บทที่ 9. การวิเคราะห์เกี่ยวกับการถดถอย Regression Analysis

การวิเคราะห์เกี่ยวกับการถดถอย

Paul A. Samuelson

Economic Theory is mainly concerned with RELATIONS among VARIABLES. Demand and Supply relations, cost functions, production functions, and many others are familiar to every student who has taken a course in Economics. In fact, the entire body of Economic Theory can be regarded as a collection of Relations among Variables.

การวิเคราะห์เกี่ยวกับการถดถอยเป็นวิธีการสืบสวนทางสถิติ (Statistical Investigations) อย่างหนึ่งที่ใช้ในการทำนาย (Prediction) หรือการประมาณค่าของตัวแปรตัวหนึ่ง จากตัวแปรอื่น ๆ (หนึ่งตัวหรือมากกว่า) ซึ่งมีความสัมพันธ์กัน จากการวิเคราะห์เกี่ยวกับการถดถอยเราจะได้ตัวแบบทางคณิตศาสตร์ (Mathematical Model) ซึ่งแสดงถึงความสัมพันธ์ระหว่างตัวแปรที่กล่าวมาแล้ว ผลที่ได้จากการทำนาย และการประมาณค่าก็คำนวนได้จากตัวแปรแบบนี้

การถดถอย (Regression) เป็นศัพท์เทคนิคตัวหนึ่งในภาษาสถิติ ซึ่งมาจากผลงานของ Sir Francis Galton ที่พิมพ์ในปี ค.ศ.1885 เรื่อง Regression Toward Mediocrity in Heredity Stature ในวารสารชื่อ Journal of the Anthropological Institute ศึกษาถึงความสัมพันธ์ระหว่างส่วนสูงเฉลี่ยของลูก (Adult children) กับของพ่อแม่ เขาพบว่าลูก ๆ ซึ่งมีพ่อแม่ที่สูงกว่าส่วนเฉลี่ย จะไม่สูงเท่าพ่อแม่ของเขา และลูก ๆ ที่มีพ่อแม่เตี้ยจะต่ำกว่าส่วนเฉลี่ยแต่ไม่เตี้ยเท่าพ่อแม่ของเขา เขาจึงสรุปว่ามีแนวโน้มที่จะถดถอย (Tendency to "regress") ไปหาความสูงเฉลี่ยของประชากร เขาจึงเรียกชื่อเส้นที่อธิบายถึงส่วนเฉลี่ยของค่าตัวแปรหนึ่งว่า เส้นถดถอย (Line of regression) ในปัจจุบันนี้ การวิเคราะห์เกี่ยวกับการถดถอยจะเป็นเทคนิคที่ใช้สรุปความสัมพันธ์ (relationship) ระหว่างตัวแปรในเชิงตัวแบบทางคณิตศาสตร์

9.1 ความสัมพันธ์ระหว่างตัวแปร (Relations between Variables)

เราจะกำหนดความสัมพันธ์ระหว่างตัวแปร X และ Y ว่าเป็นเซ็ทของค่า X และ Y ทุก ๆ คุณลักษณะตามสมการที่กำหนดให้ ตัวอย่างเช่น ถ้าสมการที่กำหนดให้เป็น

$$Y = \beta_1 + \beta_2 X_i ; \beta_1, \beta_2 \quad \text{เป็นตัวคงที่}$$

แล้วความสัมพันธ์ระหว่าง X และ Y คือ เซ็ท (x,y) ซึ่งประกอบด้วยค่าเป็นไปได้ทั้งหมดของ

X และ Y ที่สอดคล้องกับสมการนั้น แบบของสมการมีชื่อตามความสัมพันธ์ เช่น สมการเส้นตรงจะอธิบายถึงความสัมพันธ์แบบเชิงเส้น (linear) สมการเอ็กซ์โพเนนเชียลจะอธิบายถึงความสัมพันธ์แบบเอ็กซ์โพเนนเชียลเป็นต้น

แนวความคิดของความสัมพันธ์เกี่ยวข้องกันมากกับความคิดในเรื่องโดเมน (domain) และพิสัย (Range) ถ้าความสัมพันธ์ระหว่าง X และ Y มีสมการเป็น

$$Y = f(X)$$

แล้วโดเมนของความสัมพันธ์คือ เซ็ทของค่าที่เป็นไปได้ทั้งหมดของ X และ พิสัยคือเซ็ทของค่าที่สมนัยที่เป็นไปได้ทั้งหมดของ Y

ความสัมพันธ์ระหว่างตัวแปรสามารถแบ่งออกได้เป็น 2 แบบ คือ ความสัมพันธ์แบบกำหนดได้ และความสัมพันธ์เชิงสุ่ม

(1) ความสัมพันธ์แบบกำหนดได้ (Deterministic or Nonstochastic Relation) สัมพันธ์ระหว่าง X และ Y จะเป็นแบบกำหนดได้ ถ้าสมาชิกแต่ละตัวของโดเมนจับคู่ได้กับสมาชิกเพียงตัวเดียวของพิสัย นั่นคือ ความสัมพันธ์ระหว่าง X และ Y ที่แสดงได้ด้วย

$$Y = f(X)$$

จะเป็นความสัมพันธ์แบบกำหนดได้ ถ้ากำหนดค่าหนึ่งของ X มาให้แล้วจะมีค่าของ Y ที่สมนัยเพียงตัวเดียวเท่านั้น อย่างไรก็ตามตัวแปร X และ Y ทั้งสองอาจจะเป็นแบบคงที่ (Nonstochastic) นั่นคือค่าของ X และ Y จะสามารถควบคุมหรือทำนายได้ หรือตัวแปรทั้งสองอาจจะไม่คงที่ (stochastic) ฉะนั้นจะเห็นได้ว่าความสัมพันธ์จะเป็นแบบกำหนดได้ ถึงแม้ว่าตัวแปรทั้งสองที่เกี่ยวข้องกัน **เป็นแบบไม่คงที่ อย่างไรก็ตามถ้าตัวแปรทั้งสองเป็นแบบไม่คงที่ ในเมื่อความสัมพันธ์เป็นแบบ** กำหนดได้นั้น ความน่าจะเป็นเงื่อนไขของ Y เมื่อกำหนด X ให้หรือ P(Y/X) จะรวมอยู่ที่จุดเดียว (Degenerate) หรือเป็น 1

(2) ความสัมพันธ์เชิงสุ่ม (Stochastic Relation) ความสัมพันธ์ระหว่าง X และ Y เป็นแบบสุ่ม ถ้าแต่ละค่าของ X มีการแจกแจงน่าจะเป็นของค่า Y เกิดขึ้น ดังนั้นสำหรับค่า X ใด ๆ ที่กำหนดให้จะมีตัวแปร Y ซึ่งเป็นค่าเฉพาะบางค่าหรือเป็นค่าที่อยู่ภายในช่วงด้วยความน่าจะเป็นค่าหนึ่งที่อยู่ระหว่าง 0 กับ 1

ความแตกต่างระหว่างความสัมพันธ์ของทั้งสองแบบแสดงให้เห็นได้ด้วยตัวอย่างดังต่อไปนี้สมมติเราทำการทดลองเพื่อที่จะพิจารณาปริมาณอุปสงค์ของสัม ณ ระดับราคาต่าง ๆ กัน ให้ Q_t เป็นจำนวนส้มที่ขายได้ (กิโล) ณ เวลา t และให้ P_t เป็นราคา (บาท) ส้มที่ขาย ณ ราคาที่กำหนดจะมีอุปสงค์ดังนี้

ราคา P_t	5	4	3	2	1
อุปสงค์ Q_t	1	3	5	7	9

จากผลที่ได้สามารถเขียนอยู่ในรูปสมการอุปสงค์ได้ดังนี้

$$Q_t = 11 - 2P_t$$

ความสัมพันธ์ระหว่างราคาและจำนวนอุปสงค์ในเวลาใด ๆ จะเป็นว่า "ถ้าส้มราคากิโลละ 5 บาท จะขายได้ 1 ก.ก. ถ้า 4 บาท จะขายได้ 3 ก.ก._____" ความสัมพันธ์แบบนี้จะเป็นความสัมพันธ์แบบกำหนดได้เพราะแต่ละราคามีจำนวนส้มที่ขายได้สอดคล้องเพียงจำนวนเดียว

แต่ถ้าผลที่ได้จากการทดลองนั้นเป็นดังนี้

ราคา P_t	จำนวนขาย Q_t	
5	ขายได้ 0 ก.ก. เป็น 25% ของจำนวนครั้งที่นำออกขาย ขายได้ 1 ก.ก. เป็น 50% ของจำนวนครั้งที่นำออกขาย ขายได้ 2 ก.ก. เป็น 25% ของจำนวนครั้งที่นำออกขาย	$E(Q_t) = 1$ $V(Q_t) = 0.5$
4	ขายได้ 2 ก.ก. เป็น 25% ของจำนวนครั้งที่นำออกขาย ขายได้ 3 ก.ก. เป็น 50% ของจำนวนครั้งที่นำออกขาย ขายได้ 4 ก.ก. เป็น 25% ของจำนวนครั้งที่นำออกขาย	$E(Q_t) = 3$ $V(Q_t) = 0.5$
⋮	⋮	⋮
1	ขายได้ 8 ก.ก. เป็น 25% ของจำนวนครั้งที่นำออกขาย ขายได้ 9 ก.ก. เป็น 50% ของจำนวนครั้งที่นำออกขาย ขายได้ 10 ก.ก. เป็น 25% ของจำนวนครั้งที่นำออกขาย	$E(Q_t) = 9$ $V(Q_t) = 0.5$

ดังนั้นสมการอุปสงค์จะเขียนได้เป็น

$$Q_t = 11 - 2P_t + \varepsilon_t$$

เมื่อ ε_t เป็นตัวแปรเชิงสุ่มที่มีการแจกแจงน่าจะเป็น (ไม่ว่าราคาที่กำหนดจะเป็นอะไร) ดังนี้

ε_t	-1	0	1
$P(\varepsilon_t)$	.025	0.50	0.25

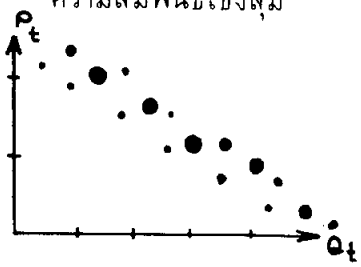
และมี $E(\varepsilon_t) = 0$ และ $V(\varepsilon_t) = 0.5$ ตัวแปรนี้มักเรียกว่า ตัวคลาดเคลื่อนเชิงสุ่ม (Random disturbance or Random Error Term) เพราะมันทำให้ความสัมพันธ์แบบกำหนดได้คลาดเคลื่อนนั่นเอง ความสัมพันธ์

$$Q_t = 11-2P_t + \varepsilon_t$$

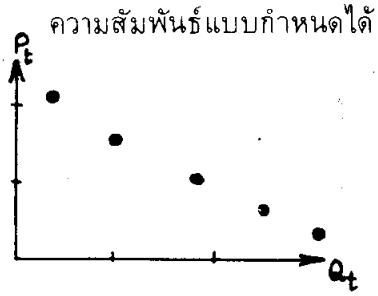
เป็นความสัมพันธ์เชิงสุ่มแบบหนึ่ง เนื่องจากมีตัวคลาดเคลื่อนทำให้ปริมาณอุปสงค์หลายค่า ณ แต่ละราคาปริมาณอุปสงค์นี้จะเกิดขึ้นด้วยความน่าจะเป็นอันหนึ่งที่กำหนดให้

ความสัมพันธ์ทั้งสองอย่างนี้แสดงให้เห็นด้วยรูปดังนี้

ความสัมพันธ์เชิงสุ่ม



ความสัมพันธ์แบบกำหนดได้



ลองพิจารณาปัญหาเกี่ยวกับความไม่เป็นอิสระ (dependence) ระหว่างตัวแปรสองตัวที่เกี่ยวข้องในเชิงความสัมพันธ์

ประการแรก พิจารณาความสัมพันธ์แบบกำหนดได้ที่แสดงด้วยฟังก์ชัน

$$Y = f(X)$$

ถ้า f(X) ไม่คงที่ตลอดทุก ๆ ค่าของ X นั่นคือ f(X) ไม่คงที่ตลอดสำหรับทุกค่าของโดเมนแล้ว เราจะพูดว่า Y ขึ้นอยู่กับ X (ในลักษณะฟังก์ชั่น) หรือ Y จะขึ้นอยู่กับ X ถ้าการเปลี่ยนแปลงของ X หมายถึงการเปลี่ยนแปลงของ Y ด้วย ในรูป 2 มิติที่ค่าของ Y วัดได้ตามแนวตั้ง และค่าของ X ตามแนวนอนนั้น Y จะขึ้นอยู่กับ X ถ้าทุกจุดอยู่ในเส้นตรงที่อยู่ในแนวนอนที่ไม่ขนานกับแกน X

ประการที่สอง Y ขึ้นอยู่กับ X (ในลักษณะฟังก์ชั่น) ถ้าการแจกแจงน่าจะเป็นของ Y ไม่เท่ากัน สำหรับทุก ๆ ค่าของ X หรือ Y ขึ้นอยู่กับ X จะเกิดขึ้นเมื่อค่าเฉลี่ยของ Y เปลี่ยนแปลงเมื่อ X มีค่าต่าง ๆ กัน อย่างไรก็ตาม Y จะขึ้นอยู่กับ X ถึงแม้ค่าเฉลี่ยของ Y จะคงที่สำหรับของ X แต่ลักษณะอื่น ๆ ของการแจกแจงของ Y จะต้องเปลี่ยนแปลงไปกับ X เช่น ถ้าความแปรปรวนของ Y เพิ่มขึ้นขณะที่เพิ่ม X อันนี้ ก็ทำให้ Y ขึ้นอยู่กับ X ได้

จากตัวอย่างเกี่ยวกับโค้งอุปสงค์ของสัมนั้น มีสิ่งที่น่าสนใจคือปริมาณอุปสงค์เปลี่ยนไปตามราคาแต่ความแปรปรวนไม่เปลี่ยนแปลง

P_t	5	4	3	2	1
$E(Q_t)$	1	3	5	7	9
$V(Q_t)$	.5	.5	.5	.5	.5

ทั้งนี้เพราะ ณ $P_t = 5$,

$$E(Q_t) = 0(.25) + 1(.05) + 2(.25) = 1$$

$$V(Q_t) = (0-1)^2(.25) + (1-1)^2(.50) + (2-1)^2(.25)$$

$$= 0.5$$

$P_t = 4$, $E(Q_t) = 3$

$V(Q_t) = 0.5$

กรณีทั่ว ๆ ไปของความไม่เป็นอิสระหรือความพึ่งพา (Dependence) นั้น ทั้งค่าเฉลี่ยและความแปรปรวนของ Y จะเปลี่ยนไปตามการเปลี่ยนแปลงของ X

ในทางเศรษฐศาสตร์ ความสัมพันธ์ต่าง ๆ มักจะกล่าวในรูปของความสัมพันธ์แบบกำหนดได้ ทั้งนี้ไม่ใช่ว่านักเศรษฐศาสตร์มีความเชื่อว่าความบังเอิญหรือโคลก (Chance) จะไม่เกิดขึ้นในความสัมพันธ์ทางเศรษฐกิจ แต่เขาพิจารณาว่าตัวคลาดเคลื่อนเชิงสุ่มนี้มีความสำคัญน้อยกว่าอิทธิพลเชิงระบบ (Systemetic INfluences) และการนำเอาตัวคลาดเคลื่อนเชิงสุ่มเข้าไปยุ่งในความสัมพันธ์ทางเศรษฐกิจ จะทำให้งานของนักทฤษฎีสลับซับซ้อนขึ้นอีก อย่างไรก็ตามการเน้นถึงการทดสอบทฤษฎีเศรษฐศาสตร์นั้นก็หมายถึงว่าเขามีความเชื่อว่าองค์ประกอบเชิงสุ่ม (Stochastic factors) มีอยู่จริง และถ้าความสัมพันธ์ตามทฤษฎีจริง ๆ เป็นแบบกำหนดได้แล้วปัญหาเกี่ยวกับการทดสอบสมมติฐานก็จะไม่เกิดขึ้น และเราจะวัดค่าของพารามิเตอร์ที่ไม่ทราบค่า ได้โดยการวัดที่มีประสิทธิภาพและความเที่ยงตรงมากกว่าที่เราจะพิจารณาถึงเรื่องการทดสอบ ตัวอย่าง เช่น เราพิจารณาทฤษฎีว่า Y ขึ้นอยู่กับ X แบบเชิงเส้น ถ้าความสัมพันธ์ระหว่าง Y กับ X เป็นแบบกำหนดได้ เราก็วัดค่าของ X และ Y เพียง 2 คู่เท่านั้น เราก็จะได้ค่าของพารามิเตอร์ได้ และถ้าเส้นตรงที่เชื่อมระหว่าง 2 จุดอยู่ในแนวนอน เราก็ปฏิเสธทฤษฎีและทฤษฎีนั้นควรจะได้รับการปรับปรุงแก้ไขให้ดีขึ้น แต่ถ้าไม่อยู่ในแนวนอนเราก็จะหาจุดตัดแกนและความชันของเส้นตรงได้โดยอาศัยกราฟ

อย่างไรก็ตามถ้าความสัมพันธ์ระหว่าง X และ Y เป็นแบบกำหนดไม่ได้หรือเป็นแบบสุ่ม การสังเกตเกี่ยวกับค่าของตัวแปรทั้งสองจะทำได้โดยใช้ตัวอย่าง และจากตัวอย่างนั้นสามารถใช้ทดสอบข้อเสนอเกี่ยวกับประชากรหรือประมาณความชันกับจุดตัดแกนได้อีกด้วย

9.2 ตัวแบบถดถอยเชิงเส้นแบบเชิงเดี่ยว (Simple Linear Regression model)

รูปแบบที่ง่ายที่สุดของความสัมพันธ์เชิงสุ่มระหว่างสองตัวแปร X และ Y จะเรียกว่า **ตัวแบบถดถอยเชิงเส้น** ตัวแบบนี้จะอยู่ในรูป

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

ในเมื่อ Y เป็นตัวแปรตาม (ซึ่งมีค่าเฉลี่ย Y เมื่อกำหนด X ให้) จะมีความสัมพันธ์เชิงเส้นกับ X นั่นคือ

$$E(Y_i) = \beta_1 + \beta_2 X_i$$

X เป็นตัวแปรอิสระหรือตัวแปรอธิบายได้ (Explanatory Variable) สำหรับ X หรือ Y สามารถสังเกตหรือวัดได้ β_1, β_2 เป็นพารามิเตอร์ของการถดถอย (Regression Parameters) ที่ไม่ทราบค่า

ε เป็นตัวคลาดเคลื่อน (Random disturbance) ซึ่งไม่สามารถวัดและสังเกตได้ และ

เป็นตัวแปรเชิงสุ่ม ที่เรายังไม่ทราบความจริงเกี่ยวกับขบวนที่ทำให้เกิดเทอมนี้ เหตุผลที่เราใส่เทอมนี้เข้าไปในสมการมีอยู่ 2 ประการคือ

(1) ตัวแปรที่ถูกตัดทิ้งไป (Omitted variables) ตามตัวแบบที่เราเห็นได้ว่าค่าของ Y ขึ้นอยู่กับค่าของ X เท่านั้น แต่ความจริงแล้ว Y อาจขึ้นอยู่กับตัวแปรอื่น ๆ ที่ไม่ปรากฏในตัวแบบ เราจึงใช้ ε แทนตัวแปรที่ถูกตัดทิ้งไป และตัวแปรเหล่านี้อาจจะเป็นทั้งประเภทที่วัดค่าได้หรือวัดค่าไม่ได้

(2) ความสัมพันธ์ที่แท้จริง (True Relationships) ตัวแบบนี้เป็นตัวแบบเชิงเส้นซึ่งความจริงแล้วเราไม่ทราบได้ว่าความสัมพันธ์ระหว่าง X และ Y อยู่ในลักษณะเชิงเส้นหรืออย่างอื่น ความสัมพันธ์ที่แท้จริงอาจจะอยู่ในรูปของแบบกำลังสอง (Quadratic) โปลีโนเมียล (Polynomial) หรือเอ็กซ์โพเนนเชียลก็ได้ ดังนั้นจึงใช้เป็นตัวแทนความคลาดเคลื่อนระหว่างตัวแบบที่ใช้กับความสัมพันธ์ที่แท้จริง

ตัวแบบถดถอยที่เป็นแบบสุ่มนั้น หมายถึง X แต่ละค่าจะมีประชากรย่อย (Subpopulation) ของ Y เกี่ยวข้องอยู่ด้วย นั่นคือ

X	Y
X_1	$Y_{11}, Y_{12}, Y_{13}, \cdots\cdots$
X_2	$Y_{21}, Y_{22}, Y_{23}, \cdots\cdots$
⋮	

หรือการแจกแจงของ X แต่ละค่าจะมีการแจกแจงน่าจะเป็นของค่า Y ซึ่งอาจจะเป็นแบบปกติหรือไม่ทราบว่าเป็นอะไร นั่นก็คือค่าของ Y ไม่สามารถพยากรณ์ได้ถูกต้อง ความไม่แน่นอนเกี่ยวกับ Y เกิดขึ้นเนื่องจากตัวคลาดเคลื่อน ε โดยที่ร่วมอยู่ในบางส่วนของ Y การแจกแจงของ Y และคุณลักษณะต่าง ๆ จะพิจารณาได้จากค่าของ X และการแจกแจงของ Y รูปแบบทางคณิตศาสตร์ของความสัมพันธ์ระหว่าง X และ Y ที่กำหนดขึ้นมานั้นเราสามารถประมาณค่าของพารามิเตอร์ได้ด้วยค่าสังเกต X และ Y ของตัวอย่าง จากการใช้ค่าประมาณของพารามิเตอร์เราก็จะประมาณค่าของตัวคลาดเคลื่อน ในแต่ละคู่ของ X และ Y ได้

เราจะเห็นได้ว่าข้อกำหนด (Specification) ของตัวแบบถดถอยไม่รวมแต่สมการถดถอยเท่านั้น ยังรวมข้อกำหนดของการแจกแจงน่าจะเป็นของ ε และข้อความที่บ่งว่าค่าของตัวแปรอธิบายได้ X พิจารณาได้อย่างไร ข้อกำหนดต่าง ๆ เหล่านี้เรียกว่า ข้อสมมติเบื้องต้น (Basic Assumptions) ข้อสมมติมีส่วนสำคัญเพราะคุณสมบัติของตัวประมาณค่าบางอย่างเช่น แบบกำลังสองต่ำสุด (Least square Estimator) ขึ้นอยู่กับข้อสมมติเหล่านั้น

ข้อสมมติเบื้องต้นเกี่ยวกับตัวแบบถดถอยเชิงเส้นชนิดเชิงเดี่ยวมีดังนี้

1. เป็นแบบปกติ (Normality) นั่นคือ ε มีการแจกแจงปกติ

2. ค่าเฉลี่ยเป็นศูนย์ (Zero Mean) นั่นคือ ε_i มีค่าเฉลี่ยเป็น 0 นั่นเอง หรือ $E(\varepsilon_i) = 0$

3. ความแปรปรวนเป็นเอกภาพ (Homoskedasticity) นั่นคือ ความแปรปรวนของ ε_i มีค่าคงที่ นั่นคือ $V(\varepsilon_i) = E(\varepsilon_i)^2 = \sigma^2$

ถ้าขาดคุณสมบัติข้อนี้เรียกว่า ความแปรปรวนไม่เป็นเอกภาพ (Heteroskedasticity)

4. ไม่มีความสัมพันธ์กัน (Nonautoregression) นั่นคือ ε_i และ ε_jของข้อมูลตัวที่ i และ j เป็นอิสระต่อกัน นั่นคือ

$$COV(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0; i, j = 1, 2, \ldots n, (i \neq j)$$

5. ตัวแปรอิสระกำหนดได้ (Nonstochastic X) นั่นคือ X_i เป็นตัวแปรแบบกำหนดได้ หรือแบบคงที่ ซึ่งมีค่าจำกัดในตัวอย่างซ้ำ ๆ และสำหรับขนาดตัวอย่างใดจะได้ว่า $\sum(X - \bar{X})^2/n$ เป็นจำนวนเลขนับได้ (finite number) ซึ่งค่าต่างจากศูนย์ ($\neq 0$)

ตัวแบบถดถอยเชิงเส้นชนิดเชิงเดี่ยวที่ประกอบด้วยข้อสมมติเบื้องต้น 5 ข้อที่กล่าวมานั้นเรียกกันว่าตัวแบบถดถอยเชิงเส้นปกติแบบ คลาสสิค (Classical Normal Linear Regression Model, CNLRM)

ข้อสมมติต่าง ๆ มีความหมายดังนี้ สองข้อแรกกล่าวว่า สำหรับแต่ละค่าของ X_i ตัวคลาดเคลื่อน ε_i จะแจกแจงปกติรอบ 0 และ ε_i เป็นตัวแปรเชิงสุ่มที่มีค่าจาก $-\infty$ ถึง $+\infty$ นั่นคือ ε_i แจกแจงสมมาตรรอบค่าเฉลี่ยของมัน การแจกแจงของ ε_i พิจารณาได้จากพารามิเตอร์ 2 ตัวคือ ค่าเฉลี่ยและความแปรปรวน

สำหรับข้อสามเกี่ยวกับแปรปรวนเอกภาพ ซึ่งหมายความว่าตัวคลาดเคลื่อน ε_i ทุกตัวมีความแปรปรวนเท่ากันแต่ไม่ทราบค่า ดังเช่นในเรื่องฟังก์ชั่นของการผลิต ข้อสมมตินี้หมายความว่าความผันแปรในผลผลิตจะเหมือนกันหมดถึงแม้ว่าจะมีจำนวนแรงงานต่างกัน

จากข้อ 1 ถึง 3 จึงเขียนได้ว่า $\varepsilon_i \sim N(0, \sigma^2)$

ข้อที่สี่ต้องการให้ตัวคลาดเคลื่อนไม่มีความสัมพันธ์กัน (autoregression) ซึ่งมีความหมายเหมือนกับว่า ผลผลิตสูงกว่าที่หวังไว้ในวันนี้จะไม่ทำให้สูง (หรือต่ำ) กว่าผลผลิตที่หวังไว้ในวันพรุ่งนี้ข้อ 3 และ 4 รวมกันจะหมายถึงว่าตัวคลาดเคลื่อนไม่มีสหสัมพันธ์กัน

ข้อสมมติสุดท้ายที่กล่าวว่าตัวแปรอธิบายได้เป็นแบบคงที่ นั่นคือ ค่าของ X_i สามารถควบคุมหรือทำนายได้และข้อความที่กล่าวว่าค่าของ X_i คงที่ในตัวอย่างซ้ำ ๆ นั้น ชี้ว่าเซ็ทของค่า X จะเป็นเช่นเดียวกันในตัวอย่าง ต่าง ๆ กัน และที่กล่าวว่า $\sum(X-\bar{X})^2/n$ เป็นจำนวนนับได้ที่ต่างจากศูนย์ นั้นหมายความว่าค่าของ X ในตัวอย่างหนึ่งต้องไม่เท่ากันหมด และค่านั้นจะไม่เพิ่มขึ้นหรือลดลงโดยปราศจากขีดจำกัดในเมื่อขนาดตัวอย่างเพิ่มขึ้น

ข้อสมมติใน CNLRM นี้ใช้สำหรับหาตัวประมาณของพารามิเตอร์เกี่ยวกับการถดถอย

โดยที่ ε_i มีการแจกแจงปกติซึ่งมีค่าเฉลี่ยเป็นศูนย์แต่ไม่ทราบค่า σ^2 หรือ $\varepsilon_i \sim N(0, \sigma^2)$ ดังนั้นตัวแบบที่บรรยายด้วยสมการ $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$ กับข้อสมมติ 5 ข้อจะมีพารามิเตอร์ที่ไม่ทราบค่า 3 เทอมคือ β_1, β_2 และความแปรปรวนของ ε หรือ σ^2

