Jurnal Matematika UNAND
Vol. **4** No. **3** Hal. 31 – 40
ISSN : 2303–2910

PEMODELAN BERAT BADAN BALITA DENGAN MENGGUNAKAN REGRESI KERNEL

AGNI HORTI MAHARANI, HAZMIRA YOZZA, YUDIANTRI ASDI

Program Studi Matematika,
Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Andalas,
Kampus UNAND Limau Manis Padang, Indonesia,
maharani.gates@ymail.com

Abstrak. Dalam analisis regresi terdapat dua pendekatan yang digunakan untuk mengestimasi fungsi regresi yaitu pendekatan parametrik dan pendekatan nonparametrik. Pendekatan parametrik digunakan apabila informasi hubungan antara variabel prediktor dengan variabel respon diketahui. Namun apabila informasi hubungan antara variabel prediktor dengan variabel respon tidak diketahui maka alternatif lain yang dapat digunakan adalah dengan pendekatan nonparametrik. Estimator kernel adalah metode yang digunakan pada penelitian ini. Penelitian ini bertujuan untuk mengetahui bentuk model regresi kernel dengan fungsi kernel Gaussian untuk memodelkan berat badan balita berdasarkan umur serta membandingkan model yang dibentuk dengan Metode Kuadrat Terkecil. Dari hasil perhitungan diperoleh nilai koefisien determinasi (R^2) dengan regresi linier sederhana adalah sebesar 0,719 dan nilai koefisien determinasi (R^2) dengan regesi nonparametrik menggunakan estimator kernel adalah sebesar 0,770606. Dari hasil perhitungan tersebut disimpulkan bahwa pada kasus ini analisis regresi nonparametrik menggunakan estimator kernel dengan fungsi kernel Gaussian mampu menghasilkan model yang jauh lebih baik pada data berat badan balita terhadap umur daripada analisis regresi dengan Metode Kuadrat Terkecil.

Kata Kunci: Analisis regresi, estimator kernel, fungsi kernel gaussian, berat badan balita

1. Pendahuluan

Analisis regresi adalah salah satu teknik analisis data yang digunakan untuk memodelkan hubungan antara variabel prediktor (X) dengan variabel respon (Y). Hubungan tersebut dapat dinyatakan dalam model yang dinamakan model regresi yaitu $y_i = m(x_i)+\varepsilon_i$. Salah satu cara untuk memperkirakan bentuk hubungan antara variabel prediktor dengan variabel respon adalah dengan melihat bentuk pola hubungan pada diagram pencar *(scatter plot)*. Dengan mengetahui pola hubungan yang terbentuk maka dapat ditentukan pendekatan yang sesuai untuk mengestimasi fungsi regresi.

Terdapat dua pendekatan yang digunakan untuk mengestimasi fungsi regresi yaitu pendekatan parametrik dan pendekatan nonparametrik. Pendekatan parametrik digunakan apabila informasi hubungan antara variabel prediktor dengan variabel respon diketahui. Namun apabila informasi hubungan antara variabel prediktor dengan variabel respon tidak diketahui maka alternatif lain yang dapat digunakan adalah dengan pendekatan nonparametrik. Pendekatan regresi nonparametrik digunakan bila tidak terdapat informasi mengenai bentuk $m(x_i)$ dan

tidak tergantung pada asumsi bentuk kurva tertentu. Diantara metode-metode pendekatan nonparametrik tersebut, estimator kernel adalah metode yang digunakan pada penelitian ini.

Estimator kernel memiliki bentuk yang lebih fleksibel dan perhitungan matematisnya mudah disesuaikan. Pada regresi kernel dikenal suatu estimator yang biasanya digunakan untuk mengestimasi fungsi regresi yaitu estimator Nadaraya-Watson. Estimator dengan pendekatan kernel tergantung pada dua parameter yaitu fungsi kernel dan *bandwidth* yang digunakan. Ada tujuh fungsi kernel antara lain *Uniform*, *Triangle*, *Epanechnikov*, *Quartic*, *Triweight*, *Gaussian*, dan *Cosinus*. *Bandwidth* adalah parameter pemulus yang berfungsi untuk mengontrol kemulusan dari kurva yang diestimasi. Metode untuk mendapatkan bandwidth yang optimal adalah dengan meminimumkan nilai *Generalized Cross Validation* (GCV). Pembahasan jurnal ini menggunakan estimator kernel dengan fungsi kernel Gaussian untuk mengestimasi model regresi nonparametrik untuk data berat badan balita berdasarkan umur.

