

**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**



# **KINH TẾ LƯỢNG**

*(Dùng cho sinh viên hệ đào tạo đại học từ xa)*

**Lưu hành nội bộ**

**HÀ NỘI - 2006**

# KINH TẾ LƯỢNG

Biên soạn : TS. TRẦN NGỌC MINH

## LỜI NÓI ĐẦU

Nếu như kinh tế vĩ mô và vi mô mô tả sự vận động của nền kinh tế thì *kinh tế lượng* cung cấp các phương pháp phân tích về mặt lượng mối quan hệ giữa các hiện tượng kinh tế cùng với sự tác động qua lại giữa chúng trên cơ sở các số liệu thu thập từ thực tế nhằm củng cố thêm các giả thiết, từ đó đưa ra các quyết định đúng đắn hơn.

Từ nhiều năm nay, cùng với sự phát triển của tin học và máy vi tính, *kinh tế lượng* đã được áp dụng rộng rãi trong kinh tế cũng trong nhiều lĩnh vực khác. Đã có nhiều tạp chí, sách giáo khoa viết về *kinh tế lượng*. Trong số đó phải kể đến các tác giả như: H. Theil (Đại học Chicago), J. Johnston, Daniel, L. Rubinfeld (Đại học California), D. Gujarati (Viện hàn lâm quân sự Hoa kỳ.). Ở Việt nam cũng đã có một số giáo trình *Kinh tế lượng* do một số tác giả viết như “*Kinh tế lượng*” của tác giả: GS.TSKH. Vũ Thiệu; của PGS.TS. Nguyễn Quang Đông, của PGS.TS. Nguyễn Khắc Minh và của tác giả Hoàng Ngọc Mậu,... Những cuốn giáo khoa *kinh tế lượng* đó được trình bày bằng công cụ thống kê toán và ngôn ngữ toán học chặt chẽ và có tính khái quát cao.

Nội dung chủ yếu của môn học này là nhằm giới thiệu:

- Cách thiết lập các mô hình toán học để mô tả mối quan hệ kinh tế, tức là nêu ra các giả thiết hay giả thiết về các mối quan hệ này giữa các biến số kinh tế (chẳng hạn như nhu cầu về số lượng hàng hoá phụ thuộc tuyến tính thuận chiều với thu nhập và ngược chiều với giá cả).
- Ước lượng các tham số nhằm nhận được số đo về sự ảnh hưởng của các biến độc lập khác nhau.
- Kiểm định tính vững chắc của các giả thuyết đó.
- Và cuối cùng, sử dụng các mô hình đó để đưa ra các dự báo hoặc dự đoán và mô phỏng các hiện tượng kinh tế.

*Kinh tế lượng* đã trở thành một bộ phận không thể thiếu được trong chương trình đào tạo cử nhân Kinh tế và Quản trị kinh doanh của các trường đại học. Sự đòi hỏi phải phân tích định lượng các hiện tượng kinh tế trong quá trình hoạch định chính sách vĩ mô, việc dự báo và dự đoán có độ tin cậy cao,... tất cả đã làm cho *kinh tế lượng* có một vai trò ngày càng quan trọng và bản thân nó cũng không ngừng được hoàn thiện và phát triển.

Với nội dung như đã nêu trên “***Sách hướng dẫn học tập môn Kinh tế lượng***” (Dùng cho sinh viên hệ đào tạo đại học từ xa) được biên soạn. Ngoài phần mở đầu, kết cấu gồm 8 chương:

Chương 1: Các khái niệm cơ bản của mô hình hồi quy hai biến.

Chương 2: Ước lượng và kiểm định giả thiết trong mô hình hai biến.

Chương 3: Mô hình hồi quy nhiều biến.

Chương 4: Hồi quy với biến độc lập là biến giả.

Chương 5: Đa cộng tuyến.

Chương 6: Phương sai của sai số thay đổi.

Chương 7; Tự tương quan.

Chương 8: Chọn mô hình và kiểm định việc chỉ định mô hình.

Với nội dung như trên, về cơ bản cuốn sách thống nhất với chương trình quy định của Bộ Giáo dục và Đào tạo cho đối tượng là Đại học Quản trị kinh doanh về môn học *kinh tế lượng*.

Mỗi chương được kết cấu làm 04 phần: Phần giới thiệu chương nhằm giới thiệu khái quát nội dung của chương và yêu cầu đối với người học khi nghiên cứu chương đó. Phần nội dung chương, được biên soạn theo trình tự, kết cấu nội dung của môn học một cách cụ thể, chi tiết, đơn giản giúp cho người học có thể nắm bắt nội dung một cách nhanh chóng. Phần tóm tắt nội dung và những vấn đề cần ghi nhớ, nhằm mục đích nhắc lại các thuật ngữ then chốt, nội dung cốt lõi của chương. Phần câu hỏi và bài tập nhằm củng cố lý thuyết và luyện tập kỹ năng ứng dụng *kinh tế lượng* vào việc giải quyết một bài toán thực tế cụ thể - Đây là phần luyện tập khi sinh viên đã học xong nội dung của chương.

Hy vọng tài liệu này góp phần tạo điều kiện thuận lợi cho sinh viên trong học tập, góp phần nâng cao chất lượng đào tạo.

Mặc dù đã có nhiều cố gắng để biên soạn, trình bày “***Sách hướng dẫn học tập môn Kinh tế lượng***” (dùng cho sinh viên hệ đào tạo đại học từ xa), nhưng không tránh khỏi những thiếu sót. Tác giả rất mong nhận được các ý kiến đóng góp của bạn đọc và các đồng nghiệp. Địa chỉ liên hệ: Bộ môn kinh tế bưu điện - Khoa QTKD1. Học viện Công nghệ Bưu chính-Viễn thông.

***Xin trân trọng cảm ơn!***

**Tác giả**

***TS. Trần Ngọc Minh***

## MỞ ĐẦU

### 1. Khái quát về kinh tế lượng

“Kinh tế lượng” được dịch từ chữ “Econometrics” có nghĩa là “Đo lường kinh tế”. Thuật ngữ này do A.Kragar Frich (Giáo sư kinh tế học người Na uy, đạt giải thưởng Nobel về kinh tế năm 1969) sử dụng lần đầu tiên vào khoảng năm 1930.

Năm 1936, Tiberger, người Hà Lan trình bày trước Hội đồng kinh tế Hà Lan một mô hình kinh tế lượng đầu tiên, mở đầu cho một phương pháp nghiên cứu mới về phân tích kinh tế. Năm 1939, ông xây dựng một số mô hình tương tự cho Mỹ.

Năm 1950, nhà kinh tế được giải thưởng Nobel là Lawrence Klein đã đưa ra một số mô hình mới cho nước Mỹ và từ đó kinh tế lượng được phát triển trên phạm vi toàn thế giới. Hiện nay Lawrence Klein cầm đầu một dự án quốc tế (Link Project) với mô hình kinh tế thế giới dùng để dự báo kinh tế thế giới hàng năm cho Liên hiệp quốc.

Kinh tế lượng là một môn khoa học về đo lường các mối quan hệ kinh tế diễn ra trong thực tế. Kinh tế lượng ngày nay là sự kết hợp giữa lý thuyết kinh tế hiện đại, thông kê toán và máy vi tính, nhằm định lượng các mối quan hệ kinh tế, dự báo khả năng phát triển hay diễn biến của các hiện tượng kinh tế và phân tích nó, làm cơ sở cho việc hoạch định các chính sách kinh tế.

### 2. Xây dựng và áp dụng mô hình kinh tế lượng:

Việc xây dựng và áp dụng mô hình kinh tế lượng được tiến hành theo các bước sau đây:

Bước 1: Nêu vấn đề lý thuyết cần phân tích và các giả thiết về mối quan hệ giữa các biến kinh tế. Chẳng hạn: Khi nghiên cứu mối quan hệ giữa mức tiêu dùng và thu nhập của các hộ gia đình. Theo lý thuyết của kinh tế học vi mô ta có thể nêu giả thiết: mức tiêu dùng của các hộ gia đình phụ thuộc theo quan hệ cùng chiều với thu nhập khả dụng của họ (Thu nhập sau khi trừ thuế và tiết kiệm).

Bước 2: Thiết lập các mô hình toán học để mô tả quan hệ giữa các biến kinh tế. Lý thuyết kinh tế học cho biết quy luật về mối quan hệ giữa các chỉ tiêu kinh tế, nhưng không nêu rõ dạng hàm. Kinh tế lượng phải dựa vào các học thuyết kinh tế để định dạng các mô hình cho các trường hợp cụ thể. Chẳng hạn, khi nghiên cứu mối quan hệ giữa lượng cầu và giá cả của một loại hàng, ta có thể dùng hàm tuyến tính hoặc hàm phi tuyến để diễn tả mối quan hệ này. Giả sử ta chọn đường cầu dạng tuyến tính thì mô hình này có dạng:

$$D = a + bp$$

Trong đó: D là lượng cầu và p là giá cả của loại hàng đó; a, b là các tham số của mô hình. D là biến phụ thuộc hay còn gọi là biến cần được giải thích và p là biến độc lập hay biến giải thích.

Bước 3: Thu thập số liệu.

Khác với các mô hình kinh tế dạng tổng quát, các mô hình kinh tế lượng được xây dựng xuất phát từ *số liệu thực tế*. Trong thống kê toán và kinh tế lượng, người ta phân biệt số liệu của tổng thể và số liệu của mẫu. Số liệu của tổng thể là số liệu của toàn bộ các đối tượng (phần tử) mà ta cần nghiên cứu. Số liệu của mẫu là số liệu của một tập hợp con được lấy ra từ tổng thể. Chẳng

hạn để nghiên cứu nhu cầu về một loại hàng hoá nào đó, thì số liệu tổng thể là số liệu về lượng hàng được mua của tất cả các hộ gia đình ở mọi nơi trong một quốc gia. Trong thực tế ta không có điều kiện để thu thập tất cả số liệu của tổng thể mà chỉ thu thập được số liệu mẫu.

**Bước 4:** Ước lượng các tham số của mô hình. Các ước lượng này là các giá trị thực nghiệm của các tham số trong mô hình. Chúng không những cho các giá trị bằng số mà còn phải thoả mãn các điều kiện, các tính chất mà mô hình đòi hỏi. Trong các trường hợp đơn giản, các tham số thường được ước lượng bằng phương pháp bình phương tối thiểu. Trong các trường hợp phức tạp thì phải dùng các phương pháp khác.

**Bước 5:** Phân tích kết quả: Dựa trên lý thuyết kinh tế để phân tích và đánh giá kết quả nhận được xem có phù hợp với lý thuyết kinh tế hay không. Kiểm định các giả thiết thống kê đối với các ước lượng nhận được (Do các ước lượng được xác định từ số liệu thống kê thực tế).

**Bước 6:** Dự báo: Nếu như mô hình phù hợp với lý thuyết kinh tế thì có thể sử dụng mô hình để dự báo sự phát triển của biến phụ thuộc trong các chu kỳ tiếp theo với sự thay đổi của biến độc lập.

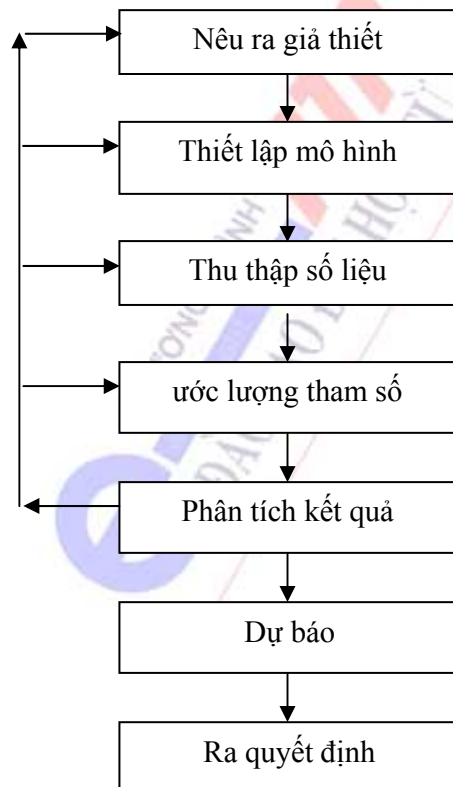
**Bước 7:** Sử dụng mô hình để kiểm tra hoặc đề ra các chính sách kinh tế.

Các bước trên đây có nhiệm vụ khác nhau trong quá trình phân tích một vấn đề kinh tế và chúng được thực hiện theo một trình tự nhất định.

Tìm ra bản chất của vấn đề kinh tế không phải là một việc đơn giản. Vì vậy quá trình trên đây phải được thực hiện lặp lại nhiều lần cho đến khi ta thu được một mô hình phù hợp.

Có thể minh hoạ quá trình phân tích kinh tế lượng bằng một sơ đồ như sau:

Sơ đồ minh hoạ quá trình phân tích kinh tế lượng.



Quá trình xây dựng và áp dụng mô hình kinh tế lượng đòi hỏi trước hết phải có sự hiểu biết về lý thuyết kinh tế học, sau đó là những kiến thức về lý thuyết xác suất và thống kê toán, cuối cùng là các phần mềm của kinh tế lượng. Các kết quả rút ra từ việc phân tích các mô hình kinh tế lượng cũng đòi hỏi phải được suy xét từ nhiều phía. Chẳng hạn các ước lượng cho thấy mối quan hệ nhân quả giữa hai chỉ tiêu kinh tế, nhưng điều đó không chứng minh hay khẳng định là trong thực tế có mối quan hệ nhân quả như vậy. Điều khẳng định phải do người nghiên cứu kinh tế lượng suy xét.

Từ khi ra đời đến nay kinh tế lượng đã cung cấp cho các nhà kinh tế một công cụ sắc bén để đo lường mối quan hệ của các biến kinh tế.

Ngày nay phạm vi ứng dụng của kinh tế lượng đã vượt quá phạm vi kinh tế, lan sang các lĩnh vực khác như xã hội học, vũ trụ học,...

Với sự đòi hỏi phải phân tích định lượng các hiện tượng kinh tế, kiểm định sự phù hợp các giả thiết trong quá trình hoạch định các chính sách, cũng như ra các quyết định tác nghiệp, việc dự báo có độ tin cậy cao,... tất cả đã làm cho kinh tế lượng có một vai trò ngày càng quan trọng, không ngừng hoàn thiện và phát triển.

Sự phát triển của máy tính và tin học đã là tăng thêm sức mạnh cho kinh tế lượng, giúp cho các nhà kinh tế kiểm chứng được các lý thuyết kinh tế có phù hợp hay không để có những quyết định đúng đắn trong hoạt động kinh doanh của doanh nghiệp và hoạch định các chính sách, các chiến lược kinh tế-xã hội.

# CHƯƠNG 1: CÁC KHÁI NIỆM CƠ BẢN CỦA MÔ HÌNH HỒI QUY HAI BIẾN

## GIỚI THIỆU

Hồi quy là một công cụ cơ bản của đo lường kinh tế. Phân tích hồi quy giải quyết những vấn đề cụ thể gì? phân tích hồi khác với các phân tích khác như thế nào? cơ sở thông tin để phân tích hồi quy là gì? vì sao phải xây dựng mô hình hồi quy?... Các vấn đề trên và bản chất của chúng sẽ được đề cập vắn tắt trong chương này. Trong chương này sẽ trình một số vấn đề cơ bản sau:

- Bản chất của phân tích hồi qui.
- Cách xử lý số liệu đầu vào.
- Hàm hồi quy tổng thể (PRF) và hàm hồi quy mẫu(SRF) trong mô hình hồi quy tuyến tính hai biến.

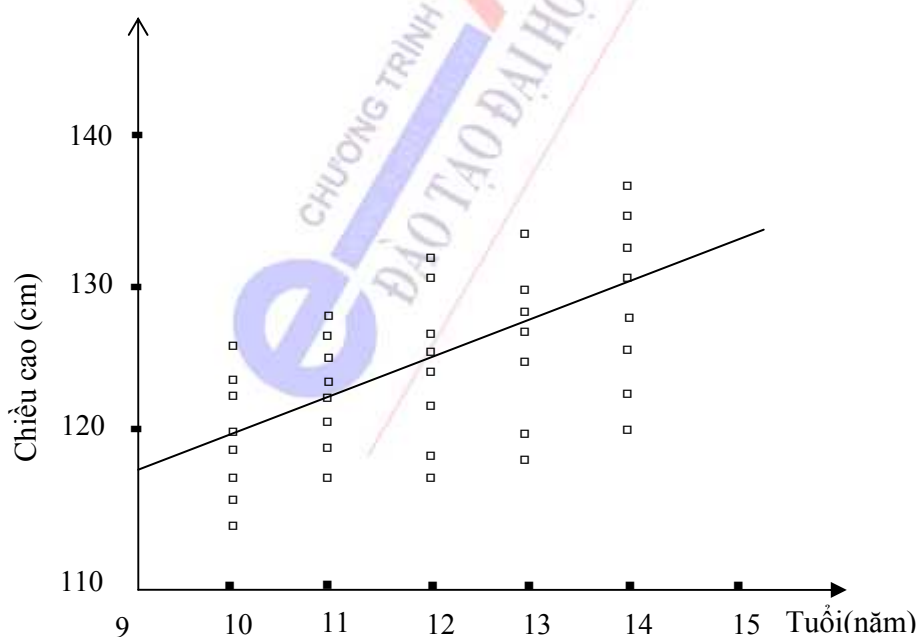
Để có thể nắm bắt được các vấn đề trên yêu cầu người học cần có kiến thức về toán cao cấp, thống kê toán, xác suất và kinh tế học.

## NỘI DUNG

### 1.1 PHÂN TÍCH HỒI QUI

**1. Định nghĩa:** Phân tích hồi quy là nghiên cứu sự phụ thuộc của một biến (biến phụ thuộc), vào một hay nhiều biến khác (các biến giải thích), với ý tưởng là ước lượng (hay dự đoán) giá trị trung bình của biến phụ thuộc trên cơ sở các giá trị biết trước của các biến giải thích.

**Ví dụ:** 1- Xét đồ thị phân tán ở hình 1.1, trong đó mô tả phân phối về chiều cao của học sinh nam tính theo độ tuổi cố định từ 9-15.

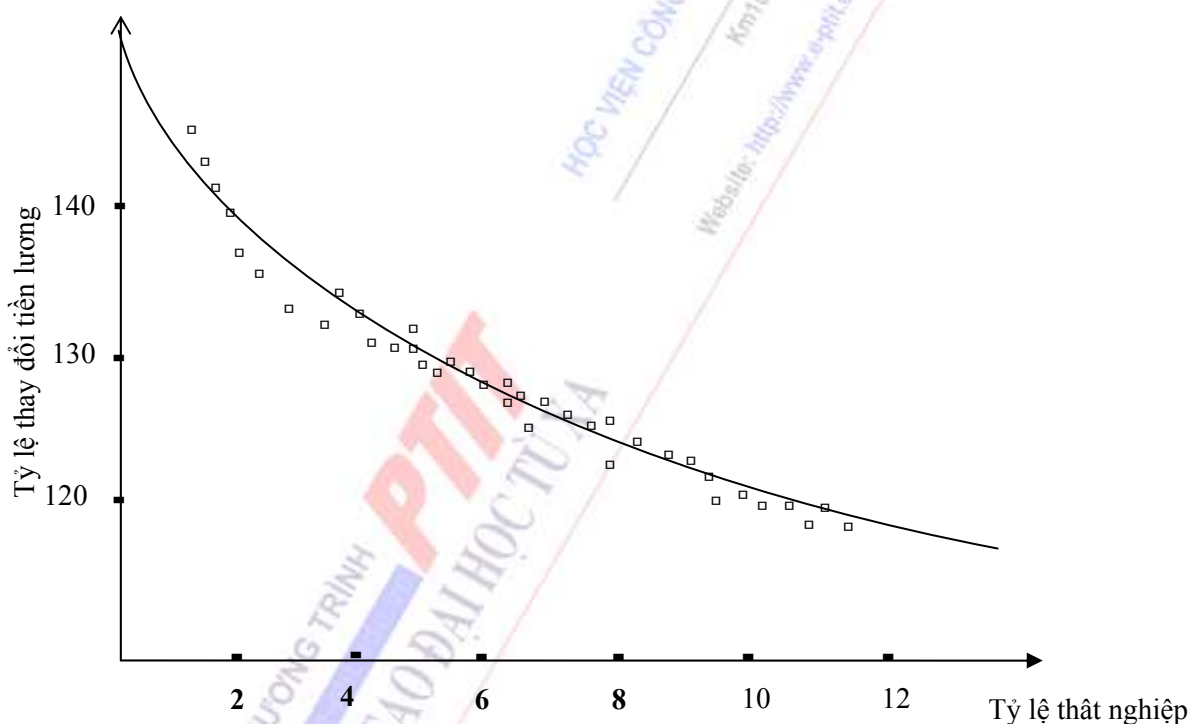


Hình 1.1: Phân phối giả thiết về chiều cao theo độ tuổi.

Rõ ràng không phải tất cả học sinh nam ở một độ tuổi nhất định có xu hướng có cùng chiều cao. Nhưng chiều cao trung bình tăng lên theo độ tuổi (tất nhiên tới độ tuổi nhất định). Như vậy, nếu biết được tuổi, ta có thể dự đoán được chiều cao trung bình tương ứng với độ tuổi đó của học sinh nam.

2- Một nhà kinh tế có thể nghiên cứu sự phụ thuộc của chi tiêu cho tiêu dùng cá nhân vào thu nhập cá nhân thực tế. Một phân tích như vậy có thể có ích trong việc ước lượng xu thế tiêu dùng biên tế (MPC), tức là, mức thay đổi trung bình về chi tiêu cho tiêu dùng khi thu nhập thực tế thay đổi một đơn vị giá trị.

3- Một nhà kinh tế lao động có thể muốn nghiên cứu tỷ lệ thay đổi tiền lương trong mối quan hệ với tỷ lệ thất nghiệp. Các số liệu trong quá khứ được biểu diễn trên đồ thị phân tán như trong hình 1.2 là một thí dụ về đường cong phillips. đường cong này liên quan đến sự thay đổi về tiền lương đối với tỷ lệ thất nghiệp. Căn cứ vào đường cong này có thể cho phép nhà kinh tế lao động dự đoán được mức thay đổi trung bình về tiền lương tại một tỷ lệ thất nghiệp cho trước.



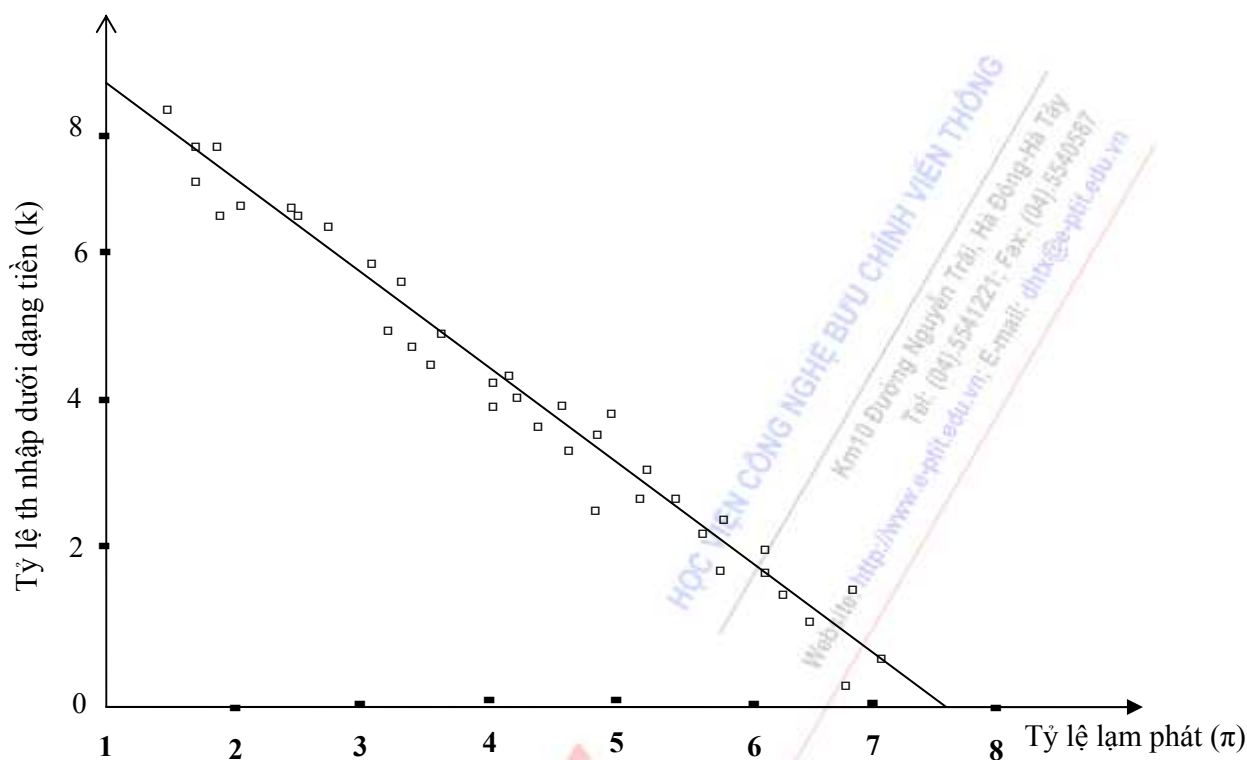
Hình 1.2: Đường cong Phillips giả thiết

Một kiến thức như thế có thể có ích trong việc phân tích quá trình lạm phát kinh tế, bởi vì sự tăng tiền lương thường được phản ánh trong giá cả gia tăng.

4- Một nhà kinh doanh độc quyền có thể định giá cả hay sản lượng (nhưng không thể cả hai), có thể muốn biết phản ứng của mức cầu đối với sản phẩm khi giá cả thay đổi. Một thử nghiệm như vậy có thể đưa tới sự ước lượng độ co giãn về giá cả (nghĩ là tính phản ứng của giá cả) đối với mức cầu của sản phẩm và có thể trợ giúp cho việc xác định mức giá tạo ra lợi nhuận cao nhất.

5- Trong kinh tế học tiền tệ, người ta biết rằng, khi các yếu tố khác không đổi, mức lạm phát ( $\pi$ ) càng cao thì tỷ lệ thu nhập mà người dân muốn giữ dưới dạng tiền mặt ( $k$ ) càng thấp.

Điều này được minh họa trong hình 1.3. Phân tích định lượng về mối quan hệ này sẽ tạo điều kiện cho nhà kinh tế tiền tệ dự đoán được lượng tiền, tính theo tỷ lệ thu nhập, mà người dân muốn giữ dưới dạng tiền mặt ở các mức.



Hình 1.3. Lượng tiền được giữ trong quan hệ với lạm phát.

6- Giám đốc tiếp thị của một công ty muốn biết mức cầu đối với sản phẩm của công ty có quan hệ như thế nào với chi phí quảng cáo. Một nghiên cứu như thế sẽ có ích cho việc xác định độ co giãn của cầu đối với chi phí quảng cáo. Tức là, tỷ lệ phần trăm thay đổi về mức cầu khi ngân sách quảng cáo thay đổi 1%. điều này có thể có ích khi xác định ngân sách quảng cáo “tối ưu”.

Trong thực tế hoạt động kinh doanh có vô số các ví dụ về sự phụ thuộc của một biến vào một hay nhiều biến khác mà người học có thể đưa ra. Các kỹ thuật phân tích hồi quy trình bày trong chương này nhằm nghiên cứu sự phụ thuộc như thế giữa các biến số.

Ta kí hiệu: Y- biến phụ thuộc (hay biến được giải thích)

Xi- biến độc lập (hay biến giải thích) thứ i.

Trong đó, biến phụ thuộc Y là đại lượng ngẫu nhiên, có quy luật phân phối xác suất nào đó. Các biến độc lập Xi không phải là biến ngẫu nhiên, giá trị của chúng được cho trước.

## **2. Nhiệm vụ của phân tích hồi qui:**

- Ước lượng giá trị trung bình của biến phụ thuộc với giá trị đã cho của biến độc lập.
- Kiểm định giả thiết về bản chất của sự phụ thuộc.
- Dự đoán giá trị trung bình của biến phụ thuộc khi biết giá trị của các biến độc lập.
- Kết hợp các vấn đề trên.

### **3- Một số vấn đề cần lưu ý trong phân tích hồi qui:**

#### *a) Phân biệt quan hệ thống kê và quan hệ hàm số:*

Vấn đề mấu chốt trong phân tích hồi qui là sự phụ thuộc thống kê của biến phụ thuộc vào một hay nhiều biến giải thích. Biến phụ thuộc là đại lượng ngẫu nhiên, có phân phối xác suất. Các biến giải thích thì giá trị của chúng đã biết. Biến phụ thuộc là ngẫu nhiên vì có rất nhiều nhân tố tác động đến nó mà ta không thể đưa tất cả các yếu tố đó vào mô hình được. ỨNG VỚI MỖI GIÁ TRỊ ĐÃ BIẾT CỦA BIẾN ĐỘC LẬP CÓ THỂ CÓ NHIỀU GIÁ TRỊ KHÁC NHAU CỦA BIẾN PHỤ THUỘC. Trong quan hệ hàm số các biến không phải là ngẫu nhiên; ứng với mỗi giá trị của biến độc lập có duy nhất một giá trị của biến phụ thuộc. Phân tích hồi qui không nghiên cứu các quan hệ hàm số.

Ví dụ: Doanh thu kinh doanh về một sản phẩm, dịch vụ nào đó phụ thuộc vào giá cả của chính doanh nghiệp, giá của các doanh nghiệp cạnh tranh khác, thị phần của chính doanh nghiệp, thị hiếu của người tiêu dùng, ... là một quan hệ thống kê; Các biến giá cả dịch vụ, thị phần, thị hiếu,... là các biến độc lập; doanh thu dịch vụ là biến phụ thuộc, là đại lượng ngẫu nhiên. Không thể dự báo một cách chính xác doanh thu cho một năm tương lai nào đó vì:

- Có thể có sai số trong dãy số thống kê.
- Có rất nhiều nhân tố khác cũng ảnh hưởng đến doanh thu của dịch vụ mà ta không thể liệt kê hết và nếu có cũng không thể tách được ảnh hưởng riêng của từng nhân tố đến biến doanh thu cho dù ta có đưa thêm vào bao nhiêu biến giải thích khác.

Trong hình học ta đều biết chu vi của hình vuông bằng 4 lần chiều dài của một cạnh, tức  $Y = 4X$ . Trong đó  $Y$  là chu vi của hình vuông và  $X$  là chiều dài của một cạnh hình vuông đó. Ở đây  $X$  và  $Y$  có mối quan hệ hàm số. ứng với mỗi giá trị của  $X$  ta chỉ có một giá trị duy nhất của  $Y$ . Phân tích hồi qui không xét các quan hệ này.

#### *b) Hàm hồi qui và quan hệ nhân quả:*

Phân tích hồi qui nghiên cứu quan hệ giữa một biến phụ thuộc với một hoặc nhiều biến độc lập khác. điều này không đòi hỏi giữa biến phụ thuộc và các biến độc lập phải có mối quan hệ nhân quả. Nếu như quan hệ nhân quả tồn tại thì nó phải được xác lập dựa trên các lý thuyết kinh tế khác. Ví dụ, luật cầu nói rằng trong điều kiện các biến (yếu tố) khác không thay đổi thì nhu cầu một loại hàng hoá tỷ lệ nghịch với giá của hàng hoá này, hay trong ví dụ trên ta có thể dự đoán doanh thu dựa vào giá cả, thị phần, thị hiếu, nhưng không thể dự báo thị hiếu khách hàng dựa trên doanh thu được.

#### *c) Hồi qui và tương quan:*

Hồi qui và tương quan khác nhau về mục đích và kỹ thuật. Phân tích tương quan trước hết là đo mức độ kết hợp tuyến tính giữa hai biến. Ví dụ, mức độ quan hệ giữa nghiệm thuốc lá và ung thư phổi, giữa kết quả thi môn lý và môn toán. Nhưng phân tích hồi qui lại ước lượng hoặc dự báo một biến trên cơ sở giá trị đã cho của các biến khác. Về kỹ thuật, trong phân tích hồi qui các biến không có tính chất đối xứng. Biến phụ thuộc là đại lượng ngẫu nhiên. Các biến giải thích thì giá trị của chúng đã được xác định. Trong phân tích tương quan không có sự phân biệt giữa các biến, chúng có tính chất đối xứng.

## **1.2. BẢN CHẤT VÀ NGUỒN SỐ LIỆU CHO PHÂN TÍCH HỒI QUI.**

Thành công của bất kỳ một sự phân tích kinh tế nào đều phụ thuộc vào việc sử dụng các số liệu thích hợp và phụ thuộc vào phương pháp xử lý các số liệu đó, do vậy phần này sẽ trình bày

đôi nét về bản chất, nguồn gốc và những hạn chế của số liệu mà ta sẽ gặp phải trong phân tích kinh tế nói chung và phân tích hồi qui nói riêng.

### **1- Các loại số liệu**

Có 3 loại số liệu: Các số liệu theo thời gian (chuỗi thời gian), các số liệu chéo và các số liệu hỗn hợp của 2 loại trên.

- Các số liệu theo thời gian là các số liệu được thu thập trong một thời kỳ nhất định. Ví dụ như các số liệu về GDP, GNP, số người thất nghiệp, lượng cung tiền, tổng giá trị sản xuất GO....có số liệu được thu thập hàng tuần, có số liệu thu thập hàng tháng, quý, năm...Các số liệu này có thể được đo bằng những con số như giá cả, thu nhập, nhưng cũng có những số liệu không đo được bằng con số, chung thường là những chỉ tiêu chất lượng như: nam, nữ, có gia đình hay chưa có gia đình, có việc làm hay chưa có việc làm, tốt xấu,...để lượng hoá các biến này, người ta thường sử dụng biến giả (dummy), chúng cũng quan trọng như các biến số được lượng hoá khác.

- Các số liệu chéo là các số liệu về một hoặc nhiều biến được thu thập tại một thời điểm ở nhiều địa phương, đơn vị khác nhau. Ví dụ các số liệu về điều tra dân số vào 0 giờ ngày 1/1/1992; các số liệu điều tra về vốn cơ bản của các xí nghiệp dệt ngày 1/10/1990 ở Việt nam,

- Các số liệu hỗn hợp theo thời gian và không gian: Ví dụ số liệu về giá vàng hàng ngày ở các thành phố Hà Nội, Thành phố HCM, Cần Thơ,...

### **2- Nguồn các số liệu**

Tập hợp các số liệu có thể được thu thập và cung cấp bởi:

- Các cơ quan Nhà nước.
- Các tổ chức quốc tế.
- Các đơn vị sản xuất, kinh doanh.
- Các cá nhân

Chúng có thể là các số liệu thực nghiệm hoặc phi thực nghiệm. Các số liệu thực nghiệm thường được thu thập trong lĩnh vực khoa học tự nhiên. Muốn thu thập số liệu về ảnh hưởng của một nhân tố đến đối tượng nghiên cứu thì cần phải cố định các nhân tố khác có tác động đến đối tượng.

Trong khoa học xã hội, các số liệu thường là phi thực nghiệm. Các số liệu về GDP, GNP, số người thất nghiệp, giá cổ phiếu,... không nằm dưới sự kiểm soát của điều tra viên. Điều này thường gây ra những vấn đề đặc biệt trong việc tìm ra những nguyên nhân chính xác ảnh hưởng đến một chỉ tiêu nào đó. Ví dụ có phải lượng cung về tiền ảnh hưởng đến GDP hay còn nguyên nhân khác?

### **3- Nhược điểm của số liệu**

Như trên đã nêu, yêu cầu về mặt chất lượng của tập hợp số liệu thu thập là phải đảm bảo tính chính xác, kịp thời, đầy đủ. Trong thực tế yêu cầu đó không phải lúc nào cũng có thể thực hiện được, vì những nguyên nhân sau đây:

- Hầu hết các số liệu trong lĩnh vực khoa học xã hội đều là số liệu phi thực nghiệm, do vậy có thể có sai số khi quan sát hoặc bỏ sót quan sát hoặc do cả hai.

- Ngay với các số liệu thu thập bằng thực nghiệm cũng có sai số trong mỗi phép đo.

- Trong các cuộc điều tra bằng câu hỏi, thường gặp tình trạng không nhận được câu trả lời hoặc có trả lời nhưng không trả lời hết các câu hỏi.

- Các mẫu số liệu trong các cuộc điều tra thường không giống nhau về kích thước nên rất khó so sánh kết quả giữa các đợt điều tra.
- Các số liệu về kinh tế thường ở mức tổng hợp cao, không cho phép đi sâu vào các đơn vị nhỏ.
- Ngoài ra một số số liệu quan trọng, cần thiết cho quá trình phân tích, đánh giá lại thuộc về bí mật quốc gia, không thể tiếp cận và thu thập được.

### 1.3 MÔ HÌNH HỒI QUI TỔNG THỂ

Ta xét ví dụ giả định sau:

**Ví dụ 1:** Giả sử ở một địa phương có 60 hộ gia đình và chúng ta quan tâm đến việc nghiên cứu mối quan hệ giữa  $Y$ - chi tiêu tiêu dùng hàng tuần của các gia đình và  $X$  – thu nhập khả dụng hàng tuần của các gia đình. Nói một cách khác là chúng ta muốn dự đoán mức trung bình của chi tiêu tiêu dùng hàng tuần khi biết thu nhập hàng tuần của hộ gia đình. Để thực hiện điều này, giả sử ta chia 60 hộ thành 10 nhóm có thu nhập tương đối như nhau, chênh lệch thu nhập giữa các nhóm là như nhau và bằng 20USD. Các số liệu về mức chi tiêu tương ứng với mức thu nhập của các hộ gia đình được ghi trong bảng 1.2

Bảng 1.2

80	100	120	140	160	180	200	220	240	260
55	65	79	80	102	110	120	135	137	150
60	70	84	93	107	115	136	137	145	152
65	74	90	95	110	120	140	140	155	175
70	80	94	103	116	130	144	152	165	178
75	85	98	108	118	135	145	157	175	180
	88		113	125	140		160	189	185
			115				162		191
325	462	445	707	678	750	685	1043	966	1211

Bảng số liệu trên được giải thích như sau:

Với thu nhập trong một tuần, chẳng hạn  $X = 100$ USD thì có 6 hộ gia đình mà chi tiêu trong tuần của các hộ gia đình trong nhóm này lần lượt là: 65; 70; 74; 80; 85 và 88. tổng chi tiêu trong tuần của 6 hộ gia đình trong nhóm này là 462USD. Như vậy mỗi cột của bảng cho ta một phân phối của chi tiêu trong tuần  $Y$  với mức thu nhập đã cho  $X$ .

Từ số liệu của bảng 1.2 ta dễ dàng tính được xác suất có điều kiện:

Chẳng hạn:  $P(Y = 85/X = 100) = 1/6$ ;  $P(Y = 90/X = 120) = 1/5$ ;....

Bảng tính các xác suất có điều kiện cho trong bảng 1.3

Trong đó:  $E(Y/X_i) = \sum_{j=1}^k Y_j P(Y = Y_j / X = X_i)$  là kỳ vọng toán có điều kiện của  $Y$  (điều kiện là  $X = X_i$ )

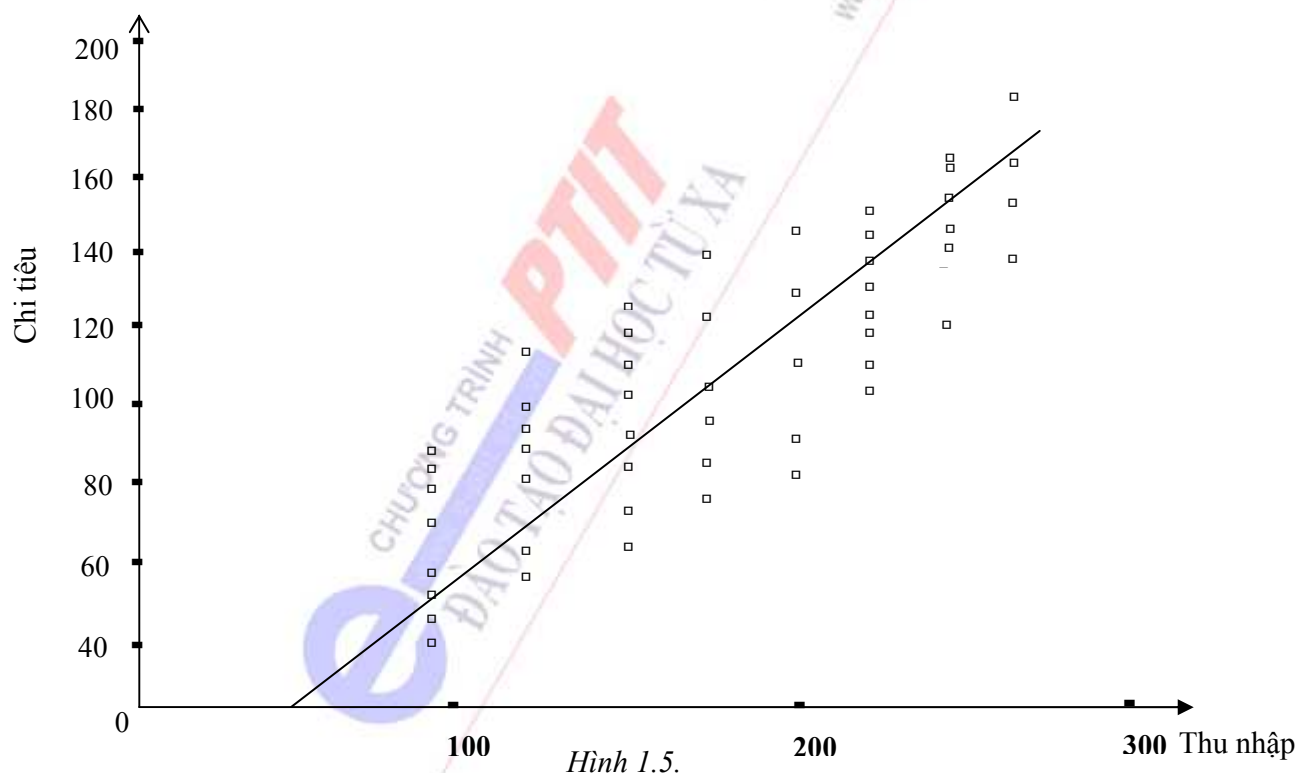
$$\text{Chẳng hạn: } E(Y/100) = 65\frac{1}{6} + 70\frac{1}{6} + 74\frac{1}{6} + 80\frac{1}{6} + 85\frac{1}{6} + 88\frac{1}{6} = 77$$

Bảng 1.3

80	100	120	140	160	180	200	220	240	260
1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/6		1/7	1/6	1/6		1/7	1/6	1/7
			1/7				1/7		1/7
65	77	89	101	113	125	137	149	161	173

$$E(Y/x_i) = \sum_{j=1}^k Y_j P(Y = Y_j / X = X_i)$$

Biểu diễn các điểm  $(X_i; Y_j)$  và các điểm  $M_i(X_i; E(Y/X_i))$  ta được đồ thị sau (Hình 1.5):



Hình 1.5.

Trên hình 1.5 ta thấy trung bình có điều kiện của mức chi tiêu trong tuần nằm trên đường thẳng có hệ số góc dương. Khi thu nhập tăng thì mức chi tiêu cũng tăng. Một cách tổng quát,  $E(Y/X_i)$  là một hàm của  $X_i$ .

$$E(Y/X_i) = f(X_i) \quad (1.1)$$

Hàm (1.1) được gọi là hàm hồi qui tổng thể (*PRF - population regression function*). Nếu PRF có một biến độc lập thì được gọi là hàm *hồi qui đơn (hồi qui 2 biến)*, nếu có từ 2 biến độc lập trở lên thì gọi là hàm *hồi qui bội*.

Hàm hồi qui tổng thể cho ta biết giá trị trung bình của biến Y sẽ thay đổi như thế nào khi biến X nhận các giá trị khác nhau.

Để xác định dạng của hàm hồi qui tổng thể người ta thường dựa vào đồ thị biểu diễn sự biến thiên của dãy các số liệu quan sát về X và Y kết hợp với việc phân tích bản chất của vấn đề nghiên cứu.

Chúng ta xét trường hợp đơn giản nhất là PRF có dạng tuyến tính.

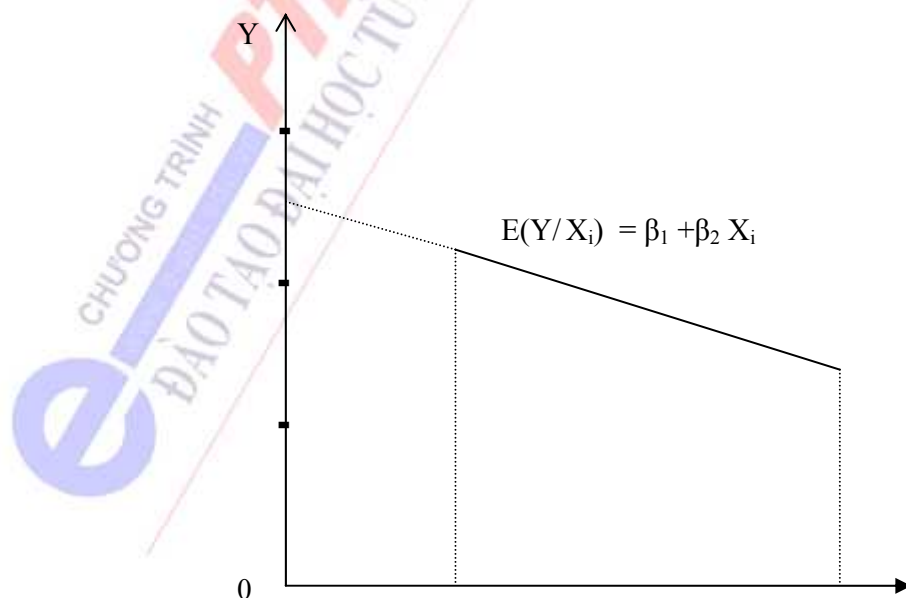
$$E(Y/X_i) = \beta_1 + \beta_2 X_i \quad (1.2)$$

Trong đó:  $\beta_1, \beta_2$  là các tham số chưa biết nhưng cố định, và được gọi là các hệ số hồi qui.

$\beta_1$  là hệ số tự do (hệ số tung độ gốc).  $\beta_1$  cho biết giá trị trung bình của biến phụ thuộc Y là bao nhiêu khi biến độc lập X nhận giá trị 0. Điều này chỉ đúng về mặt toán học, trong các trường hợp cụ thể ta phải kết hợp với lý thuyết kinh tế và điều kiện thực tế của vấn đề nghiên cứu để nêu ý nghĩa của  $\beta_1$  cho phù hợp. Trong thực tế có nhiều trường hợp  $\beta_1$  không có ý nghĩa.

Chẳng hạn, xét hàm:  $E(Y/X_i) = \beta_1 + \beta_2 X_i$

Trong đó Y là lượng hàng bán được của một loại hàng; X là giá của loại hàng đó. Trường hợp này  $\beta_1$  không phải là lượng hàng bán được trung bình khi X (giá bán) bằng 0. Vì trong thực tế không có mặt hàng nào bán với giá bằng 0. Hàm hồi qui nêu trên phản ánh mối quan hệ của lượng hàng bán được và giá bán và hàm này chỉ có ý nghĩa khi X nhận giá trị trong một khoảng ( $X_1; X_2$ ) nào đó. Ngoài khoảng này thì hàm trên không có ý nghĩa. Khi đó ta cần hiểu  $\beta_1$  chỉ là giao điểm của đường thẳng biểu diễn hàm hồi qui nêu trên với trục tung. Ta có thể minh họa bằng hình 1.6 dưới đây.



Hình 1.6

X

$\beta_2$  là *hệ số góc* (hệ số độ dốc),  $\beta_2$  cho biết giá trị trung bình của biến phụ thuộc (Y) sẽ thay đổi (tăng hoặc giảm) bao nhiêu đơn vị khi giá trị của biến độc lập (X) tăng một đơn vị với điều kiện các yếu tố khác không thay đổi.

Thật vậy: giả sử X tăng 1 đơn vị, khi đó giá trị của X sau khi tăng ( $X'_i$ ) sẽ bằng giá trị của X trước khi tăng ( $X_i$ ) cộng với 1. Tức ta có  $X'_i = X_i + 1$ . Khi đó:

$$E(Y/X'_i) = \beta_1 + \beta_2 X'_i = \beta_1 + \beta_2(X_i + 1) = \beta_1 + \beta_2 X_i + \beta_2 = E(Y/X_i) + \beta_2$$

Nếu  $\beta_2 > 0$  thì  $E(Y/X'_i) > E(Y/X_i)$  khi đó giá trị trung bình của Y sẽ tăng.

Nếu  $\beta_2 < 0$  thì  $E(Y/X'_i) < E(Y/X_i)$  khi đó giá trị trung bình của Y sẽ giảm.

$E(Y/X_i)$  là trung bình của Y với điều kiện X nhận giá trị  $X_i$ .

Thuật ngữ “tuyến tính” ở đây được hiểu theo hai nghĩa: tuyến tính đối với tham số và tuyến tính đối với các biến.

Ví dụ:  $E(Y/X_i) = \beta_1 + \beta_2 X_i^2$  là hàm tuyến tính đối với tham số. Nhưng không tuyến tính đối với biến.

$E(Y/X_i) = \beta_1 + \sqrt{\beta_2} X_i$  là hàm tuyến tính đối với biến nhưng phi tuyến đối với tham số.

Hàm hồi qui tuyến tính luôn được hiểu là tuyến tính đối với các tham số, nó có thể không tuyến tính đối với biến.

Giá trị quan sát thứ i của biến phụ thuộc Y được kí hiệu là  $Y_i$

Kí hiệu  $U_i$  là chênh lệch giữa  $Y_i$  và  $E(Y/X_i)$ :

$$U_i = Y_i - E(Y/X_i)$$

$$\text{Hay: } Y_i = E(Y/X_i) + U_i \quad (1.3)$$

## 1.4 SAI SỐ NGẪU NHIÊN VÀ BẢN CHẤT.

### 1. Sai số ngẫu nhiên

Như đã trình bày ở trên  $U_i$  là chênh lệch giữa giá trị quan sát  $Y_i$  với giá trị trung bình của nó tính theo hàm hồi qui.  $U_i$  là đại lượng ngẫu nhiên,  $U_i$  có thể nhận giá trị âm hoặc dương, người ta gọi  $U_i$  là sai số ngẫu nhiên (hoặc nhiễu) và (1.3) được gọi là hàm hồi qui tổng thể ngẫu nhiên.

### 2. Bản chất của sai số ngẫu nhiên

Sự tồn tại của  $U_i$  bởi một số lý do sau đây:

- Ngoài  $X_i$  đã được đưa vào mô hình, rất có thể còn có các biến khác chưa xem xét tới cũng có ảnh hưởng tới  $Y_i$ , nên  $U_i$  đại diện cho các biến đó.
- Thiếu số liệu của các biến đưa vào mô hình.
- Đơn giản hoá quá trình tính toán.

Tóm lại  $U_i$  giữ vai trò quan trọng trong phân tích hồi qui, chúng phải thoả mãn những điều kiện nhất định thì việc phân tích mô hình hồi qui mới thực sự có ý nghĩa.

## 1.5 HÀM HỒI QUI MẪU:

Trong thực tế, nhiều khi ta không có điều kiện để điều tra toàn bộ tổng thể. Khi đó ta chỉ có thể ước lượng giá trị bình của biến phụ thuộc từ số liệu của mẫu. Hơn nữa cũng vì lý do trên mà việc xây dựng hàm hồi qui tổng thể gây tốn kém về thời gian và kinh phí một cách không cần thiết. Trong thống kê học đã đưa ra phương pháp điều tra chọn mẫu, cho phép lấy ra từ tổng thể chung một số mẫu số liệu nhất định để nghiên cứu, phân tích và suy rộng kết quả (ước lượng) cho tổng thể chung với một xác suất tin cậy cho trước. Việc xây dựng hàm hồi qui mẫu (*SRF* - *the sample regression function*) cũng dựa trên nguyên tắc đó, nghĩa là từ số liệu mẫu ta tiến hành xây dựng hàm hồi qui mẫu và dùng nó để ước lượng các tham số cho hàm hồi qui tổng thể. Tổng thể bao gồm các số liệu mẫu thường được gọi là tổng thể mẫu.

Giả sử từ một tổng thể chung có  $N$  phần tử (đơn vị tổng thể) ta lấy ra từng mẫu  $n$  phần tử. Như vậy sẽ có tất cả  $C_N^n$  cách lấy mẫu, trong đó  $C_N^n$  là tổ hợp chập  $n$  của  $N$  phần tử được xác định theo công thức:

$$C_N^n = \frac{N!}{n!(N-n)!} \quad (1.4)$$

Như vậy, có bao nhiêu lần chọn mẫu, ta có bấy nhiêu hàm hồi qui mẫu. Vấn đề đặt ra là đường hồi qui mẫu nào là thích hợp với PRF. Câu hỏi này chưa trả lời được bởi lẽ PRF chưa biết. Cũng giống như ước lượng một tham số, ta sẽ ước lượng PRF bằng SRF mà SRF này có tính chất: tuyến tính, không chệch và có phương sai nhỏ nhất.

Nếu hàm hồi qui tổng thể có dạng tuyến tính thì hàm hồi qui mẫu có dạng:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \quad (1.5)$$

Trong đó:  $\hat{Y}$  : là ước lượng điểm của  $E(Y/X_i)$

$\hat{\beta}_1$  : là ước lượng điểm của  $\beta_1$ .

