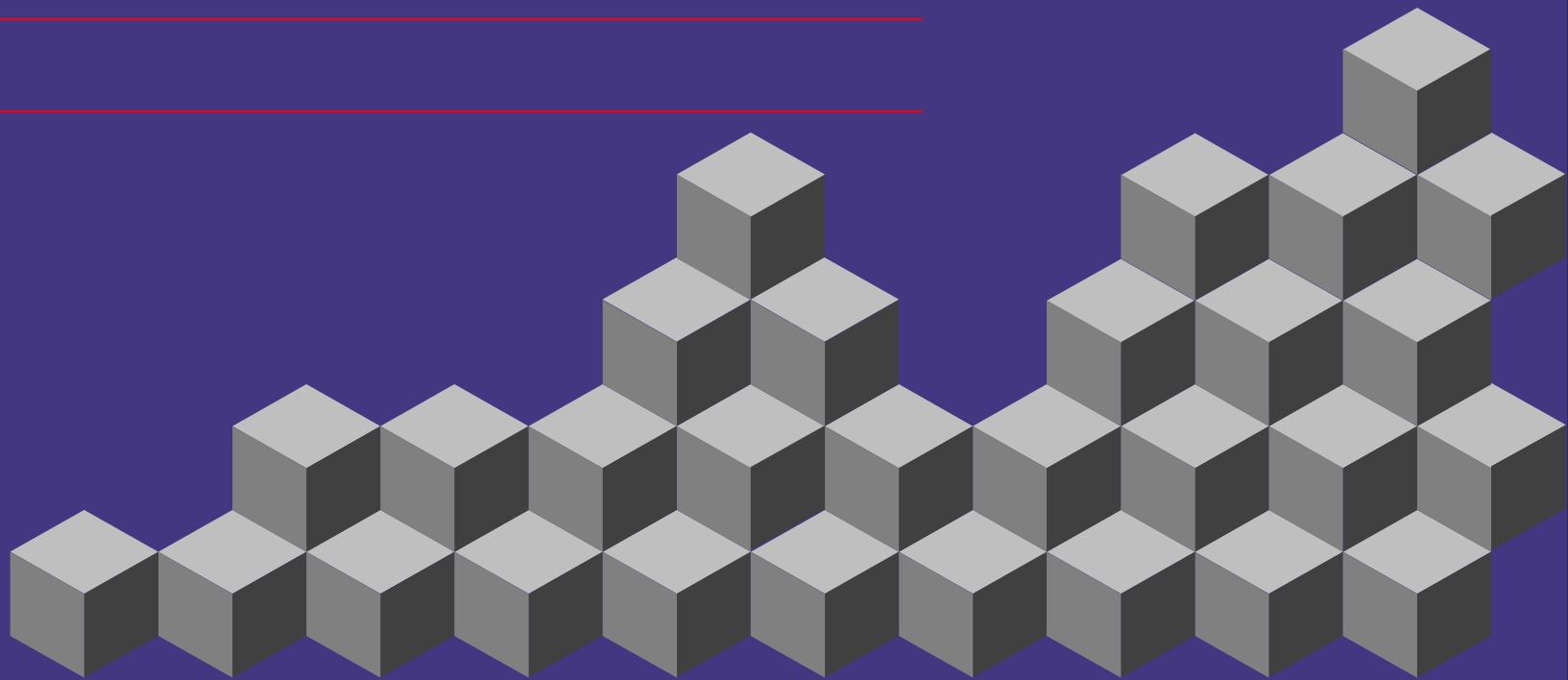
CHƯƠNG 2

HỒI QUY ĐƠN BIẾN



HỒI QUY ĐƠN BIẾN

MỤC TIÊU

1. Biết được phương pháp ước lượng bình phương nhỏ nhất để ước lượng hàm hồi quy tổng thể dựa trên số liệu mẫu
2. Hiểu các cách kiểm định những giả thiết
3. Sử dụng mô hình hồi quy để dự báo

