การคัดเลือกตัวแปรและการประมาณค่าสัมประสิทธิ์แบบเบส์เชิงประจักษ์สำหรับตัวแบบ Cox's proportional hazard ที่ข้อมูลมีมิติสูง

นางสาวอรณิชา ห่อนบุญเหิม

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2555

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย
บทคัดย่อและแฟ้มข้อมูลฉบับเต็มของวิทยานิพนธ์ตั้งแต่ปีการศึกษา 2554 ที่ให้บริการในคลังปัญญาจุฬาฯ (CUIR)
เป็นแฟ้มข้อมูลของนิสิตเจ้าของวิทยานิพนธ์ที่ส่งผ่านทางบัณฑิตวิทยาลัย
The abstract and full text of theses from the academic year 2011 in Chulalongkorn University Intellectual Repository(CUIR)
are the thesis authors' files submitted through the Graduate School.

EMPIRICAL BAYES VARIABLE SELECTION AND ESTIMATION FOR THE COX'S PROPORTIONAL HAZARD MODEL WITH HIGH DIMENSIONAL DATA

Miss Onnicha Honboonherm

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science Program in Statistics

Department of Statistics

Faculty of Commerce and Accountancy

Chulalongkorn University

Academic Year 2012

Copyright of Chulalongkorn University

หัวข้อวิทยานิพนธ์	การคัดเลือกตัวแปรและการประมาณค่าสัมประสิทธิ์แบบเบส์เชิงประจักษ์สำหรับตัวแบบ Cox's proportional hazard ที่ข้อมูลมีมิติสูง
โดย	นางสาวอรณิชา ห่อนบุญเหิม
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	อาจารย์ ดร. วิฐรา พึ่งพาพงศ์

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะพาณิชยศาสตร์และการบัญชี
(รองศาสตราจารย์ ดร.พสุ เดชะรินทร์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(อาจารย์ ดร. วิฐรา พึ่งพาพงศ์)

.....กรรมการ
(รองศาสตราจารย์ ดร. เสกสรร เกียรติสุไพบูลย์)

.....กรรมการภายนอกมหาวิทยาลัย
(อาจารย์ ดร. อรุณี กำลัง)

อรณิชา ห่อนบุญเหิม: การคัดเลือกตัวแปรและการประมาณค่าสัมประสิทธิ์แบบเบส์เชิงประจักษ์สำหรับตัวแบบ Cox's proportional hazard ที่ข้อมูลมีมิติสูง (EMPIRICAL BAYES VARIABLE SELECTION AND ESTIMATION FOR THE COX'S PROPORTIONAL HAZARD MODEL WITH HIGH DIMENSIONAL DATA). อ.ที่ปรึกษาวิทยานิพนธ์หลัก: อ.ดร. วิฐรา พึ่งพาพงศ์, 78 หน้า.

ในการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์สำหรับตัวแบบ Cox's proportional hazard ที่ข้อมูลมีมิติสูง สามารถวิเคราะห์ได้หลายวิธี ในการศึกษาครั้งนี้เราจะใช้วิธีการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์แบบเบส์เชิงประจักษ์ร่วมกับวิธีการทำซ้ำแบบมีเงื่อนไขฐานนิยมและมัธยฐาน ซึ่งเป็นเทคนิคที่สามารถคำนวณได้ง่ายและรวดเร็ว งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาผลกระทบของอัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ, ร้อยละของข้อมูลเซ็นเซอร์และค่าสัมประสิทธิ์ถดถอยเริ่มต้นจากค่าอัตราความผิดพลาดในการตรวจจับเชิงบวกและเชิงลบ

ข้อมูลที่ใช้ในการศึกษาคือเวลาในการอยู่รอดที่มีการแจกแจงแบบไวบูลล์ ที่ค่าสัมประสิทธิ์ส่วนใหญ่เป็นศูนย์กำหนดให้ขนาดตัวอย่างเท่ากับ100 ต่อตัวแปรอิสระ 300, 500และ1,000 ที่ร้อยละของข้อมูลเซ็นเซอร์10%, 50%และ70% พบว่า i) อัตราความผิดพลาดในการตรวจจับเชิงบวกและเชิงลบจะต่ำเมื่อระดับของข้อมูลเซ็นเซอร์ต่ำ ii) อัตราความผิดพลาดในการตรวจจับเชิงบวกและเชิงลบจะต่ำเมื่อจำนวนตัวแปรอิสระน้อย และiii) อัตราความผิดพลาดในการตรวจจับเชิงบวกและเชิงลบจะต่ำเมื่อค่าสัมประสิทธิ์ถดถอยเริ่มต้นเป็นค่าจริง vi) วิธีแบบเบส์เชิงประจักษ์ดีกว่าวิธีแบบ penalized lasso

ภาควิชา สถิติ ลายมือนิสิต .....

สาขาวิชา สถิติ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....

ปีการศึกษา 2555

5481721926: MAJOR STATISTICS
KEYWORDS: EMPIRICAL BAYES/COX'S PROPORTIONAL HAZARD MODEL/HIGH DIMENSIONAL DATA/ SPARSE VARIABLE

ONNICHA HONBOONHERM: EMPIRICAL BAYES VARIABLE SELECTION AND ESTIMATION FOR THE COX'S PROPORTIONAL HAZARD MODEL WITH HIGH DIMENSIONAL DATA. ADVISOR: VITARA PUNGPAPONG, Ph.D., 78 pp.

Cox's proportional hazard model with high-dimensional data cans analyses in several ways. In this study we will use empirical Bayes variable selection methods combined with iterated conditional modes/medians (ICM/M) algorithm which is empirically faster and easy to implement. The objective of this dissertation is to study the effects from the ratio of sample size to the number of independent variables, the percentages of censored data and the value of initial coefficient by comparing false positive and false negative rate.

The data in this study is survival times with Weibull distribution. Simulate sparse data with 100 sample size and 300, 500and1000 independent variables. The levels of percentages of censored data are 10%, 50% and 70%. Based on the false positive and false negative rate, the finding are following: i) the false positive and false negative rate will decrease as low percentage of censored data, ii) false positive and false negative rate will decrease as the number of variables is small and iii) false positive and false negative will decrease as the initial value of coefficients is true coefficients iv) empirical Bayes method is better than the Lasso method.

Department: Statistics Student's Signature .....
Field of Study: Statistics Advisor's Signature .....
Academic Year: 2012

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จได้ด้วยความช่วยเหลือ และดูแลเอาใจใส่อย่างดียิ่งของอาจารย์ ดร. วิฐรา พึ่งพาพงศ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้วิจัยขอกราบขอบพระคุณท่านอาจารย์เป็นอย่างสูงที่ให้คำแนะนำ และคอยให้คำปรึกษามากมายเกี่ยวกับวิทยานิพนธ์ด้วยดีเสมอมา

ผู้วิจัยขอกราบขอบพระคุณ รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา ประธานกรรมการสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร. เสกสรร เกียรติสุไพบูลย์และอาจารย์ ดร.อรุณี กำลังกรรมการสอบวิทยานิพนธ์ ที่กรุณาให้คำแนะนำ ตรวจสอบและแก้ไขวิทยานิพนธ์ฉบับนี้ให้สมบูรณ์ยิ่งขึ้น

สุดท้ายนี้ ผู้วิจัยขอกราบขอบพระคุณครอบครัว ที่ช่วยส่งเสริม สนับสนุนและให้กำลังใจเสมอมาจนสำเร็จการศึกษา และขอขอบคุณเพื่อนๆ ที่คอยให้ความช่วยเหลือ ปรึกษาและคอยให้กำลังใจผู้วิจัยมาโดยตลอด

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง

บทคัดย่อภาษาอังกฤษ ..... จ

กิตติกรรมประกาศ ..... ฉ

สารบัญ ..... ช

สารบัญตาราง ..... ฌ

สารบัญภาพ ..... ฎ

บทที่ 1 บทนำ 1

1.1 ความเป็นมาและความสำคัญของปัญหา 1

1.2 วัตถุประสงค์ของการวิจัย 5

1.3 ข้อตกลงเบื้องต้น 5

1.4 คำจำกัดความที่ใช้ในงานวิจัย 6

1.5 ขอบเขตงานวิจัย 7

1.6 เกณฑ์ที่ใช้ในการตัดสิน 9

1.7 ขั้นตอนดำเนินการวิจัย 10

1.8 ประโยชน์ที่คาดว่าจะได้รับ 11

บทที่ 2 ทฤษฎีและตัวสถิติที่เกี่ยวข้อง 12

2.1 ตัวแบบ Cox's proportional hazard 12

2.1.1 แนวคิดพื้นฐาน 13

2.1.2 การประมาณค่า 18

2.1.3 การคัดเลือกตัวแปร 24

2.2 การคัดเลือกแบบเบส์เชิงประจักษ์ 27

2.3 การคัดเลือกแบบเบส์เชิงประจักษ์สำหรับตัวแบบถดถอยเชิงเส้น 29

2.3.1 วิธีการทำซ้ำแบบมีเงื่อนไขฐานนิยมและมัธยฐาน (ICM/M) 31

หน้า

บทที่ 3 วิธีดำเนินการศึกษา 34
3.1 แผนการทำงาน 34
3.2 ขั้นตอนการทำงาน 35

บทที่ 4 ผลการวิเคราะห์ข้อมูล 47
4.1 อัตราความผิดพลาดในการตรวจจับเชิงบวก 49
4.2 อัตราความผิดพลาดในการตรวจจับเชิงลบ 57
4.3 เส้นโค้ง ROC และ พื้นที่ใต้เส้นโค้ง 65

บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะ 72
5.1 สรุปผล 72
5.2 ข้อเสนอแนะ 74

รายการอ้างอิง 76
บรรณานุกรม 77
ประวัติผู้เขียนวิทยานิพนธ์ 78

สารบัญตาราง

ตารางที่ หน้า

4.1.1 แสดงอัตราความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยของข้อมูลจำลองขนาด100 ค่าระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso และวิธี Lasso เมื่อให้อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคงที่ โดยที่ค่าเบี่ยงเบนมาตรฐานแสดงไว้ในวงเล็บ 50

4.1.2 แสดงอัตราความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยของข้อมูลจำลองขนาด100 ค่าระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso และวิธี Lasso เมื่อให้ร้อยละของข้อมูลเซ็นเซอร์คงที่ โดยที่ค่าเบี่ยงเบนมาตรฐานแสดงไว้ในวงเล็บ 53

4.2.1 แสดงอัตราความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยของข้อมูลจำลองขนาด100 ค่าระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso และวิธี Lasso เมื่อให้อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคงที่ โดยที่ค่าเบี่ยงเบนมาตรฐานแสดงไว้ในวงเล็บ 58

4.2.2 แสดงอัตราความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยของข้อมูลจำลองขนาด100 ค่าระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso และวิธี Lasso เมื่อให้ร้อยละของข้อมูลเซ็นเซอร์คงที่ โดยที่ค่าบี่ยงเบนมาตรฐานแสดงไว้ในวงเล็บ 61

4.3.1 แสดงค่า sensitivity และ 1-specificity ที่ใช้สำหรับสร้างเส้นโค้งROC เมื่อให้อัตราส่วนระหว่างขนาดตัวอย่างต่อ

ตารางที่ หน้า

ตัวแปรอิสระคงที่ เมื่อขนาดตัวอย่างต่อตัวแปรอิสระ
คือ 100:300, 100:500 และ100:1,000 66

4.3.2 แสดงค่า sensitivity และ 1-specificity ที่ใช้สำหรับสร้าง
เส้นโค้งROC เมื่อให้ร้อยละของข้อมูลเซ็นเซอร์คงที่ ที่
ระดับ10%, 50%และ70% 69

สารบัญภาพ

ภาพที่		หน้า
2.1.	แสดงลักษณะของfailure rateของเวลาการอยู่รอด	14
2.2	แสดงลักษณะของ failure rate ของเวลาการอยู่รอด ทีมีการแจกแจงแบบไวบูลล์ โดยกำหนดshape parameter มีค่าต่าง ๆ	15
2.3	แสดงตัวอย่างข้อมูลเซ็นเซอร์แบบต่าง ๆ	17
3.1	แผนผังการเขียนโปรแกรมจำลองเวลาการอยู่รอด	38
3.2	แผนผังการเขียนโปรแกรมการคัดเลือกแบบเบส์เชิงประจักษ์	44
4.1.1	แสดงการเปรียบเทียบอัตราความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso และวิธี Lassoเมื่อให้อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคงที่	51
4.1.2	แสดงการเปรียบเทียบอัตราความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso และวิธี Lasso เมื่อให้ร้อยละของข้อมูลเซ็นเซอร์คงที่	54
4.2.1	แสดงการเปรียบเทียบอัตราความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยระหว่างการคัดเลือกตัวแปร ด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริง กับค่าประมาณที่ได้จากวิธี Lasso และวิธี Lasso เมื่อให้อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคงที่	59
4.2.2	แสดงการเปรียบเทียบอัตราความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริง กับค่าประมาณที่ได้จากวิธี Lasso และวิธี Lasso เมื่อให้	

ภาพที่ หน้า

ร้อยละของข้อมูลเซ็นเซอร์คงที่ 62
4.3.1 แสดงเส้นโค้ง ROC และพื้นที่ใต้เส้นโค้ง เมื่อให้
อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระ
คงที่ ที่ระดับ 100:300, 100:500 และ 100:1000 67
4.3.2 แสดงเส้นโค้ง ROC และพื้นที่ใต้เส้นโค้ง เมื่อให้
ร้อยละของข้อมูลเซ็นเซอร์คงที่ ที่ระดับ 10%, 50%
และ 70% 70

บทที่ 1

บทนำ

1.1. ความเป็นมาและความสำคัญของปัญหา

เนื่องจากในปัจจุบันการวิเคราะห์การถดถอยเชิงเส้น (Regression Analysis) เข้าไปมีบทบาทในวงการต่างๆมากมายทั้งในวงการเศรษฐศาสตร์ การเงินการแพทย์หรือแม้แต่วงการด้านวิศวกรรมและอุตสาหกรรม การวิเคราะห์การอยู่รอด (Survival Analysis) ถือเป็นการวิเคราะห์การถดถอยเชิงเส้นประเภทหนึ่ง ตัวแบบ Cox's proportional hazard เป็นตัวแบบเชิงเส้นที่ใช้ในการวิเคราะห์การอยู่รอด เนื่องจากตัวแบบดังกล่าวมีลักษณะแบบกึ่งพารามิเตอร์ (semi-parameter)ที่มีความโดดเด่นคือรวมเอาความยืดหยุ่นได้ของวิธีที่ไม่ใช่พารามิเตอร์และประสิทธิภาพในการประมวลผลของวิธีที่ใช้พารามิเตอร์ อีกทั้งยังสามารถคำนวณหาค่าอัตราความเสี่ยง (hazard ratio) เมื่อค่าความเสี่ยงดังกล่าวเป็นค่าคงที่ไม่ขึ้นกับเวลา (constant rate over time) โดยที่ไม่จำเป็นต้องระบุถึงฟังก์ชัน hazard baseline ดังนั้นตัวแบบ Cox's proportional hazard จึงเป็นตัวแบบที่ใช้งานง่าย สะดวกและได้รับความนิยมสูง

