

PHÂN TÍCH DỮ LIỆU BẰNG PHẦN MỀM SPSS 12.0*

PHẦN 4

Nội dung chính trong phần này:

1. Khai báo các thông số của biến
2. Tạo biến giả
3. Hồi quy OLS kết hợp với phương pháp **Stepwise**

* SPSS 12.0 là sản phẩm đã đăng ký của SPSS Inc.

1. Khai báo các thông số của các biến trong bộ dữ liệu

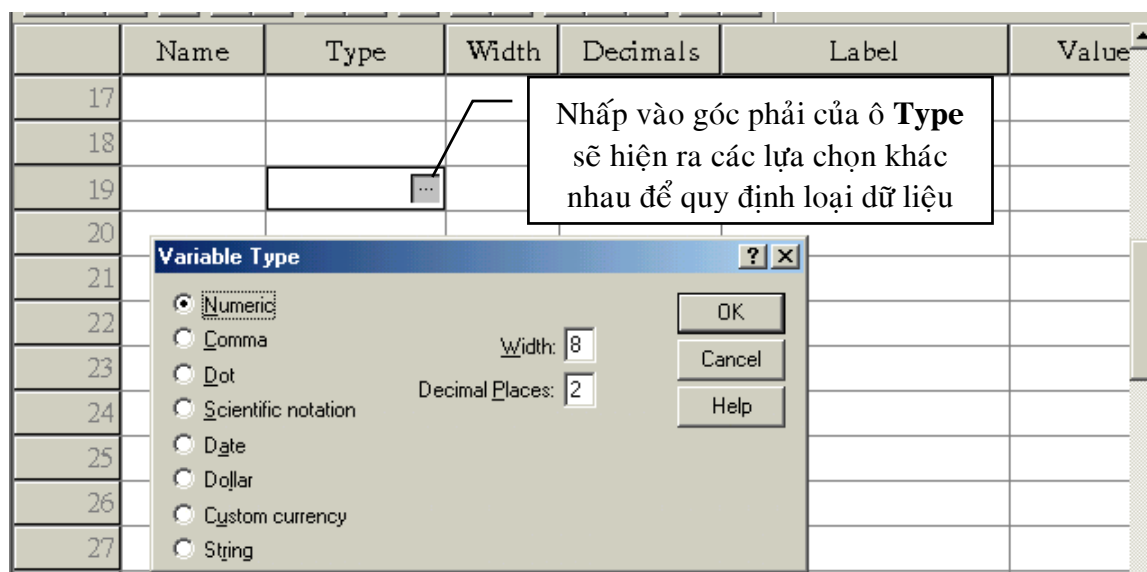
Khi sử dụng SPSS 12.0, chúng ta thường thấy hai Sheet: **Data View** và **Variable View**. **Data View** chứa dữ liệu còn **Variable View** chứa các thông tin của các biến trong dữ liệu. Các thông tin này bao gồm:

- a. **Name:** tên của biến.
- b. **Type:** loại dữ liệu của biến.
- c. **Width:** số lượng ký tự hay số lượng chữ số được hiển thị.
- d. **Decimals:** số lượng chữ số thập phân.
- e. **Label:** nhãn của biến.
- f. **Values:** nhãn hoặc giá trị của các quan sát trong biến (phát huy tác dụng tốt trong thống kê mô tả).
- g. **Missing:** số lượng quan sát bị khuyết.
- h. **Columns:** chiều rộng của cột.
- i. **Align:** vị trí (nếu là số thì sẽ là bên phải, còn là ký tự sẽ nằm bên trái)
- j. **Measure:** thang đo

Chi tiết cho một số thông tin quan trọng của biến:

- a. **Name:** tên của biến. Tên biến phải bắt đầu bằng một chữ và có độ dài tối đa là 64 ký tự (không sử dụng các ký tự đặc biệt, không kết thúc tên biến bằng dấu chấm ".").
- b. **Type:** loại dữ liệu của biến.

