ISBN : 978 – 979 – 16353 – 6 – 3

S – 2

Analisis Regresi Spline Kuadratik

Oleh: Agustini Tripena
Program Studi Matematika, Fakultas Sains dan Teknik,
Univesitas Jenderal Soedirman, Purwokerto
tripena1960@yahoo.co.id

Abstrak

Regresi spline merupakan salah satu model dengan pendekatan nonparametrik, yang merupakan modifikasi dari fungsi polinomial tersegmen. Bentuk estimator spline sangat dipengaruhi oleh nilai parameter penghalus $\boldsymbol{\lambda}$ yang pada hakekatnya adalah penentuan lokasi titik-titik knot. Pemilihan optimal dalam regresi spline berarti pemilihan lokasi titik-titik knot. Oleh karena itu, penentuan titik knot optimal merupakan persoalan yang sangat penting dalam estimasi regresi spline.

Data yang digunakan dalam penelitian ini adalah data tentang pengaruh penambahan waktu terhadap perubahan konduktansi. Tujuan dari penelitian adalah mengetahui estimasi model regresi spline kuadrat dengan empat titik knot. Titik knot yang optimal K_1 = 15, K_2 = 55, K_3 = 70 dan K_4 = 75. Nilai GCV model regresi spline kuadrat optimal sebesar 0,00003611717. Model spline kuadrat adalah

$$\hat{y}_i = 2{,}96965762 - 0{,}013640268x_i - 0{,}000429993x_i^2 + 0{,}000618387(x_i - 15)_+^2 - 0{,}000132437(x_i - 55)_+^2 + 0{,}000160092(x_i - 70)_+^2 - 0{,}000168288(x_i - 75)_+^2$$

Kata kunci: regresi nonparametrik, spline kuadratik, titik knots, GCV.

1. PENDAHULUAN

Dalam analisis regresi, pola hubungan antara dua variabel atau lebih tidak selalu berpola parametrik seperti linier, kuadrat, kubik dan yang lainnya tetapi terdapat banyak kasus dimana pola hubungan antar variabel berpola nonparametrik (Eubank,1988). Pendekatan regresi parametrik digunakan jika bentuk kurva regresi diketahui, sedangkan pendekatan regresi nonparametrik digunakan apabila informasi mengenai bentuk dan pola hubungan antara variabel prediktor dengan variabel respon tidak diketahui (Budiantara, 2005).

Regresi spline adalah suatu pendekatan ke arah pengepasan data dengan tetap memperhitungkan kemulusan kurva. Spline merupakan model polynomial yang tersegmen. Sifat tersegmen inilah yang memberikan fleksibelitas yang lebih baik daripada model polynomial biasa. Sifat ini memungkinkan model regresi spline menyesuaikan diri secara efektif terhadap karakteristik lokal dari data. Penggunaan spline difokuskan kepada adanya perilaku atau pola data, yang pada daerah tertentu, mempunyai karakteristik yang berbeda dengan daerah lain. Pencocokan data dapat dilakukan dengan melihat titik-titik pada data yang mengalami suatu perubahan ekstrim pada suatu daerah sehingga pola data pada masing-masing daerah mengalami perbedaan. Regresi spline linier biasanya diaplikasikan pada data dengan pola yang

 ISBN : 978 – 979 – 16353 – 6 – 3

masih sederhana sedangkan spline kuadrat dan kubik biasanya diaplikasikan pada data dengan pola data yang lebih kompleks.

Bentuk estimator spline sangat dipengaruhi oleh nilai parameter penghalus λ (Budihantara, 2000). Bentuk estimator spline juga dipengaruhi oleh lokasi dan banyaknya titik-titik knot. Eubank (1988) menyimpulkan bahwa pemilihan λ optimal dalam regresi spline pada hakekatnya merupakan pemilihan lokasi titik knot. Penentuan titik knot optimal untuk memilih model regresi spline terbaik didasarkan pada nilai *GCV (Generalized Cross-Validation)*. Permasalahan yang timbul adalah bagaimana mengestimasi dan bagaimana menentukan banyaknya titik knot dan lokasi titik knot serta bagaimana cara memilih model regresi spline kuadrat terbaik menggunakan kriteria GCV (λ)?. Tujuan dari penelitian ini adalah memperoleh estimasi model regresi spline kuadrat dan menentukan banyaknya titik knot dan lokasi titik knot serta memilih model regresi spline terbaik menggunakan kriteria $GCV(\lambda)$. Manfaat penelitian adalah bagaimana cara menerapkan pendekatan spline untuk memilih model regresi yang dapat menggambarkan pola pada data.

