Chủ đề 2: Mô hình hồi quy tuyến tính đơn - Những vấn đề cơ bản

I. Bản chất của phân tích hồi qui

1. Khái niệm:

- Phân tích hồi qui là nghiên cứu sự phụ thuộc của một biến(biến phụ thuộc hay còn gọi là biến được giải thích) vào một hay nhiều biến khác(biến độc lập hay còn gọi là biến giải thích) với ý tưởng cơ bản là ước lượng(hay dự đoán) giá trị trung bình của biến phụ thuộc trên cơ sở các giá trị đã biết của biến độc lập.

- Một số ví dụ:

Vd1: Công ty địa ốc rất quan tâm đến việc liên hệ giữa giá bán một ngôi nhà với các đặc trưng của nó như kích thước, diện tích sử dụng, số phòng ngủ và phòng tắm, các loại thiết bị gia dụng, có hồ bơi hay không, cảnh quan có đẹp không,...

I. Bản chất của phân tích hồi qui

1. Khái niệm:

- Một số ví dụ:

Vd2: Cho đến nay việc hút thuốc lá là nguyên nhân chính gây tử vong do ung thư phổi được ghi chép cẩn thận. Một mô hình hồi qui tuyến tính đơn cho vấn đề này là:

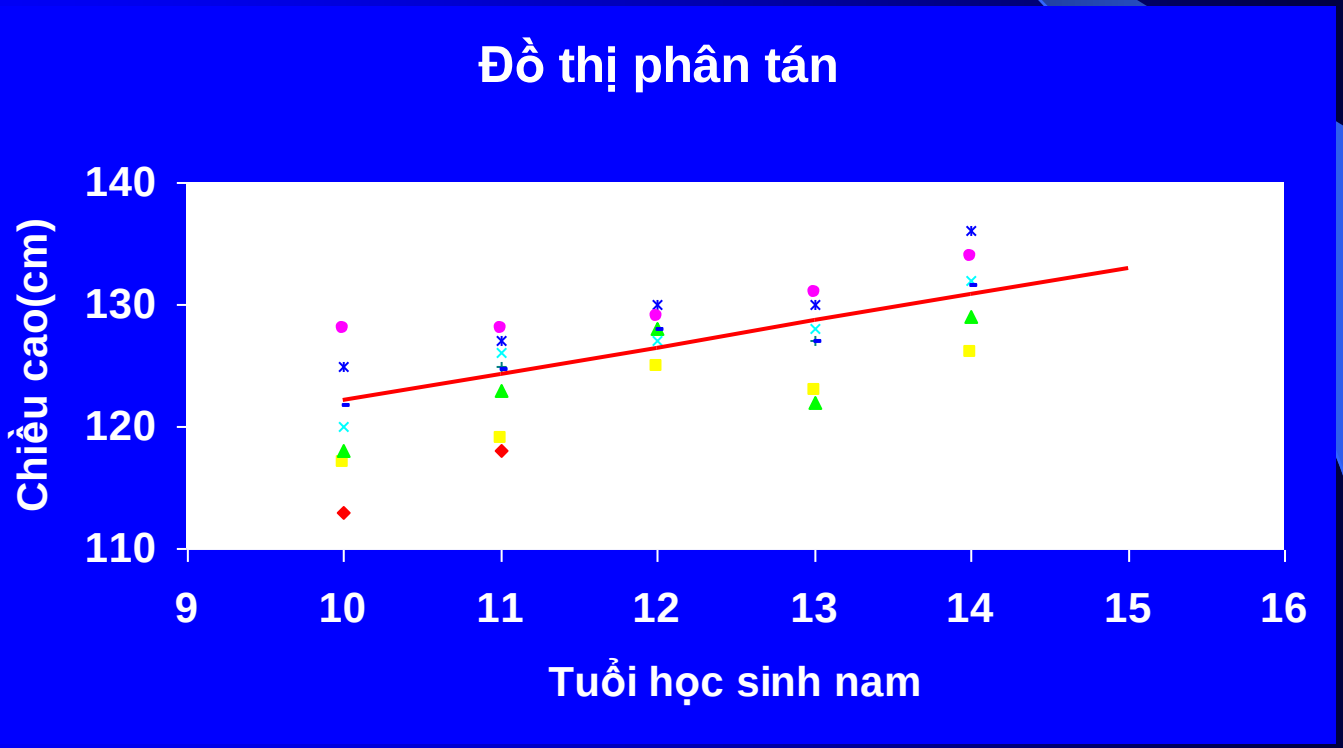
$$DEATHS = \alpha + \beta.SMOKING + u$$

I. Bản chất của phân tích hồi qui

1. Khái niệm:

- Một số ví dụ:

Vd3: Ta xem xét đồ thị phân tán sau đây mô tả phân phối về chiều cao của học sinh nam tính theo những độ tuổi cố định.



I. Bản chất của phân tích hồi qui

1. Khái niệm:

- Một số ví dụ:

- Vd4: Gám đốc tiếp thị của một công ty có thể 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 rất có ích cho việc xác định độ co dãn 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%. Kiến thức này rất có ích cho việc xác định ngân sách quảng cáo tối ưu.

- Vd5: Sau cùng một nhà nông học có thể quan tâm tới việc nghiên cứu sự phụ thuộc của sản lượng lúa vào nhiệt độ, lượng mưa, nắng, phân bón,...

I. Bản chất của phân tích hồi qui

1. Khái niệm:

Chúng ta có thể đưa ra vô số ví dụ như trên về sự phụ thuộc của một biến vào một hay nhiều biến khác. Các kỹ thuật phân tích hồi qui thảo luận trong chương này nhằm nghiên cứ sự phụ thuộc như thế giữa các biến số.

