Mô hình hồi qui tuyến tính:
đánh giá mô hình

Tuan V. Nguyen

Garvan Institute of Medical Research

University of New South Wales (UNSW Sydney), Australia

University of Technology, Sydney (UTS), Australia

Ton Duc Thang University, Vietnam

Đánh giá mô hình hồi qui tuyến tính

- Khái niệm 'residual' và phương sai
- RMSE – residual mean squared error
- Hệ số xác định (coefficient of determination)

Cái nhìn tổng thể

Những gì chúng ta quan sát (đo lường được) là kết quả của hiện tượng thật (bản chất, hệ thống) và sai sót ngẫu nhiên

Observed = Systematic + Error

Systematic = mô hình

Mô hình hồi qui tuyến tính và phần dư

- Y_i là giá trị quan sát cho cá nhân *i*
- Mô hình cho Y_i

$$Y_i = a + bX_i + e_i$$

$$Y_i = \hat{Y} + e_i$$

Nói cách khác (bỏ *i*)

Giá trị trung bình (kì vọng): $E(Y) = \hat{Y} = a + bX$

Phần dư : $e = Y - \hat{Y}$

Chiều cao của con (Y) và cha mẹ (X)

- Mô hình tiên lượng

$$\widehat{\text{Child}} = 23.94 + 0.646*\text{Parent}$$

- Dao động dư (residual)

$$e = \text{Child} - \widehat{\text{Child}}$$

tính cho mỗi đối tượng

Residual và predicted values dùng R

```
m = lm(child ~ parent, data=galton)
res = resid(m)
pred = predict(m)

> cbind(parent, child, pred, res)
    parent child     pred          res
1     70.5  61.7 69.50502 -7.80501621
2     68.5  61.7 68.21244 -6.51243505
3     65.5  61.7 66.27356 -4.57356330
4     64.5  61.7 65.62727 -3.92727272
5     64.0  61.7 65.30413 -3.60412743
6     67.5  62.2 67.56614 -5.36614446
7     67.5  62.2 67.56614 -5.36614446
8     67.5  62.2 67.56614 -5.36614446
9     66.5  62.2 66.91985 -4.71985388
10    66.5  62.2 66.91985 -4.71985388
11    66.5  62.2 66.91985 -4.71985388
```

Phân tích phương sai (analysis of variance)

