

THIẾT KẾ NGHIÊN CỨU TỔNG THỂ

6.1. Giới thiệu khái niệm

6.1.1. Khái niệm

Thiết kế nghiên cứu là tầm nhìn của nhà nghiên cứu về kết quả nghiên cứu với các chuẩn mực về độ chặt chẽ và tính phổ quát, cùng quy trình và nguồn lực tương ứng nhằm đạt các chuẩn mực đó. Thiết kế nghiên cứu phải gắn chặt với câu hỏi nghiên cứu và khung lý thuyết. Thiết kế nghiên cứu là trung tâm kết nối các hoạt động cơ bản của dự án nghiên cứu: xác định mẫu, thước đo, quy trình thu thập dữ liệu, phương pháp phân tích dữ liệu để trả lời cho câu hỏi nghiên cứu.

Thiết kế nghiên cứu có thể dưới dạng đơn lẻ hoặc kết hợp. Thiết kế nghiên cứu đơn lẻ là việc sử dụng một phương pháp nghiên cứu cụ thể (ví dụ: khảo sát hoặc thực nghiệm). Thiết kế kết hợp là việc sử dụng nhiều hơn một phương pháp.

6.1.2. Vai trò của thiết kế nghiên cứu

Vai trò đảm bảo chuẩn mực nghiên cứu: Thiết kế nghiên cứu giúp đảm bảo các dữ liệu thu thập cho phép trả lời câu hỏi một cách rõ ràng và chặt chẽ nhất có thể.

Nói khác đi, thiết kế nghiên cứu chính là việc trả lời câu hỏi: Cần dữ liệu gì để trả lời câu hỏi một cách thuyết phục nhất?

Vai trò kế hoạch: Thiết kế nghiên cứu có vai trò giúp nhà nghiên cứu chuẩn bị nguồn lực và lên kế hoạch cho các hoạt động một cách phù hợp nhất. Khi lựa chọn một thiết kế, các nhà nghiên cứu đã tính toán trước hạn chế của nghiên cứu và xác định liệu hạn chế này chấp nhận được không.

6.1.3. Phân biệt thiết kế nghiên cứu với đề cương báo cáo

Thiết kế nghiên cứu là việc hoạch định quy trình, phương pháp và nguồn lực nhằm trả lời câu hỏi nghiên cứu một cách chặt chẽ, thuyết phục. Trong khi đó đề cương báo cáo chỉ đơn thuần là việc dự kiến các mục trình bày trong báo cáo. Thiết kế quan tâm tới quá trình thực hiện nghiên cứu, còn đề cương báo cáo quan tâm tới trình bày kết quả nghiên cứu. Thiết kế là vấn đề tư duy vì nó phải thể hiện sự gắn kết giữa mục tiêu nghiên cứu với quy trình, phương pháp, nguồn lực. Báo cáo nghiên cứu thiên về vấn đề trình bày: Cùng một nghiên cứu nhưng báo cáo có thể khác nhau tùy theo đối tượng. Trước khi tiến hành nghiên cứu, các nhà nghiên cứu nên xây dựng và bảo vệ thiết kế

nghiên cứu. Đề cương báo cáo chỉ nên dừng ở mức độ tham khảo.

6.2. Các yêu cầu chính trong thiết kế nghiên cứu

6.2.1. Tính chặt chẽ

Một nghiên cứu không chỉ đơn giản là đi tìm dữ liệu và bằng chứng phù hợp với giả thuyết hay luận điểm định trước. Tính chặt chẽ đòi hỏi nghiên cứu phải tìm đủ bằng chứng/dữ liệu để bác bỏ hoặc kiểm soát các giả thuyết “cạnh tranh” khác.

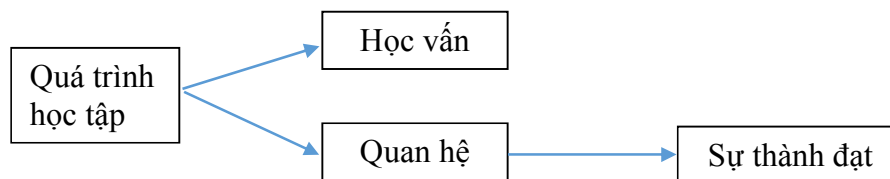
Ví dụ dưới đây minh họa điều này:

Một người nghiên cứu muốn kiểm định vai trò của vốn con người tới sự thành đạt của các cá nhân. Tác giả thu thập dữ liệu và kết quả minh chứng rằng học vấn càng cao (vốn con người cao) thì sự thăng tiến trong công việc càng cao (sự thành đạt). Nghiên cứu này được mô phỏng như sau:



Sơ đồ 6.1. Mô phỏng học vấn càng cao thì càng thành đạt

Nếu nghiên cứu dừng ở đây sẽ chưa đảm bảo sự chặt chẽ vì chưa tính tới các giả thuyết “cạnh tranh”. Một trong những giả thuyết cạnh tranh có thể là *Học vấn* không hề ảnh hưởng tới *Sự thành đạt*. Chẳng qua người có học vấn cao là người có quan hệ nhiều hơn với những người thành đạt hơn. Như vậy không phải vốn con người (học vấn) mà là *vốn xã hội (quan hệ) mang lại sự thành đạt*. Giả thuyết này có thể được mô phỏng như sau:



Sơ đồ 6.2. Mô phỏng quá trình học tập dẫn tới sự thành đạt.

6.2.2. Tính khái quát

Một trong những chuẩn mực của nghiên cứu khoa học là tính phổ biến của kết quả nghiên cứu. Tính khái quát hóa của nghiên cứu đòi hỏi kết quả nghiên cứu phải có khả năng suy rộng. Có 3 loại tổng quát hóa cơ bản sau:

- Khái quát cho tổng thể đối tượng nghiên cứu: Kết quả từ một mẫu nghiên cứu liệu có thể suy rộng cho tổng thể nghiên cứu hay không?

