BAB II

REGRESI DAN KORELASI

2.1. Pengertian Umum

Analisis regresi merupakan studi mengenai hubungan fungsional suatu variabel yang diterangkan (variabel bebas) ,dengan satu atau lebih variabel yang menerangkan (variabel bebas) sedangkan Analisis korelasi merupakan studi yang berfungsi untuk mengukur kuatnya hubungan (derajat hubungan) antara variabel-variabel tersebut.

Pada pembicaraan bab ini dibatasi hanya hubungan dua variabel pengamatan.

Dua variabel tersebut diasumsikan berasal dari populasi berdistribusi normal, sedangkan untuk Analisis regresinya hanya dibawakan hubungan yang linier misalnya antara variabel X dan Y.

Bentuk persamaan regresi adalah :

$$Y = \alpha + \beta X + \varepsilon$$

Model regresi sederhana merupakan hal yang khusus dari model regresi berganda (Multiple regresion), dimana kita dapat menyatakan hubungan fungsional antara variabel Y dan variabel-variabel X, jika kita mempunyai P variabel bebas, yaitu x_1, x_2, ..., x_p dan satu variabel tak bebas Y (P variabel prediktor dan satu variabel respon), maka himpunan data - data tersebut adalah (X_{i1}, X_{i2}, ...,X_{ip} ; Y_i) untuk i = 1,2 ... n dan dapat dituliskan secara vektor ($\underset{\sim}{X}_1$, $\underset{\sim}{X}_2$,...,$\underset{\sim}{X}_p$, $\underset{\sim}{Y}$).

vektor ($\underset{\sim}{X}_1$, $\underset{\sim}{X}_2$,...,$\underset{\sim}{X}_p$, $\underset{\sim}{Y}$).

Hubungan fungsional diatas merupakan permukaan respon (respon surface) atau Hyper plane).

Hubungan Y dan X_1, X_2, ... , X_p dapat dinyatakan :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

Koefisien-koefisien β akan diestemasi berdasarkan data hasil penggunaan sampel random sedangkan prosedur estemasi tergantung asumsi-asumsi mengenai variabel X dan sesatan random ε.

Beberapa asumsi yang dipakai adalah :

1. Nilai harapan (Expected value) dari setiap sesatan random sama dengan nol atau E $(\varepsilon_i) = 0$ dan variansinya = σ^2 atau $\nu(\varepsilon_i) = \sigma^2$
2. Harga sesatan random (ε_i) tidak berkorelasi terhadap sesatan random lainnya (ε_j), jadi :
   $E(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$
3. ε_i mengikuti distribusi normal dengan mean nol dan variasi σ^2 atau $\varepsilon_i \sim NID(0, \sigma^2)$.
4. Variabel-variabel X_{i1} ,X_{i1}, ... , X_{ip} merupakan bilangan real tanpa mengandung kesalahan (bernilai fixed).

2.2. Beberapa Ukuran-ukuran Mengenai Data

Untuk mendapatkan gambaran yang jelas tentang data mengenai suatu persoalan selain daripada data itu

disajikan dalam tabel dan diagram, masih diperlukan ukuran-ukuran yang merupakan wakil kumpulan data tersebut.

Dalam pembahasan selanjutnya akan digunakan ukuran-ukuran antara lain :

Mean sampel, $\overline{X}$ didefinisikan apabila diberikan data pengamalan sampel , x_1 ,x_2 , ... , x_n dari variabel random,maka mean sampel, dinotasikan $\overline{X}$ adalah :

$$\overline{X} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Sedangkan variansi sampel dinotasikan S^2 didefinisikan sebagai :

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Untuk simpangan baku atau standar deviasi S adalah akar positif dari S^2.

Didalam Analisis regresi maupun korelasi sering digunakan metoda (transformasi) yang akan menghemat waktu dan memudahkan dalam penghitungan.

Metoda transformasi perhitungan itu antara lain :

$z_i = x_i - \overline{X}$ yaitu selisih harga variabel X dengan meannya. Selanjutnya apabila harga variabel X mempunyai simpangan baku (standar deviasi) S, maka harga variabel standar T (harga variabel X yang telah distandarisasi) dinyatakan dengan :

$$t_i = \frac{z_i}{S}$$

Disamping menghemat waktu dan memudahkan perhitungan, dengan metoda transformasi akan diperoleh kelebihan-kelebihan, yaitu :

1. jika diberikan data pengamatan yang telah ditransformasikan ke bentuk variabel Z dan variabel T maka jumlah data variabel Z = 0 sedangkan variansi variabel T = 1.

