 http://dx.doi.org/10.29313/.v7i1.26272

Penerapan Metode *Restricted Cubic Spline* pada Kasus Data Riwayat Pasien Terinfeksi Virus Corona

Rainahda Syamsidah*, Abdul Kudus

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

*rainahdasys@gmail.com, Abdul.kudus@unisba.ac.id

Abstract. In general, regression analysis requires a linear data relationship pattern, yet the data relationship pattern is not always linear. The nonlinearity of the data requires a suitable model to be used, the use of the restricted cubic spline (RCS) method in regression analysis modeling is very good, because the RCS method can follow the data patterns flexibly, and in the modeling process an optimum node point is used to produce a good model. The RCS method was used to estimate the relationship pattern of the case history of patients infected with the corona virus, then the regression model using the RCS method was compared with the linear regression model and the quadratic regression model, RCS 5 Nodes, RCS 7 Nodes. In this thesis, the best model was obtained, namely the regression model using the RCS 7 knots method, selecting the best model using MSE as the criteria for selecting the model.

Keywords: Linear, Regression, RCS.

Abstrak. Analisis regresi umumnya membutuhkan pola hubungan data yang linier, tetapi pola hubungan data pada kenyataannya tidak selalu mebentuk linier. Ketidaklineran data memerlukan model yang cocok untuk digunakan, penggunaan metode *restricted cubic spline* (RCS) dalam pemodelan analisis regresi sangat baik, karena metode RCS dapat mengikuti pola data dengan fleksibel, dalam proses pemodelan digunakan titik simpul optimum agar menghasilkan model yang baik. Metode RCS digunakan untuk mengestimasi pola hubungan kasus data riwayat pasien terinfeksi virus *corona*, kemudian model regresi dengan metode RCS dibandingkan dengan model regresi linier dan model regresi kuadratik, RCS 5 Simpul, RCS 7 Simpul. Dalam skripsi ini diperoleh model terbaik yaitu model regresi dengan menggunakan metode RCS 7 Simpul, pemilihan model terbaik menggunakan MSE sebagai kriiteria pemilihan model.

Kata Kunci: Linier, Regresi, RCS.

1. Pendahuluan

Analisis regresi merupakan suatu metode statistik yang paling sering digunakan dalam berbagai penelitian tujuan dari analisis regresi yaitu untuk memprediksi atau peramalan data, selain itu digunakan untuk mengetahui pola hubungan antara variabel respon dengan variabel prediktor.

Asumsi terhadap pola hubungan pada analisis regresi biasanya linier namun pada kenyataannya pola hubungan data tidak selalu membentuk linier dalam mengatasi pola hubungan yang tidak linier dapat digunakan metode spline, dimana metode ini merupakan fungsi polinom yang tergabung secara mulus (*smooth*) dan dibatasi oleh beberapa simpul. Ada

beberapa metode spline yang dapat digunakan dalam mengatasi pola hubungan yang tidak linier diantaranya yaitu, spline linier, *b*-spline, kubik spline, dan lain-lain.

Spline kubik mempunyai sifat yang sangat baik dlaam mengatasi ketidaklinieran dan kemampuan yang baik dalam menyesaikan bentuk kurva yang sangat melengkung, spline kubik dapat sangat mulus dengan menambah titik simpul pada fungsi. Namun spline kubik mempunyai sifat anomaly pada ekor kurva yaitu pada simpul pertama dan setelah simpul terakhir sehingga diperlukan metode yang dapat mengatasi perilaku tersebut yaitu dengan menggunakan metode *restricted cubic spline* (Stone and Koo, 1985)

Saat ini dunia sedang digemparkan oleh virus baru yaitu virus *corona*, virus ini pertama kali ditemukan di kota Wuhan China pada akhir desember 2019 hingga tanggal 16 januari 2021 sebanyak 94.414.806 orang diseluruh dunia terinfeksi virus *corona* dan menyebabkan kematian mencapai 2.020.119 orang. WHO menyebutkan bahwa virus sudah menyebar hingga 219 negara dan teritori dan resmi menyatakan bahwa penyakit yang disebabkan oleh *virus corona* sebagai *Coronavirus disease 2019* (Covid-19).

