การวิเคราะห์การถดถอย

Models are to be used, but not to belived

Henry Theil

ตัวแปรทางจิตวิทยาและการศึกษามักจะมีความสัมพันธ์กัน เช่น ผลการเรียนและระดับสติปัญญา หรือผลการเรียน, ระดับสติปัญญา, แบะการยอมรับตนเอง เป็นต้น วิธีการทางสถิติที่ศึกษาถึงความสัมพันธ์ระหว่างตัวแปรได้ชื่อว่า การวิเคราะห์การถดถอย และการวิเคราะห์สหสัมพันธ์ (Regression and Correlation Analysis) สำหรับการวิเคราะห์สหสัมพันธ์จะได้กล่าวในบทต่อไป

การวิเคราะห์การถดถอยเป็นวิธีการศึกษาทางสถิติ (Statistical Investigations) ที่ศึกษาถึงความสัมพันธ์ระหว่างตัวแปรตัวหนึ่งที่เรียกว่าตัวแปรตาม (Dependent Variable) กับตัวแปรอื่น ๆ (หนึ่งตัวหรือมากกว่า) ที่เรียกว่าตัวแปรอิสระ Independent Variables) ว่าตัวแปรเหล่านั้นมีความสัมพันธ์กันอย่างไร (How) และความสัมพันธ์ระหว่างตัวแปรต่าง ๆ นั้นสามารถสรุปได้ในรูปของตัวแปร หรือสมการทางคณิตศาสตร์ (Mathematical Model or Equation) ได้ จากตัวแบบที่สร้างขึ้นนี้เราสามารถใช้ทำนายหรือพยากรณ์ (Predict of forecast) ค่าของตัวแปรตามได้ถ้าเราทราบค่าของตัวแปรอิสระต่าง ๆ ที่เกี่ยวข้อง

ถ้าให้ Y เป็นตัวแปรตาม และ $X_1, X_2, \ldots, X_k$ เป็นตัวแปรอิสระ แล้วความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระสามารถเขียนได้เป็น

$$Y = f(X_1, X_2, \ldots, X_k) + e$$

ซึ่งจะให้เป็นความสัมพันธ์แบบสุ่ม (Random Relationship) นั่นคือแต่ละค่าของตัวแปรอิสระจะมีค่าของตัวแปรตามหลาย ๆ ค่า หรือ แต่ละค่าของ X_i ต่าง ๆ จะมีการแจกแจงน่าจะเป็นของตัวแปรตาม Y เกิดขึ้น

ความสัมพันธ์ของตัวแปรต่าง ๆ ดังข้างบนนี้จะอยู่ในรูปใดรูปหนึ่งของสมการทางคณิตศาสตร์ ซึ่งเราสามารถแบ่งเป็น 2 รูปฟอร์มใหญ่ ๆ ดังนี้

(1) รูปฟอร์มเชิงเส้น (Linear Form) ซึ่งเป็น รูปฟอร์มที่มีตัวปรกตาม Y ผันแปรเป็นปฏิภาค (Proportional) กับตัวแปรอิสระ $X_1, X^2 \ldots X^h$ ดังสมการต่อไปนี้

$$Y = \beta o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

นั่นก็คือค่าของเฉลี่ยของตัวแปรตาม Y เมื่อกำหนด X_i ต่าง ๆ E(Y/X) จะมีความสัมพันธ์เชิงเส้นกับตัวแปรอิสระ X_i ต่าง ๆ ดังนี้

$$E(Y/X) = \beta o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

(2) รูปฟอร์มที่ไม่เป็นเชิงเส้น (Nonlinear Form) ซึ่งมีรูปแบบของสมการต่างจากรูปฟอร์มแรกนั่นเอง เช่น แบบพหุนาม (polynomial) แบบกำลังสอง (quadratic) แบบยกกำลัง (Exponential) เป็นต้น

สำหรับตัวแบบถดถอยที่ประกอบด้วยสมการทางคณิตศาสตร์เป็นแบบเชิงเส้น จะเรียกว่าตัวแบบถดถอยเชิงเส้น (Linear Regression Model) และถ้ามีตัวแปรอิสระเพียงตัวเดียวเท่านั้น นั่นคือมีสมการเป็น

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

จะเรียกว่าตัวแบบถดถอยเชิงเส้นชนิดธรรมดา (Simple Linear Regression Model) แต่ถ้ามีตัวแปรอิสระมากกว่าหนึ่งตัวดังสมการ

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

ซึ่งเป็นรูปฟอร์มแรกนั้น เราจะเรียกตัวแบบที่ประกอบด้วยสมการในรูปฟอร์มเชิงเส้นของหลายตัวแปรอิสระว่า ตัวแบบถดถอยเชิงเส้นชนิดพหุคูณ (Multiple Linear Regression Model)

9.1 ตัวแบบถดถอยเชิงเส้นชนิดธรรมดา (Simple Linear Regression Model)

ตัวแบบถดถอยเชิงเส้นชนิดธรรมดาเป็นตัวแบบถดถอยที่ง่ายที่สุดประกอบด้วยความสัมพันธ์ระหว่างตัวแปรสองตัว (X, Y) โดยมี Y เป็นตัวแปรตาม และ X เป็นตัวแปรอิสระ ดังสมการ

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

ซึ่งหมายความว่าค่าเฉลี่ยของ Y เมื่อกำหนด X หรือ E(Y/X) มีความสัมพันธ์กับ X ในแบบเส้นตรง นั่นคือ

$$E(Y/X) = \beta_0 + \beta_1 X$$

ดังนั้นสมการของตัวแบบจึงเขียนใหม่ได้เป็น

$$Y = E(Y/X) + \varepsilon$$

ในเมื่อ Y, X เป็นตัวแปรตามและอิสระที่สามารถวัดหรือสังเกตได้

β_0, β_1 เป็นพารามิเตอร์ถดถอย (Regression Parameters) ที่ไม่ทราบค่า

ε เป็นความคลาดเคลื่อนเชิงสุ่ม (Random error or disturbance) ที่ไม่สามารถวัดหรือสังเกตได้ และ ε นี้เป็นตัวแปรเชิงสุ่มที่เรายังไม่ทราบความจริงเกี่ยวกับขบวนการที่ทำให้เกิดเทอมนี้ การที่เราใส่ ε เข้าไปในสมการนั้นมีเหตุผลที่ต้องพิจารณา 2 ประการ คือ

(1) ตัวแปรที่ถูกทิ้งไป (Omitted Variables) ในสมการนั้นเราจะเห็นว่าค่าของ Y เท่านั้น ความจริงแล้ว Y อาจจะขึ้นอยู่กับตัวแปรอื่น ๆ ที่ไม่ปรากฏในสมการ ซึ่งตัวแปรเหล่านั้น

อาจจะเป็นประเภทที่วัดค่าได้หรือวัดค่าไม่ได้ ดังนั้นเราจึงใช้ ε แทนตัวแปรที่ไม่ปรากฏในสมการ

(2) ความสัมพันธ์แท้จริง (True Relationship) ความจริงความสัมพันธ์ระหว่าง X กับ Y จริง ๆ นั้นเราไม่ทราบว่าเป็นแบบใด แต่เราให้เป็นแบบเชิงเส้น ดังนั้นจึงมีความคลาดเคลื่อนเกิดขึ้น นั่นคือเน้นความคลาดเคลื่อนระหว่างสมการที่ใช้กับสมการที่แท้จริง ซึ่งเราจะแทนด้วย ε

สมการ $y = \beta_0 + \beta_1 x + \varepsilon$ นี้จะหมายความว่าแต่ละค่า X จะมีประชากรย่อย (Subpopulation) ของ Y ด้วย

x	Y
x_1	$y_{11}, y_{12}, y_{13}, \ldots\ldots$
x_2	$y_{21}, y_{22}, y_{23}, \ldots\ldots$
$\vdots$	
x_k	$y_{k_1}, y_{k_2}, y_{k_3}, \ldots\ldots$

หรือแต่ละค่าของ X จะมีการแจกแจงน่าจะเป็นของ Y ซึ่งจะเป็นแบบปกติ (Normal) หรือแบบอื่น ๆ ก็ได้ ดังนั้นค่าของ Y จึงไม่สามารถพยากรณ์ได้ถูกต้อง

ตัวแบบถดถอยนั้นมีข้อกำหนด (Specification) ไม่เพียงแต่สมการถดถอยที่กล่าวมาเท่านั้น ยังรวมถึงข้อสมมติ (Assumptions) อื่น ๆ เกี่ยวกับตัวคลาดเคลื่อน ε และตัวแปรอิสระ X ด้วย ดังนี้

(1) สำหรับแต่ละค่าของ X ตัวคลาดเคลื่อน ε จะมีการแจกแจงปกติ (Normality) ที่มีค่าจาก $-\infty$ ถึง $+\infty$

(2) และมีค่าเฉลี่ยเป็นศูนย์ (Zero means) นั่นคือ $E(\varepsilon) = 0$

(3) ความแปรปรวนของ ε จะเท่ากันหมด หรือมีความเป็นเอกภาพ (Homoskedasticity) นั่นคือ $V(\varepsilon) = \sigma^2$

ดังนั้น $\varepsilon \sim N(0, \sigma^2)$

(4) ตัวคลาดเคลื่อน ε_i และ $\varepsilon_j \; (i \neq j)$ จะไม่มีความสัมพันธ์กัน (Nonautocorrelation) หรือเป็นอิสระกันนั่นเอง นั่นคือ

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad (i, j = 1, 2, \ldots, n \,;\, i \neq j)$$

(5) ตัวแปรอิสระ X เป็นตัวแปรที่กำหนดได้ นั่นคือสามารถควบคุมหรือทำนายได้ โดยที่ค่าของมันอาจจะเลือกมาโดยแบบสุ่ม หรือ เลือกอย่างมีจุดหมาย และสำหรับตัวอย่าง

ขนาดหนึ่งนั้น ความแปรปรวนของตัวแปร X จะต้องเป็นจำนวนนับได้ (finite number) ที่ต่างจากศูนย์ หรือค่าของ X ในตัวอย่างหนึ่งจะไม่เท่ากันหมด และค่านั้นจะไม่เพิ่มขึ้นหรือลดลงโดยปราศจากขีดจำกัดในเมื่อขนาดตัวอย่างเพิ่มขึ้น

ข้อสมมติทั้ง 5 นี้มีส่วนสำคัญเกี่ยวกับการประมาณค่า เพราะตัวประมาณค่าบางอย่างขึ้นอยู่กับข้อสมมติเหล่านี้ เช่น ตัวประมาณค่าแบบกำลังสองต่ำสุด (Least Square Estimator)

ตัวแบบถดถอยเชิงเส้นที่ประกอบด้วยข้อสมมติ 5 ข้อ เหล่านี้มักจะเรียกว่าตัวแบบถดถอยเชิงเส้นแบบปกติ (Classical Normal Linear Regression Model, CNLRM)

ข้อสมมติใน CNLRM นี้ใช้สำหรับหาตัวประมาณค่าของพารามิเตอร์ถดถอย โดยที่ ε มีการแจกแจงแบบปกติที่มีค่าเฉลี่ย 0 และความแปรปรวน σ^2 หรือ $\varepsilon \sim N(0, \sigma^2)$ ตัวแบบถดถอยที่บรรยายด้วยสมการ $y = \beta_0 + \beta_1 x + \varepsilon$ กับข้อสมมติ 5 ข้อนี้จึงมีพารามิเตอร์ที่ไม่ทราบค่า 3 เทอม คือ β_0, β_1, และ σ^2

นอกจากการศึกษาถึงข้อกำหนดของตัวแบบถดถอยที่บรรยายด้วยสมการถดถอยและข้อสมมติทั้ง 5 นั้นแล้ว เรายังสนใจลักษณะบางอย่างของตัวแบบอีก โดยเฉพาะการแจกแจงน่าจะเป็นของตัวแปรตาม Y ดังนี้

(1) ค่าเฉลี่ย Y จะหาได้จากสมการถดถอย

$$
\begin{aligned}
E(y/x) &= E(\beta_0 + \beta_1 x + \varepsilon) \\
&= \beta_0 + \beta_1 x
\end{aligned}
$$

(β_0, β_1 เป็นพารามิเตอร์, X เป็นตัวคงที่, และ $E(\varepsilon) = 0$)

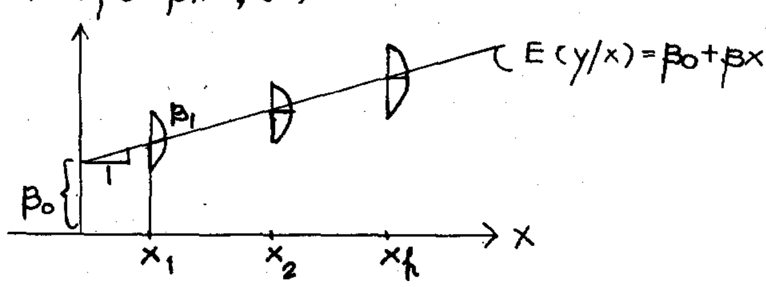
ดังนั้นสมการถดถอยจึงเขียนใหม่ได้เป็น

$$y = E(y/x) + \varepsilon$$

(2) ความแปรปรวนของ Y จะเป็น

$$V(y/x) = V(\varepsilon) = \alpha^2 .$$

จากสมการถดถอยเราจะเห็นว่า Y เป็นฟังก์ชั่นเชิงเส้นของ ε โดยที่ ε แจกแจงแบบปกติ เราจึงได้ว่าY ก็มีการแจกแจงแบบปกติที่มีค่าเฉลี่ย ($\beta_0 + \beta_1 x$) และความแปรปรวน α^2 นั่นคือ $y \sim N(\beta_0 + \beta_1 x, \alpha^2)$ ซึ่งแสดงได้โดยกราฟ ดังนี้



เราจะเห็นได้ว่าค่าเฉลี่ยของการแจกแจงทั้งหมด หรือ E(Y/X) จะอยู่บนเส้นตรงเดียวกัน และแต่ละการแจกแจงมีความแปรปรวนเท่ากันหมด

สมการ E(Y/X) = $\beta_0 + \beta_1 X$ ซึ่งแสดงถึงความสัมพันธ์ระหว่างค่าเฉลี่ยของ Y กับค่าของ X นั้นจะเรียกว่าเส้นถดถอยประชากร (Population Regression Line)

β_0 เป็นจุดตัดแกน (Intercept) ซึ่งจะวัดค่าเฉลี่ยของ Y ที่สมนัยกับ X ที่เป็นศูนย์

β_1 เป็นความชันของเส้นถดถอยซึ่งจะวัดการเปลี่ยนแปลงในค่าเฉลี่ยของ Y ที่สมนัยกับการเปลี่ยนแปลง**หนึ่ง**หน่วยในค่าของ X นั่นคือ $\beta_1 = \dfrac{\Delta Y}{\Delta X}$

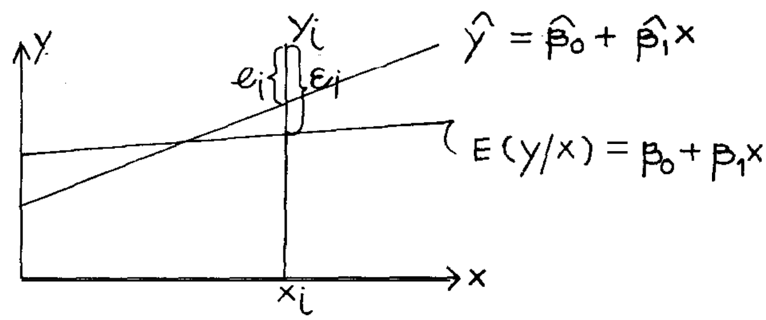
สำหรับเส้นถดถอยประชากรนั้นเรามักไม่ทราบค่าคงที่หรือพารามิเตอร์ β_0, β_1 จึงต้องประมาณค่าโดยอาศัยข้อมูลจากตัวอย่าง แล้วเราจะได้เส้นถดถอยตัวอย่าง (Sample Regression Line) ซึ่งทำหน้าที่เป็นตัวประมาณค่าของเส้นถดถอยประชากร ถ้า $\hat{\beta}_0$ และ $\hat{\beta}_1$ เป็นตัวประมาณค่าของ β_0 และ β_1 แล้วเส้นถดถอยตัวอย่างจะเป็น

