Bài giảng 18:
Phân tích tương quan

Nguyễn Văn Tuấn

Garvan Institute of Medical Research, Australia

Đại học Tôn Đức Thắng, Việt Nam

Sex	Height	LeftArm	RtArm	LeftFoot	RtFoot	LeftHand	RtHand	HeadCirc	nose
Male	69	25.5	25.5	27	26.5	9.5	9	58.5	5.5
Male	79	28	25	29	27.5	9	9	54	5
Male	75	27	27.5	31	32	3.75	3.75	62.5	5
Male	69	25	25.5	25.5	25.5	10	8	58.5	5.5
Male	65	25	25	23.5	23	9.5	9.4	57	4.4
Male	79	30.5	30.5	28	28	8.5	8.5	58.5	4.5
Male	72	26.5	26.5	28.5	28.5	9.5	9.5	60.5	4.2
Male	69.5	26.5	27	27	27	10	10.5	58.5	5
Male	73	28	28.4	30.6	31.4	8.5	8.9	57.4	6.4
Male	71.5	28.6	28.1	27.4	28.5	9.8	9.4	56.5	4.1
Male	69.5	27	27	27	27	9	9	58.5	5
Male	73	23	23	28.5	27.5	8.5	9	59.5	5
Male									6
Male									5
Male									6
Male	71	23.5	23	27	27.5	9.5	10	57	5.5
Male	72	28.5	28	26.5	27.5	11.5	11.5	58	6.5
Male	66	24	25	25.5	26	8	8.5	57	5
Male	71	26	26	29	28	9	9	55	5
Male	67	25.5	25.5	27.2	27	7.7	7.7	54	5.5
Male	71	29.5	28.5	29	28.5	9	9.5	57.5	5.5
Male	72	25.5	25	28	28	9.5	9.5	56	5.3
Male	72	25	26	28.5	29	8.5	9	61	4.5
Male	73.5	26	28.5	29	30	9.5	9.9	62	5
Male	73	30	29	23.5	24	9	9.5	58	6
Female	62	22.5	23	21.5	21	8.3	7.3	56.5	5.1
Female	66	24	25	24	25	8.5	8.5	53	6
Female	64.5	22.5	21.5	23.3	23	7.5	7.5	56	4

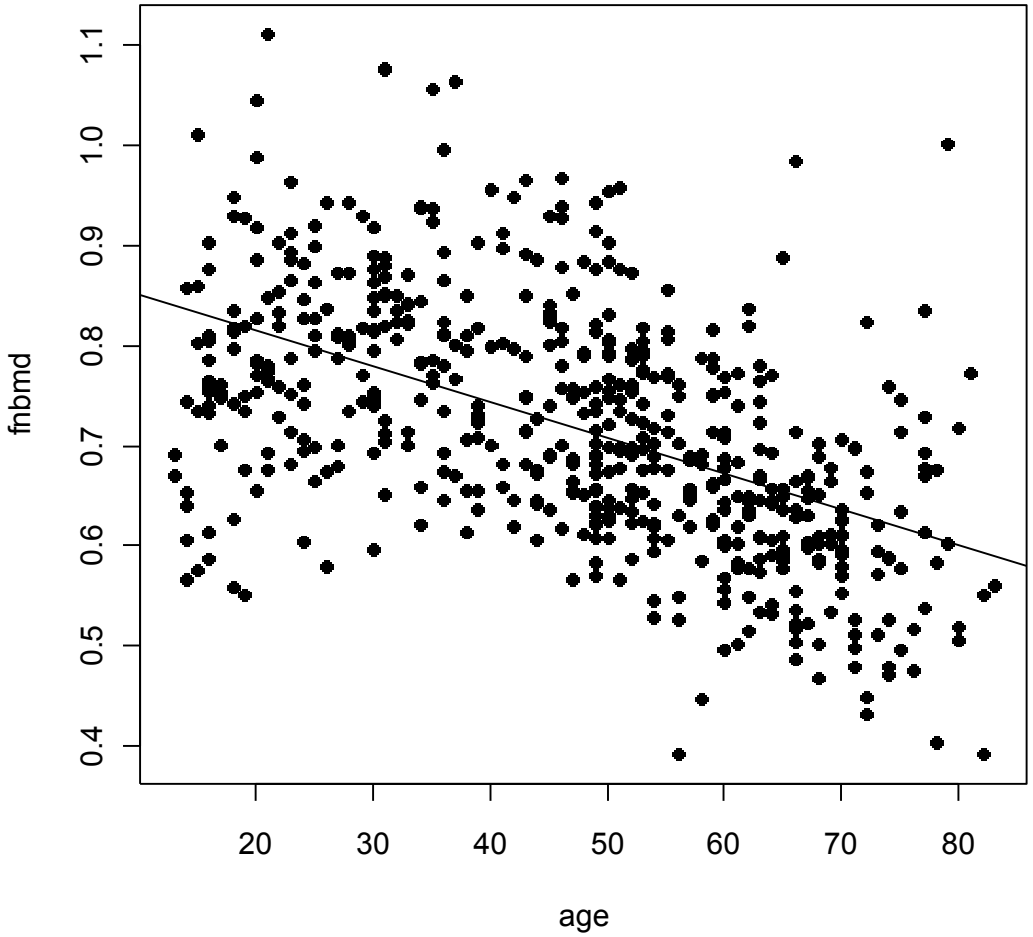
Có mối tương quan giữa các biến ?