2. Regresi Nonparametrik

Regresi nonparametrik merupakan suatu metode statistika yang digunakan untuk mengetahui hubungan antara variabel respon dengan variabel prediktor yang tidak diketahui bentuk fungsinya dan hanya diasumsikan mulus (*smooth*), sehingga regresi nonparametrik sangat mempertahankan fleksibilitasnya . Model regresi nonparametrik secara umum adalah sebagai berikut (Eubank, 1988) :

$$y_i = f(x_i) + \varepsilon_i \,, \qquad i = 1,2,\dots,n \tag{1}$$

dengan, y_i: variabel respon; x_i: variabel prediktor; $f(x_i)$: fungsi regresi; ε_i: galat (*error*) yang berdistribusi normal, independen dengan mean nol dan variansi σ^2

2.1 Fungsi *Spline polynomial truncated*

Fungsi spline berorde m dengan titik-titik knot $K_1, K_2, \dots, K_{sj}$ didefinisikan sebagai sembarang fungsi f yang disajikan dalam bentuk (Eubank, 1988) :

$$f(x) = \beta_0 + \sum_{j=1}^{m} \beta_j x^j + \sum_{k=1}^{N} \beta_{j+k}(x - K_k)_+^m \tag{2}$$

dengan fungsi sesepenggal (*truncated*) sebagai berikut:

$$(x - K_k)_+^m = \begin{cases} (x - K_k)^m & ,\text{jika } \; x \geq K_k \\ 0 & ,\text{jika } \; x < K_k \end{cases}$$

dengan, β : parameter model; β_0: *intersep*; β_{j+k}: *slope* pada peubah x *truncated* knot ke-k pada spline berorde m; x : variabel respon; K_k : knot ke-k; N: banyaknya knot dalam variabel respon ke-j; β_j adalah konstanta real dan $K_1, K_2, \ldots, K_N$ adalah titik knot. Spline adalah jumlahan dari fungsi polinomial berderajat m dengan *truncated* derajat m, fungsi spline dapat disajikan dalam bentuk sebagai berikut (Hardle, 1990):

$$f(x) = \beta_0 + \beta_1 x + \cdots + \beta_m x^m + \sum_{k=1}^{N} \beta_{j+k}(x - K_k)_+^m \qquad (3)$$

Berdasarkan bentuk matematis fungsi spline, dapat dikatakan bahwa spline merupakan model polinomial yang sepotong-sepotong (*Piecewise Polynomial*) dan, spline masih bersifat kontinu pada knot-knotnya. Knot diartikan sebagai suatu titik fokus dalam fungsi spline, sehingga kurva yang dibentuk tersegmen pada titik tersebut dan untuk setiap fungsi m, titik knot dapat dinyatakan dengan kombinasi linier. Fungsi spline merupakan suatu gabungan fungsi polinomial dimana penggabungan beberapa polinomial tersebut pada knot-knot dengan suatu cara yang menjamin sifat kontinuitas. Spline adalah potongan polinomial mulus yang masih memungkinkan memiliki sifat tersegmen (Eubank, 1988).

