REGRESI MULTIPLE

April 7, 2024

Sains Data IKOPIN
Pengantar Model Linear
github.com/sevinurafni/SD2201

Model Regresi Multiple

Kasus pada kehidupan nyata hubungan antara variabel bukan hanya terdiri dari dua variabel. Tetapi juga dapat terjadi hubungan antara tiga variabel atau bahkan lebih. Misalnya, apakah kecenderungan berprestasi ini hanya cukup ditinjau dari motivasi saja? mengapa tidak dari ciri kepribadian, kecerdasan dan barangkali juga dari tingkat pendidikan. Apabila ingin dicari hubungan antar variabel-variabel tersebut maka, harus digunakan suatu model yang disebut persamaan regresi multiple dengan persamaan matematikanya adalah

$$Y = \alpha_0 + \alpha_1 X_{1i} + ... + \alpha_k X_{ki} + \varepsilon_n \tag{1}$$

apabila ditulis dalam bentuk matriks diperoleh

$$Y = \alpha X + \varepsilon \tag{2}$$

dengan,

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \vdots & X_{nk} \end{bmatrix}, \qquad \alpha = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}, \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{3}$$

dengan metode kuadra terkecil dapat dicari taksiran dari α sebagai berikut:

$$\alpha = (X^\top X)^{-1}(X^\top Y) \tag{4}$$

Untuk mengetahui apakah koefisien regresi mempunyai arti atau tiddak maka, perlu diperlakukan pengujian dengan menggunakan daftar Anova, sedangkan bentuk hipotesisnya adalah:

$H_0 : \alpha^* \rightarrow$ model yang digunakan $\quad \hat{Y} = \alpha_0 + \varepsilon$

$H_1 : \alpha^* \rightarrow$ model yang digunakan $\quad \hat{Y} = \alpha X + \varepsilon$

Sumber Variasi	Dk	Jk
Regresi pada $X_1, \cdots, X_k$	k	$\alpha^\top (X^\top Y)\alpha$
Residu	n-k-1	$(Y^\top Y) - \alpha^\top (X^\top Y)\alpha$
Total	n-1	$(Y^\top Y)$

Mencari nilai varians b dengan:

$$\hat{V}(b) = \begin{bmatrix} \hat{V}(b_0) & cov(b_0,b_1) & \cdots & cov(b_0,b_k) \\ cov(b_0,b_1) & \hat{V}(b_1) & \cdots & cov(b_1,b_k) \\ \cdots & \cdots & \cdots & \cdots \\ cov(b_k,b_0) & cov(b_k,b_1) & \cdots & \hat{V}(b_k) \end{bmatrix} \sigma^2 = (X^T X)^{-1}\sigma^2 \tag{5}$$

Contoh:

Lima rumah tangga petani dari suatu daerah pertanian dipilih sebagai sampel acak untuk diteliti tentang pengaruh pendapatan (X_1) dan kekayaan (X_2) terhadap tingkat komsumsinya (Y). Dari hasil peneltian diperoleh data sebagai berikut:

X_1	80	110	90	60	60
X_2	120	60	60	30	180
Y	74	98	80	53	57

Tentukanlah:

a. Persamaan regresi multiple

b. Uji keberartian koefisien regresi

Penyelesaian:

Apabila data di atas ditulis dalam bentuk matrik maka diperoleh:

$$Y = \begin{bmatrix} 74 \\ 98 \\ 80 \\ 53 \\ 57 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 80 & 120 \\ 1 & 110 & 60 \\ 1 & 90 & 60 \\ 1 & 60 & 30 \\ 1 & 60 & 180 \end{bmatrix}$$

$$(Y^\top Y) = \begin{bmatrix} 74 & 98 & 80 & 53 & 57 \end{bmatrix} \begin{bmatrix} 74 \\ 98 \\ 80 \\ 53 \\ 57 \end{bmatrix} = 27538$$

$$(X^\top X) = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 80 & 110 & 90 & 60 & 60 \\ 120 & 60 & 60 & 30 & 180 \end{bmatrix} \begin{bmatrix} 1 & 80 & 120 \\ 1 & 110 & 60 \\ 1 & 90 & 60 \\ 1 & 60 & 30 \\ 1 & 60 & 180 \end{bmatrix}$$

$$= \begin{bmatrix} 5 & 400 & 450 \\ 400 & 33800 & 34200 \\ 450 & 34200 & 54900 \end{bmatrix}$$

$$(X^\top Y) = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 80 & 110 & 90 & 60 & 60 \\ 120 & 60 & 60 & 30 & 180 \end{bmatrix} \begin{bmatrix} 74 \\ 98 \\ 80 \\ 53 \\ 57 \end{bmatrix}$$

$$= \begin{bmatrix} 362 \\ 30500 \\ 31410 \end{bmatrix}$$

a.

