REGRESI KERNEL DENGAN METODE NADARAYA WATSON

SKRIPSI

Diajukan Kepada Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Negeri Yogyakarta
Untuk Memenuhi Sebagian Persyaratan
Guna Memenuhi Gelar Sarjana Sains

Oleh:
ESTY
07305144023

PROGRAM STUDI MATEMATIKA
JURUAN PENDIDIKAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS NEGERI YOGYAKARTA
2014

REGRESI KERNEL DENGAN METODE NADARAYA WATSON

Oleh :

Esty

NIM. 07305144023

ABSTRAK

Dalam analisis regresi terdapat dua jenis pendekatan dalam menentukan kurva regresi, yaitu pendekatan parametrik dan nonparametrik. Regresi kernel merupakan salah satu model dengan pendekatan nonparametrik yang tidak menggunakan asumsi tertentu mengenai bentuk kurva regresi maupun distribusi galat. Tujuan dari penelitian ini adalah menjelaskan penggunaan regresi kernel untuk mengestimasi kurva regresi serta aplikasinya. Metode yang digunakan dalam regresi kernel adalah metode estimasi Nadaraya-Watson dengan menggunakan fungsi Kernel Gaussian. Konsep estimasi Nadaraya-Watson bertujuan untuk mengestimasi kurva regresi yang tidak cocok dengan datanya, tetapi juga memiliki derajat kemulusan tertentu, dimana kemulusan kurva regresi dipengaruhi oleh pemilihan *bandwith* (h) yang optimal yaitu nilai h yang menghasilkan nilai terkecil dari *CV (Cross Validation).* Perhitungannya menggunakan bantuan sofware MATLAB 7.10 dan untuk menentukan nilai CV menggunakan sofware excel

Langkah-langkah untuk menentukan estimasi kernel dengan metode Nadaraya Watson adalah: (1) menghitung nilai bobot kernel dari data yang diketahui, (2) menghitung nilai $\hat{m}_h(x)$ dengan menggunakan rumus Nadaraya Watson, (3) menghitung nilai *Cross Validation* (CV_h), (4) memilih nilai *bandwith* yang menghasilkan *Cross Validation* terkecil. Contoh penerapan dari skripsi ini diambil dari permasalahan yang dialami oleh PT PLN mengenai penurunan tegangan tenaga listrik. Adapun data yang digunakan adalah besarnya penurunan tegangan sesaat pada durasi setiap 0,5 detik sebanyak 25 pengamatan. Hasil dari penerapan regresi kernel dengan metode estimasi Nadaraya-Watson memperoleh grafik regresi yang sangat mendekati plot data asli dengan nilai h optimalnya adalah h = 1,8 dengan $1 \leq h \leq 2$ dan nilai $CV_h = 0,803$. Sehingga regresi kernel dengan metode Nadaraya Watson adalah metode yang baik untuk mengestimasi grafik regresi yang belum diketahui fungsinya.

Kata kunci : Nadaraya Watson, fungsi Gaussian, bandwith

BAB I

PENDAHULUAN

A. Latar Belakang

Analisis regresi merupakan suatu metode statistika yang dapat digunakan untuk mengetahui hubungan antara suatu variabel terikat (dependen) Y terhadap satu atau lebih variabel bebas (independen) X sehingga memperoleh persamaan dan menggunakan persamaan tersebut untuk membuat perkiraan atau prediksi. Untuk sebuah sampel berukuran n data pengamatan (X_1, Y_1), ... , (X_n, Y_n), hubungan antara variabel-variabel tersebut dapat dinyatakan dengan model regresi $Y=m(X)+\varepsilon$. Dimana m adalah fungsi matematik yang disebut sebagai fungsi regresi yang belum diketahui dan ε adalah error. Dalam regresi parametrik, model regresi ada dua yaitu model regresi linear dan nonlinear. Model regresi linear merupakan metode statistika yang digunakan untuk menganalisis hubungan linear antara satu variabel atau lebih variabel bebas $(X_1, X_2, X_3, \ldots, X_p)$ dengan variabel terikat (Y). Model regresi non linear adalah menganalisis hubungan non linear antara dua variabel yaitu variabel bebas dan variabel terikat. Beberapa bentuk dari regresi linear diantaranya regresi linear sederhana maupun regresi linear berganda yang digunakan untuk memperoleh model hubungan linear antara variabel-variabel bebas dengan variabel terikat sepanjang tipe datanya adalah interval atau rasio.

