REGRESI LINIER DAN KORELASI

Variabel dibedakan dalam dua jenis dalam analisis regresi:

Variabel bebas atau variabel prediktor -> variabel yang mudah didapat atau tersedia. Dapat dinyatakan dengan $X_1, X_2, \ldots, X_k$ Dengan $k \geq 1$

Variabel tak bebas atau variabel respon -> variabel yang terjadi karena variabel bebas. Dapat dinyatakan dengan Y.

Contoh: fenomena yang meliputi hasil panen padi dengan volume pupuk yang digunakan, sebaiknya diambil variabel bebas X = volume pupuk dan variabel takbebas Y = hasil panen padi.

Persamaan regresi secara umum:

$$\mu_{y,x_1,x_2,\ldots,x_k} = f(X_1, X_2, \ldots, X_k | \theta_1, \theta_2, \ldots, \theta_m)$$

Dengan $\theta_1, \theta_2, \ldots, \theta_m$ parameter-parameter yang ada dalam regresi itu.

Jenis-jenis regresi adalah

1. Regresi linier: regresi linier sederhana dan regresi linier berganda
2. Regresi non linier: regresi eksponensial, regresi parabola kuadratik, regresi parabola kubik, regresi logistik, regresi geometrik

1. Regresi Linier

A. Regresi Linier Sederhana

Merupakan regresi dengan satu variabel bebas, regresi dengan variabel bebas X dan variabel takbebasnya Y atau dinamakan juga regresi Y atas X, bentuk persamaannya:

$$\mu_{y.x} = \theta_1 + \theta_2 X$$

Regresi linear sederhana berdasarkan sampel, maka θ_1 ditaksir dengan a dan θ_2 ditaksir dengan b diperoleh:

$$\hat{Y} = a + bX$$

Cara penentuan nilai a dan b dapat dilakukan dengan dua cara:

1. Metode tangan bebas

Metode Tangan Bebas adalah metode penentuan persamaan regresi kira-kira menggunakan diagram pencar. Jika fenomena meliputi sebuah variabel bebas X dan variabel tak bebas Y, maka data yang didapat digambarkan pada diagram dengan sumbu datar menyatakan X dan sumbu tegak menyatakan Y.

Jika letak titik-titik sekitar garis lurus maka untuk menentukan persamaan regresinya, dapat dicari dengan menggunakan dua titik yang dilalui garis tersebut, kemudian dicari persamaan garisnya, yaitu jika garis melewati titik-titik (x_1, y_1) dan (x_2, y_2) maka:

$$\frac{y - y_1}{y_2 - y_1} = \frac{x - x_1}{x_2 - x_1}$$

Penetuan regresi dengan cara ini bersifat tidak tunggal, artinya tiap orang akan memberikan perkiraan yang berbeda bergantung pada pertimbangan pribadi masing-masing.

2. Metode kuadrat terkecil

Seperti dikatakan sebelumnya regresi dengan variabel bebas X dan variabel takbebas Y dimana model regresi linier untuk populasi yaitu

$$\mu_{y.x} = \theta_1 + \theta_2 X$$

akan ditaksir harga-harga θ_1 dan θ_2 oleh a dan b sehingga didapat persamaan regresi menggunakan data sampel:

$$\hat{Y} = a + bX$$

Dengan

$$a = \frac{(\sum Y_i)\left(\sum X_i^2\right) - (\sum X_i)(\sum X_i Y_i)}{n \sum X_i^2 - (\sum X_i)^2}$$

$$b = \frac{n \sum X_i Y_i - (\sum X_i)(\sum Y_i)}{n \sum X_i^2 - (\sum X_i)^2}$$

Dimana n adalah jumlah sampel, X_i adalah data variabel bebas ke – i dan Y_i adalah data variabel takbebas ke – i.

Jika terlebih dahulu dihitung koefisien b, maka koefisien a dapat pula ditentukan oleh rumus:

$$a = \bar{Y} - b\bar{X}$$

dimana $\bar{X}$ dan $\bar{Y}$ adalah rata-rata untuk masing-masing variabel X dan Y.

Koefisien b dinamakan koefisien arah regresi linier dan menyatakan perubahan rata-rata variabel Y untuk setiap perubahan variabel X sebesar satu unit. Perubahan ini merupakan pertambahan jika b bertanda positif dan penurunan atau pengurangan jika bertanda negatif.