โดยทั่วไปการประมาณค่าสัมประสิทธิ์การถดถอยในตัวแบบ Cox's proportional hazard สามารถทำได้โดยวิธีการประมาณค่าภาวะน่าจะเป็นสูงสุด (Maximum likelihood estimation (MLE)) ซึ่งจำเป็นที่จะต้องมีขนาดตัวอย่างอย่างน้อยเท่ากับจำนวนตัวแปรอิสระจึงจะสามารถหาตัวประมาณ MLE ได้ ในการศึกษานี้ เราสนใจการประมาณค่าสัมประสิทธิ์กรณีที่ตัวแปรอิสระมีขนาดใหญ่กว่าขนาดตัวอย่าง รวมไปถึงการคัดเลือกตัวแปรอิสระที่เหมาะสมเข้ามาในตัวแบบ แต่เนื่องจากความก้าวหน้าทางเทคโนโลยีสารสนเทศ ข้อมูลมีการจัดเก็บได้รวดเร็ว ประกอบกับต้นทุนของอุปกรณ์จัดเก็บข้อมูลต่ำลง ทำให้สามารถจัดเก็บข้อมูลได้ในปริมาณมาก ด้วยสาเหตุนี้ ข้อมูลที่มีมิติสูงสามารถพบเห็นได้โดยทั่วไป และการวิเคราะห์ข้อมูลเหล่านี้ได้มีการศึกษากันอย่างแพร่หลายในรอบหลายปีที่ผ่านมาของกลุ่มนักสถิติ ดังนั้นในหัวข้อนี้เราจะกล่าวถึงเทคนิคการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธ์ความถดถอยสำหรับตัวแบบ Cox's proportional hazard ที่ข้อมูลมีมิติสูง โดยวิธีที่ใช้ในการวิเคราะห์ข้อมูลลักษณะนี้ที่ได้รับความนิยมในปัจจุบันแบ่งออกเป็นสองวิธีใหญ่ๆ คือวิธี Penalized likelihoodและวิธีแบบเบส์

วิธีคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์สำหรับการวิเคราะห์สมการถดถอยสำหรับข้อมูลที่มีมิติสูงด้วยวิธี Penalized likelihood ค่าสัมประสิทธิ์ความถดถอยสามารถหาได้จากการหาค่าประมาณของสัมประสิทธิ์ที่ทำให้ Penalized likelihood มีค่าสูงสุด หรือเขียนได้ในรูป

$$\widehat{\beta} = \arg\min_{\beta}\left(l(\beta) + P_{\lambda}(\beta)\right), \lambda \geq 0$$

เมื่อ $l(\beta) = -\log likelihood$, $P_{\lambda}(\beta)$ คือ Penalty function และ λ คือ tuning parameter โดยที่ $\lambda \geq 0$ จากสมการข้างต้น หากเราเลือก Penalty function ที่เหมาะสมจะสามารถทำให้สัมประสิทธิ์ส่วนใหญ่เท่ากับศูนย์ ซึ่งเสมือนกับการเลือกตัวแปรเข้ามาในตัวแบบ

Tibshirani (1996) ได้เสนอวิธี Lasso โดยใช้ $l_1 - norm$ สำหรับ Penalty function หรือเขียนได้ในรูป $P_{\lambda}(\beta) = \lambda\sum_{j=1}^{p}\left|\beta_j\right|$ สำหรับ $l_1 - norm$ Penalty function นี้ เปรียบเสมือนการหาค่าภาวะน่าจะเป็นสูงสุด โดยมีข้อจำกัด คือ $\sum_{j=1}^{p}\left|\beta_j\right| \leq s$ เมื่อ $s > 0$ จากข้อจำกัดนี้ทำให้ค่าสัมประสิทธิ์บางค่ามีค่าเป็นศูนย์ ในงานวิจัยของ Tibshirani (1996) ได้ศึกษาเปรียบเทียบวิธีLasso กับวิธีแบบเป็นขั้นตอน (stepwise) ผลการศึกษาพบว่าวิธี Lasso ให้ผลที่แม่นยำและถูกต้องกว่าวิธีเป็นขั้นตอน

ต่อมา Fan and Li พบว่าข้อเสียของวิธี Lasso ของ Tibshirani คือค่าประมาณของสัมประสิทธิ์ถดถอยที่ได้จากวิธีดังกล่าวมีความเอนเอียง (bias) Fan and Li (2001) จึงได้เสนอวิธี SCAD ในการสร้าง Penalty function ขึ้นมาใหม่เพราะทั้งสองท่านเชื่อว่าการปรับรูปแบบของ Penalty function สามารถช่วยลดความเอนเอียง (bias) ที่เกิดขึ้นให้ต่ำลงได้ จึงได้ปรับแก้ไขและเสนอแนะให้ฟังก์ชันดังกล่าวอยู่รูป $P_{\lambda}(\beta) = \sum_{j=1}^{p} P_{\lambda,j}(\beta_j; a)$

โดยจะแบ่งค่า Penalty function ออกเป็น 3 รูปแบบ โดยแต่ละรูปแบบจะขึ้นกับค่า a และค่า λ เมื่อ $a > 2$ และ $\lambda \geq 0$ ดังพิจารณาได้ต่อไปนี้

กรณีที่ 1: ถ้า $\left|\beta_j\right| < \lambda$ ค่า Penalty function จะเขียนได้ในรูป

$$P_{\lambda}(\beta_j) = \lambda\left|\beta_j\right|$$

กรณีที่ 2: ถ้า $\lambda < \left|\beta_j\right| \leq a\lambda$ ค่า Penalty function จะเขียนได้ในรูป

$$P_{\lambda}(\beta_j) = -\left(\beta_j^2 - 2a\lambda\left|\beta_j\right| + \lambda^2\right)/\left[2(a-1)\right]$$

และกรณีที่ 3: ถ้า $\left|\beta_j > a\lambda\right|$ ค่า Penalty function จะเขียนได้ในรูป

$$P_\lambda\left(\beta_j\right) = \left(a+1\right)\lambda^2/2$$

ทั้งนี้ Fan and Li (2001) แนะนำให้ใช้ค่า $a = 3.7$ กับ Penalty function ข้างต้น ซึ่งการใช้ค่า a ดังกล่าวจะทำให้ค่าประมาณสัมประสิทธิ์ที่คำนวณได้มีค่าบางส่วนเป็นศูนย์ อีกทั้งยังช่วยลดค่าความเอนเอียงให้ต่ำลงเมื่อเทียบกับค่าประมาณที่ได้จากวิธี Lasso

ต่อมาในปี2006 Zou ได้เสนอวิธี Adaptive lasso โดยพัฒนามาจากวิธี Lasso โดยยังคงใช้ $l_1 - norm$ สำหรับสร้าง Penalty function แต่ได้เพิ่มเงื่อนไขเข้ามา โดยการให้ค่าน้ำหนัก (weight)ที่แตกต่างกันของพารามิเตอร์แต่ละตัว ดังนั้น Penalty function ตัวใหม่นี้สามารถเขียนได้ในรูป $P_\lambda\left(\beta\right) = \lambda\sum_{j=1}^{p}\widehat{w}_j\left|\beta_j\right|$ และจาก Penalty functionนี้ Zou (2006) ได้กล่าวถึงคุณสมบัติ Oracleของตัวประมาณค่าจากวิธี Adaptive lasso ว่า เมื่อขนาดตัวอย่างเข้าสู่ค่าอนันต์ Adaptive lassoจะยังคงรักษาประสิทธิภาพในการคัดเลือกตัวแปรเข้าสู่ตัวแบบเสมือนกับว่าทราบตัวแบบที่แท้จริง (true model) และจากคุณสมบัตินี้ทำให้วิธี Adaptive lasso มีความแตกต่างในทางที่ดีขึ้นจากวิธีเดิมคือวิธี Lasso

แนวคิดที่กล่าวถึงในข้างต้นทั้งหมดเป็นการหาค่าสัมประสิทธิ์สำหรับการวิเคราะห์ความถดถอยกรณีที่ตัวแปรตามมีการแจกแจงแบบปกติ อย่างไรก็ตาม แนวคิดดังกล่าวสามารถขยายไปยังตัวแบบ Cox's proportional hazardได้ดังเช่นที่ปรากฏในงานวิจัยของ Tibshirani (1997), Zhang and Lu (2007) และ Fan and Li (2002)

นอกเหนือจากวิธี Penalized likelihood แล้ว วิธีคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์สำหรับการวิเคราะห์สมการถดถอยสำหรับข้อมูลที่มีมิติสูงด้วยวิธีแบบเบส์ถือเป็นอีกทางเลือกหนึ่งที่ได้รับความนิยมและสนใจในกลุ่มนักสถิติ, การแพทย์, วิศวกรและนักวิชาการสาขาต่างๆ ในวงกว้าง เนื่องจากคุณสมบัติของวิธีการแบบเบส์ที่ใช้ข้อมูลเดิม (prior) ร่วมกับการใช้ข้อมูลปัจจุบัน (likelihood) เพื่อใช้ในการพยากรณ์ค่าในอนาคต (posterior) ทำให้ข้อมูลที่ได้มีความแม่นยำและน่าเชื่อถือมากกว่าวิธีที่ใช้เพียงข้อมูลปัจจุบัน (likelihood)เพียงอย่างเดียว

ในงานวิจัยของ Johnstone and Silverman (2004) ได้เสนอวิธี Empirical Bayes thresholding เพื่อใช้ในการสร้าง Threshold แบบสุ่มสำหรับข้อมูลอิสระที่มีการแจกแจงแบบปกติ โดยการให้ prior สำหรับค่าเฉลี่ยของข้อมูลแต่ละตัวในรูปของการแจกแจงแบบผสมระหว่างส่วนที่

ค่าพารามิเตอร์เป็นศูนย์และส่วนที่ค่าพารามิเตอร์ไม่เท่ากับศูนย์ และจาก prior ดังกล่าว ทำให้สามารถสร้างการแจกแจง posterior ที่มีลักษณะการแจกแจงแบบผสมระหว่างส่วนที่ค่าพารามิเตอร์เป็นศูนย์และส่วนที่ค่าพารามิเตอร์ไม่เท่ากับศูนย์เช่นเดียวกันกับตัว prior ดังนั้นหากเราเลือกใช้ตัวประมาณอย่างเหมาะสม เช่น ค่ามัธยฐาน (posterior median) จะทำให้ค่าพารามิเตอร์บางส่วนมีค่าเป็นศูนย์

โดยทั่วไปแล้วเทคนิคที่ใช้เป็นเครื่องมือสำหรับวิธีการวิเคราะห์แบบ Bayesian กันอย่างแพร่หลายคือวิธีมาคอร์ฟเชน มัลติคาโร (MCMC) แต่เป็นที่ทราบกันดีว่าวิธีดังกล่าวใช้เวลานานในการรอให้ข้อมูลมีค่าความผันแปรเฉลี่ยมีลักษณะคงที่ตลอดช่วงเวลา (converge) โดยเฉพาะในกรณีที่จำนวนพารามิเตอร์มีขนาดใหญ่

Pungpapong และคณะ (2012) ได้เสนอวิธีการทำซ้ำแบบมีเงื่อนไขฐานนิยม/มัธยฐาน (ICM/M) ซึ่งวิธีดังกล่าวเป็นที่ใช้ในการหาค่าพารามิเตอร์ของตัวแบบที่พิจารณา โดยมีแนวคิดคือ สัมประสิทธิ์ถดถอยสามารถคำนวณได้จากค่ามัธยฐานของการแจกแจงของฟังก์ชันความน่าจะเป็นภายใต้เงื่อนไขที่ว่าค่าพารามิเตอร์และค่าสัมประสิทธิ์ตัวอื่นๆที่ไม่ใช่ตัวที่ต้องการหาค่าประมาณเป็นค่าคงที่ (ทราบค่า) ส่วนพารามิเตอร์ตัวอื่นๆที่เข้ามาเกี่ยวข้องในตัวแบบขณะที่เราพิจารณาหาค่าสัมประสิทธิ์ถดถอยสามารถคำนวณได้จากค่าฐานนิยมของการแจกแจงของฟังก์ชันความน่าจะเป็น ภายใต้เงื่อนไขที่ว่าค่าพารามิเตอร์และค่าสัมประสิทธิ์ตัวอื่นๆที่ไม่ใช่ตัวที่ต้องการหาค่าประมาณเป็นค่าคงที่ (ทราบค่า) จากแนวคิดนี้จะเห็นว่าวิธีICM/M คำนวณได้ง่ายและรวดเร็วกว่าวิธีมาคอร์ฟเชน มัลติคาโร (MCMC) โดยงานวิจัยนี้ได้ศึกษาวิธีการคัดเลือกตัวแปรแบบเบส์เชิงประจักษ์ มีเทคนิคที่ช่วยในการทำงานคือวิธีการทำซ้ำแบบมีเงื่อนไขฐานนิยมและมัธยฐานกับตัวแบบการถดถอยโลจิสติก (Binary logistic regression)

ในการศึกษาครั้งนี้ผู้วิจัยจึงมีความสนใจที่จะนำแนวคิดจะใช้วิธีการคัดเลือกตัวแปรแบบเบส์เชิงประจักษ์ที่มีเทคนิคที่ช่วยในการทำงานคือวิธีการทำซ้ำแบบมีเงื่อนไขฐานนิยมและมัธยฐาน (ICM/M) มาต่อยอดในการเลือกตัวแปรอิสระเข้าสู่ตัวแบบ รวมไปถึงขั้นตอนการประมาณค่าสัมประสิทธิ์ความถดถอย (β) สำหรับตัวแบบ Cox's proportional hazard ที่ข้อมูลมีมิติสูง โดยจะจำลองข้อมูลในลักษณะที่มีอัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระและร้อยละของข้อมูลเซ็นเซอร์ในระดับต่างๆ เพื่อทำการทดสอบว่าวิธีการดังกล่าวเหมาะสมหรือให้แนวโน้มของผลลัพธ์ที่ดี (อัตราความผิดพลาดในการตรวจจับเชิงบวกและลบต่ำ) กับข้อมูลในลักษณะใด

1.2. วัตถุประสงค์การวิจัย

1. เพื่อศึกษาการคัดเลือกตัวแปรอิสระและการประมาณค่าสัมประสิทธิ์สำหรับตัวแบบ Cox's proportional hazard ที่ข้อมูลมีมิติสูงด้วยวิธีแบบเบส์เชิงประจักษ์ที่ทำงานร่วมกับวิธีการทำซ้ำแบบมีเงื่อนไขฐานนิยมและมัธยฐาน
2. เพื่อศึกษาผลกระทบของอัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระและร้อยละของข้อมูลเซ็นเซอร์ที่ระดับต่าง ๆ ต่อตัวแบบ Cox's proportional hazard ที่ข้อมูลมีมิติสูง โดยพิจารณาจากอัตราความผิดพลาดในการตรวจจับเชิงบวกและอัตราความผิดพลาดในการตรวจจับเชิงลบ