NỘI DUNG

1. Mô hình
2. Phương pháp bình phương nhỏ nhất (OLS)
3. Khoảng tin cậy
4. Kiểm định giả thiết
5. Dự báo

2.1 MÔ HÌNH

Mô hình hồi quy tuyến tính hai biến

PRF dạng xác định

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

dạng ngẫu nhiên

- $Y_i = E(Y/X_i) + U_i = \beta_1 + \beta_2 X_i + U_i$

SRF *dạng xác định*

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

- *dạng ngẫu nhiên*

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

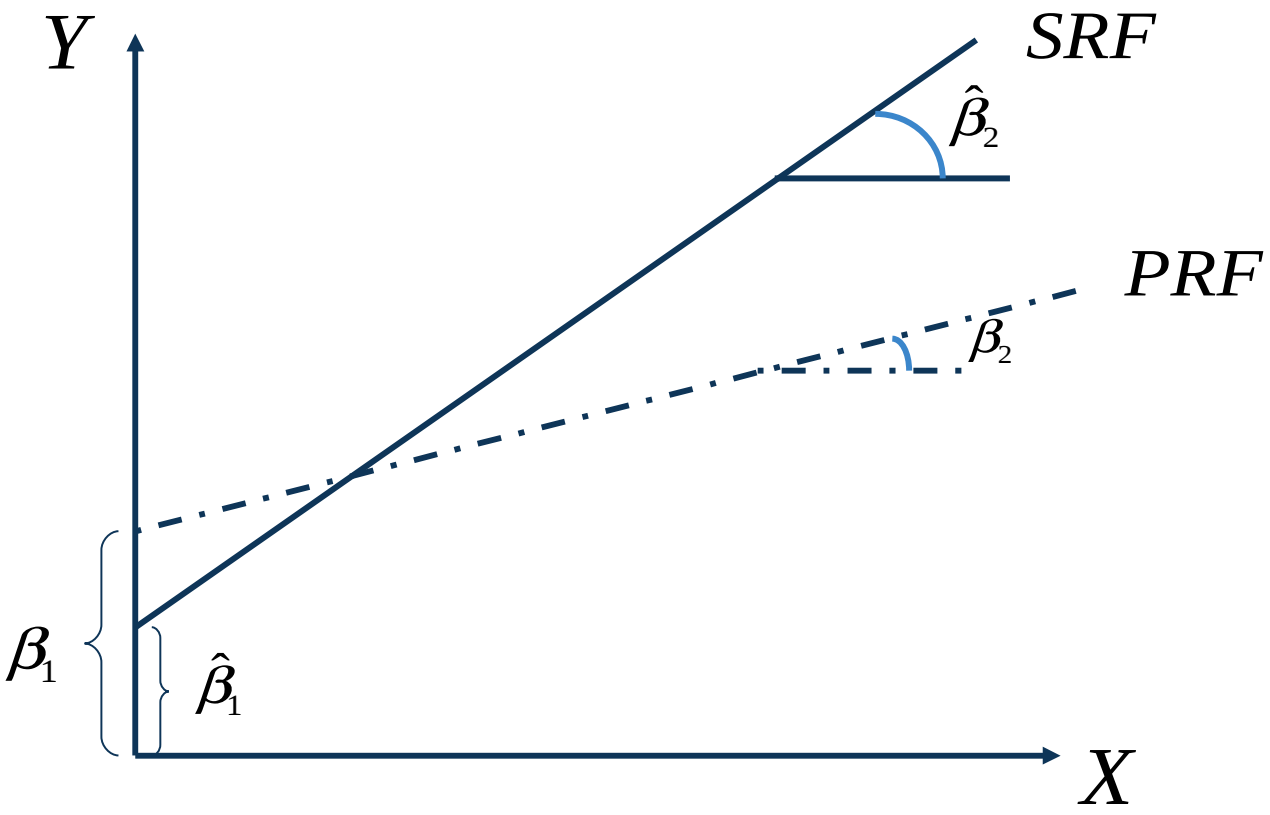
2.1 MÔ HÌNH

Trong đó

- $\hat{\beta}_1$: Ước lượng cho β_1
- $\hat{\beta}_2$: Ước lượng cho β_2
- $\hat{Y}_i$: Ước lượng cho E(Y/Xi)

- Sử dụng phương pháp bình phương nhỏ nhất thông thường (OLS) để tìm $\hat{\beta}_1$, $\hat{\beta}_2$

2.1 MÔ HÌNH



Hình 2.1: Hệ số hồi quy trong hàm hồi quy PRF và SRF

2.2 PHƯƠNG PHÁP OLS

Giả sử có n cặp quan sát (X_i, Y_i). Tìm giá trị $\hat{Y}_i$ sao cho $\hat{Y}_i$ gần giá trị Y_i nhất, tức $e_i = |Y_i - \hat{Y}_i|$ càng nhỏ càng tốt.

- Hay, với n cặp quan sát, muốn

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

2.2 PHƯƠNG PHÁP OLS

- Bài toán thành tìm $\hat{\beta}_1$, $\hat{\beta}_2$ sao cho f → min

Điều kiện để đạt cực trị là:

$$\frac{\partial\left(\sum_{i=1}^{n} e_i^2\right)}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right) = 0$$

$$\frac{\partial\left(\sum_{i=1}^{n} e_i^2\right)}{\partial \hat{\beta}_2} = -2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right) X_i = 0$$

2.2 PHƯƠNG PHÁP OLS

Hay

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

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

2.2 PHƯƠNG PHÁP OLS

- Giải hệ, được

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

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

$$x_i = X_i - \overline{X}$$

$$y_i = Y_i - \overline{Y}$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

2.2 PHƯƠNG PHÁP OLS

- Với

$$\bar{Y} = \frac{\sum Yi}{n} \qquad \bar{X} = \frac{\sum Xi}{n}$$

là trung bình mẫu (theo biến)

$$x_i = X_i - \bar{X} \qquad y_i = Y_i - \bar{Y}$$

gọi là độ lệch giá trị của biến so với giá trị trung bình mẫu

CÁC TỔNG BÌNH PHƯƠNG ĐỘ LỆCH

- TSS (Total Sum of Squares - Tổng bình phương sai số tổng cộng)

$$TSS = \sum (Y_i - \overline{Y})^2 = \sum Y_i^2 - n.(\overline{Y})^2 = \sum y_i^2$$