Nội dung

- Vài ví dụ
- Lí thuyết
- Sử dụng R
- Tóm lược

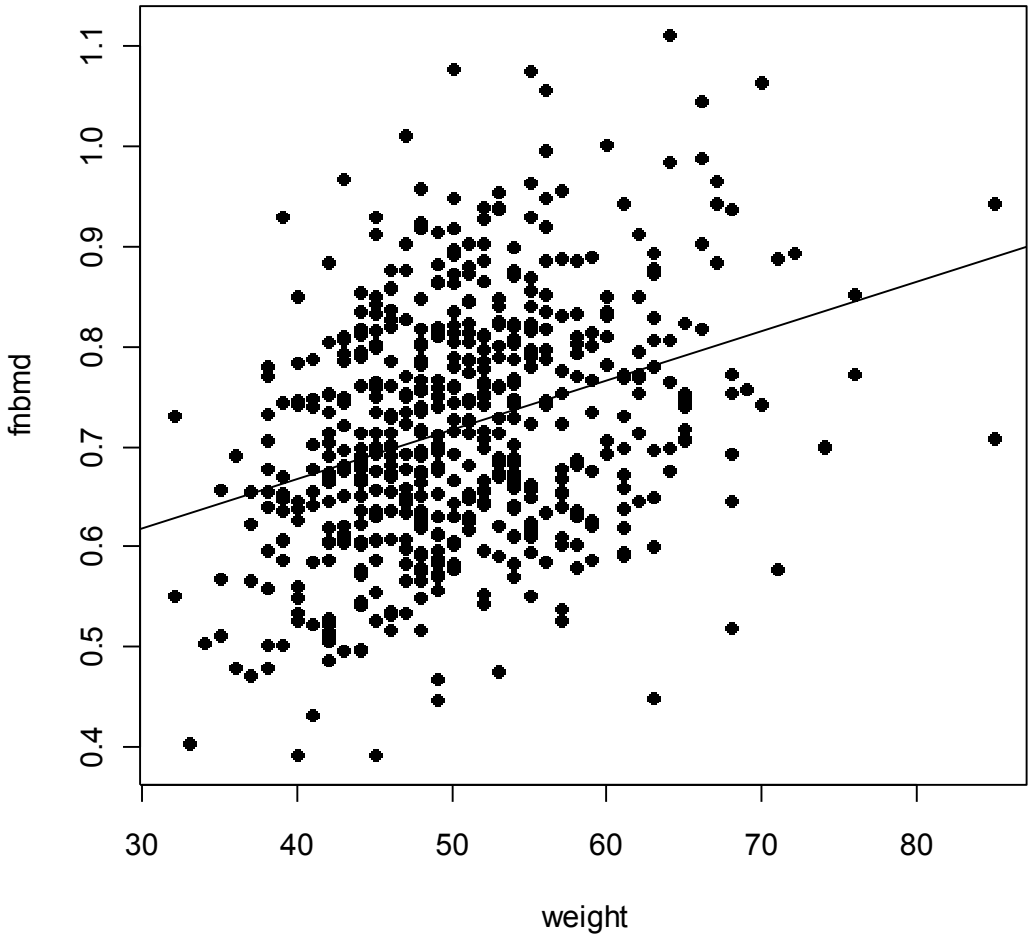
Mối tương quan giữa mật độ xương và tuổi

```
plot(fnbmd ~ age, pch=16)
abline(lm(fnbmd ~ age))
```