Contoh:

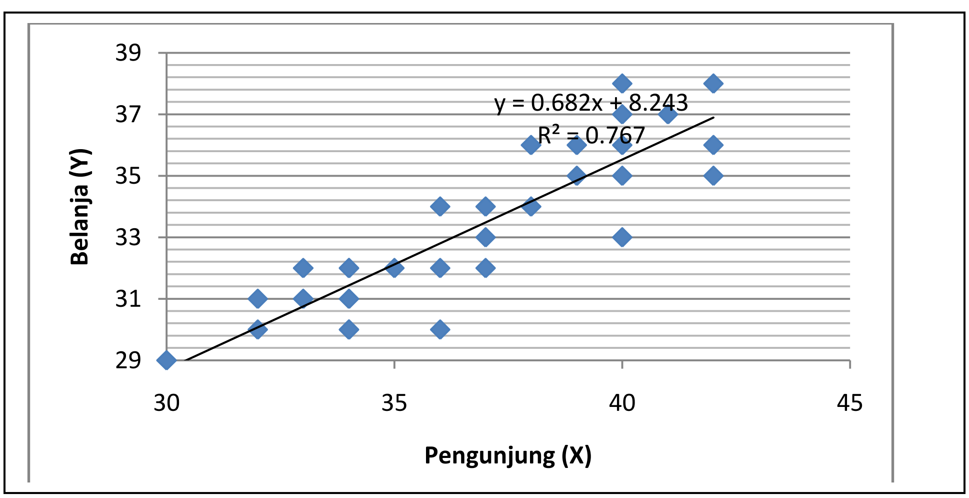
Data berikut melukiskan hasil pengamatan mengenai banyak orang yang datang (X) dan banyak orang yang berbelanja (Y) di sebuah toko selama 30 hari.

Daftar 2.1

Banyak Pengunjung dan yang Berbelanja

Di Sebuah Toko Selama 30 Hari

Pengunjung (X_i)	Berbelaja (Y_i)	Pengunjung (X_i)	Berbelanja (Y_i)
34	32	42	38
38	36	41	37
34	31	32	30
40	38	34	30
30	29	36	30
40	35	37	33
40	33	36	32
34	30	37	34
35	32	39	35
39	36	40	36
33	31	33	32
32	31	34	32
42	36	36	34
40	37	37	32
42	35	38	34



Gambar 2.1

Jawab:

X_i	Y_i	X_iY_i	X_i^2	X_i	Y_i	X_iY_i	X_i^2
34	32	1088	1156	42	38	1596	1764
38	36	1368	1444	41	37	1517	1681
34	31	1054	1156	32	30	960	1024
40	38	1520	1600	34	30	1020	1156
30	29	870	900	36	30	1080	1296
40	35	1400	1600	37	33	1221	1369
40	33	1320	1600	36	32	1152	1296
34	30	1020	1156	37	34	1258	1369
35	32	1120	1225	39	35	1365	1521
39	36	1404	1521	40	36	1440	1600
33	31	1023	1089	33	32	1056	1089
32	31	992	1024	34	32	1088	1156
42	36	1512	1764	36	34	1224	1296
40	37	1480	1600	37	32	1184	1369
42	35	1470	1764	38	34	1292	1444

Setelah dijumlahkan didapat:

$\sum X_i = 1105$, $\sum Y_i = 1001$, $\sum X_iY_i = 37094$ dan $\sum X_i^2 = 41029$

Maka diperoleh:

$$a = \frac{(1001)(41029) - (1105)(37094)}{30(41029) - (1105)^2} = 8{,}2437$$

$$b = \frac{30(37094) - (1105)(1001)}{30(41029) - (1105)^2} = 0{,}6821$$

Sehingga persamaan linier Y atas X adalah

$$\hat{Y} = 8{,}2437 + 0{,}6821X$$

Artinya untuk setiap X bertambah dengan seorang maka rata-rata pembeli Y bertambah dengan 0,68 orang.

Berbagai Varians Sehubungan Dengan Regresi Linear Sederhana

Untuk analisis selanjutnya tentang regresi linier sederhana, beberapa asumsi harus diambil.

