Bài giảng 21: Hồi qui tuyến tính đơn giản: Kiểm tra giả định

Nguyễn Văn Tuấn

Garvan Institute of Medical Research, Australia

Đại học Tôn Đức Thắng, Việt Nam

Nội dung

- Giả định trong hồi qui tuyến tính
- Định nghĩa *dao động dư - residual*
- Những chỉ số liên quan đến dao động dư
- Hàm R

Giả định

Mô hình hồi qui tuyến tính

- Mô hình cho quần thể

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

- Dữ liệu thực tế

$$y = a + bx + e$$

Giả định

- Mối liên quan giữa X và Y là tuyến tính (linear) về *tham số*
- *X* không có sai số ngẫu nhiên
- Giá trị của *Y* độc lập với nhau (vd, Y_1 không liên quan với Y_2) ;
- **Sai số ngẫu nhiên (e):**
  - phân bố chuẩn,
  - trung bình 0,
  - phương sai bất biến

$$\varepsilon \sim N(0, \sigma^2)$$

Residuals – độ dao động dư

- Mô hình cho dữ liệu

$$y = a + bx + e$$

Giá trị trung bình: E(y) = yhat = a + bx

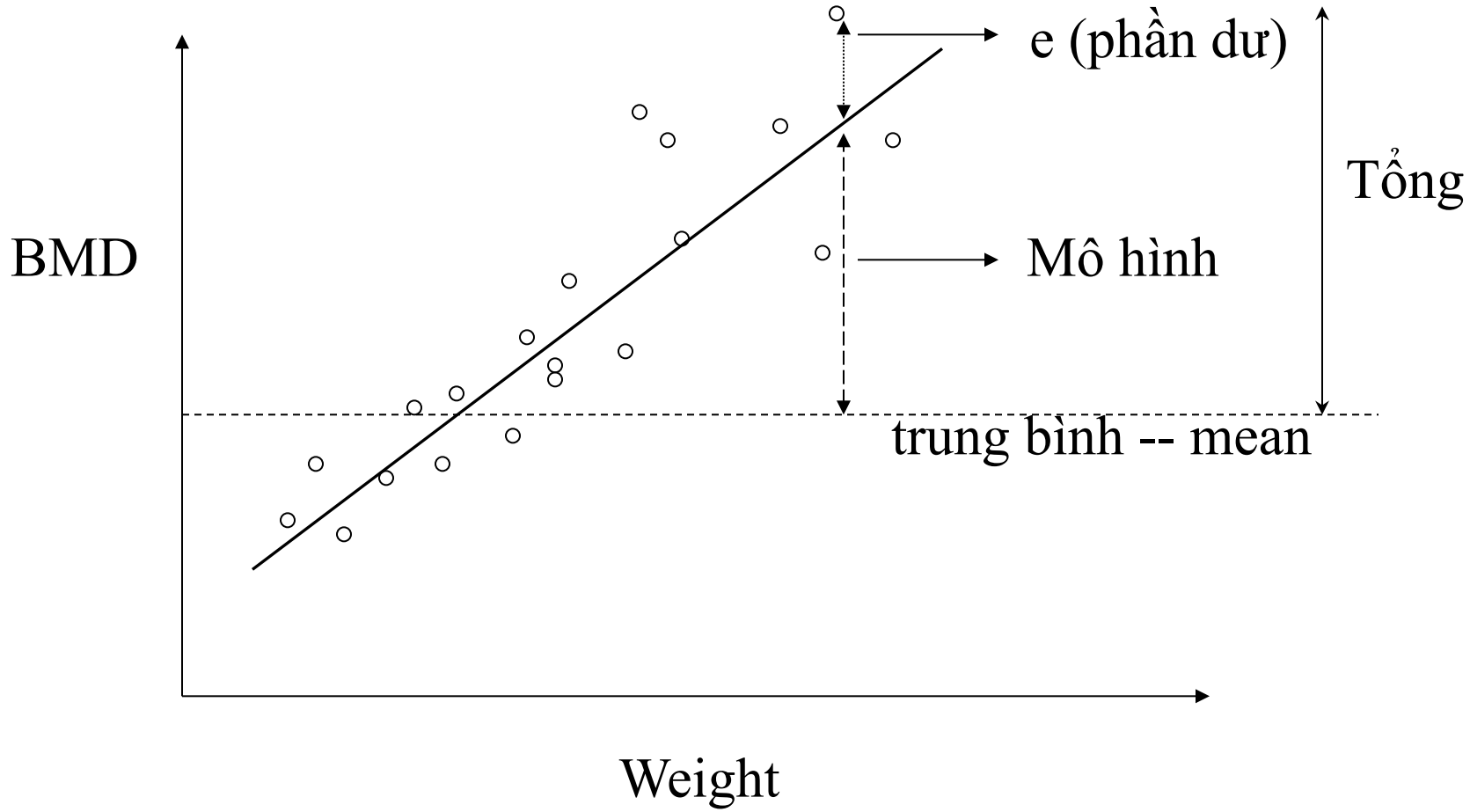
Phần dư : e = y – E(y)

- Phát biểu "thường dân"

Dữ liện quan sát = Mô hình tiên lượng + Phần dư (nhiễu)