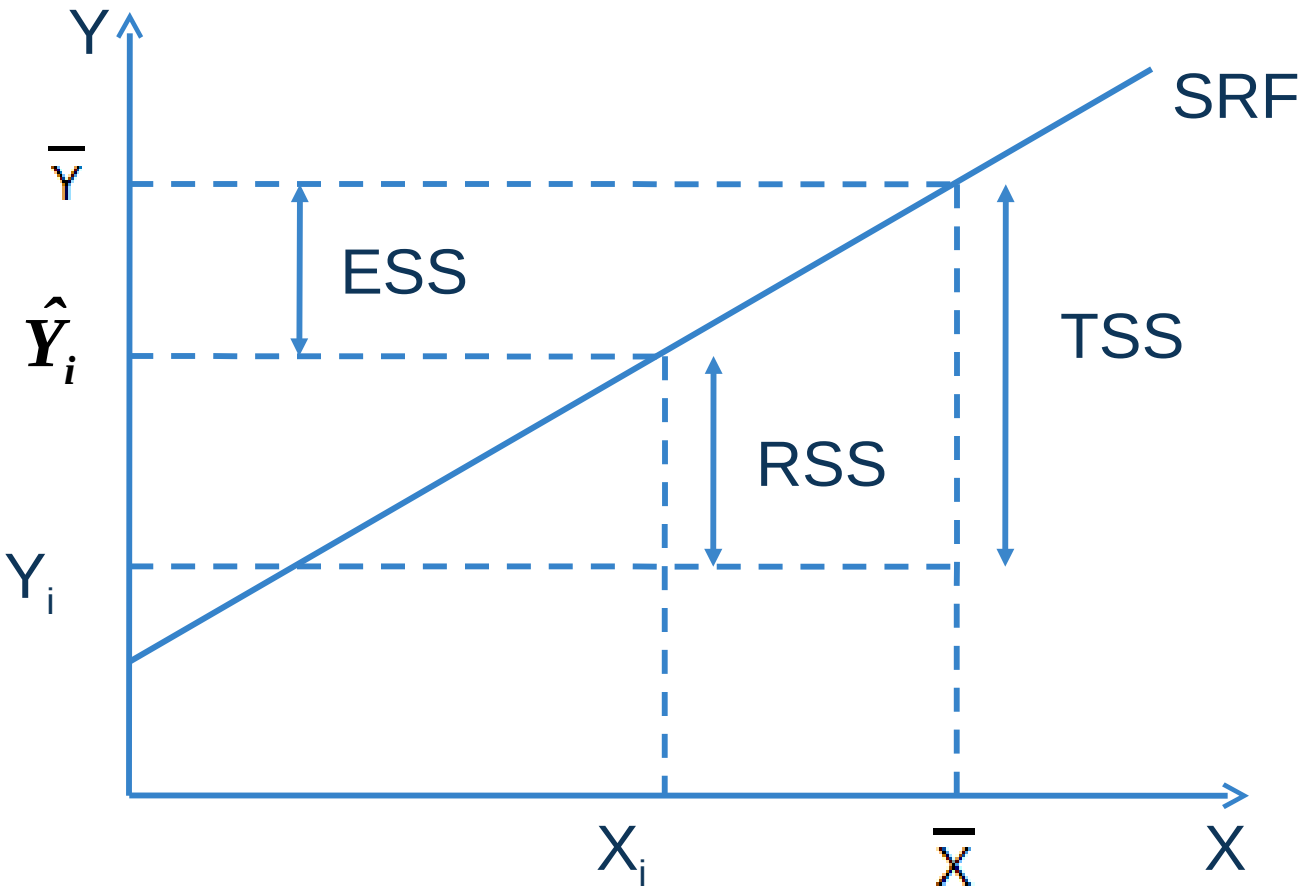
- ESS: (Explained Sum of Squares - Bình phương sai số được giải thích)

$$ESS = \sum (\hat{Y}_i - \overline{Y})^2 = (\hat{\beta})^2 \sum x_i^2$$

- RSS: (Residual Sum of Squares - Tổng bình phương sai số)

$$RSS = \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum y_i^2 - \hat{\beta}_2^2 \sum x_i^2$$

CÁC TỔNG BÌNH PHƯƠNG ĐỘ LỆCH



Hình 2.3: Ý nghĩa hình học của TSS, RSS và ESS

HỆ SỐ XÁC ĐỊNH R^2

Ta chứng minh được: TSS = ESS + RSS

hay $$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

- Hàm SRF phù hợp tốt với các số liệu quan sát (mẫu) khi $\hat{Y}_i$ gần Y_i . Khi đó ESS lớn hơn RSS.
- Hệ số xác định R^2: đo mức độ phù hợp của hàm hồi quy mẫu.

HỆ SỐ XÁC ĐỊNH R^2

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

Trong mô hình 2 biến

$$R^2 = \frac{\hat{\beta}_2^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2}$$

TÍNH CHẤT CỦA HỆ SỐ XÁC ĐỊNH R^2

$0 \leq R^2 \leq 1$

Cho biết % sự biến động của Y được giải thích bởi các biến số X trong mô hình.

$R^2 = 1$: đường hồi quy phù hợp hoàn hảo

$R^2 = 0$: X và Y không có quan hệ

Nhược điểm: R^2 tăng khi số biến X đưa vào mô hình tăng, dù biến đưa vào không có ý nghĩa.

=>Sử dụng R^2 điều chỉnh (adjusted R^2 - $\overline{R}^2$) để quyết định đưa thêm biến vào mô hình.

HỆ SỐ XÁC ĐỊNH ĐIỀU CHỈNH $\overline{R}2$

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

- Khi đưa thêm biến vào mô hình mà $\overline{R}^2$ tăng thì nên đưa biến vào và ngược lại.

HỆ SỐTƯƠNG QUAN r

Hệ số tương quan r: đo mức độ chặt chẽ của quan hệ tuyến tính giữa 2 đại lượng X và Y.

$$r = \frac{\sum_{i=1}^{n} y_i x_i}{\sqrt{\sum_{i=1}^{n} y_i^2 \sum_{i=1}^{n} x_i^2}}$$

TÍNH CHẤT HỆ SỐTƯƠNG QUAN r

$-1 \leq r \leq 1$

Có tính chất đối xứng: $r_{XY} = r_{YX}$

Nếu X, Y độc lập theo quan điểm thống kê thì hệ số tương quan giữa chúng bằng 0.

r đo sự kết hợp tuyến tính hay phụ thuộc tuyến tính, không có ý nghĩa để mô tả quan hệ phi tuyến.