Pertama, mengingat hasil pengamatan variabel takbebas Y belum tentu sama besarnya dengan harga diharapkan, yakni $\hat{Y}$ yang didapat dari regresi hasil pengamatan, maka terjadi perbedaan $e = \left|Y - \hat{Y}\right|$, biasa disebut kekeliruan prediksi atau galat prediksi. Dalam populasi, galat prediksi ini dimisalkan berbentuk variabel acak yang mengikuti distribusi normal dengan rata-rata nol dan varians σ_ϵ^2.

Kedua, untuk setiap harga X yang diberikan, variabel tak bebas Y independen dan berdistribusi normal denga rata-rata $(\theta_1 + \theta_2 X)$ dan varians $\sigma^2_{Y.X}$. Varians $\sigma^2_{Y.X}$ dimisalkan sama untuk setiap X dan karenanya dapat dinyatakan oleh σ^2_{ϵ} yang biasa pula dinamakan *varians kekeliruan taksiran* sedangkan $\sigma_{y.x}$ dikenal dengan *kekeliruan baku taksiran*.

Berpegang kedua asumsi di atas, maka varians σ^2_{ϵ} ditaksir oleh rata-rata kuadrat penyimpangan sekitar regresi atau disebut juga rata-rata kuadrat residu, dinyatakan oleh varians s^2_e yaitu

$$s^2_{Y.X} = s^2_e = \frac{\sum\left(Y_i - \widehat{Y}_i\right)^2}{(n-2)}$$

Dengan Y = variabel tak bebas hasil pengamatan dan $\widehat{Y}$ = didapat dari regresi berdasarkan sampel, dan n = ukuran sampel.

Dapat ditulis juga

$$s^2_{Y.X} = \left(\frac{n-1}{n-2}\right)(s^2_Y - b^2 s^2_X)$$

Dengan s^2_Y dan s^2_X masing-masing menyatakan varians untuk variabel-variabel Y dan X.

Varians koefisien b:

$$s^2_b = \frac{s^2_{Y.X}}{\sum(X_i - \bar{X})^2}$$

Varians koefisien a:

$$s^2_a = s^2_{Y.X}\left\{\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right\}$$

Varians ramalan rata-rata Y untuk X_0 yang diketahui:

$$s^2_{\hat{Y}} = s^2_{Y.X}\left\{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right\}$$

Varians ramalan individu Y untuk X_0 yang diketahui

$$s^2_{\hat{Y}} = s^2_{Y.X}\left\{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right\}$$

Untuk rumus-rumus di atas $\sum(X_i - \bar{X})^2$ dapat diganti oleh $\left\{\sum X_i^2 - \frac{(\sum x_i)^2}{n}\right\}$

KED

Contoh:

Untuk dengan data dalam daftar 2.1, kita dapat menghitung varians-varians di atas. Kita perlu $\bar{X} = 36{,}8$; $s_x^2 = 11{,}32$; $s_y^2 = 6{,}86$ dan$\sum(X_i - \bar{X})^2 = 328{,}2$, diperoleh b = 0,68 dan n = 30 didapat

$$s_{Y.X}^2 = \frac{30-1}{30-2}\{6{,}86 - (0{,}68)^2(11{,}32)\} = 1{,}684$$

$$s_b^2 = \frac{1{,}684}{328{,}2} = 5{,}13 \times 10^{-3}$$

$$s_a^2 = 1{,}684\left\{\frac{1}{30} + \frac{36{,}8^2}{328{,}2}\right\} = 7{,}005$$

Varians ramalan rata-rata Y untuk $X_0 = 40$ adalah

$$s_{\hat{Y}}^2 = 1{,}684\left\{\frac{1}{30} + \frac{(40-36{,}8)^2}{328{,}2}\right\} = 0{,}1087$$

Varians ramalan individu Y untuk $X_0 = 40$ adalah

$$s_{\hat{Y}}^2 = 1{,}684\left\{1 + \frac{1}{30} + \frac{(40-36{,}8)^2}{328{,}2}\right\} = 1{,}7927$$

Korelasi

Koefisien korelasi (r): ukuran hubungan linier peubah X dan Y

Nilai r berkisar antara -1 dan 1.

