Thiết kế thí nghiệm hoàn toàn ngẫu nhiên

Đỗ Trọng Hợp

Khoa Khoa Học và Kỹ Thuật Thông Tin

Đại Học Công Nghệ Thông Tin TP. Hồ Chí Minh

Dose-response modeling

- Thí nghiệm tỉ lệ của gan chuột so với cơ thể sau khi dùng 4 loại thức ăn

food	type 1	type 2	type 3	type 4
weight ratio	3.75	3.58	3.60	3.92
n_i	7	8	6	8

- Thí nghiệm ảnh hưởng của mưa axit đến mầm cây bạch dương

pH	4.7	4.0	3.3	3.0	2.3
weight	.337	.296	.320	.298	.177
n	48	48	48	48	48

- Khi yếu tố chính là biến định lượng, ta gọi các mức (level) của treatment là liều lượng (dose).

- Các kết quả **trung bình** của mỗi treatment có thể được biểu diễn dưới dạng hàm số của dose x_i:

$$\mu_i = \mu + \alpha_i = f(x_i, \beta_i)$$

trong đó μ là trung bình toàn bộ dân số, μ_i là trung bình nhóm i, **α_i là ảnh hưởng của nhóm i**, và β_i là các hệ số (chưa xác định) của hàm f.

Hồi quy đa thức (polynomial regression)

- Dạng chung

$$\mu_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_{k-1} x_i^{k-1}$$ với k là số nhóm

- Ví dụ các dạng thường dùng:

  - Dạng hằng số (kết quả không phụ thuộc vào treatment)

$$\mu_i = \beta_0$$

  - Dạng đường thẳng (kết quả phụ thuộc vào treatment theo phương trình đường thẳng)

$$\mu_i = \beta_0 + \beta_1 x_i$$

  - Dạng đường cong bậc 2 (kết quả phụ thuộc vào treatment theo phương trình bậc 2)

$$\mu_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

Linear regression

- Giả sử tồn tại một mối quan hệ tuyến tính giữa treatment và kết quả theo dạng $\mu_i = \beta_0 + \beta_1 x_i$

- Từ dữ liệu thí nghiệm (x_i và y_i), tính ra phương trình hồi quy ước lượng mối quan hệ trên:

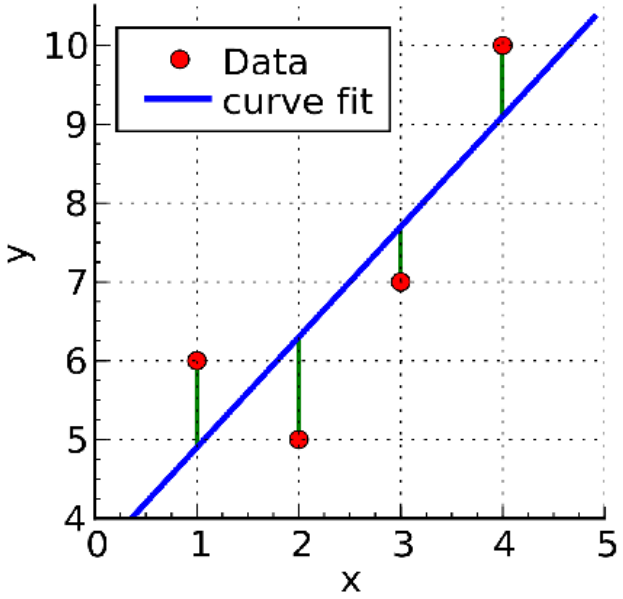
$$\hat{y}_i = b_0 + b_1 x_i + \varepsilon_i$$

- Các hệ số cần ước lượng: b_0 là chặn (intercept) và b_1 là hệ số góc (slope)

- Phần dư (residual) của phương trình hồi quy: **residual = y_i - $\hat{y}_i$**

- Tổng bình phương sai số của phương trình hồi quy :

$$SSE = \Sigma[y_i - \hat{y}_i]^2 = \Sigma[y_i - (b_0 + b_1 x_i)]^2$$



- Phương pháp **bình phương tối thiểu (Least Square)** xác định các hệ số b0 và b1 sao cho SS nhỏ nhất

Linear regression

- Để SSE (chưa xác định) tối thiểu thì

$$\partial SS / \partial b_0 = \Sigma -2 [Y_i - b_0 - b_1 X_i] = 0$$

$$\partial SS / \partial b_1 = \Sigma 2 [Y_i - b_0 - b_1 X_i] X_i = 0$$

$$\Rightarrow \quad n b_0 + b_1 \Sigma X = \Sigma Y$$

$$\Sigma X b_0 + b_1 \Sigma X^2 = \Sigma XY$$

- Tính các hệ số b0, b1 và standard error của b0 và b1

$$\bar{X} = \Sigma X / n \quad ; \bar{Y} = \Sigma Y / n \quad (\text{ký hiệu } \Sigma Y = \sum_{i=1}^{n} Y_i)$$

$$SSxx = \Sigma (X - \bar{X})^2 = \Sigma X^2 - \frac{(\sum X)^2}{n}$$

$$SSyy = \Sigma (Y - \bar{Y})^2 = \Sigma Y^2 - \frac{(\sum Y)^2}{n}$$

$$SSxy = \Sigma (X - \bar{X})*(Y - \bar{Y}) = \Sigma XY - \frac{(\sum X * \sum Y)}{n}$$

$$\Rightarrow \quad b_1 = SSxy / SSxx \text{ và phương sai } s_{b1}^2 = \frac{s_E^2}{SSxx}$$

$$b_0 = \bar{Y} - b_1 * \bar{X}$$

$$s_{b0}^2 = s_E^2 * \left[\frac{1}{n} + \frac{\bar{X}^2}{SSxx} \right]$$

với: $$s_E^2 = \frac{SSyy - \left(\frac{SSxy^2}{SSxx} \right)}{(n-2)}$$

- Lưu ý: SE tính ở trên là residual standard error. Về ý nghĩa $SE = \sqrt{\frac{SSE}{df}}$ (với **df=n-2**) nhưng ta có thể tính SE qua SSyy, SSxy, SSxx mà không cần các hệ số b_0, b_1

