BAB II

PEMILIHAN VARIABEL DALAM PERSAMAAN REGRESI LINIER

II.1. REGRESI SEMUA KEMUNGKINAN

Prosedur regresi semua kemungkinan (*All Possible Regression*) membutuhkan penelaahan analisa semua persamaan. Persamaan-persamaan ini dievaluasi untuk mendapatkan persamaan regresi "terbaik". Jika diasumsikan intersep (β_0) dimasukkan dalam persamaan, dan jika ada K kandidat regresor, maka ada 2^k total persamaan yang akan diestimasi dan akan diuji. Sebagai contoh, jika K=4, maka akan ada 2^4=16 persamaan yang mungkin, sementara jika K=10, maka ada 2^{10} =1024 persamaan. Sangat jelas banyaknya persamaan yang diuji bertambah dengan cepat dengan bertambahnya kandidat regresor.

Langkah-langkah untuk memilih variabel menggunakan metode regresi semua kemungkinan adalah sebagai berikut :

1. Mengestimasi parameter β dengan menggunakan metode *Kuadrat terkecil* untuk setiap kemungkinan dari variabel bebas x yang diberikan.
2. Menghitung $\hat{y}$ dengan meggunakan estimasi parameter β yang dihasilkan pada langkah 1 untuk setiap kemungkinan variabel yang diberikan.
3. Menghitung **Jumlah Kuadrat Error** (SS_E) untuk tiap-tiap kemungkinan variabel yang diberikan, dengan persamaan :

$$SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.1}$$

dan menghitung **Rata-rata kuadrat error (MS_E)** dengan persamaan:

$$MS_E = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{dk} \qquad (2.2)$$

dengan dk adalah derajat kebebasan.

4. MS_E yang dihasilkan oleh semua kemungkinan variabel dipilih yang paling minimal, dan variabel pada MS_E yang paling minimal adalah variabel yang dipilih untuk model persamaan regresi.

Untuk memperjelas prosedur regresi semua kemungkinan, maka diberikan contoh membentuk model persamaan regresi dengan menggunakan data Hald dengan 4 variabel bebas dari 13 observasi seperti Tabel 2.1 dibawah ini:

Observasi i	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}
1	78,5	7	26	6	60
2	74,3	1	29	15	52
3	104,3	11	56	8	20
4	87,6	11	31	8	47
5	95,9	7	52	6	33
6	109,2	11	55	9	22
7	102,7	3	71	17	6
8	72,5	1	31	22	44
9	93,1	2	54	18	22
10	115,9	21	47	4	26
11	83,8	1	40	23	34
12	113,3	11	66	9	12
13	109,4	10	68	8	12

Sumber : Draper and Smith 1981

Tabel 2.1. Data hald .

Dengan metode kuadrat terkecil dihitung estimasi kuadrat terkecil yang ditunjukkan seperti pada Tabel 2.2 berikut :

Variabel dlm model	β_0	β_1	β_2	β_3	β_4
x_1	81,479	1,869			
x_2	57,424		0,789		
x_3	110,203			-1,256	
x_4	117,568				-0,738
x_1x_2	52,577	1,468	0,662		
x_1x_3	72,349	2,312		0,494	
x_1x_4	103,097	1,440			-0,614
x_2x_3	72,075		0,731	-1,008	
x_2x_4	94,160		0,311		-0,457
x_3x_4	131,282			-1,200	-0,724
$x_1x_2x_3$	48,194	1,696	0,657	0,250	
$x_1x_2x_4$	71,648	1,452	0,416		-0,237
$x_2x_3x_4$	203,642		-0,923	-1,448	-1,557
$x_1x_3x_4$	111,684	1,052		-0,410	-0,643
$x_1x_2x_3x_4$	62,405	1,551	0,510	0,102	-0,144

Tabel 2.2. Estimasi kuadrat terkecil untuk regresi semua kemungkinan.

ntuk pengujian persamaan regresi, akan dihitung Jumlah kuadrat residu (SS_E) dan rata-rata kuadrat residu (MS_E).

