บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การใช้เทคนิคการวิเคราะห์ความถดถอยในงานวิจัยส่วนใหญ่ มักจะเลือกใช้วิธีกำลังสองต่ำสุด เนื่องจากเป็นวิธีการที่ใช้ง่าย และเป็นวิธีที่มีประสิทธิภาพภายใต้ข้อตกลงเบื้องต้นของการวิเคราะห์ความถดถอย แต่ในการวิเคราะห์ข้อมูลเพื่อจะนำไปสู่การสรุปผลที่ถูกต้องนั้น สิ่งสำคัญเบื้องต้น ส่วนหนึ่งขึ้นอยู่กับข้อมูลที่นำมาวิเคราะห์ ถ้าหากข้อมูลที่นำมาวิเคราะห์นั้นเป็นข้อมูลที่ไม่สมบูรณ์ (Incomplete Data) คือ การที่ไม่ทราบค่าสังเกตของตัวแปรที่สนใจ หรือค่าสังเกตนั้นมีลักษณะเป็นข้อมูลที่ถูกตัดทิ้ง (Censored Data) ย่อมทำให้เกิดผลกระทบจากการนำข้อมูลลักษณะดังกล่าวมาใช้เพื่อวิเคราะห์และสรุปผล เช่น จะทำให้ได้ตัวประมาณที่เอนเอียง นอกจากนี้ยังมีผลกระทบอันเนื่องมาจากจำนวนของข้อมูลที่ถูกตัดทิ้ง คือ เมื่อจำนวนข้อมูลที่ถูกตัดทิ้งมีจำนวนมากขึ้น ความแม่นยำจากการประมาณค่าย่อมจะลดน้อยลง ดังนั้นจะเกิดความคลาดเคลื่อนจากการประมาณมากขึ้น

กรณีที่ข้อมูลถูกตัดทิ้งบางส่วนนั้น คือการที่ไม่ทราบค่าที่แท้จริงต้องบันทึกค่าเท่าที่สังเกตได้ ตัวอย่างเช่น ข้อมูลความเสียหายทางด้านประกันภัย สมมติกำหนดจำนวนเงินประกันไว้สูงสุด 100,000 บาท เพราะฉะนั้น ในกรณีที่ค่าเสียหายจริงสูงกว่า 100,000 บาท บริษัทจะจ่าย 100,000 บาท และบันทึกตัวเลข 100,000 บาท เราเรียกตัวเลข 100,000 บาท ที่บันทึกไว้นี้ว่าข้อมูลถูกตัดทิ้ง ซึ่งจะเห็นได้ว่าตัวเลขข้อมูลนี้ไม่ใช่ตัวเลขแสดงค่าเสียหายที่เป็นจริง ข้อมูลที่ถูกตัดทิ้งบางส่วนอาจจะเกิดขึ้นได้จากหลายสาเหตุ ซึ่งนอกจากตัวอย่างที่กล่าวมาแล้วอาจมีกรณีอื่น ๆ อีก เช่น ไม่สามารถบันทึกค่าจริงของข้อมูลได้ อันเนื่องมาจากได้ทำการทดลองเพียงช่วงระยะเวลาหนึ่ง เมื่อหยุดทำการทดลองจึงไม่สามารถทราบค่าจริงของข้อมูลได้ ข้อมูลในลักษณะถูกตัดทิ้งนี้พบได้บ่อยในด้านการแพทย์ การทดลองด้านอุตสาหกรรม และการประกันภัย

ลักษณะข้อมูลที่ถูกตัดทิ้งบางส่วนอาจเกิดได้หลายรูปแบบ เช่น จะเกิดข้อมูลถูกตัดทิ้งทางซ้าย (Left Censoring) ที่ C ถ้าไม่ทราบค่าที่แท้จริงของค่าสังเกตเฉพาะค่าที่น้อยกว่า C หรือข้อมูลถูกตัดทิ้งทางขวา (Right Censoring) ที่ C ถ้าไม่ทราบค่าที่แท้จริงของค่าสังเกตเฉพาะค่าที่มากกว่า C