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

หรือ
$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + e$$

ในเมื่อ Y เป็นค่าทำนายหรือค่าที่เฉลี่ยได้ (Predicted or fitted value) ของ Y ค่าสังเกตของ Y ส่วนมากจะไม่อยู่บนเส้นถดถอยตัวอย่างทีเดียว ดังนั้นค่าของ Y และ Y จึงต่างกัน ความแตกต่างนี้เรียกว่าเศษเหลือ (Residual) ซึ่งจะแทนด้วย

โดยทั่วไป e จะต่างจาก ε เพราะว่า $\hat{\beta}_0$ และ $\hat{\beta}_1$ ต่างจากค่าจริงของ β_0 และ β_1 เราสามารถแสดงให้เห็นได้ว่า e เป็นค่าประมาณของ ε ดังนี้



9.2 การประมาณค่าของพารามิเตอร์ถดถอย

การประมาณค่าของพารามิเตอร์ในตัวแบบถดถอยเชิงเส้นก็เหมือนกับการประมาณค่าพารามิเตอร์ในการแจกแจงน่าจะเป็นของตัวแปรตาม Y จากข้อสมมติของตัวแบบเราได้ว่า

$$Y \sim N(\beta_0 + \beta_1 X, \alpha^2)$$

ดังนั้นการประมาณค่าของพารามิเตอร์ถดถอย β_0, β_1 จึงเหมือนกับการประมาณค่าเฉลี่ยของ Y หรือ E(Y/X) ซึ่งเป็นเส้นถดถอยประชากรนั่นเอง การประมาณค่า β_0, β_1 เราทำได้หลายวิธีคือ

(1) วิธีกำลังสองต่ำสุด (Least Squares Method)

(2) วิธีน่าจะเป็นสูงสุด (Maximum Likelihood Method)

(3) วิธีกึ่งเฉลี่ย (SemiAverage Method)

(4) วิธีเฉลี่ยเคลื่อนที่ (Moving Average Method)

เราจะขอกล่าวเฉพาะวิธีกำลังสองต่ำสุด ซึ่งเป็นวิธีที่ดีและใช้กันมาก

9.3 วิธีกำลังสองต่ำสุด (Least Squares Method)

วิธีประมาณค่าแบบนี้มีหลักการว่า "ต้องหาเส้นถดถอยที่ทำให้ผลรวมของผลต่างกำลังสองระหว่างค่าทำนาย (Predicted Value) กับค่าจริง (Actual Value) นั้นมีค่าน้อยที่สุด" นั่นคือ เราต้องหาเส้นถดถอยที่ทำให้ $\Sigma(y-\hat{y})^2$ มีค่าน้อยที่สุด

สำหรับ $\Sigma(y-\hat{y})^2$ นั้น ได้ชื่อว่าผลรวมของเศษเหลือกำลังสอง (Residual Sum Squares) หรือผลรวมของความคลาดเคลื่อนจากเส้นถดถอยกำลังสอง (Sum Squares of Error about Regression Line) และใช้เป็นมาตราวัดการปรับที่ดี (Goodness-of-Fit Measure)

เนื่องจากเราทำให้ผลรวมของระยะทางกำลังสองนั้นน้อยที่สุด วิธีการนี้จึงได้ชื่อว่า วิธีกำลังสองน้อยสุด (Least Squares Method) ภายใต้ข้อสมมติที่กล่าวมาแล้วนั้น ทฤษฎีเกาส์-มาร์คอฟ (Gauss-Markoff Theorem) จะบอกให้เราทราบว่าเส้นถดถอยที่ได้จากวิธีกำลังสองต่ำสุดจะให้ค่าประมาณของพารามิเตอร์ในเส้นถดถอยประชากร (β_0, β_1) ที่ไม่เอียงเฉและเป็นแบบเชิงเส้นที่ดีที่สุด (Best Linear Unbiased Estimate, BLUE)

เราลองพิจารณา $S_0 = \Sigma(y-\hat{y})^2$ ซึ่งจะเห็นได้ว่า

$$S_0 = \Sigma(y-\hat{y})^2$$
$$S_1 = \Sigma[y-(\hat{\beta}_0+\hat{\beta}_1 x)]^2$$

นั้นขึ้นอยู่กับ β_0, β_1 ดังนั้นในการที่จะทำให้ผลต่างกำลังสอง S_0 น้อยที่สุดก็ต้องหาค่าที่เหมาะ ๆ ของ $\hat{\beta}_0, \hat{\beta}_1$ โดยอาศัยวิธีการทางแคลคูลัส (Calculus) คือหาอนุพันธ์บางส่วน (Partial derivative) ของ So เทียบกับ $\hat{\beta}_0, \hat{\beta}_1$ แล้วให้เท่ากับศูนย์ นั่นคือ

$$\frac{\partial S_0}{\partial \hat{\beta}_0} = 0, \quad \frac{\partial S_0}{\partial \hat{\beta}_1} = 0$$

แล้วเราจะได้สมการปกติกำลังสองต่ำสุด (Lest Squares Normal Equations) สองสมการดังนี้

$$\sum y = \hat{\beta}_0 n + \hat{\beta}_1 \sum x$$

$$\sum xy = \hat{\beta}_0 \sum x + \hat{\beta}_1 \sum x^2$$

จากสมการปกตินี้เราได้ค่า $\hat{\beta}_0, \hat{\beta}_1$ ที่ต้องการเป็น

$$\hat{\beta}_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$= \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$= \sum y/n - \hat{\beta}_1 (\sum x/n)$$

จาก $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ นี้เราจะได้ $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$ ซึ่งหมายความว่าเส้นถดถอยตัวอย่างผ่านจุด $(\bar{X}, \bar{Y})$ และ $\hat{\beta}_0$, $\hat{\beta}_1$ จะวัดจุดตัดแกนและความชันของเส้นถดถอยตัวอย่างตามลำดับ

ตัวอย่าง ในการศึกษาถึงความสัมพันธ์ระหว่างขนาดของครอบครัว (family Size) และค่าครองชีพ (Cost of living) โดยใช้ตัวอย่างของ 5 ครอบครัวที่สุ่มมา จากการสอบถามจำนวนบุตรในครอบครัว (X) และรายจ่ายที่จำเป็นแต่ละเดือน (Y) ได้ข้อมูล (สมมติ) มาดังนี้

ครอบครัว	1	2	3	4	5	ผลรวม
ขนาดครอบครัว (X)	2	3	5	7	8	25
ค่าครอบชีพ (Y)	2	2	3	5	8	20
XY	4	6	15	35	64	124
X^L	4	9	25	49	64	151
Y^2	4	4	9	25	64	160

ให้ความสัมพันธ์จริงระหว่างขนาดครอบครัวและค่าเฉลี่ยของค่าครองชีพเป็น

$$E(y/x) = \beta_0 + \beta_1 x \text{ หรือ } y = \beta_0 + \beta_1 x + \varepsilon$$

ซึ่งสามารถประมาณด้วยตัวอย่างได้เป็น

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \qquad \text{หรือ} \quad y = \hat{\beta}_0 + \beta_1 x + e$$

สำหรับ $\hat{\beta}_0, \hat{\beta}_1$ ประมาณด้วยวิธีกำลังสองต่ำสุดได้เป็น

$$\hat{\beta}_1 = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{5(124) - 25(20)}{5(151) - (25)^2} = \frac{620 - 500}{755 - 625}$$

$$= 120/130 = 12/13$$

และ $$\hat{\beta}_0 = \Sigma y/n - \hat{\beta}_1(\Sigma x/n)$$

$$= 20/5 - (12/13)(25/5)$$

$$= 4 - 60/13 = -8/13$$

ดังนั้นเราได้สมการถดถอยตัวอย่างเป็น

$$\hat{y} = -8/13 + (12/13)x$$

ถ้าครอบครัวหนึ่งมีบุตร 5 คน จะมีค่าครองชีพเป็น

$$\hat{y} = -8/13 + (12/13)(5)$$

$$= -8/13 + 60/13 = 4$$

และถ้ามีบุตร 8 คน ก็จะมีค่าครองชีพเป็น

$$\hat{y} = -8/13 + (12/13)(8)$$

$$= 88/13 = 6.77$$

9.4 คุณสมบัติของตัวประมาณค่าแบบกำลังสองต่ำสุด

ภายใต้ข้อสมมติของ CNLRM เราจะพบคุณสมบัติของตัวประมาณค่าแบบกำลังสองต่ำสุด ดังนี้

(1) ตัวประมาณค่า $\hat{\beta}_0$, $\hat{\beta}_1$ จะเป็นฟังก์ชั่นเชิงเส้นของค่าสังเกตตัวอย่างของ Y นั่นคือ

$$\hat{\beta}_1 = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(X-\bar{X})^2} = \Sigma \frac{(X-\bar{X})}{\Sigma(X-\bar{X})^2} Y$$

$$= \Sigma kY \; ; \quad k = \frac{X-\bar{X}}{\Sigma(X-\bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} = \Sigma\left(\frac{1}{n} - k\bar{X}\right)Y$$

$$= \Sigma KY \; ; \quad K = \frac{1}{n} - k\bar{X}$$

(2) ตัวประมาณค่า $\hat{\beta}_0$, $\hat{\beta}_1$ จะไม่มีความเอียงเฉของพารามิเตอร์ β_0 , β_1 นั่นคือ

$$E(\hat{\beta}_0) = \beta_0 ; \quad E(\hat{\beta}_1) = \beta_1$$

(3) ความแปรปรวนของ $\hat{\beta}_0$, $\hat{\beta}_1$ และความแปรปรวนร่วมของ $\hat{\beta}_0$ กับ $\hat{\beta}_1$ จะเป็น

$$\sigma^2_{\hat{\beta}_0} = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X-\bar{X})^2}\right]$$

$$\sigma^2_{\hat{\beta}_1} = \sigma^2 / \Sigma(X-\bar{X})^2$$

และ

$$Co(\hat{\beta}_0, \hat{\beta}_1) = -\bar{X}\left[\sigma^2 / \Sigma(X-\bar{X})^2\right]$$

ตัวประมาณค่าที่ไม่เอียงเฉของ σ^2 คือ S^2 ซึ่งกำหนดไว้ว่า

$$S^2 = \frac{1}{n-2} \Sigma(Y-\hat{Y})^2$$

$$= \frac{1}{n-2}\left[\Sigma Y^2 - \hat{\beta}_0 \Sigma Y - \hat{\beta}_1 \Sigma XY\right]$$

(4) ตัวประมาณค่าจะมีประสิทธิภาพซึ่งเป็นไปตามทฤษฎีเกาส์-มาร์คอฟที่ว่า "ตัวประมาณแบบกำลังสองต่ำสุดจะเป็นตัวประมาณค่าที่ดีที่สุดชนิดไม่เอียงเฉและเป็นแบบเชิงเส้น (Best Linear Unbiased Estimales, BLUE)" นั่นคือในกลุ่มของตัวประมาณเชิงเส้นที่ไม่เอียงเฉด้วยกันนั้น ตัวประมาณแบบกำลังสองต่ำสุดจะมีความแปรปรวนน้อยที่สุด

(5) ตัวประมาณค่า $\hat{\beta}_0$, $\hat{\beta}_1$ ต่างก็มีการแจกแจงแบบปกติ นั่นคือ

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2_{\hat{\beta}_0})$$

และ

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$$

9.5 การอ้างอิงหรืออนุมานทางสถิติเกี่ยวกับการถดถอย

9.5.1 การประมาณช่วงเชื่อมั่นของพารามิเตอร์

เนื่องจากตัวประมาณค่า $\hat{\beta}_0$, $\hat{\beta}_1$ มีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเป็น β_0 , β_1 และความแปรปรวนเป็น $\sigma^2_{\hat{\beta}_0}$, $\sigma^2_{\hat{\beta}_1}$ ดังนั้น

$$\frac{\hat{\beta}_0 - \beta_0}{\sigma_{\hat{\beta}_0}} \quad \text{และ} \quad \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}}$$

จึงมีการแจกแจงแบบปกติมาตรฐาน $N(0,1)$

โดยที่ความแปรปรวนของ $\hat{\beta}_0$, $\hat{\beta}_1$ ยังขึ้นอยู่กับ σ^2 ที่ไม่ทราบค่า จึงต้องประมาณ σ^2 ด้วย S^2 ซึ่งเราจะได้ค่าประมาณของความแปรปรวนของ $\hat{\beta}_0$, $\hat{\beta}_1$ ดังนี้

$$S^2_{\hat{\beta}_0} = S^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X-\bar{X})^2}\right)$$

$$S^2_{\hat{\beta}_1} = S^2 / \Sigma(X-\bar{X})^2$$

สำหรับรากที่สองของ $S^2_{\hat{\beta}_0}$ และ $S^2_{\hat{\beta}_1}$ หรือ $S_{\hat{\beta}_0}$ และ $S_{\hat{\beta}_1}$ นี้เรียกว่า ความคลาดเคลื่อนมาตรฐาน (Standard Error) ของ $\hat{\beta}_0$ และ $\hat{\beta}_1$ ซึ่งจะใช้เป็นมาตราวัดความเที่ยงตรงของ $\hat{\beta}_0$ และ $\hat{\beta}_1$ ส่วน S ซึ่งเป็นรากที่สองของ S^2 จะเรียกว่าความคลาดเคลื่อนมาตรฐานของค่าทำนายหรือค่าประมาณ (Standard error of estimate) ซึ่งใช้เป็นมาตรวัดการปรับที่ดี (Measure of goodness-of-fit) ได้

จากค่าประมาณของความแปรปรวนของ $\hat{\beta}_0$ และ $\hat{\beta}_1$ เราจึงได้การแจกแจงของ

$$\frac{\hat{\beta}_0 - \beta_0}{S_{\hat{\beta}_0}} \text{ และ } \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}}$$

ซึ่งมีการแจกแจงแบบ Student t องศาความเป็นอิสระ n-2

ดังนั้นเราจึงสามารถสร้างช่วงเชื่อมั่น $100(1-\alpha)\%$ของพารามิเตอร์ β_0 , β_1 ได้โดยอาศัยการแจกแจงของ $\frac{\hat{\beta}_0 - \beta_0}{S_{\hat{\beta}_0}}$ และ $\frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}}$

นั่นคือเราจะได้ช่วงเชื่อมั่น $100(1-\alpha)\%$สำหรับ β_0, β_1 เป็น

$$\beta_0 = \hat{\beta}_0 \pm t^{(n-2)}_{\alpha/2} S_{\hat{\beta}_0}$$

และ

$$\beta_1 = \hat{\beta}_1 \pm t^{(n-2)}_{\alpha/2} S_{\hat{\beta}_1}$$

ส่วนช่วงเชื่อมั่น $100(1-\alpha)\%$สำหรับ σ^2 นั้นจะอาศัยการแจกแจงของฟังก์ชั่นของ S^2 หรือการแจกแจงของซึ่งมีการแจกแจงแบบไคสแควร์, องศาความเป็นอิสระ n-2 ดังนั้นเราได้ช่วงเชื่อมั่น $100(1-\alpha)\%$สำหรับ σ^2 จะเป็น

$$\sigma^2 = (n-2)S^2/\chi^{2(n-2)}_{\alpha/2} \quad , (n-2)S^2/\chi^{2(n-2)}_{1-\alpha/2}$$

ตัวอย่าง จากตัวอย่างเกี่ยวกับขนาดครอบครัวและค่าครองชีพ เราประมาณความคลาดเคลื่อนของค่าทำนาย S ได้เป็น

$$S^2 = \frac{1}{n-2}\Sigma(Y-\hat{Y})^2$$

$$= \frac{1}{5-2}(3.85)$$

$$= 1.28$$

โดยที่ $\Sigma(Y-\hat{Y})^2$หาได้จาก

X	Y	$\hat{Y} = -8/13 + (12/13)X$	$(Y-\hat{Y})^2$
2	2	$-8/13 + (12/13)(2) = 1\frac{3}{13}$	$100/169$
3	2	$-8/13 + (12/13)(3) = 2\frac{2}{13}$	$4/169$
5	3	$-8/13 + (12/13)(5) = 4$	1
7	5	$-8/13 + (12/13)(7) = 5\frac{11}{13}$	$121/169$
8	8	$-8/13 + (12/13)(8) = 6\frac{10}{13}$	$1\frac{67}{169}$
			$3\frac{143}{169}$
			$= 3.85$

