BAB III

REGRESI SPLINE

3.1 Fungsi Pemulus Spline

Fungsi regresi nonparametrik yang telah dituliskan pada bab sebelumnya yaitu

$$Y_i = f(x_i) + \varepsilon_i$$

dimana $f(x)$ merupakan fungsi pemulus yang tidak spesifik, dengan ε_i adalah faktor pengganggu. Menurut Fahmeir dan Tuhtz (1994 : 152) taksiran kurva pemulus $\hat{f}(x_i)$ diperoleh dari data observasi (x_i, y_i) dengan $i = 1,2 \ldots, n$. Fungsi $f(x_i)$ merupakan kurva regresi yang tidak diketahui bentuknya, tetapi $f(x_i)$ hanya diasumsikan mulus (*smooth*), dalam arti termuat dalam suatu ruang fungsi tertentu khususnya ruang Sobolev atau ditulis $f \epsilon W_2^p[a, b]$ dengan

$$W_2^p[a, b] = \left\{ f : \int_a^b \left[f^{(p)}(t) \right]^2 dt < \infty \right\} \quad (3.1)$$

untuk suatu p bilangan positif, dan e_i sesatan random yang diasumsikan berdistribusi normal dengan rata-rata nol dan variansi σ^2(Wahba, 1990 : 10)..

Untuk mendapatkan taksiran kurva regresi f digunakan optimasi :

$$\text{Min}_{f \epsilon W_2^p[a,b]} \sum_{i=1}^{n} (Y_i - f(x_i))^2 \quad (3.2)$$

dengan suatu syarat,

$$g(f) = \int_a^b \left[f^{(p)}(t)\right]^2 dt \leq \rho, \quad \rho \geq 0 \tag{3.3}$$

Taksiran ini ekuivalen dengan *Penalized Least Square* (PLS) yaitu penyelesaian optimasi

$$\text{PLS} = n^{-1} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b (f^p(x_i)^2 \, dx \tag{3.4}$$

(Wahba, 1990 : 18)

Dari persamaan (3.4), $\sum_{i=1}^{n} (y_i - f(x_i))^2$ merupakan *The residual Sum of Square* (RSS) atau jumlah kuadrat sisaan, yang merupakan sebuah fungsi jarak antara data dan taksiran. Sedangkan $\lambda \int_a^b (f^p(x_i)^2$ merupakan *Penalized Roughness of The Function*, yakni ukuran kemulusan atau kekasaran kurva dalam memetakan data dimana $0 < \lambda < 1$ adalah parameter penghalus pengontrol keseimbangan antara kecocokan data dan kemulusan kurva (*penalty*). Lebar λ (dari interval) disebut parameter penghalus. Jika λ besar (interval kecil), maka akan diperoleh penaksir dengan bias yang besar tetapi memiliki variansi yang kecil (*oversmoothing*) atau penaksir kurva yang diperoleh akan semakin mulus. Sebaliknya jika λ kecil (interval besar), maka akan diperoleh penaksir dengan bias yang kecil namun variansinya besar (*undersmoothing*). Dengan kata lain ukuran standar jumlah kuadrat galat akan mendominasi kriteria penaksiran kurva, sehingga mengakibatkan kurva menjadi sangat fluktuatif (Simanjuntak, 2009).

Pada persamaan (3.4), pemilihan λ yang optimal sangat penting untuk mendapatkan model penaksir kurva yang baik. Pada nilai λ yang besar maka kurvanya kasar atau sebaiknya, untuk nilai λ yang kecil maka kurvanya akan menjadi mulus (*smooth*), dimana fungsi yang mulus terlihat jelas secara

geometrik, ketika gradien kurva pada titik-titik knot tertentu tidak berubah dengan cepat (Eubank, 1999 : 239).

3.2 Regresi Spline

Spline merupakan potongan $(piecewise)$ polinomial orde ke-m yang memiliki sifat tersegmen kontinu sehingga efektif menjelaskan karakteristik lokal dari fungsi data. Dalam spline digunakan *truncated power basis* dengan k knot, misalnya $K_1, K_2, \dots, K_k$, yaitu :

$$1, t, \dots, (t - K_1)_+^m, \dots, (t - K_k)_+^m,$$

dimana m menunjukkan orde polinomial dari *truncated power basis*, dan untuk orde $m = 0,1,2$ dan 3 secara berturut-turut merupakan *truncated power basis* konstan, linear, kuadratik dan kubik. (Wu dan Zhang, 2006 : 51)

Taksiran kurva $f(x)$ adalah $\hat{f}_\lambda(x)$ yakni penaksir kurva yang mulus, diperoleh melalui model regresi polinomial. Dengan mempertimbangkan sifat-sifat fungsi spline, yang merupakan modifikasi dari regresi polinomial, maka untuk mendapatkan model taksiran dari kurva digunakan regresi spline.

