Phân tích dữ liệu với R

Trần Tuấn Anh
Khoa QTKD – Trường Đại học Mở TPHCM

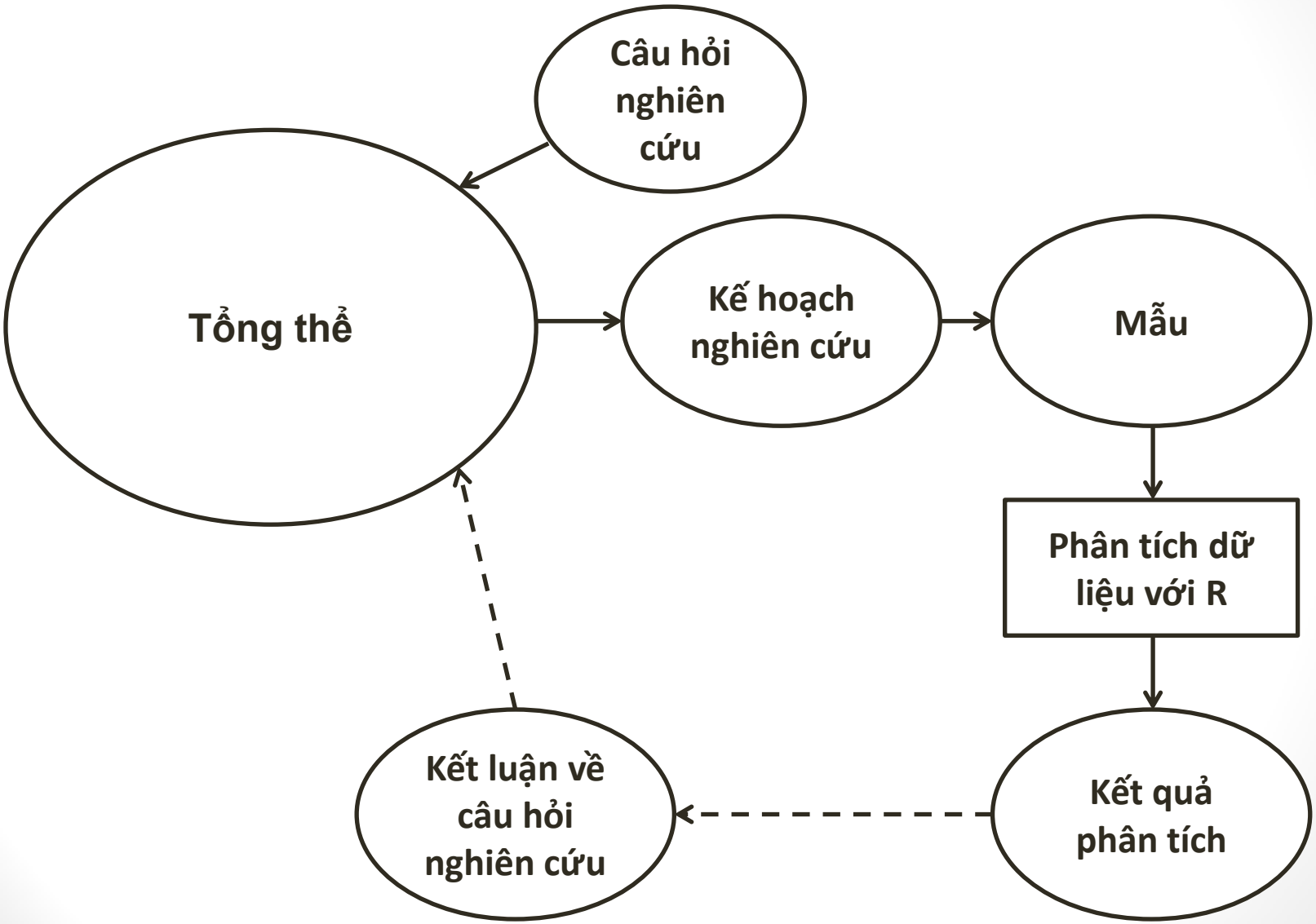
Mục tiêu

- Giới thiệu R.
- Trình bày một số thí dụ phân tích dữ liệu với R.
- Thí dụ lập trình phân tích dữ liệu với R.

