บทที่ 5
การถดถอย และ สหสัมพันธ์
(Regression and Correlation)

5.1 การถดถอยเชิงเส้นตรง (Linear Regression)

ใน linear regression จะมีตัวแปรตาม (dependent variable) หรือผลตอบสนอง Y 1 ตัว ซึ่งเป็นตัวที่ไม่ถูกควบคุมในการทดลอง ผลตอบสนองนี้ขึ้นอยู่กับตัวแปรอิสระ (independent variable) 1 ตัวหรือมากกว่า 1 ตัว คือ $x_1, x_2, \ldots, x_k$ ซึ่งถูกวัดอย่างถูกต้องไม่มีข้อผิดพลาด ตัวแปรอิสระเหล่านี้เป็นตัวถูกควบคุมในการทดลอง ดังนั้นตัวแปรอิสระ $x_1, \ldots, x_k$ จึงไม่เป็นตัวแปรเชิงสุ่ม (random variable) แต่เป็นปริมาณคงที่ (fixed quantities) ซึ่งถูกกำหนดก่อนโดยผู้ทำการทดลอง และไม่มีรูปการแจกแจงความน่าจะเป็น รูปความสัมพันธ์ของข้อมูลจากการทดลองซึ่งอยู่ในรูปของ prediction equation เราเรียกว่า Regression equation

ในกรณีที่มีตัวแปรตาม Y 1 ตัว และตัวแปรอิสระ x 1 ตัว เราเรียก regression ของ Y on x แต่ถ้ามีตัวแปรอิสระ k ตัว เราเรียก regression ของ Y on $x_1, \ldots, x_k$ ใน Chemical engineer เราจะพิจารณาเกี่ยวกับปริมาณไฮโดรเจนที่สูญเสียไปจากตัวอย่างของธาตุที่เราสนใจเมื่อเราเก็บไว้ในที่เก็บในสภาวะต่าง ๆ กัน ในกรณีนี้อาจจะมีตัวแปรอิสระ 2 ตัว คือเวลาในการเก็บ : x_1 คิดเป็นชั่วโมง และอุณหภูมิของที่เก็บ : x_2 คิดเป็นองศาเซลเซียส Y จะเป็นปริมาณของไฮโดรเจนที่สูญเสียไป

ในบทนี้เราจะพิจารณา Simple linear regression คือในกรณีที่มีตัวแปรอิสระเพียง 1 ตัว

ตัวอย่างสุ่มขนาด n $\{ (x_i, y_i), i = 1,2,\ldots,n \}$ ในตัวอย่างที่มี x ค่าเดียวกัน เราคาดว่าค่า y จะแปรค่าไป ดังนั้นค่า y ในค่าสังเกต (x_i, y_i) จะเป็นค่าของตัวแปรเชิงสุ่ม Y จากนี้ไปเราจะใช้ $Y|x$ แทนตัวแปรเชิงสุ่ม Y ที่คู่กับค่า x เราเขียนการแจกแจงความน่าจะเป็น (Probability distribution) ของ $Y|x$ ด้วย $f(y|x)$ ดังนั้นถ้า $x = x_i$ สัญลักษณ์ $Y|x_i$ จะแทนตัวแปรเชิงสุ่ม Y_i

เทอม Linear regression มีความหมายว่า mean ของ $Y|x$ ($\mu_{Y|x}$) มีความเกี่ยวข้องกับ x ในรูปของเส้นตรงที่มีสมการ

$\mu_{Y|x} = \beta_0 + \beta_1 x$

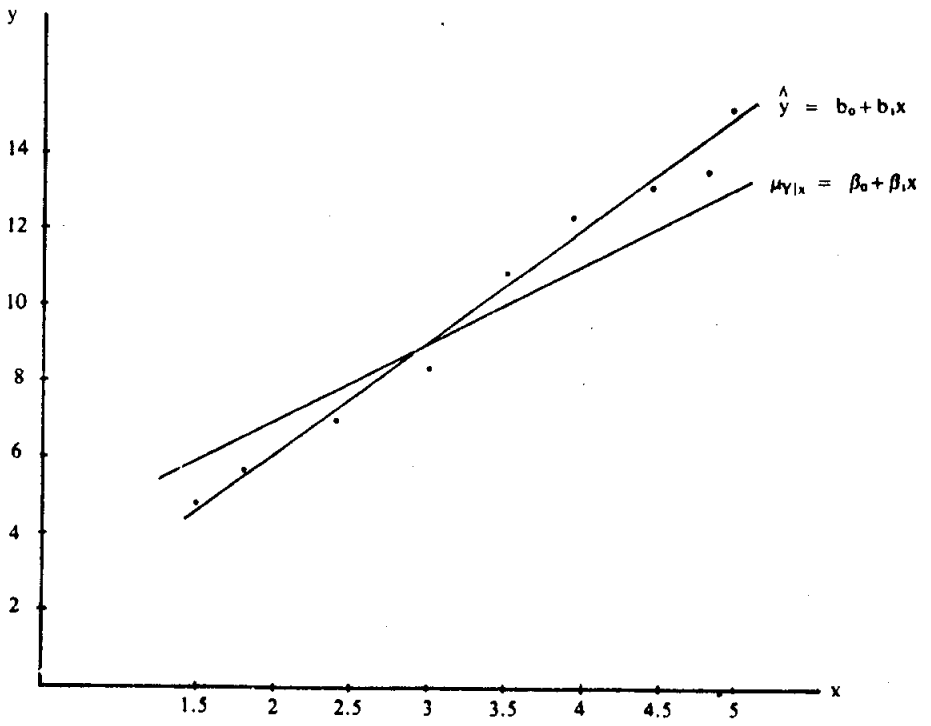
โดยที่ β_0 และ β_1 เป็นพารามิเตอร์ที่เราจะประมาณค่าโดยใช้ข้อมูลจากตัวอย่าง ค่าประมาณของมันเขียนแทนด้วย b_0 และ b_1 ตามลำดับ และค่าประมาณของผลตอบสนองคือ $\hat{y}$ เราหาได้จาก Sample regression line หรือ prediction equation

$\hat{y} = b_0 + b_1 x$

พิจารณาข้อมูลจากการทดลองจากตารางที่ 5.1 เราสามารถ plot ได้ดังรูปที่ 5.1 ในรูปของ Scatter diagram เราจะเห็นว่าสมมติฐานเกี่ยวกับความสัมพันธ์เชิงเส้นตรงนั้นสมเหตุสมผล

ตารางที่ 5.1

i	1	2	3	4	5	6	7	8	9
x_i	1.5	1.8	2.4	3.0	3.5	3.9	4.4	4.8	5.0
y_i	4.8	5.7	7.0	8.3	10.9	12.4	13.1	13.6	15.3



รูปที่ 5.1 Scatter diagram และ Regression lines

ในกรณีที่มี x 1 ตัว และ Y 1 ตัว ข้อมูลอยู่ในรูป (x_i, y_i), $i = 1 \ldots, n$ ถ้าค่าของ x ถูกควบคุมนั่นคือเราออกแบบการทดลอง โดยกำหนดค่าของ x_i ก่อนแล้วจึงสังเกตค่า y_i

สมมุติว่า mean $\mu_{Y|x}$ ทั้งหมดอยู่บนเส้นตรงเดียวกัน ตัวแปรเชิงสุ่ม $Y_i = Y|x_i$ อาจเขียนในรูปตัวแบบเชิงเส้นตรง (Straight line model)

$$Y_i = \mu_{Y|x_i} + E_i = \beta_0 + \beta_1 x_i + E_i ,$$

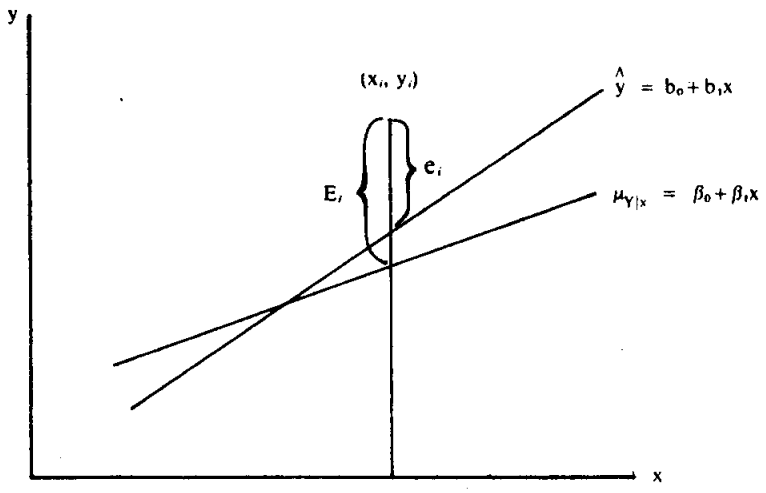
โดยที่ random error E_i ต้องมี mean เป็น 0 แต่ละ observation (x_i, y_i) ต้องสอดคล้องกับสมการ

$$y_i = \beta_0 + \beta_1 x_i + E_i$$

ดังนั้นเราประมาณ regression line ด้วย Sample regression line

$$\hat{y} = b_0 + b_1 x$$

แต่ละ observation (x_i, y_i) สอดคล้องกับ $y_i = b_0 + b_1 x_i + e_i$ นั่นคือ $e_i = y_i - \hat{y}_i$ เราเรียก e_i ว่า residual จากรูป 5.2 แสดงความแตกต่างของ e_i และ E_i



