“การประยุกต์การวิเคราะห์ถดถอยในการวิจัยเชิงประจักษ์”

โดย

ดร.เทียนฉาย กีระนันทน์

1. ความนำ

ในการวิจัยเชิงประจักษ์ (empirical research) ทางสาขาวิชาเศรษฐศาสตร์นั้น โดยปกติแล้วมักจะนิยมที่จะใช้เครื่องมือทางคณิตศาสตร์ และสถิติ เข้ามาช่วย เพื่อให้การวิเคราะห์นั้นง่ายขึ้น และส่งผลให้การวิเคราะห์เป็นไปอย่างแม่นยำสมเหตุผลที่สามารถอธิบายได้มากขึ้น เครื่องมือทางสถิติที่สำคัญที่นิยมใช้กันอยู่ ก็คือการวิเคราะห์ถดถอย (regression analysis) โดยจะเห็นได้ชัดว่า การวิเคราะห์ถดถอยนั้น ได้รับการพัฒนาทางเทคนิควิธีการให้สามารถประยุกต์ และใช้เป็นเครื่องมือที่ตอบคำถามในเชิงสถิติได้กว้างขวางยิ่งขึ้นเรื่อยมา โดยเฉพาะในสาขาวิชาเศรษฐศาสตร์ จนกระทั่งได้รับการพัฒนาขึ้นเป็นส่วนหนึ่งของเศรษฐศาสตร์วิเคราะห์ ที่เรียกกันว่า เศรษฐมิติ (econometrics)

จากเหตุ และผลของการวิเคราะห์ และจากความสามารถในการอธิบายความสัมพันธ์ระหว่างตัวแปรต่างๆ ที่เป็นเหตุ กับตัวแปรที่เป็นผลของเหตุการณ์หนึ่งๆ นี้เอง การวิเคราะห์ถดถอยจึงได้รับความสนใจจากศาสตร์อื่นๆ ทางสังคมศาสตร์เพิ่มมากขึ้น (พร้อมๆ กับที่ศาสตร์สาขาอื่นๆ ทางสังคมศาสตร์นั้น ก็ได้ค้นหา และพัฒนาเครื่องมือทางสถิติวิธีอื่นๆ ขึ้นมาด้วย) ความพยายามที่จะใช้เครื่องมือทางสถิติเข้าช่วยในการวิเคราะห์และอธิบายพฤติกรรมของมนุษย์ในทางสังคมศาสตร์หลายๆ สาขานี้เอง ที่เป็นผลให้มีการพัฒนา ส่วนของสาขาวิชาต่างๆ ขึ้นมาตามลำดับ เช่น สังคมวิทยาได้พัฒนา สังคมมิติ (Sociometrics) หรือทางจิตวิทยาได้พัฒนา psychometrics ขึ้นเป็นต้น

แม้ว่าเครื่องมือทางสถิติที่มีอยู่ และที่กำลังได้รับการค้นคว้าพัฒนาขึ้นใหม่นี้ จะมีอยู่มากมายหลายวิธี และแต่ละเครื่องมือก็มีข้อดี ข้อเสียแตกต่างกันไป โดยที่คงจะต้องยอมรับกันก่อนด้วยว่า แต่ละวิธีนั้นต่างมีจุดอ่อนด้วยกันทั้งนั้นก็ตาม การวิเคราะห์ถดถอยก็ยังคงได้รับการยอมรับว่า เป็นเครื่องมือทางสถิติที่ช่วยตอบคำถามในการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร และแม้แต่ในการคาดการณ์เหตุการณ์ในอนาคตได้เป็นอย่างมาก และมีข้อได้เปรียบที่ดีและกว้างขวางกว่า เครื่องมืออีกหลายๆ วิธีที่ใช้กันอยู่ในทางสังคมศาสตร์

ตำรา และคู่มือที่แสดง และอธิบายการวิเคราะห์ถดถอยในปัจจุบันนี้มีอยู่มาก โดยเฉพาะที่เป็นตำราภาษาอังกฤษ ที่ได้รับการยอมรับว่าดีถึงขนาดที่ใช้เป็นแม่บท และอ้างอิงได้ เช่นตำราที่เขียนเป็นความรู้เบื้องต้นโดย Wonnacott and Wonnacott หรือที่เขียนอธิบายลึกซึ้งสูงขึ้น โดย Klein, Theik, Goldberger, Kmenta และ

Johnston[1] เป็นต้น ตำราเหล่านี้ส่วนใหญ่ หรือเกือบทั้งหมดจะอธิบายหลัก และทฤษฎีตลอดจนการพิสูจน์ที่มาของกฎ และหลักในการวิเคราะห์ถดถอย ซึ่งมักจะมีเนื้อหาความที่เป็นรายละเอียด และเป็นทฤษฎีที่ต้องใช้ทั้งความรู้ทางคณิตศาสตร์ และแคลคูลัสค่อนข้างมาก นับว่าเป็นประโยชน์อย่างยิ่งสำหรับผู้ที่ศึกษาสาขาเศรษฐมิติเป็นสาขาหลักในทางเศรษฐศาสตร์ อย่างไรก็ดี ในแง่ของนักวิจัย หรือผู้ที่ศึกษาในสาขาวิชาเศรษฐศาสตร์สาขาอื่นๆ ที่ไม่ได้เน้นหนักทางเศรษฐมิติโดยเฉพาะ แต่ประสงค์จะใช้การวิเคราะห์ถดถอย เป็นเครื่องมือทางสถิติสำหรับงานค้นคว้าวิจัยของตนแล้ว มักจะประสบปัญหา กับการทำความเข้าใจหลักการ และทฤษฎีเหล่านี้อยู่เสมอ และค่อนข้างมาก ตำราที่อธิบายตรงไปตรงมาถึงการใช้ การประยุกต์ วิธีการใช้ และข้อจำกัดของการวิเคราะห์ถดถอย ที่สามารถอ่าน และทำความเข้าใจได้อย่างธรรมดาๆ โดยง่าย โดยไม่ต้องทำความเข้าใจกับ matrix algebra และแคลคูลัส ก่อนด้วยนั้น เกือบจะเรียกได้ว่าหาไม่ได้เลย การวิเคราะห์ถดถอยจึงดูเป็นเรื่องยาก สำหรับผู้ที่ขาดพื้นฐานทาง matrix algebra และแคลคูลัส และยิ่งกว่านั้น ยังทำให้เกิดความรู้สึกว่า ผู้ที่ใช้การวิเคราะห์ถดถอยนั้น เป็นพวกที่เดินไม่ติดดิน เพราะอธิบายสิ่งที่ง่ายให้ยาก และดูเหมือนจะใช้อธิบายพฤติกรรมธรรมดาๆ ไม่ได้เลย ซึ่งอันที่จริงแล้วการวิเคราะห์ถดถอยนั้น ไม่ใช้เรื่องยากที่จะเข้าใจ และนำมาประยุกต์ในฐานะที่เป็นเครื่องมือทางสถิติ โดยเฉพาะในสมัยนี้ ซึ่งมีคอมพิวเตอร์ที่เอื้อในด้านบริการเกี่ยวกับโปรแกรมสำเร็จรูป (package program) ที่ผู้ใช้สามารถเรียกใช้เครื่องมือการวิเคราะห์ถดถอยได้โดยง่าย และประโยชน์ที่จะได้รับจากการวิเคราะห์ถดถอยก็มีมากเกินพอ และคุ้มค่าที่จะเสียเวลาทำการศึกษา และเข้าใจ

ประเด็นสำคัญ สำหรับผู้ใช้การวิเคราะห์ถดถอย เป็นเครื่องมือโดยไม่สนใจที่จะศึกษาให้ลึกซึ้งถึงหลักและทฤษฎีเศรษฐมิตินั้น จะอยู่ที่ว่าจะต้องแน่ใจว่าตนเข้าใจหลักการวิธีการ คำอธิบาย และความสามารถในการวิเคราะห์ของการวิเคราะห์ถดถอยเป็นอย่างดีมากพอ มิฉะนั้นก็จะตกอยู่ในภาวะที่ใช้เครื่องมือผิดวิธี โดยอาจคิดไปได้ว่า การวิเคราะห์ถดถอยคือของวิเศษ ที่ใช้ทำการวิเคราะห์อะไรก็ได้

บทความนี้จึงมีวัตถุประสงค์ ที่จะแนะนำเทคนิควิธีการในการวิเคราะห์ถดถอยอย่างง่ายๆ โดยใช้ภาษาบรรยายธรรมดา และหลีกเลี่ยงคณิตศาสตร์และแคลคูลัสเท่าที่จะทำได้ เพื่อให้เป็นประโยชน์แก่ผู้สนใจที่อยากจะทราบถึงวิธีการเศรษฐมิติขั้นต้น โดยไม่ประสงค์จะอ่านตำราที่มีความยาก และเป็นคณิตศาสตร์ และโดยเฉพาะสำหรับผู้ที่ประสงค์จะใช้การวิเคราะห์ถดถอย เป็นเครื่องมือในการวิจัย หรือวิเคราะห์ทางสังคมศาสตร์บ้าง อย่างไรก็ตาม การเขียนอธิบายในบทความนี้ ยังมีข้อจำกัดซึ่งผู้อ่านควรตั้งเป็นข้อสังเกตไว้ก่อนด้วยว่า ตัวอย่างที่จะใช้ หรือกรณีวิจัยเชิงประจักษ์ที่จะอ้างถึงในบทความนี้ จะเป็นเรื่องที่เกี่ยวข้องกับตัวแปรทางเศรษฐศาสตร์และประชากรเป็นส่วนใหญ่ ซึ่งไม่ได้มีเหตุผลอื่นใดมากไปกว่าเป็นเพราะความสนใจของผู้เขียนในสาขานี้เองเท่านั้น ในตอนแรกของบทความนี้จะได้บรรยายถึงลักษณะ และความหมายของการวิเคราะห์ถดถอยในการวิจัยเชิง

---

[1] Ronald J. Wonnacott and Thomas H. Wonnecott , Econometrics (New York : John Wiley & Sons, 1970) ; Lawrence R.Klein, An Introduction to Econometrics (New jersey : Prentice-Hall, 1962) ; Henri Theil, Principles of Econometrics (New York : John Wiley & Sons, 1971) ; Arthur S.Goldberger, Econometic Theory (New York : John Wiley & Sons, 1964) ; Jan Kmenta, Elements of Econometrics (New York : Macmillan, 1979) ; J.Johnston, Econometrics Methods (New York : McGraw-Hill, 1963)

ประจักษ์ทางสังคมศาสตร์ และในตอนท้ายสุดจะได้สรุปถึงแนวทางการเตรียมตัว และแก้ปัญหาในการประยุกต์การวิเคราะห์ถดถอยเมื่อพบปัญหานั้นๆ ในการวิเคราะห์

๒. <u>ลักษณะและความหมายของการวิเคราะห์ถดถอย</u>

ในเบื้องต้นนี้ อาจอธิบายได้อย่างง่ายๆ ว่า การวิเคราะห์ถดถอยเป็นการศึกษาวิเคราะห์ เพื่อหาความสัมพันธ์ในเชิงสถิติของตัวแปร 2 ตัวแปร โดยที่ตัวแปรหนึ่งเป็นตัวแปรอิสระ และอีกตัวแปรหนึ่งเป็นตัวแปรตามในรูปของฟังก์ชั่นก็คือ

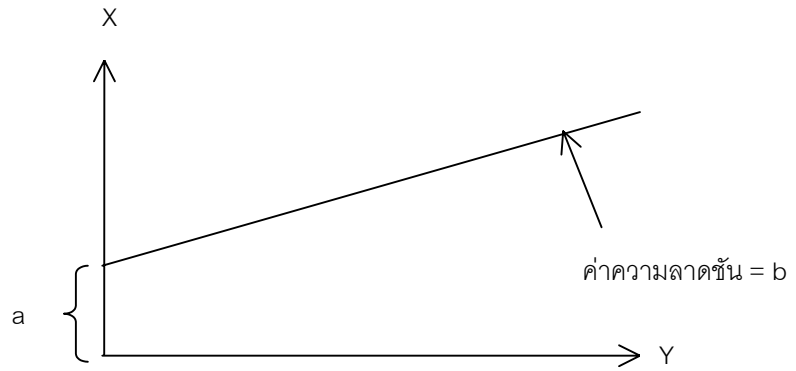
X เป็นฟังก์ชั่นของ Y หรือ $X = f(Y)$

หรือในรูปของสมการถดถอยก็คือ

$$X = a + bY$$

เมื่อ X เป็นตัวแปรตาม และ Y เป็นตัวแปรอิสระ ซึ่งหมายความเท่าๆ กันว่า X ขึ้นอยู่กับ Y

ในการแสดงการอธิบายในรูปกร๊าฟก็อาจเขียนได้ว่า (ถ้า a และ b มีค่าเป็นบวก)



ค่าของพารามิเตอร์ a คือค่าคงที่ที่อยู่นอกเหนืออิทธิพลของตัวแปรอิสระ เมื่อเริ่มต้นที่ศูนย์ (หรือ intersect) ส่วนค่าของ b คือค่าความลาดชัน (Slope) ของเส้นที่แสดงความสัมพันธ์ระหว่าง X กับ Y

ทั้งนี้ในการวิจัยเชิงประจักษ์ เมื่อผู้วิจัยทราบค่าของ X_i และ Y_i จากข้อมูลตัวอย่างจำนวนหนึ่งแล้ว ก็จะสามารถประมาณค่าของ a และ b ได้ (ส่วนวิธีการหาค่าของ a และ b ในทางสถิตินั้น อาจดูวิธีคำนวณได้จากตำราคู่มือทางเศรษฐมิติเบื้องต้นทั่วไป)

