บทที่ 2

ความถดถอยเชิงเส้นอย่างง่าย

(Simple Linear Regression)

2.1 ข้อตกลงของสมการถดถอย

ในงานวิจัยที่ต้องอาศัยเทคนิคการวิเคราะห์ความถดถอยเป็นเครื่องมือช่วยเสาะหาหรือสอบย้ำคำตอบนั้น ประเด็นที่สนใจก็คือพฤติกรรมในปัจจุบันและอนาคตของตัวแปร Y คำว่าพฤติกรรมในปัจจุบันหมายถึงโครงสร้าง หรือ ส่วนประกอบของตัวแปร Y ส่วนพฤติกรรมในอนาคตหมายถึงค่าพยากรณ์ในอนาคตของตัวแปร Y การมีความรู้พฤติกรรมทั้งปัจจุบันและอนาคตของตัวแปรที่สนใจมีผลให้นักวิจัยสามารถวิเคราะห์องค์ประกอบ เค้าโครง ตลอดจนแนวโน้มของตัวแปร Y อันจะมีประโยชน์โดยตรงต่อการวางแผนงานเพื่อให้การบริหารงานเป็นไปด้วยความเรียบร้อยและเหมาะสม

โดยปกติตัวแปร Y จะมีค่าสังเกตของตัวเองและสามารถผันแปรค่าไปได้ตามธรรมชาติอยู่แล้ว สิ่งที่นักวิจัยต้องการก็คือ การที่ Y มีค่าเท่าที่เป็นอยู่หรือแปรค่าไปได้นั้นมีอะไรเป็นสาเหตุ โดยตั้งความหวังไว้ว่าถ้าทราบสาเหตุ (X's) หรือสามารถควบคุมพฤติกรรมของสาเหตุ[1] เหล่านั้นได้ก็จะสามารถทราบค่าของ Y และควบคุมพฤติกรรมของ Y ทั้งอดีต-ปัจจุบันและอนาคตได้ ปัญหาก็คือนักวิจัยสามารถเสาะหาสาเหตุเหล่านั้นได้ครบถ้วนหรือไม่ สามารถควบคุมพฤติกรรมของสาเหตุรวมทั้งวัดค่าออกมาเป็นตัวเลขได้เพียงใด ตัวแปร Y และสาเหตุแห่งความผันแปรมีโครงสร้างความสัมพันธ์ทางคณิตศาสตร์ในรูปใด ผู้วิจัยสามารถระบุความสัมพันธ์นั้นถูกต้องตามความสัมพันธ์ที่เป็นอยู่จริงได้หรือไม่ สำหรับในกรณีของ Simple Linear Regression นักวิจัยจะ

---

[1] เนื่องจากงานวิเคราะห์ความถดถอยมีเป้าหมายเพื่อที่จะศึกษาพฤติกรรมในอดีต-ปัจจุบัน และอนาคตของตัวแปร Y โดยพยายามเสาะหาสาเหตุที่สามารถใช้เป็นดัชนีช่วยชี้ความผันแปรในค่าของ Y โดยจำแนกออกมาในรูปของตัวแปรอิสระ X's ด้วยเหตุนี้ประเด็นแห่งความสนใจของงานวิเคราะห์ความถดถอยอีกประการหนึ่งที่อาจถือว่าสืบเนื่องมาจากประเด็นแรกก็คือ การมุ่งศึกษาความสัมพันธ์ในรูปสาเหตุผลลัพธ์ (Causual Relationship) ระหว่างตัวแปร Y กับ X_j และ Y กับ X's และขอชี้ให้เห็นข้อสังเกตไว้ที่นี้ประการหนึ่งด้วยว่า การศึกษาความสัมพันธ์ระหว่างตัวแปรนั้นเราสามารถใช้เทคนิคการวิเคราะห์สหสัมพันธ์เป็นเครื่องมือได้ แต่ค่าสหสัมพันธ์จะเป็นเพียงดัชนีที่วัดระดับความสัมพันธ์ระหว่างตัวแปรเท่านั้น ไม่สามารถอธิบายในเชิงสาเหตุ-ผลลัพธ์ ได้.

เสาะหาปัจจัยภายนอกที่เห็นว่ามีอิทธิพลต่อ Y เพียงตัวหนึ่งตัวเดียวโดยเอาเฉพาะ-ปัจจัยที่วัดค่าได้ และมีอิทธิพลต่อความผันแปรของ Y มากที่สุด ซึ่งในขั้นต้นจะระบุความสัมพันธ์ทางคณิตศาสตร์ $Y = f(X)$ ในรูปของสมการเชิงเส้นคือ $Y = \alpha + \beta X$ เมื่อ α คือจุดตัดบนแกน Y และ $\beta = \frac{\Delta Y}{\Delta X}$ = Slope (ความชัน) ซึ่งใช้วัดระดับอิทธิพลที่ตัวแปรอิสระ X มีต่อความผันแปรของ Y ภายหลังเมื่อตรวจสอบความเหมาะสมแล้วพบว่าความสัมพันธ์ $Y = f(X)$ มิใช่ความสัมพันธ์เชิงเส้น นักวิจัยค่อยเปลี่ยนรูปความสัมพันธ์เป็นรูปอื่นที่เห็นว่าถูกต้องกว่าต่อไป

อย่างไรก็ตาม การระบุความสัมพันธ์ $Y = f(X)$ ในรูปของสมการเชิงเส้นคือ $Y = \alpha + \beta X$ นั้นนับว่าหละหลวมอยู่มาก นอกจากนี้การระบุตัวแปรอิสระ X ไว้เพียงตัวเดียวซึ่งแม้ว่าจะมีความผูกพันกับ Y สูงเพียงใดก็มิใช่ว่า X จะดูดซับเอาความเคลื่อนไหวทั้งปวงของ Y ไว้ได้ครบถ้วน ปัจจัยอื่น ๆ เช่น ตัวแปรอิสระ X's ที่คัดทิ้งไปก็ยังคงส่งอิทธิพลต่อความเคลื่อนไหวของ Y อยู่เป็นปกติทำให้ค่า Y จริง กับค่า Y ที่พึงคาดหมายได้จากสมการ $Y = \alpha + \hat{\beta}X$ คลาดเคลื่อนกัน หรือพูดโดยนัยกลับกันได้ว่าค่า Y จริง กับ $\hat{Y}$ ($\hat{Y} = \hat{\alpha} + \hat{\beta}X$) จะคลาดเคลื่อนกันอยู่แล้วเป็นปกติ ความคลาดเคลื่อนจะเกิดขึ้นได้ต้องมีสาเหตุ และสาเหตุหนึ่งที่นับว่าสำคัญก็คือการระบุปัจจัยภายนอกหรือตัวแปรอิสระ X's ไว้ไม่ครบถ้วน นอกจากนี้ยังมีสาเหตุอื่น ๆ อีกหลายประการที่ล้วนส่งผลสะท้อนให้ Y และ $\hat{Y}$ มีความคลาดเคลื่อนกันได้ ซึ่งสามารถสรุปเป็นหมวดหมู่ดังนี้

1. การระบุตัวแปรอิสระไม่ครบถ้วน

โดยปกติปัจจัยภายนอกที่มีผลให้ค่าของ Y ผันแปรไปได้นั้นมีอยู่มากมายจนนับไม่ถ้วน บางปัจจัยมีอิทธิพลมาก บางปัจจัยมีอิทธิพลน้อย บางปัจจัยวัดค่าเป็นตัวเลขได้ บางปัจจัยวัดค่าเป็นตัวเลขไม่ได้ แม้ว่าเราจะพยายามทุกวิถีทางที่จะระบุปัจจัยภายนอกหรือตัวแปรอิสระที่สามารถควบคุมความเคลื่อนไหวของ Y ให้ได้มากที่สุดเท่าที่จะมากได้[1] แต่ใช่ว่าจะสามารถใช้ตัวแปรอิสระเหล่านั้นได้ทั้งหมดทั้งนี้ก็ด้วยเหตุผลดังกล่าว ตัวแปรที่พิจารณาเห็นว่ามีอิทธิพลน้อยหรือแม้ว่าจะมีอิทธิพลมากแต่วัดค่าไม่ได้นักวิจัยก็จำเป็นต้องคัดทิ้ง หรือแม้ว่าตัวแปรอิสระที่ผ่านการคัดเลือกและกลั่นกรองมาดีแล้ว แต่มีจำนวนมากเกินไป (k) เมื่อเทียบกับจำนวนตัวอย่าง

---

[1] ถือว่านักวิจัยระบุตัวแปรอิสระได้ครบถ้วน แต่ถ้านักวิจัยระบุตัวแปรอิสระไม่ครบถ้วน อาจเพราะไม่รู้หรือคาดคิดไม่ถึงว่าปัจจัยนั้น ๆ จะมีอิทธิพลต่อ Y กรณีนี้ก็นับว่าเป็นปัญหา เพราะปัจจัยที่ไม่ได้ระบุไว้จะส่งอิทธิพลต่อ Y อยู่เช่นเดียวกับปัจจัยอื่นเพียงแต่มิได้ปรากฏโฉมหน้าอยู่ในแบบจำลองเท่านั้น

(ฑ) นักวิจัยก็ต้องคัดทิ้งไปอีกเพราะจะมีผลเสียต่อคุณภาพของงานเนื่องจาก df ของ Residnal (n-k) มีค่าต่ำเกินไป โดยเฉพาะอย่างยิ่งงานวิเคราะห์ความถดถอยเชิงเส้นอย่างง่ายนักวิจัยจะคัดเลือกเอาไว้เฉพาะตัวแปรอิสระที่พิจารณาเห็นว่ามีอิทธิพลต่อ Y อย่างแท้จริงเพียงตัวหนึ่งตัวเดียวเท่านั้น ซึ่งกรณีนี้จะเป็นกรณีที่เราคัดตัวแปรอินระอื่น ๆ ทิ้งไปมากกว่ากรณีของ Multiple Reqression ตัวแปรที่ถูกคัดทิ้งไปนั้นแม้จะไม่ปรากฏโฉมหน้าอยู่ในแบบจำลองก็จริง แต่ยังคงส่งอิทธิพลต่อความผันแปรของ Y อยู่เป็นปกติ ด้วยเหตุนี้ค่า $\hat{Y}$ จึงยังคงคลาดเคลื่อนไปจากค่า Y เพราะ Y มีความผันแปรค่าไปเพราะการผสมผสานของอิทธิพลของตัวแปรอิสระ หรือปัจจัยภายนอกทุกรายการขณะที่ $\hat{Y}$ มีความผันแปรค่าไปเพราะอิทธิพลของตัวแปรอิสระเพียงตัวเดียว (กรณี Simple Regression) หรืออิทธิพลของตัวแปรอิสระเพียงบางส่วน (กรณี Multiple Regression)

2. พฤติกรรมเชิงสุ่มของมนุษย์

มีงานวิจัยจำนวนมิใช่น้อยที่ต้องอาศัยข้อมูลจากการสังเกต การสำรวจหรือการบันทึกผลการตอบสนองจากพฤติกรรมของมนุษย์ พฤติกรรมของมนุษย์เป็นสิ่งที่คาดหมายได้ยากและมีอัตราการผันแปรไปได้ค่อนข้างสูงเมื่อเทียบต่อวัตถุทดลอง ดังนั้นข้อมูลที่พึงได้รับจากมนุษย์จึงมีลักษณะที่คลาดเคลื่อนอยู่มากเพราะมี Error ผนวกอยู่ในข้อมูล ผลลัพธ์จึงปรากฏออกมาในลักษณะที่ทำให้ค่า Y และ $\hat{Y}$ คลาดเคลื่อนจากกัน ตัวอย่างเช่นงานทางเศรษฐศาสตร์ที่มุ่งศึกษาลักษณะการใช้จ่ายของครัวเรือน (Y) นั้นโดยปกติลักษณะการใช้จ่ายของครัวเรือนจะเปลี่ยนแปลงไปถ้าหากว่าราคาสินค้าเปลี่ยนแปลงไป และครอบครัวมีรายได้เปลี่ยนแปลง แต่เนื่องจากพฤติกรรมของมนุษย์เป็นเรื่องที่ยากจะคาดหมายได้ และมักเปลี่ยนแปลงไปโดยสุ่มเสมอเราจึงมักพบอยู่บ่อยครั้งว่าครอบครัวมีลักษณะการใช้จ่ายผิดแปลกไปจากเดิมทั้ง ๆ ที่สินค้าและรายได้ของครอบครัวยังไม่ไม่การเปลี่ยนแปลง จึงเห็นได้ว่าถ้าจะให้สามารถควบคุมความเคลื่อนไหวของ Y ได้ดีนักวิจัยควรผนวกเอาพฤติกรรมเชิงสุ่มของมนุษย์ไว้ในแบบจำลองนอกเหนือไปจากการมีเฉพาะ X's ด้วย

3. การระบุโครงสร้างความสัมพันธ์ทางคณิตศาสตร์ Y = f (X) หรือ Y = f (X's) ผิดพลาด

การระบุโครงสร้างความสัมพันธ์ทางคณิตศาสตร์ (Mathimatical Form of Model) เป็นสิ่งจำเป็นที่สุดเพราะเป็นเรื่องของการกำหนดฟังก์ชั่นของตัวแปร Y ถ้ากำหนดฟังก์ชั่นผิด ค่าของ Y กับ $\hat{Y}$ จะคลาดเคลื่อนจากกัน อันจะมีผลเสียหายต่องานทั้งในแง่ของโครงสร้างและการพยากรณ์

เช่น Y และ X ควรสัมพันธ์กันในรูปของ Exponential คือ $Y = \alpha e^{\beta X}$ แต่นักวิจัยกำหนดความสัมพันธ์ในรูปของสมการเชิงเส้น $Y = \alpha + \beta X$ ความผิดพลาดในลักษณะนี้จะมีผลให้ $\hat{Y}$ คลาดเคลื่อนไปจากค่า Y

4. ความผิดพลาดของยอดรวมข้อมูล

ในบางครั้งนักวิจัยจำเป็นต้องหายอดรวมข้อมูลสำหรับใช้เป็นข้อมูลของตัวแปรอิสระหรือตัวแปรตาม เช่นหารายได้รวมของประชาชนในท้องถิ่นที่สำรวจด้วยการนำรายได้ของทุกครัวเรือนมารวมกัน หาผลผลิตรวมของสินค้า ก.ของประเทศด้วยการนำผลผลิตสินค้า ก.จากทุกโรงงานมารวมกัน หายอดรวมของปริมาณการบริโภคสินค้า ข.ด้วยการนำปริมาณการบริโภคของทุกครัวเรือนที่บริโภคสินค้าข.มารวมกัน เป็นต้น ขอให้สังเกตว่าการหายอดรวม (Aggregration) ดังกล่าวเราจะกระทำโดยวิธีง่าย ๆ คือนำเอาข้อมูลมารวมกันโดยตรง แต่สิ่งที่ก่อให้เกิดปัญหาก็คือธรรมชาติของข้อมูลจากต่างแหล่งย่อมมีเงื่อนไขและแบบแผนเฉพาะตัวอีกทั้งมีวิธีเก็บรวบรวมรวมถึงระยะเวลาจัดเก็บต่างกัน เช่น ข้อมูลปริมาณการผลิตสินค้า ก. แต่ละโรงงานจะมีนโยบายการผลิตต่างกัน มีวิธีควบคุมคุณภาพต่างกัน มีระยะเวลาจัดเก็บข้อมูลสั้นยาวไม่เท่ากันและมีเทคนิคการจัดเก็บข้อมูลรวมถึงความละเอียดถี่ถ้วนต่าง ๆ กัน เมื่อเป็นเช่นนี้ข้อมูลจากต่างแหล่งย่อมมีลักษณะเฉพาะตัวต่างกัน การนำข้อมูลต่างแหล่งมารวมกันโดยตรงจึงทำให้นักวิจัยผนวกเอา Error เข้าไปในยอดรวมด้วย ดังนี้เป็นต้น

5. ความผิดพลาดจากการวัดค่า (Measurement Error หรือ Error of Observation)

คำว่า ความผิดพลาดจากการวัดค่า เราหมายถึงความผิดพลาดจากเฉพาะการวัดค่าของตัวแปร Y เท่านั้น ไม่รวมถึงตัวแปรอิสระ X's เพราะถือว่าสามารถควบคุมค่าของ X's ได้ ความผิดพลาดจากการวัดค่านับว่าเป็นปัจจัยสำคัญประการหนึ่งที่ทำให้ค่าสังเกต Y คลาดเคลื่อนไปจากความเป็นจริง และเมื่อข้อมูลเดิมของ Y ผิดพลาดเสียแล้วแต่ต้นผลลัพธ์ เช่นโครงสร้าง และค่าในอนาคตของ Y ย่อมคลาดเคลื่อนไปจากความเป็นจริงไปด้วย

ด้วยเหตุผลประการดังกล่าว และด้วยความมุ่งหวังที่จะให้สมการ $Y = f(X\text{'s})$ สามารถใช้วิเคราะห์โครงสร้าง และพยากรณ์ค่าของ Y ในอนาคตได้อย่างถูกต้อง เราจึงสมควรนำเอา Error จากแหล่งต่าง ๆ เหล่านี้มาผนวกไว้ในแบบจำลองด้วย โดยให้ Error จากแหล่งต่าง ๆ รวม

ตัวกันอยู่ในที่เดียวกันให้ชื่อเป็นตัวแปรตัวใหม่ว่า u โดยค่าของ u เกิดจากการผสมผสานของ Error ที่พึงมีจากทุกแหล่งดังกล่าวข้างต้น สมการใหม่คือ

$$Y = \alpha + \beta X + u$$

จึงเป็นสมการที่ระบุความสัมพันธ์ระหว่างตัวแปรได้ครบถ้วน และจะไม่มี Error อีกต่อไป เพราะ Error ทั้งปวงที่พึงมีจะปรากฏโฉมหน้าร่วมกันในรูปของตัวแปร u และถือว่าสมการนี้คือสมการที่แสดงความสัมพันธ์ระหว่างตัวแปรที่แท้จริง และด้วยเหตุที่ u เป็นแหล่งรวมของปัจจัยต่าง ๆ มากมายหลายรายการรวมตลอดถึง Error ด้วยดังนั้นตัวแปร u จึงเป็นตัวแปรที่ไม่อาจวัดค่า หรือบอกค่าสังเกตได้ (u เป็น Unobservable Variable) แม้ค่าสังเกตจริงของ u จะมีอยู่ตามธรรมชาติแล้วก็ตาม อีกทั้งมีลักษณะของความเคลื่อนไหวที่นักวิจัยไม่อาจควบคุมได้ทั้งมีค่าแปรเปลี่ยนได้ในลักษณะที่นักวิจัยไม่อาจคาดหมายทั้งขนาดเครื่องหมายและทิศทางได้ และเพื่อให้งานวิเคราะห์ความถดถอยดำเนินต่อไปได้โดยมิต้องถูกก่อกวน (Disturb) ด้วยตัวแปร u ดังกล่าวเราจึงจำเป็นต้องหาหนทางควบคุมพฤติกรรมของ u ไว้ในกรอบที่เห็นว่าเหมาะสมและอำนวยประโยชน์แก่งานทั้งในเชิงทฤษฎีและเชิงปฏิบัติดังนี้

1. u เป็นตัวแปรเชิงสุ่ม

เหตุที่ u มีความเคลื่อนไหวได้อย่างเสรีทั้งมีค่าที่คาดหมายไม่ได้พฤติกรรมของตัวแปร u จึงอยู่นอกกฎเกณฑ์โดยที่นักวิจัยไม่อาจควบคุมได้ และการที่ u มีความเคลื่อนไหวได้อิสระเกินไปนี้เองที่ส่งผลให้ค่าของ Y ผันผวนไปจนยากที่จะคาดหมายได้ เพราะเมื่อเรากำหนดค่า X ให้คงที่เท่ากับ Xo ซ้ำ ๆ กัน [1] ค่าของ Y จะไม่คงที่ซ้ำค่าเดิม แต่จะผันผวนไปได้ที่เป็นเช่นนี้ก็เพราะ ณ. ค่า $X = X_0$ ค่าของ u จะไม่คงที่เท่ากับ u_0 แต่ค่าของ u จะผันแปรไปอย่างเสรีจนเราไม่อาจคาดหมายได้ ผลสะท้อนจึงปรากฏสู่ค่าของ Y คือค่าของ Y จะผันผวนไปอย่างเสรีเช่นกัน ด้วยเหตุนี้เราจึงมีความจำเป็นต้องควบคุมการเคลื่อนไหวของ u ให้อยู่ในกรอบที่พอจะควบคุมได้และพอจะคาดหมายพฤติกรรมได้โดยพยายามผ่อนปรนเกณฑ์การควบคุมนั้นให้ใกล้เคียงกับธรรมชาติเดิมของ u มากที่สุด วิธีแรกก็คือกำหนดให้ u เป็นตัวแปรเชิงสุ่ม การกำหนดให้ u

[1] เช่น ใส่ปุ๋ยลงในแปลงทดลองขนาดมาตรฐาน r แปลง ๆ ละ 1 กิโลกรัมเท่า ๆ กัน แล้ววัดปริมาณผลผลิตของพืชจากแปลงต่าง ๆ ทั้ง r แปลง ถ้าไม่มีอิทธิพลของตัวแปร u คอยรบกวนอยู่ผลผลิตจากทุกแปลงจะต้องเท่ากัน

เป็นตัวแปรเชิงสุ่มจะมีผลให้ u ถูกควบคุมด้วยทฤษฎีความน่าจะเป็น การเคลื่อนไหวของ u จะเป็นไปภายใต้กรอบของค่าที่พึงเป็นไปได้ของตัวแปรสุ่มนั้น ๆ อีกทั้งทำให้เราสามารถคาดหมายค่าของ u และความผันผวนของ u ได้ด้วย และเมื่อเราถือว่า u เป็นตัวแปรสุ่มผลที่ตามมาก็คือ Y จะเป็นตัวแปรเชิงสุ่มด้วยเพราะ Y เป็นฟังก์ชันเชิงเส้น (Linear Function) ของ u ในรูป $Y = \alpha + \beta X + u$ หรือในกรณีทั่วไป $Y = \beta_1 + \sum_{j=2}^{k} \beta_j X_j + u$ ทำให้เราสามารถหาการแจกแจง รวมถึงค่าคาดหมาย และความแปรปรวนของ Y ได้ นอกจากนี้ เนื่องจาก $\hat{\beta} = (X'X)^{-1} X'Y$ ซึ่งแสดงว่า $\hat{\beta}$ เป็น Linear Function ของ Y ผลที่ตามมาก็คือ $\hat{\beta}$ เป็นตัวแปรเชิงสุ่มด้วย

2. ณ. วาระใด ๆ ค่าเฉลี่ยของตัวแปร u จะเท่ากับ 0 **เสมอ** ($E(u_i) = 0; i = 1,2,\ldots,n$)

ข้อตกลงนี้มีความหมายว่า โดยถัวเฉลี่ยแล้วค่าของ u_i มีค่าเท่ากับ 0 กล่าวคือ u_i ควรมีค่าน้อย ๆ อาจจะมีค่าเป็นบวกบ้างลบบ้าง แต่เมื่อหักลบกันแล้วจะหักล้างกันหมดไป หรือเมื่อนำมาถัวเฉลี่ยแล้วจะมีค่าเท่ากับ 0 $E(u_i) = 0$ หมายความว่าเมื่อทำการทดลองซ้ำ ๆ กัน ณ. ค่า $X_i = X_{io}$ จะได้ค่า u_i ต่าง ๆ กัน u_i บางค่าอาจเป็นลบบางค่าอาจเป็นบวก แต่โดยถัวเฉลี่ยแล้วค่าของ u_i ณ. $X_i = X_{io}$ จะเท่ากับ 0

เหตุที่ $E(u_i) = 0$ ผลสะท้อนที่ติดตามมาก็คือ $\mu_{y|x} = E(Y_i) = E(\alpha+\beta X_i+u_i) = \alpha+\beta X_i$ หรือในกรณีทั่วไป $\mu_{y|x}$'s $= E(\beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i) = \beta_1 + \sum_{j=2}^{k} \beta_i X_{ji}$ ซึ่งแสดงว่าค่าเฉลี่ยของ Y_i ณ $X_i = X_{io}$ จะปรากฏอยู่บนเส้นตรง $E(Y) = \alpha+\beta X$ หรือโดยนัยกลับกันสมการ $E(Y) = \alpha+\beta X$ จะพุ่งผ่านค่าเฉลี่ยของ Y_i เสมอ และในกรณีทั่วไปสมการของ Surface คือ $E(Y) = \beta_1 + \sum_{j=2}^{k} \beta_j X_j$ จะพุ่งผ่านค่าเฉลี่ยของ Y_i เสมอ[1]

3. ณ วาระที่ i และ j ใด ๆ ที่ $i \neq j$ $E(u_i u_j) = 0$ **หรือตัวแปรสุ่ม u ในต่างวาระกันจะต้องเป็นอิสระต่อกัน**

