

COURSE

# Tương quan & Hồi quy tuyến tính

---

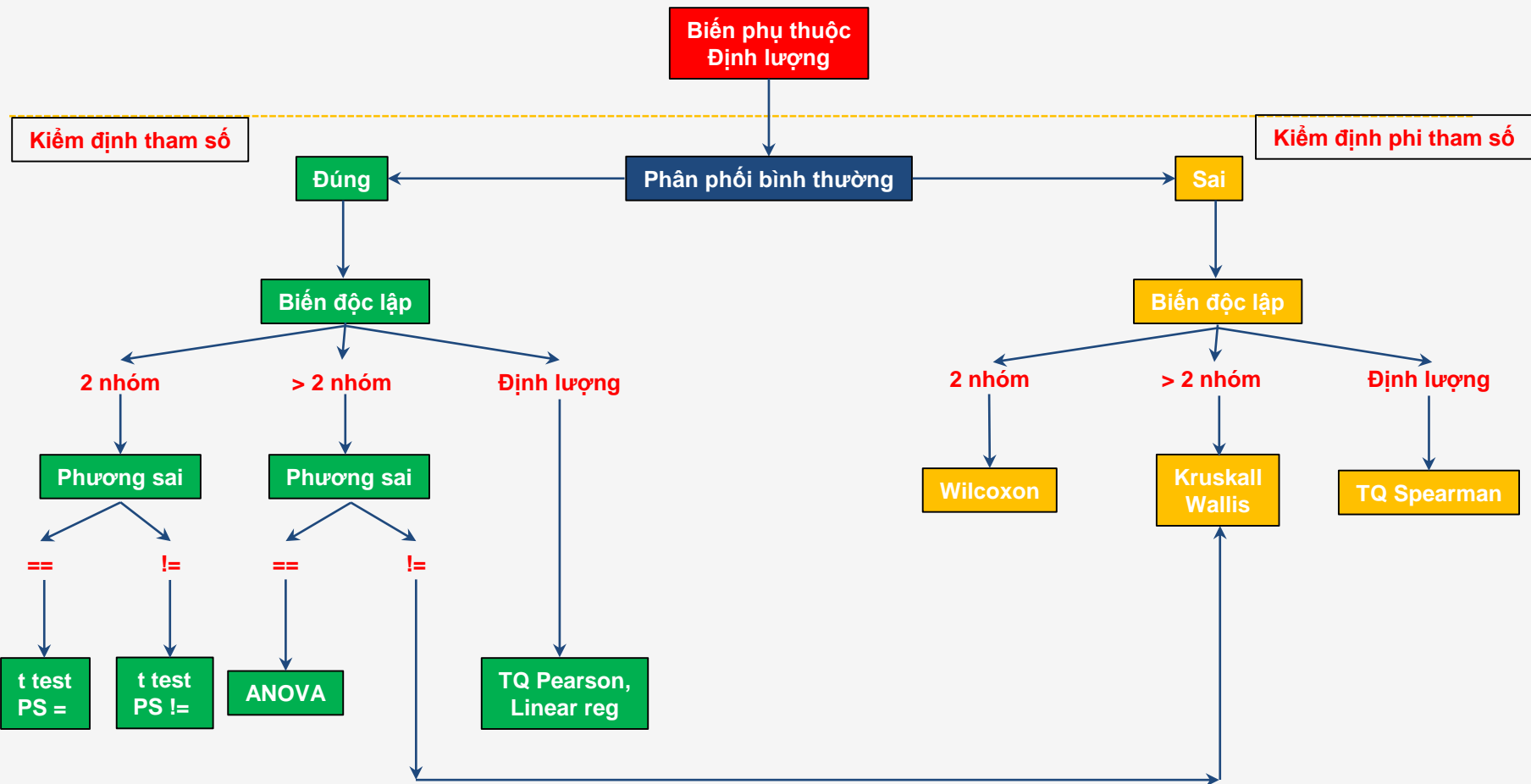
Lớp phân tích thống kê cơ bản

Khương Quỳnh Long  
Hà Nội, 06-08/06/2020

# Nội dung

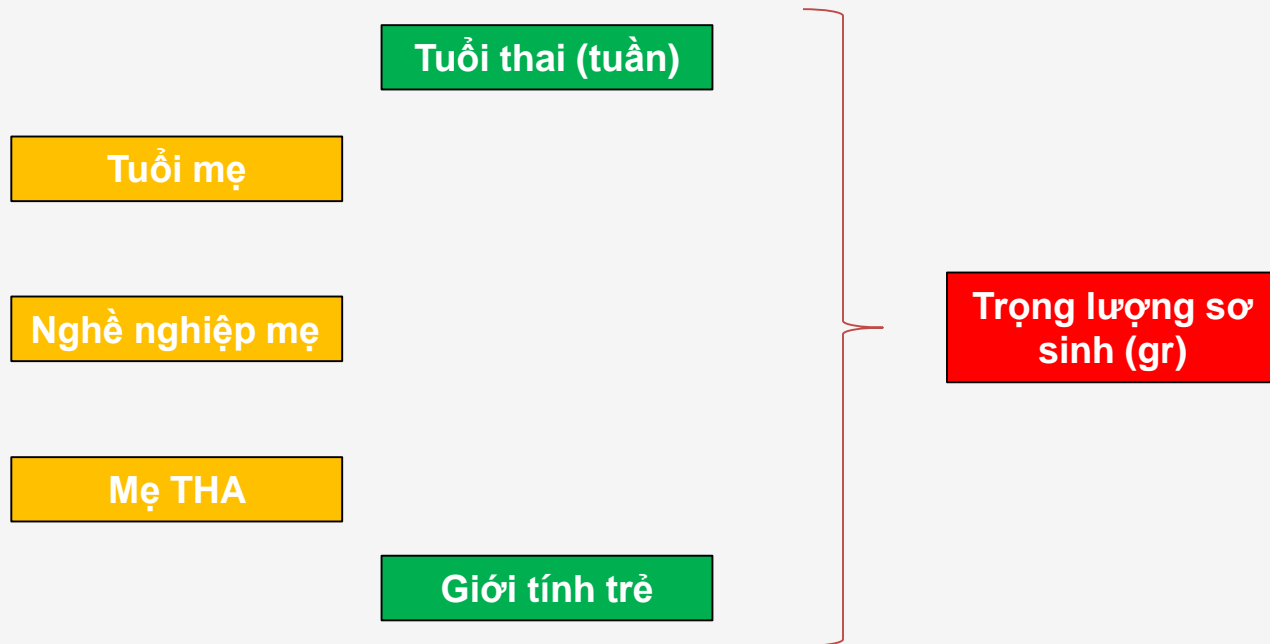
---

1. Phân tán đồ
2. Hệ số tương quan
3. Hồi quy tuyến tính đơn biến
4. Hồi quy tuyến tính đa biến



# Tình huống nghiên cứu

- Nghiên cứu nhằm khảo sát các yếu tố ảnh hưởng tới trọng lượng sơ sinh của trẻ từ ivf (*dữ liệu tlsosinh.dta*)