1.3. ข้อตกลงเบื้องต้น

1. ให้ $h(T|X)$ คือ ตัวแบบ Cox's proportional hazard ซึ่งเขียนได้ในรูป

$$h(T|X) = h_0(T)\exp\{\beta' X\}$$

ให้ $T = (t_1, t_2, ..., t_n)$ คือ เวกเตอร์ของRight-censored timeของตัวอย่างขนาด n

$X = (x_1^T, x_2^T, ..., x_n^T)^T$ คือ เมทริกซ์ของตัวแปรอิสระขนาด $n \times p$

$\beta = (\beta_1, \beta_2, ..., \beta_p)$ คือ เวกเตอร์ของสัมประสิทธิ์ความถดถอยขนาด $p \times 1$

$h_0(T)$ คือ ฟังก์ชัน hazard baseline หรือฟังก์ชัน hazard เมื่อตัวแปรอิสระทุกตัวเป็นศูนย์ $(X = 0)$

โดยตัวแบบ Cox's proportional hazard ประกอบไปด้วย

ฟังก์ชันการอยู่รอด (survival function)

$$s(T) = P(T > t) = 1 - F(t) = \exp\left\{-H_0(t)e^{x^T\beta}\right\}$$

เมื่อ $H_0(t) = \int_0^t h_0(u)du$

และฟังก์ชันการเสี่ยงอันตราย

$$h(T) = \lim_{\Delta t \to 0} \frac{P\{\text{an individual of age } t \text{ fails in the time interval } (t, t+\Delta t)\}}{\Delta t}$$

$$= P(T = t) = \left\{ h_0(t) e^{x^T \beta} \right\}$$

ดังนั้น likelihood function ของตัวแบบ Cox's proportional hazard จึงเท่ากับ

$$\prod_{i=1}^{n} \left[P(T = t)^{\zeta_i = 1} \cdot P(T > t)^{\zeta_i = 0} \right] = \prod_{i=1}^{n} \left[\left\{ h_0(t_i) e^{x_i^T \beta} \right\}^{\zeta_i} \cdot \exp\left\{ -H_0(t_i) e^{x_i^T \beta} \right\} \right]$$

โดยที่ $\zeta_i = \begin{cases} 1 & ; t_i \le c_i \\ 0 & ; t_i > c_i \end{cases}$ เมื่อ c_i คือเวลาเซ็นเซอร์ของ t_i

2. ให้เวลาการอยู่รอด (survival time) T มีการแจกแจงแบบไวบูลล์ ที่เขียนได้ในรูป

$$T = \left(-\frac{\log(U)}{\lambda \exp(\beta' X)} \right)^{1/\nu}$$

เมื่อ Scale parameter $\nu > 0$, Shape parameter $\lambda > 0$ และ $U \sim U[0,1]$

ในการศึกษาครั้งนี้ให้ $\nu = 10$, $\lambda = 1$

3. ให้ prior ของวิธีแบบเบส์เชิงประจักษ์เขียนในรูปการแจกแจงแบบผสมที่อยู่ในรูป

$$\beta \sim (1 - \omega) \delta_0(\beta) + \omega \gamma(\beta | \alpha)$$

เราจะใช้ค่าใช้ $\gamma(.)$ ในรูปความหนาแน่น Laplace ที่ $\alpha = 0.5$ ตามคำแนะนำของ Johnstone และ Silverman (2004)

4. ค่าประมาณสัมประสิทธิ์เริ่มต้นจะพิจารณาที่ 2 กรณีคือ กรณีที่ค่าสัมประสิทธิ์เริ่มต้นเป็นค่าสัมประสิทธิ์ที่แท้จริง(True beta) และกรณีที่ค่าประมาณสัมประสิทธิ์เริ่มต้นเป็นค่าที่ประมาณจากวิธี Lasso

5. ชุดข้อมูลที่สร้างขึ้น คือกรณีที่แตกต่างกัน 9 กรณี ที่มีอัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระและร้อยละของข้อมูลสูญหายแตกต่างกันออกไป

1.4. **คำจำกัดความที่ใช้ในงานวิจัย**

ในงานวิจัยนี้มีคำจำกัดความที่ใช้ในงานวิจัยดังนี้

อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระระดับสูง

คือ อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระที่มีขนาดตัวอย่างเท่ากับ10 และตัวแปรอิสระที่มีขนาดเท่ากับ 300

อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระระดับกลาง

คือ อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระที่มีขนาดตัวอย่างเท่ากับ100 และตัวแปรอิสระที่มีขนาดเท่ากับ 500

อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระระดับต่ำ

คือ อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระที่มีขนาดตัวอย่างเท่ากับ100 และตัวแปรอิสระที่มีขนาดเท่ากับ1,000

ขนาดของตัวแปรอิสระน้อย

คือ ตัวแปรอิสระขนาด 300

ขนาดของตัวแปรอิสระกลาง

คือ ตัวแปรอิสระขนาด 500

ขนาดของตัวแปรอิสระมาก

คือ ตัวแปรอิสระขนาด 1,000

ร้อยละของข้อมูลเซ็นเซอร์

คือ (จำนวนข้อมูลตัวอย่างที่ไม่อยู่ในช่วงการทดลอง/จำนวนตัวอย่างทั้งหมด) $\times 100\%$

โดยในการศึกษาครั้งนี้จะแบ่งร้อยละของข้อมูลเซ็นเซอร์ออกเป็น 3 ระดับ อันประกอบไปด้วยระดับต่ำ คือ ข้อมูลเซ็นเซอร์ที่ 10%, ระดับกลาง คือ ข้อมูลเซ็นเซอร์ที่ 50%และ ระดับสูง คือ ข้อมูลเซ็นเซอร์ที่ 70%

1.5. ขอบเขตของงานวิจัย

1. ตัวแปร T มีการแจกแจงแบบไวบูลล์ที่อยู่ในรูป

$$T = \left(-\frac{\log(U)}{\exp(\beta' X)} \right)^{1/10}$$

เมื่อ $U \sim U[0,1]$

ในการศึกษาครั้งนี้จะพิจารณาที่จำนวนของ X และ β มีขนาดเท่ากับ 300, 500 และ1,000

2. ตัวแปรอิสระ $x_i \overset{iid}{\sim} N(0,1)$

3. พิจารณาขนาดตัวอย่างที่มีขนาดเท่ากับ 100 และตัวแปรอิสระที่มีจำนวนเท่ากับ 300, 500และ1,000 ดังนั้นจะได้อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระเป็น 3 กรณี คือ 100:300, 100:500และ100:1,000
4. พิจารณาข้อมูลที่ค่าพารามิเตอร์ β ส่วนใหญ่มีค่าเป็นศูนย์ (sparse data) โดยที่ในทุกกรณีจะกำหนดให้ค่า β มีค่าดังนี้

   β ตัวที่ 1 ถึง 10 มีค่าเท่ากับ 5

   β ตัวที่ 101 ถึง 110 มีค่าเท่ากับ 2 นอกนั้นให้มีค่าเท่ากับศูนย์
5. จำลองกรณีศึกษาทั้งหมด 9 กรณี ที่มีอัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระและร้อยละของข้อมูลเซ็นเซอร์ที่แตกต่างกัน ดังนี้

ขนาดตัวอย่าง	จำนวนตัวแปรอิสระ	ข้อมูลเซ็นเซอร์
100	300	10%
		50%
		70%
	500	10%
		50%
		70%
	1,000	10%
		50%
		70%

6. การวิจัยครั้งนี้จะทำการจำลองข้อมูลให้มีลักษณะแตกต่างกันตามข้อกำหนดข้างต้น เพื่อนำแต่ละชุดข้อมูลเข้าสู่ขั้นตอนการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ด้วยวิธีแบบเบส์เชิงประจักษ์ที่ทำงานร่วมกับวิธีการทำซ้ำแบบมีเงื่อนไขฐานนิยมและมัธยฐาน โดยวิธีดังกล่าวจะกำหนดการจำลองซ้ำสูงสุดของข้อมูลแต่ละกรณีไว้ที่จำนวน 1,000 รอบ

1.6. เกณฑ์ที่ใช้ในการตัดสิน

เกณฑ์ที่ใช้ในการตัดสินว่าข้อมูลชุดใดเหมาะสมกับวิธีการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์แบบเบส์เชิงประจักษ์ที่ทำงานร่วมกับวิธีการทำซ้ำแบบมีเงื่อนไขฐานนิยม/มัธยฐานสำหรับตัวแบบ Cox's propostional hazard ที่ข้อมูลมีมิติสูง คือ การตรวจสอบอัตราความผิดพลาดในการตรวจจับเชิงบวก (False positive rate), อัตราความผิดพลาดในการตรวจจับเชิงลบ (False negative rate) และการตรวจสอบพื้นที่ใต้เส้นโค้ง Receiver Operator Characteristic (ROC) ของข้อมูลที่จำลองขึ้นมาทั้งหมด 9 กรณี โดยกรณีที่ให้อัตราความผิดพลาดในการตรวจจับเชิงบวกและเชิงลบต่ำจะถือว่ากรณีนั้นมีความเหมาะสมในการใช้การคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ด้วยวิธีการข้างต้น โดยที่

1. อัตราความผิดพลาดในการตรวจจับเชิงบวก (False positive rate) คือ ความน่าจะเป็นที่จะเกิดความผิดพลาดจากข้อสรุปที่ค่าประมาณสัมประสิทธิ์ถดถอยเชิงเส้นมีค่าไม่เท่ากับศูนย์ เมื่อค่าสัมประสิทธิ์ถดถอยที่แท้จริงมีค่าเท่ากับศูนย์ ถือเป็นความน่าจะเป็นของการเกิดความคลาดเคลื่อนประเภทที่1 ซึ่งสามารถคำนวณได้จากสูตร

$$\frac{\sum_{j=1}^{p} 1_{\{\hat{\beta}_j \neq 0 \, and \, \beta_j = 0\}}}{\sum_{j=1}^{p} 1_{\{\hat{\beta}_j \neq 0\}}}$$

เมื่อ p คือจำนวนตัวแปรอิสระ

2. อัตราความผิดพลาดในการตรวจจับเชิงลบ (False negative rate) คือ ความน่าจะเป็นที่จะเกิดความผิดพลาดจากข้อสรุปที่ค่าประมาณสัมประสิทธิ์ถดถอยเชิงเส้นมีค่าเท่ากับศูนย์ เมื่อค่าสัมประสิทธิ์ถดถอยที่แท้จริงมีค่าไม่เท่ากับศูนย์ ถือเป็นความน่าจะเป็นของการเกิดความคลาดเคลื่อนประเภทที่ 2 ซึ่งสามารถคำนวณได้จากสูตร

$$\frac{\sum_{j=1}^{p} 1_{\{\hat{\beta}_j = 0 \, and \, \beta_j \neq 0\}}}{\sum_{j=1}^{p} 1_{\{\hat{\beta}_j = 0\}}}$$

เมื่อ p คือจำนวนตัวแปรอิสระ

3. การพิจารณาเส้นโค้ง ROC คือการสร้างกราฟความสัมพันธ์ระหว่าง true positive rate (Sensitivity) กับ false positive rate (1 – Specificity) เพื่อเลือกจุดตัด (cut - off point) ที่เหมาะสม นอกจากนี้การสร้าง ROC curve ยังช่วยในการเปรียบเทียบประสิทธิภาพของการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ได้ด้วย โดยเปรียบ เทียบพื้นที่ใต้เส้นโค้งของแต่ละวิธี พื้นที่ใต้โค้งที่มากกว่าแสดงถึงประสิทธิภาพที่สูงกว่า โดยที่

$$\text{True positive rate (Sensitivity)} = \frac{\sum_{j=1}^{p} \mathbf{1}_{\{\hat{\beta}_j = 0 \, and \, \beta_j = 0\}}}{\sum_{j=1}^{p} \mathbf{1}_{\{\beta_j = 0\}}}$$

และ

$$\text{False positive rate (1 – Specificity)} = \mathbf{1} - \left(\frac{\sum_{j=1}^{p} \mathbf{1}_{\{\hat{\beta}_j \neq 0 \, and \, \beta_j \neq 0\}}}{\sum_{j=1}^{p} \mathbf{1}_{\{\beta_j \neq 0\}}} \right)$$

1.7. ขั้นตอนการดำเนินการวิจัย

1. กำหนดให้ขนาดตัวอย่างแต่ละชุดเป็น100 จำนวนตัวแปรอิสระเป็น 300, 500และ 1000
2. สร้างตัวแปรอิสระ x ที่มีการแจกแจงปกติ ที่ค่าเฉลี่ยเท่ากับ 0 และความคลาดเคลื่อนเท่ากับ 1 เท่ากับจำนวนขนาดตัวอย่างคูณกับขนาดตัวแปรอิสระในแต่กรณี
3. กำหนดค่าพารามิเตอร์ β ให้มีค่าส่วนใหญ่เป็นศูนย์โดยให้ค่า β ตัวที่1 ถึง 10 มีค่าเท่ากับ 5, β ตัวที่101 ถึง 110 มีค่าเท่ากับ 2 นอกนั้นให้มีค่าเท่ากับศูนย์
4. กำหนดร้อยละของข้อมูลเซ็นเซอร์เป็น 10%, 50%และ70%
5. นำข้อมูลที่ได้ในแต่ละชุดมาสร้างเวลาในการอยู่รอดที่มีการแจกแจงแบบไวบูลล์ ที่มีสูตรและเงื่อนไขตามที่กำหนด
6. ดังนั้นจากขั้นตอนและเงื่อนไขจากข้อ (1) - (5) ชุดข้อมูลที่ถูกจำลองขึ้นจะมีทั้งหมด 9 กรณี
7. กำหนดค่าเริ่มต้นของสัมประสิทธิ์ถดถอยเป็น 2 กรณี คือให้ค่าสัมประสิทธิ์เริ่มต้นเป็นค่าสัมประสิทธิ์ที่แท้จริงและให้ค่าสัมประสิทธิ์เริ่มต้นหาจากวิธี Lasso
8. นำชุดข้อมูลเวลาการอยู่รอดที่จำลองขึ้นมาทั้ง 9 กรณี เข้าสู่ขั้นตอนการคัดเลือกตัวแปร

และประมาณค่าสัมประสิทธิ์ด้วยวิธีแบบเบส์เชิงประจักษ์

9. นำค่าประมาณสัมประสิทธิ์ถดถอยที่ได้ในแต่ละกรณี (รวมถึงกรณีที่เป็นค่าเริ่มต้นของสัมประสิทธิ์ถดถอยทั้ง 2 กรณี) มาคำนวณค่าอัตราความผิดพลาดในการตรวจจับเชิงบวกและอัตราความผิดพลาดในการตรวจจับเชิงลบ
10. สรุปผลที่ได้จากการทดลอง

