BAB II

REGRESI GANDA

2.1. MODEL REGRESI GANDA

Model regresi linier yang sederhana adalah suatu model dengan satu variabel bebas x, dimana hubungan antara variabel bebas x dengan variabel tak bebasnya, y, berupa suatu garis lurus. Secara matematis dapat ditulis sebagai berikut

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (2.1)$$

$$i = 1,2,\ldots,n$$

dengan asumsi

$$E(\varepsilon_i) = 0$$

$$var(\varepsilon_i) = \sigma^2$$

$$cov(\varepsilon_i, \varepsilon_j) = 0 \quad , \quad i \neq j$$

Suatu model regresi yang melibatkan lebih dari satu variabel bebas disebut model regresi linier ganda. Model ini secara matematis dapat ditulis sebagai berikut

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{2i} + \ldots + \beta_p x_{ip} + \varepsilon_i \qquad (2.2)$$

$$i = 1,2,\ldots,n$$

dengan asumsi sama seperti pada model regresi linier sederhana.

Persamaan-persamaan di atas disebut persamaan-persamaan linier, karena fungsi antara β_0, β_1, β_2,..., β_p yaitu parameter-parameter yang tidak diketahui, adalah berupa fungsi linier.

Parameter β_j, $j = 0,1,2,\ldots,p$ disebut koefisien

regresi, yang menyatakan perubahan yang diharapkan pada variabel takbebas y, jika x_j berubah untuk x_i $(i \neq j)$ tetap. Karena alasan tersebut, maka parameter β_j, $j = 1,2,\ldots,p$ kadang-kadang disebut sebagai koefisien regresi parsial.

Untuk menaksir koefisien regresi pada persamaan (2.2) digunakan metode kwadrat terkecil (*least square*). Misalkan ada n persamaan $(n > k)$ dan y_i adalah respon ke-i, x_{ij} adalah variabel bebas ke-j, maka y_i dapat ditulis sebagai berikut

$$y_i = \beta_o + \sum_{i=1}^{p} \beta_j x_{ij} + \varepsilon_i \qquad (2.3)$$

Fungsi kwadrat kesalahannya adalah

$$S(y,\beta_o,\beta_1,\ldots,\beta_p) = \sum_{i=1}^{n} \varepsilon_i^2$$

$$= \sum_{i=1}^{n} (y_i - \beta_o - \sum_{j=1}^{p} \beta_j x_{ij})^2 \qquad (2.4)$$

Jika fungsi S diminimumkan terhadap $\beta_o,\beta_1,\ldots,\beta_p$ maka akan didapat taksiran kwadrat terkecil dari $\beta_o,\beta_1,\ldots,\beta_p$. Yaitu dengan cara penurunan parsial persamaan (2.4) sebagai berikut

$$\frac{\delta S}{\delta \beta}\Big|_{\hat{\beta}_o\hat{\beta}_1\ldots\hat{\beta}_p} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_o - \sum_{j=1}^{p} \hat{\beta}_j x_{ij}) = 0$$

dan

$$\frac{\delta S}{\delta \beta}\Big|_{\hat{\beta}_o\hat{\beta}_1\ldots\hat{\beta}_p} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_o - \sum_{j=1}^{p} \hat{\beta}_j x_{ij}) x_{ij} = 0$$

(2.5)

Dari persamaan (2.5) didapat persamaan-persamaan normal sebagai berikut :

$$n\hat{\beta}_o + \hat{\beta}_1 \Sigma x_{i1} + \hat{\beta}_2 \Sigma x_{i2} + \ldots + \hat{\beta}_p \Sigma x_{ip} = \Sigma y_i$$

$$\hat{\beta}_o \Sigma x_{i1} + \hat{\beta}_1 \Sigma x_{i1}^2 + \hat{\beta}_2 \Sigma x_{i1}x_{i2} + \ldots + \hat{\beta}_p \Sigma x_{i1}x_{ip} = \Sigma x_{i1}y_i$$

$$\hat{\beta}_0 \Sigma x_{ip} + \hat{\beta}_1 \Sigma x_{ip} x_{i1} + \hat{\beta}_2 \Sigma x_{ip} x_{i2} + \ldots + \hat{\beta}_p \Sigma x_{ip}^2 = \Sigma x_{ip} y_i \qquad (2.5a)$$

Solusi untuk k=p+1 persamaan-persamaan normal di atas adalah merupakan taksiran kwadrat terkecil untuk $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ yaitu $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$.

