PREDIKSI KASUS BERPENGARUH DALAM REGRESI LINIER

Oleh

ANISA

91 03 063

JURUSAN MATEMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

UJUNG PANDANG

1997

PREDIKSI KASUS BERPENGARUH DALAM REGRESI LINIER

S K R I P S I

Untuk melengkapi tugas-tugas dan
memenuhi syarat-syarat untuk
mencapai gelar sarjana.

OLEH

A N I S A

91 03 163

JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN
UJUNG PANDANG

1997

Kupersembahkan untuk :

Ibunda tercinta
Kakak-kakak tersayang
dan sahabat-sahabatku ...

PREDIKSI KASUS BERPENGARUH DALAM REGRESI LINIER

Disetujui oleh:

Pembimbing Utama

Drs. Alimin Bado, MS.
Nip. 130 604 514

Pembimbing Pertama

Drs. Nirwan Ilyas, MSi.
Nip. 131 658 823

Ujung Pandang, Mei 1997

KATA PENGANTAR

بسم الله الرحمن الرحيم

Assalamu Alaikum Wr. Wb.

Tak ada yang lebih patut untuk dilakukan selain memanjatkan rasa syukur kehadirat Allah Yang Maha Kuasa, karena atas berkah dan rahmat-Nya jualah sehingga penulisan skripsi dlngan judul "*Prediksi Kasus Berpengaruh Dalam Regresi Linier*" ini, dapat disusun sebagaimana adanya.

Melalui tulisan ini, penulis menghaturkan terima kasih yang setinggi-tingginya kepada Bapak Drs. Alimin Bado, MS selaku Pembimbing Utama dan Bapak Drs. Nirwan Ilyas, MSi selaku Pembimbing Pertama dan sekaligus Penasehat Akademik yang banyak memberikan petunjuk dan bimbingan serta nasehat-nasehat yang sangat berarti bagi penulis.

Juga tak lupa penulis haturkan terima kasih yang sebesar-besarnya kepada :

1. Bapak Ketua Jurusan Matematika, FMIPA Universitas Hasanuddin dan segenap Staf Dosen yang telah membimbing dan membekali ilmu kepada penulis.
2. Ibunda Hj. Sitti Nur yang tercinta dan segenap keluarga yang telah memberikan dukungan berupa moril dan materil serta kasih sayang dan doa yang tulus sejak lahir sampai saat ini.

3. Sahabatku Hj. Sitti Ratna, Sri Andayani, Sitti Nurlaela, dan Irma Yahya serta rekan-rekan mahasiswa Matematika Universitas Hasanuddin atas segala bantuannya selama ini.

4. Sahabatku Hamiyah, Widarni, A. Nurlinda, Rahmatullah, Syahrir, Hasdin dan Muliardi atas dukungan dan perhatiannya.

Semoga Allah SWT memberikan rahmat dan hidayah-Nya kepada mereka.

Penulis mengakui bahwa skripsi ini masih jauh dari kesempurnaan baik dalam hal isi maupun bentuk penyajiannya, oleh karena itu dengan tangan terbuka penulis menanti saran dan kritikan demi perbaikan penulisan dimasa mendatang.

Ujung Pandang, Mei 1997

Penulis

ABSTRAK

Yang dimaksud dengan *kasus* adalah satu unit data dari sekumpulan data yang diamati. Tulisan ini memusatkan perhatian pada masalah menilai pengaruh dan mendeteksi *kasus* tunggal yang berpengaruh dalam regresi linier. Suatu *kasus* dikatakan berpengaruh jika *kasus* dihapus dari data menyebabkan perubahan-perubahan yang cukup besar dalam taksiran dari β. Pengukuran besarnya pengaruh penghapusan *kasus* terhadap $\hat{\beta}$ dapat dilakukan dengan beberapa cara yaitu *Volume Elipsoida Kepercayaan*, *Diagnostik Andrews dan Pregibon*, dan *Peramalan Pengaruh*.

ABSTRACT

A case is one unit data from a group of data that is observed. This discussion will focus the problem on how to measure the influence and to detect single case that have influence in linear regression.

A case is called an influenced case if the case is removed or deleted from a group of data and causes large enough changes in β estimation. The measurement of influence of removing or deleting a case to $\hat{\beta}$ can be done in several ways, they are *Ellipsoid Confidence Volume*, *Andrews and Pregibon Diagnostic* and *Predictive Influence*.

DAFTAR ISI

Halaman

DAFTAR LAMBANG

LAMBANG	ARTI
β	beta (vektor parameter $p' \times 1$)
$\hat{\beta} = b$	taksiran dari β
ε_i	epsilon ke-i (vektor galat ukuran $n \times 1$
$\hat{\sigma} = S$	taksiran dari σ (simpangan baku)
$Sd(b)$	simpangan baku untuk b
H	matriks topi ukuran $n \times n$
I_n	matriks identitas ukuran $n \times n$
R	koefisien korelasi
X	matriks data ukuran $n \times p'$
y	vektor respon ukuran $n \times 1$
db	derajat bebas
log	logaritma dari
x_i	nomor kasus
m	jumlah kasus yang dihapus (dalam tulisan ini $m = 1$)
n	jumlah kasus seluruhnya
p'	jumlah parameter termasuk parameter konstannya
$\sim$	berdistribusi
$x \equiv y$	x ekivalen dengan y
$F(1-\alpha; p', n-p')$	Distribusi F dengan derajat bebas p' dan $n-p'$ dan taraf kepercayaan $1-\alpha$
$N(0, \sigma^2 I)$	Distribusi normal berdimensi n dengan mean o dan variansi $\sigma^2 I$

BAB I

PENDAHULUAN

A. Latar Belakang dan Alasan Memilih Judul

Suatu aspek penting dalam mencocokkan model regresi linier dengan kuadrat terkecil adalah menilai pengaruh atau pentingnya setiap *kasus* dalam kecocokan model. Pengertian *kasus* disini adalah satu unit data dari sekumpulan data yang diamati. Suatu *kasus* akan dinilai berpengaruh jika *kasus* tersebut dihapus terjadi perubahan yang cukup besar dalam hasil analisa, khususnya taksiran koefisien-koefisien dari model yang sesuai. Menemukan *kasus* yang demikian menjadi salah satu bagian penting dalam pemodelan yang berhubungan dengan statistik.

Pandang model umum untuk regresi linier $Y = X\beta + \varepsilon$, dengan Y merupakan vektor respon $nx1$, X merupakan matriks nxp' dengan rank p', β merupakan vektor parameter $p'x1$, dan ε merupakan vektor galat $nx1$ dan diasumsikan $\varepsilon \sim N(0, \sigma^2 I)$. Suatu unsur dari vektor Y akan menunjuk pada suatu pengamatan, dan suatu baris matriks X bersama pengamatan yang berkaitan akan menunjuk pada suatu *kasus*.

Gagasan utama dalam analisis pengaruh adalah memberikan sedikit gangguan (perturbasi) pada perumusan masalah dan kemudian memantau bagaimana perturbasi itu merubah hasil analisis. Dalam analisis pengaruh kita asumsikan bahwa modelnya sudah benar dan kita kaji kekekaran hasil yang diperoleh apabila diadakan perturbasi. Jika hasil tersebut dapat berubah apabila suatu *kasus* dihapus, maka kegunaan dari model dapat diragukan. Skema perturbasi yang diambil dalam tulisan ini adalah penghapusan *kasus* satu persatu. Berdasarkan hasil tersebut diatas, penulis bermaksud menuangkan dalam bentuk tulisan dengan judul :

"PREDIKSI KASUS BERPENGARUH DALAM REGRESI LINIER"

B. Perumusan Masalah

Adapun masalah yang akan dibahas disini adalah bagaimana menentukan pengaruh suatu *kasus* pada model regresi linier, dengan menggunakan penghapusan *kasus* satu persatu, dan dibatasi pada penentuan *kasus-kasus* tunggal yang berpengaruh.

C. Tujuan

Penulisan ini bertujuan untuk menentukan *kasus-kasus* berpengaruh dalam model regresi linier, sehingga hasil dari penentuan *kasus* tersebut dapat digunakan lebih lanjut oleh peneliti.