Linear regression

X	Y	XY	$X^2 = XX$	$Y^2 = YY$
35	114	3 990	1 225	12 996
45	124	5 580	2 025	15 376
55	143	7 865	3 025	20 449
65	158	10 270	4 225	24 964
75	166	12 450	5 625	27 556
275	705	40 155	16 125	101 341
55	141			
	Cf:	38 775	15 125	99 405
	SS:	1 380	1 000	1 936

n = 5

$b_1 = SSxy / SSxx = 1380 / 1000 = 1{,}38$

$b_0 = \bar{Y} - b_1 * \bar{X} = 141 - 1{,}38 * 55 = 65{,}1$

$$s_E^2 = \frac{SSyy - \left(\frac{SSxy^2}{SSxx}\right)}{(n-2)} = 10{,}5333 \Rightarrow s_E = 3{,}245$$

$$s_{b1}^2 = \frac{s_E^2}{SSxx} = 0{,}01053 \Rightarrow s_{b1} = 0{,}103$$

$$s_{b0}^2 = s_E^2 * \left[\frac{1}{n} + \frac{\bar{X}^2}{SSxx}\right] = 33{,}970 \Rightarrow s_{b0} = 5{,}828$$

t-test cho Linear regression

- Tính các hệ số b0, b1 của phương trình hồi quy $\hat{y}_i = b_0 + b_1 x_i$
- Tính SE_{b0} và SE_{b1}
- Tính các giá trị t ứng với b0 và b1
  - $t_{b0} = b0/SE_{bo}$; $t_{b1} = b1/SE_{b1}$
- Tính p-value **2 đầu** của t_{b0} và t_{b1} với df=n-2 cho cả hai giá trị t (với n là tổng số đối tượng)
- Nếu p-value>0.05 thì hệ số tương ứng không có ý nghĩa thống kê. Nếu p-value của t_{b1} = $P(t \leq -|t_{b1}| \text{ or } t \geq |t_{b1}|)$>0.05 thì b1 không có ý nghĩa thống kê, tức là kết quả không có tương quan tuyến tính với x
- Ta có thể kiểm tra giả thuyết “y có tương quan tuyến tính với x theo hệ số b1=B” qua giá trị $t=(b1-B)/SE_{b1}$ và sau đấy tính p-value 2 đầu cho giá trị t này

Khoảng tin cậy cho linear regression

- Tính giá trị $t_{0.025}$ cho khoảng tin cậy 95% (df=n-2)
- Khoảng tin cậy của b0

  $\beta_0 = b_0 \pm t_{0.025} * SE_{b0}$

- Khoảng tin cậy của b1 (nếu khoảng tin cậy chứa 0 thì kết luận y không tương quan tuyến tính với x)

  $\beta_1 = b_1 \pm t_{0.025} * SE_{b1}$

- Khoảng tin cậy của μ_i (trung bình của dân số y_i ứng với treatment x_i)

  $\mu_i = \hat{y}_i \pm t_{0.025} * SE_{\mu}$ với $s_\mu^2 = s_E^2 * \left[\frac{1}{n} + \frac{(X_i - \overline{X})^2}{SSxx}\right]$

- Lưu ý:
  - với phương trình trên, khoảng tin cậy của μ_i sẽ có tâm là giá trị $\hat{y}_i$ tính bởi pt hồi quy
  - **Ta có thể dùng phương trình hồi quy để dự đoán giá trị y tại x bất kì. Khoảng tin cậy cho dự đoán này vẫn được tính theo cách trên**

Ví dụ

X	Y
35	114
45	124
55	143
65	158
75	166

- SE = 3.245 ; SE_{b1} = 0.103 ; SE_{b0} = 5.828 ; $t_{0.025,\ df=3}$ = 3.182 ; b1 = 1.38 ; b0 = 65.1
- $t_{b0} = b_0/SE_{b0}$ = 65.1/5.828 = 11.2 → p-value = P(t ≤ -12.2 | t ≥ 11.2) = 0.0015
- $t_{b1} = b_1/SE_{b1}$ = 1.38/0.103 = 13.45 → p-value = P(t ≤ -13.45 | t ≥ 13.45) = 0.0009
- Khoảng tin cậy

$\beta_1 = b_1 \pm t_{0.05} * s_{b1} = 1{,}38 \pm 3{,}182 * 0{,}103 =$ từ 1,05 đến 1,71

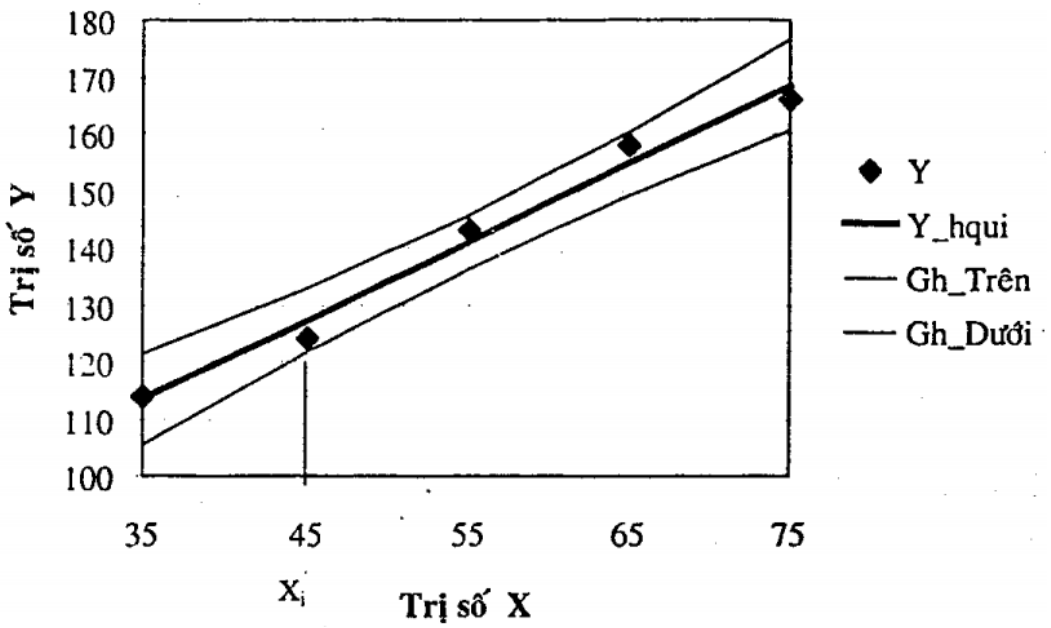
$\Rightarrow$ X và Y có quan hệ tuyến tính.

