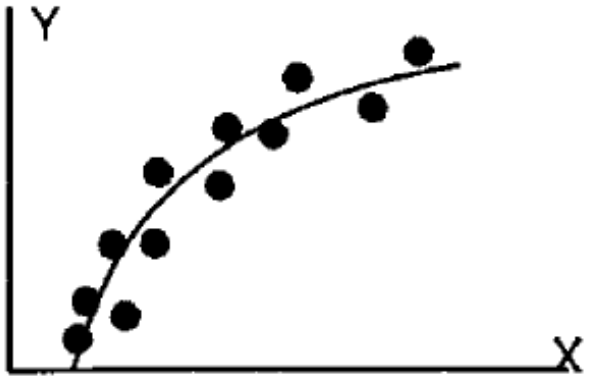
Chương 8: HỒI QUY TUYẾN TÍNH ĐƠN

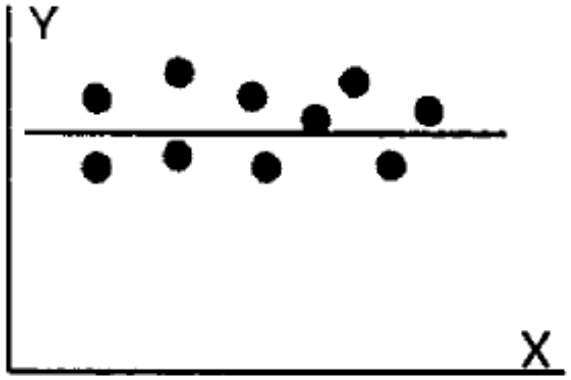
Việc phân tích hồi quy là nghiên cứu mối liên hệ phụ thuộc của một biến (gọi là biến phụ thuộc) vào một hay nhiều biến khác (gọi là các biến độc lập); với ý tưởng ước lượng giá trị trung bình (tổng thể) của biến phụ thuộc trên cơ sở biết trước giá trị các biến độc lập (qua mẫu).

Lý thuyết hồi quy đơn nghiên cứu bài toán dự báo biến ngẫu nhiên Y theo một biến ngẫu nhiên X. Biến X được gọi là biến độc lập, hay gọi là biến giải thích. Y gọi là biến phụ thuộc, hay biến được giải thích. Người ta tìm cách thay Y bởi hàm f(X) sao cho "chính xác nhất".

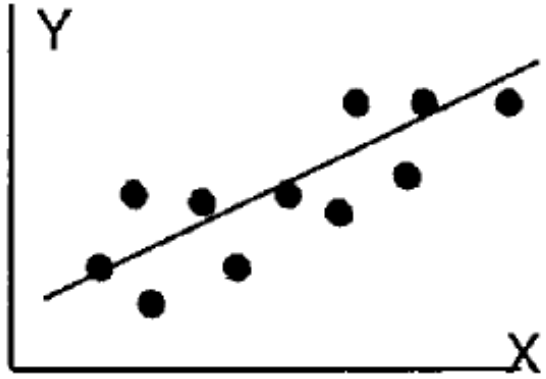
Trong mối liên hệ hàm số, với mỗi một giá trị X ta tìm được duy nhất một giá trị Y. Tuy nhiên trong bài toán hồi quy, sự phụ thuộc của Y vào X mang tính thống kê: một giá trị X có thể có tương ứng nhiều giá trị Y khác nhau, bởi vì ngoài yếu tố chính là X, biến Y có thể còn chịu tác động bởi một số yếu tố khác.