BAB II

SISA DAN MATRIKS TOPI

Pandang model umum regresi linier berikut :

$$Y = X\beta + \varepsilon \qquad (2.1)$$

dimana

Y merupakan vektor respon $nx1$

X merupakan matriks rank penuh nxp' $(n>p'=p+1)$

β merupakan vektor parameter $p'x1$

ε merupakan vektor galat $nx1$ dan diasumsikan

$\varepsilon \sim N(0, \sigma^2 I)$

Penaksir kuadrat terkecil untuk β diperoleh dengan meminimumkan $\varepsilon^T\varepsilon$, dimana :

$$\varepsilon^T\varepsilon = (Y-X\beta)^T(Y-X\beta) = (Y^T-\beta^TX^T)(Y-X\beta)$$

$$= Y^TY-\beta^TX^TY-Y^TX\beta+\beta^TX\ X\beta$$

$$= Y^TY-2\beta^TX^TY+\beta^TX^TX\beta \qquad (2.2)$$

karena β^TX^TY adalah suatu matriks berukuran $1x1$ atau suatu skalar,sehingga transposnya $(\beta^TX^TY)^T = Y^TX\beta$ mempunyai nilai yang sama.

Syarat perlu untuk meminimumkan (2.2) adalah :

$$\frac{d(\varepsilon^T\varepsilon)}{d\beta} = -2X^TY + 2X^TX\beta = 0 \qquad (2.3)$$

maka

$$X^TX\beta = X^TY \qquad (2.4)$$

Persamaan (2.3) ini disebut sebagai *Persamaan Normal*.

Karena X mempunyai rank penuh (rank nxp') maka X^TX non singulir (X^TX mempunyai invers) dan dengan mengganti β dengan $\hat{\beta}$, maka (2.4) mempunyai penyelesaian tunggal yaitu :

$$\hat{\beta} = b = (X^TX)^{-1}X^TY \qquad (2.5)$$

sehingga taksiran respon Y adalah $\hat{Y} = X\hat{\beta}$, dengan demikian maka taksiran untuk pengamatan ke-i dari n data adalah :

$$\hat{Y}_i = x_i^T\hat{\beta} \qquad (2.6)$$

dimana x_i^T menyatakan baris ke-i dari matriks X.

A. SISA

Sisa mempunyai peranan yang penting dalam diagnostik regresi karena sisa membawa informasi yang penting mengenai pendekatan terhadap asumsi. Dalam analisis, prosedur grafik atau plot memperlihatkan

sifat-sifat umum dari sisa yang sama baiknya dengan pengujian yang biasa digunakan untuk memeriksa kelayakan asumsi.

1. Sisa Biasa

Sisa biasa (selanjutnya disebut *sisa*) $\hat{\varepsilon} = e$ didefenisikan sebagai :

$$\begin{aligned} e &= Y - \hat{Y} \\ &= Y - X(X^T X)^{-1} X^T Y \\ &= Y - HY \\ &= (I - H)Y \end{aligned} \tag{2.7}$$

dimana

$$H = X(X^T X)^{-1} X^T \tag{2.8}$$

adalah matriks ukuran $n \times n$ dan disebut sebagai *matriks topi (hat matrix)*.

Hubungan antara e dan ε dapat diperoleh dengan mensubtitusikan $(X\beta + \varepsilon)$ pada Y dalam (2.7) yaitu :

$$\begin{aligned} e &= (I - H)(X\beta + \varepsilon) \\ &= X\beta + \varepsilon - HX\beta - H\varepsilon \\ &= \varepsilon - H\varepsilon \qquad \text{karena } H = X(X^T X)^{-1} X^T \\ &= (I - H)\varepsilon \end{aligned} \tag{2.9}$$

atau dalam bentuk skalar, untuk $i=1,2,\ldots,n$, maka :

$$e_i = \varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j \qquad (2.10)$$

Dari (2.10) terlihat bahwa hubungan antara e dan ε hanya bergantung pada H. Jika h_{ij} cukup kecil,maka e menaksir ε cukup baik, sebaliknya kegunaan dari e terbatas.

2. Sisa Terstudent Internal

Sisa terstudent internal (selanjutnya disebut *sisa terstudent*) didefenisikan sebagai :

$$r_i = \frac{e_i}{S\sqrt{1 - h_{ii}}} \qquad (2.11)$$

dimana $S^2 = \hat{\sigma}^2 = \dfrac{\Sigma e_i^2}{(n-p')}$ adalah taksiran dari rata-rata kuadrat sisa dan taksiran ini menggunakan semua data termasuk kasus ke-i, sedangkan h_{ii} adalah unsur diagonal ke-i dari matriks topi H.

r_i digunakan sebagai pengganti e_i dalam prosedur-prosedur grafik/plot. Untuk keperluan diagnostika, nilai r_i yang besar menunjukkan bahwa pengamatan ke-i kemungkinan

merupakan pencilan. *Belsley* memberikan kriteria bahwa pengamatan ke-i merupakan pencilan jika $|r_i| > 2$ kali simpangan bakunya.

3. Sisa Terstudent Eksternal

Sisa terstudent eksternal pada dasarnya analog dengan sisa terstudent, hanya saja pengamatan ke-i tidak diikutsertakan sehingga disebut sisa terstudent eksternal. Pendefenisian *Sisa terstudent eksternal* adalah sebagai berikut :

$$t_i = \frac{e_i}{S_{(i)}\sqrt{1 - h_{ii}}} \qquad (2.12)$$

dimana $S^2_{(i)} = \hat{\sigma}^2_{(i)}$ adalah taksiran dari σ^2 yang dihitung tanpa kasus ke-i, yang didefenisikan dengan

$$S^2_{(i)} = \frac{(n-\rho')S^2 - e_i^2/(1-h_{ii})}{(n - \rho' - 1)}$$

$$= \frac{(n-\rho')S^2 - r_i^2 S^2}{(n - \rho' - 1)}$$

$$= S^2 \left(\frac{n - \rho' - r_i^2}{n - \rho' - 1} \right) \qquad (2.13)$$

Distribusi dari t_i adalah *t-student* dengan derajat kebebasan $(n-p'-1)$. Hubungan antara t_i dan r_i dapat diperoleh dengan mensubtitusikan (2.13) kedalam (2.12) :

$$t_i = \frac{e_i}{S\left(\frac{n-p'-r_i^2}{n-p'-1}\right)^{1/2}\sqrt{1-h_{ii}}}$$

$$= r_i\left(\frac{n-p'-1}{n-p'-r_i^2}\right)^{1/2} \qquad (2.14)$$

yang menunjukkan bahwa t_i^2 merupakan transformasi monoton dari r_i^2.

B. MATRIKS TOPI

Dari (2.8) dapat ditunjukkan bahwa matriks topi tersebut bersifat *simetri* karena $H^T = H$, dan juga *idenpoten* karena $H^2 = H$. *John W.Tukey* memberi nama H dengan "**matriks topi**" karena H memetakan Y ke $\hat{Y}$, $\hat{Y} = HY$.

Unsur ke-(i,j) dari H adalah :

$$h_{ij} = x_i^T (X^T X)^{-1} x_j = h_{ji} \qquad (2.15)$$

Sedangkan unsur-unsur diagonal dari H adalah :

$$h_{ii} = x_i^T (X^T X)^{-1} x_i \qquad (2.16)$$

dimana x_i^T adalah baris ke-i dari X.

Karena H bersifat simetri dan idenpoten, maka :

$$h_{ij} = h_{ji} \quad \text{dan untuk } i = 1,2,\ldots,n$$

$$h_{ii} = \sum_{j=1}^{n} h_{ij}\, h_{ji} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2 \qquad (2.17)$$

Karena $\sum_{j \neq i} h_{ij}^2 \geq 0$ maka $h_{ii}^2 - h_{ii} \leq 0$, ini berakibat

$$0 \leq h_{ii} \leq 1 \qquad (2.18)$$

Misalkan *trace* $A \equiv tr\,(A)$ menyatakan jumlah unsur diagonal dari suatu matriks bujursangkar A, maka :

$$\Sigma\, h_{ii} = tr\,(H) = tr\,[X(X^TX)^{-1}X^T]$$

karena $tr\,(AB) = tr\,(BA)$, maka :

$$\Sigma\, h_{ii} = tr\,[(X^TX)^{-1}X^TX]$$

$$= tr\,(I_{p'}) = p' = rank\ X \qquad (2.19)$$

dimana p' menyatakan banyaknya parameter dalam model.

