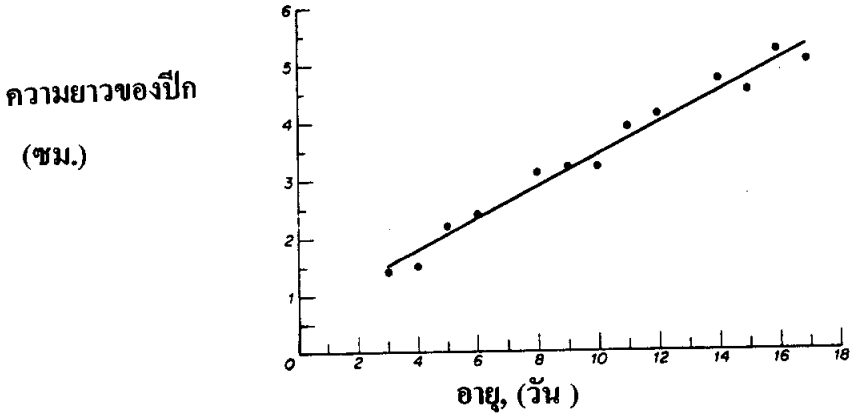
บทที่ 10
ความถดถอยและสหสัมพันธ์

1. Regression VS Correlation

การศึกษาความสัมพันธ์ระหว่างตัวแปรคู่ ในลักษณะที่ตัวแปรตัวหนึ่ง **คือตัวแปรตาม** (Dependent Variable) คือเป็นฟังก์ชันกับตัวแปรอีกตัวหนึ่ง ซึ่งเรียกว่า **ตัวแปรกำหนด** (Independent Variable) เช่น ความสัมพันธ์ระหว่างอายุและความดันโลหิต ความดันโลหิตเป็นตัวแปรตาม และอายุเป็นตัวแปรกำหนด แต่อายุจะเป็นตัวแปรตามไม่ได้ เพราะอายุจะถูกกำหนดโดยความดันโลหิตไม่ได้ ความสัมพันธ์ที่มีตัวแปรกำหนด และตัวแปรตาม เรียกว่า **ความถดถอย** หรือ regression ถ้ามีตัวแปรกำหนด 1 ตัว เรียกว่า simple regression ถ้ามีมากกว่า 1 ตัว เช่น นอกจากอายุแล้ว ยังใช้น้ำหนักของบุคคลด้วย เรียกว่า **ความถดถอยเชิงซ้อน** (multiple regression) ส่วน simple linear regression หมายถึงความสัมพันธ์ระหว่างตัวแปรคู่นั้น อยู่ในรูปเชิงเส้น (linear) ซึ่งจะทราบได้โดยการพิจารณากราฟข้อมูลซึ่งแสดงจุด (x_i, y_i) จำนวน n คู่อันดับ ดังรูปที่ 10.1

รูปที่ 10.1 แสดงความสัมพันธ์ระหว่างอายุและความยาวของปีกของลูกนก 13 ตัว



ส่วน**สหสัมพันธ์** (correlation) คือการศึกษาความสัมพันธ์ระหว่างตัวแปร 2 ตัว โดยไม่มีตัวแปรกำหนดหรือตัวแปรตาม ทั้ง 2 ตัวเป็นตัวแปรเชิงสุ่ม (regression มีตัวแปรตามเป็นตัวแปรเชิงสุ่ม)

correlation มุ่งศึกษาว่าเมื่อตัวแปรตัวหนึ่ง เปลี่ยนแปลงไประดับหนึ่ง ตัวแปรอีกตัวหนึ่งจะเปลี่ยนแปลง หรือไม่แปลี่ยนแปลง และเป็นการเปลี่ยนแปลงแบบใด เช่น ความยาวของแขนและขา จะมีสหสัมพันธ์แบบเชิงบวกคือผู้มีแขนยาว มักจะมีขายาวด้วย แต่เราไม่พูดว่าความยาวของแขนถูกกำหนดโดยความยาวของขา

เนื่องจากการศึกษา Regression และ Correlation มีสูตรค่อนข้างมาก จึงรวบรวมไว้ด้วยกัน พร้อมตัวอย่าง และจะอธิบายหลักการภายหลัง สูตรและข้อกำหนดต่างๆ มีดังนี้

(1) $Y_i = \alpha + \beta X_i$ คือสมการเส้นตรง หรือเส้นถดถอยของ **ประชากร** ซึ่งแสดงความสัมพันธ์ระหว่าง X และ Y ในรูปเชิงเส้นตรง

(2) $Y_i = a + bX_i$ คือสมการถดถอย **ตัวอย่าง** ซึ่งสร้างจากข้อมูล (x_i, y_i) ที่เก็บรวบรวมมา n คู่อันดับ ทำหน้าที่เป็นค่าประมาณของเส้นถดถอยประชากร การหาค่า a และ b ใช้หลัก "กำลังสองน้อยที่สุด" (least squares estimate) ซึ่งมีคุณสมบัติดังนี้

2.1 $\Sigma(Y_i-\hat{Y}_i) = 0$ และ $\Sigma(Y_i-\hat{Y}_i)^2 = \text{min. } (Y_i-\hat{Y}_i)^2$

$\Sigma(Y_i- \hat{Y}_i)^2$ เรียกว่า SSE(error) หรือ residual SS และ$\Sigma(Y_i-\hat{Y}_i)^2/(n-2)$ คือ SS/df = MSE หรือ $S^2_{y/x}$ ซึ่งจะกล่าวถึงในหัวข้อ ANOVA

2.2 b คือ regession coefficient เป็นค่าประมาณของ β b คือความลาดชันของเส้นถดถอย โดยหาจากสูตร

$$b = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(X-\bar{X})^2} \text{ หรือ } b = \frac{\Sigma XY - (\Sigma X)(\Sigma Y)/n}{\Sigma X^2 - (\Sigma X)^2/n}$$