Phần dư = giá trị quan sát – giá trị tiên lượng

Thể hiện qua hình học



$$SS_{total} = SS_{reg} + SS_{error}$$

Mục đích của phân tích độ dao động dư

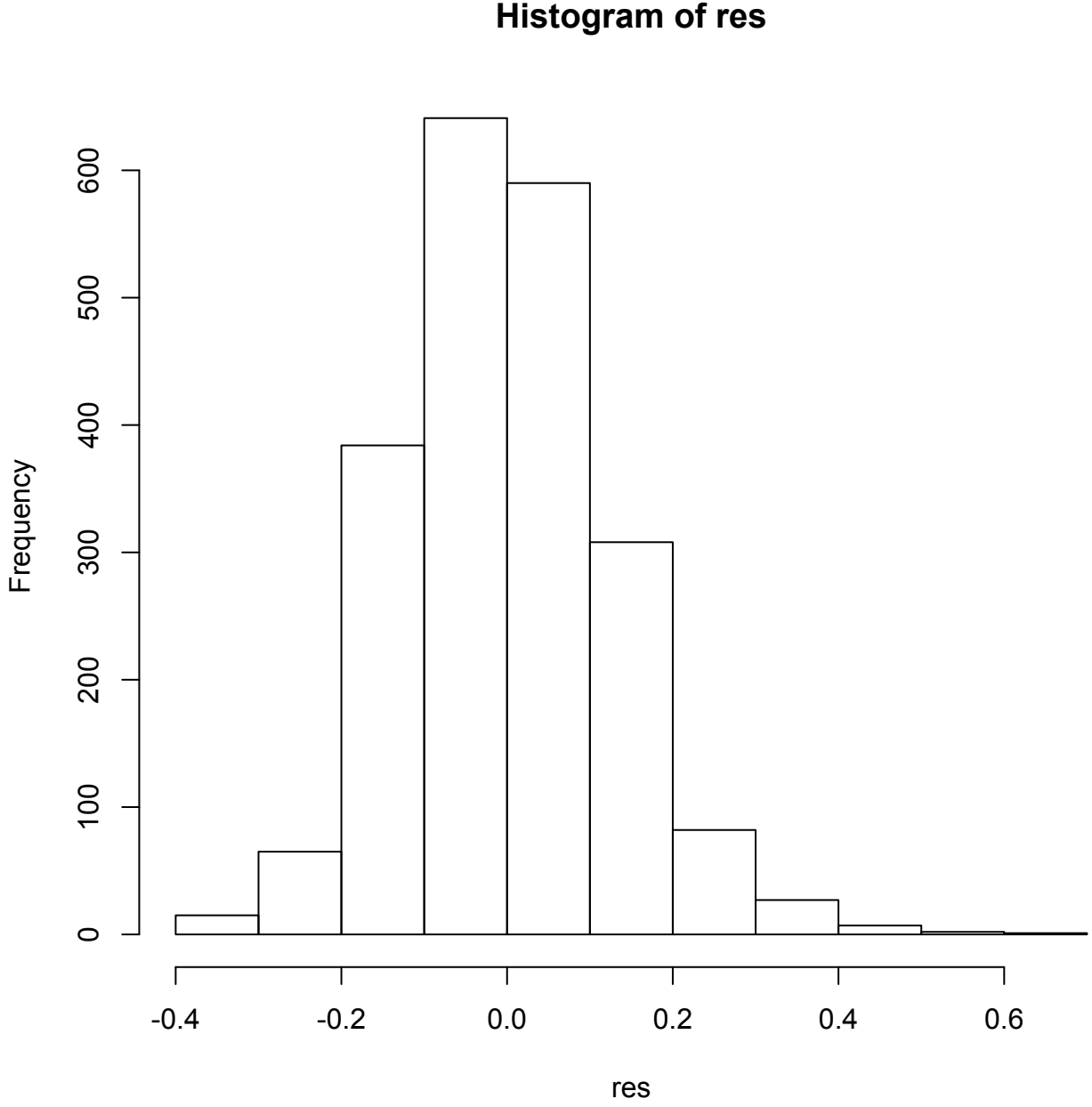
- Kiểm định phân bố chuẩn (normal distribution)
- Phương sai có phải bất biến với X ?
- Độc lập?
- Có giá trị nào là "ngoại vi" (outlier) hay có ảnh hưởng (influential observation)