Pendekatan nonparametrik merupakan pendekatan regresi yang sesuai untuk pola data yang tidak diketahui bentuknya, atau tidak terdapat informasi masa lalu tentang pola data (I Nyoman Budiantara, 2010: 1). Model regresi nonparametrik yaitu kurva regresi berdasarkan pendekatan nonparametrik diwakili oleh suatu model**.** Dalam regresi nonparametrik fungsi regresi umumnya hanya diasumsikan termuat dalam suatu ruang fungsi yang berdimensi tak hingga.

Menurut Lilis Laome, (2010: 1) dalam jurnalnya yang berjudul Perbandingan Model Regresi Nonparametrik dengan Regresi Spline dan Kernel memberikan kesimpulan ada beberapa metode pendekatan regresi nonparametrik dan di antara metode-metode yang paling sering digunakan yaitu metode nonparametrik dengan pendekatan spline dan kernel. Kedua metode tersebut memiliki keunggulan masing-masing. Dalam pendekatan kernel perhitungan matematisnya mudah disesuaikan, sedangkan pendekatan spline dapat menyesuaikan diri secara efektif terhadap data sehingga didapatkan hasil yang mendekati kebenaran.

I Nyoman Budiantara (2010: 1) mengungkapkan bahwa terdapat beberapa teknik untuk mengestimasi kurva regresi dalam regresi nonparametrik, yaitu estimator kernel dan histogram, *spline*, Deret Fourier dan Wavelets, dan Deret barisan estimasi orthogonal. Menurut Siana Halim, Indriati Bisono (2006: 74) dalam jurnalnya yang berjudul Fungsi-Fungsi Kernel pada Metode Regresi Nonparametrik dan Aplikasinya pada *Priest River Experimental Forest's Data* memberikan kesimpulan jika asumsi terhadap sebuah model parametrik dibenarkan, maka fungsi regresi dapat diestimasi dengan cara yang lebih efisien

jika dibandingkan dengan menggunakan sebuah metode nonparametrik. Tetapi jika asumsi terhadap model parametrik ini salah, maka hasilnya akan memberikan kesimpulan yang salah terhadap fungsi regresi.

Menurut I Komang Gede Sukarsa, (2012:21) dalam jurnalnya yang berjudul estimator kernel dalam model regresi nonparametrik mengungkapkan bahwa regresi kernel adalah teknik statistik nonparametrik untuk mengestimasi nilai $E(Y|X) = m(X)$ atau $y = m(X)$ dalam suatu variabel. Tujuan regresi kernel yaitu untuk memperoleh hubungan nonlinear antara X dengan Y.

Menurut Lilis Laome, untuk mencapai suatu pendekatan fungsi regresi nonparametrik perlu mengestimasi ekspektasi bersyarat $m(X)$ dengan menggunakan metode Nadaraya–Watson. Sehingga dapat diketahui besarnya bias dan variansnya.

Terdapat beberapa jenis fungsi kernel, antara lain kernel uniform, kernel triangle, kernel epanechnikov, kernel gaussian, kernel kuartik dan kernel cosinus (Hardle, 1990). Dalam regresi kernel, pemilihan parameter pemulus (*bandwidth*) jauh lebih penting dibandingkan dengan memilih fungsi kernel. Dalam regresi kernel yang menjadi permasalahan adalah pemilihan *bandwidth,* bukan pada pemilihan fungsi kernel. Fungsi kernel yang umum digunakan adalah Kernel Gaussian. Pada pembahasan skripsi ini akan digunakan metode Nadaraya Watson untuk mengestimasi model regresi nonparametrik dengan fungsi berdistribusi normal.

B. Rumusan Masalah

Berdasarkan latar belakang masalah di atas maka dapat dirumuskan permasalahan sebagai berikut :

1. Bagaimana regresi kernel dengan metode estimasi Nadayara-Watson dalam fungsi kernel Gaussian?
2. Bagaimana penerapan dalam penggunaan metode estimasi Nadaraya-Watson?

C. Tujuan Penulisan

Berdasarkan rumusan masalah tersebut maka tujuan penulisan ini adalah sebagai berikut :

1. Menjelaskan regresi kernel dengan metode estimasi Nadaraya-Watson dalam fungsi kernel Gaussian.
2. Menjelaskan penggunaan metode estimasi Nadaraya-Watson.