หมายถึงการเปลี่ยนแปลงของ Y เมื่อ X เปลี่ยนไป 1 หน่วย

a คือจุดตัดที่แกน Y หรือ Y-intercept เป็นค่าประมาณของ α วิธีหาค่า a ใช้หลักการว่า เส้นถดถอยต้องผ่านจุด$(\bar{x}, \bar{y})$ ดังนั้นจึงแทนค่า $(\bar{x}, \bar{y})$ ในสมการประชากร จะได้

$\bar{Y} = \alpha + \beta\bar{X}$ ดังนั้น $\alpha = \bar{Y} - \beta\bar{X}$ และค่าประมาณของ α คือ

$\hat{\alpha} = a = \bar{Y} - b\bar{X}$

(3) การพยากรณ์ค่าต่างของ Y_i โดยกำหนดค่า X_i ได้จากการแทนค่า X_i ในสมการ $Y = a + bX_i$ ซึ่งจะมีค่าพยากรณ์อยู่ 3 แบบด้วยกันคือ

3.1 ค่าพยากรณ์นั้นคือ**ค่าเฉลี่ย**ของ Y_i เมื่อกำหนด X_i จะมี

$$S_{\hat{\bar{y}}} = \sqrt{S^2_{y/x}(1/n + (X_i-\bar{X})^2/\Sigma(X-\bar{X})^2}$$

3.2 ค่าพยากรณ์นั้นคือค่า Y ใดๆ เมื่อกำหนด X_i จะมี

$$S_{\hat{y}} = \sqrt{S^2_{y/x}(1 + 1/n + (X_i-\bar{X})^2/\Sigma(X-\bar{X})^2}$$

จะสังเกตได้ว่า $S_{\hat{y}}$ มีค่าไม่คงที่ จะเปลี่ยนแปลงตามค่าของ X_i จะมีค่าน้อยที่สุด เมื่อพยากรณ์ $X_i = \bar{X}$ จะได้ $S_{\hat{\bar{y}}}$ (ค่าเฉลี่ย) $= S^2_{y/x}/n$ และ $S_{\hat{y}}$ (y ใดๆ) $= \sqrt{S^2_{y/x}(1 + 1/n)}$ และ $S_{\hat{y}}$ จะมีค่าสูงขึ้นเรื่อยๆ เมื่อ X_i ถอยห่าง ห่างจาก $\bar{X}$

(4) **การทดสอบความสัมพันธ์เชิงเส้น**

ก่อนที่จะใช้สมการ $Y = a + bX$ พยากรณ์ค่า Y_i ควรศึกษาก่อนว่า X และ Y มีความสัมพันธ์แบบเชิงเส้น หรือไม่ นั่นคือ ตรวจสอบว่า $Y = a + bX$ สมควรเป็นสมการพยากรณ์ หรือไม่ นั่นคือ ทดสอบว่า $Y = a + bX$ ทำหน้าที่เป็นค่าประมาณที่ดีของ $Y = \alpha + \beta X$ หรือไม่ เนื่องจาก $\alpha + \beta X$ คือสมการเส้นตรง การตรวจสอบความสัมพันธ์เชิงเส้น คือการตรวจสอบว่า $\beta = 0$? มีวิธีการตรวจสอบ 2 วิธีคือ t-test และ F-test (ทำ ANOVA) มีวิธีการดังนี้

4.1 t-test : $H_0 : \beta = 0$ (X และ Y ไม่มีความสัมพันธ์แบบเชิงเส้น)

(1) $H_a : \beta \neq 0$ (X และ Y มีความสัมพันธ์แบบเชิงเส้น แต่ไม่ได้กำหนดทิศทาง)

หรือ (2) $H_a : \beta > 0$ (X และ Y มีความสัมพันธ์แบบเชิงเส้น และเป็นแบบเชิงบวก)

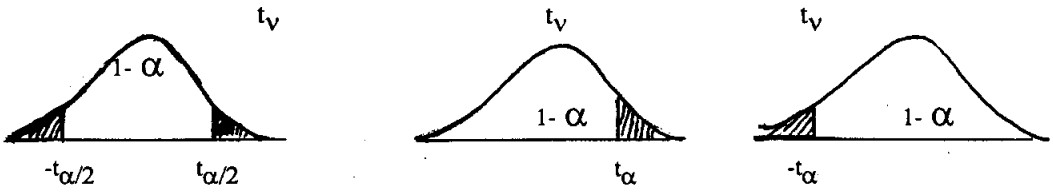
หรือ (3) $H_a : \beta < 0$ (X และ Y มีความสัมพันธ์แบบเชิงเส้น และเป็นแบบเชิงลบ)

ตัวสถิติที่ใช้ทดสอบ คือ

$T = b/S_b \sim T_{(n-2)}$, $S_b = \sqrt{S^2_{y.x}/\Sigma(X-\bar{X})^2}$, $S^2_{y/x} = MSE$

โดยเขตวิกฤต จะมี 3 อย่างคือ

(1) H_a : $\beta \neq 0$ (2) H_a: $\beta > 0$ (3) H_a : $\beta < 0$



4.2 **F-test** (ANOVA) $H_0 : \beta = 0, H_a : \beta \neq 0$

ตารางที่ 10.1 แสดง ANOVA ของ simple regression

Source	df	SS	MS	F
Linear Regression (R)	1	SSR	SSR/1	F = MSR/MSE
Residual or Error	n-2	SSE	SSE/(n-2)	เทียบกับค่าตาราง ที่ $f_{1,n-2}$
Total	n-1	SST		

$SST = \Sigma(Y-\bar{Y})^2 = \Sigma Y^2 - (\Sigma Y)^2/n$

$SSR = \{\Sigma(X-\bar{X})(Y-\bar{Y})\}^2/\Sigma(X-\bar{X})^2$ หรือ $b\Sigma(X-\bar{X})(Y-\bar{Y})$ หรือ $b\Sigma(X-\bar{X})^2$