2. Landasan Teori

2.1. *Regresi Parametrik*

Analisis regresi adalah suatu metode statistika yang dapat digunakan untuk menganalisis hubungan antara variabel prediktor (X) dan variabel respon (Y). Hubungan antar kedua variabel tersebut dapat digambarkan oleh suatu model matematika yang dinamakan model regresi. Analisis regresi yang digunakan untuk membentuk model hubungan antara variabel respon dengan satu atau lebih variabel prediktor, dimana hubungan antara variabel tersebut linier maka disebut dengan regresi linier. Persamaan model untuk regresi linier dapat ditulis sebagai berikut :

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}.$$

Jika suatu analisis regresi linier mengkaji hubungan linier antara satu variabel prediktor dengan satu variabel respon, maka analisis regresi linier tersebut dinamakan analisis regresi linier sederhana. Model regresi linier sederhana secara umum adalah sebagai berikut:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Regresi parametrik mengasumsikan bentuk fungsi regresi tertentu dan distribusi galatnya harus memenuhi asumsi tertentu seperti normalitas, heteroskedastisitas, dan lain-lain. Terpenuhinya asumsi tersebut sangat berpengaruh terhadap keabsahan analisis yang dilakukan terutama yang terkait dengan sebaran, seperti pengujian hipotesis mengenai parameter.

2.1.1. *Metode Kuadrat Terkecil*

Metode kuadrat terkecil *(Ordinary Least Square)* merupakan suatu metode yang banyak digunakan untuk menduga parameter dari model regresi parametrik. Dengan metode ini, penduga parameter model diperoleh dengan cara meminimumkan Jumlah Kuadrat Sisaan (JKS), sehingga dengan Metode Kuadrat Terkecil

diperoleh dugaan parameter β_1 dan β_0 dari persamaan regresi linier sederhana sebagai berikut:

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$
$$b_0 = \bar{y} - b_1 \bar{x}.$$

Setelah b_1 dan b_0 diperoleh, maka dugaan model regresi linier sederhana adalah sebagai berikut:

$$\widehat{y_i} = b_0 + b_1 x_i.$$

2.1.2. *Asumsi Normalitas (Kenormalan Data)*

Salah satu asumsi yang harus dipenuhi dalam analisis regresi adalah bahwa galat harus menyebar menurut sebaran normal. Model regresi yang baik adalah yang memiliki nilai galat yang terdistribusi normal. kenormalan sering dilakukan dengan statistik uji Kolmogorov-Smirnov. Pengujian Hipotesis untuk uji Kolmogorov-Smirnov dapat dinyatakan sebagai berikut:

H_0 : data mengikuti sebaran normal
H_1 : data tidak mengikuti sebaran normal

Statistik uji yang digunakan adalah

$$D = \max |F_0(x) - S_n(x)|$$

dimana

$F_0(x)$: nilai fungsi kumulatif sebaran yang diharapkan
$S_n(x)$: nilai fungsi sebaran kumulatif yang diamati dari suatu sampel.

Kriteria untuk pengujian ini adalah tolak H_0 jika nilai D_{hitung} lebih dari nilai D_{tabel}. Jika $= 0.05$ dengan banyak pengamatan n maka D_{tabel} sebagai berikut:

$$D_{tabel} = \frac{1,36}{\sqrt{n}}$$

Jika H_0 ditolak maka disimpulkan bahwa galat tidak menyebar menurut sebaran normal dan sebaliknya jika H_0 tidak ditolak maka disimpulkan bahwa galat menyebar menurut sebaran normal.

2.1.3. *Asumsi Homogenitas (Kehomogenan Ragam)*

Pengujian asumsi homogenitas bertujuan untuk memperlihatkan data pengamatan memiliki varians yang sama atau tidak. Salah satu cara yang dapat dilakukan untuk memeriksa apakah kondisi asumsi kehomogenan ragam terpenuhi atau tidak dengan melakukan plot antara nilai residual (e_i) dengan nilai dugaan ($\widehat{y_i}$). Jika titik-titik sisaan menyebar secara acak dan tidak membentuk suatu pola tertentu maka dapat dikatakan bahwa kehomogenan dari varians galat telah terpenuhi.

