



PHÂN TÍCH ĐỊNH LƯỢNG

NỘI DUNG BÀI GIẢNG

- Một số khái niệm cơ bản
- Kiểm định thang đo
- Phân tích nhân tố
- Phân tích hồi quy

PHƯƠNG PHÁP NGHIÊN CỨU ĐỊNH LƯỢNG

Mục đích
nghiên
cứu định
lượng

Các câu
hỏi và giả
thiết
trong
nghiên
cứu định
lượng

Sử dụng
lý thuyết
trong
nghiên
cứu định
lượng

Ví dụ về
nghiên
cứu có sử
dụng
phương
pháp định
lượng

Phát biểu mục đích trong nghiên cứu định lượng

Nghiên cứu định lượng là điều tra thực nghiệm có hệ thống về các hiện tượng quan sát được qua số liệu thống kê toán học hoặc kỹ thuật vi tính.
(Internet)

Phát biểu mục đích trong nghiên cứu định lượng



Mục tiêu nghiên cứu định lượng là để phát triển và sử dụng mô hình toán học, lý thuyết hoặc các giả thuyết liên quan đến các hiện tượng.

Phát biểu mục đích trong nghiên cứu định lượng



Nghiên cứu định lượng được xác định qua nhiều thuật ngữ biến số, so sánh và mối quan hệ của các biến ảnh hưởng đến nhau.

Phát biểu mục đích trong nghiên cứu định lượng



Biến số là những đại lượng hay đặc tính có thể thay đổi từ người này sang người khác hay từ thời điểm này sang thời điểm khác.

Câu hỏi và giả thuyết trong nghiên cứu định lượng

Câu hỏi nghiên cứu là gì?



- Là những lời phát biểu nghi vấn hay câu hỏi mà nhà điều tra/ nghiên cứu cố gắng trả lời
- Sử dụng: Trong nghiên cứu KH - XH

Câu hỏi và giả thuyết trong nghiên cứu định lượng

Giả thuyết trong nghiên cứu là gì?



- Là những tiên đoán đưa ra về những mối quan hệ của các biến.
- Là những ước lượng bằng số của tổng thể, dựa trên dữ liệu thu thập từ các mẫu của tổng thể
- Sử dụng trong các thí nghiệm so sánh các nhóm

Sử dụng lý thuyết trong nghiên cứu định lượng

Định nghĩa
một lý
thuyết?

Lý thuyết là “một tập hợp các cấu trúc khái niệm có tương quan với nhau, các định nghĩa và những lời xác nhận hay lời tuyên bố mà trình bày một quan điểm có hệ thống về các hiện tượng bằng cách nêu rõ những mối quan hệ giữa các biến, với mục đích là giải thích hiện tượng tự nhiên” - Kerlinger (1979)

Các lý thuyết hình thành khi các nhà nghiên cứu kiểm định 1 lời tuyên đoán nhiều lần, trong nhiều môi trường

Các lý thuyết được tìm thấy trong các ngành thuộc lĩnh vực khoa học và xã hội



Sử dụng lý thuyết trong nghiên cứu định lượng

Hình thức của các lý thuyết

Trình bày lý thuyết dưới hình thức các giả thuyết liên kết với nhau

Phát biểu một lý thuyết như là một chuỗi những lời phát biểu “nếu...thì”, giải thích tại sao người ta kỳ vọng các biến độc lập ảnh hưởng đến/ gây ra các biến phụ thuộc

Trình bày lý thuyết dưới mô hình trực quan

Sử dụng lý thuyết trong nghiên cứu định lượng

Nhà nghiên cứu kiểm định hay xác minh một lý thuyết

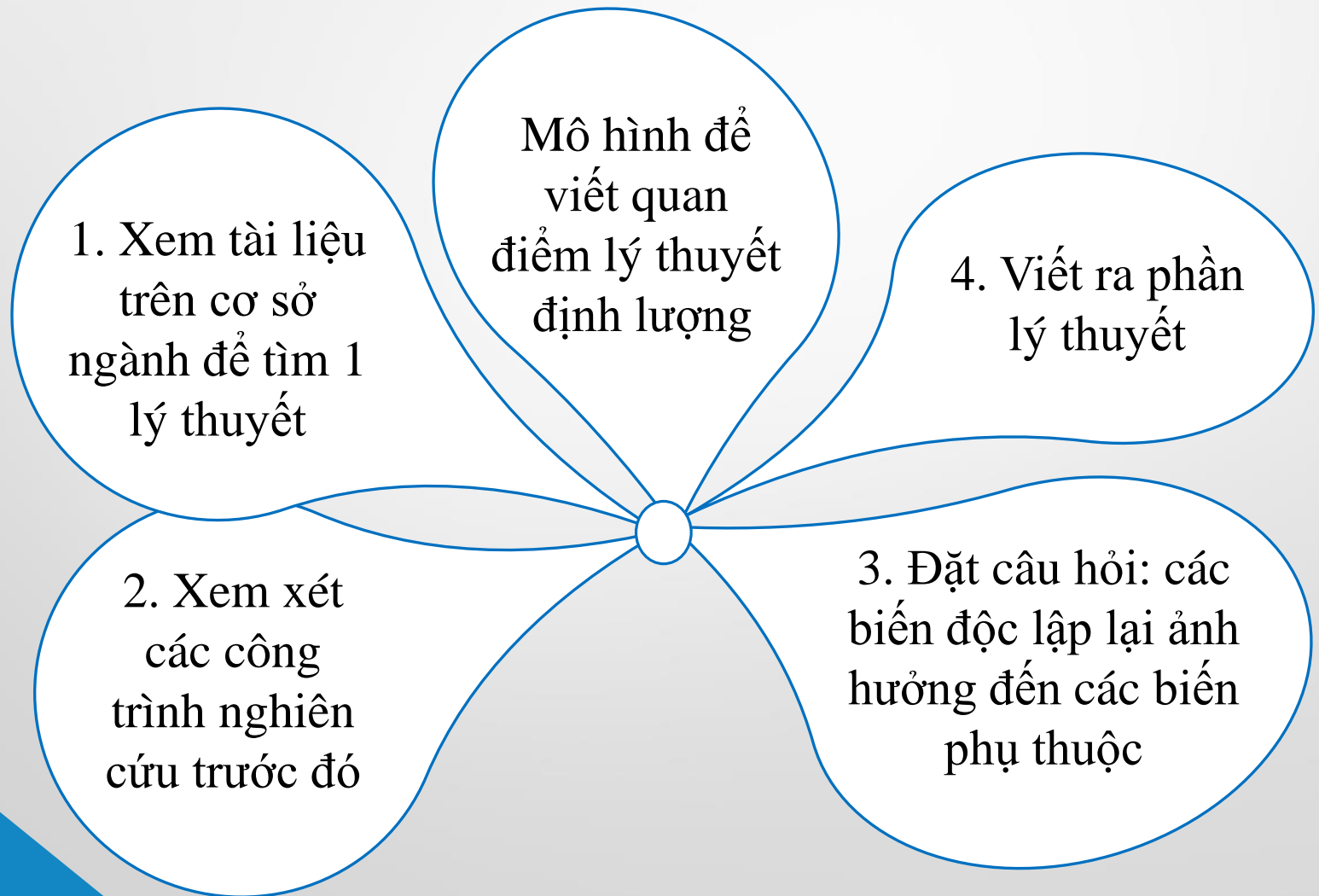
Nhà nghiên cứu kiểm định các giả thuyết hay câu hỏi từ nghiên cứu lý thuyết

Nhà nghiên cứu định nghĩa các biến và đưa các biến vào hoạt động, các biến này được rút ra từ lý thuyết nói trên

Nhà nghiên cứu đo lường hay quan sát các biến bằng cách sử dụng một công cụ để thu nhận những số điểm

**Cách
tiếp cận
suy
diễn**

Sử dụng lý thuyết trong nghiên cứu định lượng



Phân tích dữ liệu định lượng

1. Giới thiệu các loại biến trong mô hình
2. Mô hình EFA
3. Dữ liệu cho phân tích nhân tố
4. Phương pháp phân tích EFA trên SPSS

Giới thiệu các loại biến trong mô hình

Biến quan sát

- Là biến có thể ước lượng, đo lường được
- Hình dạng quy ước

Biến quan
sát

Biến tiềm ẩn

- Là biến được đo lường thông qua các biến quan sát
- Hình dạng quy ước

Biến tiềm ẩn

Khái niệm và Đo lường

- Việc đo lường một số khái niệm có thể không phức tạp về phương pháp ví dụ như: mức thu nhập, mức chi tiêu, thời gian xem truyền hình.
- Một số khái niệm phức tạp, trừu tượng đòi hỏi cần có quá trình chi tiết hóa khái niệm (construct operationalization) và thiết kế đo lường (measurement design) và kiểm tra kỹ lưỡng. Ví dụ:
 - Ý định hướng tới hành vi tiêu dùng sản phẩm/dịch vụ;
 - Mức độ hài lòng của nhân viên (employee satisfaction).

Đo lường và thang đo Likert

- Một trong những hình thức đo lường được sử dụng phổ biến nhất trong nghiên cứu kinh tế xã hội là thang đo Likert, được Rensis Likert (1932) giới thiệu. Loại thang đo này có 5 mức độ phổ biến.
- Phương pháp của Likert: Đưa ra một danh sách các khía cạnh có thể đo lường cho một khái niệm và tìm ra những tập hợp các mục hỏi để đo lường tốt các khía cạnh khác nhau của khái niệm. Nếu như khái niệm mang tính đơn khía cạnh thì chỉ cần tìm ra một tập hợp. Nếu khái niệm đó là đa khía cạnh thì cần nhiều tập hợp các mục hỏi

Thang đo đơn hướng và đa hướng

BẢNG 3.3 Kết quả phân tích nhân tố EFA của khái niệm “chất lượng dịch vụ đào tạo”

Biến quan sát	Các nhân tố chính	Trọng số	% biến thiên giải thích được	Cronbach α
F1	Hoạt động đào tạo		33.849	0.726
CL_1	Chương trình đào tạo phù hợp tốt với yêu cầu của thực tiễn.	0.600		
CL_2	Nội dung môn học được đổi mới, đáp ứng tốt yêu cầu đào tạo.	0.620		
CL_3	Phương pháp giảng của GV phù hợp với yêu cầu của từng môn học.	0.652		
CL_4	Giảng viên có kiến thức sâu về môn học đảm trách.	0.673		
CL_5	Cách đánh giá và cho điểm sinh viên công bằng.	0.583		
CL_6	Tổ chức thi cử, giám thị coi thi nghiêm túc.	0.565		
F2	Cơ sở vật chất		7.377	0.746
CL_8	Cơ sở vật chất trường đáp ứng tốt nhu cầu đào tạo và học tập.	0.639		
CL_9	Phòng máy tính đáp ứng tốt nhu cầu thực hành của sinh viên.	0.680		
CL_10	Cơ sở vật chất thư viện tốt.	0.798		
CL_11	Nhân viên thư viện phục vụ tốt.	0.698		
F3	Dịch vụ hỗ trợ và phục vụ		9.166	0.811
CL_13	Dịch vụ y tế đáp ứng tốt sinh viên có nhu cầu.	0.645		
CL_14	Tư vấn đáp ứng tốt nhu cầu chọn lựa và học tập của sinh viên.	0.718		
CL_15	Dịch vụ tài chính hỗ trợ tốt sinh viên có nhu cầu.	0.782		
CL_17	Dịch vụ ăn uống giải khát phù hợp với nhu cầu sinh viên.	0.638		
CL_19	Nhân viên giáo vụ, thanh tra nhiệt tình phục vụ sinh viên.	0.567		
CL_20	Nhà trường và khoa thường xuyên lắng nghe ý kiến sinh viên.	0.579		

Các bước xây dựng thang đo Likert

1. Nhận diện và đặt tên biến/khái niệm muốn đo lường
2. Lập ra một danh sách các phát biểu/ câu hỏi để biểu thị. Có thể lấy từ lý thuyết có liên quan, đọc sách báo, ý kiến chuyên gia, thực nghiệm.
3. Xác định loại trả lời: đồng ý – không đồng ý; ủng hộ – phản đối; phù hợp – không phù hợp; đúng – không đúng...
4. Số lượng mức độ luôn là số lẻ: 3, 5 hay 7 mức độ.
5. Kiểm tra toàn bộ các mục hỏi bằng cách khảo sát thử
6. Phân tích mục hỏi trong danh sách để tìm ra một tập hợp các mục hỏi giúp đo lường được một khía cạnh của khái niệm/biến muốn nghiên cứu trong mô hình.

Phân tích các mục hỏi

Nhằm tìm ra và giữ lại những mục hỏi có ý nghĩa giúp đo lường được một khía cạnh của khái niệm nghiên cứu.

1. Xác định điểm các câu trả lời (nhớ chú ý những câu đối nghĩa cần mã hóa lại).
2. Kiểm tra tương quan giữa các mục hỏi (Cronbach's alpha).
3. Kiểm tra vai trò của từng mục hỏi (Item-total correlation).

Bước 2: Tương quan giữa các mục hỏi

Là kiểm tra xem các mục hỏi liên quan chặt chẽ với nhau đến đâu. Điều này liên quan đến hai phép tính toán:

- Tương quan giữa bản thân các mục hỏi
- Tương quan của điểm số của từng mục hỏi với điểm số toàn bộ các mục hỏi cho mỗi người trả lời (alpha).

Bước 2 Tính toán Cronbach's Alpha

- Hệ số Cronbach's alpha: là con số thể hiện mức độ các mục hỏi tương quan chặt chẽ với nhau tới mức nào (Xác định mối tương quan với biến tổng)
- Yêu cầu: Hệ số alpha $\geq 0,6$.

$$\alpha = \frac{K}{K-1} \left(1 - \frac{\sum_{i=1}^K \sigma_{Xi}^2}{\sigma_Y^2} \right)$$

K : Là số biến đưa vào phân tích.

σ_Y^2 : Phương sai của biến tổng

σ_{Xi}^2 : Phương sai của biến quan sát thứ i

Bước 3 Kiểm tra vai trò từng mục hỏi

Mục tiêu: Cho biết mục hỏi nào cần được bỏ đi và mục hỏi nào cần được giữ lại.

Các biến quan sát có hệ số tương quan biến tổng (item total correlation) nhỏ hơn 0,3 sẽ bị loại bỏ



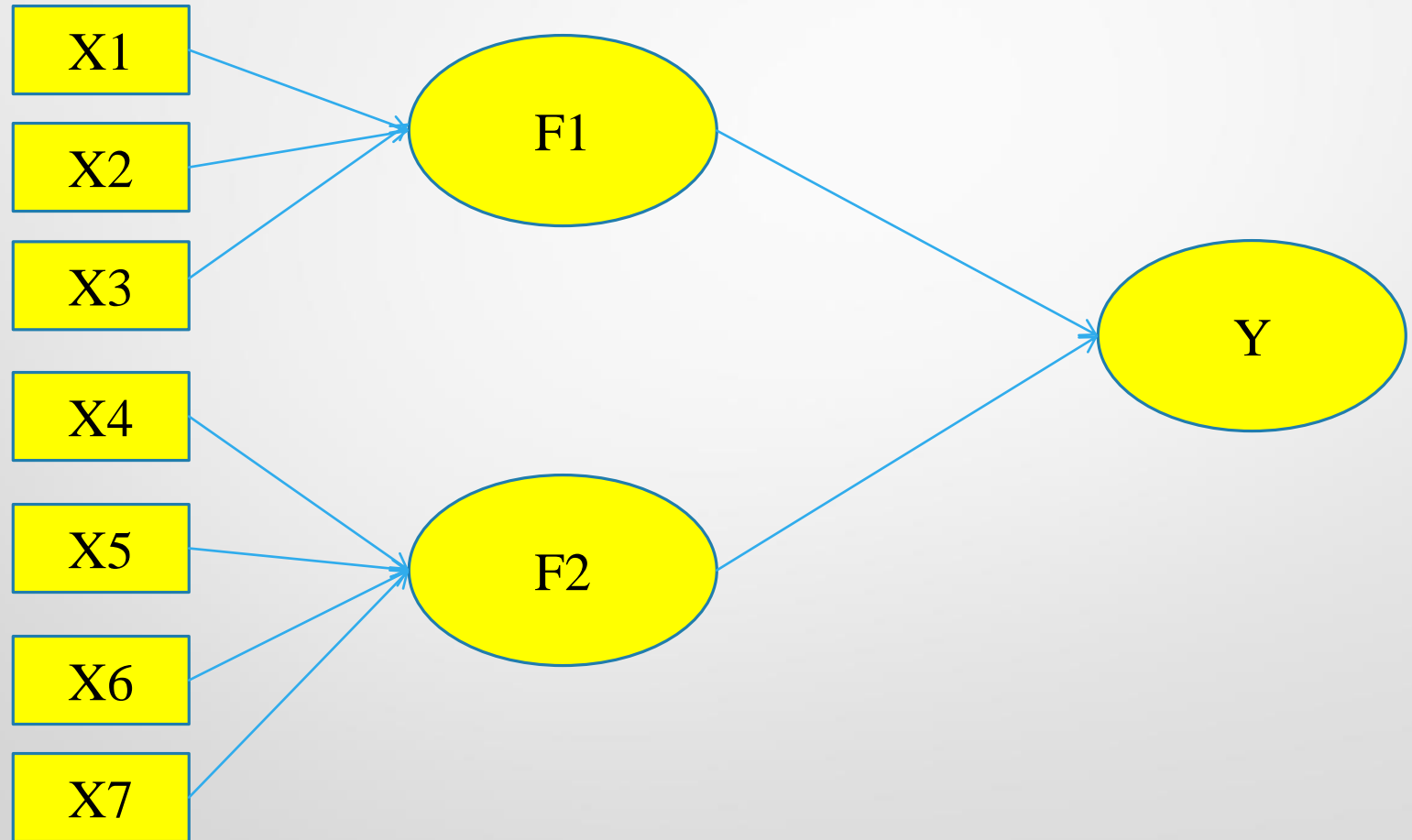
PHÂN TÍCH NHÂN TỐ

Factor Analysis

Khái niệm và ứng dụng

- Phân tích nhân tố là một nhóm các thủ tục được sử dụng chủ yếu để thu gọn và tóm tắt các dữ liệu.
 - Trong nghiên cứu và phân tích, người nghiên cứu có thể thu thập được một số lượng biến khá lớn và hầu hết các biến này có liên hệ với nhau
- ⇒ Cần giảm bớt số lượng biến đến mức người nghiên cứu có thể sử dụng được.

