Chương 2: Mô hình hồi quy đơn

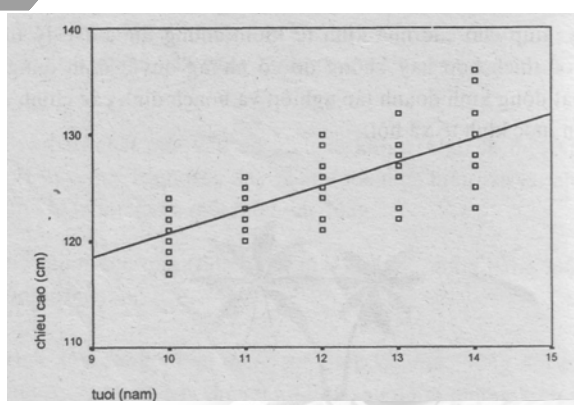
I. Bản chất của phân tích hồi quy:

1. Khái niệm:

   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) để 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:

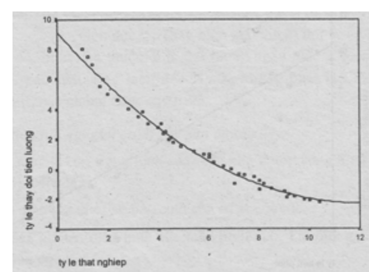
1- Quan hệ giữa chiều cao của học sinh nam tính theo những độ tuổi cố định



Hình 2.1: Phân phối giả thiết về chiều cao theo độ tuổi

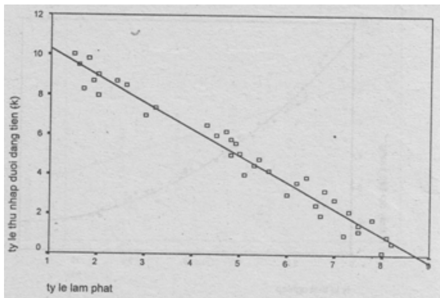
2- Sự phụ thuộc của chi tiêu cho tiêu dùng vào thu nhập thực tế

3- Tỷ lệ thay đổi tiền lương trong mối quan hệ với tỷ lệ thất nghiệp



Hình 2.2: Đường cong Phillips giả thiết

4- Mức lạm phát và tỷ lệ thu nhập người dân giữ dưới dạng tiền mặt



Hình 2.3: Lượng tiền được giữ trong quan hệ với lạm phát

5- 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.

6- Một nhà nông học quan tâm tới việc nghiên cứu sự phụ thuộc của sản lượng lúa vào nhiệt độ, lượng mưa, nắng, phân bón…

⇒Ký hiệu:

Y – Biến phụ thuộc (biến được giải thích)

X – Biến giải thích (biến độc lập)

2. Các mối quan hệ trong phân tích hồi quy

a. Quan hệ thống kê và quan hệ hàm số:

- Quan hệ thống kê thể hiện ở sự phụ thuộc thống kê của biến phụ thuộc vào các biến giải thích.

⇒ Biến phụ thuộc là đại lượng ngẫu nhiên có phân phối xác suất

⇒ Các biến giải thích có giá trị biết trước

⇒ Ứng với mỗi giá trị của biến giải thích có thể có nhiều giá trị khác nhau của biến giải thích

- Quan hệ hàm số:

⇒ Các biến không phải là ngẫu nhiên

⇒ Ứng với mỗi giá trị của biến giải thích có một giá trị của biến phụ thuộc

⇒ Phân tích hồi quy không nghiên cứu các quan hệ hàm số

Ví dụ:

Sự phụ thuộc của năng suất lúa vào nhiệt độ, lượng mưa, lượng phân bón … là một quan hệ thống kê

Tính chu vi hình vuông bằng 4 lần chiều dài y = 4x là quan hệ hàm số

b. Hồi quy và quan hệ nhân quả:

Phân tích hồi quy nghiên cứu quan hệ phụ thuộc của Y vào X

=> Không đòi hỏi giữa Y và X phải có quan hệ 2 chiều (nhân quả)

c. Hồi quy và tương quan:

- Phân tích tương quan đo mức độ kết hợp tuyến tính giữa hai biến
- Phân tích hồi quy ước lượng, dự báo một biến trên cơ sở giá trị đã cho của các biến khác
- Trong phân tích hồi quy, khác với tương quan, các biến không có tính đối xứng

