Bài 11: HỒI QUY - TƯƠNG QUAN

1. Mô hình hồi quy tuyến tính đơn

Mô hình hồi quy tuyến tính đơn cho các cặp dữ liệu (x_i, y_i) như sau

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

các hệ số β_0, β_1 chưa biết và sẽ được ước lượng từ dữ liệu.

Người ta dung phương pháp bình phương bé nhất để tìm các ước lượng này. Cụ thể

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})\sum(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Ví dụ 1: Nhịp tim tối đa

Nhịp tim tối đa của một người được cho rằng (theo kinh nghiệm) có mối quan hệ với tuổi tác theo phương trình sau

$$Max = 220 - Age$$

Người ta khảo sát nhịp tim tối đa của 15 người có độ tuổi khác nhau. Dữ liệu sau được ghi lại:

Age	18	23	25	35	65	54	34	56	72	19	23	42	18	39	37
Max Rate	202	186	187	180	156	169	174	172	153	199	193	174	198	183	178

Hãy biểu diễn dữ liệu trên và vẽ đường hồi quy đơn Max Rate theo Age trên cùng một đồ thị?

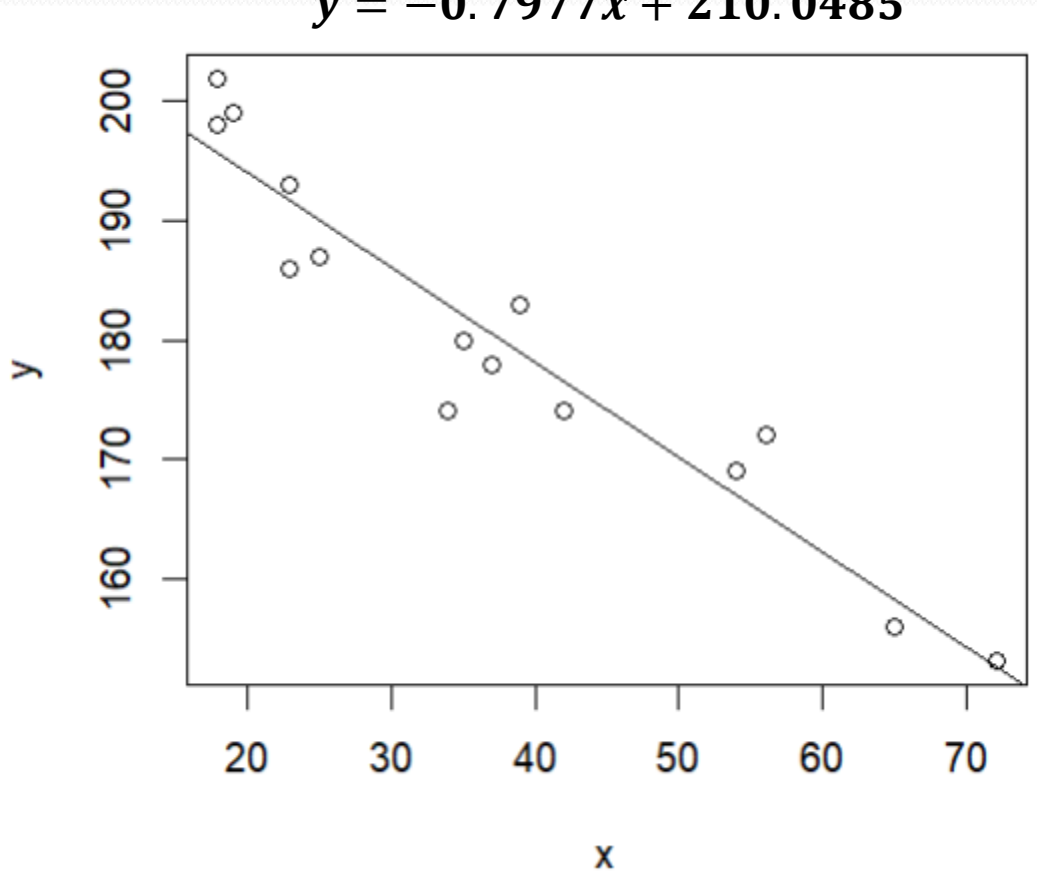
```
x <- c(18,23,25,35,65,54,34,56,72,19,23,42,18,39,37) #nhập dữ liệu
y <- c(202,186,187,180,156,169,174,172,153,199,193,174,198,183,178)


plot(x,y) # vẽ đồ thị

abline(lm(y ~ x)) # vẽ đường hồi quy
```