(5) **การทดสอบความลาดชัน** (ซึ่งคืออัตราการเปลี่ยนแปลงของ Y ต่อ 1 หน่วยของ X)

$H_0 : \beta = \beta^*$, $H_a : \beta \neq \beta^*$ หรือ $\beta > \beta^*$ หรือ $\beta < \beta^*$ ($\beta^* \neq 0$)

ต้องใช้ **t-test** วิธีเดียว (ใช้ F-test ไม่ได้)

ตัวสถิติทดสอบคือ $T = (b - \beta^*)/S_b \sim t_{(n-2)}$

โดยเขตวิกฤตจะต้องสอดคล้องกับ H_a ว่าเป็นการทดสอบ ด้านเดียวหรือ 2 ด้าน

(6) **สัมประสิทธิ์การตัดสินใจ : r^2** (coefficient of determination) r^2 เป็นค่าประมาณของ ρ^2 = SSR/SST, $0 \leq r^2 \leq 1$

การอธิบายผล นิยมคูณด้วย 100 เพื่อให้เป็นเปอร์เซนต์ หมายถึง สัดส่วนของความผันแปรทั้งหมดของ Y ที่สามารถอธิบายโดยสมการถดถอย (หรือตัวแปร X)

ดังนั้น $1 - r^2 = SSE/SST$ เรียกว่า **coefficient of nondetermination** หมายถึง สัดส่วนความผันแปรของ Y ที่ไม่สามารถอธิบายโดยสมการถดถอย

(7) **การสร้างช่วงเชื่อมั่น**

7.1 $(1-\alpha)100\%$ ช่วงเชื่อมั่นของ α คือ $a \pm t_{\alpha/2,(n-2)}S_a$

7.2 $(1-\alpha 100\%$ ช่วงเชื่อมั่นของ β คือ $b \pm t_{\alpha/2,(n-2)}S_b$

7.3 $(1-\alpha)100\%$ ช่วงเชื่อมั่นของค่าพยากรณ์ Y_i เมื่อกำหนด X_i

$$= \hat{Y}_i \pm t_{\alpha/2,(n-2)}S_y$$

(8) การทดสอบสมมติฐานเกี่ยวกับค่าพยากรณ์ Y_i หรือ $\mu_{y/x}$, $H_0 : \mu_{y/x} = \mu_0$,

$H_a : \mu_{y/x} \neq \mu_0$, $\mu_{y/x} > \mu_0$, $\mu_{y/x} < \mu_0$ ใช้ได้แต่ t-test (F-test ใช้ไม่ได้) ตัวสถิติทดสอบคือ $T = (Y_i - \mu_0)/S_{\hat{y}} \sim t_{(n-2)}$

(9) assumptions ของ การศึกษา regression มีดังนี้

9.1 ณ แต่ละค่าของ X_i จะมีประชากรของค่า Y_i ซึ่งมีการแจกแจงแบบ**ปกติ** ด้วยค่าเฉลี่ย $\mu_{y/x}$ และความแปรปรวน $\sigma^2_{y/x}$

9.2 ประชากร ของ Y ณ ค่า X ต่างๆ มีความแปรปรวนเท่ากัน (เป็นเอกภาพ)

นั่นคือ $\sigma^2_{y/x_1} = \sigma^2_{y/x_2} = \ldots = \sigma^2_{y/x_K} = \sigma^2_{y/x}$

โดย $\sigma^2_{y/x} = S^2_{y/x} = SSE/(n-2) = MSE$

9.3 ค่าต่างๆ ของ X ต้องไม่มีความผิดพลาด

9.4 X เป็นตัวแปรที่สามารถกำหนดค่าล่วงหน้าได้ แต่ Y เป็นตัวแปรเชิงสุ่ม คือกำหนดค่าล่วงหน้าไม่ได้ และอิทธิพลต่างๆ ของ Y สามารถนำมารวมกันได้

(10) การพยากรณ์ค่าของ X (Inverse prediction)

ปกติ จะพยากรณ์ค่า Y จากค่า X_i (ข้อ 3) แต่ถ้าให้ Y เป็นตัวแปรกำหนด และ X เป็นตัวแปรตาม

สมการประมาณค่าคือ $\hat{X}_i = (\hat{Y}_i - a)/b$

(11) ข้อควรระวัง

11.1 สมการถดถอย ไม่ใช่สมการแสดงเหตุและผล แต่เป็นสมการที่มุ่งหมายแสดงความสัมพันธ์ระหว่าง X และ Y โดย b แสดงอัตราการเปลี่ยนแปลงของ Y ต่อ 1 หน่วยของ X

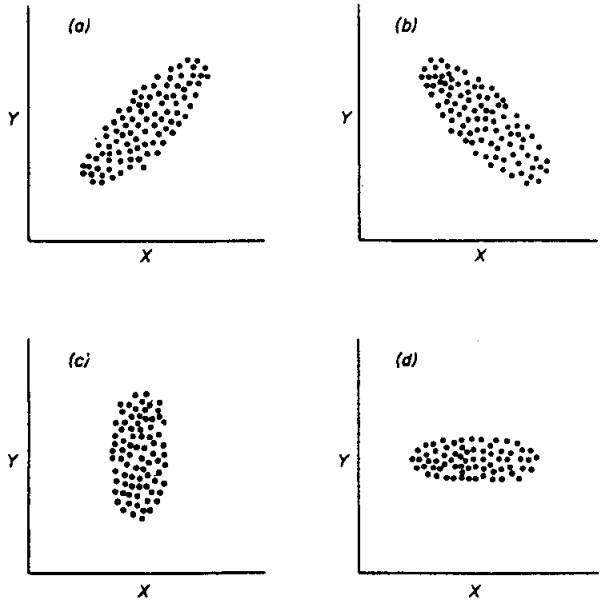
11.2 ค่าของ x_i และ y_i ต้องเป็นตัวเลข (เชิงปริมาณ)

