Chương 4
Kiểm định giả thuyết thống kê với phương trình hồi qui đơn biến

TS. Đinh Thị Thanh Bình
Khoa Kinh Tế Quốc Tế- Đại học Ngoại thương

KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ

- Ví dụ như một viện nghiên cứu nông nghiệp cho rằng giống lúa mới SYM05 có năng suất trung bình 9 tấn/ha. Để đánh giá nhận định này, ta thiết lập giả thiết sau:

$$\mathbf{H_0}: \mu = 9$$

$$\mathbf{H_1}: \mu \neq 9$$

  - Với μ là năng suất trung bình thực tế của giống lúa này
  - $\mu_0 = 9$ là năng suất trung bình của giống lúa này theo báo cáo của viện nghiên cứu.

2

KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ

- H_0 gọi là giả thiết thống kê (giả thiết không- null hypothesis)
- H_1 gọi là giả thiết đối (alternative hypothesis).
- Nếu sau khi kiểm định ta chấp nhận H_0 (xem H_0 là đúng) thì đánh giá nhận định của viện nghiên cứu là đúng. Còn nếu ta bác bỏ H_0 (xem H_0 là sai) thì cho rằng nhận định của viện nghiên cứu là sai.

KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ

- Để kiểm định giả thiết xem chấp nhận hay bác bỏ H_0 thì người ta phải dựa vào kết quả khảo sát trên mẫu và đưa ra quyết định dựa trên mẫu. Có bốn trường hợp có thể xảy ra:

Quyết định chủ quan / Thực tế khách quan	**Bác bỏ H_0**	**Chấp nhận H_0**
H_0 sai	**Đúng**	**Sai lầm loại II**
H_0 đúng	**Sai lầm loại I**	**Đúng**

KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ

- Xác suất xảy ra sai lầm loại I thường được xét nhỏ hơn hoặc bằng một giá trị số α cho trước, và α gọi là mức ý nghĩa của kiểm định. Xác suất xảy ra sai lầm loại II thường ký hiệu là β:

  P(sai lầm loại I) = P(bác bỏ H_0/H_0 đúng) ≤ α

  P(sai lầm loại II) = P(chấp nhận H_0/H_0 sai) = β

- *Tư tưởng của kiểm định là tìm cơ sở để bác bỏ giả thiết* H_0. Nếu có đủ cơ sở để bác bỏ thì ta bác bỏ H_0, còn nếu không có đủ cơ sở để bác bỏ thì ta phải chấp nhận H_0.

5

1. Phân bố xác suất của các ước lượng OLS

Giả thiết 6: Sai số u độc lập với các biến X và có phân phối chuẩn:

$$u \sim N(0, \sigma^2)$$

Định lý 4.1: Với giả thiết từ 1-6,

$$\beta_j \sim Normal[(\beta_j, Var(\beta_j)]$$

$$\Rightarrow \quad (\beta_j - \beta_j) / sd(\beta_j) \sim Normal(0,1)$$

Định lý 4.2: Với giả thiết từ 1-6,

$$(\beta_j - \beta_j) / se(\beta_j) \sim t_{n-k-1}$$

trong đó k là số lượng biến độc lập

2. Kiểm định giả thuyết về hệ số hồi quy

- Có ba dạng giả thuyet kiểm định như sau về hệ số hồi quy:

- Hai phía: $\begin{cases} H_0 : \beta_i = \beta_i^* \\ H_1 : \beta_i \neq \beta_i^* \end{cases}$

- Phía phải: $\begin{cases} H_0 : \beta_i \leq \beta_i^* \\ H_1 : \beta_i > \beta_i^* \end{cases}$

- Phía trái: $\begin{cases} H_0 : \beta_i \geq \beta_i^* \\ H_1 : \beta_i < \beta_i^* \end{cases}$

- Trong đó, β_i nhận giá trị là β_0 hoặc β_1 (trong phạm vi mô hình hồi quy đơn mà ta đang xét).
- β_i^* là giả thiết về giá trị thực của β_i,

Các thông số cần thiết

- Thống kê T
- Mức ý nghĩa α
- Hệ số tin cậy $(1-\alpha)$
- Giá trị tới hạn (critical value): c

2.1. Ước lượng khoảng: một vài tư tưởng

- Ta biết rằng $\hat{\beta}_0$ và $\hat{\beta}_1$ là ước lượng điểm (point estimators) của β_0 và β_1 nhưng do các dao động của việc lấy mẫu lặp lại nên các ước lượng điểm có thể khác với giá trị thực mặc dù trung bình giá trị của các ước lượng $\hat{\beta}_0$ và $\hat{\beta}_1$ bằng với giá trị thực β_0 và β_1.
- Do đó người ta muốn xây dựng một khoảng xung quanh giá trị ước lượng điểm với lòng tin rằng giá trị thực sẽ nằm trong khoảng đó với một độ tin cậy nhất định.