Mô hình phân tích nhân tố khám phá (EFA)



Dữ liệu cho phân tích nhân tố

- Phần thông tin cơ bản: ...
- Phần dữ liệu cho phân tích nhân tố: ...

Các tham số thống kê trong EFA

- Điều kiện áp dụng EFA: các biến có tương quan với nhau
 - ⇒ Barlett test of sphericity: kiểm định có tương quan hay không
 - ⇒ Giả thuyết H_0 : không có tương quan giữa các biến quan sát.

	v1	v2	v3	v4	v5	v6
v1	1					
v2	0	1				
v3	0	0	1			
v4	0	0	0	1		
v5	0	0	0	0	1	
V6	0	0	0	0	0	1

Các tham số thống kê trong EFA

- Hệ số Kaiser-Mayer-Olkin (KMO): KMO thích hợp khi $0,5 \leq KMO \leq 1$

=> Các tương quan đủ lớn đến mức có thể áp dụng EFA.

$$KMO_j = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} a_{ij}^{2*}} \quad KMO = \frac{\sum_{i \neq j} \sum_{j \neq i} r_{ij}^2}{\sum_{i \neq j} \sum_{j \neq i} r_{ij}^2 + \sum_{i \neq j} \sum_{j \neq i} a_{ij}^{2*}}$$

where a_{ij}^* is the anti-image correlation coefficient.

Các tham số thống kê trong EFA

- Correlation matrix (ma trận tương quan): ma trận chứa tất cả các hệ số tương quan cặp giữa các cặp biến trong phân tích.

	v1	v2	v3	v4	v5	v6
v1	1	0.039	0.321	0	0.314	-0.097
v2	0.039	1	-0.13	0.534	0.352	0.593
v3	0.321	-0.13	1	-0.432	0.474	0.037
v4	0	0.534	-0.432	1	0.077	0.345
v5	0.314	0.352	0.474	0.077	1	0.279
v6	-0.097	0.593	0.037	0.345	0.279	1

Các tham số thống kê trong EFA

- Communality (phần chung): lượng biến thiên của 1 biến được giải thích chung với các biến khác (cũng là phần biến thiên được giải thích bởi các nhân tố chung).
 - Eigenvalue: phần biến thiên được giải thích bởi mỗi nhân tố so với biến thiên toàn bộ. nếu phần biến thiên được giải thích này lớn thì nhân tố rút ra có ý nghĩa tóm tắt thông tin tốt.
- ⇒ Eigenvalue lớn hơn 1 thì nhân tố rút ra có ý nghĩa tóm tắt thông tin tốt

Xoay các nhân tố

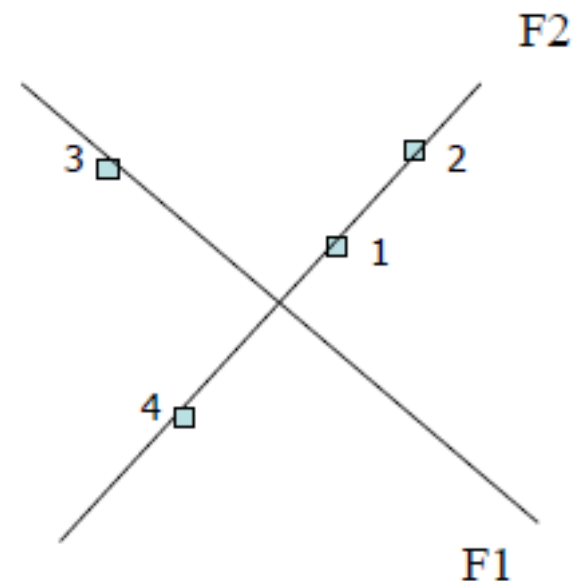
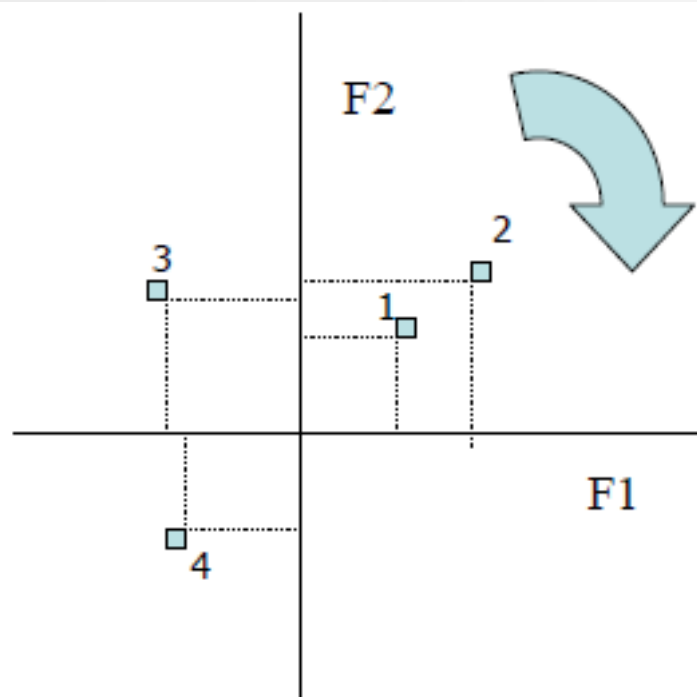
- Mỗi biến gốc cần có hệ số tải nhân tố lớn (từ 0,3 trở lên) đối với chỉ một nhân tố được rút ra.
- Thỉnh thoảng có một vài biến có hệ số lớn đối với hơn một nhân tố hoặc có nhiều nhân tố có hệ số lớn trong cùng một biến

⇒ Việc giải thích sẽ trở nên khó khăn.

⇒ Xoay nhân tố

Có hai thủ tục cơ bản Oblique, Orthogonal.

Xoay nhân tố



	Factor 1	Factor 2
x1	0.5	0.5
x2	0.8	0.8
x3	-0.7	0.7
x4	-0.5	-0.5

	Factor 1	Factor 2
x1	0	0.6
x2	0	0.9
x3	-0.9	0
x4	0	-0.9

Đặt tên và giải thích các nhân tố

- Việc giải thích các nhân tố được thực hiện trên cơ sở nhận ra các biến có hệ số (factor loading) lớn ở cùng một nhân tố.
- Chúng ta có thể tóm tắt các dữ liệu thu thập được theo đặc điểm của nhóm nhân tố chúng ta đưa ra.

Nhân số - Factor score

- Phân tích nhân tố trên SPSS cho thể cho ta tính được nhân số (factor score) nhân tố mới gồm:
 1. Nhân số được chuẩn hóa
 2. Nhân số không chuẩn hóa

Cách tính Factor Score chuẩn hóa

- Các nhân tố có thể được diễn tả như những kết hợp tuyến tính của các item.

$$F_i = W_{i1}X_1 + W_{i2}X_2 + W_{i3}X_3 + \dots + W_{ik}X_k$$

F_i : Giá trị ước lượng cho trị số của nhân tố thứ i ,
($i = 1, m$) với m là số nhân tố EFA rút ra.

W_{ik} : Trọng số nhân tố (factor score coefficient)
 k là số item cấu thành nhân tố, ($k = 1, n$)

Các giá trị của các X_k là đã được chuẩn hóa.

Nhân số không chuẩn hóa

- Nhân số (factor score) không chuẩn hóa có thể được dùng để tính giá trị trung bình, phân tích T_test, ANOVA)

Cách tính Factor Score chưa chuẩn hóa

Từ Factor score coefficient matrix (ma trận trọng số nhân tố), viết được phương trình thể hiện từng nhân tố như là kết hợp của các biến gốc.

	F1	F2
ngua sau rang	-0.012	0.341
lam trang rang	0.411	0.039
lam khoe nuu rang	-0.116	0.471
lam hoi tho thom tho	0.356	-0.172
lam sach cau rang	0.171	0.420
lam rang bong hon	0.355	0.065

$$F_1 = -0,012X_1 + 0,411X_2 - 0,116X_3 + 0,356X_4 + 0,171X_5 + 0,355X_6$$

$$F_2 = 0,341X_1 + 0,039X_2 + 0,471X_3 - 0,172X_4 + 0,420X_5 + 0,065X_6$$

EFA và Cronbach's Alpha

Tính cái nào trước?

- Chú ý trong trường hợp nghiên cứu lặp lại hay nhà nghiên cứu sử dụng scale đã được chứng minh trong các nghiên cứu trước thì trước tiên tính Cronbach's Alpha cho từng tập biến đo các khía cạnh, sau khi Cronbach đạt mới sang EFA.
- Trong trường hợp nghiên cứu mà thang đo được xây dựng lần đầu tiên, khi Nhà nghiên cứu chưa biết chính xác có bao nhiêu thành phần trong thang đo đó thì cần làm EFA trước để xem xét sau đó Cronbach để đánh giá chất lượng

Thực hành phân tích trên SPSS

- Dùng dữ liệu bài tập *phan tich nhan to.sav*
- Quy trình phân tích:
 1. Thực hiện trên SPSS
 2. Giải thích bảng kết quả
 3. Tính các nhân số (nhân tố - Factor score)
 4. Đánh giá thang đo
 5. Các phân tích khác sau khi phân tích nhân tố
 6. Ứng dụng thêm

Đọc bảng kết quả

- Những bảng kết quả quan trọng
 1. Bảng hệ số KMO (kaiser – Maiyer – Olkin)
 2. Bảng Component matrix
 3. Bảng Rotated component matrix
 4. Bảng Total variance explained

Đọc bảng kết quả - KMO

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.584
Bartlett's Test of Sphericity	Approx. Chi-Square	55.275
	df	15
	Sig.	.000

- **Mục đích:** xem xét mức độ thích hợp của EFA
- **Sử dụng:** KMO thích hợp khi $0,5 \leq KMO \leq 1$

Đọc bảng kết quả - Component matrix

Component Matrix^a

	Component	
	1	2
ngua sau rang	.050	.618
lam trang rang	.891	-.007
lam khoe nuu rang	-.143	.872
lam hoi tho thom tho	.726	-.377
lam sach cau rang	.462	.726
lam rang bong hon	.775	.050

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

- Mục đích: Xác định số nhân tố và biến quan sát giải thích cho nhân tố.
- Biến quan sát được chọn là biến có hệ số tải nhân tố $\geq 0,45$.

Đọc bảng kết quả Rotated component matrix

Rotated Component Matrix^a

	Component	
	1	2
ngua sau rang	-.014	.620
lam trang rang	.886	.086
lam khoe nuu rang	-.233	.852
lam hoi tho thom tho	.761	-.300
lam sach cau rang	.384	.770
lam rang bong hon	.766	.130

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

- **Mục đích:** Làm cho một biến quan sát chỉ có thể giải thích cho một nhân tố (factor) mà thôi.
- **Hậu quả:** Không làm biến đổi phương sai được giải thích bởi mô hình

Độc bảng kết quả Total variance explained

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.157	35.957	35.957	2.157	35.957	35.957	2.154	35.896	35.896
2	1.813	30.214	66.172	1.813	30.214	66.172	1.817	30.276	66.172
3	.912	15.206	81.378						
4	.490	8.168	89.546						
5	.350	5.829	95.375						
6	.278	4.625	100.000						

Khả
năng
giải
thích
của mô
hình

Extraction Method: Principal Component Analysis.

Kiểm định độ tin cậy Cronbach's Alpha

Reliability Analysis: Statistics

Descriptives for:

- ☐ Item
- ☐ Scale
- ☒ Scale if item deleted

Inter-Item:

- ☐ Correlations
- ☐ Covariances

Summaries:

- ☐ Means
- ☐ Variances
- ☐ Covariances
- ☐ Correlations

ANOVA Table:

- ☒ None
- ☐ F test
- ☐ Friedman chi-square
- ☐ Cochran chi-square

☐ Hotelling's T-square ☐ Tukey's test of additivity

☐ Intraclass correlation coefficient

Model: Two-Way Mixed Type: Consistency

Confidence interval: 95 % Test value: 0

Continue
Cancel
Help

Reliability Statistics

Cronbach's Alpha	N of Items
.737	3

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
lam trang rang	9.77	5.417	.689	.495
lam hoi tho thom tho	10.00	7.294	.484	.739
lam rang bong hon	11.03	5.146	.546	.690

Thực hành ví dụ khác

- Hãy thực hiện tương tự cho hai bài tập

1. Khảo sát nhân viên.sav

2. Chất lượng khoa cho thạc sĩ và sư hai long của học viên.sav

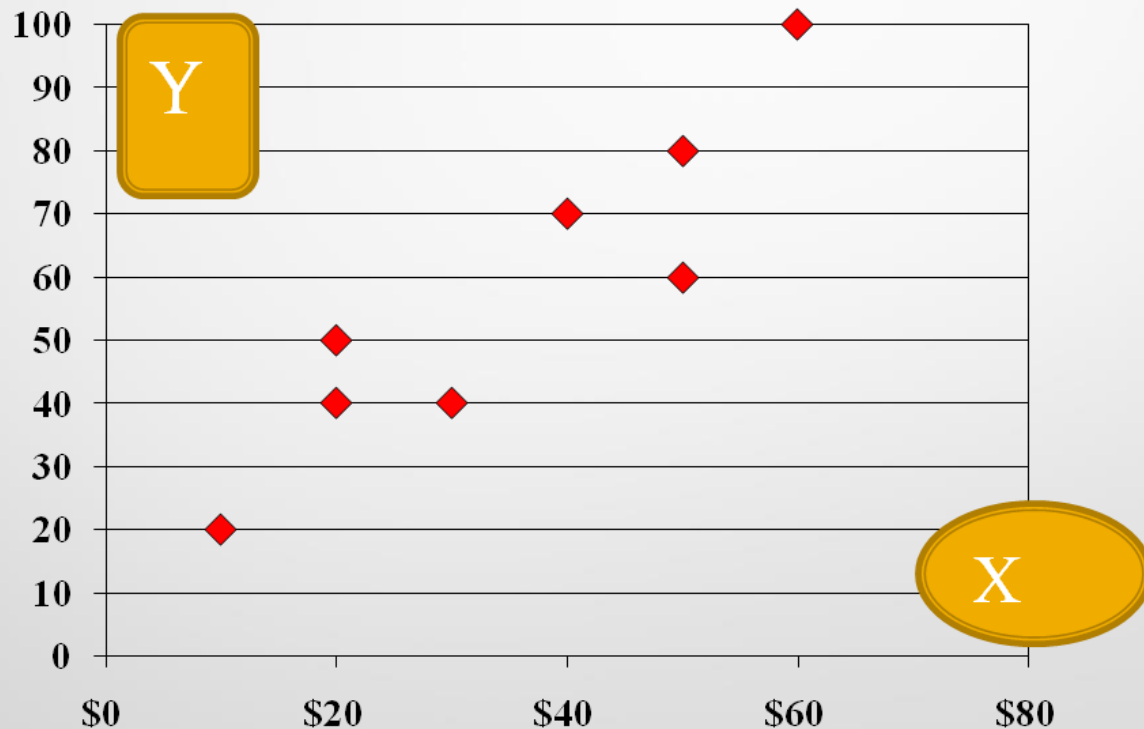
3. Long trung thành đối với sản phẩm sữa

PHÂN TÍCH HỒI QUY

1. Tương quan
2. Hồi quy
3. Quy trình xây dựng mô hình trên SPSS
4. Các loại kiểm định trong mô hình
5. Ý nghĩa hệ số hồi quy
6. Dự báo với mô hình hồi quy
7. Xử lí các lỗi của hồi quy

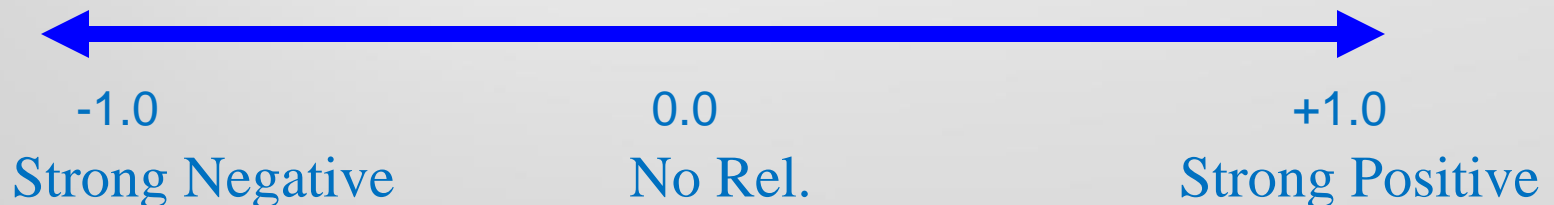
Correlation

- Là mối quan hệ tuyến tính giữa hai biến (X và Y) (r_{XY})

