

# Bài giảng 5: Lập chương trình phân tích

**Nguyễn Văn Tuấn**

Garvan Institute of Medical Research, Australia  
Đại học Tôn Đức Thắng, Việt Nam

# Bối cảnh

- Bạn đã thực hiện xong nghiên cứu
- Dữ liệu đã được nhập vào máy tính (Excel hay Epi info)
- **Phải lên kế hoạch phân tích trước khi làm bất cứ phân tích nào!**

# Lập kế hoạch phân tích

- Kiểm tra sự nhất quán của dữ liệu (consistency)
- Mục tiêu phân tích
- Phác hoạ bảng biểu
- Xác định package R cần thiết
- Tiến hành phân tích và ghi chép

**Kiểm tra tính nhất quán**

# Tính nguyên bản (integrity) RẤT quan trọng

- Độ chính xác và nguyên bản của dữ liệu" rất quan trọng
- Thời gian
  - Chuẩn bị dữ liệu: 90%
  - phân tích: 10%

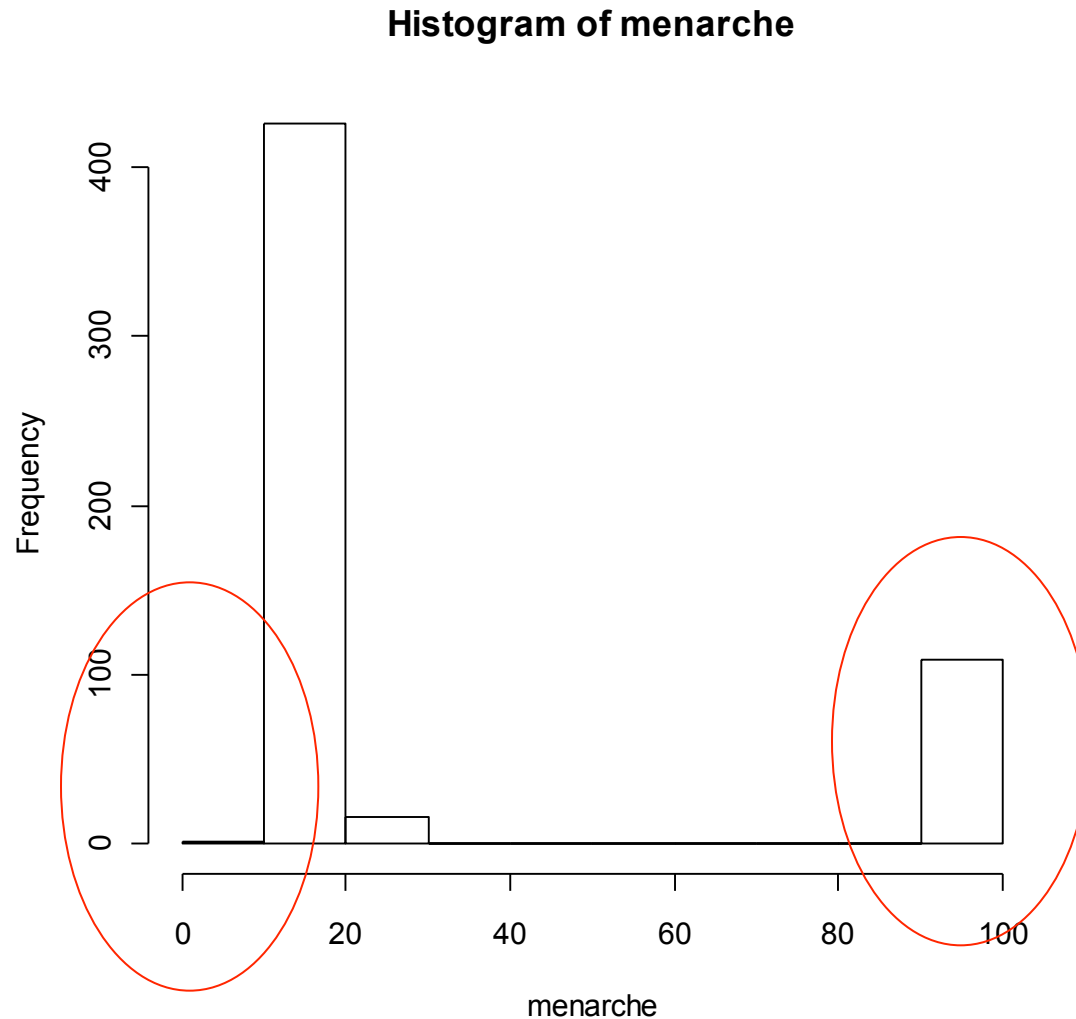
# Những bước cần thiết cho phân tích

- **Tạo ra data file**
  - Dữ liệu gốc
  - Ghi chép cẩn thận
- **"Làm sạch" dữ liệu**
- **Xử lý dữ liệu**
- **Tạo ra một data file cho phân tích**
- **Ghi chú những phân tích quan trọng**

# Kiểm tra phân bố

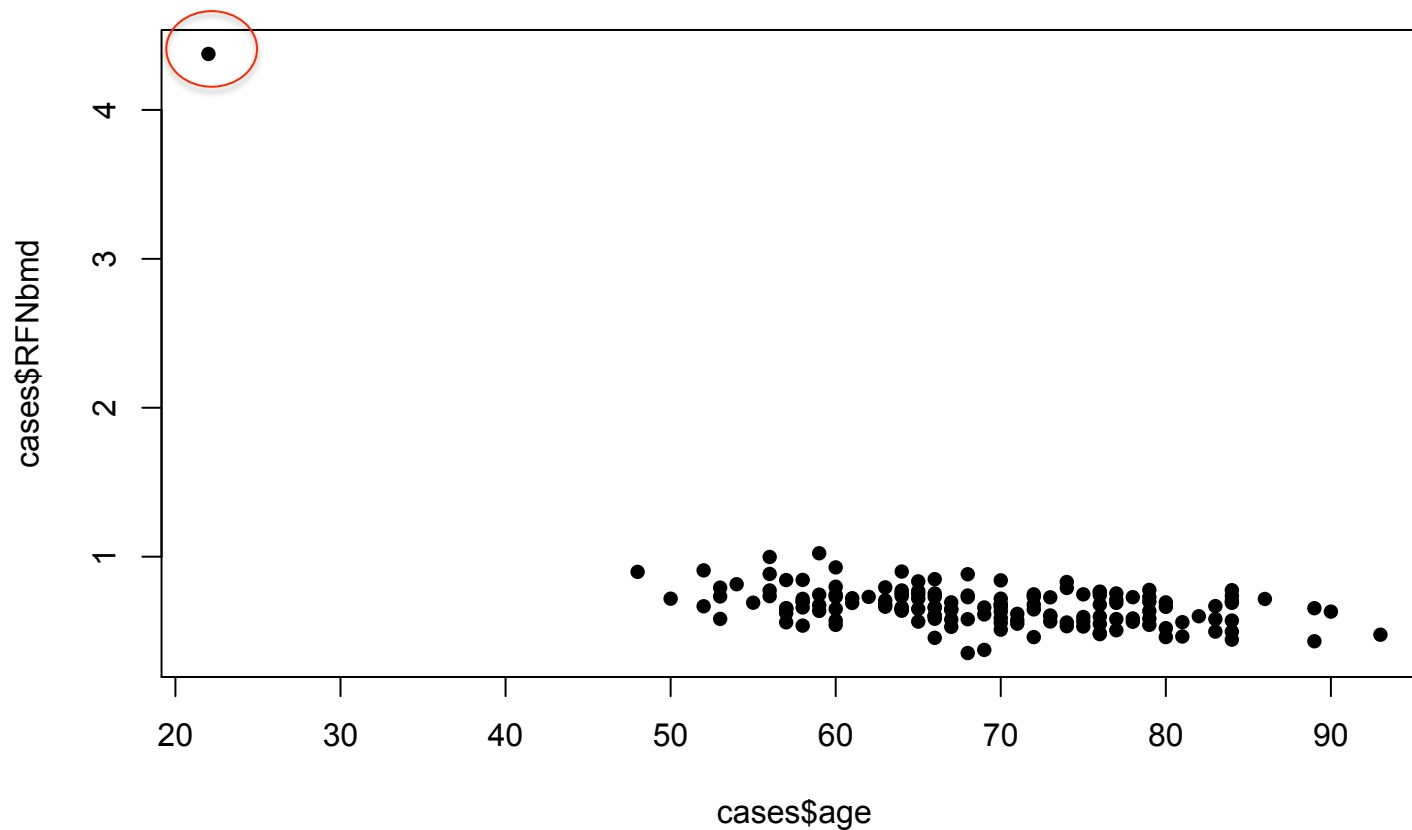
- Dùng histogram để kiểm tra phân bố
  - Dữ liệu (biến số) tuân theo luật phân bố chuẩn?
- Dùng scatterplot để kiểm tra những dữ liệu ngoại biên hay bất thường