→ Cách làm này gọi là ước lượng khoảng.

Khoảng tin cậy của hệ số β_1

- Với các giả thiết 1-6, ta có:

$$T = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = \frac{(\hat{\beta}_1 - \beta_1)\sqrt{\sum (x_i - \bar{x})^2}}{\sigma}$$

$$T \sim t_{n-k-1} \Leftrightarrow T \sim t_{n-2}$$

Khoảng tin cậy của hệ số β_1

- Xác định giá trị tới hạn $c_{\alpha/2}$ để $(1-\alpha)$ diện tích trong phân phối của T nằm giữa $c_{\alpha/2}$ và $-c_{\alpha/2}$

$$P(-c_{\alpha/2} < T < c_{\alpha/2}) = 1-\alpha$$

- Khoảng tin cậy chứa β_1 với xác suất bằng $(1-\alpha)$ là:

$$\beta_1 \mp c_{\alpha/2} se(\beta_1)$$

Khoảng tin cậy của hệ số β_1

- Khoảng tin cậy bên phải:

$$(\beta_1 - c_\alpha se(\beta_1), +\infty)$$

- Khoảng tin cậy bên trái:

$$(-\infty, \beta_1 + c_\alpha se(\beta_1))$$

Khoảng tin cậy của hệ số β_0

- Tương tự như trên ta có thể xây dựng được khoảng tin cậy cho hệ số β_0 như sau:

$$\beta_0 \mp c_{\alpha/2} se(\beta_0)$$

Trong đó:

$$se(\beta_0) = \frac{\sigma n^{-1} (\sum_{i=1}^{n} X_i^2)^{1/2}}{(\sum_{i=1}^{n} (X_i - \overline{X})^2)^{1/2}}$$

Khoảng tin cậy của hệ số β_0

- Khoảng tin cậy bên phải:

$$(\beta_0 - c_\alpha se(\beta_0), +\infty)$$

- Khoảng tin cậy bên trái:

$$(-\infty, \beta_0 + c_\alpha se(\beta_0))$$

Kết luận của phương pháp khoảng tin cậy

- **Đối với kiểm định hai phía:** Nếu giá trị β_i^* không rơi vào khoảng này thì ta bác bỏ giả thiết H_0.

$$[\hat{\beta}_j \mp c_{\alpha/2} se(\hat{\beta}_j)]$$

- **Đối với kiểm định phía phải:** Nếu giá trị β_i^* không rơi vào khoảng này thì ta bác giả thiết H_0.

$$[\hat{\beta}_j - c_{\alpha} se(\hat{\beta}_j), +\infty]$$

- **Đối với kiểm định phía trái:** Nếu giá trị β_i^* không rơi vào khoảng này thì ta bác giả thiết H_0.

$$[-\infty, \hat{\beta}_j + c_{\alpha} se(\hat{\beta}_j)]$$

2.2. Phương pháp giá trị tới hạn

- **Bước 1**: Tính giá trị $T_0 = \dfrac{\hat{\beta}_j - \beta^*_j}{se(\hat{\beta}_j)}$

- **Bước 2**: Tra bảng t-student với mức ý nghĩa α/2 (nếu là kiểm định hai phía) hoặc mức ý nghĩa α (nếu là kiểm định một phía) để có giá trị tới hạn $c_{\alpha/2}$ hoặc c_α

- **Bước 3**: So sánh T_0 với giá trị tới hạn. Quy tắc quyết định như sau:

2.2. Phương pháp giá trị tới hạn

Quy tắc quyết định

Loại giả thuyết	H_0	H_1	Miền bác bỏ H_0
Hai phía	$\beta_j = \beta_j^*$	$\beta_j \# \beta_j^*$	$\lvert T_0 \rvert > c_{\alpha/2}$
Phía phải	$\beta_j \le \beta_j^*$	$\beta_j > \beta_j^*$	$T_0 > c_\alpha$
Phía trái	$\beta_j \ge \beta_j^*$	$\beta_j < \beta_j^*$	$T_0 < -c_\alpha$

2.3. Phương pháp giá trị p-value

- **Bước 1**: tính giá trị $T = \dfrac{\hat{\beta}_j - \beta_j^*}{se(\hat{\beta}_j)}$

- **Bước 2**: tính p-value = P ($|T| > t_0$), trong đó T là đại lượng ngẫu nhiên có phân phối t-student với (n-2) bậc tự do. t_0 là giá trị cụ thể của T.