3. Nguồn số liệu cho phân tích hồi quy

3.1. Các loại số liệu:

Gồm: Số liệu theo thời gian (chuỗi TG), số liệu chéo và số liệu hỗn hợp

- Số liệu theo TG: là các số liệu được thu thập trong một thời kỳ nhất định
- Số liệu chéo: là các số liệu được thu thập tại một thời điểm, thời kỳ ở nhiều địa phương, đơn vị khác nhau.
- Số liệu hỗn hợp theo thời gian và không gian

4.2. Nguồn số liệu:

- Do các cơ quan nhà nước, tổ chức quốc tế, công ty hoặc cá nhân thu thập
- Gồm các số liệu thực nghiệm hoặc phi thực nghiệm

4.3. Nhược điểm của số liệu:

- Hầu hết số liệu trong khoa học xã hội là các số liệu phi thực nghiệm
- Các số liệu thực nghiệm có thể có sai số trong phép đo
- Trong điều tra có thể không nhận được câu trả lời hoặc không trả lời hết
- Các mẫu điều tra có cỡ mẫu khác nhau nên khó khăn trong so sánh kết quả các cuộc điều tra
- Các số liệu kinh tế thường ở mức tổng hợp cao không cho phép đi sâu vào các đơn vị nhỏ
- Số liệu bí mật quốc gia khó tiếp cận

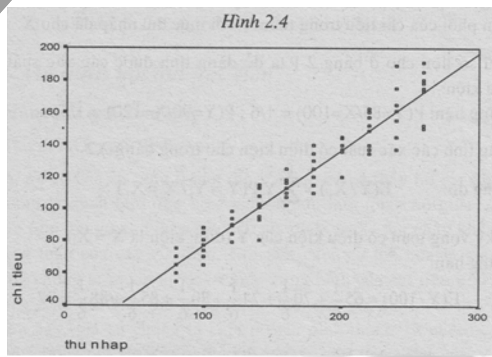
II. Các khái niệm cơ bản trong hồi quy đơn

1. Hàm hồi quy tổng thể:

   Ví dụ 2: Nghiên cứu sự phụ thuộc của Y – chi tiêu tiêu dùng hàng tuần và X – thu nhập khả dụng hàng tuần của các gia đình ở một địa phương có 60 gia đình.

Mức TN	**80**	**100**	**120**	**140**	**160**	**180**	**200**	**220**	**240**	**260**
Các mức chi tiêu	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
TB	*65*	*77*	*89*	*101*	*113*	*125*	*137*	*149*	*161*	*173*

Bảng 2.1. Ví dụ về thu nhập và chi tiêu của 60 hộ gia đình



⇒Trung bình có điều kiện của mức chi tiêu trong tuần nằm trên đường thẳng có hệ số góc dương: $E(Y/Xi) = \sum Y_j P(Y = Y_j/X = X_i)$

⇒ $E(Y/Xi)$ là một hàm của X: $E(Y/Xi) = f(Xi)$: Hàm hồi quy tổng thể PRF

⇒Hàm PRF cho biết giá trị trung bình của Y khi biến X nhận một giá trị nhất định

⇒Để xác định dạng của hàm hồi quy tổng thể người ta dựa vào đồ thị biểu diễn biến thiên kết hợp với phân tích bản chất của vấn đề nghiên cứu

⇒Nếu PRF có dạng tuyến tính: $E(Y/X_i) = \beta_1 + \beta_2 X_i$

β_1 là hệ số tự do, cho biết giá trị trung bình của Y khi X bằng 0

β_2 là hệ số góc, cho biết giá trị trung bình của biến Y sẽ thay đổi bao nhiêu đơn vị khi X tăng một đơn vị

CM: $X'_i = X_i + 1$. Khi đó: $E(Y/X'_i) = \beta_1 + \beta_2 X'_i$

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

⇒"Tuyến tính" được hiểu theo hai nghĩa:

Tuyến tính đối với tham số: $E(Y/X_i) = \beta_1 + \beta_2 X^2_i$

Tuyến tính đối với biến: $E(Y/X_i) = \beta_1 + \sqrt{\beta_2} X_i$

Hàm hồi quy tuyến tính được hiểu là tuyến tính đối với các tham số

2. Sai số ngẫu nhiên:

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

  U_i là đại lượng ngẫu nhiên và được gọi là sai số ngẫu nhiên.
- U_i tồn tại vì các lý do sau:

  U_i được sử dụng như yếu tố đại diện cho tất cả các biến giải thích không được đưa vào mô hình

- Các biến không biết rõ
- Các biến không có số liệu
- Các biến có ảnh hưởng rất nhỏ
- Các biến không được đưa vào vì lý do muốn có một mô hình đơn giản nhất có thể

3. Hàm hồi quy mẫu:

Hàm hồi quy được xây dựng trên cơ sở một mẫu được gọi là hàm hồi quy mẫu SRF

Bảng 2.3. Mẫu thứ nhất

Y	70	65	90	95	110	115	120	140	155	150
X	80	100	120	140	160	180	200	220	240	260

Bảng 2.4. Mẫu thứ hai

Y	55	88	90	80	118	120	145	157
X	80	100	120	140	160	180	200	220

⇒Từ hai mẫu xây dựng được hai hàm hồi quy mẫu là SRF1 và SRF2

⇒Hàm hồi quy mẫu tuyến tính có dạng: $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

⇒Dạng ngẫu nhiên: $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i = \hat{Y}_i + e_i$

III. **Ước lượng và kiểm định giả thiết trong hồi quy đơn**

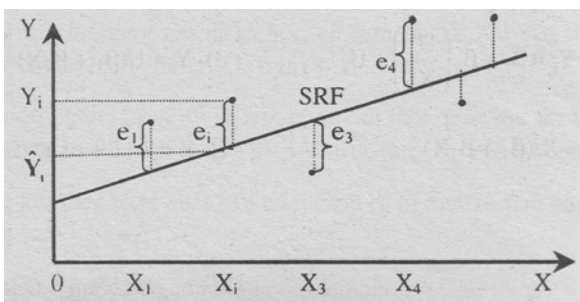
1. Phương pháp bình phương nhỏ nhất OLS

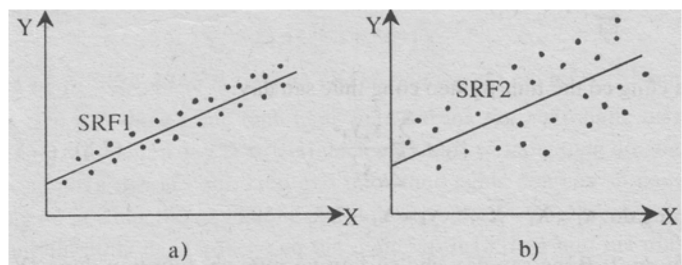
Giả sử có mẫu gồm n cặp quan sát (Y_i, X_i), i = 1..n.

Cần tìm hàm $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ sao cho càng sát với giá trị thực càng tốt

⇒ Tức là: $e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$ min

⇒ Do e_i có thể dương hoặc âm nên ta lấy tổng bình phương của e_i đạt min: $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_2)^2$ min





Do Y_i, X_i đã biết nên $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_2)^2$ là hàm của $\hat{\beta}_1, \hat{\beta}_2$

=> Ta có: $f(\hat{\beta}_1, \hat{\beta}_2) = \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_2)^2 \min$

$$\begin{cases} \dfrac{\partial f(\hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_1} = \sum 2\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)(-1) = 0 \\ \dfrac{\partial f(\hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_2} = \sum_{i=1}^{n} 2\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)(-X_i) = 0 \end{cases} \Leftrightarrow \begin{cases} n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \\ \hat{\beta}_1 \sum_{i=1}^{n} X_i + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i X_i \end{cases}$$

$$\Rightarrow \hat{\beta}_2 = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n(\bar{X})^2}; \hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

$$x_i = X_i - \bar{X}; y_i = Y_i - \bar{Y} \Rightarrow \hat{\beta}_2 = \frac{\sum y_i x_i}{\sum x_i^2}$$

Ví dụ 2: Bảng sau cho số liệu về mức chi tiêu tiêu dùng (Y-USD/tuần) và thu nhập hàng tuần (X-USD/tuần) của 10 gia đình. Hãy ước lượng hàm hồi quy tuyến tính 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

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

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

$$\Rightarrow \hat{\beta}_2 = \frac{\sum Y_i X_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n(\bar{X})^2} = \frac{205500 - 10x170x111}{322000 - 10x(170)^2} = 0,5091$$

$$\hat{\beta}_1 = 1111 - 0,5091x170 = 24,4545$$

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

2. Các giả thiết của mô hình hồi quy tuyến tính

Chất lượng của các ước lượng phụ thuộc:

