A. Sukmana

Koefisien Determinasi R^2 pada Model Regresi Linear

Intisari

Tulisan ini membahas sifat-sifat dari koefisien determinasi sampel R^2 *pada model regresi linear , diantaranya pada kondisi apa nilai* R^2 *(yang biasa muncul di output program komputer yang membahas analisa regresi) memiliki makna terutama untuk menilai kecocokan model , serta bagaimana membangun selang kepercayaan untuk koefisien determinasi populasi .*

I. Pendahuluan

Pada umumnya setiap perangkat lunak aplikasi statistika yang membahas model regresi memuat koefisien determinasi sampel R^2 pada keluarannya (output). Seringkali pemakai langsung menggunakan koefisien determinasi tersebut untuk melihat apakah model regresi tersebut sudah cukup *baik* atau *tidak* . Nilai R^2 yang menuju 1 biasanya diinterpretasikan bahwa model regresi sudah '*baik* ' sedangkan jika nilai R^2 menuju 0 diinterpretasikan model regresi '*tidak baik*'. Padahal pada kenyataannya tidak sesederhana itu untuk menyimpulkan suatu model '*baik / tidak*', banyak faktor yang harus diperhatikan agar nilai R^2 memiliki arti .

Pada tulisan ini akan kita bahas koefisien determinasi R^2 , pada kondisi apa nilai koefisien R^2 memiliki arti sehingga dapat kita gunakan untuk menilai kecocokan model , seberapa baik R^2 menghampiri nilai koefisien determinasi populasi serta terakhir membahas bagaimana membangun selang kepercayaan untuk koefisien determinasi populasi ρ^2 .

II. Koefisien Determinasi R^2

Pada model regresi dikenal apa yang disebut *variabel tak bebas* atau *variabel respon* dan beberapa *variabel bebas* atau *faktor* . Faktor bisa ditentukan oleh peneliti (*fixed model*) atau acak (*random model)* yang memiliki distribusi tertentu .

Model regresi linear dapat ditulis :

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i$$

$X_1,\ldots,X_p$ disebut faktor dan $\beta_0,\ldots,\beta_p$ disebut koefisien regresi . ε menyatakan semua hal yang belum diterangkan oleh faktor $X_1,\ldots,X_p$, bersifat acak dan berdistribusi normal dengan mean 0 dan variansi σ^2 .

Untuk memudahkan pembahasan model regresi linear tersebut kita tulis dalam bentuk matriks $Y = X\beta + e$ yang terdiri dari p buah faktor .

Y adalah matriks ukuran nx1 yang menyatakan vektor respon

X adalah matriks ukuran nx(p+1) yang menyatakan matriks faktor

β adalah matriks ukuran (p+1)x1 yang menyatakan koefisien regresi

e adalah matriks *galat* (error) ukuran nx1 yang memiliki E(e)=0 dan Var(e)=s^2

Kita gunakan metoda kuadrat terkecil untuk memperoleh taksiran dari koefisien regresi . Dari persamaan normal yang diperoleh koefisien regresi :

$$\hat{\beta}=(X'X)^{-1}X'y$$

Jumlah Kuadrat Regresi (*JKR*) didefinisikan sebagai berikut :

$$JKR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=(\hat{Y}-1\bar{y})'(\hat{Y}-1\bar{y})$$

dan Jumlah Kuadrat Total (*JKT*) didefinisikan sebagai berikut :

$$JKT=\sum_{i=1}^{n}(y_i-\bar{y})^2=(Y-1\bar{y})'(Y-1\bar{y})$$

Koefisien determinasi sampel dari model regresi tersebut dituliskan sebagai berikut :

$$R^2=\frac{JKR}{JKT}$$

yang menyatakan proporsi dari total variasi yang dapat diterangkan oleh model regresi tersebut . Makin besar nilai *JKR* maka semakin banyak proporsi variasi yang dapat diterangkan oleh model, pertanyaannya apakah jika $\mathbf{R^2}$ mendekati nilai 1 maka model regresi sudah 'baik ' ? .

Untuk menjawab pertanyan tersebut , ada dua hal yang harus kita periksa. Yang pertama adalah perilaku asimtotik dari $\mathbf{R^2}$ dan yang kedua apakah $\mathbf{R^2}$ merupakan taksiran terbaik untuk koefisien determinasi populasi ρ^2 .(ingat ! : *penaksir terbaik* adalah penaksir takbias (unbiased) yang variansinya minimum [Hogg,1978]).