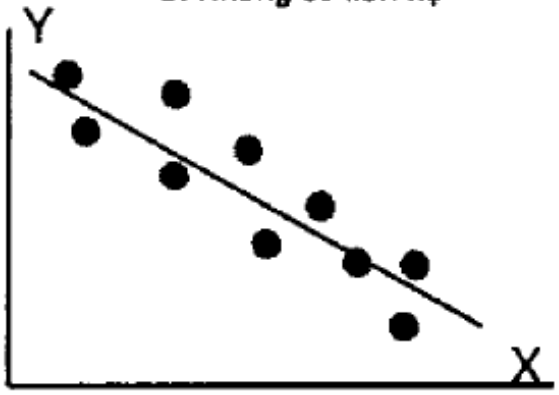
a. Liên hệ phi tuyến



b. Không có liên hệ



c. Liên hệ tuyến tính thuận



d. Liên hệ tuyến tính nghịch

Các chấm trên hình là các điểm dữ liệu phân tán, mỗi chấm là một sự kết hợp giữa Y và X cho ta một cặp giá trị cụ thể. Các đường liền nét trong hình là đường lý thuyết cho ta thấy dạng liên hệ giữa hai biến số.

Hình vẽ trích từ tài liệu tham khảo (7)

Định nghĩa:

Hàm hồi quy của Y theo X chính là kỳ vọng có điều kiện của Y đối với X, tức là E(Y|X).

Hàm hồi quy tuyến tính đơn có dạng: $f_Y(X) = E(Y|X) = \beta_0 + \beta_1 X$.

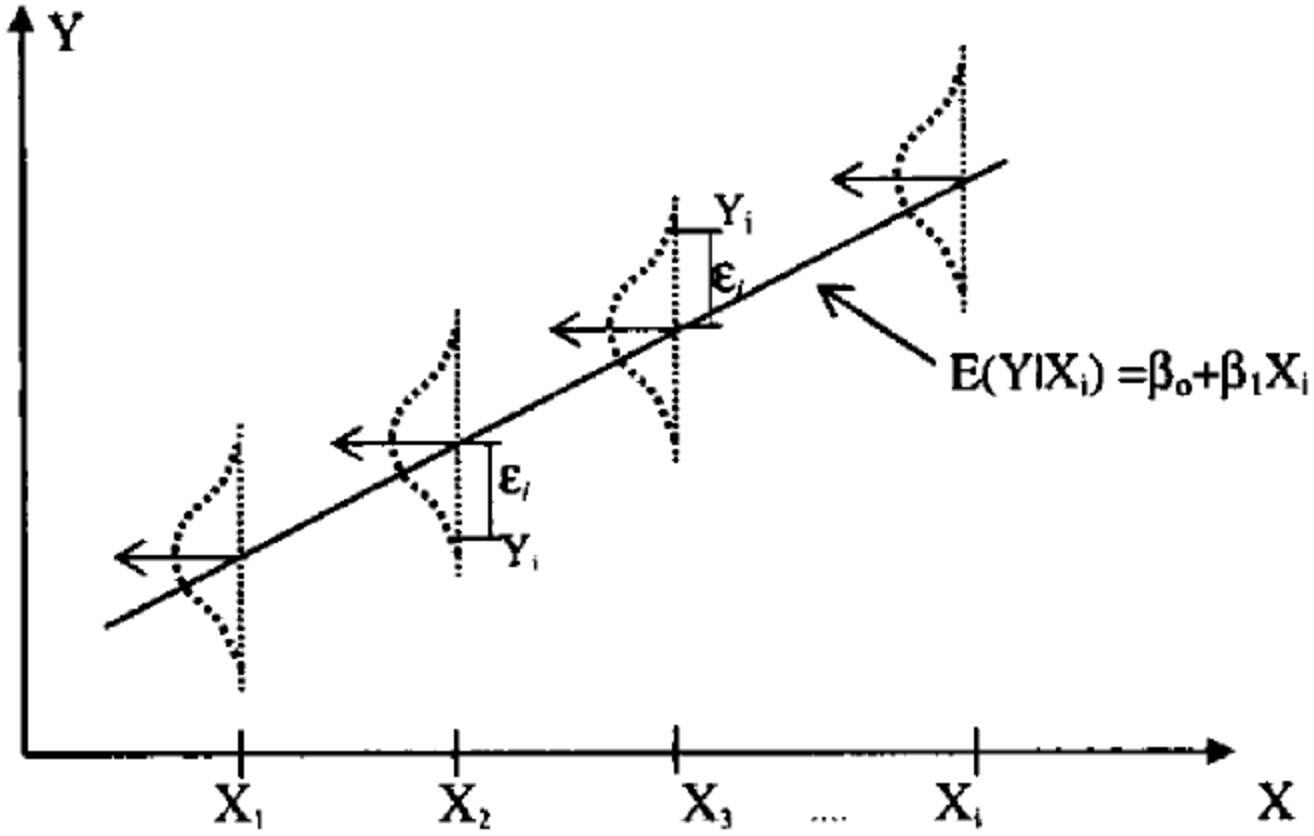
Mô hình hồi quy tuyến tính đơn:

Mô hình hồi quy tuyến tính đơn giả định có các tham số β_0 ; β_1 và σ^2 sao cho với mỗi giá trị x của biến độc lập, biến Y phụ thuộc vào x theo phương trình $Y = \beta_0 + \beta_1 x + \varepsilon$;

ở đây, biến ε là sai số ngẫu nhiên có phân phối chuẩn $N(0; \sigma^2)$.

Nếu có 1 mẫu quan sát của mô hình là: $\{ (X_1,Y_1); (X_2,Y_2);\dots (X_n,Y_n)\}$ thì ta có thể biểu diễn $Y_i = \beta_0 + \beta_1.X_i + \varepsilon_i$ *i = 1,2, … n*

với các sai số ngẫu nhiên $\{\varepsilon_i\}_i$ (nhiễu) là độc lập với nhau, tuân theo quy luật phân phối chuẩn $N(0; \sigma^2)$. (σ^2 là hằng số)



Đường thẳng hồi qui nối liền các giá trị trung bình của Y tại các giá trị khác nhau của biến độc lập X_i, ký hiệu $E(Y/X_i)$

Hình vẽ trích từ tài liệu tham khảo (7)

Từ giả định của sai số ngẫu nhiên, suy ra $Y/_{Xi} \sim N(\beta_0 + \beta_1 X_i ; \sigma^2)$

Một số yêu cầu:

1. Tìm các đặc trưng mẫu 2 chiều.

2. Tìm covarian, hệ số tương quan mẫu và ý nghĩa (chương 3).

3. Ước lượng các hệ số đường hồi quy tuyến tính (Tìm phương trình của đường hồi quy tuyến tính mẫu Y theo X và dự đoán).

4. Tìm hệ số xác định R^2

5. Ước lượng độ lệch chuẩn σ (sai số chuẩn).

6. Tìm khoảng tin cậy cho các hệ số β_0 ; β_1 của đường hồi quy tuyến tính .

7. Kiểm định sự phù hợp của của đường hồi quy tuyến tính, kiểm định các hệ số β_0 ; β_1 (BTL)

8. Tìm khoảng tin cậy cho các giá trị dự đoán của Y theo X. (BTL)

9. Kiểm tra sự phù hợp của mô hình hồi quy tuyến tính (BTL).

1. Một số đặc trưng mẫu:

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \quad \overline{(x^2)} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right)$ $\underline{\underline{k.h}}$ μ_{xx}

$$\widetilde{s_X}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \equiv \overline{(x^2)} - (\bar{x})^2; \quad s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i; \quad \overline{(y^2)} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i^2\right);$

$$\widetilde{s_y}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \equiv \overline{(y^2)} - (\bar{y})^2; \quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

- $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \quad \underline{\underline{STAT}} \quad \frac{\sum xy}{n}$

- $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$ $\underline{\underline{STAT}}$ $n \times \left(\widehat{s_x}\right)^2$

