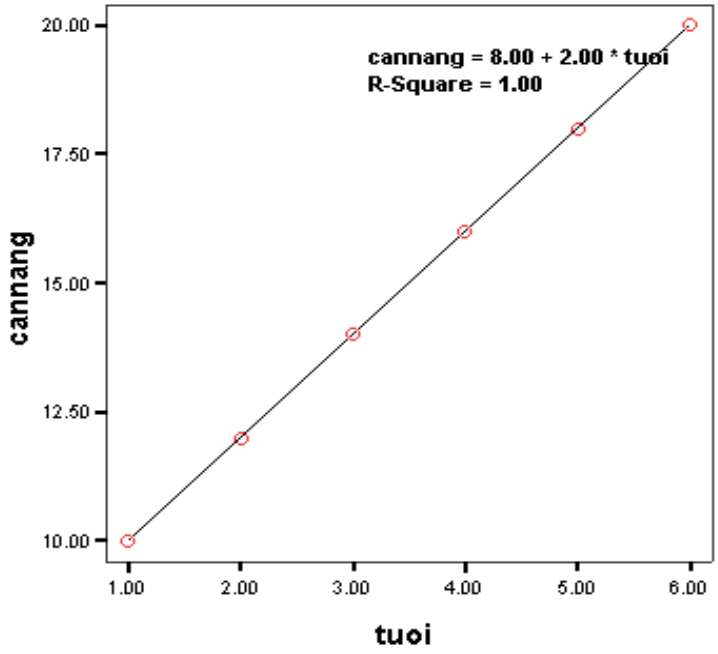
PHÂN TÍCH HỒI QUI TUYẾN TÍNH ĐƠN GIẢN

17.1 Phương trình hồi qui tuyến tính

Phân tích hồi qui tuyến tích đơn giản (Simple Linear Regression Analysis) là tìm sự liên hệ giữa 2 biến số liên tục: biến độc lập (biến dự đoán) trên trục hoành x với biến phụ thuộc (biến kết cục) trên trục tung y. Sau đó vẽ một đường thẳng hồi qui và từ phương trình đường thẳng này ta có thể dự đoán được biến y (ví dụ: cân nặng) khi đã có x (ví dụ: tuổi)

Ví dụ 1: Ta có 1 mẫu gồm 6 trẻ từ 1-6 tuổi, có cân nặng như bảng sau:

Tuổi	Cân nặng (kg)
1	10
2	12
3	14
4	16
5	18
6	20



Nối các cặp (x,y) này ta thấy có dạng 1 phương trình bậc nhất: **y=2x+8**

(trong đó 2 là độ dốc và 8 là điểm cắt trên trục tung y khi x=0). Trong thống kê phương trình đường thẳng (bậc nhất) này được viết dưới dạng:

$y= \beta x + \alpha$ [1]

Đây là phương trình hồi qui tuyến tính, trong đó β gọi là độ dốc (slope) và α là chặn (intercept), điểm cắt trên trục tung khi x=0.

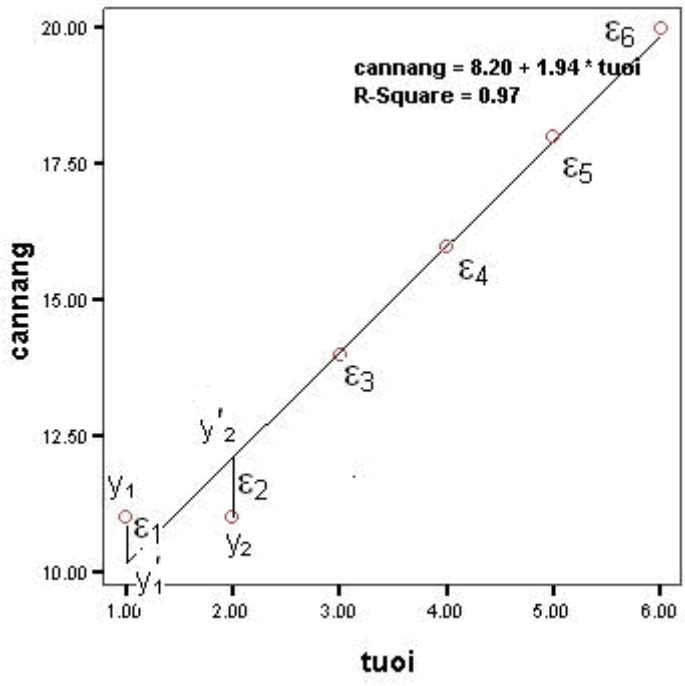
Thực ra phương trình hồi qui tuyến tính này chỉ có trên lý thuyết, nghĩa là các trị số của x_i (i=1,2,3,4,5,6) và y_i tương ứng, liên hệ với nhau 100% (hoặc hệ số tương quan R=1)

Trong thực tế hiếm khi có sự liên hệ 100% này mà thường có sự sai lệch giữa trị số quan sát y_i và trị số y_i' ước đoán nằm trên đường hồi qui.

17.1.1 Mô hình hồi qui tuyến tính

Ví dụ 2: Ta có 1 mẫu gồm 6 trẻ em khác có cân nặng theo bảng sau:

Tuổi	Cân nặng (kg)
1	11
2	11
3	14
4	16
5	18
6	20



Khi vẽ đường thẳng hồi qui, ta thấy các trị số quan sát y_3, y_4, y_5, y_6 nằm trên đường thẳng, còn y_1 và y_2 không nằm trên đường thẳng này và sự liên hệ giữa x_i và y_i

không còn là 100% mà chỉ còn 97% vì có sự sai lệch tại y_1 và y_2. Sự sai lệch này trong thống kê gọi là phần dư (residual) hoặc errors.

Gọi $y_1, y_2, y_3, y_4, y_5, y_6$ là trị số quan sát và $y'_1, y'_2, y'_3, y'_4, y'_5, y'_6$ là trị số ước đoán nằm trên đường hồi qui, $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5, \varepsilon_6$ là phần dư.