D. Manfaat

Manfaat dari penulisan skripsi ini adalah :

1. Bagi penulis

   Dapat memberikan gambaran dan ilmu pengetahuan tentang penggunaan regresi kernel dengan metode Nadaraya-Watson.

2. Bagi Jurusan Pendidikan Matematika

   Dapat dijadikan sebagai referensi maupun informasi tambahan perpustakaan Jurusan Pendidikan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Yogyakarta.

BAB II
LANDASAN TEORI

Pada BAB II ini akan dibahas mengenai Analisis Regresi, Regresi Parametrik, Regresi Nonparametrik, Estimasi Kernel, Sifat - Sifat Estimator, Fungsi Densitas Peluang dan Deret Taylor. Pembahasan - pembahasan tersebut akan dijadikan sebagai landasan teori pada bab selanjutnya.

A. Analisis Regresi

Analisis regresi adalah suatu metode statistika yang dapat digunakan untuk menganalisis hubungan antara suatu variabel terikat (dependen) Y terhadap satu atau lebih variabel bebas (independen) X. Hubungan antar kedua variabel tersebut dapat digambarkan oleh suatu kurva regresi dengan bentuk fungsi regresi tertentu.

Diberikan n pengamatan $X_i, Y_i \; ; \; i = 1, 2, \dots, n \; ; \; X_i \in R; \; Y_i \in R$. Hubungan antara X_i dan Y_i diasumsikan mengikuti model regresi :

$$Y_i = f\ x_i\ + \varepsilon_i \, ; i = 1, 2, \dots, n. \qquad \dots\dots\dots\dots\dots\dots \quad (2.1)$$

dengan :

$f\ x_i$: kurva regresi
ε_i : variabel galat

Dalam penggunaan regresi terdapat beberapa asumsi galat yang harus dipenuhi. Asumsi-asumsi galat yang harus dipenuhi adalah sebagai berikut:

1. Galat-galat ε merupakan variabel acak dengan mean nol dan variansi σ^2 atau $E\ \varepsilon_i\ = 0$dan $V\ \varepsilon_i\ = \sigma^2$.
2. Galat-galat ε ($\varepsilon_i$dan ε_j, $i \neq j$) tidak berkorelasi (saling bebas) sehingga $cov\ \varepsilon_i,\ \varepsilon_j\ = 0, i \neq j$.
3. Galat-galat ε berdistribusi normal.

Menurut Eubank (1988: 3) dan Hardle (1990: 4) terdapat dua jenis pendekatan dalam menentukan kurva regresi yaitu pendekatan parametrik dan pendekatan non parametrik atau regresi non parametrik.

B. Regresi Parametrik

Apabila dalam analisis regresi bentuk kurva regresi telah diketahui, maka model regresi tersebut dinamakan model regresi parametrik (Hardle, 1990: 4). Regresi parametrik merupakan metode statistika yang digunakan untuk mengetahui hubungan antara variabel bebas dan variabel terikat, dengan asumsi bahwa bentuk kurva regresi diketahui.

Pendekatan parametrik mengasumsikan bentuk fungsi regresi tertentu dan distribusi galatnya harus memenuhi asumsi tertentu seperti normalitas, homokedastisitas, tidak terjadi autokorelasi dan multikoliniearitas. Asumsi-asumsi tersebut sangat berpengaruh terhadap model regresi. Dalam model regresi parametrik, terdapat dua model yaitu model linear dan non linear.

1. Model Regresi Linear

Analisis regresi linear merupakan model statistika yang digunakan untuk menganalisis hubungan linier antara satu variabel atau lebih variabel bebas ($X_1, X_2, \ldots, X_p$) dengan variabel terikat (Y). Secara matematis dapat ditulis dalam model regresi linear sebagai berikut:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon_i. \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.2)$$

dengan :

Y_i : variabel terikat dalam pengamatan ke-i
$\beta_0$danβ_j : parameter
X_j : variabel bebas dari pengamatan ke-j
ε_i : variabel galat acak

Pada kasus di mana model regresi pada persamaan (2.2) hanya dibentuk oleh satu variabel bebas maka disebut dengan Regresi Linear Sederhana (*Simple Linear Regression*). Persamaannya menjadi:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i. \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.3)$$

Asumsi-asumsi dalam analisis regresi linear sederhana adalah sebagai berikut:

1. Galat memiliki ragam yang konstan.
2. Galat menyebar normal.
3. Galat bersifat saling bebas.

Sedangkan untuk variabel bebas lebih dari satu $p > 1$ disebut Regresi Linier Berganda. Dari persamaan (2.2) dapat diubah menjadi:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon_i.$$

