II. TINJAUAN PUSTAKA

2.1 Konsep Dasar Matriks

2.1.1 Matriks

Matriks adalah suatu susunan bilangan berbentuk segi empat. Bilangan-bilangan dalam susunan itu disebut anggota dalam matriks tersebut. Suatu matriks A mempunyai unsur yang dilambangkan dengan a_{ij} dengan *i* menyatakan banyaknya baris dan *j* menyatakan banyak kolom. Suatu matriks A dapat juga dilambangkan dengan **A**=‖ a_{ij} ‖ (Anton, 1987).

2.1.2 Transpose Matriks

Jika A adalah sebarang matriks m x n, maka transpose **A** dinyatakan dengan $\mathbf{A}^{\mathbf{T}}$, didefinisikan dengan matriks n x m yang didapatkan dengan mempertukarkan baris dan kolom dari **A**, yaitu kolom pertama dari $\mathbf{A}^{\mathbf{T}}$ adalah baris pertama dari **A**, kolom kedua dari $\mathbf{A}^{\mathbf{T}}$ adalah baris kedua dari **A**, dan seterusnya (Anton, 1987).

2.1.3 Matriks Simetris

Suatu matriks bujur sangkar **A** disebut simetris jika $\mathbf{A}=\mathbf{A}^{\mathbf{T}}$ (Anton, 1987).

2.1.4 Invers Matriks

Jika **A** adalah matriks bujur sangkar, dan jika sebuah matriks **B** yang berukuran sama bisa didapat sedemikian sehingga **AB=BA=I**, maka A disebut bisa dibalik dan **B** disebut invers dari **A** (Anton, 1987).

2.1.5 Matriks Diagonal

Jika a_{ii} adalah elemen pada diagonal ke-*i* dari matriks **A** berukuran n x n, dan misalkan a_{ij} adalah unsur-unsur diluar diagonal, jika $a_{ij} = 0$ untuk semua $i \neq j$, maka **A** dinamakan matriks diagonal. Biasanya matriks diagonal dilambangkan dengan **D** (Anton, 1987).

2.2 Analisis Regresi Linier

Analisis regresi linier adalah salah satu analisis statistika yang dapat digunakan untuk menyelidiki atau membangun model hubungan linier antara beberapa variabel. Analisis regresi yang mempelajari pola hubungan antara satu variabel tak bebas dan satu variabel bebas disebut analisis regresi linier sederhana (*simple linear regression*).

Model regresi linier sederhana biasa ditulis sebagai berikut :

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Dimana : β_0 adalah intersep atau perpotongan dengan sumbu tegak.

β_1 adalah kemiringan atau gradiennya.

Y adalah variabel tak bebas.

X adalah variabel bebas.

ε adalah galat (*error term*) .

Regresi linier berganda (*multiple linear regression*) merupakan suatu model regresi yang melibatkan satu variabel tak bebas dan lebih dari satu variabel bebas. Model regresi linier berganda dalam bentuk umum yaitu :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i \,, \qquad \text{i=1,2,..,n}$$

Bila dirinci untuk setiap pengamatan :

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \ldots + \beta_p X_{1p} + \varepsilon_1$$

$$Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \ldots + \beta_p X_{2p} + \varepsilon_2$$

$$\vdots \quad \vdots \quad \vdots \quad \vdots \quad \ldots \quad \vdots \quad \vdots$$

$$Y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \ldots + \beta_p X_{np} + \varepsilon_n$$

Dengan cara matriks dapat ditulis sebagai berikut :

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1p} \\ 1 & X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & X_{n1} & X_{n2} & \ldots & X_{np} \end{bmatrix} + \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Dalam notasi matriks ditulis sebagai berikut:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

Dengan: $\boldsymbol{Y}$ adalah vektor *n* x 1 variabel tak bebas.

$\boldsymbol{X}$ adalah matriks *n* x (*p*+1) variabel bebas.

$\boldsymbol{\beta}$ adalah vektor (*p*+1) x 1 parameter yang diduga.

$\boldsymbol{\varepsilon}$ adalah *n* x 1 vektor galat atau *error term*

(Myers, 1990).

