regression is a regression model that combines parametric components and nonparametric components in one model where the nonparametric components are approximated by spline regression. The estimation method that is generally used to estimate spline semiparametric regression model is least square method. However, the estimation constructed by this method is sensitive to outliers, causing the estimation of parameter values to be biased and the interpretation of results to be inaccurate. In overcoming this problem, the outliers cannot be eliminated because the outliers can contain important information that cannot be provided by other observations. Therefore, we need an estimation method that is resistant to outliers. It is called robust method. The robust method used in this study is M-estimation method. M-estimation method estimates parameters by minimizing the objective function of the residual. The result shows that M-estimation method produces a smaller GCV (Generalized Cross Validation) value than least square method's. Thus, the parameter estimators generated by M-estimation method are better than least square method's.
Keywords: M-Estimation, *Outlier*, *Robust*, Spline Semiparametric Regression

Abstrak. Regresi semiparametrik *spline* merupakan model regresi yang mengombinasikan komponen parametrik dan komponen nonparametrik dalam satu model dimana komponen nonparametriknya didekatkan dengan regresi *spline*.Metode estimasi yang umumnya digunakan untuk mengestimasi model regresi semiparametrik *spline* adalah metode kuadrat terkecil (l*east square*). Namun estimasi yang dikonstruksikan dengan metode tersebut sensitif terhadap *outlier* sehingga menyebabkan estimasi nilai parameter menjadi bias dan interpretasi hasil menjadi tidak akurat. Dalam mengatasi hal tersebut, *outlier* tidak dapat dihilangkan begitu saja karena *outlier* dapat mengandung informasi penting yang tidak dapat diberikan oleh pengamatan lain. Oleh karena itu, dibutuhkan suatu metode estimasi yang kokoh terhadap outlier, yaitu metode *robust*. Metode robust yang digunakan dalam penelitian ini adalah metode estimasi M. Metode estimasi M mengestimasi parameter dengan cara meminimumkan fungsi objektif dari residual. Hasil penelitian menunjukkan bahwa metode estimasi M menghasilkan nilai GCV (*Generalized Cross Validation*) yang lebih kecil dibandingkan nilai GCV yang diperoleh dari metode kuadrat terkecil. Dengan demikian, estimator parameter yang dihasilkan oleh metode estimasi M lebih baik dibandingkan metode kuadrat terkecil.
Kata-kata kunci: Estimasi M, *Outlier*, Regresi Semiparametrik *Spline*, *Robust*

1. PENDAHULUAN

Analisis regresi merupakan sebuah metode dalam statistika yang digunakan untuk mengetahui pola hubungan antara satu atau lebih variabel prediktor (*independent variable*) dengan satu atau lebih variabel respon (*dependent variable*).Hubungan fungsional antara variabel prediktor dan variabel respon dapat dijelaskan dalam sebuah kurva yang dinamakan kurva regresi. Pada umumnya, kurva regresi didekatkan dengan regresi parametrik atau regresi nonparametrik bergantung pada bentuk polanya. Regresi parametrik mengasumsikan bentuk kurva diketahui berdasarkan teori, informasi sebelumnya, atau sumber-sumber lain yang dapat memberikan pengetahuan secara terperinci. Sementara regresi nonparametrik digunakan untuk pendekatan model yang tidak diketahui bentuk pola kurva regresinya, hanya diasumsikan termuat dalam suatu ruang fungsi tertentu, dimana pemilihan ruang fungsi tersebut biasanya dimotivasi oleh sifat kemulusan (*smoothness*) yang dimiliki oleh fungsi. Dalam prakteknya, terdapat data yang sebagian diketahui bentuk pola kurvanya dan sebagian lagi tidak. Untuk mengatasi kondisi tersebut, dibuatlah regresi semiparametrik yang merupakan kombinasi antara regresi parametrik dan regresi nonparametrik.

Ada beberapa teknik estimasi dalam regresi nonparametrik, diantaranya pendekatan histogram, estimator *spline*, estimator *kernel*, estimator deret *orthogonal*, estimator deret *Fourier*, dan analisis wavelet. Menurut Budiantara, et al [1], di antara teknik-teknik estimasi tersebut, estimator *spline* merupakan teknik yang mempunyai interpretasi statistik dan visual yang sangat khusus dan sangat baik. Spline merupakan potongan polinomial tersegmen (*piecewise polynomial*) yang dihubungkan oleh titik-titik *knot*. Titik *knot* merupakan titik perpaduan bersama yang menjelaskan terjadinya perubahan perilaku dari fungsi *spline* pada interval-interval yang berbeda [3].

Estimasi parameter model regresi *spline*, baik itu nonparametrik maupun semiparametrik, umumnya diselesaikan dengan metode kuadrat terkecil (*Least Square*). Prinsip dari metode ini adalah meminimumkan kuadrat residual (error). Penggunaan metode ini mengasumsikan bentuk fungsi *spline* berupa polinomial *truncated* dan memberikan kemudahan interpretasi melalui model statistik. Estimator yang dihasilkan dari metode kuadrat terkecil akan bersifat tak bias dan efisien (*Best Linear Unbiased Estimator*/BLUE) jika komponen residual atau error memenuhi beberapa asumsi klasik, yaitu kenormalan, kehomogenan ragam, dan tidak terjadi autokorelasi [8]. Namun menurut Gao dan Shi [4], estimasi yang dikonstruksikan dengan menggunakan metode kuadrat terkecil sensitif terhadap *outlier* dan distribusi errornya memiliki varian yang tidak terbatas. Untuk mengatasi hal tersebut, diperlukan suatu metode estimasi yang kokoh (*robust*) terhadap *outlier*, yaitu metode regresi robust. Ada beberapa metode regresi *robust* yang dapat digunakan untuk mengatasi outlier, salah satunya yaitu metode estimasi M. Pada prinsipnya, metode estimasi M dilakukan dengan meminimumkan fungsi objektif dari residual (error).