$\beta_0 = b_0 \pm t_{0.05} * s_{b0} = 65{,}1 \pm 3{,}182 * 5{,}828 =$ từ 46,5 đến 83,6

Khoảng tin cậy của $\mu_{y.x}$ (trung bình của dân số Y_i ứng với trị số X_i)

Ví dụ: X = 45; $\hat{Y}_{45} = 65{,}1 + 1{,}38 * 45 = 127{,}2.$

X	Y	$\hat{Y}$	Giới hạn Trên	Giới hạn Dưới	s_μ^2	s_μ
35	114	113,4	121,4	105,4	6,32	2,51
45	124	127,2	132,9	121,5	3,16	1,78
55	143	141,0	145,6	136,4	2,11	1,45
65	158	154,8	160,5	149,1	3,16	1,78
75	166	168,6	176,6	160,6	6,32	2,51



- Nhận xét: từ công thức $s_\mu^2 = s_E^2 * \left[\frac{1}{n} + \frac{(X_i - \overline{X})^2}{SSxx}\right]$ ta thấy SS_μ lớn hơn khi X_i xa giá trị trung bình của X (= 55 trong ví dụ này) và do đó khoảng tin cậy sẽ rộng hơn

Linear regression sử dụng R

```
> x <- c(35 , 45 , 55 , 65 , 75)
> y <- c(114, 124, 143, 158, 166)
> relation <- lm(formula = y ~ x)
> print(relation)
Call:
lm(formula = y ~ x)
Coefficients:
(Intercept)          x
     65.10         1.38

> print(summary(relation))
Call:
lm(formula = y ~ x)
```

Residuals:

1	2	3	4	5
0.6	-3.2	2.0	3.2	-2.6

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	65.1000	5.8284	11.17	0.001538
x	1.3800	0.1026	13.45	**0.000889**

Residual standard error: 3.246 on 3 degrees of freedom

Multiple R-squared: 0.9837, Adjusted R-squared: 0.9782

F-statistic: 180.8 on 1 and 3 DF, p-value: **0.0008894**

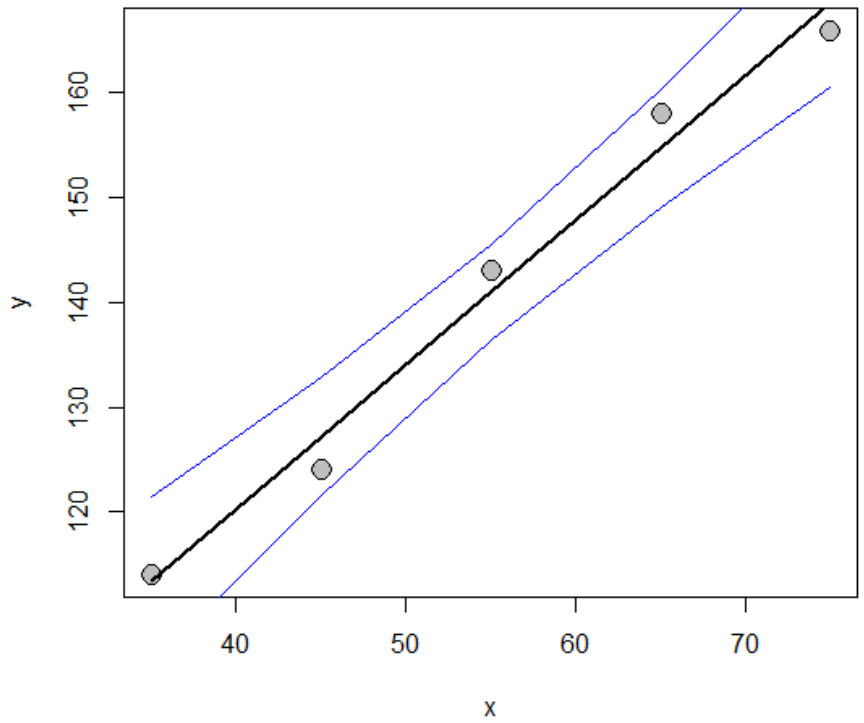
```
> res=resid(relation)
> sqrt(sum(res^2)/3) =  3.24551
```

SSE=Σresidual^2 $SE = \sqrt{\frac{SSE}{df}}$

```
> confint(relation, level = 0.95)
                2.5 %      97.5 %
(Intercept) 46.551497   83.648503
x            1.053379    1.706621
```

```
> pre=predict(relation,interval = 'confidence')
     fit      lwr      upr
1  113.4 105.3995 121.4005
2  127.2 121.5428 132.8572
3  141.0 136.3809 145.6191
4  154.8 149.1428 160.4572
5  168.6 160.5995 176.6005
```

```
> plot(x, y, cex = 1.75, pch = 21, bg = 'gray')
> lines(x,pre[1:5,1], col = 'black', lwd = 2)
> lines(x,pre[1:5,2],col='blue')
> lines(x,pre[1:5,3],col='blue')
```