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

⇒ Các giả thiết liên quan đến X_i và U_i gồm:

- GT1: Biến giải thích là phi ngẫu nhiên
- GT2: Kỳ vọng của yếu tố ngẫu nhiên U_i bằng 0, tức là: $E(U_i/X_i) = 0$
- GT3: Các U_i có phương sai bằng nhau: $Var(U_i/X_i) = Var(U_j/X_j) = \sigma^2$
- GT4: Không có sự tương quan giữa các U_i: $Cov(U_i,U_j)=0$
- GT5: U_i và X_i không tương quan với nhau: $Cov(U_i,X_j)=0$

⇒ Định lý Gauss-Markov: Với các giả thiết 1-5 các ước lượng OLS là các ước lượng tuyến tính, không chệch và có phương sai nhỏ nhất.

3. Phương sai và sai số chuẩn của các ước lượng

Các ước lượng hệ số tự do và hệ số góc là đại lượng ngẫu nhiên, với các mẫu khác nhau ta có các giá trị ước lượng khác nhau.

$$\mathrm{var}\left(\hat{\beta}_1\right) = \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2; se\left(\hat{\beta}_1\right) = \sqrt{\mathrm{var}\left(\hat{\beta}_1\right)};$$

$$\mathrm{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2}; se(\hat{\beta}_2) = \sqrt{\mathrm{var}(\hat{\beta}_2)}$$

Trong đó: $\sigma^2 = \mathrm{var}(U_i)$ và được ước lượng bằng: $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$

$TSS = \sum y_i^2 = \sum\left(Y_i - \bar{Y}\right)^2$

4. Hệ số xác định và hệ số tương quan

TSS là tổng bình phương của tất cả các sai lệch giữa Y_i với giá trị trung bình

$$TSS = \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \sum Y_i^2 - n\left(\bar{Y}\right)^2$$

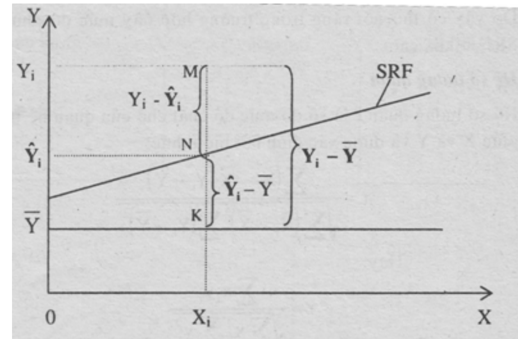
ESS là tổng bình phương của tất cả các sai lệch giữa giá trị của biến Y tính theo hàm hồi quy mẫu với giá trị trung bình

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

RSS 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 của nó từ hàm hồi quy mẫu

$$RSS = \sum e_i^2 = \sum\left(Y_i - \hat{Y}_i\right)^2$$

TSS = ESS + RSS



Hình 2.6

- Hệ số xác định: R^2 = ESS/TSS => đo mức độ phù hợp của hàm hồi quy =>

$$R^2 = \frac{\left(\sum x_i y_i\right)^2}{\left(\sum x_i^2\right)\left(\sum y_i^2\right)}$$

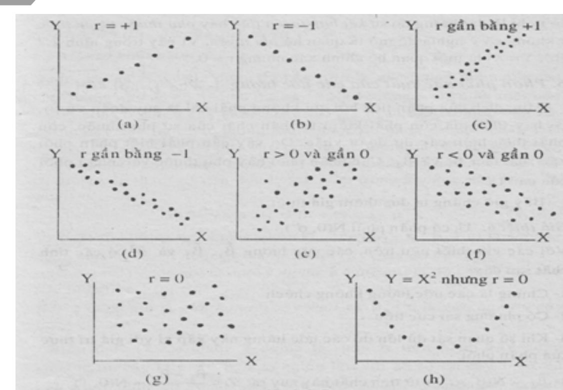
- $0 \le R^2 \le 1$

=>Nếu tất cả các giá trị quan sát của Y nằm trên SRF thì RSS = 0, ESS = TSS và $R^2 = 1$(hàm hồi quy rất phù hợp)