2.3 Matriks HAT

Alat pendiagnosa yang memberikan informasi titik data yang mengandung *leverage* tinggi adalah matriks HAT. Matriks HAT didefinisikan berikut :

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

Matriks HAT memainkan peranan penting dalam mengidentifikasi pengamatan berpengaruh. **H** menentukan varian dan kovarian dari $\hat{y}$ dan e, dimana Var ($\hat{y}$) = $\sigma^2\mathbf{H}$ dan var (e) = $\sigma^2(\boldsymbol{I} - \mathbf{H})$ (Montgomery, Peck & Vinning, 2006).

Elemen diagonal h_{ii} dari matriks $\boldsymbol{H}$ didefinisikan sebagai

$$h_{ii} = \boldsymbol{x_i^T}\,(\boldsymbol{X^T X})^{-1}\boldsymbol{x_i}, \qquad i=1,2,\ldots,n$$

Diagonal HAT memberikan ukuran jarak yang terbakukan dari titik x_i ke pusat data dari *x* yaitu $\bar{x}$. Nilai diagonal HAT yang tinggi menunjukkan pengamatan yang ekstrim pada *x* (Myers, 1990).

Nilai diagonal HAT berada antara 0 dan 1, $0 < h_{ii} < 1$. Jika *X* memiliki rank penuh, maka $\sum_{i=1}^{n} h_{ii} = p$. Sehingga rata-rata dari elemen diagonal h_{ii} adalah *p/n*. Disarankan menggunakan 2*p/n* sebagai titik kritis untuk h_{ii}. $h_{ii} > 2p/n$ memiliki potensi untuk berpengaruh kuat pada hasil regresi. Jika pengamatan ke-*i* mempunyai nilai h_{ii} yang melebihi 2*p/n*, maka pengamatan tersebut dikatakan titik *leverage* yang tinggi (Belsley, Kuh, & Welsch, 1980).

2.4 Analisis Residual

2.4.1 Residual

Salah satu dari metode pendiagnosa gangguan pada model (pencilan) adalah dengan kuadrat terkecil residual $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$. Dugaan parameter regresi dengan metode kuadrat terkecil dari $\boldsymbol{\beta}$ adalah $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y}$.

Vektor residual adalah $\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$

$$= (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}})\boldsymbol{Y}$$

$$= (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$$

Matriks varian-kovarian dari residual adalah

$$Var\,(\boldsymbol{e}) = Var\,(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}$$

$$= (\boldsymbol{I}-\boldsymbol{H})\,Var\,\boldsymbol{Y}(\boldsymbol{I}-\boldsymbol{H})^{\boldsymbol{T}}$$

$$= (\boldsymbol{I}-\boldsymbol{H})(\sigma^2\boldsymbol{I_n})(\boldsymbol{I}-\boldsymbol{H})^{\boldsymbol{T}}$$

$$= \sigma^2(\boldsymbol{I}-\boldsymbol{H})^{\boldsymbol{2}}$$

Karena bersifat idempoten $(\boldsymbol{I}-\boldsymbol{H})^{\boldsymbol{2}} = (\boldsymbol{I}-\boldsymbol{H})$, maka $Var\,\boldsymbol{e} = \sigma^2(\boldsymbol{I}-\boldsymbol{H})$ (Myers,1990).

2.4.2 *R-Student*

Diberikan penduga alternatif yaitu akar nilai tengah kuadrat galat yang dihitung dengan menghilangkan pengamatan ke-i . Ini dinotasikan dengan S_{-i} , yaitu

$$S_{-i} = \sqrt{\frac{(n-p)s^2 - e_i^2/(1-h_{ii})}{n-p-1}}$$

Jumlah kuadrat galat tanpa menggunakan pengamatan ke-*i* berbeda dari jumlah kuadrat galat menggunakan semua data dengan kuantitas $\frac{e_i^2}{1-h_{ii}}$. Penduga S_{-i} digunakan menggantikan σ menghasilkan eksternal residual *student* yang sering disebut *R-student*, dengan rumus

$$t_{\boldsymbol{i}} = \frac{e_i}{s_{-i}\sqrt{1-h_{ii}}}$$

Daerah kritis untuk *R- student* yaitu membandingkannya dengan distribusi-*t* berderajat bebas *n-p*-1 yang dapat dilihat pada tabel distribusi-*t*. Nilai *R-student* lebih besar dari nilai *t*-tabel menunjukkan pengamatan merupakan suatu pencilan (Myers, 1990).