ANOVA for regression

- Ta có thể kiểm định giả thuyết H0: $\beta_1 = 0$ (tức là y không có tương quan với x) bằng ANOVA

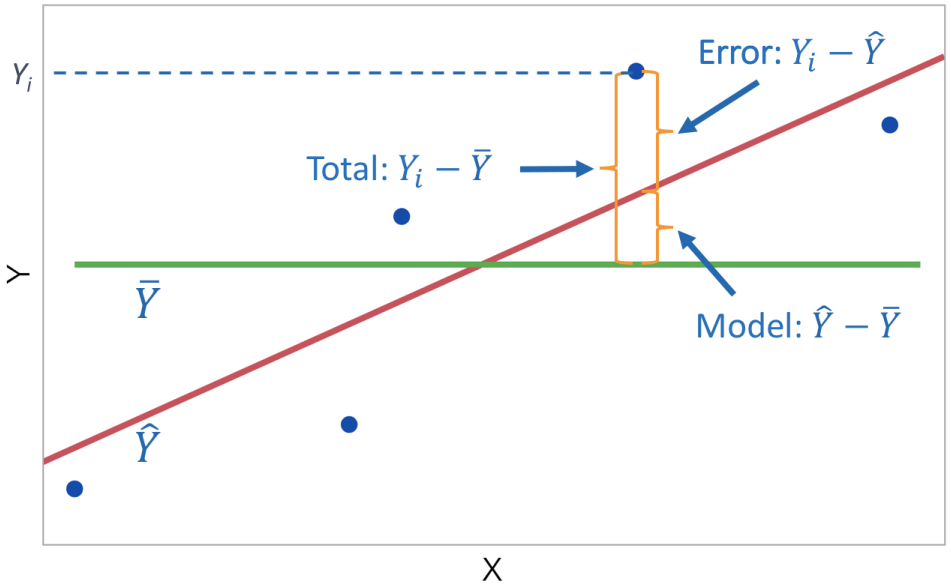
- Cơ sở:

$$y_i - ybar = \hat{y}_i - ybar + y_i - \hat{y}_i$$

$$\rightarrow \Sigma (y_i - ybar)^2 = \Sigma (\hat{y}_i - ybar)^2 + \Sigma (y_i - \hat{y}_i)^2$$

$$SSTo = SSReg + SSE$$

với $df = n-1 = 1 + n-2$



- SSReg là explained variation (giải thích bởi regression)
- SSE là unexplainied variation (sai số ngẫu nhiên)

- Nếu H0 đúng thì tất cả biến thiên trong dữ liệu đều do ngẫu nhiên (F sẽ xấp xỉ 1)

- $F = \dfrac{\dfrac{SSReg}{1}}{\dfrac{SSE}{(n-2)}}$ tuân theo phân bố F với 1 và n-2 bậc tự do. Nếu P(>F) < 0.05 → bác bỏ H0 (kết luận x và y có quan hệ)

ANOVA for regression

$$\text{SSTo} = \text{SSyy} = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n}$$

$$\text{SSReg} = (\text{SSxy})^2 / \text{SSxx} = \frac{\left[\Sigma XY - \frac{\Sigma X * \Sigma Y}{n}\right]^2}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}$$

$$\text{SSE} = \text{SSTo} - \text{SSReg}$$

Source of Variation	df	Sum of Squares	Mean Square	f
Regression	1	SSR	SSR	$\frac{\text{SSR}}{\text{SSE}/(n-2)}$
Error	$n-2$	SSE	$s^2 = \frac{\text{SSE}}{n-2}$	
Total	$n-1$	SST		

ANOVA Table for Simple Linear Regression

- Từ bảng ANOVA ta có thể tính thêm
  - Hệ số xác định R^2 = SSReg/SSTo
  - Hệ số tương quan $\sqrt{R^2}$

ANOVA for regression in R

```
> x <- c(35 , 45 , 55 , 65 , 75)
> y <- c(114, 124, 143, 158, 166)
> relation <- lm(formula = y ~ x)
> anova(relation)
```

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	1904.4	1904.40	180.8	0.0008894
Residuals	3	31.6	10.53		

```
Coefficients:   Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)     65.1000    5.8284       11.17     0.001538
x               1.3800     0.1026       13.45     0.000889
---
Residual standard error: 3.246 on 3 degrees of freedom
Multiple R-squared:  0.9837,        Adjusted R-squared:  0.9782
F-statistic: 180.8 on 1 and 3 DF,   p-value: 0.0008894
```

