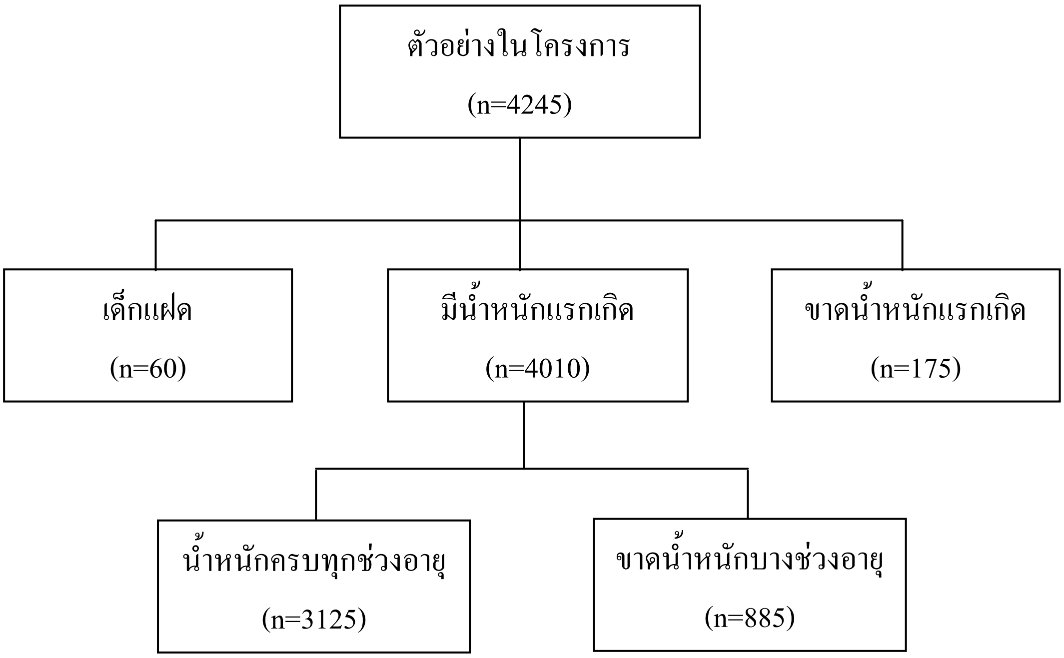
บทที่ 3

วิธีการวิจัย

3.1 ประชากรและตัวอย่าง

ประชากรในการศึกษาครั้งนี้ คือ เด็กไทยที่มีอายุแรกเกิดถึง 2 ปี ในช่วงปี พ.ศ. 2543-2544 โดยโครงการวิจัยระยะยาวในเด็กไทยทำการวิจัยแบบเฝ้าดูโดยไม่มีกระบวนการแทรกแซงตัวอย่าง แบบติดตามไปข้างหน้าในระยะยาว ศึกษาตัวแปรหลายระดับครอบคลุมประชากรเป้าหมายทั้งหมดในพื้นที่ศึกษา (Prospective Cohort Population-based Multilevel Study) การเลือกพื้นที่ศึกษาเป็นแบบเจาะจง โดยพิจารณาจากความเป็นไปได้ในการดำเนินงานระยะยาวต่อเนื่อง และ เป็นพื้นที่ที่มีลักษณะหลากหลายทั้งภูมิประเทศ ประชากร อาชีพ ทั้งนี้คัดเลือกพื้นที่ดำเนินการใน 5 แห่ง ประกอบด้วยพื้นที่ระดับอำเภอ 4 แห่งคือ 1) อำเภอพนมทวน จังหวัดกาญจนบุรี 2) อำเภอเทพา จังหวัดสงขลา 3) อำเภอกระนวน จังหวัดขอนแก่น 4) อำเภอเมือง จังหวัดน่าน ซึ่งมารดามาฝากครรภ์และคลอดที่โรงพยาบาลประจำอำเภอนั้นๆ 5) กรุงเทพมหานคร มีวิธีการเลือกตัวอย่างแตกต่างจากพื้นที่อื่น กล่าวคือ ครอบคลุมประชากรชาวกรุงเทพมหานครในทุกเขตซึ่งมารดามาฝากครรภ์และคลอดที่ โรงพยาบาลรามาธิบดี โรงพยาบาลราชวิถี และโรงพยาบาลพญาไท และสมัครใจเข้าร่วมโครงการ จากทั้ง 5 พื้นที่ มีจำนวนตัวอย่างในโครงการทั้งหมด 4,245 คน แต่ในการศึกษานี้ทำการวิเคราะห์เฉพาะเด็กที่เกิดเป็นลูกเดี่ยวที่มีชีวิตอยู่จนถึงอายุ 2 ปี ซึ่งในจำนวน 4,245 คนนี้เป็นเด็กแฝด 60 คน เด็กที่ไม่มีข้อมูลน้ำหนักแรกเกิด 175 คน และมีข้อมูลน้ำหนักแรกเกิด 4,010 คน ดังนั้นเด็กที่ค่าน้ำหนักเก็บได้ตามกำหนดเวลาครบทั้ง 5 จุด คือที่แรกเกิด อายุ 6 เดือน 1 ปี 1 ปีครึ่ง และ 2 ปี ตามลำดับ ซึ่งเป็นตัวอย่างที่นำมาใช้ในงานวิจัยครั้งนี้มีจำนวนทั้งสิ้น 3,125 คน ดังรายละเอียดในภาพที่ 3.1