$\hat{\beta}_2$  : là ước lượng điểm của  $\beta_2$ .

Dạng ngẫu nhiên của (1.5):

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i \quad (1.6)$$

Trong đó:  $e_i$  là ước lượng điểm của  $U_i$  và gọi là phần dư.

## TÓM TẮT NỘI DUNG CHƯƠNG 1

Bản chất của hồi quy là nghiên cứu sự phụ thuộc của một biến (biến phụ thuộc), vào một hay nhiều biến khác (biến giải thích), với ý tưởng là ước lượng (hay dự đoán) giá trị trung bình của biến phụ thuộc trên cơ sở các giá trị biết trước của các biến giải thích.

Nhiệm vụ của phân tích hồi quy là ước lượng giá trị trung bình của biến phụ thuộc với giá trị đã cho của biến độc lập; Kiểm định giả thiết về bản chất của sự phụ thuộc; Dự báo giá trị của biến phụ thuộc khi biết giá trị của các biến độc lập và kết hợp các vấn đề trên.

Phân tích hồi quy chỉ nghiên cứu chỉ nghiên cứu quan hệ thống kê giữa các biến.

Để có kết quả sát với thực tế cần phân biệt các loại số liệu và ưu nhược điểm và cách xử lý nguồn số liệu.

Hàm hồi qui tuyến luôn được hiểu là tuyến tính đối với các tham số, nó có thể không tuyến tính đối với biến.

Hàm hồi quy tổng thể là hàm được nghiên cứu trên toàn bộ tổng thể. Hàm hồi quy mẫu là hàm được xây dựng trên cơ sở một mẫu. Sử dụng hàm hồi quy mẫu ta ước lượng được giá trị trung bình của biến phụ thuộc từ số liệu của một mẫu.

Hàm hồi qui tổng thể ngẫu nhiên:  $Y_i = E(Y/X_i) + U_i$

Hàm hồi qui mẫu dạng ngẫu nhiên:  $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$

## CÂU HỎI VÀ BÀI TẬP ÔN CHƯƠNG I

### I. Câu hỏi lý thuyết:

- Hãy đưa ra một ví dụ về mối liên hệ thống kê giữa biến phụ thuộc với một hay một số biến độc lập trong thực tế kinh doanh của ngành BC-VT?
- Phân biệt sự khác nhau giữa hàm hồi qui tổng thể và hàm hồi qui mẫu? Sử dụng hàm hồi qui mẫu để ước lượng giá trị trung bình của biến phụ thuộc có những ưu nhược điểm gì?

### II. Bài tập:

- Dữ liệu của Y (chi tiêu tiêu dùng cá nhân) và X (tổng sản phẩm quốc nội GDP), từ 1980 – 1991, tất cả tính bằng tỷ đô la năm 1987 như sau:

Năm	Y	X	Năm	Y	X
1980	2447,1	3776,3	1986	2969,1	4404,5
1981	2476,9	3843,1	1987	3052,2	4539,9
1982	2503,7	3760,3	1988	3162,4	4718,6
1983	2619,4	3906,6	1989	3223,3	4838,0
1984	2746,1	4148,5	1990	3260,4	4877,5
1985	2865,8	4279,8	1991	3240,8	4821,0

a/ Hãy vẽ đồ thị phân tán với trục tung là Y và trục hoành là X và cho nhận xét?

b/ Ngoài GDP còn có yếu tố nào, hay các biến nào có thể ảnh hưởng đến chi tiêu tiêu dùng cá nhân?

- Các mô hình sau đây có tuyến tính theo các tham số hay tuyến tính theo các biến? Mô hình nào là mô hình hồi qui tuyến tính?

a)  $Y_i = \beta_1 + \beta_2 \left( \frac{1}{X_i} \right) + U_i$       b)  $Y_i = \beta_1 + \beta_2 \ln X_i + U_i$

c)  $\ln Y_i = \beta_1 + \beta_2 X_i + U_i$       d)  $\ln Y_i = \ln \beta_1 + \beta_2 \ln X_i + U_i$

e)  $\ln Y_i = \beta_1 - \beta_2 \left( \frac{1}{X_i} \right) + U_i$       f)  $Y_i = \beta_1 + \beta_2^3 X_i + U_i$

- Hãy biến đổi các mô hình sau đây về mô hình hồi qui tuyến tính:

$$a) Y_i = \frac{1}{1 + e^{\hat{\beta}_1 + \hat{\beta}_2 X_i + U_i}}$$

$$b) Y_i = e^{\hat{\beta}_1 + \hat{\beta}_2 X_i + U_i}$$

$$c) \ln Y_i = \hat{\beta}_1 + \hat{\beta}_2 \frac{1}{X} + U_i$$

$$d) Y = \frac{1}{\beta_1 + \beta_2 X}$$

$$e) Y = \frac{X}{\beta_1 + \beta_2 X}$$

$$f) Y = \frac{1}{1 + \exp(-\beta_1 - \beta_2 X)}$$

## CHƯƠNG 2: ƯỚC LƯỢNG VÀ KIỂM ĐỊNH GIẢ THIẾT TRONG MÔ HÌNH HỒI QUI HAI BIẾN

### GIỚI THIỆU

Trong chương này sẽ trình bày vấn đề ước lượng hàm hồi qui tổng thể (PRF) trên cơ sở số liệu của một mẫu. Thực chất là xác định các tham số trong hàm hồi qui mẫu, trên cơ sở các giả thiết tiến hành ước lượng và kiểm định các giả thiết, từ đó xây dựng hàm hồi qui tổng thể. Có nhiều phương pháp ước lượng hàm hồi qui tổng thể. Trong thực tế thường sử dụng phương pháp bình phương nhỏ nhất hoặc phương pháp OLS (*Ordinary Least Square*). Các nội dung chính của chương:

- Cách ước lượng các tham số của hàm hồi qui bằng phương pháp OLS.
- Các giả thiết của phương pháp OLS.
- Cách tính phương sai và sai số chuẩn của các ước lượng.
- Cách xác định hệ số tương quan và hệ số xác định, tính chất và ý nghĩa của các hệ số đó.
- Cách xác định không tin cậy của các tham số trong hàm hồi qui tổng thể và phương sai của nó.
- Phương pháp kiểm định giả thiết về các hệ số hồi qui.
- Phương pháp kiểm định sự phù hợp của hàm hồi qui. Phân tích hồi qui và phân tích phương sai.
- Ứng dụng phân tích hồi qui, trình bày kết quả và đánh giá kết quả của phân tích hồi qui.

Yêu cầu người học cần có các kiến thức về xác suất thống kê toán, toán cao cấp, lý thuyết kinh tế học.

### NỘI DUNG

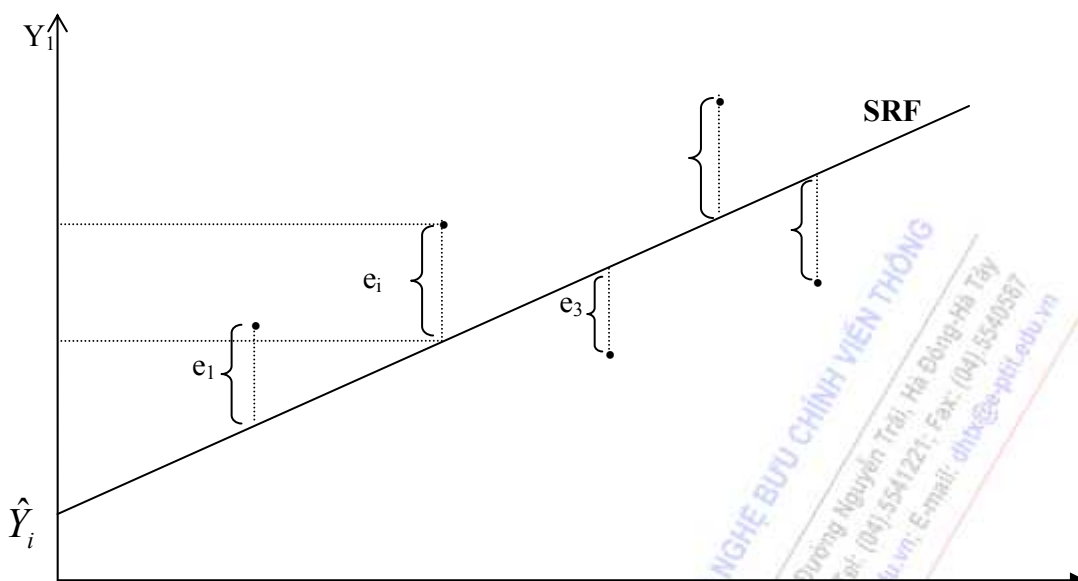
#### 2.1 PHƯƠNG PHÁP BÌNH PHƯƠNG NHỎ NHẤT.

Để tìm hàm  $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$  ta dùng phương pháp OLS do nhà toán học Đức là Carl Friedrich Gauss đưa ra. Nội dung của phương pháp như sau:

Giả sử chúng ta có một mẫu gồm  $n$  cặp quan sát  $(Y_i, X_i)$ ,  $i = 1 \div n$ . Theo phương pháp bình phương nhỏ nhất, ta phải tìm  $\hat{Y}_i$  sao cho nó càng gần với giá trị thực  $(Y_i)$  càng tốt, tức phần dư:

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i \text{ càng nhỏ càng tốt.}$$

Ta có thể minh họa bằng đồ thị sau:



Hình 2.1

Do  $e_i$  ( $i = \overline{1, n}$ ) có thể dương, có thể âm, nên ta cần tìm SRF sao cho tổng bình phương của các phần dư đạt cực tiểu. Tức  $\hat{\beta}_1, \hat{\beta}_2$  phải thỏa mãn điều kiện:

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \Rightarrow \min \quad (*)$$

Điều kiện (\*) có nghĩa là tổng bình phương các sai lệch giữa giá trị thực tế quan sát được ( $Y_i$ ) và giá trị tính theo hàm hồi quy mẫu ( $\hat{Y}_i$ ) là nhỏ nhất. Về mặt hình học, biểu thức (\*) phản ánh tổng bình phương các khoảng cách từ các điểm quan sát tới đường hồi quy mẫu là nhỏ nhất, tức đường hồi quy mẫu với  $\hat{\beta}_1, \hat{\beta}_2$  thỏa mãn điều kiện (\*) sẽ là đường thẳng “gần nhất” với tập hợp các điểm quan sát, do vậy nó được coi là đường thẳng “tốt nhất”, “phù hợp nhất” trong lớp các đường hồi quy mẫu có thể dùng để ước lượng hàm (2.1)

Do  $Y_i, X_i$  ( $i = \overline{1, n}$ ) đã biết, nên  $\sum_{i=1}^n (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$  là hàm số của  $\hat{\beta}_1, \hat{\beta}_2$ .

Vì vậy, ta cần tìm  $\hat{\beta}_1, \hat{\beta}_2$  sao cho:

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

Tức  $\hat{\beta}_1, \hat{\beta}_2$  là nghiệm của hệ phương trình sau:

$$\begin{cases} \frac{\partial f(\hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_1} = \sum_{i=1}^n 2(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)(-1) = 0 \\ \frac{\partial f(\hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_2} = \sum_{i=1}^n 2(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)(-X_i) = 0 \end{cases}$$

Hay:

$$\begin{cases} n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^n X_i = \sum_{i=1}^n Y_i \\ \hat{\beta}_1 \sum_{i=1}^n X_i + \hat{\beta}_2 \sum_{i=1}^n X_i^2 = \sum_{i=1}^n X_i Y_i \end{cases} \quad (2.1)$$

Hệ phương trình (2.1) gọi là hệ phương trình chuẩn. Giải hệ phương trình này ta được:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \cdot \bar{Y}}{\sum_{i=1}^n X_i^2 - n (\bar{X})^2} \quad (2.2)$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} \quad (2.3)$$

Ta cũng có thể tính  $\hat{\beta}_2$  theo công thức sau đây:

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} \quad (2.4)$$

Trong đó:  $x_i = X_i - \bar{X}$ ;  $y_i = Y_i - \bar{Y}$

**Thí dụ 1:** Bảng sau đây cho số liệu về mức chi tiêu tiêu dùng (Y-đôla/tuần) và thu nhập hàng tuần (X-đôla/tuần) của một mẫu gồm 10 gia đình. Giả sử Y và X có mối quan hệ tương quan tuyến tính. Hãy ước lượng hàm hồi quy của Y theo X.

$Y_i$	70	65	90	95	110	115	120	140	155	150
$X_i$	80	100	120	140	160	180	200	220	240	260

**Giải:** Từ các số liệu quan sát của X và Y cho ở bảng trên ta tính được:

$$\sum Y_i = 1110; \sum X_i = 1700; \sum X_i^2 = 322000; \sum X_i Y_i = 205500;$$

$$\bar{Y} = \frac{1110}{10} = 111; \quad \bar{X} = \frac{1700}{10} = 170$$

$$\sum_{i=1}^n x_i y_i = \sum_{i=1}^n X_i Y_i - n \bar{X} \cdot \bar{Y} = 205500 - 10 \times 170 \times 111 = 16800$$

$$\sum_{i=1}^n x_i^2 = \sum_{i=1}^n X_i^2 - n (\bar{X})^2 = 322000 - 10 \cdot (170)^2 = 33000$$

Vậy:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \frac{16800}{33000} = 0,5091 \quad (2.6)$$

$$\hat{\beta}_1 = 111 - 0,509091(170) = 24,4545$$

Vậy hàm hồi qui tuyến tính mẫu của chi tiêu cho tiêu dùng theo thu nhập là:

$$\hat{Y}_i = 24,4545 + 0,5091X_i$$

Giá trị  $\hat{\beta}_1 = 24,4545$  là tung độ gốc của đường hồi qui mẫu, chỉ mức chi tiêu tiêu dùng trung bình hàng tuần khi mà thu nhập hàng tuần bằng 0. Tuy nhiên đây là sự giải thích máy móc số hạng tung độ gốc. Trong phân tích hồi qui, cách giải thích theo nghĩa đen của số hạng tung độ gốc như thế này không phải lúc nào cũng có ý nghĩa, mặc dù trong ví dụ chúng ta đang xét, nó có thể được lập luận rằng một hộ gia đình không có bất cứ thu nhập nào (do thất nghiệp, bị sa thải,...) có thể duy trì mức chi tiêu tiêu dùng tối thiểu (hoặc từ vay mượn, hoặc từ tiết kiệm,...). Nhưng nói chung người ta phải sử dụng độ nhạy cảm trong việc giải thích số hạng tung độ gốc đối với X nhận các giá trị trong một khoảng nào đó khi quan sát. Với ví dụ mà ta đang xét thì không thể coi số 0 là một trong các giá trị quan sát của X.

Giá trị  $\hat{\beta}_2 = 0,5091$  chỉ ra rằng, xét các giá trị của X nằm trong khoảng (80; 260), khi thu nhập tăng 1 USD/tuần thì chi tiêu tiêu dùng của hộ gia đình tăng trung bình khoảng 0,51 USD/tuần.

## 2.2 CÁC GIẢ THIẾT CƠ BẢN CỦA PHƯƠNG PHÁP BÌNH PHƯƠNG NHỎ NHẤT.

Trong phân tích hồi qui, mục đích của chúng ta là ước lượng, dự báo về tổng thể, tức là ước lượng  $E(Y/X_i)$ ,  $\hat{\beta}_1$ ,  $\hat{\beta}_2$  tìm được bằng phương pháp OLS là các ước lượng điểm của  $\beta_1$ ,  $\beta_2$ . Chúng ta chưa biết chất lượng của các ước lượng này như thế nào.

Chất lượng của các ước lượng phụ thuộc vào:

- Dạng hàm của mô hình được lựa chọn.
- Phụ thuộc vào các  $X_i$  và  $U_i$ .
- Phụ thuộc vào kích thước mẫu.

Về dạng của mô hình chúng ta sẽ đề cập ở phần sau. ở đây chúng ta sẽ nói về các giả thiết đối với  $X_i$  và  $U_i$ . Theo các giả thiết này thì các ước lượng tìm được bằng phương pháp OLS là tuyến tính, không chệch và có phương sai nhỏ nhất.

**Giả thiết 1** Các biến độc lập  $X_i$  là phi ngẫu nhiên, tức là giá trị của chúng được xác định trước. Giả thiết này là đương nhiên, vì phân tích hồi qui được đề cập là phân tích hồi qui có điều kiện, phụ thuộc vào các giá trị  $X_i$  đã cho.

**Giả thiết 2** Kỳ vọng của yếu tố ngẫu nhiên  $U_i$  bằng 0., tức là  $E(U_i/X_i) = 0$ . Giả thiết này có nghĩa là các yếu tố không có trong mô hình,  $U_i$  đại diện cho chúng và không có ảnh hưởng một cách có hệ thống đến giá trị trung bình của  $\hat{Y}_i$ . Có thể nói các giá trị  $U_i$  dương triệt tiêu với các giá trị  $U_i$  âm sao cho trung bình của chúng ảnh hưởng lên  $\hat{Y}_i$  bằng 0.

**Giả thiết 3** Các  $U_i$  ( $i = \overline{1, n}$ ) có phương sai bằng nhau, tức là:

$$\text{Var}(U_i/X_i) = \text{var}(U_j/X_j) = \sigma^2 \quad \forall i \neq j. \quad (2.7)$$

Giả thiết này có nghĩa là phân phối có điều kiện của Y với giá trị đã cho của X có phương sai bằng nhau, các giá trị cá biệt của Y xoay quanh giá trị trung bình với mức độ chênh lệch như nhau.

**Giả thiết 4** Không có sự tương quan giữa các  $U_i$ :

$$\text{Cov}(U_i, U_j) = 0 \quad \forall i \neq j. \quad (2.8)$$

Giả thiết này có nghĩa  $U_i$  là ngẫu nhiên. Sai số ở quan sát này không ảnh hưởng tới sai số ở quan sát khác.

**Giả thiết 5**  $U_i$  và  $X_i$  không tương quan với nhau:

$$\text{Cov}(U_i, X_i) = 0 \quad (2.9)$$

Giả thiết 5 là cần thiết vì nếu  $U$  và  $X$  có tương quan với nhau thì ta không thể tách ảnh hưởng riêng biệt của chúng đến  $Y$ , trong khi đó  $U$  lại đại diện cho các yếu tố không có mặt trong mô hình. Giả thiết 5 sẽ thỏa mãn nếu  $X$  là phi ngẫu nhiên.

**Định lý Gauss- Markov:** Với các giả thiết từ 1÷ 5 của phương pháp OLS, các ước lượng của phương pháp bình phương nhỏ nhất sẽ là các ước lượng tuyến tính, không chệch và có phương sai nhỏ nhất trong lớp các ước lượng tuyến tính không chệch.

Đối với hàm hồi quy 2 biến, theo định lý trên thì  $\hat{\beta}_1, \hat{\beta}_2$  tương ứng là các ước lượng tuyến tính, không chệch và có phương sai nhỏ nhất của  $\beta_1, \beta_2$ .

## 2.3 ĐỘ CHÍNH XÁC CỦA CÁC ƯỚC LƯỢNG BÌNH PHƯƠNG NHỎ NHẤT.

Theo phương pháp OLS, các ước lượng  $\hat{\beta}_1, \hat{\beta}_2$  được xác định theo công thức (2.6). Các ước lượng này là đại lượng ngẫu nhiên, với các mẫu khác nhau ta có ước lượng khác nhau. Vì phương sai hay độ lệch chuẩn đặc trưng cho độ phân tán của đại lượng ngẫu nhiên, nên ta dùng chúng để đo chất lượng của các ước lượng. Với các giả thiết của phương pháp OLS, phương sai và độ lệch chuẩn của các ước lượng được xác định bởi các công thức sau:

$$\text{Var}(\hat{\beta}_2) = \frac{\delta^2}{n \sum_{i=1}^n x_i^2} \quad (2.12) \quad \text{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^n X_i^2}{n \sum_{i=1}^n x_i^2} \delta^2 \quad (2.10)$$

$$\text{se}(\hat{\beta}_1) = \sqrt{\text{Var}(\hat{\beta}_1)} \quad (2.11) \quad \text{se}(\hat{\beta}_2) = \sqrt{\text{Var}(\hat{\beta}_2)} \quad (2.13)$$

Trong đó:  $\delta^2 = \text{Var}(U_i)$  se: là sai số chuẩn.

Trong các công thức trên, nếu  $\delta^2$  chưa biết thì  $\delta^2$  được ước lượng bằng ước lượng không chệch của nó là  $\hat{\delta}^2$

$$\hat{\delta}^2 = \frac{\sum_{i=1}^n e_i^2}{n - 2} \quad (2.14) \quad \hat{\delta} = \sqrt{\hat{\delta}^2} \text{ là sai số tiêu chuẩn.} \quad (2.15)$$

## 2.4 HỆ SỐ $r^2$ ĐO ĐỘ PHÙ HỢP CỦA HÀM HỒI QUY MẪU.

### 1. Công thức xác định hệ số $r^2$ (Hệ số xác định):

$$\text{Ta ký hiệu: } TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - n(\bar{Y})^2 \quad (2.16)$$

TSS (*Total Sum of Squares*) là tổng bình phương của tất cả các sai lệch giữa các giá trị quan sát  $Y_i$  với giá trị trung bình của chúng.

$$ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = (\hat{\beta}_2)^2 \sum_{i=1}^n x_i^2 \quad (2.17)$$

ESS (*Explained Sum of Squares*) là tổng bình phương của tất cả các sai lệch giữa giá trị của biến Y tính theo hàm hồi qui mẫu với giá trị trung bình. Phần này đo độ chính xác của hàm hồi qui.

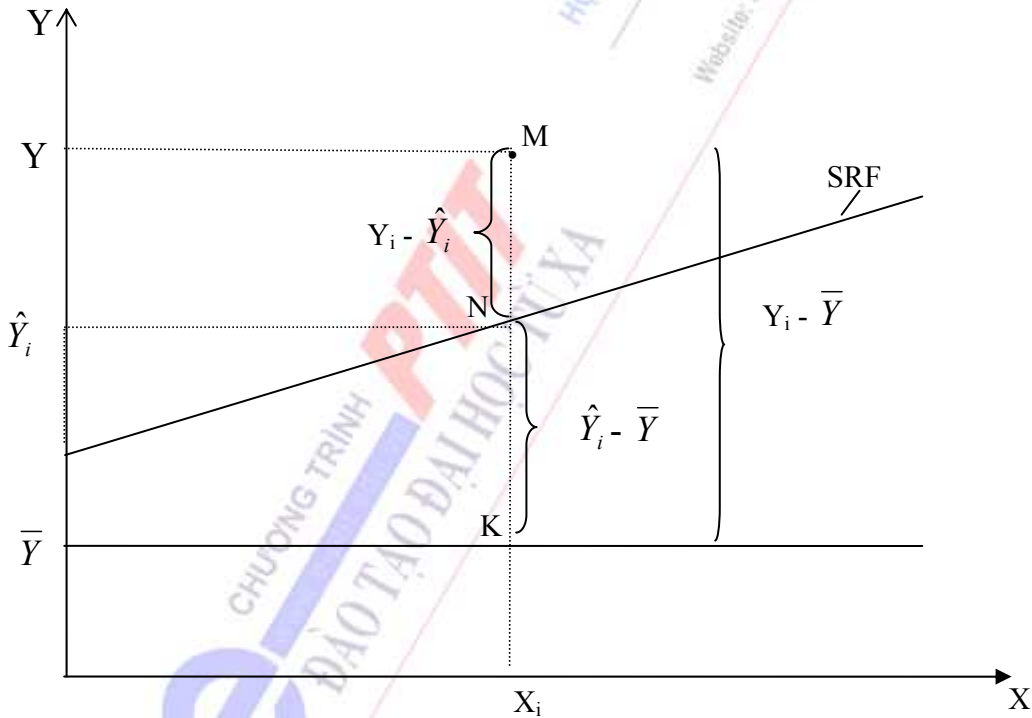
$$RSS = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (2.18)$$

RSS (*Residual Sum of Squares*) là tổng bình phương của tất cả các sai lệch giữa các giá trị quan sát của biến Y và các giá trị nhận được từ hàm hồi qui mẫu.

$$\text{Người ta đã chứng minh được rằng: } TSS = ESS + RSS \quad (2.19)$$

Nếu hàm hồi qui mẫu phù hợp tốt với các số liệu quan sát thì ESS sẽ càng lớn hơn RSS. Nếu tất cả các giá trị quan sát của Y đều nằm trên SRF thì ESS sẽ bằng TSS và do đó RSS = 0. Ngược lại, nếu hàm hồi qui mẫu kém phù hợp với các giá trị quan sát thì RSS sẽ càng lớn hơn ESS.

Về mặt hình học ta có thể minh họa điều nhận xét trên bằng hình (2.2)



$$\text{Ta định nghĩa: } r^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \text{ hoặc } r^2 = \hat{\beta}_2^2 \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n y_i^2} \quad (2.20)$$

$$\text{Trong đó: } \sum_{i=1}^n x_i^2 = \sum_{i=1}^n (X_i - \bar{X})^2; \sum_{i=1}^n y_i^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

$$\text{Ta có: } 0 \leq r^2 \leq 1 \quad (2.21)$$

Đặc biệt với  $r^2 = 1$  thì đường hồi qui mẫu phù hợp “hoàn hảo”, tất cả các sai lệch của Y (so với giá trị trung bình) đều giải thích được bởi mô hình hồi qui. Khi  $r^2 = 0$  chúng tỏ X và Y không có quan hệ.

Đại lượng  $r^2$  gọi là hệ số xác định (*coefficient of determination*) và được sử dụng để đo mức độ phù hợp của hàm hồi qui mẫu

Từ số liệu cho ở ví dụ 2 ta tính được:  $\sum_{i=1}^n Y_i^2 = 132100$

$$TSS = 132100 - 10 \cdot (111)^2 = 8890; ESS = (0,0591)^2 \cdot 33000 = 8552,73.$$

Vậy: 
$$r^2 = \frac{8552,73}{8990} = 0,9621$$

Kết quả này có nghĩa là trong hàm hồi qui mẫu, biến X (thu nhập) giải thích 96,21% sự thay đổi của biến Y (chi tiêu tiêu dùng). Do vậy có thể nói rằng trong trường hợp này mức độ phù hợp của SRF khá cao.

## **2- Công thức xác định hệ số tương quan(r):**

Nếu mô hình hồi qui mẫu có dạng  $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$  là hợp lý và đáng tin cậy, thì khả năng giữa  $X_i$  và  $Y_i$  trong tổng thể mẫu sẽ tồn tại một mối liên hệ tương quan tuyến tính. Tuy nhiên, mức độ tương quan đó thế nào là điều chúng ta cần quan tâm xem xét.

Thông thường, để xét mức độ chặt chẽ của quan hệ tuyến tính giữa X và Y, người ta sử dụng hệ số tương quan mẫu, ký hiệu là r và được xác định bằng công thức:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{X})^2 \sum_{i=1}^n (y_i - \bar{Y})^2}} \quad (2.22)$$

Hay: 
$$r = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}} \quad (2.23)$$

Có thể chứng minh được:  $r = \pm \sqrt{r^2} \quad (2.24)$

Trong trường hợp này dấu của r trùng với dấu của  $\hat{\beta}_2$

Với số liệu cho ở ví dụ 2, vì  $\hat{\beta}_2 = 0,0591 > 0$  nên:

$$r = \sqrt{0,0591} = 0,981$$

Các tính chất của hệ số tương quan:

- Hệ số tương quan có thể nhận các giá trị từ -1 đến +1.
- Nếu  $r > 0$ :  $X_i$  và  $Y_i$  có mối tương quan thuận.
- Nếu  $r < 0$ :  $X_i$  và  $Y_i$  có mối tương quan nghịch.

•  $\hat{\beta}_2 = 0$  thì  $r = 0$  và ngược lại. Ngoài ra, tham số  $\hat{\beta}_2$  và hệ số tương quan r luôn phù hợp với nhau về dấu. Vì vậy có thể căn cứ vào dấu của  $\hat{\beta}_2$  để nhận biết tính thuận nghịch của mối tương quan.

• Giá trị tuyệt đối  $|r|$  càng gần 1, thì mối liên hệ tương quan giữa  $X_i$  và  $Y_i$  càng chặt chẽ. Trường hợp  $|r| = 1$  chứng tỏ giữa hai đại lượng có quan hệ hàm số.

• Giá trị tuyệt đối  $|r|$  càng gần 0, thì mối quan hệ tương quan giữa  $X_i$  và  $Y_i$  càng lỏng lẻo. Trường hợp  $|r| = 0$  chứng tỏ giữa hai đại lượng không có quan hệ tương quan tuyến tính, hoặc chúng độc lập với nhau.

## 2.5 PHÂN BỐ XÁC SUẤT CỦA YẾU TỐ NGẪU NGHIÊN:

Mục đích của phân tích hồi qui không phải là chỉ suy đoán về  $\beta_1, \beta_2$  hay PRF mà còn phải kiểm tra bản chất của sự phụ thuộc, còn phải thực hiện các dự đoán khác. Nếu tiến hành lấy mẫu nhiều lần, ta sẽ nhận được tập hợp nhiều giá trị khác nhau của  $\hat{\beta}_1, \hat{\beta}_2$  tạo thành một đại lượng ngẫu nhiên. Ta cần phải tìm hiểu và nghiên cứu luật phân phối xác suất của hai loại tham số này. Các phân phối này phụ thuộc vào phân phối của  $U_i$ . để giải quyết vấn đề này, ta cần bổ sung thêm giả thiết sau:

Giả thiết 6:  $U_i$  có phân phối chuẩn với kỳ vọng bằng 0 và phương sai là  $\delta^2$ , tức là  $N(0, \delta^2)$ . Khi đó các tham số mẫu  $\hat{\beta}_1, \hat{\beta}_2$  và  $\delta^2$  có các tính chất sau:

- Chúng là các ước lượng không chệch.
- Có phương sai cực tiểu.
- Khi số quan sát đủ lớn thì các ước lượng này xấp xỉ giá trị thực của phân phối.
- $\hat{\beta}_1 \sim N(\beta_1, \delta_{\beta_1}^2)$  từ tính chất này suy ra:  $Z = \frac{\hat{\beta}_1 - \beta_1}{\delta_{\beta_1}} \sim N(0,1)$
- $\hat{\beta}_2 \sim N(\beta_2, \delta_{\beta_2}^2)$  từ tính chất này suy ra:  $Z = \frac{\hat{\beta}_2 - \beta_2}{\delta_{\beta_2}} \sim N(0,1)$ .
- Đại lượng ngẫu nhiên  $\frac{(n-2)\hat{\delta}^2}{\delta^2}$  có luật phân phối  $\chi^2$  với độ tự do là  $(n-2)$ , tức là  $\frac{(n-2)\hat{\delta}^2}{\delta^2} \sim \chi^2(n-2)$ .
- Hàm hồi qui mẫu  $\hat{Y}_i$  phụ thuộc vào  $\hat{\beta}_1, \hat{\beta}_2$  và  $\hat{\delta}^2$  có phân phối chuẩn và kỳ vọng toán là đường hồi qui lý thuyết trong tổng thể chung, tức là  $\hat{Y}_i \sim N(\beta_1 + \beta_2 X_i, \delta^2)$ .

## 2.6 KHOẢNG TIN CẬY VÀ KIỂM TRA GIẢ THIẾT VỀ CÁC HỆ SỐ HỒI QUI.

### 1. Khoảng tin cậy của các hệ số hồi qui:

$\hat{\beta}_1, \hat{\beta}_2$  mà chúng ta đã tìm được ở phần trên là ước lượng điểm của  $\beta_1, \beta_2$ . Ước lượng này có độ tin cậy như thế nào? Như chúng ta đã biết, một ước lượng đơn có nhiều khả năng khác với giá trị đúng. Trong thống kê, độ tin cậy của một ước lượng điểm được đo bằng sai số chuẩn của nó. Do vậy, thay vì chỉ dựa vào ước lượng điểm, ta có thể xây dựng một khoảng xung quanh giá

trị ước lượng điểm, để xác suất mà giá trị đúng của tham số cần ước lượng nằm trong khoảng này là  $1 - \alpha$ , tức là:

$$P(\hat{\beta}_1 - \varepsilon \leq \beta_1 \leq \hat{\beta}_1 + \varepsilon) = 1 - \alpha. \text{ Hoặc } P(\hat{\beta}_2 - \varepsilon \leq \beta_2 \leq \hat{\beta}_2 + \varepsilon) = 1 - \alpha. \quad (2.25)$$

Vì  $\hat{\beta}_1, \hat{\beta}_2$  là đại lượng ngẫu nhiên nên khoảng  $(\hat{\beta}_i - \varepsilon; \hat{\beta}_i + \varepsilon)$  ( $i = 1 \div 2$ ) là khoảng ngẫu nhiên;  $1 - \alpha$  được gọi là hệ số tin cậy (hay độ tin cậy).

$\alpha$  ( $0 \leq \alpha \leq 1$ ) được gọi là mức ý nghĩa.  $\varepsilon$  được gọi là độ chính xác của ước lượng. ( $\varepsilon \geq 0$ ).

$\hat{\beta}_i - \varepsilon$  được gọi là giới hạn tin cậy dưới và  $\hat{\beta}_i + \varepsilon$  được gọi là giới hạn tin cậy trên.

Biểu thức (2.25) mang ý nghĩa là: nếu ta tiến hành xây dựng khoảng tin cậy  $(\hat{\beta}_i - \varepsilon; \hat{\beta}_i + \varepsilon)$  nhiều lần với hệ số tin cậy  $1 - \alpha$  thì tính trung bình, có  $100(1 - \alpha)$  phần trăm số lần các khoảng này chứa giá trị đúng của  $\beta_i$  với ( $i = 1 \div 2$ ).

Quá trình xác định khoảng tin cậy của  $\beta_i$  với ( $i = 1 \div 2$ ) được tiến hành như sau:

#### Bước 1:

Xác định các tham số đặc trưng của tổng thể mẫu  $\hat{\beta}_1, \hat{\beta}_2$  theo công thức (2.6), xác định phương sai và sai số chuẩn của các tham số đó bằng công thức từ (2.10) đến (2.13).

#### Bước 2:

Từ giả thiết 6 ta có  $Z = \frac{\hat{\beta}_i - \beta_i}{\delta_{\beta_i}} \sim N(0,1)$  với  $i = 1 \div 2$ . Tạo biến  $t_1, t_2$  là các biến chuẩn hoá đối với  $\hat{\beta}_1, \hat{\beta}_2$  để đưa về dạng phân phối chuẩn hoá  $N(0,1)$ . Trong đó:

$$t_1 = \frac{\hat{\beta}_1 - \beta_1}{\text{Se}(\hat{\beta}_1)} \sim T(n-2)$$

$$\text{Và } t_2 = \frac{\hat{\beta}_2 - \beta_2}{\text{Se}(\hat{\beta}_2)} \sim T(n-2) \quad (2.26)$$

#### Bước 3: Xác định và đánh giá khoảng tin cậy:

Các giá trị khác nhau của  $t_1, t_2$  là những đại lượng ngẫu nhiên tuân theo luật phân phối chuẩn, còn các giá trị lý thuyết tương ứng của chúng được liệt kê trong bảng T, có ký hiệu  $t_{\alpha/2(n-2)}$ , được hiểu là: với độ tin cậy  $p$  hoặc ở ngưỡng xác suất  $\alpha = 1 - p$  nào đó, ta tra bảng T ứng với giá trị  $\alpha/2$  và độ tự do  $(n-2)$  sẽ nhận được giá trị tới hạn  $t_{\alpha/2(n-2)}$  dùng để xác định khoảng tin cậy của  $\hat{\beta}_1, \hat{\beta}_2$ . Vì  $t_{\alpha/2(n-2)}$  là giá trị tới hạn nên  $p = 1 - \alpha$  chính là xác suất để các giá trị  $t_1$ , hoặc  $t_2$  không vượt quá giá trị này. Có nghĩa là:  $|t_i| \leq t_{\alpha/2(n-2)} \rightarrow P(t_i \leq t_{\alpha/2(n-2)}) = 1 - \alpha$  hay:

$$P(-t_{\alpha/2(n-2)} \leq t_i \leq t_{\alpha/2(n-2)}) \text{ với } i = 1 \div 2 \quad (2.27)$$

$$\text{Vì } t_i = \frac{\hat{\beta}_i - \beta_i}{\text{Se}(\hat{\beta}_i)} \text{ nên suy } P(-t_{\alpha/2(n-2)} \leq \frac{\hat{\beta}_i - \beta_i}{\text{Se}(\hat{\beta}_i)} \leq t_{\alpha/2(n-2)}) = 1 - \alpha$$

với  $i = 1 \div 2$ . Biến đổi biểu thức trong ngoặc ta nhận được;

$$(\hat{\beta}_i - t_{\alpha/2(n-2)} * Se(\hat{\beta}_i) \leq \beta_i \leq \hat{\beta}_i + t_{\alpha/2(n-2)} * Se(\hat{\beta}_i)) \text{ với } i = 1 \div 2. \quad (2.28)$$

Đây chính là khoảng tin cậy hay miền chấp nhận của  $\beta_i$  ( $i = 1 \div 2$ ) với xác suất tin cậy  $1 - \alpha$ . Ta có thể viết ngắn gọn như sau:

Với hệ số tin cậy  $1 - \alpha$ , khoảng tin cậy của  $\beta_i$  ( $i = 1 \div 2$ ) là:

$$\hat{\beta}_i \pm t_{\alpha/2(n-2)} Se(\hat{\beta}_i) \text{ với } (i=1 \div 2) \quad (2.29)$$

Trong đó:  $t_{\alpha/2}$  là giá trị của đại lượng ngẫu nhiên T phân phối theo qui luật Student với bậc tự do  $(n-2)$  sao cho  $P|T| > t_{\alpha/2} = \alpha$ .

Có thể tra  $t_{\alpha/2}$  bằng hàm TINV trong Excel. Ví dụ: Với bậc tự do là  $n-2 = 8$ ,  $\alpha = 5\%$  thì;

$$t_{(0,025,8)} = \text{TINV}(0,05, 8) = 2,201.$$

Ví dụ: Với số liệu trong ví dụ 2, ở phần trên ta tính được:  $RSS = 8890 - 8552,73 = 337,27$

Vậy:

$$Var(\hat{\beta}_2) = \frac{42,15875}{33000} = 0,0012775$$

$$se(\hat{\beta}_2) = \sqrt{Var(\hat{\beta}_2)} = \sqrt{0,0012775} = 0,035742$$

$$Var(\hat{\beta}_1) = \frac{322000}{10 \cdot 33000} \cdot 42,15875 = 41,13672$$

$$se(\hat{\beta}_1) = \sqrt{Var(\hat{\beta}_1)} = \sqrt{41,13672} = 6,4138$$

Với độ tin cậy 95% thì  $t_{\alpha/2(n-2)} = t_{0,025(8)} = 2,201$

Vậy khoảng tin cậy của  $\beta_1$ , là:

$$24,4545 \pm 2,201 \cdot 6,4138 \text{ hay } 10,33773 < \beta_1 < 38,57127$$

Khoảng tin cậy của  $\beta_2$  là:

$$0,5091 \pm 2,201 \cdot 0,035742 \text{ hay } 0,43043 < \beta_2 < 0,58777.$$

Kết quả trên có nghĩa là: Với điều kiện các yếu tố khác không đổi, khi thu nhập tăng 1 USD/tuần thì chi tiêu tiêu dùng trung bình của một hộ gia đình tăng trong khoảng từ 0,43043 đến 0,58777 USD/tuần.

## **2. Khoảng tin cậy của $\delta^2$ :**

Ở giả thiết 6 ta có  $\chi^2 = \frac{(n-2)\delta^2}{\sigma^2}$  đây là đại lượng ngẫu nhiên phân phối theo qui luật  $\chi^2(n-2)$ . Do đó khoảng tin cậy của  $\delta^2$  (với hệ số tin cậy  $1 - \alpha$ ) được xác định từ biểu thức:

$$P\left(\chi_{1-\alpha/2}^2 \leq \frac{(n-2)\hat{\delta}^2}{\delta^2} \leq \chi_{\alpha/2}^2\right) = 1 - \alpha$$

Hay:

$$P\left(\frac{(n-2)\hat{\delta}^2}{\chi_{\alpha/2}^2} \leq \delta^2 \leq \frac{(n-2)\hat{\delta}^2}{\chi_{1-\alpha/2}^2}\right) = 1 - \alpha \quad (2.30)$$

Trong đó:  $\chi_{\alpha/2}^2, \chi_{1-\alpha/2}^2$  là các giá trị của đại lượng ngẫu nhiên  $\chi^2$  phân phối theo luật “khi bình phương” với bậc tự do là  $n-2$  thỏa mãn điều kiện:  $P(\chi^2 > \chi_{\alpha/2}^2) = \alpha/2$ ;

$P(\chi^2 > \chi_{1-\alpha/2}^2) = 1 - \alpha/2$ . Để tìm các giá trị này ta tra bảng (Phần phụ lục) hoặc dùng hàm CHINV trong Excel. Giá trị của phân phối  $\chi_a^2(k)$  được cho trong bảng phần phụ lục.

### 3. Kiểm định giả thiết về các hệ số hồi qui:

Kiểm định giả thiết thống kê được phát biểu đơn giản như sau: *Kết quả tìm được dựa trên số liệu thu thập từ thực tế có phù hợp với một giả thiết nêu ra hay không?* Từ “phù hợp” được dùng ở đây được hiểu là “đủ” sát với giá trị giả thiết nêu ra để ta không bác bỏ giả thiết đã nêu. Như vậy, nếu căn cứ vào một lý thuyết hay kinh nghiệm từ trước là ta tin rằng hệ số góc ( $\beta_2$ ) trong ví dụ 1 bằng 0,8 thì giá trị quan sát ( $\hat{\beta}_2$ ) là 0,5091 được tính từ mẫu có phù hợp với giả thiết phát biểu không?

Nếu phù hợp ta không bác bỏ giả thiết; nếu không phù hợp thì ta bác bỏ giả thiết nêu trên.

Trong thống kê toán, giả thiết phát biểu (giả thiết cần kiểm định) được gọi là *giả thiết không* và kí hiệu là  $H_0$ . Một mệnh đề đối lập với  $H_0$  được gọi là *giả thiết đối* và được kí hiệu là  $H_1$ . Chẳng hạn, giả thiết không là:  $H_0: \beta_2 = 0,8$ ; khi đó giả thiết đối có thể là  $H_1: \beta_2 \neq 0,8$ , hoặc  $H_1: \beta_2 > 0,8$ , hoặc  $H_1: \beta_2 < 0,8$ .

Lý thuyết kiểm định xây dựng các qui tắc hay thủ tục để quyết định bác bỏ hay không bác bỏ giả thiết không. Có hai cách tiếp cận bổ sung lẫn nhau để xây dựng qui tắc đó, gọi là khoảng tin cậy và kiểm định ý nghĩa. Cả hai phương pháp này đều dựa trên cơ sở: Đã xác định được qui luật phân phối xác suất của đại lượng ngẫu nhiên được dùng là tiêu chuẩn kiểm định (thống kê kiểm định). Phần lớn các giả thiết mà ta tiến hành kiểm định là đưa ra các phát biểu hay khẳng định liên quan đến (các) giá trị hay (các) tham số đặc trưng của thống kê kiểm định.

#### a/ Kiểm định giả thiết bằng phương pháp khoảng tin cậy:

Để minh họa cho phương pháp này, ta trở lại với ví dụ tiêu dùng – thu nhập đã xét ở phần trên. Giả sử ta cho rằng giá trị đúng của  $\beta_2$  là 0,3 tức ta kiểm định giả thiết  $H_0: \beta_2 = 0,3$ ; với  $H_1: \beta_2 \neq 0,3$ .

Giả thiết đối nêu trên là giả thiết hai phía.  $\hat{\beta}_2$  quan sát được có “phù hợp” với giả thiết  $H_0$  hay không? Để trả lời câu hỏi này, ta có thể căn cứ vào khoảng tin cậy của  $\beta_2$  mà ta đã tìm được ở phần trên:

$$(0,43043 < \beta_2 < 0,58777)$$

Như vậy: với xác suất  $1 - \alpha$ , khoảng ngẫu nhiên  $(\hat{\beta}_2 - \varepsilon; \hat{\beta}_2 + \varepsilon)$  chứa được  $\beta_2$ . Vì  $1 - \alpha$  khá lớn, nên theo nguyên lý xác suất lớn, ta có thể coi biến cố  $(\hat{\beta}_2 - \varepsilon; \hat{\beta}_2 + \varepsilon)$  hầu như chắc chắn xảy ra trong một phép thử. Với mẫu cụ thể (được coi là một phép thử) ta sẽ tìm được khoảng  $(\hat{\beta}_2 - \varepsilon; \hat{\beta}_2 + \varepsilon)$ . Nếu thấy giá trị  $\beta_2$  nằm trong khoảng này thì không bác bỏ  $H_0$ ; ngược lại nếu  $\beta_2$  nằm ngoài khoảng này thì bác bỏ  $H_0$ . Ta có thể minh họa qui tắc trên bằng hình sau:

Bác bỏ giả thiết $H_0$ nếu $\beta_2$ nằm trong miền này	Các giá trị của $\beta_2$ nằm trong khoảng này là hợp lý theo $H_0$ với độ tin cậy $1 - \alpha$ . Do vậy không bác bỏ $H_0$ nếu $\beta_2$ nằm trong miền này	Bác bỏ giả thiết $H_0$ nếu $\beta_2$ nằm trong miền này
$\hat{\beta}_2 - t_{\alpha/2} Se(\hat{\beta}_2)$		$\hat{\beta}_2 + t_{\alpha/2} Se(\hat{\beta}_2)$

**Qui tắc quyết định:** Thiết lập một khoảng tin cậy (với hệ số tin cậy  $1 - \alpha$ ) cho  $\beta_2$ . Nếu  $\beta_2$  (theo  $H_0$ ) nằm trong khoảng tin cậy này thì không bác bỏ giả thiết  $H_0$ ; Nếu  $\beta_2$  nằm ngoài khoảng này thì ta bác bỏ  $H_0$ .

Theo qui tắc này, trong ví dụ giả thiết  $H_0$  là:  $\beta_2 = 0,3$ . Vì  $\beta_2$  nằm ngoài khoảng (0,43043; 0,58777). Do vậy ta bác bỏ giả thiết  $H_0$ . (với mức ý nghĩa 5%).

**Kiểm định một phía hay một đuôi:**

Đôi khi ta có một tiên nghiệm hay kỳ vọng lý thuyết mạnh rằng giả thiết đối là một phía hay theo một hướng chứ không phải theo hai phía như vừa xét ở trên. Chẳng hạn, trong ví dụ tiêu dùng – thu nhập, nếu dựa vào lý thuyết kinh tế hay một công trình nghiên cứu thực nghiệm trước đây cho thấy xu thế tiêu dùng biên lớn hơn 0,3, khi đó ta có thể nêu giả thiết đối như sau:  $H_1: \beta_2 > 0,3$ . Thủ tục kiểm định giả thiết này có thể được suy ra một cách dễ dàng từ (2.28), nhưng trong thực tế, để kiểm định giả thiết này, ta thường áp dụng phương pháp kiểm định ý nghĩa.

**b/ Kiểm định giả thiết bằng phương pháp kiểm định ý nghĩa:**

Kiểm định ý nghĩa là một thủ tục mà các kết quả của mẫu được sử dụng để kiểm chứng tính đúng đắn hay sai lầm của một giả thiết không.

Chẳng hạn cần kiểm định giả thiết  $H_0: \beta_2 = \beta_2^*$ ;  $H_1: \beta_2 \neq \beta_2^*$ .

Quyết định chấp nhận hay bác bỏ  $H_0$  dựa vào giá trị của thống kê kiểm định thu được từ số liệu của mẫu.

Trong giả thiết 6, ta có đại lượng ngẫu nhiên

$$t_i = \frac{\hat{\beta}_2 - \beta_2}{Se(\hat{\beta}_2)}$$

Tuân theo phân phối  $t$  với  $n-2$  bậc tự do. Nếu giá trị của  $\beta_2$  đúng như giả thiết không đã nêu thì giá trị của  $t$  có thể được tính từ mẫu đã cho,  $t$  đóng vai trò là thống kê kiểm định.. Từ đó ta có khoảng tin cậy như sau:

$$(-t_{\alpha/2} \leq \frac{\hat{\beta}_2 - \beta_2^*}{Se(\hat{\beta}_2)} \leq t_{\alpha/2}) = 1 - \alpha \quad (2.31)$$

Với  $\beta_2^*$  là giá trị của  $\beta_2$  theo  $H_0$ .

(2.31) biểu thị khoảng chứa  $t$  với xác suất  $1 - \alpha$  nếu như  $\beta_2 = \beta_2^*$ . Theo ngôn ngữ kiểm định giả thiết, khoảng  $(-t_{\alpha/2}; t_{\alpha/2})$  thiết lập ở (2.31) được gọi là miền chấp nhận của giả thiết không (với mức ý nghĩa  $\alpha$ ). Vùng nằm ngoài miền chấp nhận được gọi là miền bác bỏ của  $H_0$ .  $t_{\alpha/2}$  được gọi là giá trị tới hạn;  $\alpha$  được gọi là mức ý nghĩa của kiểm định ( $\alpha$  chính là xác suất mắc phải sai lầm loại I, sai lầm mắc phải khi ta bác bỏ giả thiết không khi nó đúng, Nguyên nhân mắc sai lầm là do  $\alpha$ . Bên cạnh đó có sai lầm loại II, đó là sai lầm khi giả thiết  $H_0$  sai nhưng lại được chấp nhận. Chúng ta muốn sao cho xác suất mắc sai lầm loại II là tối thiểu. Nếu gọi  $\lambda$  là xác suất mắc sai lầm loại II, khi đó xác suất bác bỏ giả thiết  $H_0$  sai là  $1 - \lambda$ , nói cách khác,  $1 - \lambda$  là xác suất không mắc phải sai lầm loại II và  $1 - \lambda$  được gọi là lực kiểm định).

Do ta sử dụng phân phối  $t$ , nên thủ tục kiểm định này thường được gọi là kiểm định  $t$ . Theo phương pháp kiểm định ý nghĩa, một thống kê được xem là có ý nghĩa về mặt thống kê nếu giá trị của thống kê kiểm định nằm ở miền bác bỏ, trong trường hợp này, *giả thiết không* bị bác bỏ. Tương tự, một thống kê được xem là không có ý nghĩa về mặt thống kê nếu giá trị của thống kê kiểm định nằm ở miền chấp nhận. Trong tình huống này, *giả thiết không* không bị bác bỏ.

Trong vụ đang xét, vì có 10 quan sát, nên số bậc tự do bằng 8. Với mức ý nghĩa  $\alpha$ , chẳng hạn là 5%, tra bảng ta tìm được giá trị tới hạn  $t_{\alpha/2} = 2,201$ . Vậy miền chấp nhận giả thiết  $H_0: \beta_2 = 0,3$  (với giả thiết đối  $H_1: \beta_2 \neq 0,3$ ) là:  $(-2,201 < t < 2,201)$ . Theo các kết quả đã tính được trong ví dụ 2, ta có:

$$t = \frac{\hat{\beta}_2 - \beta_2^*}{Se(\hat{\beta}_2)} = \frac{0,5091 - 0,3}{0,035742} = 5,85$$

Vì giá trị của  $t$  nằm ở miền , vì vậy ta bác bỏ giả thiết  $H_0$ .

Cần chú ý là, thủ tục kiểm định mô tả ở trên là kiểm định *hai phía* hay *hai đuôi* (vì miền bác bỏ nằm ở hai phía của miền chấp nhận). Kiểm định này được áp dụng khi giả thiết đối có dạng:  $\beta_2 \neq \beta_2^*$ . Nhưng nếu giả thiết đối có dạng:  $\beta_2 > \beta_2^*$  hoặc  $\beta_2 < \beta_2^*$  thì ta sử dụng kiểm định một phía.

Nếu  $H_1: \beta_2 > \beta_2^*$  thì miền bác bỏ nằm về phía bên phải miền chấp nhận.

Nếu  $H_1: \beta_2 < \beta_2^*$  thì miền bác bỏ nằm về phía bên trái miền chấp nhận.

Ta có thể tóm tắt quy tắc quyết định đối với kiểm định giả thiết về  $\beta_2$  như sau:

Loại giả thiết	Giả thiết $H_0$	Giả thiết $H_1$	Miền bác bỏ
Hai phía	$\beta_2 = \beta_2^*$	$\beta_2 \neq \beta_2^*$	$ t  > t_{\alpha/2}$
Phía phải	$\beta_2 \leq \beta_2^*$	$\beta_2 > \beta_2^*$	$t > t_\alpha$
Phía trái	$\beta_2 \geq \beta_2^*$	$\beta_2 < \beta_2^*$	$t < -t_\alpha$

Tương tự ta có quy tắc kiểm định giả thiết đối với  $\beta_1$

Loại giả thiết	Giả thiết $H_0$	Giả thiết $H_1$	Miền bác bỏ
Hai phía	$\beta_1 = \beta_1^*$	$\beta_1 \neq \beta_1^*$	$ t  > t_{\alpha/2}$
Phía phải	$\beta_1 \leq \beta_1^*$	$\beta_1 > \beta_1^*$	$t > t_\alpha$
Phía trái	$\beta_1 \geq \beta_1^*$	$\beta_1 < \beta_1^*$	$t < -t_\alpha$

Nếu ta kiểm định giả thiết  $H_0: \beta_2 = 0$  với giả thiết đối  $H_1: \beta_2 \neq 0$  thì có nghĩa là ta kiểm định giả thiết cho rằng biến X không ảnh hưởng đến biến Y.