รูปที่ 5.2 เปรียบเทียบ random error E_i กับ residual e_i

เราสามารถหาค่า b_0 และ b_1 ซึ่งเป็นค่าประมาณของ β_0 และ β_1 ที่ทำให้ Sum of squares of residuals มีค่าน้อยที่สุด เรามักเรียก Residual sum of squares ว่า Sum of squares of errors และเขียนแทนด้วย SSE วิธีการ minimization ที่ใช้หาค่าประมาณของพารามิเตอร์ เราเรียกว่า Method of least squares

ดังนั้นเราจะหาค่า b_0 และ b_1 ที่ทำให้ SSE มีค่าน้อยที่สุด

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

ทำ Partial differentiation กับ SSE เทียบกับ b_0 และ b_1

$$\frac{\partial(SSE)}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) \qquad (1)$$

$$\frac{\partial(SSE)}{\partial b_1} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)\, x_i \qquad (2)$$

ให้สมการ (1) = 0 ; $-2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$

$$\Rightarrow \quad -\sum_{i=1}^{n} y_i + nb_0 + b_1 \sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow \quad nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad (3)$$

และ ให้สมการ (2) = 0 ; $-2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)\, x_i = 0$

$$\Rightarrow \quad -\sum_{i=1}^{n} x_i y_i + b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\Rightarrow \quad b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \qquad (4)$$

สมการ (3) และ (4) เรารวมเรียกว่า Normal equations แก้สมการทั้ง 2 จะได้

$$b_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} \quad \text{และ}$$

$$b_0 = \bar{y} - b_1\bar{x} \text{ โดยที่ } \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \text{ และ } \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

ตัวอย่างที่ 5.1 จงประมาณเส้น regression line จากข้อมูลในตารางที่ 5.1

$n = 9$

$\sum_{i=1}^{9} x_i = 30.3 \qquad \sum_{i=1}^{9} y_i = 91.1$

$\sum_{i=1}^{9} x_i^2 = 115.11 \qquad \sum_{i=1}^{9} x_i y_i = 345.09$

$\bar{x} = 3.3667 \qquad \bar{y} = 10.1222$

$$\therefore b_1 = \frac{9(345.091) - (30.3)(91.1)}{(9)(115.11) - (30.3)^2} = 2.9303$$

$$b_0 = 10.1222 - (2.9303)(3.3667) = 0.2568$$

$\therefore$ Sample regression line คือ $\hat{y} = 0.2568 + 2.9303x$

แทนค่า $x_1 = 1.5$ และ $x_9 = 5.0$ ในสมการนี้ ได้ $\hat{y}_1 = 4.7$ และ $\hat{y}_9 = 14.9$ เส้น Sample regression line ในรูปที่ 5.1 ได้จากการลากเส้นเชื่อมจุด $(x_1, \hat{y}_1) = (1.5, 4.7)$ และ $(x_9, \hat{y}_9) = (5.0, 14.9)$

5.2 การหาช่วงความเชื่อมั่นและการทดสอบสมมติฐาน (Confidence intervals and Tests of Significance)

ข้อสมมุติเพิ่มเติมเกี่ยวกับ error term ใน model $Y_i = \beta_0 + \beta_1 x_i + E_i$ (ที่เพิ่มเติม จากข้อสมมุติที่ว่า E_i เป็นตัวแปรเชิงสุ่มซึ่งมี mean เป็น 0) คือ E_i แต่ละตัวมีการแจกแจงเป็น Normal distribution ซึ่งมี variance เท่ากันหมด คือ $= \sigma^2$ (common variance) และ $E_1, \ldots, E_n$ เป็นอิสระต่อกัน การที่เราเพิ่มเติมข้อสมมุติว่า E_i's มีการแจกแจงเป็น Normal ทำให้เราสามารถสร้าง C.I. ของ β_0 และ β_1 และของ $\mu_{Y/x}$ ได้

เราต้องคำนึงอยู่เสมอว่าค่าของ b_0 และ b_1 ซึ่งเป็นค่าประมาณของ β_0 และ β_1 นั้นเราได้จากตัวอย่างขนาด n ค่าประมาณของ β_0 และ β_1 จะมีค่าต่าง ๆ กัน จากตัวอย่างขนาด n แต่ละ

ตัวอย่าง ดังนั้นค่าของ b_0 และ b_1 จากตัวอย่างหนึ่งจะเป็นค่าของตัวแปรเชิงสุ่ม B_0 และ B_1 ตามลำดับ เพราะว่าค่าของ x คงที่ ค่าของ B_0 และ B_1 จะขึ้นอยู่กับการเปลี่ยนแปลงค่าของ Y หรือขึ้นอยู่กับค่าของตัวแปรเชิงสุ่ม $Y_1, Y_2, \ldots, Y_n$ ข้อสมมุติเกี่ยวกับการแจกแจงของ E_i's ทำให้ Y_i's, i = 1, ..., n ต่างมีการแจกแจงที่เป็นอิสระต่อกัน (independently distributed) ด้วย และแต่ละตัวต่างมีการแจกแจงเป็น $N(\beta_0 + \beta_1 x_i, \sigma^2)$ และเพราะว่า

$$B_1 = \frac{n\Sigma x_i Y_i - \Sigma x_i \Sigma Y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

$$= \frac{\Sigma(x_i - \bar{x})(Y_i - \bar{Y})}{\Sigma(x_i - \bar{x})^2}$$

$$= \frac{\Sigma(x_i - \bar{x})Y_i - \bar{Y}\Sigma(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2}$$

$$= \frac{\Sigma(x_i - \bar{x})Y_i}{\Sigma(x_i - \bar{x})^2} \qquad [\ \Sigma(x_i - \bar{x}) = 0\ ]$$

นั่นคือ B_1 เป็น linear function ของตัวแปรเชิงสุ่ม $Y_1, \ldots, Y_n$ ที่มีสัมประสิทธิ์ (coefficients)

$$C_i = \frac{(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2}$$

ดังนั้น B_1 จะมีการแจกแจงเป็น Normal distribution ที่มี mean

$$\mu_{B_1} = E(B_1) = \frac{\Sigma(x_i - \bar{x})\,E(Y_i)}{\Sigma(x_i - \bar{x})^2}$$

$$= \frac{\Sigma(x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\Sigma(x_i - \bar{x})^2}$$

$$= \frac{\Sigma(x_i - \bar{x})(\beta_0 + \beta_1\bar{x} + \beta_1 x_i - \beta_1\bar{x})}{\Sigma(x_i - \bar{x})^2}$$

$$= \frac{\Sigma(x_i - \bar{x})\,[\,(\beta_0 + \beta_1\bar{x}) + \beta_1(x_i - \bar{x})\,]}{\Sigma(x_i - \bar{x})^2}$$

$$= \frac{(\beta_0 + \beta_1\bar{x})\Sigma(x_i - \bar{x}) + \beta_1\Sigma(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} = \beta_1$$

และ variance $\sigma^2_{B_1} = \dfrac{\Sigma(x_i - \bar{x})^2\,\sigma^2_{Y_i}}{[\,\Sigma(x_i - \bar{x})^2\,]^2}$

$$= \frac{\sigma^2}{\Sigma(x_i - \bar{x})^2}$$

พิจารณา $B_0 = \bar{Y} - B_1\bar{x}$

$$= \frac{1}{n}\sum_{i=1}^{n} Y_i - B_1\bar{x}$$

$$E(B_0) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) - \bar{x}E(B_1)$$

$$= \frac{1}{n}\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i) - \bar{x}\beta_1$$

$$= \frac{1}{n}(n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i) - \bar{x}\beta_1$$

$$= \beta_0 + \beta_1\bar{x} - \bar{x}\beta_1$$

$$= \beta_0$$

$$Var(B_0) = Var(\bar{Y}) + Var(B_1\bar{x})$$ [$\bar{Y}$ และ B_1 เป็นอิสระต่อกัน]

$$= \frac{1}{n^2}\sum_{i=1}^{n} Var(Y_i) + \bar{x}^2\, Var(B_1)$$

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

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

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

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

$$= \left[\frac{\sum_{i=1}^{n} x_i^2}{n\,\Sigma(x_i - \bar{x})^2}\right]\sigma^2$$

