BAB III

KECOCOKAN MODEL REGRESI LINIER GANDA

3.1. Uji Hipotesa dalam Regresi Linier Ganda.

Dalam permasalahan model regresi linier ganda, uji hipotesa mengenai parameter-parameter model selalu berguna dalam mengukur kecocokan model. Dalam bagian ini, kita akan membuat beberapa prosedur uji hipotesa yang penting, selanjutnya kita akan memerlukan asumsi kenormalan yang sudah dibicarakan pada Bab II.

3.1.1. Uji untuk Signifikan Persamaan Regresi.

Uji untuk signifikan persamaan regresi adalah uji untuk menentukan jika ada hubungan linier antara respon Y dan beberapa variabel prediktor X_1 , X_2 , ... , X_k. Hipotesa yang tepat adalah :

$$H_0 : \beta_1 = \beta_2 = \quad = \beta_k = 0$$

$$H_1 : \beta_j \neq 0 \qquad (3.1.1)$$

Menolak $H_0 : \beta_j = 0$ menyatakan bahwa sedikitnya satu variabel dalam variabel prediktor X_1, X_2 , ... , X_k memperbesar signifikan untuk model. Prosedur ujinya sama seperti dalam regresi linier sederhana jumlah total kuadrat (SYY) adalah jumlahan antara sebuah jumlah kuadrat regresi (SSR) dan jumlah kuadrat residual (SSE).

$$SYY = SSR + SSE \qquad (3.1.2)$$

dimana

$$SYY = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

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

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

Rumus untuk menghitung SSR adalah dari rumus (2.7) yaitu

$$SSE = Y^1 Y - \hat{\beta}^1 X^1 Y$$

dan karena

$$SYY = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{(\Sigma Y_i)^2}{n}$$

kita menulis rumus (2.7) sebagai

$$SSE = \underset{\sim}{Y}' \underset{\sim}{Y} - \frac{(\Sigma Y_i)^2}{n} - [\underset{\sim}{\hat{\beta}}' X' \underset{\sim}{Y} - \frac{(\Sigma Y_i)^2}{n}]$$

dan dari rumus (3.12) SSE = SYY - SSR, sehingga jumlah kuadrat regresi (SSR) adalah :

$$SSR = \underset{\sim}{\hat{\beta}}' X' \underset{\sim}{Y} - \frac{(\sum_{i=1}^{n} Y_i)^2}{n} \qquad (3.13)$$

dan jumlah kuadrat total (SYY) adalah :

$$SYY = \underset{\sim}{Y}' \underset{\sim}{Y} - \frac{(\sum_{i=1}^{n} Y_i)^2}{n} \qquad (3.14)$$

Menurut teorama cochran yang berbunyi :

"Jika seluruh n pengamatan Y_i berasal dari distribusi normal yang sama dengan rata-rata $\underline{Y} = x\,\beta$ dan variabel σ^2 , dan SYY diuraikan kedalam k jumlah kuadrat SSR, masing-masing dengan derajad kebebasan , maka $\frac{SSr}{\sigma^2}$ merupakan distribusi λ^2 dengan derajat kebebasan jika:

Jumlah derajat kebebasan = n - 1

maka

$$(1)\quad \frac{SSE}{\sigma^2} = \sum_{i=1}^{n} \frac{(Y_i - \bar{Y}_i)^2}{\sigma^2} \sim \chi^2_{n-k-1}$$

dimana banyaknya derajat kebebasan untuk χ^2 adalah n-k-1.

$$(2)\quad \frac{SSR}{\sigma^2} = \sum_{i=1}^{n} \frac{(\hat{Y}_i - \bar{Y})^2}{\sigma^2} \sim \chi^2_{k}$$

dimana banyaknya derajat kebebasan untuk χ^2 adalah banyaknya variabel prediktor dalam model karena SSE dan SSR adalah independen, prosedur uji untuk $H_0 : \beta_j = 0$ adalah menghitung :

$$F_0 = \frac{SSR \,/\, k}{SSE \,/\, n-k-1} = \frac{MSR}{MSE}$$

dan menolak H_0 jika $F_0 > F_{\alpha,\, k,n-k-1}$

3.1.2. Uji Koefisien Regresi Individual.

Hipotesa untuk menguji signifikan beberapa koefisien regresi individual, misalnya β_j adalah :

