บทที่ 1

บทนำ

1.1. ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันการพยากรณ์ได้เข้าไปมีบทบาทในการวิจัยสาขาต่าง ๆ และวิธีที่นิยมใช้กันมากที่สุดวิธีหนึ่งในการพยากรณ์ก็คือ การวิเคราะห์การถดถอย (regression analysis) ซึ่งเป็นวิธีการทางสถิติที่ใช้ในการศึกษาถึงความสัมพันธ์ของตัวแปร 2 กลุ่ม กลุ่มหนึ่งคือ ตัวแปรอิสระ (independent variables) ซึ่งอาจจะเป็นตัวแปรเชิงปริมาณหรือตัวแปรเชิงคุณภาพก็ได้ และอีกกลุ่มหนึ่งคือ ตัวแปรตาม (dependent variable) ซึ่งส่วนมากจะพบว่าเป็นตัวแปรเชิงปริมาณ แต่ก็มีปัญหาหลายปัญหาที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพ เช่น ปัญหาทางด้านการแพทย์ ทางการศึกษา ทางสังคมศาสตร์ และชีวสถิติ เป็นต้น ซึ่งปัญหาต่าง ๆ ดังกล่าวข้างต้น ตัวแปรตามมักจะเป็นตัวแปรเชิงคุณภาพที่มีค่าเป็นไปได้ 2 ค่า เช่น สำเร็จ-ล้มเหลว เป็นโรค-ไม่เป็นโรค ตาย-ไม่ตาย และซื้อ-ไม่ซื้อ เป็นต้น จะเรียกตัวแปรตามในลักษณะนี้ว่า ตัวแปรตามทวิ (dichotomous dependent variable)

การวิเคราะห์การถดถอยที่มีตัวแปรตามเป็นตัวแปรทวิ และขึ้นอยู่กับตัวแปรอิสระ ซึ่งในที่นี้จะเรียกว่า ตัวแปรอธิบาย (explanatory variable) จะยกตัวอย่างดังนี้คือ ในเรื่องของการศึกษาเกี่ยวกับปริมาณสารพิษที่เป็นอันตรายต่อชีวิตเมื่อสะสมอยู่ในร่างกาย โดยอาจจะทำการทดลองกับสัตว์ทดลอง จะมีตัวแปรอธิบาย X หนึ่งตัวและมี Y เป็นตัวแปรตามทวิ

ให้ X_j แทนค่าของตัวแปรอธิบายX ณ ระดับที่ j ; $j = 1,2,\ldots,k$
ในตัวอย่างนี้คือปริมาณสารพิษที่สะสมอยู่ในร่างกายในระดับที่ j เมื่อแบ่งระดับของปริมาณสารพิษออกเป็น k ระดับ

Y_{ij} แทนค่าของตัวแปรตามทวิ Y ค่าที่ i สำหรับ $X = X_j$; $i = 1,2,..,n$
ในตัวอย่างนี้คือหนูทดลองตัวที่ i ที่มีปริมาณสารพิษอยู่ในร่างกายในระดับที่ j มีค่าเป็นไปได้ 2 ค่าคือ ตายหรือไม่ตาย หรืออาจจะกำหนดดังนี้

$Y_{ij} = 1$ เมื่อหนูทดลองตัวที่ i มีปริมาณสารพิษอยู่ในร่างกายในระดับที่ j ตาย
$= 0$ เมื่อหนูทดลองตัวที่ i มีปริมาณสารพิษอยู่ในร่างกายในระดับที่ j ไม่ตาย

