# II. TINJAUAN PUSTAKA

## 2.1 Analisis Regresi

Analisis regresi adalah suatu metode analisis data yang menggambarkan hubungan antara variabel respon dengan satu atau beberapa variabel prediktor. Misalkan X adalah variabel prediktor dan Y adalah variabel respon untuk n data pengamatan berpasangan $\{(xi, yi)\}_{i=1}^{n}$, maka hubungan antara variabel prediktor dan variabel respon tersebut dapat dinyatakan sebagai berikut:

$y_i = f(x_i) + \varepsilon_i$ ; i=1,2,3,…,n (2.1)

Dengan $\varepsilon_i$ adalah galat yang diasumsikan independen dengan mean 0 dan variansi $\sigma^2$ (konstan). $f(x_i)$ disebut sebagai fungsi regresi atau kurva regresi (Hardle,1994).

## 2.2 Regresi Nonparametrik

Menurut Eubank (1998), regresi nonparametrik merupakan pendekatan metode regresi dimana bentuk kurva dari fungsi regresinya tidak diketahui. Kurva fungsi diasumsikan termuat dalam ruang fungsi tertentu. Model regresi nonparametrik adalah sebagai berikut:

$y_i = m(x_i) + \varepsilon_i$ ; $i$=1,2,3,…, n (2.2)

$m(x_i)$ merupakan kurva fungsi regresi yang tidak diketahui bentuknya dengan $x_i$ merupakan variabel independen. $\varepsilon_i$ adalah galat yang diasumsikan independen dengan mean 0 dan variansi $\sigma^2$ (konstan).

Estimasi fungsi regresi nonparametrik dilakukan berdasarkan data pengamatan dengan menggunakan teknik *smoothing*. Terdapat beberapa teknik *smoothing* dalam model regresi nonparametrik antara lain penduga kernel, deret orthogonal, penduga spline, deret fourier, dan wavelet (Eubank, 1998).

## 2.3 Penduga Densitas Kernel

Menurut Hardle (1994), Penduga densitas kernel merupakan pengembangan dari estimator histogram. Penduga kernel diperkenalkan oleh Rosenblatt (1956) dan Parzen (1962) sehingga disebut penduga densitas kernel Rosenblatt-Parzen. Secara umum kernel K dengan parameter pemulus (*bandwidth*) h didefinisikan sebagai:

$K_h(x) = \frac{1}{h} K(\frac{x}{h})$ untuk $-\infty < x < \infty$ dan h>0 (2.3)

Serta memenuhi :

(i) K(x) $\geq$ 0, untuk semua x

(ii) $\int_{-\infty}^{\infty} K(x)dx = 1$

(iii) $\int_{-\infty}^{\infty} x^2 K(x)dx = \sigma^2 > 0$

(iv) $\int_{-\infty}^{\infty} xK(x)dx = 0$

Beberapa jenis fungsi kernel antara lain:

1. Kernel Uniform : K(x) =$\frac{1}{2}$ ;|x| ≤1 , 0 selainnya

2. Kernel Triangle : K(x) = (1 - |x|) ;|x| ≤ 1 , 0 selainnya

3. Kernel Epanechnikov : K(x) =$\frac{3}{4}$(1 - $x^2$) ;|x| ≤ 1 , 0 selainnya

4. Kernel Kuartik : K(x) =$\frac{15}{16}(1 - x^2)^2$ ;|x| ≤ 1 , 0 selainnya

5. Kernel Triweight : K(x) =$\frac{35}{32}$(1 - $x^2$)$^3$ ;|x| ≤ 1 , 0 selainnya

6. Kernel Cosinus : K(x) =$\frac{\pi}{4}\cos\left(\frac{\pi}{2}x\right)$ ;|x| ≤ 1 , 0 selainnya

7. Kernel Gaussian : K(x) =$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$ $-\infty < x < \infty$

Estimator densitas kernel dari untuk fungsi densitas f(x) didefinisikan sebagai:

$$f_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \quad (2.4)$$

Dari persamaan (3) terlihat bahwa $f_h(x)$ tergantung pada fungsi kernel K dan parameter h. Bentuk bobot kernel ditentukan oleh fungsi kernel K, sedangkan ukuran bobotnya ditentukan oleh parameter pemulus h yang disebut bandwidth.

