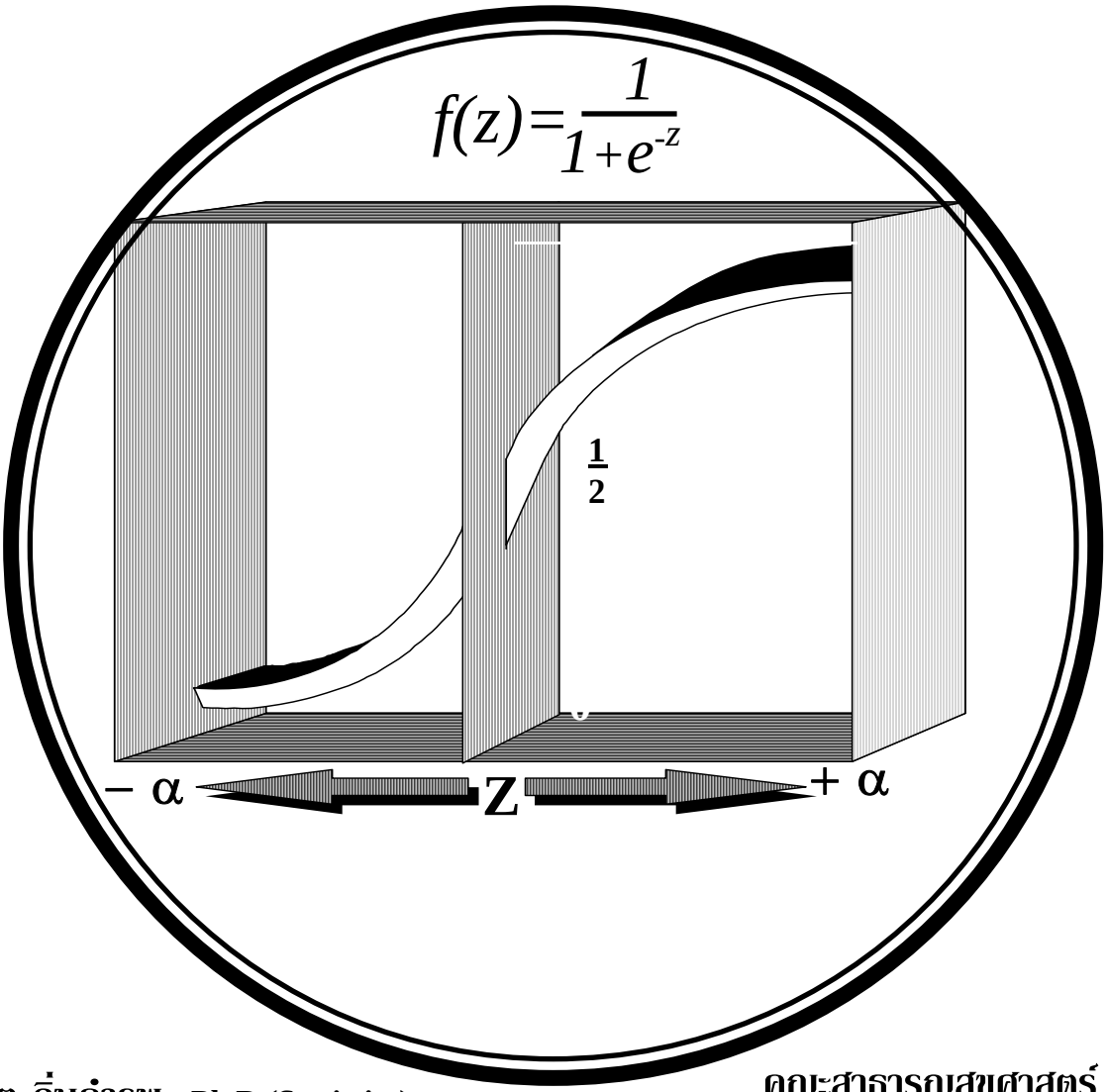
การวิเคราะห์ข้อมูลการวิจัยทางวิทยาศาสตร์สุขภาพโดยใช้

การถดถอยลอจีสติก

ANALYSIS OF DATA IN HEALTH SCIENCE RESEARCH USING

LOGISTIC REGRESSION



บัณฑิต ถิ่นคำรพ, Ph.D.(Statistics)
ภาควิชาชีวสถิติและประชากรศาสตร์

คณะสาธารณสุขศาสตร์
มหาวิทยาลัยขอนแก่น

การวิเคราะห์ข้อมูลการวิจัยทางวิทยาศาสตร์สุขภาพโดยใช้

การถดถอยลอจีสติก

ANALYSIS OF DATA IN HEALTH SCIENCE RESEARCH USING LOGISTIC REGRESSION

เรียบเรียงโดย

บัณฑิต ถิ่นดำรพ

วท.บ. (สาธารณสุขศาสตร์) เกียรตินิยม
M.P.H. (Epidemiology)
Grad. Dip. Medical Statistics
Ph.D. (Statistics)

จัดพิมพ์ที่คณะสาธารณสุขศาสตร์ มหาวิทยาลัยขอนแก่น

คำนำ

ปัญหาด้านการแพทย์และสาธารณสุข อันได้แก่ โรคและภัยต่าง ๆ ล้วนมีสาเหตุจากปัจจัยหลายอย่างสัมพันธ์ เกื้อหนุนซึ่งกันและกัน การศึกษาวิจัยเพื่อหาสาเหตุ โดยวิเคราะห์จากปัจจัยเดียว จึงไม่เหมาะสมกับลักษณะของปัญหา แต่ต้องพิจารณาผลกระทบจากหลายปัจจัยพร้อม ๆ กัน ดังนั้น การอธิบายความสัมพันธ์ระหว่างปัจจัยใด ๆ กับการเกิดโรคหรือปัญหาสาธารณสุขต่าง ๆ จะต้องมีการควบคุมผลกระทบจากปัจจัยอื่น ๆ ด้วย วิธีการทางสถิติที่ใช้วิเคราะห์ข้อมูลดังกล่าว เรียกว่า การวิเคราะห์ตัวแปรพหุ (Multivariable analysis) ในการนี้ วิธีที่สำคัญ เป็นที่นิยม และใช้แพร่หลายในงานวิจัย คือ การวิเคราะห์การถดถอยลอจีสติก (Logistic Regression)

อนึ่ง ปัจจุบันการวิจัยด้านการแพทย์ และสาธารณสุขเพื่ออธิบายความสัมพันธ์ระหว่างปัจจัยต่าง ๆ ที่มีการพิจารณาผลกระทบจากหลายปัจจัย พร้อม ๆกันนั้น ได้เริ่มรู้จักกันแพร่หลาย ยังผลให้ความต้องการ การเรียนรู้วิธีการทางสถิติ เพื่อการวิเคราะห์ข้อมูลที่ถูกต้องเหมาะสมนั้น มีมากขึ้นเป็นลำดับ แต่ตำราภาษาไทย ที่กล่าวถึงเรื่องดังกล่าว มีน้อยมาก กอปรกับความซับซ้อนของเนื้อหาวิชา เกี่ยวกับวิธีการทางสถิติดังกล่าว และพื้นฐานความรู้ด้านสถิติที่จำกัดของผู้เรียน ยังผลให้ยากต่อการทำความเข้าใจ ผู้เขียนจึงเรียบเรียงตำรานี้ขึ้น โดยกำหนดรูปแบบให้เป็นตำราที่ผู้อ่านสามารถศึกษาได้ด้วยตนเอง เพื่ออำนวยโอกาสการเรียนรู้วิธีการวิเคราะห์ข้อมูล การแปลความหมาย และการใช้ประโยชน์ อันจะยังผลให้เกิดการวิเคราะห์ข้อมูลการวิจัยด้านการแพทย์และสาธารณสุขที่ถูกต้อง เหมาะสมกับลักษณะของปัญหา สร้างองค์ความรู้ที่ถูกต้องเพื่อประโยชน์ต่อการแก้ปัญหาสุขภาพอนามัยของปวงชน ต่อไป

ผู้เขียนยินดีน้อมรับคำแนะนำ ตลอดข้อคิดเห็นต่าง ๆ เกี่ยวกับหนังสือเล่มนี้ ด้วยความเต็มใจยิ่ง

บัณฑิต ถิ่นคำรพ

เมษายน 2543

กิตติกรรมประกาศ

ตำราเล่มนี้ เกิดจากแรงบันดาลใจ ในอันที่จะอำนวยโอกาสให้ผู้อ่าน สามารถเรียนรู้ด้วยตนเอง จากแนวการเขียนตำรา โดย David G. Kleinbaum ซึ่งเป็นผู้ที่มีผลงานเขียนหลายเล่ม ที่เกี่ยวข้องกับการวิเคราะห์ข้อมูล โดยใช้การถดถอยลอจีสติก จึงใคร่ขอระลึกเกียรติคุณท่านผู้นี้ ไว้ ณ ที่นี้

ผู้เขียนใคร่ขอขอบคุณ อาจารย์ภาควิชาชีวสถิติและประชากรศาสตร์ มหาวิทยาลัยขอนแก่น โดยเฉพาะอย่างยิ่ง รองศาสตราจารย์อรุณ จิรวัฒน์กุล ซึ่งได้กรุณาให้คำแนะนำอันทรงค่ายิ่ง พร้อมกันนี้ ขอขอบคุณ รองศาสตราจารย์มาลินี เหล่าไพบูลย์ ผู้ช่วยศาสตราจารย์จารุวรรณ โชคคณาพิทักษ์ ผู้ช่วยศาสตราจารย์ยุพา ถาวรพิทักษ์ ผู้ช่วยศาสตราจารย์ดร. จิราพร เขียวอยู่ และ ผู้ช่วยศาสตราจารย์นิคม ถนอมเสียง ที่ได้มีส่วนอย่างมาก ในการให้ข้อคิดเห็น ปรับปรุงแก้ไข ตำราเล่มนี้

นอกจากนี้ ผู้เขียนขอขอบคุณ นักศึกษามหาบัณฑิตสาธารณสุขศาสตร์ สาขาชีวสถิติ รุ่นที่ 2 ถึง 5 ของคณะสาธารณสุขศาสตร์ มหาวิทยาลัยขอนแก่น ที่เป็นผู้อ่าน ทดลองใช้ และปรับปรุงแก้ไขฉบับร่าง

คำแนะนำการศึกษาด้วยตนเอง

ตำราเล่มนี้ มีทั้งหมด 5 บท แต่ละบทประกอบด้วย 4 ส่วน คือ ส่วนแรกเป็นคำอธิบายต้นบทว่าด้วยวัตถุประสงค์ เนื้อหา และกิจกรรม ส่วนที่สองคือเนื้อหา ส่วนที่สามคือแบบฝึกหัด และส่วนที่สี่คือเอกสารอ้างอิงประจำบท ส่วนที่สองซึ่งเป็นเนื้อหาวิชาการนั้น นำเสนอเป็นสองคอลัมน์ คอลัมน์ด้านซ้ายเป็นภาพสไลด์ที่ใช้สรุปประเด็นสำคัญ ๆ และคอลัมน์ด้านขวาเป็นรายละเอียดที่อธิบายภาพสไลด์ด้านซ้าย

ผู้อ่านที่มีพื้นความรู้ด้านสถิติ สามารถทบทวนความรู้ได้ โดยเพียงกวาดสายตาผ่านคอลัมน์ด้านซ้ายและเลือกอ่านรายละเอียดด้านขวาเฉพาะประเด็นที่ต้องการ ส่วนท่านที่มีพื้นฐานความรู้ด้านสถิติจำกัด ใคร่ขอแนะนำให้อ่านรายละเอียดด้านขวา พร้อมกับดูภาพสไลด์ด้านซ้าย เพื่อทราบประเด็นสำคัญของเนื้อหาที่อ่าน

ส่วนที่สามของแต่ละบทเป็นแบบฝึกหัด การนำเสนอแบบฝึกหัดท้ายบทนี้ จะมีขั้นตอนการปฏิบัติที่ผู้เรียนพึงฝึกหัดให้เกิดทักษะ พร้อมคำอธิบายในแต่ละกิจกรรมดังกล่าว เพื่อให้ผู้เรียนเกิดแนวคิด และเข้าใจสิ่งที่กล่าวในบทเรียนก่อนหน้านี้ยิ่งขึ้น นอกจากนั้น แบบฝึกหัดยังมีส่วนที่เป็นคำถามเพื่อประเมินความเข้าใจผู้เรียนด้วย ผู้เรียนจะได้รับประโยชน์จากการลงมือทำ แบบฝึกหัดท้ายบทอย่างมาก หากพยายามทำด้วยตนเองก่อนเปิดดูคำเฉลยที่ภาคผนวก 1 ซึ่งจัดทำไว้โดยละเอียด

คำแนะนำในภาคผนวกที่ 2 ที่ให้ไว้ท้ายเล่ม สำหรับผู้อ่านที่ต้องการข้อมูลที่ใช้ในการวิเคราะห์ และโปรแกรมคอมพิวเตอร์

โปรแกรมคอมพิวเตอร์ที่ใช้ในตำราเล่มนี้คือ STATA® ทั้งนี้เพราะสามารถวิเคราะห์ได้ครอบคลุมค่าสถิติที่กล่าวในที่นี้ได้ทั้งหมด

อย่างไรก็ตาม ทุกแบบฝึกหัด มีคำเฉลยโดยละเอียด ที่เสนอไว้ในภาคผนวก 2 คำเฉลยดังกล่าว ได้แสดงผลที่ได้จากการวิเคราะห์โดยใช้คอมพิวเตอร์ เหมือนกับที่ได้จากจอภาพคอมพิวเตอร์ ดังนั้น ท่านที่ต้องการเพียงอ่านทำความเข้าใจ ก็สามารถทำได้โดยไม่จำเป็นต้องใช้คอมพิวเตอร์

เชื่อมั่นอย่างยิ่งว่า ตำราเล่มนี้ จะเป็นส่วนสำคัญในความสำเร็จของท่าน ในการศึกษาแนวทางการวิเคราะห์ข้อมูลโดยใช้การถดถอยลอจีสติก นอกจากนั้น ตำราต่าง ๆ ที่ระบุในเอกสารอ้างอิง เป็นอีกแหล่งความรู้หนึ่งที่จะช่วยให้ท่านรอบรู้เรื่องดังกล่าวนี้มากยิ่งขึ้น

สารบัญ

บทนำ (Introduction)

วัตถุประสงค์ : เพื่อให้ผู้เรียนสามารถ

1. บอกลักษณะปัญหาการวิจัยที่เหมาะสมกับการใช้ Logistic Regression ในการวิเคราะห์ข้อมูล

2. วิเคราะห์ข้อมูลอย่างหยาบ (Crude analysis) ได้

3. วิเคราะห์ข้อมูลแบบ Stratified Analysis ได้

4. อธิบาย Confounding effect และ Interaction effect ในการวิเคราะห์ข้อมูลเพื่อหาปัจจัยเสี่ยงได้

เนื้อหา :

1. ตัวแปรต้นและตัวแปรตาม สำหรับการวิเคราะห์ข้อมูลโดยใช้ Logistic Regression

2. Bivariate analysis กับ Logistic Regression

3. Stratified analysis กับ Logistic Regression : บทบาทต่อ Confounding และ Interaction effect

กิจกรรม :

1. ฟังบรรยายประกอบแผ่นใส พร้อมบันทึกเนื้อหาสำคัญลงในชุดการเรียนการสอน บทที่1

2. ทำแบบฝึกหัด

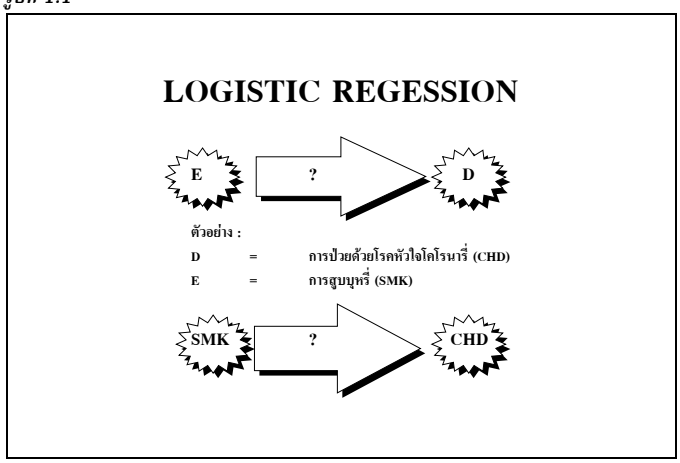
3. อภิปรายและสรุปเนื้อหา พร้อมเขียนสรุปท้ายบทลงกรอบว่างที่ให้ไว้ท้ายบท

สิ่งที่นำเสนอ	คำอธิบาย

1. คำถามวิจัยที่ต้องใช้ Logistic Regression

1.1 ตัวแปรต้นและตัวแปรตาม

รูปที่ 1.1

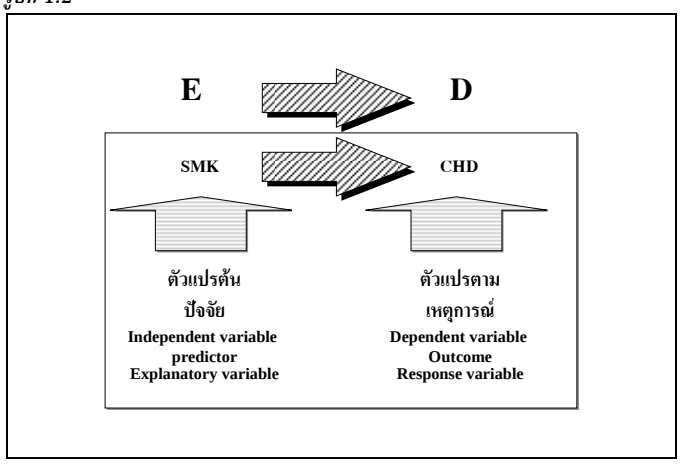


Logistic Regression เป็นวิธีการทางสถิติที่นิยมใช้มากในการวิจัยทางด้านการแพทย์และสาธารณสุข

เป้าหมายหลักของการวิเคราะห์ด้วยวิธีนี้คือเพื่ออธิบายความสัมพันธ์ระหว่างปัจจัย [Exposure (E)] และปัญหาที่ต้องการศึกษา ในที่นี้กำหนดให้เป็น “โรค” [Disease (D)]

ตัวอย่างเช่น ต้องการทราบความสัมพันธ์ระหว่างการป่วยด้วยโรคหัวใจโคโรนารี่ [Coronary Heart Disease (CHD)] กับการสูบบุหรี่ [Smoking (SMK)]

รูปที่ 1.2



กล่าวในทางสถิติ D คือตัวแปรตาม (Dependent Variable) ซึ่งค่าของ D จะเปลี่ยนไปตามอิทธิพลของตัวแปรต้น ตัวแปรตามนี้ มีชื่อเรียกอื่นอีกได้แก่ Outcome หรือ Response variable หรือเหตุการณ์ (event)

ส่วน E คือตัวแปรต้น (Independent Variable) เป็นตัวแปรที่ผู้ทำการศึกษาต้องการศึกษาว่ามีอิทธิพลต่อตัวแปรตามหรือไม่ ตัวแปรต้นนี้ มีชื่อเรียกอื่นอีกหลายชื่อได้แก่ Predictor หรือ Explanatory Variable หรือ ปัจจัย (factor)

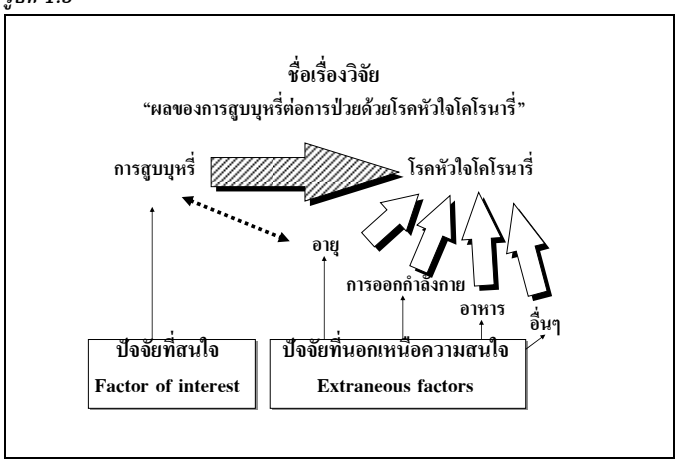
1.2 ตัวแปรที่นอกเหนือความสนใจ

ตัวอย่าง เช่น การเกิดโรคหัวใจโคโรนารี่ (CHD) เป็นตัวแปรตามโดยมีการสูบบุหรี่ (SMK) เป็นตัวแปรต้น หมายถึงเราสนใจศึกษาว่า “การสูบบุหรี่มีผลต่อการป่วยด้วย

โรคหัวใจโคโรนารี่หรือไม่”

คำถามดังกล่าว อาจเขียนเป็นชื่อเรื่องที่ทำวิจัยว่า “ผลของการสูบบุหรี่ต่อการป่วยด้วยโรคหัวใจโคโรนารี่ (Effect of smoking on coronary heart disease)” แต่การศึกษาที่ไม่คำนึงถึงปัจจัย (factor) อื่นใดนอกจากการสูบบุหรี่ มีโอกาสได้องค์ความรู้ที่ผิด (ดูรายละเอียดเพิ่มเติมใน บัณฑิต ถิ่นคำรพ, 2542) เพราะโรคหัวใจโคโรนารี่ มีสาเหตุมาจากหลายปัจจัย เช่นอายุ เพศการออกกำลังกาย อาหาร ฯลฯ และปัจจัยเหล่านี้ยังอาจเกี่ยวข้องซึ่งกันและกัน หรืออาจสัมพันธ์กับการสูบบุหรี่ด้วย

รูปที่ 1.3



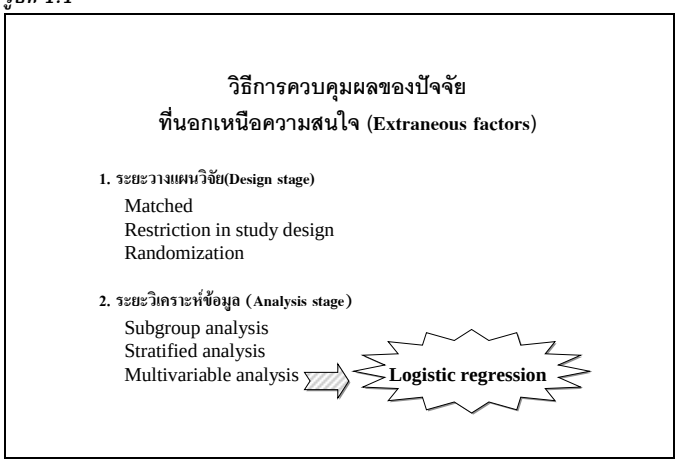
ปัจจัยที่นอกเหนือความสนใจ มีชื่อเรียกอื่นเช่น

- Extraneous factors
- Extraneous variables
- Covariates
- Controlled variables
- Confounders (แต่ชื่อนี้ไม่เหมาะ)

นั่นคือ การศึกษานี้ต้องมีการควบคุมผลของปัจจัยที่นอกเหนือความสนใจ (Extraneous factors) จึงจะสามารถบอกได้ถูกต้องเกี่ยวกับผลของการสูบบุหรี่ต่อการป่วยด้วยโรคหัวใจโคโรนารี่

1.3 การควบคุมผลของตัวแปรที่นอกเหนือความสนใจ

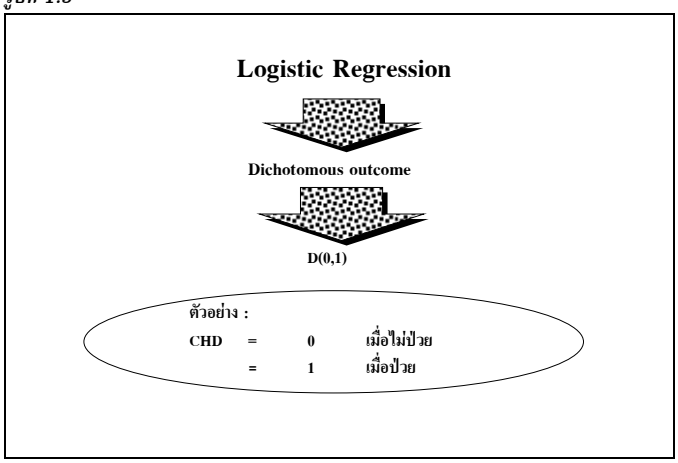
รูปที่ 1.4



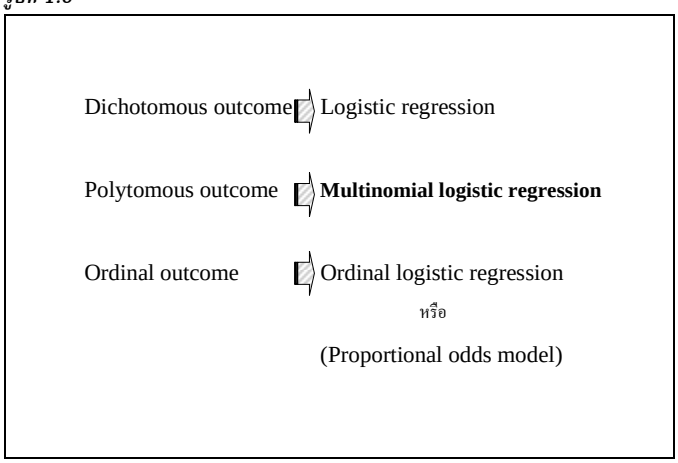
การควบคุมผลของปัจจัยอื่นกระทำได้ 2 กระบวนการหลัก ๆ คือ 1) โดยการวางรูปแบบการวิจัย กระทำก่อนการเก็บข้อมูล เช่น การจับคู่ (Matching) การจำกัดกลุ่มที่ศึกษา (Restriction) การจัดหน่วยตัวอย่างลงกลุ่มที่เปรียบเทียบโดยสุ่ม (Randomization) และ 2) โดยการวิเคราะห์ กระทำหลังการเก็บข้อมูล ได้แก่การวิเคราะห์แยกตามกลุ่มย่อย (Subgroup analysis) การวิเคราะห์ชั้นภูมิ (Stratified analysis) และการวิเคราะห์ตัวแปรพหุ(Multivariable analysis) การวิเคราะห์โดย Logistic regression เป็นวิธีการหนึ่งของ

1.4 Logistic regression ประเภทต่าง ๆ

รูปที่ 1.5



รูปที่ 1.6



Logistic Regression ใช้สำหรับวิเคราะห์ข้อมูลที่ตัวแปรตามเป็นตัวแปรแจงนับ (Categorical variable) เมื่อตัวแปรตาม มีค่าได้เพียงสองค่า หรือที่เรียกว่า ตัวแปรทวินาม (Dichotomous variable) คือ เมื่อไม่เกิดเหตุการณ์ D กำหนดให้ค่า D = 0 และเมื่อเกิดเหตุการณ์ D กำหนดให้ค่า D = 1

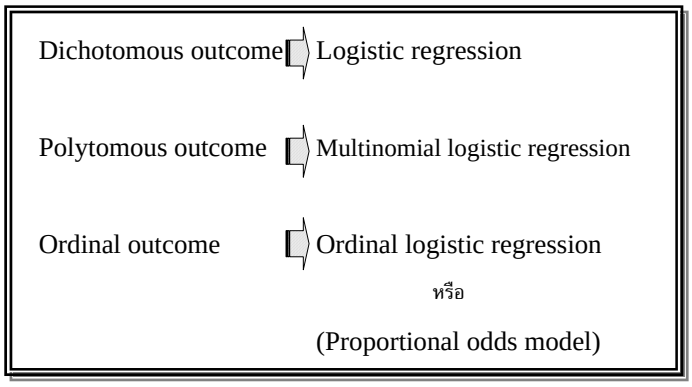
ตัวอย่าง เช่น ถ้า D คือการเกิดโรค Dจะมีค่าเป็น 0 หมายถึงไม่ป่วย และมีค่าเป็น 1 หมายถึงป่วย

ถ้า D คือผลการรักษา D จะมีค่าเป็น 0 หมายถึงไม่หาย และมีค่าเป็น 1 หมายถึงหาย เป็นต้น

ดังนั้น กรณีศึกษาความสัมพันธ์ระหว่างการสูบบุหรี่ กับการป่วยด้วยโรคหัวใจโคโรนารี ข้อมูลของตัวแปรตาม CHD = 0 เมื่อผู้เมื่อผู้ที่เราศึกษานั้น**ไม่ป่วย**เป็นโรคหัวใจโคโรนารี่ และ CHD = 1 เมื่อผู้ที่เราศึกษานั้น**ป่วย**เป็นโรคหัวใจโคโรนารี่ เป็นต้น

อย่างไรก็ตาม กรณีตัวแปรตามที่มีค่ามากกว่าสองค่าเช่น การเลือกใช้บริการรักษาพยาบาลเมื่อเจ็บป่วย (1 = ซื้อยากินเอง 2 = รักษาที่คลินิกเอกชน และ 3 = รักษาที่สถานบริการของรัฐ) ข้อมูลเป็นเพียงตัวเลขที่ใช้เป็นรหัสแทนกลุ่มและเรียงลำดับค่าไม่ได้ สามารถใช้ลอจีสติกรีเกรสชั่นที่เรียกว่า Multinomial logistic regression แต่ถ้าค่านั้นเรียงลำดับได้เช่น ความรุนแรงของโรค (1 = น้อย 2 = ปานกลาง และ 3 = มาก) จะใช้ Ordinal logistics regression (หรือ Proportional odds model) ซึ่งมีหลักการเดียวกัน แต่ในที่นี้กล่าวเฉพาะ Logistic regression ซึ่งหมายถึงกรณีตัวแปรตามมีได้

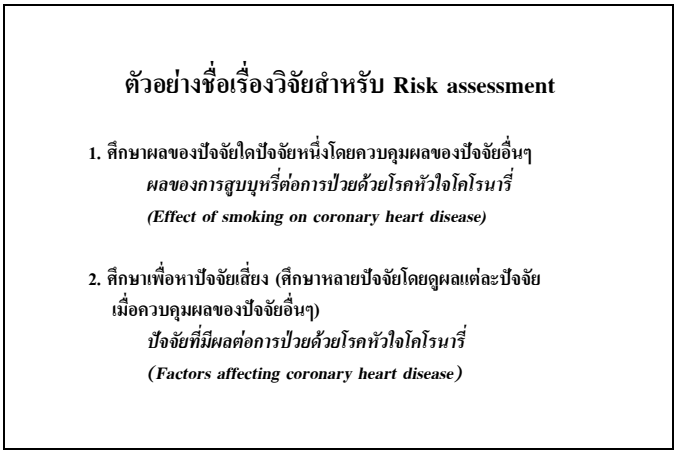
เพียงสองค่า (เป็น Dichotomous)



1.5 คำถามวิจัย

งานวิจัยที่ต้องใช้ Logistic regression วิเคราะห์ข้อมูลมี 2 ประเภทคือ วิจัยเพื่อหา

รูปที่ 1.7



ปัจจัยเสี่ยง (Risk assessment) และ วิจัยเพื่อสร้างสมการทำนายเหตุการณ์ (Prediction) ชุดการเรียนนี้กล่าวถึงเฉพาะการหาปัจจัยเสี่ยง

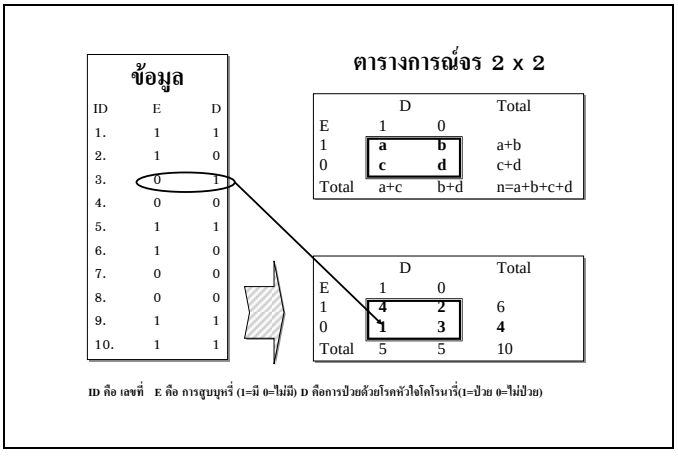
กรณีวิจัยเพื่อหาปัจจัยเสี่ยงยังมีลักษณะย่อยอีก 2 ลักษณะคือ มีและไม่มีตัวแปรที่สนใจ (Factor of interest) ตัวอย่างข้างต้น มีปัจจัยที่สนใจคือ การสูบบุหรี่ ถ้าไม่มีปัจจัยที่สนใจ เรื่องนี้จะมีชื่อเรื่องวิจัยว่า "ปัจจัยที่มีผลต่อการป่วยด้วยโรคหัวใจโคโรนารี่ (Factors affecting coronary heart disease)"

2. การวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรต้น (E) กับตัวแปรตาม (D)

ประเด็นหลักของการวิเคราะห์โดยใช้ Logistic regression นั้นคือการอธิบายความสัมพันธ์ระหว่างตัวแปรต้นกับตัวแปรตามที่มี***<u>การควบคุม</u>***ผลกระทบของปัจจัยอื่น ๆ ต่อไปนี้จะเป็นการทำความเข้าใจกับอิทธิพลของตัวแปรที่นอกเหนือความสนใจ และทำความเข้าใจความหมายของ***<u>การควบคุม</u>*** อิทธิพลของตัวแปรเหล่านั้น

2.1 Crude analysis

รูปที่ 1.8



การศึกษาผลของตัวแปรต้นตัวเดียวกับตัวแปรตามโดยไม่คำนึงถึงผลกระทบของปัจจัยอื่น ซึ่งเรียกว่า การวิเคราะห์อย่างหยาบ (Crude analysis) บางคนเรียก Bivariate analysis เพราะวิเคราะห์ตัวแปรเพียงสองตัว เป็นขั้นตอนแรกที่ต้องทำในทุกกรณีการศึกษา

เพื่อง่าย ให้ดูการศึกษาความสัมพันธ์ระหว่างตัวแปรต้นที่เป็น Dichotomous กับตัวแปรตามที่เป็น Dichotomous เช่นกัน มีวิธีดำเนินการดังนี้

เริ่มจากสร้างตารางความสัมพันธ์ระหว่าง E กับ D ได้ตารางการณ์จร (Contingency table) ในรูปตาราง 2×2 เช่น SMK มีสองกลุ่มคือ 0

การวิเคราะห์ความสัมพันธ์

1. การวัดระดับความสัมพันธ์ (Measure of association)

1.1) Relative Risk

1.2) Odds Ratio

2. การทดสอบความสัมพันธ์(Test of association)

มีสถิติทดสอบหลายประเภท

ขึ้นอยู่กับปัญหาที่จะทดสอบ เช่น

Chi-square test

Fisher's exact test

McNemar test

Binomial probability test เป็นต้น

Note: *Chi-square ไม่ valid หากมีเซลล์ที่มีค่า Expected value น้อยกว่า 5 มีจำนวนมากกว่า 20% ของจำนวนเซลล์ที่มีทั้งหมดในตาราง ในกรณีนี้ ต้องใช้ Fisher's Exact test เป็นสถิติทดสอบ*

รูปที่ 1.9

Crude analysis

SMK \ CHD	1	0	รวม
1	42	203	245
0	7	107	114
รวม	49	310	359 ➡ ขนาดตัวอย่าง (n)

ค่าสถิติที่ต้องคำนวณ:-

Cohort study	Cross-sectional และ Case-control study
X^2 = 7.99	X^2 = 7.99
ที่ df = 1 ได้ค่า p-value = 0.005	ที่ df = 1 ได้ค่า p-value = 0.005
RR = 2.8	OR = 3.2
95%CI : 1.3 ถึง 6.0	95% CI : 1.4 ถึง 7.1

คือไม่สูบบุหรี่ และ 1 คือสูบบุหรี่ และการป่วยด้วยโรคหัวใจโคโรนารี่มีสองกลุ่มคือ 0 คือไม่ป่วย และ 1 คือป่วย ตารางที่ได้จึงมี 4 ช่อง แต่ละช่องเรียกว่าเซลล์ (cell) ภายในเซลล์ใส่จำนวนคนที่มีลักษณะสอดคล้องกับค่าตัวแปรดังกล่าว สถิติที่ใช้ในการวิเคราะห์ความสัมพันธ์ระหว่าง E กับ D จำแนกเป็นสองประเภทคือ การทดสอบความสัมพันธ์ และการประมาณค่าขนาดความสัมพันธ์

การทดสอบความสัมพันธ์ (Test of association) ใช้สถิติไค-สแคว (χ^2-test) เว้นแต่เมื่อมีเซลล์ที่มีค่า Expected value น้อยกว่า 5 เกิน 20% ของจำนวนเซลล์ทั้งหมด ให้ใช้ Fisher's Exact test แทน

ขนาดความสัมพันธ์ (Measure of association) ใช้ Relative Risk (RR.) ถ้าเป็นการศึกษาแบบ Cohort study หรือใช้ Odds Ratio (OR) ถ้าเป็นการศึกษาแบบ Cross-sectional หรือ Case-control Study

ตัวอย่างการศึกษาความสัมพันธ์ระหว่างการสูบบุหรี่กับการป่วยด้วยโรคหัวใจโคโรนารี่ในคน 359 คน ทำการศึกษาแบบ Cohort study โดยติดตามคนปกติที่สูบบุหรี่จำนวน 245 คน (a+b) และที่ไม่สูบบุหรี่ จำนวน 114 คน (c+d) เป็นระยะเวลา 10 ปี พบว่ากลุ่มที่สูบบุหรี่ มีผู้ป่วย 42 คน (a) ส่วนที่ไม่สูบบุหรี่มีผู้ป่วย 7 คน (c) การศึกษารูปแบบนี้กำหนดจำนวนรวมตามแถวล่วงหน้า (Row totals are fixed)

ถ้าเป็นการศึกษาแบบ Cross-sectional study จะเป็นการสุ่มตัวอย่างคนจากชุมชนมาเท่ากับ 359 (n) แล้วตรวจโรคหัวใจและสอบถามพฤติกรรมการสูบบุหรี่ของแต่ละคน จากนั้นนำมาจำแนกกลุ่มเป็น 4 กลุ่ม ลงในตารางคือ a=42 b=203 c=7 และ d=107 การศึกษารูปแบบนี้กำหนดจำนวนรวมทั้งหมด

สูตรที่ควรทราบ: จากตาราง 2 x 2

	D+	D−	
E+	a	b	a+b
E−	c	d	c+d
	a+c	b+d	a+b+c+d=N

1. การทดสอบความสัมพันธ์ (Test of association)

$$\chi^2 = \frac{N\left[|ad-bc| - \frac{n}{2}\right]^2}{(a+b)(a+c)(b+d)(c+d)}$$

$$P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!a!b!c!d!}$$

2. มาตรวัดระดับความสัมพันธ์ (Measure of association)

2.1 Relative Risk

$$RR. = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

$$95\%CI.RR. = RR.\exp\left[\pm 1.96\sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}}\right]$$

2.2 Odds Ratio

$$OR. = \frac{ad}{bc}$$

$$95\%CI.OR. = OR.\exp\left[\pm 1.96\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right]$$

ล่วงหน้า (Grand totals is fixed)

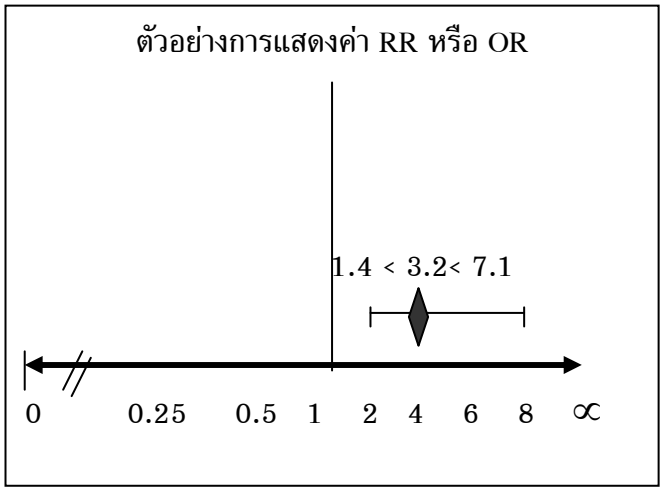
ประเภทสุดท้ายคือ การศึกษาแบบ Case-control เริ่มต้นด้วยหาผู้ป่วย 49 คน (a+c) แล้วสุ่มตัวอย่างผู้ที่ไม่ป่วย 310 คน (b+d) จากนั้นจึงถามประวัติการสูบบุหรี่ แล้วแจงนับจำนวนคนลงในตาราง การศึกษารูปแบบนี้กำหนดจำนวนรวมตามคอลัมน์ล่วงหน้า (Column totals are fixed)

สถิติทดสอบสำหรับการวิเคราะห์ความสัมพันธ์ระหว่าง SMK กับ CHD ใช้ χ^2-test เหมือนกันได้ (เว้นแต่ขนาดตัวอย่างน้อยต้องใช้ Fisher's exact test แทน) ไม่ว่าการศึกษานั้นเป็นรูปแบบ Cohort study หรือ Cross-sectional study หรือ Case-control study จากตัวอย่าง ค่า χ^2 = 7.99 นำไปเปิดตารางค่าสถิติการแจกแจงแบบไค-สแควร์ที่ขั้นความอิสระเท่ากับ (R-1)(C-1) = (2-1)(2-1) = 1 ได้ค่า p-value เท่ากับ 0.005 ถ้ากำหนดระดับนัยสำคัญเท่ากับ 0.05 แสดงว่าตัวแปรทั้งสองมีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติ (Significance)

โปรดสังเกตว่า การทดสอบทางสถิติมิได้บอกระดับความสัมพันธ์ บอกแต่เพียงว่าการศึกษาพบความสัมพันธ์ดังกล่าวเป็นความบังเอิญหรือไม่ ระดับความสัมพันธ์ ทราบได้โดยคำนวณหา RR. ในกรณี Cohort study ซึ่งได้เท่ากับ 2.8 คำนวณจากสูตร RR. = [a/(a+c)]/[b/(b+c)] แปลความหมายได้ว่า ***"ผู้ที่สูบบุหรี่มีความเสี่ยงต่อการป่วยโรคหัวใจโคโรนารี่สูงเป็น 2.8 เท่าของผู้ที่ไม่สูบบุหรี่"***

กรณีเป็นการศึกษาแบบ Cross-sectional หรือ แบบ Case-control study ไม่สามารถหา RR. ได้ โดยตรง จึงประมาณค่า RR. โดยใช้ค่า OR และแปลความหมาย

คล้ายกัน ค่า OR คำนวณได้เท่ากับ 3.2 จากสูตร OR = ad/bc

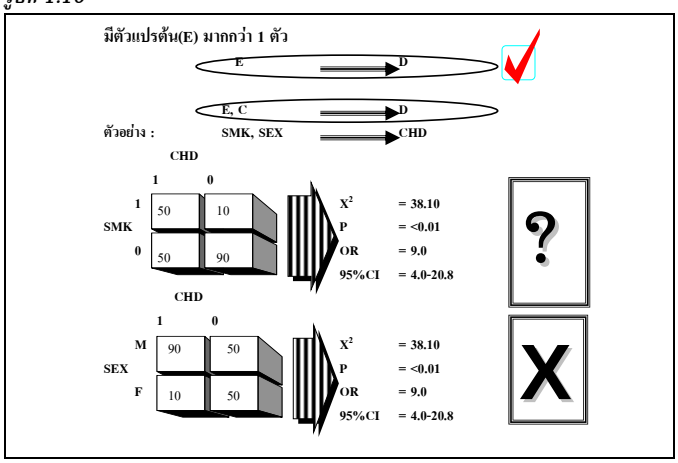


โปรดสังเกตว่าช่วงความเชื่อมั่นที่ระดับ 95% [95% Confidence Interval (95%CI.)] ของ OR ได้เท่ากับ 1.4 – 7.1 ซึ่งไม่คร่อมค่า 1 นั่นคือ ระดับความสัมพันธ์ที่ได้นี้ ต่างจาก 1 อย่างมีนัยสำคัญทางสถิติ ซึ่งให้ความหมายในตัวเองว่า ***การสูบบุหรี่ มีความสัมพันธ์กับการป่วยด้วยโรคหัวใจโคโรนารี่ อย่างมีนัยสำคัญทางสถิติ*** ที่ระดับนัยสำคัญ 0.05 แต่ที่มีประโยชน์ยิ่งกว่า p-value คือ RR หรือ OR บอกขนาดความสัมพันธ์ด้วย ในงานวิจัยที่ตัวอย่างมาก ๆ ผลที่บอกว่ามีนัยสำคัญทางสถิติ (p-value < 0.05) อาจมีค่า RR หรือ OR ที่ต่ำมากก็ได้

ดังนั้น ค่า 95% CI ของ RR. หรือของ OR จึงควรแสดงไว้ทุกครั้งในการวิเคราะห์ข้อมูล

ที่กล่าวแล้ว เป็นการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรต้นหนึ่งตัวแปรกับตัวแปรตาม ซึ่งมีตัวแปรเดียวเสมอ จากตัวอย่างข้างต้นถ้าคำถามวิจัยสนใจศึกษาผลของการสูบบุหรี่ เพศมักเกี่ยวข้องกับการสูบบุหรี่และอาจสัมพันธ์กับการป่วยด้วยโรคหัวใจโคโรนารี่ด้วย

รูปที่ 1.10



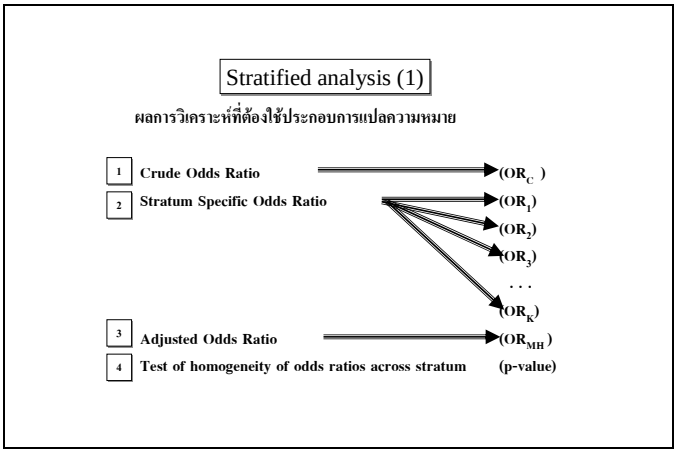
กรณีมีตัวแปรต้นมากกว่าหนึ่งเช่นนี้ การวิเคราะห์ความสัมพันธ์ทีละคู่ เช่น SMK กับ CHD และ SEX กับ CHD นั้น อาจนำไปสู่ข้อสรุปที่ผิดพลาด จากตัวอย่าง คนที่สูบบุหรี่ มีโอกาสป่วยด้วยโรคหัวใจโคโรนารี่ 9 เท่าของผู้ไม่สูบบุหรี่ ในขณะที่ ชาย มีโอกาสป่วยด้วยโรคดังกล่าว สูงเป็น 9 เท่าของเพศหญิงเช่นกัน ประเด็นคือผลที่ได้จากการศึกษาความสัมพันธ์ระหว่าง SMK กับ CHD ยังเป็นที่สงสัยอยู่ว่าจริงตามผลที่วิเคราะห์ได้หรือไม่ เพราะไม่ได้

ควบคุมผลกระทบจากเพศและตัวแปรอื่นใด อีกประการหนึ่งการวิเคราะห์ความสัมพันธ์ระหว่าง SEX กับ CHD นั้นไม่ได้อยู่ในความสนใจของการศึกษา

2.2 Stratified analysis

การวิเคราะห์ความสัมพันธ์ระหว่าง E กับ D โดยควบคุมอิทธิพลของ C (เมื่อ C แทน Covariate) จากตัวอย่างข้างต้น C คือ SEX ทำได้โดยการวิเคราะห์จำแนกชั้นภูมิ (Stratified analysis) กล่าวคือดูความสัมพันธ์ระหว่างการ E กับ D จำแนกตามกลุ่มของ C [ผู้อ่านที่สนใจศึกษาทฤษฎีเกี่ยวกับการวิเคราะห์แบบ Stratified analysis ดูได้ใน Kleinbaum, Kupper, and Morgenstern (1986) หน้า 321-376 หรือที่ง่ายกว่านั้นใน Fleiss (1981) หน้า 160-187]

รูปที่ 1.11



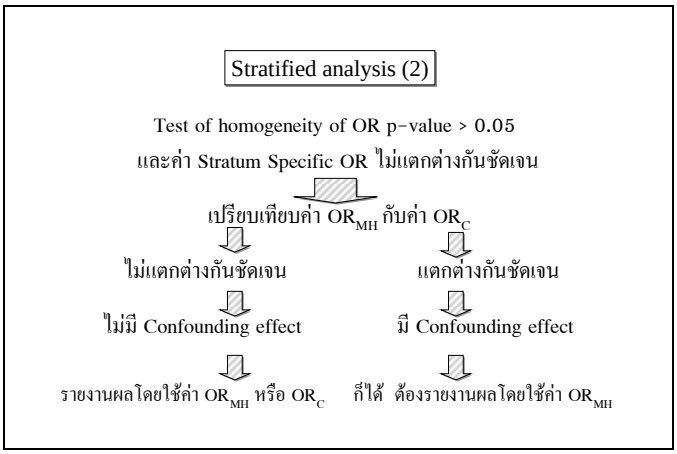
แต่ถ้าข้าม ก็ไม่ทำให้ยากต่อการเข้าใจการประยุกต์ใช้วิธีการดังกล่าวนี้แต่อย่างใด)

ค่าสถิติที่พึงต้องทราบคือค่า : ❶ Crude odds ratio (OR_C) คือค่าที่ได้จาก Crude analysis ตามที่กล่าวข้างต้น, ❷ Stratum specific odds ratio (OR_1, OR_2, ..., OR_K เมื่อตัวแปร C มี K กลุ่ม), ❸Adjusted odds ratio (OR_{MH} เมื่อ MH ห้อยให้เกียรติผู้คิดค้นวิธีการนี้คือ Mantel-Haenszel), และ ❹P-value จากการทดสอบว่าค่า Stratum specific odds ratio เท่ากันหรือไม่ (Test of homogeneity of odds ratios across stratum ชื่อ Woolf's test)

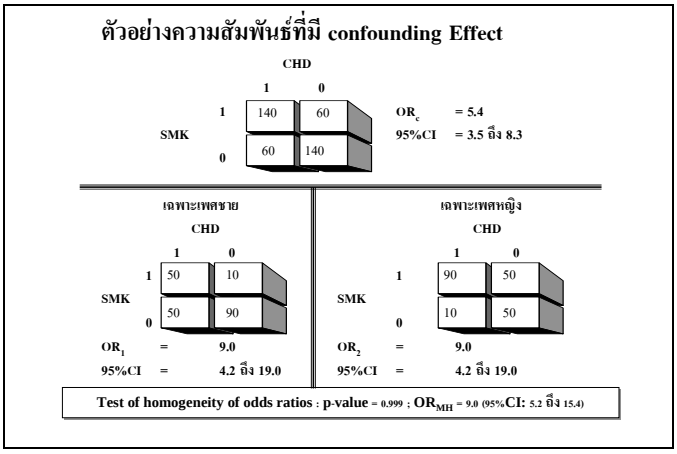
2.2.1 Confounding Effect

จากค่าทั้ง 4 ค่า ให้ดูที่ข้อ❹เป็นอันดับแรก ถ้าค่า p-value > 0.05 หมายถึงค่า Stratum specific OR ของแต่ละ Strata แตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ แต่

รูปที่ 1.12



รูปที่ 1.13



เนื่องจากอำนาจการทดสอบของสถิตินี้ต่ำ ลำดับต่อไปจึงให้ดูค่า Stratum specific OR ที่ข้อ❷ ถ้าหากแต่ละค่าไม่ต่างกันมาก (ใช้วิจารณญาณ) นั่นหมายถึงเราสามารถเฉลี่ยค่า Stratum specific OR เหล่านั้นได้ ดังนั้นลำดับต่อไปจึงดูค่า OR_{MH} ในข้อ❸

ถ้าหากค่า OR_{MH} ต่างจาก OR_C ในข้อ❶อย่างชัดเจน (ใช้วิจารณญาณ) หมายถึงมี Confounding effect โดยที่ C เป็น Confounder ต้องอธิบายความสัมพันธ์ระหว่าง E กับ D โดยใช้ OR_{MH}