โดยที่ลักษณะของข้อมูลของตัวแปรอธิบาย X และตัวแปรตามทวิ Y จะเป็นดังนี้

ระดับของปริมาณสารพิษ	ผลที่เกิดกับหนูทดลอง
$X = X_1$	$Y_{11} = 0$ หรือ 1
	$Y_{21} = 0$ หรือ 1
	$\vdots$
	$Y_{n_1 1} = 0$ หรือ 1
$X = X_2$	$Y_{12} = 0$ หรือ 1
	$Y_{22} = 0$ หรือ 1
	$\vdots$
	$Y_{n_2 2} = 0$ หรือ 1
$\vdots$	
$X = X_k$	$Y_{1k} = 0$ หรือ 1
	$Y_{2k} = 0$ หรือ 1
	$\vdots$
	$Y_{n_k k} = 0$ หรือ 1

จากลักษณะของข้อมูลข้างต้นจะเห็นว่าถ้าทำการเก็บรวบรวมข้อมูล ค่าของตัวแปรตามทวิ Y ก็จะมีแต่เลข 0 กับเลข 1 ซึ่งในทางปฏิบัติมักจะไม่ทำการจดบันทึกข้อมูลในลักษณะนี้ แต่จะทำการจดบันทึกจำนวนสิ่งที่สนใจ ในที่นี้ก็คือจำนวนหนูทดลองที่ตายในแต่ละระดับของปริมาณสารพิษที่มีอยู่ในร่างกาย แล้วทำการหาค่าความน่าจะเป็นหรือค่าสัดส่วนของสิ่งที่สนใจดังนี้

ระดับของปริมาณสารพิษ	จำนวนหนูทดลองทั้งหมด	จำนวนหนูทดลองที่ตาย	ค่าสัดส่วนของหนูทดลองที่ตาย
X_1	n_1	r_1	$p_1 = \frac{r_1}{n_1}$
X_2	n_2	r_2	$p_2 = \frac{r_2}{n_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_k	n_k	r_k	$p_k = \frac{r_k}{n_k}$

ดังนั้นจากข้อมูลที่เก็บรวบรวมได้ แทนที่จะศึกษาความสัมพันธ์ระหว่างตัวแปรตามทวิ Y กับ ตัวแปรอธิบาย X ด้วยตัวแบบการถดถอยเชิงเส้น $Y = \beta_0 + \beta_1 X + \varepsilon$ ซึ่งมีข้อตกลง

เบื้องต้นว่า $E(\varepsilon)$ จะต้องมีค่าเท่ากับ 0 นั่นก็คือ $E(Y|X)=\beta_0+\beta_1X$ ก็จะทำการศึกษาความสัมพันธ์ระหว่างค่าสัดส่วนของสิ่งที่สนใจ (p_j) ซึ่งก็คือค่าเฉลี่ยของ Y เมื่อกำหนด X ($E(Y|X)$) กับตัวแปรอธิบาย X แต่เนื่องจากค่า p_j มีค่าอยู่ระหว่าง (0,1) ซึ่งแตกต่างจากตัวแบบถดถอยเชิงเส้นที่ค่าของ $E(Y|X)$ จะมีค่าอยู่ในช่วง $(-\infty,\infty)$ ดังนั้นถ้าจะสร้างความสัมพันธ์ระหว่างค่า p_j กับ X_j ในรูปของ $p_j = \beta_0+\beta_1X_j$ นั้นก็ควรจะมีการแปลงค่าของ p_j ให้มีค่าอยู่ในช่วง $(-\infty,\infty)$ ซึ่งมีผู้เสนอวิธีการแปลงไว้หลายวิธี เช่น การแปลงแบบโพบิท (probit transformation) การแปลงแบบโลจิสติก (logistic transformation) เป็นต้น สำหรับในการวิจัยครั้งนี้จะศึกษาเฉพาะกรณีการแปลงแบบโลจิสติก ซึ่งเป็นแบบที่นิยมกันมาก เมื่อทำการแปลงค่าของ p_j โดยการแปลงแบบโลจิสติกจะเป็นดังนี้

