Bài 5 PHÂN TÍCH HỒI QUY

I- NỘI DUNG

Khi nghiên cứu một tổng thể có thể theo dõi đồng thời nhiều biến. Trong chương này chỉ xem xét các biến định lượng, thí dụ trọng lượng và chiều dài trứng gà; trọng lượng, chiều cao, vòng ngực của thanh niên; chiều dài, cân nặng, trọng lượng buồng trứng của cá, chiều cao cây, đường kính bắp, trọng lượng chất khô, năng suất ngô v.v . .

Thường chia các biến ra thành 3 nhóm :

Biến mà chúng ta chủ động cho thay đổi để theo dõi ảnh hưởng của chúng đến các biến khác. Đó là lượng phân bón, lượng thuốc sử dụng, lượng thức ăn bổ sung, mật độ cấy, số ngày tính từ một thời điểm nào đó (từ khi ngừng phun thuốc, từ khi bắt đầu thu hoạch, từ khi bắt đầu bảo quản . . .). Gọi các biến này là ***biến chủ động***.

Biến liên quan đến ngoại cảnh, nhìn chung loại biến này vượt khỏi tầm kiểm tra và chúng ta chỉ ghi lại một cách thụ động, tuy nhiên phải lưu tâm vì chúng ảnh hưởng đến kết quả nghiên cúu như: lượng bức xạ, lượng mưa, số giờ nắng, độ ẩm . . . Gọi các biến này là ***biến kèm theo hay biến liên quan.***

Các biến chúng ta quan tâm, chúng là đối tượng theo dõi, là mục đích nghiên cứu và thường là kết quả của thí nghiệm như năng suất, lượng chất khô, trọng lượng 1000 hạt, lượng tăng trọng hàng tháng, sản lượng sữa, hàm lượng vitamin ... Gọi các biến này là ***biến kết quả.***

Sau khi thu được số liệu về các biến người ta ***muốn thiết lập các mối quan hệ giữa các biến***. Các quan hệ này dựa trên số liệu thu được qua theo dõi, qua thí nghiệm nên có tính chất thực nghiệm(Empirical). Nó giúp tìm hiểu quan hệ thực sự có ***tính quy luật*** giữa các biến chứ ***không chứng minh*** cho quy luật đó.

Có 2 bài toán liên quan chặt chẽ với nhau

a- Xác định các hệ số đánh giá mối quan hệ giữa 2 biến X, Y (thí dụ hệ số tương quan, tỷ số tương quan . . .) hay tổng quát hơn đánh giá mối quan hệ giữa một biến Z và một bộ k biến $X_1, X_2, \ldots, X_k$ (thí dụ hệ số tương quan bội, hệ số tương quan riêng . . .).

b-Theo dõi biến kết quả Z và một bộ k biến $X_1, X_2, \ldots, X_k$ tìm hàm $f(X_1, X_2, \ldots X_k)$ sao cho $f(X_1, X_2, \ldots X_k)$ gần Z nhất (theo một tiêu chuẩn nào đó). Hàm này có thể gọi một cách chung nhất là hàm hồi quy của Z theo bộ k biến $X_1, X_2, \ldots, X_k$

Trước hết chúng ta xem xét trường hợp 2 biến X, Y.

A- HỒI QUY TUYẾN TÍNH ĐƠN (Simple linear regression)

a1- Sắp xếp số liệu

Theo dõi một biến X (có thể thuộc loại biến chủ động hoặc biến liên quan) và biến kết quả Y.

Quan sát được n cặp (x_i,y_i), khi có ít số liệu có thể để số liệu dưới dạng 2 cột hay 2 hàng, nếu nhiều hơn có thể sắp dưới dạng có tần số, nếu nhiều nữa thì chia khoảng cả X và Y để sắp thành bảng hai chiều.

a) Sắp thành hàng

X	x_1	x_2	. . .	x_n
Y	y_1	y_2	. . .	y_n

b) Sắp thành hàng có tần số

X	x_1	x_2	. . .	x_k	
Y	y_1	y_2	. . .	y_k	
m	m_1	m_2	. . .	m_k	n

c) Sắp thành cột và sắp thành cột có tần số

X	Y
x_1	y_1
x_2	y_2
. . .	. . .
x_n	y_n

X	Y	m
x_1	y_1	m_1
x_2	y_2	m_2
. . .	. . .	. . .
x_k	y_k	m_k
	Tổng	n

d/ Sắp thành bảng X gồm k lớp, Y gồm l lớp với các điểm giữa x_i và y_j

X \ Y	y_1	y_2	. . .	y_l
x_1	m_{11}	m_{12}	. . .	m_{1l}
x_2	m_{21}	m_{22}	. . .	m_{2l}
. . .	. . .	. . .	. . .	. . .
x_k	m_{k1}	m_{k2}	. . .	m_{kl}

Từ dạng bảng có thể dễ dàng chuyển thành dạng cột hay hàng có tần số và ngược trở lại chuyển từ dạng cột hay hàng có tần số thành bảng.

Ở phần sau các công thức tính toán chỉ đúng khi số liệu viết dưới dạng **hai cột không có tần số, khi có tần số thì phải thêm tần số vào các công thức.**

a2- Mô hình hồi quy tuyến tính đơn

Vẽ các cặp số liệu quan sát được (x_i, y_i) trên hệ tọa độ Đề các. Dựa trên hình vẽ có thể nêu ra nhiều dạng quan hệ thực nghiệm giữa 2 biến X, Y, thí dụ quan hệ đường thẳng, quan hệ hàm bậc hai, quan hệ lôgarít, quan hệ mũ . . . Nếu nhiều số liệu trong một lần khảo sát hoặc nhiều lần khảo sát thì có thể lựa chọn dạng quan hệ phù hợp, nhưng nếu ít số liệu thì quan hệ nào cũng có vẻ hợp lý. Như vậy để chọn mối quan hệ thực nghiệm hợp lý giữa X và Y cần có nhiều quan sát hoặc lặp lại nhiều lần khảo sát.

Trước hết chúng ta xem xét loại quan hệ đơn giản nhất giữa X và Y là quan hệ đường thẳng, còn gọi là quan hệ tuyến tính (linear).Trong quan hệ này chúng ta coi Y phụ thuộc bậc nhất vào X.

Mô hình của quan hệ này như sau:

$$\mathbf{Y_i = a + b X_i + \varepsilon_i} \qquad i = 1,n \qquad (1)$$

ε_i là sai số ngẫu nhiên, hình thành từ nhiều nguồn, ngoài tầm kiểm tra của hệ thống nghiên cứu (sai số rất nhỏ trong điều kiện thí nghiệm, sai số của dụng cụ, sai số khi theo dõi, ghi chép kết quả . . .).

a là tung độ gốc, còn b là hệ số góc (độ dốc) của đường hồi quy

Bây giờ cần tính các tham số a,b để đường thẳng tìm được, về một khía cạnh nào đó, có thể coi là tốt nhất.