## 2.4 Metode Nadaraya-Watson

Menurut Hardle (1991), jika terdapat n data pengamatan $\{(X_i, Y_i)\}_{i=1}^{n}$ yang memenuhi persamaan (2) dimana $X_i \in R$ dan $Y_i \in R$, maka penduga m(x) adalah:

$$\hat{m}(x) = E(Y|X = x) = \int \frac{yf(X,Y)}{f(X=x)}dy \quad (2.5)$$

Penyebut diduga dengan menggunakan penduga densitas kernel

$$f_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)$$

Fungsi densitas peluang bersama diduga dengan perkalian kernel, yaitu :

$\acute{f}_{h1,h2}(x,y) = \frac{1}{n}\sum_{i=1}^{n} K_{h1}(x - X_i)K_{h2}(y - Y_i)$

Sehingga, pembilang dari penduga nadaraya menjadi :

$$\begin{aligned}\int y\hat{f}_{h_1,h_2}(x,y)dy &= n^{-1}\sum_{i=1}^{n} K_{h_1}(x - X_i)\int yK_{h_2}(y - Y_i)dy \\ &= n^{-1}\sum_{i=1}^{n} K_{h_1}(x - X_i)\int \frac{y}{h_2}K\left(\frac{y - Y_i}{h_2}\right)dy \\ &= n^{-1}\sum_{i=1}^{n} K_{h_1}(x - X_i)\int (sh_2 + Y_i)K(s)\,ds \\ &= n^{-1}\sum_{i=1}^{n} K_{h_1}(x - X_i)Y_i\end{aligned}$$

Bentuk penduga Nadaraya-Watson dapat ditulis :

$$\hat{m}(x_i) = \frac{\frac{1}{n}\sum_{j=1}^{n} K_h(x_i - X_j)Y_j}{\frac{1}{n}\sum_{k=1}^{n} K_h(x_i - X_k)}$$

$$\hat{m}(x_i) = \frac{\frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x_i - X_j}{h}\right)Y_j}{\frac{1}{nh}\sum_{k=1}^{n} K\left(\frac{x_i - X_k}{h}\right)}$$

$$\hat{m}(x_i) = \frac{\sum_{j=1}^{n} K\left(\frac{x_i - X_j}{h}\right)}{\sum_{k=1}^{n} K\left(\frac{x_i - X_k}{h}\right)}Y_j \qquad (2.6)$$

$\hat{m}(x_i) = \sum_{j=1}^{n} W_{ij}(x_i)Y_j$

Sehingga, $\hat{Y} = WY$, dimana

$$W_{ij} = \frac{K\left(\frac{x_i - X_j}{h}\right)}{\sum_{k=1}^{n} K\left(\frac{x_i - X_k}{h}\right)} \qquad (2.7)$$

Matriks W disebut juga dengan *Hat Matrix* dari penduga m(x). Persamaan (2.6) ditemukan oleh Nadaraya dan Watson (1964), sehingga disebut estimator Nadaraya-Watson.

Pengaruh fungsi kernel kurang signifkan dibandingkan dengan pengaruh *bandwidth* h. Nilai-nilai ekstrim dari h mengakibatkan :

- Jika $h \to 0$, maka untuk x=xi, $m(x_i) \to \frac{K(0)}{K(0)} Y_i = Y_i$

Jadi *bandwidth* h sangat kecil, estimator akan menuju ke data

- Jika $h \to \infty$ maka $K(\frac{x-X_i}{h}) \to K(0)$, akibatnya

$m(x_i) \to \frac{n^{-1}\sum_{i=1}^{n} K(0)Y_i}{n^{-1}\sum_{i=1}^{n} K(0)} = \frac{n^{-1}K(0)\sum_{i=1}^{n} Y_i}{n^{-1}(nK(0))} = n^{-1}\sum_{i=1}^{n} Y_i$

Jadi *bandwidth* (h) sangat besar, estimator akan sangat mulus dan menuju rata-rata dari variabel respon.