$$\ln\left(\frac{p_j}{1-p_j}\right) = \beta_0+\beta_1X_j$$

$$\frac{p_j}{1-p_j} = e^{\beta_0+\beta_1X_j}$$

$$p_j = (1-p_j)e^{\beta_0+\beta_1X_j}$$

$$p_j + p_j\, e^{\beta_0+\beta_1X_j} = e^{\beta_0+\beta_1X_j}$$

$$p_j(1+e^{\beta_0+\beta_1X_j}) = e^{\beta_0+\beta_1X_j}$$

$$p_j = \frac{e^{\beta_0+\beta_1X_j}}{1+e^{\beta_0+\beta_1X_j}}$$

เรียกตัวแบบที่ได้ข้างต้นว่าตัวแบบถดถอยโลจิสติก (logistic regression model)

ในปัจจุบันตัวแบบถดถอยโลจิสติกน่าจะมีความสำคัญมากขึ้น เนื่องจากในวงการต่าง ๆ เช่น ในวงการธุรกิจมักจะไม่เปิดเผยรายละเอียดเท่าใดนัก และมักจะพอใจที่จะให้ข้อมูลเพียงใช่หรือไม่ใช่เท่านั้น ดังนั้นถ้าผู้วิจัยมีความเข้าใจและรู้เรื่องตัวแบบถดถอยโลจิสติกเป็นอย่างดีแล้วจะทำให้สามารถวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพ แต่อาจจะพบปัญหาในเรื่องของการประมาณค่าพารามิเตอร์ β_0 และ β_1 ซึ่งเป็นสิ่งสำคัญที่สุด ถ้าสามารถประมาณค่าพารามิเตอร์ได้ใกล้เคียงค่าจริงแล้วก็จะทำให้สามารถพยากรณ์ความน่าจะเป็นหรือค่าสัดส่วนของการเกิดสิ่งที่สนใจได้อย่างถูกต้องมากยิ่งขึ้น

ในการประมาณค่าพารามิเตอร์ β_0 และ β_1 แบบจุดของตัวแบบถดถอยโลจิสติกนั้น สามารถทำได้หลายวิธี และมีผู้ทำวิจัยหลายท่านได้ทำการศึกษาไว้แล้ว แต่ในการประมาณค่าแบบช่วงนั้นยังมีผู้ศึกษาไว้ไม่มากนัก ดังนั้นผู้วิจัยจึงสนใจที่จะทำการประมาณค่าพารามิเตอร์แบบช่วง โดยเน้นที่ค่าสัมประสิทธิ์การถดถอย ซึ่งจะเปรียบเทียบช่วงการประมาณที่ได้จากวิธีการประมาณค่าแบบจุด และวิธีการประมาณค่าความแปรปรวนของตัวประมาณค่าที่แตกต่างกัน เนื่องจากการประมาณค่าแบบช่วงนั้น จะขึ้นอยู่กับตัวประมาณค่าแบบจุด ความแปรปรวนของตัวประมาณค่า และการแจกแจงตัวอย่างของตัวประมาณค่า ซึ่งในการประมาณการแจกแจงตัวอย่างของตัวประมาณค่านั้นจะอาศัยการแจกแจงในลักษณะเดียวกับการแจกแจงแบบที

1.2 วัตถุประสงค์ของการวิจัย

เพื่อเปรียบเทียบช่วงความเชื่อมั่นสำหรับค่าสัมประสิทธิ์การถดถอยของตัวแบบถดถอยโลจิสติกว่า วิธีการประมาณค่าแบบจุดและวิธีการประมาณค่าความแปรปรวนของตัวประมาณค่าวิธีใดจะให้ช่วงการประมาณที่ครอบคลุมค่าพารามิเตอร์ และมีช่วงการประมาณค่าที่มีความยาวเฉลี่ยน้อยที่สุดในสถานการณ์ต่าง ๆ โดยจะทำการศึกษา 3 วิธีดังต่อไปนี้