Penelitian sebelumnya mengenai estimasi model regresi *spline* yang mengandung *outlier* telah dilakukan oleh Gao dan Shi [4] yang menggunakan metode estimasi M untuk mengestimasi regresi nonparametrik dan semiparametrik *spline* pada data simulasi dengan menggunakan fungsi B spline. Selain itu, Lee dan Oh [7] dalam penelitiannya mendefinisikan estimasi M untuk model regresi *spline* terpenalti dengan mengganti

metode kuadrat terkecil terpenalti (*penalized least square*) dengan metode estimasi M yang bersesuaian dengan tetap menjaga bentuk dari spline dan parameter penaltinya.

Dalam penelitian ini, akan dibahas tentang pengestimasian model regresi semiparametrik spline yang mengandung *outlier* dengan menggunakan metode estimasi M *robust* dan fungsi *spline* yang digunakan yaitu fungsi *truncated spline*. Tujuan dari penelitian ini yaitu menentukan estimator parameter model regresi semiparametrik spline menggunakan metode estimasi M robust dan membandingkan hasilnya dengan estimator yang diperoleh dengan hanya menggunakan metode kuadrat terkecil (*least square*).

2. TINJAUAN PUSTAKA

2.1. **Regresi Semiparametrik Spline.** Regresi semiparametrik merupakan suatu metode yang mengkombinasikan model regresi parametrik dan model regresi nonparametrik. Secara umum, model regresi semiparametrik dapat ditulis dalam bentuk sebagai berikut

$$y_i = \mathbf{X}_i^T \boldsymbol{\beta} + m(t_i) + \varepsilon_i; i = 1, 2, \ldots, n \tag{2.1}$$

dengan $\mathbf{X}_i^T = (1, x_{i1}, x_{i2}, \ldots, x_{it})$ merupakan komponen parametrik ke$-i$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_r)^T$ merupakan vektor parameter model parametrik yang tidak diketahui, $m(t_i)$ adalah fungsi regresi ke$-i$ yang tidak diketahui (sebagai komponen nonparametrik), dan ε_i adalah error ke$-i$ dimana $\varepsilon_i \sim (0, \sigma^2)$.

Model semiparametrik dalam beberapa literatur juga dikenal sebagai model liner parsial (*partially linear model*). Model ini lebih fleksibel daripada model linear karena keberadaan komponen parametrik dan nonparametrik akan mengakomodasi hubungan antara respon dengan prediktor yang bersifat linear, dan hubungan antara respon dan prediktor yang bersifat nonlinear.

Salah satu teknik estimasi yang dapat digunakan untuk mengestimasi regresi nonparametrik yang tidak diketahui bentuk kurvanya yaitu estimator spline. Eubank [3] menjelaskan *spline* merupakan potongan polinomial tersegmen (*piecewise polynomial*) yang dihubungkan oleh titik-titik *knot*. Titik *knot* merupakan titik perpaduan bersama yang menjelaskan terjadinya perubahan perilaku dari fungsi *spline* pada interval-interval yang berbeda. Jadi, kurva fungsi *spline* yang dibentuk tersegmen pada titik-titik tersebut.

Secara umum, fungsi *spline* berorde $k-1$ dengan titik knot $K1, K2, \ldots, K_p$ dapat dinyatakan dengan persamaan berikut

$$m(t) = \sum_{j=0}^{k-1} \alpha_j t^j + \sum_{l=1}^{p} \alpha_{(k-1)+l} (t - K_l)_+^{k-1} \tag{2.2}$$

dengan

$$(t - K_l)_+^{k-1} = \begin{cases} (t - K_l)^{k-1}, & \text{jika } t \geq K_l \\ 0, & \text{jika } t < K_l \end{cases}$$

merupakan fungsi potongan (*truncated*), t adalah variabel prediktor, $\alpha j, j = 0, 1, \ldots, k-$ 1dan $\alpha(k-1) + l, l = 1, 2, \ldots, p$ menyatakan konstanta bernilai real, serta $a < K1 <$

$K2 << Kp < b$ dengan a adalah nilai minimum dari variabel t dan b adalah nilai maksimum dari variabel t. Jika nilai $k = 2, 3$, dan 4 disubstitusikan ke persamaan 2.2, maka diperoleh fungsi *spline* secara berturut-turut dinamakan fungsi *spline* linear, *spline* kuadratik, dan *spline* kubik [3].