Ví dụ 1: Nhịp tim tối đa

```
x <- c(18,23,25,35,65,54,34,56,72,19,23,42,18,39,37) #nhập dữ liệu
y <- c(202,186,187,180,156,169,174,172,153,199,193,174,198,183,178)


plot(x,y) # vẽ đồ thị

abline(lm(y ~ x)) # vẽ đường hồi quy

lm(y ~ x) # các giá trị cơ bản của phân tích hồi quy

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
   210.0485      -0.7977
```

Ví dụ 1: Nhịp tim tối đa

$$\hat{y} = -0.7977x + 210.0485$$



Hình: Hồi quy nhịp tim tối đa (Max rate) theo tuổi (Age)

Ví dụ 1: Nhịp tim tối đa

```
result = lm(y ~ x)
summary(result)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9258 -2.5383  0.3879  3.1867  6.6242

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 210.04846    2.86694   73.27  < 2e-16 ***
x            -0.79773    0.06996  -11.40 3.85e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.578 on 13 degrees of freedom
Multiple R-squared:  0.9091, Adjusted R-squared:  0.9021
F-statistic:   130 on 1 and 13 DF,  p-value: 3.848e-08
```

Ví dụ 1: Nhịp tim tối đa

Nếu muốn lấy thông tin về thặng dư (residuals) ta dùng **resid**, về hệ số (coefficients) ta dùng **coef**. Ví dụ

```
coef(result) # hoặc sử dụng result[['coefficients']]

(Intercept)           x
210.0484584  -0.7977266


res = resid(result) # hoặc result[['residual']]
summary(res)

 Min.   1st Qu.  Median    Mean 3rd Qu.    Max.
-8.9258 -2.5383  0.3879  0.0000  3.1867  6.6242
```

2. Phân tích thặng dư

Ta dựa vào đồ thị biểu diễn thặng dư theo giá trị hồi quy ($\hat{y}$) và đồ thị Normal Q-Q.

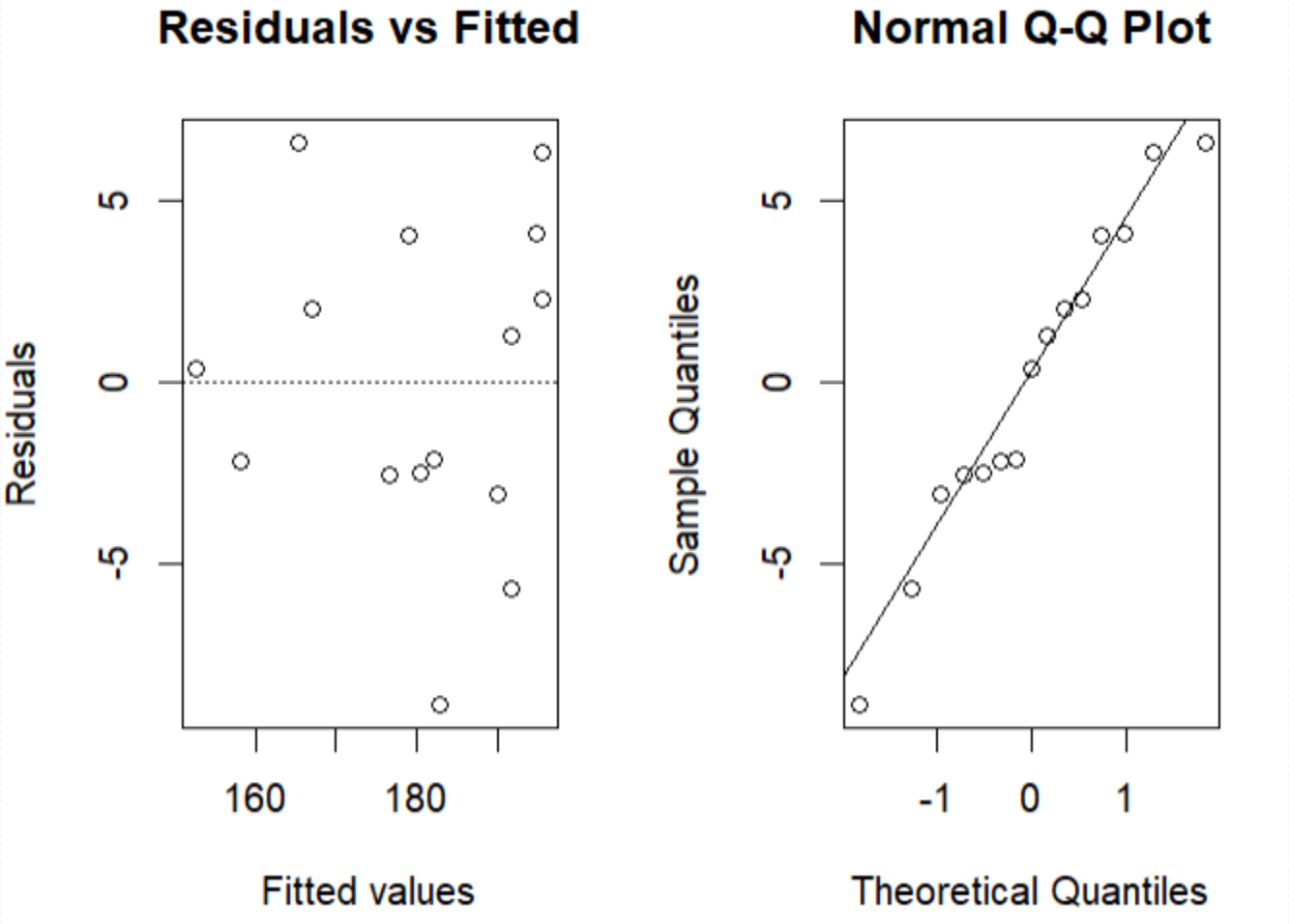
```
coef(result) # hoặc sử dụng result[['coefficients']]

(Intercept)           x
210.0484584  -0.7977266


res = resid(result) # hoặc result[['residual']]
summary(res)

 Min.   1st Qu.  Median    Mean 3rd Qu.    Max.
-8.9258 -2.5383  0.3879  0.0000  3.1867  6.6242
```