Người ta gọi bài toán này là **ước lượng tham số của đường hồi quy.**

Tùy theo tiêu chuẩn đặt ra thế nào là đường tốt nhất để đưa ra cách ước lượng a, b.

Sau đây là cách trình bầy khái niệm hồi quy trong lý thuyết giải tích và cách trình bầy khái niệm hồi quy trong lý thuyết xác suất.

a3- Phương pháp bình phương bé nhất (Least square method)

Phương pháp này đưa ra tiêu chuẩn đường thẳng tốt nhất là đường có tổng bình phương sai số nhỏ nhất. Cách tính như sau:

a) Lập tổng bình phương sai số $S = \sum (y_i - a\,x_i - b)^2$

b) Chọn a, b sao cho S nhỏ nhất

Bài toán ở đây là bài toán tìm cực trị của hàm 2 biến (Hàm S phụ thuộc 2 ẩn số a và b, còn các x_i, y_i là các số đã biết) do đó phải tính đạo hàm riêng theo a và theo b, sau đó cho các đạo hàm riêng bằng không, từ đó thu được 2 phương trình với 2 ẩn số:

$$\begin{aligned} an &+ b\sum x_i = \sum y_i \\ a\sum x_i &+ b\sum x_i^2 = \sum x_i y_i \end{aligned} \qquad (2)$$

Giải hệ này được a và b. Có nhiều cách giải hệ 2 phương trình này.

Nếu dùng định thức để giải ta có:

$$b = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2}; \quad a = \frac{(\sum y_i)(\sum x_i^2) - (\sum x_i)(\sum x_i y_i)}{n\sum x_i^2 - (\sum x_i)^2}$$

Thường hay viết đường hồi quy dưới dạng:

$$y - \bar{y} = b(x - \bar{x})$$

$$b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \qquad (3)$$

(Sau khi tính b nếu muốn tính a thì có thể dùng công thức đơn giản sau:

$$a = \bar{y} - b\,\bar{x} \qquad (4)$$

và viết phương trình dưới dạng: y = a + bx)

Đường thẳng tìm ra đơn thuần là đường "gần các điểm (x_i, y_i)" nhất chứ không đề cập đến luật phân phối của các sai số e_i , do đó không có các kiểm định đối với a, b, không có đánh giá về sai số khi dùng đường thẳng hồi quy để dự báo giá trị y tương ứng với một giá trị x đã cho.

a4- Hồi quy và tương quan trong lý thuyết xác suất

Trong lý thuyết xác suất **hệ số tương quan** giữa 2 biến ngẫu nhiên đồng thời X và Y được định nghĩa như sau:

$$\rho(X,Y) = \frac{M\{(X - MX)(Y - MY)\}}{\sqrt{M(X - MX)^2 \times M(Y - MY)^2}} \qquad (5)$$

Hệ số tương quan ρ(X,Y) có các tính chất sau:

a) Hệ số ρ nằm từ -1 đến 1 ($|\rho| <= 1$)

b) Hệ số ρ bằng và chỉ bằng ± 1 khi Y là hàm tuyến tính của X(Y = aX+b)

c) Nếu X và Y độc lập thì ρ bằng không nhưng nếu ρ = 0 thì chưa chắc X,Y đã độc lập.

d) Hệ số ρ không thay đổi khi thực hiện các biến đổi tuyến tính đối với X và Y

($X = c_1U + d_1$ $Y = c_2 V + d_2$)

Trong thực nghiệm hệ số tương quan được tính theo công thức:

$$r_{XY} == \frac{SPEXY}{\sqrt{SCEX \times SCEY}} \qquad (6)$$

Với $SCEX = \sum_i (x_i - \bar{x})^2$ $SCEY = \sum_i (y_i - \bar{y})^2$ $SPEXY = \sum_i (x_i - \bar{x})(y_i - \bar{y})$

$$\text{hay} \quad r_{XY} = \frac{SXY - \frac{(SX \times .SY)}{n}}{\sqrt{(SXX - \frac{(SX)^2}{n})(SYY - \frac{(SY)^2}{n})}} \qquad (7)$$

$SX = \Sigma x_i$; $SXX = \Sigma x_i^2$; $SY = \Sigma y_i$; $SYY = \Sigma y_i^2$; $SXY = \Sigma x_i y_i$

Nếu tính các thống kê cho 2 biến X và Y thì:

Phương sai của X $s_X^2 = \frac{SCEX}{n-1} = \frac{SXX - \frac{(SX)^2}{n}}{n-1}$

Phương sai của Y $s_Y^2 = \frac{SCEY}{n-1} = \frac{SYY - \frac{(SY)^2}{n}}{n-1}$

Hiệp phương sai của X và Y $\text{cov}ar(X,Y) = \frac{SPEXY}{n-1} = \frac{SXY - \frac{SX \times SY}{n}}{n-1}$

Khi đó hệ số tương quan tính theo công thức $r = \frac{Co\,\text{var}(X,Y)}{s_X \times s_Y}$

Hệ số tương quan thực nghiệm r_{XY} có các tính chất tương tự như ρ và thường được tóm tắt như sau:

a) $|r_{xy}| \le 1$

Nếu $r_{xy} > 0$ tương quan dương, tức là khi X tăng thì Y có khuynh hướng tăng

Nếu $r_{xy} < 0$ tương quan âm, tức là khi X tăng thì Y có khuynh hướng giảm

b) nếu Y = a + bX (Y là hàm tuyến tính của X) thì $r_{XY} = \pm 1$, ngược lại nếu $r_{XY} = \pm 1$ thì Y = a + bX, r gần về phía ± 1 gọi là tương quan mạnh, r gần về phía 0 thì gọi là tương quan yếu.

c) Nếu X và Y độc lập về xác suất thì $r_{XY} = 0$ (gọi là không tương quan).

d) Hệ số tương quan r_{xy} bất biến đối với các biến đổi tuyến tính của X và Y.

Trường hợp hai biến ngẫu nhiên X Y phân phối chuẩn 2 chiều(Binormal) ***(là phân phối thường gặp khi khảo sát đồng thời hai biến ngẫu nhiên)*** thì hệ số tương quan **ρ(X,Y)** có mặt trong **hàm mật độ xác suất** và các đường mức (đường có mật độ φ(x, y) = C) là các elip đồng tâm với tâm (MX, MY). Các elip này bầu bĩnh nếu abs(ρ(X,Y)) nhỏ và dẹt nếu abs(ρ(X,Y)) lớn.