$H_0 : \beta_j = 0$

$H_1 : \beta_j \neq 0$

Jika $H_0 : \beta_j = 0$ diterima, maka menunjukkan bahwa variabel yang bisa dikeluarkan dari model. Seperti telah dibicarakan pada Bab II bahwa sesuatu random C_i di asumsikan berdistribusi normal dan independen dengan mean 0 dan variansi σ^2 Oleh karena itu pengamatan - pengamatan Y_i merupakan distribusi normal dan independen dengan mean $\beta_0 + \sum_{j=1}^{k} \beta_j X_j$ dan variansi σ^2 karena estimator kuadrat terkecil $\hat{\beta}$ adalah kombinasi linear

dengan pengamatan Y_i, maka $\hat{\beta}$ merupakan berdistribusi normal dengan mean vektor β dan kovarian matrik $\delta^2 (X^{'} X)^{-1}$ Hal ini menyatakan bahwa distribusi pada beberapa koefisien regresi $\hat{\beta}_j$ adalah normal dengan mean $\hat{\beta}_j$ dan variansi $\delta^2 C_{jj}$ dimana C_{jj} adalah elemen diagonal ke j dalam matrik $(X^{'}X)^{-1}$

Konswekwensinya setiap statistik.

$$t_0 = \frac{\hat{\beta}_j - 0}{\sqrt{\hat{\delta}^2 C_{jj}}}$$

Adalah berdistribusi t dengan n-k-1 derajat bebas. Hipotesa nol $H_0 : \beta_j = 0$ adalah ditolak jika $| t_0 | > t_{0\ \alpha/2,\ n-k-1}$.

Perhatikan bahwa ini sesungguhnya uji parsial. sebab koefisien regresi $\hat{\beta}_j$ tergantung pada semua variabel-variabel prediktor yang lain X_i $(i \neq j)$ yang berada dalam model. Jadi ini adalah sebuah uji besarnya pengaruh X_j yang diberikan kepada variabel-variabel yang lain yang ada dalam model.

Kita mungkin juga secara langsung bisa menentukan besarnya pengaruh untuk jumlah kuadrat regresi (SSR) sebuah variabel prediktor. Untuk misalnya, X_j, sumbangan yang diberikan untuk variabel prediktor yang lain X_i $(i \neq j)$ yang termuat dalam model dengan menggunakan metode "jumlah kuadrat ekstra". Prosedur ini bisa juga digunakan untuk menyelidiki pengaruh sebuah subset variabel-variabel prediktor untuk model. Anggap model regresi dengan k variabel-variabel prediktor.

$\underset{\sim}{Y} = \underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$ dimana $\underset{\sim}{Y}$ adalah matrik vektor (nx1), X adalah matrik (nxp) , $\underset{\sim}{\beta}$ adalah matrik nx1 dan p = k+1 kita ingin menentukan, jika beberapa subset r < k variabel -variabel prediktor memberikan sumbangan dalam hubungan yang nyata untuk model regresi. Misalkan vektor koefisien regresi di partisi sebagai berikut :

$$\underset{\sim}{\beta} = \begin{vmatrix} \underset{\sim}{\beta}_1 \\ \text{----} \\ \underset{\sim}{\beta}_2 \end{vmatrix}$$

dimana $\underset{\sim}{\beta}_1$ adalah matrik (p-r) x 1 dan $\underset{\sim}{\beta}_2$ adalah matrik (rx1) Kita ingin menguji hipotesa.

$H_0 : \underset{\sim}{\beta}_2 = 0$

$H_1 : \underset{\sim}{\beta}_2 \neq 0$ (3.15)

Model bisa ditulis sebagai

$\underset{\sim}{Y} = X\,\underset{\sim}{\beta} + \underset{\sim}{\epsilon} = X_1\underset{\sim}{\beta}_1 + X_2\underset{\sim}{\beta}_2 + \underset{\sim}{\epsilon}$ (3.16)

dimana X_i adalah matrik n x (p-r) yang mewakili kolom X yang berkaitan dengan $\underset{\sim}{\beta}_1$ dan X_2 adalah matrik nxr yang mewakili kolom yang berkaitan dengan $\underset{\sim}{\beta}_2$ Persamaan ini dinamakan "full model". Untuk "full model" kita tahu bahwa $\hat{\underset{\sim}{\beta}} = (X'X)^{-1} X'\underset{\sim}{Y}$ Jumlah kuadrat regresi untuk model ini adalah.