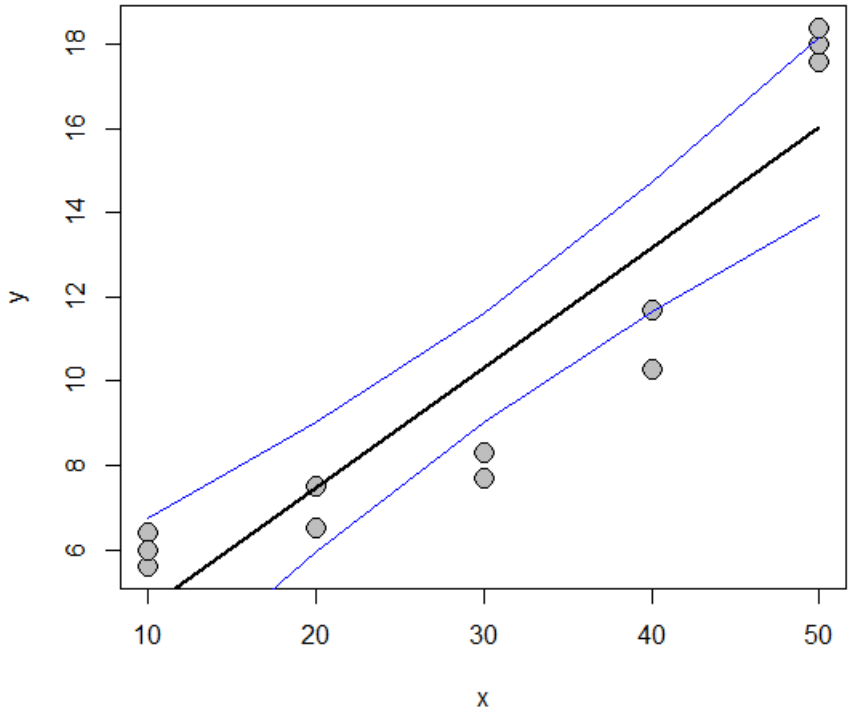
- Tính hệ số xác định

R^2 = SSReg/SSTo = SSReg / (SSReg+SSE) = 1904.4/(1904.4+31.6) = 0.9836777

- Thực tế R^2 luôn tăng khi số biến tăng (hồi quy đa biến), Adjusted R-squared là giá trị được điều chỉnh lại

Lack of fit testing (kiểm tra tính phù hợp của phương trình hồi quy)

- Cứ với n cặp số liệu (xi,yi) là lập được pt: $\hat{y}_i = b_0 + b_1 x_i$
- t-test và F-test cho biết hệ số b1 có ý nghĩa thống kê hay không
- Nếu b1 có ý nghĩa thống kê, kết luận x và y có quan hệ



- Vấn đề:
  - Làm sao biết phương trình hồi quy thể hiện quan hệ giữa x và y một cách phù hợp nhất?
  - Nói cách khác, nếu phương trình hồi quy không thể hiện đầy đủ mối quan hệ giữa x và y thì kết quả này chỉ là ngẫu nhiên hay do chính mô hình hồi quy không phù hợp?

- Lack of fit testing
  - H0: there is no lack of fit (tức là mô hình hồi quy phù hợp)

  - Nếu p-value < 0.05 → bác bỏ H0 (tức là kết luận không phù hợp)

Lack of fit testing

```
> x <- c(10,10,10,20,20,30,30,40,40,50,50,50)
> y <- c(6.4,5.6,6.0,7.5,6.5,8.3,7.7, 11.7, 10.3, 17.6, 18.0, 18.4)
> relation <- lm(formula = y ~ x)
> summary(relation)
```

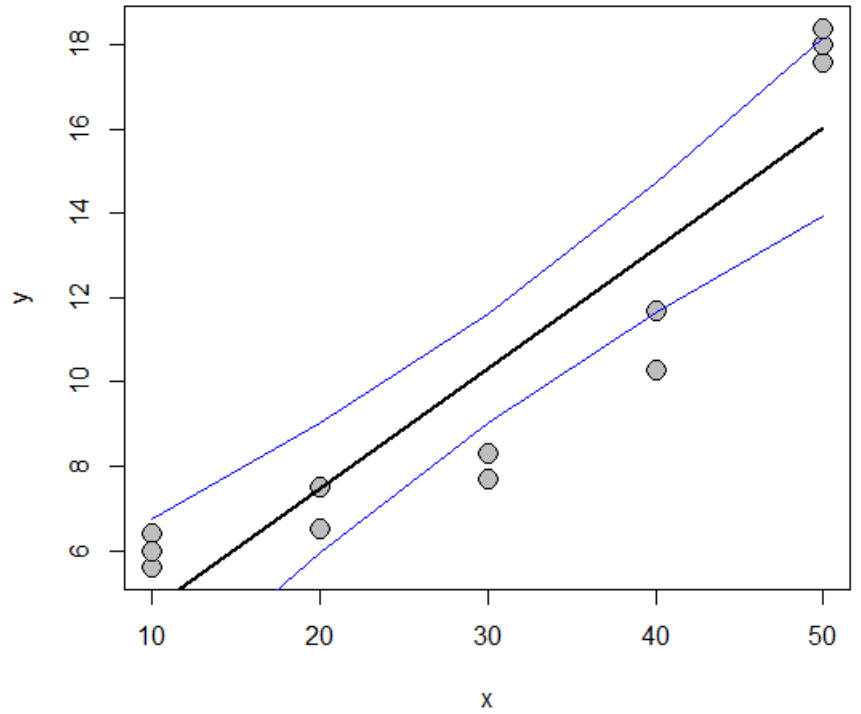
x	y	x	y	x	y	x	y	x	y
10	6.4	20	7.5	30	8.3	40	11.7	50	17.6
10	5.6	20	6.5	30	7.7	40	10.3	50	18.0
10	6.0							50	18.4

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.76190	1.27869	1.378	0.198
x	0.28571	0.03798	7.522	**2.01e-05**

Residual standard error: 2.01 on 10 degrees of freedom

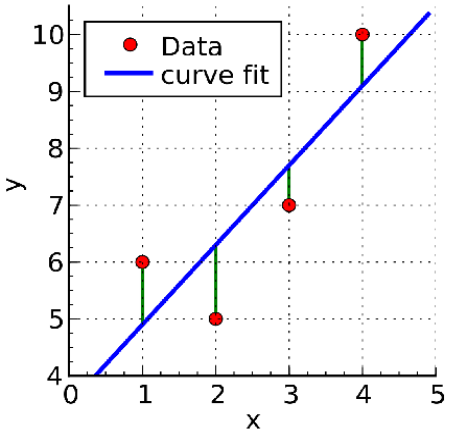
Multiple R-squared: **0.8498**, Adjusted R-squared: 0.8348