Trường hợp phân phối chuẩn hai chiều (Binormal) hồi quy tuyến tính Y theo X được hiểu như sau:

Cho X một giá trị cố định $X = x_0$ rồi tính kỳ vọng có điều kiện của Y tại x_0 (ký hiệu là $M(Y/X=x_0)$).

Khi cho x_0 thay đổi thì điểm có tọa độ $(x_0, M(Y/X= x_0))$ sẽ chạy trên một đường thẳng gọi là đường hồi quy tuyến tính Y theo X.

Ngược trở lại khi cố định $Y= y_0$ có thể tính kỳ vọng có điều kiện của X theo Y tại y_0 (ký hiệu là $M(X/Y=y_0)$). Khi cho y_0 thay đổi thì điểm có tọa độ $(y_0, M(X/Y=y_0))$ sẽ chạy trên một đường thẳng gọi là đường hồi quy tuyến tính X theo Y.

Như vậy khi có cặp biến ngẫu nhiên phân phối chuẩn hai chiều ta có hai đường thẳng hồi quy lý thuyết: Hồi quy tuyến tính Y theo X và hồi quy tuyến tính X theo Y. Đó chính là hai đường kỳ vọng có điều kiện.

Hồi quy tuyến tính lý thuyết Y theo X có phương trình $\mathbf{y = \alpha + \beta x}$

$$\text{với} \quad \beta = \rho \frac{\sigma_Y}{\sigma_X} \qquad ; \qquad \alpha = MY - MX \tag{8}$$

Hồi quy tuyến tính lý thuyết X theo Y có phương trình **x = γ + δy**

$$\text{với} \quad \delta = \rho \frac{\sigma_X}{\sigma_Y} \qquad ; \qquad \gamma = MX - MY \tag{9}$$

Hồi quy tuyến tính thực nghiệm Y theo X có phương trình y = a + bx

$$\text{với} \quad b = r \frac{s_Y}{s_X} \qquad ; \qquad a = \bar{y} - b\bar{x} \tag{10}$$

Hồi quy tuyến tính thực nghiệm X theo Y có phương trình x = c + dy

$$\text{với} \quad d = r \frac{s_Y}{s_X} \qquad c = \bar{x} - d\,\bar{y} \tag{11}$$

Hệ số tương quan r và các hệ số hồi quy a, b, c, d là các ước lượng của các tham số ρ, α, β, γ, δ . Có thể kiểm định các giả thiết về các ước lượng này cũng như đánh giá sai số mắc phải khi dùng hồi quy tuyến tính để dự báo. Các vấn đề này trùng với các vấn đề sẽ trình bầy ở phần tiếp theo.

Trường hợp hai biến ngẫu nhiên X, Y không phân phối chuẩn hai chiều thì đường kỳ vọng có điều kiện **y = f(x) = M(Y/x**) là đường hồi quy lý thuyết của Y theo X và là đường tốt nhất theo nghĩa **bình phương trung bình**, tức là khi dùng f(x) thay cho Y thì độ lệch bình phương trung bình sẽ nhỏ nhất so với mọi hàm g(x)

($M[Y- f(x)]^2 \le M[Y - g(x)]^2$ với mọi g(x))

Trong trường hợp tổng quát y = f(x) = M(Y/x) không phải đường thẳng và đường tuyến tính **y = a + bx** tính theo (8) chỉ là đường tốt nhất theo nghĩa bình phương trung bình trong **lớp các hàm tuyến tính của y theo x.**

a5- Trường hợp X không phải biến ngẫu nhiên

Xét trường hợp biến X không ngẫu nhiên. Giả sử khi $X = x_i$ thì Y là biến ngẫu nhiên phân phối chuẩn có kỳ vọng là hàm bậc nhất $a + bx_i$ và phương sai σ^2. Nói cách khác Y được tính theo mô hình (1)

$$y_i = a + bx_i + \varepsilon_i$$

với giả thiết các ε_i **độc lập, phân phối chuẩn N(0, σ^2).**

Các hệ số a và b của đường thẳng y = a + bx được tính theo hệ phương trình (2) hay theo công thức (10). Hai cách tính cho cùng một kết quả. Vì các sai số ε_i độc lập, phân phối chuẩn $N(0,\sigma^2)$ nên các hệ số a,b và hệ số tương quan r_{xy} tính như trên đều mắc sai số.

Ứng với mỗi giá trị x_i tính giá trị tương ứng của đường hồi quy

$$\hat{y}_i = a + bx_i$$

Gọi độ lệch (còn gọi là phần dư) $e_i = y_i - \hat{y}_i$

Đem bình phương độ lệch e_i, cộng lại rồi chia cho (n - 2) được:

$$se^2 = \frac{\sum_i e_i^2}{(n-2)}$$

Phương sai $\boldsymbol{\sigma^2}$ (giả thiết e_i phân phối chuẩn $N(0, \sigma^2)$) được ước lượng bằng se^2.

Có thể tính se^2 qua công thức sau: $se^2 = \frac{(1-r_{XY}^2)\sum_i (y_i - \bar{y})^2}{(n-2)}$

se được gọi là sai số ngẫu nhiên của 1 quan sát, se có bậc tự do là (n-2).

Sai số của hệ số b $\quad sb = \frac{se}{\sqrt{SCEX}}$

Sai số của hệ số a $\quad sa = se\sqrt{\frac{SSX}{nSCEX}} = se\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SCEX}}$

Kiểm định giả thiết H_0: a = 0 bằng giá tri $T_{tna} = a/sa$

Kiểm định giả thiết H_0: b = 0 bằng giá trị $T_{tnb} = b/sb$

Cả hai giá trị thực nghiệm trên đều so với giá trị tới hạn $T_{lt} = t(\alpha, n-2)$.

Khi cho một giá trị x_0 ngoài các giá trị x_i đã cho có thể tính giá trị tương ứng theo đường hồi quy, gọi là **giá trị dự báo trung bình** $y_0 = a + bx_0$.

Giá trị này mắc sai số:

$$s(\bar{y}_0) = se\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SCEX}}$$

Khoảng tin cậy $y_0 \pm s(\bar{y}_o)$ gọi là khoảng ước lượng (CI)

Nếu dùng y_0 làm **giá trị dự báo cho y tại x_0** thì sai số của dự báo:

$$sydb(\bar{y}_0) = se\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SCEX}}$$

Khoảng tin cậy $y_0 \pm sydb(\bar{y}_o)$ gọi là khoảng dự báo (PI)

Đối với giá trị r_{xy} người ta dùng các biến đổi để đưa về biến chuẩn sau đó ước lượng và kiểm định.