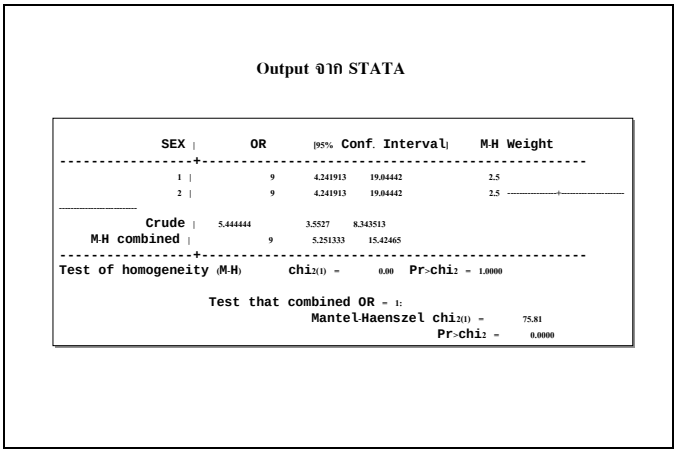
ตรงข้าม ถ้าหาก OR_{MH} ไม่ต่างจาก OR_C การอธิบายความสัมพันธ์นั้นสามารถใช้ OR_{MH} หรือ OR_C อย่างใดอย่างหนึ่ง แต่แนะนำให้ใช้ OR_{MH} เพราะสื่อกับผู้อ่านผลวิจัยได้โดยตรงว่าได้ควบคุมผลของ C แล้ว เว้นแต่ช่วงเชื่อมั่นของ OR_C จะแคบกว่า OR_{MH}

ตัวอย่างเช่น การศึกษาพบว่า ผู้สูบบุหรี่ มีความเสี่ยงต่อการป่วยสูงเป็น 5.4 เท่าของผู้ไม่สูบบุหรี่ เป็นการสรุปจากค่า OR_C แต่เมื่อแยกวิเคราะห์ตามเพศ (SEX) จะได้ OR_1 เป็นค่าของ Strata ที่ 1 คือเพศชาย เท่ากับ 9 และ OR_2 เป็นค่าของ Strata ที่ 2 คือเพศหญิง เท่ากับ 9 ผลการทดสอบ $OR_1 = OR_2$ ได้ค่า p-value = 0.999 และ OR_{MH} = 9.0

ค่าสถิติข้างต้น ค่า Test of homogeneity of odds ratios และค่า Stratum-specific odds ratios บ่งชี้ว่าค่าเหล่านี้ไม่แตกต่างกัน ดังนั้นจึงพิจารณาค่า OR_{MH} เปรียบเทียบกับ OR_C พบว่าแตกต่างกันชัดเจน (9.0 กับ 5.4 ตามลำดับ) บอกให้เราทราบว่ามี Confounding effect เกิดขึ้น โดยที่เพศเป็น Confounder ของความสัมพันธ์ระหว่างการสูบบุหรี่กับการป่วยด้วยโรคหัวใจโคโรนารี่ ต้องใช้ OR_{MH} อธิบายความสัมพันธ์ที่ศึกษา ดังนี้

"เมื่อควบคุมผลกระทบจากเพศแล้ว การ

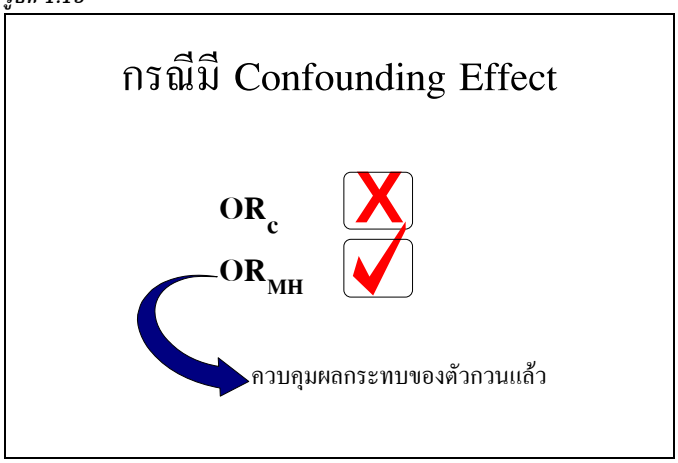
รูปที่ 1.14



สูบบุหรี่มีความสัมพันธ์กับการป่วยด้วยโรคหัวใจโคโรนารี่อย่างมีนัยสำคัญทางสถิติ (p-value < 0.001) กล่าวคือ ผู้สูบบุหรี่มีความเสี่ยงต่อการป่วยด้วยโรคหัวใจโคโรนารี่ สูงเป็น 9 เท่าของผู้ไม่สูบบุหรี่ (95%CI: 5.2 ถึง 15.4)"

ค่า p-value ที่รายงานในที่นี้เป็น Mantel-Haenszel chi-square test ที่ทดสอบว่า OR_{MH} ต่างจาก 1 หรือไม่ และผลทั้งหมดที่กล่าวแล้วมีรายงานในผลการวิเคราะห์โดยใช้โปรแกรม STATA (ดูรายละเอียดในแบบฝึกหัดที่ 1)

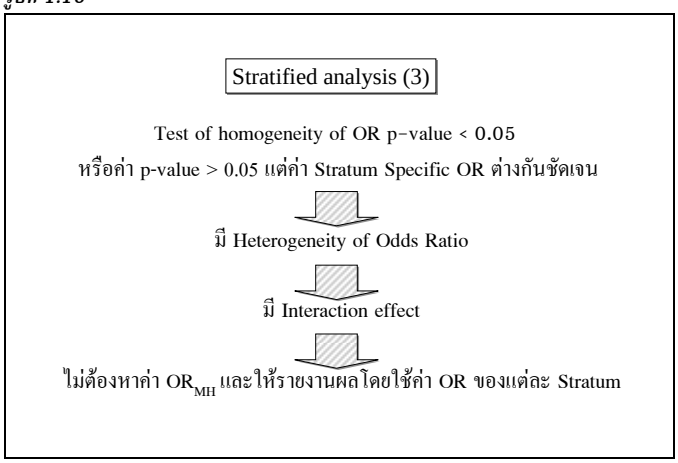
รูปที่ 1.15



โดยสรุป เมื่อมีตัวกวน หรือมี Confounding Effect การอธิบายความสัมพันธ์ระหว่าง E กับ D จะต้องไม่ใช้ค่าระดับความสัมพันธ์อย่างหยาบ (OR_C) เพราะไม่ถูกต้อง คือ อาจสูงหรือต่ำกว่าความเป็นจริง จะต้องใช้ค่าที่ควบคุมผลกระทบจากตัวกวนแล้ว หรือ OR_{MH} เท่านั้น และค่า OR_{MH} นี้แท้จริงคือค่าเฉลี่ยแบบถ่วงน้ำหนักของค่า OR ของแต่ละกลุ่มของตัวแปร C นั่นเอง

2.2.2. Interaction effect

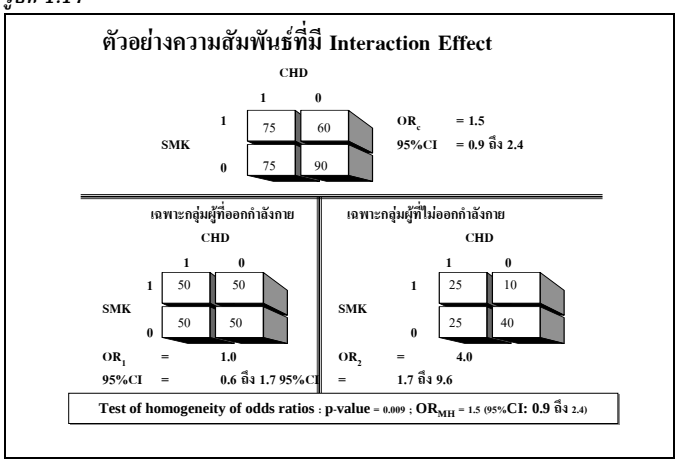
รูปที่ 1.16



ในทางตรงข้าม ถ้าหากค่า p-value < 0.05 หรือค่าของ Stratum Specific Odds Ratio ไม่เท่ากันอย่างชัดเจน บ่งชี้ว่ามี Heterogeneity of odds ratio แสดงว่ามี Interaction effect กล่าวคือ ความสัมพันธ์ระหว่าง E กับ D ขึ้นอยู่กับแต่ละระดับของ C หรือ C เป็น Effect modifier ของความสัมพันธ์ระหว่าง E กับ D

ตัวอย่างการศึกษาความสัมพันธ์ระหว่าง

รูปที่ 1.17



การสูบบุหรี่ (SMK) กับการป่วยด้วยโรคหัวใจโคโรนารี่(CHD) พบว่า ผู้สูบบุหรี่มีความเสี่ยงต่อการเกิดโรคหัวใจโคโรนารีสูงเป็น 1.5 เท่าของผู้ไม่สูบบุหรี่ แต่เมื่อจำแนกดูความสัมพันธ์ดังกล่าวในผู้ออกกำลังกาย (ACT) ความเสี่ยงของผู้สูบบุหรี่เท่ากับ 1 เมื่อเทียบกับ ในผู้ไม่ออกกำลังกาย เท่ากับ 4 บ่งชี้ว่า ความสัมพันธ์ระหว่างการสูบบุหรี่ (SMK) กับการป่วยด้วยโรคหัวใจโคโรนารี่ (CHD) ขึ้นอยู่กับว่าเป็นผู้ออกกำลังกาย (ACT) หรือไม่

เราเรียกตัวแปร C ใด ๆ ที่ยังผลให้เกิด Interaction effect ระหว่าง E กับ D ว่า Effect Modifier จากตัวอย่าง การออกกำลังกาย เป็น Effect Modifier ของความสัมพันธ์ระหว่างการสูบบุหรี่ (SMK) กับการป่วยด้วยโรคหัวใจโคโรนารี่ (CHD)

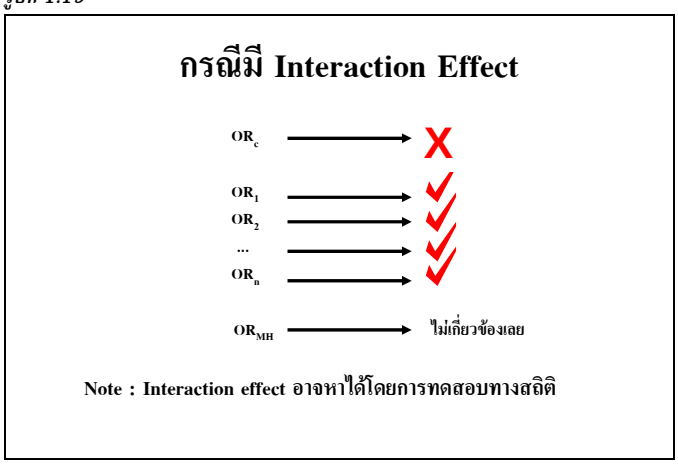
รูปที่ 1.18

Output จาก STATA

```
          ACT |      OR   [95% Conf. Interval]    M-H Weight
--------------+-------------------------------------------------
            1 |       1    .5754505    1.737769         12.5
            2 |       4    1.663145    9.596975          2.5
--------------+-------------------------------------------------
        Crude |     1.5    .9504315    2.367332
  M-H combined |     1.5    .949093     2.370685
--------------+-------------------------------------------------
Test of homogeneity (M-H)       chi2(1) =     6.75  Pr>chi2 = 0.0094

                Test that combined OR = 1:
                        Mantel-Haenszel chi2(1) =     3.07
                                 Pr>chi2 =   0.0796
```

ผลทั้งหมดที่กล่าวแล้วจากตัวอย่างข้างต้นเกี่ยวกับการศึกษาความสัมพันธ์ระหว่างการสูบบุหรี่ (SMK) กับการป่วยด้วยโรคหัวใจโคโรนารี่(CHD) เมื่อพิจารณาผลของการออกกำลังกาย (ACT) มีรายงานในผลการวิเคราะห์โดยใช้โปรแกรม STATA (ดูรายละเอียดในแบบฝึกหัดที่ 1)

รูปที่ 1.19

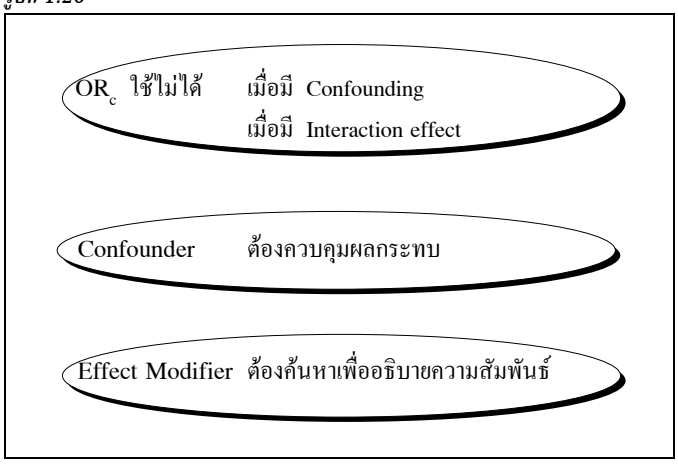


เมื่อมี Interaction effect การอธิบายความสัมพันธ์ระหว่าง E กับ D โดยใช้ค่า Crude Odds Ratio นั้นไม่เหมาะสม ต้องอธิบายโดยใช้ Stratum Specific Odds Ratio เท่านั้น กล่าวคือ อธิบายความสัมพันธ์ดังกล่าวจำแนกตามแต่ละระดับของ Effect Modifier ส่วนค่า Adjusted Odds Ratio นั้น ไม่มีส่วนเกี่ยวข้องใด ๆ เลย

เมื่อแต่ละกลุ่มมีความเสี่ยงแตกต่างกันไป การอธิบายความสัมพันธ์โดยใช้ค่าเฉลี่ยของขนาดความสัมพันธ์ (OR_{MH}) จึงไม่มี

ความหมายใด ๆ ตัวอย่างสุดขั้วกว่าที่แสดงข้างต้นนี้เช่น E มีผลเป็น Risk effect (คือ OR มากกว่า 1) ในกลุ่ม C=1 แต่มีผลเป็น Protective effect (คือ OR น้อยกว่า 1) ในกลุ่ม C=2 กรณีนี้ต้องอธิบายความสัมพันธ์ดังกล่าว จำแนกตามระดับของ C เพราะองค์ความรู้นี้นำไปสู่การแก้ไขปัญหาที่แตกต่างกันไปตามกลุ่มของ C

รูปที่ 1.20

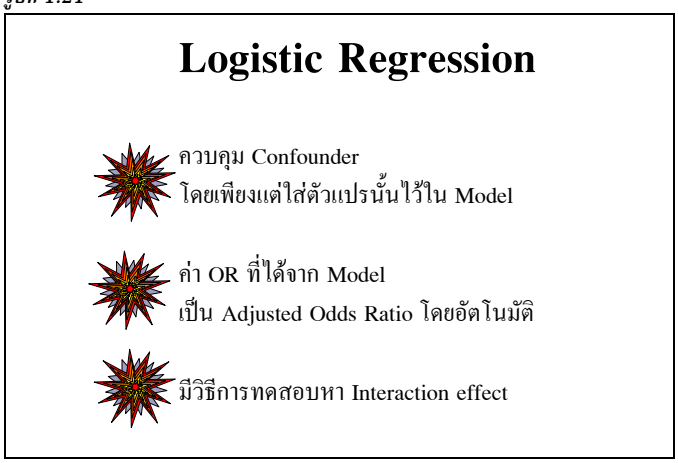


โดยสรุป การวิเคราะห์ข้อมูลใด ๆ จะต้องควบคุมผลกระทบของตัวแปรที่เป็น Confounder พร้อมกับค้นหาว่ามี Effect Modifier หรือไม่ การใช้ผลจากการวิเคราะห์คราวละสองตัวแปร จะได้เพียงค่าระดับความสัมพันธ์อย่างหยาบ ซึ่งจะนำไปสู่การอธิบายความสัมพันธ์ระหว่าง E กับ D ไม่ถูกต้อง

การวิเคราะห์แบบ Stratified analysis นี้ใช้ได้เฉพาะกรณีทั้งตัวแปรต้น E และตัวแปรตาม D เป็น Dichotomous โดยที่ตัวแปรควบคุม C เป็น Dichotomous ตามตัวอย่างที่ผ่านมา หรือ เป็น Polytomous คือมีมากกว่า 2 กลุ่มก็ได้ ดังนั้นจำนวน Strata เท่ากับจำนวนระดับของตัวแปร C

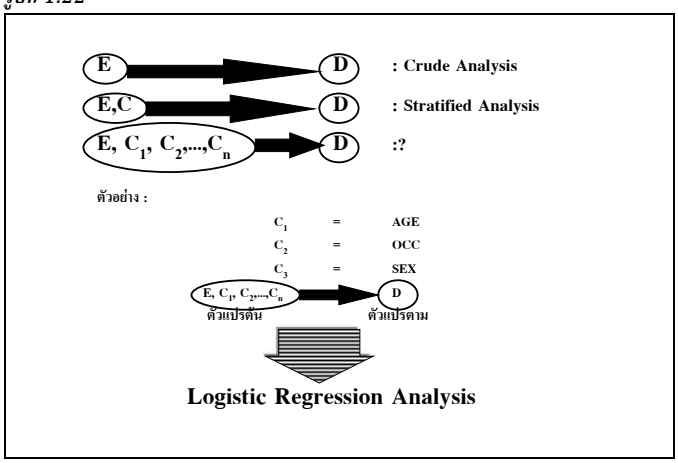
3. บทบาทของ Logistic regression

รูปที่ 1.21



ทั้งกรณีมี Confounding effect และ Interaction effect ตามกรณีที่กล่าวข้างต้น สามารถใช้ Logistic Regression วิเคราะห์ได้โดยเพียงแต่ให้ตัวแปรที่พบว่าเป็น Confounder เข้าไปใน Logistic regression model ก็เป็นการควบคุมผลกระทบจากตัวแปรดังกล่าว ได้ค่า OR ก็เป็น Adjusted Odds Ratio โดยอัตโนมัติ นอกจากนี้ ยังสามารถทดสอบหา ว่ามี Interaction effect หรือไม่ได้เช่นกัน ซึ่งจะกล่าวในรายละเอียดในชุดการเรียนต่อไป

รูปที่ 1.22



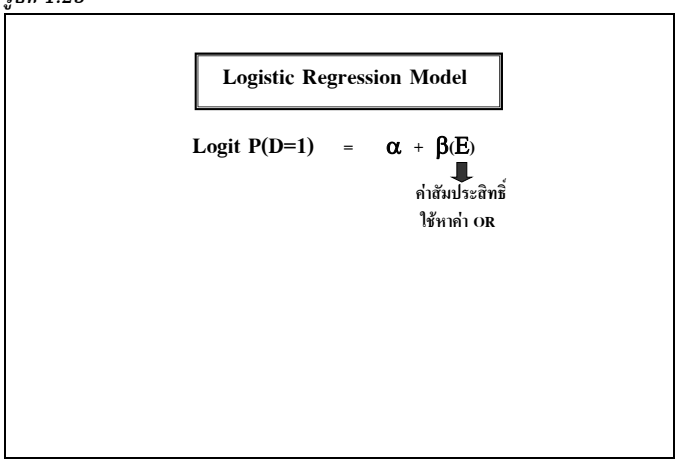
ที่กล่าวไปแล้วข้างต้น เป็นการวิเคราะห์ความสัมพันธ์ระหว่าง E กับ D โดยไม่ควบคุมผลกระทบจากตัวแปรอื่นใด หรือ Crude analysis และโดยควบคุมผลกระทบจากตัวแปรอื่น (C) เพียงตัวแปรเดียวโดยวิธี Stratified analysis

อย่างไรก็ตาม ในความเป็นจริงนั้น C มักมีมากกว่า 1 ตัวแปร เช่นอายุ (AGE) อาชีพ (OCC) เพศ (SEX) ฯลฯ บางตัวแปรเช่นอายุเป็นตัวแปรต่อเนื่อง กรณีเช่นนี้การวิเคราะห์แบบStratified analysis ไม่สามารถทำได้ จึงต้องใช้ Logistic Regression

จาก Model นี้ สามารถคำนวณหา OR ได้โดยตรงจากค่าสัมประสิทธิ์ใน Model รวมทั้งค่า 95% CI ของทั้งสองค่าดังกล่าว และการทดสอบนัยสำคัญทางสถิติ

ผลที่ได้จาก Logistic regression จะเท่ากันและได้ข้อสรุปเหมือนกันกับกล่าวแล้วข้างต้น

รูปที่ 1.23



ต่อไปนี้เป็นการแสดงให้เห็นว่าผลเหมือนกันอย่างไร โดยใช้โปรแกรม STATA (อักษรทึบคือคำสั่ง อักษรธรรมดาคือผล ที่ล้อมรอบโดยวงรีคือค่า OR และ 95%CI ที่สนใจ)

รูปที่ 1.24

ข้อมูลจากตัวอย่างกรณีมี Confounding effect นำมา วิเคราะห์ Crude analysis โดยใช้ STATA

```
. cc CHD SMK

                |  SMK                    |           Proportion
                |   Exposed   Unexposed   |    Total     Exposed
----------------+-------------------------+-------------------------
          Cases |       140          60   |      200      0.7000
       Controls |        60         140   |      200      0.3000
----------------+-------------------------+-------------------------
          Total |       200         200   |      400      0.5000
                |                         |
                |      Point estimate     | [95% Conf. Interval]
                |-------------------------+-------------------------
     Odds ratio |          5.444444       |    3.5527    8.343513  (Cornfield)
Attr. frac. ex. |           .8163265      |    .7185239    .8801464  (Cornfield)
Attr. frac. pop |           .5714286      |
                +-------------------------+-------------------------
                          chi2(1) =    64.00  Pr>chi2 = 0.0000
```

จากตัวอย่างการศึกษาความสัมพันธ์ระหว่างการสูบบุหรี่ (SMK) กับการป่วยด้วยโรคหัวใจโรโรนารี่ (CHD) โดยมีเพศ (SEX) เป็น Confounder ที่แสดงข้างต้น ผลจาก Crude analysis โดยวิเคราะห์ข้อมูลจากตารางการณ์จรธรรมดา ได้ $OR_C = 5.4$

รูปที่ 1.25

ข้อมูลจากตัวอย่างกรณีมี Confounding effect นำมาวิเคราะห์ Univariate Logistic regression
โดยใช้ STATA (โปรดสังเกตว่าให้ผลเช่นเดียวกันกับ Crude analysis)

```
. logistic CHD  SMK

Logit estimates                                   Number of obs   =        400
                                                  LR chi2(1)      =      65.83
                                                  Prob > chi2     =     0.0000
Log likelihood = -244.34572                       Pseudo R2       =     0.1187

------------------------------------------------------------------------------
         CHD | Odds Ratio   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         SMK |   5.444444   1.188075     7.766   0.000     3.549819    8.350279
------------------------------------------------------------------------------
```

Crude analysis วิเคราะห์โดย Logistic regression คือมีตัวแปรต้นเพียงตัวเดียว (คือ SMK) ใน Model ให้ผลเช่นเดียวกันกับผลข้างต้น คือ OR_C = 5.4 (โปรดสังเกตว่า CHD คือตัวแปรตาม ต้องอยู่ใน Model ในทุกกรณี)

รูปที่ 1.26

ข้อมูลจากตัวอย่างกรณีมี Confounding effect นำมา วิเคราะห์ Stratified analysis โดยใช้ STATA

```
. cc  CHD  SMK, by(SEX)

           SEX |      OR     [95% Conf. Interval]    M-H Weight
---------------+--------------------------------------------------
             1 |       9     4.241913   19.04442          2.5 (Cornfield)
             2 |       9     4.241913   19.04442          2.5 (Cornfield)
---------------+--------------------------------------------------
         Crude |   5.444444     3.5527   8.343513            (Cornfield)
  M-H combined |       9     5.251333   15.42465
---------------+--------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     0.00  Pr>chi2 = 1.0000

                Test that combined OR = 1:
                        Mantel-Haenszel chi2(1) =     75.81
                                         Pr>chi2 =    0.0000
```

Stratified analysis ตามวิธีการที่กล่าวข้างต้น โดย SEX คือ Stratified variable ได้ OR_{MH} = 9.0 หมายถึง OR แสดงระดับความสัมพันธ์ระหว่าง SMK กับ CHD เมื่อควบคุมผลของ SEX

รูปที่ 1.27

ข้อมูลจากตัวอย่างกรณีมี Confounding effect นำมาวิเคราะห์ Multiple Logistic regression
โดยใช้ STATA (โปรดสังเกตว่าให้ผลเช่นเดียวกันกับ Stratified analysis)

```
. logistic CHD  SMK SEX

Logit estimates                                   Number of obs   =        400
                                                  LR chi2(2)      =      81.40
                                                  Prob > chi2     =     0.0000
Log likelihood = -236.55918                       Pseudo R2       =     0.1468

------------------------------------------------------------------------------
         CHD | Odds Ratio   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         SMK |          9   2.473863     7.994   0.000     5.251333    15.42465
         SEX |        .36   .0989545    -3.717   0.000      .2100533    .6169862
------------------------------------------------------------------------------
```

Multiple logistic regression ให้ผลเช่นเดียวกันกับผลข้างต้น คือ OR_{SMK} = 9.0 ซึ่งได้จาก Model ที่มีทั้งตัวแปร SMK และ SEX หมายถึง OR แสดงระดับความสัมพันธ์ระหว่าง SMK กับ CHD เมื่อควบคุมผลของ SEX

สรุปท้ายบท :

เอกสารอ้างอิงประจำบทที่ 1

บัณฑิต ถิ่นคำรพ. (2541). ความสำคัญและความจำเป็นของการวิเคราะห์ข้อมูลตัวแปรเชิงพหุ. *วารสารระบาดวิทยาภาคตะวันออกเฉียงเหนือ*. 3(3) :20-25.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions*. 2nd edition. New York: John Willey & Sons.

Kleinbaum, D.G., Kupper, L.L., and Morgenstern, H. (1982). *Epidemiologic research: principles and qualitative methods*. London: Lifetime Learning Publications.

แบบฝึกหัดที่ 1

1. จากบทที่ 1 ที่กล่าวข้างต้น ผลการวิเคราะห์ตามรูปที่ 1.24 ถึง 1.27 ได้จากการวิเคราะห์ข้อมูลจากรูปที่ 1.13 ซึ่งต้องป้อนข้อมูลเข้าโปรแกรม STATA ก่อน

1.1 จงทำความเข้าใจแผนภูมิต่อไปนี้
ต่อไปนี้เป็นภาพแสดงความสัมพันธ์ระหว่างข้อมูลที่แจงนับลงตารางแล้ว และข้อมูลที่จะลงในแฟ้มข้อมูล STATA จากนั้นแสดงขั้นตอนการนำข้อมูลเข้า

ข้อมูลจากที่แจงนับลงในตารางแล้ว

MALE

	CHD+	CHD-	Total
SMK+	50	10	60
SMK-	50	90	140
Total	100	100	200

FEMALE

	CHD+	CHD-	Total
SMK-	90	50	140
SMK-	10	50	60
Total	100	100	200

กำหนดโครงสร้างแฟ้มข้อมูล

ตัวแปร	รหัส ความหมาย
SEX	1=Male, 2=Female
SMK	0=No, 1=Yes
CHD	0=No, 1=Yes

ข้อมูลที่ต้องป้อนลงใน STATA

sex	smk	chd	n
1	1	1	50
1	1	0	10
1	0	1	50
1	0	0	90
2	1	1	90
2	1	0	50
2	0	1	10
2	0	0	50

ลักษณะแบบสอบถามที่เป็นที่มาของข้อมูลชุดนี้

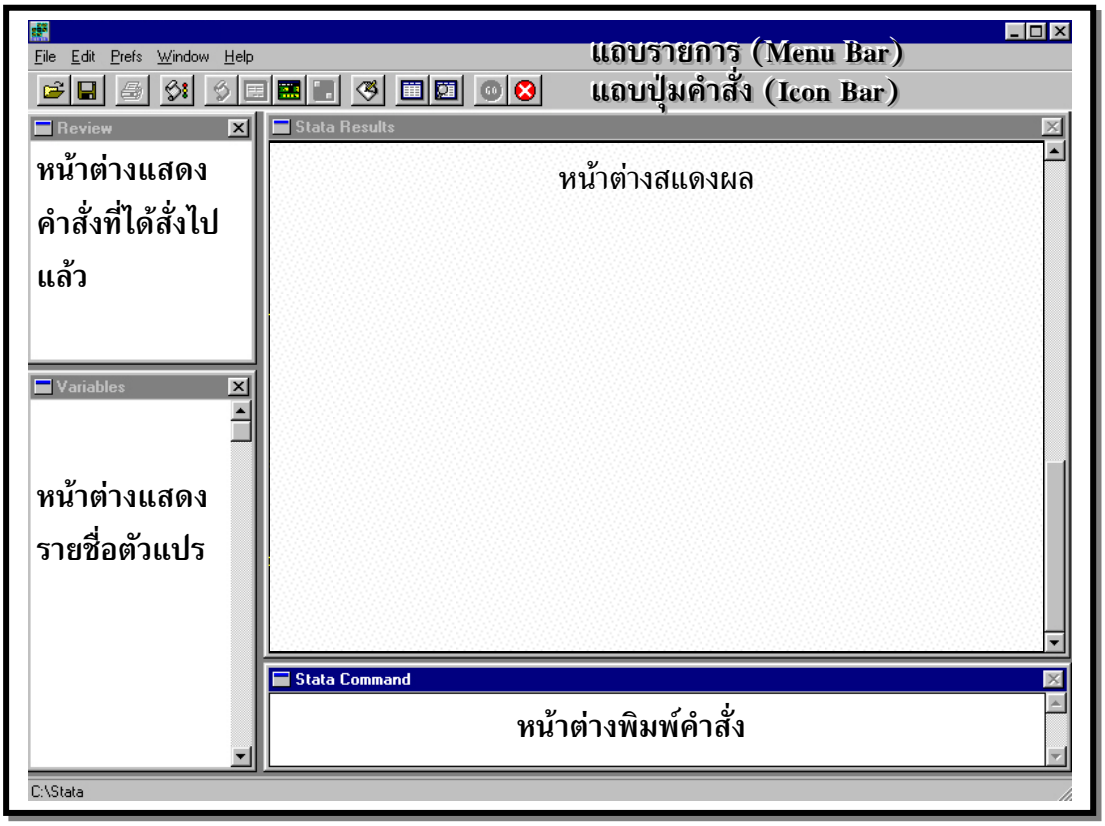
แบบสอบถาม **เลขที่ 001**

คำถาม	รหัส
1. เพศ [ ]1. ชาย [ ]2. หญิง	SEX [ ]
2. การสูบบุหรี่ [ ]0. ไม่สูบ [ ]1. สูบ	SMK [ ]
3. การป่วยด้วยโรคหัวใจโคโรนารี่ [ ]0. ไม่ป่วย [ ]1. ป่วย	CHD [ ]

1.2 จงดำเนินการนำข้อมูลเข้าคอมพิวเตอร์ตามขั้นตอนต่อไปนี้

ขั้นตอนการป้อนข้อมูลเข้า STATA (ต่อไปนี้ ทุกคำสั่งของ STATA จะแสดงไว้หลังจุด แต่การพิมพ์คำสั่งลงใน STATA ไม่ต้องพิมพ์จุด)

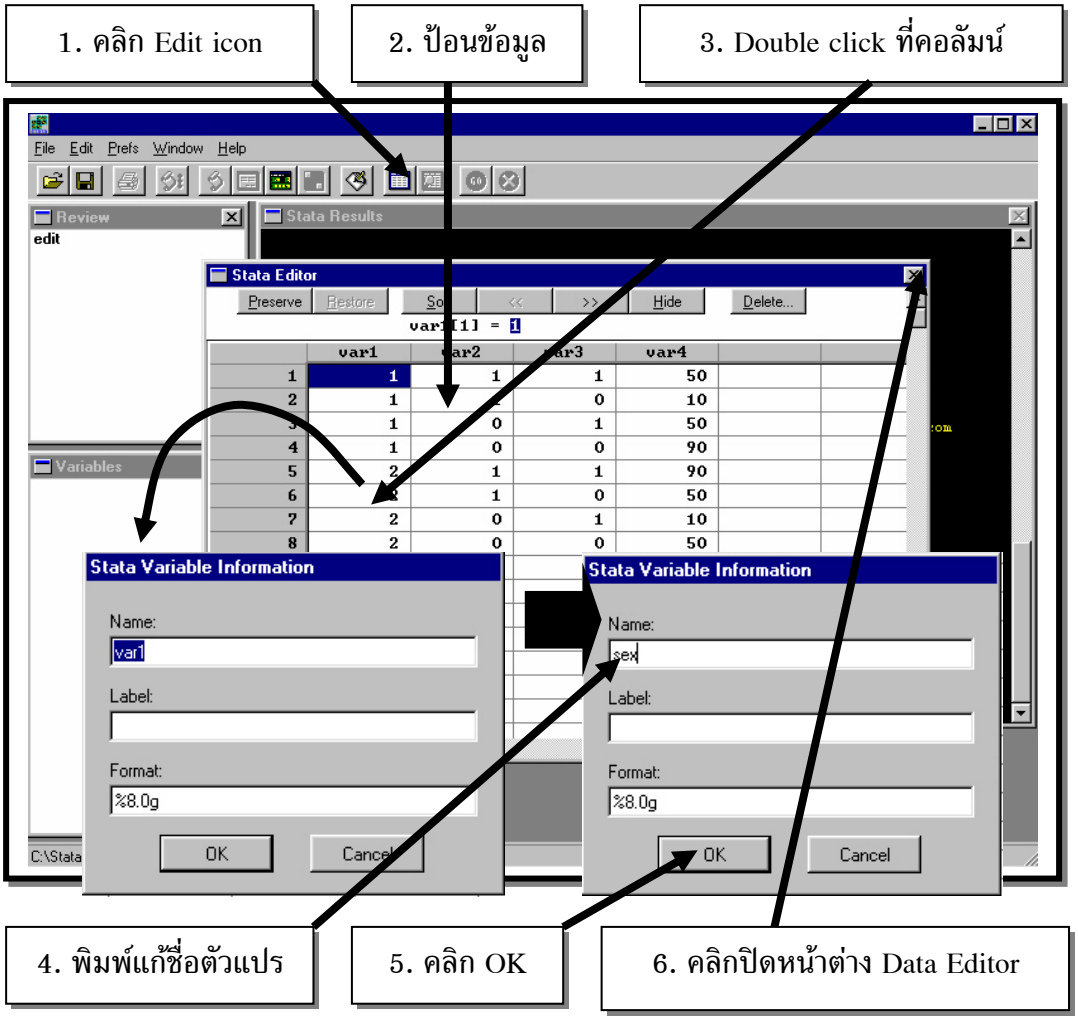
1.2.1 เปิดโปรแกรม STATA ได้หน้าจอดังนี้



1.2.2 ที่หน้าต่างพิมพ์คำสั่ง พิมพ์

.edit <ENTER> จะได้หน้าต่างป้อนข้อมูล STATA Editor

1.2.3 พิมพ์ข้อมูลจากข้อ 1.1 ลงไปจนครบทุกคอลัมน์ แต่ละคอลัมน์ ใช้เมาส์คลิกสองครั้งเพื่อเปลี่ยนชื่อตัวแปรตามที่ต้องการ (ตัวแปรที่กำหนดโดยอัตโนมัติจาก STATA คือ var<X> เมื่อ X คือหมายเลขตั้งแต่ 1 เป็นต้นไป) จากนั้นคลิกมุขวาบนสุดของหน้าต่าง STATA Editor เพื่อจบการป้อนข้อมูล กลับมาที่หน้าต่างพิมพ์คำสั่ง มีลำดับการทำงานตามรูปข้างล่างนี้



1.2.4 แฟ้มข้อมูลที่ได้ 1 record แทนหลายคน (ตัวแปร n คือจำนวนคน) ถ้าต้องการให้เป็นคนละ 1 record เราต้องสั่ง

.expand n <ENTER> แล้วก็

.drop n <ENTER> เพื่อลบตัวแปร n ทิ้งไปเนื่องจากไม่จำเป็นต้องใช้แล้ว

1.2.5 ดูข้อมูล

.list <ENTER>

คำถาม: ท่านมีข้อมูลทั้งสิ้น ..... records

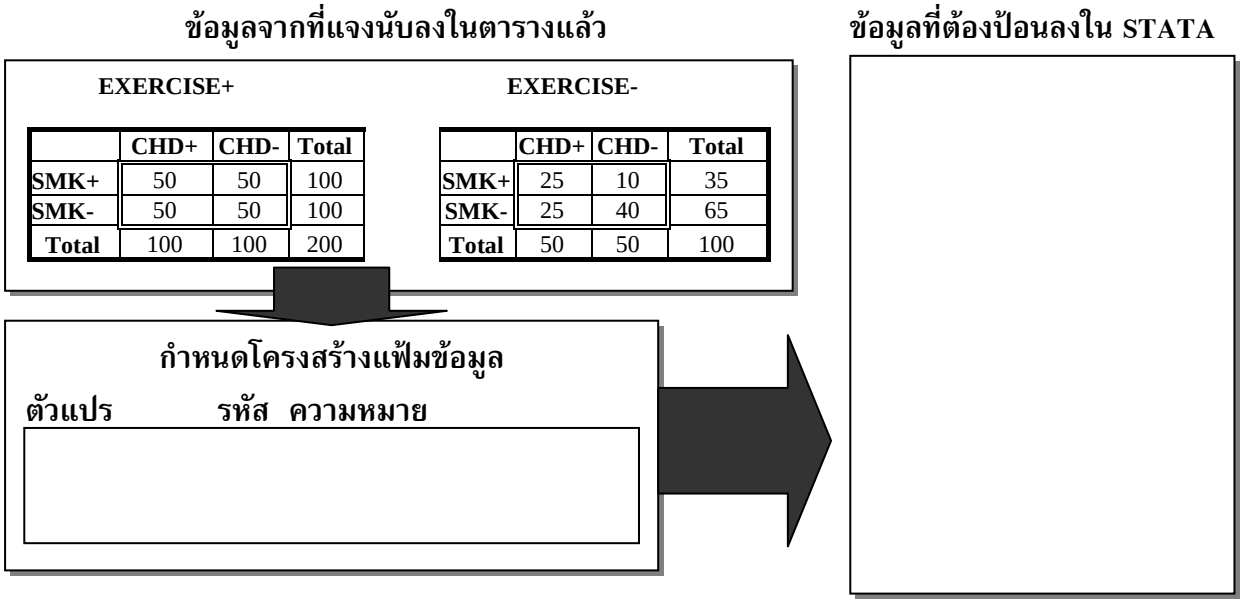
1.2.6 ถ้าต้องการ Save แฟ้มข้อมูลไว้ใช้คราวต่อไป พิมพ์ save ตามด้วยชื่อแฟ้ม แล้วกด<ENTER> เช่น

.save example1 <ENTER> เราจะได้แฟ้ม EXAMPLE1.DTA

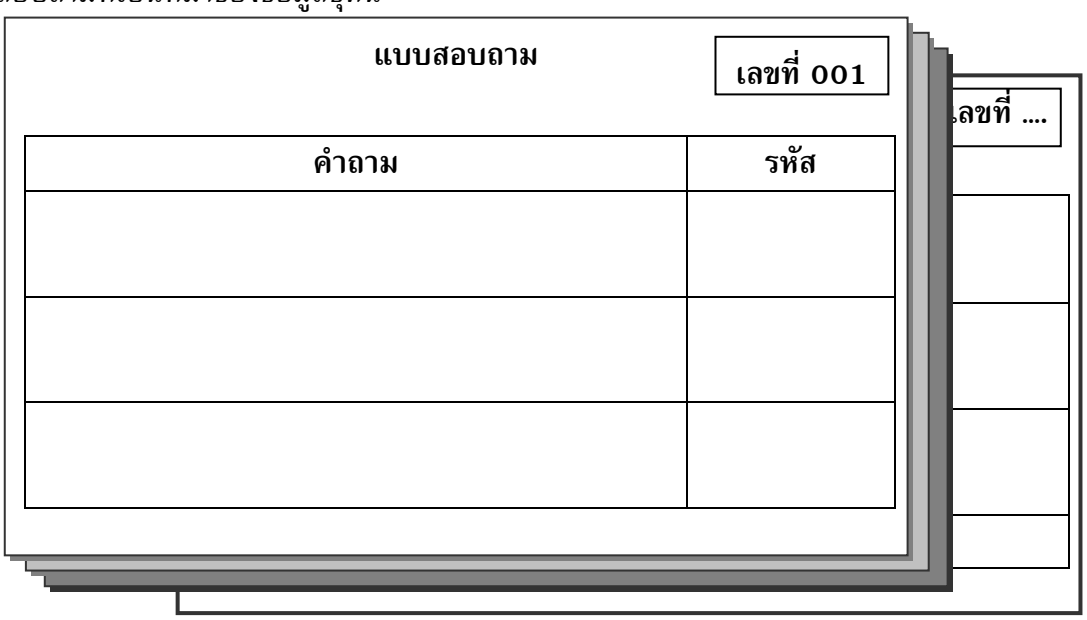
1.3 จงวิเคราะห์ข้อมูลซ้ำตามที่แสดงไว้ตามรูป 1.24 ถึง 1.27 พร้อมเปรียบเทียบผลที่ท่านทำได้ได้กับรูปเหล่านั้น

2. จงวิเคราะห์ข้อมูลจากรูปที่ 1.17 ตามลำดับดังนี้

2.1 จงกรอกคำตอบลงช่องกำหนดโครงสร้าง ข้อมูลที่ต้องป้อนลงใน STATA และร่างแบบสอบถาม



ลักษณะแบบสอบถามที่เป็นที่มาของข้อมูลชุดนี้



2.2 จงป้อนข้อมูลลงใน STATA แล้วบันทึกเก็บไว้ในแฟ้ม EXAMPLE2.DTA

2.3 คำสั่งเพื่อให้ได้ผลตามรูปที่ 1.18 คือ

คำตอบ: .....

2.5 คำสั่งสำหรับวิเคราะห์โดยใช้ Logistic regression ให้ได้ผลคล้ายกับข้อ 2.3 คือ

คำตอบ: .....

3. วิเคราะห์ข้อมูล ANC

ข้อมูลตามตารางข้างล่างนี้ ได้จากการศึกษาบทบาทของอาสาสมัครสาธารณสุข (อสม) กับผลการคลอดของสตรี ซึ่งเป็นการศึกษาแบบ Cohort Study ที่เปรียบเทียบสองพื้นที่ คือ พื้นที่ทดลองกับพื้นที่ที่มีการบริการตามปกติ (พื้นที่ควบคุม) พื้นที่ทดลอง กำหนดให้มีการเยี่ยมก่อนและหลังคลอดโดย อสม ส่วนพื้นที่ควบคุม ไม่มีกิจกรรมดังกล่าว เพียงแต่รอรับการฝากครรภ์ตามปกติที่สถานีอนามัย วัตถุประสงค์ของการศึกษา เพื่อทราบว่า ***การเยี่ยมก่อนและหลังคลอด มีผลต่อการรอดชีพของทารกในระยะเวลา 28 วันหลังคลอด หรือไม่*** ตลอดระยะเวลา 1 ปีนั้น มีการคลอด 939 ครั้ง ในพื้นที่ทดลองและ 944 ครั้งในพื้นที่ควบคุม แต่เพื่อใช้ทำแบบฝึกหัดนี้ ใช้ข้อมูล 65 ราย จากผู้มีการคลอดมีชีพ แล้วทารกตายภายใน 28 วัน (Neonatal death) และ 400 ราย ที่มีการคลอดเป็นทารกรอดชีพ (แฟ้มข้อมูลชื่อ LOGISTIC.DTA ในแผ่นดิสเกตต์ที่แจก)

ข้อมูล

ID	DEAD	AREA	MALPRES	BWT	MAGE	DCHILD
1	1	1	0	2600	30	0
2	1	1	0	2900	29	1
3	1	1	0	3100	25	0
4	1	1	0	3000	21	0
----------ข้าม 460 เร็คคอร์ด----------						
465	0	1	0	3200	22	1

โครงสร้างแฟ้มข้อมูล LOGISTIC.DTA

ชื่อตัวแปร	**คำอธิบาย**	**รหัส**
ID	เลขที่แบบสอบถาม	1 ถึง 465 ตามลำดับ
DEAD	การรอดชีพของทารก เมื่อ 28 วันหลังคลอด	0= รอด 1=ตาย
AREA	พื้นที่ ที่มารดาอาศัย	0=พื้นที่ควบคุม 1=พื้นที่ทดลอง
MALPRES	ท่าของทารกขณะคลอด	0=ปกติ 1= ผิดปกติ
BWT	น้ำหนัก ทารกแรกเกิด (กรัม)	ตามที่ระบุ
MAGE	อายุมารดา(ปี)	ตามที่ระบุ
DCHILD	จำนวนเด็กเกิดมีชีพของมารดา ซึ่งขณะนี้เด็กนั้นเสียชีวิตแล้ว (คน)	ตามที่ระบุ

ข้อ 3.1 ถึง 3.7 เป็นการทำความเข้าใจกับข้อมูล

3.1 ตัวแปรใดบ้างที่เป็น Dichotomous

ตอบ ....

3.2 ตัวแปรใดบ้างที่เป็น Continuous

ตอบ .....

3.3 ตัวแปรตาม (D) คือ .....

3.4 ตัวแปรต้นที่สนใจ (E) คือ .....

3.5 ตัวแปรต้นที่ต้องการควบคุมผลกระทบต่อความสัมพันธ์ที่มุ่งอธิบาย (C) คือ .....

3.6 ถ้าแบบสอบถามที่ได้ถูกนำไปใช้ในการศึกษาครั้งนี้ เป็นดังที่ให้ไว้ข้างล่างนี้ ข้อมูลสำหรับทารกรายแรกตามที่ให้ไว้ข้างต้น จะปรากฏในแบบสอบถามนี้ อย่างไร (กรอกข้อมูลลงในแบบสอบถาม พร้อมลงรหัส)

แบบสอบถาม เลขที่.....	
คำถาม	**สำหรับลงรหัส**
1. สถานภาพการรอดชีพของทารกเมื่ออายุครบ 28 วัน [ ]1. ตาย [ ]0. รอด	DEAD[ ]
2. พื้นที่ที่มารดาอาศัย [ ]1. พื้นที่ทดลอง [ ]0. พื้นที่ควบคุม	AREA[ ]
3. ท่าของทารกขณะคลอด [ ]1. ผิดปกติ [ ]0. ปกติ	MALPRES[ ]
4. นำหนักทารกแรกเกิด.....กรัม	BWT[ ][ ][ ][ ]
5. อายุมารดา.....ปี	AGE[ ][ ]
6. จำนวนเด็กเกิดมีชีพของมารดา ซึ่งขณะนี้ เด็กนั้นเสียชีวิตแล้ว.....คน	DCHILD[ ]

3.7 แบบสอบถามสำหรับข้อมูลที่นำมาวิเคราะห์ในครั้งนี้ มีทั้งหมดกี่แผ่น ตอบ ..... ชุด

ข้อ 3.8 และ 3.9 เป็นการวิเคราะห์ข้อมูล *(โปรดทราบว่า การวิเคราะห์ต่อไปนี้ ใช้ Odds Ratio เป็นมาตรวัดความสัมพันธ์โดยตลอด ไม่ใช้ Relative risk ทั้งที่เป็นการศึกษาแบบ Cohort study ทั้งนี้เนื่องจากสถิติที่เหมาะสมในการตอบคำถามวิจัยนี้คือ Logistic regression)*

3.8 Crude (Bivariate) analysis

3.8.1 ภาพรวมการวิเคราะห์

		DEAD	
		ตาย	รอด
AREA	พื้นที่ทดลอง	37	204
	พื้นที่ควบคุม	28	196

OR_C = ...

3.8.2 ผลการวิเคราะห์จาก STATA

```
. cc  dead area
                 | area                  |            Proportion
                 |   Exposed   Unexposed |     Total     Exposed
-----------------+-----------------------+----------------------
           Cases |        37          28 |        65      0.5692
        Controls |       204         196 |       400      0.5100
-----------------+-----------------------+----------------------
           Total |       241         224 |       465      0.5183
                 |                       |
                 |      Point estimate   |  [95% Conf. Interval]
                 |-----------------------+----------------------
      Odds ratio |         1.269608      |  .7512221    2.145309  (Cornfield)
 Attr. frac. ex. |         .2123552      | -.3311642    .5338668  (Cornfield)
 Attr. frac. pop |         .1208791      |
                 +-----------------------------------------------
                           chi2(1) =     0.79  Pr>chi2 = 0.3754
```

3.9 Stratified analysis

3.9.1 Stratified analysis โดยให้ MALPRES เป็น Stratified variable

3.9.1.1 ภาพรวมการวิเคราะห์

Stratum 1 เมื่อ MALPRES = 0			
		DEAD	
		ตาย	รอด
ARE	พื้นที่ทดลอง	18	196
	พื้นที่ควบคุม	26	190

OR_1 = ...

Stratum 2 เมื่อ MALPRES = 1			
		DEAD	
		ตาย	รอด
ARE	พื้นที่ทดลอง	19	8
	พื้นที่ควบคุม	2	6

OR_2 = ...

ทดสอบความแตกต่างโดย Woolf's test for heterogeneity of Odds Ratios

ได้ p-value = 0.015 (ดูผลจาก Computer output ในข้อ 3.9.1.2)

3.9.1.2 ผลการวิเคราะห์จาก STATA

```
. cc  dead area, by(malpres)

         malpres |       OR      [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               0 |   .6711146     .3590862   1.254787      11.85116 (Cornfield)
               1 |      7.125     1.297704   37.58284      .4571429 (Cornfield)
-----------------+-------------------------------------------------
           Crude |   1.269608     .7512221   2.145309               (Cornfield)
    M-H combined |   .9108184     .5136778   1.615001
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     5.91  Pr>chi2 = 0.0151

                  Test that combined OR = 1:
                              Mantel-Haenszel chi2(1) =      0.11
                                                Pr>chi2 =    0.7453
```

3.9.1.3 เขียนผลที่ได้จากการวิเคราะห์ ได้ดังนี้

OR_C = ...

OR_1 = ...

OR_2 = ...

Woolf's test p-value = ...

OR_{MH} = ...

3.9.1.4 เขียนสรุปผลการวิเคราะห์

3.9.2 Stratified analysis โดยให้ DCHILD เป็น Stratified variable ทั้งนี้ จำเป็นต้องแปลงตัวแปร DCHILD ที่เดิมเป็นตัวแปรต่อเนื่อง ให้เป็นตัวแปรแจงนับ โดยให้เป็น 2 กลุ่มคือ มี กับไม่มีเด็กเกิดมีชีพแล้วตายในเวลาต่อมา จากนั้น ทำคล้ายข้อ 3.9.1 ได้ผลดังนี้

3.9.2.1 ภาพรวมการวิเคราะห์

	เมื่อ DCHILD = 0		**เมื่อ DCHILD = 1**		**ตารางรวม (หยาบ)**	
	DEAD		DEAD		DEAD	
	รอด	ตาย	รอด	ตาย	รอด	ตาย
พื้นที่ควบคุม	167	21	29	7	196	28
พื้นที่ทดลอง	170	17	34	20	204	37

OR_1 = ... **OR_2 = ...** **OR_C = 1.27**

3.9.2.2 ผลการวิเคราะห์จาก STATA

```
. replace dchild = 1 if dchild > 1 & dchild ~= .
(22 real changes made)

. cc  dead area, by( dchild)

          dchild |       OR      [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               0 |   .7952381     .4090156   1.546609          9.52 (Cornfield)
               1 |   2.436975     .9184562   6.424761      2.644444 (Cornfield)
-----------------+-------------------------------------------------
           Crude |   1.269608     .7512221   2.145309               (Cornfield)
    M-H combined |   1.152137     .6695645   1.982513
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     3.35  Pr>chi2 = 0.0673

                   Test that combined OR = 1:
                               Mantel-Haenszel chi2(1) =      0.26
                                                Pr>chi2 =    0.6098
```

3.9.2.3 เขียนผลที่ได้จากการวิเคราะห์ ได้ดังนี้

OR_C = ...

OR_1 = ...

OR_2 = ...

Woolf's test p-value = ...

OR_{MH} = ...