วิธีที่ 1 ใช้วิธีกำลังสองน้อยที่สุดถ่วงน้ำหนัก (Weighted Least Squares Method) ในการประมาณค่าพารามิเตอร์แบบจุด และประมาณค่าความแปรปรวนของตัวประมาณ

วิธีที่ 2 ใช้วิธีความควรจะเป็นสูงสุด (Maximum Likelihood) ในการประมาณค่าพารามิเตอร์แบบจุด และใช้สารสนเทศของฟิชเชอร์ (Fisher Information) ในการประมาณค่าความแปรปรวนของตัวประมาณ

วิธีที่ 3 ใช้วิธีความควรจะเป็นสูงสุด ในการประมาณค่าพารามิเตอร์แบบจุด และใช้วิธีแจคไนฟ์ (Jacknife) ในการประมาณค่าความแปรปรวนของตัวประมาณ

1.3 ข้อตกลงเบื้องต้น

สำหรับขั้นตอนในการเปรียบเทียบค่าความยาวเฉลี่ยของช่วงความเชื่อมั่นที่ได้จากการจำลองนั้นจะเปรียบเทียบเฉพาะในกรณีที่ให้ระดับความเชื่อมั่นไม่ต่ำกว่าค่าสัมประสิทธิ์ความเชื่อมั่นที่กำหนดเท่านั้น

1.4 ขอบเขตของการวิจัย

1.4.1 ตัวแปรอธิบาย (X) ที่ใช้ในการวิจัยมี 1 ตัว แบ่งเป็นระดับต่าง ๆ k ระดับ โดยกำหนดให้ k เป็น 3 , 5 , 7 และ 10

1.4.2 จะทำการสร้างค่าระดับของตัวแปรอธิบาย (X) จากความสัมพันธ์ซึ่งประยุกต์จากที่ C.J.Swanepole และ C.C.Frangos ได้เสนอไว้ดังต่อไปนี้

$$X_j = X_{j-1} + \frac{1}{k-1} \quad ; \quad j = 2,3,\ldots,k$$

กำหนดให้ $X_1 = 0.1$

1.4.3 ในแต่ละระดับของตัวแปรอธิบาย จะใช้ขนาดตัวอย่างเท่ากับ n โดยกำหนดให้ n เป็น 15 , 20 , 30 , 40 และ 50

1.4.4 กำหนดค่าพารามิเตอร์เริ่มต้น $\beta_0 = 0$, $\beta_0 = 1.0$ และ $\beta_0 = -2.0$ ส่วนค่าของ β_1 นั้น ในแต่ละค่าของ β_0 จะทำการเปลี่ยนแปลงค่าของ β_1 จาก 1.0 เป็นค่าอื่น ๆ โดยการเพิ่มหรือลดครั้งละ 0.5

1.4.5 กำหนดค่าสัมประสิทธิ์ความเชื่อมั่น $(1-\alpha)100\%$ เป็น 90% , 95% และ 99%

1.4.6 กำหนดจำนวนครั้งสำหรับการประมาณการแจกแจงของตัวประมาณค่าเท่ากับ 1000 ครั้ง

1.4.7 ในการวิจัยครั้งนี้สร้างแบบการจำลองข้อมูล โดยใช้เทคนิคมอนติคาร์โลซิมูเลชั่น (Monte Carlo Simulation Technique) จากเครื่องคอมพิวเตอร์ด้วยโปรแกรมภาษาฟอร์แทรน 77 (FORTRAN 77) ทำการจำลองข้อมูลซ้ำ 500 รอบ ในแต่ละสถานการณ์ที่ทำการศึกษา

1.5 สมมติฐานของการวิจัย

ภายใต้ทุก ๆ สถานการณ์ที่ทำการวิจัย วิธีที่ทำการศึกษาทั้ง 3 วิธีจะให้ค่าสัมประสิทธิ์ความเชื่อมั่นไม่ต่ำกว่าระดับที่กำหนด คือ 0.90 , 0.95 และ 0.99 แต่ในวิธีที่ 2 คือประมาณค่าพารามิเตอร์แบบจุดด้วยวิธีความควรจะเป็นสูงสุด และประมาณค่าความแปรปรวนของตัวประมาณด้วยสารสนเทศของฟิชเชอร์ จะให้ค่าความยาวเฉลี่ยของช่วงความเชื่อมั่นน้อยที่สุด

