

ESTIMASI MODEL REGRESI SEMIPARAMETRIK MENGGUNAKAN ESTIMATOR KERNEL UNIFORM (Studi Kasus: Pasien DBD di RS Puri Raharja)

Anna Fitriani^{§1}, I Gusti Ayu Made Srinadi², Made Susilawati³

¹Jurusan Matematika, Fakultas MIPA, Universitas Udayana [Email: annafitrianipangestu@gmail.com]

²Jurusan Matematika, Fakultas MIPA, Universitas Udayana [Email: srinadiigustiayumade@yahoo.co.id]

³Jurusan Matematika, Fakultas MIPA, Universitas Udayana [Email: susilawati.made@gmail.com]

[§]*Corresponding Author*

ABSTRACT

Semiparametric regression model approach is a model approach that combines parametric regression models and nonparametric regression. On semiparametric regression, most explanatory variables are parametric and nonparametric others are. Independent variables that satisfy parametric assumptions can be predicted by linear regression analysis method, whereas that does not meet the parametric assumptions alleged by the method nonparametrik. Teknik smoothing (smoothing) nonparametric regression curve on the components used in this study using uniform kernel function. Estimation of optimal semiparametric regression curve is determined by the size of the weight or bandwidth (h) is optimal. Selection of the optimal bandwidth will produce a smooth regression curve estimation in accordance with the pattern data. Selection of the optimum bandwidth is determined based on the criteria that the minimum value of GCV. The purpose of this study was to determine the estimated regression function semiparametric dengue cases using kernel estimators uniform. The response of the data used is old data recovery of patients with Dengue Hemorrhagic Fever (DHF). There are six independent variables such as age (years), body temperature ($^{\circ}\text{C}$), pulse (beats / min), hematocrit (%), platelets ($\times 10^3/\mu\text{l}$), and duration of fever (day). Age, body temperature, pulse, platelets, and duration of fever is a component of parametric and nonparametric hematocrit is a component. Bandwidth (h) the optimal minimum GCV obtained based on the criteria of 0,005. MSE value is generated using multiple linear regression analysis of 0,031. While the semiparametric regression of 0,00437119.

Keywords: Semiparametric Regression, Kernel, Bandwidth, GCV

1. PENDAHULUAN

Analisis regresi digunakan untuk melihat pengaruh variabel bebas (prediktor) terhadap variabel terikat (respon) dengan terlebih dahulu melihat pola hubungan variabel tersebut. Pendekatan model regresi semiparametrik merupakan pendekatan model yang mengkombinasikan model regresi parametrik dan regresi nonparametrik. Pada regresi semiparametrik, sebagian variabel penjelasnya bersifat parametrik dan sebagian lain bersifat nonparametrik. Regresi semiparametrik digunakan jika pola hubungan antara sekumpulan variabel bebas dan variabel terikat

diketahui dan ada pula yang tidak diketahui. Diberikan model regresi semiparametrik:

$$Y_i = X_i^T \gamma + m(t_i) + \varepsilon_i ; i = 1, 2, \dots, n \quad (1)$$

y_i adalah variabel respon ke-i, X_i adalah komponen parametrik, $m(t_i)$ adalah fungsi regresi yang tidak diketahui, dan ε_i adalah galat acak (*random error*), dimana $\varepsilon_i \sim N(0, \sigma^2)$. Terdapat beberapa teknik *smoothing* dalam model regresi nonparametrik salah satunya adalah kernel. Penduga kernel fleksibel dan secara matematik mudah diselesaikan (Härdle [2]). Pada penduga kernel yang terpenting adalah pemilihan parameter pemulus (*bandwidth*) yang optimal untuk mendapatkan

kurva regresi yang optimal. Kasus yang dapat menggunakan regresi semiparametrik salah satunya adalah menduga model regresi dengan respon lama kesembuhan pasien Demam Berdarah Dengue (DBD). Lama kesembuhan pasien ada kemungkinan dipengaruhi oleh nyeri demam, suhu tubuh dan tekanan darah, dehidrasi, penurunan kadar trombosit, perdarahan, *shock*. Sebaran data dari variabel yang diketahui bentuk kurva regresinya diduga menggunakan regresi parametrik, sedangkan peubah yang tidak diketahui bentuk kurva regresinya dan tidak ingin terikat dengan asumsi tertentu, dapat diduga dengan regresi nonparametrik. Berdasarkan uraian tersebut, peneliti tertarik untuk menentukan model regresi semiparametrik menggunakan penduga kernel *uniform*.