หรือจะหา S^2 ได้จาก

$$S^2 = \frac{1}{n-2}\left[\Sigma Y^2 - \hat{\beta}_0\Sigma Y - \hat{\beta}_1\Sigma XY\right]$$

$$= \frac{1}{5-2}\left[106 - (-8/13)(20) - (12/13)(124)\right]$$

$$= (1/3)(3.85)$$

$$= 1.28$$

ดังนั้น $S = \sqrt{1.28} = 1.13$

เพราะฉะนั้น

$$S^2_{\hat{\beta}_0} = S^2\left(1/n + \frac{\bar{X}^2}{\Sigma(X-\bar{X})^2}\right)$$

$$= 1.28\left(1/5 + (25/5)^2/26\right)$$

$$= 1.28(1.16)$$

$$= 1.4848$$

$$S_{\hat{\beta}_0} = \sqrt{1.4848}$$

$$= 1.22$$

ในเมื่อ $\Sigma(X-\bar{X})^2 = \Sigma X^2 - (\Sigma X)^2/n = 151 - (25)^2/5 = 26$

$$S^2_{\hat{\beta}_1} = S^2/\Sigma(X-\bar{X})^2$$

$$= 1.28/26 = 0.49$$

$$S_{\hat{\beta}_1} = \sqrt{0.49} = 0.22$$

ช่วงเชื่อมั่น 95% สำหรับ β_0, β_1, และ σ^2 จะเป็น

$$\beta_0 = \hat{\beta}_0 \pm t^{(5-2)}_{.05/2}\, S_{\hat{\beta}_0}$$

$$= -8/13 \pm (3.182)(1.22)$$

$$= -0.62 \pm 3.88$$

$$= -4.50\ ,\ 3.26$$

$$\beta_1 = \hat{\beta}_1 \pm t^{(3)}_{.025}\, S_{\hat{\beta}_1}$$

$$= 12/13 \pm (3.182)(0.22)$$

$= 0.92 \pm 0.70$

$= 0.22 \quad , \quad 1.62$

และ $\sigma^2 = (n-2)S^2/\chi^{2(5-2)}_{.05/2} \quad , \quad (n-2)S^2/\chi^{2(5-2)}_{1-.05/2}$

$= (5-2)(1.28)/9.348 \quad , \quad (5-2)(1.28)/.216$

$= 3.85/9.348 \quad , \quad 3.85/.216$

$= 0.412 \quad , \quad 17.824$

9.5.2 การทดสอบสมมติฐานเกี่ยวกับการถดถอย

(1) ทดสอบเกี่ยวกับพารามิเตอร์ $\beta_0, \beta_1, \sigma^2$ ที่ระบุค่าไว้ ถ้าเราต้องการทดสอบสมมติฐานเกี่ยวกับค่าของ $\beta_0, \beta_1, \sigma^2$ ที่ระบุไว้ว่าเป็น $\beta_{00}, \beta_{10}, \sigma_0^2$ โดยอาศัยตัวอย่างขนาด n เราก็ทำได้ดังนี้

(ก) สมมติฐานเกี่ยวกับ β_0

สมมติฐาน $H_0 : \beta_0 = \beta_{00}$

และ Ha จะเป็นอย่างใดอย่างหนึ่งใน 3 แบบ ดังนี้

$H_a : \beta_0 < \beta_{00} \quad , \quad H_a : \beta_0 > \beta_{00}$, หรือ $H_a : \beta_0 \neq \beta_{00}$

ตัวสถิติทดสอบ $T = (\hat{\beta}_0 - \beta_0)/S_{\hat{\beta}_0}$

ซึ่งมีการแจกแจงแบบ Student t, องศาความเป็นอิสระ n-2 ถ้า $H_a : \beta_0 = \beta_{00}$ เป็นจริง

(ข) สมมติฐานเกี่ยวกับ β_1

สมมติฐาน $H_0 : \beta_1 = \beta_{10}$

และ Ha จะเป็นอย่างใดอย่างหนึ่งใน $H_a : \beta_1 < \beta_{10}, H_a : \beta_1 > \beta_{10}$ หรือ $H_a : \beta_1 \neq \beta_{10}$

ตัวสถิติทดสอบ $T = (\hat{\beta}_1 - \beta_{10})/S_{\hat{\beta}_1}$

ซึ่งมีการแจกแจงแบบ Student t, องศาความเป็นอิสระ n-2 ถ้า $H_0 : \beta_1 = \beta_{10}$ เป็นจริง

(ค) สมมติฐานเกี่ยวกับ σ^2

สมมติฐาน $H_0 : \sigma^2 = \sigma_0^2$

และ Ha จะเป็นอย่างใดอย่างหนึ่งใน $H_a : \sigma^2 < \sigma_0^2$, $H_a : \sigma^2 > \sigma_0^2$, หรือ $H_a : \sigma^2 \neq \sigma_0^2$

ตัวสถิติทดสอบ $\chi^2 = (n-2)S^2/\sigma^2$

ซึ่งมีการแจกแจงไคสแควร์, องศาความเป็นอิสระ n-2 ถ้า $H_0 : \sigma^2 = \sigma_0^2$ เป็นจริง

ตัวอย่าง จากตัวอย่างที่แล้วมา ถ้าเรามีสมมติฐานเกี่ยวกับ β_0, β_1 และ σ^2 เป็นดังนี้

(1) β_0 มีค่าน้อยกว่า 0

(2) β_1 มีค่าเท่ากับ 1

(3) σ มีค่ามากกว่า 1

เราทำการทดสอบโดยทดได้อย่างที่สุ่มมาแล้วนั้นได้ดังนี้

(1) $H_0 : \beta_0 = 0$, $H_a : \beta_0 < 0$

$$T = \frac{\hat{\beta}_0 - \beta_{00}}{S_{\hat{\beta}_0}} = \frac{-8/13 - 0}{1.22} = -0.51$$

ค่าวิกฤติ ณ $\alpha = .05$ หรือ $t^{(5-2)}_{.05} = 2.353$ จึงยอมรับ $H_0 : \beta_0 = 0$

นั่นคือปฏิเสธคำกล่าวที่ว่า β_0 มีค่าน้อยกว่า 0

(2) $H_0 : \beta_1 = 1$, $H_a : \beta_1 \neq 1$

$$T = \frac{\hat{\beta}_1 - \beta_{10}}{S_{\hat{\beta}_1}} = \frac{12/13 - 1}{0.22}$$

$$= -0.08/0.22 = -0.37$$

ค่าวิกฤต ณ $\alpha = .05$ หรือ $t^{(5-2)}_{.05/2} = 3.182$ จึงยอมรับ $H_0 : \beta_1 = 1$

3) $H_0 : \sigma^2 = 1$; $H_a : \sigma^2 > 1$

$$X^2 = \frac{(n-2)S^2}{\sigma_0^2} = \frac{(5-2)(1.28)}{1}$$

$$= 3.85$$

ค่าวิกฤต ณ $\alpha = .05$ หรือ $X^{2(5-2)}_{.05} = 7.815$ จึงยอมรับ $H_0 : \sigma^2 = 1$

(2) การทดสอบความสัมพันธ์ระหว่างตัวแปร จากสมการถดถอยประชากร

$$E(Y/X) = \beta_0 + \beta_1 X$$

นี้เราจะเห็นได้ว่าค่าเฉลี่ยของ Y กับ X จะมีความสัมพันธ์กันหรือไม่นั้นขึ้นอยู่กับค่าของพารามิเตอร์ β_1 ถ้า β_1 เป็นศูนย์หรือใกล้กับศูนย์ แล้วไม่ว่า X จะเปลี่ยนแปลงไปอย่างไร ค่าเฉลี่ยของ Y ก็จะมีค่าใกล้ ๆ β_0 นั่นคือ Y ไม่ผันแปรตาม X หรือ ความสัมพันธ์ระหว่างตัวแปร X และ Y จึงกำหนดสมมติฐานที่จะทดสอบไว้ดังนี้

$$H_0 : \beta_1 = 0 \quad , \quad H_a : \beta_1 \neq 0$$

แต่ถ้าเรามีความรู้มาก่อนว่า β_1 จะเป็นบวกหรือลบ แล้ว Ha ก็ควรจะเป็นไปตามที่ทราบ

มาก่อนนั่นคือ $H_a : \beta_1 > 0$ หรือ $H_a : \beta_1 < 0$

ตัวสถิติทดสอบสำหรับสมมติฐานนี้ก็คือ

$$T = \hat{\beta}_1 / S_{\hat{\beta}_1}$$

ซึ่งมีการแจกแจงแบบ Student t, องศาความเป็นอิสระ n-2

ตัวอย่าง จงทดสอบความสัมพันธ์ระหว่างขนาดครอบครัวกับค่าครองชีพในตัวอย่างที่แล้วมาว่ามีจริงหรือไม่

สมมติฐาน $H_a : \beta_1 = 0$; $H_a : \beta_1 \neq 0$

ตัวสถิติทดสอบ $T = \hat{\beta}_1 / S_{\hat{\beta}_1}$

เกณฑ์ตัดสินใจ ปฏิเสธ Ho ณ $\alpha = .05$ ถ้า

$$T > t_{.05}^{(5-2)} = 3.182$$

คำนวณตัวสถิติทดสอบ $T = \frac{12/13}{0.22} = 4.18$

สรุปผล ค่าของ T มากกว่า 3.182 จึงปฏิเสธ Ho ซึ่งแสดงว่าขนาดครอบครัวกับค่าครองชีพมีความสัมพันธ์กันจริง

(3) การทำนายหรือการพยากรณ์ (Prediction or Forecasting)

สมการถดถอยที่ประมาณได้นั้น เราจะใช้สำหรับทำนายค่าของตัวแปรตาม Y โดยระบุค่าของตัวแปรอิสระ X ให้เป็น X_o วิธีการทำนายตัวแปรตามเราทำได้ 2 แบบ คือ

- การทำนายค่าเฉลี่ยของ Y เมื่อกำหนด Xo ให้, ECY/Xo
- การทำนายค่า Y ใด ๆ เมื่อกำหนด Xo ให้, Yo

ก. การทำนายค่าเฉลี่ยของ Y เมื่อกำหนด Xo ได้ ตัวประมาณค่าที่ดีของ $E(Y/X_o)$ ก็คือ $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ นั่นเอง ตัวประมาณค่า $\hat{Y}$ จะมีการแจกแจงแบบปกติที่มีค่าเฉลี่ย $\beta_0 + \beta_1 X$ และความแปรปรวน $\sigma^2_{\hat{Y}}$

$$\sigma^2_{\hat{Y}} = \sigma^2\left(1/n + \frac{(X_o - \bar{X})^2}{\Sigma(X-\bar{X})^2}\right)$$

ค่าประมาณที่ดีของ $\sigma^2_{\hat{Y}}$ คือ $S^2_{\hat{Y}}$

$$S^2_{\hat{Y}} = S^2\left(1/n + \frac{(X_o - \bar{X})^2}{\Sigma(X-\bar{X})^2}\right)$$

โดยอาศัยการแจกแจงของ $\left\{\hat{Y} - (\beta_0 + \beta_1 X)\right\} / S_{\hat{Y}}$

ซึ่งมีการแจกแจงแบบ Student t, องศาความเป็นอิสระ n-2 เราก็จะสามารถสร้างช่วงเชื่อมั่น

สำหรับค่าเฉลี่ยของ Y เมื่อกำหนด Xo ให้ $E(Y/X_0)$ ได้เป็น

$$E(Y/X_0) = \hat{Y} \pm t_{\alpha/2}^{(n-2)} S_{\hat{Y}}$$

$$= (\hat{\beta}_0 + \hat{\beta}_1 X) \pm t_{\alpha/2}^{(n-2)} \left\{ S\sqrt{1/n + \frac{(X_0-\bar{X})^2}{\Sigma(X-\bar{X})^2}} \right\}$$

ในบางครั้งเรามีสมมติฐานเกี่ยวกับค่าที่ระบุไว้ (μ_0) ของค่าเฉลี่ย Y ที่กำหนด X_0 ได้ นั่นคือ Ho เป็นดังนี้

$$H_0 : E(Y/X_0) = \mu_0$$

ตัวสถิติทดสอบจะเป็น

$$T = (\hat{Y} - \mu_0)/S_{\hat{Y}}$$

ซึ่งมีการแจกแจงแบบ Student t, องศาความเป็นอิสระ n-2

ข. การทำนายค่า Y ใด เมื่อกำหนด X_0 ให้ หรือ Y_0 เมื่อกำหนดค่าของตัวแปรอิสระ X เป็น X_0 แล้วเราจะทำนายค่าของ Y ซึ่งจะให้เป็น Y_0 โดยที่ Y_0 เป็นแปรเชิงสุ่มมีค่ากระจายรอบจุดบนเส้นถดถอย ประชากรซึ่งสอดคล้องกับ X_0 และเราไม่ทราบค่าของมันก่อนทำการทดลอง ถ้าเราทราบพารามิเตอร์ แล้วตัวทำนายของ Y_0 จะเป็นค่าเฉลี่ยของมัน หรือ $E(Y/X_0) = P_0 + P_1X_0$ ซึ่งเป็นจุดหนึ่งบนเส้นถดถอยประชากร ตัวทำนายนี้จะดีที่สุดในแง่ที่ว่าความแปรปรวนของ Yo สอบ $E(Y/X_0)$ จะน้อยกว่าความแปรปรวนรอบจุดอื่นใด ค่าของ Y_0 จะแจกแจงแบบปกติที่มีค่าเฉลี่ย $P_0 + P_1X_0$ และความแปรปรวน σ^2 ซึ่งเป็นความแปรปรวนอันสืบเนื่องมาจากตัวคลาดเคลื่อนที่ปรากฏอยู่ในสมการถดถอย

แต่ที่จริงแล้วเราไม่ทราบค่า $E(Y/X_0)$ เพราะไม่ทราบเส้นถดถอยประชากรจึงต้องประมาณค่าด้วยเส้นถดถอยตัวอย่าง จุดที่สอดคล้องกับเส้นถดถอยตัวอย่างหรือ $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$ จะเป็นตัวประมาณค่าที่ดีของ Y_0 ค่าจริงของ Y_0 กับค่าทำนายของ Y_0 (หรือ $\hat{Y}_0$) ส่วนมากจะต่างกัน นั่นคือเกิดความคลาดเคลื่อนขึ้น ทั้งนี้ก็เพราะ

- ตัวคลาดเคลื่อนเชิงสุ่ม E_0 ซึ่งจะทำให้ค่าของ Y_0 ไม่เท่ากับ $E(Y/X_0)$ นั่นคือ Y_0 จะไม่อยู่บนเส้นถดถอยประชากร ความคลาดเคลื่อนซึ่งเกิดจากตัวคลาดเคลื่อนนี้จะประจำอยู่กับกลไกต่าง ๆ ที่ทำให้เกิดค่าของตัวแปรตาม Y และไม่มีอะไรที่สามารถจะทำให้มันลดน้อยลงได้

- ความคลาดเคลื่อนตัวอย่าง (Sampling error) ซึ่งจะทำให้เส้นถดถอยตัวอย่างไม่เป็นเส้นเดียวกับเส้นถดถอยประชากรความคลาดเคลื่อนแบบนี้จะลดลงถ้าเราเพิ่มความเที่ยงตรงในการประมาณเส้นถดถอยประชากรด้วยการเพิ่มขนาดตัวอย่างให้มากขึ้น

ความแตกต่างระหว่างค่าจริงของ Y_0 กับค่าทำนาย $\hat{Y}_0$ นี้ จะเรียกว่าความคลาดเคลื่อนพยากรณ์ (Forecast error) ซึ่งจะมีความสัมพันธ์กับความคลาดเคลื่อนอื่นดังกล่าวมาแล้วดังนี้

$$Y_0 - \hat{Y}_0 = \{Y_0 - E(Y/X_0)\} + \{E(Y/X_0) - \hat{Y}_0\}$$

ความคลาดเคลื่อนพยากรณ์นี้จะเป็นตัวแปรเชิงสุ่มที่มีการแจกแจงแบบปกติ และมีค่าเฉลี่ยและความแปรปรวนดังนี้

$$E(Y_0 - \hat{Y}_0) = 0$$

$$\begin{aligned}\sigma_f^2 &= V(Y_0 - \hat{Y}_0)\\ &= \sigma^2 + \sigma_{\hat{Y}_0}^2\\ &= \sigma^2\left[1 + 1/n + \frac{(X_0 - \bar{X})^2}{\Sigma(X-\bar{X})^2}\right]\end{aligned}$$