1.6 คำจำกัดความ

1.6.1 สัมประสิทธิ์ความเชื่อมั่น (confidence coefficient) หมายถึง ความน่าจะเป็นที่ช่วงสุ่มจะครอบคลุมค่าของพารามิเตอร์ในประชากร

1.6.2 ช่วงความเชื่อมั่น (confidence interval) หมายถึง ช่วงตัวอย่างที่คำนวณจากข้อมูลตัวอย่างหนึ่งชุดใด ๆ ซึ่งใช้ในการประมาณค่าพารามิเตอร์แบบช่วง

1.7 ขั้นตอนการดำเนินการวิจัย

ขั้นตอนในการวิจัยมีดังนี้ คือ

1.7.1 สร้างข้อมูลที่จะใช้ในการวิจัย

1.7.2 ประมาณค่าพารามิเตอร์แบบจุด และความแปรปรวนของตัวประมาณค่าตามวิธีที่ต้องการศึกษาทั้ง 3 วิธี

1.7.3 ประมาณการแจกแจงตัวอย่างของตัวประมาณค่าและสร้างช่วงการประมาณค่า

1.7.4 คำนวณค่าระดับความเชื่อมั่นของช่วงความเชื่อมั่นจากการจำลอง

1.7.5 คำนวณค่าความยาวเฉลี่ยของช่วงความเชื่อมั่น

1.7.6 ทำการเปรียบเทียบช่วงความยาวเฉลี่ยของแต่ละวิธี ในแต่ละสถานการณ์

1.7.7 สรุปผลการวิจัยในแต่ละสถานการณ์

1.8 เกณฑ์การตัดสินใจ

ในการพิจารณาว่าวิธีที่ทำการศึกษา วิธีใดจะดีที่สุดนั้นจะแบ่งเป็น 2 ขั้นตอนดังนี้

1.8.1 ระดับความเชื่อมั่นของช่วงความเชื่อมั่น

ในการพิจารณาค่าระดับความเชื่อมั่นของช่วงความเชื่อมั่น เกณฑ์ในการพิจารณาว่าสัมประสิทธิ์ความเชื่อมั่นที่ได้จากการจำลองมีค่าไม่ต่ำกว่าค่าสัมประสิทธิ์ความเชื่อมั่นที่กำหนดจะอาศัยการทดสอบสมมติฐาน โดยใช้ตัวสถิติ Z มีรูปแบบดังนี้

$$H_0 \ : \ p \geq p^*$$

$$H_1 \ : \ p < p^*$$

$$-Z_{\alpha_0} < \frac{\hat{p} - p^*}{\sqrt{\frac{p^*(1-p^*)}{n}}} < 1$$

$$p^* - Z_{\alpha_0}\sqrt{\frac{p^*(1-p^*)}{n}} < \hat{p} < 1$$

เพราะฉะนั้น ช่วงในการยอมรับสมมติฐานหลัก คือ

$$\left(p^* - Z_{\alpha_0}\sqrt{\frac{p^*(1-p^*)}{n}} \ , \ 1 \right)$$

โดยที่ α_0 คือระดับนัยสำคัญ หรือความผิดพลาดประเภทที่ 1 (Type I error)ที่กำหนดในการทดสอบสมมติฐาน

p คือสัมประสิทธิ์ความเชื่อมั่นของช่วงความเชื่อมั่น

$\hat{p}$ คือสัมประสิทธิ์ความเชื่อมั่นของช่วงความเชื่อมั่นที่ได้จากการจำลอง

p^* คือสัมประสิทธิ์ความเชื่อมั่นของช่วงความเชื่อมั่นที่กำหนด (0.90 , 0.95 , 0.99)