นอกจากการอธิบายในเบื้องต้นนี้แล้ว การวิเคราะห์ถดถอย ยังอาจแสดงการวิเคราะห์หาค่าความสัมพันธ์ ระหว่างตัวแปรตามหนึ่งตัวแปร เมื่อมีตัวแปรอิสระที่เป็นตัวกำหนดหลายๆ ตัวแปรได้ เราเรียกการวิเคราะห์ถดถอยประเภทนี้ว่า การวิเคราะห์ถดถอยพหุ (multiple regression analysis) กล่าวคือเมื่อ X ขึ้นอยู่กับ Y_1 , Y_2, Y_3, ... , Y_n ซึ่งอาจเขียนในรูปฟังก์ชั่นได้ว่า

$$X = f(Y_1, Y_2, Y_3, \ldots, Y_n)$$

หรือเขียนในรูปสมการถดถอย

$$X = a + b_1Y_1 + b_2Y_2 + b_3Y_3 + \ldots + b_nY_n)$$

ในกรณีของการถดถอยพหุเช่นนี้ การแสดงคำอธิบายด้วยกร๊าฟ ก็อาจทำได้แต่จะยากมาก และซับซ้อนเพราะเหตุที่จะมีแกน (dimension) มากกว่า 2 แกน และการแสดงความสัมพันธ์ระหว่างตัวแปรตาม กับชุดของตัวแปรอิสระจะเป็น plain แทนที่จะเป็นเส้นเหมือนในกรณีการถดถอยธรรมดา

ประเด็นสำคัญที่สุดก็คือ การตีความหมายของความสัมพันธ์ระหว่างตัวแปรตาม กับชุดของตัวแปรอิสระ ในกรณีของการถดถอยพหุ ซึ่งมักจะตีความหมายผิดไปเสมอๆ แท้จริงแล้วจะหมายความว่า “เมื่อพิจารณาตัวแปรอิสระหลายๆ ตัว ที่มีอิทธิพลต่อตัวแปรตามพร้อมๆ กันแล้ว ถ้าหากกำหนดให้ตัวแปรอิสระอื่นๆ ทั้งหมดคงที่ การแปรผันในตัวแปรอิสระตัวหนึ่งตัวใด โดยเฉพาะจะมีอิทธิพลในเชิงสถิติต่อการผันแปรในตัวแปรตามหรือไม่ และมากน้อยเพียงใด” เช่นจากรูปสมการถดถอยข้างต้น อาจตีความหมายเป็นตัวอย่างได้ว่า ถ้าหากกำหนดให้ Y_1 , Y_2, Y_3, ... , Y_n แล้ว การแปรผันใน Y_1 จะมีอิทธิพลต่อการผันแปรใน X หรือไม่ เพียงใด

ในกรณีดังเช่นที่อธิบายความหมายข้างต้นนี้ จึงแน่นอนว่า ถ้าหากผู้วิจัยจะเปลี่ยนชุดของตัวแปรอิสระแล้ว อิทธิพลที่ตัวแปรอิสระตัวหนึ่ง จะสามารถอธิบายตัวแปรตามได้ ก็จะเปลี่ยนแปลงไปด้วย เช่น

ก. $X = f(Y_1, Y_2, Y_3)$ กับ

$X = f(Y_1, Y_3, Y_4)$

อิทธิพลที่การแปรผันใน Y_1 จะมีผลต่อการผันแปรใน X จะแตกต่างกันหรือ

ข. $X = f(Y_1, Y_2, Y_3)$

$X = f(Y_1, Y_3)$

กรณี ข. นี้ อิทธิพลที่การแปรผันใน Y_1 จะมีค่าการแปรผันใน X จะเปลี่ยนแปลงไปเช่นกัน [2]

การวิเคราะห์ถดถอยนั้น นอกจากจะแสดงความสัมพันธ์เชิงสถิติว่า ตัวแปรตามมีความสัมพันธ์ หรือได้รับอิทธิพลในการแปรผันจากตัวแปรอิสระ เช่นเดียวกับการวิเคราะห์สหสัมพันธ์ (correlation) ตามธรรมดาแล้ว การวิเคราะห์ถดถอยยังแสดงถึงทิศทาง (direction) ของความสัมพันธ์ หรืออิทธิพลนั้นๆ ด้วย กล่าวคือ ในการวิเคราะห์ถดถอย จะสามารถบอกได้ว่าตัวแปรนั้น จะแปรผันในทิศทางเดียวกัน (คือมีค่าเป็นบวก) หรือจะแปรผันในทางกลับกัน (คือมีค่าเป็นลบ) กับตัวแปรอิสระแต่ละตัวในชุดของตัวแปรอิสระทั้งหมดได้ด้วย เช่น

$$X = a - b_1Y_1 + b_2Y_2$$

ก็แสดงว่า X นั้นแปรกลับกันกับ Y_1 และแปรตาม Y_2 หรือถ้าหาก Y_1 เพิ่มขึ้น X จะลดลง แต่ถ้า Y_1 ลดลง X จะกลับเพิ่มขึ้น ในทำนองกลับกัน X จะเพิ่มขึ้น เมื่อ Y_2 เพิ่มขึ้น และจะกลับลดลงเมื่อ Y_2 ลดลง

นอกจากทิศทางของความสัมพันธ์ดังกล่าวแล้ว การวิเคราะห์ถดถอยยังแสดงถึงขนาด (magnitude) ของความสัมพันธ์ระหว่างตัวแปรตาม กับตัวแปรอิสระอีกด้วย กล่าวคือ ในการวิเคราะห์ถดถอยนั้น จะประมาณค่า regression coefficient (คือ b_1 และ b_2 ในสมการตัวอย่างที่แล้วนั้น) ค่าของ coefficient ที่ประมาณได้สำหรับบทบาท หรืออิทธิพลของตัวแปรอิสระแต่ละตัวนั้น จะแสดงถึงขนาดของบทบาท หรืออิทธิพลที่ตัวแปรอิสระนั้นๆ มีต่อตัวแปรตาม ค่าของ coefficient โดดๆ จะแสดงถึงค่าความลาดชันของเส้นถดถอย ซึ่งกล่าวง่ายๆ ก็คือ Y_i จะมีอิทธิพลเพียงใดในการกำหนด X

อย่างไรก็ดี ในเรื่องของขนาดของความสัมพันธ์ ที่ได้จากการวิเคราะห์ถดถอยนี้ ในทางเศรษฐศาสตร์แล้ว จะใช้ประโยชน์ได้กว้างขวางมากในอีกแง่หนึ่ง กล่าวคือ จะไม่เพียงแต่ประมาณค่า coefficient เพื่อแสดงขนาดของความสัมพันธ์ตามปกติ เพราะจะให้ความหมายในเชิงนโยบายค่อนข้างจำกัด แต่จะคำนวณต่อเนื่องไปถึงค่าความยืดหยุ่น (elasticity) ณ ค่าเฉลี่ยของตัวแปรด้วย กล่าวคือ อาจคำนวณค่าความยืดหยุ่นได้โดย

---

[2] ประเด็นนี้อาจสังเกตได้ว่า ผลของการประมาณค่าที่วิเคราะห์ได้จากสมการถดถอยจะแตกต่างกันไปตามรูปแบบของชุดของตัวแปรอิสระที่นำเข้ามาพิจารณา ซึ่งการกำหนดชุดของตัวแปรอิสระนี้ จัดอยู่ในกระบวนการสร้างแบบจำลอง (model specification) ดังนั้น การที่ผู้วิจัยจะสร้างแบบจำลองที่เหมาะสม และถูกต้องได้ จึงเป็นศิลปของผู้วิจัยส่วนหนึ่ง และเป็นความรู้ประสบการณ์ กับความสามารถของผู้วิจัยส่วนหนึ่ง การสร้างแบบจำลองที่ไม่ถูกต้องเหมาะสม จึงเกิดขึ้นได้เสมอกับนักวิจัยใหม่ๆ และในหลายๆ กรณีก็เป็นจุดที่สามารถชี้ให้เห็นถึงข้อบกพร่องของการวิจัยทั้งโครงการได้เหมือนกัน ส่วนการสร้างแบบจำลอง โดยเลือกตัวแปรอิสระที่เหมาะสมนั้น จะต้องมาจากทฤษฎี แนวความคิด ตลอดจนผลงานวิจัยในอดีต ซึ่งผู้วิจัยจะต้องศึกษา และทำวรรณกรรมปริทัศน์ไว้เป็นอย่างดีก่อนที่จะสร้างแบบจำลองด้วย

$$\eta_{Y on X} = \frac{\frac{\Delta X}{\overline{X}}}{\frac{\Delta Y_i}{\overline{Y_i}}}$$

$$= \frac{\Delta X}{\Delta Y_i} \times \frac{Y_i}{\overline{X}}$$

ซึ่งค่าของ $\Delta X / \Delta Y_i$ ก็คือค่า coefficient ที่ประมาณได้จากการวิเคราะห์ถดถอย และค่า $\overline{X}$ คือค่าเฉลี่ยของ X กับ $\overline{Y_i}$ คือค่าเฉลี่ยของ Y_i ทั้งนี้จะสื่อความหมายว่า ถ้าหาก Y_i เปลี่ยนแปลงไป 100 เปอร์เซ็นต์ (หรือเท่าตัว) แล้ว X จะเปลี่ยนแปลงไปในทางเดียวกัน หรือกลับกันเป็นกี่เปอร์เซ็นต์ ประโยชน์ในการคำนวณค่าความยืดหยุ่นนี้ก็คือ ใช้ในการพิจารณาถึงนโยบายทางเศรษฐกิจได้โดยตรง เช่น ถ้าต้องการเปลี่ยนแปลงค่า X สัก 20 เปอร์เซ็นต์ จะต้องเพิ่ม หรือลด Y_i สักกี่เปอร์เซ็นต์

ตัวอย่างเช่น จากการศึกษาหนึ่ง[3] พบว่า ระดับการศึกษาของบุตรคนโต (SCO) ขึ้นอยู่กับการศึกษาของบิดา (SM) และการศึกษาของมารดา (SF) โดยประมาณค่าจากการวิเคราะห์ถดถอยได้ว่า

SC = 4.8034 + 0.3900 SM + 0.3765 SF

โดยที่ค่าเฉลี่ยปรากฏดังนี้ $\overline{SC}$ = 7.2950 $\overline{SM}$ = 3.9075 และ $\overline{SF}$ = 2.5700 ผลจากการวิเคราะห์ถดถอยนี้ แสดงว่าระดับการศึกษาของบิดา และระดับการศึกษาของมารดา มีอิทธิพลในการกำหนดระดับการศึกษาของบุตรคนโตในทิศทางเดียวกัน กล่าวคือ โดยเฉลี่ยแล้ว ถ้าหากการศึกษาของบิดาเพิ่มขึ้นหนึ่งเท่าตัวหรือ 100 เปอร์เซ็นต์ (คือจากประมาร 4 ปีเป็น 8 ปี) แล้ว ระดับการศึกษาของบุตรคนโตจะเพิ่มขึ้นได้ 0.3900 x (3.9075 / 7.2950) หรือ 20.89 เปอร์เซ็นต์ และโดยเฉลี่ยอีกเช่นกันเมื่อการศึกษาของมารดาเพิ่มขึ้น 100 เปอร์เซ็นต์แล้ว ระดับการศึกษาของบุตรคนโตจะเพิ่มขึ้น 0.3765 x (2.5700/7.2950) หรือ 13.26 เปอร์เซ็นต์ ซึ่งข้อค้นพบจากการวิเคราะห์ในทำนองนี้ จะเป็นสิ่งที่ให้ความหมายโดยตรงต่อผลในการกำหนดนโยบาย ซึ่งอาจนำไปใช้ได้จากการวิจัยนั้นๆ

ประเด็นสุดท้ายที่ผู้วิจัย จะต้องคำนึงถึงในการวิเคราะห์ถดถอย ก็คือ การทดสอบการประมาณที่ได้ในเชิงสถิติทั้งหมด ค่าสถิติที่ได้จากการวิเคราะห์ถดถอยที่สำคัญๆ และจะต้องใช้เป็นประจำเป็นปกตินั้นมีอยู่ด้วยกัน 3 ประการด้วยกัน กล่าวคือ

---

[3] Thienchay Kiranandana , The Demand for Children : An Application of the New Home Economics Approach to Thai Data, (Bangkok : Institute of Population Studies, Paper # 27, 1978) ,P.79.

ก. t-statistics จะเป็นสถิติที่ใช้ทดสอบค่านัยสำคัญทางสถิติของค่าประมาณที่คำนวณได้สำหรับ Coefficient ในตัวแปรอิสระแต่ละตัวแปร ในกรณีที่ค่า t-statistics ที่ได้จากการประมาณค่า coefficient นั้น เมื่อทดสอบแล้วไม่ปรากฏว่ามีนัยสำคัญทางสถิติ จะให้ความหมายว่า การแปรผันในตัวแปรอิสระนั้นๆ มีผลกระทบต่อการแปรผันในตัวแปรตามไม่แตกต่างไปจากศูนย์ ซึ่งก็เท่ากับว่าตัวแปรอิสระนั้นๆ ไม่น่าจะมีผลกระทบ หรือมีอิทธิพลต่อตัวแปรตาม แต่ถ้าหากค่า t-statistic ที่ได้นั้น เมื่อทดสอบแล้วปรากฏว่ามีนัยสำคัญ ณ ระดับนัยสำคัญที่สูงมากพอ จะให้ความหมายเพียงแต่ว่า ภายใต้ชุดของตัวแปรอิสระทั้งชุดที่กำหนดในแบบจำลองนั้น เมื่อนำเข้ามาพิจารณาร่วมกันแล้ว และเมื่อตัวแปรอิสระทั้งหลายถูกกำหนดไว้ ตัวแปรอิสระเฉพาะตัวที่ประมาณค่า Coefficient นั้นน่าจะมีอิทธิพล และบทบาทในการแปรผันของตัวแปรตาม อย่างมีนัยสำคัญทางสถิติ อย่างไรก็ตามในปัจจุบันเมื่อการประมาณค่าในการวิเคราะห์ถดถอยส่วนใหญ่ จะกระทำได้ง่ายๆ โดยใช้โปรแกรมสำเร็จรูปทางคอมพิวเตอร์เข้าช่วย ก็ควรสังเกตด้วยว่าในหลายกรณีที่โปรแกรมสำเร็จรูปบางโปรแกรม จะไม่ให้ค่า t-statistics มาโดยตรง แต่จะให้ค่าความคลาดเคลื่อนมาตรฐาน (Standard error) มาแทน ซึ่งก็จะหาค่า t-statistics ได้โดยการหารค่าประมาณของ Coefficient ด้วยค่าความคลาดเคลื่อนมาตรฐานนั่นเอง