ภาพที่ 3.1 แผนภาพแสดงขนาดตัวอย่าง

เมื่อเปรียบเทียบเด็กที่รวมและไม่รวมอยู่ในการวิเคราะห์ในการศึกษาครั้งนี้ พบว่าน้ำหนักแรกเกิดโดยเฉลี่ยของเด็กที่รวมอยู่ในการศึกษาครั้งนี้ (3,059.3 กรัม) และไม่รวมอยู่ในการศึกษาครั้งนี้ (3,048.7 กรัม) ทั้ง 2 กลุ่มไม่แตกต่างกัน และเมื่อทดสอบด้วย Pearson’s Chi-squared test พบว่า อายุครรภ์ และการกระจายของกลุ่มตัวอย่างในพื้นที่ศึกษาของเด็กทั้งสองกลุ่มต่างกันอย่างมีนัยสำคัญทางสถิติ โดยมีสัดส่วนของเด็กในพื้นที่กรุงเทพมหานครไม่ได้รวมอยู่ในการวิเคราะห์มากที่สุด (42.4%) และกลุ่มที่ไม่ได้รวมอยู่ในการวิเคราะห์มีเด็กที่คลอดก่อนกำหนด คือ อายุครรภ์น้อยกว่า 37 สัปดาห์ (11.2%) สูงกว่าอย่างมีนัยสำคัญทางสถิติ แสดงในตารางที่ 3.1

ตารางที่ 3.1 การเปรียบเทียบลักษณะของเด็กที่รวมและไม่รวมในการศึกษานี้

ตัวแปรที่วิเคราะห์	เด็กที่รวมในการศึกษาครั้งนี้ (3,125)	เด็กที่ไม่รวมในการศึกษาครั้งนี้ (885)	p-value
น้ำหนักแรกเกิด			0.143
น้อยกว่า 2500 g	242(7.7)	82(9.3)	
2500 g ขึ้นไป	2883(92.3)	803(90.7)	

ตารางที่ 3.1 (ต่อ)

ตัวแปรที่วิเคราะห์	เด็กที่รวมในการศึกษาครั้งนี้ (3,125)	เด็กที่ไม่รวมในการศึกษาครั้งนี้ (885)	p-value
อายุครรภ์			0.009
น้อยกว่า 37 สัปดาห์	242(8.2)	95(11.2)	
37 สัปดาห์ขึ้นไป	2710(91.8)	753(88.8)	
เพศ			0.542
ชาย	1544(49.4)	448(50.6)	
หญิง	1581(50.6)	437(49.4)	
พื้นที่			<0.001
พนมทวน	655(21.0)	116(13.1)	
เทพา	814(26.0)	191(21.6)	
กระนวน	737(23.6)	96(10.8)	
น่าน	655(21.0)	107(12.1)	
กรุงเทพฯ	264(8.4)	375(42.4)	
ศาสนา			0.682
พุทธ	2461(79.3)	687(79.1)	
อิสลาม	541(17.4)	155(17.8)	
คริสต์	19(0.6)	8(0.9)	
อื่นๆ	81(2.6)	19(2.2)	

3.2 ตัวแปรที่ศึกษา

ตัวแปรที่ศึกษาแบ่งตามช่วงเวลาเก็บข้อมูล ประกอบด้วย พื้นฐานครอบครัวและมารดา น้ำหนักของเด็ก ชนิดของนมที่ให้เด็ก อายุที่เริ่มให้อาหารเสริม และการเจ็บป่วย ซึ่งมีการเก็บรวบรวมข้อมูลที่มีความสมบูรณ์ที่สุดดังแสดงไว้ในตารางที่ 3.2 และสามารถจัดเป็นกลุ่มตามลักษณะของตัวแปรได้เป็น 3 กลุ่ม ดังรายละเอียดในตารางที่ 3.3

ตารางที่ 3.2 การเก็บข้อมูล ณ ช่วงเวลาต่างๆ

ตัวแปร	ขณะตั้งครรภ์	แรกเกิด	6 เดือน	12 เดือน	18 เดือน	24 เดือน
พื้นฐานครอบครัวและมารดา	✓					
น้ำหนักของเด็ก		✓	✓	✓	✓	✓
ชนิดของนมที่ให้เด็ก		✓	✓	✓	✓	✓
อายุที่เริ่มให้อาหารเสริม			✓	✓		
การเจ็บป่วย			✓	✓		✓

ตารางที่ 3.3 ตัวแปรที่ศึกษา

ตัวแปรข้อมูลพื้นฐานของมารดา และครอบครัว	
อายุมารดา	จำนวนครั้งของการตั้งครรภ์
น้อยกว่า 20 ปี	ครั้งแรก
20- < 30 ปี	ครั้งที่ 2
30- < 40 ปี	3 - 5 ครั้ง
40 ปีขึ้นไป	มากกว่า 5 ครั้ง
ระดับการศึกษามารดา	สภาพสังคมเศรษฐกิจ
อ่านไม่ออกเขียนไม่ได้	ยากจน
ประถมศึกษาและอื่นๆ	ค่อนข้างยากจน
มัธยมศึกษา ปวช. ปวส.	ปานกลาง
ปริญญาตรีและสูงกว่า	ค่อนข้างดี และดี
อาชีพมารดา	พื้นที่อยู่อาศัย
ผู้มีธุรกิจเป็นของตนเอง	พนมทวน
ผู้ปฏิบัติงานวิชาชีพ	เทพา
คนใช้แรงงาน	กระนวน
พนักงานบริษัท	น่าน
ไม่ได้ปฏิบัติงานเชิงเศรษฐกิจ	กรุงเทพ