Regresi ganda akan lebih mudah dimengerti bila dinyatakan dalam bentuk matrik. Hal ini akan memberikan suatu penggambaran yang terpadu dari model, data dan hasil. Secara umum suatu vektor atau matrik dalam regresi dinotasikan dengan X, ε, β dan sebagainya, sedangkan elemen-elemennya dinotasikan dengan x_{ij}, ε_i, β_j, dan sebagainya. Persamaan model regresi (2.3) dapat ditulis dalam notasi matrik sebagai berikut :

$$Y = X\beta + \varepsilon \qquad (2.6)$$

dimana

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Secara umum, Y adalah suatu vektor ukuran (n×1) dari pengamatan, X adalah suatu matrik ukuran (n×k) dari tingkat

variabel regresor, β adalah suatu vektor ukuran (k×1) dari koefisien regresi, dan ε adalah suatu vektor ukuran (n×1) dari kesalahan random.

Akan dicari vektor taksiran kwadrat terkecil β yang meminimumkan fungsi kwadrat jumlah $S(\beta)$, yaitu :

$$S(\beta) = \sum \varepsilon_i^2 = \varepsilon^T \varepsilon$$
$$= (Y - X\beta)^T (Y - X\beta)$$
$$= Y^T Y - \beta^T X^T Y - Y^T X\beta + \beta^T X^T X\beta$$
$$= Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta$$

Karena $\beta^T X^T Y$ adalah suatu matrik (1×1), atau suatu skalar, dan transposnya, $(\beta^T X^T Y)^T = Y^T X\beta$ adalah juga suatu skalar yang sama. Estimator kwadrat terkecil harus memenuhi :

$$\frac{\delta S}{\delta \beta}\Big|_{\hat{\beta}} = -2X^T Y + 2X^T X\hat{\beta} = 0$$

yang mana bisa ditulis menjadi

$$X^T X\hat{\beta} = X^T Y \qquad (2.7)$$

Persamaan (2.7) adalah persamaan normal kwadrat terkecil. Bentuk ini identik dengan persamaan (2.5a). Untuk mencari penyelesaiannya maka kedua bagian dari persamaan (2.7) dikalikan dengan invers dari $X^T X$. Sehingga taksiran kwadrat terkecil untuk β adalah

$$\hat{\beta} = (X^T X)^{-1} X^T Y \qquad (2.8)$$

dan ini menunjukkan bahwa $(X^T X)^{-1}$ harus ada. Matrik $(X^T X)^{-1}$ ini selalu ada jika variabel-variabel bebasnya adalah bebas linier, yaitu jika tidak ada suatu kolom dalam X yang merupakan kombinasi linier dari kolom-kolom yang lain.

Mudah dimengerti bahwa bentuk matrik dari persamaan

normal (2.7) adalah identik dengan bentuk skalar (2.5a). Secara detail, persamaan (2.7) dapat ditulis

$$\begin{bmatrix} n & \Sigma x_{i1} & \Sigma x_{i2} & \cdots & \Sigma x_{ip} \\ \Sigma x_{i1} & \Sigma x_{i1}^2 & \Sigma x_{i1}x_{i2} & \cdots & \Sigma x_{i1}x_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ \Sigma x_{ik} & \Sigma x_{ip}x_{i1} & \Sigma x_{ip}x_{i2} & \cdots & \Sigma x_{ip}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} \Sigma y_i \\ \Sigma x_{i1}y_i \\ \\ \Sigma x_{ip}y_i \end{bmatrix}$$

Jika perkalian matrik di atas dijalankan, maka akan diperoleh bentuk skalar dari persamaan normal (2.5a). Dalam displai ini dapat dilihat bahwa matrik X^TX adalah suatu matrik simetris ukuran (k×k) dan X^TY adalah suatu vektor kolom (k×1). Perlu dicatat struktur khusus dari matrik X^TX. Elemen diagonal dari matrik X^TX adalah jumlah kwadrat dari elemen-elemen dalam kolom dari X, dan elemen-elemen diluar diagonal adalah jumlah cross-product dari kolom-kolom dari X.

Model regresi yang cocok sehubungan dengan variabel bebas $x_i^T = [1\ x_1\ x_2\ \ldots\ x_p\ ]$ adalah

$$\begin{aligned} \hat{y}_i &= x_i^T \hat{\beta} \\ &= \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_j \end{aligned}$$

Vektor dari nilai yang cocok $\hat{y}_i$ sehubungan dengan nilai pengamatan y_i adalah

$$\begin{aligned} \hat{Y} &= X\hat{\beta} \\ &= X(X^TX)^{-1}X^TY \\ &= PY \end{aligned} \qquad (2.8)$$

Matrik $P = X(X^TX)^{-1}X^T$ yang berukuran nxn sering disebut matrik prediksi, karena matrik tersebut memetakan vektor

nilai pengamatan y_i menjadi vektor nilai yang cocok $\hat{y}_i$, yaitu $\hat{Y} = PY$.