Persamaan regresi dugaan untuk model Regresi Linear Berganda adalah

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p. \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.4)$$

2. Regresi Polinomial

Salah satu contoh tipe dari model parametrik adalah persamaan regresi polinomial dimana parameter-parameter tersebut adalah koefisien dari variabel bebas (Hardle, 1990: 4). Menurut Sembiring (1995: 231), polinom banyak digunakan dalam menghampiri suatu kurva, artinya suatu kurva selalu dapat dihampiri oleh suatu deret polinom. Regresi polinomial adalah bentuk khusus dari model regresi linier umum dalam parametrik yang dibentuk dengan menjumlahkan pengaruh masing-masing variabel bebas yang dipangkatkan sampai orde ke-r. Secara umum, model ditulis sebagai berikut:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \sum_{i=0}^{r} \beta_{ij} X_j^r + \varepsilon_i. \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.5)$$

dengan :

Y_i : Variabel terikat dalam pengamatan ke-r.
X_j^i : Variabel bebas ke-j dengan orde ke-r.
β_{ij} : Koefisien regresi yang bersesuaian dengan variabel bebas ke-j dengan orde ke-r.
ε_i :Variabel galat acak.

Asumsi-asumsi yang harus dipenuhi dalam regresi polinomial, diantaranya adalah:

1. $E\ \varepsilon_i\ = 0.$
2. $Cov\ \varepsilon_i, \varepsilon_j\ = 0,\ i \neq j$ (tidak terjadi autokorelasi).
3. Ragam galat homogen (tidak terjadi heteroskedesitas).
4. Tidak terjadi korelasi antar variabel bebas (multikolinearitas).
5. Galat berdistribusi normal.

C. Regresi Nonparametrik

Statistik nonparametrik dapat digunakan pada data yang memiliki distribusi normal ataupun tidak. Istilah nonparametrik pertama kali diperkenalkan oleh Wolfowitz pada tahun 1942. Pendekatan nonparametrik merupakan pendekatan regresi yang sesuai untuk pola data yang tidak diketahui bentuknya, atau tidak terdapat informasi masa lalu tentang pola data (Budiantara, 2010). Menurut Hardle (1990: 5) 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 merupakan suatu metode regresi untuk mengetahui pola hubungan antara satu variabel bebas ($X_1, X_2, \ldots, X_p$) dengan variabel terikat Y. Regresi nonparametrik tidak membutuhkan asumsi mengenai bentuk kurva

regresi maupun distribusi galat. Oleh karena itu, regresi nonparametrik bersifat lebih fleksibel terhadap perubahan pola data (Eubank, 1988: 3).

Regresi nonparametrik yang hanya memiliki satu variabel disebut regresi nonparametrik sederhana. Regresi nonparametrik tersebut dimodelkan sebagai berikut:

$$Y = f\ x\ + \varepsilon. \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.6)$$

dengan :

Y : variabel terikat.
$f\ x$: fungsi regresi nonparametrik.
ε : variabel galat acak.

Prosedur dalam statistika yang digunakan untuk menganalisis data ditentukan oleh skala pengukuran yang digunakan ketika melakukan pengamatan. Pengukuran adalah sekumpulan aturan untuk menetapkan suatu bilangan yang mewakili obyek, sifat, karakteristik, atribut atau tingkah laku. Skala adalah perbandingan antar benda yang menghasilkan bobot nilai yang berbeda. Skala pengukuran adalah kesepakatan yang digunakan untuk menentukan panjang pendeknya interval sehingga memiliki data yang kunatitatif.

Berdasarkan tingkatannya, terdapat empat macam skala pengukuran (Daniel, 1989), yaitu:

1. Skala nominal

   Skala nominal merupakan skala yang paling lemah di antara keempat skala pengukuran yang ada. Skala nominal juga disebut skala klasifikasi karena skala ini digunakan untuk mengklasifikasi suatu objek, orang atau sifat menggunakan angka-angka atau lambang-lambang berdasarkan nama atau predikat. Sebagai contoh, angka 1 digunakan untuk menyebut kelompok barang-barang yang cacat dan 0 untuk barang-barang yang tidak cacat dari suatu proses produksi. Angka 0 dan 1 digunakan sebagai lambang untuk membedakan antara barang-barang yang cacat dan tidak cacat. Dengan demikian, barang-barang yang tidak cacat dengan angka 0 dan barang-barang yang tidak cacat dengan angka 1 tanpa mengubah makna. Data semacam ini disebut data hitung atau data frekuensi.