2.2 Fungsi Spline Linier

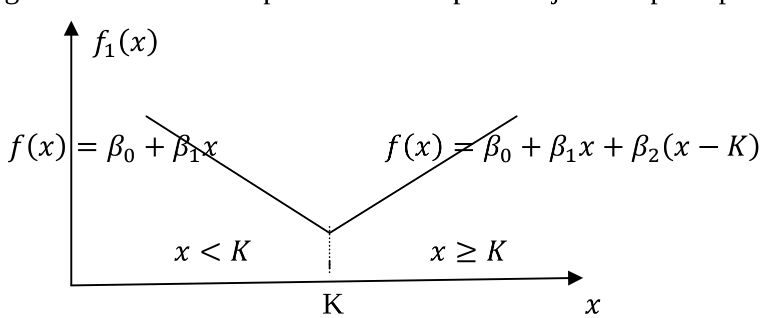
Fungsi spline linier merupakan fungsi spline dengan satu orde. Fungsi spline linier dengan satu titik knots (K) dapat disajikan dalam bentuk:

$$f_1(x) = \beta_0 + \beta_1 x + \beta_2 (x - K)_+^1 \qquad (4)$$

Fungsi ini dapat pula disajikan menjadi (Tripena, 2005) :

$$f_1(x) = \begin{cases} \beta_0 + \beta_1 x & , x < K \\ \beta_0 + \beta_1 x + \beta_2 (x - K), & x \geq K \end{cases} \qquad (5)$$

Grafik spline linier dengan satu titik knots pada $x = K$ dapat disajikan seperti pada Gambar 1 berikut:



Gambar 1. Fungsi spline linier dengan satu titik knot pada $x = K$

 ISBN : 978 – 979 – 16353 – 6 – 3

Fungsi spline dengan empat titik knot pada $(x = K_1, x = K_2, x = K_3, x = K_4)$

$$f_4(x) = \beta_0 + \beta_1 x + \beta_2 (x - K_1)_+^1 + \beta_3 (x - K_2)_+^1 + \beta_4 (x - K_3)_+^1 + \beta_5 (x - K_4)_+^1$$

Fungsi $f_4(x)$ dapat disajikan dalam bentuk:

$$f_4(x) = \begin{cases} \beta_0 + \beta_1 x & , x < K_1 < K_2 < K_3 < K_4 \\ \beta_0 + \beta_1 x + \beta_2 (x - K_1) & , K_1 \leq x < K_2 \leq K_3 < K_4 \\ \beta_0 + \beta_1 x + \beta_2 (x - K_1) + \beta_3 (x - K_2) & , K_2 \leq x < K_3 < K_4 \\ \beta_0 + \beta_1 x + \beta_2 (x - K_1) + \beta_3 (x - K_2) + \beta_4 (x - K_3), & K_3 \leq x < K_4 \\ \beta_0 + \beta_1 x + \beta_2 (x - K_1) + \beta_3 (x - K_2) + \beta_4 (x - K_3) + \beta_5 (x - K_4) & , x \geq K_4 \end{cases}$$

2.3 Regresi Spline

Menurut Eubank (1988), estimasi terhadap $f(x)$ adalah $f_\lambda(x)$ yakni estimator yang mulus. Dengan mempertimbangkan sifat-sifat fungsi spline yang merupakan modifikasi dari regresi polinomial, maka untuk mendapatkan model estimasi dari y digunakan regresi spline.

Bentuk umum regresi spline orde ke-m dengan satu variabel respon x sesuai persamaan (5) adalah $y = \beta_0 + \sum_{j=1}^{m} \beta_j x^j + \sum_{k=1}^{N} \beta_{j+k} (x - K_k)_+^m + \varepsilon$ (6)

Fungsi estimasinya adalah

$$f_\lambda = \mathbf{X}_\lambda \mathbf{b}_\lambda = \mathbf{X}_\lambda (\mathbf{X}'_\lambda \mathbf{X}_\lambda)^{-1} \mathbf{X}'_\lambda \mathbf{y} = \mathbf{H}_\lambda \mathbf{y}, \in \Lambda \quad (7)$$

dengan $\mathbf{H}_\lambda = \mathbf{X}_\lambda (\mathbf{X}'_\lambda \mathbf{X}_\lambda)^{-1} \mathbf{X}'_\lambda$, $\mathbf{H}_\lambda$ bersifat simetris dan definit positif

Sasmitoadi (2005) menyebutkan bahwa terdapat 2 strategi untuk menyelesaikan permasalahan yaitu pertama memilih banyaknya knot yang relatif sedikit, sedangkan strategi yang kedua adalah kebalikannya, yakni menggunakan knot yang relatif banyak.