Ví dụ 1. Kết quả nghiên cứu từ một mẫu gồm 200 sinh viên đại học liệu có thể suy rộng cho tổng thể là sinh viên đại học được hay không? hoặc rộng hơn nữa, liệu có

thể suy rộng cho trí thức trẻ được không? Điều này phụ thuộc rất nhiều tính đại diện của mẫu nghiên cứu, trong đó quy trình chọn mẫu và quy mô mẫu có ý nghĩa quyết định.

- Khái quát cho các bối cảnh nghiên cứu khác nhau: Kết quả nghiên cứu có thể suy rộng cho các bối cảnh khác nhau được hay không?

Ví dụ 2. Một nghiên cứu cảm nhận và trải nghiệm của người dân về tham nhũng ở 10 tỉnh, thành phố có thể suy rộng cho các tỉnh thành phố trong cả nước hay không? Điều này phụ thuộc vào việc 10 tỉnh, thành phố được nghiên cứu có thể đại diện cho 63 tỉnh thành phố về cảnh kinh tế, xã hội, văn hóa hay không?

- Khái quát cho các thời điểm khác nhau: Liệu kết quả nghiên cứu có trường tồn theo thời gian không? Điều này phụ thuộc rất nhiều liệu thời gian cho làm thay đổi khung cảnh và làm thay đổi kết quả nghiên cứu hay không. Những nghiên cứu mang tính mô tả (ví dụ: mô tả thực trạng nền kinh tế hay giá trị văn hóa) không có tính trường tồn cao. Những nghiên cứu hướng vào mối quan hệ có tính quy luật có tính trường tồn cao hơn. Tuy nhiên, khi bối cảnh thay đổi lớn thì kết quả cũng có thể thay đổi.

6.2.3. Tính khả thi

Không có nghiên cứu nào có nguồn lực vô hạn. Ngoài ra, kết quả nghiên cứu còn phụ thuộc vào dữ liệu sẵn có. Vì vậy, nếu thiết kế nghiên cứu vượt ra ngoài khả năng về nguồn lực và tiếp cận dữ liệu thì cũng không có ý nghĩa thực thi. Vì vậy, các nhà nghiên cứu cần cân đối giữa hai yêu cầu trên (tính chặt chẽ và tính khái quát hóa) với nguồn lực và khả năng tiếp cận dữ liệu trong thiết kế của mình.

6.3. Giới thiệu một số thiết kế nghiên cứu

6.3.1. Các bước thiết kế nghiên cứu

Bước 1. Chọn đề tài nghiên cứu

Bước 2. Xác định câu hỏi nghiên cứu

Bước 3. Mô tả thiết kế nghiên cứu để thực hiện

- Cách thu thập số liệu.
- Những thông tin cần thu thập để trả lời câu hỏi nghiên cứu.
- Các phương pháp đề tài sử dụng để thu thập dữ liệu.
- Ưu và nhược điểm của thiết kế nghiên cứu này.

6.3.2. Ví dụ về một thiết kế nghiên cứu cụ thể

Xét một thiết kế cụ thể được cho dưới bảng sau

Mục tiêu nghiên cứu	Giả thuyết nghiên cứu	Dữ liệu cần thiết	Phương pháp thu thập	Hạn chế
Mục tiêu 1: Nghiên cứu mối quan hệ học vấn và thu nhập	Giả thuyết nghiên cứu: Học vấn càng cao càng có thu nhập cao Giả thuyết cạnh tranh 1: Càng nhiều mối quan hệ càng có thu nhập cao Giả thuyết cạnh tranh 2: Sự trợ giúp của bố mẹ càng nhiều càng có thu nhập cao.	Dữ liệu ở cấp độ từng cá nhân: Thu nhập Học vấn Mối quan hệ xã hội (số lượng, nhóm,...) Sự trợ giúp của bố mẹ (tài chính, định hướng nghề nghiệp) Tuổi, giới tính, dân tộc,...	Khảo sát ngẫu nhiên một mẫu người trưởng thành.	Dữ liệu khảo sát không cho biết mối quan hệ nhân quả: Vì học vấn cao, có nhiều mối quan hệ nên thu nhập cao hay ngược lại.
Mục tiêu 2: Nghiên cứu mối quan hệ giữa học vấn và sự thăng tiến trong công việc				

Bảng 6.1. Ví dụ về một thiết kế.

6.4. Câu hỏi thảo luận

Câu hỏi 1. Thiết kế nghiên cứu tổng thể là gì? Nêu những yêu cầu chính khi thiết kế nghiên cứu tổng thể.

Câu hỏi 2. Giả thuyết cạnh tranh là gì? Tại sao khi thiết kế nghiên cứu chúng ta phải quan tâm tới giả thuyết cạnh tranh? Làm thế nào để đảm bảo giả thuyết cạnh tranh trọng được kiểm soát trong thiết kế nghiên cứu?

Câu hỏi 3. Thiết kế nghiên cứu tổng thể kết hợp nhiều phương pháp có ưu/nhược điểm gì? Hãy nêu một ví dụ cụ thể về một thiết kế nghiên cứu kết hợp và bình luận điểm mạnh, điểm yếu của thiết kế này.

Câu hỏi 4. Giả thuyết cạnh tranh là gì? Tại sao khi thiết kế nghiên cứu chúng ta phải quan tâm tới giả thuyết cạnh tranh?

Câu hỏi 5. Một người cho rằng “ Hướng đổi mới phương pháp dạy học hiện nay là các nghiên cứu phải áp dụng nghiên cứu định tính”. Anh/chị có đồng tình với nhận định đó hay không? Tại sao?