3.9.2.4 เขียนสรุปผลการวิเคราะห์

3.9.3 Stratified analysis โดยให้ BWT เป็น Stratified variable

เนื่องจาก BWT เป็นข้อมูลต่อเนื่อง ก่อนทำคล้ายข้อ 3.9.1 ต้องแปลงข้อมูลตัวแปร BWT ให้อยู่ในรูปข้อมูลแจงนับก่อน ในที่นี้จัดกลุ่มน้ำหนักเป็นดังนี้

1 = น้อยกว่า 2500

2 = 2500-2999

3 = 3000 หรือมากกว่า

3.9.3.2 ผลการวิเคราะห์จาก STATA

```
. gen bwtg = .
(465 missing values generated)
. replace bwtg = 1 if bwt < 2500
(39 real changes made)
. replace bwtg = 2 if bwt >= 2500 & bwt < 3000
(140 real changes made)
. replace bwtg = 3 if bwt >= 3000
(286 real changes made)
. cc  dead area, by(bwtg)

            bwtg |       OR      [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               1 |        .49     .1209808    1.95487      2.564103 (Cornfield)
               2 |   2.097917     .8631402   5.078711      3.428571 (Cornfield)
               3 |   .8043011     .3620455   1.788335      6.503497 (Cornfield)
-----------------+-------------------------------------------------
           Crude |   1.269608     .7512221   2.145309               (Cornfield)
    M-H combined |   1.094738     .6363404   1.883351
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(2) =     3.74  Pr>chi2 = 0.1541

                  Test that combined OR = 1:
                              Mantel-Haenszel chi2(1) =      0.11
                                              Pr>chi2 =    0.7438
```

3.9.3.3) เขียนผลที่ได้จากการวิเคราะห์

OR_C = ...

OR_1 = ...

OR_2 = ...

OR_3 = ...

Woolf's test p-value = ...

OR_{MH} = ...

3.9.3.4 เขียนสรุปผลการวิเคราะห์

3.10 Multivariable analysis ควรวิเคราะห์อย่างไร

แนวคิดพื้นฐานของการวิเคราะห์โดยใช้ Logistic Regression และการใช้ประโยชน์

วัตถุประสงค์: เพื่อให้ผู้อ่านสามารถ

1. บอกความหมายของปัญหาที่ต้องใช้การวิเคราะห์แบบ Multivariable analysis ได้

2. บอกความแตกต่างระหว่าง Simple Linear Regression กับ Logistic Regression ได้

3. อธิบายหลักการของ Logistic Model ได้

4. คำนวณค่า Odds Ratio (OR) จาก Logistic Model อย่างง่ายได้

เนื้อหา :

1. แนวคิดพื้นฐานของ Logistic Regression Model

2. การใช้ประโยชน์สมการ Logistic Regression Model

กิจกรรม :

1. ฟังบรรยายประกอบแผ่นใส พร้อมบันทึกเนื้อหาสำคัญลงในชุดการเรียนการสอน บทที่1

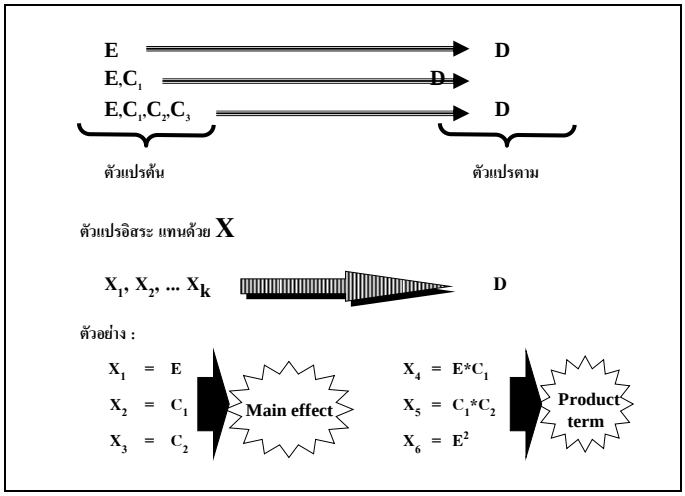
2. ทำแบบฝึกหัด

3. อภิปรายและสรุปเนื้อหา พร้อมเขียนสรุปท้ายบทลงกรอบว่างที่ให้ไว้ท้ายบท

สิ่งที่นำเสนอ	คำอธิบาย

1. แนวคิดพื้นฐานเกี่ยวกับ Logistic Model

รูปที่ 2.1

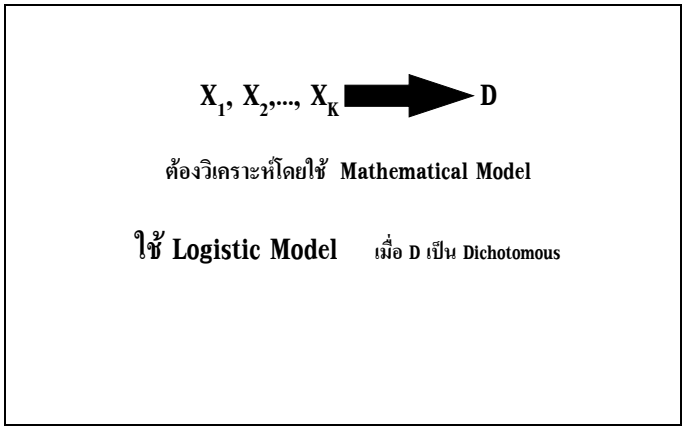


ในบทที่ 1 ได้แสดงแนวทางในการวิเคราะห์ข้อมูลที่เป็นปัญหาง่าย ๆ ได้แก่ ความสัมพันธ์ระหว่างตัวแปรสองตัว คือ ปัจจัย (E) กับ โรค (D) กรณีมีปัจจัยควบคุม (C) เข้ามาเพิ่มอีกหนึ่งตัวแปร แต่ความเป็นจริงยังมีปัจจัยอย่างอื่นอีกที่ต้องควบคุมผลไปพร้อม ๆ กัน คือมีตัวแปรต้นมากกว่าหนึ่งตัวแปร นอกจากนี้ ตัวแปรต้นยังรวมถึงผลิตผล (Product term) ของตัวแปรที่มีอยู่แล้ว อีกด้วย

เพื่อให้ง่ายต่อการทำความเข้าใจ เราให้ X (ตัวพิมพ์ใหญ่) แทนตัวแปรต้นทั้งหมด ดังนั้น X_1 แทนตัวแปรตัวที่ 1 X_2 แทนตัวแปรตัวที่ 2 จนกระทั่งถึง X_k แทนตัวแปรตัวที่ K ตัวอย่างเช่น เรามีตัวแปรต้น 6 ตัว โดยให้ X_1 แทนปัจจัย E ให้ X_2 แทน C_1 ให้ X_3 แทน C_2 ให้ X_4 แทนผลคูณระหว่าง E กับ C_1 ให้ X_5 แทนผลคูณระหว่าง C_1 กับ C_2 และให้ X_6 แทน E ยกกำลังสองเป็นต้น

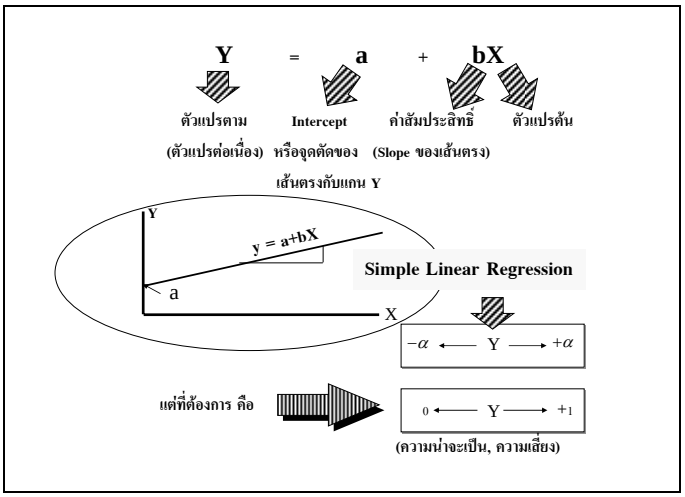
X_4และ X_5 และ X_6 เรียกว่า Product term เฉพาะ X_4 กับ X_5 คือ Interaction term เนื่องจากเราสนใจว่าจะมี Interaction effect จากตัวแปร C_1 กับ C_2 หรือไม่ นอกจากนั้น Product term ยังมีอีกหลายรูปแบบ ซึ่งจะได้กล่าวต่อไป

รูปที่ 2.2



เมื่อคำถามวิจัยเกี่ยวข้องกับปัญหาที่มีหลายสาเหตุ (Multi-factorial Outcome) จำเป็นต้องใช้แบบจำลองทางคณิตศาสตร์ (Mathematical model) ในการวิเคราะห์ข้อมูล ซึ่งมีอยู่หลายวิธี แต่ถ้าตัวแปรตาม (D) เป็นตัวแปร Dichotomous แบบจำลองทางคณิตศาสตร์ที่เหมาะสม คือ Logistic Model

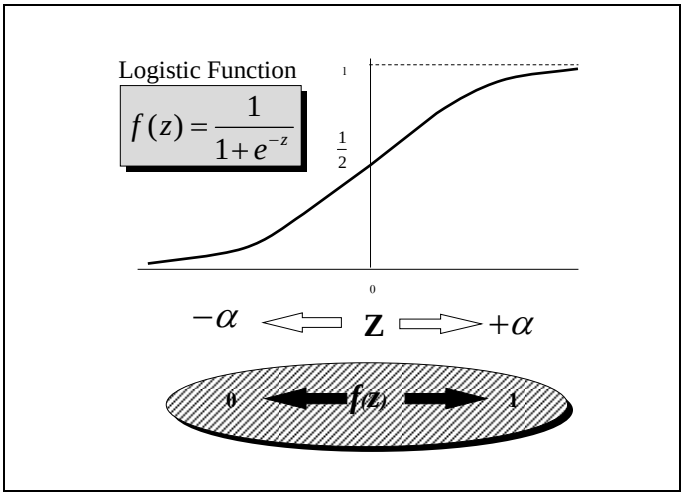
รูปที่ 2.3



แบบจำลองทางคณิตศาสตร์ที่เป็นที่คุ้นเคยกันได้แก่สมการการถอดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression) ซึ่งก็คือสมการเส้นตรง $Y = a + bX$ ที่มี a เป็นค่าของ Y เมื่อ X เป็น 0 หรือจุดที่เส้นตรงตัดแกน Y ที่เรียกว่า Intercept นั่นเอง มีค่าความลาดเอียงหรือ Slope เท่ากับ b ใช้แสดงความสัมพันธ์ระหว่างตัวแปรตาม Y กับตัวแปรต้น X โดยที่ Y เป็นตัวแปรต่อเนื่อง มีค่าได้ทุกค่า ตั้งแต่ $-\infty$ จนถึง $+\infty$

แต่ถ้า Y เป็นข้อมูลแจงนับ เฉพาะอย่างยิ่งเป็น Dichotomous คือมีค่าได้ 2 ค่า เช่น ไม่ 1 ก็ 0 เป็นต้น จะใช้วิธีการวิเคราะห์ดังกล่าวไม่ได้ นอกจากนี้ กรณี Dichotomous เราสนใจค่าความน่าจะเป็นมากกว่าค่าจริงของตัวแปร เช่น “โรค” เราสนใจโอกาสที่จะป่วย หรืออธิบายเป็นความเสี่ยงซึ่งเป็นความน่าจะเป็น ที่มีค่าอยู่ระหว่าง 0 ถึง 1 เท่านั้น

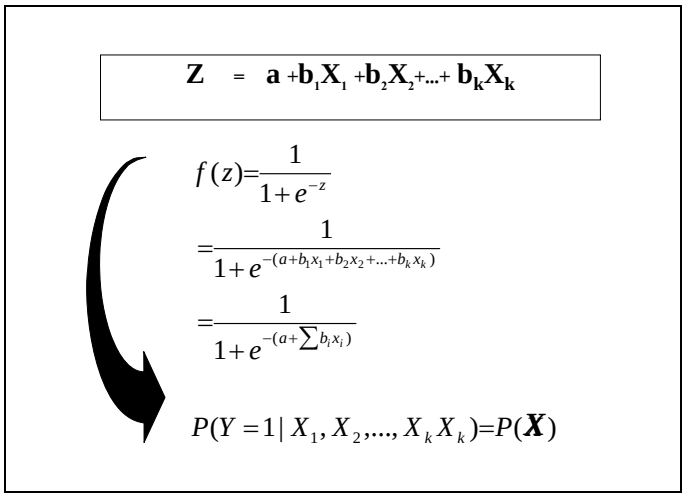
รูปที่ 2.4



ปัญหานี้ แก้ได้โดยอาศัยคุณสมบัติของฟังก์ชั่นทางคณิตศาสตร์ที่เรียกว่า Logistic function เขียนแทนด้วย $f(z) = 1/(1+e^{-z})$ ถ้าหาก plot กราฟค่าของฟังก์ชั่นนี้ เมื่อแทนค่า Z ด้วยทุก ๆ ค่าที่เป็นไปได้ จะได้กราฟ เป็นรูปตัว S มีค่าอยู่ระหว่าง 0 กับ 1

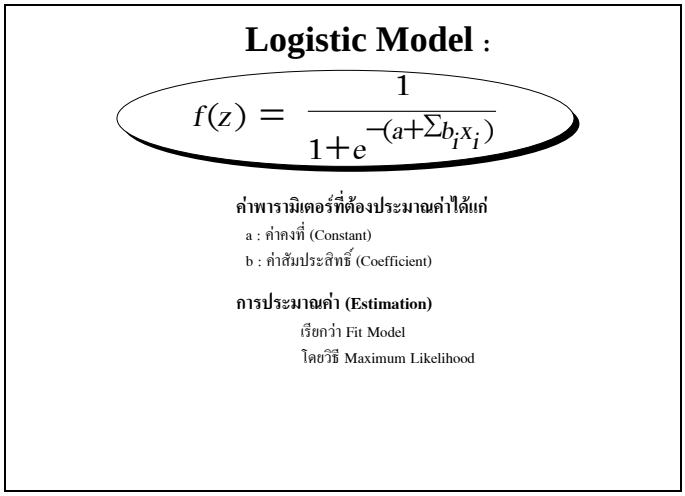
<u>หมายเหตุ</u> $e^{-z} = 2.7183^{-z}$
$= \text{Exponential}(-z)$
$= \text{EXP}(-z)$

รูปที่ 2.5



Logistic Model จะสร้างโดยอาศัย Logistic function โดยให้ Z เป็นสมการเชิงเส้นของทุกตัวแปรรวมกัน แล้วแทนค่า Z ใน Logistic function คือ $f(z)=\frac{1}{1+e^{-z}}$ ได้ Logistic Model คือ $f(z)=\frac{1}{1+e^{-(a+\sum b_ix_i)}}$ ซึ่งจะได้ว่า f(z) ก็คือความน่าจะเป็นที่จะเกิดโรค เมื่อมีตัวแปรต้นตามที่ระบุ จึงเขียนแทนด้วย P (D=1 X_1, X_2,..., X_k) แต่เพื่อให้ง่ายจึงแทนด้วย P(**X**) โปรดสังเกตตัว **X** เป็นตัวพิมพ์ใหญ่ทึบ

รูปที่ 2.6

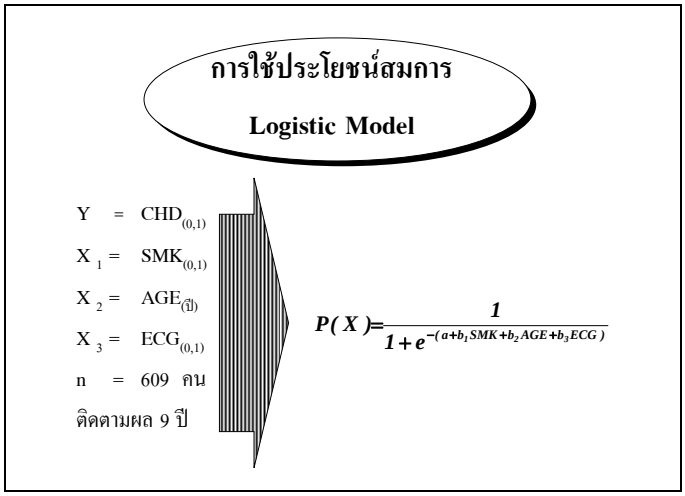


ดังนั้น รูปแบบของ Logistic Model จึงเขียนเป็น $P(\mathbf{X})= 1/[1+e^{-(a+\sum bixi)}]$ มีค่าพารามิเตอร์ ที่ต้องประมาณค่าคือค่า a ซึ่งเป็นค่าคงที่ (Constant) และค่า b_i ซึ่งเป็นค่าสัมประสิทธิ์ (Coefficient)

การใช้ข้อมูลที่มีอยู่เพื่อประมาณค่าพารามิเตอร์ เรียกว่า การ “Fit Model” วิธีการทางคณิตศาสตร์ ที่ใช้ในการประมาณค่าพารามิเตอร์เรียกว่า Maximum Likelihood

2. การใช้ประโยชน์ Logistic Model

รูปที่ 2.7



ตัวอย่าง (จาก Kleinbaum, 1994) การศึกษาแบบ Cohort Study จากตัวอย่าง 609 คน ตัวแปรตามคือ การป่วยด้วยโรคหัวใจโคโรนารี่ (CHD) เป็นตัวแปร Dichotomous มีค่าเป็น 0 ถ้าไม่ป่วย และมีค่าเป็น 1 ถ้าป่วย ตัวแปรต้นมี 3 ตัวแปร คือการสูบบุหรี่ (SMK) มีค่าเป็น 0 ถ้าไม่สูบและเป็น 1 ถ้าสูบ อายุ (AGE) เป็นตัวแปรต่อเนื่อง มีหน่วยเป็นปี และค่า Electrocardiogram (ECG) มีค่าเป็น 0 ถ้าไม่ผิดปกติและเป็น 1 ถ้าผิดปกติ เขียนความสัมพันธ์ในรูป Logistic Model เป็น:

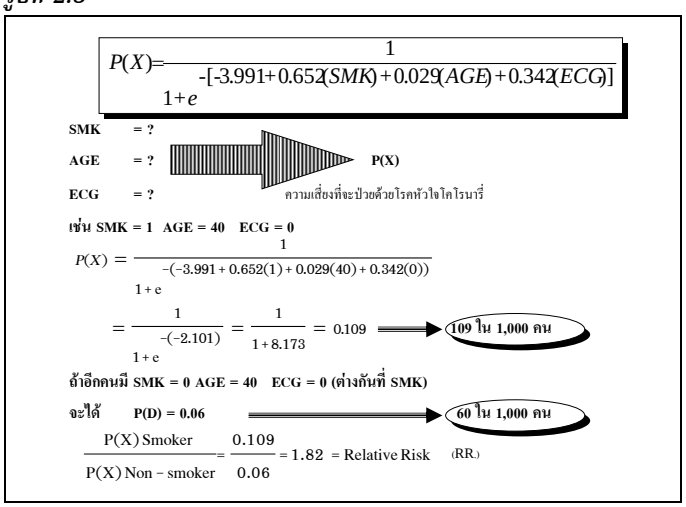
$$P(X)=\frac{1}{1+e^{-(a+b_1SMK+b_2AGE+b_3ECG)}}$$

การประมาณค่าพารามิเตอร์ a และ b โดยวิธี Maximum Likelihood มีขั้นตอนที่สลับซับซ้อนมากจึงต้องอาศัยคอมพิวเตอร์ที่มีโปรแกรมที่ใช้วิเคราะห์ Logistic Regression ได้เช่น SAS, BMDP, SPSS, STATA, GLIM, EGRET ฯลฯ ในที่นี้ใช้ STATA (ดูรายละเอียดใน StataCorp., 1999)

ต่อไปนี้คือตัวอย่างผลการวิเคราะห์ข้อมูล จากการวิเคราะห์ ได้ค่า a = -3.911 b_1 = 0.652 b_2 = 0.029 และ b_3 = 0.342

จากนั้น แทนค่าพารามิเตอร์ลงใน Logistic Model ผลที่ได้ คือสมการทางคณิตศาสตร์ แสดงความสัมพันธ์ระหว่างตัวแปรต้นทั้งสาม และตัวแปรตาม CHD ถ้าเราแทนค่าของ SMK AGE และ ECG ของบุคคลใดลงไปใน Model ก็จะได้โอกาสหรือความเสี่ยงที่จะป่วย ด้วยโรคหัวใจโคโรนารี่ เรียกว่า Predicted Risk เช่นถ้าต้องการทราบว่าใครก็ตามที่สูบบุหรี่ที่มีอายุ 40 ปี และมี ECG ปกติ จะมีโอกาสป่วยเท่าไร ก็แทนค่า SMK=1 AGE=40 และ ECG=0 ลงไปใน Model

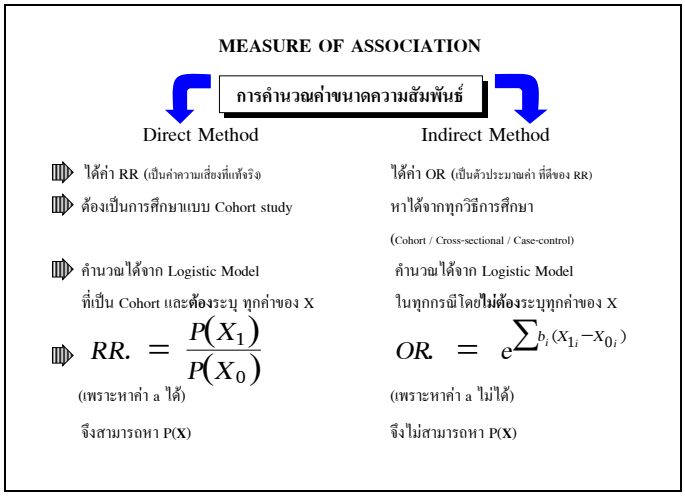
รูปที่ 2.8



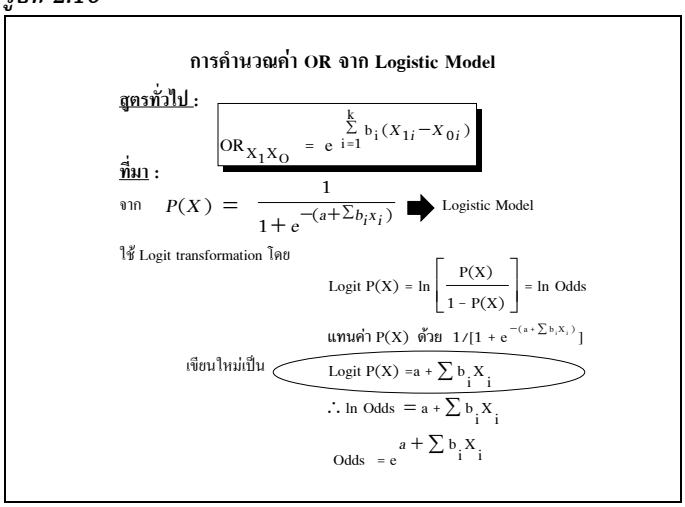
คำนวณได้ผลลัพธ์ 0.109 นั่นคือโอกาสที่บุคคลดังกล่าวจะป่วยเท่ากับ 0.109 หรือกล่าวอีกนัยหนึ่งได้ว่าผู้ที่มีอายุ 40 ปี ที่สูบบุหรี่และมี ECG ปกติ ในจำนวนทั้งหมด 1000 คน จะมีผู้ที่ป่วย 109 คน ใน ระยะเวลา 9 ปี (follow-up time) ในทำนองเดียวกัน ถ้าเปรียบเทียบกับผู้ไม่สูบบุหรี่ (SMK=0) แต่ให้ AGE และ ECG เหมือนกัน จะมีโอกาสป่วย 0.06 หรือจะพบผู้ป่วย 60 คนใน 1000 คน นั่นเอง เมื่อนำค่าที่ได้มาเปรียบเทียบกัน เป็นการหาค่าอัตราส่วน (Ratio) โดยนำ Risk ของผู้สูบบุหรี่ตั้งแล้วหารด้วย Risk ของผู้ไม่สูบบุหรี่

ค่าที่ได้เท่ากับ 0.109/ 0.060=1.82 ซึ่งก็คือ Risk Ratio หรือ Relative Risk (RR.) ใน

รูปที่ 2.9



รูปที่ 2.10



ตัวอย่างนี้ เป็น RR. ที่ควบคุมผลของอายุและ ECG แล้ว (Adjusted RR.)

RR. เป็นมาตรวัดระดับความสัมพันธ์โดยตรง (Direct measure) แม้สามารถคำนวณได้ จาก Logistic Regression Model แต่ต้องมาจากการศึกษาแบบ Cohort study และต้องระบุทุกค่าของ **X** จึงทำให้มีข้อจำกัด ตรงข้ามถ้าได้จากการศึกษาแบบอื่น ต้องคำนวณค่า OR ค่านี้คำนวณได้จากทั้งการศึกษาแบบ Cohort study แบบ Case-control และ แบบ Cross-sectional study การคำนวณไม่จำเป็นต้องระบุทุกค่าของ **X** แต่ระบุเฉพาะค่าตัวแปรที่ต้องการเปรียบเทียบ เมื่อได้ค่า OR แล้วจึงประมาณว่าเป็นค่า RR. ภายใต้ข้อกำหนดบางอย่างคือ “กรณีความชุกของโรคต่ำ ๆ (หายาก) และตัวอย่างที่ศึกษาเป็นตัวแทนของประชากร ค่า OR จะเท่ากับ RR.” เป็นต้น และแปลความหมายเหมือนกับค่า RR.

อนึ่งด้วยข้อมูลจากการศึกษาแบบ Case-control และ Cross-sectional ไม่สามารถนำข้อมูลมาคำนวณหาค่าคงที่ (a) ได้ ***แม้คอมพิวเตอร์จะให้ค่าออกมาก็ไม่ต้องนำเสนอ เพราะไม่ถูกต้อง*** เป็นเหตุผลที่คำนวณค่า RR. ไม่ได้กรณีการศึกษาแบบดังกล่าว

การคำนวณค่า OR จาก Logistic Regression Model คือการเปรียบเทียบ Odds ค่าหนึ่งกับค่า Odds อีกค่าหนึ่ง ถ้าอยู่ในรูปตัวเลขธรรมดา คือนำมาหารกัน เป็นอัตราส่วน (Ratio) แต่เมื่ออยู่ใน Logistic function ต้องเปลี่ยนรูปเป็น Logit เสียก่อน จะได้ Model เป็นผลรวมเชิงเส้น (Linear sum) ของค่าสัมประสิทธิ์ดังนี้ Logit P(**X**) = a+$\sum$biXi และ Logit P(**X**) คือ Log ฐาน e หรือ

Natural Log ของ Odds (ln อ่านว่า “ลอน”)

ในทางคณิตศาสตร์ ส่วนกลับของ log คือ Exponential แทนด้วย e ดังนั้นค่า $e^{a+\sum biXi}$ หรือ Exponential ของ $a+\sum biXi$ จึงเป็นค่า Odds

รูปที่ 2.11 (ต่อรูปที่ 10)

$$OR_{X_1X_0} = \frac{\text{odds ของ } X_1}{\text{odds ของ } X_0}$$

$$= \frac{e^{(a+\sum b_i X_{1i})}}{e^{(a+\sum b_i X_{0i})}}$$

$$= e^{[(a+\sum b_i X_{1i})]-[(a+\sum b_i X_{0i})]}$$

$$= e^{(a-a+\sum b_i (X_{1i}-X_{0i}))}$$

$$\therefore\ OR_{X_1X_0} = e^{\sum b_i (X_{1i}-X_{0i})}$$ ⇨ สูตรทั่วไป

เมื่อต้องการ Odds Ratio ก็ต้องเป็นผลหารของสองOdds แต่ผลหารของ Exponential คือ Exponential ของผลต่าง กล่าวคือ $(e^a/e^b) = e^{a-b}$ สุดท้ายได้สูตรการคำนวณ OR จาก Logistic Regression Model เป็น OR ของ X_1 เปรียบเทียบกับ X_0 เท่ากับ Exponential ของ $\sum b_i(x_{1i}-X_{0i})$ ดังนั้น สูตรทั่วไปในการคำนวณค่า OR คือ

$$OR_{X_1X_0} = e^{\sum b_i (X_{1i}-X_{0i})}$$

รูปที่ 2.12

ตัวอย่างการคำนวณค่า OR

เมื่อ X = (SMK, AGE, ECG) และให้

X_1 = (SMK =1, AGE =40, ECG =0)

X_0 = (SMK =0, AGE =40, ECG =0)

ในทางปฏิบัติ ไม่ระบุค่านี้ แต่ fixed

จาก Logit P(X) $= a + b_1SMK + b_2AGE + b_3ECG$

จาก $OR_{X_1X_O} = e^{\sum_{i=1}^{k} b_i (X_{1i}-X_{0i})}$

$= e^{b1+0+0}$

$= e^{b1}$

เมื่อ $b_1 = 0.652$ ⇨ $e^{0.652} = 1.92$ ⇨ OR = 1.92

ตัวอย่างการคำนวณ OR จาก Logistic Regression Model โดยใช้สูตรทั่วไป เมื่อ **X** เป็นลักษณะของคนคนหนึ่ง เช่น X_1 กำหนดให้เป็นคนที่มีอายุ 40 ปี มี ECG ปกติ ที่สูบบุหรี่ ส่วน X_0 กำหนดให้เป็นคนที่มีอายุเท่ากัน และ ECG ปกติเหมือนกัน แต่ไม่สูบบุหรี่ เมื่อแทนค่าลงในสูตร หา OR ค่า X_1 จะถูกหักลบด้วย X_0 ดังนั้นค่าของ AGE กับ ECG จะหักลบกันเป็นศูนย์ ยังคงเหลือเฉพาะค่าของ SMK ซึ่งได้ผลเท่ากับ 1-0 เท่ากับ 1 เมื่อคูณกับ b_1 ก็ได้เท่ากับ b_1 ค่า Exponential ของ b_1 ก็คือ OR ของการสูบบุหรี่ (SMK) กับการป่วยด้วยโรคหัวใจโคโรนารี่ (CHD) อย่างไรก็ตาม ในทางปฏิบัติ ค่าของ AGE กับ ECG จะไม่ถูกระบุ แต่จะให้เหมือนกันทั้งใน X_1 และ X_0 ซึ่ง

เป็นการควบคุมผลจากตัวแปรเหล่านั้น

สรุปท้ายบท :

เอกสารอ้างอิงประจำบทที่ 2

Kleinbaum, D.G. (1994). *Logistic Regression: A self-learning text*. New York Springer-Verlag.

StataCorp. (1999). *Stata statistical software: Release 6.0*. College Station. TX: Stata Corporation.

แบบฝึกหัดที่ 2

1. ในชุดการเรียนที่ 2 มีสมการที่สำคัญอยู่ 3 สมการ ให้เขียนลงในช่องที่กำหนดให้ถูกต้อง

1.1) Logistic Function :

1.2) Logistic Model :

1.3) Logit transformation :

2. การคำนวณค่า Odds Ratio (OR) จาก Bivariate model

2.1) ค่า OR หาได้จาก Logistic Model ที่ได้จากข้อมูลการศึกษาประเภทใด
(กาเครื่องหมายหน้าข้อที่ถูก)

- ☐ Cohort study
- ☐ Case-control study
- ☐ Cross-sectional Study

Output จาก STATA ได้ดังนี้

```
. logit  dead area

Iteration 0:   log likelihood =  -188.1264
Iteration 1:   log likelihood = -187.73286
Iteration 2:   log likelihood = -187.73214

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(1)      =       0.79
                                                  Prob > chi2     =     0.3745
Log likelihood = -187.73214                       Pseudo R2       =     0.0021

------------------------------------------------------------------------------
        dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |   .2387081   .2697124      0.885   0.376      -.2899185    .7673346
       _cons |   -1.94591   .2020295     -9.632   0.000      -2.341881    -1.54994
------------------------------------------------------------------------------
```

2.2) ค่า OR หาได้โดย

(i) สูตรที่ใช้ $OR_{x1,Xo}$ =

(ii) ทารกในพื้นที่ทดลอง เป็น X_1 =.....

ทารกในพื้นที่ควบคุม เป็น X_0 =.....

(iii) แทนค่า $OR_{X1,X0}$ =

ใช้ STATA คำนวณได้ดังนี้

```
. logistic  dead area

. logistic  dead area

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(1)      =       0.79
                                                  Prob > chi2     =     0.3745
Log likelihood = -187.73214                       Pseudo R2       =     0.0021

------------------------------------------------------------------------------
    dead | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    area |   1.269608    .342429      0.885   0.376       .7483246    2.154017
------------------------------------------------------------------------------
```

(iv) แปลความหมาย.....

2.3) ข้อมูลในรูปตาราง 2 x 2

		DEAD 1	DEAD 0
AREA	1		
	0		

OR = ...

2.4) ค่าที่ได้จาก Logistic Model กับค่าที่ได้จากตาราง 2 x 2 เหมือนกัน หรือต่างกันอย่างไร

2.5) Logistic Model ข้างต้น เขียนในรูป Logit transformation ได้ดังนี้

3. การคำนวณค่า Odds Ratio (OR) จาก Multivariable model

3.1) ผลจากการวิเคราะห์ด้วย STATA เพื่ออธิบายความสัมพันธ์ระหว่าง AREA กับ DEAD โดยควบคุมผลของ DCHILD

```
. logit  dead area  dchild

Iteration 0:   log likelihood =  -188.1264
Iteration 1:   log likelihood = -179.19055
Iteration 2:   log likelihood = -177.85094
Iteration 3:   log likelihood =  -177.8476

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(2)      =      20.56
                                                  Prob > chi2     =     0.0000
Log likelihood =  -177.8476                       Pseudo R2       =     0.0546

------------------------------------------------------------------------------
       dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
------------+-----------------------------------------------------------------
       area |   .1419709   .2775134      0.512   0.609      -.4019454    .6858872
     dchild |   1.321847    .287817      4.593   0.000       .7577363    1.885958
      _cons |  -2.255301   .2251246    -10.018   0.000      -2.696537   -1.814065
------------------------------------------------------------------------------
```

3.2) ค่า OR หาได้โดย

(I) สูตรที่ใช้ $OR_{x1,Xo}$ =

(ii) ทารกในพื้นที่ทดลอง โดยให้ DCHILD คงที่ เป็น X_1 =.....

ทารกในพื้นที่ควบคุม โดยให้ DCHILD คงที่ เป็น X_0 =.....

(iii) แทนค่า $OR_{x1,Xo}$ =

ใช้ STATA คำนวณได้ดังนี้

```
. logistic  dead area  dchild

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(2)      =      20.56
                                                  Prob > chi2     =     0.0000
Log likelihood =  -177.8476                       Pseudo R2       =     0.0546

------------------------------------------------------------------------------
       dead | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
------------+-----------------------------------------------------------------
       area |   1.152543   .3198462      0.512   0.609       .6690173    1.985533
     dchild |   3.750343   1.079412      4.593   0.000       2.133441    6.592668
------------------------------------------------------------------------------
```

(iv) แปลความหมาย

3.3) ค่าที่ได้จาก Logistic Model กับค่าที่ได้จากStratified analysis จากแบบฝึกหัดที่ 1 ข้อ 3.9.2.2 เหมือนกันหรือต่างกันอย่างไร

3.4) Logistic Model ข้างต้น เขียนในรูป Logit transformation ได้ดังนี้

4. ในแบบฝึกหัดที่ 1 ข้อ 3.9.2.1 ได้วิเคราะห์แบบ Stratified analysis เพื่อหาความสัมพันธ์ระหว่าง AREA กับ DEAD โดยควบคุมผลกระทบจาก MALPRES ถ้าใช้ Logistic Regression วิเคราะห์เพื่อได้ตามวัตถุประสงค์เดียวกัน ควรเป็น Model อย่างไร (เขียนในรูป Logit transformation)

5. ต่อไปนี้เป็นรายละเอียดความหมายของ Output จาก STATA ในการวิเคราะห์โดยคำสั่ง logistic และ logit ผู้อ่านสามารถข้ามส่วนนี้ได้โดยไม่เสียความต่อเนื่อง
(ดัดแปลงจาก http://www.ats.ucla.edu/stat/stata/output/lognoframe.htm วันที่ 2 มกราคม 2544)

ผลจากคำสั่ง (*ตัวทึบในผลจากคำสั่งจาก **A ถึง **N** สำหรับแสดงตำแหน่งที่อธิบายความหมาย*)**

```
. logistic hicrime maleteen south educ police59

Logit estimates                                   Number of obsA   =         47
                                                  LR chi2(4)B      =      13.93
                                                  Prob > chi2C     =     0.0075
Log likelihoodD = -18.606959                      Pseudo R2E       =     0.2724

------------------------------------------------------------------------------
hicrimeF | Odds RatioB  Std. Err.H      zI    P>|z|J      [95% Conf. Interval]K
---------+--------------------------------------------------------------------
maleteen |   1.086959   .0478646      1.894   0.058       .9970804    1.184939
   south |   .3272305   .4449077     -0.822   0.411       .0227796     4.70068
    educ |   1.023187   .5723757      0.041   0.967       .3418133    3.062818
police59 |   1.059909   .0222633      2.770   0.006        1.01716    1.104455
------------------------------------------------------------------------------
```

```
. logit

Logit estimates                                   Number of obsA   =         47
                                                  LR chi2(4)B      =      13.93
                                                  Prob > chi2C     =     0.0075
Log likelihoodD = -18.606959                      Pseudo R2E       =     0.2724

------------------------------------------------------------------------------
hicrimeF |      Coef.L  Std. Err.M      zI    P>|z|J      [95% Conf. Interval]N
---------+--------------------------------------------------------------------
maleteen |   .0833837   .0440353      1.894   0.058      -.0029239    .1696914
   south |  -1.117091   1.359616     -0.822   0.411      -3.781888    1.547707
    educ |   .0229224   .5594047      0.041   0.967      -1.073491    1.119335
police59 |   .0581834   .0210049      2.770   0.006       .0170147    .0993522
   _cons |  -17.70177   9.495993     -1.864   0.062      -36.31357    .9100364
------------------------------------------------------------------------------
```

คำอธิบาย

[**A**] This is the number of observations being analyzed.

[**B**] This is the likelihood ratio chi-square with 4 degrees of freedom. One degree of freedom is used for each predictor variable in the logistic regression model. The likelihood-ration chi-square is defined as 2(L1 – L0), where L0 represents the log likelihood for the ″constant-only″ model and L1 is the log likelihood for the full model with constant and predictors. In this example, L0 = -25.573407 (which doesn′t show up in the output) and L1 = -18.606959 (which is found in item d below). Thus, the likelihood-ratio chi-square = 2*(-25.573407 – (-18.606959)) = 13.93.

[**C**] This is the p-value associated the chi-square with 4 degrees of freedom. The value of .0075 indicates that the model as a whole is statistically significant.

[**D**] This is the values of the log likelihood for the model including the constant and all of the predictors that was computed using the maximum-likelihood logit model.

[**E**] Technically, R2 cannot be computed the same way in logistic regression as it is in OLS regression. The pseudo-R2, in logistic regression, is defined as (1 - L1)/L0, where L0 represents the log likelihood for the "constant-only" model and L1 is the log likelihood for the full model with constant and predictors.

[**F**] This column starts with the name of the response variable (hicrime) and then lists the names of the predictor variables (maleteen south educ police59).

[**G**] The odds ratio column gives the amount of change expected in the odds ratio when there is a one unit change in the predictor variable with all of the other valiables in the model held constant. An odds ratio close to 1.0 suggest that there is no change due to the predictor variable.

In this example, the odds ratio for police59 is 1.059909. Thus, you would predict that the odds ratio would change by 1.059909 for every one unit change in police59 when maleteen, south and educ are held constant.

For a more detailed explanation of odds rations see the Stata FAQ: How do I interpret odds ratios in logistic regression?

[**H**] The standard error for the odds ratio is obtained from the logistic regression coefficient and its standard error using the formula:

se(odds ratio) = exp(coef.)*se(coef.)

[**I**] This column contains the z-statistic testing the logistic coeffieient.

In the case of the logit command, z = (coef.)/(Std. Err). For this example, z(police59) = .0581834/.0210049 = 2.770.

Stata uses the same z-test value computed for the logistic coefficient as the test of the odds ratio.

[**J**] This column contains the two-tail p-value for the z-test. Stata uses the same p-value computed testing the hypothesis, H0: b = 0, for both the logistic coefficients and for the odds ratios.

[**K**] This column contains the 95% confidence intervals for the odds ratios. Significant effects are suggested when confidence intervals do not contain 1.0. In this example, the only interval that would be considered significant at the .05 level is the one for police59. All of the other confidence intervals contain the value 1.0.

[**L**] The coefficient column gives the values for the logistic regression coefficients. These coefficients indicate the amount of change expected in the log odds when there is a one unit change in the predictor variable with all of the other variables in the model held constant. A coefficient close to 0 suggest that there is no change due to the predictor variable.

There is a relationship between the logistic coefficients and the odds ratios, odds ration = exp(coefficient). In this example the logistic coefficienct for police59 is .0581834, exp(.0581834) = 1.0599094, which is very close to the value of the odds ratio for police59.

Also in this example, the logistic coefficient for police59 is .0581834. Thus, you would predict that the log odds for hicrime would change by .0581834 for every one unit change in police59 when maleteen, south and educ are held constant.

The logistic coefficients can be used in a manner very similar to regression coefficient to generate predicted values. In this example,

```
predicted = -17.70177 + .0833837*maleteen -1.117091*south + .0229224*educ + .0581834*police59
```

You would get the same results in you used the predict command with the xb option.

[**M**] This column contains the standard error for the logistic regression coefficient which is used to compute the z-test for the coefficient.

[**N**] This column contains the 95% confidence intervals for the logistic regression coefficients. Significant effects are suggested when confidence intervals do not contain 0. In this example, the only interval that would be considered significant at the .05 level is the one for police59. All of the other confidence intervals contain the value 0.

การคำนวณค่า Odds Ratio ใน Logistic Regression Model และการแปลความหมาย

วัตถุประสงค์ : เพื่อให้ผู้เรียนสามารถ

1. อธิบายแนวคิดพื้นฐานการคำนวณค่า Odds Ratio จาก Logistic Regression ได้
2. คำนวณค่า Odds Ratio จาก Logistic Model ในกรณีต่าง ๆ ที่ระบุใน เนื้อหาได้ถูกต้อง
3. แปลความหมายค่า Odds Ratio ได้ถูกต้อง

เนื้อหา :

1. บทนำ
2. การคำนวณค่า Odds Ratio จาก Model ที่มีเฉพาะ Main effect หรือ Additive model ในกรณีต่อไปนี้
    - 2.1) เมื่อปัจจัยที่ศึกษาเป็นตัวแปรแจงนับ (Categorical Variable)
        - 2.1.1) มีค่าได้ 2 ค่า (Dichotomous) จำแนกตามการให้รหัสได้แก่
            - 2.1.1.1) ให้รหัสเป็น 0,1
            - 2.1.1.2) ให้รหัสเป็นแบบอื่น (Arbitrary coding of E)
        - 2.1.2) มีค่าได้มากว่า 2 ค่า
    - 2.2) เมื่อปัจจัยที่ศึกษาเป็นตัวแปรอันดับ (Ordinal Scaled Variable)
    - 2.3) เมื่อปัจจัยที่ศึกษาเป็นตัวแปรต่อเนื่อง (Continuous variable)
    - 2.4) การคำนวณค่า Odds Ratio กรณีมีหลาย Exposing factor
3. การคำนวณค่า Odds Ratio จาก Model ที่มี Interaction term หรือ Multiplicative model ในกรณีต่อไปนี้
    - 3.1) มี Interaction term ถึงอันดับที่สอง (Second order term)
    - 3.2) มี Interaction term ถึงลำดับที่สาม (Third order term)

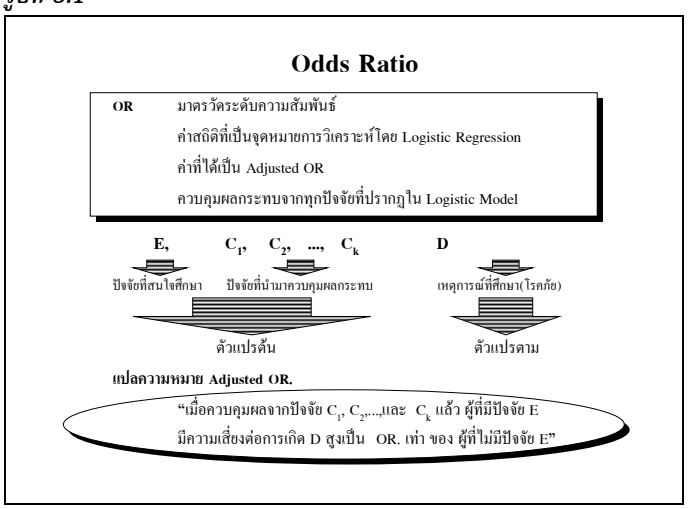
กิจกรรม :

1. ฟังบรรยายประกอบแผ่นใส พร้อมบันทึกเนื้อหาสำคัญลงในชุดการเรียนการสอน บทที่1
2. ทำแบบฝึกหัด
3. อภิปรายและสรุปเนื้อหา พร้อมเขียนสรุปท้ายบทลงกรอบว่างที่ให้ไว้ท้ายบท

สิ่งที่นำเสนอ	คำอธิบาย

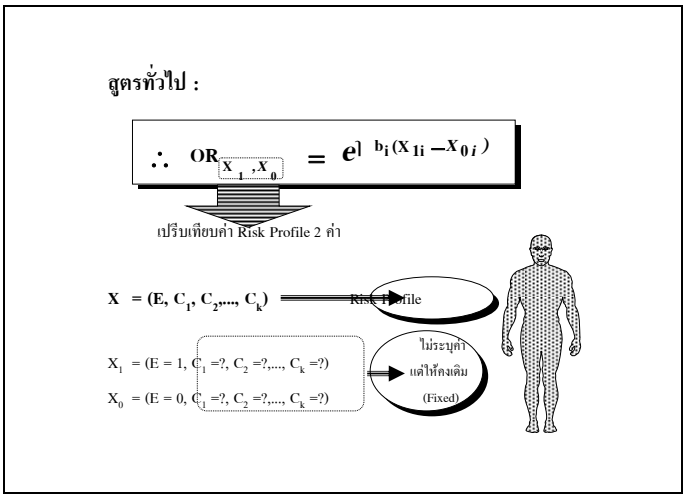
1. บทนำ

รูปที่ 3.1



เป้าหมายของการวิเคราะห์โดยใช้ Logistic Regression คือ ประมาณค่าระดับความสัมพันธ์ (Magnitude of association) ระหว่างปัจจัยที่ศึกษา (E) กับปัญหาที่ศึกษาซึ่งมักหมายถึงโรคภัยต่าง ๆ (D) โดยควบคุมผลกระทบจากตัวแปรอื่น ๆ ค่า OR ที่ได้จาก Logistic Regression เป็นค่าระดับความสัมพันธ์ที่ควบคุมผลจากทุกตัวแปรที่ปรากฏใน Model จึงเรียกว่าเป็น Adjusted OR

รูปที่ 3.2



จากสูตรทั่วไปของการคำนวณค่า OR จาก Logistic Regression Model บ่งชี้ว่า ต้องเปรียบเทียบค่า Risk Profile 2 ค่าเสมอ

ค่า Risk Profile หมายถึงลักษณะของบุคคลที่เราจะเปรียบเทียบกัน ซึ่งระบุค่าปัจจัยทุกปัจจัยที่ปรากฏใน Model การระบุค่าของปัจจัยเพื่อการเปรียบเทียบดังกล่าว จะระบุเพียงค่าของปัจจัยที่จะหาความสัมพันธ์กับ D เพียงปัจจัยเดียว เช่น E โดยให้ค่าทั้ง Risk Profile ที่หนึ่ง (X_1) และที่ศูนย์ (X_0) ตามที่ต้องการ ส่วนปัจจัยอื่น ๆ ทั้งหมด เช่น C_1 C_2 จนถึง C_k ไม่ต้องระบุค่าแต่คงเดิม (Unspecified but fixed) นั่นคือการควบคุมผลกระทบของ C_i

ถ้า E เป็น Dichotomous จะระบุให้เป็นค่า

1 สำหรับ Risk Profile ที่หนึ่ง ส่วน Risk Profile ที่ศูนย์มักเป็นค่าที่ใช้เป็นฐานของการเปรียบการเปรียบเทียบ หรือค่าอ้างอิง (Reference) เช่นให้เป็นค่า O เป็นต้น เมื่อแทนค่า Risk Profile ลงในสูตรคำนวณหา OR ค่าของตัวแปร C_i ทั้งหมดจึงหักลบกันเหลือ 0 คงเหลือปรากฏในสูตรเฉพาะค่าของ E เท่านั้น

การคำนวณค่า OR ที่ได้กล่าวแล้วนี้ กระทำได้เมื่อ E เป็น Dichotomous และไม่มี Interaction term ใน Model (ไม่มี Interaction effect) เท่านั้น กล่าวคือเป็น Model ที่ประกอบขึ้นจาก Main Effect

รูปที่ 3.3

จาก Model :

Logit P(X) = a + b_1E + b_2C_1 + b_3C_2 + ... + $b_{k+1}C_k$

แทนค่า Risk Profile ลงในสูตร OR

OR = $e^{b_1(1-0)+b_2(?-?)+b_3(?-?)+\ldots+b_{k+1}(?-?)}$

OR = $e^{b_1(1)+b_2(0)+b_3(0)+\ldots+b_{k+1}(0)}$

OR = e^{b_1}

ดังนั้น OR = Exponential ของค่าสัมประสิทธิ์ใน Model เมื่อ :-

- E (0,1) ⟶ (Dichotomous Variable)
- Model มีเฉพาะ Main effect ⟶ (E, C_1, C_2, ..., C_k)

  คือไม่มี Interaction term (เช่น E*C_1, E*C_k เป็นต้น)

2. การคำนวณค่า OR จาก Additive Model

เพื่อง่ายต่อการทำความเข้าใจ ตลอดบทเรียนนี้จะเขียน Logistic Model ในรูป Logit transformation และให้ตัวแปรต้นทุกตัวแทนด้วย X_i ตั้งแต่ X_1 จนถึง X_k เมื่อตัวแปรดังกล่าวมีจำนวน K ตัวแปรและสูตรทั่วไปทางการหาค่า OR ยังคงเขียนเช่นเดิมโปรดสังเกตความแตกต่างระหว่าง X_i กับ **X** ที่เป็น Risk Profile ซึ่ง **X** เป็นตัวอักษรตัวหนา

ส่วนที่ 2 นี้กล่าวถึงการคำนวณค่า OR จาก Logistic Regression Model ที่มีเพียง Main effect คือผลของตัวแปรใด ๆไม่ขึ้นกับตัวแปรอื่น เราเรียก Model ประเภทนี้ว่า Additive model ค่า OR ที่ได้เป็นค่าที่ใช้อธิบายความสัมพันธ์ของแต่ละตัวแปรเมื่อควบคุมผลจากตัวแปรอื่นทั้งหมด

รูปที่ 3.4

จาก Model :

Logit P(X) = a + b_1E + b_2c_1 + b_3c_2 + ... + $b_{k+1}c_k$

ให้ E = x_1, c_1=x_2, c_2=x_3 ,... c_k=x_{k+1}

เขียนใหม่เป็น

Logit P(X) = a + b_1x_1 + b_2x_2 +... + b_kx_k

สูตรทั่วไปในการหาค่า OR

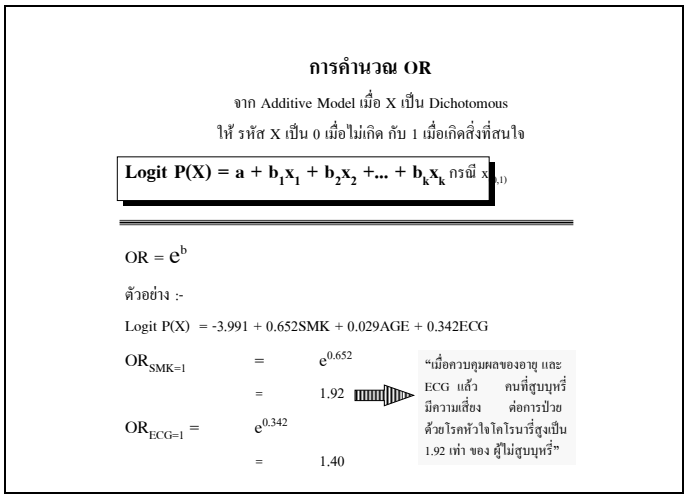
$$\therefore \quad OR_{X_1,\ X_0} = e^{\sum b_i(X_{1i}-X_{0i})}$$

