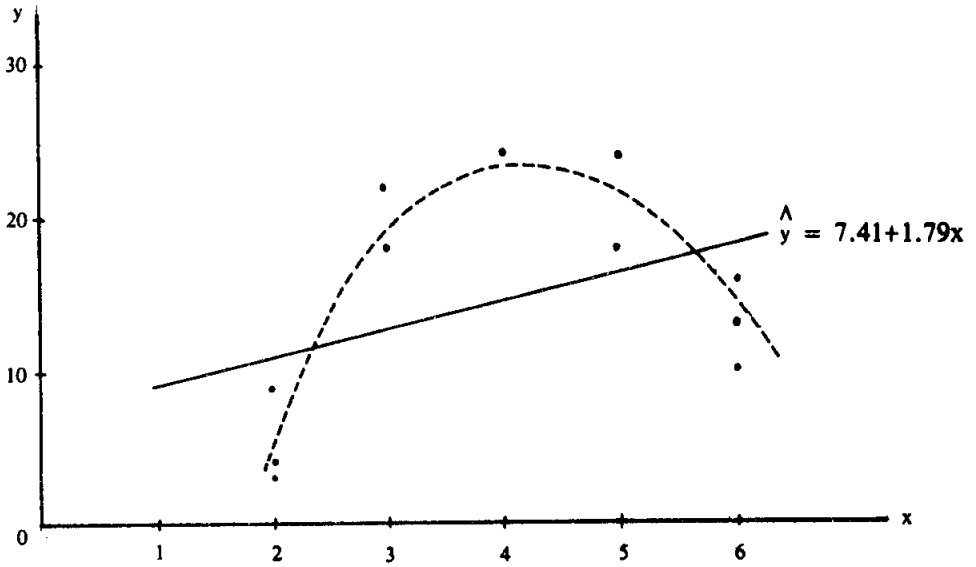


รูปที่ 5.5 แสดง sample regression line และเส้นแสดงความสัมพันธ์ในรูปเส้นโค้ง ซึ่งดูเข้ากับจุดทั้ง 11 จุดมากกว่าเส้นตรง

หมายเหตุ

ในการทำ Regression Analysis เราควร plot scatter diagram เป็นอย่างแรกเพื่อเป็นแนวทางในการตั้งตัวแบบที่จะใช้แสดงความสัมพันธ์ระหว่าง X และ Y แต่ตัวอย่างนี้ต้องการใช้เป็นตัวอย่างในการทดสอบ lack of fit จึง plot scatter diagram ทีหลังเพื่อตรวจสอบกัน

ตัวอย่างที่ 5.9 ปฏิกิริยาทางเคมีที่วัดได้ ณ อุณหภูมิต่าง ๆ กันมีดังนี้

x(°C)	y(%)	x(°C)	y(%)
150	77.4	250	88.9
150	76.7 } T_1 = 232.3	250	89.2 } T_3 = 267.8
150	78.2	250	89.7
200	84.1	300	94.8
200	84.5 } T_2 = 252.3	300	94.7 } T_4 = 285.4
200	83.7	300	95.9

จงหา sample regression line และทดสอบ lack of fit

จากข้อมูลข้างต้น n = 12

$\Sigma x_i = 2700$, $\Sigma x_i^2 = 645000$, $\bar{x} = 225$

$\Sigma y_i = 1037.8$, $\Sigma y_i^2 = 90265.52$, $\bar{y} = 86.4833$

$\Sigma x_i y_i = 237875$

$S_{xx} = 645000 - (2700)^2/12 = 37500$

$S_{yy} = 90265.52 - (1037.8)^2/12 = 513.1167$

$S_{xy} = 237875 - (2700)(1037.8)/12 = 4370$

$$b_1 = \hat{\beta}_1 = \frac{4370}{37500} = .1165$$

$$b_0 = \hat{\beta}_0 = 86.4833 - (.1165)(225) = 60.2708$$

ดังนั้น sample regression line คือ $\hat{y} = 60.2708 + .1165x$

จะทดสอบ 1) $H_0 : \beta_1 = 0$ และ

2) H_0 : ไม่มี lack of fit

คำนวณ $SST = S_{yy} = 513.1167$

$SSR = b_1 S_{xy} = (.1165)(4370) = 509.1050$

$SSE = SST - SSR = 4.0117$

คำนวณ SS(p.e.) โดยหา $T_1 = 232.3$, $T_2 = 252.3$, $T_3 = 267.8$ และ $T_4 = 285.4$

$$SS(p.e.) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{k} \frac{(\sum_{j=1}^{n_i} y_{ij})^2}{n_i}$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{k} \frac{T_i^2}{n_i}$$

$$= 90265.52 - \frac{(232.3)^2 + (252.3)^2 + (267.8)^2 + (285.4)^2}{3}$$

$$= 2.66$$

SS(lof) = SSE − SS(p.e.)

= 4.0117 − 2.66 = 1.3517

ANOVA

S.V.	d.f.	SS	MS	f_c
Regression	1	509.1050	509.1050	f_1 = 1531.60**
Error	10	4.0117	.40117	
Lack of fit	2	1.3517	.6758	f_2 = 2.03 (n.s.)
Pure error	8	2.6600	.3324	
Total	11	513.1167		

ทดสอบ 1) $H_0 : \beta_1 = 0$

2) $H_1 : \beta_1 \neq 0$

3) $\alpha = .05$

4) CR : $F > f_{(1,10),.05} = 4.9646$

5) $f_c = f_1 = 1531.60$

6) $f_c > 4.9646$ เราปฏิเสธ H_0 ที่ $\alpha = .05$ (เราปฏิเสธ H_0 ที่ $\alpha = .01$ ด้วย)

นั่นคือ มีความสัมพันธ์ระหว่าง X และ Y

ทดสอบ 1) H_0 : ไม่มี lack of fit

2) H_1 : มี lack of fit

3) $\alpha = .05$

4) CR : $F > f_{(2,8),.05} = 4.459$

5) $f_c = f_2 = 2.03$

6) $f_c < 4.459$ เราไม่สามารถปฏิเสธ H_0 ที่ $\alpha = .05$ ได้ นั่นคือไม่มี lack of fit

ถ้าคำนวณ $r^2 = \frac{SSR}{SST} = .9922$

$100r^2\% = 99.22\%$ ของความแปรปรวนของ Y เนื่องมาจากความสัมพันธ์กับ X ในรูปเชิงเส้นตรง

สรุปได้ว่า ตัวแบบเชิงเส้นตรงที่ใช้อยู่เหมาะสมดีแล้ว

5.7 สหสัมพันธ์ (Correlation) : เครื่องมือวัดความสัมพันธ์เชิงเส้นตรง

เมื่อเรามีตัวแปรเชิงสุ่ม Y ซึ่งขึ้นอยู่กับ X ที่ถูกควบคุมโดยผู้ทำการทดลอง และทำการวิเคราะห์เพื่อดูอิทธิพลของ X หรือประโยชน์ของมันในการทำนายค่า Y เราจะใช้วิธีการของ regression แต่ผู้ทำการทดลองอาจตั้งจุดมุ่งหมายที่จะทราบความสัมพันธ์ของตัวแปรเชิงสุ่ม 2 ตัว โดยที่เราไม่สนใจว่าตัวใดเป็นเหตุและตัวใดเป็นผล เราจะใช้วิธีการของ correlation

โครงสร้างของข้อมูลประกอบด้วย ตัวอย่างสุ่มจาก Bivariate Normal distribution ขนาด n : $(X_1, Y_1), \ldots, (X_n, Y_n)$ และต่างคู่กันจะเป็นอิสระต่อกัน

5.7.1 Bivariate Normal distribution

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}Q}, \quad -\infty < x < \infty \text{ และ } -\infty < y < \infty$$

$$\text{โดยที่ } Q = \left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2$$

ρ คือสัมประสิทธิ์สหสัมพันธ์ของประชากร (Population correlation coefficient)

$$\rho = \text{corr}(X, Y)$$
$$= \frac{\text{cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$$

5.7.2 พิสูจน์ว่า $-1 \leq \rho \leq 1$

ให้ $\rho = \text{corr}(X, Y)$

ถ้า standardized random variable 2 ตัวคือ

$$Z_1 = \frac{X - \mu_X}{\sigma_X} = \frac{X}{\sigma_X} - \frac{\mu_X}{\sigma_X} \qquad Z_2 = \frac{Y - \mu_Y}{\sigma_Y} = \frac{Y}{\sigma_Y} - \frac{\mu_Y}{\sigma_Y}$$

$corr(Z_1, Z_2) = \rho$ (เพราะว่าสัมประสิทธิ์ของ X และ Y ใน Z_1 และ Z_2 มีเครื่องหมายเหมือนกัน)

$\because$ Variance ของตัวแปรเชิงสุ่มใด ๆ ≥ 0

$$Var(Z_1 - Z_2) = Var(Z_1) + Var(Z_2) - 2\,Cov(Z_1, Z_2) \geq 0$$

$$Var(Z_1) + Var(Z_2) \geq 2\,Cov(Z_1, Z_2)$$

$$1 + 1 \geq \frac{2\,Cov(Z_1, Z_2)}{\sigma_{Z_1}\sigma_{Z_2}}$$

$$1 \geq Corr(Z_1, Z_2) = \rho \qquad \text{.....(1)}$$

$$Var(Z_1 + Z_2) = Var(Z_1) + Var(Z_2) + 2\,Cov(Z_1, Z_2) \geq 0$$

$$Var(Z_1) + Var(Z_2) \geq -2\,Cov(Z_1, Z_2)$$

$$1 + 1 \geq -2\rho$$

$$-1 \leq \rho \qquad \text{.....(2)}$$

จาก (1) และ (2) สรุปได้ว่า $-1 \leq \rho \leq 1$

5.7.3 สัมประสิทธิ์สหสัมพันธ์ของตัวอย่าง (Sample correlation coefficient หรือ Pearson's product moment)

$r = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$, โดยที่ S_{xx}, S_{yy} และ S_{xy} ได้กล่าวถึงในหัวข้อ 5.2

R เป็น Unbiased estimator ของ ρ ซึ่งมี mean $E(R) = \rho$ และ variance $Var(R) = \left(\dfrac{1 - \rho^2}{n - 2}\right)$

คุณสมบัติของ r (ค่าของตัวแปรเชิงสุ่ม R)