Như vậy $\varepsilon_1 = y_1 - y'_1$

$\varepsilon_2 = y_2 - y'_2$

$\varepsilon_3 = y_3 - y'_3$

$\varepsilon_4 = y_4 - y'_4$

$\varepsilon_5 = y_5 - y'_5$

$\varepsilon_6 = y_6 - y'_6$

Khi đó phương trình hồi qui tuyến tính được viết dưới dạng tổng quát như sau:

$y' = \beta x_i + \alpha_i + \varepsilon_i$ [2]

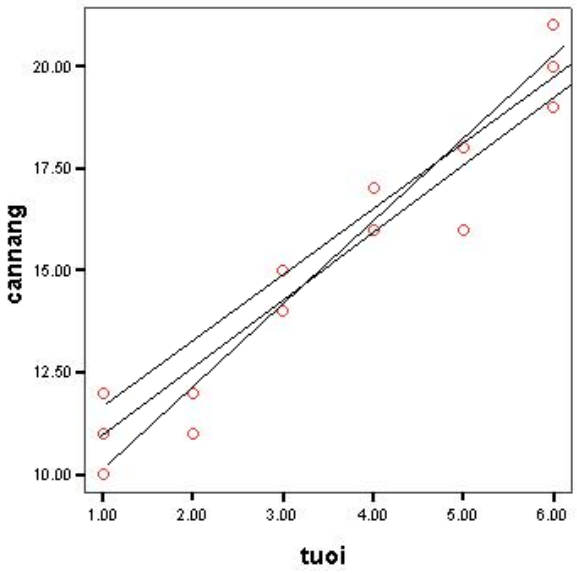
Như vậy nếu phần dư ε_i càng nhỏ sự liên hệ giữa x,y càng lớn và ngược lại. Phần liên hệ còn đượi gọi là phần hồi qui. Mô hình hồi qui tuyến tích được mô tả như sau:

Dữ liệu= Hồi qui (Regression) + Phần dư (Residual)

17.1.2 Ước tính hệ số tương quan β và chặn α

Muốn vẽ được phương trình hồi qui tuyến tính cần phải ước tính được độ dốc β và chặn α trên trục tung.

Ví dụ 3: Nếu chúng ta chọn một mẫu thực tế gồm 30 em từ 1-6 tuổi và kết quả cân nặng tương ứng của 30 em được vẽ trong biểu đồ sau:



Lúc này ta không thể nối 30 điểm trên biểu đồ mà phải vẽ 1 đường thẳng đi càng gần với tất cả các điểm càng tốt. Như vậy 3 đường thẳng ở biểu đồ ta chọn đường thẳng nào?. Nguyên tắc chọn đường thẳng nào đi gần cả 30 điểm, có nghĩa làm sao để tổng các phần dư $\Sigma\varepsilon_i$ nhỏ nhất:

$\Sigma\ \varepsilon_i = \Sigma\ (y_i - \beta x - \alpha)$

và tổng bình phương của phần dư:

$\Sigma\ (\varepsilon_i)^2 = \Sigma\ (y_i - \beta x - \alpha)^2$

Đây là phương trình bậc 2 theo x. Trong toán học, muốn tìm trị cực tiểu của 1 phương trình bậc 2, người ta lấy đạo hàm và cho đạo hàm triệt tiêu (bằng 0) sẽ tìm được trị cực tiểu của x. Giải phương trình này, ta sẽ tính được 2 thông số β và α và từ 2 thông số này ta sẽ vẽ được đường thẳng hồi qui. Phương pháp này trong toán học gọi là **phương pháp bình phương nhỏ nhất** (least square method).

Giải phương trình trên ta có:

$$\beta = r\frac{Sy}{Sx}$$

(r là hệ số tương quan; Sy là độ lệch chuẩn của y và Sx là độ lệch chuẩn của x)

$$r = \frac{1}{n-1}\Sigma\left(\frac{xi - \overline{x}}{Sx}\right)\left(\frac{yi - \overline{y}}{Sy}\right)$$

$\alpha = y - \beta x$

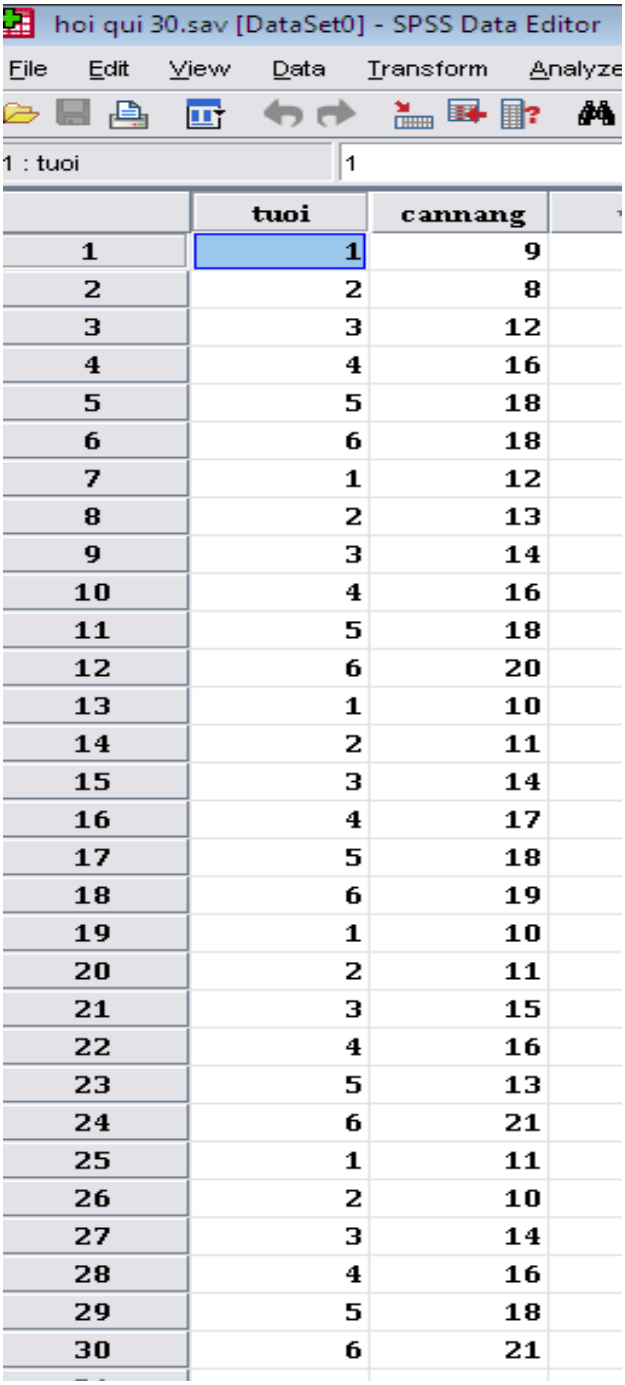
và phương trình hồi qui tuyến tính của y theo x (bình phương nhỏ nhất) là:

$y' = \beta x_i + \alpha$

17.2 Phân tích hồi qui tuyến tính trong SPSS

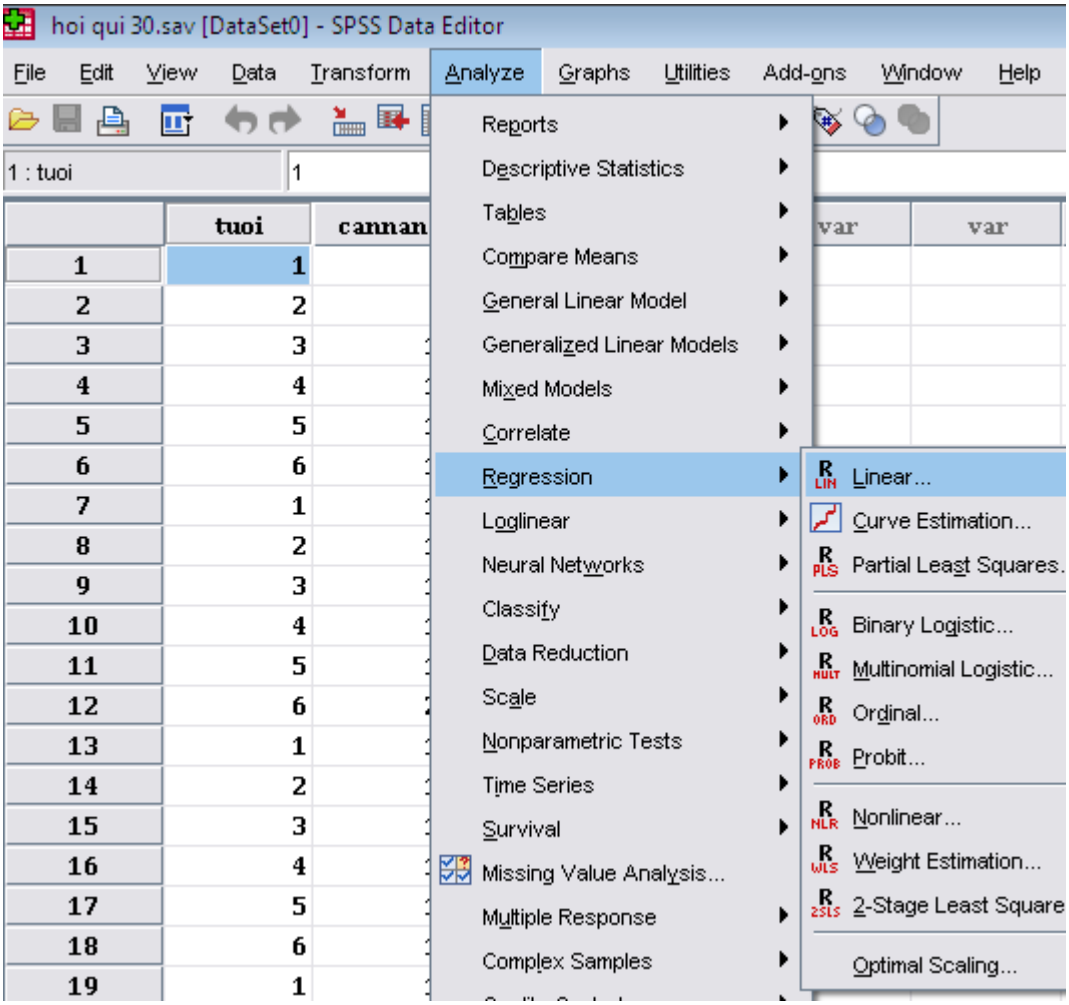
Nhập số liệu tuổi và cân nặng cân được của 30 trẻ 1-6 tuổi vào SPSS:

Cột 1: tuổi; cột 2: cân nặng



	tuoi	cannang
1	1	9
2	2	8
3	3	12
4	4	16
5	5	18
6	6	18
7	1	12
8	2	13
9	3	14
10	4	16
11	5	18
12	6	20
13	1	10
14	2	11
15	3	14
16	4	17
17	5	18
18	6	19
19	1	10
20	2	11
21	3	15
22	4	16
23	5	13
24	6	21
25	1	11
26	2	10
27	3	14
28	4	16
29	5	18
30	6	21