Besarnya unsur-unsur diagonal dari H memainkan peran yang penting dalam analisis kasus. Untuk model yang memuat parameter konstan, $h_{ii} \geq 1/n$, dimana $i=1,2,\ldots,n$. Maka (2.18) menjadi :

$$1/n \leq h_{ii} \leq 1 \qquad (2.20)$$

Kasus-kasus yang terpencil akan mempunyai nilai h_{ii} yang relatif besar. Nilai h_{ii} yang besar menunjukkan bahwa pengamatan ke-i mungkin sebagai *titik leverage tinggi*. *Belsley* memberikan kriteria bahwa pengamatan ke-i adalah *titik leverage tinggi*, jika $h_{ii} \geq 2p'/n$.

Berikut ini diberikan suatu teorema yang sangat membantu dalam perhitungan matriks topi selanjutnya yaitu teorema *Sherman-Morrison-Woodbury*. Sebelumnya perhatikan kedua lemma berikut ini.

Lemma 1 :

Untuk setiap matriks $(I+P)$ yang tak singulir maka berlaku :

$$(I+P)^{-1} = I - P\,(I+P)^{-1} \qquad (2.21)$$

$$= I - (I+P)^{-1}\,P \qquad (2.22)$$

Bukti :

$$(I+P)\,(I+P)^{-1} = (I+P)^{-1} + P\,(I+P)^{-1} = I$$

$$(I+P)^{-1}\,(I+P) = (I+P)^{-1} + (I+P)^{-1}\,P = I$$

sehingga :

$$(I+P)^{-1} = I - P\,(I+P)^{-1}$$

$$(I+P)^{-1} = I - (I+P)^{-1}\,P$$

Lemma 2 :

Bila $(I+PQ)$ dan $(I+QP)$ tak singulir maka :

$$P\,(I+QP)^{-1} = (I+PQ)^{-1}\,P \qquad (2.23)$$

Bukti :

$$(P+PQP) = P\,(I+QP) = (I+PQ)\,P$$

$$P = (I+PQ)^{-1}\,(I+PQ)\,P$$

$$P = (I+PQ)^{-1}\,P\,(I+QP)$$

dengan demikian maka :

$$P\,(I+QP)^{-1} = (I+PQ)^{-1}\,P$$

Teorema Sherman-Morrison-Woodbury :

Misalkan A dan D adalah suatu matriks tak singulir masing-masing berukuran m dan n serta B dan C matriks sebarang berukuran $m \times n$. Bila inversnya ada maka berlaku :

$$(A+BDC^{T})^{-1} = A^{-1} - A^{-1}B\,(D^{-1}+C^{T}A^{-1}B)^{-1}\,C^{T}A^{-1} \qquad (2.24)$$

Bukti :

Karena A tak singulir, maka $(A+BDC^T) = A\,(I+A^{-1}BDC^T)$.

Selanjutnya, $(A+BDC^T)^{-1} = (I+A^{-1}BDC^T)^{-1}\,A^{-1}$ (2.25)

Tulis $P = A^{-1}BDC^T$, maka :

$$(A+BDC^T)^{-1} = (I+P)^{-1}\,A^{-1}$$

Dari lemma 1, diperoleh :

$$(A+BDC^T)^{-1} = (I - (I+P)^{-1}\,P)\,A^{-1}$$

$$= (I - (I+A^{-1}BDC^T)^{-1}\,A^{-1}BDC^T)\,A^{-1}$$

$$= A^{-1} - (I+A^{-1}BDC^T)^{-1}\,A^{-1}BDC^TA^{-1} \qquad (2.26)$$

Tulis $P = A^{-1}$ dan $Q = BDC^T$

$$(A+BDC^T)^{-1} = P - (I+PQ)^{-1}\,PQP$$

dengan lemma 2, diperoleh :

$$(A+BDC^T)^{-1} = P - P(I+QP)^{-1}\,QP$$

$$(A+BDC^T)^{-1} = A^{-1} - A^{-1}\,(I+BDC^TA^{-1})^{-1}\,BDC^TA^{-1}$$

dengan menggunakan lemma 2 berulang kali, diperoleh :

$$(A+BDC^T)^{-1} = A^{-1} - A^{-1}\,(I+BDC^TA^{-1})^{-1}\,BDC^TA^{-1}$$

$$= A^{-1} - A^{-1}B\,(I+DC^TA^{-1}B)^{-1}\,DC^TA^{-1}$$

$$= A^{-1} - A^{-1}BD\,(I+C^TA^{-1}BD)^{-1}\,C^TA^{-1}$$

Akibat 1 :

Jika pada teorema diatas dipilih $A = X^T X$ berukuran $p \times p'$ dengan rank penuh, $D = I_n$, $B = -x_i$ dan $C^T = x_i^T$ (x_i kolom ke-i dari matriks X), maka berlaku :

$$\left(X_{(i)}^T X_{(i)}\right)^{-1} = (X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - x_i^T (X^T X)^{-1} x_i} \qquad (2.27)$$

Bukti :

Karena $X_{(i)}^T X_{(i)} = X^T X - x_i x_i^T$ berarti :

$$\left[X_{(i)}^T X_{(i)}\right]^{-1} = \left[X^T X - x_i x_i^T \right]^{-1}$$

$$= (X^T X)^{-1} - (X^T X)^{-1}(-x_i)[1 + x_i^T (X^T X)^{-1}(-x_i)]^{-1} x_i^T (X^T X)^{-1}$$

$$= (X^T X)^{-1} + (X^T X)^{-1} x_i [1 - x_i^T (X^T X)^{-1} x_i]^{-1} x_i^T (X^T X)^{-1}$$

Karena $x_i^T (X^T X)^{-1} x_i$ berukuran 1×1, maka :

$$[1 - x_i^T (X^T X)^{-1} x_i]^{-1} = \frac{1}{1 - x_i^T (X^T X)^{-1} x_i}$$

sehingga :

$$\left(X_{(i)}^T X_{(i)}\right)^{-1} = (X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - x_i^T (X^T X)^{-1} x_i}$$

$$\left(X_{(i)}^T X_{(i)}\right)^{-1} = (X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - x_i^T (X^T X)^{-1} x_i}$$

Misalkan h_{ij}^* adalah h_{ij} dengan tidak mengikut sertakan pengamatan ke-i pada perhitungan $(X^T X)^{-1}$, atau :

$$h_{ij}^* = x_i^T (X_{(i)}^T X_{(i)}) x_j$$

$$= x_i^T \left\{ (X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - x_i^T (X^T X)^{-1} x_i} \right\} x_j$$

$$= x_i^T (X^T X)^{-1} x_j + \frac{x_i^T (X^T X)^{-1} x_i x_i^T (X^T X)^{-1} x_j}{1 - x_i^T (X^T X)^{-1} x_i}$$

$$= h_{ij} + \frac{h_{ii} h_{ij}}{1 - h_{ii}}$$

$$= \frac{h_{ij}}{1 - h_{ii}} \qquad (2.28)$$

sehingga untuk $i = j$, diperoleh :

$$h_{ii}^* = \frac{h_{ii}}{1 - h_{ii}} \qquad (2.29)$$

BAB III

PREDIKSI KASUS BERPENGARUH DALAM REGRESI LINIER

Pada bab ini disajikan beberapa metode untuk mendeteksi seberapa besar pengaruh dari suatu kasus. Pertama, dengan membandingkan volume dari elipsoida-elipsoida kepercayaan yang didasarkan pada sampel penuh dan yang direduksi. Pengukuran yang kedua berhubungan dengan diagnostik Andrews dan Pregibon. Selanjutnya kita meninjau prosedur peramalan Bayes, dimana distribusi peramalan dari observasi-observasi mendatang akan dibandingkan.

A. PREDIKSI KASUS BERPENGARUH DENGAN VOLUME ELIPSOIDA KEPERCAYAAN

Salah satu pengukuran yang mungkin atas ketidakpastian pada penaksiran sebuah vektor dari parameter-parameter yaitu *volume elipsoida kepercayaan*.