Untuk mengetahui model pendekatan analisis regresi yang sesuai dalam menggambarkan pola hubungan pada data kasus riwayat pasien terinfeksi virus corona maka akan dilakukan pemodelan regresi linier, kuadratik, dan menggunakan metode *restricted cubic spline*.

2. Landasan Teori

Restricted Cubic Spline

Spline adalah salah satu metode regresi dimana polynomial terbagi menjadi beberapa segmen dan dibatasi oleh titik simpul dan kontinu sehingga sifatnya lebih fleksibel dibandingkan dengan polynomial biasa. Titik simpul pada regresi spline merupakan kejadian terjadinya perubahan perilaku di suatu fungsi pada selang yang berbeda, spline dikatakan lebih fleksibel karena kurva dapat menyesuaikan terhadap perilaku data (Eubank, 1988).

Fungsi kubik spline mengikuti banyaknya simpul, apabila kubik spline memiliki *k* simpul, maka fungsi akan membutuhkan sebanyak k+3 koefisien regresi selain intersep. Secara umum fungsi kubik spline digambarkan pada persamaan berikut:

$$f(x) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 (X - k_1)_+^3 + \beta_5 (X - k_2)_+^3 + \cdots + \beta_m (X - k_m)_+^3$$

Fungsi kubik spline memiliki kelemahan yaitu perilaku yang buruk pada ekor simpul yaitu pada bagian sebelum simpul pertama dan setelah simpul terakhir sehingga perlu membatasi fungsi agar linier pada bagian ekor, maka digunakan metode *restricted cubic spine* (RCS), selain itu RCS hanya perlu mengestimasi beta sebanyak k-1 parameter selan intersep, dengan mengestimasi $\beta_0, \ldots, \beta_{k-1}$, model kubik spline yang digunakan yaitu sebagai berikut:

$$f(x) = \beta_0 + \beta_1 X + \beta_2 (X - t_1)_+^3 + \beta_3 (X - t_2)_+^3 + \cdots + \beta_{k+1} (X - t_k)_+^3$$

Dimana estimasi $\beta_0, \ldots, \beta_{k-1}$ dapat menggunakan matriks berikut:

$$\beta \quad = (B(\lambda)' \, B(\lambda)) \quad B(\lambda)' y$$

Matriks β didapatkan dengan mentransformasikan nilai x menggunakan perhitungan berikut dengan j = 1,2,…,k-2

$$X_{j+1} = \left(X - t_j\right)_+^3 - \frac{(X - t_{k-1})_+^3 \left(t_k - t_j\right)}{(t_k - t_{k-1})} + \frac{(X - t_k)_+^3 (t_{k-1} - t_j)}{(t_k - t_{k-1})}$$

Apabila nilai β telah diestimasi maka nilai dua β terakhir dapat dihitung menggunakan persamaan berikut:

$\beta_k = [\beta_2(t_1 - t_k) + \beta_3(t_2 - t_k) + \ldots + \beta_{k-1}(t_{k-2} - t_k)]/(t_k - t_{k-1})$

$\beta_{k+1} = [\beta_2(t_1 - t_{k-1}) + \beta_3(t_2 - t_{k-1}) + \ldots + \beta_{k-1}(t_{k-2} - t_{k-1})]/(t_{k-1} - t_k)$.

Pemilihan Titik Simpul

Penduga spline yang diperoleh dari hasil optimasi merupakan penduga linier, dimana penduga

spline bergantung pada parameter pemulus
λ dan harus memilih parameter pemulus yang optimal (Budiantara, 2004). Selain menggunakan parameter penghalus dapat menggunakan titik simpul yang mana cenderung lebih disukai dibandingkan dengan pemilihan parameter penghalus.

Harrel (2015) mengatakan bahwa pemilihan simpul bergantung pada banyaknya simpul k], banyaknya data menentukkan banyaknya titik simpul yang harus dipilih apabila data bersampel kecil atau < 30 maka dapat menggunakan $k \leq 3$, sedangkan apabila ukuran sampel besar atau > 30 maka dapat menggunakan k > 3[5].