2.5 Pengamatan Berpengaruh

Menurut Belsley, Kuh, & Welsch (1980), suatu pengamatan berpengaruh adalah sesuatu yang secara individu atau bersama-sama dengan beberapa pengamatan lain, mempengaruhi nilai terhitung dari berbagai pendugaan (koefisien regresi, standar galat, nilai-*t* dan lain-lain) dibandingkan pada pengamatan yang lain.

Untuk menguji pengaruhnya satu demi satu pengamatan berpengaruh tersebut dihilangkan. Baris-baris pengamatan yang dihilangkan relatif menghasilkan perubahan besar pada nilai terhitung dan dianggap berpengaruh. Dengan pengujian dari prosedur ini, dapat dilihat dampak masing-masing baris pengamatan pada koefisien dugaan dan nilai prediksi ($\hat{y}$) , residual dan dugaan parameter varian-kovarian matriks.

Suatu pengamatan tidak mempunyai dampak yang sama pada semua hasil regresi. Suatu pengamatan mungkin mempunyai pengaruh pada $\hat{\beta}$, pengaruh pada penduga ragam dari $\hat{\beta}$, kecocokan nilai (*fitted value*), atau *goodness-of-fit* statistik (Chatterjee & Hadi, 1986).

2.6 DFBETAS

Diberikan matriks $(\boldsymbol{X^T X})$ berukuran *p* x *p* dan jika $\boldsymbol{x_i^T}$ baris ke-*i* pada $\boldsymbol{X}$.

$\boldsymbol{X^T X - x_i x_i^T}$ adalah matriks $\boldsymbol{X^T X}$ dengan baris ke-*i* dihilangkan.

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

Atau dapat ditulis sebagai

$$(\boldsymbol{X_{-i}^T X_{-i}})^{-1} = (\boldsymbol{X^T X})^{-1} + \frac{(\boldsymbol{X^T X})^{-1} \boldsymbol{x_i x_i^T} (\boldsymbol{X^T X})^{-1}}{1 - \boldsymbol{h_{ii}}}$$

Dimana $\boldsymbol{X_{-i}}$ diperoleh dengan menghapus baris ke-*i* dari $\boldsymbol{X}$. Juga diberikan $h_{ii} = \boldsymbol{x_i (X^T X)^{-1} x_i^T}$ dan diasumsikan $h_{ii} < 1$.

Dari formula di atas dihasilkan berikut

$$\boldsymbol{(X_{-i}^T X_{-i})^{-1} x_i^T} = \frac{\boldsymbol{(X^T X)^{-1} x_i^T}}{\boldsymbol{1 - h_{ii}}} \qquad \ldots (*)$$

Diketahui bahwa

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

$$\boldsymbol{\hat{\beta}_{-i} = (X_{-i}^T X_{-i})^{-1} X_{-i}^T Y_{-i}}$$

Sehingga

$$\boldsymbol{\hat{\beta} - \hat{\beta}_{-i} = (X_{-i}^T X_{-i})^{-1} x_i^T [Y_i - X\hat{\beta}]} \qquad \ldots. (**)$$

Substitusi (**) ke dalam (*), sehingga dihasilkan

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{-\boldsymbol{i}} = \frac{(\boldsymbol{X^T X})^{-1}\boldsymbol{x_i^T}}{\mathbf{1} - \boldsymbol{h_{ii}}}\left[\boldsymbol{Y_i} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right]$$

Sehingga ukuran jarak antara b dan b_{-i} sebagai berikut:

$$DFBETA_i = b - b_{-i} = \frac{(\boldsymbol{X^T X})^{-1}\boldsymbol{x_i^T e_i}}{\mathbf{1} - \boldsymbol{h_{ii}}}$$

Jika $R = (\boldsymbol{X^T X})^{-1}\boldsymbol{x_i}^{\boldsymbol{T}}$, maka

$$b_j - b_{j.-i} = \frac{r_{j.i}\, e_i}{1-h_{ii}}$$

Karena

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{X^T X})^{-1}\boldsymbol{x^T x}(\boldsymbol{X^T X})^{-1} = (\boldsymbol{X^T X})^{-1}$$

Maka diperoleh

$$var(b_j) = \sigma^2 \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{r_{jj}^2}$$

(Belsley, Kuh & Welsch, 1980).

