Kiểm định Giả thuyết trong Hồi quy Đa biến

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

19-22/12/2023

Hồi quy tuyến tính đa biến

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + u$$

- y gọi là biến phụ thuộc/biến được giải thích.
- x_1, x_2, ... là biến độc lập/biến giải thích.
- u là sai số, bao gồm tất cả những yếu tố khác ảnh hưởng đến y nhưng không nằm trong x_1, x_2.
- β_0, β_1, β_2. ... là các tham số trong mô hình.

Các giả định đối với hồi quy đa biến

1. Tuyến tính theo tham số.
2. Chọn mẫu ngẫu nhiên.
3. Không có cộng tuyến hoàn hảo giữa các biến giải thích.
4. Trung bình có điều kiện của sai số bằng 0:

$$E(u|x_1, ..., x_k) = 0$$

$\Rightarrow$ Ước lượng của OLS là không chệch.

$$E(\hat{\beta}) = \beta$$

Giả định phương sai của sai số không đổi (homoskedasticity)

5. Với các giá trị của các biến giải thích cho trước, phương sai của sai số là một hằng số:

$$Var(u|x_1, ..., x_k) = \sigma^2$$

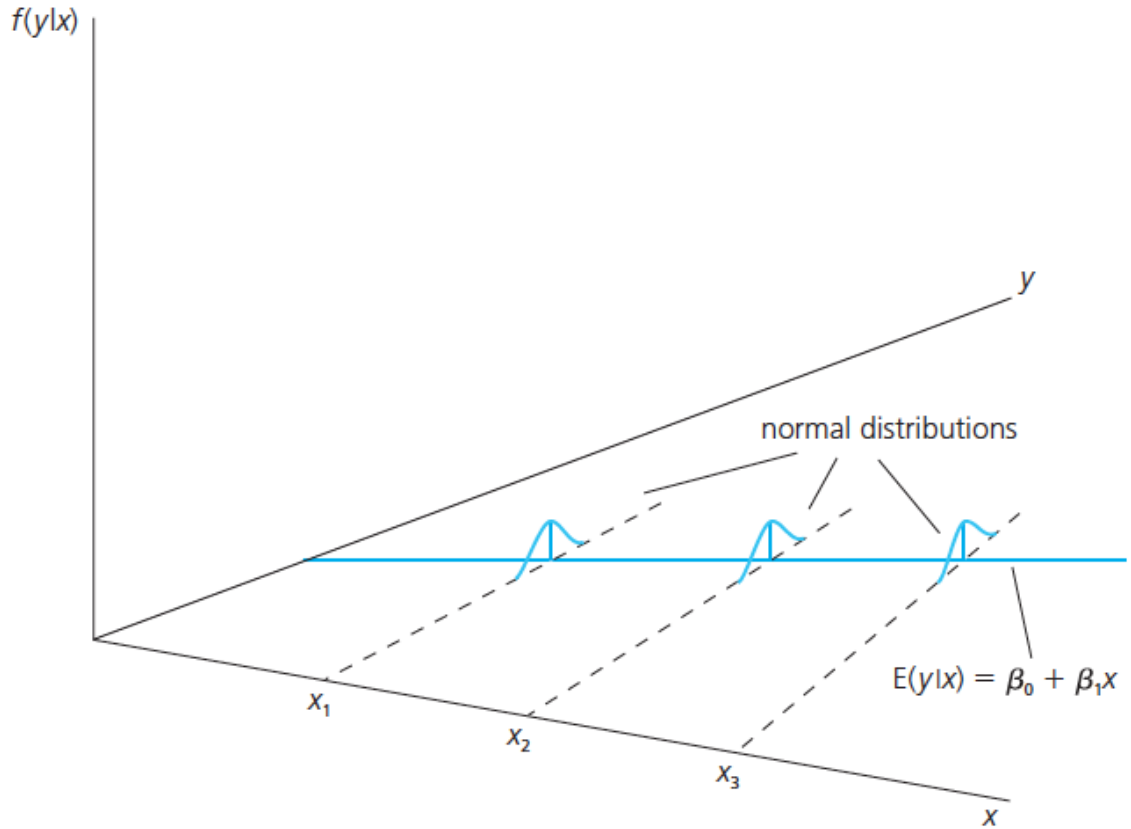
Với các giả định 1-5, ước lượng của OLS là ước lượng tuyến tính, không chệch, và hiệu quả nhất (**Best Linear Unbiased Estimator - BLUE**).