เราจะเห็นได้ว่าความแปรปรวนของความคลาดเคลื่อนพยากรณ์จะน้อยลงได้ ถ้า

- ขนาดตัวอย่างโตขึ้น
- การกระจายของตัวแปรอิสระ X ในตัวอย่างมีมาก หรือ $\Sigma(X-\bar{X})^2$ โตมากนั่นเอง
- ระยะห่างระหว่าง X_0 กับค่าเฉลี่ยตัวอย่าง $\bar{X}$ น้อยลง

สองข้อแรกนี้จะแสดงถึงความจริงที่ว่า การพยากรณ์หรือค่าประมาณของเส้นถดถอยประชากรจะยิ่งดีขึ้น ถ้าความแปรปรวนของความคลาดเคลื่อนพยากรณ์มีค่าน้อยลง สำหรับข้อที่สามนั้นหมายความว่า การพยากรณ์จะดีขึ้น ถ้าค่าของ X (หรือ X_0) ใกล้ $\bar{X}$ มากขึ้นซึ่งตรงกับข้อยืนยันที่ว่าเราสามารถพยากรณ์ได้ดีภายในช่วงหรือนิสัยของประสบการณ์ที่เป็นค่าของตัวแปรอิสระ X ในตัวอย่างมากกว่าภายนอกช่วงประสบการณ์ จุดกลางของนิสัยคือ $\bar{X}$ ดังนั้น ความเชื่อมั่นของการพยากรณ์จะยิ่งน้อยลงถ้าเราพยากรณ์ด้วยจุดที่อยู่ห่างจากจุดกลางของนิสัย $\bar{X}$ มากเท่าขึ้น

โดยทั่วไปเราไม่ทราบ σ_f^2 จึงต้องประมาณด้วย S_f^2 ซึ่งเป็นตัวประมาณค่าที่ดี และ S_f^2 กำหนดไว้ดังนี้

$$S_f^2 = S^2\left(1 + 1/n + \frac{(X_0-\bar{X})^2}{\Sigma(X-\bar{X})^2}\right)$$

จากการแจกแจงของ

$$(\hat{Y} - \hat{Y}_0)/S_f$$

ซึ่งมีการแจกแจงแบบ Student t, องศาความเป็นอิสระ n-2 เราก็สามารถสร้างช่วงพยากรณ์ $100(1-\alpha)\%$ ของ Y_0 ได้เป็น

$$\begin{aligned}Y_0 &= \hat{Y}_0 \pm t_{\alpha/2}^{(n-2)} S_f\\ &= [\hat{\beta}_0 + \hat{\beta}_1 X] \pm t_{\alpha/2}^{(n-2)}\end{aligned}$$

ตัวอย่าง จากความสัมพันธ์ระหว่างขนาดครอบครัวและค่าครองชีพ จะใช้ทำนายค่าครองชีพจากขนาดครอบครัว (หรือจำนวนบุตร)

ถ้าครอบครัวที่มีบุตร 4 คน แล้วค่าครองชีพเฉลี่ยที่มีความเชื่อมั่น 95% จะเป็น

$$E(Y/X=4) = \hat{Y} \pm t^{(3)}_{.025} S_{\hat{Y}}$$
$$= [-8/13 + (12/13)(4)] \pm 3.182(0.55)$$
$$= 3.08 \pm 1.75$$
$$= 1.33\ ,\ 4.83$$

ในเมื่อ

$$S_{\hat{Y}} = S\sqrt{1/n + \frac{(X_0-\bar{X})^2}{\Sigma(X-\bar{X})^2}} = 1.13\sqrt{1/5 + (4-5)^2/26}$$
$$= 1.13\sqrt{0.24} = 1.13(0.49)$$
$$= 0.55$$

ถ้าครอบครัวหนึ่งมีบุตร 6 คน แล้วช่วงพยากรณ์ 95% ของค่าครองชีพสำหรับครอบครัวนี้จะเป็น

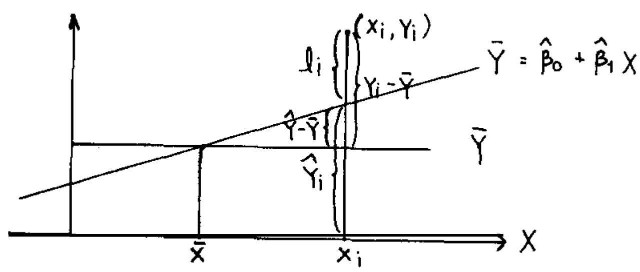
$$Y_0 = \hat{Y}_0 \pm t^{(n-2)}_{\alpha/2} S_f$$
$$= [-8/13 + (12/13)(6)] \pm 3.182(1.25)$$
$$= 4.92 \pm 3.98$$
$$= 0.94\ ,\ 8.90$$

ในเมื่อ Sf =

$$S_f = S\sqrt{1 + 1/n + (X_0-\bar{X})^2/\Sigma(X-\bar{X})^2}$$
$$= 1.13\sqrt{1 + 1/5 + (6-5)^2/26} = 1.13\sqrt{1.24}$$
$$= 1.13(1.11) = 1.25$$

9.6 การแยกความผันแปรของตัวแปรตาม

ความผันแปรของตัวแปรตามนั้นเราหมายถึงการเปลี่ยนแปลงในค่าสังเกต Y ของตัวอย่างจากค่าหนึ่ง ๆ ไปยังค่าอื่น ๆ เราแยกความผันแปรในตัวแปรตามก็เพื่อต้องการเพิ่มความรู้เกี่ยวกับค่าประมาณที่หามาได้ ลองพิจารณาความผันแปรของ Y ในแผนภาพกระจาย (Scatter diagram) ซึ่งเป็นกราฟของค่าสังเกต Y ที่สมนัยกับค่าของ X ดังต่อไปนี้



จากแผนภาพนี้เราจะเห็นได้ว่า ความผันแปรใน Y นี้ส่วนหนึ่งเนื่องมาจากการเปลี่ยนแปลงใน X (ซึ่งนำไปสู่การเปลี่ยนแปลงในค่าเฉลี่ย Y) และอีกส่วนหนึ่งเนื่องมาจากผลกระทบของตัวคลาดเคลื่อน ถ้าตัวแปรตาม Y ไม่มีความผันแปร แล้วค่าของมันทั้งหมด (ซึ่งจะเท่ากัน) จะอยู่บนเส้นนอน และทุกค่าจะเท่ากับค่าเฉลี่ยของมัน $\bar{Y}$ เส้นนอนจะสมนัย $\bar{Y}$ ตามความจริงแล้วค่าสังเกตของ Y จะกระจายรอบ ๆ เส้นนี้ ดังนั้นความผันแปรของ Y จึงวัดด้วยระยะทางของค่าสังเกต Y ที่เบี่ยงเบนจากค่าเฉลี่ย $\bar{Y}$ มาตราวัดที่ดีของระยะทางนี้เราจะใช้ผลรวมของค่ากำลังสองของมัน ตามปกติเราเรียกว่าผลรวมกำลังสองทั้งหมด (Total Sum Squared deviation, SST) นั่นคือ

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

ถ้าพิจารณาค่าสังเกตค่าหนึ่งของ Y หรือ Y_i ที่สมนัยกับค่าหนึ่งของ X หรือ X_i เราจะเห็นว่าระยะทางของ (X_i, Y_i) ที่ห่างจาก $\bar{Y}$ นี้แบ่งออกได้เป็น 2 ส่วน คือส่วนหนึ่งเป็นระยะทางที่สังเกตได้ถึงเส้นถดถอยตัวอย่าง และอีกส่วนหนึ่งเป็นระยะทางเส้นถดถอยถึงค่าเฉลี่ย $\bar{Y}$ นั่นคือ

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

สำหรับค่าสังเกตทุกค่าในตัวอย่างขนาด n เราจะได้ว่า

$$\Sigma_i(Y_i-\bar{Y})^2 = \Sigma_i(Y_i-\hat{Y}_i)^2 + \Sigma_i(\hat{Y}_i-\bar{Y})^2$$

$$SST = SSE + SSR$$

นั่นคือความผันแปรของตัวแปรตาม Y แบ่งออกได้เป็น 2 ส่วน คือส่วนแรก (SSR) บรรยายถึงความผันแปรของค่าที่ทำนายได้ของ Y หรือความผันแปรที่เนื่องจากผลกระทบของ X และส่วนที่สอง (SSE) บรรยายถึงความผันแปรของเศษเหลือของการถดถอย (Regression Residnals) หรือความผันแปรเนื่องจากผลกระทบของตัวคลาดเคลื่อน

สำหรับ SST และ SSR สามารถกระจายออกได้เป็น

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

$$SSR = \Sigma(\hat{Y}-\bar{Y})^2 = \hat{\beta}_1^2\Sigma(X-\bar{X})^2$$

$$= \hat{\beta}_1^2[\Sigma X^2 - (\Sigma X)^2/n]$$

9.7 สัมประสิทธิ์การตัดสินใจ (Coefficient of Determination, R^2)

การแยกความผันแปรของตัวแปรตาม Y นี้จะได้มาตราวัดอันหนึ่งที่ชื่อว่าสัมประสิทธิ์การตัดสินใจ , R^2 , ซึ่งใช้วัดการปรับที่ดี (Goodness-of-fit) ของเส้นถดถอยต่อค่าสังเกตจาก

ตัวอย่าง และใช้วัดเปอร์เซ็นต์ของความผันแปรในตัวแปรตาม Y ที่เนื่องมาจากตัวแปรอิสระ X สัมประสิทธิ์การตัดสินใจ R^2 กำหนดไว้ดังนี้

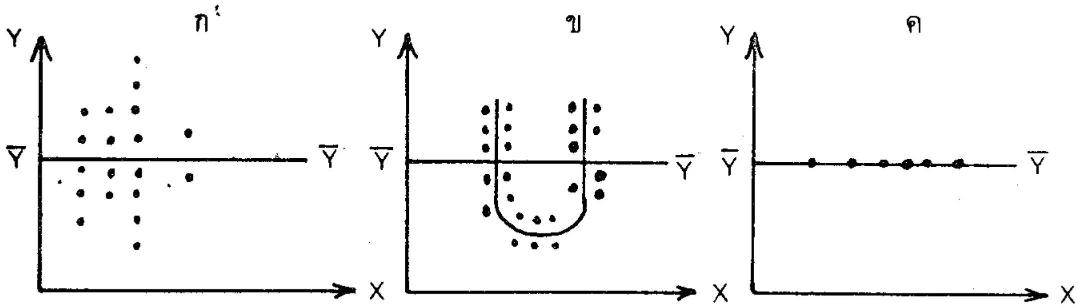
$$R^2 = \frac{SSR}{SST}$$

หรือ $$R^2 = 1 - \frac{SSE}{SST}$$

สัมประสิทธิ์การตัดสินใจ , R^2 จะเป็นมาตราวัดที่ใช้บรรยายว่า "เส้นถดถอยตัวอย่างปรับเข้าได้กับข้อมูลที่สังเกตได้ดีเพียงไร" และค่าของมันจะอยู่ระหว่าง 0 กับ 1 นั่นคือ

$$0 \leq R^2 \leq 1$$

ถ้า $R^2 = 0$ แสดงว่าการปรับเลวที่สุด และถ้า $R^2 = 1$ ก็แสดงว่าการปรับดีที่สุด สิ่งที่น่าสนใจกรณี $R^2 = 0$ ก็คือเส้นถดถอยอยู่ในแนวนอน หรือ $\beta_1 = 0$ และการที่เส้นถดถอยจะเป็นเส้นนอนได้นั้นมีเหตุผลหลายประการ ดังรูป



กรณี ก. ค่าสังเกตกระจายแบบสุ่มรอบ $\bar{Y}$

ข. ค่าสังเกตกระจายรอบเส้นโค้ง ทั้งที่เส้นถดถอย เป็นเส้นนอน กรณีนี้ X และ Y มีความสัมพันธ์แบบไม่เป็นเชิงเส้น ซึ่งจะทำให้เส้นตรงปรับเข้ากับข้อมูลได้ไม่ดี

ค. ค่าสังเกตทั้งหมดเท่ากันไม่คำนึงว่า X จะมีค่าใด ๆ กรณีเป็นกรณียกเว้น เมื่อทุกค่าของ Y คงที่จึงไม่มีความผันแปร และแยกความผันแปรไม่ได้ ดังนั้นค่าของ R^2 จึงพิจารณาไม่ได้ (Indeterminate)

ความผันแปรใน Y (SST) ที่แยกออกเป็น SSR และ SSE นั้นจะทำได้กับวิธีที่หาเส้นถดถอยตัวอย่างโดยวิธีกำลังสองต่ำสุดเท่านั้น สำหรับวิธีอื่นจะทำไม่ได้ แต่อย่างไรก็ตาม R^2 ก็ยังสามารถใช้กับวิธีประมาณค่าแบบอื่น ๆ ได้โดยกำหนด R^2 ดังนี้

$$R^2 = 1 - SSE/SST$$

กรณีที่เส้นถดถอยได้จากวิธีอื่น R^2 ก็ไม่สามารถแปลได้ว่าเป็นมาตราวัดของสัดส่วน (เปอร์-เซ็นต์) ของความผันแปรในตัวแปรตามที่เนื่องมาจากการถดถอยตัวอย่าง (หรือตัวแปรอิสระ) แต่ R^2 จะใช้เป็นมาตราวัด "การปรับที่ดีที่สุด"

ถ้า R^2 มีค่าต่ำ เราจะถือว่าการปรับเส้นถดถอยเข้ากับค่าสังเกตจะไม่ค่อยดี หรือความผันแปรในตัวแปรอิสระจะไม่มีผลกระทบต่อตัวแปรตาม หรืออาจกล่าวได้ว่าอิทธิพลของตัวแปรอิสระต่อตัวแปรตามมีน้อยมาก เมื่อเทียบกับอิทธิพลของตัวคลาดเคลื่อนหรือยังกล่าวได้อีกว่าสมการถดถอยที่กำหนดไว้ไม่ถูกต้อง นั่นคือ R^2 เป็นตัวชี้ถึงความถูกต้องของข้อกำหนด (Specification) ของตัวแบบ

สัมประสิทธิ์การตัดสินใจ R^2 ซึ่งได้จากตัวอย่างนั้น จะเป็นตัวประมาณค่าของสัมประสิทธิ์การตัดสินใจประชากร ρ^2 และเป็นตัวประมาณค่าที่เอียงเฉทางบวก (Positively Biased estimator) นั่นคือ

$$E(R^2) > \rho^2$$

เราแก้ความเอียงเฉด้วยองศาความเป็นอิสระซึ่งจะได้สัมประสิทธิ์การตัดสินใจที่ปรับปรุง (Adjusted Coefficient of Determination) $\bar{R}^2$ ดังนี้

$$\begin{aligned}
\bar{R}^2 &= 1 - \frac{SSE/(n-2)}{SST/(n-1)} \\
&= 1 - \frac{S^2_{Y/X}}{S^2_Y} \quad ; \quad S^2_{Y/X} = \Sigma(Y-\hat{Y})^2, \quad S^2_Y = \Sigma(Y-\bar{Y})^2 \\
&= 1 - (1-R^2)\frac{n-1}{n-2}
\end{aligned}$$

สำหรับค่าของ $\bar{R}^2$ นั้นอาจจะเป็นลบ (-) ได้ (ถ้า $S^2_{Y/X}$ มากกว่า S^2_Y) และเราจะแปลความหมายเหมือนกับ $\bar{R}^2$ เป็นศูนย์

ตัวอย่าง จากการศึกษาถึงความสัมพันธ์ของขนาดครอบครัว และค่าครองชีพในตัวอย่างที่แล้วมา เราจะแยกความผันแปรของตัวแปรตาม Y และหาสัมประสิทธิ์การตัดสินใจได้เป็น

$$\begin{aligned}
SST &= \Sigma(Y-\bar{Y})^2 \\
&= \Sigma Y^2 - (\Sigma Y)^2/n \\
&= 106 - (20)^2/5 && = 26 \\
SSR &= \Sigma(\hat{Y}-\bar{Y})^2 \\
&= \hat{\beta}_1^2 \Sigma(X-\bar{X})^2 \\
&= \hat{\beta}_1^2 [\Sigma X^2 - (\Sigma X)^2/n] \\
&= (12/13)^2 [151 - (25)^2/5] \\
&= 26(12/13)^2 && = 22.15 \\
SSE &= SST - SSR \\
&= 26 - 22.15 && = 3.85 \\
SSE &= \Sigma(Y-\hat{Y})^2 && = 3.85 \quad \text{[ดังที่หาไว้ตอนแรก]}
\end{aligned}$$