# Kiểm tra số liệu ngoại biên (outliers)





# Dùng biểu đồ tương quan



**Xác định mục tiêu phân tích**

# Những phục tiêu phổ biến

- Xác định mục tiêu: bạn muốn làm gì?
- Mô tả một biến số
- So sánh giữa các nhóm
- Tìm mối liên quan giữa các biến
- Mô hình tiên lượng

# Phân tích mô tả một biến số

- Biến liên tục
  - Phân bố (**hist**, **boxplot**, **plot**)
  - Trung bình, độ lệch chuẩn, trung vị, v.v. (**psych package: describe, describe.by**)
- Biến phân nhóm (categorical variable)
  - Tỷ lệ (**table**, **gmodels: CrossTable**)
  - Đếm
  - Khoảng tin cậy 95%

# So sánh giữa các nhóm

- Biến liên tục
  - T-test, ANOVA (`t.test`, `anova`, `lm`)
- Biến phân nhóm
  - Z-test, Chi-squared analysis (`Cross.Table`, `chisq.test`, `fisher.test`)
  - Binomial test (`prop.test`)

# Tìm mối tương quan

- Biến liên tục
  - Hồi qui tuyến tính đơn giản (**lm**)
  - Hồi qui tuyến tính đa biến (**lm**)
- Biến phân nhóm
  - Hồi qui logistic (**glm**)
  - Hồi qui nhị phân (**glm**)
  - Hồi qui Cox's (**survfit**, **survdiff**, **coxph**)

**Ví dụ**

# Nghiên cứu về vitamin D

- Nghiên cứu cắt ngang (cross-sectional study)
- N = 558 đối tượng, 222 nam + 336 nữ, tuổi 13–83
- Mục tiêu:
  - Tỷ lệ thiếu vitamin D
  - Yếu tố nguy cơ thiếu vitamin D
  - Yếu tố có liên quan đến vitamin D



# Các biến liên quan

## Số liệu về nhân trắc

- id
- sex
- dob
- address
- region
- age
- menarche
- height
- weight
- bmi
- fracture
- alcohol, coffee, tea
- season

## Số liệu về sinh hoá và xương

- estradiol
- testo
- vitd
- pth
- xlap
- fnbmd
- hipbmd
- lsbmd

# Dữ liệu trong Excel (.csv format)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	id	sex	address	region	age	menarche	estradiol	testo	vitd	xlap	pth	height	weight	bmi	fnbmd	hipbmd	lsbmd	fracture
2	DD 002	2	DONG DA	1	30	17.00	278.30	0.519	24.43	0.293	33.73	146	38	18	0.598	0.645	0.754	2
3	DD 018	2	DONG DA	1	29	16.00	33.41	0.097	28.38		35.67	157	50	20	0.744	0.789	1.039	1
4	DD 020	1	DONG DA	1	31	16.00	34.19	9.94	23.06	0.53	19.62	160	55	21	1.075	1.04	1.016	2
5	DD 032	1	DONG DA	1	49	99.00	38.97	9.4	28.75	0.202	18.41	175	70	23	0.742	0.925	1.079	2
6	DD 035	2	DONG DA	1	60	99.00	304.89	0.033	19.41		1.20	164	63	23	0.6	0.641	0.638	2
7	DD 039	2	DONG DA	1	75	15.00	30.25	0.264	9.77		4.02	150	44	20	0.496	0.523	0.77	2
8	DD 042	2	DONG DA	1	63	15.00	20.75	0.243	39.20		24.94	155	54	22	0.669	0.737	0.88	2
9	DD 045	1	DONG DA	1	69	99.00	33.61	8.17	35.40	0.314	21.58	157	57	23	0.603	0.767	0.851	2
10	DD 046	1	DONG DA	1	71	99.00	53.73	6.5	29.10	0.41	38.41	170	64	22	0.7	0.842	0.836	1
11	DD 047	1	DONG DA	1	65	99.00	49.51	7.13	25.21	0.277	39.96	160	55	21	0.657	0.776	0.882	2
12	DD 048	1	DONG DA	1	80	99.00	45.25	3.71	34.31	0.178	57.13	159	65	26	0.718	0.929	1.214	2
13	DD 049	1	DONG DA	1	65	16.00	36.05	4.84	16.73	0.794	32.46	169	71	25	0.888	0.979	1.113	2
14	DD 050	1	DONG DA	1	65	99.00	52.59	13.85	23.43	0.747	41.30	169	71	25	0.578	0.725	0.749	2
15	DD 051	1	DONG DA	1	70	99.00	39.90	12.41	4.00	0.721	32.56	165	61	22	0.592	0.714	0.818	2
16	DD 052	1	DONG DA	1	64	16.00	42.81	10.23	23.59	0.595	35.85	163	60	23	0.694	0.853	0.97	2
17	DD 062	2	DONG DA	1	30	99.00	75.19	0.389	18.40		22.31	155	52	22	0.742	0.807	0.979	2
18	DD 080	1	DONG DA	1	35	99.00	38.97	9.4	28.75	0.202	23.62	173	56	19	1.057	0.965	1.117	2
19	DD 086	1	DONG DA	1	34	99.00	35.40	6.6	25.11	0.768	28.83	168	59	21	0.621	0.763	0.888	2
20	DD 091	1	DONG DA	1	79	17.00	38.97	7.15	23.83	0.533	50.40	160	60	23	1.002	1.125	1.478	2
21	DD 093	1	DONG DA	1	63	99.00	30.96	8.09	23.84	0.572	31.97	165	45	17	0.587	0.679	0.761	2
22	DD 094	1	DONG DA	1	77	99.00	31.90	7.91	25.34	0.674	24.34	169	55	19	0.614	0.673	0.648	2

