Hồi quy Đa biến
Kiểm định Giả thuyết

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

24/12/2019

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

Tương tự như các điều kiện của hồi quy đơn 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:

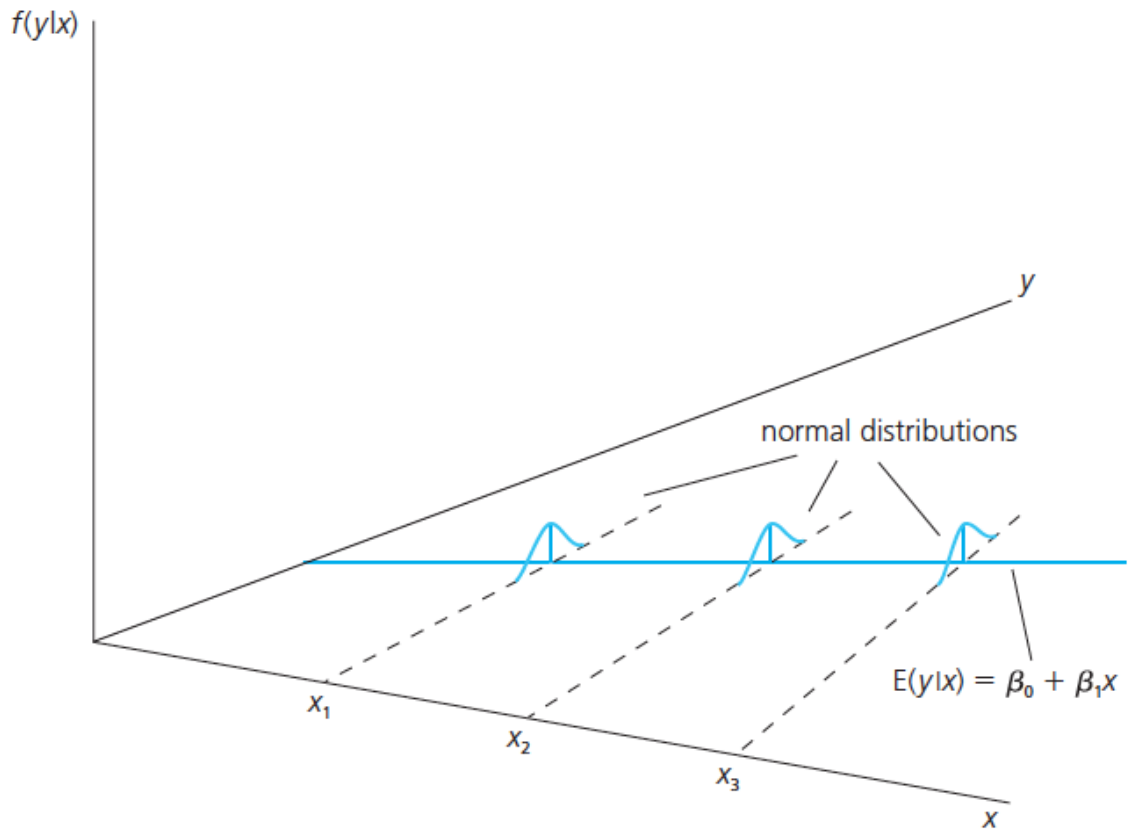
$$Var(u|x_1, ..., x_k) = \sigma^2$$

- 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.
  - Trong tất cả các ước lượng tuyến tính, OLS có phương sai của ước lượng là nhỏ nhất.
  - Không chệch, $E(\hat{\beta}) = \beta$.

Giả định về phân phối mẫu của sai số

6. Sai số u độc lập với các biến giải thích, 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 $\hat{\beta}$ là:

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

Viết dưới dạng chuẩn hóa:

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

Một số công thức đáng lưu ý

Đối với ước lượng của tham số β_j tương ứng với biến giải thích x_j, và mẫu dữ liệu có n quan sát:

$$sd(\hat{\beta}_j) = \frac{\sigma^2}{SST_j * (1 - R_j^2)}$$

trong đó, tổng biến thiên của x_j được tính như sau:

$$SST_j = \sum_{i=1}^{n} (x_{ij} - \overline{x_j})^2$$

và R_j^2 là hệ số thích hợp của hồi quy x_j lên các biến giải thích còn lại.