Apabila fungsi *spline* pada persamaan 2.2 disubstitusikan ke model regresi semiparametrik pada persamaan 2.1, maka akan diperoleh model sebagai berikut

$$y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \sum_{j=0}^{k-1} \alpha_j t_i^j + \sum_{l=1}^{p} \alpha_{(k-1)+l} (t_i - K_l)_+^{k-1} + \varepsilon_i \tag{2.3}$$

Model tersebut dikenal sebagai model regresi semiparametrik spline. Persamaan 2.3 dapat disajikan lebih sederhana dalam bentuk matriks berikut

$$\mathbf{Y} = \mathbf{C}\boldsymbol{\omega} + \boldsymbol{\varepsilon} \tag{2.4}$$

dengan $\mathbf{C}$ merupakan matriks yang berisi variabel-variabel prediktor dari komponen parametrik dan komponen nonparametrik, termasuk variabel pada fungsi *truncated* sedangkan ω adalah matriks yang terdiri dari parameter-parameter yang belum diketahui nilainya dan akan diestimasi, serta ε merupakan matriks yang berisi residual dari tiap pengamatan. Berdasarkan model regresi pada persamaan 2.4, diperoleh estimator parameter ω dengan menggunakan metode kuadrat terkecil (*least square*), yaitu $\hat{\boldsymbol{\omega}} = (\boldsymbol{C^T C})^{-1} \boldsymbol{C^T Y}$.

2.2. **Pemilihan Model Regresi Spline Terbaik.** Pemilihan model regresi spline terbaik merupakan hal yang sangat penting. Eubank [3] menyatakan bahwa pemilihan model regresi spline terbaik diperoleh dari titik knot yang optimal. Titik knot yang optimal dilihat berdasarkan nilai *Generalized Cross Validation* (GCV) paling minimum yang diperoleh dari setiap model. Adapun fungsi untuk menghitung GCV didefinisikan sebagai berikut

$$GCV(K) = \frac{MSE(K)}{\left[n^{-1} \operatorname{tr}(\boldsymbol{I} - \boldsymbol{H})\right]^2} \tag{2.5}$$

dengan $\mathrm{MSE}(K) = n^{-1} \sum_{j=1}^{n} (y_i - \hat{y}_i)^2$, K merepresentasikan titik-titik knot $(K_1, K_2, \ldots, K_p)$, n menyatakan banyaknya data amatan, $\boldsymbol{I}$ merupakan matriks identitas, dan $\boldsymbol{H}$ adalah matriks *hat*.

2.3. **Outlier.** Rousseeuw dan Leroy [10] mendefinisikan outlier sebagai observasi yang menyimpang jauh dari pola yang terbentuk pada keseluruhan data. Outlier muncul disebabkan oleh berbagai kemungkinan, diantaranya kesalahan prosedur dalam memasukan data atau pengkodean, muncul dalam range nilai yang ada tetapi bila dikombinasi dengan variabel lain menjadi ekstrim (disebut multivariate outliers), dan juga outlier terjadi dari proses luar biasa yang menunjukkan keunikan dari observasi.

Dalam statistika ruang, outlier harus dilihat tehadap posisi dan sebaran data yang lainnya sehingga dapat ditentukan apakah outlier tersebut perlu dihilangkan atau tidak. Namun sebelumnya, perlu dilakukan identifikasi terlebih dahulu untuk mengetahui ada tidaknya outlier pada data. Soemartini [11] menjelaskan terdapat beberapa metode

yang dapat digunakan untuk mengidentifikasi outlier pada data, salah satunya yaitu metode Residual Studentized.

Pendeteksian outlier menggunakan metode ini dilakukan dengan melakukan perhitungan yang melibatkan residual dan variansi. Residual Studentized didefinisikan sebagai berikut

$$t_i = \frac{\epsilon_i}{\sqrt{\hat{\sigma}_i^2 \left(1 - h_{ii}\right)}} \tag{2.6}$$

dimana $\hat{\sigma}_i^2$ merupakan variansi yang dihitung tanpa mengikutsertakan pengamatan ke-i, dengan nilai

$$\hat{\sigma}_i^2 = \frac{(n-p)\hat{\sigma}^2}{n-p-1} - \frac{\epsilon_i^2}{(n-p-1)\left(1-h_{ii}\right)}; \text{ dan}$$
$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \epsilon_i^2}{n-p}$$

serta h_{ii} merupakan nilai pengaruh (*leverage value*) yang diperoleh dari elemen-elemen diagonal pada matriks *hat* $\boldsymbol{H}$ dengan $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$ dan $\boldsymbol{X}$ merupakan matriks yang berisi nilai-nilai variabel prediktor dalam pengamatan. Nilai h_{ii} berada di antara 0 dan 1 dengan $\sum_{i=1}^{n} h_{ii} = p$ dimana p merupakan banyaknya parameter pada persamaan regresi yang terbentuk termasuk *intercept* dan n adalah banyaknya data yang diamati. Suatu pengamatan ke- i diidentifikasi sebagai *outlier* jika memiliki nilai $|t_i| > t_{\alpha/2,n-p-1}$.

2.4. **Regresi Robust.** Rousseeuw dan Leroy [10] menyatakan bahwa regresi robust merupakan suatu metode estimasi yang robust (kokoh) terhadap kemunculan outlier. Sementara menurut Draper dan Smith [2], regresi robust merupakan metode regresi yang digunakan ketika distribusi dari sisaan tidak berdistribusi normal. Biasanya untuk membuat distribusi dari sisaan berdistribusi normal, digunakan teknik transformasi. Namun seringkali transformasi tidak dapat menghilangkan atau melemahkan pengaruh outlier sehingga pada akhirnya estimasi menjadi bias. Maka dari itu, sangat tepat jika menggunakan metode regresi robust yang tahan terhadap pengaruh outlier sehingga diperoleh hasil yang lebih baik.

