การเปรียบเทียบวิธีการประมาณสำหรับการวิเคราะห์การถดถอยเชิงเส้นพหุเมื่อตัวแปรตามและตัวแปรอิสระมีการสูญหายแบบนอนอิกนอร์เรเบิล

Comparison of The Estimation Methods for The Multiple Linear Regression Model with Nonignorable – Missing Dependent and Independent Variables

วริษฐา กณิกนันต์[1*] และ อนุภาพ สมบูรณ์สวัสดี[2]

Warittha Kaniknant and Anupap Somboonsavatdee

[1] นิสิตปริญญาโท สาขาวิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย

[2] อาจารย์ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย

Abstract

The objectives of this research were to compare the estimation methods for the multiple linear regression model with nonignorable-missing dependent and independent variables. The methods for estimating missing data are K-Nearest Neighbor (KNN) , EM Algorithm (EM) and Predictive Mean Matching (PMM) , three levels of missing proportion of data of 10%, 20%, 30% and three levels of nonignorable missingness of none, medium, high are studied from the simulation. Based on the size of average mean square error (AMSE), the best methods for estimating missing data are least AMSE. The findings are the followings : i) all estimation methods perform better as the sample size increases, ii) all estimation methods perform worse as the standard deviation of errors, the missing proportion, or level of nonignorable missingness increase, iii) KNN method performs best when the standard deviation of error is medium and high (30 and 90), iv) EM method performs best when the standard deviation of error is not high (10) except data of unequal variance case.

Keyword: ***Missing data*****,** ***multiple linear regression, nonignorable-missing***

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการประมาณข้อมูลสูญหายสำหรับการวิเคราะห์การถดถอยเชิงเส้นพหุ เมื่อตัวแปรตามและตัวแปรอิสระมีการสูญหายแบบนอนอิกนอร์เรเบิล วิธีการประมาณค่าสูญหายที่ใช้ในงานวิจัยนี้คือ วิธี K-Nearest Neighbor (KNN) วิธี EM Algorithm (EM) และวิธี Predictive Mean Matching (PMM) ซึ่งข้อมูลที่ใช้ในการศึกษาได้จากการจำลองโดยมีสัดส่วนของการสูญหาย 3 ระดับคือ 10% 20% และ 30% และมีระดับการสูญหายแบบนอนอิกนอร์เรเบิล 3 ระดับคือ ไม่มี ปานกลาง และสูง โดยจะทำการเปรียบเทียบประสิทธิภาพของแต่ละวิธีการด้วยค่าเฉลี่ย

ใหญ่ขึ้น ii) วิธีการประมาณทุกวิธีจะแย่ลงเมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน สัดส่วนของการสูญหาย และระดับการสูญหายแบบนอนอิกนอร์เรเบิลเพิ่มสูงขึ้น iii) วิธีการ KNN จะเป็นวิธีการประมาณที่ดีที่สุด เมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมีขนาดปานกลางและสูง (30 และ 90) iv) วิธีการ EM จะเป็นวิธีการที่ดีที่สุดเมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนน้อย (10) ยกเว้นในกรณีที่ข้อมูลตัวแปรอิสระมีความแปรปรวนเท่ากัน

คำสำคัญ: *ข้อมูลสูญหาย, การวิเคราะห์การถดถอยเชิงเส้นพหุ, การสูญหายแบบนอนอิกนอร์เรเบิล*

