Bài giảng 20: Hồi qui tuyến tính đơn giản: Hiểu kết quả phân tích

Nguyễn Văn Tuấn

Garvan Institute of Medical Research, Australia

Đại học Tôn Đức Thắng, Việt Nam

Mục tiêu

- Mô hình

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

- Ước tính 2 tham số α và β
- Dùng dữ liệu thí nghiệm / thực tế

Ước tính bằng R

- Chúng ta muốn ước tính mối liên quan giữa BMD và trọng lượng
- Mô hình hồi qui tuyến tính:

$$BMD = \alpha + \beta * weight + \varepsilon$$

Dùng R

```
lm(bmd ~ weight)
```

Phân tích 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)
summary(m1)
anova(m1)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.36371 -0.08817 -0.00733  0.07918  0.64354

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.3818873  0.0138582   27.56   <2e-16 ***
wt          0.0063760  0.0001939   32.89   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

Residual standard error: 0.1259 on 2120 degrees of
freedom (94 observations deleted due to missingness)
Multiple R-squared:  0.3379,  Adjusted R-squared:  0.3375
F-statistic:  1082 on 1 and 2120 DF,  p-value: < 2.2e-16
```

Diễn giải kết quả

```
Residual standard error: 0.1259 on 2120 degrees of freedom

Multiple R-squared:  0.3379,     Adjusted R-squared:  0.3375

F-statistic:  1082 on 1 and 2120 DF,  p-value: < 2.2e-16
```

Ý nghĩa của:

- R squared
- Adjusted R squared
- F-test

Câu hỏi

- Mô hình này "tốt" hay "xấu"
- Tiêu chí để định nghĩa là "Tốt"
- Tốt có nghĩa là mô hình phản ảnh giá trị quan sát
  - Giá trị tiên lượng (predicted values) gần với giá trị quan sát (observed values)

Phân tích phương sai

Phân tích phương sai

- BMD = a + b*weight + e
- Observed variation = model + random

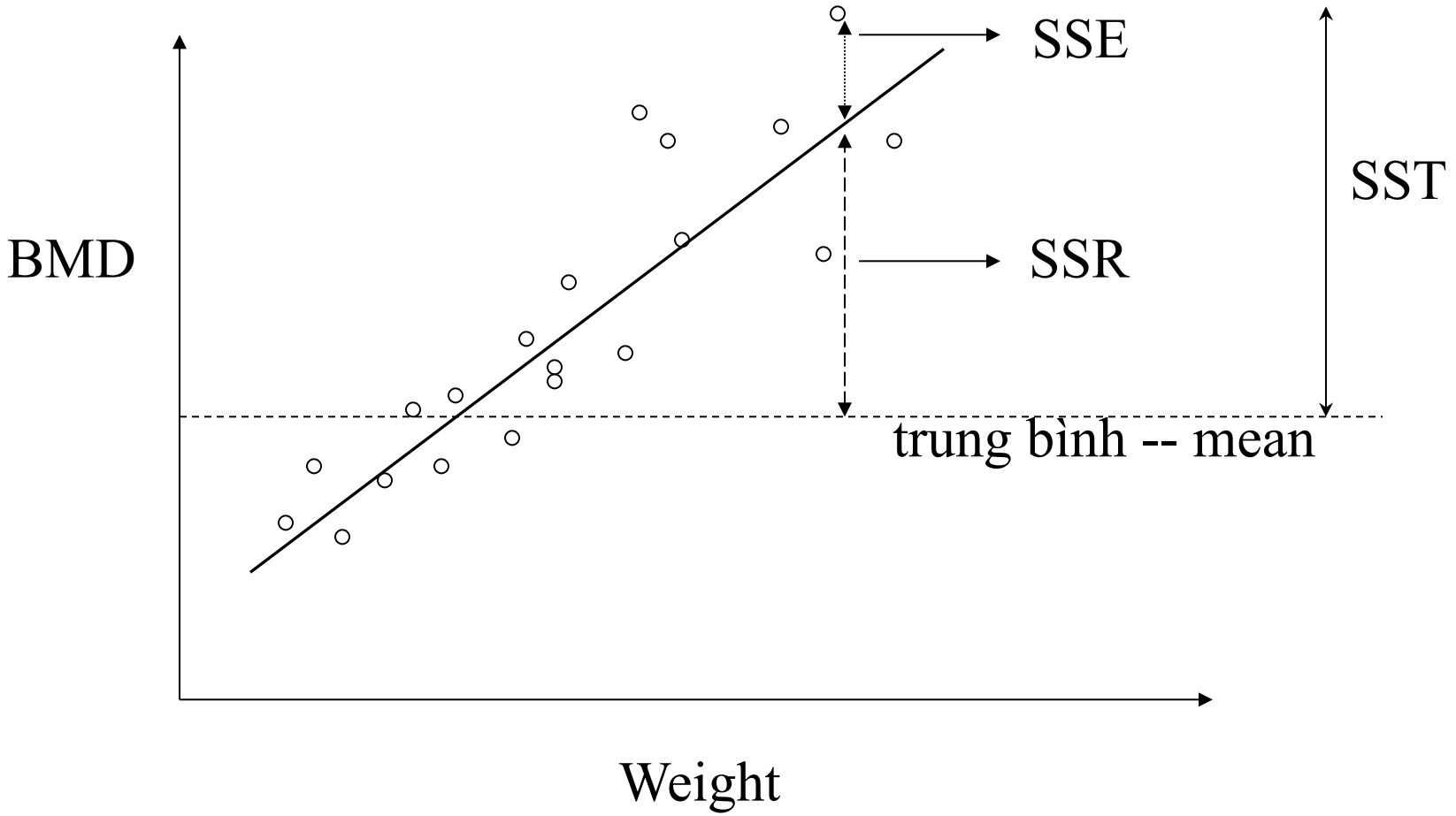
"Variation" = sum of squares

- SS_{total} = total sum of squares

  SS_{reg} = sum of squares due to the regression model

  SS_{error} = sum of squares due to random component

Thể hiện qua hình học



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

Phân định nguồn phương sai