Một số nhận xét

- Tổng biến thiên SST_j của x_j càng lớn thì độ lệch chuẩn của $\hat{\beta}_j$ càng nhỏ. Để ước lượng càng chính xác thì yêu cầu có dữ liệu x_j thay đổi giữa các quan sát.
  - Dữ liệu điều tra ngẫu nhiên đảm bảo x_j khác nhau.
  - Không thể ước lượng được β_j nếu x_j không thay đổi. Ví dụ không thể ước lượng tỷ suất thu nhập của việc đi học nếu tất cả các quan sát có số năm đi học giống nhau là 12 năm.
- R_j^2 càng nhỏ, hay là x_j càng ít tương quan với các biến còn lại, thì độ lệch chuẩn của $\hat{\beta}_j$ càng nhỏ, và ước lượng $\hat{\beta}_j$ càng chính xác.

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

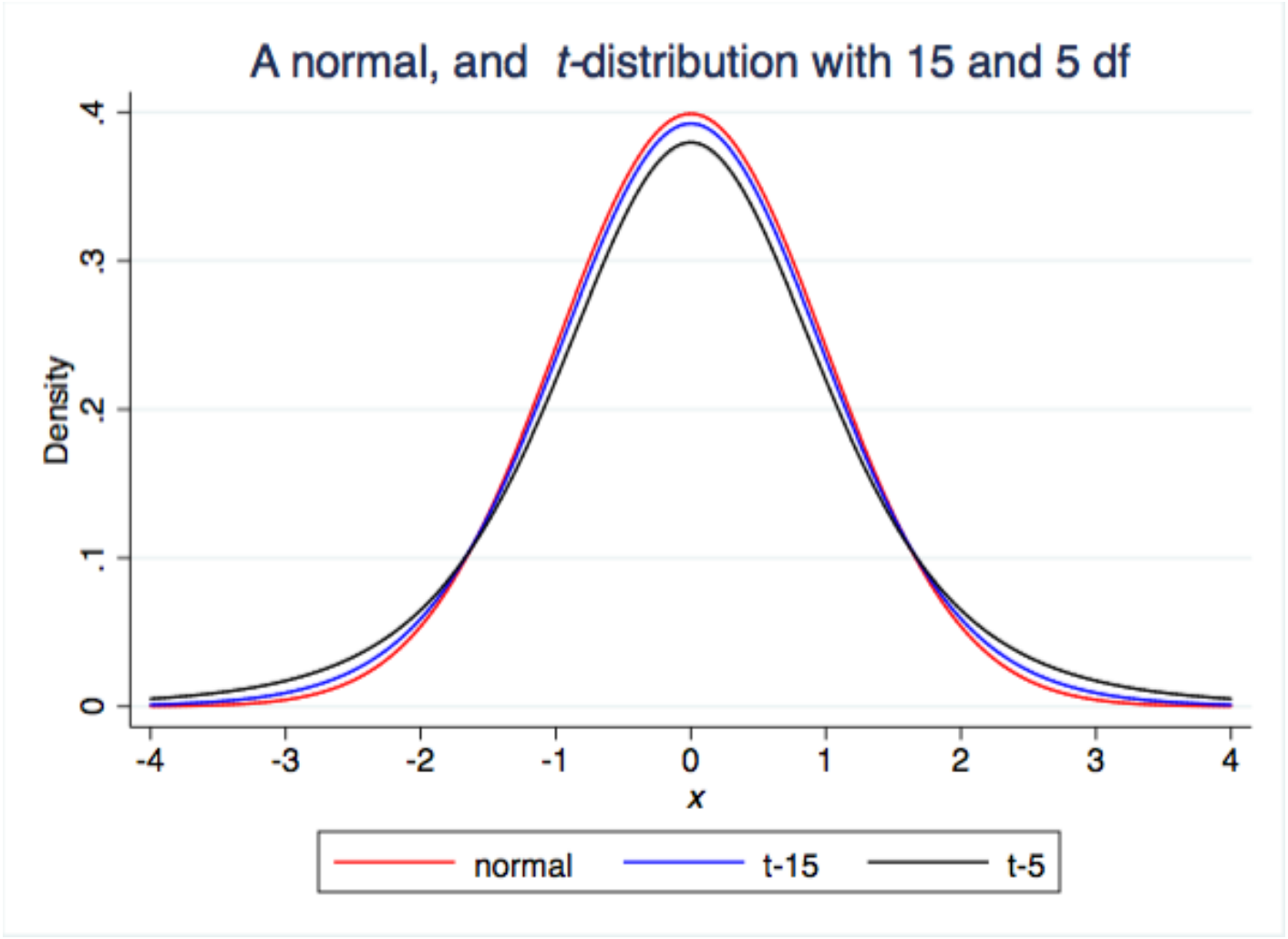
Từ các giả định CLRM, nhưng không biết phương sai σ^2 của sai số từ tổng thể (mặc dù biết là không đổi), các trị kiểm định của β_j dựa trên phân phối t-student được tính như sau:

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

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.