บทนำ

ข้อมูลสูญหายคือปัญหาหลักอย่างหนึ่งที่มักเกิดขึ้นได้เสมอจากการเก็บรวบรวมข้อมูลเพื่อทำการวิจัยในด้านต่างๆ ถ้าหากเพิกเฉยต่อปัญหาเหล่านี้แล้ว ย่อมส่งผลกระทบต่อการวิเคราะห์ข้อมูล ซึ่งอาจนำไปสู่ข้อสรุปที่ผิดพลาดได้ เทคนิคการพยากรณ์ที่นิยมใช้คือ การวิเคราะห์การถดถอยเชิงเส้นพหุ (Multiple Linear Regression) ซึ่งในการวิเคราะห์ข้อมูลจะประกอบไปด้วยตัวแปรอิสระ (Independent Variable) ตั้งแต่ 2 ตัวขึ้นไป และตัวแปรตาม (Dependent Variable) ซึ่งตัวแปรอิสระและตัวแปรตามจะมีความสัมพันธ์กันในรูปแบบเชิงเส้น ดังนั้น หากเกิดเหตุการณ์ที่ข้อมูลสูญหายทั้งในตัวแปรอิสระและตัวแปรตาม ก็ย่อมส่งผลให้เกิดปัญหาในการวิเคราะห์ขึ้นมาทันที วิธีการแก้ไขปัญหาดังกล่าว สามารถทำได้หลายวิธี ซึ่งวิธีการที่ง่ายที่สุดคือการตัดข้อมูลบางส่วนที่ไม่สมบูรณ์ทิ้งไป แต่วิธีการนี้จะทำให้สูญเสียสาระสำคัญบางอย่างที่อาจส่งผลต่อข้อสรุปก็เป็นได้ ดังนั้นจึงมีการคิดค้นวิธีการต่างๆ เพื่อนำมาใช้ในการประมาณค่าสูญหาย ทั้งนี้ความมีประสิทธิภาพของแต่ละวิธีการจะมากหรือน้อยก็ขึ้นอยู่กับความเหมาะสมของลักษณะข้อมูลที่จะนำไปใช้ด้วย ดังนั้นเราจึงสนใจที่จะศึกษาเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายแบบต่างๆ ที่เกิดขึ้นทั้งในตัวแปรอิสระและตัวแปรตาม ในกรณีที่มีการสูญหายแบบนอนอิกนอร์เรเบิลซึ่งเป็นการสูญหายที่เกิดขึ้นเมื่อความน่าจะเป็นของการสูญหายของตัวแปรนั้นไม่มีความสัมพันธ์กับตัวแปรอื่น แต่จะมีความสัมพันธ์กับตัวมันเอง

วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาวิธีการประมาณค่าสูญหายของตัวแปรอิสระและตัวแปรตามที่มีความสัมพันธ์กันอย่างมีเงื่อนไขในกรณีที่มีการสูญหายแบบ Nonignorable
2. เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายเมื่อตัวแปรอิสระและตัวแปรตามเมื่อมีการสูญหายแบบ Nonignorable ทั้ง 3 วิธี ได้แก่ วิธี K-Nearest Neighbor Imputation วิธี EM Algorithm และวิธี Predictive Mean Matching Imputation (PMM)

ข้อตกลงเบื้องต้น

1. การศึกษาในครั้งนี้จะสนใจกรณีที่ตัวแปรอิสระ (x) และตัวแปรตาม (y) มีความสัมพันธ์กันภายใต้การถดถอยเชิงเส้นพหุ (Multiple Linear Regression) ซึ่งมีรูปแบบดังนี้

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \qquad ;i = 1,2,...,m,m+1,...,n$$

เมื่อ y_i แทน ค่าสังเกตของตัวแปรตามของข้อมูลตัวที่ i

x_{ij} แทน ค่าสังเกตของข้อมูลตัวที่ i ของตัวแปรอิสระตัวที่ j เมื่อ j=1,2,3

β_p แทน สัมประสิทธิ์การถดถอยตัวที่ p เมื่อ p = 0,1,2,3

ε_i แทน ค่าความคลาดเคลื่อนของข้อมูลตัวที่ i

n แทน จำนวนค่าสังเกตทั้งหมด

m แทน จำนวนค่าสังเกตที่ทราบค่า

n - m แทน จำนวนค่าสังเกตที่สูญหาย

2. ความคลาดเคลื่อนเป็นตัวแปรสุ่มที่มีการแจกแจงแบบปกติ $\varepsilon_i = N(0, \sigma^2)$

3. $\varepsilon_i, \varepsilon_k$ ไม่มีสหสัมพันธ์กัน นั่นคือ $E(\varepsilon_i, \varepsilon_k) = 0$ เมื่อ $i \neq k$

4. การสูญหายของข้อมูลเกิดขึ้นที่ตัวแปรอิสระตัวใดตัวหนึ่ง และตัวแปรตาม อย่างมีความสัมพันธ์กันตามรูปแบบที่กำหนด โดยจะแบ่งช่วงของตัวแปรอิสระและตัวแปรตามออกเป็น 3 ช่วง และกำหนดให้ร้อยละของการสูญหายในแต่ละช่วงแตกต่างกันไป