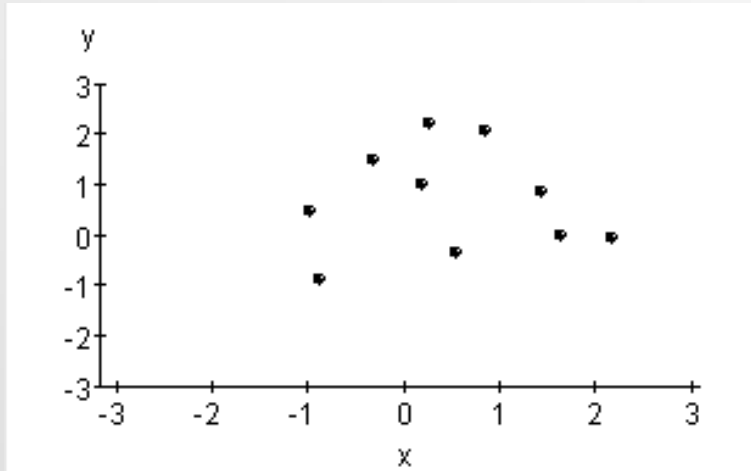


Hệ số tương quan

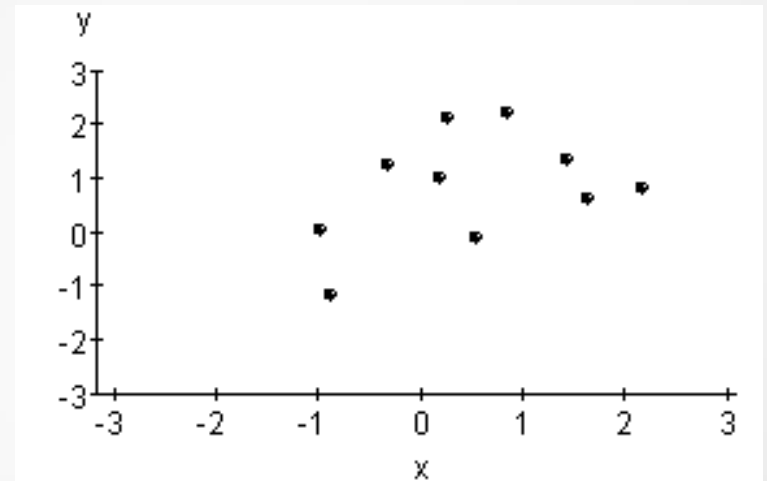
- “r”: Hệ số tương quan...
 - Độ mạnh của mối quan hệ (mạnh, yếu, hoặc không có quan hệ)
 - Các loại quan hệ
 - Đồng biến – X và Y biến thiên cùng chiều
 - Nghịch biến – X và Y biến thiên ngược chiều
- Khoảng biến thiên của r từ -1 đến 1



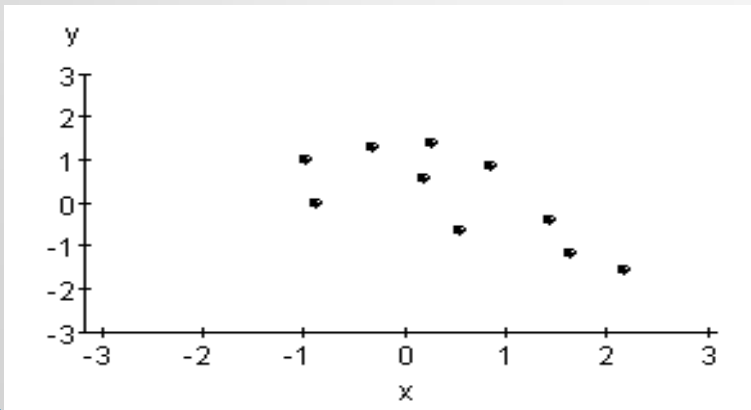
Thực hành với đồ thị phân tán



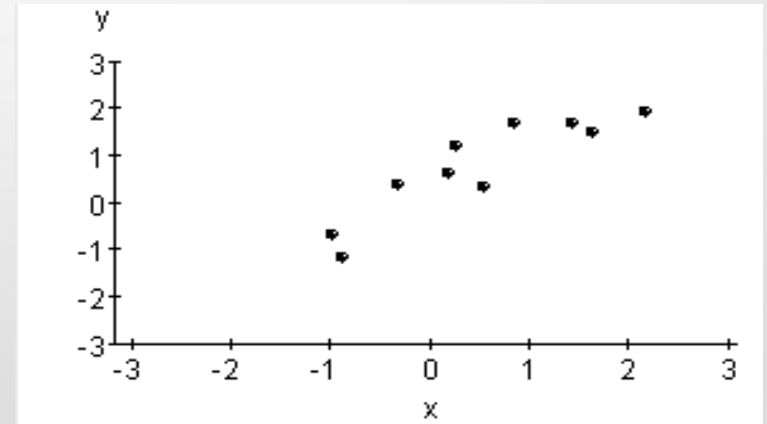
$r =$.__ __



$r =$.__ __

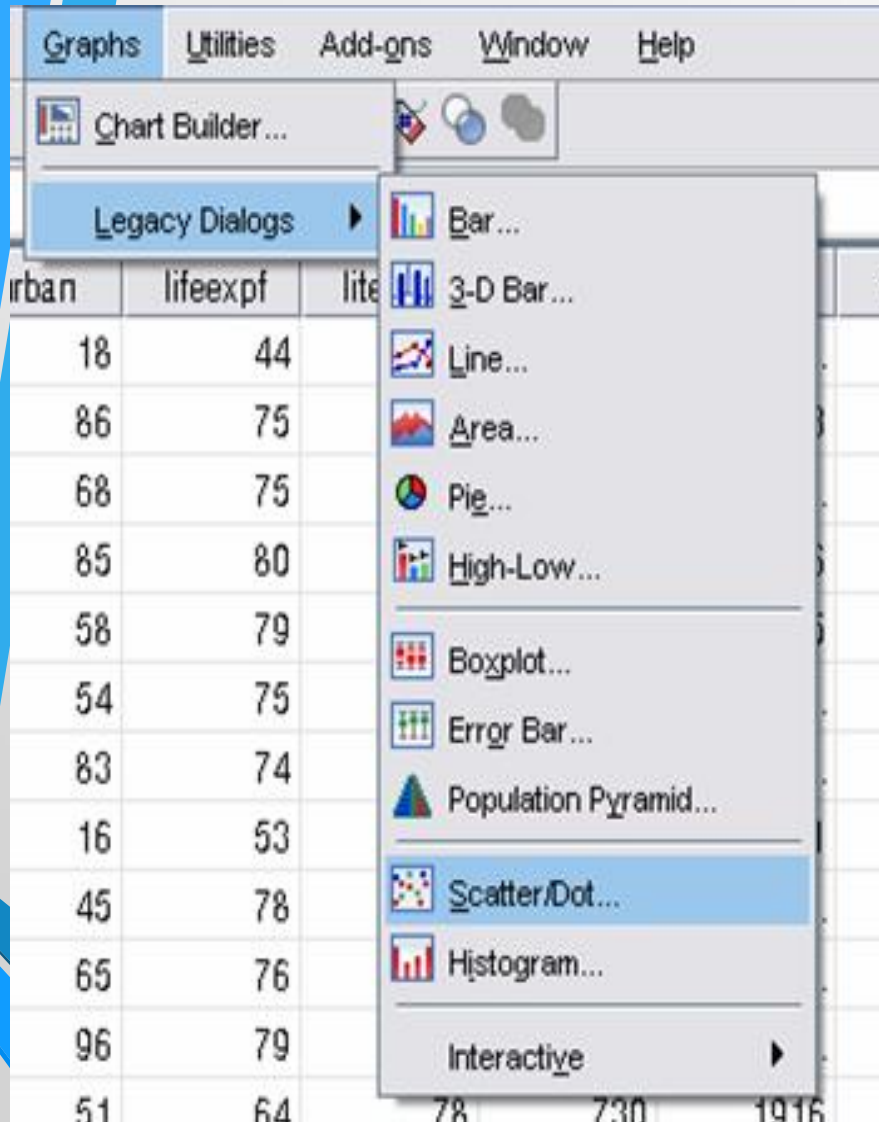


$r =$.__ __

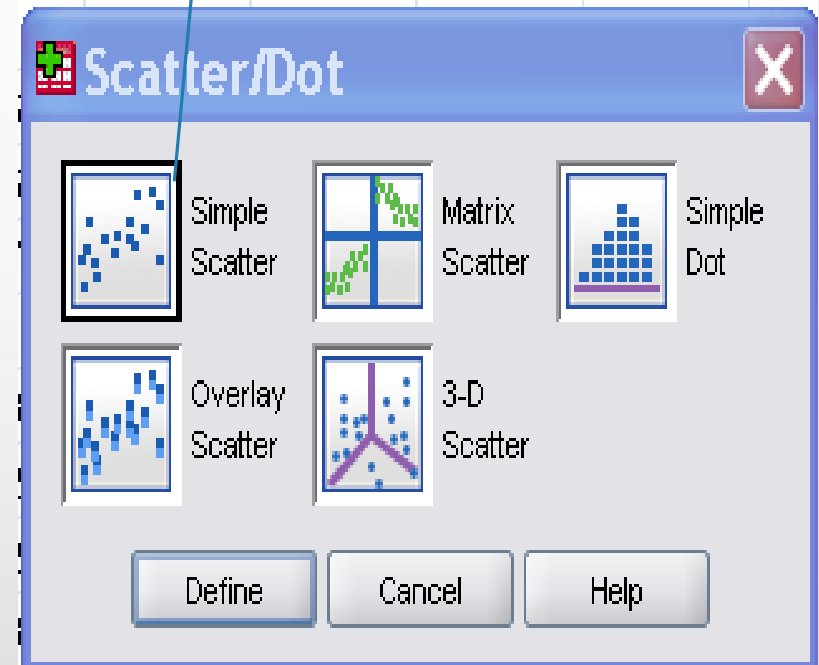


$r =$.__ __

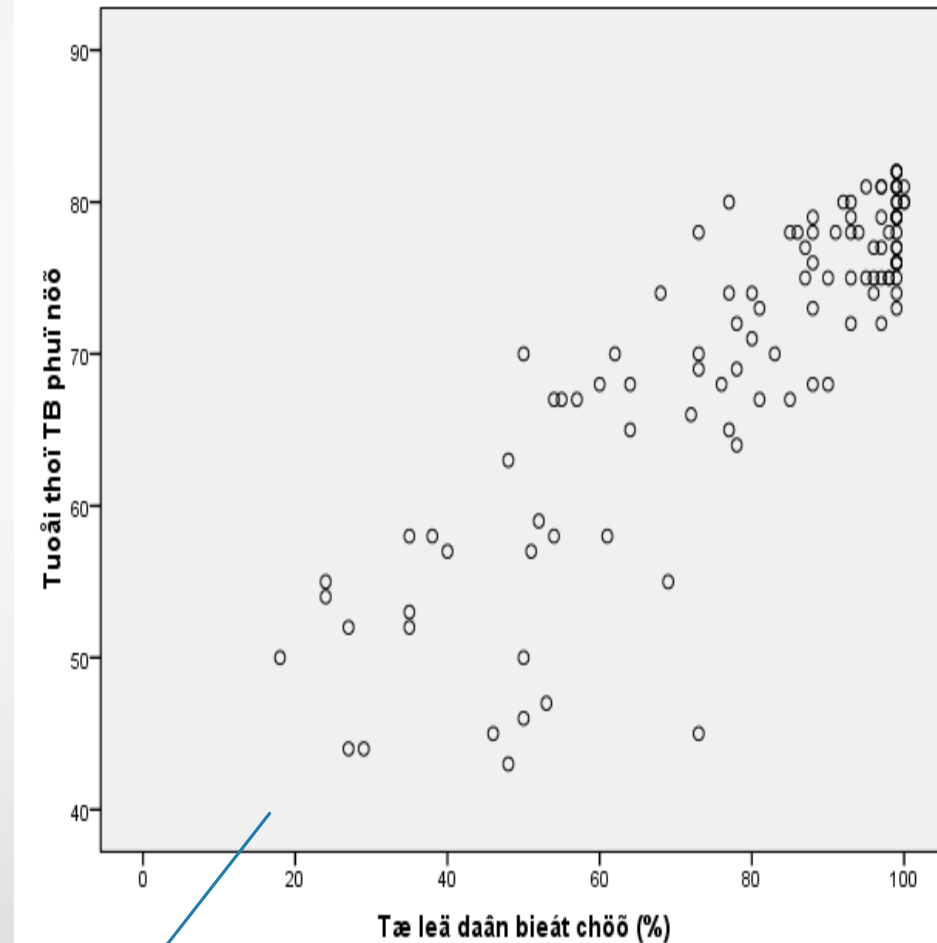
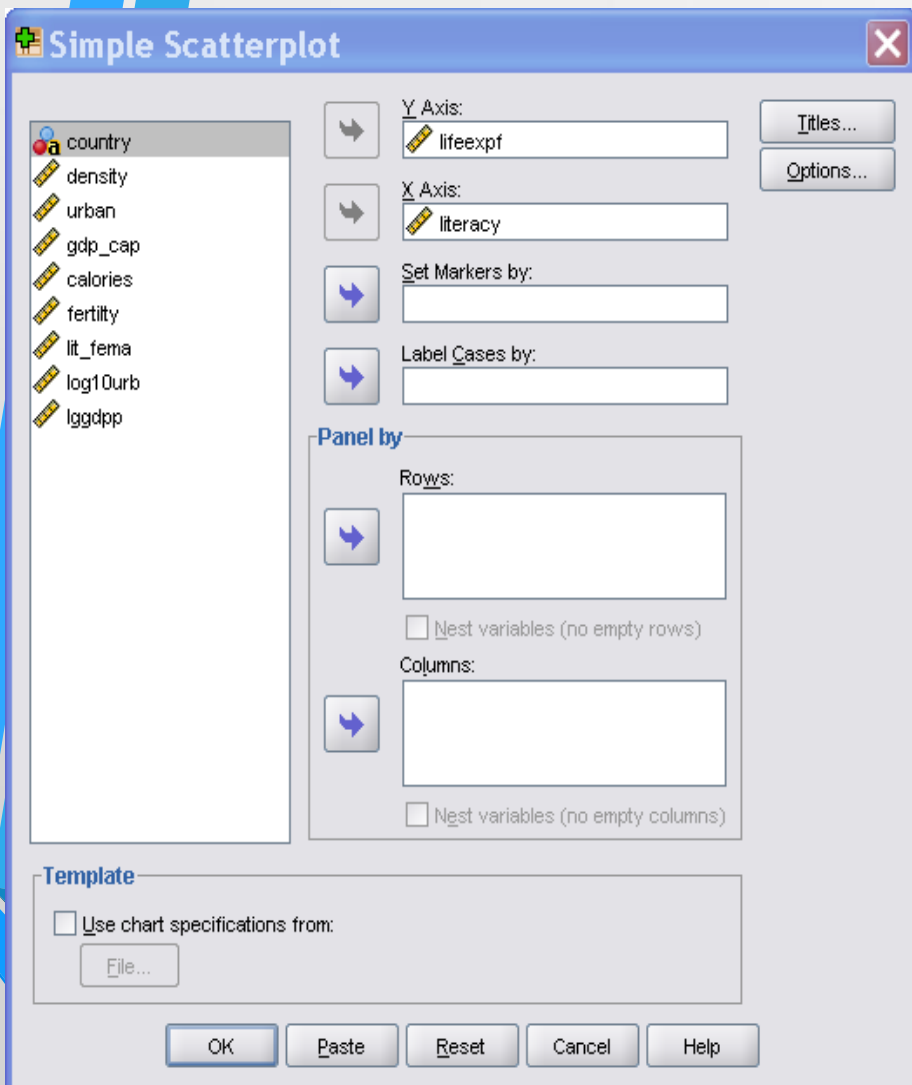
Thực hành trên SPSS



Click here



Thực hành trên SPSS



Nhận xét mối quan hệ

XÁC ĐỊNH HỆ SỐ TƯƠNG QUAN

Correlations

Statistics=Pearson Correlation

	Tuổi thọ TB phụ nữ	Mật độ dân số (người/km ²)	Tỉ lệ dân sống ở vùng đô thị (%)	Tỉ lệ dân biết chữ (%)	GDP tính trên đầu người (USD)	Calori nạp hàng ngày TB 1 người
Tuổi thọ TB phụ nữ	1	.128	.743**	.865**	.642**	.775**
Mật độ dân số (người/km ²)	.128	1	.223*	.031	.201*	.067
Tỉ lệ dân sống ở vùng đô thị (%)	.743**	.223*	1	.650**	.605**	.692**
Tỉ lệ dân biết chữ (%)	.865**	.031	.650**	1	.552**	.682**
GDP tính trên đầu người (USD)	.642**	.201*	.605**	.552**	1	.751**
Calori nạp hàng ngày TB 1 người	.775**	.067	.692**	.682**	.751**	1

** . Correlation is significant at the 0.01 level (2-tailed).

* . Correlation is significant at the 0.05 level (2-tailed).