Ví dụ: Với số liệu ở ví dụ 2, ta kiểm định giả thiết  $H_0: \beta_2 = 0$  với giả thiết đối  $H_1: \beta_2 \neq 0$  với mức ý nghĩa 5%.

Phân trên ta đã tính được:  $\hat{\beta}_2 = 0.5091$  và  $Se(\hat{\beta}_2) = 0,035742$ ; Vậy:

$$t = \frac{0,5091 - 0}{0,035742} = 14,243$$

Với mức ý nghĩa  $\alpha = 5\%$  và bậc tự do  $n - 2 = 8$  thì  $t_{0,025} = 2,201$ . Vì  $|t| = 14,243 > t_{0,025}$  nên ta bác bỏ giả thiết  $H_0$ . Tức biến thu nhập (X) thực sự có ảnh hưởng tới biến chủ yếu (Y).

### c/ Kiểm định giả thiết về $\delta^2$

Giả sử ta cần kiểm định giả thiết:  $H_0: \delta^2 = \delta_0^2$ ;  $H_1: \delta^2 \neq \delta_0^2$  với mức ý nghĩa  $\alpha$ .

Quy tắc kiểm định giả thiết trên có thể tóm tắt ở bảng sau:

Loại giả thiết	Giả thiết $H_0$	Giả thiết $H_1$	Miền bác bỏ
Hai phía	$\delta^2 = \delta_0^2$	$\delta^2 \neq \delta_0^2$	$\chi^2 > \chi_{\alpha/2}^2$ hoặc $\chi^2 < \chi_{1-\alpha/2}^2$
Phía phải	$\delta^2 \leq \delta_0^2$	$\delta^2 > \delta_0^2$	$\chi^2 > \chi_\alpha^2$
Phía trái	$\delta^2 \geq \delta_0^2$	$\delta^2 < \delta_0^2$	$\chi^2 < \chi_{1-\alpha}^2$

Ví dụ: Với số liệu ở ví dụ 2, ta hãy kiểm định giả thiết  $H_0: \delta^2 = 85$  với  $H_1: \delta^2 \neq 85$  với mức ý nghĩa 5%.

ở phần trên ta đã tính được  $\hat{\sigma}^2 = 42,15875$ ; Vậy:

$$\chi^2 = \frac{(10-2)42,15875}{85} = 3,968$$

Dùng bảng Excel, ta tìm được:

$$\chi^2_{\alpha/2} = \chi^2_{0,025(8)} = \text{CHINV}(0,02; 5; 8) = 17,5345$$

$$\chi^2_{1-\alpha/2} = \chi^2_{0,975(8)} = \text{CHINV}(0,97; 5; 8) = 2,1797$$

Vì  $\chi^2 = 3,968$  không thuộc miền bác bỏ, vì vậy không bác bỏ giả thiết  $H_0$ .

#### d/ Một số chú ý khi kiểm định giả thiết

+ Ý nghĩa của việc “chấp nhận” và “bác bỏ” một giả thiết

Khi áp dụng phương pháp kiểm định ý nghĩa, chẳng hạn kiểm định  $t$ , ta kết luận: chấp nhận *giả thiết không* khi giá trị  $t$  nằm ở miền chấp nhận. điều đó không có nghĩa *giả thiết không* là đúng. Tại sao? để trả lời câu hỏi này ta trở lại ví dụ về tiêu dùng-thu nhập và giả sử  $H_0: \beta_2 = 0,5$ . Khi đó dễ dàng tính được  $t = 0,25$ . Với  $\alpha = 5\%$  ta chấp nhận  $H_0$ . Nhưng bây giờ hãy giả sử  $H_0: \beta_2 = 0,48$ , áp dụng công thức, ta sẽ tính được  $t = 0,82$ , và như vậy theo qui tắc kiểm định, ta cũng kết luận là “chấp nhận  $H_0$ ”. Giả thiết nào đúng trong hai *giả thiết không* này? điều đó ta không biết. Do vậy, khi nói “chấp nhận *giả thiết không*” ta phải luôn nhận thức rằng, còn nhiều *giả thiết không* nữa cũng có thể hoàn toàn phù hợp với số liệu. Trong thực hành, tốt hơn là ta nên kết luận rằng có thể chấp nhận *giả thiết không* hoặc là nói “chưa có cơ sở để bác bỏ *giả thiết không*” chứ không nên nói là chấp nhận nó.

+ Lập *giả thiết không* và *giả thiết đối*

Với các *giả thiết không* và *giả thiết đối* cho trước thì việc kiểm định chúng là dễ dàng. Nhưng làm sao có thể thiết lập được các giả thiết này? Không hề có một qui tắc bất di bất dịch nào. Thường thì tình huống trong nghiên cứu sẽ gợi ý về tính chất của *giả thiết không* và *giả thiết đối*. Ví dụ, xét mô hình hồi qui:  $E_i = \beta_1 + \beta_2\sigma_i$ , trong đó  $E_i$  là suất sinh lợi kỳ vọng của chứng khoán  $i$ ;  $\sigma_i$  là độ lệch chuẩn của suất sinh lợi. Do suất sinh lợi và rủi ro được dự đoán có quan hệ đồng biến, vì vậy, *giả thiết đối* tự nhiên cho *giả thiết không* ( $\beta_2 = 0$ ) sẽ là  $\beta_2 > 0$ . Tức là, ta sẽ không xem xét các giá trị  $\beta_2 < 0$ .

Nhưng khi xem xét trường hợp mức cầu tiền tệ. Một trong các yếu tố ảnh hưởng tới mức cầu tiền tệ là thu nhập. Các nghiên cứu trước đây về hàm cầu tiền tệ chỉ ra rằng độ co giãn của mức cầu tiền tệ đối với thu nhập (tỷ lệ thay đổi % về mức cầu tiền tệ khi thu nhập thay đổi 1%) thường nằm trong khoảng từ 0,7 đến 1,3. Do vậy trong một nghiên cứu mới về mức cầu tiền tệ, nếu ta lập *giả thiết không* là hệ số co giãn của mức cầu tiền tệ đối với thu nhập là bằng 1 (tức là  $H_0: \beta_2 = 1$ ) thì *giả thiết đối* có thể là:  $H_1: \beta_2 \neq 1$ .

Như vậy, có thể dựa vào các kỳ vọng lý thuyết hay nghiên cứu kinh nghiệm trước đây hoặc cả hai để thiết lập các giả thiết. Nhưng mặc dù các giả thiết được lập như thế nào đi nữa thì điều vô cùng quan trọng là nhà nghiên cứu phải thiết lập các giả thiết trước khi điều tra thực nghiệm. Nếu không, nhà nghiên cứu sẽ phạm phải việc lập luận vòng quanh hay cố ước đoán cho phù hợp với kết quả thực nghiệm. Tức là, nếu thiết lập các giả thiết sau khi xem xét các kết quả thực nghiệm, ta có thể muốn thiết lập các giả thiết để biện minh cho kết quả tìm được. Phải tránh cách làm này bằng mọi giá, ít nhất là để tạo sự khách quan trong nghiên cứu.

+ Lựa chọn mức ý nghĩa  $\alpha$

Khi tiến hành kiểm định giả thiết, việc ta bác bỏ hay không bác bỏ *giả thiết không* phụ thuộc nhiều vào  $\alpha$ , mức ý nghĩa hay xác suất phạm phải sai lầm loại I (xác suất bác bỏ giả thiết đúng). Tại sao  $\alpha$  hay được cố định ở mức 1%, 5% hay nhiều nhất là 10%. Trong thực tế, việc ấn định mức ý nghĩa  $\alpha$  không phải là bất khả xâm phạm; mọi giá trị khác cũng có thể được lựa chọn. Nhưng việc lựa chọn giá trị thích hợp của  $\alpha$  sẽ không cần thiết nếu ta sử dụng *giá trị p* của thông kê kiểm định. *Giá trị p* sẽ được đề cập ở mục tiếp theo.

+ Mức ý nghĩa chính xác: *Giá trị p*

Từ số liệu của một mẫu, ta tính được giá trị của tiêu chuẩn kiểm định (ví dụ tổng kê t). Trong ví dụ tiêu dùng-thu nhập nêu trên, khi kiểm định giả thiết  $H_0: \beta_2 = 0$  với  $H_1: \beta_2 \neq 0$ , ta đã tính được  $t = 14,243$ . Khi đó ta có thể tính được:  $P(|t| > 14,243)$ . Xác suất này được gọi là *giá trị p* (giá trị xác suất). Nó cũng được gọi là *mức ý nghĩa quan sát* hay *mức ý nghĩa chính xác* mà *giả thiết không* có thể bị bác bỏ. Các phần mềm kinh tế lượng đều cho *giá trị p* trong bảng kết quả.

Ở trên chúng ta đã biết, nếu số liệu không hỗ trợ giả thiết không, thì  $|t|$  tính được theo *giả thiết không* sẽ “lớn” và như vậy *giá trị p* ứng với  $t$  sẽ “nhỏ”. Nói cách khác, với cỡ mẫu cho trước, khi  $|t|$  tăng lên, *giá trị p* giảm đi, và do vậy ta có thể bác bỏ *giả thiết không* với mức tin cậy càng cao.

*Mối quan hệ giữa giá trị p và mức ý nghĩa  $\alpha$ .*

Nếu ta tạo thói quen cố định  $\alpha$  bằng giá trị  $p$  của thông kê kiểm định (ví dụ tổng kê t), thì không hề có mâu thuẫn giữa hai giá trị. Nói cách khác, ta nên từ bỏ cách cố định  $\alpha$  một cách tùy ý và đơn giản là chọn giá trị  $p$  của thông kê kiểm định. Người nghiên cứu tự quyết định có bác bỏ *giả thiết không* tại giá trị  $p$  tính được hay không?. Nếu trong một ứng dụng, giá trị  $p$  của thông kê kiểm định là 0,145 và nếu người nghiên cứu muốn bác bỏ *giả thiết không* tại mức ý nghĩa (chính xác) này thì cứ việc thực hiện. Không có gì sai nếu chấp nhận xác suất sai lầm nếu bác bỏ giả thiết không khi giả thiết đó đúng 14,5%. Tương tự, nếu trong ví dụ tiêu dùng- thu nhập, nếu ta kiểm định giả thiết  $H_0: \beta_1 = 0$  với  $H_1: \beta_1 \neq 0$  và sử dụng phần mềm Stata, thì giá trị  $p$  tương ứng sẽ là 0,005. không có gì sai nếu nhà nghiên cứu muốn chọn mức ý nghĩa là 5%, tức không muốn xác suất phạm phải sai lầm nhiều hơn 5 trong 1000 lần.

## 2.7 KIỂM ĐỊNH SỰ PHÙ HỢP CỦA HÀM HỒI QUI. PHÂN TÍCH HỒI QUI VÀ PHƯƠNG SAI.

Phần này sẽ trình bày việc phân tích hồi qui theo quan điểm của phân tích phương sai, nó cung cấp cho chúng ta một cách khác, hữu ích trong việc giải quyết vấn đề phán đoán thống kê.

ở phần trên, ta đã biết:  $TSS = ESS + RSS$  và  $r^2 = \frac{ESS}{TSS}$ , nên:

$$ESS = r^2 * TSS \text{ và } RSS = (1 - r^2)TSS.$$

Do  $\beta_2$  có phân phối  $N(\beta_2, \frac{\sigma^2}{\sum x_i^2})$  nên  $\frac{\hat{\beta}_2 - \beta_2}{\sigma} \sqrt{\sum x_i^2} \sim N(0,1)$

$$\text{Và } S_1 = \frac{(\hat{\beta}_2 - \beta_2)^2}{\sigma^2} \sum x_i^2 \sim \chi^2(1); S_2 = \frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^n e_i^2}{\sigma^2} \sim \chi^2(n-2)$$

$$\text{Nên } F = \frac{S_1 / 1}{S_2 / (n-2)} = \frac{(\hat{\beta}_2 - \beta_2)^2 \sum_{i=1}^n x_i^2}{\sum_{i=1}^n e_i^2 / (n-2)} = \frac{(\hat{\beta}_2 - \beta_2)^2 \sum_{i=1}^n x_i^2}{\hat{\sigma}^2} \sim F(1, n-2)$$

Chúng ta kiểm định giả thiết:  $H_0: \beta_2 = 0$  với  $H_1: \beta_2 \neq 0$ .

Để kiểm định giả thiết trên ta áp dụng quy tắc kiểm định sau:

$$\text{Tính } F = \frac{(\hat{\beta}_2)^2 \sum_{i=1}^n x_i^2}{\hat{\sigma}^2}. \text{ Nếu } F > F_\alpha(1, n-2) \text{ thì bác bỏ giả thiết } H_0. \text{ Mặt khác:}$$

$$F = \frac{(\hat{\beta}_2)^2 \sum_{i=1}^n x_i^2}{\hat{\sigma}^2} = \frac{\text{ESS}/1}{\text{RSS}/(n-2)} = \frac{r^2 \text{TSS}/1}{(1-r^2) \text{TSS}/(n-2)} = \frac{r^2(n-2)}{1-r^2}$$

Cho nên quá trình phân tích phương sai cho phép ta đưa ra các phán đoán thống kê về độ thích hợp của hàm hồi qui. Có thể tóm tắt quá trình phân tích phương sai bằng bảng sau:

Nguồn biến thiên	Tổng bình phương	Bậc tự do	Phương sai
Từ hàm hồi qui (ESS)	$\sum_{i=1}^n \hat{y}_i^2 = (\hat{\beta}_2)^2 \sum_{i=1}^n x_i^2$	1	$(\hat{\beta}_2)^2 \sum_{i=1}^n x_i^2$
Từ các yếu tố ngẫu nhiên	$\sum_{i=1}^n e_i^2$	n-2	$\frac{\sum_{i=1}^n e_i^2}{n-2} = \hat{\sigma}^2$
TSS	$\sum_{i=1}^n y_i^2$	n-1	

Với số liệu cho ở ví dụ 2, hãy kiểm định giả thiết:

$$H_0: \beta_2 = 0 \text{ với } H_1: \beta_2 \neq 0.$$

Phân trên ta đã tính được  $r^2 = 0,96206$ , vậy:

$$F = \frac{r^2(n-2)}{1-r^2} = \frac{0,96206(10-2)}{1-0,96206} = 202,86$$

Giá trị p tương ứng với F rất nhỏ ( $< 0,00005$ ) nên ta bác bỏ giả thiết  $H_0$ . Ta có thể kết luận với mức tin cậy cao rằng X(thu nhập) thật sự có tác động tới Y (chi tiêu tiêu dùng).

## 2.8. ỨNG DỤNG PHÂN TÍCH HỒI QUY: VẤN ĐỀ DỰ BÁO:

Trên cơ sở số liệu mẫu ở ví dụ 2, ta có hàm hồi qui mẫu:

$$\hat{Y}_i = 24,4545 + 0,5091X_i$$

Ta có thể dùng hàm hồi qui mẫu này để “dự đoán” hay “dự báo” chi tiêu cho tiêu dùng (Y) trong tương lai ứng với một mức thu nhập (X) cho trước.

Có hai loại dự báo:

- Dự báo trung bình có điều kiện của Y với giá trị  $X = X_0$ .

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

### 1. Dự báo giá trị trung bình:

Giả sử  $X = X_0$ , ta muốn dự báo  $E(Y/X_0) = \beta_1 + \beta_2 X_0$ . Đường hồi qui mẫu cho ta ước lượng điểm của  $E(Y/X_0)$  là:

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$$

Trong đó,  $\hat{Y}_0$  là ước điểm, không chệch và có phương sai nhỏ nhất của  $E(Y/X_0)$ . Tuy nhiên  $\hat{Y}_0$  vẫn sai khác so với giá trị thực của nó.

$\hat{Y}_0$  có phân phối chuẩn với kỳ vọng toán là  $\beta_1 + \beta_2 X_0$  và phương sai là:

$$Var(\hat{Y}_0) = \sigma^2 \left[ \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n x_i^2} \right]; \quad Se(\hat{Y}_0) = \sqrt{var(\hat{Y}_0)} \quad (2.32)$$

Với hệ số tin cậy  $1-\alpha$ , dự báo khoảng của  $E(Y/X_0)$  là:

$$[\hat{Y}_0 - t_{\alpha/2} * se(\hat{Y}_0) < E(Y/X_0) < \hat{Y}_0 + t_{\alpha/2} * se(\hat{Y}_0)] \quad (2.33)$$

### 2. Dự báo giá trị riêng biệt:

Nếu chúng ta muốn dự báo giá trị riêng biệt ( $Y_0$ ) khi  $X = X_0$  với hệ số tin cậy  $1-\alpha$  thì áp dụng công thức:

$$\hat{Y}_0 \pm t_{\alpha/2} * se(Y_0 - \hat{Y}_0) \quad (2.34)$$

Trong đó:  $t_{\alpha/2}$  là giá trị của đại lượng ngẫu nhiên  $T \sim T(n-2)$  thỏa mãn điều kiện:  $P(|T| > t_{\alpha/2}) = \alpha$ .

$$\begin{aligned} \hat{Y}_0 &= \hat{\beta}_1 + \hat{\beta}_2 X_0; \quad se(Y_0 - \hat{Y}_0) = \sqrt{var(Y_0 - \hat{Y}_0)} \\ var(Y_0 - \hat{Y}_0) &= \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n x_i^2} \right] \end{aligned} \quad (2.35)$$

Ví dụ: Với số liệu cho ở ví dụ 2, hãy dự báo giá trị trung bình và giá trị cá biệt của chi tiêu cho tiêu dùng khi thu nhập ở mức 100USD/tuần với hệ số tin cậy 95%?

Giải: Ta có:  $\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0 = 24,4545 + 0,5091 * 100 = 75,3636$

$$Var(\hat{Y}_0) = 42,15875 \left[ \frac{1}{10} + \frac{(100 - 170)^2}{33000} \right] = 10,4758$$

→  $se(\hat{Y}_0) = 3,2366$ ; Với hệ số tin cậy 95% và bậc tự do là 8 thì  $t_{\alpha/2} = t_{0,025} = 2,201$ . Vậy dự báo khoảng của chi tiêu cho tiêu dùng khi thu nhập ở mức 100USD/tuần với hệ số tin cậy 95% là:

$$75,3636 \pm 2,201 * 3,2366 \text{ Hay : } (68,24 < E(Y/X_0 = 100 < 82,487)$$

Để dự báo giá trị riêng biệt, trước hết ta tính:

$$\text{var}(Y_0 - \hat{Y}_0) = 42,15875 \left[ 1 + \frac{1}{10} + \frac{(100 - 170)^2}{33000} \right] = 52,63457$$

→  $\text{se}(Y_0 - \hat{Y}_0) = 7,25497$ . Vậy dự báo khoảng chi tiêu tiêu dùng khi thu nhập ở mức 100USD/tuần với hệ số tin cậy 95% là:

$$75,3636 \pm 2,201 * 7,25497 ; \text{ Hay } (60,781 < Y_0 < 91,332)$$

So sánh kết quả này với kết quả về dự báo khoảng của giá trị trung bình ta thấy khoảng tin cậy của giá trị riêng biệt ( $Y_0$ ) rộng hơn khoảng tin cậy của  $E(y/X_0)$ .

## 2.9 TRÌNH BÀY KẾT QUẢ PHÂN TÍCH HỒI QUI:

Có nhiều cách khác nhau để trình bày các kết quả của phân tích hồi qui, ở đây ta sẽ sử dụng cách trình bày như sau (vận dụng ví dụ tiêu dùng - thu nhập):

$$\begin{aligned} \hat{Y}_i &= 24,4545 + 0,5091 X_i & r^2 &= 0,9621 \\ \text{se} &= (6,4138) \quad (0,0357) & \text{df} &= 8 \\ t &= (3,813) \quad (14,243) & F(1,8) &= 202,87 \\ p &= (0,005) \quad (0,000) & p &= (0,0000) \end{aligned} \quad (2.36)$$

*Cần lưu ý:*

- Các giá trị ghi ở dòng 3 của (2.36) được tính theo công thức:

$$t_j = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \quad (j=1,2)$$

- Giá trị p ghi ở dòng thứ 4 của (2.36) có nghĩa là: với bậc tự do là 8 thì:  $P(|T| > 3,813) = 0,005$ .

$P(|T| > 14,2405) = 0,0005$ ; Vì giá trị 0,0005 rất nhỏ nên có thể làm tròn là 0,000. Như vậy khi máy báo kết quả giá trị p bằng 0,000 thì ta cần hiểu giá trị này là một con số rất nhỏ (nhỏ hơn 0,0005) và đã được làm tròn là 0,000. Tương tự:  $P(F > 202,87) < 0,00005$  và xác suất này được làm tròn là 0,0000.

- Nhìn vào giá trị p ta có thể kết luận chấp nhận hay bác bỏ giả thiết  $H_0: \beta_j = 0; H_1: \beta_j \neq 0 (j=1,2)$ . Theo kết quả trên, với mức ý nghĩa khá nhỏ (chẳng hạn 1%) thì ta bác bỏ giả thiết:

$$H_0: \beta_1 = 0; H_1: \beta_1 \neq 0 \text{ và } H_0^*: \beta_2 = 0; H_1^*: \beta_2 \neq 0$$

(vì giá trị p đều nhỏ hơn 0,01, tức giá trị của thống kê kiểm định đều nằm ở miền bác bỏ).

- Sau khi ước lượng được mô hình hồi qui tổng thể và tính được các thông số hồi qui, ta cần đánh giá về sự thích hợp của mô hình. Mô hình phù hợp tới đâu? Để trả lời câu hỏi này, ta cần một số tiêu chí:

*Thứ nhất:* Dấu của các hệ số hồi qui ước lượng có phù hợp với lý thuyết hay tiên nghiệm không? Một sự tiên nghiệm là  $\beta_2$ , xu hướng tiêu dùng biên (MPC) trong hàm tiêu dùng phải dương. Trong ví dụ 2,  $\hat{\beta}_2 = 0,5091$  (là một ước lượng điểm của  $\beta_2$ ) là số dương.

*Thứ hai:* Theo lý thuyết kinh tế thì mối quan hệ giữa chi tiêu và thu nhập không những chỉ đồng biến mà còn phải có ý nghĩa thống kê thì trong ví dụ đang xét có thoả mãn không? Như trên

ta đã tiến hành kiểm định,  $\beta_2$  không những dương mà còn khác 0 đáng kể về mặt thống kê. Lập luận cũng đúng cho tung độ gốc.

*Thứ ba:* Mô hình giải thích biến thiên trong chỉ tiêu tiêu dùng tốt đến đâu? Ta có thể dùng  $r^2$  để trả lời câu hỏi này, trong ví dụ  $r^2 = 0,962$ . tức là rất gần 1, như vậy mức độ phù hợp của mô hình khá tốt.

Ngoài ra, ta cần kiểm tra xem mô hình có thoả mãn các giả thiết của mô hình hồi qui tuyến tính cổ điển. Vấn đề này sẽ được đề cập ở các chương sau.

## TÓM TẮT NỘI DUNG CHƯƠNG 2

Mô hình hồi qui hai biến là mô hình đơn giản nhất trong lớp mô hình hồi qui tuyến tính, trong mô hình chỉ có một biến phụ thuộc và một biến độc lập, giá trị của biến độc lập cho trước. Để xác định giá trị trung bình của biến phụ thuộc; Trước hết ta căn cứ vào một mẫu có sẵn, sử dụng phương pháp OLS để ước lượng các tham số của mẫu; Xác định hàm hồi qui mẫu, kiểm tra tính hợp lý của các tham số (các hệ số trong hàm SRF). Nếu phù hợp ta tiến hành xác định phương sai và sai số chuẩn đối với các ước lượng trên, từ đó xác định hệ số tương quan và hệ số xác định để kiểm tra mức độ chặt chẽ của quan hệ giữa hai biến và mức độ phù hợp của hàm SRF.

Trên cơ sở hàm hồi qui mẫu, các giả thiết của phương pháp OLS và giả thiết về phân phối xác suất của các ước lượng, tiến hành xác định khoảng tin cậy của các tham số trong hàm PRF; Kiểm định giả thiết về các hệ số hồi qui; Kiểm định sự phù hợp của hàm hồi qui. Phân tích hồi qui và phân tích phương sai. Sau khi đã có kết luận về mức độ tin cậy của hàm hồi qui, ta tiến hành dự báo giá trị của biến phụ thuộc khi biết giá trị của biến độc lập trong tương lai (kế hoạch). Cuối cùng ta trình bày kết quả và đánh giá các kết quả của phân tích hồi qui.

## CÂU HỎI VÀ BÀI TẬP ÔN CHƯƠNG 2

### I- Câu hỏi

1. Trình bày nội dung của phương pháp bình phương nhỏ nhất? Trong mô hình hồi qui hai biến các hệ số hồi qui có ý nghĩa như thế nào?
2. Tại sao cần phải đưa ra các giả thiết đối với phương pháp bình phương nhỏ nhất? Nêu các giả thiết và ý nghĩa của từng giả thiết?
3. Cách xác định phương sai và sai số chuẩn đối với các ước lượng của hàm hồi qui mẫu?
4. ý nghĩa của hệ số xác định và hệ số tương quan?
5. Cách xác định khoảng tin cậy của  $\beta_1$ ,  $\beta_2$  và  $\delta^2$ ?
6. Phương pháp kiểm định giả thiết về các hệ số hồi qui và phương sai của hàm hồi qui tổng thể?
7. Nêu ý nghĩa của *giá trị p* trong kiểm định giả thiết về các hệ số hồi qui?
8. Trình bày phương pháp kiểm định sự phù hợp của hàm hồi qui. Phân tích hồi qui và phân tích phương sai?
9. Các loại dự báo biến phụ thuộc khi biết dạng hàm hồi qui mẫu và giá trị của biến độc lập?
10. Cách trình bày kết quả phân tích hồi qui? Các chỉ tiêu cần đánh giá đối với kết quả của phân tích hồi qui?

## II- Bài tập:

1. Bảng dưới đây cho các cặp biến phụ thuộc và độc lập. Trong mỗi trường hợp hãy cho biết quan hệ giữa hai biến là: cùng chiều, ngược chiều hay không xác định? Hãy giả thích?

<u>Biến phụ thuộc</u>	<u>Biến độc lập</u>
a- Vốn đầu tư	Lãi suất
b- Tiết kiệm cá nhân	Lãi suất
c- Cầu về tiền	GDP
d- Sản lượng	Vốn cơ bản (hoặc lao động)
e- Lượng cầu về xe máy	Giá xăng
f- Lượng điện tiêu thụ của hộ gia đình	Giá ga

2. Quan sát về thu nhập (X-USD/tuần) và chi tiêu (Y-USD/tuần) của 10 người, ta thu được số liệu sau:

$X_i$	31	50	47	45	39	50	35	40	45	50
$Y_i$	29	42	38	30	29	41	23	36	42	48

- Ước lượng hàm hồi qui tuyến tính:  $Y_i = \beta_1 + \beta_2 X_i + U_i$ .
- Nêu ý nghĩa kinh tế của các hệ số hồi qui đã ước lượng được. Các giá trị có phù hợp với lý thuyết kinh tế hay không?
- Tìm khoảng tin cậy của  $\beta_1$ ,  $\beta_2$  với độ tin cậy 95%?
- Kiểm định giả thiết  $H_0: \beta_2 = 0$ ;  $H_1: \beta_2 \neq 0$  với mức ý nghĩa 5%?
- Tính  $r^2$  và đánh giá mức độ phù hợp của mô hình?
- Dự báo chi tiêu của một người có mức thu nhập 40USD/tuần?



Ký hiệu:  $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}; \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}; \quad U = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix}$

$$X = \begin{bmatrix} 1 & X_{21} & X_{31} & \dots & X_{k1} \\ 1 & X_{22} & X_{32} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2n} & X_{3n} & \dots & X_{kn} \end{bmatrix}$$

Khi đó ta có: (3.1) có thể viết dưới dạng ma trận như sau:

$$Y = X\beta + U.$$

### 3.2 CÁC GIẢ THIẾT:

Phần này sẽ nhắc lại các giả thiết mô hình hồi qui tuyến tính cổ điển được trình bày bằng cách phát biểu bình thường và bằng ngôn ngữ ma trận.

Ký hiệu  $E(U) = E \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{bmatrix} = \begin{bmatrix} E(U_1) \\ E(U_2) \\ \vdots \\ E(U_n) \end{bmatrix}; \quad UU^T = \begin{bmatrix} U_1^2 & U_1U_2 & U_1U_3 & \dots & U_1U_n \\ U_2U_1 & U_2^2 & U_2U_3 & \dots & U_2U_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ U_nU_1 & U_nU_2 & U_nU_3 & \dots & U_n^2 \end{bmatrix}$

Giả thiết 1:  $E(U_i) = 0 \quad \forall i$  hay  $E(U) = 0$

Giả thiết 2:  $E(U_i, U_j) = \begin{cases} 0 & \forall i \neq j \\ \sigma^2 & i = j \end{cases}$

Hay:  $E(UU^T) = \sigma^2 I$  ( $I$  là ma trận đơn vị cấp  $n$ )

Giả thiết 3:  $X_2, X_3, \dots, X_k$  đã được xác định hay ma trận  $X$  đã được xác định.

Giả thiết 4: Không có hiện tượng đa cộng tuyến giữa các biến giải thích hay hạng của ma trận

$X$  bằng  $k$ :  $R(X) = k$ .

Giả thiết 5:  $U_i \sim N(0, \sigma^2) \quad \forall i$  hay  $U \sim N(0, \sigma^2 I)$ .

(Dùng để kiểm định giả thiết và tìm khoảng tin cậy).

Giả thiết 2 có thể viết đầy đủ như sau:

$$\begin{aligned}
 E(UU^T) &= E \begin{bmatrix} U_1^2 & U_1U_2 & U_1U_3 & \dots & U_1U_n \\ U_2U_1 & U_2^2 & U_2U_3 & \dots & U_2U_n \\ \dots & \dots & \dots & \dots & \dots \\ U_nU_1 & U_nU_2 & U_nU_3 & \dots & U_n^2 \end{bmatrix} = \\
 &= \begin{bmatrix} E(U_1^2) & E(U_1U_2) & E(U_1U_3) & \dots & E(U_1U_n) \\ E(U_2U_1) & E(U_2^2) & E(U_2U_3) & \dots & E(U_2U_n) \\ \dots & \dots & \dots & \dots & \dots \\ E(U_nU_1) & E(U_nU_2) & E(U_nU_3) & \dots & E(U_n^2) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix} \\
 E(UU^T) &= \sigma^2 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & 1 & \dots & 0 \\ 0 & \dots & \dots & \dots & 1 \end{bmatrix} = \sigma^2 I \quad \text{Với } I \text{ là ma trận đơn vị cấp } n. \quad (3.2)
 \end{aligned}$$

Ma trận (3.2) gọi là ma trận hiệp phương sai của  $U_i$ . Các phần tử trên đường chéo chính là hiệp phương sai của  $U_i$ , các phần tử ngoài đường chéo chính là hiệp phương sai.

Giả thiết 4 nói rằng hạng của ma trận  $X$  bằng số cột của ma trận này, nghĩa là các cột của ma trận  $X$  là độc lập tuyến tính. Hay nói cách khác đi không có hiện tượng cộng tuyến giữa các biến độc lập, về mặt toán học có nghĩa không tồn tại các số  $\lambda_2, \dots, \lambda_k$  trong đó có ít nhất một  $\lambda_i \neq 0$  để:  $\lambda_2 X_{2i} + \lambda_3 X_{3i} + \dots + \lambda_k X_{ki} = 0 \quad \forall i = \overline{1, n}$

### 3.3 ƯỚC LƯỢNG CÁC THAM SỐ - OLS.

Hàm hồi quy mẫu SRF có dạng:

$$\begin{aligned}
 \hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_k X_{ki} \\
 Y_i &= \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_k X_{ki} + e_i
 \end{aligned}$$

Hay:  $Y = X\hat{\beta} + e$

Trong đó:  $e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = Y - X\hat{\beta}$

Các ước lượng OLS được tìm bằng cách:

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \dots - \hat{\beta}_k X_{ki})^2 \Rightarrow \text{Min}$$

$\sum_{i=1}^n e_i^2$  là tổng bình phương các phần dư (RSS).

Ta ký hiệu  $X^T Y^T, \hat{\beta}^T, e^T$  tương ứng là các ma trận chuyển vị của  $X, Y, \hat{\beta}, e$ . Tức là:

$$X^T = \begin{bmatrix} 1 & \dots & 1 & \dots & 1 \\ X_{21} & & X_{22} & & X_{2n} \\ \vdots & & \vdots & & \vdots \\ X_{k1} & & X_{k2} & \dots & X_{kn} \end{bmatrix} \quad Y^T = (Y_1, Y_2, \dots, Y_n).$$

$$\hat{\beta}^T = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k) \quad e^T = (e_1, e_2, \dots, e_n). \text{ Khi đó:}$$

$$\begin{aligned} e^T e &= \sum_{i=1}^n e_i^2 = (Y - X\hat{\beta})^T (Y - X\hat{\beta}) = (Y^T - \hat{\beta}^T X^T)(Y - X\hat{\beta}) = \\ &= Y^T Y - \hat{\beta}^T X^T Y - Y^T X \hat{\beta} + \hat{\beta}^T X^T X \hat{\beta} = Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T X \hat{\beta} \\ &\quad (\text{Vì } \hat{\beta}^T X^T Y = Y^T X \hat{\beta}) \end{aligned}$$

Hệ phương trình chuẩn có dạng:

$$\frac{\partial (e^T e)}{\partial \hat{\beta}} = 0 \Rightarrow -2X^T Y + 2X^T X \hat{\beta} \Rightarrow X^T Y = X^T X \hat{\beta} \quad \text{Vậy } \hat{\beta} = (X^T X)^{-1} X^T Y$$

Trong đó ma trận  $(X^T X)$  có dạng như sau:

$$X^T X = \begin{bmatrix} n & \sum X_{2i} & \sum X_{3i} & \dots & \sum X_{ki} \\ \sum X_{2i} & \sum X_{2i}^2 & \sum X_{2i} X_{3i} & \dots & \sum X_{2i} X_{ki} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum X_{ki} & \sum X_{ki} X_{2i} & \sum X_{ki} X_{3i} & \dots & \sum X_{ki}^2 \end{bmatrix}$$

**Thí dụ 3.1:** Có số liệu quan sát của một mẫu cho ở bảng số liệu dưới đây (bảng 3.1). Trong đó:

$Y$ : là lượng hàng bán được của một loại hàng (tấn/tháng).

$X_2$ : là thu nhập của người tiêu dùng (triệu đồng/năm)

$X_3$ : là giá bán của loại hàng này (ngàn đồng/kg)

$Y_i$	20	18	19	18	17	17	16	15	13	12
$X_{2i}$	8	7	8	8	6	6	5	5	4	3
$X_{3i}$	2	3	4	4	5	5	6	7	8	8

Tìm hàm hồi quy:  $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$

Giải: Từ bảng số liệu đó cho, ta tính được các tổng:

$$\sum Y_i = 165; \quad \sum X_{2i} = 60; \quad \sum X_{3i} = 52; \quad \sum Y_i^2 = 2781; \quad \sum X_{2i}^2 = 388; \quad \sum X_{3i}^2 = 308;$$

$$\sum X_{2i} X_{3i} = 282; \quad \sum Y_i X_{2i} = 1029; \quad \sum Y_i X_{3i} = 813;$$

Ma trận nghịch đảo:

$$(X^T X)^{-1} = \begin{bmatrix} 10 & 60 & 52 \\ 60 & 388 & 282 \\ 52 & 282 & 308 \end{bmatrix}^{-1} = \frac{1}{1528} \begin{bmatrix} 39980 & -3816 & -3256 \\ -3816 & 376 & 300 \\ -3256 & 300 & 280 \end{bmatrix}$$

$$\hat{\beta} = \frac{1}{1528} \begin{bmatrix} 39980 & -3816 & -3256 \\ -3816 & 376 & 300 \\ -3256 & 300 & 280 \end{bmatrix} \begin{bmatrix} 165 \\ 1029 \\ 813 \end{bmatrix} = \begin{bmatrix} 22908 / 1528 \\ 1164 / 1528 \\ -900 / 1528 \end{bmatrix}$$

Hay  $\hat{\beta} = \begin{bmatrix} 14,99215 \\ 0,76178 \\ -0,58901 \end{bmatrix}$

Vậy hàm hồi qui cần tìm là:

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

### 3.4. MA TRẬN HIỆP PHƯƠNG SAI CỦA CÁC ƯỚC LƯỢNG:

Để kiểm tra giả thiết, tìm khoảng tin cậy, cũng như thực hiện các suy luận thống kê khác cần phải tìm  $\text{Var}(\hat{\beta}_i)$ ;  $i = \overline{1, k}$  và  $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j)$ . Phương pháp ma trận cho phép chúng ta tìm chúng một cách dễ dàng.

Ma trận hiệp phương sai của  $\hat{\beta}$

$$\text{Cov}(\hat{\beta}) = \begin{bmatrix} \text{Var}(\hat{\beta}_1) & \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) & \dots & \text{Cov}(\hat{\beta}_1, \hat{\beta}_k) \\ \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) & \text{Var}(\hat{\beta}_2) & \dots & \text{Cov}(\hat{\beta}_2, \hat{\beta}_k) \\ \dots & \dots & \dots & \dots \\ \text{Cov}(\hat{\beta}_k, \hat{\beta}_1) & \text{Cov}(\hat{\beta}_k, \hat{\beta}_2) & \dots & \text{Var}(\hat{\beta}_k) \end{bmatrix}$$

Để tìm  $\text{Cov}(\hat{\beta})$  ta áp dụng công thức:

$$\hat{\beta} = \delta^2 (X^T X)^{-1}$$

Trong công thức trên  $(X^T X)^{-1}$  là ma trận nghịch đảo của ma trận  $(X^T X)$ ,  $\delta^2$  là  $\text{Var}(U_i)$ , nhưng chưa biết nên ta phải dùng ước lượng không chệch của  $\delta^2$  là:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{(n-k)} \text{ trong đó } k \text{ là số biến của mô hình.}$$

Thí dụ: Với số liệu cho ở thí dụ 3.1 hãy tìm ma trận hiệp phương sai của  $\hat{\beta}$

Giải: ở phần trên ta đã tính được ma trận  $(X^T X)^{-1}$ , ta cần tính  $\hat{\sigma}^2$ .

Ta có:

$$\begin{aligned} TSS &= Y^T Y - n(\bar{Y})^2 = \sum Y_i^2 - n(\bar{Y})^2 = 2781 - 10(16,5)^2 = 58,5 \\ ESS &= \hat{\beta}^T (X^T Y) - n(\bar{Y})^2 = \\ &= (14,99215 \times 0,76178 - 0,58901) \begin{bmatrix} 165 \\ 1029 \\ 813 \end{bmatrix} - 10(16,5)^2 = 56,211 \\ \rightarrow RSS &= 58,5 - 56,211 = 2,289 \\ \hat{\sigma}^2 &= \frac{RSS}{(n-k)} = \frac{2,289}{10-3} = 0,327 \\ Cov(\hat{\beta}) &= \frac{0,327}{1528} \begin{bmatrix} 39980 & -3816 & -3256 \\ -3816 & 376 & 300 \\ -3256 & 300 & 280 \end{bmatrix} = \begin{bmatrix} 8,55593 & -0,81664 & -0,6968 \\ -0,81664 & 0,080466 & 0,0642 \\ -0,6968 & 0,0642 & 0,05992 \end{bmatrix} \end{aligned}$$

### 3.5 CÁC TÍNH CHẤT CỦA CÁC ƯỚC LƯỢNG BÌNH PHƯƠNG NHỎ NHẤT:

Véc tơ  $\hat{\beta}$  thu được bằng phương pháp OLS với 5 giả thiết đã nêu ở trên có các tính chất:

- Tuyến tính.
- Không chệch
- Có phương sai nhỏ nhất.

Thật vậy:  $\hat{\beta} = (X^T X)^{-1} X^T Y$ . do  $(X^T X)^{-1}$  là ma trận các số cố định, cho nên  $\hat{\beta}$  là hàm của Y- hàm tuyến tính.

Hàm hồi quy tổng thể có dạng:  $Y = X\beta + U$

Cho nên  $\hat{\beta} = (X^T X)^{-1} X^T (X\beta + U) = \beta + (X^T X)^{-1} X^T U$ .

$$E(\hat{\beta}) = E(\beta) + (X^T X)^{-1} X^T E(U) = \beta + 0.$$

Điều này chứng tỏ  $\hat{\beta}$  là ước lượng không chệch của  $\beta$ . Bây giờ ta giả sử rằng  $\beta^*$  là ước lượng tuyến tính không chệch bất kỳ khác của  $\beta$ ,  $\beta^*$  có thể viết dưới dạng:

$$\beta^* = [(X^T X)^{-1} X^T + C]Y = [(X^T X)^{-1} X^T + C](X\beta + U) = \beta + CX\beta + (X^T X)^{-1} X^T U + CU.$$

Trong đó C là một ma trận.

Do  $\beta^*$  là ước lượng không chệch của  $\beta$  nên  $CX = 0$

$$\rightarrow \beta^* - \beta = (X^T X)^{-1} X^T U + CU.$$

$$Cov(\beta^*) = E[(\beta^* - \beta)(\beta^* - \beta)^T] = E[(X^T X)^{-1} X^T U + CU] [(X^T X)^{-1} X^T U + CU]^T.$$

$$= (X^T X)^{-1} E(UU^T) + CC^T E(UU^T) = \delta^2 (X^T X)^{-1} + \delta^2 CC^T = \text{Cov}(\hat{\beta}) + \delta^2 CC^T.$$

Do  $CC^T$  là ma trận không âm cho nên  $\text{Var}(\beta_i^*) \geq \text{Var}(\hat{\beta}_i)$  do tính chất bất kỳ của  $\beta^*$ , điều này chứng tỏ  $\hat{\beta}$  có phương sai nhỏ nhất.

### 3.6 ƯỚC LƯỢNG HỢP LÝ TỐI ĐA (ML):

Với các giả thiết đã trình bày ở trên thì  $Y_i \sim N(\beta_1 + \beta_2 X_{2i} + \dots + \beta_k X_{ki}, \delta^2)$  và các ước lượng của  $\beta_1, \beta_2, \dots, \beta_k$  là  $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$  thu được bằng OLS và ML là như nhau. Nhưng ước

lượng của  $\delta^2$  từ ML:  $\hat{\delta}^2 = \frac{\sum_{i=1}^n e_i^2}{n-k}$  là ước lượng chệch.

### 3.7 HỆ SỐ XÁC ĐỊNH BỘI VÀ HỆ SỐ XÁC ĐỊNH BỘI ĐÃ ĐIỀU CHỈNH:

Trong mô hình hồi quy hai biến,  $r^2$  đo độ thích hợp của hàm hồi quy. Nó chính là tỷ lệ của toàn bộ sự biến động của biến phụ thuộc Y do biến giải thích X gây ra. Trong mô hình hồi quy bội tỷ lệ của toàn bộ sự khác biệt của biến Y do tất cả các biến giải thích  $X_2, X_3, \dots, X_k$  gây ra được gọi là hệ số xác định bội, ký hiệu  $R^2$ .

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

$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n y_i^2}$$

$0 \leq R^2 \leq 1$ . Nếu  $R^2 = 1$ , có nghĩa là đường hồi quy giải thích 100% sự thay đổi của Y. Nếu  $R^2 = 0$ , có nghĩa là mô hình không giải thích sự thay đổi nào của Y.

$R^2$  là hàm không giảm của biến số giải thích có trong mô hình. Dễ dàng thấy rằng  $\sum_{i=1}^n y_i^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2$  không phụ thuộc vào số biến giải thích có trong mô hình, nhưng  $\sum_{i=1}^n e_i^2$  là hàm giảm của số này. Do đó, nếu tăng số biến giải thích trong mô hình thì  $R^2$  cũng tăng. Vấn đề đặt ra là khi nào thì đưa thêm biến giải thích vào mô hình?

Không thể dùng  $R^2$  làm tiêu chuẩn để xem xét việc đưa thêm hay không đưa thêm một biến giải thích mới vào mô hình. Bởi vì  $R^2$  còn phụ thuộc vào số bậc tự do của  $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$  và  $\sum_{i=1}^n (Y_i - \bar{Y})^2$  tương ứng là  $(n-k)$  và  $(n-1)$ . Trong đó k là số các tham số (kể cả hệ số chặn) của mô hình.

Người ta dùng hệ số xác định bội đã điều chỉnh, ký hiệu là  $\bar{R}^2$  để cân nhắc khi xem xét việc thêm biến giải thích mới vào mô hình. Nó được xác định bằng công thức:

$$\overline{R^2} = 1 - \frac{\sum_{i=1}^n e_i^2 / (n-k)}{\sum_{i=1}^n y_i^2 / (n-1)} = 1 - (1 - R^2) \frac{n-1}{n-k}$$

$\overline{R^2}$  có các tính chất sau:

• Nếu  $k > 1$ ,  $\overline{R^2} \leq R^2 \leq 1$ , điều này có nghĩa là nếu số biến giải thích tăng lên thì  $\overline{R^2}$  tăng chậm hơn so với  $R^2$ .

•  $R^2 \geq 0$ , nhưng  $\overline{R^2}$  có thể âm. Như vậy khi  $\overline{R^2}$  còn tăng thì ta còn phải đưa thêm biến mới vào mô hình.  $\overline{R^2}$  còn có thể tăng khi mà hệ số của biến mới trong hàm hồi quy khác không. Khi nào biết được hệ số của biến mới trong hàm hồi quy khác không? Khi mà giả thiết:

$$H_0: \beta_k = 0$$

$$H_1: \beta_k \neq 0$$

Nếu giả thiết  $H_0$  bị bác bỏ thì biến  $X_k$  sẽ được đưa vào mô hình. Trong thực tế, ta nên dùng  $R^2$  hay  $\overline{R^2}$ ? Dùng  $\overline{R^2}$  tốt hơn  $R^2$  bởi vì  $R^2$  có khuynh hướng cho ra một bức tranh quá lạc quan về độ thích hợp của hồi qui, đặc biệt là khi số lượng các biến giải thích không quá nhỏ so với số lượng các lần quan sát.

### 3.8 MA TRẬN TƯƠNG QUAN:

Giả sử ta có mô hình hồi qui bội:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + U_i$$

Ký hiệu  $r_{ij}$  là hệ số tương quan giữa biến thứ  $i$  và thứ  $j$ . Nếu  $i=j$  thì  $r_{ij}$  là hệ số tương quan giữa các biến  $Y$  và biến  $X_j$ .

$$r_{ij} = \frac{\sum_{i=1}^n (y_i x_{ij})^2}{\sum_{i=1}^n y_i^2 \sum_{i=1}^n x_{ji}^2}; \quad r_{ij}^2 = \frac{\left( \sum_{i=1}^n x_{ti} x_{ji} \right)^2}{\sum_{i=1}^n x_{ti}^2 \sum_{i=1}^n x_{ji}^2}$$

Trong đó:  $x_{ji} = X_{ji} - \overline{X_j}$

Dễ dàng nhận thấy rằng:  $r_{ii} = r_{ji}$ ;  $r_{jj} = 1$ .

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} & \dots & r_{1k} \\ r_{21} & r_{22} & r_{23} & \dots & r_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ r_{k1} & r_{k2} & r_{k3} & \dots & r_{kk} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & r_{13} & \dots & r_{1k} \\ r_{21} & 1 & r_{23} & \dots & r_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ r_{k1} & r_{k2} & r_{k3} & \dots & 1 \end{bmatrix}$$

### 3.9 HỆ SỐ TƯƠNG QUAN RIÊNG PHẦN:

Ta đã biết hệ số tương quan  $r$  đo mức độ phụ thuộc tuyến tính giữa hai biến. đối với mô hình hồi quy 3 biến:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + U_i$$

Ta định nghĩa:

$r_{12,3}$  là hệ số tương quan giữa biến  $Y$  và  $X_2$  trong khi  $X_3$  không đổi.

$r_{13,2}$  là hệ số tương quan riêng giữa biến  $Y$  và  $X_3$  trong khi  $X_2$  không đổi.

$r_{23,1}$  là hệ số tương quan riêng giữa  $X_2$  và  $x_3$  trong khi  $Y$  không đổi.

Ta có thể chứng minh các công thức sau:

$$r_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}};$$

$$r_{13,2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}};$$

$$r_{23,1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}}$$

Hệ số tương quan riêng đã được định nghĩa như trên được gọi là hệ số tương quan bậc nhất. Từ “bậc” ở đây ngụ ý chỉ số hạng sau dấu phẩy vì thế  $r_{12,34}$  là hệ số tương quan riêng bậc 2; còn  $r_{12}, r_{13}$  là các hệ số tương quan bậc không.

Giữa hệ số xác định bội và các hệ số tương quan bậc không và hệ số tương quan bậc nhất có các mối liên hệ sau:

$$R^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2};$$

$$R^2 = r_{12}^2 + (1 - r_{12}^2)r_{13,2}^2;$$

$$R^2 = r_{13}^2 + (1 - r_{13}^2)r_{12,3}^2$$

Ma trận  $R$  nói ở trên được gọi là ma trận hệ số tương quan riêng cấp 0

### 3.10 KIỂM ĐỊNH GIẢ THIẾT VÀ KHOẢNG TIN CẬY CỦA CÁC HỆ SỐ HỒI QUY RIÊNG- KIỂM ĐỊNH T.

Với giả thiết  $U \sim N(0, \delta^2)$  ta có thể kiểm định giả thiết, tìm khoảng tin cậy cho các hệ số hồi quy riêng.

$$\hat{\beta} \sim N(\beta, \delta^2(X^T X)^{-1})$$

Thành phần  $\hat{\beta}_i$  có phân phối chuẩn với kỳ vọng  $\beta_i$  và phương sai bằng  $\delta^2$  nhân với phần tử nằm trên dòng thứ  $i$  và cột  $i$  của ma trận  $(X^T X)^{-1}$  hay chính là phần tử thứ  $i$  trên đường chéo chính của ma trận  $\text{Cov}(\hat{\beta})$ . Tuy nhiên do  $\delta^2$  chưa biết, nên ta phải dùng ước lượng không chệch của  $\delta^2$  là:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-k}$$

Khi đó  $t = \frac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)}$  có phân bố  $t(n-k)$ . Với tiêu chuẩn này có thể tìm khoảng tin cậy, kiểm định giả thiết về các hệ số hồi quy riêng.

Khoảng tin cậy với hệ số tin cậy  $1-\alpha$  của  $\beta_i$  được xác định:

$$P\left(-t_{\alpha/2}(n-k) < \frac{\hat{\beta}_i - \beta_i}{Se(\hat{\beta}_i)} < t_{\alpha/2}(n-k) Se(\hat{\beta}_i)\right)$$

Do đó:

$$(\hat{\beta}_i - t_{\alpha/2}(n-k)Se(\hat{\beta}_i) < \beta_i < \hat{\beta}_i + t_{\alpha/2}(n-k)Se(\hat{\beta}_i)); \forall i = \overline{1, k}$$

Chúng ta có thể kiểm định giả thiết:  $\beta_i = \beta_i^*$

Tiêu chuẩn dùng để kiểm định:  $t = \frac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)} \sim t(n-k)$

Tuỳ theo giả thiết  $H_1$ , chúng ta có các miền bác bỏ sau đây:

Loại giả thiết	$H_0$	$H_1$	Miền bác bỏ
Hai phía	$\beta_i = \beta_i^*$	$\beta_i \neq \beta_i^*$	$ t  > t_{\alpha/2}(n-k)$
Bên trái	$\beta_i \geq \beta_i^*$	$\beta_i < \beta_i^*$	$t < -t_{\alpha/2}(n-k)$
Bên phải	$\beta_i \leq \beta_i^*$	$\beta_i > \beta_i^*$	$t > t_{\alpha/2}(n-k)$

Nếu  $\beta_i^* = 0$ , chúng ta muốn kiểm định biến độc lập  $X_i$ , không ảnh hưởng đến biến phụ thuộc.

**Kiểm định giả thiết**  $H_0: \beta_2 = \beta_3 = \dots = \beta_k$  Hay  $R^2 = 0$

$H_1$ : Không phải tất cả các hệ số hồi quy riêng đồng thời bằng 0.

Đại lượng ngẫu nhiên  $F = \frac{ESS/(k-1)}{RSS/(n-k)}$  tuân theo phân phối  $F$  với  $k-1$  và  $n-k$  bậc tự do.

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} = \frac{(n-k)ESS}{(k-1)RSS} = \frac{(n-k)}{k-1} \frac{ESS}{TSS - ESS} = \frac{(n-k)}{(k-1)} \frac{ESS/TSS}{1 - (ESS/TSS)} =$$

Ta có:

$$= \frac{(n-k)}{(k-1)} \frac{R^2}{1 - R^2} = \frac{R^2/(k-1)}{(1 - R^2)/(n-k)}$$

Như vậy giữa  $F$  và  $R^2$  có liên quan với nhau.  $R^2$  càng lớn thì  $F$  càng lớn. Khi  $R^2 = 0$  thì  $F = 0$ . Khi  $R^2 = 1$  thì  $F$  vô hạn. Do vậy việc kiểm định giả thiết  $H_0: \beta_2 = \beta_3 = \dots = \beta_k = 0$  cũng là kiểm định giả thiết cho rằng  $R^2$  (trong tổng thể) bằng 0.