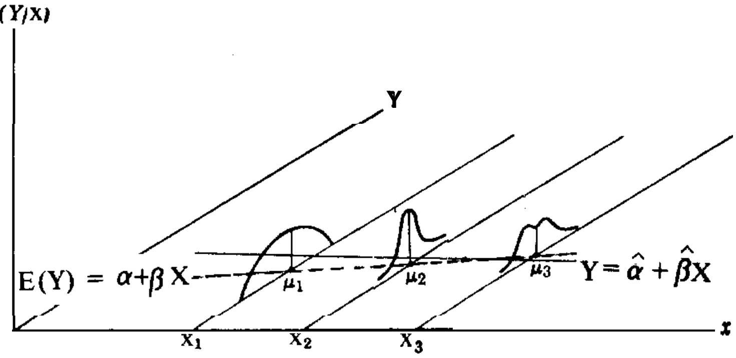
ข้อตกลงนี้มีผลให้ Y_i และ Y_j เป็นอิสระต่อกัน หรือไม่ขึ้นอยู่แก่กันและกัน โค้งของ Y_i และ Y_j จึงปรากฏได้ในรูปของ Marginal Density มิใช่ Joint Density ทำให้สามารถแยกวัดค่า

---

[1] $\mu_{y|x}$ หมายถึงค่าเฉลี่ยของ Y_i เมื่อกำหนดให้ X_i คงที่เท่ากับ X_{io} แล้วทดลองซ้ำ ๆ กัน ณ $X_i = X_{io}$ **ผลลัพธ์**คือได้ค่า Y_i ที่แตกต่างกัน (อาจมีบางค่าซ้ำกันได้) ค่าเฉลี่ยของ Y_i ในกรณีนี้ก็คือ $\mu_{y|x}$ หรือ μ_i

Y_i ณ ค่า $X_i = X_{io}$ ได้โดยมิต้องคำนึงถึงผลกระทบที่จะพึงเกิดขึ้นกับ Y_j

จากข้อตกลงทั้งสามประการทำให้สามารถเสนอรูปทรงทางเรขาคณิตให้เห็นดังภาพและขอให้สังเกตว่าเรายังไม่ได้กำหนดรูปแบบการแจกแจงของ u ไว้ว่าจะต้องมีลักษณะใด โค้งของ u_i จึงเป็นไปได้หลายลักษณะ ลักษณะที่ได้ควบคุมไว้แล้วที่สามารถเห็นได้จากภาพก็คือ $\mu_{y|x}$ วางอยู่บนเส้นตรง $E(Y) = \alpha+\beta X$ หรืออีกนัยหนึ่งเส้นตรง $E(Y) = \alpha+\beta X$ ลากผ่านจุดที่ $\mu_{y|x}$ ปรากฏอยู่[1] และโค้งของ u_i แยกจากกัน มิได้มีลักษณะเชื่อมติดต่อกันใน **Bivariate Nornal หรือ Multivariate Density** ใด ๆ



ภาพ 2.1 แสดงให้เห็นว่า u เป็นตัวแปรเชิงสุ่มที่เป็นอิสระต่อกัน และเส้นตรง $E(Y) = \alpha+\beta X$ เป็น True Line[2] จะลากผ่านจุดที่เป็นค่าเฉลี่ยของ Y_i เมื่อกำหนดค่า X_i ซ้ำ ๆ กัน ณ ค่า $X_i = X_{io}$; $i = 1, 2, \ldots, n$
เส้นตรง $Y = \hat{\alpha} + \hat{\beta}X$ คือ Estimated Line ซึ่งสร้างขึ้นโดยอาศัยค่าสังเกต Z_i $i = 1, 2, \ldots, n$ ขอให้สังเกตแนวที่เส้น $Y = \hat{\alpha}+\hat{\beta}X$ เบี่ยงเบนไปจากจุด $\mu_{y|x}$ ระยะเบี่ยงเบนไปนี้ คือ e_i ซึ่งจะใช้เป็นค่าประมาณของ u_i เมื่อ $u_i = Y_i - (\alpha - \beta X_i)$ ขณะที่ $e_i = Y_i - (\hat{\alpha} - \hat{\beta} X_i)$

---

[1] เมื่อ u_i เป็นตัวแปรเชิงสุ่ม Y_i ย่อมเป็นตัวแปรเชิงสุ่มด้วย เพราะ Y_i เป็น Linear Function ของ u_i ดังนั้นถ้าเรากำหนดการแจกแจงให้ u เป็นรูปใด Y_i ก็จะมีการแจกแจงรูปนั้นด้วย แต่จะมี mean และ Variance ต่างกัน

[2] โดยปกติเราจะไม่มีโอกาสทราบ True Line ได้เลย สิ่งที่จะทราบได้จะมีเพียงเฉพาะ Estimated Line ซึ่งประมาณได้จากกลุ่มสังเกต Z_i เท่านั้น ด้วยเหตุนี้ Error ที่เราจะพบได้จึงมีเฉพาะเพียง e_i มิใช่ u_i True Line คือ $E(Y) = \alpha + \beta X$ และ True Error คือ u_i จึงเป็นสิ่งที่มีอยู่ในอุดมคติ สิ่งที่พบจริงมีเพียง Estimated Line คือ $Y = \hat{\alpha}+\hat{\beta}X$ และ Estimated Error คือ e_i

4. ณ วาระใด ๆ ความแปรปรวนของตัวแปรเชิงสุ่ม u_i จะมีค่าคงที่เสมอ

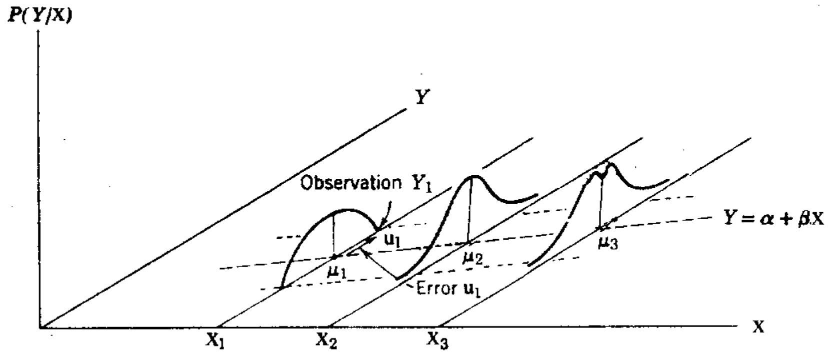
จากข้อตกลง 3 ประการขั้นต้น แม้ว่าจะทำให้เราสามารถควบคุมตัวแปรสุ่ม u_i ได้ (ซึ่งหมายถึงการควบคุมตัวแปรเชิงสุ่ม Y_i ได้ในขณะเดียวกัน) แต่ก็ยังไม่เพียงพอเพราะค่าของ u_i ณ วาระที่ $X_i = X_{io}$; $i = 1, 2, \ldots, n$ มีลักษณะของการผันแปรค่าไปได้มาก การบังคับให้ u_i มีค่าเฉลี่ยหยุดอยู่ที่จุด 0 มิได้ทำให้ u_i ขาดเสรีภาพในการผันแปรค่าไปมากนัก บางวาระของ X เช่นเมื่อให้ X คงที่เท่ากับ X_1 และทดลองซ้ำ ๆ เพื่อวัดค่าของ Y_1 ค่าของ u_1 จะผันแปรไปเพียงเล็กน้อยอันมีผลให้ Y_1 ผันแปรค่าไปเล็กน้อย ขณะเดียวกัน เมื่อให้ X คงที่ที่ X_2 แล้วทดลองซ้ำ ๆ เพื่อวัดค่า Y_2 กลับปรากฏว่า u_2 ผันแปรค่าไปอย่างกว้างขวางทำให้ Y_2 ผันแปรค่าไปได้มากตามไปด้วย ฐานโค้งของ $f(y_1|x_1)$ กับ $f(y_2|x_2)$ จึงกว้างแคบต่างกัน (ดูรูป 2.1) ในทำนองเดียวกันเมื่อเราทดลองหยุดค่า X ให้คงที่เท่ากับ $X_3, X_4, \ldots, X_n$ ค่าของ $u_3, u_4, \ldots, u_n$ ก็จะผันแปรไปโดยอิสระภายในกรอบของ $E(u_3) = 0, E(u_4) = 0, \ldots, E(u_n) = 0$ ซึ่งมีผลให้ค่าของ $Y_3, Y_4, \ldots, Y_n$ ผันแปรไปโดยอิสระ ฐานโค้งของ $f(y_3|x_3), f(y_4|x_4), \ldots, f(y_n|x_n)$ จึงกว้างแคบต่างกัน เมื่อเหตุการณ์ทำนองนี้เกิดขึ้น เส้นสมการคาดหมายซึ่งมีสิทธิที่จะเบี่ยงเบนไปจากเส้นจริง (True Line) ภายในกรอบของฐานโค้งจึงบิดเบนและเปลี่ยนค่าความชันไปได้ จึงทำให้มีเส้นสมการคาดหมายที่พึงเป็นไปได้มากมายหลายเส้น ความน่าเชื่อถือในความถูกต้องของสมการคาดหมายเส้นหนึ่งเส้นใดจึงมีได้น้อย[1]

เพื่อปิดกั้นปัญหาดังกล่าวและเพื่อให้สมการคาดหมายที่พึงเป็นไปได้ทุกสมการมีลักษณะของความถูกต้องน่าเชื่อถือ เราจึงจำเป็นต้องกำหนดหรือควบคุมให้ฐานโค้งของตัวแปรสุ่ม u_i ทุกตัวมีความกว้างคงที่เท่ากับ σ^2 นั่นคือกำหนดให้ $V(u_i) = \sigma_U^2$; $i = 1, 2, \ldots, n$ ผลสะท้อนที่ติดตามมาก็คือ $V(Y_i) = V(\alpha+\beta X_i+u_i) = 0+0+V(u_i) = \sigma^2$ หรือในกรณีทั่วไป $V(Y_i) = V(\beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i) = 0+V(u_i) = \sigma^2$ แสดงว่าถ้าเรากำหนดให้โค้งของตัวแปรสุ่ม u_i มีฐานกว้างคงที่เท่ากันในทุกค่าของ i ; $i = 1, 2, \ldots, n$ แล้วโค้งของ Y_i ก็จะมีฐานกว้างคงที่เท่ากันเท่ากับ σ^2 ในทุกค่าของ i ; $i = 1, 2, \ldots, n$ ด้วยเช่นกัน

---

[1] เส้นสมการคาดหมายที่น่าเชื่อถือคือเส้นสมการที่พุ่งผ่านเข้าไปในกลุ่มของค่าสังเกตอย่างเหมาะสมและใช้ข้อมูลข้อสนเทศทุกประการที่มีอยู่ ถ้าฐานโค้งของ $f(y_i|x_i)$ กว้างแคบต่างกันสิ่งที่เป็นไปได้ก็คือจะมีเส้นสมการคาดหมายจำนวนมากจนนับไม่ถ้วน (Infinility Many) ที่ไม่มีลักษณะของสมการที่น่าเชื่อถือตามนัยนี้ ปัญหานี้เรียกว่าปัญหาของ Heteroscedasticity

ผลลัพธ์จากการควบคุมฐานโค้งของ u_i ให้กว้างคงที่เท่ากัน ทำให้ภาพ 2.1 กลายมาเป็นลักษณะใหม่ดังนี้



ภาพ 2.2 แสดงโค้งของตัวแปรสุ่ม u_i (หรือ Y_i) ที่เป็นอิสระต่อกันและมีฐานโค้งกว้างคงที่เท่ากันเท่ากับ σ^2 [1] โดยที่เส้นสมการจริงจะผ่านไปในจุดต่าง ๆ ที่เป็นค่าเฉลี่ยของ Y_i เมื่อกำหนดให้ X_i คงที่เท่ากับ X_{io} (หรือ $E(u_i) = 0$) โค้งในภาพยังคงมีรูปร่างที่แตกต่างกันไปเพราะเรายังไม่ได้ควบคุมลักษณะการแจกแจงของ u_i

๕. ณ วาระใด ๆ ความแปรปรวนของตัวแปรสุ่ม u_i จะมีการแจกแจงแบบ $N(0, \sigma^2)$

เพื่อที่จะให้โค้งของ $f(y|x)$ มีลักษณะเดียวกันตลอดทั้งยังสามารถพัฒนาเทคนิคการประมาณค่าพารามิเตอร์ และทดสอบสมมุติฐานเกี่ยวกับพารามิเตอร์ต่าง ๆ ได้ด้วย นักวิจัยพึงกำหนดข้อตกลงเกี่ยวกับตัวแปรเชิงสุ่ม u_i เพิ่มเติมเพื่อสนองความต้องการดังกล่าว โดยกำหนดเป็นข้อตกลงว่า $u_i \sim N(0, \sigma_u^2)$ การตกลงให้ $u_i \sim N(0, \sigma_u^2)$, $i = 1, 2, \ldots, n$ มีผลให้ $Y_i \sim N(\alpha+\beta X_i, \sigma_u^2)$ เมื่อ $Y_i = \alpha+\beta X_i+u_i$; $i = 1, 2, \ldots, n$ และ $Y_i \sim N(\beta_1 + \sum_{j=2}^{k} \beta_j X_{ji}, \sigma_u^2)$ หรือ $Y \sim N(X\beta, \sigma_u^2 I_n)$ เมื่อ $Y_i = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$; $j = 1, 2, \ldots, n$ นอกจากนี้การกำหนดให้ $u_i \sim N(0, \sigma_u^2)$; $i = 1, 2, \ldots, n$ ยังมีผลเชื่อมโยงไปถึงการใช้ t-test, F-test, Z-test, χ^2-Test

[1] ในสัญญลักษณ์ σ_u^2 Subscript "u" ใช้เป็นดัชนีแสดงให้เห็นว่า σ^2 นี้คือความแปรปรวนของ u เท่านั้น มิได้มีความหมายเป็นอย่างอื่น ถ้าเข้าใจเช่นนี้แล้ว จะตัด Subscript u ทิ้งเสียก็ได้

และ test อื่น ๆ ที่อาศัยการแจกแจงแบบปกติเป็นรากฐาน[1] อนึ่งขอให้สังเกตว่าข้อตกลงที่ 5 นี้กำหนดเพียงลักษณะการแจกแจงของ u_i เท่านั้น ส่วน $E(u_i)$ และ $V(u_i)$ ยังคงเป็นไปเช่นที่กำหนดไว้เดิมในข้อ 2 และ 4 แต่ข้อตกลงที่ 5 นี้จะมีความจำเป็นเฉพาะเมื่อนักวิจัยสนใจการประมาณค่า β โดยอาศัยวิธี Maximum Likelihood เท่านั้น การประมาณค่า β โดยอาศัยวิธี Ordinary Least Square ยังไม่มีความจำเป็นต้องอาศัยข้อตกลงนี้ หากอาศัยเพียงเฉพาะข้อตกลง 1-4 เท่านั้น

---

[1] ถ้าข้อตกลงนี้ไม่เป็นจริง คือ u ไม่มีการแจกแจงแบบปกติ ปัญหาที่ตามมาคือไม่อาจใช้ t-test , F-test และ test อื่น ๆ ที่อิงการแจกแจงแบบปกติได้ ทั้งไม่อาจหา Confident Region ได้ วิธีประมาณสมการถดถอยในกรณีที่ u ไม่มีการแจกแจงแบบปกติคือ Robust Regression

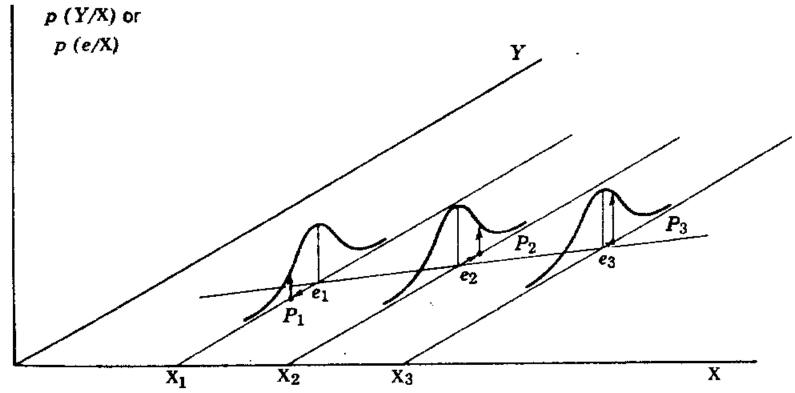
อนึ่งการทดสอบว่า u มีการแจกแจงแบบปกติหรือไม่กระทำได้หลายวิธีเช่น

1. วัดหาค่า Sknewness (γ_1),Kurtosis (γ_2) โดยที่ถ้า $\gamma_1 = 0$ แสดงว่าโค้งสมมาตร ถ้า $\gamma_2 = 3$ แสดงว่าโค้งมีความลาดชันเช่นโค้งปกติ

2. ใช้ χ^2 - test (Goodness of Fit) หรือ Kolmogorov - Smirnov test

3. ใช้ Shapiro - Wilk Statistics $W = \{\sum_i^m a_{n-i+1}(y_{n-i+1} - y_i)\}^2 / \sum_i^n (y_i - \bar{y})^2$ โดยที่ $m = [\frac{n}{2} + 1]$ และ a_{n-i+1} คือค่าจากตารางที่เสนอโดย Shapiro - Wilk อ่านรายละเอียดใน Maddala, G.S., Ibid., p.305-308 และรายละเอียดเรื่องนี้ในบทที่ 5 ซึ่งจะกล่าวถึง Nonnormal Error โดยเฉพาะ

จากการที่เรากำหนดข้อตกลงที่ 5 เกี่ยวกับลักษณะการแจกแจงของตัวแปรสุ่ม u คือ $u_i \sim N(0, \sigma_u^2)$ ดังกล่าวทำให้ภาพ 2.1 และ 2.2 กลายมาเป็นภาพ 2.3 ดังต่อไปนี้



ภาพ 2.3 แสดงโค้งของตัวแปรเชิงสุ่ม u_i (หรือ Y_i) ตามข้อตกลงว่า $E(u_i) = 0$ $V(u_i) = \sigma_u^2$ $Cov(u_i u_j) = 0$ และ $u_i \sim N(0, \sigma_u^2)$ จากภาพขอให้สังเกตว่าเส้นสมการจริงจะพุ่งผ่านไปในจุดต่าง ๆ ที่เป็นค่าเฉลี่ยของ Y_i โดยถือว่าการทดลองกระทำซ้ำ ๆ กัน $X_i = X_{io}$ และโค้งทุกโค้งของ Y_i จะเป็นโค้งปกติที่มีฐานโค้งกว้างเท่ากัน (หรือมียอดโค้งสูงเท่ากัน) สำหรับเส้นสมการคาดหมายที่เป็นไปได้จะเป็นเส้นตรงที่พุ่งอยู่ภายในกรอบของฐานโค้งเท่านั้น

6. ตัวแปรเชิงสุ่ม u_i จะต้องไม่เกี่ยวข้องกับตัวแปรอิสระ นั่นคือ $Cov(u_i X_j) = 0$; $i = 1, 2, \ldots, n$; $j = 2, 3, \ldots, k$

ข้อตกลงนี้มีเจตนาเพื่อบังคับให้สมการถดถอยที่ต้องการยังคงหมายถึงสมการ $Y = f(X)$ สำหรับกรณี Simple Regression และยังคงหมายถึง สมการ $Y = f(X's)$ **ในกรณี Multiple Regression** ไม่ให้กลับเป็น $X = f(Y)$ หรือ $X_j = f(Y)$ ทั้งนี้เพราะโดยปกติแล้วตัวแปรบางคู่มีลักษณะของสาเหตุ และผลลัพธ์ (Cansual Relationship) สู่กันและกัน ซึ่งตัวแปรใดตัวแปรหนึ่งมีสิทธิ์ที่จะเป็นได้ทั้งในฐานะของปัจจัยภายนอก หรือ ตัวแปรอิสระ (สาเหตุ) และในฐานะของตัวแปรตาม (ผลลัพธ์) เช่น ความสูง-น้ำหนัก ปริมาณความชื้น-อุณหภูมิ และอื่น ๆ

ข้อตกลงว่า $Cov(u_i X_j) = 0$ มีความหมายเป็น 2 นัยดังนี้

1. ตัวแปรอิสระที่เราละเลยหรือตัดทิ้ง ซึ่งยังคงส่งอิทธิพลต่อตัวแปร Y ในรูปของตัวแปรสุ่ม u นั้นจะต้องไม่เกี่ยวข้องผูกพันกับตัวแปรอิสระ X_j ; $j = 2, 3, \ldots, k$ ที่ระบุไว้ เพราะถ้าตัวแปรอิสระที่เราละเลยหรือตัดทิ้งเหล่านั้นมีความเกี่ยวข้องผูกพันอยู่กับ X_j ย่อมมีผลให้ u_i และ X_j มีความเกี่ยวข้องกัน

2. ตัวแปร X ต้องมีฐานะเป็นตัวแปรอิสระได้เพียงฐานะเดียว จะมีฐานะเป็นตัวแปรตามไม่ได้[1] ที่เป็นเช่นนี้เพราะเราถือว่า X ต้องเป็นตัวคงที่ (Mathematical Variable หรือ Non - Stochastic Variable) ที่สามารถควบคุมการทดลองโดยซ้ำค่า $X = X_{io}$ ได้ถึง r ครั้ง ($r \geqslant 1$) แล้ววัดค่า Y_i ผลลัพธ์ที่พบก็คือ ณ $X = X_{io}$ ค่าของ Y_i จะไม่คงที่เท่ากันทั้ง r ครั้ง แสดงว่ามี Error ในการวัดค่า Y_i เกิดขึ้นแล้ว (นอกเหนือไปจากอิทธิพลของปัจจัยภายนอกที่แฝงเร้นอยู่) ซึ่งเหตุการณ์ทำนองนี้มีผลให้ค่าของ Y_i ที่สอดคล้องกับ $X = X_{io}$ มีค่าที่ผันแปรไปได้โดยสุ่ม (เพราะ u_i เป็นตัวแปรสุ่ม และ Y_i เป็น Linear Function ของ u_i) โดยมีค่าเกาะกลุ่มอันค่อนข้างหนาแน่นที่จุด ๆ หนึ่งคือ $\mu_{y|x_{io}}$ และเมื่อเปลี่ยนค่า X_{io} ไปสู่ค่าอื่น ๆ ค่าของ Y_i ที่สอดคล้องกับ ค่า X_{io} ที่เปลี่ยนแปลงไปก็มีกลุ่มค่าใหม่กลุ่มละ r ค่าโดยแต่ละกลุ่มมีค่าเฉลี่ยเท่ากับ $\mu_{y|x_{io}}$; $i = 1, 2, \ldots, n$ โดยที่ $\mu_{y|x_{io}}$; $i = 1, 2, \ldots, n$ จะวางอยู่บนเส้นตรง $E(Y) = \alpha + \beta X$ เสมอดังภาพ 2.1 ซึ่งแสดงว่าการกำหนดเป็นข้อตกลงว่า $E(u_i X_j) = 0$; $i = 1, 2, \ldots, n$; $j = 2, 3, \ldots, k$ เป็นการบังคับให้สมการถดถอยที่ต้องการยังคงมีลักษณะของ $Y = f(X)$ หรือ $Y = f(X\text{'s})$ อย่างเดิมไม่เปลี่ยนแปลง

โดยนัยกลับกัน ถ้าหากการวัดค่า Y ไม่มี Error แต่การวัดค่า X กลับมี Error ลักษณะความสัมพันธ์จะเป็นไปในทำนองเดียวกันกับที่อธิบายมาแล้วเพียงแต่กลับ X เป็น Y กลับ Y เป็น X ความสัมพันธ์ทางคณิตศาสตร์จึงกลับกันคือ $X = f(Y)$ หรือ $X_j = f(Y)$

อย่างไรก็ตาม ในกรณีทั่ว ๆ ไปที่มิใช่กรณีของงานทดลอง แต่เป็นงานวิจัยทางสังคม หรืองานใดก็ตามที่เราไม่อาจควบคุมค่าของตัวแปรอิสระ X's ให้คงที่ได้ ซึ่งในกรณีนั้น X's จะมีค่าที่ผันผวนไปได้และมีลักษณะของตัวแปรสุ่มอย่างสมบูรณ์[2] ในกรณีเช่นนี้ข้อตกลง $E(u_i X_j) = 0$ มี

---

[1] แสดงว่าความสัมพันธ์นี้ต้องไม่ใช่ความสัมพันธ์ของแบบจำลองในรูประบบสมการ (Model of Simultaneous Equation หรือ Multi-Equation System) ซึ่งตัวแปรอิสระตัวหนึ่งตัวใดอาจเป็นได้ทั้งตัวแปรอิสระ (ปัจจัยภายนอก) หรือตัวแปรตาม (ปัจจัยภายใน) เช่น ในวิชาเศรษฐมิติ.