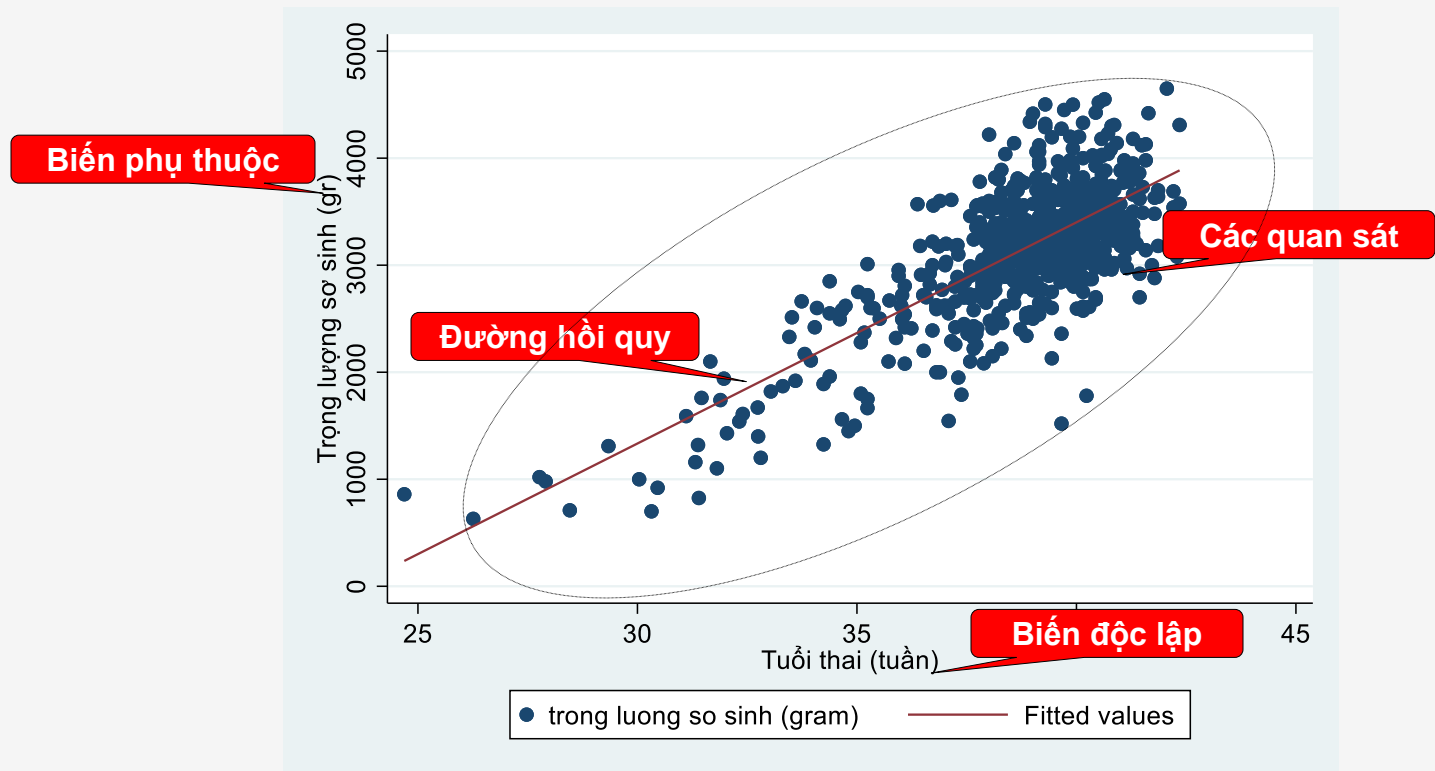


# Câu hỏi?

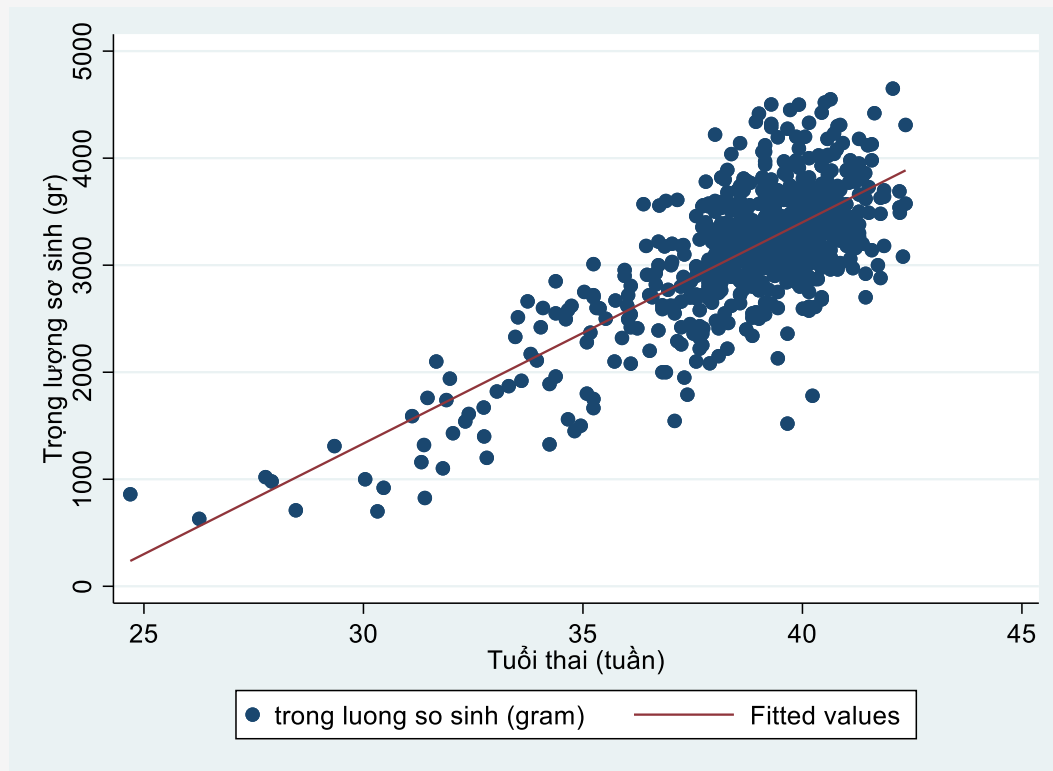
---

- Mối liên quan giữa tuổi thai và trọng lượng sơ sinh?

# Phân tán đồ (scatter plot)

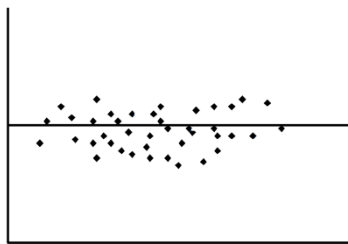


- Nhận xét?