HỆ SỐTƯƠNG QUAN r

Có thể chứng minh được

$$r = \pm\sqrt{R^2}$$

và r cùng dấu với $\hat{\beta}_2$

VD: $\hat{Y}_i = 6{,}25 + 0{,}75X_i$

Với $R^2 = 0{,}81 \Rightarrow r = 0{,}9$

2.3 Các giả thiết của phương pháp OLS

- Giả thiết 1: Các giá trị X_i được xác định trước và không phải là đại lượng ngẫu nhiên
- Giả thiết 2: Kỳ vọng hoặc trung bình số học của các sai số là bằng 0 (zero conditional mean), nghĩa là $E(U/X_i) = 0$

2.3 Các giả thiết của phương pháp OLS

- Giả thiết 3: Các sai số U có phương sai bằng nhau (homoscedasticity)

$$Var(U/X_i) = \sigma^2$$

- Giả thiết 4: Các sai số U không có sự tương quan, nghĩa là

$$Cov(U_i, U_{i'}) = E(U_i U_{i'}) = 0, \text{ nếu } i \neq i'$$

2.3 Các giả thiết của phương pháp OLS

- Giả thiết 5: Các sai số U độc lập với biến giải thích

$$Cov(U_i, X_i) = 0$$

- Giả thiết 6: Đại lượng sai số ngẫu nhiên có phân phối chuẩn $U_i \sim N(0, \delta^2)$

Định lý Gauss-Markov

Định lý: Với những giả thiết (từ 1 đến 5) của mô hình hồi quy tuyến tính cổ điển, mô hình hồi quy tuyến tính theo phương pháp bình phương nhỏ nhất là ước lượng tuyến tính không chệch tốt nhất.

2.4 TÍNH CHẤT CÁC ƯỚC LƯỢNG OLS

$\hat{\beta}_1$, $\hat{\beta}_2$ được xác định một cách duy nhất với n cặp giá trị quan sát (X_i , Y_i)

$\hat{\beta}_1$, $\hat{\beta}_2$ là các đại lượng ngẫu nhiên, với các mẫu khác nhau, giá trị của chúng sẽ khác nhau

Đo lường độ chính xác các ước lượng bằng ***sai số chuẩn*** (standard error – se).

Sai số chuẩn của các ước lượng OLS

var: phương sai

se: sai số chuẩn

σ^2: phương sai nhiễu của tổng thể

$$\sigma^2 = Var(U_i)$$

-> thực tế khó biết được giá trị σ^2 -> dùng ước lượng không chệch

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$$

Sai số chuẩn của các ước lượng OLS

$\hat{\beta}_1$ — $\text{var}(\hat{\beta}_1) = \dfrac{\sum X_i^2}{\sum n} \cdot \dfrac{\sigma^2}{x_i^2}$ — $se(\hat{\beta}_1) = \sqrt{\text{var}(\hat{\beta}_1)}$

$\hat{\beta}_2$ — $\text{var}(\hat{\beta}_2) = \dfrac{\sigma^2}{\sum x_i^2}$ — $se(\hat{\beta}_2) = \sqrt{\text{var}(\hat{\beta}_2)}$

Sai số chuẩn của các ước lượng OLS

$$\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{n-2}}$$

Sai số chuẩn của hồi quy: là độ lệch tiêu chuẩn các giá trị Y quanh đường hồi quy mẫu

Tính chất đường hồi quy mẫu SRF

1. SRF đi qua trung bình mẫu $\overline{Y} = \hat{\beta}_1 + \hat{\beta}_2 \overline{X}$
2. $\overline{\hat{Y}} = \overline{Y}$
3. $\sum e_i = 0$
4. $\sum \hat{Y} e_i = 0$
5. $\sum e_i X_i = 0$

2.4 KHOẢNG TIN CẬY CỦA HỆ SỐ HỒI QUY

Xác suất của khoảng $(\hat{\beta}_i - \varepsilon_i, \hat{\beta}_i + \varepsilon_i)$ chứa giá trị thực của β_i là $1 - \alpha$ hay:

$$P(\hat{\beta}_i - \varepsilon_i \leq \beta_i \leq \hat{\beta}_i + \varepsilon_i) = 1 - \alpha.$$

với

$$\varepsilon_i = t_{(\alpha/2, n-2)} SE(\hat{\beta}_i)$$

2.4 KHOẢNG TIN CẬY CỦA HỆ SỐ HỒI QUY

- $(\hat{\beta}_i - \varepsilon_i, \hat{\beta}_i + \varepsilon_i)$: khoảng tin cậy,
- ε_i : độ chính xác của ước lượng,
- $1 - \alpha$: hệ số tin cậy,
- α $(0 < \alpha < 1)$: mức ý nghĩa,
- $t(\alpha/2, n-2)$: giá trị tới hạn (tìm bằng cách tra bảng số t-student)
- n: số quan sát

2.4 KHOẢNG TIN CẬY CỦA σ^2

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

hay

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

$\chi^2_{1-\alpha/2}$ $\chi^2_{\alpha/2}$: giá trị của đại lượng ngẫu nhiên phân phối theo quy luật χ^2 với bậc tự do n-2 thỏa điều kiện

$$P(\chi^2 > \chi^2_{1-\alpha/2}) = 1-\alpha; P(\chi^2 > \chi^2_{\alpha/2}) = \alpha/2$$