1) $-1 \leq r \leq 1$

$r = 1$ เกิดขึ้นเมื่อจุดทั้ง n จุดอยู่บนเส้นตรงที่มี slope เป็น +

$r = -1$ เกิดขึ้นเมื่อจุดทั้ง n จุดอยู่บนเส้นตรงที่มี slopeเป็น

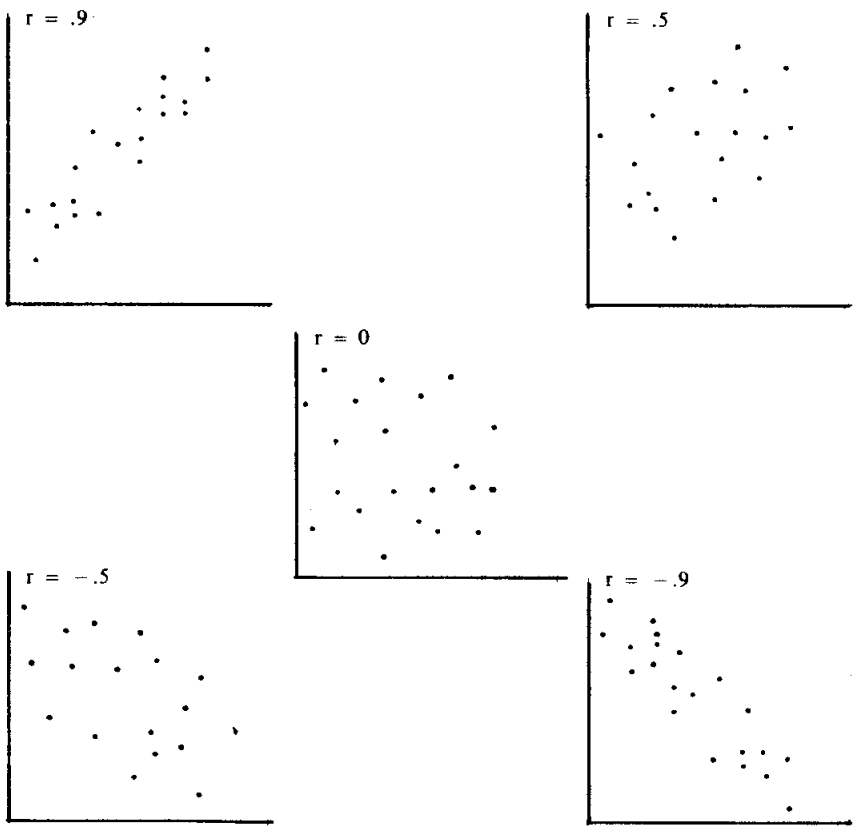
$r = \pm 1$ แสดงว่า X และ Y มีความสัมพันธ์เชิงเส้นตรงอย่างสมบูรณ์

$r = 0$ หมายความว่า X และ Y ไม่มีความสัมพันธ์กันในเชิงเส้นตรง แต่อาจมีความสัมพันธ์กันในรูปอื่น หรือไม่มีความสัมพันธ์กันเลย

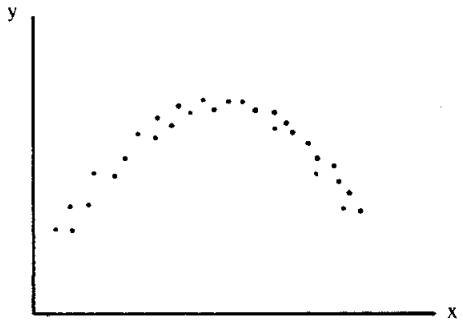
2) ค่าของ r ใช้วัดอัตราความมากน้อยของความสัมพันธ์เชิงเส้นตรง เครื่องหมายของ r บอกทิศทางของความสัมพันธ์ นั่นคือ ถ้า r เป็น + ค่าของตัวหนึ่งเพิ่ม ค่าของอีกตัวหนึ่งจะเพิ่มขึ้นด้วย และถ้า r เป็น – ค่าของทั้ง 2 ตัว จะมีทิศทางตรงกันข้าม

3) r^2 เป็นอัตราส่วนของความแปรปรวนของ Y ที่อธิบายได้ด้วย regression line ที่คำนวณได้ด้วยวิธีการของ least square method หรืออาจกล่าวว่าอัตราส่วนของความแปรปรวนของ Y ที่เนื่องมาจากอิทธิพลของความสัมพันธ์เชิงเส้นตรงกับ X

4) $Corr(aX + b, cY + d) = Corr(X, Y)$ ถ้า a และ c มีเครื่องหมายเดียวกัน



รูปที่ 5.6 แสดงค่าของ r และความมากน้อยของการกระจัดกระจายของจุด



รูปที่ 5.7 แสดงความสัมพันธ์ของ X และ Y ในรูปเส้นโค้ง ถ้าคำนวณค่า r, ค่าเกือบเป็นศูนย์

5.8 ความสัมพันธ์ระหว่าง Sample regression coefficient (b_1) และ Sample correlation coefficient (r)

$$b_1 = \frac{S_{xy}}{S_{xx}}$$

$$r = \frac{b_1\sqrt{S_{xx}}}{\sqrt{S_{yy}}}$$ นั่นคือ r และ b_1 จะมีเครื่องหมายเหมือนกัน

$$r^2 = \frac{b_1^2 S_{xx}}{S_{yy}}$$

$$\because \quad SSE = SST - SSR$$

$$= S_{yy} - b_1 S_{xy}$$

$$= S_{yy} - b_1^2 S_{xx}$$

$$\frac{SSE}{S_{yy}} = 1 - \frac{b_1^2 S_{xx}}{S_{yy}}$$

$$b_1^2 \frac{S_{xx}}{S_{yy}} = 1 - \frac{SSE}{S_{yy}}$$

$$r^2 = 1 - \frac{SSE}{S_{yy}}$$

$$= \frac{S_{yy} - SSE}{S_{yy}}$$

$$= \frac{SSR}{SST}$$

ถ้า SSE = 0, $r^2 = 1$ นั่นคือ $r = -1$ หรือ $r = +1$ ซึ่งกรณีนี้เกิดเมื่อจุดทุกจุดอยู่บนเส้นตรง

นั่นคือ ถ้า SSE $\to 0$, $r^2 \to 1$ นั่นคือ ถ้าค่า r^2 ยิ่งสูง ค่า SSE จะยิ่งน้อย เราจึงใช้ r^2 เป็นมาตรการช่วยชี้ว่าความแปรปรวนของ Y ที่เนื่องมาจากอิทธิพลของความสัมพันธ์เชิงเส้นตรงกับ X มีมากน้อยเท่าใด

5.9 การทดสอบสมมติฐานและการหาช่วงความเชื่อมั่นเกี่ยวกับ ρ

5.9.1 ทดสอบ $H_0 : \rho = 0$, $H_1 : \rho \neq 0$

ตัวสถิติที่ใช้ในการทดสอบ คือ $T = \dfrac{R - \rho}{\sqrt{\dfrac{1 - R^2}{n - 2}}} \sim t_{n-2}$

ถ้า H_0 จริง $t_c = \dfrac{r\sqrt{n-2}}{\sqrt{1 - r^2}}$ เป็นค่าของตัวสถิติ T ซึ่งมีการกระจายเป็น t-distribution ที่มี d.f. $= n - 2$

5.9.2 ทดสอบ $H_0 : \rho = \rho_0 (\rho_0 \neq 0)$

$\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ เป็นค่าของตัวแปรเชิงสุ่ม ซึ่งมีการกระจายเป็น Normal distribution ที่มี mean $\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ และ variance $\frac{1}{n-3}$

(ln = log ฐาน e = 2.3026 $\log_{10}$)

ถ้า H_0 จริง $z_c = \dfrac{\sqrt{n-3}}{2}\left[\ln\left(\dfrac{1+r}{1-r}\right) - \ln\left(\dfrac{1+\rho_0}{1-\rho_0}\right)\right]$

$= \dfrac{\sqrt{n-3}}{2}\ln\left[\dfrac{(1+r)(1-\rho_0)}{(1-r)(1+\rho_0)}\right]$ จะเป็นค่าของ $Z \sim N(0, 1)$

ตาราง A10 เราสามารถอ่านค่าของ $z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ เราจึงควรใช้สูตร

$$z_c = \frac{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\sqrt{\frac{1}{n-3}}}$$

$$= \frac{z_r - z_{\rho_0}}{\sqrt{\frac{1}{n-3}}}$$

โดยที่ $z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ และ $z_{\rho_0} = \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)$

5.9.3 การหา $100(1-\alpha)\%$ C.I. ของ ρ

จาก $Z = \frac{z_r - z_\rho}{\sqrt{\frac{1}{n-3}}} \sim N(0, 1)$

$$\Pr\left[-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

$$\Pr\left[-z_{\frac{\alpha}{2}} \leq \frac{z_r - z_\rho}{\sqrt{\frac{1}{n-3}}} \leq z_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

$$\Pr\left[z_r - z_{\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}} \leq z_\rho \leq z_r + z_{\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}}\right] = 1 - \alpha$$

$100(1-\alpha)\%$ C.I. ของ z_ρ คือ $z_r \pm z_{\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}}$

ในการหา C.I. ของ ρ อาจทำได้โดยใช้ตาราง A10 เพื่อเปลี่ยนค่า $z_r - z_{\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}}$ และ $z_r + z_{\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}}$ ไปเป็น correlation coefficients หรืออาจใช้ Hyperbolic tangent function จะได้ $100(1-\alpha)\%$ C.I. ของ ρ คือ

$$\left[\tanh\left(z_r - z_{\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}}\right), \tanh\left(z_r + z_{\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}}\right)\right]$$

ตัวอย่างที่ 5.10 ในการศึกษาถึงความสัมพันธ์ของปริมาณน้ำฝนที่ตกลงมากับปริมาณอากาศเสีย (air pollution) ที่จางหายไปหลังจากฝนตก ได้ผลดังนี้

x : ปริมาณฝนตกต่อวัน (.01 นิ้ว)	4.3	4.5	5.9	5.6	6.1	5.2	3.8	2.1	7.5
y : ปริมาณของอากาศสกปรกที่หายไป (ไมโครกรัม/ตารางเมตร)	126	121	116	118	114	118	132	141	108

$n = 9, \quad \Sigma x_i = 45 \quad , \Sigma x_i^2 = 244.26 \quad , \bar{x} = 5$

$\Sigma y_i = 1094 \quad , \Sigma y_i^2 = 133786 \quad , \bar{y} = 121.56$

$\Sigma x_i y_i = 5348.2$

$S_{xx} = 19.26$

$S_{yy} = 804.22223$

$S_{xy} = -121.8$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{-121.8}{\sqrt{(19.26)(804.22223)}}$$

$$= -.9786$$

(ถ้าหาค่า b_0 , $b_1 = -6.3240$)