Selisih antara nilai pengamatan y_i dan nilai yang cocok $\hat{y}_i$ disebut residual *(sisaan)*, dengan notasi e_i, dimana $e_i = y_i - \hat{y}_i$. Residual-residual (sebanyak n) dapat ditulis secara lebih cocok dalam bentuk matrik sebagai :

$$\begin{aligned} e &= Y - \hat{Y} \\ &= Y - X\hat{\beta} \\ &= Y - X(X^TX)^{-1}X^TY \\ &= Y - PY \\ &= (I - P)Y \end{aligned} \qquad (2.9)$$

2.2. ASUMSI DASAR

Selain asumsi-asumsi yang telah disebutkan di muka, hasil dari metode kwadrat terkecil dan analisis statistik harus mengikuti asumsi-asumsi sebagai berikut :

1. Asumsi kelinieran

   Asumsi ini dinyatakan secara implisit pada definisi model (2.1), yaitu setiap nilai y_i dapat ditulis sebagai fungsi linier dari baris ke-i pada matrik X, ditulis x_i^T.

   $$y_i = x_i^T\beta + \varepsilon_i \qquad , i = 1,2,\ldots,n.$$

2. Asumsi perhitungan

   Untuk mendapatkan taksiran β yang tunggal, $(X^TX)^{-1}$ harus ada, atau secara ekwivalen rank(X) = k.

3. Asumsi distribusi

Analisis statistik yang berdasarkan pada metode kwadrat terkecil (misal : uji-t, uji-F dan sebagainya) mengasumsikan bahwa :

a. X diukur tanpa kesalahan.

b. ε_i tidak bergantung pada x_i^T.

c. $\varepsilon \sim N_n(0,\sigma^2 I)$.

4. Asumsi yang tertera secara implisit

Seluruh pengamatan dapat dipercaya dan mempunyai peranan yang sama dalam menentukan hasil metode kwadrat terkecil dan pengaruhnya.

Jika asumsi-asumsi di atas dipenuhi, maka teori kwadrat terkecil memberikan hasil-hasil yang diketahui sebagai berikut :

1. Vektor $\hat{\beta}$ ukuran (k×1) mempunyai sifat-sifat sebagai berikut :

a. $E(\hat{\beta}) = \beta$ (2.10a)

yaitu $\hat{\beta}$ adalah taksiran tak bias untuk β.

b. $\hat{\beta}$ adalah taksiran tak bias linier terbaik (the best linear unbiased estimator/BLUE) untuk β, yaitu diantara kelas-kelas dari taksiran tak bias linier untuk β, $\hat{\beta}$ mempunyai variansi terkecil. Variansi untuk $\hat{\beta}$ adalah

$$var(\hat{\beta}) = \sigma^2(X^TX)^{-1} \quad (2.10b)$$

c. $\hat{\beta} \sim N_k(\beta,\sigma^2(X^TX)^{-1})$,

dimana $N_k(\mu,\Sigma)$ menggambarkan suatu distribusi normal multivariat berdimensi k dengan mean μ

(suatu vektor ukuran (k×1)) dan variansi Σ (suatu matrik (k×k)).

2. Vektor (n×1) dari nilai prediksi

$$\hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY = PY \qquad (2.11)$$

mempunyai sifat-sifat sebagai berikut

a. $E(\hat{Y}) = X\beta$ (2.11a)

b. $var(\hat{Y}) = \sigma^2 P$ (2.11b)

c. $\hat{Y} \sim N_n(X\beta, \sigma^2 P)$ (2.11c)

3. Vektor (n×1) dari residual

$$e = Y - \hat{Y} = Y - PY$$
$$= (I - P)Y \qquad (2.12)$$

mempunyai sifat-sifat sebagai berikut :

a. $E(e) = 0$ (2.12a)

b. $var(e) = \sigma^2(I-P)$ (2.12b)

c. $e \sim N_n(0, \sigma^2(I-P))$ (2.12c)

4. Suatu taksiran tak bias dari σ^2 diberikan oleh

$$\hat{\sigma}^2 = \frac{e^Te}{n-k}$$
$$= \frac{Y^T(I-P)Y}{n-k} \qquad (2.13)$$

dimana e^Te adalah jumlah kwadrat sisaan.