Nội dung trình bày

- Giới thiệu R
- Giao diện và cú pháp cơ bản
- Tính toán cơ bản và biểu đồ
- Phân tích dữ liệu bằng thống kê: kiểm định giả thuyết, ANOVA, hồi qui
- Thí dụ lập trình trong R

Phân tích dữ liệu với R



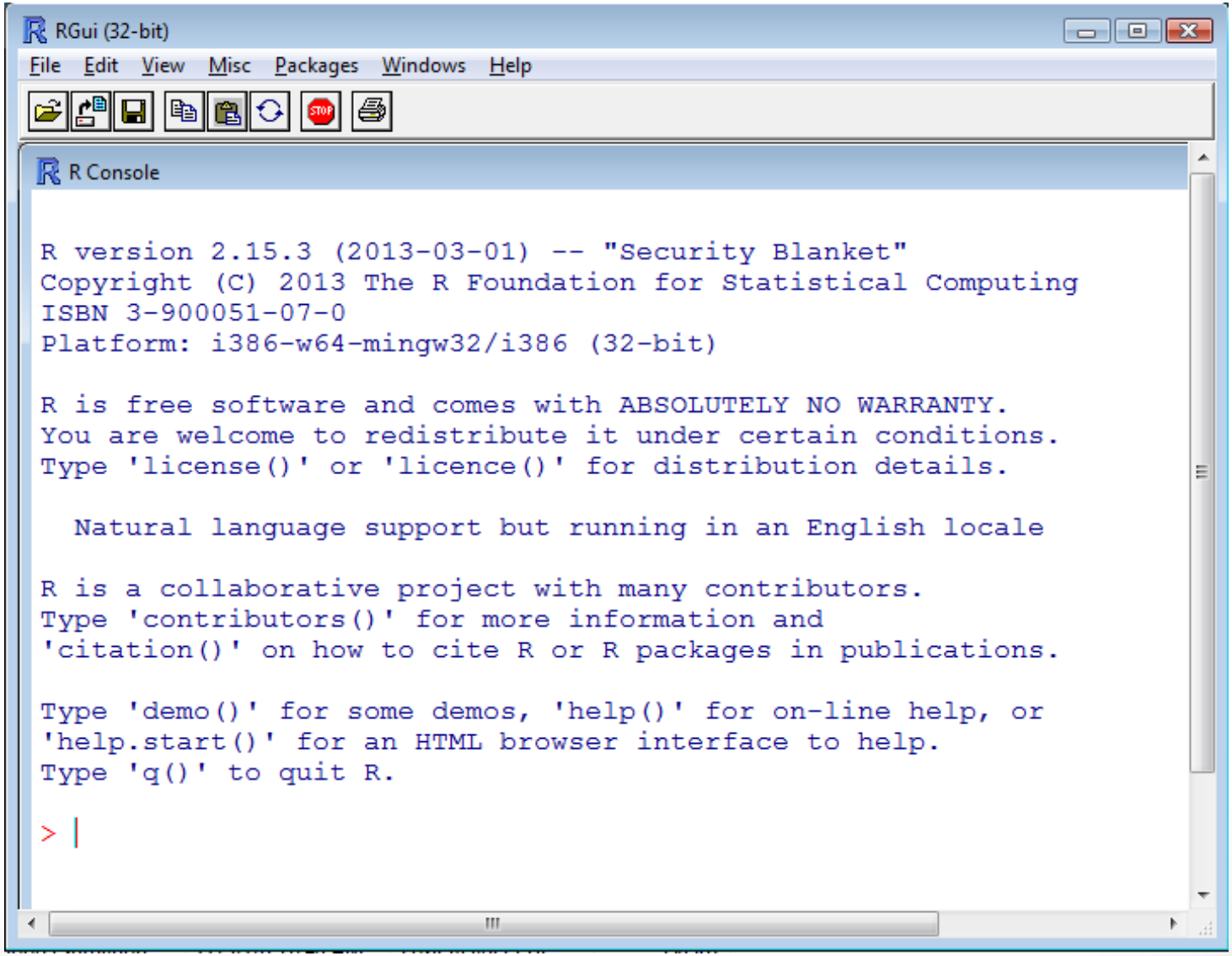
Vì sao dùng R?

- Miễn phí.
- Phân tích thống kê và biểu đồ.
- Chạy trên nhiều hệ điều hành.
- Có khả năng lập trình.
- Được cập nhật, nâng cấp.