ตารางที่ 3.3 (ต่อ)

ตัวแปรการให้นม และอาหารเสริม	
ระยะเวลาการให้นมมารดา	อายุที่เริ่มให้ไข่แดง
น้อยกว่า 4 เดือน	ก่อน 4 เดือน
4-6 เดือน	4-6 เดือน
7-12 เดือน	มากกว่า 6 เดือน
นานกว่า 12 เดือน	อายุที่เริ่มให้ตับ
อายุที่เริ่มให้นมผสม	ก่อน 4 เดือน
ก่อน 4 เดือน	4-6 เดือน
4-6 เดือน	มากกว่า 6 เดือน
7-12 เดือน	อายุที่เริ่มให้เนื้อปลา/ไก่/หมู/วัว
ยังไม่ให้ที่อายุ12 เดือน	ก่อน 4 เดือน
อายุที่เริ่มให้ข้าว/ธัญพืช	4-6 เดือน
ก่อน 4 เดือน	มากกว่า 6 เดือน
4-6 เดือน	อายุที่เริ่มให้ผักผล/ผักใบ
มากกว่า 6 เดือน	ก่อน 4 เดือน
อายุที่เริ่มให้กล้วย/มะละกอ	4-6 เดือน
ก่อน 4 เดือน	มากกว่า 6 เดือน
4-6 เดือน	อายุที่เริ่มให้ไข่ทั้งฟอง
มากกว่า 6 เดือน	ก่อน 4 เดือน
	4-6 เดือน
	มากกว่า 6 เดือน
ตัวแปรข้อมูลทั่วไปของเด็ก	
เพศ	การเจ็บป่วยในช่วง 2 ปีแรก
ชาย	ไม่เคย
หญิง	เคย
อายุครรภ์	น้ำหนักแรกเกิด
น้อยกว่า 37 สัปดาห์	น้อยกว่า 2500 กรัม
37 สัปดาห์ขึ้นไป	2500 กรัมขึ้นไป

3.3 การจัดการข้อมูล

1) ข้อมูลพื้นฐานของครอบครัว มารดา และเด็ก

ในกรณีของตัวแปรเชิงปริมาณ ได้แก่ อายุมารดา จำนวนครรภ์ของมารดา อายุครรภ์ และจำนวนครั้งที่เด็กป่วยต้องรับไว้ในโรงพยาบาล นำมาใช้โดยต้องทำการแปลงข้อมูลให้เป็นเชิงคุณภาพ ในส่วนที่เป็นข้อมูลเชิงคุณภาพเดิม ได้แก่ การศึกษามารดา อาชีพมารดา พื้นที่อยู่อาศัย เพศเด็ก สภาพสังคมเศรษฐกิจ ศาสนา ทำการตรวจสอบความครบถ้วนสมบูรณ์ของข้อมูลแล้วนำมาใช้ได้เลย แต่สำหรับตัวแปรอิสระที่นำไปวิเคราะห์ความสัมพันธ์บางตัวต้องทำการจัดกลุ่มใหม่ เนื่องจากมีจำนวนตัวเลือกมากเกินไป และปัญหาคำตอบของบางตัวเลือกมีความถี่น้อยเกินไปจึงต้องจัดกลุ่มตัวเลือกใหม่โดยลดจำนวนตัวเลือกลงตามความเหมาะสมของตัวเลือกนั้นๆ

2) ข้อมูลการให้นมและอาหารเสริม

การให้นมมารดา มีการเก็บข้อมูลไว้หลายช่วงอายุ ผู้วิจัยคัดเลือกข้อมูลในช่วงเวลาที่มีการเก็บรวบรวมข้อมูลมีความสมบูรณ์ที่สุดใน ช่วงแรกเกิด 6 เดือน 12 เดือน 18 เดือน และ 24 เดือน ตามลำดับ แปลงข้อมูลจาก 5 ช่วงอายุ ให้เหลือเพียงตัวแปรเดียว คือ อายุที่หยุดให้นมมารดา หมายถึง ระยะเวลาการเลี้ยงลูกด้วยนมมารดา มีหน่วยเป็นเดือน ในส่วนของตัวแปรของการให้นมผสม ในโครงการฯ ได้ทำการเก็บข้อมูล 3 ช่วงอายุ คือ 6 เดือน 12 เดือน และ 18 เดือน และการให้อาหารเสริม เก็บข้อมูล 2 ช่วงอายุ คือ 6 เดือน และ 12 เดือน แปลงข้อมูล ให้เหลือเพียงตัวแปรเดียว คือ อายุที่เริ่มให้นมผสม และอายุที่เริ่มให้อาหารเสริม มีหน่วยเป็นเดือนที่เริ่ม