1) $H_0 : \rho = 0$

2) $H_1 : \rho \neq 0$

3) $\alpha = .05$

4) CR : $T < -t_{7, .025} = -2.365$ และ $T > 2.365$

5) $t_c = \dfrac{-.9786\sqrt{7}}{\sqrt{1-(-.9786)^2}} = -12.9032$

6) $t_c < -2.365$ (ตกใน CR) เราปฏิเสธ H_0 ที่ $\alpha = .05$ นั่นคือ X และ Y มีความสัมพันธ์กันเชิงเส้นตรง

ตัวอย่างที่ 5.11 จากการศึกษาถึงความสัมพันธ์ของคะแนน Midterm กับคะแนนสอบไล่ในวิชาสถิติเบื้องต้นว่ามีความสัมพันธ์กันหรือไม่ ได้สุ่มนักศึกษามา 20 คนได้คะแนนดังนี้

x : คะแนน Midterm	35	60	55	35	35	50	30	60	50	20
y : คะแนนสอบไล่	60	80	60	80	75	90	60	105	60	30

x : คะแนน Midterm	55	45	40	60	40	60	50	55	50	35
y : คะแนนสอบไล่	90	75	80	80	45	80	80	95	100	75

$n = 20,\ \Sigma x_i = 920 \quad , \Sigma x_i^2 = 44900$

$\Sigma y_i = 1500 \quad , \Sigma y_i^2 = 118850$

$\Sigma x_i y_i = 71600$

$S_{xx} = 2580$

$S_{yy} = 6350$

$S_{xy} = 2600$

ก. 1) $H_0 : \rho = .9$

2) $H_1 : \rho \neq .9$

3) $\alpha = .05$

4) CR : $Z < -1.96$ และ $Z > 1.96$

5) $r = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \dfrac{2600}{\sqrt{(2580)(6350)}} = .642$

จากตาราง A10 : $z_r = z_{.642} = .7616$

$z_{\rho_0} = z_{.9} = 1.4722$

$$z_C = \frac{.7616 - 1.4722}{\sqrt{1/17}} = -2.93$$

หรือ $$z_C = \frac{\sqrt{n-3}}{2} \ln \left[\frac{(1+r)(1-\rho_0)}{(1-r)(1+\rho_0)} \right]$$

$$= \frac{\sqrt{17}}{2} \ln \left[\frac{(1.642)(.1)}{(.358)(1.9)} \right]$$

$$= \frac{\sqrt{17}}{2} \ln (.2413996)$$

$$= (2.0615528)(-1.4213017)$$

$$= -2.930085 \approx -2.93$$

6) $z_c = -2.93 < -1.96$ เราปฏิเสธ H_0 ที่ $\alpha = .05$

ข. หา 90% C.I. ของ ρ

ก่อนอื่นหา 90% C.I. ของ $z_\rho = z_r \pm z_{\frac{\alpha}{2}} \sqrt{\dfrac{1}{n-3}}$

$$= .7616 \pm \frac{1.645}{\sqrt{17}}$$

$$= .7616 \pm (1.645)(.2425)$$

$$= (.3627, 1.1605)$$

1) ใช้ตาราง A10 เปลี่ยนค่า .3627 และ 1.1605 เป็น correlation coefficients

$\therefore$ 90% C.I. ของ ρ คือ (.348,.821)

2) ใช้ Hyperbolic tangent function

90% C.I. ของ ρ คือ (tanh (.3627), tanh (1.1605))

$$= (.34759, .8212028) = (.348, .821)$$

5.10 การทดสอบการเท่ากันของ correlation coefficient 2 ตัว ($H_0 : \rho_1 = \rho_2$)

โครงสร้างของข้อมูล

x_{1i}	x_{11}	x_{12}	...	x_{1n_1}
y_{1i}	y_{11}	y_{12}	...	y_{1n_1}

คำนวณ $r_1 = \dfrac{S_{xy}^{(1)}}{\sqrt{S_{xx}^{(1)} S_{yy}^{(1)}}}$

x_{2i}	x_{21}	x_{22}	...	x_{2n_2}
y_{2i}	y_{21}	y_{22}	...	y_{2n_2}

คำนวณ $r_2 = \dfrac{S_{xy}^{(2)}}{\sqrt{S_{xx}^{(2)} S_{yy}^{(2)}}}$

1) $H_0 : \rho_1 = \rho_2$

2) $H_1 : \rho_1 \neq \rho_2$

3) กำหนด α

4) CR : $Z > z_{\alpha/2}$ และ $Z < -z_{\alpha/2}$

5) $z_c = \dfrac{z_1 - z_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$

โดยที่ $z_1 = \dfrac{1}{2} \ln\left(\dfrac{1+r_1}{1-r_1}\right)$,

$z_2 = \dfrac{1}{2} \ln\left(\dfrac{1+r_2}{1-r_2}\right)$,

$\sigma_1^2 = \dfrac{1}{n_1 - 3}$ และ $\sigma_2^2 = \dfrac{1}{n_2 - 3}$

6) สรุปผล

ตัวอย่างที่ 5.12 ในการศึกษาถึงความแตกต่างของสัมประสิทธิ์สหสัมพันธ์ของคะแนนวิชาคำนวณ (X) และคะแนนวิชาภาษาอังกฤษ (Y) ของนักเรียนจาก 2 โรงเรียน ได้ทำการสุ่มนักเรียนชั้น ม.1 12 คนจากโรงเรียนที่ 1 และ 10 คนจากโรงเรียนที่ 2 ผลปรากฏดังนี้

โรงเรียนที่ 1	x_{1i}	41	39	53	67	61	67	46	50	55	72	63	59
	y_{1i}	29	19	30	27	28	27	22	29	24	33	25	20
โรงเรียนที่ 2	x_{2i}	56	38	52	40	65	61	64	64	53	51		
	y_{2i}	34	21	25	24	32	29	27	26	24	25		

จงทดสอบที่ α = .05 ว่าสัมประสิทธิ์สหสัมพันธ์ (ของคะแนนทั้ง 2 วิชา) ของ 2 โรงเรียนนี้ไม่แตกต่างกัน

$n_1 = 12, \Sigma x_{1i} = 673 \quad , \Sigma x_{1i}^2 = 38985$

$\Sigma y_{1i} = 313 \quad , \Sigma y_{1i}^2 = 8359$

$\Sigma x_{1i} y_{1i} = 17759$

$S_{xx}^{(1)} = 1240.9167$

$S_{yy}^{(1)} = 194.9167$

$S_{xy}^{(1)} = 204.9167$

$$r_1 = \frac{S_{xy}^{(1)}}{\sqrt{S_{xx}^{(1)} S_{yy}^{(1)}}} = .417$$

จากตาราง A10 : $z_1 = z_{.417} = .4441$

$n_2 = 10, \Sigma x_{2i} = 544 \quad , \Sigma x_{2i}^2 = 30432$

$\Sigma y_{2i} = 267 \quad , \Sigma y_{2i}^2 = 7269$

$\Sigma x_{2i} y_{2i} = 14750$

$S_{xx}^{(2)} = 838.4$

$S_{yy}^{(2)} = 140.1$

$S_{xy}^{(2)} = 225.2$

$$r_2 = \frac{S_{xy}^{(2)}}{\sqrt{S_{xx}^{(2)} S_{yy}^{(2)}}} = .657$$

จากตาราง A10 : $z_2 = z_{.687} = .7875$

1) $H_0 : \rho_1 = \rho_2$

2) $H_1 : \rho_1 \neq \rho_2$

3) $\alpha = .05$

4) CR : $Z < -z_{.025} = -1.96$ และ $Z > 1.96$

5) $$z_c = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

$$= \frac{.4441 - .7875}{\sqrt{\frac{1}{9} + \frac{1}{7}}}$$

$$= \frac{-.3434}{.5039526} = -.6814$$

6) $\because z_c = -.6814$ **ไม่ตกอยู่ใน CR เราไม่สามารถปฏิเสธ H_0 ที่ $\alpha = .05$ ได้นั่นคือสัมประสิทธิ์สหสัมพันธ์ (ของคะแนนทั้ง 2 วิชา) ของ 2 โรงเรียนนี้ไม่แตกต่างกันอย่างมี**นัยสำคัญ

5.11 Serial correlation

เมื่อเราสังเกตค่าของตัวแปรตัวหนึ่ง ณ เวลาที่ต่าง ๆ กัน สิ่งสำคัญที่เราควรพิจารณาคือโอกาสที่จะเกิดความสัมพันธ์ระหว่างค่าสังเกตที่ติดตามกันมา หรืออาจกล่าวได้ว่าขาดคุณสมบัติของการสุ่ม (randomness) คำว่า "serial correlation" นี้หมายถึงความสัมพันธ์ระหว่างค่าสังเกตที่ติดต่อกันตามลำดับเวลา หรือตามลำดับอื่น ๆ ของการเก็บรวบรวมข้อมูล

กำหนดให้ $x_1, \ldots, x_n$ แทนค่าสังเกตของตัวแปร X ณ เวลา n จุดที่ต่อเนื่องกัน เพื่อที่จะตรวจดูว่าสมาชิกที่อยู่ต่อเนื่องกันนั้นมีความสัมพันธ์กันหรือไม่ เราจับคู่ค่าสังเกตที่อยู่ติดกันคือ (x_{i-1}, x_i) โดยที่ $i = 2, \ldots, n$ ดังนั้นเราจะได้ $(n-1)$ คู่ ดังนี้คือ

$$(x_1, x_2), (x_2, x_3), \ldots, (x_{i-1}, x_i), \ldots, (x_{n-1}, x_n)$$

เพื่อช่วยการสำรวจในขั้นต้น เราอาจสร้าง Scatter diagram ซึ่งประกอบด้วย $(n-1)$ จุดข้างต้น โดยให้สมาชิกตัวแรกของ (x_{i-1}, x_i) เป็นค่าบนแกน x และสมาชิกตัวที่สองเป็นค่าบนแกน y

สัมประสิทธิ์สหสัมพันธ์ซึ่งคำนวณจากค่าสังเกต $(n-1)$ คู่ข้างบนเราเรียกว่า first serial correlation หรือ first-order autocorrelation หรือ lag 1 correlation

Lag 1 correlation หรือ first-order autocorrelation

$$r_1 = \frac{\sum_{i=2}^{n} (x_{i-1} - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

พิจารณา $\sum_{i=2}^{n} (x_{i-1} - \bar{x})(x_i - \bar{x}) = \sum_{i=2}^{n} (x_{i-1}x_i - x_{i-1}\bar{x} - x_i\bar{x} + \bar{x}_2)$

$$= \sum_{i=2}^{n} x_{i-1}x_i - \bar{x}\left[\sum_{i=2}^{n} x_{i-1} + \sum_{i=2}^{n} x_i\right] + (n-1)\bar{x}^2$$

$$= \sum_{i=2}^{n} x_{i-1}x_i - \bar{x}\left[\sum_{i=1}^{n-1} x_i + \sum_{i=2}^{n} x_i\right] + (n-1)\bar{x}^2$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n$$