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

hoặc

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

- 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).

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

- Nếu $\beta_j \neq 0$ thì biến x_j 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, significance level), kiểm định một giả thuyết là xem xét liệu chúng ta có bác bỏ được giả thuyết khi thực tế giả thuyết là đúng với xác suất α.
  - 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 là mức ý nghĩa của kiểm định.
  - Sai lầm loại II liên quan đến sức mạnh thống kê (power of the test, $1 - \beta$). Sức mạnh thống kê là xác suất bác bỏ H_0 khi H_1 đúng.

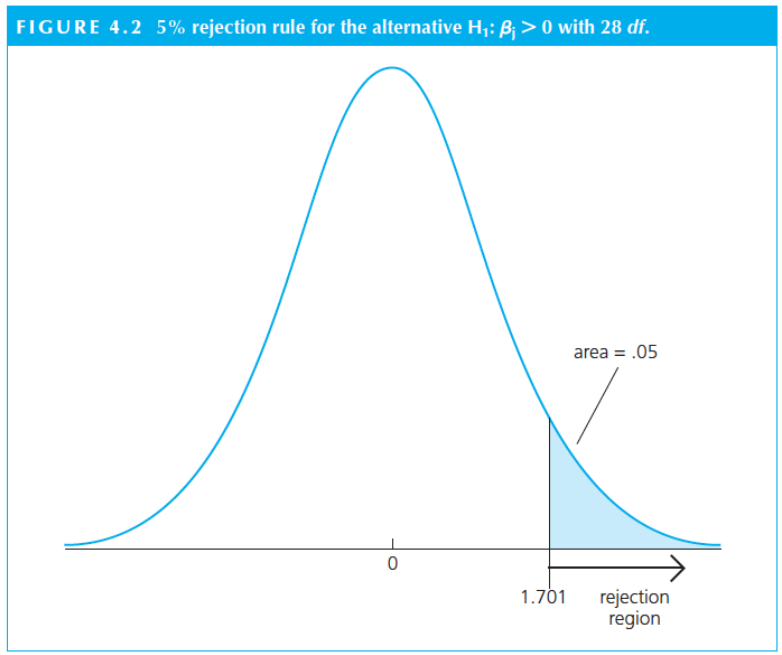
		Giả thuyết Ho	
		Đúng	**Sai**
Quyết định	**Không bác bỏ**	$1 - \alpha$ **[Đúng]**	β **[Sai]**
	Bác bỏ	α **[Sai]**	$1 - \beta$ **[Đúng]**

Kiểm định 1 phía (1-sided test)

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

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

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



Kiểm định 1 phía (1-sided test) (2)

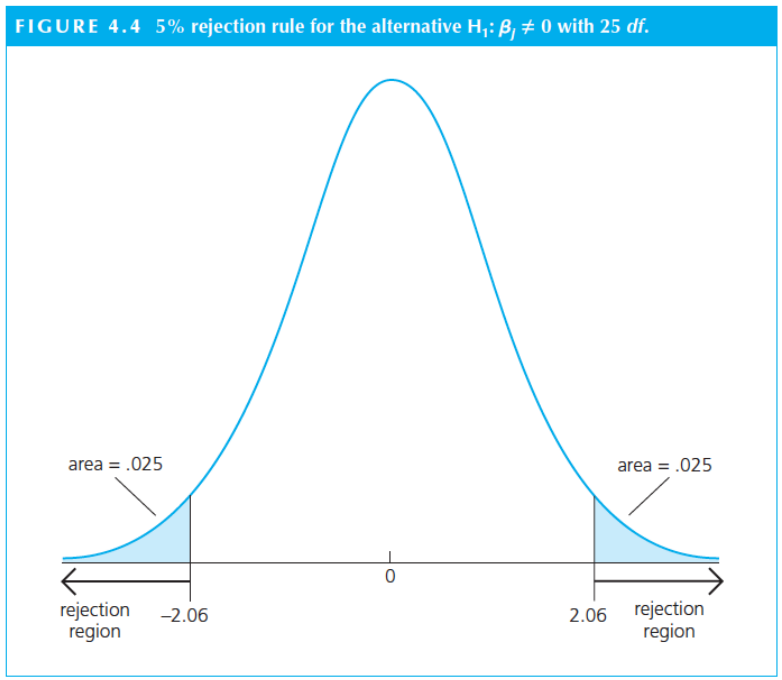
- H_0: Giả thuyết không (null hypothesis), $\beta_j \geq 0$
- H_1: Giả thuyết thay thế (alternative hypothesis), $\beta_j < 0$