- **Bước 3**: nếu cho trước mức ý nghĩa α, quy tắc quyết định sẽ là:
  - Kiểm định hai phía: p-value < α: bác bỏ H_0
  - Kiểm định một phía: p-value/2 < α: bác bỏ H_0

3. Kiểm định giả thuyết về phương sai của nhiễu

- Phương pháp tiến hành kiểm định giả thiết tương tự như kiểm định giả thiết về hệ số hồi quy. Bảng 2.06 trình bày một cách tóm tắt các loại giả thiết, phương pháp kiểm định và quy tắc quyết định.
- Trong giả thiết H_0, σ_0^2 là giá trị số cho trước và:

$$T = \frac{(n-2)\hat{\sigma}^2}{\sigma_0^2} \sim \chi^2_{n-2}$$

$$p-value = P(T > t_0 \mid H_0)$$

3.1. Khoảng tin cậy của phương sai

- Phương sai của tổng thể chính là phương sai của thành phần nhiễu u_i mà ta kí hiệu là σ^2.
- Với giả thiết về phân phối chuẩn của nhiễu, ta có thống kê:

$$T = (n-2)\frac{\hat{\sigma}^2}{\sigma^2} \square \chi^2_{n-k-2}$$

3.2. Khoảng tin cậy của phương sai

- Xác định giá trị tới hạn $c_{\alpha/2}$ để $(1-\alpha)$ diện tích trong phân phối của T nằm giữa $c_{1-\alpha/2}$ và $c_{\alpha/2}$

$$P(c_{1-(\alpha/2)} < T < c_{\alpha/2}) = 1-\alpha$$

- Khoảng tin cậy $(1-\alpha)$ chứa σ^2 là:

$$(n-2)\frac{\sigma^2}{c_{\alpha/2}} \mp (n-2)\frac{\sigma^2}{c_{1-\alpha/2}} c_{1-(\alpha/2)}$$

Bảng 4.1 Kiểm định giả thiết về phương sai của nhiễu

giả thiết	H_0	H_1	Phương pháp	Miền bác bỏ H_0
Hai phía	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	Khoảng tin cậy	$\sigma_0^2 \notin [(n-2)\frac{\hat{\sigma}^2}{c_{\alpha/2}}, (n-2)\frac{\sigma^2}{c_{1-\alpha/2}}]$
			Giá trị tới hạn	$T > c_{\alpha/2}$ **hoặc** $T < c_{1-\alpha/2}$
			p-value	*p-value* < α/2 **hoặc** *p-value* > 1- α/2
Phía phải	$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$	Khoảng tin cậy	$\sigma_0^2 \notin [(n-2)\frac{\hat{\sigma}^2}{c_\alpha}, +\infty]$
			Giá trị tới hạn	$T > c_\alpha$
			p-value	*p-value* < α
Phía trái	$\sigma^2 = \sigma_0^2$	$\sigma^2 < \sigma_0^2$	Khoảng tin cậy	$\sigma_0^2 \notin [-\infty, (n-2)\frac{\hat{\sigma}^2}{c_{1-\alpha}}]$
			Giá trị tới hạn	$T < c_{1-\alpha}$
			p-value	*p-value*> 1- α

4. Kiểm định sự phù hợp của mô hình hồi quy

- 5.1. Các tổng bình phương độ lệch
- 5.2. Hệ số xác định (đơn)
- 5.3. Kiểm định sự phù hợp của mô hình hồi quy