[2] ถ้า x เป็นตัวแปรเชิงสุ่ม เรานิยมกำหนดให้ $f_{yx}(.)$ เป็น Bivariate Normal Density และถ้า x's เป็นตัวแปรเชิงสุ่ม เรานิยมกำหนดให้ $f_{yx_2x_3\ldots x_k}(.)$ เป็น Multivariate Normal Density

ความหมายว่า ตัวแปรสุ่ม u_i และ X_j มี ฟังก์ชั่นการแจกแจงที่เป็นอิสระต่อกัน ซึ่งถ้าเป็นไปตามข้อตกลงนี้ $\hat{\beta}_j$ จะเป็น Consistent Estimator[1] แต่ถ้าหากว่า $E(u_i X_j) \neq 0$ จะมีผลให้ $\hat{\beta}_j$ เป็น Inconsistent Estimator

2.2 การประมาณค่าพารามิเตอร์ α และ β

2.2.1 หลักเกณฑ์ทั่วไป

เนื่องจากความสัมพันธ์ $Y = \alpha + \beta X + u$ เป็นความสัมพันธ์ที่แท้จริงที่แสดงถึงความผันแปรในค่าของ Y และเป้าหมายที่นักวิจัยต้องการก็คือสมการเส้นตรงจริง (True Line) $Y = \alpha + \beta X$ แต่เนื่องจากความสัมพันธ์และสมการเส้นตรงดังกล่าวเป็นสิ่งที่เราไม่อาจทราบได้ว่าวางอยู่ ณ. ตำแหน่งใดใน Space ดังนั้นสิ่งที่สามารถกระทำได้มีเพียงการคาดหมายเพื่อหาที่ตั้งของเส้นสมการดังกล่าว โดยอาศัยความรู้จากข้อมูลที่มีอยู่

ให้ $\hat{\alpha}$ และ $\hat{\beta}$ คือค่าประมาณของ α และ β โดยที่ค่าประมาณนี้กระทำโดยอาศัยข้อมูลจากค่าสังเกต $Z_i = (Y_i, X_i)$; $i = 1, 2, \ldots, n$ ดังนั้นเส้นสมการประมาณค่าที่ต้องการคือ $Y = \hat{\alpha} + \hat{\beta}X$ โดยที่ $\hat{Y}_i$ คือ ค่าประมาณของ Y_i ที่ได้รับจากสมการประมาณค่าดังกล่าวนั้นคือ $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$; $i = 1, 2, \ldots, n$ และเนื่องจากเส้นสมการจริง $E(Y) = \alpha + \beta X$ จะสามารถใช้แทนความสัมพันธ์ $Y = \alpha + \beta X + u$ ได้ดีก็ต่อเมื่อผลรวมของ Error มีค่าต่ำที่สุด กล่าวคือ Σu_i มีค่าต่ำที่สุด แต่เนื่องจาก u_i เป็นตัวแปรสุ่มที่วัดค่าไม่ได้ และเส้นสมการจริง $E(Y) = \alpha + \beta X$ เป็นสมการในอุดมคติ (Ideal Line) ซึ่งไม่อาจทราบตำแหน่งที่ตั้งได้ ด้วยเหตุนี้สิ่งที่พอจะกระทำได้ก็คือการใช้สมการคาดหมาย $Y = \hat{\alpha} + \hat{\beta}X$ แทนความสัมพันธ์ $Y = \alpha + \beta X + u$ โดยถือว่าสมการ $Y = \hat{\alpha} + \hat{\beta}X$ จะใช้แทนความสัมพันธ์ $Y = \alpha + \beta X + u$ ได้ดีและคาดหมายความสัมพันธ์จริงคือ $E(Y) = \alpha + \beta X$ ได้อย่างถูกต้อง ถ้าผลรวมของค่าประมาณของ Error คือ Σe_i มีค่าต่ำสุด เมื่อ $e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\alpha} + \hat{\beta}X_i)$; $i = 1, 2, \ldots, n$ ด้วยเหตุนี้เส้นสมการประมาณค่าที่พึงปรารถนาจึงเป็นเส้นสมการที่พัฒนาขึ้นจากแนวคิดเรื่อง e_i นี่เอง โดยถือหลักเกณฑ์ว่า "เส้นสมการประมาณค่าที่ดีที่สุด (Best Fitted Line) คือ เส้นสมการที่พุ่งผ่านไปในกลุ่มของสังเกตในลักษณะที่มีผลให้ระยะห่างระหว่างจุด (Coordinate) ของค่าสังเกตต่าง ๆ กับเส้นสมการประมาณค่ารวมกันทุกระยะมีค่าต่ำที่สุด"

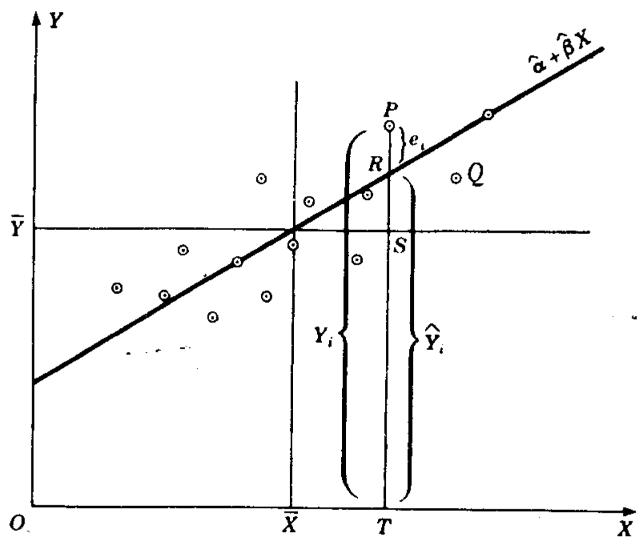
---

[1] ถ้า $\hat{\theta}$ เป็น Estimator ของ θ แล้ว $\hat{\theta}$ จะเป็น Consistent Estimator ถ้า $\lim_{n \to \infty} |\theta - \hat{\theta}| = 0$ หรือ $\lim_{n \to \infty} V(\hat{\theta}) = 0$ กล่าวคือ Error Varinance จะมีค่าลดลงเมื่อใช้กลุ่มตัวอย่างขนาดใหญ่

จากหลักเกณฑ์ประการดังกล่าวมีข้อที่นักศึกษาพึงสังเกตเป็น 3 ประการคือ

1. การพิจารณาว่าเส้นสมการประมาณค่าใดเป็นสมการที่เหมาะสมที่สุดนั้น เราพิจารณาจากระยะห่างระหว่างค่าสังเกต (สนใจเฉพาะค่า Y) จากกลุ่มตัวอย่างกับเส้นสมการประมาณค่า (Assumed Best Fitted Line) มิใช่ระยะห่างระหว่างค่าสังเกตจากกลุ่มตัวอย่างกับเส้นสมการจริง (True Line) ที่เป็นเช่นนี้เพราะเส้นสมการจริงเป็นสมการในอุดมคติที่เราไม่ทราบตำแหน่งที่ตั้งใน Space ได้ การวัดระยะห่างระหว่างค่าสังเกตกับเส้นสมการจริงจึงไม่อาจกระทำได้

2. ระยะห่างระหว่างค่าสังเกต Y_i (ถือว่าเป็นค่าจริง) กับค่าประมาณจากสมการประมาณค่า คือ $\hat{Y}_i$ คือ ระยะทางที่วัดจากจุด Coordinate ของค่าสังเกตมายังเส้นสมการประมาณค่าตามแนวที่ตั้งฉากกับแกน X กล่าวคือ $e_i = Y_i - \hat{Y}_i; i = 1, 2, \ldots, n$ ถ้าจุด Coordinate วางอยู่เหนือเส้นสมการประมาณค่าให้ถือว่า e_i มีค่าบวก ถ้าจุด Coordinate อยู่ใต้เส้นสมการประมาณค่าให้ถือว่า e_i มีค่าลบ ดังภาพ 2.4



ภาพ 2.4 แสดงระยะห่างระหว่างค่าสังเกตกับสมการประมาณค่า จากภาพจะพบว่า e_i มีค่าเป็นบวกเพราะจุด P อยู่เหนือเส้นสมการ $Y = \hat{\alpha} + \hat{\beta}X$ โดยที่ $e_i = Y_i - \hat{Y}_i$ เมื่อ $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ แต่ e_s จะมีเครื่องหมายเป็นลบเพราะจุด Q อยู่ใต้เส้นสมการประมาณค่า โดยที่ $e_s = Y_s - \hat{Y}_s$ เมื่อ $\hat{Y}_s = \hat{\alpha} + \hat{\beta}X_s$ (ขอให้สังเกตว่าไม่มีเส้นสมการจริงปรากฏในภาพ)

3. คำว่า **"เส้นสมการพุ่งผ่านไปในกลุ่มของค่าสังเกต"** หมายความว่าสมการประมาณค่าที่ดีจะต้องเป็นสมการที่อาศัยข้อสนเทศจากค่าสังเกตทุกจุดจะสนใจเฉพาะบางส่วนและละเลยบางส่วนมิได้

2.2.2 แนวคิดในการประมาณค่า α และ β

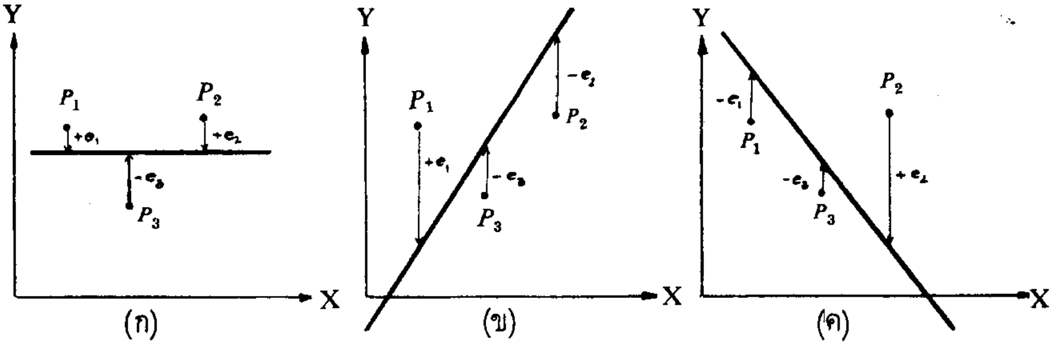
จากหลักเกณฑ์ในตอน 2.2.1 ทำให้เราสามารถประมาณค่าพารามิเตอร์ได้หลายแนวทาง แนวทางที่ขัดแย้งกับหลักเกณฑ์ดังกล่าวหรือไม่ก่อให้เกิด Unique Solution จะถือว่าเป็นแนวทางที่ไม่เหมาะสม ให้เร่งพัฒนาแนวทางแบบอื่นที่เหมาะสมกว่าต่อไป

แนวทางที่ 1 α และ β ที่มีผลให้ Σe_i มีค่าต่ำที่สุด

แนวทางนี้พัฒนามาจากแนวความคิดว่า ถ้าเส้นสมการประมาณค่าพุ่งทับไปบนจุด Coordinate ของค่าสังเกตทุกจุด ผลรวมของ Error จะมีค่าต่ำที่สุดเท่ากับ 0 เพราะถ้าเส้นสมการพุ่งทับไปบนจุด Coordinate ทุกจุด ระยะ e_i จะเท่ากับ 0 ทุกค่าซึ่งมีผลให้ $\Sigma e_i = 0$

แต่อย่างไรก็ตาม สถานการณ์ที่เส้นสมการพุ่งผ่านทับไปบนจุด Coordinate ของค่าสังเกตทุกจุดเป็นสิ่งที่ไม่ปรากฏขึ้นได้บ่อยครั้ง เหตุการณ์ที่ปรากฏขึ้นเป็นปกติก็คือเส้นสมการพุ่งทับไปบนจุด Coordinate เพียงบางส่วน และเฉียดผ่านจุด Coordinate อื่น ๆ ซึ่งในสถานการณ์เช่นนี้ หากเรายังคงยึดถือเอาหลักเกณฑ์ว่า $\Sigma e_i = 0$ เป็นแนวทางในการประมาณค่า α และ β ก็จะเป็นการเปิดทางให้นักวิจัยนำเอาเครื่องหมายประจำตัวของ e มาร่วมพิจารณาด้วย โดยนักวิจัยจะพยายามกำหนดเครื่องหมายบวกและลบให้แก่ e ในลักษณะที่ทำให้ผลรวมของ e_i หักล้างกันหมดไปสิ่งที่ตามมาก็คือเราจะได้ Fitted Model มากมายหลายสมการที่มีคุณสมบัติครบถ้วนด้วยกันทุกสมการจนไม่อาจแน่ใจได้ว่าสมการใดคือสมการที่สามารถใช้เป็นตัวแทนของ True Li... ได้วิธีนี้จึงนับว่ามีช่องโหว่อยู่ และไม่สมควรนำมาใช้เป็นหลักในการประมาณค่า α และ β

ขอให้พิจารณาภาพ 2.5 สมมุติว่ามีค่าสังเกต 3 ชุดคือ p_1, p_2 และ p_3 และ e_1, e_2, e_3 คือระยะห่างระหว่างจุด p_1, p_2 และ p_3 กับเส้นสมการประมาณค่า การกำหนดเครื่องหมายให้แก่ระยะ e_1, e_2, e_3 ในลักษณะที่มีผลให้ $\Sigma e_i = 0$ ครั้งหนึ่งจะมีผลให้ได้สมการที่เหมาะสม... หนึ่งเสมอ



ภาพ 2.5 แสดงเส้นสมการประมาณค่าที่พึงเป็นไปได้เมื่อใช้หลักเกณฑ์การประมาณค่า α และ β จากสมการ $\Sigma e_i = 0$ ขอให้สังเกตว่า ในกรณีนี้เส้นสมการประมาณค่าที่เหมาะสมที่สุดตามเงื่อนไข $\Sigma e_i = 0$ จะไม่ Unique และไม่อาจจำแนกได้อย่างชัดเจนว่าสมการใด Best Fit สมการใด Bad Fit ขอให้สังเกต ภาพ (ก) จะพบว่าสมการในภาพ (ก) เป็นสมการที่ดีที่สุด เพราะ e มีค่าต่ำ แต่วิธี $\Sigma e_i = 0$ ก็มิได้ยืนยันความจริงข้อนี้

แนวทางที่ 2 α และ β ที่มีผลให้ $\sum_i^n |e_i|$ มีค่าต่ำที่สุด (Least Absolute Residual : LAR)

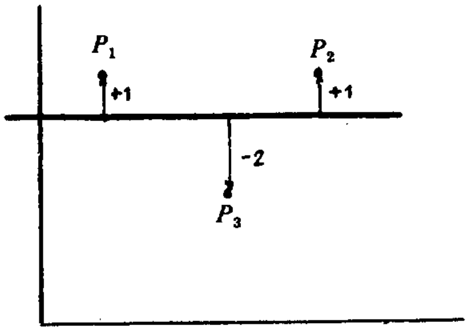
แนวทางนี้พัฒนาขึ้นมาเพื่อปกปิดช่องไหว่เรื่องการกำหนดเครื่องหมายบวกลบให้แก่ e ในแนวทางที่ 1 โดยวิธีนี้จะไม่เปิดโอกาสให้เกิด Infinitely Many Fitted Line ด้วยการนำเอาเฉพาะขนาดของ e_i มาใช้ แล้วหาทางประมาณค่า α และ β ที่มีผลให้ $\Sigma|e_i|$ มีค่าต่ำที่สุด[1]

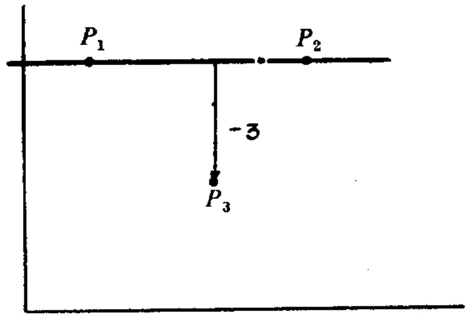
---

[1] ในปัจจุบันได้มีความพยายามที่จะพัฒนาวิธีนี้ให้สามารถใช้ได้กว้างขวางขึ้นเพื่อใช้แทนวิธี OLS และพบว่าในบางกรณีวิธี LAR ให้อำนาจพยากรณ์สูงกว่าวิธี OLS วิธีนี้มีชื่อเรียกหลายชื่อเช่น LAR, LAE (Least Absolute Error) และ MAD (Minimum Absolute Deviation) โดยปกติปัญหาเรื่องพีชคณิตมิใช่ปัญหาสำคัญ ปัญหาสำคัญก็คือ เราไม่อาจทราบ Sampling Distribution ของ $\hat{\alpha}$ และ $\hat{\beta}$ (ในกรณี Simple Regression) และ $\hat{\beta}_j$, j = 1, 2, ..., k (ในกรณี Multiple Regression) ทั้งไม่ทราบค่าของ σ^2 ได้ วิธีนี้เป็นหนทางเลือกเมื่อเกิดปัญหาขัดแย้งกับข้อตกลงว่า $u \sim N(0,\sigma^2)$ กล่าวคือเมื่อทดสอบพบว่า u ไม่แจกแจงแบบปกติ ปัญหาต่าง ๆ ในการทดสอบสมมุติฐาน และการประมาณค่าด้วยช่วงเชื่อมั่นจะติดตามมา วิธี LAR นับว่าเป็นวิธีหนึ่งที่พอจะแก้ปัญหานี้ได้ นอกจากนี้วิธี LAR ยังสามารถแก้ปัญหาข้อมูลมี Outlier อย่างได้ผล

อ่านรายละเอียดใน Maddala, G.S., *Ibid.* p.90, 305-315

วิธีนี้แม้จะแก้ปัญหาเรื่องเครื่องหมายของ e ได้แต่ก็มีช่องทางให้เกิดสมการที่ไม่เหมาะสม เพราะอาจใช้ข้อสนเทศ จากกลุ่มตัวอย่างค่าสังเกตไม่ครบทุกชุด พิจารณา ภาพ 2.6 จะเห็นว่าใน **ภาพ (ก)** $\Sigma|e_i|$ = 4 ขณะที่ภาพ (ข) $\Sigma|e_i|$ = 3 ซึ่งแสดงว่า สมการถดถอยในภาพ (ข) เหมาะสมกว่าสมการในภาพ (ก) แต่โดยข้อเท็จจริงแล้ว สมการในภาพ (ข) ขาดความเหมาะสมเพราะอาศัยข้อสนเทศจากตัวอย่างเพียง 2 ชุดคือ p_1 และ p_2 โดยตัด p_3 ออกจากการพิจารณา ซึ่งมิใช่หลักเกณฑ์ที่ถูกต้อง นอกเหนือไปจากนี้วิธีการตามแนวทางที่ 2 มักจะก่อให้เกิดความยุ่งยากในการวิเคราะห์ โดยเฉพาะในเชิงพีชคณิต





ภาพ 2.6 แสดงสมการถดถอยที่ประมาณค่าโดยอาศัยหลักเกณฑ์ "$\sum_i^n|e_i|$ มีค่าต่ำที่สุด" ภาพ (ก) มี $\Sigma|e_1|$ = 4 ขณะที่ภาพ (ข) มี $\Sigma|e_i|$ = 3 แต่สมการในภาพ (ก) ดีกว่าเพราะใช้ข้อสนเทศจากตัวอย่างทุกชุด

แนวทางที่ 3 α และ β ที่มีผลให้ Σe_i^2 มีค่าต่ำที่สุด (Ordinary Least Square : OLS)

แนวทางนี้พัฒนาขึ้นมาเพื่อขจัดปัญหาอันเป็นช่องโหว่ของแนวทางที่ 1 และแนวทางที่ 2 กล่าวคือ แก้ปัญหาเรื่องเครื่องหมายของ e โดยการยกกำลังสองแทนการใช้ค่าสัมบูรณ์ ในขณะเดียวกัน Σe_i^2 มิได้เปิดโอกาสให้มีช่องทางที่จะใช้ข้อสนเทศจากเฉพาะข้อมูลบางส่วน ดังนั้น สมการที่ได้จากวิธีนี้จึง Unique และใช้ข้อสนเทศจากตัวอย่างครบทุกชุด วิธี OLS จึงเป็นวิธีที่

ให้ผลลัพธ์ตรงตามต้องการ และเราสามารถพิสูจน์ได้ว่าตัวประมาณค่าที่ได้มาจากวิธี OLS เป็นตัวประมาณค่าที่ดีที่สุด (Best Linear Unbiased Estimator : BLUE)[1]

นอกจากแนวทางทั้ง 3 ประการดังกล่าวแล้ว ยังมีแนวทางอื่น ๆ อีก เช่น วิธี Maximum Likelihood วิธีของเบยส์ วิธีของ Aitken และอื่น ๆ อีกมาก วิธีเหล่านี้เป็นวิธีที่พัฒนาแล้วและโดยส่วนใหญ่พัฒนาขึ้นมาเพื่อใช้แทนหรือเปรียบเทียบกับวิธี OLS ซึ่งรายละเอียดเหล่านี้จะได้กล่าวถึงในโอกาสต่อไป

2.2.3 การประมาณค่า α และ β โดยวิธี OLS หรือ CLS[2]

จากสมการความสัมพันธ์จริงระหว่างตัวแปร Y และ X คือ $Y = \alpha + \beta X + u$ สมมุติว่า $\hat{\alpha}$ และ $\hat{\beta}$ คือตัวประมาณค่าของ α และ β ตามลำดับ เมื่อแทนที่ $\hat{\alpha}$ และ $\hat{\beta}$ ลงในสมการความสัมพันธ์จริง จะได้ความสัมพันธ์คาดหมาย $Y = \hat{\alpha} + \hat{\beta}X + e$ หรือ $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i + e_i$; $i = 1, 2, \ldots, n$[3] และสมการประมาณค่าที่ต้องการคือ

$$Y = \hat{\alpha} + \hat{\beta}X$$

---

[1] อาจมีผู้ถามว่าเพราะเหตุใดจึงไม่ใช้ $\sum_i^n e_i^4$ หรือ $\sum_i^n e_i^6$ หรือ $\sum_i^n e_i^r$; $r = 3, 4, \ldots$, เรื่องนี้เป็นสิ่งที่ตอบได้ยากยกเว้นจะมีการพิสูจน์ แต่สิ่งหนึ่งที่พอจะชี้ให้เห็นได้ก็คือ $\sum_i^n e_i^r$ มีวิธีการทางพีชคณิตที่ยุ่งยากกว่า $\sum_i^n e_i^2$ และเราไม่อาจแน่ใจได้ว่าตัวประมาณค่าที่ได้จะเป็น BLUE หรือไม่ทั้งไม่อาจแน่ใจได้ว่าจะมีคุณสมบัติของ Good Estimator เช่น Efficience, Suficience, Unbias และอื่น ๆ หรือไม่ เรื่องนี้ยังเป็นสิ่งที่รอการพิสูจน์อยู่

[2] OLS = Ordinary Least Square, CLS = Classical Least Square

[3] โดยทั่วไปเรานิยมเขียนสมการประมาณค่าเป็น $\hat{Y} = \hat{\alpha} + \hat{\beta}X$ และความสัมพันธ์คาดหมายเป็น $\hat{Y} = \hat{\alpha} + \hat{\beta}X + e$ การใส่เครื่องหมาย "^" บน Y มีเจตนาเพื่อชี้ให้เห็นถึงความแตกต่างระหว่างสมการจริงกับ สมการประมาณค่าและความสัมพันธ์จริงกับความสัมพันธ์คาดหมาย แต่ผู้เขียนจะไม่ใช้วิธีการดังกล่าว ในทางปฏิบัติเราไม่นิยมใส่เครื่องหมาย "^" เหนือตัวแปร Y เช่นเรานิยมเขียน $Y = 2 + 1.28X$ มากกว่าที่จะใช้ $\hat{Y} = 2 + 1.28X$ แต่ถ้าเป็นค่าเดี่ยว ๆ ของ Y คือ Y_i ผู้เขียนจะใช้ $\hat{Y}_i$ สำหรับแทนค่าพยากรณ์ของ Y_i (ทั้งโดยนัยของ Interpolation และ Extrapolation) ขอให้นักศึกษาเข้าใจตามนี้ด้วย

ให้ e_i = ความคลาดเคลื่อนระหว่างค่าสังเกตของ Y_i กับค่าประมาณของ Y_i คือ $\hat{Y}_i$ ที่ประมาณค่าจากสมการ $Y = \hat{\alpha} + \hat{\beta}X$ กล่าวคือ

$$e_i = Y_i - \hat{Y}_i ; i = 1, 2, \ldots, n$$

พิจารณา $e_i = Y_i - \hat{Y}_i$ จะพบว่า

$$e_i = Y_i - \hat{Y}_i$$

$$= Y_i - \hat{\alpha} - \hat{\beta}X_i ; i = 1, 2, \ldots, n$$

$$\text{ดังนั้น } \sum_i^n e_i^2 = \sum_i^n (Y_i - \hat{\alpha} - \hat{\beta}X_i)^2$$

และ $\hat{\alpha}$ และ $\hat{\beta}$ ที่มีผลให้ $\sum^n e_i^2$ มีค่าต่ำที่สุดคือ $\hat{\alpha}$ และ $\hat{\beta}$ ที่เป็น Solution ของระบบสมการซึ่งเรียกว่า Normal Equation ดังนี้

$$\frac{\partial}{\partial\hat{\alpha}} \Sigma e_i^2 = 0 , \quad \frac{\partial}{\partial\hat{\beta}} \Sigma e_i^2 = 0$$

$$-2 \sum_i^n (Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0 \text{ หรือ } n\hat{\alpha} - \hat{\beta}\sum_i^n X_i = \sum_i^n Y_i \qquad \ldots\ldots\ldots\ (1)$$

$$-2X_i \sum_i^n (Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0 \text{ หรือ } \hat{\alpha} \sum_i^n X_i - \hat{\beta} \sum_i^n X_i^2 = \sum_i^n X_i Y_i \qquad \ldots\ldots\ldots\ (2)$$

โดยอาศัยเทคนิคการแก้สมการวิธีใดวิธีหนึ่ง เช่นการแทนค่า (Substitution) การขจัดตัวแปร (Systematic Elimination) การใช้ Cramer's Rule หรือการใช้ส่วนกลับของ Matrix จะพบว่า

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$\hat{\beta} = \sum_i^n (X_i - \bar{X})(Y_i - \bar{Y}) / \sum_i^n (X_i - \bar{X})^2 = \sum_i^n x_i y_i / \sum_i^n x_i^2$$