# Kế hoạch phân tích

- Kiểm tra *data consistency*
- Kiểm tra *distribution* của vitamin D và hormones
- Thống kê mô tả

# Kế hoạch phân tích

- Xác định tiêu chuẩn "vitamin D deficiency" – gọi là *def*
- Tìm tần số của *def* theo:
  - giới tính, độ tuổi, nông thôn và thành phố, mùa
- Tìm yếu tố nguy cơ cho *def*
- Tìm mối tương quan giữa vitamin D và PTH. Có một ngưỡng PTH nào mà vitamin D ổn định

# Phương pháp thống kê

Checking <i>data consistency</i>	Graphical analysis
Checking the <i>distribution</i> of vitamin D and hormones	Histogram
<i>Descriptive statistics</i> of all variables for men and women separately	T-test for continuous variables and z test for categorical variables
Prevalence of vit D deficiency for men and women, and various factors	Counting, Chi square test
Find risk factors for <i>def</i>	Logistic regression
Find the correlation between vitamin D and PTH. Is there a threshold whereby vitamin D is stable	Linear regression

# Chúng ta muốn có ...

Table 1. Characteristics of participants

Variable	Men	Women	P-value
Age			
Height			
Weight			
BMI			
25(OH)D			
PTH			
Etc.			

# Chúng ta muốn có ...

Table 2. Prevalence of vitamin D deficiency

Risk factor	N	Prevalence	95% CI
Sex Men Women			
Age <20 20-29 30-39 40-49 50-59 60+			
Season Winter Non-winter			
Residence Rural Urban			

# Chúng ta muốn có ...

Table 3. Risk factors for vitamin D deficiency

Risk factor	Odds ratio	95% CI	P-value
Age			
Sex			
BMI			
Season			
Residence			
Etc.			

**Which determinants are statistically significant ?**

**How much variance of vit D deficiency can be explained by these risk factors?**

**Which one is important?**



# Chúng ta muốn có ...

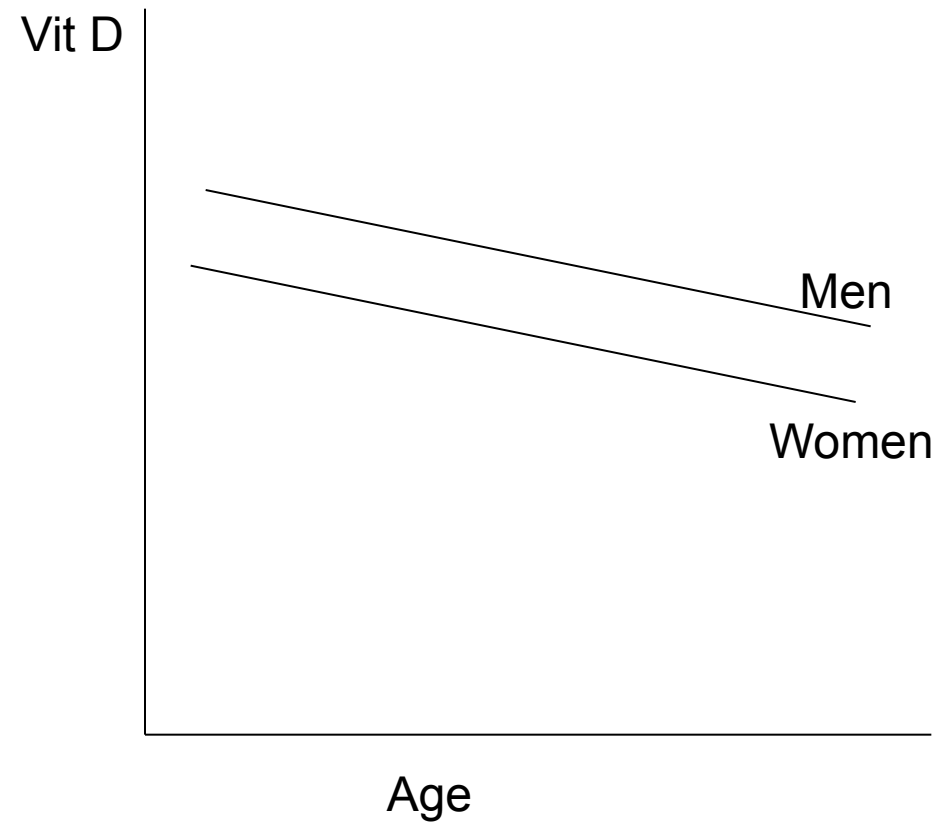
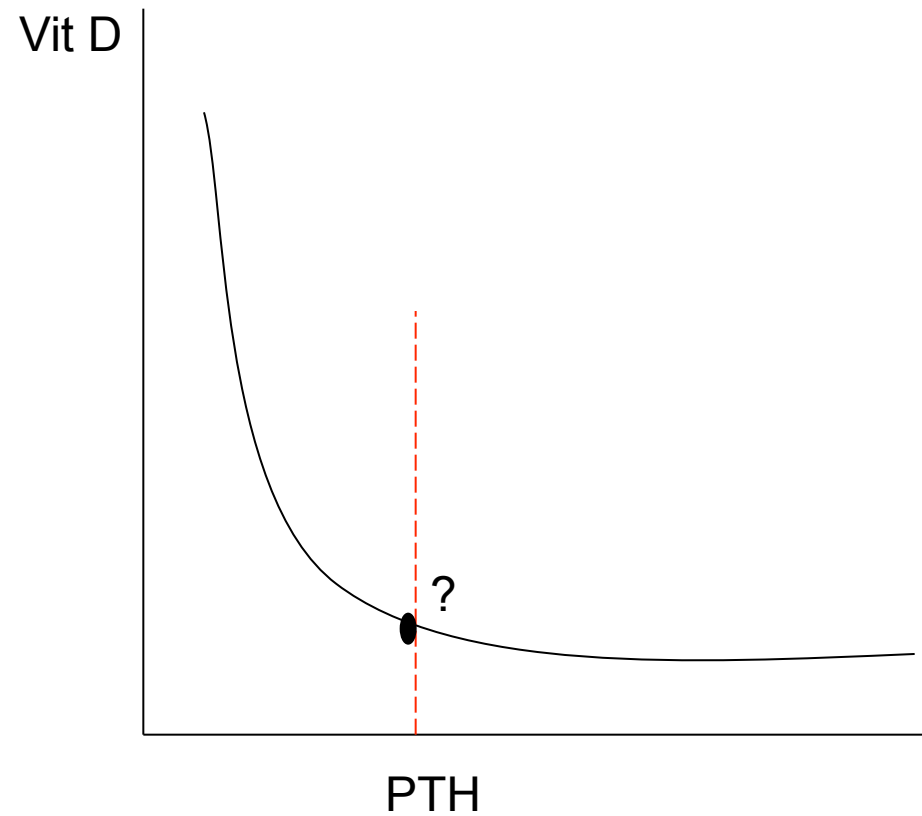
Table 4. Determinants of 25(OH) levels

Determinant	Unit	Regression coefficient	P-value
Age			
Sex			
BMI			
Season			
Residence			
Etc.			