=>Nếu hàm hồi quy kém phù hợp thì RSS càng lớn và R^2 tiến tới 0

- Vd2: $\sum Y_i^2 = 132100$, TSS = 132100 – 10.$(111)^2$ = 8890, ESS = $(0{,}509091)^2$.33000 = 8552,73 => R^2 = 8552,73/8890 = 0,9621
- Hệ số tương quan: đo mức độ chặt chẽ của quan hệ tuyến tính giữa X và Y:

$$r = \frac{\sum x_i y_i}{\sqrt{\left(\sum x_i^2\right)\left(\sum y_i^2\right)}} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}; r = \pm\sqrt{R^2}$$



Hình 2.7

5. Khoảng tin cậy của β_1, β_2 và σ^2

5.1. Một số khái niệm:

- Ước lượng điểm có thể không phải là giá trị thực

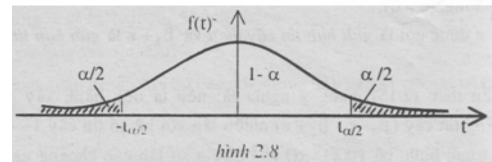
=> xây dựng một khoảng xung quan giá trị ước lượng điểm: $P\left(\hat{\beta}_2 - \varepsilon \le \beta_2 \le \hat{\beta}_2 + \varepsilon\right) = 1 - \alpha$

- Khoảng $\left(\hat{\beta}_2 - \varepsilon; \hat{\beta}_2 + \varepsilon\right)$: khoảng ngẫu nhiên; 1 – α: hệ số tin cậy; α (0 < α < 1): mức ý nghĩa, ε: độ chính xác của ước lượng.

  $\hat{\beta}_2 - \varepsilon$: giới hạn dưới; $\hat{\beta}_2 + \varepsilon$:giới hạn trên

5.2 Khoảng tin cậy của β_2

- Chứng minh được: $t = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} \sim T(n-2)$
- Thiết lập khoảng tin cậy: $P(-t_{\alpha/2} \le t \le t_{\alpha/2}) = 1 - \alpha$ trong đó $t_{\alpha/2}$ thoả mãn: $P(|t| < t_{\alpha/2}) = 1 - \alpha$
- Minh hoạ:



hình 2.8

- Thay t vào: $P(-t_{\alpha/2}(n-2) \le \frac{\hat{\beta}_2 - \beta_2}{se\left(\hat{\beta}_2\right)} \le t_{\alpha/2}(n-2)) = 1 - \alpha$

  $\Rightarrow P\left(\hat{\beta}_2 - t_{\alpha/2} se(\hat{\beta}_2) \le \beta_2 \le \hat{\beta}_2 + t_{\alpha/2} se(\hat{\beta}_2)\right) = 1 - \alpha$

- Với hệ số tin cậy 1 – α, khoảng tin cậy của β_2 là:

  $\left(\hat{\beta}_2 - t_{\alpha/2} se(\hat{\beta}_2); \hat{\beta}_2 + t_{\alpha/2} se(\hat{\beta}_2)\right)$

5.3. Khoảng tin cậy của β_1

- Tương tự: $\Rightarrow P\left(\hat{\beta}_1 - t_{\alpha/2} se(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2} se(\hat{\beta}_1)\right) = 1 - \alpha$
- Khoảng tin cậy của β_1 là:

  $\left(\hat{\beta}_1 - t_{\alpha/2} se(\hat{\beta}_1); \hat{\beta}_1 + t_{\alpha/2} se(\hat{\beta}_1)\right)$

- Để tìm $t_{\alpha/2}$ ta tra bảng ở phần phụ lục hoặc dùng hàm trong excel. Vd: với số bậc tự do là n – 2 = 8, α = 5% thì $t_{0,025}$ = TINV(0,05,8) = 2,306
- Vd2:

  RSS = TSS – ESS = 8890 – 8552,73 = 337,27

  $\hat{\sigma}^2 = \frac{337,27}{10-2} = 42,15875; var(\hat{\beta}_2) = \frac{42,15875}{33000} = 0,0012775$

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

  $Var(\hat{\beta}_1) = \frac{322000}{10x33000} 42,15875 = 41,13672$

  $se(\hat{\beta}_1) = \sqrt{var(\hat{\beta}_2)} = \sqrt{41,13672} = 6,4138; t_{\alpha/2}(n-2) = t_{0,025}(8) = 2,306$

  $\Rightarrow \beta_1 \in (24,4545 \pm 2,306x6,4138) \Leftrightarrow 9,6643 < \beta_1 < 39,2448$