Vào menu: **>Analyze> Regression> Linear**



Bảng 17.1 Tóm tắt mô hình

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.918[a]	.843	.837	1.49794

a. Predictors: (Constant), tuoi

Hệ số tương quan R=0,918 và R^2=0,843

Bảng 17. 2 Phân tích ANOVA với biến phụ thuộc là cân nặng

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	336.140	1	336.140	149.808	.000[a]
	Residual	62.827	28	2.244		
	Total	398.967	29			

a. Predictors: (Constant), tuoi

b. Dependent Variable: cannang

Tổng bình phương phần hồi qui (Regression)=336,14

Tổng bình phương phần dư (Residual)=62,8

Trung bình bình phương hồi qui: 336,14/ 1 (bậc tự do)=336,14

Trung bình bình phương phần dư: 62,8/ 28(bậc tự do=n-2)=2,24

$$\mathbf{F=\frac{336,14}{2,24}=149,8 \text{ và } p<0,000}$$

Bảng 17.3 Hệ số tương quan β và chặn α

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	7.773	.624		12.464	.000
	tuoi	1.960	.160	.918	12.240	.000

a. Dependent Variable: cannang

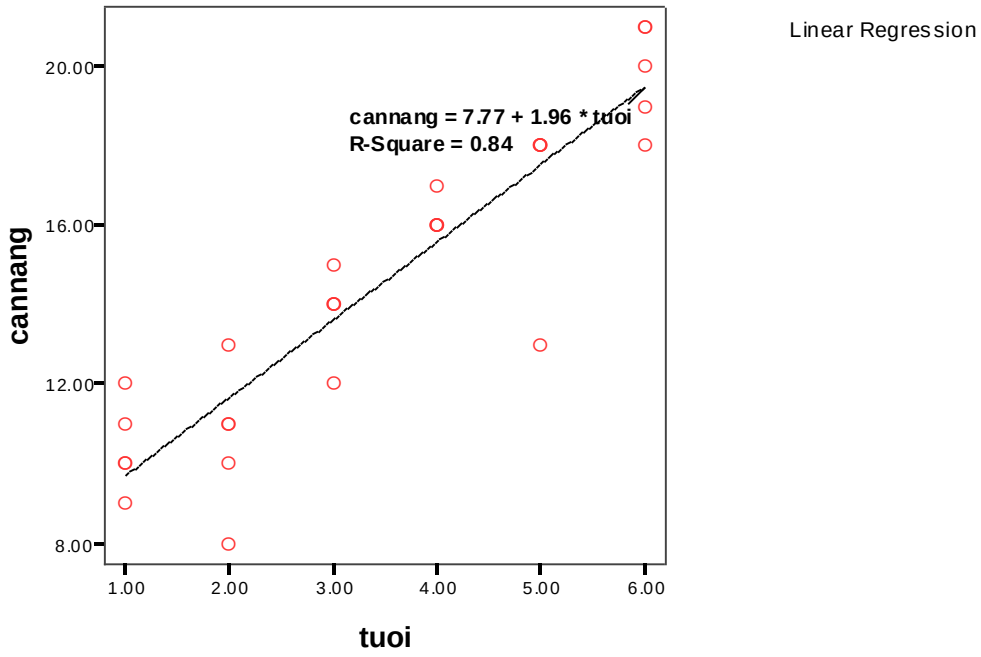
Kết quả bảng 3 cho biết hệ số tương quan β (độ dốc) = 1,96 và điểm cắt tại trung tung là α=7.773

Phương trình đường thẳng hồi qui được viết:

Cân nặng= 7,77 + 1,96 x tuổi

Như vậy khi em bé tăng lên 1 tuổi thì cân nặng tăng lên 1,96 kg

Vẽ đường thẳng hồi qui trong SPSS



Từ phương trình này ta có thể ước đoán được cân nặng theo tuổi của trẻ, tuy nhiên nằm trong một giới hạn nào đó chẳng hạn như từ 1-12 tuổi, vì sau tuổi này cân nặng trẻ sẽ tăng vọt trong thời kỳ dậy thì và không còn liên hệ tuyến tính với tuổi nữa.
Ví dụ muốn ước đoán cân nặng của trẻ từ quần thể nghiên cứu này:
7 tuổi ⇨ Cân nặng= 7,77 + 1,96 x7 = 21,49 kg
8 tuổi ⇨ Cân nặng= 7,77 + 1,96 x8 = 23,45 kg

17. 3 Các giả định trong phân tích hồi qui tuyến tính

Phân tích hồi qui tuyến tính không chỉ là việc mô tả các dữ liệu quan sát được trong mẫu (sample) nghiên cứu mà cần phải suy rộng cho mối liên hệ trong dân số (population). Vì vậy, trước khi trình bày và diễn dịch mô hình hồi qui tuyến tính cần phải dò tìm vi phạm các giả định. Nếu các giả định bị vi phạm thì các kết quả ước lượng không đáng tin cậy được.

Các giả định cần thiết trong hồi qui tuyến tính:

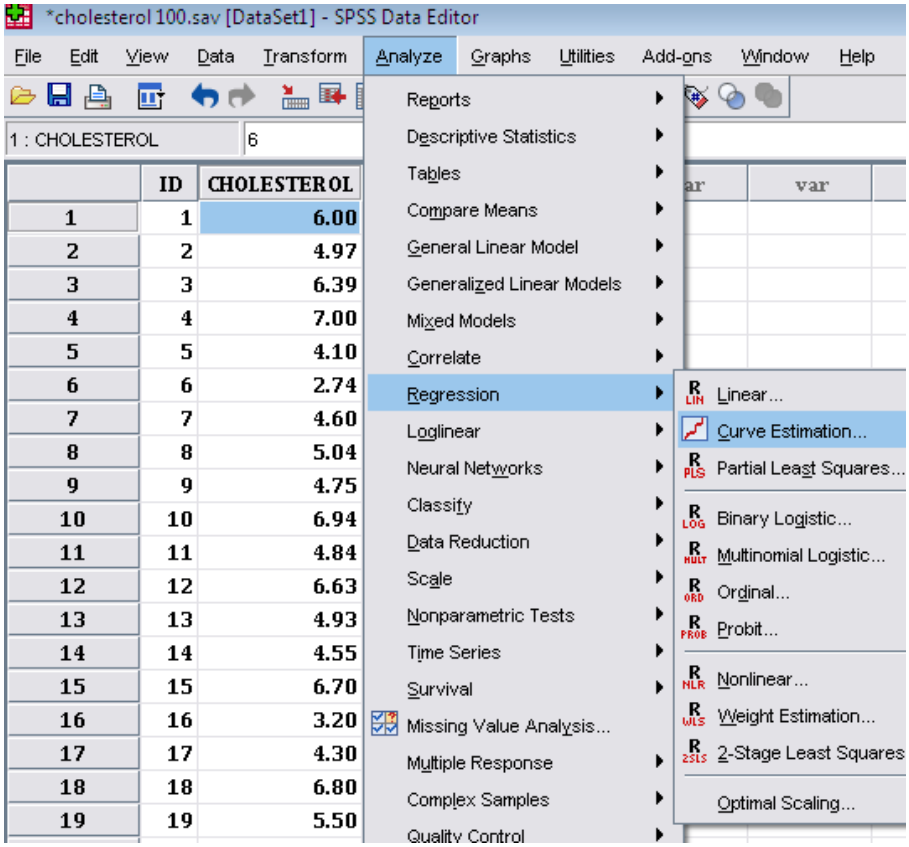
1. x_i là biến số cố định, không có sai sót ngẫu nhiên trong đo lường.
2. Phần dư (trị số quan sát trừ cho trị số ước đoán) phân phối theo luật phân phối chuẩn
3. Phần dư có trị trung bình bằng 0 và phương sai không thay đổi cho mọi trị x_i
4. Không có tương quan giữa các phần dư

Ví dụ: Một nghiên cứu tìm sự tương quan giữa cholesterol máu với bề dày lớp nội trung mạc (NTM) của động mạch cảnh đo được trên siêu âm với dữ liệu ghi nhận ở 100 BN như sau:

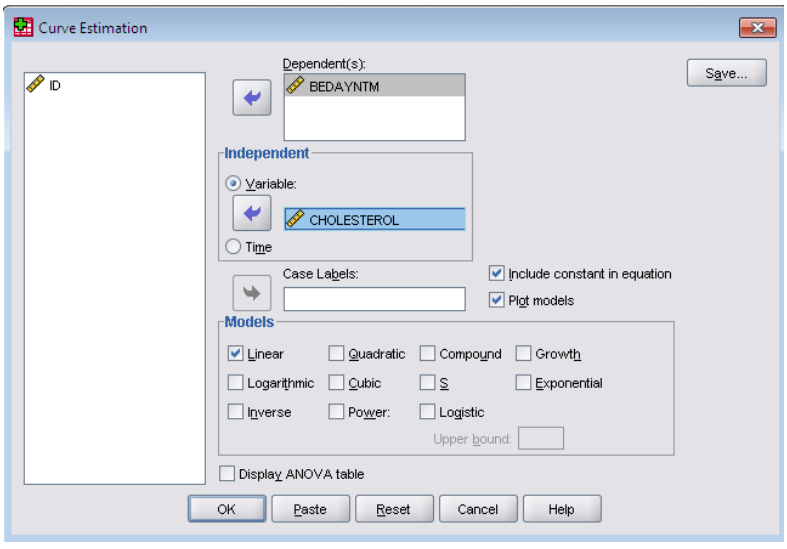
ID	CHOLESTEROL	BEDAYNTM	ID	CHOLESTEROL	BEDAYNTM	ID	CHOLESTEROL	BEDAYNTM
1	6.00	1.95	35	5.84	0.88	69	7.00	0.82
2	4.97	1.33	36	6.91	0.97	70	6.60	1.00
3	6.39	0.83	37	5.01	0.90	71	5.75	1.70
4	7.00	2.00	38	4.00	0.89	72	4.70	2.30
5	4.10	1.30	39	4.88	0.80	73	4.61	0.89
6	2.74	1.16	40	4.20	1.13	74	6.30	0.97
7	4.60	1.00	41	4.47	1.20	75	2.00	0.70
8	5.04	1.00	42	6.90	0.90	76	2.50	1.10
9	4.75	0.80	43	4.71	0.81	77	4.79	1.01
10	6.94	1.60	44	5.70	0.80	78	5.31	1.15
11	4.84	0.65	45	3.00	0.74	79	3.80	0.92
12	6.63	1.00	46	5.06	2.66	80	7.13	1.10
13	4.93	0.97	47	4.61	0.89	81	5.50	0.80
14	4.55	0.73	48	4.15	0.79	82	4.20	0.70
15	6.70	1.10	49	5.30	0.80	83	3.30	1.00
16	3.20	1.10	50	4.10	0.56	84	5.90	0.80
17	4.30	1.10	51	3.00	0.80	85	4.73	0.89
18	6.80	0.80	52	2.57	1.20	86	3.00	0.60
19	5.50	0.99	53	6.78	0.82	87	5.88	1.50
20	6.80	1.00	54	5.62	0.90	88	5.39	0.70
21	5.74	1.13	55	8.07	1.00	89	6.15	1.10
22	6.90	1.00	56	3.00	1.15	90	3.94	0.81
23	7.00	1.70	57	3.31	1.16	91	3.83	0.70
24	3.40	0.90	58	4.73	0.97	92	4.93	0.71
25	4.92	0.89	59	4.00	0.80	93	7.00	2.70
26	6.08	0.80	60	3.60	1.67	94	8.18	1.13
27	6.25	0.81	61	5.30	1.06	95	8.16	1.70
28	5.40	1.20	62	6.00	1.10	96	7.20	0.90
29	6.54	0.82	63	6.49	0.80	97	5.24	1.16
30	3.91	0.89	64	7.00	1.70	98	4.40	1.00
31	5.30	1.19	65	7.48	0.99	99	5.20	0.97
32	2.60	0.97	66	5.19	1.16	100	5.20	2.30
33	6.85	0.97	67	3.00	0.62			
34	3.75	0.97	68	6.70	1.00			

Biểu đồ phân tán (scatter) là một phương tiện tốt để đánh giá mức độ đường thẳng phù hợp với dữ liệu quan sát.