1.8. **ผลที่คาดว่าจะได้รับจากงานวิจัย**

1. เพื่อเป็นทางเลือกในการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์สำหรับตัวแบบ Cox's proportional hazardที่ข้อมูลมีมิติสูง
2. สามารถบอกผลกระทบในการคัดเลือกตัวแปรและการประมาณค่าสัมประสิทธิ์แบบเบส์เชิงประจักษ์โดยพิจารณาอัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ และร้อยละของข้อมูลเซ็นเซอร์ที่ระดับต่างๆต่อตัวแบบ Cox's proportional hazard ที่ข้อมูลมีมิติสูง
3. เพื่อเป็นแนวทางในการศึกษาเพิ่มเติมและเปรียบเทียบอัตราความผิดพลาดในการตรวจจับเชิงบวกและอัตราความผิดพลาดในการตรวจจับเชิงลบในการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ในข้อมูลที่มีค่าส่วนใหญ่เป็นศูนย์ในตัวแบบ Cox's proportional hazard ในสถานการณ์อื่นๆ

บทที่ 2

ทฤษฎีและตัวสถิติที่เกี่ยวข้อง

ตัวแบบ Cox's proportional hazard เป็นตัวแบบหนึ่งที่อยู่ในเรื่องของการวิเคราะห์การอยู่รอด การวิเคราะห์การอยู่รอดคือการศึกษาเวลาที่จะทำให้เหตุการณ์ใดเหตุการณ์หนึ่งเกิดขึ้น ซึ่งเดิมทีจะสนใจเพียงการแจกแจงของเวลาในการอยู่รอด(The distribution of survival time)เท่านั้น แม้การวิเคราะห์การอยู่รอดจะเป็นที่รู้จักดีในเรื่องการประมาณลักษณะการแจกแจงการอยู่รอดแบบไม่มีเงื่อนไข แต่สิ่งที่น่าสนใจที่สุดของเรื่องการวิเคราะห์การอยู่รอดก็คือการศึกษาความสัมพันธ์ระหว่างการอยู่รอดกับตัวแปรอิสระที่ส่งผลต่อการอยู่รอด ดังนั้นกระบวนการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบด้วยวิธีที่เหมาะสมจึงถือเป็นหัวใจหลักของการศึกษาเกี่ยวกับตัวแบบ Cox's proportional hazard ในหัวข้อนี้จะนำเสนอเกี่ยวกับรายละเอียดและลักษณะทั่วไปของตัวแบบ Cox's proportional hazard วิธีการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ในกรณีทั่วไป รวมไปถึงในกรณีที่ข้อมูลมีลักษณะที่ขนาดตัวอย่างน้อยกว่าจำนวนตัวแปรอิสระหรือที่เรียกว่าข้อมูลที่มีมิติสูง(High dimensional data)โดยที่ค่าสัมประสิทธิ์ส่วนใหญ่ของตัวแปรอิสระเป็นศูนย์

2.1. ตัวแบบ Cox's proportional hazard

ตัวแบบ Cox's proportional hazards เป็นวิธีการวิเคราะห์ time-to-event curves ที่ได้รับความนิยมมากที่สุด เนื่องจากเป็นตัวแบบ แบบกึ่งพารามิเตอร์ที่สามารถใช้ในการหาค่าประมาณอัตราส่วนความเสี่ยง(Hazard Ratio) ที่มีค่าความเสี่ยงที่คงที่ไม่ขึ้นกับเวลา ค่าดังกล่าวสามารถนำไปใช้งานได้อย่างกว้างขวางเช่นการหาค่าความเสี่ยงของการเป็นมะเร็ง โดยค่าความเสี่ยงควรมีค่าคงที่ไม่ปรับไปเปลี่ยนมาตามเวลาเพื่อให้การรักษาทำได้สะดวก แต่ในขณะเดียวกันก็ควรเป็นค่าที่เหมาะสม ซึ่งการจะได้ค่าดังกล่าวต้องพิจารณาจากตัวแบบที่แสดงความสัมพันธ์ระหว่างการอยู่รอดกับตัวแปรอิสระที่ส่งผลต่อการอยู่รอด โดยตัวแบบดังกล่าวถูกศึกษาในรูปของตัวแบบ Cox's proportional hazard แนวคิดนี้ถูกเสนอโดย David Cox, 1972 มีรายละเอียดดังนี้

2.1.1. แนวคิดพื้นฐาน

ให้ T แทนเวลาในการอยู่รอด เมื่อพิจารณาให้ T เป็นตัวแปรอิสระที่มีฟังก์ชันการแจกแจงสะสมคือ $P(t) = \Pr(T \le t)$ และฟังก์ชันความหนาแน่นของความน่าจะเป็น $p(t) = \frac{\partial P(t)}{\partial t}$ ในการเก็บข้อมูลเวลาการอยู่รอด สามารถแบ่งลักษณะข้อมูลเป็น 2 กลุ่มคือ กรณีเหตุการณ์ที่สนใจไม่เกิดขึ้นภายในช่วงเวลาที่กำหนดหรือสูญหายขณะทดลอง เราสามารถเขียนความน่าจะเป็นของเหตุการณ์ดังกล่าวด้วยฟังก์ชันการอยู่รอด (Survival function) ซึ่งเขียนได้ในรูป

$$S(t) = \Pr(T > t) = 1 - P(t)$$

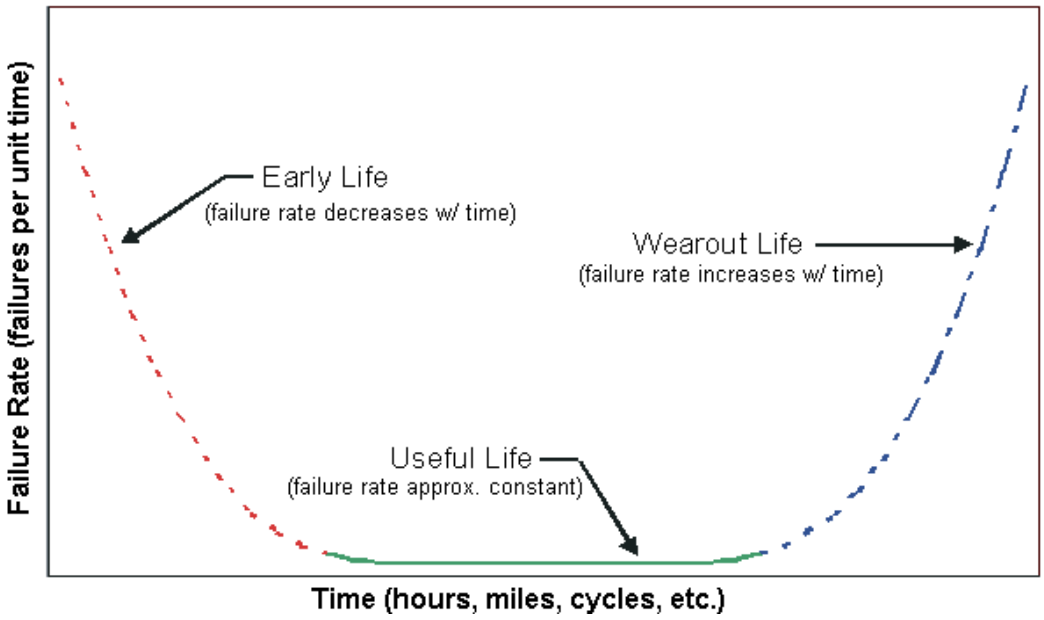
และกรณีที่ 2 เหตุการณ์ที่สนใจเกิดขึ้นภายในช่วงเวลาที่กำหนด ณ เวลา t เราสามารถเขียนความน่าจะเป็นของเหตุการณ์ที่บ่งบอกความเสี่ยงดังกล่าวได้จากฟังก์ชันความเสี่ยง (hazard function) ที่เขียนได้ในรูป

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr\left[(t \le T < t + \Delta t) \mid T \ge t \right]}{\Delta t}$$

ถ้า $f(t)$ คือฟังก์ชันความหนาแน่นของความน่าจะเป็น เราสามารถเขียนความสัมพันธ์ของฟังก์ชันการอยู่รอดและฟังก์ชันความเสี่ยงได้ในรูป

$$h(t) = \frac{f(t)}{S(t)}$$

เนื่องจากในเวลาการอยู่รอดเป็นการศึกษาลักษณะการแจกงแจงของความน่าจะเป็นของการอยู่รอด ซึ่งจะมีฟังก์ชันความเสี่ยง(hazard function) หรือlog-hazard แสดงความเสี่ยงของเหตุการณ์อยู่ในตัวแบบเสมอ โดยลักษณะของความเสี่ยงสามารถแบ่งเป็น3 ลักษณะดังภาพ



(ภาพจาก Reliability HotWire, eMagazine issue14, April2002)

ภาพที่ 2.1 แสดงลักษณะของfailure rateของเวลาการอยู่รอด

ช่วงแรกหรือช่วงเส้นจุด(dot line)สีแดงจะแสดงถึงความเสี่ยงของการเกิดเหตุการณ์จะลดลงตามเวลา ยกตัวอย่างเช่นในกลุ่มตัวอย่างที่เป็นผู้ป่วยที่ได้รับยาที่เหมาะสม เมื่อเหตุการณ์ที่สนใจคือเหตุการณ์ที่จะเกิดการทรุดลงของอาการในกลุ่มตัวอย่าง ดังนั้นโอกาสที่กลุ่มตัวอย่างจะเสี่ยงต่ออาการทรุดลงจะลดลงตามเวลาของการได้รับยา ช่วงที่ 2 หรือช่วงเส้นตรง(line)สีเขียวเป็นช่วงที่ความเสี่ยงของการเกิดเหตุการณ์คงที่ไม่ขึ้นกับเวลา เช่นกลุ่มตัวอย่างคือบุคคลทั่วไป ดังนั้นโอกาสที่กลุ่มตัวอย่างจะเสี่ยงต่ออาการทรุดลงจะคงที่ไม่ขึ้นหรือลงตามเวลา และช่วงสุดท้ายหรือช่วงเส้นประ(dash line)สีน้ำเงินแสดงถึงความเสี่ยงของการเกิดเหตุการณ์จะเพิ่มขึ้นตามเวลา ยกตัวอย่างเช่นในกลุ่มตัวอย่างที่เป็นหลอดไฟหรือถนน เมื่อเหตุการณ์ที่สนใจคือเหตุการณ์ที่จะเกิดการทรุดลงหรือหมดอายุการใช้งานในกลุ่มตัวอย่าง ความเสี่ยงของการเกิดเหตุการณ์จะเพิ่มขึ้นตามระยะเวลา หรืออายุการใช้งาน

การสร้างเวลาการอยู่รอดที่มีความเสี่ยงดังกล่าว สามารถทำได้โดยการกำหนดการแจกแจงเวลา เช่น เมื่อกำหนดให้ค่าความเสี่ยงคงที่ $h(t)=\lambda$ ก็คือการกำหนดการแจกแจงของเวลาการอยู่รอดให้มีการแจกแจงแบบเอ็กซ์โปเนนเชียลที่มีความหนาแน่น $f_0(t)=\lambda\exp\{-\lambda t\}$ โดยที่ $\lambda>0$ คือ scale parameter หรือ เมื่อต้องการกำหนดให้ความเสี่ยงไม่คงที่ สามารถกำหนดการแจกแจงของเวลาการอยู่รอดได้ในรูปการแจกแจง Gompertz และไวบูลล์

เมื่อให้เวลาการอยู่รอดมีการแจกแจงแบบ Gompertz จะมี hazard functionในรูป

$$h(t) = \exp(\lambda + \rho t)$$

หรือ log-hazard คือ $\log h(t) = \lambda + \rho t$

โดยที่ $\rho \in (-\infty, \infty)$ คือ shape parameter และ $\lambda > 0$ คือscale parameter

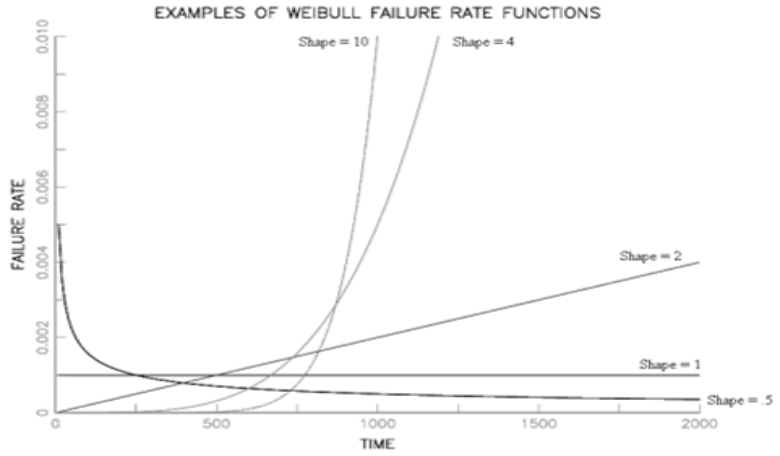
หรือเมื่อให้เวลาการอยู่รอดมีการแจกแจงแบบไวบูลล์(Weibull) จะมีhazard functionในรูป

$$h(t) = \lambda \nu t^{(\nu-1)}$$

หรือ log-hazard คือ $\log h(t) = \log(\lambda\nu) + (\nu - 1)\log(t)$

โดยที่ $\lambda > 0$ คือ scale parameter $\nu > 0$ คือ shape parameter

เวลาการอยู่รอดที่มีการแจกแจงแบบไวบูลล์สามารถสร้างความเสี่ยงของเหตุการณ์ให้มีลักษณะได้ทั้ง 3 รูปแบบคือเพิ่มขึ้นตามเวลา, คงที่ตลอดช่วงเวลา และลดลงตามเวลา โดยจะขึ้นอยู่กับการกำหนดค่าscale paremeter



(ภาพจากNIST/SEMATECH e-Handbook of Statistical Methods)

ภาพที่ 2.2 แสดงลักษณะของ failure rate ของเวลาการอยู่รอดที่มีการแจกแจงแบบไวบูลล์ โดยกำหนดshape parameter มีค่าต่าง ๆ

โดยที่ตัวแบบของ log-hazard สามารถเขียนในรูปเชิงเส้นได้ดังนี้

$$\log h_i(t) = \alpha + \beta_1(x_{i1}) + \beta_2(x_{i2}) + \ldots + \beta_k(x_{ik})$$

หรืออาจเขียนในรูป

$$h_i(t) = \exp\left(\alpha + \beta_1(x_{i1}) + \beta_2(x_{i2}) + \ldots + \beta_k(x_{ik})\right)$$

โดยที่ x คือตัวแปรอิสระของค่าสังเกต i

α คือค่า log- hazard baseline เมื่อค่าสังเกตทุกค่าเป็น0 ค่าความเสี่ยงจะเท่ากับค่า hazard baseline

โดยที่ในขบวนการเกี่ยวกับการอยู่รอดจะไม่มีค่าคลาดเคลื่อนในตัวแบบ
ให้ $\alpha(t) = \log h_0(t)$ ตัวแบบ Cox สามารถเขียนอยู่ในรูป

$$\log h_i(t) = \alpha(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

หรืออาจเขียนได้ในรูป

$$h_i(t) = h_0(t)\exp\left(\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}\right)$$