Pertama kita bahas perilaku asimtotik dari $\mathbf{R^2}$. Jika ukuran sampel dibuat besar ($n\to\infty$) maka $\mathbf{R^2}$ dapat dihampiri sebagai berikut :

$$R^2 \sim \frac{\beta' S_X \beta}{\beta' S_X \beta+\sigma^2}$$

dimana S_X menyatakan matriks kovariansi dari sampel. Perhatikan nilai $\mathbf{R^2}$ akan membesar jika $\beta' S_X \beta$ sangat besar dibandingkan dengan σ^2 , ditulis dalam notasi penjumlahan : $\sum_{j=1}^{p}\beta_j{}^2 s_{x_j}^2$. Hal ini disebabkan oleh nilai $s_{x_j}^2$ yang besar atau dengan kata lain setiap faktor x sangat berbeda sehingga variansinya membesar. Dengan kata lain nilai $\mathbf{R^2}$ yang mendekati 1 tidaklah berarti bahwa model regresi sudah baik.

Koefisien determinasi dari populasi didefinisikan sebagai berikut:

$$\rho^2=\frac{Var(\beta_1x_1+\beta_2x_2+...+\beta_px_p)}{Var(y)}$$

atau

$$\rho^2=\frac{\beta'\Sigma_X\beta}{\beta'\Sigma_X\beta+\sigma^2}$$

dimana Σ_X menyatakan matriks kovariansi dari **X** acak .

Jika matriks **X** adalah matriks acak , dimana masing-masing baris yang merupakan hasil pengamatan acak berdistribusi normal multivariat dengan mean μ_X , kovariansi

Σ_X dan saling bebas dengan vektor galat **e** , maka $E(R^2)=\rho^2$. Dengan kata lain jika **X** adalah matriks acak maka $\mathbf{R^2}$ merupakan *penaksir takbias* dari koefisien determinasi populasi sehingga $\mathbf{R^2}$ dapat digunakan untuk memeriksa kecocokan model .

Jika **X** bukan matriks acak maka $E(R^2)=\frac{p}{n-1}$ pada $\rho^2=0$. Jadi $\mathbf{R^2}$ merupakan *penaksir bias* dari koefisien determinasi $\rho^2=0$ dengan bias positif sebesar $\frac{p}{n-1}>0$ sehingga $\mathbf{R^2}$ tidak dapat digunakan untuk menilai kecocokan model .

Bias dari taksiran koefisien determinasi $\rho^2=0$ tersebut dapat dihilangkan dengan cara melakukan penyesuaian :

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

sehingga $E(R^2_{adj})=0$ atau R^2_{adj} merupakan taksiran takbias untuk koefisien determinasi $\rho^2=0$.

Meskipun R^2_{adj} adalah taksiran takbias , tetapi hanya berlaku jika koefisien determinasi dari populasi sama dengan nol .

Dari sini kita lihat bahwa nilai $\mathbf{R^2}$ tidak banyak bermakna untuk menilai kecocokan model apabila matriks rancangannya (**X**) bukan matriks acak .

III. Selang Kepercayaan untuk Koefisien Determinasi

Biasanya kita tidak puas jika hanya berhenti pada taksiran titik sehingga kita perlu membahas taksiran selang (selang kepercayaan) untuk koefisien determinasi yang memberikan hasil yang lebih baik. Sebelum membentuk selang kepercayaan, kita perlu mengetahui distribusi dari $\mathbf{R^2}$.

Perhatikan bentuk berikut :

$$\frac{R^2}{1-R^2}=\frac{\frac{JKR}{JKT}}{\frac{JKG}{JKT}}=\frac{JKR}{JKG}$$

Jumlah kuadrat regresi (JKR) dan jumlah kuadrat galat(JKG) saling bebas , dan masing-masing memiliki distribusi khi-kuadrat .

JKR memiliki distribusi khi-kuadrat nonsentral dengan parameter λ dan derajat kebebasan p , sedangkan JKG memiliki distribusi khi-kuadrat dengan derajat kebebasan n-p-1 . Nisbah dari JKR dan JKT memiliki distribusi **F** .

Jika **X** bukan matriks acak maka parameter khi-kuadrat nonsentral adalah :

$\lambda=\frac{\beta' X' X\beta}{\sigma^2}$ sedangkan jika **X** matriks acak dimana baris dari **X** berdistribusi normal multivariat dengan mean μ_X dan matriks kovariansi Σ_X. maka parameter khi-kuadrat nonsentral adalah

$$\lambda=\frac{\beta'\Sigma_X\beta}{\sigma^2}\chi^2_{n-1}=\frac{\rho^2}{1-\rho^2}\chi^2_{n-1}$$

Menghitung nilai eksak dari distribusi **F** diatas tidaklah mudah , *Gurland(1968)*

mengemukakan metoda penghampiran untuk distribusi $\mathbf{R^2}$.