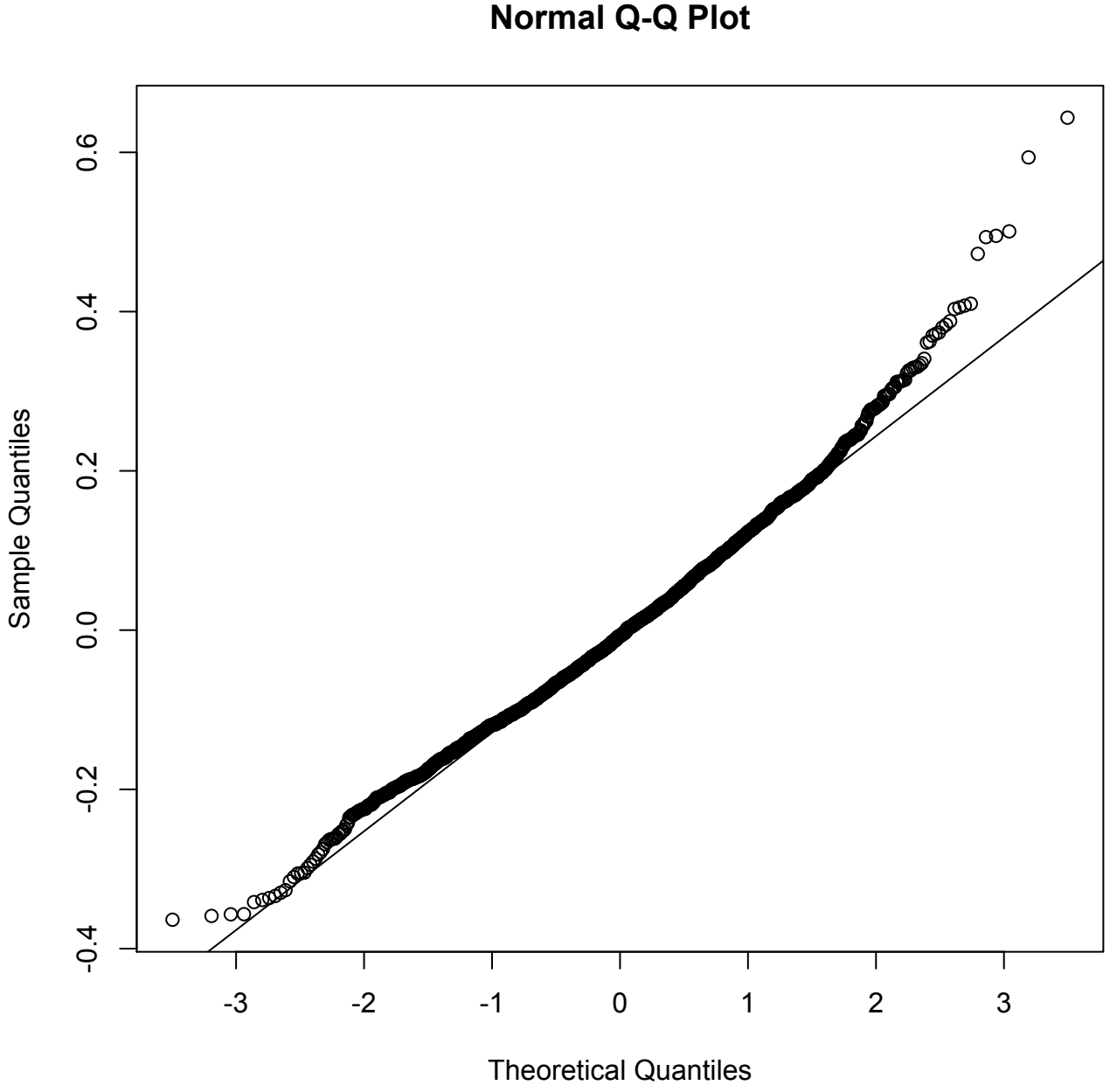
Ví dụ bằng R

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

attach(dat)