4.1. CÁC TỔNG BÌNH PHƯƠNG ĐỘ LỆCH

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

$$SST = \sum (Y_i - \overline{Y})^2$$

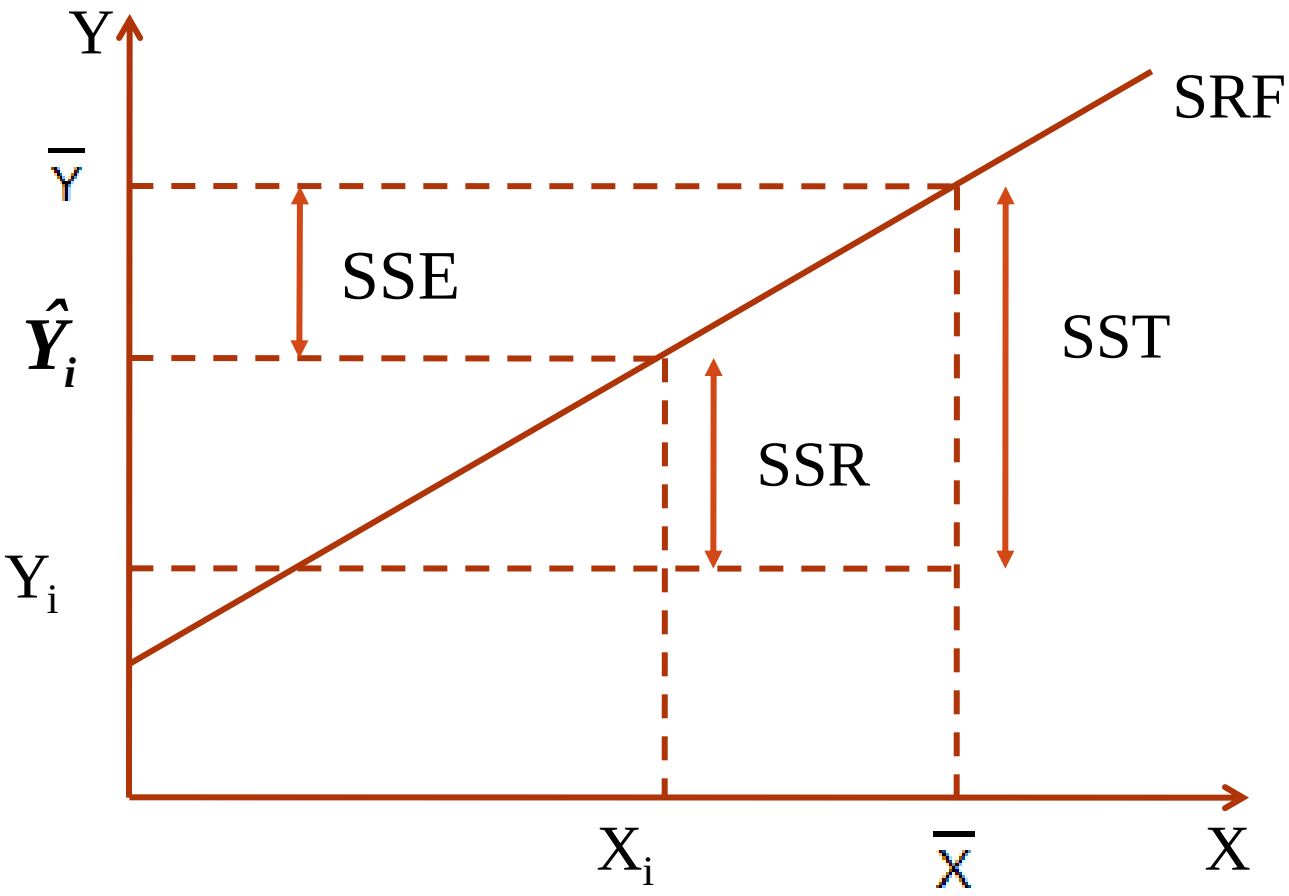
- SSE: (Explained Sum of Squares - Bình phương sai số được giải thích)

$$SSE = \sum (\hat{Y}_i - \overline{Y})^2$$

- SSR: (Residual Sum of Squares - Tổng bình phương các phần dư)

$$SSR = \sum_{i=1}^{n} \left(Y_i - Y_i \right)^2 = \sum u_i^2$$

Hình 4.2: Ý nghĩa hình học của SST, SSR và SSE



4.2. HỆ SỐ XÁC ĐỊNH R^2

Ta chứng minh được: SST = SSE + SSR

$$1 = \frac{SSE}{SST} + \frac{SSR}{SST}$$

4.2. HỆ SỐ XÁC ĐỊNH R^2

Hệ số xác định R^2: đo mức độ phù hợp của hàm hồi quy mẫu.