2. Skala ordinal

   Skala ordinal merupakan skala yang membedakan kategori berdasarkan tingkat atau urutan. Skala ordinal merupakan skala pengukuran yang lebih teliti daripada skala nominal. Dengan menggunakan skala ordinal dapat dibedakan benda atau peristiwa yang satu dengan yang lainnya berdasarkan jumlah relatif beberapa karakteristik tertentu. Misalnya membagi tinggi badan sampel ke dalam tiga kategori: tinggi, sedang dan pendek. Skala ordinal juga sering disebut sebagai peringkat.

3. Skala interval

   Apabila suatu skala mempunyai sifat skala ordinal dan jarak antara dua angka pada skala diketahui maka skala interval dapat diterapkan. Dalam pengukuran menggunakan skala interval, rasio dua interval yang mana pun tidak tergantung pada unit pengukuran dan titik manapun, keduanya dipilih sembarang. Contoh pengukuran interval adalah pengukuran temperatur dalam derajat Farenheit dan Celcius. Titik nol yang tidak bernilai mutlak dan unit pengukuran dalam mengukur suhu adalah sembarang dan berlainan dalam kedua skala pengukuran tersebut. Meskipun demikian, skala pengukuran menggunakan derajat Farenheit dan Celcius mengandung informasi yang sama banyaknya dan sama jenisnya karena keduanya berhubungan linear, artinya yang terbaca pada skala yang satu dapat ditransformasi untuk hal yang sama pada skala yang lain.

4. Skala rasio

   Apabila suatu skala memiliki ciri – ciri suatu skala interval dan memiliki suatu titik nol mutlak sebagai titik asalnya maka skala tersebut dinamakan skala rasio. Dalam suatu skala rasio, perbandingan antara suatu titik skala tidak tergantung pada unit pengukuran. Data hasil pengukuran menggunakan skala rasio dapat dijumlahkan secara aljabar, misalnya rasio antara dua berat dalam ons sama dengan rasio antara dua berat dalam gram. Skala rasio merupakan skala dengan tingkat pengukuran paling tinggi.

D. Fungsi Densitas Peluang

Definisi 2.1 (Lee J. Bain dan Max Engelhardt, 1991)

Variabel acak *X* disebut variabel acak kontinu jika terdapat fungsi *f(x)* yang disebut dengan fungsi densitas peluang dari *x*, maka

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

Teorema 2.1 (Lee J. Bain dan Max Engelhardt, 1991)

Fungsi *f(x)* adalah fungsi densitas peluang dari variabel acak kontinu *X* jika dan hanya jika memenuhi

$$\int_{-\infty}^{\infty} f(x)dx = 1. \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.7)$$

Untuk setiap bilangan real *x* dan

$$f(x) \geq 0. \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.8)$$

Bukti Teorema 2.1

$$\int_{-\infty}^{\infty} f(x)dx = \lim_{x \to \infty} F(x)$$

$$= 1.$$

Terbukti persamaan (2.7)

$f(x)$ merupakan fungsi densitas peluang pada *X* sehingga terdapat $F(x)$

$$\lim_{x \to -\infty} F(x) = 0$$

$$f(x) \geq 0$$

Terbukti persamaan (2.8).

Definisi 2.2 (Lee J. Bain dan Max Engelhardt, 1991)

Distribusi dengan fungsi densitas peluang *f(x)* dikatakan simetris terhadap *c* jika *f(c - x) = f(c + x)* untuk semua *x*.

Dari definisi (2.2), jika c = 0 maka diperoleh

$$f(0-x)=f(0+x)$$

$$\Leftrightarrow f(-x)=f(x). \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.9)$$

Definisi 2.3 (Lee J. Bain dan Max Engelhardt, 1991)

Dalam fungsi densitas peluang jika X dan Y adalah peubah acak diskrit atau kontinu dengan fungsi densitas bersama $f(x,y)$, sehingga kondisi fungsi densitas bersama dari Y relatif terhadap $X=x$ didefinisikan

$$f(y|x)=\begin{cases}\dfrac{f(x,y)}{f(x)}, & f(x)>0\\ 0, & f(x)\le 0\end{cases} \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.10)$$

Definisi 2.4 (Lee J. Bain dan Max Engelhardt, 1991)