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

$\underline{\underline{STAT}}$ $\sum xy - n \times \bar{x} \times \bar{y}$

- $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$ $\underline{\underline{STAT}}$ $n \times \left(\widehat{s_y}\right)^2$

2. Covarian và hệ số tương quan của mẫu:

(xem định nghĩa và ý nghĩa ở chương 3)

- $\text{Covarian} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{S_{xy}}{n} \equiv \overline{xy} - \bar{x}.\bar{y}$
- $r_{XY} = \frac{S_{xy}}{\sqrt{S_{xx}.S_{yy}}} \equiv \frac{\overline{x.y} - \bar{x}.\bar{y}}{\tilde{s}_x . \tilde{s}_Y}$ STAT r

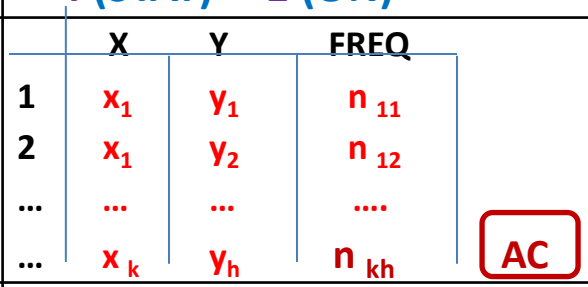
r_{XY} là một ước lượng của hệ số tương quan ρ giữa X,Y.

Khi $|r_{XY}| \leq 0.3 \Rightarrow$ X, Y không có mối quan hệ tuyến tính hoặc mối quan hệ tuyến tính rất yếu.

Khi $0.3 < |r_{XY}| \leq 0.5 \Rightarrow$ X, Y có mối quan hệ tuyến tính rất yếu.

Khi $0.5 < |r_{XY}| \leq 0.8 \Rightarrow$ X, Y có quan hệ tuyến tính trung bình.

Khi $0.8 < |r_{XY}| \Rightarrow$ X, Y có quan hệ tuyến tính mạnh.

Các bước thực hiện	Máy CASIO fx 570 ES (PLUS)...	Máy CASIO fx 580 vn....
Vào chế độ thống kê hai biến.	MODE -- 3 (STAT) -- 2 (A+BX)	MODE -- 6 - 2
Mở cột tần số (nếu máy chưa mở)	SHIFT -- MODE (SETUP) -- ▼ -- -- 4 (STAT) -- 1 (ON)	SHIFT -- MODE - ▼ - 3 - 1
Nhập dữ liệu		
Đọc kết quả $\bar{x}; \bar{y}$	SHIFT – 1 (STAT)- 4 (VAR) – --- 2 ($\bar{x}$) -- = *Tương tự ta chọn* $\bar{y}$	OPTN – 2 – (▼)
Đọc kết quả $\hat{s}_x; \hat{s}_y$	SHIFT – 1 (STAT)- 4 (VAR) – --- 3 (σ_X) -- = *Tương tự ta chọn* σ_Y	
Đọc kết quả $\overline{xy}$	SHIFT – 1 (STAT)- 3 (SUM) – -- 5 ($\sum xy$) -- ÷ -- n -- =	
Đọc kết quả R_{XY}	SHIFT – 1 (STAT)-6(REG)–3 (r) --=	OPTN – 3

Ví dụ 1: Xét bảng tương quan mẫu 2 chiều (X,Y) thu được khi người ta sơ chế một loại nông sản, ở đây X (đơn vị: phút) biểu diễn thời gian chế biến, và Y (đơn vị: %) thể hiện mức suy giảm lượng đường trong sản phẩm. Hãy tính các đặc trưng mẫu và hệ số tương quan mẫu.