2.1 กรณีมีตัวแปรต้นหนึ่งตัวซึ่งมีค่าได้สองค่า ให้รหัสเป็น 0,1

การคำนวณค่า OR จาก Model ที่ประกอบด้วยเฉพาะ Main effect เท่านั้น และตัวแปรต้นที่สนใจนั้นเป็น Dichotomous มีการให้รหัสเป็น 0 เมื่อไม่มีปัจจัย (Non-exposed) และเป็น 1 เมื่อมีปัจจัย (Exposed) กรณีดังกล่าว ค่า OR จะเท่ากับค่า Exponential ของค่าสัมประสิทธิ์ของตัวแปรที่สนใจ เช่น Model ที่ได้จากการศึกษาปัจจัยที่มีผลต่อการป่วยด้วยโรคหัวใจโคโรนารี่ ถ้าสนใจดูศึกษาดูผลของการสูบบุหรี่ เราอธิบายผลดังกล่าวด้วยค่า OR ซึ่งเท่ากับ Exponential ของค่าสัมประสิทธิ์ของตัวแปร SMK ซึ่งเท่ากับ 0.652 ได้ค่า OR เท่ากับ 1.92 แปลความหมายได้ว่า ***"เมื่อควบคุมผลของอายุและผลจาก ECG แล้ว คนที่สูบบุหรี่มีความเสี่ยงต่อการป่วยด้วยโรคหัวใจโคโรนารี่ สูงเป็น 1.9 เท่าของผู้ที่ไม่สูบบุหรี่"***

ในทำนองเดียวกัน ถ้าสนใจดูผลของ ECG ต่อการป่วย ก็หาได้เช่นเดียวกัน

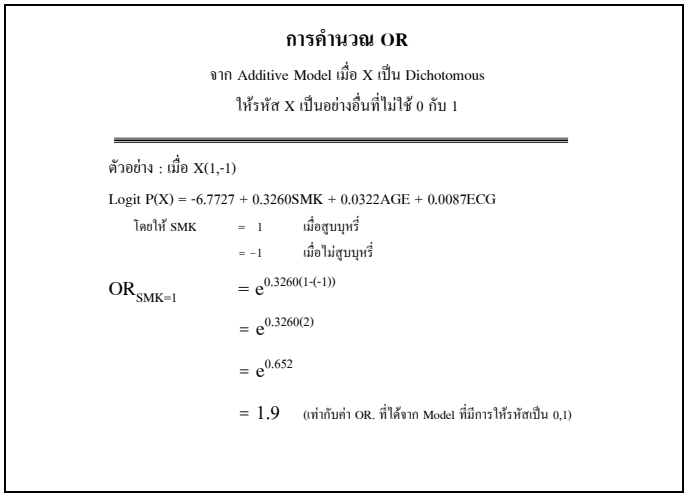
รูปที่ 3.5



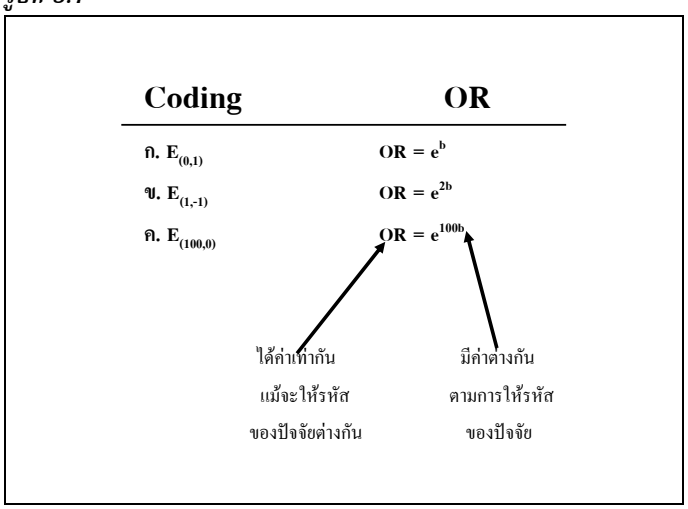
2.2 กรณีมีตัวแปรต้นหนึ่งตัวซึ่งมีค่าได้สองค่า รหัสไม่เป็น 0,1

รณีทีปัจจัยที่สนใจมีการให้รหัสเป็นแบบอื่น (Arbitrary coding of E) เช่น แทนที่จะเป็น 0 กับ 1 แต่เป็น 1 กับ -1 โดยให้เท่ากับ 1 เมื่อมีปัจจัย (Exposed) และเท่ากับ -1 เมื่อไม่มีปัจจัย (Non-exposed) วิธีการคำนวณค่า OR ใช้สูตรเดิม แต่ผลต่างของ X_1 กับ X_0 จะไม่ใช้ 1 เหมือนเดิม หากแต่เป็น 2 (คือ 1-(-1) เท่ากับ 1+1 เท่ากับ 2) คูณค่าสัมประสิทธิ์ แล้วหาค่า Exponential ได้ผลลัพธ์เป็นค่า OR ซึ่งจะไม่แตกต่างจากที่ได้จาก Model ที่มีการให้รหัสเป็น 0 กับ 1 แต่อย่างใด

รูปที่ 3.6



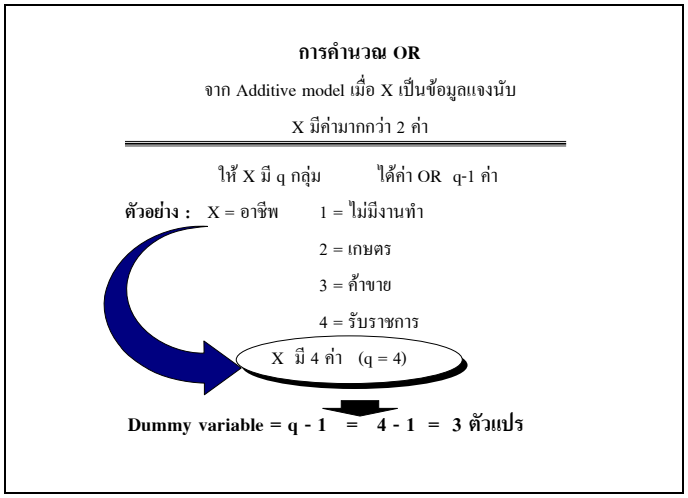
รูปที่ 3.7



โปรดสังเกตค่าสัมประสิทธิ์ของ Logistic Regression Model จากข้อมูลชุดเดียวกัน และประกอบด้วย ตัวแปรชุดเดียวกันเมื่อให้รหัสตัวแปรต่างกัน ค่าสัมประสิทธิ์จะต่างกันไปในแต่ละ Model ที่ได้ แต่ค่า OR ที่คำนวณได้จะเท่ากัน ไม่ว่าจะให้รหัสตัวแปรรูปแบบใดก็ตาม

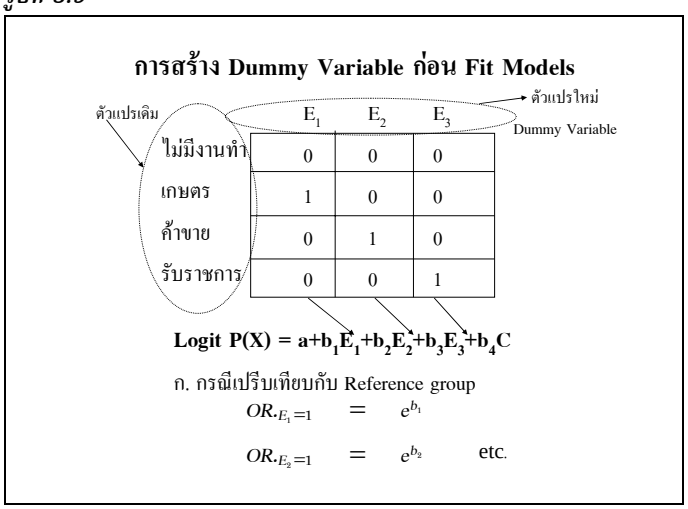
2.3 กรณีมีตัวแปรต้นหนึ่งตัวซึ่งมีมากกว่าสองค่า

รูปที่ 3.8



กรณีที่ปัจจัยที่สนใจ (E) เป็นตัวแปรแจงนับ (Categorical variable) ที่มีค่าได้มากกว่า 2 ค่าและไม่มีอันดับ หรือ Polytomous variable เช่นอาชีพ มี 4 กลุ่ม (Categories) เป็นต้น การอธิบายความสัมพันธ์ระหว่าง E ประเภทนี้กับ D จะต้องคำนวณค่า OR ของแต่ละกลุ่มเปรียบเทียบกับอีกกลุ่มหนึ่งที่เป็นกลุ่มอ้างอิง (Reference group) จะได้ค่า OR ทั้งหมด q-1 ค่า เมื่อ q คือจำนวนกลุ่มทั้งหมดของตัวแปร E นั่นคือต้องมีตัวแปรสำหรับ E อยู่ q-1 ตัวแปรใน Model ตัวแปรเหล่านี้ถูกสร้างขึ้นใหม่ เรียกว่า Dummy Variable โดยสร้างขึ้นก่อนการ Fit model เว้นแต่โปรแกรมคอมพิวเตอร์บางโปรแกรม มีการสร้าง Dummy Variable ให้โดยอัตโนมัติ เมื่อระบุว่าตัวแปรใดเป็นตัวแปรแจงนับที่ต้องสร้าง Dummy variable

รูปที่ 3.9



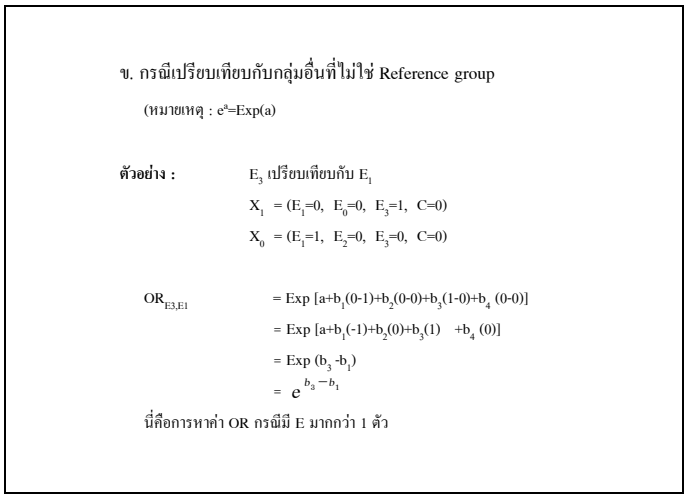
ตัวอย่างการสร้าง Dummy Variable เช่น กรณีอาชีพ (OCC) มี 4 กลุ่ม ได้แก่ ไม่มีงานทำ, เกษตรกรรม, ค้าขาย, และรับราชการ จะได้ Dummy Variable ทั้งหมด 3 ตัวแปร คือ E_1 (1=เกษตรกร 0=ไม่ใช่เกษตรกร) E_2(1= ค้าขาย 0=ไม่ใช่ค้าขาย) และ E_3 (1=รับราชการ 0=ไม่ใช่รับราชการ) การ Fit Model ต้องใส่ตัวแปร E_1 E_2 และ E_3 เข้าไปใน Model มิใช่ ตัวแปร OCC

หลังจาก Fit Model ถ้าแทนค่า E_1 E_2 และ E_3 ทุกตัวด้วยค่า 0 จะหมายถึงบุคคลนั้นไม่มีงานทำถ้าแทนค่า E_1 เท่ากับ 1 นอกนั้นเป็น 0 ก็จะหมายถึงมีอาชีพเกษตรกรรม เป็นต้น ดังนั้น กลุ่มที่จะให้เป็นกลุ่มเปรียบเทียบ (Reference group) จะถูกให้รหัสเป็น 0

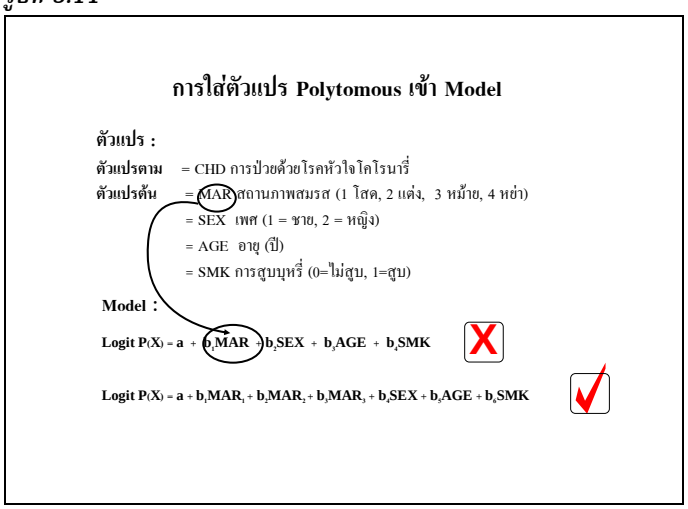
จะเห็นว่าเมื่อสร้าง Dummy Variable ตัวแปรที่ได้จะเป็น Dichotomous Variable ตัวหนึ่ง การหาค่า OR จึงกระทำได้เช่นเดียวกันกับกรณีตัวแปร $E_{(0,1)}$ ตามที่กล่าวแล้วข้างต้น คือเพียงแต่หาค่า Exponential ของค่าสัมประสิทธิ์ของกลุ่มที่สนใจเพียงค่าเดียว

กรณีต้องการเปรียบเทียบระหว่างกลุ่มที่ไม่ใช่ Reference group เช่นเปรียบเทียบเกษตรกร กับ ข้าราชการ ค่า OR จะได้จากค่า Exponential ของผลลัพธ์จากค่าสัมประสิทธิ์ของกลุ่มที่สนใจ ลบด้วยค่าสัมประสิทธิ์ของกลุ่มที่ต้องการเปรียบเทียบ จากนั้นใช้หลักการคำนวณเหมือนที่กล่าวข้างต้นคือใช้สูตรทั่วไปในการหา OR แล้วกำหนด Risk profile ของสองกลุ่มที่ต้องการเปรียบเทียบ แล้วแทนค่าในสูตร วิธีการนี้ใช้กับตัวแปรอื่นที่ไม่ใช่ Dummy Variable ได้เช่นกัน ซึ่งจะกล่าวในกรณีคำนวณค่า OR กรณีมี E มากกว่า 1 ตัวท้ายบทนี้

รูปที่ 3.10



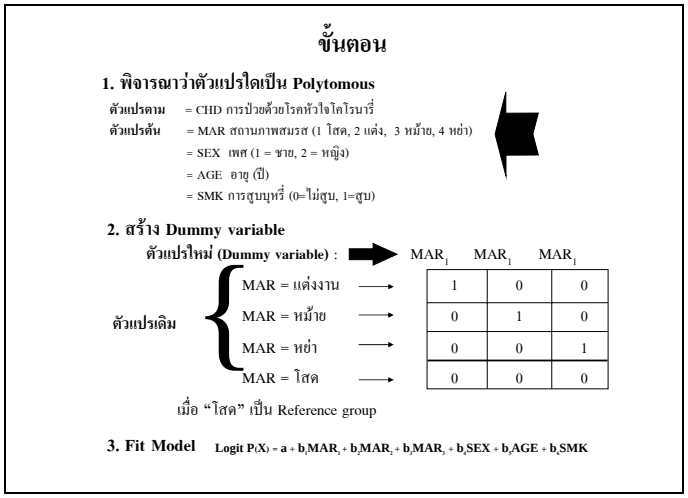
รูปที่ 3.11



ตัวอย่างการคำนวณ OR กรณีตัวแปรต้นที่เป็นตัวแปรแจงนับ และมีค่ามากกว่า 2 ค่า สิ่งแรกที่ต้องทราบคือตัวแปรใดมีลักษณะดังกล่าว จากตัวอย่างนี้ คือสถานภาพสมรส (MAR) ซึ่งมี 4 ค่า นอกเหนือจากนี้เป็นตัวแปร Dichotomous ทั้งหมดยกเว้นอายุ (AGE) ซึ่งเป็นตัวแปรต่อเนื่อง

ถ้าหากวิเคราะห์ข้อมูลโดยที่ไม่สร้าง Dummy variable ก่อน Model ที่ได้จะมีตัวแปร MAR อยู่ในรูปตัวแปรเพียงตัวเดียวอยู่ร่วมกับตัวแปรอื่น ๆ คือ เพศ (SEX) อายุ (AGE) และการสูบบุหรี่ (SMK) กรณีเช่นนี้เป็นการวิเคราะห์ที่กำหนดให้ MAR เป็นเสมือนตัวแปรต่อเนื่อง ซึ่งไม่ถูกต้อง

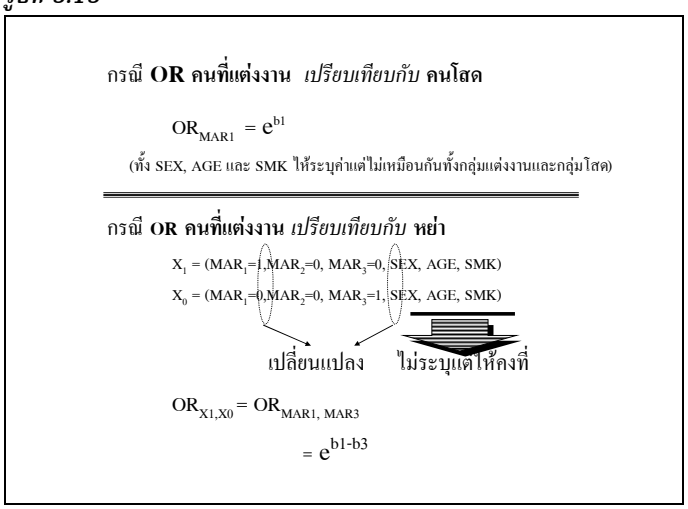
รูปที่ 3.12



กรณี X เป็นตัวแปร Polytomous ถ้ามีการสร้าง Dummy Variable ก่อนการ Fit Model ผลที่ได้จะมีตัวแปรที่บอกว่าเป็นอาชีพใด (MAR) จำนวน 3 ตัวแปร ปรากฏใน Model ร่วมกับตัวแปรอื่น ๆ ถือว่าเหมาะสมกับลักษณะข้อมูล

การสร้างDummy Variable มีหลายวิธี แต่ที่นิยมคือ วิธีที่เรียกว่า Partial method หรือ Reference cell coding ตามตัวอย่างที่แสดงในที่นี้ ทั้งนี้เพื่อทราบว่า กลุ่มใดหรือค่าใดในตัวแปรดังกล่าวนี้ถูกกำหนดให้เป็นฐานการเปรียบเทียบ (Reference group) ของกลุ่มอื่น ๆ เป็นกรณีตัวอย่างนี้ กลุ่มที่เป็นฐานคือ กลุ่มผู้ที่ยังไม่แต่งงาน (โสด)

รูปที่ 3.13



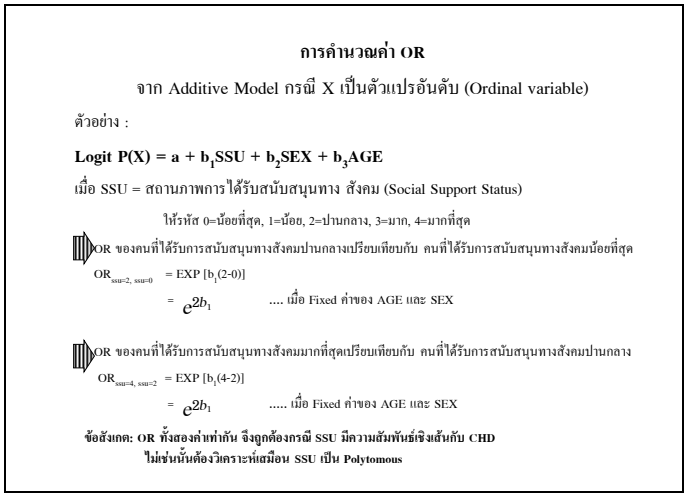
Note: *ก่อน Fit Model ต้องทราบ:*

- *ตัวแปรใด > 2 categories*
- *มีการให้รหัส Dummy variable เป็นแบบใด*
- *กลุ่มใดเป็น Reference group*

ดังนั้นค่า OR ของคนที่แต่งงานแล้ว เขียนแทนด้วย OR_{MAR1} เท่ากับค่า Exponential ของค่าสัมประสิทธิ์ของ MAR_1 นั่นคือ e^{b1} ค่าที่ได้นี้แปลความหมายได้ว่า “ผู้ที่มี เพศ อายุ และการสูบบุหรี่ เหมือนกัน คนที่แต่งงานจะมีความเสี่ยงต่อการป่วยด้วยโรคหัวใจโคโรนารี่ สูงเป็น $[e^{b1}]$ เท่าของ คนที่ไม่แต่งงาน” ในทำนองเดียวกัน ถ้าต้องการเปรียบเทียบระหว่างคนที่แต่งงานกับคนที่หย่ากันแล้ว เขียนแทนด้วย $OR_{MAR1,MAR3}$ เท่ากับ ค่า Exponential ของผลจากค่าสัมประสิทธิ์ของคนที่แต่งงาน (b_1) ลบด้วยค่าสัมประสิทธิ์ของคนที่หย่า (b_3) แปลความหมายได้ว่า “เมื่อให้เพศ อายุ และการสูบบุหรี่เหมือนกัน คนที่แต่งงานมีโอกาสป่วยด้วยโรคหัวใจโคโรนารี่สูงเป็น $[e^{b1-b3}]$ เท่าของคนที่หย่า”

2.4 กรณีมีตัวแปรต้นเป็นตัวแปรอันดับ

รูปที่ 3.14



กรณีที่ตัวแปรต้น เป็นตัวแปรอันดับ (Ordinal Variable) เช่นสถานภาพการได้รับสนับสนุนทางด้านสังคม (Social Support Status) เขียนแทนด้วย SSU ในการศึกษาปัจจัยที่มีผลต่อโรคหัวใจโคโรนารี่ต่อเนื่องจากตัวอย่างก่อนหน้านี้ โดยที่ค่าของตัวแปรมีค่าตั้งแต่ 0 คือได้รับการสนับสนุนทางสังคมน้อยที่สุด จนกระทั่งถึง 4 คือได้รับการสนับสนุนทางสังคมมากที่สุด

จะสังเกตได้ว่า SSU มีลักษณะคล้าย MAR จากตัวอย่างก่อนหน้านี้แต่ต่างกันตรงที่ค่าของ SSU สามารถเรียงลำดับได้ในขณะที่ค่าของ MAR ไม่สามารถเรียงลำดับได้

เมื่อต้องการหาความสัมพันธ์ของ SSU กับการป่วยด้วยโรคหัวใจโคโรนารี่ (CHD)

สามารถใส่ตัวแปร SSU ใน Model เพียงตัวแปรเดียวได้ถ้าพบว่าความสัมพันธ์ระหว่างSSU กับ CHD เป็นเชิงเส้น (Linear relationship) โดยไม่จำเป็นต้องมีการสร้างตัวแปรใหม่ (Dummy variable) การหาค่า OR จึงยังคงใช้สูตรทั่วไปในการคำนวณ (คือ OR = Exp $[\sum b_i(X_{1i}-X_{0i})]$ เมื่อต้องการอธิบายผลของ SSU ต่อ CHD โดยควบคุมผลจากทั้ง AGE และ SEX ยังคงเป็นการเปรียบเทียบ Odds จาก Risk Profile 2 ลักษณะที่ไม่ระบุค่า AGE และ SEX (มีค่าเดียวกันทั้งสอง Risk Profile) แต่ระบุเฉพาะค่า SSU ตามที่ต้องการจะเปรียบเทียบ

รูปที่ 3.15

$$\text{Logit } P(X) = 2.6341 - 0.4540SSU + 0.2016AGE + 1.010SEX$$

$$\begin{aligned} OR_{SSU=2,SSU=0} &= \text{Exp } [b_1(2-0) + b_2(0) + b_3(0)] \\ &= \text{Exp } [b_1(2)] \\ &= e^{2b_1} \\ &= e^{2(-0.4540)} \\ &= e^{-0.908} \\ &= 0.4 \end{aligned}$$

ตัวอย่างเช่น ค่า OR ของคนที่ได้รับการสนับสนุนทางสังคมปานกลางเปรียบเทียบกับคนที่ได้รับการสนับสนุนทางสังคมน้อยที่สุด หรือเขียนแทนด้วย $OR_{SSU2,SSU0}$ ค่า Risk Profile ที่แทนค่าลงในสูตรทั่วไปในการหาค่า OR จะยังคงเหลือเพียงผลต่างของค่า SSU ที่ต้องการเปรียบเทียบกัน นอกเหนือจากนี้คือค่าของ AGE และ SEX จะมีผลต่างเท่ากับ 0

ในกรณีตัวอย่างดังกล่าวผลต่างของค่า SSU เท่ากับ 2 คือ SSU ปานกลาง ซึ่งมีค่าคะแนนเท่ากับ 2 ลบด้วย 0 คือ SSU น้อยที่สุด ได้ผลลัพธ์เท่ากับ 2 ดังนั้น ค่า OR จึงเท่ากับค่า Exponential ของสองคูณด้วยค่าสัมประสิทธิ์ของ SSU นั่นคือ Exp^{2b1} การแปลความหมายยังคงเหมือนกรณีอื่น ๆ ก่อนหน้านี้

ในตัวอย่าง ค่าสัมประสิทธิ์ติดลบ (-0.454) จะยังผลให้ได้ค่า OR ต่ำกว่า 1 (คือ 0.4) เรียกปัจจัยที่มีลักษณะเช่นนี้ว่าปัจจัยป้องกัน (Protective factor)

2.5 กรณีตัวแปรต้นเป็นตัวแปรต่อเนื่อง

รูปที่ 3.16

การคำนวณค่า OR

จาก Additive Model กรณี X เป็นตัวแปรต่อเนื่อง

(Continuous variables)

ตัวอย่าง :

Logit P(x) $= a+b_1SBP+b_2SMH+ b_3ECG$

$OR_{SBP200,SBP120}$ $= Exp [b_1(200-120)+b_2(0)+ b_3(0)]$

$= Exp [b_1(80)+0+0]$

$= Exp [b_1(80)]$

$= e^{80b_1}$

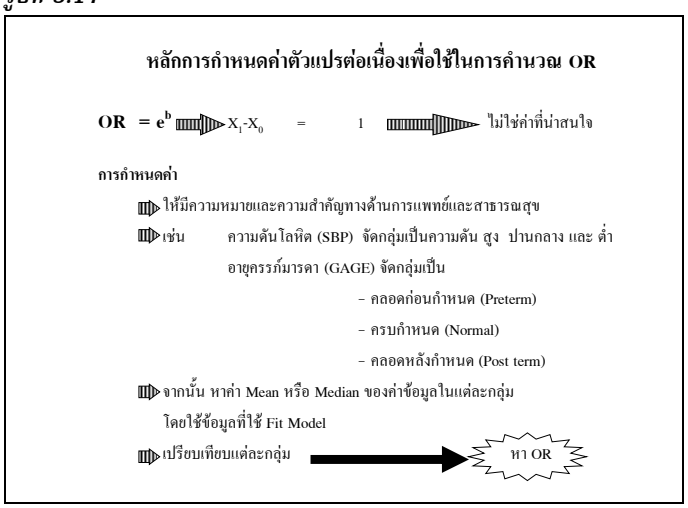
การคำนวณค่า OR จาก Model ที่ไม่มี Product term ใด ๆ ในกรณีที่ตัวแปรต้นเป็นตัวแปรต่อเนื่อง (Continuous variable) ทำได้เช่นเดียวกันกับกรณีตัวแปรต้นเป็นตัวแปรอันดับ (Ordinal variable) ตามที่กล่าวแล้วข้างต้น ตัวอย่างเช่นความดันโลหิต (Systolic Blood Pressure เขียนแทนด้วย SBP) 200 mmHg ตัวแปรประเภทนี้สามารถใส่เข้าไปใน Model ได้โดยตรง ยกเว้นกรณีพบว่าความสัมพันธ์กับตัวแปรตามไม่เป็นแบบเชิงเส้นต่อกันซึ่งต้องจัดกลุ่มก่อนแล้ววิเคราะห์แบบตัวแปรแจงนับ (จะกล่าวในรายละเอียดในบทที่ 5) ค่า OR คำนวณได้จากการกำหนดค่า Risk Profile สองค่าโดยให้ตัวแปรอื่น ๆ คงที่และมีค่าที่แตกต่างกันเพียงตัวแปรต่อเนื่องที่สนใจนี้เท่านั้น ตัวอย่างเช่น ต้องการหาค่า OR ของผู้ที่มีความดันโลหิต 200 mmHg เปรียบเทียบกับ ผู้ที่มีความดันโลหิต 120 mmHg ค่าความแตกต่างระหว่างสอง Risk Profile ของตัวแปรต้นอื่น ๆ จะเท่ากับ 0 ยกเว้นของ SBP ซึ่งเท่ากับ 200 ลบด้วย 120 เท่ากับ 80 ได้ OR เท่ากับ ค่า Exponential ของผลคูณระหว่าง 80 กับค่าสัมประสิทธิ์ของ SBP คือ b1 แปลความหมายได้ว่า “คนที่สูบบุหรี่และมีผลของ ECG เหมือนกันนั้น คนที่มีความดัน 200 mmHg จะมีโอกาสป่วยด้วยโรคหัวใจโคโรนารี่เป็น e^{80b1} เท่าของคนที่มีความดันโลหิต 120 mmHg” ทั้งนี้ OR นี้จะมีค่าเท่ากับของสอง Risk profile อื่นใดที่มีค่าความดันต่างกัน 80 mmHg ด้วยเหตุนี้จึงถูกต้องกรณีมี Linear relationship ตามที่กล่าวข้างต้น

ผลการวิเคราะห์จากคอมพิวเตอร์ส่วนมากให้ค่า OR โดยคำนวณจาก e^b ในทุกตัวแปร

ผลที่ได้ใช้ได้กรณีตัวแปรนั้นเป็นตัวแปร Dichotomous เท่านั้น นอกเหนือจากนี้ต้องใช้อย่างระมัดระวัง ซึ่งจะกล่าวในรายละเอียดถัดจากนี้ไป

กรณีตัวแปรดังกล่าวเป็นตัวแปรต่อเนื่อง ค่า e^b หมายถึงค่า Odds ของการป่วยเปรียบเทียบกันค่าสองค่าที่ต่างกันหนึ่งหน่วย เช่นOdds ต่อการเป็น CHD ของผู้ที่มีความดันโลหิต 200 เทียบกับของผู้ที่มีความดัน 199 เป็นต้น ค่า Odds สองค่าหารกันกลายเป็น Odds ratio จึงแปลความหมายได้ว่า ***"ทุก ๆ หนึ่ง mmHg ที่เพิ่มขึ้น Odds ของการป่วยจะคูณด้วย*** e^b" ซึ่งเข้าใจยากและไม่ใช่ค่าที่น่าสนใจ กรณีนี้ควรกำหนดค่าที่มีความสำคัญทางด้านการแพทย์และสาธารณสุข ซึ่งควรอิงกับทฤษฎี เช่น ความดันต่ำ ปกติ และสูง เป็นต้น หรืออายุครรภ์มารดา เป็นกลุ่มคลอดก่อนกำหนด ครบกำหนด และช้ากว่ากำหนด เป็นต้น แล้วใช้ค่าเฉลี่ยของแต่ละกลุ่มจากข้อมูลที่ใช้ในการ Fit Model นั้น แทนค่า X ใน Model เพื่อการเปรียบเทียบหาค่า OR ต่อไป

รูปที่ 3.17

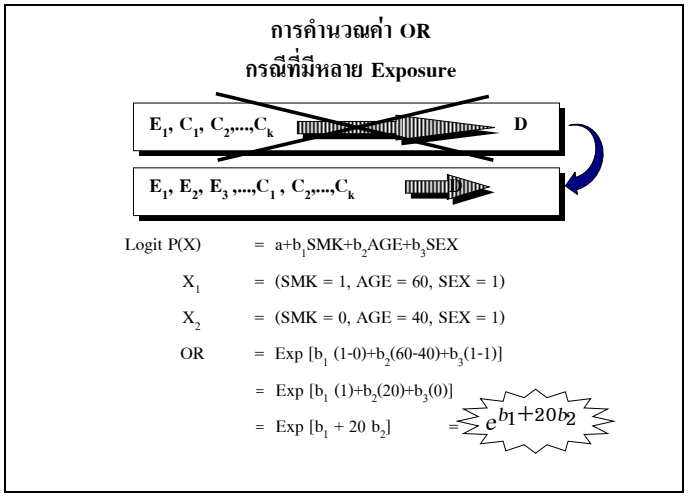


2.6 กรณีมีตัวแปรต้นมากกว่าหนึ่งตัว

การคำนวณค่า OR จาก Main Effect Model กรณีที่กล่าวแล้วทั้งหมด เป็นการวิเคราะห์เพื่ออธิบายความสัมพันธ์ของตัวแปรต้นตัวแปรเดียว (E) กับตัวแปรตามที่สนใจ (D) โดยที่ตัวแปรอื่น ๆ นอกนั้นเป็นตัวแปรควบคุม (C)

ในบางกรณีผู้วิจัยอาจสนใจอธิบายความสัมพันธ์ระหว่าง E กับ D โดยที่ E มีมากกว่า 1 ตัวแปร กล่าวคือลักษณะของคนที่นำมาเปรียบเทียบกัน หรือ Risk Profile ทั้ง 2 ชุด ที่นำมาเปรียบเทียบกันนั้น อาจมีการระบุค่าของตัวแปรมากกว่าหนึ่งตัว เช่น ต้องการคำนวณ OR ระหว่างผู้ที่สูบบุหรี่ (SMK=1)

รูปที่ 3.18



อายุ 60 ปี (AGE=60) และเป็นเพศชาย (SEX = 1) เปรียบเทียบกับผู้ที่ไม่สูบบุหรี่ (SMK=0)อายุ 40 ปี (AGE=40) และเป็นเพศชาย (SEX=1) โปรดสังเกตว่าตัวแปรควบคุมคือ SEX เท่านั้น อีกสองตัวแปรเป็นตัวแปรที่สนใจ (E) ขั้นแรก แทนค่าดังกล่าวลงในสูตรทั่วไปของการหาค่า OR ค่าความแตกต่างระหว่างสอง Risk Profile มีถึง 2 ตัวแปร คือของ SMK เท่ากับ 1-0 เท่ากับ 1 และค่าของ AGE เท่ากับ 60-40 เท่ากับ 20 ส่วนค่าของ SEX หักลบกันเหลือ 0 ยังผลให้ได้ค่า OR เท่ากับค่า Exponential ของผลจากค่าสัมประสิทธิ์ของ SMK บวกกับผลลัพธ์ของ 20 คูณด้วยค่าสัมประสิทธิ์ของ AGE แปลความหมายได้ว่า “ในคนเพศเดียวกันผู้ที่มีอายุ 60 ปี และสูบบุหรี่มีความเสี่ยงต่อโรคหัวใจโคโรนารี่สูงเป็น $[e^{b1+20b2}]$ เท่าของผู้ที่มีอายุ 40 ปีที่ไม่สูบบุหรี่” กรณีเช่นที่กล่าวนี้โอกาสใช้อาจมีไม่มากนัก แต่ต้องการแสดงให้เห็นว่าสามารถกระทำได้เช่นเดียวกับกรณี E เพียง 1 ตัวแปร อย่างไรก็ตาม ผู้วิจัยอาจต้องใช้เพื่ออธิบายความสัมพันธ์ที่จำเพาะกับกลุ่มของตัวแปรอิสระตามวิธีการนี้ในบางกรณี

3. การคำนวณค่า ORจาก Multiplicative model

รูปที่ 3.19

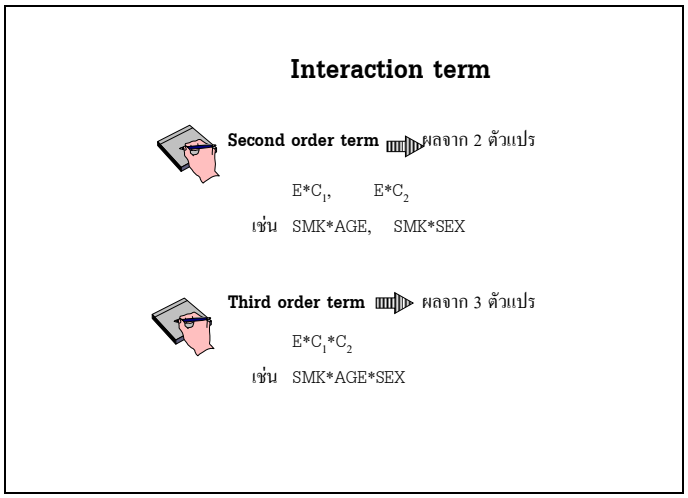
ที่กล่าวมาแล้ว เป็นการคำนวณค่า OR จาก Additive model ค่า OR ที่ได้เป็นค่าที่ใช้อธิบายความสัมพันธ์ของแต่ละตัวแปรเมื่อควบคุมผลจากตัวแปรอื่นทั้งหมด

บางกรณีผลของตัวแปรใด ๆขึ้นกับตัวแปรอื่นด้วย (มี Joint effect ระหว่างตัวแปร) Model ที่ได้มี Product term ด้วยอย่างน้อยหนึ่งตัว ซึ่งเป็น Interaction term เรียก Model

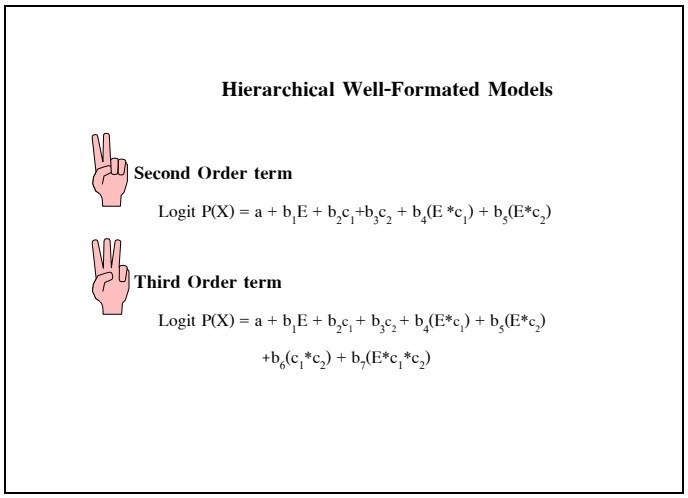
ประเภทนี้ว่า Multiplicative model

กลุ่มตัวแปรที่แสดงผลร่วมกันตามที่กล่าวจะต้องจัดให้อยู่ในรูป Interaction term ก่อนใส่ใน Model (คือผลคูณของสองตัวแปรขึ้นไป) ถ้ามีเพียงสองตัวแปรประกอบกันขึ้นเป็นหนึ่ง term เรียกว่า Second order term ถ้ามีสามตัวแปรประกอบกันขึ้นเป็นหนึ่ง term เรียกว่า Third order term แต่มากกว่านี้ไม่มีปรากฏ เพราะแม้จะทำได้ในเชิงสถิติ แต่แปลความหมายอย่างสลับซับซ้อนมาก มีความจำเพาะเกินความจำเป็นจึงไร้ประโยชน์ แม้แต่ Third order term ยังพบได้น้อยมากที่ผู้วิจัยจะคงไว้ใน Model เนื่องจากแปลความหมายให้สื่อเข้าใจได้ยาก

รูปที่ 3.20



รูปที่ 3.21



กรณีมี Interaction effect การเขียนรูป Model จะต้องรวมตัวแปรเดี่ยว หรือ Main effect และ Interaction term ที่อยู่ต่ำกว่าลำดับที่สูงที่สุดของ Model เสมอ ทั้งนี้จะกล่าวถึงเหตุผลและรายละเอียดอื่นๆ ในบทที่ 4 ตัวอย่างเช่น กรณี Model มีอันดับสูงสุดเป็น Third order term ใน Model นอกจากต้องมี Main effect ทุกตัวแปรแล้ว ยังต้องมี Second order term ทุก term ที่ประกอบกันขึ้นเป็น Third order term เป็นต้น หลักการนี้เรียกว่า Hierarchical Well-formatted model การกำหนดรูปแบบ Model ตามนี้ จะยังผลให้ค่าสัมประสิทธิ์ที่ได้นั้น ไม่ขึ้นอยู่กับการให้รหัสตัวแปร กล่าวคือไม่ว่าจะให้รหัสค่าตัวแปรต่างวิธีกันอย่างไร ก็จะให้ค่าสัมประสิทธิ์เท่ากันและค่า OR เท่ากัน ซึ่งถือเป็นสิ่งจำเป็น

การคำนวณค่า OR กรณีมี Interaction term ยังคงใช้หลักการเดิมตามที่ได้กล่าวมาแล้ว กล่าวคือต้องระบุ Risk Profile 2 ชุด เพื่อเปรียบเทียบกัน จากนั้นแทนค่าลงในสูตรทั่วไปในการคำนวณหา OR

3.1 กรณีอันดับสูงสุดเป็น Second order term

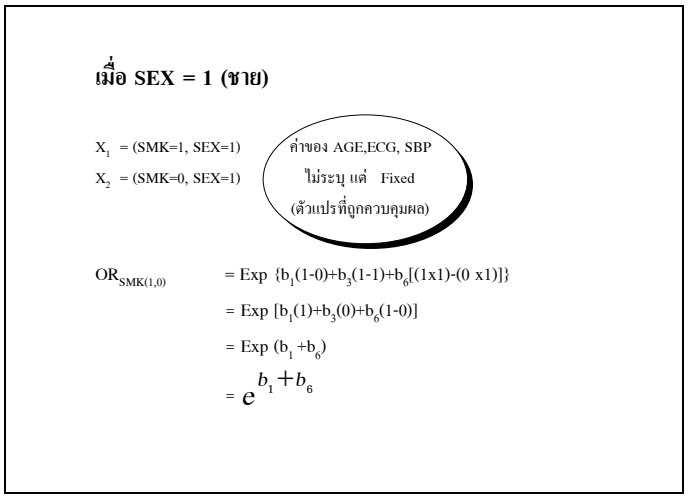
กรณีที่มี Interaction term อันดับสูงสุดเป็น Second Order term จากตัวอย่างการศึกษาเพื่อหาความสัมพันธ์ระหว่างการสูบบุหรี่ (SMK) กับการป่วยด้วยโรคหัวใจโคโรนารี่ (CHD) โดยมีตัวแปรต้นอื่น ๆ ที่ต้องการควบคุมผลกระทบ รวมทั้งตัวแปรที่พบว่าเป็นตัวกวน (Confounder) จำนวนหนึ่ง เช่น อายุ (AGE) เพศ (SEX) ผลการตรวจคลื่นไฟฟ้าหัวใจ (ECG) และระดับความดันโลหิต (SBP) เป็นต้น

ตัวแปรใดที่เป็นตัวกวนเมื่อใส่ใน Model ก็คือผลของตัวแปรนั้นได้รับการควบคุมแล้ว แต่ถ้าตัวแปรใดเป็น Effect Modifier จะถูกใส่ไว้ใน Model ในรูป Interaction term เช่นพบว่า SEX เป็น Effect Modifier จะได้ term SMK*SEX ใน Model ดังนั้นค่า OR ของ SMK กับ CHD จะไม่เท่ากับที่ได้จาก Model ที่ไม่มี Interaction term ทั้งนี้เพราะยังมีอีก term หนึ่งที่ยังติดค่า SMK อยู่ นั่นคือ Interaction term ที่เขียนเป็น SMK*SEX สิ่งนี้แสดงให้ทราบว่าค่า OR_{SMK} ขึ้นกับค่า SEX กล่าวคือขึ้นอยู่กับว่าเป็นชายหรือหญิง

ดังนั้น ถ้าต้องการหาค่า OR_{SMK} ของชายหรือ SEX = 1 ขั้นแรกต้องระบุ Risk Profile 2 ชุด เพื่อเปรียบเทียบกัน ใน Model ค่าที่ต้องระบุมีเพียงค่าของ SMK กับ SEX โดยที่ SMK ต้องให้ต่างกัน (เพื่อให้ได้สอง Odds ไว้หารกัน) ส่วน SEX ให้เหมือนกัน โดยต้องระบุค่าในตัวอย่างนี้คือให้เท่ากับ 1 ในทั้งสอง Risk profile (เพื่อจำกัดให้ได้ค่า OR ของ SMK ต่อ CHD ในเฉพาะกลุ่มของ SEX) ในขณะที่ตัวแปรอื่น ๆ นอกเหนือจากนี้ ไม่ต้องระบุค่าแต่ความหมาย คือ Fixed (เพื่อควบคุมผลกระทบนั่นเอง)

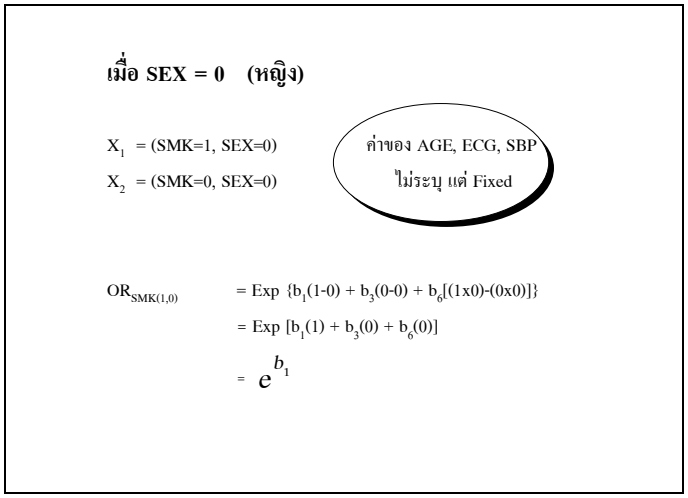
รูปที่ 3.22

รูปที่ 3.23



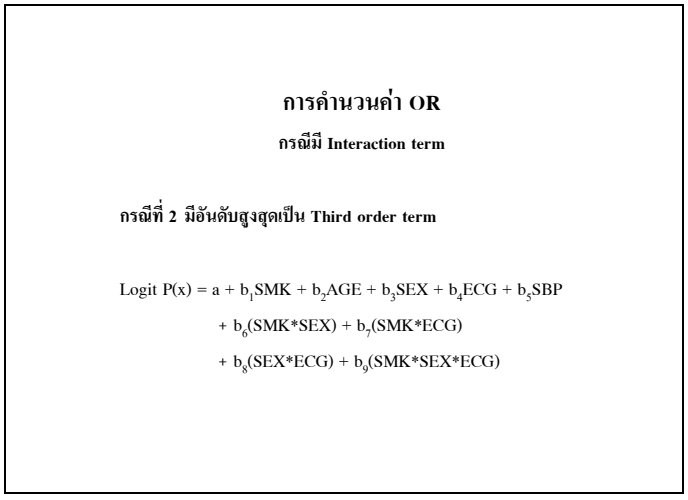
จากนั้นแทนค่าลงในสูตรทั่วไปของการหา OR โดยที่ผลต่างของ Risk Profile ในแต่ละตัวแปร $(X_{1i}-X_{0i})$ จะมีเพียงค่าผลต่างของตัวเลข 2 ตัวตามที่กำหนด ยกเว้นของ Interaction term ที่เป็นผลต่างระหว่างผลคูณ กล่าวคือ $(X_{11}X_{12}-X_{01}\ X_{02})$ หรือผลคูณของค่า Interaction term เมื่อเป็นค่าของ Risk Profile ที่ 1 (X_1) ลบด้วยของ Risk Profile ที่ 2 (X_2) นั่นเอง ค่า OR_{SMK} เท่ากับค่า Exponential ของผลบวกระหว่างค่าสัมประสิทธิ์ของ SMK กับค่าสัมประสิทธิ์ของ Interaction term (SMK*SEX) ดังนั้น ค่า OR_{SMK} กรณีเพศชายเท่ากับ $e^{b1\ +\ b6}$ ในทำนองเดียวกันค่า OR_{SMK} ของเพศหญิงก็หาได้เมื่อกำหนดค่า SEX=0 แล้วคำนวณโดยวิธีเดียวกัน ได้ค่า OR = e^{b1} แปลความหมายได้ว่า ***"ในกลุ่มผู้ชายที่มีอายุเท่ากันและผลของ ECG เหมือนกันนั้น ผู้ที่สูบบุหรี่มีความเสี่ยงต่อการป่วยด้วยโรคหัวใจ โคโรนารี่ สูงเป็น $e^{b1\ +\ b6}$ เท่าของผู้ไม่สูบบุหรี่"*** และ ***"ในกลุ่มผู้หญิงที่มีอายุเท่ากันและผลของ ECG เหมือนกันนั้น ผู้ที่สูบบุหรี่ มีความเสี่ยงต่อการป่วยด้วยโรคหัวใจโคโรนารี่ สูงเป็น e^{b1} เท่าของผู้ที่ไม่สูบบุหรี่"***

รูปที่ 3.24



3.2 กรณีอันดับสูงสุดเป็น Third order term

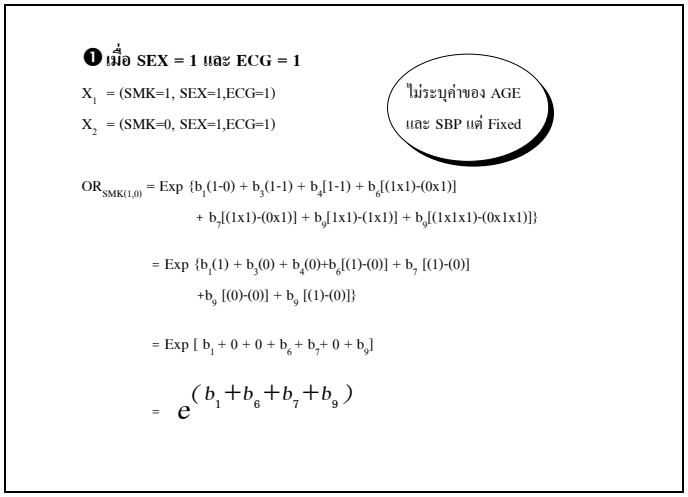
รูปที่ 3.25



กรณีอันดับสูงสุดเป็น Third order term นั้นตามหลักการของ Hierarchical Well-formatted Model ทุก ๆ term ที่เป็น Second order term ที่ประกอบกันเป็น Third order term ต้องระบุใน Model ด้วย ซึ่งมีทั้งหมด 3 term เช่นกรณี Third order term เป็น (SMK*SEX*ECG) ต้องมี Second order term ทั้งหมด 3 term ได้แก่ (SMK*SEX)

(SMK*ECG) และ (SEX*ECG) ที่เหลือคือ Main effect ซึ่งอย่างน้อยต้องมี SMK SEX และ ECG ใน Model แต่ตัวแปรอื่น เช่น AGE อาจอยู่ใน Model ถ้าต้องการควบคุมผลกระทบ

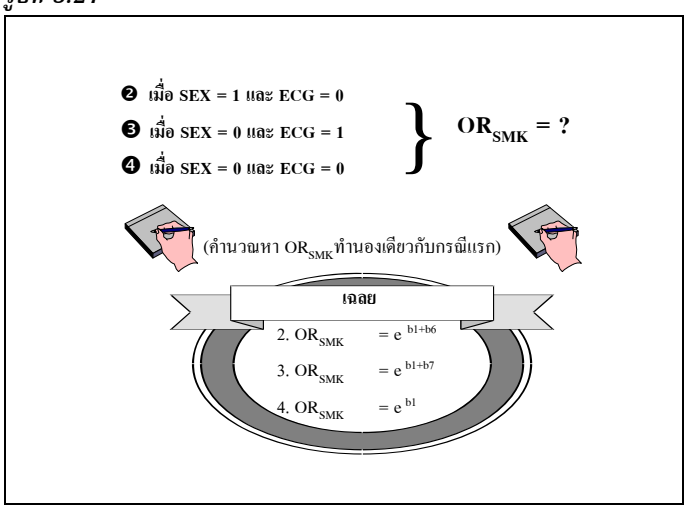
รูปที่ 3.26



กรณีนี้ การอธิบายความสัมพันธ์ระหว่าง SMK กับ CHD จะต้องอธิบายจำแนกตามกลุ่มของ SEX และค่าของ ECG เช่น ค่า OR_{SMK} ของกลุ่ม**<u>ชาย</u>**ที่มี **<u>ECG</u>** ผิดปกติ