   karena :

$$\sum_{h=1}^{n} Z_h = \sum_{h=1}^{n} (X_h - \bar{X}) = \sum_{h=1}^{n} X_h - \frac{n}{n} \sum_{h=1}^{n} X_h = 0$$

   maka

$$\bar{T} = \sum_{h=1}^{n} \frac{T_h}{n} = \sum_{h=1}^{n} \frac{Z_h}{nS} = \frac{1}{nS} \sum_{h=1}^{n} Z_h = \frac{1}{nS} \cdot 0 = 0$$

   sehingga variansi T adalah :

$$S_T^2 = \sum_{h=1}^{n} \frac{T_h^2}{n} = \frac{1}{n} \sum_{h=1}^{n} \left(\frac{Z_h}{S} \right)^2 = \frac{1}{nS^2} \sum_{h=1}^{n} (X_h - \bar{X})^2$$

$$= \frac{1}{nS^2} \, n \, S^2 = 1$$

2. Apabila diberikan harga variabel Xi dan Xj, yaitu $x_{1i}, x_{2i}, \ldots, x_{ni}$ dan $x_{1j}, x_{2j}, \ldots, x_{nj}$. makacovariansi Xi dan Xj ditulis Cov [Xi, Xj] adalah :

$$\text{Cov } [X_i, X_j] = \sum_{h=1}^{n} \frac{(x_{hi} - \bar{X}_i)(x_{hj} - \bar{X}_j)}{n} = \sum_{h=1}^{n} \frac{z_{hi} \, z_{hj}}{n}$$

3. Apabila harga variabel Xi dan Xj ditranformasikan

berturut - turut menjadi variabel T_i dan T_j , yaitu t_{i1}, t_{i2}, ... , t_{in} dan t_{j1}, t_{j2}, ... ,t_{jn} maka koefisien korelasi antara variabel X_i dengan variabel X_j ditulis r_{ij} adalah :

$$r_{ij} = \frac{Cov\,[X_i, X_j]}{S_i\ S_j} = \sum_{h=1}^{n} \frac{z_{hi}\ z_{hj}}{n\ S_i\ S_j} = \sum_{h=1}^{n} \frac{t_{hi}\ t_{hj}}{n}$$

2.3. Regresi Sederhana

Misalkan kita mengambil variabel bebas (prediktor) X_1 yang fixed (ditetapkan tidak mempunyai error) dan variabel respon X_2.

Kemudian harga Variabel X_1 ditransformasikan ke bentuk variabel Z_1 , maka menurut sifat yang telah dijelaskan pada bagian akhir sub bab 2.2. jumlah harga variabel Z_1 sama dengan nol, secara matrik ditulis $\underset{\sim}{U}'\underset{\sim}{Z}_1 = 0$, dimana $\underset{\sim}{Z}_1$ adalah vektor kolom dari n harga dari yariabel prediktor X_1 dan $\underset{\sim}{U}'$adalah vektor baris dengan elemen-elemennya sama dengan satu atau vektor jumlah (Sum vektor).

Garis regresi mempunyai bentuk umum :$\hat{X}_2 = a + b\ Z_1$. Persamaan ini kita ubah ke bentuk matrik yaitu diberikan vektor $\underset{\sim}{Z}_1$ berukuran n x 1, kemudian ditentukan $\underset{\sim}{\hat{X}}_2$ sedemikian sehingga masing-masing pasangan elemen yang bersesuaian memenuhi persamaan linier sebagai berikut :

$$\hat{\underset{\sim}{X}}_2 = a\,\underset{\sim}{U} + b\,\underset{\sim}{Z}_1 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.1.)$$

dimana a dan b tetapan sekalar sedangkan $\underset{\sim}{U}$ adalah vektor kolom dengan elemen-elemennya sama dengan satu.

Vektor $\underset{\sim}{Z}_1$ dan $\hat{\underset{\sim}{X}}_2$ menyatakan n titik pada bidang (X_1, X_2) yang terletak pada garis regresi, sebaliknya $\underset{\sim}{Z}_1$ dan $\underset{\sim}{X}_2$ ($\underset{\sim}{X}_2$ vektor berukuran n x 1) menentukan n titik yang menggambarkan data-data pengamatan. Selisih harga-harga X_2 dengan garis regresi dinyatakan dengan vektor $(\underset{\sim}{X}_2 - \hat{\underset{\sim}{X}}_2)$. atau juga bisa dikatakan merupakan jarak titik-titik (harga-harga) pengamatan ke garis regresi yang diukur sepanjang sumbu X_2.