$SSR(\hat{\underset{\sim}{\beta}}) = \hat{\underset{\sim}{\beta}}' X'\underset{\sim}{Y}$ (p derajad bebas)

dan

$$MSE = \frac{\underset{\sim}{Y}'\underset{\sim}{Y} - \hat{\underset{\sim}{\beta}}' X'\underset{\sim}{Y}}{n - p}$$

Untuk mendapatkan kontribusi suku-suku dalam $\underset{\sim}{\beta}_2$ untuk regresi ialah mencocokan model dengan mengasumsikan bahwa hipotesa nol $H_0 : \underset{\sim}{\beta}_2 = 0$ adalah benar."Reduced model" ini adalah

$$\underset{\sim}{Y} = X \underset{\sim}{\beta} + \underset{\sim}{\epsilon} \qquad (3.17)$$

Estimator kuadrat terkecil β_1 dalam "reduced model" adalah

$$\hat{\underset{\sim}{\beta}}_1 = (X_1'X_1)^{-1} X_1'\underset{\sim}{Y}$$

Jumlah kuadrat regresi adalah

$$SSR(\underset{\sim}{\beta}_1) = \hat{\underset{\sim}{\beta}}_1^1 X_1^1 \underset{\sim}{Y} \quad (\text{p-r derajad bebas}) \qquad (3.18)$$

Sumbangan jumlah kuadrat regresi dari $\underset{\sim}{\beta}_2$ yang diberikan bahwa $\underset{\sim}{\beta}_1$ sudah ada dalam modelnya adalah :

$$SSR(\underset{\sim}{\beta}_2/\underset{\sim}{\beta}_1) = SSR(\underset{\sim}{\beta}) - SSR(\underset{\sim}{\beta}_1) \qquad (3.14)$$

Dengan p-1-(p-r-1) = r derajat bebas, jumlah kuadrat ini disebut "Jumlah kuadrat ekstra" dari $\underset{\sim}{\beta}_2$, sebab jumlah kuadrat ekstra ini mengukur pertambahan jumlah kuadrat regresi yang merupakan hasil dari pertambahan variabel-variabel prediktor $X_{k-r+1}, X_{k-r+2}, \ldots, X_k$, untuk sebuah model yang sudah memuat $X_1, X_2, \ldots, X_{k-r}$.

Sekarang $SSR(\underset{\sim}{\beta}_2/\underset{\sim}{\beta}_1)$ adalah independen dengan MSE dan hipotesa nol $\underset{\sim}{\beta}_2 = 0$ bisa diuji dengan statistik.

$$F_0 = \frac{SSR\ (\beta_2/\beta_1)/r}{MSE(\beta_1\beta_2)}$$

Jika $F_0 > F_{\alpha,r,n-p}$ kita tolak H_0 , kesimpulan bahwa sedikit-dikitnya satu dari parameter-parameter X_{k-r+1} , X_{k-r+2}, ... , $X_k$dalam X_2 memberikan sumbangan bahwa ada hubungan nyata pada model.

Beberapa ahli menamakan uji dalam (318) adalah sebuah uji F parsial, sebab uji untuk mengukur kotribusi parsial X_2 pada variabel prediktor dalam X_1 yang sudah ada dalam model. Untuk menggambarkan kegunaan prosedur ini, pandang model :

$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \epsilon$

Jumlah kuadrat :

$SSR\ [\beta_1/\beta_0\ ,\ \beta_2\ ,\ \beta_3]$

$SSR\ [\beta_2/\beta_0\ ,\ \beta_1\ ,\ \beta_3]$

$SSR\ [\beta_3/\beta_0\ ,\ \beta_1\ ,\ \beta_2]$

adalah jumlah kuadrat regresi dengan derajat bebas tunggal yang mengukur kontribusi setiap variabel prediktor X_j , j= 1 , 2 , 3 , untuk model yang variabel prediktor lainnya sudah ada dalam model. Sehingga kita bisa menaksir harga X_j yang dimasukan ke sebuah model yang tidak memuat variabel prediktor ini.

Pada umumnya, kita bisa mendapatkan jumlah kuadrat regresi :

$SSR\ [\beta_j/\beta_0\ ,\ \beta_1\ ,\ \ldots\ ,\ \beta_{j-1}\ ,\ \beta_{j+1}\ ,\ \ldots\ ,\ \beta_k\ ]$

$1 \leq j \leq k$

yang merupakan pertambahan jumlah kuadrat regresi dari X_j yang dimasukkan ke sebuah model yang sudah ada X_1,

$X_2, \ldots, X_{j-1}, X_{j+1}, X_k$.

Jadi variabel X_j dianggap seolah-olah sebagai variabel terakhir yang dimasukkan ke dalam model. Dan ide ini digunakan untuk memilih kumpulan variabel yang terbaik untuk digunakan dalam model.