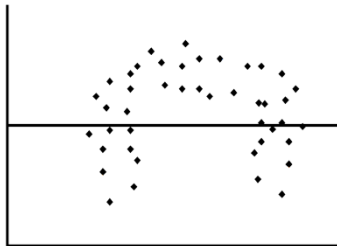


# Phân tán đồ (scatter plot)

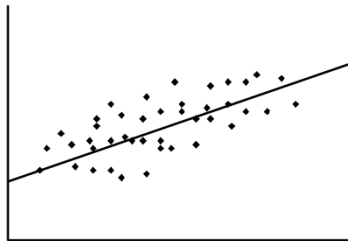
(a) không tương quan



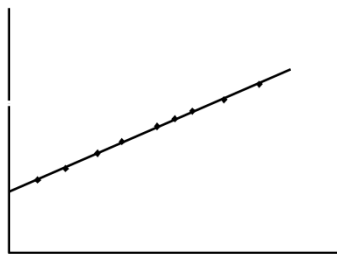
(b) không có mối liên hệ tuyến tính



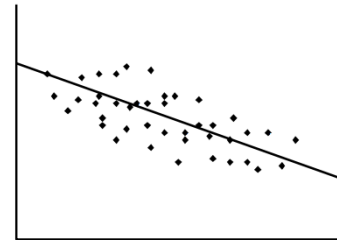
(c) Tương quan thuận



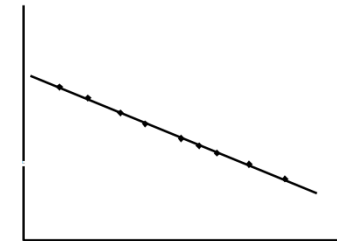
(d) Tương quan thuận (hoàn toàn)



(e) Tương quan nghịch



(f) Tương quan nghịch (hoàn toàn)





# Phân tán đồ (scatter plot)

---

- Thực hiện bằng Stata

- Cơ bản

```
twoway (scatter biếnphụthuộc biểndộclập)
```

```
twoway (scatter tlosinh tuoi thai)
```

- Thêm đường hồi quy

```
twoway (scatter biếnphụthuộc biểndộclập) (lfit biếnphụthuộc biểndộclập)
```

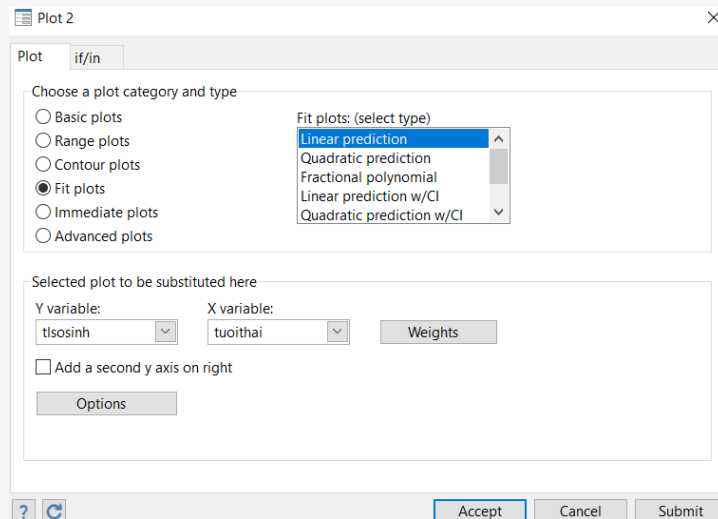
```
twoway (scatter tlosinh tuoi thai) (lfit tlosinh tuoi thai)
```

- Thêm các label, legend...

```
twoway (scatter tlosinh tuoi thai) (lfit tlosinh tuoi thai), ytitle(Trọng  
lượng sơ sinh (gr)) xtitle(Tuổi thai (tuần))
```

# Phân tán đồ (scatter plot)

- Thực hiện bằng Stata
- Graphics → twoway (scatter, line, etc.)
- Create → chọn biến X (độc lập) và Y (phụ thuộc) → Accept
- Create → chọn Fit plots → chọn biến X và Y → Accept
- Chọn OK
- Tùy chỉnh label, legend...



# Hệ số tương quan

- Hệ số thể hiện mối liên hệ tuyến tính giữa 2 biến định lượng
- Tương quan **Pearson**

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

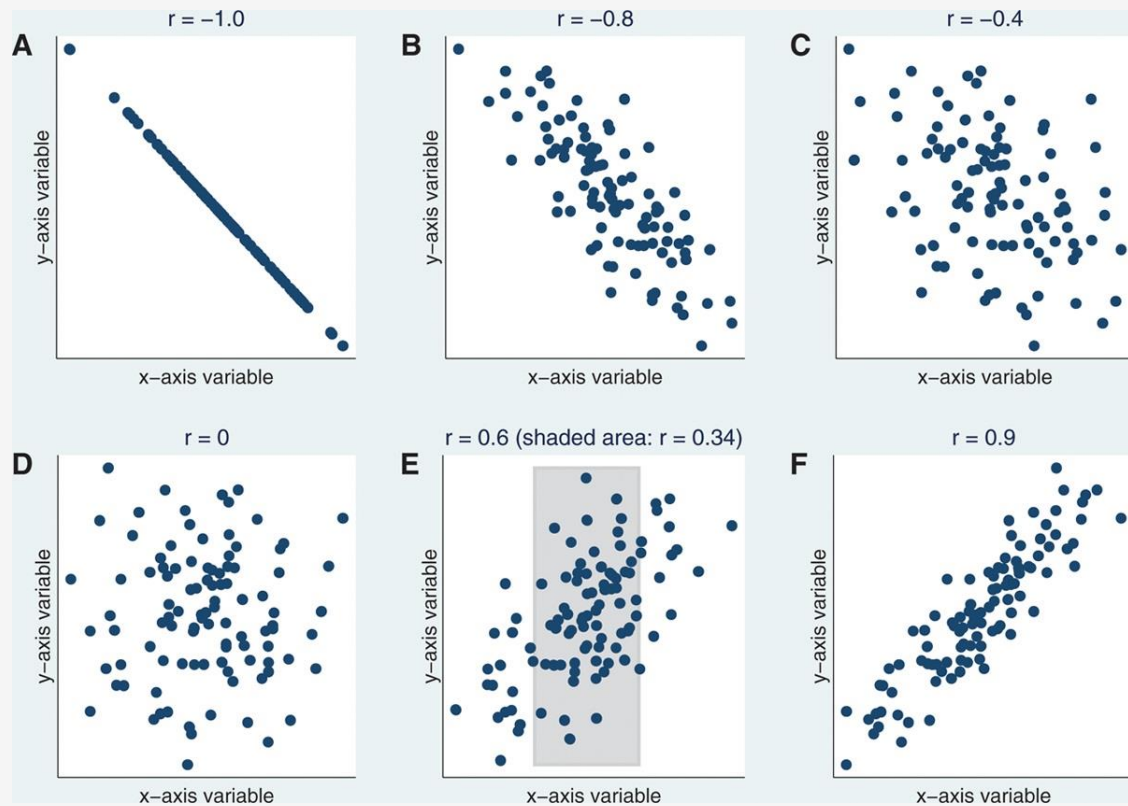
- Điều kiện sử dụng hệ số tương quan **Pearson**:
  - ✓ Biến kết cuộc (phụ thuộc) có **phân phối bình thường**
  - ✓ Biến độc lập và phụ thuộc có **tương quan tuyến tính**