3) ข้อมูลน้ำหนัก

ข้อมูลการเจริญเติบโตด้านน้ำหนัก ใช้ค่ามาตรฐานของน้ำหนักตามเกณฑ์อายุ (z-score of weight for age) เรียกอีกอย่างว่า ค่า z-score ของน้ำหนักตามเกณฑ์อายุ ในโครงการฯ ได้ทำการเก็บข้อมูล 5 ช่วงอายุ คือ แรกเกิด 6 เดือน 12 เดือน 18 เดือน และ 24 เดือน ตามลำดับ โดยค่า z-score ของน้ำหนักตามเกณฑ์อายุ มีสูตรคำนวณดังนี้ (WHO, 1999)

$$\text{z-score น้ำหนักตามเกณฑ์อายุ} = \frac{\text{น้ำหนัก-ค่ามัธยฐานน้ำหนักของประชากรมาตรฐาน}}{\text{ค่าเบี่ยงเบนของประชากรมาตรฐาน}}$$

ในการคำนวณค่า z-score น้ำหนักตามเกณฑ์อายุ และการแปลงเป็นภาวะโภชนาการน้ำหนักตามเกณฑ์อายุนั้นใช้เกณฑ์อ้างอิง น้ำหนัก ส่วนสูง และเครื่องชี้วัดโภชนาการอื่นของประชาชนไทยอายุ 1 วัน – 19 ปี พ.ศ. 2542 (กองโภชนาการ, 2542) ซึ่งภาวะโภชนาการน้ำหนักตามเกณฑ์อายุของเด็กนั้น ได้แบ่งออกเป็น 2 กลุ่ม คือ กลุ่มน้ำหนักน้อยกว่าเกณฑ์ (z-score of weight for age <- 2 SD) และกลุ่มน้ำหนักไม่น้อยกว่าเกณฑ์ (z-score of weight for age $\geq$ -2 SD)

ดังนั้นการเจริญเติบโตด้านน้ำหนัก วัดด้วยค่า z-score น้ำหนักตามเกณฑ์อายุ จัดเป็นตัวแปรตามเชิงปริมาณ ส่วนภาวะน้ำหนักน้อยกว่าเกณฑ์ วัดด้วยภาวะโภชนาการน้ำหนักตามเกณฑ์อายุ จัดเป็นตัวแปรตามเชิงคุณภาพแบบทวินาม

3.3 การวิเคราะห์ข้อมูล

การวิเคราะห์ข้อมูลแบ่งเป็น 2 ลักษณะ คือ 1) การวิเคราะห์ข้อมูลด้วยวิธี Generalized Estimating Equations สำหรับข้อมูลการวิจัยระยะยาว โดยมีตัวแปรตาม 2 ตัวแปร คือ ภาวะน้ำหนักน้อยกว่าเกณฑ์ และการเจริญเติบโตด้านน้ำหนัก เพื่อศึกษาผลของการเลี้ยงลูกด้วยนมมารดา และการให้อาหารเสริมที่มีต่อการเจริญเติบโตด้านน้ำหนัก และภาวะน้ำหนักน้อยกว่าเกณฑ์ 2) การวิเคราะห์ข้อมูลภาคตัดขวาง โดยมีการเลี้ยงทารกด้วยนมมารดาเป็นตัวแปรตาม ซึ่งในกรณีนี้จัดการตัวแปรระยะเวลาการให้นมมารดาเป็นตัวแปรเชิงคุณภาพแบบทวินาม เพื่อศึกษาผลของภาวะน้ำหนักน้อยกว่าเกณฑ์ ที่มีต่อ ระยะเวลาการเลี้ยงลูกด้วยนมมารดา โดยมีขั้นตอนการวิเคราะห์ ดังนี้ 1) Descriptive Statistics: วิเคราะห์ข้อมูลเบื้องต้นโดยใช้ ความถี่ และร้อยละ เมื่อนำเสนอข้อมูลเชิงคุณภาพ ใช้ค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐาน เมื่อนำเสนอข้อมูลเชิงปริมาณ จัดกลุ่มใหม่ในกรณีที่ความถี่ของกลุ่มน้อยเกินไป กรณีที่ตัวแปรอิสระมีค่าขาดหายไป แทนด้วยค่ามัธยฐานของตัวแปรนั้น 2) Univariate Analysis: ศึกษาความสัมพันธ์ระหว่างภาวะน้ำหนักน้อยกว่าเกณฑ์ กับตัวแปรอิสระเชิงคุณภาพทีละตัวแปรใช้ Pearson’s Chi-Squared test และศึกษาความแตกต่างค่าเฉลี่ยการเจริญเติบโตด้านน้ำหนัก กับตัวแปรอิสระเชิงคุณภาพทีละตัวแปรใช้ t-test และ ANOVA 3) Multivariate Analysis: สร้างตัวแบบความสัมพันธ์กรณีตัวแปรตามเป็นตัวแปรเชิงปริมาณโดย Multiple Regression และกรณีตัวแปรตามเป็นตัวแปรเชิงคุณภาพโดย Logistic

Regression 4) Analysis of Longitudinal Data: สร้างตัวแบบความสัมพันธ์โดยใช้ GEE เมื่อคำนึงถึงความสัมพันธ์ภายในตัวแปรตามและใช้เกณฑ์ QIC ในการเลือกตัวแบบที่เหมาะสม วิเคราะห์ข้อมูลโดยใช้โปรแกรม STATA Version 7