Jika X dan Y adalah distribusi bersama dari variabel acak, maka nilai harapan dari Y relatif terhadap $X=x$ adalah

$E(Y|x)=\sum yf(y|x)$, jika X dan Y diskrit. …… (2.11)

$E(Y|x)=\int_{-\infty}^{\infty} yf(y|x)dy$, jika X dan Y kontinu. …… (2.12)

Berdasarkan persamaan (2.10) dan (2.12) diperoleh nilai harapan bersyarat dari variabel Y relatif terhadap X.

$$E(Y|x)=\int_{-\infty}^{\infty} yf(y|x)dy$$

$$=\int_{-\infty}^{\infty} y\frac{f(x,y)}{f(x)}dy$$

$$=\int_{-\infty}^{\infty}\frac{yf(x,y)dy}{f(x)}.$$

E. Estimasi Kernel

Regresi nonparametrik dalam statistika digunakan untuk memperkirakan nilai harapan bersyarat dari variabel acak, yang bertujuan untuk menemukan hubungan nonlinier antara sepasang variabel acak Y dan X untuk mendapatkan dan menggunakan bobot yang sesuai.

Dalam setiap regresi nonparametrik, nilai harapan bersyarat dari variabel relatif terhadap variabel Y relatif terhadap variabel X dapat ditulis $E\,Y\,X = m(X)$. Dimana *m* adalah fungsi yang tidak diketahui. Untuk mengestimasi *m* dapat menggunakan kernel sebagai fungsi pembobotan.

Diberikan *n* sampel random X_i, *i=1, 2, 3, . . . , n*, maka karakteristik dasar yang menggambarkan sifat dari suatu variabel acak adalah fungsi densitas *f* dari variabel acak tersebut. Berdasarkan sampel acak ini akan diestimasi fungsi densitas *f* yang tidak diketahui dengan pendekatan kernel. Kernel K di definisikan (Hardle, 1990).

$$K_h(x)=h^{-1}K\left(\frac{x}{h}\right). \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.13)$$

Dengan *K* adalah fungsi Kernel dan *h* adalah *bandwidth*. Penghalusan dengan pendekatan kernel yang dikenal sebagai penghalusan kernel *(kernel smoother)* sangat bergantung pada fungsi kernel dan bandwidth. (Lilis Laome, 2010).

Menurut (Siana Halim, 2006) terdapat tiga macam estimasi kernel, yaitu:

1. Nadaraya Watson
2. Priestley chao
3. Gasser Muller Kernel

Sedangkan estimasi kernel yang paling sering digunakan adalah Nadaraya Watson yang hasilnya dapat memperoleh grafik yang mendekati data sebenarnya.

F. Sifat-sifat Estimator

Pada umumnya, semakin banyak observasi dalam data sampel, semakin tinggi akurasi suatu estimator. Oleh karena itu, sifat-sifat yang dibutuhkan oleh estimator dapat digolongkan menjadi dua kelompok tergantung pada besar kecilnya ukuran sampel, yaitu sifat sampel kecil dan sifat sampel besar (Gunawan Sumodiningrat, 2007: 40). Sifat-sifat sampel kecil atau sampel terbatas (*finite*) mengacu pada sifat-sifat distribusi sampel suatu estimator yang didasarkan pada ukuran sampel yang tetap (*fixed sample size*). Sifat-sifat sampel besar adalah sifat-sifat distribusi sampel suatu estimator yang diperoleh dari sampel yang banyaknya mendekati tak berhingga (*infinite*).

1. Sifat estimator untuk sampel kecil

Kriteria utama suatu estimator yang baik untuk sampel kecil adalah :

a. Tak bias (*Unbiasedness*)

Bias (penyimpangan) dari suatu estimator adalah perbedaan antara nilai harapan dan nilai parameter yang sebenarnya. Secara matematik, bias = $E(\theta) - \theta$.

Definisi 2.5 (Lee J. Bain dan Max Engelhardt, 1991)

Jika *X* adalah variabel acak kontinu dengan fungsi densitas *F*(*x*), maka nilai harapan didefinisikan dengan

$$E(X)=\int_{-\infty}^{\infty} xf(x)dx. \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.14)$$

Suatu estimator dikatakan tidak bias, apabila $E\,\theta - \theta = 0$. Oleh karena itu, dapat dikatakan bahwa θ adalah sebuah estimator yang tidak bias (*unbiased estimator*) terhadap θ apabila $E\,\theta = \theta$. Jika biasnya positif maka estimator tersebut mengalami "bias ke atas" dan jika biasnya negatif maka estimator tersebut mengalami "bias ke bawah".