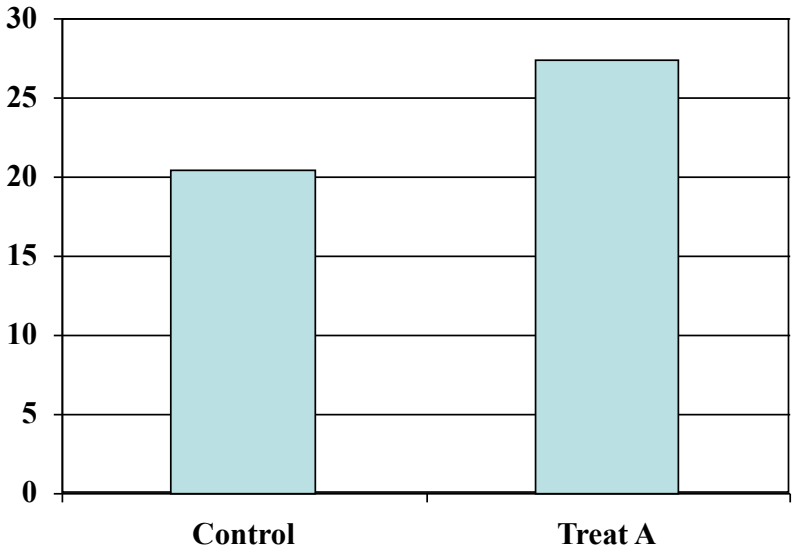
Weight and femoral neck bone density

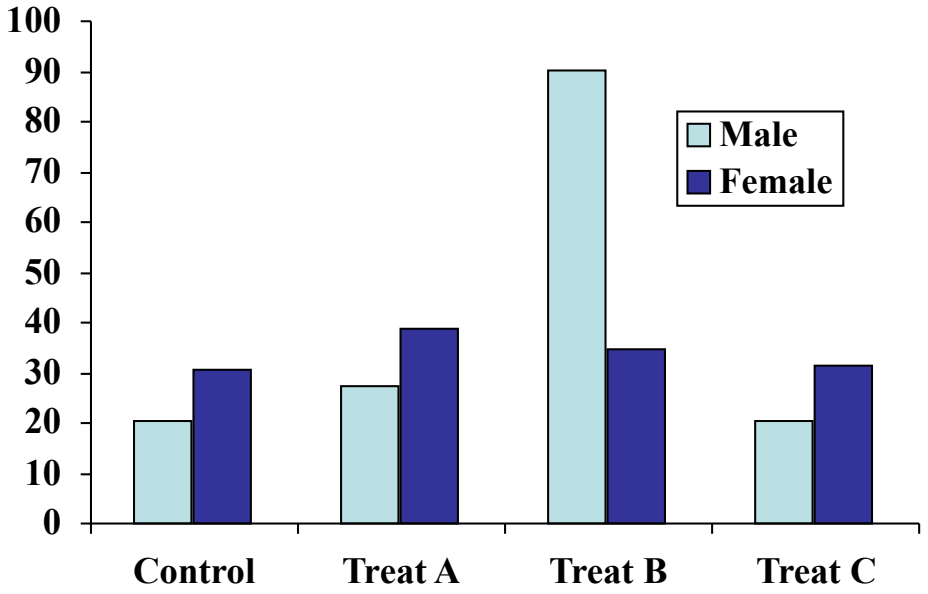
```
plot(fnbmd ~ weight, pch=16)
abline(lm(fnbmd ~ weight))
```



Khi nào cần phân tích tương quan

t-test hay ANOVA



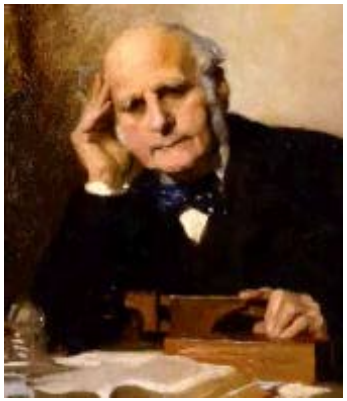


Khi biến độc lập (independent variable) là biến phân nhóm (*categorical variable)*

Khi nào cần phân tích tương quan

- *Tương quan* giữa 2 biến liên tục
- Mức độ "co-variation"
- Tiên lượng?

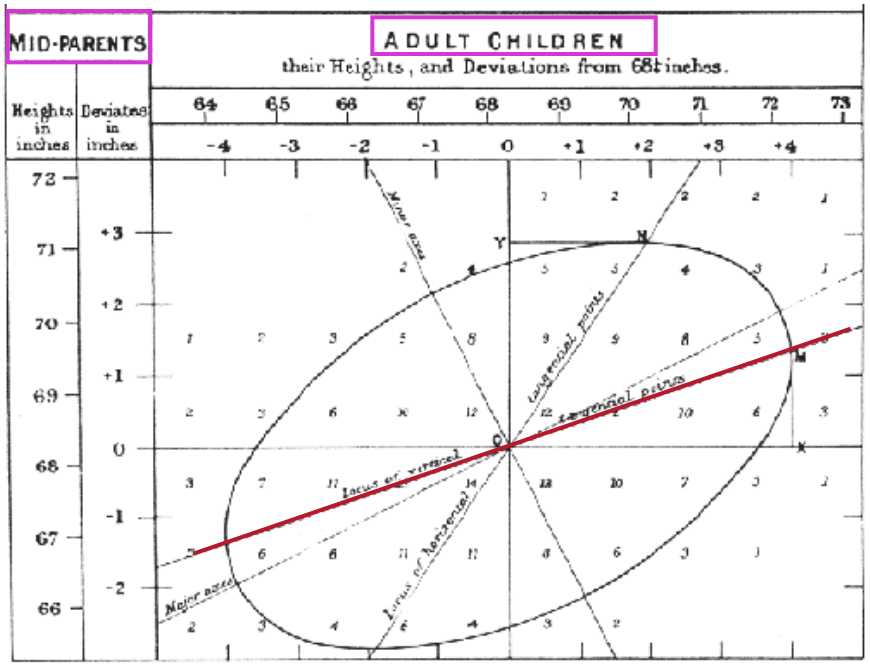
Sir Francis Galton (16/2/1822 – 17/1/1911)



Research interest:

"Those qualifications of intellect and disposition which … lead to reputation"