ขอบเขตและวิธีการวิจัย

1. สร้างข้อมูลตัวแปรอิสระ (x_1, x_2, x_3) มีการแจกแจงแบบปกติ (Normal Distribution) ซึ่งมีฟังก์ชันการแจกแจงคือ

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2} \right), -\infty < x < \infty$$

ในการศึกษาครั้งนี้เราจะศึกษาการสูญหายของข้อมูลตัวแปรอิสระตัวใดตัวหนึ่งโดยคำนึงถึงระดับของความแปรปรวนที่มีขนาดเล็ก กลาง และใหญ่ ซึ่งจะแบ่งลักษณะการแจกแจงของตัวแปรอิสระออกเป็น 2 รูปแบบดังต่อไปนี้

1) $X_1 \sim N(0,300), X_2 \sim N(0,300)$ และ $X_3 \sim N(0,300)$ โดยที่ ศึกษาการสูญหายในกรณีที่ตัวแปรอิสระมีความแปรปรวนเท่ากัน

2) $X_1 \sim N(0,100), X_2 \sim N(0,300)$ และ $X_3 \sim N(0,500)$ โดยที่ ศึกษาการสูญหายในกรณีที่ตัวแปรอิสระมีความแปรปรวนแตกต่างกัน

โดยจะกำหนดให้ตัวแปรอิสระทั้ง 3 ตัวไม่มีความสัมพันธ์กันนั่นคือ มีค่าสหสัมพันธ์เท่ากับ 0

2. สร้างค่าความคลาดเคลื่อน (ε) มีการแจกแจงแบบปกติ ที่มีค่าเฉลี่ยเท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 10 30 และ 90 เมื่อพิจารณาจากค่าสัมประสิทธิ์ความแปรผัน (Coefficient of Variation) ที่ 75% 100% และ 225% ตามลำดับ

3. สร้างตัวแปรตามที่เกิดจากความสัมพันธ์ระหว่างตัวแปรอิสระภายใต้การถดถอยเชิงเส้นพหุ คือ

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \quad ; i = 1, 2, ..., m, m+1, ..., n$$

โดยกำหนดให้ $\beta_0 = 42$ และ $\beta_1 = \beta_2 = \beta_3 = 1$ เนื่องจากในการศึกษาครั้งนี้ต้องการเปรียบเทียบชุดข้อมูลของตัวแปรอิสระที่มีความแปรปรวนแตกต่างกัน ดั้งนั้นถ้าหากเปลี่ยนค่า β_1, β_2 และ β_3 จะส่งผลให้ตัวแปรตามที่ได้จากชุดข้อมูลของตัวแปรอิสระแต่ละชุดมีความแปรปรวนแตกต่างกันด้วย ดังนั้นเพื่อควบคุมความแปรปรวนของตัวแปรตามให้มีขนาดเท่ากัน จึงกำหนดให้ $\beta_1 = \beta_2 = \beta_3 = 1$

4. ในแต่ละสถานการณ์จะกำหนดขนาดตัวอย่าง 3 ขนาดคือ 50 100 และ 200
5. การสูญหายของข้อมูลที่เกิดขึ้นที่ตัวแปรตาม และตัวแปรอิสระตัวใดตัวหนึ่งเท่านั้น โดยจะทำการแบ่งตัวแปรอิสระ และตัวแปรตามออกเป็น 3 ช่วงด้วยอัตราส่วนเท่าๆ กันและให้แต่ละช่วงมีร้อยละของการสูญหายแตกต่างกัน โดยจะสร้างตัวแปรที่มีการแจกแจงแบบเบอร์นูลลี่ ด้วยความน่าจะเป็น 0.1 0.2 และ 0.3 เพื่อให้เกิดการสูญหายเฉลี่ยในช่วงต้น ช่วงกลาง และช่วงปลาย ตามลำดับ แล้วจะจับคู่กับข้อมูลตัวแปรอิสระและตัวแปรตาม
6. การสูญหายของข้อมูลเกิดขึ้นอย่างมีความสัมพันธ์กันระหว่างตัวแปรตามและตัวแปรอิสระ และเป็นการสูญหายแบบ Nonignorable กล่าวคือจะเกิดการสูญหายในตัวแปรใดตัวแปรหนึ่งแบบ Nonignorable ถ้าความน่าจะเป็นของการสูญหายของตัวแปรนั้นไม่มีความสัมพันธ์กับค่าของตัวแปรอื่นๆ แต่จะมีความสัมพันธ์กับค่าของตัวมันเอง