หรือ $SSE = \Sigma Y^2 - \hat{\beta}_0 \Sigma Y - \hat{\beta}_1 \Sigma XY$

$= 106 - (-8/13)(20) - (12/13)(124)$

$= 3.85$

$R^2 = \frac{SSR}{SST} = \frac{22.15}{26}$

$= 0.85$

หรือ $R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{3.85}{26}$

$= 1 - 0.15$

$= 0.85$

และ $\bar{R}^2 = 1 - \frac{SSE/(n-2)}{SST/(n-1)}$

$= 1 - \frac{3.85/(5-2)}{26/(5-1)}$

$= 1 - 1.283/6.5 = 1 - .197$

$= 0.803$

9.8 การวิเคราะห์ความแปรปรวนในการถดถอย

สมมติฐานเกี่ยวกับความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตามนั้นสามารถจะตรวจสอบหรือทดสอบได้โดยอาศัยการวิเคราะห์ความแปรปรวน ถ้าไม่มีความสัมพันธ์ระหว่างตัวแปรทั้งสองแล้วความผันแปรของตัวแปรตามจะไม่มีผลกระทบต่อการเปลี่ยนแปลงของตัวแปรอิสระ แต่จะผันแปรเพราะตัวคลาดเคลื่อนอย่างเดียว ซึ่งหมายความว่าผลรวมกำลังสองเนื่องจากการถดถอย (SSR) มีค่าต่างจากศูนย์ก็เพราะเราใช้ตัวอย่างไม่ใช่ประชากรทั้งหมด

จาก $SSR = \beta_1^2 \Sigma (X - \bar{X})^2$ นั้น ถ้า $\hat{\beta}_1 = 0$ แล้ว ค่าของ SSR ในประชากรควรจะเป็นศูนย์ด้วย หรือ $\rho^2 = 0$ ดังนั้นสมมติฐานที่ว่าไม่มีความสัมพันธ์ระหว่างตัวแปรอิสระ X และตัวแปรตาม Y จึงสามารถเขียนได้เป็น

$$H_0 : \rho^2 = 0$$

และจากความสัมพันธ์ $SST = SSR + SSE$

เมื่อ SSR เป็นศูนย์แล้ว SST จะเท่ากับ SSE ดังนั้นถ้าไม่มีความสัมพันธ์ระหว่าง X และ Y แล้วอัตราส่วน

$$F = \frac{SSR/1}{SSE/(n-2)} = \frac{MSR}{MSE}$$

จะต่างจากศูนย์ก็เพราะการใช้ตัวอย่างเท่านั้น เราสามารถแสดงได้ว่า ภายใต้ $H_0 : \rho^2 = 0$ นั้น ตัวสถิติ F = MSR/MSE จะมีการแจกแจงแบบ Snececor F, องศาความเป็นอิสระ 1, n-2

จากการใช้ตัวสถิติทดสอบ F นั้นเราได้ใช้ความผันแปรจากแหล่งต่าง ๆ คือ จากการถดถอย (หรือตัวแปรอิสระ) และความคลาดเคลื่อน เทคนิคเช่นนี้ได้ชื่อว่าเทคนิคการวิเคราะห์ความแปรปรวน ซึ่งเราสามารถสรุปได้ในตารางวิเคราะห์ความแปรปรวน ดังนี้

แหล่งความผันแปร SOV	df	SS	MS	F
X	1	SSR	MSR	MSR/MSE
ความคลาดเคลื่อน	n-2	SSE	MSE	
รวม	n-1	SST		

เมื่อตัวสถิติทดสอบ F มีค่ามากกว่าค่าวิกฤต $F_{\alpha}^{(1,n-2)}$ ณ ระดับนัยสำคัญ α เราจะปฏิเสธ $H_0 : \rho^2 = 0$ ซึ่งแสดงว่าตัวแปรอิสระ X และตัวแปรตามมีความสัมพันธ์กัน นั่นคือ $H_a : \rho^2 > 0$ เป็นจริง

ตัวอย่าง จากตัวอย่างที่แล้ว ๆ มา เราสามารถทดสอบความสัมพันธ์ระหว่างขนาดครอบครัวและค่าครองชีพด้วยเทคนิคการวิเคราะห์ความแปรปรวนได้ดังนี้

สมมติฐาน $H_0 : \rho^2 = 0$; $H_a : \rho^2 > 0$

ค่าของตัวสถิติต่าง ๆ คำนวณได้ดังนี้

SST = 26, SSR = 22.15, SSE = 3.85

MSR = SSR/1 = 22.15

MSE = SSE/(n-2) = 3.85/(5-2) = 1.28

ตารางวิเคราะห์ความแปรปรวนจะเป็น

SOV	df	SS	MS	F
ขนาดครอบครัว X	1	22.15	22.15	22.15/1.28 = 17.30
ความคลาดเคลื่อน	5-2 = 3	3.85	1.28	
รวม	5-1 = 4	26.00		

ณ ระดับนัยสำคัญ $\alpha = .05$ ค่าวิกฤต $F_{.05}^{(1,3)} = 10.13$ จึงปฏิเสธ $H_0 : \rho^2 = 0$ ซึ่งแสดงว่าขนาดครอบครัวมีความสัมพันธ์กับค่าครองชีพจริง

9.9 การนำเสนอเกี่ยวกับการถดถอย (Presentation of Regression Results)

การนำเสนอเกี่ยวกับผลที่ได้จากการวิเคราะห์การถดถอยโดยอาศัยข้อมูลจากตัวอย่าง

นั้นจะประกอบด้วย

- สมการถดถอยตัวอย่างที่ประมาณ
- ความคลาดเคลื่อนมาตรฐานของตัวประมาณค่า $(S_{\hat{\beta}_0}, S_{\hat{\beta}_1})$
- สัมประสิทธิ์การตัดสินใจ R^2

นั่นคือการนำเสนอจะเป็นดังนี้

$$\hat{Y} = \underset{(S_{\hat{\beta}_0})}{\hat{\beta}_0} + \underset{(S_{\hat{\beta}_1})}{\hat{\beta}_1}X \qquad R^2 = \ldots\ldots$$

หรือ

$$Y = \underset{(S_{\hat{\beta}_0})}{\hat{\beta}_0} + \underset{(S_{\hat{\beta}_1})}{\hat{\beta}_1}X + e \qquad R^2 = \ldots\ldots$$

ตามตัวอย่างที่ผ่านมาแล้วนั้น เราเสนอผลของการวิเคราะห์ถดถอยได้เป็น

$$\hat{Y} = \underset{(1.22)}{-8/13} + \underset{(0.22)}{(12/13)}X \qquad R^2 = 0.85$$

หรือ

$$Y = \underset{(1.22)}{-8/13} + \underset{(0.22)}{(12/13)}X + e \qquad R^2 = 0.85$$

บางครั้งเราก็แสดง MSE และ F ไว้ด้วย ดังนี้

$$\hat{Y} = \underset{(1.22)}{-8/13} + \underset{(0.22)}{(12/13)}X \qquad \begin{aligned} R^2 &= 0.85 \\ MSE &= 1.28 \\ F &= 17.30 \end{aligned}$$

9.10 ตัวแบบถดถอยเชิงเส้นชนิดพหุคูณ (Multiple Linear Regression Model)

ในทางการศึกษาและจิตวิทยานั้นความสัมพันธ์ระหว่างตัวแปรต่าง ๆ ส่วนมากสามารถอธิบายได้ด้วยตัวแบบถดถอยชนิดพหุคูณ เช่น ผลสัมฤทธิ์ด้านการอ่าน ความถนัดด้านใช้คำ และแรงจูงใจมีความสัมพันธ์กันเป็นต้น ตัวแบบถดถอยชนิดพหุคูณนี้จะแสดงให้เห็นว่าการเปลี่ยนแปลงในตัวแปรหนึ่ง (ตัวแปรตาม) นั้นสามารถอธิบายได้ด้วยการเปลี่ยนแปลงของตัวแปรอื่น ๆ หลายตัว (ตัวแปรอิสระ) ความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระต่าง ๆ เหล่านี้ อธิบายได้ดังสมการถดถอยเชิงเส้นชนิดพหุคูณ ดังนี้

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots\cdots \beta_kX_k + \varepsilon$$

ในเมื่อ Y เป็นตัวแปรตาม และ $X_1, X_2, \ldots, X_k$ เป็นตัวแปรอิสระ, กับ E เป็นตัวคลาดเคลื่อนเชิงสุ่ม

สำหรับข้อสมมติของตัวแบบถดถอยชนิดพหุคูณ จะเป็นดังนี้

1. ตัวคลาดเคลื่อนเป็นตัวแปรเชิงสุ่มที่มีการแจกแจงแบบปกติ

2. และมีค่าเฉลี่ยเป็นศูนย์ นั่นคือ $E(\varepsilon) = 0$

3. ความแปรปรวนของ ε จะเท่ากันหมด นั่นคือ $V(\varepsilon) = \sigma^2$

4. ตัวคลาดเคลื่อน ε_i และ ε_j $(i \neq j)$ จะไม่มีความสัมพันธ์กัน (Nonautocorre lation or Serial independence) นั่นคือ

$$Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i, \varepsilon_j) = 0 \text{ สำหรับ } i \neq j$$

5. ตัวแปรอิสระแต่ละตัวเป็นตัวแปรที่กำหนดได้ (Nonstochaslic) และจะไม่มีความสัมพันธ์เชิงเส้นกัน (No Perfect Multicollinear) สำหรับค่าสังเกตของมันอาจจะเลือกมาแบบสุ่มหรืออย่างมีจุดหมาย

6. จำนวนค่าสังเกตต้องมากกว่าจำนวนสัมประสิทธิ์ในสมการถดถอยที่ต้องประมาณค่า นั่นคือ $n > k+1$

จากข้อกำหนดของตัวแบบถดถอยชนิดพหุคูณที่กำหนดมานั้น เราจะได้ว่า Y มีการแจกแจงแบบปกติเช่นเดียวกับตัวแบบถดถอยชนิดธรรมดา สำหรับค่าเฉลี่ยและความแปรปรวนของมันคือ

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots\cdots \beta_k X_k$$

$$V(Y) = \sigma^2$$

สำหรับพารามิเตอร์ถดถอย $\beta_0, \beta_1, \cdots, \beta_k$ สามารถแปลความได้ดังนี้

β_0 เป็นค่าเฉลี่ยของ Y เมื่อตัวแปรอิสระทุกตัวเป็นศูนย์

$\beta_a \,/ a = 1, 2, \ldots, k$ เป็นการเปลี่ยนแปลงของค่าเฉลี่ย Y หรือ ECY ที่สอดคล้องกับการเปลี่ยนแปลงหนึ่งหน่วยของตัวแปรอิสระ ตัวที่ a โดยให้ตัวแปรอิสระที่เหลือคงที่ นั่นคือ

$$\beta_a = \frac{\Delta Y}{\Delta X_a} \quad ; \quad a = 1, 2, \ldots, k$$

และ β_a จะวัดผลกระทบ (effect) ของ X_a ต่อ $E(Y)$ เมื่อตัวแปรอิสระอื่น ๆ (นอกจาก X_a) คงที่ เรามักเรียก β_a ต่าง ๆ ว่า ความชันของการถดถอย (Regression Slopes) หรือสัมประสิทธิ์การถดถอยบางส่วน (Partical Regression Coefficients) ส่วน β_0 จะเรียกว่าจุดตัดแกน หรือค่าคงที่ของการถดถอย (Regression Constant)

9.11 การประมาณค่าของพารามิเตอร์ถดถอยด้วยวิธีกำลังสองต่ำสุด

การประมาณค่าพารามิเตอร์ถดถอยด้วยวิธีกำลังสองต่ำสุดในสมการถดถอยชนิด

พหุคูณก็ทำได้เช่นเดียวกับในสมการถดถอยชนิดธรรมดาคือ จะต้องหาเส้นถดถอยชนิดพหุคูณที่ทำให้ผลรวมของผลต่างกำลังสองระหว่างค่าทำนายและค่าจริงมีค่าน้อยที่สุด นั่นจะต้องทำให้ S_o

$$S_o = \Sigma(Y-\hat{Y})^2$$
$$= \Sigma[Y-(\hat{\beta}_0+\hat{\beta}_1X_1+\hat{\beta}_2X_2+\ldots\ldots\hat{\beta}_kX_k)]^2$$

มีค่าน้อยที่สุด

ค่าของ S_o จะขึ้นอยู่กับ $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ ดังนั้น ค่าที่เหมาะ ของ $\hat{\beta}_0, \hat{\beta}_1, \ldots\ldots, \hat{\beta}_k$ ที่ทำให้ S_o มีค่าน้อยที่สุดนั้นสามารถหาได้โดยวิธีการหาแคลคูลลัส นั่นคือ หาอนุพันธ์บางส่วนของ S_o เทียบกับ $\beta_0, \beta_1, \ldots\ldots, \beta_k$ แล้วเทียบให้เท่ากับศูนย์ แล้วจะได้สมการปกติแบบกำลังสองต่ำสุด ดังนี้

$$\Sigma Y = \hat{\beta}_0 n + \hat{\beta}_1\Sigma X_1 + \ldots\ldots \hat{\beta}_k\Sigma X_k$$
$$\Sigma X_1Y = \hat{\beta}_0\Sigma X_1 + \hat{\beta}_1\Sigma X_1^2 + \ldots\ldots \hat{\beta}_k\Sigma X_1X_k$$
$$\Sigma X_2Y = \hat{\beta}_0\Sigma X_2 + \hat{\beta}_1\Sigma X_1X_2 + \ldots\ldots \hat{\beta}_k\Sigma X_2X_k$$
$$\Sigma X_kY = \hat{\beta}_0\Sigma X_k + \hat{\beta}_1\Sigma X_1X_k + \ldots\ldots \hat{\beta}_k\Sigma X_k^2$$

จากสมการปกติเหล่านี้สามารถหาค่าของ $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ ที่ต้องการได้

สำหรับสมการปกติข้างบนนี้ เรามีวิธีการเขียนได้จากสมการต่อไปนี้

(ก) $Y = \hat{\beta}_0 + \hat{\beta}_1X_1 + \ldots\ldots \hat{\beta}_kX_k$

สมการ (1) หาผลรวมของสมการ (ก)

$$\Sigma Y = \hat{\beta}_0 n + \hat{\beta}_1\Sigma X_1 + \ldots\ldots \hat{\beta}_k\Sigma X_k$$

(2) คูณสมการ (ก) ด้วย X_1 แล้วหาผลรวม

$$\Sigma X_1Y = \hat{\beta}_0\Sigma X_1 + \hat{\beta}_1\Sigma X_1^2 + \ldots\ldots \hat{\beta}_k\Sigma X_1X_k$$

(3) คูณสมการ (ก) ด้วย X_2 แล้วหาผลรวม

$$\Sigma X_2Y = \hat{\beta}_0\Sigma X_2 + \hat{\beta}_1\Sigma X_1X_2 + \ldots\ldots \hat{\beta}_k\Sigma X_2X_k$$
$$\vdots$$

($k+1$) คูณสมการ (ก) ด้วย X_h แล้วหาผลรวม

$$\Sigma X_kY = \hat{\beta}_0\Sigma X_k + \hat{\beta}_1\Sigma X_1X_k + \ldots\ldots + \hat{\beta}_k\Sigma X_k^2$$

จากสมการแรกเมื่อหารด้วยขนาดตัวอย่าง n เราจะได้

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}_1 - \hat{\beta}_2\bar{X}_2 - \cdots - \hat{\beta}_k\bar{X}_k$$

เมื่อแทน $\hat{\beta}_0$ ลงในสมการที่เหลือ เราจะได้

$$\begin{aligned}
m_{y1} &= m_{11}\hat{\beta}_1 + m_{12}\hat{\beta}_2 + \cdots + m_{1k}\hat{\beta}_k \\
m_{y2} &= m_{21}\hat{\beta}_1 + m_{22}\hat{\beta}_2 + \cdots + m_{2k}\hat{\beta}_k \\
&\vdots \\
m_{yk} &= m_{k1}\hat{\beta}_1 + m_{k2}\hat{\beta}_2 + \cdots + m_{kk}\hat{\beta}_k
\end{aligned}$$