2.4 Pemilihan Model Regresi Spline dengan λ yang Optimal

Sesuai tujuan dari pendekatan regresi nonparametrik, yakni ingin didapatkan kurva mulus yang mempunyai λ optimal menggunakan data amatan sebanyak n, maka diperlukan ukuran kinerja atas estimator yang dapat diterima secara universal. Eubank (1988) menyebutkan, ukuran kinerja atas estimator tersebut adalah *Generalized Cross-Validation* (GCV). Menurut Budihantara (2005), GCV merupakan modifikasi dari *Cross-Validation* (CV). *Cross-Validation* (CV) merupakan suatu metode untuk memilih model berdasarkan pada kemampuan prediksi dari model tersebut. Fungsi GCV didefinisikan sebagai:

ISBN : 978 – 979 – 16353 – 6 – 3

$$GCV(\lambda) = n^{-1} \frac{\sum_{i=1}^{n}\left(y_i - f_\lambda(x_i)\right)^2}{\left(1 - n^{-1}Tr(\mathbf{H}_\lambda)\right)^2} \quad (8)$$

dengan $Tr(\mathbf{H}_\lambda) < n$. Kriteria $GCV(\lambda)$ diharapkan memiliki nilai yang minimum, sehingga model regresi spline dapat dikatakan memiliki nilai λ yang optimal.

3. METODE PENELITIAN

3.1 Metode dan Data Penelitian

Metode yang digunakan dalam penelitian ini adalah studi literatur. Data yang digunakan dalam penelitian ini berupa data sekunder yang diperoleh dari sistem akuisisi data untuk memantau Para Hidrolisis Alkali dari Ester dengan mengukur perubahan konduktansi dari campuran reaksi dengan waktu.

3.2.Analisis hasil

Tahapan-tahapan yang dilakukan dalam penelitian ini sebagai berikut:

a. Memilih lokasi dan banyaknya titik knot

Mengestimasi model regresi spline untuk empat knot sekaligus penempatan titik-titik knot tersebut, yakni: odel regresi spline Kuadrat

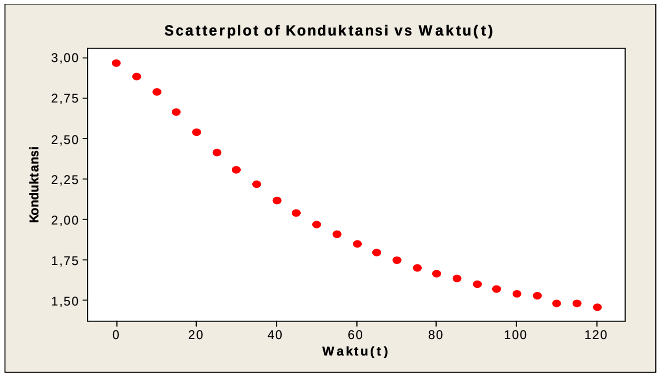
b. Menghitung nilai $GCV(\lambda)$ untuk masing-masing model regresi spline.

c. Memilih model regresi spline terbaik diantara model-model yang didapatkan berdasarkan kriteria $GCV(\lambda)$ minimum. Paket *software S-PLUS* 2000 dan *MINITAB* 14 digunakan untuk memudahkan dalam hal komputasi.

4. HASIL DAN PEMBAHASAN

4.1 Pemilihan Lokasi dan Banyaknya Titik Knot

Model regresi spline terbaik dipilih berdasarkan lokasi dan banyaknya titik knot yang optimal. Bentuk estimator regresi spline sangat dipengaruhi oleh lokasi titik knot dan banyaknya titik-titik knot tersebut. Penentuan lokasi titik knot yang berbeda pada data akan menghasilkan model regresi spline yang berbeda pula. Lokasi titik knot tersebut akan berpengaruh terhadap nilai kriteria $GCV(\lambda)$ dari model regresi spline yang dibentuk. Penyebaran data pengaruh lama waktu (detik) yang diberikan sebagai variabel bebas terhadap perubahan konduktansi sebagai variabel tak bebas dapat dilihat dari bentuk plot yang disajikan pada Gambar 1.