แสดงว่าค่าประมาณของระยะจุดตัดบนแกน Y ของเส้นตรงจริงคือ $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$ และค่าประมาณของความชันคือ $\hat{\beta} = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$ และสมการ $Y = \hat{\alpha} + \hat{\beta}X$ คือสมการที่สามารถใช้ประมาณค่าสมการจริงโดยอาศัยข้อสนเทศ จากค่าสังเกต Z ทั้งสิ้น n คือ $Z_i = (Y_i, X_i) ; i = 1, 2, \ldots, n$

อย่างไรก็ตามในบางโอกาสเราอาจมีความจำเป็นต้องลดขนาดข้อมูลเดิมให้เล็กลงเพื่อให้สามารถวิเคราะห์ได้ง่ายขึ้น วิธีการที่สามารถสนองความต้องการดังกล่าวคือการแปลงข้อมูลจากข้อมูลเดิม (Initial Data หรือ Deviation from Zero) เป็นข้อมูลให้ที่มีลักษณะของ Deviation from Mean กล่าวคือ แปลงจาก Y_i เป็น $y_i = Y_i - \bar{Y}$ และแปลงจาก X_i เป็น $x_i = X_i - \bar{X}$ ดังวิธีการต่อไปนี้

จากสมการ $Y_i = \alpha + \beta X_i + u_i$; $i = 1,2,\ldots,n$ .....(1)

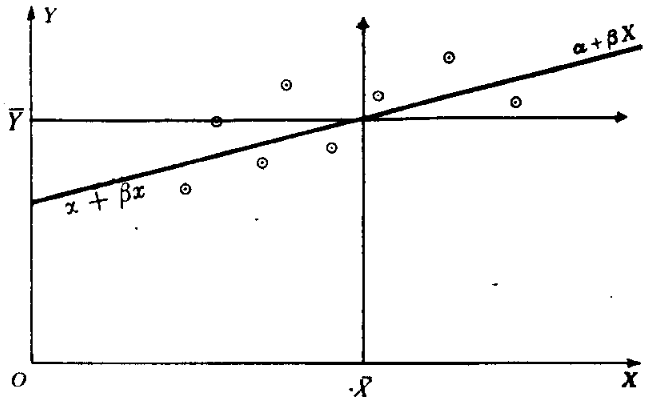
รวม (1) ในทุกค่าของ i แล้วหารตลอดด้วย n

$$\bar{Y} = \alpha + \beta\bar{X} + \bar{u} \quad .....(2)$$

(1)-(2) $$Y_i - \bar{Y} = \beta(X_i - \bar{X}) + u_i - \bar{u}$$ [1]

หรือ $$y_i = \beta x_i + (u_i - \bar{u}) \quad .....(3)$$

สมการที่ (3) คือ True Relationship เมื่อข้อมูลได้รับการแปลงรูปสู่ลักษณะของ Deviation from Mean ซึ่งในทางคณิตสาสตร์ สมการ (1) และ (3) ก็คือสมการเดียวกัน ต่างกันเพียงสมการ (3) คือสมการที่เทียบต่อแกนใหม่ที่ย้ายแกนมาสู่ระบบที่มีจุดกำเนิดเป็น $(\bar{X}, \bar{Y})$ ดังภาพ



ดังนั้น e−ē จึงเป็นค่าประมาณของ u−ū แต่ ē = o [2] ดังนั้น $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}x_i$

$$\text{ดังนั้น } \sum_i^n e_i^2 = \sum_i^n (y_i - \hat{\beta}x_i)^2$$

$$\frac{\partial}{\partial\hat{\beta}} \sum_i^n e_i^2 = 0 \quad \Rightarrow \quad -2x_i \sum_i^n (y_i - \hat{\beta}x_i) = 0 \text{ นั่นคือ } \hat{\beta} = \sum_i^n x_i y_i \Big/ \sum_i^n x_i^2$$

---

[1] $\bar{u} \neq 0$ แต่ $E(u_i) = 0$ และ $E(\bar{u}) = 0$ ขอให้เข้าใจว่า $\bar{u} = \frac{1}{n}\sum_i^n u_i$ เป็น Sample Mean มิใช่ Population Mean

[2] จาก $\frac{\partial}{\partial\alpha} \sum e_i^2 = 0 \Rightarrow -2\sum_i^n (Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0$ แสดงว่า $\sum_i^n (Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0$ หรือ $\sum_i^n e_i = 0$ นั่นคือ $\frac{1}{n}\sum_i^n e_i = 0$ หรือ $\bar{e} = 0$ แสดงว่า $e_i - \bar{e} = e_i$ คือตัวประมาณค่าของ u−ū

ดังนั้นสมการประมาณค่าในกรณี Deviation from Mean คือ

$$y = \hat{\beta}x$$

ข้อสังเกต

สำหรับสมการประมาณค่าที่ประมาณค่าพารามิเตอร์จากข้อมูลที่แปลงรูปในลักษณะของ Deviation from Mean นั้นมีข้อสังเกตดังนี้

1. การประมาณค่าพารามิเตอร์กระทำโดยประมาณเพียงเฉพาะ β เท่านั้น

2. สมการประมาณค่าคือ $y = \hat{\beta}x$ มิได้หมายความว่า $\hat{\alpha}$ หายไปจากสมการ ความจริง $\hat{\alpha}$ ยังคงปรากฏอยู่เพียงแต่แฝงอยู่ในรูปของ X และ Y ถ้าเราย้อนสมการ $y = \hat{\beta}x$ สู่รูปของข้อมูลเดิม $\hat{\alpha}$ ก็จะปรากฏให้เห็นดังนี้

$$\begin{aligned} \text{จาก} \quad y &= \hat{\beta}x \quad \text{หรือ } y_i = \hat{\beta}x_i \\ \Longrightarrow \quad Y_i - \bar{Y} &= \hat{\beta}(X_i - \bar{X}) \\ Y_i &= (\bar{Y} - \hat{\beta}\bar{X}) + \hat{\beta}X_i \\ \text{นั่นคือ} \quad Y_i &= \hat{\alpha} + \hat{\beta}X_i \end{aligned}$$

2.3 คุณสมบัติของ OLS-Esimator

2.3.1 $\hat{\alpha}$ และ $\hat{\beta}$ เป็นตัวประมาณค่าที่ปราศจากอคติ (Unbiasedness)

ก. $E(\hat{\beta}) = \beta$

พิจารณา $\hat{\beta} = \dfrac{\sum_i^n x_i y_i}{\sum_i^n x_i^2} = \sum_i^n \left(\dfrac{x_i}{\sum x_i^2}\right) y_i$

ให้ $w_i = \dfrac{x_i}{\sum_i^n x_i^2}$; $i = 1, 2, \ldots, n$

ดังนั้น $\hat{\beta} = \sum_i^n w_i y_i$

$$= \sum_i^n w_i y_i^{1} = \sum_i^n w_i(\alpha+\beta X_i+u_i)$$

$$= \alpha\sum_i^n w_i+\beta \sum_i^n w_iX_i+\sum_i^n w_iu_i$$

$$= \beta\sum_i^n w_iX_i+\sum_i^n w_iu_i$$

$$= \beta\sum_i^n w_ix_i+\sum_i^n w_iu_i^{2}$$

$$= \beta+\sum_i^n w_iu_i^{3}$$

ดังนั้น $E(\hat{\beta}) = \beta+\sum_i^n w_iE(u_i) = \beta$

แสดงว่า $\hat{\beta} = \dfrac{\sum_i^n x_iy_i}{\sum_i^n x_i^2}$ เป็นตัวประมาณค่าปราศจากอคติของ β

ข. $E(\hat{\alpha}) = \alpha$ (ให้พิสูจน์เองเป็นแบบฝึกหัด)

2.3.2 $\hat{\alpha}$ และ $\hat{\beta}$ มีความแปรปรวนต่ำที่สุด (Minimum Variance) หรืออีกนัยหนึ่ง $\hat{\alpha}$ และ $\hat{\beta}$ เป็นตัวประมาณค่าที่ดีที่สุด (Best Estimator)

ก. $V(\hat{\beta}) = ?$

พิจารณา $\hat{\beta}$ พบว่า $\hat{\beta} = \beta+\sum_i^n w_iu_i$ โดยที่ $w_i = \dfrac{x_i}{\sum_i^n x_i^2}$

---

[1] $\sum_i^n w_iy_i = \sum_i^n w_i(Y_i-\bar{Y}) = \sum_i^n w_iY_i-\bar{Y}\sum_i^n w_i$

แต่ $\sum_i^n w_i = \sum_i^n \left(\frac{x_i}{\Sigma x_i^2}\right) = \left(\frac{1}{\Sigma x_i^2}\right)\sum_i^n x_i = \left(\frac{1}{\Sigma x_i^2}\right)\sum_i^n(X_i-\bar{X}) = 0$

ดังนั้น $\sum_i^n w_iy_i = \sum_i^n w_iY_i$ ซึ่งแสดงว่า $\hat{\beta} = \sum_i^n w_iY_i$ คือ Linear Function ของตัวแปรสุ่ม Y_i

[2] $\sum_i^n w_ix_i = \sum_i^n w_i(X_i-\bar{X}) = \sum_i^n w_iX_i-\bar{X}\sum_i^n w_i = \sum_i^n w_iX_i$

[3] $\sum_i^n w_ix_i = \sum_i^n \left(\frac{x_i}{\Sigma x_i^2}\right)x_i = \sum_i^n \left(\frac{x_i^2}{\Sigma x_i^2}\right) = \left(\frac{1}{\Sigma x_i^2}\right)\sum_i^n x_i^2 = 1$

ดังนั้น $V(\hat{\beta}) = E(\hat{\beta}-\beta)^2 = E(\beta+\sum_i^n w_iu_i-\beta)^2 = E(\sum_i^n w_iu_i)^2$

$$= E(\sum_i^n w_i^2u_i^2+\sum_i^n\sum_{\substack{j\\i\neq j}}^n w_iw_ju_iu_j)$$

$$= \sum_i^n w_i^2E(u_i^2)+\sum_i^n\sum_{\substack{j\\i\neq j}}^n w_iw_jE(u_iu_j)$$

ตามข้อตกลงที่ 3 และ 4 ระบุว่า $E(u_iu_j) = 0$ และ $E(u_i^2) = \sigma^2$

ดังนั้น $V(\hat{\beta}) = \sigma^2\sum_i^n w_i^2 = \sigma^2\sum_i^n \left(\frac{x_i^2}{\sum_i^n x_i^2}\right)^2 = \sigma^2 \frac{1}{\left(\sum_i^n x_i^2\right)^2}\cdot\sum_i^n x_i^2$

นั่นคือ $V(\hat{\beta}) = \frac{\sigma^2}{\Sigma x_i^2}$

ให้ $\hat{\beta}^*$ เป็นตัวประมาณค่าที่ปราศจากอคติอีกตัวหนึ่งของ β กล่าวคือ ให้ $\hat{\beta}^* = \sum_i^n d_iY_i = \sum_i^n (w_i+c_i)Y_i$ ซึ่งเราจะพบว่า[1]

$$\hat{\beta}^* = \sum_i^n (w_i+c_i)Y_i = \sum_i^n d_i(\alpha+\beta X_i+u_i)$$

$$= \alpha\sum_i^n d_i+\beta\sum_i^n d_iX_i+\sum_i^n d_iu_i$$

ดังนั้น $E(\hat{\beta}^*) = \alpha\sum_i^n d_i+\beta\sum_i^n d_iX_i \neq \beta\ ;\ E\sum_i^n d_iu_i = \sum_i^n d_iE(u_i) = 0$

แต่เรากำหนดให้ $\hat{\beta}^*$ เป็น Unbiased Estimator ของ β ดังนั้น $\hat{\beta}^*$ จะเป็น Unbiased Estimator ของ β ได้ก็ต่อเมื่อ $\sum_i^n d_i = 0$ และ $\sum_i^n d_iX_i = 1$ แสดงว่า $\sum_i^n d_i = 0$ และ $\sum_i^n d_iX_i = 1$ คือเงื่อนไขที่มีผลให้ $\hat{\beta}^*$ เป็น Unbiased Estimator

จากเงื่อนไขว่า $\Sigma d_i = 0$ แสดงว่า $\sum_i^n (w_i+c_i) = 0 \Rightarrow \sum_i^n c_i = 0$

---

[1] เดิม $\beta = \sum_i^n w_iy_i = \sum_i^n w_iY_i$ เมื่อ w_i คือตัวถ่วงน้ำหนักของ y_i หรือ Y_i การหาค่าประมาณตัวใหม่ของ β สามารถกระทำได้โดยง่ายเพียงแต่เปลี่ยนตัวถ่วงน้ำหนักที่กำหนดให้กับ y_i หรือ Y_i เสียใหม่ ในที่นี้ตัวถ่วงน้ำหนักที่กำหนดขึ้นใหม่คือ $d_i = (w_i+c_i)$

และจากเงื่อนไข$\Sigma d_i X_i = 1$ แสดงว่า $\sum_i^n (w_i + c_i)X_i = 1 \Rightarrow \sum_i^n c_i X_i = 0$[1]

แสดงว่า $\hat{\beta}$ จะเป็น Unbiased Estimator ของ β ได้เฉพาะเมื่อ $\sum_i^n c_i = 0$ และ $\sum_i^n c_i X_i = 0$ เท่านั้น

พิจารณา $V(\hat{\beta}^*)$ จะพบว่า[2]

$$V(\hat{\beta}^*) = E(\hat{\beta}^* - \beta)^2$$

$$= E(\alpha\sum_i^n d_i + \beta\sum_i^n d_i X_i + \sum_i^n d_i u_i - \beta)^2$$

แต่ $\because \sum_i^n d_i = 0$ และ $\sum_i^n d_i X_i = 1$ ดังนั้น

$$V(\hat{\beta}^*) = E(0 + \beta + \sum_i^n d_i u_i - \beta)^2 = E(\sum_i^n d_i u_i)^2$$

$$= E(\sum_i^n d_i^2 u_i^2 + \sum_{\substack{i \\ i \neq j}}^n \sum_j^n d_i d_j u_i u_j)$$

$$= \sigma^2 \sum_i^n d_i^2$$

$$= \sigma^2 \sum_i^n (w_i + c_i)^2$$

$$= \sigma^2 \sum_i^n w_i^2 + \sigma^2 \sum_i^n c_i^2 + 2\sigma^2 \sum_i^n w_i c_i$$

$$= \frac{\sigma^2}{\Sigma x_i^2} + \sigma^2 \sum_i^n c_i^2 + 0^3$$

นั่นคือ $V(\hat{\beta}^*) = \dfrac{\sigma^2}{\Sigma x_i^2} + \sigma^2 \sum_i^n c_i^2 = V(\hat{\beta}) + \sigma^2 \sum_i^n c_i^2$

ดังนั้น $V(\hat{\beta}^*) - V(\hat{\beta}) = \sigma^2 \sum_i^n c_i^2$

---

1 $\because \sum_i^n w_i = 0$ และ $\sum_i^n w_i X_i = \sum_i^n w_i x_i = 1$

2 ใช้รูป $V(\hat{\beta}^*) = E(\hat{\beta}^* - \beta)^2$ ได้เพราะ $E(\hat{\beta}^*) = \beta$ ซึ่ง $E(\hat{\beta}^*) = \beta$ ได้เฉพาะเมื่อ $\sum_i^n d_i = 0$ และ $\sum_i^n d_i X_i = 1$

3 $\sum_i^n w_i c_i = \sum_i^n \left(\dfrac{x_i}{\sum_i^n x_i^2}\right) c_i = \left(\dfrac{1}{\sum_i^n x_i^2}\right) \sum_i^n x_i c_i = \left(\dfrac{1}{\sum_i^n x_i^2}\right)\left(\sum_i^n c_i X_i - \bar{X} \sum_i^n c_i\right)$ แต่เนื่องจาก $\sum_i^n c_i = 0$ และ $\sum_i^n c_i X_i = 0$

ดังนั้น $\sum_i^n w_i c_i = 0$

พิจารณาผลต่างระหว่าง $V(\hat{\beta}^*)$ กับ $V(\hat{\beta})$ คือ $\sigma^2\sum_i^n c_i^2$ จะพบว่า $c_i^2 > 0$ ดังนั้น $\sigma^2\sum_i^n c_i^2$ จึงเป็นปริมาณบวก จึงเป็นการยืนยันให้เห็นได้ว่า $V(\hat{\beta}^*) \geqslant V(\hat{\beta})$ ซึ่งแสดงว่า $\hat{\beta}$ ซึ่งประมาณได้จากวิธี OLS จะมีความแปรปรวนต่ำกว่า $\hat{\beta}^*$ ซึ่งประมาณค่าโดยวิธีอื่น ๆ หรือนัยหนึ่ง $\hat{\beta}_{OLS}$ เป็นตัวประมาณค่าที่ดีที่สุด (Best Estimator)

ข. $V(\hat{\alpha}) = ?$[1]

$$
\begin{aligned}
V(\hat{\alpha}) &= E(\hat{\alpha}-\alpha)^2 = E(\bar{Y}-\hat{\beta}\bar{X}-\alpha)^2 \\
&= E\left(\frac{1}{n}\sum_i^n Y_i - \hat{\beta}\bar{X} - \alpha\right)^2 \\
&= E\left\{\frac{1}{n}\sum_i^n(\alpha+\beta X_i+u_i) - \hat{\beta}\bar{X} - \alpha\right\}^2 \\
&= E\left\{\alpha+\beta\bar{X}+\frac{1}{n}\sum_i^n u_i - \hat{\beta}\bar{X} - \alpha\right\}^2 \\
&= E\left\{\frac{1}{n}\sum_i^n u_i - \bar{X}(\hat{\beta}-\beta)\right\}^2
\end{aligned}
$$

แต่เนื่องจาก $\hat{\beta} = \beta + \sum_i^n w_i u_i$ ดังนั้น

$$
\begin{aligned}
V(\hat{\alpha}) &= E\left\{\frac{1}{n}\sum_i^n u_i - \bar{X}(\beta+\sum_i^n w_i u_i - \beta)\right\}^2 \\
&= E\left\{\frac{1}{n}\sum_i^n u_i - \bar{X}\sum_i^n w_i u_i\right\}^2 \\
&= E\left\{\sum_i^n\left(\frac{1}{n}-\bar{X}w_i\right)u_i\right\}^2 \\
&= E\left\{\sum_i^n\left(\frac{1}{n}-\bar{X}w_i\right)^2 u_i^2 + \sum_i^n\sum_{\substack{j \\ i\neq j}}^n\left(\frac{1}{n}-\bar{X}w_i\right)\left(\frac{1}{n}-\bar{X}w_j\right)u_iu_j\right\} \\
&= \sum_i^n\left(\frac{1}{n}-\bar{X}w_i\right)^2 E(u_i^2) + 0 \\
&= \sigma^2\sum_i^n\left(\frac{1}{n}-\bar{X}w_i\right)^2 \\
&= \sigma^2\sum_i^n\left(\frac{1}{n^2}+\bar{X}^2w_i^2-\frac{2\bar{X}w_i}{n}\right)
\end{aligned}
$$

---

[1] $\hat{\alpha} = \bar{Y}-\hat{\beta}\bar{X} = \bar{Y}-\bar{X}\sum_i^n w_iY_i = \frac{1}{n}\sum_i^n Y_i - \sum_i^n \bar{X}w_iY_i = \sum_i^n\left(\frac{1}{n}-\bar{X}w_i\right)Y_i$ แสดงว่า $\hat{\alpha}$ คือ Linear Function ของตัวแปรสุ่ม Y_i

$$= \frac{\sigma^2}{n} + \frac{\sigma^2\bar{X}^2}{\sum_i^n x_i^2} \quad ^1$$

$$= \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_i^n x_i^2} \right)$$

$$= \sigma^2 \left(\frac{\sum_i^n x_i^2 + n\bar{X}^2}{n\sum_i^n x_i^2} \right)$$

$$= \sigma^2 \left(\frac{\sum_i^n X_i^2 - n\bar{X}^2 + n\bar{X}^2}{n\sum^n x_i^2} \right) \quad ^2$$

$$= \sigma^2 \frac{\sum_i^n X_i^2}{n\sum^n x_i^2}$$

นั่นคือ $$V(\hat{\alpha}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_i^n x_i^2} \right) \text{ หรือ } \sigma^2 \frac{\sum_i^n X_i^2}{n\sum_i^n x_i^2}$$

ในทำนองเดียวกันเราสามารถพิสูจน์ได้ว่า $\hat{\alpha}_{OLS}$ เป็นตัวประมาณค่าที่ดีที่สุดของ α (ขอเว้นไว้เป็นแบบฝึกหัด)

ผลการพิสูจน์ในตอน 2.3.1 และ 2.3.2 ทำให้สามารถสรุปได้ว่า OLS-Estimator เป็น Best Linear Unbiased Estimator (BLUE)[3] ของ Regression Parameter (α, β)

สำหรับ Cov ($\hat{\alpha}$, $\hat{\beta}$) ซึ่งแสดงถึงความสัมพันธ์ระหว่าง Regression Coefficient และโดยปกติ Cov ($\hat{\alpha}$, $\hat{\beta}$) $\neq 0$ ซึ่งแสดงว่าจะต้องมี Joint Confidence Region ระหว่าง Regression Coefficient

---

[1]
(1) $\sum_i^n \frac{1}{n^2} = \frac{n}{n^2} = \frac{1}{n}$

(2) $\sum_i^n \bar{X}^2 w_i^2 = \bar{X}^2 \sum_i^n w_i^2 = \bar{X}^2 \sum_i^n \left(\frac{x_i}{\sum x_i^2}\right)^2 = \bar{X}^2 \frac{\sum x_i^2}{(\sum x_i^2)^2} = \bar{X}^2 \left(\frac{1}{\sum x_i^2}\right)$

(3) $\sum_i^n \bar{X} w_i = \bar{X} \sum_i^n \left(\frac{x_i}{\sum x_i^2}\right) = \bar{X} \left(\frac{1}{\sum x_i^2}\right) \sum_i^n x_i = 0$

[2] $\sum x_i^2 = \sum_i^n (X_i - \bar{X})^2 = \sum_i^n X_i^2 - n\bar{X}^2$

[3] ทฤษฎี Gauss-Markov "ในกลุ่มของ Linear Unbiased Estimator ของ α และ β นั้น OLS-Estimator **จะเป็นตัวประมาณค่าที่ดีที่สุด**"

เสมอนั้นเราสามารถพิสูจน์ให้เห็นโครงสร้างได้ดังนี้

$$\begin{aligned}
\mathrm{Cov}(\hat{\alpha}, \hat{\beta}) &= E(\hat{\alpha} - \alpha)(\hat{\beta} - \beta) \\
&= E\left\{ \sum_i^n \left(\frac{1}{n} - \bar{X}w_i \right)u_i \right\}\left\{ \sum_i^n w_i u_i \right\} \\
&= E\left\{ \sum_i^n \left(\frac{1}{n} - \bar{X}w_i \right)w_i u_i^2 + \sum_i^n \sum_{\substack{j \\ i \neq j}}^n \left(\frac{1}{n} - \bar{X}w_i \right)w_i \left(\frac{1}{n} - \bar{X}w_j \right)w_j u_i u_j \right\} \\
&= \sigma^2 \sum_i^n \left(\frac{1}{n} - \bar{X}w_i \right)w_i \\
&= \sigma^2 \left(\sum_i^n \frac{w_i}{n} - \bar{X}\sum_i^n w_i^2 \right) = \frac{-\bar{X}}{\sum_i^n x_i^2}\,\sigma^2
\end{aligned}$$

ข้อสังเกต

ในที่นี้มีข้อที่ควรที่จะชี้ให้เห็นเป็นที่สังเกตเกี่ยวกับความสำคัญของ BLUE ไว้เป็น 3 ประการดังนี้

1. ในบรรดาตัวประมาณค่าทั้งหลายนั้นเราสามารถพิสูจน์ให้เห็นโดยง่ายว่า Linear Estimator เป็นตัวประมาณค่าที่น่าพึงพอใจหลายประการ โดยเฉพาะอย่างยิ่งคือนำไปใช้ในทางปฏิบัติได้ง่ายกว่าตัวประมาณค่าแบบอื่น นอกจากนี้การหา Sampling Distribution จะโดยอาศัย mgf-Technique, cdf-Technique หรือ Transformation Technique ก็กระทำได้โดยง่าย

ในกรณีของ $\hat{\alpha}$ และ $\hat{\beta}$ ซึ่งเป็น Linear Function ของตัวแปรสุ่ม Y_i (หรือ $\hat{\alpha}$ และ $\hat{\beta}$ เป็น Linear Combination ของตัวแปรสุ่ม Y_i) เราทราบว่า $Y_i \sim N(\alpha + \beta X_i, \sigma^2)$ และ

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = \sum_i^n \left(\frac{1}{n} - \bar{X}w_i \right)Y_i$$

$$\text{และ} \quad \hat{\beta} = \sum_i^n w_i Y_i$$

โดยอาศัย mgf-Technique เราสามารถพิสูจน์เพื่อหา Sampling Distribution ของ $\hat{\alpha}$ และ $\hat{\beta}$ ได้โดยง่าย กล่าวคือเนื่องจาก $Y_i \sim N(\alpha + \beta X_i, \sigma^2)$

$$\text{ดังนั้น} \qquad M_{Y_i}(t) = e^{(\alpha+\beta X_i)t + \sigma^2 \frac{t^2}{2}}$$

$$\begin{aligned}
\therefore M_{\hat{\alpha}}(t) &= \prod_i^n M_{Y_i}\left(\left(\frac{1}{n} - \bar{X}w_i \right)t \right) \\
&= \prod_i^n \left\{ e^{(\alpha+\beta X_i)\left(\frac{1}{n} - \bar{X}w_i \right)t + \sigma^2 \left(\frac{1}{n} - \bar{X}w_i \right)^2 \frac{t^2}{2}} \right\}
\end{aligned}$$

$$= e^{\alpha t+\sigma^2(\Sigma X_i^2 / n\Sigma x_i^2)t^2/2}$$

แสดงว่า $\hat{\alpha} \sim N(\alpha, \sigma^2 \frac{\Sigma X_i^2}{n\Sigma x_i^2})$