นอกจากศึกษาถึงข้อกำหนดของตัวแบบถดถอยที่บรรยายด้วยสมการถดถอยและข้อสมมติเบื้องต้นแล้วเรายังสนใจลักษณะบางอย่างของตัวแบบอีก โดยเฉพาะสนใจการแจกแจงน่าจะเป็นของตัวแปรตาม Y_i

(1) ค่าเฉลี่ยของ Y_i จะหาได้จากสมการถดถอย

$$E(Y_i) = E(\beta_1 + \beta_2 X_i + \varepsilon_i) = \beta_1 + \beta_2 X_i$$

ตามข้อกำหนด β_1, β_2 เป็นพารามิเตอร์ X_i เป็นแบบคงที่ และค่าเฉลี่ยของ ε_i เป็นศูนย์ ดังนั้นตัวแบบถดถอยจึงเขียนใหม่ได้เป็น $Y_i = E(Y_i) + \varepsilon_i$

(2) ความแปรปรวนของ Y_i จะหาได้ดังนี้

$$\begin{aligned} V(Y_i) &= E((Y_i) - E(Y_i))^2 = E((\beta_1 + \beta_2 X_i + \varepsilon_i) - (\beta_1 + \beta_2 X_1))^2 \\ &= E(\varepsilon_i^2) = \sigma^2 \\ V(Y_i) &= V(\varepsilon_i) = \sigma^2 \end{aligned}$$

ดังนั้น $V(Y_i) = V(\varepsilon_i) = \sigma^2$

จากสมการถดถอยจะเห็นได้ว่า Y_i เป็นฟังก์ชันเชิงเส้นของ ε_i โดยมี ε_i แจกแจงแบบปกติ ตามทฤษฎีเกี่ยวกับตัวแบบเชิงสุ่มแบบปกติเราจะได้ว่า Y_i ก็มีการแจกแจงปกติที่มีค่าเฉลี่ย $(\beta_1 + \beta_2 X_i)$ และความแปรปรวน σ^2 ด้วย นั่นคือ

$$Y_i \sim N(\beta_1 + \beta_2 X_i, \sigma^2)$$

เมื่อแสดงด้วยกราฟ จะเห็นได้ว่าค่าเฉลี่ยของการแจกแจงทั้งหมดอยู่บนเส้นตรง และแต่ละการแจกแจงจะมีความแปรปรวนเท่ากันหมด

จาก $E(Y_i) = \beta_1 + \beta_2 X_i$ ซึ่งให้ค่าเฉลี่ยของ Y สำหรับแต่ละค่าของ X สมการนี้ได้ชื่อว่า **เส้นถดถอยประชากร** (Population Regression Line) ซึ่งเขียนใหม่ได้เป็น $Y_i = \beta_1 + \beta_2 X_i$

โดยที่ β_1 เป็นจุดตัดแกนของเส้นซึ่งจะวัดค่าเฉลี่ยของ Y ที่สมนัยกับค่า X ที่เป็นศูนย์ ความชันของเส้น β_2 จะวัดการเปลี่ยนแปลงในค่าเฉลี่ยของ Y ที่สมนัยกับการเปลี่ยนแปลงหนึ่งหนึ่งในค่าของ X หรือ

$$\beta_2 = \Delta Y / \Delta X = dy/dx$$

เช่นถ้า Y แทนการบริโภครวม และ X เป็นรายได้รวม แล้ว β_1 จะวัดระดับยังชีพของการบริโภค

และ β_2 จะวัดความโน้มเอียงที่จะบริโภคเพิ่ม (MPC)

โดยที่ค่าของพารามิเตอร์เหล่านี้ไม่ทราบ เส้นถดถอยประชากรจึงไม่ทราบด้วย เมื่อประมาณค่า β_1, β_2 ด้วยตัวอย่าง เราจะได้**เส้นถดถอยตัวอย่าง** (Sample Regression Line) ซึ่งทำหน้าที่เป็นค่าประมาณของเส้นถดถอยประชากร ถ้า $\hat{\beta}_1$ และ $\hat{\beta}_2$ เป็นตัวประมาณค่าของ β_1 และ β_2 แล้ว เส้นถดถอยตัวอย่างจะเป็นดังนี้

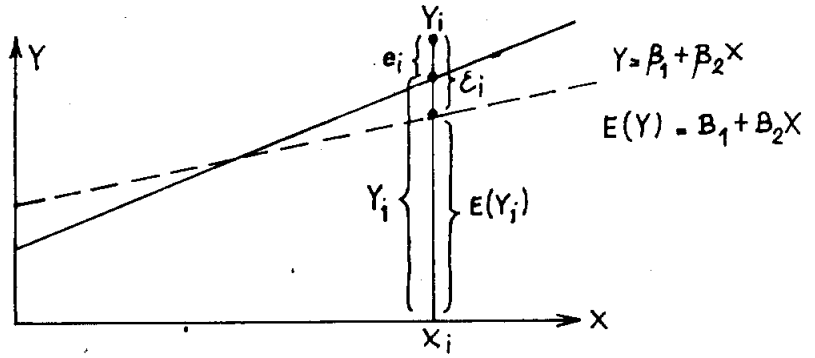
$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

เมื่อ $\hat{Y}_i$ เป็นค่าปรับ (fitted valve) ของ Y_i ค่าสังเกตของ Y_i ส่วนมากจะไม่อยู่บนเส้นถดถอยตัวอย่างทีเดียว ดังนั้นค่าของ Y_i และ $\hat{Y}_i$ จึงต่างกัน ความแตกต่างนี้เรียกว่า ตัวเศษ (Residual) ซึ่งจะแทนตัว e_i ดังนั้นเราจะได้ 2 สมการซึ่งต่างกันดังนี้

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i \text{(ประชากร)}$$

และ $$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i \text{(ตัวอย่าง)}$$

โดยทั่ว ๆ ไป e_i ต่างจาก ε_i เพราะว่า $\hat{\beta}_1$ และ $\hat{\beta}_2$ ต่างจากค่าจริงของ β_1 และ β_2 ความจริงเราสามารถแสดงให้เห็นได้ว่า ตัวเศษ e_i เป็นค่าประมาณของตัวคลาดเคลื่อน ε_i (หรือกล่าวได้ว่าการแจกแจงของ e_i จะใช้ประมาณพารามิเตอร์ของการแจกแจงของ ε_i) ดังรูปต่อไปนี้



9.3 การประมาณค่าของพารามิเตอร์เกี่ยวกับการถดถอย

การประมาณค่าของพารามิเตอร์ในตัวแบบถดถอยเชิงเส้นก็เหมือนกับการประมาณค่าพารามิเตอร์ในการแจกแจงน่าจะเป็นของตัวแปรตาม Y_i โดยข้อสมมติของตัวแบบเราได้ว่า

$$Y_i \sim N(\beta_1 + \beta_2 X_i, \sigma^2)$$

ดังนั้นการประมาณค่าของพารามิเตอร์ถดถอย β_1, β_2 จึงเหมือนกับการประมาณค่าเฉลี่ยของ Y_i การประมาณทำได้หลายวิธี คือ

(1) วิธีกำลังสองต่ำสุด (Least Square Method)

(2) วิธี BLUE (Best Linear Unbiased Estimation Method)

(3) วิธีน่าจะเป็นสูงสุด (Maximum Likelihood Method)

(4) วิธีกึ่งเฉลี่ย (Mehod of SemiAverage)

(5) วิธีเฉลี่ยเคลื่อนที่ (Moving Average Method)

เราจะกล่าวสัก 4 วิธี และจะเปรียบเทียบตัวประมาณและคุณสมบัติของมัน ทั้งนี้เพื่อต้องการหาตัวประมาณที่มีคุณสมบัติที่พึงประสงค์มากที่สุด ซึ่งตัวประมาณเช่นนั้นจะใช้ในการทดสอบสมมติฐานเกี่ยวกับตัวแบบถดถอยและทำการพยากรณ์อีกด้วย

9.3.1 วิธีกำลังสองต่ำสุด (Least Square Method)

หลักการประมาณแบบกำลังสองต่ำสุดคือการทำให้ผลรวมของความเบี่ยงเบนกำลังสอง (Squared Deviation) ของค่าสังเกตุที่ห่างจากค่าเฉลี่ยของมันให้มีค่าน้อยที่สุด นั่นคือเราต้องหาค่าของเฉลี่ยที่ทำให้ผลรวมตามที่ต้องการมีค่าน้อยเท่าที่เป็นไปได้ ในกรณีของเราต้องทำให้ผลรวมของความเบี่ยงเบนกำลังสอง (So) มีค่าน้อยที่สุด ผลรวม So กำหนดได้ดังนี้

$$S_o = \sum_1^n (Y_i - E(Y_i))^2$$
$$= \sum (Y_i - (\beta_1 + \beta_2 X_i))^2 \quad = \sum \varepsilon_i^2$$

การหาค่าของ β_1, β_2 ที่จะทำให้ So มีค่าน้อยที่สุดนั้นทำได้โดยการหาอนุพันธ์ So เทียบกับ β_1 และ β_2 ดังนี้

$$\delta S_o / \delta \beta_1 = \delta / \delta \beta_1 \sum (Y_i - \beta_1 + \beta_2 X_i))^2$$
$$= 2 \sum (Y_i - \beta_1 - \beta_2 X_i)(-1)$$

และ $$/\delta S_o / \delta \beta_2 = 2 \sum (Y_i - \beta_1 - \beta_2 X_i)(-X_i)$$

แล้วให้ $\frac{\delta So}{\delta \beta_1} = o$ และ $\frac{\delta So}{\delta \beta_2} = o$ แล้วใส่หมวก (^) บน β_1 และ β_2 เป็น $\hat{\beta}_1$ และ $\hat{\beta}_2$ เพื่อชี้ว่าผลที่ได้จากสมการนี้เป็นตัวประมาณแบบกำลังสองต่ำสุดของ β_1 และ β_2 เราจะได้

$$\sum 2 (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)(-1) = 0$$
$$\sum 2 (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)(-X_i) = 0$$

จากสองสมการนี้เราได้สมการที่มีชื่อว่า สมการปกติแบบกำลังสองต่ำสุด (Least Square Normal Equations) ดังนี้

$$\sum Y_i = \hat{\beta}_1 n + \hat{\beta}_2 \sum X_i$$
$$\sum X_i Y_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2$$

สำหรับ $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$ เมื่อ e_i แทนตัวเศษแบบกำลังสองต่ำสุด (Least Square

Residuals) เราสามารถเขียนสมการปกติได้ง่าย ๆ ดังนี้

$$\sum e_i = 0$$
$$\sum X_i e_i = 0$$

จากสม การปกติเราแก้สมการหาค่า $\hat{\beta}_1$ และ $\hat{\beta}_2$ ได้ดังนี้

$$\hat{\beta}_2 = \frac{n\sum(XY) - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$$

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

หรือในรูปอื่น ๆ ดังนี้

$$\hat{\beta}_2 = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}$$

$$= \sum xy / \sum x^2 \qquad x = X-\bar{X},\ y = Y-\bar{Y}$$

$$\hat{\beta}_1 = \sum Y/n - \hat{\beta}_2 \sum X/n$$

จาก $\hat{\beta}_1 = \bar{Y} - \beta_2\bar{X}$ หรือ $\bar{Y} = \hat{\beta}_1 + \beta_2\ \bar{x}$ ซึ่งหมายความว่าเส้นถดถอยตัวอย่าง $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ ผ่านจุด $(\bar{X}, \bar{Y})$ และค่าของ $\hat{\beta}_1$ และ $\hat{\beta}_2$ จะวัดจุดตัดแกนและความชันของเส้นถดถอยตัวอย่างตามลำดับ

ตัวอย่าง ราคาและปริมาณอุปสงค์ของส้มชั้นดีที่ขายในตลาดเวียงจันทร์ในเวลา 12 วัน ติดต่อกันเป็นดังนี้

ราคาต่อกิโล (กีบ)	100	90	80	70	70	70	70	65	60	60	55	50
ปริมาณอุปสงค์ (กิโล)	55	70	90	100	90	105	80	110	125	115	130	130

ให้ X_i เป็นราคาที่ขาย และ Y_i เป็นปริมาณที่ขายได้ในวันที่ i สำหรับฟังก์ชั่นอุปสงค์ ให้อยู่ในรูป

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

และให้ข้อสมมติของ CNLRM สอดคล้องด้วย แล้วเราจะได้ค่าประมาณแบบกำลังสองต่ำสุดของ β_1 และ β_2 ดังนี้

$$\hat{\beta}_2 = -3550/2250 = -1.578$$
$$\hat{\beta}_1 = 100-(-1.578)(70) = 210.460$$

ในเมื่อ $\bar{X} = 70$ $\quad \sum X^2 = 80450$ $\quad \sum xy = -3350$

$\bar{Y} = 100$ $\quad \sum Y^2 = 126300$ $\quad \sum x^2 = 2250$

$\sum X = 840 \quad \sum XY = 80450 \quad \sum y^2 = 6300$

$\sum Y = 1200$

ดังนั้นเส้นถดถอยตัวอย่างที่ประมาณได้เป็น

$$\hat{Y}_i = 210.460 + (-1.578)(X_i)$$

โดยที่ฟังก์ชั่นเป็นแบบเชิงเส้น ความยืดหยุ่น (ราคา) ของอุปสงค์ η จะแตกต่างกัน ณ ที่ราคาต่าง ๆ นั่นคือ

$$\eta_X = \hat{\beta}_2 (X_i / Y_i)$$

สำหรับที่ราคาเฉลี่ย ($X = 70$) แล้ว $\eta_{\bar{X}}$ ประมาณได้เป็น

$$\eta_{\bar{X}} = (-1.578)(70/100) = -1.05$$

9.3.2 วิธีประมาณค่าแบบ BLUE (Best Linear Unbiased Estimation)

วิธีประมาณค่าแบบ BLUE นี้ต้องการตัวประมาณในรูปผสมเชิงเส้น (Linear Combation) ของค่าสังเกตจากตัวอย่างที่เป็นตัวประมาณที่ไม่เอียงเฉ มีความแปรปรวนน้อยกว่าตัวประมาณค่าไม่เอียงเฉแบบเชิงเส้นตัวอื่น ๆ เราจะใช้วิธีนี้หาตัวประมาณของ β_2 แบบ BLUE ซึ่งจะให้เป็น $\tilde{\beta}_2$

จากเงื่อนไขของความเป็นเชิงเส้น (Linearity) แล้ว $\tilde{\beta}_2$ จะอยู่ในรูป

$$\tilde{\beta}_2 = \sum a_i Y_i$$

ในเมื่อ a_i ($i = 1, 2, \ldots n$) เป็นตัวคงที่ซึ่งจะต้องพิจารณา

$$\begin{aligned} E(\tilde{\beta}_2) &= E(\sum a_i Y_i) = \sum a_i E(Y_i) \\ &= \sum a_i (\beta_1 + \beta_2 X_i) \\ &= \beta_1 \sum a_i + \beta_2 \sum a_i X_i \end{aligned}$$

ถ้า $\tilde{\beta}_2$ ไม่เอียงเฉเราต้องให้ $\sum a_i = 0$ และ $\sum a_i X_i = 1$ และต้องการให้ $\tilde{\beta}_2$ มีความแปรปรวนน้อยกว่าตัวประมาณค่าอื่น ๆ ที่สอดคล้องกับเงื่อนไขข้างต้นนี้

$$V(\tilde{\beta}_2) = V(\sum a_i Y_i) = \sigma^2 \sum a_i^2$$

เราต้องหาค่า $a_1, a_2, \ldots, a_n$ ที่มีคุณสมบัติว่า $\sum a_i = 0$, $\sum a_i X_i = 1$ และ $\sigma^2 \sum a^2$ มีค่าน้อยที่สุดเท่าที่เป็นไปได้ นั่นคือ ทำให้ $\sigma^2 \sum a_i^2$ มีค่าน้อยที่สุดโดยมีเงื่อนไขว่า $\sum a_i = 0$ และ $\sum a_i X_i = 1$ ปัญหาเช่นนี้เป็นเรื่องของการหาค่าต่ำสุดโดยมีข้อจำกัด ซึ่งแก้ปัญหาได้ด้วยวิธี Lagrange

Multiplier ดังนี้ เราสร้างฟังก์ชั่นใหม่เป็น

$$H = \sigma^2 \sum a_i^2 - \lambda_1 \sum a_i - \lambda_2 (\sum a_i X_i - 1)$$

ซึ่งเป็นฟังก์ชั่นที่ต้องทำให้น้อยที่สุดโดยมี λ_1 และ λ_2 เป็น Lagrange Multiplier แก้สมการนี้โดยหาอนุพันธ์ของ H เทียบกับ $a_1, a_2, \ldots, a_n$ และ λ_1, λ_2 แล้วให้แต่ละเทอมเท่ากับศูนย์ นั่นคือ

$$\frac{\delta H}{\delta a_1} = 0, \quad \frac{\delta H}{\delta a_2} = 0, \ \ldots, \quad \frac{\delta H}{\delta a_n} = 0$$

$$\frac{\delta H}{\delta \lambda_1} = 0, \quad \frac{\delta H}{\delta \lambda_2} = 0 .$$

จะได้ $$a_i = \frac{- \sum X_i + n X_i}{n \sum X_i^2 - (\sum X_i)^2} \quad , i = 1,2,\ldots n$$

a_i (i = 1, 2,..., n) เป็นตัวคงที่ซึ่งทำให้ β_2 เป็นตัวประมาณค่าไม่เอียงเฉ และทำให้มีความแปรปรวนน้อยที่สุด (วิธีทดสอบว่ามีความแปรปรวนน้อยที่สุดก็โดยอาศัยวิธีทางแคลคูลัส) เมื่อแทนค่า a_i ลงใน

$$\tilde{\beta}_2 = \sum a_i Y_i \qquad \text{แล้วเราจะได้}$$

$$\tilde{\beta}_2 = \frac{- \sum X_i + n X_i}{n \sum X_i^2 - (\sum X_i)^2} Y_i$$

$$= \frac{n \sum XY - (\sum X)(\sum Y)}{n \sum X^2 - (\sum X)^2}$$

นั่นคือ $$\tilde{\beta}_2 = \hat{\beta}_2$$

วิธีของ BLUE ไม่เพียงแต่จะหาตัวประมาณได้เท่านั้น ยังสามารถหาความแปรปรวนของตัวประมาณได้อีกด้วย

สำหรับ $V(\tilde{\beta}_2)$ เราจะหาได้จาก $V(\tilde{\beta}_2) = \sigma^2 \sum a_i^2$

$$V(\tilde{\beta}_2) = \sigma^2 \sum a_i^2$$

โดยที่ $$\sum a_i^2 = \sum a_i a_i = \sum a_i \left\{ \frac{- \sum X_i + n X_i}{n \sum X_i^2 - (\sum X_i)^2} \right\}$$

$$= \frac{- \sum X_i \sum a_i + n \sum a_i X_i}{n \sum X_i^2 - (\sum X_i)^2}$$

และ $\sum a_i = 0$ และ $\sum a_i X_i = 1$ เราจะได้

$$\sum a_i^2 = n/\{n\sum X_i^2 - (\sum X_i)^2\} = 1/\sum(X-\bar{X})^2$$

นั่นคือ $$V(\tilde{\beta}_2) = \sigma^2/\sum(X-\bar{X})^2 = V(\hat{\beta}_2)$$

ซึ่งแสดงว่าความแปรปรวนจากวิธี BLUE และ LSE จะเท่ากัน

ในทำนองเดียวกันเราสามารถหา BLUE ของ $\tilde{\beta}_1$ และความแปรปรวนของ $\tilde{\beta}_1$ หรือ $V(\tilde{\beta}_1)$ ได้ดังนี้

$$\tilde{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X}$$
$$V(\tilde{B}_1) = \sigma^2(1/n + (\bar{X}^2)/\sum(X-\bar{X})^2)$$
$$= \sigma^2/n + \bar{X}^2V(\tilde{\beta}_2) = V(\hat{\beta}_1)$$

ซึ่งจะเห็นได้ว่า $\tilde{\beta}_1$ และ $\hat{\beta}_1$ จะเหมือนกันอีก

9.3.3 วิธีประมาณค่าแบบน่าจะเป็นสูงสุด (MLE, Maximum Likelihood Estimation)

ตัวประมาณค่าแบบน่าจะเป็นสูงสุดของพารามิเตอร์ในประชากรที่กำหนดเราจะพิจารณาจากค่าของพารามิเตอร์ที่จะให้ (Generate) ค่าสังเกตของตัวอย่างบ่อยที่สุด การหาตัวประมาณแบบนี้เราต้องพิจารณาฟังก์ชั่นน่าจะเป็น (Likeli-hood Function) ของค่าสังเกตในตัวอย่าง และแล้วจะทำให้ฟังก์ชั่นนี้โดยเทียบกับพารามิเตอร์ไม่ทราบค่านั้นมีค่ามากที่สุด ในกรณีตัวแบบถดถอยตัวอย่างประกอบด้วยค่าสังเกตของตัวแปร Y จำนวน n ตัว คือ $Y_1, Y_2, \ldots, Y_n$ ตัวแปรเหล่านี้แจกแจงแบบปกติที่มีค่าเฉลี่ย $\beta_1 + \beta_2X_1, \beta_1 + \beta_2X_2, \ldots, \beta_1 + \beta_2X_n$ และมีความแปรปรวนเท่ากันคือ σ^2 นั่นคือ

$$Y_i \sim N(\beta_1 + \beta_2X_i; \sigma^2)$$

และเมื่อให้ค่าสังเกตเหล่านี้เป็น $y_1, y_2, \ldots, y_n$ แล้วฟังก์ชั่นน่าจะเป็น L คือ

$$L = f(y_1, y_2, \ldots, y_n)$$
$$= f(y_1)\,f(y_2)\ldots f(y_n)$$

โดยที่ $y_1, y_2, \ldots, y_n$ เป็นอิสระกัน

พารามิเตอร์ที่จะทำให้ L มีค่ามากสุดก็เช่นเดียวกับพารามิเตอร์ที่จะทำให้ log L มีค่ามากที่สุดด้วย ดังนั้นเราจะต้องทำให้

$$\log L = \log\{f(y_1)\,f(y_2)\ldots f(y_n)\}$$

มีค่ามากที่สุด แต่ $Y_i \sim N(\beta_1 + \beta_2X_i; \sigma^2)$ เราจึงได้

$$\log f(y_i) = (1/2)\log(2\pi\sigma^2) - (1/2)(y_i-\beta_1-\beta_2 x_i)^2/\sigma^2$$

นั่นคือ $$\log L = -(n/2)\log(2\pi) - (n/2)\log\sigma^2 - (1/2\sigma^2)\sum(y_i-\beta_1-\beta_2 x_i)^2$$

สมการนี้มีพารามิเตอร์ไม่ทราบค่า 3 ตัวคือ β_1, β_2, σ^2 เราจะหาค่าของพารามิเตอร์นี้ ซึ่งค่าเหล่านี้จะทำให้ log L มีค่ามากสุดได้เมื่ออนุพันธ์ของ log L เทียบกับพารามิเตอร์แต่ละตัวจะเป็นศูนย์

ดังนั้น $\partial \log L/\partial \beta_1 = 0$, $\partial \log L/\partial \beta_2 = 0$

และ $\partial \log L/\partial\sigma^2 = 0$

จากสองสมการแรกเราได้สมการปกติแบบน่าจะเป็นมากสุด ดังนี้

$$\sum Y_i = \hat{\beta}_1 n + \hat{\beta}_2 \sum X_i$$
$$\sum X_i Y_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2$$

ซึ่งเป็นเช่นเดียวกับสมการปกติแบบกำลังสองต่ำสุดที่กล่าวมาแล้ว นั่นก็คือ **MLE's**ของ β_1, β_2 ได้เป็นเช่นเดียวกับ LSE นั่นคือ

$$\hat{\beta}_1 = \hat{\beta}_1 = \bar{Y} - \beta_1\bar{X}$$
$$\hat{\beta}_2 = \hat{\beta}_2 = \frac{n\sum XY - \sum X\sum Y}{n\sum X^2 - (\sum X)^2}$$

จากสมการที่สามเราจะได้ MLE**ของ** σ^2 นั่นคือ

$$\hat{\sigma}^2 = (1/n)\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

$$\hat{\sigma}^2 = (1/n)\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$
$$= (1/n)\sum e_i^2$$

จะเห็นได้ว่า MLE ของความแปรปรวนของตัวคลาดเคลื่อนเท่ากับความแปรปรวนจากตัวอย่างตัวเศษ แบบกำลังสองต่ำสุด (Least Square Residuals)

จากวิธีทั้งสามที่กล่าวมาพอสรุปผลได้ว่าค่าประมาณของพารามิเตอร์เกี่ยวกับการถดถอยเป็นเช่นเดียวกัน หรือกล่าวได้ว่าภายใต้ข้อสมมติของ CNLRM ตัวประมาณค่าพารามิเตอร์ของการถดถอย โดยวิธี LSE, BLUE และ MLE เป็นแบบเดียวกัน อย่างไรก็ตามวิธี LSE ให้แต่เพียงสูตรของตัวประมาณ β_1 และ β_2 สำหรับวิธี BLUE จะให้สูตรของตัวประมาณของ β_1, β_2 และความแปรปรวนของมันอีกด้วย ส่วนวิธี MLE จะให้สูตรสำหรับตัวประมาณของ β_1,β_2 และ σ^2

9.3.4 การประมาณด้วยวิธีกึ่งเฉลี่ย Method of Semi Average

ในทางปฏิบัติก็ใช้วิธีนี้กัน เราจะลองพิจารณาวิธีนี้กับวิธีกำลังสองต่ำสุดดู วิธีกึ่งเฉลี่ยต้องเรียงลำดับค่าสังเกตุของ n ตามขนาด แล้วแบ่งค่าสังเกตุเป็น 2 ส่วนเท่า ๆ กัน (หรือเกือบเท่ากัน) แล้วหาค่าเฉลี่ยของ X และ Y ตามที่แยกไว้ นั่นคือเราจะได้จุด 2 จุด (ส่วนละจุดนั่นเอง) เส้นถดถอยที่ประมาณได้ก็คือเส้นที่ลากผ่าน 2 จุดนี้

เพื่อความสะดวกเรากำหนดให้จำนวนค่าสังเกตเป็นจำนวนคู่ให้ $\bar{X}_1$ และ $\bar{Y}_1$ เป็นค่าเฉลี่ยของตัวอย่างของค่า X และ Y ส่วนแรก กับ $\bar{X}_2$ และ $\bar{Y}_2$ เป็นค่าเฉลี่ยในส่วนที่สอง

ให้ b_1, b_2 เป็นตัวประมาณแบบกึ่งเฉลี่ยของ β_1 และ β_2 แล้ว b_1, b_2 คำนวณได้จาก

$$\bar{Y}_1 = b_1 + b_2\bar{X}_1$$

และ $$\bar{Y}_2 = b_1 + b_2\bar{X}_2$$

ทั้งนี้เนื่องจากเส้นถดถอยที่ประมาณจะผ่านจุด $(\bar{X}_1, \bar{Y}_1)$ และ $(\bar{X}_2, \bar{Y}_2)$

ดังนั้น $$b_1 = \bar{Y}_1 - b_2\bar{X}_1$$

$$b_2 = (\bar{Y}_1 - \bar{Y}_2)/(\bar{X}_1 - \bar{X}_2)$$

ค่าคาดหวังและความแปรปรวนของ b_2 หาได้ดังนี้

จากตัวแบบ $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$

เราจะได้ $$\bar{Y}_1 = \beta_1 + \beta_2\bar{X}_1 + \bar{\varepsilon}_1$$

$$\bar{Y}_2 = \beta_1 + \beta_2\bar{X}_2 + \bar{\varepsilon}_2$$

ในเมื่อ $\bar{\varepsilon}_1, \bar{\varepsilon}_2$ เป็นค่าเฉลี่ยของตัวคลาดเคลื่อนในค่าสังเกตส่วนแรกและหลัง