2.5 KIỂM ĐỊNH GIẢ THIẾT

1. Kiểm định giả thiết về hệ số hồi quy

- **Hai phía:**

$$H_0 : \beta_i = \beta_i^*$$

$$H_1 : \beta_i \neq \beta_i^*$$

- **Phía phải:**

$$H_0 : \beta_i \leq \beta_i^*$$

$$H_1 : \beta_i > \beta_i^*$$

- **Phía trái:**

$$H_0 : \beta_i \geq \beta_i^*$$

$$H_1 : \beta_i < \beta_i^*$$

1. Kiểm định giả thiết về hệ số hồi quy

$$H_0 : \beta_i = \beta_i^*$$

$$H_1 : \beta_i \neq \beta_i^*$$

Cách 1: Phương pháp giá trị tới hạn

Bước 1: Tính t

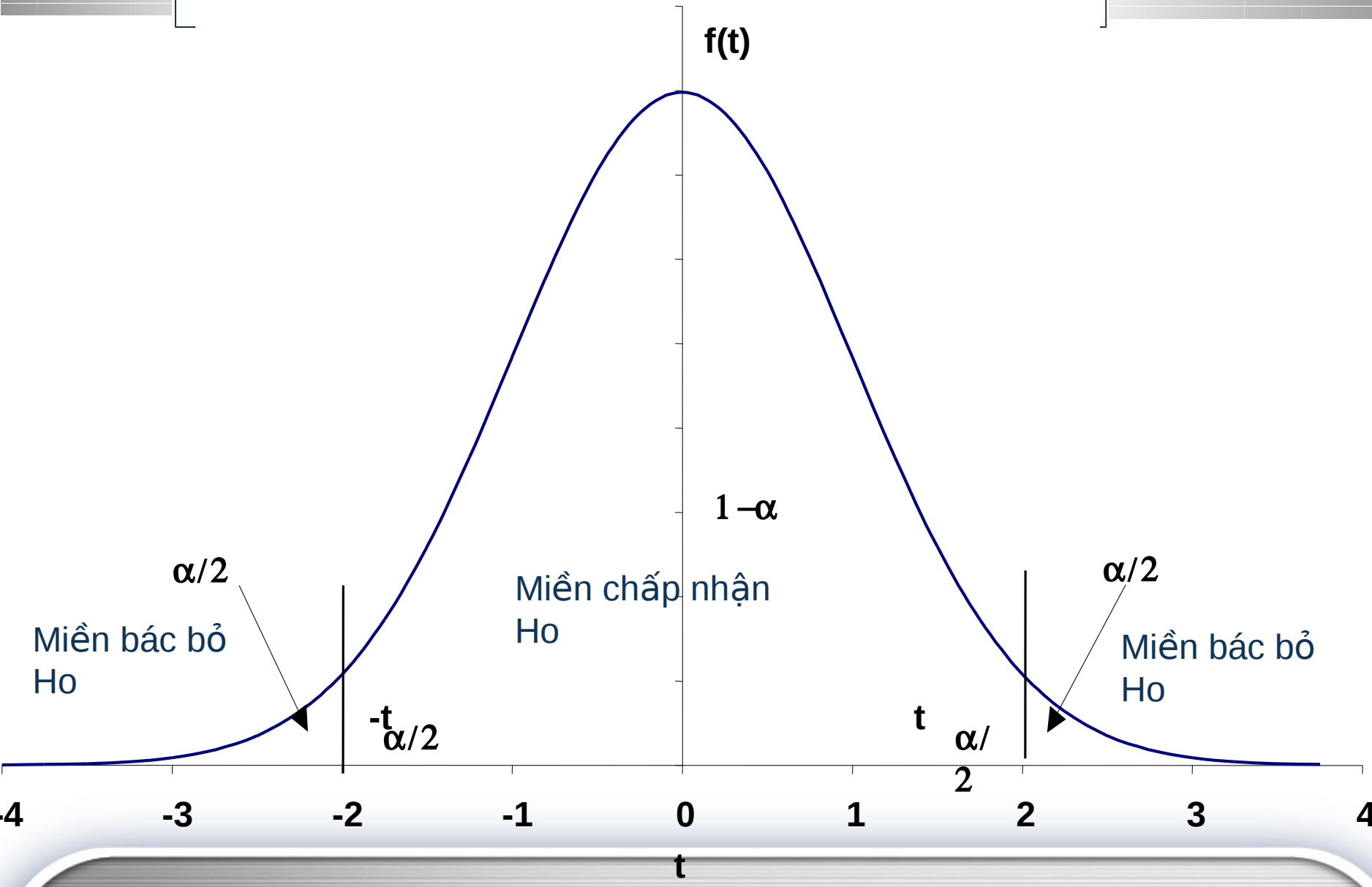
$$t = \frac{\hat{\beta}_2 - \beta_2^*}{SE(\hat{\beta}_2)}$$

Bước 2: Tra bảng t-student để có giá trị tới hạn $t_{(n-2,\alpha/2)}$

Bước 3: Quy tắc quyết định

Nếu $|t| > t_{(n-2,\alpha/2)}$ bác bỏ H_0.