$$\alpha = (X^\top X)^{-1}(X^\top Y)$$

$$= \begin{bmatrix} 68598.10^4 & -657.10^4 & -153.10^4 \\ -657.10^4 & 72.10^3 & 9.10^3 \\ -153.10^4 & 9.10^3 & 9.10^3 \end{bmatrix} \begin{bmatrix} 362 \\ 30500 \\ 31410 \end{bmatrix} = \begin{bmatrix} -1{,}0365 \\ 0{,}8849 \\ 0{,}0294 \end{bmatrix}$$

sehingga persamaan regresinya adalah $\hat{y} = -1{,}0365 + 0{,}8849x_1 + 0{,}0294x_2$

b.

Sumber Variasi	Dk	Jk	RJk
Regresi pada X_1, X_2	2	27537,691	13768,8455
Residu	2	0,309	0,1545
Total	4	27538	

$F_{perhitungan} = \frac{13768{,}8455}{0{,}1545} = 89118{,}7411$
$F_\alpha; 2:2 = 19.0$

H_0 : ditolak maka H_1 diterima atau model $\hat{Y} = \alpha X + \varepsilon$ diterima

Kesimpulan Dalam Regresi Multiple

Dengan statistika kita berusaha untuk menyimpulkan populasi. Untuk itu sifat populasi dipelajari berdasarkan data yang diambil baik secara sampling maupun sensus. Sifat populasi yang akan ditinjau hanya mengenai parameter populasi dan sample yang digunakan secara acak.

Cara pengambilan kesimpulan yang dibicarakan adlah interval taksiran dan pengujian jipotesis. Sehubungan dengan masalah regresi multiple maka yang dibahas adalah mengenai pengujian hipotesis dan interval taksiran tentang koefisien regresi multiple rata-rata respon dan nilai respon yang tunggal.

A. Pengujian Hipotesis

Menguji keberartian koefisien regresi multiple dari persamaan regresi:

$$Y = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \varepsilon_n \tag{6}$$

maka hipotesisnya berbentuk:

$H_0 : \beta_j = 0$ (tidak ada hubungan/pengaruh X_j terhadap Y)

$H_1 : \beta_j \neq 0$ (ada hubungan/pengaruh X_j terhadap Y)

Statistik yang digunakan adalah

$$t = \frac{\beta_j}{\sqrt{C_{(j+1),(j+1)}}\ \sigma} \tag{7}$$

dengan:

$b_j - \beta_{jo}$: adalah nilai koefisien β_j

$\sqrt{C_{(j+1),(j+1)}}$: Unsur ke $(j+1)$ diagonal $(X'X) - 1$

terima H_0 jika:

$$-t_{\frac{\alpha}{2}} < t < t_{\frac{\alpha}{2}}$$

dengan

$$t_{\frac{\alpha}{2}} = t_{\frac{\alpha}{2}}; n - k - 1$$

Pengujian ini disebut uji-t untuk menguji signifikansi secara individual dari koefisien regresi dari masing-masing variabel independen. Ini memberikan informasi apakah variabel independen tersebut secara signifikan berkontribusi terhadap variabel dependen. Uji t menghasilkan nilai t-statistik, di mana nilai tersebut dibandingkan dengan nilai kritis dari distribusi t untuk menentukan apakah koefisien regresi tersebut signifikan secara statistik atau tidak.

Kemudian dilakukan juga Analisis Varians (ANOVA) regresi atau yang dikenal uji-f. Digunakan untuk menguji signifikansi keseluruhan model regresi, yaitu apakah setidaknya satu variabel independen secara bersama-sama mempengaruhi variabel dependen. Uji F menghasilkan nilai F-statistik, yang juga dibandingkan dengan nilai kritis dari distribusi F untuk menentukan apakah model regresi secara keseluruhan signifikan atau tidak.

hipotesisnya berbentuk:

$$H_0 : \beta_j = 0$$
$$H_1 : \text{min ada satu } \beta_j \neq 0$$

berarti:
H_0: variabel dependen tidak mempunyai hubungan linear dengan variabel independen.
H_1: variabel dependen mempunyai hubungan linear dengan variabel independen.