# Hệ số tương quan

---

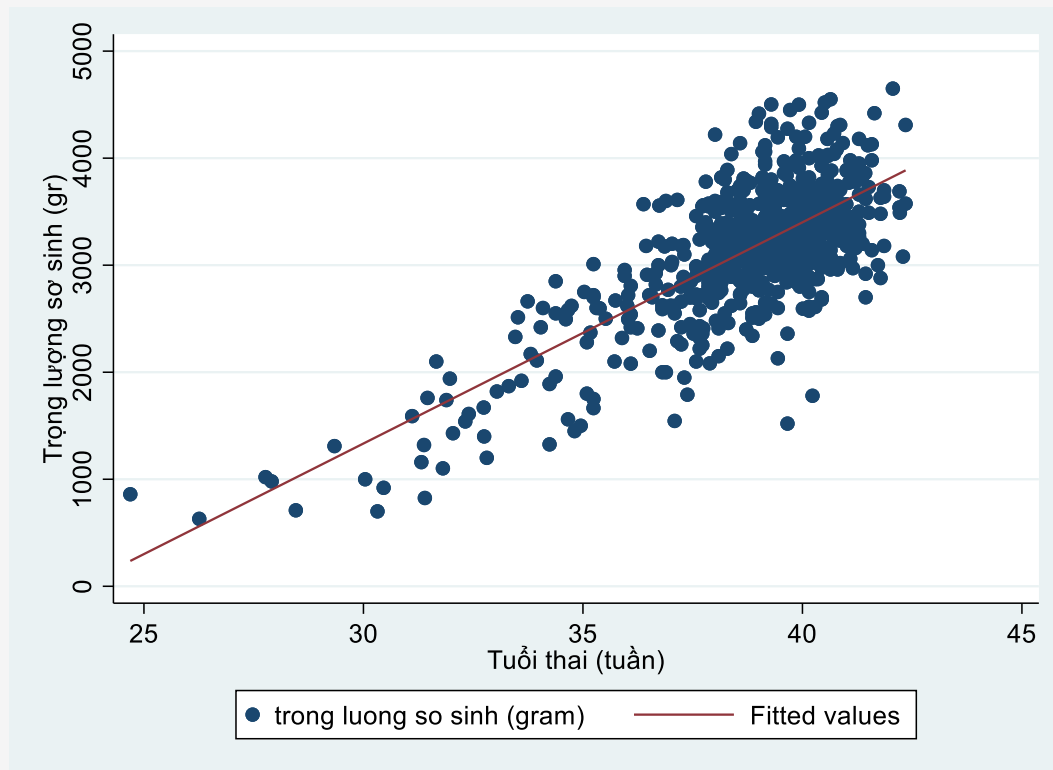
- Hệ số tương quan:
  - ✓  $r \in [-1 : 1]$
  - ✓  $r > 0$ : tương quan thuận;
  - ✓  $r < 0$ : tương quan nghịch
  - ✓  $r = 0$ : không tương quan
- Mức độ tương quan (trị số tuyệt đối của  $r$ )<sup>1</sup>
  - ✓ 0.00–0.10: không đáng kể
  - ✓ 0.10–0.39: tương quan yếu
  - ✓ 0.40–0.69: tương quan vừa
  - ✓ 0.70–0.89: tương quan mạnh
  - ✓ 0.90–1.00: tương quan rất mạnh

# Hệ số tương quan



# Hệ số tương quan

$r = 0.74$ , nhận xét?



# Hệ số tương quan

---

- Thể hiện mối liên hệ tuyến tính, nếu  $r = 0$ 
  - ✓ Không có mối liên hệ giữa 2 biến
  - ✓ Có mối liên hệ nhưng không phải tuyến tính
- Không có đơn vị đo lường
- Hai chiều (nếu X hoặc Y thay đổi thì biến còn lại thay đổi như thế nào)
- Không thay đổi bởi những phép biến đổi tuyến tính
- Có thể có cùng giá trị r nhưng phân tán đồ khác nhau
- Đơn biến, không kiểm soát được nhiều

# Hệ số tương quan

---

- $R^2$ : tỷ lệ biến thiên của biến phụ thuộc được giải thích bởi biến độc lập
- Ví dụ:  $r = 0.74 \rightarrow R^2 = 0.55$   
 $\rightarrow$  tuổi thai giải thích được 55% sự biến thiên của trọng lượng sơ sinh



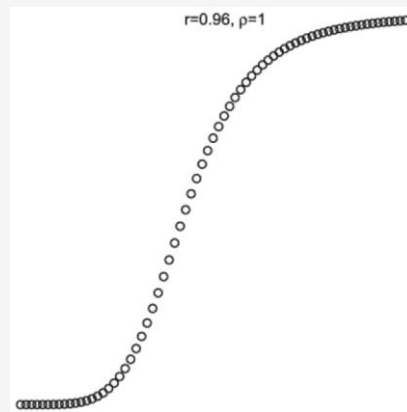
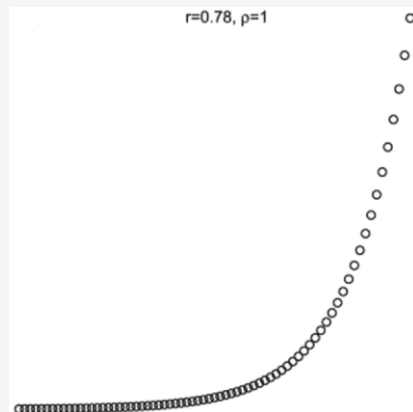
# Hệ số tương quan