# Phân tích hồi qui tuyến tính

m1 = lm(fnbmd ~ wt)
```

Residual

```
m1 = lm(fnbmd ~ wt)
res = resid(m1)
hist(res)
qqnorm(res); qqline(res)
```

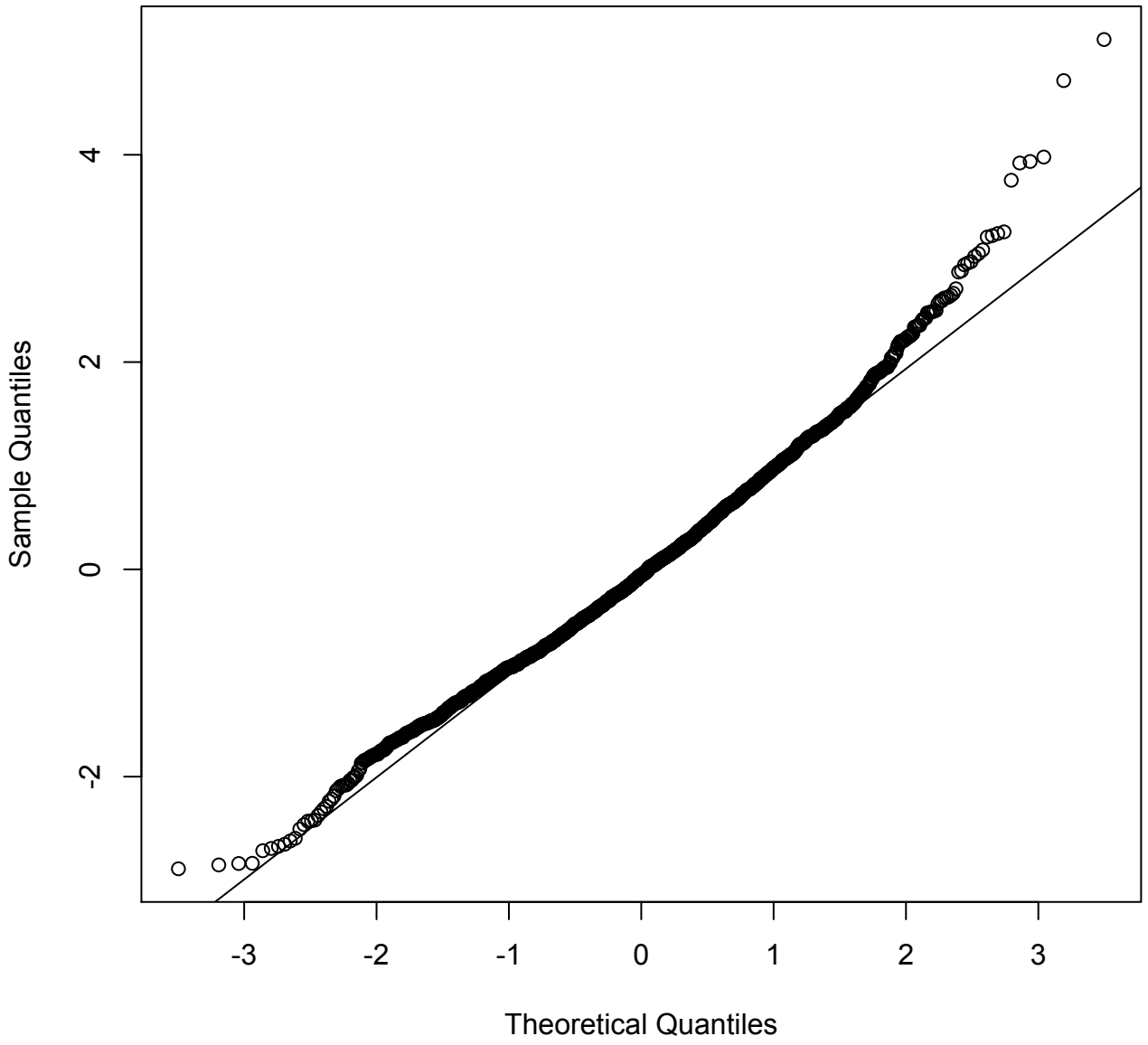

Histogram of res
Frequency
0
100
200
300
400
500
600
-0.4
-0.2
0.0
0.2
0.4
0.6
res


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-0.4
-0.2
0.0
0.2
0.4
0.6
-3
-2
-1
0
1
2
3

Standardized residuals

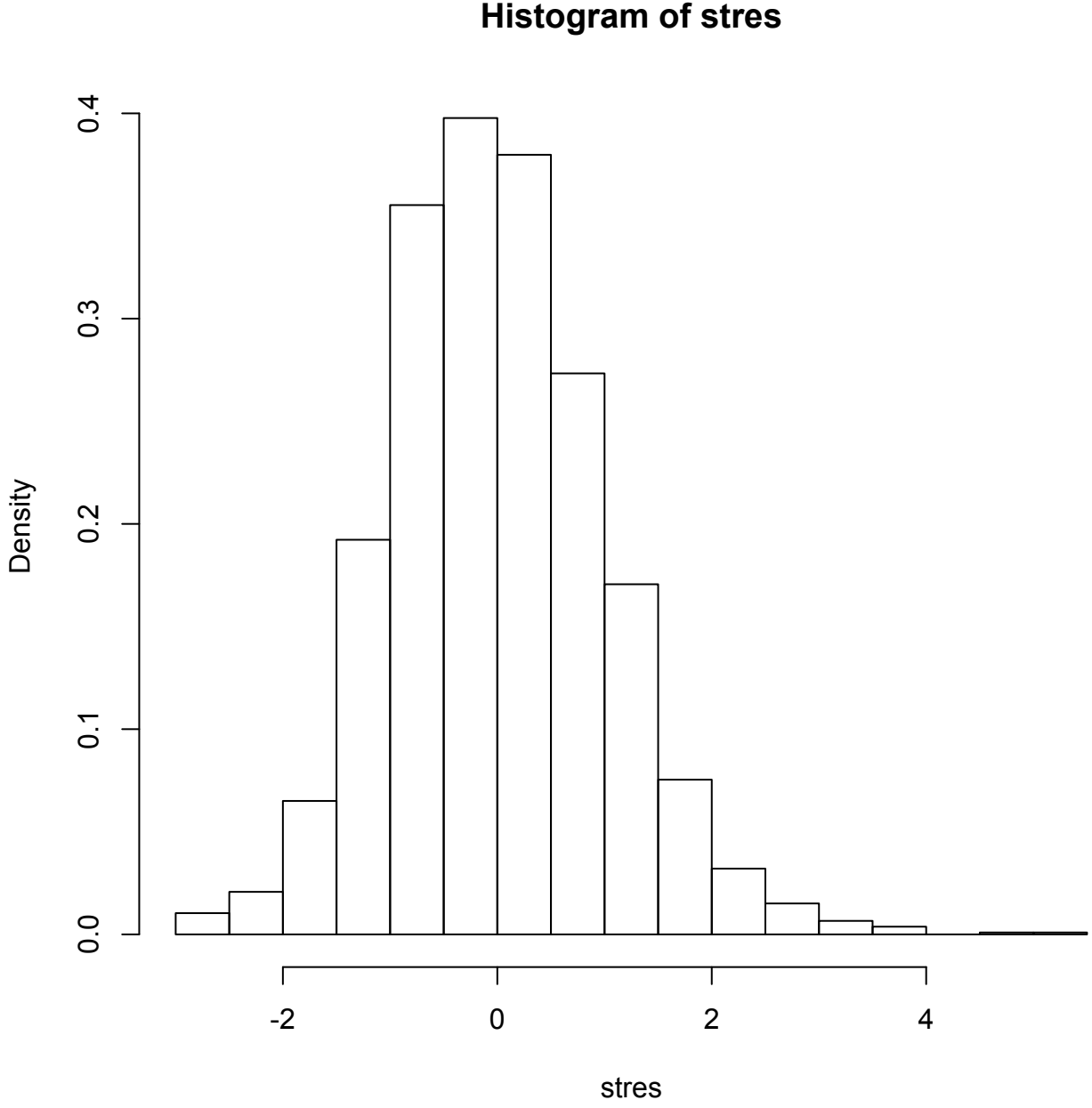
$$Standardized\ Residual\ i = \frac{Residual\ i}{Standard\ Deviation\ of\ Residual\ i}$$