ตัวแปรเชิงสุ่ม B_0 ก็มีการแจกแจงเป็น Normal distribution ที่มี mean $\mu_{B_0} = \beta_0$ และ variance

$$\sigma^2_{B_0} = \left[\frac{\Sigma x_i^2}{n\,\Sigma(x_i - \bar{x})^2} \right] \sigma^2$$

B_0 และ B_1 เป็น Unbiased estimator ของ β_0 และ β_1 ตามลำดับ ทั้งนี้เพราะ $E(B_0) = \beta_0$ และ $E(B_1) = \beta_1$

Unbiased estimate ของ σ^2 คือ $s^2 = \frac{SSE}{n-2} = MSE$ ซึ่งมี d.f. $= (n-2)$ s^2 เป็นค่าค่าหนึ่งของตัว Estimator S^2

เพื่อง่ายแก่การแสดงสูตรต่าง ๆ เราจะใช้สัญลักษณ์ต่อไปนี้

$$S_{xx} = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}$$

$$S_{yy} = \Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}$$

$$S_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n}$$

พิจารณา

$$\begin{aligned} SSE &= \Sigma(y_i - b_0 - b_1 x_i)^2 \\ &= \Sigma(y_i - \bar{y} + b_1\bar{x} - b_1 x_i)^2 \\ &= \Sigma\left[(y_i - \bar{y}) - b_1(x_i - \bar{x})\right]^2 \\ &= \Sigma(y_i - \bar{y})^2 - 2b_1\Sigma(x_i - \bar{x})(y_i - \bar{y}) + b_1^2\Sigma(x_i - \bar{x})^2 \\ &= S_{yy} - 2b_1 S_{xy} + b_1^2 S_{xx} \\ &= S_{yy} - b_1 S_{xy} \qquad \left[\because b_1 = \frac{S_{xy}}{S_{xx}}\right] \end{aligned}$$

เนื่องจาก B_1 เป็นตัวแปรเชิงสุ่มที่มีการแจกแจงเป็น Normal distribution $(\beta_1, \sigma^2/S_{xx})$

หรือ $\frac{B_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0, 1)$ และ $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$

ดังนั้น

$$\begin{aligned} T &= \frac{(B_1 - \beta_1) / \sqrt{Var(B_1)}}{\sqrt{\frac{(n-2)S^2}{\sigma^2} / (n-2)}} \\ &= \frac{(B_1 - \beta_1) / (\sigma / \sqrt{S_{xx}})}{S/\sigma} \\ &= \frac{B_1 - \beta_1}{S / \sqrt{S_{xx}}} \end{aligned}$$

จะมีการแจกแจงเป็น t_{n-2}

เราใช้ตัวสถิติ T ในการสร้าง $100(1-\alpha)\%$ C.I. ของ β_1

$100(1-\alpha)\%$ C.I. ของ β_1 คือ $b_1 \pm t_{n-2,\frac{\alpha}{2}} \frac{s}{\sqrt{S_{xx}}}$, โดยที่ $s = \sqrt{MSE}$

ตัวอย่างที่ 5.2 หา 95% C.I. ของ β_1 ในตัวแบบเชิงเส้นตรง $\mu_{Y|x} = \beta_0 + \beta_1 x$ โดยใช้ข้อมูลจากตาราง 5.1

ในตัวอย่างที่ 5.1 เราได้ $\Sigma x_i = 30.3$, $\Sigma x_i^2 = 115.11$, $\Sigma y_i = 91.1$ และ $\Sigma x_i y_i = 345.09$ จากตาราง 5.1 $\Sigma y_i^2 = 1036.65$

ดังนั้น $S_{xx} = 115.11 - \frac{(30.3)^2}{9} = 13.10$

$$S_{yy} = 1036.65 - \frac{(91.1)^2}{9} = 114.52$$

$$S_{xy} = 345.09 - \frac{(30.3)(91.1)}{9} = 38.39$$

$$b_1 = 2.9303$$

ดังนั้น $MSE = s^2 = \frac{S_{yy} - b_1 S_{xy}}{n-2}$

$$= \frac{114.52 - (2.9303)(38.39)}{7} = .2894$$

$$s = \sqrt{MSE} = \sqrt{.2894} = .5380$$

$$\sqrt{S_{xx}} = \sqrt{13.10} = 3.6194$$

$$t_{7,.025} = 2.365$$

ดังนั้น 95% C.I. ของ β_1 คือ $2.9303 \pm 2.365 \frac{(.5380)}{3.6194}$

$$= 2.9303 \pm .3515$$

$$= (2.5788, 3.2818)$$

ในการทดสอบ $H_0: \beta_1 = \beta_1'$ เราใช้ t-distribution ที่มี d.f. = n − 2 ในการสร้าง Critical region และการตัดสินใจขึ้นอยู่กับค่าของ $t_c = \frac{b_1 - \beta_1'}{s / \sqrt{S_{xx}}}$

ตัวอย่างที่ 5.3 จากตัวอย่าง 5.1 จงทดสอบ $H_0 : \beta_1 = 2.5$, ที่ $\alpha = .01$ กำหนด $H_1 : \beta_1 > 2.5$

1) $H_0 : \beta_1 = 2.5$

2) $H_1 : \beta_1 > 2.5$

3) $\alpha = .01$

4) CR : $T > t_{7,.01} = 2.998$

5) $t_c = \dfrac{2.9303 - 2.5}{.5380 / 3.6194} = 2.8948$

6) $\because t_c = 2.8948 < 2.998$ (ไม่ตกอยู่ใน CR) เราไม่สามารถปฏิเสธ H_0 ได้ที่ $\alpha = .01$ นั่นคือยอมรับว่า $\beta_1 = 2.5$

ในทำนองเดียวกันเราสามารถแสดงได้ว่า $T = \dfrac{B_0 - \beta_0}{S\sqrt{\Sigma x_i^2 / n S_{xx}}} \sim t_{n-2}$

ดังนั้นในการสร้าง $100(1-\alpha)\%$ C.I. ของ β_0 และการทดสอบ $H_0 : \beta_0 = \beta_0'$ เราจะใช้ t-distribution

$$100(1-\alpha)\% \text{ C.I. ของ } \beta_0 \text{ คือ } b_0 \pm t_{n-2,\frac{\alpha}{2}} \frac{s\sqrt{\Sigma x_i^2}}{\sqrt{n S_{xx}}}$$

ในการทดสอบสมมติฐาน $H_0 : \beta_0 = \beta_0'$ โดยใช้ H_1 ที่เหมาะสม เราใช้ t-distribution ที่มี d.f. $= n-2$ ในการสร้าง Critical region และการตัดสินใจขึ้นอยู่กับค่าของ

$$t_c = \frac{b_0 - \beta_0'}{s\sqrt{\Sigma x_i^2 / n S_{xx}}}$$

สิ่งที่สำคัญสำหรับผู้ทำการทดลอง คือ การหา C.I. ของ mean response ($\mu_{Y/x}$) ณ ค่าของ x ซึ่งไม่ใช่ค่าที่สังเกตมาก่อน

สมมุติว่าเราสนใจ mean ของ Y ที่ $x = x_0$ (μ_{Y/x_0})

เราประมาณค่าของ $\mu_{Y/x_0} = \beta_0 + \beta_1 x_0$ ด้วย $\hat{Y}_0 = B_0 + B_1 x_0$

เพราะว่า $E(\hat{Y}_0) = E(B_0 + B_1 x_0) = \beta_0 + \beta_1 x_0$ ดังนั้น $\hat{Y}_0$ เป็น Unbiased estimator ของ μ_{Y/x_0} ที่มี variance

$$\sigma^2_{\hat{Y}_0} = Var(B_0 + B_1x_0) = Var[\bar{Y} + B_1(x_0 - \bar{x})]$$

$$= Var(\bar{Y}) + (x_0 - \bar{x})^2 Var(B_1)$$

$\because B_0 + B_1x_0 = \bar{Y} - B_1\bar{x} + B_1x_0 = \bar{Y} + B_1(x_0 - \bar{x})$

($\because \bar{Y}$ และ B_1 เป็นอิสระต่อกัน)

$$= \frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \frac{\sigma^2}{\Sigma(x_i - \bar{x})^2}$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} \right]$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

ดังนั้นเราสามารถสร้าง $100(1 - \alpha)$ % C.I. ของ mean response μ_{Y/x_0} ได้จากตัวสถิติ

$$T = \frac{\hat{Y}_0 - \mu_{Y/x_0}}{S\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

$100(1 - \alpha)$ % C.I. ของ μ_{Y/x_0} คือ $\hat{y}_0 \pm t_{n-2,\frac{\alpha}{2}} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$