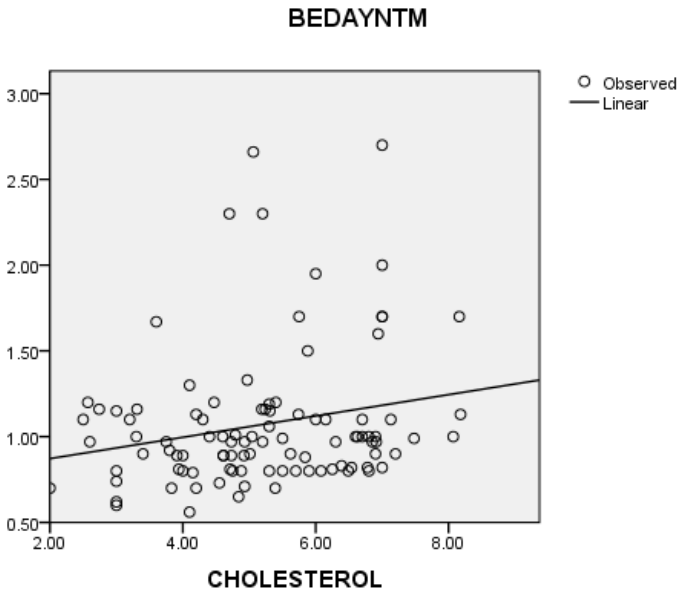
Vào menu: **Analyze> Curve Estimation**



Vào màn hình Curve Estimation



Nhắp chuyển BEDAYNTM (Bề dày nội trung mạc) vào ô Dependent (s) và CHOLESTEROL vào ô Variable. Đánh dấu nháy vào các ô Include constant in equation, ô Plot models và ô Linear (nếu muốn ước lượng sự liên hệ giữa 2 biến theo dạng phương trình bậc 2 thì đánh thêm dấu nháy vào ô Quadratic). Nhấn OK, ta có biểu đồ sau:

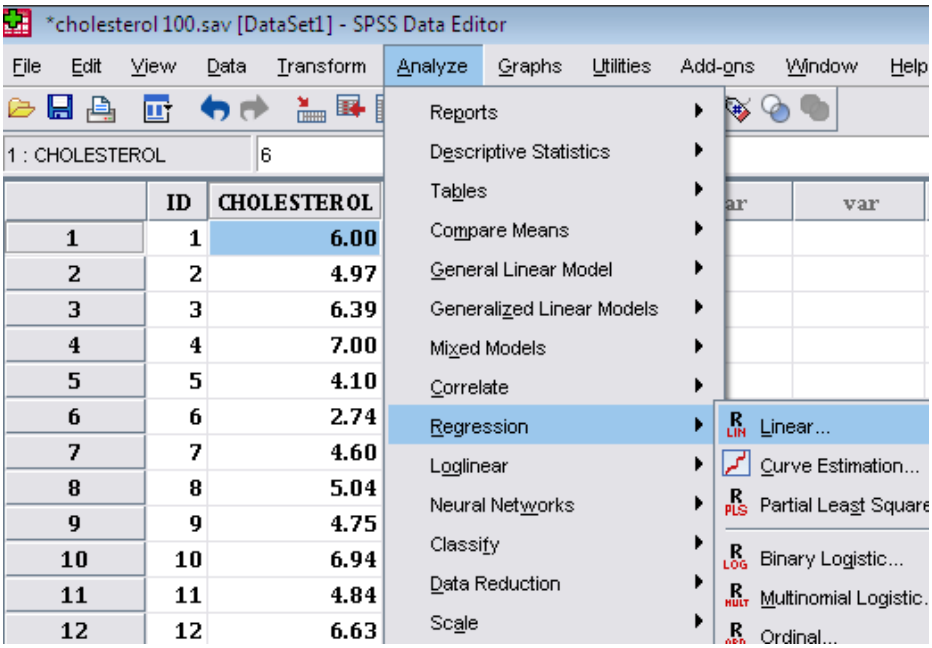


Đây là phương trình hồi qui tuyến tính với y= 0,748 + 0,062x

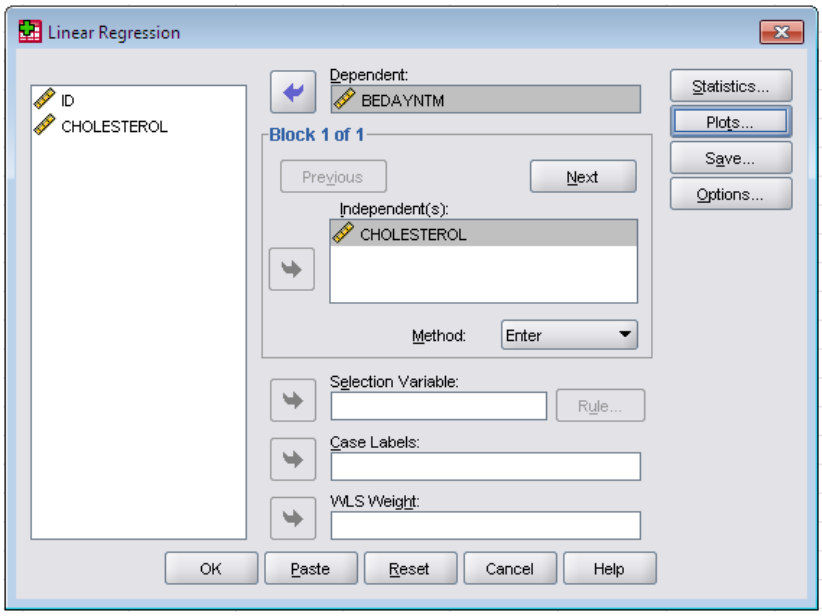
Giả định x (cholesterol máu) là một biến cố định, không có sai sót trong đo lường. Giả định này không có vấn đề nếu bệnh nhân được đo ở một phòng thí nghiệm chuẩn.