ดังนั้น $$\bar{Y}_1 - \bar{Y}_2 = (\bar{X}_1 - \bar{X}_2)\beta_2 + (\bar{\varepsilon}_1 - \bar{\varepsilon}_2)$$

เราจะได้ $$E(b_2) = E(\bar{Y}_1 - \bar{Y}_2)/(\bar{X}_1 - \bar{X}_2)$$

$$= E\left\{\frac{\beta_2(\bar{X}_1 - \bar{X}_2) + (\bar{\varepsilon}_1 - \bar{\varepsilon}_2)}{\bar{X}_1 - \bar{X}_2}\right\} = \beta_2$$

นั่นคือ b_2 เป็นตัวประมาณแบบกึ่งเฉลี่ยที่ไม่เอียงเฉของ β_2

$$V(b_2) = E(b_2 - E(b_2))^2 = 4\sigma^2/n(\bar{X}_1 - \bar{X}_2)^2$$

ซึ่งจะเห็นได้ว่า $V(b_2) \geq V(\hat{\beta}_2)$

ตัวอย่าง วิธีประมาณแบบกำลังสองต่ำสุดได้เปรียบกว่าวิธีกึ่งเฉลี่ยดังตัวอย่างของสัมดังนี้ เรา

มี 12 ค่าสังเกตุ จึงแบ่งได้พวกละ 6 ค่า ซึ่งจะได้

$$\bar{X}_1 = 60 \qquad \bar{X}_2 = 80$$
$$\bar{Y}_1 = 115 \qquad \bar{Y}_2 = 85$$
$$b_2 = (115-85)/(60-80) = -1.5$$
$$b_1 = 115-(-1.5)60 = 205$$

เส้นถดถอยที่ประมาณด้วยวิธีกึ่งเฉลี่ย คือ

$$\tilde{Y}_i = 205 - 1.5X_i$$

ความแปรปรวนของ b_2 คือ $V(b_2) = 4\sigma^2/12(60-80)^2$

$$= \sigma^2/1200$$

ดังนั้น $V(b_2)/V(\hat{\beta}_2) = (\sigma^2/1200)/(\sigma^2/2250) = 1.875$

เราจะเห็นได้ว่าวิธีกึ่งเฉลี่ยจะให้ค่าประมาณของความชันที่มีความแปรปรวนมากกว่าความแปรปรวนของตัวประมาณแบบกำลังสองต่ำสุดถึง 87.5 % นั่นคือ การประมาณหรือการเดาเกี่ยวกับความชันของเส้นถดถอยประชากรที่ทำขึ้นจากวิธีกึ่งเฉลี่ยจะแผ่กระจายรอบค่าจริงมากกว่าที่ทำจากแบบวิธีกำลังสองต่ำสุด

9.4 คุณสมบัติของตัวประมาณค่าแบบกำลังสองต่ำสุด

ภายใต้ข้อสมมตของ CNLRM เราลองพิจารณาคุณสมบัติของตัวประมาณค่าแบบกำลังสองต่ำสุดเราจะพบว่า

(1) ตัวประมาณค่าแบบวิธีกำลังสองต่ำสุด $\hat{\beta}_1, \hat{\beta}_2$ เป็นฟังก์ชั่นเชิงเส้นของค่าสังเกตจากตัวอย่างของ Y นั่นคือ

$$\hat{\beta}_2 = \sum(X_i - \bar{X})(Y_i - \bar{Y})/\sum(X_i - \bar{X})^2$$
$$= \sum(X_i-\bar{X})Y_i/\sum(X_i - \bar{X})^2$$
$$= \sum k_i Y_i \ , \ k_i = (X_i - \bar{X})/\sum(X_i - \bar{X})^2$$
$$= f(Y_i)$$

$$\hat{\beta}_1 = \bar{Y} - \hat{B}_2\bar{X} = \sum Y_i/n - \sum k_i Y_i \bar{X}$$
$$= \sum(1/n - k_i\bar{X})Y_i = \sum K_i Y_i \ , \quad K_i = 1/n - k_i\bar{X}$$
$$= f(Y_i)$$

(2) ตัวประมาณแบบกำลังสองต่ำสุดจะไม่มีความเอียงเฉ นั่นคือ

$$E(\hat{\beta}_1) = \beta_1 \qquad E(\hat{\beta}_2) = \beta_2$$

(3) ความแปรปรวนของ $\hat{\beta}_1, \hat{\beta}_2$ และความแปรปรวนของมันจะเป็น

$$V(\hat{\beta}_1) = \sigma^2(1/n + \bar{X}^2/\sum(X - \bar{X})^2)$$
$$V(\hat{\beta}_2) = \sigma^2/\sum(X - \bar{X})^2$$
$$Cov(\hat{\beta}_1, \hat{\beta}_2) = -\bar{X}\sigma^2/\sum(X - \bar{X})^2$$

นั่นคือ $\hat{\beta}_1, \hat{\beta}_2$ ไม่เป็นอิสระกัน

ตัวประมาณค่าของ σ^2 ที่ไม่เอียงเฉ คือ S^2

$$S^2 = \sum e_i^2/(n-2)$$
$$= (\sum Y^2 - \hat{\beta}_1 \sum Y - \hat{\beta}_2 \sum XY)/(n-2)$$

และ $\hat{\beta}_1, \hat{\beta}_2$ ต่างก็เป็นอิสระกับ S^2

(4) ตัวประมาณค่าแบบกำลังสองต่ำสุดมีประสิทธิภาพ (Efficient) และขอบเขตต่ำสุด (Cramer-Rao Lower Bounds) สำหรับตัวประมาณค่าไม่เอียงเฉของ β_1, β_2 กำหนดได้ดังนี้

$$\begin{bmatrix} \sigma^2(1/n + \bar{X}^2/\sum(X-\bar{X})^2) & -\bar{X}\sigma^2/\sum(X-\bar{X})^2 & 0 \\ -X\sigma^2/\sum(X - \bar{X})^2 & \sigma^2/\sum(X-\bar{X})^2 & 0 \\ 0 & 0 & 2\sigma^4/n \end{bmatrix}$$

จะเห็นได้ว่าตามแนวทะแยง 2 ตัวแรกจะเป็นสูตรของ $V(\hat{\beta}_1)$ และ $V(\hat{\beta}_2)$

(5) ทฤษฎีเกาส์มาร์คอฟ (Gauss-Markoff Theorem) ตัวประมาณโดยวิธีกำลังสองต่ำสุดจะเป็นตัวประมาณเชิงเส้นที่ไม่เอียงเฉที่ดีที่สุด (Best Linear Unbiased Estimators) นั่นคือในชั้นของตัวประมาณเชิงเส้นที่ไม่เอียงเฉด้วย ตัวประมาณโดยวิธีกำลังสองต่ำสุดมีความแปรปรวนน้อยที่สุด

(6) ตัวประมาณโดยวิธีกำลังสองต่ำสุดจะเป็นตัวประมาณแบบน่าจะเป็นมากที่สุด (MLE) และมันจะมีคุณสมบัติของ Asymptotic นั่นคือ มีคุณสมบัติของ Asymptotic ในแง่ไม่เอียงเฉ คงเส้นคงวา (Consistent) และประสิทธิภาพ

(7) ตัวประมาณโดยวิธีกำลังสองต่ำสุด $\hat{\beta}_1$ และ $\hat{\beta}_2$ ต่างก็มีการแจกแจงแบบปกติ

9.5 การอ้างอิงเชิงสถิติ (Statistical Inference)

เราได้ศึกษาถึงวิธีหาตัวประมาณค่าของพารามิเตอร์และได้ทราบถึงคุณสมบัติของตัวประมาณค่ามาแล้วต่อไปเราจะพิจาณาถึงการใช้ตัวแบบถดถอยในการทดสอบสมมติฐาน และประมาณค่าเกี่ยวกับพารามิเตอร์ของการถดถอย และใช้สำหรับทำนายอีกด้วย

9.5.1 การแจกแจงของ $\hat{\beta}_1$ และ $\hat{\beta}_2$ ที่เราทราบคุณสมบัติมาแล้วว่า

การประมาณแบบกำลังสองต่ำสุด $\hat{\beta}_1$ และ $\hat{\beta}_2$

(1) ไม่เอียงเฉ นั่นคือ ค่าเฉลี่ยของมันเท่ากับค่าจริง

(2) $$V(\hat{\beta}_1) = \sigma^2 (1/n + \bar{X}^2 / \sum(X-\bar{X})^2)$$
$$V(\hat{\beta}_2) = \sigma^2 / \sum (X-\bar{X})^2$$

(3) โดยที่ทั้ง $\hat{\beta}_1$ และ $\hat{\beta}_2$ เป็นตัวผสมเชิงเส้น (Linear Combinations) ของตัวแปรปกติที่เป็นอิสระ $Y_1, Y_2, \ldots, Y_n$ ดังนั้น $\hat{\beta}_1$ และ $\hat{\beta}_2$ จึงแจกแจงปกติด้วย นั่นคือ

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2(1/n + \bar{X}^2/\sum(X-\bar{X})^2)$$
$$\hat{\beta}_2 \sim N(\beta_2, \sigma^2/\sum(X-\bar{X})^2)$$

ถ้าเราพิจารณาความแปรปรวนของ $\hat{\beta}_1$ และ $\hat{\beta}_2$ เราจะพบว่า

(1) ถ้าความแปรปรวนของตัวคลาดเคลื่อน σ^2 โตขึ้นแล้วความแปรปรวนของ $\hat{\beta}_1$ และ $\hat{\beta}_2$ ก็จะโตขึ้นด้วย

ในข้อนี้หมายความว่าการกระจัดกระจายของตัวคลาดเคลื่อนรอบ ๆ เส้นถดถอยประชากรยิ่งมากเท่าไร การกระจัดกระจายของ "การทำนายหรือการเดา" ของเราเกี่ยวกับค่าของพารามิเตอร์ยิ่งมากเท่านั้นด้วย ถ้าตัวคลาดเคลื่อนทั้งหมดเกาะกลุ่มอย่างมากอยู่ที่ค่าเฉลี่ยของมัน นั่นคือ ตัวคลาดเคลื่อนทั้งหมดเท่ากับ 0 การประมาณจะได้เช่นเดียวกับพารามิเตอร์ (ในเมื่อเราใช้ค่าสังเกตอย่างน้อยสองค่าต่าง ๆ ของตัวแปรตาม)

(2) ถ้าการกระจัดกระจายในค่าของตัวแปรที่อธิบายได้ยิ่งมากความแปรปรวนของ $\hat{\beta}_1$ และ $\hat{\beta}_2$ ยิ่งน้อยลง

ในข้อนี้ขึ้นอยู่กับความจริงที่ว่า การกระจัดกระจายของค่าของ X มากขึ้นเท่าไร $\sum (X-\bar{X})^2$ จะมากขึ้นด้วย โดยที่จริงเรามีอิสระที่จะเลือกค่าของ X จำนวนหนึ่งที่อยู่ภายในช่วงบางช่วง เช่น ช่วงจาก a ถึง b ($o<a<b$) แล้วการเลือกที่เหมาะ ๆ นั้นจะเป็นการเลือกที่ทำให้ครึ่งหนึ่งของ X เท่ากับ a และอีกครึ่งหนึ่งเท่ากับ b การเลือกเช่นนี้จะทำให้ $\sum (X-\bar{X})^2$ มีค่ามากที่สุด

(3) ถ้าค่าของ X ทั้งหมดเท่ากัน นั่นคือ $X_1 = X_2 = \ldots = X_n$ ความแปรปรวนของ $\hat{\beta}_1$ และ $\hat{\beta}_2$ จะโตมาก (Infinity Large)

ในข้อสามนี้มาจากความจริงที่ว่า ถ้าค่าทั้งหมดของตัวแปรอธิบายได้เท่ากัน ค่าของ $\sum (X-\bar{X})^2$ จะเท่ากับศูนย์ และจำนวนจำกัดจำนวนหนึ่งหารด้วยศูนย์จะมีค่ามากมาย หรือกล่าวอย่างหนึ่งได้ว่า ถ้าค่าสังเกตุทั้งหมดของ Y อยู่ตามเส้นตรงแนวดิ่ง (ค่าทั้งหมดสอดคล้องกับค่าเดียวของ X) เราจะไม่สามารถอ้างอิง หรือสรุปผลเกี่ยวกับความชันและจุดตัดแกนของเส้นถดถอยได้

เลย

(4) ความแปรปรวนของ β_1 จะน้อยที่สุด เมื่อ $\bar{X} = 0$ โดยที่ $\sum (X-\bar{X})^2 = 0$

ข้อสุดท้ายนี้ในทางปฏิบัติมีความสำคัญน้อย เพราะสนใจแต่เพียงความแปรปรวนของ β_1 เท่านั้น และกล่าวว่าถ้า พิสัย (Range) ของ X เป็นค่าลบ และค่าบวก ซึ่งทำให้ผลรวมเท่ากับศูนย์ (o) แล้ว $V(\hat{\beta}_1)$ จะน้อยที่สุด นั่นคือ $V(\hat{\beta}_1) = \sigma^2/n$

ตัวอย่าง จากตัวอย่างเกี่ยวกับอุปสงค์ของส้ม เราทราบว่า $\bar{X} = 70$ และ $\sum (X-\bar{X})^2 = 2250$ ความแปรปรวนของ $\hat{\beta}_1, \hat{\beta}_2$

$$\begin{aligned} V(\hat{\beta}_1) &= \sigma^2(1/n + \bar{X}^2/\sum(X-\bar{X})^2) \\ &= \sigma^2(1/12 + 70^2/2250) = \sigma^2 \ (2.26111) \\ V(\hat{\beta}_2) &= \sigma^2/\sum(X-\bar{X})^2 = \sigma^2/2250 \\ &= 0.000444 \end{aligned}$$

ถ้าค่าของ X มีดังนี้ $X_1 = X_2 = \ldots = X_6 = 100$ และ $X_7 = X_8 = \ldots = X_{12} = 50$ แล้ว $\bar{X} = 75$ และ $\sum (X-\bar{X})^2 = 7500$ เราจะได้

$$\begin{aligned} V(\hat{\beta}_1) &= \sigma^2(1/2 + 75^2/7500) = 0.833333 \\ V(\hat{\beta}_2) &= \sigma^2/7500 = 0.000143 \end{aligned}$$

เปรียบเทียบความแปรปรวนทั้งสองกรณีเราจะได้

$$\frac{V(\hat{\beta}_1) \text{ กรณีแรก}}{V(\hat{\beta}_1) \text{ กรณีหลัง}} = (2.261111)/(0.833333) = 2.713$$

และ

$$\frac{V(\hat{\beta}_2) \text{ กรณีแรก}}{V(\hat{\beta}_2) \text{ กรณีหลัง}} = (0.000444)/(0.000143) = 3.338$$

นั่นคือความแปรปรวนของ $\hat{\beta}_1$ และ $\hat{\beta}_2$ ในกรณีแรกเป็น $2\frac{1}{2}$ เท่า และ $3\frac{1}{3}$ เท่า ของความแปรปรวนในกรณีที่สอง

จะเห็นได้ว่าประสิทธิภาพที่ได้จากการเลือกที่ดีในค่าของ X จะเป็นสิ่งที่น่าพิจารณา แต่ในทางปฏิบัติเป็นไปไม่ได้เพราะไม่มีโอกาสได้เลือก หรือสุ่มตัวอย่างเอง ส่วนมากก็ได้รับผลสำเร็จมาแล้ว

9.5.2 ความแปรปรวนร่วมของ $\hat{\beta}_1$ และ $\hat{\beta}_2$

เราจะพิจารณาถึงความสัมพันธ์ระหว่าง $\hat{\beta}_1$ และ $\hat{\beta}_2$ จากการใช้ $\hat{\beta}_1, \hat{\beta}_2$ ประมาณ β_1, β_2

นั้นความคลาดเคลื่อนจากตัวอย่าง (Sampling Errors) จะเกิดขึ้น และเราประสงค์ที่จะทราบค่าความคลาดเคลื่อนของสองตัวนี้จะมีเครื่องหมายเหมือนกันหรือไม่ นั่นคือเราต้องการหาเครื่องหมายของความแปรปรวนร่วมของ $\hat{\beta}_1$ และ $\hat{\beta}_2$ หรือของ Cov $(\hat{\beta}_1, \hat{\beta}_2)$

$$\begin{aligned} \text{Cov}\ (\hat{\beta}_1, \hat{\beta}_2) &= E(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2) = -\bar{X}E(\hat{\beta}_2 - \beta_2)^2 \\ &= -\bar{X}V(\hat{\beta}_2) = -\bar{X}\sigma^2 / \sum (X - \bar{X})^2 \end{aligned}$$

จากผลอันนี้เราจะได้ว่า เมื่อ $\bar{X}$ เป็นบวก ความคลาดเคลื่อนของตัวอย่างของ $\hat{\beta}_1$ และ $\hat{\beta}_2$ จะมีเครื่องหมายตรงข้าม นั่นคือ ค่าประมาณที่มากของค่าจริงของ β_1 มีความเกี่ยวพันกับ ค่าที่ต่ำของค่าจริงของ β_2 และในทางกลับกัน

9.5.3 การประมาณความแปรปรวนของ $\hat{\beta}_1$ และ $\hat{\beta}_2$

ภายใต้ข้อสมมติของ CNLRM เราได้ตัวประมาณแบบกำลังสองต่ำสุดที่มีคุณสมบัติที่ดีของตัวประมาณ แต่มันก็ยังขึ้นอยู่กับขนาดของความแปรปรวน อย่างไรก็ตามเราก็สบายใจจากคุณสมบัติที่ว่าไม่มีตัวประมาณที่ไม่เอียงเฉตัวอื่น ๆ ที่มีความแปรปรวนน้อยกว่า ถ้าความแปรปรวนโตมาก การคาดคะเนเกี่ยวกับค่าจริงของพารามิเตอร์จะไกลจากจุดหมายมากที่ผ่านมาเราหาความแปรปรวนของ $\hat{\beta}_1$ และ $\hat{\beta}_2$ ไว้แล้วแต่ยังคิดอยู่ในรูปของพารามิเตอร์ σ^2 ที่ไม่ทราบค่า ดังนั้นเราจึงประเมินผลการคาดคะเนไม่ได้ อย่างไรก็ตามเราสามารถประมาณ σ^2 ได้เพราะจากวิธีประมาณค่าแบบมากสุด (MLE) เราได้

$$\hat{\sigma}^2 = (1/n) \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

และ $\hat{\sigma}^2$ นี้มีคุณสมบัติแบบ Asymptotic ($n \to \infty$) ที่ต้องการทั้งหมด

$\hat{\sigma}^2$ เป็นตัวประมาณค่าที่เอียงเฉของ σ^2 ทั้งนี้เพราะ

$$E(\hat{\sigma}^2) = \frac{n-2}{n}\sigma^2 \neq \sigma^2$$

ถ้าเราจะหาตัวประมาณที่ไม่เอียงเฉของ σ^2 ได้โดยคูณ $n/(n-2)$ เข้าทั้งสองข้างจะได้

$$\frac{n}{n-2} E(\hat{\sigma}^2) = \sigma^2$$

หรือ $$E\left(\frac{n}{n-2}\right)(1/n) \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 = \sigma^2$$

ซึ่งจะได้ $$E \sum (Y - \hat{\beta}_1 - \hat{\beta}_2 X)^2 / (n-2) = \sigma^2$$

ดังนั้นประมาณไม่เอียงเฉของ σ^2 คือ S^2

$$S^2 = \sum(Y_i - \beta_1 - \beta_2 X_i)^2/(n-2)$$
$$= \sum e_i^2/(n-2)$$

เมื่อ $n \to \infty$ เราจะได้ว่า $\frac{1}{n-2} \cong \frac{1}{n}$ ดังนั้น S^2 จึงเท่ากับ σ^2 แบบ Asymptotic และ S^2 มีคุณสมบัติที่ดีแบบ Asymptotic

เพื่อความสะดวกในการคำนวณและการจำเราทำสูตรของ S^2 ให้เป็นรูปดังนี้

$$(n-2)S^2 = \sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$
$$= \sum Y^2 - \hat{\beta}_2 \sum XY - n\bar{Y}\hat{\beta}_1$$

ดังนั้น $$S^2 = (\sum Y^2 - \beta_1 \sum Y - \beta_2 \sum XY) / (n-2)$$

เราก็สามารถจะหาตัวประมาณของ $V(\hat{\beta}_1)$ และ $V(\hat{\beta}_2)$ ได้ซึ่งเป็นตัวประมาณที่ไม่เอียงเฉ กับมีคุณสมบัติที่ดีแบบ Asymptotic อีกด้วย เราจะได้ $S^2_{\hat{\beta}_1}$ และ $S^2_{\hat{\beta}_2}$ เป็นตัวประมาณของ $V(\hat{\beta}_2)$ และ $V(\hat{\beta}_1)$ ซึ่งจะมีสูตรดังนี้

$$S^2_{\hat{\beta}_1} = S^2(1/n + \bar{X}^2/\sum(X-\bar{X})^2$$

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

รากที่สองของตัวประมาณค่าเหล่านี้ คือ $S_{\hat{\beta}_1}$ และ $S_{\hat{\beta}_2}$ ซึ่งเรียกว่า ความคลาดเคลื่อนมาตราฐาน (Standard Error) ของ $\hat{\beta}_1$ และ $\hat{\beta}_2$ และจะใช้เป็นมาตราวัดความเที่ยงตรงของ $\hat{\beta}_1$ และ $\hat{\beta}_2$

9.5.4 ช่วงเชื่อมั่นของ β_1, β_2 และ σ^2

สิ่งที่แสดงถึงความเที่ยงตรง (Precision) ของ $\hat{\beta}_1$, $\hat{\beta}_2$ และ S^2 ก็คือการสร้างช่วงเชื่อมั่นของ β_1, β_2 และ σ^2

สำหรับ $S^2 = \sum e_i^2/(n-2)$

เราทราบว่า $(n-2)S^2/\sigma^2 \sim \chi^2(n-2)$

ซึ่งเราจะได้ช่วงเชื่อมั่น $100(1-\alpha)\%$ ของ σ^2 เป็น

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

สำหรับ β_2 เราทราบว่า $\hat{\beta}_2 \sim N(\beta_2, \sigma^2_{\hat{\beta}_2})$ นั่นคือ

$$(\hat{\beta}_2 - \beta_2)/\sigma_{\hat{\beta}_2} \sim N(0,1)$$

และเราทราบว่า

$$(n-2)S^2_{\hat{B}_2} / \sigma^2_{\hat{B}_2} \sim \chi^{2(n-2)}$$

ดังนั้นเราจะได้

$$(\hat{\beta}_2 - \beta_2)/ S_{\hat{B}_2} \sim t^{(n-2)}$$

ในทำนองเดียวกันเราจะได้

$$(\hat{\beta}_1-\beta_1)/ S_{\hat{\beta}_1} \sim t^{(n-2)}$$

จากการแจกแจงดังกล่าวเราสามารถสร้างช่วงความเชื่อมั่น $100(1 - \alpha)$ % ของ β_1, และ β_2 ได้เป็น

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

$$\hat{B}_2 - t^{(n-2)}_{\alpha/2} S_{\hat{\beta}_2} < \beta_2 < \hat{\beta}_2 + t^{(n-2)}_{\alpha/2} S_{\hat{\beta}_2}$$

ในเมื่อ $(1 - \alpha)$ เป็นความน่าจะเป็นที่ช่วงเชื่อมั่นที่ระบุไว้นี้จะครอบคลุมค่าจริงของพารามิเตอร์ถดถอย β_1 β_2 ซึ่งหมายถึงระดับความเชื่อมั่น (Level of Confidence) ด้วย ตามปกติใช้ระดับ 95 % หรือ 99%

ตัวอย่าง สร้างช่วงเชื่อมั่น 95 % ของ δ^2, β_1 และ β_2 จากตัวอย่างเรื่องส้มซึ่งมี

$$\bar{X} = 70 \quad \bar{Y} = 100 \quad \sum xy = -3550$$

$$\sum x^2 = 2250 \quad \sum y^2 = 6300 \quad \hat{\beta}_1 = 210.460$$

$$\hat{\beta}_2 = -1.578$$

$$S^2 = (\sum y^2 - \beta_2 \sum xy)/(n-2)$$

$$= (6300-(-1.578)(-3550))/(12-2) = 69.8$$

$$S^2_{\hat{\beta}_1} = 69.8(1/12 + 70^2/2250) = 157.825555$$

$$S^2_{\hat{\beta}_2} = 69.8/2250 = 0.031022$$

$$S_{\hat{\beta}_1} = 12.563 \qquad S_{\hat{\beta}_2} = 0.176$$

$$t^{(12-2)}_{.025} = 2.228 \qquad \chi^{2(12-2)}_{.025} = 20.48$$

$$\chi^{2(12-2)}_{.975} = 3.25$$

ดังนั้นช่วงเชื่อมั่น 95% ของ σ^2, β_1 และ β_2 คือ

$$(12-2)69.8/20.48 < \sigma^2 < (12-2)69.8/3.25$$
$$210.460-2.228(12.563) < \beta_1 < 210.460+2.228(12.563)$$
$$-1.578-2.228(0.176) < \beta_2 < -1.578+2.228(0.176)$$

นั่นคือ
$$34.082 < \sigma^2 < 214.769$$
$$182.470 < \beta_1 < 238.450$$
$$-1.970 < \beta_2 < -1.186$$

9.5.5 ช่วงเชื่อมั่นของ $E(Y_i)$

เราจะพิจารณาความเที่ยงตรง (Precision) ของเส้นถดถอยตัวอย่างที่ใช้แทนเส้นถดถอยประชากร $E(Y_i) = \beta_1 + \beta_2 X_i$ ซึ่งกำหนดไว้สำหรับค่า X ใด ๆ ภายในพิสัยบางอย่างตัวประมาณค่าของเส้นถดถอยของประชากรก็คือเส้นถดถอยตัวอย่าง $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

โดยที่ $E(\hat{Y}_i) = \beta_1 + \beta_2 X_i$ (มันจึงเป็นตัวประมาณไม่เอียงเฉของ $E(Y_i)$ และโดยที่ $\hat{\beta}_1, \hat{\beta}_2$ เป็นตัวประมาณที่ดีของ β_1, β_2 เราก็อาจกล่าวได้ว่า $\hat{Y}_i$ ก็เป็นตัวประมาณที่ดีที่สุดของ $E(Y_i) = \beta_1 + \beta_2 X_i$

เมื่อพิจารณาจุดที่กำหนดใด ๆ บนเส้นถดถอยประชากร เราจะพบความแปรปรวนของตัวประมาณ $\hat{Y}_i$ ของมันคือ $\sigma^2_{\hat{Y}_i}$ ซึ่งหาได้ดังนี้

$$\sigma^2_{\hat{Y}_i} = E(\hat{Y}_i - E(\hat{Y}_i))^2$$
$$= \sigma^2(1/n + (X-\bar{X})^2/\sum(X-\bar{X})^2)$$

เราจะพิจารณาต่อไปถึงการแจกแจงของ $\hat{Y}_i$ ซึ่งทำได้ง่าย ๆ ดังนี้ เนื่องจาก $(\hat{\beta}_1 + \hat{\beta}_2 X_i)$ เป็นตัวผสมเชิงเส้นของตัวแปรเชิงสุ่ม $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ ที่มีการแจกแจงเป็นอิสระแบบปกติ มันจึงแจกแจงปกติด้วย นั่นคือ

$$\hat{Y}_i \sim N(\beta_1 + \beta_2 X_i, \sigma^2_{\hat{Y}_i})$$

หรือ
$$\frac{\hat{Y}_i - (\beta_1 - \beta_2 X_i)}{\sigma_{\hat{Y}_i}} \sim N(0,1)$$