Model regresi spline orde ke-2 adalah

$$y_i = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \sum_{j=1}^{k} \lambda_j (X - K_j)_+^m + \varepsilon_i \qquad (3.8)$$

Model regresi spline orde ke-2 pada persamaan (3.8) biasanya disebut dengan model regresi spline kuadratik. Sedangkan Model regresi spline orde ke-3 adalah

$$y_i = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \beta_3 X^3 + \sum_{j=1}^{k} \lambda_j (X - K_j)_+^m + \varepsilon_i \qquad (3.9)$$

Model regresi spline pada persamaan (3.9) biasanya disebut model regresi spline kubik. Dengan demikian bentuk umum regresi spline orde ke-m adalah

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j X^j + \sum_{j=1}^{k} \lambda_j (X - K_j)_+^m + \varepsilon_i \qquad (3.10)$$

$$y_i = \hat{y} + \varepsilon_i$$

Selanjutnya model regresi spline dapat ditulis menjadi :

$$y_i = \beta_1 x^1 + \cdots + \beta_m x^m + \lambda_1 (X - K_1)_+^m + \cdots + \lambda_k (X - K_k)_+^m + \varepsilon_i \qquad (3.11)$$

dengan menggunakan data sebanyak n, maka bentuk matriks dari persamaan (3.11) ditulis sebagai berikut :

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 & \cdots & x_1^m & (x_1 - K_1)_+^m & \cdots & (x_1 - K_k)_+^m \\ x_2 & \cdots & x_2^m & (x_2 - K_1)_+^m & \cdots & (x_1 - K_k)_+^m \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_n & \cdots & x_n^m & (x_n - K_1)_+^m & \cdots & (x_n - K_k)_+^m \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_m \\ \lambda_1 \\ \vdots \\ \lambda_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

3.3 Penaksiran Parameter

Fungsi kurva pada model regresi nonparametrik, seperti yang telah dituliskan pada persamaan (2.7) dapat dinyatakan sebagai berikut :

$$f(x_i) = \sum_{r=1}^{\infty} \beta_r x_r$$

sehingga model regresi nonparametrik menjadi

$$y_i = \sum_{r=1}^{\infty} \beta_r x_r(x_i) + \varepsilon_i \quad ; \; i = 1,2,\ldots,n \qquad (3.12)$$

Karena $\sum_{r=1}^{\infty}|\beta_r|^2 < \infty$ dan β_r menuju nol, maka terdapat bilangan λ sedemikian sehingga fungsi f dapat didekati dengan

$$f(x_i) = \sum_{r=1}^{\lambda} \beta_r x_r$$

sehingga model menjadi

$$y_i = \sum_{r=1}^{\lambda} \beta_r x_r(x_i) + \varepsilon_i \quad ; \quad i = 1,2,\dots,n \tag{3.13}$$

Penaksir kurva mulus regresi nonparametrik $f(x)$ harus mempunyai λ optimal. Misalkan $f(x)$ terdapat pada kelas penaksir, $C(\Lambda) = \{f_\lambda : \lambda \epsilon \Lambda\}$ dengan Λ mewakili beberapa himpunan indeks. Untuk mempermudah, akan diasumsikan elemen $C(\Lambda)$ merupakan penaksir linear. Artinya bahwa, untuk setiap λ tedapat matriks $H(\lambda)$ berukuran $(n \times n)$ sehingga

$$\hat{f}_\lambda = H(\lambda)y \tag{3.14}$$

dengan $H(\lambda)$ merupakan matriks *hat,* yaitu matriks yang bersifat simetris dan semi definit positif.

3.3.1 Penaksiran Kurva Regresi

Penaksiran kurva regresi nonparametrik pada suatu data menggunakan pendekatan spline didasarkan pada *tuncated power basis*. Telah diberikan persamaan regresi nonparametrik pada persamaan (2.7), dimana Y_i merupakan variabel respon ke-i, $f(x_i)$ merupakan fungsi yang tidak diketahui dengan x_i merupakan variabel prediktor ke-i dan nilai ε_i adalah faktor pengganggu yang

tidak dapat dijelaskan oleh model, sedangkan n menyatakan banyak objek yang diamati. Penaksiran kurva regresi dilakukan dengan menyelesaikan optimasi,

$$\underset{f \epsilon W_2^p[a,b]}{\text{Min}} \sum_{i=1}^{n} (Y_i - f(x_i))^2$$

Misalkan, $f(x_i) = \varphi(k)\boldsymbol{B}$ dimana, $\varphi(k)$ merupakan matriks yang dapat ditulis sebagai berikut :

$$\begin{bmatrix} x_1 & \cdots & x_1^m & (x_1 - K_1)_+^m & \cdots & (x_1 - K_k)_+^m \\ x_2 & \cdots & x_2^m & (x_2 - K_1)_+^m & \cdots & (x_1 - K_k)_+^m \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_n & \cdots & x_n^m & (x_n - K_1)_+^m & \cdots & (x_n - K_k)_+^m \end{bmatrix} \tag{3.15}$$