Ukuran ke-robust-an suatu metode regresi robust dapat dilihat dari breakdown point dan efisiensinya. Menurut Hubert dan Debruyne [6], breakdown point adalah fraksi terkecil dari data atau presentase dari outlier yang dapat menyebabkan nilai estimator menjadi tak berhingga ketika semua data pada fraksi tersebut diubah menjadi tak berhingga. Kemungkinan tertinggi breakdown point untuk sebuah estimator adalah 0,5 . Jika breakdown point lebih dari 0,5 berarti estimasi model regresi tidak dapat menggambarkan informasi dari mayoritas data. Sementara itu, efisiensi menjelaskan seberapa baiknya suatu metode regresi robust dibandingkan dengan metode estimasi kuadrat terkecil tanpa outlier. Semakin tinggi efisiensi dan breakdown point suatu metode regresi robust maka semakin robust (resisten) pula metode tersebut terhadap outlier.

2.5. **Estimasi M.** Estimasi M merupakan salah satu metode estimasi dalam regresi robust yang pertama kali diperkenalkan oleh Huber pada tahun 1973. Huruf "M" menunjukkan bahwa estimasi M merupakan estimasi yang berdasarkan "tipe maksimum likelihood". Pradewi dan Sudarno [9] menjelaskan pada prinsipnya estimasi M merupakan estimasi yang meminimumkan fungsi objektif ρ dari residual (ϵ_i). Secara matematis dituliskan sebagai berikut

$$\min_{\hat{\omega}} \sum_{i=1}^{n} \rho\left(\epsilon_i\right) = \min_{\hat{\omega}} \sum_{i=1}^{n} \rho\left(y_i - \hat{y}_i\right)$$

dengan $\hat{\omega} = \left(\beta_0, \beta_1, \alpha_1, \ldots, \alpha_{k-1}, \alpha_{(k-1)+1}, \ldots, \alpha_{(k-1)+p}\right)^T$. Untuk mempermudah pengestimasian, digunakan suatu skala invariant dari estimator $\hat{\sigma}$ yang akan diestimasi menggunakan standarisasi residual seperti berikut

$$\min_{\hat{\omega}} \sum_{i=1}^{n} \rho\left(u_i\right) = \min_{\hat{\omega}} \sum_{i=1}^{n} \rho\left(\frac{\epsilon_i}{\hat{\sigma}}\right) = \min_{\hat{\omega}} \sum_{i=1}^{n} \rho\left(\frac{y_i - \hat{y}_i}{\hat{\sigma}}\right) \tag{2.7}$$

dimana $\rho\left(u_i\right)$ adalah fungsi simetris dari residual atau fungsi yang memberikan konstribusi pada masing-masing residual pada fungsi objektif dan $\hat{\sigma}$ diperoleh dari nilai s selaku skala estimasi robust yang dipilih dengan cara

$$s = \frac{\text{median}\left\{\left|\epsilon_i - \text{median}\left(\epsilon_i\right)\right|\right\}}{0,6745}.$$

Pemilihan konstanta 0,6745 membuat sedemikian hingga s merupakan suatu estimator yang mendekati tak bias dari σ jika n besar dan error berdistribusi normal.

Meminimumkan persamaan 2.7 dilakukan dengan mencari turunan parsial pertama dari ρ terhadap $\boldsymbol{\omega}$, kemudian hasilnya disamadengankan dengan 0 sebagai berikut

$$\begin{aligned} \frac{\partial \sum_{i=1}^{n} \rho\left(\frac{y_i - \hat{y}_i}{\hat{\sigma}}\right)}{\partial \boldsymbol{\omega}} &= 0 \\ \sum_{i=1}^{n} \hat{y}_i' \psi\left(\frac{y_i - \hat{y}_i}{\hat{\sigma}}\right) &= 0 \end{aligned} \tag{2.8}$$

Fungsi turunan parsial pertama ρ dinamakan fungsi pengaruh (influence) dan dinotasikan dengan ψ. Selanjutnya dari fungsi influence, dapat dicari fungsi pembobot, yaitu

$$w\left(u_i\right) = \frac{\psi\left(\frac{\epsilon_i}{\hat{\sigma}}\right)}{\left(\frac{\epsilon_i}{\hat{\sigma}}\right)} = \frac{\psi\left(\frac{y_i - \hat{y}_i}{\hat{\sigma}}\right)}{\left(\frac{y_i - \hat{y}_i}{\hat{\sigma}}\right)}$$

Dengan memisalkan $w_i = w\left(u_i\right)$, maka maka persamaan 2.8 dapat ditulis sebagai

$$\sum_{i=1}^{n} \hat{y}_i' w_i \left(\frac{y_i - \hat{y}_i}{\hat{\sigma}} \right) = 0 \tag{2.9}$$

dengan

$$w_i = \begin{cases} \frac{\psi\left(\frac{y_i - \hat{y}_i}{\hat{\sigma}}\right)}{\left(\frac{y_i - \hat{y}_i}{\hat{\sigma}}\right)} & , \text{ jika } y_i \neq \hat{y}_i \\ 1 & , \text{ jika } y_i = \hat{y}_i \end{cases} \tag{2.10}$$