Nếu số quan sát không nhỏ lắm có thể kiểm định giả thiết không tương quan

H_0: $r_{xy} = 0$ bằng giá trị Student $T_{tnr} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

(so abs(Ttn) với ngưỡng $T_{lt} = t(\alpha, n-2)$).

Kiểm định giả thiết r = 0 và kiểm định giả thiết b = 0 tương đương vì $b = r\frac{s_y}{s_x}$

Thường lập bảng phân tích phương sai để tách riêng tổng bình phương SCEY thành hai phần: phần biến động do hồi quy tuyến tính và phần biến động do sai số ngẫu nhiên (đôi khi còn nói là biến động của các điểm trên đường hồi quy tuyến tính $(x_i, \hat{y}_i)$ và biến động của các điểm thực nghiệm (x_i, y_i) quanh đường hồi quy)

Phần do hồi quy tuyến tính được tính theo công thức

$SSl = (SPEXY)^2/SCEX$ (hay r^2 SCEY)

Phần do sai số: SSE hay SSR = SCEY - SSl (hay $(1 - r^2)$ SCEY)

Bảng phân tích phương sai

Nguồn biến động	Tổng B P	Bậc tự do	Trung bình	Ftn
Do hồi quy tuyến tính	SSl	1	smr	Smr/Sme
Sai số	SSE	n -2	sme = SSE /(n-2)	
Toàn bộ	SCEY	n - 1	se^2	

So F_{tn} với F_{lt} ở mức tin cậy α và các bậc tự do 1, n-2 để kiểm định xem đường hồi quy có đáng tin cậy hay không (biến động do hồi quy vượt xa biến động ngẫu nhiên do sai số).

Phép kiểm định này hoàn toàn tương đương với kiểm định Student của giả thiết H_0:b = 0 vì $F_{tn} = T^2_{tnb}$

a6 - Một số đường cong có thể biến đổi thành dạng tuyến tính.

Trong nông nghiệp thường gặp các đường sau:

a) $Y = ae^{bX}$ lấy lôgarít được $LnY = Lna + bX$

đặt $U = LnY$ $A = Lna$ ta có $U = A + bX$

b) $Y = ab^X$ lấy lôgarít được $LnY = Lna + X\,Lnb$

đặt $U = LnY$ $A = Lna$ $B = Lnb$ có $U = A + BX$

c) $Y = 1/(a + bX)$ đặt $U = 1/Y$ có $U = a + bX$

d) $Y = a + b/X$ đặt $V = 1/X$ có $Y = a + bV$

Như vậy là bằng một số phép biến đổi có thể đưa đường cong về dạng tuyến tính nhưng những giả thiết về sai số e_i trong mô hình ban đầu không còn đúng khi biến đổi do đó phải có các giả thiết mới về sai số e_i trong mô hình đã biến đổi. Nếu giả thiết phù hợp ta tính được đường hồi quy tuyến tính sau đó có thể sử dụng ở dạng biến đổi hoặc biến đổi ngược để trở lại biến ban đầu. thí dụ có

$Y = a\,e^{bX}$ sau khi biến đổi lôgarít được $U = A + B\,X$

(U = LnY A = Lna B = b)

giả sử tìm được đường hồi quy $U = 4{,}45791 - 0{,}40342X$

Biến đổi ngược a = antilog 4,45791 = 86,31 có hồi quy ban đầu

$Y = 86{,}31e^{-0{,}40342\,X}$

B- HỒI QUY BỘI TUYẾN TÍNH

Gọi biến phụ thuộc là Y, các biến độc lập là $X_1, X_2, \ldots, X_p$.

Có thể viết hồi quy bội tuyến tính dưới dạng ma trận như sau:

gọi **Y** (n x 1) là vectơ các giá trị Y ,

b (p +1 x 1) là vectơ hệ số b_i i = 0, p

X ma trận (n x p +1) các quan sát ($X_{0i} = 1$, $X_{1i}, X_{2i}, \ldots, X_{pi}$)

e (n x 1) là vectơ các sai số

(giả thiết phân phối chuẩn, độc lập, phương sai không đổi)

$V(y) = V(e) = \sigma^2 I_n$ (I_n là ma trân đơn vị cấp n)

Hồi quy bội tuyến tính có dang:

$$Y = b_0 + b_1X_1 + b_2 X_2 + \ldots + b_p X_p + e \qquad (b1)$$

$$\mathbf{Y = Xb + e} \qquad (b2)$$

Dùng phương pháp bình phương bé nhất tính được các hệ số b_i như sau:

$$\mathbf{b} = (\mathbf{X'X})^{-1} (\mathbf{X'Y}) \qquad (b3)$$

(Đem ma trận chuyển vị **X'** nhân với vectơ Y ta được **X'Y** sau đó tính tích của hai ma trận $(\mathbf{X'X})^{-1}$ **(X'Y)) .**

Nếu dùng các biến quy tâm **y , x_1, x_2,. . . , x_p** thì có thể bỏ bớt hệ số b_0 trong vectơ **b** và gọi **x** là ma trân các giá trị quy tâm (x_{1i}, x_{2i}, . . ., x_{pi})

$$\mathbf{y = x\,b + e} \qquad (b4)$$

$$\mathbf{b = (x'x)^{-1}(x'y)} \qquad (b5)$$

Sau đó tính b_0 theo công thức $\mathbf{b_0} = \bar{Y} - (\mathbf{b_1}\bar{X}_1 + \mathbf{b_2}\bar{X}_2 + \ldots + \mathbf{b_p}\bar{X}_p)$

Có hồi quy rồi chúng ta tính các giá trị theo hồi quy $\hat{Y}_i$ rồi lần lượt tính:

Tổng bình phương toàn bộ SSTO = $\sum y^2$ với n - 1 bậc tự do

Tổng bình phương do hồi quy $SSR = \sum_i (\hat{Y}_i - \bar{Y})^2$ với p bậc tự do

(Thường tính SSR theo công thức sau**:**

SSR = $b_1\sum x_1y + b_2\sum x_2y + \ldots + b_p\sum x_py$ = **b(x'y)**)

Tổng bình phương các sai số $SSE = \sum_i (Y_i - \hat{Y}_i)^2$ với (n - p - 1) bậc tự do

(hoặc tính bằng hiệu số SSE = SSTO - SSR)

Tỷ số SSR/ SSTO là hệ số xác định D, căn của D là hệ số tương quan bội R

Bảng phân tích phương sai

Nguồn Biến động	Btd	Tổng bình phương	Trung bình	Ftn
Hồi quy bội tt	p	SSl = R^2 SSt	sml	sml/se^2
Sai số	n-p-1	SSE = (1 - R^2)SSTO	sme=se^2	
Toàn bộ	n - 1	SSTO		