```
> m1 = lm(fnbmd ~ wt)
> anova(m1)
```

```
Analysis of Variance Table

Response: fnbmd
             Df Sum Sq Mean Sq F value    Pr(>F)
wt            1 17.153 17.1528  1081.7 < 2.2e-16 ***
Residuals 2120 33.616  0.0159
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Total SS = 17.15 + 33.62 = **50.77**
  - Do weight: 17.15
  - Phần chưa giải thích được (residuals): 33.62

Hệ số xác định R^2

```
> m1 = lm(fnbmd ~ wt)
> anova(m1)
```

```
Analysis of Variance Table

Response: fnbmd
             Df Sum Sq Mean Sq F value    Pr(>F)
wt            1 17.153 17.1528  1081.7 < 2.2e-16 ***
Residuals  2120 33.616  0.0159
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Total SS = 17.15 + 33.62 = 50.77
- R^2 = 17.15 / 50.77 = 0.34

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.36371 -0.08817 -0.00733  0.07918  0.64354

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.3818873  0.0138582   27.56   <2e-16 ***
wt          0.0063760  0.0001939   32.89   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

Residual standard error: 0.1259 on 2120 degrees of
freedom (94 observations deleted due to missingness)
Multiple R-squared:  0.3379,  Adjusted R-squared:  0.3375
F-statistic:  1082 on 1 and 2120 DF,  p-value: < 2.2e-16
```

Ý nghĩa của R^2

```
Residual standard error: 0.1259 on 2120 degrees of freedom (94
observations deleted due to missingness)
Multiple R-squared:  0.3379,     Adjusted R-squared:  0.3375
```

- Coefficient of determination $R^2 = 0.34$
- Diễn giải: **Approximately 34% of BMD variance could be accounted for by body weight**
- **Những khác biệt về trọng lượng giải thích khoảng 34% những khác biệt về mật độ xương**

Hệ số xác định điều chỉnh (adjusted R^2)

- Định nghĩa dễ hiểu nhất:

$$\mathbf{R^2_{adj} = 1 - (MS_{error} / MS_{total})}$$

MS_{error} : mean square due to error

MS_{total} : mean square (total)

Hệ số xác định điều chỉnh (adjusted R^2)

```
> m1 = lm(fnbmd ~ wt)
> anova(m1)
```

```
Analysis of Variance Table

Response: fnbmd
             Df Sum Sq Mean Sq F value    Pr(>F)
wt            1 17.153 17.1528  1081.7 < 2.2e-16 ***
Residuals  2120 33.616  0.0159
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- MS_{total} = (17.15 + 33.62) / 2121 = 0.0239
- MS_{error} = 0.0159
- R^2_{adj} = 1 – (0.0159 / 0.0239) = 0.337

Phương sai của BMD sau khi điều chỉnh cho wt

```
> m1 = lm(fnbmd ~ wt)
> anove(m1)
```

```
Analysis of Variance Table

Response: fnbmd
             Df Sum Sq Mean Sq F value    Pr(>F)
wt            1 17.153 17.1528  1081.7 < 2.2e-16 ***
Residuals 2120 33.616  0.0159
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Mean square (MS) = sum of squares / (degrees of freedom)
- MS(residuals) = 33.616 / 2120 = 0.0159

Phương sai của BMD sau khi đã điều chỉnh cho trọng lượng là 0.0159

(trước khi điều chỉnh là 0.0239)

Tiên lượng tốt hơn

- Mô hình là: BMD = 0.38 + 0.0064*weight
- Nếu không biết trọng lượng, BMD trung bình là 0.83 g/cm^2
- Nếu biết trọng lượng, chúng ta tiên lượng tốt hơn:
- Weight = 50 kg, BMD = 0.38 + 0.0064*50 = 0.70 g/cm^2

  Weight = 60 kg, BMD = 0.38 + 0.0064*60 = 0.76 g/cm^2

Tóm tắt

- Phân tích phương sai là một phần của phân tích hồi qui tuyến tính
- Tổng bình phương = mô hình + ngẫu nhiên

$$SS_{total} = SS_{model} + SS_{random}$$

- Hệ số xác định

$$R^2 = SS_{model} / SS_{total}$$