- Ước lượng của β là hàm tuyến tính của biến phụ thuộc (Linear).
- Trong tất cả các ước lượng tuyến tính, OLS có phương sai của ước lượng là nhỏ nhất (Best).
- Không chệch (Unbiased), $E(\hat{\beta}) = \beta$.

Giả định về phân phối mẫu của sai số

6. Sai số u đồng nhất, độc lập với các biến giải thích (independent, identically distributed - *iid*), và có phân phối chuẩn với giá trị trung bình là 0 và phương sai σ^2.

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



Mô hình hồi quy tuyến tính cổ điển (Classical Linear Regression Model - CLRM)

Nếu thỏa các giả định 1-6 thì mô hình được coi là mô hình hồi quy tuyến tính cổ điển.

- Ước lượng của β là BLUE.
- Phân phối mẫu của ước lượng của β là:

$$\hat{\beta} \sim N(\beta, Var(\beta))$$

Viết dưới dạng chuẩn hóa:

$$\frac{\hat{\beta} - \beta}{sd(\hat{\beta})} \sim N(0, 1)$$

Phân phối mẫu của ước lượng $\hat{\beta}_j$

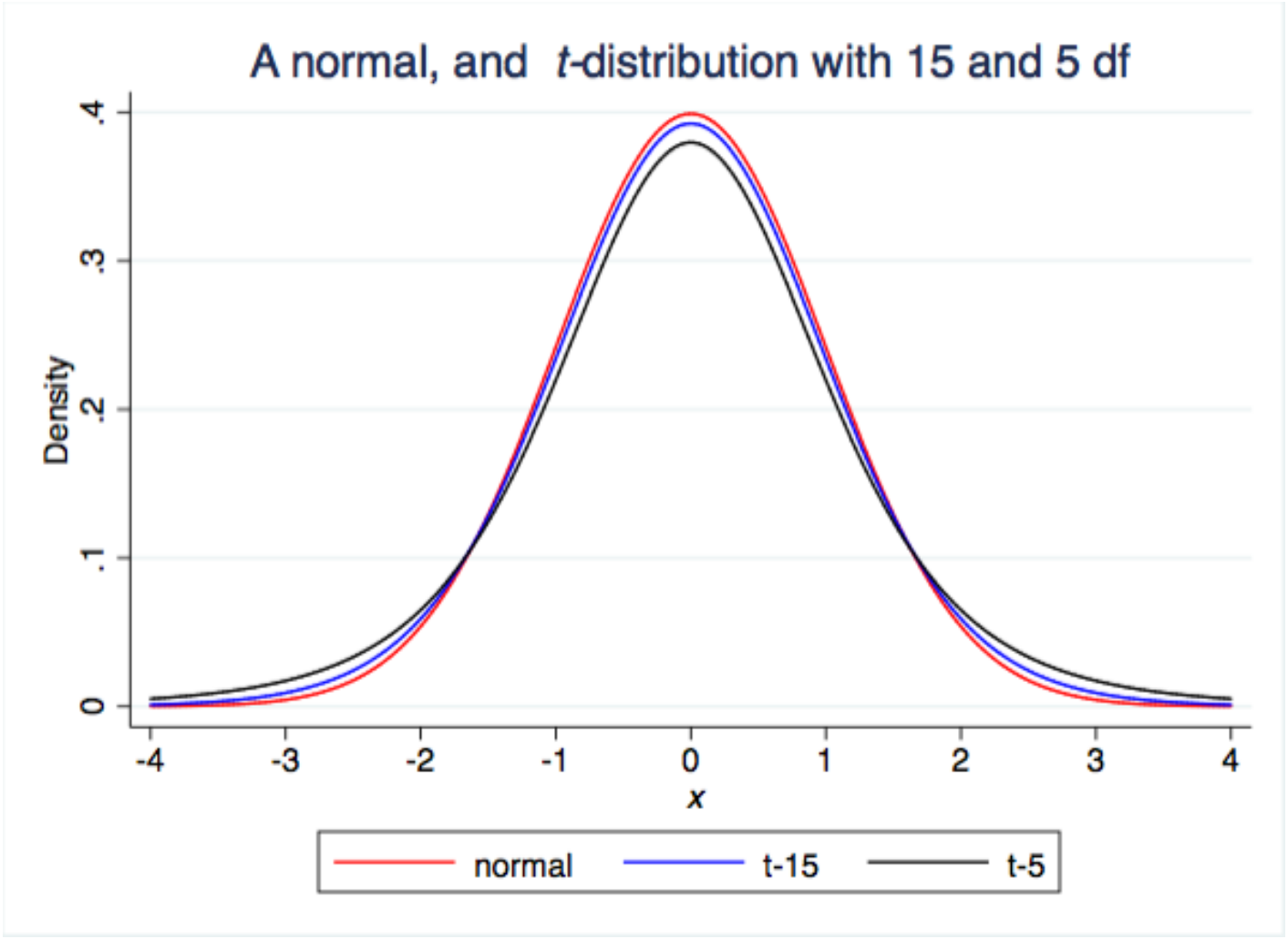
Từ các giả định CLRM, nhưng không biết phương sai σ^2 của sai số từ tổng thể (mặc dù biết là không đổi), các trị kiểm định của β_j dựa trên phân phối t-student được tính như sau:

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

với n là số quan sát trong mô hình, k là số biến giải thích.

- Công thức này sẽ cho phép kiểm định các giả thuyết về giá trị của ước lượng trong mô hình hồi quy.
- $\hat{\beta}_j$ và $se(\hat{\beta}_j)$ được tính từ phương pháp OLS với hồi quy đa biến.

Phân phối t và phân phối chuẩn



Giả thuyết và kiểm định giả thuyết

- Giả thuyết 1 phía, ví dụ nữ có thu nhập thấp hơn nam trong mô hình ước lượng tỷ suất thu nhập của việc đi học.

$$H_0 : \beta \leq 0 \quad vs. \quad H_1 : \beta > 0$$

hoặc

$$H_0 : \beta \geq 0 \quad vs. \quad H_1 : \beta < 0$$

- Giả thuyết 2 phía, ví dụ số năm đi học có tác động đến thu nhập (chiều hướng tác động có thể là âm hoặc dương).

$$H_0 : \beta = 0 \quad vs. \quad H_1 : \beta \neq 0$$

- Nếu $\beta \neq 0$ thì biến x được gọi là có ý nghĩa thống kê trong mô hình.

Mức ý nghĩa và sai lầm khi thực hiện kiểm định giả thuyết

- Dựa trên một mức ý nghĩa cho trước (α), kiểm định một giả thuyết là xem xét liệu chúng ta có bác bỏ được giả thuyết với xác suất α trong khi thực tế giả thuyết là đúng.
  - Ví dụ thực hiện một kiểm định ở mức ý nghĩa $\alpha = 5\%$ có nghĩa là chúng ta chấp nhận xác suất là 5% sai lầm khi bác bỏ giả thuyết H_0.
- **Sai lầm loại I và sai lầm loại II**
  - Sai lầm loại I (false positive) là mức ý nghĩa của kiểm định, α.
  - Sai lầm loại II (false negative) là β.

		Giả thuyết Ho ($\tau = 0$)	
		Đúng ($\tau = 0$)	**Sai ($\tau \neq 0$)**
Quyết định	**Không bác bỏ Ho**	$1 - \alpha$ **[Đúng]**	β **[Sai]**
	Bác bỏ Ho	α **[Sai]**	$1 - \beta$ **[Đúng]**

Vai trò của mức ý nghĩa α trong kiểm định giả thuyết

α thể hiện việc chấp nhận sai lầm loại I ($H_0 : \tau = 0$ đúng nhưng bị bác bỏ). Có nghĩa là chúng ta kết luận lầm có tác động nhưng trên thực tế không có tác động.