ซึ่งในการกำหนดสัดส่วนของการสูญหายจะแบ่งตัวแปรตามและตัวแปรอิสระออกเป็น 3 ช่วง และจะกำหนดให้สัดส่วนของการสูญหายของข้อมูลแตกต่างกันตามระดับของการสูญหายแบบ Nonignorable โดยจะกำหนดให้ช่วงของตัวแปรตามและตัวแปรอิสระที่มีค่ามากจะมีสัดส่วนของการสูญหายมากกว่าช่วงของตัวแปรตามและตัวแปรอิสระที่มีค่าน้อย ซึ่งจะส่งผลให้แต่ละช่วงมีความน่าจะเป็นในการสูญหายที่สูง-ต่ำ แตกต่างกันไป

ระดับการสูญหายแบบ Nonignorable จะแบ่งออกเป็น 3 ระดับคือ ไม่มี ปานกลาง และสูง ซึ่งในแต่ละช่วงจะมีอัตราส่วนของการสูญหายดังต่อไปนี้

ไม่มี	1 : 1 : 1
ปานกลาง	7 : 10 : 13
สูง	4 : 10 : 16

7. สร้างตัวแปรที่ทำให้เกิดการสูญหาย ซึ่งตัวแปรตามและตัวแปรอิสระจะมีการสูญหายอย่างมีความสัมพันธ์กันภายใต้ทฤษฎีความน่าจะเป็นแบบมีเงื่อนไข ความน่าจะเป็นของการสูญหายโดยเฉลี่ยจะกำหนดให้เท่ากับ 0.1 0.2 และ 0.3 โดยให้แต่ละช่วงของตัวแปรอิสระมีการสูญหายดังต่อไปนี้
8.

ความน่าจะเป็นของการสูญหายโดยเฉลี่ย	ระดับการสูญหายแบบ Nonignorable	ความน่าจะเป็นของการสูญหายในแต่ละช่วง			odd
		ช่วงต้น (%)	ช่วงกลาง (%)	ช่วงปลาย (%)	
0.1	ไม่มี	0.10 (10)	0.10 (10)	0.10 (10)	1

	ปานกลาง	0.07 (7)*	0.10 (10)	0.13 (13)	2
	สูง	0.04 (4)	0.10 (10)	0.16 (16)	4
0.2	ไม่มี	0.20 (20)	0.20 (20)	0.20 (20)	1
	ปานกลาง	0.14 (14)	0.20 (20)	0.26 (26)	2
	สูง	0.08 (8)	0.20 (20)	0.32 (32)	4
ความน่าจะเป็นของการสูญหายโดยเฉลี่ย	**ระดับการสูญหายแบบ Nonignorable**	**ความน่าจะเป็นของการสูญหายในแต่ละช่วง**			**odd**
		ช่วงต้น (%)	**ช่วงกลาง (%)**	**ช่วงปลาย (%)**	
0.3	ไม่มี	0.30 (30)	0.30 (30)	0.30 (30)	
	ปานกลาง	0.21 (21)	0.30 (30)	0.39 (39)	2
	สูง	0.12 (12)	0.30 (30)	0.48 (48)	4

เนื่องจากข้อมูลตัวแปรตามและตัวแปรอิสระมีการสูญหายอย่างมีความสัมพันธ์กันเราจึงต้องคำนวณหาความน่าจะเป็นของการสูญหายของตัวแปรตามในแต่ละช่วง ผลดังตารางต่อไปนี้