(12) **สัมประสิทธิ์สหสัมพันธ์ : r** (Coefficient of correlation)

เป็นสัมประสิทธิ์ที่ใช้วัด **association** ระหว่าง 2 ตัวแปร (ไม่มีตัวแปรกำหนดหรือตัวแปรตาม)

$r = \sum(X-\bar{X})(Y-\bar{Y}) / \sqrt{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2}$, $-1 \leq r \leq +1$

เครื่องหมายของ r ขึ้นอยู่กับตัวตั้ง หรือ $\sum(X-\bar{X})(Y-\bar{Y})$ ถ้าเป็น + หรือ - แสดงว่ามีสหสัมพันธ์แบบตามกัน หรือ แย้งกัน ถ้า $\sum(X-\bar{X})(Y-\bar{Y}) = 0$ จะให้ $r = 0$ แสดงว่า X และ Y **ไม่มี** สหสัมพันธ์



รูปที่ 10.2 แสดงสหสัมพันธ์ระหว่าง X และ Y แบบต่างๆ
รูป (a) สหสัมพันธ์เชิงบวก
(b) สหสัมพันธ์เชิงลบ
(c), (d) ไม่มีสหสัมพันธ์

r ไม่มีหน่วยการวัด (เช่นเดียวกับ b) โดย b วัดการเปลี่ยนแปลงเชิงปริมาณระหว่าง 2 ตัวแปร ($-\infty < b < \infty$) r วัดความมากน้อยของสหสัมพันธ์ (intensity of association) ระหว่าง 2 ตัวแปร ส่วน r^2 วัดสัดส่วนความผันแปรของ Y (มีหน่วยเป็น %) ว่ามีมากน้อยแค่ไหนที่อยู่ในขอบเขตของ regression Y on X ถ้าทราบค่า r จะหา r^2 ได้โดยการยกกำลังสองค่า r

และถ้าทราบค่า r^2 ก่อน (จาก ANOVA) จะหาค่า r ได้ โดยการถอดรากที่สอง แต่ต้องระวังเครื่องหมาย เพราะ $r^2 = \pm r$ จะต้องเลือกเครื่องหมายเพียงอันเดียวโดยการสังเกตจากเครื่องหมายของ b หรือเครื่องหมายของ $\Sigma(X-\bar{X})(Y-\bar{Y})$ การอธิบายค่า r ค่อนข้างลำบาก เช่น ถ้า r = .50 อาจสับสนคิดว่า X และ Y มีสหสัมพันธ์ 50% ซึ่งผิด เพราะค่าของ r ไม่อธิบายเป็นเปอร์เซนต์ ซึ่งความจริง r = .50, จะให้ $r^2 = .25$ แสดงว่า X อธิบายความผันแปรของ Y ได้เพียง 25% เท่านั้น

(13) **การทดสอบนัยสำคัญของ r**

จากข้อ (12) จะเห็นว่า r = .50 จะให้ r^2 เพียง .25 ถ้า r = .30, $r^2 = .09 = 9\%$ ซึ่งน่าสงสัยว่า แท้จริงความสัมพันธ์ในประชากรอาจมีหรือไม่มี ดังนั้น จะต้องทดสอบนัยสำคัญของค่า r นั่นคือ ตรวจสอบว่า $\rho = 0$? สมมติฐานคือ $H_0 : \rho = 0$ (ไม่มีสหสัมพันธ์ระหว่าง X และ Y), $H_a : \rho \neq 0$ หรือ $\rho > 0$ หรือ $\rho < 0$ ตัวสถิติที่ใช้ทดสอบคือ t-test โดย $T = r/S_r \sim t_{(n-2)}$

ในเมื่อ $S_r = \sqrt{(1-r^2)/(n-2)}$

ข้อสังเกต ข้อมูลจะสนับสนุน H_a : มีสหสัมพันธ์ระหว่าง X และ Y ขึ้นอยู่กับค่าของ $r >> 0$, และ n ต้องไม่เล็กเกินไป

(14) **สหสัมพันธ์แบบอันดับ** (Rank Correlation)

การศึกษา สหสัมพันธ์จากข้อมูล (x_i,y_i) จะต้องมีข้อสมมติว่า (x_i,y_i) มาจากประชากรที่มีการแจกแจงร่วมกันแบบ bivariate normal ถ้าไม่ใช้ข้อสมมตินี้ คือ ใช้วิธี Non-parameter หาสัมประสิทธิ์ r_S เรียกว่า Spearman rank correlation (1904) หรือใช้วิธีของ Kendall (1962) แต่วิธีของ Spearman ง่ายกว่า โดยมี

$$r_S = 1 - 6\Sigma d_i^2/n(n^2-1), \quad -1 \leq r \leq 1$$

d_i คือผลต่างอันดับของ x_i, y_i ค่า r_S จะใช้ทดสอบนัยสำคัญ $H_0 : \rho = 0$ โดยการเปรียบเทียบกับค่าจากตาราง Spearman หรือ ใช้ t-test ในข้อ 13

(15) **Multiple Regression and Correlation**

15.1 $Y_i = \alpha + \beta X_i$ เรียกว่า Simple linear regression, Y และ X คือ dependent และ independent variable ตามลำดับ

15.2 $Y_i = \alpha + \beta X_{1i} + \beta X_{2i}$ เรียกว่า Multiple regression, Y เป็น dependent variable ส่วน X_1, X_2 เป็น independent variable จำนวน 2 ตัว

15.3 $Y_i = \alpha + \beta X_i + \beta X_i + ... + \beta X_k$ คือ สมการ Multiple regression ทั่วๆ ไปที่มี Y เป็น dependent variable, X_1, X_2, ..., X_k เป็น independent variable **k** ตัว และ β_1, β_2, ..., β_k คือ partial regression coefficients