Để kiểm định giả thiết trên, ta áp dụng qui tắc kiểm định sau đây:

- Tính F theo công thức:  $F = \frac{R^2 / (k-1)}{(1-R^2) / (n-k)}$

- Với mức ý nghĩa  $\alpha$ , tra bảng phân phối Fisher-Snedecor với bậc tự do  $n_1 = (k-1)$  và  $n_2 = (n-k)$  để tìm giá trị  $F_{\alpha}(k-1; n-k)$ . Trong đó n là số quan sát, k là số biến trong mô hình hồi quy (kể cả biến phụ thuộc).  $F_{\alpha}(k-1; n-k)$  là giá trị thỏa mãn điều kiện:

$$P[F > F_{\alpha}(k-1, n-k)] = \alpha$$

- Nếu  $F > F_{\alpha}(k-1; n-k)$  thì ta bác bỏ giả thiết  $H_0$ , tức là các hệ số hồi quy không đồng thời bằng 0 (hay  $R^2 \neq 0$  có ý nghĩa). Ngược lại, nếu  $F < F_{\alpha}(k-1; n-k)$  thì ta không bác bỏ giả thiết  $H_0$ , tức là các hệ số hồi quy đồng thời bằng 0 (hay  $R^2 = 0$  có ý nghĩa).

### 3.11 HỒI QUY CÓ ĐIỀU KIỆN RÀNG BUỘC - KIỂM ĐỊNH F:

Giả sử chúng ta có hàm hồi quy:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + U_i \quad (*)$$

Bây giờ ta kiểm định giả thiết:

$$H_0: \beta_{k-m+1} = \beta_{k-m+2} = \dots = \beta_k = 0$$

Với giả thiết này hàm hồi quy có dạng:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_{k-m} X_{k-m,i} + U_i$$

(được gọi là hàm hồi quy thu hẹp hay hàm hồi quy có điều kiện ràng buộc).

Ký hiệu:  $e_R$ : Véc tơ phần dư từ hàm hồi quy có điều kiện ràng buộc.

$e_{UR}$ : Véc tơ phần dư từ hàm hồi quy ban đầu (không điều kiện ràng buộc).

m: Số bị loại khỏi mô hình ban đầu (Số điều kiện ràng buộc).

n: Số quan sát.

Khi đó tiêu chuẩn kiểm định giả thiết  $H_0$ :

$$\frac{(e_R^T e_R - e_{UR}^T e_{UR}) / m}{e_{UR}^T e_{UR} / (n-k)} \sim F(m, (n-k))$$

Nếu  $F > F_{\alpha}(m, (n-k))$  thì giả thiết  $H_0$  bị bác bỏ.

### 3.12 DỰ BÁO:

Chúng ta có thể sử dụng mô hình hồi quy vào dự báo: dự báo giá trị trung bình và dự báo giá trị cá biệt.

Cho  $X^0 = \begin{bmatrix} 1 \\ X_2^0 \\ X_3^0 \\ \vdots \\ X_k^0 \end{bmatrix}$

Dự báo giá trị trung bình:  $E(Y/X^0)$

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \dots \hat{\beta}_k X_k = X^T \hat{\beta}$$

Với  $X = X^0$  ta có  $(\hat{Y}_0 / X^0) = X^{0T} \hat{\beta} \Rightarrow \text{var}(\hat{Y}_0 / X^0) = X^{0T} \text{var}(\hat{\beta}) X^0$

$$\text{var}(\hat{Y}_0 / X^0) = \sigma^2 X^{0T} (X^T X)^{-1} X^0 \text{ vì } \text{var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

Nhưng  $\sigma^2$  chưa biết nên phải dùng ước lượng không chệch là  $\hat{\sigma}^2$  của nó:

$$\text{Var}(\hat{Y} / X^0) = \hat{\sigma}^2 X^{0T} (X^T X)^{-1} X^0$$

$$\text{Se}(\hat{Y}_0 / X^0) = \sqrt{\hat{\sigma}^2 X^{0T} (X^T X)^{-1} X^0}$$

$$\hat{Y}_0 - t_{\alpha/2}(n-k) \text{se}(\hat{Y}_0 / X^0) \leq E(Y / X^0) \leq \hat{Y}_0 + t_{\alpha/2}(n-k) \text{se}(\hat{Y}_0 / X^0)$$

Dự báo giá trị cá biệt:

$$Y_i = X^T \hat{\beta} + e_i \Rightarrow \text{Var}(Y_0 / X^0) = \text{Var}(X^{0T} \hat{\beta}) + \sigma^2$$

$$\text{Var}(Y_0 / X^0) = \hat{\sigma}^2 \left[ 1 + X^{0T} (X^T X)^{-1} X^0 \right]$$

$$\text{Se}(Y_0 / X^0) = \sqrt{\hat{\sigma}^2 \left[ 1 + X^{0T} (X^T X)^{-1} X^0 \right]}$$

$$\hat{Y}_0 - t_{\alpha/2}(n-k) \text{Se}(Y_0 / X^0) \leq (Y_0 / X^0) \leq \hat{Y}_0 + t_{\alpha/2}(n-k) \text{Se}(Y_0 / X^0)$$

Với việc trình bày mô hình hồi quy bằng ngôn ngữ ma trận đã cung cấp cho chúng ta một công nghệ mà nhờ đó có thể sử dụng kỹ thuật tính toán, tự động hoá toàn bộ quá trình tính toán, phân tích và dự báo. Mô hình hồi quy nhiều biến được giải một cách nhanh chóng nhờ phần mềm MFIT3 và MFIT4 hoặc phần mềm Stata.

### 3.13 Một số dạng của hàm hồi quy.

Dạng của hàm hồi quy là một vấn đề quan trọng, một trong những nhân tố có tính chất quyết định đối với kết quả nghiên cứu. Tuy vậy, vấn đề “dạng của hàm hồi quy” lại không có một cơ sở lý thuyết đủ mạnh để có thể khẳng định dạng của hàm hồi quy là dạng này mà không phải là dạng khác. Dạng của mô hình hồi quy là một vấn đề thực nghiệm.

Một trong những phương pháp thường được dùng là biểu diễn các số liệu lên hệ toạ độ. Nếu như đồ thị chỉ ra quan hệ giữa hai biến là tuyến tính thì dạng hàm của mô hình là tuyến tính, nếu quan hệ được chỉ ra là hàm bậc 2,3 (phi tuyến), v.v... thì dạng hàm của mô hình được chọn một cách tương ứng. Phương pháp này được sử dụng trong mô hình hồi quy giản đơn. Nó sẽ không hữu ích nếu chúng ta có mô hình hồi quy bội.

Ở đây sẽ trình bày một số dạng hàm đơn giản hay gặp trong ứng dụng thực tiễn.

#### 3.13.1 Hàm có hệ số co giãn không đổi- hàm Cobb-Douglas

Hàm cobb-Douglas có dạng:  $Y = \beta_1 X^{\beta_2}$

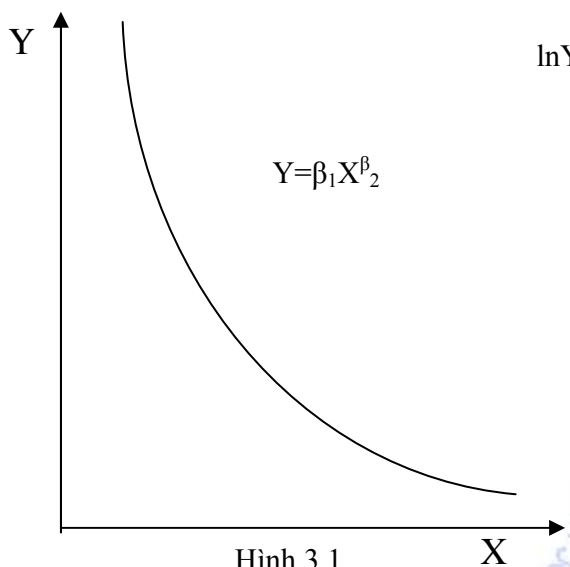
Hàm này là hàm phi tuyến đối với  $X$  và phi tuyến đối với tham số  $\beta_2$ . Tuy nhiên có thể biến đổi về dạng tuyến tính đối với tham số. Lấy ln hai vế, ta có:

$$\ln Y = \ln \beta_1 + \beta_2 \ln X$$

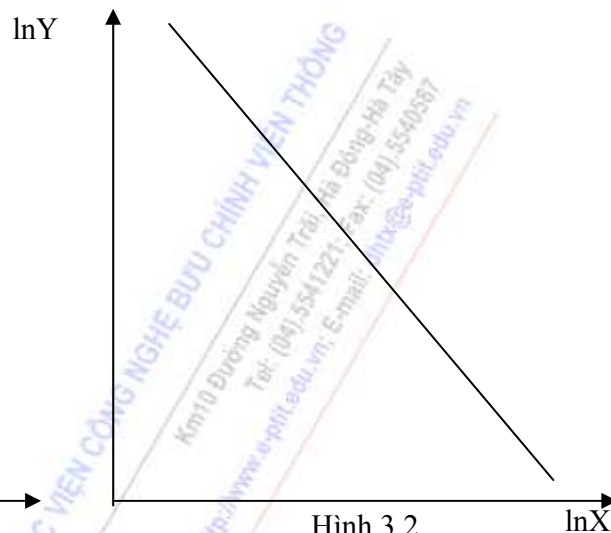
$$\text{Đặt } \beta_1' = \ln \beta_1; \quad Y' = \ln Y; \quad X' = \ln X$$

Ta có:  $Y' = \beta_1' + \beta_2 X'$

Đây là mô hình giản đơn mà ta đã biết. Ta có thể minh hoạ hàm ban đầu và hàm sau khi biến đổi bằng đồ thị:



Hình 3.1



Hình 3.2

Hàm Cobb-Douglas có thể mở rộng cho trường hợp có nhiều biến giải thích:

$$Y = \beta_1 X_2^{\beta_2} X_3^{\beta_3} \dots X_m^{\beta_m}$$

Bằng phép biến đổi, ta có:

$$\ln Y = \ln \beta_1 + \beta_2 \ln X_2 + \beta_3 \ln X_3 + \dots + \beta_m \ln X_m$$

Chúng ta dễ dàng có hàm tuyến tính đối với các tham số. Trong hàm Cobb-Douglas, hệ số co giãn của  $Y$  đối với  $X_i$  bằng  $\beta_i$ .

### 3.13.2 Hàm có dạng: $Y_t = \beta(1+r)^t$

trong đó  $t$  là biến thời gian.

Hàm này thường dùng để đo sự tăng trưởng của yếu tố  $Y_t$  theo thời gian,  $r$  là tỷ lệ tăng trưởng.

Ở năm (thời kỳ)  $t = 0$ , ta có  $Y_0 = \beta$ , do đó  $Y_t = Y_0(1+r)^t$

Biến đổi hàm về dạng tuyến tính đối với tham số:

$$\ln Y_t = \ln Y_0 + t \ln(1+r).$$

Đặt:  $Y'_t = \ln Y_t$ ;  $\beta_0 = \ln Y_0$ ;  $\beta_1 = \ln(1+r)$

Khi đó:  $Y'_t = \beta_0 + \beta_1 t$

Dễ dàng ước lượng được hàm này và từ đó tìm được  $Y_0$  và  $r$ .

### 3.13.3 Hàm dạng Hypecbol: $Y = \beta_1 + \frac{\beta_2}{X}$

Hàm này là phi tuyến đối với X, nhưng tuyến tính đối với các tham số. Sau đây là một số trường hợp quan trọng của hàm này:

a)  $\beta_1, \beta_2 > 0$ , khi đó đồ thị có dạng hình 3.3:

Trong trường hợp này có mức tiệm cận dưới, dù có tăng đến đâu. Y không thể nhỏ hơn  $\beta_1$ .

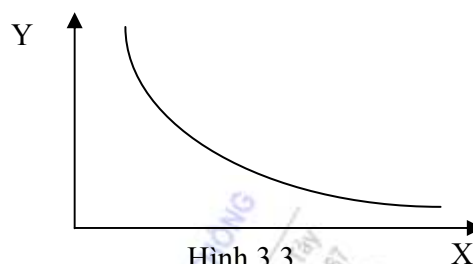
Hàm này thường được dùng khi phân tích chi phí trung bình để sản xuất ra một đơn vị sản phẩm.

b)  $\beta_1 > 0; \beta_2 < 0$

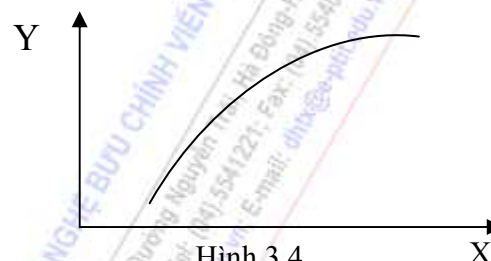
Trong trường hợp này có mức tiệm cận trên. Đây gọi là đường cong Engel, nghiên cứu mức chi tiêu phụ thuộc vào thu nhập.

c)  $\beta_1 < 0; \beta_2 > 0$

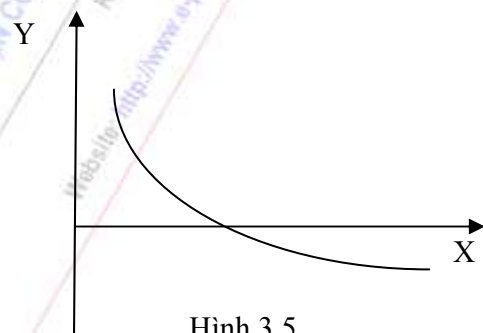
Đây là đường cong Phillips



Hình 3.3



Hình 3.4



Hình 3.5

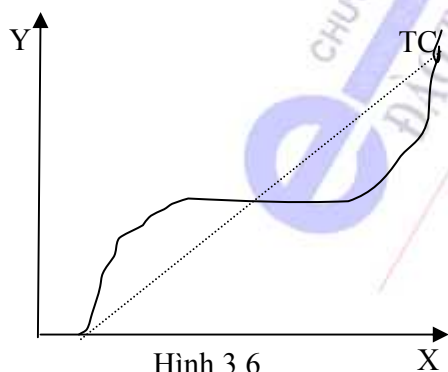
### 3.13.4 Hàm có dạng đa thức:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 \text{ - Mô hình đa thức bậc 2}$$

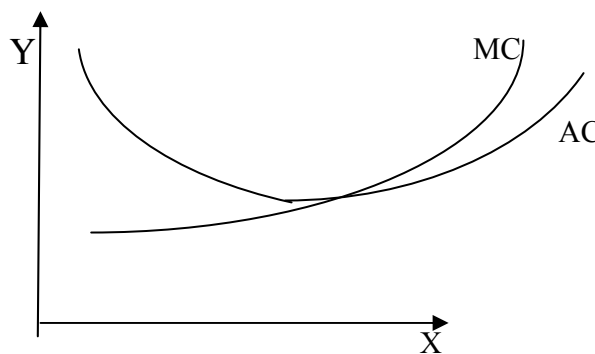
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 \text{ - Mô hình đa thức bậc 3}$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_k X^k \text{ - Mô hình đa thức bậc k}$$

Hàm này thường được sử dụng để nghiên cứu quan hệ giữa chi phí và số lượng sản phẩm được sản xuất ra trong một thời kỳ nhất định. Chẳng hạn Y – tổng chi phí; X- Số sản phẩm. Nếu như xây dựng được hàm này thì ta dễ dàng tìm được chi phí trung bình và chi phí biên.



Hình 3.6



Hình 3.7

TC: Tổng chi phí; MC: Chi phí biên; AC: Chi phí trung bình.

Trên đây là một số dạng của mô hình hồi quy. Tất nhiên còn rất nhiều dạng khác. Trong thực tế để vận dụng mô hình này hay mô hình khác trước hết phải hiểu được quan hệ giữa các biến, tính chất của mô hình (các dạng hàm) muốn vận dụng.

## TÓM TẮT NỘI DUNG CHƯƠNG 3

Mô hình hồi quy bội xét ảnh hưởng của nhiều biến độc lập đến một biến phụ thuộc, trên cơ sở các giả thiết giống như giả thiết trong phần hồi quy hai biến, chỉ thêm giả thiết là các biến độc lập không có quan hệ tuyến tính (không có hiện tượng đa cộng tuyến). Để ước lượng các tham số của mô hình ta dùng phương pháp OLS thông thường. Sau khi xác định được các tham số ta tiến hành tìm ma trận hiệp phương sai, hệ số xác định bội, ma trận tương quan, hệ số tương quan riêng phần và kiểm tra giả thiết cũng như tìm khoảng tin cậy của các hệ số hồi quy riêng-Kiểm định T. Kiểm định giả thiết về sự phù hợp của hàm hồi quy, phân tích phương sai và cuối cùng nếu hàm hồi quy là phù thì tiến hành dự báo giá trị trung bình và cá biệt. Chương này cũng đưa ra một số dạng hàm hay gặp trong thực tế.

## CÂU HỎI VÀ BÀI TẬP ÔN CHƯƠNG 3

### I. Câu hỏi lý thuyết:

- Viết mô hình hồi quy tuyến tính k biến dưới dạng ma trận? Giải thích các thành phần trong mô hình?
- Nêu các giả thiết đối với mô hình hồi quy tuyến tính k biến? Phương pháp ước lượng các tham số của mô hình?
- Trình bày công thức xác định hệ số hồi quy riêng và hệ số hồi quy bội? Ma trận tương quan, hệ số tương quan riêng phần và ma trận hiệp phương sai? ý nghĩa của từng chỉ tiêu trên?
- Phương pháp tìm khoảng tin cậy của các hệ số hồi quy và cách kiểm định các giả thiết đối với mô hình hồi quy tuyến tính k biến?
- Phương pháp dự báo giá trị trung bình và giá trị cá biệt đối với mô hình hồi quy tuyến tính k biến?
- Nêu các dạng hàm phổ biến và giải thích ý nghĩa kinh tế của các tham số trong mô hình có dạng hàm Cobb-Douglas, hàm có dạng đa thức?

### II. Bài tập:

- Bảng sau đây cho số liệu về doanh thu (Y), chi phí chào hàng ( $X_2$ ), chi phí quảng cáo ( $X_3$ ) trong năm 2002 của 12 khu vực bán hàng của một công ty (đơn vị: triệu đồng). Hãy ước lượng hàm hồi quy tuyến tính của Y phụ thuộc vào  $X_2$ ,  $X_3$  và trả lời các câu hỏi sau:

STT	1	2	3	4	5	6	7	8	9	10	11	12	$\Sigma$
Y	1270	1490	1060	1626	1020	1800	1610	1280	1390	1440	1590	1380	16956
$X_{2i}$	100	106	60	160	70	170	140	120	116	120	140	150	1452
$X_{3i}$	180	248	190	240	150	260	250	160	170	230	220	150	2448

- Tìm ước lượng phương sai của yếu tố ngẫu nhiên?

- b) Tìm ước lượng phương sai của các hệ số hồi quy mẫu
- c) Xác định hệ số hồi quy bội  $R^2$  và hệ số hồi quy bội có điều chỉnh  $\bar{R}^2$ ?
- d) Tìm khoảng tin cậy của các hệ số hồi quy với mức ý nghĩa  $\alpha = 5\%$
- e) Kiểm định giả thiết đồng thời,  $H_0: \beta_2 = \beta_3 = 0$ ? Cho biết ý nghĩa của kết quả?
- f) Dự báo doanh thu ta nên dùng hàm nào trong các hàm sau đây:

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + U_i \quad (1)$$

$$Y_i = \beta_1 + \beta_2 X_{3i} + U_i \quad (2)$$

$$Y_i = \gamma_1 + \gamma_2 X_{2i} + \gamma_3 X_{3i} + U_i \quad (3)$$

g). Dự báo doanh thu trung bình của công ty khi chi phí chào hàng là 121 triệu đồng và chi phí quảng cáo là 204 triệu đồng với hệ số tin cậy 95%?

2. Bảng sau đây cho số liệu về biến phụ thuộc (Y) có quan hệ tuyến tính với các biến độc lập  $X_2$ ,  $X_3$  (đơn vị: triệu đồng). Hãy ước lượng hàm hồi quy tuyến tính của Y phụ thuộc vào  $X_2$ ,  $X_3$  và trả lời các câu hỏi sau:

<b>Y</b>	40	44	46	48	52	58	60	68	74	80
<b><math>X_2</math></b>	6	10	12	14	16	18	22	24	26	32
<b><math>X_3</math></b>	4	4	5	7	9	12	14	20	21	24

- a) Giải thích ý nghĩa của các hệ số hồi quy nhận được?
- b) Biến  $X_2$  ( $X_3$ ) có ảnh hưởng đến biến phụ thuộc Y hay không?
- c) Tìm khoảng tin cậy cho các hệ số hồi quy riêng?
- d) Giải thích ý nghĩa của hệ số  $R^2$  nhận được?
- e) Cả  $X_2$  và  $X_3$  đều không ảnh hưởng đến Y?
- f) Có thể bỏ biến  $X_3$  ra khỏi mô hình được không? Vì sao?
- g) Hãy ước lượng mô hình bằng phương pháp ma trận?
- h) Dự báo giá trị trung bình và cá biệt khi  $X_2 = 20$ ;  $X_3 = 15$ ?

## CHƯƠNG 4: HỒI QUY VỚI BIẾN ĐỘC LẬP LÀ BIẾN GIẢ

### GIỚI THIỆU

Trong các mô hình hồi quy tuyến tính mà chúng ta đã xem xét từ các chương trước cho đến nay thì các biến giải thích đều là các biến số lượng. Các biến đó có thể nhận giá trị bằng số. Chẳng hạn tiền lương của cán bộ, doanh số bán ra của một cửa hàng, chi tiêu cho quảng cáo, cung tiền,...là những biến số lượng. Như nhưng trong thực tế có nhiều trường hợp các biến giải thích (hoặc thậm chí cả biến phụ thuộc) là biến chất lượng (biến định tính). Trong chương này ta sẽ nghiên cứu hồi quy khi biến giải thích là biến chất lượng.

Để thực hành được đối với dạng mô hình hồi quy biến giả đòi hỏi người học cần nắm vững phương pháp OLS, cách chọn số biến,..

### NỘI DUNG

#### 4.1. BẢN CHẤT CỦA BIẾN GIẢ - MÔ HÌNH TRONG ĐÓ BIẾN GIẢI THÍCH LÀ BIẾN GIẢ

Biến chất lượng như đã nói ở trên thường chỉ ra có hoặc không có một thuộc tính nào đó, chẳng hạn như nam hay nữ; khu vực tư nhân hay nhà nước,...vấn đề đặt ra là làm thế nào để lượng hóa được những thuộc tính như vậy. Trong phân tích hồi quy người ta sử dụng kỹ thuật gọi là kỹ thuật biến giả. Kỹ thuật này cho phép ta lượng hóa được những thuộc tính như vậy. Chẳng hạn để giải thích cho việc một số thanh niên vào trường đại học, một số khác thì không, chúng ta tạo ra biến giả mà nhận giá trị là 1 nếu thanh niên vào đại học và nhận giá trị là không nếu thanh niên đó không vào đại học. Chúng ta cũng sẽ chỉ ra biến giả có thể được sử dụng như thế nào trong phạm vi hồi quy để giải thích cho sự kiện là có những quan sát trong phạm trù (thuộc tính) đã cho gắn với một tập các tham số hồi quy còn các quan sát khác trong phạm trù thứ 2 (hoặc thứ 3) lại gắn với những tham số hồi quy khác. Biến giả được sử dụng trong mô hình hồi quy giống như biến số lượng thông thường.

Giả sử một công ty sử dụng hai quá trình sản xuất (kí hiệu quá trình sản xuất A và quá trình sản xuất B) để sản xuất ra một loại sản phẩm. Giả sử sản phẩm thu được từ mỗi một quá trình sản xuất là đại lượng ngẫu nhiên có phân phối chuẩn và có kỳ vọng khác nhau nhưng phương sai như nhau. Chúng ta có thể biểu thị quá trình sản xuất đó như một phương trình hồi quy

$$Y_i = \beta_1 + \beta_2 D_i + U_i \quad (4.1)$$

Trong đó  $Y_i$  là sản lượng sản phẩm gắn với quá trình thứ  $i$

$D_i$  là biến giả nhận 1 trong 2 giá trị:

$$D_i = \begin{cases} 1 & \text{nếu sản lượng sản phẩm thu được từ quá trình sản xuất A} \\ 0 & \text{nếu sản lượng sản phẩm thu được từ quá trình sản xuất B} \end{cases}$$

Mô hình hồi quy trên đây giống như mô hình hồi quy 2 biến mà chúng ta đã gặp trước đây chỉ khác là biến số lượng  $X$  được thay bằng biến giả  $D$ . Căn cứ vào mô hình này chúng ta có thể

biết được sản lượng trung bình do quá trình sản xuất A có khác với sản lượng trung bình do quá trình sản xuất B tạo ra hay không?

Hệ số chặn  $\beta_1$  của hồi quy tuyến tính đo sản lượng trung bình gắn với quá trình sản xuất B, trong khi đó độ dốc  $\beta_2$  của đường hồi quy đo sự khác nhau về sản lượng sinh ra do việc thay đổi từ quá trình sản xuất B đến quá trình sản xuất A.

Điều này có thể thấy bằng 2 cách lấy giá trị kỳ vọng cả 2 vế của phương trình (4.1) ứng với  $D_i = 0$  và  $D_i = 1$ :

$$E(Y_i | D_i = 0) = \beta_1$$

$$E(Y_i | D_i = 1) = \beta_1 + \beta_2$$

Kiểm định giả thiết  $H_0: \beta_2 = 0$  cung cấp kiểm định về giả thiết là không có sự khác nhau về sản lượng do quá trình sản xuất A và B tạo ra. Điều này dễ làm được như đã chỉ ra trước đây.

Thủ tục biến giả có thể dễ dàng mở rộng cho trường hợp có nhiều hơn 2 phạm trù. Chẳng hạn trong thí dụ ở trên ta giả thiết có 3 quá trình sản xuất khác nhau có thể sử dụng để sản xuất ra sản phẩm và người ta hi vọng giải thích cho vấn đề là sản lượng được sản xuất ra cho mỗi quá trình có thể không như nhau. Trong trường hợp này ta sẽ đưa vào 2 biến giả là  $D_1$  và  $D_2$ . Chúng ta sẽ xét mô hình:

$$Y_i = \beta_1 + \beta_2 D_{1i} + \beta_3 D_{2i} + U_i \quad (4.2)$$

Trong đó:

$$D_1 = \begin{cases} 1 & \text{nếu sản lượng sản phẩm thu được từ quá trình sản xuất A} \\ 0 & \text{nếu sản lượng sản phẩm thu được từ quá trình khác} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{nếu sản lượng sản phẩm thu được từ quá trình sản xuất B} \\ 0 & \text{nếu sản lượng sản phẩm thu được từ quá trình khác} \end{cases}$$

Như vậy 3 quá trình sản xuất này được biểu thị dưới dạng các kết hợp sau của các giá trị biến giả:

Quá trình sản xuất	$D_1$	$D_2$
A	1	0
B	0	1
C	0	0

Bằng việc lấy kỳ vọng cho mỗi một trong 3 trường hợp này chúng ta có thể giải thích kết quả hồi quy:

$$E(Y_i | D_1 = 1; D_2 = 0) = \beta_1 + \beta_2;$$

$$E(Y_i | D_1 = 0; D_2 = 1) = \beta_1 + \beta_3$$

$$E(Y_i | D_1 = 0; D_2 = 0) = \beta_1$$

Hệ số chặn của hồi quy biểu thị giá trị kỳ vọng của sản lượng do quá trình sản xuất C tạo ra. Hệ số góc thứ nhất đo sự thay đổi trung bình về sản lượng do việc chuyển từ quá trình sản xuất C

sang quá trình sản xuất A và hệ số góc thứ 2 tức là  $\beta_3$  đo thay đổi trung bình về sản lượng khi thay đổi từ quá trình sản xuất C sang quá trình sản xuất B.

Kiểm định giả thiết  $H_0: \beta_2 = 0$  có nghĩa là không có sự khác nhau giữa quá trình sản xuất A và quá trình sản xuất C. Giả thiết  $H_0: \beta_3 = 0$  cũng có ý nghĩa tương tự nhưng lại so sánh 2 quá trình sản xuất B và C.

**Thí dụ 4.1:** Để xem xét kết quả sản lượng do 2 quá trình sản xuất A và B có khác nhau hay không người ta tiến hành lấy một mẫu được cho trong bảng dưới đây. Hãy phân tích kết quả hồi quy thu được.

Dựa vào các kết quả dưới đây ta thấy rằng sản lượng trung bình 1 ca của quá trình sản xuất B ước lượng là 18000 kg =  $\hat{\beta}_1$ , còn sản lượng trung bình 1 ca đã được ước lượng của quá trình sản xuất A là 21,280 kg =  $\hat{\beta}_1 + \hat{\beta}_2$

**Bảng 4.1**

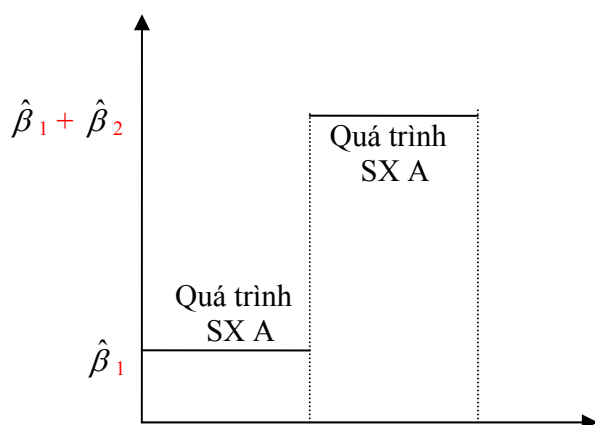
Quá trình sản xuất A là 1, quá trình sản xuất B là 0	Sản lượng trong 1 ca hoạt động
1	22,0
0	19,0
0	18,0
1	21,0
0	18,5
1	21,0
1	20,5
0	17,0
0	17,5
1	21,2

Kết quả hồi quy như sau:

$$Y_i = 18 + 3.2 D_i$$

Biến	Hệ số	Sai lệch tiêu chuẩn	t
D	3,28	0,44	7,439
Const	18	0,32	57,74

$R^2 = 0.8737$ .  $\hat{\beta}_2$  có ý nghĩa về mặt thống kê, kết quả chỉ ra rằng sản lượng trung bình của 2 quá trình đó là khác nhau. Hồi quy trên có thể mô tả trên hình 4.1.



Hình 4.1

Trước khi chuyển sang mục sau ta cần một số chú ý:

1. Để phân biệt 2 phạm trù nam hoặc nữ hay quá trình sản xuất A hoặc B người ta dùng một biến giả. Để phân biệt 3 phạm trù người ta dùng 2 biến giả. Một cách tổng quát để phân biệt N phạm trù người ta dùng N-1 biến giả. Số biến giả thấp hơn số phạm trù là 1 để tránh tính đa cộng tuyến hoàn hảo. Để phân biệt 3 quá trình sản xuất A, B và C ta chỉ sử dụng 2 biến giả  $D_1$  và  $D_2$ , nếu ta đưa thêm một biến giả  $D_3$  nữa chẳng hạn.

$$D_3 = \begin{cases} 1 & \text{nếu sản lượng sản phẩm thu được từ quá trình sản xuất C} \\ 0 & \text{nếu sản lượng sản phẩm thu được từ quá trình khác} \end{cases}$$

thì việc đưa thêm  $D_3$  vào không cho thêm thông tin mà lại gặp đa cộng tuyến hoàn hảo (xem chương V). Trong trường hợp này thì ước lượng bình phương nhỏ nhất của các tham số hồi quy không thể thu được. Có hiện tượng đa tuyến tính hoàn hảo vì:

$$D_3 = 1 - D_1 - D_2$$

2. Phạm trù được gán giá trị không được coi là phạm trù cơ sở. Phạm trù được gọi là cơ sở theo nghĩa việc so sánh được tiến hành với phạm trù này. Như vậy trong mô hình trên quá trình sản xuất C là phạm trù cơ sở, nghĩa là nếu ta ước lượng hồi quy (5.2) với

$$D_1 = 0 ; D_2 = 0 \text{ thì chỉ có quá trình sản xuất C, hệ số chặn sẽ là } \hat{\beta}_1$$

3. Hệ số  $\beta_2$  gắn với biến giả  $D_1$  được gọi là hệ số chặn chênh lệch, vì nó cho biết giá trị của số hạng chặn của phạm trù nhận giá trị bằng 1 sẽ khác bao nhiêu với hệ số chặn của phạm trù cơ sở.

## 4.2. HỒI QUY VỚI MỘT BIẾN LƯỢNG VÀ MỘT BIẾN CHẤT.

Trong mục này ta sẽ xét mô hình hồi quy chỉ có một biến lượng và một biến chất với số phạm trù nhiều hơn hoặc bằng 2. Trường hợp có nhiều biến lượng và một biến chất thì thủ tục cũng được xét tương tự như ta sẽ làm dưới đây chỉ khác là số biến lượng sẽ tăng lên. Để dễ theo dõi trong mục này ta chia ra làm 2 trường hợp: trường hợp 1 khi biến chất chỉ có 2 phạm trù, trường hợp 2 khi biến chất có nhiều hơn 2 phạm trù.

### 4.2.1. Trường hợp khi biến chất chỉ có 2 phạm trù

Trong trường hợp này, mô hình hồi quy sẽ đơn giản vì theo chú ý ở trên khi biến chất có 2 phạm trù thì chỉ cần đặt 1 biến giả là đủ. Thí dụ ta xét mô hình sau:

$$Y_i = \beta_1 + \beta_2 D_i + \beta_3 X_i + U_i \quad (4.3)$$

Trong đó:

$Y_i$ : là tiền lương hàng tháng của một công nhân cơ khí  $i$

$X_i$ : là bậc thợ của công nhân  $i$ ,

$D_i = 1$  nếu công nhân  $i$  làm việc trong khu vực tư nhân

$D_i = 0$  nếu công nhân  $i$  làm việc trong khu vực quốc doanh.

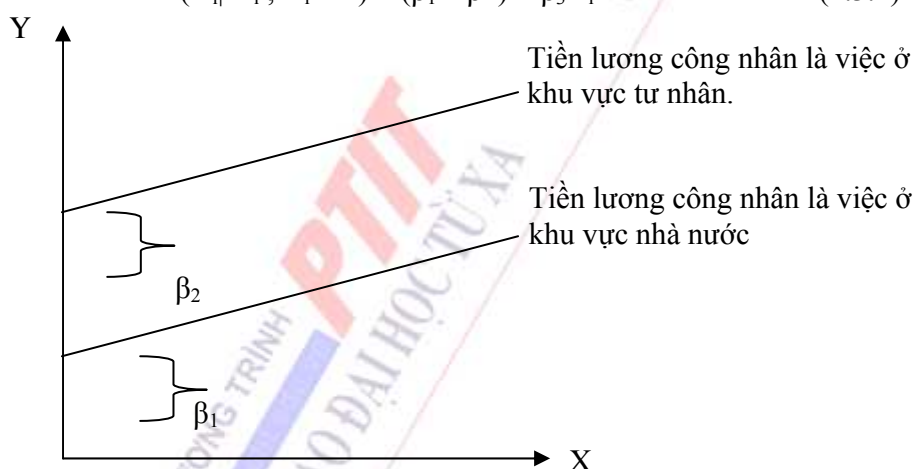
Mô hình có một biến lượng đó là bậc thợ của người công nhân và một biến chất chỉ rõ công nhân đó làm việc thuộc khu vực nào. Nếu ta giả thiết  $E(U_i) = 0$  thì (4.3) có thể cho ta thấy liệu tiền lương của người công nhân làm việc ở khu vực tư nhân có khác tiền lương của người công nhân làm việc ở khu vực nhà nước không nếu các điều kiện khác không thay đổi. Bằng cách lấy kỳ vọng cả 2 vế (4.3) ta được:

Tiền lương trung bình của người công nhân cơ khí làm việc trong khu vực nhà nước:

$$E(Y_i | X_i, D_i = 0) = \beta_1 + \beta_3 X_i \quad (4.3.1)$$

Tiền lương trung bình của người công nhân cơ khí làm việc trong khu vực tư nhân:

$$E(Y_i | X_i, D_i = 1) = (\beta_1 + \beta_2) + \beta_3 X_i \quad (4.3.2)$$



**Hình 4.2**

Hình 4.2 chỉ cho chúng ta thấy rằng tiền lương của công nhân cơ khí làm việc trong khu vực tư nhân và nhà nước tính theo bậc thợ có cùng độ dốc  $\beta_3$  nhưng lại khác nhau về hệ số chặn. Nói một cách khác mô hình này giả thiết rằng mức lương trung bình của người công nhân ngành cơ khí làm việc ở khu vực tư nhân khác với mức tiền lương trung bình của công nhân cơ khí làm việc ở khu vực nhà nước nhưng tốc độ tăng lương trung bình theo bậc thì như nhau.

Nếu giả thiết về tốc độ đã nêu trên là có giá trị thì kiểm định giả thiết rằng 2 hồi quy (4.3.1) và (4.3.2) có cùng hệ số chặn có thể tiến hành dễ dàng bằng cách ước lượng hồi quy (4.3) và chú rằng ý nghĩa về mặt thống kê của  $\beta_2$  đã được ước lượng trên cơ sở kiểm định  $t$ . Nếu  $t$  chỉ ra rằng

$\beta_2$  là có ý nghĩa về mặt thống kê thì chúng ta từ bỏ giả thiết  $H_0$  là tiền lương của công nhân cơ khí ở 2 khu vực kinh tế là như nhau.

#### 4.2.2. Trường hợp khi biến chất có nhiều hơn 2 phạm trù.

Khi biến chất có nhiều hơn 2 phạm trù thì vấn đề cũng không phức tạp hơn nhiều bởi vì theo chú ý ở trên nếu số phạm trù là  $N$  thì ta đưa vào mô hình hồi quy  $N-1$  biến giả làm biến giải thích.

Thí dụ căn cứ vào số liệu chéo người ta muốn hồi quy thu nhập hàng năm của một cán bộ giảng dạy đại học đối với tuổi nghề giảng dạy và vùng mà anh ta giảng dạy. Vì biến vùng là biến chất, trên thực tế chúng ta có thể căn cứ vào 3 vùng khác nhau trong cả nước là Bắc, Trung, Nam. Như vậy trong trường hợp này, biến chất của ta có 3 phạm trù, theo chú ý ở trên chúng ta sẽ đưa vào mô hình hồi quy 2 biến giả.

Giả sử rằng cả 3 hồi quy có cùng độ dốc nhưng khác nhau hệ số chặn, chúng ta có mô hình sau:

$$Y_i = \beta_1 + \beta_2 D_{1i} + \beta_3 D_{2i} + \beta_4 X_i + U_i \quad (4.4)$$

Trong đó:

$Y_i$ : là thu nhập hàng năm của một giảng viên đại học

$X_i$ : là tuổi nghề của giảng viên

$$D_1 = \begin{cases} 1 & \text{nếu giảng viên } i \text{ thuộc một trường đại học ở miền Bắc} \\ 0 & \text{nếu giảng viên thuộc một trường không phải ở miền Bắc} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{nếu giảng viên } i \text{ thuộc một trường đại học ở miền Nam} \\ 0 & \text{nếu giảng viên thuộc một trường không phải ở miền khác.} \end{cases}$$

Như vậy, ta coi giảng viên thuộc một trường đại học ở miền Trung là phạm trù cơ sở, hệ số chặn chênh lệch  $\beta_2$  và  $\beta_3$  cho chúng ta biết chênh của các phạm trù khác với chênh của phạm trù cơ sở bao nhiêu. Chúng ta có thể tính được nếu giả thiết  $E(U_i) = 0$  thì từ (4.4) ta có:

Thu nhập trung bình của một cán bộ giảng dạy ở một trường đại học ở miền Trung:

$$E(Y_i | D_1 = 0; D_2 = 0, X_i) = \beta_1 + \beta_4 X_i \quad (4.4.1)$$

Thu nhập trung bình của một cán bộ giảng dạy ở một trường đại học ở miền Bắc:

$$E(Y_i | D_1 = 1; D_2 = 0, X_i) = (\beta_1 + \beta_2) + \beta_4 X_i \quad (4.4.2)$$

Thu nhập trung bình của một cán bộ giảng dạy ở một trường đại học ở miền Nam:

$$E(Y_i | D_1 = 1; D_2 = 1, X_i) = (\beta_1 + \beta_3) + \beta_4 X_i \quad (4.4.3)$$

Giả sử  $\beta_1 > 0$  ta có minh họa sau:

Sau khi ước lượng hồi quy (4.4) chúng ta dễ thấy rằng liệu có sự khác nhau về thu nhập của cán bộ giảng dạy ở các miền khác nhau của đất nước không.

Thí dụ: Căn cứ vào số liệu chéo về doanh thu trên đầu người từ các tỉnh và thu nhập quốc dân tính theo đầu người của các tỉnh đó. Người ta muốn hồi quy  $\log R$  ( $\log$  thu nhập/người) đối với  $\log G$  ( $\log$  thu nhập quốc dân/người) và tính vùng của mỗi tỉnh. Theo cách phân chia đã biết 1 tỉnh thuộc một trong 3 vùng Bắc (N), Trung (C) hoặc Nam (S). Số liệu cho trong bảng 4.2

**Bảng 4.2**

Tỉnh	Vùng	G	R	Tỉnh	Vùng	G	R
H.Giang	N	53099	15433	Q.Ngãi	C	921501	51778
T.Q	N	1026246	32534	B.Định	C	974381	71581
C.Bắc	N	939592	25915	P.Yên	C	1162995	69315
L.Sơn	N	1105916	79225	K.Hòa	C	2195192	204879
L.Châu	N	1173320	27660	N.Thuận	C	506939	44556
L.Cai	N	819802	35570	B.Thuận	C	1825380	105927
Y.Bái	N	1506025	46010	K.Tum	C	1181003	65642
B.Thái	N	1565816	48172	G.Lai	C	1024570	69518
S.La	N	643444	29338	Đ.Lắc	C	1725293	83590
Q.Ninh	N	2195642	311755	L.Đồng	C	1493055	107289
H.Bình	N	2040215	72980	HCM	S	4997041	962414
V.Phú	N	1229249	74594	Đ.Nai	S	2370994	239017
H.Bắc	N	1021338	47667	VT-BR	S	1991269	907834
H.Nội	N	3631625	1062935	S.Bé	S	1491414	192057
H.Phòng	N	2070000	160291	T.Ninh	S	1553681	204342
H.Tây	N	1142000	45778	L.An	S	1762790	147537
H.Hưng	N	1303451	105546	T.Giang	S	1057203	109833
T.Bình	N	1162747	51436	A.Giang	S	1760437	142684
N.Hà	N	1181747	53902	Đ.Tháp	S	1370071	100655
N.Bình	N	1062831	42545	B.Tre	S	1532974	82319
T.Hóa	C	1200650	102550	V.Long	S	2021798	116666
Ng.An	C	930844	59498	T.Vinh	S	2024467	66530
H.Tĩnh	C	910650	30163	S.Trăng	S	1571656	84334
Q.Bình	C	992594	41673	Cần Thơ	S	1760304	121340
Q.Trị	C	943805	70579	K.Giang	S	2329861	154636
T.T.Huế	C	1009223	90856	M.Hải	S	1736902	142079
QN-ĐN	C	1367673	198297				

Mô hình hồi quy là:

$$\text{Log } R_i = \beta_1 + \beta_2 D_{1i} + \beta_3 D_{2i} + \beta_4 \log G_i + U_i$$

Trong đó:

Log R: ln (thu nhập/người)

Log G: ln (thu nhập quốc dân/người)

$$D_1 = \begin{cases} 1 & \text{nếu tỉnh đó thuộc miền Bắc} \\ 0 & \text{nếu tỉnh đó không phải là tỉnh miền Bắc} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{nếu tỉnh đó thuộc miền Nam} \\ 0 & \text{nếu tỉnh đó không phải là tỉnh miền Nam} \end{cases}$$

Kết quả hồi quy như sau:

$$\log R_i = -15,040082 - 0,42196 D_1 - 0,33754 D_2 + 1,94474 \log G_i$$

Ta có thể viết 3 phương trình hồi quy:

$$\text{Miền Trung: } \log R_i = -15,04 + 1,95 \log G_i$$

$$\text{Miền Bắc : } \log R_i = -15,46 + 1,95 \log G_i$$

$$\text{Miền Nam: } \log R_i = -15,83 + 1,95 \log G_i$$

Theo chú ý ở trên  $\beta_1$  cho ta biết  $\log$  (doanh thu) thay đổi khi chuyển từ tỉnh miền Trung đến một tỉnh miền Bắc.

$\beta_2$  cho ta biết doanh thu thay đổi khi chuyển từ miền Trung đến một tỉnh miền Nam.

#### 4.3 HỒI QUY VỚI MỘT BIẾN LƯỢNG VÀ HAI BIẾN CHẤT.

Trong mục này ta xét hồi quy với 1 biến lượng và 2 biến chất. Nguyên tắc đã trình bày ở trên được mở rộng cho trường hợp này. Số biến giả được đưa vào mô hình hồi quy phụ thuộc vào số biến chất và các phạm trù mà mỗi biến chất có.

Thí dụ: Chúng ta hãy quay lại thí dụ về thu nhập của một giảng viên đại học, bây giờ chúng ta giả thiết thêm rằng giới tính của giảng viên cũng là một nhân tố quan trọng để xác định thu nhập của giảng viên. Như vậy ta có 2 biến chất:

- Vùng có 3 phạm trù
- Giới tính có 2 phạm trù.

Mô hình sẽ có dạng:

$$Y_i = \beta_1 + \beta_2 D_{1i} + \beta_3 D_{2i} + \beta_4 D_{3i} + \beta_5 X_i + U_i \quad (4.5)$$

Trong đó:

$Y_i$ : là thu nhập hàng năm của giảng viên đại học

$X_i$ : là tuổi nghề của giảng viên

$$D_1 = \begin{cases} 1 & \text{nếu giảng viên thuộc một trường đại học miền Bắc} \\ 0 & \text{nếu giảng viên không thuộc một trường miền Bắc} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{nếu giảng viên thuộc một trường đại học miền Nam} \\ 0 & \text{nếu giảng viên không thuộc một trường miền N} \end{cases}$$

$$D_3 = \begin{cases} 1 & \text{nếu giảng viên là nam} \\ 0 & \text{nếu giảng viên là nữ.} \end{cases}$$

Phạm trù cơ sở bây giờ là giảng viên thuộc một trường đại học miền Trung. Giả sử  $E(U_i) = 0$  chúng ta có thể thu được kết quả sau bằng cách lấy kỳ vọng có điều kiện cả 2 vế (4.5)

Thu nhập trung bình của 1 giảng viên nữ ở một trường đại học miền Trung

$$E(Y_i | D_i = 0; D_2 = 0, D_3 = 0, X_i) = \beta_1 + \beta_5 X_i$$

Thu nhập trung bình của 1 giảng viên nam ở một trường đại học miền Trung

$$E(Y_i | D_i = 0; D_2 = 0, D_3 = 1, X_i) = (\beta_1 + \beta_4) + \beta_5 X_i$$

Thu nhập trung bình của 1 giảng viên nữ ở một trường đại học miền Bắc

$$E(Y_i | D_i = 1; D_2 = 0, D_3 = 0, X_i) = (\beta_3 + \beta_2) + \beta_5 X_i$$

Thu nhập trung bình của 1 giảng viên nam ở một trường đại học miền Bắc

$$E(Y_i | D_i = 1; D_2 = 0, D_3 = 1, X_i) = (\beta_1 + \beta_2 + \beta_4) + \beta_5 X_i$$

Thu nhập trung bình của 1 giảng viên nữ ở một trường đại học miền Nam

$$E(Y_i | D_i = 0; D_2 = 1, D_3 = 0, X_i) = (\beta_1 + \beta_3) + \beta_5 X_i$$

Thu nhập trung bình của 1 giảng viên nam ở một trường đại học miền Nam

$$E(Y_i | D_i = 0; D_2 = 1, D_3 = 1, X_i) = (\beta_1 + \beta_3 + \beta_4) + \beta_5 X_i$$

## 4.4 SO SÁNH HAI HỒI QUY

### 4.4.1. Tư tưởng cơ bản

Tất cả các mô hình hồi quy đã xem xét đến nay chúng ta đều giả thiết rằng biến lượng ảnh hưởng đến hệ số chặn nhưng không ảnh hưởng đến hệ số góc các hồi quy của các nhóm con khác nhau. Nhưng nếu hệ số góc khác nhau thì việc kiểm định về tính khác nhau của hệ số chặn sẽ ít có ý nghĩa. Một khía cạnh khác nữa là các tập số liệu mà chúng ta sử dụng là các tập số liệu đơn nhưng liệu một mô hình đã cho có thể áp dụng cho hai tập số liệu hay không? Chẳng hạn khi nghiên cứu mối quan hệ giữa tiết kiệm và thu nhập trước và sau khi chuyển đổi kinh tế của nước ta. Vì thế chúng ta cần phát triển phương pháp chung để tìm xem liệu 2 hồi quy có khác nhau hay không? Sự khác nhau nếu có thì ở hệ số chặn, hệ số góc hay cả hai.

Thí dụ: ta có hồi quy:

Thời kỳ trước cải tổ:

$$Y_i = \lambda_1 + \lambda_2 X_i + U_{1i} \quad (4.5a)$$

Với  $i = 1, n_1$

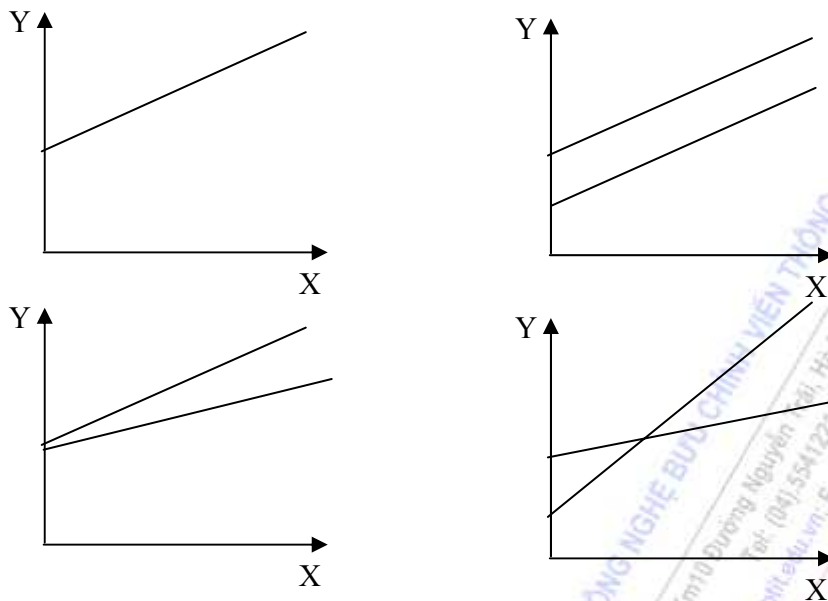
Thời kỳ sau cải tổ:

$$Y_j = \gamma_1 + \gamma_2 X_j + U_{2j} \quad (4.5b)$$

Với  $j = 1, n_2$

Có 4 khả năng xảy ra đối với 2 hồi quy này:

1.  $\lambda_1 = \gamma_1$  và  $\lambda_2 = \gamma_2$ , nghĩa là 2 hồi quy đồng nhất, trên đồ thị chúng chồng khít lên nhau, điều này chỉ ra ở hình 4.4a
2.  $\lambda_1 \neq \gamma_1$  và  $\lambda_2 = \gamma_2$ , nghĩa là 2 hồi quy có cùng hệ số góc, điều này chỉ ra ở hình 4.4b
3.  $\lambda_1 = \gamma_1$  và  $\lambda_2 \neq \gamma_2$ , nghĩa là 2 hồi quy có cùng hệ số chặn nhưng khác nhau về hệ số góc, điều này chỉ ra ở hình 4.4c
4.  $\lambda_1 \neq \gamma_1$  và  $\lambda_2 \neq \gamma_2$ , nghĩa là 2 hồi quy hoàn toàn khác nhau, điều này chỉ ra ở hình 4.4d



Để kiểm định sự bằng nhau của hệ số hồi quy chúng ta có thể sử dụng một trong hai kỹ thuật:

Một là kiểm định Chow và hai là sử dụng biến giả. Sau đây chúng ta xét cả hai kỹ thuật.

#### 4.4.2. So sánh 2 hồi quy - kiểm định Chow

Một trong những phương pháp phổ biến để kiểm định sự khác nhau giữa hai hồi quy là kiểm định của Chow. Kiểm định này dựa trên những giả thiết sau:

a. Các nhiễu  $U_{1i}$  và  $U_{2j}$  có phân phối chuẩn có kỳ vọng bằng 0 và phương sai không đổi và đều bằng  $\sigma^2$ :

$$U_{1i} \sim N(0, \sigma^2)$$

$$U_{2j} \sim N(0, \sigma^2)$$

b. Các  $U_{1i}$  và  $U_{2j}$  có phân phối độc lập.

Với giả thiết đã cho thì thủ tục kiểm định Chow như sau:

Bước 1: Kết hợp tất cả các quan sát của 2 thời kỳ lại ta được  $n = n_1 + n_2$  quan sát rồi ước lượng hồi quy gộp. Mô hình gộp của (4.5a) và (4.5b) lúc này có thể viết dưới dạng:

$$Y_i = \beta_1 + \beta_2 X_i + U_i \quad (4.6)$$

Từ hồi quy này chúng ta thu được tổng bình phương các phần dư là RSS với hệ số bậc tự do  $n_1 + n_2 - k$  (trong đó  $k$  là tham số được ước lượng, trong mô hình (4.6) thì  $k = 2$ )

Bước 2: Ước lượng riêng từng hồi quy (4.5a) và (4.5b) và thu được tổng bình phương các phần dư tương ứng từ mô hình (4.5a) là  $RSS_1$  và từ mô hình (4.5b) là  $RSS_2$  và bậc tự do tương ứng  $n_1 - k$  và  $n_2 - k$ .

