Regresi Linear Sederhana
dan
Korelasi

Analisa Regresi Linear

Analisa regresi digunakan untuk meramalkan nilai dari satu peubah (peubah Terikat) berdasarkan peubah yang yang lain (peubah bebas).

Peubah Terikat: dituliskan sebagai **Y**
Peubah Bebas: dituliskan sebagai $\mathbf{X_1, X_2, \ldots, X_k}$

Jika hanya terdapat satu peubah bebas, maka ia disebut regresi linear sederhana, yang modelnya adalah sebagai berikut:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Yang kita lakukan adalah memperkirakan β_0 dan β_1 dari data yang telah dikumpulkan.

Analisa Regresi Linear

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Peubah:

X = Peubah Bebas (Harus tersedia)

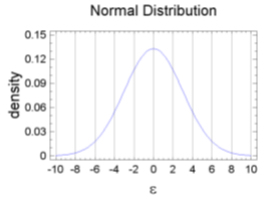
Y = Peubah Terikat (akan diperkirakan)

Parameter:

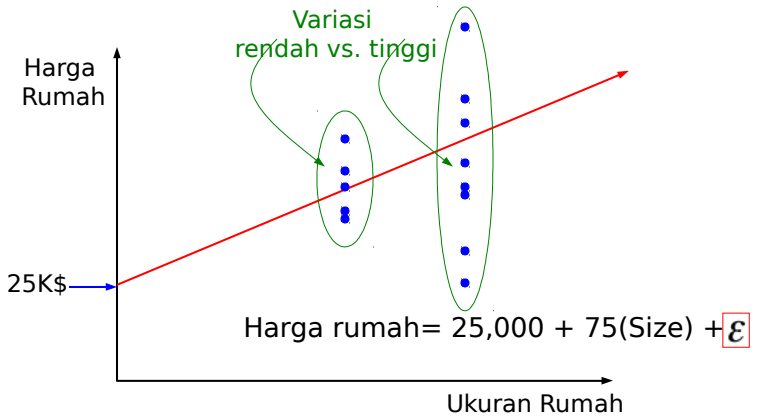
β_0 = Y-Intercept

β_1 = Slope

ε ~ Peubah Acak yang mengikuti distribusi Normal (μ_ε = 0, σ_ε = ???) [Noise]

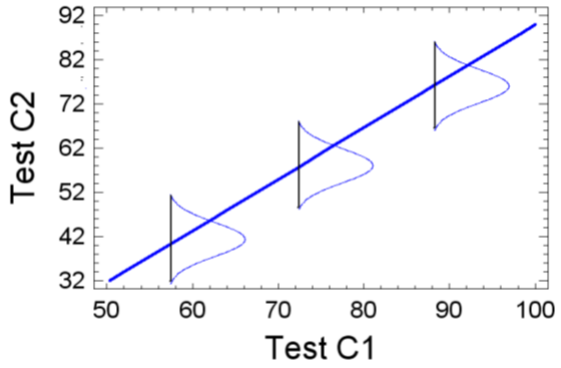


Akibat dari nilai σ_ε yang membesar



Teori Model Linear

$$y = \beta_0 + \beta_1 x + \varepsilon$$



1. Membangun Model – mengumpulkan Data

Nilai ujian 2= β_0 +β1*(Nilai ujian1)

Dari Data:

Estimasi β_0

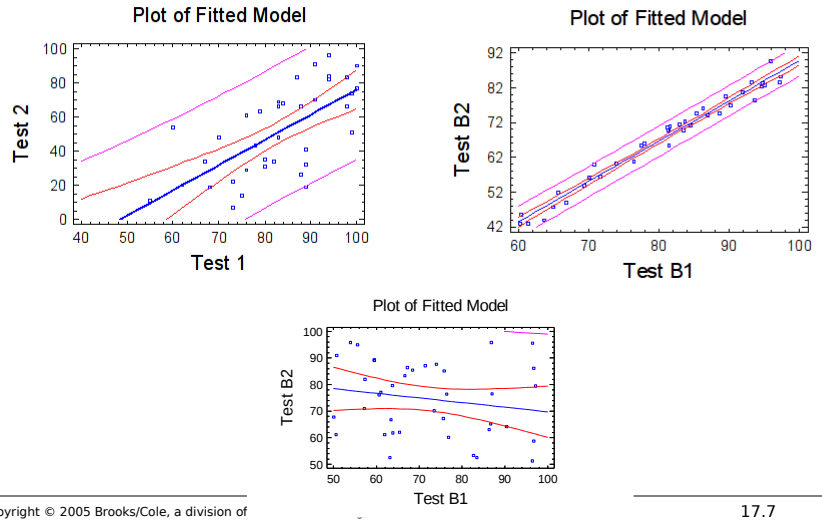
Estimasi β_1

Estimasi σ_ε

Student	Test 1	Test 2
1	50	32
2	51	33
3	52	34
4	53	35
5	54	36
6	55	37
7	56	39
8	57	40
9	58	41
10	59	42
11	60	43
12	61	44
13	62	46
14	63	47
15	64	48
16	65	49
17	66	50
18	67	51
19	68	53
20	69	54
21	70	55
22	71	56
23	72	57

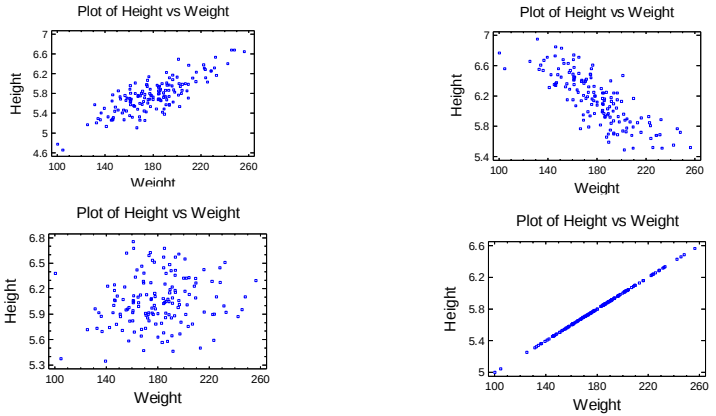
Linear Regression Analysis...

$$y = \beta_0 + \beta_1 x + \varepsilon$$



Analisa Korelasi... "-1 ≤ ρ < 1"

Jika kita hanya ingin mengetahui apakah trdapat relasi (hubungan) antara dua peubah, maka gunakan analisa korelasi. contoh: Berat badan dan Tinggi badan.



Analisa Korelasi... "$-1 \leq \rho < 1$"

Jika koefisien korelasi dekat ke +1 artinya terdapat hubungan positif yang kuat antar dua peubah.

Jika koefisien korelasi dekat ke -1 artinya terdapat hubungan negatif yang kuat antar dua peubah.

ika koefisien korelasi dekat ke 0 artinya tidak terdapat hubungan antar dua peubah.

Pada analisa korelasi, bisa dilakukan uji hipotesia

$H_0: \rho = 0$

Regresi: Model ... X=ukuran rumah, Y=charga rumah

Model Deterministik: sebuah atau kumpulan persamaan yang memperbolehkan kita untuk **memperkirakan secara keseluruhan** nilai dari peubah terikat yang dipengaruhi olehpeubah bebas.

$y = \$25{,}000 + (75\$/ft^2)(x)$

Daerah lingkaran: $A = \pi * r^2$

Model Probabilistik: sebuah metode yang digunakan untuk **menangkap keacakan** yang merupakan bagian dari proses sebenarnya yang terjadi.

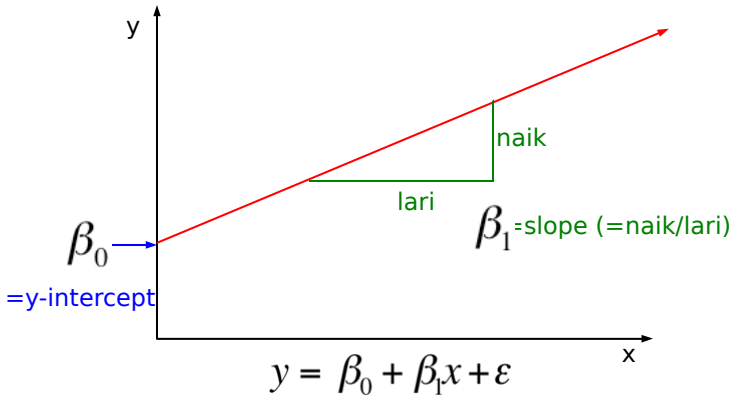
$y = 25{,}000 + 75x + \varepsilon$

Contoh. Apakah rumah dengan ukuran yang sama akan terjual dengan harga yang sama?