Nếu $|t| \leq t_{(n-2,\alpha/2)}$ chấp nhận H_0.


f(t)
1 –α
α/2
Miền bác bỏ
Ho
Miền chấp nhận
Ho
α/2
Miền bác bỏ
Ho
-t
α/2
t
α/
2
-4
-3
-2
-1
0
1
2
3
4
t

Cách 2: Phương pháp khoảng tin cậy

Khoảng tin cậy của β_i:

$$\beta_i \in (\hat{\beta}_i - \varepsilon_i ; \hat{\beta}_i + \varepsilon_i) \quad \varepsilon_i = t_{(n-2,1-\alpha/2)} SE(\hat{\beta}_i)$$

với mức ý nghĩa α trùng với mức ý nghĩa của H_0

Quy tắc quyết định

- Nếu $\beta_i^* \in (\hat{\beta}_i - \varepsilon_i ; \hat{\beta}_i + \varepsilon_i)$ chấp nhận H_0
- Nếu $\beta_i^* \notin (\hat{\beta}_i - \varepsilon_i ; \hat{\beta}_i + \varepsilon_i)$ bác bỏ H_0

1. Kiểm định giả thiết về hệ số hồi quy

Cách 3: Phương pháp p-value

Bước 1: Tính

$$t_i = \frac{\hat{\beta}_i - \beta_i^*}{SE(\hat{\beta}_i)}$$

Bước 2: Tính $P(T > |t_i|) = p$

Bước 3: Quy tắc quyết định

- Nếu $p \leq \alpha$: Bác bỏ H_0
- Nếu $p > \alpha$: Chấp nhận H_0

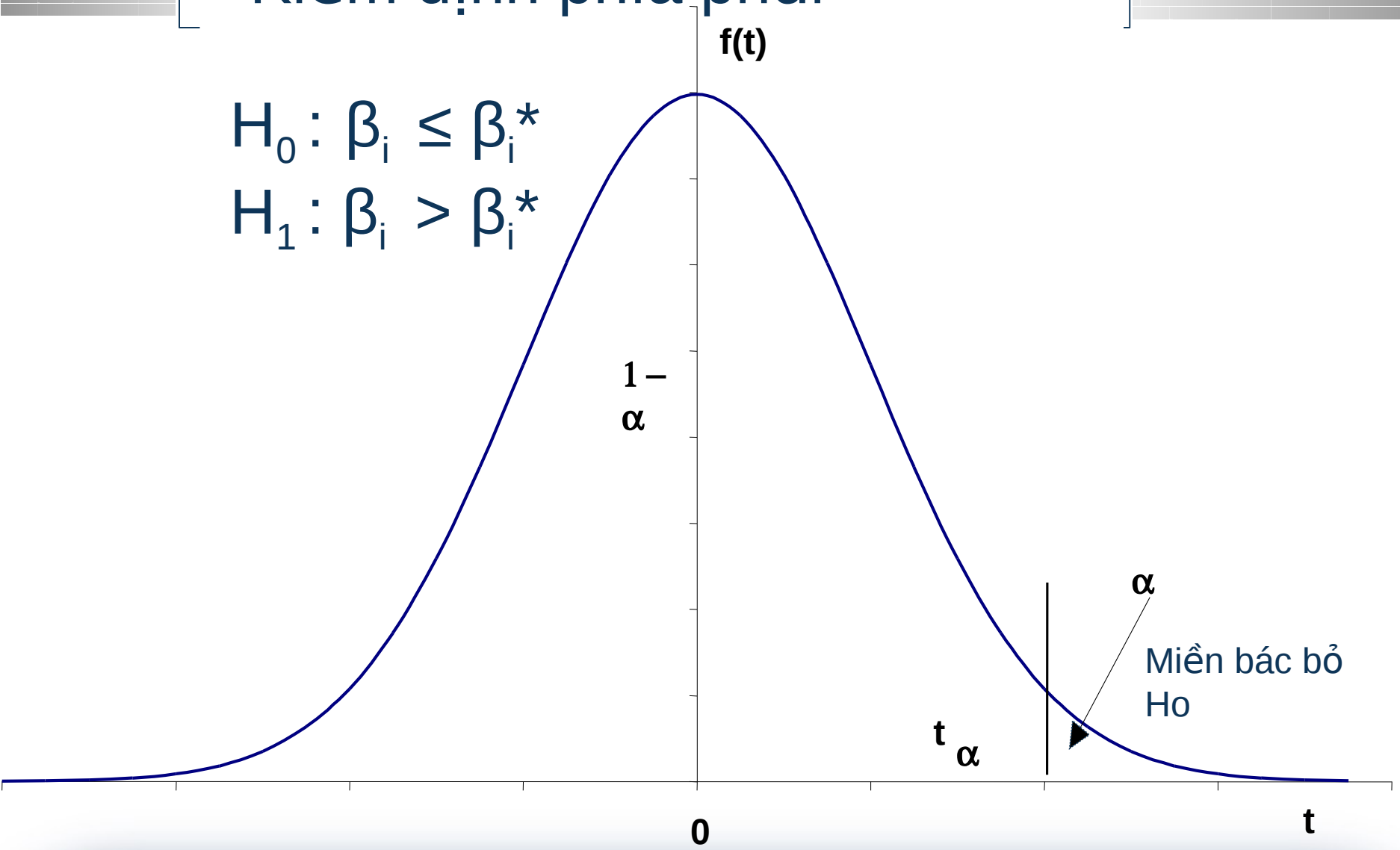
1. Kiểm định giả thiết về hệ số hồi quy

Loại GT	H_0	H_1	Miền bác bỏ
Hai phía	$\beta_i = \beta_i^*$	$\beta_i \neq \beta_i^*$	$\|t\| > t\alpha_{/2}$ (n-2)
Phía phải	$\beta_i \leq \beta_i^*$	$\beta_i > \beta_i^*$	$t > t\alpha$ (n-2)
Phía trái	$\beta_i \geq \beta_i^*$	$\beta_i < \beta_i^*$	$t < t\alpha$ (n-2)