ดังนั้นในการคำนวณโดยใช้เครื่องคิดเลขแนะนำให้ใช้สูตร

$$r_1 = A/B$$

โดยที่ $A = \sum_{i=2}^{n} x_{i-1}x_i - \bar{x}\left[\sum_{i=1}^{n-1} x_i + \sum_{i=2}^{n} x_i\right] + (n-1)\bar{x}^2$

และ $B = \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n$

r_1 เป็นค่าค่าหนึ่งของตัวแปรเชิงสุ่ม R_1 ซึ่งเป็นตัวประมาณค่า (Estimator) ของ ρ_1 (population first-order autocorrelation) ถ้าตัวอย่างมีขนาดใหญ่ การทดสอบสมมติฐาน H_0 : $\rho_1 = 0$ (ข้อมูลไม่มีความสัมพันธ์กันในตัวข้อมูลเอง) เราจะใช้ความจริงที่ว่าเราอาจประมาณการแจกแจงของ $\sqrt{n}\, R_1$ เป็น $N(0, 1)$

สรุปขั้นในการทดสอบสมมติฐานเกี่ยวกับ ρ_1 (เมื่อตัวอย่างมีขนาดใหญ่)

1) $H_0 : \rho_1 = 0$

2) H_1 : 1) $\rho_1 \neq 0$

2) $\rho_1 > 0$

3) $\rho_1 < 0$

3) กำหนด α

4) CR : 1) $Z > z_{\alpha/2}$ และ $Z < -z_{\alpha/2}$

2) $Z > z_{\alpha}$

3) $Z < -z_{\alpha}$

5) $z_c = \sqrt{n}\, r_1$

6) สรุป

ก) ถ้า z_c ตกใน CR. เราปฏิเสธ H_0 ที่ระดับนัยสำคัญ $= \alpha$ นั่นคือข้อมูลชุดนี้มีความสัมพันธ์กันในตัวข้อมูลเอง

ข) ถ้า z_c ไม่ตกอยู่ใน CR. เราไม่ปฏิเสธ H_0 ที่ระดับนัยสำคัญ $= \alpha$ นั่นคือข้อมูลชุดนี้ไม่มีความสัมพันธ์กันในตัวข้อมูลเอง หรือข้อมูลมีคุณสมบัติของการสุ่ม (randomness)

ตัวอย่างที่ 5.13 โรงงานผลิตแผ่นพลาสติกพยายามที่จะควบคุมน้ำหนักของแผ่นพลาสติกให้ใกล้เคียงกับ 125 ปอนด์ต่อหนึ่งหน่วยพื้นที่ ข้อมูลทั้ง 28 ค่าต่อไปนี้ซึ่งเป็นค่าที่ต่อเนื่องกันนั้นได้มาจากการหาค่าน้ำหนักที่ต่างไปจาก 125

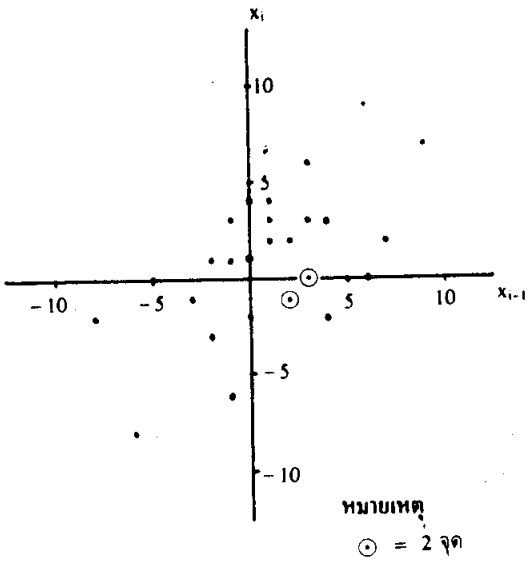
ค่าสังเกตตัวที่	ค่าสังเกต									
1 – 10	6	9	7	2	–1	–6	–8	–2	1	3
11 – 20	0	1	4	3	3	0	4	–2	–3	–1
21 – 28	1	2	2	–1	3	6	0	–2		

ก) จงสร้าง Scatter diagram

ข) จงทดสอบ $H_0 : \rho_1 = 0$, $H_1 : \rho_1 \neq 0$ ที่ $\alpha = .05$

ก) 27 จุดที่จะใช้สร้าง Scatter diagram คือ

(6, 9), (9, 7), (7, 2), ..., (6, 0), (0, −2)



ข) $n = 28, \sum_{i=1}^{n} x_i = 31 , \sum_{i=1}^{n} x_i^2 = 409 , \bar{x} = 1.1071$

$$\sum_{i=1}^{n-1} x_i = \sum_{i=1}^{n} x_i - x_n = 31-(-2) = 33$$

$$\sum_{i=2}^{n} x_i = \sum_{i=1}^{n} x_i - x_1 = 31-6 = 25$$

$$\sum_{i=2}^{n} x_{i-1}x_i = (6)(9)+(9)(7)+\ldots+(6)(0)+(0)(-2) = 244$$

$$\sum_{i=2}^{n} x_{i-1}x_i - \bar{x}\left[\sum_{i=2}^{n} x_i + \sum_{i=1}^{n-1} x_i\right] + (n-1)\bar{x}^2 = 212.8813$$

$$\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n = 409 - \frac{(31)^2}{28} = 374.6786$$

$$\therefore r_1 = \frac{212.8813}{374.6786} = .5682$$

1) $H_0 : \rho_1 = 0$

2) $H_1 : \rho_1 \neq 0$

3) $\alpha = .05$

4) CR : $Z > z_{.025} = 1.96$ และ $Z < -1.96$

5) $z_c = \sqrt{n}\, r_1 = \sqrt{28}(.5682) = 3.0066$

6) $\because z_c > 1.96$, เราปฏิเสธ H_0 ที่ $\alpha = .05$ นั่นคือข้อมูลชุดนี้มีความสัมพันธ์กันในตัวข้อมูลเอง

5.12 Model checking และ Multiple linear regression

5.12.1 Examining the residuals

วิธีการนี้ไม่ได้ใช้สำหรับตัวแบบเชิงเส้นตรงเท่านั้น แต่อาจใช้ได้กับ model อื่น ๆ ด้วย เมื่อเราได้ทำการ fit ตัวแบบด้วยวิธีกำลังสองน้อยที่สุดแล้ว ส่วนของความแปรปรวนของ y ซึ่งไม่สามารถอธิบายได้ด้วยตัวแบบจะรวมอยู่ใน residuals

$$e_i = y_i - \hat{y}_i,\ i = 1, \ldots, n$$

โดยที่ y_i คือค่า observed y และ $\hat{y}_i = b_0 + b_1 x_i$

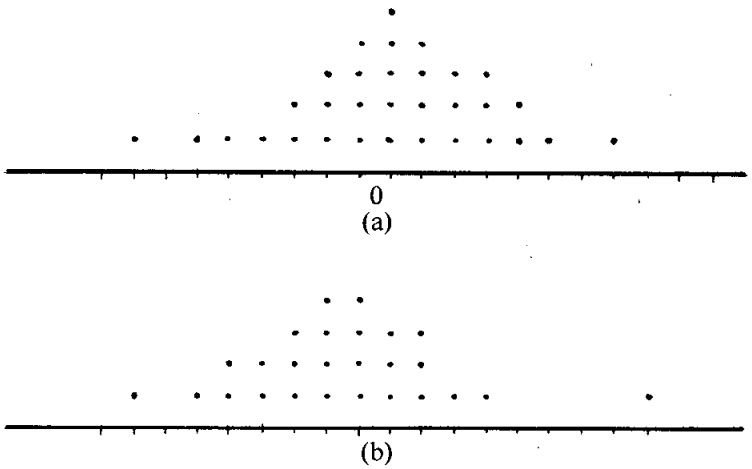
ในการวิเคราะห์การถดถอยที่ผ่านมาเรามีข้อสมมติเกี่ยวกับ random error (E_i) ว่า E_i เป็นอิสระต่อกัน และต่างก็มีการแจกแจงเป็น normal distribution ที่มี mean = 0 และ constant variance σ^2

นั่นคือ $E_i \sim NID(0,\sigma^2)$

ถ้าข้อสมมติเป็นจริงดังกล่าว เราจึงจะเชื่อถือผลการวิเคราะห์ได้ เมื่อตัวแบบถูกต้องเราใช้ $e_i = \hat{E}_i$ ดังนั้นในการตรวจสอบสมมติฐานข้างต้น เราจะสำรวจ residuals โดยวิธีการทางกราฟ

1. Histogram **หรือ** Dot diagram

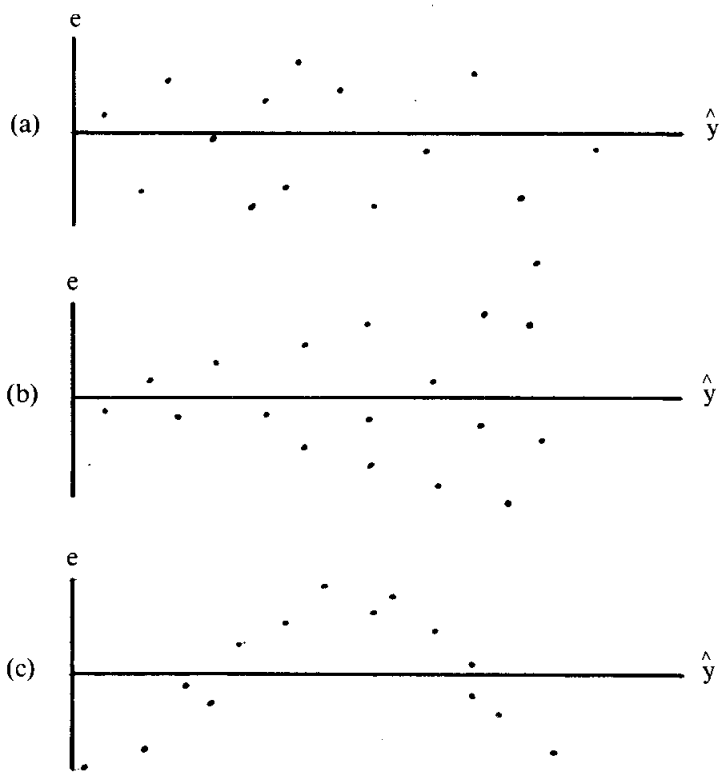
เพื่อดูลักษณะรวม ๆ ของ residuals เราจัด residuals เป็นแบบ grouped data สำหรับข้อมูลจำนวนมาก แล้ว plot histogram สำหรับข้อมูลจำนวนน้อยเราใช้วิธี plot dot diagram บนเส้นตรง ตัวอย่างการ plot ในรูป (a) และ (b) รูป (a) แสดงว่าข้อมูลมีการแจกแจงคล้ายแบบปกติที่มี mean ใกล้ 0 ส่วนรูป (b) อาจพอสรุปได้ว่าข้อมูลมีการแจกแจงแบบปกติ แต่มีค่าสังเกตตัวหนึ่งอยู่ห่างไปจากกลุ่ม (wild observation) ในสถานการณ์เช่นนี้ควรจะได้พิจารณาข้อมูลตัวนั้นอีกครั้งเพื่อให้แน่ใจว่าข้อมูลนี้จะมีอิทธิพลต่อการ fit model หรือไม่อย่างไร (ซึ่งอาจพบว่าเป็นข้อมูลผิดปกติที่อาจตัดออกไปได้)