Đặt  $\overline{RSS} = RSS_1 + RSS_2$  với bậc tự do  $n_1 + n_2 - 2k$

Bước 3: Sử dụng tiêu chuẩn F như sau:

$$F = \frac{(RSS - \overline{RSS}) / k}{\overline{RSS} / (n_1 + n_2 - 2k)} \quad (4.7)$$

Với bậc tự do  $k, n_1 + n_2 - 2k$ . Nếu giá trị  $F$  tính được vượt giá trị  $F$  tới hạn thì ta từ bỏ giả thiết rằng 2 hồi quy là như nhau. Điều này có nghĩa là hai tập số liệu không gộp được.

#### 4.4.3. So sánh hai hồi quy - thủ tục biến giả

Bằng kiểm định Chow cho phép ta so sánh 2 hồi quy. Sau đây ta sẽ trình bày thủ tục biến giả mà cũng sẽ cho ta gộp tất cả  $n_1$  và  $n_2$  quan sát lại với nhau và ước lượng hồi quy sau:

$$Y_i = \beta_1 + \beta_2 D_i + \beta_3 X_i + \beta_4 (D_i X_i) + U_i \quad (4.8)$$

Giả sử  $Y_i$  và  $X_i$  là tiết kiệm và thu nhập trước và sau khi chuyển đổi kinh tế.

$$D_i = \begin{cases} 1 & \text{đối với quan sát rơi vào trước thời kỳ chuyển đổi} \\ 0 & \text{đối với quan sát rơi vào từ chuyển đổi trở về sau} \end{cases}$$

Để thấy được ứng dụng của mô hình (4.8) ta lấy kỳ vọng có điều kiện cả 2 vế của phương trình này với giả thiết  $E(U_i) = 0$  chúng ta thu được:

$$E(Y_i | D_i = 0, X_i) = \beta_1 + \beta_3 X_i \quad (4.8a)$$

$$E(Y_i | D_i = 1, X_i) = (\beta_1 + \beta_2) + (\beta_3 + \beta_4) X_i \quad (4.8b)$$

Với ý nghĩa các kí hiệu đã cho ở trên ta có thể giải thích như sau:

(4.8a) là hàm tiết kiệm trung bình cho thời kỳ trước chuyển đổi kinh tế diễn ra. (4.8b) là hàm tiết kiệm trung bình từ khi chuyển đổi kinh tế về sau. Hai hàm này hoàn toàn giống như hai hàm đã cho (4.5a) và (4.5b) cho nên ước lượng (4.8) cũng tương đương với việc ước lượng hàm (4.5a) và (4.5b)

Trong (4.8) thì  $\beta_2$  chính là hệ số chặn biểu thị sự khác nhau của tiết kiệm giữa hai thời kỳ còn  $\beta_4$  chính là hệ số độ dốc khác nhau chỉ ra rằng hệ số góc của hàm tiết kiệm trước khi chuyển đổi khác với hệ số góc của thời kỳ từ khi chuyển đổi kinh tế là bao nhiêu.

#### 4.5 ẢNH HƯỞNG CỦA TƯƠNG TÁC GIỮA CÁC BIẾN GIẢ.

Chúng ta xét mô hình sau:

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 X_i + U_i \quad (4.9)$$

Trong đó:

$Y_i$ : là chỉ tiêu hàng năm về quần áo

$X_i$ : là thu nhập

$$D_2 = \begin{cases} 1 & \text{nếu là nữ} \\ 0 & \text{nếu là nam} \end{cases}$$

$$D_3 = \begin{cases} 1 & \text{nếu là sinh viên} \\ 0 & \text{nếu là công nhân viên} \end{cases}$$

Ngụ ý của mô hình trên đây là ảnh hưởng chênh lệch của biến giả giới tính  $D_2$  là hằng số qua 2 tầng lớp sinh viên và công nhân viên và ảnh hưởng chênh lệch  $D_3$  cũng là hằng số qua 2 giới. Điều này có nghĩa là nếu chỉ tiêu trung bình về quần áo của nữ cao hơn của nam thì điều này cũng đúng với cả hai tầng lớp. Tương tự cũng có thể nói rằng chỉ tiêu trung bình về quần áo của sinh viên nhiều hơn của công nhân viên thì điều đó cũng đúng dù nó là nam hay nữ.

Trong nhiều áp dụng thì giả thiết như vậy không duy trì được. Nữ sinh có thể tiêu dùng nhiều quần áo hơn nam công nhân viên. Nói một cách khác có thể có ảnh hưởng tương tác giữa hai biến chất  $D_2$  và  $D_3$  và do đó ảnh hưởng của chúng lên trung bình  $Y$  có thể không phải là phép cộng đơn giản như trên mà là nhân như mô hình sau:

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 (D_{2i} D_{3i}) + \beta_5 X_i + U_i \quad (4.10)$$

Từ (4.10) ta thu được chỉ tiêu trung bình về quần áo của một nữ sinh sẽ là:

$$E(Y | D_2 = 1, D_3 = 1, X_i) = (\beta_1 + \beta_2 + \beta_3 + \beta_4) + \beta_5 X_i$$

Trong đó  $\beta_2$  là ảnh hưởng chênh lệch của nữ,  $\beta_3$  là ảnh hưởng chênh lệch của sinh viên, còn  $\beta_4$  là ảnh hưởng chênh lệch của nữ sinh.  $\beta_4$  cho ta biết chỉ tiêu trung bình về quần áo của một nữ sinh khác với chỉ tiêu trung bình của nam và của nữ công nhân viên. Điều này cho biết sự tương tác giữa các biến giả làm thay đổi ảnh hưởng của 2 thuộc tính đã được xem xét một cách riêng rẽ như thế nào.

Để kiểm định hệ số  $\beta_4$ , sự tương tác của biến giả có ý nghĩa về mặt thống kê hay không, có thể kiểm định t.

#### 4.6. SỬ DỤNG BIẾN GIẢ TRONG PHÂN TÍCH MÙA

Như chúng ta đã biết nhiều chuỗi thời gian trong kinh tế có tính chất vụ rất rõ, chẳng hạn doanh số bán ra của cửa hàng quần áo vào ngày Tết, doanh số bán ra của cửa hàng văn phòng phẩm vào đầu năm học,... Thông thường người ta muốn loại nhân tố mùa khỏi chuỗi thời gian để người ta có thể tập trung vào các thành phần khác của chuỗi thời gian như khuynh hướng tăng hoặc giảm hoàn toàn đều đặn theo một thời kỳ thời gian dài. Quá trình loại thành phần khỏi chuỗi thời gian thu được như vậy gọi là chuỗi thời gian đã được điều chỉnh theo mùa. Có một số phương pháp điều chỉnh theo mùa của chuỗi thời gian, trong mục này ta xét phương pháp biến giả.

Hoàn toàn như đã trình bày ở trên việc đưa biến giả vào để loại yếu tố khỏi chuỗi thời gian được thực hiện, dựa trên các giả thiết:

1. Yếu tố chỉ ảnh hưởng đến hệ số chặn của hồi quy,
- hoặc 2. Yếu tố ảnh hưởng cả đến hệ số góc.

Ứng với mỗi giả thiết, mô hình được xem xét cũng khác nhau. Để thuận tiện trình bày, ta xét:

Giả sử mối liên hệ giữa thu nhập và chỉ tiêu cho quần áo, dụng cụ gia đình, người ta thu nhập được ngẫu nhiên kích thước  $n$  và người ta cho rằng mỗi một quý có thể biểu thị mẫu theo mùa vì thế người ta đề nghị mô hình sau:

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 D_{4i} + \beta_5 X_i + U_i \quad (4.11)$$

Trong đó:

$Y_i$ : biểu thị chỉ tiêu của người tiêu dùng về các loại hàng nói trên

$X_i$ : là thu nhập của người tiêu dùng  $i$

$$D_2 = \begin{cases} 1 & \text{nếu quan sát nằm ở quý 2} \\ 0 & \text{nếu quan sát nằm ở quý khác} \end{cases}$$

$$D_3 = \begin{cases} 1 & \text{nếu quan sát nằm ở quý 2} \\ 0 & \text{nếu quan sát nằm ở quý khác} \end{cases}$$

$$D_4 = \begin{cases} 1 & \text{nếu quan sát nằm ở quý 2} \\ 0 & \text{nếu quan sát nằm ở quý khác} \end{cases}$$

Trong mô hình trên ta giả thiết biến có 4 phạm trù, nên ta dùng 3 biến giả, phạm trù cơ bản là quý I. Như vậy nếu có ảnh hưởng theo mùa của từng quý khác nhau thì hệ số chặn  $\beta_2$ ,  $\beta_3$  và  $\beta_4$  khác nhau có ý nghĩa về mặt thống kê. Mỗi một hệ số chặn cho ta biết chi tiêu trung bình ở mỗi quý khác với quý I như thế nào. Với giả thiết  $E(U_i) = 0$  ta có:

Chi tiêu trung bình về các khoản đã kể trên trong quý I là:

$$E(Y | D_2 = 0, D_3 = 0, D_4 = 0, X_i) = \beta_1 + \beta_5 X_i$$

Chi tiêu trung bình về các khoản đã kể trên trong quý II là:

$$E(Y | D_2 = 1, D_3 = 0, D_4 = 0, X_i) = (\beta_1 + \beta_2) + \beta_5 X_i$$

Chi tiêu trung bình về các khoản đã kể trên trong quý III là:

$$E(Y | D_2 = 0, D_3 = 1, D_4 = 0, X_i) = (\beta_1 + \beta_3) + \beta_5 X_i$$

Chi tiêu trung bình về các khoản đã kể trên trong quý IV là:

$$E(Y | D_2 = 0, D_3 = 0, D_4 = 1, X_i) = (\beta_1 + \beta_4) + \beta_5 X_i$$

Bây giờ ta giả sử rằng có sự ảnh hưởng tương tác giữa mùa và thu nhập lên chi tiêu, nói cách khác là có sự ảnh hưởng lên cả hệ số góc của hồi quy. Sử dụng phương pháp tương tự như ta đã trình bày ở trên ta đi đến mô hình:

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 D_{4i} + \beta_5 X_i + \beta_6 (D_{2i} X_i) + \beta_7 (D_{3i} X_i) + \beta_8 (D_{4i} X_i) + U_i \quad (4.12)$$

Như vậy việc phân tích thời vụ có thể sử dụng 2 mô hình (4.11) và (4.12). Tuy nhiên mô hình (4.12) tổng quát hơn, để tránh sự không thích hợp ta nên sử dụng mô hình (4.12). Qua việc ước lượng hồi quy (4.12) chúng ta có thể biết được hệ số góc nào có ý nghĩa, hệ số góc nào không có ý nghĩa.

## TÓM TẮT NỘI DUNG CHƯƠNG 4

Để biểu thị có hay không một tính chất nào đó hoặc biểu thị mức độ khác nhau của một tiêu thức thuộc tính nào đó của biến định tính, người ta sử dụng biến giả. Kỹ thuật sử dụng biến giả trong phân tích hồi quy là gán cho các biến định tính giá trị 0 hoặc 1 để biểu thị mức độ khác nhau của tiêu thức nghiên cứu. Để phân biệt 2 mức độ (2 phạm trù) người ta dùng một biến giả, để phân biệt 3 mức độ người ta dùng hai biến giả. Một cách tổng quát, để phân biệt m mức độ người ta dùng m-1 biến giả. Với quy định: phạm trù được gán giá trị 0 được coi là phạm trù cơ sở. Mọi phạm trù khác được so sánh với phạm trù cơ sở.

Hồi quy biến giả có thể xảy ra các trường hợp sau:

- Hồi quy với một biến định lượng và một biến định tính:
  - Biến định tính có hai phạm trù
  - Biến định tính có nhiều hơn hai phạm trù.
- Hồi quy với một biến định lượng và hai biến định tính

Việc ước lượng các tham số trong mô hình hồi quy được thực hiện bằng phương pháp OLS thông thường.

Hồi quy biến giả có thể sử dụng trong trường hợp cần loại nhân tố mùa vụ ra khỏi chuỗi thời gian. Ngoài ra để kiểm định tính ổn định cấu trúc của các mô hình hồi quy người ta sử dụng kiểm định Chow và rút ngắn kiểm định Chow bằng cách sử dụng biến giả. Ngoài ra kỹ thuật biến giả còn được sử dụng để nghiên cứu ảnh hưởng của tương tác giữa các biến giả.

## CÂU HỎI VÀ BÀI TẬP ÔN CHƯƠNG 4

### I. Câu hỏi lý thuyết:

- Các biến sau đây là biến định lượng hay định tính:
  - GDP
  - Khủng hoảng dầu mỏ năm 1973?
  - Xuất khẩu của Việt nam sang các nước ASIAN?
  - Số sinh viên diện chính sách trong đợt thi tuyển vào trường?
  - Lương bình quân của các nhà quản trị Marketing trong các doanh nghiệp cung cấp dịch vụ BC-VT?
- Trình bày giả thiết và thủ tục kiểm định Chow khi so sánh hai hồi quy?
- Nêu phương pháp so sánh hai hồi quy - thủ tục biến giả?
- Cách sử dụng biến giả trong phân tích mùa?

### II. Bài tập

1. Để nghiên cứu nhu cầu của một loại hàng người ta tiến hành khảo sát giá cả và lượng hàng bán được ở 20 khu vực bán hàng và thu được các số liệu cho trong bảng dưới đây:

$Y_i$	$X_i$	$Z_i$	$Y_i$	$X_i$	$Z_i$
20	2	1	14	5	0
19	3	0	14	6	1
18	3	1	13	6	0
18	4	0	12	7	1
17	4	1	12	7	0
17	3	1	15	5	1
16	4	0	16	4	0
16	4	1	12	7	1
15	5	1	10	8	0
15	5	1	11	8	1

Trong đó:  $Y$  là lượng hàng bán được (tấn/tháng)

$X$  là giá bán (ngàn đồng/Kg)

$$Z = \begin{cases} 0 & \text{Nếu khu vực bán ở nông thôn} \\ 1 & \text{Nếu khu vực bán ở thành phố} \end{cases}$$

a) Tìm các hàm hồi quy:  $\hat{Y}_i = A_0 + A_1 X_i$  (1)

$\hat{Y}_i = B_0 + B_1 X_i + B_2 Z_i$  (2)

b) Cho biết ý nghĩa của các hệ số hồi quy  $B_1$  và  $B_2$ ?

c) Dùng hệ số xác định hồi quy bội điều chỉnh kết hợp với kiểm định giả thiết hệ số hồi quy của biến  $Z$  bằng 0 để kết luận xem có nên đưa biến  $Z$  vào mô hình không?

d) Dùng hàm (1) để dự báo hàng bán được trung bình của một khu vực khi giá bán là 7 ngàn đồng/kg với độ tin cậy 95%?

2. Bảng dưới đây là số liệu giả thiết về thu nhập hàng năm của giáo viên đại học, số năm kinh nghiệm giảng dạy:

Lương khởi điểm (Y) (Ngàn USD)	Số năm kinh nghiệm giảng dạy (X)	Giới tính (1=nam; 0 = nữ)
23,0	1	1
19,5	1	0
24,0	2	1
21,0	2	0
25,0	3	1
22,0	3	0
26,5	4	1
23,1	4	0
25,0	5	0
28,0	5	1
29,5	6	1
26,0	6	0
27,5	7	0
31,5	7	1
29,0	8	0

a) Giới tính có ảnh hưởng đến thu nhập của giáo viên đại học hay không?

b) Dự báo mức thu nhập của một giáo viên nam có số năm kinh nghiệm giảng dạy 8 năm với độ tin cậy 95%?

c) Dự báo mức thu nhập của một giáo viên nữ có số năm kinh nghiệm giảng dạy 9 năm với độ tin cậy 98%?

3. Có số liệu về tiết kiệm và thu nhập cá nhân ở một quốc gia từ năm 1986 đến năm 2004 (Triệu USD) được chia làm hai thời kỳ như bảng sau:

Thời kỳ I	Tiết kiệm	Thu nhập	Thời kỳ II	Tiết kiệm	Thu nhập
1986	0,36	8,8	1995	0,59	15,5
1987	0,21	9,4	1996	0,90	16,7
1988	0,08	10,	1997	0,95	17,7
1989	0,20	10,6	1998	0,82	18,6
1990	0,10	11,0	1999	1,04	19,7
1991	0,12	11,9	2000	1,53	21,1
1992	0,41	12,7	2001	1,94	22,8
1993	0,50	13,5	2002	1,75	23,9
1994	0,43	14,3	2003	1,99	25,2

Hãy sử dụng kiểm định Chow để kiểm định xem hàm tiết kiệm có bị thay đổi cấu trúc giữa hai thời kỳ hay không?

## CHƯƠNG 5: ĐA CỘNG TUYẾN

### GIỚI THIỆU

Trong mô hình phân tích hồi quy bội, chúng ta đã giả thiết giữa các biến giải thích của mô hình không có hiện tượng đa cộng tuyến. Nhưng nếu giả thiết đó bị vi phạm thì hậu quả sẽ như thế nào và là thế nào để phát hiện ra rằng giả thiết này bị vi phạm và biện pháp để khắc phục. Đó là nội dung của chương này. tìm hiểu chương này giúp cho sinh hiểu được bản chất và nguyên nhân của đa cộng tuyến, từ đó có thể phát hiện và loại bỏ các mối quan hệ tuyến tính hoàn hảo hoặc chính xác giữa các biến giải thích. Đảm bảo cho việc lựa chọn các biến giải thích là hoàn toàn độc lập.

### NỘI DUNG

#### 5.1 BẢN CHẤT CỦA ĐA CỘNG TUYẾN:

Trong mô hình hồi quy bội, chúng ta đã giả thiết giữa các biến giải thích không có hiện tượng cộng tuyến. Thuật ngữ đa cộng tuyến có nghĩa là sự tồn tại mối quan hệ tuyến tính “hoàn hảo” hoặc chính xác giữa một số hoặc tất cả các biến giải thích trong mô hình hồi quy. Nghiêm khắc mà nói thì đa cộng tuyến đề cập đến sự tồn tại của nhiều hơn một mối quan hệ tuyến tính chính xác và cộng tuyến là nói đến sự tồn tại duy nhất một mối quan hệ tuyến tính. Nhưng trong thực tế, đa cộng tuyến thường được dùng cho cả hai trường hợp.

Trường hợp lý tưởng là biến  $X_i$  trong môi trường hồi quy không có tương quan với nhau; mỗi một biến  $X_i$  chứa một thông tin riêng về  $Y$ , thông tin này không chứa trong bất kỳ biến  $X_i$  nào khác. Trong thực hành, khi điều này xảy ra ta không gặp hiện tượng đa cộng tuyến.

Trường hợp ngược lại, chúng ta gặp đa cộng tuyến hoàn hảo. giả sử ta phải ước lượng hàm hồi quy  $Y$  gồm  $k$  biến giải thích  $X_1, X_2, \dots, X_k$ ,

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + U_i$$

đa cộng tuyến hoàn hảo xảy ra khi một biến giải thích được biểu diễn dưới dạng tổ hợp tuyến tính của các biến giải thích còn lại đối với mọi điểm của tập hợp số liệu. Hoặc có thể phát biểu: Đa cộng tuyến hoàn hảo giữa các biến giải thích  $X_1, X_2, \dots, X_k$  xảy ra nếu điều kiện sau đây được thoả mãn:

$$\lambda_1 X_1 + \lambda_2 X_2 + \dots + \lambda_k X_k = 0 \quad (5.1)$$

trong đó:  $\lambda_1, \lambda_2, \dots, \lambda_k$  là các hằng số không đồng thời bằng không.

$$\text{Nếu } \lambda_2 X_{2i} + \lambda_3 X_{3i} + \dots + \lambda_k X_{ki} + V_i = 0 \quad (5.2)$$

với  $V_i$  là sai số ngẫu nhiên thì ta có hiện tượng đa cộng tuyến không hoàn hảo giữa các biến giải thích. Nói cách khác, là một biến giải thích nào đó có tương quan chặt chẽ với một số biến giải thích khác.

Thí dụ ta có dữ liệu như sau:

$X_2$	10	15	18	24	30
$X_3$	50	75	90	120	150
$X_3^*$	52	75	97	129	152

Có thể thấy rõ ràng là  $X_{3i} = 5X_{2i}$ , vì vậy có cộng tuyến hoàn hảo giữa  $X_2$  và  $X_3$  và  $r_{23} = 1$ , nhưng giữa  $X_2$  và  $X_3^*$  không có cộng tuyến hoàn hảo, hai biến này có tương quan chặt (cộng tuyến không hoàn hảo), hệ số tương quan giữa chúng là 0.9959.

## 5.2 ƯỚC LƯỢNG KHI CÓ ĐA CỘNG TUYẾN HOÀN HẢO:

Trường hợp đa cộng tuyến hoàn hảo, các hệ số hồi quy không xác định và các sai số chuẩn của chúng là vô hạn. Hiện tượng này có thể giải thích dưới dạng mô hình hồi quy ba biến. Sử dụng dạng độ lệch, trong đó tất cả các biến có thể diễn tả bằng độ lệch của chúng so với giá trị trung bình. Mô hình hồi quy ba biến có thể viết dưới dạng sau:

$$y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + e_i \quad (5.3)$$

$$\text{Trong đó: } y_i = Y_i - \bar{Y} ; \quad x_{2i} = X_{2i} - \bar{X}_2 ; \quad x_{3i} = X_{3i} - \bar{X}_3 \quad (5.4)$$

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i ; \quad \bar{X}_2 = \frac{1}{n} \sum_{i=1}^n X_{2i} ; \quad \bar{X}_3 = \frac{1}{n} \sum_{i=1}^n X_{3i} \quad (5.5)$$

Theo tính toán trong chương hồi quy bội ta thu được các ước lượng:

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \quad (5.6)$$

$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \quad (5.7)$$

Giả sử  $X_{3i} = \lambda X_{2i}$ , trong đó  $\lambda$  là hằng số khác không, thay vào (5.6) ta được:

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\lambda^2 \sum x_{2i}^2) - (\lambda \sum y_i x_{2i})(\lambda \sum x_{2i}^2)}{(\sum x_{2i}^2)(\lambda^2 \sum x_{2i}^2) - \lambda^2 (\sum x_{2i}^2)^2} = \frac{0}{0} \quad (5.8)$$

Đây là biểu thức không xác định. Tương tự như vậy ta cũng có thể chỉ ra  $\hat{\beta}_3$  cũng không xác định.

Vì sao chúng ta lại thu được kết quả như ở (5.8). Nhớ lại ý nghĩa của  $\hat{\beta}_2$ . Như đã biết,  $\hat{\beta}_2$  cho biết mức độ thay đổi về giá trị trung bình của  $Y$  khi  $X_2$  thay đổi 1 đơn vị, với điều kiện  $X_3$  được giữ cố định. Nhưng nếu  $X_2$  và  $X_3$  cộng tuyến hoàn hảo thì không có cách nào để giữ cố định  $X_3$ , điều đó có nghĩa là không thể tách ảnh hưởng của  $X_2$  và  $X_3$  khỏi mẫu đã cho. Trong kinh tế lượng điều này phá hủy toàn bộ ý định tách ảnh hưởng riêng của từng biến lên biến phụ thuộc.

Để thấy được sự khác biệt này, chúng ta hãy thay  $X_{3i} = \lambda X_{2i}$  vào (5.3) ta được:

$$y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 (\lambda x_{2i}) + e_i = (\hat{\beta}_2 + \lambda \hat{\beta}_3) x_{2i} + e_i = \hat{\alpha} x_{2i} + e_i$$

$$\text{Trong đó: } \hat{\alpha} = \hat{\beta}_2 + \lambda \hat{\beta}_3$$

Áp dụng công thức tính ước lượng bằng phương pháp OLS thông thường ta được:

$$\hat{\alpha} = \hat{\beta}_2 + \lambda \hat{\beta}_3 = \frac{\sum x_{2i} y_i}{\sum x_{2i}^2}$$

Như vậy dù  $\alpha$  được ước lượng một cách duy nhất thì cũng không thể xác định được  $\hat{\beta}_2, \hat{\beta}_3$  từ một phương trình hai ẩn.

Như vậy, trong trường hợp đa cộng tuyến hoàn hảo, không thể có lời giải duy nhất cho các hệ số hồi quy riêng. Ta chỉ có thể có được lời giải duy nhất cho tổ hợp tuyến tính các hệ số này. Chú rằng, trong trường hợp đa cộng tuyến hoàn hảo, phương sai và sai số chuẩn của  $\hat{\beta}_2, \hat{\beta}_3$  là vô hạn.

### 5.3 ƯỚC LƯỢNG TRONG TRƯỜNG HỢP CÓ ĐA CỘNG TUYẾN KHÔNG HOÀN HẢO:

Đa cộng tuyến hoàn hảo là trường hợp đặc biệt, rất ít khi xảy ra. Trong các số liệu liên quan đến chuỗi thời gian, thường xảy ra đa cộng tuyến không hoàn hảo.

Xét mô hình hồi quy ba biến dạng độ lệch, ta giả thiết:

$$x_{3i} = \lambda x_{2i} + V_i$$

với  $\lambda \neq 0$  và  $V_i$  là sai số ngẫu nhiên sao cho:  $\sum x_{2i} V_i = 0$

Trong trường hợp này, các hệ số hồi quy  $\beta_2, \beta_3$  có thể ước lượng được. Chẳng hạn thay  $x_{3i} = \lambda x_{2i} + V_i$  vào (5.6) ta có:

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\lambda^2 \sum x_{2i}^2 + \sum V_i^2) - (\lambda \sum y_i x_{2i} + \sum y_i V_i)(\lambda \sum x_{2i}^2)}{(\sum x_{2i}^2)(\lambda^2 \sum x_{2i}^2 + \sum V_i^2) - \lambda^2 (\sum x_{2i}^2)^2} \quad (5.9)$$

Trong trường hợp này không có lý do gì để nói rằng (5.9) là không ước lượng được.

### 5.4 HẬU QUẢ CỦA ĐA CỘNG TUYẾN:

Trong trường hợp có tồn tại đa cộng tuyến gần hoàn hảo (đa cộng tuyến cao) thì có thể gặp một số tình huống sau:

#### 5.4.1 Phương sai và hiệp phương sai của các ước lượng OLS lớn:

Chúng ta xét mô hình (5.3), theo công thức tính phương sai và hiệp phương sai của các ước lượng  $\hat{\beta}_2, \hat{\beta}_3$  ta có:

$$Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)}; \quad (5.11)$$

$$Var(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)} \quad (5.12)$$

$$\text{Và } Cov(\hat{\beta}_2, \hat{\beta}_3) = \frac{-r_{23} \sigma^2}{(1 - r_{23}^2) \sqrt{\sum x_{2i}^2 \sum x_{3i}^2}} \quad (5.13)$$

Trong đó  $r_{23}$  là hệ số tương quan giữa  $X_2$  và  $X_3$ .

Từ (5.11) và (5.12) ta thấy khi  $r_{23}$  tiến đến 1 (nghĩa là sự cộng tuyến gia tăng) thì phương sai của  $\hat{\beta}_2, \hat{\beta}_3$  sẽ tăng dần đến vô hạn. Từ (5.13) ta thấy khi  $r_{23}$  tăng dần đến 1 thì  $Cov(\hat{\beta}_2, \hat{\beta}_3)$  cũng sẽ tăng về giá trị tuyệt đối.

#### 5.4.2 Khoảng tin cậy rộng hơn:

Như ta đã biết trong chương 3, không tin cậy của  $\beta_2, \beta_3$  (với hệ số tin cậy  $1-\alpha$ ) là:

$$\hat{\beta}_2 \pm t_{\alpha/2} se(\hat{\beta}_2); \quad \hat{\beta}_3 \pm t_{\alpha/2} se(\hat{\beta}_3)$$

$$\text{Trong đó: } se(\hat{\beta}_2) = \frac{\delta^2}{\sqrt{(1-r_{23}^2) \sum x_{2i}^2}}; \quad se(\hat{\beta}_3) = \frac{\delta^2}{\sqrt{(1-r_{23}^2) \sum x_{3i}^2}};$$

Như vậy khi  $r_{23}$  càng gần 1 thì khoảng tin cậy của  $\beta_2, \beta_3$  càng rộng. Do đó trường hợp có đa cộng tuyến hoàn hảo thì số liệu của mẫu có thể thích hợp với tập các giả thiết khác nhau. Vì thế xác suất chấp nhận giả thiết sai tăng lên (tức là tăng sai lầm loại II).

#### 5.4.3 Tỷ số t không có ý nghĩa:

Như đã biết, khi kiểm định giả thiết  $H_0: \beta_2 = 0$ , chúng ta đã sử dụng tỷ số  $t = \frac{\hat{\beta}_2}{se(\hat{\beta}_2)}$  và so

sánh với giá trị ước lượng của t với giá trị tới hạn của t. Nhưng khi có đa cộng tuyến gần hoàn hảo thì sai số tiêu chuẩn ước lượng được sẽ rất cao vì vậy sẽ làm cho tỷ số t nhỏ đi. Kết quả là sẽ làm tăng khả năng chấp nhận giả thiết  $H_0$ .

#### 5.4.4 $R^2$ cao nhưng tỷ số t ít ý nghĩa:

Để giải thích điều này, ta hãy xét mô hình hồi quy k biến như sau:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + U_i$$

Trong trường hợp có đa cộng tuyến gần hoàn hảo, như ta đã chỉ ra ở trên, ta có thể tìm được một hoặc một số hệ số góc riêng là không có ý nghĩa về mặt thống kê trên cơ sở kiểm định t. Nhưng trong khi đó  $R^2$  lại có thể rất cao, nên bằng kiểm định F, chúng ta có thể bác bỏ giả thiết  $H_0: \beta_2 = \beta_3 = \dots = \beta_k = 0$ . Mâu thuẫn này cũng là tín hiệu của đa cộng tuyến.

#### 5.4.5 Các ước lượng OLS và các sai số tiêu chuẩn của chúng trở nên rất nhạy đối với những thay đổi nhỏ trong số liệu:

Chỉ cần đa cộng tuyến không hoàn hảo thì việc ước lượng các hệ số hồi quy có thể thực hiện được, nhưng các giá trị ước lượng và sai số chuẩn của chúng trở nên vô cùng nhạy ngay cả với thay đổi nhỏ nhất trong số liệu.

#### 5.4.6 Dấu của các ước lượng của hệ số hồi quy có thể sai:

Khi có đa cộng tuyến gần hoàn hảo thì có thể thu được ước lượng của các hệ số hồi quy trái với điều chúng ta mong đợi. Chẳng hạn lý thuyết kinh tế cho rằng đối với hàng hoá bình thường khi thu nhập tăng, cầu hàng hoá tăng, nghĩa là khi hồi quy thu nhập là một trong các biến giải thích còn cầu hàng hoá là biến được giải thích, nếu xảy ra hiện tượng đa cộng tuyến gần hoàn hảo thì ước lượng của hệ số của biến thu nhập có thể mang dấu âm- mâu thuẫn với lý thuyết kinh tế.

#### 5.4.7 Thêm vào hay bớt đi các biến cộng tuyến với các biến khác, mô hình sẽ thay đổi về độ lớn của các ước lượng hoặch dấu của chúng:

Tóm lại: dấu hiệu chủ yếu của đa cộng tuyến mà ta đã nói ở trên là làm tăng sai số chuẩn. Sai số chuẩn lớn hơn phản ánh sự biến thiên của hệ số hồi quy từ mẫu này đến mẫu khác cao hơn, do đó một sự thay đổi nhỏ trong số liệu hay trong mô hình hồi quy (như thêm vào hay bớt đi một biến) sẽ gây nên sự thay đổi lớn cho các hệ số.

Như vậy chúng ta đã biết một số hậu quả của đa cộng tuyến. Nhưng điều quan trọng là làm thế nào để phát hiện được sự tồn tại của đa cộng tuyến để có biện pháp khắc phục.

#### 5.5 CÁCH PHÁT HIỆN SỰ TỒN TẠI CỦA ĐA CỘNG TUYẾN:

Để phát hiện sự tồn tại của đa cộng tuyến, chúng ta căn cứ vào các dấu hiệu sau đây:

##### 5.5.1 Hệ số $R^2$ lớn nhưng tỷ số t nhỏ.

Trong trường hợp  $R^2$  cao (thường  $R^2 > 0,8$ ) mà tỷ số t thấp như trên đã chú ý đó chính là dấu hiệu của đa cộng tuyến.

##### 5.5.2 Tương quan cặp giữa các biến giải thích cao:

Nếu hệ số tương quan cặp giữa các biến giải thích cao (vượt 0,8) thì có khả năng tồn tại đa cộng tuyến.

Hệ số tương quan này được tính bằng công thức sau:

$$R_{xz} = \frac{\sum (X_i - \bar{X})(Z_i - \bar{Z})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Z_i - \bar{Z})^2}}$$

Trong đó: X và Z là hai biến giải thích nào đó trong mô hình.

Tuy nhiên tiêu chuẩn này thường không chính xác. Có trường hợp tương quan cặp không cao nhưng vẫn có đa cộng tuyến. Ví dụ, ta xét ba biến giải thích  $X_1, X_2, X_3$  như sau:

$$X_1 = (1,1,1,1,1, \quad 0,0,0,0,0 \quad 0,0,0,0,0 \quad 0,0,0,0,0)$$

$$X_2 = (0,0,0,0,0, \quad 1,1,1,1,1, \quad 0,0,0,0,0 \quad 0,0,0,0,0)$$

$$X_3 = (1,1,1,1,1, \quad 1,1,1,1,1, \quad 0,0,0,0,0 \quad 0,0,0,0,0)$$

Rõ ràng  $X_1 = X_2 + X_3$ , nghĩa là có đa cộng tuyến hoàn hảo, tuy nhiên tương quan cặp là:

$$r_{12} = -0,3333; \quad r_{13} = r_{23} = 0,5774$$

##### 5.5.3 Sử dụng mô hình hồi quy phụ:

Hồi quy phụ là hồi quy một biến giải thích  $X_i$  nào đó theo các biến còn lại.  $R_i^2$  được tính từ hồi quy này ta ký hiệu là  $R_i^2$ .

Mối liên hệ giữa  $F_i$  và  $R_i^2$ :

$$F_i = \frac{R_i^2 / (k-2)}{(1-R_i^2) / (n-k+1)} \quad (5.14)$$

$F_i$  tuân theo luật phân phối F với  $(k-2)$  và  $(n-k+1)$  bậc tự do. Trong đó n cỡ mẫu, k là số biến giải thích kể cả hệ số chặn trong mô hình.  $R_i^2$  là hệ số xác định trong hồi quy của biến  $X_i$  theo các biến X khác. Nếu  $F_i$  tính được vượt điểm tới hạn  $F_{\alpha}(k-2, n-k+1)$  ở mức ý nghĩa đã cho thì có nghĩa

là  $X_i$  có liên hệ tuyến tính với các biến  $X$  khác; trong trường hợp đó ta giữ lại các biến trong mô hình. Nếu  $F_i$  có ý nghĩa về mặt thống kê chúng ta cũng phải quyết định liệu biến  $X_i$  nào sẽ bị loại khỏi mô hình. Một trở ngại của kỹ thuật hồi quy phụ là gánh nặng tính toán. Nhưng ngày nay nhiều chương trình máy tính đã có thể đảm đương công việc tính toán này.

#### 5.5.4 Sử dụng nhân tử phóng đại phương sai(VIF):

Nhân tử phóng đại phương sai gắn với biến  $X_i$ , ký hiệu là  $VIF(X_i)$ .

Đối với hàm hồi quy có hai biến giải thích  $X_2$  và  $X_3$ , VIF được định nghĩa như sau:

$$VIF = \frac{1}{1 - r_{23}^2} \quad (5.15)$$

VIF cho thấy phương sai của hàm ước lượng tăng nhanh như thế nào khi có đa cộng tuyến. Khi  $r_{23} = 1$  thì VIF tiến đến vô hạn. Nếu không có đa cộng tuyến giữa  $X_2$  và  $X_3$  thì  $VIF = 1$ .

Từ định nghĩa này ta có thể diễn tả (5.11) và (5.12) như sau;

$$Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2} VIF; \quad (5.16)$$

$$Var(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2} VIF \quad (5.17)$$

để có khái niệm về phương sai và hiệp phương sai tăng như thế nào khi  $r_{23}$  tăng, ta hãy xem bảng (5.1) sau đây:

Giá trị của $r_{23}$	VIF	$Var(\hat{\beta}_2)$	$Cov(\hat{\beta}_2, \hat{\beta}_3)$
0,00	1,00	1A	0
0,5	1,33	1,33A	0,67B
0,70	1,96	1,96A	1,37B
0,80	2,78	2,78A	2,22B
0,90	5,76	5,76A	4,73B
0,95	10,26	10,26A	9,74B
0,97	16,92	16,92A	16,41B
0,99	50,25	50,25A	49,75B
0,995	100,00	100A	99,5B
0,999	500,00	500A	499,5B

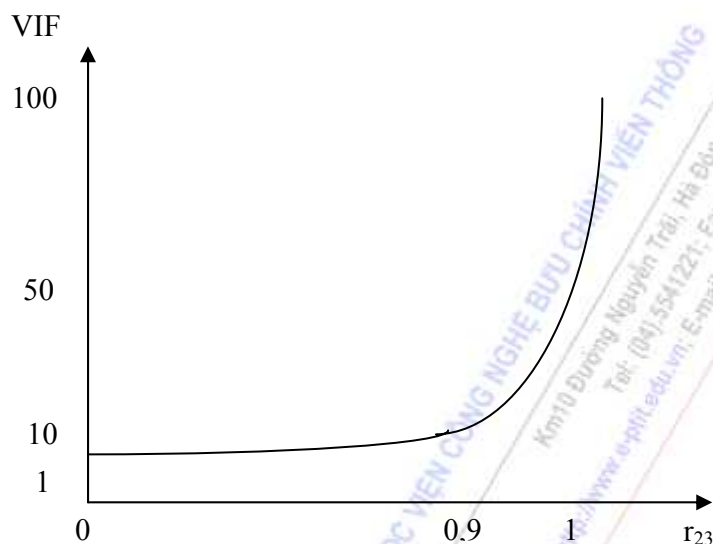
Ghi chú:

$$A = \frac{\sigma^2}{\sum x_{2i}^2}; \quad B = \frac{-\sigma^2}{\sqrt{\sum x_{2i}^2 \sum x_{3i}^2}}$$

Từ kết quả tính toán ở trên, ta thấy gia tăng  $r_{23}$  ảnh hưởng nghiêm trọng đến phương sai và hiệp phương sai [ớc lượng của các hàm ước lượng OLS. Khi  $r_{23} = 0,5$ ,  $var(\hat{\beta}_2) = 1,33$  lần khi  $r_{23} =$

0,00, nhưng khi  $r_{23} = 0,95$  thì  $\text{var}(\hat{\beta}_2)$  lớn gấp 10 lần khi không có đa cộng tuyến. Và khi  $r_{23}$  tăng từ 0,95 đến 0,995 đã làm phương sai ước lượng tăng 100 lần so với không có cộng tuyến. ảnh hưởng nghiêm trọng này cũng thấy ở  $\text{Cov}(\hat{\beta}_2, \hat{\beta}_3)$ .

Đồ thị phản ánh mối quan hệ giữa VIF và  $r_{23}$  như sau:



Hình 5.1

Trên đồ thị ta nhận thấy, khi  $r_{23}$  tăng từ 0,9 đến 1 thì VIF tăng rất nhanh. Khi  $r_{23} = 1$  thì VIF là vô hạn.

Có nhiều chương trình máy tính cho biết giá trị của VIF đối với các biến độc lập của mô hình hồi quy.

## 5.6 BIỆN PHÁP KHẮC PHỤC:

### 5.6.1 Sử dụng thông tin tiên nghiệm:

Một trong các cách tiếp cận để giải quyết vấn đề đa cộng tuyến là phải sử dụng thông tin tiên nghiệm hoặc thông tin từ nguồn khác để ước lượng các hệ số riêng.

Ví dụ: Ta muốn ước lượng hàm sản xuất của một quá trình sản xuất nào đó có dạng:

$$Q_t = AL_t^\alpha K_t^\beta e^{U_i} \quad (5.18)$$

Trong đó  $Q_t$  là lượng sản phẩm được sản xuất thời kỳ  $t$ ;  $L_t$  là số lao động sử dụng ở kỳ  $t$ ;  $K_t$  là vốn thời kỳ  $t$ ;  $U_i$  là sai số ngẫu nhiên;  $A, \alpha, \beta$  là các tham số mà chúng ta cần ước lượng. Lấy ln hai vế (5.18) ta được:

$$\ln Q_t = \ln A + \alpha \ln L_t + \beta \ln K_t + U_i$$

Đặt:  $\ln Q_t = Q_t^*$ ;  $\ln A = A^*$ ;  $\ln L_t = L_t^*$ ;  $\ln K_t = K_t^*$  ta được:

$$Q_t^* = A^* + \alpha L_t^* + \beta K_t^* + U_i \quad (5.19)$$

Giả sử  $K$  và  $L$  có tương quan rất cao, dĩ nhiên điều này dẫn đến phương sai của các ước lượng của các hệ số cơ giản của hàm sản xuất lớn. Giả sử từ một nguồn thông tin nào đó mà ta biết được rằng ngành công nghiệp này thuộc ngành có lợi tức không đổi theo quy mô, nghĩa là

$\alpha + \beta = 1$ . Với thông tin này, cách xử lý của chúng ta sẽ là thay  $\beta = 1 - \alpha$  vào (5.19) và thu được:

$$Q_t^* = A^* + \alpha L_t^* + (1 - \alpha) K_t^* + U_i \quad (5.20)$$

$$\text{Từ đó ta được: } Q_t^* - K_t^* = A^* + \alpha(L_t^* - K_t^*) + U_i \quad (5.21)$$

Đặt:  $Q_t^* - K_t^* = Y_t^*$ ;  $(L_t^* - K_t^*) = Z_t^*$ ; Ta có:

$$Y_t^* = A^* + \alpha Z_t^* + U_i \quad (5.22)$$

Thông tin tiên nghiệm đã giúp ta giảm số biến độc lập xuống còn một biến  $Z_t^*$ .

Sau khi thu được ước lượng  $\hat{\alpha}$  của  $\alpha$  thì  $\hat{\beta}$  tính được từ điều kiện  $\hat{\beta} = 1 - \hat{\alpha}$

### 5.6.2 Thu thập thêm số liệu hoặc lấy thêm mẫu mới:

Vì vấn đề đa cộng tuyến là một đặc tính của mẫu, có thể là trong một mẫu khác, các biến cộng tuyến không nghiêm trọng như trong mẫu đầu tiên. Vì vậy, đôi khi ta chỉ cần tăng cỡ mẫu cũng có thể làm giảm bớt vấn đề cộng tuyến.

Ví dụ: Trong mô hình 3 biến chúng ta đã thấy:

$$Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)}$$

Khi cỡ mẫu tăng,  $\sum x_{2i}^2$  nói chung sẽ tăng, vì vậy, đối với bất kỳ  $r_{23}$  nào cho trước, phương sai của  $\hat{\beta}_2$  sẽ giảm, kéo theo sai số chuẩn giảm, điều này giúp cho ta ước lượng  $\beta_2$  chính xác hơn.

### 5.6.3 Loại trừ một biến giải thích ra khỏi mô hình:

Bước 1: Xem cặp biến giải thích nào có quan hệ chặt chẽ. Giả sử  $X_2, X_3, \dots, X_k$  là các biến độc lập;  $Y$  là biến phụ thuộc và  $x_2$  và  $X_3$  có tương quan chặt chẽ với nhau.

Bước 2: Tính  $R^2$  đối với các hàm hồi quy: có mặt cả hai biến; không có mặt một trong hai biến.

Bước 3: Ta loại biến mà giá trị  $R^2$  tính được khi không có mặt biến đó là lớn hơn.

Ví dụ:  $R^2$  của hàm có mặt của cả hai biến  $X_2$  và  $X_3$  là 0,94;  $R^2$  của mô hình không có biến  $X_2$  là 0,87;  $R^2$  của mô hình không có biến  $X_3$  là 0,92 thì loại biến  $X_3$  ra khỏi mô hình.

### 5.6.4 Sử dụng sai phân cấp một:

Giả sử chúng ta có số liệu chuỗi thời gian biểu thị mối liên hệ giữa biến  $Y$  và các biến phụ thuộc  $X_2$  và  $X_3$  theo mô hình sau:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + U_t \quad (5.23)$$

Trong đó  $t$  là thời gian. Phương trình trên đúng với  $t$  thì cũng đúng với  $t-1$ , nghĩa là:

$$Y_{t-1} = \beta_1 + \beta_2 X_{2t-1} + \beta_3 X_{3t-1} + U_{t-1} \quad (5.24)$$

Từ (5.23) và (5.24) ta được:

$$Y_t - Y_{t-1} = \beta_2(X_{2t} - X_{2t-1}) + \beta_3(X_{3t} - X_{3t-1}) + U_t - U_{t-1} \quad (5.25)$$

Đặt:  $y_t = Y_t - Y_{t-1}$ ;  $x_{2t} = X_{2t} - X_{2t-1}$ ;  $x_{3t} = X_{3t} - X_{3t-1}$ ;  $V_t = U_t - U_{t-1}$

$$\text{Ta có: } y_t = \beta_2 x_{2t} + \beta_3 x_{3t} + V_t \quad (5.26)$$

Mô hình hồi quy dạng (5.26) thường là giảm tính nghiêm trọng của đa cộng tuyến vì dù  $X_2$  và  $X_3$  có thể tương quan cao nhưng không có lý do tiên nghiệm nào chắc chắn rằng sai phân của chúng cũng tương quan cao.

Tuy nhiên biến đổi sai phân bậc nhất sinh ra một số vấn đề. Chẳng hạn như số hạng sai số  $V_t$  trong (5.26) có thể không thoả mãn giả thiết của mô hình hồi quy tuyến tính cổ điển là các sai số ngẫu nhiên không tương quan.

### 5.6.5 Giảm tương quan trong các hàm hồi quy đa thức:

Nét đặc biệt của hồi quy đa thức là các biến giải thích xuất hiện với lũy thừa khác nhau trong mô hình hồi quy. Trong thực hành, để giảm tương quan trong hồi quy đa thức, người ta thường sử dụng dạng độ lệch (lệch so với giá trị trung bình). Nếu việc sử dụng dạng độ lệch mà vẫn không giảm đa cộng tuyến thì người ta có thể phải xem xét đến kỹ thuật “đa thức trực giao”.

### 5.6.6 Một số biện pháp khác:

Ngoài các biện pháp kể trên, người ta còn sử dụng một số biện pháp khác nữa để giải quyết vấn đề đa cộng tuyến. Như:

- Hồi quy thành phần chính.
- Sử dụng các ước lượng từ bên ngoài.
- Hồi quy ngọn sóng.

Nhưng tất cả các biện pháp đã trình bày ở trên có thể là giải pháp cho vấn đề đa cộng tuyến như thế nào còn phụ thuộc vào bản chất của tập số liệu và tính nghiêm trọng của vấn đề đa cộng tuyến.

## TÓM TẮT NỘI DUNG CHƯƠNG 5

Đa cộng tuyến là sự tồn tại ít nhất một mối quan hệ tuyến tính giữa một biến giải thích nào đó với tất cả hay với một biến khác có mặt trong mô hình. Khi có đa cộng tuyến hoàn hảo, không thể có lời giải duy nhất cho các hệ số hồi quy riêng. Ta chỉ có thể có được lời giải duy nhất cho tổ hợp tuyến tính của các hệ số này. Trường hợp đa cộng tuyến không hoàn hảo thường hay gặp trong thực hành. Đa cộng tuyến gây ra nhiều hậu quả như là tăng sai số chuẩn, dấu của các ước lượng về hệ số hồi quy có thể sai,... Vì vậy cần có các biện pháp để phát hiện và khắc phục hiện tượng đa cộng tuyến, phần này giới thiệu 4 cách để phát hiện đa cộng tuyến, tùy từng trường hợp cụ thể có thể vận dụng một trong bốn cách đó. Có 5 biện pháp cơ bản để khắc phục hiện tượng đa cộng tuyến.

## CÂU HỎI VÀ BÀI TẬP ÔN CHƯƠNG 5

### I. Câu hỏi:

1. Giải thích hiện tượng đa cộng tuyến? Trong hồi quy bội có những loại đa cộng tuyến nào?
2. Trường hợp đa cộng tuyến hoàn hảo và không hoàn hảo thì các ước lượng về hệ số hồi quy có gì khác nhau?
3. Trình bày hậu quả của đa cộng tuyến?

4. Để phát hiện đa cộng tuyến thường dùng những phương pháp nào?
5. Trình bày các biện pháp khắc phục đa cộng tuyến?

## II. Bài tập:

1. Xét một tập hợp số liệu lý thuyết cho ở bảng dưới đây:

Y	-10	-8	-6	-4	-2	0	2	4	6	8	10
X <sub>2</sub>	1	2	3	4	5	6	7	8	9	10	11
X <sub>3</sub>	1	3	5	7	9	11	13	15	17	19	21

Giả sử bạn môn áp dụng mô hình sau cho các số liệu ở bảng trên:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + U_i$$

- a) Bạn có thể ước lượng 3 tham số chưa biết hay không? Tại sao?
- b) Nếu không, hàm tuyến tính nào bạn có thể ước lượng?

2. Cho các biến số: C là tiêu dùng; Y là thu nhập sau thuế; L là tài sản dễ chuyển thành tiền. dựa trên 38 quan sát, ta ước lượng được các mô hình sau:

$$C = -7,160 + 0,95213Y + e \quad R^2 = 0,9933 \quad (1)$$

(t) (-1,93) (73,25)

$$C = -10,627 + 0,168166Y + 0,37252L + e \quad R^2 = 0,9953 \quad (2)$$

(t) (-3,25) (9,60) (3,96)

$$L = 9,307 + 0,76207Y + e \quad R^2 = 0,9758 \quad (3)$$

(t) (1,8) (37,2)

- a) Xét xem giữa Y và L có tương quan với nhau hay không?
- b) Có thể bỏ biến L trong mô hình đi được không?
- c) Nếu loại bỏ đi hai quan sát cuối cùng, ước lượng lại mô hình ta có các kết quả sau:

$$C = -6,980 + 0,95145Y + e \quad R^2 = 0,9925 \quad (4)$$

(t) (-1,74) (67,04)

$$C = -13,391 + 0,63258Y + 0,45065L + e \quad R^2 = 0,9951 \quad (5)$$

(t) (-3,71) (8,12) (4,24)

$$L = 9,307 + 0,76207Y + e \quad R^2 = 0,9758 \quad (6)$$

(t) (2,69) (37,80)

Nhận xét về mức độ đa cộng tuyến?

## CHƯƠNG 6: PHƯƠNG SAI CỦA SAI SỐ THAY ĐỔI

### GIỚI THIỆU

Một trong những giả thiết quan trọng của mô hình hồi quy tuyến tính cổ điển là các nhiễu ngẫu nhiên  $U_i$  trong hàm hồi quy tổng thể là có phương sai không đổi và bằng  $\sigma_i^2$ ? Nhưng trong thực tế, nếu giả thiết này bị vi phạm, tức là phương sai của  $U_i$  là  $\sigma_i^2$  (phụ thuộc vào từng quan sát), thì điều gì sẽ xảy ra? Làm thế nào để biết được rằng giả thiết này bị vi phạm? Cách khắc phục như thế nào? Đó là một loạt các câu hỏi mà chúng ta sẽ cho các câu trả lời trong chương này.