3.4.1 Odds Ratio

Odds Ratio (OR) เป็นสถิติที่ใช้วัดระดับความสัมพันธ์ ระหว่างตัวแปร y กับ x กรณีที่ y และ x เป็นตัวแปรเชิงคุณภาพที่มี 2 กลุ่ม ตัวอย่างเช่น ให้ y เป็นภาวะโภชนาการด้านน้ำหนัก (y = 1 คือ น้ำหนักน้อยกว่าเกณฑ์ และ y = 0 คือ น้ำหนักไม่น้อยกว่าเกณฑ์) และ x เป็นเพศ (x = 1 คือ เพศชาย และ x = 2 คือ เพศหญิง) สามารถคำนวณค่า Odds Ratio ได้ดังนี้

$$\mathrm{OR} = (ad)/(bc) \tag{3.1}$$

โดยที่ a, b, c, d เป็นความถี่ แสดงในตารางที่ 3.4

ตารางที่ 3.4 ตารางการจรณ์ 2x2

	$y=1$	$y=0$
$x=1$	a	b
$x=0$	c	d

ถ้า OR = 1 หมายถึง ไม่มีความสัมพันธ์ระหว่าง ภาวะโภชนาการด้านน้ำหนักกับเพศ ส่วนกรณีที่ OR = 1.5 หมายถึง เพศชายมีความเสี่ยงที่จะมีภาวะน้ำหนักน้อยกว่าเกณฑ์เป็น 1.5 เท่าของเพศหญิง

เนื่องจาก Odds Ratio มีการแจกแจงแบบเบ้ ดังนั้นค่าความคลาดเคลื่อนมาตรฐาน (Standard Error) ของ Odds Ratio สามารถคำนวณในรูปของลอการิทึม โดยมีสมการเป็นดังนี้

$$\mathrm{se}(\ln \mathrm{OR}) = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \tag{3.2}$$

ซึ่งสามารถนำมาคำนวณช่วงความเชื่อมั่น95% ของ Odds Ratio (95%Confidence Interval of Odds Ratio) สามารถคำนวณได้ดังนี้

$$\left[\mathrm{OR}\exp\left(-1.96\,\mathrm{se}\right),\mathrm{OR}\exp\left(1.96\,\mathrm{se}\right)\right]$$

3.4.2 Multiple Regression Model

สร้างตัวแบบความสัมพันธ์ระหว่างตัวแปรอิสระกับการเจริญเติบโตด้านน้ำหนัก สำหรับตัวแปรตามในที่นี้ คือ z-score ของน้ำหนักตามเกณฑ์อายุใช้ Multiple Regression โดยมีตัวแบบเป็น

$$\boldsymbol{\mu}_i = E(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \mathbf{x}_i^T \boldsymbol{\beta} \tag{3.3}$$

โดยที่ Y_i คือตัวแปรตาม ของค่าสังเกตที่ i โดยที่ $i=1,\ldots N$, $\mathbf{x}_i$ คือ เวกเตอร์ของตัวแปรอิสระ p ตัว $\mathbf{x}_i = [1, x_{i1}, \ldots, x_{ip}]^T$ และ $\boldsymbol{\beta}$ คือ เวกเตอร์ของค่าสัมประสิทธิ์การถดถอยที่ต้องการประมาณ $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_p]^T$

3.4.3 Logistic Regression Model

สร้างตัวแบบความสัมพันธ์ระหว่างตัวแปรอิสระกับภาวะน้ำหนักน้อยกว่าเกณฑ์ ใช้ Logistic Regression โดยมีตัวแบบเป็น

$$\ln\left(\pi/1-\pi\right) = \beta_0 + \beta_1 x_1 + \ldots \beta_i x_i \ldots + \beta_p x_p \tag{3.4}$$

ค่า $\beta_0, \beta_1 \ldots \beta_i \ldots$ และ β_p ก็คือ ค่าสัมประสิทธิ์ของสมการ เช่นเดียวกับสมการ Multiple Regression เรียกว่า Logistic Regression Coefficient การวิเคราะห์ Logistic Regression จากข้อมูลที่ได้จากการวิจัย จะได้ค่าประมาณของ Logistic Regression Coefficient ของประชากร ซึ่ง $\mathrm{logit}(E(y_i)) = \ln(\pi/1-\pi)$ เมื่อ π คือ สัดส่วนของการมีภาวะน้ำหนักน้อยกว่าเกณฑ์

3.4.4 Generalized Estimating Equations (GEE)

GEE พัฒนามาจาก GLM สำหรับวิเคราะห์ข้อมูลที่มีความสัมพันธ์ ลักษณะสำคัญของ GEE คือเมตริกซ์ของความแปรปรวนร่วมที่ไม่ใช่เมตริกซ์ทแยงมุมในขณะที่ใน GLM เมตริกซ์ของความแปรปรวนร่วมเป็นเมตริกซ์ทแยงมุม วิธีการ GEE สามารถประมาณค่าสัมประสิทธิ์การถดถอย (Regression Coefficients) และค่าความคลาดเคลื่อนมาตรฐาน (Standard Error) ได้ในลักษณะเดียวกับวิธีการถดถอยอื่นๆ