ในเมื่อ

$$\begin{aligned}
m_{yj} &= \Sigma(Y-\bar{Y})(X-\bar{X}_k) = \Sigma y x_j \quad ; \quad j = 1, 2, \ldots, k \\
m_{ij} &= \Sigma(X-\bar{X}_i)(Y-\bar{Y}_j) = \Sigma x_i x_j \quad ; \quad i, j = 1, 2, \ldots, k
\end{aligned}$$

จากสมการเหล่านี้เราสามารถหาค่า $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ ได้โดยวิธีแก้สมการทางพีชคณิต

สำหรับกรณีที่มีตัวแปรอิสระ 2 ตัว เราจะมีสมการของตัวแบบถดถอยเชิงเส้นเป็นดังนี้

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

นั่นคือสมการถดถอยตัวอย่างเป็น

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

ในเมื่อ $\hat{\beta}_0, \hat{\beta}_1,$ และ $\hat{\beta}_2$ หาได้โดยวิธีประมาณค่าแบบกำลังสองต่ำสุดดังนี้

$$\hat{\beta}_1 = \frac{m_{Y1}m_{22} - m_{Y2}m_{12}}{m_{11}m_{22} - m_{12}^2}$$

$$\hat{\beta}_2 = \frac{m_{Y2}m_{11} - m_{Y1}m_{12}}{m_{11}m_{22} - m_{12}^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} - \hat{\beta}_2\bar{X}_2$$

ตัวอย่าง จากการศึกษาผลสัมฤทธิ์ด้านวิทยาศาสตร์ (Y) ความถนัดทางคณิตศาสตร์ (X_1) และแรงจูงใจ (X_2) เพื่อที่จะใช้ทำนายผลสัมฤทธิ์ด้านวิทยาศาสตร์จากความถนัดทางคณิตศาสตร์ และแรงจูงใจ ได้ข้อมูลมาดังนี้

Y	2	1	1	1	5	4	7	6	7	8	3	3	6	6	10
X_1	2	2	1	1	3	4	5	5	7	6	4	3	6	6	8
X_2	4	4	4	3	6	6	3	4	3	3	5	5	9	8	6

Y	9	6	6	9	10
X_1	9	10	9	4	4
X_2	7	5	5	7	7

สมการถดถอยตัวอย่างที่ต้องการประมาณ จะเป็น

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

จากข้อมูลเหล่านี้เราได้

$\Sigma Y = 110$ $\quad \Sigma X_1 = 99$ $\quad \Sigma X_2 = 104$

$\Sigma Y^2 = 770$ $\quad \Sigma X_1^2 = 625$ $\quad \Sigma X_2^2 = 600$

$\Sigma X_1 Y = 645$ $\quad \Sigma X_2 Y = 611$ $\quad \Sigma X_1 X_2 = 530$

$$m_{y1} = \Sigma x_1 y = \Sigma X_1 Y - (\Sigma X_1)(\Sigma Y)/n$$
$$= 645 - 99(110)/20 = 645 - 544.50 = 100.50$$

$$m_{y2} = \Sigma x_2 y = \Sigma X_2 Y - (\Sigma X_2)(\Sigma Y)/n$$
$$= 611 - 104(110)/20 = 611 - 572 = 39.00$$

$$m_{12} = \Sigma x_1 x_2 = \Sigma X_1 X_2 - (\Sigma X_1)(\Sigma X_2)/n$$
$$= 530 - 99(104)/20 = 538 - 514.80 = 23.20$$

$$m_{11} = \Sigma x_1^2 = \Sigma X_1^2 - (\Sigma X_1)^2/n$$
$$= 625 - (99)^2/20 = 625 - 490.05 = 134.95$$

$$m_{22} = \Sigma x_2^2 = \Sigma X_2^2 - (\Sigma X_2)^2/n$$
$$= 600 - (104)^2/20 = 600 - 540.80 = 59.20$$

$$m_{yy} = \Sigma y^2 = \Sigma Y^2 - (\Sigma y)^2/n$$
$$= 770 - (110)^2/20 = 770 - 605 = 165.00$$

ดังนั้น

$$\hat{\beta}_1 = \frac{m_{y1} m_{22} - m_{y2} m_{12}}{m_{11} m_{22} - m_{12}^2}$$

$$= \frac{100.50(59.20) - 39.00(23.20)}{134.95(59.20) - (23.20)^2}$$

$$= \frac{5949.60 - 904.80}{7989.04 - 538.24}$$

$$= \frac{5044.80}{7450.80} = .6771$$

$$\hat{\beta}_2 = \frac{m_{y2}m_{11} - m_{y1}m_{12}}{m_{11}m_{22} - m_{12}^2} = \frac{39.00(134.95) - 100.50(23.20)}{134.95(59.20) - (23.20)^2}$$

$$= \frac{5263.05 - 2331.60}{7989.04 - 538.24}$$

$$= \frac{2931.41}{7450.80} = .3934$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} - \hat{\beta}_2\bar{X}$$

$$= 110/20 - (.6771)(99/20) - (.3934)(104/20) = .1027$$

เพราะฉะนั้นสมการถดถอยตัวอย่างที่ต้องการ คือ

$$\hat{Y} = .1027 + .6771(X_1) + .3934(X_2)$$

สำหรับตัวประมาณค่า $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ นี้จะเป็นตัวประมาณค่าที่ไม่เอียงเฉของพารามิเตอร์ $\beta_0, \beta_1, \beta_2$ นั่นคือ

$$E(\hat{\beta}_0) = \beta_0 , \quad E(\hat{\beta}_1) = \beta_1 , \quad E(\hat{\beta}_2) = \beta_2$$

และความแปรปรวนของตัวประมาณค่าจะเป็น

$$V(\hat{\beta}_1) = \frac{\sigma^2 m_{22}}{m_{11}m_{22} - m_{12}^2}$$

$$V(\hat{\beta}_2) = \frac{\sigma^2 m_{11}}{m_{11}m_{22} - m_{12}^2}$$

$$V(\hat{\beta}_0) = \sigma^2\left[1/n + \frac{\bar{X}_1^2 m_{22} + \bar{X}_2^2 m_{11}}{m_{11}m_{22} - m_{12}^2}\right]$$

$$Cov(\hat{\beta}_1, \hat{\beta}_2) = \frac{-\sigma^2 m_{12}}{m_{11}m_{22} - m_{12}^2}$$

เนื่องจาก σ^2 นั้นโดยทั่วไปเราไม่ทราบค่า ตัวประมาณค่าที่ไม่เอียงเฉของมันคือ S^2

$$S^2 = 1/n\text{-}3\,[Y - \hat{Y}]^2$$

$$= 1/n\text{-}3\,[m_{yy} - \hat{\beta}_1 m_{y1} - \hat{\beta}_2 m_{y2}]$$

สำหรับตัวแปรอิสระ k ถ้าเราจะได้

$$S^2 = 1/n\text{-}k\text{-}1\ \Sigma(Y - \hat{Y})^2$$

$$= 1/n\text{-}k\text{-}1\,[m_{yy} - \hat{\beta}_1 m_{y1} - \hat{\beta}_2 m_{y2} - \cdots - \hat{\beta}_k m_{yk}]$$

ตัวอย่าง จากตัวอย่างที่แล้วมา เราหา $\hat{Y}$ และ $Y-\hat{Y}$ ได้ดังนี้

Y	X_1	X_2	$\hat{Y} = .1027 + (.6771)X_1 + (.3934)X_2$	$Y-\hat{Y}$
2	2	4	$.1027 + .6771(2) + .3934(4) = 3.0305$	-1.0305
1	2	4	$.1027 + .6771(2) + .3934(4) = 3.0305$	-2.0305
1	1	4	$.1027 + .6771(1) + .3934(4) = 2.3534$	-1.3534
			⋮	
6	9	5	$.1027 + .6771(9) + .3934(5) = 8.1636$	-2.1636
9	4	7	$.1027 + .6771(4) + .3934(7) = 5.5649$	3.4351
10	4	7	$.1027 + .6771(4) + .3934(7) = 5.5649$	4.4351

แล้วเราจะได้ $\Sigma(Y-\hat{Y})^2 = 81.6091$

ดังนั้น
$$S^2 = 1/n-3\ \Sigma(Y-\hat{Y})^2$$
$$= 1/20-3\ (81.6091)$$
$$= 4.8005$$

หรือ
$$S^2 = 1/n-3\left[m_{yy} - \hat{\beta}_1 m_{y1} - \hat{\beta}_2 m_{y2}\right]$$
$$= 1/20-3\left[165 - .6771(100.50) - .3934(39.00)\right]$$
$$= 4.8005$$

เพราะฉะนั้นจะได้ค่าประมาณของ $V(\hat{\beta}_0)$, $V(\hat{\beta}_1)$, $V(\hat{\beta}_2)$ และ $Cov(\hat{\beta}_1 \hat{\beta}_2)$ ดังนี้

$$S^2_{\hat{\beta}_1} = \frac{4.8005\ (59.20)}{(134.95)(59.20)-(23.20)^2} = 0.038$$
$$S_{\hat{\beta}_1} = .195$$
$$S^2_{\hat{\beta}_2} = \frac{4.8005\ (134.95)}{(134.95)(59.20)-(23.20)^2} = 0.087$$
$$S_{\hat{\beta}_2} = .295$$
$$S^2_{\hat{\beta}_0} = 4.8005\left[1/20 + \frac{(99/20)^2(59.20) + (104/20)^2(134.95)}{134.95(59.20)-(23.20)^2}\right]$$
$$= 4.8005(0.05 + 0.2951)$$
$$= 1.6565 \qquad \therefore S_{\hat{\beta}_0} = 1.287$$
$$S_{\hat{\beta}_1\hat{\beta}_2} = \frac{-4.8005\ (23.20)}{134.95(59.20)-(23.20)^2} = -.01495$$

9.12 ช่วงเชื่อมั่นของพารามิเตอร์ถดถอย

ตัวประมาณค่า $\hat{\beta}_i$; $i = 0,1,2,\ldots,k$ จะมีการแจกแจงแบบปกติที่มีค่าเฉลี่ย β_i และความแปรปรวน $V(\beta_i)$ ดังนั้นเราจึงได้ว่า

$$(\hat{\beta}_i - \beta_i)/S_{\hat{\beta}_i} \sim t^{(n-k-1)} \quad ; \quad i = 0,1,2,\ldots,k$$

ในเมื่อ $S_{\hat{\beta}_i}$ เป็นค่าประมาณของความคลาดเคลื่อนมาตรฐานของ $\hat{\beta}_i$

จากการแจกแจงข้างบนนี้เราจึงสร้างช่วงเชื่อมั่น แบบ $100(1-\alpha)\%$ สำหรับ β_i ; $i = 0,1,2,\ldots,k$ ได้เป็น

$$\beta_i = \hat{\beta}_i \pm t_{\alpha/2}^{(n-k-1)} S_{\hat{\beta}_i}$$

ตัวอย่าง จากตัวอย่างที่ผ่านมา เราประมาณช่วงเชื่อมั่น 95% สำหรับ β_0, β_1 และ β_2 ได้เป็น

$$\begin{aligned}
\beta_0 &= \hat{\beta}_0 \pm t_{.025}^{(17)} S_{\hat{\beta}_0} \\
&= .1027 \pm 2.110(1.287) = .1027 \pm 2.716 \\
&= -2.6133 \quad , \quad 2.8187 \\
\beta_1 &= \hat{\beta}_1 \pm t_{.025}^{(17)} S_{\hat{\beta}_1} \\
&= .6771 \pm 2.110(0.195) = .6771 \pm .4114 \\
&= .2657 \quad , \quad 1.0885
\end{aligned}$$

และ

$$\begin{aligned}
\beta_2 &= \hat{\beta}_2 \pm t_{.025}^{(17)} S_{\hat{\beta}_1} \\
&= .3934 \pm 2.110(.295) = .3934 \pm .62245 \\
&= -.229 \quad , \quad 1.016
\end{aligned}$$

9.13 การทำนายหรือการพยากรณ์ (Prediction or Forecasting)

(1) แถบเชื่อมั่น (Confidence band) สำหรับค่าเฉลี่ย $E(Y)$ เมื่อกำหนดค่าของตัวแปรอิสระชุดหนึ่ง $(X_{01}, X_{02}, \ldots, X_{0k})$ เราสามารถสร้างช่วงเชื่อมั่นหรือแถบเชื่อมั่นของค่าเฉลี่ย

ได้จากตัวประมาณค่า $\hat{Y}$ ซึ่งมีการแจกแจงแบบปกติ โดยมีค่าเฉลี่ย และความแปรปรวนดังนี้

$$\begin{aligned}
E(\hat{Y}_0) &= \beta_0 + \beta_1 X_{01} + \beta_2 X_{02} + \cdots + \beta_k X_{0k} = E(Y_0) \\
V(\hat{Y}_0) &= \sigma^2/n + \sum_i^k x_i^2 V(\beta_i) + 2\sum_{i<j}^k x_i x_j \,Cov(\hat{\beta}_i, \hat{\beta}_j) \\
x_i &= X_{0i} - \bar{X}_i \quad ; \quad i = 1,2,\ldots,k
\end{aligned}$$

ตัวประมาณค่าของ $V(\hat{Y}_0)$ ที่ไม่เอียงเฉ คือ $S_{\hat{Y}}$ ซึ่งได้จากการแทน σ^2 ด้วย S^2 ใน $V(\hat{Y})$ นั่นเอง และเราทราบว่า

$$(\hat{Y} - E(\hat{Y}))/S_{\hat{Y}} \sim t^{(n-k-1)}$$

ดังนั้นเมื่อกำหนดค่าของตัวแปรอิสระชุดหนึ่งให้ เราสามารถสร้างแถบเชื่อมั่นแบบ สำหรับค่าเฉลี่ย $E(Y)$ ได้เป็น

$$E(Y) = \hat{Y} \pm t_{\alpha/2}^{(n-k-1)} S_{\hat{Y}}$$

(2) ค่าทำนาย Y_0 เมื่อกำหนดค่าของตัวแปรอิสระชุดหนึ่งให้ตัวพยากรณ์หรือตัวทำนายที่ดีที่สุดของ Y_0 ก็คือ $E(Y_0)$ เพราะว่าความแปรปรวนของ Y_0 รอบ $E(Y_0)$ น้อยกว่ารอบจุดอื่นใด แต่เราไม่ทราบค่า $E(Y_0)$ จึงใช้ค่าประมาณ $\hat{Y}_0$ แทน เนื่องจากความคลาดเคลื่อนพยากรณ์ (Forecast error) หรือ $(Y_0 - \hat{Y}_0)$ นั้นเป็นตัวแปรเชิงสุ่มที่แจกแจงปกติ และมีค่าเฉลี่ยกับความแปรปรวนดังนี้

$$E(Y_0 - \hat{Y}_0) = 0$$

$$V(Y_0 - \hat{Y}_0) = \sigma^2 + \sigma^2/n + \sum X_i^2 V(\beta_i) + 2\sum_{i<j}^{k} X_i X_j \cdot Cov(\hat{\beta}_i, \beta_i)$$

$$X_i = X_{0i} - \bar{X}_i \quad ; \quad i = 1, 2, \ldots, k$$

ตัวประมาณค่าที่ไม่เอียงเฉของ $V(Y_0 - \hat{Y}_0)$ หรือ $\hat{\sigma}_f^2$ คือ S_f^2 ซึ่งได้จากการแทน σ^2 ด้วย S^2 และเราจะได้ว่า

$$(Y_0 - \hat{Y}_0)/S_f \sim t^{(n-k-1)}$$

จากผลอันนี้เราก็สามารถสร้างช่วงพยากรณ์ แบบ $100(1-\alpha)\%$ สำหรับค่าจริงของ Y_0 ได้เป็น

$$Y_0 = \hat{Y} \pm t_{\alpha/2}^{(n-k-1)} S_f$$

สำหรับตัวสถิติ $(Y_0 - \hat{Y}_0)/S_f$ นี้ เราสามารถใช้ทดสอบสมมติฐานที่ว่าค่าสังเกตใหม่ เช่นค่าสังเกตตัวที่ $n+1$ มาจากประชากรเดียวกับค่าสังเกต n ค่าที่ใช้ประมาณค่าพารา-มิเตอร์ถดถอย