ซึ่งเป็นตัวแบบแบบกึ่งพารามิเตอร์เนื่องจากไม่ว่าค่า hazard baselineจะอยู่ในรูปแบบใดก็ตาม ตัวแปรอิสระก็จะยังคงอยู่ในรูปเชิงเส้น
พิจารณาค่าสังเกตที่ i และ j ซึ่งมีค่าของตัวแปรอิสระแตกต่างกัน จะได้ว่าตัวอัตราส่วนความเสี่ยง (Hazard Ratio) เขียนได้ในรูป

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

และ

$$\eta_j = \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \beta_k x_{jk}$$

ดังนั้นอัตราส่วนความเสี่ยงของค่าสังเกตทั้ง 2 คือ

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)e^{\eta_i}}{h_0(t)e^{\eta_j}}$$

$$= \frac{\eta_i}{\eta_j}$$

จะเห็นว่าค่าที่ได้ ไม่ขึ้นกับเวลา t เราเรียกว่าค่าอัตราส่วนความเสี่ยง(Hazard Ratio) ของตัวแบบ Cox's proportional hazard model โดยค่าที่ได้จะบอกถึงว่ามีโอกาสที่สิ่งๆหนึ่งจะหมดอายุขัยมากกว่าหรือน้อยกว่าเมื่อเทียบกับอีกสิ่งหนึ่งเป็นจำนวนเท่าใด

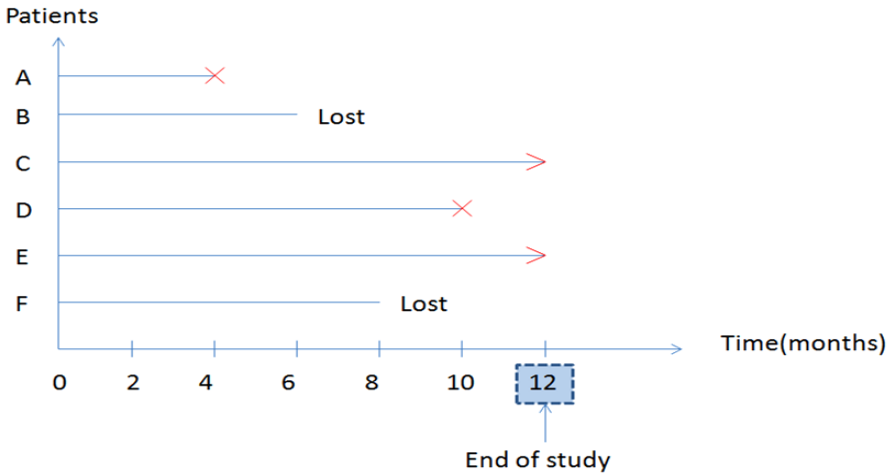
โดยในตัวแบบ Cox's proportional hazard model มีค่าภาวะน่าจะเป็นแบบเต็มรูปแบบ (the full likelihood) อยู่ในรูป

$$L = \prod_{i=1}^{n}\left[P(T=t)^{\zeta_i=1}\cdot P(T>t)^{\zeta_i=0}\right] = \prod_{i=1}^{n}\left[\left\{h_0(t_i)e^{x_i^T\beta}\right\}^{\zeta_i}\cdot\exp\left\{-H_0(t_i)e^{x_i^T\beta}\right\}\right]$$

เมื่อ $H_0(t) = \int_0^t h_0(u)\,du$ โดยที่ $y_1 < y_2 < \cdots < y_D$ เป็นค่า t_i ที่แตกต่างกัน

และ $\Delta\hat{h}_0\left(y_j\right)=\dfrac{d_j}{\sum_{i:t_j\geq y_j} e^{x_i^T\beta}}, d_j=\sum_{i:t_i=y_i}\varsigma_i$

ในการเก็บข้อมูลเวลาการอยู่รอด สิ่งหนึ่งที่มักถูกกล่าวถึงคู่กันคือข้อมูลเซ็นเซอร์ จากตัวอย่างแสดงการทดลองเพื่อเก็บเวลาการอยู่รอดโดยมีกลุ่มตัวอย่างคือผู้ป่วย 6 ท่าน คือ A, B, C, D, E และ F โดยเหตุการณ์ที่สนใจคือการทรุดลงของอาการของผู้ป่วย โดยระยะเวลาการทดลองคือ 12 เดือน



ภาพที่ 2.3 แสดงตัวอย่างข้อมูลเซ็นเซอร์แบบต่าง ๆ

จากภาพจะเห็นว่ามีผู้ป่วยเพียง2 ท่านคือ Aและ D เท่านั้นที่สามารถเก็บเวลาการอยู่รอดที่แท้จริงได้ ส่วนที่เหลือมีรายละเอียดดังนี้ ผู้ป่วย B และ Fแทนลักษณะของผู้ป่วยที่เป็นกลุ่มตัวอย่างแต่ไม่มาติดต่อจนสิ้นสุดการทดลอง เราถือว่าเวลาการอยู่รอดของผู้ป่วยทั้ง2รายคือเวลาสุดท้ายที่เข้ามาติดต่อ (last contact) เรียกเวลานี้ว่าเวลาเซ็นเซอร์ ส่วนกรณีของผู้ป่วยCและEเป็นผู้ป่วยที่ไม่มีอาการเปลี่ยนตามที่เราสนใจตั้งแต่เริ่มการทดลองจนสิ้นสุดการทดลอง เราถือว่าเวลาการอยู่รอดของผู้ป่วยทั้งสองคือ12 เดือนหรือเวลาสิ้นสุดการทดลอง

2.1.2. การประมาณค่าสัมประสิทธิ์

ค่าที่คำนวณได้จากตัวแบบ Cox's proportional hazard นี้สามารถนำไปประยุกต์ใช้งานได้หลากหลายเช่นหาอัตราการเกิดโรค อัตราการเสื่อมของถนน ซึ่งเมื่อสิ่งที่กล่าวมาเป็นค่าคงที่ที่ไม่ขึ้นกับเวลาการรักษาและดูแลย่อมทำได้ง่ายและสะดวก การจะหาค่าอัตราคงที่นี้ได้จำเป็นต้องทราบค่าสัมประสิทธิ์ถดถอยในตัวแบบ เนื่องจากตัวแบบ Cox's proportional hazardถือว่าเป็นตัวแบบเชิงเส้นประเภทหนึ่งดังนั้นวิธีการประมาณค่าสัมประสิทธิ์ที่ง่ายและนิยมโดยทั่วไปคือวิธีการหาภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimate (MLE)), วิธีนิวตัน-รัฟสัน (Newton-Raphson)และวิธีกำลังสองน้อยที่สุดถ่วงน้ำหนักแบบการย้อนซ้ำโดยแต่ละวิธีมีแนวคิดคือ

ให้ $T=\left(t_1,t_2,...,t_n\right)$ คือ เวกเตอร์ของRight-censored timeของตัวอย่างขนาด n ตัวอย่าง

$X=\left(x_1^T,x_2^T,...,x_n^T\right)^T$ เป็นเมทริกซ์ขนาด $n\times p$ ของตัวแปรอิสระจำนวน p ตัว

$$\rho_T\left(t;\theta;\phi\right)=\exp\left\{\frac{t\theta-b\left(\theta\right)}{a\left(\phi\right)}+c\left(t,\phi\right)\right\}$$

โดยที่ $\theta=\left(\theta_1,...,\theta_n\right)$ คือพารามิเตอร์ของตัวแบบ, ϕ คือพารามิเตอร์ของการกระจาย, T มีค่าเฉลี่ยเท่ากับ μ ซึ่งขึ้นกับค่า $x_1,x_2,...,x_p$ ให้ $X=\left(x_1^T,x_2^T,...,x_n^T\right)^T$ คือเมทริกซ์ของตัวแปรอิสระขนาด $n\times p$ ส่วน $a\left(.\right),b\left(.\right)$ และ $c\left(.\right)$ เป็นฟังก์ชันเฉพาะเจาะจงที่ได้จาก

ให้ $\eta=\alpha+x_1\beta_1+...+x_p\beta_p$ และ $g\left(\mu\right)=\eta$ เราเรียก $g\left(\cdot\right)$ ว่า canonical link function ดังนั้นจะได้ $\eta=\theta$ โดยที่

$$E\left(T\right)=\mu=b'\left(\theta\right)$$

และ

$$Var\left[T\right]=b''\left(\theta\right)a\left(\phi\right)$$

จากการหาค่าประมาณพารามิเตอร์โดยวิธีภาวะน่าจะเป็นสูงสุด จากค่าสังเกตที่อิสระกันจำนวน n ค่าที่มีความหนาแน่น $\rho_T\left(t;\theta;\phi\right)$ ค่า log-likelihood จะเขียนได้ในรูป

$$l\left(\beta_0,\beta\right)=l\left(\theta\right)=\sum_{i=1}^{n}\log\rho_{t_i}\left(t_i\mid\theta_i,\phi\right)=\sum_{i=1}^{n}\left(\frac{t_i\theta_i-b\left(\theta_i\right)}{a\left(\phi\right)}+c\left(t_i,\phi\right)\right)$$

เมื่อพิจารณา log-likelihood ของค่าสังเกตที่ i คือ

$$l_i = \log \rho_{t_i}\left(t_i \mid \theta_i, \phi\right) = \left(\frac{t_i\theta_i - b(\theta_i)}{a(\phi)} + c\left(t_i, \phi\right) \right)$$

เนื่องจากตัวแบบเชิงเส้นวางนัยทั่วไปประกอบไปด้วยส่วนประกอบ 3 ส่วน คือ ส่วนประกอบเชิงสุ่ม $\left(\rho_{t_i}\left(t_i \mid \theta_i, \phi\right)\right)$, ส่วนประกอบที่มีระบบ $\left(\eta_i = X\beta\right)$ และส่วนประกอบlink function $\left(\eta_i = g\left[\mu_i\right]\right)$ โดยที่ฟังก์ชัน g ซึ่งทำให้ $g\left(\mu_i\right) = \theta_i$ เรียกว่า “canonical link function” ดังนั้นการหาค่าMLE จาก log-likelihood หาได้จากการหาอนุพันธ์แบบกฎลูกโซ่คือ

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j}$$

โดยที่ $\frac{\partial l_i}{\partial \theta_i} = \frac{t_i - b'(\theta)}{a(\phi)}$

แต่จาก Cox&Hinkley (1974) พบว่า

(1) $E\left(\frac{\partial l}{\partial \theta}\right) = 0$ และ $-E\left(\frac{\partial^2 l}{\partial \theta^2}\right) = E\left(\frac{\partial l}{\partial \theta}\right)^2$ ดังนั้น

$$b'(\theta_i) = E(T_i) = \mu_i$$

$$\frac{\partial l_i}{\partial \theta_i} = \frac{t_i - \mu_i}{a(\phi)}$$

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i)$$

และ $$\frac{\partial^2 l}{\partial \theta^2} = \frac{-b''(\theta_i)}{a(\phi)}$$

(2) $\frac{b''(\theta_i)}{a(\phi)} = E\left[\frac{T - \mu}{a(\phi)}\right]^2 = Var\left(\frac{T_i}{\left[a(\phi)\right]^2}\right)$ ดังนั้น

$$b''(\theta_i) = \frac{Var(T_i)}{a(\phi)}$$

$$\therefore \quad \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)} = \frac{a(\phi)}{Var(T_i)}$$

เนื่องจาก $\eta_i = \sum_j \beta_j x_{ij}$ ดังนั้น $\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$

และจาก(1) และ (2) จะได้

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\left(t_i - \mu_i\right)}{a(\phi)} \cdot \frac{a(\phi)}{Var(T_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot x_{ij} = \frac{\left(t_i - \mu_i\right)x_{ij}}{Var(T_i)} \cdot \left(\frac{\partial \mu_i}{\partial \eta_i}\right)$$

ดังนั้นสมการภาวะน่าจะเป็นคือ

$$\frac{\partial l_i}{\partial \beta_j} = U_j = \sum_i^N \frac{(t_i - \mu_i) x_{ij}}{Var(T_i)} \cdot \left(\frac{\partial \mu_i}{\partial \eta_i} \right) = 0,\ j = 1, 2, ..., p$$

จากสมการ log-likelihood ที่ได้มักจะไม่อยู่ในรูปสมการเชิงเส้น วิธีที่เหมาะสมที่ใช้แก้ปัญหาดังกล่าวเช่นวิธีนิวตัน-รัฟสันและวิธีFisher scoring

วิธีนิวตัน-รัฟสัน (Newton-Raphson) เป็นวิธีในการหาค่าประมาณแบบหนึ่งที่ใช้ได้ในกรณีที่ T (ตัวแปรตาม) และ X (ตัวแปรอิสระ)มีความสัมพันธ์กันทั้งในลักษณะเชิงเส้นและไม่เป็นเชิงเส้นมีขบวนการทำงานและแนวคิดคือ

เมื่อพิจารณาข้อมูลกลุ่มตัวอย่างของผู้ป่วยจำนวน 9 รายโดยที่ $x_{(i)}$ คือปัจจัยที่คาดว่าทำให้เกิดโรคหรือตัวแปรอิสระ และ $t_{(i)}$ คือเวลาการอยู่รอดหรือเวลาที่เก็บได้ตั้งแต่เริ่มการทดลองจนผู้ป่วยตาย หน่วยเป็นวัน ดังนี้

i	1	2	3	4	5	6	7	8	9
$t_{(i)}$	6	98	189	374	1002	1205	2065	2201	2421
$x_{(i)}$	31.4	21.5	27.1	22.7	35.7	30.7	26.5	28.3	27.9

หมายเหตุ: ชุดข้อมูลนี้จะพิจารณาที่กรณีไม่มีข้อมูลเซ็นเซอร์

จากข้อมูลข้างต้นสามารถสร้างตัวแบบ Cox's hazard ได้ในรูป

$$h(t, x) = h_0(t, \alpha) e^{\beta x}$$

ขั้นตอนการทำงานของวิธีนิวตัน-รัฟสัน มีดังนี้

1. กำหนด $k = 0$
2. เลือกค่าเริ่มต้น $\beta^{(0)} = 0$
3. หาค่า $\beta^{(k+1)} = \dfrac{U\left(\beta^{(k)}\right) + \beta^{(k)}}{I\left(\beta^{(k)}\right)}$

เมื่อให้ $U(\beta) = \left(\dfrac{\partial l(\beta)}{\partial \beta_1}, ..., \dfrac{\partial l(\beta)}{\partial \beta_p} \right)$ โดยที่ $l(\beta)$ คือฟังก์ชัน log-likelihood

และ $I(\beta) = -\begin{pmatrix} \frac{\partial^2 l(\beta)}{\partial \beta_1^2} & \cdots & \frac{\partial^2 l(\beta)}{\partial \beta_1 \partial \beta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 l(\beta)}{\partial \beta_1 \partial \beta_p} & \cdots & \frac{\partial^2 l(\beta)}{\partial \beta_p^2} \end{pmatrix}$

แต่เนื่องจากในตัวแบบที่พิจารณามี β เพียงตัวเดียวดังนั้นสร้าง log-likelihood และอนุพันธ์อันดับหนึ่งและสองได้ในรูป

$$l(\beta) = \beta \sum_{i=1}^{9} x_{(i)} - \sum_{i=1}^{9} \log\left(\sum_{j \in R(t_{(i)})} e^{x\beta_j} \right)$$