F-statistic: 56.58 on 1 and 10 DF, p-value: **2.011e-05**



Kết quả t-test và F-test đều đưa ra kết luận có quan hệ

Lack of fit F-test



- SSE = $\Sigma[y_i - \hat{y}_i]^2$ là sai số không giải thích được bởi mô hình

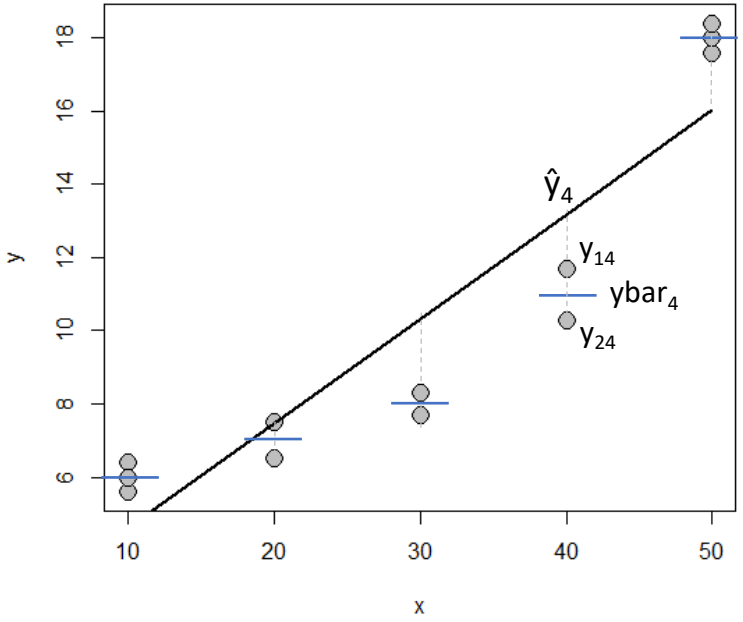
- SSE = SSEp + SSLf trong đó

  - SSEp là sai số thuần (pure error)

$$SSEp = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - ybar_j)^2 \quad \text{với} \quad df = \sum_{j=1}^{k}(n_j - 1) = n - k$$

  - SSLf là sai số do không phù hợp (lack of fit error)

    SSLf = SSE - SSEp với df = k-2



- H0: mô hình phù hợp (there is no lack of fit)

- F = [SSLf/(k-2)] / [SSEp/(n-a)] tuân theo phân phối F với k-2 và n-k bậc tự do

- Nếu P(>F) < 0.05 → bác bỏ H0 → kết luận mô hình không phù hợp

ANOVA for lack of fit F-test

Nguồn	df	SS	MS	$F_{tính}$	$F_{bảng}$
Hồi qui	1	SSReg	MSReg		
Dư (Residual)	n –2	SSE			
• Không phù hợp	a –2	SSLf	MSLf	$\frac{MSLf}{MSEp}$	
• Sai số thuần	n – a	SSEp	MSEp		
Tổng	n –1	SSTo			

Lần lập lại ↓	$10 = X_1$	$20 = X_2$	$30 = X_3$	$40 = X_4$	$50 = X_5$
1	6,4	7,5	8,3	11,7	17,6
2	5,6	6,5	7,7	10,3	18,0
n_j	6,0				18,4
Tổng :	18,0	14	16,0	22,0	54
Số lần lập lại n_j :	3	2	2	2	3
Trung bình :	6,0	7,0	8,0	11,0	18

Nguồn biến động	df	SS	MS	$F_{tính}$	$F_{bảng}$ α=0,05	$F_{bảng}$ α=0,01
Hồi qui	1	228,57	228,57			
Dư		40,39				
• Không phù hợp	3	38,09	12,697	38,64	4,35	8,45
• Sai số thuần	7	2,30	0,3286			
Tổng	11	268,96				

SSE = 40.39
SSEp = 2.3
SSLf = SSE - SSEp = 38.09

F = (38.09/3) / (2.3/7) = 38.64

P(F>38.64) = 0.0001 → bác bỏ H0 (kết luận mô hình không phù hợp)

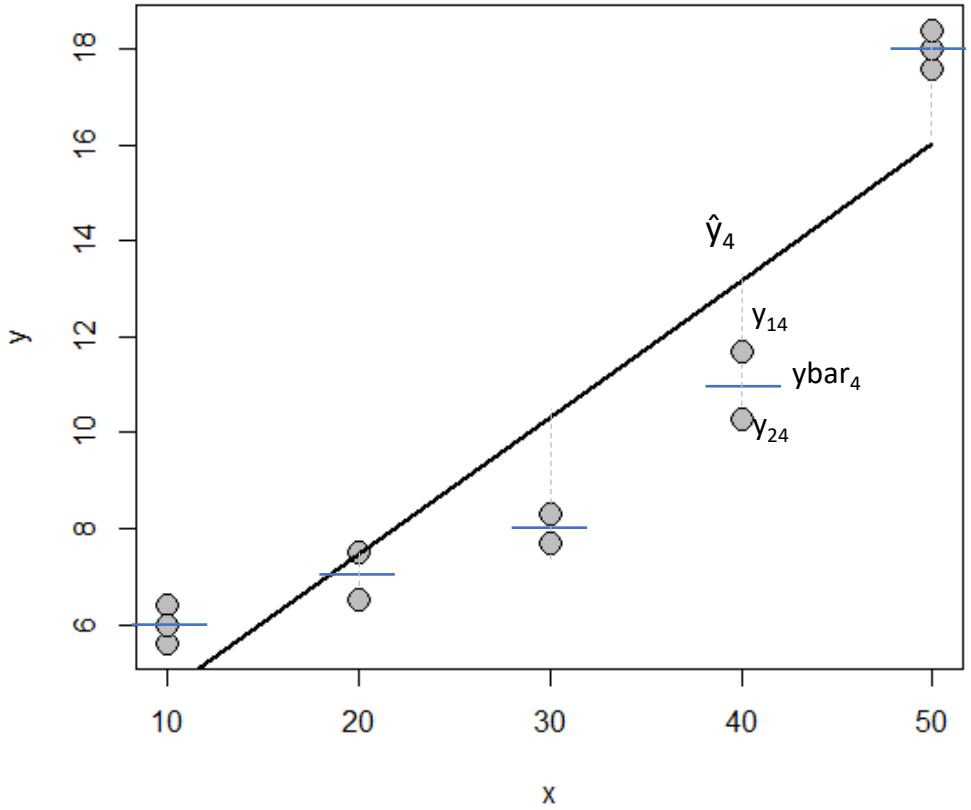
Lack of fit F-test in R

```
> x <- c(10,10,10,20,20,30,30,40,40,50,50,50)
> y <- c(6.4,5.6,6.0,7.5,6.5,8.3,7.7, 11.7, 10.3, 17.6, 18.0, 18.4)
> relation <- lm(formular = y ~ x)
> library(alr3)
> pureErrorAnova(relation)
```