Mức ý nghĩa $\alpha \Rightarrow$ Độ tin cậy $1 - \alpha$

- Nếu áp tiêu chuẩn cao (hạn chế chấp nhận sai lầm), hàm ý kiểm định $H_0 : \tau = 0$ sẽ khó bị bác bỏ hơn, thì α phải nhỏ (ví dụ với $\alpha = 0.01 < 0.05 < 0.1$). Điều này đồng nghĩa với giá trị cực trị (critical value) trong kiểm định giả thuyết t_c càng lớn khi α càng nhỏ.
- Ngược lại, nếu sẵn sàng chấp nhận sai lầm ở mức độ cao thì α có thể nhận giá trị lớn, ví dụ 0.05 hay 0.1. Giá trị cực trị t_c càng nhỏ khi α càng lớn.

Sai lầm loại II và sức mạnh thống kê $1 - \beta$ trong kiểm định giả thuyết

Sai lầm loại II (β) là xác suất không bác bỏ $H_0 : \tau = 0$ khi trên thực tế giả thuyết này là sai, có nghĩa là chúng ta kết luận lầm rằng $\tau = 0$ trong khi trên thực tế $\tau \neq 0$ (false negative).

- Sức mạnh thống kê (power of the test) $1 - \beta$ là xác suất phát hiện được tác động ($\tau \neq 0$) khi trên thực tế tác động là có thực ($\tau \neq 0$).

Vai trò của α và β khi thiết kế nghiên cứu và diễn giải kết quả

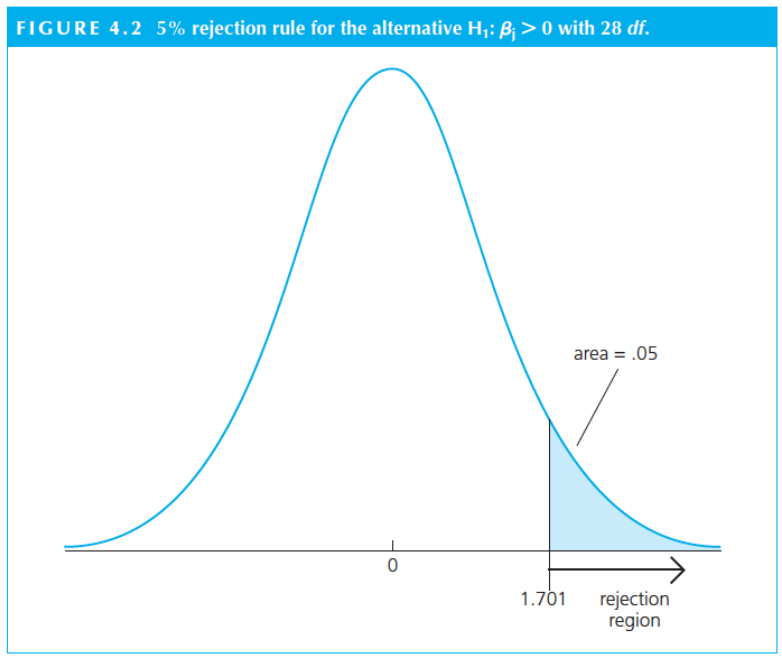
- Nếu chú trọng ý nghĩa thống kê của tham số ước lượng thì cần tham số có ý nghĩa thống kê cao (α thấp, chẳng hạn ở mức ý nghĩa 0.05 hay 0.01 thay vì 0.1).
- Tăng số quan sát trong mô hình $\Rightarrow$ Tăng số bậc tự do, $df = n - k - 1 \Rightarrow$ Giảm sai lầm loại I.
- Chúng ta cần β trong giai đoạn thiết kế nghiên cứu và tính toán số cỡ mẫu tối thiểu. Nếu tác động trên thực tế càng nhỏ thì cần cỡ mẫu càng lớn để phát hiện được tác động.