Sai số của 1 quan sát hay còn gọi là độ lệch chuẩn se

Sai số bình phương của các hệ số b_i (i = 1, p)

$$(Sb_i)^2 = C_{ii} . se^2$$

với $\mathbf{C_{i\,i}}$ là phần tử (i,i) trên đường chéo của $\mathbf{(x'x)^{-1}}$

Khi cho bộ số (X_{10},X_{20}, . . .,X_{p0}), hay nói vắn tắt cho vectơ quan sát $\mathbf{X_0}$ ta có giá trị dự báo trung bình Y_{TB} theo (b2) hoặc giá trị y_{tb} theo (b4)

Khoảng tin cậy $Y_{TB} \pm t_{0,05}\, s_e \sqrt{\frac{1}{n} + x_0^{'} (x'x)^{-1} x_0}$

Gía trị dự báo Y_{DB} có khoảng tin cậy:

$$Y_{DB} \pm t_{0,05}\, s_e \sqrt{1 + \frac{1}{n} + x_0^{'} (x'x)^{-1} x_0}$$

Thí dụ

X1	X2	X3	Y
126	121	10.0	2.3
146	140	15.2	3.0
124	121	13.3	2.3
157	160	10.9	3.0
140	137	12.7	2.0
161	162	12.9	3.2
119	111	11.0	2.4
131	124	10.5	3.0
185	175	15.4	2.3
159	158	10.6	2.6
156	145	9.0	2.8
194	175	13.7	2.7
216	197	13.2	3.3
Tổng			
2014	1926	158.4	34.9
Tbình			
155	148	12.2	2.7

Ma trận **X'X** (tính cả Y**)**

322194.00	307144.00	24826.50	5470.10
307144.00	293240.00	23716.70	5229.30
24826.50	23716.70	1978.14	425.20
5470.10	5229.30	425.20	95.65

Ma trận **x' x** (tính cả y)			
10178.923	8762.1538	286.6846	63.2846
8762.1538	7895.6923	249.1308	58.7308
286.6846	249.1308	48.0970	-0.0431
63.2846	58.7308	-0.0431	1.9569

Ma trận tương quan			
1.000	0.977**	0.410	0.448
0.977**	1.000	0.404	0.472
0.410	0.404	1.000	-0.004
0.448	0.472	-0.004	1.000

Ma trận nghich đảo $\mathbf{(x'x)^{-1}}$

0.00220931	-0.00243406	-0.00056097
	0.00283306	-0.00016626
		0.02499565

Bảng phân tích phương sai

Nguồn B Đ	BTD	Tổng B P	T bình	Ftn
Hồi quy	3	0,5306115477	0,1768703849	1,12
Sai số	9	1,142631192215	0,1584791024	
Toàn bộ	12	1,956923076923		

Hệ số xác định $R^2 = 0.27$ Hồi quy $Y = 1{,}9 - 0{,}003\ X_1 + 0{,}01\ X_2 - 0{,}05\ X_3$

Các hệ số và sai số

Biến	Hệ số	Sai số	Ttn
X_1	-0.00311469	0.018711740	0.17
X_2	0.01235684	0.021189185	0.58
X_3	-0.04633590	0.062938765	0.74

Bảng các giá trị quan sát Y và giá trị hồi quy $\hat{y}$

Y	2.3	3.0	2.3	3.0	2.0	3.2	2.4	3.0	2.3	2.6	2.8	2.7	3.3
$\hat{y}$	2.5	2.5	2.4	2.9	2.6	2.8	2.4	2.5	2.8	2.9	2.8	2.8	3.1

C- HỒI QUY ĐA THỨC

Theo dõi quan hệ giữa biến độc lập X và biến phụ thuộc Y ngoài dạng đơn giản nhất là tuyến tính còn có:

Dạng hồi quy bậc hai

$$Y = b_0 + b_1 X + b_2 X^2$$

Dạng hồi quy bậc ba

$$Y = b_0 + b_1 X + b_2 X^2 + b_3 X^3$$

Dạng đa thức bậc m

$$Y = b_0 + b_1 X + b_2 X^2 + \ldots + b_m X^m \qquad (1)$$

Đối với các hồi quy đa thức dùng phương pháp bình phương bé nhất có thể lập được hệ phương trình chuẩn để tìm các hệ số.

Có một cách khác là dùng ngay hồi quy bội tuyến tính để giải. Muốn vậy ta chỉ việc đặt $X_1 = X$ $X_2 = X^2$ $X_3 = X^3$ v v ...

Sau đó tính hồi quy bội tuyến tính đối với các biến X_1, X_2, ...

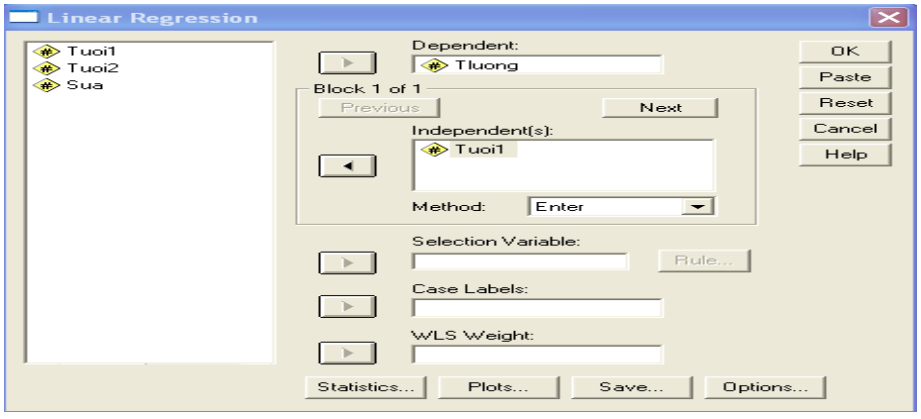
Trường hợp số liệu X cách đều người ta hay viết hồi quy đa thức (1) dưới dang hồi quy của các đa thức trực giao.