ตัวอย่างที่ 5.4 ใช้ข้อมูลจากตารางที่ 5.1 จงสร้าง 95% C.I. ของ $\mu_{Y/x=2}$

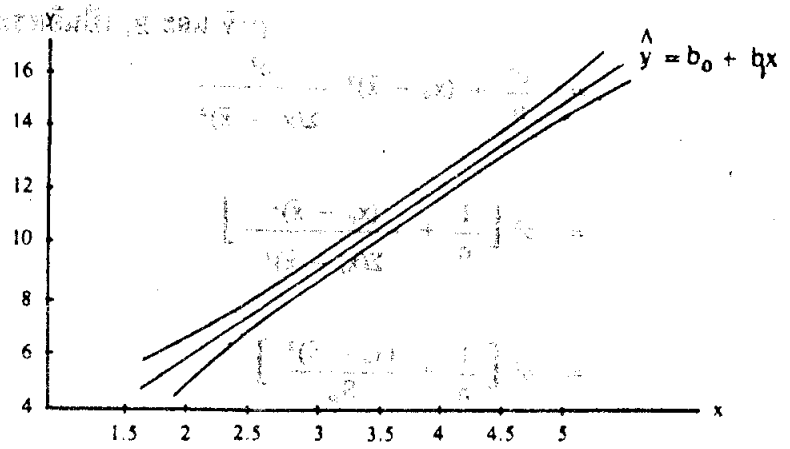
$x_0 = 2, \hat{y}_0 = .2568 + (2.9303)(2) = 6.1174$

จาก $\bar{x} = 3.3667, S_{xx} = 13.10, s = .5380$ และ $t_{7,.025} = 2.365$

95 % C.I. ของ $\mu_{Y/2}$ คือ $6.1174 \pm 2.365(.5380)\sqrt{\frac{1}{9} + \frac{(2 - 3.3667)^2}{13.10}}$

$= (5.4765, 6.7583)$

เราอาจทำการคำนวณทำนองเดียวกันโดยใช้ค่าต่าง ๆ กันของ x_0 เพื่อหา C.I. ของ μ_{Y/x_0} รูปที่ 5.3 แสดง estimated regresion line, upper และ lower confidence limits ของ μ_{Y/x_0}



รูปที่ 5.3 Confidence limits ของ $\mu_{Y/x}$

Confidence interval อีกชนิดหนึ่งที่มักจะสับสนกับ Confidence interval ของ $\mu_{Y/x}$ และการแปลความหมายมักผิดพลาดบ่อย ๆ คือ prediction interval ของผลในอนาคต ความจริงในหลาย ๆ สถานการณ์ prediction interval มักจะมีประโยชน์ต่อนักวิทยาศาสตร์, วิศวกร มากกว่า Confidence interval ของ $\mu_{Y/x}$

สมมุติเราเริ่มต้นด้วย distribution ของผลต่างของ $\hat{y}_0$ ที่ได้จาก regression line ที่คำนวณได้ และค่าจริงของ y_0 ที่ $x = x_0$

เราอาจพิจารณาค่า $\hat{y}_0 - y_0$ ว่า เป็นค่าค่าหนึ่งของตัวแปรเชิงสุ่ม $\hat{Y}_0 - Y_0$ ซึ่งมีการแจกแจงเป็น Normal distribution ซึ่งมี mean

$$\begin{aligned}
\mu_{\hat{Y}_0 - Y_0} &= E(\hat{Y}_0 - Y_0) \\
&= E[B_0 + B_1x_0 - (\beta_0 + \beta_1x_0) + E_0)] \\
&= (\beta_0 + \beta_1x_0) - (\beta_0 + \beta_1x_0) + E(E_0) \\
&= 0
\end{aligned}$$

และ variance

$$\mathrm{Var}(\hat{Y}_0 - Y_0) = \mathrm{Var}(B_0 + B_1x_0 + E_0)$$

$$= \mathrm{Var}(\bar{Y} + B_1(x_0 - \bar{x}) + E_0)$$

$$= \mathrm{Var}(\bar{Y}) + (x_0 - \bar{x})^2\,\mathrm{Var}(B_1) + \mathrm{Var}(E_0)$$

$$= \frac{\sigma^2}{n} + \frac{(x_0 - \bar{x})^2\sigma^2}{S_{xx}} + \sigma^2$$

$$= \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

ดังนั้น $100(1 - \alpha)$ % C.I. ของ y_0 สามารถสร้างได้จากตัวสถิติ

$$T = \frac{\hat{Y}_0 - Y_0}{S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

$100(1-\alpha)$ % C.I. ของ individual y_0 คือ $\hat{y}_0 \pm t_{n-2,\,\alpha/2}\, s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$

ตัวอย่างที่ 5.5 ใช้ข้อมูลจากตารางที่ 5.1 จงสร้าง 95% C.I. ของ y_0 ที่ $x_0 = 2$

$n = 9,\ x_0 = 2,\ \bar{x} = 3.3667,\ \hat{y}_0 = 6.1171,\ S_{xx} = 13.10$

$s = .5380,\ t_{7,.025} = 2.365$

95% C.I. ของ y_0 ที่ $x_0 = 2$ คือ $6.1171 \pm (2.365)(.5380)\sqrt{1 + \frac{1}{9} + \frac{(2-3.3667)^2}{13.10}}$

$= (4.6927,\ 7.5421)$

ข้อสังเกต

C.I. ของ μ_{Y/x_0} และ individual y_0 จะสั้นที่สุด เมื่อ $x_0 = \bar{x}$ และจะยาวขึ้น ๆ เมื่อ x_0 ยิ่งห่างไปจาก $\bar{x}$

5.3 Analysis of variance approach

จากหัวข้อ 5.2 $S_{yy} = b_1S_{xy} + SSE$

เราอาจเขียนใหม่เป็น SST = SSR + SSE นั่นคือ แบ่ง SST ออกเป็น 2 ส่วน คือ ส่วนที่ 1 SSR = Regression sum of squares เป็นส่วนของความแปรปรวนของ Y ที่เนื่องมาจากอิทธิพลของ x หรือส่วนของความแปรปรวนของ Y ที่อธิบายได้โดยตัวแบบเชิงเส้นตรงที่เราใช้ และส่วนที่ 2 SSE = Error sum of squares เป็นส่วนของความแปรปรวนของ Y ที่เนื่องมาจาก random error

เนื่องจาก SSR และ SSE เป็นค่าของตัวแปรเชิงสุ่ม Chi-square 2 ตัว ที่เป็นอิสระต่อกัน และมี d.f. = 1 และ (n − 2) ตามลำดับ ดังนั้น SST จะเป็นค่าของตัวแปรเชิงสุ่ม Chi-square ที่มี d.f. = n − 1 (ดู 2.13.1.2)

ในการทดสอบ $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$ เราคำนวณ

$$f_c = \frac{SSR/1}{SSE/(n-2)} = \frac{SSR}{MSE} = \frac{SSR}{s^2}$$

เราจะปฏิเสธ H_0 ที่ระดับนัยสำคัญ α เมื่อ $f_c > f_{(1,n-2),\alpha}$

ในทางปฏิบัติเราคำนวณ $SST = S_{yy}$

$SSR = b_1S_{xy}$

และ $SSE = SST - SSR$

สรุปค่าต่าง ๆ ที่คำนวณได้ลงในตาราง ANOVA

ANOVA

S.V.	d.f.	SS	MS	f_c
Regression	1	SSR	MSR = SSR	SSR/s^2
Error	n − 2	SSE	$MSE = s^2$	
Total	n − 1	SST		

ถ้า $f_c > f_{(1,n-2),\alpha}$ เราจะปฏิเสธ H_0 และสรุปว่า ค่าความแปรปรวนของ Y ที่เนื่องมาจาก x มีมากพอที่จะยอมรับว่า x เป็นเหตุของผล Y

ถ้า $f_c < f_{(1,n-2),\alpha}$ เราจะ ยอมรับ H_0 และสรุปว่าค่าความแปรปรวนของ Y ที่เนื่องมาจาก x มีไม่มากพอที่จะยอมรับว่า x เป็นเหตุของผล Y

ข้อสังเกต

ในการทดสอบ $H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$ โดยใช้ t - test ตัวสถิติที่ใช้ในการทดสอบคือ

$$\Gamma = \frac{B_1}{S/\sqrt{S_{xx}}} \text{ ดังนั้น } t_c = \frac{b_1}{s/\sqrt{S_{xx}}}$$

พิจารณา $t_c^2 = \dfrac{b_1^2}{s^2/S_{xx}} = \dfrac{b_1^2 S_{xx}}{s^2} = \dfrac{b_1 S_{xy}}{s^2} = \dfrac{SSR}{s^2} = f_c$

ความสัมพันธ์ระหว่าง t_ν และ $f_{(1,\nu)}$ คือ

$$t^2_{\nu,\frac{\alpha}{2}} = f_{(1,\nu),\alpha}$$

5.4 สรุปสูตรในการวิเคราะห์การถดถอย (Regression Analysis)