Untuk mendapatkan pengukuran yang umum, susunlah X sedemikian hingga kolom terakhir $q \leq p'$ dari X merupakan koefisien-koefisien dari partisi $X = (X_1, X_2)$ dengan X_2

matriks ukuran $n \times q$. Defenisikan pula $C = (0 \ , \ I_q)$ sedemikian hingga $\varphi = C\beta$ adalah koefisien vektor. Elipsoida kepercayaan $(1-\alpha) \times 100\%$ untuk φ yang didasarkan pada $\hat{\varphi} = C\hat{\beta}$ adalah :

$$\varepsilon(\varphi) = \left\{ \varphi^* \mid (\varphi^* - \hat{\varphi})^T [C(X^TX)^{-1}C^T]^{-1} (\varphi^* - \hat{\varphi}) \leq q\hat{\sigma}^2 F(1-\alpha; q, n-p') \right\} \quad (3.1)$$

Jika sebuah subset dari m kasus yang berindeks I dihapus, maka elipsoidanya yang didasarkan pada $\hat{\varphi}_{(I)} = C\hat{\beta}_{(I)}$ adalah

$$\varepsilon_{(I)}(\varphi) = \left\{ \varphi^* \mid (\varphi^* - \hat{\varphi}_{(I)})^T [C(X_{(I)}^T X_{(I)})^{-1}C^T]^{-1} \times (\varphi^* - \hat{\varphi}_{(I)}) \leq q\hat{\sigma}^2_{(I)} F(1-\alpha; q, n-p'-m) \right\} \quad (3.2)$$

Volume dari kedua elipsoida tersebut adalah :

$$vol(\varepsilon(\varphi)) = (q\hat{\sigma}^2 F_q)^{q/2} \times (\det[C(X^TX)^{-1}C^T])^{1/2} \quad (3.3)$$

dan

$$vol(\varepsilon_{(I)}(\varphi)) = (q\hat{\sigma}^2_{(I)} F_q^m)^{q/2} \times (\det[C(X_{(I)}^T X_{(I)})^{-1}C^T])^{1/2} \quad (3.4)$$

dimana $F_q = F(1-\alpha; q, n-p')$ dan $F_q^m = F(1-\alpha; q, n-p'-m)$.

Logaritma perbandingan antara (3.4) dan (3.3) yang dinyatakan dengan $VR_I(\varphi)$ adalah :

$$VR_I(\varphi) = \log \left[\frac{vol\ (\varepsilon_{(I)}(\varphi))}{vol\ (\varepsilon(\varphi))} \right]$$

$$= \log \left[\frac{\left(q\hat{\sigma}^2_{(I)}\ F^m_q \right)^{q/2} \left\{ det\left[C(X^T_{(I)} X_{(I)})^{-1} C^T \right] \right\}^{1/2}}{\left(q\hat{\sigma}^2\ F_q \right)^{q/2} \left\{ det\left[C(X^T X)^{-1} C^T \right] \right\}^{1/2}} \right]$$

$$= \log \left[\left(\frac{q\hat{\sigma}^2_{(I)} F^m_q}{q\hat{\sigma}^2 F_q} \right)^q \frac{det\left[C(X^T_{(I)} X_{(I)})^{-1} C^T \right]}{det\left[C(X^T X)^{-1} C^T \right]} \right]^{1/2}$$

$$= 1/2 \log \left[\frac{det\left[C(X^T_{(I)} X_{(I)})^{-1} C^T \right]}{det\left[C(X^T X)^{-1} C^T \right]} \left(\frac{\hat{\sigma}^2_{(I)} F^m_q}{\hat{\sigma}^2 F_q} \right) \right] \qquad (3.5)$$

Dari (2.13) diperoleh hubungan antara $\sigma^2_{(I)}$ dan $\hat{\sigma}^2$ yaitu :

$$\frac{\hat{\sigma}^2_{(I)}}{\hat{\sigma}^2} = \frac{(n-p'-r^2_I)}{(n-p'-m)} \qquad (3.6)$$

Misalkan $det\ A = |A|$ menyatakan determinan dari matriks A, $|A| = \frac{1}{|A^{-1}|}$ dan $|AB| = |A|\ |B|$, maka :

$$\frac{det\left[\,C(X^T_{(I)}X_{(I)})^{-1}C^T\right]}{det\left[\,C(X^TX)^{-1}C^T\right]} = \frac{|C|\;|(X^T_{(I)}X_{(I)})^{-1}|\;|C^T|}{|C|\;|(X^TX)^{-1}|\;|C^T|}$$

$$= \frac{|C|\;|C^T|}{|X^T_{(I)}X_{(I)}|}\;\frac{|X^TX|}{|C|\;|C^T|} \qquad (3.7)$$

Karena

$$|X^T_{(I)}X_{(I)}| = |X^TX - X_I^TX_I|$$

$$= |X^TX|\;|I - X_I(X^TX)^{-1}X_I^T|$$

$$= |X^TX|\;|I - H_I| \qquad (3.8)$$

maka (3.7) menjadi :

$$\frac{det\left[\,C(X^T_{(I)}X_{(I)})^{-1}C^T\right]}{det\left[\,C(X^TX)^{-1}C^T\right]} = \frac{|C|\;|C^T|}{|X^TX|\;|I-H_I|}\;\frac{|X^TX|}{|C|\;|C^T|}$$

$$= \frac{|I-U_I|}{|I-H_I|} \qquad (3.9)$$

dimana $U = X_I(X_I^TX_I)^{-1}X_I^T$ dan U_I dan H_I adalah submatriks $m \times m$ dari U dan H.

Dengan menggabungkan hasil-hasil tersebut dan menyederhanakannya, maka (3.5) menjadi :

$$VR_I(\rho) = 1/2 \log \left[\frac{|I-U_I|}{|I-H_I|} \left(\frac{n-\rho'-r_I^2}{n-\rho'-m} \times \frac{F_q^m}{F_q} \right)^q \right]$$

$$= 1/2 \log |I-U_I| - 1/2 \log |I-H_I|$$

$$+ 1/2 \log \left(\frac{n-\rho'-r_I^2}{n-\rho'-m} \times \frac{F_q^m}{F_q} \right)^q$$

$$= -1/2 \log |I-H_I| + 1/2 \log |I-U_I|$$

$$- q/2 \log \left(\frac{n-\rho'-m}{n-\rho'-r_I^2} \times \frac{F_q}{F_q^m} \right) \qquad (3.10)$$

Karena dalam penulisan ini dibatasi pada penentuan kasus tunggal yang berpengaruh, maka $m = 1$. Jika irisannya termasuk dalam model, maka $q = \rho'$, $C = I$, $|I-H_I| = 1-h_{ii}$, $|I-U_I| = 1$, maka (3.10) menjadi :

$$VR_i = \log \left[\frac{vol\ (\varepsilon_{(i)})}{vol\ (\varepsilon)} \right]$$

$$= -(1/2) \log (1-h_{ii}) + (1/2) \log 1$$

$$- (\rho'/2) \log \left[\frac{n-\rho'-1}{n-\rho'-r_i^2} \frac{F_{p'}}{F_{p'}^1} \right] \qquad (3.11)$$

Sebagai alternatifnya, jika irisannya diabaikan, maka

$C = (0, I_p)$, $|I - H_I| = 1 - h_{ii}$, $|I - U_I| = 1 - 1/n$, maka (3.10)

menjadi :

$$VR_i' = \log \left[\frac{vol\ (\varepsilon_{(i)}(\varphi))}{vol\ (\varepsilon(\varphi))} \right]$$

$$= -(1/2) \log(1 - h_{ii}) + (1/2) \log (1 - 1/n)$$

$$- (p/2) \log \left[\frac{n - p' - 1}{n - p' - r_i^2} \; \frac{F_p}{F_p^1} \right] \qquad (3.12)$$

Nilai pengukuran volume logaritma dapat bernilai negatif atau positif. Pengukuran negatif berarti bahwa penghapusan kasus akan menurunkan volume dan karenanya meningkatkan ketelitian. Hal ini terjadi jika r_i^2 besar tetapi h_{ii} kecil. Nilai positif dari rasio ini menyatakan bahwa volumenya lebih besar untuk data yang direduksi dan ketelitiannya menjadi menurun. Hal ini akan terjadi jika h_{ii} besar. Pengukuran volume ini menyeimbangkan antara efek-efek sisa dan potensial.

B. PREDIKSI KASUS BERPENGARUH MENGGUNAKAN DIAGNOSTIK ANDREWS DAN PREGIBON

Sebuah metode alternatif untuk mendeteksi kasus berpengaruh pada regresi linier diberikan oleh *Andrews dan Pregibon*.