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{9} x_{(i)} - \sum_{i=1}^{9} \left(\frac{\sum_{j \in R(t_{(i)})} \beta x_j}{\sum_{j \in R(t_{(i)})} e^{\beta x_j}} \right)$$

$$\frac{\partial^2 l(\beta)}{\partial \beta^2} = -\sum_{i=1}^{9} \left(\frac{\left(\sum_{j \in R(t_{(i)})} x^2 e^{\beta x_j} \right)\left(\sum_{j \in R(t_{(i)})} e^{\beta x_j} \right) - \left(\sum_{j \in R(t_{(i)})} x e^{\beta x_j} \right)\left(\sum_{j \in R(t_{(i)})} x^2 e^{\beta x_j} \right)}{\left(\sum_{j \in R(t_{(i)})} e^{\beta x_j} \right)^2} \right)$$

ดังนั้นจะได้ $U\left(\beta^{(0)}\right) = U(0) = -2.51$ และ $I\left(\beta^{(0)}\right) = I(0) = 77.13$

คำนวณค่า $\beta^{(k+1)} = \beta^{(1)}$

$$\beta^{(1)} = \frac{U\left(\beta^{(0)}\right) + \beta^{(0)}}{I\left(\beta^{(0)}\right)} = \frac{-2.51 + 0}{77.13} = -0.0326$$

4. คำนวณค่า $\theta_i^{(k+1)}; i = 1, 2, ..., p$ ทีละค่าจนครบทั้ง p ตัว ซึ่งในกรณีนี้ $p = 1$
5. ทำขั้นที่1-3ซ้ำจนกว่าค่าประมาณ β จะมีค่าความผันแปรเฉลี่ยมีลักษณะคงที่ตลอดช่วงเวลา (converge) ดังนี้

$U\left(\beta^{(1)}\right) = U(-0.0326) = -0.069$ และ $I\left(\beta^{(1)}\right) = I(-0.0326) = 72.83$

$$\beta^{(2)} = \frac{U\left(\beta^{(1)}\right) + \beta^{(1)}}{I\left(\beta^{(1)}\right)} = \frac{-0.069 - 0.0326}{72.83} = -0.0335$$

$$U\left(\beta^{(2)}\right) = U\left(-0.0335\right) = -0.000061 \text{ และ } I\left(\beta^{(2)}\right) = I\left(-0.0335\right) = 72.70$$

$$\beta^{(3)} = \frac{U\left(\beta^{(2)}\right) + \beta^{(2)}}{I\left(\beta^{(2)}\right)} = \frac{-0.000061 - 0.0335}{72.70} = -0.0335$$

เนื่องจาก $\beta^{(2)} = \beta^{(3)}$ จึงถือว่าข้อมูลลู่เข้าเรียบร้อยแล้ว ดังนั้นค่า $\beta = -0.0335$

อีกทางเลือกที่ใช้ในการหาค่าประมาณทั้งในกรณีที่ความสัมพันธ์ของตัวแปรตาม (T) และตัวแปรอิสระ (X) เป็นแบบเชิงเส้นและไม่เชิงเส้นคือแนวคิดของFisher scoring โดยการหาค่าคาดหวัง ของobserve information matrix เขียนได้ในรูป

$$E\begin{pmatrix} \frac{\partial^2 l(\beta)}{\partial \beta_1^2} & \cdots & \frac{\partial^2 l(\beta)}{\partial \beta_1 \partial \beta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 l(\beta)}{\partial \beta_1 \partial \beta_p} & \cdots & \frac{\partial^2 l(\beta)}{\partial \beta_p^2} \end{pmatrix}$$

หรือเท่ากับ $-I(\beta)$ ในวิธีนิวตัน-รัฟสัน จากนั้นสร้างตัวถ่วงน้ำหนัก(weight(w_i))สำหรับตัวแปรอิสระแต่ละตัวและสร้างตัวแปรค่าสังเกตเทียม(pseudo observation(z_i)) แล้วจึงนำแนวคิดของวิธีนิวตัน-รัฟสันที่อาศัยเทคนิคการย้อนซ้ำเพื่อหาค่าที่เหมาะสมมาประยุกต์ใช้ Green(1984)เรียกกระบวนการดังกล่าวว่ากระบวนการย้อนซ้ำ(IRLS process) หรือที่รู้จักในชื่อวิธีกำลังสองน้อยสุดถ่วงน้ำหนักแบบย้อนซ้ำ(Iterative Reweight Least Square(IRLS)) ขั้นตอนการย้อนซ้ำหาค่า $\beta^{(k+1)}$ จะสิ้นสุดลงเมื่อค่าประมาณที่ได้ลู่เข้า(convergence) ซึ่งมีขั้นตอนการคิดดังนี้ จากการหาค่าประมาณของวิธีนิวตัน-รัฟสันคือ

$$\beta^{(k+1)} = \frac{U\left(\beta^{(k)}\right) + \beta^{(k)}}{I\left(\beta^{(k)}\right)}$$

ให้ $H = -I(\beta)$ จากการใช้สูตร Fisher scoring H อาจเขียนได้ในรูปเมทริกซ์ M ในรูป $E(H) = M$ จะได้ว่า

$$\beta^{(k+1)} = \beta^{(k)} + \left(M^{(k)}\right)^{-1} U\left(\beta^{(k)}\right)$$

หรือ $$\left(M^{(k)}\right)\beta^{(k+1)} = \left(M^{(k)}\right)\beta^{(k)} + U\left(\beta^{(k)}\right) \quad \text{.....(ก)}$$

โดยที่ $M = X'WX$ เมื่อ W คือเมทริกซ์ของน้ำหนัก

ดังนั้น $$X'W^{(k)}X\beta^{(k+1)} = X'W^{(k)}X\beta^{(k)} + U\left(\beta^{(k)}\right)$$

เมื่อ $W = diag\{w_i\}$ และตัวถ่วงน้ำหนัก(weight(w_i))สำหรับตัวแปรอิสระแต่ละตัว คือ

$$w_i^{-1} = \left(\frac{d\eta_i}{d\mu_i}\middle|\eta_i = \eta_i^{(0)}\right)^2 V_i^{(0)}$$

เมื่อ $V_i^{(0)} = Var\left(T\middle|\mu_i^{(0)}\right)$

จะเห็นว่าสมการ(ก)เป็นวิธีประมาณภาวะน่าจะเป็นสูงสุดย้อนซ้ำแบบFisher scoring หรืออาจเขียนให้อยู่ในรูป

$$\sum_j\left[\sum_i \frac{x_{ia}x_{ij}}{Var(T_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \beta^{(k)}{}_j\right] + \sum_i \frac{\left(t_i - \mu_i^{(k)}\right)x_{ia}}{Var(T_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)$$

ดังนั้น $$\beta^{(k+1)} = \left(X'W^{(k)}X\right)^{-1} X'W^{(k)}Z^{(k)}$$

เมื่อ $Z^{(k)} = \eta^{(k)} + \left(T - \mu^{(k)}\right)\left(\frac{d\eta}{d\mu}\middle|\eta = \eta^{(k)}\right)$ และตัวแปรค่าสังเกตเทียม(pseudo observation(z_i)) คือ

$$z_i = \eta_i^{(0)} + \left(T - \mu_i^{(0)}\right)\left(\frac{d\eta_i}{d\mu_i}\middle|\eta = \eta^{(0)}\right)$$

โดยให้ $a(\phi) = \phi = 1$ และ $Var(z_i) = w_i^{-1}$

พิจารณาฟังก์ชันCox's hazard ที่อยู่ในรูป

$$h(T|X) = \exp\{\beta_0(T) + \beta' X\}$$

$$= h_0(T)\exp\{\beta' X\}$$

สามารถสร้างตัวแปรถ่วงน้ำหนัก(weight(w_i)) ได้ในรูป

$$\widehat{w}_i^{(k)} = \widehat{H}_0^{(k)}(t_i)e^{x_i^T\widehat{\beta}^{(k)}}$$

และสร้างตัวแปรค่าสังเกตเทียม(pseudo observation(z_i)) ได้ในรูป

$$\widehat{z}_i^{(k)} = x_i^T\widehat{\beta}^{(k)} + \frac{1}{\widehat{w}_i^{(k)}}\left(\xi_i - \widehat{w}_i^{(k)}\right)$$

เมื่อ $H_0(t) = \sum_{j:y_j \le t_j} \Delta h_0(y_j)$

$y_1 < y_2 < \cdots < y_D$ เป็นค่า t_i ที่แตกต่างกันและ

$\Delta\widehat{h}_0(y_j) = \frac{d_j}{\sum_{i:t_j \ge y_j} e^{x_i^T\beta}}, d_j = \sum_{i:t_i=y_i} \varsigma_i$

หลังจากสร้างตัวแปรถ่วงน้ำหนัก(weight(w_i)) และตัวแปรค่าสังเกตเทียม(pseudo observation(z_i)) ก็เข้าสู่ขั้นตอนของวิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนักแบบย้อนซ้ำ โดยสามารถทำได้ดังนี้

1. กำหนด $k = 0$
2. เลือกค่าเริ่มต้น $\beta^{(k)}$
3. อัพเดทค่า $w_i^{(k)}$ และ $z_i^{(k)}$ ซึ่งหาได้จากสูตรข้างต้น
4. หาค่า $\beta^{(k+1)}$ จากสูตร

$$\beta^{(k+1)} = \left(X'W^{(k)}X\right)^{-1}X'W^{(k)}Z^{(k)}$$

5. ทำซ้ำขั้นตอนที่ 3 – 4จนกว่าค่าประมาณ β จะมีค่าความผันแปรเฉลี่ยมีลักษณะคงที่ตลอดช่วงเวลา (converge)

2.1.3 การคัดเลือกตัวแปร

เนื่องจากในการพิจารณาตัวแบบถดถอยเชิงเส้น จำเป็นต้องเลือกตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามเข้าสู่ตัวแบบการเลือกสมการถดถอยที่เหมาะสมอาจพิจารณาได้ 2 แนวทาง แนวทางแรก คือ การคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบโดยใช้เทคนิคที่มีการพัฒนาขึ้นมา

เพื่อพิจารณาเซตย่อยของตัวแบบการถดถอยเพื่อลดภาระในการต้องพิจารณาตัวแบบที่เป็นไปได้ทั้งหมด โดยการเพิ่มตัวแปรอิสระเข้าและ/หรือลดตัวแปรอิสระออกจากตัวแบบการถดถอย วิธีที่นิยมใช้กันอย่างแพร่หลาย ได้แก่

1. การเลือกตัวแปรโดยวิธีนำตัวแปรเข้าทั้งหมด (Enter Regression)
2. การเลือกตัวแปรโดยวิธีเพิ่มตัวแปร (Forward Selection)
3. การเลือกตัวแปรโดยวิธีเพิ่มตัวแปรอิสระแบบขั้นตอน (Stepwise Regression)
4. การเลือกตัวแปรโดยวิธีลดตัวแปร (Backward Elimination)

โดยแต่ละวิธีมีแนวคิดคร่าวๆดังนี้

1. วิธีการเลือกแบบคัดเลือกเข้า (enter selection) เป็นการคัดเลือกตัวแปรเข้าทั้งหมด โดยใช้วิจารณญาณของผู้วิจัยเองว่าจะคัดเลือกตัวแปรพยากรณ์ใดบ้างเข้าสมการ

2. วิธีการเลือกแบบก้าวหน้า (forward selection) เลือกตัวแปรพยากรณ์ที่มีสหสัมพันธ์กับตัวแปรเกณฑ์สูงที่สุดเข้าสมการก่อน ส่วนตัวแปรที่เหลือจะมีการคำนวณหาสหสัมพันธ์แบบแยกส่วน ถ้าตัวแปรใดมีความสัมพันธ์กันสูงก็นำเข้าสมการต่อไป

3. วิธีการคัดเลือกแบบขั้นตอน (stepwise selection) ขั้นแรกจะเลือกตัวแปรพยากรณ์ที่มีค่าสัมประสิทธิ์สหสัมพันธ์กับตัวแปรเกณฑ์สูงที่สุดเข้าสมการก่อน (backward selection) จากนั้นก็จะทดสอบตัวแปรที่ไม่ได้อยู่ในสมการว่าจะมีตัวแปรพยากรณ์ตัวใดบ้างมีสิทธิ์เข้ามาอยู่ในสมการด้วยวิธีการคักเลือกแบบก้าวหน้า (forward selection)

4. วิธีการเลือกแบบถอยหลัง (backward selection) นำตัวแปรพยากรณ์ทั้งหมดเข้าสมการ จากนั้นหาสหสัมพันธ์ หากทดสอบค่าสัมพันธ์แล้วน้อยที่สุดเอาออก

โดยมีข้อตกลงเบื้องต้น คือ

1. ประชากรมีการแจกแจงแบบปกติ
2. ตัวแปรพยากรณ์กับตัวแปรเกณฑ์มีความสัมพันธ์กันเชิงเส้นตรง
3. ความแปรปรวนของความคลาดเคลื่อนมีความคงที่ทุกค่าการสังเกต
4. ตัวแปรที่นำมาใช้พยากรณ์ไม่ควรมีความสัมพันธ์กันสูงเกินไป

เนื่องจาก จากที่ได้อธิบายมาข้างต้นแล้วว่าตัวแบบ Cox's proportional hazard ไม่ได้มีความสัมพันธ์กับตัวแปรตามในลักษณะเชิงเส้นตรง ซึ่งจากข้อตกลงดังกล่าวทำให้วิธีคัดเลือกตัวแปรที่กล่าวมาไม่สามารถนำมาใช้กับตัวแบบ Cox's proportional hazard ได้ ดังนั้นวิธีทั้ง4 จึงไม่ใช่วิธีที่เหมาะสม

แนวทางที่สองโดยใช้ค่าสถิติเป็นเกณฑ์ในการพิจารณาคัดเลือกความเหมาะสมของทุกตัวแบบที่เป็นไปได้ (All Possible Regression) ค่าสถิติที่นิยมใช้และมีในโปรแกรมสำเร็จรูปทั่วไป ได้แก่ ค่าสัมประสิทธิ์ตัวกำหนด (Coefficient of determination: R^2) ค่าสัมประสิทธิ์ตัวกำหนดปรับแล้ว (Adjusted Coefficient of determination: R^2_{adj}) โดยทั่วไปมักเข้าใจว่าตัวแบบที่เหมาะสมที่สุดคือเมื่อ R^2_{adj} มีค่าสูงสุด แต่ความเป็นจริงแล้วตัวแบบที่ให้ R^2_{adj} สูงที่สุดมักเกิดปัญหามีจำนวนตัวแปรอิสระมากเกินจำเป็น(over-fit) (Sheather, 2009) นอกจากวิธีการตรวจ R^2_{adj} อาจพิจารณาจากค่าสถิติของมอลโลวส์ (Mallows's C_p Statistic) แต่วิธีดังกล่าวจะเกิดปัญหาเมื่อขนาดตัวอย่างเล็ก เกณฑ์ข้อสนเทศของอาไคเคะ (Akaike's Information Criterion: AIC) เกณฑ์ข้อสนเทศของอาไคเคะ ที่มีแนวคิดว่าตัวแบบที่ให้AICน้อยสุดแสดงว่าตัวแบบนั้นเหมาะสมที่สุด ซึ่งเกณฑ์ดังกล่าวใช้ได้ดีเมื่อขนาดตัวอย่างมีขนาดใหญ่ แต่ในกรณีที่เป็นข้อมูลที่มีมิติสูง จะทำให้ตัวแบบที่ได้รับการคัดเลือกมีจำนวนตัวแปรอิสระมากเกินพอดี(over-fit) ซึ่งถูกแก้ไขโดย Hurvich&Tsai, 1989 ที่สร้างวิธี AIC_c ขึ้นมาแต่ก็พบปัญหาเมื่อขนาดตัวอย่างใหญ่ขึ้นกลับกลายเป็นว่าตัวแบบที่ได้รับเลือกจะมีจำนวนตัวแปรอิสระเกินพอดี(over-fit) ข้อสนเทศของชวาร์ซ (Schwarz Bayesian Criteria : SBC)ที่ใช้ค่าสูงสุดของความน่าจะเป็นภายหลัง(posterior probability)มาพิจารณา ค่าความคลาดเคลื่อนกาลังสองเฉลี่ย (Mean Square Error: MSE)