Tabel Anova yang digunakan adalah sebagai berikut:

Sumber Variasi	Dk	Jk	Rjk
Regresi pada $\beta_1, \beta_2, \cdots, \beta_k \vert \beta_0$	k	$\beta^\top (X^\top Y) - \frac{(\sum Y_i)^2}{n}$	$\frac{\text{Jk regres}}{k}$
Kekeliruan	n-k-1	$(Y^\top Y) - \beta^\top (X^\top Y)$	$\frac{\text{Jk Kekeliruan}}{n-k-1}$
Total	n-1	$(Y^\top Y) - \frac{(\sum Y_i)^2}{n}$	

Statistik yang digunakan adalah

$$F = \frac{\text{RJK regresi}}{\text{RJK residu}} \tag{8}$$

Tolak H_0 jika:

$$F > F\alpha; n-k-1$$

Contoh:

Dari hasil penelitian diperoleh data sebagai berikut:

Y	X1	X2	X3
25.5	1.74	5.3	10.8
31.2	6.32	5.42	9.4
25.9	6.22	8.41	7.2
38.4	10.52	4.63	8.5
18.4	1.19	11.6	9.4
26.7	1.22	5.85	9.9
26.4	4.1	6.62	8
25.9	6.32	8.72	9.1
32	4.08	4.42	8.7
25.2	4.15	7.6	9.2
39.7	10.15	4.83	9.4
35.7	1.72	3.12	7.6
26.5	1.7	5.3	8.2

Apakah koefisien regresi mempunyai arti atau tidak?

Penyelesaian:

Untuk menentukan persamaan regresi multiple, maka nilai β diperoleh dari:

$$\beta = (X^\top X)^{-1}(X^\top Y)$$

$$= \begin{bmatrix} 13 & 59.43 & 81.82 & 115.4 \\ 59.43 & 394.7255 & 360.6621 & 522.078 \\ 81.82 & 360.6621 & 576.7264 & 728.31 \\ 115.4 & 522.078 & 728.31 & 1035.96 \end{bmatrix}^{-1} \begin{bmatrix} 377.5 \\ 1877.567 \\ 2246.661 \\ 3337.780 \end{bmatrix}$$

$$= \begin{bmatrix} 8.064794635 & -0.08259270531 & -0.09419511495 & -0.7905268759 \\ -0.08259270531 & 0.008479816238 & 0.001716687178 & 0.003720020321 \\ -0.09419511495 & 0.001716687178 & 0.01662942431 & -0.002063307812 \\ -0.7905268759 & 0.003720020321 & -0.002063307812 & 0.08860128617 \end{bmatrix} \begin{bmatrix} 377.5 \\ 1877.567 \\ 2246.661 \\ 3337.780 \end{bmatrix}$$

$$= \begin{bmatrix} 39.15734995 \\ 1.016100441 \\ -1.861649203 \\ -0.3432604926 \end{bmatrix}$$

maka persamaan regresinya adalah

$$\hat{y} = 39.15734995 + 1.016100441x_1 - 1.861649203x_2 - 0.3432604926x_3$$

Daftar Anova yang diperoleh:

Sumber Variasi	Dk	Jk	Rjk
Regresi pada $\beta_1, \beta_2, \beta_3 \mid \beta_0$	3	399,451	133,1504
Kekeliruan	9	38,6797	4,2977
Total	12	438,1308	

untuk menguji apakah koefisien regresi mempunyai arti atau tidak, maka hipotesisnya adalah

1). $H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

2). $H_0 : \beta_2 = 0$
$H_1 : \beta_2 \neq 0$

3). $H_0 : \beta_3 = 0$
$H_1 : \beta_3 \neq 0$

Statistik ujinya adalah

$$t = \frac{\beta_j}{\sqrt{C_{(j+1),(j+1)}}\ \sigma}$$

1). $t = \frac{1,0161}{\sqrt{0,00848}\ \ 2,0731} = 5,32$

2). $t = \frac{-1,8616}{\sqrt{0,016629}\ \ 2,0731} = -6,96$

3). $\mathrm{t} = \frac{0,3433}{\sqrt{0,088601}\ \ 2,0731} = -0,56$

sedangkan $t_{0.05} : 9 = 1,833$

Untuk 1 dan 2 H_0 ditolak, untuk 3 H_0 diterima, berarti koefisien regresi untuk x_3 tidak mempunyai arti, sehingga model regresinya menjadi:

$$\hat{y} = 39.15734995 + 1.016100441x_1 - 1.861649203x_2$$

B. Konfedensi Interval

1). Untuk rata-rata respon
suatu rentang nilai yang diperkirakan akan mencakup nilai rata-rata sebenarnya dari populasi dengan tingkat kepercayaan tertentu. Ini digunakan untuk memberikan perkiraan tentang di mana rata-rata populasi sebenarnya mungkin berada berdasarkan sampel yang diambil.