นอกจากนี้ยังมีค่า OR_{SMK} ของกลุ่มชายที่มี ECG ปกติ OR_{SMK}ของกลุ่มหญิงที่มี ECG ผิดปกติ และ OR_{SMK} ของกลุ่มหญิงที่มี ECG ปกติ รวมเป็นทั้งหมด 4 ค่า หรือได้ค่า OR ทั้งหมดเท่ากับผล 2×2 ค่า (ผลคูณของจำนวนระดับของ Effect Modifier)

รูปที่ 3.27

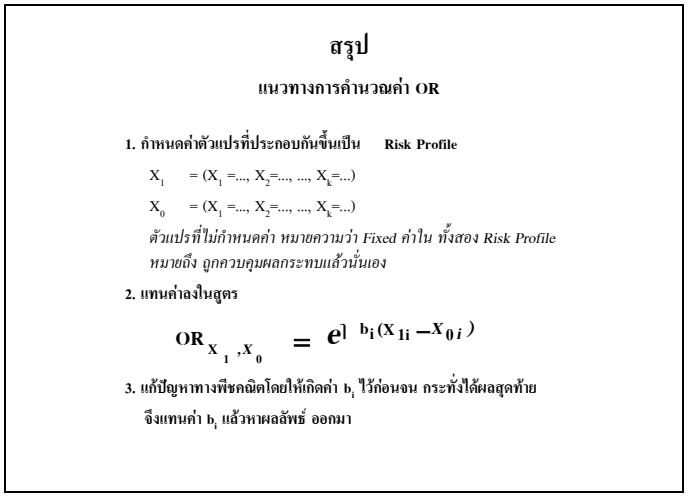


การคำนวณค่า OR_{SMK} ทั้ง 4 ค่าดังกล่าว ใช้หลักการเดียวกันกับที่คำนวณกรณีมีเพียง Second order term ตามที่กล่าวข้างต้น กล่าวคือต้องระบุค่า Risk Profile 2 ค่า คือ X_1 กับ X_0 เพื่อเปรียบเทียบกัน แล้วแทนค่าลงในสูตรทั่วไปของการหา OR แก้ปัญหาทางพีชคณิตจนกระทั่งได้ค่าสุดท้ายติดค่าไว้ในรูปค่า Exponential จากนั้นหาค่าสุทธิโดยแทนค่าสัมประสิทธิ์ เพื่อหาค่า Exponential เป็นค่าของ OR ในที่สุด

ตัวอย่างเช่นกรณีสุดท้ายตามที่กล่าวข้างต้น คือค่า OR ในกลุ่มผู้หญิงที่มีค่า ECG ปกติ เท่ากับ e^{b1} เป็นต้น แปลความหมาย ได้ว่า “ในกลุ่มผู้หญิงที่มีอายุและความดันโลหิตเท่ากัน และผล ECG ปกติ คนที่สูบบุหรี่มีความเสี่ยงต่อการป่วยด้วยโรคหัวใจโคโรนารี่สูงเป็น e^{b1} เท่าของคนที่ไม่สูบบุหรี่”

4. สรุปการคำนวณค่า OR

รูปที่ 3.28



โดยสรุป การคำนวณค่า OR จาก Logistic Regression model ไม่ว่ากรณีใด ๆ มีแนวทางทั่วไปดังนี้คือ

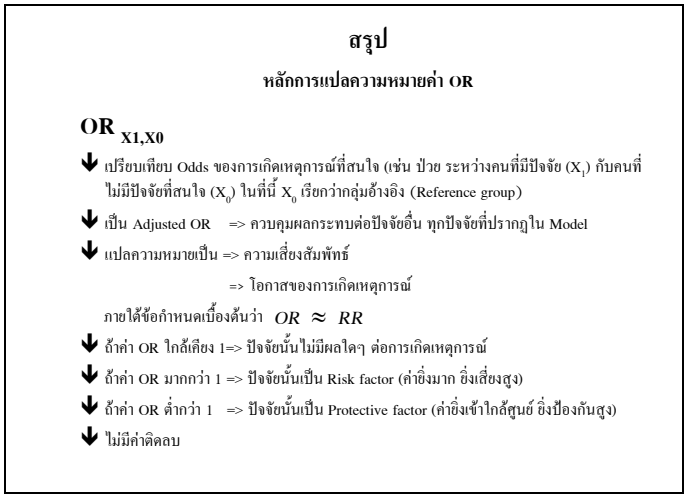
ขั้นแรกต้องกำหนดค่าตัวแปรที่ประกอบกันขึ้นเป็น Risk Profile ทั้ง 2 ชุดที่จะเปรียบเทียบกันก่อน ค่าของตัวแปรใดที่ไม่ได้ระบุค่าแต่มีอยู่ใน Model และให้คงเดิมในทั้งสองProfile หมายถึงผลของตัวแปรเหล่านั้นได้รับการควบคุมแล้ว

ขั้นที่สองแทนค่าลงในสูตรทั่วไปของการคำนวณค่า OR แล้วแก้ปัญหาทางพีชคณิต โดยยังคงติดค่าสัมประสิทธิ์ไว้ก่อน

ขั้นที่สามแทนค่าสัมประสิทธิ์ลงไปแล้วหาค่า Exponential จากผลดังกล่าว ได้เป็นค่า OR

5. การแปลความหมายของค่า OR

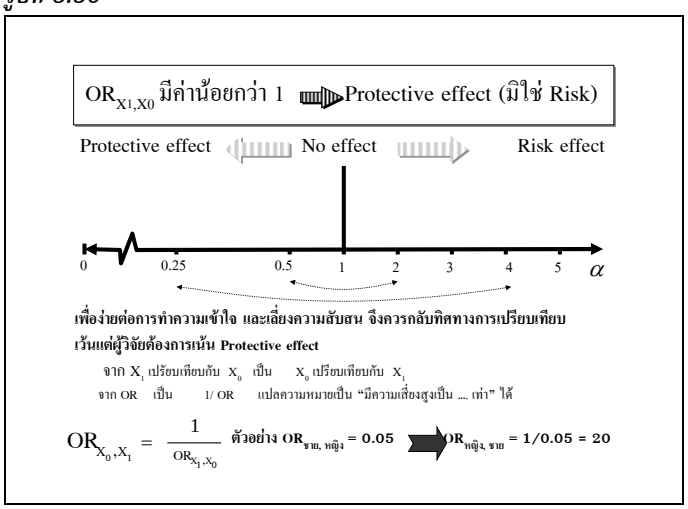
รูปที่ 3.29



การได้มาซึ่งค่า OR เป็นการเปรียบเทียบค่า Odds ของการเกิดเหตุการณ์ที่ศึกษา (เช่นการป่วย) ระหว่างคนที่มีปัจจัย (Exposed หรือ X_1) กับคนที่ไม่ปัจจัย (Non-exposed หรือ X_0) ค่า OR ที่ได้นั้น ได้ควบคุมผลกระทบจากทุกตัวแปรที่ปรากฏใน Model แล้วจึงเรียกว่า Adjusted Odds Ratio การแปลความหมายนิยมแปลเป็นความเสี่ยง หรือ โอกาสที่จะเกิดเหตุการณ์ที่ศึกษา ว่า ***"คนที่มีปัจจัย มีความเสี่ยงที่จะเกิดเหตุการณ์เป็นกี่เท่าของคนที่ไม่มีปัจจัย เมื่อควบคุมผลกระทบจากตัวแปรอื่น ๆ ทุกตัวแปร ที่อยู่ใน Model"*** ทั้งนี้ เป็นการแปลความหมายภายใต้ข้อกำหนดเบื้องต้นว่าค่า OR เป็นตัวประมาณค่าที่ดีที่สุดของค่า RR

กรณีที่ค่า OR มากกว่า 1 หมายถึงปัจจัยดังกล่าวเป็นปัจจัยเสี่ยง (Risk factor) กรณี OR น้อยกว่า 1 หมายถึงปัจจัยดังกล่าวเป็น

รูปที่ 3.30



ปัจจัยป้องกัน (Protective factor)

กรณีตัวแปรใดเป็น Protective factor ผู้วิจัยอาจแสดงค่า OR ตามที่ได้ (คือค่าที่น้อยกว่า 1) เพื่อแสดงให้เห็นว่าตัวแปรนั้นมี Protective effect แต่มักสื่อความหมายได้ยาก เพราะเมื่อแปลความหมายว่า "มีความเสี่ยงเป็นกี่เท่า" แต่ความจริงคือไม่ได้มีความเสี่ยง หรือแม้แต่กล่าวว่า "ป้องกันเป็นกี่เท่า" ก็ไม่สื่อความหมาย ที่สำคัญคือพบว่ามีความเข้าใจที่ผิดได้ง่ายเพราะคนส่วนมากเข้าใจว่าช่วงระหว่าง 0 ถึง 1 เท่ากันกับช่วงระหว่าง 1 ถึง 2 ซึ่งค่า OR ไม่เป็นจริงตามนั้น (ดูรายละเอียดการแปลความหมายใน Jaeschke et al., 1995)

ตัวอย่างเช่นค่า OR 1.05 กับ 1.5 แตกต่างกันน้อยมาก (ความเสี่ยงประมาณเท่าๆกัน) แต่ถ้า 0.05 กับ 0.5 ดูเหมือนแตกต่างไม่มาก แต่ความจริงเมื่อกลับกลุ่มอ้างอิงแล้วค่า OR เท่ากับ 1/0.05 = 20 และ 1/0.5 = 2 ตามลำดับ (ความเสี่ยงประมาณ 20 เท่าเทียบกับ 2 เท่า)

ดังนั้นผู้วิจัยจึงมักให้กลุ่มอ้างอิงเป็นกลุ่มที่มีความเสี่ยงต่ำ เพื่อให้ได้ค่า OR มากกว่า 1การเปลี่ยนกลุ่มอ้างอิงก็คือการผกผันค่า OR ซึ่งเท่ากับหนึ่งหารด้วยค่า OR นั่นเอง

สรุปท้ายบท :

เอกสารอ้างอิงประจำบทที่ 3

Jaeschke, R., Guyatt, G., Shannon, H., Walter, S. Cook, D. Heddle, N. (1995). Assessing the effects of treatment: measures of association . *Canadian Medical Association Journal*. 152: 351-357

Kleinbaum, D.G. (1994). *Logistic Regression: A self-learning text*. New York Springer-Verlag.

แบบฝึกหัดที่ 3

1. จากแบบฝึกหัดที่ 1 ข้อ 2.2.3 ตัวแปร BWT ซึ่งเป็นตัวแปรต่อเนื่อง ได้รับการแปลงเป็นตัวแปรแจงนับในชื่อตัวแปร BWTG

1.1 จากองค์ความรู้ด้านการแพทย์ ทราบว่า ทารกที่มีน้ำหนักแรกเกิดต่ำหรือสูงเกินปกติ จะมีโอกาสตายสูงกว่าทารกที่น้ำหนักแรกเกิดปกติ ดังนั้น Dummy Variable สำหรับตัวแปร BWTG จึงควรเขียนเป็นตารางแสดงค่าได้ดังนี้ โปรดเติมรหัสตัวแปรลงในช่องว่าง

ให้ BWTG = 2 เป็นกลุ่มอ้างอิง (Reference group)

	ตัวแปรใหม่ (Dummy Variable)	
ตัวแปรเดิม	**BWTD1**	**BWTD3**
BWTG = 1	.....	.....
BWTG = 2	.....	.....
BWTG = 3	.....	.....

1.2 จาก Output ที่ใช้ STATA วิเคราะห์ โปรดเขียน Logistic regression model

```
. tab  bwtg, gen(bwtd)

       bwtg |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         39        8.39        8.39
          2 |        140       30.11       38.49
          3 |        286       61.51      100.00
------------+-----------------------------------
      Total |        465      100.00

. logit  dead area  bwtd1 bwtd3

Iteration 0:   log likelihood =  -188.1264
Iteration 1:   log likelihood =  -180.5665
Iteration 2:   log likelihood = -179.79369
Iteration 3:   log likelihood = -179.79198
Iteration 4:   log likelihood = -179.79198

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(3)      =      16.67
                                                  Prob > chi2     =     0.0008
Log likelihood = -179.79198                       Pseudo R2       =     0.0443

------------------------------------------------------------------------------
        dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |   .0911082   .2777323      0.328   0.743      -.4532371    .6354536
       bwtd1 |   .6092406     .40922      1.489   0.137      -.1928159    1.411297
       bwtd3 |  -.8628492   .2980022     -2.895   0.004      -1.446923   -.2787755
       _cons |  -1.483589   .2677199     -5.542   0.000       -2.00831   -.9588672
------------------------------------------------------------------------------
```

1.3 ค่า OR เพื่ออธิบายความสัมพันธ์ระหว่างน้ำหนักแรกเกิดกับการตายของทารก หาได้โดย

1.3.1) OR สำหรับ ทารกน้ำหนักแรกเกิดต่ำกว่าปกติ โดยใช้ทารกน้ำหนักแรกเกิดปกติเป็นกลุ่มอ้างอิง

(I) สูตรที่ใช้ $OR_{x1,Xo}$ =

(ii) ทารกน้ำหนักแรกเกิดต่ำกว่าปกติ

เป็น X_1 =

.....

ทารกน้ำหนักแรกเกิดปกติ

เป็น X_0 =

.....

(iii) แทนค่า $OR_{x1,Xo}$ =

1.3.2) OR สำหรับ ทารกน้ำหนักแรกเกิดสูงกว่าปกติ โดยใช้ทารกน้ำหนักแรกเกิดปกติเป็นกลุ่มอ้างอิง

(I) สูตรที่ใช้ $OR_{x1,Xo}$ =

(ii) ทารกน้ำหนักแรกเกิดสูงกว่าปกติ

เป็น X_1 =

.....

ทารกน้ำหนักแรกเกิดปกติ เป็น

เป็น X_0 =

.....

(iii) แทนค่า $OR_{x1,Xo}$ =

1.3.3) ใช้ STATA คำนวณได้ดังนี้

```
. logistic  dead area  bwtd1 bwtd3

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(3)      =      16.67
                                                  Prob > chi2     =     0.0008
Log likelihood = -179.79198                       Pseudo R2       =     0.0443

------------------------------------------------------------------------------
        dead | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |   1.095388   .3042245      0.328   0.743       .6355674    1.887878
       bwtd1 |   1.839034   .7525696      1.489   0.137       .8246337    4.101271
       bwtd3 |   .4219581   .1257445     -2.895   0.004       .2352932    .7567098
------------------------------------------------------------------------------
```

2. จากการวิเคราะห์แบบ Stratified analysis ในแบบฝึกหัดที่ 1 ข้อ 3.9.2.1 และคำถามที่ 4 ของแบบฝึกหัดที่ 2 พบว่าท่าการคลอดของทารก (MALPRES) เป็น Effect Modifier ของความสัมพันธ์ระหว่างการตายของทารก (DEAD)กับพื้นที่ที่ศึกษา (AREA)

2.1) จงเขียน Model ในรูปของ Logit transformation

2.2) ผลการวิเคราะห์จาก STATA ได้ดังนี้

```
. gen a_mal = area * malpres

. logit  dead area malpres a_mal

Iteration 0:   log likelihood =  -188.1264
Iteration 1:   log likelihood = -170.29205
Iteration 2:   log likelihood = -162.58391
Iteration 3:   log likelihood = -162.10446
Iteration 4:   log likelihood = -162.10316

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(3)      =      52.05
                                                  Prob > chi2     =     0.0000
Log likelihood = -162.10316                       Pseudo R2       =     0.1383

------------------------------------------------------------------------------
        dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.3988154   .3230824     -1.234   0.217      -1.032045    .2344146
     malpres |   .8903152   .8428469      1.056   0.291      -.7616343    2.542265
       a_mal |   2.362425   .9739987      2.425   0.015       .4534228    4.271427
       _cons |  -1.988928   .2091045     -9.512   0.000      -2.398765    -1.57909
------------------------------------------------------------------------------
```

2.3) คำนวณค่า OR เพื่ออธิบายความสัมพันธ์ระหว่างพื้นที่ที่ศึกษากับการตายของทารก ใน 28 วันหลังคลอด ดังนี้

จาก Logistic Model:

Logit P(X) = −1.989 − 0.399AREA + 0.890MALPRES + 2.362A_MAL

กรณี MALPRES = 0

$OR_{(AREA1,0)}$ = Exp{[−1.989−0.399(1)+0.890(0)+2.362(1)(0)]
−[−1.989−0.399(0)+0.890(0)+2.362(0)(0)]}
= Exp(−0.399)
= .....

กรณี MALPRES = 1

$OR_{(AREA1,0)}$ = Exp{[−1.989−0.399(1)+0.890(1)+2.362(1)(1)]
−[−1.989−0.399(0)+0.890(1)+2.362(0)(1)]}
= Exp(−0.399 + 2.362)
= Exp(1.963)
= .....

2.4) จงเปรียบเทียบค่า OR ที่ได้นี้ กับที่ได้จากการวิเคราะห์โดย Stratified analysis ตามที่แสดงในข้อ 3.9.2.1 ของแบบฝึกหัดที่ 1 พร้อมให้ความเห็น

2.5) จงเปรียบเทียบค่า OR ที่ได้นี้ กับที่ได้จากการวิเคราะห์โดย STATA ซึ่งเป็นตัวอย่างของผลจากการวิเคราะห์โดยโปรแกรมสถิติทั่วไปที่ค่า OR เป็นค่า Exponential ของค่า Coefficient (ดู Output ในข้อ 2.3) พร้อมให้ความเห็น

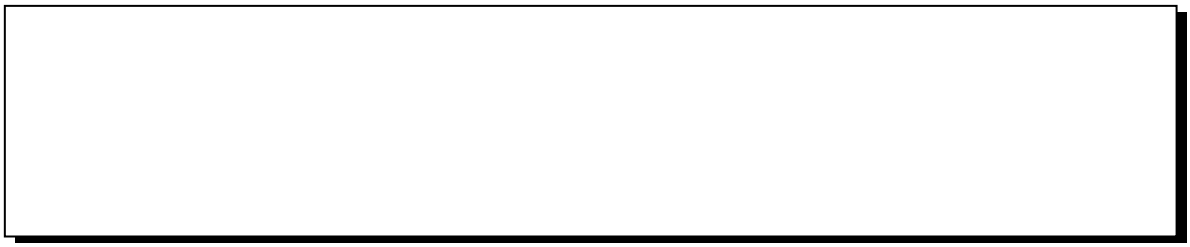
```
. logistic  dead area malpres a_mal

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(3)      =      52.05
                                                  Prob > chi2     =     0.0000
Log likelihood = -162.10316                       Pseudo R2       =     0.1383

------------------------------------------------------------------------------
        dead | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |   .6711146   .2168253     -1.234   0.217       .3562775    1.264168
     malpres |   2.435897   2.053089      1.056   0.291       .4669028    12.70842
       a_mal |   10.61667   10.34062      2.425   0.015       1.573689     71.6238
------------------------------------------------------------------------------
```

3. สมมติว่า จากการวิเคราะห์ พบว่าใน Model จำเป็นต้องใส่ Interaction term AREA*MALPRESS*MAGE เนื่องจากมีผลที่มีนัยสำคัญต่อ Model

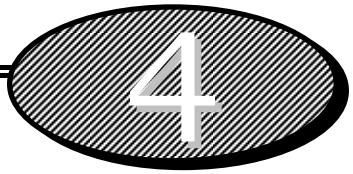
3.1) Term นี้เรียกว่า (กาเครื่องหมายหน้าข้อที่ถูก)

- ☐ Main effect
- ☐ Second order term
- ☐ Third order term

3.2) จากหลักการเขียน Model แบบ Hierarchical Well-formatted Model จงเขียน Logistic Regression Model ในรูปของ Logit transformation



การคำนวณช่วงความเชื่อมั่นของ Odds Ratio ที่ได้จาก Logistic Regression Model



วัตถุประสงค์ : เพื่อให้ผู้เรียนสามารถ

1. อธิบายแนวคิดพื้นฐานการคำนวณค่าช่วงความเชื่อมั่นของ Odds Ratio ได้

2. คำนวณค่าช่วงความเชื่อมั่น ของค่า Odds Ratio ทั้งในกรณีที่มีหรือไม่มี Interaction ได้

เนื้อหา :

1. แนวทางการนำเสนอข้อมูล จากผลการวิเคราะห์โดยใช้ Logistic Regression

2. การประมาณค่า ช่วงความเชื่อมั่นกรณี ไม่มี Interaction term

3. การประมาณค่าช่วงความเชื่อมั่น กรณีมี Interaction term

กิจกรรม :

1. ฟังบรรยายประกอบแผ่นใส พร้อมบันทึกเนื้อหาสำคัญลงในชุดการเรียนการสอน บทที่1

2. ทำแบบฝึกหัด

3. อภิปรายและสรุปเนื้อหา พร้อมเขียนสรุปท้ายบทลงกรอบว่างที่ให้ไว้ท้ายบท

สิ่งที่นำเสนอ	คำอธิบาย

ผลลัพธ์สุดท้ายของการวิเคราะห์โดยใช้ Logistic Regression แม้จะออกมาในรูป Model ทางคณิตศาสตร์ แต่การนำเสนอข้อมูลขึ้นอยู่กับวัตถุประสงค์หลักของการวิจัย วัตถุประสงค์หลักของการวิจัยที่ต้องใช้ Logistic Regression วิเคราะห์ข้อมูล มี 2 ประเภท ได้แก่ เพื่อได้ Best Predicted Model สำหรับนำไปใช้คาดคะเนความเสี่ยงต่อการเกิดเหตุการณ์ที่ศึกษาเมื่อระบุค่าตัวแปรต้น (ต้องเป็นข้อมูลจาก Cohort study) และอีกประเภทหนึ่ง คือ เพื่อหาปัจจัยเสี่ยง (Risk assessment) ซึ่งเป็นกรณีที่กล่าวถึงในที่นี้ (หนังสือที่ดีเกี่ยวกับแนวทางการนำเสนอผลการศึกษาวิจัยคือ Lang and Secic, 1997)

เนื่องจาก Risk assessment เป็นเป้าหมายพื้นฐานของการวิเคราะห์โดยใช้ Logistic Regression ที่พบได้บ่อย และสอดคล้องกับวิธีการศึกษาที่ไม่จำเพาะเพียง Cohort study แต่ยังใช้ได้กับ Case-control และ cross-sectional study ด้วย แนวทางการนำเสนอข้อมูลในกรณีดังกล่าวนี้ อันดับแรกอาจเสนอ Model ที่ได้ ทั้งนี้เพื่อให้ผู้อ่านได้ทราบว่าเราควบคุมตัวแปรอะไรบ้าง แต่ที่ต้องนำเสนอสำหรับตัวแปรต้นแต่ละตัวแปรคือค่า OR ค่า 95%CI และค่า p-value เนื่องจากการศึกษาหนึ่งมักมีหลายตัวแปรต้น จึงควรนำเสนอเป็นตาราง

ในตาราง ควรนำเสนอค่าสัดส่วน (%) ค่า Crude OR และ Adjusted OR สำหรับค่าช่วงความเชื่อมั่น (Confidence Interval) ซึ่งนิยม 95%CI และ p-value นั้น ให้แสดงเฉพาะของ Adjusted OR

รูปที่ 4.1

แนวทางการนำเสนอข้อมูลจากการวิเคราะห์

โดยใช้ Logistic Regression

(ตัวอย่าง Case-control study กรณีมีวัตถุประสงค์เพื่อหาปัจจัยเสี่ยง)

นำเสนอข้อมูล Measure of association (มีใช่ค่า Coefficient) ดังตารางหุ่นต่อไปนี้

ปัจจัย	Case (n=....)	Control (n=....)	OR (95%CI)	
			(Unadjusted)	(Adjusted)
1.	%	%		
2.	%	%		
3.	%	%		
...				

รูปที่ 4.2

ตัวอย่าง

การนำเสนอข้อมูลจากการวิเคราะห์โดยใช้ Logistic Regression

"...สมการการถดถอยลอจีสติค แสดงความสัมพันธ์ระหว่างการสูบบุหรี่ กับ การป่วย ด้วยโรคหัวใจโคโรนารี่ โดยควบคุมผลของอายุ และ เพศ เป็น ดังนี้

Logit P(X) = 1.421 + 1.609SMK + 0.095SEX + 0.301AGE

ผู้สูบบุหรี่ มีความเสี่ยงต่อการป่วยด้วยโรคหัวใจโคโรนารี่เป็น 5 เท่าของผู้ไม่สูบบุหรี่(95%CLOR: 2.1 ถึง 11.7) พบว่า เพศ และ อายุ มีความสัมพันธ์กับ การป่วยด้วยโรคหัวใจโคโรนารี่ โดย ... (ตารางที่ 1)"

ตารางที่ 1 Adjusted Odds Ratio แสดงความสัมพันธ์ระหว่างการสูบบุหรี่ กับ การป่วย ด้วยโรคหัวใจโคโรนารี่

ปัจจัย	Case (n = 150)	Control (n = 150)	OR(95%CI) (Unadjusted)	(Adjusted)
1. การสูบบุหรี่				
- สูบ	80.7%	30.0%	9.7(5.5 ถึง 17.3)	5(2.1 ถึง 11.7)
- ไม่สูบ	19.3%	70.0%		
2. เพศ				
- ชาย	73.3%	63.3%	1.6(1.0 ถึง 2.7)	1.1(0.8 ถึง 3.7)
- หญิง	36.7%	33.7%		
...	...	...	...	...

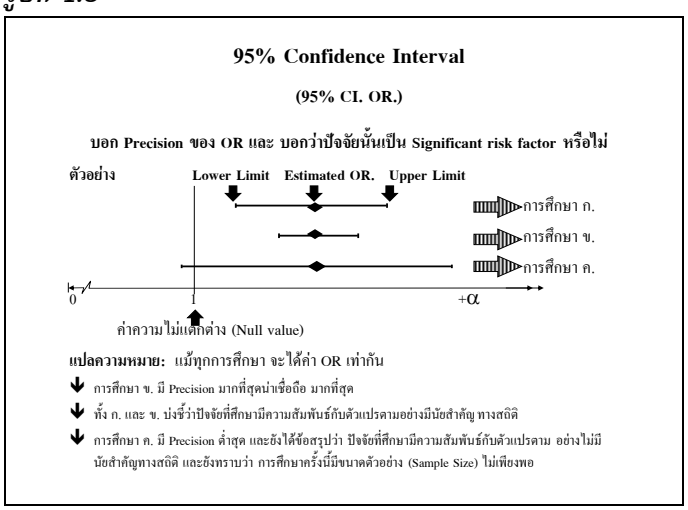
รูปแบบตารางคอลัมน์สำหรับแสดงค่าสัดส่วนอาจแตกต่างกันไปตาม Study design ที่เป็นตัวอย่างในที่นี้เป็น Case-control study ตารางสำหรับการศึกษาแบบอื่นโปรดดูใน แบบฝึกหัด และ Lang and Secic (1997)

ดังนั้นค่า 95%CI ของ OR ควรมีการคำนวณและนำเสนอคู่กับค่า OR ทุกครั้ง ค่า 95% CI ของ OR ใดๆ ประกอบด้วยค่าขีดจำกัดล่าง (Lower Limit) และขีดจำกัดบน (Upper Limit) ถ้าค่าทั้งสองห่างกันมาก หมายถึงช่วงความเชื่อมั่นกว้าง จะบอกได้ว่าค่า OR ที่ได้ไม่กระชับ (Precision ต่ำ) ซึ่งไม่ดีเท่าช่วงความเชื่อมั่นที่แคบ (Precision สูง)

ตัวอย่างเทียบเคียงความกว้างของช่วงความเชื่อมั่นกับความน่าเชื่อถือ เช่น บอกว่าผู้เขียนมีอายุระหว่าง 30-35 ปี กับอีกคนหนึ่งบอกว่าผู้เขียนมีอายุระหว่าง 2-80 ปี ค่าแรกย่อมให้ความหมายมากกว่า (รู้ว่าเป็นวัยกลางคน) ส่วนค่าหลังนั้น แม้จะไม่มีโอกาสผิดพลาดเลย แต่ผู้อ่านไม่สามารถคาดเดาได้เลยว่ายังเป็นทารก หรือชราภาพแล้ว จึงเป็นคำตอบที่หาความรู้ไม่ได้

นอกจากนั้นช่วงความเชื่อมั่นยังให้ข้อสรุปเหมือนกันกับการทดสอบสมมติฐาน (Hypothesis testing) อีกด้วย กล่าวคือ ถ้าหากช่วงของความเชื่อมั่น คร่อมค่า 1 ซึ่งเป็นค่าความไม่แตกต่าง หรือ Null value ของมาตรวัดระดับความเสี่ยง จะสามารถสรุปได้ว่า ตัวแปรนั้นสัมพันธ์กับตัวแปรตามอย่างไม่มีนัยสำคัญทางสถิติ หรือ p-value >0.05 ได้ แต่ถ้าช่วงความเชื่อมั่นไม่คร่อมค่า 1 แสดงว่าตัวแปรนั้นสัมพันธ์กับตัวแปรตามอย่างมีนัยสำคัญทางสถิติ (p-value <0.05) ผู้อ่านศึกษารายละเอียดใน Guyatt et al.(1995)

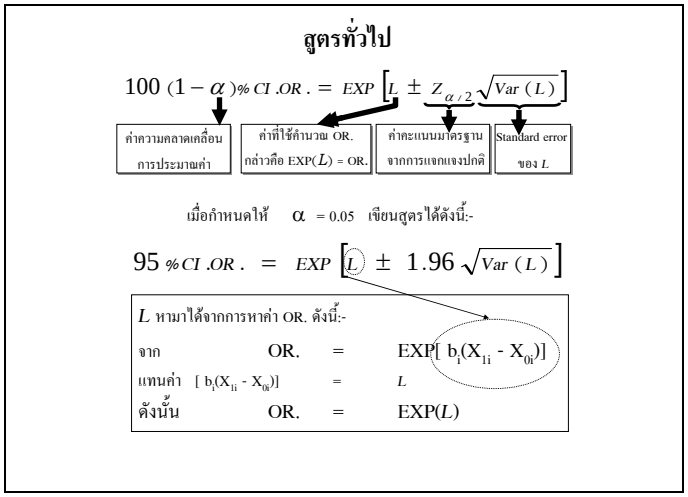
รูปที่ 4.3



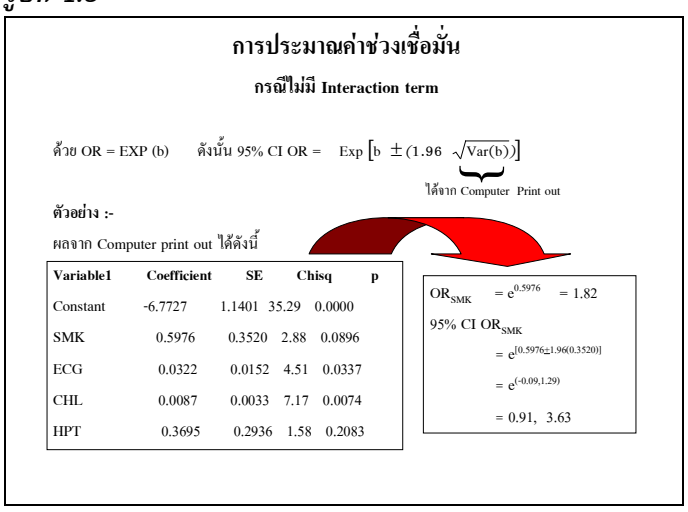
ค่าช่วงความเชื่อมั่นในระดับ 95% ของ OR หาได้จากสูตรทั่วไปในการหาช่วงความเชื่อมั่น เมื่อกำหนดให้โอกาสผิดพลาดของการประมาณค่า(α) เท่ากับ 0.05 ค่าระดับความเชื่อมั่นจึงเท่ากับ 95% ตรงกับค่ามาตรฐานการแจกแจงแบบปกติ $Z_{0.05/2}$ เท่ากับ 1.96 ดังนั้นจึงต้องทราบค่า OR ซึ่งได้กล่าวถึงวิธีการคำนวณแล้วในชุดการเรียนที่ 3 และต้องทราบค่า SE (ย่อมาจาก Standard Error) ของ OR ได้มาจากค่า square root ของ variance ของ OR หรือเขียนแทนด้วย $\sqrt{\text{Var (OR)}}$

กรณี Logistic regression model ค่า OR หาจากสูตรทั่วไป OR = EXP[$\sum b_i(X_1-X_0)$] เพื่อง่ายจึงให้ $L = \sum b_i(X_1-X_0)$ ดังนั้น OR = EXP(L) ดังนั้นค่า SE ของ OR = $\sqrt{Var(L)}$ และจะแตกต่างกันไปตามค่า L ที่ได้

รูปที่ 4.4



รูปที่ 4.5



ตามที่กล่าวข้างต้น การคำนวณค่า OR มีวิธีการที่แตกต่างกันขึ้นอยู่กันว่า Model นั้นมี Interaction term หรือไม่ การประมาณค่าช่วงความเชื่อมั่นก็เช่นกัน

ถ้าไม่มี Interaction term ค่า Variance ของ L ได้จากผลของคอมพิวเตอร์ที่ print out ออกมาโดยตรง (คือค่า SE ของแต่ละค่าสัมประสิทธิ์) ซึ่งใช้แทนค่า $\sqrt{\text{Var (L)}}$ ในสูตรได้โดยตรง

แต่ถ้ามี Interaction term การคำนวณค่า Variance ของ L จะต้องให้เครื่องคำนวณ Variance-covariance matrix จาก Final model ก่อน แล้วนำค่าเหล่านั้นมาคำนวณด้วยมือเพิ่มเติมอีก ตามวิธีการที่จะได้กล่าวต่อไป จากนั้นจึงนำไปแทนค่าในสูตรหาช่วงเชื่อมั่นต่อไป กระบวนการดังกล่าวมีความยุ่งยากแต่ควรค่าแก่การทำความเข้าใจ ในชีวิตจริงเราใช้คอมพิวเตอร์คำนวณให้ ซึ่งได้แสดงไว้ในแบบฝึกหัด (ศึกษารายละเอียดใน StataCorp.,

1999).

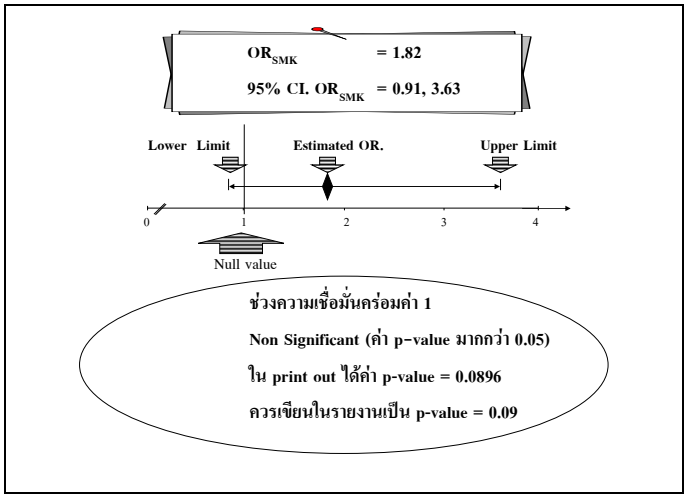
กรณีไม่มี Interaction ไม่ยุ่งยากต่อการคำนวณด้วยมือ และคอมพิวเตอร์ให้ไว้เสมอ ตัวอย่างนี้แสดงการคำนวณ จากการศึกษาปัจจัยที่มีผลต่อการป่วยด้วยโรคหัวใจโคโรนารี่ (CHD) โดยปัจจัยที่ศึกษาได้แก่ การสูบบุหรี่ (SMK) ผลการตรวจคลื่นไฟฟ้าหัวใจ (ECG) ปริมาณโมเลสเตอรอล (CHL) และการมีความดันโลหิตสูง (HPT) ผลจากคอมพิวเตอร์ให้ค่า SE ในแต่ละค่าของค่าสัมประสิทธิ์ (นอกจากนั้นยังให้ค่าการทดสอบความมีนัยสำคัญของแต่ละตัวแปรโดยให้ค่า Chi-square พร้อมด้วย ค่า p-value ซึ่งการทดสอบดังกล่าวอธิบายรายละเอียดในบทที่ 5)

จากผลดังกล่าว เราสามารถคำนวณค่า OR และช่วงความเชื่อมั่นได้โดยแทนค่าลงในสูตรทั่วไปตามที่กล่าวแล้วข้างต้น เช่น ค่า OR ของ SMK ได้เท่ากับ 1.82 และค่า 95% CI ได้เท่ากับ 0.91 ถึง 3.63

โปรแกรมคอมพิวเตอร์มักให้ทางเลือกให้ผู้วิเคราะห์ว่าจะให้ผลเป็นค่าสัมประสิทธิ์หรือ OR ดังนั้น ในชีวิตจริงจึงมักใช้ผลจากคอมพิวเตอร์โดยตรง ไม่ต้องคำนวณด้วยมือ ซึ่งได้แสดงไว้ในแบบฝึกหัด (ศึกษารายละเอียดใน StataCorp., 1999).

จากผลที่ได้ แปลความหมายได้ว่า ***"เมื่อควบคุมผลจากผลการตรวจคลื่นไฟฟ้าหัวใจ ค่าปริมาณโมเลสเตอรอล และการมีความดันโลหิตสูงแล้ว คนที่สูบบุหรี่ มีความเสี่ยงต่อการป่วยด้วยโรคหัวใจโคโรนารี่ สูงเป็น 1.82 เท่าของคนที่ไม่สูบบุหรี่ และเชื่อมั่น 95% ว่าค่าดังกล่าวในประชากรทั่วไป จะอยู่ระหว่าง 0.91 ถึง 3.63 อย่างไรก็ตามค่าสัมพันธ์ดังกล่าว ไม่มีนัยสำคัญทางสถิติ (p-value = 0.0896)"*** โปรดสังเกตว่า แม้ช่วงความเชื่อมั่นคร่อมค่าแต่มีความโน้มเอียงเป็น

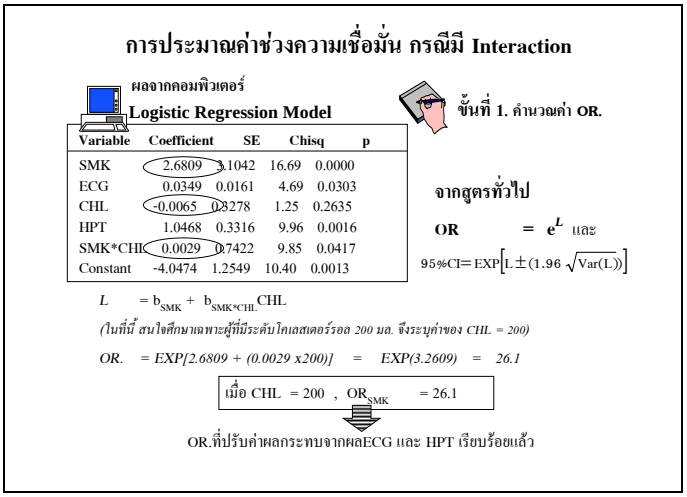
รูปที่ 4.6



ปัจจัยเสี่ยงมากกว่าเป็นปัจจัยป้องกัน

ถ้าการศึกษาในลักษณะเดียวกันนี้แล้วเพิ่มจำนวนขนาดตัวอย่างให้มากขึ้น จะยังผลให้ช่วงความเชื่อมั่นแคบเข้า (มี precision สูงขึ้น) โอกาสที่จะไม่คร่อมค่า 1 มีมากขึ้น นั่นคือจะมีโอกาสได้ผลการทดสอบเป็น "มีนัยสำคัญ (Significance)" หรือ p-value ที่ได้จะมีค่าน้อยกว่า 0.05 ค่าระดับความสัมพันธ์ และช่วงเชื่อมั่นให้ความหมายได้มากกว่า p-value เราจึงควรให้ความสำคัญต่อค่า OR และ 95%CI มากกว่าที่จะสรุปว่ามีหรือไม่มีนัยสำคัญทางสถิติเพียงอย่างเดียว

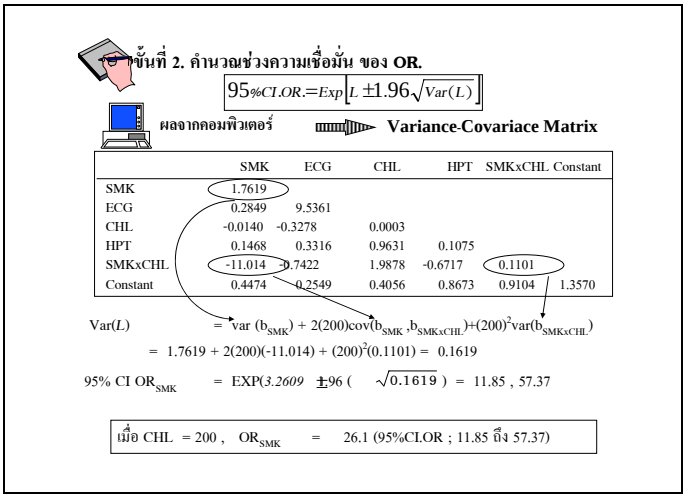
รูปที่ 4.7



ในกรณีที่มี Interaction ค่าของ Var(L) จะใช้เฉพาะค่า SE ที่ได้จากคอมพิวเตอร์**ไม่ได้** เนื่องจากมีการคำนวณค่าสัมประสิทธิ์โดยใช้ค่าจากตัวแปรมากกว่าสองตัวแปร คือตัวแปรที่ Interaction กัน จึงยังผลให้ค่าสัมประสิทธิ์ที่ได้นั้นมีความสัมพันธ์กับค่าสัมประสิทธิ์ตัวอื่น ๆ ดังนั้นค่า Var(L) จึงต้องคำนวณจากทั้งค่า variance และค่า covariance ของค่าสัมประสิทธิ์ด้วย

จากตัวอย่างที่กล่าวข้างต้น เมื่อ Fit Model ใหม่พบว่ามี Interaction ระหว่าง SMK กับ CHL ขั้นแรกให้คำนวณค่า OR ของ SMK จำแนกตามระดับของ CHL แต่เนื่องจาก CHL เป็นข้อมูลต่อเนื่อง (ปริมาณโฆเลสเตอรอล) จึงต้องกำหนดระดับที่เราสนใจ สมมติให้เป็น 200 220 และ 240 เป็นต้น ดังนั้น เมื่อ CHL เท่ากับ 200 ได้ค่า OR_{SMK} เท่ากับ 26.1 เป็นต้น จากนั้นคำนวณค่าช่วงความเชื่อมั่นโดยต้องคำนวณค่า Var(L) ก่อน

รูปที่ 4.8



ในกรณีนี้ค่า L คือ $b_{SMK} + b_{SMK*CHL}$ ดังนั้น ค่า Var (L) เท่ากับ $var(b_{SMK}) + 2(200)cov(b_{SMK},b_{SMK*CHL}) + (200)^2$ var ($b_{SMK*CHL}$) ทั้งสองค่านี้ ได้จากคอมพิวเตอร์ ค่าที่ได้ใน Computer print out เรียกว่า Variance Covariance Matrix นำผลที่ได้ไปคำนวณค่า Var (L) ตามที่กล่าวข้างต้น จากนั้นคำนวณช่วงความเชื่อมั่นของค่า OR ของการสูบบุหรี่เมื่อ CHL = 200

รูปที่ 4.9

ตัวอย่างสูตรสำหรับคำนวณค่า Var(L) ในแต่ละ Model

กรณี X_1เป็น Dichotomous (1,0)

Model: Logit P(X) = $a + b_1X_1 + b_2X_2 + b_3X_1X_2$

เมื่อ $L = b_1X_1 + b_3(X_1X_2)$

$$Var(L) = Var(b_1) + (X_2)^2var(b_3) + 2(X_2)cov(b_1,b_3)$$

Model: Logit P(X) = $a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_1X_2 + b_5X_1X_3$

เมื่อ $L = b_1X_1 + b_4X_1X_2 + b_5X_1X_3$

$$Var(L) = Var(b_1) + (X_2)^2var(b_4) + (X_3)^2var(b_5) + 2X_2cov(b_1,b_4) + 2X_3cov(b_1,b_5) + 2X_2X_3Cov(b_4,b_5)$$

(ศึกษารายละเอียดใน Kleinbaum, 1994. หน้า 141 - 144)

สูตรสำหรับคำนวณค่า Var(L) เพื่อประกอบการคำนวณค่าช่วงความเชื่อมั่น กรณีมี Interaction นั่น จะแตกต่างกันไปตามลักษณะของ Model ที่ได้ ทั้งนี้ต้องอาศัยพื้นฐานความเข้าใจทางพีชคณิตในการได้มาซึ่งสูตรดังกล่าว ตัวอย่างสูตรสำหรับคำนวณค่า Var(L) สำหรับแต่ละลักษณะของ Model (ในที่นี้แสดงเพียงค่า L) ที่แสดงในที่นี้เพียงบางลักษณะจะเห็นว่า มีความยุ่งยากเพิ่มมากขึ้นตามจำนวน Interaction term ที่มีใน Model ยิ่งถ้า Interaction term สูงกว่าระดับที่สอง (Second order term) การคำนวณ Var(L) ยิ่งยุ่งยากซับซ้อนมากขึ้นเป็นลำดับ นอกจากนี้แม้คำนวณออกมาได้ก็แปลความหมายได้ลำบาก เข้าใจได้ยาก ผู้วิจัยจึงควรหลีกเลี่ยงการคงไว้ซึ่ง Interaction term ใน Model เว้นแต่เป็นกรณีที่จำเป็นจริง ๆ เช่นเป็นเป้าหมายหลักของการวิจัยนั้นที่จะอธิบายความสัมพันธ์ของสองตัวแปร ตามตัวแปรที่เป็น Effect Modifier ที่สำคัญมาก เป็นต้น (ศึกษารายละเอียดใน Kleinbaum, 1994. หน้า 141 - 144)

สรุปท้ายบท :

เอกสารอ้างอิงประจำบทที่ 4

Guyatt, G., Jaeschke, R., Heddle, N., Cook, D., Shannon, H., and Walter S. (1995). Interpreting study results: confidence intervals. *Canadian Medical Association Journal.* 152:169-173.

Kleinbaum, D.G. (1994). *Logistic Regression: A self-learning text*. New York Springer-Verlag.

Lang, TA., Secic, M. (1997). *How to report statistics in medicine: annotated guidelines for authors*, editors, and reviewers. Philadelphia: American College of Physician.

StataCorp. (1999). *Stata statistical software: Release 6.0*. College Station. TX: Stata Corporation.

แบบฝึกหัดที่ 4

1. จากข้อมูลการศึกษาที่ใช้สำหรับเป็นแบบฝึกหัด ตั้งแต่บทที่ 1 เป็นต้นมา (แฟ้มข้อมูล LOGISTIC.DAT) จงคำนวณช่วงความเชื่อมั่นของค่า OR. ที่ระดับความเชื่อมั่น 95% พร้อมทั้งแปลความหมาย ในกรณีต่อไปนี้

1.1) จงแสดงความสัมพันธ์ระหว่าง ***การเยี่ยมก่อนและหลังคลอดโดย อสม และการรอดชีพของทารกในระยะเวลา 28 วัน หลังคลอด*** ในรูปตาราง 2 x 2 ที่เว้นว่างให้เติมข้อมูลพร้อมตอบคำถามต่อไปนี้

1.1.1) โปรดใส่ค่าของ a b c และ d พร้อมแสดงการคำนวณค่า OR. และช่วงความเชื่อมั่น ที่ระดับ 95% โดยแทนค่าในสูตรและคำนวณด้วยมือ

		การตาย 1	
พื้นที่ทดลอง	1	a	b
พื้นที่ควบคุม	0	c	d

OR. = ad /bc

OR. = ...

$$(1-\alpha)100\%CI.OR. = OR.exp\left[\pm Z_{\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right]$$

1.1.2) จงใช้ STATA วิเคราะห์ข้อมูลในรูปตาราง 2 × 2 บันทึกผลที่ได้ พร้อมกับแปลความหมาย (***คำชี้แนะ*** *ใช้คำสั่งในกลุ่ม epitab คือคำสั่ง cc*)

1.2.3 จงสรุปผลการวิเคราะห์ โดยเขียนเป็นข้อความเสมือนกับที่จะให้ปรากฏในบทคัดย่อของงานวิจัยชิ้นนี้

1.2) จงเขียน Logistic Regression Model ที่ใช้สำหรับวิเคราะห์ความสัมพันธ์ ตามข้อ 1.1 โดยเขียนในรูป Logit transformation (ให้เขียนเป็นสัญลักษณ์ คือติดค่า b และชื่อตัวแปรไว้)

1.3) จง Fit Model ข้างต้นโดยใช้ STATA แล้วเขียน Logistic Regression Model โดยใช้ผลจากการ Fit Model แทนค่าสัมประสิทธิ์ (***คำชี้แนะ*** *ใช้คำสั่ง logit*)

1.4) จงคำนวณค่า OR. และช่วงความเชื่อมั่น ที่ระดับ 95% โดยใช้สูตรทั่วไปในการคำนวณจาก Model

1.5) จงคำนวณค่า OR. และช่วงความเชื่อมั่น ที่ระดับ 95% โดยใช้ STATA (***คำชี้แนะ*** *ใช้คำสั่ง logistic*)

1.6) เปรียบเทียบค่า OR และ 95%CI ที่ได้จากข้อ 1.1.1 ข้อ 1.1.2 ข้อ 1.4 และข้อ 1.5 พร้อมอภิปรายสิ่งที่พบ

2. จงคำนวณช่วงความเชื่อมั่นของค่า OR. ที่ใช้ในการอธิบายความสัมพันธ์ระหว่างอายุมารดา (MAGE) กับการตายของทารก (DEAD) เปรียบเทียบระหว่างมารดาอายุ 30 ปี กับ 20 ปี ตามขั้นตอนต่อไปนี้

2.1) ใช้ STATA ในการ Fit Model
จงเขียนคำสั่งที่ใช้ และบันทึกผลที่ได้

2.2) คำนวณค่า $OR_{MAGE30,20} = e^{L}$

ในที่นี้ L =.....

ดังนั้น OR. =.....

จากสูตรทั่วไปในการคำนวณค่า 95% CI

95% CI OR. = $\exp [L \pm 1.96\sqrt{\text{var}(L)}]$

เมื่อ $\text{Var}(L)$ = ...

ดังนั้น 95% CI OR. = ...
.....

แปลความหมาย .....

3. จากแบบฝึกหัดที่ 3 จงคำนวณช่วงความเชื่อมั่นของค่า OR. ที่ใช้ในการอธิบายความสัมพันธ์ระหว่างพื้นที่ที่ศึกษา (AREA) กับการตายของทารก (DEAD) เมื่อมี Interaction effect โดย ท่าการคลอดของทารก (MALPRES) เป็น Effect Modifier

3.1) ใช้ STATA ในการ Fit Model พร้อมบันทึกผลที่ได้

3.2) จงคำนวณค่า OR.

$OR_{(MALPRES=0)}$ =.....

$OR_{(MALPRES=1)}$ =.....

3.3) จงทำความเข้าใจการคำนวณช่วงเชื่อมั่น OR กรณีมี Interaction effect โดย***คำนวณด้วยมือ***ดังต่อไปนี้ (ผู้อ่านสามารถข้ามข้อ 3.3 ได้โดยไม่เสียความต่อเนื่อง)

3.3.1 หาค่า Variance Covariance Matrix โดยใช้ STATA คำสั่งที่ใช้ คือ

.logit dead area malpres a_m =>*เพื่อ Fit Model ก่อน (เมื่อ a_m คือ Interaction term ระหว่าง area กับ malpres)*

.matrix V = get(VCE) =>*สร้างตัวแปร V เก็บข้อมูล Matrix (โปรดสังเกต ตัวพิมพ์ใหญ่ หรือ เล็กมีผลต่างกันในSTATA)*

.matrix list V =>*ได้**Variance-covariance Matrix***

ดังผลต่อไปนี้

```
. logit  dead  area  malpres  a_m

Iteration 0:   log likelihood =  -188.1264
Iteration 1:   log likelihood = -170.29205
Iteration 2:   log likelihood = -162.58391
Iteration 3:   log likelihood = -162.10446
Iteration 4:   log likelihood = -162.10316

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(3)      =      52.05
                                                  Prob > chi2     =     0.0000
Log likelihood = -162.10316                       Pseudo R2       =     0.1383

------------------------------------------------------------------------------
        dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.3988154   .3230824     -1.234   0.217      -1.032045    .2344146
     malpres |   .8903152   .8428469      1.056   0.291      -.7616343    2.542265
         a_m |   2.362425   .9739987      2.425   0.015       .4534228    4.271427
       _cons |  -1.988928   .2091045     -9.512   0.000      -2.398765    -1.57909
------------------------------------------------------------------------------

. matrix  V =  get(VCE)

. matrix  list  V

symmetric V[4,4]
              area     malpres         a_m       _cons
   area   .10438226
malpres    .0437247   .71039085
    a_m  -.10438226  -.71039085   .94867342
  _cons   -.0437247   -.0437247    .0437247   .0437247
```

นี่คือ Variance-covariance Matrix

3.3.2คำนวณค่าช่วงความเชื่อมั่นของ OR.