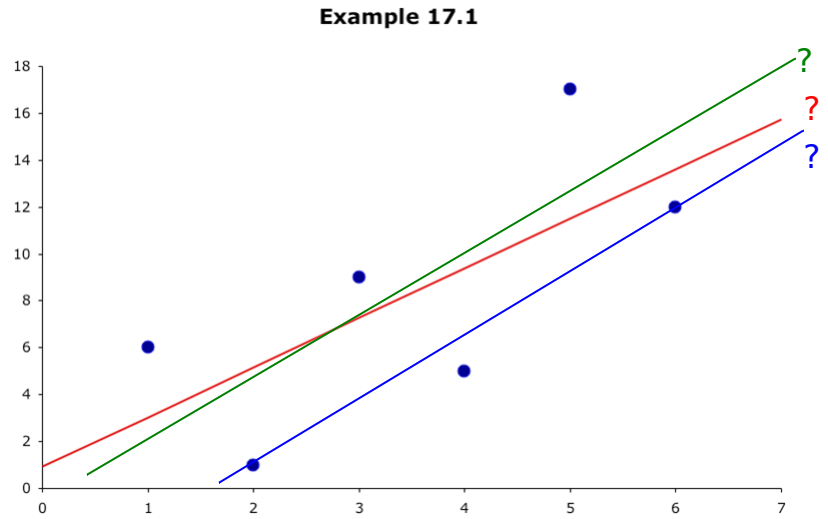
Model Regresi Linear Sederhana

Arti dari β_0 dan β_1

$\beta_1 > 0$ [slope positif] $\beta_1 < 0$ [slope negatif]



Yang mana garis terbaik?



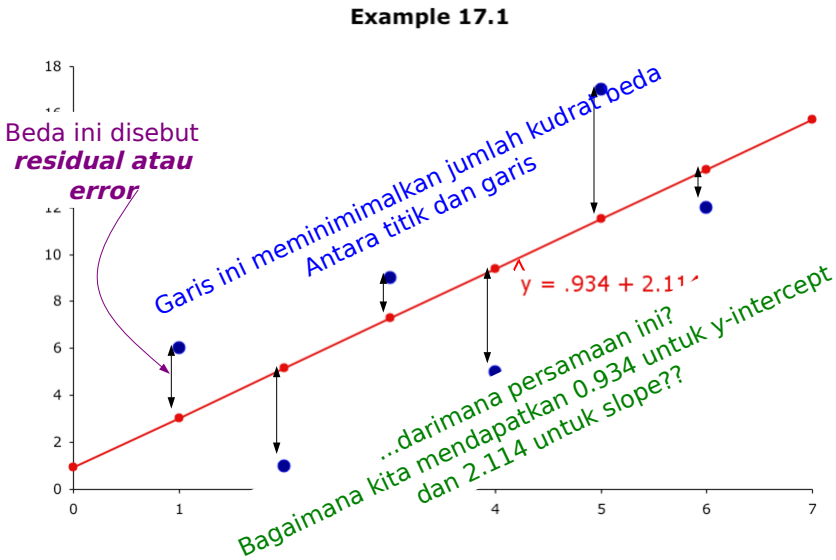
Memperkirakan Koefisien

Dengan dasar yang sama untuk memperkirakan μ dengan $\bar{x}$, perkirakan β_0 dengan b_0 dan β_1 dengan b_1, y-intercept dan slope dengan metode ***least squares*** atau ***garis regresi*** diberikan oleh:

$$\hat{y} = b_0 + b_1 x \qquad y = \beta_0 + \beta_1 x$$

(Penggunakan metode least squares dan menghasilkan garis lurus yang meminimilkan jumlah beda kuadrat antara titik sebenarnya dengan garis regresi)