```
m1 = lm(fnbmd ~ wt)
stdres = rstandard(m1)
hist(stdres)
qqnorm(stdres); qqline(stdres)
```


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-2
0
2
4
-3
-2
-1
0
1
2
3

Studentized residuals

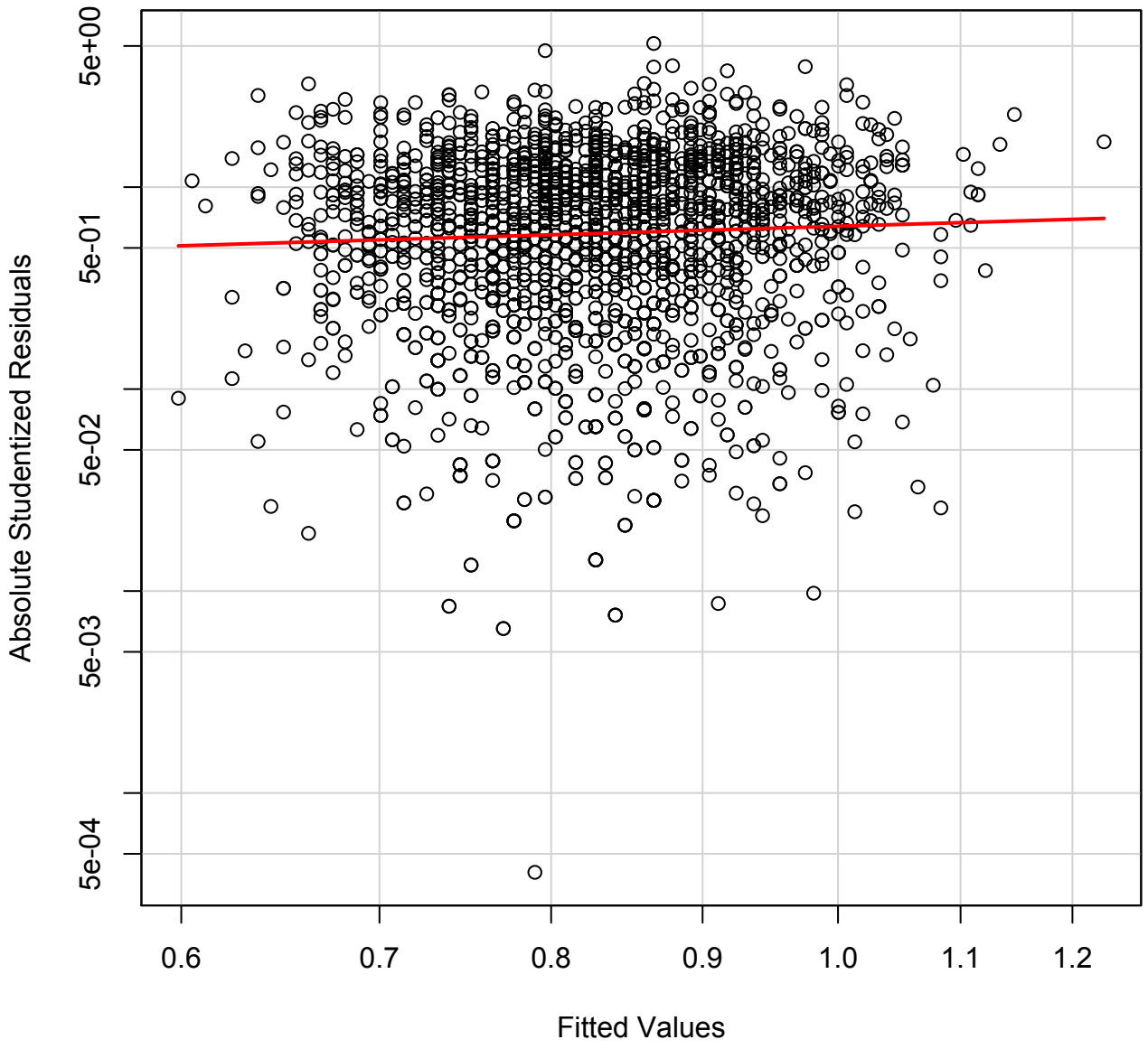
```
m1 = lm(fnbmd ~ wt)

library(MASS)

stres = studres(m1)