Semakin kecil nilai *bandwidth* h, maka grafik akan semakin kurang mulus namun memiliki bias yang kecil. Sebaliknya semakin besar nilai *bandwidth* h, maka grafik akan sangat mulus tetapi memiliki bias yang besar. Karena tujuan estimasi kernel adalah memperoleh kurva yang mulus namun memiliki nilai MSE yang tidak terlalu besar, perlu dipilih nilai h optimal untuk mendapatkan grafik optimal. Salah satu cara memilih parameter pemulus optimal adalah dengan menggunakan metode *Generalized Cross Validation (GCV)*.

## 2.5 Pemilihan Bandwidth h optimal

Menurut Hardle (1991), *Bandwidth* h adalah parameter pemulus yang berfungsi untuk mengontrol kemulusan dari kurva yang diestimasi. *Bandwidth* yang terlalu kecil akan menghasilkan kurva yang *under-smoothing* yaitu sangat kasar dan sangat fluktuatif, dan sebaliknya *bandwidth* yang terlalu lebar akan menghasilkan kurva yang *over-smoothing* yaitu sangat mulus, tetapi tidak sesuai dengan pola data.

Oleh karena itu perlu dipilih *bandwidth* yang optimal. Metode untuk mendapatkan h optimal dapat diperoleh dengan menggunakan kriteria *Generalized Cross Validation (GCV),* yang didefinisikan sebagai berikut:

$$GCV(h) = \frac{MSE(h)}{[1-\frac{trace(W)}{n}]^2} \quad (2.8)$$

Dengan MSE (h) $= n^{-1}\sum_{i=1}^{n}(y_i - m(x_i))^2$dan $W$ adalah hat matriks berukuran nxn yang memenuhi $[m_h(x_1), m_h(x_2), \dots, m_h(x_n)]^t = WY$

Nilai *bandwidth* h optimal akan diperoleh jika nilai akan menghasilkan nilai *Generalized Cross Validation* minimal (Craven dan Wahba, 1979).

.

## 2.6 Fungsi Periodik

Menurut Tolstov (1962), suatu fungsi f(x) dikatakan periodik jika terdapat konstanta T>0, sehingga memenuhi f(x+T)=f(x) untuk setiap x anggota domain f(x). Selanjutnya T disebut dengan periode dari fungsi f(x). Jika T adalah periode dari suatu fungsi f(x), maka ...,-2T, -T,2T,3T ... juga merupakan periode dari fungsi f(x).

Salah satu contoh fungsi periodik adalah f(x)=sin (x) dengan periode $2\pi$, karena sin(x+$2\pi$) = sin(x).

## 2.7 Deret Fourier

Menurut Tolstov (1962), jika fungsi f(x) terdefinisi pada interval $[-L, L]$ dan diluar selang ini oleh $f(x \pm 2L) = f(x)$, sehingga f(x) merupakan fungsi

periodik dengan periode 2L. $f(x)$ dapat direpresentasikan dengan deret perluasan fourier sebagai berikut :

$$f(x) = \frac{1}{2}a_0 + \sum_{j=1}^{\infty} a_j \cos\left(\frac{2\pi jx}{2L}\right) + b_j \sin\left(\frac{2\pi jx}{2L}\right) \quad (2.9)$$

$$f(x) = \frac{1}{2}a_0 + \sum_{j=1}^{\infty} a_j \cos\left(\frac{\pi jx}{L}\right) + b_j \sin\left(\frac{\pi jx}{L}\right)$$

dengan

$$a_j = \frac{1}{L}\int_{-L}^{L} f(x) cos\left(\frac{j\pi x}{L}\right) dx$$

$$b_j = \frac{1}{L}\int_{-L}^{L} f(x) sin\left(\frac{j\pi x}{L}\right) dx$$

$j = 1,2,3, \ldots.$

Nilai $\frac{2\pi}{T}$ (dengan T adalah periode f(x)) merupakan faktor pengali agar x dalam satuan radian.