### NỘI DUNG

#### 6.1. NGUYÊN NHÂN CỦA PHƯƠNG SAI CỦA SAI SỐ THAY ĐỔI

##### 6.1.1. Phương sai của sai số thay đổi là gì?

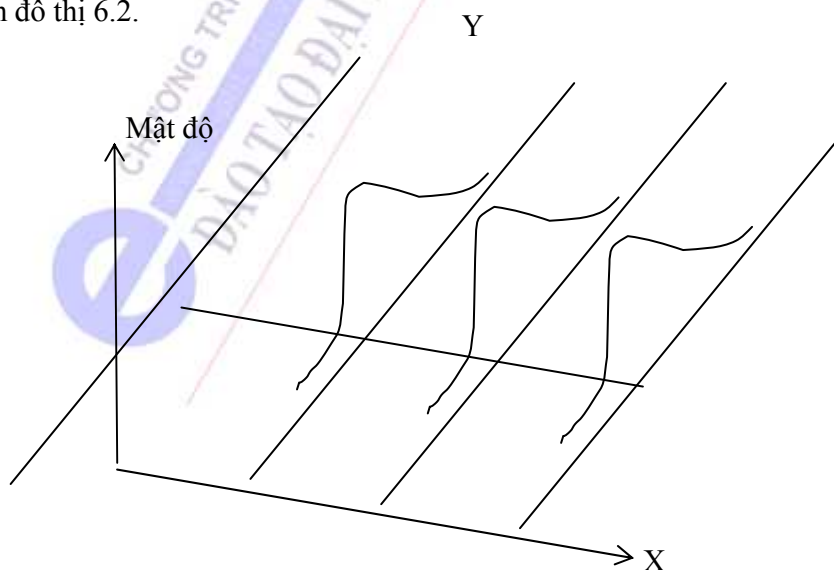
Khi nghiên cứu mô hình hồi quy tuyến tính cổ điển, chúng ta đã đưa ra giả thiết rằng: phương sai của mỗi một nhiễu ngẫu nhiên  $U_i$  trong điều kiện giá trị đã cho của biến giải thích  $X_i$  là không đổi, nghĩa là

$$\text{Var}(U_i | X_i) = E[U_i - E(U_i)]^2 = E(U_i)^2 = \sigma^2 \quad (6.1)$$

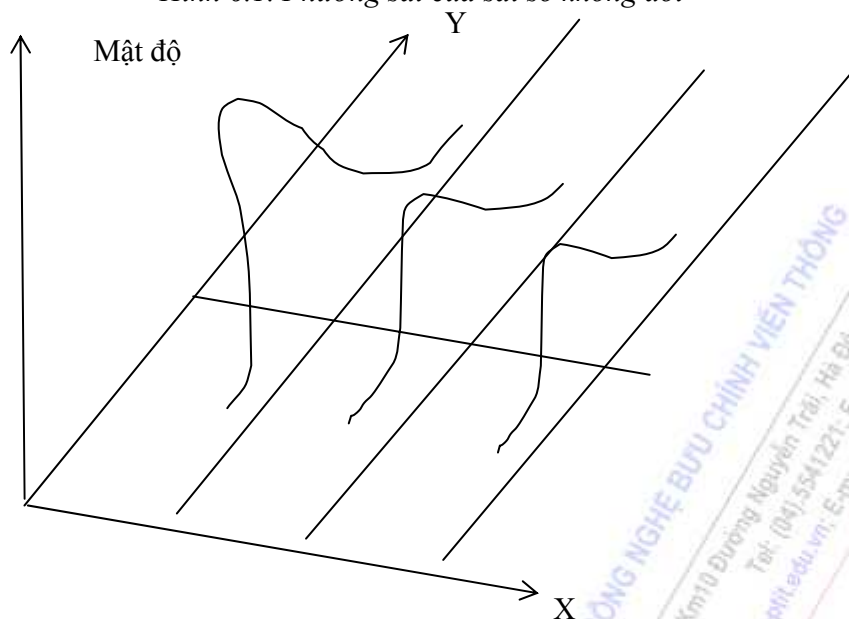
$i = 1, 2, \dots, n$

Về mặt đồ thị thì mô hình hồi quy 2 biến có phương sai không đổi được minh họa ở hình 6.1.

Ngược với trường hợp trên là trường hợp: phương sai có điều kiện của  $Y_i$  thay đổi khi  $X_i$  thay đổi, nghĩa là:  $E(U_i)^2 = \sigma_i^2$  (trong đó các  $\sigma_i^2$  khác nhau). Thí dụ khi nghiên cứu mối quan hệ giữa lỗi mắc phải do đánh máy trong một thời kỳ đã cho với số giờ thực hành, thì người ta nhận thấy số giờ thực hành đánh máy càng tăng thì lỗi sai trung bình mắc phải càng giảm. Điều này mô tả bằng hình đồ thị 6.2.



Hình 6.1. Phương sai của sai số không đổi



Hình 6.2. Phương sai của sai số thay đổi.

### 6.1.2. Nguyên nhân của phương sai của sai số thay đổi

Phương sai thay đổi có thể do một trong các nguyên nhân sau:

- Do bản chất của các mối liên hệ kinh tế: Có nhiều mối quan hệ kinh tế chứa đựng hiện tượng này. Chẳng hạn mối quan hệ giữa thu nhập và tiết kiệm, thông thường thu nhập tăng thì mức độ biến động của tiết kiệm cũng tăng.
- Do kỹ thuật thu thập số liệu được cải tiến,  $\sigma^2$  dường như giảm. Kỹ thuật thu thập số liệu càng được cải tiến, sai lầm phạm phải càng ít hơn.
- Do con người học được hành vi trong quá khứ. Chẳng hạn, lỗi của người đánh máy càng ít nếu thời gian thực hành càng tăng...
- Phương sai của sai số thay đổi cũng xuất hiện khi có các quan trắc ngoại lai. Quan sát ngoại lai là các quan sát khác biệt rất nhiều (quá nhỏ hoặc quá lớn) với các quan sát khác trong mẫu. Việc đưa vào hay loại bỏ các quan sát này ảnh hưởng rất lớn đến phân tích hồi quy.
- Một nguyên nhân khác là mô hình định dạng sai. Có thể do bỏ sót biến thích hợp hoặc dạng giải tích của hàm là sai.

### 6.2. ƯỚC LƯỢNG BÌNH PHƯƠNG NHỎ NHẤT KHI PHƯƠNG SAI CỦA SAI SỐ THAY ĐỔI.

Trong mục này chúng ta hãy xem điều gì sẽ xảy ra đối với các ước lượng bình phương bé nhất và phương sai của chúng nếu phương sai của sai số thay đổi nhưng vẫn giữ nguyên các giả thiết khác của mô hình hồi quy tuyến tính cổ điển? Để trả lời cho vấn đề đó ta xét mô hình hai biến sau:

$$Y_i = \beta_1 + \beta_2 X_i + U_i \quad (6.2)$$

Áp dụng công thức thông thường của phương pháp bình phương nhỏ nhất đã cho ở chương trước để tính  $\hat{\beta}_2$  ta được:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{n \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n \sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \quad (6.3)$$

Còn phương sai là

$$\begin{aligned} \text{Var}(\hat{\beta}_2) &= E(\hat{\beta}_2 - \beta_2)^2 \\ &= E\left(\sum_{i=1}^n k_i U_i\right)^2 \\ &= E(k_1^2 U_1^2 + k_2^2 U_2^2 + k_n^2 U_n^2 + 2k_1 k_2 U_1 U_2 + \dots + 2k_{n-1} k_n U_{n-1} U_n) \\ &= E(k_1^2 U_1^2 + k_2^2 U_2^2 + \dots + k_n^2 U_n^2) \\ &\quad (\text{do giả thiết không tương quan}) \\ &= k_1^2 E(U_1^2) + k_2^2 E(U_2^2) + k_n^2 E(U_n^2) \\ &= \sum_{i=1}^n k_i^2 \sigma_i^2 = \sum_{i=1}^n \left[ \left( \frac{x_i}{\sum_{i=1}^n x_i} \right)^2 \right] \sigma_i^2 \end{aligned} \quad (6.4)$$

$$\text{trong đó } k_i = \frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{x_i}{\sum_{i=1}^n x_i^2}$$

Nhưng phương sai của  $\hat{\beta}_2$  trong trường hợp  $E(U_i)^2 = \sigma^2$  là:

$$\text{Var}(\hat{\beta}_2) = \sigma^2 / \sum_{i=1}^n (X_i - \bar{X})^2 \quad (6.5)$$

Đĩ nhiên nếu  $\sigma_i^2 = \sigma^2$  thì (6.4) và (6.5) trùng nhau.

Như ta đã biết  $\hat{\beta}_2$  là ước lượng tuyến tính không chệch tốt nhất của  $\beta_2$  nếu các giả thiết của mô hình hồi quy tuyến tính cổ điển được thoả mãn. Nhưng liệu nó có còn là ước lượng tuyến tính không chệch tốt nhất trong trường hợp giả thiết phương sai của sai số không thay đổi không được thoả mãn nữa không?

Để chứng tỏ rằng  $\hat{\beta}_2$  vẫn là ước lượng tuyến tính không chệch của  $\beta_2$ . Nhưng liệu  $\hat{\beta}_2$  có vẫn là ước lượng hiệu quả nữa không? Liệu phương sai tính được từ (6.4) có phải là phương sai cực tiểu không? Nếu không thì cái gì là ước lượng tuyến tính không chệch tốt nhất trong trường hợp này? Để trả lời câu hỏi đó ta xét mục sau:

### 6.3. PHƯƠNG PHÁP BÌNH PHƯƠNG NHỎ NHẤT TỔNG QUÁT

Để giải đáp cho câu hỏi ở mục trên ta cần phải xét phương pháp bình phương nhỏ nhất tổng quát. Trước khi đi vào nội dung cụ thể, chúng ta trình bày "Phương pháp bình phương nhỏ nhất có trọng số".

#### 6.3.1. Phương pháp bình phương nhỏ nhất có trọng số

Xét mô hình 2 biến

$$Y_i = \beta_1 + \beta_2 X_i + U_i$$

Như ta đã biết phương pháp bình phương nhỏ nhất không có trọng số: cực tiểu tổng bình phương các phần dư:

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \quad (6.6)$$

để thu được các ước lượng.

Còn phương pháp bình phương nhỏ nhất có trọng số cực tiểu tổng bình phương các phần dư có trọng số:

$$\sum_{i=1}^n W_i e_i^2 = \sum_{i=1}^n W_i (Y_i - \beta_1^* - \beta_2^* X_i)^2 \quad (6.7)$$

Trong đó  $\beta_1^*$  và  $\beta_2^*$  là các ước lượng bình phương nhỏ nhất có trọng số, ở đây các trọng số  $W_i$  là như sau:

$$W_i = 1/\sigma_i^2 \quad (\forall i) \quad (\sigma_i^2 > 0)$$

Nghĩa là trọng số tỷ lệ nghịch với phương sai của  $U_i$  với điều kiện  $X_i$  đã cho, trong đó

$$\text{Var}(U_i | X_i) = \text{Var}(Y_i | X_i) = \sigma_i^2$$

Vì phân cả 2 vế của phương trình (6.7) theo  $\beta_1^*$  và  $\beta_2^*$  ta được:

$$\frac{\partial \sum_{i=1}^n W_i e_i^2}{\partial \beta_1^*} = 2 \sum_{i=1}^n W_i (Y_i - \beta_1^* - \beta_2^* X_i)(-1) \quad (6.8)$$

$$\frac{\partial \sum_{i=1}^n W_i e_i^2}{\partial \beta_2^*} = 2 \sum_{i=1}^n W_i (Y_i - \beta_1^* - \beta_2^* X_i)(-X_i)$$

Cho các đạo hàm riêng bằng không ta thu được hệ phương trình chuẩn:

$$\sum_{i=1}^n W_i Y_i = \beta_1^* \sum_{i=1}^n W_i + \beta_2^* \sum_{i=1}^n W_i X_i \quad (6.9)$$

$$\sum_{i=1}^n W_i X_i Y_i = \beta_1^* \sum_{i=1}^n W_i X_i + \beta_2^* \sum_{i=1}^n W_i X_i^2$$

giải hệ này ta được:

$$\beta_1^* = \bar{Y}^* - \beta_2^* \bar{X}^* \quad (6.10)$$

$$\text{và } \beta_2^* = \frac{(\sum_{i=1}^n W_i)(\sum_{i=1}^n W_i X_i Y_i) - (\sum_{i=1}^n W_i X_i)(\sum_{i=1}^n W_i Y_i)}{(\sum_{i=1}^n W_i)(\sum_{i=1}^n W_i X_i^2) - (\sum_{i=1}^n W_i X_i)^2} \quad (6.11)$$

$$\text{Trong đó } \bar{Y}^* = \sum_{i=1}^n W_i Y_i / \sum_{i=1}^n W_i \text{ và } \bar{X}^* = \sum_{i=1}^n W_i X_i / \sum_{i=1}^n W_i$$

Rõ ràng rằng khi  $W_i = w$  ( $\forall i$ ) thì trung bình có trọng số bằng trung bình thông thường

### 6.3.2. Phương pháp bình phương nhỏ nhất tổng quát

Bây giờ ta quay lại ước lượng bình phương nhỏ nhất của  $\beta_2$  đã cho ở trên là  $\hat{\beta}_2$ .  $\hat{\beta}_2$  vẫn là ước lượng tuyến tính không chệch lệch nhưng không phải là tốt nhất. Vì sao? Nguyên nhân của hiện tượng đó là do một giả thiết của mô hình cổ điển không được thoả mãn đó là giả thiết phương sai của sai số không đổi bị vi phạm.

Vậy làm thế nào để khắc phục tình trạng đó? Để trả lời cụ thể cho câu hỏi này chúng ta phải phân biệt từng trường hợp đã biết hoặc chưa biết phương sai (xem ở mục cuối ở chương này). Ở đây chúng ta chỉ trình bày một phương pháp tổng quát để đưa một mô hình không thoả mãn giả thiết đó, để làm cơ sở cho việc xem xét ảnh hưởng của việc vi phạm giả thiết này.

Xét mô hình 2 biến  $Y_i = \beta_1 + \beta_2 X_i + U_i$ , trong đó tất cả các giả thiết của mô hình hồi quy tuyến tính cổ điển được thoả mãn trừ giả thiết phương sai của sai số không đổi. Phương trình này có thể viết lại dưới dạng

$$Y_i = \beta_1 X_{0i} + \beta_2 X_i + U_i \quad (6.12)$$

Trong đó:  $X_{0i} = 1 (\forall i)$ . Với mỗi  $i$ , chia cả 2 vế của (6.12) cho  $\sigma_i$  ( $\sigma_i > 0$ ) ta được:

$$\frac{Y_i}{\sigma_i} = \beta_1 \frac{X_{0i}}{\sigma_i} + \beta_2 \frac{X_i}{\sigma_i} + \frac{U_i}{\sigma_i} \quad (6.13)$$

Đặt  $\frac{X_{0i}}{\sigma_i} = X_{0i}^*$ ,  $\frac{X_i}{\sigma_i} = X_i^*$ ,  $\frac{U_i}{\sigma_i} = U_i^*$  và ta cũng sử dụng ký hiệu  $\beta_1^*$  và  $\beta_2^*$  của ước

lượng bình phương nhỏ nhất thông thường  $\beta_1$  và  $\beta_2$

Vậy mô hình đã được biến đổi có dạng:

$$Y_1^* = \beta_1^* X_{0i}^* + \beta_2^* X_i^* + U_i^* \quad (6.14)$$

Mục tiêu của việc biến đổi mô hình gốc là gì? Để thấy được điều này chúng ta xét số hạng sai số đã được biến đổi  $U_i^*$ . Ta có

$$\text{Var}(U_i^*) = E(U_i^*)^2 = \frac{1}{\sigma_i^2} E(U_i)^2 = \frac{\sigma_i^2}{\sigma_i^2} = 1 (\forall i)$$

Vậy  $U_i^*$  có phương sai không đổi.

Vì chúng ta vẫn giữ lại tất cả các giả thiết khác của mô hình hồi quy tuyến tính cổ điển. Thêm vào đó  $U_i^*$  thoả mãn cả giả thiết phương sai không đổi. nên nếu chúng ta tiếp tục phương pháp bình phương nhỏ nhất cho mô hình biến đổi (6.14) thì các ước lượng sinh ra từ đó sẽ là các ước lượng tuyến tính không chệch tốt nhất.

Thủ tục biến đổi các biến gốc theo cách đã trình bày ở trên trong đó các biến đã được biến đổi thoả mãn các giả thiết của mô hình cổ điển và sau đó áp dụng phương pháp bình phương nhỏ nhất vào chúng, được gọi là *phương pháp bình phương nhỏ nhất tổng quát*.

Thủ tục ước lượng  $\beta_1^*$  và  $\beta_2^*$  là như sau:

Trước hết ta viết hàm hồi quy mẫu của (6.14) dưới dạng:

$$\frac{Y_i}{\sigma_i} = \hat{\beta}_1^* \left( \frac{X_{0i}}{\sigma_i} \right) + \hat{\beta}_2^* \left( \frac{X_i}{\sigma_i} \right) + \left( \frac{e_i}{\sigma_i} \right)$$

$$\text{Hay } Y_i^* = \hat{\beta}_1^{**} X_{0i}^* + \hat{\beta}_1^* X_i^* + e_i^* \quad (6.15)$$

Để thu được ước lượng bình phương nhỏ nhất tổng quát, ta cực tiểu hàm:

$$\sum_{i=1}^n e_i^{*2} = \sum_{i=1}^n (Y_i^* - \hat{\beta}_1^{**} X_{0i}^* - \hat{\beta}_1^* X_i^*)^2$$

Hay:

$$\sum_{i=1}^n \frac{e_i^2}{\sigma_i^2} = \sum_{i=1}^n \left( \frac{Y_i}{\sigma_i} - \hat{\beta}_1^* \frac{X_{0i}}{\sigma_i} - \hat{\beta}_2^* \frac{X_i}{\sigma_i} \right)^2 = \sum_{i=1}^n \frac{1}{\sigma_i^2} (Y_i - \hat{\beta}_1^* - \hat{\beta}_2^* X_i)^2$$

Đặt  $W_i = \frac{1}{\sigma_i^2}$  ta quay về dạng (6.7), cho nên ta có thể viết ngay được các ước lượng:

$$\hat{\beta}_1^* = \bar{Y}^* - \hat{\beta}_2^* \bar{X}^* \quad (6.16)$$

$$\hat{\beta}_2^* = \frac{\left( \sum_{i=1}^n W_i \right) \left( \sum_{i=1}^n W_i X_i Y_i \right) - \left( \sum_{i=1}^n W_i X_i \right) \left( \sum_{i=1}^n W_i Y_i \right)}{\left( \sum_{i=1}^n W_i \right) \left( \sum_{i=1}^n W_i X_i^2 \right) - \left( \sum_{i=1}^n W_i X_i \right)^2} \quad (6.17)$$

$$\text{Var}(\hat{\beta}_2^*) = \frac{\sum_{i=1}^n W_i}{\left( \sum_{i=1}^n W_i \right) \left( \sum_{i=1}^n W_i X_i^2 \right) - \left( \sum_{i=1}^n W_i X_i \right)^2} \quad (6.18)$$

#### 6.4. HẬU QUẢ CỦA PHƯƠNG SAI CỦA SAI SỐ THAY ĐỔI

Mục này ta sẽ xét xem hậu quả của giả thiết phương sai của sai số không đổi không được thoả mãn có ảnh hưởng như thế nào đến các ước lượng thu được.

Chúng ta sẽ chỉ ra rằng:

- Các ước lượng bình phương nhỏ nhất vẫn là không chệch nhưng không hiệu quả.
- Ước lượng của các phương sai sẽ bị chệch, như vậy làm mất hiệu lực khi kiểm định.

Vì quan tâm của chúng ta chủ yếu là hệ số góc  $\beta_2$  cho nên để đơn giản ta xét mô hình không có hệ số chặn sau:

$$Y_i = \beta_i X_i + U_i \quad (6.19)$$

Trong đó  $U_i$  là nhiễu là nhiễu ngẫu nhiên thoả mãn các điều kiện:

- $E(U_i) = 0$
- $\text{Cov}(U_i, U_j) = 0$
- $\text{Var}(U_i) = \sigma_i^2$

Theo phương pháp bình phương nhỏ nhất ta được ước lượng bình phương nhỏ nhất của  $\beta_2$  là:

$$\hat{\beta}_2^* = \frac{\left( \sum_{i=1}^n X_i Y_i \right)}{\sum_{i=1}^n X_i^2} = \sum_{i=1}^n k_i Y_i ; \text{ trong đó } k_i = \frac{X_i}{\sum_{i=1}^n X_i^2}$$

Vậy  $\hat{\beta}_2$  vẫn tuyến tính theo  $Y_i$ . Mặt khác từ  $Y_i = \beta_2 X_i + U_i$  ta suy ra:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2} = \frac{\sum_{i=1}^n X_i (\beta_2 X_i + U_i)}{\sum_{i=1}^n X_i^2} = \beta_2 + \frac{\sum_{i=1}^n X_i U_i}{\sum_{i=1}^n X_i^2}$$

Vì  $E(U_i) = 0$  và  $X$  không phải là ngẫu nhiên nên  $E(\hat{\beta}_2) = \beta_2$ , vậy  $\hat{\beta}_2$  là ước lượng không chệch của  $\beta_2$ .

Ta tính được:

$$\text{Var}(\hat{\beta}_2^*) = \frac{\sum_{i=1}^n X_i^2 \sigma_i^2}{\left( \sum_{i=1}^n X_i^2 \right)^2} \quad (6.20)$$

(cách làm tương tự như đã nói ở trên).

Bây giờ chúng ta thực hiện đánh trọng số cho quan sát thứ  $i$  là  $\frac{1}{Z_i}$  trong đó  $Z_i$  thỏa mãn điều kiện  $Z_i^2 = \sigma_i^2 / \sigma^2$  ( $\sigma^2$  là hằng số). (Lưu ý rằng phép biến đổi ở đây tổng quát hơn ở trên một chút vì chỉ cần đặt  $\sigma^2 = 1$  ta được ngay  $Z_i = 1/W_i$ ). Ta sử dụng  $\hat{\beta}^*$  để chỉ ước lượng tham số của mô hình đã biến đổi. Lúc đó (6.19) có thể viết lại là:

$$\frac{Y_i}{Z_i} = \beta_2 \frac{X_i}{Z_i} + \frac{U_i}{Z_i}$$

$$\text{Đặt } V_i = \frac{U_i}{Z_i}, \text{ khi đó}$$

$$E(V_i)^2 = E\left(\frac{U_i}{Z_i}\right)^2 = \frac{E(U_i^2)}{Z_i^2}; \quad E(V_i)^2 = \frac{\sigma_i^2}{Z_i^2} = \sigma^2$$

Hồi quy mẫu của (6.19) có dạng:

$$\frac{Y_i}{Z_i} = \hat{\beta}_2^* \frac{X_i}{Z_i} + V_i$$

Ước lượng bình phương nhỏ nhất của (6.19) như đã biết đó là ước lượng bình phương nhỏ nhất có trọng số và ta ký hiệu là  $\hat{\beta}_2^*$  thì:

$$\hat{\beta}_2^* = \frac{\sum_{i=1}^n (Y_i / Z_i)(X_i / Z_i)}{\sum_{i=1}^n (X_i / Z_i)^2} = \beta_2 + \frac{\sum_{i=1}^n (X_i / Z_i) V_i}{\sum_{i=1}^n (X_i / Z_i)^2} \quad (6.21)$$

Lấy kỳ vọng 2 vế của (6.19) ta có  $E(\hat{\beta}_2^*) = \beta_2$

Như vậy  $\hat{\beta}_2^*$  là ước lượng không chệch của  $\beta_2$ . Ta sẽ chỉ ra rằng  $\hat{\beta}_2^*$  hiệu quả hơn  $\hat{\beta}_2$ .

Chúng ta có

$$\text{Var}(\hat{\beta}_2^*) = \frac{\sigma^2}{\sum_{i=1}^n (X_i / Z_i)^2}$$

Thay  $\sigma_i^2 = \sigma^2 Z_i$  vào (6.18) ta được:

$$\text{Var}(\hat{\beta}_2^*) = \frac{\sigma^2 \sum_{i=1}^n X_i^2 Z_i^2}{\left( \sum_{i=1}^n X_i^2 \right)^2}$$

Lập tỉ số

$$\frac{\text{Var}(\hat{\beta}_2^*)}{\text{Var}(\hat{\beta}_2)} = \frac{(\sum_{i=1}^n X_i^2)^2}{\sum_{i=1}^n (X_i^2 / Z_i^2) \sum_{i=1}^n X_i^2 Z_i^2}$$

Đặt  $a_i = X_i Z_i$ ;  $b_i = X_i / Z_i$  lúc đó:

$$\frac{\text{Var}(\hat{\beta}_2^*)}{\text{Var}(\hat{\beta}_2)} = \frac{(\sum_{i=1}^n a_i b_i)^2}{\sum_{i=1}^n a_i^2 \sum_{i=1}^n b_i^2}$$

Theo bất đẳng thức Bunhiacopski cho  $n$  số tùy ý thì:

$$\sum_{i=1}^n a_i^2 \sum_{i=1}^n b_i^2 \geq \left( \sum_{i=1}^n a_i b_i \right)^2;$$

và dấu bằng xảy ra khi và chỉ khi

$$\frac{a_1}{b_1} = \frac{a_2}{b_2} = \dots = \frac{a_n}{b_n}.$$

Áp dụng vào (6.20) ta được

$$\frac{\text{Var}(\hat{\beta}_2^*)}{\text{Var}(\hat{\beta}_2)} \leq 1$$

Nghĩa là  $\text{Var}(\hat{\beta}_2^*) \leq \text{Var}(\hat{\beta}_2)$ , dấu bằng chỉ xảy ra khi và chỉ khi

$$\frac{a_1}{b_1} = \frac{X_1 Z_1}{X_1 / Z_1} = Z_1^2 = \text{const}$$

nghĩa là  $\sigma_i^2$  không đổi, vậy ước lượng  $\hat{\beta}_2$  không hiệu quả. bây giờ ta quay lại với ước lượng của

phương sai của  $\hat{\beta}_2$  như ta đã biết, nó được ước lượng bởi công thức sau:  $\frac{RSS}{n-1} \frac{1}{\sum_{i=1}^n X_i^2}$ . Trong đó

RSS là tổng bình phương các phần dư thu được từ mô hình ước lượng bình phương nhỏ nhất. Ta tính kỳ vọng của RSS:

$$\begin{aligned} E(RSS) &= E\left[\sum (Y_i - \hat{\beta}_2 X_i)^2\right] \\ &= \sum_{i=1}^n \sigma_i^2 - \frac{\sum_{i=1}^n X_i^2 \sigma_i^2}{\sum_{i=1}^n X_i^2} = \frac{\sum_{i=1}^n \sigma_i^2 \sum_{i=1}^n X_i^2 - \sum_{i=1}^n X_i^2 \sigma_i^2}{\sum_{i=1}^n X_i^2} \end{aligned} \quad (6.22)$$

Lưu ý rằng nếu  $\sigma_i^2 = \sigma^2$  ( $\forall i$ ) thì  $(RSS) = (n - 1)\sigma^2$ . Chúng ta ước lượng phương sai  $\hat{\beta}_2$  mà giá trị kỳ vọng của nó là:

$$E\left(\frac{RSS}{n-1} \frac{1}{\sum X_i^2}\right) = \frac{1}{(n-1)\sum X_i^2} E(RSS) = \frac{\sum \sigma_i^2 \sum X_i^2 - \sum \sigma_i^2 X_i^2}{(n-1)(\sum X_i^2)^2}$$

Trong khi đó phương sai thực là:  $\frac{\sum X_i^2 \sigma_i^2}{(\sum X_i^2)^2}$ .

Như vậy phương sai đã được ước lượng cũng là ước lượng chệch. Bây giờ giả thiết rằng  $\sigma_i^2$  và  $X_i^2$  có tương quan dương (điều này thường xảy ra với các số liệu kinh tế) mà thỏa mãn điều kiện

$$\sum_{i=1}^n X_i^2 \sigma_i^2 > \frac{1}{n} \sum_{i=1}^n X_i^2 \sum_{i=1}^n \sigma_i^2$$

thì giá trị kỳ vọng của phương sai đã được ước lượng nhỏ hơn phương sai thực. Như vậy chúng ta sẽ ước lượng quá thấp phương sai thực của ước lượng bình phương nhỏ nhất và sẽ thu được khoảng tin cậy hẹp hơn khoảng tin cậy thực. Điều này sẽ làm ảnh hưởng kiểm định giả thiết về  $\beta_2$ . Hay nói cách khác là khoảng tin cậy và các kiểm định giả thiết dựa trên phân phối t và F không còn đáng tin cậy nữa. Vì vậy nếu sử dụng thủ tục kiểm định giả thiết thông thường có thể dẫn đến những kết luận sai lầm. Điều này sẽ dẫn đến hậu quả không lường trước được trong thực tiễn. Đó chính là lý do vì sao chúng ta phải nghiên cứu vấn đề này. Nhưng làm thế nào để biết được rằng phương sai của sai số thay đổi hay không?

## 6.5. PHÁT HIỆN PHƯƠNG SAI CỦA SAI SỐ THAY ĐỔI

Như chúng ta đã thấy về mặt lý thuyết thì dễ dàng chỉ ra hậu quả của hiện tượng phương sai của sai số thay đổi, nhưng việc phát hiện ra hiện tượng này trong thực tế thì cũng không phải là vấn đề đơn giản. Vì sao vậy? Bởi vì chúng ta biết được  $\sigma_i^2$  chỉ khi chúng ta có toàn bộ tổng thể tương ứng với những giá trị X được chọn, nhưng điều này hầu như hiếm xảy ra, nghĩa là chúng ta ít khi có được toàn bộ tổng thể để nghiên cứu. Thông thường chúng ta chỉ có được mẫu rút ra được từ tổng thể muốn nghiên cứu mà thôi. Như vậy chúng ta chỉ có những giá trị đơn của Y ứng với những giá trị đã cho của biến X, và ta lại không có cách nào để xác định phương sai  $\sigma_i^2$  từ giá trị đơn của Y. Vậy thì làm thế nào để phát hiện ra phương sai của sai số thay đổi?

Cũng như trong trường hợp "đa cộng tuyến", chúng ta không có một phương pháp chắc chắn để phát hiện ra phương sai của sai số thay đổi. Chúng ta chỉ có vài công cụ để chẩn đoán có thể giúp chúng ta phát hiện ra hiện tượng này. Sau đây chúng ta hãy xét một vài cách chẩn đoán.

### 6.5.1. Bản chất của vấn đề của nghiên cứu

Thông thường bản chất của vấn đề nghiên cứu gợi ý cho chúng ta rằng có thể xảy ra hiện tượng: Phương sai của sai số thay đổi hay không? Trên thực tế thì ở số liệu chéo liên quan đến các đơn vị không thuần nhất hay xảy ra hiện tượng phương sai của sai số thay đổi. Chẳng hạn trong nghiên cứu số liệu chéo của chi phí trung bình của sản xuất tùy theo sản lượng sản phẩm được sản xuất ra, trong mẫu gồm những doanh nghiệp có quy mô khác nhau, người ta thấy rằng dường như phương sai của sai số thay đổi.

### 6.5.2. Xem xét đồ thị của phần dư

Đồ thị của sai số của hồi quy, phần dư đối với giá trị của biến độc lập  $X$  hoặc giá trị dự đoán  $\hat{Y}$  sẽ cho ta biết liệu phương sai của sai số có thay đổi hay không. Phương sai của phần dư được chỉ ra bằng độ rộng của biểu đồ phân rải của phần dư khi  $X$  tăng. Nếu độ rộng của biểu đồ rải của phần dư tăng hoặc giảm khi  $X$  tăng thì giả thiết về phương sai hằng số có thể không được thỏa mãn.

*Thí dụ:* Thí dụ sau đây là biểu hiện quan hệ của chi tiêu cho tiêu dùng ( $Y$ ) và thu nhập ( $X$ ) hàng tháng của 20 hộ gia đình ở một vùng nông thôn.

Bảng 6.1. Chi tiêu cho tiêu dùng ( $Y$ ) và thu nhập ( $X$ ) - ĐV 10.000đ

Gia đình	Chi tiêu Y	Thu nhập X	Gia đình	Chi tiêu Y	Thu nhập X
1	19,9	2,3	11	8,0	8,1
2	31,2	32,3	12	33,1	34,5
3	31,8	33,6	13	33,5	38,0
4	12,1	12,1	14	13,1	14,1
5	40,7	42,3	15	14,8	16,4
6	6,1	6,2	16	21,6	24,1
7	38,6	44,7	17	29,3	30,1
8	25,5	26,1	18	25,0	28,3
9	10,3	10,3	19	17,9	18,2
10	38,8	40,2	20	19,8	20,1

Căn cứ vào số liệu trên đây ta sử dụng phương pháp bình phương nhỏ nhất và ước lượng được hàm:

$$\hat{Y} = 0,847 + 0,899X$$

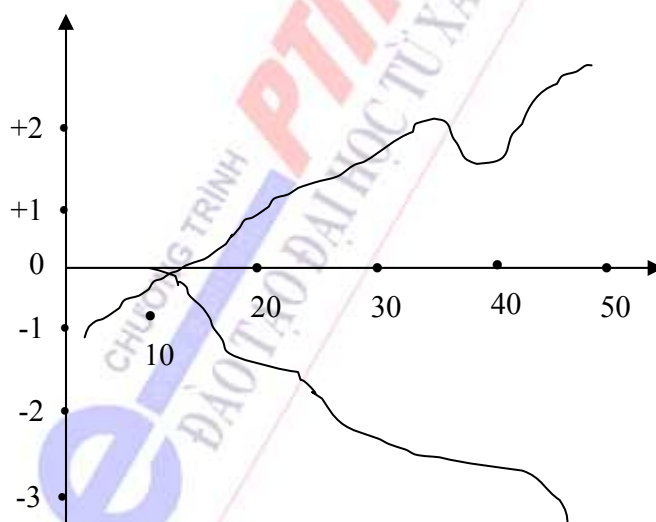
Bảng 6.2 Phần dư đối với hàm tiêu dùng được ước lượng từ tập số liệu đã cho ở bảng 6.1

Quan sát	Giá trị của X	Phần dư	Quan sát	Giá trị của X	Phần dư
6	6,2	-0,32	8	26,1	1,18
11	8,1	-0,13	18	28,3	-1,30
9	10,3	0,19	17	30,1	1,38
4	12,1	0,37	2	32,2	1,30
14	14,1	-0,43	12	34,5	1,23
15	1,4	-0,80	3	36,6	-1,96
19	18,2	0,69	13	38,0	1,52
20	20,1	0,88	10	40,2	1,80
1	22,3	-1,00	5	42,3	1,81
16	24,1	-0,92	7	44,7	-2,45

Các giá trị dự toán và phần dư được tính trong bảng 6.2.

Bảng 6.2 sắp xếp theo thứ tự của giá trị của các quan sát tăng dần từ nhỏ đến lớn và các phần dư tương ứng.

Với giá trị đã cho trong bảng: đồ thị của phần dư đối với X được cho ở hình 6.3



Hình 6.3

Biểu đồ phần dư đối với X cho chúng ta thấy rằng độ rộng của biểu đồ rải tăng lên khi X tăng, cho nên có chứng cứ để cho rằng phương sai của sai số thay đổi khi X tăng. Chú ý rằng đôi khi người ta vẽ đồ thị của phần dư bình phương đối với X.

Nhưng có một vấn đề thực hành mà ta cần bàn tới là nếu chúng ta xem xét hồi quy bội có nhiều hơn một biến giải thích thì chúng ta phải làm thế nào? Liệu có thể dùng đồ thị nữa không?

Một trong các cách có thể làm là vẽ đồ thị của phần dư (hoặc phần dư bình phương). Vì  $\hat{Y}_i$  là tổ hợp tuyến tính của các giá trị của  $X$  nên đồ thị của phần dư bình phương đối với  $\hat{Y}_i$  có thể chỉ ra một mẫu gợi ý cho ta có tồn tại hiện tượng phương sai của sai số thay đổi hay không?

### 6.5.3. Kiểm định PARK

PARK đã hình thức hoá phương pháp đồ thị cho rằng  $\sigma_i^2$  là hàm nào đó của biến giải thích  $X$ . Dạng hàm mà ông đề nghị là:

$$\sigma_i^2 = \sigma^2 X_i^{\beta_2} e^{v_i} \quad (6.23)$$

lấy ln của 2 vế ta được

$$\ln \sigma_i^2 = \ln \sigma^2 + \beta_2 \ln X_i + v_i \quad (6.24)$$

Trong đó  $v_i$  là chưa biết nên PARK đã đề nghị sử dụng  $e_i^2$  thay cho  $\sigma_i^2$  và ước lượng hồi sau:

$$\ln e_i^2 = \ln \sigma^2 + \beta_2 \ln X_i + v_i \quad (6.25)$$

trong đó  $\beta_1 = \ln \sigma^2$ ,  $e_i^2$  thu được từ hồi quy gốc.

Như vậy để thực hiện kiểm định Park ta sẽ tiến hành các bước sau:

1. Ước lượng hồi quy gốc, cho dù có hoặc không tồn tại hiện tượng phương sai của sai số thay đổi.

2. Từ hồi quy gốc thu được các phần dư  $e_i$  sau đó bình phương chúng được  $e_i^2$  rồi đến lấy  $\ln e_i^2$ .

3. Ước lượng hồi quy (6.23) trong đó biến giải thích ( $X_i$ ) là biến giải thích trong hồi quy gốc, nếu có nhiều biến giải thích có thể ước lượng hồi quy đối với mỗi biến giải thích, hoặc có thể ước lượng hồi quy đối với mỗi biến giải thích, hoặc có thể ước lượng hồi quy đối với  $\hat{Y}_i$  làm biến giải thích, trong đó  $\hat{Y}_i$  là  $Y_i$  đã được ước lượng.

4. Kiểm định giả thiết  $H_0: \beta_2 = 0$  có thể bị bác bỏ, trong trường hợp này ta phải tìm cách khắc phục.

5. Nếu giả thiết  $H_0: \beta_2 = 0$  được chấp thuận thì  $\beta_1$  trong hồi quy 9.23 có thể được giải thích như là giá trị của phương sai không đổi ( $\beta_1 = \ln \sigma^2$ )

*Thí dụ:* Căn cứ vào số liệu đã cho ở bảng 6.2

"Phần dư đối với hàm tiêu dùng được ước lượng từ tập số liệu đã cho ở bảng 6.1". Ước lượng hồi quy (6.23) kết quả là như sau:

$$\text{Góc } \ln e^2 = -8,407406 + 2,617445 \ln X_i$$

Bảng 6.3

Biến	Hệ số	Sai lệch chuẩn	t	p
lnX	2,614445	0,218363	11,906	0,000
cons	-8,407406	0,6911656	-12,164	0,000

Nhìn vào kết quả ta thấy có mối liên hệ có ý nghĩa giữa  $\ln e_i^2$  và  $\ln X_i$  nên giả thiết  $H_0: H_0: \beta_2 = 0$  bị bác bỏ ở mức ý nghĩa 5% nghĩa là phương sai của sai số thay đổi, giống như kết luận đã rút ra từ đồ thị phần dư.

#### 6.5.4. Kiểm định Glejser

Kiểm định Glejser cũng tương tự như kiểm định Park. Sau khi thu được phần dư  $e_i$  từ hồi quy theo phương pháp bình phương nhỏ nhất, Glejser đề nghị hồi quy giá trị tuyệt đối của  $e_i$ ,  $|e_i|$  đối với biến  $X$  nào mà có thể có kết hợp chặt chẽ với  $\sigma_i^2$ . Trong thực nghiệm Glejser sử dụng các dạng hàm sau:

$$|e_i| = \beta_1 + \beta_2 X_i + v_i \quad (6.26)$$

$$|e_i| = \beta_1 + \beta_2 \sqrt{X_i} + v_i \quad (6.26')$$

$$|e_i| = \beta_1 + \beta_2 \frac{1}{X_i} + v_i \quad (6.26'')$$

$$|e_i| = \beta_1 + \beta_2 \frac{1}{\sqrt{X_i}} + v_i \quad (6.27)$$

$$|e_i| = \sqrt{\beta_1 + \beta_2 X_i} + v_i \quad (6.28)$$

$$|e_i| = \sqrt{\beta_1 + \beta_2 X_i^2} + v_i \quad (6.28')$$

Trong đó  $v_i$  là sai số.

Giả thiết  $H_0$  trong mỗi trường hợp đã nêu trên là không có phương sai của sai số thay đổi, nghĩa là  $H_0: \beta_2 = 0$ . Nếu giả thiết này bị bác bỏ thì có thể hiện tượng phương sai của sai số thay đổi. Cần lưu ý rằng kiểm định Glejser cũng có vấn đề như kiểm định Park. Glejser và Quandt đã chỉ ra rằng sai số  $v_i$  trong hồi quy của Glejser có một số vấn đề, như giá trị kỳ vọng của nó khác không, nó có tương quan chuỗi. Tuy nhiên Glejser đã cho rằng trong mẫu lớn thì 4 mô hình đầu cho ta kết quả tốt trong việc vạch ra hiện tượng phương sai của sai số thay đổi (2 mô hình cuối cùng còn có vấn đề vì là phi tuyến theo tham số, do đó, không thể ước lượng được bằng thủ tục bình phương nhỏ nhất thông thường).

Do vậy mà kiểm định Glejser được sử dụng như là một công cụ để chẩn đoán trong mẫu lớn.

*Thí dụ:* Sử dụng số liệu đã cho về chi tiêu của tiêu dùng phụ thuộc vào thu nhập ở bảng (6.1) và phần dư tách được ở bảng (6.2) chúng ta tiến hành kiểm định Glejser kết quả như sau:

- Đối với dạng (6.24) ta thu được kết quả sau

$$|e_i| = -0.2093825 + 0.0511835X_i$$

Kết quả chi tiết ở bảng 6.4.

Bảng 6.4

Biến	Hệ số	Sai lệch chuẩn	t	p
X	0,511835	0,0033863	15,115	0,000
cons	-0,209383	0,0941067	-2,225	0,039

- Đối với dạng (6.25) ta được

$$|e_i| = -1.232191 + 0.4782725\sqrt{X_i}$$

Bảng 6.5

Biến	Hệ số	Sai lệch chuẩn	t	p
$\sqrt{X_i}$	0,4725735	0,0369738	12,867	0,000
cons	-1,2232191	0,18561	-6,639	0,000

- Đối với dạng (6.26) ta được

$$|e_i| = -1.826248 - 13.77976 \frac{1}{X_i}$$

Kết quả chi tiết cho ở bảng sau:

Bảng 6.6

Biến	Hệ số	Sai lệch chuẩn	t	p
1/X	13,77976	2,387942	-5,771	0,000
cons	1,826248	0,1549618	11,785	0,000

- Đối với dạng (6.27) ta được:

$$|e_i| = 2.825054 - 7.849832 \frac{1}{\sqrt{X_i}}$$

Bảng 6.7

Biến	Hệ số	Sai lệch số	t	p
$\frac{1}{\sqrt{X_i}}$	-7,848932	1,049445	-7,479	0,000
cons	2,825054	0,2437281	11,591	0,000

Nhìn vào 4 bảng ta đều thấy có mối liên hệ có ý nghĩa  $|\hat{e}_i|$  và biến giải thích  $X_i$  cho nên chúng ta thấy rằng:

Cả 4 kết quả đều cho ta cùng một kết luận là giả thiết  $H_0: \beta_2 = 0$  bị bác bỏ với mức ý nghĩa 5% nghĩa là có hiện tượng phương sai của sai số thay đổi. Các kết luận này cũng giống như kết luận kiểm định Park.

### 6.5.5. Kiểm định tương quan hạng của Spearman.

*Định nghĩa:* Hệ số tương quan hạng Spearman  $r_s$  được xác định như sau:

$$R_s = 1 - 6 \left[ \frac{\sum d_i^2}{n(n-1)} \right] \quad (6.29)$$

Trong đó  $d_i$  = hiệu của các hạng được gán cho 2 đặc trưng khác nhau cùng một phần tử  $i$  và  $n$  = số các phần tử được xếp hạng. Thí dụ cho xếp hạng của 10 học sinh theo kết quả của kỳ thi giữa kỳ và kỳ thi cuối năm là như sau:

Bảng 6.8

<b>Phân tử</b> <b>Hạng</b>	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>	<b>E</b>	<b>F</b>	<b>G</b>	<b>H</b>	<b>I</b>	<b>J</b>
Hạng: giữa kỳ	1	3	7	10	9	5	4	8	2	6
Hạng: cuối kỳ	3	2	8	7	9	6	5	10	1	4

Hệ số tương quan hạng có thể dùng để phát hiện ra phương sai của sai số thay đổi.

Chúng ta xét mô hình

$$Y_i = \beta_0 + \beta_1 X_i + U_i.$$

Thủ tục kiểm định như sau:

*Bước 1:* Ước lượng hồi quy trên tập số liệu đối với  $Y$  và  $X$  thu được phần dư  $e_i$ .

*Bước 2:* Xếp hạng  $|e_i|$  và  $X_i$  theo thứ tự giảm hoặc tăng, tính  $d = \text{hạng } |e_i| - \text{Hạng } X_i$ , sau đó tính hệ số tương quan hạng Spearman.

*Bước 3:* Giả sử hệ số tương quan hạng của tổng thể là  $\rho_1$  bằng 0 và  $n > 8$  thì ý nghĩa của hệ số tương quan hạng mẫu  $r_s$  có thể kiểm định bằng tiêu chuẩn  $t$  sau:

$$t = \frac{r_s \sqrt{n-2}}{\sqrt{1-r_s^2}} \quad (6.30)$$

với bậc tự do  $df = n - 2$

Nếu giá trị  $t$  tính được mà vượt điểm tới hạn  $t$ , chúng ta có thể chấp nhận giả thiết phương sai của sai số thay đổi; ngược lại chúng ta từ bỏ giả thiết về phương sai sai số thay đổi. Nếu mô hình hồi quy có biến giải thích thì hệ số tương quan hạng có thể tính giữa  $|e_i|$  với mỗi một biến  $X$  riêng và có thể kiểm định ý nghĩa thống kê bằng tiêu chuẩn ở trên.

### 6.6. BIỆN PHÁP KHÁC PHỤC

Như chúng ta đã biết phương sai của sai số thay đổi chẳng những phá huỷ còn làm cho các ước lượng đó không còn là ước lượng hiệu quả nữa. Vì thế biện pháp khắc phục là hết sức cần thiết. Việc chữa chạy căn bệnh này phụ thuộc chủ yếu vào liệu  $\sigma_i^2$  được biết hay chưa. Ta phân biệt 2 trường hợp.

### 6.6.1. $\sigma^2_i$ đã biết

Khi  $\sigma^2_i$  đã biết, chúng ta có thể dễ dàng khắc phục căn bệnh đó bằng cách sử dụng phương pháp bình phương nhỏ nhất có trọng số đã trình bày ở trên.

### 6.6.2. $\sigma^2_i$ chưa biết

Trong nghiên cứu kinh tế việc biết trước  $\sigma^2_i$  nói chung là hiếm. Vì vậy nếu chúng ta muốn sử dụng phương pháp bình phương nhỏ nhất có trọng số thì chúng ta cần có những giả thiết nhất định về  $\sigma^2_i$  và biến đổi mô hình hồi quy gốc sao cho mô hình đã được biến đổi này thỏa mãn giả thiết phương sai của sai số không đổi. Phương pháp bình phương nhỏ nhất sẽ được áp dụng cho mô hình đã được biến đổi như đã chỉ ra trước đây, phương pháp bình phương nhỏ nhất áp dụng cho tập số liệu đã được biến đổi.

Chúng ta sẽ minh họa cho các phép biến đổi này qua việc sử dụng mô hình hồi quy 2 biến mà ta gọi là mô hình gốc:

$$Y_i = \beta_1 + \beta_2 X_i + U_i$$

Giả sử mô hình này thỏa mãn các giả thiết của mô hình hồi quy tuyến tính cổ điển trừ giả thiết phương sai của sai số không đổi. Chúng ta xét 1 số giả thiết sau về phương sai của sai số. Những dạng này tuy chưa bao quát được tất cả nhưng phổ biến.

**Giả thiết 1:** Phương sai của sai số tỉ lệ với bình phương của biến giải thích:

$$E(U_i^2) = \sigma^2 X_i^2 \quad (6.31)$$

Nếu bằng phương pháp đồ thị hoặc cách tiếp cận Park hoặc Glejser... chỉ cho chúng ta rằng có thể phương sai  $U_i$  tỉ lệ với bình phương của biến giải thích  $X$  thì chúng ta có thể biến đổi mô hình gốc theo cách sau:

Chia 2 vế của mô hình gốc cho  $X_i (X_i \neq 0)$

$$\frac{Y_i}{X_i} = \frac{\beta_1}{X_i} + \beta_2 + \frac{U_i}{X_i} = \beta_1 \frac{1}{X_i} + \beta_2 + V_i; \quad (6.32)$$

Trong đó  $V_i = \frac{U_i}{X_i}$  là số hạng nhiễu đã được biến đổi, và rõ ràng rằng  $E(V_i)^2 = \sigma^2$ , thực vậy:

$$E(V_i)^2 = E\left(\frac{U_i}{X_i}\right)^2 = \frac{1}{X_i^2} E(U_i)^2 = \frac{\sigma^2 X_i^2}{X_i^2} = \sigma^2$$

Như vậy tất cả các giả thiết của mô hình hồi quy tuyến tính cổ điển được thỏa mãn đối với (6.38) vậy ta có thể áp dụng phương pháp bình phương nhỏ nhất cho phương trình đã được biến đổi (6.38). Hồi quy  $\frac{Y_i}{X_i}$  theo  $\frac{1}{X_i}$ . Chú ý rằng trong hồi quy đã được biến đổi thì số hạng chặn  $\beta_2$

là hệ số góc trong phương trình hồi quy gốc và hệ số góc  $\beta_1$  là số hạng chặn trong mô hình hồi quy gốc. Do đó để trở lại mô hình gốc chúng ta phải nhân cả hai vế của (6.38) đã ước lượng với  $X_i$ .

**Giả thiết 2:** Phương sai của sai số tỉ lệ với biến giải thích  $X$

$$E(U_i)^2 = \sigma^2 X_i$$

Nếu sau khi ước lượng hồi quy bằng phương pháp bình phương nhỏ nhất thông thường, chúng ta vẽ đồ thị của phần dư này đối với biến giải thích  $X$  và quan sát thấy hiện tượng chỉ ra

phương sai của sai số liên hệ tuyến tính với biến giải thích hoặc bằng cách nào đó có thể tin tưởng như vậy thì mô hình gốc sẽ được biến đổi như sau:

Với mỗi  $i$  chia cả 2 vế của mô hình gốc cho  $\sqrt{X_i}$  (Với  $X_i > 0$ )

$$\frac{Y_i}{\sqrt{X_i}} = \frac{\beta_1}{\sqrt{X_i}} + \beta_2 \sqrt{X_i} + \frac{U_i}{\sqrt{X_i}} = \beta_1 \frac{1}{\sqrt{X_i}} + \beta_2 \sqrt{X_i} + V_i \quad (6.33)$$

Trong đó  $v_i = \frac{U_i}{\sqrt{X_i}}$  và có thể thấy ngay rằng  $E(v_i) = 0$ .

**Chú ý:** Mô hình (7.42) là mô hình không có hệ số chặn cho nên ta sẽ sử dụng mô hình hồi quy qua gốc để ước lượng  $\beta_1$  và  $\beta_2$ , sau khi ước lượng (6.42) chúng ta sẽ trở lại mô hình gốc bằng cách nhân cả hai vế (6.42) với  $\sqrt{X_i}$ .

**Giả thiết 3:** Phương sai của sai số tỉ lệ với bình phương của giá trị kỳ vọng của  $Y$ , nghĩa là  $E(U_i^2) = \sigma^2 E(Y_i)^2$ .

Khi đó thực hiện phép đổi biến số như sau:

$$\begin{aligned} \frac{Y_i}{E(Y_i)} &= \frac{\beta_1}{E(Y_i)} + \frac{\beta_2}{E(Y_i)} X_i + \frac{U_i}{E(Y_i)} \\ &= \beta \frac{1}{E(Y_i)} + \beta_2 \frac{1}{E(Y_i)} X_i + V_i \end{aligned} \quad (6.34)$$

Trong đó:  $V_i = \frac{U_i}{E(Y_i)}$ ,  $Var(V_i) = \sigma^2$ .

Nghĩa là nhiều  $V_i$  có phương sai không đổi. Điều này chỉ ra rằng hồi quy (6.44) thỏa mãn giả thiết phương sai không đổi của mô hình hồi quy tuyến tính cổ điển.

Tuy nhiên phép biến đổi (6.44) vẫn chưa thực hiện được vì bản thân  $E(Y_i)$  phụ thuộc vào  $\beta_1$  và  $\beta_2$  trong khi đó  $\beta_1$  và  $\beta_2$  lại chưa biết.

Nhưng chúng ta biết  $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$  là ước lượng của  $E(Y_i)$ . Do đó có thể tiến hành theo 2 bước sau:

**Bước 1:** Ước lượng hồi quy (6.31) bằng phương pháp bình phương bé nhất thông thường, thu được  $\hat{Y}_i$ . Sau đó sử dụng  $\hat{Y}_i$  để biến đổi mô hình gốc thành dạng như sau:

$$\frac{Y_i}{\hat{Y}_i} = \beta_1 \left( \frac{1}{\hat{Y}_i} \right) + \beta_2 \left( \frac{X_i}{\hat{Y}_i} \right) + V_i \quad (6.35)$$

Trong đó  $V_i = \frac{U_i}{\hat{Y}_i}$

**Bước 2:** Ước lượng hồi quy (7.45), dù  $\hat{Y}_i$  không chính xác là  $E(Y_i)$ , chúng chỉ là ước lượng vững nghĩa là khi cỡ mẫu tăng lên vô hạn thì chúng hội tụ đến  $E(Y_i)$  vì vậy phép biến đổi (7.45) có thể sử dụng trong thực hành khi cỡ mẫu tương đối lớn.