ก. เมื่อ MALPRES = 0

95% CI. OR. = $\text{Exp}[-0.399 \pm 1.96\sqrt{0.1044}]$

= Exp(-1.032) ถึง Exp(0.234)

= 0.36 ถึง 1.26

ข. เมื่อ MALPRES = 1

$$\text{var}(L) = b_{AREA} + b_{A\_MAL}$$
$$= \text{var}(b_{AREA}) + 2\text{MALPRES}\ \text{cov}(b_{AREA}, b_{A\_MAL}) + (\text{MALPRES})^2 \text{var}(b_{A\_MAL})$$
$$= 0.1044 + [2(1)(-0.01044)] + [(1)^2(0.9487)]$$
$$= 0.8443$$

$$95\%\ \text{CI. OR.} = \text{Exp}[1.963 \pm 1.96\sqrt{0.8443}]$$
$$= \text{Exp}(0.162) \text{ ถึง } \text{Exp}(3.764)$$
$$= 1.18 \text{ ถึง } 43.12$$

3.4) จงศึกษาการใช้ STATA ช่วยคำนวณ OR และ 95%CI กรณีมี Interaction effect ตามกรณีข้างต้น พร้อมเปรียบเทียบผลที่ได้ตามข้อ 3.3 (*สามารถใช้คำสั่ง lincom หรือ effmod ก็ได้*)

คำสั่ง lincom ได้ผลดังนี้

```
. lincom  area

 ( 1)  area = 0.0

------------------------------------------------------------------------------
    dead | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     (1) |   .6711146   .2168253    -1.234   0.217       .3562775    1.264168
------------------------------------------------------------------------------


. lincom  area +  a_m

 ( 1)  area + a_m = 0.0

------------------------------------------------------------------------------
    dead | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     (1) |      7.125   6.546829      2.137   0.033       1.176673    43.14337
------------------------------------------------------------------------------
```

อีกทางเลือกหนึ่งคือคำสั่ง effmod ได้ผลดังนี้

```
. effmod  dead area, cov( malpres) int( a_m 0)

OR and 95% CI for a Logistic Regression Model with Interaction
------------------------------------------------------------------

 Disease:     dead
 Exposure:    area
 Confounders: malpres
 Interaction Terms and Stratum Values:
              a_m: 0
 Exposed-Unexposed= 1

   l= -.39881537
   Var(l)= .10438226

 Odds Ratio (95% CI) for dead vs. area:  0.671  (0.356, 1.264)
```

```
. effmod  dead area, cov( malpres) int( a_m 1)

OR and 95% CI for a Logistic Regression Model with Interaction
------------------------------------------------------------

 Disease:     dead
 Exposure:    area
 Confounders: malpres
 Interaction Terms and Stratum Values:
              a_m: 1
 Exposed-Unexposed= 1

   l= 1.9636097
   Var(l)= .84429116

 Odds Ratio (95% CI) for dead vs. area:  7.125  (1.177, 43.143)
```

3.5) จงแปลความหมายผลที่ได้

กลวิธีการสร้าง Logistic Regression Model

วัตถุประสงค์ : เพื่อให้ผู้เรียนสามารถ

1. อธิบายแนวคิดและวิธีการในการสร้างModelได้

2. สร้างModelได้อย่างถูกต้องเหมาะสม

เนื้อหา :

1. การสร้างModel (Model fitting Strategies)
    1.1) แนวคิดของการสร้างModel
    1.2) เป้าหมายของการสร้างModel
    1.3) ข้อพึงระวัง - Multicolinearity
        - Multiple testing
        - Outlier
        - Non-linear relationship

2. ขั้นตอนในการสร้างModel
    2.1) การระบุตัวแปรใน Initial Model
    2.2) การตัดตัวแปรต้นออกไปจาก Model

กิจกรรม :

1. ฟังบรรยายประกอบแผ่นใส พร้อมบันทึกเนื้อหาสำคัญลงในชุดการเรียนการสอน บทที่1

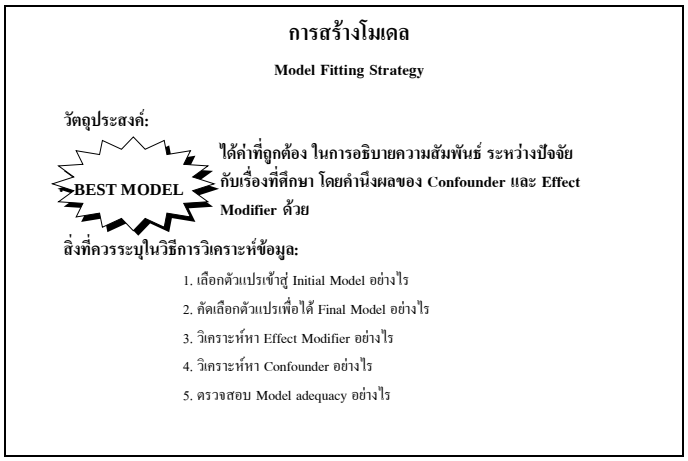
2. ทำแบบฝึกหัด

3. อภิปรายและสรุปเนื้อหา พร้อมเขียนสรุปท้ายบทลงกรอบว่างที่ให้ไว้ท้ายบท

สิ่งที่นำเสนอ	คำอธิบาย

1. การสร้าง Model

รูปที่ 5.1



จุดหมายสูงสุดของการใช้ Logistic Regression ในการวิเคราะห์ข้อมูล เพื่อให้ได้ Model ที่ดีที่สุด กล่าวคือ เป็น Model ที่ให้ค่าที่ถูกต้อง ที่จะนำไปใช้ในการอธิบายความสัมพันธ์ระหว่างปัจจัย (E) กับเหตุการณ์ที่ศึกษา (D) โดยคำนึงผลกระทบจาก ปัจจัยอื่นๆ (C) ทั้งที่เป็น Confounder หรือ Effect Modifier กล่าวอีกนัยหนึ่งคือเพื่อให้ได้ค่าสัมประสิทธิ์ที่ถูกต้อง หรือ OR ที่ถูกต้องนั่นเอง

ปัญหาที่พบบ่อยสิ่งแรกคือ ผู้วิจัยส่วนมากไม่ได้บันทึกหรือบอกวิธีการในการสร้างModel ในรายงานการวิจัย เพียงแต่เสนอผลสุดท้ายคือ Final Model อย่างเดียว ซึ่งทำให้ผู้อ่านไม่สามารถประเมินความถูกต้องเหมาะสมหรือความน่าเชื่อถือของModelได้

ดังนั้น ภายใต้หัวข้อวิธีการวิเคราะห์ข้อมูลของรายงานการวิจัย ควรบันทึกขั้นตอนที่ใช้ในการสร้างModelไว้ โดยควรระบุประเด็นต่อไปนี้

1. เลือกตัวแปรเข้าสู่โมเดลเริ่มต้นอย่างไร
2. คัดเลือกตัวแปรเพื่อได้โมเดลสุดท้ายอย่างไร
3. วิเคราะห์หา Effect Modifier อย่างไร
4. วิเคราะห์หา Confounder อย่างไร
5. ตรวจสอบความถูกต้องของโมเดลอย่างไร

1.1 เป้าหมายการสร้าง Model

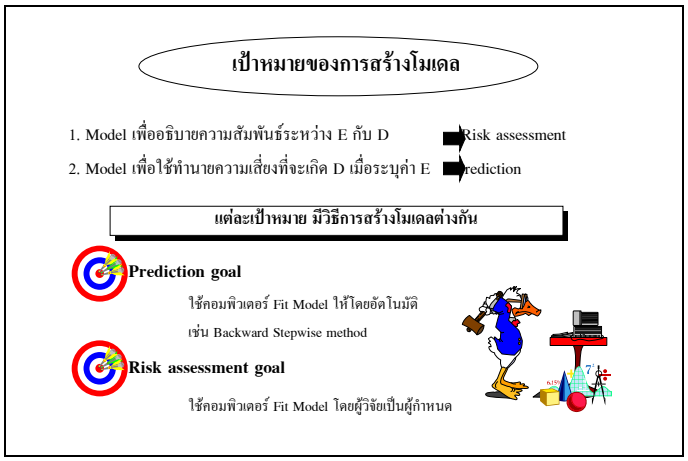
โดยทั่วไป เป้าหมายของการสร้าง Model มีอยู่ 2 อย่าง ได้แก่ สร้าง Model เพื่ออธิบายความสัมพันธ์ระหว่าง E กับ D เรียกเป้าหมายนี้ว่า Risk assessment goal และอีกเป้าหมายหนึ่งคือสร้าง Model เพื่อใช้ในการทำนายโอกาสที่จะเกิดเหตุการณ์ D เมื่อระบุค่าของ E

หรือแทนค่าด้วย Risk Profile ของบุคคลที่ใช้ทำนาย เรียกเป้าหมายนี้ว่า Prediction goal

เป้าหมายแรกสามารถได้จากรูปแบบการศึกษาใดก็ได้ ส่วนเป้าหมายหลัง ต้องมีรูปแบบการศึกษาเป็นแบบ Cohort study เท่านั้น

แต่ละเป้าหมายก็มีวิธีการสร้าง Model ต่างกัน ถ้าเป็น Prediction goal จะสามารถใช้วิธีการมาตรฐานในโปรแกรมคอมพิวเตอร์ได้ เช่น Backward Elimination หรือ Forward Inclusion หรือ Backward Stepwise Method เป็นต้น แต่ถ้าเป็น Risk Assessment Goal ซึ่งเป็นเป้าหมายที่กล่าวถึงในที่นี้เนื่องจากใช้มากในงานวิจัยทางการแพทย์และสาธารณสุขนั้น ไม่สามารถให้โปรแกรมคอมพิวเตอร์สร้าง Model ให้เองโดยอัตโนมัติได้ แต่ผู้วิจัยต้องทำเองโดยต้องใช้องค์ความรู้เกี่ยวกับเรื่องที่ศึกษาและวิจารณญาณในการเลือกตัวแปรเข้าในและออกจาก Model วิจารณญาณดังกล่าวเกี่ยวข้องกับการพิจารณาว่าตัวแปรใดมีความสำคัญทางการแพทย์และสาธารณสุข ตัวแปรใดที่พบว่าเป็นหรืออาจเป็น Confounder หรือ Effect Modifier เหล่านี้ไม่มีโปรแกรมคอมพิวเตอร์ใดที่สามารถกระทำได้

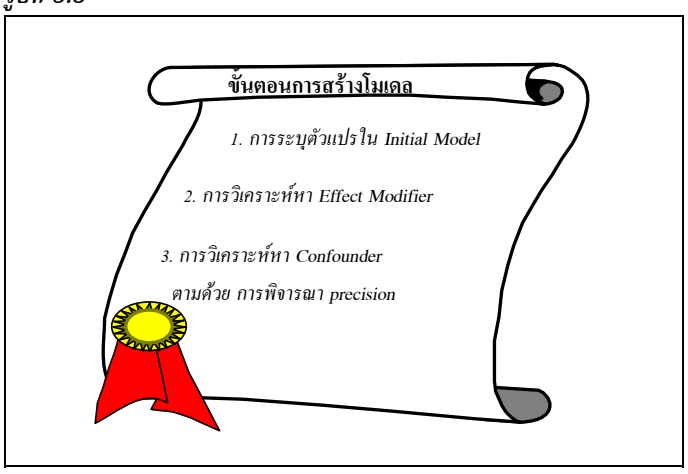
รูปที่ 5.2



1.2 ขั้นตอนการสร้าง Model

ขั้นตอนหลักในการสร้าง Model มี 3 ขั้นตอนตามลำดับ เริ่มจากการระบุตัวแปรสำหรับเป็นModelเริ่มต้น (Initial Model) จากนั้น วิเคราะห์ว่ามีตัวแปรใดเป็น Effect Modifier ถ้าตัวแปรใดเป็น Effect Modifier แล้ว ก็ไม่ต้องคำนึงว่าตัวแปรดังกล่าวจะเป็น confounder หรือไม่ คือแม้ผลของตัวแปรนั้นโดด ๆจะไม่มีนัยสำคัญแต่ตัวแปรนั้นมีผลร่วมกับตัวแปรอื่นอย่างมีนัยสำคัญก็ต้องให้อยู่ใน Model แต่ถ้าไม่พบว่ามีตัวแปรใดเป็น

รูปที่ 5.3



Effect Modifier ขั้นต่อไปจึงเป็นการวิเคราะห์ว่ามีตัวแปรใดเป็น Confounder หรือไม่ ตามด้วยการพิจารณา Precision ของสัมประสิทธิ์ที่คำนวณได้ โดยที่ Precision นี้สามารถพิจารณาจากความกว้างแคบของช่วงความเชื่อมั่น

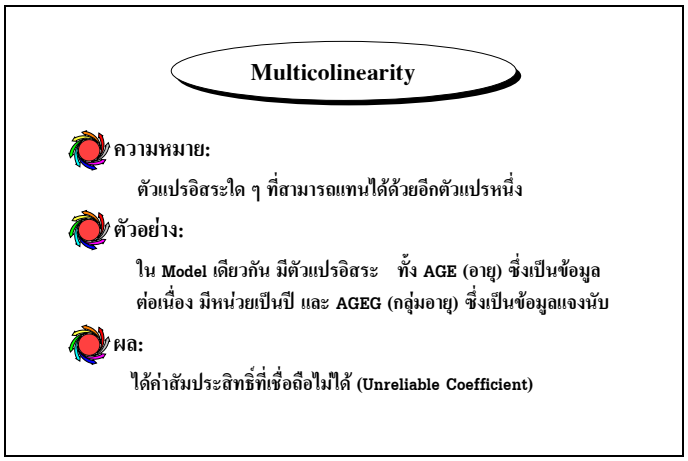
1.3 ข้อพึงระวังในการสร้าง Model

รูปที่ 5.4

อย่างไรก็ตาม การสร้างModelยังมีข้อพึงระวังบางประการได้แก่ เรื่อง Multicolinearity เรื่องMultiple testing เรื่อง Outlier และ เรื่อง Non-linear relationship ในกรณีที่ตัวแปรต้นเป็นตัวแปรอันดับ (ordinal) หรือต่อเนื่อง (continuous)

1.3.1 Multicolinearity

รูปที่ 5.5



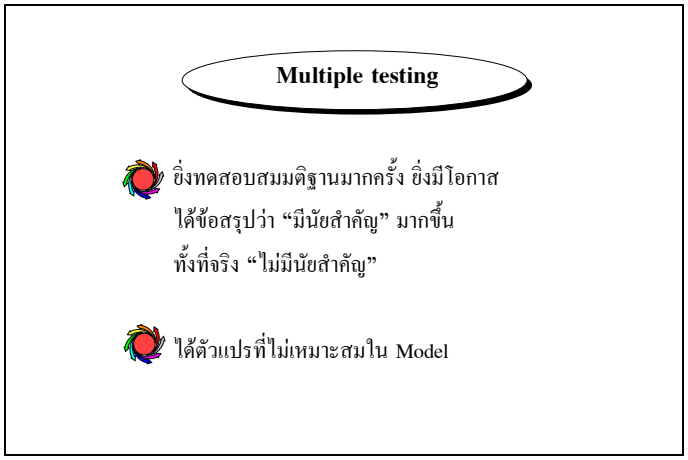
Multicolinearity จะเกิดขึ้นได้ในกระบวนการสร้าง Model เมื่อมีตัวแปรต้นใด ๆ ใน Model สามารถแทนด้วยอีกตัวแปรหนึ่งใน Model เดียวกัน เช่นมีทั้งตัวแปร AGE และ AGEG ใน Model เดียวกัน โดยที่ตัวแปร AGE เป็นข้อมูลต่อเนื่อง (อายุ) แต่ AGEG เป็นข้อมูลแจงนับ (กลุ่มอายุ) อีกตัวอย่างเช่น ในกลุ่มตัวอย่างที่ศึกษา ทุกคนที่สูบบุหรี่จะเป็นผู้ดื่มสุรา และไม่มีใครที่ไม่ดื่มสุราเป็นผู้สูบบุหรี่ ถ้าทั้งสองตัวแปรอยู่ใน Model เดียวกัน จะเกิด Multicolinearity ยังผลให้ ได้ค่าสัมประสิทธิ์ที่เชื่อถือไม่ได้ ดังนั้นจึงจำเป็นต้องมีการตรวจสอบ Multicolinearity ในการ Fit model เสมอ และพิจารณาตัดตัวแปรที่มีความสำคัญน้อยในกลุ่มที่ก่อให้เกิด Multicolinearity ออกไปและคงไว้ใน Model เพียงตัวแปรเดียว

ที่มีความสำคัญหรือที่สนใจ

โดยทั่วไป การตรวจสอบดังกล่าวสามารถใช้วิธีการทางสถิติทดสอบได้ จึงสามารถทำได้โดยใช้คอมพิวเตอร์ (STATA มีระบบตรวจสอบ Multicolinearity โดยอัตโนมัติ) แต่การตัดตัวแปรใดออกจาก Model ต้องเป็นวิจารณญาณของผู้วิจัยเท่านั้น

1.3.2 Multiple Testing

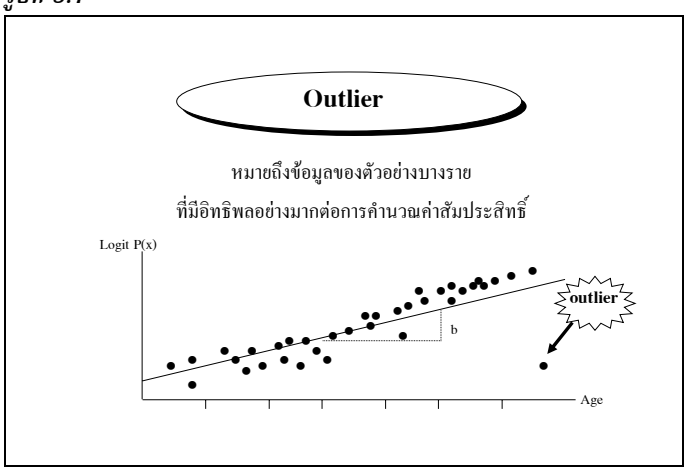
รูปที่ 5.6



ในกระบวนการสร้าง Model ซึ่งมีการคัดเลือกตัวแปรนั้น จะมีการทดสอบความมีนัยสำคัญทางสถิติหลายครั้ง ยิ่งจำนวนครั้งของการทดสอบยิ่งมาก โอกาสที่จะได้ผลว่าตัวแปรนั้นมีนัยสำคัญก็ยิ่งมีมาก ทั้งที่แท้จริงตัวแปรดังกล่าว ไม่มีนัยสำคัญทางสถิติ จะยังผลให้ได้ Model ที่กอปรขึ้นด้วยตัวแปรที่ไม่ได้มีผลต่อ Model ทำให้ได้ Model ที่ไม่เหมาะสม จึงควรหลีกเลี่ยงถ้าไม่จำเป็น แต่ถ้าหลีกเลี่ยงไม่ได้ ก็มีวิธีทางสถิติที่พอช่วยปรับผลดังกล่าวได้บ้าง แต่ก็ยังไม่มีวิธีการใดที่ดีที่สุด

1.3.3 Outlier

รูปที่ 5.7



Outlier หมายถึงข้อมูลของบุคคลใดหรือ Record ใดที่มีอิทธิพลอย่างมากต่อการคำนวณค่าสัมประสิทธิ์ กล่าวคือการมีหรือไม่มีข้อมูลของ record ดังกล่าว จะยังผลให้ค่าสัมประสิทธิ์เปลี่ยนแปลงไปอย่างมาก

Outliner มักเป็นค่าข้อมูลที่ต่างจากค่าอื่นมาก ๆ จึงควรตรวจสอบ outlier ในกระบวนการสร้าง Model ว่ามีหรือไม่ เพื่อนำมาประกอบการพิจารณาว่า จะคงไว้ หรือตัดออกจากการ Fit Model ต่อไป

การตรวจสอบ outlier มีหลายวิธี ที่นิยมทำเรียกว่า Measure of influence ผู้อ่านสามารถศึกษารายละเอียดใน Hosmer and Lemeshow (1989) ซึ่งสามารถทำได้โดยใช้โปรแกรม STATA (ดูรายระเอียดใน Stata Corp, 1999

Volume 2 H-O หน้า 218 ถึง 222)

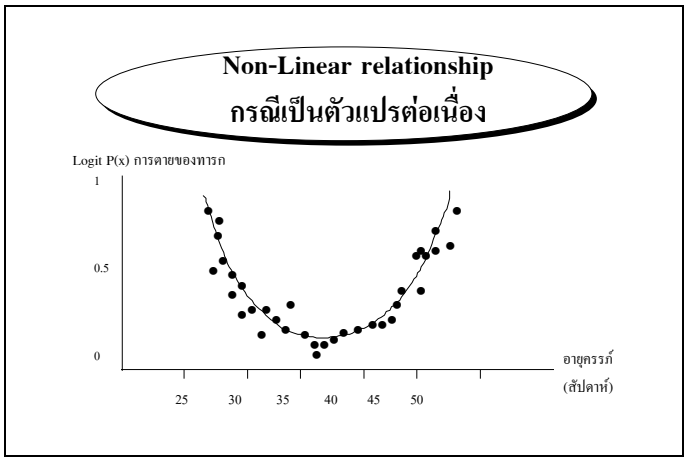
1.3.4 Non-linear Relationship

ตัวแปรต้นใดที่เป็นตัวแปรอันดับหรือตัวแปรต่อเนื่อง การนำเข้าไปใน Model มีทางเลือกสองทางคือนำเข้าโดยตรงโดยใส่ตัวแปรนั้นเข้าไปตัวแปรเดี่ยว ๆ และนำเข้าโดยสร้างตัวแปรใหม่ก่อน เฉพาะอย่างยิ่งนำมาจัดกลุ่มก่อน

กรณีตัวแปรอันดับ การนำเข้า Model โดยตรงจะดีกว่าการสร้าง dummy variable ก่อนเพราะจำนวนตัวแปรใน Model น้อยกว่า กรณีข้อมูลต่อเนื่องก็เช่นเดียวกัน การนำเข้า Model โดยตรงจะดีกว่าการนำมาจัดกลุ่มเป็นข้อมูลแจงนับก่อนนำเข้า Model เพราะทุกค่าข้อมูลได้รับการพิจารณา

อย่างไรก็ตาม ถ้าตัวแปรดังกล่าวไม่มีความสัมพันธ์เชิงเส้นตรงกับตัวแปรตาม การนำเข้า Model โดยตรงนั้นไม่ถูกต้อง จะต้องเข้าในรูป dummy variable กรณีที่เป็นตัวแปรอันดับหรืออาจจัดกลุ่มก่อนสร้าง dummy variable แล้วแต่กรณี ส่วนกรณีตัวแปรต่อเนื่องต้องจัดกลุ่มเป็นข้อมูลแจงนับก่อนแล้วสร้าง dummy variable (กรณีจัดเป็นสองกลุ่มจะได้ dichotomous variable ก็ไม่ต้องสร้าง dummy variable) การจัดกลุ่มต้องกระทำด้วยความระมัดระวัง แนวทางเบื้องต้นมีกล่าวใน Mazumdar and Glassman (2000)

รูปที่ 5.8



ตัวอย่าง Linear relationship มีลักษณะเช่นเดียวกันกับกราฟที่แสดงในเรื่อ outlier (รูปที่ 5.7) ส่วนกรณี Non-linear relationship มีตัวอย่างเช่นความสัมพันธ์ระหว่างอายุครรภ์มารดา (Gestational age) กับการตายของทารก เมื่อพิจารณาความเสี่ยงของการตายของทารก ในที่นี้คือ Logit P(**X**) พบว่า อายุครรภ์ต่ำกว่า 20 สัปดาห์ (Pre-term) กับอายุครรภ์สูงกว่า 36 สัปดาห์ (Post-

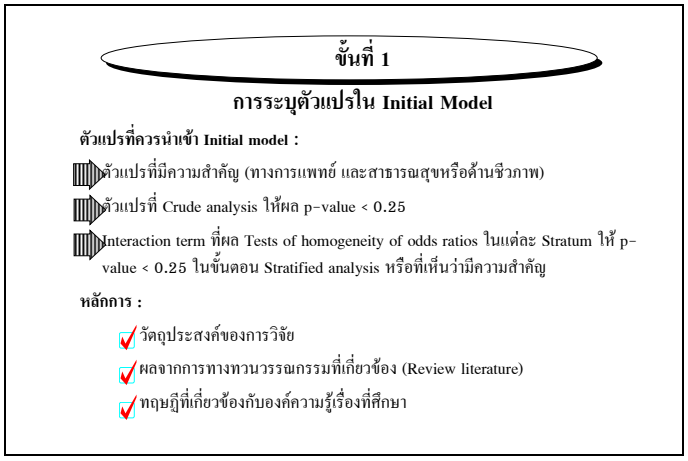
term) มีความเสี่ยงสูงมากกว่าที่อายุครรภ์ปกติ (20-36 สัปดาห์) แสดงถึงความสัมพันธ์ดังกล่าว ไม่เป็นเส้นตรงดังที่แสดงในรูปที่ 5.8 จะสังเกตได้ว่าใช้ Logit transformation คือ logit P(**X**) ซึ่งเท่ากับผลรวมเชิงเส้นของค่าสัมประสิทธิ์ของทุกตัวแปรต้น ตามที่กล่าวในบทต้นๆ เมื่อไม่เป็นเชิงเส้น Model ที่ได้จึงไม่ถูกต้อง

ดังนั้นกรณีเช่นนี้ ถ้านำข้อมูลตัวแปรต้นดังกล่าวมานำเสนอเป็นกราฟเส้นแสดงความสัมพันธ์กับ Logit P(**X**) ซึ่งเป็นค่าที่หาได้จากการแทนค่า Risk profile ใน Model จะทราบได้ว่า มีความสัมพันธ์เชิงเส้นต่อกันหรือไม่ ตัวอย่างที่แสดงในที่นี้ แสดงความสัมพันธ์ที่ไม่เป็นเชิงเส้นต่อกันอย่างชัดเจน ต้องพิจารณาจัดกลุ่มตัวแปรต่อเนื่องให้อยู่ในรูปตัวแปรแจงนับ ก่อนนำเข้า Model

การตรวจสอบ Linear relationship ควรกระทำในระยะ Crude analysis ตามที่กล่าวแล้วในบทที่ 1 เพื่อตัดสินใจว่าต้องมีการจัดกลุ่มตัวแปรก่อนเข้าใน Model หรือไม่ ตามรายเอียดที่จะกล่าวในหัวข้อต่อไป

2. การกำหนด Model เริ่มต้น

รูปที่ 5.9



ขั้นตอนแรกของการ Fit Model คือการสร้าง Initial Model เพื่อเริ่มต้นพิจารณาตัดตัวแปรที่ไม่มีผลออกไป ในขั้นนี้ ควรยึดวัตถุประสงค์ของการวิจัย ประกอบกับความสำคัญของตัวแปรโดยควรระบุตัวแปรที่มีความสำคัญ ไม่ว่าจะเป็นทางด้านการแพทย์และสาธารณสุข หรือ ทางด้านชีวภาพ โดยอาศัยองค์ความรู้จากการทบทวนวรรณกรรมที่เกี่ยวข้องหรือจากทฤษฎีที่เกี่ยวข้องกับองค์ความรู้ที่ศึกษา นอกจากนั้น ทุกตัวแปรที่ผลการวิเคราะห์ Crude analysis และ Stratified analysis ได้ค่า p-value < 0.25 (ตาม

รายละเอียดที่กล่าวในบทที่ 1) ควรนำเข้าใน Initial model ด้วย

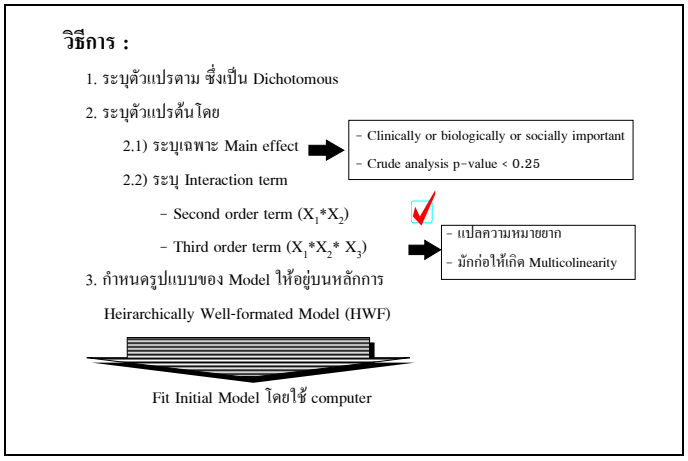
อย่างไรก็ตาม ถ้าพบว่าตัวแปรที่เลือกเข้าไปใน Model นั้นก่อให้เกิด Multicolinearity จะต้องตัดออกไปแม้จะมีความสำคัญทางด้านการแพทย์หรือการสาธารณสุขก็ตาม

นอกจากนั้นยังต้องพิจารณาเรื่อง Non-linear relationship ตามที่กล่าวข้างต้น (ดูแนวทางในแบบฝึกหัด)

วิธีการของการระบุตัวแปรสำหรับ Initial Model ทำได้โดยระบุตัวแปรตามซึ่งมีได้เพียงตัวแปรเดียว จากนั้นระบุตัวแปรต้น ซึ่งมีได้มากกว่าหนึ่งตัวแปร โดยระบุให้มากที่สุดเท่าที่จะมากได้แต่ต้องอยู่บนหลักการที่กล่าวข้างต้น และต้องไม่มากจนเกินไปเมื่อเปรียบเทียบกับขนาดตัวอย่าง ทั้งหมดนี้เรียกว่า Main effect Model

ขั้นต่อไป ระบุ Interaction term ซึ่งเป็น Product term ระหว่างสองตัวแปรต้น (เรียกว่า Second order term) หรือ สามตัวแปรต้น (เรียกว่า Third order term) ทั้งนี้ต้องยึดถือหลักการที่กล่าวข้างต้น กล่าวคือ ระบุเฉพาะที่มีองค์ความรู้เดิมบ่งชี้ว่าอาจมี Interaction effect และมีความสำคัญทางด้านการแพทย์และสาธารณสุข แต่ถ้ามากกว่า Second order term มักไม่นิยมเพราะจะทำให้ยากต่อการแปลความหมาย นอกจากนั้น การมีมากกว่า Second order term ยังมักก่อให้เกิด Multicolinearity ได้อีกด้วย

รูปที่ 5.10



รูปที่ 5.11

ขั้นต่อไป ผู้วิจัยต้องกำหนดรูปแบบ Model ให้อยู่ในรูป Heirarchically Well-formatted Model กล่าวคือ ในการกำหนดตัวแปรใน Model นั้น เมื่อระบุให้ Product term ใดอยู่ใน Model แล้ว ตัวแปรที่เป็น Main effect หรือที่เป็น Product term ที่อยู่ในลำดับที่ต่ำกว่าทุกตัวจะต้องอยู่ใน Model ด้วย ตัวอย่างที่แสดงในที่นี้ เพื่อชี้ให้เห็นความแตกต่างระหว่าง Model

ตัวอย่าง **Heirarchically Well-formated Model (HWF) :**

1. Logit P(**X**) = $a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_1X_3$

✓ เป็น HWF Model

2. Logit P(**X**) = $a + b_1X_1 + b_2X_2 + b_3X_3X_4$

✗ ไม่เป็น HWF Modelเพราะ ไม่มี Main effect ของ X_3 กับ X_4 ใน Model

3. Logit P(**X**) = $a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_3X_4 + b_6X_2X_3X_4$

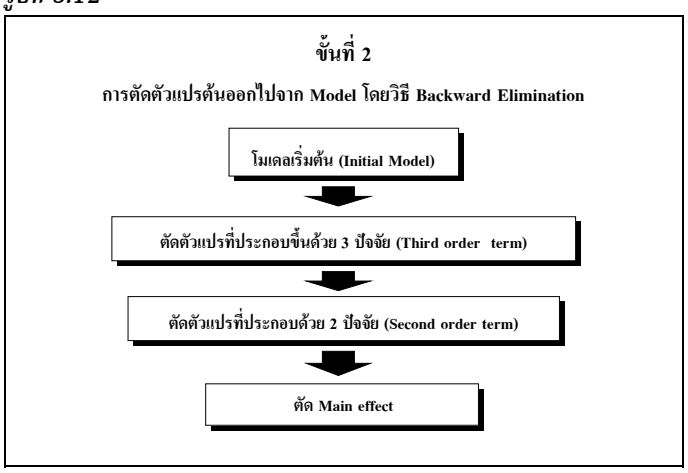
✗ ไม่เป็น HWF Modelเพราะ ไม่มี product term ต่อไปนี้คือ X_2X_3 และ X_2X_4

ที่เป็นและไม่เป็น Heirarchically Well-formatted Model รูปแบบนี้จำเป็นเพราะมีผลอย่างยิ่งต่อการคำนวณค่าสัมประสิทธิ์ซึ่งถือว่าเป็นหัวใจของการวิเคราะห์ข้อมูล

3. การคัดเลือกตัวแปรออกจาก Model

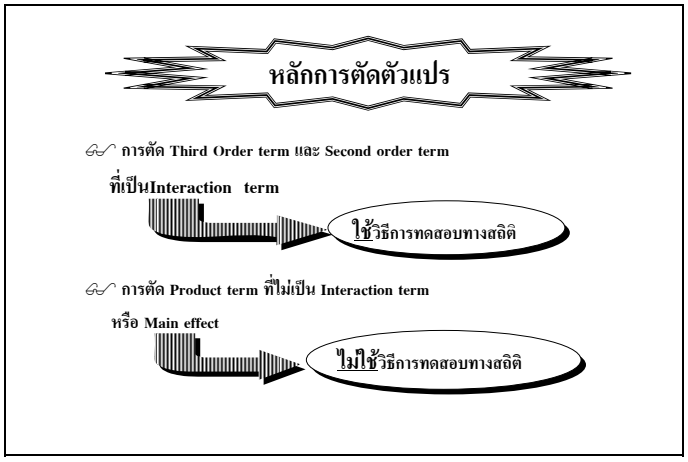
ขั้นตอนที่สองของการสร้าง Model คือการตัดตัวแปรต้นที่ไม่มีผลต่อตัวแปรตามออกจาก Model หรือเรียกว่า Backward elimination ทั้งนี้ต้องอยู่บนหลักการของ Heirarchical Principle ตัวอย่างเช่น เมื่อมี Model เริ่มต้นที่มีThird order term ตัวแปรที่จะต้องพิจารณาตัดออกจาก Model เป็นตัวแรกคือ Third order term จากนั้นพิจารณาตัดตัวแปร Second order term และสุดท้ายจึงพิจารณาตัด Main effect ที่ไม่มีผลต่อ Model ออกไป

รูปที่ 5.12



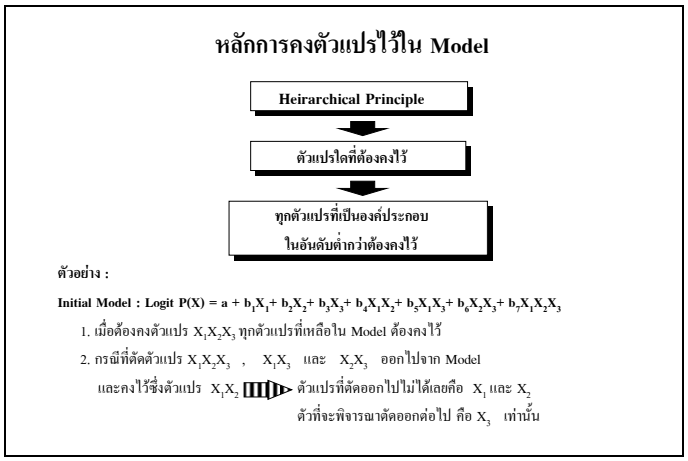
การตัดตัวแปรที่เป็น Interaction term ซึ่งอาจเป็น Third order term หรือ Second order term นั้น ใช้ผลจากวิธีการทดสอบทางสถิติเป็นเกณฑ์ ซึ่งจะกล่าวรายละเอียดในลำดับต่อไป ส่วนการตัดตัวแปรที่เป็น Main effect นั้น ใช้วิจารณญาณของผู้วิจัย ไม่ต้องใช้การทดสอบทางสถิติ เพราะพิจารณาเสมือนเป็น Confounder ซึ่งไม่มีการทดสอบทางสถิติ

รูปที่ 5.13



รูปที่ 5.14

จากวิธีการที่กล่าวข้างต้น บางตัวแปรจะถูกตัดออกไปด้วยเหตุที่ตัวแปรดังกล่าวไม่มีความสำคัญหรือไม่มีผลใด ๆ ต่อ Model กล่าวคือจะใส่หรือไม่ใส่ตัวแปรนั้นก็ไม่ทำให้ค่า



รูปที่ 5.15

สัมประสิทธิ์ที่คำนวณได้แตกต่างไป จากนั้นจึงพิจารณาตัวแปรที่อยู่ในอันดับรองลงไปตามหลักของ Heirarchical Principle

ในทำนองเดียวกัน เมื่อพบว่ามีบางตัวแปรที่พบว่ามีผลต่อ Model กล่าวคือ ถ้าตัดตัวแปรนั้นออกไป จะยังผลให้ค่าสัมประสิทธิ์ที่คำนวณได้นั้นแตกต่างไปจากเดิมมาก กรณีเช่นนี้ต้องคงตัวแปรดังกล่าวไว้ใน Model

เมื่อต้องคงตัวแปรที่เป็น Product term ไว้ ตัวแปรที่เป็นองค์ประกอบในอันดับรองลงไปทุกตัวจะต้องคงไว้แม้จะไม่มีนัยสำคัญทางสถิติ

เมื่อตัวแปรย่อยที่เป็น dummy variable ของตัวแปรหนึ่งต้องคงไว้ ตัวแปรย่อยตัวอื่นๆ ทุกตัวของตัวแปรนั้นต้องคงไว้ แม้ไม่มีนัยสำคัญทางสถิติ

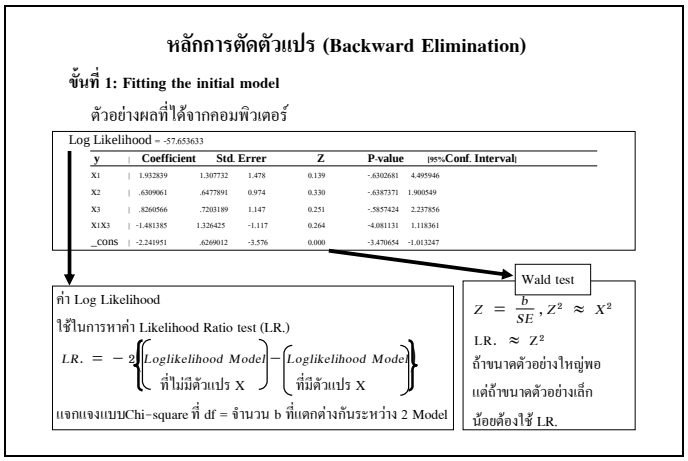
ดังนั้น ในขั้นตอนการตัดตัวแปรในลำดับถัดไป จะไม่ยุ่งเกี่ยวกับทุกตัวแปรที่ต้องคงไว้ตามหลักการดังกล่าว

ตัวอย่าง ให้ Initial model เป็นดังนี้

$$\text{Logit } P(\mathbf{X}) = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_1X_2 + b_5X_2X_3 + b_6X_1X_3 + b_7X_1X_2X_3$$

1. เมื่อต้องคงตัวแปร $X_1X_2X_3$ ทุกตัวแปรที่เหลือใน Model ต้องคงไว้
2. กรณีที่ตัดตัวแปร $X_1X_2X_3$, X_1X_3 และ X_2X_3 ออกไปจาก Model และคงไว้ซึ่งตัวแปร X_1X_2 ตัวแปรที่ตัดออกไปไม่ได้เลยคือ X_1 และ X_2 ตัวที่จะพิจารณาตัดออกต่อไป คือ X_3 เท่านั้น

หลักการตัดตัวแปรออกจาก Model นั้น ขั้นแรกให้ Fit Model ที่เป็น Initial model ก่อน จากผลจากคอมพิวเตอร์ สิ่งแรกที่ต้องพิจารณามีตัวเลข 2 จำนวนคือ (1) ค่า p-value จาก

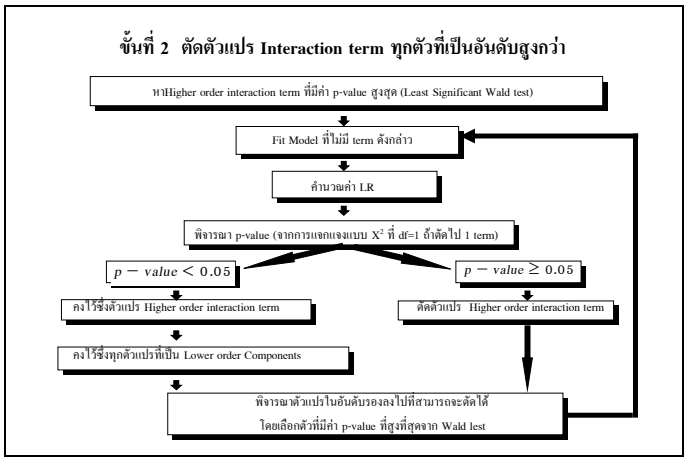


รูปที่ 5.16

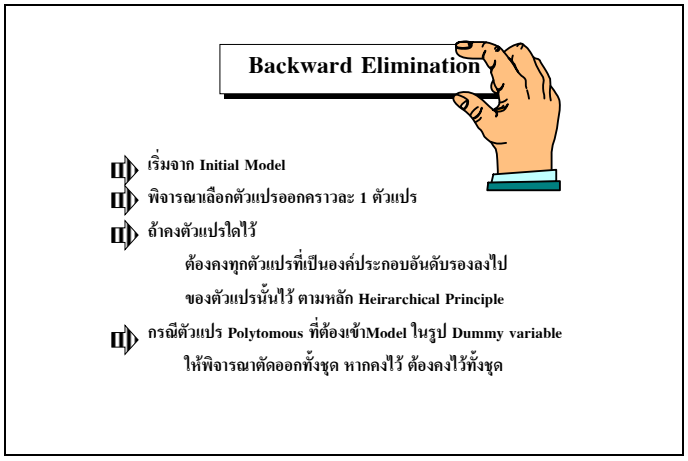
สถิติทดสอบ Wald test ซึ่งแทนด้วย Z ใน Output จาก STATA และ (2) ค่า Log Likelihood ทั้งสองค่านี้บอกเราว่าตัวแปรใดมีนัยสำคัญต่อ Model

ค่า p-value จาก Wald test เป็นการทดสอบสมมติฐานว่าค่าสัมประสิทธิ์แตกต่างจาก 0 หรือไม่ เป็นค่าที่สะดวกเพราะรู้ผลได้ภายใน Model เดียวที่ Fit ออกมาว่าตัวแปรใดมีผลอย่างมีนัยสำคัญต่อตัวแปรตาม แต่ถ้าจำนวนตัวอย่างน้อย ค่า Wald test จะไม่ถูกต้องเพราะ Wald test มีข้อกำหนดเบื้องต้นว่าข้อมูลต้องแจกแจงแบบปกติซึ่งมักไม่ได้ ตามนั้นถ้าตัวอย่างมีน้อย ในขณะที่ค่า Log Likelihood สามารถนำไปคำนวณค่า Likelihood Ratio (LR) ซึ่งมีค่าต่ำสุดเท่ากับศูนย์ จึงมีการกระจายแบบ Chi-square และใช้ในกรณีการแจกแจงไม่เป็นปกติแทน Wald test ได้ เพียงแต่ต้อง Fit Model ถึงสอง Model เพื่อได้ค่า Log Likelihood จากสอง Model โดย Model แรกใส่ตัวแปรที่จะทดสอบ และ Model ที่สองไม่ใส่ตัวแปรดังกล่าว ผลต่างของค่า Log Likelihood จากสอง Model คูณด้วย (-2) ได้เป็นค่า LR นำไปเปิดตาราง χ^2 ที่ค่า degree of freedom เท่ากับผลต่างของจำนวนสัมประสิทธิ์ระหว่างสอง Model ได้ค่า p-value สำหรับนำไปประกอบการตัดสินใจว่า ตัวแปรดังกล่าว มีผลต่อ Model หรือไม่ จึงเรียกวิธีการทดสอบความมีนัยสำคัญของตัวแปรแบบนี้ว่า Likelihood Ratio test (บางตำราเรียกว่า Deviance หรือ G^2)

จากผลที่ได้จากการ Fit Initial Model พิจารณาค่า P-value ของ Wald tests ของตัวแปรที่เป็น Interaction term ที่มีอันดับสูงสุดก่อน ถ้าหากอันดับเดียวกันนั้นมีหลาย term ให้เลือกตัด term ที่มีค่า p-value สูงที่สุด โดย Fit Model ใหม่ เป็น Model ที่สองที่ไม่มี term



รูปที่ 5.17



รูปที่ 5.18

ดังกล่าว แล้วนำค่า Log Likelihood ที่ได้ไปลบออกจาก ค่า Log Likelihood ที่ได้จาก Model ที่หนึ่ง นำผลลัพธ์ที่ได้ คูณด้วย -2 เป็นค่า LR เปิดตาราง Chi Square ที่ df = 1 (เนื่องจาก Model ทั้งสอง มีจำนวนสัมประสิทธ์ต่างกัน 1 ตัว) ได้ค่า p-value ซึ่งเป็นผลจาก LR test ถ้าหาก p-value < 0.05 ต้องคงตัวแปรดังกล่าวไว้ และคงไว้ซึ่งตัวแปรทุกตัวแปรที่เป็นองค์ประกอบในอันดับต่ำกว่าตัวแปรดังกล่าว แต่ถ้าหาก p-value ≥0.05 ก็สามารถตัดตัวแปรนี้ออกไปได้

จากนั้นจึงพิจารณาตัวแปรที่เหลือให้ตัดได้ โดยเลือกตัดตัวแปรที่มีอันดับสูงสุดและมี p-value ของ Wald test สูงสุดก่อนเป็นอันดับแรกแล้วดำเนินการเช่นเดียวกับที่กล่าวข้างต้น

ทำเช่นนี้ต่อไปจนกระทั่งไม่สามารถตัดตัวแปรใดออกจากModel ได้อีก เนื่องจากทุกตัวแปร มีผลต่อ Model อย่างมีนัยสำคัญทางสถิติ จึงถือว่าได้ Final Model

อย่างไรก็ตาม ตัวแปรที่เป็น Main effect บางตัวแปรอาจต้องคงไว้ในModel ทั้งที่พบว่ามีผลต่อModel อย่างไม่มีนัยสำคัญ (p≥0.05) เพื่อต้องการให้ผู้อ่านทราบว่า ได้ควบคุมผลกระทบของตัวแปรดังกล่าวแล้ว

วิธีการ Fit Model ตามที่กล่าวข้างต้นนี้ เรียกว่า Backward Elimination โดยพิจารณาเอาตัวแปรออกคราวละ 1 ตัวแปรแต่ถ้าต้องคงตัวแปรนั้นไว้ ทุกตัวแปรที่เป็นองค์ประกอบอันดับรองของตัวแปรดังกล่าวจะต้องคงไว้ตามหลักการของ Heirarchical Principle

ตัวอย่างการ Fit Model ของการศึกษาความสัมพันธ์ระหว่างปัจจัย X_1 กับ D โดยมีปัจจัยX_3 ถึง X_4 เป็นตัวแปรที่มีความสำคัญทางด้านชีวภาพต่อ D และมี Interaction Effect เมื่อ Initial Model เขียนในรูป logit form เป็น Logit P(x) = $a+b_1X_1 + b_2X_2 +$

ตัวอย่าง : Backward Elimination

1. Model ที่ 1 : Initial Model

Log Likelihood = -57.653633

y	Coefficient	Std. Errer	Z	P-value	[95%Conf.	Interval]
X1	1.932839	1.307732	1.478	0.139	-.6302681	4.495946
X2	.6309061	.6477891	0.974	0.330	-.6387371	1.900549
X3	.8260566	.7203189	1.147	0.251	-.5857424	2.237856
X4	-.3261456	.6569414	-0.496	0.620	-1.613727	.9614357
X1X3	-1.481385	1.326425	-1.117	0.264	-4.081131	1.118361
X1X4	-.4069955	1.004148	-0.405	0.685	-2.375089	1.561098
_cons	-2.241951	.6269012	-3.576	0.000	-3.470654	-1.013247

ค่าที่จะตัดออกคือ X1X4

รูปที่ 5.19

2. Model ที่ 2 : ไม่มีตัวแปร X1X4

Log Likelihood = -57.735607

y	Coefficient	Std. Error	Z	P-value	[95%Conf.	Interval]
X1	1.839284	1.276507	1.441	0.150	-.6626245	4.341192
X2	.6436386	.6455916	0.997	0.319	-.6216976	1.908975
X3	.8429361	.7207969	1.169	0.242	-.5697999	2.255672
X4	-.5031502	.5043087	-0.998	0.318	-1.491577	.4852766
X1X3	-1.553466	1.305413	-1.190	0.234	-4.112029	1.005097
_cons	-2.207755	.6191707	-3.566	0.000	-3.421307	-.9942028

LR = -2(-57.74 -(-57.65)) = 0.18

$X^2_{df=1}$ ได้ค่า p = 0.6892 → มากกว่า 0.05 ดังนั้น สามารถตัดตัวแปร X1X4 ได้

ตัวแปรที่จะตัดออกลำดับต่อไป คือ X1X3

รูปที่ 5.20

3. Model ที่ 3 : ไม่มีตัวแปร X1X3

Log Likelihood = -59.839966

y	Coefficient	Std. Error	Z	P-value	[95%Conf.	Interval]
X1	.5128482	.6508643	0.788	0.431	-.7628223	1.788519
X2	.6767462	.6568664	1.030	0.303	.6106883	1.964181
X3	.4361177	.597445	0.730	0.465	-.7348529	1.607088
X4	-.5383285	.5042896	-1.067	0.286	-1.526718	.450061
_cons	-1.9106	.5103662	-3.744	0.000	-2.910899	-.9103008

LR = -2(-59.84 -(-57.74)) = 4.2

$X^2_{df=1}$ ได้ค่า p = 0.04 น้อย[illegible]0.05 ดังนั้น สามารถตัดตัวแปร X1X3

ต้องคงไว้ใน Model พร้อมทั้ง X1 และ X3 → หลัก Heirarchical well-formated model

ตัวแปรที่จะตัดลำดับต่อไปคือ X2

รูปที่ 5.21

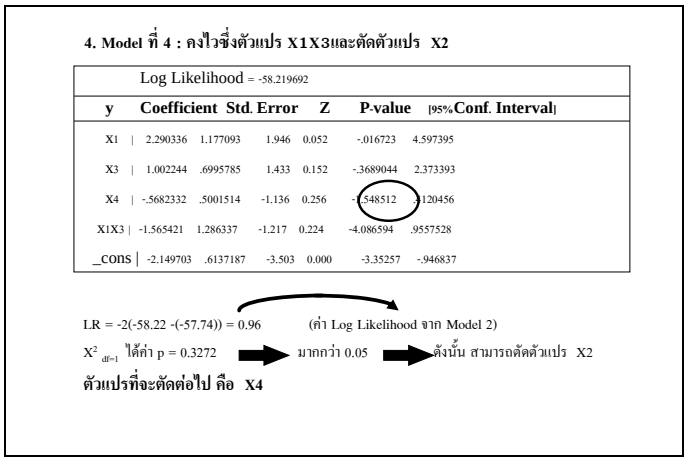
$b_3X_3 + b_4X_4 + b_5X_1X_3 + b_6X_1X_4$