Yang menjadi dasar permasalahan adalah menentukan nilai a dan b, yang mana berfungsi untuk menentukan, dimana garis lurus (garis regresi) digambarkan dalam kumpulan titik. Selanjutnya kita tentukan standar dari garis kecocokan terbaik yang dapat dipilih. Standar klasik merupakan standar kwadrat terkecil, yang tak lain adalah jumlah kwadrat deviasi yang harus diminimumkan. Notasi jumlah ini dengan " f ", kemudian dengan adanya deviasi yang dikumpulkan dalam vektor $(\underset{\sim}{X}_2 - \hat{\underset{\sim}{X}}_2)$ didapat

$$f = (\underset{\sim}{X}_2 - \hat{\underset{\sim}{X}}_2)'(\underset{\sim}{X}_2 - \hat{\underset{\sim}{X}}_2)$$

$$= (\underset{\sim}{X}_2 - b\,\underset{\sim}{Z}_1 - a\,\underset{\sim}{U})'(\underset{\sim}{X}_2 - b\,\underset{\sim}{Z}_1 - a\,\underset{\sim}{U}) \quad \ldots (2.2)$$

dimana digunakan persamaan (2.1) untuk mensubstitusi $\hat{\underset{\sim}{X}}_2$, sehingga didapat f sebagai berikut :

$$f = \underset{\sim}{X}_2'\,\underset{\sim}{X}_2 - b\,\underset{\sim}{X}_2'\,Z_1 - a\,\underset{\sim}{X}_2'\,\underset{\sim}{U} - b\,\underset{\sim}{Z}_1'\,\underset{\sim}{X}_2 + b^2\underset{\sim}{Z}_1'\,\underset{\sim}{Z}_1 +$$

$$a\ b\ \underset{\sim}{Z}_2'\ \underset{\sim}{U} - a\ \underset{\sim}{U}\ \underset{\sim}{X}_2' + a\ b\ \underset{\sim}{U}'\ \underset{\sim}{Z}_1 + a^2\ \underset{\sim}{U}'\ \underset{\sim}{U}$$

$$= \underset{\sim}{X}_2'\ \underset{\sim}{X}_2 + b^2\ \underset{\sim}{Z}_1'\ Z_1 + a^2\ U'\ U - 2\ a\ \underset{\sim}{U}'\underset{\sim}{X}_2 - 2\ b\ \underset{\sim}{Z}_1'\ \underset{\sim}{X}_2$$

$$+2\ a\ b\ \underset{\sim}{U}'\ \underset{\sim}{Z}_1$$

$$= \underset{\sim}{X}_2'\ \underset{\sim}{X}_2 + b^2\ \underset{\sim}{Z}_1'\ \underset{\sim}{Z}_1 + a^2 n - 2\ a\ \underset{\sim}{U}'\ \underset{\sim}{X}_2 - 2\ b\ \underset{\sim}{Z}_1'\ \underset{\sim}{X}_2 \quad ..(2.3)$$

Untuk $\underset{\sim}{U}'\underset{\sim}{U} = n$ dan batasan $\underset{\sim}{U}'\underset{\sim}{Z}_2 = 0$, sehingga batasan itu bisa dihilangkan, untuk mendapatkan harga ekstrim, maka f diderivatifkan parsial ke a dan b, kemudian disamakan dengan nol.

$$\frac{\partial f}{\partial a} = 2\ an - 2\ \underset{\sim}{U}'\ \underset{\sim}{X}_2 = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.4)$$

$$\frac{\partial f}{\partial b} = 2\ b\ \underset{\sim}{Z}_1'\ \underset{\sim}{Z}_1 - 2\ \underset{\sim}{Z}_1'\ \underset{\sim}{X}_2 = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots(2.5)$$

Dari persamaan (2.4) didapatkan $a = \frac{\underset{\sim}{U}'\underset{\sim}{X}_2}{n} = \overline{X}_2$, yaitu mean dari variabel X_2 dan dari persamaan (2.5) didapatkan

$$b = \frac{\underset{\sim}{Z}_1'\ \underset{\sim}{X}_2}{\underset{\sim}{Z}_i'\underset{\sim}{Z}_i} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.6)$$