Các giả định còn lại thực hiện trong SPSS như sau:

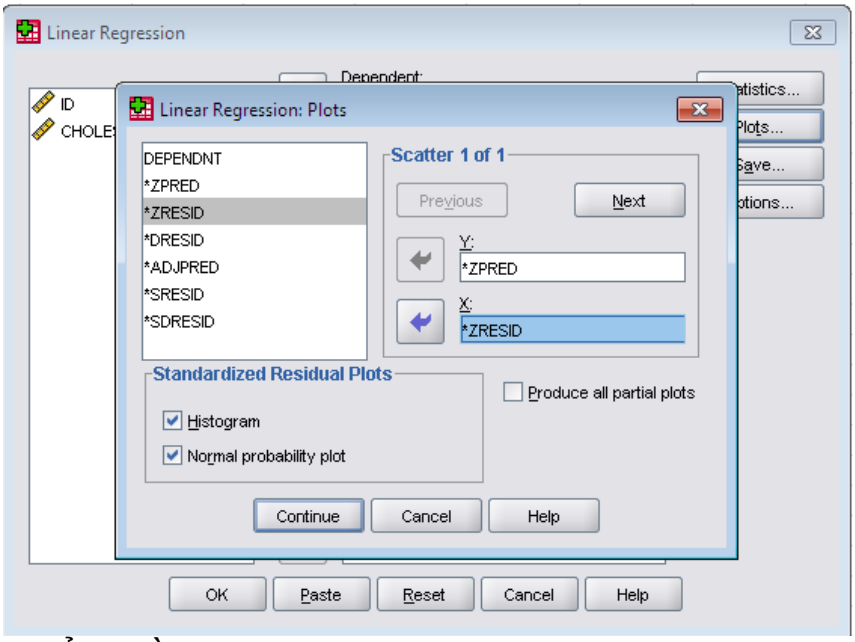
Vào menu: **Analyze> Regression> Linear...**



Vào màn hình Linear, Nhấp chuyển BEDAYNTM qua ô Dependent và CHOLESTEROL qua ô Independent(s)



Nhấn nút Plots, mở hộp thoại Plots:



Nhấp chuyển phần dư *ZRESID vào ô X (trục hoành) và giá trị dự đoán vào ô Y (trục tung) để xem phân dư có phân bố ngẫu nhiên và phương sai có cố định cho mọi trị của x_i. Nhấn dấu nháy vào ô Histogram và ô Normal probability plot để xem phần dư có phân phối chuẩn.

Nhấn Continue, sau đó nhấn OK cho kết quả sau:

Model Summary[a]

a. Dependent Variable: BEDAYNTM

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.748	.151		4.954	.000
	CHOLESTEROL	.062	.028	.219	2.218	.029

a. Dependent Variable: BEDAYNTM

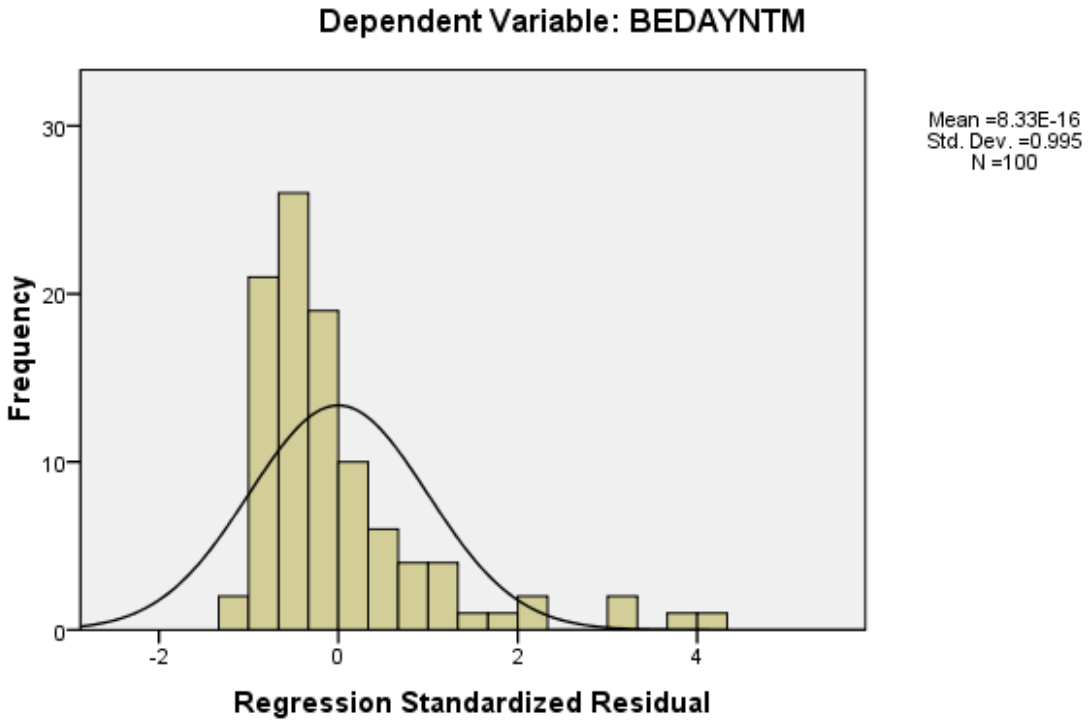
Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	.8722	1.2562	1.0710	.08828	100
Residual	-.44269	1.59765	.00000	.39405	100
Std. Predicted Value	-2.252	2.098	.000	1.000	100
Std. Residual	-1.118	4.034	.000	.995	100