เช่นเดียวกันกับตัวสถิติ $(Y_0 - M_0)/S_{\hat{Y}}$ ซึ่งใช้ทดสอบสมมติฐานเกี่ยวกับค่าเฉลี่ย

ที่ระบุไว้ (M_0) ได้

ตัวอย่าง จากตัวอย่างที่แล้ว ๆ มา ถ้าเด็กคนหนึ่งมีถนัดทางคณิตศาสตร์เป็น $X_{01} = 3$ และแรงจูงใจเป็น $X_{02} = 4$ เราก็สามารถพยากรณ์ผลสัมฤทธิ์ด้านวิทยาศาสตร์ได้ดังนี้

$$X_{01} = 3 \qquad X_{02} = 4$$

$$\hat{Y}_0 = .1027 + .6771(3) + .3934(4)$$

$$= 3.7076$$

$$S_f^2 = S^2 + S^2/n + (X_{01} - \bar{X}_1)^2 S^2_{\hat{\beta}_1} + (X_{02} - \bar{X}_2)^2 S^2_{\hat{\beta}_2} + 2(X_{01} - \bar{X}_1)(X_{02} - \bar{X}_2) S_{\hat{\beta}_1\hat{\beta}_2}$$

$$= 4.8005 + 4.8005/20 + (3 - 99/20)^2(.038) + (4 - 104/20)^2(.087) + 2(3 - 99/20)(4 - 104/20)(-.01495)$$

$$= 5.241$$

$$S_f = 2.2895$$

ดังนั้นช่วงพยากรณ์ 95% สำหรับ Y_0 จะเป็น

$$Y_0 = \hat{Y}_0 \pm t^{(17)}_{.025} S_f$$

$$= 3.7076 \pm 2.110(2.2895)$$

$$= -.8714 \ , \ 8.2866$$

9.14 การแยกความผันแปรของตัวแปรตาม

ความผันแปรในตัวอย่างของตัวแปรตาม Y สามารถแยกออกได้เช่นเดียวกับการถดถอยชนิดธรรมดา ดังนี้

$$\Sigma(Y - \bar{Y})^2 = \Sigma(\hat{Y} - \bar{Y})^2 + \Sigma(Y - \hat{Y})^2$$

หรือ $$SST = SSR + SSE$$

ในเมื่อ $$SSR = \Sigma(\hat{Y} - \bar{Y})^2$$

$$= \hat{\beta}_1 m_{Y1} + \hat{\beta}_2 m_{Y2} + \cdots\cdots + \hat{\beta}_k m_{Yk}$$

$$= \hat{\beta}_1^2 m_{11} + \hat{\beta}_2^2 m_{22} + \cdots\cdots + \hat{\beta}_k^2 m_{kk}$$

เราจะเห็นว่า SSR เป็นค่าประมาณของผลกระทบของการถดถอยต่อตัวแปรตาม Y ซึ่งประกอบด้วยผลกระทบของ X_1, ผลกระทบของ X_2, ..., และผลกระทบของ X_k ตามลำดับ

9.15 สัมประสิทธิ์การตัดสินใจแบบพหุคูณ (Coefficient of Multiple Determination)

สัมประสิทธิ์การตัดสินใจแบบพหุคูณกำหนดไว้ดังนี้

$$R^2 = SSR/SST$$
$$= 1 - SSE/SST$$

ในกรณีปรับปรุงแก้ไข เราได้

$$\bar{R}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-k-1)}$$
$$= 1 - (1-R^2)\frac{n-1}{n-k-1}$$

จะเป็นมาตราวัดความมากน้อย (How Well) ที่เส้นถดถอยสามารถปรับเข้ากับข้อมูลตัวอย่าง และยังใช้เปรียบเทียบการปรับที่ดีที่สุด (Goodness of fit) ของสมการถดถอยหลาย ๆ สมการที่ผันแปรไปกับจำนวนตัวแปรอิสระและจำนวนค่าสังเกต

ตัวอย่าง จากตัวอย่างที่แล้ว ๆ มา เราแยกความผันแปรของตัวแปรตาม Y ได้ดังนี้

$$SSR = \beta_1 m_{y1} + \beta_2 m_{y2}$$
$$= (.6711)(100.50) + (.3934)(39.00)$$
$$= 83.3912$$
$$SSE = \Sigma(Y-\hat{Y})^2 = 81.6091$$
$$SST = \Sigma(Y-\bar{Y})^2$$
$$= m_{yy} = 165.00$$

เมื่อคำนวณสัมประสิทธิ์การตัดสินใจจะได้

$$R^2 = SSR/SST = 83.3912/165$$
$$= .5054$$

หรือ
$$\bar{R}^2 = 1 - \frac{81.6091/(20-2-1)}{165/(20-1)}$$
$$= 1 - 4.801/8.4 = .447$$

ซึ่งแสดงว่า 50.54% ของความผันแปรสำหรับตัวแปรตาม Y จะเนื่องมาจากผลกระทบของความผันแปรใน X_1 และความผันแปรใน X_2

9.16 การทดสอบสมมติฐานเกี่ยวกับการถดถอยเชิงเส้นชนิดพหุคูณ

ในการวิเคราะห์การถดถอยสำหรับตัวแบบถดถอยเชิงเส้นชนิดพหุคูณนั้นมีสมมติฐานที่น่าสนใจจะทดสอบดังนี้

ก. สมมติฐานเกี่ยวกับค่าของ β_i ; $i = 0, 1, 2, \ldots, k$ นั่นคือ ระบุค่าของ β_i ว่าเป็น β_{io} ดังนั้นสมมติฐาน Ho จึงเป็น

$$H_0 : \beta_i = \beta_{io} \quad ; \quad i = 0, 1, 2, \ldots, k$$

ตัวสถิติทดสอบ คือ $T = (\hat{\beta}_i - \beta_{io}) / S_{\hat{\beta}_i}$; $i = 0, 1, 2, \ldots, k$ ซึ่งมีการแจกแจงแบบ

Student t, องศาความเป็นอิสระ n-k-1

ส่วนมากเราสนใจทดสอบ $H_0 : \beta_i = 0$; $i = 0, 1, 2, \ldots, k$ แต่ถ้า I = 0 จะหมายความว่าระนาบถดถอย (Regression Plane) ผ่านจุดกำเนิด และถ้า i = 1,2,..., h จะหมายถึงตัวแปรอิสระ X_i ไม่มีอิทธิพลต่อค่าเฉลี่ยของตัวแปรตาม E(Y) ซึ่งถ้าปฏิเสธ Ho ไม่ได้ เราก็จะสรุปว่า X_i เป็นตัวแปรที่ไม่เกี่ยวข้องในสมการถดถอย

สำหรับสมมติฐาน $H_0 : \beta_i = 0$; $H_a : \beta_i \neq 0$ $(i = 0, 1, 2, \ldots, k)$ นั้นเรามีวิธีทดสอบดังนี้

(1) ทดสอบโดยใช้ความคลาดเคลื่อนมาตรฐาน (Standard Error Test) นั่นคือใช้ความคลาดเคลื่อนมาตรฐานของตัวประมาณค่า $(S_{\hat{\beta}_i})$ เปรียบเทียบกับค่าของตัวประมาณค่า

เราจะปฏิเสธ $H_0 : \beta_i = 0$ หรือค่าประมาณของพารามิเตอร์ β_i มีนัยสำคัญเชิงสถิติหรือแตกต่างจากศูนย์อย่างมีนัยสำคัญ ถ้า $S_{\hat{\beta}_i} < \frac{1}{2} \hat{\beta}_i$

การทดสอบวิธีนี้เป็นการทดสอบโดยประมาณและมีระดับนัยสำคัญ 5%

(2) ทดสอบด้วยตัวสถิติ T (Student's Test) ใช้การเปรียบเทียบค่าของตัวสถิติ T

$$T = \hat{\beta}_i / S_{\hat{\beta}_i} \quad ; \quad i = 0, 1, 2, \ldots, k$$

จากค่าสังเกตของตัวอย่าง กับ ค่าทางทฤษฎีจากตาราง Student t ที่มีองศาความเป็นอิสระ n-k-1 ณ ระดับนัยสำคัญที่ต้องการ

สมมติฐาน $H_0 : \beta_i = 0$ จะได้รับการปฏิเสธ ณ ระดับนัยสำคัญ α ถ้า $|T| > t_{\alpha/2}^{(n-k-1)}$ นั่นคือ ถ้า $i = 1, 2, \ldots, k$ เมื่อปฏิเสธ $H_0 : \beta_i = 0$ ก็จะหมายความว่าตัวแปรอิสระ X_i จะเป็นตัวอธิบายความผันแปรในตัวแปรตาม Y

ตัวอย่าง จากตัวอย่างที่แล้ว ๆ มา เราจะทดสอบสมมติฐาน

$$H_0 : \beta_i = 0 \qquad i = 0, 1, 2$$

$$H_a : \beta_i \neq 0$$

$$T_0 = \hat{\beta}_0 / S_{\hat{\beta}_0}$$

$$= .1027 / 1.287 \qquad = 3.478$$

โดยที่ $|T_0|$ โตกว่าค่าวิกฤต $t_{.025}^{(17)} = 2.110$ จึงปฏิเสธ $H_0 : \beta_0 = 0$

$$T_1 = \hat{\beta}_1 / S_{\hat{\beta}_1} = .6771/0.195 = 3.472$$

โดยที่ $|T_1|$ โตกว่าค่าวิกฤต $t_{.025}^{(17)} = 2.110$ จึงปฏิเสธ $H_0 : \beta_1 = 0$

$$T_2 = \hat{\beta}_2 / S_{\hat{\beta}_2} = .3934/0.295 = 1.3336$$

โดยที่ $|T_2|$ โตกว่าค่าวิกฤต $t_{.025}^{(17)} = 2.110$ จึงปฏิเสธ $H_0 : \beta_2 = 0$ ไม่ได้ นั่นคือ X_2 ไม่มีอิทธิพลต่อ Y

เพราะฉะนั้นระนาบถดถอยจะไม่ผ่านจุดกำเนิด และตัวแปรอิสระ X_1 จะเป็นตัวอธิบายความผันแปรในตัวแปรตาม Y แต่ X_2 ไม่เป็น

ข. สมมติฐานเกี่ยวกับตัวแปรอิสระทุกตัวมีอิทธิพลต่อค่าเฉลี่ยของตัวแปรตาม Y หรือนัยสำคัญทั้งหมดของการถดถอย (Overall Significance of Regression) สมมติฐานที่จะทดสอบ Ho จะเป็น

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

หรือ $$H_0 : \rho_1^2 = 0$$

ถ้า Ho เป็นจริงก็จะแสดงว่าความผันแปรของตัวแปรตาม Y ไม่ได้มีผลจากการเปลี่ยนแปลงของตัวแปรอิสระตัวใดตัวหนึ่ง แต่เนื่องจากความคลาดเคลื่อนอย่างเดียว นั่นคือความผันแปรเนื่องจากการถดถอย (SSR) จะต่างจากศูนย์ก็เพราะการสุ่มตัวอย่างเท่านั้น

ตัวสถิติทดสอบจึงใช้ F

$$F = \frac{SSR/k}{SSE/(n-k-1)} = MSR/MSE$$

ซึ่งมีการแจกแจงแบบ Snedecor F, องศาความเป็นอิสระ h, n-h-1

ดังนั้นจะปฏิเสธ Ho ถ้าค่าสังเกตจากตัวอย่างของตัวสถิติทดสอบ F มากกว่าค่าวิกฤตจากตาราง Snedecor F ที่มีองศาความเป็นอิสระ h, n-k-1 และระดับนัยสำคัญ ที่ต้องการ

ตัวสถิติทดสอบ F สามารถแปลงให้อยู่ในเทอมของสัมประสิทธิ์กับตัดสินใจได้เป็น

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

ในการทดสอบสมมติฐานดังกล่าวนี้สามารถสรุปได้ในตารางวิเคราะห์ความแปรปรวนดังนี้

SOV	df	SS	MS	F
$X_1, X_2, \ldots, X_k$ Residual	k n-k-1	SSR SSE	MSR MSE	MSR/MSE
Total	n-1	SST		

ตัวอย่าง ตามตัวอย่างเรื่องผลสัมฤทธิ์ทางวิทยาศาสตร์ เราอาศัยเทคนิคการวิเคราะห์ความแปรปรวนทดสอบนัยสำคัญทั้งหมดของการถดถอยได้ดังนี้

$$SST = 165.00 \qquad SSR = 83.3912$$

$$SSE = 81.6091$$

SOV	df	SS	MS	F
X_1, X_2 Residual	2 17	83.3912 81.6091	41.6956 4.8005	8.686
Total	19	165.00		

ณ ระดับนัยสำคัญ $\alpha = .05, .01$ ค่าวิกฤต $F_{.05}^{(2,17)} = 3.59$ และ $F_{.01}^{(2,17)} = 6.01$ เราจึงปฏิเสธ $H_0 : \beta_1 = \beta_2 = 0$ ณ ระดับนัยสำคัญ $\alpha = .01$ นั่นคือทั้ง X_1 และ X_2 มีอิทธิพลต่อตัวแปรตาม Y

ค. สมมติฐานเกี่ยวกับอิทธิพลต่อตัวแปรตามของตัวแปรอิสระที่เพิ่มเข้าไปในการถดถอย (Improvement of fit obtained from additional independent variables) จุดประสงค์ของการเพิ่มตัวแปรอิสระเข้าไปในการถดถอยก็เพื่อจะเพิ่มความถูกต้อง (Accunacy) ของการทำนายตัวแปรตาม หรือเป็นการลดความคลาดเคลื่อน (Rednce deriation or error) จากการทำนายหรือการถดถอย

ลองพิจารณาข้อเสนอของการถดถอย 2 แบบต่อไปนี้

ข้อเสนอแรกกล่าวว่าสมการถดถอยจะเป็น

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

ข้อเสนอที่สองกล่าวว่าตัวแปรตาม Y นอกจากจะขึ้นกับ $X_1, X_2, \ldots, X_u$ แล้วยังขึ้นอยู่

กับตัวแปรอิสระอื่น ๆ อีก คือ $X_{k+1}, X_{k+2}, \ldots, X_q$ t ดังนั้นสมการถดถอยจะเป็น

$$Y = \beta_0 + \beta_1 X_1 + \ldots\ldots + \beta_k X_k + \beta_{k+1} X_{k+1}, + \ldots\ldots, \beta_q X_q + \varepsilon$$

ดังนั้นข้อเสนอแรกจะแยกความผันแปรของตัวแปรตามได้เป็น

$$SST = SSR_k + SSE_k$$

และข้อเสนอหลังจะเป็น $SST = SSR_q + SSE_q$

ซึ่งทั้งสองข้อเสนอนั้นจะมีความผันแปรของตัวแปรตาม (SST) เท่ากัน

ถ้าเราต้องการทดสอบว่าตัวแปรอิสระที่เพิ่มเข้าไป $X_{h+1}, X_{h+2}, \ldots, X_q$ นั้นจะมีอิทธิพลต่อตัวแปรตามหรือจะเพิ่มความถูกต้องในการทำนายหรือไม่ เราก็ทำได้โดยการทดสอบสมมติฐาน Ho

$$H_o : \beta_{k+1} = \beta_{k+2} = \ldots\ldots = \beta_k = 0$$

ถ้าตัวแปรอิสระที่เพิ่มเข้าไปนั้นจะไม่มีอิทธิพลต่อตัวแปรตาม แล้ว SSR_k กับ SSR_q จะแตกต่างกันเพราะความคลาดเคลื่อนจากการสุ่มตัวอย่างเท่านั้น นั่นคือ SSR_k และ SSR_q ในประชากรจะเป็นจำนวนเดียวกัน

ถ้า Ho เป็นจริง แล้วเราสามารถแสดงได้ว่า

$$F_a = \frac{(SSR_q - SSR_k)/(q-k)}{SSE_q/(n-q-1)}$$

จะมีการแจกแจงแบบ Shedeca F, องศาความเป็นอิสระ $q-k$ และ $n-q-1$

ดังนั้นเราจึงใช้ F_a เป็นตัวสถิติทดสอบ สมมติฐาน Ho ณ ระดับนัยสำคัญ α ที่ต้องการได้