Thuật ngữ chính chương 6	
Tiếng Anh	Tiếng Việt
Assumptions	Giả thiết
Data collection process	Quy trình thu thập dữ liệu
Data analysis	Phân tích dữ liệu
Data over time	Dữ liệu theo thời gian
Data collection	Thu thập dữ liệu
External validity	Tính khái quát hóa
Factor scale	Thang đo nhân tố
For example	Ví dụ
General	Khái quát
Internal validity	Tính chặt chẽ
Main design	Thiết kế chính
Models	Mô hình
New points of research	Điểm mới của nghiên cứu
Research outline	Đề cương nghiên cứu
Report	Báo cáo
Research report	Báo cáo nghiên cứu
Research restrictions	Hạn chế nghiên cứu
Research summary	Tóm tắt nghiên cứu
Research models	Mô hình nghiên cứu
Scale	Thang đo
Scheme	Kế hoạch

XỬ LÝ VÀ PHÂN TÍCH DỮ LIỆU ĐỊNH LƯỢNG CHO NGHIÊN CỨU

7.1. Tổng quan kiến thức về thống kê và sử dụng các kỹ thuật thống kê

7.1.1. Phân tích mô tả và khám phá

7.1.1.1. Thống kê mô tả

Thống kê mô tả cung cấp các chỉ số cơ bản của biến số với dữ liệu của mẫu nghiên cứu. Hầu hết các nghiên cứu định lượng đều cần cung cấp các chỉ số thống kê mô tả để giúp người đọc hiểu về dữ liệu sử dụng. Các chỉ số và cách trình bày có thể khác nhau với biến định lượng và biến định danh.

Đối với các biến có giá trị liên tục (biến định lượng). Các nhà nghiên cứu thường cung cấp các chỉ số như giá trị trung bình, giá trị lớn nhất, giá trị nhỏ nhất và độ lệch chuẩn của biến. Trong một số trường hợp thì cần thêm giá trị trung vị.

Đối với các biến định danh. Các nhà nghiên cứu thường cung cấp các chỉ số như tần suất, tỷ lệ phần trăm trong tổng số, giá trị trung vị, giá trị yếu vị.

Ví dụ 1. Cho bảng thống kê mô tả của các biến định lượng

	Y	X2	X3	
	Y	X2	X3	
Mean	141.3333	20.41667	12.16667	
Median	141.5000	20.50000	12.00000	
Maximum	180.0000	26.00000	17.00000	
Minimum	102.0000	15.00000	6.000000	
Std. Dev.	23.20789	4.187825	3.352972	
Skewness	-0.218444	-0.034924	-0.453123	
Kurtosis	2.301560	1.401529	2.378514	
Jarque-Bera	0.339345	1.279994	0.603763	
Probability	0.843941	0.527294	0.739426	
Sum	1696.000	245.0000	146.0000	
Sum Sq. Dev.	5924.667	192.9167	123.6667	
Observations	12	12	12	

Bảng 7.1. Thống kê mô tả.

Trong đó

- Mean: Trung bình của lượng biến.

- Median: Trung vị của lượng biến.
- Maximun: Giá trị lớn nhất của lượng biến.
- Minimun: Giá trị nhỏ nhất của lượng biến.
- Std. Dev: Độ lệch chuẩn của lượng biến.
- Skewness: Hệ số bất đối xứng.
- Kurtosis: Hệ số nhọn.
- Jarque – Bera: Giá trị thống kê Jarque – Bera dùng để kiểm định phân phối chuẩn.
- Probability: Giá trị xác suất của thống kê Jarque – Bera dùng để kiểm định phân phối chuẩn.
- Sum: Tổng các giá trị của lượng biến.
- Sum Sq. Dev: Độ lệch chuẩn của tổng.
- Observations: Tổng số quan sát.

7.1.1.2. Ma trận hệ số tương quan

Các biến số có thể có tương quan với nhau, ma trận hệ số tương quan là một công cụ ban đầu để giúp các tác giả và người đọc quan sát về mối tương quan của từng cặp biến. Công đoạn này cũng giúp các tác giả nhận biết các hiện tượng bất thường hoặc đề phòng trường hợp đa cộng tuyến khi các biến độc lập có tương quan lớn.

Ví dụ 2. Cho ma trận tương quan các biến định lượng như sau:

Correlation				
	Y	X2	X3	
	Y	X2	X3	
Y	1.000000	0.782281	0.904627	
X2	0.782281	1.000000	0.480173	
X3	0.904627	0.480173	1.000000	

Bảng 7.2. Ma trận tương quan giữa các biến.

Ý nghĩa. Ma trận tương quan cho biết mối tương quan giữa các biến trong mô hình. Ví dụ hệ số tương quan của X_2 và X_3 là 0,480173; hệ số tương quan của Y và X_2 là 0,782281; hệ số tương quan của Y và X_3 là 0,904627.

7.1.1.3. Phân tích nhân tố khám phá

Phân tích nhân tố khám phá là một phương pháp phân tích thống kê dùng để rút gọn một tập gồm nhiều biến quan sát có liên hệ với nhau thành một tập biến (gọi là các nhân tố) ít hơn để chúng có ý nghĩa hơn nhưng vẫn chứa đựng hầu hết nội dung thông

tin của tập biến ban đầu (Hair và cộng sự, 1998).

Phân tích nhân tố khám phá thường được sử dụng nhiều nhất trong nghiên cứu khảo sát khi mà tác giả phải dùng nhiều câu hỏi để thu thập thông tin một vấn đề trừu tượng hơn, đặc biệt là những thông tin về tâm lý, thái độ, thậm chí hành vi. Kể cả khi chúng ta sử dụng thước đo đã được các tác giả trước phát triển và kiểm định thì cũng vẫn nên thực hiện kỹ thuật này xem liệu các mệnh đề/câu hỏi có “nhóm” đúng theo thước đo ban đầu hay không. Kết quả phân tích nhân tố là cơ sở để tạo biến số cho các phân tích tiếp theo.