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

Kiểm định 2 phía (2-sided test)

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

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



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

- Mức ý nghĩa α hoặc độ tin cậy (confidence level, $1-\alpha$): 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 $\hat{\beta}_j$ 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 VHLSS 2010, ước lượng lại mô hình tỷ suất thu nhập của đi học:

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

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

Source	SS	df	MS		
Model	1753.70541	8	219.213176	Number of obs	= 7,552
Residual	4040.86526	7,543	.535710627	F(8, 7543)	= 409.20
				Prob > F	= 0.0000
				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

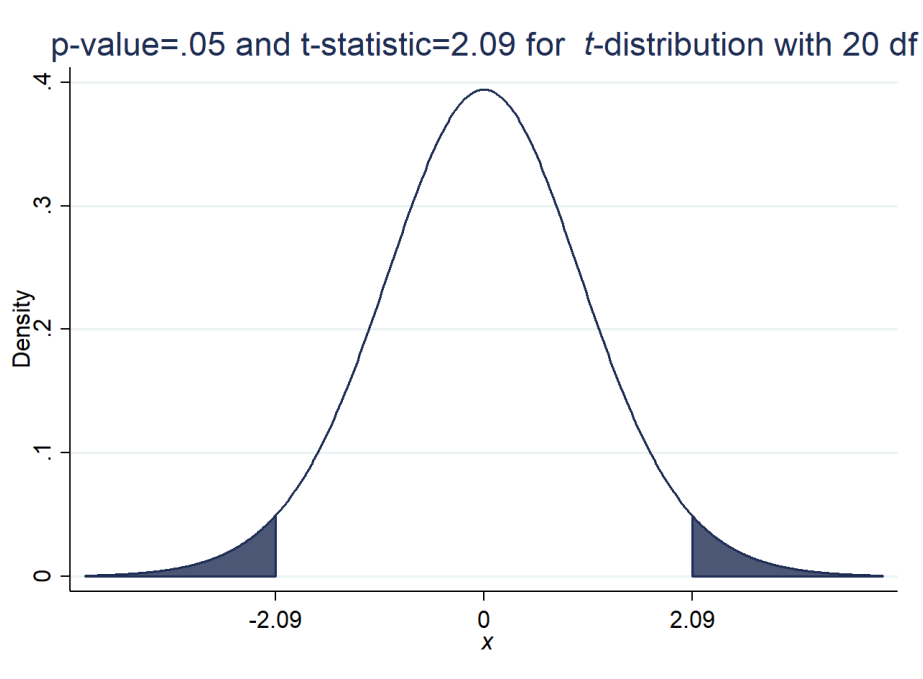
- Kiểm định hai phía: $H_0 : \beta_1 = 0$
  - Trị kiểm định $t_{\beta_1} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = 33.76$
  - Giá trị cực trị $t^{7543}_{.025} = 1.9603$ và $t^{7543}_{.005} = 2.5765$
  - $|t_{\beta_1}| > t_{critical}$ nên bác bỏ giả thuyết $H_0 \Rightarrow$ đi học có tác động đến thu 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 trị kiểm định lớn hơn 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 20 df, 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

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

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

- 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_j có ý nghĩa thống kê.

Hồi quy Đa biến
Cấu trúc Hàm và Lựa chọn Mô hình

Các loại kiểm định giả thuyết

1. Giả thuyết đơn: kiểm định đối với một tham số của mô hình.
2. Kiểm định điều kiện ràng buộc đối với các tham số.
3. Giả thuyết bội: kiểm định đồng thời nhiều tham số.
4. Kiểm định cấu trúc hàm.
5. Kiểm định mô hình gộp.
6. Kiểm định ước lượng từ hai mô hình riêng biệt.

Kiểm định điều kiện ràng buộc với tham số

Ví dụ ta muốn kiểm định H_0 là tỷ suất thu nhập của đi học bằng với tỷ suất thu nhập của kinh nghiệm làm việc, $\beta_1 = \beta_2$, trong mô hình:

$$\begin{aligned} log(income) =& \beta_0 + \beta_1 yoeduc + \beta_2 yoexper + \beta_3 married \\ &+ \beta_4 school + \beta_5 public + \beta_6 foreign + \beta_7 official + u \end{aligned}$$

Trị kiểm định được tính như sau:

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{se(\hat{\beta}_1 - \hat{\beta}_2)}$$

Có 2 cách thực hiện trong Stata:

1. test yoeduc = yoexper
   (1) yoeduc - yoexper = 0
   F(1, 7544) = 982.41
   Prob > F = 0.0000

2. Tạo ra biến mới sum = yoeduc + yoexper, ước lượng mô hình với biến sum, và kiểm định $\theta = 0$:

$$\begin{aligned} log(income) =& \beta_0 + \theta yoeduc + \beta_2 sum + \beta_3 married \\ &+ \beta_4 school + \beta_5 public + \beta_6 foreign + \beta_7 official + u \end{aligned}$$

Lưu ý trị kiểm định F-stat đối với một ràng buộc bằng trị kiểm định t-stat bình phương.