- Child = a + b*Parent + e
- Observed variation = model + random

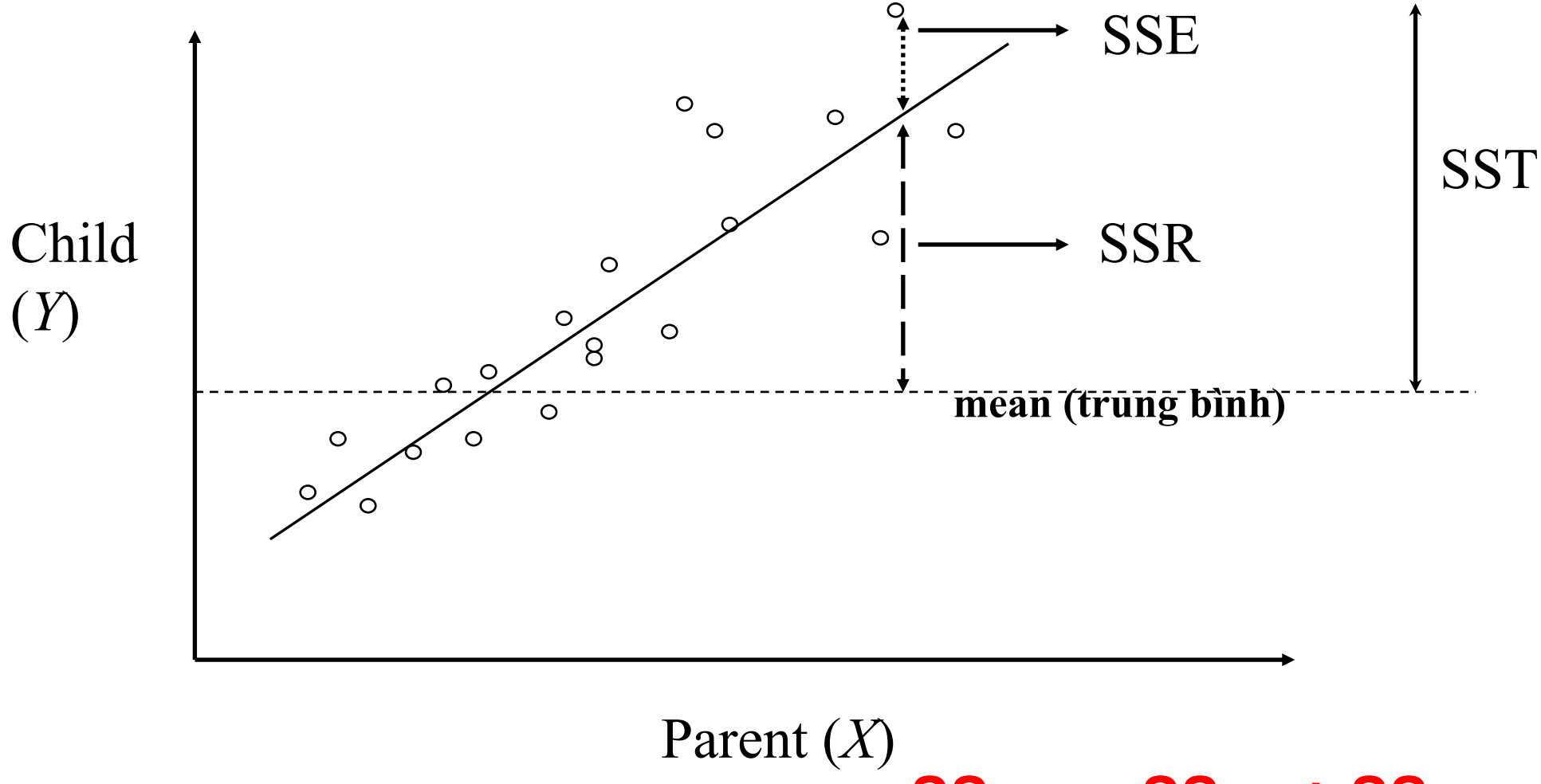
"Variation" = sum of squares

- 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 phân tích phương sai



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

Phân tích nguồn của sum of squares

```
m = lm(child ~ parent, data=galton)
anova(m)
```

```
Analysis of Variance Table

Response: child
            Df Sum Sq Mean Sq F value    Pr(>F)
parent       1 1236.9 1236.93  246.84 < 2.2e-16 ***
Residuals  926 4640.3    5.01
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Total SS = 1237 + 4640 = **5877**
  - Do "parent": 1237
  - Do residuals: 4640

Hệ số xác định - coefficient of determination (R^2)

```
m = lm(child ~ parent, data=galton)
anova(m)
```

```
Analysis of Variance Table

Response: child
            Df Sum Sq Mean Sq F value    Pr(>F)
parent       1 1236.9 1236.93  246.84 < 2.2e-16 ***
Residuals  926 4640.3    5.01
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Total SS = 1237 + 4640 = **5877**