HÌNH 1



c. **Measure:** thang đo

Trong SPSS 12.0 có 3 loại thang đo: Scale, Nominal và Ordinal.

- **Scale:** cho biết dữ liệu là những con số định lượng (ví dụ: thu nhập, tuổi, chiều cao ...).
- **Nominal:** dữ liệu là chữ hoặc con số định tính (ví dụ: nam, nữ, hay 0, 1).
- **Ordinal:** dữ liệu là chữ hoặc con số định tính **nhưng chú ý đến thứ bậc, mức độ cao thấp, nặng nhẹ...** (ví dụ: thấp, trung bình, cao; kịch liệt phản đối, phản đối, nhất trí, nhất trí cao).

2. **Tạo biến giả**

Giả sử chúng ta có bộ dữ liệu sau được import từ Excel:

HÌNH 2

	hieu	doanh_thu	loai_xe	don_gia	quy	var	var
1	Toyota	21500	oto	21500	1		
2	Toyota	28400	oto	28400	2		
3	Toyota	32000	oto	32000	3		
4	Toyota	42000	oto	42000	4		
5	Toyota	23990	oto	23990	2		
6	Toyota	33950	xe tai	33950	3		
7	Toyota	62000	oto	62000	1		
8	Honda	26990	oto	26990	4		

Dữ liệu này là các quan sát ngẫu nhiên của một cửa hàng bán ô tô và xe tải trong năm.

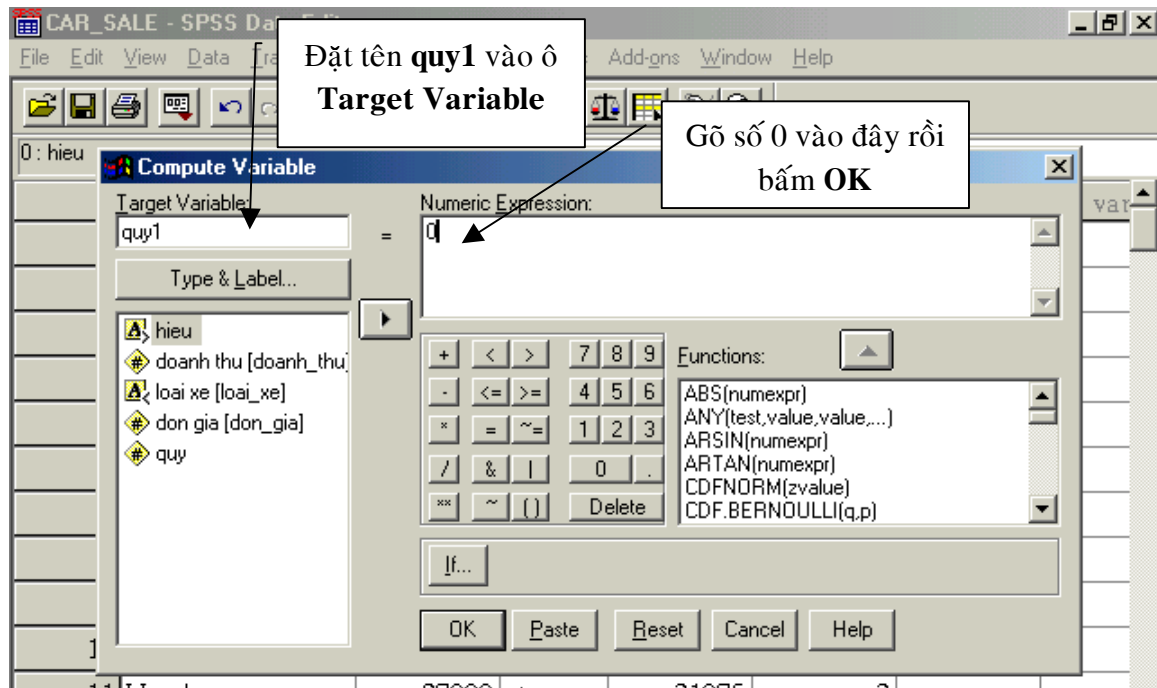
Trong đó:

- hieu: tên của nhà sản xuất.
- doanh_thu: doanh thu trong ngày quan sát (USD).
- loai_xe: loại xe ô tô hay xe tải.
- don_gia: đơn giá (USD).
- quy: quý mà quan sát đổ rơi vào.