Kiểm định giả thuyết bội (multiple hypothesis test)

- Kiểm định đồng thời nhiều ràng buộc, ví dụ trong mô hình tỷ suất thu nhập ta muốn kiểm định số năm kinh nghiệm làm việc và số năm kinh nghiệm làm việc bình phương đồng thời không có tác động đến thu nhập.

$$log(income) = \beta_0 + \beta_1 yoeduc + \beta_2 yoexper + \beta_3 yoexper^2 + \beta_4 married + \beta_5 school + \beta_6 public + \beta_7 foreign + \beta_8 official + u$$

$$H_0 : \beta_2 = 0, \beta_3 = 0$$

so với H_1: ít nhất một trong các đẳng thức không đạt.

- Kiểm định giả thuyết bội khác với kiểm định từng biến riêng rẽ. Có thể các biến β_2 và β_3 không có ý nghĩa thống kê nhưng không đồng thời bằng không.

Mô hình gốc (còn gọi là mô hình không bị ràng buộc - **unrestricted model**) là:

$$
\begin{aligned}
&log(income) = \\
&\beta_0 + \beta_1 yoeduc + \beta_2 yoexper + \beta_3 yoexper^2 + \beta_4 married \\
&+ \beta_5 school + \beta_6 public + \beta_7 foreign + \beta_8 official + u
\end{aligned}
$$

Mô hình bị ràng buộc (**restricted model**) theo giả thuyết là:

$$
\begin{aligned}
log(income) =& \beta_0 + \beta_1 yoeduc + \beta_4 married \\
&+ \beta_5 school + \beta_6 public + \beta_7 foreign + \beta_8 official + u
\end{aligned}
$$

- Để kiểm định giả thuyết bội ta dựa vào tổng bình phương của phần dư SSR.
- Mô hình càng nhiều biến thì SSR càng nhỏ.
- Sự khác biệt giữa SSR của mô hình bị ràng buộc (SSR_R) và mô hình không bị ràng buộc (SSR_U) có thể dùng để kiểm định của việc thiếu biến trong mô hình.
- Trị kiểm định có phân phối $F_{q,n-k-1}$, với q là số ràng buộc của mô hình bị ràng buộc:

$$F = \frac{(SSR_R - SSR_U)/q}{SSR_U/(n-k-1)}$$

- Kiểm định F còn gọi là kiểm định Wald.

Ví dụ với mô hình tỷ suất thu nhập

- $H_0 : \beta_2 = 0, \beta_3 = 0 \Rightarrow q = 2,\ n - k - 1 = 7543$

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

Source	SS	df	MS		
Model	1753.70541	8	219.213176	Number of obs	= 7,552
Residual	4040.86526	7,543	.535710627	F(8, 7543)	= 409.20
				Prob > F	= 0.0000
				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

```
.
. test yoexper = yoexpersq = 0

 ( 1)  yoexper - yoexpersq = 0
 ( 2)  yoexper = 0

       F(  2,  7543) =  310.83
            Prob > F =    0.0000
```

Dùng kiểm định bội để xác định cấu trúc hàm

- R^2, R^2_{adj} đã được sử dụng để lựa chọn biến số và cấu trúc hàm số.
- F-test cũng có thể sử dụng để kiểm định cấu trúc hàm số trong các mô hình lồng ghép (nested models). Ví dụ mô hình (1) được lồng ghép trong mô hình (2):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \quad (1)$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + u \quad (2)$$

  - Kiểm định $H_0 : \beta_3 = \beta_4 = 0$ để biết liệu hai mô hình trên là tương đương hay không. Nếu bác bỏ H_0 thì mô hình (1) được lồng ghép trong mô hình (2).
- Nếu kiểm định tất cả các tham số trong mô hình $\Rightarrow$ ý nghĩa thống kê của mô hình tổng quát (overall significance of the regression).
  - Trong mô hình tỷ suất thu nhập, trị kiểm định $F_{8,7543} = 409.02$, p-value $= 0.000$.

Kiểm định khác biệt giữa các nhóm trong cùng một mô hình - Chow test

Chúng ta muốn kiểm định liệu mô hình tỷ suất thu nhập của việc đi học giống nhau giữa nhóm nam và nữ.