---

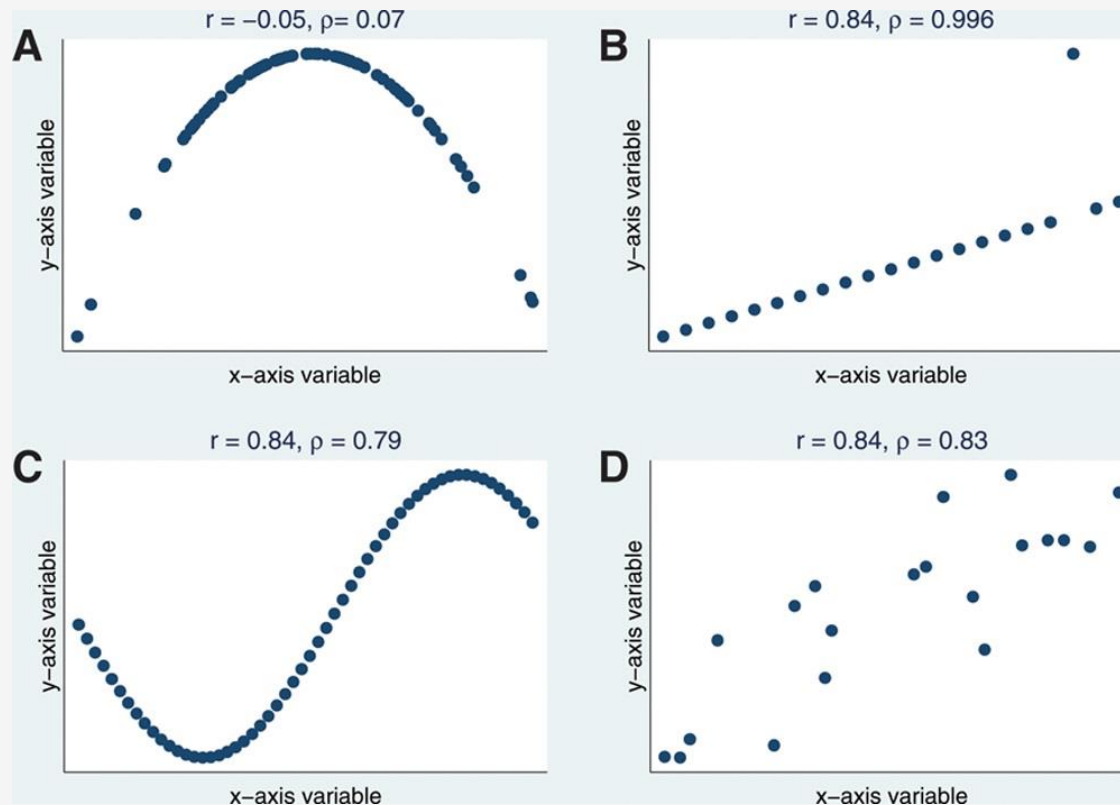
- Kiểm định ý nghĩa cho hệ số tương quan:
- $H_0: r = 0$ : không có tương quan tuyến tính
- $H_a: r \neq 0$ : có tương quan tuyến tính
- Kết luận có ý nghĩa thống kê dựa vào giá trị p
- Tóm lại:
  - ✓ Kết luận **mức độ, xu hướng** tương quan  $\rightarrow r$
  - ✓ Kết luận có **ý nghĩa thống kê**  $\rightarrow p$

# Hệ số tương quan

- Hệ số tương quan **Spearman** ( $\rho$ : rho)
- Đo lường sự tương quan giữa 2 biến:
  - ✓ Phân phối không bình thường
  - ✓ Biến thứ tự
  - ✓ Tương quan monotonic



# Tương quan Pearson vs Spearman

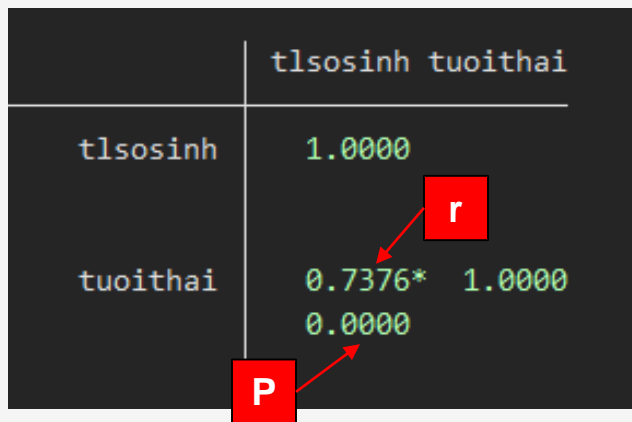


# Hệ số tương quan

- Thực hiện bằng Stata (tương quan Pearson)

```
pwcorr cácbiến số, sig star(5)
```

```
pwcorr tlsosinh tuoi thai, sig star(5)
```



	tlsosinh	tuoi thai
tlsosinh	1.0000	
tuoi thai	0.7376* 0.0000	1.0000

# Hệ số tương quan

- Thực hiện bằng Stata (tương quan Spearman)

spearman **cácbiến số**

spearman **tlsosinh tuoi thai**

```
. spearman tlsosinh tuoi thai  
  
Number of obs =      641  
Spearman's rho =      0.5700  
  
Test of Ho: tlsosinh and tuoi thai are independent  
Prob > |t| =      0.0000
```

**P**

**$\rho$**

# Hồi quy tuyến tính

# Mục tiêu

---

- Tìm phương trình để **diễn giải** mối liên quan giữa biến độc lập và phụ thuộc
  - ✓ Nếu biến  $x$  thay đổi thì biến  $y$  thay đổi như thế nào?
- Đưa ra mô hình **tiên lượng** (dự báo)
  - ✓ Với giá trị của  $x = \dots$  thì  $y$  là bao nhiêu?
- Hiệu chỉnh các yếu tố gây nhiễu (đa biến)

# Hồi quy tuyến tính đơn biến

---

- Phương trình:

$$Y = \alpha + \beta X + \varepsilon$$

- $\alpha$ : Điểm chặn/hằng số (intercept)
- $\beta$ : Hệ số góc (slope)
- $\varepsilon$ : Sai số ngẫu nhiên/phần dư (random error/residual)



# Hồi quy tuyến tính đơn biến

---

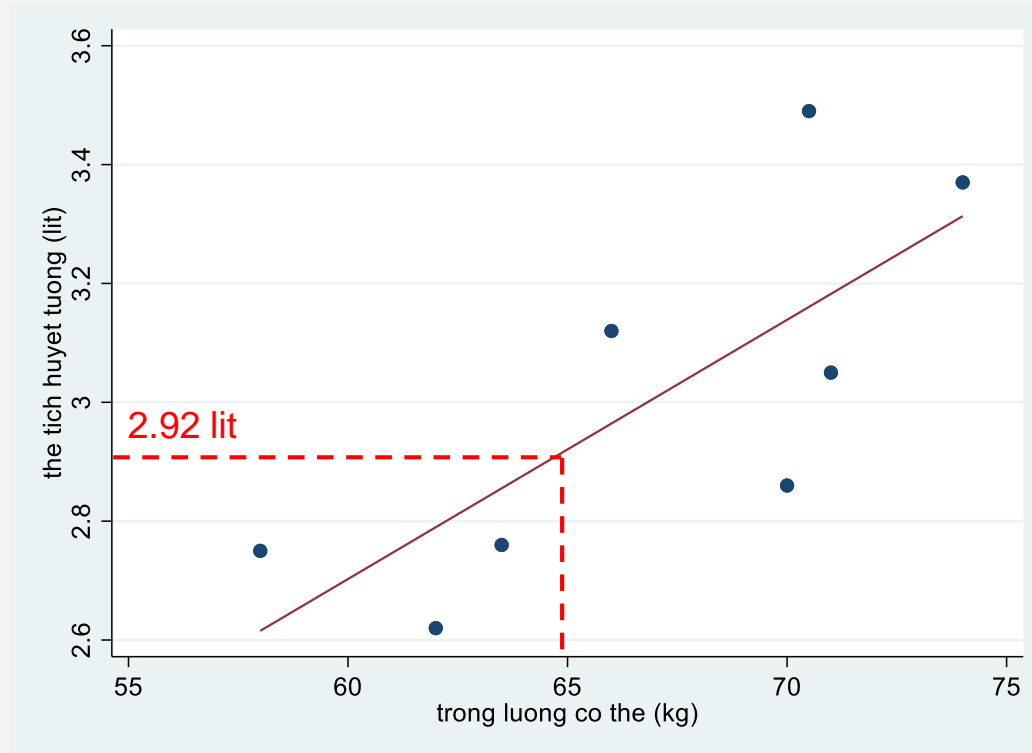
Ví dụ:

Đối tượng	Trọng lượng (kg)	Thể tích huyết tương (lit)
1	58.0	2.75
2	70.0	2.86
3	74.0	3.37
4	63.5	2.76
5	62.0	2.62
6	70.5	3.49
7	71.0	3.05
8	66.0	3.12

1. Nếu trọng lượng cơ thể tăng 1 kg thì thể tích huyết tương thay đổi bao nhiêu lit?
2. Một người có cân nặng 65 kg thì thể tích huyết tương là bao nhiêu lit?

# Hồi quy tuyến tính đơn biến

Thể tích huyết tương =  $0.0857 + 0.04362 \times \text{trọng lượng cơ thể}$



# Hồi quy tuyến tính đa biến

---

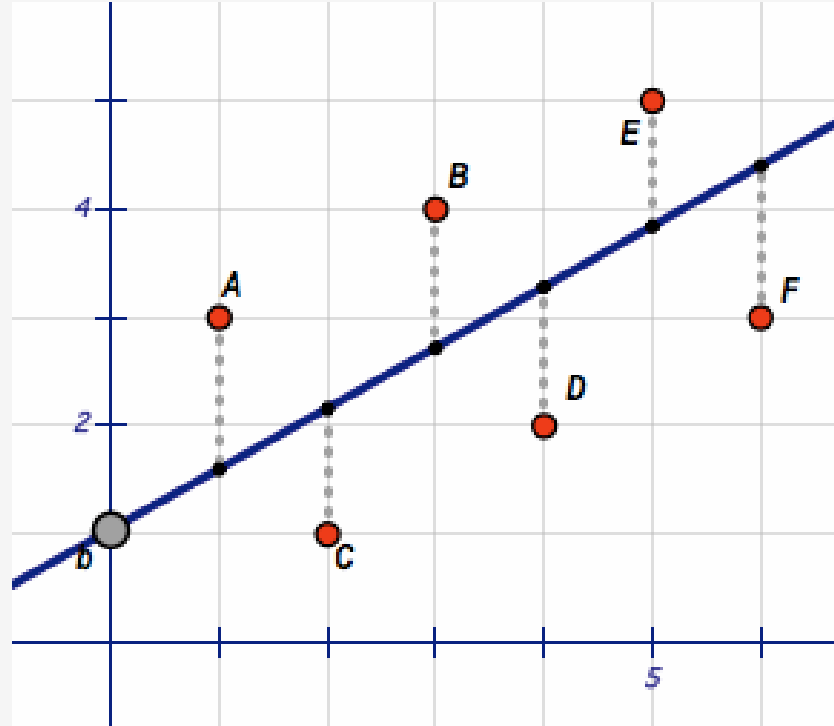
- Phương trình:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon$$

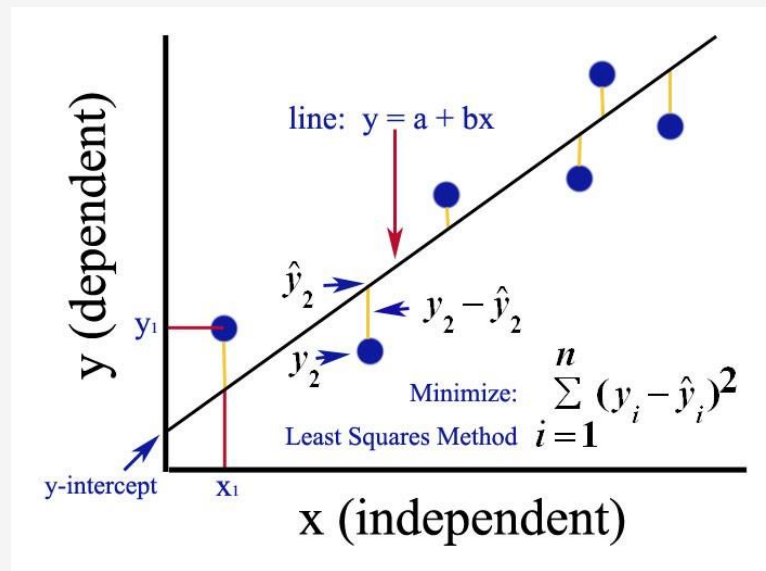
- **Diễn giải:** Khi các  $x_2, \dots, x_n$  không thay đổi, biến  $x_1$  thay đổi 1 đơn vị thì biến  $y$  thay đổi bao nhiêu đơn vị?
- **Tiên lượng:** Với các thông tin của  $x_1, x_2, \dots, x_n$  thì  $y$  là bao nhiêu?

# Phương trình hồi quy

- Tìm đường thẳng hồi quy như thế nào?



# Phương trình hồi quy



- Phương pháp bình phương tối thiểu (Ordinary Least Squares - OLS)
- Tìm đường thẳng hồi quy sao cho tổng bình phương sai số (error) là nhỏ nhất

# Phương trình hồi quy

---

- Hệ số  $\beta$

$$b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} = r \frac{s_y}{s_x}$$

- Điểm chặn  $\alpha$

$$a = \bar{y} - b\bar{x}$$

# Hồi quy tuyến tính

- Thực hiện bằng Stata

```
regress biếnphụthuộc biếndộclập
```

```
regress tlsosinh tuoi thai
```

Source	SS	df	MS	Number of obs	=	641
				F(1, 639)	=	762.25
Model	148354317	1	148354317	Prob > F	=	0.0000
Residual	124365805	639	194625.673	R-squared	=	0.5440
				Adj R-squared	=	0.5433
Total	272720122	640	426125.19	Root MSE	=	441.16
tlsosinh	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
tuoi thai	206.6412	7.484572	27.61	0.000	191.9439	221.3386
_cons	-4865.245	290.0814	-16.77	0.000	-5434.873	-4295.617

- Phương trình

Trọng lượng sơ sinh (gram) =  $-4865 + 206 \times \text{Tuổi thai (tuần)}$

# Hồi quy tuyến tính

---

Trọng lượng sơ sinh (gram) =  $-4865 + 206 \times \text{Tuổi thai (tuần)}$

Ý nghĩa:

- ✓ Diễn giải: Khi tuổi thai tăng lên 1 tuần thì trọng lượng sơ sinh tăng thêm 206 gram
- ✓ Tiên lượng:
  - Một người mang thai 38 tuần  $\rightarrow$  tl sơ sinh dự đoán là  $-4865 + 206 \times 38 = 2963$  gram
  - Một người mang thai 30 tuần  $\rightarrow$  tl sơ sinh dự đoán là?