Garis Least Squares



Garis Least Squares Line

Nilai $\mathbf{b_1}$ dan $\mathbf{b_0}$ f line…

$$\hat{y} = b_0 + b_1 x$$

…dihitung sebagai berikut:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

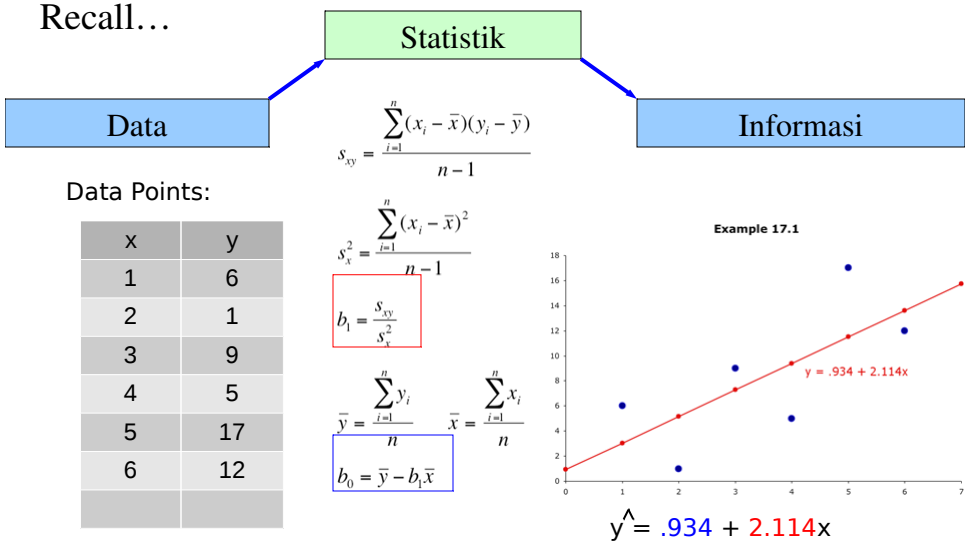
$$b_1 = \frac{s_{xy}}{s_x^2}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Garis least Square

Recall…



$\hat{y} = .934 + 2.114x$

Garis Least Squares

	X	Y	X - Xbar	Y - Ybar	(X-Xbar)*(Y-Ybar)	$(X - Xbar)^2$
	1	6	-2.500	-2.333	5.833	6.250
	2	1	-1.500	-7.333	11.000	2.250
	3	9	-0.500	0.667	-0.333	0.250
	4	5	0.500	-3.333	-1.667	0.250
	5	17	1.500	8.667	13.000	2.250
	6	12	2.500	3.667	9.167	6.250
Sum =	**21**	**50**	**0.000**	**0.000**	**37.000**	**17.500**

Xbar =	**3.500**	
Ybar =	**8.333**	
s_{xy} =	**7.400**	37.00/(6-1)
s_x^2 =	**3.500**	17.5/(6-1)
b_1 =	**2.114**	7.4/3.5
b_0 =	**0.933**	8.33 - 2.114*3.50

Syarat yang harus dipenuhi

Dalam menggunakan metode regresi, syarat berikut harus dipenuhi, jika tidak maka model yang didapat tidak valid. Syarat tersebut adalah:

1. Distribusi peluang dari ε adalahnormal.
2. Mean dari distribusi ε adalah 0, yaitu The mean of the $E(\varepsilon) = 0$.
3. Standar deviasi dari ε yaitu σ_ε, adalah konstan berapapun nilai dari x.

• Nilai ε yang berhubungan dengan nilai y tertentu adalah saling bebas dengan nilai ε yang berhubungan dengan y yang lain.

Melihat kecocokan model

Metode least square akan selalu menghasilkan garis lurus, walaupun sebenarnya tidak ada hubungan antara kedua peubah, atau hubungan kedua peubah bukanlah hubungan linear (misal kuadrat, atau log).