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

- Chúng ta đã ước lượng mô hình trên cho nhóm nam và nữ riêng biệt và quan sát thấy tỷ suất thu nhập của việc đi học với nhóm nữ cao hơn nhóm nam.
- Câu hỏi: Sự khác biệt có ý nghĩa thống kê hay không?

Trị kiểm định của Chow-test $F_{k+1,n-2(k+1)}$ được tính như sau:

$$F = \frac{[SSR_p - (SSR_1 + SSR_2)]/(k+1)}{[SSR_1 + SSR_2]/(n-2(k+1))}$$

trong đó

- Giả thuyết H_0: Tất cả các tham số ước lượng của mô hình nam và nữ là giống nhau.
- k là số biến giải thích trong mô hình (+1 do thêm tham số tung độ gốc)
- SSR_p, SSR_1, SSR_2 là tổng bình phương phần dư của hồi quy gộp toàn bộ dữ liệu, của nhóm nam, và nhóm nữ.

Ví dụ với mô hình tỷ suất thu nhập

$$F = \frac{[4040.8653 - (2234.8287 + 1649.6582)]/(8+1)}{[2234.8287 + 1649.6582]/(7552 - 2(8+1))} = 33.699694$$

- Giá trị cực trị của $F_{k+1,n-2(k+1)}$ tại mức tin cậy 99% là F(9,7534,.99) = 2.4096768 $\Rightarrow$ Bác bỏ H_0.

Thực hiện Chow-test bằng kiểm định bội

$$\begin{aligned} log(income) =& \beta_0 + \beta_1 yoeduc + \beta_2 yoexper + \beta_3 yoexpersq + \beta_4 married \\ &+ \beta_5 school + \beta_6 public + \beta_7 foreign + \beta_8 official \\ &+ \mathbf{male} * [\beta_\mathbf{0} + \beta_\mathbf{1} \mathbf{yoeduc} + ... + \beta_\mathbf{8} \mathbf{official}] + \mathbf{u} \end{aligned}$$

- Tạo biến tương tác giữa biến giới tính (male) với các biến giải thích.
- Ước lượng mô hình gộp bao gồm 2(k+1) = 18 biến giải thích.
- Nếu không có sự khác biệt giữa các nhóm nam và nữ thì các tham số ứng với các biến tương tác sẽ đồng thời bằng không.
- Dùng kiểm định bội đối với mô hình ràng ruộc (nhóm nam và nữ giống nhau) và không ràng buộc (nam và nữ khác nhau).
- Đối chiếu với cách kiểm định dựa trên SSR phía trên.

Thực hiện Chow-test với một số biến giải thích

Ví dụ chúng ta chỉ muốn kiểm định tỷ suất thu nhập của việc đi học giữa hai nhóm nam và nữ.

- Tạo tương tác giữa biến male với biến yoeduc, $dyoeduc = male * yoeduc$.
- Đưa 2 biến male và dyoeduc vào mô hình và ước lượng.
- Kiểm định $H_0 : male = dyoeduc = 0$.
- Nếu H_0 bị bác bỏ nghĩa là $male \neq 0$ (tung độ gốc khác nhau) hoặc $dyoeduc \neq 0$ (hệ số góc khác nhau), hoặc cả hai.

Hồi quy với Biến Định tính
(Regression with Qualitative Variables)

Biến định tính là gì

- Còn được gọi là biến giả (dummy variable)
- Là biến mô tả trạng thái (nam/nữ, đi làm/đi học, làm nông/công chức)
- Có thể là biến nhị phân (có/không) hoặc biến nhóm (categorical variable - có nhiều hơn 2 trạng thái giá trị, ví dụ phương tiện đi lại là ô tô/xe máy/xe đạp/đi bộ)
- Đa số trường hợp các biến định tính không thể xếp được thứ bậc (ví dụ làm việc trong khu vực nhà nước/tư nhân/nước ngoài).
- Một số trường hợp biến định tính có thể xếp được thứ bậc, ví dụ bằng cấp cao nhất có được là gì, từ không có bằng cấp, bằng tiểu học, THCS,THPT, cao đẳng, đại học, thạc sỹ, tiến sỹ.