ในการวิจัยครั้งนี้สนใจที่จะศึกษา กรณีที่ตัวแปรตามในสมการถดถอยเชิงเส้นอย่างง่ายมีค่าถูกตัดทิ้งทางขวาเท่านั้น การทดลองหรือการศึกษาที่ทำให้ตัวแปรตามมีค่าถูกตัดทิ้งทางขวา เช่น การทดลองเกี่ยวกับความทนทาน หรืออายุการใช้งานของฉนวนกันความร้อนว่าจะขึ้นอยู่กับอุณหภูมิหรือความร้อนที่ได้รับหรือไม่ ในการทดลองนี้ตัวแปรตาม คือ อายุการใช้งานของฉนวน และตัวแปรอิสระ คือ อุณหภูมิหรือความร้อน โดยให้อุณหภูมิหรือความร้อนแก่ฉนวน แล้วบันทึกเวลา หรือจำนวนชั่วโมงที่ฉนวนนั้นจะเสื่อมสภาพ ในระหว่างการทดลองฉนวนอันที่มีการเสื่อมสภาพ จะเป็นข้อมูลที่ไม่ถูกตัดทิ้ง (Uncensored Data) เมื่อสิ้นสุดการทดลอง ฉนวนที่ยังคงอยู่ในสภาพใช้งานได้ดีจะเป็นฉนวนที่ไม่ทราบอายุการใช้งานที่แท้จริง เพราะฉะนั้นตัวเลขอายุการใช้งานที่บันทึกไว้เมื่อสิ้นสุดการทดลองจะเป็นข้อมูลที่ถูกตัดทิ้งทางขวา

ตัวอย่างข้อมูลที่ถูกตัดทิ้งทางขวาในการทดลองทางด้านการแพทย์ เช่น ในโครงการทดลองเปลี่ยนหัวใจคนไข้[1] Stanford Heart Transplantation Program เมื่อต้องการทดลองว่าจำนวนวันที่มีชีวิตอยู่รอดของคนไข้หลังการผ่าตัดเปลี่ยนหัวใจ ขึ้นอยู่กับอายุของคนไข้ที่เข้ารับการรักษาหรือไม่ ในการทดลองนี้ตัวแปรตาม คือ จำนวนวันที่มีชีวิตอยู่รอดของคนไข้หลังการผ่าตัด และตัวแปรอิสระ คืออายุของคนไข้ที่เข้ารับการรักษา โดยเริ่มทดลองเมื่อ 1 ตุลาคม ค.ศ.1967 และสิ้นสุดการทดลอง ในวันที่ 1 เมษายน ค.ศ. 1974 ในระหว่างการทดลองมีคนไข้ 69 คน เข้ารับการเปลี่ยนหัวใจ และสังเกตว่าคนไข้จะมีชีวิตอยู่รอดหลังจากเปลี่ยนหัวใจแล้ว โดยนับเป็นจำนวนวันที่มีชีวิตอยู่รอด ดังนั้นคนไข้ที่เสียชีวิตในระยะเวลาของการทดลองจะเป็นข้อมูลที่ไม่ที่ถูกตัดทิ้ง แต่คนไข้ที่มีชีวิตอยู่รอดเมื่อสิ้นสุดการทดลองจะไม่ทราบอายุที่มีชีวิตอยู่รอดหลังการเปลี่ยนหัวใจที่แน่นอนได้ จะทราบก็แต่อายุที่อยู่รอดเมื่อสิ้นสุดการทดลองเท่านั้น ข้อมูลส่วนนี้จะเป็นข้อมูลที่ถูกตัดทิ้งทางขวา

---

[1] Rupert G. Miller. "Least Squares Regression with Censored Data." Biometrika (1976),63,3,456-458

ทางด้านการประกันภัย เช่น การประกันภัยรถยนต์ ข้อมูลค่าเสียหาย (Claim) ให้เป็นตัวแปรตาม สำหรับตัวแปรอิสระ เช่น อายุคนขับ, ขนาดซีซีของรถ, ประเภทรถ และขนาดบรรทุก เป็นต้น ปกติข้อมูลค่าเสียหายทางด้านการประกันภัยจะจัดระดับค่าความเสียหายที่มีหน่วยเป็นบาท และจะกำหนดจำนวนเงินค่าเสียหายที่บริษัทรับประกันสูงสุดไม่เกิน C_k ถ้า T เป็นจำนวนเงินค่าเสียหายจริงมีค่ามากกว่าจำนวนเงินสูงสุดที่ต้องรับผิดชอบตามกรมธรรม์ คือ $T > C_k$ บริษัทรับประกันจะจ่ายเงินค่าเสียหายจำนวน C_k เท่านั้น และบันทึกการจ่ายเงินค่าเสียหายจำนวน C_k ในกรณีนี้บริษัทรับประกันจะไม่ทราบจำนวนเงินค่าเสียหายส่วนที่เกินกว่าจำนวนเงินที่ต้องรับผิดชอบตามกรมธรรม์ที่แน่นอนได้ ข้อมูลลักษณะเช่นนี้เป็นข้อมูลที่ถูกตัดทิ้งทางขวา