ความน่าจะเป็นของการสูญหายโดยเฉลี่ย	**ระดับการสูญหายแบบ Nonignorable**	**ความน่าจะเป็นของการสูญหายในแต่ละช่วง**			**odd**
		ช่วงต้น $P(y=1\|x=1), P(y=1\|x=0)$	**ช่วงกลาง** $P(y=1\|x=1), P(y=1\|x=0)$	**ช่วงปลาย** $P(y=1\|x=1), P(y=1\|x=0)$	
0.1	ไม่มี	$\frac{1}{10},\frac{1}{10}$	$\frac{1}{10},\frac{1}{10}$	$\frac{1}{10},\frac{1}{10}$	1
	ปานกลาง	$\frac{14}{107},\frac{7}{107}$	$\frac{2}{11},\frac{1}{11}$	$\frac{26}{113},\frac{13}{113}$	2
	สูง	$\frac{4}{28},\frac{1}{28}$	$\frac{4}{13},\frac{1}{13}$	$\frac{16}{37},\frac{4}{37}$	4
0.2	ไม่มี	$\frac{2}{10},\frac{2}{10}$	$\frac{2}{10},\frac{2}{10}$	$\frac{2}{10},\frac{2}{10}$	1
	ปานกลาง	$\frac{14}{57},\frac{7}{57}$	$\frac{2}{6},\frac{1}{6}$	$\frac{26}{63},\frac{13}{63}$	2
	สูง	$\frac{8}{31},\frac{2}{31}$	$\frac{4}{8},\frac{1}{8}$	$\frac{32}{49},\frac{8}{49}$	4
0.3	ไม่มี	$\frac{3}{10},\frac{3}{10}$	$\frac{3}{10},\frac{3}{10}$	$\frac{3}{10},\frac{3}{10}$	1
	ปานกลาง	$\frac{42}{121},\frac{21}{121}$	$\frac{6}{13},\frac{3}{13}$	$\frac{78}{139},\frac{39}{139}$	2
	สูง	$\frac{12}{34},\frac{3}{34}$	$\frac{12}{19},\frac{3}{19}$	$\frac{48}{61},\frac{12}{61}$	4

8. การศึกษาในครั้งนี้จะทำการจำลองข้อมูลภายใต้สถานการณ์ต่างๆ ที่เป็นไปตามเงื่อนไขข้างต้นที่แตกต่างกันโดยใช้เทคนิคการจำลองแบบมอนติคาร์โล (Monte Carlo Simulation Technique) ทำการจำลองในแต่ละสถานการณ์เป็นจำนวน 5,000 รอบ
9. ประมาณค่าข้อมูลเพื่อแทนที่ข้อมูลที่สูญหายในตัวแปรอิสระและตัวแปรตามด้วย วิธี K-Nearest Neighbor Imputation (KNN) , วิธี EM Algorithm และวิธี Predictive Mean Matching Imputation (PMM)

10. ประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุด้วยวิธีกำลังสองน้อยสุดแบบสามัญ (Ordinary Least Squares Method : OLS)
11. สร้างสมการถดถอยเชิงเส้นพหุจากค่าสัมประสิทธิ์การถดถอยเพื่อใช้ในการพยากรณ์
12. คำนวณหาค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี
13. สรุปผลการวิจัยที่ได้ในแต่ละสถานการณ์

เกณฑ์ที่ใช้ในการตัดสินใจ

เกณฑ์ที่ใช้ในการตัดสินใจว่าวิธีประมาณค่าสูญหายวิธีใดที่ให้ค่าประมาณใกล้เคียงกับค่าจริงมากที่สุดนั้นจะพิจารณาจากค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองระหว่างค่าพยากรณ์กับค่าจริง (Average mean square error : AMSE) ซึ่งวิธีการที่ให้ค่า AMSE ต่ำสุดจะเป็นวิธีการประมาณค่าสูญหายที่ดีที่สุด โดยสามารถคำนวณได้จากสูตรดังต่อไปนี้

$$MSE_t = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{ti})^2}{n}$$

$$AMSE = \frac{1}{5,000}\sum_{i=1}^{5,000} MSE_t$$

เมื่อ y_i แทน ค่าจริงของข้อมูลตัวแปรตามตัวที่ i

$\hat{y}_{ti}$ แทน ค่าพยากรณ์ของข้อมูลตัวแปรตามตัวที่ i จากการทำซ้ำรอบที่ t

MSE_t แทน ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าพยากรณ์ตัวแปรตามจากการทำซ้ำรอบที่ t

$AMSE$ แทน ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าพยากรณ์ตัวแปรตามจากการทำซ้ำทั้งหมด 5,000 รอบ