Persamaan 2.9 dapat disajikan lebih sederhana dalam bentuk matriks seperti berikut

$$\boldsymbol{C}^T \boldsymbol{W} \boldsymbol{C} \boldsymbol{\omega} = \boldsymbol{C}^T \boldsymbol{W} \boldsymbol{Y} \tag{2.11}$$

dimana $\boldsymbol{W}$ merupakan matriks diagonal berukuran $n \times n$ dari bobot dengan elemen-elemen diagonal $w_1, w_2, \cdots, w_n$ diberikan oleh persamaan 2.10. Persamaan 2.11 dapat digunakan sebagai alat untuk mendapatkan estimasi M sehingga estimasi parameter menjadi

$$\hat{\boldsymbol{\omega}} = \left(\boldsymbol{C}^T \boldsymbol{W} \boldsymbol{C} \right)^{-1} \boldsymbol{C}^T \boldsymbol{W} \boldsymbol{Y} \tag{2.12}$$

Fungsi pembobot pada estimasi M yang sering digunakan adalah fungsi pembobot Huber. Fungsi pembobot Huber didefinisikan sebagai berikut

$$w\left(u_i\right) = \begin{cases} 1 & , |u_i| \leq c \\ \frac{c}{|u_i|} & , |u_i| > c \end{cases}$$

dengan c adalah tuning constant. Tuning constant dalam regresi robust menentukan ke-robust-an penduga terhadap outlier dan efisiensi penduga dalam ketiadaan outlier. Huber dan Ronchetti [5] menjelaskan jika diambil $\alpha = 5\%$, maka estimasi M Huber akan efektif digunakan dengan nilai $c = 1,345$. Dengan demikian, fungsi pembobot Huber menjadi

$$w\left(u_i\right) = \begin{cases} 1 & , |u_i| \leq 1,345 \\ \frac{1,345}{|u_i|} & , |u_i| > 1,345 \end{cases} . \tag{2.13}$$

3. METODE PENELITIAN

3.1. **Sumber Data.** Data yang digunakan pada penelitian ini adalah data sekunder yang diperoleh dari publikasi Badan Pusat Statistika (BPS) Provinsi Jawa Timur, yakni data produksi padi di Jawa Timur Tahun 2017. Unit observasi yang digunakan dalam penelitian ini yaitu 38 kabupaten/kota di Jawa Timur. Variabel penelitian yang digunakan terdiri dari satu variabel respon dan dua variabel prediktor. Adapun variabel respon yang digunakan yaitu jumlah produksi padi menurut kabupaten/kota di Jawa Timur sedangkan variabel-variabel prediktor yang diduga mempengaruhi produksi padi

tersebut yaitu luas panen padi dan produktivitas padi menurut kabupaten/kota di Jawa Timur.

3.2. **Metode Analisis.** Adapun langkah-langkah yang dilakukan dalam metode analisis penelitian ini adalah sebagai berikut

(1) Menetapkan komponen parametrik dan komponen nonparametrik berdasarkan bentuk pola yang tergambar pada scatterplot dari masing-masing variabel prediktor dengan variabel respon,
(2) Memilih model regresi semiparametrik spline terbaik berdasarkan titik knot optimal yang diperoleh dari nilai GCV paling minimum,
(3) Mengidentifikasi adanya outlier pada data dengan menggunakan metode Residual Studentized,
(4) Melakukan estimasi M dengan langkah-langkah sebagai berikut:
    (a) Mengestimasi parameter regresi menggunakan metode kuadrat terkecil (least square) sehingga diperoleh $\hat{y}_i^0$ dan residual ϵ_i^0, dengan $\epsilon_i^0 = y_i - \hat{y}_i^0$, yang diperlakukan sebagai nilai awal;
    (b) Dari nilai-nilai residual tersebut, dihitung $\hat{\sigma}^0$ dan fungsi pembobot awal yaitu w_i^0 yang dihitung berdasarkan fungsi pembobot Huber pada persamaan 2.13 ;
    (c) Mencari estimasi pada iterasi h ($h = 1, 2, \ldots$) dengan menggunakan persamaan $\hat{\boldsymbol{\omega}}^h = \left(\boldsymbol{C}^T\boldsymbol{W}^{h-1}\boldsymbol{C}\right)^{-1}\boldsymbol{C}^T\boldsymbol{W}^{h-1}\boldsymbol{Y}$ dimana $\boldsymbol{W}^{h-1}$ merupakan matriks diagonal dengan elemen-elemen diagonalnya adalah w_i^{h-1} sehingga estimasi parameter pada iterasi pertama ($h = 1$) menggunakan ϵ_i^0 dan w_i^0;
    (d) Menghitung $\sum_{i=1}^n \left|\epsilon_i^1\right|$ atau $\sum_{i=1}^n \left|y_i - \hat{y}_i^1\right|$; dan
    (e) Mengulang langkah 2 sampai langkah 4 hingga diperoleh $\sum_{i=1}^n \left|\epsilon_i^h\right|$ yang konvergen, yakni selisih antara $\hat{\boldsymbol{\omega}}^{\boldsymbol{h+1}}$ dan $\hat{\boldsymbol{\omega}}^{\boldsymbol{h}}$ sama dengan 0 atau mendekati 0 ,
(5) Membandingkan metode kuadrat terkecil dan metode estimasi M robust berdasarkan nilai GCV dari masing-masing metode,
(6) Membuat kesimpulan dari hasil analisis.