3.1.3. Keadaan Khusus Kolom orthogonal dalam Matrik X.

Pandang model (3.16)

$$\underset{\sim}{Y} = X \underset{\sim}{\beta} + \underset{\sim}{\epsilon}$$

$$= X_1\beta_1 + X_2\beta_2 + \underset{\sim}{\epsilon} \qquad (3.21)$$

Seperti dalam metode jumlah kuadrat ekstra, kita bisa mengukur pengaruh variabel prediktor X_2 pada X_1 dengan menghitung SSR (β_2/β_1).

Pada umumnya, kita tidak bisa bicara mengenai besarnya jumlah kuadrat dari β_2 (SSR(β_2), tanpa memperhatikan keterkaitan jumlah kuadrat regresi ini dengan variabel prediktor X_1. Namun jika kolom-kolom X_1 adalah orthogonal dengan kolom-kolom X_2, kita dapat menentukan sebuah jumlah kuadrat dan β_2 yang bebas dari beberapa ketergantungan pada variabel-variabel prediktor dalam X_1.

Untuk mengambarkan bentuk ini persamaan normal $(X^1X)\,\hat{\beta} = X^1 Y$. Untuk model (3.2.1) adalah :

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ \hline X_1'X_1 & X_1'X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \hline \beta_2 \end{bmatrix} = \begin{bmatrix} X_1'Y \\ \hline X_2'Y \end{bmatrix}$$

Sekarang, jika kolom X_1 adalah orthogonal dengan kolom

X_2, maka $X_1'X_2 = 0$ dan $X_2'X_1 = 0$, persamaan normal menjadi

$X_1'X_1\hat{\beta}_1 = X_1'Y$, $X_2'X_2\hat{\beta}_2 = X_2'Y$

dengan penyelesaian

$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'Y$

$\hat{\beta}_2 = (X_2'X_2)^{-1}X_2'Y$

Perhatikan bahwa estimator kuadrat terkecil β_1 adalah $\hat{\beta}_1$, tanpa memperhatikan apakah X_2 ada atau tidak didalam model, dan estimator kuadrat terkecil β_2 adalah $\hat{\beta}_2$, tanpa memperhatikan ada atau tidak X_1 didalam model.

Jumlah kuadrat regresi untuk "full model" adalah :

$$SSR(\beta) = \hat{\beta}'X'Y$$

$$= [\hat{\beta}_1 \ \hat{\beta}_2] \begin{vmatrix} X_1'Y \\ X_2'Y \end{vmatrix}$$

$$= \hat{\beta}_1' X_1' Y + \hat{\beta}_2' X_2' Y$$

$$= Y' X_1 (X_1' X_1)^{-1} X_1' Y + Y' X_2 (X_2' X_2)^{-1} X_2' Y \qquad (3.2.2)$$

Jumlah kuadrat untuk β_2 dan β_1 adalah : $SSR(\beta_1) = \hat{\beta}_1' X_1' Y = Y' X_1 (X_1'X_1)^{-1} X_1'Y$

$SSR(\beta_2) = \hat{\beta}_2' X_2' Y = Y' X_2 (X_2'X_2)^{-1} X_2'Y$ (3.2.3)

Bandingkan (3.2.2) dengan (3.2.3) tampak bahwa :

$SSR(\beta) = SSR(\beta_1) + SSR(\beta_2)$ (3.2.4)

Oleh karena itu

$SSR(\beta_2/\beta_1) = SSR(\beta) - SSR(\beta_1)$

$= SSR(\beta_2)$

dan

$SSR(\beta_1/\beta_2) = SSR(\beta) - SSR(\beta_2)$

$= SSR(\beta_1)$

Konsekuensinya , $SSR(\beta_1)$ mengukur kontribusi variabel-variabel prediktor X_1 ke model dengan tanpa syarat, dan $SSR(\beta_2)$ mengukur kontribusi variabel-variabel prediktor ke model dengan tanpa syarat juga. Menyebabkan kita bisa mempunyai 2 penilaian dalam menetapkan pengaruh setiap variabel prediktor jika variabel prediktornya orthogonal, kumpulan data percobaan sering kali di disain untuk mempunyai variabel-variabel orthogonal.

Sebagai contohnya dalam sebuah model regresi dengan variabel-variabel prediktor orthogonal, pandang model :

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

dimana matriks X adalah :

$$X = \begin{array}{c} \begin{matrix} \underline{\beta_0} & \underline{\beta_1} & \underline{\beta_2} & \underline{\beta_3} \end{matrix} \\ \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

Mudah dilihat bahwa kolom-kolom x adalah orthogonal. jadi $SSR(\beta_j)$, j = 1 , 2 , 3 , mengukur kontribusi variabel-variabel prediktor X_j untuk model tanpa

memperhatian apakah beberapa dari variabel-variabel prediktor yang lain termuat dalam model.