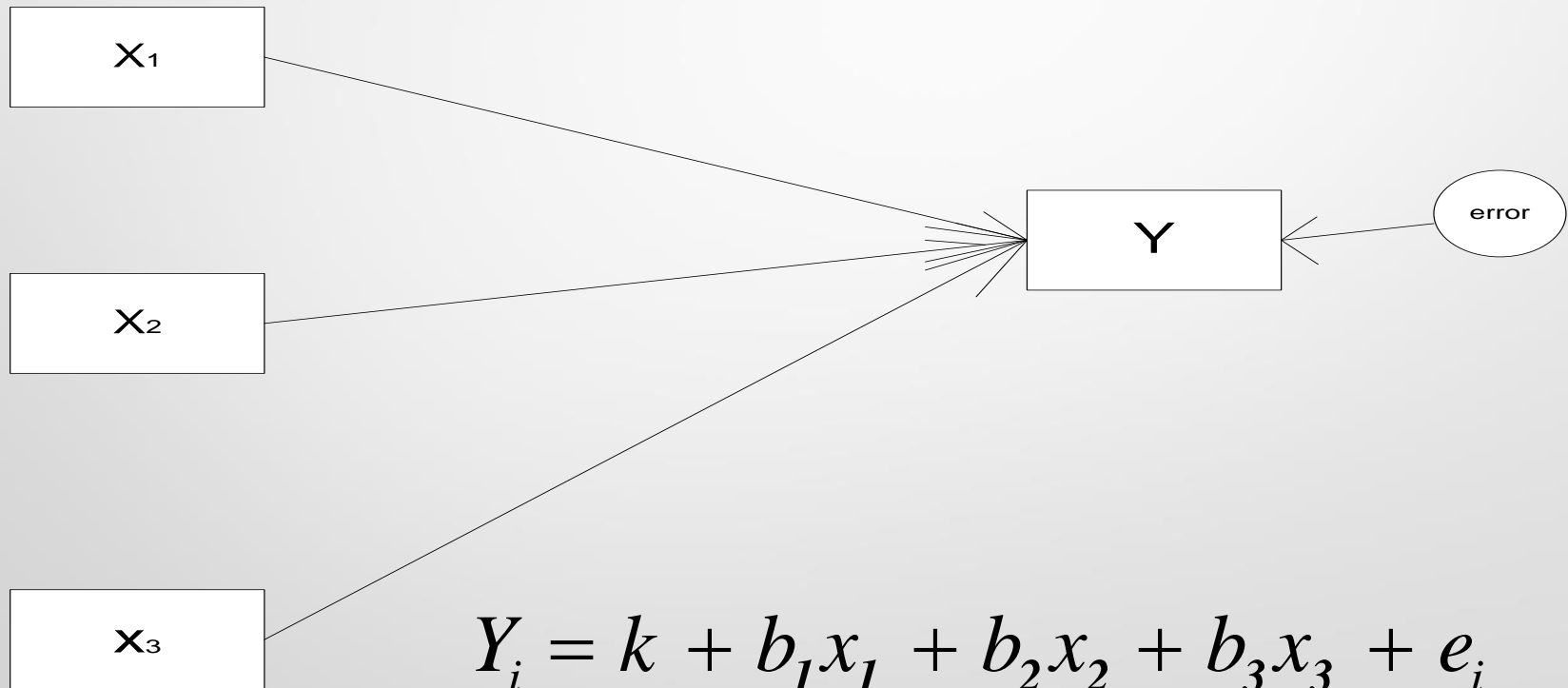
KIỂM ĐỊNH MỐI QUAN HỆ TUYẾN TÍNH

- Kiểm định mối quan hệ tuyến tính giữa các biến.
 - $H_0: r_{xy} = 0$: hai biến không có mối quan hệ tuyến tính phụ thuộc nhau
 - $H_1: r_{xy} \neq 0$: hai biến có mối quan hệ tuyến tính phụ thuộc nhau

	Hệ số tương quan Tuổi thọ TB phụ nữ	Giá trị sig Tuổi thọ TB phụ nữ
Tuổi thọ TB phụ nữ	1	.186
Mật độ dân số (người/km ²)	.128	1
Tỉ lệ dân sống ở vùng đô thị (%)	.743**	.000
Tỉ lệ dân biết chữ (%)	.865**	.000
GDP tính trên đầu người (USD)	.642**	.000
Calori nạp hàng ngày TB 1 người	.775**	.000

HỒI QUY TRỰC TIẾP

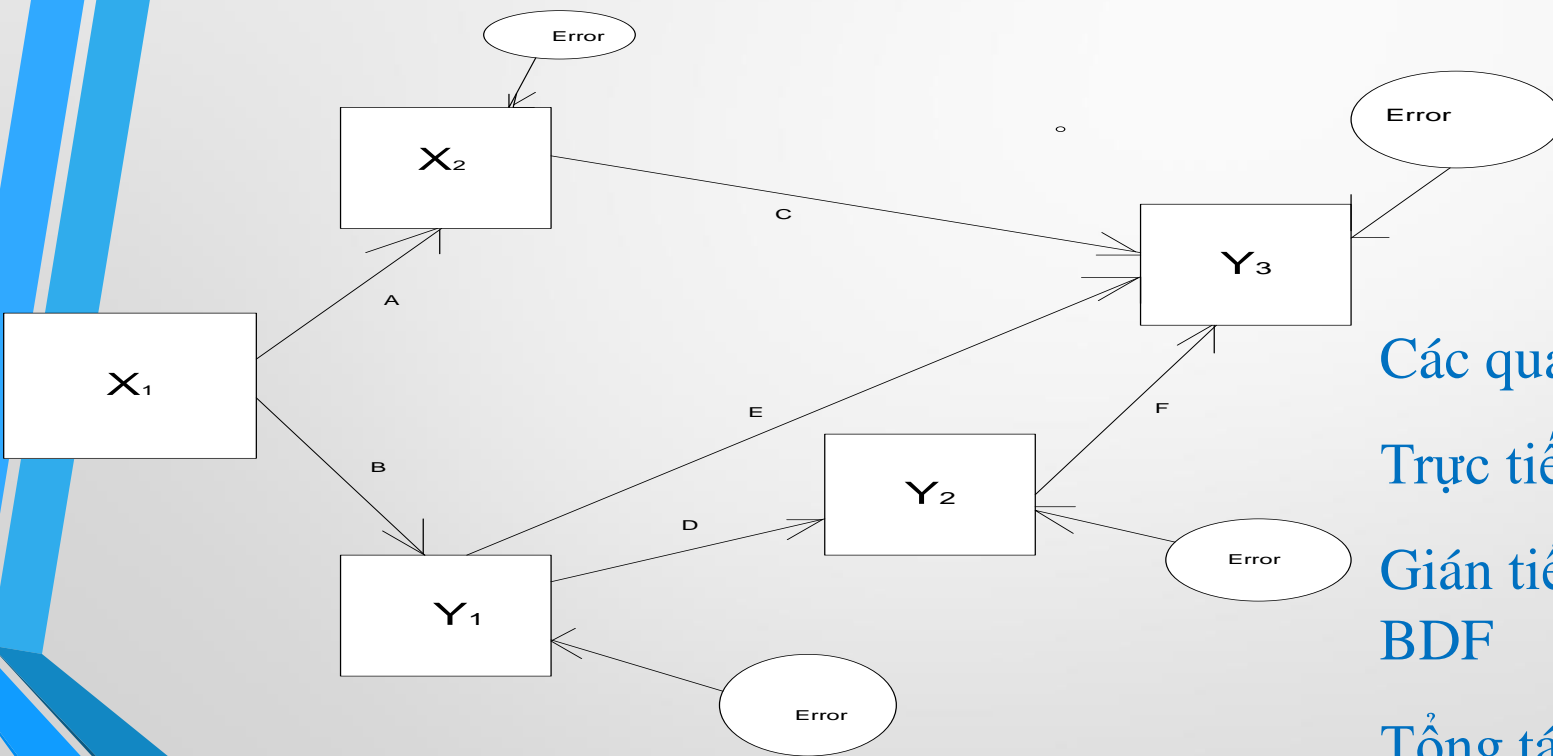
Path Diagram of A Linear Regression Analysis



HỒI QUY TỔNG HỢP

A Path Analysis

Decomposition of Effects into Direct, Indirect, Spurious, and Total Effects



Các quan hệ hồi quy

Trực tiếp Y_3 : C, E, F

Gián tiếp Y_3 : BE, BDF

Tổng tác động = Trực tiếp + gián tiếp

Direct Effects:
Paths C, E, F

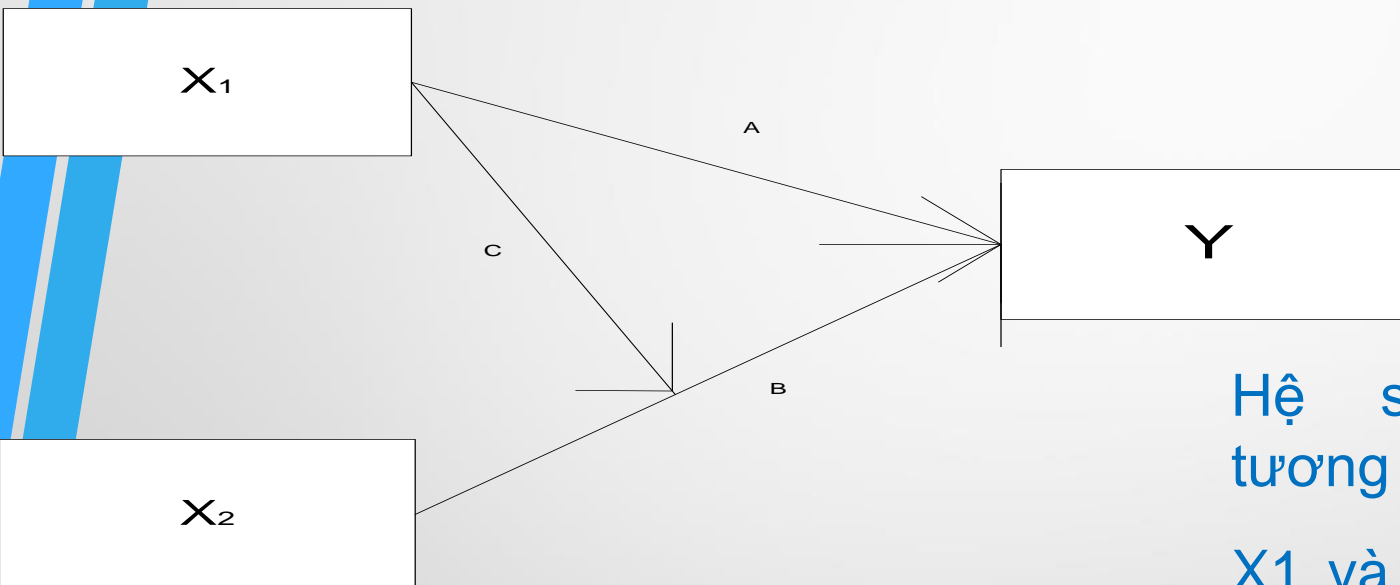
Indirect Effects:
Paths
AC, BE, DF

Total Effects:
Sum of Direct and
Indirect Effects

Spurious effects are due to
common (antecedent) causes

HỒI QUY TƯƠNG TÁC

Interaction Analysis



Hệ số tác động
tương tác: C

X_1 và X_2 tương tác
lẫn nhau cùng tác
động lên Y .

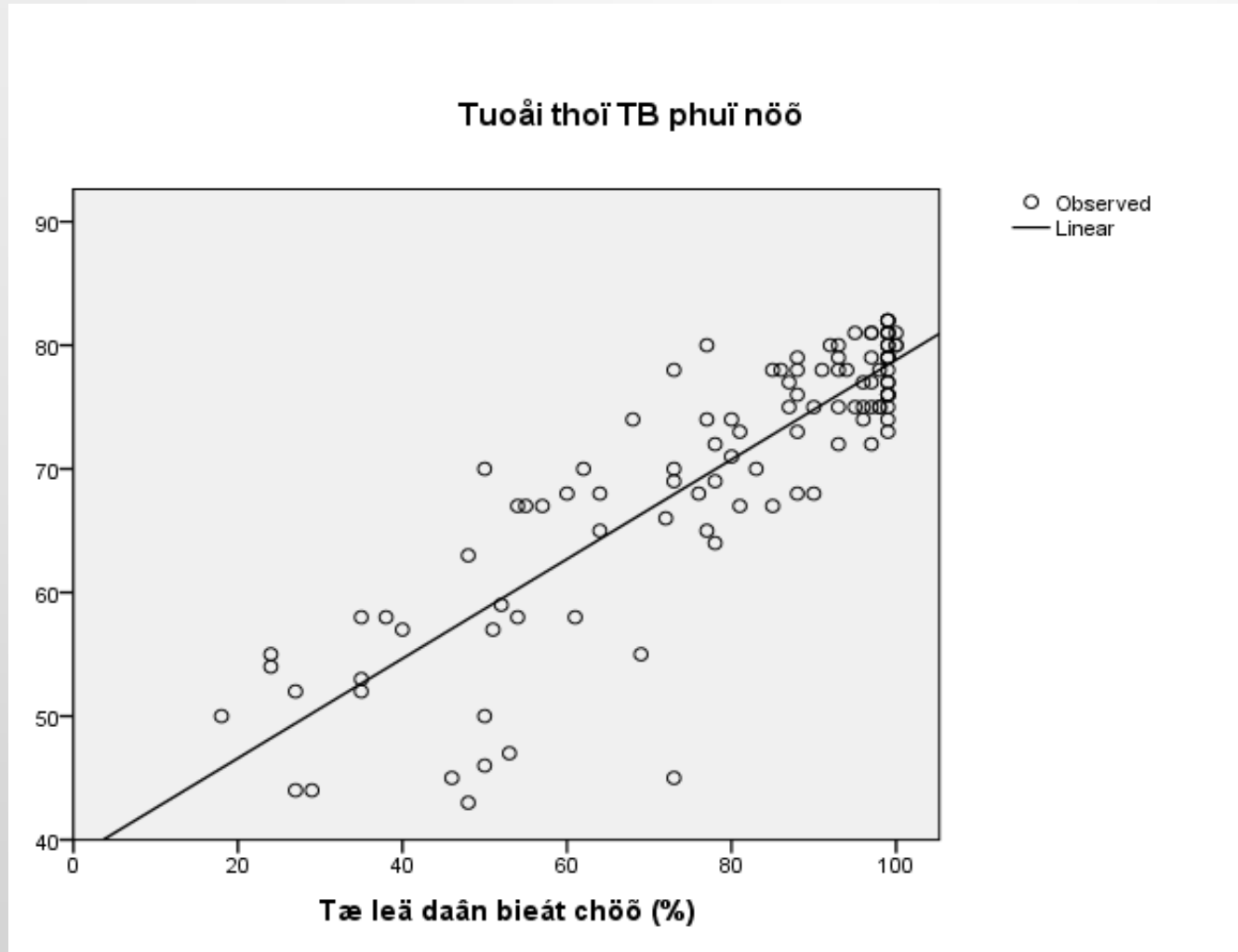
$$Y = K + aX_1 + BX_2 + CX_1 \cdot X_2$$

Các loại quan hệ giữa biến phụ thuộc và biến độc lập

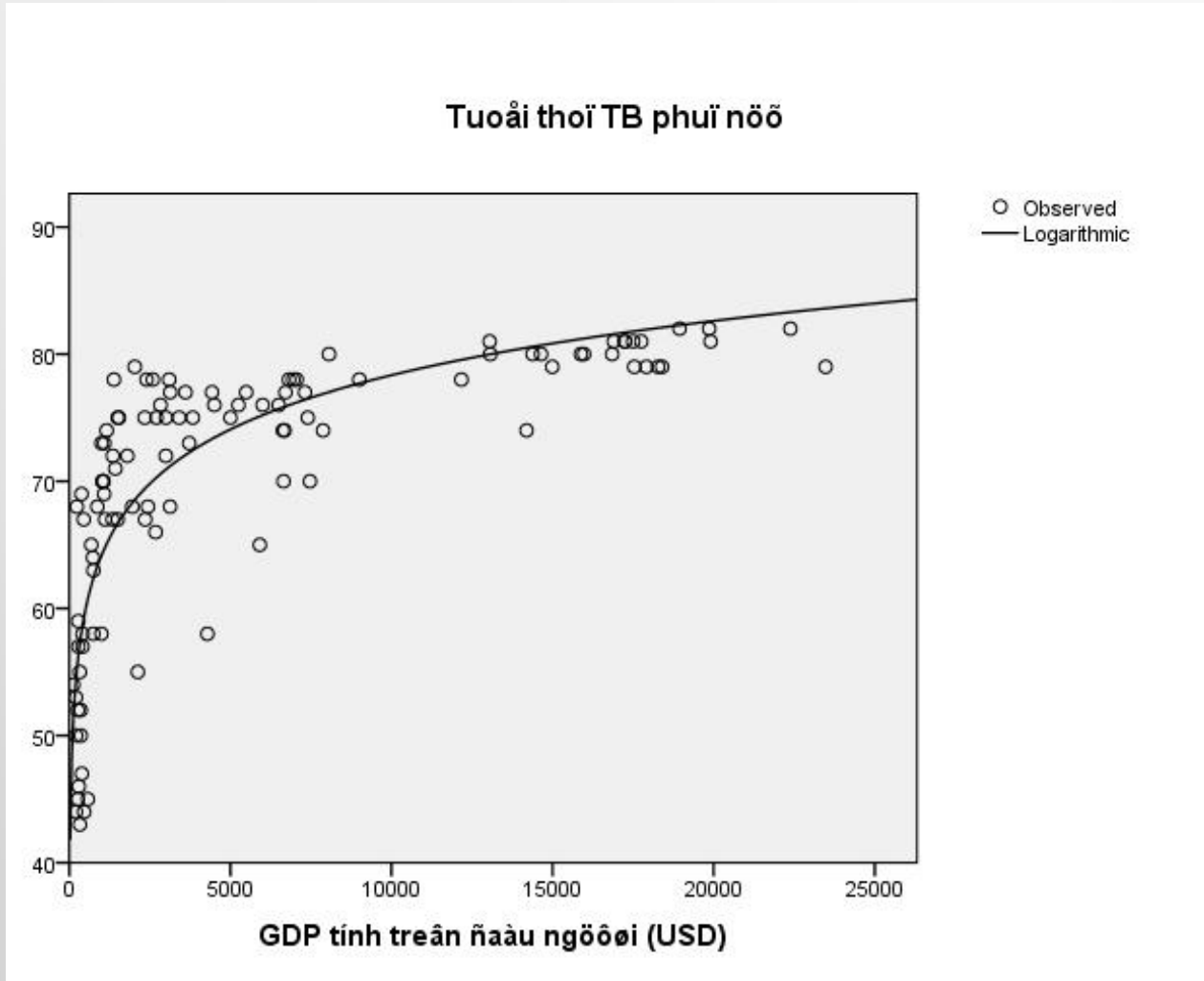
1. Quan hệ tuyến tính (linear)
2. Quan hệ logarithmic
3. Quan hệ hàm nghịch đảo (inverse)
4. Quan hệ parapol (quadratic)
5. Quan hệ hàm bậc 3 (cubic)
6. Quan hệ hàm mũ (Power)
7. Quan hệ logistic
8. Quan hệ hàm tăng trưởng (growth)
9. Quan hệ san bằng hàm mũ (exponential)