2. TINJAUAN PUSTAKA

2.1 Regresi Parametrik

Regresi parametrik merupakan metode statistika yang digunakan untuk mengetahui pola hubungan antara variabel prediktor dengan variabel respon, dengan asumsi bahwa telah diketahui bentuk fungsi regresinya. Bentuk umum regresi linier ditulis sebagai berikut

$$y_i = \gamma_0 + \gamma_j X_i + \varepsilon_i, i, j = 1, 2, \dots, n \quad (2)$$

2.2 Regresi Nonparametrik

Untuk n pengamatan yang independen, (t_i, y_i) dimana $i = 1, 2, \dots, n$ maka model regresi secara umum dapat ditulis dengan:

$$y_i = m(t_i) + \varepsilon_i, i = 1, 2, \dots, n \quad (3)$$

y_i adalah variabel respon ke- i , $m(t_i)$ adalah fungsi regresi yang tidak diketahui bentuk kurva regresinya dan ε_i adalah *error random* atau galat acak yang diasumsikan independen dan identik dengan rataan 0 dan keragaman σ^2 .

2.3 Regresi Semiparametrik

Model regresi semiparametrik dapat ditulis sebagai berikut:

$$y_i = X_i^T \gamma + m(t_i) + \varepsilon_i, i = 1, 2, \dots, n \quad (4)$$

y_i adalah variabel respon ke- i , X_i adalah komponen parametrik, $m(t_i)$ adalah fungsi regresi yang tidak diketahui bentuk kurva regresinya dan ε_i adalah galat acak dengan $\varepsilon_i \sim N(0, \sigma^2)$.

2.4 Regresi Nonparametrik Kernel

Regresi kernel adalah teknik statistika nonparametrik untuk menduga fungsi regresi $m(t_i)$ pada model regresi nonparametrik $y_i = m(t_i) + \varepsilon_i$ dengan $i = 1, 2, \dots, n$. Penduga fungsi regresi semiparametrik adalah sebagai berikut:

$$\hat{m}(t) = \sum_{i=1}^n w_{hi}(t) y_i \quad (7)$$

dengan

$$w_{hi}(t) = \frac{K_h(t - t_i)}{\sum_{i=1}^n K_h(t - t_i)}$$

$$\begin{aligned} &= \frac{\frac{1}{h} K\left(\frac{t - t_i}{h}\right)}{\frac{1}{h} \sum_{i=1}^n K\left(\frac{t - t_i}{h}\right)} \\ &= \frac{K\left(\frac{t - t_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{t - t_i}{h}\right)} \end{aligned} \quad (8)$$

Penduga (8) diusulkan oleh Nadaraya dan Watson, sehingga penduga ini sering disebut penduga Nadaraya-Watson (Härdle [2]). Pada regresi kernel, ukuran penduganya ditentukan oleh *bandwidth* (h).

2.5 Pemilihan Bandwidth Optimal

Bandwidth (h) adalah parameter pemulus (*smoothing*) yang berfungsi untuk mengontrol kemulusan dari kurva yang diestimasi. Proses pemilihan *bandwidth* yang sesuai (parameter *smoothing*) adalah bagian yang penting dari regresi nonparametrik. Telah diketahui secara umum, bahwa permasalahan utama pada kernel *smoothing* bukan terletak pada pemilihan kernel tetapi pada pemilihan *bandwidth* (Hastie dan Tibshirani [3]). *Bandwidth* yang terlalu kecil akan menghasilkan kurva yang *under-smoothing* yaitu sangat kasar dan fluktuatif. Sebaliknya, *bandwidth* yang terlalu lebar akan menghasilkan kurva yang *over-smoothing* yaitu sangat mulus. (Härdle [2]). Oleh karena itu

perlu dipilih *bandwidth* yang optimal untuk menghasilkan kurva optimal. Salah satu metode untuk mendapatkan h optimal adalah dengan menggunakan kriteria *Generalized Cross Validation* (GCV) (Eubank [1]), yang didefinisikan sebagai berikut:

$$GCV(h) = \frac{MSE}{\left(\frac{1}{n} \text{tr}(I - H(h))\right)^2} \quad (9)$$

dengan:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - m_h(t_i))^2 \quad (10)$$

2.6 Koefisien Determinasi

Koefisien determinasi (R^2) merupakan besaran yang digunakan untuk mengukur kelayakan model regresi dan menunjukkan besar kontribusi X terhadap perubahan Y . Semakin tinggi nilai R^2 semakin baik model regresi yang terbentuk:

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{SSR}{SST} \quad (11)$$

Nilai R^2 terletak antara 0 dan 1. Model dikatakan lebih baik jika R^2 semakin mendekati nilai 1.

2.7 Demam Berdarah Dengue

Demam berdarah dengue merupakan penyakit yang disebabkan oleh virus dengue. Pada penderita demam berdarah dengue biasanya ditemukan perdarahan pada kulit, perdarahan dari gusi, hidung, usus, dan lain lain. Bila tidak ditangani segera, demam berdarah dengue dapat menyebabkan kematian. Selain menyebabkan demam berdarah dengue, inveksi virus dengue juga menyebabkan demam dengue. Setelah tergigit nyamuk pembawa virus, maka inkubasi akan berlangsung antara 3 sampai 15 hari sampai gejala demam dengue muncul. Adapun indikasi atau gejala demam berdarah adalah nyeri demam, suhu tubuh dan tekanan darah, dehidrasi, penurunan kadar trombosit, perdarahan, *shock*.