Kiểm định 1 phía (1-sided test)

- H_0: Giả thuyết không (null hypothesis), $\beta \leq 0$
- H_1: Giả thuyết thay thế (alternative hypothesis), $\beta > 0$

Mục đích của kiểm định là để bác bỏ H_0 dựa trên nguyên tắc bác bỏ (rejection rule):

$$t_{\hat{\beta}} > t_{critical} \Rightarrow \quad Reject \quad H_0$$



Kiểm định 1 phía (1-sided test) (2)

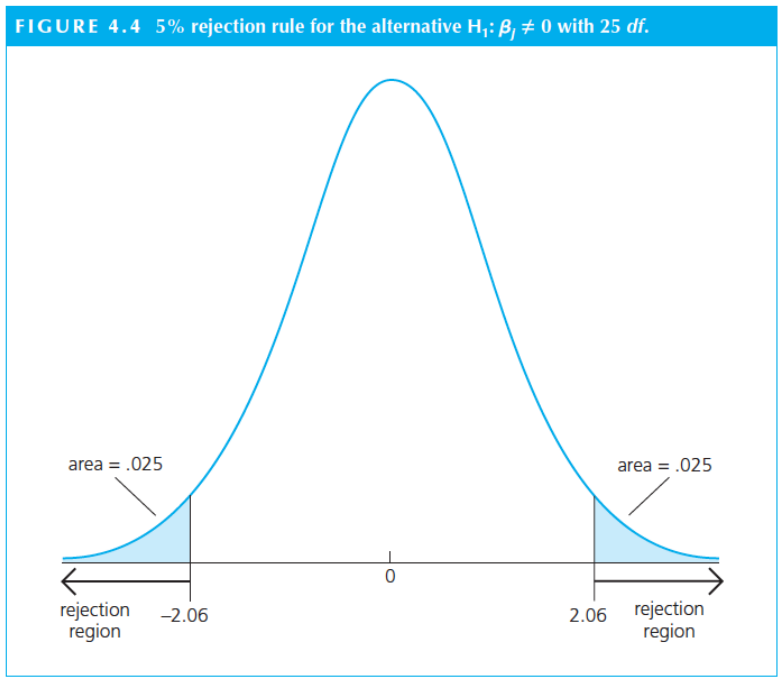
- H_0: Giả thuyết không (null hypothesis), $\beta \geq 0$
- H_1: Giả thuyết thay thế (alternative hypothesis), $\beta < 0$

$$t_{\hat{\beta}} < t_{critical} \Rightarrow \quad Reject \quad H_0$$

Kiểm định 2 phía (2-sided test)

- H_0: Giả thuyết không (null hypothesis), $\beta = 0$
- H_1: Giả thuyết thay thế (alternative hypothesis), $\beta \neq 0$

$$|t_{\hat{\beta}}| > t_{critical} \Rightarrow \quad Reject \quad H_0$$



Giá trị cực trị và độ tự do của trị kiểm định

- Mức ý nghĩa α (significance level) hoặc độ tin cậy $1 - \alpha$ (confidence level): Để bác bỏ giả thuyết ở độ tin cậy 99% khó hơn ở độ tin cậy 95% và càng khó hơn ở độ tin cậy 90%.
- Độ tự do $df = n - k - 1$: số quan sát n càng nhiều thì phân phối mẫu của tham số ước lượng $\hat{\beta}$ càng gần với phân phối chuẩn và khả năng bác bỏ giả thuyết càng dễ. k là số biến giải thích trong mô hình.

Giá trị cực trị

- Với kiểm định một phía, cần tìm t_{α}^{df} tương ứng với độ tự do df và mức ý nghĩa α cho trước. Ví dụ:
  - Với $df = 30$, $\alpha = 90\%$, $\alpha = 95\%$, $\alpha = 99\%$ thì $t_{.10}^{30} = 1.3104$, $t_{.05}^{30} = 1.6973$, $t_{.01}^{30} = 2.4573$.