hist(stres, freq=FALSE)
```


Histogram of stres
Density
0.0
0.1
0.2
0.3
0.4
-2
0
2
4
stres

Kiểm tra phương sai

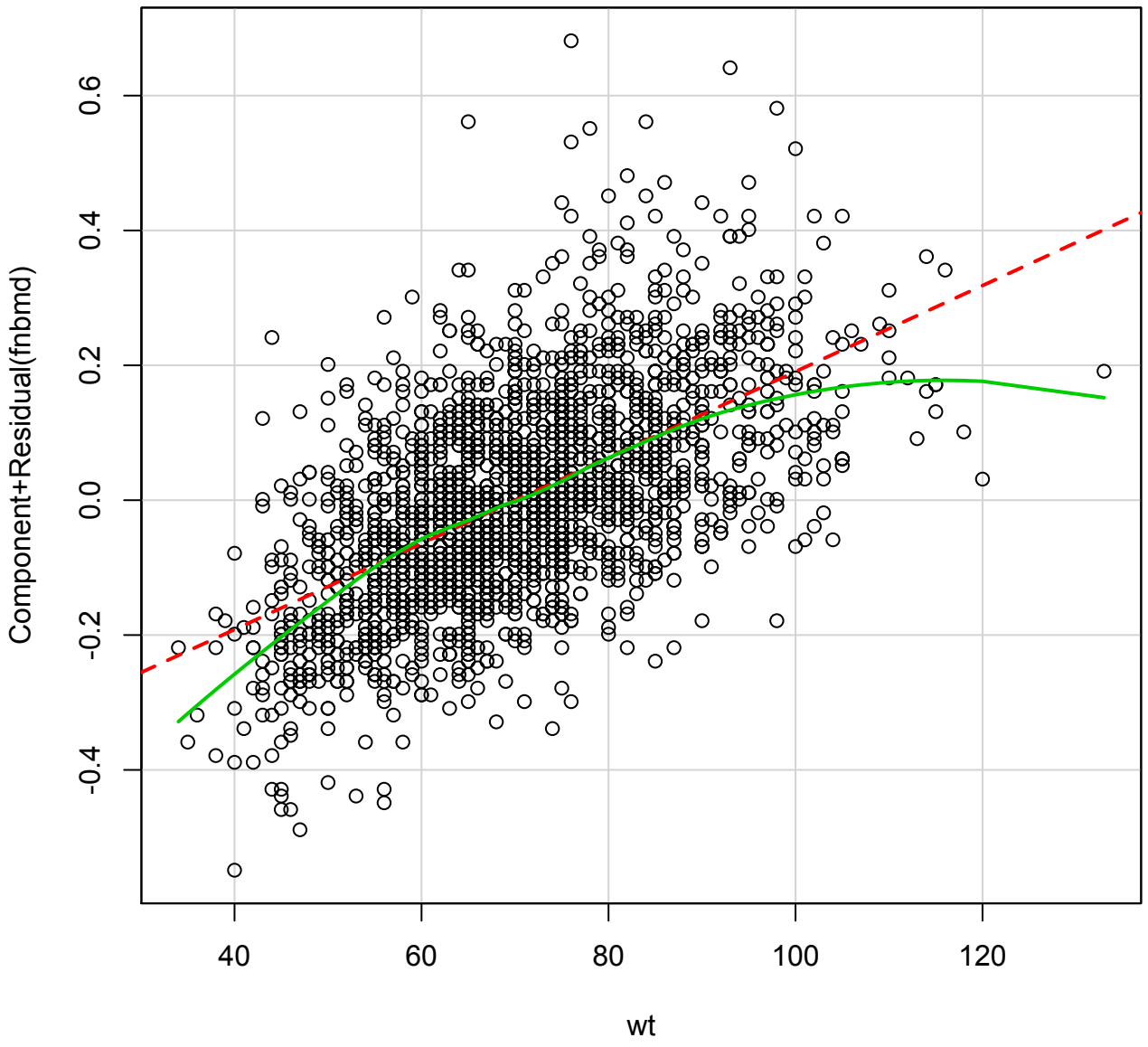
```
m1 = lm(fnbmd ~ wt)
library(car)
ncvTest(m1)
# studentized residuals và giá trị
tiên lượng
spreadLevelPlot(m1)
```

```
> ncvTest(m1)

Non-constant Variance Score Test

Variance formula: ~ fitted.values

Chisquare = 9.115887    Df = 1  p=0.002533991
```


Spread-Level Plot for
m1
Absolute Studentized Residuals
5e+00
5e-01
5e-02
5e-03
5e-04
0.6
0.7
0.8
0.9
1.0
1.1
1.2
Fitted Values

Kiểm tra non-linearity

```
m1 = lm(fnbmd ~ wt)
library(car)
# component + residual plot
crPlots(m1)
# Ceres plots
ceresPlots(m1)
```


Component+Residual(fnbmd)
0.6
0.4
0.2
0.0
-0.2
-0.4
40
60
80
100
120
wt

Kiểm tra independence

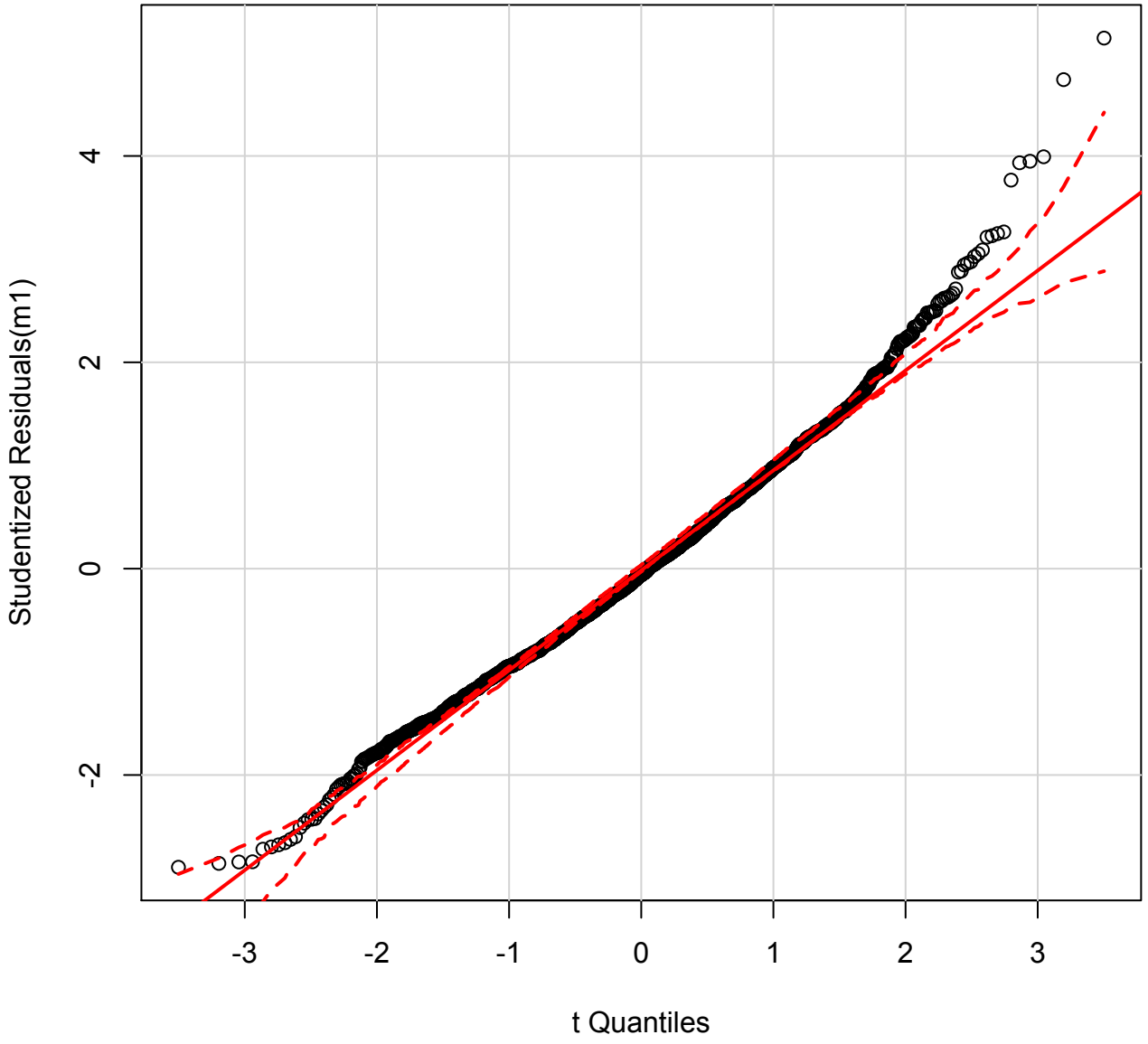
```
m1 = lm(fnbmd ~ wt)
library(car)
durbinWatsonTest(m1)
```