**Giả thiết 4:** Phép biến đổi loga

Đôi khi thay cho việc dự đoán về  $\sigma^2_i$  người ta định dạng lại mô hình. Chẳng hạn thay cho việc ước lượng hồi quy gốc có thể chúng ta sẽ ước lượng hồi quy:

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + U_i \quad (6.36)$$

Việc ước lượng hồi quy (7.46) có thể làm giảm phương sai của sai số thay đổi do tác động của phép biến đổi loga. Một trong ưu thế của phép biến đổi loga là hệ số góc  $\beta_2$  là hệ số co giãn của Y đối với X

*Thí dụ:* Với số liệu đã cho ở bảng 7.1 chỉ tiêu cho tiêu dùng Y và thu nhập X. Như ta đã biết ước lượng bình phương nhỏ nhất thông thường của hàm hồi quy:

$$Y_i = \beta_1 + \beta_2 X_i + U_i$$

Kết quả hồi quy của  $\log X_i$  căn cứ vào số liệu ở bảng (6.10) ta được:

$$*\log Y_i = 0,0757 + 0,95621 \log X_i$$

Bảng 6.10. Phần dư từ phương trình log

Quan sát	logX	Phần dư	Quan sát	logX	Phần dư
6	1,82	-0,12	8	3,26	0,44
11	2,09	0,04	18	3,34	-0,53
9	2,33	0,27	17	3,40	0,47
4	2,49	0,34	2	3,48	0,42
14	2,65	-0,33	12	3,54	0,38
15	2,70	-0,56	3	3,60	-0,59
19	2,90	0,35	13	3,64	-0,42
20	3,00	0,41	10	3,69	0,51
1	3,10	-0,54	5	3,74	0,50
16	3,18	-0,46	7	3,80	-0,56

Từ bảng này ta có thể nhận thấy rằng không xảy ra tình trạng khi X tăng phần dư tăng. Không xuất hiện hiện tượng phương sai của sai số thay đổi ở đây. Dĩ nhiên ta có thể sử dụng bất kỳ kiểm định nào đã trình bày ở trên.

Trước khi kết thúc chương trình này chúng tôi xin lưu ý bạn đọc một số vấn đề sau:

Hiện tượng mà chúng ta đang bàn đến là tương đối phổ biến, cho nên biện pháp khắc phục nó rất là quan trọng. Nhưng biện pháp khắc phục thực chất là toa thuốc cho con bệnh, bệnh có chữa được không, không chỉ là toa thuốc có hay hay không, mà trước hết là chẩn đoán đúng bệnh. Vì vậy, cả hai vấn đề chẩn đoán và chữa đều quan trọng.

Vì thế cần phải lưu ý một số vấn đề:

- Khi nghiên cứu mô hình có nhiều biến giải thích thì việc chọn biến nào để biến đổi cần phải xem xét cẩn thận.
- Phép biến đổi log không dùng được khi các giá trị X hoặc Y là âm

• Có thể xảy ra tình trạng là bản thân biến gốc không tương quan nhưng tỷ số của các biến lại có thể tương quan: Chẳng hạn xét mô hình:

$$Y_i = \beta_1 + \beta_2 X_i + U_i$$

Giữ  $Y_i$  và  $X_i$  có thể không tương quan nhưng trong mô hình được biến đổi dưới dạng:

$$\frac{Y_i}{X_i} = \beta_1 \left( \frac{1}{X_i} \right) + \beta_2 \text{ thì } \frac{Y_i}{X_i} \text{ và } \frac{1}{X_i} \text{ Lại là tương quan}$$

• Khi  $\sigma^2_i$  chưa biết nó được ước lượng từ một trong các cách biến đổi trên. Tất cả các kiểm định t, F mà chúng ta sử dụng chỉ có hiệu lực trong những mẫu lớn. Do đó chúng ta phải cẩn thận khi giải thích các kết quả dựa trên các phép biến đổi khác nhau trong các mẫu nhỏ.

## TÓM TẮT NỘI DUNG CHƯƠNG 6

Trong mô hình hồi quy tuyến tính cổ điển, một trong những giả thiết quan trọng là các sai số ngẫu nhiên  $U_i$  trong hàm hồi quy tổng thể có phương sai không đổi. Nhưng trong thực tế do một số nguyên nhân khách quan và chủ quan có thể làm cho giả thiết này bị vi phạm. Nếu giả thiết này bị vi phạm sẽ dẫn đến các hệ số ước lượng vẫn là ước lượng tuyến tính không chệch nhưng không còn là ước lượng hiệu quả nữa (ước lượng có phương sai nhỏ nhất) và ước lượng của các phương sai sẽ bị chệch, do đó các kiểm định mức ý nghĩa, khoảng tin cậy dựa theo phân phối t và F không còn đáng tin cậy nữa. Để phát hiện ra phương sai của sai số thay đổi có thể sử dụng một trong các biện pháp sau:

- Xem xét đồ thị của phần dư  $e_i$  hoặc  $e_i^2$ .
- Kiểm định Park.
- Kiểm định Glejser.
- Kiểm định tương quan hạng của Spearman.

Sau khi kiểm định nếu phát hiện có hiện tượng phương sai của sai số thay đổi thì tiến hành khắc phục theo hai trường hợp sau:

- Trường hợp đã biết  $\delta_i^2$  ta áp dụng phương pháp bình phương nhỏ nhất có trọng số.
- Trường hợp  $\delta_i^2$  chưa biết, ta cũng khắc phục bằng phương pháp bình phương nhỏ nhất có trọng số nhưng cần có những giả thiết nhất định về  $\delta_i^2$  và biến đổi mô hình gốc sao cho mô hình đã được biến đổi thoả mãn giả thiết phương sai không đổi.

## CÂU HỎI VÀ BÀI TẬP ÔN CHƯƠNG 6

### I. Câu hỏi lý thuyết:

1. Giải thích các khái niệm sau:
  - a) Thế nào là hiện tượng phương sai không đồng đều? Nguyên nhân.
  - b) Tính chất của các ước lượng khi phương sai của sai số thay đổi?
2. Phương pháp bình phương nhỏ nhất có trọng số (WLS) và phương pháp bình phương nhỏ nhất tổng quát?
3. Phương pháp phát hiện phương sai thay đổi?

4. Cách khắc phục hiện tượng phương sai của sai số thay đổi?

**II. Bài tập:**

Cho các số liệu về chi tiêu cho tiêu dùng (Y) và thu nhập (X) hàng tháng của 20 hộ gia đình ở một vùng nông thôn (Đơn vị: 10.000 đồng):

Y	X	Y	X	Y	X
39,9	44,6	51	52,2	29,6	32,8
62,4	64,6	20,6	20,6	43,2	48,2
63,6	67,2	77,6	80,4	58,6	60,2
24,2	24,2	16,0	16,2	50,0	56,6
81,4	84,6	66,2	69,0	35,8	36,4
12,2	12,4	67,0	76,0	39,6	40,2
77,2	89,4	26,2	28,2		

1. Dùng phương pháp bình phương có trọng số để ước lượng hàm hồi quy:

$$\frac{Y_i}{\delta_i} = \frac{\beta_1}{\delta_i} + \frac{\beta_2 X_i}{\delta_i} + \frac{U_i}{\delta_i}$$

2. Có xảy ra hiện tượng phương sai thay đổi đối với mô hình hồi quy đang xét hay không? (Dùng đồ thị, kiểm định Park và Glejser).

## CHƯƠNG 7: TỰ TƯƠNG QUAN

### GIỚI THIỆU

Một trong các giả thiết của mô hình hồi quy tuyến tính cổ điển là không có tự tương quan giữa các sai số ngẫu nhiên  $U_i$  trong hàm hồi quy tổng thể. Nhưng trong thực tế liệu hiện tượng đó có thể xảy ra hay không? Nguyên nhân của hiện tượng đó là gì? Nếu có hiện tượng tự tương quan thì liệu có còn áp dụng phương pháp OLS nữa hay không? Làm thế nào để biết rằng hiện tượng tự tương quan xảy ra? Cách khắc phục như thế nào?...Đó là một loạt vấn đề mà chúng ta cần giải quyết trong chương này.

Trong chương này giới thiệu cho người học hiểu về hiện tượng tự tương quan trong mô hình hồi quy tuyến tính. Khi xảy ra hiện tượng tự tương quan thì hậu quả của nó sẽ như thế nào? Cách phát hiện và biện pháp để khắc phục. Vì vậy yêu cầu người học cần nắm vững phương pháp OLS và các giả thiết của nó khi hồi quy tuyến tính. Với một mô hình cụ thể cần phải nhận biết được có hiện tượng tự tương quan hay không? và có cách xử lý phù hợp.

### NỘI DUNG

#### 7.1 BẢN CHẤT VÀ NGUYÊN NHÂN CỦA HIỆN TƯỢNG TỰ TƯƠNG QUAN.

##### 7.1.1 Tự tương quan là gì?

Thuật ngữ tự tương quan có thể hiểu là sự tương quan giữa các thành phần của chuỗi quan sát được sắp xếp theo thứ tự thời gian (trong các số liệu chuỗi thời gian) hoặc không gian (trong số liệu chéo).

Trong mô hình hồi quy tuyến tính cổ điển, ta giả thiết rằng không có tương quan giữa các nhiễu  $U_i$ , nghĩa là:

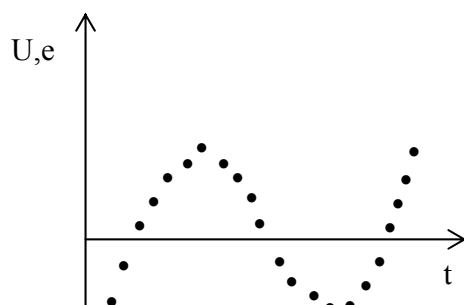
$$\text{Cov}(U_i, U_j) = 0 \quad (i \neq j) \quad (7.1)$$

Nói một cách khác, mô hình cổ điển giả thiết rằng sai số ứng với quan sát nào đó không bị ảnh hưởng bởi sai số ứng với một quan sát khác.

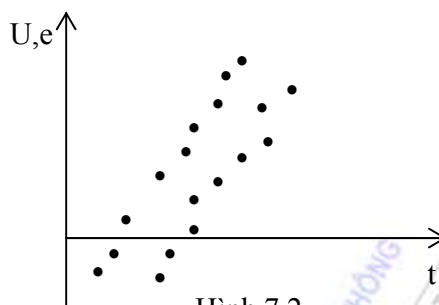
Tuy nhiên, trong thực tế có thể xảy ra hiện tượng mà sai số của các quan sát lại phụ thuộc nhau. Nghĩa là:

$$\text{Cov}(U_i, U_j) \neq 0 \quad (i \neq j) \quad (7.2)$$

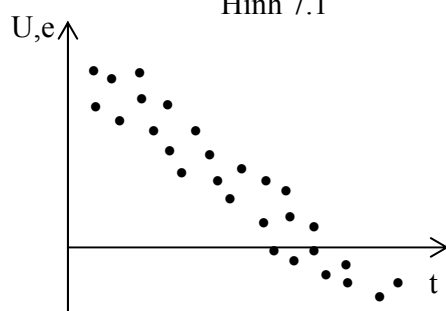
Hãy xét các đồ thị dưới đây với trục tung là  $U_i$  (hoặc  $e_i$ ), trục hoành là thời gian. Trong đó  $U_i$  chỉ nhiễu của tổng thể còn  $e_i$  chỉ là phần dư.



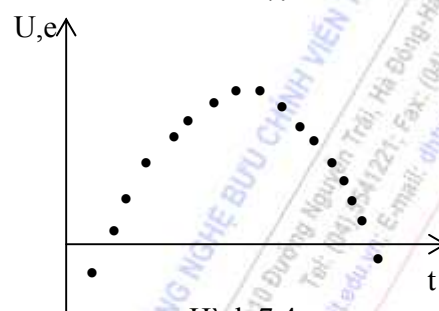
Hình 7.1



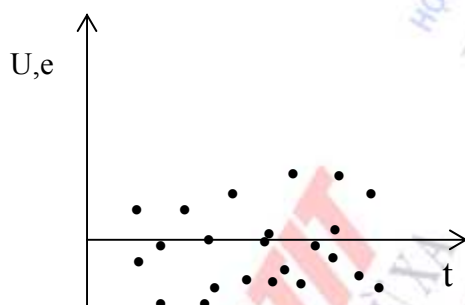
Hình 7.2



Hình 7.3



Hình 7.4



Hình 7.5

Từ hình (8.1) đến (8.4) cho thấy rằng có một dạng phụ thuộc giữa các  $U_i$  (hoặc  $e_i$ ). Hình 7.1 cho thấy dạng chu kỳ; Hình (7.2) và (7.3) cho thấy các xu hướng tuyến tính đi lên hay đi xuống của các sai số; Hình (7.4) cho thấy các sai số có hai dạng: xu hướng tuyến tính và bình phương. Chỉ có hình (7.5) là cho thấy dạng không có hệ thống, ủng hộ cho giả thiết không có tự tương quan trong mô hình hồi quy tuyến tính cổ điển.

### 7.1.2 Nguyên nhân của tự tương quan.

a. Nguyên nhân khách quan:

- Quán tính: Nét nổi bật của hầu hết các chuỗi thời gian trong kinh tế là quán tính. Chúng ta đều biết các chuỗi thời gian như: Tổng sản phẩm, chỉ số giá, thất nghiệp,... mang tính chu kỳ. Chẳng hạn ở giai đoạn đầu của thời kỳ khôi phục kinh tế, tổng sản phẩm có xu hướng đi lên, do đó giá trị của chuỗi ở điểm sau thường cao hơn điểm trước và khi hồi quy chuỗi thời gian, các quan sát kế tiếp có nhiều khả năng phụ thuộc vào nhau.

- Hiện tượng mạng nhện: Người ta thấy rằng việc cung nhiều mặt hàng nông sản biểu hiện hiện tượng “mạng nhện”, trong đó lượng cung phản ứng lại với giá có trễ một khoảng thời gian, vì các quyết định cung cần phải mất một khoảng thời gian để thực hiện, người ta gọi đó là thời kỳ thai nghén.

- Các độ trễ: Trong phân tích chuỗi thời gian, chúng ta có thể gặp hiện tượng biến phụ thuộc ở thời kỳ  $t$  phụ thuộc vào chính biến đó ở thời kỳ  $t-1$  và các biến khác. Chẳng hạn khi nhiên cứu mối quan hệ giữa tiêu dùng và thu nhập, chúng ta thấy rằng tiêu dùng ở thời kỳ hiện tại chẳng những phụ thuộc vào thu nhập mà còn phụ thuộc vào tiêu dùng ở thời kỳ trước đó, nghĩa là:

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 Y_{t-1} + U_t \quad (7.3)$$

trong đó:  $Y_t$ : tiêu dùng thời kỳ  $t$ .

$X_t$ : Thu nhập ở thời kỳ  $t$ .

$Y_{t-1}$ : Tiêu dùng ở thời kỳ  $t-1$ .

$\beta_i$  ( $i=1,2,3$ ) các hệ số.

$U_t$ : Sai số ngẫu nhiên.

Chúng ta có thể lý giải mô hình (7.3) như sau: Người tiêu dùng thường không thay đổi thói quen tiêu dùng,... như vậy nếu chúng ta bỏ qua số hạng trễ trong (7.3) thì sai số sẽ mang tính hệ thống do ảnh hưởng của tiêu dùng ở thời kỳ trước lên tiêu dùng ở thời kỳ hiện tại.

B. Nguyên nhân chủ quan:

- Xử lý số liệu: Trong phân tích thực nghiệm, số liệu thô thường được xử lý. Chẳng hạn trong hồi quy chuỗi thời gian gắn với các số liệu quý, các số liệu này thường được suy ra từ số liệu tháng bằng cách cộng 3 quan sát theo tháng rồi chia cho 3. Việc lấy trung bình làm trơn các số liệu và làm giảm sự dao động trong số liệu tháng. Do vậy đồ thị số liệu quý trơn tru hơn nhiều so với số liệu tháng.

Chính sự làm trơn này có thể dẫn tới sai số hệ thống trong các sai số ngẫu nhiên và gây ra sự tương quan.

Một kiểu xử lý khác là phép nội suy và ngoại suy số liệu. Chẳng hạn tổng điều tra dân số được tiến hành 10 năm 1 lần, lần cuối cùng vào năm 1997. Nếu cần số liệu cho 1 năm, nằm giữa 2 cuộc điều tra, cách phổ biến là nội suy, kỹ thuật có thể gây ra sai số hệ thống mà không có số liệu gốc.

- Sai lệch do lập mô hình: Đây là nguyên nhân thuộc về lập mô hình. Có hai loại sai lầm có thể gây ra hiện tượng tự tương quan.

*Một là:* Không đưa đủ các biến ảnh hưởng cơ bản vào mô hình.

*Thí dụ: xét mô hình:*

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + U_t \quad (7.4)$$

trong đó:  $Y$  là cầu về mặt hàng  $A$

$X_2$ : Giá mặt hàng  $A$ .

$X_3$ : Thu nhập của người tiêu dùng.

$X_4$ : Giá mặt hàng  $B$  có liên quan.

$t$ : là thời gian.

$U_i$ : Sai số ngẫu nhiên.

Nhưng vì lý do nào đó chúng ta đưa vào mô hình chỉ có 2 biến độc lập là  $X_2$  và  $X_3$

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + V_t \quad (7.5)$$

Vậy nếu (7.4) là mô hình đúng thì khi ta tiến hành hồi quy hàm (7.5) cũng tương đương và cho  $V_t = \beta_4 X_{4t} + U_t$ . Nhưng vì việc tăng giá hàng B có ảnh hưởng đến nhu cầu của hàng A nên thành phần nhiễu  $V_t$  sẽ có sai số hệ thống và tạo nên tự tương quan.

Hai là: Dạng hàm sai.

Thí dụ: Giả sử mô hình đúng của chi phí biên và sản lượng là:

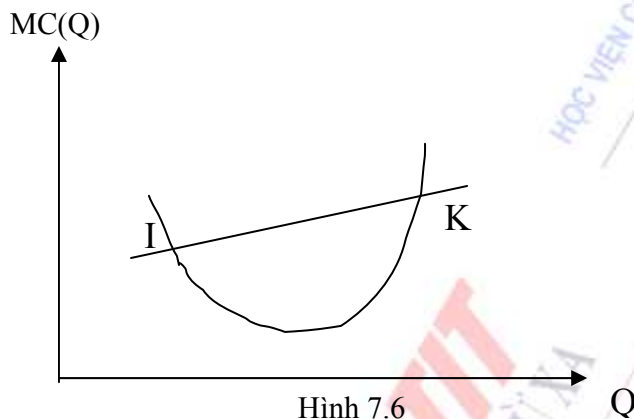
$$(MC)_i = \beta_1 + \beta_2 Q_i + \beta_3 Q_i^2 + U_i \quad (7.6)$$

trong đó: MC là chi phí biên; Q là sản lượng sản phẩm, dịch vụ.

Nhưng ta lại ước lượng mô hình có dạng;

$$(MC)_i = \alpha_1 + \alpha_2 Q_i + V_i \quad (7.7)$$

Đồ thị của (7.6) và (7.7) được biểu diễn ở hình 7.6:



Nhìn vào hình vẽ ta thấy các điểm nằm trên đoạn IK của đường hồi quy (7.7) cho ước lượng quá cao chi phí biên đúng, còn các điểm nằm ngoài đoạn này cho ước lượng thấp hơn. Khi đó các số hạng nhiễu  $V_i$  được xác định như sau:

$$V_i = \beta_3 Q_i^2 + U_i \quad (7.8)$$

Và do đó nó bị ảnh hưởng có tính hệ thống của sản lượng đối với chi phí biên. Vậy  $V_i$  có tự tương quan do sử dụng hàm không chính xác.

## 7.2 ƯỚC LƯỢNG BÌNH PHƯƠNG NHỎ NHẤT KHI CÓ TỰ TƯƠNG QUAN.

Giả sử tất cả các giả thiết đối với mô hình hồi quy tuyến tính cổ điển đều thỏa mãn trừ giả thiết không tương quan giữa các nhiễu  $U_i$ . Trong trường hợp này điều gì sẽ xảy ra đối với các ước lượng OLS và phương sai của chúng?

Để đơn giản ta xét mô hình:

$$Y_t = \beta_1 + \beta_2 X_t + U_t \quad (7.9)$$

trong đó: t là kí hiệu quan sát ở thời điểm t (giả thiết ta đang nghiên cứu số liệu chuỗi thời gian).

Ta giả thiết các nhiễu được tạo ra như sau:

$$U_t = \rho U_{t-1} + \varepsilon_t \quad (-1 < \rho < 1) \quad (7.10)$$

trong đó  $\rho$  được gọi là hệ số tự tương quan;  $\varepsilon_t$  là nhiễu ngẫu nhiên thoả mãn các giả thiết của mô hình hồi quy tuyến tính cổ điển:

$$E(\varepsilon_t) = 0 \quad (\forall t); \text{Cov}(\varepsilon_t, \varepsilon_{t+s}) = 0 \quad (\forall s \neq 0); \text{Var}(\varepsilon_t) = \delta^2.$$

Lược đồ (7.10) được gọi là lược đồ tự hồi quy bậc nhất Markov. Ta ký hiệu lược đồ này là AR(1).

$$\text{Nếu } U_t \text{ có dạng: } U_t = \rho_1 U_{t-1} + \rho_2 U_{t-2} + \varepsilon_t \quad (7.11)$$

thì ta có lược đồ tự hồi quy bậc 2 và ký hiệu là AR(2)

Chú ý rằng hệ số  $\rho$  trong (7.10) có thể giải thích là hệ số tự tương quan bậc nhất hay đúng hơn là hệ số tự tương quan trễ một thời kỳ.

Bây giờ hàm ước lượng OLS của  $\beta_2$ , như thường lệ là:

$$\hat{\beta}_2 = \frac{\sum x_t y_t}{\sum x_t^2} \quad (7.12)$$

Nhưng phương sai của nó trong lược đồ AR(1), bây giờ là:

$$\text{Var}(\hat{\beta}_2)_{AR(1)} = \frac{\delta^2}{\sum_{t=1}^n x_t^2} + \frac{2\delta^2}{\sum_{t=1}^n x_t^2} \left[ \rho \frac{\sum_{t=1}^{n-1} x_t x_{t+1}}{\sum_{t=1}^n x_t^2} + \rho^2 \frac{\sum_{t=1}^{n-2} x_t x_{t+2}}{\sum_{t=1}^n x_t^2} + \dots + \rho^{n-1} \frac{x_1 x_n}{\sum_{t=1}^n x_t^2} \right] \quad (7.13)$$

Nếu không có tự tương quan thì:

$$\text{Var}(\hat{\beta}_2) = \frac{\delta^2}{\sum_{t=1}^n x_t^2} \quad (7.14)$$

Ta thấy (7.13) bằng (7.14) cộng với một số hạng phụ thuộc vào  $\rho$ . Nếu  $\rho = 0$  thì:

$$\text{Var}(\hat{\beta}_2)_{AR(1)} = \text{Var}(\hat{\beta}_2)$$

Nếu tiếp tục dùng phương pháp OLS và điều chỉnh công thức phương sai thông thường bằng việc sử dụng lược đồ AR(1) thì có thể chứng minh được rằng:

- $\hat{\beta}_2$  vẫn là ước lượng tuyến tính không chệch.
- $\hat{\beta}_2$  không còn là ước lượng hiệu quả nữa, do đó nó không còn là ước lượng không chệch tốt nhất.

### 7.3 ƯỚC LƯỢNG TUYẾN TÍNH KHÔNG CHỆCH TỐT NHẤT KHI CÓ TỰ TƯƠNG QUAN.

Giả sử chúng ta xét mô hình hai biến và có quá trình AR(1) bằng phương pháp OLS tổng quát đã xét từ chương trước, ta thu được:

$$\hat{\beta}_2^{OLS} = \frac{\sum_{t=2}^n (x_t - \rho x_{t-1})(y_t - \rho y_{t-1})}{\sum_{t=2}^n (x_t - \rho x_{t-1})^2} + C \quad (7.15)$$

trong đó  $C$  là hiệu số hiệu chỉnh có thể bỏ qua trong thực tế.

Và phương sai của nó được cho bởi:

$$Var(\hat{\beta}_2^{OLS}) = \frac{\delta^2}{\sum_{t=2}^n (x_t - \rho x_{t-1})^2} + D \quad (7.16)$$

trong đó  $D$  cũng là hệ số hiệu chỉnh có thể bỏ qua trong thực tế.

Như ta đã nhận xét trong chương trước, phương pháp OLS tổng quát cho phép chúng ta phối hợp được những thông tin bổ sung vào thủ tục ước lượng, một cách trực tiếp, bằng phép đổi biến. Như vậy, ước lượng OLS tổng quát của  $\hat{\beta}_2$  trong (7.15) phối hợp được tham số tự tương quan  $\rho$  vào công thức ước lượng. đó chính là lý do vì sao ước lượng bình phương nhỏ nhất tổng quát là ước lượng tuyến tính không chệch tốt nhất. Còn ước lượng bình phương nhỏ nhất thông thường thì không. Ước lượng bình phương nhỏ nhất tổng quát tận dụng được nhiều nhất thông tin có được. Khi  $\rho = 0$ , không có thông tin bổ sung cần được xem xét và vì vậy cả hai hàm ước lượng tổng quát và thông thường là như nhau.

## 7.4 HẬU QUẢ CỦA VIỆC SỬ DỤNG PHƯƠNG PHÁP BÌNH PHƯƠNG NHỎ NHẤT THÔNG THƯỜNG KHI CÓ TỰ TƯƠNG QUAN.

Khi có hiện tượng tự tương quan, nếu chúng ta cứ sử dụng phương pháp OLS thì sẽ như thế nào? Cũng như ở chương 6, ta có một số nhận xét sau:

- Các ước lượng OLS vẫn là các ước lượng tuyến tính, không chệch, nhưng chúng không phải là ước lượng hiệu quả nữa. Nói cách khác, ước lượng OLS không phải là ước lượng tuyến tính không chệch tốt nhất nữa.

- Phương sai ước lượng được của các ước lượng OLS thường là chệch. Khi tính phương sai và sai số tiêu chuẩn của các ước lượng OLS thường cho những giá trị thấp hơn các giá trị thực và do đó làm cho giá trị của  $t$  lớn, dẫn đến kết luận sai khi kiểm định. Do đó kiểm định  $t$  và  $F$  không còn tin cậy nữa.

- $\hat{\delta}^2 = \frac{RSS}{df}$  là ước lượng chệch của  $\delta^2$  và trong một số trường hợp là chệch về phía dưới.

- Giá trị ước lượng  $R^2$  có thể không tin cậy khi dùng để thay thế cho giá trị thực của  $R^2$ .

- Phương sai và sai số chuẩn của các giá trị dự báo không được tin cậy (không hiệu quả).

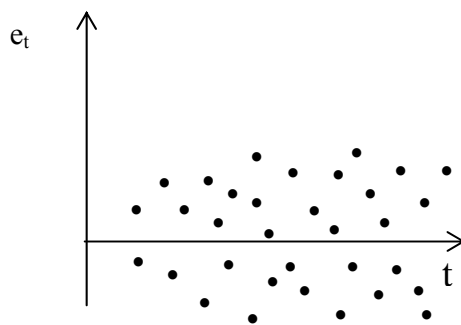
Như vậy, hậu quả của hiện tượng tự tương quan cũng tương tự như hậu quả của hiện tượng phương sai thay đổi là vấn đề nghiêm trọng trong thực hành. Vì vậy, nếu trong số liệu quan sát có hiện tượng tự tương quan thì phải tìm cách phát hiện và khắc phục nó.

## 7.5 PHÁT HIỆN CÓ TỰ TƯƠNG QUAN.

### 7.5.1 Phương pháp đồ thị:

Giả thiết không có tự tương quan trong mô hình hồi quy tuyến tính cổ điển gắn với các nhiễu  $U_t$ , nhưng  $U_t$  không quan sát được, ta chỉ có thể quan sát được các phần dư  $e_t$ . Mặc dù  $e_t$  không hoàn toàn giống  $U_t$  nhưng quan sát các phần dư  $e_t$  có thể gợi ý cho ta những nhận xét về  $U_t$ .

Có nhiều cách khác nhau để xem xét các phần dư. Chẳng hạn chúng ta có thể đơn thuần vẽ đồ thị của  $e_t$  theo thời gian như hình 7.7:



Hình 7.7

Nhìn vào đồ thị, ta thấy phần dư không biểu thị một kiểu mẫu nào khi thời gian tăng lên, nó phân bố một cách ngẫu nhiên xung quanh trung bình của chúng.

Một cách khác là vẽ đồ thị của phần dư chuẩn hoá theo thời gian. Lưu ý rằng, theo giả thiết của mô hình hồi quy tuyến tính cổ điển thì  $U_t \sim N(0, \sigma^2)$ , suy ra:  $\frac{U_t}{\sigma} \sim N(0,1)$ . Vì vậy, nếu chia  $e_t$  cho  $\hat{\sigma}_t$  ta được các phần dư chuẩn hoá. Với mẫu có kích thước khá lớn thì  $\frac{e_t}{\hat{\sigma}_t}$  có phân phối xấp xỉ  $N(0,1)$ . Xem xét đồ thị phần dư chuẩn hoá cũng gợi cho ta ý tưởng về các nhiễu  $U_t$  có phải là ngẫu nhiên không.

Người ta cũng có thể vẽ đồ thị của  $e_t$  đối với  $e_{t-1}$ , một loại kiểm định bằng thực nghiệm lược đồ AR(1).

### 7.5.2 Kiểm định đoạn mạch:

Kiểm định đoạn mạch là một phép kiểm định thống kê giúp ta xác định xem có thể coi một dãy các ký hiệu, các khoản mục, các số liệu có phải là kết quả của một quá trình mang tính ngẫu nhiên hay không.

Ta xét thí dụ về một mô hình chuỗi thời gian, từ số liệu của một mẫu ta đã ước lượng được mô hình và thu được chuỗi các phần dư như sau:

-23	30	12	-10	-5	-17	-22	57	43	-23	31	42
50	61	-28	-52	10	34	28	55	60	32	88	-75
-22	-56	-89	-34	-20	-2	-5	29	12	45	77	78
91	25	60	-25	45	45	30	-59	-60	-40	-75	-25
-34	-66	-90	10	-20							

Nhìn vào dãy các phần dư ta thấy đầu tiên có một phần dư âm, kế tiếp là phần dư dương,...(nếu theo thứ tự từ trái qua phải, từ trên xuống dưới). Ghi lại các dấu + (hoặc dấu -) theo thứ tự từ dãy các phần dư trên, ta có:

-	+	+	-	-	-	-	+	+	-	+	+
+	+	-	-	+	+	+	+	+	+	+	-

.....

Một đoạn mạch là một dãy các phần tử giống nhau mà ở sát trước và sát sau là các phần tử khác chúng hoặc không có phần tử nào. Chiều dài của một đoạn mạch là số phần tử của nó.

Đặt:  $n$  là tổng số quan sát.

$n_1$  là số ký hiệu + (số phần dư dương)

$n_2$  là số ký hiệu âm (số phần dư âm)

$N$  là số đoạn mạch.

Giả thiết kiểm định:  $H_0$ : các phần dư là độc lập.

$H_1$ : các phần dư không độc lập.

Với giả thiết  $n_1 \geq 10$  và  $n_2 \geq 10$  thì  $N$  có phân phối chuẩn với kỳ vọng  $E(N)$  và phương sai  $\delta_N^2$  được tính như sau:

$$E(N) = \frac{2n_1n_2}{n_1 + n_2} + 1 \quad (7.17)$$

$$\delta_N^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)} \quad (7.18)$$

Độ lệch chuẩn:

$$\delta = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}} \quad (7.19)$$

Nếu giả thiết  $H_0$  có thể chấp nhận được thì ta sẽ kỳ vọng số đoạn mạch  $N$  thu được nằm trong khoảng  $[E(N) \pm Z_\alpha \delta_N]$  với hệ số tin cậy  $1-\alpha$ . Với  $Z_\alpha$  là giá trị của  $Z \sim N(0,1)$  thỏa mãn:  $P(|Z| > Z_\alpha) = \alpha$

Vậy quy tắc quyết định như sau:

Chấp nhận giả thiết  $H_0$  (với mức ý nghĩa  $\alpha$ ) nếu  $N \in [E(N) \pm Z_\alpha \delta_N]$  và bác bỏ giả thiết  $H_0$  nếu  $N \notin [E(N) \pm Z_\alpha \delta_N]$ .

Trở lại thí dụ trên, ta tính được  $n_1 = 27$ ;  $n_2 = 26$ .

$$E(N) = \frac{2 * 27 * 26}{27 + 26} + 1 = 27,49$$

$$\delta_N^2 = \frac{2 * 27 * 26(2 * 27 * 26 - 27 - 26)}{(27 + 26)^2(27 + 26 - 1)} = 3,6$$

Với mức ý nghĩa  $\alpha = 5\%$  thì  $Z_{0,05} = 1,96$ , vậy:

$$E(N) - 1,96\delta_N = 27,49 - 1,96 * 3,6 = 20,434$$

$$E(N) + 1,96\delta_N = 27,49 + 1,96 * 3,6 = 34,546$$

Ni N = 15 (20,434; 34,546) nên ta bác bỏ giả thiết  $H_0$  cho rằng phần dư là độc lập, chuỗi thời gian cần phải được điều chỉnh.

### 7.5.3 Kiểm định $\chi^2$ về tính độc lập của các phần dư.

để kiểm định  $\chi^2$  về tính độc lập của các phần dư ta sử dụng bảng tiếp liên sau:

Bảng 7.1

	Phần dư dương tại t	Phần dư âm tại t	Tổng( $R_i$ )
Số phần dư dương tại t-1	$A_{11}$ ( $E_{11}$ )	$A_{12}$ ( $E_{12}$ )	$R_1$
Số phần dư âm tại t-1	$A_{21}$ ( $E_{21}$ )	$A_{22}$ ( $E_{22}$ )	$R_2$
Tổng( $C_j$ )	$C_1$	$C_2$	$n$

Trong đó:  $A_{11}$  là số phần dư dương tại t-1 và t.

$A_{12}$  là số phần dư dương tại t-1 và âm tại t.

$A_{21}$  là số phần dư âm tại t-1 và dương tại t.

$A_{22}$  là số phần dư âm tại t-1 và âm tại t.

$R_1 = A_{11} + A_{12}$ ;  $R_2 = A_{21} + A_{22}$ ;  $C_1 = A_{11} + A_{21}$ ;  $C_2 = A_{12} + A_{22}$ ;

$n$  là tổng số phần dư ở t và t-1.

$E_{ij}$  là tần số lý thuyết ở ô chứa  $A_{ij}$  ( $i, j = 1, 2$ ).

để kiểm định giả thiết về tính độc lập của các phần dư ta có thể tiến hành kiểm định giả thiết  $H_0$ : Các hàng và cột độc lập với nhau; Với giả thiết đối:  $H_1$ : Các hàng và cột không độc lập với nhau.

Để kiểm định giả thiết  $H_0$  nêu trên ta dùng tiêu chuẩn kiểm định  $\chi^2$ :

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(A_{ij} - E_{ij})^2}{E_{ij}} \quad (7.20)$$

Nếu giả thiết  $H_0$  đúng, tức các phần dư có phân bố độc lập thì thống kê  $\chi^2$  đã nói ở trên sẽ có phân phối theo quy luật “khi bình phương” với bậc tự do là:  $df = (2-1)(2-1) = 1$ .

Quy tắc ra quyết định là nếu giá trị của thống kê  $\chi^2$  đã tính được vượt quá giá trị  $\chi^2$  tới hạn với 1 bậc tự do ở mức ý nghĩa cho trước (chẳng hạn  $\alpha = 5\%$ ) thì ta có thể bác bỏ giả thiết  $H_0$  về tính độc lập của các phần dư, ngược lại ta sẽ thừa nhận nó.

Cách tính thống kê  $\chi^2$ ;

Nếu giả thiết  $H_0$  đúng thì các hàng và cột độc lập với nhau và khi đó:  $E_{ij} = nP_{ij}$ , trong đó  $P_{ij}$  là xác suất để phần dư thuộc ô chứa  $A_{ij}$  ở bảng trên.

$$P_{11} = \frac{R_1}{n} * \frac{C_1}{n}; P_{12} = \frac{R_1}{n} * \frac{C_2}{n}; P_{21} = \frac{R_2}{n} * \frac{C_1}{n}; P_{22} = \frac{R_2}{n} * \frac{C_2}{n}$$

Trong đó:

$\frac{R_1}{n}$  là ước lượng xác suất để một phần dư là dương tại t-1.

$\frac{R_2}{n}$  là ước lượng xác suất để một phần dư là âm tại t-1.

$\frac{C_1}{n}$  là ước lượng xác suất để một phần dư là dương tại t.

$\frac{C_2}{n}$  là ước lượng xác suất để một phần dư là âm tại t.

Vậy ta có:

$$E_{ij} = \frac{R_i C_j}{n} \quad (i, j = 1, 2) \quad (7.21)$$

#### 7.5.4 Kiểm định d của Durbin-Watson:

Phương pháp kiểm định có ý nghĩa nhất để phát hiện ra tự tương quan là kiểm định d của Durbin-Watson.

Thống kê d của Durbin-Watson được định nghĩa như sau:

$$d = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2} \quad (7.22)$$

là tỷ số giữa tổng bình phương sai lệch của các phần dư kế tiếp nhau với RSS. Lưu ý trên tử số của thống kê d số quan sát là n-1 vì một quan sát bị mất đi khi lấy ký hiệu các quan sát kế tiếp.

Người ta đã chứng minh khi n đủ lớn thì:

$$d \approx 2(1 - \hat{\rho}) \quad (7.23)$$

Trong đó:

$$\hat{\rho} = \frac{\sum_{i=1}^n e_i e_{i-1}}{\sum_{i=1}^n e_i^2} \quad (7.24)$$

là một ước lượng của hệ số tương quan  $\rho$ . Vì  $-1 \leq \rho \leq 1$  nên ta suy ra:  $0 \leq d \leq 4$

đây là các biên cho d, bất cứ giá trị nào của d ước lượng được phải nằm trong giới hạn này.

Từ phương trình (7.23) ta thấy rằng:  $\hat{\rho} = 0 \rightarrow d \approx 2$ , tức là nếu không tồn tại tương quan chuỗi thì d được kỳ vọng là 2. Do đó theo quy tắc ngón tay cái nếu d gần bằng 2 thì có thể giả định rằng không có tự tương quan bậc nhất. Nếu  $\hat{\rho} = +1$  nghĩa là có tương quan dương hoàn hảo trong phần dư thì  $d \approx 0$ . Do đó d càng gần 0 thì càng chứng tỏ có sự tương quan thuận chiều.

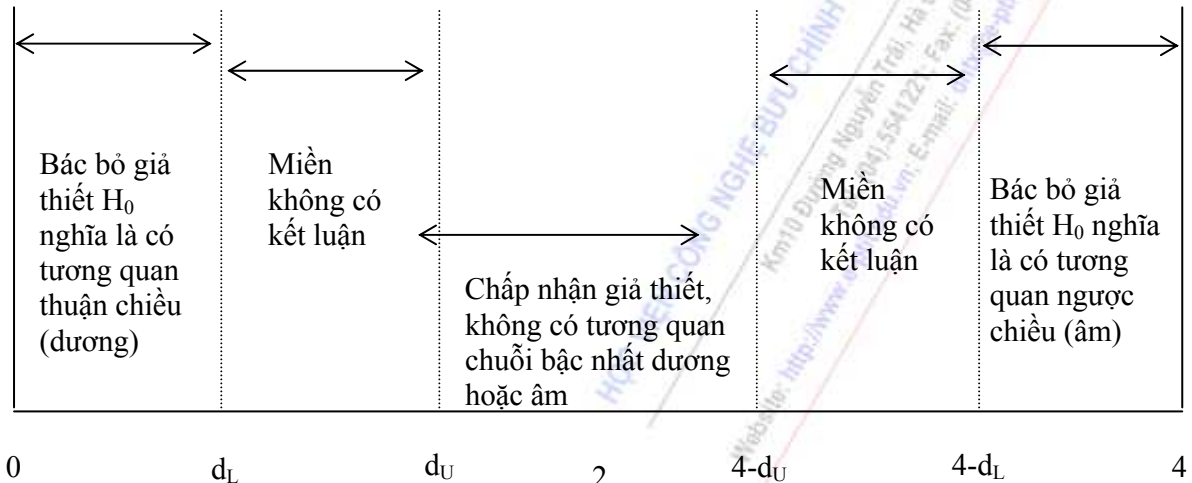
Nếu  $d = -1$  thì có sự tương quan ngược chiều hoàn hảo giữa các phần tử dư kế tiếp nhau và khi đó  $d \approx 4$ . Vì vậy d càng gần 4 thì càng chứng tỏ có sự tương quan chuỗi ngược chiều.

Nếu các giả thiết của kiểm định d thoả mãn thì có thể trình bày quy tắc ra quyết định như sau:

Bảng 7.2 Kiểm định d - Durbin-Watson. Quy tắc ra quyết định

Giả thiết $H_0$	Quyết định	Nếu
Không có tự tương quan dương	Bác bỏ	$0 < d < d_L$
Không có tự tương quan dương	Không quyết định	$d_L \leq d \leq d_U$
Không có tự tương quan âm	Bác bỏ	$4-d_L < d < 4$
Không có tự tương quan âm	Không quyết định	$4-d_U \leq d \leq 4-d_L$
Không có tự tương quan dương hoặc âm	Không bác bỏ	$d_U < d < 4-d_U$

Các quy tắc ra quyết định có thể minh họa ở hình 7.8:



Hình 7.8

- Ước lượng hồi quy bằng phương pháp OLS thông thường và thu được phần dư  $e_t$ .
- Tính giá trị của thống kê  $d$  theo công thức (7.22).
- Với cỡ mẫu đã cho  $n$  và số biến giải thích, tìm các giá trị tới hạn  $d_U$  và  $d_L$  được cho trong bảng giá trị  $d$  (phụ lục).

Theo các quy tắc ra quyết định đã cho trong bảng (7.2)

Nếu giá trị của  $d$  thuộc miền không có quyết định, tức ta không thể kết luận có tương quan hay không. Khi đó ta kết luận như thế nào? để giải quyết vấn đề này đã có một số cải biên kiểm định  $d$ . Dưới đây là quy tắc kiểm định cải biên thường được áp dụng để kiểm tự tương quan bậc nhất.

1.  $H_0: \hat{\rho} = 0$ ;  $H_1: \hat{\rho} > 0$ . Nếu  $d < d_U$  thì bác bỏ  $H_0$  và chấp nhận  $H_1$  (với mức ý nghĩa  $\alpha$ ), nghĩa là có tự tương quan dương.
2.  $H_0: \hat{\rho} = 0$ ;  $H_1: \hat{\rho} < 0$ . Nếu  $(4-d) < d_U$  thì bác bỏ giả thiết  $H_0$ , nghĩa là có tự tương quan âm.
3.  $H_0: \hat{\rho} = 0$ ;  $H_1: \hat{\rho} \neq 0$ . Nếu  $d < d_U$  hoặc  $(4-d) < d_U$  thì bác bỏ giả thiết  $H_0$  chấp nhận  $H_1$  (với mức ý nghĩa  $2\alpha$ ), tức là có tự tương quan (dương hoặc âm).

Các phần mềm kinh tế lượng đã tự động tính giá trị  $d$ . Chẳng hạn, với số liệu cho ở thí dụ 2.1 thì  $d = 2,49331$ . Tra bảng  $d$  ta được  $d_L = 0,812$ ;  $d_U = 1,58$ . Ta thấy  $4-d_U < d < 4-d_L$  nên theo

kiểm định d. Durbin-Watson không cho kết luận. Nếu áp dụng kiểm định cải biên ta thấy  $4 - d = 1,5067 < d_U = 1,58$  nên bác bỏ  $H_0$ , tức có sự tương quan âm.

## 7.6 CÁC BIỆN PHÁP KHẮC PHỤC.

Vì khi có sự tương quan chuỗi, các ước lượng OLS là không hiệu quả. Sau đây đưa ra một số biện pháp khắc phục hiện tượng này, nhưng các biện pháp lại dựa trên sự hiểu biết về bản chất của sự phụ thuộc qua lại giữa các nhiễu. Chính ta phân biệt hai tình huống:

- Tự tương quan đã biết.
- Tự tương quan chưa biết.

### 7.6.1 Trường hợp đã biết cấu trúc của tự tương quan:

Vì các nhiễu  $U_t$  không quan sát được nên tính chất của tương quan chuỗi thường là vấn đề suy đoán hoặc là do những đòi hỏi cấp bách của thực tiễn. Trong thực hành, người ta thường giả sử rằng  $U_t$  theo mô hình tự hồi quy bậc nhất, nghĩa là:

$$U_t = \rho U_{t-1} + \varepsilon_t \quad (7.25)$$

Trong đó:  $|\rho| < 1$  và  $\varepsilon_t$  thỏa mãn các giả thiết của phương pháp OLS (trung bình bằng 0, phương sai không đổi và không tự tương quan). Giả sử (7.25) là đúng thì vấn đề tương quan chuỗi có thể được giải quyết thỏa đáng nếu hệ số tương quan  $\rho$  là đã biết. để làm sáng tỏ vấn đề này, ta xét mô hình hai biến:

$$Y_t = \beta_1 + \beta_2 X_t + U_t \quad (7.26)$$

Nếu (7.26) đúng với  $t$  thì cũng đúng với  $t-1$ , nên:

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + U_{t-1} \quad (7.27)$$

Nhân 2 vế của (7.27) với  $\rho$  ta được:

$$\rho Y_{t-1} = \rho \beta_1 + \rho \beta_2 X_{t-1} + \rho U_{t-1} \quad (7.28)$$

Trừ (7.26) cho (7.28) ta được:

$$\begin{aligned} Y_t - Y_{t-1} &= \beta_1(1 - \rho) + \beta_2(X_t - \rho X_{t-1}) + (U_t - \rho U_{t-1}) = \\ &= \beta_1(1 - \rho) + \beta_2(X_t - \rho X_{t-1}) + V_t \end{aligned} \quad (7.29)$$

Đặt:  $\beta_1^* = \beta_1(1 - \rho)$ ;  $\beta_2^* = \beta_2$

$$Y_t^* = Y_t - Y_{t-1}; \quad X_t^* = X_t - \rho X_{t-1}$$

Khi đó (7.29) có thể viết dưới dạng:

$$Y_t^* = \beta_1^* + \beta_2^* X_t^* + V_t \quad (7.30)$$

Vì  $V_t$  thỏa mãn các giả thiết của phương pháp OLS đối với các biến  $Y^*$  và  $X^*$  nên các ước lượng tìm được sẽ là các ước lượng tuyến tính không chệch tốt nhất.

Phương trình hồi quy (7.29) được gọi là phương trình sai phân tổng quát.

Việc ước lượng hồi quy  $Y^*$  đối với  $X^*$  có hay không có hệ số chặn phụ thuộc vào hồi quy gốc có hệ số chặn hay không. Trong quy trình lấy sai phân chúng ta bị mất một quan sát vì quan sát đầu tiên không có quan sát đứng trước nó. để tránh mất mát một quan sát này thì quan sát đầu của  $Y$  và  $X$  được biến đổi như sau:

$$Y_1^* = Y_1 \sqrt{1 - \rho^2}; \quad X_1^* = X_1 \sqrt{1 - \rho^2}$$

### 7.6.2 Trường hợp $\rho$ chưa biết:

Trong mục này ta xét một số thủ tục ước lượng  $\rho$ .

1. *Phương pháp sai phân cấp 1*: Như ta đã biết  $-1 \leq \rho \leq 1$  nghĩa là  $\rho$  nằm giữa  $[-1,0)$  hoặc  $(0,1]$  cho nên người ta có thể bắt đầu từ các giá trị ở các đầu mút của các khoảng đó. Nghĩa là ta có thể giả thiết rằng:  $\rho = 0$ , tức không có tương quan chuỗi. Hoặc  $\rho = \pm 1$ , tức có tương quan dương hoặc âm hoàn hảo.

Trên thực tế, khi ước lượng hồi quy người ta thường giả thiết không có tự tương quan rồi sau đó tiến hành kiểm định Durbin- Watson hay các kiểm định khác để xem các giả thiết này có đúng hay không. Tuy nhiên nếu  $\rho = 1$  thì phương trình sai phân tổng quát (7.29) quy về phương trình sai phân cấp 1:

$$Y_t - Y_{t-1} = \beta_2(X_t - X_{t-1}) + (U_t - U_{t-1}) = \beta_2(X_t - X_{t-1}) + \varepsilon_t$$

$$\text{Hay: } \Delta Y_t = \beta_2 \Delta X_t + \varepsilon_t \quad (7.31)$$

Trong đó:  $\Delta$  là toán tử sai phân cấp 1. Để ước lượng hồi quy (7.31) ta sẽ sử dụng mô hình hồi quy qua gốc tọa độ. Giả sử mô hình ban đầu là:

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 t + U_t \quad (7.32)$$

trong đó  $t$  là biến xu thế, còn  $U_t$  theo sơ đồ tự hồi quy bậc nhất.

Thực hiện phép biến đổi sai phân cấp 1 đối với (7.32) ta được:

$$\Delta Y_t = \beta_2 \Delta X_t + \beta_3 + \varepsilon_t \quad (7.33)$$

Phương trình (7.33) có hệ số chặn dưới dạng sai phân cấp 1. Nhưng ta chú ý rằng  $\beta_3$  là hệ số của biến xu thế trong mô hình ban đầu. vì vậy, nếu có số hạng chặn ở sai phân cấp 1 thì điều đó có nghĩa là có một số hạng xu thế tuyến tính trong mô hình gốc và số hạng chặn thực ra là hệ số của biến xu thế.

Thí dụ, nếu  $\beta_3$  trong (7.33) là dương thì điều đó có nghĩa là có xu thế tăng trong  $Y$  sau khi đã tính đến ảnh hưởng của tất cả các biến khác.

Nếu  $\rho = -1$  nghĩa là có tương quan âm hoàn toàn. (đây không phải là trường hợp điển hình của các chuỗi thời gian trong kinh tế), phương trình sai phân tổng quát bây giờ có dạng: (suy từ 7.29):

$$Y_t + Y_{t-1} = 2\beta_1 + \beta_2(X_t - X_{t-1}) + \varepsilon_t$$

$$\text{Hay: } \frac{Y_t + Y_{t-1}}{2} = \beta_1 + \beta_2 \frac{X_t + X_{t-1}}{2} + \frac{\varepsilon_t}{2} \quad (7.34)$$

Mô hình này được gọi là mô hình hồi quy trung bình trượt (2 thời kỳ) vì chúng ta hồi quy giá trị của một trung bình trượt đối với một trung bình trượt khác.

Phép biến đổi sai phân cấp 1 đã giới thiệu trên đây rất phổ biến trong kinh tế lượng ứng dụng vì nó dễ thực hiện. Nhưng lưu ý rằng phép biến đổi này giả thiết rằng  $\rho = +1$ , nghĩa là các nhiễu có tương quan dương hoàn toàn. Nếu điều ta giả thiết không xảy ra thì điều đó có khi còn tồi tệ hơn bản thân căn bệnh. Nhưng là thế nào để biết  $\rho = +1$  là đúng? Để trả lời câu hỏi này ta xét tiếp mục dưới đây.

a. Ước lượng  $\rho$  dựa trên thống kê  $d$  - Durbin- Watson.

Trong phần kiểm định  $d$  chúng ta đã biết các công thức:

$$d \approx 2(1 - \hat{\rho}) \quad (7.35)$$

$$\text{hoặc:} \quad \hat{\rho} = 1 - \frac{d}{2} \quad (7.36)$$

Công thức này gợi cho ta cách thức đơn giản để thu được ước lượng của  $\rho$  từ thống kê  $d$ . Từ (7.35) chỉ ra rằng giả thiết sai phân cấp 1 với  $\hat{\rho} = +1$  chỉ đúng khi  $d = 0$  hoặc xấp xỉ bằng 0. Cũng vậy khi  $d = 2$  thì  $\hat{\rho} = 0$  và khi  $d = 4$  thì  $\hat{\rho} = -1$ . Do đó thống kê  $d$  cung cấp cho ta phương pháp để thu được giá trị của  $\rho$ .

Chúng ta cần lưu ý rằng quan hệ (7.36) chỉ là quan hệ xấp xỉ và có thể không đúng với mẫu nhỏ. Đối với các mẫu nhỏ có thể sử dụng thống kê  $d$  cải biên của Theil-Nagar.

Theil-Nagar đã đề xuất rằng trong các mẫu nhỏ, thay cho việc ước lượng  $\rho$  như là  $(1-d/2)$ , có thể ước lượng như là:

$$\hat{\rho} = \frac{n^2 \left(1 - \frac{d}{2}\right) + k^2}{n^2 - k^2}$$

trong đó  $n$  là tổng số quan sát;  $d$  là Durbin- Watson,  $d$  và  $k$  là số các hệ số (bao gồm cả tung độ gốc) cần phải ước lượng. Khi  $n$  đủ lớn, ước lượng  $\rho$  này là bằng với ước lượng thu được bởi công thức đơn giản  $(1-d/2)$

Khi  $\rho$  đã được ước lượng thì có thể biến đổi tập số liệu như đã chỉ ra ở (7.30) và tiến hành ước lượng theo phương pháp OLS thông thường. Nhưng vấn đề đặt ra là các ước lượng thu được có phải là các ước lượng tuyến tính không chệch và có phương sai nhỏ nhất? Chú ý rằng trong phương trình sai phân tổng quát chứa  $\rho$  chứ không phải là  $\hat{\rho}$ , nhưng khi tiến hành hồi quy theo phương pháp OLS ta sử dụng  $\hat{\rho}$ . Ở đây ta có thể áp dụng một quy tắc thực hành là: Khi ta sử dụng một ước lượng thay cho giá trị đúng thì các hệ số thu được từ phương pháp OLS có thuộc tính tối ưu thông thường chỉ tiệm cận có nghĩa là chúng chỉ có thuộc tính đó đối với mẫu có kích thước lớn. Các kết luận khi tiến hành các kiểm định cũng chỉ đúng một cách tiệm cận. Vì vậy đối với các mẫu nhỏ chúng ta cần cẩn thận khi giải thích các kết quả ước lượng.