Cài đặt

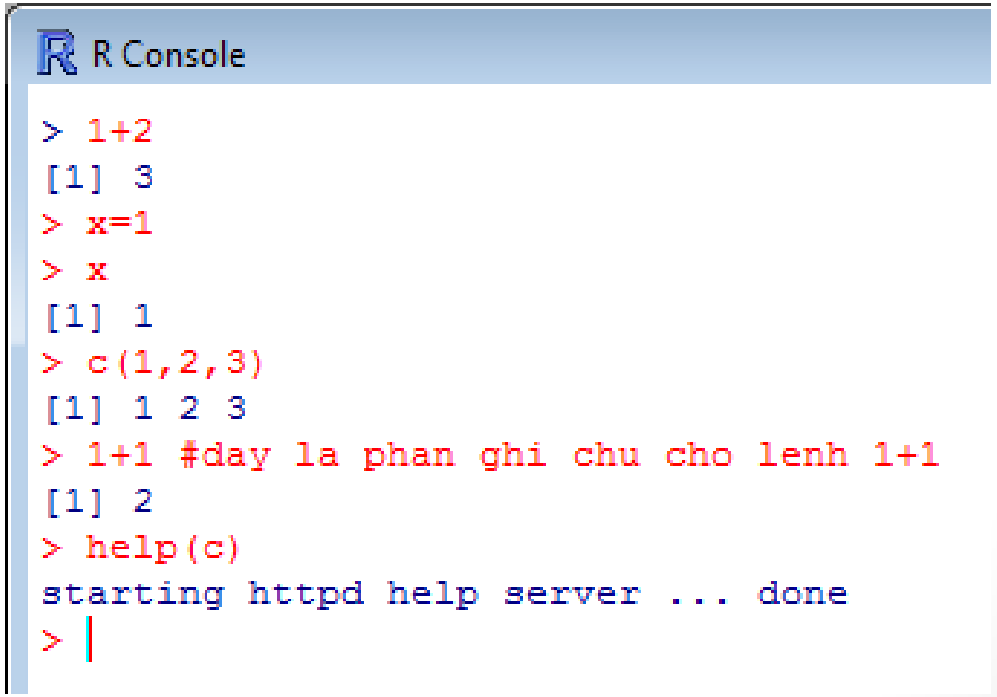
- Đường liên kết tải R về máy:
  - http://cran.r-project.org/mirrors.html
- Chọn vị trí gần nhất.

Giao diện


RGui (32-bit)
File Edit View Misc Packages Windows Help
R Console
R version 2.15.3 (2013-03-01) -- "Security Blanket"
Copyright (C) 2013 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-w64-mingw32/i386 (32-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>

Cú pháp cơ bản

- Tương tác dòng lệnh
- Tính số học
- Gán biến
- Gọi hàm
- Ghi chú
- Trợ giúp



Kiểu dữ liệu cơ bản

- Numeric
- Integer
- Complex
- Logical
- character

```
> x = 10.5
> x
[1] 10.5
> class(x)
[1] "numeric"
```

```
> z = 1 + 2i
> z
[1] 1+2i
> class(z)
[1] "complex"
```

```
> x = 1; y = 2
> z = x > y
> z
[1] FALSE
> class(z)
[1] "logical"
```

```
> u=TRUE
> v=FALSE
> u&v       # u AND v
[1] FALSE
> !u        #u OR v
[1] FALSE
> !u        #phu dinh u
[1] FALSE
```

```
> ho = "Nguyen"
> ten = "Minh"
> paste(ho,ten)
[1] "Nguyen Minh"
```

Dữ liệu trong R

- Vector
- Matrix
- List
- Data Frame

```
> c(2, 3, 5)
[1] 2 3 5
```

```
> A = matrix(
+   c(2, 4, 3, 1, 5, 7),
+   nrow=2,
+   ncol=3,
+   byrow = TRUE)
```

```
> A
     [,1] [,2] [,3]
[1,]    2    4    3
[2,]    1    5    7
```