โดยทั่วไป $\sigma^2_{\hat{Y}_i}$ หาค่าไม่ได้เพราะมีพารามิเตอร์ σ^2 ติดอยู่ เราจึงใช้ S^2 ประมาณ σ^2 เพราะ S^2 เป็นตัวประมาณค่าแบบ Asymptotic ที่ไม่เอียงเฉคงเส้นคงวา และมีประสิทธิภาพ แล้วเราจะได้ตัวประมาณของ $\sigma^2_{\hat{Y}_i}$ คือ $S^2_{\hat{Y}_i}$ ซึ่งมีคุณสมบัติที่ต้องการเหมือนกัน

ดังนั้น $S^2_{\hat{Y}_i}$ กำหนดไว้ดังนี้

$$S^2_{\hat{Y}_i} = S^2(1/n + (X_i - \bar{X})^2 / \sum (X_i - \bar{X})^2$$

แล้วเราจะได้ว่า $$\frac{\hat{Y}_i - (\beta_1 + \beta_2 X_i)}{S_{\hat{Y}_i}} \sim t^{(n-2)}$$

และช่วงเชื่อมั่น $100(1 - \alpha)$ % ของ $(\beta_1 + \beta_2 X_i)$ หรือ $E(Y_i)$ คือ

$$\hat{Y}_i - t^{(n-2)}_{\alpha/2} S_{\hat{Y}_i} < (\beta_1 + \beta_2 X_i) < \hat{Y}_i + t^{(n-2)}_{\alpha/2} S_{\hat{Y}_i}$$

ในเมื่อ $(1 - \alpha)$ เป็นระดับความเชื่อมั่นที่เราต้องการ โดยที่ช่วงเชื่อมั่นนี้สามารถคำนวณได้สำหรับค่าใด ๆ ของ X_i ที่อยู่ในโดเมน (Applicable Domain) เราสามารถสร้างช่วงเชื่อมั่นสำหรับจุดใด ๆ บนเส้นถดถอยประชากรได้ และแล้วเราก็สามารถสร้างแถบเชื่อมั่น (Confidence Band) ของเส้นถดถอยทั้งหมดได้

ตัวอย่าง จากตัวอย่างอุปสงค์ของส้ม เราจะสร้างแถบเชื่อมั่น 95% สำหรับโค้งอุปสงค์ของประชากร เราได้เส้นถดถอยตัวอย่างที่ประมาณแล้วคือ

$$\hat{Y}_i = 210.460 - 1.578 X_i$$

และค่าประมาณของความแปรปรวนของ $\hat{Y}_i$ คือ

$$S^2_{\hat{Y}_i} = 69.8(1/12 + (X_i - 70)^2/2250)$$

สำหรับ $t^{(12-2)}_{.025}$ มีค่าเท่ากับ 2.228 เราจะได้แถบเชื่อมั่น 95% ของ $E(Y_i)$ ดังนี้

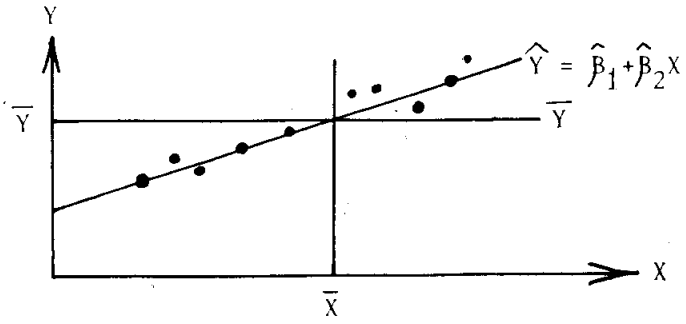
X_i	Y_i	S_{Y_i}	$2.228 S_{Y_i}$	ช่วงเชื่อมั่น 95% ขีดจำกัดขั้นต่ำ	ขีดจำกัดขั้นสูง
0	210.46	12.57	28.01	182.45	238.47
10	194.66	10.84	24.15	170.51	218.81
20	178.88	9.14	20.36	158.52	199.24
30	163.10	7.45	16.60	146.50	179.70
40	147.32	5.81	12.94	134.38	160.26

120	36.86	7.45	16.60	20.26	53.46
120	21.08	9.14	20.36	0.72	41.44

จากตารางข้างบนในแถวทั้งสุดท้าย 2 แถว จะแทนช่วงที่เราหวังว่ามันจะครอบคลุมค่าของประชากร ช่วงที่แคบที่สุดก็คือ ช่วงที่สมนัยกับ $X_i = X = 70$ ช่วงจะกว้างขึ้นเมื่อ X_i เคลื่อนห่างจากค่าเฉลี่ย $\bar{X}$

9.6 การแยกความผันแปรของตัวอย่างในตัวแปรตาม (Decomposition of the Sample Variation of Y)

ความผันแปรของตัวแปรตามนั้นเราหมายถึงการเปลี่ยนแปลงใน Y จากค่าสังเกตของตัวอย่างค่าหนึ่ง ๆ ไปยังค่าอื่น ๆ แนวความคิดที่จะแยกความผันแปรในตัวแปรตามก็เพื่อต้องการจะเพิ่มความรู้เกี่ยวกับผลการประมาณที่หามาได้แล้ว ตัวอย่างเช่น พิจารณาความผันแปรของ Y ในรูปต่อไปนี้ เราจุด (Plot) ค่าสังเกตุจากตัวอย่างของ Y ที่สมนัยค่าของ X เราจะได้กราฟที่เรียกว่า "แผนภาพการกระจาย" (Scatter diagram) ดังรูป



จากรูปเราให้ 10 ค่าสังเกตของ Y ที่สมนัยกับ 10 ค่าต่าง ๆ ของ X แล้วเราจะได้คำถามว่า "ทำไมค่าของ Y จึงต่างกันจากค่าสังเกตหนึ่ง ๆ และเนื่องมาจากอะไร" คำตอบ (ตามตัวแบบถดถอยที่สมมติ) ก็คือ ความผันแปรใน Y นี้ส่วนหนึ่งเนื่องจากการเปลี่ยนแปลงใน X (ซึ่งนำไปสู่การเปลี่ยนแปลงในค่าเฉลี่ยของ Y) และอีกส่วนหนึ่งเนื่องจากผลกระทบ (Effect) ของตัวคลาดเคลื่อน

คำถามต่อไปนี้คือ ความผันแปรที่สังเกตได้ใน Y นี้ เนื่องมาจากความผันแปรใน X สักเท่าใดและจากผลของตัวคลาดเคลื่อนสักเท่าใด คำตอบจะทำได้ก็โดยอาศัยมาตรการบางอย่างที่จะกล่าวต่อไปนี้

ประการแรกเราจะให้คำนิยามของคำว่า "ความผันแปรของตัวอย่างใน Y" ก่อน ถ้าไม่มีความผันแปรค่าของ Y ทั้งหมดจะอยู่บนเส้นนอน ถ้าค่าของ Y ทุกค่าเท่ากันหมดแล้ว Y ทุกค่าควรจะเท่ากับค่าเฉลี่ยของมัน $\bar{Y}$ เส้นนอนจะเป็นเส้นหนึ่งที่สมนัยกับ $\bar{Y}$ ตามความเป็นจริงค่าสังเกตของ Y จะกระจายรอบ ๆ เส้นนี้ ดังนั้นความผันแปรของ Y ควรจะวัดด้วยระยะทางของค่าสังเกตของ Y

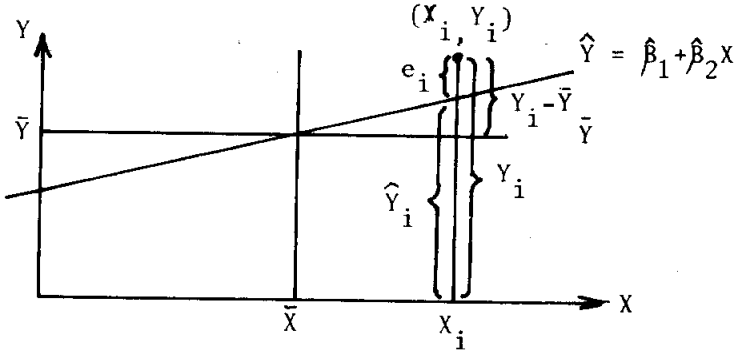
ที่เบี่ยงเบนจาก $\bar{Y}$ มาตรการที่ดีของระยะทางนี้เราจะใช้ผลรวมของค่ากำลังของมัน ปกติเราจะเรียกว่า ผลรวมกำลังทั้งหมด (Total Sum, Square, SST) นั่นคือเรานิยามว่า

$$SST = \sum (Y_i - \bar{Y})^2 = \sum y_i^2$$

ขั้นต่อไปเราจะแยก SST ออกเป็นสองส่วน คือ ส่วนหนึ่งเป็นความผันแปรของ Y ซึ่งอธิบายถึงความผันแปรของ X และอีกส่วนหนึ่งสมมติว่าเป็นความผันแปรใน Y ที่บ่งถึงสาเหตุเชิงสุ่มอย่างหนึ่ง (Random Causes)

ลองพิจารณารูปที่ผ่านมา สมมติว่าเส้นถดถอยตัวอย่างได้จากวิธีกำลังสองต่ำสุด เส้นนี้จะให้ผลรวมกำลังสอง SSE มีค่าน้อยที่สุด บางทีเรียกเส้นนี้ว่า เส้นที่ปรับได้ดีที่สุด (Line of the Best Fit)

ถ้าพิจารณาค่าสังเกตค่าหนึ่งของ Y หรือ Y_i ที่สมนัยกับค่าของ X หรือ X_i เราสนใจระยะทางของ (X_i, Y_i) ที่ห่างจากเส้น $\bar{Y}$ ดังรูป



เราจะเห็นว่าระยะทางนี้แบ่งได้เป็น 2 ส่วน ส่วนหนึ่งเป็นระยะทางที่สังเกตุได้ถึงเส้นถดถอยตัวอย่างและอีกส่วนหนึ่งเป็นระยะทางจากเส้นถดถอยถึง $\bar{Y}$ นั่นคือ

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

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

นี่เป็นค่าสังเกตุค่าเดียว เราต้องการมาตรการที่สรุปสำหรับทุก ๆ ค่าสังเกตุของตัวอย่าง เราจึงยกกำลังสองและรวมผลซึ่งจะได้

$$\underset{SST}{\sum (Y_i - \bar{Y})^2} = \underset{SSR}{\sum (\hat{Y} - \bar{Y})^2} + \underset{SSE}{\sum e_i^2}$$

เราจะได้ $SSR = \sum (\hat{Y} - \bar{Y})^2 = \hat{\beta}_2 \sum (X_i - \bar{X})^2$

ดังนั้นความผันแปรของตัวอย่างใน Y (SST) แบ่งแยกได้เป็น 2 ส่วน ส่วนแรก (SSR)

บรรยายถึงความผันแปรของค่าที่ปรับได้ (Fitted Values) ของ Y หรือ บรรยายถึงผลกระทบ (Estimated Effect) ของ X ในความผันแปรของ Y และส่วนอื่น (SSE) บรรยายถึงความผันแปรของเศษจากการถดถอย (Regression Residuals) หรือบรรยายถึงผลกระทบของตัวคลาดเคลื่อน

9.7 สัมประสิทธิ์การตัดสินใจ (Coefficient of Determination)

การแยกความผันแปรของ Y นี้จะได้มาตราวัดอันหนึ่งเกี่ยวกับการปรับที่ดีที่สุด (Goodness of fit) นั่นคือมาตราวัดที่เรียกว่า สัมประสิทธิ์การตัดสินใจ (Coefficient of Determination, R^2) ซึ่งเป็นสัดส่วนของความผันแปรของ Y ที่เนื่องมาจากความผันแปรของ X และ สัมประสิทธิ์การตัดสินใจ R^2 นิยามไว้ว่า

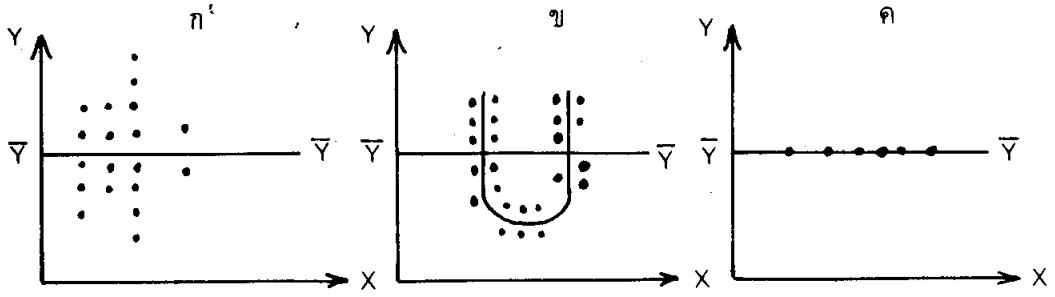
$$R^2 = SSR/SST = \hat{\beta}_2^2 \sum (X_i - \bar{X})^2 / \Sigma (Y_i - \bar{Y})^2$$

หรือ $$R^2 = 1 - SSE/SST = 1 - \sum e_i^2 / \sum y_i^2$$

R^2 เป็นมาตราวัดที่ใช้บรรยายว่า "เส้นถดถอยตัวอย่างปรับเข้าได้กับข้อมูลที่สังเกตได้ดีเพียงไร" เราจะเห็นได้ว่า R^2 ไม่เป็นลบ หรือมากกว่า 1 นั่นคือ $0 \leq R^2 \leq 1$

ถ้า $R^2 = 0$ แสดงว่าการปรับที่เลวที่สุด และถ้า $R^2 = 1$ แสดงว่าการปรับที่ดีที่สุด

มีสิ่งที่น่าสนใจเมื่อ $R^2 = 0$ นั่นคือ เส้นถดถอยอยู่ในแนวนอน หรือ $\beta_2 = 0$ แต่เส้นถดถอยจะเป็นเส้นนอนนั้นมีเหตุหลายประการ ดังรูป



กรณี (ก) นั้นค่าสังเกตกระจายแบบสุ่มรอบ Y

กรณี (ข) ค่าสังเกตกระจายรอบเส้นโค้งทั้งที่เส้นถดถอยเป็นเส้นนอน กรณีนี้ X, Y มีความสัมพันธ์กัน แต่ความสัมพันธ์ไม่ใช่เชิงเส้นเลย ดังนั้นเส้นตรงจึงปรับเข้ากับข้อมูลไม่ดีเลย

กรณี (ค) ค่าสังเกตทั้งหมดเท่ากันไม่คำนึงว่า X จะมีค่าใด ๆ กรณีนี้เป็นกรณียกเว้นเมื่อทุกค่าของ Y คงที่จึงไม่มีความผันแปรที่อธิบายได้ การแยกความผันแปรจึงทำไม่ได้ ดังนั้นค่าของ R^2 ในกรณีนี้จึงพิจารณาไม่ได้ (Indeterminate)

นอกจากนี้สิ่งที่น่าสังเกตเกี่ยวกับความผันแปรของตัวอย่าง Y คือ การแยก SST เป็น SSR และSSE นั้น ทำได้กับวิธีที่จะหาเส้นถดถอยโดยวิธีกำลังสองต่ำสุด แต่ถ้าหาโดยวิธีอื่นจะแยก SST เป็น SSR กับ SSE ไม่ได้ แต่อย่างไรก็ตาม R^2 ก็ยังสามารถใช้ได้กับวิธีประมาณโดยวิธีอื่น ๆ ได้ กำหนด R^2 ดังนี้

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

เป็นผลรวมของเศษซึ่งเป็นส่วนเบี่ยงเบนจากเส้นถดถอย ซึ่งไม่คำนึงว่าจะประมาณโดยวิธีอะไร แต่ถ้าเส้นถดถอยได้จากวิธีอื่นที่ไม่ใช่วิธีกำลังสองต่ำสุด R^2 ก็ไม่สามารถแปลได้ว่าเป็นมาตราวัดของสัดส่วนของความผันแปรของ Y ที่เนื่องมาจากการถดถอยของตัวอย่าง ในกรณีนี้ R^2 ควรจะใช้เป็นมาตราวัดของ "การปรับที่ดีที่สุด (Goodness of Fit)"

ถ้าถือ R^2 เป็น ตัวสถิติเชิงพรรณาที่หาได้จากตัวอย่างที่กำหนด แล้ว R^2 มีค่าต่ำ เราจะถือว่าการปรับเส้นถดถอยเข้ากับค่าสังเกตจะไม่ค่อยดี หรือ X เป็นตัวแปรอธิบายได้ที่ไม่ดี ในแง่ที่ว่าความผันแปรใน X ไม่มีผลกระทบต่อ Y หรืออาจกล่าวได้ว่าอิทธิพลของ X ต่อ Y อ่อนมากเมื่อเทียบกับอิทธิพลของตัวคลาดเคลื่อน หรือยังกล่าวได้อีกว่าสมการถดถอยที่กำหนดไม่ถูกต้อง นั่นคือ R^2 เป็นตัวชี้ถึงความถูกต้องของข้อกำหนด (Specification) ของตัวแบบ

สำหรับประชากรมาตราที่วัดการปรับที่ดีที่สุด ก็คือ สัมประสิทธิ์การตัดสินใจของประชากร ρ^2 และเราประมาณได้ด้วย R^2 และ R^2 นั้นเป็นตัวประมาณค่าที่เอียงเฉทางบวก นั่นคือ $E(R^2) > \rho$ เมื่อแก้ความเอียงเฉเราจะได้R^2 ที่ปรับปรุงแล้วเป็น $\bar{R}^2$ ซึ่งทำได้โดยใช้องศาความเป็นอิสระนั่นคือ

$$\bar{R}^2 = 1 - \frac{SSE/(n-2)}{SST/(n-1)} = 1 - S^2_{y.x}/S^2_y$$

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

แต่เมื่อใช้ตัวอย่างขนาดโต $E(R^2) \cong \rho^2$ นั่นคือ R^2 เป็นตัวประมาณไม่เอียงเฉแบบ Asymptotic ของ ρ^2 และ $R^2 \cong \bar{R}^2$

ถ้า Y อธิบายด้วย X อย่างสมบูรณ์แล้ว $\sum e^2 = 0$ และ $R^2 = 1$ ถ้าไม่มีความสัมพันธ์ระหว่าง Y และ X, $\beta_2 = 0$ แล้วทั้ง $S^2_{y.x}$ และ S^2_y จะเป็นค่าประมาณของจำนวนเดียวกัน อัตราส่วนจะประมาณ 1 นั่นคือ $R^2 = 0$ อย่างไรก็ตามอาจเป็นไปได้ที่ $S^2_{yx} > S^2_y$ ซึ่งจะทำให้ R^2 เป็นลบได้เพราะ $\frac{n-1}{n-2} > 1$ และ SSE/SST เท่ากับหรือใกล้ ๆ 1 ถ้า $\bar{R}^2$ เป็นลบก็ให้แปลความหมายเช่นเดียวกับ $R^2 = 0$ ดังนั้น $\bar{R}^2 \leq R^2$

ตัวอย่าง จากตัวอย่างเรื่องส้ม เราสามารถแยก SST และหาค่า R^2 ได้ดังนี้

จากข้อมูลเราได้ $\sum y^2 = 6300$; $\sum x^2 = 2250$; $\hat{\beta}_2 = -1.578$

$$SST = \sum y^2 = 6300$$

$$SSR = \hat{\beta}^2 \sum x^2 = (-1.5781)^2 (2250) = 5602$$

$$SSE = SST - SSR = 698$$

$$R^2 = SSR/SST = 5602/6300$$

$$= 0.889$$

เราหมายความว่า 88.9 ของความผันแปรของ Y เนื่องมาจากความผันแปรของค่าที่ปรับได้ (Fitted Value) ของ Y คือ Y หรือเนื่องมาจากความผันแปรใน X นั่นเอง

9.8 การทดสอบสมมติฐาน (Test of Hypotheses)

เราหันมาสนใจในการใช้ตัวแบบถดถอยสำหรับสอบสมมติฐาน แบบของสมมติฐานที่จะทดสอบโดยใช้ตัวแบบนี้ช่วยก็คือ "ไม่มีความสัมพันธ์ระหว่างตัวแปรที่อธิบายได้กับตัวแปรตาม" สมมติฐานนี้จะได้รับการแปลความหมายได้อย่างชัดเจน ถ้าเราระบุถึงส่วนประกอบของสมมติฐานที่เกี่ยวข้องนั่นคือเป็นการกล่าวถึงข้อสมมติเกี่ยวกับประชากรที่เราประสงค์จะทดสอบนั่นเอง ข้อสมมติเหล่านี้เป็นหลักสำคัญของตัวแบบถดถอยเชิงเส้นปกติแบบคลาสสิค (CNLRM) ซึ่งระบุไว้ด้วยข้อความดังนี้

(1) $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$

(2) Normality : ε_i แจกแจงปกติ

(3) Zero mean : $E(\varepsilon_i) = 0$

(4) Homoskedasticity : $V(\varepsilon_i) = E(\varepsilon_i)^2 = \sigma^2$

(5) Nonautoregression : $Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i, \varepsilon_j) = 0; i \neq j$

(6) Nonstochastic X : $\sum (X-\bar{X})^2 \neq 0$

ภายใต้สมมติฐานที่เกี่ยวข้องซึ่งกล่าวมานี้ ความสัมพันธ์ระหว่าง X และ Y กำหนดไว้ว่าเป็นการพึ่งพิงเชิงเส้น (Linear Dependence) ของค่าเฉลี่ยของ Y_i ต่อ X_i นั่นคือ $E(Y_i) = \beta_1 + \beta_2 X_i$ ดังนั้นข้อความที่ว่า "ไม่มีความสัมพันธ์ระหว่าง X และ Y" จึงแปลความหมายได้ว่า ค่าเฉลี่ยของ Y_i จะไม่เป็นการพึ่งพิงเชิงเส้นกับ X_i นั่นคือเส้นถดถอยประชากรจะอยู่ในแนวนอน อาจจะกล่าวได้ง่าย ๆ ว่า β_2 เท่ากับศูนย์ เพราะฉะนั้นสมมติฐานหลัก (Ho) ที่ว่าไม่มีความสัมพันธ์ระหว่าง X และ Y ก็คือ

$$Ho : \beta_2 = 0$$

ถ้าเราไม่มีความรู้เกี่ยวกับค่าของพารามิเตอร์ถดถอยมาก่อนแล้วสมมติฐานรอง (Ha) จะเป็น

$$Ha : \beta_2 \neq 0$$

แต่ถ้าเรารู้ก่อนว่า จะเป็นบวกหรือลบ สมมติฐานรองควรเปลี่ยนแปลงแก้ไขตามที่ทราบมาก่อน การทดสอบ Ho นั้นเราจะต้องหาตัวสถิติทดสอบ (Test Statistic) ขึ้นมาและพิจารณาเขตปฏิเสธ หรือเขตวิกฤต (Rejection or Critical Region) ตัวสถิติทดสอบได้มาจากตัวประมาณแบบกำลังสองต่ำสุดของ β_2 ซึ่งมีคุณสมบัติที่ดีภายใต้ข้อสมมติที่กำหนดให้

เราทราบมาแล้วว่า ภายใต้สมมติฐานที่ว่า β_2 เท่ากับศูนย์ นั่น

ดังนั้นตัวสถิติทดสอบที่เหมาะ คือ $$T = (\hat{\beta}_2 - \beta_2)/S_{\hat{\beta}_2} \sim t^{(n-2)}$$
$$T = (\hat{\beta}_2 - \beta_2)/S_{\hat{\beta}_2}$$

เขตกั้น (boundary) ระหว่างเขตยอมรับและเขตปฏิเสธพิจารณาได้จากตาราง t ที่มีระดับนับสำคัญและจำนวน df ที่กำหนดคือ α และ n-2 สำหรับการทดสอบสองทาง (Two-tailed Test) เรากำหนดเขตยอมรับได้ดังนี้

$$-t^{(n-2)}_{\alpha/2} \leq T = (\hat{\beta}_2 - \beta_2)/S_{\hat{\beta}_2} \leq t^{(n-2)}_{\alpha/2}$$

ตัวอย่าง พิจารณาอุปสงค์ของส้มที่ประมาณด้วยตัวแบบถดถอยเชิงเส้นจากข้อมูลที่กำหนดที่แล้วมาสมมติฐานหลัก Ho ที่ว่า ไม่มีความสัมพันธ์ระหว่างราคา X และปริมาณอุปสงค์ Y หรืออาจจะพูดว่าราคาไม่มีอิทธิพลต่อปริมาณอุปสงค์ของส้ม นั่นคือ

$$Ho : \beta_2 = 0$$

และสมมติฐานรองเป็น $$Ha : \beta_2 = 0$$

ถ้าเราต้องการทดสอบสมมติฐาน ณ ระดับนัยสำคัญ 0.01 โดยที่การแจกแจง t มีค่าจาก α ถึง α ค่าใด ๆ ของ T จึงคงเส้นคงวา (Consistent) กับ Ho แต่ถ้า Ho เป็นจริงค่าของ t ที่จะห่างจากศูนย์มากจึงไม่น่าจะเป็นจริง การตัดสินใจที่ใช้ระดับนัยสำคัญ 1% หมายความว่าถ้า T เบี่ยงเบนจากศูนย์มีค่ามากและจะเกิดขึ้นด้วยความจะเป็นถึง 1% เราจะปฏิเสธ Ho เรามี 12 ค่าสังเกต ค่าที่เหมาะสมของตัวสถิติทดสอบ T สำหรับการทดสอบสองทางของเรา คือ

$$-3.169 \leq T = \hat{\beta}_2/S_{\hat{\beta}_2} \leq 3.169$$

จากการคำนวณที่แล้วมาเราได้ $\hat{\beta}_2 = -1.578$, $S_{\hat{\beta}_2} = 0.176$

$$T = \hat{\beta}_2/S_{\hat{\beta}_2} = -1.578/(0.176) = -8.965$$

ซึ่งอยู่นอกเขตยอมรับ ฉะนั้นจึงปฏิเสธสมมติฐานที่ว่าไม่มีความสัมพันธ์ระหว่าง X และ Y ในกรณีเช่นนี้ เราอาจต้องการใช้สมมติฐานรองทางด้านเดียว เพื่อดูค่าของ β_2 เพราะถ้าค่าของ β_2 เป็นบวก

ก็จะขัดกับหลักพื้นฐานทางทฤษฎีเศรษฐศาสตร์ ดังนั้นควรจะใช้การทดสอบแบบเฉียงขาด (Sharper Test) ต่อสมมติฐานรองด้านเดียว นั่นคือ

$$Ha : \beta_2 < 0$$

และเขตยอมรับจะเป็น $-t_{\alpha/2}^{(n-2)} < T = \hat{\beta}_2 / S_{\hat{\beta}_2} < t_{\alpha/2}^{(n-2)}$

หรือ $-2.764 < T < 2.764$

โดยที่-8.965 อยู่นอกบริเวณเขตยอมรับนี้ คำตัดสินใจที่จะปฏิเสธ Ho จึงไม่เปลี่ยนแปลง

9.9 การวิเคราะห์ความแปรปรวนกับการถดถอย (Analysis of Variance of Regression)

สำหรับสมมติฐานที่ว่า "ไม่มีความสัมพันธ์ระหว่าง X และ Y" นั้นสามารถทดสอบได้โดยการใช้ตัวสถิติทดสอบอื่นที่ต่างจากตัวสถิติทดสอบ T เราจะเห็นว่าถ้าสมมติฐาน Ho เป็นจริงแล้ว ความผันแปรของ Y จากค่าสังเกตหนึ่งต่อค่าสังเกตหนึ่งจะไม่มีผลกระทบต่อการเปลี่ยนแปลงของ X แต่ผันแปรเพราะตัวคลาดเคลื่อนเชิงสุ่มอย่างเดียว นี่ก็หมายความว่า ผลรวมกำลังสองเนื่องจากการถดถอย (SSR) มีค่าต่างจากศูนย์ก็เพราะเราสังเกตตัวอย่างไม่ใช่ประชากรทั้งหมด จากสูตรที่กล่าวมาแล้วเราได้

$$SSR = \hat{\beta}_2^2 \sum (X - \bar{X})^2$$

ถ้า $\beta_2 = 0$ ค่าของ SSR ในประชากรควรจะเป็นศูนย์ และจากความสัมพันธ์ที่ว่า

$$SST = SSR + SSE$$

นั้น เมื่อ SSR เป็นศูนย์ SST จึงเท่ากับ SSE ดังนั้นถ้าไม่มีความสัมพันธ์ระหว่าง X และ Y แล้ว อัตราส่วน SSR/SSE จะต่างจากศูนย์ก็เพียงเพราะการสุ่มตัวอย่างเท่านั้น เราสามารถแสดงให้เห็นได้ว่า ภายใต้สมมติฐาน Ho : $\beta_2 = 0$ ตัวสถิติ F