ตัวแบบสำหรับการวิเคราะห์ข้อมูลระยะยาวมีสามตัวแบบหลัก สำหรับตัวแบบของ GEE คือ Marginal Model หรือ Population Average ซึ่งเป็นตัวแบบที่อธิบายค่าเฉลี่ยในระดับประชากร โดยเมื่อมีการเปลี่ยนแปลงไปตามกาลเวลานั้นจะมีผลอย่างไรต่อตัวแปรตามโดยเฉลี่ย ทั้งนี้ต้องกำหนดรูปแบบสหสัมพันธ์ของตัวแปรตามภายในหน่วยตัวอย่างเดียวกัน (Intra-Subject Correlation) ระหว่างตัวแปรตามในเวลาต่างๆ ซึ่งสหสัมพันธ์ระหว่างตัวแปรตาม แต่ละครั้ง (Two-Point Intra-Subject Correlation) ในหน่วยตัวอย่างเดียวกันมักจะมีผลการวัดครั้งติดๆ กันได้ค่าใกล้เคียงกัน ดังนั้นถ้าเอาผลการวัดครั้งแรกกับครั้งที่สองของทุกหน่วยตัวอย่างมาหาค่าสหสัมพันธ์ก็จะได้ค่าสหสัมพันธ์ระดับหนึ่งซึ่งมักจะมีค่าไม่เท่ากับ 0 เขียนเป็นสัญลักษณ์ว่า p_{12} แต่การวัดมีหลายครั้งมาก นอกจากครั้งที่ 1 กับ 2 ยังมีครั้งที่ 1 กับ 3, 1 กับ 4, 1 กับ ... และครั้งที่ 2 กับ 3, 2 กับ 4, 2 กับ 5 ...ฯลฯ ถ้าปรับผลของปัจจัย (x) ต่างๆ ต่อผลลัพธ์ที่เกิดขึ้นแต่ละค่าก็จะกลายเป็นความคลาดเคลื่อน (Residual) สหสัมพันธ์จึงเปลี่ยนจากสหสัมพันธ์ของตัวแปรตามกลายเป็นสหสัมพันธ์ของ Residuals ระหว่างครั้งต่างๆ ในงานวิจัยนี้คำว่า Working Correlation จะหมายถึง Intra-Subject Correlation of Residuals

วิธีการ GEE สามารถประมาณค่าสัมประสิทธิ์การถดถอย และค่าคลาดเคลื่อนมาตรฐาน (Liang and Zeger, 1986) ได้ในลักษณะเดียวกับวิธีการ Weighted Least Squares ซึ่งวิธีการ GEE มีรายละเอียดเกี่ยวกับตัวแบบดังนี้

กรณีข้อมูลการวิจัยระยะยาว $(y_{it},\mathbf{x}_{it})$ เก็บข้อมูล ณ ช่วงเวลา t, t = 1, 2, 3, …n_i จากหน่วยตัวอย่างที่ i , i = 1, 2, 3, …k ซึ่งทำให้มีค่าสังเกตทั้งหมดเป็น $N=\sum_{i=1}^{k} n_i$ โดยที่ y_{it} เป็นค่าสังเกตของตัวแปรตาม และ $\mathbf{x}_{it}$ เป็นเวกเตอร์ของค่าสังเกตของตัวแปรอิสระขนาด $p\times 1$

ดังนั้น $\mathbf{y}_i=\left(y_{i1},...,y_{in_i}\right)^T$ เป็นเวกเตอร์ขนาด $n_i\times 1$ และ $X_i=\left(\mathbf{x}_{i1},...,\mathbf{x}_{in_i}\right)^T$ เป็นเมตริกซ์ขนาด $n_i\times p$ สำหรับหน่วยตัวอย่างที่ i

ค่าคาดหวังของ y_{it} คือ

$$E(y_{it}) = \boldsymbol{\mu}_{it} \tag{3.5}$$

และ

$$g(\mu_{it}) = \mathbf{x}_{it}\boldsymbol{\beta}, \quad g(\mu_i) = X_i\boldsymbol{\beta} \tag{3.6}$$

โดยที่ $g(\mu_{it})$ เป็น Link Function ใน Generalized Linear Model ซึ่งพิจารณาจากความสัมพันธ์ระหว่างค่าเฉลี่ยและฟังก์ชั่นเชิงเส้นของตัวแปรอิสระ และ $\boldsymbol{\beta}$ คือเวกเตอร์ขนาด $p\times1$ ของสัมประสิทธิ์การถดถอย

ความแปรปรวนของ y_{it} แสดงให้อยู่ในรูปฟังก์ชั่นของ μ_{it} ได้เป็น

$$\operatorname{var}(y_{it}) = v(\mu_{it})/\varphi \tag{3.7}$$

โดย $v(\mu_{it})$ คือ ฟังก์ชั่นความแปรปรวน (Variance Function) และ φ เป็น Scale Parameter โดยทั้ง 2 ตัวสามารถประมาณค่าได้

เวกเตอร์ของสัมประสิทธิ์การถดถอย $\boldsymbol{\beta}$ สามารถประมาณได้จากสมการ

$$\sum_{i=1}^{k} D_i^T V_i^{-1} S_i = \mathbf{0} \tag{3.8}$$

เมื่อ $S_i = \mathbf{y}_i - \boldsymbol{\mu}_i$ เป็นเมตริกซ์ขนาด $n_i \times p$, $\boldsymbol{\mu}_i = \left(\mu_{i1},\ldots,\mu_{in_i}\right)^T$,

โดย $D_i = \dfrac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$ เป็นเมตริกซ์ขนาด $n_i \times p$

$$D_i = \begin{bmatrix} \dfrac{\partial \mu_{i1}}{\partial \beta_1} & \cdots & \dfrac{\partial \mu_{i1}}{\partial \beta_p} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial \mu_{in_i}}{\partial \beta_1} & \cdots & \dfrac{\partial \mu_{in_i}}{\partial \beta_p} \end{bmatrix}$$