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	228.571	228.571	695.652	2.883e
Residuals	10	40.395	4.040		
Lack of fit	3	38.095	12.698	38.647	0.0001002
Pure Error	7	2.300	0.329		



Quadratic regression

- Giả sử tồn tại một mối quan hệ giữa treatment và kết quả theo dạng $\mu_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$

- Từ dữ liệu thí nghiệm (x_i và y_i), tính ra phương trình hồi quy ước lượng mối quan hệ trên:

$$\hat{y}_i = b_0 + b_1 x_i + b_2 x_i^2 + \varepsilon_i$$

- Các hệ số cần ước lượng: b_0 là chặn (intercept), b_1 là hệ số ảnh hưởng tuyến tính, và b_2 là hệ số ảnh hưởng bậc 2

- Phần dư (residual) của phương trình hồi quy: **residual = $y_i - \hat{y}_i$**

- Tổng bình phương sai số của phương trình hồi quy :

$$SSE = \Sigma[y_i - \hat{y}_i]^2 = \Sigma[y_i - (b_0 + b_1 x_i + b_2 x_i^2)]^2$$

- Phương pháp **bình phương tối thiểu (Least Square)** xác định các hệ số b0, b1, b2 sao cho SS nhỏ nhất

t-test for Quadratic regression

- Tính các hệ số b_0, b_1, b_2 của phương trình hồi quy $\hat{y}_i = b_0 + b_1 x_i + b_2 x_i^2$
- SE là residual standard error ($y_i - \hat{y}_i$) với **df=n-3**
- Tính SE_{b0} , SE_{b1} , SE_{b2} **(df=3)**

- Tính các giá trị t ứng với b0 , b1 , b2
  - $t_{b0} = b0/SE_{bo}$; $t_{b1} = b1/SE_{b1}$; $t_{b2} = b2/SE_{b2}$

- Tính p-value **2 đầu** với **df=n-3** cho cả các giá trị t (với n là tổng số đối tượng)

- Nếu p-value>0.05 thì hệ số tương ứng không có ý nghĩa thống kê (kết luận không có linear hoặc quadratic effect)

- Ta có thể kiểm tra giả thuyết "y có tương quan với x theo hệ số b_1=B1, b_2=B2" qua giá trị $t=(b_1-B1)/SE_{b1}$ và $t=(b_2-B2)/SE_{b2}$ và sau đấy tính p-value 2 đầu cho các giá trị t này

F-test and lack of fit test

	df	SS	MS	F	P
x	1	SSx	SSx	MSx/MSE	P(>F) (df=1,n-3)
x^2	1	SSx2	SSx2	MSx2/MSE	P(>F) (df=1,n-3)
regression	2	SSReg=SSx+SSx2	SSReg/2	MSReg/MSE	P(>F) (df=2,n-3)
Residuals	n-3	SSE=SSLf+SSp	SSE/(n-3)		
Lack of fit	k-3	SSLf	SSLf/k-3	MSLf/MSEp	P(>F) (df=k-3,n-k)
Pure error	n-k	SSEp	SSEp/n-k		
Total	n-1	SSTo			