Distribusi $\chi_p^2(\lambda)$ didekati oleh $a\chi_\nu^2$ dimana **a** dan ν dipilih sedemikian sehingga dua momen pertama memberikan hasil yang sama dengan distribusi semula . Menurut *Gurland(1968)* nilai **a** dan ν yang memenuhi kriteria diatas adalah :

$$a = \frac{(n-1)k(k+2)+p}{(n-1)k+p}$$

dan

$$\nu = \frac{(n-1)k+p}{a}$$

dengan

$$k = \frac{\rho^2}{1-\rho^2}$$

Dari hasil hampiran tersebut diperoleh bahwa nisbah JKR/JKG memiliki distribusi $F_{\nu,n-p-1}$.

Dengan memanfaatkan hubungan berikut :

$$F_{\nu,n-p-1} \approx \frac{(n-p-1)(1-\rho^2)}{p+(n-p-1)\rho^2}\frac{R^2}{1-R^2}$$

kita dapat menentukan selang kepercayaan 100(1-α)% untuk ρ^2 yaitu :

$$[R^2_{\frac{\alpha}{2}}, R^2_{1-\frac{\alpha}{2}}]$$

dengan

$$R^2_\alpha = \frac{(n-p-1)R^2-(1-R^2)pF_\alpha}{(n-p-1)[R^2+(1-R^2)F_\alpha]}$$

F_α diperoleh dari tabel F dengan derajat kebebasan ν dan n-p-1 .

Masalah yang muncul kemudian adalah menentukan derajat kebebasan ν , karena nilai $\nu = \frac{[(n-p-1)\rho^2+p]^2}{n-1-(n-p-1)(1-\rho^2)^2}$ bergantung pada parameter yang tidak diketahui ρ^2.

Penyelesaian secara *heuristik* berikut dapat dilakukan untuk memperoleh nilai derajat kebebasan ν:

Ambil $\rho^2 = R^2$, untuk mendapatkan derajat kebebasan ν yang menentukan batas R^2_α . Ulangi hal yang sama dengan mengambil $\rho^2 = R^2_\alpha$ lakukan *iterasi* sampai konvergen ke suatu nilai .

Tabel berikut [Helland] membandingkan antara nilai hampiran hasil algoritma

α=5%	p=3	n=13	α=5%	p=40	n=141
R	R_α	R_α hampiran	R	R_α	R_α hampiran
0.7566	0.1	0.0997	0.6209	0.1	0.0999
0.7744	0.2	0.2006	0.6399	0.2	0.1999
0.7998	0.2	0.3009	0.6692	0.3	0.2999
0.8291	0.4	0.4008	0.7063	0.4	0.3998
0.8598	0.5	0.5007	0.7490	0.5	0.4997
0.8903	0.6	0.6004	0.7957	0.6	0.5998
0.9199	0.7	0.7003	0.8450	0.7	0.6999
0.9481	0.8	0.8000	0.8959	0.8	0.8
0.9748	0.9	0.8999	0.9477	0.9	0.9

heuristik diatas dengan nilai eksaknya .

Jika kita bandingkan antara nilai eksak dari R_α dengan hampirannya, dapat kita lihat cara penghampiran tersebut sudah baik dan dapat kita gunakan .

IV. Penutup

Dari hasil pembahasan diatas diperoleh kesimpulan bahwa koefisien determinasi sampel $\mathbf{R^2}$ tidak banyak berguna untuk

menilai kecocokan model apabila sampel yang diambil tidak acak , sebagai alternatif dapat digunakan taksiran selang yang memberikan hasil yang lebih baik walau tidak mudah menginterpretasikannya .

Pustaka

1. Helland,I.S. *On the Interpretation and Use of* R^2 *in Regression Analysis* , (Research Report) Department of Mathematics and Statistics Agricultural University of Norway
2. Hogg.R. *Introduction to Mathematical Statistics* , Collier Mac-millan, 1977.
3. Jobson. *Applied Multivariate Data Analysis Vol.I : Regression and Experimental Design,* Springer-Verlag, 1991.
4. Sen,A. *Regression Analysis : Theory , Methods and Aplication* , Springer-Verlag, 1990 .
5. Weisberg,S. , *Applied Linear Regression* , Wiley, 1980.

Penulis

A. Sukmana adalah dosen Jurusan Matematika Universitas Katolik Parahyangan.