- Không nhầm lẫn với biến số đếm rời rạc, ví dụ biến số con cái trong gia đình không phải là biến định tính.
- Thống kê mô tả biến định tính khác với biến định lượng.
  - Cần xác định nhóm tham chiếu (baseline/reference group) và nhóm được tham chiếu. Ví dụ với biến giới tính thì có thể đặt nhóm tham chiếu là nữ và nhóm được tham chiếu là nam.
  - Giá trị trung bình diễn giải xác suất xảy ra một sự kiện.
  - Giá trị lớn nhất và nhỏ nhất không có ý nghĩa kinh tế.
  - Sai số chuẩn liên quan đến xác suất quan sát được sự kiện.
  - Hệ số tương quan mẫu (correlation coefficient) không có ý nghĩa.
  - Thường dùng biến định tính để phân tách và so sánh giữa các nhóm, ví dụ nhóm nam và nữ.

Xử lý biến định tính

Sử dụng lại bộ dữ liệu VHLSS 2010.

- Cần hiểu cách mã hóa biến trong bảng dữ liệu.
- Có thể gộp biến nhóm thành biến nhị phân.
- Có thể tách biến nhóm thành nhiều biến nhị phân.
- Bẫy biến giả (dummy trap): Một biến định tính có n giá trị thì có thể tách ra tối đa là $n-1$ biến giả. Nếu tách làm n biến giả đưa vào mô hình sẽ có hiện tượng đa cộng tuyến hoàn hảo.

Hồi quy với biến định tính

Ước lượng mô hình tỷ suất thu nhấp của đi học với các biến định tính là có gia đình, học trường công, làm nhà nước, làm nước ngoài, là công chức:

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

Giải thích ý nghĩa của biến định tính

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

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

```
.
```

Diễn giải ý nghĩa của tham số ước lượng đối với biến định tính

- Nếu biến phụ thuộc là **thu nhập** thì tham số ước lượng là tác động tăng thêm của nhóm được tham chiếu so với nhóm tham chiếu.

- Nếu biến phụ thuộc là **log của thu nhập** thì diễn giải tham số ước lượng tùy thuộc vào biến giải thích là biến liên tục hay biến rời rạc.
  - Với **biến liên tục**, ví dụ số năm đi học *yoeduc*, hệ số ước lượng là % tăng thêm của thu nhập. Ví dụ 1 năm đi học làm tăng thu nhập 9.26%.

- Với **biến rời rạc**, ví dụ các biến định tính, hoặc nếu có biến số con trong gia đình, thì:
  - Nếu β nhỏ, β có thể coi là phần trăm tăng thêm của biến phụ thuộc.
  - Công thức tính chính xác đối với tác động của biến rời rạc lên biến phụ thuộc **log(Y)** là:

$$\frac{Y_1 - Y_0}{Y_0} = e^{\beta} - 1$$

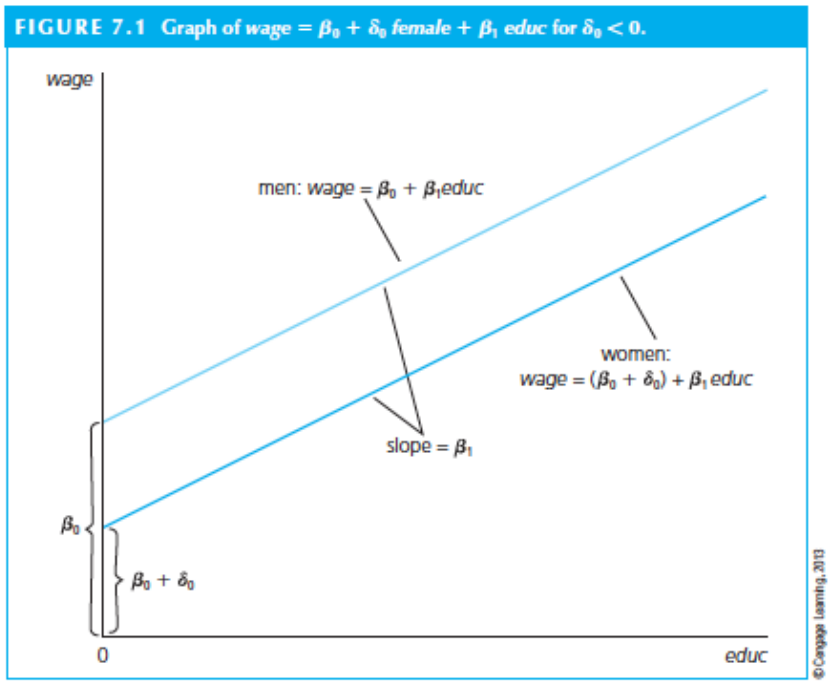
- Trong ví dụ trên:
  - Làm việc trong khu vực nước ngoài thu nhập cao hơn khu vực tư là: $2.718^{.45} - 1 = .5682$ hay 56.82% (chứ không phải là 45%).
  - Làm việc trong khu vực nhà nước thu nhập thấp hơn khu vực tư là: $2.718^{-1043} - 1 = -.099$ hay 9.9%.
  - Nếu coi *yoeduc* là biến rời rạc thì với mỗi năm học tăng thêm thu nhập là $2.718^{.0926} - 1 = .097$ hay 9.7%.