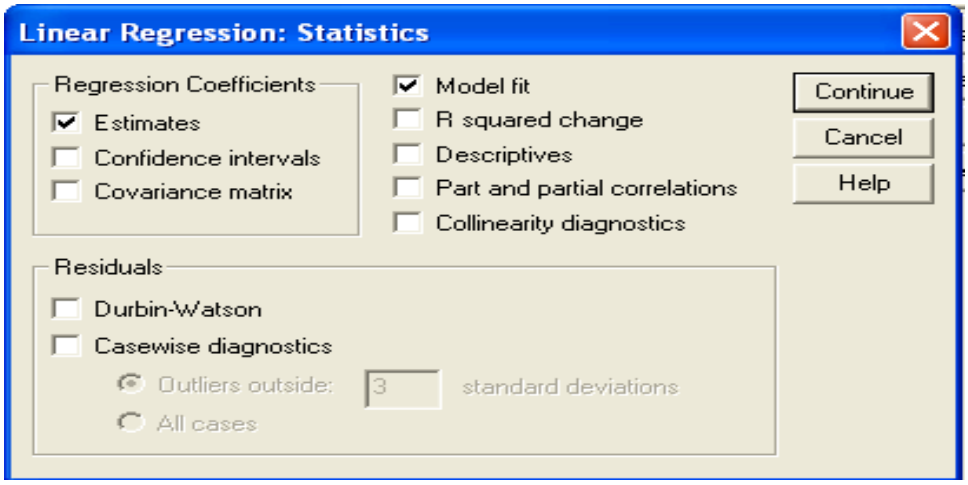
II-XỬ LÝ TRONG SPSS

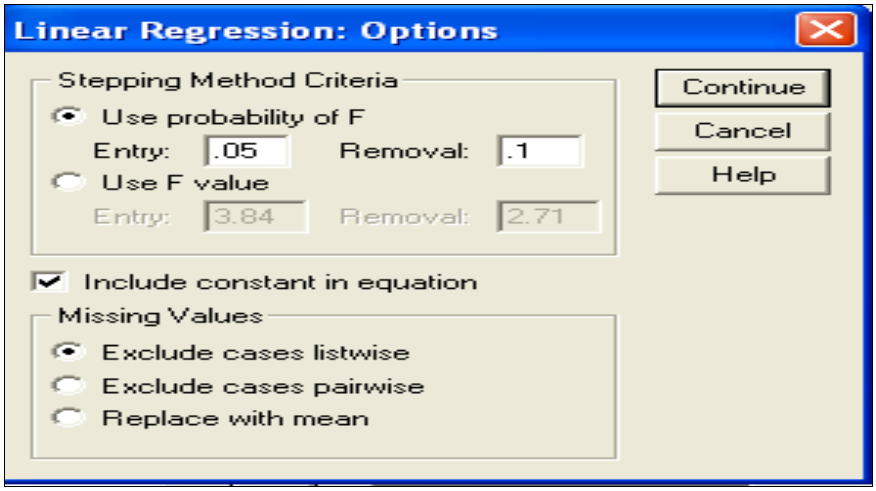
Mở tệp Baitap4

A- Hồi quy tuyến tính đơn

Vào Analyse Regression Linear Chọn Tluong (trọng lượng của bê) vào Dependent, chọn Tuoi1 (tuổi của bê tính theo tháng) vào Independent. Chọn Enter ở Method







Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.978(a)	.957	.948	12.147

a Predictors: (Constant), Tuoi1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	16225.724	1	16225.724	109.974	.000[a]
	Residual	737.704	5	147.541		
	Total	16963.429	6			

a. Predictors: (Constant), Tuoi1

b. Dependent Variable: Tluong

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	12.949	7.663		1.690	.152
	Tuoi1	12.867	1.227	.978	10.487	.000

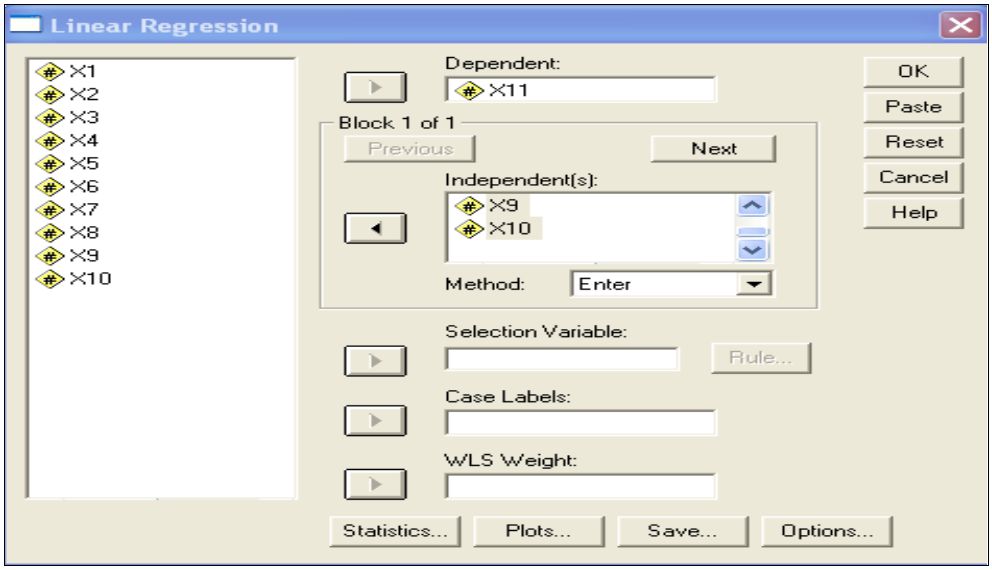
a. Dependent Variable: Tluong

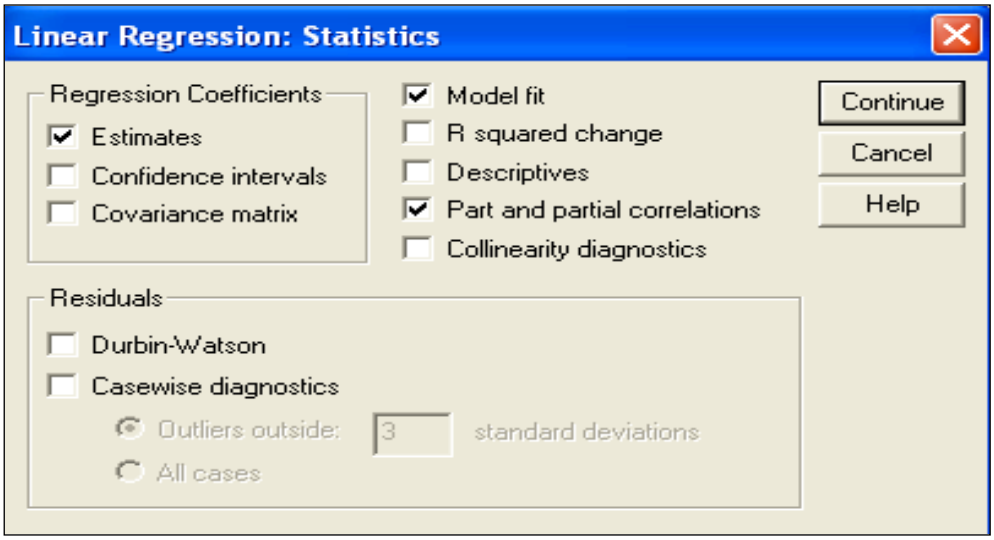
B- Hồi quy bội tuyến tính

Mở tệp caythong Analyse Regresion Linear

Dependent : X11 Independent: X1- X10

Method Enter





Variables Entered/Removed(b)