$r = -1$	Hubungan linier sempurna tak langsung. Titik-titik yang ditentukan oleh (X_i, Y_i) seluruhnya terletak pada garis regresi dan harga X besar berpasangan dengan Y kecil dan X kecil berpasangan dengan Y besar
$r = +1$	Hubungan linier sempurna langsung. Titik-titik yang ditentukan oleh (X_i, Y_i) seluruhnya terletak pada garis regresi dan harga X besar berpasangan dengan Y besar dan X kecil berpasangan dengan Y kecil
$r < 1$	Korelasi tak langsung atau korelasi negatif
$r > 1$	Korelasi langsung atau korelasi positif
$r = 0$	Tidak terdapat hubungan linier antara variabel X dan Y

Untuk keperluan perhitungan koefisien korelasi r berdasarkan sekumpulan data (X_i, Y_i) berukuran n dapat digunakan rumus:

$$r = \frac{n\sum X_iY_i - (\sum X_i)(\sum Y_i)}{\sqrt{\{n\sum X_i^2 - (\sum X_i)^2\}\{n\sum Y_i^2 - (\sum Y_i)^2\}}}$$

Bentuk lain dapat pula digunakan:

$$r = \sqrt{1 - \frac{s_{y.x}^2}{s_y^2}}$$

dengan $s_{y.x}$ = kekeliruan baku taksiran dan s_y = simpangan baku untuk variabel Y.

Jika persamaan regresi linier Y atas X telah ditentukan dan sudah didapat koefisien arah,b, maka koefisien determinasi, r^2, dapat ditentukan oleh rumus:

$$r^2 = \frac{b\{n\sum X_i Y_i - (\sum X_i)(\sum Y_i)\}}{n\sum Y_i^2 - (\sum Y_i)^2}$$

atau dapat juga menggunakan formula:

$$r = \frac{bs_x}{s_y}$$

dengan s_x simpangan baku untuk variabel X dan s_y simpangan baku untuk variabel Y.

Jika b_1 adalah koefisien arah regresi Y atas X dan b_2 adalah koefisien arah regresi X atas Y untuk data yang sama, maka

$$r^2 = b_1 b_2$$

Rumus ini menyatakan bahwa koefisien korelasi r adalah rata-rata ukur daripada koefisien-koefisien arah b_1 dan b_2.

Contoh:

Untuk dengan data dalam daftar 2.1, untuk menentukan korelasi diperlu $\bar{X} = 36{,}8$; $s_x^2 = 11{,}32$; $s_y^2 = 6{,}86$ dan$\sum(X_i - \bar{X})^2 = 328{,}2$, diperoleh b = 0,68 dan n = 30 didapat

$$r = 0.68\sqrt{\frac{11.32}{6.86}} = 0.87$$

B. Regresi Linier Berganda

Banyak data pengamatan yang terjadi sebagai akibat lebih dari dua variabel. Misalnya, rata-rata pertambahan berat daging sapi (Y) bergantung pada berat permulaan (X_1), umur sapi ketika pengamatan dimulai dilakukan (X_2), berat makanan yang diberikan setiap hari (X_3) dan mungkin masih ada faktor lain.

Akan ditentukan hubungan antara Y dan $X_1, X_2, \dots, X_k$ sehingga didapat regresi Y atas $X_1, X_2, \dots, X_k$. Yang akan ditinjau di sini hanyalah garis regresi sederhana ialah yang dikenal dengan *regresi linier ganda*. Model tersebut ditaksir oleh:

$$\hat{Y} = a_0 + a_1 X_1 + a_2 X_2 + \cdots + +a_k X_k$$

Koefisien-koefisien $a_0, a_1, \dots, a_k$ ditentukan dengan menggunakan metode kuadrat terkecil.