2.3. MATRIK PREDIKSI

Pada bagian ini akan dibahas berbagai hal yang berhubungan dengan matrik prediksi P. Pada (2.8) telah diperoleh bahwa :

$$\hat{Y} = PY$$

dimana $P = X(X^TX)^{-1}X^T$ (2.14)

Sehingga P disebut matrik prediksi, karena matrik tersebut jika dikalikan dengan Y akan menghasilkan nilai prediksi untuk Y, yaitu $\hat{Y}$. Matrik prediksi ini mempunyai peranan yang penting dalam analisis regresi dan teknik analisis multivariat yang lainnya.

Elemen-elemen dari matrik P adalah :

$$p_{ij} = x_i^T(X^TX)^{-1}x_j \quad (2.15.a)$$

$$p_{ii} = x_i^T(X^TX)^{-1}x_i \quad (2.15.a)$$

dimana

p_{ii} adalah elemen diagonal ke-i dari matrik P

x_i^T adalah baris ke-i dari matrik data X

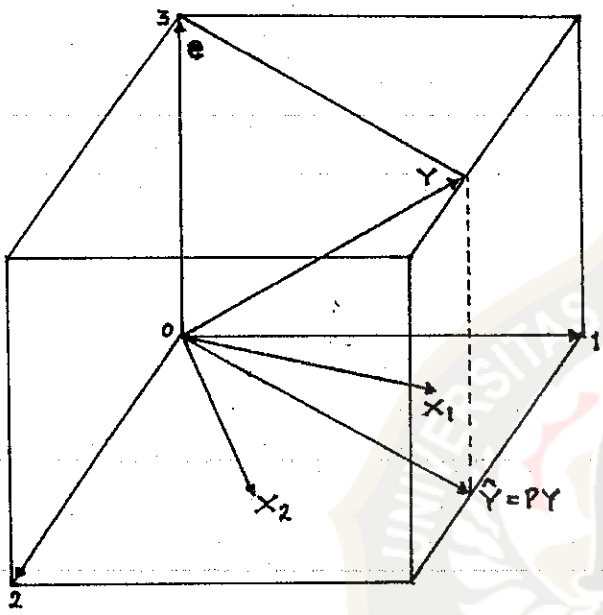
x_j adalah kolom ke-j dari matrik data X

$i = 1,2,\ldots,n \qquad j = 1,2,\ldots,k$

Secara geometris, bila vektor y dan kolom-kolom dalam X digambarkan sebagai titik-titik dalam ruang euclidian berdimensi n, maka titik $X\beta$ (yang diperoleh sebagai kombinasi linier dari vektor-vektor kolom) akan membentuk suatu subruang berdimensi k. Vektor nilai yang cocok $\hat{y}$ adalah suatu titik pada ruang tersebut yang paling dekat dengan y dan juga merupakan proyeksi orthogonal y pada subruang tersebut, sehingga P adalah matrik proyeksi. Untuk lebih jelasnya dapat dilihat pada contoh berikut ini :

Contoh 2.5.1. Suatu kasus dimana Y diregresikan melalui titik asal pada dua prediktor X_1 dan X_2, dengan n=3. Dalam

gambar 2.1 dapat dilihat bahwa vektor dari nilai prediksi $\hat{Y} = PY$ adalah proyeksi orthogonal dari Y pada subruang berdimensi 2 yang dibentuk oleh X_1 dan X_2. Demikian juga vektor sisaan $e = (I-P)Y$ adalah proyeksi orthogonal dari y pada komplemen subruang dimensi satu yang tegak lurus



$$X = (X_1 : X_2) = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ o & o \end{bmatrix}$$

$$\text{rank}(X) = 2, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

$$\hat{Y} = \begin{bmatrix} y_1 \\ y_2 \\ 0 \end{bmatrix}, \quad e = \begin{bmatrix} 0 \\ 0 \\ y_3 \end{bmatrix}$$

Gambar 2.1. Suatu contoh yang menggambarkan bahwa $\hat{Y}$ adalah proyeksi orthogonal dari Y pada ruang yang dibentuk ole X1 dan X2.

(orthogonal) pada X_1 dan X_2. Di sini juga nampak bahwa e tegak lurus pada $\hat{Y}$ ($\hat{Y}$ terletak pada subruang yang dibentuk oleh $(X_1 : X_2)$ dan e terletak pada komplemen orthogonalnya).