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

  X_j *- biến độc lập(hay biến giải thích) thứ j*

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

I. Bản chất của phân tích hồi qui

2. Phân tích hồi qui giải quyết các vấn đề sau:

- 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.

I. Bản chất của phân tích hồi qui

3. Phân biệt các quan hệ trong phân tích hồi qui:

- Quan hệ thống kê và quan hệ hàm số
- Hồi qui và nhân quả
- Hồi qui và tương quan

II.Hàm hồi qui đơn

1.Hàm hồi qui tổng thể:

Xét ví dụ giả định sau: Giả sử ở một địa phương có cả thảy 60 gia đình và chúng ta quan tâm đến việc nghiên cứu mối quan hệ giữa:

Y-Tiêu dùng trong tuần của các gia đình

X-Thu nhập khả dụng trong tuần của các hộ gia đình.

Các số liệu giả thuyết cho ở bảng sau:

II.Hàm hồi qui đơn

1.Hàm hồi qui tổng thể:

Thu nhập và chi tiêu trong một tuần của tổng thể

Y \ X	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
Tổng	325	462	445	707	678	750	685	1043	966	1211

II.Hàm hồi qui đơn

1.Hàm hồi qui tổng thể:

Các số liệu ở bảng trên được giải thích như sau:

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

II.Hàm hồi qui đơn

1.Hàm hồi qui tổng thể:

Từ số liệu cho ở bảng trên ta dễ dàng tính được các xác suất có điều kiện:

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

Từ đó ta có bảng các xác suất có điều kiện và kỳ vọng toán có điều kiện của Y điều kiện là $X=X_i$

Kỳ vọng toán có điều kiện(trung bình có điều kiện) của Y với điều kiện là $X=X_i$ được tính theo công thức sau:

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

II.Hàm hồi qui đơn

1.Hàm hồi qui tổng thể:

Xác suất có điều kiện P(Y/X) và kỳ vọng có điều kiện $E(Y/X_i)$

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
-	1/6	-	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