7.1.1.4. Phân tích độ tin cậy của thước đo

Phân tích độ tin cậy cho phép chúng ta xác định thuộc tính của thước đo mà chủ yếu là liệu các mệnh đề/câu hỏi của thước đo có “thống nhất” với nhau hay không. Thông thường các tác giả sử dụng chỉ số Cronbach’ alpha từ 0,7 trở lên, song giá trị tối thiểu để thước đo có thể sử dụng được là 0,63 (DeVellis, 1990).

Phân tích độ tin cậy của thước đo thường được sử dụng cùng với phân tích nhân tố khám phá để quyết định các mệnh đề/câu hỏi cho từng thước đo. Lý tưởng nhất thước đo đủ cả ba điều kiện:

- Các câu hỏi/mệnh đề của thước đo được phát triển dựa trên lý thuyết hoặc đã được các tác giả trước xây dựng và kiểm định.
- Các câu hỏi/mệnh đề của thước đo “nhóm” cùng với nhau khi thực hiện phân tích nhân tố khám phá.
- Các câu hỏi/mệnh đề có chỉ số Cronbach’s alpha từ 0,7 trở lên, hoặc ít nhất cũng là 0,63.

7.1.2. So sánh nhóm

Một dạng nghiên cứu định lượng khá thông dụng là so sánh sự khác biệt giữa các nhóm về một hoặc một số chỉ số nào đó. Dưới đây là liệt kê các công cụ chính:

- T – test (kiểm định t) : được sử dụng để so sánh hai giá trị trung bình.
- ANOVA và ANCOVA (Analysis of Covariance) : Khi có nhiều hơn hai nhóm cần so sánh thì sử dụng ANOVA. Khi so sánh các nhóm, đồng thời kiểm soát tác động của một biến liên tục khác thì sử dụng ANCOVA.
- MANOVA (Multivariate Analysis of Variance): Tương tự như ANOVA nhưng được sử dụng khi có nhiều hơn một biến phụ thuộc và các biến phụ thuộc lại tương quan chặt với nhau.

7.2. Hồi quy tuyến tính cho phân tích dữ liệu định lượng

7.2.1. Mô hình hồi quy đơn

7.2.1.1. Hàm hồi quy tổng thể

Giả sử ta có các bộ số liệu $(X_i, Y_{i,j})$ cho tổng thể, với $i = 1, 2, \dots, n$; $j = 1, 2, \dots, m(i)$. Ứng với mỗi giá trị của X , $X = X_i$, với $i = 1, 2, \dots, n$, ta có thể có nhiều giá trị của Y tương ứng nên quan hệ của Y theo X không là quan hệ “hàm số”. Tuy nhiên, ứng với mỗi giá trị của X , $X = X_i$, ta có duy nhất giá trị trung bình $E(Y | X = X_i)$, nên quan hệ này trở thành quan hệ hàm số

$$E(Y | X = X_i) \equiv f(X_i)$$

và hàm số này được gọi là *hàm hồi quy tổng thể*, PRF (Population Regression Functions) mà trong trường hợp này, ta còn gọi là *hàm hồi quy đơn (hồi quy hai biến)*, do nó chỉ có một biến độc lập. Trường hợp có nhiều hơn một biến độc lập, ta gọi là *hàm hồi quy bội*.

Trước hết, giả sử PRF là hàm tuyến tính

$$E(Y | X = X_i) = \beta_1 + \beta_2 X_i$$

mà ta còn viết là

$$E(Y | X) = \beta_1 + \beta_2 X,$$

trong đó β_1 và β_2 là các tham số chưa biết nhưng cố định, được gọi là các *hệ số hồi quy*; β_1 gọi là *hệ số tự do* hay *hệ số chặn*, β_2 gọi là *hệ số góc* (nó cho biết tỷ lệ thay đổi của Y đối với X).

Dạng ngẫu nhiên

$$Y = \beta_1 + \beta_2 X + \varepsilon,$$

trong đó ε là một đại lượng ngẫu nhiên.

7.2.1.2. Hàm hồi quy mẫu

Hàm hồi quy quy tuyến tính mẫu có dạng

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X,$$

trong đó $\hat{Y}$ là ước lượng điểm của $E(Y | X)$, $\hat{\beta}_1$ là ước lượng điểm của β_1 và $\hat{\beta}_2$ là ước lượng điểm β_2 .

7.2.1.3. Phương pháp bình phương bé nhất

Phương pháp bình phương nhỏ nhất, OLS (Ordinary Least Square), do nhà toán học Đức Carl Fredrich Gauss đưa ra. Với phương pháp này, kèm theo một vài giả thiết, các ước lượng thu được có một số tính chất đặc biệt mà nhờ đó nó trở thành phương pháp hồi quy mạnh và phổ biến nhất.

Nội dung phương pháp OLS

Giả sử $Y_i = \beta_1 + \beta_2 X_i$ là PRF cần tìm. Ta tìm cách ước lượng nó bằng cách xây dựng SRF dạng

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

từ một mẫu gồm n quan sát (X_i, Y_i) , với $i = 1, 2, \dots, n$.

Khi đó, ứng với mỗi i , sai biệt giữa giá trị chính xác, Y_i , và giá trị ước lượng, $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$, là $e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$, mà ta gọi là các *phần dư*.

Phương pháp OLS nhằm xác định các tham số $(\hat{\beta}_1, \hat{\beta}_2)$ sao cho tổng bình phương các phần dư, $RSS = e_1^2 + e_2^2 + \dots + e_n^2 = \sum_{i=1}^n e_i^2$, là đạt nhỏ nhất.