#### *b. Thủ tục lặp Cochrance- Orcutt để ước lượng $\rho$ :*

Một cách khác để ước lượng  $\rho$  từ thống kê  $d$ - Durbin- Watson là phương pháp Cochrance-Orcutt. Phương pháp này sử dụng các phần dư  $e_t$  đã được ước lượng để thu được thông tin về  $\rho$  chưa biết.

Ta xét mô hình hai biến sau:

$$Y_t = \beta_1 + \beta_2 X_t + U_t \quad (7.37)$$

Giả sử  $U_t$  được sinh ra từ lược đồ AR(1):

$$U_t = \rho U_{t-1} + \varepsilon_t \quad (7.38)$$

Các bước ước lượng  $\rho$  được tiến hành như sau:

Bước 1: Ước lượng mô hình (7.37) bằng phương pháp OLS và thu được các phần dư  $e_t$ .

Bước 2: Sử dụng các phần dư để ước lượng hồi quy:

$$e_t = \rho e_{t-1} + V_t \quad (7.39)$$

**Bước 3:** Sử dụng  $\hat{\rho}$  thu được từ (7.39) để ước lượng phương trình sai phân tổng quát (7.29).  
Tức phương trình:

$$Y_t - \hat{\rho} Y_{t-1} = \beta_1(1 - \hat{\rho}) + \beta_2(X_t - \hat{\rho} X_{t-1}) + (U_t - \hat{\rho} U_{t-1})$$

$$\text{Đặt: } Y_t^* = Y_t - \hat{\rho} Y_{t-1}; X_t^* = X_t - \hat{\rho} X_{t-1}; \beta_1^* = \beta_1(1 - \hat{\rho}); \beta_2^* = \beta_2.$$

Ta ước lượng hồi quy:

$$Y_t^* = \beta_1^* + \beta_2^* X_t^* + V_t \quad (7.40)$$

**Bước 4:** Chúng ta chưa biết trước rằng  $\hat{\rho}$  thu được từ (7.39) có phải là ước lượng tốt nhất của  $\rho$  hay không. Ta thế giá trị của ước lượng  $\beta_1^*$  và  $\beta_2^*$  thu được từ (7.40) vào hồi quy gốc (7.37) và thu được các phần dư mới  $e_t^*$ :

$$e_t^* = Y_t - (\beta_1^* + \beta_2^* X_t) \quad (7.41)$$

Ước lượng phương trình hồi quy tương tự với (7.39):

$$e_t^* = \hat{\rho}^* e_{t-1}^* + W_t \quad (7.42)$$

$\hat{\rho}^*$  là ước lượng vòng hai của  $\rho$ .

Thủ tục này tiếp tục cho đến khi các ước lượng kế tiếp nhau của  $\rho$  khác nhau một lượng rất nhỏ, chẳng hạn nhỏ hơn 5% hoặc 0,5%. Trong thực tế dùng 3-4 bước lặp là đủ.

**c. Phương pháp Durbin- Watson 2 bước để ước lượng  $\rho$ .**

Để minh họa phương pháp này, chúng ta viết lại phương trình sai phân tổng quát dưới dạng sau:

$$Y_t = \beta_1(1 - \rho) + \beta_2 X_t - \rho \beta_2 X_{t-1} + \rho Y_{t-1} + \varepsilon_t \quad (7.43)$$

Durbin đã đề xuất thủ tục 2 bước như sau để ước lượng  $\rho$ :

**Bước 1:** Coi (7.43) như là một mô hình hồi quy bội, hồi quy  $Y_t$  theo  $X_t$ ,  $X_{t-1}$  và  $Y_{t-1}$  và coi giá trị ước lượng được đối với hệ số hồi quy của  $Y_{t-1}$  ( $= \hat{\rho}$ ) là ước lượng của  $\rho$ . Mặc dầu là ước lượng chệch nhưng ta có ước lượng vững của  $\rho$ .

**Bước 2:** Sau khi thu được  $\hat{\rho}$ , hãy biến đổi  $Y_t^* = Y_t - \hat{\rho} Y_{t-1}$  và  $X_t^* = X_t - \hat{\rho} X_{t-1}$  và ước lượng hồi quy (7.30) với các biến đã được biến đổi như trên.

Như vậy theo phương pháp này thì bước 1 là để ước lượng  $\rho$  còn bước 2 là để thu được các tham số.

Thí dụ: Có các số liệu về mối quan hệ giữa 2 đại lượng kinh tế ( $Y$  là biến phụ thuộc;  $X$  là biến độc lập) trong vòng 24 năm từ năm  $t=1$  đến  $t=24$  như sau:

t	Y	X	t	Y	X
1	104,66	5,63	13	143,33	4,83
2	108,53	5,46	14	144,66	4,73
3	97,30	5,63	15	152,33	4,46
4	95,96	5,60	16	178,33	4,20

5	98,83	5,83	17	192,00	3,83
6	97,23	5,76	18	186,00	3,90
7	99,06	5,56	19	188,00	3,86
8	113,66	5,63	20	193,33	3,70
9	117,00	5,46	21	187,66	3,66
10	119,66	5,26	22	175,33	3,83
11	124,33	5,06	23	178,00	3,93
12	133,00	5,08	24	187,66	3,96

Mô hình hồi quy được chọn cho nghiên cứu thực nghiệm là:

$$\ln Y_t = \beta_1 \beta_2 \ln X_t + U_t$$

giả sử tất cả các giả thiết OLS được thỏa mãn. Ước lượng hồi quy trên ta được:

Variable	Coeff	Std err	t
X	-1,5375	0,0711	-21,612
-Cons	7,3084	0,1110	65,825

$$r^2 = 0,955; \quad d = 0,9021$$

Từ phương trình hồi quy đã ước lượng được và giá trị của thống kê  $d$ , ta xét xem liệu có tương quan chuỗi hay không?

Với  $n = 24$ ;  $k = 1$  và  $\alpha = 0,05$  ta tính được  $d_U = 1,45$ ;  $d_L = 1,27$ . Vậy giá trị của thống kê  $d$  bé hơn  $d_L$  cho nên ta kết luận: Có tương quan thuận chiều. Như vậy ta không thể tin vào các sai số chuẩn đã được ước lượng và các tỷ số  $t$ , cho nên cần phải có biện pháp khắc phục. Việc khắc phục lại phụ thuộc vào  $\rho$  và  $\rho$  được ước lượng bằng một số phương pháp đã nêu trên. Kết quả như sau:

Phương pháp sử dụng	Giá trị $\hat{\rho}$
d- Durbin- Watson	0,5490
d- Theil - Nagra	0,5598
Cochrance- Orcutt	
Bước lặp 1	0,54571
Bước lặp 2	0,57233
Bước lặp 3	0,57836
Bước lặp 4	0,57999
Durbin 2 bước	0,79520

Lưu ý thủ tục lặp Cochrance- Orcutt dừng ở bước 4 vì giữa bước 3 và 4 không khác nhau nhiều.

Theo kết quả trên các thủ tục đều cho kết quả gần giống nhau, riêng thủ tục Durbin 2 bước cho kết quả hoàn toàn khác. Vậy trọng thực tế chọn phương pháp nào để ước lượng  $\rho$ ?

Thực tế ta thấy rằng nếu chúng ta có mẫu lớn (chẳng hạn trên 60 quan sát) thì chọn phương pháp nào cũng không gây ra sự khác biệt nhiều lắm vì chúng đều mang lại kết quả tương

tự nhau. Nhưng điều này sẽ không đúng khi các mẫu nhỏ, trong trường hợp này kết quả sẽ phụ thuộc vào phương pháp được chọn. Nhưng liệu có phương pháp nào đáng được ưa chuộng hơn hay không? Không có câu trả lời trong trường hợp này vì qua mô phỏng Monte-Carlo thì người ta thấy rằng không thiên vị một phương pháp nào. Tuy nhiên trong thực tế phương pháp thường được sử dụng là phương pháp lặp Cochran- Orcutt mà ngày nay đã được đưa vào chương trình máy tính.

## TÓM TẮT NỘI DUNG CHƯƠNG 7

Trong hồi quy tương quan cổ điển chúng ta giả thiết giữa các sai số ngẫu nhiên không có sự tương quan với nhau. Nhưng trong thực tế có thể xảy ra hiện tượng mà sai số ở quan sát này lại có thể phụ thuộc vào sai số ở quan sát khác. Nguyên nhân là có thể do quán tính, do hiện tượng mạng nhện, do các độ trễ, do xử lý số liệu và do lập mô hình chưa chính xác. Từ đó gây nên nhiều hậu quả, như: Các ước lượng OLS không còn là các ước lượng tuyến tính không chệch tốt nhất nữa; Phương sai của các ước lượng thường là chệch và do đó làm cho giá trị của kiểm định t lớn, dẫn đến kết luận sai khi kiểm định; Phương sai mẫu sẽ là ước lượng chệch của phương sai chung;... Vì vậy khi tiến hành hồi quy cần phát hiện xem có hiện tượng tự tương quan hay không với một mẫu cho trước. Bằng cách sử dụng phương pháp đồ thị; phương pháp kiểm định đoạn mạch, phương pháp kiểm định  $\chi^2$  và kiểm định d- Durbin- Watson. Nếu có, cần sử dụng các biện pháp thích hợp để khắc phục, đó là phương pháp đổi biến nếu biết trước cấu trúc của tự tương quan; hoặc trong trường hợp  $\rho$  chưa biết thì dùng phương pháp sai phân cấp 1, bao gồm sử dụng thống kê d- Durbin- Watson, phương pháp d-Theil-Nagra, phương pháp Cochran- Orcutt. Trong các phương pháp dùng để xử lý hiện tượng tự tương quan thì phương pháp thủ tục lặp Cochran- Orcutt thường hay được sử dụng nhiều nhất.

## CÂU HỎI VÀ BÀI TẬP ÔN TẬP CHƯƠNG 7

### I. Câu hỏi lý thuyết:

1. Nêu bản chất của hiện tượng tự tương quan: định nghĩa, nguyên nhân.
2. Tính chất của các ước lượng khi có hiện tượng tự tương quan. Các kiểm định t,  $\chi^2$ , F có đem lại thông tin chính xác không?
3. Trình bày ngắn gọn các phương pháp phát hiện và khắc phục hiện tượng tự tương quan.

### II. Bài tập:

1. Giả sử có các phần dư hồi quy về năng suất và tiền lương theo các tháng cho ở bảng, hãy sử dụng tiêu chuẩn  $\chi^2$  để kiểm định tính độc lập của các phần dư:

Năm	Tháng	Phần dư $e_t$	Phần dư tại $e_{t-1}$
2003	1	-1,2116	
	2	-1,1274	-1,2116
	3	-0,7908	-1,1274
	4	-1,1368	-0,7908
	5	-0,8954	-1,1368
	6	-0,1489	-0,8954
	7	-0,2873	-0,1489
	8	0,2270	-0,2873
	9	0,9983	0,2270
	10	2,2334	0,9983
	11	2,7557	2,2334
	12	2,1971	2,7557
2004	1	2,5384	2,1971
	2	2,1576	2,5384
	3	2,6559	2,1576
	4	1,4226	2,6559
	5	1,4465	1,4226
	6	0,5656	1,4465
	7	0,9530	0,5656
	8	0,2954	0,9530
	9	-0,2459	0,2954
	10	-4,5021	-0,2459
	11	-2,8772	-4,5021
	12	-4,0882	-2,8772

## CHƯƠNG 8: CHỌN MÔ HÌNH VÀ KIỂM ĐỊNH VIỆC CHỈ ĐỊNH MÔ HÌNH

### GIỚI THIỆU

Trong các chương trước chúng ta đã xét một số mô hình hồi quy tuyến tính một phương trình, chẳng hạn như hàm tiêu dùng, hàm sản xuất, hàm doanh thu thuế... Trong khi xem xét các mô hình đó, ta ngầm giả định rằng mô hình đã chọn là mô hình đúng, nghĩa là mô hình phản ánh đúng hiện tượng đang nghiên cứu. Nói một cách khác chúng ta đã ngầm giả định rằng không có sai lầm trong việc chọn mô hình, không có việc thiếu biến thích hợp hoặc không có các biến không cần thiết trong mô hình. Vấn đề mà chúng ta quan tâm đến là vấn đề ước lượng mô hình khi mô hình đã được chỉ định, và tìm các phương pháp ước lượng để đem lại ước lượng vững và hiệu quả của các tham số.

Nhưng nếu có sai lầm trong việc chỉ định mô hình, có nghĩa là thay cho việc ước lượng một mô hình đúng ta lại ước lượng một mô hình không đúng. Nếu điều này xảy ra thì hậu quả sẽ như thế nào? Chúng ta có thể gặp những loại sai lầm chỉ định nào? Làm thế nào để phát hiện được sai lầm chỉ định? Đó là những nội dung mà chúng ta sẽ đề cập đến trong chương này, tuy nhiên do khuôn khổ của cuốn sách mà chúng tôi chỉ trình bày những vấn đề cơ bản nhất và không quá phức tạp.

Trước khi trả lời những câu hỏi trên đây, chúng ta cần có tiêu chuẩn nào đó để giải đáp cho câu hỏi: "Thế nào là một mô hình đúng?". Mặc dầu trong thực tế việc tìm kiếm một mô hình đúng là một điều cực kỳ khó khăn.

- Nắm vững lý thuyết của các chương trước, đặc biệt là các giả thiết của phương pháp OLS các sai lầm khi hồi quy thường mắc phải.

- Nắm được cách khắc phục các sai lầm khi chọn mô hình sai.

### NỘI DUNG

#### 8.1. CÁC THUỘC TÍNH CỦA MỘT MÔ HÌNH TỐT

Để đánh giá, xem xét một mô hình có thể căn cứ vào các tiêu chuẩn sau của A.C Harvey:

##### 1. Tính kiệm

Một mô hình không bao giờ có thể thu tóm toàn bộ thực tại, việc trừu tượng hóa và đơn giản hóa là cần thiết bởi vì mô hình là sự biểu diễn đơn giản nhưng hoàn chỉnh của hiện thực. Nguyên tắc kiệm cho rằng hãy giữ cho mô hình càng đơn giản càng tốt.

##### 2. Tính đồng nhất

Nghĩa là với một tập dữ liệu đã cho, các tham số ước lượng phải có giá trị thống nhất

##### 3. Tính thích hợp.

Vì mục đích của phân tích hồi quy là giả thích sự biến động của biến phụ thuộc bằng biến giải thích của mô hình càng nhiều càng tốt, nên một mô hình sẽ được coi là mô hình tốt nếu có  $R^2$  hoặc  $\bar{R}^2$  càng gần 1 thì được coi là càng thích hợp.

#### 4. Tính vững về mặt lý thuyết

Trong việc xây dựng mô hình ta phải có một cơ sở lý thuyết nào đó, nếu không sẽ dẫn đến kết quả sai.

#### 5. Khả năng dự đoán.

Tiêu chuẩn thực tiễn của chân lý ở đây thể hiện ở sức dự đoán của mô hình phù hợp với thực tế. Chẳng hạn trong việc lựa chọn giữa mô hình của trường phái tiền tệ và mô hình trường phái Keynes, ta sẽ chọn mô hình nào mà những dự đoán lý thuyết của nó được thực tiễn chứng minh.

### 8.2. CÁC LOẠI SAI LẦM CHỈ ĐỊNH.

#### 8.2.1 Bỏ sót một biến thích hợp.

Trong việc xây dựng mô hình có thể chúng ta phạm sai lầm là bỏ sót một hay một số biến thích hợp mà đáng lẽ chúng phải có mặt trong mô hình

Việc bỏ sót biến như vậy sẽ gây ra hậu quả như thế nào đối với thủ tục ước lượng bình phương nhỏ nhất thông thường.

Giả sử mô hình “đúng” biểu thị mối liên hệ kinh tế giữa biến phụ thuộc  $Y$  và các biến  $X_2$  và  $X_3$  có dạng:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + U_t \quad (8.1)$$

Trong đó các  $\beta_1, \beta_2$  và  $\beta_3$  là các hệ số,  $U_t$  là sai số ngẫu nhiên ;  $t$  biểu thị thời gian.

Nhưng vì một lý do nào đó ta ước lượng mô hình

$$Y_t = \alpha_1 + \alpha_2 X_{2t} + V_t \quad (8.2)$$

Chúng ta xem xét việc bỏ biến  $X_{3t}$  gây tác hại như thế nào

1. Nếu  $X_3$  tương quan với biến đưa vào  $X_2$  thì  $\hat{\alpha}_1, \hat{\alpha}_2$  là các ước lượng chệch của  $\beta_1$  và  $\beta_2$ , nghĩa là:

$$E(\hat{\alpha}_1) \neq \beta_1 \text{ và } E(\hat{\alpha}_2) \neq \beta_2$$

Nếu gọi  $\beta_{32}$  là hệ số góc trong hàm hồi quy của biến bị bỏ sót  $X_3$  đối với biến  $X_2$  thì có thể chỉ rằng:

$$E(\hat{\alpha}_2) = \beta_2 + \beta_3 \beta_{32} \quad (8.3)$$

$$E(\hat{\alpha}_1) = \beta_1 + \beta_3 (\bar{X}_3 - \beta_{32} \bar{X}_2) \quad (8.4)$$

Nếu  $\beta_3$  và  $\beta_{32}$  dương thì  $\hat{\alpha}_2$  sẽ chệch lên, về trung bình nó sẽ ước lượng cao  $\beta_2$

Nếu  $\beta_3 > 0, \beta_{32} < 0$  hoặc ngược lại thì  $\hat{\alpha}_2$  sẽ chệch xuống, về trung bình nó sẽ ước lượng thấp hơn  $\beta_2$

Tương tự  $\hat{\alpha}_1$  sẽ chệch lên nếu  $\beta_3 (\bar{X}_3 - \beta_{32} \bar{X}_2) > 0$  và chệch xuống nếu  $\beta_3 (\bar{X}_3 - \beta_{32} \bar{X}_2) < 0$ .

2.  $\hat{\alpha}_1$  và  $\hat{\alpha}_2$  không phải là ước lượng vững.

3. Nếu  $X_2$  và  $X_3$  không tương quan thì  $\beta_{32} = 0$ , khi đó  $E(\hat{\alpha}_1) = \beta_2$  nghĩa là  $\hat{\alpha}_2$  là ước lượng không chệch của  $\beta_2$  và đồng thời nó cũng là ước lượng vững. Trong khi đó  $\hat{\alpha}_1$  là ước lượng chệch của  $\beta_1$ .

4. Phương sai sai số ước lượng từ (8.2) là một ước lượng chệch của phương sai sai số đúng. Nói cách khác, phương sai sai số ước lượng từ mô hình đúng (8.1) và phương sai sai số ước lượng từ mô hình chỉ định sai sẽ không như nhau.

5. Phương sai ước lượng  $\hat{\alpha}_2 (= \hat{\sigma}^2 / \sum x_{2t}^2)$  là ước lượng chệch của phương sai của ước lượng đúng  $\beta_2$ . Thậm chí cả khi  $X_2$  và  $X_3$  không tương quan, phương sai này vẫn là ước lượng chệch vì có thể chỉ ra rằng:

$$E[\text{var}(\hat{\alpha}_2)] = \text{var}(\hat{\beta}_2) + \frac{\beta_3^2 \sum x_{3t}^2}{(n-2) \sum x_{2t}^2} \quad (8.5)$$

Nghĩa là giá trị kỳ vọng của phương sai của  $\hat{\alpha}_2$  không bằng phương sai  $\hat{\beta}_2$  vì số hạng thứ hai trong phương trình (8.5) dương. Về trung bình  $\text{var}(\hat{\alpha}_2)$  sẽ ước lượng cao phương sai đúng của  $\beta_2$ .

6. Kết quả là khoảng tin cậy thông thường và các thủ tục kiểm định giả thiết không còn đáng tin cậy nữa,

Trong trường hợp phương trình (8.5) thì khoảng tin cậy sẽ rộng hơn, do đó có thể có khuynh hướng là ta thường chấp nhận giả thiết rằng các giá trị thực của hệ số bằng không (hoặc giả thiết  $H_0$  khác) hơn là so với tình huống thực.

*Thí dụ:* Ta hãy xét mối quan hệ giữa chi tiêu cho nhập khẩu và thu nhập sau thuế của một nước trong thời kỳ 20 năm. Số liệu cho ở bảng dưới đây:

Bảng 9.1. Số liệu chi tiêu cho nhập khẩu

Năm	Chi tiêu cho nhập khẩu	Thu nhập sử dụng được	Năm	Chi tiêu cho nhập khẩu	Thu nhập sử dụng được
1	35,7	1551,3	11	247,1	2167,4
2	144,6	1599,8	12	277,9	2212,6
3	150,9	1668,1	13	253,6	2214,3
4	166,2	1728,4	14	258,7	2248,6
5	190,7	1797,4	15	249,5	2261,5
6	218,2	1916,3	16	282,2	2331,9
7	211,8	1896,9	17	251,1	2469,8
8	187,9	1931,7	18	367,9	2542,8
9	299,9	2001,0	19	412,3	2640,9
10	159,4	2066,6	20	439,0	2686,3

Giả sử mô hình ‘ đúng ‘ biểu thị quan hệ giữa chỉ tiêu cho nhập khẩu và thu nhập sử dụng và biến xu thế là như sau:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + U_t \quad (8.6)$$

Trong đó:  $Y$ : chỉ tiêu cho thu nhập

$X_2$ : thu nhập có thể sử dụng được

$X_3$ : Biến xu thế (biểu thị thời gian hay xu thế), nó lấy giá trị từ 1,2,...,20

Phương trình (8.6) cho ta biết rằng có một biến khác  $X_3$  cũng ảnh hưởng đến chỉ tiêu nhập khẩu ngoài thu nhập có thể sử dụng được. Nó có thể là một biến như dân số, thị hiếu, công nghệ.....mà ta biểu thị bằng một biến thời gian hoặc xu thế.

Nhưng thay cho việc ước lượng mô hình đúng (8.6) ta lại ước lượng mô hình sau đây:

$$Y_t = \alpha_1 + \alpha_2 X_{2t} + V_t \quad (8.7)$$

Kết quả hồi quy (8.7) như sau:

$$\hat{Y}_t = -261,0914 + 0,2452312 X_{2t} \quad (8.8)$$

Nếu ước lượng mô hình (8.6) ta được kết quả:

$$\hat{Y}_t = -859,9217 + 0,649962 X_{2t} - 23,19518 X_{3t} \quad (8.9)$$

Đối với mô hình (8.8), tức là mô hình được chọn sai, ta có:

$$\frac{d\hat{Y}_t}{dX_{2t}} = 0,2452312$$

Còn ở mô hình (8.9), là mô hình đúng, thì:

$$\frac{d\hat{Y}_t}{dX_{2t}} = 0,6499$$

Nghĩa là khuynh hướng biên của chỉ tiêu cho nhập khẩu là 0,2452312 (đối với mô hình 8.6). Như vậy mô hình này cho biết khi thu nhập tăng 1 đơn vị thì chỉ tiêu cho nhập khẩu tăng 0,2452312 đơn vị. Trong khi đó mô hình đúng (8.9) lại cho biết rằng khi thu nhập tăng 1 đơn vị thì chỉ tiêu cho nhập khẩu tăng 0,6499 đơn vị. Như vậy, mô hình chọn sai ước lượng thấp khuynh hướng biên của tiêu dùng và nó bị chệch xuống.

Như vậy việc bỏ sót biến  $X_3$  chẳng những đã bỏ qua ảnh hưởng của  $X_3$  đối với  $Y$  mà cả ảnh hưởng của  $X_3$  đối với  $X_2$ .

Ngoài ra hệ số chặn, phương sai và sai số chuẩn của hai mô hình (8.8) và (8.9) cũng sai khác nhau rất nhiều. Như vậy hậu quả của việc bỏ sót biến là rất tai hại.

### 8.2.2. Đưa vào những biến không thích hợp

Trong mục này chúng ta xét trường hợp khi đưa vào mô hình những biến không thích hợp thì hậu quả sẽ như thế nào?

Để làm sáng tỏ vấn đề này ta hãy xét trên 2 mô hình đơn giản sau:

Giả sử rằng:

$$Y_i = \beta_1 + \beta_2 X_{2i} + U_i \quad (8.10)$$

là mô hình chỉ định đúng, nhưng người ta đã đưa thêm vào một biến thừa  $X_3$  và ước lượng mô hình sau:

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + V_i \quad (8.11)$$

Trong trường hợp này, sai lầm chỉ định là gì? Hậu quả của việc ước lượng mô hình (8.11) thay cho ước lượng mô hình (8.10) là như sau:

a. Các ước lượng OLS của mô hình (8.10) vẫn là các ước lượng không chệch và vững, nghĩa là ta vẫn có:

$$E(\hat{\alpha}_1) = \beta_1, E(\hat{\alpha}_2) = \beta_2 \text{ và } E(\hat{\alpha}_3) = 0$$

b. Ước lượng của  $\sigma^2$  thu được từ mô hình hồi quy (8.11) là ước lượng vững.

c. Tuy nhiên các ước lượng thu được từ (8.11) là không hiệu quả, các phương sai của chúng sẽ lớn hơn các phương sai của các ước lượng thu được từ mô hình chỉ định đúng (8.10). Có thể chỉ ra sự không hiệu quả tương đối của  $\hat{\alpha}_2$ . Vì từ phương pháp bình phương nhỏ nhất thông thường chúng ta có:

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2}$$

$$\text{và } \text{var}(\hat{\alpha}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)}$$

Do đó:

$$\frac{\text{var}(\hat{\alpha}_2)}{\text{var}(\hat{\beta}_2)} = \frac{1}{1 - r_{23}^2}$$

Vì  $0 \leq r_{23}^2 \leq 1 \Rightarrow \text{var}(\hat{\alpha}_2) \geq \text{var}(\hat{\beta}_2)$  nghĩa là phương sai của  $\hat{\alpha}_2$  lớn hơn phương sai của  $\hat{\beta}_2$  dù  $E(\hat{\alpha}_2) = \beta_2$ .

Từ nhận xét trên ta thấy rằng các khoảng tin cậy dựa trên các sai số tiêu chuẩn của các ước lượng thu được từ mô hình chỉ định sai (8.11) sẽ lớn hơn các khoảng tin cậy dựa trên các sai số tiêu chuẩn của các ước lượng từ mô hình đúng (8.10), dù các khoảng tin cậy dựa trên các sai số tiêu chuẩn đó là chấp nhận được đối với các thủ tục kiểm định giả thiết.

### 8.2.3 Dạng hàm không đúng

Bây giờ ta xét một loại sai lầm chỉ định khác, đó là sai lầm trong việc chỉ định mô hình. Như vậy nếu mắc phải sai lầm trong trường hợp này có nghĩa là đáng lẽ ta ước lượng mô hình đúng thì ta lại ước lượng một mô hình chỉ định sai.

Ta hãy quay lại thí dụ đã xét về chi tiêu cho nhập khẩu. Giả sử mô hình đúng là mô hình có dạng:

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + V_i \quad (8.12)$$

nhưng vì một lý do gì đó mà người ta ước lượng mô hình:

$$\ln Y_t = \alpha_1 + \alpha_2 \ln X_{2t} + \alpha_3 \ln X_{3t} + V_t \quad (8.13)$$

Việc chọn dạng hàm sai dẫn đến hậu quả như thế nào? Muốn phân tích cụ thể ta sử dụng số của bảng (8.1) nêu ở phần trên để ước lượng mô hình (8.13). Ta được kết quả như sau:

$$\ln \hat{Y}_t = -23,72662 + 3,897514 \ln X_{2t} - 0,0526215 X_{3t}$$

Kết quả hồi quy cho thấy: Hệ số  $\hat{\alpha}_2 = 3,8975$  có nghĩa là khi thu nhập sử dụng được tăng 1% thì chi tiêu cho nhập khẩu sẽ tăng 3,8975%. Còn trong khi đó hệ số co giãn tính từ mô hình tuyến tính (8.9) lại cao hơn nhiều. Như vậy nếu mô hình (8.6) là đúng thì các kết luận rút ra từ mô hình (8.13) sẽ không đúng với thực tế và do vậy sẽ dẫn đến những kết luận sai lầm.

### 8.3. PHÁT HIỆN NHỮNG SAI LẦM CHỈ ĐỊNH - CÁC KIỂM ĐỊNH VỀ SAI LẦM CHỈ ĐỊNH.

Trên đây đã trình bày hậu quả của sai lầm chỉ định, nhưng vấn đề đặt ra là làm thế nào để phát hiện ra những sai lầm chỉ định để có thể tìm những biện pháp khắc phục chúng. Mục này sẽ trình bày một số kiểm định để phát hiện sai lầm chỉ định.

#### 8.3.1 Phát hiện ra sự có mặt của các biến không cần thiết.

Giả sử ta có mô hình hồi quy sau:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + U_i \quad (8.14)$$

Trước hết lưu ý rằng, nếu lý thuyết cho rằng tất cả các biến  $X_2, X_3, X_4, X_5$  đều quyết định  $Y$  thì ta phải giữ chúng ở trong mô hình cho dù sau kiểm định thực nghiệm ta nhận thấy rằng hệ số của một biến  $X$  nào đó không có ý nghĩa thống kê. Tuy nhiên, nếu trong mô hình có biến “kiểm tra”, mà ta để chúng ở trong mô hình để tránh sự chệch do bỏ sót thì chúng ta phải tiến hành kiểm định xem sự có mặt của nó ở trong mô hình có thực sự cần thiết hay không. Giả sử  $X_5$  là biến mà ta không biết chắc chắn thuộc vào mô hình hay không thì cách đơn giản là ước lượng hồi quy (9.14) và kiểm định hệ số của  $X_5$  có bằng không hay không.

Giả thiết  $H_0: \beta_5 = 0$

Giả thiết đối  $H_1: \beta_5 \neq 0$

Thống kê kiểm định là  $t = \frac{\hat{\beta}_5}{Se(\hat{\beta}_5)}$  phân phối t Student với  $n-5$  bậc tự do. Nếu giá trị  $t$  tính

được không vượt quá giá trị tới hạn  $t$  ở mức ý nghĩa đã chọn thì ta không bác bỏ giả thiết  $H_0$ . Nếu ta bác bỏ giả thiết  $H_0$  thì biến  $X_5$  có thể thuộc vào mô hình.

Trường hợp ta không chắc chắn rằng cả  $X_4$  và  $X_5$  có thực sự cần thiết ở trong mô hình hay không thì giả thiết  $H_0: \beta_4 = \beta_5 = 0$ . Khi đó ta sử dụng kiểm định F đã biết.

Như vậy việc phát hiện ra sự có mặt của các biến không cần thiết trong mô hình không phải là việc khó khăn, tuy nhiên chúng ta cần lưu ý rằng trong việc tiến hành các kiểm định về chỉ định này, ta đã có một mô hình trong đầu mà ta thừa nhận là mô hình đúng. Khi đã cho mô hình đó, ta có thể tìm ra một hay một số biến  $X$  có thực sự là thích hợp không bằng các kiểm định  $t$  và  $F$ .

#### 8.3.2. Kiểm định các biến bị bỏ sót

Giả sử chúng ta có mô hình hồi quy tuyến tính sau:

$$Y_t = \beta_0 + \beta_1 X_t + U_t \quad (8.15)$$

Để kiểm định xem mô hình có bị chỉ định sai do thiếu một biến  $Z_t$  hay không ta phải ước lượng mô hình:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 Z_t + U_t$$

và kiểm định giả thiết  $H_0: \beta_2 = 0$

Nếu ta sẵn có số liệu về  $Z_t$  thì không có vấn đề gì, những cái cần làm là hồi quy  $Y_t$  đối với  $X_t$  và  $Z_t$  và kiểm định hệ số của  $Z_t$  có bằng không hay không.

Trường hợp không có sẵn quan sát về  $Z$  thì chúng ta sẽ sử dụng xấp xỉ  $\hat{Z}$  đối với  $Z$ . Khi đó phép kiểm định thích hợp đối với các biến bị bỏ sót là ước lượng mô hình:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 \hat{Z}_t + V_t$$

và kiểm định giả thiết  $\beta_2 = 0$ . Trong trường hợp chưa biết biến  $Z$  thì có thể thực hiện bằng các cách sau:

### 1. Kiểm định RESET của Ramsey:

Ramsey đã đề xuất sử dụng  $\hat{Y}_t^2, \hat{Y}_t^3$  làm xấp xỉ cho  $Z_t$ . Trong đó  $\hat{Y}_t$  là giá trị dự đoán của  $Y_t$  từ hồi quy của  $Y_t$  đối với  $X_t$ .

Các bước kiểm định RESET như sau:

**Bước 1:** hồi quy  $Y_t$  theo  $X_t$  và thu được  $\hat{Y}_t$  (ta gọi mô hình này là mô hình cũ).

**Bước 2:** Hồi quy  $Y_t$  đối với  $X_t, \hat{Y}_t^2, \hat{Y}_t^3$  (mô hình hồi quy này được gọi là mô hình mới) và kiểm định giả thiết cho rằng các hệ số của  $\hat{Y}_t^2, \hat{Y}_t^3$  bằng 0.

**Bước 2:** Tính

$$F = \frac{(R_{new}^2 - R_{old}^2) / m}{(1 - R_{new}^2) / (n - k)}$$

Trong đó:  $R_{old}^2$  là hệ số xác định của mô hình cũ.

$R_{new}^2$  là hệ số xác định của mô hình mới.

$m$  là số biến độc lập mới được đưa thêm vào mô hình.

$k$  là số hệ số của mô hình mới.

Nếu  $n$  khá lớn  $F$  có phân phối  $F(m, n-k)$ .

**Bước 4:** Nếu  $F$  có ý nghĩa tại mức 5% ta có thể chấp nhận giả thiết cho rằng mô hình (8.15) được xác định không đúng.

### 2. Kiểm định d. Durbin- Watson:

Thủ tục kiểm định này gồm các bước:

**Bước 1:** Ước lượng mô hình ban đầu, Chẳng hạn:

$$Y_i = \beta_0 + \beta_1 X_i + U_i$$

Từ kết quả ước lượng này, ta thu được các phần dư  $e_i$ .

**Bước 2:** Nếu ta nghi ngờ biến  $Z$  đã bị bỏ sót, sắp xếp các phần dư theo thứ tự tăng dần của biến  $Z$ , trường hợp không có số liệu của biến  $Z$  ta có thể sắp xếp  $e_i$  theo một trong các biến độc lập.

**Bước 3:** Tính  $d$ :

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

**Bước 4:** Kiểm định:

$H_0$ : Dạng hàm đúng (không có tự tương quan).

$H_1$ : Dạng hàm sai (có tự tương quan)

Dựa vào bảng Durbin-Watson với mức ý nghĩa để kết luận về  $H_0$ .

Thí dụ: các biến số:  $Y$ ,  $X_2$  và  $X_3$ ,  $n = 20$

Ta ước lượng được các mô hình sau:

$$\hat{Y}_t = 9,7702 + 0,52372 X_2 + 0,69302 X_3, R^2 = 0,7815$$

tính được các  $\hat{Y}_t$  và  $e_t$

- Kiểm định Ramsey 's Reset: trước hết ước lượng mô hình:

$$Y_t = \beta_0 + \beta_2 X_{2t} + \beta_3 X_{3t} + \alpha_2 \hat{Y}_t^2 + v_t$$

Ta thu được:

$$\hat{Y}_t = 7,7604 + 0,995 X_2 + 1,318 X_3 - 0,083405 \hat{Y}_t, R^2 = 0,78296$$

$$F = \frac{(R_{new}^2 - R_{old}^2) / m}{(1 - R_{new}^2) / (n - k)} = \frac{(0,78296 - 0,7815) / 1}{(1 - 0,78296) / (20 - 4)} = 0,133872$$

Trong khi đó:  $F_{0,05}(1,16) = 4,49$ .  $F = 0,133872 < F_{0,05}(1,16)$

Không bác bỏ giả thiết  $H_0$ .

### 3. Kiểm định về tính phân bố chuẩn của $U$

Để sử dụng các kiểm định  $T$ , kiểm định  $F$ ,  $\chi^2$ , trong hầu hết các trường hợp chúng ta giả thiết rằng các yếu tố ngẫu nhiên  $U_i$  có phân bố chuẩn. Do tổng thể chưa biết cho nên ta cũng không biết  $U_i$  và do đó cần phải thông qua  $e_i$  để đoán nhận.

Để kiểm định  $e_i$  có phân bố chuẩn hay không người ta có thể dùng  $\chi^2$ . Ngày nay hầu hết các phần mềm kinh tế lượng thường sử dụng kiểm định Jarque - Bera (JB)

$$JB = n \left[ \frac{S^2}{6} + \frac{(K-3)^2}{24} \right],$$

trong đó  $S$  là hệ số bất đối xứng,  $K$  là độ nhọn. Trong trường hợp tổng quát  $S$  và  $K$  được tính như sau:

$$S = \left[ \frac{\sum_{i=1}^n (X_i - \bar{X})^3 / n}{S_{X^3}} \right], K = \left[ \frac{\sum_{i=1}^n (X_i - \bar{X})^4 / n}{S_{X^4}} \right]$$

Với n khá lớn JB có phân bố xấp xỉ  $\chi^2(2)$ . Xét cặp giả thiết:

$H_0$ : U có phân bố chuẩn

$H_1$ : U không có phân bố chuẩn

$H_0$  sẽ bị bác bỏ nếu  $JB > \chi^2$ , trường hợp ngược lại không có cơ sở bác bỏ  $H_0$

Thí dụ: Khi ước lượng một mô hình hồi quy với  $n = 20$  ta thu được:

$$\sum e_i = 1,352, \sum e_i^3 / 20 = -0,0029656; \sum e_i^4 / 20 = 0,010699$$

Từ đó tính được:

$$S_e^2 = 1,352/20 = -0,0676; S_e = 0,26; S_e^3 = 0,017576; S_e^4 = 0,0045697$$

$$\text{Hệ số } S = (\sum e_i^3 / 20) / S_e^3 = -0,16876$$

$$K = (\sum e_i^4 / 20) / S_e^4 = 2,3412608$$

$$JB = 20(S^2/6 + (K-3)^2/24) = 0,4566$$

Với  $\alpha = 5\%$ ,  $\chi^2(2) = 3,84$ . trong trường hợp này không có cơ sở bác bỏ  $H_0$

## TÓM TẮT NỘI DUNG CHƯƠNG 8

Trong khi xem xét các mô hình hồi quy tuyến tính, ta ngầm giả định rằng mô hình đã chọn là mô hình đúng - Mô hình phản ánh đúng hiện tượng đang nghiên cứu. Nói cách khác là không có sai lầm trong việc chọn mô hình, không thiếu các biến cần thiết hoặc không thừa các biến không cần thiết trong mô hình. đây là mô hình chỉ định. Nhưng nếu có sai lầm trong việc chỉ định mô hình thì cần có biện pháp để phát hiện sai lầm. Trước hết cần hiểu thế nào là một mô hình tốt, đó là mô hình thỏa mãn 5 tiêu chuẩn của A.C. Harvy. Các loại sai lầm chỉ định bao gồm: Bỏ sót biến thích hợp; Đưa vào mô hình những biến không thích hợp; Dạng hàm không đúng.

Để phát hiện sự có mặt của các biến không cần thiết có thể sử dụng thống kê kiểm định t và F.

Để phát hiện các biến bị bỏ sót có thể sử dụng kiểm định RESET của RAMSEY, kiểm định Durbin-Watson hoặc kiểm định về tính phân bố chuẩn của U.

## CÂU HỎI VÀ BÀI TẬP ÔN CHƯƠNG 8

### I. Câu hỏi lý thuyết:

1. Giải thích 5 thuộc tính của một mô hình hồi quy tuyến tính tốt?
2. Trình bày các loại sai lầm thường gặp phải khi chọn mô hình?
3. Cách phát hiện các sai lầm và kiểm định/
4. Trình bày kiểm định giả thiết phân phối chuẩn của U?

## II. Bài tập:

1. Xét mối quan hệ giữa chi phí quảng cáo tiếp thị với mức cầu đối với sản phẩm dịch vụ của công ty trong thời kỳ 20 năm, người ta có số liệu sau:

Đơn vị: Triệu đồng

Năm	Chi tiêu cho quảng cáo, TT	Cầu về sản phẩm dịch vụ	Năm	Chi tiêu cho quảng cáo, TT	Cầu về sản phẩm dịch vụ
1	35,7	1551,3	11	247,1	2167,4
2	144,6	1599,8	12	277,9	2212,6
3	150,9	1668,1	13	253,6	2214,3
4	166,2	1728,4	14	258,7	2248,6
5	190,7	1797,4	15	249,5	2261,5
6	218,2	1916,3	16	282,2	2331,9
7	211,8	1896,9	17	251,1	2469,8
8	187,9	1931,7	18	367,9	2542,8
9	299,9	2001,0	19	412,3	2640,9
10	159,4	2066,6	20	439,0	2686,3

Giả sử mô hình đúng biểu thị quan hệ giữa chi tiêu cho quảng cáo tiếp thị và mức cầu của sản phẩm dịch vụ và biến xu thế có dạng:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + U_t \quad (1)$$

Trong đó:  $Y_t$ : Mức cầu về sản phẩm dịch vụ

$X_2$ : Chi tiêu cho quảng cáo tiếp thị.

$X_3$ : Biến xu thế (biểu thị thời gian hay xu thế), nó lấy giá trị từ 1, 2, ..., 20

Yêu cầu:

a) Nếu bỏ biến  $X_3$  và chọn hàm  $Y_t = \alpha_1 + \alpha_2 X_{2t} + V_t$  (2) để ước lượng. Hãy nhận xét về việc bỏ sót biến  $X_3$ .

b) Thay vì ước lượng hàm (1) ta lại chọn hàm dạng:

$$\ln Y_t = \alpha_1 + \alpha_2 \ln X_{2t} + \alpha_3 X_{3t} + V_t \quad (3)$$

Hãy nhận xét hậu quả của việc chọn hàm (3)?

2. Cho số liệu về tổng chi phí (Y) và sản lượng (X) ở bảng sau:

X	1	2	3	4	5	6	7	8	9	10
Y	193	226	240	244	257	260	274	297	350	420

Yêu cầu: Xác định hàm hồi quy tuyến tính của Y theo X. Sử dụng kiểm định RESET và Durbin Watson để kiểm định xem mô hình có bị chọn sai do thiếu biến Z hay không ?

3. Cho số liệu cho ở bảng sau:

<b>X</b>	70	65	90	95	110	115	120	140	155	150
<b>Y</b>	80	100	120	140	160	180	200	220	240	260

Biết hàm hồi quy tuyến tính mẫu có dạng:

$$\hat{Y}_i = 24,4545 + 0,5091X_i$$

Yêu cầu: Kiểm định giả thiết về phân phối chuẩn của U với  $\alpha = 5\%$  và  $\chi^2(2) = 3,84$ .

## MỤC LỤC

<b>LỜI NÓI ĐẦU .....</b>	<b>1</b>
<b>MỞ ĐẦU .....</b>	<b>3</b>
1. Khái quát về kinh tế lượng .....	3
2. Xây dựng và áp dụng mô hình kinh tế lượng: .....	3
<b>CHƯƠNG 1: CÁC KHÁI NIỆM CƠ BẢN CỦA MÔ HÌNH HỒI QUI HAI BIẾN .....</b>	<b>6</b>
GIỚI THIỆU .....	6
NỘI DUNG .....	6
1.1 PHÂN TÍCH HỒI QUI .....	6
1.2. BÀN CHẤT VÀ NGUỒN SỐ LIỆU CHO PHÂN TÍCH HỒI QUI .....	9
1.3 MÔ HÌNH HỒI QUI TỔNG THỂ .....	11
1.4 SAI SỐ NGẪU NHIÊN VÀ BÀN CHẤT .....	14
1.5 HÀM HỒI QUI MẪU: .....	15
TÓM TẮT NỘI DUNG CHƯƠNG 1 .....	15
CÂU HỎI VÀ BÀI TẬP ÔN CHƯƠNG I .....	16
I. Câu hỏi lý thuyết: .....	16
II. Bài tập: .....	16
<b>CHƯƠNG 2: ƯỚC LƯỢNG VÀ KIỂM ĐỊNH GIẢ THIẾT TRONG MÔ HÌNH HỒI QUI HAI BIẾN .....</b>	<b>18</b>
GIỚI THIỆU .....	18
NỘI DUNG .....	18
2.1 PHƯƠNG PHÁP BÌNH PHƯƠNG NHỎ NHẤT .....	18
2.2 CÁC GIẢ THIẾT CƠ BẢN CỦA PHƯƠNG PHÁP BÌNH PHƯƠNG NHỎ NHẤT .....	21
2.3 ĐỘ CHÍNH XÁC CỦA CÁC ƯỚC LƯỢNG BÌNH PHƯƠNG NHỎ NHẤT .....	22
2.4 HỆ SỐ $r^2$ ĐO ĐỘ PHÙ HỢP CỦA HÀM HỒI QUI MẪU .....	22
2.5 PHÂN BỐ XÁC SUẤT CỦA YẾU TỐ NGẪU NGHIÊN: .....	25
2.6 KHOẢNG TIN CẬY VÀ KIỂM TRA GIẢ THIẾT VỀ CÁC HỆ SỐ HỒI QUI .....	25
2.7 KIỂM ĐỊNH SỰ PHÙ HỢP CỦA HÀM HỒI QUI. PHÂN TÍCH HỒI QUI VÀ PHƯƠNG SAI .....	33
2.8. ỨNG DỤNG PHÂN TÍCH HỒI QUI: VẤN ĐỀ DỰ BÁO: .....	34
2.9 TRÌNH BÀY KẾT QUẢ PHÂN TÍCH HỒI QUI: .....	36
TÓM TẮT NỘI DUNG CHƯƠNG 2 .....	37
CÂU HỎI VÀ BÀI TẬP ÔN CHƯƠNG 2 .....	37
I- Câu hỏi .....	37
II- Bài tập: .....	38
<b>CHƯƠNG 3: MÔ HÌNH HỒI QUI NHIỀU BIẾN (HỒI QUI BỘ) .....</b>	<b>39</b>
GIỚI THIỆU .....	39
NỘI DUNG .....	39
3.1 MÔ HÌNH HỒI QUI TUYẾN TÍNH K BIẾN .....	39
3.2 CÁC GIẢ THIẾT: .....	40
3.3 ƯỚC LƯỢNG CÁC THAM SỐ - OLS .....	41
3.4. MA TRẬN HIỆP PHƯƠNG SAI CỦA CÁC ƯỚC LƯỢNG: .....	43
3.5 CÁC TÍNH CHẤT CỦA CÁC ƯỚC LƯỢNG BÌNH PHƯƠNG NHỎ NHẤT: .....	44
3.6 ƯỚC LƯỢNG HỢP LÝ TỐI ĐA (ML): .....	45
3.7 HỆ SỐ XÁC ĐỊNH BỘỊ VÀ HỆ SỐ XÁC ĐỊNH BỘỊ DẪ ĐIỀU CHỈNH: .....	45

3.8 MA TRẬN TƯƠNG QUAN:	46
3.9 HỆ SỐ TƯƠNG QUAN RIÊNG PHẦN:	47
3.10 KIỂM ĐỊNH GIẢ THIẾT VÀ KHOẢNG TIN CẬY CỦA CÁC HỆ SỐ HỒI QUY RIÊNG- KIỂM ĐỊNH T.	47
3.11 HỒI QUY CÓ ĐIỀU KIỆN RÀNG BUỘC - KIỂM ĐỊNH F:	49
3.12 DỰ BÁO:	49
3.13 Một số dạng của hàm hồi quy	50
TÓM TẮT NỘI DUNG CHƯƠNG 3	53
CÂU HỎI VÀ BÀI TẬP ÔN CHƯƠNG 3	53
I. Câu hỏi lý thuyết:	53
II. Bài tập:	53
<b>CHƯƠNG 4: HỒI QUY VỚI BIẾN ĐỘC LẬP LÀ BIẾN GIẢ</b>	<b>55</b>
GIỚI THIỆU	55
NỘI DUNG	55
4.1. BẢN CHẤT CỦA BIẾN GIẢ - MÔ HÌNH TRONG ĐÓ BIẾN GIẢ THÍCH LÀ BIẾN GIẢ	55
4.2. HỒI QUY VỚI MỘT BIẾN LƯỢNG VÀ MỘT BIẾN CHẤT	58
4.3 HỒI QUY VỚI MỘT BIẾN LƯỢNG VÀ HAI BIẾN CHẤT	62
4.4 SO SÁNH HAI HỒI QUY	63
4.5 ẢNH HƯỞNG CỦA TƯƠNG TÁC GIỮA CÁC BIẾN GIẢ	65
4.6. SỬ DỤNG BIẾN GIẢ TRONG PHÂN TÍCH MÙA	66
TÓM TẮT NỘI DUNG CHƯƠNG 4	67
CÂU HỎI VÀ BÀI TẬP ÔN CHƯƠNG 4	68
I. Câu hỏi lý thuyết:	68
II. Bài tập	68
<b>CHƯƠNG 5: ĐA CỘNG TUYẾN</b>	<b>71</b>
GIỚI THIỆU	71
NỘI DUNG	71
5.1 BẢN CHẤT CỦA ĐA CỘNG TUYẾN:	71
5.2 ƯỚC LƯỢNG KHI CÓ ĐA CỘNG TUYẾN HOÀN HẢO:	72
5.3 ƯỚC LƯỢNG TRONG TRƯỜNG HỢP CÓ ĐA CỘNG TUYẾN KHÔNG HOÀN HẢO:	73
5.4 HẬU QUẢ CỦA ĐA CỘNG TUYẾN:	73
5.5 CÁCH PHÁT HIỆN SỰ TỒN TẠI CỦA ĐA CỘNG TUYẾN:	75
5.6 BIỆN PHÁP KHẮC PHỤC:	77
TÓM TẮT NỘI DUNG CHƯƠNG 5	79
CÂU HỎI VÀ BÀI TẬP ÔN CHƯƠNG 5	79
I. Câu hỏi:	79
II. Bài tập:	80
<b>CHƯƠNG 6: PHƯƠNG SAI CỦA SAI SỐ THAY ĐỔI</b>	<b>81</b>
GIỚI THIỆU	81
NỘI DUNG	81
6.1. NGUYÊN NHÂN CỦA PHƯƠNG SAI CỦA SAI SỐ THAY ĐỔI	81
6.2. ƯỚC LƯỢNG BÌNH PHƯƠNG NHỎ NHẤT KHI PHƯƠNG SAI CỦA SAI SỐ THAY ĐỔI	82
6.3. PHƯƠNG PHÁP BÌNH PHƯƠNG NHỎ NHẤT TỔNG QUÁT	83
6.4. HẬU QUẢ CỦA PHƯƠNG SAI CỦA SAI SỐ THAY ĐỔI	86
6.5. PHÁT HIỆN PHƯƠNG SAI CỦA SAI SỐ THAY ĐỔI	89
6.6. BIỆN PHÁP KHẮC PHỤC	95
TÓM TẮT NỘI DUNG CHƯƠNG 6	99

CÂU HỎI VÀ BÀI TẬP ÔN CHƯƠNG 6 .....	99
I. Câu hỏi lý thuyết: .....	99
II. Bài tập: .....	100
<b>CHƯƠNG 7: TỰ TƯƠNG QUAN.....</b>	<b>101</b>
GIỚI THIỆU .....	101
NỘI DUNG .....	101
7.1 BẢN CHẤT VÀ NGUYÊN NHÂN CỦA HIỆN TƯỢNG TỰ TƯƠNG QUAN.....	101
7.2 ƯỚC LƯỢNG BÌNH PHƯƠNG NHỎ NHẤT KHI CÓ TỰ TƯƠNG QUAN.....	104
7.3 ƯỚC LƯỢNG TUYẾN TÍNH KHÔNG CHỆCH TỐT NHẤT KHI CÓ TỰ TƯƠNG QUAN. ....	105
7.4 HẬU QUẢ CỦA VIỆC SỬ DỤNG PHƯƠNG PHÁP BÌNH PHƯƠNG NHỎ NHẤT THƯỜNG KHI CÓ TỰ TƯƠNG QUAN. ....	106
7.5 PHÁT HIỆN CÓ TỰ TƯƠNG QUAN.....	106
7.6 CÁC BIỆN PHÁP KHẮC PHỤC.....	112
TÓM TẮT NỘI DUNG CHƯƠNG 7 .....	117
CÂU HỎI VÀ BÀI TẬP ÔN TẬP CHƯƠNG 7.....	117
I. Câu hỏi lý thuyết: .....	117
II. Bài tập: .....	117
<b>CHƯƠNG 8: CHỌN MÔ HÌNH VÀ KIỂM ĐỊNH VIỆC CHỈ ĐỊNH MÔ HÌNH .....</b>	<b>119</b>
GIỚI THIỆU .....	119
NỘI DUNG .....	119
8.1. CÁC THUỘC TÍNH CỦA MỘT MÔ HÌNH TỐT .....	119
8.2. CÁC LOẠI SAI LẦM CHỈ ĐỊNH. ....	120
8.3. PHÁT HIỆN NHỮNG SAI LẦM CHỈ ĐỊNH - CÁC KIỂM ĐỊNH VỀ SAI LẦM CHỈ ĐỊNH. ....	124
TÓM TẮT NỘI DUNG CHƯƠNG 8 .....	127
CÂU HỎI VÀ BÀI TẬP ÔN CHƯƠNG 8 .....	127
I. Câu hỏi lý thuyết: .....	127
II. Bài tập: .....	128
<b>MỤC LỤC.....</b>	<b>130</b>

# KINH TẾ LƯỢNG

Mã số: 497KTL370

Chịu trách nhiệm bản thảo

TRUNG TÂM ĐÀO TẠO BƯU CHÍNH VIỄN THÔNG 1