3. METODE PENELITIAN

Data yang digunakan dalam penelitian ini adalah data sekunder yang diambil di Rumah Sakit Puri Raharja Denpasar Bali. Populasi dari penelitian ini adalah pasien DBD yang pernah menjalani rawat inap di Rumah Sakit Puri Raharja. Sampel dari penelitian ini berasal dari data rekam medis pasien DBD periode bulan Januari sampai bulan Maret 2015. Peubah respons (Y) yaitu lama kesembuhan pasien DBD (hari) dan peubah bebas (X_i) yaitu umur (U), suhu tubuh (S), nadi (N), trombosit (PLT), lama demam (LD) serta kadar hematokrit (HCT).

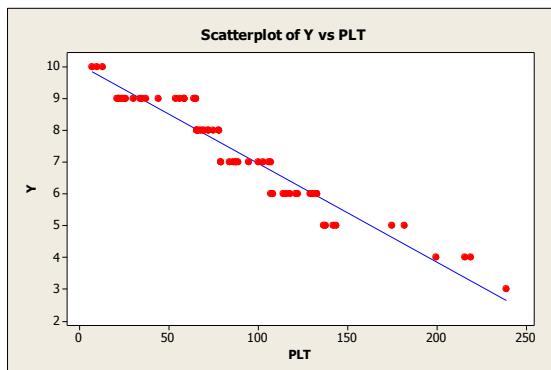
- a. Estimasi kurva regresi dengan pendekatan estimator kernel
 - Mendefinisikan penduga \hat{y} untuk y dan \hat{m} untuk m pada model regresi semiparametrik (4)
 - Estimator $\hat{m}(t_i)$ dicari dengan menggunakan persamaan (7) dan diperoleh model estimasi.
 - Estimator \hat{y} dicari dengan menggunakan meminimumkan kuadrat galat dan diperoleh model estimasi.
- b. Penentuan parameter pemulus menggunakan fungsi kernel uniform dan ditentukan dengan GCV pada persamaan (9)
- c. Menghitung penduga y_i .
- d. Menerapkan pada data sekunder.
- e. Interpretasi model yang diperoleh.

4. HASIL DAN PEMBAHASAN

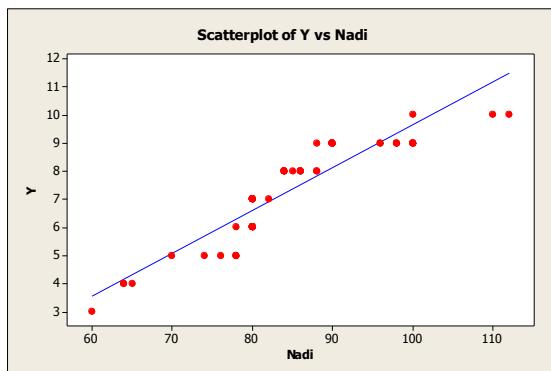
Data yang digunakan dalam penelitian ini adalah data sekunder yang diambil di Rumah Sakit Puri Raharja Denpasar Bali. Populasi dari penelitian ini adalah pasien DBD yang pernah menjalani rawat inap di Rumah Sakit Puri Raharja. Sampel dari penelitian ini berasal dari data rekam medis pasien DBD periode bulan Januari sampai bulan Maret 2015. Peubah respons (Y) yaitu lama kesembuhan pasien DBD (hari) dan peubah bebas (X_i) yaitu umur (U), suhu tubuh (S), nadi (N), trombosit (PLT), lama demam (LD) serta kadar hematokrit (HCT).

4.1 Penentuan Komponen Parametrik dan Komponen Nonparametrik

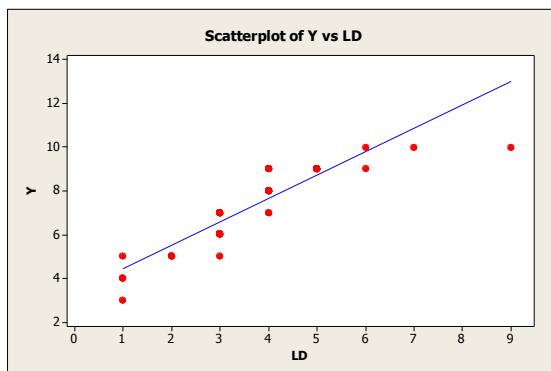
Apabila plot antara variabel bebas dengan variabel respon memiliki hubungan linear, maka variabel bebas tersebut merupakan komponen parametrik. Namun, apabila plot antara variabel bebas dengan variabel respon tersebut memiliki hubungan nonlinier, dan sulit untuk menduga bentuk kurva regresinya, variabel bebas tersebut dipilih sebagai komponen nonparametriknya.