2.1. Tạo 3 biến giả thể hiện Quý 1, Quý 2 và Quý 3

Vào Menu **Transform**, chọn **Compute**

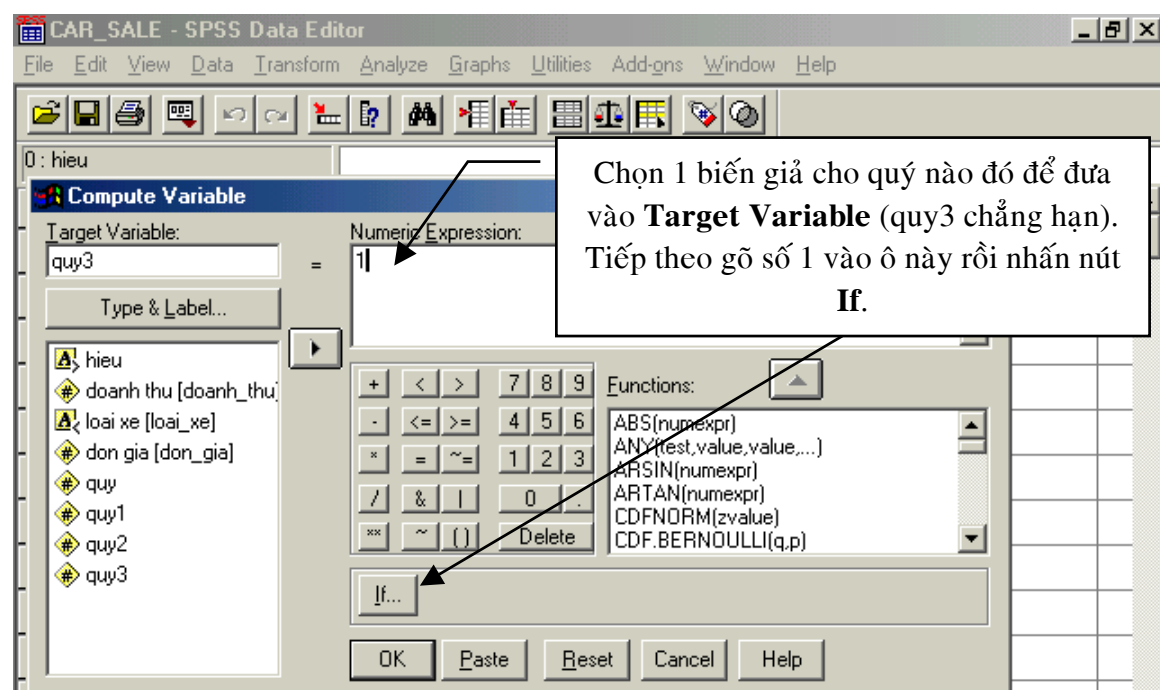
HÌNH 3



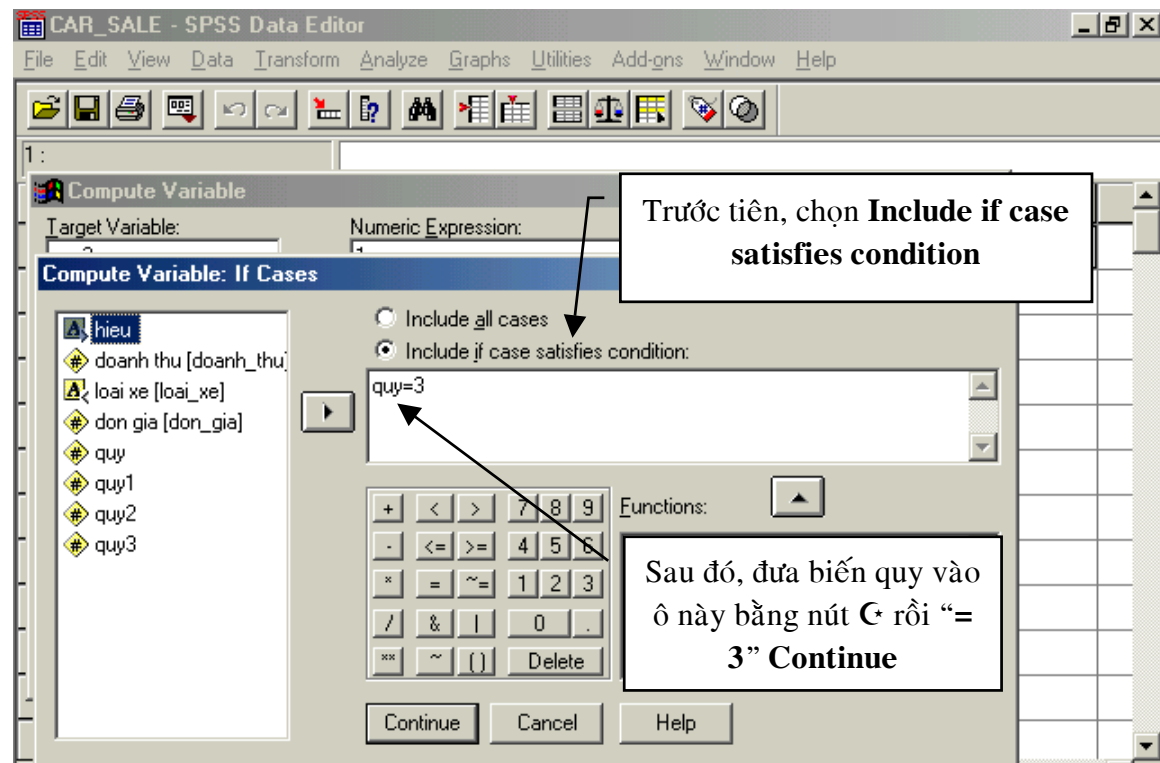
Lần lượt thực hiện như thế cho **quy2** và **quy3**. Lúc này trong dữ liệu đã xuất hiện biến **quy1**, **quy2** và **quy3** với tất cả các giá trị đều bằng 0.

Tiếp theo, lại trở vào **Compute**.