2.1.4. *Koefisien Determinasi*

Koefisien determinasi didefinisikan sebagai berikut:

$$R = 1 - \frac{\Sigma_{i=1}^{n}(y_i - \widehat{y_i})^2}{\Sigma_{i=1}^{n}(y_i - \overline{y})^2}$$

Koefisien determinasi (R^2) merupakan besaran yang digunakan untuk mengukur kelayakan model regresi dan menunjukkan besarnya kontribusi X terhadap perubahan Y. Semakin tinggi nilai R^2 (mendekati 1) semakin baik model yang terbentuk.

2.2. *Regresi Nonparametrik*

Pendekatan nonparametrik merupakan pendugaan model yang dilakukan berdasarkan pendekatan yang tidak terikat asumsi bentuk kurva regresi tertentu. Kurva regresi yang sesuai dengan pendekatan nonparametrik diwakili oleh model yang disebut dengan model regresi nonparametrik. Regresi nonparametrik adalah suatu teknik analisis data dalam statistika yang menjelaskan hubungan antara variabel prediktor dan variabel respon yang tidak diketahui bentuk fungsinya karena sebelumnya tidak ada informasi tentang bentuk kurva regresi $m(x)$.

2.3. *Regresi Kernel*

Regresi kernel adalah sebuah teknik nonparametrik dalam statistik untuk menduga nilai harapan bersyarat dari variabel acak. Dengan tujuan untuk menemukan hubungan nonlinier antara sepasang variabel acak X dan Y. Dalam regresi nonparametrik, nilai harapan bersyarat dari variabel Y bila X diketahui ditulis $E(Y|X)$ atau $E(Y|X = x) = \int \frac{f(x,y)}{f(y)} dy$ dimana m adalah fungsi yang tidak diketahui.

2.3.1. *Fungsi Kernel*

Secara umum, kernel K dengan parameter pemulus *(bandwidth)* h didefinisikan sebagai berikut:

$$K_h(u) = \frac{1}{h} K(\frac{u}{h}) \text{ untuk } -\infty < u < \infty \text{ dan } h > 0.$$

$K_h(u)$ dengan $u = x - x_i$ adalah fungsi kernel jika memenuhi sifat-sifat berikut:

(i) $K(u) = 0$, untuk semua u,
(ii) $\int_{-\infty}^{\infty} K(u)du = 1$,
(iii) $\int_{-\infty}^{\infty} uK(u)du = 0$,
(iv) $\int -\infty^{\infty} u^2 K(u)du = \sigma^2 > 0$.

Pada Tabel 1 berikut diberikan beberapa fungsi kernel. dimana I merupakan fungsi indikator yang didefinisikan sebagai berikut:

$$I(|u| \leq 1) = \begin{cases} 0, & \text{untuk } |u| > 1, \\ 1, & \text{untuk } |u| \leq 0 \end{cases}$$

Tabel 1. Macam-Macam Fungsi Kernel

Kernel	$K(u)$
Uniform	$K(u) = \frac{1}{2} I(\|u\| \leq 1)$
Triangle	$K(u) = (1 - \|u\|) I(\|u\| \leq 1)$
Epanechnikov	$K(u) = \frac{3}{4}(1 - u^2) I(\|u\| \leq 1)$
Quartic	$K(u) = \frac{15}{16}(1 - u^2)^2 I(\|u\| \leq 1)$
Triweight	$K(u) = \frac{35}{22}(1 - u^2)^3 I(\|u\| \leq 1)$
Gaussian	$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) I(\|u\| < \infty)$
Cosinus	$K(u) = \frac{\pi}{4} \cos\left(\frac{\pi}{2}u\right) I(\|u\| \leq 1)$

2.3.2. *Pemilihan Bandwidth Optimal*

Bandwidth dari kernel adalah parameter bebas yang menunjukkan pengaruh yang kuat pada perkiraan yang dihasilkan. Salah satu metode unuk mendapatkan h optimal adalah dengan menggunakan kriteria *Generalized Cross Validation* (GCV) yang didefinsikan sebagai berikut:

$$GCV(h) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \widehat{m}(x_i)}{1 - n^{-1} \Sigma_{i=1}^{n} w_i x_i} \right)^2$$