และ $V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2} / \varphi$ โดยที่ $A_i = \mathrm{diag}\{v(\mu_{it})\}$ เป็นเมตริกซ์ทแยงมุม ขนาด $n_i \times n_i$ ที่ประกอบด้วยฟังก์ชั่นความแปรปรวน $v(\mu_{it})$ ในลำดับทแยงมุมที่ t

$$A_i = \begin{bmatrix} v(\mu_{i1}) & 0 & \cdots & \cdots & 0 \\ 0 & v(\mu_{i2}) & 0 & \cdots & 0 \\ \cdots & 0 & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & v(\mu_{in_i}) \end{bmatrix}$$

ให้ $R_i(\alpha)$ เป็นเมตริกซ์ของความสัมพันธ์ของ y_i หรือเรียกอีกอย่างว่า Working Correlation Matrix ขนาด $n_i \times n_i$ กรณี $R_i(\alpha)$ มีโครงสร้างของสหสัมพันธ์แบบ Stationary

$$R_i(\alpha) = \begin{bmatrix} 1 & \alpha_1 & 0 & \cdots & 0 \\ \alpha_1 & 1 & \alpha_1 & \cdots & 0 \\ 0 & \alpha_1 & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \alpha_1 \\ 0 & 0 & 0 & \alpha_1 & 1 \end{bmatrix}$$

เพราะฉะนั้น $$V_i = \frac{1}{\varphi} \begin{bmatrix} v(\mu_{i1}) & \alpha_1\sqrt{v(\mu_{i1})} & 0 & \cdots & 0 \\ \alpha_1\sqrt{v(\mu_{i1})} & v(\mu_{i2}) & \alpha_1\sqrt{v(\mu_{i2})} & \cdots & 0 \\ \cdots & \alpha_1\sqrt{v(\mu_{i2})} & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \alpha_1\sqrt{v(\mu_{in_i})} \\ 0 & 0 & 0 & \alpha_1\sqrt{v(\mu_{in_i})} & v(\mu_{in_i}) \end{bmatrix}$$

ดังนั้น $V_i = \mathrm{cov}(y_i)$ ไม่ได้เป็นฟังก์ชั่นของ $\boldsymbol{\beta}$ เพียงอย่างเดียวแต่เป็นฟังก์ชั่นของ α ด้วย ในการแก้สมการ (3.8) φ, α และ $\boldsymbol{\beta}$ สามารถประมาณได้โดยการทำย้อนซ้ำ ดังต่อไปนี้

1. คำนวณค่าประมาณพารามิเตอร์ $\boldsymbol{\beta}$ ด้วยวิธี GLM

2. นำค่า $\boldsymbol{\beta}$ ที่ประมาณได้ในข้อที่ 1 มาคำนวณหาค่าประมาณ φ, α โดยใช้ Standardized Residuals

$$r_{it} = (\mathbf{y}_{it} - \hat{\boldsymbol{\mu}}_{it}) / v(\hat{\boldsymbol{\mu}}_{it})^{\frac{1}{2}} \tag{3.9}$$

โดย φ สามารถคำนวณได้จากสูตร

$$\varphi = \frac{1}{(Np)\sum_{i=1}^{k}\sum_{t=1}^{n_i} r_{it}^2} \tag{3.10}$$

และพารามิเตอร์ของสหสัมพันธ์ α สามารถประมาณค่าได้โดยขึ้นอยู่กับการกำหนดรูปแบบสหสัมพันธ์ (Working Correlation) แสดงในตารางที่ 3.5

3. คำนวณค่าประมาณ $V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2} / \varphi$

4. คำนวณค่าประมาณพารามิเตอร์ β ใหม่อีกครั้งจากสมการ

$$\boldsymbol{\beta}_{r+1} = \boldsymbol{\beta}_r + \left[\sum_{i=1}^{k} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right]^{-1} \left[\sum_{i=1}^{k} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} V_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)\right] \tag{3.11}$$

โดยผลที่ได้จากสมการ (3.11) นำไปใช้เป็นค่าเริ่มต้นของขั้นตอนที่ 2-4 อีกครั้งในการวนซ้ำโดยการถ่วงน้ำหนักค่ากำลังสองน้อยที่สุดใหม่ จนกว่าจะได้ค่าที่เปลี่ยนแปลงไปต่ำที่สุดในการประมาณค่าพารามิเตอร์จากการ เลือกตัวแบบถดถอยที่ดีที่สุด (Gardner et al., 1995; Hardin and Hilbe, 2003; Liang and Zeger, 1986; McCullagh and Nelder, 1989) เมื่อขนาดที่เปลี่ยนไปเมื่อเทียบกับรอบก่อนหน้ามีค่าเข้าใกล้ 0 นั่นคือค่าประมาณพารามิเตอร์ (β_r และ SE) มีค่าคงที่ ซึ่งการกำหนดรูปแบบของความสัมพันธ์ที่ถูกต้องจะทำให้เพิ่มประสิทธิภาพของการประมาณค่าอีกด้วย (Fitzmaurice, 1995; Hardin and Hilbe, 2003) โดยเฉพาะเมื่อความสัมพันธ์ภายในกลุ่มมีค่ามาก (Diggle et al., 2002; Zorn, 2001) อย่างไรก็ตาม ค่าคลาดเคลื่อนของตัวแบบที่ได้จากการกำหนดโครงสร้างความสัมพันธ์ยังคงใช้ได้ เพราะว่า ค่าประมาณที่ได้เปลี่ยนแปลงไปไม่มาก เพราะฉะนั้นประสิทธิภาพที่ได้เพิ่มขึ้นมาจากการกำหนดโครงสร้างของความสัมพันธ์ที่ถูกต้องนั้นโดยทั่วไปแล้วเพิ่มขึ้นเพียงเล็กน้อยเท่านั้น (Liang and Zeger, 1986)