- **<u>SS_E dan MS_e Untuk 1 variabel</u>**

Pada **Tabel 2.2** untuk variabel dalam model x_4, didapatkan penaksir untuk y dengan persamaan :

$$\hat{y}=117,568-0,738\,x_4 \tag{2.3}$$

Dengan memasukkan nilai x_4 pada Tabel 2.1 ke persamaan (2.3) didapatkan nilai estimasi untuk y seperti pada Tabel 2.3 berikut:

y	ŷ	y-ŷ	(y-ŷ)²
78,5	73.27	5.22	27.24
74.3	79.18	-4.88	23,81
104.3	102.80	1.49	2.22
87.6	82.87	4.72	22.27
95.9	93.20	2.69	7.23
109.2	101.32	7.87	61.93
102.7	113.13	-10.43	108.78
72.5	85.08	-12.58	158.25
93.1	101.32	-8.22	67.56
115.9	98.37	17.52	306.95
83.8	92.47	-8.67	75.16
113.3	108.70	4.59	21.06
109.4	108.70	0.69	0.47
			Σ=883.86

Tabel 2.3. Estimasi y untuk variabel x_4

Dari Tabel 2.3 didapatkan $SS_E = \Sigma(y-\hat{y})^2 = 883.86$.

Dengan derajat kebebasan(dk)= n-p-1=11, maka MS_E = 883,86/11=80,35

Analog untuk mencari SS_E dan MS_E dengan menggunakan x_1,x_2,x_3.

- **SS$_E$ dan MS$_E$ Untuk 2 variabel**

Pada **Tabel 2.2** untuk variabel dalam model x_1 dan x_4, didapatkan penaksir untuk y dengan persamaan :

$$\hat{y} = 103,097 + 1,440x_1 - 0,614x_4 \qquad (2.4)$$

Dengan memasukkan nilai x_1 dan x_4 pada Tabel 2.1 ke persamaan (2.4) didapatkan nilai estimasi untuk y seperti pada Tabel 2.4 berikut:

y	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$
78,5	76,33	2,16	4,66
74,3	72,61	1,68	2,85
104,3	106,65	-2,35	5,55
87,6	90,08	-2,48	6,15
95,9	92,91	2,98	8,92
109,2	105,42	3,77	14,21
102,7	103,73	-1,03	1,06
72,5	77,52	-5,02	25,23
93,1	92,47	0,62	0,39
115,9	117,37	-1,47	2,17
83,8	83,66	0,13	0,01
113,3	111,56	1,73	2,99
109,4	110,12	-0,72	0,59
			Σ=74,76

Tabel 2.4. Estimasi y untuk variabel x_1, x_4

Dari Tabel 2.4 didapatkan $SS_E = \Sigma(y-\hat{y})^2 = 74,76$

Dengan derajat kebebasan(dk)= n-p-1=10 maka MS_E =74,76/10 =7,47

Analog untuk mencari SS_E dan MS_E dengan menggunakan x_1x_2, x_1x_3, x_2x_3, x_2x_4, x_3x_4.

- **SS_E dan MS_E Untuk 3 variabel**

Pada **Tabel 2.2** untuk variabel dalam model x_1, x_2, x_4 didapatkan penaksir untuk y dengan persamaan :

$$\hat{y} = 71,648 + 1,4519x_1 + 0,4161x_2 - 0,2385x_4 \quad (2.5)$$

Dengan memasukkan nilai x_1, x_2 dan x_4 pada Tabel 2.1 ke persamaan (2.5) didapatkan nilai estimasi untuk y seperti pada tabel berikut:

y	ŷ	y − ŷ	$(y-\hat{y})^2$
78,5	78,43	0,06	0,00
74,3	72,86	1,43	2,05
104,3	106,19	-1,89	3,57
87,6	89,40	-1,80	3,24
95,9	95,64	0,25	0,06
109,2	105,30	3,89	15,19
102,7	104,12	-1,42	2,04
72,5	75,59	-3,09	9,55
93,1	91,81	1,28	1,64
115,9	115,54	0,35	0,12
83,8	81,70	2,09	4,40
113,3	112,24	1,05	1,11
109,4	111,62	-2,22	4,94
			Σ=47,97

Tabel 2.5. Estimasi y untuk variabel x_1, x_2, x_4

Dari Tabel 2.5 didapatkan $SS_E = \Sigma(y-\hat{y})^2 = 47,97$

Dengan derajat kebebasan(dk)= n-p-1=9 maka $MS_E = 47,97 / 9 = 5,33$

Analog untuk mencari SS_E dan MS_E dengan menggunakan $x_1x_2x_3$, $x_1x_3x_4$, $x_2x_3x_4$.