และ $$M_{\hat{\beta}}(t) = \prod_{i}^{n} M_{Y_i}(w_i t)$$

$$= \prod_{i}^{n} \{ e^{(\alpha+\beta X_i)w_i t+\sigma^2 w_i^2 t^2/2} \}$$

$$= e^{\beta t+(\sigma^2/\Sigma x_i^2)t^2/2}$$

แสดงว่า $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\Sigma x_i^2})$

2. คำว่าปราศจากอคติ (Unbiasedness) หมายความว่าโดยถัวเฉลี่ยแล้วค่าของตัวประมาณค่า (คิดจาก All Possible Sample of Size n) จะมีค่าเท่ากับพารามิเตอร์ของประชากรกลุ่มที่สนใจนั้น ซึ่งแสดงว่าถ้าพิจารณาเฉพาะตัวอย่างชุดหนึ่งชุดใดเพียงชุดเดียว (Single Sample) เราย่อมไม่อาจแน่ใจได้ว่าค่าประมาณที่ได้รับจากกลุ่มตัวอย่างชุดดังกล่าวจะมีค่าเท่ากับพารามิเตอร์ที่สนใจหรือไม่ ความจริงเรื่องนี้จึงยืนยันให้เห็นว่า Unbiasedness เพียงประการเดียวมิได้ให้ประโยชน์ต่อการอนุมานทางสถิติมากนัก เพราะอาจเป็นไปได้ที่ตัวประมาณค่าที่ปราศจากอคติดังกล่าวมี Error Variance สูงมากซึ่งมีผลให้ค่าประมาณขาดความน่าเชื่อถือ ในกรณีเช่นนี้ Biased Estimator ที่มี Mean Square Error ต่ำมากจะเป็นตัวประมาณค่าที่เหมาะสมกว่า[1] ด้วยเหตุผลดังกล่าวคุณสมบัติ Unbiaseness จึงต้องผสมผสานกับคุณสมบัติอื่นโดยเฉพาะอย่างยิ่งคือ Minimum Variance (Best Estimator) จึงจะได้ตัวประมาณค่าที่น่าพึงพอใจ

คุณสมบัติ Minimum Variance (Best) คือคุณสมบัติที่ชี้ให้เห็นว่าในบรรดาตัวประมาณค่าทั้งหลายของพารามิเตอร์ที่สนใจนั้น ตัวประมาณค่าที่ดีที่สุดคือตัวประมาณค่าที่ให้ Error Variance ต่ำที่สุด คุณสมบัติข้อนี้ถ้าพิจารณาเฉพาะตัวก็นับว่าควรจะน่าพึงปรารถนา แต่โดยข้อเท็จจริงแล้วยังมีช่องโหว่เพราะตัวประมาณค่าที่มี Minimum Variance อาจใช้ประมาณค่าพารามิเตอร์ได้ไม่ถูกต้องเพราะเป็น Biased Estimator ดังนั้นคุณสมบัติข้อนี้จึงควรผสมกับคุณสมบัติอื่นโดยเฉพาะ Unbiasedness จึงจะเป็นคุณสมบัติร่วมของตัวประมาณค่าที่น่าพึงพอใจ

3. OLS-Estimator เป็นตัวประมาณค่าที่มีคุณสมบัติสอดคล้องกับคุณสมบัติร่วมคือเป็นทั้ง Linear Estimator เป็น Best Estimator และเป็น Unbiased Estimator ด้วยเหตุนี้ OLS-

---

[1] เช่น Ridge Estimator ซึ่งเป็น Biased Estimator แต่มี Mean Square Error ต่ำกว่า OLS-Estimator

Estimator **จึงมีคุณสมบัติที่น่าพึงพอใจ ทั้งยังสอดคล้องกับ Guass-Markov Theorem อีกด้วย**

ตัวอย่างที่ 2.1 **(Guass-Markov** Theorem) **ให้** $\hat{\alpha} = \sum_i^n c_iY_i$ **โดยที่** $Y_i = \alpha+\beta X_i+u_i$ **จงอาศัย** **Lagrange Multiplier Method คำนวณหาค่าของตัวถ่วงน้ำหนัก** c_i **ที่มีผลให้** $\hat{\alpha}$ **เป็น BLUE**

วิธีทำ พิจารณา Linear Estimator ของ α คือ $\hat{\alpha} = c_1Y_1+c_2Y_2+\ldots+c_nY_n = \sum^n c_iY_i$ จะพบว่า

$$E(\hat{\alpha}) = E(\sum_i^n c_iY_i) = E\{\sum^n c_i(\alpha+\beta X_i+u_i)\} = \alpha\sum_i^n c_i+\beta\sum_i^n c_iX_i$$

แสดงว่า $\hat{\alpha}$ จะเป็น Unbiased Estimator ได้ก็ต่อเมื่อ $\sum^n c_i = 1$ และ $\sum_i^n c_iX_i = 0$ หรือนัยหนึ่ง $\sum^n c_i = 1$ และ $\sum_i^n c_iX_i = 0$ คือเงื่อนไข (Constraint Condition) ที่ทำให้ $\hat{\alpha}$ เป็น Unbiased Estimator

ปัญหาก็คือ $c_i = ?$; i = 1, 2,...,n

พิจารณา $V(\hat{\alpha})$ จะพบว่า $V(\hat{\alpha}) = V(\sum_i^n c_iY_i) = \sum^n c_i^2V(Y_i) = \sigma^2\sum^n c_i^2$

ให้ λ_1 และ λ_2 เป็น Lagrange Multiplier

$$\text{ดังนั้น}\quad F = \sigma^2\sum_i^n c_i^2 - 2\lambda_1(\sum_i^n c_i - 1) - 2\lambda_2\sum_i^n c_iX_i$$ [1]

ดังนี้ $\hat{\alpha}$ จะมี **Variance ต่ำที่สุด (Best) และเป็น Unbiased Estimator** ก็ต่อเมื่อ c_i มีค่าสอดคล้องกับระบบสมการ $\frac{\partial F}{\partial c_i} = 0$, $\frac{\partial F}{\partial \lambda_1} = 0$, $\frac{\partial F}{\partial \lambda_2} = 0$ และพบว่า

$$\frac{\partial F}{\partial c_i} = 0 \Rightarrow 2\sigma^2c_i - 2\lambda_1 - 2\lambda_2X_i = 0\ ;\ i = 1,2,\ldots,n \qquad \ldots\ldots\ldots(1)$$

$$\frac{\partial F}{\partial \lambda_1} = 0 \Rightarrow -2(\sum^n c_i - 1) = 0 \text{ หรือ } \sum_i^n c_i = 1 \qquad \ldots\ldots\ldots(2)$$

---

[1] โดยปกติเราจะเสนอ F ดังนี้คือ $F = \sigma^2\sum_i^n c_i^2+\lambda_1(\sum_i^n c_i-1)+\lambda_2\sum_i^n c_iX_i$ แต่เมื่อทดลองดิฟเฟอเรนชิเอทเทียบต่อ c_i พบว่าเทอมแรกคือ $2\sigma^2\sum_i^n c_i$ ส่วนเทอมอื่น ๆ ไม่มี 2 คูณ ทั้งยังมีเครื่องหมายบวกทำให้ติดเครื่องหมายลบเมื่อย้ายข้าง ดังนั้นเพื่อประโยชน์ในทางพีชคณิต (ผลลัพธ์ยังคงเดิม) จึงเสนอ F เป็น $F = \sigma^2\sum_i^n c_i^2 - 2\lambda_1(\sum_i^n c_i - 1) - 2\lambda_2\sum_i^n c_iX_i$

$$\frac{\partial F}{\partial \lambda_2} = 0 \Rightarrow -2\sum^{n} c_i X_i = 0 \text{ หรือ } \sum^{n} c_i X_i = 0 \qquad \text{..........(3)}$$

จาก (1) พบว่า

$$c_i = \frac{1}{\sigma^2}(\lambda_1 + \lambda_2 X_i) \; ; \; i = 1,2,\ldots,n \qquad \text{.......... (4)}$$

[1]

จาก (4) ให้รวมตลอดในทุกค่าของ i จะพบว่า

$$\sum_i^n c_i = \frac{1}{\sigma^2}(n\lambda_1 + \lambda_2 \sum_i^n X_i) \qquad \text{.......... (5)}$$

แต่จาก (2) เราทราบว่า $\sum_i^n c_i = 1$ แทนค่า $\sum_i^n c_i = 1$ ใน (5) จะได้

$$\lambda_1 = \frac{\sigma^2}{n} - \lambda_2 \bar{X} \qquad \text{..........(6)}$$

แทน (6) ใน (4) จะได้

$$c_i = \frac{1}{\sigma^2}\left(\frac{\sigma^2}{n} - \lambda_2 \bar{X} + \lambda_2 X_i\right) \; ; \; i = 1,2,\ldots,n \qquad \text{..........(7)}$$

จาก (7) ให้คูณตลอดด้วย X_i แล้วรวมตลอดในทุกค่าของ i จะพบว่า

$$\sum_i^n c_i X_i = \frac{1}{\sigma^2}\left(\frac{\sigma^2}{n}\sum_i^n X_i - \lambda_2 \sum_i^n \bar{X} X_i + \lambda_2 \sum_i^n X_i^2\right)$$

$$= \frac{1}{\sigma^2}\left\{\sigma^2 \bar{X} + \lambda_2 \left(\sum_i^n X_i^2 - n\bar{X}^2\right)\right\}$$

$$\sum_i^n c_i X_i = \bar{X} + \frac{\lambda_2}{\sigma^2}\sum_i^n x_i^2 \qquad \text{.......... (8)}$$

แต่จาก (3) เราทราบว่า $\sum_i^n c_i X_i = 0$ ดังนั้นจาก (8) จะพบว่า

$$\lambda_2 = -\frac{\sigma^2 \bar{X}}{\sum_i^n x_i^2} \qquad \text{..........(9)}$$

unu $\lambda_2 = -\frac{\sigma^2 \bar{X}}{\sum_i^n x_i^2}$ ใน (7) พบว่า

---

[1] ขอให้สังเกตว่าสมการที่ (4) ให้ค่า c_i ติดอยู่ในเทอมของ λ_1 และ λ_2 ถ้าสามารถหาค่า λ_1 และ λ_2 จากระบบสมการมาแทนค่าได้ ก็จะได้ค่า c_i ตามต้องการ

$$c_i = \frac{1}{\sigma^2}\left\{\frac{\sigma^2}{n} + \left(-\frac{\sigma^2\bar{X}}{\sum_i^n x_i^2}\right)(X_i - \bar{X})\right\} ; i = 1, 2, \ldots, n$$

นั่นคือ

$$c_i = \left(\frac{1}{n} - \bar{X}\frac{x_i}{\sum_i^n x_i^2}\right)$$

$$= \frac{1}{n} - \bar{X}w_i ; i = 1,2,3,\ldots,n$$

ดังนั้นตัวถ่วงน้ำหนัก c_i ที่มีผลให้ $\hat{\alpha} = \sum_i^n c_iY_i$ เป็น BLUE ของ α คือ $c_i = \left(\frac{1}{n} - \bar{X}w_i\right)$; $i = 1, 2, \ldots, n$

และเมื่อทดลองแทนที่ c_i ดังกล่าวลงใน Linear Function $\sum_i^n c_iY_i$ จะพบว่า

$$\hat{\alpha} = \sum_i^n c_iY_i = \sum_i^n\left(\frac{1}{n} - \bar{X}w_i\right)Y_i$$

$$= \bar{Y} - \bar{X}\sum_i^n w_iY_i$$

$$= \bar{Y} - \bar{X}\sum_i^n w_iy_i{}^{1}$$

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

ซึ่งแสดงว่า $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$ เป็น BLUE และมีผลตรงกันกับตัวประมาณค่าที่ประมาณได้โดยอาศัย Gauss – Markov Theorem

หมายเหตุ ให้นักศึกษาพิสูจน์หาค่า c_i ที่มีผลให้ $\hat{\beta} = \sum_i^n c_iY_i$ เป็น BLUE เป็นแบบฝึกหัด

2.4 ค่าประมาณของ σ^2 และสัมประสิทธิ์แห่งการตัดสินใจ

2.4.1 ค่าประมาณของ σ^2

จากตอน 2.3 เราได้พิสูจน์มาแล้วว่า $V(\hat{\beta}) = \frac{\sigma^2}{\sum_i^n x_i^2}$, $V(\hat{\alpha}) = \sigma^2\frac{\sum_i^n X_i^2}{n\sum_i^n x_i^2}$

---

[1] (1) $\sum_i^n w_iy_i = \sum_i^n w_i(Y_i - \bar{Y}) = \sum_i^n w_iY_i - \bar{Y}\sum_i^n w_i = \sum_i^n w_iY_i ; \sum_i^n w_i = 0$

(2) $\sum_i^n w_iy_i = \sum_i^n\left(\frac{x_i}{\sum x_i^2}\right)y_i = \frac{\sum x_iy_i}{\sum x_i^2} = \hat{\beta}$

และ $\text{Cov}(\hat{\alpha}, \hat{\beta}) = -\frac{\bar{X}}{\Sigma x_i^2}\sigma^2$ ปัญหาก็คือ σ^2 เป็น Unknown Parameter ทำให้ไม่สามารถนำนำสูตร $V(\hat{\alpha})$ และ $V(\hat{\beta})$ ไปใช้งานในทางปฏิบัติ โดยเฉพาะในงานทดสอบสมมุติฐาน และงานประมาณค่าพารามิเตอร์ α และ β ด้วยช่วงเชื่อมั่นได้ ดังนั้นในขั้นต้นนี้เราจึงมีความจำเป็นต้องประมาณค่า σ^2 โดยอาศัยข้อมูลจากค่าสังเกต $Z_i = (Y_i, X_i)$; $i = 1, 2, \ldots, n$ เสียก่อน

พิจารณา $e_i = Y_i - \hat{Y}_i = y_i - \hat{y}_i$ จะพบว่า $e_i = y_i - \hat{\beta}x_i$

แต่ $y_i = \beta x_i + (u_i - \bar{u})$

$$\text{ดังนั้น } \sum_i^n e_i^2 = \sum_i^n \{ \beta x_i + (u_i - \bar{u}) - \hat{\beta}x_i \}^2 \text{[1]}$$

$$= \sum_i^n \{ -(\hat{\beta} - \beta)x_i + (u_i - \bar{u}) \}^2$$

$$= \sum_i^n \{ (\hat{\beta} - \beta)^2 x_i^2 + (u_i - \bar{u})^2 - 2(\hat{\beta} - \beta)x_i(u_i - \bar{u}) \}$$

$$= (\hat{\beta} - \beta)^2 \sum_i^n x_i^2 + \sum_i^n (u_i - \bar{u})^2 - 2(\hat{\beta} - \beta) \cdot \sum_i^n x_i(u_i - \bar{u})$$

$$\text{ดังนั้น } E\left(\sum_i^n e_i^2 \right) = \sum_i^n x_i^2 E(\hat{\beta} - \beta)^2 + E\sum_i^n (u_i - \bar{u})^2 - 2E\{(\hat{\beta} - \beta)\sum_i^n x_i(u_i - \bar{u})\}$$

$$= A + B - C$$

$$A = \sum_i^n x_i^2 E(\hat{\beta} - \beta)^2 = \sum_i^n x_i^2 \left(\frac{\sigma^2}{\sum_i^n x_i^2} \right) = \sigma^2$$

$$B = E\sum_i^n (u_i - \bar{u})^2 = E\left\{ \sum_i^n u_i^2 - \frac{(\sum_i^n u_i)^2}{n} \right\}$$

$$= E\left\{ \sum_i^n u_i^2 - \frac{1}{n}\left(\sum_i^n u_i^2 + \sum_{i \neq j}^n \sum^n u_i u_j \right) \right\}$$

---

[1] $\sum_i^n e_i^2 = \sum_i^n (e_i - \bar{e})^2$ เพราะ $\bar{e} = 0$ แสดงว่า $\sum_i^n e_i^2$ ก็คือ Sum Square Error ซึ่งสามารถเปลี่ยนเป็น Mean Square Error ได้เพียงแต่นำ df มาหาร และนักศึกษาจะพบจากการพิสูจน์ว่า df มีค่าเท่ากับ $n-2$ เลข 2 ในที่นี้แสดงจำนวนพารามิเตอร์คือ α และ β

$$= n\sigma^2 - \frac{1}{n}(n\sigma^2 + 0)$$

$$= (n-1)\sigma^2$$

$$C = 2E\left\{(\hat{\beta}-\beta)\sum_i^n x_i(u_i - \bar{u})\right\}$$

$$= 2E\left\{(\beta + \sum_i^n w_i u_i - \beta)(\sum_i^n x_i u_i - \bar{u}\sum_i^n x_i)\right\}$$

$$= 2E\left\{(\sum_i^n w_i u_i)(\sum_i^n x_i u_i)\right\} \quad ; \quad \sum_i^n x_i = 0$$

$$= 2E\left\{\left(\frac{1}{\sum_i^n x_i^2} \cdot \sum_i^n x_i u_i\right)\left(\sum_i^n x_i u_i\right)\right\} \quad ; \quad w_i = \frac{x_i}{\sum_i^n x_i^2}$$

$$= \frac{2}{\sum_i^n x_i^2} E\left(\sum_i^n x_i^2 u_i^2 + \sum_i^n \sum_{\substack{j \\ i \neq j}}^n x_i x_j u_i u_j\right)$$

$$= \frac{2}{\sum_i^n x_i^2}\left(\sigma^2 \sum_i^n x_i^2\right)$$

$$= 2\sigma^2$$

ดังนั้น $E(\sum_i^n e_i^2) = A + B - C$

$$= \sigma^2 + (n-1)\sigma^2 - 2\sigma^2$$

$$= (n-2)\sigma^2$$

แสดงว่า $\hat{\sigma}^2 = \frac{1}{n-2}\sum_i^n e_i^2$ [1]

หรือ $\frac{1}{n-2}\sum_i^n e_i^2$ เป็น Unbiased Estimator ของ σ^2

นั่นคือ ตัวประมาณค่าของ σ^2 คือ $\hat{\sigma}^2 = \frac{1}{n-2}\sum_i^n e_i^2$ เมื่อ e_i คือ Residnal กล่าวคือ

$$e_i = Y_i - \hat{Y}_i$$

$$= Y_i - \hat{\alpha} - \hat{\beta}X_i$$

หรือ $e_i = y_i - \hat{y}_i$

$$= y_i - \hat{\beta}x_i$$

---

[1] ถ้า $E(t) = \theta$ แสดงว่า $\hat{\theta} = t$ เช่น $E(\bar{X}) = \mu$ แสดงว่า $\hat{\mu} = \bar{X}$ และ $E(S^2) = \sigma^2$ แสดงว่า $\hat{\sigma}^2 = S^2$ เป็นต้น

พิจารณา $\sum^n e_i^2$ จะพบว่าเราสามารถหา Computational Formular ได้ดังนี้

$$\sum^n e_i^2 = \sum^n (y_i - \hat{\beta} x_i)^2$$

$$= \sum^n y_i^2 - 2\hat{\beta} \sum^n x_i y_i + \hat{\beta}^2 \sum^n x_i^2$$

แต่ $$\hat{\beta}^2 \sum_i^n x_i^2 = \hat{\beta} \cdot \hat{\beta} \sum_i^n x_i^2 = \hat{\beta} \left(\frac{\sum x_i y_i}{\sum x_i^2} \right) \sum_i^n x_i^2 = \hat{\beta} \sum_i^n x_i y_i$$

ดังนั้น $$\sum_i^n e_i^2 = \sum_i^n y_i^2 - \hat{\beta} \sum_i^n x_i y_i$$

นั่นคือ $$\hat{\sigma}^2 = \frac{1}{n-2} \sum_i^n e_i^2$$

$$= \frac{1}{n-2} \left(\sum_i^n y_i^2 - \hat{\beta} \sum_i^n x_i y_i \right)$$

และเมื่อแทนค่า σ^2 ด้วย $\hat{\sigma}^2$ จะได้

$$\hat{V}(\hat{\beta}) = \frac{\hat{\sigma}^2}{\sum x_i^2}, \quad \hat{V}(\hat{\alpha}) = \hat{\sigma}^2 \frac{\sum X_i^2}{n \sum x_i^2} \quad \text{และ} \quad \hat{Cov}(\hat{\alpha}, \hat{\beta}) = \frac{-\bar{X}}{\sum x_i^2} \hat{\sigma}^2$$

2.4.2 **สัมประสิทธิ์แห่งการตัดสินใจ** (Coefficient of Determination : r^2)

สัมประสิทธิ์แห่งการตัดสินใจ คือดัชนีที่ใช้วัดปริมาณความผันแปรของตัวแปรสุ่ม Y ที่ตัวแปรอิสระสามารถดูดซับ (Absorb) เอาไว้ได้ หรือในความหมายเชิงปฏิบัติ r^2 คือดัชนีที่ใช้วัดความถูกต้องน่าเชื่อถือของสมการถดถอย

พิจารณา Residual $e_i = y_i - \hat{y}_i$ จะพบว่า $y_i = \hat{y}_i + e_i$ เมื่อ $\hat{y}_i = \hat{\beta} x_i$

ดังนั้นความผันแปรของตัวแปรสุ่ม Y คือ $\sum_i^n (Y_i - \bar{Y})^2 = \sum_i^n y_i^2$ สามารถแยกออกเป็นส่วน ๆ ได้โดยอาศัยสมการ $y_i = \hat{\beta} x_i + e_i$ ดังนี้

$$\sum_i^n y_i^2 = \sum_i^n (\hat{y}_i + e_i)^2 = \sum_i^n \hat{y}_i^2 + \sum_i^n e_i^2 + 2\sum_i^n \hat{y}_i e_i$$

พิจารณา $\sum^n \hat{y}_i e_i$ จะพบว่า

$$\sum_i^n \hat{y}_i e_i = \sum_i^n \hat{y}_i(y_i - \hat{\beta}x_i) = \sum_i^n y_i\hat{y}_i - \hat{\beta}\sum_i^n x_i\hat{y}_i$$

$$= \sum_i^n (\hat{\beta}x_i)y_i - \hat{\beta}\sum_i^n x_i\hat{y}_i$$

$$= \hat{\beta}\sum_i^n x_iy_i - \hat{\beta}\sum_i^n x_i\hat{y}_i$$

$$= \hat{\beta}\sum_i^n x_iy_i - \hat{\beta}\sum_i^n x_i(\hat{\beta}x_i)$$

$$= \hat{\beta}\sum_i^n x_iy_i - \hat{\beta}\sum_i^n x_i^2\left(\frac{\sum x_iy_i}{\sum x_i^2}\right)$$

$$= \hat{\beta}\sum_i^n x_iy_i - \hat{\beta}\sum_i^n x_iy_i = 0$$

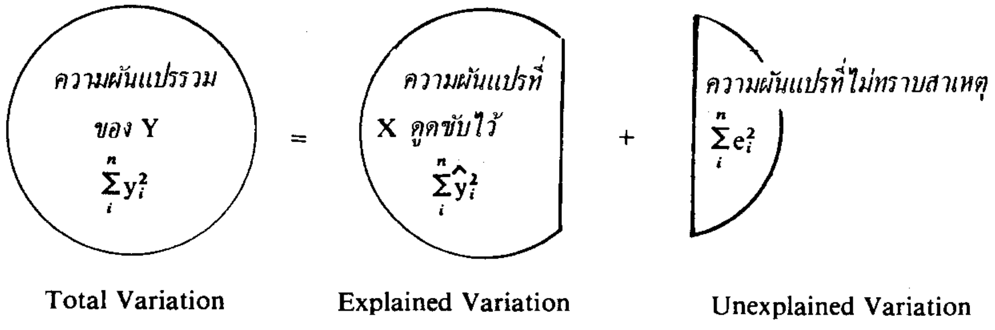
ดังนั้น $\sum_i^n y_i^2 = \sum_i^n \hat{y}_i^2 + \sum_i^n e_i^2$ หรือ $\sum_i^n (Y_i - \bar{Y})^2 = \sum_i^n (\hat{Y}_i - \bar{Y})^2 + \sum_i^n e_i^2$

แสดงว่า ความผันแปรของตัวแปรสุ่ม Y สามารถจำแนกออกเป็น 2 ส่วนคือ ความผันแปรที่วัดค่าได้ด้วยสมการประมาณค่า $Y = \hat{\alpha} + \hat{\beta}X$ หรือถ้ามองลึกลงไปได้ถึงโครงสร้างก็คือ ความผันแปรที่ตัวแปรสุ่ม X สามารถดูดซับเอาไว้ส่วนหนึ่ง และความผันแปรที่เราไม่ทราบสาเหตุแน่ชัดอีกส่วนหนึ่ง[1]