Kiểm định phía phải

$H_0 : \beta_i \leq \beta_i^*$

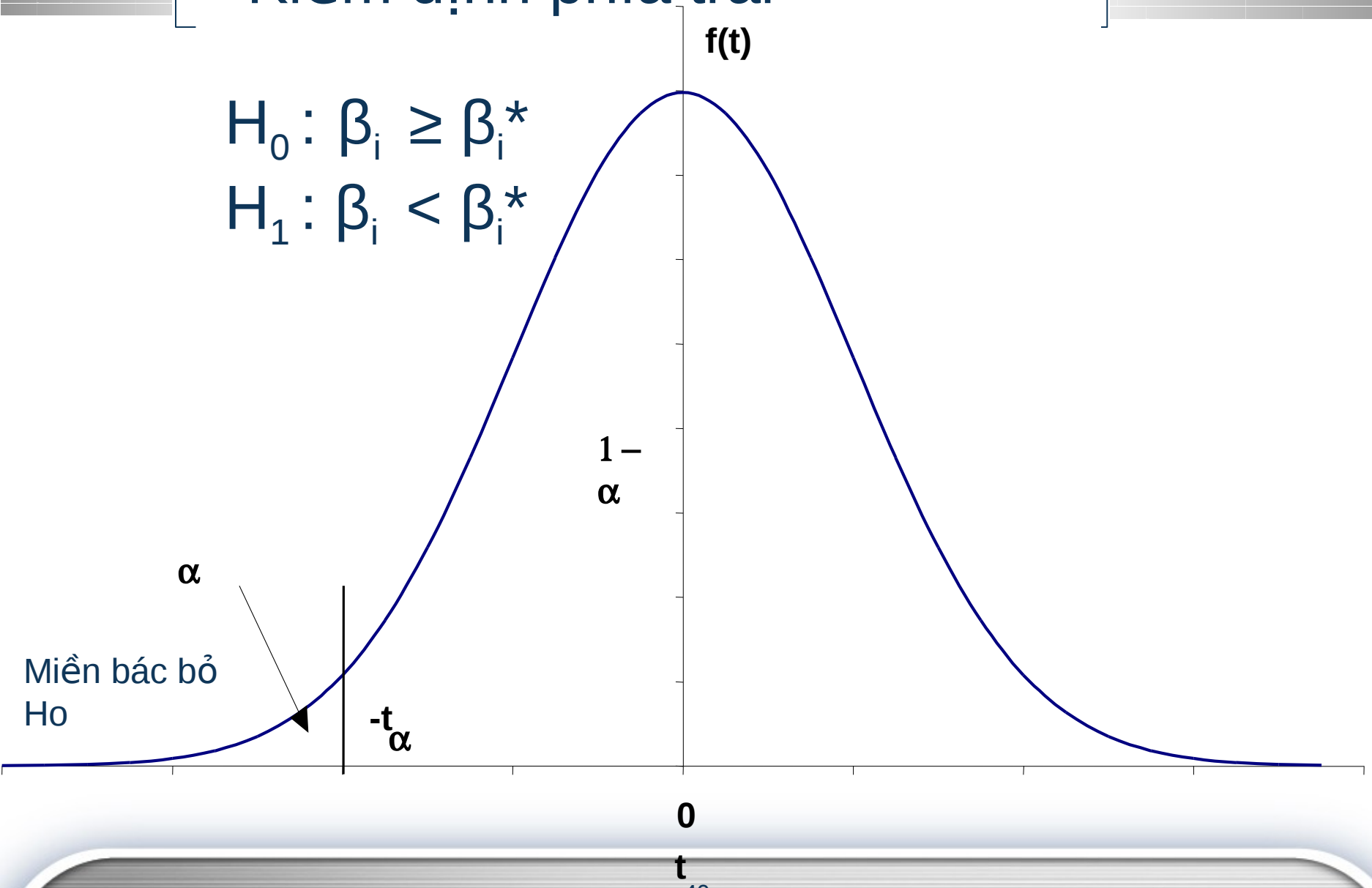
$H_1 : \beta_i > \beta_i^*$



Kiểm định phía trái

$H_0 : \beta_i \geq \beta_i^*$

$H_1 : \beta_i < \beta_i^*$



2. Kiểm định sự phù hợp của mô hình

Kiểm định giả thiết $H_0: R^2 = 0$

(tương đương $H_0: \beta_2 = 0$)

với mức ý nghĩa α hay độ tin cậy $1 - \alpha$

Bước 1:

Tính

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

a. Phương pháp giá trị tới hạn

Bước 2: Tra bảng F với mức ý nghĩa α và hai bậc tự do (1, n-2)

Bước 3: Quy tắc quyết định

- Nếu $F > F_\alpha(1,n-2)$: Bác bỏ H_0
- Nếu $F \leq F_\alpha(1,n-2)$: Chấp nhận H_0