Dữ liệu trong R

- Vector
- Matrix
- List
- Data Frame

```
> n=c(2,3,5)
> s=c("aa","bb","cc","dd","ee")
> b=c(TRUE, FALSE, TRUE, FALSE, FALSE)
> x=list(n,s,b,3)
> x
[[1]]
[1] 2 3 5

[[2]]
[1] "aa" "bb" "cc" "dd" "ee"

[[3]]
[1]  TRUE FALSE  TRUE FALSE FALSE

[[4]]
[1] 3
```

```
> n=c(2,3,5)
> id=c(1,2,3,4,5)
> ten=c("Dong","Tay","Nam","Bac","Trung")
> dangky=c(TRUE,TRUE,FALSE,TRUE,FALSE)
> dfr=data.frame(id,ten,dangky)
> dfr
  id   ten dangky
1  1  Dong   TRUE
2  2   Tay   TRUE
3  3   Nam  FALSE
4  4   Bac   TRUE
5  5 Trung  FALSE
```

```
> doanhso=c(120,200,250,320,180)
> dfr=data.frame(dfr,doanhso)
> dfr
  id   ten dangky doanhso
1  1  Dong   TRUE      120
2  2   Tay   TRUE      200
3  3   Nam  FALSE      250
4  4   Bac   TRUE      320
5  5 Trung  FALSE      180
> mean(dfr$doanhso)
[1] 214
```

Lấy dữ liệu từ ngoài

- từ tập tin text

```
chol = read.table("chol.txt", header=TRUE)
```

- từ tập tin Excel

```
library(gdata)
df1 = read.xls(“file01.xls”, header=TRUE)
```

- từ tâp tin SPSS

```
library(foreign)
df1 = read.spss(“file01.sav”, to.data.frame=TRUE)
```

- từ tâp tin Minitab

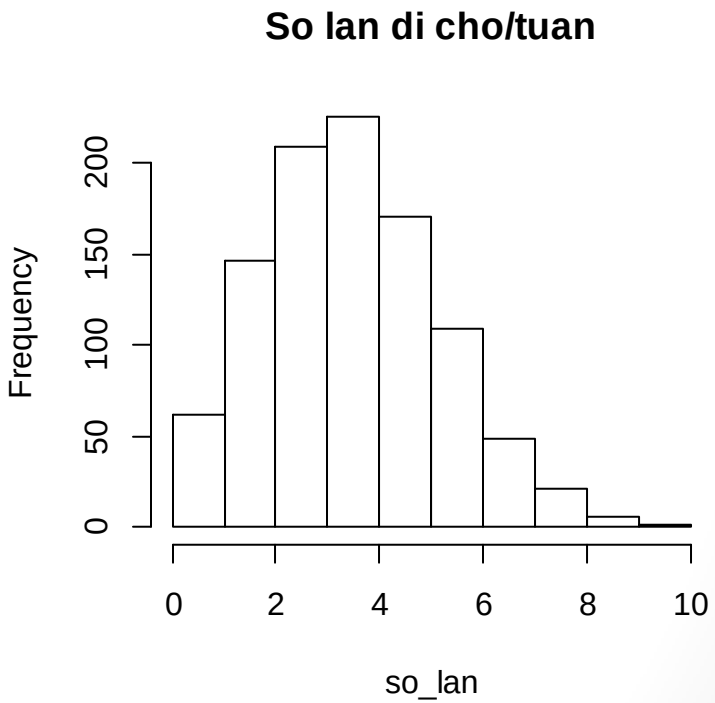
```
library(foreign)
df1 = read.mtp("file01.mtp")
```

Tính toán cơ bản

- Thí dụ mô phỏng
- Chọn mẫu ngẫu nhiên
- Thí dụ biểu đồ

Mô phỏng hàm nhị thức

```
> b=rbinom(1000,20,0.20)
> table(b)
b
  0   1   2   3   4   5   6   7   8   9  10
 11  60 123 225 212 172 101  61  31   3   1
> hist(b,main="So lan di cho/tuan")
> so_lan=rbinom(1000,20,0.2)
> table(so_lan)
so_lan
  0   1   2   3   4   5   6   7   8   9  10
 11  51 146 209 225 171 109  49  21   6   2
> hist(so_lan,main="So lan di cho/tuan")
```