HÌNH 4



HÌNH 5

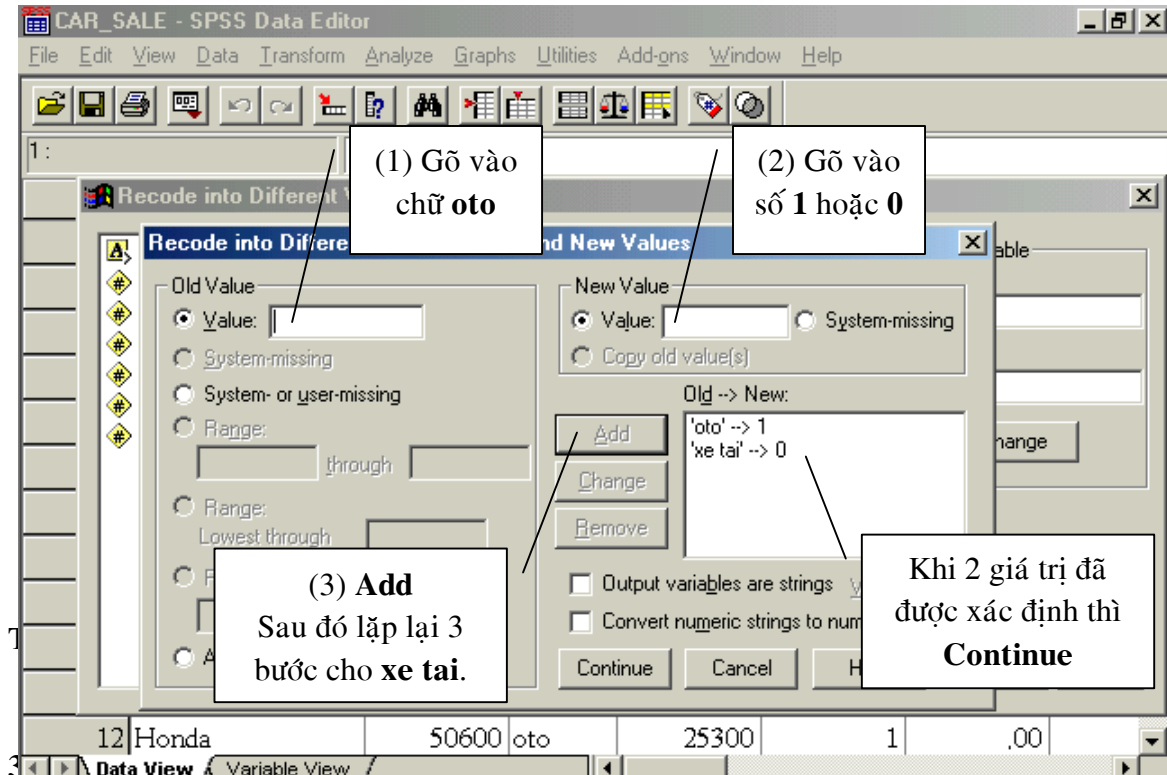
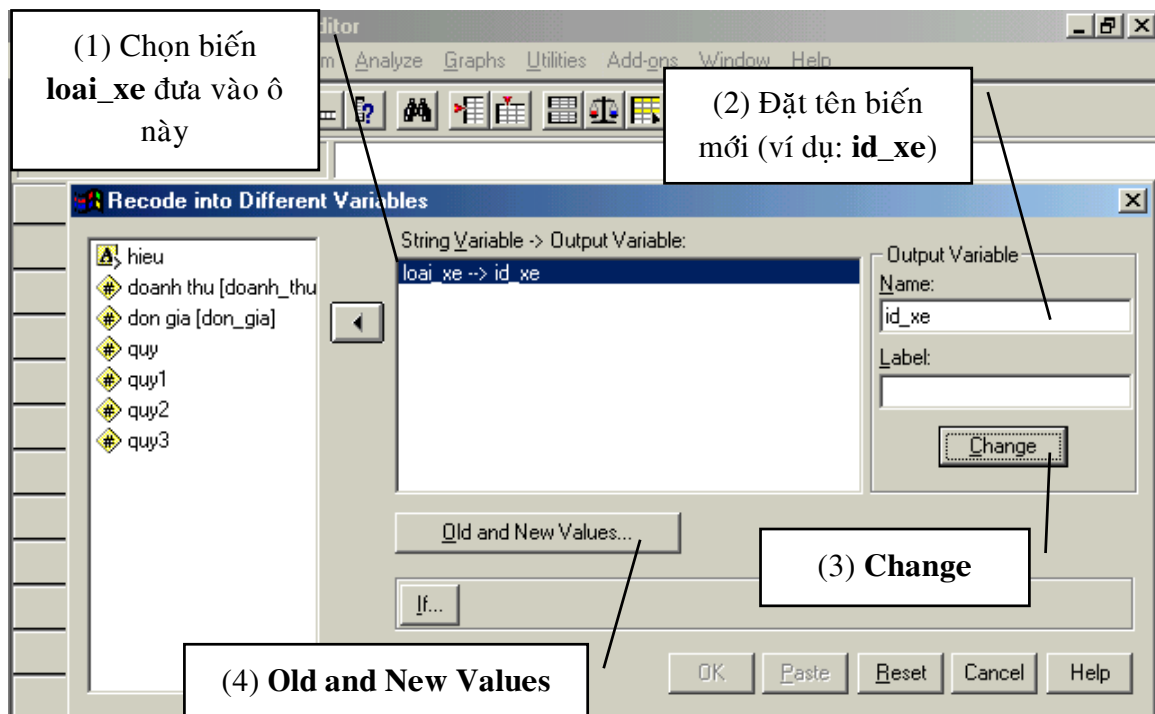


Trở lại hộp thoại phía trước, tiếp tục chọn OK để hoàn tất tạo biến giả **quy3**, và tiếp tục làm tương tự cho quy1 và quy2.