Didn't have data on "intelligence" so instead studied HEIGHT



Galton's conclusions:

- **Nature dominates:** "families of reputation were much more likely than ordinary families to produce offspring of ability"
- **Recommended "judicious marriages during several generations" to "produce a highly gifted race of men"**
- His "genetic utopia": **"Bright, healthy individuals were treated and paid well, and encouraged to have plenty of children. Social undesirables were treated with reasonable kindness so long as they worked hard and stayed celibate."**

- Although a self-proclaimed genius, who wrote that he could read @2½, write/do arithmetic @4, and was comfortable with Latin texts @8, **he couldn't figure out how to model these data(!)**
- He went to JD Dickson, a mathematician at Cambridge, who formalized the relationship by developing what we now know as **linear regression**

Một chút lí thuyết

Làm thế nào để mô tả mối tương quan tuyến tính?

- Gọi X và Y là 2 biến ngẫu nhiên từ n quan sát
- Đo lường độ biến thiên: **phương sai (variance)**

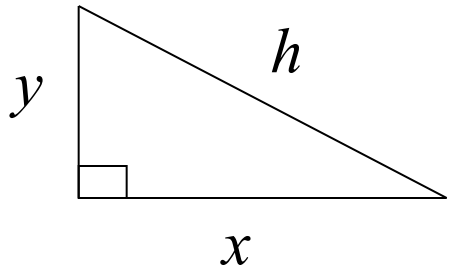
$$\mathrm{var}(x) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \qquad \mathrm{var}(y) = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}$$

- **Chúng ta cần một thước đo độ "hiệp biến" giữa X và Y**
- **Covariance là trung bình của tích số X và Y**

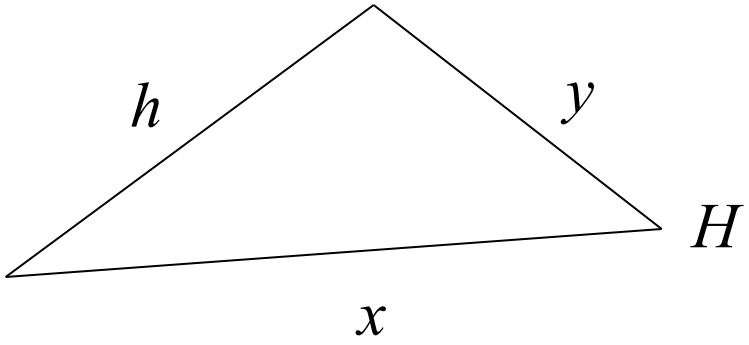
$$\mathrm{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Variance và covariance: hình học

Tính độc lập của 2 biến *x* và *y* có thể thể hiện qua hình học:



$$h^2 = x^2 + y^2$$



$$h^2 = x^2 + y^2 - 2xy\cos(H)$$

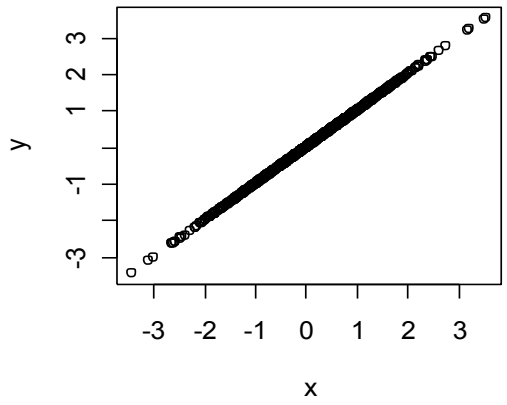
Covariance

Ý nghĩa của phương sai và hiệp biến

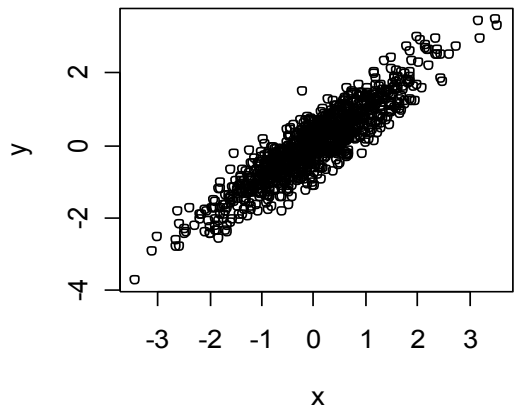
- Phương sai lúc nào cũng là số dương
- Hiệp biến có thể dương hay âm (vì là bình phương của tích số X và Y)
  - Nếu covariance = 0, *X và Y độc lập*
  - Nếu covariance > 0, X và Y biến thiên cùng chiều
  - Nếu covariance < 0, X và Y biến thiên nghịch chiều
- **Covariance = thước đo và độ liên quan**