Ada beberapa cara yang dapat digunakan dalam memilih titik simpul optimum diantaranya adalah menggunakan metode *Cross Generalized Validation* dan dapat menggunakan nilai kuantil data, pada penelitian ini akan digunakan nilai kuantil data sebagai berikut:

Tabel 1. Nilai Kuantil untuk Simpul

k	Kuantil						
3			0.1	0.5	0.9		
4			0.05	0.35	0.65	0.95	
5		0.05	0.275	0.5	0.725	0.95	
6	0.05	0.23	0.41	0.59	0.77	0.95	
7	0.025	0.1833	0.3417	0.5	0.6583	0.8167	0.975

Akaike Information Criterion (AIC)

Dalam memilih titik simpul optimum dapat digabandingkan dengan menggunkan nilai AIC, pemilihan simpul terbaik yaitu yang modelnya memiliki nilai AIC terkecil[6] berikut merupakan persamaan yang digunakan dalam menentukan titik simpul optimum.

$$AIC = LR\chi^2 - 2p$$

***Mean Square Error* (MSE)**

Mean Square Error (MSE) yaitu fungsi risiko yang merupakan perbedaan antara nilai dugaan dengan nilai sebenarnya, pada penelitian ini digunakan nilai MSE untuk menentukan model terbaik, perolehan nilai MSE berdasarkan persamaan berikut:

$$MSE = \frac{\sum_{i=y}^{n}\left(Y_i - \hat{Y}\right)^2}{db}$$

3. Hasil Penelitian dan Pembahasan

Bahan yang digunkaan merupakan data sekunder yang diperoleh dari *website www.kaggle.com* yaitu sebanyak 1500 data, variabel yang digunakan dalam kasus data pasien terinfeksi virus *corona* yaitu sebagai berikut:

Tabel 2. Variabel Data Riwayat Pasien Terinfeksi Virus Corona

No	Variabel	Deksripsi	Satuan
1	Y	Lama bertahan hidup	hari
2	X	Usia	tahun

Sumber: kaggle, 2020.

Karakteristik Lama Waktu Pasien Bertahan Hidup dan Usia Pasien

Berikut merupakan karakteristik dari lama waktu pasien bertahan hidup dan usia pasien yang akan disajikan pada Tabel 3.

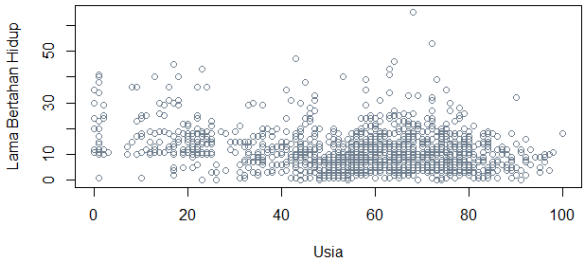
Tabel 3. Lama Waktu Bertahan Hidup Pasien Terinfeksi Virus Corona

	Lama Pasien Bertahan Hidup	Usia Pasien
Min	0	0
Max	65	100
Mean	12	53

Berdasarkan Tabel 3 secara rata-rata lama waktu pasien bertahan hidup yaitu 12 hari setelah terinfeksi virus corona dan lama waktu pasien bertahan hidup paling singkat yaitu nol hari artinya pasien meninggal dunia bersamaan dengan hari dimana pasien didiagnosis terinfeksi virus corona dan lama waktu pasien berahan hidup paling lama yaitu selama 65 hari dari hari pertama pasien didiagnosis terinfeksi virus corona.

Secara rata-rata usia pasien terinfeksi virus corona secara rata-rata berada pada usia 52 tahun, usia pasien paling muda yang terinfeksi virus corona yaitu nol tahun dan usia pasien paling tua yang terinfeksi virus corona yaitu 100 tahun.

Berikut merupakan *scatterplot* pola hubungan lama waktu pasien bertahan hidup dengan usia pasien.



Gambar 1. Gambar Scatterplot Lama Waktu Pasien Bertahan Hidup dengan Usia Pasien

Berdasarkan Gambar 1 terlihat bahwa pola hubungan data berpencar membentuk suatu fungsi yang menunjukkan bahwa pola hubungan data tidak sederhana, fungsi tersebut tidak dapat secara sederhana diwakili oleh suatu fungsi regresi parametrik seperti linier, kuadratik, dan lain-lain.