Hồi quy chỉ xét đối với hồi quy tuyến tính (đối với tham số).
Những mối quan hệ phi tuyến đều phải chuyển về quan hệ tuyến tính)

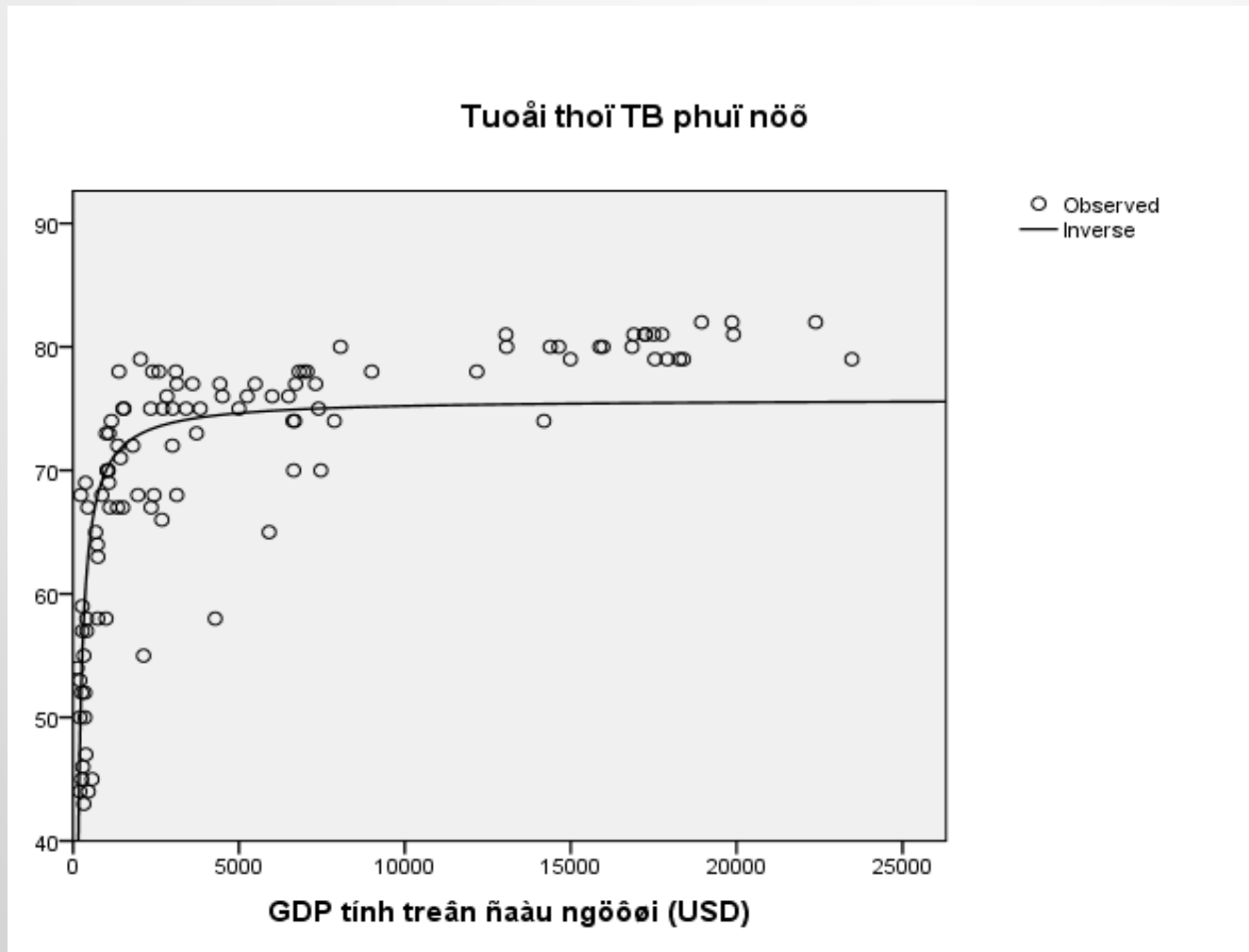
Mối quan hệ tuyến tính (linear)



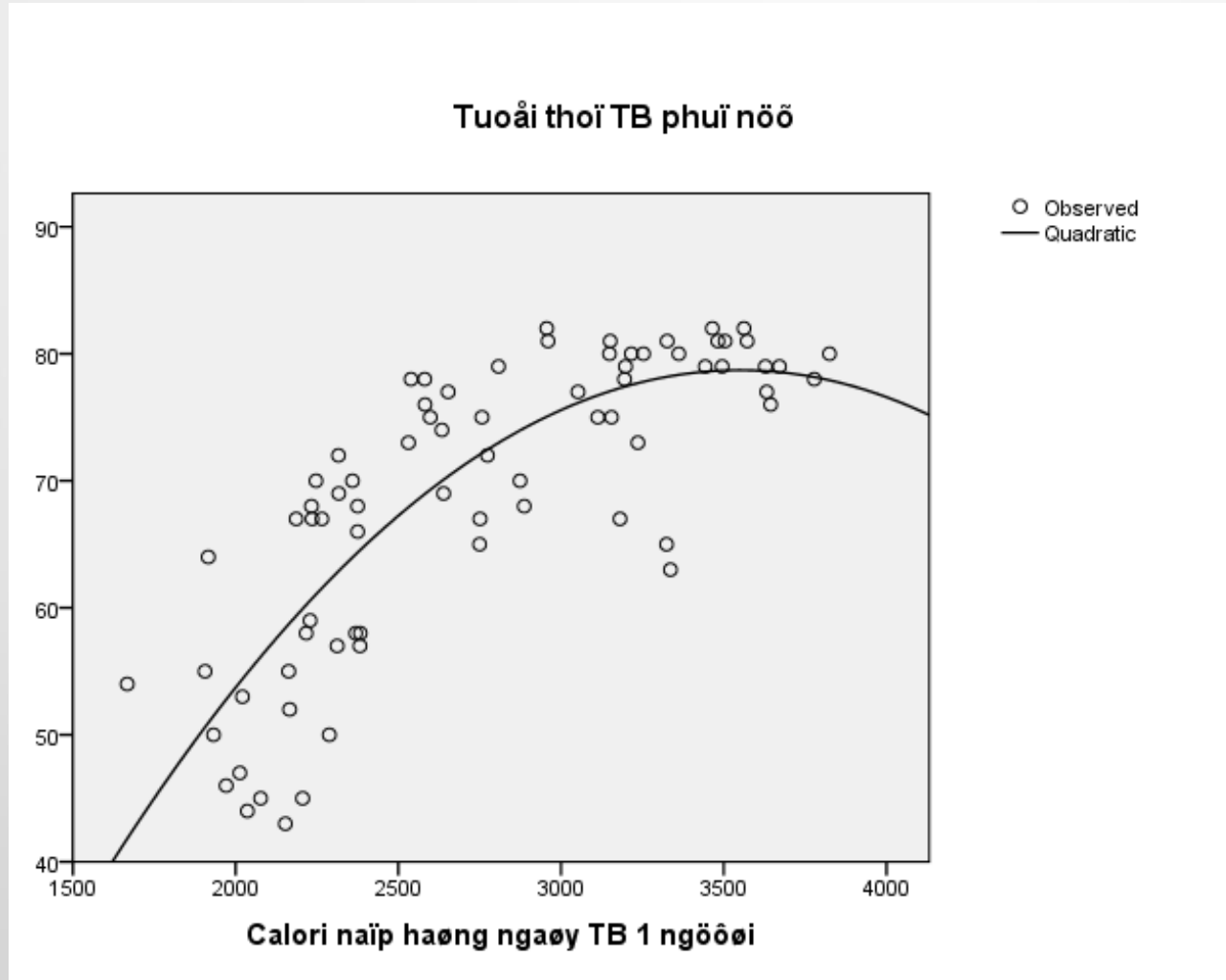
Quan hệ logarithmic



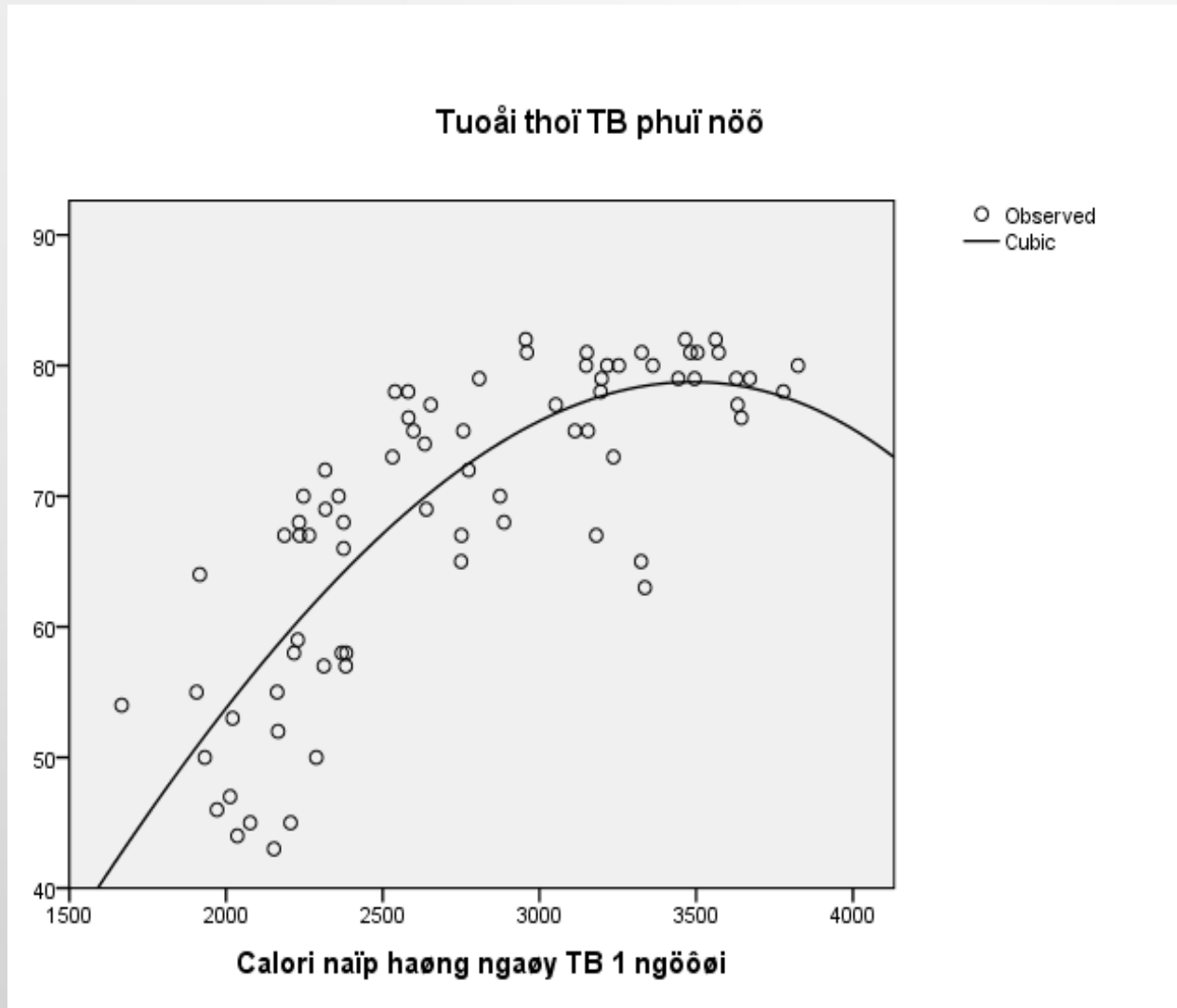
Quan hệ nghịch đảo(inverse – hypecpol)



Quan hệ hàm bậc hai (Quadratic)



Quan hệ hàm bậc 3 (cubic)



THỰC HÀNH: NGHIÊN CỨU CÁC NHÂN TỐ TÁC ĐỘNG ĐẾN MỨC LƯƠNG HIỆN TẠI

- BÀI TẬP: EMPLOYEE DATA.SAV.

Yêu cầu: Hãy xây dựng mô hình hồi quy mô tả những nhân tố (mối quan hệ) tác động đến mức lương hiện tại của người lao động trong công ty theo 2 mô hình sau.

- MH1: $\text{Lifeexpf} = a + b1 * \text{calories} + b2 * \text{gdp_gap}$
- MH2: $\text{Lifeexpf} = a + b1 * \text{calories} + b2 * \text{Ln}(\text{gdp_gap})$
- MH3: Xây dựng mô hình hồi quy với Lifeexpf là biến phụ thuộc và tất cả các biến còn lại là biến độc lập⁶⁵

Ý NGHĨA HỆ SỐ HỒI QUY

- Phương trình hồi quy mô hình 1 được viết như sau:

$$\text{Lifeexpf} = 32,77 + 0,012 * \text{calories} + 0 * \text{gdp_gap}$$

- B1: Trong điều kiện các nhân tố khác không đổi (2 nước giống nhau trừ lượng calories), nếu nước này có lượng calories nạp vào hàng ngày cao hơn nước kia 1 calories thì tuổi thọ của người dân nước này cao hơn nước kia là 0,012 tuổi.
- B2: Nếu chọn $\alpha=5\%$, ta đề xuất loại biến **gdp_gap** ra khỏi mô hình vì biến này không có tác động đến tuổi thọ

BẢNG KẾT QUẢ

Hệ số R^2 và R^2 hiệu chỉnh

- R^2 là khả năng giải thích của mô hình. Nếu $R^2 = 0,706 \Leftrightarrow 80,4\%$) thì mô hình có khả năng giải thích được 70,6% giá trị thực tế.
- **Chú ý: mô hình hồi quy đa biến độc lập (hồi quy bội) ta dùng R^2 hiệu chỉnh để nêu khả năng giải thích của mô hình. (69,8%)**

Model Summary

	Model
	1
R	.840
R Square	.706
Adjusted R Square	.698
Std. Error of the Estimate	6.275

DỰ BÁO BẰNG MÔ HÌNH HỒI QUY

DỰ BÁO CHO 3 NGƯỜI
CÓ ĐIỀU KIỆN SAU

Việt Nam	Gdp_gap	Calories
PA1	1000	2400
PA2	1200	2500
PA3	1500	2700

Nhập dữ liệu của 3
phương án trên vào quan
sát thứ 110-111-112

Tại hộp thoại **linear
regression - save**

Linear Regression: Save

Predicted Values

- ☒ Unstandardized
- ☐ Standardized
- ☐ Adjusted
- ☐ S.E. of mean predictions

Residuals

- ☐ Unstandardized
- ☐ Standardized
- ☐ Studentized
- ☐ Deleted
- ☐ Studentized deleted

Distances

- ☐ Mahalanobis
- ☐ Cook's
- ☐ Leverage values

Prediction Intervals

- ☐ Mean
- ☐ Individual

Confidence Interval: 95 %

Influence Statistics

- ☐ DfBeta(s)
- ☐ Standardized DfBeta(s)
- ☐ DfFit
- ☐ Standardized DfFit
- ☐ Covariance ratio

KẾT QUẢ DỰ BÁO

	country	lifeexpf	gdp_cap	calories	PRE_1
107	Venezuela	76	2829	2582	65.66015
108	Vietnam	68	230	2233	60.46090
109	Zambia	45	573	2077	58.64869
110	VietnamPA1	.	1000	2400	62.78769
111	VietnamPA2	.	1200	2500	64.09218
112	VietnamPA3	.	1500	2700	66.66714

THỰC HÀNH

Xây dựng mô hình 3 và dự báo tuổi thọ BQ của phụ nữ cho 3 phương án của Việt nam và năm 2008

	country	density	urban	lifeexpf	literacy	gdp_cap	calories	fertility	lit_fema
107	Venezuela	22.0	91	76	88	2829	2582	3.0	87
108	Vietnam	218.0	20	68	88	230	2233	3.3	83
109	Zambia	11.0	42	45	73	573	2077	6.7	65
110	Vietnampa1	250.0	25	.	90	1000	2400	2.0	90
111	Vietnampa2	300.0	27	.	92	1200	2500	3.0	92
112	Vietnampa3	350.0	30	.	95	1500	2700	4.0	95
113	Vietnam2008	270.0	28	.	95	1024	2700	3.0	90

GIẢ THIẾT CỦA MÔ HÌNH HỒI QUY BỘI

1. Các biến độc lập (giải thích) được biết trước
2. Các biến độc lập không tương quan với nhau
 $[\text{cov}(x_i, x_j) = 0]$
3. Các sai số giữa giá trị thực tế và giá trị dự báo (phần dư, resid, u_i) không tự tương quan với nhau
 $[\text{cov}(u_i, u_j) = 0]$
4. Phần dư (resid) có phương sai không đổi.
 $\text{Var}(\text{resid}) = \text{constant}$
5. Phần dư (resid) giữa giá trị dự báo và giá trị thực tế tuân theo phân phối chuẩn.