ข. F-statistics ซึ่งนอกเหนือจากค่า t-statistics ที่จะใช้ในการทดสอบหาความสัมพันธ์อย่างมีนัยสำคัญทางสถิติ ระหว่างตัวแปรอิสระ กับตัวแปรตามแต่ละคู่แล้ว ค่า F-statistics จะเป็นค่าสถิติที่อาจใช้ทดสอบความเหมาะสมของแบบจำลอง หรือชุดของตัวแปรอิสระทั้งชุดที่กำหนดขึ้นใช้ทดสอบนั้นด้วย การที่ทดสอบค่า F-statistics ที่คำนวณได้นั้นแล้วปรากฏว่ามีนัยสำคัญทางสถิติ ก็จะพอช่วยเพิ่มน้ำหนักในการอธิบายได้ว่าแบบจำลอง หรือชุดของตัวแปรอิสระนั้นๆ ใช้ได้ตามสมควร แต่ถ้าเมื่อทดสอบแล้วปรากฏว่า ไม่ปรากฏว่ามีนัยสำคัญทางสถิติ ก็จะเท่ากับว่าแบบจำลอง หรือชุดของตัวแปรอิสระที่กำหนดขึ้น อธิบายการผันแปรในตัวแปรตามนั้น ยังไม่เหมาะสมนัก หรือยังใช้ไม่ได้ดีนัก [4]

ค. ค่า R^2 ซึ่งจะช่วยบอกถึงระดับความสามารถในการอธิบาย ซึ่งแบบจำลอง หรือชุดของตัวแปรอิสระที่กำหนดนั้น จะสามารถอธิบายการแปรผันในตัวแปรตามได้มากน้อยเพียงใด สิ่งที่ควรสังเกตในที่นี้ก็คือ ถึงแม้ค่า R^2 จะมีประโยชน์อยู่มากในการบอกระดับความสามารถในการอธิบายการแปรผัน แต่ค่า R^2 ก็มิใช่สิ่งที่สำคัญ และจำเป็นมากนักในแง่ที่ว่า ไม่จำเป็นเลยที่จะต้องพอใจกับการวิเคราะห์ถดถอยที่ให้ค่า R^2 สูงๆ หรือมีค่าใกล้เคียง 1.0 (R^2 จะมีค่าระหว่าง 0 ถึง 1) ขณะเดียวกัน ถ้าหากการวิเคราะห์ถดถอยนั้น ให้ค่า R^2 ต่ำมากๆ ก็ไม่ใช่เรื่องที่น่าวิตกนัก ทั้งนี้เพราะเหตุผลง่ายๆ ว่า ในทางเศรษฐศาสตร์นั้น จริงอยู่แม้ว่าเราต้องการอธิบายสาเหตุของการแปรผันในตัวแปรตามให้ได้ แต่ในหลายๆ กรณี เราต้องการเพียงที่จะชี้ให้เห็นว่า อิทธิพลของตัวแปรอิสระเฉพาะบางตัวนั้น มี หรือ ไม่มี และมากน้อยเพียงใด ดังนั้น ถ้าผลปรากฏว่าตัวแปรอิสระที่เป็นหลักๆ นั้น จะสามารถอธิบายการแปรผันในตัวแปรตามได้เพียงไม่กี่เปอร์เซ็นต์ เพียงเท่านั้น ก็เป็นที่พอใจแล้ว แต่ตรงกันข้ามกับในกรณีที่ต้องการวิเคราะห์หาสาเหตุของการผันแปรในตัวแปรตามให้ได้ ในกรณีเช่นนั้นก็จะต้องพยายามสร้าง

---

[4] ควรสังเกตด้วยว่า การทดสอบค่า F-statistic นั้น จะต้องใช้ค่า degree of freedom 2 ค่าพร้อมๆ กัน ค่าหนึ่งสำหรับจำนวนตัวแปรอิสระ และอีกค่าหนึ่งสำหรับจำนวนตัวอย่างที่ใช้ ส่วนวิธีการทดสอบ และวิธีการใช้นั้น น่าจะศึกษาทบทวนได้จากตำราทางสถิติเบื้องต้นได้

แบบจำลองให้ครอบคลุมชุดของตัวแปรอิสระให้มากที่สุด เพื่อให้ได้ค่า R^2 สูงๆ ซึ่งจะแสดงว่าประสบผลสำเร็จในการวิเคราะห์หาสาเหตุนั้นๆ ได้ ยกตัวอย่างเช่น การวิเคราะห์อุปสงค์ต่อสินค้า ซึ่งโดยทฤษฎีเศรษฐศาสตร์นีโอคลาสสิคแล้ว ปัจจัยสำคัญที่เป็นตัวกำหนดอุปสงค์ต่อสินค้า ก็คือ รายได้ของผู้บริโภค และราคาของสินค้านั้น (คือ income effect และ price effect) และแท้จริงแล้วทฤษฎีนั้นยังได้กำหนดไว้ว่า รสนิยม (taste) ของผู้บริโภคคงที่โดยตลอด ซึ่งสิ่งที่สะท้อนถึงรสนิยมของผู้บริโภคนั้น มีมากมายหลายประการ เป็นต้นว่า การศึกษา ถิ่นที่อยู่อาศัย การเข้าถึงสื่อประเภทต่าง ๆ อายุ อาชีพต่างๆ เหล่านี้ต่างเป็นตัวกำหนดสำคัญที่ทำให้ผู้บริโภคทั้งหลายมีรสนิยมแตกต่างกันไปได้ทั้งสิ้น ดังนั้น ถ้าเราสนใจที่เน้นเฉพาะผลของรายได้ และผลของราคาเท่านั้น การวิเคราะห์ถดถอยก็อาจให้ค่า R^2 ค่อนข้างจะต่ำ ซึ่งเท่ากับว่ารายได้ และราคาสินค้านั้น อธิบายอุปสงค์ต่อสินค้าได้ไม่กี่เปอร์เซ็นต์ ส่วนที่เหลือจะอธิบายได้โดยรสนิยม ซึ่งเรากำหนดไว้ต่างหาก แต่ถ้าหากต้องการจะหาตัวกำหนดอุปสงค์ต่อสินค้าทั้งหมด ก็จะต้องเพิ่มตัวแปรอิสระที่สำคัญๆ ที่มีผลต่อรสนิยมของผู้บริโภคไว้ในแบบจำลองให้หมด แน่นอนว่า เมื่อเป็นเช่นนั้นแล้วค่า R^2 ก็จะสูง เพราะจะสามารถอธิบายการแปรผันในอุปสงค์ได้มากกว่า และด้วยเหตุนี้เองจึงเป็นกลยุทธอย่างหนึ่ง ในการวิเคราะห์ถดถอยสำหรับผู้ที่จะใช้สถิติ เป็นเครื่องมืออย่างง่ายๆ โดยการใส่ตัวแปรอิสระให้มากตัวที่สุดเข้าไว้ในแบจำลอง ยิ่งถ้าหากมีจำนวนตัวแปรอิสระมากขึ้นเท่าใด ค่า R^2 ก็จะเพิ่มมากขึ้นเท่านั้น

ตัวอย่างผลการวิเคราะห์ถดถอยอาจแสดงได้ดังนี้ [5]

$$SC = 4.8034 + 0.3900\,SM + 0.3765\,SF$$
$$(5.9711) \qquad (4.5672)$$

$R^2 = 0.5602$; $F(2, 397) = 90.7865$

ทั้งนี้ตัวเลขในวงเล็บคือค่า t-statistics ซึ่งทดสอบทางสถิติแล้ว จะพบว่าค่าประมาณของ coefficient ทั้งสองค่านั้นมีนัยสำคัญทางสถิติ ส่วนค่าของ F-statistics ณ ระดับ degree of freedom ที่ 2 และ 397 เมื่อทดสอบแล้ว ก็มีนัยสำคัญทางสถิติเช่นกัน และค่า $R^2 = 0.5602$ แสดงว่า การศึกษาของบิดา และมารดานั้นมีอิทธิพลในการอธิบายระดับการศึกษาของบุตรคนโตได้ถึงประมาณร้อยละ 56 (ซึ่งเท่ากับว่าบุตรคนโตจะมีการศึกษา ณ ระดับใดนั้น 56 เปอร์เซ็นต์ขึ้นอยู่กับการศึกษาของบิดา และของมารดา ส่วนอีกประมาณ 44 เปอร์เซ็นต์นั้นจะขึ้นอยู่กับปัจจัยอื่นๆ

๓.<u>ปัจจัยพื้นฐานในการวิเคราะห์ถดถอย</u>

ถึงแม้ว่าการวิเคราะห์ถดถอยในปัจจุบัน จะไม่ใช่เรื่องยาก เพราะเพียงแต่ผู้ใช้จะเข้าใจหลักการ และขอบเขตของการวิเคราะห์ถดถอย ก็สามารถใช้โปรแกรมสำเร็จรูป สำหรับการคำนวณโดยคอมพิวเตอร์เข้าช่วย ก็จะสามารถ

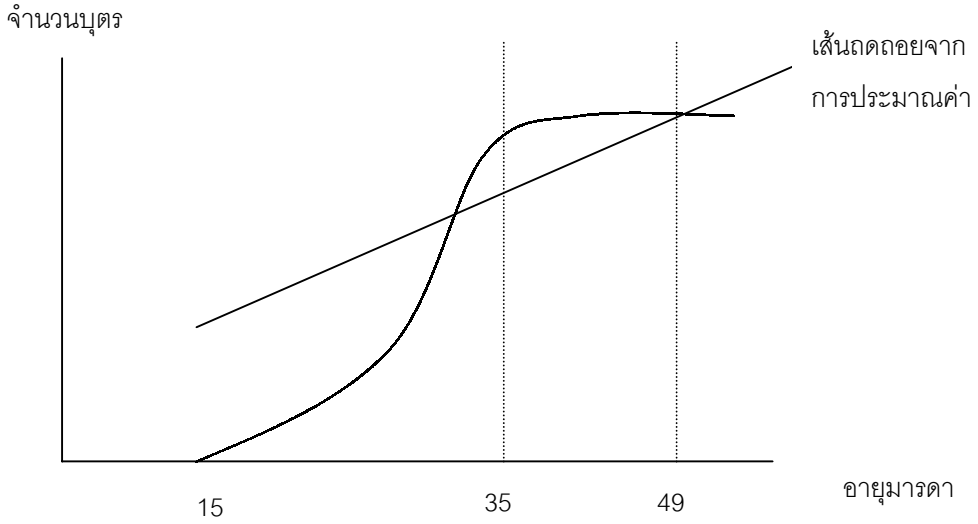
---

[5] Thienchay kiranandana, <u>ibid</u>

ได้ผลการวิเคราะห์อย่างง่ายดายก็ตาม แต่การวิเคราะห์ถดถอยนั้น ก็ยังมีข้อจำกัดอยู่บางประการ ซึ่งผู้ใช้จะต้องพึงระมัดระวัง เพราะมิฉะนั้นแล้วอาจเกิดข้อคลาดเคลื่อน หรืออคติ ขึ้นได้ในผลของการวิเคราะห์ และจะทำให้การตีความหมายตลอดจนการประยุกต์อธิบายผลการวิเคราะห์นั้น ให้ความหมายที่ผิด หรือคลาดเคลื่อนไปทั้งหมดได้ ในส่วนนี้จะได้พยายามสรุปประเด็นสำคัญๆ ที่เป็นข้อจำกัดในการวิเคราะห์ถดถอย และก่อให้เกิดปัญหาได้เสมอๆ โดยเฉพาะเมื่อผู้ใช้การวิเคราะห์แบบนี้ ไม่ได้ให้ความระมัดระวังเท่าที่ควรไว้ล่วงหน้า

3.1 <u>ปัญหาจากการวิเคราะห์ที่มิใช่เชิงเส้น (non – linearity)</u>

ปัญหานี้มักจะเกิดขึ้นบ่อยที่สุด ทั้งนี้อาจจะเป็นเพราะผู้ใช้การวิเคราะห์ถดถอยมองข้ามข้อสมมติสำคัญสำหรับการวิเคราะห์ถดถอยไปเสีย ข้อสมมุติสำคัญดังกล่าวนั้นก็คือ การวิเคราะห์ถดถอยนั้น สมมุติอยู่ว่าเป็นการผูกแสดงความสัมพันธ์เชิงเส้น (linear relationship) ระหว่างตัวแปรอิสระ กับตัวแปรตามเท่านั้น ดังนั้นในการประมาณค่า coefficient นั้น การวิเคราะห์ถดถอยจึงให้ค่าประมาณเชิงเส้นตรง สำหรับความสัมพันธ์ระหว่างตัวแปรอิสระ กับตัวแปรตามโดยตลาด อย่างไรก็ดีโดยข้อเท็จจริงแล้ว มักจะปรากฏเสมอๆ ว่าความสัมพันธ์ระหว่างตัวแปรอิสระตัวใดตัวหนึ่งกับตัวแปรตามนั้น อาจจะไม่ได้เป็นความสัมพันธ์เชิงเส้น (ตรง) เสมอไป การวิเคราะห์ถดถอยก็จะประมาณค่า coefficient ในรูปของเส้น (ตรง) อยู่ดี จึงมีผลให้คลาดเคลื่อนไปจากที่ควรจะพึงเป็น เช่นในการพิจารณาถึงความสัมพันธ์ระหว่างจำนวยบุตร กับอายุมารดา โดยทั่วไปแล้วในแง่ของการเจริญพันธ์สะสม (cumulative fertility) เราจะพบว่าจำนวนบุตร จะมีความสัมพันธ์ในทางบวก กับอายุของมารดา กล่าวคือ ยิ่งมารดาอายุมากขึ้น จะมีแนวโน้มที่จะมีบุตรมากคนขึ้นเสมอ แต่ทว่าความสัมพันธ์ระหว่างจำนวนบุตร กับอายุของมารดานี้ มิได้เป็นความสัมพันธ์เชิงเส้ง (ตรง) ธรรมดา เพราะเหตุที่สภาวะทางธรรมชาติของร่างกายสตรีนั้น ความสามารถในการมีบุตร (fecundity) จะเริ่มประมาณอายุ 15 ปี และจะสิ้นสุดเมื่ออายุประมาณ 49 ปี ประกอบกับสภาวะทางสังคม และอื่นๆ ที่สตรีมักจะสมรถในเกณฑ์อายุ 20 ปีเศษโดยเฉลี่ยและจะมีบุตร จนกระทั่งล่วงเข้าถึงอายุ 35 ปี โดยประมาณไปแล้ว ความสามารถในการมีบุตรจะชะลอลง ขณะเดียวกันก็มักจะพบว่าสตรีอายุเกินกว่า 35 ปี โดยประมาณมักจะไม่ค่อยมีบุตรอีก ความสัมพันธ์ระหว่างจำนวนบุตร กับอายุมารดา จึงมีลักษณะที่เป็นความสัมพันธ์เชิงเส้นโค้ง (non-linear) ดังแสดงในแผนภาพได้ว่าปกติ