Sebagai awal, ambil pengaruh dari sebuah pencilan pada Y dan baris pencilan dari X secara terpisah. Pertama, penghapusan dari sebuah kasus pencilan pada Y akan cenderung berakibat pada reduksi bertanda pada jumlahan kuadrat sisa. Karenanya jumlahan kuadrat sisa merupakan diagnostik untuk mendeteksi kasus berpengaruh yang timbul karena sebuah pencilan dalam Y. Kedua, pengaruh dari baris ke-X paling sedikit direfleksikan oleh perubahan pada $|X^TX|$ jika baris tersebut dihapus. Jika $|X^TX|$ sangat berubah jika x_i dihapus, maka kasus ke-(y_i, x_i^T) akan mempunyai pengaruh yang besar pada $\hat{\beta}$.

Karena adanya dua macam diagnostik yaitu sisa dan matriks H yang terpisah dan sering memberikan keputusan yang bertentangan, maka Andrews dan Pregibon (1978) menyederhanakan dan menformulasikan diagnostik tunggal yang didasarkan pada sisa dan matriks H, yaitu :

$$R_I = \frac{(n-\rho'-m)\,\hat{\sigma}^2_{(I)}\,|X^T_{(I)}X_{(I)}|}{(n-\rho')\,\hat{\sigma}^2\,|X^TX|}$$

$$= \frac{(n-\rho'-m)}{(n-\rho')} \times \frac{(n-\rho'-r_I^2)}{(n-\rho'-m)}\,|I-H_I|$$

$$= \frac{(n-\rho'-r_I^2)}{(n-\rho')}\,|I-H_I|$$

$$= \left\{ 1 - \frac{r_I^2}{(n-\rho')} \right\} |I-H_I| \qquad (3.13)$$

sebagai pengukuran dari pengaruh gabungan kasus-kasus berindeks I.

Untuk tujuan perbandingan, maka tepat sekali untuk mengambil $-(1/2)\log R_I$, yaitu :

$$AP_I = -(1/2)\log R_I = -(1/2)\log\left[\frac{(n-\rho'-r_I^2)}{(n-\rho')}\,|I-H_I|\right]$$

$$= -(1/2)\log|I-H_I| - (1/2)\log\left(\frac{n-\rho'-r_I^2}{n-\rho'}\right)$$

$$= -(1/2)\log|I-H_I| + (1/2)\log\left(\frac{n-\rho'}{n-\rho'-r_I^2}\right) \qquad (3.14)$$

Karena $m = 1$, maka (3.14) menjadi :

$$AP_i = -(1/2) \log (1-h_{ii}) + (1/2) \log \left(\frac{n-\rho'}{n-\rho'-r_i^2} \right) \qquad (3.15)$$

Statistik ini akan besar untuk kasus-kasus berpengaruh dan dapat dibandingkan dengan rasio volume yang didasarkan pada elipsoida berdimensi-ρ' dari (3.7). Kedua statistik ini berbeda pada tanda dan bobot relatif dari kedua bentuk, dan oleh sebuah faktor $\frac{-(1)}{(n-\rho'-r_i^2)}$ pada logaritma kedua. Jika $(n-\rho')$ cukup besar untuk mengabaikan faktor terakhir ini, maka statistik-statistik tersebut menggunakan informasi yang sama tetapi menggabungkan dengan cara yang berbeda.

C. PREDIKSI KASUS BERPENGARUH MENGGUNAKAN PERAMALAN PENGARUH

Pada bagian ini, kita menggunakan metode Bayes untuk menaksir pengaruh kasus-kasus pada peramalan observasi-observasi dimasa mendatang. Metode ini dikembangkan oleh *Johnson dan Geisser*, dengan menggunakan

divergensi Kullback-Leibler untuk mengukur selisih antara kepadatan ramalan didasarkan pada himpunan data penuh dan data yang direduksi. Pertama-tama diasumsikan bahwa σ^2 diketahui dan akhirnya perluasan metodologi pada situasi yang lebih umum dimana σ^2 tidak diketahui.

1. Fungsi Ramalan Pengaruh Dengan σ^2 Diketahui

Misalkan Y menyatakan sebuah vektor-n dari variabel acak yang dapat dinyatakan dengan model linier (2.1) dan diasumsikan bahwa galat-galat ε mengikuti distribusi normal berdimensi-n, $N_n(0,\sigma^2 I)$. Diberikan nilai pengamatan y dari Y, yang bertujuan untuk meramalkan Y_f yang menyatakan sebuah vektor berdimensi-q dari observasi yang akan datang dan dinyatakan oleh model linier :

$$Y_f = X_f\beta + \varepsilon_f$$

dimana ε_f berdistribusi $N(0, \sigma^2 I)$, X_f adalah matriks $q \times p'$ dan β sebagaimana pada (2.1) adalah vektor parameter $p' \times 1$.

Diasumsikan bahwa improper prior untuk β adalah $p(\beta)d\beta \propto d\beta$. Kepadatan posterior $p(\beta|y)$ untuk β yang diberikan oleh $Y = y$ adalah :

$$p(\beta|y) = \frac{f(y|X\beta,\ \sigma^2 I)\ p(\beta)}{\int f(y|X\beta,\ \sigma^2 I)\ p(\beta)d\beta} \qquad (3.16)$$

Kepadatan ramalan untuk Y_f yang diberikan oleh y, X, X_f dan σ^2 adalah :

$$\int f(y_f|X_f\beta,\ \sigma^2 I)\ p(\beta|y)d\beta = f(y_f|X_f\hat{\beta},\ \sigma^2(I+X_f(X^TX)^{-1}X_f^T)) \qquad (3.17)$$

$N_q(X_f\hat{\beta},\ \sigma^2(I+X_f(X^TX)^{-1}X_f^T))$, dari (3.17) adalah kepadatan ramalan dari observasi-observasi mendatang, dan kepadatan ramalan untuk data yang direduksi adalah $N_q(X_f\hat{\beta}_{(I)},\ \sigma^2(I+X_f(X_{(I)}^TX_{(I)})^{-1}X_f^T))$.

Jika $f_1 = N_n(\mu_1,\Sigma_1)$ dan $f_2 = N_n(\mu_2,\Sigma_2)$ dan diasumsikan bahwa Σ_1 dan Σ_2 adalah definit positif, maka *divergensi Kullback-Leibler* adalah :

$$2d(f_1,f_2) = (\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2) + \log\left(|\Sigma_2|/|\Sigma_1|\right) + tr(\Sigma_1\Sigma_2^{-1}) - n \qquad (3.18)$$

dimana f_1 adalah kepadatan ramalan untuk data yang direduksi dan f_2 adalah kepadatan ramalan untuk data penuh.

Bentuk pertama dari ruas kanan (3.18) menyatakan jarak antara pusat-pusat f_1 dan f_2 yang relatif terhadap

kontour-kontour dari kepadatan konstan f_2. Bentuk kedua membandingkan volume elipsoida yang didasarkan pada dua distribusi dan akan sama dengan nol hanya jika volumenya sama. Bentuk ketiga yaitu $tr\,(\Sigma_1\Sigma_2^{-1})$ dipandang sebagai bentuk sisa yang membandingkan struktur-struktur eigen dari Σ_1 terhadap Σ_2. Metode ini lebih luas dibanding yang lain, karena menggabungkan beberapa aspek analisis kedalam sebuah pengukuran tunggal.

$d(f_1, f_2)$ selanjutnya disebut sebagai *fungsi ramalan pengaruh* (*Predictive Influence Function* atau *PIF*).

Johnson dan Geisser menyarankan menggunakan X sebagai pengganti X_f, agar PIF dapat digunakan sebagai diagnostik, mengingat sulitnya untuk menentukan X_f.

Jika $X_f = X$, maka dapat dituliskan d_I untuk $d(f_1, f_2)$, dimana $f_2 = N_n\left(X\hat{\beta},\ \sigma^2\,(I+X(X^TX)^{-1}X^T)\right) = N_n\left(X\hat{\beta},\ \sigma^2\,(I+H)\right)$ dan $f_1 = N_n\left(X\hat{\beta}_{(I)},\ \sigma^2\,(I+X(X_{(I)}^TX_{(I)})^{-1}X^T)\right)$.