R^2 = 1237 / 5877 = 0.21

```
summary(m)


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.94153    2.81088   8.517   <2e-16 ***
parent       0.64629    0.04114  15.711   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.239 on 926 degrees of freedom
Multiple R-squared:  0.2105,  Adjusted R-squared:  0.2096
F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

Diễn giải: Approximately 21% of childen's height variance could be accounted for by parental height

Adjusted R^2

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

MS_{error} : mean square due to error

MS_{total} : mean square (total)

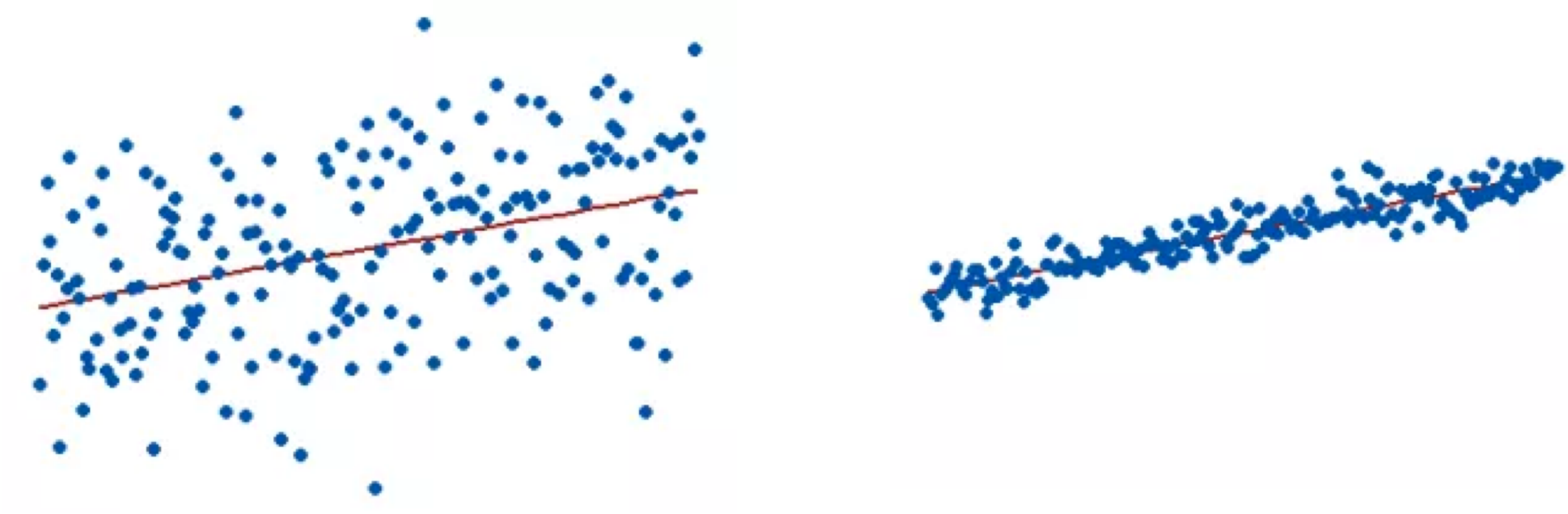
```
Analysis of Variance Table

Response: child
           Df Sum Sq Mean Sq F value    Pr(>F)
parent      1 1236.9 1236.93  246.84 < 2.2e-16 ***
Residuals 926 4640.3    5.01
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