Matrik prediksi P mempunyai sifat-sifat khusus yang berhubungan dengan hukum-hukum dalam aljabar matrik, yaitu :

Sifat 2.5.1. Matrik P dan (I-P) adalah matrik yang simetris

dan idempoten. (matrik P dikatakan simetris bila $P^T = P$ dan matrik P dikatakan idempoten bila P.P = P)

Bukti :

$$P^T = [X(X^TX)^{-1}X^T]^T$$
$$= X[(X^TX)^{-1}]^TX^T$$
$$= X(X^TX)^{-1}X^T$$
$$= P.$$

karena P simetris dan I juga simetris maka jelas (I-P) juga simetris.

$$PP = \{X(X^TX)^{-1}X^T\}\{X(X^TX)^{-1}X^T\}$$
$$= X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T$$
$$= X(X^TX)^{-1}X^T$$
$$= P.$$

$$(I-P)(I-P) = I - 2P + PP$$
$$= I - 2P + P$$
$$= (I-P)$$

Sifat 2.5.2. Ambil X matrik ukuran (n×k), maka

$$trace[P] = rank(P) = k.$$

Bukti :

Berdasarkan asumsi ke-2 maka rank(X) = k. Maka X^TX adalah matrik definit positip berukuran (k×k), sehingga X^TX adalah nonsingular, dan $rank[X^TX] = k$, demikian juga $rank[X^TX]^{-1} = k$. Jadi $rank[P] = rank[X(X^TX)^{-1}X^T] = rank[X^TX]^{-1} = k$.

Matrik P adalah simetris dan idempoten. Diberikan PP = P maka $Px = \lambda x$ $(x \neq 0)$ mengimplikasikan bahwa $\lambda x^Tx = x^TPx = x^TP^2x =$

$(Px)^T(Px) = \lambda^2 x^T x$, dan $\lambda(\lambda-1) = 0$. Oleh karena itu eigenvaluenya adalah 0 dan 1. Maka jumlah eigenvalue matrik P sama dengan rank[P] = k, dan trace[P] = $\sum_{i=1}^{n} \lambda_i$ = k = rank[P].

Sifat 2.5.3. Ambil $X = (X_1 : X_2)$ dimana X_1 adalah suatu matrik berukuran (n×r) dengan rank = r dan X_2 adalah suatu matrik berukuran (n×(k-r)) dengan rank = k-r. Andaikan bahwa $P_1 = X_1(X_1^T X_1)^{-1} X_1^T$ adalah matrik prediksi untuk X_1 dan $W = (I-P_1)X_2$. adalah proyeksi dari X_2 pada komplemen orthogonal dari X_1. Akhirnya diandaikan bahwa $P_2 = W(W^T W)^{-1} W^T$ adalah matrik prediksi untuk W. Maka matrik prediksi P dapat diekspresikan sebagai

$$X(X^T X)^{-1} X^T = X_1(X_1^T X_1)^{-1} X_1^T + (I-P_1)X_2(X_2^T(I-P_1)X_2)^{-1} X_2^T(I-P_1) \quad (2.16)$$

atau

$$P = P_1 + P_2 \quad (2.16a)$$

Bukti :

Diketahui $X = (X_1 : X_2)$ maka $P = X(X^T X)^{-1} X^T$ dapat ditulis dalam bentuk

$$P = [X_1 \; X_2] \begin{bmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{bmatrix} \begin{bmatrix} X_1^T \\ X_2^T \end{bmatrix} \quad (2.17)$$

Dengan menghitung invers dari $(X^T X)$dalam bentuk dipartisikan, didapat :

$$(X^T X)^{-1} = \begin{bmatrix} (X_1^T X_1)^{-1} + (X_1^T X_1)^{-1} X_1^T X_2 M X_2^T X_1 (X_1^T X_1)^{-1} & X_1^T X_1)^{-1} X_1^T X_2 M \\ -M X_2^T X_1 (X_1^T X_1)^{-1} & M \end{bmatrix} \quad (2.17a)$$

dimana $M = [X_2^T X_2 - X_2^T X_1 (X_1^T X_1)^{-1} X_1^T X_2]^{-1}$

$$= [X_2^T(I-X_1(X_1^TX_1)^{-1}X_1^T)X_2]^{-1}$$

$$= [X_2^T(I-P_1)X_2]^{-1}$$

Dengan mensubstitusikan (2.17a) ke (2.17), didapatkan

$$P = P_1+P_1X_2MX_2^TP_1-P_1X_2MX_2^T-X_2MX_2^TP_1+X_2MX_2^T$$

$$= P_1+(I-P_1)X_2MX_2^T(I-P_1)$$

$$= P_1+(I-P_1)X_2[X_2^T(I-P_1)X_2]^{-1}X_2^T(I-P_1)$$

$$= P_1 + P_2$$

Sifat 2.5.3. di atas menunjukkan bahwa matrik prediksi P dapat didekomposisi menjadi jumlah dua (atau lebih) matrik prediksi.

sifat 2.5.4. Untuk i= 1,2,...,n dan j= 1,2,...,n, maka

(a) $0 \leq p_{ii} \leq 1$ untuk semua i

(b) $-0,5 \leq p_{ij} \leq 0.5$ untuk setisp i≠j

(c) jika $p_{ii} = 0$ atau 1 maka $p_{ij} = 0$.