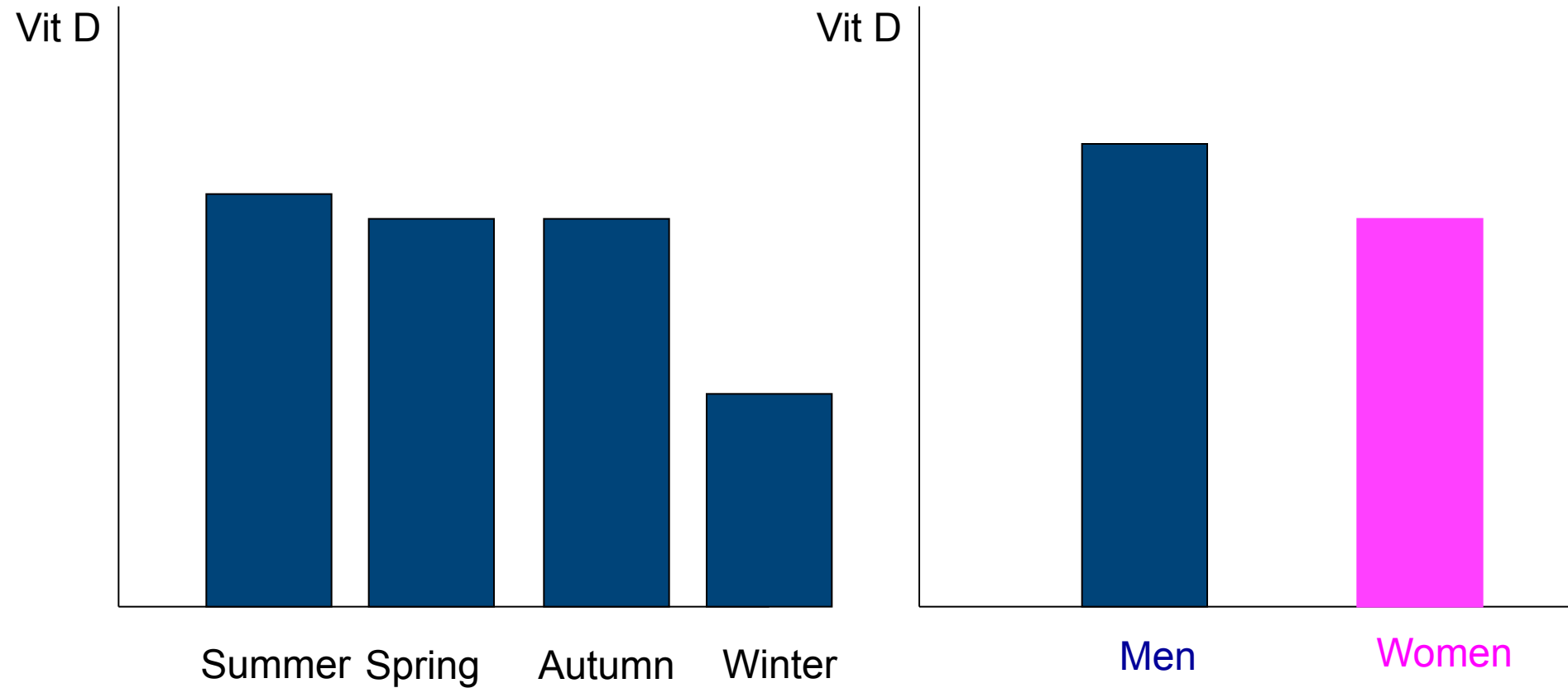
**Which determinants are statistically significant ?**

**How much variance of vit D can be explained by these determinants?**

# Chúng ta muốn có ...



# Chúng ta muốn có ...



# Quy trình phân tích

- Sắp xếp dữ liệu trong Excel
- Dùng R packages cho phân tích
  - **pysch** (for descriptive analyses)
  - **rms** (for logistic regression analysis)
  - **Hmisc** (for data manipulation)
  - **xtab** (for contingency tables)

# Đọc dữ liệu vào R

```
# Tell R where the data are stored
```

```
setwd("C:/Documents and Settings/Tuan/My Documents/_Current  
Projects/_Vietnam/Huong/Vitamin D")
```

```
# Reading the data into a dataset called vd
```

```
vd = read.csv("vitaminD.csv", header=T, na.strings=" ")
```

```
# What variables are in vd
```

```
names(vd)
```

```
# Tell R that we will work on vd from now on
```

```
attach(vd)
```

# Kiểm tra consistency

```
# Let's get some basic statistics for all variables
```

```
# First, we load the package psych
```

```
library(psych)
```

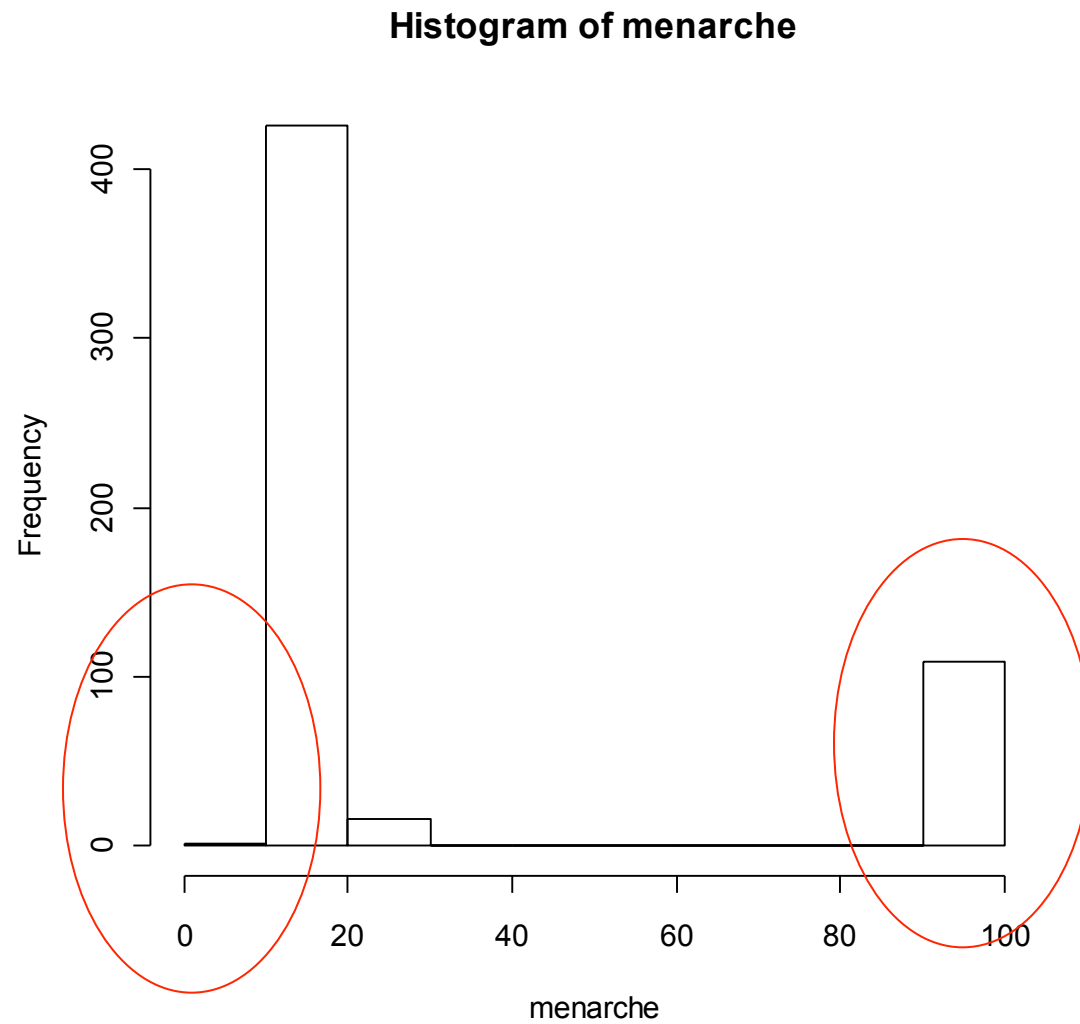
```
# then get statistics
```

```
describe(vd)
```

```
> describe(vd, skew=FALSE)
```