MS_{total} = (1237 + 4640) / 927 = 6.34

MS_{error} = 5.01

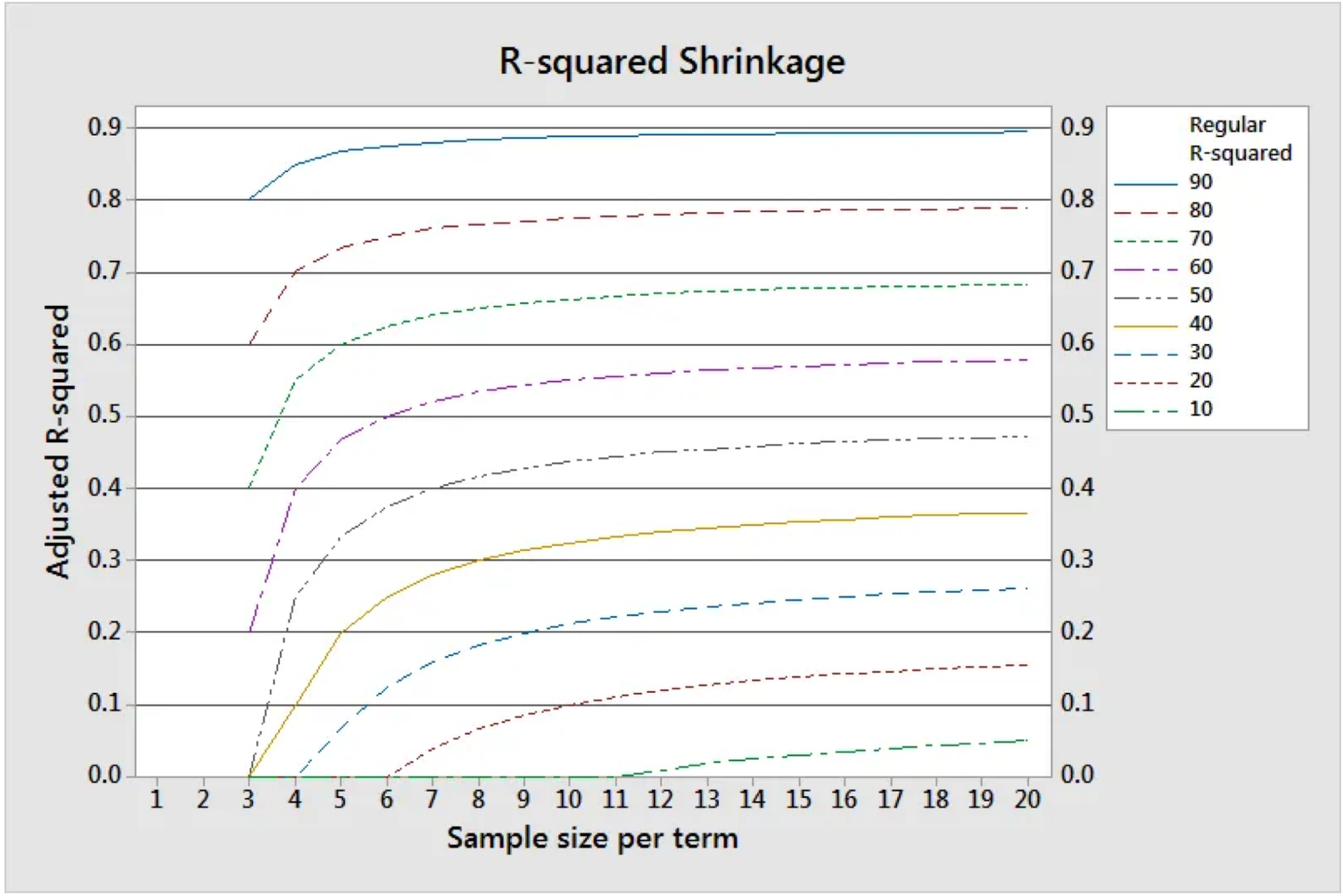
R^2_{adj} = 1 – (5.01 / 6.34) = 0.21

Cẩn thận với R^2



Hai mô hình có cùng slope, nhưng rất khác R^2. Mô hình bên phải có $R^2 = 0.15$, mô hình bên phải có $R^2 = 0.85$

Cẩn thận với R^2



https://statisticsbyjim.com

RMSE là gì?

- RMSE = root mean square error
- MSE là phương sai của Y sau khi đã hiệu chỉnh cho X
- RMSE là độ lệch chuẩn của Y sau khi đã hiệu chỉnh cho X

Phương sai của chiều cao con TRƯỚC khi hiệu chỉnh

```
> var(galton$child)
[1] 6.340029
```

Phương sai của chiều cao con SAU khi hiệu chỉnh

```
anova(m)

Response: child
           Df  Sum Sq Mean Sq F value    Pr(>F)
parent      1  1236.9 1236.93  246.84 < 2.2e-16 ***
Residuals 926  4640.3    5.01
```

Đánh giá mô hình hồi qui tuyến tính

- RMSE: phản ảnh độ lệch chuẩn của Y sau khi đã hiệu chỉnh cho mô hình hồi qui tuyến tính
- R^2 (hệ số xác định): tỉ lệ mà mô hình có thể giải thích được những khác biệt về *Y* giữa các đối tượng