Sehingga selain melihat koefisien dari garis least square, harus dilihat pula seberapa cocok (benar) model yang dipilih. Untuk melihat kecocokan ini, maka harus dilihat nilai dari sum of squares for errors (SSE).

Sum of Squares for Error (SSE)

Sum of squares for error dihitung sebagai berikut:

$$SSE = (n-1)\left(s_y^2 - \frac{s_{xy}^2}{s_x^2}\right)$$

Dan digunakan untuk menghitung estimasi ***standard error*** :

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$$

Jika S_ε maka semua titik akan berada pada garis regresi.

Standard Error...

Xm17-02.xls [Read-Only]

	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	*Regression Statistics*					
4	Multiple R	0.8052				
5	R Square	0.6483				
6	Adjusted R Square	0.6447				
7	Standard Error	0.3265				
8	Observations	100				
9						
10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	Regression	1	19.26	19.26	180.64	0.00
13	Residual	98	10.45	0.11		
14	Total	99	29.70			
15						
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	
17	**Intercept**	**17.25**	0.182	94.73	0.0000	
18	**Odometer**	**-0.0669**	0.0050	-13.44	0.0000	

S_ε

Jika S_ε bernilai kecil, maka model sangat cocok untuk data yang dipakai. Jika tidak maka model tidak sesuai.

Standard Error

Untuk mengetahui apakah nilai Standard Error cukup kecil, bandingkan dengan nilai dari mean sampel peubah terikat. ($\bar{y}$).

Pada contoh,

s_ε = .3265 and

$\bar{y}$ = 14.841

Bisa dikatakan bahwa nilai standard Error cukup kecil, sehingga model cukup bagus.

Menguji Slope

Jika tidak terdapat hubungan linear antara dua peubah, maka garis regresi seharusnya berbentuk garis horisontal, artinya slope seharusnya bernilai nol (0).

Untuk melihat apakah hubungan kedua peubah adalah linear, maka kita uji menggunakan hipotesis sebagai berikut::

$H_1: \beta_1 \neq 0$

Null hypothesis adalah:

$H_0: \beta_1 = 0$

Lihat kembali bab hipotesis!

Menguji Slope

Untuk menguji hipotesis maka digunakan statistik berikut:

H0: $\beta_1 = 0$

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

Dimaka S_{b1} adalah standar deviasi dari b_1, didefinisikan:

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1)s_x^2}}$$

Jika error berdistribusi normal, maka statistik diatas mengikuti distribusi Student **tdengan df** n–2. Daerah penolakan biasanya menggunakan 2 sisi.

Menguji Slope

Uji hipotesis untuk melihat apakah slope secara signifikan berbeda dari "0" (dengan tingkat kepercayaan 5%)

Yang diuji adalah:

H_1 $\beta_1 \neq 0$

H_0: $\beta_1 = 0$

Daerah penolakan adalah:

$$t < -t_{\alpha/2,\nu} = -t_{.025,98} \approx -1.984 \quad or \quad t > t_{\alpha/2,\nu} = t_{.025,98} \approx 1.984$$

ATAU lihat p-value.

Menguji Slope

Atau dapatkan selang kepercayaan dari slope. Ingat bahwa perkiraan β_1 adalah b_1.

Selang kepercayaan diberikan oleh:

$$b_1 \pm t_{\alpha/2} s_{b_1} \quad \nu = n - 2$$

Sehingga: $b_1 \pm t_{\alpha/2} s_{b_1} = -.0669 \pm 1.984(.00497) = -.0669 \pm .0099$

Maka perkiraan selang dari koefisien slope adalah

–.0768 dan –.0570

Koefisien Determinasi

Selain melihat apakah kedua peubah mempunyai hubungan linear, penting juga untuk melihat ukuran kekuatan hubungan antara keduanya. Untuk itu perlu dilihat ***coefisien determinasi*** – R^2.

