บทที่ 1

บทนำ

1.1 ความสำคัญและที่มาของการวิจัย

การวิเคราะห์การถดถอย (Regression analysis) เป็นการศึกษาลักษณะของความสัมพันธ์ระหว่างตัวแปรสองกลุ่ม คือตัวแปรที่ต้องการศึกษา หรือตัวแปรตาม (Dependent variable) กับตัวแปรที่แทนปัจจัยต่าง ๆ ที่มีผลกระทบกับตัวแปรตาม เรียกว่าตัวแปรอิสระ (Independent variable) โดยลักษณะความสัมพันธ์นี้สามารถเขียนในรูปของฟังก์ชัน หรือตัวแบบการถดถอย (Regression model) ตัวแบบการถดถอยนอกจากใช้อธิบายความผันแปรของตัวแปรตามอันเนื่องมาจากการเปลี่ยนแปลงของตัวแปรอิสระแล้ว ยังสามารถใช้พยากรณ์ค่าของตัวแปรตามเมื่อทราบค่าของตัวแปรอิสระในตัวแบบได้อีกด้วย

ในการวิเคราะห์การถดถอยเชิงเส้นนั้น ตัวแปรตามจะต้องเป็นตัวแปรสุ่มชนิดต่อเนื่อง และมีการแจกแจงปรกติ ส่วนตัวแปรอิสระจะเป็นตัวแปรเชิงปริมาณ หรือเชิงคุณภาพก็ได้ ในกรณีที่ตัวแปรตามเป็นจำนวนนับ หรือความถี่ ซึ่งเป็นตัวแปรไม่ต่อเนื่อง ถ้าใช้การวิเคราะห์การถดถอยเชิงเส้น จะทำให้เกิดปัญหา 2 ประการ คือค่าคาดหวังของตัวแปรตามอาจจะมีค่าน้อยกว่าศูนย์ ซึ่งไม่เป็นไปตามข้อสมมติเบื้องต้นที่ว่า ค่าคาดหวังของจำนวนนับจะต้องมากกว่าหรือเท่ากับศูนย์ และข้อมูลมีความแปรปรวนของค่าคลาดเคลื่อนไม่คงที่ ซึ่งไม่เป็นไปตามข้อสมมติเบื้องต้นที่ว่า ความแปรปรวนของค่าคลาดเคลื่อนคงที่ทุก ๆ ระดับของตัวแปรอิสระ (Gardner et al., 1995)

เป็นที่ทราบว่าตัวแปรตามที่เป็นจำนวนของเหตุการณ์หรือลักษณะที่สนใจ (ซึ่งในวิทยานิพนธ์ฉบับนี้ขอเรียกสั้น ๆ ว่าจำนวนนับ หรือข้อมูลแบบนับ) ที่เกิดขึ้นในช่วงเวลา หรือขอบเขตที่กำหนดจะมีการแจกแจงปัวซง (Poisson distribution) ดังนั้นความสัมพันธ์ระหว่างตัวแปรตามนี้กับตัวแปรอิสระ จึงสามารถอธิบายได้ด้วยการวิเคราะห์การถดถอยปัวซง (Poisson regression analysis) อย่างไรก็ตามตัวแปรเชิงสุ่มปัวซงมีข้อสมมติเบื้องต้นที่ต้องพิจารณาคือ จำนวนของเหตุการณ์หรือลักษณะที่สนใจที่เกิดขึ้นใน 1 หน่วยเวลา หรือในขอบเขตหนึ่งจะต้องเป็นอิสระกับจำนวนของเหตุการณ์หรือลักษณะที่สนใจที่เกิดขึ้นใน 1 หน่วยเวลาหรือในขอบเขตอื่น และความแปรปรวนมีค่าเท่ากับค่าเฉลี่ย แต่ในทางปฏิบัติข้อสมมติเบื้องต้นเกี่ยวกับการเท่ากันของค่าความแปรปรวนและค่าเฉลี่ยมักไม่เกิดขึ้น ถ้าพบว่าข้อมูลที่เป็นจำนวนนับให้ค่าความแปรปรวนมากกว่า