สำหรับการคัดลือกตัวแปรสำหรับตัวแบบCox's proportional hazard สามารถทำได้จากวิธีที่ประยุกต์มาจากการคัดเลือกตัวแปรเชิงเส้นเช่นวิธี selection, the best-subset selection (Forward), backward selection, stepwise และจากการใช้ค่าสถิติเป็นเกณฑ์ในการพิจารณาคัดเลือกความเหมาะสมของทุกตัวแบบที่เป็นไปได้ (All Possible Regression) ตามที่กล่าวมาแล้วข้างต้น แต่วิธีที่กล่าวมาล้วนแต่ซึ่งจะเกิดปัญหาเมื่อข้อมูลมีมิติสูงโดยข้อมูลส่วนใหญ่มีค่าเป็นศูนย์ วิธีที่สามารถคัดเลือกตัวแปรสำหรับตัวแบบCox's proportional hazard ในกรณีที่ข้อมูลมีมิติสูงโดยข้อมูลส่วนใหญ่มีค่าเป็นศูนย์ ทำได้ทั้งวิธี boot strap(Sauerbrei and Schumacher,1992), SIS, ISIS, Sure-screening variable selection, penalized likelihoodที่สามารถสร้างpenalty functionได้ในหลายรูปแบบตามที่กล่าวมาแล้วในบทที่ 1 และวิธีแบบเบส์ ในการศึกษาครั้งนี้ผู้ศึกษาจะเลือกใช้วิธีแบบเบส์ที่สร้างpriorจากข้อมูลค่าสังเกตหรือที่รู้จักกันในชื่อวิธีแบบเบส์เชิงประจักษ์ ซึ่งจะอธิบายแนวคิดและวิธีการในหัวข้อถัดไป

2.2. การคัดเลือกตัวแปรแบบเบส์เชิงประจักษ์

ในปี 2004 Johnstone and Silverman ได้นำเสนอวิธีการในการคัดเลือกตัวแปรโดยใช้หลักการแบบเบส์เชิงประจักษ์ (empirical Bayes methodology) และเพิ่มเงื่อนไขเกี่ยวกับการสร้างthreshold ที่เหมาะสมเข้ามา โดยนำเสนอแนวคิดดังกล่าวในการหาค่าประมาณของข้อมูลที่เป็นลำดับ (sequence) แบบมีมิติสูงที่มีค่าส่วนใหญ่เป็นศูนย์ และพบว่าวิธีนี้ทำงานได้ดีในกรณีดังกล่าว ซึ่งมีรายละเอียดดังนี้

ให้ X_i แทนค่าสังเกต ครั้งที่ i ดังนั้นเมื่อพิจารณาค่าสังเกตจำนวน n ค่าใดๆ เขียนแทนด้วย $X=(X_1,X_2,...,X_n)$ โดยที่

$$X_i=\mu_i+\varepsilon_i \quad ,\varepsilon_i \sim N(0,1)$$

จากสมการข้างต้นจะเห็นว่าการประมาณค่าภาวะน่าจะเป็นสูงสุดสำหรับของ μ_i คือค่า X_i แต่เมื่อให้ μ คือเวกเตอร์ค่าเฉลี่ยของ $\mu=(\mu_1,\mu_2,...,\mu_n)$ การหา $\hat{\mu}$ ก็จะยุ่งยากขึ้น โดยเฉพาะในกรณีที่ จำนวน μ_i มีจำนวนเยอะมากๆ ปัญหานี้รู้จักกันในชื่อปัญหาแบบหลายตัวแปร (Multivariated problem) วิธีหนึ่งที่สามารถนำมาใช้การแก้ไขปัญหาดังกล่าวก็คือแนวคิดของเบส์เชิงประจักษ์ โดยการกำหนดการแจกแจงของ prior จากค่าประมาณของค่าสังเกต (X_i) หรือก็คือ μ_i นั้นเองJohnstone and Silverman จึงได้นำแนวคิดดังกล่าวมาใช้ โดยprior สามารถเขียนรูปของการแจกแจงแบบผสม โดยที่ตัวสถิติที่เพียงพอ (X_i) ของตัวที่ต้องการประมาณ (μ_i) ต้องมีการแจกแจงปกติที่มีค่าเบี่ยงเบนมาตรฐานเท่ากับ 1 ดังนี้

$$f_{prior}(\mu_i)\sim(1-\omega)\delta_0(\mu_i)+\omega\gamma(\mu_i)$$

เมื่อ $\delta_0(.)$ แทน diract delta function ซึ่งเป็นฟังก์ชันความหนาแน่นที่ค่าประมาณมีค่าเป็นศูนย์ โดยที่ $\mu_i=0$ ที่ความน่าจะเป็น $(1-\omega)$ และ $\mu_i\neq 0$ ที่ความน่าจะเป็น ω และ $\gamma(.)$ คือฟังก์ชันความหนาแน่นที่ค่าประมาณมีค่าไม่เท่ากับศูนย์ โดยเสนอรูปแบบการแจกแจงสำหรับ $\gamma(.)$ ไว้ 2 รูปแบบคือการแจกแจง Laplace

$$\gamma(\mu\mid\alpha)=\frac{1}{2}\alpha\exp\{-\alpha|\mu|\},a>0$$

และการแจกแจง quasi-Cauchy

$$\gamma(\mu)=(2\pi)^{-\frac{1}{2}}\left\{1-|\mu|\frac{1-\Phi(|\mu|)}{\phi(\mu)}\right\}$$

Johnstone and Silvermanเสนอวิธีหาค่าประมาณ μ_i เมื่อกำหนด X_i ด้วยค่ามัธยฐานของ posterior (Posterior median) ภายใต้กฎของThreshold และด้วยการประมาณดังกล่าวจะทำให้ค่าประมาณ μ_i ที่ได้ มีค่าส่วนใหญ่เป็นศูนย์ ซึ่งเหมาะสมกับข้อมูลที่ทำการทดสอบ

โดยให้ $\left(\hat{\mu}\left(X;\hat{\omega}\right)\right)$ คือค่ามัธยฐานของการแจกแจงของ posterior ที่อยู่ในเงื่อนไขของ Threshold (τ)ที่ว่า สำหรับค่า ω ที่คงที่ $\exists\tau\left(\omega\right)>0,\hat{\mu}\left(x,\omega\right)=0\leftrightarrow\left|x_i\right|\leq\tau\left(\omega\right)$
กำหนดให้ $X\sim N\left(\mu,1\right)$ และ μ มีการแจกแจงแบบผสม ดังนั้นการแจกแจงของ posterior ก็จะอยู่ในรูปของการแจกแจงผสมระหว่างส่วนที่ค่าประมาณพารามิเตอร์มีค่าเป็นศูนย์และส่วนที่ค่าประมาณพารามิเตอร์มีค่าไม่เท่ากับศูนย์เช่นเดียวกับการแจกแจงของprior จากเงื่อนไขของ Threshold ข้างต้นจำเป็นต้องทราบค่า ω เพื่อนำไปประกอบการพิจารณาหาค่า $\hat{\mu}\left(X;\hat{\omega}\right)$
จากความสัมพันธ์ของ $X_i=\mu_i+\varepsilon_i$ ดังนั้นความหนาแน่นส่วนริมของค่าสังเกต X_i สามารถเขียนได้ในรูป

$$\left(1-\omega\right)\phi\left(X_i\right)+\omega g\left(X_i\right)$$

เมื่อ g คือความหนาแน่นรวมที่เกิดจากความหนาแน่น γ กับ ϕ
เนื่องจาก $X_i\sim N\left(\mu_i,1\right)$ ดังนั้น ϕ คือความหนาแน่นของการแจกแจงปกติ

จาก $X_i\sim P\left(X_i\middle|\mu_i\right)$และ $\mu_i\sim P\left(\mu_i\middle|\omega\right)$ภาวะน่าจะเป็นส่วนริมของค่าสังเกต X_i คือ
$P\left(X_i\middle|\omega\right)=\int_{\mu}P\left(X_i\middle|\mu_i\right)\cdot P\left(\mu_i\middle|\omega\right)d\mu$
ดังนั้นเมื่อกำหนดให้ γ มีความหนาแน่นแบบLaplace จะได้

$$g\left(X_i\right)=\int_{-\infty}^{\infty}\left(\frac{1}{2}\alpha\exp\left\{-\alpha\left|\mu\right|\right\}\right)\cdot\left(\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(X_i-\mu\right)^2\right\}\right)d\mu$$

และ $\hat{\omega}$ หาก็สามารถหาได้จากภาวะน่าจะเป็นส่วนริมสูงสุด หรือจากล็อกของภาวะน่าจะเป็นส่วนริมสูงสุด

$$l\left(\omega\right)=\sum_{i=1}^{n}\left\{\log\left(1-\omega\right)\phi\left(X_i\right)+\omega g\left(X_i\right)\right\}$$

เมื่อหาค่า $\hat{\omega}$ ได้แล้วจึงนำค่าที่ได้ไปใส่ใน prior และเข้าสู่ขั้นตอนพิจารณาเงื่อนไขระหว่างค่ามัธยฐานของposterior (Posterior median$\left(\hat{\mu}\right)$) และเขตกั้น (Threshold ($\tau\left(\omega\right)$))เพื่อใช้ในการพิจารณาหาค่าของ $\hat{\mu}\left(X;\hat{\omega}\right)$ μ ต่อไป

2.3. การคัดเลือกแบบเบส์เชิงประจักษ์สำหรับตัวแบบถดถอยเชิงเส้น

จากหัวข้อที่แล้วได้นำเสนอแนวคิดการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์โดยวิธีการแบบเบส์เชิงประจักษ์(Johnstone and Silverman (2004)) งานวิจัยดังกล่าวได้อธิบายการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ในกรณีที่ข้อมูลเป็นลำดับที่มีค่าส่วนใหญ่เป็นศูนย์ (sparse sequences) Pungpapongและคณะ (2012) ขยายแนวคิดดังกล่าวเข้าสู่ตัวแบบถดถอยเชิงเส้น โดยศึกษากรณีการถดถอยโลจิสติกแบบชนิดbinary logistic ซึ่งอธิบายรายละเอียดไว้ดังนี้ ให้ฟังก์ชันภาวะน่าจะเป็น(Likelihood function) เขียนแทนด้วย

$$L\left(Y;X\beta;\phi\right)$$

โดยที่ Y คือ เวกเตอร์สุ่มขนาด $n\times 1$

X คือ เมทริกซ์ขนาด $n\times p$ ของตัวแปรอิสระจำนวน p ตัว

β คือ เวกเตอร์ของสัมประสิทธิ์ถดถอยขนาด $p\times 1$

ϕ คือ Auxiliary parameters หรือพารามิเตอร์อื่นๆที่อยู่ในตัวแบบ

จากฟังก์ชันภาวะน่าจะเป็นจะเห็นว่าพารามิเตอร์สำหรับตัวแบบนี้มี 2 ตัวคือพารามิเตอร์ β และพารามิเตอร์ ϕ จากแนวคิดในการหาค่าประมาณด้วยวิธีแบบเบส์เชิงประจักษ์ (Johnstone and Silverman (2004)) คือต้องสร้างprior ของพารามิเตอร์ที่ต้องการประมาณค่า

ถ้าให้ prior อยู่ในรูปของการแจกแจงแบบผสม จะเขียนได้ในรูป

$$\beta \sim \left(1-\omega\right)\delta_0\left(\beta\right)+\omega g\left(\beta\right)$$
$$\phi \sim \left(1-\varsigma\right)\delta_0\left(\phi\right)+\varsigma g\left(\phi\right)$$

จาก prior ของพารามิเตอร์ β มี hyperparameterคือ ω ส่วนprior ของพารามิเตอร์ ϕ มี hyperparameterคือ ς

ถ้าให้ตัวแปรบ่งชี้(Latent variable) ที่เขียนแทนด้วย $\delta_j = 1_{\{\beta_j \neq 0\}}$ เป็นเวกเตอร์ขนาด $p\times 1$, ψ_1 เป็นพารามิเตอร์ใน prior ของ β , ψ_2 เป็นพารามิเตอร์ใน prior ของ δ และ ψ_3 เป็นพารามิเตอร์ใน prior ของ ϕ โดยที่ $\psi = \left(\psi_1^t, \psi_2^t, \psi_3^t\right)^t$ ดังนั้น

$$\beta \sim \pi(\beta \mid \delta, \psi_1) \times \pi(\delta \mid \psi_2)$$
$$\phi \sim \pi(\phi \mid \psi_3)$$

ภาวะน่าจะเป็นจึงเขียนได้ในรูป

$$L(Y; X\beta; \phi)\pi(\beta \mid \delta, \psi_1)\pi(\delta \mid \psi_2)\pi(\phi \mid \psi_3)$$

ดังนั้นภาวะน่าจะเป็นสำหรับค่า ψ คือ

$$L(\psi) = \iint L(Y; X\beta; \phi)\pi(\beta \mid \delta, \psi_1)\pi(\delta \mid \psi_2)\pi(\phi \mid \psi_3)\, d\beta d\phi$$

โดยที่ค่าประมาณของพารามิเตอร์ ψ หาได้จากภาวะน่าจะเป็นสูงสุดส่วนริม คือ

$$\hat{\psi} = \arg\max_{\psi}\left(L(\psi)\right)$$

จากการหาค่าประมาณร่วมของ $\hat{\beta} = \hat{\beta}(Y, X, \hat{\psi})$ กับ $\hat{\phi} = \hat{\phi}(Y, X, \hat{\psi})$ ด้วยการหาตัวประมาณแบบเบส์(Bayes estimator) โดยการminimize ค่าคาดหวังของฟังก์ชันการสูญเสีย หรืออีกนัยหนึ่งคือ maximizeค่าคาดหวังของposteriorของฟังก์ชันที่เหมาะสม ซึ่งมีค่าเท่ากับค่าสูงสุดของค่าประมาณposterior เขียนได้ในรูป

$$\left(\hat{\beta}, \hat{\phi}\right) = \arg\min_{\hat{\beta}, \hat{\phi}}\left\{E\left[E\left[L\left(\tilde{\beta}(Y, X, \hat{\psi}), \tilde{\phi}(Y, X, \hat{\psi}); \beta, \phi\right) \middle| \beta, \phi, \hat{\psi}\right] \middle| \hat{\psi}\right]\right\}$$