ขั้นที่หนึ่งคือการ Fit Model ที่เป็น Initial Model พิจารณาตัวแปรที่จะตัดออกเป็นอันดับแรก ถ้ามีตัวแปรที่มีอันดับสูงสุดใน Model มากกว่า 1 ตัวแปร ให้เลือกตัดตัวแปรที่มีค่า p-value ของ Wald test สูงที่สุดเป็นอันดับแรก ในที่นี้คือ X_1X_4 ซึ่งมี p-value สูงสุด (0.685) จากนั้น Fit Model ที่สองต่อไป

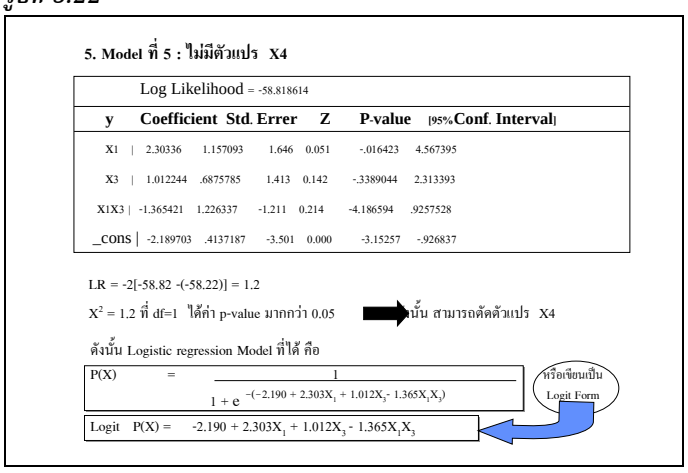
ขั้นต่อไปคือ นำค่า Log Likelihood ที่ได้จาก Model ที่สอง ลบออกจากที่ได้จาก Model ที่หนึ่งแล้วคูณด้วย -2 ได้ค่า LR เท่ากับ 0.18 เปิดตาราง X^2 ที่ degree of freedom (df) เท่ากับ 1 (คือ b_6 นั่นเอง) ได้ค่า p-value = 0.6892 ซึ่งมากกว่า 0.05 แสดงว่า ตัวแปร X_1X_4 ซึ่งเป็น Interaction term นั้น ส่งผลต่อ Model อย่างไม่มีนัยสำคัญทางสถิติ สามารถตัดออกไปได้

จากนั้นพิจารณาตัวแปรในอันดับรองลงมาเพื่อตัดออกในที่นี้ยังคงเป็น Second order term อีกตัวหนึ่งคือ X_1X_3 จึง Fit Model ที่สามโดยไม่มีตัวแปร X_1X_3 คำนวณค่า LR. โดยเปรียบเทียบกับModel ที่ 2 ได้ค่า p-value < 0.05 จึงไม่สามารถตัดตัวแปรนี้ออกไปได้ ต้องคงไว้ในModel ผลจากกรณีนี้ยังกำหนดให้ต้องคงตัวแปร พร้อมกับตัวแปร X_1 และ X_3ไว้ใน Model ด้วย ตามหลักการของ Heirarchical Principle

Model สำหรับเริ่มต้นในขั้นต่อไป คือ Model ที่ 2 ซึ่งมี X_1X_3 อยู่ใน Model ดังนั้นยังคงเหลือเพียง แปร X_2 และ X_4 เท่านั้นที่จะพิจารณาตัดออกในขั้นต่อไป ในที่นี้เลือกตัวแปร X_2 เพราะมีค่า p-value สูงที่สุดในอันดับ



รูปที่ 5.22



รูปที่ 5.23

เดียวกัน (0.319) และ Fit Model ที่สี่ คำนวณค่า LR. โดยต้องเปรียบเทียบกับ Model ที่ 2 ได้ ค่า p-value > 0.05 จึงตัดออกจาก Model ได้

เหลือเพียงตัวแปร X_4 ที่จะพิจารณาตัดออกเป็นตัวแปรสุดท้าย โดยดำเนินการ ตามวิธีการที่กล่าวข้างต้น และได้ Final Model

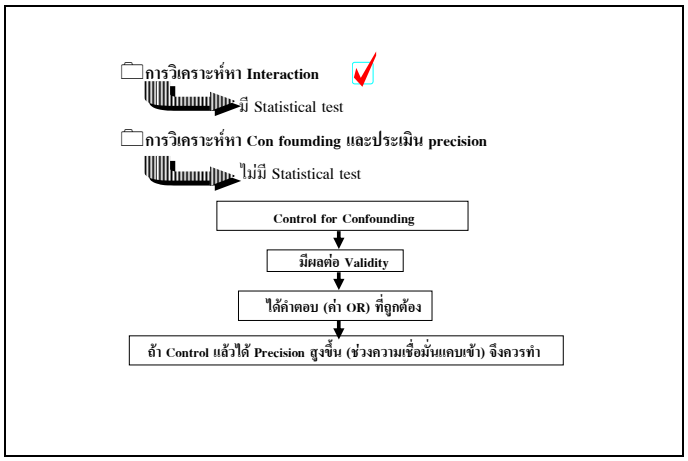
เพื่อนำไปคำนวณหา OR เพื่อนำไปอธิบายความสัมพันธ์ในเรื่องที่ศึกษา ต่อไป

อย่างไรก็ตาม หากการทบทวนวรรณกรรมบ่งชี้ว่ามีบางตัวแปรที่เป็น หรือ อาจเป็นตัวกวน แม้ตัวแปรดังกล่าว ส่งผลต่อ Model อย่างไม่มีนัยสำคัญทางสถิติ ผู้วิจัยอาจต้องคงไว้ใน Model ทั้งนี้ให้พิจารณาความกระชับ (Precision) ของค่า OR ด้วย ซึ่งจะได้กล่าวในรายละเอียดต่อไป

การหา p-value จาก LR test ตามที่กล่าวแล้วข้างต้น สามารถใช้โปรแกรม STATA คำนวณได้ (ดูรายละเอียดในแบบฝึกหัด)

4. การวิเคราะห์หา Interaction Effect และ Confounding Effect

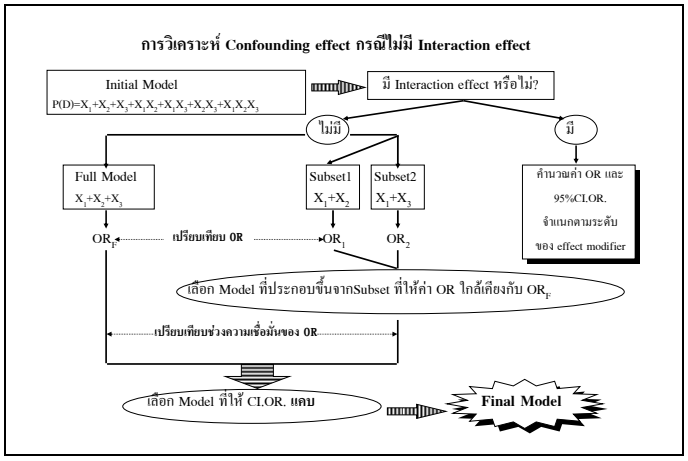
จากขั้นตอนตามที่กล่าวข้างต้น ถ้าหากมี Interaction term ใดจำเป็นต้องคงไว้ใน Model ด้วยเหตุที่มีนัยสำคัญต่อ Model (p-value <0.05) แสดงว่ามี Interaction Effect ในทางกลับกัน ถ้าผลการทดสอบได้ค่า p>0.05 แสดงว่า Interaction term ไม่มีผลต่อ Model สามารถตัดออกไปได้เนื่องจากไม่มี Interaction Effect ด้วยวิธีการที่กล่าวนี้ ทำให้เราทราบ Effect Modifier นั่นคือ Interaction Effect สามารถทดสอบได้ทางสถิติโดยใช้ Likelihood



Ratio test

ในทางตรงข้าม กรณีการวิเคราะห์หา Confounder นั้น ไม่สามารถทดสอบทางสถิติได้ เพราะเกี่ยวข้องกับ Validity มิได้เกี่ยวข้องกับ Random error เหมือน Interaction Effect การควบคุมผลของ confounder จะยังผลให้ได้คำตอบที่ถูกต้อง ซึ่งบางครั้งอาจไม่กระชับเท่าที่ควร หรือ Precision อาจไม่สูงนัก จึงต้องพิจารณา Precision ด้วยหลังจากที่ควบคุม Confounding Effect

รูปที่ 5.24

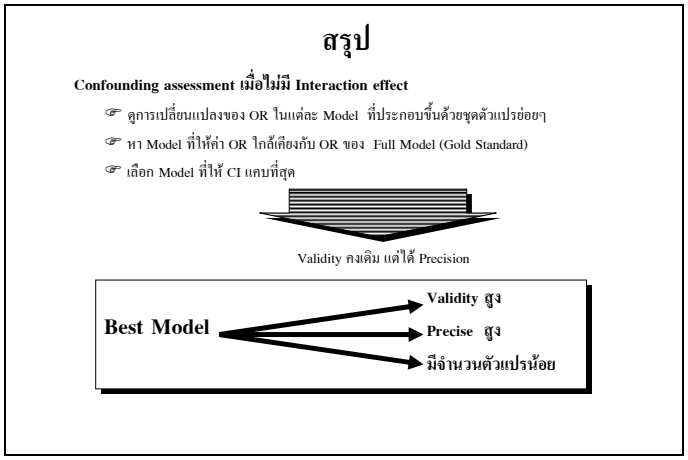


การวิเคราะห์ว่ามี Confounding effect หรือไม่ จะกระทำหลังจากที่ได้วิเคราะห์หา Interaction Effect เสมอ ในกรณีที่ไม่มี Interaction Effect ใน Model จะคงเหลือกลุ่มตัวแปรที่เป็น Main effect ที่ไม่เป็นองค์ประกอบของ Interaction term ค่า OR จาก Model นี้ ถือว่าดีที่สุด เพราะได้ควบคุมผลจากตัวแปรที่มีศักยภาพเป็นตัวกวนได้ (Potential confounder) ทุกตัวแปรแล้ว จึงถือว่าเป็น Model ที่เป็น ***Gold standard*** แต่เนื่องจาก Model ที่ดีที่สุด ควรมีตัวแปรไม่มากนัก จึงต้องพิจารณาตัดตัวแปรอื่นต่อไปดังต่อไปนี้

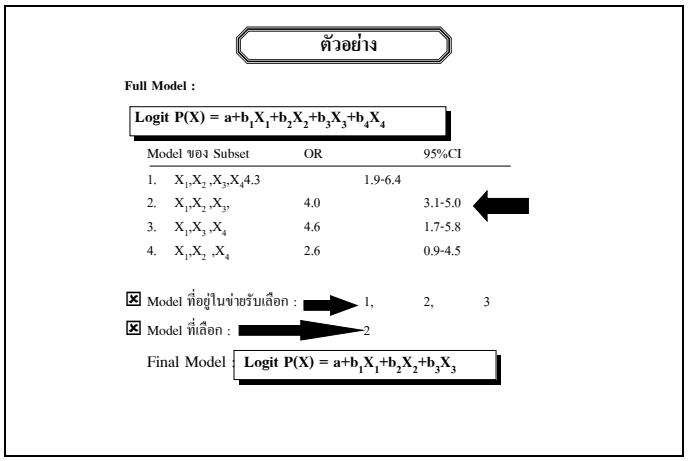
ขั้นแรก ให้ระบุชุดของ Potential Confounder (เป็น Subset ของตัวแปรที่เหลือใน Model) แล้ว Fit Model นำผลที่ได้มาคำนวณค่า OR เปรียบเทียบ กับค่า OR ที่เป็น Gold Standard เลือก Model ของ Subset ที่ให้ค่า OR ใกล้เคียงกับ Gold standard จากนั้นคำนวณช่วงความเชื่อมั่นของค่า OR แล้วนำไปเปรียบเทียบกับช่วงความเชื่อมั่นค่า OR จาก Gold Standard ถ้า ช่วงของ Model ใดแคบก็เลือก Model นั้นเป็น ***Final Model***

โดยสรุป หลักการของการวิเคราะห์ Confounding effect ก็คือการจัดกลุ่มย่อย (Subset) ของตัวแปรที่เราระบุใน Initial Model ว่าเป็น Confounder ที่ต้องนำเข้า Model เพื่อควบคุมผลกระทบ โดยอาศัยองค์ความรู้จากวรรณกรรมและทฤษฎีที่เกี่ยวข้องเป็นพื้นฐาน จากนั้น Fit Model ของแต่ละชุดของตัวแปรดังกล่าวเลือก Model ที่ให้ค่า OR ใกล้เคียงกับ Model ที่มีทุกตัวแปร (Gold standard) จากนั้นพิจารณา precision ถ้าช่วงความเชื่อมั่นของ OR จากModel ใดแคบกว่าก็เลือก Model นั้นเป็น Final Model แต่ช่วงความเชื่อมั่นยังคงไม่แตกต่างกันมากนัก ผู้วิจัยอาจเลือก Model ของ Subset หรือ Full Model ก็ได้ ขึ้นอยู่กับความประสงค์ กล่าวคือถ้าผู้วิจัยต้องการให้ผู้อ่านทราบว่าได้ควบคุมตัวแปรใดบ้างก็เลือก Full Model แต่โดยทั่วไป Best Model ต้องมีตัวแปรน้อยที่สุดเท่าที่จำเป็น ดังนั้นถ้าไม่มีความประสงค์ดังกล่าวจึงควรเลือก Model Subset เพราะมีจำนวนตัวแปรน้อยกว่า

รูปที่ 5.25



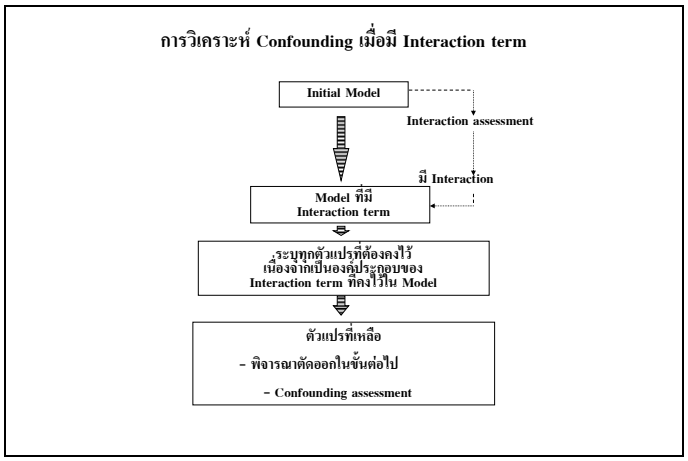
รูปที่ 5.26



ตัวอย่าง หลังจากที่ได้ผ่านการวิเคราะห์ หา Interaction แล้ว พบว่าไม่มี Interaction จึงคงเหลือเพียง Model ที่ประกอบขึ้นด้วย Main effect ได้แก่ X_1, X_2, X_3 และ X_4 ทั้งสี่ตัวแปรนี้ เป็นตัวแปรที่องค์ความรู้เดิมบ่งชี้ว่าเป็น Confounder เพื่อหา Best Model จึงจัดกลุ่มย่อยของตัวแปรได้ค่าต่างกันหลายชุด (n ชุด) จากนั้น Fit Model แล้วคำนวณค่า OR และ 95% CI ของ OR ในแต่ละ Model ดังนั้น Model ที่ 1 ก็คือ Full Model ค่า OR จึงถือเป็น Gold Standard นอกจากนี้เป็น Model ของ Subset ตัวแปรทั้ง 4 เฉพาะModel ↑

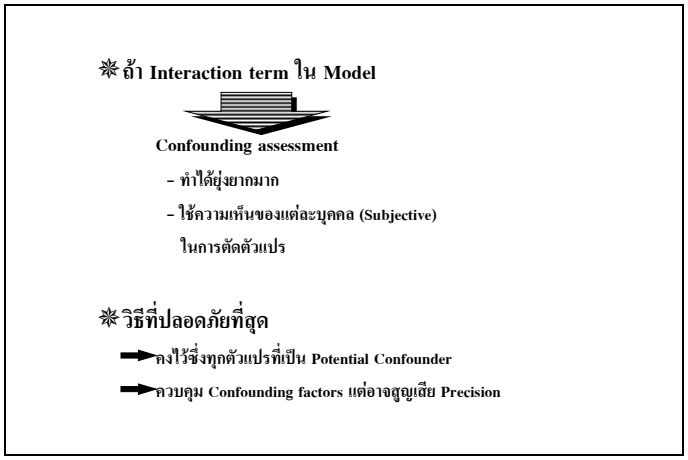
และ → เท่านั้นที่ให้ผล Valid คือ OR ไม่ต่างจาก Gold standard แต่เมื่อพิจารณาช่วงความเชื่อมั่น พบว่าModel ↑ มีช่วงแคบกว่า นั่นคือ Precision สูงกว่า จึงถูกเลือกเป็น Final Model

รูปที่ 5.27



กรณีที่พบว่ามี Interaction effect ในการศึกษาใด Modelที่ผ่านการวิเคราะห์ Interaction แล้วจะคงเหลือ Interaction term อยู่ในModel จากหลัก Heirarchical principle ต้องคงไว้ซึ่งตัวแปรทุกตัวแปรที่เป็นองค์ประกอบของ term ดังกล่าวที่อยู่ในอันดับรองลงมา ตัวแปรที่นอกเหนือจากนี้เท่านั้นที่จะพิจารณาตัดออกจาก Model ต่อไป

รูปที่ 5.28



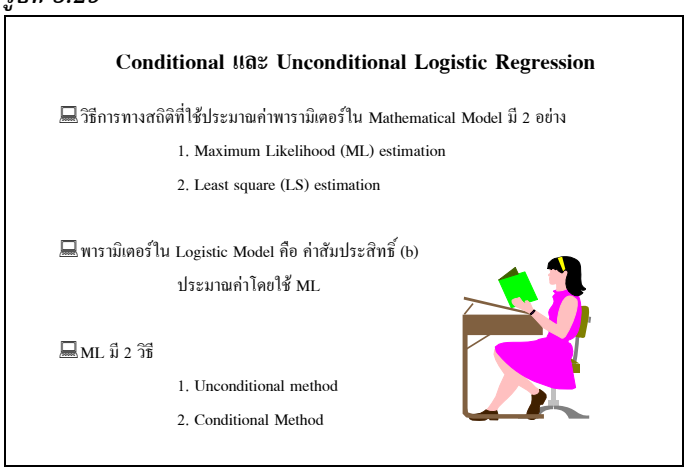
ในกรณีนี้ เมื่อถึงขั้นตอนสุดท้ายของการ Fit Model ซึ่งเป็น Confounding assesment นั้น ทำได้ยุ่งมากมาก นอกจากนั้นยังใช้ความรู้สึกค่อนข้างมาก ในการตัดสินใจว่าจะตัดตัวแปรใดออก ดังนั้น วิธีที่ปลอดภัยที่สุดคือให้คงตัวแปรที่องค์ความรู้บ่งชี้ว่าเป็น Confounder ทุกตัวแปรไว้ใน Model ซึ่งเป็นการควบคุมผลของตัวแปรดังกล่าว วิธีการนี้จะประกัน ผลด้าน Validity แต่อาจไม่ได้ผลด้าน Precision กล่าวคือได้คำตอบที่ตรงตามความเป็นจริง แต่อาจขาดความกระชับ หรือช่วงความเชื่อมั่นอาจกว้างกว่าที่ตัดบางตัวแปรออกไป

อย่างไรก็ตาม หากต้องการวิเคราะห์หา

Confounding effect ในกรณีนี้ ยังคงใช้หลักเดียวกันกับกรณีที่ไม่มี Interaction กล่าวคือ เปรียบเทียบค่า OR จาก Model ที่เป็น Subset กับ Full Model หรือค่า OR ที่เป็น Gold standard ซึ่งจะไม่กล่าวรายละเอียดในที่นี้ ผู้อ่านสารมารถศึกษารายละเอียดวิธีการคำนวณได้ใน Kleinbaum, 1994 หน้า 203-218

5. Conditional Logistic Regression และ Unconditional Logistic Regression

รูปที่ 5.29



การวิเคราะห์ข้อมูลโดยใช้ Logistic Regression ต้องอาศัยโปรแกรมคอมพิวเตอร์ ซึ่งมีวิธีการให้เลือกสองวิธีคือ Conditional และ Unconditional ทั้งสองวิธี เป็นวิธีการทางสถิติที่ใช้ในการประมาณค่าพารามิเตอร์

ค่าพารามิเตอร์ ของ Mathematical Model ใด ๆ มีวิธีการประมาณค่า ที่ใช้กันแพร่หลาย 2 วิธีคือ Maximum Likelihood (ML.) กับ Least Square (LS.) ค่าพารามิเตอร์ใน Model หรือค่าสัมประสิทธิ์ (Coefficient) คือแทนด้วย b สำหรับการวิเคราะห์โดยใช้ Logistic Regression นั้น ใช้ ML. ในการประมาณค่าเป็นวิธีการที่ยุ่งยากซับซ้อนมากจึงต้องใช้คอมพิวเตอร์เท่านั้น ถ้าการศึกษาใดมีตัวแปรที่ศึกษาไม่มากนักเมื่อเทียบกับขนาดตัวอย่าง ซึ่งยังผลให้มีจำนวนสัมประสิทธิ์ที่จะประมาณค่ามีจำนวนน้อยเมื่อเทียบกับขนาดตัวอย่าง สามารถใช้วิธี Unconditional วิเคราะห์ได้

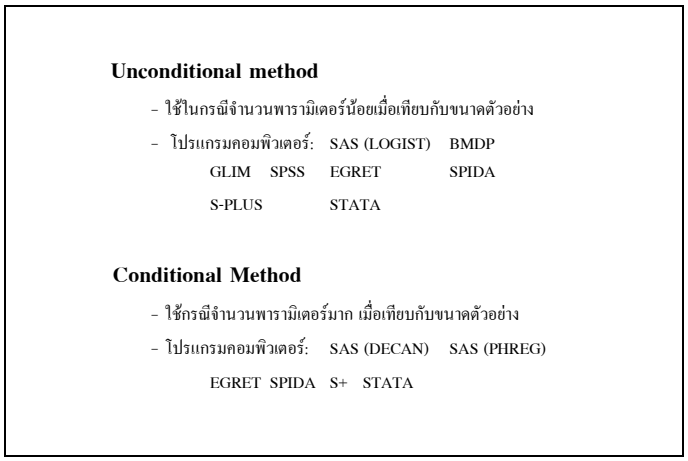
Unconditional method

- ใช้ในกรณีจำนวนพารามิเตอร์น้อยเมื่อเทียบกับขนาดตัวอย่าง
- โปรแกรมคอมพิวเตอร์ SAS (LOGIST) BMDP
  GLIM SPSS EGRET SPIDA
  S-PLUS STATA

Conditional Method

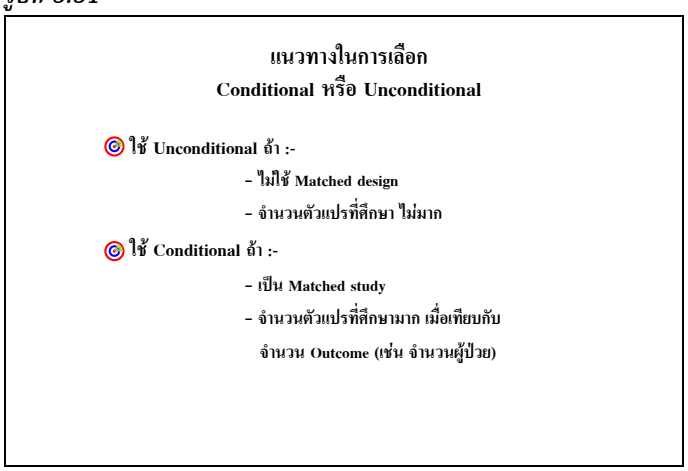
- ใช้กรณีจำนวนพารามิเตอร์มาก เมื่อเทียบกับขนาดตัวอย่าง
- โปรแกรมคอมพิวเตอร์ SAS (DECAN) SAS (PHREG)
  EGRET SPIDA S+ STATA

แต่ถ้าจำนวนสัมประสิทธิ์มีมากเมื่อเทียบกับขนาดตัวอย่าง ต้องใช้ Conditional ไม่เช่นนั้นค่า OR ที่ได้จะสูงกว่าความเป็นจริง (Overestimate Odds Ratio) ทั้งสองวิธีต้องใช้



โปรแกรมคอมพิวเตอร์ต่างกันตามตัวอย่างที่ให้ ซึ่งผู้วิจัยต้องเลือกวิธีการให้เหมาะสมกับการศึกษาของตนเอง

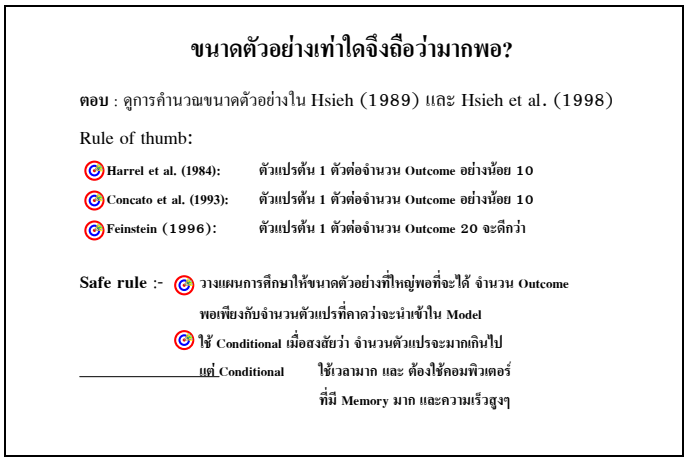
รูปที่ 5.31



แนวทางในการเลือกวิธีดังกล่าว มีดังนี้ คือ ให้เลือกใช้ Unconditional เมื่อการศึกษานั้นไม่เป็นแบบ Matched study แต่ต้องมีจำนวนตัวแปรไม่มากนัก ถ้าหากเป็น Matched study หรือแม้ไม่ใช่ Matched study แต่มีจำนวนตัวแปรมาก ต้องใช้ Conditional

เหตุที่เมื่อเป็น Matched study ต้องใช้ Conditional ก็เพราะการวิเคราะห์ Matched data ต้องสร้างตัวแปร (Dummy variable) จากตัวแปรที่ Matched เท่ากับจำนวนตัวอย่างลบด้วย 1 ซึ่งยังผลให้มีจำนวนตัวแปรมาก เมื่อเปรียบเทียบกับขนาดตัวอย่าง

รูปที่ 5.32



จำนวนตัวแปรที่นำเข้า Model มากเท่าไรจึงถือว่ามากนั้น เป็นประเด็นของขนาดตัวอย่าง การคำนวณขนาดตัวอย่างสำหรับ Logistic regression นั้นยุ่งยากซับซ้อน ผู้อ่านสามารถศึกษาใน Hsieh (1989) และ Hsieh et al. (1998) หลักการหยาบๆ โดย Harrel et al. (1984): และ Concato et al. (1993) ระบุว่า ตัวแปรต้น 1 ตัวต่อจำนวน Outcome อย่างน้อย 10 เช่นการศึกษาปัจจัยที่มีผลต่อการป่วยด้วยโรคหัวใจ ถ้าผู้วิจัยคาดว่าต้องมีประมาณ 5 ตัวแปรต้นใน Model จะต้องเก็บตัวอย่างที่สามารถให้ได้จำนวนผู้ป่วยอย่างน้อย 50 ราย ดังนั้นถ้าอัตราป่วยในชุมชนคือ 5 รายต่อประชากรพันคน จะต้องใช้ขนาดตัวอย่างถึง

10,000 คนเป็นต้น Feinstein (1996) ยังได้แสดงให้เห็นว่าตัวแปรต้น 1 ตัวต่อจำนวน Outcome 20 จะปลอดภัยกว่า

กรณีมีจำนวน Outcome น้อยเมื่อเทียบกับจำนวนตัวแปรต้นใน Model ถ้าใช้ Unconditional ML. จะไม่เหมาะสม คือจะเกิด Over fitting และยังผลให้ได้ค่า OR ที่มากเกินความเป็นจริง ดังนั้นเพื่อความปลอดภัยจึงควรวางแผนการศึกษาที่มั่นใจว่าจะได้จำนวน Outcome ให้มากพอกับจำนวนตัวแปรต้นที่คาดว่าจะต้องนำเข้าใน Model แต่ถ้าไม่สามารถให้เป็นไปตามนั้นได้ให้เลือก Conditional ML. เมื่อสงสัยว่าจำนวนตัวแปรอาจจะมาก เพราะ Conditional ML. แต่นั่นหมายถึงต้องลงทุนด้านเวลา และต้องอาศัยคอมพิวเตอร์ที่มีหน่วยความจำ (RAM) ที่ใหญ่พอและมีความเร็วสูง ๆ

สรุปท้ายบท :

เอกสารอ้างอิงประจำบทที่ 5

Concato, J., Feinstein, A.R., and Holford, T.R., (1993). The risk of determining risk in multivariable models. *Annals of Internal Medicine.* **118**:201-210.

Feinstein, A.R. (1996). *Multivariable analysis: an introduction.* Yale university Press: New Haven.

Harrell, F.E., Lee, K.L., Califf, R.M., Pryor, D.B., and Rosati, R.A. (1984). Rgression modelling strategies for improve prognostic modelling. *Statistics in Medicine.* **3**:143-152.

Hsieh, F. Y. (1989). Sample size tables for logistic regression. *Stat Med* **8,** 795-802.

Hsieh, F. Y., Bloch, D. A., & Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression. *Stat Med* **17,** 1623-34.

Hosmer, D.W., and Lemeshow, S. (1989). *Applied Logistic Regression*. New York: John Wiley & Sons.

Kleinbaum, D.G. (1994). *Logistic Regression: A self-learning text*. New York Springer-Verlag.

Mazumdar, M., and Glassman, J. R. (2000). Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. *Stat Med* **19,** 113-32.

StataCorp. (1999). *Stata statistical software: Release 6.0*. College Station. TX: Stata Corporation.

แบบฝึกหัดที่ 5

แบบฝึกหัดรวบยอด

1. จงทำความเข้าใจการสร้างโมเดล (Model fitting strategies) การนำเสนอ และการแปลผล โดยใช้ข้อมูลที่วิเคราะห์ในแบบฝึกหัดที่ 1 ถึง 4 เป็นลำดับดังต่อไปนี้ (แฟ้มข้อมูลชื่อ EXAMPLE.DTA ในแผ่นดิสเกตต์ที่แจก โดยแตกต่างจากข้อมูลที่ใช้ในแบบฝึกหัดที่ 1 ถึง 4 เพียงเล็กน้อย เพื่อให้มีประสบการณ์กับ Polytomous variable) จากนั้นวิเคราะห์ข้อมูลในโจทย์ที่ให้ไว้ตามข้อ 2 และ 3

ทบทวนปัญหา

ตัวแปรในแฟ้มข้อมูลมีดังต่อไปนี้

ชื่อตัวแปร	**คำอธิบาย**	**รหัส**
ID	เลขที่แบบสอบถาม	1 ถึง 465 ตามลำดับ
DEAD	การรอดชีพของทารก เมื่อ 28 วันหลังคลอด	0= รอด 1=ตาย
AREA	พื้นที่ ที่มารดาอาศัย	0=พื้นที่ควบคุม 1=พื้นที่ทดลอง
MALPRES	ท่าของทารกขณะคลอด	0=ปกติ 1= ผิดปกติ
BWT	น้ำหนัก ทารกแรกเกิด (กรัม)	ตามที่ระบุ
MAGE	อายุมารดา(ปี)	ตามที่ระบุ
PLACE	สถานที่คลอด	0=โรงพยาบาล 1=สถานีอนามัย 2=ที่บ้าน 3=ข้างถนน(ขณะเดินทาง)

คำถามวิจัยคือ ″AREA″ มีผลอย่างไรต่อ ″DEAD″ เมื่อควบคุมผลกระทบจากปัจจัยอื่น ๆ (ได้แก่ MALPRES BWT MAGE และ PLACE) ปัญหานี้เป็น Risk assessment goal ที่มี exposure of interest คือ AREA

ต่อไปนี้เป็นขั้นตอนการวิเคราะห์เต็มรูปแบบ ผลจาก STATA แสดงเป็นรูปแบบตัวอักษรต่างไปจากคำบรรยาย โดย**อักษรตัวหนาตามหลังจุด** คือคำสั่ง ที่แสดงไว้ก่อนผลทุกครั้ง เพื่อท่านสามารถทำซ้ำได้

ขั้นที่ 1 Exploring the data and univariate analysis

คำสั่ง ″list″ สำหรับดูข้อมูลดิบ

```
. list  dead area malpres bwt mage place

          dead      area    malpres   bwt        mage      place
  1.       1         1         0      2600        30         0
  2.       1         1         0      2900        29         1
  3.       1         1         0      3100        25         0

       --- ข้าม 460 records ---

464.       0         1         0      3500        30         0
465.       0         1         0      3200        22         1
```

คำสั่ง "summarize" สำหรับดูรายการตัวแปรและสรุปข้อมูล

. summarize

Variable	Obs	Mean	Std. Dev.	Min	Max
dead	465	.1397849	.3471372	0	1
area	465	.5182796	.5002039	0	1
malpres	465	.0752688	.2641087	0	1
bwt	465	3010.695	437.7349	1850	4000
mage	465	25.52473	5.362298	17	42
place	465	.2408602	.5273217	0	3

คำสั่ง "tab" ตามด้วยตัวแปรตัวเดียวสำหรับแจกแจงข้อมูล

. tab dead

dead	Freq.	Percent	Cum.
0	400	86.02	86.02
1	65	13.98	100.00
Total	465	100.00	

คำสั่ง "ci" สำหรับคำนวณค่าช่วงเชื่อมั่น

. ci dead

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]	
dead	465	.1397849	.0160981	.1081507	.1714192

สรุปผลการวิเคราะห์ข้างต้น *"จากมารดา 465 คน มีทารกตายภายในเดือนแรกหลังคลอด 65 ราย คิดเป็นอัตราตาย (neonatal dead rate) 14.0% (95%CI: 10.8% ถึง 17.1%)".*

ขั้นที่ 2 Bivariate (crude) analysis

2.1 Crude effect of AREA on DEAD

. cs dead area, or

	area		
	Exposed	Unexposed	Total
Cases	37	28	65
Noncases	204	196	400
Total	241	224	465
Risk	.153527	.125	.1397849
	Point estimate		[95% Conf. Interval]
Risk difference	.028527		-.0342996 .0913535
Risk ratio	1.228216		.778466 1.937803
Attr. frac. ex.	.1858108		-.2845776 .4839517
Attr. frac. pop	.1057692		
Odds ratio	1.269608		.7512221 2.145309 (Cornfield)

chi2(1) = 0.79 Pr>chi2 = 0.3754

ตัวอย่างการรายงานผลจากการวิเคราะห์ข้างต้น *"จากมารดาที่อยู่ในพื้นที่ทดลอง 241 คน อัตราตายทารกตายภายในเดือนแรกหลังคลอดเท่ากับ 15.4% ในขณะที่มารดาที่อยู่ในพื้นที่ควบคุม 224 คน อัตราตายทารกตายภายในเดือนแรกหลังคลอดเท่ากับ 12.5% มารดาที่อยู่ในพื้นที่ทดลองมีโอกาสที่ทารกตายภายในเดือนแรกหลังคลอดเป็น 1.26 เท่าของผู้ที่อยู่ในพื้นที่ควบคุม (95%CI: 0.8 ถึง 2.1) อย่างไรก็ตาม ความสัมพันธ์ดังกล่าวไม่มีนัยสำคัญทางสถิติ (p-value = 0.375).*

2.2 Crude effect of MALPRES on DEAD

```
. cs dead malpres, or
```

```
                 | malpres                   |
                 |   Exposed   Unexposed  |     Total
-----------------+---------------------------+----------
           Cases |        21          44  |        65
        Noncases |        14         386  |       400
-----------------+---------------------------+----------
           Total |        35         430  |       465
                 |                         |
            Risk |        .6    .1023256  |  .1397849
                 |                         |
                 |      Point estimate     |  [95% Conf. Interval]
                 |-------------------------+----------------------
 Risk difference |         .4976744        |  .3328653    .6624835
      Risk ratio |         5.863636        |  3.972854    8.654289
 Attr. frac. ex. |         .8294574        |  .7482918    .8844504
 Attr. frac. pop |         .2679785        |
      Odds ratio |         13.15909        |  6.309044    27.44195  (Cornfield)
                 +-------------------------------------------------
                             chi2(1) =    66.67  Pr>chi2 = 0.0000
```

2.3 Crude effect of BWT on DEAD

สี่คำสั่งต่อไปนี้เพื่อสร้างตัวแปรใหม่สำหรับจัดกลุ่มน้ำหนักแรกเกิด

```
. gen bwtg = .
(465 missing values generated)

. replace bwtg = 1 if bwt < 2500
(39 real changes made)

. replace bwtg = 2 if bwt >= 2500 & bwt <= 3000
(213 real changes made)

. replace bwtg = 3 if bwt > 3000
(213 real changes made)
```

คำสั่งต่อไปนี้เพื่อให้ได้ค่าสัดส่วนการตายจำแนกตามกลุ่มน้ำหนักแรกเกิด

```
. tab  bwtg dead, row chi2 exact
```

```
           |         dead
      bwtg |         0          1 |     Total
-----------+----------------------+----------
         1 |        27         12 |        39
           |     69.23      30.77 |    100.00
-----------+----------------------+----------
         2 |       175         38 |       213
           |     82.16      17.84 |    100.00
-----------+----------------------+----------
         3 |       198         15 |       213
           |     92.96       7.04 |    100.00
-----------+----------------------+----------
     Total |       400         65 |       465
           |     86.02      13.98 |    100.00

          Pearson chi2(2) =  20.3082   Pr = 0.000
           Fisher's exact =                 0.000
```

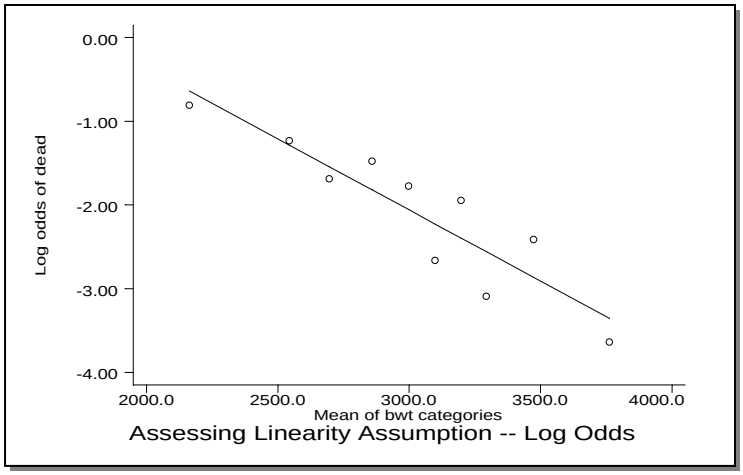
สองคำสั่งต่อไปนี้เพื่อให้ได้ค่า OR โดยกำหนดให้กลุ่มน้ำหนักแรกเกิดน้อยกว่า 2500 กรัม เป็นกลุ่มอ้างอิง

```
. csi 38 12 175 27, or

                 |   Exposed   Unexposed  |     Total
-----------------+------------------------+----------
           Cases |        38          12  |        50
        Noncases |       175          27  |       202
-----------------+------------------------+----------
           Total |       213          39  |       252
                 |                        |
            Risk |  .1784038    .3076923  |  .1984127
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |        -.1292886       | -.2829945    .0244174
      Risk ratio |         .5798122       |  .3338618     1.00695
 Prev. frac. ex. |         .4201878       |   -.00695    .6661382
 Prev. frac. pop |         .3551587       |
      Odds ratio |         .4885714       |  .2294889    1.037412  (Cornfield)
                 +-------------------------------------------------
                             chi2(1) =     3.46  Pr>chi2 = 0.0627
```

```
. csi 15 12 198 27, or

                 |   Exposed   Unexposed  |     Total
-----------------+------------------------+----------
           Cases |        15          12  |        27
        Noncases |       198          27  |       225
-----------------+------------------------+----------
           Total |       213          39  |       252
                 |                        |
            Risk |  .0704225    .3076923  |  .1071429
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |        -.2372698       |  -.386141   -.0883985
      Risk ratio |         .2288732       |  .1161831    .4508654
 Prev. frac. ex. |         .7711268       |  .5491346    .8838169
 Prev. frac. pop |         .6517857       |
      Odds ratio |         .1704545       |  .0730813    .3964905  (Cornfield)
                 +-------------------------------------------------
                             chi2(1) =    19.40  Pr>chi2 = 0.0000
```

กรณีตัวแปรต่อเนื่องที่ไม่มีเกณฑ์ในการจัดกลุ่ม อาจใช้คำสั่ง "lintrend" สำหรับดูว่ามีความสัมพันธ์เชิงเส้นกับตัวแปรตามหรือไม่ เพื่อเป็นแนวทางตัดสินใจว่าจะนำตัวแปรดังกล่าวเข้าใน Model ในรูปตัวแปรต่อเนื่องหรือจัดกลุ่มที่เหมาะสมก่อนนำเข้าต่อไป

```
. lintrend dead bwt, groups(12) plot(log) xlab ylab

The proportion and log odds of dead by categories of bwt

  (Note: 12 bwt categories of equal sample size;
     Uses mean bwt value for each category)

      bwt       min        max      d    total    dead  logodds
   2162.8      1850       2400     12       39    0.31    -0.81
   2544.1      2500       2600     14       62    0.23    -1.23
   2695.3      2650       2700      5       32    0.16    -1.69
   2858.1      2750       2900      8       43    0.19    -1.48
   2998.0      2950       3000     11       76    0.14    -1.78
   3099.1      3060       3100      3       46    0.07    -2.66
   3196.9      3150       3200      4       32    0.12    -1.95
   3293.5      3250       3300      1       23    0.04    -3.09
   3473.7      3380       3500      6       73    0.08    -2.41
   3761.5      3600       4000      1       39    0.03    -3.64
```



จะเห็นว่ามี Linear relationship ชัดเจน เพื่อง่ายต่อการแปลความหมาย จึงจัดกลุ่มใหม่เป็นสองกลุ่ม ซึ่งสามารถกระทำได้อย่างเหมาะสมเพราะ Linear relationship ดังกล่าว และความหมายทางด้านการแพทย์และสาธารณสุข

. replace bwtg = .

(465 real changes made, 465 to missing)

. replace bwtg = 1 if bwt < 2500

(39 real changes made)

. replace bwtg = 0 if bwt >= 2500

(426 real changes made)

โปรดสังเกตว่าเรากำหนดให้ 1 = Low birth weight และ 0 = Normal birth weight แล้วดู Crude effect โดย

```
. cs dead bwtg, or

                 | bwtg                    |
                 |   Exposed   Unexposed   |     Total
-----------------+-------------------------+----------
           Cases |        12          53   |        65
        Noncases |        27         373   |       400
-----------------+-------------------------+----------
           Total |        39         426   |       465
                 |                         |
            Risk |  .3076923    .1244131   |  .1397849
                 |                         |
                 |      Point estimate     |  [95% Conf. Interval]
                 |-------------------------+----------------------
 Risk difference |         .1832792        |  .0350754    .3314829
      Risk ratio |         2.473149        |  1.449993    4.218275
 Attr. frac. ex. |         .5956573        |  .3103413    .7629363
 Attr. frac. pop |          .1099675       |
      Odds ratio |         3.127883        |  1.513083     6.47943  (Cornfield)
                 +-------------------------------------------------
                            chi2(1) =     9.98  Pr>chi2 = 0.0016
```

2.4 Crude effect of MAGE on DEAD

เนื่องจาก MAGE เป็น continuous variable จึงตรวจสอบว่ามีความสัมพันธ์เชิงเส้นกับ DEAD หรือไม่ คล้ายกับที่ทำกับ BWT ซึ่งพบว่ามี จึง dichotomize เพื่อสอดคล้องกับความรู้ที่มีเกี่ยวกับว่า teenage pregnancy มีความเสี่ยงสูง

```
. gen mageg = .
(465 missing values generated)

. replace mageg = 1 if mage < 20
(46 real changes made)

. replace mageg = 0 if mage >= 20
(419 real changes made)


. cs dead mageg, or

                 | mageg                    |
                 |   Exposed   Unexposed    |     Total
-----------------+--------------------------+----------
           Cases |         7           58   |        65
        Noncases |        39          361   |       400
-----------------+--------------------------+----------
           Total |        46          419   |       465
                 |                          |
            Risk |  .1521739    .1384248    |  .1397849
                 |                          |
                 |      Point estimate      |  [95% Conf. Interval]
                 |--------------------------+----------------------
 Risk difference |          .0137491        | -.0951896    .1226878
      Risk ratio |          1.099325        |  .5336421    2.264657
 Attr. frac. ex. |          .0903512        | -.8739152     .558432
 Attr. frac. pop |          .0097301        |
      Odds ratio |          1.117153        |  .4874978    2.567507  (Cornfield)
                 +-------------------------------------------------
                              chi2(1) =     0.07  Pr>chi2 = 0.7985
```

2.5 Crude effect of PLACE on DEAD

PLACE เป็น polytomous variable ก่อนอื่นทดสอบความสัมพันธ์อย่างหยาบก่อน พร้อมทั้งจะได้ค่าสัดส่วนเพื่อพรรณนาข้อมูลด้วย ดังนี้

```
. tab place dead, row chi2 exact

           |         dead
     place |         0          1 |     Total
-----------+----------------------+----------
         0 |       337         38 |       375
           |     89.87      10.13 |    100.00
-----------+----------------------+----------
         1 |        47         21 |        68
           |     69.12      30.88 |    100.00
-----------+----------------------+----------
         2 |        11          4 |        15
           |     73.33      26.67 |    100.00
-----------+----------------------+----------
         3 |         5          2 |         7
           |     71.43      28.57 |    100.00
-----------+----------------------+----------
     Total |       400         65 |       465
           |     86.02      13.98 |    100.00

          Pearson chi2(3) =  24.0179   Pr = 0.000
           Fisher's exact =                 0.000
```

โปรดสังเกตว่าเซลล์ตัวหนามีค่าน้อย ซึ่งจะส่งผลต่อ Model แน่นอน จึงต้องนำไปรวมกับกลุ่มอื่น โดยที่ความหมายไม่เปลี่ยนมากนักดังนี้

```
. replace place = 2 if place == 3
(7 real changes made)
```

```
. tab place dead, row chi2 exact
```

```
           |         dead
     place |         0          1 |     Total
-----------+----------------------+----------
         0 |       337         38 |       375
           |     89.87      10.13 |    100.00
-----------+----------------------+----------
         1 |        47         21 |        68
           |     69.12      30.88 |    100.00
-----------+----------------------+----------
         2 |        16          6 |        22
           |     72.73      27.27 |    100.00
-----------+----------------------+----------
     Total |       400         65 |       465
           |     86.02      13.98 |    100.00

          Pearson chi2(2) =  24.0035  Pr = 0.000
           Fisher's exact =                 0.000
```

แม้จะยังมีเซลล์ที่มีค่าน้อยอยู่ แต่ก็วิเคราะห์ตามนี้ไปก่อน ถ้าพบปัญหาจึงกลับมาจัดกลุ่มใหม่อีกครั้ง ลำดับต่อไปหาค่าสัดส่วนและ OR ดังนี้

```
. csi 21 38 47 337, or
```

```
                 |   Exposed   Unexposed  |     Total
-----------------+------------------------+----------
           Cases |        21          38  |        59
        Noncases |        47         337  |       384
-----------------+------------------------+----------
           Total |        68         375  |       443
                 |                        |
            Risk |  .3088235    .1013333  |  .1331828
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |         .2074902       |  .0935114     .321469
      Risk ratio |         3.047601       |  1.912134    4.857333
 Attr. frac. ex. |          .671873       |   .477024    .7941257
 Attr. frac. pop |         .2391412       |
      Odds ratio |         3.962486       |  2.156189    7.289677  (Cornfield)
                 +-------------------------------------------------
                             chi2(1) =    21.47  Pr>chi2 = 0.0000
```

```
. csi 6 38 16 337, or
```

```
                 |   Exposed   Unexposed  |     Total
-----------------+------------------------+----------
           Cases |         6          38  |        44
        Noncases |        16         337  |       353
-----------------+------------------------+----------
           Total |        22         375  |       397
                 |                        |
            Risk |  .2727273    .1013333  |  .1108312
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |         .1713939       | -.0171971     .359985
      Risk ratio |         2.691388       |  1.276449     5.67478
 Attr. frac. ex. |         .6284444       |  .2165766    .8237817
 Attr. frac. pop |          .085697       |
      Odds ratio |         3.325658       |  1.268007    8.771802  (Cornfield)
                 +-------------------------------------------------
                             chi2(1) =     6.19  Pr>chi2 = 0.0128
```

โดยสรุป ขั้นที่สองนี้ทำให้เราได้ข้อมูลเพื่อนำเสนอผลการศึกษา เฉพาะอย่างยิ่งค่าสัดส่วน (ร้อยละของการตายทารก) ค่า OR ค่า 95%CI และค่า p-value จากค่า p-value นี้ เราสามารถใช้กรองตัวแปรเข้า Model ได้ด้วย

ขั้นที่ 3 Stratified analysis

3.1 Effect of MALPRES on the association between AREA and DEAD

```
. cc dead area, by(malpres)

             malpres |       OR      [95% Conf. Interval]    M-H Weight
-----------------+------------------------------------------------
               0 |   .6711146     .3590862   1.254787      11.85116 (Cornfield)
               1 |      7.125     1.297704   37.58284      .4571429 (Cornfield)
-----------------+------------------------------------------------
           Crude |   1.269608     .7512221   2.145309               (Cornfield)
    M-H combined |   .9108184     .5136778   1.615001
-----------------+------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     5.91  Pr>chi2 = 0.0151

                  Test that combined OR = 1:
                              Mantel-Haenszel chi2(1) =      0.11
                                                Pr>chi2 =    0.7453
```

ผลชี้ว่า MALPRES เป็น Effect modifier จึงวิเคราะห์ข้อมูลเพื่อได้ค่าสถิตินำเสนอในผลการศึกษา ดังนี้

```
. tab area dead if malpres == 0, row chi2 exact

           |         dead
      area |         0          1 |     Total
-----------+----------------------+----------
         0 |       190         26 |       216
           |     87.96      12.04 |    100.00
-----------+----------------------+----------
         1 |       196         18 |       214
           |     91.59       8.41 |    100.00
-----------+----------------------+----------
     Total |       386         44 |       430
           |     89.77      10.23 |    100.00

          Pearson chi2(1) =   1.5385   Pr = 0.215
           Fisher's exact =                 0.265
   1-sided Fisher's exact =                 0.140
```

```
. tab area dead if malpres == 1, row chi2 exact

           |         dead
      area |         0          1 |     Total
-----------+----------------------+----------
         0 |         6          2 |         8
           |     75.00      25.00 |    100.00
-----------+----------------------+----------
         1 |         8         19 |        27
           |     29.63      70.37 |    100.00
-----------+----------------------+----------
     Total |        14         21 |        35
           |     40.00      60.00 |    100.00

          Pearson chi2(1) =   5.2932   Pr = 0.021
           Fisher's exact =                 0.039
   1-sided Fisher's exact =                 0.030
```

3.2 Effect of BWTG on the association between AREA and DEAD

```
. cc dead area, by(bwtg)

            bwtg |       OR      [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               0 |   1.339792     .7542532   2.379341      9.934272 (Cornfield)
               1 |        .49     .1209808    1.95487      2.564103 (Cornfield)
-----------------+-------------------------------------------------
           Crude |   1.269608     .7512221   2.145309               (Cornfield)
    M-H combined |   1.165453     .6827702   1.989367
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     1.62  Pr>chi2 = 0.2026

                  Test that combined OR = 1:
                              Mantel-Haenszel chi2(1) =      0.32
                                              Pr>chi2 =    0.5728
```

3.3 Effect of MAGEG on the association between AREA and DEAD

```
. cc dead area, by(mageg)

           mageg |       OR      [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               0 |   1.599555     .9133475   2.800384      9.661098 (Cornfield)
               1 |   .1571429     .0305953   .8318563      3.043478 (Cornfield)
-----------------+-------------------------------------------------
           Crude |   1.269608     .7512221   2.145309               (Cornfield)
    M-H combined |   1.254014     .7442425   2.112957
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     5.93  Pr>chi2 = 0.0149

                  Test that combined OR = 1:
                              Mantel-Haenszel chi2(1) =      0.75
                                              Pr>chi2 =    0.3867
```