  $\Rightarrow \beta_2 \in (0,5091 \pm 2,306x0,035742) \Leftrightarrow 0,4268 < \beta_2 < 0,5914$

5.4. Khoảng tin cậy của σ^2:

- CM được:

$$\chi^2 = (n-2)\frac{\hat{\sigma}_2}{\sigma^2} \sim \chi^2(n-2) \Rightarrow P\left(\chi^2_{1-\alpha/2} \le (n-2)\frac{\hat{\sigma}^2}{\sigma^2} \le \chi^2_{\alpha/2}\right) = 1-\alpha$$

$$\Rightarrow P\left(\frac{(n-2)\hat{\sigma}^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-2)\hat{\sigma}^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha$$

- Để tìm các giá trị này tra bảng phần phụ lục hoặc dùng hàm CHIINV của excel: CHIINV(0,025,7)=16,0128

6. Kiểm định giả thiết về các hệ số hồi quy:

- KĐGT nhằm trả lời câu hỏi: “Kết quả tìm được dựa trên số liệu thu thập có phù hợp với một giả thiết nêu ra hay không?”
- Có hai cách KĐGT: Dựa vào khoảng tin cậy và dựa vào kiểm định ý nghĩa.

6.1. Kiểm định giả thiết - Phương pháp khoảng tin cậy:

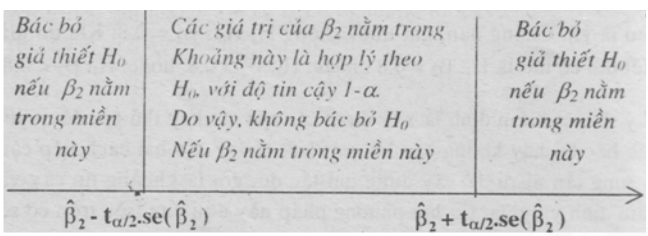
- Từ số liệu của Vd 2, kiểm định GT: H_0: $\beta_2 = 0{,}3$ với H_1: $\beta_2 \neq 0{,}3$.

=> Căn cứ vào khoảng tin cậy, ta thấy: $0{,}4268 < \beta_2 < 0{,}5914$

$\Rightarrow$Quy tắc KĐ:

Thiết lập một khoảng tin cậy với hệ số tin cậy 1 – α cho β_2.

Nếu β_2 nằm trong khoảng này thì không bác bỏ H_0; ngược lại nằm ngoài thì bác bỏ H_0



6.2. Kiểm định giả thiết: Phương pháp kiểm định ý nghĩa

- KĐGT: H_0: $\beta_2 = \beta^*_2$ với H1: $\beta_2 \neq \beta^*_2$
- Ta đã có: $P(-t_{\alpha/2} \le \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} \le t_{\alpha/2}) = 1-\alpha$

- Nếu $\beta_2 = \beta^*_2$ thì: $P(-t_{\alpha/2} \le \frac{\hat{\beta}_2 - \beta^*_2}{se(\hat{\beta}_2)} \le t_{\alpha/2}) = 1-\alpha$

$\Rightarrow$ Như vậy: $(-t_{\alpha/2}; t_{\alpha/2})$ được gọi là ***miền chấp nhận***;

$\Rightarrow$ Vùng nằm ngoài được gọi là ***miền bác bỏ***;

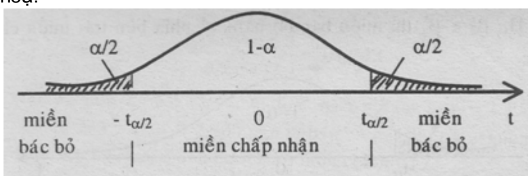
$\Rightarrow$ $t_{\alpha/2}$: ***giá trị tới hạn***; α: mức ý nghĩa của kiểm định.

$\Rightarrow$ Quy tắc quyết định:

- Tính $t = \left(\hat{\beta}_2 - \beta^*_2\right) / se\left(\hat{\beta}_2\right)$
- Nếu t thuộc khoảng $(-t_{\alpha/2}; t_{\alpha/2})$ thì chấp nhận H_0
- Nếu t ngoài khoảng $(-t_{\alpha/2}; t_{\alpha/2})$ thì bác bỏ H_0

$\Rightarrow$ Do sử dụng phân phối t nên thủ tục KĐ này được gọi là kiểm định t

⇒ Minh hoạ:



Hình 2.9

⇒ VD2: H_0: $\beta_2 = 0,3$ với H_1: $\beta_2 \neq 0,3$.
Số bậc tự do là n - 2 = 8; với α = 5% tra bảng ta có $t_{\alpha/2}$ = 2,306. Vậy miền chấp nhận H_0 là -2,306 < t < 2,306.