## 2.8 Estimator Fourier

Diberikan n data pengamatan $\{(x_i, y_i)\}_{i=1}^{n}$ yang memenuhi persamaan (2). Jika $X_i \in [-L, L]$ dan $Y_i \in R$, dan diasumsikan periode m(x) adalah $T = 2L$, maka penduga m(x) dapat didekati oleh deret fourier yang didefinisikan sebagai berikut:

$$\widehat{m}(x) = \frac{1}{2}a_0 + \sum_{j=1}^{J} a_j \cos\left(\frac{2\pi jx}{2L}\right) + b_j \sin\left(\frac{2\pi jx}{2L}\right) \quad (2.10)$$

Dengan $a_0$ , $a_j$ dan $b_j$ adalah koefisien Fourier (Bowman dan Azzalini, 1997).

Tingkat kemulusan estimator deret Fourier ditentukan oleh pemilihan parameter pemulus J. Semakin kecil parameter pemulus *J*, semakin mulus estimasinya dan

semakin besar parameter pemulus *J*, semakin kurang mulus estimasi dari *f*. Oleh karena itu, perlu dipilih J yang optimal.

## 2.9 Pemilihan Parameter Pemulus (J) Optimal

Pada pemodelan regresi nonparametrik dengan menggunakan deret Fourier, hal yang perlu diperhatikan adalah menentukan nilai J. Salah satu metode yang dapat digunakan adalah metode *Generalized Cross Validation (GCV).* Penentuan J optimal akan menghasilkan nilai koefisien determinasi ($R^2$) yang tinggi.

*Generalized Cross Validation(GCV)* didefiniskan sebagai berikut:

$$GCV(J) = \frac{MSE(J)}{[1-(\frac{trace(A_J)}{n})]^2} \tag{2.11}$$

dengan MSE(J) $= n^{-1}\sum_{i=1}^{n}(y_i - \widehat{m(x_i)})^2$ dan $A_J$ adalah matriks berukuran nxn yang memenuhi $\widehat{m}(X) = A_J Y$ dan disebut juga *Hat Matrixs.* Nilai GCV terkecil akan menghasilkan nilai J yang optimal (Craven dan wahba, 1979).

## 2.10 Ukuran Kebaikan *Bandwidth* Optimal

Kebaikan suatu penduga dapat dilihat dari tingkat kesalahannya. Semakin kecil tingkat kesalahan suatu pendugaan maka semakin baik estimasinya. Menurut Chatterjee (2007), kriteria untuk mentukan estimator terbaik dalam model regresi antara lain nilai *Mean Square Error (MSE)* dan nilai koefisien determinasi *R-Square* ($R^2$).

MSE didefinisikan sebagai berikut :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \ . \quad (2.12)$$

Sedangkan koefisen determinasi didefinisikan sebagi berikut :

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (2.13)$$

$y_i$ adalah data variabel respon ke-i, $\bar{y}$ adalah mean data variabel respon, sedangkan $\hat{y}_i$ adalah nilai hasil estimasi variabel respon ke-i. *Sum of Square Regression (SSR)* adalah jumlah kuadrat simpangan hasil dugaan terhadap rata-rata variabel respon. Sedangkan *Sum of Square Total (SST )* adalah jumlah kuadrat simpangan variabel respon. SSR berfungsi untuk mengukur kualitas variabel prediktor sebagai prediktor variabel respon. Sehingga, koefisien determinasi dapat diartikan sebagai proporsi keragaman total variabel respon yang diukur oleh variabel prediktor.