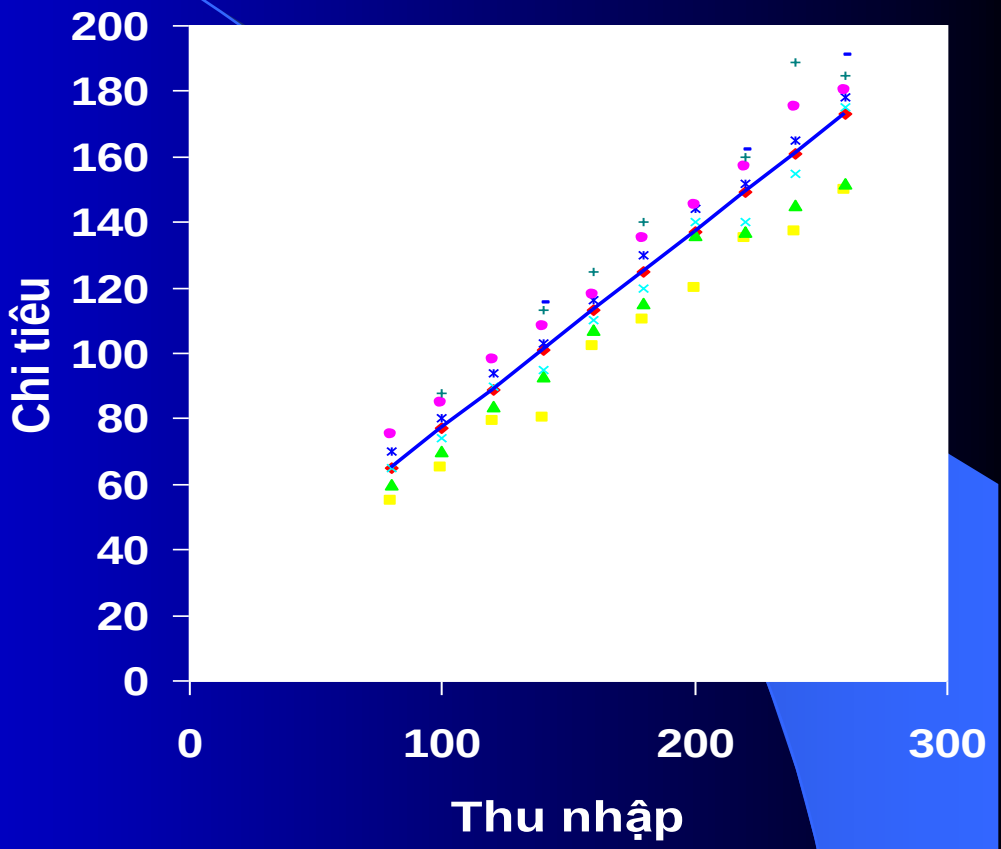
II.Hàm hồi qui đơn

1.Hàm hồi qui tổng thể:

- Biểu diễn các điểm $(X_i;Y_j)$ và các điểm $(X_i; E(Y/X_i))$ ta được đồ thị như hình bên.

Theo hình bên 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.

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



II.Hàm hồi qui đơn

1.Hàm hồi qui tổng thể:

Hàm (*) được gọi là hàm hồi qui tổng thể (PRF-Population Regression Function). Nếu PRF có một biến độc lập thì được gọi là *hồi qui đơn*, nếu có từ hai biến độc lập trở lên được gọi là *hồi qui bội*.

- Ý nghĩa của hàm PRF:

Hàm hồi qui tổng thể (PRF) 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 hàm của PRF 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 vấn đề nghiên cứu.

II.Hàm hồi qui đơn

1.Hàm hồi qui tổng thể:

- Ý nghĩa của hàm PRF:

Chúng ta xét trường hợp đơn giản nhất là PRF có dạng tuyến tính: $E(Y/X_i) = \beta_1 + \beta_2 X_i$.

Trong đó : β_1, β_2 là các tham số chưa biết nhưng cố định, và được gọi là các hệ số hồi qui.