รูปที่ 5.8 Dot diagram of residuals

2. Residual vs. predicted value ($\hat{y}$)

การ plot e_i กับ $\hat{y}_i$ นั้น จะช่วยตรวจจับว่าข้อสมมติเกี่ยวกับ constant error variance ว่าจริงหรือไม่



รูปที่ 5.9 Plot of residuals vs. predicted values

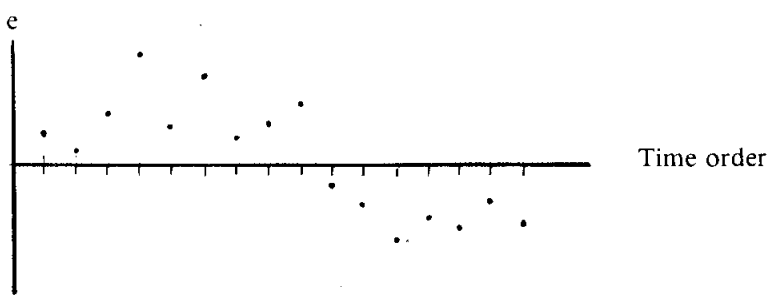
จากรูป (a) จะเห็นว่าจุดกระจายเป็นรูป horizontal band แสดงว่าไม่มีการบ่งชี้ว่าขาดคุณสมบัติของ constant error variance

รูป (b) จะเห็นว่าจุดกระจายในลักษณะที่เมื่อ $\hat{y}$ มีค่ามากขึ้น e_i จะกระจายมากขึ้น แสดงว่า error variance มีแนวโน้มจะมากขึ้นเมื่อค่าของ y_i มากขึ้น ควรตรวจสอบคุณสมบัติของ constant error variance ทั้งนี้เรามีวิธีการแก้โดยการแปลงข้อมูลเพื่อทำให้ variance คงที่

รูป (c) จะเห็นว่าจุดกระจายอย่างมีแบบแผน (systematic pattern) แทนที่จะกระจายอยู่อย่างสุ่มรอบ ๆ แกน $\hat{y}$ เริ่มด้วยการเพิ่มขึ้นอย่างคงที่แล้วลดลง ลักษณะนี้จะบ่งชี้ให้เราสงสัยว่าตัวแบบไม่พอเพียงที่จะอธิบายข้อมูล เราอาจเพิ่มเทอมกำลังสอง หรือพิจารณาลองใช้ตัวแบบที่ไม่ใช่ตัวแบบเชิงเส้นตรง (nonlinear model)

3. Residual vs. Time order

ในการวิเคราะห์การถดถอย การขาดข้อสมมติที่ถือว่าร้ายแรงคือการที่ error E_i ไม่เป็นอิสระต่อกัน การขาดความเป็นอิสระต่อกันมักจะเกิดขึ้นเมื่อประยุกต์ใช้ทางธุรกิจและเศรษฐศาสตร์ โดยเกิดขึ้นเมื่อมีการรวบรวมข้อมูลตามลำดับเวลาเพื่อทำการวิเคราะห์การถดถอยในการทำนายแนวโน้มใด ๆ ก็ตามเมื่อทำการทดลองในเวลาที่แปรไปอย่างมาก การ plot residuals กับลำดับเวลา (time order) มักจะทำให้เราตรวจจับการขาดคุณสมบัติของความเป็นอิสระได้ตัวอย่างในรูปที่ 5.10 เป็นรูปที่มีแบบแผน แสดงว่ากลุ่มของค่าที่สูงจะถูกตามมาด้วยกลุ่มของค่าที่ต่ำ



รูปที่ 5.10 Plot of residuals vs. time order

สิ่งนี้บ่งชี้ว่า residual ที่อยู่ต่อเนื่องกันจะมีความสัมพันธ์กัน ทำให้เราสงสัยได้ว่าเกิดการขาดคุณสมบัติของความเป็นอิสระ เราอาจตรวจสอบความเป็นอิสระของ error โดยการ plot residual $(n-1)$ คู่ในรูปคู่ลำดับ (e_{i-1}, e_i), $i = 2, \ldots, n$ ถ้าการกระจายของจุดไม่มี

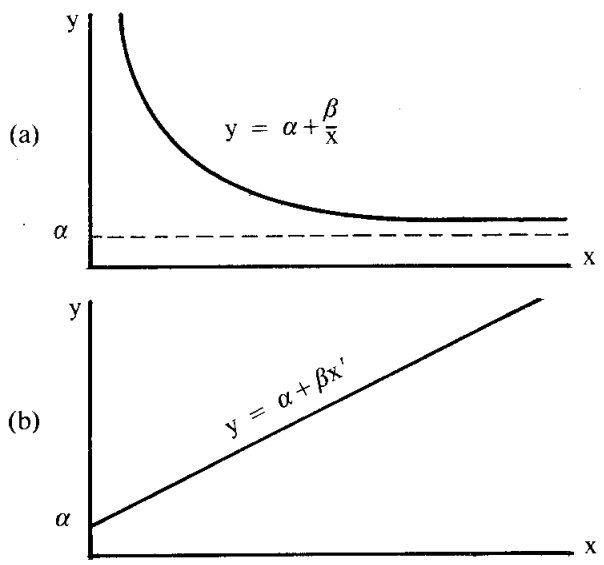
รูปแบบที่ชัดเจน ก็ถือได้ว่าข้อมูลเป็นอิสระกัน ถ้ามีการจับกลุ่มกันอยู่ในแนวเส้นตรง เราจะถือว่าขาดความเป็นอิสระระหว่างข้อมูลที่อยู่ติดกัน วิธีการนี้ได้กล่าวมาแล้วในหัวข้อ 5.11.

5.12.2 ความสัมพันธ์ที่ไม่เป็นเส้นตรงและการแปลงให้เป็นเส้นตรง

(Nonlinear relations and Linearizing Transformations)

เท่าที่ผ่านมาเราได้ศึกษาถึงรูปความสัมพันธ์ระหว่าง X กับ Y ที่เป็นสมการเส้นตรง หรือใช้ตัวแบบที่เป็นสมการเส้นตรง อย่างไรก็ดีในชีวิตจริงรูปความสัมพันธ์ที่เราสนใจนั้นไม่เป็นเส้นตรงเสมอไป จากการ plot scatter diagram จะพบว่ารูปความสัมพันธ์อาจไกลไปจากเส้นตรงมาก และเมื่อทำการคำนวณค่า r^2 จะได้ค่าเล็ก ๆ เมื่อเราใช้สมการเส้นตรง หรือเมื่อทดสอบ lack of fit ก็จะได้ผลสรุปว่ามี lack of fit วิธีการทางสถิติที่จะต้องปฏิบัติกับความสัมพันธ์ที่ไม่เป็นเส้นตรงนั้นยุ่งยากกว่าวิธีการจัดการกับความสัมพันธ์เชิงเส้นตรง ทั้งนี้ยกเว้นกับการจัดการกับตัวแบบที่เรียกว่า polynomial regression model ซึ่งจะกล่าวในหัวข้อ 5.12.4

ในบางสถานการณ์ เราอาจทำการแปลงตัวแปร X และ/หรือ Y เพื่อทำให้รูปความสัมพันธ์ใหม่ใกล้เคียงกับรูปเชิงเส้นตรง เราอาจเขียนตัวแบบเชิงเส้นตรงให้อยู่ในรูปของตัวแปรที่ถูกแปลงแล้ว และทำการวิเคราะห์กับข้อมูลที่ถูกแปลงแล้ว ในการวิเคราะห์นี้ควรจะมีการตรวจสอบคุณสมบัติของ residuals ในตัวแบบที่ถูกแปลงแล้วด้วย เพราะว่าข้อสมมุติที่ว่าตัว error เป็นอิสระต่อกัน และมีการแจกแจงเป็นแบบปกติที่มี variance คงที่นั้นจะต้องใช้กับตัวแบบที่ถูกแปลงแล้วเช่นกัน



รูปที่ 5.11

จากรูปที่ 5.11

ถ้า x = เวลาที่ฝึกฝน

y = เวลาที่ทำงานเสร็จ

จะเห็นได้ว่าเวลาที่ทำงานเสร็จจะลดลงเมื่อใช้เวลาที่ฝึกฝนเพิ่มขึ้น แต่เมื่อถึงจุดหนึ่งจะเห็นได้ว่า y ไม่ลดลงอีกแม้ว่าจะเพิ่มระยะเวลาที่ฝึกฝน

ดังนั้นเราอาจพิจารณาความสัมพันธ์ในรูป

$y = \alpha + \frac{\beta}{x}$ ซึ่งมีรูปความสัมพันธ์ดังในรูป (a)

รูป (a) แสดงว่าความสัมพันธ์ไม่เป็นรูปเส้นตรง

แทนที่เราจะใช้ x เราจะทำการแปลงโดยใช้ $x' = \frac{1}{x}$ ดังนั้นตัวแบบจะกลายเป็น

$y = \alpha + \beta x'$ ซึ่งมีรูปความสัมพันธ์ดังในรูป (b)

การวิเคราะห์ข้อมูลจะทำกับ (y_i, x_i'), $i = 1, \ldots, n$ โดยที่ $x_i' = \frac{1}{x_i}$

ในการ plot scatter diagram เพื่อดูว่าควรจะแปลงตัวแปรอย่างไรเพื่อให้ได้ความสัมพันธ์เชิงเส้นตรงนั้น อาจ plot (x_i, y_i), $i = 1, \ldots, n$ ลงบนกระดาษกราฟเช่น semilog paper (แกนใดแกนหนึ่งอยู่ใน log scale) หรือ double-log paper (ทั้ง 2 แกนอยู่ใน log scale) เพื่อดูว่าควรแปลงตัวใดตัวหนึ่งหรือทั้ง 2 ตัวจึงจะเหมาะสม ตัวอย่างเช่นความสัมพันธ์ของข้อ (a) ในตารางที่ 5.3 นั้น เมื่อ plot ลงบน semilog paper กราฟจะเป็นเส้นตรง บางครั้งถ้า diagram แสดงว่า ถ้า y เพิ่มเร็วมากเมื่อเทียบกับ x เราอาจใช้ $\sqrt{y}$ หรือกำลังเศษส่วนอื่น ๆ ของ y เพื่อทำให้รูปความสัมพันธ์เป็นเส้นตรง