เนื่องจากข้อมูลส่วนที่ถูกตัดทิ้ง มีความสำคัญที่จะต้องนำมาใช้วิเคราะห์เพื่อวางแผนหรือปรับปรุงแผน และเพื่อการตัดสินใจ จึงควรทำการประมาณค่าที่ถูกตัดทิ้ง โดยใช้วิธีการที่มีความเหมาะสมจะได้ค่าประมาณที่มีค่าใกล้เคียงกับค่าจริง และมีจำนวนข้อมูลเพียงพอที่จะนำไปศึกษาตามวัตถุประสงค์ เป็นต้นว่า เพื่อสร้างตัวแบบที่เหมาะสม หรือเพื่อศึกษาถึงการแจกแจงของตัวแปรตาม ตัวอย่างทางด้านการประกันภัย เมื่อทำการประมาณจำนวนเงินค่าเสียหาย T ที่มากกว่า C_k จะมีข้อมูลค่าความเสียหายส่วนที่ได้จากการประมาณของค่าที่ถูกตัดทิ้ง ข้อมูลเหล่านี้รวมทั้งข้อมูลส่วนที่ไม่ถูกตัดทิ้ง สามารถนำข้อมูลทั้งหมดมาศึกษารูปแบบการแจกแจงของค่าเสียหาย (Loss Distribution) และอัตราค่าความเสียหาย (Loss Ratio) ซึ่งข้อมูลเหล่านี้จะให้ประโยชน์ในการสร้างตารางอัตราค่าเบี้ยประกันภัย เป็นต้น

ในการวิจัยครั้งนี้จะประมาณค่าพารามิเตอร์ β_i, i=0,1 จากสมการถดถอยเชิงเส้นอย่างง่าย โดยใช้วิเคราะห์กับข้อมูลที่ค่าสังเกตของตัวแปรตามมีค่าถูกตัดทิ้ง ซึ่งการนำข้อมูลที่มีค่าถูกตัดทิ้งมาใช้ประมาณค่าพารามิเตอร์นั้น อาจจะกระทำได้ในลักษณะต่อไปนี้

กรณีแรก จะถือว่าค่าที่ถูกตัดทิ้งเสมือนเป็นค่าที่ไม่ถูกตัดทิ้ง แล้วจึงทำการวิเคราะห์จากข้อมูลทั้งหมด

กรณีที่สอง ไม่สนใจข้อมูลส่วนที่เป็นค่าถูกตัดทิ้ง นั่นคือ จะทำการวิเคราะห์ข้อมูลเฉพาะกับข้อมูลที่ไม่ถูกตัดทิ้งเท่านั้น ดังนั้น จำนวนข้อมูลที่นำมาวิเคราะห์จึงมีขนาดตัวอย่างน้อยกว่ากรณีแรก

จากทั้งสองกรณีดังกล่าว เมื่อประมาณค่าพารามิเตอร์ ด้วยวิธีกำลังสองต่ำสุดจะทำให้ได้ตัวประมาณที่เอนเอียง และโดยเฉลี่ยการประมาณค่าจะต่ำกว่าความเป็นจริง หรือจะทำให้ได้ช่วงความเชื่อมั่นแคบกว่าความเป็นจริง[2] ดังนั้น ในการประมาณค่าพารามิเตอร์ กรณีที่ข้อมูลของตัวแปรตามมีค่าที่ถูกตัดทิ้งนั้น ควรที่จะศึกษาวิธีการอื่น ๆ ที่มีประสิทธิภาพมากกว่าวิธีกำลังสองต่ำสุด นั่นคือ สนใจที่จะศึกษาวิธีการประมาณด้วยภาวะน่าจะเป็นสูงสุด และวิธีการของบัคเลย์และเจมส์ และใช้วิเคราะห์กับข้อมูลกรณีแรก คือ จะถือว่าค่าสังเกตที่ถูกตัดทิ้งเสมือนค่าสังเกตที่ไม่ถูกตัดทิ้ง