3.2. Koefisien Determinasi Ganda.

Koefisien determinasi ganda juga bisa untuk menaksir kecocokan model, didefinisikan sebagai :

$$R^2 = \frac{SSR}{SYY} = 1 - \frac{SSE}{SYY} \qquad (3.28)$$

R^2 adalah sebuah pengukuran pengurangan dalam varibilitas Y yang diperoleh dengan menggunakan variabel-variabel prediktor X_1 , X_2 , ... , X_k.

Seperti dalam regresi linier sederhana, kita mempunyai $0 \leq R^2 \leq 1$.

Namun besarnya harga R^2 tidak menjamin bahwa model regresi adalah bagus. Penambahan sebuah variabel prediktor akan selalu menambah R^2 , tanpa memperhatikan apakah ada atau tidak kontribusi penambahan variabel prediktor untuk model.

Akar kuadrat positif R^2 adalah koefisiensi korelasi ganda antara Y dan sekumpulan variabel-variabel prediktor X_1 , X_2 , ... , X_k.

Sehingga R adalah sebuah pengukuran hubungan linier antara Y dan X_1 , X_2 , ... , X_k.

Sifat-sifat R^2 :

- $R^2 = 0$, jika semua β_1 , β_2 , β_3 , ... , $\beta_k = 0$. karena SSR = 0 atau $\hat{Y}_i = \bar{Y}_i$.
- $R^2 = 1$, jika semua Y_i terletak pada permukaan respone (karena $Y_i = \hat{Y}_i \ \forall_i$ atau SSE = 0.

- Karena $0 \leq SSE \leq SYY$, maka $0 \leq R^2 \leq 1$.

3.3. Analisa Residual..

Residual-residual e_i dari model regresi ganda memainkan peranan penting dalam menduga kecocokan model. Ada beberapa metoda analisa residual yang lain yang berguna dalam regresi ganda.

3.3.1. Plot Residual Biasa.

Didefinisikan residual-residual sebagai $e_i = Y_i - \hat{Y}_i$, $i = 1 , 2 , \ldots , n$. dimana Y_i adalah harga dari suatu pengamatan dan Y_i adalah harga dari suatu model yang cocok, karena sebuah residual mungkin bisa memperlihatkan deviasi antara data dan model yang cocok,maka plot residual biasa adalah sebuah pengukuran variabilitas yang tidak bisa diterangkan dengan model regresi, dan juga plot residual juga sangat tepat dalam menaksir harga-harga pengamatan yang salah.

3.3.2. Plot Residual Parsial.

Plot-plot ini dirancang untuk menampilkan ketepatan hubungan antara residual-residual dan variabel prediktor X_j. Didefinisikan residual parsial ke i untuk variabel prediktor X_j sebagai :

$$e_{ij} = Y_i - \hat{\beta}_1 X_{i1} - \ldots - \hat{\beta}_{j-1} X_{i,j-1} - \hat{\beta}_{j+1} X_{i,j+1} - \ldots - \hat{\beta}_k X_{ik}$$

$$= e_i + \Sigma \hat{\beta}_j X_{ij} \qquad i = 1 , 2 , \ldots , n$$

Plot e_{ij} pada X_{ij} disebut sebuah plot residual parsial, plot-plot ini dikemukakan oleh Zekiel dan Pox [1959] dan Lansen dan Meleary [1972].

Seperti pada plot residual biasa, plot residual parsial berguna dalam menyelidiki ketidaksamaan varians. Namun karena plot-plot residual parsial menggambarkan hubungan antara Y dan variabel-variabel prediktor X_j. Setelah pengaruh variabel-veriabel prediktor $X_i (i \# j)$ lainnya telah dikeluarkan, maka plot residual parsial lebih jelas memperlihatkan pengaruh X_j pada respon Y dengan adanya variabel-variebel prediktor lainnya.

Mengingat persamaan regresi linier melalui e_{ij} pada X_{ij}, kemungkinan garis kuadrat terkecil untuk persamaan regresi ini adalah β_j dan cara mencari $\hat{\beta}_j$ adalah sama dengan harga $\hat{\beta}_j$ pada "full model" dengan k variabel. Oleh karena itu plot residual parsial akan selalu mempunyai kemiringan $\hat{\beta}_j$ lebih baik daripada kemiringan zero pada plot residual biasa.

Penggambaran ini memperbolehkan peneliti lebih mudah menaksir asal mula linearitas atau timbulnya ketidaksamaan varians.