(d) $p_{ii} + \dfrac{e_i^2}{e^Te} \leq 1$

Bukti :

Dengan mengingat sifat 2.4.1. maka elemen diagonal ke-i dari matrik P dapat ditulis sebagai

$$p_{ii} = \sum_{j=1}^{n} p_{ij}^2 = p_{ii}^2 + \sum_{j\neq i} p_{ij}^2. \qquad (2.18)$$

dari persamaan di atas didapat bahwa $0 \leq p_{ii} \leq 1$, untuk semua i. Jadi (a) terbukti.

Persamaan (2.18) juga bisa ditulis dalam bentuk

$$p_{ii} = p_{ii}^2 + p_{ij}^2 + \sum_{r\neq i,j} p_{ir}^2. \qquad (2.19)$$

dari persamaan ini didapat bahwa

$$p_{ij}^2 \leq p_{ii}(1 - p_{ii})$$

atau $$-\sqrt{p_{ii}(1-p_{ii})} \leq p_{ij} \leq \sqrt{p_{ii}(1-p_{ii})}$$

dan karena $0 \leq p_{ii} \leq 1$ maka didapat

$$-0,5 \leq p_{ij} \leq 0,5$$

Jadi (b) terbukti.

Dari persamaan (2.18) jelas bahwa jika $p_{ii} = 0$ atau 1 maka $p_{ij} = 0$, untuk semua $i \neq j$.

Jadi (c) terbukti.

Didefinisikan $Z = (X:Y)$, $P_x = X(X^TX)^{-1}X^T$, dan $P_z = Z(Z^TZ)^{-1}Z^T$.

Berdasarkan (2.16) maka diperoleh

$$P_z = P_x + \frac{(I-P_x)YY^T(I-P_x)}{Y^T(I-P_x)Y}$$

$$= P_x + \frac{ee^T}{e^Te}$$

dan karena elemen diagonal dari matrik P_z adalah kurang atau sama dengan satu, maka (d) terbukti.

Contoh 2.5.2. Misalkan pencocokan garis lurus pada suatu himpunan data yang terdiri dari lima titik; empat titik terletak pada x=1 dan satu titik terletak pada x=4. Dalam kasus ini maka diperoleh

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 4 \end{bmatrix},$$

$$X^TX = \begin{bmatrix} 5 & 8 \\ 8 & 20 \end{bmatrix},$$

$$(X^TX)^{-1} = \frac{1}{36}\begin{bmatrix} 20 & -8 \\ -8 & 5 \end{bmatrix},$$

$$P = X(X^TX)^{-1}X^T = \begin{bmatrix} 0,25 & 0,25 & 0,25 & 0,25 & 0 \\ 0,25 & 0,25 & 0,25 & 0,25 & 0 \\ 0,25 & 0,25 & 0,25 & 0,25 & 0 \\ 0,25 & 0,25 & 0,25 & 0,25 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

dan

$$(I-P) = \begin{bmatrix} 0,75 & -0,25 & -0,25 & -0,25 & 0 \\ -0,25 & 0,75 & -0,25 & -0,25 & 0 \\ -0,25 & -0,25 & 0,75 & -0,25 & 0 \\ -0,25 & -0,25 & -0,25 & 0,75 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$\text{trace}[P] = \sum p_{ii} = 2 = \text{rank}[P]$.

$\text{trace}[I-P] = n-k = 5-2 = 3$.

tampak bahwa $p_{5.5}=1$.