X	Y				
	30	35	40	45	50
2	4				
4		7	3		
6		1	16	4	
8			2	10	3
10				4	6

Hướng dẫn nhập dữ liệu:

X	Y	Freq
2	30	4
4	35	7
4	40	3
6	35	1
6	40	16
6	45	4
8	40	2
8	45	10
8	50	3
10	45	4
10	50	6

- Các đặc trưng mẫu:

$n = 60$

$\bar{x} = 6,5667 \quad \hat{s}_x = 2,2536 \quad s_X = 2,2727$

$\bar{y} = 41,6667 \quad \hat{s}_y = 5,4518 \quad s_Y = 5,4978$

$\overline{x.y} = 284,5$

- Hệ số tương quan mẫu:

$$r_{XY} = \frac{\overline{xy} - \bar{x} \times \bar{y}}{\hat{s}_x \times \hat{s}_y} = 0,8863$$

- Hệ số đường hồi quy tuyến tính mẫu (phần sau):

$$B = \frac{\overline{xy} - \bar{x} \times \bar{y}}{\hat{s}_x^2} = 2,1440; \quad A = \bar{y} - B.\bar{x} = 27,5881$$

Ví dụ 2:

Khi theo dõi kết quả thực hành của sinh viên, người ta có được số liệu mẫu sau đây. Tìm các đặc trưng mẫu và tìm hệ số tương quan của X,Y.

Thời gian thí nghiệm *(phút)* **X**	3	4	3	5	6	4	6	5	7	8
Khối lượng sản phẩm tạo thành *(gram)* **Y**	7	8	7,3	9	10,5	8,2	10,8	9,5	11	12

$n = 10$ $\quad \overline{xy} = 50,1$

$\bar{x} = 5,1$ $\quad \hat{s}_x = 1,5780$ $\quad s_X = 1,6633$

$\bar{y} = 9,33$ $\quad \hat{s}_y = 1,6181$ $\quad s_Y = 1,7056$

(Cần ghi công thức tính:)

$r_{XY} = 0,9858$

$(B = 1,0108 \quad A = 4,1741)$

Ví dụ 3: Việc áp dụng kỹ thuật để xử lý sau thu hoạch đối với các trái thanh long thương phẩm giúp thời gian bảo quản của trái được lâu hơn. Người ta muốn tìm sự liên hệ của biến ngẫu nhiên Y là hàm lượng acid hữu cơ trong trái thanh long (đơn vị đo: %) với biến ngẫu nhiên X là thời gian bảo quản trái cây (đơn vị đo: tuần).

Một mẫu khảo sát gồm 7 trái với số liệu đo được như sau:

X (tuần)	0	1	1	2	2	3	3
Y (mg%)	0.53	0.45	0.42	0.3	0.35	0.19	0.17

a) Tìm hiệp phương sai và hệ số tương quan mẫu r_{XY} .

b) Tìm phương trình đường hồi quy tuyến tính mẫu Y theo X.

c) Hãy dự đoán hàm lượng acid hữu cơ hiện tại của một trái thanh long đã được thu hoạch từ 1.5 tuần trước.

d) Có một trái thanh long mà người ta đo được hàm lượng acid hữu cơ của nó là 0.4 %. Hãy dự đoán số tuần mà trái thanh long đó đã được bảo quản.

e) Tìm hệ số xác định R^2

f) Hãy ước lượng độ lệch chuẩn σ.

g) Tìm khoảng ước lượng 95% cho các hệ số của đường hồi quy tuyến tính biểu diễn Y theo X.

Hỗ trợ từ excel:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	**-0.98602**					
R Square	**0.97224**					
Adjusted R Square	0.966684033					
Standard Error	**0.02444**					
Observations	7					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression (SSR)	1	**0.10458**	0.104585	175.0938276	**4.4E-05**	
Residual (SSE)	5	**0.00299**	0.000597			
Total (SST)	6	**0.10757**				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept (a)	**0.54769**	0.017933968	30.53938	7.0636E-07	**0.50159**	**0.5938**
X (b)	**-0.11865**	0.008966984	-13.2323	4.40466E-05	**-0.1417**	**-0.0956**