Untuk mendapatkan bentuk sederhana dari d_I, maka subtitusikan f_1 dan f_2 kedalam bentuk (3.18) satu persatu. Pertama, perubahan pada pusat adalah :

$$(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2) = \left(X\hat{\beta}_{(I)}-X\hat{\beta}\right)^T\left(\sigma^2\,(I+H)\right)^{-1}\left(X\hat{\beta}_{(I)}-X\hat{\beta}\right)$$

$$(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2) = \left[X(\hat{\beta}_{(I)}-\hat{\beta})\right]^T\left[(I+H)^{-1}/\sigma^2\right]\left[X(\hat{\beta}_{(I)}-\hat{\beta})\right]$$

$$= (\hat{\beta}_{(I)}-\hat{\beta})^T X^T\left[(I+H)^{-1}/\sigma^2\right) X(\hat{\beta}_{(I)}-\hat{\beta})$$

$$= (\hat{\beta}_{(I)}-\hat{\beta})^T X^T\left[(I-(1/2)H)/\sigma^2\right) X(\hat{\beta}_{(I)}-\hat{\beta})$$

$$= (\hat{\beta}_{(I)}-\hat{\beta})^T\left[(X^TX-(1/2)X^THX)/\sigma^2\right)(\hat{\beta}_{(I)}-\hat{\beta})$$

$$= (\hat{\beta}_{(I)}-\hat{\beta})^T\left[(X^TX-(1/2)X^TX)/\sigma^2\right)\ (\hat{\beta}_{(I)}-\hat{\beta})$$

$$= (\hat{\beta}_{(I)}-\hat{\beta})^T\left(X^TX/2\sigma^2\right)(\hat{\beta}_{(I)}-\hat{\beta}) \qquad (3.19)$$

Persamaan (3.19) tidak lain adalah *Jarak Cook* $D_I(X^TX, 2\sigma^2)$, yang merupakan jarak dari β ke $\hat{\beta}_{(I)}$ relatif terhadap X^TX, dengan σ^2 sebagai pengganti dari $\hat{\sigma}^2$.

Berikutnya perubahan volume diukur dengan :

$$\log\left(\ |\Sigma_2|\ /\ |\Sigma_1|\ \right) = \log|\Sigma_2| - \log|\Sigma_1|$$

$$= \log|I+H| - \log\left|I+X(X_{(I)}^TX_{(I)})^{-1}X^T\right|$$

Karena H adalah matriks simetri dan idenpoten dengan rank ρ', maka $I+H$ juga matriks simetri dan idenpoten dengan rank ρ', sehingga $I+H$ dapat didiagonalisir (dijadikan matriks diagonal) dengan nilai-nilai eigen pada diagonal utama yaitu $2^{\rho'}$ dan $1^{n-\rho'}$, sehingga $|I+H| = 2^{\rho'}$.

Sedangkan bentuk $|I+X(X^{T}_{(I)}X_{(I)})^{-1}X^{T}|$ dapat disederhanakan dengan melihat kembali (2.29) yaitu :

$$|I+X(X^{T}_{(I)}X_{(I)})^{-1}X^{T}| = 2^{p'}|I+(1/2)H_{I}(I-H_{I})^{-1}|$$

maka

$$\log\left(|\Sigma_2| / |\Sigma_1|\right) = -\log|I+(1/2)H_{I}(I-H_{I})^{-1}| \qquad (3.20)$$

Bentuk akhir dari d_I adalah :

$$tr\left(\Sigma_1\Sigma_2^{-1}\right) = tr\left[\left(I+X(X'_{(I)}X_{(I)})^{-1}X'\right)\left(I+H\right)^{-1}\right]$$

$$= tr\left[\left(I+X(X'_{(I)}X_{(I)})^{-1}X'\right)\left(I-(1/2)H\right)\right]$$

$$= tr\left[I+X(X'_{(I)}X_{(I)})^{-1}X'-(1/2)X(X'_{(I)}X_{(I)})^{-1}X'H - 1\ 2)H\right]$$

$$= tr\left[I+(1/2)X(X'_{(I)}X_{(I)})^{-1}X'-(1/2)H\right]$$

$$= n-(p'/2)+(1/2)\,tr\left[X(X'_{(I)}X_{(I)})^{-1}X'\right]$$

$$= n-(p'/2)+(1/2)\,tr\left[XX'(X'_{(I)}X_{(I)})^{-1}\right]$$

$$= n+(1/2)\,tr\left[H_{I}(I-H_{I})^{-1}\right] \qquad (3.21)$$

Dari (3.19), (3.20) dan (3.21) maka d_I dapat dinyatakan sebagai :

$$d_I = D_I(X^T X, 4\sigma^2) - (1/2)\log\left| I + (1/2) H_I (I-H_I)^{-1} \right|$$

$$+ (1/4)\, tr\left[H_I (I-H_I)^{-1} \right] \qquad (3.22)$$

dimana PIF d_I hanya bergantung pada e_I dan H_I. Pendekatan ramalan memanfaatkan pembentukan blok yang sama seperti sebelumnya, tetapi perbedaan utamanya terletak pada bagaimana pendekatan ramalan menggabungkan informasi-informasi ini untuk menghasilkan satu pengukuran menyeluruh yang umum.

2. Fungsi Ramalan Pengaruh Dengan σ^2 Tidak Diketahui

Jika σ^2 tidak diketahui, maka kepadatan ramalan merupakan kepadatan student multivariat daripada kepadatan normal multivariat. Jika $S_n(v,\mu,\Sigma)$ menyatakan sebuah kepadatan student berdimensi-n, v derajat bebas, lokasi parameter μ dan matriks dispersi Σ, maka kepadatan ramalan berdasarkan himpunan data penuh dan data yang dikurangi dengan mengatur $X_f = X$ adalah :

$$S_n\left(n-p' , X\hat{\beta} , \hat{\sigma}^2(I+H) \right)$$

dan

$$S_n\left(n-m-p' , X\hat{\beta}_{(I)} , \hat{\sigma}^2_{(I)}(I+X(X^T_{(I)} X_{(I)})^{-1}X^T) \right)$$

Karena PIF yang didasarkan kepadatan ini cukup sulit untuk dipelajari, maka kepadatan normal digunakan untuk mendekati kepadatan student, dimana kepadatan ramalan untuk data penuh adalah $f = N_n\left(X\hat{\beta}, ((n-\rho')/(n-\rho'-2))\hat{\sigma}^2(I+H)\right)$ dan kepadatan ramalan untuk data yang direduksi didefenisikan sebagai $f_{(I)} = N_n\left(X\hat{\beta}_{(I)}, ((n-m-\rho')/(n-m-\rho'-2))\right.$ $\left.\hat{\sigma}^2_{(I)}(I+X(X^T_{(I)}X_{(I)})^{-1}X^T)\right)$.

Bentuk sederhana dari $\tilde{d}_I$ diperoleh dengan mensubtitusikan f_1 dan f_2 kedalam (3.18) satu persatu.

Pertama, perubahan pada pusat adalah :

$$(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2) = \left(X\hat{\beta}_{(I)}-X\hat{\beta}\right)^T\left(\frac{(n-\rho')}{(n-\rho'-2)}\hat{\sigma}^2(I+H)\right)^{-1}\left(X\hat{\beta}_{(I)}-X\hat{\beta}\right)$$

$$= \left(\frac{n-\rho'-2}{n-\rho'}\right)\left(X\hat{\beta}_{(I)}-X\hat{\beta}\right)^T\left(\hat{\sigma}^2(I+H)\right)^{-1}\left(X\hat{\beta}_{(I)}-X\hat{\beta}\right)$$

$$= \left(\frac{n-\rho'-2}{n-\rho'}\right) D_I(X^TX\,,\,4\hat{\sigma}^2) \qquad (3.23)$$

karena tiga suku terakhir sama dengan yang terdapat pada (3.19).