Tung độ gốc trong mô hình hồi quy

Với biến giới tính *male* trong mô hình:

$$log(income) = \beta_0 + \beta_1 yoeduc + \beta_2 yoexper + ... + \sigma_0 \mathbf{male} + u$$

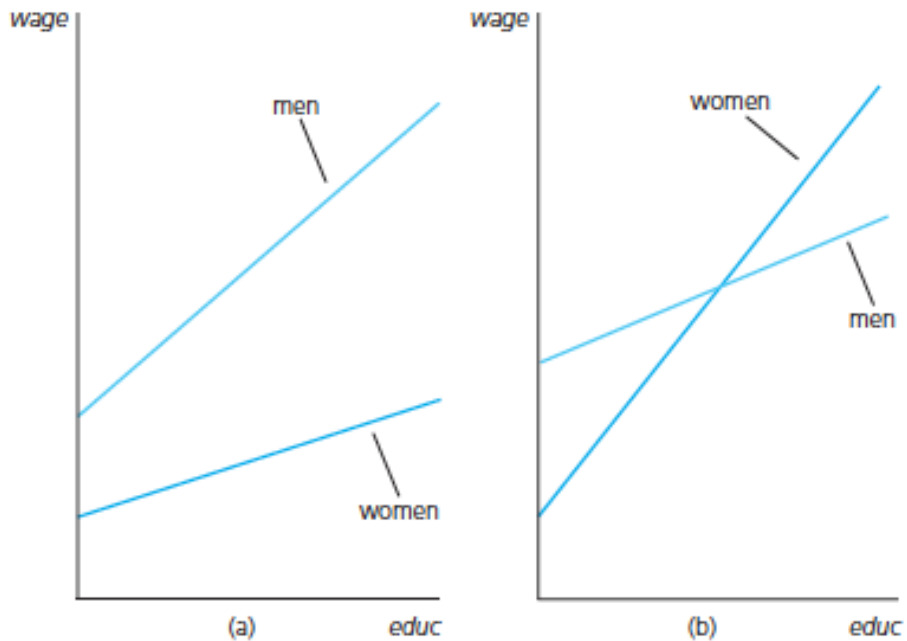
- Tung độ gốc là β_0 với nhóm nữ, và $\beta_0 + \sigma_0$ với nhóm nam
- Hệ số góc là β_1 giống nhau với cả hai nhóm (đường hồi quy song song)
- Nếu $\sigma_0 = 0$ thì hai đường hồi quy trùng nhau



Tung độ gốc và hệ số góc trong mô hình hồi quy với biến tương tác

$$log(income) = \beta_0 + \beta_1 yoeduc + \beta_2 yoexper + ... + \sigma_0 \mathbf{male} + \sigma_1 \mathbf{male} * \mathbf{yoeduc} + u$$

- Tung độ gốc là β_0 với nhóm nữ, và $\beta_0 + \sigma_0$ với nhóm nam
- Hệ số góc là β_1 với nhóm nữ, và $\beta_1 + \sigma_1$ với nhóm nam.
- Hai đường hồi quy chỉ trùng nhau khi σ_0 và σ_1 đồng thời bằng 0.



Kiểm định khác biệt theo nhóm

- Tung độ gốc khác nhau $\Rightarrow$ t-test nếu $\sigma_0 = 0$
- Tung độ gốc và hệ số góc khác nhau $\Rightarrow$ F-test nếu $\sigma_0 = \sigma_1 = 0$
- Tất cả các tham số của hai nhóm khác nhau $\Rightarrow$ Chow test

Ôn tập các loại kiểm định

- Kiểm định đơn: $H_0 : \sigma_0 = 0$

$$t_{\hat{\sigma}_0} \sim t_{n-k-1}$$

- Kiểm định bội: $H_0 : \sigma_0 = \sigma_1 = 0$

$$F = \frac{(SSR_R - SSR_U)/q}{SSR_U/(n-k-1)} \sim F_{q,n-k-1}$$

- Kiểm định khác biệt nhóm (tất cả các tham số):

$$H_0 : \sigma_0 = \sigma_1 = ... = \sigma_k = 0$$

$$F = \frac{[SSR_p - (SSR_1 + SSR_2)]/(k+1)}{[SSR_1 + SSR_2]/(n-2(k+1))} \sim F_{k+1,n-2(k+1)}$$