Vài hệ số tương quan

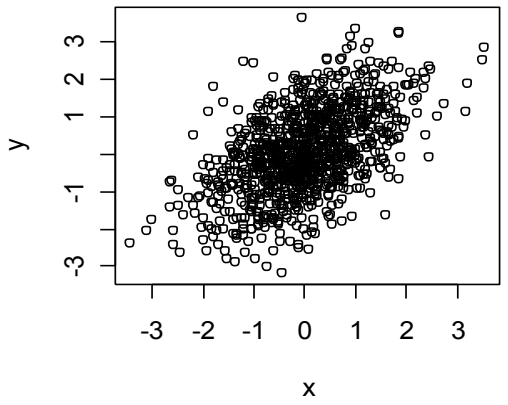
r = 0.99



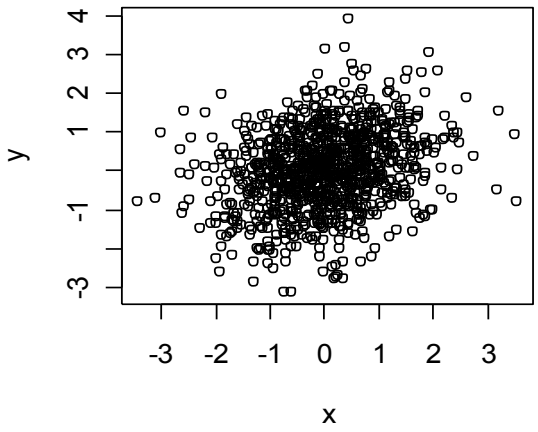
r = 0.90



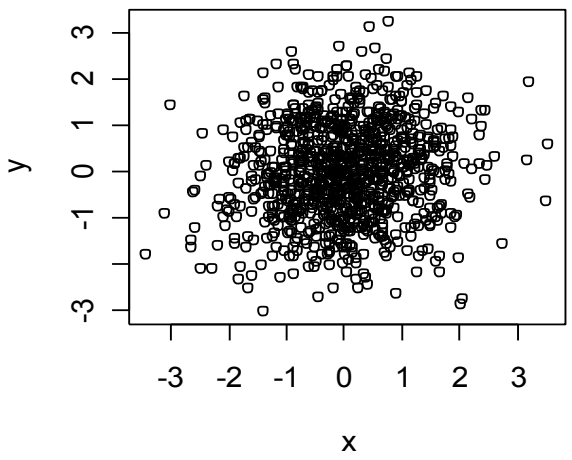
r = 0.50



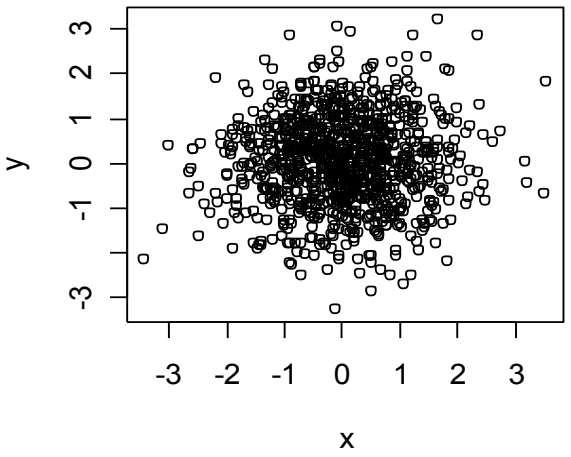
r = 0.25



r = 0.10



r = 0.01



Mục tiêu của phân tích tương quan

- Ước tính hệ số tương quan - *coefficient of correlation (r)*
- Kiểm định xem giả thuyết $r = 0$?

Ước tính hệ số tương quan

- Covariance có đơn vị đo lường (X * Y).
- **Coefficient of correlation** (*r*) giữa *X* và *Y* là một **standardized covariance – không có đơn vị đo lường**
- *r* định nghĩa như sau:

$$r = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x) \times \mathrm{var}(y)}} = \frac{\mathrm{cov}(x, y)}{SD_x \times SD_y}$$

Kiểm định giả thuyết r = 0

- Giả thuyết vô hiệu: H_o: *r* = 0 và H_A: *r* không bằng 0.
- Fisher's z-transformation: hoán chuyển *r* → z