Bentuk yang kedua adalah :

$$\log\left(|\Sigma_2|/|\Sigma_1|\right) = \log|\Sigma_2| - \log|\Sigma_1|$$

$$= \log\left|\frac{(n-\rho')}{(n-\rho'-2)}(I+H)\right|$$

$$- \log\left|\frac{(n-m-\rho')}{(n-m-\rho'-2)}(I+X(X^{T}_{(I)}X_{(I)})^{-1}X^{T}\right|$$

Dengan melihat kembali persamaan (3.20), maka bentuk ini dapat disederhanakan menjadi :

$$\log\left(|\Sigma_2|/|\Sigma_1|\right) = (-1/2)\log\left|\, I + (1/2)\, H_I(I-H_I)^{-1}\,\right|$$

$$+ (n/2)\left(k_I - \log(k_I) - 1\right) \qquad (3.24)$$

dimana

$$k_I = \left(\frac{n-\rho'-2}{n-m-\rho'-2}\right)\left(1 - \frac{r_I^2}{n-\rho'}\right) \qquad (3.25)$$

Jika (3.6) dan (3.21) disubtitusikan kedalam bentuk terakhir ini, maka diperoleh :

$$tr\left(\Sigma_1\Sigma_2^{-1}\right) = (1/4)k_I\, tr\left(H_I(I-H_I)^{-1}\right) \qquad (3.26)$$

Dari persamaan (3.23), (3.24), (3.25) dan (3.26), maka $\tilde{d}_I$ dapat ditulis sebagai :

$$\tilde{d}_I = \left(\frac{n-\rho'-2}{n-\rho'}\right) D_I(X^{T}X, 4\hat{\sigma}^2) + (1/4)\, k_I\, tr\left[H_I(I-H_I)^{-1}\right]$$

$$- (1/2)\log\left|I+(1/2)H_I(I-H_I)^{-1}\right| + (n/2)\left(k_I - \log(k_I) - 1\right)$$

(3.27)

Terpisah dari konstanta-konstanta, maka selisih antara $\tilde{d}_I$ dan d_I terletak pada keberadaan k_I dalam pengukuran sebelumnya. Karena k_I adalah fungsi menurun dari r_I^2, maka k_I akan kecil jika kasus-kasus berindeks I tidak sesuai dengan model yang diasumsikan.

Karena $m = 1$, maka persamaan (3.24) memberikan

$$\tilde{d}_i = \frac{p'(n-p'-2)}{4(n-p')} D_i(X^TX, 4\hat{\sigma}^2) + (k_i/4)\left[h_{ii}/(1-h_{ii})\right] - (1/2)\log\left(1+(1/2)h_{ii}/(1-h_{ii})\right) + (n/2)\left(k_i - \log(k_i) - 1\right) \quad (3.28)$$

dan (3.25) menjadi :

$$k_i = \left(\frac{n-p'-2}{n-m-p'-2}\right)\left(1 - \frac{r_i^2}{n-p'}\right).$$

Jadi $\tilde{d}_i$ hanya bergantung pada n, p', r_i^2 dan h_{ii}, dan akan sensitif terhadap pemindahan kasus-kasus dengan nilai r_i^2 yang besar.

BAB IV

P E N E R A P A N

A. Model dan Hasil-Hasil Regresi

Pada bagian ini diambil sekumpulan data yang diperoleh dari catatan kesehatan karyawan anggota tetap dari klub kesehatan suatu perusahaan. Model yang digunakan adalah sebagai berikut :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \qquad (4.1)$$

dimana

X_1 = berat badan dalam pon

X_2 = laju nadi (istirahat) per menit

X_3 = kekuatan lengan dan kaki (berat angkatan seorang karyawan) dalam pon

X_4 = waktu dalam uji coba lari 1/4 mil dalam detik

Y = waktu lari sejauh 1 mil dalam detik

Data lengkap dapat dilihat pada Tabel 4.1. Perhitungan dengan metode kuadrat terkecil menghasilkan suatu persamaan taksiran dari model regresi linier sebagai berikut :

$$\hat{Y} = -3,62 + 1,27X_1 - 0,525X_2 - 0,505X_3 + 3,90X_4$$

Tabel 4.1. Data Klub Kesehatan

Kasus	X_1	X_2	X_3	X_4	Y
1	217	67	260	91	481
2	141	52	190	66	292
3	152	58	203	68	338
4	153	56	183	70	357
5	180	66	170	77	396
6	193	71	178	82	429
7	162	65	160	74	345
8	180	80	170	84	469
9	205	77	188	83	425
10	168	74	170	79	358
11	232	65	220	72	393
12	146	68	158	68	346
13	173	51	243	56	279
14	155	64	198	59	311
15	212	66	220	77	401
16	138	70	180	62	267
17	147	54	150	75	404
18	197	76	228	88	442
19	165	59	188	70	368
20	125	58	160	66	295
21	161	52	190	69	391
22	132	62	163	59	264
23	257	64	313	96	487
24	236	72	225	84	481
25	149	57	173	68	374
26	161	57	173	65	309
27	198	59	220	62	367
28	245	70	218	69	469
29	141	63	193	60	252
30	177	53	183	75	338

Sumber : Chatterje dan Hadi, "Sensitivity Analysis in Linear Regression," 1988.

Tabel 4.2 memperlihatkan hasil hitung parameter regresi dari data klub kesehatan perusahaan.

Tabel 4.2. Hasil Hitung Parameter Regresi, Data Klub Kesehatan

variabel	*b*	*sd(b)*
konstanta	-3,620	56,100
X_1	1,268	0,287
X_2	-0,525	0,863
X_9	-0,505	0,246
X_4	3,903	0,748
S = 28,67	R^2 = 85,3%	F = 36,3	*db* = (5,25)

Sumber : Data diolah dengan MINITAB

B. Pemeriksaan Sisa dan Leverage

Sisa biasa e_i dan sisa terstudent r_i dan t_i yang dihasilkan dalam kecocokan model linier, ditunjukkan dalam Tabel 4.3. Dari tabel tampak bahwa kasus 30 mempunyai sisa cukup besar, r_{30} = -2,325, melebihi kriteria 2σ = 2,088. Dari kenyataan ini, maka kasus 30 dicurigai sebagai pencilan.

Nilai-nilai leverage h_{ii}, ditunjukkan dalam Tabel 4.4. Dari tabel dapat dilihat bahwa kasus 23 dan 28 merupakan titik dengan leverage tinggi, karena $h_{23,23}$ = 0,513, $h_{28,28}$ = 0,387, melebihi kriteria $2p'/n$ = 0,333.

Tabel 4.3. Sisa Biasa dan Sisa Terstudent, Data Klub Kesehatan

Kasus	e_i	r_i	t_i
1	20,863	0,835	0,847
2	-17,452	-0,643	-0,649
3	16,515	0,596	0,601
4	15,290	0,548	0,551
5	- 8,570	-0,308	-0,308
6	- 4,898	-0,178	-0,178
7	-30,619	-1,122	-1,152
8	44,462	1,838	1,983
9	-19,811	-0,741	-0,750
10	-34,963	-1,326	-1,377
11	-33,246	-1,327	-1,379
12	14,647	0,530	0,534
13	- 5,745	-0,235	-0,235
14	21,465	0,812	0,823
15	-18,883	-0,676	-0,682
16	-18,633	-0,734	-0,743
17	32,665	1,334	1,387
18	7,490	0,292	0,292
19	15,180	0,534	0,538
20	- 6,169	-0,224	-0,224
21	44,487	1,717	1,833
22	-15,105	-0,554	-0,558
23	-18,167	-0,905	-0,921
24	9,049	0,335	0,336
25	40,642	1,509	1,586
26	-27,861	-1,016	-1,039
27	19,733	0,747	0,756
28	39,600	1,848	1,996
29	-26,741	-1,021	-1,043
30	-55,224	-2,325	-2,642

Sumber : Data diolah dengan MINITAB

Tabel 4.4. Statistik-Statistik Pengaruh, Data Klub Kesehatan

Kasus	r_i	h_{ii}	D_i	VR_i	AP_i	$\bar{d}_i$
1	0,835	0,249	0,047	0,084	0,068	0,244
2	-0,643	0,126	0,012	0,064	0,033	0,277
3	0,596	0,090	0,007	0,058	0,024	0,286
4	0,548	0,079	0,005	0,057	0,020	0,302
5	-0,308	0,090	0,002	0,069	0,021	0,375
6	-0,178	0,110	0,001	0,077	0,026	0,401
7	-1,122	0,084	0,023	0,016	0,030	-0,021
8	1,838	0,221	0,175	-0,051	0,086	-0,545
9	-0,741	0,147	0,019	0,063	0,393	0,241
10	-1,326	0,129	0,050	-0,003	0,046	-0,157
11	-1,327	0,214	0,093	0,025	0,068	-0,092
12	0,530	0,099	0,006	0,063	0,025	0,314
13	-0,235	0,298	0,005	0,127	0,077	0,446
14	0,812	0,162	0,026	0,062	0,044	0,211
15	-0,676	0,070	0,007	0,048	0,020	0,245
16	-0,734	0,231	0,033	0,086	0,062	0,281
17	1,334	0,248	0,114	0,034	0,078	-0,065
18	0,292	0,227	0,005	0,105	0,057	0,414
19	0,534	0,047	0,003	0,051	0,013	0,299
20	-0,224	0,114	0,001	0,077	0,027	0,395
21	1,717	0,119	0,074	-0,056	0,055	-0,537
22	-0,554	0,122	0,009	0,067	0,031	0,311
23	-0,905	*0,513*	0,174	*0,173*	*0,164*	0,470
24	0,335	0,143	0,004	0,081	0,035	0,382
25	1,509	0,072	0,034	-0,035	0,037	-0,365
26	-1,016	0,084	0,019	0,026	0,028	0,054
27	0,747	0,167	0,023	0,068	0,045	0,246
28	1,848	*0,388*	*0,394*	-0,000	0,138	-0,257
29	-1,021	0,163	0,041	0,045	0,048	0,092
30	*-2,325*	0,193	0,220	*-0,165*	0,099	*-1,135*