dan

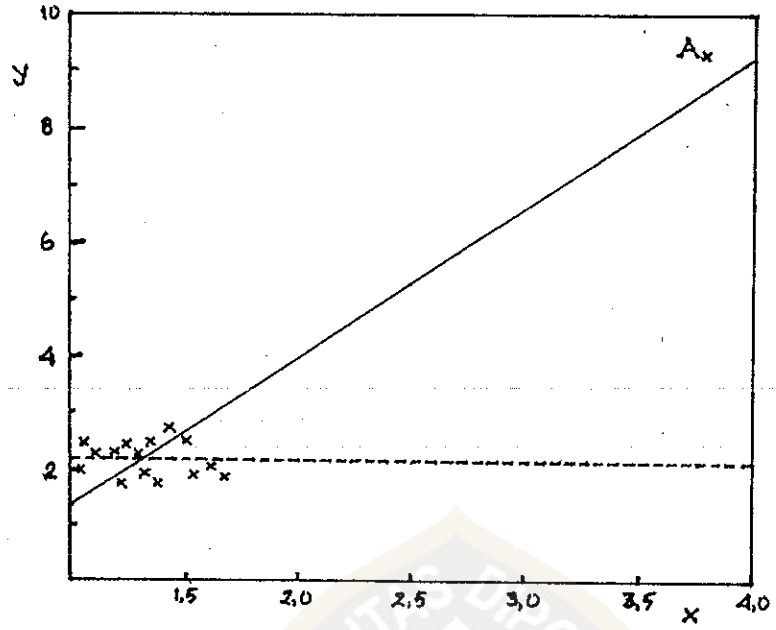
$p_{5.j} = 0$, untuk $j= 1,2,3,4$. (sifat 2.4.3c).

2.4. PENGAMATAN BERPENGARUH

Dalam menganalisa data, teknik-teknik analisis yang biasa digunakan adalah metode-metode statistik untuk menganalisa data yang mempengaruhi model regresi secara seimbang. Padahal mungkin terdapat data atau pengamatan yang mempengaruhi model secara tak seimbang, sehingga metode-metode penaksiran parameter, uji hipotesa dan lain sebagainya tidak memberikan hasil yang baik serta kecukupan model masih kurang.

Data haruslah dianalisa secara seksama dengan memberikan perhatian khusus pada data atau pengamatan yang

diperkirakan sebagai yang berpengaruh pada model regresi.



Gambar 2.6.1 Suatu contoh titik pengamatan berpengaruh

———— : regresi y pada x, seluruh data, $R^2 = 0.90$

--------- : regresi dengan pengamatan A dihapus, $R^2 < 0.01$

Mungkin saja karakter dari regresi dapat ditentukan hanya berdasarkan beberapa pengamatan berpengaruh ini, dan yang lain diabaikan.

Berikut ini diberikan contoh yang cukup ekstrim untuk menggambarkan pengamatan yang berpengaruh.

Contoh 2.6.1. Dari gambar 2.6.1 nampak bahwa jika titik A dipindahkan atau dihapus dari data, hasil analisa dapat berubah banyak sekali, sebagaimana yang digambarkan oleh kedua garis regresi yang dihitung dengan dan tanpa titik A. Tampak bahwa R^2 mengalami perubahan yang cukup mencolok.

Titik outlier, titik high-leverage dan titik yang

berpengaruh merupakan tiga konsep yang saling berhubungan. Berikut ini akan dibahas bagaimana interaksi diantara ketiga konsep tersebut.

Outlier. Dalam kerangka regresi linier, outlier didefinisikan sebagai pengamatan yang mempunyai nilai residual absolut yang besar dibandingkan dengan pengamatan-pengamatan yang lain dalam himpunan data.

High-Leverage. Titik high-leverage adalah suatu pengamatan yang mempunyai harga p_{ii} yang besar dibandingkan dengan pengamatan-pengamatan yang lain dalam himpunan data. Suatu pengamatan yang terisolasi dalam ruang-X akan merupakan titik high-leverage. Suatu titik high-leverage dapat dipandang sebagai outlier dalam ruang-X. Konsep dari high-leverage sepenuhnya berhubungan dengan variabel prediktor dan tidak dengan variabel respon.

Pengamatan Berpengaruh. Pengamatan berpengaruh adalah pengamatan-pengamatan yang secara individu maupun berkelompok terlalu mempengaruhi pencocokan model regresi dibandingkan dengan pengamatan-pengamatan yang lain dalam himpunan data. Definisi ini nampak subyektif, tapi memberikan implikasi bahwa pengamatan-pengamatan dapat diurutkan dengan cara yang sesuai menurut beberapa ukuran dari pengaruhnya.

Dalam hubungan ketiga konsep tersebut ada 4 hal yang perlu dicatat, yaitu :

1. Titik outlier belum tentu merupakan pengamatan berpengaruh.

2. Pengamatan berpengaruh belum tentu suatu outlier.
3. Jika tidak terdapat residual yang besar pada hasil analisis regresi, bukan berarti model regresi yang ada telah cocok, tapi mungkin terdapat pengamatan dengan residual yang besar tertutup oleh pengamatan yang lain. Pada kenyataannya terdapat kecenderungan, titik dengan high-leverage mempunyai sisaan yang kecil dan mempengaruhi pencocokan model secara tak seimbang.
4. Seperti halnya pada outlier, titik high-leverage belum tentu berpengaruh dan pengamatan berpengaruh tidak harus titik high-leverage. Tapi, bagaimanapun titik high-leverage kemungkinan besar adalah berpengaruh.