ด้วยเหตุนี้อัตราส่วนที่ตัวแปร X สามารถดูดซับความผันแปรของตัวแปรสุ่ม Y ไว้ได้จึงวัดออกมาในรูปของ $\frac{\sum \hat{y}_i^2}{\sum y_i^2}$ ค่าอัตราส่วน (Ratio) นี้จะมีค่าสูงสุดเท่ากับ 1 หรือ 100 เมื่อ $\sum_i^n \hat{y}_i^2 = \sum_i^n y_i^2$ ซึ่งในกรณีนี้ย่อมยืนยันให้เห็นว่าตัวแปร X ที่เรากำหนดขึ้นนั้นสามารถใช้ควบคุมความเคลื่อนไหวของ Y ได้ทั้งหมด และอัตราส่วนนี้จะมีค่าลดต่ำกว่า 1 ตามลำดับถ้าหากตัวแปร X สามารถควบคุมความเคลื่อนไหวของ Y ได้น้อยลง ปริมาณที่ขาดจำนวนไปจากจำนวนเต็ม 1 หรือ 100% คือปริมาณที่แสดงถึงความผันแปรที่เราไม่ทราบสาเหตุที่แน่ชัดในความเปลี่ยนแปลงของ Y ดังไดอะแกรม

---

[1] $\sum_i^n y_i^2 = \sum_i^n \hat{y}_i^2 + \sum_i^n e_i^2$ แสดงว่า Sum Square Total = Sum Square Regression + Sum Square Residual (error)



ขอให้สังเกตว่า ปริมาณความผันแปรทั้งสองส่วนที่ประกอบกันเป็นความผันแปรรวมของตัวแปรสุ่ม Y นั้นจะมีลักษณะของ Linear Additive ด้วยเหตุนี้ถ้า $\sum_i^n \hat{y}_i^2$ มีค่าสูงขึ้น $\sum_i^n e_i^2$ จะมีค่าลดลง และถ้า $\sum_i^n \hat{y}_i^2$ มีค่าลดลง $\sum_i^n e_i^2$ จะมีค่าสูงขึ้น สิ่งที่นักวิจัยปรารถนาคือ $\sum_i^n \hat{y}_i^2$ ที่มีค่าสูง ซึ่งสะท้อนให้เห็นว่าแบบจำลองที่เรากำหนดขึ้นนั้นสอดคล้องกับธรรมชาติของข้อมูลที่มีอยู่มาก ในกรณีของ Multiple Regression นักวิจัยจะพยายามหาหนทางเพิ่มปริมาณของ $\sum_i^n \hat{y}_i^2$ ให้สูงขึ้น ด้วยวิธีต่าง ๆ ด้วยการตรวจสอบ Lack of Fit หรือ Partial Correlation

พิจารณา $\frac{\sum \hat{y}_i^2}{\sum y_i^2}$ จะพบว่า

$$\frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{\sum(\hat{\beta} x_i)^2}{\sum y_i^2} = \frac{\hat{\beta}^2 \sum x_i^2}{\sum y_i^2} = \left(\frac{\sum x_i y_i}{\sum x_i^2}\right)^2 \left(\frac{\sum x_i^2}{\sum y_i^2}\right)$$

$$= \frac{\left(\sum x_i y_i\right)^2}{\sum x_i^2 \sum y_i^2} = \left(\frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}\right)^2$$

$$= r^2 = (\text{สปส..สหสัมพันธ์})^2$$

ดังนั้น r^2 จึงเป็นดัชนีที่ช่วยตัดสินว่าสมการถดถอยของเรามีคุณภาพดีเพียงใด นักศึกษาจะใช้สูตรใดสูตรหนึ่งข้างบนนี้คำนวณหาค่า r^2 ก็ได้แต่โดยทั่วไปเรานิยมใช้

$$r^2 = \frac{\hat{\beta} \sum x_i y_i}{\sum y_i^2}$$

เพราะมีโครงสร้างที่เชื่อมโยงกันได้กับกรณีของ **Multiple Regression** ซึ่งจะได้กล่าวถึงเรื่องนี้โดยละเอียดอีกครั้งหนึ่งในบทที่ 3

หมายเหตุ

การทราบค่าสหสัมพันธ์ r นอกจากทำให้ทราบค่า r^2 ซึ่งเป็นดัชนีแห่งการตัดสินใจได้แล้ว เรายังสามารถนำ r และ r^2 ไปใช้ประโยชน์ได้หลายประการดังนี้

1. ใช้คำนวณหา $\hat{\beta}$

จาก $\hat{\beta} = \dfrac{\sum_i^n x_i y_i}{\sum_i^n x_i^2}$ ให้นำ $\sqrt{\sum_i^n y_i^2}$ คูณตลอดทั้งเศษและส่วน จะพบว่า

$$\hat{\beta} = \frac{\sum_i^n x_i y_i}{\sum_i^n x_i^2} = \frac{\sum_i^n x_i y_i}{\sum_i^n x_i^2} \cdot \frac{\sqrt{\sum_i^n y_i^2}}{\sqrt{\sum_i^n y_i^2}}$$

$$= \frac{\sum_i^n x_i y_i}{\sqrt{\sum_i^n x_i^2 \sum_i^n y_i^2}} \cdot \frac{\sqrt{\sum_i^n y_i^2}}{\sqrt{\sum_i^n x_i^2}} = r\frac{S_y}{S_x}$$

แสดงว่าถ้าเราทราบค่า r และทราบค่าส่วนเบี่ยงเบนมาตรฐานของตัวแปรสุ่ม Y และตัวแปรอิสระ X เราย่อมสามารถคำนวณหาค่าของ $\hat{\beta}$ ได้ นอกจากนี้ $\hat{\beta} = r\frac{S_y}{S_x}$ ยังชี้ให้เห็นอีกด้วยว่า $\hat{\beta}$ คือดัชนีที่พอจะใช้วัดค่าสหสัมพันธ์ระหว่างตัวแปร Y และ X ได้โดยประมาณ และชี้ให้เห็นถึงลักษณะของ **Causual Relationship** ได้ในขณะที่ r จะบอกเพียงดีกรีของความสัมพันธ์เท่านั้น ไม่อาจมองในรูปของ **Causual Relationship** ได้

2. **ใช้ประมาณค่า** Mean Square Error ($\hat{\sigma}^2$) และ Mean Square Regression

จาก $$\hat{\sigma}^2 = \frac{\sum_i^n e_i^2}{n-2}$$

แต่เราทราบว่า $$\sum_i^n y_i^2 = \sum_i^n \hat{y}_i^2 + \sum_i^n e_i^2 \quad \text{หรือ} \quad 1 = \frac{\sum_i^n \hat{y}_i^2}{\sum_i^n y_i^2} + \frac{\sum_i^n e_i^2}{\sum_i^n y_i^2}$$

แสดงว่า $$r^2 = 1 - \frac{\sum_i^n e_i^2}{\sum_i^n y_i^2} \quad \text{หรือ} \quad \sum_i^n e_i^2 = \sum_i^n y_i^2(1-r^2)$$

นั่นคือ $$\hat{\sigma}^2 = \frac{\sum_i^n y_i^2(1-r^2)}{n-2}$$

และจาก $$SSR = \sum_i^n \hat{y}_i^2$$

จะพบว่า
$$\begin{aligned}
\sum_i^n \hat{y}_i^2 &= \sum_i^n (\hat{\beta} x_i)^2 = \hat{\beta}^2 \sum_i^n x_i^2 \\
&= \left(\frac{\sum_i^n x_i y_i}{\sum_i^n x_i^2}\right)^2 \sum_i^n x_i^2 \\
&= \frac{\left(\sum_i^n x_i y_i\right)^2}{\sum_i^n x_i^2}
\end{aligned}$$

คูณทั้งเศษและส่วนด้วย $\sum_i^n y_i^2$ จะได้

$$\begin{aligned}
SSR &= \frac{\left(\sum_i^n x_i y_i\right)^2}{\sum_i^n x_i^2} \cdot \frac{\sum_i^n y_i^2}{\sum_i^n y_i^2} \\
&= \sum_i^n y_i^2 \left(\frac{\sum_i^n x_i y_i}{\sqrt{\sum_i^n x_i^2 \sum_i^n y_i^2}}\right)^2 \\
&= r^2 \sum_i^n y_i^2
\end{aligned}$$

ดังนั้นตาราง ANOVA จึงสามารถเสนอได้ดังนี้

SOV	df	SS	MS	F
X	1	$\Sigma\hat{y}^2 = \hat{\beta}\sum_i^n x_iy_i = r^2\sum_i^n y_i^2$	$r^2\sum_i^n y_i^2$	$\frac{r^2(n-2)}{1-r^2}$
Residual	$n-2$	$\sum_i^n e_i^2 = \sum_i^n y_i^2 - \hat{\beta}\sum_i^n x_iy_i = \sum_i^n y_i^2(1-r)^2$	$\frac{\sum_i^n y_i^2(1-r^2)}{n-2}$	
Total	$n-1$	$\sum_i^n y_i^2$		

สำหรับกรณีนี้เราจะใช้ $F = \frac{\hat{\beta}\sum_i^n x_iy_i}{(\sum_i^n y_i^2 - \hat{\beta}\sum_i^n x_iy_i)/n-2}$ หรือ $F = \frac{r^2(n-2)}{1-r^2}$ ก็ได้

F ดังกล่าวจะนำไปใช้สำหรับทดสอบสมมติฐาน

$$H_0 : \beta = 0 \quad VS \quad H_1 : \beta \neq 0$$

แต่เนื่องจาก $t^2 = F$[1] ดังนั้นการทดสอบสมมติฐาน

$$H_0 : \beta = 0 \quad VS \quad H_1 : \beta \neq 0$$

จึงสามารถตัดสินใจได้โดยใช้

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

ซึ่ง $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ คือตัวสถิติที่ใช้ทดสอบ $H_0 : \rho = 0 \quad VS \quad H_1 : \rho \neq 0$

---

[1] จากนิยาม $F = \frac{U/\nu_1}{V/\nu_2}$ และ $t = \frac{Z}{\sqrt{V/\nu_2}}$ เมื่อ $U \sim \chi^2(\nu_1)$, $V \sim \chi^2(\nu_2)$ และ $Z \sim N(0, 1)$

จาก $F = \frac{U/\nu_1}{V/\nu_2}$ ถ้า $\nu_1 = 1$ แสดงว่า $\frac{U}{\nu_1} = U = (\frac{x_i-\mu}{\sigma})^2 = Z^2$

ดังนั้น $F = \frac{U/\nu_1}{V/\nu_2} = \frac{Z^2}{V/\nu_2} = (\frac{Z}{\sqrt{V/\nu_2}})^2 = t^2$

แสดงว่า $F_{\nu_1,\nu_2} = t^2_{\nu_2}$ เมื่อ $\nu_1 = 1$

เรื่องนี้จึงเป็นการยืนยันคำพูดที่กล่าวมาแล้วในข้อ 1. ได้เป็นอย่างดี

อนึ่ง จาก $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ เราสามารถจัดรูปให้เป็น $t = \frac{\hat{\beta}}{s_{\hat{\beta}}} = \frac{\hat{\beta}}{\hat{\sigma}/\sqrt{\Sigma x_i^2}}$ ซึ่งเป็นรูปของ t – test สำหรับทดสอบสมมติฐาน

$$H_0 : \beta = 0 \text{ VS } H_1 : \beta \neq 0$$

ซึ่งเราคุ้นเคยได้ดังนี้

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \hat{\beta}\frac{s_x}{s_y}\cdot\frac{\sqrt{n-2}}{\sqrt{1-r^2}}\ ^{1}$$

$$= \hat{\beta}\sqrt{\frac{\Sigma x_i^2}{\Sigma y_i^2}}\frac{\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$= \frac{\hat{\beta}\sqrt{\sum_i^n x_i^2}}{\sqrt{\sum_i^n y_i^2(1-r^2)/n-2}}$$

แต่
$$\hat{\sigma}^2 = \frac{\sum_i^n y_i^2(1-r^2)}{n-2}$$

ดังนั้น
$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{\hat{\beta}\sqrt{\sum_i^n x_i^2}}{\hat{\sigma}} = \frac{\hat{\beta}}{\hat{\sigma}/\sqrt{\Sigma x_i^2}} = \frac{\hat{\beta}}{s_{\hat{\beta}}}$$

ความจริงเหล่านี้จึงเป็นเครื่องยืนยันว่า $t = \frac{\hat{\beta}}{s_{\hat{\beta}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ และ $F = \frac{r^2(n-2)}{1-r^2}$ ต่างก็เป็น Test Statistics สำหรับทดสอบว่าตัวแปร X และ Y ไม่มีส่วนสัมพันธ์หรือกระทบกระเทือนสืบเนื่องกัน ($\beta = 0$ หรือ $\rho = 0$) ได้เช่นเดียวกัน

2.5 การอนุมานเชิงสถิติ

2.5.1 การประมาณค่า α และ β ด้วยช่วงเชื่อมั่น

จาก Sampling Distribution ของ $\hat{\beta}$ คือ $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum_i^n x^2})$ และโดยอาศัยทฤษฎีการโน้ม

---

[1] $\because\ \hat{\beta} = r\frac{s_y}{s_x}$ ดังนั้น $r = \hat{\beta}\frac{s_x}{s_y} = \hat{\beta}\sqrt{\frac{\Sigma x_i^2}{\Sigma y_i^2}}$

สู่เกณฑ์กลาง (Central Limit Theorem-CLT) เราจะพบว่า

$$Z = \frac{\hat{\beta}-\beta}{\sigma/\Sigma x_i^2} \sim N(0, 1)$$

และสามารถพิสูจน์ได้ว่า $\dfrac{\hat{\beta}-\beta}{\hat{\sigma}/\sqrt{\Sigma x_i^2}} \sim t_{(n-2)}$ [1] เมื่อ $\hat{\sigma}^2 = \dfrac{\sum_i^n e_i^2}{n-2}$

ดังนั้น $\Pr\left\{t_{(n-2,\alpha/2)} \leqslant \dfrac{\hat{\beta}-\beta}{\hat{\sigma}/\sqrt{\Sigma x_i^2}} \leqslant t_{(n-2,1-\alpha/2)}\right\} = 1-\alpha$

หรือ $\Pr\left\{\hat{\beta}+t_{(n-2,\alpha/2)}\dfrac{\hat{\sigma}}{\sqrt{\Sigma x_i^2}} \leqslant \beta \leqslant \hat{\beta}+t_{(n-2,1-\alpha/2)}\dfrac{\hat{\sigma}}{\sqrt{\Sigma x_i^2}}\right\} = 1-\alpha$

นั่นคือ ช่วงเชื่อมั่น $100(1-\alpha)$ % ที่คาดว่าค่าจริงของ β จะปรากฏอยู่คือ

$$\left(\hat{\beta}+t_{(n-2,\alpha/2)}\frac{\hat{\sigma}}{\sqrt{\Sigma x_i^2}}\ ,\ \hat{\beta}+t_{(n-2,1-\alpha/2)}\frac{\hat{\sigma}}{\sqrt{\Sigma x_i^2}}\right)$$

ในทำนองเดียวกัน ช่วงเชื่อมั่น $100(1-\alpha)$ % ที่คาดว่าค่าจริงของ α จะปรากฏอยู่คือ

$$\left(\hat{\alpha}+t_{(n-2,\alpha/2)}\ \hat{\sigma}\sqrt{\frac{\Sigma X_i^2}{n\Sigma x_i^2}} < \alpha < \hat{\alpha}+t_{(n-2,1-\alpha/2)}\ \hat{\sigma}\sqrt{\frac{\Sigma X_i^2}{n\Sigma x_i^2}}\right)$$

---

1 $\dfrac{\hat{\beta}-\beta}{\hat{\sigma}/\sqrt{\Sigma x_i^2}} = \dfrac{(\hat{\beta}-\beta)\sqrt{\Sigma x_i^2}}{\hat{\sigma}} = \dfrac{(\hat{\beta}-\beta)\sqrt{\Sigma x_i^2}/\sigma}{\hat{\sigma}/\sigma} = \dfrac{\dfrac{\hat{\beta}-\beta}{\sigma/\sqrt{\Sigma x_i^2}}}{\hat{\sigma}/\sigma} = \dfrac{Z}{\hat{\sigma}/\sigma}$

พิจารณาตัวแปรสุ่ม $\dfrac{\hat{\sigma}^2}{\sigma^2}$ จะพบว่า $\dfrac{\hat{\sigma}^2}{\sigma^2} = \dfrac{(n-2)\hat{\sigma}^2}{(n-2)\sigma^2} = \dfrac{(n-2)\hat{\sigma}^2}{\sigma^2}\cdot\dfrac{1}{n-2}$

แต่ตัวแปรสุ่ม $U = \dfrac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-2)}$

ดังนั้นตัวแปรสุ่ม $\dfrac{\hat{\sigma}^2}{\sigma^2} = U/(n-2)$

นั่นคือ $\dfrac{\hat{\beta}-\beta}{\hat{\sigma}/\sqrt{\Sigma x_i^2}} = \dfrac{Z}{\sqrt{U/n-2}} = t$ ที่มี df $= n-2$

2.5.2 การทดสอบสมมติฐานเกี่ยวกับพารามิเตอร์ α และ β

จากความเรื่อง Sampling Distribution ของ $\hat{\alpha}$ และ $\hat{\beta}$ คือ

$$\hat{\alpha} \sim N\left(\alpha, \sigma^2 \frac{\sum X_i^2}{n \sum x_i^2}\right) \quad \text{และ} \quad \hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum x_i^2}\right)$$

และโดยอาศัย Maximum Likelihood Ratio Test – MLRT เราสามารถพิสูจน์ได้ว่า

(1) ปฏิเสธสมมติฐานหลัก $H_0 : \alpha = 0$ VS $H_1 : \alpha \neq 0$ เมื่อ[1]

$$|t_c| > t_{n-2,1-\alpha/2} \text{ เมื่อ } t_c = \frac{\hat{\alpha}}{\hat{\sigma}\sqrt{\frac{\sum X_i^2}{n \sum x_i^2}}}$$

(2) ปฏิเสธสมมติฐานหลัก $H_0 : \beta = 0$ VS $H_1 : \beta \neq 0$ เมื่อ

$$|t_c| > t_{n-2,1-\alpha/2} \text{ เมื่อ } t_c = \frac{\hat{\beta}}{\hat{\sigma}/\sqrt{\sum x_i^2}}$$

---

[1] การทดสอบสมมติฐานลักษณะอื่นที่แตกต่างไปจากนี้ เช่น $\alpha = \alpha_0$, $\beta = \beta_0$ จะได้กล่าวถึงในตอน 3.2.5.3

ตัวอย่าง 2.2 ก. ถ้า r คือสหสัมพันธ์ระหว่างตัวแปร Y และ X จากค่าสังเกต $Z_i = (Y_i, X_i)$; $i = 1, 2, \ldots, n$ จงแสดงให้เห็นว่าสหสัมพันธ์ระหว่างตัวแปร U และ V จากค่าสังเกต $Z_i = (aY_i+b, cX_i+d)$ ยังคงมีสูตรโครงสร้างเป็น r เช่นเดียวกับกรณีสหสัมพันธ์ระหว่างตัวแปร X และ Y

ข. ผลจากการวิเคราะห์ข้อมูลชุดหนึ่งพบว่า $\hat{y} = 1.2x$ และ $\hat{x} = 0.6y$ คือสมการถดถอย y on x และ x on y ตามลำดับ โดยที่ $x = X - \bar{X}$ และ $y = Y - \bar{Y}$ จงคำนวณหา r_{xy} และ s_x/s_y

วิธีทำ

ก. เนื่องจาก $r = \dfrac{\Sigma(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\Sigma(X_i-\bar{X})^2\Sigma(Y_i-\bar{Y})^2}}$ คือสหสัมพันธ์ระหว่าง X และ Y

เมื่อให้ $U_i = cX_i+d$ และ $V_i = aY_i+b$; $i = 1, 2, \ldots . n$ จะพบว่า

$$\bar{U} = c\bar{X}+d \text{ และ } \bar{V} = a\bar{Y}+b$$

ดังนั้น $$r_{UV} = \frac{\Sigma(U_i-\bar{U})(V_i-\bar{V})}{\sqrt{\Sigma(U_i-\bar{U})^2\Sigma(V_i-\bar{V})^2}} = \frac{ca\,\Sigma(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{c^2\Sigma(X_i-\bar{X})^2a^2\Sigma(Y_i-\bar{Y})^2}} = r$$

แสดงว่า Linear Function ของตัวแปรสุ่ม X และ Y มิได้มีผลให้ค่าสหสัมพันธ์เปลี่ยนแปลง

ข. $\hat{y} = 1.2x$ และ $\hat{x} = 0.6y$ จงคำนวณหา r_{xy} และ s_x/s_y

จาก $\hat{y} = 1.2x$ แสดงว่า $$\hat{\beta} = \frac{\sum_i^n x_iy_i}{\sum^n x_i^2} = 1.2$$

จาก $\hat{x} = 0.6y$ แสดงว่า $$\hat{\beta}' = \frac{\sum_i^n y_ix_i}{\sum_i^n y_i^2} = 0.6$$

ดังนั้น $$r_{xy} = \sqrt{\hat{\beta}'\hat{\beta}} = \sqrt{\frac{(\Sigma x_iy_i)^2}{\Sigma x_i^2\,\Sigma y_i^2}}$$

$$= \sqrt{(1.2)(0.6)} = \sqrt{.72} = 0.848$$

และ $$s_x/s_y = \sqrt{\hat{\beta}'/\hat{\beta}} = \sqrt{\frac{0.6}{1.2}} = 0.707$$

ตัวอย่าง 2.3 จากการสุ่มตัวอย่างเพื่อวัดเปอร์เซ็นต์ของมันเนย (X) และปริมาณสารอื่น ๆ ที่มิใช่มันเนย (Y) ในนมโค โดยทำการสำรวจตัวอย่างนมโคจากฟาร์มโคนมจาก 2 ท้องที่ คือ ท้องที่อำเภอมวกเหล็ก จ.สระบุรี และท้องที่อำเภอโพธาราม จ.ราชบุรี ปรากฏข้อมูลดังนี้ (ข้อมูลสมมติ)

อำเภอมวกเหล็ก : $\Sigma X_i = 51.13$ $\Sigma Y_i = 117.25$ $\Sigma x_i^2 = 1.27$ $\Sigma x_i y_i = 1.84$
$(n = 16)$

อำเภอโพธาราม : $\Sigma X_i = 37.20$ $\Sigma Y_i = 78.75$ $\Sigma x_i^2 = 1.03$ $\Sigma x_i y_i = 1.10$
$(n = 10)$

จงวิเคราะห์สมการถดถอย $Y = \alpha + \beta X + u$ ในทั้งสองท้องที่แล้วเปรียบเทียบดูว่าสมการทั้งสองมีความชันต่างกันหรือไม่

วิธีทำ เพราะว่า $\hat{\beta} = \dfrac{\sum_i^n x_i y_i}{\sum_i^n x_i^2}$, $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$, $\hat{\sigma}^2 = \dfrac{\sum_i^n e_i^2}{n-2} = \dfrac{1}{n-2}\left(\sum_i^n y_i^2 - \hat{\beta}\sum_i^n x_i y_i\right)$

$$\hat{V}(\hat{\beta}) = \frac{\hat{\sigma}^2}{\sum_i^n x_i^2}, \quad \hat{V}(\hat{\alpha}) = \hat{\sigma}^2 \frac{\sum^n X_i^2}{n\sum_i^n x_i^2} = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_i^n x_i^2}\right)$$

$$r^2 = \frac{\hat{\beta}\sum_i^n x_i y_i}{\sum_i^n y_i^2}, \quad \sum_i^n y_i^2 = \sum_i^n Y_i^2 - n\overline{Y}^2$$

ก. ในกรณีอำเภอมวกเหล็กจะพบว่า

$$\overline{X} = \frac{51.13}{16} = 3.195, \quad \overline{Y} = \frac{117.25}{16} = 7.328, \quad \hat{\beta} = \frac{1.84}{1.27} = 1.448$$

$$\hat{\alpha} = 7.328 - (1.448)(3.195) = 2.70$$

$$\hat{\sigma}^2 = \frac{1}{14}\{4.78 - (1.448)(1.84)\} = 0.151$$

$$\hat{V}(\hat{\beta}) = \frac{0.151}{1.27} = 0.119, \quad \sqrt{\hat{V}(\hat{\beta})} = 0.345$$

$$\hat{V}(\hat{\alpha}) = (0.151)\left(\frac{1}{16} + \frac{(3.195)^2}{1.27}\right) = 1.223, \quad \sqrt{\hat{V}(\hat{\alpha})} = 1.106$$

$$r^2 = \frac{(1.448)(1.84)}{4.78} = 0.557$$

นั่นคือ $\quad Y = 2.70 + 1.448X \; ; \; r^2 = 0.557$

(1.106) (0.345)
(2.441) (4.197)

หมายเหตุ ค่าในวงเล็บแถวที่ 1 หมายถึง Standard Error แถวที่ 2 หมายถึง ค่า t

ข. กรณีอำเภอโพธาราม

$$\bar{X} = \frac{37.2}{10} = 3.72, \quad \bar{Y} = \frac{78.75}{10} = 7.875, \quad \hat{\beta} = \frac{1.10}{1.03} = 1.068$$

$$\hat{\alpha} = 7.875 - (1.068)(3.72) = 3.902$$

$$\hat{\sigma}^2 = \frac{1}{8}\{(2.48 - (1.068)(1.10)\} = 0.163$$

$$\hat{V}(\hat{\beta}) = \frac{0.163}{1.03} = 0.158 \; ; \; \sqrt{\hat{V}(\hat{\beta})} = 0.397$$

$$\hat{V}(\hat{\alpha}) = (0.163)\left(\frac{1}{10} + \frac{(3.72)^2}{1.03}\right) = 2.206, \quad \sqrt{\hat{V}(\hat{\alpha})} = 1.485$$

$$r^2 = \frac{(1.068)(1.10)}{2.48} = 0.473$$

นั่นคือ $\quad Y = 3.902 + 1.068X \; ; \; r^2 = 0.473$

(1.485) (0.397)
(2.628) (2.69)