Untuk regresi linier ganda dua variabel bebas:

$$\hat{Y} = a_0 + a_1 X_1 + a_2 X_2$$

maka untuk mengetahui koefisien-koefisiennya harus menyelesaikan persamaan-persamaan berikut:

$$\boxed{\begin{gathered} \sum Y_i = a_0 n + a_1 \sum X_{1_i} + a_2 \sum X_{2_i} \\ \sum Y_i X_{1_i} = a_0 \sum X_{1_i} + a_1 \sum X_{1\,i}^2 + a_2 \sum X_{1_i} X_{2_i} \\ \sum Y_i X_{2_i} = a_0 \sum X_{2_i} + a_1 \sum X_{1_i} X_{2_i} + a_2 \sum X_{2\,i}^2 \end{gathered}}$$

persamaan (*)

$$a_0 = \bar{Y} - a_1 \overline{X_1} - a_2 \overline{X_2}$$

$$a_1 = \frac{\left(\sum x_{1\,i}^2\right)\left(\sum x_{1_i} y_i\right) - \left(\sum x_{1_i} x_{2_i}\right)\left(\sum x_{2_i} y_i\right)}{\left(\sum x_{1\,i}^2\right)\left(\sum x_{2\,i}^2\right) - \left(\sum x_{1_i} x_{2_i}\right)^2}$$

$$a_2 = \frac{\left(\sum x_{1\,i}^2\right)\left(\sum x_{2_i} y_i\right) - \left(\sum x_{1_i} x_{2_i}\right)\left(\sum x_{1_i} y_i\right)}{\left(\sum x_{1\,i}^2\right)\left(\sum x_{2\,i}^2\right) - \left(\sum x_{1_i} x_{2_i}\right)^2}$$

Koefisien a_1 merupakan perubahan rata-rata Y unutuk setiap perubahan satuan dalam variabel X_1 apabila $X_2, X_3, \dots, X_k$ semua dianggap tetap, begitu juga a_2 merupakan perubahan rata-rata Y unutuk setiap perubahan satuan dalam variabel X_2 apabila $X_1, X_3, \dots, X_k$ semua dianggap tetap dan begitu seterusnya. Jelas bahwa di sini setiap koefisien hanya memberikan gambaran parsial apa yang terjadi pada Y untuk perubahan X yang berhubungan dengan koefisien dimaksud. Karenanya, koefisien-koevisien $a_1, a_2, \dots, a_k$ disebut pula *koefisien regresi parsil.*

Untuk regresi linier ganda variabel, maka ukuran kekelirua yang digunakan adalah:

$$s_{y.1,2,\dots,k}^2 = \frac{\sum\left(Y_i - \widehat{Y_i}\right)^2}{n - k - 1}$$

dimana Y_i = nilai data hasil pengamatan dan $\widehat{Y_i}$ = nilai harapan yang didapat dari persamaan regresi.

Korelasi Linier Berganda

Koefisien Determinasi Sampel untuk Regresi Linier Berganda diberi notasi sebagai berikut $R_{y.12}^2$

Sedangkan Koefisien korelasi adalah **akar positif** koefisien determinasi atau

$$r_{y.12} = \sqrt{R^2_{y.12}}$$

Rumus:

$$R^2_{y.12} = 1 - \frac{JKG}{(n-1)s_y^2}$$

Keterangan:

JKG: Jumlah Kuadrat Galat

s_y^2: Jumlah Kuadrat y (terkoreksi)

Dimana:

$$s_y^2 = \frac{n\sum y^2 - (\sum y)^2}{n(n-1)}$$

$$JKG = \sum y^2 - a_0 \sum y - a_1 \sum x_1 y - a_2 \sum x_2 y$$

(sumber: thomasyunigunarto)

Berikut adalah data Volume Penjualan (juta unit) Mobil dihubungkan dengan variable biaya promosi (X_1 dalam juta rupiah/tahun) dan variable biaya penambahan asesoris (X_2 dalam ratusan ribu rupiah/unit)

x_1	x_2	y	x_1x_2	x_1y	x_2y	x_1^2	x_2^2	y^2
2	3	4	6	8	12	4	9	16
3	4	5	12	15	20	9	16	25
5	6	8	30	40	48	25	36	64
6	8	10	48	60	80	36	64	100
7	9	11	63	77	99	49	81	121
8	10	12	80	96	120	64	100	144
$\sum x_1 = 31$	$\sum x_2 = 40$	$\sum y = 50$	$\sum x_1x_2 = 239$	$\sum x_1y = 296$	$\sum x_2y = 379$	$\sum x_1^2 = 187$	$\sum x_2^2 = 306$	$\sum y^2 = 470$

Tetapkan persamaan regresi linier berganda: $a_0 + a_1X_1 + a_2X_2$

Substitusi kedalam persamaan:

(i) $6a_0 + 31a_1 + 40a_2 = 50$
(ii) $31a_0 + 187a_1 + 239a_2 = 296$
(iii) $40a_0 + 239a_1 + 306a_2 = 379$

Gunakan eliminasi dan substitusi untuk mendapatkan nilai a_0, a_1, a_2 yaitu:

Sehingga diperoleh:

$a_0 = \frac{3}{4}, a_1 = \frac{1}{2}, a_3 = \frac{3}{4}$

Maka bentuk regresi linier berganda:

$$\hat{Y} = \frac{3}{4} + \frac{1}{2}X_1 + \frac{3}{4}X_2$$

Koefisien korelasi adalah

$$s_y^2 = \frac{6(470) - 50^2}{6(5)} = 10.667$$

$$JKG = 470 - 0.75(50) - 0.5(296) - 0.75(379) = 0.25$$

$$R_{y.12}^2 = 1 - \frac{0.25}{5(10.667)} = 0.99$$

2. Regresi Non Linier

1. Parabola kuadratik: $\hat{Y} = a + bX + cX^2$
   Dengan menggunakan metode kuadrat terkecil, maka a, b, dan c dapat dihitung dari sistem persamaan:
$$\sum Y_i = na + b\sum X_i + c\sum X_i^2$$
$$\sum X_iY_i = a\sum X_i + b\sum X_i^2 + c\sum X_i^3$$
$$\sum X_i^2Y_i = a\sum X_i^2 + b\sum X_i^3 + c\sum X_i^4$$

2. Parabolik kubik: $\hat{Y} = a + bX + cX^2 + dX^3$
   Untuk menentukan nilai a, b, c, dan d gunakan sistem persamaan berikut:
$$\sum Y_i = na + b\sum X_i + c\sum X_i^2 + d\sum X_i^3$$
$$\sum X_iY_i = a\sum X_i + b\sum X_i^2 + c\sum X_i^3 + d\sum X_i^4$$
$$\sum X_i^2Y_i = a\sum X_i^2 + b\sum X_i^3 + c\sum X_i^4 + d\sum X_i^5$$
$$\sum X_i^3Y_i = a\sum X_i^3 + b\sum X_i^4 + c\sum X_i^5 + d\sum X_i^6$$

3. Eksponen : $\hat{Y} = ab^X$
   Besar nilai a dan b ditentukan menggunakan persamaan:
$$\log a = \frac{\sum \log Y_i}{n} - (\log b)\left(\frac{\sum X_i}{n}\right)$$
$$\log b = \frac{n(\sum X_i \log Y_i) - (\sum X_i)(\sum \log Y_i)}{n\sum X_i^2 - (\sum X_i)^2}$$

4. Geometrik: $\hat{Y} = aX^b$
   Besar nilai a dan b ditentukan menggunakan persamaan:

$$\log a = \frac{\sum \log Y_i}{n} - b\frac{\sum \log X_i}{n}$$
$$b = \frac{n(\sum \log X_i \log Y_i) - (\sum \log X_i)(\sum \log Y_i)}{n\sum log^2 X_i - (\sum \log X_i)^2}$$

5. Logistik: $\hat{Y} = \frac{1}{ab^X + c}$

$$\log a = \frac{\sum \log\left(\frac{1}{Y_i}\right)}{n} - (\log b)\left(\frac{\sum X_i}{n}\right)$$
$$\log b = \frac{n\left(\sum X_i \log\left(\frac{1}{Y_i}\right)\right) - (\sum X_i)\left(\sum \log\left(\frac{1}{Y_i}\right)\right)}{n\sum X_i^2 - (\sum X_i)^2}$$

6. Hiperbola: $\hat{Y} = \frac{1}{a + bX}$

   Jika $\hat{Y}$ tidak ada yang bernilai nol, maka a dan b adalah

$$a = \frac{\left(\sum \frac{1}{Y_i}\right)(\sum X_i^2) - (\sum X_i)\left(\sum X_i \frac{1}{Y_i}\right)}{n\sum X_i^2 - (\sum X_i)^2}$$
$$b = \frac{n\sum X_i \frac{1}{Y_i} - (\sum X_i)\left(\sum \frac{1}{Y_i}\right)}{n\sum X_i^2 - (\sum X_i)^2}$$