# Kiểm định giả thuyết

- Kiểm định ý nghĩa thống kê cho hệ số hồi quy:
  - ✓  $H_0: \beta = 0$
  - ✓  $H_a: \beta \neq 0$
- Kết luận có ý nghĩa thống kê dựa vào giá trị p

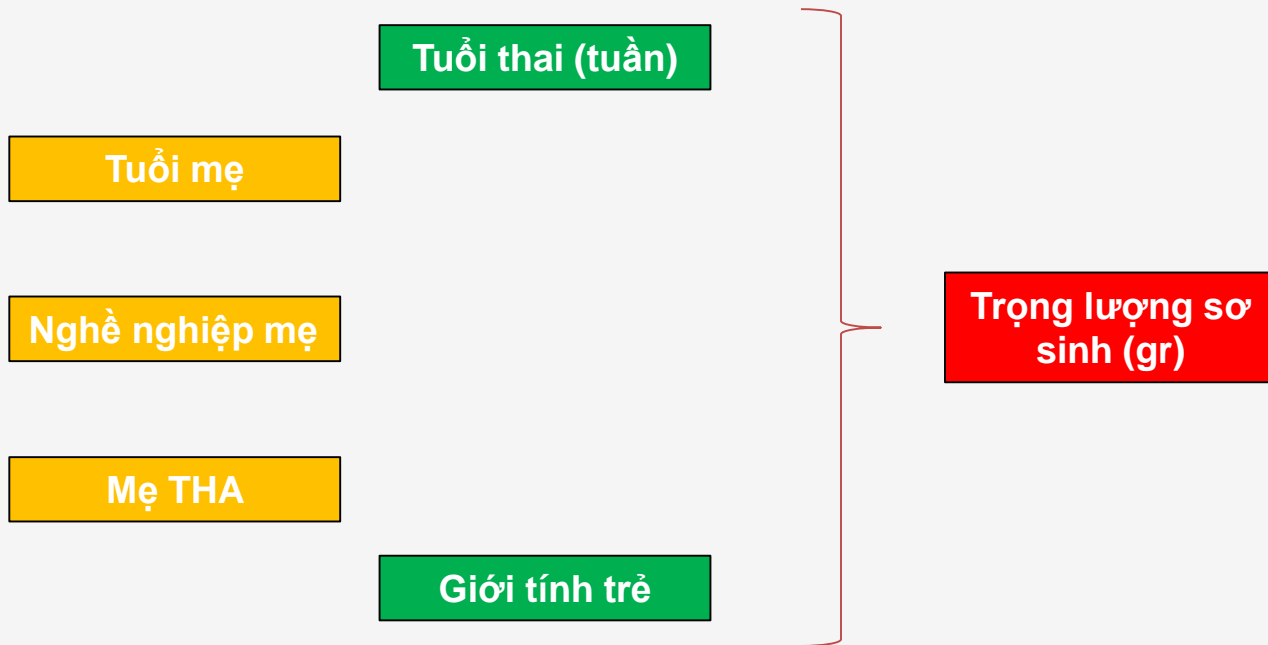
Source	SS	df	MS	Number of obs	=	641
Model	148354317	1	148354317	F(1, 639)	=	762.25
Residual	124365805	639	194625.673	Prob > F	=	0.0000
Total	272720122	640	426125.19	R-squared	=	0.5440
				Adj R-squared	=	0.5433
				Root MSE	=	441.16

tlssosinh	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
tuoithai	206.6412	7.484572	27.61	0.000	191.9439	221.3386
_cons	-4865.245	290.0814	-16.77	0.000	-5434.873	-4295.617

# Hồi quy tuyến tính đa biến

- Nghiên cứu nhằm khảo sát các yếu tố ảnh hưởng tới trọng lượng sơ sinh của trẻ



# Hồi quy tuyến tính đa biến

```
regress tlsosinh tuoime tang_ha tuoi thai gioi i.nghenghiep
```

Chú ý: biến phân nhóm phải thêm **i.** trước tên biến

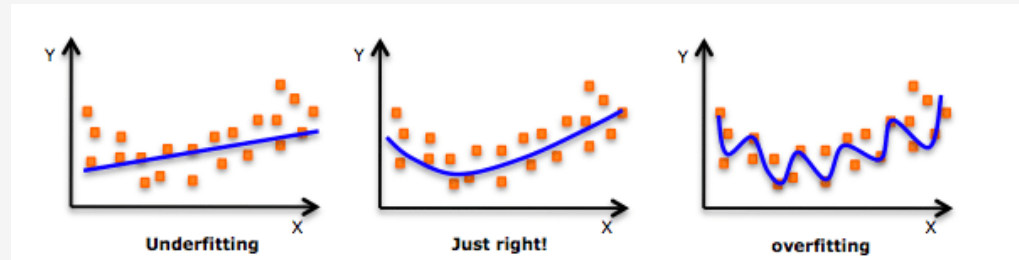
Source	SS	df	MS	Number of obs = 641		
				F(6, 634) = 142.76		
Model	156720174	6	26120028.9	Prob > F = 0.0000		
Residual	115999948	634	182965.218	R-squared = 0.5747		
				Adj R-squared = 0.5706		
Total	272720122	640	426125.19	Root MSE = 427.74		

tl	sosinh	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
	tuoime	1.715159	4.395607	0.39	0.697	-6.916551	10.34687
	tang_ha	-141.4826	50.67361	-2.79	0.005	-240.991	-41.9742
	tuoi thai	201.1081	7.485773	26.87	0.000	186.4082	215.808
	gioi	166.1027	33.94635	4.89	0.000	99.44177	232.7635
nghenghiep							
	2	156.1827	50.31545	3.10	0.002	57.37762	254.9878
	3	185.3339	48.71592	3.80	0.000	89.6698	280.998
	_cons	-4918.722	330.6725	-14.87	0.000	-5568.068	-4269.376

# Hồi quy tuyến tính đa biến

- Chọn biến số đưa vào mô hình:
  - ✓ Có  $2^k - 1$  mô hình khả dĩ (k: số biến số độc lập)
- Mô hình quá nhiều biến  $\rightarrow$  overfitting
  - ✓ ít nhất 10 đối tượng cho mỗi biến trong mô hình (thường là 30-50)<sup>1</sup>
- Mô hình quá ít biến  $\rightarrow$  underfitting
  - ✓ Mô hình kém chính xác, nhiễu



1. Maxwell, S. E. (2000). Sample size and multiple regression analysis. Psychological Methods, 5(4), 434.

# Đánh giá độ phù hợp mô hình (model fit)

$R^2$ : hệ số xác định (Coefficient of determination)