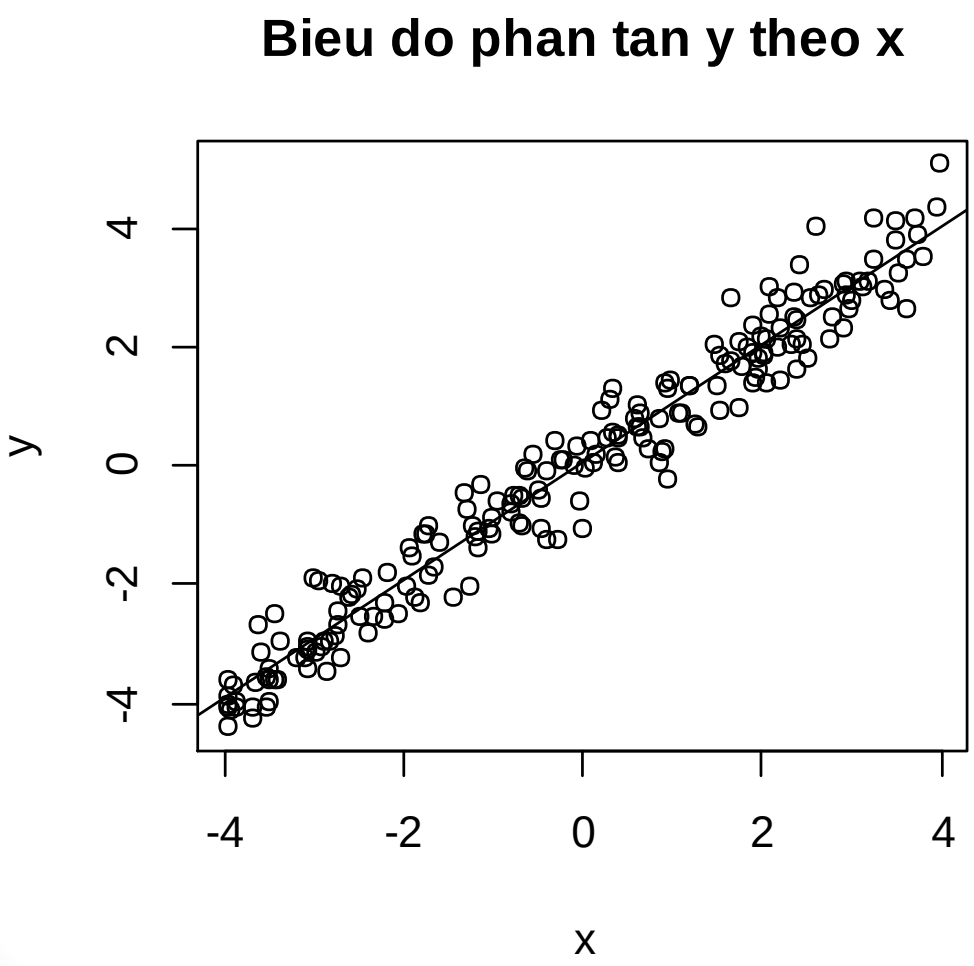


Chọn mẫu ngẫu nhiên

```
> sample(1:1000,50)
 [1] 893 215   4 126  39 140 957 424 514 101 829 815 761 919
[20] 846 485 417 194  30 301  99 510 278 811 876 449 889 243
[39] 380 405 676  89  76 752 103 871  58 994 105  85
```

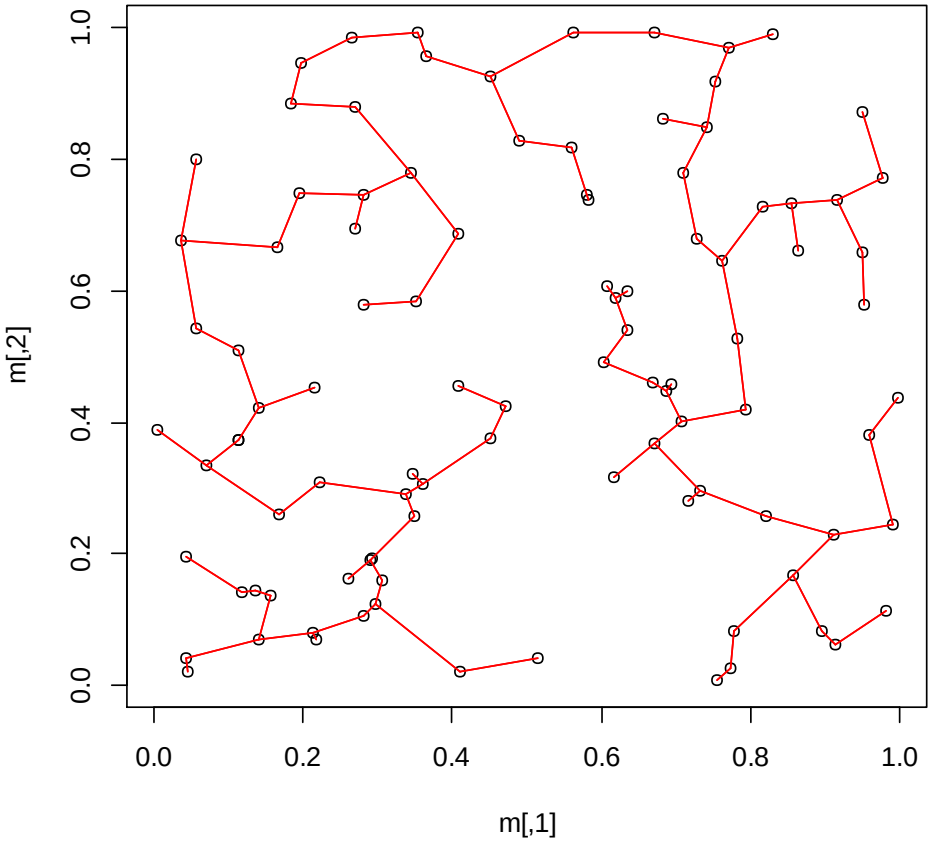
Biểu đồ

```
> n=200
> x=runif(n,-4,4)
> y=x+0.5*rnorm(n)
> plot(x,y,main="Bieu do phan tan y theo x")
> reg=lm(y~x)
> abline(reg)
```