การหาสมการประมาณค่าของความถดถอยเชิงซ้อน นิยมใช้โปรแกรมคอมพิวเตอร์ ซึ่งจะแสดง printout 2 ส่วน ส่วนแรกจะแสดงรายละเอียดต่างๆ ของการสร้างสมการถดถอย นั่นคือ จะให้ค่า α, β, S_β , r^2, $S_{y/x}$ ฯลฯ และ t-test สำหรับ $H_0 : \beta_i = 0$ โดย $T = b_i/S_b \sim t_{(n-k-1)}$ และส่วนที่ 2 จะแสดง ANOVA และค่า F เพื่อทดสอบ $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ (สมการ $Y_i = \alpha + \beta_1 X_1 + ... + \beta_k X_k$ **ไม่เหมาะสม**) และ H_a : not all β's = 0 (สมการใน H_0 เป็นสมการที่เหมาะสม)

ตารางที่ 10.2 แสดง ANOVA ของ multiple regression

SOV	df	SS	MS	F	
Multiple regression	k	SSR	MSR	F= MSR/MSE	เปิดตารางที่
Residual or error	m-k-1	SSE	MSE		$f_{k,\,(n-k-1),\alpha}$
Total	n-1				

R^2 = SSR/SST = coefficient of determination for a multiple regression

$R = \sqrt{R^2}$ = multiple correlation coefficient

(16) **Polynomial Regression**

คือรูปแแบบหนึ่งของ multiple regression แต่ใช้ตัวแปรกำหนดเพียง 1 ตัว

$Y = \alpha + \beta_1 X$ คือสมการ polynomial degree ที่ 1 เรียกว่า linear (รูป (a))

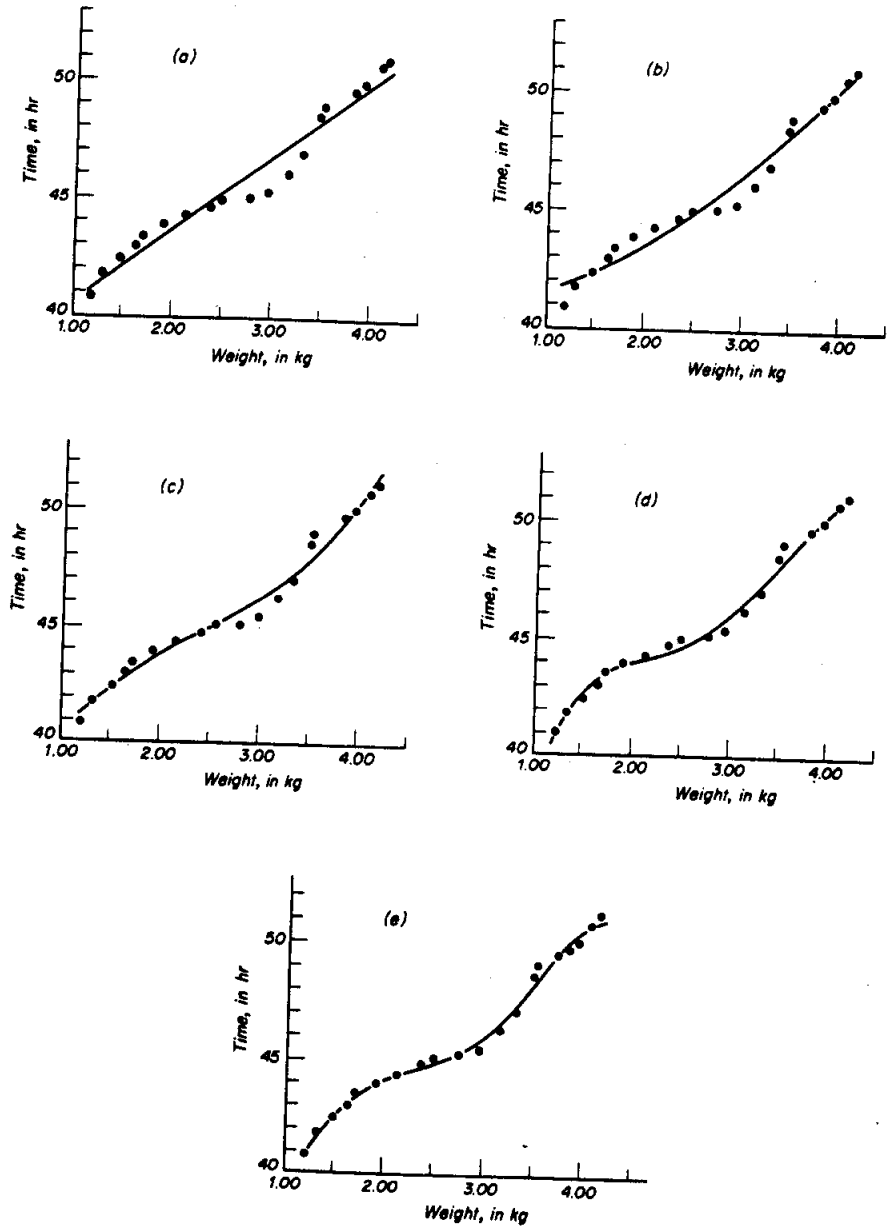
$Y = \alpha + \beta_1 X + \beta_2 X^2$ คือสมการ polynomial degree ที่ 2 เรียกว่า quadratic (รูป (b))

$Y = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$ คือสมการ polynomial degree ที่ 3 เรียกว่า cubic (รูป (c))

$Y = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4$ คือสมการ polynomial degree ที่ 4 เรียกว่า quartic (รูป (d))

$Y = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5$ คือสมการ polynomial degree ที่ 5 เรียกว่า quintic (รูป (e))

รูปที่ 10.3 เส้นทึบแสดงการนำสมการ polynomial degree ต่างๆ fit ข้อมูล จะเห็นว่า รูป (e) fit ข้อมูลได้ดีที่สุด



<u>ตัวอย่างที่ 1</u> ความยาวของปีกลูกนก 13 ตัว หลังจากฟักจากไข่ X วัน, X = อายุเป็นวัน
Y = ความยาวของปีกเป็นเซนติเมตร