Gambar 2 Plot pengaruh waktu (detik) terhadap konduktansi

Berdasarkan Gambar 1 terlihat bahwa pola data mengalami kecenderungan turun secara tajam pada beberapa detik pertama (detik ke-0 sampai 55). Namun, pada beberapa detik terakhir (detik ke-60 sampai 120) pola data mengalami penurunan yang tidak signifikan. Ada kecenderungan perubahan waktu terhadap konduktansi untuk membentuk pola tertentu.

4.2. Estimasi Regresi Spline Kuadrat

Dicobakan model spline kuadrat dengan empat titik knot adalah

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 (x_i - K_1)_+^2 + \beta_4 (x_i - K_2)_+^2 + \beta_5 (x_i - K_3)_+^2 + \beta_6 (x_i - K_4)_+^2 \varepsilon_i$$

Titik knot optimal yang bersesuaian dengan nilai GCV minimum untuk model spline kuadrat dengan empat titik knot diberikan dalam Tabel 1 sebagai berikut:

No	Titik knot				GCV
1	15	55	70	75	0,00003611717
2	15	50	80	110	0,0000361927
3	15	50	80	90	0,00003640869
4	15	50	80	105	0,00003650289
5	15	50	80	85	0,00003650649

Tabel 1 Nilai GCV model regresi spline kuadrat dengan empat titik knot

ISBN : 978 – 979 – 16353 – 6 – 3

Berdasarkan Tabel 1 didapatkan nilai GCV yang minimum untuk model spline kuadrat dengan empat titik knot sebesar 0,00003611717 yang berada pada titik knot $K_1 = 15$, $K_2 = 55$, $K_3 = 70$, dan $K_4 = 75$. Tabel 2 memberikan estimasi untuk model regresi spline kuadrat dengan empat titik knot $K_1 = 15$, $K_2 = 55$, $K_3 = 70$, dan $K_4 = 75$.

Parameter	Estimasi
β_0	2,96965762
β_1	-0,013640268
β_2	-0,000429993
β_3	0,000618387
β_4	-0,000132437
β_5	0,000160092
β_6	-0,000168288

Tabel 2 Estimasi model regresi spline kuadrat dengan empat titik knot

Sehingga diperoleh estimasi model regresi spline kuadrat dengan empat knot $K_1 = 15$, $K_2 = 55$, $K_3 = 70$, dan $K_4 = 75$ yakni

$$\hat{y}_i = 2{,}96965762 - 0{,}013640268\, x_i - 0{,}000429993\, x_i^2 + 0{,}000618387(x_i - 15)_+^2$$
$$-0{,}000132437(x_i - 55)_+^2 + 0{,}000160092(x_i - 70)_+^2 - 0{,}000168288(x_i - 75)_+^2$$

Model spline ini disajikan dalam Gambar 3.

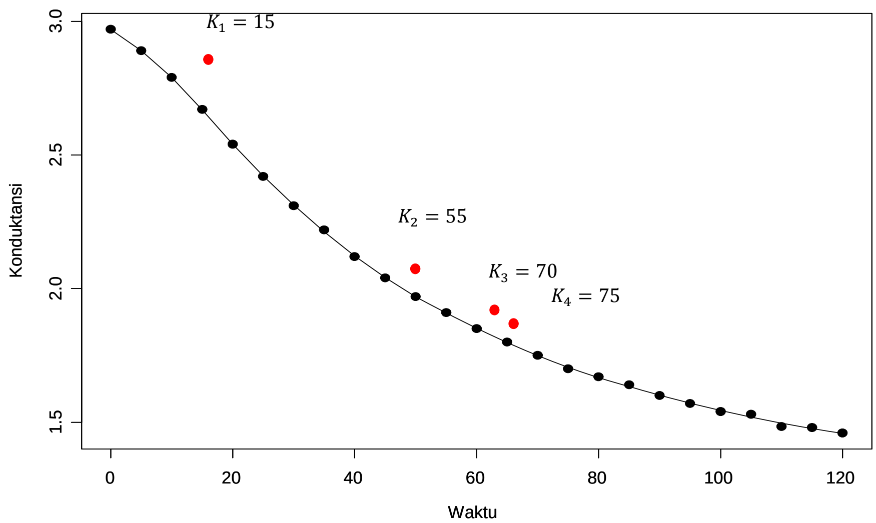
4.3.Pemilihan Model Regresi Spline Terbaik