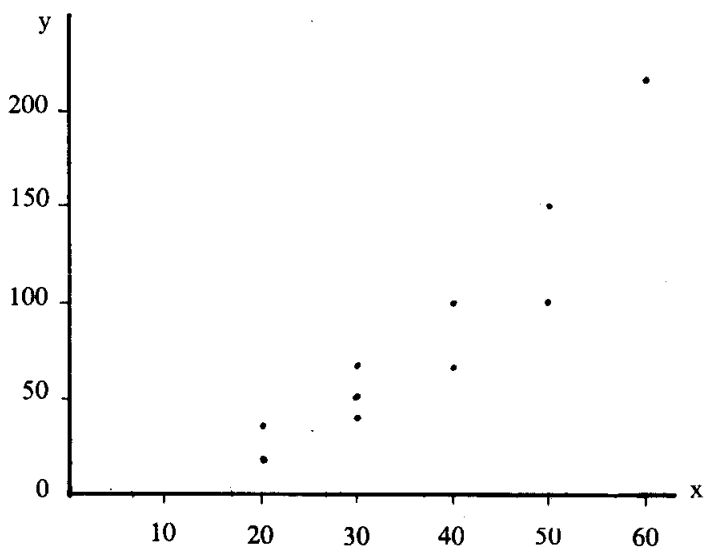
ตารางที่ 5.3

แสดงการแปลงรูปตัวแปรเพื่อให้ได้ความสัมพันธ์ในรูปเส้นตรง

Nonlinear model	Transformation	Transformed model
a) $y = ae^{bx}$	$y' = \ln y,\ x' = x$	$y' = \alpha+\beta x',\ \alpha = \ln a,\ \beta = b$
b) $y = ax^{b}$	$y' = \log y,\ x' = \log x$	$y' = \alpha+\beta x',\ \alpha = \log a,\ \beta = b$
c) $y = \frac{1}{a+bx}$	$y' = \frac{1}{y},\ x' = x$	$y' = \alpha+\beta x',\ \alpha = a,\ \beta = b$
d) $y = \frac{1}{(a+bx)^2}$	$y' = \frac{1}{\sqrt{y}},\ x' = x$	$y' = \alpha+\beta x',\ \alpha = a,\ \beta = b$
e) $\frac{1}{y} = a + \frac{b}{1+x}$	$y' = \frac{1}{y},\ x' = \frac{1}{1+x}$	$y' = \alpha+\beta x',\ \alpha = a,\ \beta = b$
f) $y = a+b\sqrt{x}$	$y' = y,\ x' = \sqrt{x}$	$y' = \alpha+\beta x',\ \alpha = a,\ \beta = b$

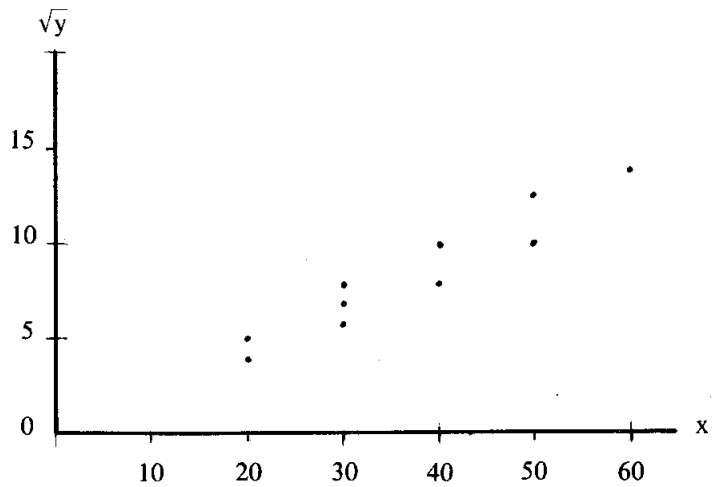
ตัวอย่างที่ 5.14 เพื่อทดสอบความสามารถสูงสุดของการหยุดของรถยนต์เมื่อเหยียบเบรคเต็มที่ รถ 10 คันได้ถูกขับด้วยอัตราความเร็วที่ระบุไว้สำหรับแต่ละคัน และวัดระยะทางที่ต้องใช้ก่อนหยุดรถได้สนิท ความเร็วที่ขับสำหรับรถแต่ละคันและระยะทางที่ใช้หยุด เป็นดังนี้

Initial speed :	x	20	20	30	30	30	40	40	50	50	60
Stopping distance :	y	16.3	26.7	39.2	63.5	51.3	98.4	65.7	104.1	155.6	217.2



จากรูปจะเห็นว่าความสัมพันธ์ไม่เป็นเส้นตรง y เพิ่มขึ้นด้วยอัตราสูงที่ค่าสูง ๆ ของ x มากกว่าที่ค่าต่ำกว่าของ x เป็นการแนะว่าควรจะแปลงโดยใช้ $y' = \sqrt{y}$ หรือ y กำลังเศษส่วนอื่น ๆ การ plot $(x_i, \sqrt{y_i})$ ทำในรูปถัดไป จะเห็นได้ว่ารูปความสัมพันธ์เข้าใกล้เส้นตรงมากกว่ารูปแรก

x	20	20	30	30	30	40	40	50	50	60
$y' = \sqrt{y}$	4.037	5.167	6.261	7.969	7.162	9.920	8.106	10.203	12.474	14.738



เมื่อใช้ model $y' = \alpha + \beta x + E$

จากข้อมูลข้างต้น $\bar{x} = 37, \bar{y}' = 8.604$

$$S_{xx} = 1610, S_{y'y'} = 97.773, S_{xy'} = 381.621$$

$$\therefore \hat{\alpha} = -.166, \hat{\beta} = .237$$

ดังนั้น prediction equation คือ $\hat{y}' = -.166 + .237x$

คำนวณ $r^2 = \dfrac{SSR}{SST} = \dfrac{\hat{\beta} S_{xy'}}{S_{y'y'}} = \dfrac{.237(381.621)}{97.773} = .925$

5.12.3 Multiple linear regression

Model : $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + E_i$, $i = 1, \ldots, n$

โดยที่

x_{1i}, x_{2i}, x_{3i} เป็น fixed values ของตัวแปรอิสระ (independent variable) 3 ตัวในหน่วยทดลองที่ i

y_i คือ response

$E_i \sim NID(0,\sigma^2)$

$\beta_0, \beta_1, \beta_2, \beta_3$ เป็นพารามิเตอร์ที่เราจะประมาณค่า

Observations : $(x_{1i}, x_{2i}, x_{3i}, y_i)$, $i = 1, \ldots, n$

เราจะประมาณค่า $\beta_0, \beta_1, \ldots, \beta_3$ ด้วยวิธีการกำลังสองน้อยที่สุด คือหาค่า $b_0, b_1, \ldots, b_3$ เพื่อทำให้ $\sum_{i=1}^{n} e_i^2$ มีค่าน้อยที่สุด

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - \ldots - b_3 x_{3i})^2$$

จากวิธีทาง calculus

$$\frac{\partial S}{\partial b_0} = 0$$

$$\frac{\partial S}{\partial b_1} = 0$$

$$\vdots$$

$$\frac{\partial S}{\partial b_3} = 0$$

จะได้ normal equations ดังนี้

$$\begin{aligned}
nb_0 + \Sigma x_{1i} b_1 + \Sigma x_{2i} b_2 + \Sigma x_{3i} b_3 &= \Sigma y_i \\
\Sigma x_{1i} b_0 + \Sigma x_{1i}^2 b_1 + \Sigma x_{1i} x_{2i} b_2 + \Sigma x_{1i} x_{3i} b_3 &= \Sigma x_{1i} y_i \\
\Sigma x_{2i} b_0 + \Sigma x_{1i} x_{2i} b_1 + \Sigma x_{2i}^2 b_2 + \Sigma x_{2i} x_{3i} b_3 &= \Sigma x_{2i} y_i \\
\Sigma x_{3i} b_0 + \Sigma x_{1i} x_{3i} b_1 + \Sigma x_{2i} x_{3i} b_2 + \Sigma x_{3i}^2 b_3 &= \Sigma x_{3i} y_i
\end{aligned}$$

หรืออยู่ในรูป

$$b_1S^2_{x_1} + b_2S_{x_1x_2} + b_3S_{x_1x_3} = S_{x_1y}$$

$$b_1S_{x_1x_2} + b_2S^2_{x_2} + b_3S_{x_2x_3} = S_{x_2y}$$

$$b_1S_{x_1x_3} + b_2S_{x_2x_3} + b_3S^2_{x_3} = S_{x_3y}$$

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - b_3\bar{x}_3$$

รูปทั่วไปของ multiple linear regression model ในรูปเมตริกซ์

Model : $\underline{Y}_{nx1} = \underline{X}_{n\,x\,(p+1)}\underline{\beta}_{(p+1)\,x\,1} + \underline{E}_{nx1}$

โดยที่ $\underline{Y}$ คือ observed vector

$$\underline{Y}' = [y_1\ y_2\ \dots\ y_n]$$

$$\underline{E}' = [E_1\ E_2\ \dots\ E_n]$$

$$\underline{\beta}' = [\beta_0\ \beta_1\ \dots\ \beta_p]$$

X = design matrix

$$X = \begin{bmatrix} 1 & x_{11} & \dots & x_{p1} \\ 1 & x_{12} & \dots & x_{p2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & & x_{pn} \end{bmatrix}_{n\,x\,(p+1)}$$