Maka dengan suatu bobot V dan menganggap $\varphi(k)$ sebagai variabel prediktor (X) diperoleh penaksir $\boldsymbol{B}$,

$$\widehat{\boldsymbol{B}} = (\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{Y} \tag{3.16}$$

dan taksiran Y,

$$\hat{Y} = \boldsymbol{X}\widehat{\boldsymbol{B}}$$

$$\hat{Y} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{Y}$$

Suatu fungsi spline dengan titik-titik knot $(K_1, K_2, \dots, K_k)$ yang didefinisikan pada persamaan (2.12) yaitu

$$f(x) = \beta_0 + \sum_{j=1}^{m} \beta_j X^j + \sum_{j=1}^{k} \lambda_j (X - K_j)_+^m$$

dengan $K_1 < \cdots < K_k$ adalah k buah knot yang tetap dan $(\beta_0, \beta_1, \dots, \beta_m, \lambda_1, \lambda_2, \dots, \lambda_k)$ adalah parameter. Dalam notasi matriks persamaan (2.12) dapat ditulis menjadi :

$$f(x) = \varphi(k)\boldsymbol{B}$$

dengan $\varphi(k)$ seperti pada persamaan (3.15) dan $\boldsymbol{B} = (\beta_0, \beta_1, \dots, \beta_m, \lambda_1, \lambda_2, \dots, \lambda_k)'$. Sehingga dari model regresi nonparametrik dapat diperoleh persamaan :

$$Y_i = \varphi(k)\boldsymbol{B} + \varepsilon_i \tag{3.17}$$

Misalkan, $\boldsymbol{Y}_i = (Y_1, Y_2, \dots, Y_n)'$, $\boldsymbol{M}_i = [\varphi_1(x_1), \varphi_2(x_2), \dots, \varphi_2(x_2)]$ dan $\boldsymbol{B}_i = (\beta_0, \beta_1, \dots, \beta_m, \lambda_1, \lambda_2, \dots, \lambda_k)'$ dan $e_i = (e_1, e_2, \dots, e_n)'$ maka persamaan (3.17) dapat ditulis sebagai berikut :

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} M_1 & 0 & \cdots & 0 \\ 0 & M_2 & \cdots & 0 \\ \vdots & \vdots & M_3 & \vdots \\ 0 & \cdots & 0 & M_4 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{3.18}$$

Bentuk penyederhanaan dari persamaan (3.18) yang ditulis dalam bentuk matriks adalah sebagai berikut :

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{\varepsilon} \tag{3.19}$$

dengan $\boldsymbol{X} = diag(M_1, M_2, \dots, M_n)^T$

Penaksir dari $\boldsymbol{B}$ dapat diperoleh dengan menyelesaikan optimasi,

$$(\boldsymbol{Y} - \boldsymbol{XB})^T \boldsymbol{V} (\boldsymbol{Y} - \boldsymbol{XB}) \tag{3.20}$$

dimana $\boldsymbol{V} = diag(V_1, V_2, \dots, V_n)$, $V_i = v_i \boldsymbol{I}$ dengan $v_i = \frac{1}{N}$; $i = 1,2, \dots, n$. Dengan kriteria Metode Kuadrat Terkecil, penyelesaian dari optimasi (3.20) adalah

$$(\boldsymbol{Y} - \boldsymbol{XB})^T \boldsymbol{V} (\boldsymbol{Y} - \boldsymbol{XB}) = \boldsymbol{Y^T V Y} - 2\boldsymbol{B^T X^T V Y} + \boldsymbol{B X^T V X B}$$