X	Y
0	0.53
1	0.45
1	0.42
2	0.3
2	0.35
3	0.19
3	0.17

3. Ước lượng các hệ số β_0; β_1 của đường hồi quy tuyến tính:

(Tìm đường hồi quy tuyến tính mẫu của Y theo X)

Giả sử ta có 1 mẫu cụ thể $\{ (x_i, y_i) \}_{i = 1,2,\ldots,n}$

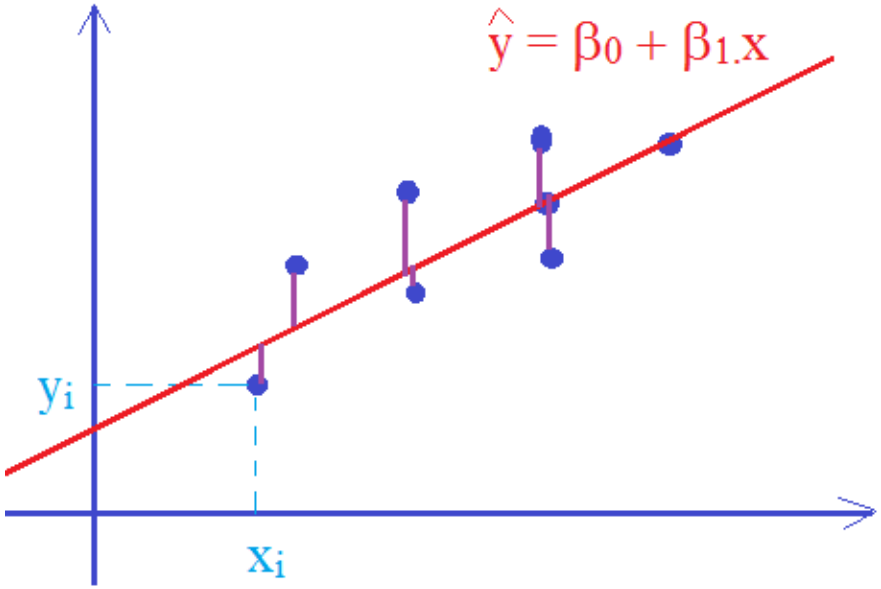
Hàm $y = a + bx$ là đường hồi quy tuyến tính mẫu của Y theo X

nếu hàm $Q = \sum_{i=1}^{n} \left(y_i - y_i \right)^2$ là nhỏ nhất.

(Phương pháp tổng bình phương bé nhất - OLS)

Tìm giá trị nhỏ nhất của hàm 2 biến (cực trị tự do):

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 . x_i) \right]^2$$



ta tìm được giá trị a, b là nghiệm.

Giá trị a là một ước lượng cho hệ số tự do β_0.

Giá trị b là một ước lượng cho hệ số số góc β_1.

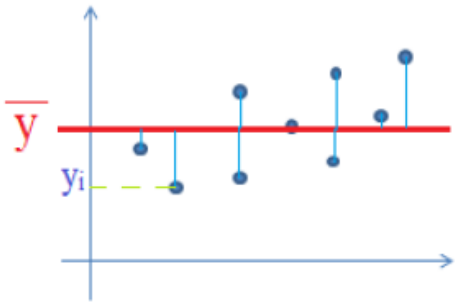
$$\begin{cases} b = \dfrac{S_{xy}}{S_{xx}} = \dfrac{\overline{xy} - \overline{x}.\overline{y}}{\widetilde{s_x}^2} & \underline{\underline{STAT}} \quad \boxed{B} \\ a = \overline{y} - b.\overline{x} & \underline{\underline{STAT}} \quad \boxed{A} \end{cases}$$

Phương trình hồi quy tìm được có thể dùng để nội suy giá trị $E(Y|X=x_0)$. Công thức dự đoán:

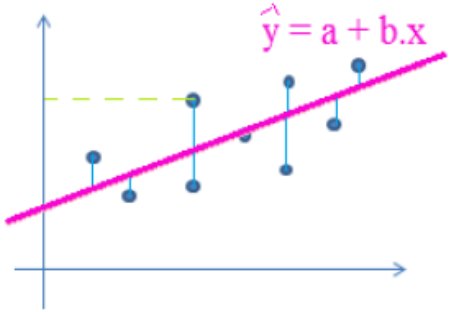
$$\widehat{y_0} = a + b\,x_0 \quad \underline{\underline{STAT}}\ x_0\ \boxed{\widehat{y}}\ \boxed{=}$$

4. Các thông số khác của đường hồi quy. Hệ số xác định R^2:

- $SST = \sum_{i=1}^{n} \left(y_i - \overline{y} \right)^2$ *Sum of Squares Total*



SST đo mức biến động các giá trị quan sát y_i xung quanh giá trị trung bình của chính mẫu. SST được tạo bởi 2 thành phần: SST = SSE + SSR.



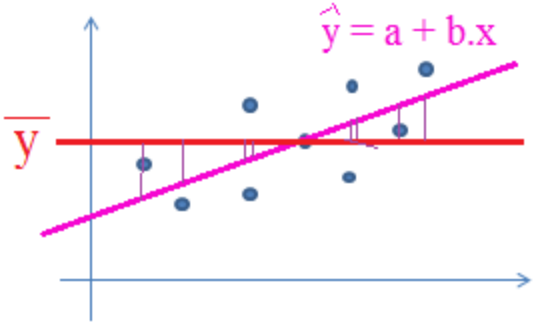
- $SSE = \sum_{i=1}^{n} \left(y_i - \widehat{y_i} \right)^2$

Sum of Squares for Error | *Sum of squares Residual*

TBP sai số ước lượng đo sự chênh lệch giữa từng giá trị quan sát với giá trị dự đoán (ước lượng). SSE được xem như sai số do những yếu tố khác ngoài X hoặc do lấy mẫu ngẫu nhiên.

- $SSR = \sum_{i=1}^{n} \left(\widehat{y_i} - \overline{y} \right)^2$ *Sum of Squares in Regression*

SSR là sai số do khác biệt giữa đường hồi quy mẫu và trung bình của Y. Sự khác biệt này được giải thích bởi sự biến động của X. SSR đo sự phân tán của dữ liệu do mô hình hồi quy gây ra. SSR càng gần tới SST thì mô hình càng phù hợp.



- $SSR = \sum_{i=1}^{n} \left(\widehat{y_i} - \overline{y} \right)^2 = \sum_{i=1}^{n} \left(a + bx_i - \overline{y} \right)^2 = \sum_{i=1}^{n} \left(\cancel{\overline{y}} - b.\overline{x} + b\,x_i - \cancel{\overline{y}} \right)^2$

$$\underline{\underline{STAT}} \quad \frac{n(\overline{x.y} - \overline{x}.\overline{y})^2}{\widehat{s}_X^{\,2}}$$

- $SSE = \sum_{i=1}^{n} \left(y_i - \widehat{y_i} \right)^2 = S_{yy} - b.S_{xy}$ = SST - SSR
- $SST = \sum_{i=1}^{n} \left(y_i - \overline{y} \right)^2 = S_{yy}$ $\underline{\underline{STAT}}$ $n \times \widehat{s}_Y^{\,2}$

Hệ số xác định R^2:

$$R^2 = \frac{SSR}{SST} \times 100\% \qquad hay \quad R^2 = \left(1 - \frac{SSE}{SST}\right) \times 100\%$$

Hệ số R^2 giải thích trong 100% sự biến động của Y so với trung bình của nó thì có bao nhiêu % là do biến X gây ra.

Trong mô hình hồi quy tuyến tính đơn, $R^2 = r_{XY}^2$.