Biểu đồ

- Cây bao trùm tối thiểu



Phân tích bằng thống kê

- Kiểm định giả thuyết
- ANOVA
- Hồi qui
- Tương quan

Kiểm định giả thuyết

- Kiểm định t 1 mẫu

  t.test(x,mu=a)

- Kiểm định Wilcoxon

  wilcox.test(x, mu=a)

Kiểm định giả thuyết

- Kiểm định t 2 mẫu

  t.test(x~g)
- Kiểm định phương sai bằng nhau

  var.test(x~g)
- Kiểm định t 2 mẫu với phương sai bằng nhau

  t.test(x~g, var.equal=T)

Kiểm định giả thuyết

- Kiểm định Wilcoxon

  wilcox.test(x~g)
- Kiểm định paired t test

  t.test(x1, x2, paired=T)
- Kiểm định Wilcoxon mẫu từng cặp

  wilcox.test(x1, x2, paired=T)

ANOVA

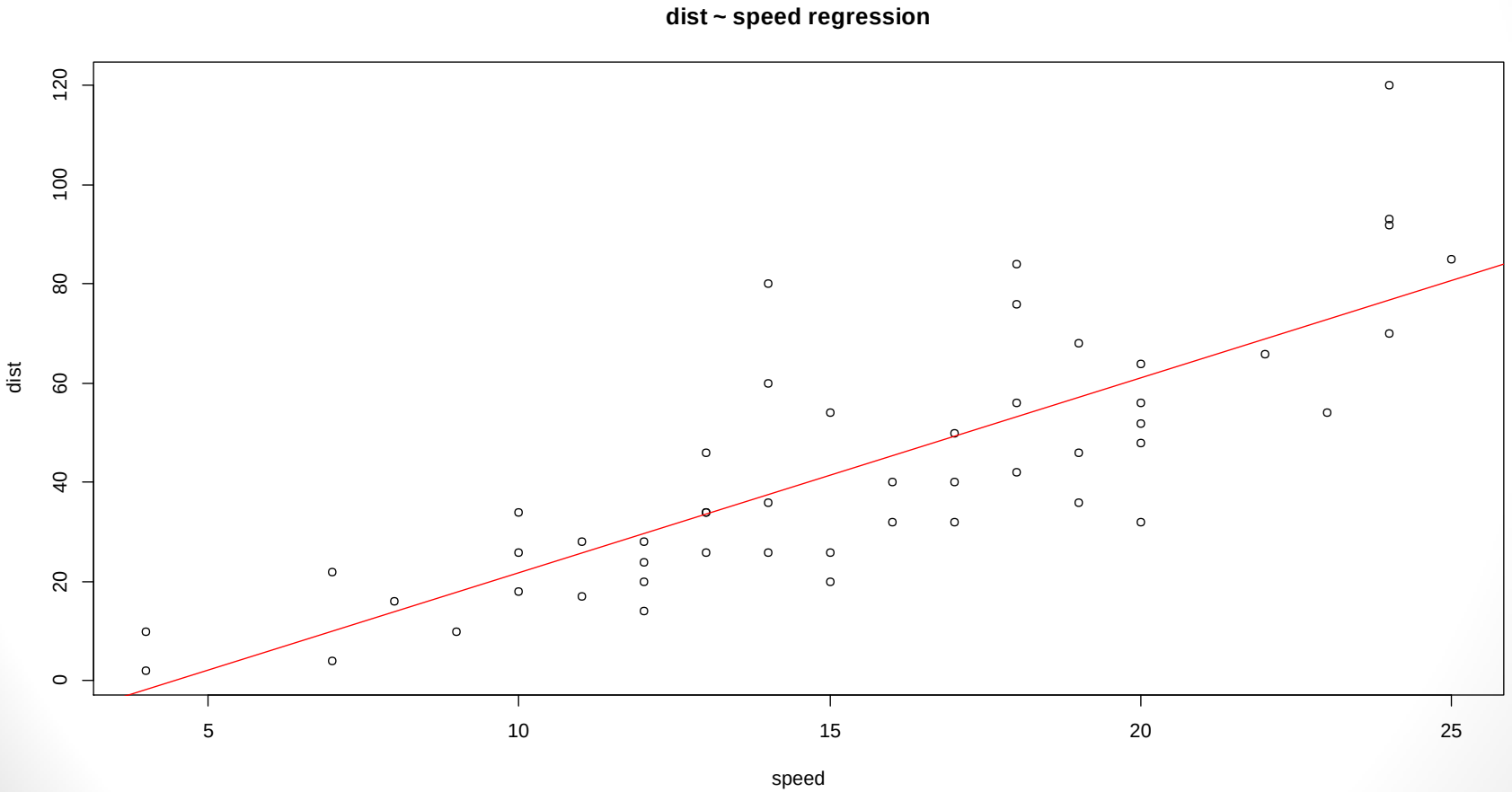
- ANOVA
  - attach(dataset)
  - summary(dataset)
  - res=aov(x1~g)
  - summary(res)
- So sánh bắt cặp
  - TukeyHSD(res)
  - plot(TukeyHSD(res),ordered=T)