3.4 Effect of PLACE on the association between AREA and DEAD

```
. cc dead area, by(place)

           place |       OR      [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               0 |   .7952381     .4090156   1.546609          9.52 (Cornfield)
               1 |       3.74      1.13241   12.16321      1.470588 (Cornfield)
               2 |   .7777778     .1316856   4.564086      1.227273 (Cornfield)
-----------------+-------------------------------------------------
           Crude |   1.269608     .7512221   2.145309               (Cornfield)
    M-H combined |   1.147927     .6675961   1.973853
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(2) =     4.84  Pr>chi2 = 0.0888

                  Test that combined OR = 1:
                              Mantel-Haenszel chi2(1) =      0.25
                                              Pr>chi2 =    0.6185
```

ขั้นที่ 4 Multivariable analysis : Logistic regression

จากผล Crude และ stratified analysis ในขั้นตอนที่ 2 และ 3 และองค์ความรู้เกี่ยวกับ Neonatal dead ทำให้เราทราบตัวแปรที่เป็น Candidate สำหรับ Initial model สามคำสั่งต่อไปนี้ เพื่อสร้างตัวแปรใหม่ซึ่งเป็น Interaction term ดังนี้

. gen a_mal = area * malpres

. gen a_mageg = area * mageg

. gen a_place = area * place

4.1. The initial model – the full model

```
. xi: logit dead  area malpres  bwtg mageg i.place  a_mal a_mageg i.a_place
i.place              Iplace_0-2   (naturally coded; Iplace_0 omitted)
i.a_place            Ia_pla_0-2   (naturally coded; Ia_pla_0 omitted)

Iteration 0:   log likelihood =  -188.1264
Iteration 1:   log likelihood = -158.90781
Iteration 2:   log likelihood = -151.19391
Iteration 3:   log likelihood = -150.81363
Iteration 4:   log likelihood =  -150.8124
Iteration 5:   log likelihood =  -150.8124

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(10)     =      74.63
                                                  Prob > chi2     =     0.0000
Log likelihood =  -150.8124                       Pseudo R2       =     0.1983

------------------------------------------------------------------------------
        dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.5413629   .4338548     -1.248   0.212      -1.391703    .3089768
     malpres |   .8913886   .9323986      0.956   0.339      -.9360792    2.718856
        bwtg |   1.117437   .4577921      2.441   0.015       .2201811    2.014693
       mageg |   1.439287   .6143028      2.343   0.019       .2352758    2.643299
    Iplace_1 |   .5058782   .6178105      0.819   0.413      -.7050082    1.716765
    Iplace_2 |   1.306483   .7715727      1.693   0.090      -.2057713    2.818738
       a_mal |   2.086607   1.073441      1.944   0.052      -.0172996    4.190513
     a_mageg |  -1.630821   1.032344     -1.580   0.114      -3.654178    .3925359
    Ia_pla_1 |   .8395218   .8137985      1.032   0.302      -.7554939    2.434537
    Ia_pla_2 |   .2971595   1.080756      0.275   0.783      -1.821084    2.415403
       _cons |   -2.38564   .2742555     -8.699   0.000      -2.923171   -1.848109
------------------------------------------------------------------------------
```

เก็บ Log-likelihood ไว้ใน Model 0 โดยสั่ง

. lrtest, saving(0)

โปรดสังเกต “xi” ก่อน “logit” คือคำสั่งเพื่อบอกให้ STATA รู้ว่ามี polytomous variables ใน model โดยที่“i.” ก่อน polytomous variable นั้น บอกให้ STATA สร้าง dummy variables ให้โดยอัตฯนมัติ (รายละเอียดคำสั่ง "logit" ดูได้จาก StataCorp (1999); หน้า 228-239 Volumn 2 : H-O)

จาก interaction terms (ตัวเอียง) ทุกตัว AREA*PLACE (ในวงรี) จะนำออกจาก Model ก่อนเพราะ p-valueสูงสุด โดยต้องนำออกไปทุกตัวที่เป็น dummy variables ของมัน ตามหลัก hierarchical well-formatted model

*4.2. Model without AREA*PLACE*

```
. xi: logit dead  area malpres  bwtg mageg i.place  a_mal a_mageg
```

```
i.place              Iplace_0-2   (naturally coded; Iplace_0 omitted)

Iteration 0:   log likelihood =  -188.1264
Iteration 1:   log likelihood = -159.76855
Iteration 2:   log likelihood = -151.72756
Iteration 3:   log likelihood = -151.36622
Iteration 4:   log likelihood = -151.36543

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(8)      =      73.52
                                                  Prob > chi2     =     0.0000
Log likelihood = -151.36543                       Pseudo R2       =     0.1954

------------------------------------------------------------------------------
        dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |   -.310655    .355062    -0.875   0.382      -1.006564    .3852538
     malpres |   .9825991   .9336054     1.052   0.293      -.8472338    2.812432
        bwtg |   1.040474   .4492353     2.316   0.021       .159989    1.920959
       mageg |    1.55312   .6067453     2.560   0.010       .3639209    2.742319
    Iplace_1 |   .9748812   .3862543     2.524   0.012       .2178366    1.731926
    Iplace_2 |   1.452425    .539014     2.695   0.007       .3959774    2.508873
       a_mal |   2.045493   1.077159     1.899   0.058      -.0657009    4.156686
     a_mageg |  -1.861715   1.004119    -1.854   0.064      -3.829751    .1063217
       _cons |  -2.487954   .2633255    -9.448   0.000      -3.004063   -1.971846
------------------------------------------------------------------------------
```

หา p-value จาก Likelihood Ratio test โดยเปรียบเทียบกับ Model แรก โดยสั่ง

```
. lrtest, using(0)
Logit:  likelihood-ratio test                     chi2(2)     =      1.11
                                                  Prob > chi2 =    0.5752
```

เก็บ Log-likelihood ของ Model นี้ ไว้ใน Model 1 โดยสั่ง

```
. lrtest, saving(1)
```

ผล LR test บอกเราว่า AREA*PLACE ไม่มีผลต่อ Model (p-value = 0.575) จึงสามารถนำจาก Model ได้ ต่อไป นำAREA*MAGEG ออกจาก Model

*4.3. Model without AREA*MAGE*

```
. xi: logit dead  area malpres  bwtg mageg i.place  a_mal
i.place              Iplace_0-2   (naturally coded; Iplace_0 omitted)

Iteration 0:   log likelihood =  -188.1264
Iteration 1:   log likelihood = -161.37036
Iteration 2:   log likelihood = -153.53573
Iteration 3:   log likelihood = -153.25674
Iteration 4:   log likelihood = -153.25615

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(7)      =      69.74
                                                  Prob > chi2     =     0.0000
Log likelihood = -153.25615                       Pseudo R2       =     0.1854

------------------------------------------------------------------------------
        dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.5590873   .3350725    -1.669   0.095      -1.215817    .0976429
     malpres |   .8722993   .9317147     0.936   0.349      -.9538279    2.698427
        bwtg |   1.047556   .4440794     2.359   0.018       .1771763    1.917935
       mageg |   .7140317   .4804309     1.486   0.137      -.2275954    1.655659
    Iplace_1 |   .9866509   .3859063     2.557   0.011       .2302884    1.743013
    Iplace_2 |   1.478023    .540374     2.735   0.006       .4189098    2.537137
       a_mal |   2.246689   1.073616     2.093   0.036       .1424403    4.350939
       _cons |  -2.382853   .2488897    -9.574   0.000      -2.870668   -1.895038
------------------------------------------------------------------------------


. lrtest, using(1)
Logit:  likelihood-ratio test                     chi2(1)     =      3.78
                                                  Prob > chi2 =    0.0518

. lrtest, saving(2)
```

AREA*MAGE ไม่มีผลต่อ Model (p-value = 0.052) จึงเอาออกไปได้ ต่อมาคือ AREA*MALPRES แต่ไม่สามารถนำออกไปได้เพราะมีผลต่อ Model (p-value = 0.036) ต่อไปพิจารณาตัวแปร Main effect ที่มี p-value สูงสุดคือ MALPRES แต่เป็นองค์ประกอบของ Interaction term จึงเอาออกไปไม่ได้ ดังนั้นจึงมีเพียง MAGEG ที่จะลองเอาออกไป

4.4. Model without MAGEG

```
. xi: logit dead  area malpres  bwtg i.place  a_mal
i.place              Iplace_0-2   (naturally coded; Iplace_0 omitted)

Iteration 0:   log likelihood =  -188.1264
Iteration 1:   log likelihood = -162.26531
Iteration 2:   log likelihood = -154.50657
Iteration 3:   log likelihood = -154.26029
Iteration 4:   log likelihood =  -154.2599

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(6)      =      67.73
                                                  Prob > chi2     =     0.0000
Log likelihood =  -154.2599                       Pseudo R2       =     0.1800

------------------------------------------------------------------------------
        dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.5157685   .3326557     -1.550   0.121      -1.167762    .1362246
     malpres |   .7955996    .925203      0.860   0.390      -1.017765    2.608964
        bwtg |   1.093564   .4429316      2.469   0.014       .2254335    1.961694
    Iplace_1 |   .8849724    .376117      2.353   0.019       .1477966    1.622148
    Iplace_2 |   1.365092   .5319488      2.566   0.010       .3224913    2.407692
       a_mal |   2.266141   1.069409      2.119   0.034        .170138    4.362143
       _cons |  -2.295464   .2367904     -9.694   0.000      -2.759565   -1.831363
------------------------------------------------------------------------------

. lrtest, using(2)
Logit:  likelihood-ratio test                     chi2(1)     =       2.01
                                                  Prob > chi2 =     0.1565
```

เราสามารถนำ MAGEG ออกไปได้ (p-value = 0.156) และที่เหลือใน Model ล้วนเป็น significant predictors ของ DEAD จึงถือว่า Model ข้างบนนี้คือ Final model.

ขั้นที่ 5 Assessing model adequacy: test for goodness of fit of the model

```
. lfit

Logistic model for dead, goodness-of-fit test

       number of observations =       465
 number of covariate patterns =        16
              Pearson chi2(9) =        17.32
                  Prob > chi2 =         0.0440
```

คำสั่ง “lfit” ให้ค่า Pearson หรือ Hosmer-Lemeshow goodness-of-fit tests แล้วแต่จะเลือก แต่ Hosmer-Lemeshow test เหมาะกว่า Pearson test เมื่อขนาดตัวอย่างไม่มากนัก เมื่อเทียบกับจำนวนตัวแปรใน Model (ดูรายละเอียดใน StataCorp (1999); หน้า 209-211 Volumn 2 : H-O) ผลข้างต้น p-value = 0.044 แสดงว่า Model เข้าได้กับข้อมูลไม่ดีนัก

ถ้าลอง Fit Model ใหม่โดยใส่ BWT และ MAGE เป็นข้อมูลต่อเนื่องจะได้ p-value = 0.465 ซึ่งแสดงว่า Model เข้าได้กับข้อมูลดี แต่แปลความหมายได้เข้าใจยาก กอปรกับค่า OR ไม่ต่างกันมากนักกับ Modelข้างต้น เราจึงเลือก Model ข้างต้นสำหรับการวิจัยนี้

Model assessment ยังสามารถทำได้หลายวิธีจากแนวทางที่เสนอโดย Hosmer and Lemeshow (1989) โดยมีเป้าหมายหลักเพื่อหา influence observations คือค่าข้อมูลที่ทำให้การประมาณค่า OR แปรเปลี่ยนมาก เป็นแนวทางหนึ่งที่ช่วยในการหาทางให้ Model เข้าได้กับข้อมูลได้ดีขึ้น STATA ทำกระบวนการดังกล่าวได้ (ดูรายละเอียดใน StataCorp, 1999 Volume 2, หน้า 200-222).

ขั้นที่ 6 Obtaining measure of associations from the model

หาค่า Odds ratios พร้อมช่วงเชื่อมั่นและ p-value จาก Model โดยใช้คำสั่ง “logistic” จากนี้ค่าของ BWTG และ PLACE (ตัวเอียง) เท่านั้นที่ใช้ได้

```
. xi: logistic dead  area malpres  bwtg i.place  a_mal
i.place              Iplace_0-2   (naturally coded; Iplace_0 omitted)

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(6)      =      67.73
                                                  Prob > chi2     =     0.0000
Log likelihood =  -154.2599                       Pseudo R2       =     0.1800

------------------------------------------------------------------------------
        dead | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |   .5970416   .1986093     -1.550   0.121       .3110624    1.145939
     malpres |   2.215769   2.050036      0.860   0.390       .3614018    13.58497
        bwtg |   2.984892   1.322103      2.469   0.014       1.252866     7.11136
    Iplace_1 |   2.422917   .9113004      2.353   0.019       1.159277    5.063957
    Iplace_2 |   3.916082   2.083155      2.566   0.010       1.380563     11.1083
       a_mal |   9.642116   10.31136      2.119   0.034       1.185468    78.42503
------------------------------------------------------------------------------
```

จากที่มี Interaction ระหว่าง AREA กับ MALPRES เราต้องหาค่าเหล่านั้นของ AREA ต่อ DEAD ในแต่ละกลุ่มของ MALPRES ผลของ AREA ในกลุ่ม MALPRES = 0 สามารถใช้จาก Output ข้างบนตรงตัวแปร “area” ผลของ AREA ในกลุ่ม MALPRES = 1 หาได้ดังนี้

```
. lincom area + a_mal

 ( 1)  area + a_mal = 0.0

------------------------------------------------------------------------------
        dead | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   5.756744   5.836508      1.726   0.084       .7891896    41.99257
------------------------------------------------------------------------------
```

ขั้นที่ 7 Summarize findings

จากมารดาทั้งหมด 465 ราย มี 65 รายที่ทารกตายใน 28 วันหลังคลอด คิดเป็นอัตราตาย 14.0% (95%CI: 10.8% ถึง 17.1%) ตามรายละเอียดในตารางที่ 1

เมื่อควบคุมผลกระทบของปัจจัยอื่นๆ แล้ว พบว่าความสัมพันธ์ระหว่างการเยี่ยมก่อนและหลังคลอดโดย อสม. กับการตายของทารกใน 28 วันหลังคลอด เมื่อเปรียบเทียบกับการให้บริการโดยเจ้าหน้าที่สาธารณสุข ณ สถานบริการสาธารณสุข พบว่า ความสัมพันธ์ดังกล่าว ขึ้นอยู่กับว่า ทารกมีการคลอดท่าผิดปกติหรือไม่ อย่างมีนัยสำคัญทางสถิติ (p-value = 0.034) กล่าวคือ ถ้าทารกคลอด***ท่าปกติ*** พบว่าทารกในพื้นที่มีการ**ให้บริการตามปกติ**โดยเจ้าหน้าที่สาธารณสุขที่สถานบริการสาธารณสุข ที่มีความเสี่ยงต่อการตายภายใน 28 วันหลังคลอดสูงเป็น 1.7 เท่าของทารกในพื้นที่ที่**มีการเยี่ยมโดย อสม.** (95%CI: 0.9 ถึง 3.2) ในทางตรงข้าม ถ้าทารกคลอด***ท่าผิดปกติ*** พบว่าทารกในพื้นที่ที่**มีการเยี่ยมโดย อสม.** มีความเสี่ยงต่อการตายภายใน 28 วันหลังคลอดสูงเป็น 5.8 เท่าของทารกในพื้นที่ที่มีการ**ให้บริการตามปกติ** โดยเจ้าหน้าที่สาธารณสุขที่สถานบริการสาธารณสุข (95%CI: 0.8 ถึง 42.0)

ตารางที่ 1 ค่า Crude และ adjusted odds ratio ของตัวแปรต่าง ๆ ต่อการตายของทารก

Factors	**จำนวน**	**ตาย (%)**	**Crude OR**	**Adjusted OR**	**95%CI for adjusted OR**	**p-value**
1. การเยี่ยมบ้านโดย อสม. ต่อการตายของทารก จำแนกตามท่าคลอด						0.034
1.1 ท่าคลอดผิดปกติ						
เยี่ยมบ้านโดย อสม	27	70	7.1	5.8	0.8 to 42.0	
รับบริการที่สถานีอนามัย	8	25	1.0	1.0		
1.2 ท่าคลอดปกติ						
เยี่ยมบ้านโดย อสม	214	8	0.7	0.6	0.3 to 1.1	
รับบริการที่สถานีอนามัย	216	12	1.0	1.0		
2. น้ำหนักแรกเกิด						0.014
น้อยกว่า 2,500 กรัม	39	31	3.1	3.0	1.2 to 7.1	
2,500 กรัมหรือมากกว่า	426	12	1.0	1.0		
3. สถานที่คลอด						0.010
โรงพยาบาล	375	10	1.0	1.0		
สถานีอนามัย	68	31	4.0	2.4	1.2 to 5.1	
ที่บ้านหรือขณะเดินทาง	22	27	3.3	3.9	1.4 to 11.1	

****

2. ตารางต่อไปนี้ เป็นข้อมูลการสูบบุหรี่และผลการทดสอบการหายใจจำแนกตามกลุ่มอายุสำหรับคนงานในโรงงานแห่งหนึ่ง

		ผลการทดสอบการหายใจ	
อายุ	การสูบบุหรี่	ปกติ	ผิดปกติ
< 40	ไม่เคย	577	34
	เคยและกำลังสูบ	682	57
40 – 59	ไมเคย	164	4
	เคยและกำลังสูบ	245	74

2.1) ถ้าหากไม่คำนึงถึงอายุแล้ว ความสัมพันธ์ระหว่างการสูบบุหรี่กับผลการทดสอบการหายใจเป็นอย่างไร

2.2) ความสัมพันธ์ตามข้อ 2.1 จำแนกตามแต่ละกลุ่มอายุเป็นอย่างไร

2.3) อายุเป็น Confounder หรือ Effect Modifier หรือไม่อย่างไร

2.4) พิจารณาจากผลตามข้อ 2.3 แล้วท่านมีแนวทางอื่นใด ในการวิเคราะห์ข้อมูลชุดนี้ และจงวิเคราะห์ตามแนวทางดังกล่าว

2.5) เขียนสรุปผลการวิเคราะห์

3. มะเร็งหลอดอาหาร (Esophageal cancer) เป็นหนึ่งในโรคมะเร็งที่พบได้บ่อยในภาคใต้ของไทย ได้มีการศึกษาโดยวิธี case-control study เพื่อหาปัจจัยเสี่ยงของโรคมะเร็งหลอดอาหาร โดยมีสมมติฐานว่า “ผู้ป่วยมะเร็งหลอดอาหารมีประวัติการสูบบุหรี่ ดื่มสุรา และทำอุตสาหกรรมยางพารา มากกว่าเพื่อนบ้านที่มีอายุและเพศเดียวกัน” โดย case หมายถึง ผู้ป่วยที่ตรวจยืนยันทางเนื้อเยื่อแล้วว่าเป็นมะเร็งที่ได้รับการวินิจฉัยในระหว่าง 1 มกราคม – 31 ตุลาคม 2531 ส่วน Control นั้นเลือกจากเพื่อนบ้านของ case ข้อมูลได้มาจากการใช้แบบสอบถามสัมภาษณ์

ข้อมูลแสดงไว้ในภาคผนวก 1 ข้อ 2 ชื่อแฟ้มเพื่อการอ้างอิงคือ “ESOPH_CA.DTA”

โครงสร้างของแฟ้มเป็นดังนี้

คอลัมน์	ชื่อตัวแปร	คำอธิบาย	รหัส
1-3	ID	เลขที่ผู้ป่วย	1 = case
4	CASE	กลุ่มที่ศึกษา	1 = case
			0 = control
5	AlC1	การดื่มสุราในปัจจุบัน	1 = ดื่ม
			2 = ไม่ดื่ม
6	ALC2	การดื่มสุราในอดีต	1 = ดื่ม
			2 = ไม่ดื่ม
7	SMK	การสูบบุหรี่ในปัจจุบัน	1 = สูบ
			2 = ไม่สูบ
8	PARA	การอยู่ในกระบวนการทำอุตสาหกรรมยางพารา	1 = อยู่
			2 = ไม่อยู่
9	CIGA	ระดับการสูบบุหรี่	0 = ไม่สูบ
			1 = 1-9 มวน/วัน
			2 = 10-19 มวน/วัน
			3 = 20+ มวน/วัน

จงวิเคราะห์ข้อมูล พร้อมเขียนสรุปผลการศึกษาที่สามารถใช้เป็นส่วนหนึ่งของรายงานการศึกษาฉบับสมบูรณ์ภายใต้หัวข้อ “ผลการศึกษา” ได้โดยให้มีความถูกต้องเหมาะสม

เอกสารอ้างอิง

บัณฑิต ถิ่นคำรพ. (2541). ความสำคัญและความจำเป็นของการวิเคราะห์ข้อมูลตัวแปรเชิงพหุ. *วารสารระบาดวิทยาภาคตะวันออกเฉียงเหนือ*. 3(3) :20-25.

Concato, J., Feinstein, A.R., and Holford, T.R., (1993). The risk of determining risk in multivariable models. *Annals of Internal Medicine.* **118**:201-210.

Feinstein, A.R. (1996). *Multivariable analysis: an introduction.* Yale university Press: New Haven.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions*. 2nd edition. New York: John Willey & Sons.

Guyatt, G., Jaeschke, R., Heddle, N., Cook, D., Shannon, H., and Walter S. (1995). Interpreting study results: confidence intervals. *Canadian Medical Association Journal.* 152:169-173.

Harrell, F.E., Lee, K.L., Califf, R.M., Pryor, D.B., and Rosati, R.A. (1984). Rgression modelling strategies for improve prognostic modelling. *Statistics in Medicine.* **3**:143-152.

Hosmer, D.W., and Lemeshow, S. (1989). *Applied Logistic Regression*. New York: John Wiley & Sons.

Hsieh, F. Y. (1989). Sample size tables for logistic regression. *Stat Med* **8,** 795-802.

Hsieh, F. Y., Bloch, D. A., & Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression. *Stat Med* **17,** 1623-34.

Jaeschke, R., Guyatt, G., Shannon, H., Walter, S. Cook, D. Heddle, N. (1995). Assessing the effects of treatment: measures of association . *Canadian Medical Association Journal*. 152: 351-357

Kleinbaum, D.G. (1994). *Logistic Regression: A self-learning text*. New York Springer-Verlag.

Kleinbaum, D.G., Kupper, L.L., and Morgenstern, H. (1982). *Epidemiologic research: principles and qualitative methods*. London: Lifetime Learning Publications.

Lang, TA., Secic, M. (1997). *How to report statistics in medicine: annotated guidelines for authors*, editors, and reviewers. Philadelphia: American College of Physician.

Mazumdar, M., and Glassman, J. R. (2000). Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. *Stat Med* **19,** 113-32.

StataCorp. (1999). *Stata statistical software: Release 6.0*. College Station. TX: Stata Corporation.

ภาคผนวก 1

เฉลย
แบบฝึกหัดที่ 1

ข้อ 1

1.2.5 ข้อมูลมีทั้งสิ้น 400 records

ข้อ 2

2.1

ข้อมูลจากที่แจงนับลงในตารางแล้ว

EXERCISE+

	CHD+	CHD-	Total
SMK+	50	50	100
SMK-	50	50	100
Total	100	100	200

EXERCISE-

	CHD+	CHD-	Total
SMK+	25	10	35
SMK-	25	40	65
Total	50	50	100

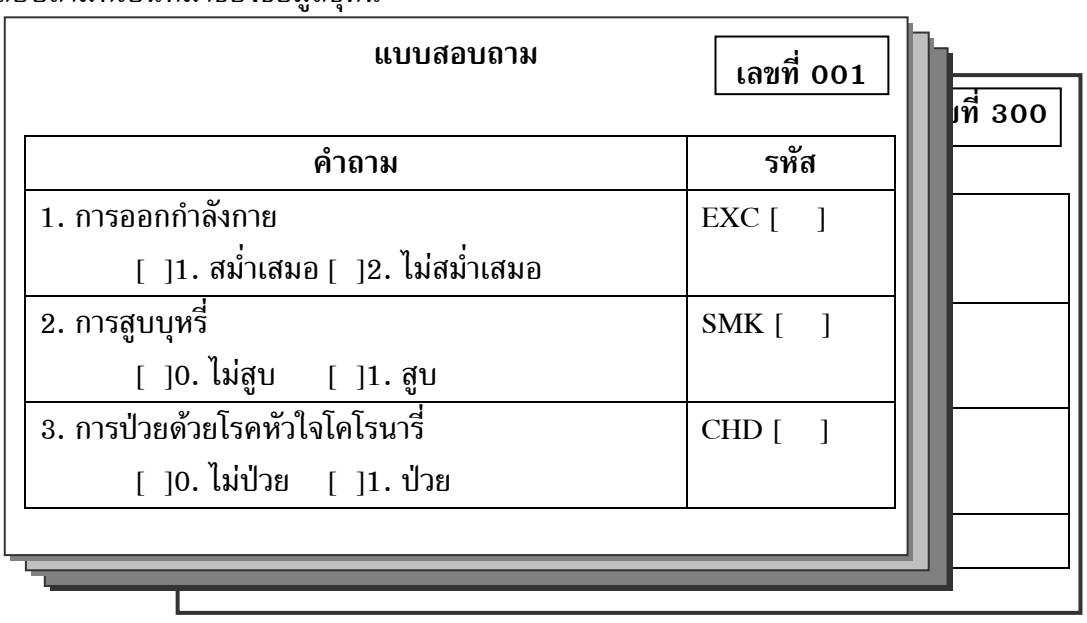
กำหนดโครงสร้างแฟ้มข้อมูล

ตัวแปร	รหัส ความหมาย	
EXC	1=Always,	2= Not always
SMK	0=No,	1=Yes
CHD	0=No,	1=Yes

ข้อมูลที่ต้องป้อนลงใน STATA

exc	smk	chd	n
1	1	1	50
1	1	0	50
1	0	1	50
1	0	0	50
2	1	1	25
2	1	0	10
2	0	1	25
2	0	0	40

ลักษณะแบบสอบถามที่เป็นที่มาของข้อมูลชุดนี้



แบบสอบถาม **เลขที่ 001**

คำถาม	รหัส
1. การออกกำลังกาย [ ]1. สม่ำเสมอ [ ]2. ไม่สม่ำเสมอ	EXC [ ]
2. การสูบบุหรี่ [ ]0. ไม่สูบ [ ]1. สูบ	SMK [ ]
3. การป่วยด้วยโรคหัวใจโคโรนารี่ [ ]0. ไม่ป่วย [ ]1. ป่วย	CHD [ ]

2.3 คำสั่งคือ .cc chd smk, by(exc)

2.5 คำสั่งคือ .logistic chd smk exc

ข้อ 3.

3.1) ตัวแปรที่เป็น Dichotomous คือ DEAD AREA MALPRES

3.2) ตัวแปรที่เป็น Continuous คือ BWT MAGE DCHILD

3.3) ตัวแปรตาม (D) คือ DEAD

3.4) ตัวแปรต้น (E) คือ AREA

3.5) ตัวแปรต้น (C) คือ MALPRES BWT MAGE DCHILD

3.6)

เลขที่...*1*....

แบบสอบถาม

คำถาม	สำหรับลงรหัส
1. สถานภาพการรอดชีพของทารกเมื่ออายุครบ 28 วัน	DEAD[*1*]
[X]1. ตาย [ ]0. รอด	
2. พื้นที่ที่มารดาอาศัย	AREA[*1*]
[X]1. พื้นที่ทดลอง [ ]0. พื้นที่ควบคุม	
3. ท่าของทารกขณะคลอด	MALPRES[*0*]
[ ]1. ผิดปกติ [X]0. ปกติ	
4. นำหนักทารกแรกเกิด.....*2600*.....กรัม	BWT[*2*][*6*][*0*][*0*]
5. อายุมารดา.....*30*.....ปี	AGE[*3*][*0*]
6. จำนวนเด็กเกิดมีชีพของมารดา ซึ่งขณะนี้ เด็กนั้นเสียชีวิตแล้ว....*0*.....คน	DCHILD[*0*]

3.7) แบบสอบถามชุดนี้ มีทั้งหมดกี่แผ่น 465 แผ่น

3.8.1 $OR_C = 1.27$

3.9.1.1 $OR_1 = 0.67$

$OR_2 = 7.13$

3.9.1.3 $OR_1 = 0.67$

$OR_2 = 7.13$

$OR_C = 1.27$

$OR_{MH} = 0.91$

Woolf's test p-value = 0.015

3.9.1.4 สรุปผลที่ได้

จากผล Stratified analysis พบว่า ความสัมพันธ์ระหว่าง DEAD กับ AREA นั้น มี Interaction effect โดยมี MALPRES เป็น Effect Modifier

กล่าวคือการตายของทารกภายใน 28 วันหลังคลอด ในพื้นที่ทั้งสองที่เปรียบเทียบกันนั้น ขึ้นอยู่กับว่าทารกนั้น คลอดท่าผิดปกติหรือไม่

ถ้าทารกคลอดท่าปกติ พบว่าทารกในพื้นที่ควบคุม มีความเสี่ยงต่อการตายภายใน 28 วันหลังคลอดสูงเป็น 1.49 เท่าของทารกในพื้นที่ทดลอง (95%CI: 0.80 ถึง 2.79)

ถ้าทารกคลอดท่าผิดปกติ พบว่าทารกในพื้นที่ทดลอง มีความเสี่ยงต่อการตายภายใน 28 วันหลังคลอดสูงเป็น 7.13 เท่าของทารกในพื้นที่ควบคุม (95%CI: 0.93 ถึง 67.28)

โปรดสังเกต: *1. ควรนำเสนอช่วงความเชื่อมั่น (95%CI.) ด้วยเสมอ โดยใช้ผลจากคอมพิวเตอร์*

2. ถ้าหากค่า OR มีค่าน้อยกว่า 1 และหากต้องการแปลความหมายเป็นรูปประโยค ควรกลับข้าง การเปรียบเทียบ และคำนวณค่า OR ใหม่ โดยเพียงนำค่าเดิมไปหารค่า 1 หรือ 1/OR นั่นเอง เช่น 1/0.69 = 1.49 เป็นต้น แต่ถ้าไม่นำเสนอประโยคแปลความหมาย ก็สามารถนำเสนอตัวเลขค่า OR โดยตรง แม้จะมีค่าน้อยกว่า 1 เพื่อให้เห็นว่ามีผลด้านป้องกัน (protective effect) หรือถ้ามากกว่า 1 ก็ชี้ว่ามีผลด้านทำให้เกิดความเสี่ยง (risk effect)

3.9.2.1 $OR_1 = 0.80$

$OR_2 = 2.44$

3.9.2.3 OR_1 = 0.80

OR_2 = 2.44

OR_C = 1.27

OR_{MH} = 1.15

Woolf's test p-value = 0.067

ผลการทดสอบความแตกต่างโดย Woolf's test for heterogeneity of Odds Ratios ได้ p-value = 0.057 ==> OR_1และOR_2 แตกต่างกันอย่างไม่มีนัยสำคัญ แสดงว่า ไม่มี Interaction effect จึงใช้ค่า OR_{MH} = 1.15 ซึ่งเหมาะสมที่จะใช้อธิบายความสัมพันธ์มากกว่าใช้ OR_C จากนั้นพิจารณาเปรียบเทียบค่า OR_{MH} กับ OR_C พบว่า ค่าใกล้เคียงกัน แสดงว่า ไม่มี Confounding effect จึงสรุปว่า DCHILD ไม่เป็นทั้ง Effect modifier และ Confounder ของความสัมพันธ์ระหว่าง DEAD กับ AREA

3.9.2.4 สรุปผลที่ได้

จำนวนเด็กเกิดมีชีพของมารดาซึ่งขณะนี้ เด็กนั้นเสียชีวิตแล้ว ไม่มีผลต่อความสัมพันธ์ระหว่างการตายของทารกภายใน 28 วันหลังคลอด ในพื้นที่ทั้งสองที่เปรียบเทียบกัน

3.9.3.3 OR_C = 1.27

OR_1 = 0.49

OR_2 = 2.1

OR_3 = 0.8

Woolf's test p-value = 0.154 แสดงว่า ไม่มี Interaction effect

OR_{MH} = 1.09 ใกล้เคียงกับ OR_C =1.27 แสดงว่า ไม่มี Confounding effect

3.9.3.4 ดังนั้น BWTG ไม่เป็นทั้ง Effect modifier และ Confounder **ของความสัมพันธ์ระหว่าง DEAD กับ AREA**

3.10 ในการอธิบายความสัมพันธ์ระหว่าง AREA กับ DEAD เราทราบว่า หากใช้ผลจากการวิเคราะห์แบบ Bivariate analysis ตอบคำถามวิจัยนี้ จะเป็นองค์ความรู้ที่ผิด เราทราบจากการวิเคราะห์แบบ stratified analysis ว่า MALPRES เป็น effect modifier และทราบว่า DCHILD และ BWT ไม่เป็นทั้งทั้ง Effect modifier และ Confounder ของความสัมพันธ์ระหว่าง DEAD กับ AREA รูปแบบความสัมพันธ์ที่พบเหล่านี้ มีความสำคัญอย่างมากในการวิเคราะห์แบบ multivariable analysis กล่าวคือ เราทราบว่าต้องสร้าง Model อย่างไร เช่นกรณีนี้ เราทราบว่า MALPRES จะต้องนำเข้าใน Model ในรูป Interaction term กับ AREA (จะกล่าวในรายละเอียดต่อไป) ส่วน DCHILD และ BWT อาจไม่จำเป็นต้องนำเข้า Model ก็ได้ อย่างไรก็ตาม ผลร่วมกัน (joint effect) ระหว่าง AREA MALPRES DCHILD และ BWT ต่อ DEAD อาจมีอยู่ ซึ่งต้องมีวิธีการวิเคราะห์ผลดังกล่าวด้วย เป็นต้น

เฉลย
แบบฝึกหัดที่ 2

ข้อ 1.

1.1) Logistic Function :

$$f(z) = \frac{1}{1+e^{-z}}$$

1.2) Logistic Model :

$$P(X) = \frac{1}{1+e^{-(a+\sum b_i X_i)}}$$

1.3) Logit transformation :

$$\ln\left[\frac{P(X)}{1-P(X)}\right] = a + \sum b_i x_i \quad \text{หรือ}$$

$$\text{Logit } P(\mathbf{X}) = a + \sum b_i X_i$$

โปรดสังเกต *P(X) ใน Model แรก (X ตัวธรรมดา) หมายถึง ค่าความน่าจะเป็น มีค่าเท่ากันกับค่า P(X) ใน Model ตามข้อ 1.2 แต่ P(**X**) ใน Model ที่สอง (**X** ตัวทึบ) หมายถึงค่า Odds*

ข้อ 2.

2.1) ค่า OR หาได้จาก Logistic Model ที่ได้จากข้อมูลการศึกษาทุกประเภทต่อไปนี้ คำตอบจึงเป็นดังนี้

☒Cohort study

☒Case-control study

☒Cross-sectional Study

2.2) ค่า OR หาได้โดย

(i) สูตรที่ใช้ $OR_{x1,Xo} = Exp\{\sum[b_i(X_{1i} - X_{0i})]\}$

(ii) ทารกในพื้นที่ทดลอง เป็น X_1 คือ AREA=1
ทารกในพื้นที่ควบคุม เป็น X_0 คือ AREA=0

(iii) แทนค่า OR_{x1,X_o}= Exp{[-1.946 + 0.239(1)] -[-1.946 + 0.239(0)]}

= Exp[0 + 0.239)]

= Exp(0.239)

= $e^{(0.239)}$

= 1.27

(iv) ทารกที่อยู่ในพื้นที่ทดลอง มีโอกาสตายภายใน 28 วันหลังคลอด สูงเป็น 1.27 เท่าของทารกที่อยู่ในพื้นที่ควบคุม กล่าวอีกนัยหนึ่ง ทารกที่อยู่ในพื้นที่ที่มีการเยี่ยมก่อนและหลังคลอดโดย อาสาสมัครสาธารณสุข (อสม.) มีความเสี่ยงต่อการตายใน 28 วันหลังคลอด สูงเป็น 1.27 เท่าของทารกที่อยู่ในพื้นที่ที่มีการให้บริการโดยเจ้าหน้าที่สาธารณสุขตามปกติที่สถานบริการสาธารณสุข

2.3) ข้อมูลในรูปตาราง 2 x 2

		DEAD 1	DEAD 0
AREA	1	37	204
	0	28	196

OR = (37 x 196) / (204 x 28) = 1.27

2.4) ค่าที่ได้จาก Logistic Model กับค่าที่ได้จากตาราง 2 x 2 เหมือนกัน

2.5) Logistic Model ข้างต้น เขียนในรูป Logit transformation ได้ดังนี้

- รูปแบบติดค่าสัมประสิทธิ์ $\text{Logit } P(\mathbf{X}) = a + b_1\text{AREA}$

- รูปแบบแทนค่าสัมประสิทธิ์ $\text{Logit } P(\mathbf{X}) = -1.94591 + .2387081(\text{AREA})$

ข้อ 3

3.2) ค่า OR หาได้โดย

(i) สูตรที่ใช้ $OR_{x1,X_o} = \text{Exp}\{\Sigma[b_i(X_{1i} - X_{0i})]\}$

(ii) ทารกในพื้นที่ทดลอง โดยให้ DCHILD คงที่

ดังนั้น X_1 คือ (AREA=1, DCHILD ไม่ระบุค่า แต่ให้คงที่)

ทารกในพื้นที่ควบคุม โดยให้ DCHILD คงที่ เป็น X_0 คือ

ดังนั้น X_0 คือ (AREA=0, DCHILD ไม่ระบุค่า แต่ให้คงที่)

(iii) แทนค่า OR_{x1,X_o} = Exp{$\sum$[-2.255301 (0 – 0) + 0.1419709 (1 – 0)
+ 1.321847(0 – 0)]}
= Exp(0.1419709)
= 1.15

(iv) แปลความหมาย “เมื่อควบคุมผลกระทบจากจำนวนเด็กเกิดมีชีพของมารดาซึ่งขณะนี้เด็กนั้นเสียชีวิตแล้ว พบว่าทารกในพื้นที่ที่มีการเยี่ยมโดย อสม. มีความเสี่ยงต่อการตายภายใน 28 วันหลังคลอดสูงเป็น 1.15 เท่าของทารกในพื้นที่ที่มีการให้บริการตามปกติโดยเจ้าหน้าที่สาธารณสุขที่สถานบริการสาธารณสุข”

3.3) ค่าที่ได้จาก Logistic Model กับค่าที่ได้จากStratified analysis จากแบบฝึกหัดที่ 1 ข้อ 3.9.2.3 นั้นเท่ากัน

3.4) Logistic Model ข้างต้น เขียนในรูป Logit transformation ได้ดังนี้

Logit P(**X**) = -2.255301 + 0.1419709(AREA) + 1.321847 (DCHILD)

ข้อ 4. ท่านไม่ได้ถูกคาดหวังว่าจะตอบได้ในตอนนี้ แต่คนทั่วไปมักเข้าใจว่า และตอบว่าเขียนได้ดังนี้

Logit P(**X**) = a + b_1AREA + b_2MALPRES

ซึ่งไม่ถูกต้อง (ดูคำตอบที่ถูกที่เฉลยแบบฝึกหัดที่ 3 ข้อ 2 ข้อย่อยที่ (2.1)

เฉลย
แบบฝึกหัดที่ 3

ข้อ 1.

1.1 ให้ BWTG = 2 เป็นกลุ่มอ้างอิง (Reference group)

	ตัวแปรใหม่ (Dummy Variable)	
ตัวแปรเดิม	**BWTD1**	**BWTD3**
BWTG = 1	1	0
BWTG = 2	0	0
BWTG = 3	0	1

1.3.1) OR สำหรับ ทารกน้ำหนักแรกเกิดต่ำกว่าปกติ โดยใช้ทารกน้ำหนักแรกเกิดปกติเป็นกลุ่มอ้างอิง

(i) สูตรที่ใช้ $OR_{x1,Xo} = Exp\{\sum[b_i(X_{1i} - X_{0i})]\}$

(ii) ทารกน้ำหนักแรกเกิดสูงกว่าปกติ
X_1 คือ (BWTD3=1, ตัวแปรอื่นไม่ระบุค่าแต่ให้คงที่)
ทารกน้ำหนักแรกเกิดปกติ
X_0 คือ (BWTD1=0, ตัวแปรอื่นไม่ระบุค่าแต่ให้คงที่)

(iii) ทำเช่นเดียวกันกับข้อ 3.2 (iii) ในแบบฝึกหัดที่ 2

$OR_{x1,Xo} = Exp(0.6092406)$
$= 1.84$

1.3.2) OR สำหรับ ทารกน้ำหนักแรกเกิดสูงกว่าปกติ โดยใช้ทารกน้ำหนักแรกเกิดปกติเป็นกลุ่มอ้างอิง

(i) สูตรที่ใช้ $OR_{x1,Xo} = Exp\{\sum[b_i(X_{1i} - X_{0i})]\}$

(ii) ทารกน้ำหนักแรกเกิดต่ำกว่าปกติ
X_1 คือ (BWTD1=1, ตัวแปรอื่นไม่ระบุค่าแต่ให้คงที่)
ทารกน้ำหนักแรกเกิดปกติ
X_0 คือ (BWTD1=0, ตัวแปรอื่นไม่ระบุค่าแต่ให้คงที่)

(iii) ทำเช่นเดียวกันกับข้อ 3.2 (iii) ในแบบฝึกหัดที่ 2

$OR_{x1,Xo} = Exp(-0.8628492)$
$= 0.42$

ข้อ 2.

2.1) จงเขียน Model ในรูปของ Logit transformation

$$\text{Logit } P(\mathbf{X}) = a + b_1\text{AREA} + b_2\text{MALPRES} + b_3\text{AREA*MALPRES}$$

2.3) กรณี MALPRES = 0, $OR_{(AREA1,0)} = 0.67$

กรณี MALPRES = 1, $OR_{(AREA1,0)} = 7.12$

2.4) OR ที่ได้นี้ กับที่ได้จากการวิเคราะห์โดย Stratified analysis มีค่าเท่ากัน

2.5) OR ที่ได้จากการวิเคราะห์โปรแกรมสถิติทั่วไปเป็นค่า Exponential ของค่า Coefficient จากโมเดล ไม่สามารถใช้ได้ถ้าหากมี Interaction effect

3.1) Term นี้เรียกว่า Third order term

3.2)

$$\text{Logit } P(\mathbf{X}) = a + b_1\text{AREA} + b_2\text{MALPRES} + b_3\text{MAGE} + b_4\text{AREA*MALPRES} + b_5\text{AREA*MAGE} + b_6\text{MALPRES*MAGE} + b_7\text{AREA*MALPRES*MAGE}$$

เฉลย
แบบฝึกหัดที่ 4

ข้อ 1.

1.1.1) คำนวณค่า OR. และช่วงความเชื่อมั่น ที่ระดับ 95% โดยคำนวณด้วยตนเอง

		DEAD 1	DEAD 0
AREA	1	37	204
	0	28	196

OR. = (37 x 196) / (204 x 28) = 1.27

$$95\%CI.OR. = 1.27 EXP\left[\pm 1.96\sqrt{\frac{1}{37}+\frac{1}{204}+\frac{1}{28}+\frac{1}{196}}\right]$$

= 0.75 – 2.15

หมายเหตุ: ใช้คำสั่ง STATA คือ .cc area dead <ENTER>

1.1.2 ผลการวิเคราะห์โดยใช้ STATA

```
. use logistic.dta

. logistic dead area

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(1)      =       0.79
                                                  Prob > chi2     =     0.3745
Log likelihood = -187.73214                       Pseudo R2       =     0.0021

------------------------------------------------------------------------------
        dead | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |   1.269608    .342429      0.885   0.376       .7483246    2.154017
------------------------------------------------------------------------------
```

1.2.3 สรุปผลการวิเคราะห์ :

ทารกที่อยู่ในพื้นที่ที่มีการเยี่ยมก่อนและหลังคลอดโดย อาสาสมัครสาธารณสุข (อสม.) มีความเสี่ยงต่อการตายใน 28 วันหลังคลอด สูงเป็น 1.27 เท่าของทารกที่อยู่ในพื้นที่ที่มีการให้บริการโดยเจ้าหน้าที่สาธารณสุขตามปกติที่สถานบริการสาธารณสุข

1.2) Logistic Regression Model ที่ใช้สำหรับวิเคราะห์ความสัมพันธ์ ตามข้อ 1.1 เขียนในรูป Logit

transformation ได้ดังนี้

$$\text{Logit } P(\mathbf{X}) = -1.946 + 0.239\text{AREA}$$

1.2.1) แสดงการคำนวณด้วยตนเอง ได้ดังนี้

ค่า OR. = Exp(0.239) = 1.27

ค่า 95% CI OR. = Exp[0.239 $\pm$ 1.96(0.27)]

เมื่อ Standard Error (SE.) = 0.27 ซึ่งได้จาก STATA ด้วยคำสั่ง .logit dead area

= Exp(-0.29) ถึง Exp(0.77)

= 0.75 ถึง 2.16

1.2.2) คำนวณโดยใช้ STATA

คำสั่งคือ .logistic dead area

ค่า OR. = 1.27

ค่า 95% CI OR. = 0.75 - 2.15

1.3) เปรียบเทียบค่าที่ได้จาก 1.1 กับ 1.2 พบว่า ค่าที่ได้เท่ากัน

ข้อ 2.

2.1) ใช้ STATA ในการ Fit Model

คำสั่งที่ใช้คือ .logit dead mage

ผลที่ได้ เป็นดังนี้

```
------------------------------------------------------------------------------
    dead |      Coef.         Std. Err.              z        P>|z|
[95% Conf. Interval]
------+-----------------------------------------------------------------
    mage |  .0170007          .0244332      0.696   0.487      -.0308874
.0648888
   _cons |  -2.254025         .64633        -3.487   0.000      -3.520808   -
.9872413
------------------------------------------------------------------------------
```

2.2) คำนวณค่า $\text{OR}_{\text{MAGE30,20}} = e^{L}$

ในที่นี้ $L = b_{(\text{MAGE30 - MAGE20})}$

ดังนั้น OR. = Exp[(30-20)(0.017)]

= Exp(0.17)

= 1.19

จากสูตรทั่วไปในการคำนวณค่า 95% CI

95% CI OR. = exp [$L \pm 1.96 \sqrt{\text{var}\ (L)}$]

เมื่อ L = $10(b_{MAGE})$

Var (L) = $\text{var}[10(b_{MAGE})]$

Var (L) = $10^2\ \text{var}(b_{MAGE})$

$\sqrt{\text{var}\ (L)}$ = $\sqrt{10^2\ \text{var}(b_{MAGE})}$

$\sqrt{\text{var}\ (L)}$ = $10 \times \sqrt{\text{var}(b_{MAGE})}$

= $10 \times SE.(b_{MAGE})$

= 10 x 0.024 ....>*(เพราะ $\sqrt{var}$ = SE. และ $SE.(b_{MAGE})$ ได้จากข้อ 2.1)*

= 0.24

ดังนั้น 95% CI OR. = Exp[(0.17) $\pm$ 1.96 x 0.24]

= Exp(-0.3) ถึง Exp(0.64)

= 0.74 ถึง 1.90

แปลความหมาย : ทารกที่มีมารดาอายุ 30 ปี มีความเสี่ยงต่อการตายใน 28 วันหลังคลอด สูงเป็น 1.19 เท่าของทารกที่มีมารดาอายุ 20 ปี (95%CI: 0.74 ถึง 1.90)

ข้อ 3.

3.1) ใช้ STATA ในการ Fit Model ผลที่ได้ เป็นดังนี้

dead	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
area	-.3988155	.3230824	-1.234	0.217	-1.032045	.2344144
malpres	.8903152	.8428469	1.056	0.291	-.7616343	2.542265
a_mal	2.362425	.9739987	2.425	0.015	.453423	4.271427
_cons	-1.988927	.2091045	-9.512	0.000	-2.398765	-1.57909

3.2) คำนวณค่า OR.

$$OR_{(MALPRES=0)} = Exp(-0.399)$$
$$= 0.67$$
$$OR_{(MALPRES=1)} = Exp(-0.399 + 2.362)$$
$$= 7.12$$

3.5) แปลความหมายผลที่ได้

ความสัมพันธ์ระหว่างการเยี่ยมก่อนและหลังคลอดโดย อสม. กับการตายของทารกใน 28 วันหลังคลอด เมื่อเปรียบเทียบกับการให้บริการโดยเจ้าหน้าที่สาธารณสุข ณ สถานบริการสาธารณสุข พบว่าความสัมพันธ์ดังกล่าว ขึ้นอยู่กับว่า ทารกมีการคลอดท่าผิดปกติหรือไม่ กล่าวคือ ถ้าทารกคลอด***ท่าปกติ*** พบว่าทารกในพื้นที่มีการ**ให้บริการตามปกติ**โดยเจ้าหน้าที่สาธารณสุขที่สถานบริการสาธารณสุข ที่มีความเสี่ยงต่อการตายภายใน 28 วันหลังคลอดสูงเป็น 1.49 เท่าของทารกในพื้นที่ที่**มีการเยี่ยมโดย อสม.** (95%CI: 0.79 ถึง 2.78) ในทางตรงข้าม ถ้าทารกคลอด***ท่าผิดปกติ*** พบว่าทารกในพื้นที่ที่**มีการเยี่ยมโดย อสม**. มีความเสี่ยงต่อการตายภายใน 28 วันหลังคลอดสูงเป็น 7.13 เท่าของทารกในพื้นที่ที่มีการ**ให้บริการตามปกติ**โดยเจ้าหน้าที่สาธารณสุขที่สถานบริการสาธารณสุข (95%CI: 1.18 ถึง 43.14)

หมายเหตุ: *โปรดสังเกตว่าค่า OR กรณีท่าคลอดปกติน้อยกว่า 1 จึงสลับให้พื้นที่ที่มีการเยี่ยมเป็นกลุ่มอ้างอิง เพื่อง่ายต่อการแปลความหมายและทำความเข้าใจ ดังนั้นจึงต้องหารค่าเหล่านั้นกับ 1 OR=1.49 มาจาก 1 / 0.67 และ 95%CI: 0.79 ถึง 2.81 มาจาก 1/ 1.26 และ 1/0.36*

ภาคผนวก 2
คำแนะนำติดต่อขอรับข้อมูลเพิ่มเติม

1. ข้อมูลสำหรับใช้ในการทำแบบฝึกหัด

สำหรับท่านที่ต้องการข้อมูลที่ใช้ในการทำแบบฝึกหัดทั้งหมด โปรดส่งจดหมายสั่งซื้อ พร้อมระบุจ่าหน้าซองส่งถึงตัวท่านเอง และธนาณัติ 50 บาท (ห้าสิบบาทถ้วน) สั่งจ่าย ปทจ. มหาวิทยาลัยขอนแก่น ในนาม ***นางรัดดา ปิติ ภาควิชาชีวสถิติและประชากรศาสตร์ คณะสาธารณสุขศาสตร์ มหาวิทยาลัยขอนแก่น อ. เมือง จ.ขอนแก่น 40002*** จากนั้น ท่านจะได้รับแผ่นดิสเกตต์ที่บรรจุข้อมูลที่ใช้ในการทำแบบฝึกหัด หรือ Download ได้ฟรี ที่ http://web.kku.ac.th/~bandit/data/

2. โปรแกรม STATA®

STATA® เป็นโปรแกรมทางสถิติที่มีแนวโน้มใช้มากขึ้นในวงการศึกษา ติดต่อขอข้อมูลเพิ่มเติม หรือสั่งซื้อได้โดยตรง ตามที่อยู่ต่อไปนี้

Stata Coorporation
702 University Drive East
College Station, TX 77840 **USA.**
Fax. 409-696-4601
ดูรายละเอียดเพิ่มเติมได้โดยตรงที่ http://www.stata.com

ให้ข้อเสนอแนะผู้เขียน:
รองศาสตราจารย์ ดร. บัณฑิต ถิ่นคำรพ
ภาควิชาชีวสถิติและประชากรศาสตร์
คณะสาธารณสุขศาสตร์ มหาวิทยาลัยขอนแก่น
อ. เมือง จ. ขอนแก่น 40002
e-Mail : karawa@kku.ac.th