-β_1: là hệ số tự do (hệ số tung độ góc). Nó cho biết giá trị trung bình của biến phụ thuộc Y bằng bao nhiêu khi biến độc lập X nhận giá trị 0. Điều này chỉ đúng về mặt lý thuyết, trong thực tế nhiều khi hệ số này không có ý nghĩa.

II.Hàm hồi qui đơn

1.Hàm hồi qui tổng thể:

- Ý nghĩa của hàm PRF:

-β_2: ***là hệ số góc*** (hệ số độ dốc) - 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.

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

II.Hàm hồi qui đơn

1.Hàm hồi qui tổng thể:

- Ý nghĩa của hàm PRF:

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

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

$E(Y/X_i) = \beta_1 + \beta_2^3 X_i$ là hàm tuyến tính đối với biến nhưng không tuyến tính với tham số.

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

II.Hàm hồi qui đơn

2. Sai số ngẫu nhiên và bản chất của nó.

Giả sử chúng ta đã có hàm hồi quy tổng thể $E(Y/X_i)$, vì $E(Y/X_i)$ là giá trị trung bình của biến Y với giá trị X_i đã biết, cho nên các giá trị cá biệt Y_i không phải bao giờ cũng trùng với $E(Y/X_i)$ mà chúng xoay quanh $E(Y/X_i)$.

Ta ký hiệu U_i là chênh lệch giữa giá trị cá biệt Y_i và $E(Y/X_i)$:

$$U_i = Y_i - E(Y/X_i) \text{ hay } Y_i = E(Y/X_i) + U_i \ (**)$$

U_i là đại lượng ngẫu nhiên, người ta gọi U_i là yếu tố ngẫu nhiên (hoặc nhiễu) và (**) được gọi là PRF ngẫu nhiên.

Nếu như $E(Y/X_i)$ là tuyến tính đối với X_i thì:

$$Y_i = \beta_1 + \beta_2 X_i + U_i$$

II.Hàm hồi qui đơn

2. Sai số ngẫu nhiên và bản chất của nó.

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

- Chúng ta có thể biết một cách chính xác biến giải thích X và biến phụ thuộc Y, nhưng chúng ta không biết hoặc biết không rõ về các biến khác ảnh hưởng đến Y. Vì vậy, U_i được sử dụng như yếu tố đại diện cho tất cả các biến không có trong mô hình.

- Ngay cả khi biết các biến bị loại khỏi mô hình là các biến nào, khi đó chúng ta có thể xây dựng mô hình hồi quy bội, nhưng có thể không có số liệu cho các biến này.

II.Hàm hồi qui đơn

2. Sai số ngẫu nhiên và bản chất của nó.

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

- Ngoài các biến giải thích đã có trong mô hình còn có một số biến khác nhưng ảnh hưởng của chúng đến Y rất nhỏ. Trong trường hợp này, chúng ta cũng sử dụng U_i đại diện cho chúng.

- Về mặt kỹ thuật và kinh tế, chúng ta mong muốn một mô hình đơn giản nhất có thể được. Nếu như chúng ta có thể giải thích được hành vi của biến Y bằng một số nhỏ nhất các biến giải thích và nếu như ta không biết tường minh những biến khác là biến nào có thể bị loại ra khỏi mô hình thì ta dùng yếu tố U_i để thay cho tất cả các biến này.

II.Hàm hồi qui đơn

3. Hàm hồi quy 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 trung bình của biến phụ thuộc Y từ số liệu của một mẫu.

Hàm hồi quy được xây dựng trên cơ sở của một mẫu được gọi là hàm hồi quy mẫu (SRF – The Sample Regression Function).

Nếu hàm PRF có dạng tuyến tính thì hàm hồi quy mẫu có dạng:

Trong đó : $\hat{\beta}_1$: là ước lượng điểm của β_1

$\hat{\beta}_2$: là ước lượng điểm của β_2

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

II.Hàm hồi qui đơn

3. Hàm hồi quy mẫu:

Dạng ngẫu nhiên của (***) là:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i \quad \text{Hay:} \quad Y_i = \hat{Y}_i + e_i$$

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