ตัวแบบเชิงเส้นตรง (Straight line model)

$Y_i = \beta_0 + \beta_1 x_i + E_i$ หรือ $\mu_{Y/x_i} = \beta_0 + \beta_1 x_i$

$i = 1, \ldots, n$

โดยที่ Y = ตัวแปรตาม (dependent variable)

x = ตัวแปรอิสระ (independent variable)

β_0 = ค่าของ $\mu_{Y/x}$ เมื่อ x = 0

β_1 = สัมประสิทธิ์การถดถอย (regression coefficient)

E_i = random error

ข้อมูล : (x_i, y_i), $i = 1, \ldots, n$

คำนวณ $S_{xx} = \Sigma x_i^2 - (\Sigma x_i)^2/n$

$S_{yy} = \Sigma y_i^2 - (\Sigma y_i)^2/n$

$S_{xy} = \Sigma x_i y_i - \Sigma x_i \Sigma y_i/n$

$\bar{x} = \Sigma x_i/n$

$\bar{y} = \Sigma y_i/n$

1) ค่าของ b_0 และ b_1 ซึ่งเป็นค่าของ B_0 และ B_1 (least square estimators และเป็น Unbiased estimators ของ β_0 และ β_1 ตามลำดับ)

$$b_1 = S_{xy}/S_{xx}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

2) Sample regression equation หรือ prediction equation

$$\hat{y} = b_0 + b_1x$$

3) $SSE = \Sigma e_i^2 = S_{yy} - b_1S_{xy} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$

4) $\hat{\sigma}^2 = s^2 = MSE = \frac{SSE}{n-2}$

5) $100(1-\alpha)\%$ C.I. ของ β_0 คือ $b_0 \pm t_{n-2,\frac{\alpha}{2}} \frac{s\sqrt{\Sigma x_i^2}}{\sqrt{n S_{xx}}}$

6) $100(1-\alpha)\%$ C.I. ของ β_1 คือ $b_1 \pm t_{n-2,\frac{\alpha}{2}} \frac{s}{\sqrt{S_{xx}}}$

7) $100(1-\alpha)\%$ C.I. ของ mean response μ_{Y/x_0} คือ

$$\hat{y}_0 \pm t_{n-2,\frac{\alpha}{2}} s \sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}$$, โดยที่ $\hat{y}_0 = b_0 + b_1x_0$

8) $100(1-\alpha)\%$ C.I. ของ individual y_0 $(y/x = x_0)$ คือ

$$\hat{y}_0 \pm t_{n-2,\frac{\alpha}{2}} s \sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}$$, โดยที่ $\hat{y}_0 = b_0 + b_1x_0$

9) $H_0 : \beta_0 = \beta'_0$ ใช้ $t_C = \dfrac{b_0 - \beta'_0}{s\sqrt{\Sigma x_i^2/n\, S_{xx}}}$ ซึ่งเป็นค่าของตัวสถิติ T ซึ่ง

ถ้า H_0 จริง มันจะมีการแจกแจงเป็น t - distribution ที่มี d.f. = n − 2

10) $H_0 : \beta_1 = \beta'_1$ ใช้ $t_C = \dfrac{b_1 - \beta'_1}{s/\sqrt{S_{xx}}}$ ซึ่งเป็นค่าของตัวสถิติ T ซึ่งถ้า H_0 จริง มันจะมีการแจกแจงเป็น t - distribution ที่มี d.f. = n − 2

11) $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$ อาจใช้ $t_C = \dfrac{b_1}{s/\sqrt{S_{xx}}}$ ซึ่งถ้า H_0 จริงมันจะเป็นค่าของตัวสถิติ T ซึ่งมีการแจกแจงเป็น t - distribution ที่มี d.f. = n − 2 หรือใช้

$f_C = \dfrac{MSR}{MSE} = \dfrac{SSR}{s^2}$ ซึ่งถ้า H_0 จริงมันจะเป็นค่าของตัวสถิติ F ซึ่งมีการแจกแจงเป็น F - distribution ที่มี d. f. = (1, n − 2) โดยที่

$$MSR = b_1 S_{xy}$$

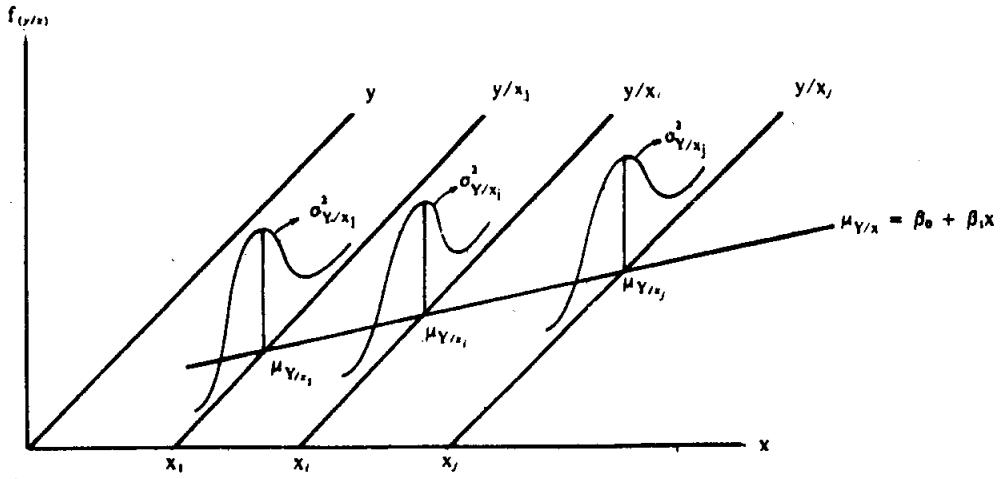
$$MSE = \frac{SSE}{n-2} = \frac{S_{yy} - b_1 S_{xy}}{n-2}$$

12) ข้อสมมุติในการวิเคราะห์การถดถอยเชิงเส้นตรง

ก) x เป็นค่าคงที่ที่เรากำหนดขึ้น (fixed constant) ถือว่า x วัดได้ถูกต้องแน่นอน ไม่มีความผิดพลาด

ข) ที่ $x = x_i$, $Y/x_i = Y_i \sim N(\mu_{Y/x_i}, \sigma^2_{Y/x_i})$ โดยที่ $\mu_{Y/x_i} = \beta_0 + \beta_1 x_i$ และ Y_i's, i = 1, ..., n เป็นอิสระต่อกัน

ค) $\sigma^2_{Y/x_1} = \sigma^2_{Y/x_2} = \ldots = \sigma^2_{Y/x_n} = \sigma^2_E = \sigma^2$ (error variance)



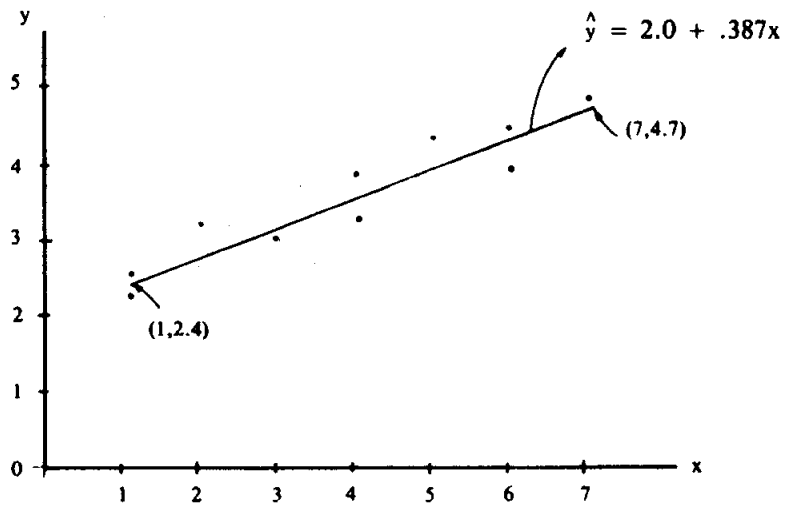
รูปที่ 5.4 แสดงการแจกแจงของตัวแปรตาม Y/x_i และ regression line

ตัวอย่างที่ 5.6 จากตารางแสดงจำนวน additive ในน้ำมันและปริมาณ Nitrogen Oxides ที่ลดลงจากไอเสียรถยนต์ 10 คัน จงวิเคราะห์ข้อมูลโดยใช้ตัวแบบเชิงเส้นตรง

$\mu_{Y/x_i} = \beta_0 + \beta_1 x_i, i = 1, \ldots, 10$

ปริมาณ additive : x	1	1	2	3	4	4	5	6	6	7
ปริมาณ Nitrogen oxides ที่ลดลง	2.1	2.5	3.1	3.0	3.8	3.2	4.3	3.9	4.4	4.8