2.2. Tạo biến giả cho loại xe

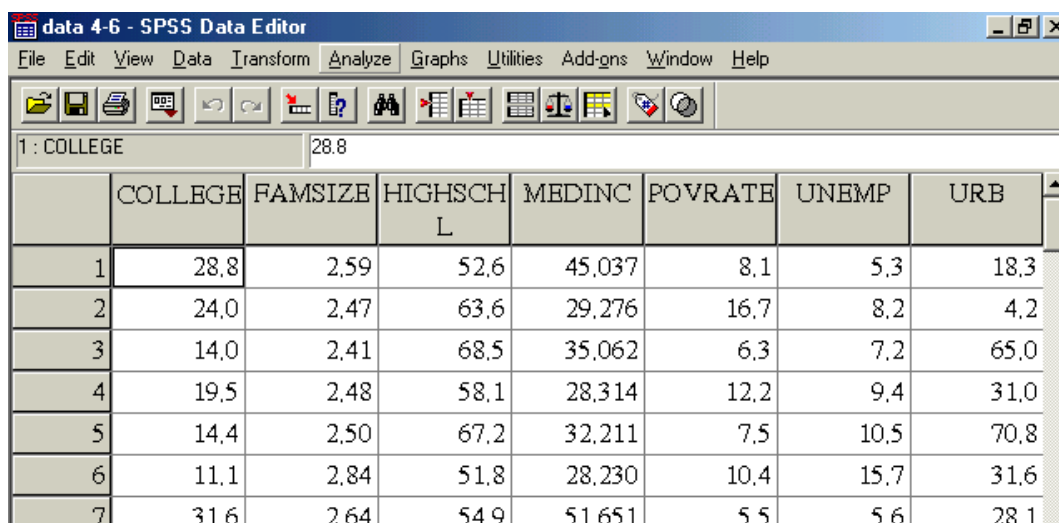
Vào **Transform, Recode, Into Different Variables**. Tức là chúng ta sẽ mã hóa lại biến **loai_xe**, và sẽ cho ra một biến mới (nếu chọn **Into Same Variables** thì SPSS sẽ biến đổi rồi thay thế luôn thông tin của biến cũ).

HÌNH 6



Lấy dữ liệu từ file DATA4-6 của Ramanathan.