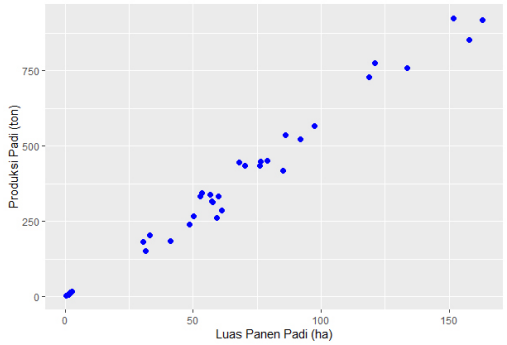
4. HASIL DAN PEMBAHASAN

4.1. **Deskripsi Data.** Berikut ini hasil deskripsi data yang menjelaskan mean, variansi, nilai minimum, dan nilai maksimum dari variabel-variabel produksi padi, luas panen padi, dan produktivitas padi.

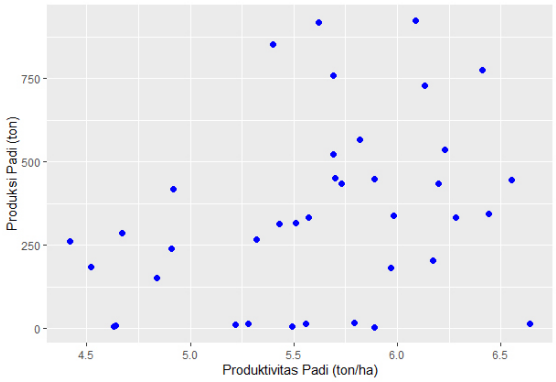
Variabel	Mean	Variansi	Minimum	Maksimum
Produksi Padi	343.696	74274.194	2.653	924.933
Luas Panen Padi	60.138	2153.585	0.451	163.093
Produktivitas Padi	5.612	0.357	4.420	6.640

TABEL 1. Deskripsi Data Penelitian

4.2. **Penentuan Komponen Parametrik dan Nonparametrik.** Berikut merupakan scatterplot yang menyatakan hubungan antara produksi padi dengan luas panen padi dan produktivitas padi.



GAMBAR 1. *Scatterplot* Produksi Padi dengan Luas Panen Padi



GAMBAR 2. *Scatterplot* Produksi Padi dengan Produktivitas Padi

Berdasarkan Gambar 1, diketahui bahwa hubungan antara produksi padi dengan luas panen padi membentuk pola yang linear. Dengan demikian, luas panen padi dapat dikategorikan sebagai komponen parametrik dan dinotasikan dengan x. Sementara dari Gambar 2 terlihat bahwa hubungan antara produksi padi dengan produktivitas padi tidak membentuk pola tertentu. Dengan demikian, produktivitas padi dapat dikategorikan sebagai komponen nonparametrik dan dinotasikan dengan t.

4.3. **Pemilihan Model Terbaik.** Model terbaik dipilih berdasarkan model yang memiliki titik knot optimal. Titik tersebut didapat dari model dengan nilai GCV paling minimum. Berikut nilai GCV dari model regresi semiparametrik spline berode 1, orde 2, dan orde 3 dengan masing-masing orde memiliki 1 titik knot, 2 titik knot, dan 3 titik knot.

Banyaknya Titik Knot	Orde	GCV
1 Titik Knot (6.6)	1	420.0855
2 Titik Knot (4.66, 4.67)	1	369.5019
3 Titik Knot (**4.64**, **4.67**, **6.6**)	**1**	**275.4717**
1 Titik Knot (6.49)	2	389.5613
2 Titik Knot (4.73, 6.47)	2	368.6773
3 Titik Knot (4.86, 4.95, 6.58)	2	345.2857
1 Titik Knot (6.34)	3	394.9961
2 Titik Knot (6.36, 6.37)	3	392.4503
3 Titik Knot (4.6, 4.72, 6.42)	3	355.2834

TABEL 2. Nilai GCV Minimum pada Tiap Model

Dari Tabel 2, terlihat bahwa nilai GCV yang paling minimum yaitu sebesar 275.4717. Nilai ini diperoleh dari model regresi berorde 1 dengan 3 titik knot optimal, yaitu $K_1 = 4.64, K_2 = 4.67$, dan $K_3 = 6.6$. Dengan demikian, model regresi semiparametrik spline terbaik adalah sebagai berikut

$$y_i = \beta_0 + \beta_1 x_i + \alpha_1 t_i + \alpha_2 (t_i - K_1)_+ + \alpha_3 (t_i - K_2)_+ + \alpha_4 (t_i - K_3)_+ + \epsilon_i.$$

Parameter-parameter pada model regresi tersebut $(\beta_0, \beta_1, \alpha_1, \alpha_2, \alpha_3, \alpha_4)$ belum diketahui nilainya. Maka dari itu, perlu dilakukan estimasi parameter dengan menggunakan metode kuadrat terkecil (least square). Pengestimasian dilakukan dengan berdasarkan pada persamaan 2.12. Hasil estimasi parameter model dengan menggunakan metode Least Square dapat dilihat pada Tabel 3.