ค่าเฉลี่ยแสดงว่าเกิดปรากฏการณ์ที่เรียกว่า Overdispersion และเรียกข้อมูลที่เป็นจำนวนนับที่มีลักษณะดังกล่าวว่า Overdispersed counts แต่ถ้าข้อมูลมีค่าความแปรปรวนน้อยกว่าค่าเฉลี่ย จะเกิดลักษณะที่เรียกว่า Underdispersion แต่ในสถานการณ์จริง Underdispersion จะเกิดขึ้นน้อยมาก (McCullagh and Nelder, 1989) ดังนั้นในวิทยานิพนธ์ฉบับนี้เราจะศึกษากระบวนการวิเคราะห์ข้อมูลแบบนับที่มีลักษณะ Overdispersion เท่านั้น

Hinde and Demètrio (1998) ได้อธิบายถึงสาเหตุที่เป็นไปได้ที่ทำให้เกิด Overdispersion ในจำนวนนับ ดังนี้

1) การจัดตัวอย่างแบบเกาะกลุ่ม (Cluster sampling) จะแบ่งประชากรออกเป็นกลุ่มย่อย ๆ โดยที่ในแต่ละกลุ่มย่อยจะมีคุณลักษณะคล้ายกัน และภายในกลุ่มเดียวกันจะมีคุณลักษณะที่แตกต่างกัน ถ้าการแบ่งกลุ่มไม่เหมาะสม อาจทำให้เกิดความผันแปรระหว่างกลุ่มซึ่งทำให้เกิดความแปรปรวนมาก

2) จำนวนนับที่เกิดขึ้นใน 1 หน่วยเวลา หรือในขอบเขตหนึ่งไม่เป็นอิสระกับจำนวนของเหตุการณ์หรือลักษณะที่สนใจที่เกิดขึ้นใน 1 หน่วยเวลาหรือในขอบเขตอื่น ทำให้เกิดความผันแปรระหว่างตัวแปรตาม

3) การรวมกลุ่มระหว่างระดับของตัวแปร ทำให้เกิดการแจกแจงแบบผสม (Compound distributions)

Hinde and Demètrio (1998) ได้กล่าวถึงผลเสียที่จะเกิดขึ้นในกรณีที่เราไม่นำ Overdispersion มาพิจารณาในการวิเคราะห์ข้อมูล คือ

1) สัมประสิทธิ์ของการถดถอยปัวซงมีค่าคลาดเคลื่อนมาตรฐาน (Standard error) ต่ำกว่าความเป็นจริง ทำให้ตัดสินใจเลือกตัวแปรอิสระเข้าตัวแบบไม่ถูกต้อง

2) ส่งผลให้การเปลี่ยนของค่า Deviance ที่เกี่ยวข้องกับการนำตัวแปรอิสระเข้าหรือออกจากตัวแบบมีค่ามากเกินไป อาจทำให้เลือกตัวแบบที่มีความซับซ้อนมากเกินไป

นักสถิติหลายท่านได้เสนอตัวแบบสำหรับประมาณค่า Overdispersion ในการถดถอยปัวซง ซึ่ง Poortema (1999) ได้รวบรวมไว้ แต่ไม่ได้เสนอวิธีการหาค่าประมาณ หรือวิธีการอนุมานเชิงสถิติ (ทดสอบสมมติฐาน และการประมาณค่าแบบช่วง) ของพารามิเตอร์ที่เกี่ยวข้อง อย่างไรก็ตามตัวแบบเชิงพาราเมตริกที่นิยมใช้มากก็คือ ตัวแบบการถดถอยทวินามแบบลบ (Negative Binomial Regression Model: NB)

การแจกแจงทวินามแบบลบมีหลายรูปแบบตามลักษณะของความสัมพันธ์ระหว่างค่าเฉลี่ย และความแปรปรวน แต่รูปแบบที่นิยมกันทั่ว ๆ ไปมากที่สุดสำหรับวิเคราะห์ข้อมูลดังกล่าวได้แก่ ตัวแบบทวินามแบบลบที่มีความแปรปรวนเป็นฟังก์ชันเชิงเส้นของค่าเฉลี่ย (Linear mean –

variance Negative Binomial model) หรือเรียกย่อ ๆ ว่า ตัวแบบ NB1 และตัวแบบทวินามแบบลบที่มีความแปรปรวนเป็นฟังก์ชันพหุนามกำลังสองของค่าเฉลี่ย (Quadratic mean – variance Negative Binomial model) หรือตัวแบบทวินามแบบลบที่มีความแปรปรวนเป็นฟังก์ชันกำลังสองค่าเฉลี่ย และเรียกย่อ ๆ ว่า ตัวแบบ NB2 ในวิทยานิพนธ์ฉบับนี้จะเน้นศึกษากระบวนการวิเคราะห์ข้อมูลทวินามแบบลบที่มีความแปรปรวนเป็นฟังก์ชันกำลังสองค่าเฉลี่ยเท่านั้น