Chú ý rằng tổng bình phương các phần dư này là hàm theo hai biến $(\hat{\beta}_1, \hat{\beta}_2)$,

$$RSS = \sum_{i=1}^n (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

với đạo hàm riêng theo các biến

$$\frac{\partial RSS}{\partial \hat{\beta}_1}(\hat{\beta}_1, \hat{\beta}_2) = -\sum_{i=1}^n 2(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)$$

$$\frac{\partial RSS}{\partial \hat{\beta}_2}(\hat{\beta}_1, \hat{\beta}_2) = -\sum_{i=1}^n 2X_i(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)$$

Giá trị nhỏ nhất của RSS, nếu có, phải đạt tại điểm dừng của nó, nghĩa là

$$\begin{cases} \frac{\partial RSS}{\partial \hat{\beta}_1}(\hat{\beta}_1, \hat{\beta}_2) = 0 \\ \frac{\partial RSS}{\partial \hat{\beta}_2}(\hat{\beta}_1, \hat{\beta}_2) = 0 \end{cases}$$

Do đó, ta nhận được hệ phương trình

$$\begin{cases} n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^n X_i = \sum_{i=1}^n Y_i \\ \hat{\beta}_1 \sum_{i=1}^n X_i + \hat{\beta}_2 \sum_{i=1}^n X_i^2 = \sum_{i=1}^n X_i Y_i \end{cases}$$

Giải hệ phương trình trên, ta được

$$\hat{\beta}_2 = \frac{n \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

và

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X},$$

trong đó $\bar{X}$, $\bar{Y}$ là các trung bình của mẫu X , Y .

Các giá trị $\hat{\beta}_1$ và $\hat{\beta}_2$ nhận được bằng các công thức trên được gọi là các *ước lượng bình phương nhỏ nhất* của β_1 và β_2 .

Ví dụ 3. Bảng sau cho số liệu về lãi suất ngân hàng (Y) và tỷ lệ lạm phát (X) trong năm 1988 ở 9 nước

X	7,2	4,0	3,1	1,6	4,8	51,0	2,0	6,6	4,4
Y	11,9	9,4	7,5	4,0	11,3	66,3	2,2	10,3	7,6

Giả sử ra rằng X và Y có quan hệ tuyến tính. Hãy ước lượng hàm hồi quy mẫu.

Giải. Lập bảng ta tính được một số đại lượng sau :

X_i	Y_i	$X_i Y_i$	X_i^2
7,2	11,9	85,68	51,84
4	9,4	37,6	16
3,1	7,5	23,25	9,61
1,6	4	6,4	2,56
4,8	11,3	54,24	23,04
51	66,3	3381,3	2601
2	2,2	4,4	4
6,6	10,3	67,98	43,56
4,4	7,6	33,44	19,36
$\sum X_i = 84,7$	$\sum Y_i = 130,5$	$\sum X_i Y_i = 3694,29$	$\sum X_i^2 = 2770,97$

Bảng 7.3. Kết quả tính toán các tổng.

Thay vào hệ phương trình trên, ta có

$$\begin{cases} 9\hat{\beta}_1 + 84,7\hat{\beta}_2 = 130,5 \\ 84,7\hat{\beta}_1 + 2770,97\hat{\beta}_2 = 3694,29 \end{cases} \Leftrightarrow \begin{cases} \hat{\beta}_1 = 2,74169 \\ \hat{\beta}_2 = 1,24941 \end{cases}$$

và từ đó suy ra hệ số hồi quy

Ta nhận được SRF: $\hat{Y} = 2,74169 + 1,24941X$.

7.2.1.4. Các giả thiết của mô hình

Để có thể dùng các công cụ của thống kê toán nhằm đánh giá chất lượng của mô hình hồi quy tuyến tính, ta cần các giả thiết sau trên biến số X và đại lượng sai số ngẫu nhiên ε

Giả thiết 1. Biến giải thích X là biến phi ngẫu nhiên, nghĩa là các giá trị của nó được hoàn toàn xác định. Giả thiết này đương nhiên được thỏa trong mô hình hồi quy tuyến tính.

Theo giả thiết này thì các giá trị của biến độc lập X là những giá trị có sẵn. Ví dụ khi tìm mối quan hệ giữa thu nhập và chi tiêu thì các mức thu nhập được xác định từ trước để xem chi tiêu thay đổi như thế nào tương ứng với mỗi mức thu nhập sẵn có này.

Giả thiết 2. Kỳ vọng của sai số ngẫu nhiên ε bằng 0, nghĩa là $E(\varepsilon_i) = E(\varepsilon_i | X = X_i) = 0$.

Giả thiết này có nghĩa là các yếu tố không xuất hiện trong mô hình, được đại diện bởi đại lượng ngẫu nhiên ε , không ảnh hưởng một cách có hệ thống đến giá trị trung bình của Y .

Với mỗi một giá trị cho trước của biến độc lập X_i có một tập hợp giá trị của biến phụ thuộc Y xoay quanh giá trị trung bình, trong đó có một số giá trị nằm trên đường trung bình và các giá trị khác nằm dưới đường trung bình tạo ra các sai số ε_i . Giả thiết 2 giả định rằng giá trị kỳ vọng của các sai số này bằng 0.

Nói cách khác, tác động dương của các sai số triệt tiêu tác động âm của những sai số khác làm cho tổng ảnh hưởng của sai số đến biến phụ thuộc là bằng 0.

Giả thiết 3. Phương sai của sai số không đổi (phương sai thuần nhất), nghĩa là $\text{Var}(\varepsilon_i) = \text{Var}(\varepsilon_j) = \sigma^2$, với mọi i, j . Giả thiết này có nghĩa là các giá trị cụ thể của Y tương ứng với cùng một giá trị của X đều được phân bố xung quanh giá trị trung bình của nó với cùng một mức độ phân tán như nhau.