ค. การทดสอบสมมติฐาน $H_0 : \beta_1 = \beta_2$ VS $H_1 : \beta_1 \neq \beta_2$

เนื่องจาก $\hat{\beta}_1 \sim N\left(\beta_1, \dfrac{\sigma_1^2}{\sum_i^{n_1} x_{1i}^2}\right)$ และ $\hat{\beta}_2 \sim N\left(\beta_2, \dfrac{\sigma_2^2}{\sum_i^{n_2} x_{2i}^2}\right)$

ดังนั้น $\hat{\beta}_1 - \hat{\beta}_2 \sim N\left(\beta_1 - \beta_2, \dfrac{\sigma_1^2}{\sum_i^{n_1} x_{1i}^2} + \dfrac{\sigma_2^2}{\sum_i^{n_2} x_{2i}^2}\right)$

ในที่นี้เราจำเป็นต้องสมมติว่า $\sigma_1^2 = \sigma_2^2 = \sigma^2$ เพื่อประโยชน์ในการพัฒนาตัวทดสอบ และด้วยเหตุที่ σ^2 เป็นพารามิเตอร์ที่เราไม่ทราบค่า จึงประมาณค่า σ^2 ด้วย MLE พบว่า

$$\hat{\sigma}^2 = \frac{(n_1-2)\hat{\sigma}_1^2+(n_2-2)\hat{\sigma}_2^2}{n_1+n_2-4}$$

และพบว่า $$\frac{(\hat{\beta}_1-\hat{\beta}_2)-(\beta_1-\beta_2)}{\hat{\sigma}\sqrt{\frac{1}{\sum_i^{n_1} x_{1i}^2}+\frac{1}{\sum_i^{n_2} x_{2i}^2}}} \sim t_{n_1+n_2-4}$$

ดังนั้นเราจึงปฏิเสธสมมติฐานหลักเมื่อ $|t_c| \geq t_{n_1+n_2-4,1-\alpha/2}$ โดยที่

$$t_c = \frac{\hat{\beta}_1-\hat{\beta}_2}{\hat{\sigma}\sqrt{\frac{1}{\sum_i^{n_1} x_{1i}^2}\quad\frac{1}{\sum_i^{n_2} x_{2i}^2}}}$$

จากข้อมูลในตอน ก. และ ข. ทำให้ได้

$$\hat{\sigma}^2 = \frac{(14)(0.151)+(8)(0.163)}{22} = 0.155$$

ดังนั้น $$t_c = \frac{1.448-1.068}{(0.393)(1.326)} = 0.738$$

และจากตาราง t พบว่า $t_{22,.975} = 2.074$ จึงเห็นได้ว่า $|t_c|$ มิได้มากกว่า $t_{n-4,1-\alpha/2}$ ด้วยเหตุนี้เราจึงไม่อาจปฏิเสธสมมติฐานหลัก และเชื่อว่าสมการในตอน ก. และ ข. มีความชันไม่ต่างกันอย่างมีนัยสำคัญ

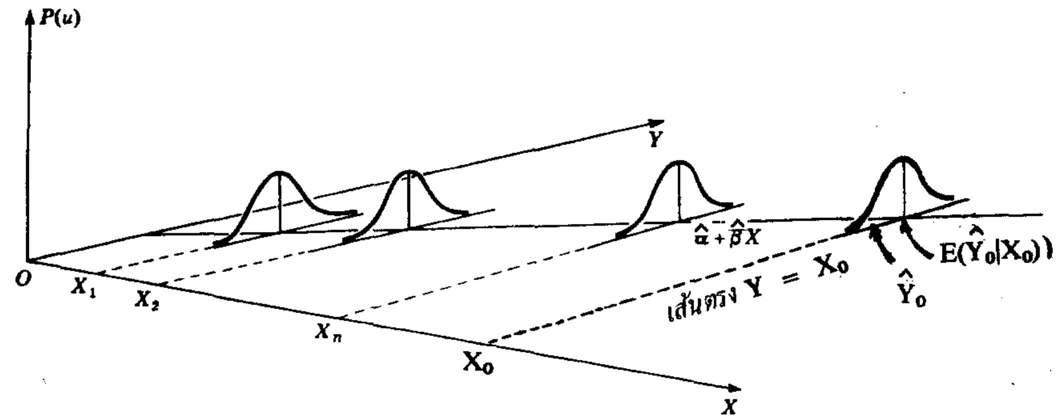
2.6 การพยากรณ์

การพยากรณ์ที่เราจะเสนอในที่นี้ผู้เขียนจะเสนอเป็น 2 นัย คือ การพยากรณ์ค่าของ $E(Y_0|X_0)$[1] เมื่อ X_0 คือค่าของ X ใน Sample Period (กรณีเรียกว่า Interpolation) หรือเป็นค่าของ X นอก Sample Period (กรณีนี้เรียกว่า Extrapolation) ก็ได้ และการพยากรณ์ค่าของ Y_0 (Individual Value) โดยปกติงานพยากรณ์ที่เกี่ยวข้องกับการวิเคราะห์ความถดถอยควรเป็นเรื่องของการพยากรณ์ค่าของ $E(Y_0|X_0)$ ทั้งนี้เพราะสมการจริงคือ $E(Y) = \alpha+\beta X$ ในขณะที่สมการคาดหมาย คือ $E(Y) = \hat{\alpha}+\hat{\beta}X$ เมื่อกำหนดให้ $X = X_0$ เราย่อมหาค่าของ $E(Y_0|X_0)$ ได้โดยง่าย

---

[1] Y_0 คือค่าที่สอดคล้องกับ X_0 ในที่นี้จะใช้ Subscript "0" สำหรับแสดงค่าพยากรณ์ของ Y และใช้สำหรับแสดงค่า X ภายนอก Sample Period หรือภายใน Sample Period ที่อยู่ระหว่าง X_i กับ X_j แต่เราอาจใช้ Subscript อื่นที่มิใช่ "0" เช่น "P" หรือ "F" ในความหมายของ Predicted หรือ Forecasted ก็ได้

เพียงแต่ลากเส้นดิ่งจากจุด $(X_0, 0)$ ไปตัดเส้นสมการ $E(Y) = \hat{\alpha}+\hat{\beta}X$ จุดตัดจะมี Coordinate เป็น $(X_0, \hat{\alpha}+\hat{\beta}X_0) = (X_0, E(\hat{Y}_0|X_0))$ ซึ่งยืนยันว่าค่าพยากรณ์ของ Y เมื่อกำหนดให้ $X = X_0$ ก็คือ $E(\hat{Y}_0|X_0)$ หรือค่าเฉลี่ยของ $\hat{Y}_0$ ที่เกิดจากการทดลองซ้ำ ๆ ณ จุดที่ $X = X_0$ รวมทั้งสิ้น r ครั้ง แต่บางครั้งนักวิจัยอาจมุ่งสนใจที่จะพยากรณ์เฉพาะค่าเดียว (Individual Value) ของ Y เมื่อกำหนดให้ $X = X_0$ ซึ่งเป็นกรณีที่แตกต่างจากกรณีที่กล่าวมาแล้ว ค่าเดียวของ Y ก็คือ Response ที่เกิดจากการทดลองให้ $X = X_0$ เพียงครั้งหนึ่งครั้งเดียว กรณีเช่นนี้ค่าของ Y อาจปรากฏ ณ ตำแหน่งใด ๆ ในแนวเส้นตรง $Y = X_0$ ก็ได้ดังภาพ ซึ่งในกรณีนี้เราจะไม่อาจแน่ใจในความถูกต้องของผลพยากรณ์ได้มากนัก ช่วงพยากรณ์สำหรับกรณีนี้จึงกว้างกว่ากรณีของการพยากรณ์ค่าเฉลี่ยของ Y เมื่อกำหนดค่า $X = X_0$ (Mean Response of Y given $X = X_0$) ดังกล่าว



ภาพ 2.8 แสดงการพยากรณ์ค่า $E(Y_0|X_0)$ *และ* Y_0 *นอก* Sample Period *ขอให้สังเกตว่าค่า* $E(Y_0|X_0)$ *จะวางอยู่บนเส้นตรง* $E(Y) = \hat{\alpha}+\hat{\beta}X$ *ส่วนค่าเดี่ยวของ* Y *คือ* $\hat{Y}_0$ *จะวางอยู่บนเส้นตรง* $Y = X_0$

2.6.1 การพยากรณ์ค่า $E(Y_0|X_0)$

ให้ X_0 คือค่าของ X นอก Sample Period (กรณี Extrapolation) หรือใน Sample Period (กร ณี Interpolation)

ให้ $\hat{Y}_0 = \sum_i^n c_i Y_i$ คือ Linear Function ของตัวแปรสุ่ม Y_i ; i = 1, 2, ... n เป็นตัวสถิติที่ใช้เป็น Pr .dictor โดยที่ c_i ; i = 1, 2, ... n เป็นตัวคงที่ใด ๆ ที่เลือกขึ้นมาในลักษณะที่มีผลให้ $\hat{Y}_0$ เป็น BLUE

เนื่องจาก $Y_i = \alpha+\beta X_i+u_i$

ดังนั้น $Y_0 = \alpha+\beta X_0+u_0$

และ $E(Y_0|X_0) = E(\alpha+\beta X_0+u_0) = \alpha+\beta X_0$ .... (ก)

พิจารณา $\hat{Y}_0 = \sum_i^n c_iY_i$

จะพบว่า
$$\hat{Y}_0 = \sum_i^n c_i(\alpha+\beta X_i+u_i)$$
$$= \alpha\sum_i^n c_i+\beta\sum_i^n c_iX_i+\sum_i^n c_iu_i$$

และ $E(\hat{Y}_0|X_0) = \alpha\sum_i^n c_i+\beta\sum_i^n c_iX_i$ .... (ข)

จาก (ก) และ (ข) จะเห็นว่า $E(\hat{Y}_0|X_0) = \alpha\sum_i^n c_i+\beta\sum_i^n c_iX_i \neq E(Y_0|X_0) = \alpha+\beta X_0$ แต่เมื่อพิจารณาให้ดีจะพบว่า $\hat{Y}_0|X_0$ จะเป็น Unbiased Estimator ของ $\alpha+\beta X_0$ ก็ต่อเมื่อ

$$\sum_i^n c_i = 1 \quad \text{และ} \quad \sum_i^n c_iX_i = X_0$$

นั่นคือ ถ้า $\sum_i^n c_i = 1$ และ $\sum_i^n c_iX_i = X_0$ แล้ว $\hat{Y}_0 = \sum_i^n c_iY_i$ จะเป็น BLUE ของ $E(Y_0|X_0)$ หรือนัยหนึ่ง $\sum_i^n c_i = 1$ และ $\sum_i^n c_iX_i = X_0$ คือเงื่อนไข (Constraint Condition) ที่ทำให้ $\hat{Y}_0 = \sum_i^n c_iY_i$ เป็น BLUE[1]

พิจารณา $V(Y_0)$ จะพบว่า

$$V(\hat{Y}_0) = E\{\hat{Y}_0-E(\hat{Y}_0|X_0)\}^2$$

แต่ $\hat{Y}_0 = \sum_i^n c_iY_i$ และ $E(\hat{Y}_0|X_0) = \alpha\sum_i^n c_i+\beta\sum_i^n c_iX_i$ ดังนั้น

---

[1] การพิสูจน์ส่วนนี้ยืนยันว่า $\hat{Y}_0 = \sum_i^n c_iY_i$ เป็น Linear Unbiased Estimator (LUE) ของ $E(Y_0|X_0)$ แต่จะเป็น LUE ได้ก็ต่อเมื่อ $\sum_i^n c_i = 1$ และ $\sum_i^n c_iX_i = X_0$ ขั้นต่อไปคือพิสูจน์เพื่อหาค่า c_i ที่สอดคล้องกับเงื่อนไขคู่นี้ที่มีผลให้ $\hat{Y}_0$ เป็น Minimum Variance (Best) Estimator

$$V(\hat{Y}_0) = E\{\sum_i^n c_iY_i - (\alpha\sum_i^n c_i + \beta\sum_i^n c_iX_i)\}^2$$

$$= E\{\sum_i^n c_i(\alpha + \beta X_i + u_i) - (\alpha\sum_i^n c_i + \beta\sum_i^n c_iX_i)\}^2$$

$$= E(\sum_i^n c_iu_i)^2$$

$$= E(\sum_i^n c_i^2u_i^2 + \sum_{i}^{n}\sum_{\substack{j \\ i\neq j}}^{n} c_ic_ju_iu_j) = \sigma^2\sum_i^n c_i^2 + 0$$

นั่นคือ $$V(\hat{Y}_0) = \sigma^2\sum_i^n c_i^2$$

ดังนั้น $V(\hat{Y}_0)$ จะมีค่าต่ำสุดภายใต้เงื่อนไขว่า $\sum_i^n c_i = 1$ และ $\sum_i^n c_iX_i = X_0$ ก็ต่อเมื่อ c_i

มีค่าสอดคล้องกับระบบสมการ $\frac{\partial F}{\partial c_i} = 0$, $\frac{\partial F}{\partial \lambda_1} = 0$, $\frac{\partial F}{\partial \lambda_2} = 0$ โดยที่

$$F = V(\hat{Y}_0) + \lambda_1(\sum_i^n c_i - 1) + \lambda_2(\sum_i^n c_iX_i - X_0)$$

$$= \sigma^2\sum_i^n c_i^2 + \lambda_1\sum_i^n c_i - \lambda_1 + \lambda_2\sum_i^n c_iX_i - \lambda_2X_0$$

ระบบสมการ $\frac{\partial F}{\partial c_i} = 0$; $i = 1, 2, \ldots, n$, $\frac{\partial F}{\partial \lambda_1} = \frac{\partial F}{\partial \lambda_2} = 0$ ปรากฏดังนี้

$$\frac{\partial F}{\partial c_i} = 0 \Rightarrow 2c_i\sigma^2 + \lambda_1 + \lambda_2X_i = 0 ; \quad i = 1, 2, \ldots, n \qquad \ldots\ldots. (1)$$

$$\frac{\partial F}{\partial \lambda_1} = 0 \Rightarrow \sum_i^n c_i - 1 = 0 \qquad \ldots\ldots. (2)$$

$$\frac{\partial F}{\partial \lambda_2} = 0 \Rightarrow \sum_i^n c_iX_i - X_0 = 0 \qquad \ldots\ldots. (3)$$

จาก (1) ให้รวมตลอดในทุกค่าของ i

$$\Rightarrow 2\sigma^2\Sigma c_i + n\lambda_1 + \lambda_2\Sigma X_i = 0 \qquad \ldots\ldots\ldots.(4)$$

และจาก (2) ให้คูณตลอดด้วย $2\sigma^2$

$$\Rightarrow 2\sigma^2\Sigma c_i = 2\sigma^2 \qquad \ldots\ldots\ldots.(5)$$

แทน (5) ใน (4)

$$\Rightarrow n\lambda_1 + \lambda_2\Sigma X_i = -2\sigma^2$$

$$\Rightarrow \lambda_1 = \frac{1}{n}(-2\sigma^2 - \lambda_2\Sigma X_i) = \frac{-2\sigma^2}{n} - \lambda_2\bar{X} \qquad \ldots\ldots\ldots.(6)$$

แทน (6) ใน (1)

$$\Rightarrow \quad 2c_i\sigma^2 - \frac{2\sigma^2}{n} - \lambda_2\bar{X} + \lambda_2 X_i = 0 \qquad ; i = 1, 2, \ldots, n.$$

$$\Rightarrow \quad 2c_i\sigma^2 = \frac{2\sigma^2}{n} - \lambda_2(X_i - \bar{X}) \qquad ; i = 1, 2, \ldots, n$$

$$\Rightarrow \quad c_i = \frac{1}{n} - \frac{\lambda_2}{2\sigma^2} x_i \qquad ; i = 1, 2, \ldots, n \qquad \ldots\ldots\ldots.(7)$$

จาก (7) ให้คูณตลอดด้วย X_i แล้วรวมตลอดในทุกค่าของ i

$$\Rightarrow \quad \Sigma c_i X_i = \bar{X} - \frac{\lambda_2}{2\sigma^2} \Sigma x_i X_i = \bar{X} - \frac{\lambda_2}{2\sigma^2} (\Sigma X_i^2 - n\bar{X}^2) \qquad \ldots\ldots\ldots.(8)$$

จาก (3) เราทราบว่า $\Sigma c_i X_i = X_0$ ดังนั้นสมการ (8) จึงกลายเป็น

$$\bar{X} - \frac{\lambda_2}{2\sigma^2} (\Sigma X_i^2 - n\bar{X}^2) = X_0$$

$$- \frac{\lambda_2}{2\sigma^2} \Sigma(X_i - \bar{X})^2 = X_0 - \bar{X}$$

$$\Rightarrow \quad \lambda_2 = \frac{-2\sigma^2(X_0 - \bar{X})}{\Sigma(X_i - \bar{X})^2} \qquad \ldots\ldots\ldots.(9)$$

แทน (9) ใน (7) จะได้ c_i ตามต้องการดังนี้

$$c_i = \frac{1}{n} + \frac{x_i}{2\sigma^2} \left(\frac{2\sigma^2(X_0 - \bar{X})}{\Sigma(X_i - \bar{X})^2} \right) \qquad ; i = 1,2,\ldots,n$$

นั่นคือ $c_i = \frac{1}{n} + \frac{(X_0 - \bar{X})x_i}{\Sigma(X_i - \bar{X})^2}$; i = 1,2,...,n **คือค่าคงที่ที่สอดคล้องกับเงื่อนไข** $\Sigma c_i = 1$ **และ** $\Sigma c_i X_i = X_0$ **และมีผลให้** $V(\hat{Y}_0)$ **มีค่าต่ำที่สุด (Best) ซึ่งเมื่อแทนที่** c_i **ดังกล่าวลงในโครงสร้าง** $\hat{Y}_0 = \Sigma c_i Y_i$ **จะพบว่า**

$$\hat{Y}_0 = \Sigma c_i Y_i = \Sigma \left\{ \frac{1}{n} + \frac{(X_0 - \bar{X})x_i}{\Sigma(X_i - \bar{X})^2} \right\} Y_i$$

$$= \bar{Y} + \frac{X_0 \Sigma x_i Y_i}{\Sigma(X_i - \bar{X})^2} - \frac{\bar{X}\Sigma x_i Y_i}{\Sigma(X_i - \bar{X})^2}$$

แต่ $\hat{\beta} = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$, $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$, $\Sigma x_i y_i = \Sigma x_i Y_i - \bar{Y}\Sigma x_i = \Sigma x_i Y_i$,

$x_i = (X_i - \bar{X})$

ดังนั้น $\hat{Y}_0 = \bar{Y} + X_0 \frac{\sum x_i y_i}{\sum x_i^2} - \frac{\bar{X}\sum x_i y_i}{\sum x_i^2}$

$$= (\bar{Y} - \hat{\beta}\bar{X}) + \hat{\beta}X_0$$

$$= \hat{\alpha} + \hat{\beta}X_0$$

นั่นคือ $\hat{Y}_0 = \hat{\alpha} + \hat{\beta}X_0$ คือ BLUE ของ $E(Y_0|X_0)$ ซึ่งจากโครงสร้างนี้ชี้ให้เห็นว่าการพยากรณ์ค่าเฉลี่ยของ Y_0 เมื่อ $X = X_0$ (คือ $E(Y_0|X_0)$) สามารถกระทำได้โดยง่าย เพียงแต่แทนที่ของ X ในสมการคาดหมายคือ $Y = \hat{\alpha} + \hat{\beta}X$ ด้วย X_0 เท่านั้น หรือนัยหนึ่งค่าพยากรณ์ของ $E(Y_0|X_0)$ นั้นสามารถหาได้โดยการลากเส้นดิ่งจากจุด $(X_0, 0)$ ไปตัดกับเส้นตรง $Y = \hat{\alpha} + \hat{\beta}X$ ค่า Ordinate ของจุดตัดก็คือค่าพยากรณ์ของ $E(Y_0|X_0)$

สำหรับ $V(\hat{Y}_0)$ สามารถหาได้ดังนี้

$$V(\hat{Y}_0) = E\{\hat{Y}_0 - E(\hat{Y}_0|X_0)\}^2 = E\{(\hat{\alpha} + \hat{\beta}X_0) - (\alpha + \beta X_0)\}^2 \quad [1]$$

$$= E\{(\hat{\alpha} - \alpha) + X_0(\hat{\beta} - \beta)\}^2$$

$$= E\{(\hat{\alpha} - \alpha)^2 + X_0^2(\hat{\beta} - \beta)^2 + 2X_0(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)\}$$

$$= E(\hat{\alpha} - \alpha)^2 + X_0^2 E(\hat{\beta} - \beta)^2 + 2X_0 E(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)$$

$$= V(\hat{\alpha}) + X_0^2 V(\hat{\beta}) + 2X_0 \operatorname{cov}(\hat{\alpha}, \hat{\beta})$$

$$= \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2}\right) + \sigma^2 \frac{X_0^2}{\sum x_i^2} + 2\sigma^2 X_0 \frac{-\bar{X}}{\sum x_i^2}$$

$$= \sigma^2\left\{\frac{1}{n} + \frac{X_0^2}{\sum x_i^2} - \frac{2\bar{X}X_0}{\sum x_i^2} + \frac{\bar{X}^2}{\sum x_i^2}\right\}$$

---

[1] จากสมการ (ข) $E(\hat{Y}_0|X_0) = \alpha\sum c_i + \beta\sum c_i X_i$ เมื่อแทนที่ c_i ด้วย $\frac{1}{n} + \frac{(X_0 - \bar{X})x_i}{\sum x_i^2}$ จะพบว่า

$$E(\hat{Y}_0|X_0) = \alpha\sum\left\{\frac{1}{n} + \left(\frac{X_0 - \bar{X}}{\sum x_i^2}\right)x_i\right\} + \beta\sum\left\{\frac{1}{n} + \frac{(X_0 - \bar{X})x}{\sum x_i^2}\right\}X_i = \alpha + \beta X_0$$

นั่นคือ $$V(\hat{Y}_0) = \sigma^2\left\{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right\}$$

และ Unbiased Estimator ของ $V(\hat{Y}_0)$ คือ $\hat{V}(\hat{Y}_0) = \hat{\sigma}^2\left\{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right\}$

โดยที่ $$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{1}{n-2}(\sum y_i^2 - \hat{\beta}\sum x_i y_i)$$

พิจารณา $\hat{V}(\hat{Y}_0) = \hat{\sigma}^2\left\{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right\}$ จะพบข้อสังเกตที่สำคัญ และการพัฒนา test Statistics ที่สืบเนื่องที่สำคัญดังนี้

1. $\hat{V}(\hat{Y}_0)$ มีค่าต่ำที่สุดเมื่อ $X_0 = \bar{X}$ คือ $\hat{V}(\hat{Y}_0) = \frac{\hat{\sigma}^2}{n}$ และ $\hat{V}(\hat{Y}_0)$ จะค่อย ๆ ทวีจำนวนสูงขึ้นเมื่อ X_0 ห่างจาก $\bar{X}$ ออกไปทั้งด้านซ้ายมือและขวามือของ $\bar{X}$ แสดงว่า Interpolation และ Extrapolation ที่กระทำโดยกำหนดให้ X_0 ห่างจาก $\bar{X}$ มาก ผลการพยากรณ์จะขาดความน่าเชื่อถือ และเห็นได้ค่อนข้างชัดเจนถ้าหากงานวิจัยนั้นเกี่ยวข้องอยู่กับข้อมูลอนุกรมเวลา คือเมื่อกำหนดแบบจำลองเป็น $Y_t = \alpha + \beta X_t + u_t$ ซึ่งในกรณีนี้ถ้าหากผู้วิจัยต้องการพยากรณ์ค่า $E(Y_0|X_0)$ เมื่อ X_0 คือค่าของ X_t ในอนาคตที่ไกลปัจจุบัน (Sample Period) มาก ๆ ผลการวิจัยจะขาดความน่าเชื่อถือเพราะ $\hat{V}(\hat{Y}_0|X_0)$ จะมีค่าสูงมาก

2. ช่วงพยากรณ์ของ $E(Y_0|X_0)$ จะค่อย ๆ กว้างขึ้นเมื่อ X_0 ค่อย ๆ คลาดเคลื่อนหรือเบนห่างจาก $\bar{X}$ กล่าวคือจาก $\hat{Y}_0 = \hat{\alpha} + \hat{\beta}X_0$ เราสามารถพิสูจน์ได้โดยอาศัย mgf-technique ได้ว่า

$$\hat{Y}_0 \sim N\left\{\alpha + \beta X_0, \sigma^2\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right)\right\}^1$$

และ $$\frac{\hat{Y}_0 - (\alpha + \beta X_0)}{\hat{\sigma}^2\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}}} \sim t_{n-2}{}^2$$