```
> durbinWatsonTest(m1)
 lag Autocorrelation D-W Statistic p-value
   1       0.03894983      1.921814    0.05
 Alternative hypothesis: rho != 0
```

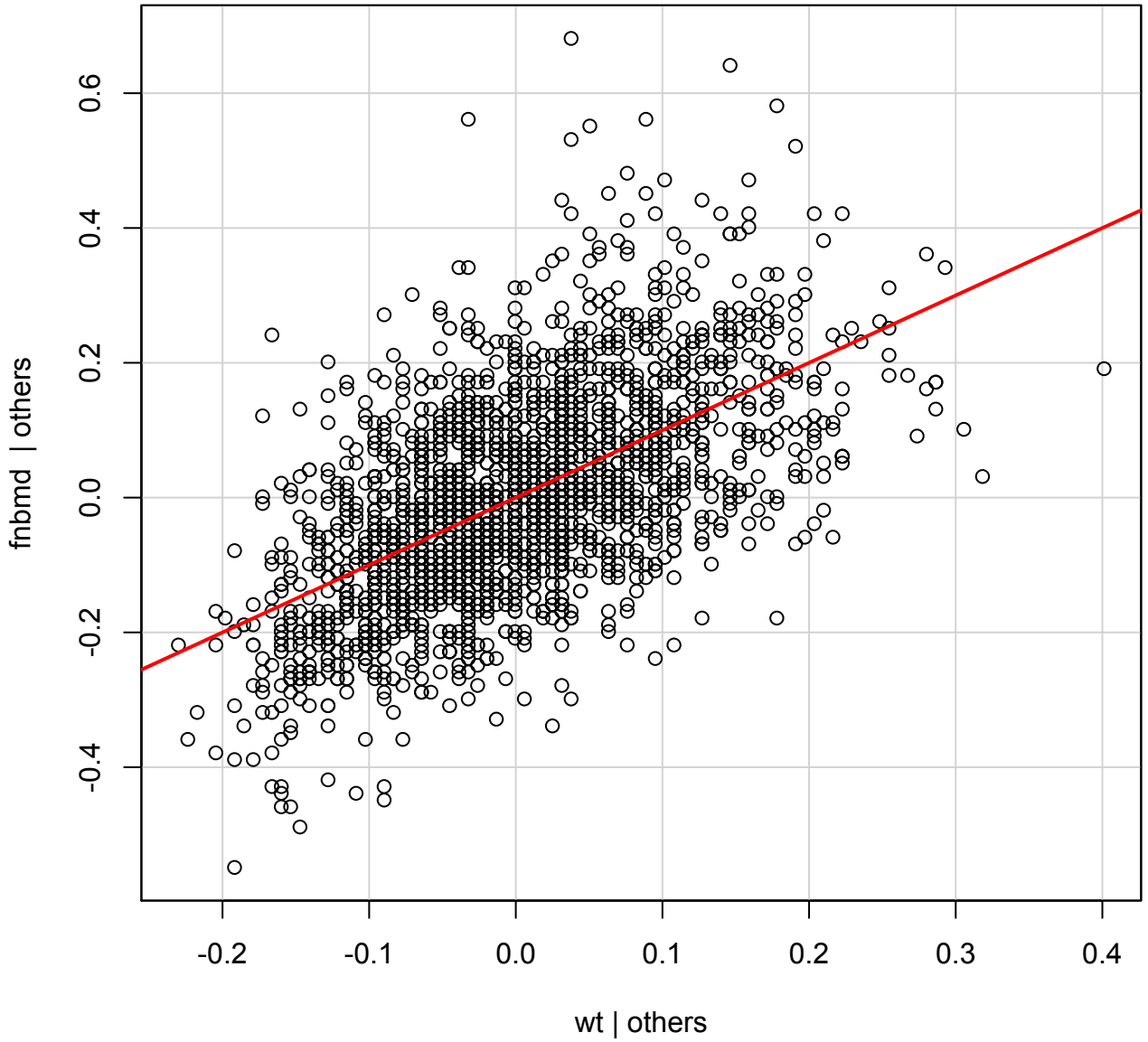
Outliers