Regresi Linier Sederhana

Estimasi nilai β yang didapatkan adalah sebagai berikut:

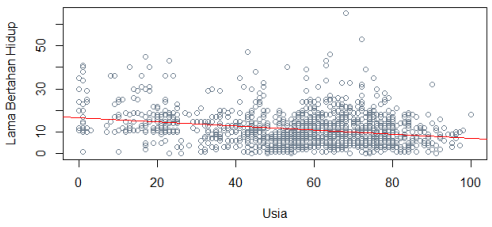
Tabel 4. Nilai Estimasi Parameter Regresi Linier

Variabel	Parameter	Estimasi Parameter
Konstan	β_0	16,8308
Usia Pasien	β_1	-0,0961

Sehingga diperoleh persamaan model regresi linier sederhana sebagai berikut:

$$\hat{Y} = 16{,}8308 - 0{,}0961X$$

dari model regresi linier di atas didapatkan nilai MSE = 7,453, dengan plot antara variabel prediktor usia pasien dan variabel respon lama waktu bertahan hidup berikut



Gambar 2. Pola Hubungan Estimasi Model Regresi Linier

Regresi Kuadratik

Estimasi nilai β yang didapatkan adalah sebagai berikut:

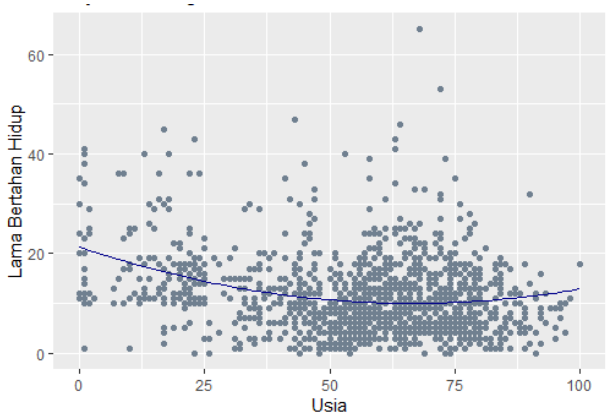
Tabel 5. Nilai Estimasi Parameter Model Regresi Kuadratik

Variabel	Parameter	Estimasi Parameter
Konstan	β_0	11,7440
Usia Pasien	β_1	-82,0379
	${\beta_1}^2$	51,6454

Sehingga model regresi kuadratik diperoleh sebagai berikut:

$$\hat{Y} = 11{,}7440 - 82{,}0379X + 51{,}6454X^2$$

dari model regresi kuadratik di atas diperoleh nilai MSE = 7,333 dengan plot hubungan antara variabel prediktor usia pasien dan variabel prediktor lama waktu pasien bertahan hidup berikut.



Gambar 3. Pola Hubungan Model Regresi Kuadratik

***Restricted Cubic Spline* 5 Simpul**

Nilai estimasi β yang diperoleh adalah berikut:

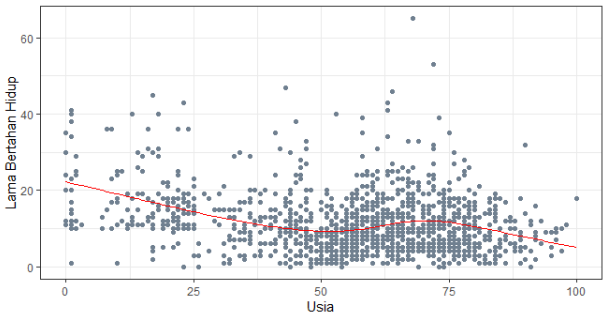
Tabel 6. Nilai Estimasi Parameter Model RCS 5 Simpul

Estimasi	Parameter	Estimasi Parameter
Konstan	$\hat{\beta}_0$	22,2312
Usia Pasien	$\hat{\beta}_1$	-0,3152
	$\hat{\beta}_2$	0,3258
	$\hat{\beta}_3$	0,9545
	$\hat{\beta}_4$	-6,9893
	$\hat{\beta}_5$	7,5129
	$\hat{\beta}_6$	-1,8039