Jika dimisalkan $P = \boldsymbol{Y^T V Y} - 2\boldsymbol{B^T X^T V Y} + \boldsymbol{B^T X^T V X B}$ maka

$$\frac{\partial P}{\partial B} = -2\boldsymbol{X^T V Y} + 2\boldsymbol{X^T V X B}$$

dan nilai optimum dari $\boldsymbol{B}$ diperoleh dari,

$$\frac{\partial \boldsymbol{P}}{\partial \boldsymbol{B}} = 0$$

$$= -2\boldsymbol{X^T V Y} + 2\boldsymbol{X^T V X B}$$

$$= -\boldsymbol{X^T V Y} + \boldsymbol{X^T V X B}$$

$$\boldsymbol{X^T V X B} = \boldsymbol{X^T V Y}$$

Jika kedua ruas dikalikan dengan $(\boldsymbol{X^T V X})^{-1}$ maka diperoleh

$$(\boldsymbol{X^T V X})^{-1} \boldsymbol{X^T V X B} = (\boldsymbol{X^T V X})^{-1} \boldsymbol{X^T V Y}$$

Sehingga penaksir dari $\boldsymbol{B}$ adalah,

$$\widehat{\boldsymbol{B}} = (\boldsymbol{X^T V X})^{-1} \boldsymbol{X^T V Y} \tag{3.21}$$

Jika $\boldsymbol{B}$ dalam (3.19) disubstitusi dengan penaksirnya yang ada pada (3.21) maka diperoleh

$$\boldsymbol{Y} = \boldsymbol{X}\widehat{\boldsymbol{B}} + \boldsymbol{\varepsilon}$$

sehingga diperoleh taksiran $\boldsymbol{Y}$ sebagai berikut :

$$\begin{aligned} \widehat{\boldsymbol{Y}} &= \boldsymbol{X}\widehat{\boldsymbol{B}} \\ &= \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}\boldsymbol{X})^{-1}\,\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}\boldsymbol{Y} \\ &= \boldsymbol{H}\boldsymbol{Y} \end{aligned} \tag{3.22}$$

dimana matriks *hat H* adalah $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}\boldsymbol{X})^{-1}\,\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}$ (Basri, 2008).

3.4 Pemilihan Model Regresi Spline

Model spline yang baik adalah model yang mampu menjelaskan hubungan antara variabel prediktor (X) dengan variabel respon (Y) dan memenuhi beberapa kriteria tertentu, antara lain mempunyai nilai *Mean Squared Error* (MSE) yang minimum dan nilai *Generalized Cross Validation* (GCV) yang minimum. Nilai MSE merupakan nilai taksiran dari varians residual sehingga model terbaik adalah model yang dengan MSE minimum yang menandakan nilai taksiran mendekati nilai sebenarnya.

3.4.1 Kriteria *Mean Square Error* (MSE)

Kriteria sederhana yang digunakan sebagai ukuran kinerja atas penaksir yang baik adalah *Mean Square Error* (MSE). Seperti yang telah dituliskan pada persamaan (2.14), yaitu :

$$\text{MSE}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_\lambda(x_i))^2$$

3.4.2 Kriteria *Generalized Cross Validation* (GCV)

Kriteria lain yang dapat digunakan sebagai ukuran kinerja atas penaksir yang baik adalah *Generalized Cross Validation.* Seperti yang telah dituliskan pada persamaan (2.15), yaitu :

$$\text{GCV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{f}_\lambda(x_i)}{1 - {}^{v}/_{n}}\right]^2$$

$$= \frac{n^{-1}\sum_{j=1}^{n}(y_i - \hat{f}_\lambda(x_i))^2}{\{n^{-1}(tr(I) - tr(H(\lambda))\}^2}$$

Matriks $H(\lambda)$ merupakan matriks *hat* yang telah diperoleh sebelumnya, yaitu $\boldsymbol{H = X(X^T V X)^{-1} X^T V.}$

3.5 Pemilihan Knot Dalam Regresi Spline

Pemilihan knot sangat penting, karena berpengaruh pada model regresi spline yang akan dipilih. Terdapat 2 strategi untuk memilih knot yang baik. Strategi pertama adalah memilih banyaknya knot yang relatif sedikit, sedangkan strategi yang kedua adalah kebalikannya, yakni menggunakan knot yang relatif banyak. Diantara kedua strategi tersebut, strategi kedua lebih banyak digunakan pada model yang sangat memperhatikan pola matematis yang ada pada data. Sedangkan strategi pertama, lebih mengarah pada alasan kesederhanaan model (Wand, 2000).

Penentuan lokasi knot yang berbeda akan menghasilkan model regresi spline yang berbeda pula. Lokasi knot tersebut akan berpengaruh terhadap nilai kriteria $MSE(\lambda)$ dan $GCV(\lambda)$ dari model regresi spline yang dibentuk. Pengaruh banyaknya knot terhadap model regresi spline adalah jika model menggunakan orde yang besar maka knot yang cukup efektif yang digunakan adalah semakin sedikit.