Tak bias merupakan sifat yang dibutuhkan namun tidak terlalu penting. Hal ini disebabkan karena sifat tak bias tidak menunjukkan apapun mengenai penyebaran dari distribusi estimator. Suatu estimator yang tidak bias namun

mempunyai varians yang besar seringkali menghasilkan estimasi yang jauh berbeda dari nilai parameter yang sebenarnya (Gunawan Sumodiningrat, 2007).

b. Varians terkecil (*least variance*) atau estimator terbaik (*best estimator*)

Sebuah estimator dikatakan sebagai estimator terbaik apabila estimator tersebut memiliki varians terkecil (*least variance*) dibandingkan dengan estimator-estimator lain yang diperoleh dengan metode berbeda.

Teorema 2.2 (Lee J. Bain dan Max Engelhardt, 1991)

Jika X adalah variabel acak kontinu, maka

$$Var(X) = E(X^2) - (E(X))^2. \qquad \ldots\ldots\ldots\ldots\ldots\ldots (2.15)$$

Bukti Teorema 2.2

$$Var(X) = E(X^2 - 2E(X)X + (E(X))^2)$$

$$= E(X^2) - 2E(X)E(X) + (E(X))^2$$

$$= E(X^2) - 2(E(X))^2 + (E(X))^2$$

$$= E(X^2) - (E(X))^2.$$

Teorema 2.2 terbukti.

c. Minimum kesalahan kuadrat rerata (*Mean-Square-Error* atau MSE)

Kesalahan kuadrat rerata atau *mean-square-error* (MSE) adalah nilai harapan dari kuadrat perbedaan antara estimator dengan parameter populasi.

$$MSE\ (\theta) = E[\theta - \theta]^2$$

$$= E[\theta - E[\theta] + E[\theta] - \theta]^2$$

$$= E[\theta - E[\theta]]^2 + E[E(\theta) - \theta]^2 + 2E[\{\theta - E[\theta]\}\{E[\theta] - \theta\}]$$

karena

$$E[\theta - E[\theta]]^2 = var(\theta) \text{ dan } [E(\theta) - \theta]^2 = [bias\ (\theta)]^2$$

dan

$$E[\{\theta - E[\theta]\}\{E[\theta] - \theta\}] = E[\theta\ E[\theta] - \{E[\theta]\}^2 - \theta\theta + \theta\ E[\theta]]$$

$$= \{E[\theta]\}^2 - \{E[\theta]\}^2 - \theta\ E[\theta] - \theta\ E[\theta]$$

$$= 0.$$

sehingga, $MSE\ (\theta) = var(\theta) + \{bias\ (\theta)\}^2$. …… (2.16)

Jadi $MSE\ X_n$ sama dengan varians ditambah bias kuadrat. Jika X_n adalah penduga yang tak bias maka $MSE\ X_n$ merupakan variannya. Dengan kata lain, MSE adalah jumlah dari dua kuantitas, yaitu varians dan bias kuadrat. Apabila salah satu dari kedua komponen ini mempunyai nilai lebih kecil dibanding komponen lainnya, maka perbedaan tersebut ditunjukkan oleh MSE.Oleh karena itu estimator yang memiliki MSE terkecil lebih baik dari kriteria minimum dari salah satu komponen MSE.

d. *Best Linear Unbiasedness Estimator* (BLUE)

Suatu estimator dikatakan BLUE apabila estimator tersebut memenuhi kriteria linier, tidak bias (*unbiased*), dan memiliki varians terkecil bila dibandingkan dengan estimator lain juga linear dan tak bias (Gunawan Sumodiningrat, 1993).

2. Sifat estimator untuk sampel besar

Sifat-sifat asimptotik berkaitan dengan estimator-estimator yang diperoleh dari sampel-sampel besar. Sampel ini mempunyai ukuran sampel n, dengan $n \rightarrow \infty$. Dalam hal ini, pengertian asimptotik menunjukkan distribusi asimptotik dari suatu estimator. Menurut Gunawan Sumodiningrat (1993), beberapa sifat distribusi asimptotik dari estimator adalah :

a. Tak bias secara asimptotik (*asymptotic unbiasedness*)

Sebuah estimator dikatakan sebagai estimator yang tak bias secara asimptotik bagi parameter yang sebenarnya apabila :

$$\lim_{n\to\infty} E\ \theta_n = \theta.$$

Subskrip n pada θ menunjukkan ukuran sampel, sehingga $\lim_{n\to\infty} E[\theta_n] - \theta = 0$.