Sehingga estimasi model *restricted cubic spline* didapatkan sebagai berikut:

$$\hat{Y} = 22{,}2312 - 0{,}3152X + 0{,}3258(X-16)_+^3 + 0{,}9545(X-43)_+^3$$
$$-6{,}9893(X-57)_+^3 + 7{,}5129(X-68)_+^3 - 1{,}8039(X-88)_+^3$$

Dari model di atas didapatkan nilai MSE = 7,206 dengan nilai AIC = 10194,0. Selanjutnya plot estimasi model regresi *restricted cubic spline* didapatkan hasil sebagai berikut:



Gambar 4. Pola Hubungan Model Regresi Restricted Cubic Spline Lima Simpul

***Restricted Cubic Spline* 7 Simpul**

Nilai estimasi β yang diperoleh adalah berikut:

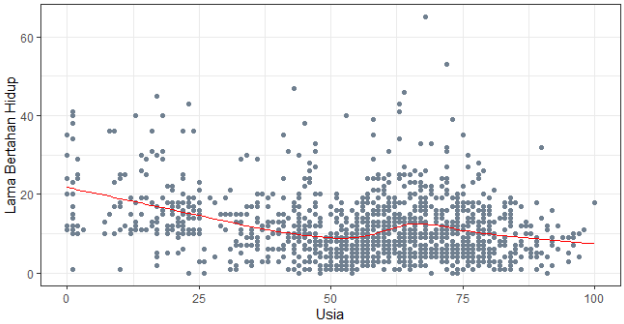
Tabel 7. Nilai Estimasi Parameter Model RCS 5 Simpul

Estimasi	Parameter	Estimasi Parameter	Parameter	Estimasi Parameter
Konstan	$\hat{\beta}_0$	21,7454	$\hat{\beta}_5$	-32,4924
Usia Pasien	$\hat{\beta}_1$	-0,2825	$\hat{\beta}_6$	40,1329
	$\hat{\beta}_2$	-0,0925	$\hat{\beta}_7$	-17,5906
	$\hat{\beta}_3$	0,7445	$\hat{\beta}_8$	2,4083
	$\hat{\beta}_4$	6,8916		

Sehingga estimasi model *restricted cubic spline* didapatkan sebagai berikut:

$$\hat{Y} = 21{,}7454 - 0{,}2825X - 0{,}0926(X-9)_+^3 + 0{,}7445(X-23)_+^3 + 6{,}8916(X-46)_+^3 - 32{,}3924(X-57)_+^3 + 40{,}1329(X-65)_+^3 - 17{,}5906(X-74)_+^3 + 2{,}4083(X-88)_+^3$$

Dari model di atas didapatkan nilai MSE = 7,19 dengan nilai AIC = 10191,0 Selanjutnya plot estimasi model regresi *restricted cubic spline* didapatkan hasil sebagai berikut:



Gambar 5. Pola Hubungan Model Regresi Restricted Cubic Spline Tujuh Simpul

Pemilihan Titik Simpul Optimum

Setelah dilakukan pemodelan regresi dengan menggunakan *restricted cubic spline* pemilihan titik simpul dengan lima simpul dan tujuh simpul dilakukan pemilihan pemilihan model dengan titik simpul optimal, pemilihan titik simpul optimal dilakukan dengan membandingkan nilai *Akaike's Information Criterion* (AIC) yang paling minimum. Nilai AIC paling minimum akan disajikan pada Tabel 8

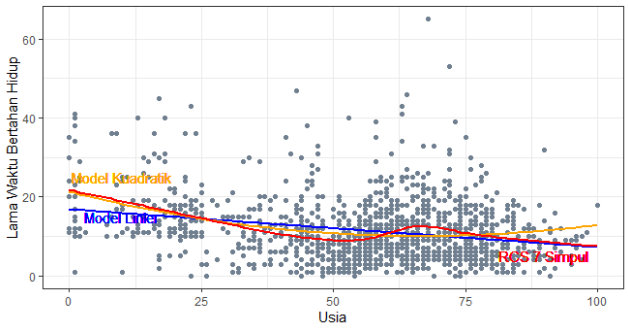
Tabel 8. Nilai AIC

Model	AIC
RCS 5 Simpul	10194,0
RCS 7 Simpul	10191,0

Berdasarkan Tabel 3.7 Diketahui bahwa model dengan titik simpul optimum dengan nila AIC paling minimum merupakan model dengan menggunakan *restricted cubic spline* 7 simpul. Maka, model yang dipilih adalah model dengan tujuh titik simpul.