- **<u>SS_E dan MS_E Untuk 4 variabel</u>**

Pada **Tabel 2.1** untuk variabel dalam model x_1, x_2, x_3 dan x_4 didapatkan penaksir untuk y dengan persamaan :

$$\hat{y} = 62,405 + 1,551x_1 + 0,510x_2 - 0,102x_4 - 0,144x_4 \quad (2.6)$$

Dengan memasukkan nilai x_1, x_2, x_3 dan x_4 pada persamaan (2.6) didapatkan nilai estimasi untuk y seperti pada tabel berikut:

y	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$
78,5	78,49	0,00	0,00
74,3	72,78	1,51	2,28
104,3	105,97	-1,67	2,79
87,6	89,32	-1,72	2,98
95,9	95,64	0,25	0,06
109,2	105,27	3,92	15,40
102,7	104,14	-1,44	2,09
72,5	75,67	-3,17	10,07
93,1	91,72	1,37	1,89
115,9	115,61	0,28	0,07
83,8	81,80	1,99	3,96
113,3	112,32	0,97	0,94
109,4	111,69	-2,29	5,26
			Σ=47,86

Tabel 2.6. Estimasi y untuk variabel x_1, x_2, x_3 x_4

Dari Tabel 2.6 didapatkan $SS_E = \Sigma(y-\hat{y})^2 = 47{,}86$

Dengan derajat kebebasan(dk)= n-p-1=8 maka MS_E = 47,86 / 8 =5,98

Hasil keseluruhan perhitungan SS_E dan MS_E adalah sebagai berikut:

Regresor dlm model	**$SS_E(p)$**	**$MS_E(p)$**
x_1	1265,68	115,06
x_2	906,33	82,39
x_3	1939,40	176,30
x_4	883,86	80,35
x_1x_2	57,90	5,79
x_1x_3	1227,07	122,70
x_1x_4	74,76	7,47
x_2x_3	415,44	41,54

Tabel 2.7. Tabel SS_E dan MS_E

Regresor dlm model	$SS_E(p)$	$MS_E(p)$
x_2x_4	868,88	86,88
x_3x_4	175,73	17,57
$x_1x_2x_3$	48,11	5,34
$x_1x_2x_4$	47,97	5,33
$x_2x_3x_4$	50,83	5,64
$x_1x_3x_4$	73,81	8,20
$x_1x_2x_3x_4$	47,86	5,98

Tabel 2.7. Tabel SS_E dan MS_E *(lanjutan)*

MS_E minimal pada persamaan yang dibentuk variabel x_1, x_2 dan x_4. Jadi persamaan :

$$\hat{y} = 71,648 + 1,4519x_1 + 0,4161x_2 - 0,2385x_4$$

sebagai persaman regresi yang terpilih.

II.2. PEMILIHAN KE DEPAN

Untuk meminimalkan kesukaran dalam perhitugan, metode diatas dikembangkan untuk mengevaluasi hanya menggunakan **sedikit** bilangan parameter dalam model regresi. Salah satu metode tersebut adalah **Seleksi Kedepan (Forward Selection).**

Prosedur Seleksi Kedepan dimulai dengan mengasumsikan di dalam model tidak ada regresor intercept(β_0). Langkah metode pemilihan ke depan adalah sebagai berikut

1. Regresor pertama dipilih untuk dimasukkan kedalam persamaan yang mempunyai ***korelasi sederhana*** (r_{iy}) terbesar dengan variabel respon y. Korelasi sederhana ini dihitung dengan rumus :

$$r_{iy} = \frac{n\sum x_i y - (\sum x_i)(\sum y)}{\sqrt{\{n\sum x_i^2 - (\sum x_i)^2\}\{n\sum y^2 - (\sum y)^2\}}} \tag{2.7}$$

Misalkan regresor pertama yang terpilih adalah $\mathbf{x_1}$, maka regresor ini harus mempunyai harga F statistik yang lebih besar dari F tabel

2. Regresor kedua dipilih untuk dimasukkan kedalam persamaan, yaitu yang mempunyai ***korelasi parsial*** tertinggi dengan y dan mempunyai $F_{stat} > F_{tabel}$. Untuk mencari korelasi parsial adalah sebagai berikut : Setelah variabel pertama terpilih misalnya $\mathbf{x_1}$, maka model persamaan regresinya adalah $y = \beta_0 + \beta_1 x_1 + \varepsilon$.