Parameter	Estimasi
$\hat{\beta}_0$	-1616.031550
$\hat{\beta}_1$	5.684927
$\hat{\alpha_1}$	348.044151
$\hat{\alpha_2}$	-2311.741403
$\hat{\alpha_3}$	2027.114158
$\hat{\alpha_4}$	-1573.797732

TABEL 3. Hasil Estimasi Parameter Menggunakan *Least Square*

Dari hasil estimasi parameter dan titik knot optimal yang diperoleh, maka model yang terbentuk adalah sebagai berikut

$$\begin{aligned}\hat{y}_i = &- 1616.031550 + 5.684927x_i + 348.044151t_i - 2311.741403\,(t_i - 4.64)_+ \\ &+ 2027.114158\,(t_i - 4.67)_+ - 1573.797732\,(t_i - 6.6)_+ .\end{aligned}$$

4.4. **Identifikasi Outlier.** Mengidentifikasi outlier menggunakan residual studentized dilakukan dengan menghitung nilai t_i berdasarkan pada persamaan 2.6. Suatu pengamatan ke- i diidentifikasi sebagai outlier jika memiliki nilai $|t_i| > t_{\alpha/2,n-p-1}$ dengan n adalah banyaknya data yang diamati dan p merupakan banyaknya parameter pada persamaan regresi yang terbentuk termasuk intercept. Dalam penelitian ini, data yang diamati ada sebanyak 38 kabupaten/kota di Jawa Timur. Sementara berdasarkan Tabel 3, diketahui bahwa parameter yang diestimasi ada sebanyak 6 parameter. Dengan menggunakan taraf signifikan $\alpha = 0.05$, maka batasan nilai untuk $t_{0.025,31}$ berdasarkan tabel-t adalah 2.0395 . Nilai inilah yang akan dibandingkan dengan hasil hitung t dari masing-masing pengamatan untuk menentukan apakah pengamatan tersebut merupakan outlier atau tidak. Hasil identifikasi outlier menggunakan residual studentized disajikan dalam Tabel 4.

Pengamatan ke-	Nilai $\boldsymbol{t_i}$
10	2.8969
22	-2.4955
24	2.4011

TABEL 4. Hasil Identifikasi Outlier dengan Residual Studentized

Berdasarkan tabel 4, diketahui bahwa pengamatan ke-10, 22, dan 24 merupakan outlier. Hal ini dikarenakan nilai $|t_i|$ yang dihasilkan dari masing-masing pengamatan lebih besar dari $t_{0.025,31} = 2.0395$.

4.5. **Estimasi Parameter Regresi Menggunakan Metode Estimasi M.** Berdasarkan hasil identifikasi outlier pada Tabel 4, dapat disimpulkan bahwa terdapat beberapa outlier pada data. Hal ini mengakibatkan estimasi parameter yang diperoleh dengan menggunakan metode least square menjadi bias. Untuk mendapatkan estimasi parameter yang tidak bias dengan tetap mempertahankan outlier yang ada, maka pengestimasian parameter kembali dilakukan dengan menggunakan metode estimasi M robust.

Dalam estimasi M robust, proses estimasi parameter dilakukan secara iteratif. Untuk iterasi awal, parameter diestimasi menggunakan metode kuadrat terkecil sehingga diperoleh residual ϵ_i^0 sebagai nilai awal. Nilai residual tersebut selanjutnya digunakan untuk mencari nilai pembobot awal w_i^0 yang akan ditambahkan ke dalam estimasi kuadrat terkecil untuk mengestimasi parameter pada iterasi berikutnya. Proses iterasi ini dilakukan terus-menerus hingga diperoleh selisih antara estimator parameter pada iterasi ke- h dengan estimator parameter pada iterasi ke- $(h-1)$ sama dengan 0 atau mendekati 0 dengan $h = 1, 2, 3, \ldots$. Berikut ini hasil iterasi estimasi parameter menggunakan estimasi M robust dengan bantuan software R.

Iterasi ke-	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\alpha_1}$	$\hat{\alpha_2}$	$\hat{\alpha_3}$	$\hat{\alpha_4}$
0	-1616.031550	5.684927	348.044151	-2311.741403	2027.114158	-1573.797732
1	-1602.803495	5.674646	345.196926	-2242.106963	1957.557111	-1487.368753
2	-1597.750919	5.670719	344.109403	-2213.438398	1929.032956	-1461.566037
3	-1596.907051	5.670063	343.927768	-2204.938761	1920.461435	-1455.311966
4	-1596.064756	5.669408	343.746471	-2201.358775	1916.975454	-1453.555194
5	-1595.548832	5.669007	343.635423	-2199.749593	1915.445068	-1453.086294
6	-1595.273316	5.668793	343.576120	-2199.002448	1914.744464	-1452.971451
7	-1595.134332	5.668685	343.546205	-2198.650630	1914.417215	-1452.948495
8	-1595.066122	5.668632	343.531523	-2198.484034	1914.262964	-1452.946820
⋮	⋮	⋮	⋮	⋮	⋮	⋮
23	-1595.002846	5.668583	343.517904	-2198.333533	1914.124124	-1452.952550
24	-1595.002845	5.668583	343.517904	-2198.333531	1914.124123	-1452.952550
25	-1595.002845	5.668583	343.517904	-2198.333530	1914.124122	-1452.952550
26	-1595.002845	5.668583	343.517904	-2198.333530	1914.124122	-1452.952550

TABEL 5. Hasil Iterasi Estimasi Parameter Menggunakan Estimasi M Robust

Berdasarkan Tabel 5, terlihat bahwa estimator parameter pada iterasi ke-26 sama dengan estimator parameter pada iterasi ke-25 sehingga proses iterasi berhenti pada iterasi ke-26. Dengan demikian, model regresi semiparametrik spline yang terbentuk adalah sebagai berikut

$$\begin{aligned}\hat{y}_i = &- 1595.002845 + 5.668583x_i + 343.517904t_i - 2198.333530\,(t_i - 4.64)_+ \\ &+ 1914.124122\,(t_i - 4.67)_+ - 1452.952550\,(t_i - 6.6)_+ .\end{aligned}$$

Dari persamaan tersebut, dilakukan perhitungan untuk mencari nilai GCV. Dengan bantuan software R, diperoleh nilai GCV untuk metode estimasi M robust yaitu sebesar 68,1755 . Selanjutnya nilai GCV ini dibandingkan dengan nilai GCV yang diperoleh menggunakan metode kuadrat terkecil untuk mengetahui estimasi mana yang lebih baik. Hasil perbandingan keduanya dapat dilihat pada Tabel 6.