Definisi ini menyatakan bahwa sebuah estimator tidak bias secara asimptotik apabila penyimpangannya menjadi nol untuk $n \rightarrow \infty$. Sebuah estimator yang tidak bias tetap tidak bias secara asimptotik, namun tidak demikian sebaliknya.

b. Konsisten (*consistency*)

Sebuah estimator, θ, disebut estimator yang konsisten bagi θ apabila memenuhi dua syarat berikut :

1. θ adalah estimator yang tidak bias secara asimptotik atau $\lim_{n\to\infty} E\ \theta_n = \theta$.
2. Varians dari θ mendekati nol jika $n \rightarrow \infty$

$$\lim_{n\to\infty} var(\theta) = 0.$$

c. Efisien secara asimptotik (*asymptotic efficiency*)

Sebuah estimator θ, adalah estimator yang efisien secara asimptotik bagi θ apabila memenuhi syarat :

1. θ adalah konsisten.
2. θ memiliki varians asimptotik yang lebih kecil dibanding dengan varians asimptotik estimator konsisten lainnya.

Terdapat suatu kesulitan dalam menentukan apakah suatu estimator yang konsisten telah memenuhi syarat kedua. Kesulitan ini disebabkan karena varians dari setiap estimator yang konsisten akan cenderung menjadi nol apabila $n \rightarrow \infty$. Sehingga, apabila akan dibuat perbandingan diantara estimator-estimator yang konsisten, maka dipilih sebuah estimator yang variansnya lebih cepat mendekati nol. Secara asimptotik, estimator ini disebut estimator yang lebih efisien.

G. Deret Taylor

Teorema 2.3 (Dale Varberg and Edwin J. Purcell, 2010)

(Rumus Taylor dengan Sisa). Andaikan *f* suatu fungsi turunan ke *(n+1)*, $f^{(n+1)}(x)$, ada untuk setiap *x* pada suatu selang terbuka *I* yang mengandung *a*. Maka untuk setiap *x* di *I*

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^n(a)}{n!}(x-a)^n + R_n(x)$$

dengan sisa (galat) $R_n(x)$ diberikan rumus:

$$R_n(x) = \frac{R^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1}$$

dan *c* suatu titik antara *x* dan *a*.

Bukti Teorema (2.3)

$R_n(x)$ didefinisikan pada *I* oleh

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^n(a)}{n!}(x-a)^n + R_n(x)$$

x sebagai suatu konstanta dan didefinisikan oleh suatu fungsi baru *g* pada *I* oleh:

$$g(t) = f(x) - f(t) + f'(t)(x-t) + \frac{f''(t)}{2!}(x-t)^2 - \cdots - \frac{f^n(t)+1}{n!}(x-t)^n - R_n(x)\frac{(x-t)^{n+1}}{(x-a)^{n+1}}$$

Jika $g(t)$ diturunkan terhadap *t* (dengan *x* tetap), maka hasilnya adalah:

$$g(t) = -\frac{f^{(n+1)}(t)}{n!}(x-t)^n + R_n(x)(n+1)\frac{(x-t)^n}{(x-a)^{n+1}} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.17)$$

Jika $g'(c) = 0$, maka

$$g'(c) = -\frac{f^{(n+1)}(c)}{n!}(x-c)^n + R_n(x)(n+1)\frac{(x-c)^n}{(x-a)^{n+1}}$$

$$0 = -\frac{f^{(n+1)}(c)}{n!}(x-c)^n + R_n(x)(n+1)\frac{(x-c)^n}{(x-a)^{n+1}}$$

$$\Leftrightarrow R_n(x)(n+1)\frac{(x-c)^n}{(x-a)^{n+1}} = \frac{f^{(n+1)}(c)}{n!}(x-c)^n$$

$$\Leftrightarrow R_n(x) = \frac{f^{(n+1)}(c)(x-c)^n}{n!} \cdot \frac{(x-a)^{n+1}}{(n+1)(x-c)^n}$$

$$= \frac{f^{(n+1)}(c)}{(n+1)n!} \cdot (x-a)^{n+1}$$

$$= \frac{f^{(n+1)}(c)}{(n+1)!} \cdot (x-a)^{n+1}.$$

Teorema 2.3 terbukti.