```
m1 = lm(fnbmd ~ wt)
library(car)
outlierTest(m1)
qqPlot(m1)
leveragePlots(m1)
```

Qqplot



Leverage plot



Outliers

```
> outlierTest(m1)
      rstudent unadjusted p-value Bonferonni p
1558 5.142598          2.9588e-07   0.00062785
1813 4.739670          2.2829e-06   0.00484420
```

Influential observations

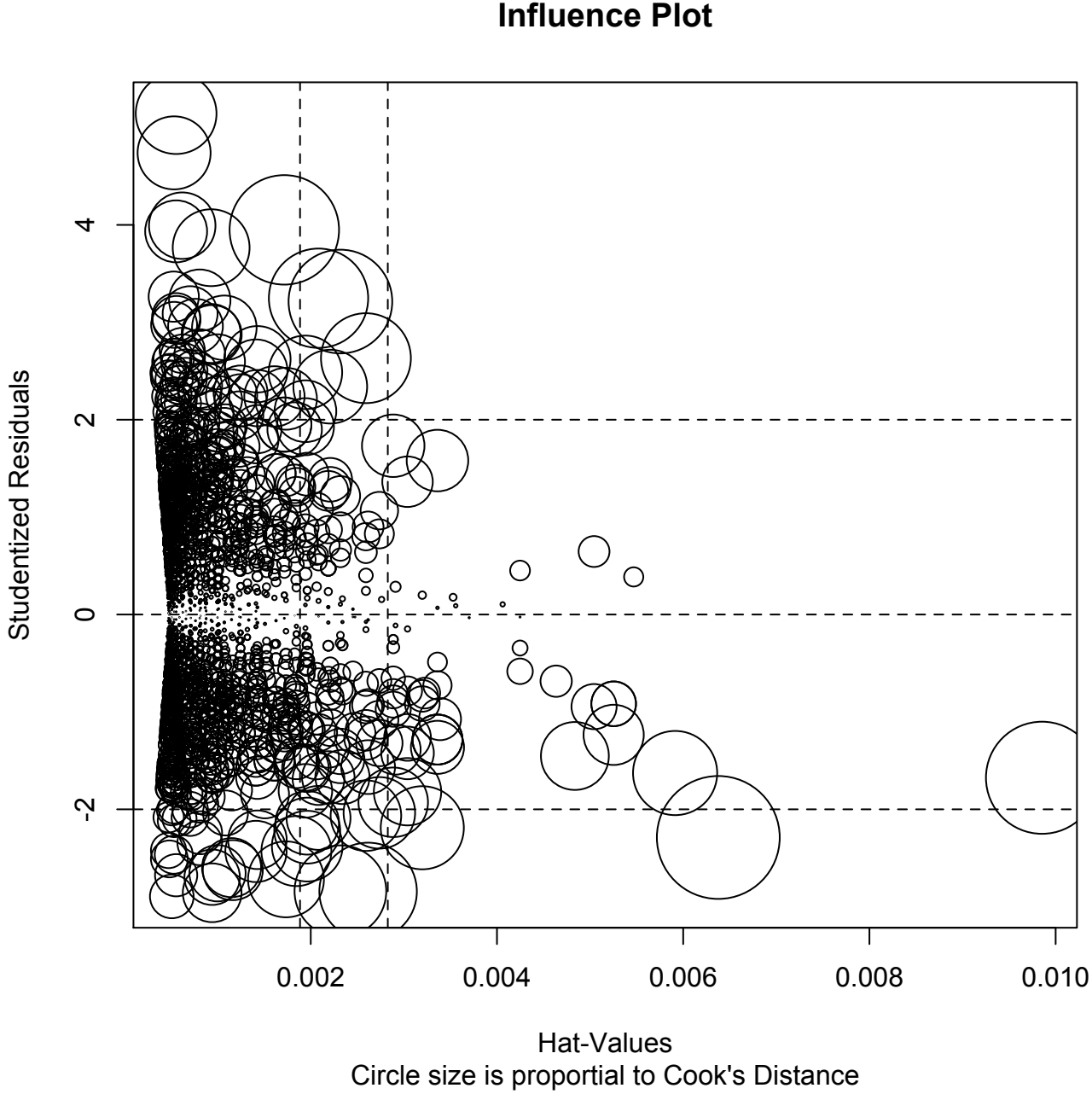
Cook's D plot

identify D values > 4/(n-k-1)

```
cutoff = 4/((nrow(dat)-length(m1$coefficients)-2))
plot(m1, which=4, cook.levels=cutoff)
```

Influence Plot

```
influencePlot(m1, id.method="identify",
main="Influence Plot", sub="Circle size is
proportial to Cook's Distance" )
```


Influence Plot
Studentized Residuals
4
2
0
-2
0.002
0.004
0.006
0.008
0.010
Hat-Values
Circle size is proportial to Cook's Distance

Tóm lược

- Residual = quan sát – tiên lượng
- Phân tích residual: kiểm định
  - phân bố chuẩn (normal distribution)?
  - Phương sai bất biến?
  - Độc lập?
  - Giá trị ngoại vi và ảnh hưởng