Nilai *bandwidth* optimal juga dapat dicari dengan menggunakan rumus berikut [8]:

$$h_{opt} = 1.06 A n^{-\frac{1}{5}}$$

dengan

$$\begin{aligned} A &= \min\{s, \frac{R}{1,34}\} \\ n &= \text{ banyak data} \\ R &= \text{ jangkauan kuartil} \\ s &= \text{ standar deviasi.} \end{aligned}$$

3. Metode Penelitian

Data yang digunakan pada penelitian ini adalah data sekunder yang diperoleh dari posyandu Kenagarian Pandam Gadang, Kecamatan Gunuang Omeh, Kabupaten Lima Puluh Kota, Sumatera Barat pada bulan November 2014. Variabel prediktor (X) yang digunakan adalah umur balita (dalam bulan) dan variabel respon (Y) yaitu berat badan balita (dalam kilogram). Banyak data yang digunakan pada penelitian ini adalah sebanyak 44 pengamatan.

Langkah-langkah yang dilakukan dalam analisis data adalah sebagai berikut:

(1) Membentuk model dengan Metode Kuadrat Terkecil

(a) Menentukan dugaan model persamaan regresi antara variabel prediktor dan variabel respon.
(b) Menentukan nilai koefisien determinasi
(c) Memeriksa keterpenuhan asumsi kenormalan dan kehomogenan ragam. Asumsi kenormalan dilakukan dengan uji Kolmogorov-Smirnov

- Menentukan hipotesis awal
  H_0 : galat menyebar normal
  H_1 : galat tidak menyebar normal
- Menetapkan taraf uji $\alpha = 0,05$
- Penetapan dan perhitungan statistik uji

$$D = \max |F_0(x) - S_n(x)|$$

- Membandingkan nilai statistik uji (D_{hitung}) dengan nilai kritisnya (D_{tabel}), kemudian ambil kesimpulan
- Pemberian interpretasi terhadap kesimpulan

Asumsi kehomogenan ragam diperlihatkan dengan melihat plot data $\hat{y}_i$ dan e_i.

(2) Analisis dengan Menggunakan Estimator Kernel

(a) Mencari estimasi parameter dari $\widehat{m}(x)$, dapat dilakukan dengan langkah sebagai berikut:

- Mencari s (standar deviasi), dengan menggunakan rumus

$$s^2 = \frac{1}{n-1}(x_i - \overline{x})^2.$$

- Mencari R
  R adalah jangkauan kuartil dimana $R = x_{\max} - x_{\min}$
- Mencari A dengan $A = \min(s, \frac{R}{1,34})$
- Mencari nilai bandwidth
  $h_{opt} = 1,06An^{(-1/5)}$
- Mencari estimasi parameter $\widehat{m}(x)$

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{(-\frac{1}{2}\left(\frac{x-x_i}{h}\right)^2 y_i)}}{\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{(-\frac{1}{2}\left(\frac{x-x_i}{h}\right)^2)}}$$

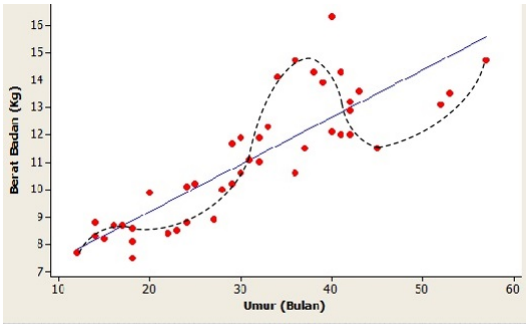
(b) Menentukan nilai koefisien determinasi.

(3) Membandingkan model regresi kernel dengan model yang dibentuk menggunakan metode kuadrat terkecil.

4. Pembahasan

4.1. *Eksplorasi Data*

Pada penelitian ini, data yang digunakan adalah data sekunder yang berasal dari posyandu Kenagarian Pandam Gadang, Kecamatan Gunuang Omeh pada bulan November 2014. Data yang diambil adalah data berat badan balita berdasarkan umur yang berada antara umur 12 bulan sampai dengan umur kurang dari 57 bulan dengan rata-rata berat badan sekitar $11,1$ kg. Untuk melihat pola sebaran data antara variabel respon (umur) dan variabel prediktor (berat badan) dilakukan dengan mengguanakan diagram pencar.