Plot scatter diagram



จาก Scatter diagram แสดงว่า การใช้ตัวแบบเชิงเส้นตรงสมเหตุสมผล
จากข้อมูลข้างต้นคำนวณ

$n = 10, \quad \Sigma x_i = 39, \ \Sigma x_i^2 = 193, \ \bar{x} = 3.9$

$\Sigma y_i = 35.1, \ \Sigma y_i^2 = 130.05, \ \bar{y} = 3.51$

$\Sigma x_i y_i = 152.7$

$S_{xx} = 193 - (39)^2/10 = 40.9$

$S_{yy} = 130.05 - (35.1)^2/10 = 6.85$

$S_{xy} = 152.7 - (39)(35.1)/10 = 15.81$

1) $\hat{\beta}_1 = b_1 = S_{xy}/S_{xx} = 15.81/40.9 = .387$

$\hat{\beta}_0 = b_0 = \bar{y} - b_1\bar{x} = 3.51 - (.387)(3.9) = 2.0$

2) Prediction equation : $\hat{y} = 2.0 + .387x$

$x_1 = 1, \ \hat{y} = 2.387$

$x_{10} = 7, \hat{y}_{10} = 4.709$

ลากเส้นเชื่อม 2 จุด (1, 2.4) และ (7, 4.7) จะได้ prediction line $\hat{y} = 2.0 + .387x$

3) $SSE = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} = 6.85 - \frac{(15.81)^2}{40.9} = .74$

4) $\hat{\sigma}^2 = MSE = s^2 = \frac{.74}{8} = .0925$

5) $t_{8,.025} = 2.306$

95% C.I. ของ β_0 คือ $2.0 \pm 2.306\sqrt{\frac{.0925(193)}{10(40.9)}}$

$= 2.0 \pm 2.306(.209)$

$= 2.0 \pm .482 = (1.518, 2.482)$

6) 95% C.I. ของ β_1 คือ $.387 \pm 2.306 \sqrt{\frac{.0925}{40.9}}$

$= .387 \pm .110$

$= (.277, .497)$

7) ถ้าจำนวน additive ที่ให้แก่รถยนต์ = 4 จงหา C.I. ของค่าเฉลี่ยของปริมาณ Nitrogen Oxides ที่จะลดลงสำหรับรถยนต์ใด ๆ ที่ได้รับ additive = 4

95% C.I. ของ $\mu_{Y/4}$ คือ $\hat{y} \pm t_{8,.025} \sqrt{\left(\frac{1}{10} + \frac{(4-\bar{x})^2}{S_{xx}}\right) s^2}$

$$\hat{y} = b_0 + b_1(4) = 2.0 + (.387)(4) = 3.548$$

$$\sqrt{.0925\left(\frac{1}{10} + \frac{(4-3.9)^2}{40.9}\right)} = .096$$

$\therefore$ 95% C.I. ของ $\mu_{Y/4}$ คือ $3.548 \pm 2.306(.096)$

$= 3.548 \pm .22$

$= (3.33, 3.77)$

8) ถ้าจำนวน additive ที่ให้แก่รถ Mustang คันหนึ่งมีปริมาณ 4.5 จงหา C.I. ของปริมาณ Nitrogen Oxides ที่จะลดลงสำหรับรถยนต์คันนี้

$$\hat{y} = b_0 + b_1(4.5) = 2.0 + (.387)(4.5) = 3.74$$

$$\sqrt{s^2\left(1 + \frac{1}{n} + \frac{(4.5-\bar{x})^2}{S_{xx}}\right)} = \sqrt{.0925\left(1 + \frac{1}{10} + \frac{(4.5-3.9)^2}{40.9}\right)} = .32$$

$\therefore$ 95% C.I. ของ y ที่ x = 4.5 คือ $3.74 \pm 2.306(.32)$

$= 3.74 \pm (.7379)$

$= (3.0021, 4.4779)$ (ความยาวของ interval = 1.4758)

ในทำนองเดียวกัน

95% C.I. ของ y ที่ $x = 3.9 = \bar{x}$ คือ $3.51 \pm 2.306(.319)$

$= 3.51 \pm .7356$

$= (2.7744, 4.2456)$ (ความยาวของ interval = 1.4712)

9) $H_0 : \beta_0 = 0$, $H_1 : \beta_0 > 0$, $\alpha = .05$, CR : $T > t_{8,.05} = 1.86$

$$t_C = \frac{b_0}{\sqrt{s^2 \Sigma x_i^2 / n\, S_{xx}}}$$

$$= \frac{2.0}{.209} = 9.569 > 1.86$$ ดังนั้นเราปฏิเสธ H_0 ที่ $\alpha = .05$

นั่นคือ Regression line ไม่ผ่านจุด origin

10) $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$, $\alpha = .05$, CR : $T > t_{8,.025} = 2.306$ และ $T < -t_{8,.025} = -2.306$

$$t_C = \frac{b_1}{\sqrt{s^2/S_{xx}}} = \frac{.387}{.0475564} = 8.14$$ ดังนั้นเราปฏิเสธ H_0 ที่ $\alpha = .05$

นั่นคือยอมรับว่า x เป็นเหตุของผล Y

11) $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$, $\alpha = .05$, CR : $F > f_{(1,8),.05} = 5.318$

$SSR = MSR = b_1 S_{xy} = 6.11$

$MSE = (S_{yy} - SSR)/8 = (6.85 - 6.11)/8 = .74/8 = .0925$ (หรือจากข้อ 4 MSE = .0925)

$$f_C = \frac{MSR}{MSE} = \frac{6.11}{.0925} = 66.054 > 5.318$$ เราปฏิเสธ H_0 ที่ $\alpha = .05$

($t_c^2 = 8.14^2 = 66.26 \approx f_c$ ความแตกต่างเนื่องมาจากการปัดเศษทศนิยม)

5.5 การทดสอบการเท่ากันของ slopes จากสมการ regression 2 สมการ

$(H_0 : \beta_1^{(1)} = \beta_1^{(2)})$: Test for Parallelism

โครงสร้างของข้อมูล

x_1 ของวิธีการที่ 1	x_{11}	x_{12}	...	x_{1i}	...	x_{1n_1}
y_1	y_{11}	y_{12}	...	y_{1i}	...	y_{1n_1}

x_2 ของวิธีการที่ 2	x_{21}	x_{22}	...	x_{2i}	...	x_{2n_2}
y_2	y_{21}	y_{22}	...	y_{2i}	...	y_{2n_2}

Model 1 : $Y_{1i} = \beta_0^{(1)} + \beta_1^{(1)} x_{1i} + E_{1i}$, $i = 1, \ldots, n_1$

Model 2 : $Y_{2i} = \beta_0^{(2)} + \beta_1^{(2)} x_{2i} + E_{2i}$, $i = 1, \ldots, n_2$

ข้อสมมุติเพิ่มเติมจากหัวข้อข้างต้น : $Var(E_{1i}) = Var(E_{2i}) = \sigma^2$

5.5.1 ทดสอบ $H_0 : \beta_1^{(1)} = \beta_1^{(2)}$

กำหนดให้ $\widehat{\beta_1^{(1)}} = b_1^{(1)}$ เป็นค่าของตัวแปรเชิงสุ่ม $B_1^{(1)}$

$\widehat{\beta_1^{(2)}} = b_1^{(2)}$ เป็นค่าของตัวแปรเชิงสุ่ม $B_1^{(2)}$

SSE(1) และ SSE(2) แทน error sum of squares จาก model 1 และ 2 ตามลำดับ

$$S_{xx}^{(1)} = \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2 = \sum_{i=1}^{n_1} x_{1i}^2 - (\sum_{i=1}^{n_1} x_{1i})^2/n_1$$

$$S_{xx}^{(2)} = \sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2 = \sum_{i=1}^{n_2} x_{2i}^2 - (\sum_{i=1}^{n_2} x_{1i})^2/n_2$$

pooled estimate ของ σ^2 คือ

$$\hat{\sigma}^2 = s_{pooled}^2 = \frac{SSE(1) + SSE(2)}{n_1 + n_2 - 4}$$

$$(B_1^{(1)} - B_1^{(2)}) \sim N(\beta_1^{(1)} - \beta_1^{(2)}, \sigma^2 (\frac{1}{S_{xx}^{(1)}} + \frac{1}{S_{xx}^{(2)}}))$$

$$t = \frac{(b_1^{(1)} - b_1^{(2)}) - (\beta_1^{(1)} - \beta_1^{(2)})}{S_{pooled}\sqrt{\frac{1}{S_{xx}^{(1)}} + \frac{1}{S_{xx}^{(2)}}}}$$ เป็นค่าของตัวสถิติ $T \sim t_{n_1+n_2-4}$