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

Trong mô hình 2 biến:

$$R^2 = \frac{\hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}$$

4.3. Hệ số xác định (đơn)

Nếu chia cả tử và mẫu của phân số trên cho mẫu n (hoặc (n-1) nếu là mẫu nhỏ) thì ta sẽ được :

$$r^2 = \hat{\beta}_1^2 \left[\frac{\sum (X_i - \overline{X})^2 / (n-1)}{\sum (Y_i - \overline{Y})^2 / (n-1)} \right] = \hat{\beta}_1^2 \left(\frac{S_x^2}{S_y^2} \right)$$

- S_x^2 và S_y^2 là phương sai mẫu của X và Y.
- r^2 đo tỷ lệ hay số phần trăm của toàn bộ sai lệch của Y với giá trị trung bình của chúng được giải thích bằng mô hình (hay biến độc lập).
- r^2 nằm trong đoạn [0,1]

4.4. Kiểm định sự phù hợp của mô hình hồi quy

- Để đánh giá mức độ thích hợp của mô hình hồi quy, nghĩa là mô hình hồi quy giải thích được bao nhiêu % sự thay đổi của biến phụ thuộc Y, thì ta sử dụng hệ số xác định r^2.
- Hệ số r^2 càng gần 1 bao nhiêu thì mô hình hồi quy càng có ý nghĩa bấy nhiêu.

4.4. Kiểm định mô hình

- Chúng ta quan tâm đến việc đánh giá xem giá trị của r^2 khác 0 có ý nghĩa thống kê hay không. Nghĩa là ta tiến hành kiểm định giả thiết:

$$\begin{cases} H_0 : R^2 = 0 \\ H_1 : R^2 \neq 0 \end{cases}$$

- Đối với mô hình hồi quy hai biến, giả thiết trên tương đương với giả thiết:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

- Ta sẽ tiến hành kiểm định giả thiết này dựa vào giá trị của F được tính theo công thức.

4.4.1. Phương pháp giá trị tới hạn

- **Bước 1**: Tính $F_0 = \dfrac{R^2 / k}{(1-R^2)/(n-k-1)}$
- **Bước 2**: Tra bảng F với mức ý nghĩa α và hai bậc tự do (1, n-k-1) ta được giá trị tới hạn $c_{\alpha, (1, n-k-1)}$
- **Bước 3**: So sánh F_0 và $c_{\alpha, (1, n-k-1)}$
  - Nếu $F_0 > c_{\alpha, (1, n-k-1)}$ bác bỏ H_0
  - Nếu $F_0 < c_{\alpha, (1, n-k-1)}$ không có cơ sở để bác bỏ H_0

4.4.2. Phương pháp giá trị p-value

- **Bước 1**: Tính $F_0 = \frac{R^2(n-k-1)}{(1-R^2)}$
- **Bước 2**: Tính p-value = P(F > F_0) với F là phân phối Fisher có hai bậc tự do là (k, n-2)
- **Bước 3**: So sánh p-value và mức ý nghĩa α
  - Nếu p-value < α : bác bỏ H_0
  - Nếu p-value > α : không có cơ sở để bác bỏ H_0

Bài tập

```
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  4,   521) =
       Model |                                         Prob > F      =  0.0000
    Residual |  4899.15523                             R-squared     =
-------------+------------------------------           Adj R-squared =  0.3105
       Total |  7160.41429   525  13.6388844           Root MSE      =

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5833233    .051656
       exper |   .0556664   .0110553
      female |  -2.067101   .2722077
     married |   .6602419   .2968513
       _cons |  -1.790662   .7512121
```

Bài tập

```
  Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------   F(  4,   521) =   60.12
       Model |  2261.25906     4  565.314766   Prob > F      =  0.0000
    Residual |  4899.15523   521  9.40336896   R-squared     =  0.3158
-------------+------------------------------   Adj R-squared =  0.3105
       Total |  7160.41429   525  13.6388844   Root MSE      =  3.0665

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5833233    .051656    11.29   0.000     .4818437    .6848029
       exper |   .0556664   .0110553     5.04   0.000     .0339479    .0773849
      female |  -2.067101   .2722077    -7.59   0.000    -2.601861   -1.532342
     married |   .6602419   .2968513     2.22   0.027     .0770693    1.243414
       _cons |  -1.790662   .7512121    -2.38   0.017    -3.266439   -.3148853
```