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

มีการแจกแจง F ที่มี df = 1, n-2

เขตยอมรับสมมติฐานที่ว่า ไม่มีความสัมพันธ์ระหว่าง X และ Y จะเป็น

$$F = \frac{SSR/1}{SSE/(n-2)} \leq F_{\alpha}^{(1,n-2)}$$

ในเมื่อ F(1, n-2) เป็นค่าของการแจกแจง F ที่มี df = 1, n-2 และสมนัยกับระดับนัยสำคัญ α

ตัวสถิติทดสอบ F จะเหมือนกับตัวสถิติทดสอบ 2 ทาง t ในแง่ที่ว่าการทดสอบทั้งสองจะให้คำตอบเช่นเดียวกัน ถ้าระดับนัยสำคัญและข้อมูลจากตัวอย่างเหมือนกัน ทั้งนี้ก็เพราะว่าในทาง

คณิตศาสตร์เราได้ว่าตัวสถิติทั้งสองมีความสัมพันธ์กันคือ

$$t^2 = F$$

ความแตกต่างของตัวสถิติทดสอบทั้งสองคือ ตัวสถิติทดสอบ F ใช้ประยุกต์กับตัวแบบถดถอยที่มีตัวแปรอธิบายได้มากกว่า 1 ตัว ในเมื่อสถิติทดสอบ t ใช้ได้กับ 1 ตัวเท่านั้น

ตัวอย่าง สำหรับตัวอย่างของส้มเราได้

$$SSR = 5602 \qquad SSE = 698$$

ดังนั้น

$$F = \frac{SSR/1}{SSE/(n-2)} = 5602/(698/10)$$

$$= 80.20$$

แต่ $F^{(1,10)}_{01}$ จากตารางมีค่าเท่ากับ 10.0 ฉะนั้นจึงปฏิเสธ Ho

จากการใช้ตัวสถิติทดสอบ F เราจะเห็นว่าเราใช้ความผันแปรจากแหล่งต่าง ๆ คือ เนื่องจากการถดถอย และจากความคลาดเคลื่อน หรือเนื่องจากความผันแปรที่อธิบายไม่ได้ เทคนิค เช่นนี้ทางสถิติเรียกว่า เทคนิคการวิเคราะห์ความแปรปรวน

จากตัวอย่างเราทำตารางวิเคราะห์ความแปรปรวนได้ดังนี้

แหล่งของความผันแปร

SOV	df	SS	MS	F
ตัวแปรอธิบายได้	k-1 =1	SSR=5602	MSR=5602	80.20
ความคลาดเคลื่อน	n-k=10	SSE=698	MSE=68.8	
ทั้งหมด	n-1=11	SST=6300		

นอกจากการทดสอบสมมติฐานเกี่ยวกับความสัมพันธ์ระหว่าง Y และ X ว่าจะยังคงมีอยู่ (Existence) หรือไม่นั้น เรายังสามารถดำเนินการทดสอบสำหรับบางค่าของสัมประสิทธิของการถดถอยได้อีก เช่นสมมติฐานที่ว่า β_1 เท่ากับศูนย์ หรือ Ho : $\beta_1 = 0$ ซึ่งก็เป็นสมมติฐานที่ว่าเส้นถดถอยประชากรผ่านจุดกำเนิดนั่นเอง ตัวสถิติทดสอบที่เหมาะสมควรเป็น

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

ซึ่งมีการแจกแจงแบบ t ที่มี df = n-2 หรือเราประสงค์จะทดสอบสมมติฐานที่ว่า β_2 เท่ากับค่าหนึ่งค่าใด β_{20} หรือ β_1 เท่ากับ β_{10} นั้นตัวสถิติทดสอบจะเป็น

$$(\hat{\beta}_2 - \beta_{20})/S_{\hat{\beta}_2} \sim t^{(n-2)}$$

$$(\hat{\beta}_1 - \beta_{10})/S_{\hat{\beta}_1} \sim t^{(n-2)}$$

9.10 การทำนาย (Prediction)

ตัวแบบถดถอยนอกจากจะใช้สำหรับประมาณค่าและทดสอบสมมติฐานแล้วเรายังใช้ในการทำนายอีกด้วย บ่อยครั้งเราสนใจทำนายหรือพยากรณ์ค่าของ Y ในเมื่อกำหนดค่าของ X ให้ ตัวอย่างเช่น ผู้จัดการร้านสนใจที่จะทราบว่าส้มเป็นจำนวนเท่าใดที่เขาคาดว่าจะขายได้ ถ้าตั้งราคาไว้ ณ ระดับหนึ่ง นั่นคือเมื่อกำหนดค่าของตัวแปรอธิบายได้ X_0 ให้ แล้วจะต้องทำนายค่าของ Y_0 โดยที่ Y_0 เป็นตัวแปรเชิงสุ่มมีค่ากระจายรอบจุดบนเส้นถดถอยของประชากร ซึ่งสอดคล้องกับ X_0 และเราไม่ทราบค่าของมันก่อนทำการทดลอง ถ้าเราทราบพารามิเตอร์แล้วตัวทำนาย (Predictor) ของ Y_0 จะเป็นค่าเฉลี่ยของมัน นั่นคือ $E(Y_0) = \beta_1 + \beta_2 X_0$ ซึ่งเป็นจุดหนึ่งบนเส้นถดถอยของประชากรตัวทำนายนี้จะดีที่สุด ในแง่ที่ว่าความแปรปรวนของ Y_0 รอบ $E(Y_0)$ น้อยกว่าความแปรปรวนรอบจุดอื่นใด ๆ ค่าของ Y_0 จะแจกแจงแบบปกติที่มีความแปรปรวน σ^2 ซึ่งเป็นความแปรปรวนอันสืบเนื่องมาจากตัวคลาดเคลื่อนที่ปรากฏอยู่ในตัวแบบถดถอยแต่ที่จริงแล้วเราไม่ทราบค่า $E(Y_0)$ เพราะเราไม่ทราบเส้นถดถอยของประชากร เราจึงต้องประมาณค่า สำหรับตัวประมาณค่านั้นคือจุดที่สอดคล้องกับเส้นถดถอยของตัวอย่าง $\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$ นี่นั่นเอง ในเมื่อ $\hat{\beta}_1$ และ $\hat{\beta}_2$ เป็นตัวประมาณที่ดีของ β_1 และ β_2 ค่าจริงของ Y_0 กับค่าทำนายของ Y_0 นั้นส่วนมากจะต่างกัน นั่นคือเกิดความคลาดเคลื่อน (error) ขึ้นทั้งนี้เนื่องจากเหตุ 2 ประการ คือ

(1) ตัวคลาดเคลื่อนเชิงสุ่ม (Random disturbance) ตัวคลาดเคลื่อนนี้จะทำให้ค่าของ Y_0 จะไม่เท่ากับ $E(Y_0)$ นั่นคือ Y_0 จะไม่อยู่บนเส้นถดถอยประชากร ความคลาดเคลื่อนที่เกิดจากตัวคลาดเคลื่อนนี้จะเป็นความคลาดเคลื่อนที่ติดแน่นประจำอยู่กับกลไก ต่าง ๆ ที่ทำให้เกิดค่าของตัวแปรตาม Y และไม่มีอะไรที่สามารถจะทำให้มันลดน้อยลงได้

(2)ความคลาดเคลื่อนของตัวอย่าง (Sampling Error) ความคลาดเคลื่อนนี้จะทำให้เส้นถดถอยตัวอย่างไม่เป็นเส้นเดียวกับเส้นถดถอยประชากร ความคลาดเคลื่อนแบบนี้จะลดลงถ้าเราเพิ่มความเที่ยงตรง (Precision) ในการประมาณเส้นถดถอยของประชากร โดยการเพิ่มขนาดของตัวอย่างให้มากขึ้น ความคลาดเคลื่อนของตัวอย่าง $E(Y_0) - \hat{Y}_0$ สามารถเขียนเป็นสมการที่สัมพันธ์กับความคลาดเคลื่อนอื่น ๆ ได้ดังนี้

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

Forecast Error = Random disturbance Error + Sampling Error

ความแตกต่างระหว่างค่าจริง (Actual value) ของ Y_0 และค่าทำนาย $\hat{Y}_0$ เรียกว่าความคลาดเคลื่อนของการพยากรณ์ (Forecast error) ซึ่งมีความสัมพันธ์กับความคลาดเคลื่อนอื่น ๆ ดังข้างบน และเราจะเห็นได้ว่าความคลาดเคลื่อนของการพยากรณ์ $Y_0 - \hat{Y}_0$ เขียนได้เป็น

$$Y_0 - \hat{Y}_0 = (\beta_1 + \beta_1 X_0 + \varepsilon_0) - (\hat{\beta}_1 + \hat{\beta}_2 X_0)$$

ซึ่งเป็นตัวผสมเชิงเส้น (Linear Combination) ของตัวแปรเชิงสุ่ม $\varepsilon_0, \varepsilon_1, \varepsilon_2 \ldots \varepsilon_n$ ที่เป็นอิสระกัน และมีการแจกแจงแบบปกติ ดังนั้น ความคลาดเคลื่อนของการพยากรณ์ จึงเป็นตัวแปรเชิงสุ่มที่แจกแจงปกติด้วย สำหรับค่าเฉลี่ยและความแปรปรวนพิจารณาได้ดังนี้

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

$$\begin{aligned}\text{ความแปรปรวน } V(Y_0-\hat{Y}_0) &= E((Y_0-\hat{Y}_0)-E(Y_0-\hat{Y}_0))^2 \\ &= E(Y_0-\hat{Y}_0)^2 \quad ; \quad E(Y_0-\hat{Y}_0) = 0 \\ &= E(Y_0-E(Y_0))^2+E(E(Y_0)-Y_0)^2 \\ \sigma^2_{f,} &= \sigma^2 + \sigma^2_{\hat{Y}_0}\end{aligned}$$

นั่นคือความแปรปรวนของความคลาดเคลื่อนของการพยากรณ์ σ^2 (Variance of forecast error) ประกอบด้วย 2 ส่วน คือ (1) ความแปรปรวนของตัวคลาดเคลื่อนเชิงสุ่ม σ^2 และ (2) ความแปรปรวนของตัวพยากรณ์ $\hat{Y}_0$ รอบตัวกลาง $E(Y_0)$ ของมันหรือเป็นความแปรปรวนของความคลาดเคลื่อนจากตัวอย่างนั่นเอง $\sigma^2_{\hat{Y}_0}$

$$\text{จาก} \quad \sigma^2_{\hat{Y}_0} = \sigma^2\left(1/n + \frac{(X_0-\bar{X})^2}{\sum(X-\bar{X})^2}\right)$$

ที่กล่าวมาแล้ว เราจะได้สูตร ความแปรปรวนของความคลาดเคลื่อนของการพยากรณ์ดังนี้

$$\begin{aligned}\sigma^2_f &= \sigma^2 + \sigma^2_{\hat{Y}_0} \\ &= \sigma^2 + \sigma^2(1/n+ (X_0-\bar{X})^2/\Sigma(X-\bar{X})^2) \\ &= \sigma^2(1+1/n+(X_0-\bar{X})^2/\Sigma(X-\bar{X})^2)\end{aligned}$$

เราจะเห็นได้ว่าความแปรปรวนของความคลาดเคลื่อนของการพยากรณ์จะน้อยลงได้ ถ้า

(1) ขนาดของตัวอย่าง n โตขึ้น

(2) การกระจัดกระจาย (Dispersion) ของตัวแปรอธิบายได้ X ในตัวอย่างมีมากนั่นคือ $\Sigma (X-\bar{X})^2$ โตขึ้น

(3) ระยะห่างระหว่าง X_0 กับค่าเฉลี่ยของตัวอย่าง $\bar{X}$ น้อยลง

สองข้อแรกนี้จะแสดงถึงความจริงที่ว่าค่าประมาณของเส้นถดถอยประชากรจะยิ่งดีขึ้นถ้าความแปรปรวนของความคลาดเคลื่อนของการพยากรณ์ยิ่งน้อย

แต่ข้อที่สามน่าสนใจเพราะมีความหมายว่าการพยากรณ์ จะดีขึ้นถ้าค่าของ X ใกล้ $\bar{X}$ มากขึ้นอันนี้ยังเป็นสิ่งที่คงเส้นคงวากับข้อยืนยันที่มีเหตุผลว่าเราสามารถพยากรณ์ได้ภายในช่วงหรือ

พิสัยของประสบการณ์ซึ่งเป็นค่าของตัวอย่างของตัวแปรอธิบายได้ X มากกว่าภายนอกช่วงประสบการณ์ และจุดกลางของพิสัยก็คือ $\bar{X}$ ดังนั้นความเชื่อมั่นของการพยากรณ์จะยิ่งน้อยลงถ้าเราพยากรณ์จุดที่ห่างจากจุดกลางของพิสัย X มากเท่าไร

โดยทั่วไป σ^2_f จะไม่ทราบจึงต้องประมาณเอา และทำได้โดยแทน σ^2 ด้วยตัวประมาณของมันคือ S^2 แล้วเราจะได้ตัวประมาณของ σ^2_f ที่ไม่เอียงเฉ คงเส้นคงวา และมีประสิทธิภาพแบบ Asymptotic และเราจะเรียกว่า S^2_f ซึ่งจะได้สูตรดังนี้

$$S_f^2 = S^2(1+1/n +(X_o-\bar{X})^2/\Sigma(X-\bar{X})^2)$$

โดยสรุปเราพบว่า ความคลาดเคลื่อนของการพยากรณ์ $Y_0 - \hat{Y}_0$ มีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเป็นศูนย์ และความแปรปรวนเป็น σ^2_f นั่นคือ

$$(Y_o-\hat{Y}_o) \sim N(o,\ \sigma_f^2)$$

หรือ $$(Y_o-\hat{Y}_o)/\sigma_f \sim N(0,1)$$

เมื่อแทน σ_f ด้วย S_f จะได้

$$(Y_o-\hat{Y}_o)/S_f \sim t^{(n-2)}$$

ผลสุดท้ายทำให้เรากล่าวข้อความเกี่ยวกับการพยากรณ์ที่มีความน่าจะเป็นกำกับด้วยได้ นั่นคือเราสามารถสร้างช่วงที่ครอบคลุมค่าจริงของ Y_0 ตามความน่าจะเป็นที่กำหนดให้ได้ ถ้าให้ระดับของความน่าจะเป็นนั้นเป็น $1-\alpha$ เมื่อ α เป็นค่าใด ๆ ที่กำหนดให้ระหว่าง 0 กับ 1 แล้วเราจะสามารถสร้างช่วงเชื่อมั่น $100(1-\alpha)$ % ของ Y_0 ได้เป็น

$$\hat{Y}_o - t^{(n-2)}_{\alpha/2}\ S_f < Y_o < \hat{Y}_o + t^{(n-2)}_{\alpha/2}\ S_f$$

ช่วงนี้สมมาตรรอบ ๆ ตัวพยากรณ์ Y_0 และคาดว่าจะครอบคลุมค่าจริงของ Y_0 ด้วยความน่าจะเป็น $(1-\alpha)$

ตัวอย่าง สมมติว่าเราต้องการทำนายอุปสงค์ของเส้น ณ ราคา 110 กีบต่อกิโล

โดยการใช้เส้นอุปสงค์ที่ประมาณได้ เราจะทำนายปริมาณอุปสงค์ของส้ม ณ ราคา 110 กีบ ได้เป็น

$$Y_o = 210.460+1.578(110)$$
$$= 36.88$$

ซึ่งมีความคลาดเคลื่อนมาตรฐาน

$$S_f = 69.8(1+1/12+(110-70)^2/2250$$
$$= 11.20$$

ช่วงความเชื่อมั่น 95 % ของ Y_0 คือ

$$Y_o = 36.88 + 2.228(11.20)$$
$$= 12,62$$

9.11 การนำเสนอผลของการถดถอย (Presentation of Regression Results)

การนำเสนอผลของการวิเคราะห์ถดถอยที่ได้จากข้อมูลตัวอย่าง (Sample Data) จะเกี่ยวข้องกับการประมาณสัมประสิทธิ์ของการถดถอย (Regression Coefficients) และความคลาดเคลื่อนมาตรฐานของตัวประมาณเหล่านี้ ผลที่นำเสนอนี้จะรวมด้วยกับสัมประสิทธิ์การตัดสินใจ (Coefficient of Determination, R^2) ซึ่งเป็นตัวชี้ให้เห็นว่าเส้นถดถอยตัวอย่างจะปรับ (fit) เข้าค่าสังเกตได้ดีเพียงไร

วิธีการนำเสนอผลนั้นจะประกอบด้วย สมการถดถอยตัวอย่างที่ประมาณได้ และมีความคลาดเคลื่อนมาตรฐานที่ประมาณได้ในวงเล็บใต้สัมประสิทธิ์นั้น ๆ ตามลำดับ และจะตามด้วยค่า R^2 นั่นคือ

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$
$$(S_{\hat{\beta}_1})(S_{\hat{\beta}_2}) \qquad R^2 = \ldots$$

ตามตัวอย่างเกี่ยวกับอุปสงค์ของส้ม เราเสนอผลของการถดถอยได้เป็น

$$\hat{Y}_i = 210.460 + 1.578(X_i)$$
$$(12.563) \quad (0.176) \qquad R^2 = 0.889$$

เครื่องหมาย ^ (hat) บน Y_i แสดงให้ทราบว่า สมการนี้เป็นจริงกับค่าที่ปรับได้ (fitted value) ของตัวแปรตาม ไม่ใช่กับค่าสังเกตที่แท้จริง (actually observed value)

เราเสนอผลได้อีกรูปหนึ่ง คือ

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$$
$$(S_{\hat{\beta}_1}) \quad (S_{\hat{\beta}_2}) \qquad R^2 = \ldots$$

ในเมื่อ e_i (ไม่ใช่ ε_i) เป็นตัวเศษแบบกำลังต่ำสุด (Least Squares Residuals) ถ้าจะเขียนว่า $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ จะไม่ถูกต้อง

นักสถิติบางคนเสนอเป็นแบบค่า $t = \hat{\beta}_1 / S_{\hat{\beta}_1}$ ซึ่งก็เป็นที่ยอมรับกันด้วยดังนี้

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \qquad R^2 = \ldots.$$
$$(\hat{\beta}_1/S_{\hat{\beta}_1}) \quad (\hat{\beta}_2/S_{\hat{\beta}_2})$$

หรือ
$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e \qquad R^2 = \ldots.$$
$$(\hat{\beta}_1/S_{\hat{\beta}_1}) \quad (\hat{\beta}_2/S_{\hat{\beta}_2})$$

จากตัวอย่างเราจึงเสนอได้เป็น $\hat{Y}_i = 210.460 - 1.578(X_i)$

$= (16.752) \quad (8.966) \qquad R^2 = 0.889$

หรือ $Y_i = 210.460 - 1.578(X_i) + e_i$

$(16.752) \quad (8.966) \qquad R^2 = 0.889$

9.12 ตัวแบบถดถอยเชิงซ้อน (Mutiple Regression Model)

ในตัวแบบถดถอยเชิงเดี่ยวที่กล่าวมาจะใช้ได้เมื่อตัวแปรอิสระมีเพียงตัวเดียว ถ้าตัวแบบประกอบด้วยตัวแปรอิสระมากกว่า 1 ตัวตัวแบบนั้นได้ชื่อว่า ตัวแบบถดถอยเชิงซ้อน (Multiple Regression Model) ในทางเศรษฐศาสตร์ความสัมพันธ์ส่วนมากอธิบายได้ด้วยตัวแบบถดถอยเชิงซ้อนเช่น ฟังก์ชั่นการผลิตนั้นผลผลิต (Output) จะเป็นฟังก์ชั่นของทรัพยากร (Input) หลายอย่าง ฟังก์ชั่นการบริโภคนั้นตัวแปรตามได้รับอิทธิพลจากรายได้และองค์ประกอบอื่น ๆ และฟังก์ชั่นอุปสงค์มีตัวแปรอิสระเป็นราคาของผลิตภัณฑ์ราคาสินค้าที่ใช้ทดแทน และรายได้ เป็นต้น

ตัวแบบถดถอยเชิงซ้อนที่แสดงถึงความสัมพันธ์ระหว่างตัวแปรนี้จะเป็นรูปทั่วไปของตัวแบบถดถอยเชิงเดี่ยว นั่นคือแสดงว่าการเปลี่ยนแปลงในตัวแปรหนึ่งนั้นสามารถอธิบายได้ด้วยการเปลี่ยนแปลงของตัวแปรอื่น ๆ หลายตัว ความสัมพันธ์เช่นนี้อธิบายได้ในแบบง่าย ๆ ของสมการถดถอยเชิงเส้นแบบเชิงซ้อน ดังรูป ฟอร์มดังนี้

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots. + \beta_k X_{ik} + \varepsilon_i$$

ในเมื่อ Y แทนตัวแปรตาม $X_2, X_3, \ldots\ldots, X_k$ เป็นตัวแปรอิสระซึ่งมีอยู่ (k-1) ตัว ถ้า k = 2 สมการนี้จะเป็นสมการถดถอยเชิงเดี่ยว เป็นความคลาดเคลื่อนเชิงสุ่ม
สมการข้างบนนี้เราอาจจะเขียนได้เป็น

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots. + \beta_k X_{ik} + \varepsilon_i$$

ในเมื่อ X_{i1} = สำหรับทุก ๆ i = 1, 2, ..., n

ข้อกำหนด (Specification) ของตัวแบบถดถอยเชิงเส้นแบบเชิงซ้อน นั้นต้องรวมถึง

ก. สมการถดถอยเชิงเส้นแบบเชิงซ้อนที่กล่าวมา คือ

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

ข. ข้อสมมติเบื้องต้น (Basic Assumptions) ดังนี้

1. ε_i แจกแจงแบบปกติ

2 $E(\varepsilon_i) = 0$

3. $E(\varepsilon_i^2) = \sigma^2$

4. $E(\varepsilon_j \varepsilon_i) = 0, (i \neq j)$

5. ตัวแปรอิสระแต่ละตัวไม่เชิงสุ่ม (Nonstochastic) และมีค่าคงที่ในตัวอย่างที่ซ้ำ ๆ กัน ได้กับต้องให้ $\sum (X_{ia} - \bar{X}_a)^2/n$ เป็นเลขจำนวนจำกัดไม่เป็นศูนย์ สำหรับตัวอย่างขนาด n และ a = 2, 3,....., k

6. จำนวนค่าสังเกตต้องมากกว่าจำนวนสัมประสิทธิ์ของการถดถอยที่ต้องประมาณค่านั่นคือ $n > k$

7. ตัวแปรอิสระใด ๆ จะไม่มีความสัมพันธ์เชิงเส้นกัน

ข้อสมมติ 1 ถึง 5 เป็นเช่นเดียวกับตัวแบบถดถอยเชิงเดี่ยวแต่ข้อสมมติ 6 เพิ่มขึ้นมาให้มีจำนวนองศาแห่งความเป็นอิสระเพียงพอในการประมาณค่า และข้อสมมติ 7 กล่าวว่าไม่มีตัวแปรอิสระใด มีสหสัมพันธ์สมบูรณ์ (Perfectly Correlated) กับตัวแปรอิสระอื่นใด หรือมีการรวมเชิงเส้น (Linear Combination) ใด ๆ ของตัวแปรอิสระอื่น ๆ ข้อสมมตินี้มีความจำเป็นสำหรับการประมาณค่า

จากข้อกำหนดของตัวแบบถดถอยเชิงซ้อนที่กำหนดมานั้น เราจะได้ว่า Y_i มีการแจกแจงปกติเช่นเดียวกับตัวแบบถดถอยเชิงเดี่ยว สำหรับค่าเฉลี่ยและความแปรปรวนของมันคือ

$$E(Y_i) = \beta_1 + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$
$$V(Y_i) = E(Y_i - E(Y_i))^2 = \sigma^2$$

สำหรับสัมประสิทธิ์การถดถอย $\beta_1, \beta_2, \ldots \beta_k$ สามารถแปลความหมายได้ดังนี้

(1) β_1 เป็นค่าเฉลี่ยของ Y เมื่อตัวแปรอิสระทุกตัวเท่ากับศูนย์

(2) β_a (a = 2, 3,....., k) เป็นการเปลี่ยนแปลงของ E(Y) ที่สอดคล้องกับการเปลี่ยนแปลงหนึ่งหน่วยของตัวแปรอิสระตัวที่ k โดยให้ตัวแปรอิสระที่เหลือคงที่ นั่นคือ

$$\beta_a = \frac{\delta E(Y_i)}{\delta X_{ia}} \quad ; \quad a = 2, 3, \ldots k$$

และ β_a จะวัดผลกระทบ (Effect) ของ X_{ik} ต่อ $E(Y_i)$ เมื่อ $X_{i2}, X_{i3}, \ldots\ldots, X_{i,k-1}$ คงที่ β_1 บางทีเรียกว่า จุดตัดแกน (Intercept) หรือค่าคงที่ของการถดถอย (Regression Constant) และ $\beta_2, \beta_3, \ldots\ldots, \beta_k$ เป็นความชันของการถดถอย (Regression Slopes) หรือ สัมประสิทธิ์การถดถอยบางส่วน (Partial Regression Coefficients)

9.13 การประมาณค่าของพารามิเตอร์ถดถอย (Estimation of Regression Parameters)

9.13.1 การประมาณแบบกำลังสองต่ำสุด (Least Squares Estimation)

สำหรับตัวประมาณค่าแบบกำลังสองต่ำสุดของสัมประสิทธิ์การถดถอยนั้นพัฒนามาจากการทำให้ผลรวมกำลังสอง So น้อยที่สุด ในเมื่อ

$$S_o = \sum_i (Y_i - \beta_1 - \beta_2 X_{i2} - \ldots - \beta_k X_{ik})^2$$
$$= \sum \varepsilon_i$$

จากแคลลูลัสเราหาอนุพันธ์ของ So เทียบกับ $\beta_1, \beta_2, \ldots\ldots, \beta_k$ แล้วเทียบกับศูนย์ นั่นคือ

$$\frac{\delta S_o}{\delta \beta_a} = 0; \quad a = 1, 2, \ldots k$$

แล้วเราจะได้สมการปกติแบบกำลังสองน้อยสุด ดังนี้

$$\sum Y_i = \beta_1 n + \beta_2 \sum X_{i2} + \ldots + \beta_k \sum X_{ik}$$
$$\sum X_{i2} Y_i = \beta_1 \sum X_{i1} + \beta_2 \sum X_{i2}^2 + \ldots + \beta_k \sum X_{i2} X_{ik}$$
$$\vdots$$
$$\sum X_{ik} Y_i = \beta_1 \sum X_{ik} + \beta_2 \sum X_{i2} X_{ik} + \ldots + \beta_k \sum X_{ik}^2$$

สำหรับสมการปกติเหล่านี้เรามีวิธีการจดจำหรือวิธีเขียน ได้ดังนี้

จาก $\quad Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{i2} + \ldots + \hat{\beta}_k X_{ik} \qquad (ก)$

เราได้สมการที่ (1) จากการหาผลรวม (Summation) ของสมการ นั่นคือ

$$\sum Y_i = \hat{\beta}_1 n + \hat{\beta}_2 \sum X_{i2} + \ldots + \hat{\beta}_k \sum X_{ik}$$

(2) คูณสมการ (ก) ด้วย X_{i2} แล้วหาผลรวม

$$\sum X_{i2} Y_i = \hat{\beta}_1 \sum X_{i2} + \hat{\beta}_2 \sum X_{i2}^2 + \ldots + \hat{\beta}_k \sum X_{i2} X_{ik}$$