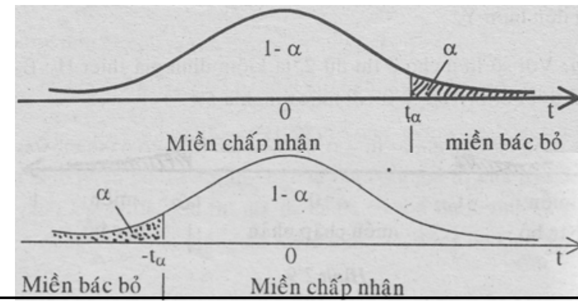
$$t = \left(\hat{\beta}_2 - \beta^*_2\right) \Big/ se\left(\hat{\beta}_2\right) = (0,509091 - 0,3)/0,035742 = 5,85$$

Vì giá trị của t nằm ở miền bác bỏ nên ta bác bỏ giả thiết H_0.

- Kiểm định một phía:
  H_0: $\beta_2 = \beta^*_2$ với H_1: $\beta_2 < \beta^*_2$ hoặc $\beta_2 > \beta^*_2$

⇒ Nếu H_1: $\beta_2 > \beta^*_2$ thì miền bác bỏ nằm bên phải;

⇒ Nếu H_1: $\beta_2 < \beta^*_2$ thì miền bác bỏ nằm bên trái



Tóm tắt quy tắc KĐGT với β_2:

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

Tương tự ta có quy tắc KĐGT với β_1:

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

- KĐGT: H_0: $\beta_2 = 0$ với H_1: $\beta_2 \neq 0$

⇒ kiểm định GT cho rằng biến X không ảnh hưởng tới biến Y

VD2:

KĐGT H_0: $\beta_2 = 0$ với H_1: $\beta_2 \neq 0$ với α = 5%

t = (0,509091 – 0)/0,035742=14,243

$t_{0,025}$ = 2,306

t > $t_{0,025}$ nên bác bỏ H_0.

⇒ biến thu nhập X có ảnh hưởng thực sự tới biến chi tiêu Y

6.3. Kiểm định giả thiết về σ^2

KĐGT: H_0: $\sigma^2 = \sigma^2_0$; H_1: $\sigma^2 \neq \sigma^2_0$ với mức ý nghĩa α

$\Rightarrow$ Quy tắc KĐ: Tính $\chi^2 = (n-2)\dfrac{\hat{\sigma}_2}{\sigma^2}$

Loại giả thiết	Giả thiết H_0	Giả thiết H_1	Miền bác bỏ
Hai phía	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\chi^2 > \chi^2_{\alpha/2}$ hoặc $\chi^2 < \chi^2_{1-\alpha/2}$
Phía phải	$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\chi^2 > \chi^2_{\alpha}$
Phía trái	$\sigma^2 = \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\chi^2 < \chi^2_{1-\alpha}$

$\Rightarrow$ VD2: KĐGT: H_0: $\sigma^2 = 85$; H_1: $\sigma^2 \neq 85$ với $\alpha = 5\%$

Ta đã có $\hat{\sigma}^2 = 42,15875$. Vậy $\chi^2 = (10 - 2).42,15875/85 = 3,968$

=> không thuộc miền bác bỏ nên ta chấp nhận H_0

$\chi^2_{\alpha/2} = \chi^2_{0,025}(8) = CHIINV(0.025,8) = 17,5345; \chi^2_{1-\alpha/2} = \chi^2_{0,975}(8) = CHIINV(0.975,8) = 2,1797$

7. Kiểm định sự phù hợp của hàm hồi quy

- CM được: $F = \dfrac{\left(\hat{\beta}_2 - \beta_2\right)^2 \sum x_i^2}{\hat{\sigma}^2} \sim F(1, n-2)$
- Kđ sự phù hợp: H_0: $R^2 = 0$; H_1: $R^2 > 0$

  <=> H_0: $\beta_2 = 0$; H_1: $\beta_2 \neq 0$.
- Quy tắc kđ:

  Tính $F = \dfrac{\left(\hat{\beta}_2\right)^2 \sum x_i^2}{\hat{\sigma}^2} = \dfrac{R^2(n-2)}{1-R^2}$

  Nếu $F > F_\alpha(1, n-2)$ thì bác bỏ H_0
- Vd2: H_0: $\beta_2 = 0$; H_1: $\beta_2 \neq 0$.