Untuk jumlah kwadrat jarak f di peroleh dengan memasukkan solusi a dan b ke persamaan (2.3), sehingga didapatkan :

$$f = \underset{\sim}{X}_2'\ \underset{\sim}{X}_2 + \frac{(\underset{\sim}{Z}_1'\ \underset{\sim}{X}_2)^2}{\underset{\sim}{Z}_1'\ \underset{\sim}{Z}_1} + n\ \overline{X}_2^2 - 2\ \frac{(\underset{\sim}{Z}_1'\ \underset{\sim}{X}_2)^2}{\underset{\sim}{Z}_1'\ \underset{\sim}{Z}_1} - 2\ n\ \overline{X}_2^2$$

$$= \left(\underset{\sim}{X}_2\ \underset{\sim}{X}_2 - n\ \overline{X}_2^2\right) - \frac{(\underset{\sim}{Z}_1'\ \underset{\sim}{X}_2)^2}{n\ \underset{\sim}{Z}_1'\ \underset{\sim}{Z}_1} \quad \ldots\ldots\ldots\ldots\ldots\ldots(2.7)$$

jika kedua ruas persamaan (2.7) dibagi n, didapatkan variansi dari garis regresi, dinotasikan $S^2_{2.1}$, variansi ini merupakan variasi residu, sehingga di dapat :

$$S^2_{2.1} = \left(\frac{\underset{\sim}{U}' \underset{\sim}{X}_2}{n} - \bar{X}_2^{\,2} \right) - \frac{\left(\underset{\sim}{Z}_1' \underset{\sim}{X}_2 \right)^2}{n \, \underset{\sim}{Z}_1' \underset{\sim}{Z}_1} \quad \ldots\ldots\ldots\ldots\ldots\ldots (2.8)$$

$\left(\frac{\underset{\sim}{U}' \underset{\sim}{X}_2}{n} - \bar{X}_2^{\,2} \right)$ merupakan variansi dari variabel X_2 $(= S^2_2)$

sedangkan $\frac{\left(\underset{\sim}{Z}_1' \underset{\sim}{X}_2 \right)^2}{n \, \underset{\sim}{Z}_1' \underset{\sim}{Z}_1}$ dapat ditulis sebagai :

$$\frac{\left(\underset{\sim}{Z}_1' \underset{\sim}{X}_2 \right)^2}{n \, \underset{\sim}{Z}_1' \underset{\sim}{Z}_1} = \frac{\left(\frac{\underset{\sim}{U}' \underset{\sim}{X}_2}{n} \right)^2}{\frac{\underset{\sim}{Z}_1 \underset{\sim}{Z}_1}{n}} = \frac{\left(\mathrm{Cov}\ (X_1, X_2) \right)^2}{S^2_1} \quad \ldots\ldots\ldots\ldots (2.9)$$

sehingga persamaan (2.8) dapat disusun menjadi :

$$S^2_2 = S^2_{2.1} + \frac{\left(\mathrm{Cov}\ (X_1, X_2) \right)^2}{S^2_1} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.10)$$

Persamaan (2.10) diatas menunjukkan bahwa variansi variabel X_2 dapat dipisahkan menjadi 2 bagian, pertama sebagai variansi residu dan garis regresi dan yang kedua adalah bagian variansi X_2 yang dapat dijelaskan (dipengaruhi) oleh regresi X_2 atas X_1.

2.4. Korelasi Sederhana

Hubungan fungsional antara variabel- variabel telah disinggung pada bagian 2.1, Persoalan berikutnya ialah berapa kuat hubungan antara variabel - variabel itu, dengan kata lain perlu ditentukan derajat hubungan antara variabel - variabel tersebut. Sedangkan studi yang membahas tentang derajat hubungan

antara variabel-variabel tersebut dikenal dengan nama Analisis korelasi.

Analisis korelasi sukar dipisahkan dari Analisis regresi yang mana pembicaraannya akan lebih luas. Oleh karena itu kita ingin menentukan 2 garis regresi dalam sub bab 2.4 ini sebagai penjelasan, yaitu garis regresi X_2 atas X_1 dan juga sebaliknya garis regresi X_1 atas X_2. Untuk mempermudah dalam pembahasannya, maka akan kita lakukan transformasi. Transformasi untuk variabel X_1 dan X_2 berturut-turut menjadi variabel T_1 dan T_2.