1.2 วัตถุประสงค์ของการวิจัย

เพื่อเปรียบเทียบความถูกต้องของการประมาณค่าตัวแปรตาม ในสมการถดถอยเชิงเส้นอย่างง่าย เมื่อค่าสังเกตของตัวแปรตามเป็นค่าถูกตัดทางขวา โดยประมาณค่าพารามิเตอร์ด้วยวิธี

1.2.1 วิธีกำลังสองต่ำสุด (Ordinary Least Squares Method)

1.2.2 วิธีการประมาณด้วยภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimation Method)

1.2.3 วิธีการของบัคเลย์และเจมส์ (Buckley and James Method)

1.3 สมมติฐานทางการวิจัย

การประมาณค่าตัวแปรตาม ในสมการถดถอยเชิงเส้นอย่างง่าย เมื่อตัวแปรตามมีค่าที่ถูกตัดทิ้งทางขวา ในแต่ละวิธีจะให้ค่าประมาณที่แตกต่างกัน วิธีการประมาณด้วยภาวะน่าจะเป็นสูงสุด จะให้ค่าประมาณตัวแปรตามที่ใกล้เคียงกับค่าจริงมากกว่าวิธีกำลังสองต่ำสุด และวิธีการของบัคเลย์และเจมส์

---

[2] Josef Schmee and Gerald J. Hahn. "A Simple Method for Regression Analysis with Censored Data." Technometrics 21(4);(1979): 417-418.

1.4 ข้อตกลงเบื้องต้น

1.4.1 ศึกษารูปแบบสมการถดถอยอย่างง่าย มีรูปแบบดังนี้

$$T_i = \beta_0 + \beta_1 X_i + \varepsilon_i \ , \ i = 1,2,\ldots,n+m$$

T_i เป็นตัวแปรตาม

X_i เป็นตัวแปรอิสระ

β_i เป็นพารามิเตอร์ที่ไม่ทราบค่า , $i = 0,1$

ε_i เป็นค่าความคลาดเคลื่อนสุ่ม

1.4.2 การแจกแจงของค่าที่ถูกตัด และค่าที่ไม่ถูกตัด เป็นอิสระต่อกัน

1.4.3 ตัวแปรตาม (Dependent Variable) เท่านั้นที่เป็นค่าถูกตัดทิ้ง

1.4.4 ศึกษากรณีของประเภทค่าที่ถูกตัดทิ้งเป็นการตัดทิ้งทางขวา

1.5 ขอบเขตของการวิจัย

1.5.1 ในการวิจัยครั้งนี้จะทำการเปรียบเทียบการประมาณค่าตัวแปรตาม ในสมการถดถอยเชิงเส้นอย่างง่าย เมื่อค่าสังเกตของตัวแปรตามเป็นค่าถูกตัดทางขวา โดยประมาณค่าพารามิเตอร์ด้วยวิธี

1.5.1.1 วิธีกำลังสองต่ำสุด

1.5.1.2 วิธีการประมาณด้วยภาวะน่าจะเป็นสูงสุด

1.5.1.3 วิธีการของบัคเลย์และเจมส์

1.5.2 ศึกษาเมื่อกรณีของค่าที่ไม่ถูกตัดทิ้ง T_i มีรูปแบบเป็น

$$T_i = \beta_0 + \beta_1 X_i + \varepsilon_i \ , \ i = 1,\ldots,n$$

กำหนดค่าของพารามิเตอร์ ค่าของ X_i และ ε_i เป็นค่าใด ๆ ซึ่งในการวิจัยครั้งนี้กำหนดพารามิเตอร์ $\beta_0 = 2$, $\beta_1 = 1$, X_i มีการแจกแจงแบบปกติ มีค่าเฉลี่ยเป็น 20 มีค่าความแปร

ปรวนเป็น 60 ε_i มีการแจกแจงแบบปกติ มีค่าเฉลี่ยเป็น 0 มีค่าความแปรปรวนเป็น 16 ทุกเงื่อนไขการวิจัย[3]