HÌNH 8



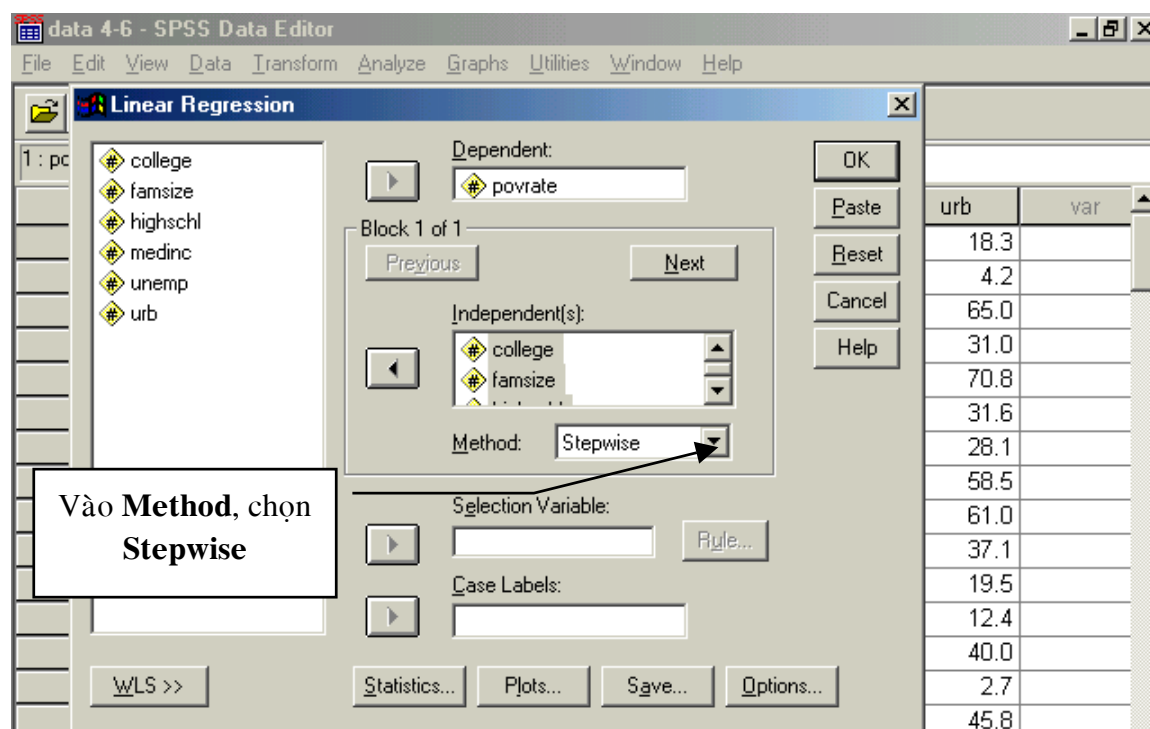
	COLLEGE	FAMSIZE	HIGHSCH L	MEDINC	POVRATE	UNEMP	URB
1	28,8	2,59	52,6	45,037	8,1	5,3	18,3
2	24,0	2,47	63,6	29,276	16,7	8,2	4,2
3	14,0	2,41	68,5	35,062	6,3	7,2	65,0
4	19,5	2,48	58,1	28,314	12,2	9,4	31,0
5	14,4	2,50	67,2	32,211	7,5	10,5	70,8
6	11,1	2,84	51,8	28,230	10,4	15,7	31,6
7	31,6	2,64	54,9	51,651	5,5	5,6	28,1

Bây giờ chúng ta sẽ hồi quy OLS kết hợp với phương pháp **Stepwise** với biến phụ thuộc là POVRATE, biến độc lập là tất cả các biến còn lại trong dữ liệu.

Tác dụng của phương pháp **Stepwise** được hiểu nôm na là giúp chúng ta tìm ra được những kết hợp của các biến độc lập sao cho kết quả hồi quy sẽ “tốt” theo hướng các giá trị thống kê t , F có ý nghĩa, và việc lựa chọn các kết hợp này sẽ được căn cứ vào khả năng làm gia tăng giá trị của R^2 .

Để bắt đầu, vào Menu **Analyze, Regression, Linear** rồi đưa biến POVRATE và ô **Dependent** và các biến còn lại vào **Independent(s)**.

HÌNH 9



Kết quả hồi quy được trình bày như sau:

Bảng 1: Trình bày thông tin cho biết SPSS đã tìm ra được bao nhiêu kết hợp tốt theo thống kê t và F. Đồng thời, các mô hình xuất hiện sau sẽ có giá trị R^2 và R^2 hiệu chỉnh lớn hơn mô hình xuất hiện trước (xem bảng 2).