	var	n	mean	sd	median	trimmed	mad	min	max	range	se
id*	1	558	278.37	159.85	279.50	278.58	204.60	1.00	554.00	553.00	6.77
sex	2	558	1.60	0.49	2.00	1.63	0.00	1.00	2.00	1.00	0.02
address*	3	558	4.65	2.17	4.50	4.68	3.71	1.00	8.00	7.00	0.09
region	4	558	2.60	1.30	2.50	2.63	2.22	1.00	4.00	3.00	0.05
age	5	558	46.63	17.79	49.00	46.82	20.76	13.00	83.00	70.00	0.75
menarche	6	552	32.61	33.04	17.00	26.84	2.97	0.00	99.00	99.00	1.41
estradiol	7	558	67.57	74.74	41.35	51.03	25.69	5.00	476.10	471.10	3.16
testo	8	558	2.67	3.26	0.50	2.19	0.68	0.02	13.85	13.83	0.14
vitd	9	558	25.69	8.62	25.83	25.64	7.31	4.00	59.87	55.87	0.36
xlap	10	244	0.39	0.27	0.31	0.35	0.21	0.02	1.57	1.55	0.02
pth	11	558	30.62	12.65	29.04	29.52	11.79	1.20	93.53	92.33	0.54
height	12	558	157.71	7.92	157.00	157.60	7.41	134.00	181.00	47.00	0.34
weight	13	558	50.70	8.08	50.00	50.23	7.41	32.00	85.00	53.00	0.34
bmi	14	558	20.33	2.63	20.00	20.18	2.97	14.00	31.00	17.00	0.11
fnbmd	15	558	0.72	0.12	0.71	0.72	0.12	0.39	1.11	0.72	0.01
hipbmd	16	558	0.81	0.12	0.81	0.81	0.11	0.29	1.17	0.88	0.01
lsbmd	17	554	0.86	0.15	0.87	0.86	0.15	0.42	1.48	1.05	0.01
fracture	18	558	4.90	16.70	2.00	2.00	0.00	1.00	99.00	98.00	0.71
alcohol	19	557	2.63	9.19	2.00	1.82	0.00	1.00	99.00	98.00	0.39
coffee	20	552	1.87	0.34	2.00	1.96	0.00	1.00	2.00	1.00	0.01
tea	21	552	1.61	0.49	2.00	1.64	0.00	1.00	2.00	1.00	0.02
ruralurban	22	558	1.69	0.46	2.00	1.74	0.00	1.00	2.00	1.00	0.02
location	23	558	0.31	0.46	0.00	0.26	0.00	0.00	1.00	1.00	0.02
season	24	558	2.18	0.81	2.00	2.22	1.48	1.00	3.00	2.00	0.03

```
> hist(menarche)
```





# Phân tích mô tả theo nhóm

`describe.by(vd, sex, skew=F, interp=FALSE)`

group: 1

	var	n	mean	sd	median	trimmed	mad	min	max	range	se
id*	1	222	296.79	166.79	302.50	301.37	197.19	3.00	554.00	551.00	11.19
sex	2	222	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00	0.00
address*	3	222	4.85	2.06	5.00	4.94	2.97	2.00	7.00	5.00	0.14
region	4	222	2.72	1.25	3.00	2.77	1.48	1.00	4.00	3.00	0.08
age	5	222	46.67	19.30	49.00	46.57	25.20	14.00	83.00	69.00	1.30
menarche	6	220	53.99	40.85	20.00	53.31	7.41	12.00	99.00	87.00	2.75
estradiol	7	222	42.76	12.46	41.33	42.19	10.90	8.95	92.33	83.38	0.84
testo	8	222	6.24	2.25	6.05	6.16	2.16	0.84	13.85	13.01	0.15
vitd	9	222	28.57	8.94	28.29	28.45	8.46	4.00	59.87	55.87	0.60
xlap	10	100	0.45	0.31	0.35	0.41	0.26	0.02	1.57	1.55	0.03
pth	11	222	31.76	13.50	29.69	30.29	12.34	8.63	93.53	84.90	0.91
height	12	222	164.20	6.03	164.00	164.07	5.93	148.00	181.00	33.00	0.40
weight	13	222	54.46	8.44	53.00	54.06	7.41	35.00	85.00	50.00	0.57
bmi	14	222	20.13	2.72	20.00	19.96	2.97	15.00	31.00	16.00	0.18
fnbmd	15	222	0.76	0.12	0.76	0.75	0.12	0.39	1.11	0.72	0.01
hipbmd	16	222	0.85	0.12	0.84	0.85	0.12	0.29	1.17	0.88	0.01
lsbmd	17	220	0.90	0.14	0.91	0.90	0.14	0.53	1.48	0.94	0.01
fracture	18	222	4.14	14.43	2.00	2.00	0.00	1.00	99.00	98.00	0.97
alcohol	19	222	1.88	6.57	1.00	1.43	0.00	1.00	99.00	98.00	0.44
coffee	20	221	1.79	0.41	2.00	1.86	0.00	1.00	2.00	1.00	0.03