Untuk setiap koefisien regresi, pendiagnosa pengaruh menyediakan satu statistik, yang memberikan nilai standar galat perubahan koefisien jika pengamatan ke-*i* dihilangkan. Rumusnya

$$DFBETAS_{j.i} = \frac{b_j - b_{j.-i}}{s_{-i}\sqrt{C_{jj}}}$$

Dimana C_{jj} adalah elemen diagonal ke-*j* dari $(\boldsymbol{X^T X})^{-1}$.

b_j adalah koefisien regresi ke-*j*.

$b_{j.-i}$ adalah koefisien regresi ke-*j* yang dihitung tanpa pengamatan ke-*i*.

Besarnya nilai $DFBETAS_{j.i}$ mengindikasikan bahwa pengamatan ke-*i* mempunyai pengaruh pada koefisien regresi ke-*j*. Untuk menghitung nilai $DFBETAS_{j.i}$ dibutuhkan suatu matriks *p* x *n*, matriks $R = (\boldsymbol{X^T X})^{-1}\boldsymbol{x_i^T}$. Dari konversi formula diatas didapat

$$DFBETAS_{j.i} = \frac{b_j - b_{j.-i}}{s_{-i}\sqrt{C_{jj}}} = \frac{b_j - b_{j.-i}}{s_{-i}\sqrt{\sum_{i=1}^{n} r_{jj}^2}}$$

$$= \frac{r_{j.i}\, e_i}{1-h_{ii}} \frac{1}{s_{-i}\sqrt{\sum_{i=1}^{n} r_{jj}^2}}$$

$$= \frac{r_{j.i}}{\sqrt{r_j^T r_j}} \frac{e_i}{s_{-i}(1-h_{ii})}$$

$$= \frac{r_{j.i}}{\sqrt{r_j^T r_j}} \frac{1}{\sqrt{1-h_{ii}}} \text{ (}R\text{-}student\text{)}$$

Ukuran kritis untuk $DFBETAS_{j.i}$ yaitu $2/\sqrt{n}$. Jika nilai $\left|DFBETAS_{j.i}\right| > 2/\sqrt{n}$ mengindikasikan pengamatan ke-*i* berpengaruh pada koefisien ke-*j* (Myers,1990).

2.7 Regresi Himpunan Bagian (*Subset*)

Ada beberapa prosedur statistik tertentu yang dapat menentukan variabel yang akan dimasukkan dalam regresi, misal ingin menentukan suatu persamaan regresi linier variabel respon tertentu *Y* terhadap variabel bebas *X*. Dalam kaitannya ada dua kriteria yang saling bertentangan:

1. Agar persamaannya bermanfaat bagi tujuan peramalan, dimasukkan sebanyak mungkin variabel *X* sehingga diperoleh nilai ramalan yang terandalkan.
2. Karena untuk memperoleh informasi dari banyak variabel bebas *X* serta pemonitorannya seringkali diperlukan biaya yang tinggi, maka diinginkan persamaan regresinya mencakup sedikit mungkin variabel *X*.

Ada beberapa algoritma yang dapat dipergunakan untuk pemilihan himpunan bagian terbaik peubah peramal dalam regresi. Algoritma dapat menghitung hanya sebagian dari semua kemungkinan regresi dalam menentukan himpunan bagian "*K* terbaik". Beberapa kriteria yang dapat digunakan untuk menentukan himpunan bagian "*K* terbaik" yaitu *adj*-R^2 maksimum dan S^2 minimum. Algoritma yang digunakan dapat menghasilkan *K* regresi terbaik dengan satu peubah peramal, *K* regresi terbaik dengan dua peubah peramal, dan seterusnya sampai persamaan regresi yang mencakup semua peubah peramal.

Misalkan ada 3 variabel X_1, X_2 dan X_3 , kelompokkan persamaan regresi kedalam 3 kelompok :

Kelompok yang terdiri atas persamaan regresi dengan 1 peubah peramal, dengan model :

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

$$Y = \beta_0 + \beta_2 X_2 + \varepsilon$$

$$Y = \beta_0 + \beta_3 X_3 + \varepsilon$$

Kelompok yang terdiri atas persamaan regresi dengan 2 peubah peramal, dengan model :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$$

Kelompok yang terdiri atas persamaan regresi dengan 3 peubah peramal, dengan model :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

(Draper & Smith, 1992).