Variables Entered/Removed(a)

Model	Variables Entered	Variables Removed	Method
1	MEDINC	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	HIGHSCHL	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	FAMSIZE	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
4	COLLEGE	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: POVRATE

Bảng 2:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.782(a)	.612	.605	2.4870
2	.895(b)	.800	.793	1.7999
3	.903(c)	.816	.805	1.7445
4	.912(d)	.831	.818	1.6870

a Predictors: (Constant), MEDINC

b Predictors: (Constant), MEDINC, HIGHSCHL

c Predictors: (Constant), MEDINC, HIGHSCHL, FAMSIZE

d Predictors: (Constant), MEDINC, HIGHSCHL, FAMSIZE, COLLEGE

Bảng 3:

ANOVA(e)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	545.424	1	545.424	88.181	.000(a)
	Residual	346.376	56	6.185		
	Total	891.799	57			
2	Regression	713.626	2	356.813	110.144	.000(b)
	Residual	178.173	55	3.240		
	Total	891.799	57			
3	Regression	727.461	3	242.487	79.679	.000(c)
	Residual	164.338	54	3.043		
	Total	891.799	57			
4	Regression	740.961	4	185.240	65.088	.000(d)
	Residual	150.838	53	2.846		
	Total	891.799	57			

a Predictors: (Constant), MEDINC

b Predictors: (Constant), MEDINC, HIGHSCHL

c Predictors: (Constant), MEDINC, HIGHSCHL, FAMSIZE

d Predictors: (Constant), MEDINC, HIGHSCHL, FAMSIZE, COLLEGE

e Dependent Variable: POVRATE

Bảng 4: Các hệ số hồi quy và thống kê t

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	23.131	1.446		15.997	.000
	MEDINC	-.374	.040	-.782	-9.390	.000
2	(Constant)	41.849	2.801		14.943	.000
	MEDINC	-.435	.030	-.909	-14.475	.000
	HIGHSCHL	-.288	.040	-.452	-7.206	.000
3	(Constant)	31.775	5.449		5.831	.000
	MEDINC	-.421	.030	-.880	-14.131	.000
	HIGHSCHL	-.235	.046	-.369	-5.111	.000
	FAMSIZE	2.434	1.141	.148	2.132	.038
4	(Constant)	19.172	7.826		2.450	.018
	MEDINC	-.552	.067	-1.154	-8.284	.000
	HIGHSCHL	-.139	.063	-.218	-2.214	.031
	FAMSIZE	5.414	1.758	.329	3.079	.003
	COLLEGE	.195	.090	.380	2.178	.034

a Dependent Variable: POVRATE

Bảng 5: Các biến bị bỏ ra trong quá trình chạy hồi quy

Excluded Variables(e)

Model		Beta In	T	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	COLLEGE	.157(a)	.998	.323	.133	.281
	FAMSIZE	.339(a)	4.809	.000	.544	.999
	HIGHSCHL	-.452(a)	-7.206	.000	-.697	.921
	UNEMP	.342(a)	3.082	.003	.384	.490
	URB	-.094(a)	-1.133	.262	-.151	.993
2	COLLEGE	-.038(b)	-.324	.747	-.044	.266
	FAMSIZE	.148(b)	2.132	.038	.279	.708
	UNEMP	.071(b)	.733	.467	.099	.386
	URB	-.010(b)	-.155	.878	-.021	.955
3	COLLEGE	.380(c)	2.178	.034	.287	.105
	UNEMP	-.055(c)	-.485	.630	-.066	.270
	URB	-.114(c)	-1.620	.111	-.217	.666
4	UNEMP	.025(d)	.212	.833	.029	.242
	URB	-.091(d)	-1.296	.201	-.177	.645

a Predictors in the Model: (Constant), MEDINC

b Predictors in the Model: (Constant), MEDINC, HIGHSCHL

c Predictors in the Model: (Constant), MEDINC, HIGHSCHL, FAMSIZE

d Predictors in the Model: (Constant), MEDINC, HIGHSCHL, FAMSIZE, COLLEGE

e Dependent Variable: POVRATE

Tuy nhiên, việc lựa chọn mô hình thích hợp cho nghiên cứu còn phụ thuộc vào nhiều yếu tố khác nữa, phương pháp Stepwise chỉ là một cách giúp có thể chúng ta tiết kiệm thời gian hay gợi ra một ý tưởng về việc kết hợp các biến độc lập trong quá trình hồi quy. Nếu như chúng ta chưa nắm vững về hồi quy bội thì sẽ không phát huy được tiện ích của phương pháp này.