1.2 วัตถุประสงค์

1. เพื่อศึกษาแนวคิดเชิงทฤษฎีของการวิเคราะห์ข้อมูล NB2
2. เพื่อศึกษากระบวนการของการวิเคราะห์ข้อมูล NB2
3. เพื่อศึกษาการประยุกต์ใช้ตัวแบบ NB2
4. เพื่อศึกษารูปแบบ และคุณสมบัติของ Robust variance ของตัวประมาณค่าพารามิเตอร์ของตัวแบบ NB2

1.3 ขอบเขตและวิธีการดำเนินการวิจัย

1. ศึกษาคุณลักษณะ รวบรวมทฤษฎีบท ตลอดจนคุณสมบัติต่าง ๆ ที่สำคัญที่เกี่ยวข้องกับตัวแบบเชิงเส้นวางนัยทั่วไป (Generalize Linear Models: glms) และการวิเคราะห์การถดถอย NB2 เช่น การหาตัวประมาณค่าพารามิเตอร์ด้วยวิธีความควรจะเป็นสูงสุด (Maximum Likelihood Method) โดยใช้ วิธีนิวตัน – ราฟสัน (Newton – Raphson) การเลือกตัวแบบ (Model Selection) โดยใช้ Akaike Information Criteria และการวินิจฉัยตัวแบบ (Model Diagnostics) โดยใช้ Half normal plot with simulated envelope

2. ศึกษาแบบทดสอบเพื่อเปรียบเทียบระหว่างการถดถอยปัวซง กับการถดถอย NB2 ได้แก่ Wald test, Likelihood ratio test และ Score tests และตรวจสอบคุณสมบัติ Score test ที่เสนอโดย Dean (1992) และโดย Wang – Shu Lu (1997) โดยใช้ Simulation study

3. พัฒนาสูตรของ Robust variance ของค่าประมาณสัมประสิทธิ์ตัวทำนายเชิงเส้น (Linear predictor coefficient Estimates) ของ NB2 และศึกษาคุณสมบัติของ Robust variance ที่ได้โดยใช้ Simulation study

1.4 ระยะเวลาการดำเนินงาน

กันยายน พ.ศ.2548 – พฤษภาคม พ.ศ.2550

ตารางที่ 1.1 แผนการดำเนินการวิจัย

ขั้นตอนการดำเนินงาน	ระยะเวลาดำเนินการวิจัย																			
	พ.ศ.2548				พ.ศ.2549												พ.ศ.2550			
	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5
1. เลือกหัวข้อวิทยานิพนธ์	●	●																			
2. ทบทวนวรรณกรรม	●	●	●	●	●																
3. จัดทำโครงร่างวิทยานิพนธ์				●	●	●															
4. เสนอโครงร่างวิทยานิพนธ์							●														
5. เตรียมเครื่องมือ NB2 fit							●	●													
6. เขียนชุดคำสั่งของ Half normal pot with simulated envelope ของ Poisson และ NB2							●	●	●	●	●										
8. ศึกษาคุณสมบัติ Score test ที่เสนอโดย Dean (1992) และ Wang – Shu Lu (1997) โดยใช้ Simulation study											●	●	●	●							
9. เขียนสูตรในการคำนวณ Robust variance ของ NB2													●	●	●						
10. ศึกษาเปรียบเทียบ Robust variance กับ Asymptotic variance โดย Simulation study														●	●	●	●	●			
11. เตรียมต้นฉบับวิทยานิพนธ์																		●	●	●	
12. สอบวิทยานิพนธ์																				●	
13. แก้ไขและส่งรายงานวิทยานิพนธ์ฉบับสมบูรณ์																				●	●

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. เข้าใจถึงแนวคิด ทฤษฎี และการประยุกต์ใช้ตัวแบบ NB2
2. เป็นแนวทางศึกษา และใช้เป็นเอกสารอ้างอิงสำหรับนักศึกษา และผู้สนใจที่จะศึกษาค้นคว้าเกี่ยวกับคุณลักษณะต่าง ๆ ของตัวแบบ NB2
3. ทำให้ทราบถึงกระบวนการที่จะดำเนินการศึกษาวิจัย
4. เสนอ Robust variance ของตัวแบบ NB2 แก่วรรณกรรมที่เกี่ยวกับการวิเคราะห์ข้อมูลจำนวนนับ