ถ้า H_0 จริง $t_C = \frac{b_1^{(1)} - b_1^{(2)}}{S_{pooled}\sqrt{\frac{1}{S_{xx}^{(1)}} + \frac{1}{S_{xx}^{(2)}}}}$ เป็นค่าของ $T \sim t_{n_1+n_2-4}$

5.5.2 $100(1 - \alpha)\%$ C.I. ของ $\beta_1^{(1)} - \beta_1^{(2)}$ คือ

$$(b_1^{(1)} - b_1^{(2)}) \pm t_{n_1+n_2-4,\frac{\alpha}{2}} S_{pooled}\sqrt{\frac{1}{S_{xx}^{(1)}} + \frac{1}{S_{xx}^{(2)}}}$$

ตัวอย่างที่ 5.7 นำวิธีเตรียมอินซูลิน (Insulin) 2 วิธี มาศึกษาดูอิทธิพลในการลดอัตราน้ำตาลในเลือดของหนู หนู 13 ตัวที่มีพันธุ์เดียวกันถูกแบ่งออกเป็น 2 กลุ่มอย่างสุ่ม โดยกลุ่มที่ 1 มีหนู 7 ตัว และอีกกลุ่มหนึ่งมีหนู 6 ตัว ฉีดอินซูลิน A แก่หนู 7 ตัว และฉีดอินซูลิน B แก่หนู 6 ตัว ปริมาณน้ำตาลในเลือดที่ลดลงเป็นดังนี้

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

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

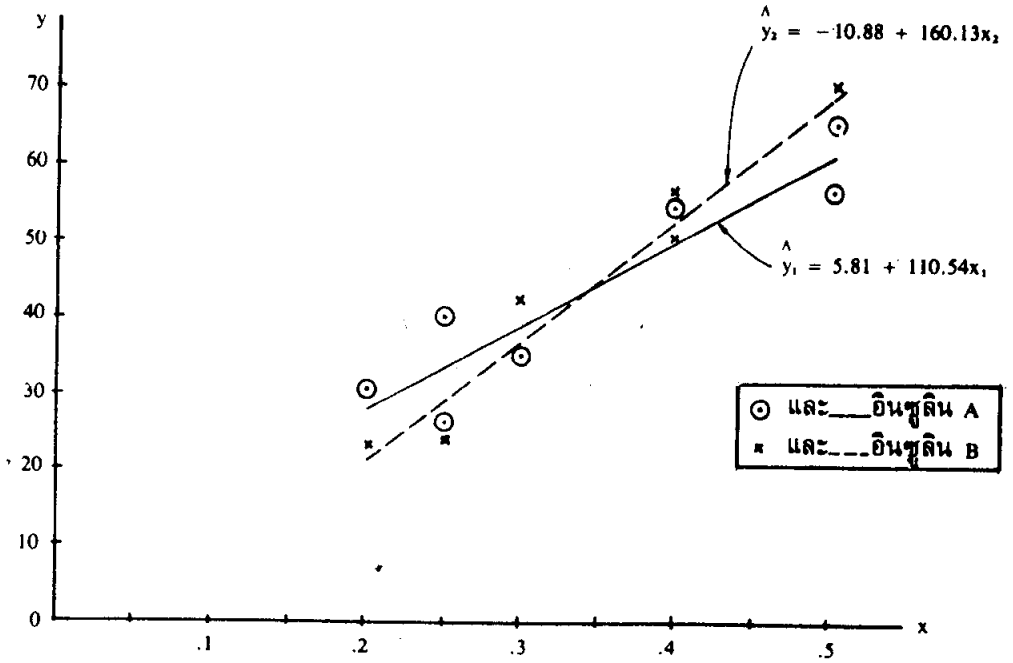
1) Plot scatter diagram สำหรับข้อมูล 2 ชุด ลงบนกระดาษกราฟแผ่นเดียวกัน ใช้สัญลักษณ์สำหรับแต่ละชุดให้ต่างกัน

2) คำนวณ sample regression line สำหรับข้อมูลแต่ละชุดแล้วลากเส้นทั้ง 2 ลงบนกระดาษกราฟในข้อ 1)

3) ข้อมูลสนับสนุนหรือไม่ว่าอัตราการลดลงของระดับน้ำตาลในเลือดเมื่อใช้อินซูลิน A และ B ไม่แตกต่างกัน นั่นคือทดสอบ $H_0 : \beta_1^{(1)} = \beta_1^{(2)}$ ให้ทดสอบที่ $\alpha = .05$

4) หา 95% C.I. ของ $\beta_1^{(1)} - \beta_1^{(2)}$

1) Plot scatter diagram



2) จะหา $b_0^{(1)}, b_1^{(1)}$ และ $\hat{y}_1 = b_0^{(1)} + b_1^{(1)}x_1$

$b_0^{(2)}, b_1^{(2)}$ และ $\hat{y}_2 = b_0^{(2)} + b_1^{(2)}x_2$

จากข้อมูลชุด A : $\Sigma x_{1i} = 2.4$, $\Sigma x_{1i}^2 = .915$, $\bar{x}_1 = .3429$

($n_1 = 7$) $\Sigma y_{1i} = 306$, $\Sigma y_{1i}^2 = 14678$, $\bar{y}_1 = 43.7143$

$\Sigma x_{1i}y_{1i} = 115.1$

$S_{xx}^{(1)} = .915 - (2.4)^2/7 = .0921$

$S_{yy}^{(1)} = 14678 - (306)^2/7 = 1301.4286$

$S_{xy}^{(1)} = 115.1 - (2.4)(306)/7 = 10.1857$

$b_1^{(1)} = 110.54$

$b_0^{(1)} = 5.81$

$\hat{y}_1 = 5.81 + 110.54x_1$

ถ้า $x_1 = .2$, $\hat{y}_1 = 27.918$

ถ้า $x_1 = .5$, $\hat{y}_1 = 61.08$

ลากเส้นตรงต่อจุด (.2, 27.918) และ (.5, 61.08) จะได้ sample regression line

$\hat{y}_1 = 5.81 + 110.54x_1$

จากข้อมูลชุด B : $\Sigma x_{2i} = 2.05$, $\Sigma x_{2i}^2 = .7625$, $\bar{x}_2 = .3417$

$(n_2 = 6)$ $\Sigma y_{2i} = 263$, $\Sigma y_{2i}^2 = 13195$, $\bar{y}_2 = 43.8333$

$\Sigma x_{2i}y_{2i} = 99.8$

$S_{xx}^{(2)} = .7625 - (2.05)^2/6 = .0621$

$S_{yy}^{(2)} = 13195 - (263)^2/6 = 1666.8333$

$S_{xy}^{(2)} = 99.8 - (2.05)(263)/6 = 9.9417$

$b_1^{(2)} = 160.13$

$b_0^{(2)} = -10.88$

$\hat{y}_2 = -10.88 + 160.13x_2$

ถ้า $x_2 = .2$, $\hat{y}_2 = 21.146$

ถ้า $x_2 = .5$, $\hat{y}_2 = 69.185$

ลากเส้นตรงต่อจุด (.2, 21.146) และ (.5, 69.185) จะได้ sample regression line

$\hat{y}_2 = -10.88 + 160.13x_2$

3) $SSE(1) = S_{yy}^{(1)} - \dfrac{(S_{xy}^{(1)})^2}{S_{xx}^{(1)}} = 1301.4286 - \dfrac{(10.1857)^2}{.0921} = 175.4735$

$$SSE(2) = S_{yy}^{(2)} - \frac{(S_{xy}^{(2)})^2}{S_{xx}^{(2)}} = 166.8333 - \frac{(9.9417)^2}{.0621} = 74.8313$$

$$\hat{\sigma}^2 = s_{pooled}^2 = \frac{175.4735 + 74.8313}{9} = \frac{250.3048}{9} = 27.8116$$

$$s_{pooled} = 5.2737$$

$$\sqrt{\frac{1}{S_{xx}^{(1)}} + \frac{1}{S_{xx}^{(2)}}} = 5.1923$$

ทดสอบ $H_0 : \beta_1^{(1)} = \beta_1^{(2)}$

$H_1 : \beta_1^{(1)} \neq \beta_1^{(2)}$

$\alpha = .05$

CR : $T < -t_{9,.025} = -2.262$ และ $T > 2.262$

$t_C = \frac{110.54 - 160.13}{5.2737(5.1923)} = -1.81 > -2.262$ เราไม่สามารถปฏิเสธ H_0

ได้ที่ $\alpha = .05$ นั่นคือ ข้อมูลสนับสนุนว่าอัตราการลดของระดับน้ำตาลในเลือดเมื่อใช้อินซูลิน-A และ B ไม่แตกต่างกัน