✓ Phương sai giải thích bởi mô hình/tổng phương sai

$R^2$  hiệu chỉnh

$$\bar{R}^2 = 1 - \left(1 - R^2\right) \left[ \frac{n-1}{n-(k+1)} \right]$$

Source	SS	df	MS	Number of obs	=	641
				F(1, 639)	=	762.25
Model	148354317	1	148354317	Prob > F	=	0.0000
Residual	124365805	639	194625.673	R-squared	=	0.5440
				Adj R-squared	=	0.5433
Total	272720122	640	426125.19	Root MSE	=	441.16

tlssinh	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
tuoithai	206.6412	7.484572	27.61	0.000	191.9439	221.3386
_cons	-4865.245	290.0814	-16.77	0.000	-5434.873	-4295.617

# Đánh giá độ phù hợp mô hình (model fit)

---

- $F = (\text{Phương sai giải thích bởi mô hình}) / (\text{Phương sai không thể giải thích bởi mô hình})$
- AIC (Akaike's Information Criterion)  
$$\text{AIC} = 2 \times (\text{Số biến} - \log\text{-likelihood})$$
- BIC (Bayesian Information Criterion)  
$$\text{BIC} = \log(n) \times \text{Số biến} - 2 \times \log\text{-likelihood}$$
- Trong Stata  
**estat ic**

# Các phương pháp lựa chọn mô hình

---

# Giả định của hồi quy tuyến tính

---

## LINE

1. **L**inear: Quan hệ tuyến tính giữa biến độc lập và phụ thuộc
2. **I**ndependence: Các sai số là độc lập
3. **N**ormality: Sai số của ước lượng có phân phối bình thường
4. **E**qual variance: Phương sai đồng nhất (homoscedasticity)

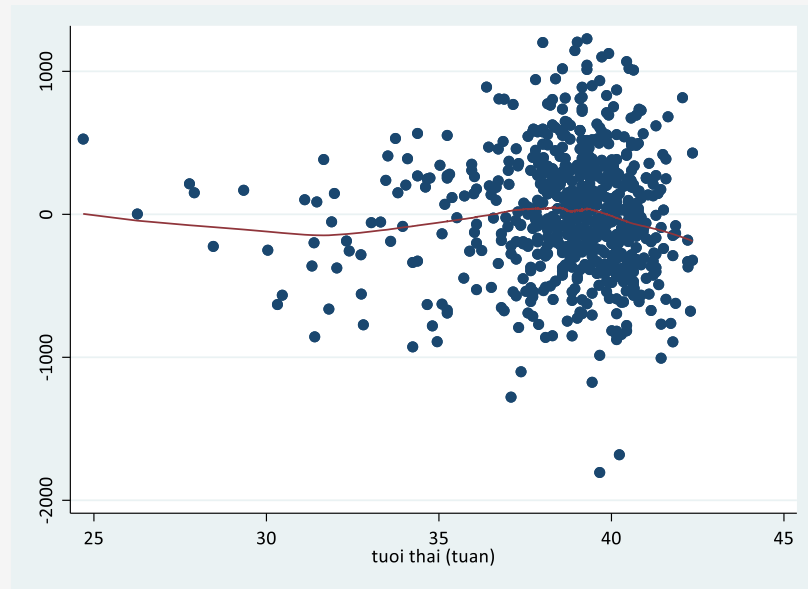


# Giả định của hồi quy tuyến tính

1. Linear: Quan hệ tuyến tính giữa biến độc lập và phụ thuộc
  - Đơn biến: phân tán đồ của biến độc lập và phụ thuộc
  - Đa biến: Phân tán đồ phần dư của mô hình và biến độc lập

`predict res, resid`

`acprplot {biếndộclập}, lowess`



# Giả định của hồi quy tuyến tính

---

2. Independence: Các sai số là độc lập

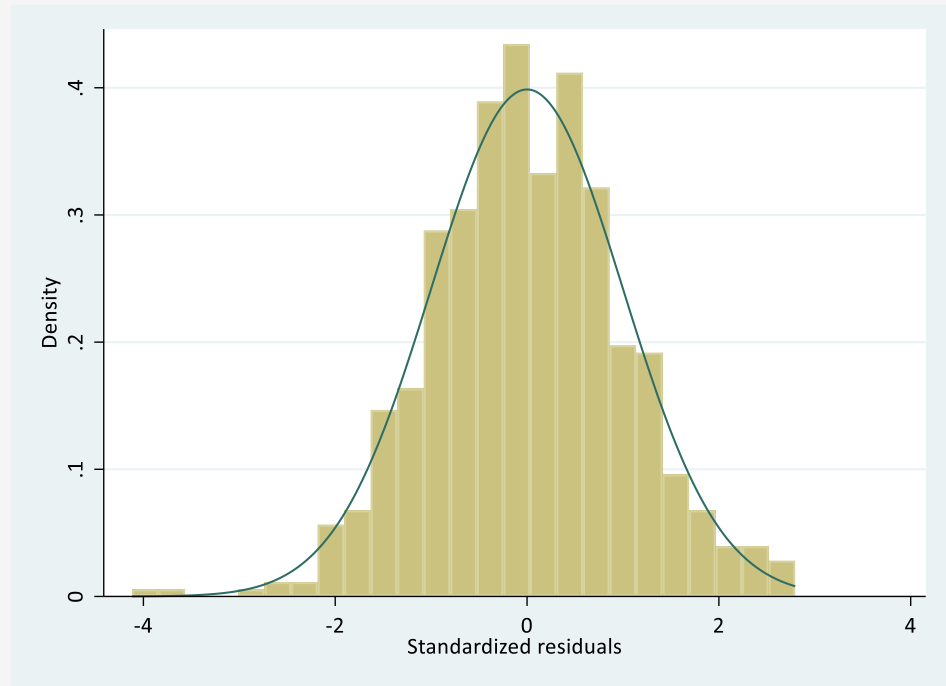
Thiết kế nghiên cứu

# Giả định của hồi quy tuyến tính

3. Normality: Sai số của ước lượng có phân phối bình thường

`predict stdres, rstandard`

`hist stdres, norm`



# Giả định của hồi quy tuyến tính

---

## 4. Equal variance: Phương sai đồng nhất (homoscedasticity)\

- Biểu đồ residual vs. fitted plot

**rvfplot**

- Heteroskedasticity test

**estat hettest**

# Nội dung đã học

---

1. Phân tán đồ
2. Hệ số tương quan
  - ✓ Ý nghĩa hệ số tương quan
  - ✓ Kiểm định ý nghĩa thống kê
  - ✓ Lựa chọn hệ số tương quan
3. Hồi quy tuyến tính
  - ✓ Ứng dụng, ý nghĩa của các tham số hồi quy (đơn biến, đa biến)
  - ✓ Đánh giá độ phù hợp mô hình
  - ✓ Giả định của hồi quy tuyến tính

# Bài tập

---

- Dữ liệu “FEV.dta”
- **age**: tuổi
- **fev**: thể tích khí thở ra gắng sức trong 1 giây đầu tiên
- **height**: chiều cao (inch)
- **sex**: 0 = Nữ; 1 = Nam
- **smoker**: đang hút thuốc
- Xác định các yếu tố ảnh hưởng tới FEV1