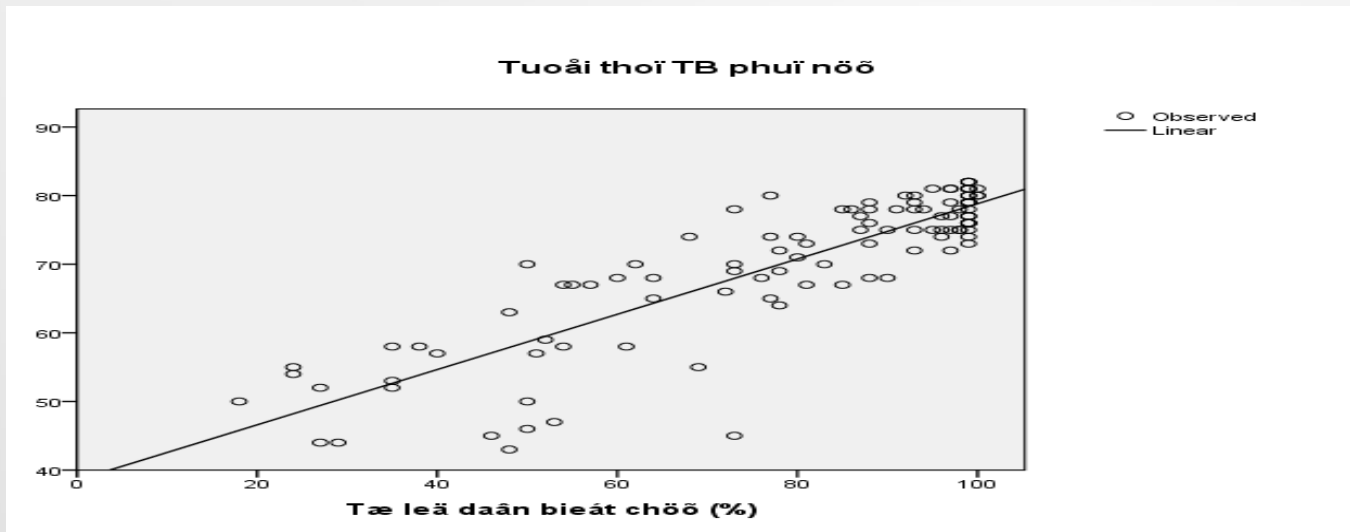
KIỂM ĐỊNH CÁC LỖI CỦA MÔ HÌNH

1. Giả định về sự liên hệ tuyến tính giữa hai biến (đồ thị scatter)
2. Khả năng tuân theo phân phối chuẩn của phần dư (residual)
3. Hiện tượng tự tương quan
4. Hiện tượng đa cộng tuyến
5. Hiện tượng phương sai thay đổi

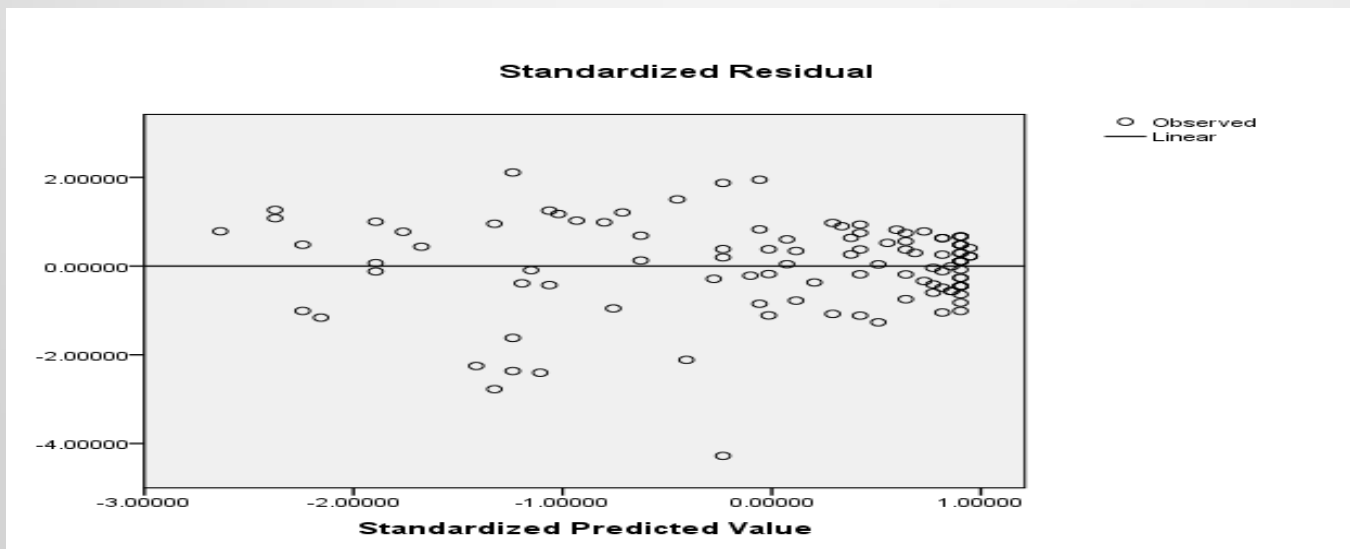
Giả định về sự liên hệ tuyến tính (đồ thị scatter)

- Biến độc lập X có thể giải thích cho biến phụ thuộc Y khi X có mối liên hệ tuyến tính với X .
 1. Đồ thị Scatter giữa X và Y có mối liên hệ nào đó với nhau
 2. Đồ thị phần dư (resid) giữa phần dư và giá trị dự báo biến thiên ngẫu nhiên

Giả định về sự liên hệ tuyến tính (đồ thị scatter)



Mối liên hệ tuyến tính



KIỂM TRA ĐỒ THỊ PHÂN TÁN PHẦN DƯ

- B1: Tính giá trị phần dư và giá trị dự báo chuẩn hoá.
- B2: Vẽ đồ thị phân tán với trục hoành là Zpr và trục tung là ZRE. (Có thể thêm vào đường xu hướng để kiểm tra mối liên hệ tuyến tính)

Linear Regression: Save

Predicted Values

- ☐ Unstandardized
- ☒ Standardized
- ☐ Adjusted
- ☐ S.E. of mean predictions

Residuals

- ☐ Unstandardized
- ☒ Standardized
- ☐ Studentized
- ☐ Deleted
- ☐ Studentized deleted

Distances

- ☐ Mahalanobis
- ☒ Cook's
- ☐ Leverage values

Influence Statistics

- ☐ DfBeta(s)
- ☐ Standardized DfBeta(s)
- ☐ DfFit
- ☐ Standardized DfFit
- ☐ Covariance ratio

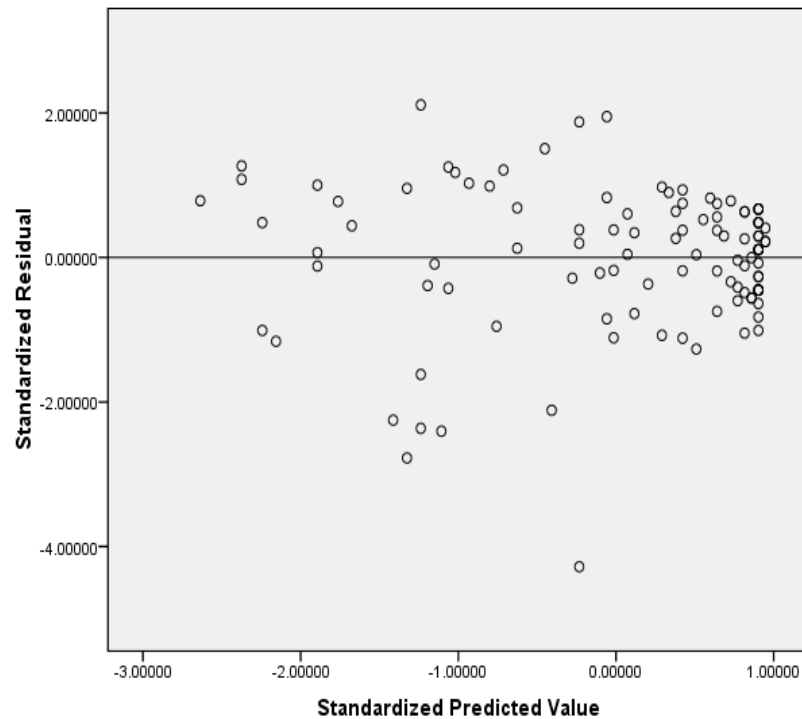
Prediction Intervals

- ☒ Mean
- ☒ Individual

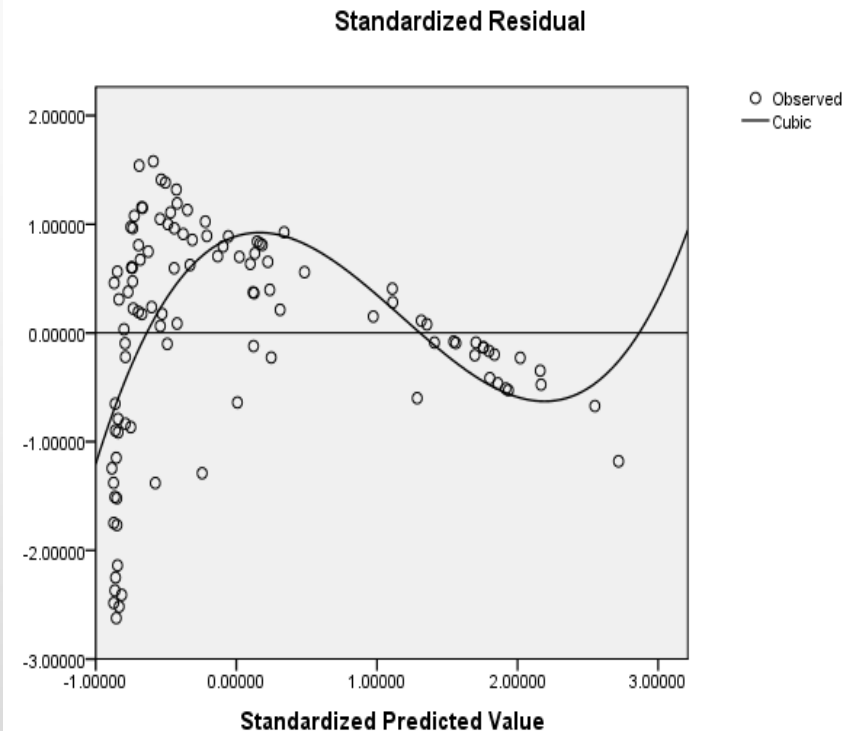
Confidence Interval: %

KIỂM TRA ĐỒ THỊ PHÂN TÁN PHẦN DƯ (QUY TRÌNH THỰC HIỆN)

Không có mối liên hệ



Có mối liên hệ (cubic)



Kiểm định phần dư có phân phối chuẩn

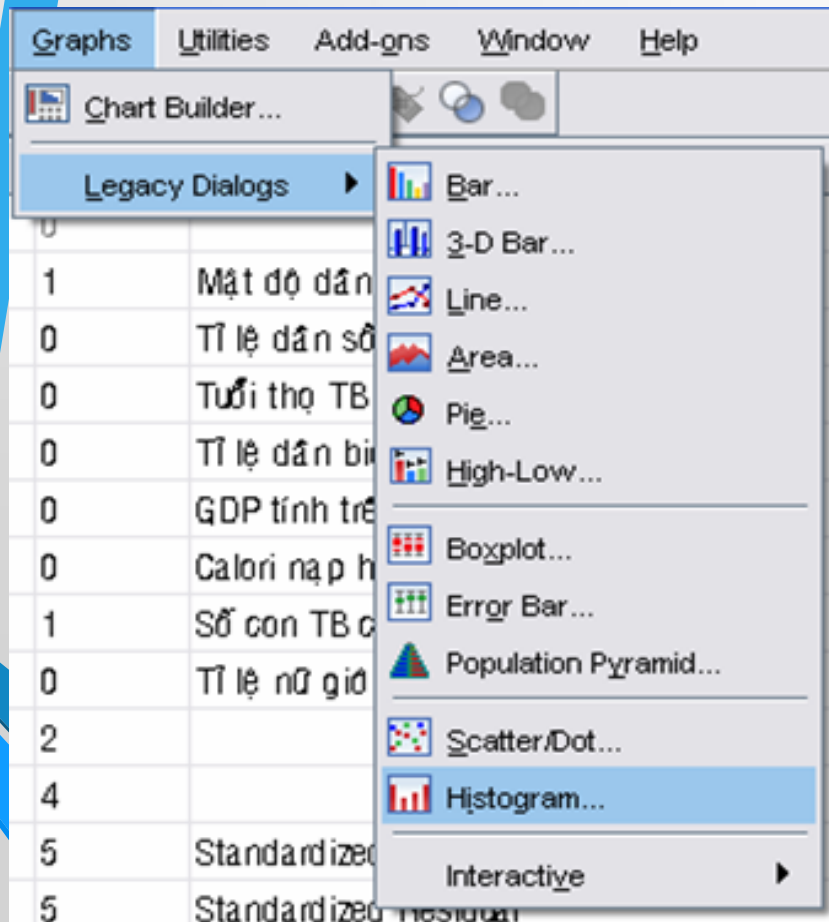
- Ta có thể kiểm tra khả năng tuân theo phân phối chuẩn của resid thông qua hai đồ thị