ในกรณีเช่นนี้ เมื่อใช้การวิเคราะห์ถดถอยธรรมดา เส้นถดถอยที่ได้จากการประมาณค่าจะมีสภาพเป็นเส้นตรงดังแสดงไว้ในแผนภาพ ซึ่งคลาดเคลื่อนไปจากสภาพที่เป็นจริง โดยเฉพาะในช่วงอายุของมารดาก่อนเกณฑ์อายุ 35 ปี กล่าวคือ จะประมาณสูงกว่าที่เป็นจริงไปมาก (Over – estimated)

คำถามสำหรับผู้วิจัย จะมีอยู่ว่าจะทราบได้อย่างไรว่า ความสัมพันธ์ระหว่างตัวแปรอิสระ กับตัวแปรตามคู่นั้นๆ เป็นความสัมพันธ์เชิงเส้น (ตรง) หรือไม่ ก่อนที่จะทำการวิเคราะห์ถดถอย ทั้งนี้เพื่อที่จะได้หลีกเลี่ยงการคลาดเคลื่อนที่จะเกิดในการประมาณค่านั้นเสียได้ สำหรับคำถามในทำนองนี้นั้น คงจะหาคำตอบได้ยากนอกเสียจากว่าผู้วิจัยจะต้องพิจารณาเอาเองจากข้อสังเกตเบื้องต้น เป็นต้นว่า [6] (ก) พิจารณาจากวรรณกรรมในอดีต โดยเฉพาะจากผลการวิจัยของผู้อื่น ที่ได้ทำไว้แล้วในทำนองคล้ายๆ กันในอดีต เพราะว่าผู้ที่ได้ทำวิจัยไว้ในอดีต อาจพบ และแก้ปัญหานั้นไว้บ้างแล้ว หรือ (ข) พิจารณาหรือประเมินจากประสบการณ์ที่ตนได้เคยพบ หรือเคยทำวิจัยไว้บ้างแล้วในอดีต ซึ่งอาจพบความสัมพันธ์ที่มิได้เป็นเชิงเส้น (ตรง) ไว้บ้างแล้ว หรือ (ค) พิจารณาจากทฤษฎี หรือแนวความคิด หรือหลักอื่นๆ ในศาสตร์สาขาที่ตนจะทำวิจัย หรือในศาสตร์อื่นที่เกี่ยวข้อง ซึ่งเป็นเรื่องยาก และต้องใช้ความละเอียดรอบคอบในการพิจารณาตีความทฤษฎี และแนวความคิดเหล่านั้นมากพอควร หรือ (ง) พิจารณาโดยทำการทดสอบอย่างง่ายๆ เช่น สร้างกร๊าฟแสดงความสัมพันธ์ของข้อมูลดิบ ระหว่าง

---

[6] ในประเด็นปัญหาข้างต้นนี้ มีข้อเสนอแนะอยู่ว่าทางหนึ่งที่ทำได้ และให้ผลดีมากพอควรก็คือ การศึกษาถึงผลการวิจัย และข้อเขียนบทความในส่วนที่เกี่ยวข้องกับเรื่องที่จะทำวิจัยนี้ล่วงหน้า โดยรอบคอบ ปกติแล้วผลการวิจัย และข้อเขียนบทความต่างๆ ที่ทำกันมาในอดีตนั้น จะเสนอรูปแบบและทิศทางของความสัมพันธ์ระหว่างตัวแปรต่างๆ ไว้อย่างกระจัดกระจาย โดยที่จะหารูปแบบ และทิศทางของความสัมพันธ์ระหว่างตัวแปรต่างๆ ให้เหมือนกับชุดที่ผู้วิจัยกำหนดขึ้นในแบบจำลองของตนนั้นได้ยาก เพราะแบบจำลองแต่ละแบบจะมีความเหมาะสมกับข้อมูลแต่ละชุด แต่ละแหล่งและแตกต่างกันไปตามแต่กรณีที่ศึกษากัน เช่น อาจพบความสัมพันธ์ระหว่างอายุ กับการเจริญพันธ์ในบทความหนึ่ง และสัมพันธ์ระหว่างรายได้ กับการเจริญพันธ์ในอีกข้อเขียนหนึ่ง เป็นต้น ผู้วิจัยมีหน้าที่ปะติดประต่อความรู้ที่ค้นพบจากแหล่งต่างๆ เหล่านั้นเข้าด้วยกันให้เป็นสาระเนื้อหาเอง

ตัวแปรคู่นั้นๆ โดยเฉพาะ ในปัจจุบันจะสร้างกร๊าฟได้ง่ายขึ้น โดยการใช้โปรแกรมสำเร็จรูปเข้าช่วย ข้อพิจารณาสุดท้ายนี้ จะช่วยได้มากในกรณีที่ผู้วิจัยเริ่มสงสัยว่าตัวแปรอิสระ กับตัวแปรตามคู่นั้นๆ มีความสัมพันธ์ที่มิใช่เชิงเส้น (ตรง) ธรรมดาๆ

การแก้ไข หรือหลีกเลี่ยงปัญหาดังกล่าวนี้ อาจทำได้หลายวิธีแล้วแต่กรณี ทั้งนี้เพราะทางออกในแต่ละวิธีนั้น มีวัตถุประสงค์ที่แตกต่างกันไป ในที่นี้จะขอยกตัวอย่างทางออกไว้ 3 วิธีพอเป็นสังเขป กล่าวคือ

ก. การแก้ปัญหาโดยการใช้วิธีแปลงรูปความสัมพันธ์ (transformation) ซึ่งเป็นวิธีธรรมดาที่พยายามหาทางแปลงรูปความสัมพันธ์ที่มิใช่เชิงเส้นนั้น ให้กลายมาเป็นความสัมพันธ์เชิงเส้นเสียก่อน แล้วใช้การวิเคราะห์ถดถอยธรรมดาๆ ได้ต่อไป ปัญหาต่อเนื่องนั้นจะอยู่ที่ว่าผู้วิจัยจะต้องทราบ และเข้าใจสภาพความสัมพันธ์ที่แท้จริงก่อน การแปลงรูปนั้นเสียก่อนว่าอยู่ในสภาพใดกันแน่ ยกตัวอย่างเช่น ถ้าหากผู้วิจัยทราบว่าความสัมพันธ์ระหว่างจำนวนบุตร (F) กับอายุมารดา (A) นั้นอยู่ในรูปฟังก์ชั่นกำลังสองคือ

$$F = f(A^2)$$

การแปลงรูปความสัมพันธ์ ให้กลายมาอยู่ในรูปของฟังก์ชั่นกำลังหนึ่งธรรมดา ก็อาจทำได้โดยใช้ log ซึ่งเรียกกันว่า log transformation กล่าวคือ take log ทั้งข้างของฟังก์ชั่น ซึ่งอาจเขียนเป็นสมการได้ว่า

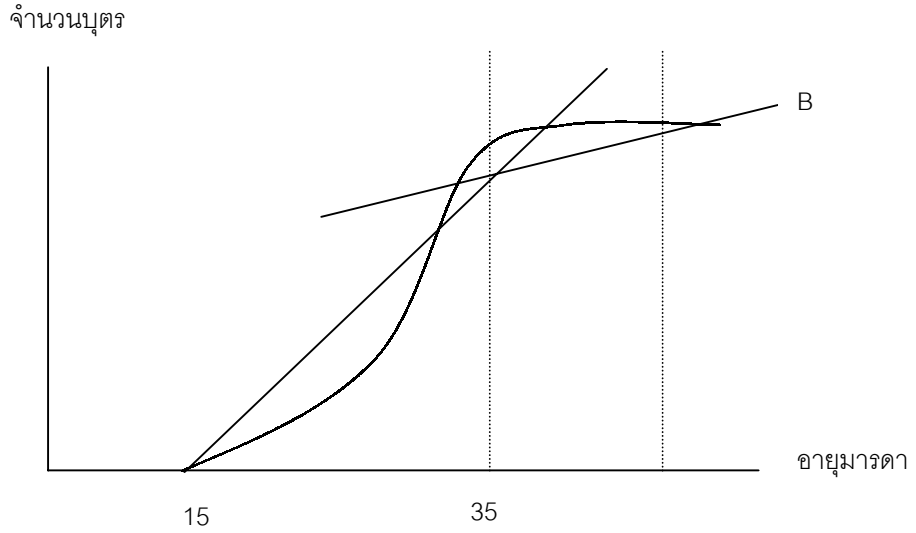
$$\log F = 2 \log A$$

log F ก็จะเป็นตัวแปรตามตัวใหม่ และ log A ก็จะเป็นตัวแปรอิสระตัวใหม่ ซึ่งความสัมพันธ์ระหว่าง log F กับ log A นั้นเป็นเชิงเส้น (ตรง) ก็จะสามารถการวิเคราะห์ถดถอยธรรมดาได้ ซึ่งในเชิงประจักษ์แล้ว ก็เพียงแต่ใช้ค่า log ของตัวแปรแทนค่าตัวแปรธรรมดาจากข้อมูลดิบเท่านั้น

ข. การแก้ปัญหาโดยการประมาณค่าสมการถดถอย 2 สมการ หรือกว่านั้นขึ้นไป กล่าวง่ายๆ ก็คือแทนที่จะประมาณค่าโดยการวิเคราะห์ถดถอยเพียง 1 สมการ เช่นในกรณีปกติก็อาจต้องประมาณค่าเป็น 2 สมการ โดยให้แต่ละสมการเหมาะสมกับความสัมพันธ์แต่ละช่วง ประเด็นสำคัญที่พึงระวังก็คือ ถ้าจะแก้ปัญหาด้วยวิธีนี้ ผู้วิจัยจะต้องทราบเคร่าๆ ล่วงหน้าก่อนว่า จุด หรือระดับที่สมการทั้ง 2 สมการ จะพบกันนั้นอยู่ที่ใด เพราะเหตุที่จุด หรือระดับนั้นเป็นจุดเริ่มต้นของสมการที่สอง ที่ต้องการประมาณค่า เช่นจากกรณีตัวอย่างที่ยกไว้นั้น ถ้าหากพอทราบว่าระดับอายุของมารดา ที่ความสัมพันธ์ระหว่างจำนวนบุตร กับอายุของมารดา เริ่มจะเปลี่ยนค่าความลาดชันอย่างมากๆ นั้น อยู่ที่ช่วงอายุประมาณ 35 ปี ก็จะสามารถประมาณค่าสมการถดถอย 2 สมการนั้นได้อย่างเหมาะสม [7] ดังแสดงในแผนภาพต่อไปนี้ เส้น A และเส้น B คือสมการถดถอย 2 สมการที่ประมาณค่าขึ้นแทนการประมาณค่าเพียงสมการเดียวตาม

---

[7] วิธีการนี้ อาจซับซ้อนมาก และควรที่ผู้วิจัย จะต้องมีพื้นความรู้ทางเศรษฐมิติพอสมควรด้วย ดูตัวอย่างการวิเคราะห์นี้ได้จาก Allen C. Kelly and Lea Malo da Silva, " The Choice of Family Size and the Compatability of Female Workforce Participation in the Low – Income Setting, " Department of Economics, Duke University, April 1977. (Mimeographed)



ค. การแก้ปัญหาด้วยการวิเคราะห์ถดถอยเป็นชุดของสมการ โดยในชุดของสมการนั้น ประกอบด้วยการวิเคราะห์ถดถอยเฉพาะกลุ่มตัวอย่างในแต่ละช่วงอายุของมารดา กล่าวคือ เป็นการใช้ age-specific fertility มากกว่าที่จะเป็น cumulative fertility ดังเช่นที่ผ่านๆ มา กรณีเช่นนี้จะให้ความหมายที่ต้องตีความแตกต่างไปจากเดิม ส่วนวิธีการในรายละเอียดจะได้กล่าวถึงในตอนต่อไป

อย่างไรก็ตาม นอกเหนือจากวิธีการทั้ง 3 วิธีที่ยกตัวอย่างมาข้างต้นนี้แล้ว ก็ยังมีกลยุทธ์อย่างอื่นอยู่บ้าง เช่น เปลี่ยนวิธีการวิเคราะห์เป็นอย่างอื่น เป็นต้นว่าเปลี่ยนการวิเคราะห์ถดถอยพหุ เป็นการวิเคราะห์จำแนกพหุ (multiple classification analysis)

3.2 <u>ปัญหาจากการที่ตัวแปรอิสระบางคู่ มีความสัมพันธ์ในเชิงสถิติต่อกัน (multicollinearity)</u>

ในการประยุกต์การวิเคราะห์ถดถอยพหุ กับงานวิจัยทางเศรษฐศาสตร์นั้น มักจะเกิดปัญหาเช่นนี้โดยเฉพาะในกรณีที่ผู้วิจัยกำหนดแบบจำลองให้ประกอบด้วยตัวแปรอิสระที่เป็นตัวกำหนดไว้หลายๆ ตัว ตัวแปรอิสระต่างๆ เหล่านั้น บางคู่อาจมีความสัมพันธ์ในเชิงสถิติกันอย่างใกล้ชิดมาก กล่าวคือ เมื่อพิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation coefficient) ระหว่างตัวแปรอิสระแต่ละคู่แล้ว จะมีค่าเป็น $\pm$ 1.0 หรือมีค่าที่ใกล้เคียงกับ $\pm$ 1.0 มากๆ