Gambar 1. Diagram pencar umur terhadap berat badan balita

4.2. *Model dengan Metode Kuadrat Terkecil*

Dengan menggunakan metode kuadrat terkecil diperoleh persamaan regresi linier sederhana yang menggambarkan hubungan

$$\widehat{y} = 5,7583 + 0,17194x$$

dengan
x : umur
y : berat badan

Dari persamaan di atas tampak bahwa nilai $b_0 = 5,7583$ dan nilai $b_1 = 0,17194$, ini berarti jika terjadi kenaikan umur sebesar 1 bulan maka nilai dugaan kenaikan rata-rata berat badan balita sebesar $0,17194$ kg. Selanjutnya diperoleh nilai koefisien determinasi yang sama yaitu sebesar $0,719$ yang berarti bahwa besarnya kontribusi pengaruh umur terhadap berat badan balita sebesar $0,719$ dan selebihnya sebesar $0,281$ dipengaruhi oleh variabel lain.

Untuk mengetahui sebaran galat maka dilakukan uji kenormalan, salah satunya yaitu dengan uji Kolmogorov-Smirnov. Berikut ini adalah hipotesis uji Kolmogorov-Smirnov :

H_0 : galat menyebar normal
H_1 : galat tidak menyebar normal

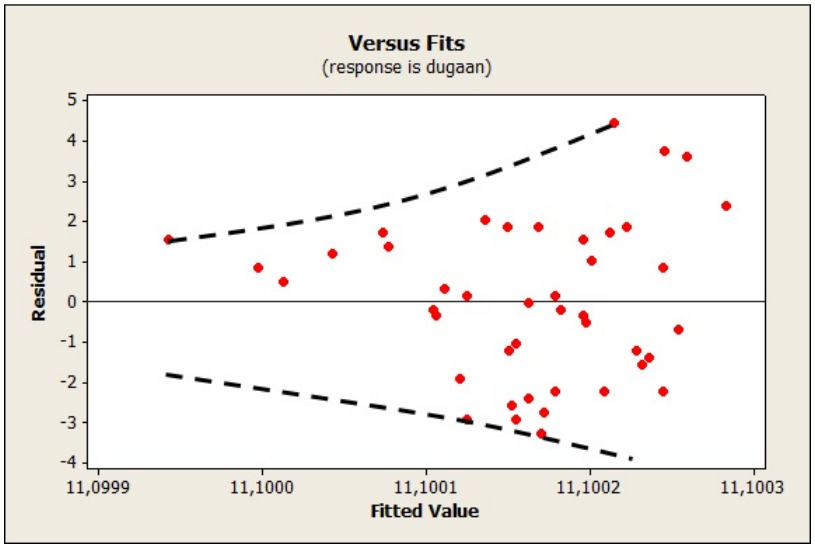
Dalam menggunakan uji Kolmogorov-Smirnov, statistik uji yang kita gunakan berdasarkan persamaan 2.1.10.

Tabel 2. Uji Normalitas

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	Df	Sig.	Statistic	Df	Sig.
VAR00001	.132	44	.051	.940	44	.024
a. Lilliefors Significance Correction						

Berdasarkan Tabel 2 diperoleh nilai $D_{hitung} = 0,132$, sementara nilai $D_{tabel} = 0,205$ dengan taraf uji $= 0,05$ dan $p-value = 0,051$ dan $n = 44$. Karena $D_{hitung} < D_{tabel}$ yaitu $0,132 < 0,205$ dan $p - value >$ yaitu $0,051 > 0,05$ maka dikatakan tidak tolak H_0 dan disimpulkan bahwa galat menyebar normal. Hal ini menunjukkan bahwa asumsi kenormalan terpenuhi.

Plot kehomogenan ragam pada Gambar 2 memperlihatkan bahwa ragam sisaannya bersifat tidak konstan dan membentuk suatu pola, dimana pola sebaran yang terbentuk adalah semakin besar nilai dugaannya maka variansi dari nilai residualnya terlihat semakin besar. Akibatnya asumsi kehomogenan ragam tidak terpenuhi.