นก		1	2	3	4	5	6	7	8	9	10	11	12	13
(วัน)	X:	3.0	4.0	5.0	6.0	8.0	9.0	10.0	11.0	12.0	14.0	15.0	16.0	17.0
(ซ.ม.)	Y:	1.4	1.5	2.2	2.4	3.1	3.2	3.2	3.9	4.1	4.7	4.5	5.2	5.0

$n = 13$, $\Sigma X = 130$, $\bar{X} = 10$, $\Sigma X^2 = 1562$, $\Sigma Y = 44.4$, $\bar{Y} = 3.42$, $\Sigma Y^2 = 171.3$, $\Sigma XY = 514.80$

$\Sigma(X-\bar{X})(Y-\bar{Y}) = \Sigma XY - (\Sigma X)(\Sigma Y)/n = 514.80 - (130)(44.4)/13 = 70.80$

$\Sigma(X-\bar{X})^2 = \Sigma X^2 - (\Sigma X)^2/n = 1562 - (130)^2/13 = 262$

$\Sigma(Y-\bar{Y})^2 = \Sigma Y^2 - (\Sigma Y)^2/n = 171.30 - (44.4)^2/13 = 19.6569$

---

(1) **จงหาสมการประมาณค่า** $Y_i = a + bX_i$

$b = \Sigma(X-\bar{X})(Y-\bar{Y})/\Sigma(X-\bar{X})^2 = 70.80/262 = .27$ ซ.ม./วัน

$a = Y - b\bar{X} = 3.42$ ซ.ม. $- (.27$ ซ.ม./วัน$)(10$ วัน$) = .72$ ซ.ม.

$Y_i = .72 + .27X_i$

(2) **จงพยากรณ์**ความยาวของปีก สำหรับลูกนกอายุ 13 วัน และอายุ 7 วัน

$\hat{Y}_{/X=13} = .72 + (.27$ ซ.ม./วัน$)(13$ วัน$) = 4.23$ ซ.ม.

$\hat{Y}_{/X=7} = .72 + (.27$ ซ.ม./วัน$)(7$ วัน$) = 2.61$ ซ.ม.

(3) **จงทดสอบ** regression coefficient

$H_0 : \beta = 0$ (อายุและความยาวของปีก**ไม่มี**ความสัมพันธ์แบบเชิงเส้น)

$H_a : \beta \neq 0$ (อายุและความยาวของปีกมีความสัมพันธ์แบบเชิงเส้น)

3.1 **ใช้ F-test**

SS(total) $= \Sigma(Y-\bar{Y})^2 = 19.6569$

SS(regression) = $\{\Sigma(X-\bar{X})(Y-\bar{Y})\}^2/\Sigma(X-\bar{X})^2 = (70.80)^2/262$

$= 19.1322$

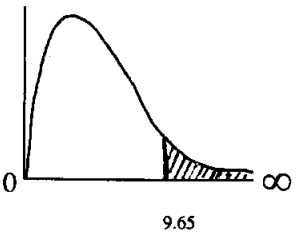
SS(residual หรือ error) = SST - SSR = 19.6569 - 19.1322

$= .5247$

ตารางที่ 10.3 แสดง ANOVA ของตัวอย่างที่ 1

Source	df	SS	MS	F
Regression	1	19.1322	19.1322	19.1322/.0477
Residual	11	.5247	.0477	= 401.1
Total	12	19.6569		

$f_{1,11,.01} = 9.65$



จึงปฏิเสธ $H_0 : \beta = 0$ ด้วย $p <<< .005$ และสรุปว่า ความยาวของปีกและอายุมีความสัมพันธ์แบบเชิงเส้น

3.2 **วิธี t-test**

$H_0 : \beta = 0$, $H_a : \beta \neq 0$, b = .27 ซ.ม./วัน

$S_b = \sqrt{S^2_{y/x}/\Sigma(X-\bar{X})^2} = \sqrt{.0477/262} = .0135$ ซ.ม./วัน

$T = b/S_b = .27/.0135 = 20 \sim t_{11}$

$t_{11,.0005} = 4.437$ จึงปฏิเสธ H_0 ด้วย p-value <<< .0005

ข้อสังเกต 1) ผลทดสอบ t-test และ F-test เมื่อใช้ α ระดับเดียวกัน ต้องเหมือนกัน

2) $\sqrt{F} = T$, $\sqrt{401.1} \approx 20$ (ความแตกต่างเนื่องจากการปัดเศษ)

(4) 95% ช่วงเชื่อมั่นของ $\beta = b \pm t_{.025,11}S_b$

$= .27 \pm (2.201)(.0135)$

$= .27 \pm .03 = .24, .30$ ซ.ม./วัน

(5) **การทดสอบ บางค่าของ** β ($\beta \neq 0$) เช่นต้องการทดสอบคำกล่าวว่าลูกนกที่มีอายุเพิ่มขึ้น 1 วัน จะมีปีกยาวขึ้น .25 ซ.ม. ? คือ $H_0 : \beta = .25$ ซ.ม., $H_a : \beta \neq .25$ ซ.ม., $T = (b - \beta^2)/S_b = (.27 - .25)/.0135 = 1.48$, $t_{.025,11} = 2.201$ ยังปฏิเสธ H_0 ไม่ได้ นั่นคือยอมรับคำกล่าวว่าเป็นจริง