Teorema 2.4.1 :

jika variabel X_1 dan X_2 telah ditransformasi ke bentuk variabel T (telah distandarisasi) maka untuk garis regresi T_2 atas T_1 berlaku hubungan, koefisien regresi (slope) b_1 sama dengan koefisien korelasi, r_{12} atau dituliskan :

$$b_1 = r_{12}$$

bukti :

Misalnya regresi T_2 atas T_1, berbentuk :

$$\hat{\underset{\sim}{T}}_2 = a + b_1 \underset{\sim}{T}_2$$

Dengan menggantikan :

Variabel (batasan) T_2 dari X_2 , T_1 dari z_1 dan b_1 dari b pada persamaan (2.3) didapatkan :

$$f = \underset{\sim}{T}_2' \underset{\sim}{T}_2 + b_1^2 \underset{\sim}{T}_1' \underset{\sim}{T}_1 + a^2 n - 2 a \underset{\sim}{U}' \underset{\sim}{T}_2 - 2 b_1 \underset{\sim}{T}_1' \underset{\sim}{T}_2$$

setelah f diderivatifkan parsial ke b_1 kemudian

disamakan nol, didapat :

$$b_1 = \frac{\underset{\sim}{T}_1' \underset{\sim}{T}_2}{\underset{\sim}{T}_1' \underset{\sim}{T}_1} = \frac{\underset{\sim}{Z}_1' \underset{\sim}{Z}_2}{S_1 S_2 \dfrac{\underset{\sim}{Z}_1' \underset{\sim}{Z}_1}{S_1^2}} = \frac{S_1 \underset{\sim}{Z}_1' \underset{\sim}{Z}_2}{S_2 \underset{\sim}{Z}_1' \underset{\sim}{Z}_2} = \frac{S_1 \dfrac{\underset{\sim}{Z}_1' \underset{\sim}{Z}_2}{n}}{S_2 \dfrac{\underset{\sim}{Z}_1' \underset{\sim}{Z}_2}{n}}$$

$$= \frac{S_1 \text{ Cov } (X_1, X_2)}{S_2 \; S_1^2} = \frac{\text{Cov } (X_1, X_2)}{S_1 \; S_2} = r_{12} \quad \ldots(2.11)$$

terbukti

Dengan menggunakan teorema (2.4.1), persamaan (2.8) dapat dituliskan menjadi :

$$S_{2.1}^2 = \left(\frac{\underset{\sim}{T}_2' \underset{\sim}{T}_2}{n} - \overline{T}_2\right) - \frac{\left(\underset{\sim}{T}_1' \underset{\sim}{T}_2\right)^2}{n \; \underset{\sim}{T}_1' \underset{\sim}{T}_2} = \frac{\underset{\sim}{T}_2' \underset{\sim}{T}_2}{n} - \overline{T}_2 - \frac{\left(\dfrac{\underset{\sim}{T}_1' \underset{\sim}{T}_2}{n}\right)^2}{\dfrac{\underset{\sim}{T}_1' \underset{\sim}{T}_1}{n}}$$

Karena mengingat sifat variabel T, maka dapat disederhanakan :

$$S_{2.1}^2 = 1 - \left(\frac{\underset{\sim}{T}_1' \underset{\sim}{T}_2}{n}\right)^2 = 1 - \left(\frac{\underset{\sim}{Z}_1' \underset{\sim}{Z}_2}{n \; S_1 S_2}\right)^2 = 1 - \left(\frac{\text{Cov } (X_1, X_2)}{S_1 \; S_2}\right)^2$$

$$= 1 - r_{12}^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.12)$$

Seperti pada persamaan (2.10), maka persamaan (2.12) dapat dituliskan

$$1 = S_{2.1}^2 + r_{12}^2 = S_{2.1}^2 + r^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.13)$$

Hal persamaan (2.13) dapat dituliskan demikian, karena akibat sifat dari matrik tranpose yaitu : $\underset{\sim}{X}_i' \underset{\sim}{X}_j = \underset{\sim}{X}_j' \underset{\sim}{X}_i$

$$r_{12} = \frac{\underset{\sim}{T}_1' \underset{\sim}{T}_2}{n} = \frac{\underset{\sim}{T}_2' \underset{\sim}{T}_1}{n} = r_{21} = r$$

Dengan melihat kembali persamaan (2.13), maka nampak bahwa r^2 merupakan bagian dari variansi X_2 (yang telah di standarisasi), yaitu bagian variansi X_2 yang dijelaskan atau dipengaruhi oleh regresi.

Seringkali r^2 ini disebut sebagai "koefisien determinasi", sebaliknya $1 - r^2$ adalah bagian variansi X_2 yang tidak dijelaskan oleh regresi atau sering disebut "variansi residu".

Dengan cara yang sama kita peroleh garis regresi untuk regresi variabel X_1 atas X_2, yaitu

$\hat{\underset{\sim}{T}}_1 = a + b_2 \underset{\sim}{T}_2$, sehingga didapat :

$b_2 = \dfrac{\underset{\sim}{T}_2' \underset{\sim}{T}_1}{n} = r$, sehingga dapat dikatakan bahwa koefisien regresi dari dua variabel yang berbentuk variabel T sama dengan koefisien korelasinya.