(r_{XY}: hệ số tương quan)

5. Ước lượng độ lệch chuẩn σ (sai số chuẩn của ước lượng):

σ^2 có ước lượng không chệch của nó là $\widetilde{\sigma}^2$

$$\widetilde{\sigma}^2 = \frac{SSE}{n-2} \quad \Rightarrow \quad \widetilde{\sigma} = \sqrt{\frac{SSE}{n-2}}$$

6. Ước lượng các hệ số hồi quy với độ tin cậy 1- α:

- Khoảng ước lượng cho tung độ gốc β_0 là $(a - \varepsilon_0 ; a + \varepsilon_0)$ với:

$$\varepsilon_0 = t_{\frac{\alpha}{2}(n-2)} \times \sqrt{\frac{\overline{(x^2)}}{S_{xx}} \times \frac{SSE}{n-2}} = t_{\frac{\alpha}{2}(n-2)} \times \frac{1}{\widehat{S_x}} \sqrt{\frac{SSE \times \overline{(x^2)}}{n(n-2)}}$$

- Khoảng ước lượng cho hệ số góc β_1 là $(b - \varepsilon_1 ; b + \varepsilon_1)$ với:

$$\varepsilon_1 = t_{\frac{\alpha}{2}(n-2)} \times \sqrt{\frac{SSE}{(n-2) \times S_{xx}}} = t_{\frac{\alpha}{2}(n-2)} \times \frac{1}{\widehat{S_x}} \sqrt{\frac{SSE}{n(n-2)}}$$

7. Kiểm định sự phù hợp của đường hồi quy tuyến tính:

ANOVA

	SS	df	MS	Tiêu chuẩn kiểm định F	Giá trị P
Regression *(Hồi quy)*	**SSR**	1	MSR = SSR	$F=\frac{MSR}{MSE}$	
Residual *(Phần dư)*	**SSE**	n-2	$MSE=\frac{SSE}{n-2}$		
Total	SST	n-1			

Kiểm định sự phù hợp của hàm hồi quy tuyến tính đơn:

Giả thiết $H_0: R^2 = 0$ hoặc $H_0: \beta_1 = 0$

$H_1: R^2 \neq 0$ $H_1: \beta_1 \neq 0$ (mô hình phù hợp)

Tiêu chuẩn kiểm định: $F=\frac{R^2}{\frac{1-R^2}{n-2}}$ hoặc $F=\frac{SSR}{\frac{SSE}{n-2}}$

Miền bác bỏ: $W_\alpha = (f_\alpha(1;n-2) ; +\infty)$

Kiểm định sự phù hợp của các hệ số đường hồi quy tuyến tính:

* Giả thiết $H_0: \beta_1 = b_o$

$H_1: \beta_1 \neq b_o$

Tiêu chuẩn kiểm định: $$T = \frac{b - b_0}{\frac{\sqrt{SSE}}{\widetilde{S_x} \times \sqrt{n(n-2)}}}$$

Miền bác bỏ: $W_\alpha = (-\infty; -t_{\frac{\alpha}{2}}(n-2)) \cup (t_{\frac{\alpha}{2}}(n-2); +\infty)$

---

* Giả thiết $H_0: \beta_0 = a_o$

$H_1: \beta_0 \neq a_o$

Tiêu chuẩn kiểm định: $$T = \frac{a - a_0}{\frac{\sqrt{SSE \times \overline{x^2}}}{\widetilde{S_x} \times \sqrt{n(n-2)}}}$$

Miền bác bỏ: $W_\alpha = (-\infty; -t_{\frac{\alpha}{2}}(n-2)) \cup (t_{\frac{\alpha}{2}}(n-2); +\infty)$

8. Dự báo giá trị trung bình của Y khi $X = x_0$.

(Khoảng ước lượng của $f_Y(x_o)$)

$$a + b \times x_o \pm t_{\frac{\alpha}{2}(n-2)} \times \hat{\sigma} \times \sqrt{\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{XX}}}$$