ปัญหาเช่นนี้สืบเนื่องมาจากข้อจำกัดของการวิเคราะห์ถดถอยเอง จากการที่ตัวแปรอิสระอย่างน้อย 2 ตัว ที่มีความสัมพันธ์กันใกล้ชิดมากขนาดที่สัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระคู่นั้นๆ มีค่าเป็น $\pm$1.0 แล้ว ในเชิงของการวิเคราะห์จะเกิดเป็น singular matrix และไม่สามารถจะ invert matrix นั้นๆ ได้ ผลการวิเคราะห์จะสิ้นสุดลงหรือไม่อาจคำนวณต่อไปถึงขั้นสุดท้ายได้

อย่างไรก็ดี ในทางปฏิบัติแล้ว กรณีที่เป็น extreme ขนาดที่ว่า ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระ 2 ตัว จะมีค่าเป็น ±1.0 นั้น จะเกิดขึ้นได้ยากมาก แต่เป็นไปได้เสมอๆ ที่ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร 2 ตัว จะมีค่าใกล้เคียงกับ ± 1.0 อย่างมากๆ เช่น 0.94 หรือ –0.96 เป็นต้น การตีความหมายของปัญหาเช่นว่านี้อย่างง่ายๆ ก็คือ ตัวแปรอิสระคู่นั้นๆ (ที่มีค่าสัมประสิทธิ์สหสัมพันธ์สูงมากๆ) ให้ความหมายเท่าๆ กับว่า ตัวแปรอิสระคู่นั้นๆ แสดงอิทธิพล หรือผลกระทบต่อตัวแปรตามได้เกือบจะเท่าๆ กัน (คือดีเท่าๆ กัน) หรืออาจใช้แทนกันได้ จึงไม่มีความจำเป็นจะต้องใช้ตัวแปรอิสระ 2 ตัวที่ให้ความหมาเกือบจะเหมือนๆ กันนั้นมาอธิบายการแปรผันในตัวแปรตามอยู่ด้วยพร้อมๆ กัน หรืออีกนัยหนึ่งก็คือ ในกรณีเช่นนั้น จะเลือกใช้ตัวแปรอิสระเพียงตัวใดตัวหนึ่งไปอธิบายการแปรผันในตัวแปรตาม ก็เพียงพอแล้ว ตัวแปรอิสระเหล่านี้ โดยปกติแล้วมักจะเป็นตัวแปร ซึ่งให้ความหมายแทนกันได้อยู่แล้วในเชิงทฤษฎี เช่น รายได้ กับทรัพย์สิน หรืออายุปัจจุบัน กับอายุแรกสมรส กับระยะเวลาสมรส อย่างไรก็ดี ตัวอย่างที่ยกมากล่าวถึงนี้ ก็ไม่ใช่กติกาตายตัวว่า จะต้องมีค่าสัมประสิทธิ์สหสัมพันธ์สูงมากในทุกๆ กรณีที่ศึกษาเสมอไป

ทางออกที่ดี เพื่อที่จะหลีกเลี่ยงปัญหาดังกล่าวนี้ก็คือ อาจทำได้เป็น 2 ขั้นตอน ขั้นตอนแรกผู้วิจัยก็ควรพึงระมัดระวังให้มาก ในการสร้างแบบจำลองขึ้นพิจารณา โดยเฉพาะในกรณีที่วรรณกรรม ทฤษฎี และผลงานวิจัยในอดีต ได้แสดงข้อค้นพบเกี่ยวกับความสัมพันธ์ทางสถิติ ระหว่างตัวแปรอิสระคู่ใดคู่หนึ่งไว้อย่างชัดเจนแล้ว กรณีเช่นนั้น หากหลีกเลี่ยงเสียได้ ก็ควรกระทำ และในขั้นที่สอง ก่อนที่จะทำการวิเคราะห์ถดถอย ก็ควรตรวจสอบสถิติจากค่าสัมประสิทธิ์สหสัมพันธ์ระหว่าง ตัวแปรอิสระแต่ละคู่เสียก่อน ด้วยความสะดวกจากโปรแกรมสำเร็จรูปทางคอมพิวเตอร์ในปัจจุบัน ผู้วิจัย จะสามารถสั่งให้หาค่า matrix ของสัมประสิทธิ์สหสัมพันธ์ได้โดยง่าย ถ้าหากพบว่าค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระคู่ใดมีค่าที่สูงมากๆ กล่าวคือ มากกว่า 0.9 หรือน้อยกว่า -0.9 แล้ว ผู้วิจัยก็พึงสังวรว่าปัญหา multicollinearity อาจเกิดขึ้นได้ และทางแก้ไขที่ง่ายที่สุดก็คือ เลือกตัวแปรอิสระเพียงตัวใดตัวหนึ่งในคู่ที่มีปัญหานั้นไว้ในแบบจำลองเพียงตัวเดียว [8]

3.3 <u>ปัญหา simultaneity</u>

ลักษณะของปัญหานี้ สำหรับผู้ที่ไม่คุ้นเคยกับเศรษฐมิติเลย อาจจะเข้าใจได้ยากพอสมควร แท้จริงแล้วผู้วิจัยจะเข้าใจลักษณะของปัญหา simultaneity ได้ไม่ยากนัก ถ้าหากทราบก่อนล่วงหน้าว่า ในการวิเคราะห์ถดถอยแบบง่ายๆ ที่เรียกกันว่า ordinary least square estimation หรือ OLS นั้นมีข้อสมมุติที่เป็นเงื่อนไขอยู่ว่าตัวแปรอิสระแต่ละตัวที่ใช้เป็นตัวกำหนดอธิบายการแปรผันในตัวแปรตามนั้น จะต้องเป็นตัวแปรซึ่งถูกกำหนดโดยตัวแปรอื่นๆ ภายนอกระบบ ที่เรียกว่าเป็น exogeneous variable กล่าวคือ ต้องถือได้ว่าตัวแปรอิสระแต่ละตัวนั้นถูกกำหนดค่า โดยเรียบร้อยแล้วจากภายนอก โดยไม่ได้มีความสัมพันธ์หรือขึ้นอยู่กับตัวแปรอิสระตัวหนึ่งตัว

---

[8] แน่นอนว่า ในการเสนอรายงานการวิจัยนั้น ผู้วิจัยคงจะต้องอธิบายเพิ่มเติม ถึงเหตุ และผลถึงการที่ต้องดัดแปลงแบบจำลองทางทฤษฎี มาเป็นแบบจำลองเชิงประจักษ์ โดยเฉพาะการที่จะดัดแปลงชุดของตัวแปรอิสระ และการวิเคราะห์เบื้องต้นที่แสดงให้เห็นถึงความจำเป็นที่จะต้องพิจารณาเลือกตัวแปรอิสระคู่นั้นๆ เก็บไว้เพียงตัวแปรเดียว

ใดภายในระบบ หรือแบบจำลองที่สร้างขึ้นนั้น ลักษณะของปัญหา Simultaneity ที่เกิดขึ้นนั้นจึงเกิดขึ้น จากการที่ตัวแปรอิสระตัวหนึ่งตัวใดในระบบ หรือในแบบจำลองที่สร้างขึ้นไม่ได้เป็น exogeneous แต่กลับเป็น endogeneous variable กล่าวคือ กลับมีความสัมพันธ์ในเชิงที่ขึ้นอยู่กับตัวแปรอิสระตัวอื่นๆ ตัวหนึ่งตัวใด หรือหลายตัวที่มีอยู่ในระบบ หรือแบบจำลองนั้นด้วย ทั้งนี้ไม่ว่าความสัมพันธ์ในเชิงนั้นจะมีต่อตัวแปรอิสระอื่นๆ ในระบบเพียงอย่างเดียว หรือจะสัมพันธ์กับทั้งตัวแปรอิสระอื่นๆ ในระบบ และตัวแปรอื่นๆ อีกที่อยู่นอกระบบ ซึ่งอาจเรียกได้ว่าเป็นความสัมพันธ์ต่อเนื่องเป็นลูกโซ่ เช่นในรูปของฟังก์ชั่น

$$X = f(Y_1, Y_2, Y_3)$$

$$Y_1 = f(Y_2, Y_4, Y_5)$$

กล่าวคือ X ขึ้นอยู่กับ Y_1, Y_2 และ Y_3 ขณะเดียวกัน Y_1 มิได้เป็น exogeneous แต่ยังคงขึ้นอยู่กับ Y_2, Y_4 และ Y_5 ซึ่ง Y_2 เป็นตัวแปรอิสระอีกตัวหนึ่งภายในระบบเดิม ส่วน Y_4 และ Y_5 เป็นตัวแปรอิสระใหม่ที่มาจากนอกระบบเดิม

ตัวอย่างที่เห็นได้ชัดที่สุดก็คือ การวิเคราะห์รายได้ประชาชาติในเศรษฐศาสตร์มหภาคอย่างง่าย กล่าวคือ รายได้ประชาชาติ (Y) ขึ้นอยู่กับรายจ่ายบริโภค (C) รายจ่ายลงทุน (I) รายจ่ายของรัฐบาล (G) มูลค่าสินค้าเข้า (M) และมูลค่าสินค้าออก (X) ทั้งนี้ในช่วงเวลาเดียวกัน หรือ

$$Y_t = f(C_t, I_t, G_t, X_t, M_t)$$

ขณะเดียวกัน รายจ่ายบริโภค (C_t) จะขึ้นอยู่กับรายได้ในช่วงเวลาก่อน ($Y_t - 1$) และรายจ่ายลงทุนโดยนิยามแล้วจะเท่ากับเงินออม (S_t) หรือ

$$C_t = f(Y_t - 1)$$

$$I_t = S_t$$

หรือในอีกตัวอย่างหนึ่ง เช่นกรณีที่วิเคราะห์ถึงการเจริญพันธ์ในแง่อุปสงค์ต่อบุตรนั้น อาจพิจารณาสร้างแบบจำลองให้การเจริญพันธ์ หรือจำนวนบุตร (C) ขึ้นอยู่กับรายได้ของครอบครัว (Y) เพื่อหาผลของรายได้ (income effect) และค่าใช้จ่ายของบุตรแต่ละคน (E) เพื่อหาผลของราคา (price effect) และอาจขึ้นอยู่กับปัจจัยอื่นๆ อีก (Z_i) ขณะเดียวกันค่าใช้จ่ายของบุตรแต่ละคน (E) ที่บิดามารดาจะจ่ายให้ได้นั้น ยังจะกลับมาขึ้นอยู่กับรายได้ของครอบครัว (Y) จำนวนบุตร (C) และปัจจัยอื่นๆ โดยเฉพาะรสนิยม (Z_j) [9]

---

[9] ดูกรณีตัวอย่าง และคำอธิบายใน เทียนฉาย กีระนันทน์ <u>ความสัมพันธ์ทางเศรษฐกิจระหว่าง การเจริญพันธ์ รายได้ และค่าใช้จ่ายเกี่ยวกับบุตร</u> , รายงานการวิจัย (กรุงเทพ : จุฬาลงกรณ์มหาวิทยาลัย , 2524).

ทั้งนี้อาจเขียนในรูปฟังก์ชั่น ก็คือ

$$C = f(Y, E, Z_t)$$

$$E = f(Y, C, Z_j)$$

ในกรณีตัวอย่างทั้ง 2 กรณีที่ยกมานี้ จะเห็นได้ชัดว่า รูปแบบของความสัมพันธ์ระหว่างตัวแปรอิสระ กับตัวแปรตามนั้น จะอยู่ในรูปแบบที่เป็นความสัมพันธ์เชิงซ้อน หรือความสัมพันธ์สองทาง กรณีเช่นนี้ในทางเศรษฐมิติแล้ว ถ้าจะวิเคราะห์โดยใช้การวิเคราะห์ถดถอยอย่างธรรมดา หรือ OLS จะเกิดอคติ (bias) ขึ้นได้ จึงควรจะมีการปรับแก้เสียก่อน

วิธีการปรับแก้เพื่อลดอคติจากการวิเคราะห์ถดถอยในทำนองนี้ ในทางปฏิบัติแล้วอาจพิจารณาได้ 2 วิธี ซึ่งให้ความหมายต่างกัน สำหรับวิธีแรกก็คือ ต้องใช้การวิเคราะห์แบบ Simultaneous equation system ทั้งนี้เพราะรูปแบบของแบบจำลองที่สร้างขึ้นนั้น มีความสัมพันธ์เชิงซ้อนอยู่แล้ว จำเป็นต้องหาทางวิเคราะห์ให้ได้ค่าสัมประสิทธิ์ หรือ coefficient ที่ถูกต้อง การวิเคราะห์นั้นอาจเป็นวิธีที่เรียกกันว่า two-stage least square หรือ three-stage least square (TTLS) เนื้อหา และเหตุผลของการวิเคราะห์แบบ TTLS นี้ค่อนข้างจะสลับซับซ้อน จึงจะขอยกเว้นไม่กล่าวไว้ในที่นี้ [10] แต่ผู้วิจัยอาจใช้โปรแกรมสำเร็จรูปสำหรับคอมพิวเตอร์ที่มีอยู่แล้ว เช่น SAS โปรแกรม ในการวิเคราะห์เพื่อให้ได้ค่าประมาณของ coefficient จาก TTLS ได้โดยตรง และง่ายมาก การตีความหมายของผลของค่าประมาณที่ได้ ก็อาจตีความหมายได้เช่นเดียวกับการใช้ OLS ธรรมดา อย่างไรก็ตาม เงื่อนไข (condition) ทางเศรษฐมิติ ที่ผู้วิจัยจะต้องแน่ใจว่าเป็นไปได้ในการใช้ TTLS ก็คือ เงื่อนไขเกี่ยวกับ rank และเงื่อนไขเกี่ยวกับ identification ทั้งนี้แบบจำลองในกรณีนี้ จะใช้การวิเคราะห์แบบ TTLS ได้ก็ต่อเมื่อผ่านเงื่อนไขของ rank condition และสมการทั้งหมดจะต้อง just-identified หรือ over-identified เท่านั้น แต่จะเป็น under-identified ไม่ได้ [11]

อีกวิธีหนึ่งเป็นการถอดสมการต่างๆ ที่แสดงความสัมพันธ์เชิงซ้อนที่มีอยู่ทั้งหมดในระบบ หรือในแบบจำลองเสียก่อน วิธีการนี้เรียกว่า reduced-form equation เช่นจากตัวอย่างข้างต้น ถ้าหากเขียนในรูปของสมการในระบบก็คือ

$$C = a + bY + cE + dZ_i$$

และ $$E = e + fY + gC + hZ_j$$

---

[10] อาจดูคำอธิบายได้จาก ดร.ชัยวุฒิ ชัยพันธ์ “การประมาณค่าพารามิเตอร์ เมื่อมีขนาดตัวอย่างน้อย : ลักษณะปัญหา วิธีแก้ และคอมพิวเตอร์โปรแกรม” เอกสารวิชาการฉบับที่ 2401 หน่วยเศรษฐศาสตร์วิจัย คณะเศรษฐศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย มิถุนายน 2524.