- Với kiểm định hai phía, cần tìm $t_{\alpha/2}^{df}$ tương ứng với độ tự do df và mức ý nghĩa α cho trước. Ví dụ:
  - Với $df = 30$, $\alpha = 10\%$, $\alpha = 5\%$, $\alpha = 1\%$ thì $t_{.05}^{30} = 1.6973$, $t_{.025}^{30} = 2.0423$, $t_{.005}^{30} = 2.75$.

- Trong stata, display invttail(df,α)

Ví dụ với mô hình tỷ suất thu nhập

Sử dụng bộ dữ liệu hh2010.dta, ước lượng mô hình tỷ suất thu nhập của đi học bằng hồi quy đa biến như sau:

$$log(income) = \beta_0 + \beta_1 yoeduc + \beta_2 yoexper + \beta_3 yoexpersq + \beta_4 married + \beta_5 school + \beta_6 public + \beta_7 foreign + \beta_8 official + u$$

```
. reg lincome yoeduc yoexper yoexpersq married publicSchool public foreign official
```

Source	SS	df	MS	Number of obs	=	7,552
				F(8, 7543)	=	409.20
Model	1753.70541	8	219.213176	Prob > F	=	0.0000
Residual	4040.86526	7,543	.535710627	R-squared	=	0.3026
				Adj R-squared	=	0.3019
Total	5794.57067	7,551	.767391162	Root MSE	=	.73192

lincome	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
yoeduc	.0926075	.0027428	33.76	0.000	.0872309	.0979841
yoexper	.061687	.0025081	24.60	0.000	.0567705	.0666035
yoexpersq	-.0012002	.0000488	-24.58	0.000	-.0012959	-.0011044
married	.0352395	.0221221	1.59	0.111	-.0081259	.078605
publicSchool	-.1145887	.0423549	-2.71	0.007	-.1976161	-.0315613
public	-.1042541	.0329488	-3.16	0.002	-.1688429	-.0396652
foreign	.4499482	.0363715	12.37	0.000	.37865	.5212464
official	.2705426	.0359373	7.53	0.000	.2000956	.3409897
_cons	8.493551	.0474837	178.87	0.000	8.40047	8.586633

Kiểm định giả thuyết về tỷ suất thu nhập của việc đi học

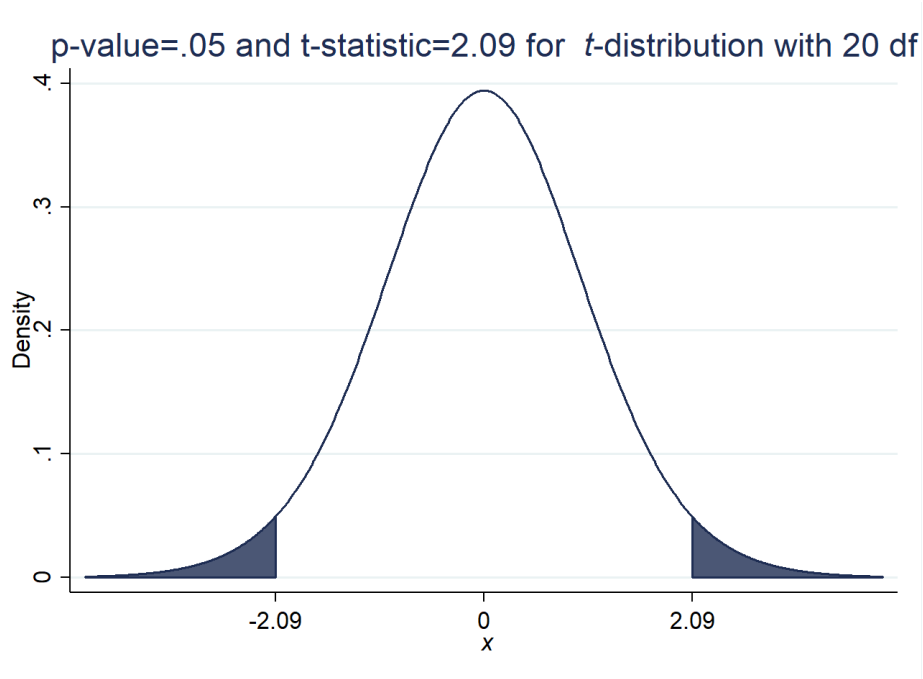
- Kiểm định hai phía: $H_0 : \beta_1 = 0$
  - Trị kiểm định $t_{\beta_1} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = 33.76$
  - Giá trị cực trị $t^{7543}_{.025} = 1.9603$ và $t^{7543}_{.005} = 2.5765$
  - $|t_{\beta_1}| > t_{critical}$ nên bác bỏ giả thuyết $H_0 \Rightarrow$ đi học có tác động đến thu nhập.