[1] เมื่อกำหนดเป็นข้อตกลงว่า $u \sim N(0, \sigma^2)$ เราสามารถพิสูจน์โดยง่ายได้ว่า $\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum x_i^2}\right)$ และ $\hat{\alpha} \sim N\left(\alpha, \sigma^2\frac{\sum X_i^2}{n\sum x_i^2}\right)$ แต่เนื่องจาก $\hat{Y}_0 = \hat{\alpha} + \hat{\beta}X_0$ แสดงว่า $\hat{Y}_0$ เป็น Linear Function ของ $\hat{\alpha}$ และ $\hat{\beta}$ ซึ่งเราสามารถพิสูจน์ได้โดยง่ายว่า $\hat{Y}_0 \sim N\left\{\alpha + \beta X_0, \sigma^2\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right)\right\}$

[2] ให้ $T = \frac{\hat{Y}_0 - E(\hat{Y}_0|X_0)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_0 - X)^2}{\sum x_i^2}}} = \frac{\{\hat{Y}_0 - E(\hat{Y}_0|X_0)\}/\sigma}{\{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_0 - X)^2}{\sum x_i^2}}\}/\sigma} = \left\{\frac{\hat{Y}_0 - E(\hat{Y}_0|X_0)}{\sigma\sqrt{\frac{1}{n} + \frac{(X_0 - X)^2}{\sum x_i^2}}}\right\}\Big/\frac{\hat{\sigma}}{\sigma} = \frac{Z}{\hat{\sigma}/\sigma} = \frac{Z}{\sqrt{\chi^2/n-2}} \sim t_{n-2}$

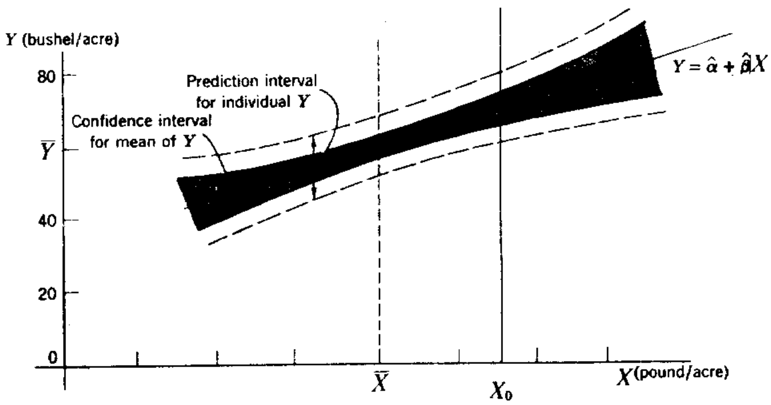
ดังนั้น $\Pr\left\{ t_{n-2,\frac{\alpha}{2}} < \dfrac{\hat{Y}_0 - (\alpha + \beta X_0)}{\hat{\sigma}\sqrt{\dfrac{1}{n} + \dfrac{(X_0 - \bar{X})^2}{\Sigma x_i^2}}} < t_{n-2,1-\frac{\alpha}{2}} \right\} = 1 - \alpha$

$$\Rightarrow \Pr\left\{ \hat{Y}_0 + t_{n-2,\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma x_i^2}} < E(Y_0|X_0) < \hat{Y}_0 + t_{n-2,1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma x_i^2}} \right\}$$

นั่นคือช่วงเชื่อมั่น หรือช่วงพยากรณ์ $(1 - \alpha)$ 100% ที่คาดว่าค่าจริงของ $E(Y_0|X_0)$ จะปรากฏอยู่คือ

$$\left(\hat{Y}_0 + t_{n-2,\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma x_i^2}}\ ,\ \hat{Y}_0 + t_{n-2,1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma x_i^2}} \right)$$

ภาพแสดงช่วงเชื่อมั่นดังกล่าวปรากฏดังนี้



ภาพ 2.9 แสดงช่วงเชื่อมั่น $(1 - \alpha)$ 100% ที่คาดว่า $E(Y_0|X_0)$ จะมีค่าปรากฏอยู่ ขอให้สังเกตว่า Random Interval นี้จะมีช่วงแคบที่สุดเมื่อ $X_0 = \bar{X}$ เพราะเมื่อ $X_0 = \bar{X}$, $\hat{V}(\hat{Y}_0)$ มีค่าต่ำที่สุด แต่เมื่อ X_0 เบนห่างจาก $\bar{X}$ มากขึ้น ทั้งด้านซ้ายและด้านขวาของ $\bar{X}$ Random Interval จะค่อย ๆ กว้างขึ้น Random Interval มีลักษณะดังภาพ เพราะถือว่ามี Error ทั้งในการประมาณค่า α และ β

จากภาพจะเห็นได้ว่าช่วงพยากรณ์ของ $E(Y_0|X_0)$ จะกว้างมาก เมื่อ X_0 เบนห่างจาก $\bar{X}$ มาก เรื่องนี้นับว่าสอดคล้องกับข้อเตือนใจที่กล่าวแล้วในตอนต้นว่า การพยากรณ์นั้นควรพยากรณ์ในระยะสั้น (Short Run) เท่านั้น ทั้งนี้เพราะในระยะสั้น ค่าของ X_0 ยังไม่ต่างจาก $\bar{X}$ มากนัก ช่วงพยากรณ์ของ $E(Y_0|X_0)$ จึงยังไม่กว้างจนทำให้การพยากรณ์ขาดความน่าเชื่อถือ

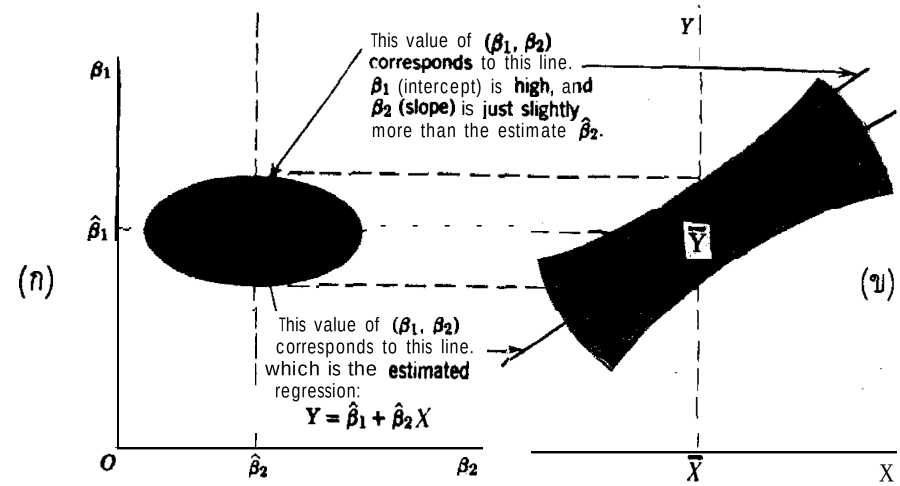
หนึ่งในกรณีของ Extrapolation หรือการพยากรณ์ค่าของ $E(Y_0|X_0)$ เมื่อ X_0 เป็นค่าของตัวแปร X นอก Sample Period นั้น อันตรายจากการพยากรณ์ผิดพลาดจะมีสูงกว่าในกรณีของ Interpolation อันตรายดังกล่าวจะเกิดขึ้นด้วยสาเหตุ 3 ประการคือ

ก. ช่วงพยากรณ์กว้างมากเกินไปจนงานพยากรณ์ขาดความน่าเชื่อถือ

ข. เราไม่อาจแน่ใจได้ว่าโครงสร้างของแบบจำลองจะยังคงมีลักษณะของ Linear Regression เช่นที่เคยเป็นอยู่หรือไม่ ทั้งนี้เพราะสถานการณ์แวดล้อมที่เปลี่ยนไป ความสัมพันธ์ทางคณิตศาสตร์ระหว่างตัวแปร Y และ X อาจเปลี่ยนแปลงไป แนวของเส้นสมการที่ต่อจากเส้นเดิมที่อยู่นอก Sample Period อาจเปลี่ยนเป็นเส้นโค้ง ซึ่งในกรณีเช่นนี้ ถ้าเราพยากรณ์ค่า $E(Y_0|X_0)$ โดยอาศัยสมการเดิมคือ $Y = \hat{\alpha} + \hat{\beta}X$ ผลการพยากรณ์จะผิดพลาด

ค. เราไม่อาจแน่ใจได้ว่าในอนาคต (นอก Sample Period) นั้นสถานการณ์ต่าง ๆ จะเอื้ออำนวยให้ข้อตกลงของสมการถดถอย (กล่าวแล้วในตอน 2.1) จะยังคงเป็นจริงอยู่ต่อไปหรือไม่

3. ช่วงเชื่อมั่นที่กว้างขึ้นแสดงว่าภายในช่วงนี้จะมี True Line ที่พึงเป็นไปได้มากมายหลายเส้นอัดอยู่ภายใน มีผลให้เขตเชื่อมั่นร่วม (Joint Confidence Region)[1] ของ (α, β) กว้างกว่าในกรณีเมื่องานพยากรณ์นั้นมีช่วงเชื่อมั่นที่แคบกว่า ความน่าเชื่อถือในค่าของ α และ β จึงต่ำกว่า



ภาพ 2.10 แสดงการเชื่อมโยงระหว่างช่วงพยากรณ์ และเขตเชื่อมั่นของ (α, β) จากภาพ (ก) แสดงให้เห็นว่า เขตเชื่อมั่นของ (α, β) เมื่อ $X_0 = \bar{X}$ จะเป็นเขตที่แคบที่สุด ขณะที่เมื่อ X_0 เบนห่างจาก $\bar{X}$ เขตเชื่อมั่นจะกว้างขึ้น เนื่องจากทุกจุดในเขตเชื่อมั่นมีสิทธิ์ที่จะเป็นค่าที่แท้จริงของ (α, β) ได้ด้วยกันทั้งหมด เขตที่แคบกว่าจึงน่าเชื่อถือกว่า ถ้างานใดมีช่วงเชื่อมั่น (ภาพ (ข)) แคบกว่าเขตเชื่อมั่นของ (α, β) จะแคบกว่างานวิเคราะห์ย่อมน่าเชื่อถือกว่า

[1] จะกล่าวถึง Confidence Contour ในตอน 3.2.5.2

2.6.2 การพยากรณ์ค่า Y_F (Individnal Value)[1]

บางครั้งนักวิจัยอาจสนใจที่จะพยากรณ์เฉพาะค่าเดี่ยว ๆ ของ Y เมื่อกำหนดให้ $X = X_0$ เท่านั้น ซึ่งแม้การพยากรณ์ในประเด็นนี้จะมีขอบเขตของประโยชน์ไม่สูงพอเช่นในการพยากรณ์ $E(Y_0|X_0)$ แต่ก็ยังคงมีประโยชน์อยู่ในแง่ที่เราสามารถใช้เป็นดัชนีชี้ว่า สมการยังคงรูปของ Linear Relation อยู่เช่นเดิม (Sample Period) หรือไม่ด้วยการทดสอบ สมมุติฐาน

$$H_0 : Y_F = \hat{Y}_F \text{ VS } H_1 : Y_F \gtrless \hat{Y}_F$$

เมื่อ Y_F คือค่าจริงของ Y เมื่อ $X = X_F$ และ $\hat{Y}_F$ คือค่าพยากรณ์ของ Y_F

ให้ $Y = \alpha+\beta X+u$ เป็นความสัมพันธ์ (True Relationship) ระหว่างตัวแปร Y และ X ดังนั้นเมื่อกำหนดให้ $X = X_F$ ค่าเดี่ยว (True Value) ของ Y_F จึงปรากฏดังนี้

$$Y_F = \alpha+\beta X_F+u_F$$

และเมื่อประมาณค่า α และ β ด้วย $\hat{\alpha}$ และ $\hat{\beta}$ ตามลำดับ จะพบว่า $\hat{Y}_F = \hat{\alpha}+\hat{\beta}X_F$ ซึ่งแสดงให้เห็นว่า Point Estimate ของ Y_F เมื่อ $X = X_F$ ก็คือ $\hat{Y}_F = \hat{\alpha}+\hat{\beta}X_F$[2]

สำหรับ Prediction Error และ Prediction Variance สามารถพิสูจน์ให้เห็นได้ดังนี้

เนื่องจาก $Y_F = \alpha+\beta X_F+u_F$ และ $\hat{Y}_F = \hat{\alpha}+\hat{\beta}X_F$ ดังนั้น Prediction Error ก็คือ

$$\begin{aligned}\hat{Y}_F - Y_F &= (\hat{\alpha}+\hat{\beta}X_F) - (\alpha+\beta X_F+u_F)\\ &= (\hat{\alpha}-\alpha)+(\hat{\beta}-\beta)X_F - u_F \qquad \text{..........(1)}\end{aligned}$$

จาก (1) พบว่า

$$E(\hat{Y}_F - Y_F) = E(\hat{\alpha}-\alpha)+X_F E(\hat{\beta}-\beta) - E(u_F) = 0$$

แสดงว่า $E(\hat{Y}_F) = Y_F$ หรือ $\hat{Y}_F = \hat{\alpha}+\hat{\beta}X_F$ เป็น Unbiased Estimator ของ Y_F ดังนั้น Prediction Variance ของ $\hat{Y}_F$ คือ

$$V(\hat{Y}_F) = E(\hat{Y}_F - Y_F)^2 = E\{(\hat{\alpha}-\alpha)+X_F(\hat{\beta}-\beta) - u_F\}^2$$

---

[1] ในที่นี้ใช้ Subscript "F" เพื่อป้องกันความสับสนกับกรณีที่ผ่านมาในตอน 6.2.1

[2] การพิสูจน์ $\hat{Y}_F = \hat{\alpha}+\hat{\beta}X_F$ เป็น BLUE ของ Y_F จะขอเว้นไว้เป็นแบบฝึกหัด แนวการพิสูจน์ให้ยึดถือวิธีที่แสดงไว้แล้วในตอน 2.6.1

$$= E(\hat{\alpha} - \alpha)^2 + X_F^2 E(\hat{\beta} - \beta)^2 + E(u_F^2) + 2X_F E\{(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)\}$$

$$-2E\{u_F(\hat{\alpha} - a)\} - 2X_F E\{u_F(\hat{\beta} - \beta)\}$$

$$= \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\Sigma x_i^2}\right) + \sigma^2 \frac{X_F^2}{\Sigma x_i^2} + \sigma^2 + 2\sigma^2 X_F \left(\frac{-\bar{X}}{\Sigma x_i^2}\right) - 0 - 0^1$$

$$= \sigma^2 + \sigma^2 \left\{ \frac{1}{n} + \frac{(X_F - \bar{X})^2}{\Sigma x_i^2} \right\}$$

นั่นคือ $$V(\hat{Y}_F) = \sigma^2 \left\{ 1 + \frac{1}{n} + \frac{(X_F - \bar{X})^2}{\Sigma x_i^2} \right\}$$

จะเห็นได้ว่า $V(\hat{Y}_F) > V(\hat{Y}_0)$ แสดงว่า ช่วงพยากรณ์ของ Y_F กว้างกว่าช่วงพยากรณ์ของ Y_0 สำหรับช่วงพยากรณ์ของ Y_F สามารถหาได้โดยอาศัย Sampling Distribution ของ Y_F ดังนี้

จาก $\hat{Y}_F = \hat{\alpha} + \hat{\beta} X_F$ ซึ่งแสดงว่า $\hat{Y}_F$ เป็น Linear Function ของ $\hat{\alpha}$ และ $\hat{\beta}$

แต่ $\hat{\alpha} \sim N\left(\alpha, \sigma^2 \frac{\Sigma X_i^2}{n\Sigma x_i^2}\right)$ และ $\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\Sigma x_i^2}\right)$ ดังนั้น โดยอาศัย mgf-technique เราสามารถพิสูจน์ได้ว่า

$$\hat{Y}_F \sim N\left[Y_F, \sigma^2 \left\{ 1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma x_i^2} \right\} \right]$$

และสามารถพิสูจน์ได้ว่า

$$\frac{\hat{Y}_F - Y_F}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma x_i^2}}} \sim t_{n-2} \quad \text{เมื่อ } \hat{\sigma}^2 = \frac{\Sigma e_i^2}{n-2}$$

ก. ช่วงเชื่อมั่น หรือช่วงพยากรณ์ $(1 - \alpha)$ 100% ที่คาดว่าค่าจริงคือ Y_F จะปรากฏอยู่คือ

$$\left(\hat{Y}_F + t_{n-2, \frac{\alpha}{2}} \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma x_i^2}} \, , \, \hat{Y}_F + t_{n-2, 1-\frac{\alpha}{2}} \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma x_i^2}} \right)$$

ข. ปฏิเสธสมมุติฐานหลัก $(H_0 : Y_F = \hat{Y}_F \text{ VS } H_1 : Y_F \neq \hat{Y}_F)$ เมื่อ $|t_c| > t_{n-2, 1-\frac{\alpha}{2}}$

โดยที่ $$t_c = (\hat{Y}_F - Y_F)/\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma x_i^2}}$$

---

[1] $E\{u_F(\hat{\beta} - \beta)\} = E(u_F\hat{\beta}) - \beta E(u_F) = E(u_F\hat{\beta}) = E\{u_F(\beta + \Sigma w_i u_i)\} = \beta E(u_F) + \{w_1 E(u_1 u_F) + w_2 E(u_2 u_F) + \ldots + w_n E(u_n u_F)\} = 0$

ในทำนองเดียวกันเราสามารถพิสูจน์ได้ว่า $E\{u_F(\hat{\alpha} - \alpha)\} = 0$

แบบฝึกหัดบทที่ 2

1. ถ้า z_i เป็นตัวแปรสุ่มที่มีการแจกแจงแบบเดียวกันที่มีความแปรปรวนเท่ากับ σ^2 และเป็นอิสระต่อกัน

ถ้า Y_i สัมพันธ์กับ X_i และ z_i ในรูป $Y_i = \alpha + \beta X_i + z_i$; $i = 1, 2, ..., n$ และ X_i มีค่าดังนี้ คือ $X_1 = 1, X_2 = 2, X_3 = 3, X_4 = 4, X_5 = 5, X_6 = 6$ ซึ่งพบว่าค่าประมาณของ β ที่ได้มาโดยวิธีอื่นที่มิใช่ OLS คือ

$$\hat{\beta}. = \frac{1}{8}[(Y_6+Y_5)-(Y_2+Y_1)]$$

จงหา $V(\hat{\beta}.)$ พร้อมทั้งเปรียบเทียบค่าที่ได้กับ $V(\hat{\beta})$ ที่ได้จากสูตร $V(\hat{\beta}) = \frac{\sigma^2}{\sum n^2}$

2. จากคู่ลำดับชุดหนึ่งพบว่า $\Sigma X = 11.34$ $\Sigma Y = 20.72$ $\Sigma X^2 = 12.16$ $\Sigma Y^2 + 84.96$ $\Sigma XY = 22.13$

จงประมาณค่าสมการถดถอย $Y = f(X)$ และ $X = f(Y)$ พร้อมทั้ง $V(\hat{\alpha})$ และ $V(\hat{\beta})$

3. ถ้าสหสัมพันธ์ระหว่าง X กับ Y คือ r จงหาสหสัมพันธ์ระหว่าง aX+b กับ cY+d เมื่อ a,b,c,d เป็นตัวคงที่ใด ๆ

4. นักวิจัยทำการบันทึกร้อยละของไขมัน (X) และร้อยละของสารอื่นที่มิใช่ไขมัน (Y) จากนมโค โดยบันทึกข้อมูลจากฟาร์มโคนม 2 แห่งโดยมีความประสงค์จะเปรียบเทียบว่าความชันของสมการถดถอย $Y = f(X)$ จากฟาร์มโคนมทั้งสองต่างกันหรือไม่ จงดำเนินการทดสอบ ข้อมูลปรากฏดังนี้

ฟาร์มโคนมที่ 1

$\Sigma x = 51.13$ $\Sigma y = 117.25$ $\Sigma x^2 = 1.27$ $\Sigma y^2 = 4.78$ $\Sigma xy = 1.84$

ฟาร์มโคนมที่ 2

$\Sigma x = 37.20$ $\Sigma y = 78.75$ $\Sigma x^2 = 1.03$ $\Sigma y^2 = 2.48$ $\Sigma xy = 1.10$.

5. ถ้า $u = ax+by$ และ $v = bx-ay$ โดยที่ $x = X-\bar{X}$ และ $y = Y-\bar{Y}$ ทั้งนี้ X และ Y มีสหสัมพันธ์เท่ากับ r แต่ u และ v ไม่มีสหสัมพันธ์ต่อกัน จงแสดงให้เห็นว่า

$$s_u s_v = (a^2+b^2)\, s_x s_y \sqrt{1-r^2}$$

6. จากข้อมูลชุดหนึ่งเราวิเคราะห์สมการถดถอยในรูปที่ข้อมูลถูกปรับด้วยค่าเฉลี่ย (deviation from mean) ของสมการ $y = f(x)$ และ $x = f(y)$ ได้เป็น $\hat{y} = 1.2\,x$ และ $\hat{x} = 0.6\,y$ จงคำนวณหา r_{xy} และ s_x/s_y

และถ้า $y = x+v$ จงคำนวณหา r_{vx} , r_{vy} และ s_v/s_y

<u>คำแนะนำ</u> เปลี่ยน $\hat{\beta}$ ให้อยู่ในรูปของ r กับ s_x, s_y และจาก $y = x+v$ แสดงว่า $v = y-x$ จากนั้นนำ y คูณตลอด แล้วหา r_{vy} และนำ x คูณตลอดแล้วหา r_{vx}

7. กำหนดให้ข้อมูลดังนี้

ตัวอย่างที่	ขนาดตัวอย่าง (n)	$\bar{x}$	$\bar{y}$	s_x	s_y	r_{xy}
1	600	5	12	2	3	0.6
2	400	7	10	3	4	0.7

จงหาสหสัมพันธ์ระหว่าง X กับ Y โดยรวม 2 ตัวอย่างเป็นตัวอย่างเดียวกัน

8. จากข้อมูลต่อไปนี้

	ปีที่											
	1	2	3	4	5	6	7	8	9	10	11	12
จำนวนการตาย ของทารก (100 คน)	60	62	61	55	53	60	63	53	52	48	49	43
ยอดจำหน่าย เบียร์ (ล้านลิตร)	23	23	25	25	26	26	29	30	30	32	33	31

จงวิเคราะห์สหสัมพันธ์ระหว่างจำนวนการตายของทารกกับยอดจำหน่ายเบียร์ จากนั้นให้กำจัดปัจจัยแนวโน้มออกจากอนุกรมใดอนุกรมหนึ่งแล้วคำนวณหาค่าสหสัมพันธ์อีกครั้ง

คำแนะนำ การกำจัดอิทธิพลของแนวโน้มกระทำได้โดยการหาผลต่างของข้อมูลคือ $W_t = X_t - X_{t-1}$; $t=2,3,\ldots n$ เราจะได้อนุกรม W ที่ปลอดจากปัจจัยแนวโน้ม

9. จากการบันทึกข้อมูลรายได้และการออมของครอบครัวต่าง ๆ จำนวน 5 ครอบครัว ปรากฏข้อมูลดังนี้

ครอบครัวที่	รายได้ (Y)	เงินออม (X)
1	8,000	600
2	11,000	1,200
3	9,000	1,000
4	6,000	700
5	6,000	300

1) จงพล็อตกราฟและประมาณค่าสมการถดถอย เงินออม = f(รายได้)

2) ประมาณค่าสมการถดถอย การบริโภค = f (รายได้)

10. บริษัท 4 บริษัทต่อไปนี้มีเงินกำไร และค่าใช้จ่ายเพื่อการวิจัยและพัฒนาดังต่อไปนี้ (หน่วยเงิน = ล้านบาท)

บริษัทที่	กำไร (P)	ค่าใช้จ่ายเพื่อ วิจัยและพัฒนา (R&D)
1	50	40
2	60	40
3	40	30
4	50	50

1) จงประมาณค่าสมการถดถอย P = f (R&D)

2) สมการที่ได้บ่งชี้หรือไม่ว่าค่าวิจัยและพัฒนามีผลต่อกำไรของบริษัท อภิปรายผล

11. จากข้อมูลในข้อ 9

1) จงหาช่วงเชื่อมั่น 95% ของ $\underline{\beta}$

2) จากข้อมูลในข้อ 9 สมมุติฐานใดต่อไปนี้ที่ไม่อาจยอมรับได้

(1) $\beta = 0$

(2) $\beta = 0.5$

(3) $\beta = 0.1$

(4) $\beta = -1$

3) จงทดสอบว่าเงินออมไม่ขึ้นอยู่กับรายได้ โดยมีสมมุติฐานรองว่าเงินออมเพิ่มขึ้นตามรายได้

4) จงหาช่วงพยากรณ์สำหรับเงินออมเฉพาะครอบครัวที่มีรายได้เท่ากับ

(1) 6,000

(2) 8,000

(3) 10,000

(4) 12,000

แล้วสรุปว่าช่วงใดใน 4 ช่วงข้างต้นแม่นยำสูงสุด และช่วงใดแม่นยำต่ำสุด และถ้าใช้รายได้เฉลี่ยและเงินออมเฉลี่ยมาวิเคราะห์จะปรากฏผลเช่นใด จงเปรียบเทียบ

12. ถ้าเราประสงค์จะวิเคราะห์สมการถดถอย การลงทุน = f(อัตราดอกเบี้ย) ท่านควรบันทึกข้อมูลเมื่อใดระหว่างเมื่อรัฐควบคุมอัตราดอกเบี้ยให้คงที่กับเมื่อรัฐปล่อยให้อัตราดอกเบี้ยลอยตัว