[11] ดูคำอธิบายใน Thienchay Kiranandana, Notes on The Economics Impacts of Child Mortality on Parents’ Fertility Behavior Paper No. 31 (Bangkok : Institute of Population Studies, 1979) , p.17.

การทำ reduced-form equation ก็คือการนำเอาสมการที่สองแทนค่าในสมการที่หนึ่ง ผลก็คือ

$$C = a + bY + c(e + fY + gC + hZ_j) + dZ_I$$

หรือเขียนได้ว่า

$$C = \left(\frac{a + ce}{1 - cg}\right) + \left(\frac{b + cf}{1 - cg}\right)Y + \left(\frac{ch}{1 - cg}\right)z_i + \left(\frac{d}{1 - cg}\right)z_j$$

เมื่อ reduce สมการในแบบจำลองลงเหลือเพียง 1 สมการแล้ว ก็อาจใช้การวิเคราะห์ถดถอยอย่างง่าย คือ OLS ได้อย่างปกติต่อไป แต่ทั้งนี้มีข้อสังเกตว่า เมื่อได้ reduce ลงแล้ว คำอธิบาย และความหมายจะเปลี่ยนไปตามสมควรทีเดียว กล่าวคือ จากสมการแสดงว่า C ขึ้นอยู่กับ Y , Z_i, Z_j เท่านั้น ดังนั้น ค่าประมาณของ coefficient ที่คำนวณได้จึงเป็นค่าสุทธิของผลทั้งหมด (โดยรวมเอาผลของ E ซึ่งมีในทางอ้อมตามความหมายนี้เข้าไว้แล้วด้วย โปรดสังเกตุว่าค่าสัมประสิทธิ์ของตัวแปรอิสระที่ยังเหลือยู่ในสมการที่ reduce แล้ว ก็จะมีค่าเปลี่ยนแปลงไปด้วย เช่นค่าสัมประสิทธิ์ของ Y จะเปลี่ยนจากเดิมคือ b มาเป็น (b + cf) / (1- cg) ดังนั้น ถ้าหากต้องการจะหาค่าผลทางตรงเฉพาะที่ Y มีต่อ C (โดยไม่รวมอิทธิพลของตัวแปรอื่นที่มีอิทธิพลต่อ C โดยผ่านทาง Y ไว้ด้วย) ก็จะต้องถอดสมการหาค่า b อีกครั้งหนึ่งในภายหลังจากที่ได้ประมาณค่าสัมประสิทธิ์ถดถอยจากวิธี OLS แล้วเสียก่อน (และที่น่าสงสัยก็คือ ใน 4 สมการมีตัวไม่รู้ค่าถึง 8 ตัว จะถอดสมการได้หรือไม่)

ผู้เขียนขอให้ข้อสังเกตว่า การปรับแก้โดยวิธีแรกที่ใช้ TTLS นั้น แม้ว่าจะใช้โปรแกรมสำเร็จรูปคำนวณให้โดยไม่ยากนัก แต่นักวิจัยมักเชื่อกันว่าผลของการประมาณค่าโดยวิธี TTLS ก็อาจช่วยให้ดีขึ้นบ้างไม่มากนัก เมื่อเปรียบเทียบกับการใช้วิธี OLS ธรรมดา (ในแต่ละสมการ) นักวิจัยหลายคนจึงตัดสินใจที่จะใช้วิธี OLS ไปเลยโดยไม่สนใจว่าอคติที่เกิดในการประมาณค่าเมื่อแบบจำลองเป็น simultaneous equation system นั้น จะมีหรือไม่และเพียงใด

๓.๔ <u>ปัญหาที่เรียกว่า heteroskedasticity</u>

โดยเนื้อหาแล้ว ลักษณะของปัญหา heteroskedasticity มักจะเกิดขึ้นเนื่องจากการใช้ข้อมูล cross-section ที่ตัวแปรตามมีค่าเป็นช่วงห่างมากระหว่างค่าที่สูงสุด กับค่าที่ต่ำสุด (เป็นต้นว่าค่าสูงสุดมีค่าเป็น 5-6 เท่า หรือกว่านั้น เมื่อเทียบกับค่าต่ำสุด) ผลของปัญหาก็คือ ค่าประมาณของ coefficient อาจคลาดเคลื่อนได้ ผู้วิจัยอาจดูได้ว่าเกิด heteroskasdaticity ก็คือ ค่าความคลาดเคลื่อนมาตรฐาน (Standard error) ของค่าประมาณของ coefficient จะมีค่าสูงผิดปกติกว่าที่ควรจะเป็น (อาจจะ 2 เท่า หรือกว่านั้น)

ทางออกสำหรับปัญหาดังกล่าวนี้ อาจทำได้ 3 วิธี กล่าวคือ วิธีแรกโดยการปล่อยทิ้งไว้เฉยๆ หรือไม่สนใจ โดยเฉพาะถ้าหากว่า ในการประมาณค่า coefficient นั้น ๆ ยังให้ค่า t-statistic ที่สูงมากพอ (เช่น 2.5 หรือสูงกว่านั้นขึ้นไป) หรือมิฉะนั้น ก็ใช้วิธีที่สอง โดยการกลับไปเริ่มวิเคราะห์ใหม่ และใช้ขนาดตัวอย่างมากขึ้นกว่าเดิม การใช้ขนาดตัวอย่างมากขึ้นกว่าเดิมในกรณีเช่นนี้ จะช่วยแก้ปัญหาที่เกิดจากช่วงห่างในค่าสูงสุด และค่าต่ำสุดในตัวแปรตามได้มาก หรือวิธีที่ 3 ก็คือ การหันไปใช้วิธีการวิเคราะห์ถดถอยที่มีประสิทธิภาพมากกว่า OLS

ธรรมดาๆ กล่าวคือ วิธี weighted least squares ซึ่งจะไม่กล่าวถึงรายละเอียดในที่นี้ แต่ผู้สนใจอาจหาอ่านได้จากตำราเศรษฐมิติโดยทั่วไป

3.5 <u>ปัญหาเกี่ยวกับความเกี่ยวพันในเวลา</u>

นอกจากปัญหาต่างๆ ข้างต้นแล้ว ในกรณีที่การศึกษาวิจัยหนึ่ง ๆ ใช้ข้อมูลแบบ time series สิ่งที่มักจะพบว่าเป็นปัญหาอยู่เสมอๆ ก็คือ ตัวแปรแต่ละตัวมักจะมีความผูกพัน หรือเกี่ยวพันในเชิงเวลา เช่น ถ้าตัวแปรหนึ่งมีค่าสูงในปีนี้ มักจะมีแนวโน้มว่าจะมีค่าสูงอีกในปีต่อๆ ไป กรณีเช่นนี้เป็นปัญหาที่เรียกกันว่า autocorrelation การตรวจสอบว่าในการวิเคราะห์มีอคติเกิดขึ้นจากปัญหา autocorrelation หรือ ไม่นั้นก็อาจดูได้จากค่าสถิติที่เรียกว่า Durbin-Watson statistics ส่วนทางแก้ของปัญหานี้เมื่อเกิดขึ้นแล้ว ไม่ใช่ของง่าย และควรที่จะต้องใช้พื้นความรู้ทางเศรษฐมิติพอสมควร ดังนั้นประเด็นนี้จึงไม่กล่าวถึงอย่างละเอียดในที่นี้

4. <u>ข้อสังเกตทั่วไปก่อนดำเนินการวิเคราะห์</u>

ในท้ายของบทความนี้ จะกล่าวถึงเรื่องทั่วๆ ไป ที่เป็นข้อสังเกตสำหรับนักวิจัยที่ไม่คุ้นเคยกับเทคนิคของการวิเคราะห์ถดถอยมากนัก (และอยากจะใช้วิธีการวิเคราะห์ถดถอย) ประเด็นต่างๆ ที่จะตั้งเป็นข้อสังเกตไว้ในที่นี้เป็นประเด็นที่พบเสมอๆ เมื่อผู้ใช้ไม่ได้ระมัดระวังเท่าที่ควร ซึ่งในบางกรณีเมื่อเกิดปัญหาแล้วยังพอแก้ไขได้ แต่ในบางกรณีก็แก้ไขไม่ได้ และสายเกินไป

4.1 ประเด็นแรก ที่ขอตั้งเป็นข้อสังเกตไว้ในเบื้องต้นนี้ก็คือ การออกแบบจำลอง (model specification) โดยปกติแล้วนักวิจัยใหม่มักจะไม่ได้ใช้ความระมัดระวังเท่าที่ควร โดยเฉพาะการใช้เหตุผล และวรรณกรรมปริทัศน์เข้ามาเสริมความคิดในการสร้างแบบจำลองที่แสดงความสัมพันธ์ระหว่างตัวแปรต่างๆ ตัวแปรอิสระ และตัวแปรตามในบางกรณีมีความสัมพันธ์กันทั้งสองทาง และเป็นเงื่อนไขของเวลาด้วย เช่นจำนวนบุตรเกิดรอด กับจำนวนบุตรที่ตาย ซึ่งเป็นตัววัดค่าความสัมพันธ์ระหว่างภาวะเจริญพันธ์ กับภาวะการตาย โดยนัยหนึ่งจำนวนบุตรเกิดจะขึ้นอยู่กับจำนวนบุตรตาย โดยบิดามารดาจะคำนึงว่า ถ้าอัตราตายของทารก และเด็กสูง ตนก็จำเป็นต้องมีบุตรให้มากคนไว้ เมื่อจำนวนหนึ่งอาจต้องเสียชีวิตไปก่อนที่จะเติบโตเป็นผู้ใหญ่ และตรงกันข้าม โดยข้อเท็จจริง เมื่อมีคนเกิดมาก ถ้าหากอัตราการตายคงที่ จะมีผลให้จำนวนคนตายสูงขึ้นแน่ๆ การสร้างแบบจำลองแสดงความสัมพันธ์ของตัวแปรนี้ จึงควรต้องใช้ความรอบคอบอย่างมาก และครอบคลุมให้หมด วิธีการที่ดีคือแยกแบบจำลองที่สร้างนี้เป็น 2 แบบ แบบหนึ่งเป็นแบบจำลองทางทฤษฎี และแนวความคิด (conceptual & theoretical model) ซึ่งจะต้องเป็นตัวที่แน่นอนถูกต้อง และเหมาะสมทางทฤษฎีที่สุด แล้วจึงสร้างเป็นแบบที่สองคือ แบบจำลองเชิงประจักษ์ (empirical model) การปรับแบบจำลองทางแนวคิดมาเป็นแบบจำลองเชิงประจักษ์นี้ ก็เพื่อเหตุผล 3 ประการคือ (ก) ให้เหมาะสมกับสภาวะแห่งความเป็นจริงในกรณีที่ศึกษา (ข) ให้เหมาะสมกับข้อมูลที่มีอยู่ และเอื้ออำนวยให้ และ (ค) ให้เหมาะกับความจำเป็นอื่นๆ ที่บังคับไว้ไม่มีทางเลือก

ที่สำคัญมากประเด็นแรกก็คือ เมื่อกำหนดแบบจำลองเชิงประจักษ์แล้ว ก็จำเป็นต้องคำนึงถึงตัวแปรต่าง ๆ ในเชิงประจักษ์ เช่น ในทฤษฎีอาจพูดถึงรายได้ซึ่งเป็นตัวแปรสำคัญมาก แต่ในทางประจักษ์แล้วเก็บข้อมูล “รายได้” ที่แน่นอนถูกต้องเชื่อถือไม่ได้ เมื่อเป็นเช่นนี้ ในแบบจำลองเชิงประจักษ์ก็ต้องหาตัวแทนของ “รายได้” ตัวแทนนี้เรียกว่า “proxy” การเลือก proxy นี้จึงเป็น “ศิลป” ของผู้วิจัยโดยแท้ว่า จะสามารถเลือก proxy มาแทนตัวแปรเชิงทฤษฎีได้เหมาะสมเพียงใด อย่างไรก็ดี การเลือก proxy นั้นจะต้องแน่ใจว่ามีข้อมูลสำหรับตัวแปรที่จะใช้เป็น proxy นั้นอยู่แล้ว (หรือจะสามารถเก็บข้อมูลได้) และสิ่งที่จำเป็นที่สุดในการวิจัยเชิงประจักษ์ ผู้วิจัยจะต้องระลึกด้วยว่าเมื่อตัวแปรใดถูกเลือกมาเป็น proxy นั้น จะไม่มีทางให้ความหมายที่ถูกต้องครบถ้วนได้เท่าตัวแปรจริงๆ ที่เป็นตัวหลักในทฤษฎีที่พบเสมอๆ ก็คือตัวแปรที่เป็น proxy นั้นจะรวมความหมายที่กว้างขวาง หรือมี implication ที่กว้างขวางมากกว่าความหมายของตัวแปรหลักทางทฤษฎี เช่นในการใช้ระดับการศึกษาเป็น proxy ของรายได้ ผู้วิจัยจะต้องเข้าใจ และยอมรับด้วยว่า ระดับการศึกษานั้นจะให้ความหมายส่วนหนึ่งแทนรายได้ได้พอสมควร (อาจไม่เต็ม 100 เปอร์เซ็นต์) แต่ยังอาจให้ความหมายทางอ้อมถึง “รายได้” และอาจให้ความหมายอื่นๆ ที่มีอิทธิพลต่อตัวแปรตามในแบบจำลองอีกด้วย ดังนั้น การสรุปผลการวิเคราะห์ในกรณีที่ต้องใช้ proxy นั้น จึงต้องระมัดระวังมาก และจะต้องไม่แสดงความเชื่อมั่นที่แน่ชัดเกินไปอีกด้วย