Hồi qui

Thí dụ:

```
data(cars)
plot(cars)
abline(lm(cars$dist ~ cars$speed), col='red')
title(main="dist ~ speed regression")
```

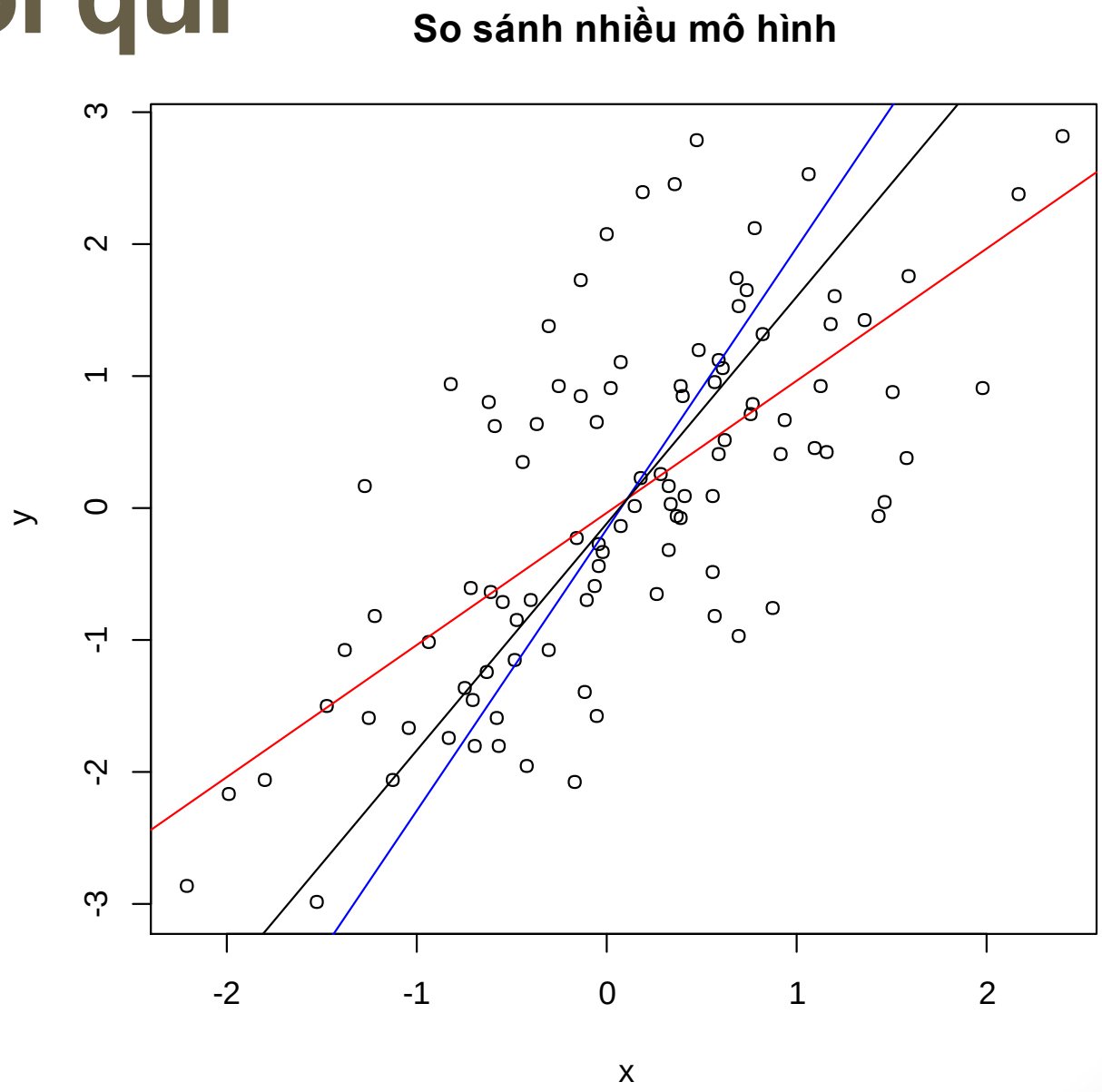
Hồi qui


dist ~ speed regression
dist
speed
0
20
40
60
80
100
120
5
10
15
20
25

Trình bày nhiều mô hình

```
set.seed(1)
x <- rnorm(100)
y <- x + rnorm(100)
plot(y~x)
r <- lm(y~x)
abline(r, col='red')
r <- lm(x ~ y)
a <- r$coefficients[1] # hệ số chặn
b <- r$coefficients[2] # hệ số góc
abline(-a/b , 1/b, col="blue")
r <- princomp(cbind(x,y))
b <- r$loadings[2,1] / r$loadings[1,1]
a <- r$center[2] - b * r$center[1]
abline(a,b)
title(main='So sánh nhiều mô hình')
```

Hồi qui



Tương quan

- **Tương quan Pearson**

```
cor(x1, x2, use="complete.obs")

cor(dataset,use="complete.obs")

cor.test(x1,x2)
```

Tương quan

- **Tương quan Spearman**

```
cor(x1,x2, use="complete.obs")

cor(dataset,use="complete.obs")

cor.test(x1,x2)
```

Tương quan

- **Tương quan Kendall**

```
cor.test(x1,x2, method="kendall")