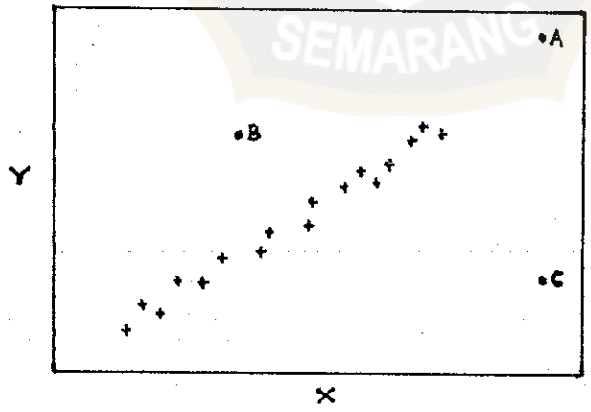
Untuk lebih jelasnya diberikan contoh-contoh berikut sebagai ilustrasi dari pernyataan-pernyataan tersebut.

Contoh 2.6.2. Misalkan ada titik-titik data yang dibentuk dengan tanda "+" dan akan ditambahkan 3 data yang ditandai dengan A, B dan C, seperti yang terlihat pada gambar 2.6.2. Jika hanya titik A yang dimasukkan ke dalam data, maka A mempunyai harga sisaan yang kecil karena Y-nya terletak di dekat garis regresi. A merupakan titik high-leverage karena merupakan outlier dalam ruang-X, tapi tidak berpengaruh besar pada pencocokan persamaan regresi. Jelaslah bahwa A merupakan contoh titik high-leverage yang bukan outlier maupun titik

berpengaruh. A tidak berpengaruh pada penaksiran koefisien regresi, tapi ,karena A suatu titik ekstrim pada ruang-X, mungkin berpengaruh pada standard error dari koefisien regresi.

Bila hanya titik B yang dimasukkan ke dalam data, maka B tidak akan menjadi titik high-leverage karena letaknya dekat dengan pusat X, tapi jelas akan merupakan outlier dan titik berpengaruh. B mempunyai sisaan yang besar, dan pemasukannya tidak akan merubah slope tapi intercept dari garis regresi. Pemasukannya juga akan merubah penaksiran variansi error, dan karena itu juga mempengaruhi variansi taksiran koefisian.

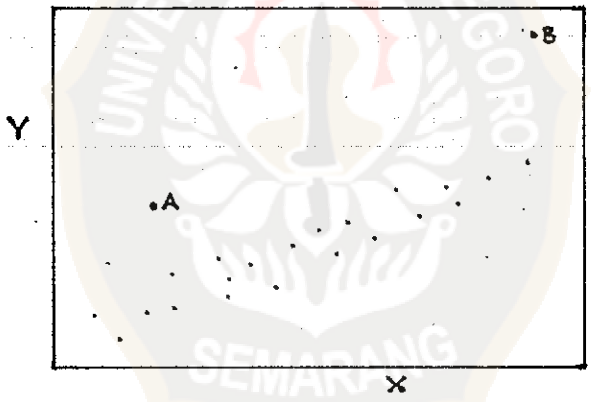
Bila hanya titik C yang dimasukkan ke dalam data, maka C akan merupakan outlier, titik high-leverage sekaligus titik berpengaruh. Sebagai outlier karena C akan memiliki harga sisaan yang besar. Merupakan titik high-leverage karena C adalah titik ekstrim dalam ruang-X. Sebagai titik berpengaruh



Gambar 2.6.2 Suatu contoh yang menggambarkan perbedaan antara outlier, titik high-leverage dan pengamatan berpengaruh

karena pemasukannya secara substansial akan merubah karakteristik dari persamaan regresi yang dicocokkan.

Contoh 2.6.3. Misalkan ada data yang diplotkan seperti pada gambar 2.6.3. Jika suatu garis lurus model regresi dicocokkan pada data tersebut, maka terlihat bahwa A adalah suatu outlier. Garis yang dicocokkan akan berubah jika pengamatan ini dihapus, tapi hanya berpengaruh sedikit pada pada $\hat{\beta}$. Titik B memiliki memiliki harga sisaan yang kecil, tapi jika dihapus maka taksiran koefisien regresi akan berubah. Jadi titik B adalah contoh pengamatan berpengaruh yang bukan outlier.



gambar 2.6.3. Tentang outlier yang bukan pengamatan berpengaruh dan pengamatan berpengaruh yang bukan outlier.