Normal equations : $(X'X)\hat{\underline{\beta}} = X'\underline{Y}$

โดยที่

$$X'X = \begin{bmatrix} n & \Sigma x_{1i} & \Sigma x_{2i} & \dots & \Sigma x_{pi} \\ \Sigma x_{1i} & \Sigma x^2_{1i} & \Sigma x_{1i}x_{2i} & \dots & \Sigma x_{1i}x_{pi} \\ \vdots & \vdots & \vdots & & \vdots \\ \Sigma x_{pi} & \Sigma x_{1i}x_{pi} & \Sigma x_{2i}x_{pi} & \dots & \Sigma x^2_{pi} \end{bmatrix}_{(p+1)\,x\,(p+1)}, \quad X'\underline{Y} = \begin{bmatrix} \Sigma y_i \\ \Sigma x_{1i}y_i \\ \vdots \\ \Sigma x_{pi}y_i \end{bmatrix}$$

$$\hat{\underline{\beta}}' = [b_0\ \ b_1\ \ \dots\ \ b_p]$$

ดังนั้น $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{Y}$

ตัวอย่างที่ 5.15 เราสนใจว่าความดันโลหิต y มีความสัมพันธ์กับน้ำหนัก (x_1) และอายุ (x_2) ในกลุ่มผู้ชายที่มีความสูงเท่า ๆ กันหรือไม่ จากชาย 13 คนซึ่งมีน้ำหนักและอายุที่เราเลือกไว้ก่อน วัดความดันโลหิตของชายทั้ง 13 คนได้ข้อมูลในตารางที่ 5.4 ให้ fit multiple regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + E_i,\ i = 1, \ldots, 13$$

ตารางที่ 5.4

แสดงน้ำหนัก อายุและความดันโลหิตของชาย 13 คน

คนที่	x_1	x_2	y
1	152	50	120
2	183	20	141
3	171	20	124
4	165	30	126
5	158	30	117
6	161	50	129
7	149	60	123
8	158	50	125
9	170	40	132
10	153	55	123
11	164	40	132
12	190	40	155
13	185	20	147

คำนวณค่าต่าง ๆ จากข้อมูลในตารางที่ 5.4 ดังนี้ (Σ คือ $\sum_{i=1}^{13}$)

$\Sigma x_{1i} = 2159$ $\quad \Sigma x_{1i}^2 = 360639$ $\quad \Sigma x_{1i}x_{2i} = 82335$

$\Sigma x_{2i} = 505$ $\quad \Sigma x_{2i}^2 = 21925$ $\quad \Sigma x_{1i}y_i = 282921$

$\Sigma y_i = 1694$ $\quad \Sigma x_{2i}y_i = 65135$

$$S^2_{x_1} = \sum_{i=1}^{13} x^2_{1i} - (\sum_{i=1}^{13} x_{1i})^2/13 = 360639 - \frac{(2159)^2}{13} = 2078.9$$

$$S^2_{x_2} = \sum_{i=1}^{13} x^2_{2i} - (\sum_{i=1}^{13} x_{2i})^2/13 = 21925 - \frac{(505)^2}{13} = 2307.7$$

$$S_{x_1x_2} = \sum_{i=1}^{13} x_{1i}x_{2i} - \frac{(\Sigma x_{1i})(\Sigma x_{2i})}{13} = 82335 - \frac{(2159)(505)}{13} = -1533.846$$

$$S_{x_1y} = \sum_{i=1}^{13} x_{1i}y_i - \frac{(\Sigma x_{1i})(\Sigma y_i)}{13} = 282921 - \frac{(2159)(1694)}{13} = 1586.7$$

$$S_{x_2y} = \sum_{i=1}^{13} x_{2i}y_i - \frac{(\Sigma x_{2i})(\Sigma y_i)}{13} = 65135 - \frac{(505)(1694)}{13} = -670.38$$

$$\bar{y} = 1694/13 = 130.308$$

$$\bar{x}_1 = 2159/13 = 166.077$$

$$\bar{x}_2 = 505/13 = 38.846$$

ดังนั้น Normal equations คือ

$$2078.9b_1 - 1533.8b_2 = 1586.7 \qquad \text{..... (1)}$$

$$-1533.8b_1 + 2307.7b_2 = -670.38 \qquad \text{..... (2)}$$

$$b_0 = 130.308 - 166.077b_1 - 38.846b_2 \qquad \text{..... (3)}$$

จาก (1) และ (2) ได้ $b_1 = 1.077,\ b_2 = .425$

$\therefore \quad b_0 = -65.10$

prediction equation คือ

$$\hat{y} = -65.10 + 1.077x_1 + .425x_2$$

หมายความว่า ความดันโลหิตเฉลี่ยจะเพิ่มขึ้น 1.077 หน่วย ถ้าน้ำหนักเพิ่มขึ้น 1 หน่วย โดยที่อายุคงที่ และถ้าน้ำหนักคงที่แล้วอายุเพิ่มขึ้น 1 ปี จะทำให้ความดันเฉลี่ยเพิ่มขึ้น .425 หน่วย

5.12.4 Polynomial regression model : Quadratic function (Second degree polynomial)

เมื่อเรา plot scatter diagram แล้วความสัมพันธ์ไม่อยู่ในรูปเส้นตรง (nonlinear) เราอาจทำการแปลงตัวแปรตามวิธีที่กล่าวมาแล้วในหัวข้อ 5.12.2 แต่บางกรณีเราไม่อาจทำดังกล่าวได้ วิธีหนึ่งที่จะจัดการกับข้อมูลได้คือใช้ตัวแปร x ที่มีกำลังสูงกว่ากำลังหนึ่ง เช่น

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i,\ i = 1, \ldots, n$$

ถ้าตัดเทอม E_i ออกจะเห็นว่าเป็น quadratic function ของตัวแปรอิสระ x เราเรียกตัวแบบรูปนี้ว่าตัวแบบการถดถอยรูป polynomial ของ y กับ x กำลังสูงสุดของ x จะเป็นตัวบอก degree หรือ order ของการถดถอยรูป polynomial และในการวิเคราะห์การถดถอยรูปนี้ไม่ได้ใช้วิธีพิเศษนอกเหนือไปจากวิธีของ multiple regression เลย

ตัวแบบที่ใช้คือ

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + E_i,\ i = 1, \ldots, n$$

โดยที่ $x_{1i} = x_i$ และ $x_{2i} = x_i^2$ นั่นเอง

ตัวอย่างที่ 5.16 จากข้อมูลที่กำหนดให้ จง fit quadratic model $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + E$ โดยวิธีกำลังสองน้อยที่สุด

x	−2	−1	0	1	2
y	.4	1.3	2.2	2.5	3.0

	$x_1 = x$	$x_2 = x^2$	y	x_1^2	x_2^2	x_1x_2	x_1y	x_2y
	−2	4	.4	4	16	−8	−.8	1.6
	−1	1	1.3	1	1	−1	−1.3	1.3
	0	0	2.2	0	0	0	0	0
	1	1	2.5	1	1	1	2.5	2.5
	2	4	3.0	4	16	8	6.0	12.0
Total	0	10	9.4	10	3.4	0	6.4	17.4

คำนวณค่าต่าง ๆ ดังนี้ $\bar{x}_1 = 0$, $S^2_{x_1} = 10$, $S_{x_1y} = 6.4$

$\bar{x}_2 = 2$, $S^2_{x_2} = 14$, $S_{x_2y} = -1.4$

$\bar{y} = 1.88$, $S_{x_1x_2} = 0$

Normal equations คือ

$$10b_1 = 6.4 \qquad \text{..... (1)}$$

$$14b_2 = -1.4 \qquad \text{..... (2)}$$

$$b_0 = 1.88 - 2b_2 \qquad \text{..... (3)}$$

ดังนั้น $b_1 = .64,\ b_2 = -.1,\ b_0 = 2.08$

prediction equation คือ

$$\hat{y} = 2.08 + .64x - .1x^2$$

ถ้าใช้วิธีเมตริกซ์

$$\text{Design matrix } X = \begin{bmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix}, \quad \underline{Y} = \begin{bmatrix} .4 \\ 1.3 \\ 2.2 \\ 2.5 \\ 3.0 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 5 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 34 \end{bmatrix}, \quad X'\underline{Y} = \begin{bmatrix} 9.4 \\ 6.4 \\ 17.4 \end{bmatrix}$$

$$\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{Y}$$

$$= \begin{bmatrix} \frac{34}{70} & 0 & -\frac{1}{7} \\ 0 & \frac{1}{10} & 0 \\ -\frac{1}{7} & 0 & \frac{5}{70} \end{bmatrix} \begin{bmatrix} 9.4 \\ 6.4 \\ 17.4 \end{bmatrix}$$

$$= \begin{bmatrix} 2.08 \\ .64 \\ -.1 \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

$$\therefore \quad \hat{y} = 2.08 + .64x - .1x^2$$

แบบฝึกหัด

5.1 พนักงานป่าไม้ต้องการทราบว่าเขาจะใช้เส้นผ่าศูนย์กลางในการทำนายความสูงของต้นไม้ได้หรือไม่ เขาได้ทำการวัดเส้นผ่าศูนย์กลางของต้นไม้ตรงส่วนสูง 4.5 ฟุตจากพื้นดิน และความสูงของต้นเมเปิลที่ให้น้ำตาล 12 ต้น ผลปรากฏ ดังนี้

x : เส้นผ่าศูนย์กลาง (นิ้ว)	.9	1.2	2.3	3.1	3.3	3.9	4.3	6.2	9.6	12.6	16.1	25.8
y : ความสูง (ฟุต)	18	26	32	36	44.5	35.6	40.5	57.5	67.3	84	67	87.5

ก) จงสร้าง Scatter diagram เพื่อดูว่าควรใช้ Straight line model หรือไม่

ข) ถ้ากำหนดให้ $x' = \log x$ และ $y' = \log y$ จงสร้าง Scatter diagram ของ (x', y')

x′	−.05	.08	.36	.49	.52	.59	.63	.79	.98	1.10	1.21	1.41
y′	1.26	1.41	1.51	1.56	1.65	1.55	1.61	1.76	1.83	1.92	1.83	1.94

ค) จาก model $Y_i' = \beta_0 + \beta_1 x_i' + E_i$ จงหา Sample regresion equation

5.2 ในการเตรียมการที่จะคัดเลือกนักศึกษาปีที่ 1 เข้าเรียนวิชาเอกคณิตศาสตร์ โดยจะใช้วิธีการสอบวิชาคณิตศาสตร์พื้นฐาน (คะแนนเต็ม 10 คะแนน) เพื่อจะศึกษาดูว่าการใช้คะแนนวิชานี้จะทำนายผลสำเร็จในการเรียนวิชาเอกได้หรือไม่ วิทยาลัยแห่งหนึ่งจึงได้เอานักศึกษาที่จบแล้ว และได้คะแนนวิชาพื้นฐานต่าง ๆ กันมา 8 คน แล้วจดคะแนนตอนจบปริญญาตรี (Grade point average) ของแต่ละคนดังนี้

x : คะแนนคณิตศาสตร์พื้นฐาน	4	6	8	6.5	3.5	8	7	5
y : คะแนนจบปริญญาตรี	2.1	3.1	3.4	2.9	2.1	3.3	2.8	2.3

กำหนดค่าที่คำนวณจากข้อมูลดังนี้

$\Sigma x_i = 46$, $\Sigma x_i^2 = 308.5$, $\bar{x} = 6$, $\Sigma x_i y_i = 137.9$