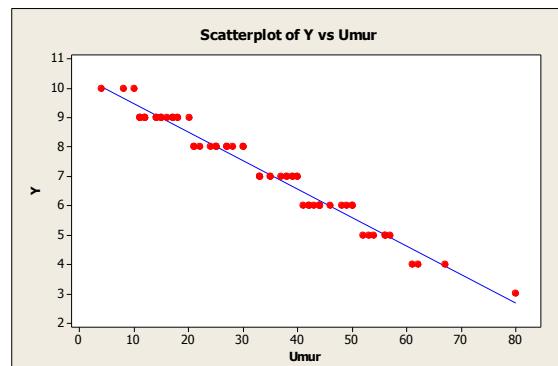
Gambar 4.1 Scatter plot Trombosit (PLT) dengan Lama kesembuhan pasien (Y)



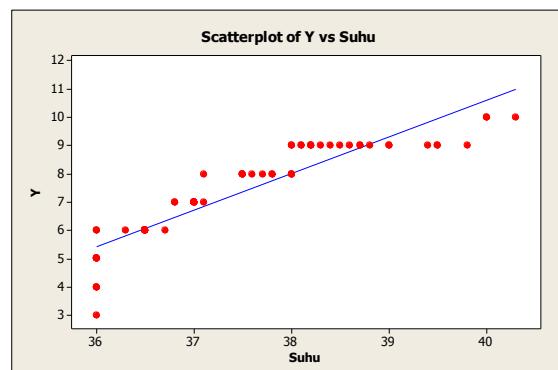
Gambar 4.2 Scatter plot antara Nadi (N) dengan Lama kesembuhan pasien (Y)



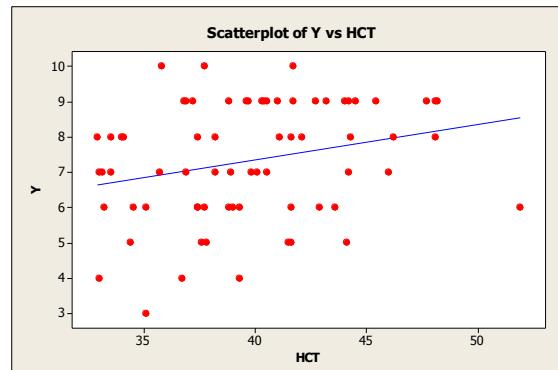
Gambar 4.3. Scatter plot antara Lama demam (LD) dengan Lama kesembuhan pasien (Y)



Gambar 4.4. Scatter plot antara Umur (U) dengan Lama kesembuhan pasien (Y)



Gambar 4.5. Scatter plot antara Suhu (S) dengan Lama kesembuhan pasien (Y)



Gambar 4.6. Scatter plot antara Kadar hematokrit (HCT) dengan Lama kesembuhan pasien (Y)

4.2 Penentuan Bandwidth h Optimal

Pemilihan *bandwidth* yang optimal ditentukan berdasarkan kriteria nilai GCV yang minimum. *Bandwidth* optimal pada penelitian diperoleh sebesar 0,005.

5. KESIMPULAN

Pada regresi semiparametrik dengan *estimator kernel uniform* tidak diperoleh model estimasi secara eksplisit seperti regresi lainnya melainkan estimasi dari titik-titik amatan. Berdasarkan kriteria nilai MSE, regresi semiparametrik lebih bagus dibandingkan dengan analisis regresi linear berganda. Nilai MSE yang dihasilkan dengan regresi semiparametrik sebesar 0.00437119 sedangkan nilai MSE yang dihasilkan dengan analisis regresi linear berganda sebesar 0.031. Variabel yang signifikan dalam model regresi ini adalah suhu, umur, dan PLT sehingga dapat dikatakan bahwa dengan tingkat kepercayaan 1% variabel yang berpengaruh terhadap lama kesembuhan pasien adalah suhu, umur dan PLT. Diperoleh model regresi semiparametrik sebagai berikut:

$$\hat{y}_i = -0.005495777 + 0.265938992 S - 0.472707138 U - 0.330655080 PLT + \hat{m}_h(t_i)$$

DAFTAR PUSTAKA

- [1] Eubank, R. 1988. *Spline Smoothing and Nonparametric Regression*. Marcel Dekker. New York.
- [2] Härdle, W. 1994. *Applied Nonparametric Regression*. Cambridge University Press. New York.
- [3] Hastie, T.J. and R.J. Tibshirani. 1990. *Generalized Additive Models*. Chapman and Hall. New York. London