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}, \; or \; R^2 = 1 - \frac{SSE}{\sum (y_i - \bar{y})^2}$$

Koefisien detrminasi adalah kuadrat dari koefisien korelasi (**r**), sehingga $R^2 = (r)^2$

Koefisien Determinasi

Nilai dari R^2 adalah 0.6483. Artinya 64.83% variasi dari peubah terikat (y) bisa dijelaskan oleh model regresi. Sisanya, yaitu 35.17% tidak mampu dijelaskan oleh model, bisa jadi karena datanya tidak mencukupi

Koefisen determinasi bukan merupakant nilai uji statistik, sehingga tidak mempunyai titik kritis yang memungkinkan kita mengambil keputusan.

Secara umum, semakin besar nilai R^2, semakin bagus modelnya.

$R^2 = 1$: Semua titik data berada dalam garis regresi.

$R^2 = 0$: Tidak terdapat hubungan linear antaya x dan y.

ANOVA

Tabel Analysis of variance (ANOVA) untuk model regresi linear sederhana diberikanoleh:

Source	degrees of freedom	Sums of Squares	Mean Squares	F-Statistic
Regression	1	SSR	MSR = SSR/1	F=MSR/MSE
Error	n–2	SSE	MSE = SSE/(n–2)	
Total	n–1	Variation in y (SST)		

		df	SS	MS	F	Significance F
10	ANOVA					
12	Regression	1	19.26	19.26	180.64	0.00
13	Residual	98	10.45	0.11		
14	Total	99	29.70			

Menggunakan Persamaan Regresi

Persamaan regresi:

$\hat{y} = 17.250 - .0669x$

Bisa digunakan untuk meramal harga mobil dengan x = 40:

$y = 17.250 - .0669x = 17.250 - .0669(40) = 14, 574$

Maka perkiraan harga mobil adalah ($14,574) .

Diagnosa Regresi

Tiga syarat (kondisi) yang harus dipenuhi untuk menggunakan analisa regresi yaitu:

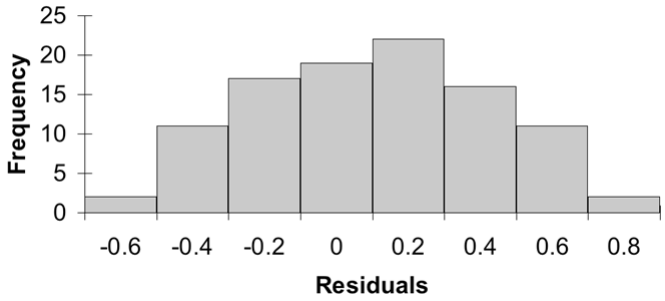
- Peubah error harus berdistribusi normal,
- Peubah error harus mempunyai varians yang konstan, &
- Errors harus saling bebas datu sama lain.

Untuk melakukan diagnosa kondisi diatas, maka harus dilakukan

➔ **analisa residual** , yaitu melihat perbedaan antara nilai data sebenarnya dengan hasil perkiraan persamaan regresi

Nonnormalitas

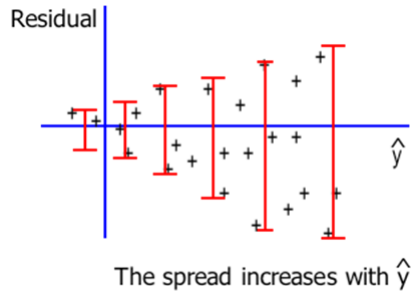
Tabulasi residual dan buat histogram mereka



Jika histogram berbentuk lonceng dengan mean disekitar nol (0), maka bisa dikatakan bahwa residual mengikuti distribusi normal. ✓

Heteroscedastisitas

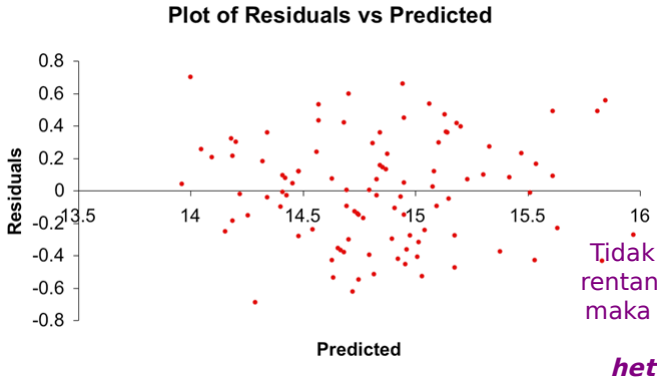
Ketika syarat (kondisi) varians yang konstan tidak terpenuhi, kondisi tersebut disebut dengan ***heteroscedastisitas***.