$\Sigma y_i = 22$, $\Sigma y_i^2 = 62.42$, $\bar{y} = 2.75$

$\Sigma(x_i - \bar{x})^2 = 20.5$, $\Sigma(y_i - \bar{y})^2 = 1.92$, $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 5.9$

สมมุติว่าความสัมพันธ์ของ X และ Y อยู่ในรูป $\mu_{Y/x} = \beta_0 + \beta_1 x$

ก) จงหา Least-squared estimated ของ β_0 และ β_1

ข) จงเขียน prediction equation

ค) จงหา point estimate ของ $\mu_{Y|7.5}$ และอธิบายความหมาย และถ้านายก.ได้คะแนนวิชาพื้นฐาน 5.5 คะแนน จงประมาณว่านายก.จะเรียนจบด้วยคะแนนเฉลี่ยเท่าใด

ง) จงทดสอบ $H_0 : \beta_1 = 0$ โดยใช้ F-test กำหนด $\alpha = .01$

จ) จงหา 100 r^2% และอธิบายความหมาย

ฉ) ถ้าความสัมพันธ์ของ X และ Y อยู่ในรูปนี้จริง ค่าประมาณ ของ $\sigma^2 = ?$

5.3 ในปี 2519 บริษัทผู้ผลิตผงซักฟอกยี่ห้อต่าง ๆ มีรายจ่ายในการโฆษณาและยอดขายดังนี้

ยี่ห้อ	x : รายจ่ายในการโฆษณา (ล้านบาท)	y : ยอดขาย (ล้านบาท)
บรีด	7	9
ริน	9	12
นิส	6	8
ควิก	5	7
แฟบ	3	4

สมมุติรูปความสัมพันธ์ คือ $\mu_{Y/x} = \beta_0 + \beta_1 x$

ก) จงเขียน prediction equation

ข) จงอธิบายค่า $\hat{\beta}_0 = b_0$ และ $\hat{\beta}_1 = b_1$ ที่หาได้

ค) ถ้าบริษัทผลิตควิกมีรายจ่ายในการโฆษณา 4 ล้านบาท จงประมาณว่าจะได้ยอดขายกี่ล้านบาท

ง) จงทดสอบที่ $\alpha = .05$ $H_0 : \beta_1 = 0$ โดยใช้ F-test

จ) ถ้าหา 95% C.I. ของ $\mu_{Y/6}$ ได้ (7.631, 8.369) จงอธิบายความหมาย

5.4 จากข้อมูลที่กำหนดให้

ก) จงสร้าง Scatter diagram

ข) จงเขียน prediction equation ของ model : $\mu_{Y/x} = \beta_0 + \beta_1 x$

ค) จงทดสอบ lack of fit ที่ $\alpha = .05$

x	1	1	1	2	3	3	4	5	5
y	9	7	8	10	15	12	19	24	21

5.5 ปริมาณสารเคมี y (กรัม) ซึ่งละลายในน้ำ 100 กรัม ณ อุณหภูมิต่าง ๆ กันเป็นดังนี้

x (°c)	y (กรัม)
0	8, 6, 8
15	12, 10, 14
30	25, 21, 24
45	31, 33, 28
60	44, 39, 42
75	48, 51, 44

ก) จงสร้าง Scatter diagram

ข) จงเขียน prediction equation ของ model : $\mu_{Y/x} = \beta_0 + \beta_1 x$

ค) จงประมาณปริมาณของสารเคมีที่จะละลายในน้ำ 100 กรัมที่อุณหภูมิ 50°C

ง) จงทดสอบที่ $\alpha = .01$ $H_0 : \beta_0 = 6$, $H_1 : \beta_0 \neq 6$

จ) จงทดสอบว่า Straight line model ที่ใช้อยู่นี้พอเพียงหรือไม่ นั่นคือทดสอบ H_0 : ไม่มี lack of fit

5.6 กำหนดข้อมูลให้ต่อไปนี้

x	2.1	2.6	2.7	2.7	2.7	3.0	3.0	3.0	3.1	3.1	3.2	3.3	3.6	3.8
y	2.6	2.8	2.9	3.6	3.6	3.1	3.4	3.1	3.9	3.1	4.1	3.6	4.0	3.6

$\Sigma x_i = 41.9$, $\Sigma y_i = 47.4$, $\Sigma x_i y_i = 143.59$

$\Sigma x_i^2 = 127.79$, $\Sigma y_i^2 = 163.26$, $n\bar{y}^2 = 160.4829$

$\Sigma (x_i - \bar{x})^2 = 2.3893$, $\Sigma (y_i - \bar{y})^2 = 2.7771$, $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 1.7286$

Simple regression model : $\mu_{Y/x} = \beta_0 + \beta_1 x$ หรือ $Y = \beta_0 + \beta_1 x + E$

กำหนด $\hat{\beta}_0 = b_0 = 1.2203$ และ

$\hat{\beta}_1 = b_1 = 0.7235$

ก) จงเขียน prediction equation หรือ Fitted regression equation

ข) จงสร้างตาราง ANOVA เพื่อทดสอบที่ $\alpha = .05$

1) $H_0 : \beta_1 = 0$

2) H_0 : ไม่มี lack of fit , เราต้องปรับปรุง model ใหม่หรือไม่

3) คำนวณ $100\ r^2\%$

5.7 สุ่มคนไข้ที่เป็นโรคชนิดหนึ่งมา 12 คน หลังจากที่ให้การรักษาด้วยยาชนิดหนึ่งแล้ว 3 วัน ได้วัดอุณหภูมิ (X) และจำนวนการเต้นของชีพจรต่อ 1 นาทีของคนไข้แต่ละคน (Y) ผลปรากฏดังนี้

x : อุณหภูมิ	99.4	100.1	98.6	97.9	101.4	102.6
y : จำนวนการเต้นของชีพจร/นาที	78	83	70	76	75	89

x : อุณหภูมิ	99.0	98.6	101.1	98.0	100.5	99.3
y : จำนวนการเต้นของชีพจร/นาที	80	77	86	79	84	87

$\Sigma x_i = 1196.5,\qquad \Sigma y_i = 964.0$

$\Sigma x_i^2 = 119324.37,\qquad \Sigma_i^2 = 77786$

$\Sigma x_i y_i = 96170.2$

ก) คำนวณค่า r (sample correlation coefficient)

ข) ทดสอบที่ $\alpha = .05$ $H_0 : \rho = 0,\ H_1 : \rho \neq 0$

ค) ทดสอบที่ $\alpha = .05$ $H_0 : \rho = .7,\ H_1 : \rho \neq .7$

ง) จงหา 95% C.I. ของ ρ

5.8 สุ่มนักเรียนมา 8 คน แต่ละคนวัด I.Q. (X) และวัดความเร็วในการอ่าน (Y) ผลปรากฏดังนี้

x : I.Q.	102	95	115	112	86	102	121	108
y : ความเร็ว	14.7	12.0	17.1	15.8	10.1	11.6	19.8	17.5

ก) คำนวณค่า r

ข) ทดสอบ $H_0 : \rho = 0,\ H_1 : \rho \neq 0$ ที่ $\alpha = .05$

ค) จงหา 95% C.I. ของ ρ

5.9 จากคะแนนวิชาคณิตศาสตร์ และคะแนนภาษาอังกฤษ ของนักศึกษา 6 คนที่สุ่มมาเป็นตัวอย่าง

x : คะแนนคณิตศาสตร์	70	92	80	74	65	83
y : คะแนนภาษาอังกฤษ	74	84	63	87	78	90

ก) จงคำนวณค่า r

ข) ทดสอบ $H_0 : \rho = 0$, $H_1 : \rho \neq 0$ ที่ $\alpha = .10$

ค) จากข้อ ข) ถ้าเราไม่สามารถปฏิเสธ H_0 เราจะสรุปผลการทดสอบว่าอย่างไร

5.10 ต่อไปนี้เป็นความสูงและน้ำหนักของมิสอเมริกา 17 คน

x : ความสูง (นิ้ว)	65	67	66	65.5	65	66.5	66	67	66
y : น้ำหนัก (ปอนด์)	114	120	116	118	115	124	124	115	116

x : ความสูง	69	67	65.5	68	67	68	69	68
y : น้ำหนัก (ปอนด์)	135	125	110	121	118	120	125	119

ก) จงสร้าง Scatter diagram

ข) คำนวณค่า r และอธิบายความหมาย

ค) ทดสอบที่ $\alpha = .10$ $H_0 : \rho = 0$

5.11 จากข้อมูลที่กำหนดให้ 2 ชุด จงทดสอบ $H_0 : \rho_1 = \rho_2$ ที่ $\alpha = .05$

x_{1i}	.20	.25	.25	.30	.40	.50	.50
y_{1i}	30	26	40	35	54	56	65

x_{2i}	.20	.25	.30	.40	.40	.50
y_{2i}	23	24	42	49	55	70

5.12 จากการวัดความถี่ของเสียงที่สี่แยกหนึ่งในเวลาที่ต่อเนื่องกันได้ข้อมูลดังนี้

65, 64, 63, 61, 60, 58, 63, 64, 62, 64, 63, 63, 62, 60, 62, 64, 66, 68, 68, 69

จงทำการสำรวจเพื่อตรวจดูว่าอาจมี lag 1 independence หรือไม่โดย

ก) สร้าง Scatter diagram

ข) คำนวณ r_1 (first serial correlation coefficient)

ค) ทดสอบ $H_0 : \rho_1 = 0$, $H_1 : \rho_1 \neq 0$ ที่ $\alpha = .05$

กำหนด $\sum_{i=1}^{n} x_i^2 = 80667$, $\sum_{i=2}^{n} x_{i-1}x_i = 76140$, $\sum_{i=1}^{n} x_i = 1269$

5.13 จากข้อมูลที่วัดต่อเนื่องกันของปริมาณซัลเฟอร์ของโลหะชนิดหนึ่งจากเครื่องทำความร้อนในบ้าน เป็นดังนี้

2.0, 2.8, 2.0, 1.9, 1.1, 1.4, 1.2, 2.4, 2.5, 2.3, 3.1, 5.2, 2.8, 2.7, 2.6, 2.4, 3.3, 3.9, 3.0, 3.8, 2.9

ก) จงสร้าง Scatter diagram

ข) จงคำนวณ r_1

ค) จงทดสอบ $H_0 : \rho_1 = 0$, $H_1 : \rho_1 \neq 0$ ที่ $\alpha = .05$

กำหนด $\sum_{i=1}^{n} x_i^2 = 163.61$, $\sum_{i=2}^{n} x_{i-1}x_i = 148.48$, $\sum_{i=1}^{n} x_i = 55.3$