Hasil yang telah diperoleh, dapat disimpulkan bahwa titik knot (K) yang paling optimal dengan nilai GCV minimum sebesar 0,00003611717 sehingga nilai GCV minimum untuk empat titik knot disajikan pada Tabel 3.

Orde	**Titik knot**				***GCV***
Spline Kuadrat	15	55	70	75	0,00003611717

Tabel 3. Nilai GCV minimum untuk empat titik knot

 ISBN : 978 – 979 – 16353 – 6 – 3



Gambar 3. Kurva estimasi regresi spline kuadrat dengan empat titik knot

4.4. Berdasarkan Tabel 3 dapat disimpulkan bahwa model terbaik adalah model regresi spline kuadrat dengan empat titik knot K_1 = 15, K_2 = 55, K_3 = 70, dan K_4 = 75

$$\hat{y}_i = 2{,}96965762 - 0{,}013640268\, x_i - 0{,}000429993\, x_i^2$$
$$+0{,}000618387(x_i - 15)_+^2 - 0{,}000132437(x_i - 55)_+^2$$
$$+0{,}000160092(x_i - 70)_+^2 - 0{,}000168288(x_i - 75)_+^2$$

Nilai koefisien determinasi (R^2) sebesar 0.92421. Hal ini berarti bahwa variabel pemberian waktu tertentu mampu menerangkan sebesar 92,421% terhadap perubahan konduktansi.

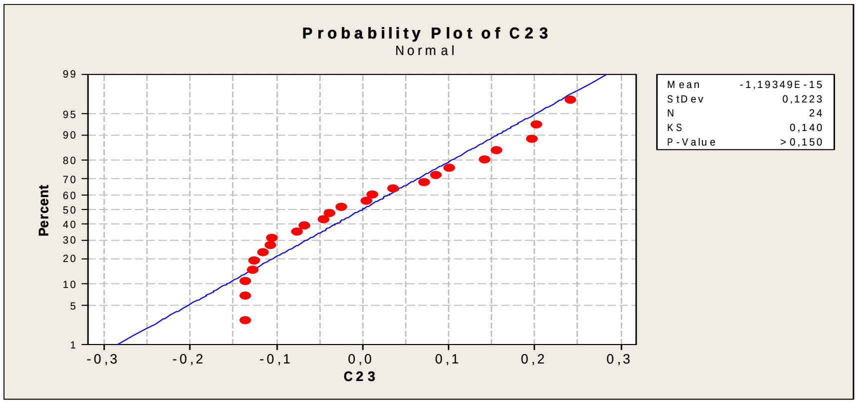
4.5. Pengujian Model Regresi Spline Terbaik

Untuk menguji asumsi ini digunakan uji Kolmogorov-Smirnov dengan hipotesis:

H_0 : *error* random ε_i berdistribusi normal

H_1 : *error* random ε_i tidak berdistribusi normal

dengan menggunakan $\alpha = 0{,}05$. Dari hasil output Minitab diperoleh nilai $P_{value} = 0{,}150 > 0{,}05$. Jadi dapat disimpulkan bahwa *error* random ε_i berdistribusi normal. Plot normalitas *error* random ε_i disajikan dalam Gambar 4.