   Selanjutnya dibuat variabel-variabel baru $x_1^*, x_2^*, \ldots, x_p^*$, dimana x_i^* (i=1,2,...,p) merupakan sisaan x_i setelah diregresikan terhadap $\mathbf{x_1}$. Peubah tidak bebasnya Y^* juga merupakan sisaan Y setelah diregresikan terhadap $\mathbf{x_1}$. Selanjutnya dihitung korelasi sederhana antar variabel baru tersebut, yaitu korelasi sederhana antara sisaan dari regresi $\hat{Y} = f(x_1)$ dengan sisaan dari regresi $\hat{x}_j = f_j(x_1)$ yaitu : $\hat{x}_j = \hat{\alpha}_{0j} + \hat{\alpha}_{1j} x_1$, j=2,3,....,p yang disebut **korelasi parsial.** Korelasi parsial dilambangkan dengan, misalnya $r_{2y.1}$, yaitu korelasi antara x_2^* dan Y^* , dan dibaca 'korelasi parsial $\mathbf{x_2}$ dengan Y setelah keduanya dikoreksi untuk peubah $\mathbf{x_1}$ '.

3. Langkah kedua diulang sampai semua variabel terproses.

Berikut contoh untuk metode Seleksi Kedepan dengan menggunakan data yang sama pada metode regresi semua kemungkinana seperti pada **Tabel 2.1.** Pada langkah pertama dihitung koefisien korelasi sederhana antara y dengan x_1, x_2, x_3, x_4 dengan **persamaan (2.7)** :

$$r_{ij} = \frac{n\sum x_i y - (\sum x_i)(\sum y)}{\sqrt{\{n\sum x_i^2 - (\sum x_i)^2\}\{n\sum y^2 - (\sum y)^2\}}}$$

didapatkan tabel korelasi sederhana :

	x_1	x_2	x_3	x_4	y
y	0,73	0,81	-0,53	-0,82	1

Diambil korelasi sederhana yang tertinggi dengan y, yaitu pada x_4 (r_{4y}=-0,82) untuk dimasukkan ke dalam persamaan regresi, sehingga didapatkan persamaan regresi :

$$\hat{y}=117{,}568-0{,}738\,x_4$$

UJI F

Dari **Tabel 2.3** didapatkan $\Sigma(y-\hat{y})^2 = 883.86$ dengan derajat kebebasan(dk) = n-p-1=11 maka :

$MS_E(\beta_0,\beta_1) = 883{,}86/11$

$SS_E(\beta_1 | \beta_0) = \beta_1(\Sigma x_4 y - (\Sigma x_4 \Sigma y)/n)$

$= -0{,}738\,(\ 34733{,}3 - (390\ .\ 1240{,}5)/13\ )$

$= 1831{,}89$

$$Fstat = \frac{SS_E(\beta_1|\beta_0)}{MS_E(\beta_0,\beta_1)} = \frac{1831{,}89}{80{,}35} = 22{,}79$$

Dari *lampiran 7* nilai $F_{tabel} = F_{.1,1,11} = 3{,}23$. Karena $F_{stat} > F_{tabel}$ maka x_4 dimasukkan dalam persamaan regresi.

Langkah kedua adalah menghitung koefisien korelasi parsial antara x_1, x_2, x_3 yaitu korelasi sederhana antara residu dari $y = \beta_0 + \beta_1 x_4$ dengan $\hat{x}_j = \alpha_{0j} + \alpha_{1j} x_4$, $j \neq 4$. Dengan menggunakan persamaan (2.7) didapatkan :

Variabel	kuadrat korelasi parsial
x_1	0,91
x_2	0,01
x_3	0,80
y	1

Terlihat korelasi yang tertinggi adalah x_1 , maka x_1 dicobakan kedalam persamaan regresi bersama x_4 : $\hat{y} = 103{,}09 + 1{,}4399x_1 - 0{,}6139x_4$

UJI F

Dari Tabel 2.3 dan Tabel 2.4 dapat dihitung

$$SS_R\ (x_1|x_4) = \sum (y - \hat{y}_{f(x_4)})^2 - \sum (y - \hat{y}_{f(x_4, x_1)})^2$$

$$= 883{,}86 - 74{,}76$$

$$= 809{,}10$$

Dari Tabel 2.4 didapatkan $MS_E = \dfrac{\sum (y - \hat{y}_{f(x1,x4)})^2}{n-p-1} = \dfrac{74{,}76}{10} = 7{,}47$

Dari *lampiran 7*, $F_{tabel} = F_{.1,1,10} = 3{,}29$ dan

$$Fstat = \frac{SS_R(x_1|x_4)}{MS_e(x_1, x_4)} = \frac{809{,}10}{7{,}47} = 108{,}22$$

karena $F_{stat} > F_{tabel}$ maka x_1 dimasukkan dalam persamaan regresi.