1.5.3 ค่าที่ถูกตัดทิ้ง $c_1, c_2, \ldots$ เป็นอิสระกัน และมีการแจกแจงเป็น

1.5.3.1 การแจกแจงแบบสม่ำเสมอ (Uniform Distribution) U(a,b) :

$$f(c) = \frac{1}{b-a} \quad , a < c < b$$

1.5.3.2 การแจกแจงแบบปกติตัดปลายทางซ้าย (Left-Truncated Normal Distribution) กำหนดค่า μ, σ^2 เพื่อให้เกิดค่าที่ถูกตัดทิ้งตามขอบเขตที่กำหนด

$$f(c^*) = \frac{f(c)}{1-F(c_o)} \quad , c_o = a$$

$$f(c) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1\,(c-\mu)^2}{2\sigma^2}\right) \quad , c > a$$

1.5.3.3 การแจกแจงแบบไวบูลล์ตัดปลายทางซ้าย (Left-Truncated Weibull Distribution) กำหนดค่า α, β เพื่อให้เกิดค่าที่ถูกตัดทิ้งตามขอบเขตที่กำหนด

$$f(c^*) = \frac{f(c)}{1-F(c_o)} \quad , c_o = a$$

$$f(c) = \alpha\beta\, c^{\beta-1} \exp(-\alpha c^{\beta}) \quad , c > a$$

1.5.3.4 เมื่อค่าถูกตัดทิ้งเป็นฟังก์ชันเชิงเส้น ในรูปแบบเดียวกันกับ T_i นั่นคือ $C_i = \beta_o + \beta_1 X_i + \varepsilon_i$, $i = 1,\ldots,m$, m คือ จำนวนค่าสังเกตที่ที่ถูกตัดทิ้ง $\beta_o = 2$, $\beta_1 = 1$, X_i มีการแจกแจงแบบปกติ มีค่าเฉลี่ยเป็น 20 มีค่าความแปรปรวนเป็น 60 และ ε_i มีการแจกแจงแบบปกติ มีค่าเฉลี่ยเป็น 0 มีค่าความแปรปรวนเป็น 25

---

[3] ผู้วิจัยได้ทดลองด้วยค่าพารามิเตอร์ ค่าเฉลี่ย และความแปรปรวน หลาย ๆ ค่า พบว่าผลสรุปไม่แตกต่างกัน

1.5.4 ศึกษาเมื่อกรณีสัดส่วนของข้อมูลที่ถูกตัดทิ้งเป็น 10%, 20%, 30% และ 40% ของขนาดตัวอย่าง

1.5.5 ศึกษาเมื่อกรณีขนาดตัวอย่างเป็น 5 ระดับคือ 10, 15, 30, 50 และ 70

การวิจัยครั้งนี้จำลองข้อมูลให้มีสถานการณ์ตามที่ต้องการศึกษาโดยใช้เทคนิคการจำลองแบบมอนติ คาร์โล (Monte Carlo Simulation Technique) จากเครื่องคอมพิวเตอร์ AMDAHL 5860 เขียนโปรแกรมด้วยภาษาฟอร์แทรน (FORTRAN 77) ทำการจำลองข้อมูลซ้ำ ๆ กัน จำนวน 1,000 ครั้ง ในแต่ละสถานการณ์

1.6 เกณฑ์การตัดสินใจ

เกณฑ์การตัดสินใจว่าการประมาณค่าพารามิเตอร์ด้วยวิธีใดใช้ได้ดีกว่าจะพิจารณาโดยการเปรียบเทียบค่าความคลาดเคลื่อนระหว่างค่าประมาณของตัวแปรตาม กับค่าจริง ในรูปของค่ารากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง, RMSE (The Square Root of Mean Squares Error) วิธีการใดให้ค่า RMSE ต่ำกว่าจะเป็นวิธีการประมาณที่ดีกว่า

1.7 ประโยชน์ที่คาดว่าจะได้รับ

1.7.1 เพื่อเป็นแนวทางให้นักวิจัยมีผลสรุปและหลักฐานในการเลือกวิธีการประมาณประมาณค่าพารามิเตอร์ในสมการถดถอยเชิงเส้นอย่างง่าย เมื่อค่าสังเกตของตัวแปรตามมีค่าที่ถูกตัดทางขวา จะทำให้ได้ค่าประมาณของตัวแปรตามที่ใกล้เคียงกับค่าจริง

1.7.2 เพื่อเป็นแนวทางในการศึกษา และเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ในสมการถดถอยเชิงเส้นอย่างง่าย เมื่อมีการตัดค่าทางขวาของตัวแปรตาม ในสถานการณ์อื่น ๆ ต่อไป เช่น เมื่อเป็นสมการถดถอยเชิงเส้นพหุคูณ