4) 95% C.I. ของ $\beta_1^{(1)} - \beta_1^{(2)}$ คือ

$$(b_1^{(1)} - b_1^{(2)}) \pm t_{9,.025}\, s_{pooled} \sqrt{\frac{1}{S_{xx}^{(1)}} + \frac{1}{S_{xx}^{(2)}}}$$

$$= (110.54 - 160.13) \pm 2.262(5.2737)(5.1923)$$

$$= -49.59 \pm 61.94$$

$$= (-111.53, 12.35)$$

จาก C.I. นี้เราไม่สามารถสรุปได้ว่า $\beta_1^{(1)}$ และ $\beta_1^{(2)}$ แตกต่างกันอย่างมีนัยสำคัญ เพราะ C.I. นี้รวมศูนย์ไว้ด้วย ผลสรุปนี้เหมือนกับผลสรุปในข้อ 3)

5.6 การทดสอบ Lack of fit

ในหัวข้อนี้จะอธิบายวิธีการทดสอบการขาดความพอดีของข้อมูลกับตัวแบบเชิงเส้นตรง (Straight - line model) แต่เราสามารถจะใช้วิธีการนี้กับตัวแบบรูปอื่น ๆ ได้

เราจะทดสอบ H_0 : ไม่มี lack of fit หรือ ตัวแบบเป็นสมการเส้นตรง

H_1 : มี lack of fit หรือ ตัวแบบไม่เป็นสมการเส้นตรง

สมมุติว่า มีค่า x ที่ต่างกัน k ค่า, สำหรับ x แต่ละค่าเราสังเกตค่า y ซ้ำ $n_1, n_2, \ldots, n_k$ ครั้ง โดยที่ n_i อย่างน้อย 1 ตัว มีค่า ≥ 2

ตารางที่ 5.2

แสดงข้อมูลที่สามารถจะนำไปใช้ในการทดสอบ Lack of fit ได้

ค่าต่าง ๆ ของ x	ค่า y's	Mean	SS	d.f.	
x_1	$y_{11}, \ldots, y_{1n_1}$	$\bar{y}_1$	SS_1	$n_1 - 1$	
x_2	$y_{21}, \ldots, y_{2n_2}$	$\bar{y}_2$	SS_2	$n_2 - 1$	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
x_k	$y_{k1}, \ldots, y_{kn_k}$	$\bar{y}_k$	SS_k	$n_k - 1$	
		Total	SS(p.e.)	$n - k$	$n = \sum_{i=1}^{k} n_i$

โดยที่ $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$

$$SS_i = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \sum_{j=1}^{n_i} y_{ij}^2 - (\sum_{j=1}^{n_i} y_{ij})^2/n_i, \text{ d.f.} = n_i - 1$$

SS(p.e.) = pure error sum of squares

$$= \sum_{i=1}^{k} SS_i$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{k} \frac{\left(\sum_{i=1}^{n_i} y_{ij}\right)^2}{n_i}, \text{ d.f.} = \sum_{i=1}^{k} (n_i - 1) = n - k$$

$$MS(p.e.) = SS(p.e.)/(n - k)$$

= mean square for pure error เป็น Unbiased estimate ของ error variance σ^2

ถ้าเรามีค่าสังเกต n ค่า เราสามารถคำนวณ regression line และค่า residual sum of squares (SSE) โดยใช้สูตร

$$SSE = S_{yy} - b_1^2 S_{xx} = S_{yy} - b_1 S_{xy}, \text{ d.f.} = n - 2$$

เราจะแยก SSE ออกเป็น 2 ส่วน คือ SS(p.e.) และอีกส่วนหนึ่งคือ SS(lof) = lack of fit sum of squares

นั่นคือ $SSE = SS(p.e.) + SS(lof)$

$\therefore$ $SS(lof) = SSE - SS(p.e.)$, d.f. $= (n - 2) - (n - k) = k - 2$

ดังนั้น $MS(lof) = SS(lof)/(k - 2)$

ถ้า model ถูกต้อง MS(lof) จะเป็นค่าประมาณของ error variance σ^2 เช่นเดียวกับ MS(p.e.) แต่ถ้า model ที่แท้จริงไม่ใช่ straight line model MS(lof) จะ overestimate σ^2 และจะมีค่ามากกว่า M.S.(p.e.)

การทดสอบ Lack of fit เราจะคำนวณค่า

$$f_c = \frac{MS(lof)}{MS(p.e.)} = \frac{(SSE - SS(p.e.))/(k - 2)}{SS(p.e.)/(n - k)}$$ ซึ่งเป็นค่าของตัวสถิติ

$F \sim F_{(k-2, n-k)}$

ถ้า $f_c > f_{(k-2, n-k), \alpha}$ เราจะปฏิเสธ H_0 และสรุปว่ามี lack of fit และเราควรจะหาตัวแบบใหม่สำหรับแสดงความสัมพันธ์ระหว่าง X และ Y

ถ้า $f_c < f_{(k-2, n-k), \alpha}$ เราไม่สามารถปฏิเสธ H_0 และสรุปว่ามีประโยชน์น้อยมากที่จะลองใช้ตัวแบบที่ยุ่งยากมากกว่าตัวแบบเชิงเส้นตรงนี้ แต่วิธีการทดสอบ Lack of fit อย่างเดียวยังไม่สามารถรับประกันความพอเพียงของตัวแบบเชิงเส้นตรงที่ใช้อยู่ เราจะใช้ขนาดของ r^2 : สัมประสิทธิ์ของการตัดสินใจ (coefficient of determination) เข้าช่วยการตัดสินด้วย

หมายเหตุ

ในกรณีที่เราปฏิเสธ H_0 $\hat{\sigma}^2$ = MS(p.e.) (แทนที่จะใช้ MSE เป็นค่าประมาณของ error variance)

ตัวอย่าง 5.8 จากข้อมูลข้างล่างนี้ จงหา sample regression equation และทดสอบ lack of fit

x	2	2	2	3	3	4	5	5	6	6	6
y	4	3	8	18	22	24	24	18	13	10	16

$n = 11$, $\Sigma x_i = 44$, $\Sigma x_i^2 = 204$ $\bar{x} = 4$

$\Sigma y_i = 160$, $\Sigma y_i^2 = 2898$, $\bar{y} = 14.545$

$\Sigma x_i y_i = 690$

$S_{xx} = 28$

$S_{yy} = 571$

$S_{xy} = 50$

$\hat{\beta}_1 = b_1 = 1.786$

$\hat{\beta}_0 = b_0 = 7.401$

sample regression line (prediction equation) : $\hat{y} = 7.401 + 1.786x$

SST = 571 , d.f. = 10

SSR = 90 , d.f. = 1

SSE = 571 − 90 = 481 , d.f. = 9

จัดข้อมูลให้อยู่ในรูปเดียวกับตารางที่ 5.2

x	y	$\bar{y}_i$	SS_i	d.f.	
2	4,3,8	5	14	2	
3	18,22	20	8	1	
4	24	24	0	0	
5	24,18	21	18	1	
6	13,10,16	13	18	2	
		Total	58	6	n = 11

$x = 2, \bar{y} = 5$ หา $SS_1 = (4-5)^2 + (3-5)^2 + (8-5)^2 = 14$

หรือ $SS_1 = (4^2 + 3^2 + 8^2) - \frac{(15)^2}{3} = 89 - 75 = 14$ เป็นต้น

$\therefore SS(p.e.) = 14 + 8 + 0 + 18 + 18 = 58$, d.f. = 6

ANOVA

S.V.	d.f.	SS	MS	f_c
Regression	1	90	90	$f_1 = \frac{90}{53.4444} = 1.684$
Error	9	481	53.4444	
Lack of fit	3	423	141	$f_2 = \frac{141}{9.6667} = 14.586^{**}$
Pure error	6	58	9.6667	
Total	10	571		

1) H_0 : ไม่มี lack of fit

2) H_1 : มี lack of fit

3) $\alpha = .01$

4) CR : $F > f_{(3,6),.01} = 9.78$

5) $f_c = f_2 = 14.586$

6) $f_c > 9.78$ เราปฏิเสธ H_0 ที่ $\alpha = .01$ นั่นคือ ตัวแบบไม่เป็นสมการเชิงเส้นตรง เราต้องปรับปรุงตัวแบบ

ในการทำ scatter diagram เพื่อเป็นการตรวจสอบกับวิธีการทดสอบ lack of fit นี้ จากรูปที่ 5.5 จะเห็นว่าตัวแบบที่แสดงความสัมพันธ์ระหว่าง X และ Y ไม่ควรเป็นเส้นตรง ควรเป็นเส้นโค้ง

ถ้าคำนวณ r^2 (coefficient of determination)

$$100r^2\% = \frac{SSR}{SST} \times 100 = \frac{90}{571} \times 100 = 15.76\%$$

นั่นคือ 15.76% ของความแปรปรวนของ Y เท่านั้น ที่เนื่องจากความสัมพันธ์กับ X ในรูปเชิงเส้นตรง