2.8 Kriteria Seleksi Model

2.8.1 *Mean Square Error* (MSE)

Mean Square Error (MSE) dapat didefinisikan sebagai perbandingan antara *Sum Square Error* (SSE) dan derajat bebas suatu galat. Misalnya diketahui model regresi sederhana sebagai berikut.

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

$$Y_i = \widehat{Y}_i + e_i$$

$$e_i = \ Y_i - \widehat{Y}_i$$

Maka SSE dapat ditulis dalam persamaan berikut.

$$SSE = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Sehingga SSE mempunyai *n*-2 derajat bebas. Kuadrat tengah galat (*Mean Square Error*) yang tepat dinotasikan oleh MSE atau S^2, dapat ditulis dalam persamaan berikut.

$$S^2 = MSE = \frac{SSE}{n-2}$$

Hal ini juga ditunjukan bahwa MSE adalah penduga tak bias dari σ^2, sehingga:

$$\text{E(MSE)}=\sigma^2$$

Sebagai nilai standar deviasi penduga σ adalah $S = \sqrt{MSE}$

(Neter & Kutner, 1990).

MSE yang disimbolkan dengn S^2 merupakan salah satu patokan yang baik digunakan dalam menilai kecocokan suatu model. Semakin kecil MSE maka model semakin baik. Ukuran ini memperhitungkan banyaknya parameter dalam model melalui pembagian dengan derajat bebasnya. S^2 mungkin membesar bila penurunan dalam SSE akibat pemasukan suatu variabel tambahan ke dalam model tidak dapat mengimbangi penurunan dalam derajat bebasnya. Menurut Sembiring (1995), rumus umum dari MSE diberikan sebagai berikut:

$$MSE = \frac{SSE}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p}$$

2.8.2 R^2 disesuaikan (*Adjusted*- R^2)

Membandingkan dua atau lebih model regresi dan himpunan bagian dari model misalkan seperti $\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ dengan $\hat{y} = \beta_0 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$ penggunaan R^2 lebih sesuai. Namun R^2 memiliki salah satu kelemahan yaitu besarnya dipengaruhi oleh banyaknya variabel dalam model. R^2 akan cenderung membesar bersama *p*, sehingga sulit menyatakan R^2 yang optimum. Untuk mengatasi kesulitan dari interpretasi R^2, maka digunakanlah statistik *Adjusted*-R^2 (R^2 yang disesuaikan). Penyesuaiannya yaitu membagi *Sum Square Error* (SSE) dan *Sum Square Total* (SST) dengan masing-masing derajat bebasnya. Menurut Sembiring (1995), rumus umum dari *Adjusted*-R^2 diberikan sebagai berikut :

$$Adj\text{-}R^2 = 1 - \frac{\frac{SSE}{n-p}}{\frac{SST}{n-1}} = 1 - \left(\frac{n-1}{n-p}\frac{SSE}{SST}\right)$$

$$= 1 - \frac{n-1}{n-p}\left(\frac{SSE}{SST}\right)$$

$$= 1 - \frac{n-1}{n-p}(1 - R^2)$$

Statistik $Adj\text{-}R^2_p$ belum tentu meningkat seiring pertambahan variabel ke dalam model. Faktanya bahwa jika *k* variabel *x* (regresor) ditambahkan pada model , $Adj\text{-}R^2_{p+k}$ akan melebihi $Adj\text{-}R^2_p$ jika dan hanya jika statistik parsial-*F* untuk uji signifikan pada penambahan *k* variabel *x* (regresor) melebihi 1. Konsekuen, satu kriteria seleksi pada model himpunan bagian (*subset*) optimum adalah dengan memilih model yang memiliki maksimum $Adj\text{-}R^2_{p.}$

Kriteria seleksi model regresi himpunan bagian selain dengan minimum MSE dapat juga dengan maksimum $Adj\text{-}R^2$. Hubungan keduanya sebagai berikut :

$$Adj\text{-}R^2 = 1 - \frac{n-1}{n-p}(1 - R^2)$$

$$= 1 - \frac{n-1}{n-p}\frac{SSE}{SST}$$

$$= 1 - \frac{(n-1)MSE}{SST}$$

Dari rumus diatas maka kriteria minimum MSE dan maksimum $Adj\text{-}R^2$ ekuivalen (Montgomery, Peck, & Vinning, 2006).