Gambar 2. Plot nilai residual (e_i) dengan nilai dugaan ($\widehat{y}$)

4.3. *Model dengan Menggunakan Estimator Kernel*

Dengan menggunakan estimator kernel diperoleh

$$\widehat{m}(x) = \frac{\sum_{i=1}^{44} \frac{1}{\sqrt{2\pi}} e^{(-\frac{1}{2}\left(\frac{x-x_i}{5,5947}\right)^2 y_i)}}{\sum_{i=1}^{44} \frac{1}{\sqrt{2\pi}} e^{(-\frac{1}{2}\left(\frac{x-x_i}{5,5947}\right)^2)}}$$

Selanjutnya diperoleh nilai koefisien determinasinya sebesar $0,770606$ yang berarti bahwa besarnya kontribusi pengaruh umur terhadap berat badan balita sebesar $0,770606$ dan selebihnya sebesar $0,229394$ dipengaruhi oleh variabel lain.

Model terbaik didapat dari perbandingan nilai R^2 dari analisis data menggunakan regresi linier dan analisis data menggunakan regresi kernel. Dari hasil perhitungan diperoleh nilai koefisien determinasi (R^2) dengan regresi linier sederhana adalah sebesar $0,719$ dan hasil perhitungan koefisien determinasi (R^2) dengan regesi nonparametrik menggunakan estimator kernel adalah sebesar $0,770606$, sesuai dengan teori koefisien determinasi maka yang dipilih adalah nilai koefisien determinasi yang terbesar.

Dengan demikian dapat disimpulkan bahwa pada kasus ini analisis regresi nonparametrik menggunakan estimator kernel dengan fungsi kernel Gaussian mampu menghasilkan model yang jauh lebih baik pada data berat badan balita terhadap umur daripada analisis regresi linier

5. Kesimpulan

Berdasarkan pembahasan sebelumnya dapat diambil kesimpulan bahwa dugaan model regresi linier dengan metode kuadrat terkecil adalah:

$$\widehat{y} = 5,7583 + 0,17194x$$

dengan nilai koefisien determinasi sebesar $0,71900$ sedangkan dugaan model regresi kernel dengan fungsi kernel Gaussian diperoleh nilai bandwidth optimal sebesar $5,597$ dan nilai koefisien determinasi sebesar $0,770606$ adalah:

$$\widehat{m}(x) = \frac{\sum_{i=1}^{44} \frac{1}{\sqrt{2\pi}} e^{(-\frac{1}{2}\left(\frac{x-x_i}{5,5947}\right)^2 y_i)}}{\sum_{i=1}^{44} \frac{1}{\sqrt{2\pi}} e^{(-\frac{1}{2}\left(\frac{x-x_i}{5,5947}\right)^2)}}$$

Dari kedua dugaan model regresi tersebut, dugaan model regresi kernel merupakan dugaan model regresi yang lebih baik untuk digunakan pada data Berat Badan Balita di Posyandu Kenagarian Pandam Gadang, Kecamatan Gunuang Omeh, Kabupaten Lima Puluh Kota jika dibandingkan dengan dugaan model regresi linier. Hal ini disebabkan karena dugaan model regresi dengan nilai koefisien determinasi terbesar merupakan model yang lebih baik.

6. Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada Bapak Dr Dodi Devianto, Ibu Dr Susila Bahri, dan Bapak Narwen M.Si yang telah memberikan masukan dan saran dalam penyempurnaan penulisan artikel ini.

Daftar Pustaka

[1] Bain, L.J and M. Engelhardt. 1987. *Introduction to Probability and Mathematical Statistics Second Edition*. Introduction to Probability and Mathematical Statistics Second Edition. Duxbury Press. California.

[2] Connover, W.J. 1980. *Practical Nonparametric Statistics*. John Wiley and Sons. New York.

[3] Eubank, R, L, 1999. *Nonparametric Regression and Smoothing Spline*. Marcel Dekker Inc. New York

[4] Hardle, W. 1994. *Applied Nonparametric Regression*. Cambridge University Press. New York.

[5] Montgomery, D,C. 1992. *Introduction to Linear Regression Analysis Second Edition*. John Wiley and Son. New York.

[6] Sembiring, R.K. 1995. *Analisis regresi*. ITB, Bandung.

[7] Siegel, S. 1992. *Statistika Nonparametrik untuk Ilmu-Ilmu Sosial*. PT. Gramedia Pustaka Utama. Jakarta.

[8] Silverman, BW. 1986. *Density Estimation for Statistics and Data Analysis*.Chapman and Hall. London