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

- **Tính phương sai của z**

$$SE(z) = \frac{1}{\sqrt{n-3}}$$

- **T-test:**

$$t = \frac{z}{SE(z)}$$

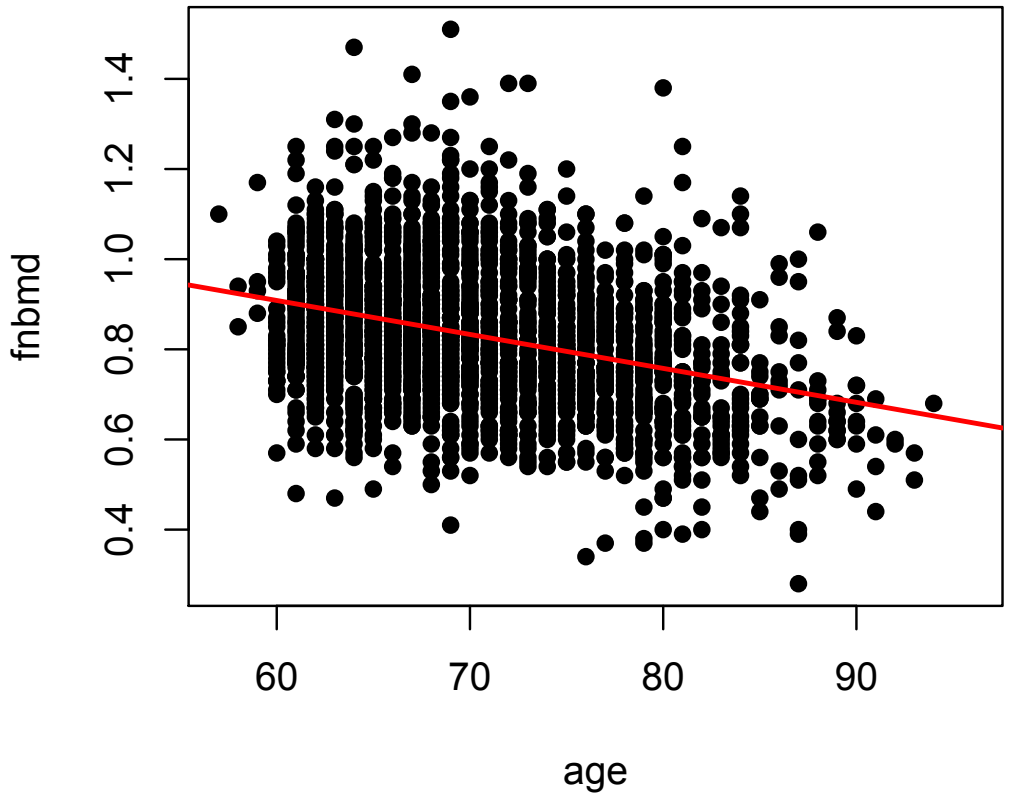
Ví dụ: mật độ xương và độ tuổi

- Nghiên cứu cắt ngang
- Bone mineral density (BMD) đo ở cổ xương đùi (femoral neck)
- Tuổi (age), trọng lượng (weight)
- Câu hỏi: có mối tương quan giữa BMD và
  - tuổi và trọng lượng
  - Mối tương quan đó có ý nghĩa thống kê?

Dùng R
(cor, cor.test)

Mật độ xương và tuổi

```
dat=read.csv("http://statistics.vn/data/does_vn07.csv",
   header=T)
women=subset(dat,gender=="Female")
plot(fnbmd ~ age, pch=16)
abline(lm(fnbmd ~ age), col="red", lwd=2)
```



Hàm cor

- Nếu không có missing values

```
cor(x, y)
```

- Nếu có missing values

```
cor(x, y,
use="pairwise.complete.obs")
```

```
cor(x, y, use="complete.obs")
```

Ví dụ R

```
dat=read.csv("http://statistics.vn/data/
  does_vn07.csv", header=T)

women=subset(dat,gender=="Female")

plot(fnbmd ~ age, pch=16)

abline(lm(fnbmd ~ age), col="red", lwd=2)

cor(age, fnbmd)

cor(age, fnbmd, use="complete.obs")
```

Ví dụ R

```
> cor(age, wt,
use="complete.obs")

[1] -0.236508
```

Hàm cor.test

```
> cor.test(fnbmd, age)

                Pearson's product-moment correlation

data:  fnbmd and age
t = -14.4162, df = 556, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal
  to 0
95 percent confidence interval:
 -0.5795310 -0.4584638
sample estimates:
       cor
-0.5216183
```

r and R^2

- r là hệ số tương quan
- R^2 là hệ số xác định (coefficient of determination)

  phản ảnh phần trăm phương sai của y có thể giải thích bởi biến x

- r(weight, BMD) = 0.33 có nghĩa là $R^2 = (0.33)^2 = 0.11$.