เราสามารถแสดงได้ว่า $SSR_q \geqslant SSR_k$ นั่นคือ F_a ไม่มีทางเป็นลบ จากผลอันนี้ทำให้เราทราบว่า $SSR_q/SST \geqslant SSR_k/SST$ หรือ $R_q^2 \geqslant R_k$ ซึ่งหมายความว่าการเพิ่มตัวแปรอิสระเข้าไปในสมการถดถอยนั้นจะเป็นการเพิ่มประสิทธิ์ของการตัดสินใจ R^2 แต่อาจจะไม่เป็นจริงสำหรับ R'^2

การทดสอบสมมติฐาน Ho : $\beta_{k+1} = \beta_{k+2} = \ldots\ldots \beta_q = 0$

ที่อาศัยเทคนิคการวิเคราะห์ความแปรปรวนนั้นจะสรุปผลได้ดังตารางต่อไปนี้

SOV	df	SS	MS	F
$X_1, X_2, \ldots\ldots, X_k$	k	SSR_k	MSR_k	
$X_1, X_2, \ldots\ldots, X_k, X_{k+1}, \ldots, X_q$	q	SSR_q	MSR_q	

SOV	df	SS	MS	F
$X_{k+1}, X_{k+2}, \ldots\ldots, X_q$	$q-k$	$SSR_q - SSR_k$	MSR_{q-k}	$\frac{MSR_{q-k}}{MSE_q}$
Residual	$n-q-1$	SSE_q	MSE_q	
Total	$n-1$	SST		

ตัวอย่าง จากตัวอย่างที่ผ่านมา ถ้าเป็นการถดถอยชนิดธรรมดาของ Y กับ X_1 อย่างเดียวเราจะได้

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

ซึ่งประมาณด้วยวิธีกำลังสองต่ำสุดได้เป็น

$$\hat{Y} = 1.8137 + .7447 X_1$$

$$\hat{\beta}_1 = \Sigma X_1 Y / \Sigma X_1^2 = 100.50/134.95 = .7447$$

$$\hat{\beta}_0 = 110/20 - .7447(99/20) = 1.8137$$

สำหรับ

$$SSR_k = (\Sigma X_1 Y)^2 / \Sigma X_1^2$$
$$= \hat{\beta}_1^2 \Sigma X_1^2$$
$$= (.7447)^2 (134.95) = 74.8444$$

$$SSE_k = SST - SSR_k$$
$$= 165 - 74.8444 = 90.1556$$

$$R^2_{Y_1} = SSR_k / SST$$
$$= 74.8444/165 = .4536$$

$$F = \frac{R^2_{Y.1}/k}{(1-R^2_{Y.1})/(n-k-1)}$$
$$= \frac{.4536/1}{(1-.4536^2)/(20-1-1)}$$
$$= .4536/.0304$$
$$= 14.9211$$

หรือ

$$F = \frac{SSR_k/k}{SSE_k/(n-k-1)} = \frac{74.8444/1}{90.1556/18}$$
$$= 14.9432$$

อัตราส่วน F นี้มีนัยสำคัญ ณ ระดับ .01 ดังนั้นการถดถอยของ Y กับ X_1 อย่างเดียวมีนัยสำคัญ

ทางสถิติ โดยที่ R^2 = .45 เราจึงพูดได้ว่า 45% ของความผันแปรใน Y เนื่องมาจาก X_1

ลองมาพิจารณาการถดถอยของ Y กับ X_2 อย่างเดียว เราจะได้สมการทดถอยตัวอย่างเป็น

$$\hat{Y} = 2.2389 + .6588 X_2$$

ในเมื่อ

$$\hat{\beta}_1 = \Sigma X_2 Y / \Sigma X_2^2$$
$$= 39.00/59.20 \qquad = .6588$$
$$\hat{\beta}_0 = 110/20 - (.6588)(104/20) \qquad = 2.0742$$
$$SSR_k = (\Sigma X_2 y)^2 / \Sigma X_2^2$$
$$= (39.00)^2/59.20 \qquad = 25.6926$$
$$SSE_k = SST - SSR_k \quad = 165.00 - 25.6926 \quad = 139.3074$$
$$R^2_{y.2} = SSR/SST \quad = 25.6926/165.00 \quad = .1557$$
$$F = \frac{SSR/k}{SSE/(n-k-1)} = \frac{25.6926/1}{139.3074/(10-1-1)}$$
$$= 3.320$$

เราจะเห็นว่าอัตราส่วน F ไม่มีนัยสำคัญ ณ ระดับ .05 ดังนั้น X_1 เป็นตัวทำนายของ Y ที่ดีกว่า X_2 เมื่อเราพิจารณาแยกกัน

ต่อไปเราจะพิจารณาว่าถ้าเพิ่ม X_2 เข้าไปในสมการถดถอย Y กับ X_1 แล้วมันจะเพิ่มความจากห้องของการทำนาย Y หรือไม่

จาก $R^2_{y.1}$ = .4536 และ $R^2_{y.12}$ = .5054 (จากตัวอย่างที่แล้ว ๆ มา) เราจะเห็นว่ามันเพิ่ม R^2 เท่ากับ .5054 — .4536 = .0518 เมื่อเราเพิ่มตัวแปร X_2 เข้าไปในสมการถดถอย การเพิ่มของ R^2 จะไม่มีนัยสำคัญ นั่นคือ X_2 ที่เพิ่มเข้าไปในสมการถดถอยจะไม่เพิ่มความถูกต้องของการทำนาย

ลองพิจารณาตัวอย่างอื่นต่อไปนี้บ้าง

ตัวอย่าง ในการศึกษาเกี่ยวกับตัวแปรทางการศึกษา 4 ตัว คือ Y, X_1, X_2 และ X_3 โดยมี Y เป็นตัวแปรตาม และ X_1, X_2, X, เป็นตัวแปรอิสระได้ข้อมูลมาดังนี้

Y	x1	X_2	x3	Y	X_1	X_2	X_3	Y	X_1	x2	x3
32	90	100	0	'35	95	101	0	44	100	91	0
30	90	104	0	38	95	93	0	46	100	93	0
35	90	102	0	37	95	91	0	47	100	96	0
33	90	104	0	40	95	89	0	47	100	91	0

35	90	96	0	41	95	88	0	47	100	94	0
34	90	96	1	37	95	97	1	43	100	93	1
29	90	110	1	41	95	91	1	47	100	91	1
32	90	105	1	36	95	96	1	45	100	91	1
36	90	109	1	35	95	101	1	48	100	89	1
34	90	102	1	40	95	91	1	47	100	91	1

ข้อมูลเหล่านี้สรุปได้เป็น

$\Sigma Y = 1170$, $\bar{Y} = 39$; $\Sigma X_1 = 2850$, $\bar{X}_1 = 95$

$\Sigma X_2 = 2880$, $\bar{X}_2 = 96$; $\Sigma X_3 = 15$, $\bar{X}_3 = 0.5$

$\Sigma Y^2 = 978$, $\Sigma X_1^2 = 500$, $\Sigma X_2^2 = 986$, $\Sigma X_3^2 = 7.50$

$\Sigma y x_1 = 650$, $\Sigma y X_2 = -778$, $\Sigma y X_3 = -1.0$

$\Sigma X_1 X_2 = -510$, $\Sigma X_1 X_3 = 0$, $\Sigma X_2 X_3 = 7.0$

ถ้าสมการถดถอยเป็น $Y = \beta_1 + \beta_2 X_2 + \varepsilon$ เราจะได้สมการถดถอยตัวอย่าง

เป็น $\hat{Y} = -84.50 + 1.30\, X_1$ $\quad R^2 = 0.8640$
$\quad\quad (9.27) \quad (0.097)$

Sob	df	SS	MS	F	
X_1	1	845	845	177.8947	$F^{(1,28)}_{.05} = 4.20$
Reoidual	28	133	4.75		
Total	29	978			

เราจะเห็นได้ว่า X_2 มีอิทธิพลต่อตัวแปรตาม Y

ถ้าเพิ่ม X_2 เข้าไป เราจะได้สมการถดถอยตัวอย่างเป็น

$\hat{Y} = -36.88 + 1.05\, X_1 - 0.25\, X_2$
(๑๙.๖๕) (0.๑3) (0.0๙) $\quad R^2 = .8926$

Sov	df	ss	Ms	P
x_1	1	845	845	

X_1, X_2	2	873	436.5	112.2109
เพิ่ม X_2	1	28	28	7.1979
Residual	27	105	3.89	
Total	29	978		

$$F_{.05}^{(2,27)} = 3.35; \quad F_{.05}^{(1,27)} = 4.21$$

เราจะเห็นว่าทั้ง X_4 และ X_2 มีอิทธิพลต่อ Y และการเพิ่ม X_2 เข้าไปนั้นมีผลต่อ Y จริง

เมื่อเพิ่ม X_3 เข้าไปอีกก็จะได้สมการถดถอยตัวอย่าง เป็น

$$\hat{Y} = \underset{(19.93)}{-36.64} + \underset{(0.13)}{1.05X_1} - \underset{(0.09)}{0.25X_2} + \underset{(0.74)}{0.10X_3} \qquad R^2 = .8937$$

SOV	df	SS	MS	F
X_1, X_2	2	873	436.5	
X_1, X_2, X_3	3	874	291.33	72.83
เพิ่ม X_3	1	1	1.00	0.25
Residual	26	104	4.00	
Total	29	978		

$$F_{.05}^{(3,26)} = 2.98; \quad F_{.05}^{(1,26)} = 4.23$$

เราจะเห็นว่า X_1, X_2, X_3 มีอิทธิพลต่อ Y แต่การเพิ่ม X^3 ไม่มีผลต่อ Y

ดังนั้นสมการถดถอยจึงมีตัวแปรอิสระเพียง X_1 และ X_2 เท่านั้น

ง. สมมติฐานเกี่ยวกับการเท่ากับสองของพารามิเตอร์ถดถอย ซึ่งมีสมมติฐานที่จะทดสอบสองดังนี้

$$Ho: \beta_i = \beta_j \qquad (i \neq j)$$

สำหรับสมมติฐานรอง Hu จะเป็นแบบสองทาง หรือทางเดียวก็ได้

ตัวสถิติทดสอบจะเป็น

$$T = (\hat{\beta}_i - \hat{\beta}_j) / S_{(\hat{\beta}_i - \hat{\beta}_j)}$$

$$S^2_{\hat{\beta}_i - \hat{\beta}_j} = S^2_{\hat{\beta}_i} - 2\, S_{\hat{\beta}_i \hat{\beta}_j}$$

จ. สมมติฐานเกี่ยวกับการเท่ากันระหว่างสัมประสิทธิการถดถอยที่ได้จากตัวอย่างต่างกัน (Test of Equality between Coefficient obtained from different samples)

ถ้าเรามี 2 ตัวอย่าง (n_1, n_2) เกี่ยวกับตัวแปรตามและตัวแปรอิสระ และเราใช้แต่ละตัวอย่างนี้ประมาณความสัมพันธ์ระหว่างตัวแปร ซึ่งเราจะได้เส้นถดถอยตัวอย่าง 2 เส้น สมมติว่าทั้งสองเส้นเป็นดังนี้

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k \quad \text{———} (1)$$

$$\hat{Y} = \hat{\delta}_0 + \hat{\delta}_1 X_1 + \hat{\delta}_2 X_2 + \cdots + \hat{\delta}_k X_k \quad \text{———} (2)$$

แล้วเราต้องการทดสอบว่าสองสมการนี้แตกต่างกันหรือไม่ นั่นคือทดสอบสมมติฐาน Ho

$$Ho: \beta_1 = \delta_1 \ , \ \beta_2 = \delta_2 \ , \ \cdots\cdots , \ \beta_k = \delta_k$$

ตัวสถิติทดสอบ จะได้จากวิธีประมาณค่าแบบกำลังสองต่ำสุดในข้อมูลชุดแรก, ในข้อมูลชุดที่สอง, และในข้อมูลชุดที่สามซึ่งได้จากสามข้อมูลชุดแรกและชุดหลังเข้าด้วยกัน ในการประมาณพารามิเตอร์ถดถอยจากข้อมูลรวม เราได้สมการถดถอยเป็น

$$\hat{Y} = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots\cdots + \alpha_k X_k$$

สำหรับความผันแปรเนื่องจากความคลาดเคลื่อน (SSE) จากตัวอย่างต่าง ๆ เราจะแทนด้วย SSE_1, SSE_2, SSE_c ดังนี้

$$SSE_1 = \sum_{i=1}^{n_1} (Y_i - \hat{Y}_i) \qquad \text{(จากตัวอย่างแรกขนาด } n_1\text{)}$$

$$SSE_2 = \sum_{i=1}^{n_2} (Y_i - \hat{Y}_i)^2 \qquad \text{(จากตัวอย่างหลังขนาด } n_2\text{)}$$

$$SSE_c = \sum_{i=1}^{n_1+n_2} (Y_i - \hat{Y}_i)^2 \qquad \text{(จากตัวอย่างรวมขนาด } n_1 + n_2\text{)}$$

แล้วเราสามารถแสดงได้ว่าตัวสถิติ F_c

$$F_c = \frac{SSE_c - (SSE_1 + SSE_2)/(k+1)}{(SSE_1 + SSE_2)/(n_1 + n_2 - 2k - 2)}$$

ซึ่งมีการแจกแจงแบบ Sredecor F, องศาความเป็นอิสระ $k + 1$ และ $(n_1 + n_2 - 2k - 2)$ ซึ่งจะใช้เป็นตัวสถิติทดสอบ Ho

ตัวสถิติทดสอบ F_c นี้จะใช้ได้ก็ต่อเมื่อ

— จำนวนค่าสังเกตในตัวอย่างหลังจะต้องมากกว่าจำนวนพารามิเตอร์ถดถอย นั่นคือ $n_2 > k+1$

— สมมติฐานที่ทดสอบจะเป็นแบบที่ว่า "พารามิเตอร์ถดถอย**ทั้งหมด**เปลี่ยนแปลงจากข้อมูลชุดหนึ่งไปยังข้อมูลอีกชุดหนึ่ง"

ฉ. สมมติฐานเกี่ยวกับการคงที่ของสัมประสิทธิ์การถดถอยเมื่อเพิ่มขนาดตัวอย่าง (Stability of Regression Coefficients when increasing the size of the sample) จุดมุ่งหมายของการทดสอบก็เพื่อจะดูว่าค่าของสัมประสิทธิ์ถดถอยจะคงที่หรือไม่เมื่อขนาดตัวอย่างเพิ่มขึ้น นั่นคือเราจะดูว่าค่าประมาณจะแตกต่างในตัวอย่างที่เพิ่มขึ้นหรือไม่ หรือว่าจะคงที่ตลอดเวลาหรือไม่ บางครั้งอาจจะมีเหตุการณ์เกิดขึ้นและทำให้เปลี่ยนโครงสร้างของความสัมพันธ์ได้ หรือทำให้พารามิเตอร์ถดถอยไม่คงที่ได้

ดังนั้นสมมติฐานที่จะทดสอบ คือ

$$Ho: \beta_1 = \alpha_1 \quad , \quad \beta_2 = \alpha_2 \quad , \ldots\ldots , \quad \beta_k = \alpha_k$$

ตัวสถิติทดสอบที่ใช้ขึ้นอยู่กับค่าสังเกตที่เพิ่มขึ้น (m) นั่นคือ

(1) ถ้าค่าสังเกตที่เพิ่มขึ้นมากกว่าจำนวนพารามิเตอร์ที่ต้องประมาณ นั่นคือ m k + 1 แล้วใช้ตัวสถิติทดสอบเช่นเดียวกับสมมติฐาน จ. ดังนี้

$$Fc = \frac{SSE_c - (SSE_1 + SS\ \ _2)/(k+1)}{SSE_1 + SSE_2)/(n + m - 2k - 2)}$$

(2) ถ้าค่าสังเกตที่เพิ่มขึ้นน้อยกว่าจำนวนพารามิเตอร์ หรือ m k + 1 เราใช้ตัวสถิติทดสอบ Fc ดังนี้

$$Fc = \frac{(SSE_c - SSE_1)/m}{SSE_1/(n-k-1)}$$

ซึ่งมีการแจกแจงแบบ Snedecor F, องศาความเป็นอิสระ m และ n−k−1