Perbandingan Model

Setelah melakukan pemodelan regresi linier, regresi kuadratik, dan regresi dengan menggunakan metode *restricted cubic spline* selanjutnya akan dilakukan perbandingan model untuk mengetahui model terbaik dari hasil analisis secara keseluruhan dapat dilihat pada gambar berikut:



Gambar 6. Perbandingan Model Regresi Linier, Kuadratik, RCS 7 Simpul

Berdasarkan gambar di atas dapat dilihat bahwa estimasi model regresi dengan menggunakan *restricted cubic spline* memberikan hasil yang sangat baik dibandingkan dengan model lainnya, fungsi lebih fleksibel dan mendekati pola hubungan data. Hal tersebut dapat diperkuat dengan membandingkan nilai MSE pada masing-masing model yang akan disajikan pada tabel berikut:

Tabel 9. Nilai MSE

No	Model	MSE
1	Linier	7,453
2	Kuadratik	7,333
3	RCS 7 Simpul	7,190

Model regresi yang digunakan berdasarkan nilai MSE paling minimum yaitu model regresi *restricted cubic spline* dengan tujuh simpul.

4. Kesimpulan

Berdasarkan hasil analisis dan pembahasan didapatkan kesimpulan sebagai berikut:

Rata-rata usia pasien yang terinfeksi virus corona berada pada usia 53 tahun dengan usia pasien paling muda yaitu nol tahun dan usia pasien paling tua yaitu 100 tahun sedangkan rata-rata lama waktu bertahan hidup pasien yang terinfeksi virus corona yaitu 12 hari dengan lama waktu bertahan hidup paling sedikit yaitu nol hari dan lama waktu bertahan hidup paling lama yaitu 65 hari.

Model regresi terbaik pada pemodelan pola hubungan kasus data riwayat pasien terinfeksi virus corona yaitu menggunakan metode *restricted cubic spline* dengan 7 titik simpul, diperoleh nilai AIC sebesar 10191,0 dan nilai MSE sebesar 7,19. Hal ini membuktikan bahwa metode *restricted cubic spline* mempunyai fleksibilitas yang tinggi, garis kurva dapat mengikuti pola data dengan baik. Model yang diperoleh oleh metode *restricted cubic spline* adalah sebagai berikut.

$$\hat{Y} = 21{,}7454 - 0{,}2825X - 0{,}0926(X-9)_+^3 + 0{,}7445(X-23)_+^3 + 6{,}8916(X-46)_+^3 - 32{,}3924(X-57)_+^3 + 40{,}1329(X-65)_+^3 - 17{,}5906(X-74)_+^3 + 2{,}4083(X-88)^3$$

5. Saran

Berdasarkan hasil analisis dan pembahasan yang telah dijelaskan adapun saran dari peneliti yaitu pada penelitian selanjutnya untuk mengetahui pola hubungan data dengan menggunaan metode *restricted cubic spline* dapat digunakan dengan menambahkan variabel lain untuk digunakan sebagai prediktor sehingga dapat digunakan pada model regresi multivariat, selain itu dapat digunakan juga model regresi logistik.

Daftar Pustaka

[1]Budiantara, I.N. 2004. *Spline : Historis, Motivasi dan Perannya dalam Regresi Nonparametrik*, Makalah Pembicara Utama pada Konferensi Nasional Matematika XII. Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Udayana, Denpasar-Bali.

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

[3] Harrell Jr, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

[4] Stone, C. J., & Koo, C. Y. (1985). *Additive splines in statistics. Proceedings of the American Statistical Association. Original pagination is p*, 45, 48.