a. Dependent Variable: BEDAYNTM

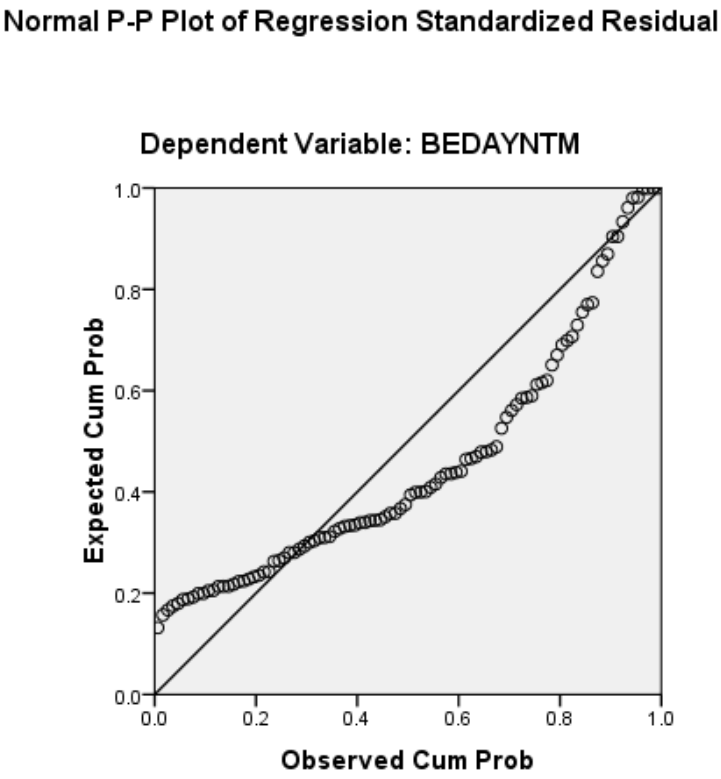
Như vậy phần dư có trung bình (mean)=0 và độ lệch chuẩn (SD)=0,394

Biểu đồ phân bố phần dư có dạng hình chuông đều 2 bên, trị trung bình gần bằng zero và SD gần bằng 1. Như vậy giả định phần dư có phân phối chuẩn không bị vi phạm.

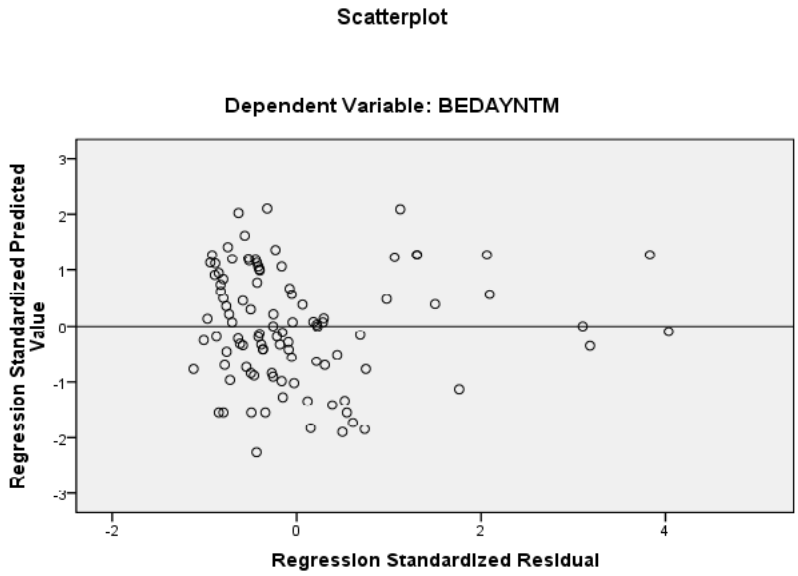


Hoặc xem biểu đồ P-P plot so sánh giữa phân phối tích lũy của phần dư quan sát (Observed Cum Prob) trên trục hoành và phân phối tích lũy kỳ

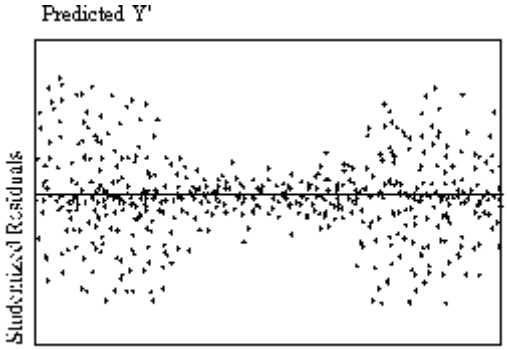
vọng (Expected Cum Prob) trên trục tung. Nếu các điểm đều nằm gần đường chéo thì phân phối phần dư được coi như gần chuẩn.



Cuối cùng để xem giả định các phương sai không đổi với mọi giá trị của x (cholesterol máu) hoặc gọi là **homoscedasticity**. Nếu các trị phần dư phân tán ngẫu nhiên quanh giá trị zero (đường ngang) thì coi như phương sai không thay đổi, và giả định về homoscedasticity không bị vi phạm.



Nếu phương sai thay đổi (lớn dần hoặc nhỏ dần theo giá trị của x) thì gọi là Heteroscedascity (giả định về phương sai cố định bị vi phạm). Ví dụ như hình dưới đây:



Tóm lại, với ví dụ trên các giả định của phân tích hồi qui tuyến tính đều thỏa mãn và ta có thể kết luận là bề dày nội trung mạc động mạch cảnh có liên hệ tuyến tính với nồng độ cholesterol máu theo phương trình :

Y (Bề dày nội trung mạc)= 0,062 x cholesterol + 0,748.

Như vậy cứ nồng độ cholesterol tăng lên 1 mmol/L thì bề dày nội trung mạc động mạch cảnh tăng lên 0,062mm.

Tài liệu tham khảo:

1. McClave J T and Sincich T. 2000. Simple linear regression in Statistics, 8th edition, Prentice-Hall, USA, pp. 505-557.
2. Moore D. S. and McCabe G. P. 1999. Looking at Data-Relationships (Chapter 2), in Introduction to the Practice of Statistics, W.H. Freeman and Company, New York, pp. 102-145.