2. Kiểm định sự phù hợp của mô hình

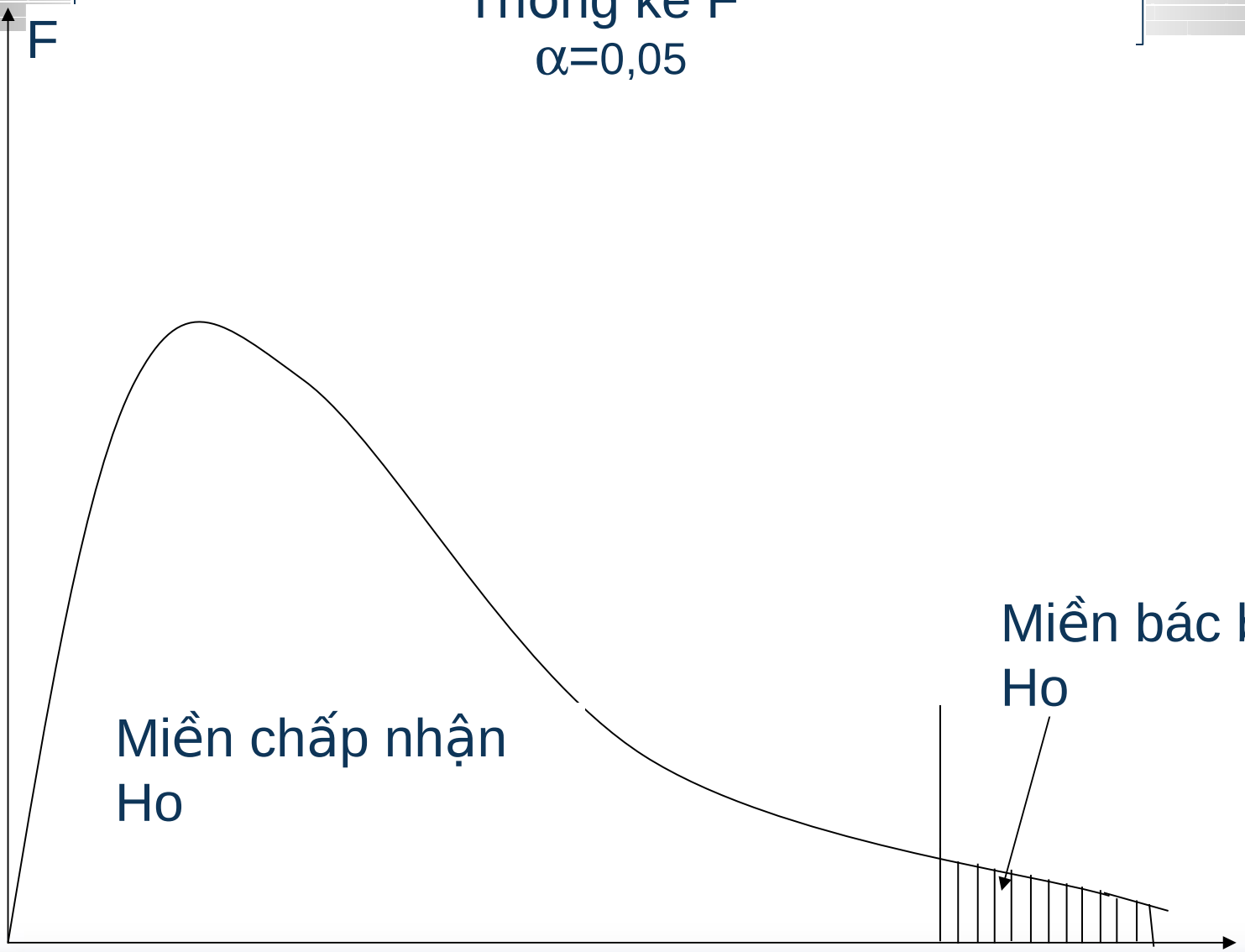
b. Phương pháp p-value

Bước 2: Tính p-value= p ($F_\alpha(1,n-2)>F$)

Bước 3: Quy tắc quyết định

- Nếu $p \leq \alpha$: Bác bỏ H_0
- Nếu $p > \alpha$: Chấp nhận H_0

Thống kê F
α=0,05



2.6 DỰ BÁO

Mô hình hồi quy

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

Cho trước giá trị $X = X_0$, dự báo giá trị trung bình và giá trị cá biệt của Y với mức ý nghĩa α hay độ tin cậy $1 - \alpha$.

*** Ước lượng điểm**

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

2.6 DỰ BÁO

* Dự báo giá trị trung bình của Y

$$E(Y / X_0) \in (\hat{Y}_0 - \varepsilon_0; \hat{Y}_0 + \varepsilon_0)$$

Với:

$$\varepsilon_0 = SE(\hat{Y}_0) t_{(n-2,\alpha/2)}$$

$$SE(\hat{Y}_0) = \sqrt{Var(\hat{Y}_0)}$$

$$Var(\hat{Y}_0) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(\overline{X} - X_0)^2}{\sum x_i^2}\right)$$

2.6 DỰ BÁO

*** Dự báo giá trị cá biệt của Y**

$$Y_0 \in (\hat{Y}_0 - \varepsilon_0^{'}; \hat{Y}_0 + \varepsilon_0^{'})$$

Với:

$$\varepsilon_0^{'} = SE(Y_0 - \hat{Y}_0) t_{(n-2,\alpha/2)}$$

$$SE(Y_0 - \hat{Y}_0) = \sqrt{Var(Y_0 - \hat{Y}_0)}$$

$$Var(Y_0 - \hat{Y}_0) = \hat{\sigma}^2 (1 + \frac{1}{n} + \frac{(\overline{X} - X_0)^2}{\sum x_i^2})$$

2.7 HỒI QUY VÀ ĐƠN VỊ ĐO CỦA BIẾN

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$$

Nếu đơn vị đo của biến X, Y thay đổi thì mô hình hồi quy mới là

$$Y_i^* = \hat{\beta}_1^* + \hat{\beta}_2^* X_i^* + e_i^*$$

Với

$$Y_i^* = k_1 Y_i;\ X_i^* = k_2 X_i \qquad \hat{\beta}_1^* = k_1 \hat{\beta}_1;\ \hat{\beta}_2^* = \frac{k_1}{k_2}\hat{\beta}_2$$

$$\text{var}(\hat{\beta}_1^*) = (k_1)^2 . \text{var}(\hat{\beta}_1);\ \text{var}(\hat{\beta}_2^*) = \left(\frac{k_1}{k_2}\right)^2 . \text{var}(\hat{\beta}_2)$$