Sumber : Data diolah dengan MINITAB dan LOTUS 123

C. Pemeriksaan Pengaruh

Ukuran-ukuran pengaruh ditunjukkan dalam Tabel 4.4. Dari tabel dapat dilihat bahwa kasus 28 adalah kasus yang berpengaruh berdasarkan *jarak Cook*, karena D_{28} = 0,394. Tetapi berdasar pada *volume elipsoida kepercayaan*, dengan α = 0,05, maka kasus yang berpengaruh adalah kasus 23 dan kasus 30, karena VR_{23} = 0,173 dan VR_{30} = -0,165, sedangkan berdasarkan pada *diagnostik Andrews dan Pregibon*, kasus 23 adalah kasus yang berpengaruh karena AP_{23} = 0,164. Berdasarkan pada *peramalan pengaruh*, kasus yang berpengaruh adalah kasus 30, karena $\bar{d}_{30}$ = -1,135 akibat dari r^2_{30} = 5,406.

Dari hasil pemeriksaan diatas diperoleh kasus-kasus dengan sisa besar, leverage tinggi dan yang berpengaruh yaitu kasus 23, 28, dan kasus 30.

Selanjutnya akan diperiksa perubahan-perubahan pada perhitungan kuadrat terkecil jika masing-masing kasus tersebut dihapus dari data satu persatu. Hasil selengkapnya dapat kita lihat pada Tabel 4.5 dan Tabel 4.6.

Tabel 4.5 menyajikan ringkasan regresi data klub kesehatan dengan penghapusan kasus 23,28,30 dan tanpa penghapusan kasus-kasus tersebut. Sedangkan pada Tabel 4.6 disajikan statistik-statistik kasus yang diperoleh dari penghapusan kasus-kasus tersebut satu persatu.

Tabel 4.5. Hasil hitung Parameter Regresi, Dengan dan Tanpa Penghapusan Kasus, Data Klub Kesehatan

variabel	*Semua data(b)*	*Kasus 23 dihapus(b)*	*Kasus 28 dihapus(b)*	*Kasus 30 dihapus(b)*
konstanta	-3,620	-32,090	-15,140	31,720
X_1	1,268	1,210	0,938	1,384
X_2	-0,525	-0,643	-0,644	-1,316
X_3	-0,505	-0,371	-0,379	-0,642
X_4	3,903	4,194	4,599	4,228
S	28,67	28,78	27,38	26,44
R^2	85,3%	84,2%	86,2%	87,9%
F	36,30	32,03	37,34	43,65
db	(5,25)	(5,24)	(5,24)	(5,24)
$F_{\alpha=0,05}$	2,60	2,62	2,62	2,62

Sumber : Data diolah dengan MINITAB

Dari tabel diatas dihasilkan suatu persamaan taksiran dari model regresi linier dengan dan tanpa penghapusan kasus sebagai berikut :

Untuk *semua data*, maka persamaan taksirannya adalah :

$$\hat{Y} = -3,62 + 1,27X_1 - 0,525X_2 - 0,505X_3 + 3,90X_4$$

Jika *kasus 23 dihapus*, maka persamaan taksirannya adalah :

$$\hat{Y} = -32,1 + 1,21X_1 - 0,643X_2 - 0,371X_3 + 4,19X_4$$

Jika *kasus 28 dihapus*, maka persamaan taksirannya adalah :

$$\hat{Y} = -15,1 + 0,938X_1 - 0,644X_2 - 0,379X_3 + 4,60X_4$$

Jika *kasus 30 dihapus*, maka persamaan taksirannya adalah :

$$\hat{Y} = 31,7 + 1,38X_1 - 1,32X_2 - 0,642X_3 + 4,23X_4$$

Dari Tabel 4.5 dapat dilihat bahwa perubahan nilai b yang disebabkan oleh penghapusan kasus 30 tampak lebih besar daripada perubahan yang disebabkan oleh penghapusan kasus 23 dan kasus 28. Dari Tabel **4.6** tampak bahwa jika kasus 30 dihapus, maka kasus 23 mempunyai statistik-statistik pengaruh yang cukup besar dibanding kasus 28.

Tabel 4.6. Statistik-Statistik Kasus, Data Klub Kesehatan

Statistik	*Semua data*	*Kasus 23 dihapus*	*Kasus 28 dihapus*	*Kasus 30 dihapus*
$h_{23,23}$	0,513	-	0,516	0,515
$h_{28,28}$	0,388	0,391	-	0,388
$h_{30,30}$	0,193	0,196	0,193	-
$D_{23,23}$	0,174	-	0,251	0,275
$D_{28,28}$	0,394	0,428	-	0,443
$D_{30,30}$	0,220	0,236	0,234	-
$VR_{23,23}$	0,173	-	0,157	0,151
$VR_{28,28}$	-0,000	-0,021	-	-0,033
$VR_{30,30}$	-0,165	-0,199	-0,204	-
$AP_{23,23}$	0,164	-	0,168	0,169
$AP_{28,28}$	0,138	0,144	-	0,145
$AP_{30,30}$	0,099	0,108	0,108	-
$\tilde{d}_{23,23}$	0,470	-	0,426	0,408
$\tilde{d}_{28,28}$	-0,257	-0,323	-	-0,375
$\tilde{d}_{30,30}$	-1,135	-1,239	-1,263	-

Sumber : Data diolah dengan MINITAB dan LOTUS 123

BAB V

KESIMPULAN

1. Untuk menentukan kasus-kasus yang berpengaruh, diperlukan pemeriksaan leverage dan pemeriksaan pengaruh selain pemeriksaan sisa.
2. Suatu titik data secara individu mungkin merupakan pencilan, titik-titik dengan leverage tinggi, atau titik yang berpengaruh.

   Kasus yang termasuk dalam salah satu kategori ini harus diperiksa dengan cermat terhadap ketelitian pada pencatatan atau pengetikan, ketelitian pada pengukuran, dan sebagainya. Jika kasus yang berpengaruh berkaitan dengan kesalahan pencatatan atau pengetikan, kesalahan pengukuran, atau kondisi-kondisi percobaan yang tidak memadai, maka kasus tersebut harus dihapus, atau jika mungkin diperbaiki.
3. Suatu kasus pencilan harus diperiksa dengan hati-hati dan teliti. Jika suatu kasus pencilan dihapus, memberikan pengaruh naik pada koefisien

korelasi yang berarti bahwa penghapusan kasus pencilan tersebut menjadikan kecocokan model makin baik.

Sedangkan penghapusan kasus dengan leverage tinggi memberikan pengaruh naik dan turun pada koefisien korelasi. Kasus dengan leverage tinggi yang tidak berpengaruh tidak menimbulkan masalah, tetapi kasus dengan leverage tinggi yang berpengaruh harus diperiksa dengan teliti.

DAFTAR PUSTAKA

Chatterjee, S. dan A. S. Hadi, 1988; "*Sensitivity Analysis in Linear Regression*", John Wiley and Sons, New York; Halaman 64.

Cook, R. D. and S. Weisberg, 1982; "*Residuals and Influence in Regression*", Chapman-Hall, London; Halaman 157-175.

Neter, J., W. Wasserman and M. H. Kutner, 1990; "*Applied Linear Regression Models*", Irwin, Homewood; Halaman 400-409.

Sembiring, R. K., 1977; "*Analisis Regresi*", Jurusan Matematika, ITB Bandung; Halaman 146-168.

Weisberg, S., 1985; "*Applied Regression Analysis*", Jon Wiley and Sons, Inc., New York; Halaman 106-125.