Metode	Nilai GCV
MKT	275.4717
Estimasi M	68,1755

Tabel 6. Perbandingan Nilai GCV MKT dan Estimasi M

Dari Tabel 6, dapat dilihat bahwa metode estimasi M memiliki nilai GCV yang lebih kecil dibandingkan nilai GCV yang dihasilkan menggunakan metode kuadrat terkecil. Dengan demikian, metode estimasi M lebih baik digunakan dalam mengestimasi parameter model regresi semiparametrik spline yang mengandung outlier.

5. PENUTUP

5.1. **Kesimpulan.** Berdasarkan hasil penelitian yang telah dilakukan, maka diperoleh kesimpulan sebagai berikut:

(1) Model regresi semiparametrik spline yang terbentuk setelah diestimasi menggunakan metode estimasi M yaitu

$$\hat{y}_i = -1595.002845 + 5.668583x_i + 343.517904t_i - 2198.333530\,(t_i - 4.64)_+ + 1914.124122\,(t_i - 4.67)_+ - 1452.952550\,(t_i - 6.6)_+\,,$$

sedangkan model regresi semiparametrik spline yang terbentuk dari hasil estimasi menggunakan metode kuadrat terkecil yaitu

$$\hat{y}_i = -1616.031550 + 5.684927x_i + 348.044151t_i - 2311.741403\,(t_i - 4.64)_+ + 2027.114158\,(t_i - 4.67)_+ - 1573.797732\,(t_i - 6.6)_+\,.$$

(2) Nilai GCV model regresi semiparametrik spline yang dihasilkan dari metode estimasi M lebih kecil dibandingkan nilai GCV model regresi yang dihasilkan dari metode kuadrat terkecil. Ini berarti metode estimasi M lebih baik dalam mengestimasi parameter model regresi semiparametrik spline yang mengandung outlier.

5.2. **Saran.** Pada penelitian ini, basis fungsi spline yang digunakan yaitu fungsi spline potongan (truncated) dengan memuat satu variabel prediktor pada masing-masing komponen parametrik dan nonparametrik. Selain itu, metode estimasi yang digunakan untuk mendapatkan nilai awal iterasi pada metode estimasi M yaitu metode kuadrat terkecil (least square). Untuk itu, pada penelitian selanjutnya dapat digunakan basis fungsi spline lain seperti B-spline atau P-spline. Penelitian juga dapat menggunakan variabel prediktor lebih dari satu pada masing-masing komponen parametrik dan nonparametrik, serta estimasi untuk nilai awal iterasi pada metode estimasi M dapat menggunakan metode estimasi kuadrat terkecil terpenalti (penalized least square).

Referensi

[1] Budiantara, I.N., Suryadi, F., Otok, B.W., dan Guritno, S., Pemodelan B-Spline dan MARS pada Nilai Ujian Masuk terhadap IPK Mahasiswa Jurusan Disain Komunikasi Visual UK. Petra Surabaya, *Jurnal Teknik Industri*, 8 (2006), 1-13.
[2] Draper, N.R. dan Smith, H., *Applied Regression Analysis*, Third Edition, John Wiley and Sons, Inc., New York, 1998.
[3] Eubank, R.L., *Spline Smoothing and Nonparametric Regression*, Second Edition, Marcel Dekker, Inc., New York, 1999.
[4] Gao, J. dan Shi, P., M-Type Smoothing Splines in Nonparametric and Semiparametric Regression Models, *Statistica Sinica*, 7 (1997), 1155-1169.
[5] Huber, P.J. dan Ronchetti, E., *Robust Statistics*, John Wiley and Sons, Inc., New York, 2009.
[6] Hubert, M. dan Debruyne, M., Minimum Covariance Determinant, *WIREs Computational Statistics*, 2 (2010), 36-43.
[7] Lee, T.C.M. dan Oh, H.S., Robust Penalized Regression Spline Fitting with Application to Additive Mixed Modeling, *Computational Statistics - Springer*, 22 (2007), 159-171.
[8] Myers, R.H, *Classical and Modern Regression with Application, Second Edition*, Duxbury/Thompson Learning, Boston, 1990.
[9] Pradewi, E.D. dan Sudarno, Kajian Estimasi-M IRLS Menggunakan Fungsi Pembobot Huber dan Bisquare Tukey pada Data Ketahanan Pangan di Jawa Tengah, *Media Statistika*, 5 (2012), 1-10.
[10] Rousseeuw, P. dan Leroy, A., *Robust Regression and Outlier Detection*, John Wiley and Sons, Inc., New York, 1987.
[11] Soemartini, *Pencilan (Outlier)*, Penerbit Universitas Padjajaran, Bandung, 2007.

Putri Nilam Cayo* (Penulis Korespondensi)
Departemen Matematika, Fakultas MIPA, Universitas Gadjah Mada, Indonesia
putri.nilam.c@mail.ugm.ac.id

Sri Haryatmi
Departemen Matematika, Fakultas MIPA, Universitas Gadjah Mada, Indonesia
s_kartiko@yahoo.com