ข้อสังเกต เมื่อใช้ $\alpha = .05$ การทดสอบในข้อ (5) สามารถสรุปผลได้จากช่วงเชื่อมั่น 95% ของ β จากข้อ (4) โดยพิจารณาว่า ถ้าช่วงเชื่อมั่นรวมค่าที่อ้าง ใน H_0 จะต้องยอมรับ H_0 แต่ถ้าไม่รวมค่าที่อ้างใน H_0 จะต้องปฏิเสธ H_0 และยอมรับ H_a ค่าที่อ้างใน H_0 คือ .25 เมื่อตรวจดูจากช่วงเชื่อมั่น $.24 \leq \beta \leq .30$ จะเห็นว่า .25 อยู่ในช่วงเชื่อมั่น จึงยอมรับว่าคำกล่าวใน H_0 เป็นจริง สำหรับค่าที่อ้าง ใน 3.1, 3.2 คือ $\beta = 0$ จะเห็นว่า 0 ไม่อยู่ในช่วงเชื่อมั่นนี้ จึงปฏิเสธ H_0 และยอมรับ H_a

(6) **ช่วงเชื่อมั่นของค่าพยากรณ์**

6.1 จากข้อ (2) ได้ $Y_{/x=13} = 4.23$ ซ.ม. ถ้าค่าพยากรณ์นี้ คือค่าเฉลี่ยความยาวของปีกลูกนกทั้งหลายที่มีอายุ 13 วัน จะมี

$$S_{\hat{y}} = \sqrt{S^2_{y/x}(1/n + (X-\bar{X})^2/\Sigma(X-\bar{X})^2}$$

$$= \sqrt{.0477(1/13 + (13-10)^2/262} = .073 \text{ ซ.ม.}$$

ดังนั้น 95% ช่วงเชื่อมั่นของ Y คือ $\hat{Y} \pm t_{.025,11} S_{\hat{y}} = 4.23 \pm (2.201)(.073)$

$$= 4.23 \pm .16 = 4.07, 4.39 \text{ ซ.ม.}$$

6.2 ถ้าต้องการพยากรณ์ค่าเดี่ยวๆ เพียงค่าเดียว คือ "สุ่มลูกนกมา 1 ตัว อายุ 13 วัน" ให้พยากรณ์ความยาวของลูกนกตัวนี้ จะมี

$$S_{\hat{y}} = \sqrt{S^2_{y/x}(1 + 1/n + (X-\bar{X})^2/\Sigma(X-\bar{X})^2} \text{ และ } \hat{Y} = 4.23 \text{ ซ.ม.}$$

$$= \sqrt{.0477(1 + 1/3 + (13-10)^2/262} = .23 \text{ ซ.ม.}$$

ดังนั้น 95% ช่วงเชื่อมั่นของ Y คือ $4.23 \pm (2.201)(.23) = 4.23 \pm 5.1 = 3.72, 4.74$ ซ.ม.

6.3 ถ้าสุ่มลูกนกอายุ 13 วันมาหลายๆ ตัว เช่น $m = 10$ ตัว จงพยากรณ์ความยาวของปีกลูกนก

กลุ่มนี้จะมี $\hat{Y}_i = 4.23$ ซ.ม. แต่

$$S_{\hat{y}} = \sqrt{S^2_{y/x}[1/m + 1/n + (X-\bar{X})^2/\Sigma(X-\bar{X})^2}$$

$= \sqrt{.0477(1/10 + 1/13 + (13-10)^2/262} = .10$ ซ.ม.

95% ช่วงเชื่อมั่น $= 4.23 \pm (2.201)(.10) = 4.23 \pm .22 = 4.01, 4.45$ ซ.ม.

(7) การทดสอบสมมติฐานเกี่ยวกับค่าพยากรณ์ Y เช่น

H_0 : ลูกนกอายุ 13 วัน จะมีปีกยาวโดยเฉลี่ยไม่เกิน 4 ซ.ม. หรือ $\mu \leq 4$ ซ.ม., $\alpha = .05$

H_a : ลูกนกอายุ 13 วัน จะมีปีกยาวโดยเฉลี่ยเกิน 4 ซ.ม. หรือ $\mu > 4$ ซ.ม.

$T = (\hat{Y} - \mu_0)/S_{\hat{y}} = (4.23 - 4)/.073 = 3.151 \sim t_{.05,11} = 1.796$

จะปฏิเสธ H_0 ด้วย p-value < .005 ($\hat{Y} = 4.23$ และ $S_{\hat{y}} = .073$ จากข้อ 6.1)

(8) **การพยากรณ์ตัวแปรกำหนด**

เช่นอยากทราบว่าลูกนกที่มีปีกยาว 4.5 ซ.ม. (Y_i) จะมีอายุกี่วัน (X_i)

$\hat{X}_i = (\hat{Y}_i - a)/b = (4.5 - .72)/.27 = 14$ วัน

(9) **coefficient of determination (r^2), และ coefficient of correlation (r)**

$r^2 = SSR/SST = 19.1322/19.6569 = .9733$ (SSR, SST จาก 3.1) หมายถึง 97.33% ของความผันแปรของความยาวของปีกลูกนก อธิบายได้ว่าเพราะลูกนกมีอายุที่ต่างกัน

$r = \sqrt{r^2} = \sqrt{.9733} = .9866$ (เนื่องจากทราบค่า r^2)

(ถ้าไม่ทราบค่า r^2, $r = \Sigma(X-\bar{X})(Y-\bar{Y})/ \sqrt{\Sigma(X-\bar{X})^2\Sigma(Y-\bar{Y})^2} = 70.80/ \sqrt{(262)(19.6569)} = .9866$)

(10) **การทดสอบนัยสำคัญของค่า r**

$H_0 : \rho = 0$, $H_a : \rho > 0$, $T = r/S_r$, $S_r = \sqrt{1-r^2/(n-2)}$

$S_r = (1 - .9733)/11 = .0493$, $T = .9866/.0493 \sim 20$

ปฏิเสธ H_0 ด้วย p-value <<< .0005 สรุปว่า $\rho > 0$

(11) **Rank correlation**

สมมติต้องการหาสหสัมพันธ์แบบอันดับ ระหว่างความยาวของปีก (X) เป็น ซ.ม. และความยาวของหาง (Y) เป็น ซ.ม. โดยมีข้อมูลจากนก 8 ตัว ดังนี้