(3) คูณสมการ (ก) ด้วย X_{i3} แล้วหาผลรวม

$$\sum X_{i3}Y_i = \hat{\beta}_1\sum X_{i3} + \hat{\beta}_2\sum X_{i2}X_{i3} + \ldots + \hat{\beta}_k\sum X_{i3}X_{ik}$$

(k) คูณสมการ (ก) ด้วย X_{ik} แล้วหาผลรวม

$$\sum X_{ik}Y_i = \hat{\beta}_1\sum X_{ik} + \hat{\beta}_2\sum X_{i2}X_{ik} + \ldots + \hat{\beta}_k\sum X_{ik}^2$$

จากสมการปกตินี้เราจะเห็นได้ว่าในการแก้สมการเพื่อหา $\hat{\beta}$ ต่าง ๆ ที่ไม่ทราบค่า นั้นจะทำไม่ได้ถ้า

(1) จำนวนตัวแปรอิสระรวมกับ 1 (นั่นคือเท่ากับ k) มากกว่าจำนวนค่าสังเกต n หรือ

(2) ตัวแปรอิสระตัวหนึ่งตัวใดเป็นการรวมเชิงเส้นที่แท้จริงของตัวแปรอิสระตัวอื่น ๆ

ในกรณีแรกนั้น คือ จำนวนสมการจะน้อยกว่าจำนวนตัวที่ไม่ทราบค่า ส่วนกรณีหลังสมการเหล่านั้นจะไม่เป็นอิสระกัน ในสมการปกติแบบกำลังสองต่ำสุดนั้น จากสมการแรกเราได้

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X}_2 - \hat{\beta}_3\bar{X}_3 - \ldots - \hat{\beta}_k\bar{X}_k$$

ในเมื่อ $Y = \sum Y_i/n, \bar{X}_a = \sum X_{ia}/n; a = 2, 3, \ldots, k$

โดยการแทน $\hat{\beta}_1$ ในสมการปกติที่เหลือเราจะได้

$$\begin{aligned}
m_{y2} &= m_{22}\hat{\beta}_2 + m_{23}\hat{\beta}_3 + \ldots + m_{2k}\hat{\beta}_k \\
m_{y3} &= m_{23}\hat{\beta}_2 + m_{33}\hat{\beta}_3 + \ldots + m_{3k}\hat{\beta}_k \\
&\vdots \\
m_{yk} &= m_{2k}\hat{\beta}_2 + m_{3k}\hat{\beta}_3 + \ldots + m_{kk}\hat{\beta}_k
\end{aligned}$$

ในเมื่อ $m_{ya} = \sum(Y_i - \bar{Y})(X_{ia} - \bar{X}_a)$

และ $m_{ja} = \sum(X_{ij} - \bar{X}_j)(X_{ia} - \bar{X}_a) \qquad j, a = 2, 3, \ldots, k$

จากสมการเหล่านี้สามารถหาค่า $\hat{\beta}_2, \hat{\beta}_3, \ldots, \hat{\beta}_k$ ได้โดยวิธีแก้สมการ แต่ยุ่งยากหน่อย

สำหรับกรณีที่มีตัวแปรอิสระ 2 ตัว (k = 3) นั่นคือ ตัวแบบถดถอยจะเป็น

$$Y_i = \beta_1 + \beta_2X_{i2} + \beta_3X_{i3} + \varepsilon_i$$

เราจะได้ค่าประมาณของ $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ ดังนี้

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X}_2 - \hat{\beta}_3\bar{X}_3$$

$$= \frac{m_{y2}m_{33} - m_{y3}m_{23}}{m_{22}m_{33} - m_{23}^2} \quad ; \quad \hat{\beta}_3 = \frac{m_{y3}m_{22} - m_{y2}m_{23}}{m_{22}m_{33} - m_{23}^2}$$

ตัวอย่าง ในการศึกษาเกี่ยวกับสัมนั้นถ้าปริมาณที่ขายได้ขึ้นอยู่กับราคาและจำนวนค่าใช้จ่ายในการโฆษณาสินค้า นั่นคือ ตัวแบบถดถอยเป็น

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

ในเมื่อ Y_i แทนจำนวนส้มที่ขายได้ (กิโล) X_{i2} แทนราคา (กีบ) ต่อกิโล และ X_{i3} แทนค่าโฆษณา (10,000 กีบ)

ข้อมูลที่ศึกษาได้จะเป็น

ปริมาณที่ขายได้	ราคา	ค่าโฆษณา
55	100	5.50
70	90	6.30
90	80	7.20
100	70	7.00
90	70	6.30
105	70	7.35
80	70	5.60
110	65	7.15
125	60	7.50
115	60	6.90
130	55	7.15
130	50	6.50

ผลสรุปของข้อมูลเป็นดังนี้

$\bar{Y} = 100$ $\quad m_{22} = 2250$ $\quad m_{y2} = -3550$

$\bar{X}_2 = 70$ $\quad m_{33} = 4.86$ $\quad m_{y3} = 12520$

$\bar{X}_3 = 6.70$ $\quad m_{23} = -54$ $\quad m_{yy} = 6300$

ค่าประมาณของสัมประสิทธิ์การถดถอย คือ

$$\hat{\beta}_2 = \frac{(-3550)(4.86) - (-54)(125.25)}{2250(4.86) - (-54)^2} = -1.326$$

$$\hat{\beta}_3 = \frac{2250(125.20) - (-54)(-3550)}{2250(4.86)-(-54)^2} = 11.237$$

$$\hat{\beta}_1 = 100-(-1.326)(70) - 11.237(6.7) = 117.532$$

ดังนั้นสมการถดถอยที่ประมาณได้ คือ

$$Y_i = 117.532-1.326X_{i2} + 11.237X_{i3} + e_i$$

สมการนี้แสดงว่าราคาของส้มลดลง 10 กีบ (เมื่อค่าใช้จ่ายโฆษณาไม่เปลี่ยนแปลง) จะเพิ่มปริมาณขายประมาณ 13 กิโล และถ้าเพิ่มค่าโฆษณา 10,000 กีบ (เมื่อราคาไม่เปลี่ยนแปลง) จะเพิ่มปริมาณขายประมาณ 11 กิโล

ตัวแบบถดถอยเชิงเส้นแบบเชิงซ้อน โดยปกติเสนอในสัญญนิยมเมทริกซ์ (Matrix Notation) ได้ดังนี้

ก. สมการ $\underline{Y} = \underline{X}\underline{B} + \underline{\varepsilon}$

ในเมื่อ

$$\underline{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \underline{X} = \begin{pmatrix} X_{11} & X_{12} \cdots X_{1k} \\ X_{21} & X_{22} \cdots X_{2k} \\ \vdots & \vdots \\ X_{n1} & X_{n2} \cdots X_{nk} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} \quad \underline{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

ข. ข้อสมมติเบื้องต้น

1. $\underline{\varepsilon} \sim N(\underline{0}, \underline{\sum})$
2. $\underline{\sum} = \sigma^2 I_n$

3. สมาชิกของ X เป็นแบบคงที่ (Nonstochastic) ซึ่งมีค่าคงที่ในตัวอย่างซ้ำ ๆ และ $(1/x)'(\underline{X}'\underline{X})$ เป็น Nonsingular โดยที่สมาชิกของมันจำกัด (Finite) สำหรับขนาดตัวอย่างใด ๆ

ในเมื่อ $\underline{0} = \begin{pmatrix} 0 \\ 0 \\ \vdots \end{pmatrix}$ $I_n = \begin{pmatrix} 1 & 0 & 0\ldots & 0 \\ 0 & 1 & 0\ldots & 0 \\ & & \vdots & \\ 0 & 0 & 0\ldots & 1 \end{pmatrix}$

สมการปกติแบบกำลังสองน้อยสุด คือ

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

นั่นคือเราจะได้

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

ในเมื่อ

$$\underline{X}'\underline{Y} = \begin{pmatrix} \sum Y_i \\ \sum X_{i2}Y_i \\ \vdots \\ \sum X_{ik}Y_i \end{pmatrix} \qquad (\underline{X}'\underline{X}) = \begin{pmatrix} n & \sum X_{i2} & \cdots & \sum X_{ik} \\ \sum X_{i2} & \sum X_{i2}^2 & \cdots & \sum X_{i2}X_{ik} \\ \vdots & \vdots & & \\ \sum X_{ik} & \sum X_{i2}X_{ik} & & \sum X_{ik}^2 \end{pmatrix}$$

$$\hat{\underline{\beta}}' = (\hat{\beta}_1 \quad \hat{\beta}_2 \cdots \hat{\beta}_k)$$

หรือสมการปกติในรูปที่แทนด้วย $\hat{\underline{\beta}}$ คือ

$$(\underset{\sim}{X}'\underset{\sim}{Y}) = (\underset{\sim}{X}'\underset{\sim}{X})\hat{\underline{\beta}}$$

ในเมื่อ

$$\hat{\underline{\beta}} = \begin{pmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} \qquad (\underset{\sim}{X}'\underset{\sim}{X}) = \begin{pmatrix} m_{22} & m_{23} & \cdots & m_{2k} \\ m_{23} & m_{33} & \cdots & m_{3k} \\ & \vdots & & \\ m_{2k} & m_{3k} & \cdots & m_{kk} \end{pmatrix} \qquad (\underset{\sim}{X}'\underset{\sim}{Y}) = \begin{pmatrix} m_{y2} \\ m_{y3} \\ \vdots \\ m_{yk} \end{pmatrix}$$

9.13.2 การประมาณค่าแบบไม่เอียงเฉเชิงเส้นดีที่สุด (Best Linear Unbiased Estimation)

ตัวประมาณแบบไม่เอียงเฉเชิงเส้นดีที่สุด (BLUE) สำหรับตัวแบบถดถอยเชิงซ้อนที่มีตัวแปรอิสระ 2 ตัว นั้นมีวิธีการเช่นเดียวกับตัวแบบถดถอยเชิงเดี่ยวนั่นเอง สมมติว่าเราจะหา BLUE ของ β_2 ในตัวแบบ

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

จากเงื่อนไขความเป็นเชิงเส้นเราจะมี $\tilde{\beta}_2 = \sum a_i Y_i$

ในเมื่อ a_i $(i = 1,2,...,n)$ เป็นค่าคงที่ ค่าเฉลี่ยของ $\tilde{\beta}_2$ คือ

$$E(\tilde{\beta}_2) = E\sum a_i Y_i = E\sum a_i(\beta_1+\beta_2 X_{i2}+\beta_3 X_{i3}+\varepsilon_i)$$
$$= \beta_1\sum a_i+\beta_2\sum a_i X_{i2}+\beta_3\sum a_i X_{i3}$$

$\tilde{\beta}_2$ จะไม่เอียงเฉถ้า

$$\sum a_i = 0 \ , \ \sum a_i X_{i2} = 1 \ , \ \sum a_i X_{i3} = 0$$

ความแปรปรวนของ $\tilde{\beta}_2$ คือ

$$V(\tilde{\beta}_2) = E(\sum a_i Y_i - E(\sum a_i Y_i))^2$$
$$= E\sum(a_i \varepsilon_i)^2 = \sigma^2\sum a_i^2$$

ซึ่งเราจะหาค่าของ a ที่จะทำให้ $\sigma^2\Sigma a^2$ มีค่าต่ำสุดโดยมีเงื่อนไขว่า

$$\sum a_i = 0, \ \sum a_i X_{i2} = 1, \ \sum a_i X_{i3} = 0$$

จากวิธี Lagrange Multiplier ได้สมการ

$$H = \sigma^2\sum a_i^2 - \lambda_1(\sum a_i) - \lambda_2(\sum a_i X_{i2} - 1) - \lambda_3(\sum a_i X_{i3})$$

หาอนุพันธ์ H เทียบกับ $a_1, a_2,...,a_i, \lambda_1, \lambda_2, \lambda_3$ และให้แต่ละอนุพันธ์เท่ากับศูนย์เราจะได้สมการที่จะหาค่า $a_1, a_2,...,\lambda_1, \lambda_2, \lambda_3$ ถึง n + 3 สมการแล้วเราจะได้ว่า

ดังนั้น
$$a_i = (-\bar{X}_2 m_{33} + \bar{X}_3 m_{23} + m_{33}X_{i2} - m_{23}X_{i3})/(m_{22}m_{33} - m_{23}^2)$$
$$\tilde{\beta}_2 = \sum a_i Y_i = (m_{y2}m_{33} - m_{y3}m_{23})/(m_{22}m_{33} - m_{23}^2)$$

นั่นคือได้เช่นเดียวกับตัวประมาณค่าแบบกำลังสองต่ำสุดของ β_2 และในทำนองเดียวกันเราสามารถแสดงว่า $\tilde{\beta}_3 = \hat{\beta}_3$ และ $\tilde{\beta}_1 = \hat{\beta}_1$ เพราะฉะนั้นตัวประมาณค่าแบบกำลังสองต่ำสุดของสัมประสิทธิ์ถดถอยจะเป็น BLUE สำหรับตัวแบบถดถอยที่มีตัวแปรอิสระจำนวนหนึ่งก็สรุปได้เช่นเดียวกัน วิธี BLUE จะใช้สูตรของความแปรปรวนของตัวประมาณค่า β สำหรับ $\tilde{\beta}_2, \tilde{\beta}_3$ เราจะได้

$$V(\tilde{\beta}_2) = \sigma^2\sum a_i^2$$
$$= \sigma^2 m_{33}/(m_{22}m_{33} - m_{23}^2)$$

$$V(\tilde{\beta}_3) = \sigma^2 m_{22}/(m_{22}m_{33} - m_{23}^2)$$

9.13.3 การประมาณค่าแบบน่าจะเป็นมากสุด (Maximum Likelihood Estimation)

ตัวประมาณค่าแบบน่าจะเป็นมากสุดของพารามิเตอร์ในตัวแบบถดถอยเชิงเส้นแบบเชิงซ้อนนั้นพัฒนามาได้เช่นเดียวกับตัวแบบถดถอยเชิงเดี่ยว นั่นคือจากฟังก์ชั่นน่าจะเป็น

$$L = -(n/2)\log(2\pi) - (n/2)\log\sigma^2 - (1/2\sigma^2)\sum(Y_i - \beta_1 - \beta_2 X_{i2} - \cdots - \beta_k X_{ik})$$

เราหาอนุพันธ์ L โดยเทียบกับ $\beta_1, \beta_2, \ldots, \beta_k$ และ σ^2 และให้แต่อนุพันธ์เท่ากับศูนย์ซึ่งจะได้สมการปกติแบบน่าจะเป็นมากสุด $k + 1$ สมการ ในสมการแรกเราจะได้ค่าประมาณของ β_1 เช่นเดียวกับวิธีกำลังสองต่ำสุด ส่วนสมการที่ $k + 1$ จะได้ตัวประมาณค่าแบบน่าจะเป็นมากสุดของ σ^2 คือ

$$\hat{\sigma}^2 = (1/n)\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{i2} - \cdots - \hat{\beta}_k X_{ik})^2 = (1/n)\sum e_i^2$$

9.14 สถิติอนุมานในการถดถอยเชิงซ้อน

(1) การประมาณค่าของความแปรปรวนของตัวประมาณค่ากำลังสองต่ำสุดในกรณีที่มีตัวแปรอิสระ $k - 1$ ตัวจะมีความแปรปรวนดังนี้

$$\sum(\underline{\hat{\beta}}) = E(\underline{\hat{\beta}} - \underline{\beta})(\underline{\hat{\beta}} - \underline{\beta}) = \sigma^2(X'X)$$

$$= \begin{bmatrix} V(\hat{\beta}_1) & \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) & \cdots & \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) \\ \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) & V(\hat{\beta}_2) & \cdots & \mathrm{Cov}(\hat{\beta}_2, \hat{\beta}_k) \\ & \vdots & & \\ \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_k) & \mathrm{Cov}(\hat{\beta}_2, \hat{\beta}_k) & \cdots & V(\hat{\beta}_k) \end{bmatrix}$$

หรือ

$$\sum\underline{\hat{\beta}} = \sigma^2 \begin{bmatrix} m_{22} & m_{23} & \cdots & m_{2k} \\ m_{23} & m_{33} & \cdots & m_{3k} \\ & \vdots & & \\ m_{2k} & m_{3k} & \cdots & m_{kk} \end{bmatrix}^{-1}$$

สำหรับ $V(\hat{\beta}_1)$ เราหาได้ดังนี้

$$V(\hat{\beta}_1) = \sum_a \bar{X}_a^2 V(\hat{\beta}_a) + 2\sum_{j<a} \bar{X}_j \bar{X}_a \mathrm{Cov}(\hat{\beta}_j, \hat{\beta}_a) + \sigma^2/n$$

$$j,\ a = 2,3,\ldots,k; \quad j < a$$

เมื่อมีตัวแปรอิสระเพียง 2 เราจะได้

$$V(\hat{\beta}_1) = \bar{X}_2^2 V(\hat{\beta}_2) + \bar{X}_3^2 V(\hat{\beta}_3) + 2\bar{X}_2\bar{X}_3 Cov(\hat{\beta}_2,\hat{\beta}_3) + \sigma^2/n$$
$$= \sigma^2(1/n + (X_2^2 m_{33} + X_3^2 m_{22} - 2X_2X_3m_{23})/(m_{22}m_{33} - m_{23}^2))$$

$$V(\hat{\beta}_2) = \sigma^2 m_{33}/(m_{22}m_{33} - m_{23}^2)$$

$$V(\hat{\beta}_3) = \sigma^2 m_{22}/(m_{22}m_{33} - m_{23}^2)$$

$$Cov(\hat{\beta}_2, \hat{\beta}_3) = -\sigma^2 m_{23}/(m_{22}m_{33} - m_{23}^2)$$

สำหรับ σ^2 ในกรณีที่มี $k-1$ ตัวแปรอิสระ ตัวประมาณค่าที่ไม่เอียงเฉได้จากการเฉลี่ยผลรวมกำลังสองของตัวเศษแบบกำลังสองต่ำสุด (Least Squares Residuals) นั่นคือ

$$S^2 = \sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2X_{i2} - \cdots - \hat{\beta}_kX_{ik})/(n-k)$$
$$= (m_{yy} - \hat{\beta}_2m_{y2} - \hat{\beta}_3m_{y3} - \cdots - \hat{\beta}_km_{yk})/(n-k)$$

ในกรณีที่ตัวอย่างขนาดโต S^2 จะเท่ากับ σ^2 นั่นคือ S^2 มีคุณสมบัติ Asymptotic ที่ดีเช่นเดียวกับ σ^2

จาก $\Sigma\hat{\beta}$ หรือ $\Sigma\hat{\beta}$ เมื่อแทน σ^2 ด้วย S^2 เราจะได้ตัวประมาณค่าแบบไม่เอียงเฉของความแปรปรวน, ความแปรปรวน ร่วม ($\hat{\Sigma\beta}$ หรือ $\hat{\Sigma\beta}$) ของตัวประมาณค่าแบบกำลังสองต่ำสุดของสัมประสิทธิ์ถดถอย

ช่วงเชื่อมั่นของสัมประสิทธิ์สามารถสร้างได้จากความรู้ที่ว่า

$$(\hat{\beta}_a - \beta_a)/S_{\hat{\beta}_a} \sim t^{(n-k)}, \qquad a = 1,2,\ldots k$$

ในเมื่อ $S_{\hat{\beta}_k}$ แทนค่าประมาณความคลาดเคลื่อนมาตราฐานของ $\hat{\beta}_k$

ตัวอย่าง จากตัวอย่างที่ผ่านมาเราสามารถสร้างช่วงเชื่อมั่น 95% สำหรับสัมประสิทธิ์ของตัวแบบถดถอยได้ดังนี้

$$S^2 = (m_{yy} - \hat{\beta}_2m_{y2} - \hat{\beta}_3m_{y3})/(n-3)$$
$$= (6300-(-1.326)(-3550) - 11.237(125.25))/(12-3$$
$$= 185.27/9 = 20.586$$

$$S^2_{\hat{\beta}_2} = \frac{S^2 m_{33}}{m_{22}m_{33} - m_{23}^2} = \frac{20.586(22.50)}{2250(4.86) - (-54)^2} = 0.021476$$

$$S^2_{\hat{\beta}_3} = \frac{S^2 m_{22}}{m_{22}m_{33} - m_{23}^2} = \frac{20.586(2250)}{2250(4.86) - (-54)^2} = 5.7761$$

$$S^2_{\hat{\beta}_1} = S^2(1/n + (\bar{X}_2^2 m_{33} + \bar{X}^2 m_{22} - 2\bar{X}_2\bar{X}_3 m_{23})/(m_{22}m_{33} - m_{23}^2))$$

$$S^2_{\hat{\beta}_1} = 20.586\left\{1/12 + \frac{70^2(4.86) + 6.7^2(2250) - 2(70)(6.7)}{2250(4.86) - (-54)^2}\right\}$$

$$= 452.1700$$

ดังนั้นค่าประมาณ ของความคลาดเคลื่อนมาตราฐานของ $\hat{\beta}_1, \hat{\beta}_2$ และ $\hat{\beta}_3$ คือ

$$S_{\hat{\beta}_1} = 21.264 \qquad S_{\hat{\beta}_2} = 0.112 \qquad S_{\hat{\beta}_3} = 2.403$$

ดังช่วงเชื่อมั่น 95% ของสัมประสิทธิ์ถดถอย β_a

$$\beta_a = \hat{\beta}_a + t^{(n-k)}_{\alpha/2} S_{\hat{\beta}_a}$$

$$\beta_1 = 177.532 \pm 2.262(21.264) = 69.43 \text{ , } 165.63$$

$$\beta_2 = -1.326 \pm 2.262(0.112) = -1.58, \quad -1.07$$

$$\beta_3 = 11.237 \pm 2.262(2.403) = 5.80, \quad 16.67$$

ในเมื่อ $t^{(12-3)}_{.025} = 2.262$

(2) ช่วงเชื่อมั่นของค่าเฉลี่ย ช่วงเชื่อมั่นของ $E(Y_i)$ นี้สร้างได้จากค่า $\hat{Y}_i$ ที่เป็นตัวประมาณค่า เราทราบว่า $\hat{Y}_i$ มีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเท่ากับ $E(Y_i)$ และความแปรปรวน $\sigma^2_{\hat{Y}_i}$

$$\sigma^2_{\hat{Y}_i} = \sum_a (X_{ia} - \bar{X}_a)^2 V(\hat{\beta}_a) + 2\sum_j \sum_{<a} (X_{ij} - \bar{X}_j)(X_{ia} - \bar{X}_a) Cov(\hat{\beta}_j, \hat{\beta}_a) + \sigma^2/n$$

$$(j, a = 2, 3, \ldots, k; \quad j < a)$$

เมื่อแทน σ^2 ด้วย S^2 เราจะได้ตัวประมาณค่าของ $\sigma^2_{\hat{Y}_i}$ แบบไม่เอียงเฉ คือ $S^2_{\hat{Y}_i}$ แถบเชื่อมั่น (Confidence Band) สำหรับ $E(Y_i)$ สร้างได้จากความรู้ที่ว่า

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

นั่นคือ เมื่อกำหนดค่าของตัวแปรอิสระชุดหนึ่งเราพิจารณาช่วงเชื่อมั่นของ $E(Y_i)$ ได้ นั่นคือ

$$E(Y_i) = \hat{Y}_i + t^{(n-k)}_{\alpha/2} S_{\hat{Y}_i}$$

(3) การแยกความผันแปรตัวอย่างในตัวแปรตามความผันแปรตัวอย่างของ Y สามารถ**แยกได้เช่นเดียวกับการถอดถอยเชิงเดี่ยวดังนั้นเราแยกความผันแปรของ Y คือ $\Sigma(Y_i - Y^2)$ ได้ดังนี้ ในกรณี** 2 ตัวแปรอิสระเราจะได้

$$\underset{\text{SST}}{\sum(Y_i - \bar{Y})^2} = \underset{\text{SSR}}{\sum(\hat{Y}_i - \bar{Y})^2} \quad \underset{\text{SSE}}{\sum(Y_i - \hat{Y}_i)^2}$$

$$\text{SSR} = \sum(\hat{\beta}_1 + \hat{\beta}_2 X_{i2} + \hat{\beta}_3 X_{i3} - \bar{Y})^2$$
$$= \hat{\beta}_2^2 m_{22} + \hat{\beta}_3^2 m_{33} + 2\hat{\beta}_2\hat{\beta}_3 m_{23}$$

จะแทนค่าที่ประมาณได้ของผลกระทบการถอยต่อ Y ซึ่งประกอบด้วย 3 เทอมแรกสมนัยกับผลกระทบของ X_{i2} เทอมที่สอง ต่อผลกระทบของ X_{i3} และเทอมที่สามต่อผลกระทบรวม (Combined Effect) ของทั้งสองตัวแปรผลกระทบรวมของ X_{i2} และ X_{i3} จะแสดงให้เห็นว่า X_{i2} และ X_{i3} อาจจะผันแปรไปด้วยกันในขอบเขตบางอย่าง ความผันแปรร่วมของ X_{i2} และ X_{i3} จะเป็นส่วนหนึ่งของความผันแปรของ Y_i สำหรับ X_{i2} และ X_{i3} แต่ละตัวที่ให้ความผันแปรใน Y_i นั้น ไม่สามารถแยกกันได้ถ้า X_{i2} และ X_{i3} ยังมีสหสัมพันธ์กัน นั่นคือ $m_{23} \neq 0$ สำหรับกรณีทั่วไปที่มี $k=1$ ตัวแปรอิสระเราจะได้

$$\text{SSR} = \sum \hat{\beta}_a^2 m_{aa} + 2\sum_{j<a} \hat{\beta}_j \hat{\beta}_a m_{ja} \quad ; \quad j,a = 2,3,\ldots k$$
$$= \hat{\underset{\sim}{\beta}}'(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\hat{\underset{\sim}{\beta}}$$

กรณีที่เราสนใจค่ารวมของ SSR โดยไม่แยกออกเป็นส่วน ๆ เราจะได้

$$\text{SSR} = \sum_{2}^{k} \hat{\beta}_a m_{ya} = \hat{\underset{\sim}{\beta}}'(\underset{\sim}{X}'\underset{\sim}{Y})$$

(4) **สัมประสิทธิ์การตัดสินใจเชิงซ้อน** (Coefficient of Multiple Determination) สัมประสิทธิ์การตัดสินใจเชิงซ้อน กำหนดได้ดังนี้

$$R^2 = \text{SSR/SST} = 1 - \text{SSE/SST}$$

ในกรณีที่ปรับปรุงแก้ไข เราได้

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k}\right)(1-R^2)$$

R^2 จะเป็นมาตราวัดความมากน้อย (How well) ที่เส้นถดถอยสามารถปรับเข้ากับข้อมูล และใช้เปรียบเทียบ "การปรับที่ดีที่สุด (Goodness of fit)" ของสมการถดถอยหลาย ๆ สมการที่ผันแปรไปกับจำนวนตัวแปรและจำนวนค่าสังเกต

ตัวอย่าง จากข้อมูลตัวอย่างที่แล้วเราแยกความผันแปร ตัวอย่างของ Y ได้ดังนี้

	แหล่ง	สูตร	ค่า	
SSR	X_{i2}	$\hat{\beta}_2^2 m_{22}$	$(-1.326)(2250) = 3956$	6179
	X_{i3}	$\hat{\beta}_3^2 m_{33}$	$(11.237)(4.86) = 614$	
	X_{i2}, X_{i3}	$2\hat{\beta}_2\hat{\beta}_3 m_{23}$	$2(-1.326)(11.237)(-54) = 1609$	
SSE		m_{YY}-SSR	6300-6179 = 121	

SST m_{YY} 6300

$$R^2 = 6179/6300 = 0.981$$

ซึ่งแสดงว่า 98% ของความผันแปรตัวอย่างของประมาณสัมที่ขายได้ เนื่องมาจากผลกระทบของความผันแปรในราคาและความผันแปรของค่าใช้จ่ายในการโฆษณา