**group: 2**

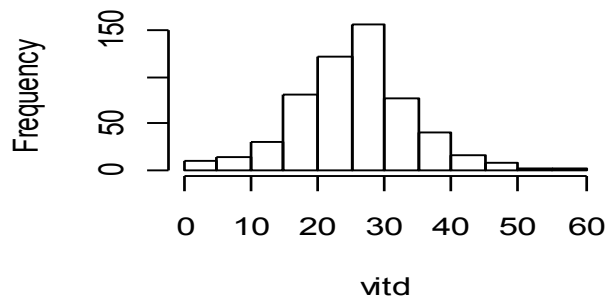
	var	n	mean	sd	median	trimmed	mad	min	max	range	se
id*	1	336	266.19	154.14	266.00	264.52	202.37	1.00	551.00	550.00	8.41
sex	2	336	2.00	0.00	2.00	2.00	0.00	2.00	2.00	0.00	0.00
address*	3	336	4.51	2.23	4.00	4.51	2.97	1.00	8.00	7.00	0.12
region	4	336	2.53	1.33	2.00	2.53	1.48	1.00	4.00	3.00	0.07
age	5	336	46.60	16.74	49.00	47.04	19.27	13.00	82.00	69.00	0.91
menarche	6	332	18.44	14.42	16.00	16.00	2.97	0.00	99.00	99.00	0.79
estradiol	7	336	83.96	92.24	41.74	65.28	36.46	5.00	476.10	471.10	5.03
testo	8	336	0.32	0.59	0.20	0.23	0.19	0.02	7.91	7.89	0.03
vitd	9	336	23.79	7.86	24.26	23.93	7.10	4.00	59.60	55.60	0.43
xlap	10	144	0.34	0.24	0.28	0.31	0.19	0.06	1.26	1.20	0.02
pth	11	336	29.87	12.03	28.56	29.03	11.32	1.20	92.87	91.67	0.66
height	12	336	153.42	5.82	154.00	153.53	5.93	134.00	168.00	34.00	0.32
weight	13	336	48.21	6.78	48.00	47.94	7.41	32.00	74.00	42.00	0.37
bmi	14	336	20.46	2.56	20.00	20.33	2.97	14.00	30.00	16.00	0.14
fnbmd	15	336	0.70	0.11	0.69	0.70	0.12	0.39	1.00	0.60	0.01
hipbmd	16	336	0.78	0.12	0.79	0.78	0.11	0.48	1.10	0.62	0.01
lsbmd	17	334	0.83	0.15	0.85	0.84	0.16	0.42	1.22	0.80	0.01
fracture	18	336	5.40	18.04	2.00	2.00	0.00	1.00	99.00	98.00	0.98
alcohol	19	335	3.12	10.56	2.00	2.00	0.00	1.00	99.00	98.00	0.58
coffee	20	331	1.92	0.28	2.00	2.00	0.00	1.00	2.00	1.00	0.02
tea	21	331	1.80	0.40	2.00	1.87	0.00	1.00	2.00	1.00	0.02
ruralurban	22	336	1.65	0.48	2.00	1.69	0.00	1.00	2.00	1.00	0.03
location	23	336	0.35	0.48	0.00	0.31	0.00	0.00	1.00	1.00	0.03
season	24	336	2.16	0.85	2.00	2.20	1.48	1.00	3.00	2.00	0.05

# Phân bố của dữ liệu

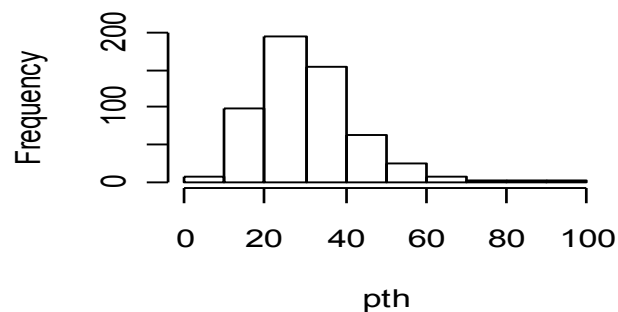
```
par(mfrow=c(2,2))
```

```
hist(vitd); hist(pth); hist(bmi); hist(age)
```

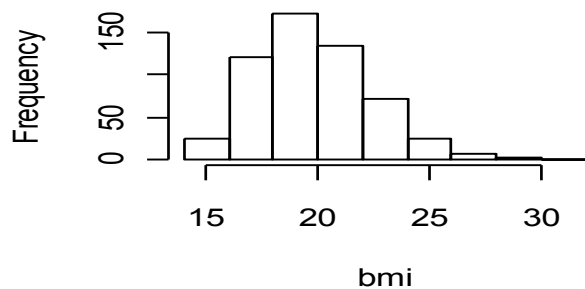
**Histogram of vitd**



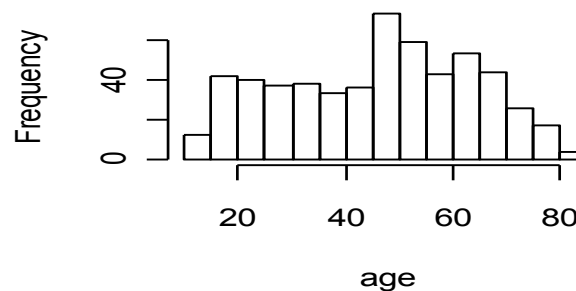
**Histogram of pth**



**Histogram of bmi**



**Histogram of age**



# Định nghĩa thiếu vitamin D

```
library(Hmisc)
insuff = cut2(vitd, 30)

def = vitd
def = replace(def, vitd < 20.0, 1)
def = replace(def, vitd >= 20, 0)

library(gmodels)
CrossTable(def, sex, digits=3, chisq=T, fisher=T)
CrossTable(def, season, digits=3, chisq=T, fisher=T)
CrossTable(def, location, digits=3, chisq=T, fisher=T)
```

# Thiếu vitamin D theo giới tính

Total Observations in Table: 558

sex	def		Row Total
	0	1	
1	187	35	222
	1.988	6.290	
	0.842	0.158	0.398
	0.441	0.261	
	0.335	0.063	
2	237	99	336
	1.313	4.156	
	0.705	0.295	0.602
	0.559	0.739	
	0.425	0.177	
Column Total	424	134	558
	0.760	0.240	

# Kiểm định thống kê

Statistics for All Table Factors

Pearson's Chi-squared test

-----  
Chi<sup>2</sup> = 13.74685      d.f. = 1      p = 0.0002091717

Pearson's Chi-squared test with Yates' continuity correction

-----  
Chi<sup>2</sup> = 13.00639      d.f. = 1      p = 0.0003104303

Fisher's Exact Test for Count Data

-----  
Sample estimate odds ratio: 2.22876

Alternative hypothesis: true odds ratio is not equal to 1

p = 0.0001784270

95% confidence interval: 1.427180 3.540137

# Thiếu vitamin D theo mùa

def			
season	0	1	Row Total
1	77	64	141
	8.479	26.828	
	0.546	0.454	0.253
	0.182	0.478	
	0.138	0.115	
2	147	30	177
	1.163	3.679	
	0.831	0.169	0.317
	0.347	0.224	
	0.263	0.054	
3	200	40	240
	1.705	5.396	
	0.833	0.167	0.430
	0.472	0.299	
	0.358	0.072	
Column Total	424	134	558
	0.760	0.240	

# Yếu tố nguy cơ thiếu vitamin D

```
library(epicalc)
```

```
logistic.display(glm(def ~ sex, family=binomial))
```

```
logistic.display(glm(def ~ location, family=binomial))
```

	OR(95%CI)	P(Wald's test)	P(LR-test)
<b>sex (cont. var.)</b>	<b>2.23 (1.45,3.43)</b>	<b>&lt; 0.001</b>	<b>&lt; 0.001</b>