Heteroscedastisitas bisa dilihat dengan cara mem-plot residual terhadap nilai perkiraan y.

Heteroscedastisitas

Jika varians dari peubah error (σ_ε^2) tidak konstan, maka terdapat "***heteroscedastisitas***". Plot daibawah adalah plot error terhadap nilai perkiraan y:



Tidak terdapat perubahan rentang pada titik-titk plot, maka bisa dikatakan tidak terdapat ***heteroscedastisitas*** ✓

Peubah error tak saling bebas

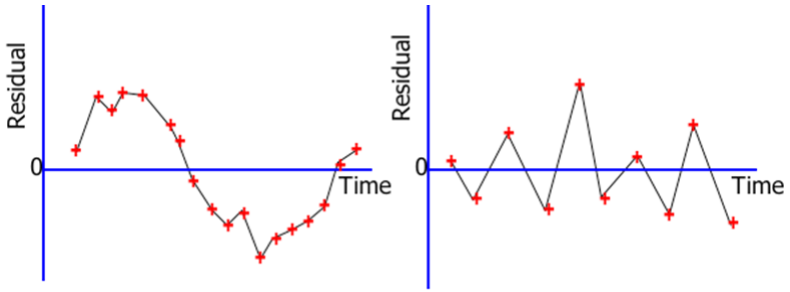
Data yang kita kumpulkan dalam bentuk tiap jam, hari, minggu akan berbentuk data deret berkala (time series).

Data yang berbentuk deret berkala, pada umumnya errornya akan saling berkorelasi. Bentuk error yang demikian dikatakan sebagai ***autokorelasi atau korelasi berseri***.

Autokorelasi bisa dilihat dengan cara menggambar residual terhadap periode waktu. Jika terdapat pola, maka syarat (kondisi) saling bebas tidak terpenuhi.

Peubah error tak saling bebas

Jika terdapat pola pada grafik residu terhadap waktu, maka terdapat autokorelasi:

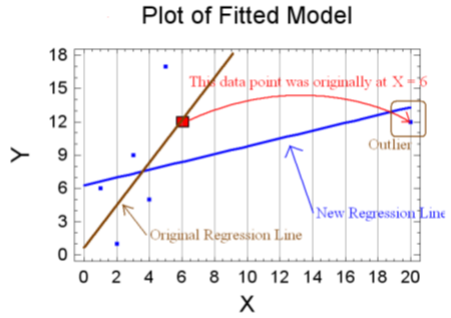


Terdapat pola karena residual positif Dan residual negatif teerpisah

Residu berkisar di nol.

Pencilan (Outlier)

Pencilan (***outlier)*** adalah observasi yang biasanya terlalu kecil atau terlalu besar.



Pencilan (Outlier)

Pencilan bisa terjadi karena:

- Terdapat kesalahan dalam mencatat data
- Titik tersebut seharusnya tidak ada dalam sampel

* Mungkin observasi tersebut memang tidak valid.

Pencilan bisa dengan mudah dilihat hari plot scatter.

Jika nilai mutlak dari residual > 2, maka kemungkinan besar titik tersebut adalah pencilan dan perlu dilihat lebih lanjut..

Pencilan harus diteliti lebih lanjut karena bida dengan mudah mempengaruhi garis least squares

Langkah Diagnosa Regresi

1. Bangun model berdasarkan teori yang telah ada.
2. Dapatkan data untuk kedua pubah yang akan dimasukkan dalam model.
3. Gambar diagram scatter untuk melihat apakah model linear sesuai. Lihat apakah terdapat pencilan (outlier).
4. Dapatkan persamaan regresi.
5. Hitung residual dan lihat apakah sudah memenuhi syarat (kondisi) model regresi
6. Perhatikan apakah model sesuai.
7. ***Jika model sesuai,*** *gunakan persamaan regresi untuk memperkirakan nilai peubah terikat*.