(5) การทดสอบสมมติฐานเกี่ยวกับการวิเคราะห์ถดถอยเชิงซ้อน ในตัวแบบถดถอยเชิงซ้อนนั้นสมมติฐานที่จะทดสอบ คือ

ก. ค่าของ β_a เท่ากับจำนวนที่กำหนด β_{ao} ไว้นั่นคือ

$$Ho : \beta_a = \beta_{ao} ; \quad a = 1,2,\ldots k$$

ตัวสถิติทดสอบที่ใช้ทดสอบ Ho คือ $T = \dfrac{\hat{\beta}_a - \beta_{ao}}{S_{\hat{\beta}_a}} \sim t^{n-k}$ $\quad (a = 2,3\ldots k)$

ส่วนมากเราสนใจที่ทดสอบสมมติฐานที่ว่า β_a เท่ากับ 0 เมื่อ a=1 สมมติฐานนั้นคือ จุดตัดแกน (Intercept) เท่ากับ 0 ซึ่งก็คือ ระนาบถดถอย (Regression Plane) ผ่านจุดกำเนิดเมื่อ a = 2, 3...k สมมติฐานที่ว่า $\beta_a = 0$ หมายความว่าตัวแปร X_{ia} ไม่มีอิทธิพลต่อค่าเฉลี่ย Y_i ถ้าปฏิเสธสมมติฐานไม่ได้เราก็สรุปว่า X_{ia} จะไม่เป็นตัวแปรที่เกี่ยวข้องในสมการถดถอย

สำหรับสมมติฐาน Ho : $\beta_a = 0$ และ Ha : $\beta_a \neq 0$ (a=1,2,3,....,k) เรามีวิธีทดสอบ สองแบบดังนี้

– แบบทดสอบความคลาดเคลื่อนมาตรฐาน (Standard Error Test) เป็นแบบทดสอบดั้งเดิมสำหรับระดับนัยสำคัญ 5% วิธีนี้ใช้เปรียบเทียบค่าของตัวประมาณค่ากับความคลาดเคลื่อนมาตราฐานของมันนั่นคือ เราจะปฏิเสธ Ho : $\beta_a = 0$ ถ้า $S_{\hat{\beta}_a} < \frac{1}{2}\hat{\beta}_a$
นั่นคือเรายอมรับว่าค่าประมาณพารามิเตอร์ β_a มีนัยสำคัญเชิงสถิติหรือแตกต่างจากศูนย์อย่างมีนัยสำคัญ

– แบบทดสอบ T (Student's Test) ใช้เปรียบเทียบค่าของตัวสถิติ T

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

จากค่าสังเกตุของตัวอย่างกับค่าทางทฤษฎีจากตาราง t ที่มีองศาแห่งความเป็นอิสระ n-k และระดับนัยสำคัญที่ต้องการ จะปฏิเสธ Ho: $\beta_a = 0$ ถ้า $|T| > t_{\alpha/2}^{(n-k)}$ นั่นคือเรายอมรับว่าตัวแปรอิสระจะเป็นตัวอธิบายความผันแปรใน Y

(ข) สมมติฐานที่เราสนใจจะทดสอบต่อไปคือ "ไม่มีตัวแปรอิสระใดมีอิทธิพลต่อค่าเฉลี่ยของ Y_i" นั่นคือ

$$\text{Ho} : \beta_2 = \beta_3 = \cdots \beta_k = 0$$

หรือ $$\text{Ho} : \rho^2 = 0$$

สำหรับ Ha : Ho ไม่จริง นั่นก็คือมีความชันของการถดถอยที่ไม่เท่ากับศูนย์อย่างน้อยหนึ่งตัว ถ้าเป็นจริงแล้วความผันแปรของ Y ไม่มีผลจากการเปลี่ยนแปลงของตัวแปรอิสระตัวหนึ่งตัวใด แต่เนื่องจากการสุ่มอย่างเดียว (Purely random) อย่างไรก็ตามถ้า Ho เป็นจริงแล้วค่าสังเกตของ SSR (ผลรวมของกำลังสองเนื่องจากการถดถอย) จะไม่เท่ากับศูนย์ก็เนื่องจากการสุ่มตัวอย่างเท่านั้น แต่เราทราบมาแล้วว่า

$$SSR = \sum_a \hat{\beta}_a^2 m_{aa} + 2 \sum_{j<a} \hat{\beta}_j \hat{\beta}_a m_{ja} \quad (j, a = 2,3,\ldots,k;\ j<a)$$

จากความจริงที่ว่าถ้า β แทนด้วย β และ Ho เป็นจริงแล้ว SSR จะเท่ากับศูนย์และSSE จะเท่ากับ SST เรายังสามารถแสดงได้อีกว่า ถ้า Ho เป็นจริงแล้วตัวสถิติ F

$$F = (SSR/(k-1)/(SSE/(n-k)))$$

ในเมื่อ $F_{k-1,n-k,}$ แทนการแจกแจง F ที่มีองค์แห่งความเป็นอิสระ k-1,n-k ถ้าค่าของตัวสถิติ F ไม่แตกต่างจากศูนย์อย่างมีนัยสำคัญแล้วอย่างก็จะไม่ให้ประจักษ์พยานที่ว่าตัวแปรอิสระมีผลกระทบใด ๆ ต่อค่าเฉลี่ยของ Y นั่นคือสมมติฐาน Ho : $\beta_2 = \beta_3 = \ldots \beta_k = 0$ จะได้รับการปฏิเสธ ถ้าค่าสังเกตจากตัวอย่างของสถิติ F มากกว่าค่าทางทฤษฎีจากตาราง F ที่มีองศาแห่งความเป็นอิสระ (k-1,n-k) และระดับนัยสำคัญ α

สำหรับค่าตัวสถิติ F สามารถคำนวนได้จาก R^2 ดังนี้

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

ในการทดสอบสมมติฐานดังกล่าวนี้ ถ้าใช้เทคนิควิเคราะห์ความแปรปรวน ก็จะสรุปผลการคำนวนได้ดังตารางต่อไปนี้

SOV	df	SS	MS	F-Ratio
$X_2, X_3, \ldots X_k$	k - 1	SSR	MSR	MSR/MSE
Residual	n - k	SSE	MSE	
Total	n - 1	SST		

ตัวอย่าง ตามตัวอย่างที่กล่าวมาแล้ว ถ้าเราต้องการทดสอบ

$$\text{Ho} : \beta_a = 0 \quad ; \quad a = 2, 3$$
$$\text{Ha} : \beta_a \neq 0$$

เราจะได้ว่า $T_2 = \hat{\beta}_2 / S_{\hat{\beta}_2} = -1.326/(0.112) = -11.839$
ดังนั้น T_2 มากกว่า $t^{(9)}_{.025} = 2.262$ จึงปฏิเสธ Ho: $\beta_2 = 0$ และ $T_3 = \hat{\beta}_3 / S_{\hat{\beta}_3} = 11.237/2.403 = 4.676$
ซึ่งมากกว่า $t^{(9)}_{.025} = 2.262$ จึงปฏิเสธ Ho : $\beta_3 = 0$

เพราะฉะนั้นตัวแปรอิสระทั้ง X_2 และ X_3 จะเป็นตัวอธิบายความผันแปรใน Y นั่นคือ ทั้ง X_2 และ Y_3 มีอิทธิพลต่อ E(Y)

ในกรณีที่ใช้เทคนิควิเคราะห์ความแปรปรวนเพื่อทดสอบอิทธิพลของ X_2 และ X_3 เราจะทำได้ดังนี้

SOV	df	SS	MS	F
X_2, X_3	2	6179	3089.5	229.875
Residual	9	121	13.44	
Total	11	6300		

เนื่องจาก F = 229.874 โตกว่า $F^{(2,9)}_{.05} = 5.79$ เราจึงปฏิเสธสมมติฐาน Ho : $\beta_2 = \beta_3 = 0$ นั่นคือ X_2 และ X_3 มีอิทธิพลต่อตัวแปรตาม Y

เราจะสังเกตได้ว่า ถ้าตัวหนึ่งตัวใดของสัมประสิทธิ์ถดถอยที่ประมาณได้ $\hat{\beta}_2, \hat{\beta}_3, \ldots, \hat{\beta}_k$ แตกต่างจากศูนย์ (0) อย่างมีนัยสำคัญ ตามการทดสอบแบบ T แล้วค่าของตัวสถิติ F ก็จะแตกต่างจากศูนย์อย่างมีนัยสำคัญด้วย ทั้งนี้การทดสอบจะต้องกระทำในระดับนัยสำคัญเดียวกัน และมีสมมติฐานรอง Ha เช่นเดียวกัน (มีบางกรณีซึ่งเกิดยากที่ค่า T ทั้งหมดมีนัยสำคัญ แต่ค่า F ไม่มีนัยสำคัญ) อย่างไรก็ตาม มันจะเป็นไปได้ที่พบว่าไม่มี $\hat{\beta}_2, \hat{\beta}_3, \ldots, \hat{\beta}_k$ ตัวหนึ่งตัวใดแตกต่างจากศูนย์อย่างมีนัยสำคัญตามการทดสอบ T แต่ในขณะเดียวกันจะปฏิเสธสมมติฐานที่ว่า Ho : $\beta_2 = \beta_3 = \ldots = \beta_k = 0$ ตามการทดสอบ F เหตุการณ์แบบนี้จะเกิดขึ้นได้เมื่อตัวแปรอิสระมีสหสัมพันธ์กันสูง ในสถานการณ์อย่างนี้อิทธิพลที่แยกกันของตัวแปรอิสระแต่ละตัวที่มีต่อตัวแปรตามนั้นจะน้อย

(Weak) ในขณะที่อิทธิพลร่วมอาจจะมาก (Strong)

ความสัมพันธ์ระหว่างค่าของตัวสถิติ T แต่ละตัวในการทดสอบสมมติฐาน

$Ho: \beta_a = 0 ; a = 2,3,....,k$

กับค่าของตัวสถิติ F ในการทดสอบสมมติฐาน

$Ho: \beta_2 = \beta_3 = ... \beta_k = 0$

สามารถทำได้ด้วยการแสดงตัวสถิติ F ในเทอมของตัวสถิติ T เราจะทำในกรณีของตัวแปรอิสระ 2 ตัว ดังนี้

เราทราบว่า $SSR = \hat{\beta}_2^2 m_{22} = \hat{\beta}_3^2 m_{33} + 2\hat{\beta}_2\hat{\beta}_3 m_{23}$

และ $S^2 = SSE/(n-3)$

เพราะฉะนั้น $F^{(2,\ n-3)} = \frac{SSR/2}{SSE/(n-3)} = \frac{\hat{\beta}_2^2 m_{22} + \hat{\beta}_3^2 m_{33} + 2\hat{\beta}_2\hat{\beta}_3 m_{23}}{2S^2}$

เมื่อให้ค่าของตัวสถิติ T ของ $\hat{\beta}_2$ และ $\hat{\beta}_3$ เป็น t_2 และ t_3 ตามลำดับ แล้วจะได้

$$t_2 = \hat{\beta}_2/S_{\hat{\beta}_2} \qquad t_3 = \hat{\beta}_3/S_{\hat{\beta}_3}$$

นั่นคือ
$$t_2^2 = \hat{\beta}_2^2(m_{22}m_{33} - m_{23}^2)/S^2 m_{33}$$
$$t_3^2 = \hat{\beta}_3^2(m_{22}m_{33} - m_{23}^2)/S^2 m_{22}$$

หรือ $\hat{\beta}_2^2 = \frac{t_2^2 S^2 m_{33}}{m_{22}m_{33} - m_{23}^2}$ $\qquad \hat{\beta}_3^2 = \frac{t_3^2 S^2 m_{22}}{m_{22}m_{33} - m_{23}^2}$

และยังได้ $\hat{\beta}_2\hat{\beta}_3 = t_2 t_3 S^2 \sqrt{m_{22}}\sqrt{m_{33}}/(m_{22}m_{33} - m_{23}^2)$

ดังนั้น $F^{(2,n-3)} = \frac{t_2^2 S^2 m_{22}m_{33} + t_3^2 S^2 m_{22}m_{33} + 2t_2 t_3 \sqrt{m_{22}}\sqrt{m_{33}} m_{23}}{2S^2(m_{22}m_{33} - m_{23}^2)}$

ถ้าเราให้ r_{23} แทนสัมประสิทธิ์ของสหพันธ์จากตัวอย่างระหว่าง X_{i2} และ X_{i3}

นั่นคือ $r_{23} = m_{23}/\sqrt{m_{22}}\sqrt{m_{33}}$

แล้วเราจะได้ $F^{(2,n-3)} = (t_2^2 + t_3^2 + 2t_2 t_3 r_{23})/2(1 - r_{23}^2)$

ซึ่งแสดงให้เห็นได้ชัดเจนว่าถ้า r_{23} ไม่ต่างจาก 1 มากนัก ค่าของ $F^{(2,n-3)}$ จะมากถึงแม้ว่าทั้ง t_2 และ t_3 จะน้อย นั่นคือ $\hat{\beta}_2$ หรือ $\hat{\beta}_3$ อาจจะไม่แตกต่างจากศูนย์อย่างมีนัยสำคัญ แต่ค่าของ $F^{(2,n-3)}$ อาจจะมีนัยสำคัญอย่างยิ่งได้

ตัวอย่าง พิจารณาตัวแบบถดถอย

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

สมมติ n = 20 และโมเมนต์ตัวอย่างที่คำนวนได้ เป็นดังนี้

$$m_{yy} = 100 \quad m_{y2} = 90 \quad m_{y3} = 90$$

$$m_{22} = 100 \quad m_{33} = 100 \quad m_{23} = 95$$

เราจะได้ $\hat{\beta}_2 = \dfrac{90(100) - 90(95)}{100(100) - (95)^2} = 0.461$

$\hat{\beta}_3 = \dfrac{90(100) - 90(95)}{100(100) - (95)^2} = 0.461$

$SSR = \hat{\beta}_2^2 m_{22} + \hat{\beta}_3^2 m_{33} + 2\hat{\beta}_2\hat{\beta}_3 m_{23}$

$= 0.461^2(100) + 0.461^2(100) + 2(0.461)(0.461)(95)$

$= 83$

$SST = m_{yy} = 100$

$SSE = SST - SSR = 100 - 83 = 17$

เพราะฉะนั้น $S^2 = SSE/(n-3) = 17/(20-3) = 1$

$S^2_{\hat{\beta}_2} = S^2 m_{22}/(m_{22}m_{33} - m_{23}^2)$

$= 1(100)/(100(100) - (95)^2) = 0.102564$

$S^2_{\hat{\beta}_3} = 1(100)/(100(100) - (95)^2) = 0.102564$

$t_2 = 0.461/\sqrt{0.102564} = 1.444;$

$t_3 = 0.461/\sqrt{0.102564} = 1.444$

โดยที่ค่า t จากตารางที่มี df = 17 และ $\alpha = .05$ เป็น 2.110 ซึ่งมากกว่า 1.444 ดังนั้น β_2 หรือ β_3 ไม่แตกต่างจากศูนย์อย่างมีนัยสำคัญ ณ ระดับนัยสำคัญ 5 % และในทางตรงกันข้ามค่าของ F

$$F = \frac{SSR/(k-1)}{SSE/(n-k)} = \frac{83/2}{17/(20-3)} = 41.5$$

ซึ่งมีค่าสูงกว่าค่าจากตาราง คือ $F^{(2,17)}_{.05} = 3.59$ มาก

ดังนั้นจากการทดสอบแบบ T เราจึงไม่สามารถปฏิเสธสมมติฐานที่ว่า Ho : $\beta_2 = 0$ หรือ สมมติฐาน Ho : $\beta_3 = 0$ และจากการทดสอบแบบ F เราปฏิเสธสมมติฐาน Ho $\beta_2 = \beta_3 = 0$ ที่เป็นเช่นนี้เนื่องจากความจริงที่ว่าตัวแปรอิสระ X_{i2} และ X_{i3} แต่ละตัวซึ่งแยกกัน จะมีผลต่อความผันแปรของ Y น้อย แต่เมื่อร่วมกันมีมาก นั่นคือ SSR แยกออกได้เป็น

$$\hat{\beta}_2^2 m_{22} = 21.25 \qquad \hat{\beta}_3^2 m_{33} = 21.25$$

$$2\hat{\beta}_2\hat{\beta}_3 m_{23} = 40.4$$

ซึ่งแสดงว่าอิทธิพลร่วมของ X_{i2} และ X_{i3} ใน SSR โตเกือบ 2 เท่าของอิทธิพลที่แยกกันของ X_{i2} หรือ X_{i3} จะสังเกตได้ว่าสหสัมพันธ์ระหว่าง X_{i2} และ X_{i3} มีดีกรีสูง นั่นคือ

$$r_{23} = m_{23}/\sqrt{m}_{22}\sqrt{m}_{33}$$

$$= 95/\sqrt{100}\ \sqrt{100} = 0.95$$

(ค) สมมติฐานเกี่ยวกับอิทธิพลที่มีต่อตัวแปรตาม Y_i ของตัวแปรอิสระที่เพิ่มเข้าไปในการถดถอย ลองพิจารณาข้อเสนอ 2 แบบ ดังนี้

ข้อเสนอแรกกล่าวว่า สมการถดถอย คือ

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

อีกข้อหนึ่งกล่าวว่า Y_i นอกจากจะขึ้นอยู่กับ $X_{i2}, X_{i3}, \ldots, X_{ik}$ แล้วยังขึ้นอยู่กับตัวแปรอิสระอื่น ๆ คือ $X_{i(k+1)}, X_{i(k+1)}, \ldots, X_{iq}$ นั่นคือ สมการถดถอยจะเป็น

$$Y_i = \beta_1 + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \beta_{k+1} X_{i(k+1)} + \cdots + \beta_q X_{iq} + \varepsilon_i$$

ในกรณีนี้ ข้อเสนอที่สองสามารถจะทดสอบได้ด้วยการทดสอบสมมติฐาน

$$\text{Ho} : \quad \beta_{k+1} = \beta_{k+2} = \cdots = \beta_q = 0$$

สำหรับแบบทดสอบที่เหมาะสมสำหรับสมมติฐานนี้ เราจะใช้สัญญลักษณ์ใหม่คือ ใช้ตัวห้อย k เพื่อแสดงถึงค่าที่เกี่ยวข้องกับเซ็ทของตัวแปรอิสระเดิม และตัวห้อย Q แสดงถึงค่าที่เกี่ยวกับเซ็ทของตัวแปรอิสระทั้งหมด

ดังนั้นเราจึงเขียนความผันแปรทั้งหมด ได้เป็น

$$SST = SSR_K + SSE_K$$

และ $$SST = SSR_Q + SSE_Q$$

ถ้าตัวแปรที่เพิ่มเข้าไปไม่เกี่ยวข้องกับการอธิบายความผันแปรของ Y_i หรือไม่มีอิทธิพลต่อ Y_i แล้ว SSR_k และ SSR_q (ในประชากร) จะเป็นจำนวนเดียวกัน และความแตกต่างที่สังเกตได้ระหว่างจำนวนทั้งสอง จะเนื่องมาจากความคลาดเคลื่อนจากการสุ่มตัวอย่าง ถ้า Ho เป็นจริง แล้วเราสามารถแสดงได้ว่า

$$F_a = \frac{(SSR_Q - SSR_K)/(q - k)}{SSE_Q/(n - q)} \sim F^{(q-k,n-q)}$$

ตัวสถิติ F_a นี้ สามารถใช้ทดสอบ Ho ณ ระดับนัยสำคัญที่ต้องการได้ เราสามารถแสดงได้ว่า

$$SSR_Q \geq SSR_K$$

ดังนั้น F_a จึงไม่มีทางเป็นลบได้ อย่างเก่งก็เท่ากับศูนย์ (0) จากผลอันนี้ทำให้เราทราบว่า

$$SSR_Q/SST \geq SSR_K/SST$$
$$R_Q^2 \geq R_K^2,$$

ซึ่งหมายความว่า ตัวแปรอิสระที่เพิ่มเข้าไปใหม่ในสมการถดถอยนั้นจะไม่มีทางลดค่าของสัมประสิทธิ์การตัดสินใจได้เลย อย่างไรก็ตาม ที่กล่าวมานี้ไม่จำเป็นต้องเป็นจริงสำหรับ R^2

การทดสอบสมมติฐาน Ho : $\beta_1 = \beta_2 = \ldots = \beta = 0$ ที่ทำโดยการวิเคราะห์ความแปรปรวนนั้นจะสรุปผลได้ดังตาราง

SOV	df	SS	MS	F-ratio
$X_2, X_3, \ldots X_k$	k - 1	SSR_K	MSR_K	
$X_2, X_3, \ldots X_k, X_{k+1}, \ldots X_q$	q - 1	SSR_Q	MSR_Q	$MS(R_Q - R_K)$
$X_{k+1}, X_{k+2}, \ldots X_q$	q - k	$SSR_Q - SSR_K$	$MS(R_Q - R_K)$	MSE_Q
Residual	n - k	SSE_Q	MSE_Q	
Total	n - 1	SST		

ตัวอย่าง สมมติว่าฟังก์ชั่นการบริโภคของประเทศหนึ่งเสนอไว้ ดังนี้

$$C_t = \beta_1 + \beta_2 Y_t + \varepsilon_t$$

และ $$C_t = \beta_1 + \beta_2 Y_t + \beta_3 C_{t-1} + \beta_4 I_t + \varepsilon_t$$

ในเมื่อ C = การบริโภค, Y = รายได้, C_{t-1} = การบริโภคของคาบเวลาที่ผ่านมา, และ L = ทรัพย์สิน

เราจะทดสอบสมมติฐาน Ho : $\beta_3 = \beta_4 = 0$

Ha : Ho ไม่เป็นจริง

ณ ระดับนัยสำคัญ 1 % โดยการใช้ตัวสถิติ F_a

ถ้าเรามีข้อมูลที่สรุปแล้ว เป็นดังนี้ $n_1 = 31$ q = 4 k = 2

$\bar{R}^2_Q = 0.984$ $\bar{R}^2_K = 0.944$

เราก็สามารถหาค่าของตัวสถิติทดสอบได้

เนื่องจาก $\bar{R}^2_K = R^2_K - \left(\frac{k-1}{n-k}\right)(1 - R^2_K)$

เราจึงได้ $R^2_K = \frac{(n-k)\bar{R}^2_K + (k-1)}{n-1} = 0.946$

ในทำนองเดียวกัน $\hat{R}^2_Q = 0.986$

ค่าของตัวสถิติทดสอบจึงเป็น

$$F_a = \frac{SSR_Q - SSR_K}{SSE_Q}\ \frac{n-q}{q-k} = \frac{R^2_Q - R^2_K}{1 - R^2_Q}\ \frac{n-q}{q-k}$$

$$= 38.57$$

แต่ค่าของ $F^{(2,27)}$ ณ $\alpha = .01$ เท่ากับ 5.49 จึงปฏิเสธ Ho นั่นคือการเพิ่มตัวแปรอิสระ (การบริโภคของคาบเวลาที่ผ่านมา และทรัพย์สิน) จะมีส่วนช่วยการอธิบายของความผันแปรมีส่วน ในการบริโภคปัจจุบัน หรือการบริโภคที่ผ่านมา และทรัพย์สิน จะมีอิทธิพลต่อการบริโภคปัจจุบัน

ตัวอย่าง ในการศึกษาเกี่ยวกับความสัมพันธ์ของตัวแปรต่าง ๆ ดังนี้

Y : ระยะทางที่รถวิ่งได้ (ไมล์) ต่อน้ำมัน 1 แกลลอน

X_2 : ระดับของอ๊อคเทน, X_3 ดัชนีของฝนตก, X_4 : สภาพถนน

ผลจากการศึกษาได้ข้อมูลมาดังนี้

Y	X_2	X_3	X_4
32	90	100	0
30	90	104	0
35	90	102	0
33	90	104	0
35	90	96	0
34	90	96	1
29	90	110	1
32	90	105	1

Y	X_2	X_3	X_4
35	95	101	0
38	95	93	0
37	95	91	0
40	95	89	0
41	95	88	0
37	95	97	1
41	95	91	1
36	95	96	1

Y	X_2	X_3	X_4
44	100	91	0
46	100	93	0
47	100	96	0
47	100	91	0
47	100	94	0
43	100	93	1
47	100	91	1
45	100	91	1

36 90 103 1	35 95 101 1	48 100 89 1
34 90 102 1	40 95 91 1	47 100 91 1

และสรุปข้อมูลเหล่านี้ได้เป็น

$$n = 30, \sum Y = 1170, \bar{Y} = 39, \sum X_2 = 2850, \bar{X}_2 = 95$$
$$\sum X_3 = 2880, \bar{X}_3 = 96, \sum X_4 = 15, \bar{X}_4 = 0.5$$
$$\sum y^2 = 978, \sum x_2^2 = 500, \sum x_3^2 = 986$$
$$\sum x_4^2 = 7.50, \sum yx_2 = 650, \sum yx_3 = -778$$
$$\sum yx_4 = -1.0, \sum x_2x_3 = +510, \sum x_2x_4 = 0$$
$$\sum x_3x_4 = 7.0$$

ถ้าสมการถดถอยเป็น

$$Y = \beta_1 + \beta_2 X_2 + \epsilon$$

เราจะได้สมการถดถอยตัวอย่างเป็น

$$\hat{Y} = -84.50 + 1.30X_2 \qquad R^2 = 0.864$$
$$(9.27) \quad (0.097)$$

SOV	df	SS	MS	F-ratio
X_2	1	845	845	845/4.75 = 117.8
Residual	28	133	4.75	$F_{05}^{(1,28)} = 4.20$
Total	29	978		

จากตารางเราจะเห็นว่า X_2 มีอิทธิพลต่อตัวแปรตาม

ถ้าเพิ่ม X_3 เราจะได้สมการถดถอยตัวอย่าง เป็น

$$\hat{Y} = -36.88 + 1.05X_2 - 0.25X_3$$
$$(19.48) \quad (0.13) \quad (0.09)$$

SOV	df	SS	MS	F-ratio
X_2	1	845	845	
X_2, X_3	2	873	436	436/3.9 = 112.7

เพิ่ม X_3	1	28	28	$28/3.9 = 7.18$
Residual	27	105	3.9	
Total	29	978	$F^{(2,27)}_{.05} = 3.35$;	$F^{(1,27)}_{.05} = 4.21$

เราจะเห็นว่า ทั้ง X_2 และ X_3 มีอิทธิพลต่อ Y และเราจะเห็นอีกว่าการเพิ่ม X_3 เข้าไปนั้นมีผลต่อ Y จริง

เมื่อเพิ่ม X_4 เข้าไปอีก ก็จะได้สมการถดถอยตัวอย่าง เป็น

$$\hat{Y} = \underset{(19.93)}{-36.64} + \underset{(0.13)}{1.05X_2} - \underset{(0.09)}{0.25X_3} + \underset{(0.74)}{0.10X_4} \qquad R^2 = .893$$

SOV	df	SS	MS	F
X_1, X_2	2	873	436	
X_1, X_2, X_3	3	874	291	$.291/4 = 72.75$
เพิ่ม X_3	1	1	1	$1/4 = 0.25$
Residual	26	104	4	
Total	29	978	$F^{(3,26)}_{.05} = 2.98$,	$F^{(1,26)}_{.01} = 4.23$

จากตารางจะเห็นว่า X_2, X_3 และ X_4 มีอิทธิพลต่อ Y แต่การเพิ่ม X_4 ไม่มีผลต่อ Y ดังนั้นสมการถดถอยตัวอย่างจึงมีตัวแปรอิสระเพียง X_2 และ X_3 เท่านั้น