- Kiểm định một phía: $H_0 : \beta_1 < 0$
  - Trị kiểm định $t_{\beta_1} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = 33.76$
  - Giá trị cực trị $t^{7543}_{.01} = 2.3268$
  - $t_{\beta_1} > t_{critical}$ nên bác bỏ giả thuyết $H_0 \Rightarrow$ đi học có tác động dương đến thu nhập.

- Kiểm định một phía: $H_0 : \beta_1 > 0$
  - Lưu ý quy tắc bác bỏ H_0 là $t_{\hat{\beta}_j} < t_{critical}$.
  - Do $t_{\beta_1} > t_{critical}$ nên không bác bỏ giả thuyết $H_0 \Rightarrow$ đi học có tác động dương đến thu nhập, giống như trên.

Sử dụng $p - value$ để kiểm định giả thuyết

$p - value$ là xác suất tích lũy quan sát được vùng phân phối có giá trị kiểm định lớn hơn giá trị tới hạn, $t > t_{critical}$.



- p-value là diện tích vùng tô đậm (đối với kiểm định 2 phía) được tính từ giá trị $t = \pm 2.09$
- Đối với phân phối t với df $= 20$, diện tích phần tô đậm tương ứng với 0.025*2 $= 0.05$.

Sử dụng $p - value$ để kiểm định giả thuyết

$p - value$ là mức ý nghĩa thấp nhất mà giả thuyết H_0 có thể bị bác bỏ.

- Với kiểm định một phía, nếu $p - value < \alpha$ thì giả thuyết H_0 bị bác bỏ ở mức ý nghĩa α hay độ tin cậy $1 - \alpha$.
- Với kiểm định hai phía, nếu $p - value < \alpha/2$ thì giả thuyết H_0 bị bác bỏ ở mức ý nghĩa α hay độ tin cậy $1 - \alpha$.
- Trong Stata, sử dụng lệnh display ttail(df,t-stat) để tính p-value/2.

Ví dụ kiểm định giả thuyết về tỷ suất thu nhập của việc đi học bằng p-value

Kiểm định hai phía: $H_0 : \beta_1 = 0$

- Tương ứng với df $= 7{,}543$ và t-stat $= 33.76$ thì p-value $= 0.000 < 0.005 \Rightarrow$ bác bỏ giả thuyết H_0 ở độ tin cậy 99% $\Rightarrow$ đi học có tác động đến thu nhập.

Khoảng tin cậy (confidence interval)

- Khoảng tin cậy $1-\alpha$ của ước lượng của tham số β được tính bằng:

$$\hat{\beta} \pm t^{df}_{\alpha/2} * se(\hat{\beta})$$

- Ví dụ khoảng tin cậy 95% của tham số yoeduc trong mô hình tỷ suất thu nhập là:

$$[\beta_{lower} - \beta_{upper}] = .0926 \pm 1.96 * .0027 = [.0872 - .0980]$$

- Khoảng tin cậy này sẽ không chứa giá trị 0 nếu ước lượng của β có ý nghĩa thống kê.

Thực hành kiểm định giả thuyết

- Ước lượng các mô hình hồi quy đa biến, lần lượt đưa các biến giải thích khác nhau vào trong mô hình.
- Lựa chọn biến và mô hình phù hợp nhất.
- Kiểm định độ nhạy của ước lượng đối với các nhóm mẫu (stratification/subgroup): theo vùng miền, độ tuổi, phân tầng theo các đặc tính nhân khẩu học (gia đình, dân tộc, số người phụ thuộc, trình độ học vấn, kinh nghiệm làm việc, thu nhập...)