Log-likelihood = -300.4421

No. of observations = 558

AIC value = 604.8842

	OR(95%CI)	P(Wald's test)	P(LR-test)
<b>location: 1 vs 0</b>	<b>3.71 (2.47,5.58)</b>	<b>&lt; 0.001</b>	<b>&lt; 0.001</b>

Log-likelihood = -287.407

No. of observations = 558

AIC value = 578.8139



# Yếu tố liên quan đến thiếu vitamin D: phân tích đơn biến

```
logistic.display(glm(def ~ bmi, family=binomial))
```

```
logistic.display(glm(def ~ factor(season), family=binomial))
```

	OR(95%CI)	P(Wald's test)	P(LR-test)
<b>bmi (cont. var.)</b>	<b>1.05 (0.98,1.13)</b>	<b>0.174</b>	<b>0.176</b>

Log-likelihood = -306.6787

No. of observations = 558

AIC value = 617.3574

	OR	lower95ci	upper95ci	Pr(> Z )
<b>factor(season)2</b>	<b>0.2455357</b>	<b>0.1468710</b>	<b>0.4104812</b>	<b>8.506640e-08</b>
<b>factor(season)3</b>	<b>0.2406250</b>	<b>0.1497158</b>	<b>0.3867354</b>	<b>4.002872e-09</b>

# Yếu tố liên quan đến thiếu vitamin D: phân tích đa biến

```
logistic.display(glm(def ~ age+sex+bmi+location+factor(season),  
  family=binomial))
```

	OR	lower95ci	upper95ci	Pr(> Z )
age	0.9821236	0.9698786	0.9945232	0.004833844
sex	1.9543516	1.2394551	3.0815883	0.003927809
bmi	0.9878399	0.9038924	1.0795839	0.787155190
location	2.4925512	0.9452428	6.5727150	0.064872828
factor(season) 2	0.5482407	0.2237126	1.3435448	0.188766835
factor(season) 3	0.5817646	0.1990562	1.7002735	0.322202383

# Yếu tố liên quan đến thiếu vitamin D: mô hình hồi qui nhị phân

```
library(MASS)  
summary(glm.nb(def ~ age+sex+location, data=vd))
```

## Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-2.060883	0.411777	-5.005	5.59e-07	***
age	-0.013404	0.004877	-2.748	0.0060	**
sex	0.503312	0.198337	2.538	0.0112	*
location	0.938540	0.175347	5.352	8.68e-08	***

# Yếu tố liên quan đến vitamin D: phân tích đa biến

```
library(leaps)
library(car)
library(MASS)
library(relaimpo)

# stepwise regression

fit = lm(vitd~age+bmi+sex+factor(season), data=vd)
step = stepAIC(fit, direction="both")
step$anova

# All Subsets Regression

fit2 = regsubsets(vitd~age+bmi+sex
  +factor(season),data=vd,nbest=5)

summary(fit2)

plot(fit2,scale="r2")

subsets(fit2, statistic="rsq")
```

# Yếu tố liên quan đến vitamin D: phân tích đa biến

```
# Calculate Relative Importance for Each Predictor
library(relaimpo)
fit = lm(vitd~age+bmi+sex+factor(season), data=vd)
calc.relimp(fit, rela=TRUE)

# Bootstrap Measures of Relative Importance (1000 samples)

library(boot)
boot = boot.relimp(fit, b = 1000), rank = TRUE,
  diff=TRUE, rela=TRUE)
booteval.relimp(boot)

# print result
plot(booteval.relimp(boot,sort=TRUE)) # plot result
```

### Relative importance metrics:

	lmg
factor(season)	0.48695621
age	0.06065485
bmi	0.02030591
sex	0.43208303

### Average coefficients for different model sizes:

	1group	2groups	3groups	4groups
age	0.04504289	0.05004166	0.04961454	0.04470275
bmi	-0.23392346	-0.14266033	-0.05454423	0.03080951
sex	-4.77558719	-4.61454708	-4.46527769	-4.34587273
`factor(season)`2	5.45826902	5.18184506	4.86820278	4.52678038
`factor(season)`3	5.93707270	5.90083953	5.81691391	5.69611558

# Yếu tố liên quan đến vitamin D: stepwise regression

```
> fit = lm(vitd~age+bmi+sex+factor(season),  
  data=vd)  
> step = stepAIC(fit, direction="both")
```

Start: AIC=2322.88

vitd ~ age + bmi + sex + factor(season)

	Df	Sum of Sq	RSS	AIC
- bmi	1	3	35095	2321
<none>			35092	2323
- age	1	324	35416	2326
- sex	1	2425	37517	2358
- factor(season)	2	2616	37708	2359

Step: AIC=2320.93

vitd ~ age + sex + factor(season)

	Df	Sum of Sq	RSS	AIC
<none>			35095	2321
+ bmi	1	3	35092	2323
- age	1	371	35466	2325
- sex	1	2422	37517	2356
- factor(season)	2	2895	37989	2361

# Yếu tố liên quan đến vitamin D: stepwise regression

```
> step$anova  
Stepwise Model Path  
Analysis of Deviance Table
```

Initial Model:

```
vitd ~ age + bmi + sex + factor(season)
```

Final Model:

```
vitd ~ age + sex + factor(season)
```

	Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1				552	35091.86	2322.882
2	- bmi	1	2.975332	553	35094.83	2320.929