4.3 ในการวิจัยที่ใช้ข้อมูลระดับจุลภาค (micro data) นั้น ควรตั้งไว้เป็นข้อสังเกตไว้ล่วงหน้าก่อนว่า การวิจัยนั้นเป็นการศึกษาหาข้อเท็จจริงในรายละเอียดระดับบุคคล หรือครัวเรือน ประกอบกับการใช้เทคนิค กับการถดถอยนั้น จะให้ผลทางสถิติได้ดียิ่งขึ้น ถ้าหากค่าของตัวแปรแต่ละตัว (ไม่ว่าจะเป็นตัวแปรอิสระ หรือตัวแปรตามก็ตาม) จะมีการแปรผันค่อนข้างสูง เมื่อเป็นเช่นนี้ ข้อมูลระดับจุลภาคที่ใช้ จึงควรหาทางให้มีการผันแปรสูงหรือมากที่สุดเท่าที่จะทำได้

ประเด็นสำคัญประการหนึ่งที่ผู้วิจัยมักจะไม่ทันสังเกต และกลับมีผลให้ค่าของตัวแปรแต่ละตัว มีค่าการผันแปรต่ำไปกว่าที่ควรจะเป็นก็คือ การกำหนดข้อมูลเป็นช่วงๆ ขึ้นใช้ เป็นต้นว่า อายุกำหนดเป็น 15-19 , 20-24, 25-29, …. หรือ ระดับการศึกษาที่สำเร็จกำหนดเป็น ป.1-ป.4, ป.5-ป.6, ม.1-ม.3, …. ซึ่งเมื่อตั้งขึ้นวิเคราะห์ข้อมูลด้วยเทคนิคสมการถดถอย ก็จะต้องเทียบค่าของข้อมูล (Scaling) ลงเป็นข้อมูลสืบเนื่อง (continuous) ทั้งนี้เพื่อให้ใช้เทคนิคดังกล่าวได้

สาเหตุที่กำหนดข้อมูลเป็นช่วงๆ ดังที่ยกตัวอย่างนั้น อาจเนื่องมาจากความสะดวกในการเก็บข้อมูล หรือจากความสะดวกในการลงรหัสข้อมูล หรือจากเหตุอื่นๆ ก็ได้ ที่พบเสมอๆ ก็คือ ผู้วิจัยมิได้คาดถึงปัญหาล่วงหน้าไว้ก่อน ดังนั้น เมื่อทำการสำรวจหาข้อมูล จึงออกแบบสอบถามอย่างง่ายๆ ขึ้น เช่น ออกแบบสอบถามด้วยการสัมภาษณ์ไว้ดังนี้

อายุ

❑ 15-19 ปี	❑ 30-34 ปี
❑ 20-24 ปี	❑ 35-39 ปี
❑ 25-29 ปี	❑ 40-44 ปี

ซึ่งผู้สัมภาษณ์ เพียงแต่กาเครื่องหมายลงในช่องที่กำหนด (ผู้วิจัยมักอ้างว่า เพื่อหลีกเลี่ยงปัญหาในกรณีที่ผู้สัมภาษณ์จำอายุตัวเองไม่ได้ หรือผู้อื่นไม่ได้) หรืออย่างเช่นในกรณีกำหนดขึ้นเพื่อความสะดวกในการลงรหัสข้อมูล โดยเฉพาะในกรณีที่ใช้คอมพิวเตอร์ช่วยในการวิเคราะห์ โดยอ้างว่าเพื่อประหยัดช่วงที่ใช้บัตรคอมพิวเตอร์ กล่าวคือ ถ้ากำหนดอายุไว้เป็นช่วง 5 ปี สตรีในวัยเจริญพันธ์อายุประมาณ 15-49 ปี ก็จะกำหนดได้เพียง 7 ช่วง ซึ่งเป็นเลขหลักเดียว และใช้ช่วงในบัตรคอมพิวเตอร์เพียง 1 ช่วง ตรงกันข้าม ถ้าหากใช้อายุจริงโดยไม่กำหนดเป็นช่วงอายุแล้ว จะเป็นเลขสองหลัก และต้องใช้ช่วงในบัตรคอมพิวเตอร์ถึงสองช่วง

ที่น่าเสียดายมากกว่านั้นก็คือ ในหลายกรณีข้อมูลดิบที่เก็บได้จากการสำรวจโดยสุ่มตัวอย่าง เป็นข้อมูลสืบเนื่องดีอยู่แล้ว แต่ผู้วิจัยกลับมาจัดระบบลงรหัสให้เป็นข้อมูลช่วงเสีย อย่างไรดี ถ้าหากเป็นการเก็บข้อมูลช่วงในขั้นการสำรวจเพื่อเก็บข้อมูลแล้ว คงจะแก้ไขได้ยาก เพราะการกลับไปถามข้อมูลที่แท้จริง (เป็นข้อมูลสืบเนื่อง) ใหม่จะต้องใช้เวลา และทรัพยากรอื่นๆ อีกมาก

ประเด็นสำคัญในที่นี้อยู่ที่ว่า เมื่อกำหนดข้อมูลไว้เป็นช่วง และต้องเทียบค่าของข้อมูลลงเป็นข้อมูลใหม่ ในขั้นของการวิเคราะห์ เราจะพบว่า ข้อมูลจะไม่มีการผันแปรในส่วนที่ควรจะมีการผันแปร และจะผันแปรในส่วนที่อาจจะไม่ผันแปรมากนักก็ได้ ยกตัวอย่างเช่น กลุ่มอายุของสตรี

15-19 ปี เทียบค่าเป็น 1

20-24 ปี เทียบค่าเป็น 2

25-29 ปี เทียบค่าเป็น 3

ในกรณีนี้สตรีอายุ 15 ปี กับสตรีอายุ 19 ปี น่าจะมีการผันแปรทางการเจริญพันธ์สูง เพราะอายุ 15 ปี เป็นปีแรกๆ เมื่อเริ่มภาวะที่อาจมีบุตรได้ (fecundity) ประกอบกับสภาพแวดล้อมทางสังคม เศรษฐกิจ และอื่นๆ ที่ทำให้เราเชื่อได้ว่า สตรีจะมีภาวะเจริญพันธ์สูงขึ้นอย่างรวดเร็วตามอายุในช่วงนั้น แต่จากการเทียบค่าข้อมูลข้างต้น สตรีอายุ 15 ปีเทียบค่าเท่ากับ 1 และสตรีอายุ 19 ปี ก็เทียบค่าเท่ากับ 1 จึงเท่ากับว่าไม่มีความแตกต่างทางอายุเลย ขณะเดียวกัน สตรีอายุ 19 ปี และสตรีอายุ 20 ปี อาจมีความแตกต่างในแง่ของผลที่จะมีต่อภาวะเจริญพันธ์น้อยมาก แต่จากการเทียบค่าข้อมูลสตรีอายุ 19 ปี เทียบค่าเท่ากับ 1 ส่วนสตรีอายุ 20 ปี เทียบค่าเท่ากับ 2 แสดงว่าข้อมูลมีการผันแปรถึงเท่าตัว

ข้อบกพร่องในเรื่องนี้ แม้ว่าจะไม่มีผลสำคัญถึงกับทำให้เกิดความคลาดเคลื่อนในผลการวิจัยอย่างร้ายแรงก็ตาม แต่ก็เป็นเรื่องที่หลีกเลี่ยงได้ ถ้าหากมีการเตรียมการล่วงหน้า และเป็นผลให้ผลทางสถิติของการวิเคราะห์นั้นละเอียดขึ้น และดีขึ้น

4.4 ในบางกรณีตัวแปรต่างๆ ในแบบจำลอง จะสื่อความหมายที่มีความแตกต่างกันในช่วงเวลาอยู่ด้วย บ่อยครั้งที่ผู้วิจัยเสนอรูปแบบความสัมพันธ์ระหว่างตัวแปรตาม กับตัวแปรอิสระ โดยไม่ได้คำนึงถึงเหตุผลในด้านความแตกต่างในช่วงเวลา ที่พบเสมอๆ เช่นรูปแบบความสัมพันธ์ระหว่างการเจริญพันธ์ กับการตาย (ตามที่ได้กล่าวไว้บ้างแล้วในตอนก่อน) กรณีเช่นนี้ อาจเป็นไปได้ว่า ในครอบครัวหนึ่งมีบุตรจำนวนหนึ่งตายไปแล้ว (จะเพราะเหตุใดก็ตาม) ครอบครัวนั้น จึงตัดสินใจมีบุตรเพิ่มขึ้นมาทดแทนจำนวนที่ตายไป หรือว่าครอบครัวนั้น จะตัดสินใจมีบุตรให้มากคนเผื่อไว้ก่อน เพราะสภาพโดยทั่วๆ ไปแล้ว อัตราการตายของทารก และเด็กสูงมาก (ในที่สุดเด็กในครอบครัวนี้ อาจไม่ตายเลย หรือตายไปไม่เท่ากับที่บิดามารดาคาดเผื่อไว้ก็ได้) หรืออีกกรณีหนึ่งที่พบเสมอๆ ก็คือ การวิจัยที่เสนอว่าความทันสมัยมีผลกระทบต่อการเจริญพันธ์ เช่นมีสมมติฐานว่าครอบครัวในเมืองน่าจะมีขนาดเล็กกว่าครอบครัวในชนบท เป็นต้น โดยลักษณะของตัวแปรบางตัวดังที่ยกตัวอย่างไว้นี้ จะมีลักษณะที่สื่อความหมายถึงลำดับเหตุการณ์ตามช่วงเวลาอยู่ด้วย ดังนั้น ในการวิจัยที่ใช้ข้อมูลแบบ Cross-section จึงจะต้องระวังให้มากกว่า ข้อมูลของครอบเรือนที่เก็บมานั้น เป็นข้อมูลที่เก็บ ณ วันที่ออกสำรวจหรือสัมภาษณ์ จากตัวอย่างแรก ถ้าจะถามเพื่อเก็บข้อมูลจำนวนเด็กเกิด และตายเท่านั้น อาจไม่เพียงพอที่จะแสดงถึงลำดับของเหตุผลที่คาดไว้ได้ หรือจากตัวอย่างที่สอง ถ้าบังเอิญเป็นครอบครัวที่เพิ่งย้ายเข้ามาอยู่ในเขตเมือง เมื่อตกเป็นตัวอย่าง ก็จะถูกนิยามว่าเป็นเขตเมือง แท้จริงแล้วครอบครัวนั้น อยู่ในเขตชนบทมาตลอดเวลา กรณีเช่นนี้ การย้ายเข้ามาอยู่ในเขตเมือง จึงเป็นการย้ายเมื่ออาจจะจบสิ้นภาวะเจริญพันธ์ไปแล้ว คือ ไม่มีบุตรอีกแล้ว และกรณีเช่นนี้ ก็คงต้องระวังในการวิจัยที่จะสรุปผลว่า “ครอบครัวในเมืองมีขนาดเล็กกว่าหรือไม่”

ทางออกที่เหมาะสมตามสมควร และควรคำนึงถึงไว้ล่วงหน้าก็คือ อาจใช้ตัวแปรคุม (controlled variable) เข้ามาช่วย ตัวแปรคุมนี้จะเข้ามาลดอิทธิพลที่อยู่นอกวงที่ต้องการออกได้หมด หรือเก็บหมด การเลือกตัวแปรคุมนี้ ก็เป็นศิลปอีกอย่างหนึ่งของผู้วิจัย ในกรณีของการวิจัยเกี่ยวกับภาวะเจริญพันธ์นั้น ตัวแปรคลุมที่เกือบจะเรียกได้ว่าหลีกเลี่ยงไม่ได้ คือตัวแปรคุมเกี่ยวกับอายุของมารดา เพราะเหตุที่มารดาอายุต่างกัน จะมีจำนวนบุตรต่างกันแน่นอน หรือกล่าวชัดเจนได้ว่า ยิ่งมารดาอายุมากยิ่งจะมีบุตรมากคน (ในแง่ของ cumulative fertility) ดังนั้น หากไม่ควบคุมด้วยอายุของมารดาแล้ว อิทธิพลของตัวแปรอิสระอื่นๆ อาจให้ความหมายที่ตีความเป็นอย่างอื่นได้ [12] นอกจากนั้น อีกวิธีหนึ่งที่อาจทำได้ ถ้าหากขนาดตัวอย่างไม่เล็กจนเกินไป ก็คือ จำกัดขอบเขตของการศึกษาวิจัยเสียให้แคบลง และชัดเจน ว่าจะรวมถึงกรณีใดบ้าง และไม่รวมกรณีใดบ้าง กรณีที่อาจมีปัญหาดังที่ยกตัวอย่างก็จะตกอยู่นอกขอบเขตของการศึกษาได้ (ถ้าหากผู้วิจัยต้องการเช่นนั้น)

---

[12] ในเรื่องอายุของมารดานี้ นักวิจัยหลายกรณีใช้ตัวแปรคุมอื่นๆ ที่มีผลคล้ายกันมาแทน เช่น ระยะเวลาของการสมรส หรืออายุแรกสมรส เรื่องนี้ไม่มีปัญหา เพราะมีความหมายแตกต่างกันไปบ้าง แต่สื่อความหมายคล้ายกัน ข้อสำคัญอยู่ที่ว่าในบางงานวิจัยใช้ตัวแปรทั้ง 3 นี้มากกว่า 1 ตัวในคราวเดียวกัน ผลที่จะเกิดปัญหาคือเรื่องของ multicollinearity ดังที่กล่าวไว้แล้ว เพราะเหตุที่ความสัมพันธ์ของ 3 ตัวแปรนี้อธิบายได้ชัดเจนมาก กล่าวคือ อายุปัจจุบัน - อายุแรกสมรส = ระยะเวลาของการสมรส