  $F=R^2(n-2)/(1-R^2)=0,96206(10-2)/(1-0,96206) = 202,86$

$\Rightarrow$ giá trị p tương ứng với F rất nhỏ (<0,0005) nên bác bỏ H_0.

8. Dự báo

- Vd2: Ta có hàm HQ mẫu: $\hat{Y}_i = 24,4545 + 0,5091X_i$

$\Rightarrow$ Có 2 loại dự báo:

+ Dự báo trung bình có điều kiện của Y với $X = X_0$;

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

- Dự báo giá trị trung bình: $E(Y/X_0) = \beta_1 + \beta_2 X_0$

$\Rightarrow$ Ước lượng điểm không chệch, có phương sai nhỏ nhất của $E(Y/X_0)$ là: $\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$

$\Rightarrow$ $\hat{Y}_0$ có phân phối chuẩn với kỳ vọng $\beta_1 + \beta_2 X_0$ và phương sai: $Var(\hat{Y}_0) = \sigma^2\left[\dfrac{1}{n} + \dfrac{\left(X_0 - \bar{X}\right)^2}{\sum x_i^2}\right]$

$\hat{\sigma}^2$

$\Rightarrow$ σ^2 chưa biết nên sử dụng UL không chệch của nó là $\hat{\sigma}^2$

$\Rightarrow$ Ta có:

$$t = \frac{\hat{Y}_0 - E(Y/X_0)}{se(\hat{Y}_0)} \sim T(n-2)$$

$$\Rightarrow P\left[-t_{\alpha/2} \leq \frac{\hat{Y}_0 - E(Y/X_0)}{se(\hat{Y}_0)} \leq t_{\alpha/2}\right] = 1-\alpha$$

$$\Rightarrow P\left[\hat{Y}_0 - t_{\alpha/2}se(\hat{Y}_0) \leq E(Y/X_0) \leq \hat{Y}_0 + t_{\alpha/2}se(\hat{Y}_0)\right] = 1-\alpha$$

$$\Rightarrow \hat{Y}_0 - t_{\alpha/2}se(\hat{Y}_0) \leq E(Y/X_0) \leq \hat{Y}_0 + t_{\alpha/2}se(\hat{Y}_0)$$

- Dự báo giá trị riêng biệt:

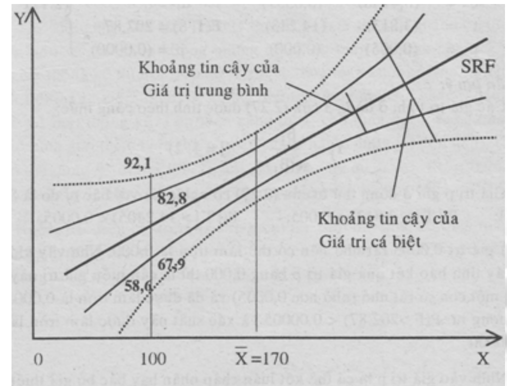
⇒Ước lượng của Y_0 là $\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$

⇒Phương sai của Y_0:

$$Var(Y_0) = \sigma^2 \left[1 + \frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{\sum x_i^2} \right]$$

⇒Khoảng tin cậy của Y_0: $\hat{Y}_0 - t_{\alpha/2} se(Y_0) \le Y_0 \le \hat{Y}_0 + t_{\alpha/2} se(Y_0)$

⇒Vd2:



Hình 2.10

9. Đánh giá các kết quả của phân tích HQ

Các tiêu chí đánh giá:

- Tiêu chí 1: dấu của các hệ số hồi quy có phù hợp với lý thuyết không?
- Tiêu chí 2: các hệ số hồi quy phải có ý nghĩa về mặt thống kê.
- Tiêu chí 3: Mô hình giải thích sự biến thiên của biến phụ thuộc tốt đến đâu => dùng R^2.
- Tiêu chí 4: Kiểm tra xem mô hình có thoả mãn các giả thiết của mô hình hồi quy tuyến tính không?