ตารางที่ 3.5 รูปแบบโครงสร้างสหสัมพันธ์ (Working Correlation)

Working Correlation		Estimator
Independent	$Corr(y_{ij},y_{ik})=\begin{cases}1 & j=k\\ 0 & j\neq k\end{cases}$	สหสัมพันธ์นี้ไม่ต้องประมาณค่า
Exchangeable	$Corr(y_{ij},y_{ik})=\begin{cases}1 & j=k\\ \alpha & j\neq k\end{cases}$	$\hat{\alpha}=\frac{1}{(N^{*}-p)\varphi}\sum_{i=1}^{k}\sum_{j\neq k} r_{ij}r_{ik}$ โดยที่ $N^{*}=\sum_{i=1}^{k} n_i(n_i-1)$
Autoregressive AR(1)	$Corr(y_{ij},y_{i,j+t})=\alpha^{t}$ $t = 0,1,2, ... ,ni-j$	$\hat{\alpha}=\frac{1}{(N^{*}-p)\varphi}\sum_{i=1}^{k}\sum_{j\leq n_i-1} r_{ij}r_{i,j+1}$ โดยที่ $N^{*}=\sum_{i=1}^{k}(n_i-1)$
Stationary	$Corr(y_{ij},y_{i,j+t})=\begin{cases}1 & t=0\\ \alpha_t & t=1\\ 0 & t>1\end{cases}$	$\hat{\alpha}=\frac{1}{(N^{*}-p)\varphi}\sum_{i=1}^{k}\sum_{j\leq n_i-1} r_{ij}r_{i,j+t}$ โดยที่ $N^{*}=\sum_{i=1}^{k}(n_i\text{-}t)$
Unstructured	$Corr(y_{ij},y_{ik})=\begin{cases}1 & j=k\\ \alpha_{jk} & j\neq k\end{cases}$	$\hat{\alpha}_{jk}=\frac{1}{(N-p)\varphi}\sum_{i=1}^{k} r_{ij}r_{ik}$

3.4.5 การเลือกตัวแบบที่เหมาะสม (Model Selection)

การเลือกตัวแบบที่เหมาะสมสำหรับ GEE ในการศึกษาครั้งนี้ใช้หลักเกณฑ์ Quasi Information Criterion (QIC) ซึ่งได้ประยุกต์มาจาก Akaike Information Criterion (AIC) ใน GLM (Pan, 2001)

$$AIC = -2L+2(p) \tag{3.12}$$

โดย L คือ Log-likelihood Function

p คือ จำนวนพารามิเตอร์ที่อยู่ในตัวแบบ

ซึ่งการเลือกตัวแบบโดยวิธี AIC ใช้สำหรับ GLM ซึ่งสมการประมาณค่าเป็น Likelihood Functions แต่สำหรับ GEE มีสมการประมาณค่าเป็น Quasi-likelihood Functions

$$\text{Quasi-likelihood Function} \qquad Q(\mu,\varphi;y)=\sum_{i}^{k}\frac{(\mathbf{y}_i-\boldsymbol{\mu}_i)}{\varphi V(\boldsymbol{\mu}_i)} \tag{3.13}$$

จากสมการ (3.13) สามารถแปลงไปเป็น สมการ (3.8) ได้โดย $\frac{\partial Q(\mu,\varphi;y)}{\partial \beta}$ เพราะฉะนั้นเกณฑ์ AIC สามารถแปลงมาใช้วิธี QIC ได้เช่นเดียวกัน สำหรับการเลือกสหสัมพันธ์ของตัวแปรตาม (Working Correlation) ใช้เกณฑ์ QIC สามารถคำนวณได้จากสมการ (3.14)โดยพิจารณาเลือกสหสัมพันธ์จากค่า QIC ต่ำที่สุด

$$QIC(R) = -2Q(\mu,\varphi;y) + 2\operatorname{trace}(A_I^{-1}V_r) \tag{3.14}$$

โดยที่ $Q(\mu,\varphi;y)$ คือ Quasi-likelihood Function

A_I^{-1} คือ เป็นเมตริกซ์เอกลักษณ์

$\hat{V}_r$ คือ $Cov(\hat{\beta})$

ในส่วนการเลือกตัวแบบที่เหมาะสมใช้เกณฑ์ QIC_u สามารถคำนวณได้โดยสมการ (3.15)โดยพิจารณาเลือกตัวแบบจากค่า QIC_u ต่ำที่สุด (Hardin and Hilbe, 2003)

$$QIC_u = -2Q(\mu,\varphi;y)+2(p) \tag{3.15}$$

โดยที่ $Q(\mu,\varphi;y)$ คือ Quasi-likelihood Function

p คือ จำนวนพารามิเตอร์ที่อยู่ในตัวแบบ