```

Thí dụ lập trình

- Tự động hóa phân tích dữ liệu.

Tính thống kê Bayesian

```
bayes <- function(x, prior.mean, prior.var)
{
n <- length(x)
sample.mean <- mean(x)
sample.var <- var(x)
numerator <- (prior.mean/prior.var) + (n*sample.mean/sample.var)
denominator <- 1/prior.var + n/sample.var
posterior.mean = numerator/denominator
posterior.var = 1/denominator
a <- "Posterior mean = "
b <- "Posterior variance = "
cat("Sample size = ", n, "\n")
cat("Sample mean = ", sample.mean, "\n")
cat("Sample var = ", sample.var, "\n")
cat("Prior mean = ", prior.mean, "\n")
cat("Prior var = ", prior.var, "\n")
cat(a, posterior.mean, "\n")
cat(b, posterior.var, "\n")
}
```

Thí dụ lập trình

bmd <- c(1.0, 1.5, 2.1, 1.7, 1.8, 0.9, 0.7)
bayes(bmd, 1.0, 0.0144)

Kết quả:

```
> bayes(bmd, 1.0, 0.0144)
Sample size =  7
Sample mean =  1.385714
Sample var =  0.2747619
Prior mean =  1
Prior var =  0.0144
Posterior mean =  1.103525
Posterior variance =  0.01053507
>
```

Chân thành cám ơn!