Giả thiết 4. Không có tương quan giữa các sai số, nghĩa là $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$, với mọi i, j .

Giả thiết này có nghĩa là sai số ở số liệu quan sát này không ảnh hưởng gì tới sai số ở số liệu quan sát khác.

Giả thiết 5. Sai số và biến giải thích không có tương quan, nghĩa là $\text{Cov}(\varepsilon, X) = 0$. Giả thiết này là cần thiết vì nó cho phép tách bạch ảnh hưởng của X và của các yếu tố không

xuất hiện trong mô hình đến các giá trị của Y. Giả thiết này đương nhiên được thỏa do X là phi ngẫu nhiên.

Giả thiết 6: ε_i tuân theo phân phối chuẩn với trị bình quân của ε_i bằng 0 và phương sai σ^2 hay $\varepsilon_i \sim N(0, \sigma^2)$.

7.2.1.5. Tính chất cho hàm hồi quy mẫu

Hàm hồi quy mẫu $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$ có các tính chất

i) SRF đi qua trung bình mẫu $(\bar{X}, \bar{Y})$, nghĩa là $\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$.

ii) Giá trị trung bình của $\hat{Y}$ bằng với giá trị trung bình quan sát Y, nghĩa là $\overline{\hat{Y}} = \bar{Y}$.

iii) Giá trị trung bình của phần dư e bằng 0, nghĩa là $\bar{e} = \frac{1}{n} \sum_{i=1}^n e_i = 0$.

iv) Phần dư e và $\hat{Y}$ không tương quan, nghĩa là $\sum_{i=1}^n e_i \hat{Y}_i = 0$.

iv) Phần dư e và X không tương quan, nghĩa là $\sum_{i=1}^n e_i X_i = 0$.

7.2.1.6. Hệ số xác định mô hình

Gọi TSS (Total Sum of Squares) là tổng bình phương các sai số giữa giá trị quan sát Y_i với giá trị trung bình của chúng,

$$TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

ESS (Explained Sum of Squares) là tổng bình phương các sai lệch giữa giá trị của Y tính theo hàm hồi quy mẫu với giá trị trung bình,

$$ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_2^2 \sum_{i=1}^n (X_i - \bar{X})^2.$$

RSS (Residual Sum of Squares) là tổng bình phương các sai lệch giữa giá trị quan sát của Y và các giá trị nhận được từ hàm hồi quy mẫu,

$$RSS = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

Khi đó,

TSS đo độ chính xác của số liệu thống kê,

ESS đo độ chính xác của hàm hồi quy so với trung bình, và

RSS đo độ chính xác của hàm hồi quy mẫu so với hàm hồi quy tổng thể.

Nhận xét rằng

$$TSS = ESS + RSS.$$

Nếu các số liệu quan sát của Y đều nằm trên SRF thì $RSS = 0$ và $ESS = TSS$.

Nếu hàm hồi quy mẫu phù hợp tốt với các số liệu quan sát thì ESS sẽ lớn so với RSS.

Ngược lại, nếu hàm hồi quy mẫu kém phù hợp với các số liệu quan sát thì RSS sẽ lớn so với ESS.

Với các nhận xét trên, ta dùng *hệ số xác định* (coefficient of determination)

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

để đo mức độ phù hợp của hàm hồi quy.

Ta có $0 \leq R^2 \leq 1$ và $R^2 = (r_{X,Y})^2$, với $r_{X,Y}$ là hệ số tương quan giữa X và Y. Do đó,

Khi $R^2 = 1$, ta được đường hồi quy “hoàn hảo”, mọi sai lệch của Y (so với trung bình) đều giải thích được bởi mô hình hồi quy.

Khi $R^2 = 0$, X và Y không có quan hệ tuyến tính.

7.2.1.7. Khoảng tin cậy và kiểm định giả thuyết

Khoảng tin cậy cho β_j .

Để ước lượng hệ số hồi quy tổng thể β_j , ta dùng thống kê sau

$$T = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim St(n-2), \quad j=1,2$$

Do $\hat{\beta}_j$ và $se(\hat{\beta}_j)$ tính được từ mẫu nên với độ tin cậy γ cho trước, ta suy ra được mức ý nghĩa $\alpha = 1 - \gamma$, tra trong bảng phân phối Student ta có giá trị tới hạn $C = t_{\frac{\alpha}{2}}^{n-2}$. Từ khoảng tin cậy của T, ta suy ra được khoảng tin cậy cho β_j là

$$\beta_j \in [\hat{\beta}_j - Cse(\hat{\beta}_j); \hat{\beta}_j + Cse(\hat{\beta}_j)], \quad j=1,2.$$

So sánh β_2 với β_2^0 cho trước.

Ta có bài toán kiểm định

$$\begin{cases} H_0 : \beta_2 = \beta_2^0 \\ H_1 : \beta_2 \neq \beta_2^0 \end{cases}$$

Nếu giả thuyết H_0 đúng, ta có thống kê sau

$$T = \frac{\hat{\beta}_2 - \beta_2^0}{\text{se}(\hat{\beta}_2)} \sim \text{St}(n-2)$$

Với nguy cơ sai lầm α cho trước, tra trong bảng phân phối Student, ta có giá trị tới hạn

$$C = t_{\frac{\alpha}{2}}^{n-2}.$$

Nếu $|T| > C$, bác bỏ giả thuyết H_0 .

Nếu $|T| \leq C$, chưa đủ cơ sở bác bỏ H_0 nên ta chấp nhận H_0 .

Đặc biệt, với $\beta_2^0 = 0$, giả thiết H_0 có nghĩa là “*biến độc lập X không ảnh hưởng gì tới biến phụ thuộc Y*”. Khi đó, thống kê tương ứng là

$$T = \frac{\hat{\beta}_2}{\text{se}(\hat{\beta}_2)} \sim \text{St}(n-2).$$

Chẳng hạn, với số liệu của ví dụ 3, ta có $T = \frac{\hat{\beta}_2}{\text{se}(\hat{\beta}_2)} = \frac{1,249407}{0,0388} \approx 32,2$.