Model	Variables Entered	Variables Removed	Method
1	X10, X7, X5, X2, X9, X1, X3, X8, X4, X6(a)	.	Enter

a All requested variables entered.
b Dependent Variable: X11

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.876(a)	.768	.663	.46833

a Predictors: (Constant), X10, X7, X5, X2, X9, X1, X3, X8, X4, X6

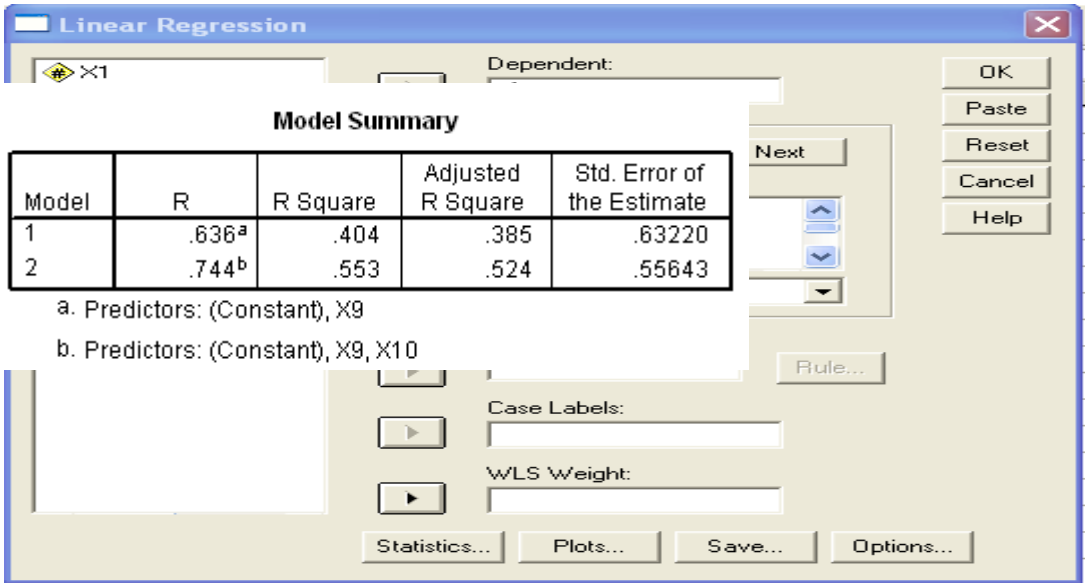
ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	15.975	10	1.597	7.283	.000[a]
	Residual	4.825	22	.219		
	Total	20.800	32			

a. Predictors: (Constant), X10, X7, X5, X2, X9, X1, X3, X8, X4, X6

b. Dependent Variable: X11

Nếu muốn sử dụng hồi quy lọc thì vào Regresion Stepwise



Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	X9	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	X10	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: X11

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.410	1	8.410	21.042	.000[a]
	Residual	12.390	31	.400		
	Total	20.800	32			
2	Regression	11.512	2	5.756	18.590	.000[b]
	Residual	9.289	30	.310		
	Total	20.800	32			

a. Predictors: (Constant), X9

b. Predictors: (Constant), X9, X10

c. Dependent Variable: X11

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.605	.406		6.413	.000
	X9	-.905	.197	-.636	-4.587	.000
2	(Constant)	2.229	.377		5.916	.000
	X9	-.830	.175	-.583	-4.733	.000
	X10	.099	.031	.390	3.165	.004

a. Dependent Variable: X11

Nếu dùng Method Backward thì kết quả như sau:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.876[a]	.768	.663	.46833
2	.876[b]	.768	.677	.45832
3	.875[c]	.765	.687	.45130
4	.870[d]	.756	.688	.45052
5	.868[e]	.754	.697	.44398

a. Predictors: (Constant), X10, X7, X5, X2, X9, X1, X3, X8, X4, X6

b. Predictors: (Constant), X10, X7, X5, X2, X9, X1, X3, X4, X6

c. Predictors: (Constant), X10, X7, X5, X2, X9, X1, X4, X6

d. Predictors: (Constant), X10, X7, X5, X2, X9, X1, X6

e. Predictors: (Constant), X10, X7, X2, X9, X1, X6

ANOVA[f]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	15.975	10	1.597	7.283	.000[a]
	Residual	4.825	22	.219		
	Total	20.800	32			
2	Regression	15.969	9	1.774	8.447	.000[b]
	Residual	4.831	23	.210		
	Total	20.800	32			
3	Regression	15.912	8	1.989	9.766	.000[c]
	Residual	4.888	24	.204		
	Total	20.800	32			
4	Regression	15.726	7	2.247	11.068	.000[d]
	Residual	5.074	25	.203		
	Total	20.800	32			
5	Regression	15.675	6	2.613	13.254	.000[e]
	Residual	5.125	26	.197		
	Total	20.800	32			

a. Predictors: (Constant), X10, X7, X5, X2, X9, X1, X3, X8, X4, X6

b. Predictors: (Constant), X10, X7, X5, X2, X9, X1, X3, X4, X6

c. Predictors: (Constant), X10, X7, X5, X2, X9, X1, X4, X6

d. Predictors: (Constant), X10, X7, X5, X2, X9, X1, X6

e. Predictors: (Constant), X10, X7, X2, X9, X1, X6

f. Dependent Variable: X11

Nếu dùng Method Forward thì kết quả tương tự như Method Stepwise

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.636[a]	.404	.385	.63220
2	.744[b]	.553	.524	.55643

a. Predictors: (Constant), X9

b. Predictors: (Constant), X9, X10

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.410	1	8.410	21.042	.000[a]
	Residual	12.390	31	.400		
	Total	20.800	32			
2	Regression	11.512	2	5.756	18.590	.000[b]
	Residual	9.289	30	.310		
	Total	20.800	32			