Pada langkah ketiga dihitung koefisien korelasi parsial $r_{x,14}$ yaitu korelasi sederhana antara residu dari $y = \beta_0 + \beta_1 x_1 + \beta_2 x_4$ dengan residu dari $\hat{x}_i = \hat{\alpha}_{0i} + \hat{\alpha}_{1i} x_1 + \hat{\alpha}_{2i} x_4$ didapatkan :

Variabel	kuadrat korelasi parsia
x_2	0,35
x_3	0,32
y	1

x_2 mempunyai korelasi terbesar, maka dicobakan kedalam persamaan regresi :

$$\hat{y} = 71{,}6482 + 1{,}4519 x_1 + 0{,}4161 x_2 - 0{,}2385 x_4$$

UJI F:

Dari Tabel 2.4 dan Tabel 2.5 dihitung

$$SS_R\,(x_2|x_1, x_4) = \sum (y - \hat{y}_{f(x_1, x_4)})^2 - \sum (y - \hat{y}_{f(x_1, x_2, x_4)})^2$$

$$= 74{,}76 - 47{,}97$$

$$= 26{,}79$$

$$MS_e(x_1, x_4) = \frac{\sum (y - \hat{y}_{f(x_1, x_2, x_4)})^2}{n - p - 1} = \frac{47{,}97}{9} = 5{,}33$$

Dari *lampiran 7*, $F_{tabel} = F_{.1,1,9} = 3{,}36$ dan

$$Fstat = \frac{SS_R(x_1|x_4)}{MS_e(x_1, x_4)} = \frac{26{,}79}{5{,}33} = 5{,}02$$

Karena $F_{stat} > F_{tabel}$ maka x_2 dimasukkan ke dalam persamaan regresi.

Sekarang prediktor yang masih tersedia adalah x_3. Misalkan x_3. dimasukkan kedalam persamaan regresi, maka persamaamya adalah :

$$\hat{y} = 62{,}405 + 1{,}551 x_1 + 0{,}510 x_2 - 0{,}102 x_4 - 0{,}144 x_4$$

<u>*Uji F :*</u>

Dari Tabel 2.5 dan Tabel 2.6 dihitung :

$$SS_R(x_3|x_1,x_2,x_4) = \sum(y - \hat{y}_{f(x_1,x_2,x_4)})^2 - \sum(y - \hat{y}_{f(x_1,x_2,x_3,x_4)})^2$$

$$= 47{,}97 - 47{,}86$$

$$= 0{,}10$$

$$MS_e(x_1,x_2,x_3,x_4) = \frac{\sum(y - \hat{y}_{f(x_1,x_2,x_3,x_4)})^2}{n-p-1}$$

$$= 5{,}98$$

Dari *lampiran 7*, $F_{tabel} = F_{.1,1,8} = 3{,}46$ dan

$$Fstat = \frac{SS_R\ (x_3|x_1,x_2,x_4)}{MS_e(x_1,x_2,x_3,x_4)} = \frac{0{,}10}{5{,}98} = 0{,}01$$

Karena $F_{stat} < F_{tabel}$ maka x_3 tidak dimasukkan ke dalam persamaan regresi. Hasil akhir seleksi kedepan adalah :

$$\hat{y} = 71{,}6482 + 1{,}4519x_1 + 0{,}4161x_2 - 0{,}2385x_4$$

Dari hasil akhir metode seleksi kedepan dapat disimpulkan bahwa metode pemilihan kedepan tidak menghitung semua kemungkinan dari variabel yang diberikan seperti pada metode regresi semua kemungkinan regresi, maka metode pemilihan kedepan lebih sederhana dan lebih singkat dibandingkan dengan metode semua kemungkinan. □