Với nguy cơ sai lầm $\alpha = 0,05$, tra bảng phân phối Student với $n-2 = 9-2 = 7$ độ tự do, ta được giá trị $C = 2,365$. Vì $|T| > C$ nên ta bác bỏ H_0 , chấp nhận H_1 và ta kết luận rằng *tỷ lệ lạm phát có ảnh hưởng đến lãi suất ngân hàng*, với nguy cơ sai lầm $\alpha = 0,05$.

7.2.1.8. Kiểm định sự phù hợp của mô hình

Trong phần này, ta khảo sát việc phân tích hồi quy theo quan điểm của phân tích phương sai. Việc phân tích này cung cấp cho ta một phương pháp hữu ích khác trong việc giải quyết vấn đề phán đoán thống kê.

Xuất phát từ tính chất

$$\text{TSS} = \text{ESS} + \text{RSS} \text{ và } R^2 = \frac{\text{ESS}}{\text{TSS}},$$

Từ thống kê này, ta cũng có thể kiểm định giả thiết

$$\begin{cases} H_0 : \beta_2 = 0 \\ H_1 : \beta_2 \neq 0 \end{cases}$$

nghĩa là kiểm định giả thiết cho rằng biến độc lập X không ảnh hưởng gì đến biến phụ thuộc Y. Chú ý rằng khi đó, nghĩa là khi (H_0) đúng, giá trị F còn có thể được tính bằng công thức

$$F = \frac{\hat{\beta}_2^2 n S_X^2}{\hat{\sigma}^2} = \frac{\frac{ESS}{1}}{\frac{RSS}{n-2}} = \frac{\frac{R^2 TSS}{1}}{\frac{(1-R^2)TSS}{n-2}} = \frac{(n-2)R^2}{1-R^2}.$$

Do đó, quá trình phân tích phương sai cho phép ta đưa ra các phán đoán thống kê về độ thích hợp của hàm hồi quy.

Chẳng hạn, với số liệu của ví dụ 3, ta có

$$F = \frac{(n-2)R^2}{1-R^2} = \frac{(9-2) \times 0,9933}{1-0,9933} \approx 1037,7761.$$

Với nguy cơ sai lầm $\alpha = 0,05$, tra bảng phân phối Fisher với độ tự do (1,7), ta được $C = f_{0,05}(1,7) = 5,59$. Vì $F > C$ nên ta bác bỏ H_0 , chấp nhận H_1 , nghĩa là biến độc lập X (tỷ lệ lạm phát) có ảnh hưởng đến biến phụ thuộc Y (lãi suất ngân hàng).

7.2.1.9. Ứng dụng phân tích hồi quy vào dự báo

Ta có thể dùng hàm hồi quy mẫu để dự báo giá trị của biến phụ thuộc Y tương ứng với một giá trị của biến độc lập X . Có hai loại dự báo

Dự báo giá trị trung bình của Y ứng với $X = X_0$.

Dự báo giá trị cá biệt của Y ứng với $X = X_0$.

Dự báo trung bình.

Với $X = X_0$, ta muốn dự báo giá trị trung bình của Y khi biết X lấy giá trị X_0 ,

$$E(Y | X = X_0) = \beta_1 + \beta_2 X_0.$$

Hàm hồi quy mẫu cho ta ước lượng điểm là $\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$, trong đó $\hat{Y}_0$ là ước lượng điểm, không chệch và có phương sai nhỏ nhất của $E(Y | X = X_0)$. Tuy nhiên, $\hat{Y}_0$ vẫn sai khác so với giá trị thực của nó. Để có khoảng tin cậy cho $E(Y | X = X_0)$, ta chú ý rằng $\hat{Y}_0$ có phân phối chuẩn với trung bình $\beta_1 + \beta_2 X_0$ và phương sai

$$\text{Var}(\hat{Y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{n S_X^2} \right].$$

Với mức ý nghĩa α , ta tìm được giá trị $C = t_{\alpha/2}^{n-2}$ trong bảng phân phối Student, suy ra ước lượng khoảng cho giá trị trung bình $E(Y | X = X_0)$

$$E(Y | X = X_0) \in \left[\hat{Y}_0 - C \text{se}(\hat{Y}_0); \hat{Y}_0 + C \text{se}(\hat{Y}_0) \right].$$

Dự báo giá trị riêng biệt Y_0 .

Nếu muốn dự báo giá trị riêng biệt Y_0 khi $X = X_0$, ta chú ý rằng $Y_0 - \hat{Y}_0$ là đại lượng ngẫu nhiên có phân phối chuẩn với trung bình là 0 và phương sai

$$\text{Var}(Y_0 - \hat{Y}_0) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{nS_X^2} \right].$$

Với mức ý nghĩa α , ta tìm được giá trị $C = t_{\alpha/2}^{n-2}$ trong bảng phân phối Student, suy ra ước lượng khoảng cho giá trị riêng biệt Y_0 .

$$Y_0 \in \left[\hat{Y}_0 - C \text{se}(Y_0 - \hat{Y}_0); \hat{Y}_0 + C \text{se}(Y_0 - \hat{Y}_0) \right].$$

7.2.2. Mô hình hồi quy tuyến tính k biến

7.2.2.1. Hàm hồi quy tổng thể

Xét hàm hồi quy tổng thể k biến dạng

$$E(Y|X_2, X_3, \dots, X_k) = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_k X_k,$$

Hay dạng ngẫu nhiên

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_k X_k + \varepsilon$$

7.2.2.2. Ước lượng tham số

Hàm hồi quy mẫu SRF có dạng

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \dots + \hat{\beta}_k X_k,$$

nghĩa là $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2,i} + \hat{\beta}_3 X_{3,i} + \dots + \hat{\beta}_k X_{k,i} + e_i$,

Với phương pháp bình phương nhỏ nhất, các hệ số $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ được chọn sao cho tổng bình phương của các phần dư (RSS) nhỏ nhất, nghĩa là

$$\text{RSS} \equiv \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2,i} - \dots - \hat{\beta}_k X_{k,i})^2$$

đạt giá trị nhỏ nhất.