(ง) สมมติฐานต่อไปคือ การเท่ากันของพารามิเตอร์ถดถอย ซึ่งมีสมมติฐานที่จะทดสอบดังนี้

$$Ho : \beta_j = \beta_a, (j \neq a)$$

ซึ่งจะทำการทดสอบสองทาง หรือทางเดียวก็ได้ ตัวอย่างเช่น ถ้าฟังก์ชั่นของการบริโภครวม เป็นดังนี้

$$C_t = \beta_1 + \beta_2 W_t + \beta_3 P_t + \beta_4 C_{t-1} = \varepsilon_t$$

ในเมื่อ C = การบริโภค, W = รายได้จากค่าจ้าง, P = รายได้อย่างอื่น

สมมติฐานที่เราสนใจก็คือ แนวโน้มที่จะบริโภคเพิ่มขึ้น (MPC) ของรายได้จากค่าจ้างจะเท่ากับของรายได้อย่างอื่น นั่นคือ

$$\beta_2 = \beta_3$$

การทดสอบสมมติฐานที่ว่าสัมประสิทธิ์ถดถอย 2 จำนวนเท่ากันนั้น เราจะต้องพิจารณาการแจกแจงของผลต่างของตัวประมาณค่าแบบกำลังสองต่ำสุด โดยทั่วไปเราจะได้

$$E(\hat{\beta}_j - \hat{\beta}_a) = \hat{\beta}_j - \beta_a$$

และ $V(\hat{\beta}_j - \hat{\beta}_a) = V(\hat{\beta}_j) + V(\hat{\beta}_a) - 2Cov(\hat{\beta}_j, \hat{\beta}_a)$

ถ้า Ho เป็นจริง $(\hat{\beta}_j - \hat{\beta}_a) \sim N(0, \sigma^2_{(\hat{\beta}_j - \hat{\beta}_a)})$

ตัวประมาณค่าที่ไม่เอียงเฉของ $\sigma^2_{(\hat{\beta}_j - \hat{\beta}_a)}$ คือ $S^2_{(\hat{\beta}_j - \hat{\beta}_a)}$ ซึ่งได้จากใช้ S^2 เป็นตัวประมาณของ σ^2 แล้วเราจะได้ว่า

$$T = (\hat{\beta}_j - \hat{\beta}_a)/S_{(\hat{\beta}_j - \hat{\beta}_a)} \sim t^{(n-k)}$$

สมมติฐานที่ว่าสัมประสิทธิ์ถดถอย 2 จำนวนเท่ากัน จะได้รับการปฏิเสธ ถ้าค่าของตัวสถิตินี้แตกต่างจากศูนย์อย่างมีนัยสำคัญ นั่นคือถ้า $|T| > t_{\alpha}^{(n-k)}$ หรือ $t_{\alpha/2}^{(n-k)}$ แล้วแต่สมมติฐานรองที่ทดสอบ

การทดสอบสมมติฐานที่ว่า สัมประสิทธิ์ถดถอยมากกว่า 2 จำนวนเท่ากันนั้น ก็สามารถทำได้ ซึ่ง โกลด์เบอร์เกอร์ ได้อธิบายไว้ในหนังสือเศรษฐมิติ ของเขาในหน้า 175

สมมติฐานที่คล้ายคลึงกับที่กล่าวมาก็คือ คำกล่าวที่ว่าผลรวมของสัมประสิทธิ์ถดถอย 2 จำนวนเท่ากับจำนวนที่กำหนด นั่นคือ $Ho: \beta_j + \beta_a = A ; j \neq a$
ซึ่งสามารถทดสอบได้ด้วยตัวสถิติทดสอบ T

$$T = (\hat{\beta}_j + \hat{\beta}_a - A)/S_{(\hat{\beta}_j + \hat{\beta}_a)}$$

ในเมื่อ $S^2_{(\hat{\beta}_j - \hat{\beta}_a)} = S^2_{\hat{\beta}_j} + S^2_{\hat{\beta}_a} + 2\,Cov(\hat{\beta}_j, \hat{\beta}_a)$

การทดสอบนี้สามารถขยายออกไปถึง ผลรวมของสัมประสิทธิ์ถดถอยมากกว่า 2 จำนวน

ตัวอย่างของผลรวม 2 จำนวน เช่นในฟังก์ชั่นการผลิตของ Cobb-Douglas

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

ในเมื่อ Y_i = log(output), X_{i2} = log(labor input), X_{i3} = log(capital input)

สมมติฐานที่สนใจก็คือ ผลตอบแทนคงที่ (Constant Returns to Scale) นั่นก็คือ

$Ho: \beta_2 + \beta_3 = 1$

สำหรับสมมติฐาน $Ho: \beta_j + \beta_a = A$ นั้น Tintner ได้แนะนำตัวสถิติทดสอบไว้คือ

$$F = \frac{SSE_2 - SSE_1}{SSE_2/(n-k)} \sim F^{(1,\ n-k)}$$

ในเมื่อ SSE_2 เป็นค่าประมาณของผลรวมกำลังสองของตัวเศษ จากฟังก์ชั่นที่มีข้อจำกัด

SSE_1 เป็นค่าประมาณของผลรวมกำลังสองของตัวเศษ จากฟังก์ชั่นที่ไม่มีข้อจำกัด

ศึกษาการทดสอบจากตัวอย่างของฟังก์ชั่นการผลิตของ Cobb-Douglas

$$Y_i = B_1 L_i^{\beta_2} K_i^{\beta_2} + \varepsilon_i$$

ในเมื่อ Y_i = ผลผลิต, L = แรงงาน, และ K = เงิน

ฟังก์ชั่นนี้ ถ้า $\beta_2+\beta_3 = 1$ เราจะได้ผลตอบแทนคงที่ สมมติว่าเมื่อใช้ตัวอย่างขนาด 30 สำหรับอุตสาหกรรมอย่างหนึ่ง เราจะได้เส้นถดถอยตัวอย่างเป็น

$$\hat{Y}_i = (\hat{\beta}_1 L_i^{0.8223} \quad K_i^{0.2324}), \quad R^2 = 0.768$$

$$m_{yy} = 180, \quad SSE_1 = 464, \quad S_{\hat{\beta}_2} = 0.03, \quad S_{\hat{\beta}_3} = 0.02$$

เราจะเห็นว่า $\hat{\beta}_2+\hat{\beta}_3 = 1.0547$ นั่นคือค่าประมาณนี้แสดงถึงผลตอบแทนเพิ่มขึ้น (Increasing Return to Scale) เราต้องการจะทดสอบความเชื่อถือของผลที่ได้นี้ นั่นคือเราต้องการทดสอบสมมติฐานหลักที่ว่า

$$Ho: \beta_2+\beta_3 = 1$$

สมมติฐานนี้ก็ทดสอบด้วยตัวสถิติทดสอบ F ดังได้กล่าวมาแล้ว

ตาม Ho เราได้แบบเป็น

$$Y_i = \beta_1 L_i^{\beta_2} K_i^{1-\beta_2} + \varepsilon_i$$

เราประมาณค่าของ β_2 ได้เป็น 0.7431 และ $1-\beta_2$ เป็น 0.2569 ดังนั้นฟังก์ชั่นการผลิตจะเป็น

$$\hat{Y}_i = \hat{\beta}_1 L_i^{0.7431} K_i^{0.2569} \qquad R^2 = 0.650$$

$$m_{yy} = 180 \qquad SSE_2 = 6.45$$

ดังนั้น จากฟังก์ชั่นที่มีข้อจำกัด $SSE_2 = 6.45$

จากฟังก์ชั่นที่ไม่มีข้อจำกัด $SSE_1 = 4.64$

เราจะได้ $$F = \frac{(SSE_2 - SSE_1)}{SSE_1/(n-k)} = \frac{6.45 - 4.64}{4.64/(30-2)}$$

$$= 1.81(28)/4.64 = 10.92$$

แต่ $F^{(1,28)}_{.05} = 4.20$ จึงปฏิเสธ Ho และสรุปได้ว่า $\beta_2+\beta_3 \neq 1$

(จ) สมมติฐานที่เราสนใจต่อมาก็คือ การเท่ากันของสมการถดถอย 2 เส้น ลองพิจารณา

$$Y_i = \beta_1+\beta_2 X_{i2}+\beta_3 X_{i3}+\ldots\ldots+\beta_k X_{ik} + \varepsilon_i$$

ซึ่งจะประมาณจากตัวอย่างขนาด n ค่าสังเกต ถ้าสมมติว่าเราได้รับค่าสังเกตเพิ่มเติม m ค่าสังเกต และ m>k และเราต้องการทดสอบสมมติฐานที่ว่า ค่าสังเกตเพิ่มเติมนั้นมาจากประชากรเดียวกับ n ค่าสังเกตแรกนั้น เช่นข้อมูลของฟังก์บริโภครวมก่อนและหลังสงคราม เป็นต้น

ถ้าเราจะยอมรับความเป็นไปได้ดีที่ว่า พารามิเตอร์ของฟังก์ชั่นบริโภครวมหลังสงครามอาจจะแตกต่างจากฟังก์ชั่นบริโภครวมก่อนสงคราม เราควรจะทดสอบสมมติฐานที่ว่า พารามิเตอร์ของฟังก์ชั่นบริโภครวมไม่เปลี่ยนแปลง ถ้าเราเขียนสมการถดถอยเป็น

$$Y_i = \beta_1 + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i \qquad (i = 1,2,\ldots n)$$

และ $$Y_i = \delta_1 + \delta_2 X_{i2} + \ldots + \delta_k X_{ik} + \varepsilon_i \qquad (i = n+1, n+2, \ldots, n+m)$$

แล้วสมมติฐานที่จะทดสอบจะเป็น

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

ตัวสถิติทดสอบจะได้จากวิธีประมาณค่าแบบกำลังสองต่ำสุดในข้อมูลชุดแรก $(i=1,2,\ldots, n)$ ในข้อมูลชุดที่สอง $(i=n+1,n+2,\ldots,n+m)$ และในข้อมูลชุดที่สามซึ่งได้จากรวมข้อมูลทั้งสองชุดเข้าด้วยกัน $(i=1,2,\ldots,n,n+1n+2,\ldots n+m)$ ในการประมาณพารามิเตอร์ถดถอยจากข้อมูลรวมเราจะเขียนสมการถดถอยเป็น

$$Y_i = \alpha_1 + \alpha_2 X_{i2} + \ldots + \alpha_k X_{ik} + \varepsilon_i \qquad (i = 1,2,\ldots n+m)$$

และแทนผลรวมกำลังสองของตัวเศษ (ความผันแปรที่อธิบายไม่ได้) แบบกำลังสองต่ำสุด เป็น

$$SSE_1 = \sum_{i=1}^{n} (Y_i - \beta_1 - \beta_2 X_{i2} - \ldots - \beta_k X_{ik})^2$$

$$SSE_2 = \sum_{i=n+1}^{n+m} (Y_i - \delta_1 - \delta_2 X_{i2} - \ldots - \delta_k X_{ik})^2$$

$$SSE_c = \sum_{i=1}^{n+m} (Y_i - \alpha_1 - \alpha_2 X_{i2} - \ldots - \alpha_k X_{ik})^2$$

แล้วเราสามารถแสดงได้ว่า

$$F_c = \frac{SSE_c - (SSE_1 + SSE_2)/k}{(SSE_1 + SSE_2)/(n+m-2k)} \sim F^{(k,n+m-2k)}$$

และสมมติฐาน Ho จะได้รับการปฏิเสธ ถ้าค่าของตัวสถิติ F_c มากกว่าค่าของ F จากตาราง นั่นคือ

$F_c > F^{(k,n+m-2k)}$

ตัวสถิติทดสอบ F_c นี้มีจุดที่น่าสนใจ 2 ประการคือ

— ตัวสถิติทดสอบนี้จะใช้ก็ต่อเมื่อจำนวนค่าสังเกตเพิ่มเติม m จะต้องมากกว่าจำนวน

ตัวแปรอิสระ + 1 นั่นคือ m>k แต่ถ้า m<k จะใช้การทดสอบอย่างอื่น

– การตั้งสมมติฐานที่จะใช้ตัวสถิตินี้ทดสอบ จะต้องเป็นแบบที่ว่า "พารามิเตอร์ถดถอย**ทั้งหมด**เปลี่ยนแปลงจากข้อมูลชุดหนึ่งไปยังข้อมูลหนึ่ง" สำหรับการทดสอบสมมติฐานเกี่ยวกับ**สัมประสิทธิ์**ถดถอยแต่เพียงบางส่วนจะไม่ขอกล่าว

ตัวอย่าง สมมติว่าเรามีตัวอย่าง 2 กลุ่ม ซึ่งเป็นข้อมูลเกี่ยวกับการบริโภค และรายได้ในคาบเวลา 1956-1965 และ 1966-1975 ดังตารางต่อไปนี้

ปี	รายได้ Y	การบริโภค C	ปี	รายได้ Y	การบริโภค C
1956	17500	12420	1966	22758	15362
57	18253	12690	67	23720	16080
58	18900	13050	68	24924	16735
59	19126	12863	69	25769	17127
60	19518	12876	70	25993	17517
61	20413	13450	71	27146	18375
62	21179	13995	72	28748	19182
63	21911	14559	73	29461	19421
64	22265	14682	74	30032	19811
65	22706	14985	75	30489	20211

ก. จากข้อมูลชุดแรกเราได้ฟังก์ชั่นการบริโภคเป็น

$$\hat{C}_1 = \hat{\beta}_1 + \hat{\beta}_2 Y$$

$$\hat{C}_1 = 3315.27 + 0.51Y \qquad R^2 = 0.958$$

$$(757.7) \quad (0.04)$$

$$SSE_1 = 323313 \qquad df = 10-2 = 8$$

ข. จากตัวอย่าง 2 หรือข้อมูลในคาบเวลา 1966-1975 เราได้ฟังก์ชั่นการบริโภคเป็น

$$C_2 = \hat{\delta}_1 + \hat{\delta}_2 Y$$

$$= 1545.57 + 0.61Y \qquad R^2 = 0.994$$

$$(421.7) \quad (0.01)$$

$$SSE_2 = 128552 \qquad df = 10-2 = 8$$

ค. จากตัวอย่างทั้งสองรวมกันเราจะได้ฟังก์ชั่นการบริโภคเป็น

$$\hat{C}_3 = \hat{\alpha}_1 + \hat{\alpha}_2 Y$$

$$= 850.23 + 0.63Y \qquad R^2 = .992$$

$$(323.7) \quad (0.01)$$

$SSE_c = 1062082$

$df = n + m - k = 10 + 10 - 2 = 18$

ดังนั้น $SSE_c - (SSE_1 + SSE_2) = 610217 \qquad df = 2$

$SSE_1 + SSE_2 = 451865 \qquad df = n+m-2k = 16$

$$F = \frac{SSE_c - (SSE_1 + SSE_2)/k}{(SSE_1 + SSE_2)/(n+m-2k)}$$

$$= \frac{610217/2}{451865/16} = 10.8$$

ค่าทางทฤษฎีของ F ณ ระดับนัยสำคัญ 5% และ df = 2,16 เป็น 3.63 ดังนั้นเราจึงปฏิเสธ Ho นั่นคือความสัมพันธ์ทั้งสองแตกต่างกันอย่างมีนัยสำคัญ หรือจะกล่าวได้ว่า ฟังก์ชั่นการบริโภคจะเปลี่ยนแปลงระหว่าง 2 คาบเวลา MPC จะเพิ่มขึ้นในคาบเวลาที่สอง (1966-1975)

(ฉ) สมมติฐานสุดท้ายที่เราจะทดสอบ คือ ความคงที่ (Stability) ของสัมประสิทธิ์ถดถอยเมื่อเพิ่มขนาดตัวอย่าง ถ้าเราต้องการทราบว่าค่าประมาณของสัมประสิทธิ์ถดถอยจะแตกต่างในตัวอย่างที่เพิ่มขยายขึ้นหรือไม่ หรือว่าจะคงที่ตลอดเวลาหรือไม่ บางครั้งอาจจะมีเหตุการณ์เกิดขึ้นและจะทำให้เปลี่ยนโครงสร้างของความสัมพันธ์ได้ เช่นมาตรการในการคุมกำเนิด การเปลี่ยนแปลงกฎหมายภาษีอากร เป็นต้น การเปลี่ยนแปลงต่าง ๆ อาจจะทำให้พารามิเตอร์ถดถอยไม่คงที่ได้และอาจจะฉับไว (Sensitive) ต่อการเปลี่ยนแปลงในส่วนประกอบของตัวอย่าง

ดังนั้นสมมติฐานที่เราจะทดสอบ คือ

Ho : $\beta_a = \alpha_a$; a = 1,2,....,k

ถ้าค่าสังเกตที่เพิ่มขึ้น (m) มากกว่าจำนวนพารามิเตอร์(หรือจำนวนตัวแปรอิสระ + 1)ในฟังก์ชั่นนั้น การทดสอบก็จะทำได้เช่นเดียวกับสมมติฐาน (จ) นั่นคือใช้ตัวสถิติทดสอบ

$$F_c = \frac{SSE_c - (SSE_1 + SSE_2)/k}{(SSE_1 + SSE_2)/(n + m - 2k)}$$

ถ้าจำนวนค่าสังเกตที่เพิ่มขึ้น (m) น้อยกว่าจำนวนพารามิเตอร์ในฟังก์ชั่น นั่นคือ m<k เราทำได้ดังนี้

— จากตัวอย่างที่ขยาย

$$SSE_c = \sum_{i=1}^{n+m} (Y - \hat{\alpha}_1 - \hat{\alpha}_2 X_{i2} - \hat{\alpha}_3 X_{i3} - \ldots - \hat{\alpha}_k X_{ik})^2$$

ซึ่งมีองศาความเป็นอิสระ $n+m-k$

จากตัวอย่างชุดเดิมเราจะได้ว่า

$$SSE_1 = \sum_{i=1}^{n} (Y - \hat{\beta}_1 - \hat{\beta}_2 X_{i2} - \hat{\beta}_3 X_{i3} - \cdots - \hat{\beta}_k X_{ik})^2$$

ซึ่งมีองศาความเป็นอิสระ $n-k$

เราสามารถแสดงได้ว่าตัวสถิติ

$$F_c = \frac{(SSE_c - SSE_1)/m}{SSE_1/(n-k)} \sim F^{(m,\ n-k)}$$

สมมติฐาน Ho เราจะได้รับการปฏิเสธ ณ ระดับนัยสำคัญ α ถ้าค่าของตัวสถิติ F_c มากกว่า $F^{(m,n-k)}$ นั่นคือเรายอมรับว่าสัมประสิทธิ์การถดถอยไม่คงที่ หรือค่าของมันเปลี่ยนแปลงในการขยายคาบเวลาของตัวอย่าง

ตัวอย่าง สมมติว่าตัวอย่างของสินค้าขาเข้า และรายได้ในคาบเวลา 1960-1971 เป็นดังนี้

ปี	สินค้าขาเข้า	รายได้ X	ปี	สินค้าเข้า	รายได้ X
1956	3748	21777	1964	4753	25886
57	4010	22418	65	5062	2686
58	3711	22308	66	5669	28134
59	4004	23319	67	5628	29091
60	4151	24180	68	5736	29450
61	4569	24893	69	5946	30705
62	4582	25310	70	6501	32372
63	4697	25799	71	6549	33152

ฟังก์ชั่นของมูลค่าสินค้าขาเข้าประมาณได้จากตัวอย่างนี้เป็น

$$\hat{Y} = -2011.85 + 0.26X \qquad R^2 = 0.984$$
$$(236.71) \quad (0.01)$$
$$SSE_1 = 208581$$

ถ้าเรารได้ค่าสังเกตอีก 4 ค่า ดังนี้

	สินค้าเข้า	รายได้
1972	6705	33764

73	7104	34411
74	7609	35429
75	8100	36200

ดังนั้นฟังก์ชั่นจะเป็น

$$\hat{Y} = -2461.38 + 0.28 \qquad R^2 = 0.983$$
$$(250.0) \quad (0.01)$$
$$SSE_2 = 573069$$

ค่าสังเกตของ F คำนวนได้เป็น

$$F_c = \frac{(SSE_2 - SSE_1)/m}{SSE_1/(n-k)} = \frac{364488/4}{573069/14}$$
$$= 6.12$$

เนื่องจาก $F^{(4,14)}_{.05} = 3.11$ เราจึงปฏิเสธ Ho นั่นคือมีความแตกต่างมีนัยสำคัญระหว่างพารามิเตอร์ของฟังก์ชั่นทั้งสอง สัมประสิทธิ์ของสินค้าขาเข้าไม่คงที่ ซึ่งฉับไวต่อการเปลี่ยนแปลงขนาดตัวอย่าง

(6) การทำนาย (Prediction) สำหรับตัวแปรอิสระชุดหนึ่งที่กำหนดให้ เราสามารถพยากรณ์ค่าของตัวแปรตามได้ นั่นคือถ้าให้ค่าของตัวแปรอิสระเป็น $X_{02}, X_{03}, \ldots, X_{ok}$ และให้ค่า Y ที่สอดคล้องของตัวแปรตามเป็น Y_o ซึ่ง Y_o เป็นค่าที่เราสนใจพยากรณ์ ตัวพยากรณ์ที่ดีที่สุดของ Y_o คือ $E(Y_o)$ เพราะว่าความแปรปรวนของ Y_o รอบ $E(Y_o)$ น้อยกว่ารอบจุดอื่น ๆ $E(Y_o)$ ไม่ทราบค่า เราจึงใช้ $\hat{Y}_o$ แทน ซึ่งเป็นค่าประมาณหรือค่าที่ปรับได้จากวิธีกำลังสองต่ำสุด

เนื่องจาก $\hat{Y}_o = \hat{\beta}_1 + \hat{\beta}_2 X_{02} + \ldots, \hat{\beta}_k X_{ok}$

และเราทราบว่า $\hat{Y}_o$ มีการแจกแจงปกติที่มีค่าเฉลี่ย และความแปรปรวน ดังนี้

$$E(\hat{Y}_o) = \beta_1 + \beta_2 X_{02} + \beta_3 X_{03} + \cdots + \beta_k X_{0k}$$

$$V(\hat{Y}_o) = \sigma^2/2 + \sum_a (X_{0a} - \bar{X}_a)^2 V(\hat{\beta}_a) + 2 \sum_{j<a} (X_{oj} - \bar{X}_j)(X_{oa} - \bar{X}_a) Cov(\hat{\beta}_j, \hat{\beta}_a)$$
$$(j, a = 2,3,\ldots k \; ; \; j<a)$$

หรือ

$$V(\hat{Y}_o) = \sigma^2 \left\{ (\underset{\sim}{X}_o - \underset{\sim}{\bar{X}})' (\underset{\sim}{X}' \underset{\sim}{X})^{-1} (\underset{\sim}{X}_o - \underset{\sim}{\bar{X}}) + 1/n \right\}$$

$$\text{ในเมื่อ } (\underset{\sim}{X}_o - \underset{\sim}{\bar{X}}) = \begin{pmatrix} X_{o2} - \bar{X}_2 \\ X_{o3} - \bar{X}_3 \\ \vdots \\ X_{ok} - \bar{X}_k \end{pmatrix}$$

สำหรับความคลาดเคลื่อนมาตรฐานของการทำนาย (Forecast Error) หรือ $(Y_o - \hat{Y}_o)$ จะเป็นตัวแปรเชิงสุ่มที่แจกแจงแบบปกติ โดยมีค่าเฉลี่ยและความแปรปรวน เป็น

$$E(Y_o - \hat{Y}_o) = 0$$

$$\sigma^2 = V(Y_o - \hat{Y}_o) = V(Y_o) + V(\hat{Y}_o) - 2Cov(Y_o, \hat{Y}_o)$$

แต่ $V(Y_o) = \sigma^2$, $V(\hat{Y}_o) = \sigma^2_{\hat{Y}_o}$, $Cov(Y_o, \hat{Y}_o) = 0$

ดังนั้น $$\sigma^2_f = \sigma^2 + \sigma^2/n + \sum (X_{oa} - \bar{X}_a)^2 V(\hat{\beta}_a) + 2 \sum_{j<a} (X_{oj} - X_j)(X_{oa} - X_a) Cov(\hat{\beta}_j, \hat{\beta}_a)$$

หรือ $$\sigma^2_f = \sigma^2 \{1 + 1/n + (\underset{\sim}{X}_o - \underset{\sim}{\bar{X}})' (\underset{\sim}{X}' \underset{\sim}{X})^{-1} (\underset{\sim}{X}_o - \underset{\sim}{\bar{X}})\}$$

ตัวประมาณค่าที่ไม่เอียงเฉของ σ^2_f คือ S^2_f ซึ่งได้จากการแทน σ^2 ด้วย S^2 แล้วเราจะได้ว่า

$$(\hat{Y}_o - Y_o)/S_f \sim t^{(n-k)}$$

จากผลอันนี้ทำให้เราสามารถสร้างช่วงพยากรณ์ที่ครอบคลุมค่าของ Y_o ด้วยความน่าจะเป็นที่ต้องการได้ นั่นคือ

$$\hat{Y}_o - t^{(n-k)}_{\alpha/2} S_f < Y_o < \hat{Y}_o + t^{(n-k)}_{\alpha/2} S_f$$

สำหรับตัวสถิติ $(Y_o - Y_o)/s_f$ นี้สามารถใช้ทดสอบสมมติฐานที่ว่า ค่าสังเกตใหม่ (เช่น ค่าสังเกตที่ $n + 1$ มาจากประชากรเดียวกับ n ค่าสังเกตที่ใช้ประมาณค่าพารามิเตอร์ถดถอย

ตัวอย่าง ตามตัวอย่างเกี่ยวกับส้ม เราประมาณสมการถดถอยที่อธิบายถึงอุปสงค์ของส้มจากราคาและค่าโฆษณา ได้เป็น

$$\hat{Y}_i = 117.532 - 1.326 X_{i2} + 11.237 X_{i3}$$

ถ้าสมมติว่าไม่มีส้มขาย แต่ส้มที่สั่งมาใหม่กว่าจะส่งมาถึงก็ใช้เวลาหลายสัปดาห์ **เมื่อส้ม**มาถึง วันแรกที่ขายได้พบว่า, $Y_o = 100$, $X_{02} = 80$, $X_{03} = 7$

ปัญหาที่จะพิจารณาก็คือ ฟังก์ชั่นอุปสงค์ได้เปลี่ยนไปหรือไม่

เราได้ $\hat{Y}_o = 117.532 - 1.326(80) + 11.237(7)$

$= 90.111$

ซึ่งเป็นค่าพยากรณ์ของ Y ความแปรปรวนที่ประมาณได้ของความคลาดเคลื่อนพยากรณ์ จะเป็น

$$S_f^2 = S^2 + S^2/n + (X_{02}-\bar{X}_2)^2 S^2_{\hat{\beta}_2} + (X_{03}-\bar{X}_3)^2 S^2_{\hat{\beta}_3} + 2(X_{02}-\bar{X}_2)(X_{03}-\bar{X}_3)\mathrm{Cov}(\hat{\beta}_2,\hat{\beta}_3)$$

จากการคำนวนก่อน ๆ เราได้

$$S^2 = 20.586 \qquad \bar{X}_2 = 70 \qquad \bar{X}_3 = 6.7$$
$$S^2_{\hat{\beta}_2} = 0.012476 \qquad S^2_{\hat{\beta}_3} = 5.7761$$
$$\mathrm{Cov}(\hat{\beta}_2,\hat{\beta}_3) = 0.138628$$

ดังนั้น
$$S_f^2 = 20.586 + 20.586/12 + (80-70)^2(.012476) + (7-6.7)^2(5.7761) + 2(80-70)(7-6.7)(.138628)$$
$$= 75.808974$$
$$S_f = 8.706$$

ช่วงเชื่อมั่น 95% สำหรับ Y_o สร้างได้ดังนี้

$$\hat{Y}_o - t^{(12-3)}_{.025} S_f < Y_o < \hat{Y}_o + t^{(12-3)}_{.025} S_f$$

$$90.111 - 2.262(8.706) < Y_o < 90.11 + 2.262(8.706)$$
$$70.418 < Y_o < 109.804$$

เราจะเห็นได้ว่าค่าสังเกตของ Y_o เป็น 100 นี้ตกอยู่ในช่วงนี้ แสดงว่าฟังก์ชั่นอุปสงค์ได้เปลี่ยนแปลงไปเลย