Quadratic regression in R

```
> x <- c(10,10,15,20,20,25,25,25,30,35)
> y <- c(73,78,85,90,91,86,87,91,74,65)
> x2 <- x^2
> relation_quad=lm(formula = y ~ x + x2)
> summary(relation_quad)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.4130	-1.1936	-0.0439	1.7100	3.6994

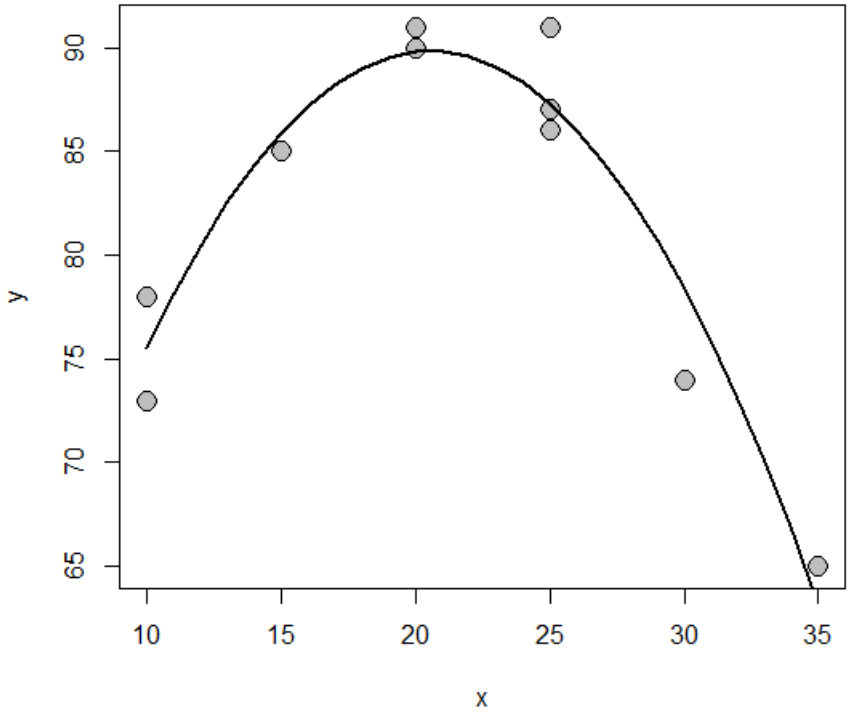
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	35.72231	6.09781	5.858	0.000625
x	5.26369	0.60569	8.690	5.35e-05
x2	-0.12802	0.01391	-9.206	3.68e-05

Residual standard error: 2.758 on 7 degrees of freedom

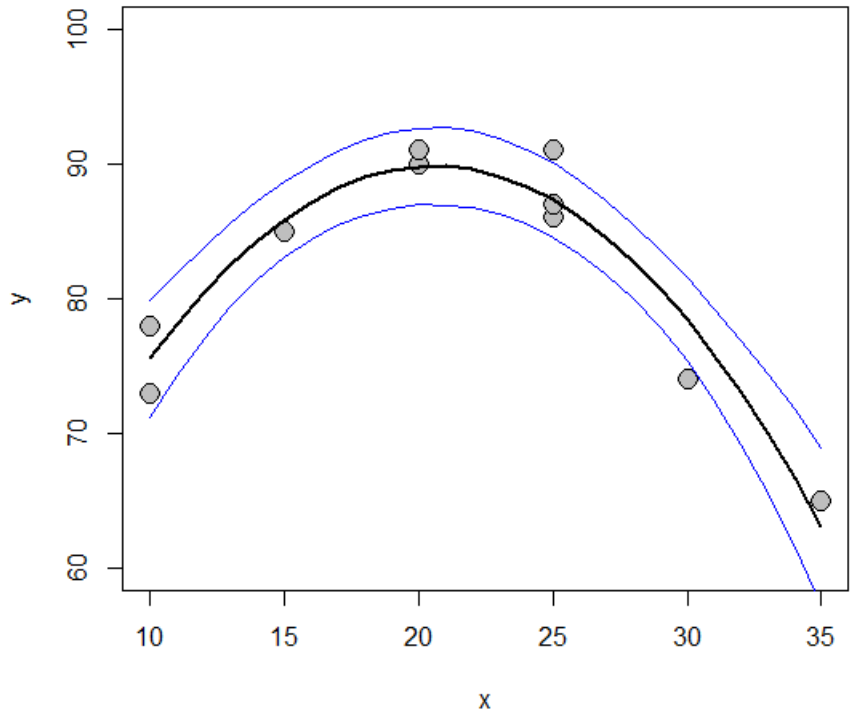
Multiple R-squared: 0.9267, Adjusted R-squared: 0.9057

F-statistic: 44.22 on 2 and 7 DF, p-value: 0.0001068



Prediction and confidence interval

```
> x_new <- 10:35
> ynew=predict(relation_quad,list(x=x_new,x2=x_new^2),interval = 'confidence')
> plot(x, y, cex = 1.75, pch = 21, bg = 'gray',ylim=c(60,100))
> lines(x_new,ynew[1:26,1], col = 'black', lwd = 2)
> lines(x_new,ynew[1:26,2], col = 'blue', lwd = 1)
> lines(x_new,ynew[1:26,3], col = 'blue', lwd = 1)
```



```
> anova(relation_quad)
Analysis of Variance Table

Response: y
           Df  SS      MS      F         Pr(>F)
x           1  28.05  28.05  3.6877   0.09628 .
x2          1 644.71 644.71 84.7589 3.68e-05
Residuals   7  53.24   7.61
---
```

```
> pureErrorAnova(relation_quad)
Analysis of Variance Table

Response: y
             Df  SS      MS      F value  Pr(>F)
x             1  28.05  28.05  4.1555   0.1111317
x2            1 644.71 644.71 95.5120 0.0006142
Residuals     7  53.24   7.61
 Lack of fit  3  26.24   8.75   1.2960  0.3907213
 Pure Error   4  27.00   6.75
---
```

Thử với đa thức bậc cao hơn

```
> relation_tri <- lm(formula = y ~ x + I(x^2) + I(x^3))
> summary(relation_tri)
Call:
lm(formula = y ~ x + I(x^2) + I(x^3))
Coefficients:
             Estimate     SE          t value  Pr(>|t|)
(Intercept) 21.186930  21.174980   1.001   0.3557
x            7.678181   3.415289    2.248   0.0656 .
I(x^2)      -0.246044   0.164729   -1.494   0.1859
I(x^3)       0.001759   0.002445    0.719   0.4990

Residual standard error: 2.858 on 6 degrees of freedom
Multiple R-squared:  0.9325,        Adjusted R-squared:  0.8987
F-statistic: 27.62 on 3 and 6 DF,  p-value: 0.000656
```

```
> anova(relation_tri)
Analysis of Variance Table

Response: y
          Df  SS      MS      F value   Pr(>F)
x          1  28.05   28.05   3.4334    0.1133279
I(x^2)     1  644.71  644.71  78.9138   0.0001133
I(x^3)     1  4.23    4.23    0.5173    0.4990478
Residuals  6  49.02   8.17
```

Thử với đa thức bậc cao hơn

```
> x <- c(10,10,10,20,20,30,30,40,40,50,50,50)
> y <- c(6.4,5.6,6.0,7.5,6.5,8.3,7.7, 11.7, 10.3, 17.6, 18.0, 18.4)
> relation_tri <- lm(formula = y ~ x + I(x^2)+ I(x^3))
> anova(relation_tri)
```

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	228.571	228.571	795.031	2.704e-09
I(x^2)	1	34.667	34.667	120.580	4.204e-06
I(x^3)	1	3.429	3.429	11.925	**0.008652**
Residuals	8	2.300	0.287