1. Đồ thị tần số Histogram

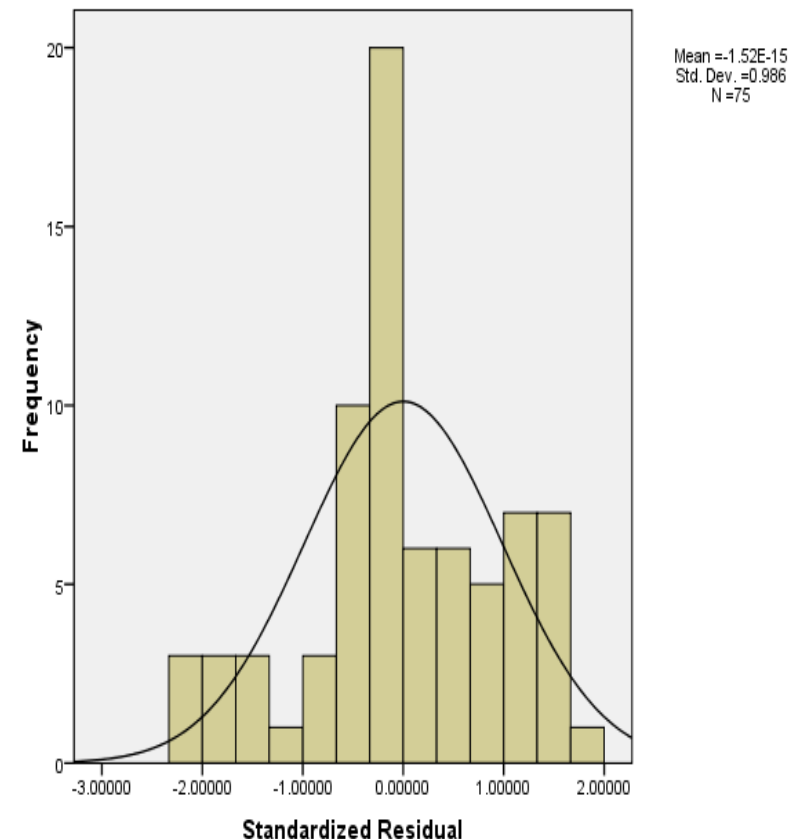
2. Đồ thị Q-Q plot

Kiểm định phần dư có phân phối chuẩn

Đồ thị histogram



Đồ thị histogram



Kiểm định phần dư có phân phối chuẩn

Đồ thị q-q plot

Đồ thị q-q plot

Linear Regression: Plots

DEPENDENT

- *ZPRED
- *ZRESID
- *DRESID
- *ADJPRED
- *SRESID
- *SDRESID

Scatter 1 of 1

Previous Next

Y: *ZRESID

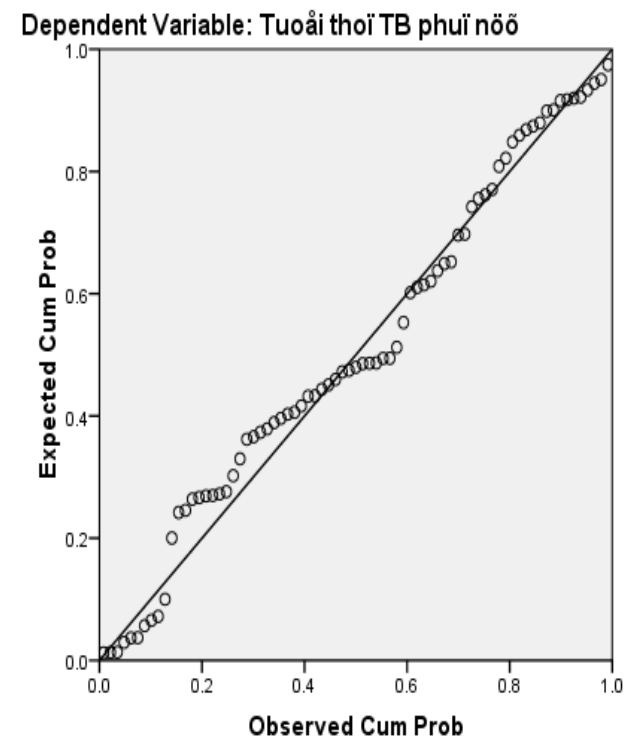
X: *ZPRED

Standardized Residual Plots

- ☒ Histogram
- ☒ Normal probability plot

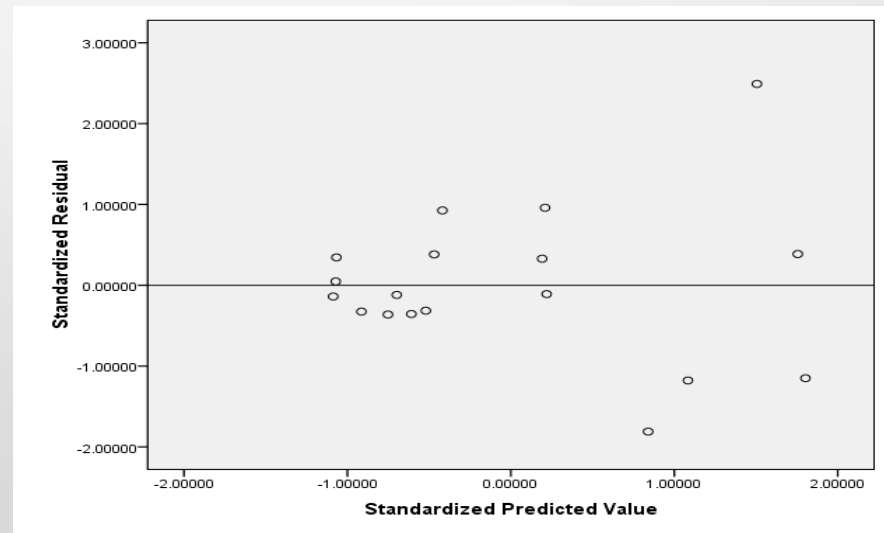
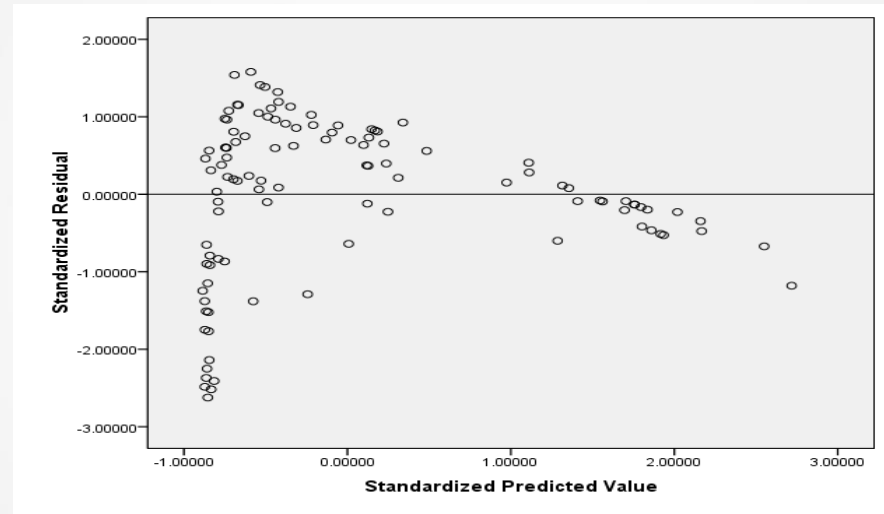
☐ Produce all partial plots

Continue Cancel Help



KIỂM ĐỊNH HIỆN TƯỢNG PHƯƠNG SAI THAY ĐỔI

- Là hiện tượng các sai số (resid) có mối tương quan với giá trị dự báo (\hat{Y}).



KIỂM ĐỊNH HIỆN TƯỢNG PHƯƠNG SAI THAY ĐỔI

Quy trình

- B1: Chạy hồi quy, lấy phần dư (resid)
- B2: Tạo biến trị tuyệt đối của resid (ABS_resid).
- B3. Kiểm định hệ số tương quan giữa biến ABS_resid với từng biến độc lập
 - H0: Không có hiện tượng phương sai thay đổi

H1: Ngược lại

Kiểm định hệ số tương quan

Correlations

Statistics=Pearson Correlation

	Tri tuyệt doi phan du	lggdpp
Tri tuyệt doi phan du	1	-.467**
lggdpp	-.467**	1

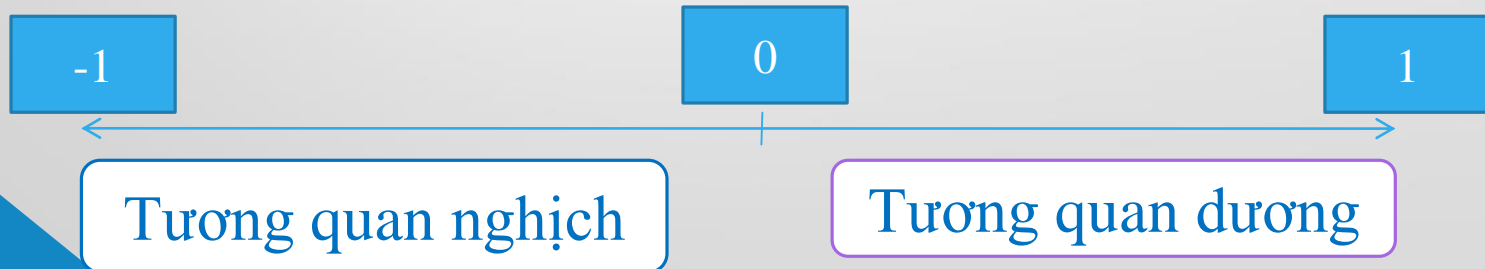
** . Correlation is significant at the 0.01 level (2-tailed).

KIỂM ĐỊNH HIỆN TƯỢNG TỰ TƯƠNG QUAN (Autocorrelation)

- Là hiện tượng các thành phần trong phần dư có mối tương quan với nhau [$\text{cov}(\varepsilon_i, \varepsilon_j) \neq 0$]

➤ Tương quan bậc 1: $\boldsymbol{\varepsilon}_i = \rho \boldsymbol{\varepsilon}_{i-1} + \boldsymbol{e}_i$

➤ Tương quan bậc p: $\varepsilon_i = \rho \varepsilon_{i-1} + \rho^2 \varepsilon_{i-2} + \dots + \rho^p \varepsilon_{i-p} + e_i$



KIỂM ĐỊNH HIỆN TƯỢNG TỰ TƯƠNG QUAN (Autocorrelation)

*Durbin – Watson d
tests first – order
autocorrelation of residuals*

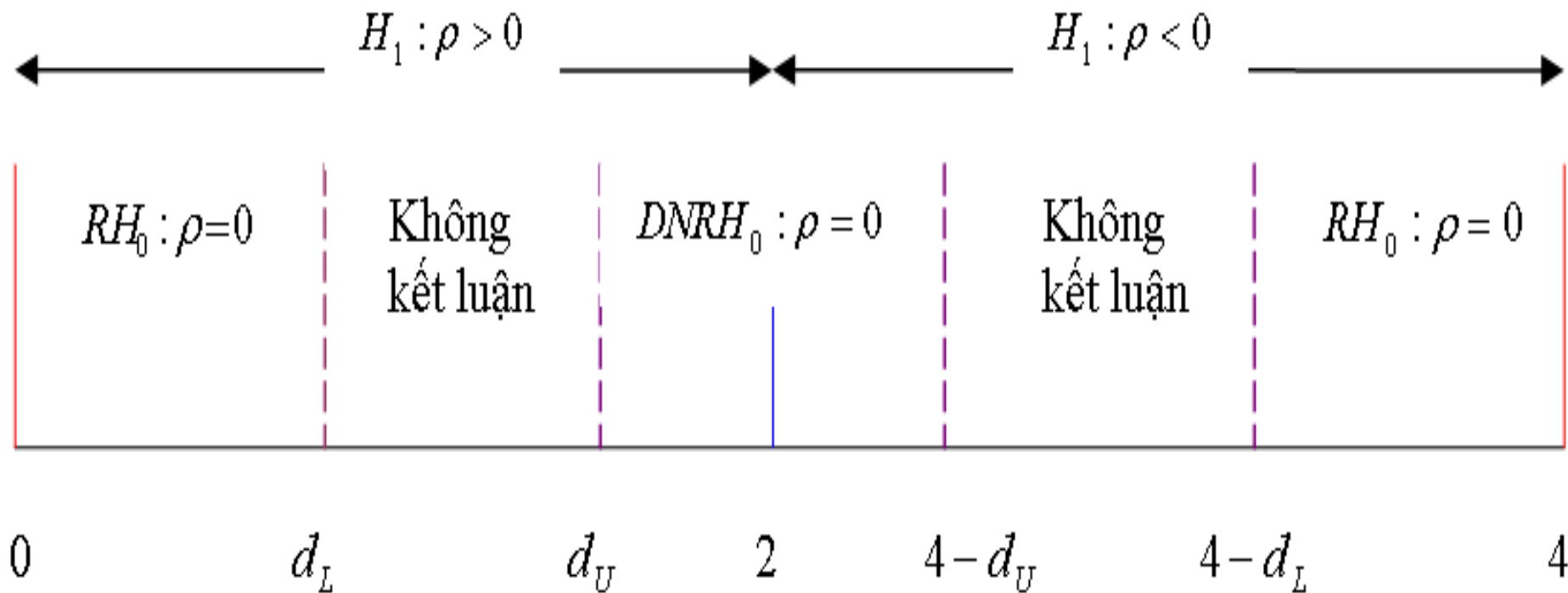
$$d = \sum_{i=1}^n \frac{(e_t - e_{t-1})^2}{e_t}$$

$$d = 2(1 - \rho)$$

KIỂM ĐỊNH HIỆN TƯỢNG TỰ TƯƠNG QUAN (Autocorrelation)

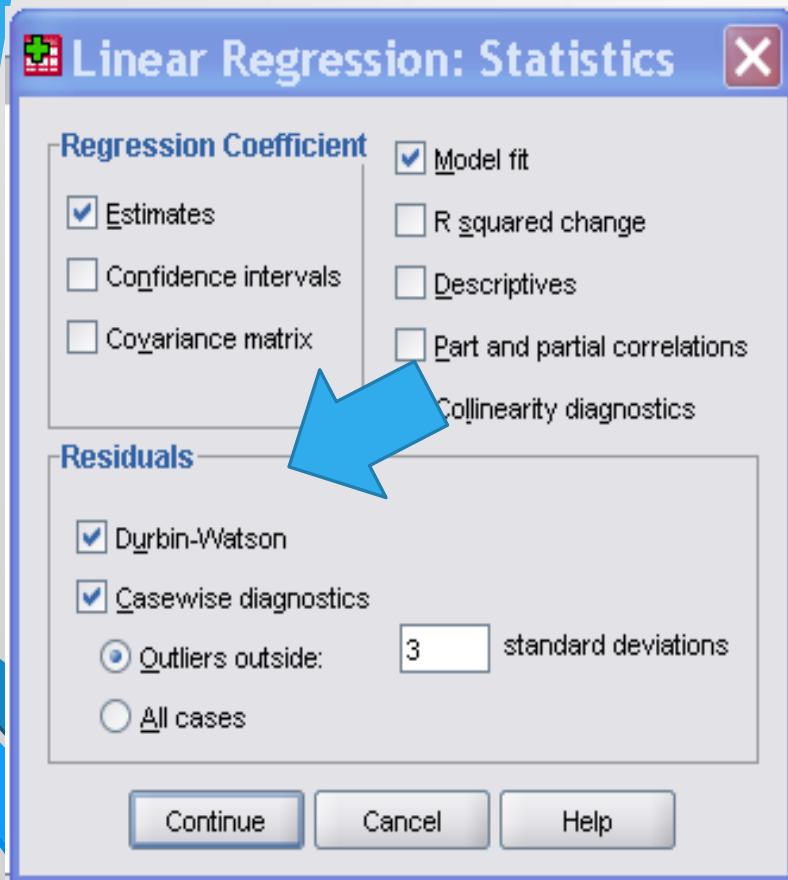
Tự tương quan thuận

Tự tương quan nghịch




KIỂM ĐỊNH HIỆN TƯỢNG TỰ TƯƠNG QUAN (Auto)

Cách phát hiện



Cách phát hiện

Model Summary^a



Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.840 ^a	.706	.698	6.275	1.783

a. Predictors: (Constant), logdpp, Calori nạp hàng ngày TB 1 người

b. Dependent Variable: Tuổi thọ TB phụ nữ

KIỂM ĐỊNH HIỆN TƯỢNG ĐA CỘNG TUYẾN (COLLINEAR)

- Có thể phát hiện hiện tượng đa cộng tuyến dựa vào các cách sau

1. Độ chấp nhận Tolerance = $1 - R_k^2$

2. Hệ số phóng đại phương sai (VIF)

$$VIF = \frac{1}{\textit{Tolerance}} = \frac{1}{1 - R_k^2}$$

3. Ma trận hệ số tương quan giữa các biến

KIỂM ĐỊNH HIỆN TƯỢNG ĐA CỘNG TUYẾN (COLLINEAR)

Linear Regression: Statistics

Regression Coefficient

☒ Estimates

☐ Confidence intervals

☐ Covariance matrix

☒ Model fit

☐ R squared change

☐ Descriptives

☐ Part and partial correlations

☒ Collinearity diagnostics

Residuals

☐ Durbin-Watson

☐ Casewise diagnostics

☒ Outliers outside: standard deviations

☐ All cases

Continue Cancel Help

Coefficients^a

		Model		
		1		
		(Constant)	Calori nạp hàng ngày TB 1 người	lggdpp
Unstandardized Coefficients	B	19.398	.005	10.469
	Std. Error	3.833	.002	2.069
Standardized Coefficients	Beta		.259	.609
t		5.061	2.153	5.059
Sig.		.000	.035	.000
Collinearity Statistics	Tolerance		.282	.282
	VIF		3.548	3.548

a. Dependent Variable: Tuổi thọ TB phụ nữ

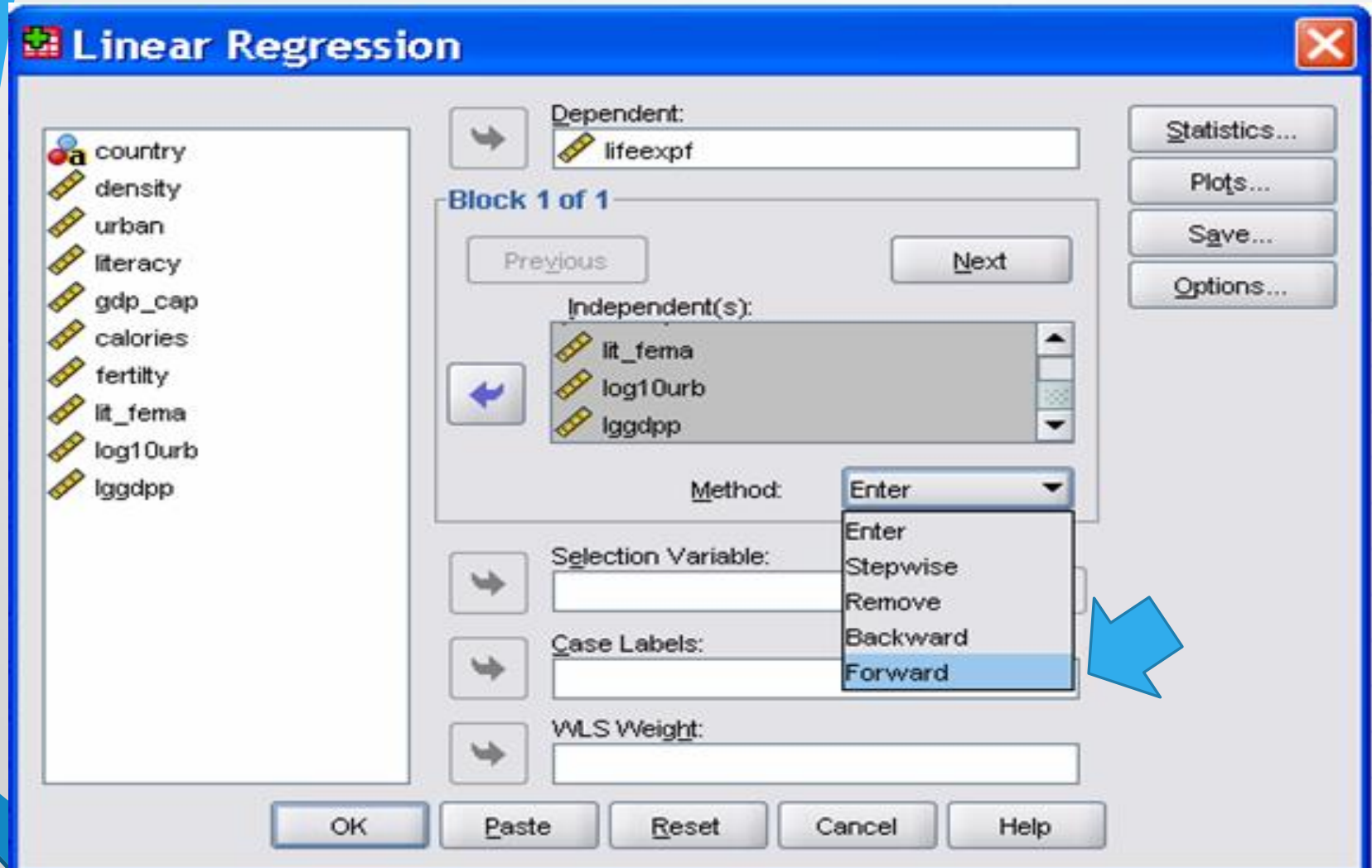
Thủ tục chọn biến nhanh

1. Thủ tục đưa vào dần (forward selection)
2. Thủ tục loại trừ dần (Backward elimination)
3. Thủ tục chọn từng bước (stepwise selection)