Ví dụ 2: Nhịp tim tối đa

Hãy phân tích giá trị thặng dư của mô hình hồi quy.

```
x <- c(18,23,25,35,65,54,34,56,72,19,23,42,18,39,37) #nhập dữ liệu
y <- c(202,186,187,180,156,169,174,172,153,199,193,174,198,183,178)

result = lm(y ~ x)

par(mfrow=c(1,2)) #chuẩn bị vẽ hai đồ thị trên 1 cửa sổ

plot(result$fitted.values,resid(result),xlab ='Fitted values',
ylab ='Residuals', main = 'Residuals vs Fitted')

#đồ thị thặng dư theo giá trị hồi quy

abline(h=0,lty=3) #đường thẳng y = 0 với nét chấm

qqnorm(res) #đồ thị Normal Q-Q

qqline(res) #đường thẳng lí thuyết trên đồ thị Normal Q-Q
```

Ví dụ 2: Nhịp tim tối đa



Hình: Phân tích thặng dư.

Ví dụ 2: Nhịp tim tối đa

Nhìn vào đồ thị **Residuals vs Fitted** ta thấy thặng dư phân tán quanh trục Ox một cách ngẫu nhiên đồng đều. Do đó thặng dư có kì vọng 0 và phương sai không đổi.

Nhìn vào đồ thị **Normal Q-Q Plot** ta thấy thặng dư gần xấp xỉ đường thẳng. Do đó thặng dư tuân theo phân phối chuẩn.

Bài tập

Giá một căn nhà (đv: 1000 USD) phụ thuộc vào số phòng ngủ trong căn nhà đó. Giả sử rằng dữ liệu sau được ghi lại cho các căn nhà ở một thành phố.

Price	300	250	400	550	317	389	425	289	389	559
No. bedrooms	3	3	4	5	4	3	6	3	4	5

(a) Vẽ đồ thị phân tán và đường hồi quy trên cùng một hệ trục tọa độ.

```
# Bai tap 1
# a) Nhap du lieu
P <- c(300,250,400,550,317,389,425,289,389,559)
NB <- c(3,3,4,5,4,3,6,3,4,5)

# Ve bieu do phan tan

plot(NB,P)

abline(lm(P ~ NB)) # ve duong hoi quy
```

Bài tập

Giá một căn nhà (đv: 1000 USD) phụ thuộc vào số phòng ngủ trong căn nhà đó. Giả sử rằng dữ liệu sau được ghi lại cho các căn nhà ở một thành phố.

Price	300	250	400	550	317	389	425	289	389	559
No. bedrooms	3	3	4	5	4	3	6	3	4	5

(b) Kiểm định giả thuyết cho rằng khi thêm một phòng ngủ thì chi phí tăng thêm 60.000 USD với đối thuyết là chi phí cao hơn.

```
# b) Hoi quy
lm(P ~ NB)

Call:
lm(formula = P ~ NB)

Coefficients:
(Intercept)           NB
       94.4         73.1
```