อภิปรายผล

ในการวิจัยครั้งนี้พบว่า กรณีที่ข้อมูลมีการกระจายน้อย การประมาณค่าสูญหายของข้อมูลด้วยวิธีการ EM จะให้ประสิทธิภาพดีกว่าวิธีการอื่นๆ แต่ถ้าข้อมูลมีการกระจายอยู่ในระดับปานกลาง และสูง วิธีการ KNN จะให้ประสิทธิภาพดีที่สุด

ถ้าหากพิจารณาค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) จะพบว่า เมื่อขนาดตัวอย่างใหญ่ขึ้นจะทำให้ค่า AMSE ของแต่ละวิธีลดลง เนื่องจากการเพิ่มขึ้นของขนาดตัวอย่างจะช่วยให้ค่าความคลาดเคลื่อนจากการพยากรณ์ลดลง ส่วนถ้าสัดส่วนของการสูญหาย ระดับของการสูญหายแบบ Nonignorable และส่วนเบี่ยงเบียนมาตรฐานของความคลาดเคลื่อนมีขนาดเพิ่มขึ้น จะทำให้ค่า AMSE ของทุกวิธีการเพิ่มสูงขึ้นด้วย เพราะถ้าหากข้อมูลมีการสูญหายมาก และมีการกระจายมากย่อมส่งผลให้ความค่าความคลาดเคลื่อนจากการพยากรณ์มากขึ้นด้วย

ข้อเสนอแนะ

จากงานวิจัยนี้ ผู้วิจัยได้ทำการศึกษากรณีเฉพาะบางกรณีเท่านั้น ซึ่งในความเป็นจริงแล้วปัญหาที่พบอาจจะอยู่นอกเหนือจากข้อสรุปของงานวิจัยชิ้นนี้ เช่น ในงานวิจัยชิ้นนี้ ถูกกำหนดให้เกิดการสูญหายที่ตัวแปรอิสระเพียงตัวใดตัวหนึ่งเท่านั้น แต่ในความเป็นจริงแล้วตัวแปรอิสระอาจสูญหายพร้อมๆ กันมากกว่าหนึ่งตัวก็เป็นไปได้ รวมทั้งยังควรศึกษาเพิ่มเติมในกรณีที่เกิดการสูญหายของชุดข้อมูลที่มีทั้งข้อมูลเชิงคุณภาพและข้อมูลเชิงปริมาณด้วย เพราะในงานวิจัยชิ้นนี้ศึกษาแต่การสูญหายของข้อมูลเชิงปริมาณเท่านั้น นอกจากนี้ยังควรศึกษาเพิ่มเติมในกรณีที่ข้อมูลเป็นข้อมูลอนุกรมเวลา เพื่อศึกษาว่าการเปลี่ยนแปลงของเวลาจะส่งผลกระทบต่อค่าสูญหายและวิธีการประมาณค่าสูญหายหรือไม่

บรรณานุกรม

ธีระพร วีระถาวร. ตัวแบบเชิงเส้น : ทฤษฎีและการประยุกต์. กรุงเทพมหานคร : พิทักษ์การพิมพ์, 2531.

ธีระพร วีระถาวร. ความน่าจะเป็นกับการประยุกต์. กรุงเทพมหานคร : นำอักษรการพิมพ์, 2537.

กัลยา วานิชย์บัญชา. การวิเคราะห์สถิติ : สถิติสำหรับการบริหารและงานวิจัย. กรุงเทพมหานคร : โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย, 2553.

อุษณีย์ วงศ์อามาตย์. การเปรียบเทียบวิธีการประมาณค่าสูญหายแบบนอนอิกนอร์เรเบิล ในการวิเคราะห์การถดถอยเชิงพหุ. วิทยานิพนธ์ ปริญญามหาบัณฑิต ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย, 2555.

เพียงออ ยีสา. การเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยเชิงเส้น. วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย, 2551.

วารุณี ตรีบำรุงศักดิ์. การพยากรณ์ด้วยวิธีการถดถอยเชิงเส้นพหุ เมื่อตัวแปรตามมีค่าสูญหาย. วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย, 2538.

Roderick J.A. Little and Donald B. Rubin. Statistical Analysis with Missing Data. Wiley,1987.

van Buuren, S., and Groothuis-Oudshoorn, K. Multivariate imputation by chained equations in r. Journal of Statistical Software 45 (December 2011): 1-67.