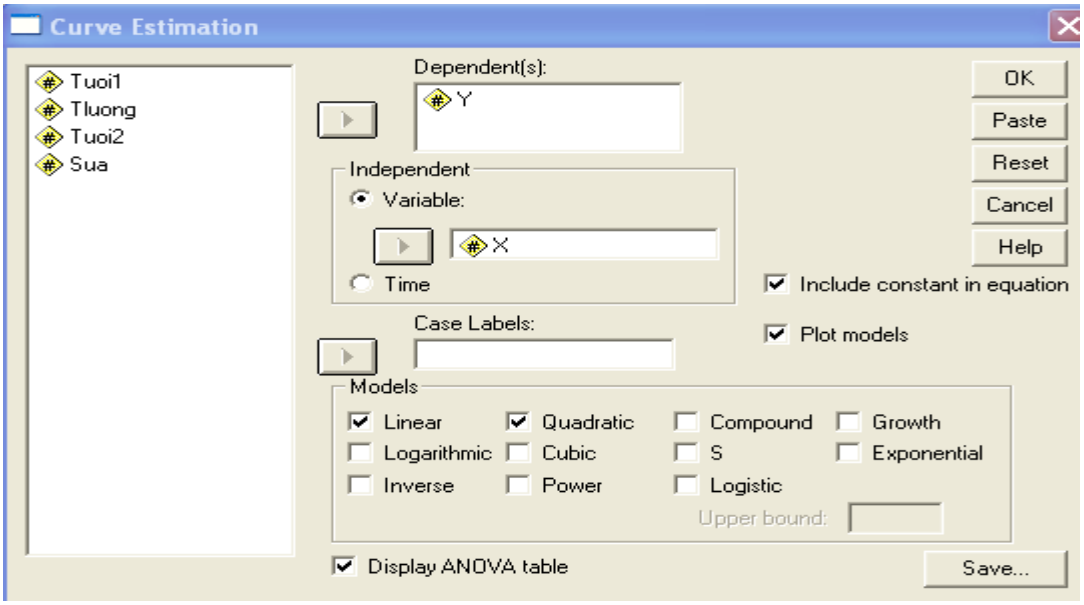
a. Predictors: (Constant), X9

b. Predictors: (Constant), X9, X10

c. Dependent Variable: X11

C- Một số hồi quy phi tuyến

Vào Analyse regression curve estimation



Chọn Y vào Dependent, X vào variable, trong Models chọn linear (bậc nhất) và Quadratic (bậc hai)

Kết quả:

Model Summary **Linear**

R	R Square	Adjusted R Square	Std. Error of the Estimate
.815	.664	.631	4.541

The independent variable is X.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	408.052	1	408.052	19.789	.001
Residual	206.198	10	20.620		
Total	614.250	11			

The independent variable is X.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
X	-.489	.110	-.815	-4.449	.001
(Constant)	18.243	2.134		8.547	.000

Model Quadrratic

R	R Square	Adjusted R Square	Std. Error of the Estimate
.927	.860	.829	3.089

The independent variable is X.

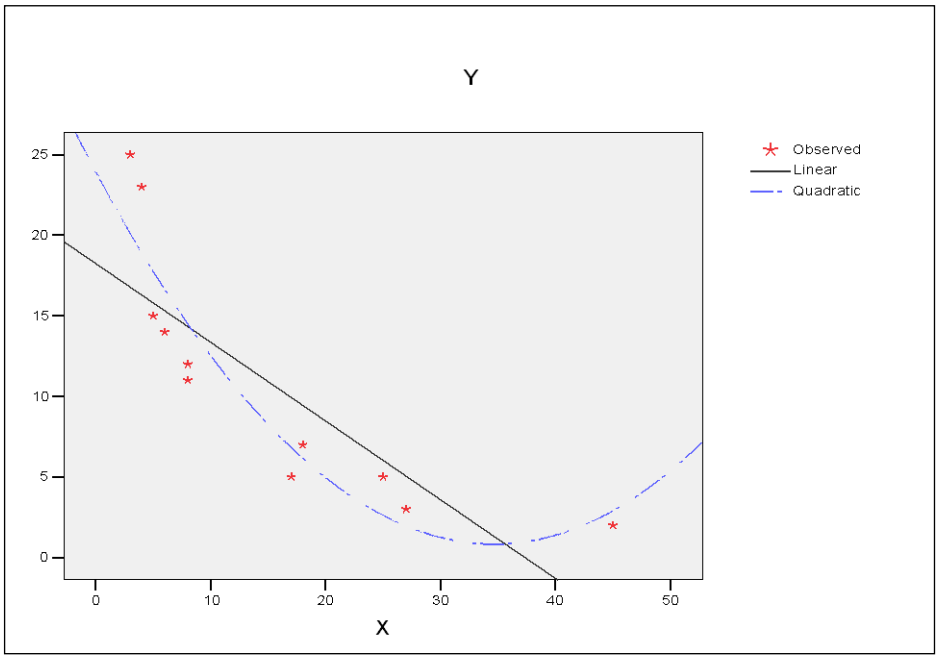
ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	528.387	2	264.194	27.692	.000
Residual	85.863	9	9.540		
Total	614.250	11			

The independent variable is X.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
X	-1.335	.250	-2.226	-5.346	.000
X ** 2	.019	.005	1.479	3.552	.006
(Constant)	23.940	2.164		11.065	.000



Trong mục Curve estimation có thể chọn các mô hình hồi quy khác nhau như:

Linear (Tuyến tính hay bậc nhất) Model whose equation is Y = b0 + b1 * x.

. The series values are modeled as a linear function of time.

Logarithmic (lôgarit) Model whose equation is Y = b0 + b1 * ln(x).

Inverse (nghịch đảo) Model whose equation is Y = b0 + b1 / x

Quadratic (bậc hai) Model whose equation is $Y = b0 + b1x + b2\ x^2$

Cubic (bậc ba) Model defined by the equation

$$Y = b0 + b1x + b2\ x^2 + b3\ x^3$$

Power (lũy thừa)Model whose equation is $Y = b0\ x^{b1}$

or $\ln(Y) = \ln(b0) + b1\ \ln(x)$

Compound. Model whose equation is $Y = b0.\ b1^x$ or $\ln(Y) = \ln(b0) + \ln(b1)\ x$

S-curve (Hình chữ S) Model whose equation is $Y = e^{(b0 + b1/x)}$

or $\ln(Y) = b0 + b1/x$.

Logistic (Lôgistic) Model whose equation is $Y = 1 / (1/u + (b0 * (b1^x))$

or $\ln(1/y-1/u) = \ln(b0) + \ln(b1)x$

where u is the upper boundary value. After selecting Logistic, specify the upper boundary value to use in the regression equation. The value must be a positive number, greater than the largest dependent variable value.

Trong Regression còn có một số loại hồi quy hay dùng trong các kiểm định hoạt tính của thuốc và kiểm định sinh học (Bioassay) nhu binary logistic, probit và có cả phương pháp tổng quát để ước lượng các hệ số trong các hồi quy phi tuyến.

Thí dụ về hồi quy dạng mũ : X1 ngày tuổi, Y1 trọng lượng phôi gà