นก :	1	2	3	4	5	6	7	8
X :	10.4	10.8	11.1	10.3	10.3	10.2	10.7	10.5
(อันดับ)	(4)	(7)	(8)	(1.5)	(3)	(1.5)	(6)	(5)
Y :	7.4	7.6	7.9	7.2	7.4	7.1	7.4	7.2
(อันดับ)	(5)	(7)	(8)	(2.5)	(5)	(1)	(5)	(2.5)

หาผลต่างของอันดับของ X และ Y

นก :	1	2	3	4	5	6	7	8
X :	4	7	8	1.5	3	1.5	6	5
Y :	5	7	8	2.5	5	1	5	2.5
d_i :	-1	0	0	-1	-2	.5	1	2.5

$\sum d_i^2 = -1^2 + 0^2 + ... + 2.5^2 = 13.5, n = 8$

$r_S = 1 - 6\sum d_i^2/n(n^2-1) = 1 - \quad 6(13.5)/8(64-1) \quad = .8393$

การตรวจสอบนัยสำคัญ ของ r_s ทำได้ 2 กรณีคือ ถ้ามีตารางของ Spearman จะเปิดที่ n = 8 จะได้ .0005 < p-value < .001 จึงปฏิเสธ H_0

หรือจะใช้ t-test จะได้

$T = r/S_r$, $S_r = \sqrt{(1 - (.8393)^2)/(6)} = .22195$

$T = .8393/.22195 = 3.78$, $t_{.005,6} = 3.707$

จึงปฏิเสธ H_0 ด้วย p-value < .005

หมายเหตุ : การจัดอันดับสำหรับค่าสังเกตที่เท่ากัน ต้องหา**ค่าเฉลี่ยของอันดับ** เช่น นกตัวที่ 4, 8 ค่าของ y เท่ากันคือ 7.2 ซึ่งจะต้องใช้ตำแหน่งที่ 2 และ 3 (ตำแหน่งที่ 1 คือ 7.1) ดังนั้นค่าเฉลี่ยคือ (2+3)/2 = 2.5

แบบฝึกหัดบทที่ 10

1. อัตรา oxygen consumption ของนก ณ ระดับอุณหภูมิต่างๆ มีดังนี้

Y = oxygen consumption ml/g/ช.ม.	5.2	4.7	4.5	3.6	3.4	3.1	2.7	1.8
X = อุณหภูมิ ($^{\circ}$C)	-18	-16	-10	-5	0	5	10	19

จงสร้างสมการถดถอย ทดสอบความสัมพันธ์เชิงเส้น และสร้างช่วงเชื่อมั่น 95% ของ β

2. จากข้อ 1. จงหาอัตรา oxygen consumption ของนกที่อยู่ในอุณหภูมิระดับ 15^{o}C และสร้าง 95% ช่วงเชื่อมั่นของค่าพยากรณ์

3. จากข้อ 1. ถ้าสุ่มนกมา 1 ตัว ซึ่งอยู่ ณ อุณหภูมิ 15^{o}C จงสร้าง 95% ช่วงเชื่อมั่นของอัตรา oxygen consumption

4. จงหาค่า r^2 และอธิบายความหมาย

5. จงทดสอบ $H_0 : \rho = 0, H_a : \rho \neq 0$

6. จงหา r_S และทดสอบนัยสำคัญ

เฉลยแบบฝึกหัดที่ 10

1. $n = 8$, $\Sigma X = -14$, $\bar{X} = -1.75$, $\Sigma X^2 = 1160$, $\Sigma(X-\bar{X})^2 = 1{,}135.5$

   $\Sigma Y = 29$, $\bar{Y} = 3.625$, $\Sigma Y^2 = 114.04$, $\Sigma(Y-\bar{Y})^2 = 8.915$

   $\Sigma XY = -150.40$, $\Sigma(X-\bar{X})(Y-\bar{Y}) = -99.65$

   $b = -99.65/1{,}135.5 = -0.0877586$

   $a = Y - b\bar{X} = 29 - (-.0877586)(-1.75) = 28.85$

   $Y = 28.85 - .08776X$, SSR = 8,745, SSE = 8.915 - 8.745 = .1698

   $S^2_{y/x} = .0283$, $S_b = \sqrt{.0283/1135.5} = .0049899$

   $H_0 : \beta = 0$, $H_a : \beta \neq 0$, $T = -.0877586/.0049899 = -17.6$

   reject H_0, p-value << .0005

   95% ช่วงเชื่อมั่นของ $\beta = -.0877 \pm 2.447(.0049899) = -.0999, -.0755$

2. $\hat{Y}_{/x=15\ C} = 28.85 - .08776(15) = 27.53$

   $S_{\hat{y}(mean)} = .1682\sqrt{1/8 + \{15 - (-1.75)\}^2/1135.5} \approx 2.82$

   95% ช่วงเชื่อมั่น $= 27.53 \pm 2.447(2.82) = 1.781, 37.25$

3. $\hat{Y}_{/x=15\ C} = 27.53$

   $S_{\hat{y}(individual)} = .16822\ \sqrt{1 + 1/8 + \{15 - (-1.75)\}^2/1135.5} \approx 2.82$

   95% ช่วงเชื่อมั่น $= 27.53 \pm 2.447(2.82) = 17.81, 37.25$

4. $r^2 = 8.745/8.915 = .9809 = 98.09\%$, อุณหภูมิ อธิบายความผันแปรของ oxygen consumption ได้ 98.09%

5. $H_0 : \rho = 0$, $H_a : \rho \neq 0$, $T = r/S_r$, $r = \sqrt{.9809} = -.9904$

   $S_r = \sqrt{(1-r^2)/(n-2)} = .0564$, $T = -.9904/.0564 = -17.6$, reject H_0, p-value << .0005

6. $r_s = 1 - 6\Sigma d_i^2/n(n^2-1)$, $\Sigma d_i^2 = 168$, $r_s = -1$, ปฏิเสธ H_0 ด้วย p-value << .0005