Ví dụ 4. Số liệu quan sát của một mẫu cho ở bảng sau. Trong đó

Y : Lượng hàng bán được của một loại hàng hóa (tấn / tháng)

X_2 : Thu nhập của người tiêu dùng (triệu / năm)

X_3 : Giá bán của loại hàng này (ngàn đồng / kg)

Y	X_2	X_3	Y	X_2	X_3
---	-------	-------	---	-------	-------

20	8	2	17	6	5
18	7	3	16	5	6
19	8	4	15	5	7
18	8	4	13	4	8
17	6	5	12	3	8

Vậy hàm hồi quy cần tìm là

$$\hat{Y}_i = 14,99215 + 0,76178X_{2i} - 0,58901X_{3i}$$

7.2.2.3. Hệ số xác định hồi quy bội

Hệ số xác định hồi quy bội R^2 có thể tính bằng một trong hai công thức sau

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}$$

Trong đó

$$TSS = Y^T \cdot Y - n(\bar{Y})^2; \quad ESS = \hat{\beta}^T \cdot X^T \cdot Y - n(\bar{Y})^2;$$

$$RSS = TSS - ESS.$$

7.2.2.4. Ma trận tương quan

Giả sử $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_k X_k$. Ký hiệu $r_{t,j}$ chỉ hệ số tương quan giữa biến thứ t và biến thứ j . Với $t=1$ thì $r_{1,j}$ là hệ số tương quan giữa biến Y và biến X_j .

$$r_{1,j} = \frac{\sum_{i=1}^n y_i x_{i,j}}{\sqrt{\sum_{i=1}^n y_i^2 \sum_{i=1}^n x_{j,i}^2}}, \quad r_{t,j} = \frac{\sum_{i=1}^n x_{t,i} x_{j,i}}{\sqrt{\sum_{i=1}^n x_{t,i}^2 \sum_{i=1}^n x_{j,i}^2}},$$

trong đó: $x_{j,i} = X_{j,i} - \bar{X}_j$.

Dễ thấy rằng $r_{t,j} = r_{j,t}$ và $r_{j,j} = 1$ và ma trận hệ số tương quan có dạng

$$R = \begin{pmatrix} 1 & r_{1,2} & \dots & r_{1,k} \\ r_{2,1} & 1 & \dots & r_{2,k} \\ \dots & \dots & \dots & \dots \\ r_{k,1} & r_{k,2} & \dots & 1 \end{pmatrix}.$$

7.2.2.5. Ma trận hiệp phương sai

Để kiểm định giả thiết, tìm khoảng tin cậy cũng như thực hiện các suy đoán thống kê khác, ta cần phải tìm $\text{var}(\hat{\beta}_j)$, với $j=1,2,\dots,k$ và $\text{cov}(\hat{\beta}_i, \hat{\beta}_j)$. Phương pháp ma trận giúp ta có thể thực hiện điều này.

Ma trận hiệp phương sai của $\hat{\beta}$ có dạng tổng quát

$$\text{cov}(\hat{\beta}) = \begin{pmatrix} \text{cov}(\hat{\beta}_1, \hat{\beta}_1) & \text{cov}(\hat{\beta}_1, \hat{\beta}_2) & \dots & \text{cov}(\hat{\beta}_1, \hat{\beta}_k) \\ \text{cov}(\hat{\beta}_2, \hat{\beta}_1) & \text{cov}(\hat{\beta}_2, \hat{\beta}_2) & \dots & \text{cov}(\hat{\beta}_2, \hat{\beta}_k) \\ \dots & \dots & \dots & \dots \\ \text{cov}(\hat{\beta}_k, \hat{\beta}_1) & \text{cov}(\hat{\beta}_k, \hat{\beta}_2) & \dots & \text{cov}(\hat{\beta}_k, \hat{\beta}_k) \end{pmatrix}$$

7.2.2.6. Khoảng tin cậy và kiểm định giả thiết

Khoảng tin cậy cho β_j .

Để ước lượng hệ số hồi quy tổng thể β_j , ta dùng thống kê sau

$$T = \frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} \sim \text{St}(n-2), \quad j=1,2,\dots,k$$

Do $\hat{\beta}_j$ và $\text{se}(\hat{\beta}_j)$ tính được từ mẫu nên với độ tin cậy γ cho trước, ta suy ra được mức ý nghĩa $\alpha=1-\gamma$, tra trong bảng phân phối Student ta có giá trị tới hạn $C=t_{\alpha/2}^{n-k}$. Từ khoảng tin cậy của T , ta suy ra được khoảng tin cậy cho β_j là

$$\beta_j \in [\hat{\beta}_j - C\text{se}(\hat{\beta}_j); \hat{\beta}_j + C\text{se}(\hat{\beta}_j)], \quad j=1,2,\dots,k.$$

Kiểm định giả thiết về các hệ số hồi quy.

Xuất phát từ thống kê

$$T = \frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} \sim \text{St}(n-k).$$

ta có thể kiểm định giả thiết

$$\begin{cases} H_0 : \beta_j = \beta_j^0 \\ H_1 : \beta_j \neq \beta_j^0 \end{cases}$$

với β_j^0 cho trước.

Từ đó, ta dễ dàng kiểm định giả thiết về hệ số hồi quy riêng.