4.5 การใช้ตัวแปรชนิดที่มีค่าแปรผันน้อย และจำกัดนั้น ผู้วิจัยน่าจะต้องใช้ความระมัดระวังมากเป็นพิเศษ โดยเฉพาะอย่างยิ่งในการตีความหมายของผลการวิเคราะห์ที่ได้รับ ตัวแปรชนิดที่กำหนดค่าแปรผันไว้เพียง 2 ค่า ที่เรียกว่า dummy variable นั้น จะใช้เฉพาะในกรณีที่ตัวแปรหนึ่งๆ ให้ความหมายอย่างกว้างๆ และผู้วิจัยทราบว่าในแต่ละค่าของตัวแปรนั้น จะแสดงถึงความแตกต่างอย่างชัดเจน เช่นการกำหนดค่าตัวแปร dummy เป็น 0 ในกรณีของชนบท และให้เป็น 1 ในกรณีของเมือง ซึ่งจะแสดงว่าผู้วิจัยแน่ใจว่า เมืองกับชนบทนั้นแตกต่างกันอย่างเห็นได้ชัด แต่ไม่มีทางที่ผู้วิจัยจะกำหนดระดับความแตกต่างให้แคบกว่านั้นได้ (กล่าวง่ายๆ ก็คือ ไม่สามารถกำหนดค่าให้กับกรณีที่มีความเป็นเมือง และชนบทผสมผสานกันอยู่ได้ เป็นต้นว่าเขตชานเมือง) การแปรผันของตัวแปร dummy จึงมีลักษณะที่เป็น extreme cases เพียง 2 ค่า คือ 0 หรือ 1

ในกรณีที่ผู้วิจัยจะใช้ค่าของตัวแปรอิสระเป็นตัวแปร dummy แล้ว ปัญหาจะไม่มีมากนัก และส่วนใหญ่ก็นิยมใช้กันในกรณีที่ไม่สามารถให้รหัสตัวแปรเชิงคุณภาพให้เป็นเชิงปริมาณอย่างอื่นได้ ตัวแปรเหล่านี้มักจะเป็น nominal scale เช่น เพศ (ชาย-หญิง) ความเชื่อ (เชื่อ-ไม่เชื่อ) แต่ทั้งนี้ก็มีข้อสังเกตด้วยว่าหากผู้วิจัยจะให้ตัวแปรอิสระหลายๆ ตัวมีค่าเป็น dummy พร้อมๆ กันแล้ว แม้ว่าจะไม่มีปัญหาในการใช้การวิเคราะห์ถดถอย แต่การตีความหมายจะยุ่งยากมาก เช่นในกรณีหนึ่ง ผู้วิจัยกำหนดการศึกษาไว้เป็นตัวแปรอิสระถึง 5 ตัวด้วยกัน โดยที่แต่ละตัวมีค่าเป็น dummy กล่าวคือ

$$Q = f(ED_1, ED_2, ED_3, ED_4, ED_5, OT)$$

เมื่อ	ED_1	= 1	if educated
		= 2	if no education
	ED_2	= 1	if primary education
		= 0	otherwise
	ED_3	= 1	if secondary education
		= 0	otherwise
	ED_4	= 1	if high school education
		= 0	otherwise
	ED_5	= 1	if college education
		= 0	otherwise

ในขณะที่ OT เป็นตัวแปรอิสระอื่นๆ อีก 5-6 ตัวแปร ค่าประมาณของ coefficient ของ ED_1 ED_5 ที่ประมาณค่าได้นั้น จะมีความหมายอย่างไร

ความรุนแรงของปัญหาในการใช้การวิเคราะห์ถดถอยนั้น จะอยู่ในกรณีที่ผู้วิจัยกำหนดค่าตัวแปรตามไว้เป็น dummy กรณีเช่นนี้ในเชิงเศรษฐศาสตร์แล้ว จะเกิดอคติอย่างแรงพอสมควร มากพอที่จะกระทบกระเทือนค่า R^2 และค่าประมาณของ coefficient ทั้งหมดได้ เหตุผลของอคติที่เกิดนั้น ก็เพียงแต่เพราะเหตุว่าตัวแปรตามมีการแปรผันในค่าของมันเองน้อยมากเกินไป (คือเพียงค่า 0 หรือ 1 เท่านั้น) ยิ่งกว่านั้น ข้อมูลที่เป็น

nominal scale โดยปกติแล้วจะไม่ให้ค่าเฉลี่ยที่มีความหมายใดๆ เลย ซึ่งจะทำให้ยากในการตีความหมายมากยิ่งขึ้นอีก (เช่นในกรณีที่ให้ ชนบท = 0 เมือง =1 และหาค่าเฉลี่ยของจำนวนตัวอย่างทั้งหมดว่าเป็น 0.67 ก็จะไม่อาจให้ความหมายที่เหมาะสมของค่าเฉลี่ยนั้นได้) การแก้ปัญหาในกรณีที่ใช้ตัวแปรตามมีค่าเป็น dummy นี้ ในการวิเคราะห์ถดถอยแล้ว จะต้องใช้เทคนิคการวิเคราะห์ขั้นสูงขึ้นไปกว่า OLS ธรรมดา ซึ่งได้แก่ Logit-Probit Analysis และต้องใช้ความรู้ทางเศรษฐมิติขั้นสูงพอสมควร

กรณีที่บรรเทากว่าตัวแปร dummy ก็คือ กรณีที่ตัวแปรมีค่าเป็น Scale ได้แก่ กรณีที่ข้อมูลมีค่ามากกว่า 2 ค่า แต่ก็ไม่ต่อเนื่อง เพียงแต่พอที่จะเปรียบเทียบความแตกต่างว่ามากกว่ากัน หรือน้อยกว่ากันได้ (แต่จะไม่สามารถบอกได้ว่า ต่างกันเพียงใด) เช่นข้อมูลที่เป็น ordinary scale เป็นต้นว่า ความพอใจ อาจจำแนกออกเป็น

พอใจมากที่สุด = 4
พอใจมาก = 3
พอใจปานกลาง = 2
ไม่พอใจ = 1
ไม่พอใจมากที่สุด = 0

ตัวแปรที่มีค่าของข้อมูลเชิงคุณภาพในลักษณะเช่นนี้ อาจแปรสภาพเป็นเชิงปริมาณได้ โดยการกำหนดค่าเป็นตัวเลขให้กับความหมาย แต่ก็มีเงื่อนไขว่าควรต้องกำหนดค่าโดยเรียงลำดับจากน้อยไปหามาก หรือจากมากไปหาน้อย และตัวเลขที่กำหนดไว้นี้ จะเป็นสื่อแสดงถึงความแตกต่างในความหมายเชิงคุณภาพของตัวแปรในขณะเดียวกันได้ การทำเป็น scaling variable ในลักษณะเช่นนี้จะช่วยมากในหลายๆ กรณีที่ผู้วิจัยมีข้อมูลเชิงคุณภาพ แต่จะต้องระวังให้มาก ในการตีความหมายผลการวิเคราะห์เช่นกัน (โปรดสังเกตด้วยว่าในหลายๆ กรณีสำหรับเชิงคุณภาพนั้น จะไม่อาจให้ scale เป็นตัวเลขที่สื่อความหมายถึง “ลำดับ” ได้ เช่น เหตุผลที่กล่าวถึงการที่รายได้แท้จริงลดลง ผู้ตอบอาจให้เหตุผลอย่างไรก็ได้ แม้แต่ว่าผู้วิจัยจะออกแบบสอบถาม เป็นคำถามปิดให้ระบุเหตุผลได้เพียง 6-7 คำตอบก็ตาม กรณีเช่นนั้น ก็ไม่อาจให้ค่าตัวเลขในเหตุผล แต่ละอย่างที่จะสื่อความหมายถึง “ลำดับ” โดยเปรียบเทียบในระหว่างเหตุผลแต่ละข้อเหล่านั้นได้เลย)

4.6 การวิเคราะห์รวมข้อมูล 2 ชุดเข้าด้วยกัน (Pooled cross-sectional data) ประเด็นนี้เป็นเทคนิคสำคัญของนักวิจัยที่ใช้การวิเคราะห์ถดถอย ซึ่งมีคำอธิบายในเชิงเศรษฐมิติอยู่มาก และค่อนข้างซับซ้อน จะไม่ขอกล่าวไว้ในที่นี้ แต่สำหรับประโยชน์ในการใช้เทคนิคนี้ ก็คือ ในกรณีที่ผู้วิจัยได้สร้างแบบจำลองไว้ และต้องการทดสอบแบบจำลองนั้นด้วยข้อมูลที่เป็น cross-section 2 ชุด ที่มีความแตกต่างกัน เป็นต้นว่า เป็นตัวอย่างที่มาจากคนละกรอบ (frame) แม้แต่ว่าขนาดตัวอย่างที่แตกต่างกันบ้าง เช่น มีข้อมูลของตัวอย่าง 2 ชุด ชุดหนึ่งเป็นของเขตเมือง อีกชุดหนึ่งเป็นของเขตชนบท เมื่อต้องการเปรียบเทียบผลการวิเคราะห์ของทั้ง 2 ชุด เช่น ต้องการเปรียบเทียบว่าคนในเมือง กับคนในชนบทมีความต้องการต่อบุตรแตกต่างกันหรือไม่ เทคนิคของการวิเคราะห์ถดถอยจะเอื้อเป็นอย่างมาก ในกรณีนี้โดยอาจทำได้ 2 วิธี

วิธีแรกนั้น ผู้วิจัยอาจทำการวิเคราะห์แบบจำลอง โดยทำ 2 ครั้ง แต่ละครั้งจะใช้ข้อมูลแต่ละชุด แล้วเปรียบเทียบค่า R^2 , F , ค่าประมาณของ coefficient แต่ละคู่ และค่า t-statistics แต่ละคู่

ส่วนวิธีที่สองนั้น ผู้วิจัยอาจรวมข้อมูล 2 ชุด เข้าด้วยกัน (โดยมีเงื่อนไขว่า ถ้าจะให้ง่ายแล้ว การให้รหัสข้อมูลทั้ง 2 ชุดนั้น จะต้องอยู่ในระบบเดียวกัน และเหมือนกันด้วย เมื่อรวมข้อมูลแล้ว ก็จะสามารถวิเคราะห์ถดถอยโดย OLS ธรรมดาเพียงครั้งเดียว แต่จะต้องสร้างตัวแปร dummy ขึ้นมา 1 ตัว เพื่อให้เป็นตัวแปรคุมที่จะแสดง และสะท้อนถึงผลของความแตกต่างในข้อมูล 2 ชุดนั้นไปในตัว

ข้อแตกต่างของ 2 วิธีนี้จะอยู่ที่ว่าวิธีแรกนั้น เป็นการแสดงถึงตัวกำหนดค่าแปรผันในตัวแปรตามสำหรับข้อมูลแต่ละชุด เช่น ตัวกำหนดความต้องการมีบุตรของประชากรเขตเมือง และตัวกำหนดความต้องการมีบุตรของประชากรในเขตชนบท เป็นต้น แต่ในกรณีที่สองนั้น จะเท่ากับวิเคราะห์ถึงตัวกำหนดความต้องการมีบุตรของประชากรทั้งหมด (ทั้งในเมือง และในชนบท) ขณะเดียวกันได้ดูว่า ความเป็นเมืองกับความเป็นชนบท (โดยผ่านตัวแปร dummy ที่กำหนดนั้น) มีผลสะท้อนถึงความต้องการมีบุตรด้วยหรือไม่ และเพียงใด

ตารางแสดงการเปรียบเทียบผลการวิเคราะห์ถดถอยของภาวะเจริญพันธ์ (โดยวิธี OLS)

	Rural	Urban	Pooled
Constant	-0.4861	-0.2026	0.1583
CL	2.9944* (5.939)	3.4681* (8.512)	3.3526* (10.523)
SR	0.5383 (1.932)	0.5563* (3.629)	0.5640* (4.090)
IF	0.2555* (5.860)	0.4653* (13.751)	0.3671* (13.801)
LM	0.2824* (21.075)	0.1776* (24.132)	0.2082* (31.545)
CU	0.7402* (2.826)	0.4883* (4.702)	0.5331* (5.310)
WE	-0.0249 (0.245)	-0.2013* (5.751)	-0.1706* (4.959)
UR			-0.4589* (4.313)
N	478	1,068	1,546
d.f.	471	1,061	1,538
R^2	0.6064	0.6018	0.6016
F – Statistics	120.9540*	267.2560*	331.8833*

<u>หมายเหตุ</u>

- แสดงว่ามีนัยสำคัญทางสถิติตั้งแต่ 5 % ขึ้นไป
- ตัวเลขในวงเล็บแสดงค่า t-statistics
- ผลการวิเคราะห์มาจาก Thienchay Kiranandana , An Economics Analysis of Fertility Determination Among Rural and Urban Thai Women, Institute of Population Studies Paper No. 20 , (Bangkok : Chulalongkorn University, 1977).

จากตารางผลการวิเคราะห์ถดถอยโดยวิธี OLS ธรรมดา ที่แสดงในหน้าก่อนจะเป็นการเปรียบเทียบ เพื่อชี้ให้เห็นถึงความแตกต่างในผลการวิเคราะห์ ระหว่างการวิเคราะห์แยกเป็น 2 สมการ สำหรับชนบท 1 สมการ และเมือง อีก 1 สมการ กับการวิเคราะห์โดยรวมตัวอย่างข้อมูลทั้งชนบท และเมืองเข้าด้วยกัน (ทั้งนี้โปรดสังเกตขนาดของตัวอย่างที่มีความแตกต่างกันค่อนข้างมากระหว่างเขตชนบท และเขตเมืองด้วย)