  11% độ khác biệt về BMD có thể giải thích bằng những khác biệt về cân nặng

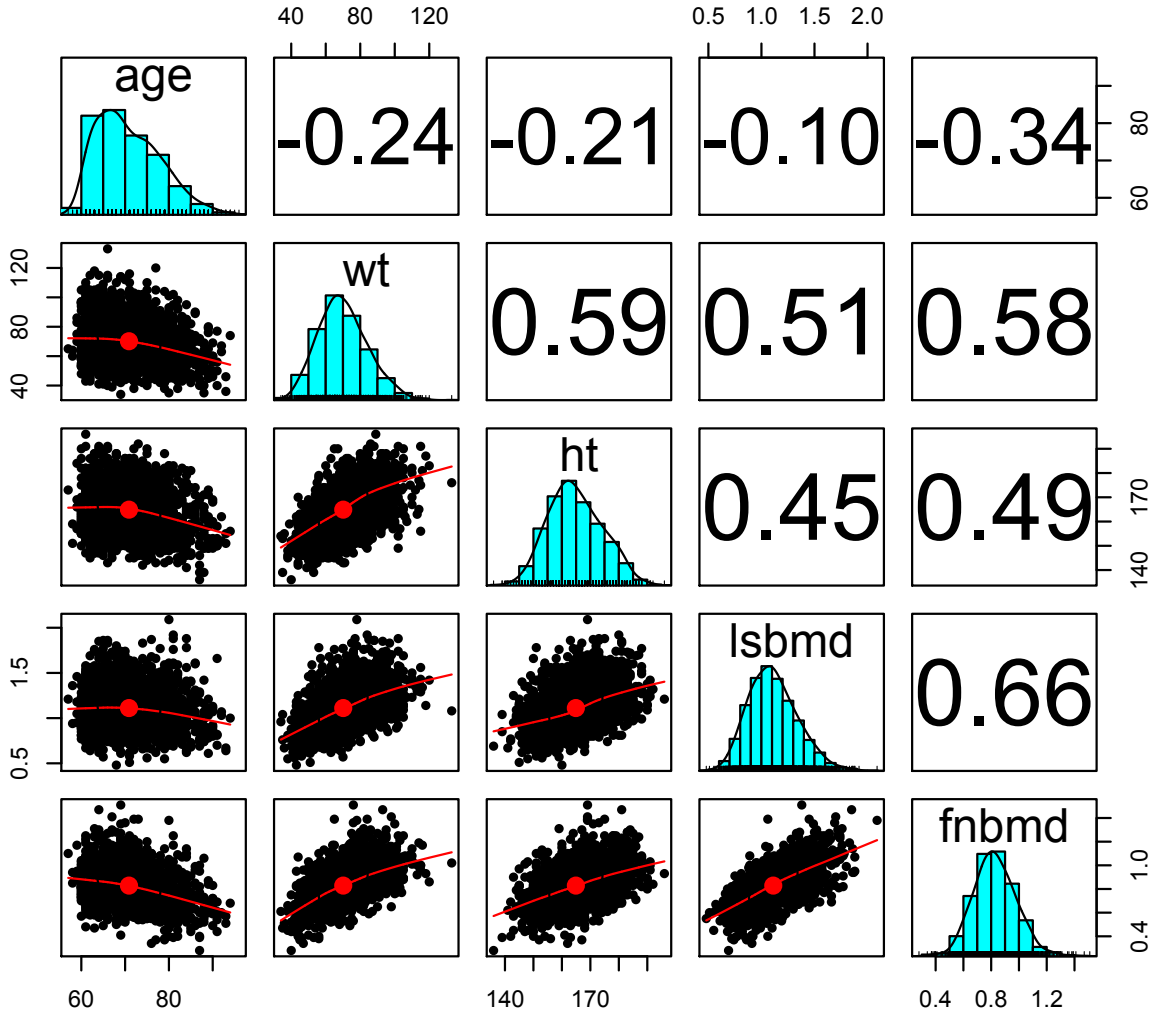
Tương quan đa biến

Hàm cor.test (psych)

- Có thể tính hệ số tương quan của nhiều biến

```
library(psych)
vars = cbind(age, wt, ht, lsbmd, fnbmd)
cor.test(vars)
pairs.panels(vars)
```

```
library(psych)
vars= cbind(age, wt, ht, lsbmd, fnbmd)
pairs.panels(vars)
```



```
> corr.test(vars)

Call:corr.test(x = vars)
Correlation matrix

          age    wt    ht lsbmd fnbmd
age     1.00 -0.24 -0.21 -0.11 -0.34
wt     -0.24  1.00  0.58  0.51  0.58
ht     -0.21  0.58  1.00  0.45  0.49
lsbmd -0.11  0.51  0.45  1.00  0.66
fnbmd -0.34  0.58  0.49  0.66  1.00
```

Tương quan từng phần (partial correlation)

Ý nghĩa của partial correlation

- Trường hợp 3 biến: Y, X1, X2
- Tương quan giữa Y và X1, nếu X2 bất biến

Ví dụ partial correlation

- Ví dụ:

$$r_{Y,1} = 0.7 \rightarrow r_{Y,1}^2 = 0.49$$

$$r_{Y,2} = 0.6 \rightarrow r_{Y,2}^2 = 0.36$$

$$R_{y.12} = 0.8 \rightarrow R_{Y.12}^2 = 0.64$$

$$0.49 + 0.36 = 0.85 \neq 0.64$$

- Tại sao? Vì X1 và X2 có liên quan với nhau

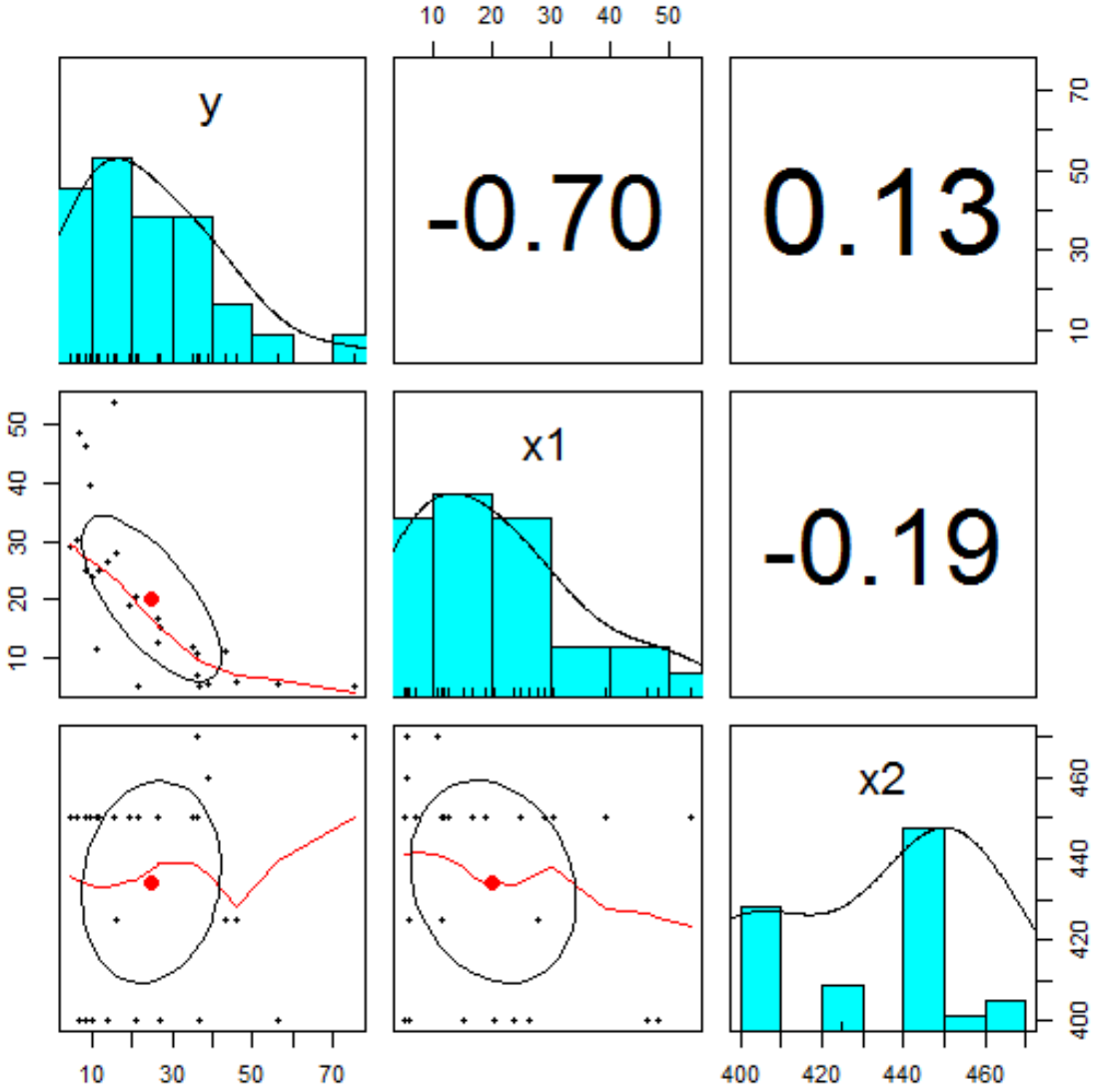
Package ppcor

- ppcor có thể ước tính partial correlation
- Công thứ chung

```
pcor.test(x, y, z, method=c("pearson",
          "kendall", "spearman")
```

Ví dụ dữ liệu thực tế

```
y = c(36.98,13.74,10.08, 8.53,36.42,26.59,19.07, 5.96, 15.52,
56.61, 26.72, 20.80, 6.99, 45.93, 43.09, 15.79, 21.60, 35.19,
26.14, 8.60, 11.63, 9.59, 4.42, 38.89, 11.19, 75.62, 36.03)

x1 = c(5.1,26.4,23.8,46.4,
7.0,12.6,18.9,30.2,53.8,5.6,15.1,20.3,48.4,
5.8,11.2,27.9,5.1,11.7,16.7,24.8,24.9,39.5,29.0, 5.5, 11.5,
5.2,10.6)

x2 = c(400,400, 400, 400, 450, 450, 450, 450, 450, 400, 400,
400,  400, 425, 425, 425, 450, 450, 450, 450, 450, 450, 450,
460, 450, 470, 470)

dd=cbind(y,x1,x2)

library(psych); pairs.panels(dd)

library(ppcor)

pcor.test(y, x1, x2, method=c("pearson", "kendall",
"spearman"))
```


y
x1
x2
-0.70
0.13
-0.19
10
20
30
40
50
70
400
420
440
460

Partial correlation

```
> pcor.test(y, x1, x2, method=c("pearson", "kendall",
"spearman"))
```

```
estimate      p.value statistic     n gp  Method
-0.698412   1.747116e-06  -4.78067 27  1 pearson
```

```
> pcor.test(y, x1, x2, method=c("pearson", "kendall",
"spearman"))
```

```
 estimate   p.value   statistic    n gp  Method
-0.004112931 0.9839242 -0.02014934 27  1 pearson
```

Tóm lược

- Hệ số tương quan – đo lường mức độ tương quan giữa hai biến liên tục
- Hàm R
  - `cor(x, y, use="complete.obs")`
  - `cor.test(x, y)`

Tương quan đa biến: trong **psych**

  - `pairs.panels(vars)`
  - `corr.test(vars)`