# Yếu tố liên quan đến vitamin D: mô hình sau cùng

```
fit = lm(vitd ~ age+sex+factor(season), data=vd)
summary(fit)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	26.66091	1.64606	16.197	< 2e-16	***
age	0.04591	0.01899	2.418	0.0159	*
sex	-4.33997	0.70246	-6.178	1.26e-09	***
factor(season)2	4.47858	0.91536	4.893	1.31e-06	***
factor(season)3	5.62941	0.84679	6.648	7.14e-11	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

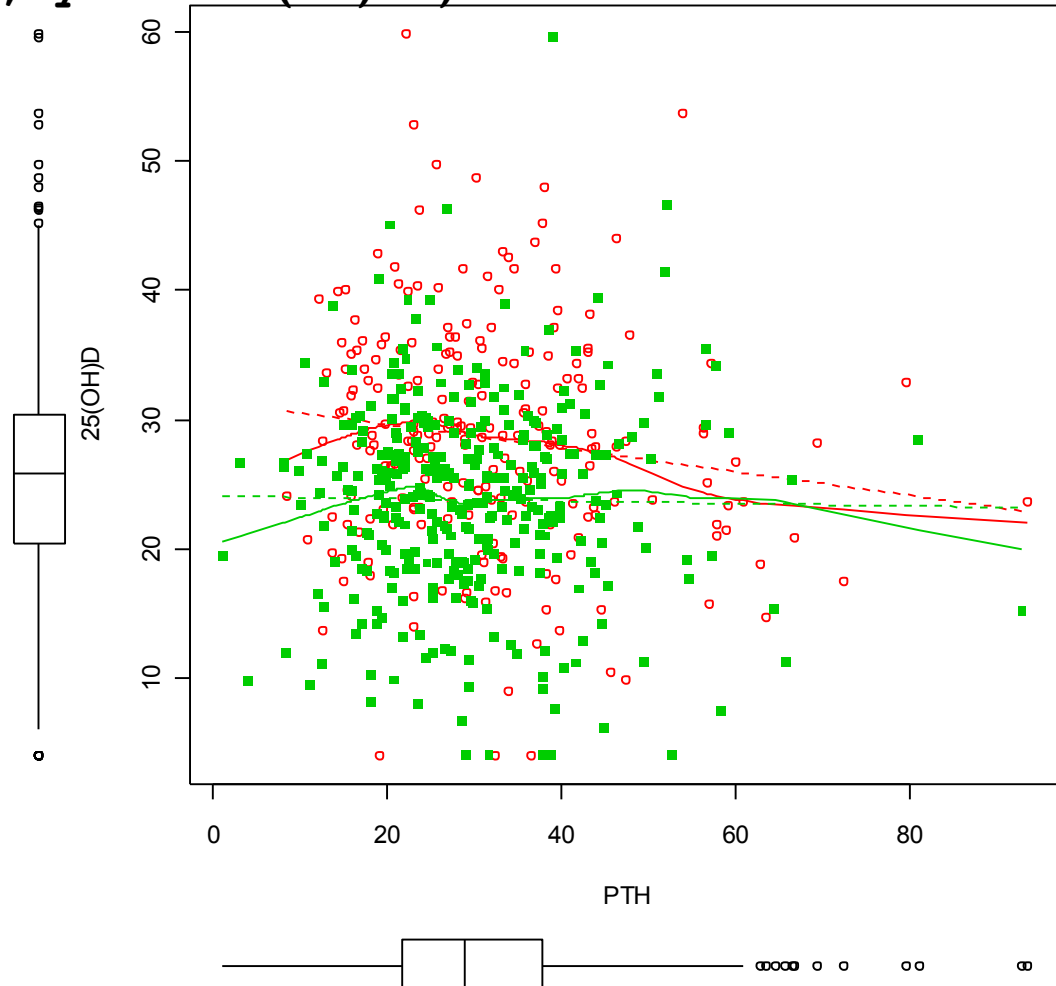
Residual standard error: 7.966 on 553 degrees of freedom

Multiple R-squared: 0.1521, Adjusted R-squared: 0.146

F-statistic: 24.81 on 4 and 553 DF, p-value: < 2.2e-16

# Mối liên quan giữa vitamin D và PTH

```
library(car)
scatterplot(vitd ~ pth | sex, pch=c(1, 15),
            xlab="PTH", ylab="25(OH)D")
```



# Mối liên quan giữa vitamin D và PTH

```
> fit = lm(vitd ~ pth, data=vd)
> summary(fit)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	26.70950	0.95613	27.935	<2e-16 ***
pth	-0.03329	0.02886	-1.153	0.249

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.618 on 556 degrees of freedom

Multiple R-squared: 0.002387, Adjusted R-squared: 0.0005927

F-statistic: 1.33 on 1 and 556 DF, p-value: 0.2492

# Mối liên quan giữa vitamin D và PTH

```
> fit = lm(vitd ~ pth+sex+pth:sex, data=vd)
> summary(fit)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	38.82787	3.09515	12.545	< 2e-16	***
pth	-0.17214	0.09074	-1.897	0.0583	.
sex	-7.36461	1.86973	-3.939	9.23e-05	***
pth:sex	0.08091	0.05586	1.448	0.1481	

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.282 on 554 degrees of freedom

Multiple R-squared: 0.08187, Adjusted R-squared: 0.0769

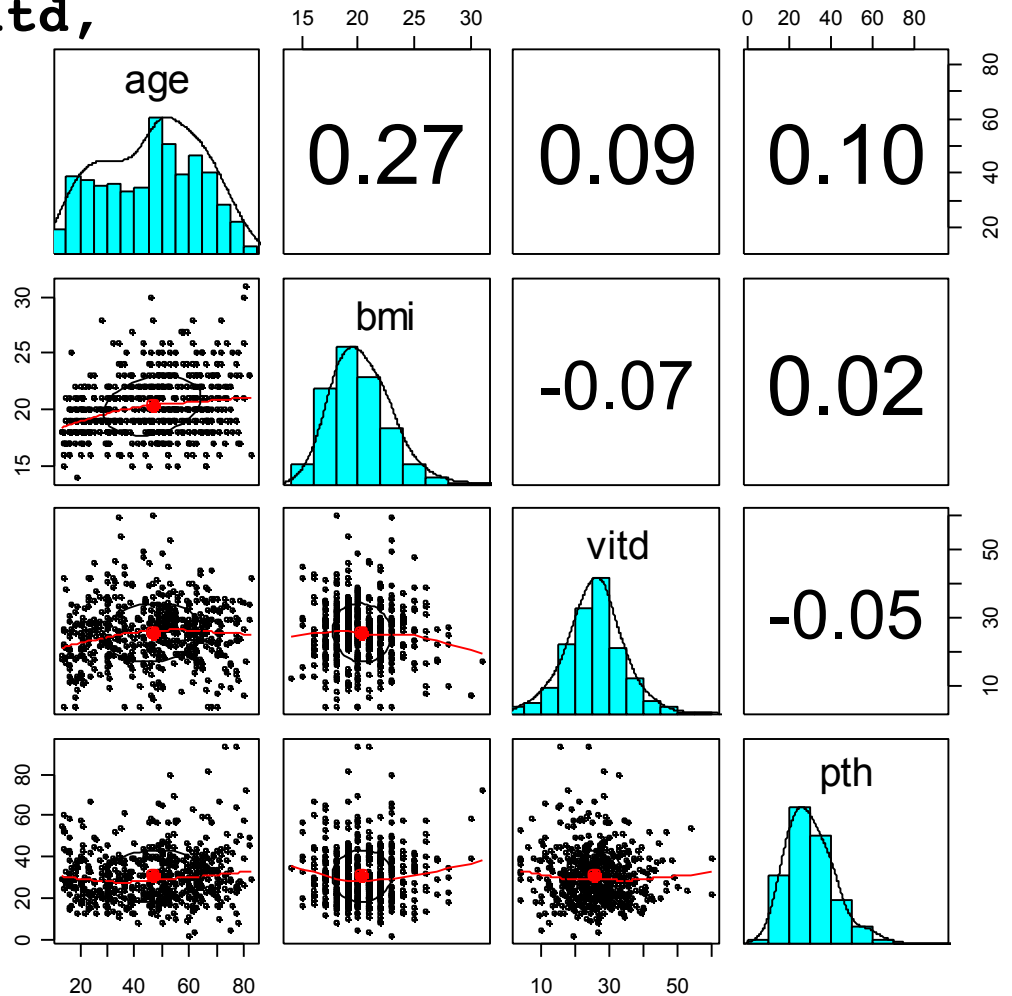
F-statistic: 16.47 on 3 and 554 DF, p-value: 2.907e-10

# Thử xem toàn cảnh ...

```
library(psych)
```

```
dd = cbind(age, bmi, vitd,  
            pth)
```

```
pairs.panels(dd)
```



# Error plot

```
error.bars.by(vitd, sex, bars=TRUE, labels=c("Men",  
"Women"), ylab = "25(OH)D", xlab=" ", ylim=c(0,30))
```