Thủ tục đưa vào dần

- **Nguyên tắc:** Dựa trên hệ số tương quan thuận (nghịch) lớn giữa biến phụ thuộc với từng biến độc lập. Biến nào lớn nhất được đưa vào trước.
- Điều kiện được đưa vào:
 1. Thỏa mãn điều kiện thống kê F (F in: thống kê đưa vào)
 2. Thỏa mãn điều kiện xác suất đưa vào (P in: xác suất để đưa vào)

Thủ tục đưa vào dần



Thủ tục đưa vào dần

Coefficients^a

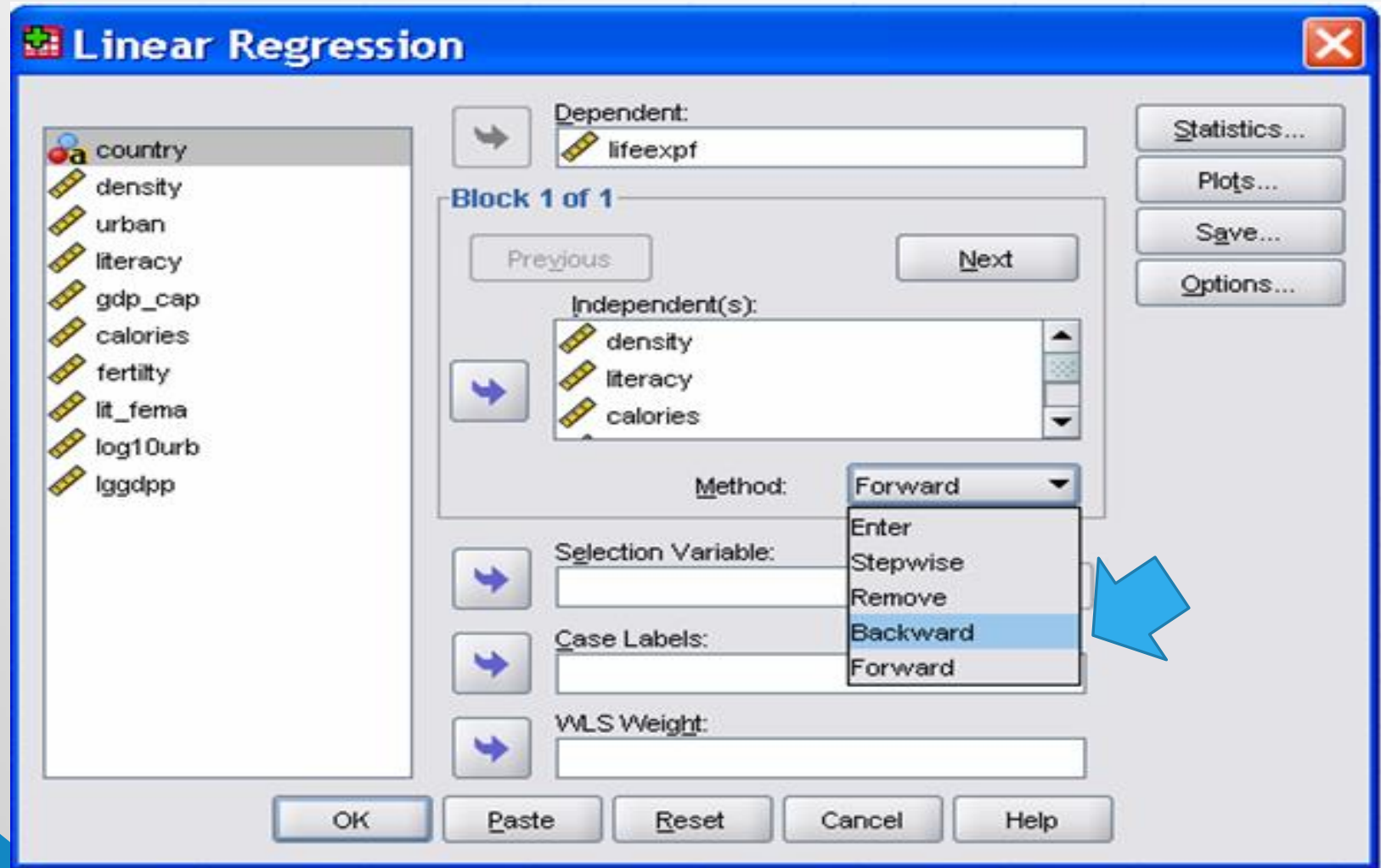
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	37.296	2.701		13.807	.000		
	Tỉ lệ dân biết chữ (%)	.410	.037	.827	11.085	.000	1.000	1.000
2	(Constant)	21.505	3.974		5.411	.000		
	Tỉ lệ dân biết chữ (%)	.283	.041	.570	6.927	.000	.588	1.701
	lggdpp	7.878	1.622	.400	4.857	.000	.588	1.701
3	(Constant)	37.656	7.137		5.276	.000		
	Tỉ lệ dân biết chữ (%)	.180	.055	.363	3.293	.002	.296	3.382
	lggdpp	7.291	1.556	.370	4.686	.000	.576	1.735
	Số con TB của 1 phụ nữ	-1.664	.624	-.278	-2.666	.010	.330	3.027
4	(Constant)	33.348	7.124		4.681	.000		
	Tỉ lệ dân biết chữ (%)	.165	.053	.332	3.104	.003	.291	3.436
	lggdpp	4.909	1.821	.249	2.696	.009	.390	2.563
	Số con TB của 1 phụ nữ	-1.557	.603	-.260	-2.583	.013	.328	3.045
	log10urb	7.635	3.318	.204	2.301	.025	.423	2.365

a. Dependent Variable: Tuổi thọ TB phụ nữ

Thủ tục loại trừ dần

- Nguyên tắc: Đưa tất cả các biến vào mô hình. Căn cứ vào biến nào có mối tương quan thấp nhất loại ra dần
- Điều kiện được loại trừ:
 1. Không thoả mãn điều kiện ở lại mô hình (F out: thống kê đưa vào)
 2. Không thoả mãn điều kiện ở lại mô hình (P out: xác suất để đưa vào)

Thủ tục loại trừ dần



Thủ tục loại trừ dần

Coefficients^a

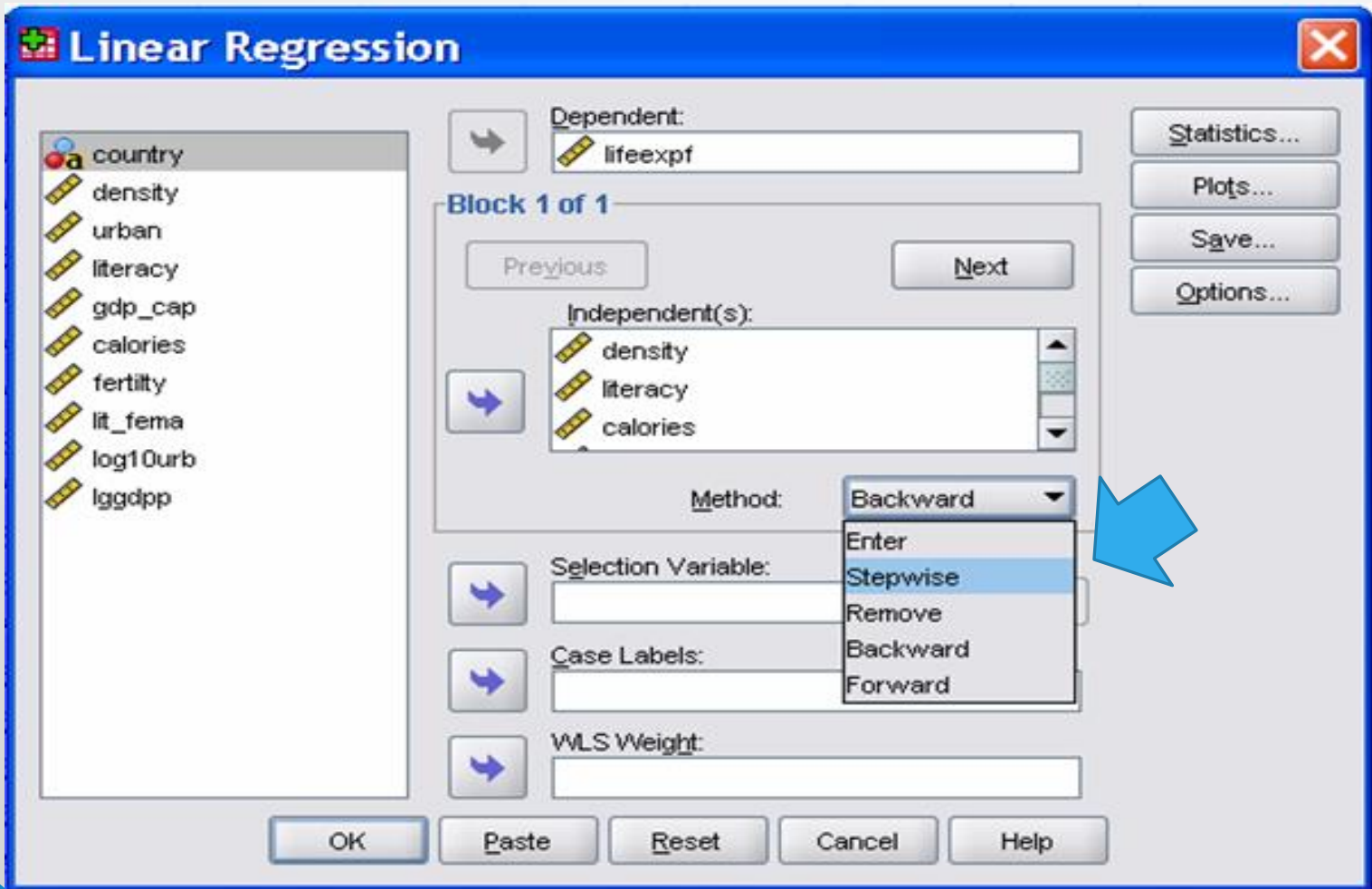
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	31.574	7.449		4.239	.000		
	Mật độ dân số (người/km2)	.000	.001	-.012	-.192	.849	.932	1.073
	Tỉ lệ dân biết chữ (%)	.117	.115	.235	1.017	.314	.064	15.733
	Calori nạp hàng ngày TB 1 người	.003	.002	.129	1.272	.209	.332	3.012
	Số con TB của 1 phụ nữ	-1.357	.647	-.227	-2.098	.041	.291	3.433
	Tỉ lệ nữ giới biết chữ (%)	.047	.095	.115	.497	.621	.064	15.634
	log10urb	6.843	3.472	.183	1.971	.054	.394	2.538
	lggdpp	3.472	2.345	.176	1.480	.145	.240	4.166
2	(Constant)	31.425	7.340		4.282	.000		
	Tỉ lệ dân biết chữ (%)	.118	.113	.239	1.045	.301	.064	15.638
	Calori nạp hàng ngày TB 1 người	.003	.002	.130	1.301	.199	.334	2.996
	Số con TB của 1 phụ nữ	-1.336	.631	-.223	-2.116	.039	.300	3.330
	Tỉ lệ nữ giới biết chữ (%)	.047	.094	.114	.500	.619	.064	15.632
	log10urb	6.890	3.431	.184	2.008	.050	.396	2.525
	lggdpp	3.396	2.290	.172	1.483	.144	.247	4.046
3	(Constant)	31.191	7.273		4.289	.000		
	Tỉ lệ dân biết chữ (%)	.168	.053	.340	3.189	.002	.290	3.446
	Calori nạp hàng ngày TB 1 người	.003	.002	.129	1.296	.201	.334	2.994
	Số con TB của 1 phụ nữ	-1.408	.610	-.235	-2.306	.025	.317	3.158
	log10urb	7.301	3.307	.195	2.208	.032	.420	2.380
	lggdpp	3.199	2.240	.162	1.428	.159	.255	3.927
4	(Constant)	33.348	7.124		4.681	.000		
	Tỉ lệ dân biết chữ (%)	.165	.053	.332	3.104	.003	.291	3.436
	Số con TB của 1 phụ nữ	-1.557	.603	-.260	-2.583	.013	.328	3.045
	log10urb	7.635	3.318	.204	2.301	.025	.423	2.365
	lggdpp	4.909	1.821	.249	2.696	.009	.390	2.563

a. Dependent Variable: Tuổi thọ TB phụ nữ

Thủ tục lựa chọn từng bước

- **Nguyên tắc:** Theo trình tự, đưa dần vào một biến theo nguyên tắc forward selection. Sau đó, xét biến này có thể tồn tại hay không theo nguyên tắc backward elimination.
- Chú ý: Để tránh trường hợp thực hiện liên tục (đưa vào rồi đưa ra), cần thiết lập:
 - $F_{in} > F_{out}$
 - hoặc $P_{in} < P_{out}$

Thủ tục lựa chọn từng bước



Thủ tục lựa chọn từng bước

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	37.296	2.701		13.807	.000		
	Tỉ lệ dân biết chữ (%)	.410	.037	.827	11.085	.000	1.000	1.000
2	(Constant)	21.505	3.974		5.411	.000		
	Tỉ lệ dân biết chữ (%)	.283	.041	.570	6.927	.000	.588	1.701
	lggdpp	7.878	1.622	.400	4.857	.000	.588	1.701
3	(Constant)	37.656	7.137		5.276	.000		
	Tỉ lệ dân biết chữ (%)	.180	.055	.363	3.293	.002	.296	3.382
	lggdpp	7.291	1.556	.370	4.686	.000	.576	1.735
	Số con TB của 1 phụ nữ	-1.664	.624	-.278	-2.666	.010	.330	3.027
4	(Constant)	33.348	7.124		4.681	.000		
	Tỉ lệ dân biết chữ (%)	.165	.053	.332	3.104	.003	.291	3.436
	lggdpp	4.909	1.821	.249	2.696	.009	.390	2.563
	Số con TB của 1 phụ nữ	-1.557	.603	-.260	-2.583	.013	.328	3.045
	log10urb	7.635	3.318	.204	2.301	.025	.423	2.365

a. Dependent Variable: Tuổi thọ TB phụ nữ

Sử dụng biến giả trong mô hình

- Sử dụng bài tập Employee data.sav xây dựng mô hình hồi quy với dự báo mức lương theo số năm kinh nghiệm và giới tính.

$$\text{Salary} = a + b1 * \text{gender} + b2 * \text{prevexp} (*)$$

B1: Mã hóa biến giới tính thành những giá trị số

B2: Xây dựng mô hình hồi quy (*) và giải thích ý nghĩa của các hệ số hồi quy

- Hãy cho biết vấn đề nào diễn ra đối với mô hình này???

Sử dụng biến giả trong mô hình

Coefficients^a

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	11845.000	2260.953		5.239	.000
	Gender	16406.073	1401.561	.479	11.706	.000
	Previous Experience (months)	-28.806	6.681	-.176	-4.312	.000

a. Dependent Variable: Current Salary

Sử dụng biến giả trong mô hình

- Quy tắc sử dụng biến giả:
 1. Đối với biến nominal, sử dụng trực tiếp biến giả để đưa vào mô hình hồi quy
 2. Đối với biến ordinal có thang đo dưới 5, sử dụng trực tiếp biến giả để đưa vào mô hình hồi quy. Trong trường hợp thang đo từ 5 trở lên, ta có thể xem biến ordinal như biến định lượng để đưa vào mô hình hồi quy.
- Sử dụng bài tập trình do hoc van.sav để dự báo mức lương (salary) theo trình độ học vấn (edu) và số năm kinh nghiệm (exp).



Xin cảm ơn!