Gambar 4. Plot normalitas residual

Pengujian selanjutnya adalah uji hipotesis untuk pemeriksanaan model, dengan rumus hipotesis sebagai berikut: H_0: variabel bebas secara bersama-sama tidak berpengaruh terhadap variabel tak bebas' H_1: variabel bebas secara bersama-sama berpengaruh terhadap variabel tak bebas. Dengan menggunakan F_{tabel}, diperoleh $F_{\alpha,p,(n-(p+1))} = F_{0.05,6,18} = 2{,}66100$. Karena $F_{Hitung} = 26.677{,}21 \geq F_{0.05,5,19} = 2{,}66100$ hal ini mengidentifikasi bahwa H_0 ditolak, artinya variabel bebas secara bersama-sama berpengaruh terhadap variabel tak bebas.

Lebih lanjut diuji koefisien-koefisien regresi yang memberikan pengaruh signifikan terhadap model regresi spline dengan menggunakan uji hipotesis sebagai berikut:

H_0: Koefisien regresi β_i tidak berpengaruh, atau $\beta_i = 0, i = 0,1,..,p$

H_1: Koefisien regresi β_i berpengaruh, atau $\beta_i \neq 0$

Uji hipotesis untuk pemeriksaan model dengan menggunakan tingkat signifikansi 5% diperoleh nilai $t_{tabel} = t_{(\alpha/2,(n-(p+1))} = t_{0,025;19} = 2{,}093$. Parameter-parameter yang berpengaruh diperoleh jika $|t_{hitung}| \geq t_{tabel}$. Jadi dapat disimpulkan bahwa model regresi spline kuadrat dengan titik knot $K_1 = 15$, $K_2 = 55$, $K_3 = 70$, dan $K_4 = 75$ cukup memadai sebagai model pendekatan untuk data pengaruh waktu terhadap perubahan konduktansi.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

a. Model regresi spline terbaik yang menggambarkan hubungan pengaruh lama waktu terhadap perubahan konduktansi adalah model regresi spline kuadrat dengan empat titik knot yang diperoleh berdasarkan nilai $GCV(\lambda)$ yang paling minimum yakni sebesar 0,00003611717.

b. Estimasi modelnya adalah sebagai berikut:

$$\hat{y}_i = 2{,}96965762 - 0{,}013640268\, x_i - 0{,}000429993\, x_i^2 + 0{,}000618387(x_i - 15)_+^2$$
$$-0{,}000132437(x_i - 55)_+^2 + 0{,}000160092(x_i - 70)_+^2 - 0{,}000168288(x_i - 75)_+^2$$

c. Titik knot optimal adalah dengan empat titik knot adalah K_1 = 15, K_2 = 50, dan K_4 = 80.

5.2 Saran

a. Penelitian selanjutnya dapat dibuat suatu program untuk menghitung nilai $GCV(\lambda)$ untuk semua kombinasi dari titik knot yang mungkin agar lebih mudah dalam perhitungan.

b. Penelitian selanjutnya dapat digunakan pendekatan yang lainnya untuk mengestimasi model regresi spline.

6. DAFTAR PUSTAKA

Budiantara, I. N, 2002. *Aplikasi Spline Estimator Terbobot* . Jurnal Teknik Industri PETRA, Surabaya.

Budiantara, I. N, 2005. *Model Keluarga Spline Polinomial Truncated Dalam Regresi Semiparametrik*. Makalah Seminar Nasional Matematika, Jurusan Matematika UNDIP Semarang.

Budiantara, I. N, 2005. *Penentuan Titik-Titik Knots dalam Regresi Spline* , Jurnal Jurusan Statistika FMIPA-ITS, Surabaya.

Eubank, R. 1988. *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New York.

Hardle, W. 1990. *Applied Nonparametric Regression.* Cambridge University Press, New York.

 ISBN : 978 – 979 – 16353 – 6 – 3

Sasmitoadi, D. 2005. *Kajian Penggunaan Knot dan Orde pada Regresi Spline*. http://doclines.files.wordpress.com/2008/10/Kajian-Penggunaan-Knot-dan-Orde-pada-Regresi-Spline-pdf1.pdf. diakses tanggal 8 Maret 2010.

Tripena, A. 2005. *Pendekatan Model Regresi Spline Linier* . Jurusan MIPA, Fakultas Sains dan Teknik, UNSOED.