Untuk meramalkan nilai respon $\hat{y}_0$ berdasarkan nilai-nilai $x_{10}, x_{20}, \cdots, x_{k0}$ adalah interval konfidensi dari rata-rata $\mu_y / x_{10}, x_{20}, \cdots, x_{k0}$. Interval konfidensi ini dapat dibentuk dari statistik:

$$t = \frac{\hat{y}_0 - \mu_y / x_{10}, x_{20}, \cdots, x_{k0}}{\sigma \sqrt{x_0^T (X^T X)^{-1} x_0}} \tag{9}$$

yang mengikuti distribusi student t dengan $dk = n - k - 1$ maka

$$\hat{y}_0 - t_{\alpha/2} \sigma \sqrt{x_0^T (X^T X)^{-1} x_0} < \mu_y / x_{10}, x_{20}, \cdots, x_{k0} < \hat{y}_0 + t_{\alpha/2} \sigma \sqrt{x_0^T (X^T X)^{-1} x_0}$$

dengan $\frac{\alpha}{2}$ merupakan nilai dari distribusi t dengan $dk = n - k - 1$

Contoh:

Dengan menggunakan contoh sebelumnya, tentukanlah konfidensi interval 95% rata-rata respon bila $x_1 = 3, x_2 = 8, x_3 = 9$.

Penyelesaian:

persamaan regresi yang diperoleh adalah

$$\hat{y} = 39.15734995 + 1.016100441 x_1 - 1.861649203 x_2 - 0.3432604926 x_3$$

bila disubstitusikan nilai $x_1 = 3, x_2 = 8, x_3 = 9$ diperoleh $\hat{y} = 24{,}2322$

$$x_0^T(X^TX)^{-1}x_0 = \begin{bmatrix}1 & 3 & 8 & 9\end{bmatrix} \begin{bmatrix} 8.064794635 & -0.08259270531 & -0.09419511495 & -0.7905268759 \\ -0.08259270531 & 0.008479816238 & 0.001716687178 & 0.003720020321 \\ -0.09419511495 & 0.001716687178 & 0.01662942431 & -0.002063307812 \\ -0.7905268759 & 0.003720020321 & -0.002063307812 & 0.08860128617 \end{bmatrix} \begin{bmatrix}1\\3\\8\\9\end{bmatrix}$$
$$= 0{,}1267$$

$\sigma^2 = 4{,}2977$ maka $\sigma = 2{,}0731$

$t_{\frac{\alpha}{2}} = t_{0{,}0025:9} = 2{,}262$

maka konfidensi intervalnya adalah

$24{,}2322 - 2{,}262(2{,}0731)\sqrt{0{,}1267} < \mu_y/x_{10}, x_{20}, \cdots, x_{k0} < 24{,}2322 + 2{,}262(2{,}0731)\sqrt{0{,}1267}$

atau

$$22{,}5633 < \mu_y/x_{10}, x_{20}, \cdots, x_{k0} < 25{,}9011$$

Artinya dengan tingkat kepercayaan 95% dapat kita katakan bahwa nilai $\mu_y/3,8,9$ ada diantara 22,5633 hingga 25,9011

2). Untuk nilai respon

Interval kepercayaan untuk nilai respons memberikan perkiraan rentang nilai yang mungkin untuk nilai respons individu dalam populasi, berdasarkan sampel yang diambil.

konfidensi interval untuk nilai respon ini dapat dibentuk dari statistik:

$$t = \frac{\hat{y}_0 - y_0}{\sigma\sqrt{x_0^T(X^TX)^{-1}x_0}} \tag{10}$$

yang mengikuti distribusi student t dengan $dk = n - k - 1$ maka

$$\hat{y}_0 - t_{\alpha/2}\sigma\sqrt{1 + x_0^T(X^TX)^{-1}x_0} < y_0 < \hat{y}_0 + t_{\alpha/2}\sigma\sqrt{1 + x_0^T(X^TX)^{-1}x_0}$$

dengan $\frac{\alpha}{2}$ merupakan nilai dari distribusi t dengan $dk = n - k - 1$

Contoh: Dengan menggunakan contoh sebelumnya, tentukanlah konfidensi interval 95% rata-rata respon bila $x_1 = 3, x_2 = 8, x_3 = 9$.

Penyelesaian:

$\hat{y} = 24{,}2322$, $\sigma = 2{,}0731$, $x_0^T (X^T X)^{-1} x_0 = 0{,}1267$, dan $t_{\frac{\alpha}{2}} = t_{0{,}0025:9} = 2{,}262$

maka

$$24{,}2322 - 2{,}262(2{,}0731)\sqrt{1{,}1267} < \mu_y / x_{10}, x_{20}, \cdots, x_{k0} < 24{,}2322 + 2{,}262(2{,}0731)\sqrt{1{,}1267}$$

atau

$$19{,}2547 < y_0 < 29{,}2097$$

Artinya dengan tingkat kepercayaan 95%, kita memprediksi bahwa nilai y_0 dengan $x_1 = 3, x_2 = 8, x_3 = 9$ akan berada diantara 19,2547 dan 29,2097.