เกณฑ์ในการพิจารณานั้นเราอาจจะพิจารณาจากค่าความผิดพลาดของการประมาณค่าแทนก็ได้ โดยจะพิจารณาว่าค่าความผิดพลาดของการประมาณค่านั้นมีค่าไม่มากกว่าค่าความผิดพลาดของการประมาณค่าที่กำหนดหรือไม่ ถ้าค่าความผิดพลาดของการประมาณค่ามีค่าไม่มากกว่าค่าความผิดพลาดของการประมาณที่กำหนดเราจะนำวิธีการนั้นไปคำนวณหาค่าความยาวเฉลี่ยของช่วงความเชื่อมั่นต่อไป ซึ่งจะได้ผลสรุปเช่นเดียวกัน โดยมีหลักการดังนี้

$$H_0 \ : \ \alpha \le \alpha^*$$
$$H_1 \ : \ \alpha > \alpha^*$$

$$0 < \frac{\hat{\alpha} - \alpha^*}{\sqrt{\frac{\alpha^*(1--\alpha^*)}{n}}} < Z_{\alpha_0}$$

$$0 < \hat{\alpha} < \alpha^* + Z_{\alpha_0}\sqrt{\frac{\alpha^*(1-\alpha^*)}{n}}$$

เพราะฉะนั้น ช่วงในการยอมรับสมมติฐานหลัก คือ

$$(0 \ , \ \alpha^* + Z_{\alpha_0}\sqrt{\frac{\alpha^*(1-\alpha^*)}{n}})$$

โดยที่ α_0 คือระดับนัยสำคัญ หรือความผิดพลาดประเภทที่ 1 (Type I error) ที่กำหนดในการทดสอบสมมติฐาน

α คือสัดส่วนที่ช่วงความเชื่อมั่นไม่ครอบคลุมค่าพารามิเตอร์

$\hat{\alpha}$ คือสัดส่วนของช่วงความเชื่อมั่นที่คำนวณได้ไม่ครอบคลุมค่าพารามิเตอร์จากการจำลองคือ $1-\hat{p}$

α^* คือสัดส่วนที่ช่วงความเชื่อมั่นไม่ครอบคลุมค่าพารามิเตอร์ที่กำหนด (0.10 , 0.05 , 0.01)

ในการวิจัยนี้จะพิจารณาเฉพาะสัมประสิทธิ์ความเชื่อมั่นที่ได้จากการจำลองมีค่าไม่ต่ำกว่าค่าสัมประสิทธิ์ความเชื่อมั่นที่กำหนดเท่านั้น

1.8.2 ความยาวเฉลี่ยของช่วงความเชื่อมั่น

สำหรับการเปรียบเทียบค่าความยาวเฉลี่ยของช่วงความเชื่อมั่นที่ได้จากการจำลองนั้นจะเปรียบเทียบเฉพาะในกรณีที่ให้ค่าระดับความเชื่อมั่นไม่ต่ำกว่าค่าสัมประสิทธิ์ความเชื่อมั่นที่กำหนดเท่านั้น

สำหรับรายละเอียดของเกณฑ์การตัดสินใจนั้นจะเสนอในบทที่ 2

1.9 ประโยชน์ของการวิจัย

1.9.1 เพื่อเป็นแนวทางในการเลือกวิธีการประมาณค่าพารามิเตอร์แบบจุด และค่าความแปรปรวนของตัวประมาณค่าในตัวแบบถดถอยโลจิสติกที่เหมาะสมสำหรับการประมาณค่าสัมประสิทธิ์ถดถอยแบบช่วงในแต่ละสถานการณ์

1.9.2 เพื่อเป็นแนวทางในการศึกษาเปรียบเทียบสำหรับตัวแบบอื่น ๆ ต่อไป