โดยให้ฟังก์ชันการสูญเสีย(Loss function) อยู่ในรูป

$$L\left(\tilde{\beta}, \tilde{\phi}; \beta, \phi\right) = \left\|\tilde{\beta}(Y, X, \hat{\psi}) - \beta\right\|_1 + \left\|\tilde{\phi}(Y, X, \hat{\psi}) - \phi\right\|_0$$

เมื่อ $\|.\|_1$ แทน $l_1 - norm$ และ $\|.\|_0$ แทน $l_0 - norm$

ฟังก์ชันความเสี่ยงแบบเบส์หาได้จากค่าคาดหวังของฟังก์ชันการสูญเสียสามารถเขียนได้ในรูป

$$R\left(\tilde{\beta}, \tilde{\phi}; \beta, \phi\right) = E\left[L\left(\tilde{\beta}, \tilde{\phi}; \beta, \phi\right)\right] = E\left[\left\|\tilde{\beta}(Y, X, \hat{\psi}) - \beta\right\|_1 + \left\|\tilde{\phi}(Y, X, \hat{\psi}) - \phi\right\|_0 \middle| \hat{\psi}\right]$$

เนื่องจาก $L(a) = \|a\| = \sum_j L(a_j)$ ดังนั้น

$$R\left(\tilde{\beta}, \tilde{\phi}; \beta, \phi\right) = E\left[\sum_j E\left[\left|\tilde{\beta}_j(Y, X, \hat{\psi}) - \beta_j\right| Y, X, \hat{\psi}\right] + \sum_j E\left[1\left\{\tilde{\phi}_j(Y, X, \hat{\psi}) \neq \phi_j\right\} \middle| Y, X, \hat{\psi}\right] \middle| \hat{\psi}\right]$$

จากฟังก์ชันความเสี่ยงแบบเบส์ที่ได้จะเห็นว่าถูกประกอบขึ้นด้วยสองส่วนคือส่วนที่เกี่ยวข้องและอยู่ในรูปของพารามิเตอร์ β (ส่วนแรก)และอีกส่วนคือส่วนที่เกี่ยวข้องและอยู่ในรูปของพารามิเตอร์ ϕ (ส่วนหลัง)ซึ่งทั้งสองตัวคือพารามิเตอร์ทั้งหมดของตัวแบบตามที่กล่าวมาก่อนหน้านี้ โดยมีเงื่อนไขว่าค่าไฮเปอร์พารามิเตอร์(พารามิเตอร์ของprior)เป็นค่าคงที่หรือค่าที่เราทราบค่าอยู่ก่อนแล้ว ดังนั้นการหาค่าพารามิเตอร์ $\hat{\beta}_j$ ด้วยวิธีแบบเบส์หาได้จากการ minimizesฟังก์ชันความเสี่ยงแบบเบส์ซึ่งมีค่าเท่ากับการหาค่าposterior median

$$\begin{aligned}\hat{\beta}_j &= \arg\min_{\beta}\left[E\left[\sum_j E\left[\left| \tilde{\beta}_j\left(Y,X,\hat{\psi}\right) - \beta_j \right| \middle| Y,X,\hat{\psi} \right] + \sum_j E\left[1\left\{ \tilde{\phi}_j\left(Y,X,\hat{\psi}\right) \neq \phi_j \right\} \middle| Y,X,\hat{\psi} \right] \middle| \hat{\psi} \right] \right] \\ &= \text{median}\left(\beta_j \middle| Y,X,\hat{\psi} \right)\end{aligned}$$

และ $\hat{\phi}_j$ หาได้จากการminimizesฟังก์ชันความเสี่ยงแบบเบส์

$$\hat{\phi}_j = \arg\min_{\phi}\left[E\left[\sum_j E\left[\left| \tilde{\beta}_j\left(Y,X,\hat{\psi}\right) - \beta_j \right| \middle| Y,X,\hat{\psi} \right] + \sum_j E\left[1\left\{ \tilde{\phi}_j\left(Y,X,\hat{\psi}\right) \neq \phi_j \right\} \middle| Y,X,\hat{\psi} \right] \middle| \hat{\psi} \right] \right]$$

ตามที่แสดงไว้โดยJohnstone and Silverman (2004) และZhang, Zhang and Wells (2010) จากฟังก์ชันการสูญเสีย (Loss function) เมื่อพิจารณาภายใต้เงื่อนไขการประมาณด้วยเขตกั้น ผลลัพธ์ที่ได้จะมีค่าประมาณ β ส่วนใหญ่เป็นศูนย์

2.3.1. วิธีการทำซ้ำแบบมีเงื่อนไขฐานนิยมและมัธยฐาน (ICM/M)

Pungpapongและคณะ(2012) ได้เสนอวิธีการทำซ้ำแบบมีเงื่อนไข ฐานนิยม/มัธยฐาน (ICM/M) โดยวิธีดังกล่าวเป็นวิธีการสำหรับช่วยในการคำนวณหาค่าประมาณพารามิเตอร์ให้รวดเร็ว ซึ่งวิธีการICM/Mมีลักษณะแนวคิดหลักเช่นเดียวกับวิธี Iterated conditional modes(ICM) ที่เสนอโดยBesag(1986) เมื่อกล่าวถึงวิธี ICM เป็นวิธีการที่มีคุณสมบัติที่เหมาะสมและมีประสิทธิภาพในการประมาณค่าพารามิเตอร์แบบทั่วไปและแบบที่ข้อมูลมีมิติสูง (High-dimensional) ได้ดีกว่าการประมาณค่าพารามิเตอร์โดยการใช้สมการถดถอยและวิธีการแบบ MCMC ดังนั้นวิธีICM/M ที่ถูกพัฒนามาจากวิธีICMจึงยังคงรักษาคุณสมบัติที่ดีเอาไว้ หรือกล่าวได้ว่าวิธีICM/M มีความเหมาะสมและมีประสิทธิภาพในการประมาณค่าพารามิเตอร์แบบที่ข้อมูลมีมิติสูง(High-dimensional)ด้วยเช่นกัน

ในหัวข้อนี้จะกล่าวถึงวิธีการและแนวคิดของการทำซ้ำแบบมีเงื่อนไข ฐานนิยม/มัธยฐาน (ICM/M) ที่ช่วยในการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์แบบเบส์เชิงประจักษ์ให้ทำได้เร็วยิ่งขึ้นตามงานวิจัยของPungpapongและคณะ

จากวิธีภาวะน่าจะเป็นสูงสุดในการประมาณค่าพารามิเตอร์ในตัวแบบแต่ละตัว สามารถหาได้จากการหาค่าประมาณภาวะน่าจะเป็นสูงสุดส่วนริม(maximum marginal likelihood estimated) ในกรณีที่ข้อมูลมีจำนวนตัวแปรอิสระมากจะส่งผลให้จำนวนค่า β มากตามไปด้วย ข้อมูลลักษณะดังกล่าวจะทำให้การแจกแจงposteriorของ β มีความซับซ้อนมากส่งผลให้การคำนวณหาค่า β ยุ่งยากและใช้เวลานาน Pungpapong et al., 2012 ได้ยกตัวอย่างเอาไว้ดังนี้ ให้ $\pi\left(\beta_j \middle| Y, X, \widehat{\psi}\right)$ แทน การแจกแจงของposterior ของ β ซึ่งประกอบไปด้วยพารามิเตอร์ β จำนวน p ตัวและพารามิเตอร์ ϕ ที่เป็นauxiliary parameter จำนวน q ตัว ซึ่งสามารถเขียนได้ในรูป

$$\pi\left(\beta_j \middle| Y, X, \widehat{\psi}\right) = \int\left[\int \pi\left(\beta_j \middle| Y, X, \beta_{-j}, \phi, \widehat{\psi}\right) d\beta_{-j}\right] d\phi$$

เมื่อ β_{-j} คือ β ทุกตัวยกเว้นตัวที่ j

ϕ_{-j} คือ ϕ ทุกตัวยกเว้นตัวที่ j

ค่าประมาณของ β_j คำนวณได้จาก posterior median

$$\widehat{\beta}_j = \widehat{\beta}_j\left(\widehat{\beta}_{-j}, \widehat{\phi}, \widehat{\psi}\right) = \text{median}\left(\beta_j \middle| Y, X, \widehat{\beta}_{-j}, \widehat{\phi}, \widehat{\psi}\right)$$

ค่าประมาณของ ϕ_j สามารถคำนวณได้จาก minimizes ฟังก์ชันความเสี่ยงแบบเบส์ซึ่งมีค่าเท่ากับการหา posterior mode

$$\widehat{\phi}_j = \widehat{\phi}_j\left(\widehat{\beta}, \widehat{\phi}_{-j}, \widehat{\psi}\right) = \text{mode}\left(\phi_j \middle| Y, X, \widehat{\beta}, \widehat{\phi}_{-j}, \widehat{\psi}\right)$$

ในที่นี้มีพารามิเตอร์เพิ่มเติมที่เข้ามาเกี่ยวข้องในการคำนวณคือ ψ ซึ่งเป็นพารามิเตอร์ชั้นที่ 2(hyperparameters) ซึ่งเป็นพารามิเตอร์ในส่วนของprior จำนวน r ตัวโดยที่ค่าประมาณของ ψ_j หาได้จากการ minimizes ฟังก์ชันความเสี่ยงแบบเบส์ซึ่งมีค่าเท่ากับการหาposterior mode

$$\begin{aligned}
\widehat{\psi}_j &= \text{mode}\left(\psi_j \middle| Y, X, \widehat{\psi}_{-j}\right) \\
&= \text{mode}\left(\iint \pi\left(\psi_j \middle| Y, X, \beta, \phi, \widehat{\psi}_{-j}\right) d\beta d\phi\right) \\
&= \text{mode}\left(\psi_j \middle| Y, X, \widehat{\beta}, \widehat{\phi}, \widehat{\psi}_{-j}\right)
\end{aligned}$$

เนื่องจากขบวนการสำหรับวิธีการนี้เป็นขบวนการทำซ้ำ เราแทนการประมาณค่าของพารามิเตอร์ทั้งสามสำหรับการทำซ้ำรอบที่ k ได้ด้วยสัญลักษณ์ $\hat{\beta}^{(k)}, \hat{\phi}^{(k)}$ และ $\hat{\psi}^{(k)}$ ซึ่งสามารถเขียนการหาค่าประมาณของพารามิเตอร์ทั้งสามรอบที่ k+1 ได้ในรูป

$$\hat{\beta}_j^{(k+1)} = \text{median}\left(\beta_j \middle| Y, X, \hat{\beta}_{1:(j-1)}^{(k+1)}, \hat{\beta}_{(j+1):p}^{(k)}, \hat{\phi}^{(k)}, \hat{\psi}^{(k)}\right)$$

$$\hat{\phi}_j^{(k+1)} = \text{mode}\left(\phi_j \middle| Y, X, \hat{\beta}^{(k+1)}, \hat{\phi}_{1:(j-1)}^{(k+1)}, \hat{\phi}_{(j+1):q}^{(k)}, \hat{\psi}^{(k)}\right)$$

$$\hat{\psi}_j^{(k+1)} = \text{mode}\left(\psi_j \middle| Y, X, \hat{\beta}^{(k+1)}, \hat{\phi}^{(k+1)}, \hat{\psi}_{1:(j-1)}^{(k+1)}, \hat{\psi}_{(j+1):r}^{(k)}\right)$$

ดังนั้นเราสามารถสรุปขั้นตอนการทำงานวิธีICM/Mได้ดังนี้

1. กำหนดสัมประสิทธิ์เริ่มต้นให้เป็น $\hat{\beta}_j^{(0)}$ จำนวน p ตัว เมื่อ $j = 1, 2, \ldots, p$

2. อัพเดทค่าประมาณ ϕ จำนวน q ตัว จาก

$$\hat{\phi}_j^{(k+1)} = \text{mode}\left(\phi_j \middle| Y, X, \hat{\beta}^{(k+1)}, \hat{\phi}_{1:(j-1)}^{(k+1)}, \hat{\phi}_{(j+1):q}^{(k)}, \hat{\psi}^{(k)}\right)$$

3. อัพเดทค่าประมาณ ψ จำนวน r ตัว จาก

$$\hat{\psi}_j^{(k+1)} = \text{mode}\left(\psi_j \middle| Y, X, \hat{\beta}^{(k+1)}, \hat{\phi}^{(k+1)}, \hat{\psi}_{1:(j-1)}^{(k+1)}, \hat{\psi}_{(j+1):r}^{(k)}\right)$$

4. อัพเดทค่าประมาณ β จำนวน p ตัว จาก

$$\hat{\beta}_j^{(k+1)} = \text{median}\left(\beta_j \middle| Y, X, \hat{\beta}_{1:(j-1)}^{(k+1)}, \hat{\beta}_{(j+1):p}^{(k)}, \hat{\phi}^{(k)}, \hat{\psi}^{(k)}\right)$$

5. ทำตามขั้นตอนที่ 2-5 จนกว่าค่าประมาณจะลู่เข้า

บทที่ 3

วิธีดำเนินการศึกษา

ในงานวิจัยครั้งนี้เป็นการศึกษาการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์แบบเบย์เชิงประจักษ์ที่ทำงานร่วมกับวิธีการทำซ้ำแบบมีเงื่อนไขฐานนิยม/มัธยฐานสำหรับตัวแบบCox's proportional hazard ในกรณีที่ข้อมูลที่ต้องการศึกษามีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง พร้อมทั้งเปรียบเทียบประสิทธิภาพการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ด้วยวิธีข้างต้นเมื่อจำลองข้อมูลที่มีลักษณะแตกต่างกัน เพื่อตรวจสอบความมีประสิทธิภาพของวิธีการแบบเบย์เชิงประจักษ์ที่ทำงานร่วมกับวิธีการทำซ้ำแบบมีเงื่อนไขฐานนิยม/มัธยฐาน โดยใช้อัตราความผิดพลาดในการตรวจจับเชิงบวกและอัตราความผิดพลาดในการตรวจจับเชิงลบ ในการจำลองข้อมูล การคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ไปจนถึงการตรวจสอบอัตราความผิดพลาดเชิงบวกและเชิงลบกล่าวคือ ทุกขั้นตอนในการทำการศึกษาผู้วิจัยทำงานโดยใช้โปรแกรม R เวอร์ชั่น 2.15.2 ซึ่งมีแผนการจำลองข้อมูลและขั้นตอนในการวิจัยดังนี้

3.1. แผนการทำงาน

ในงานวิจัยนี้จะทำการศึกษาประสิทธิภาพของขบวนการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ด้วยวิธีแบบเบย์เชิงประจักษ์ที่ทำงานร่วมกับวิธีICM/MในตัวแบบCox's proportional hazard โดยสร้างสถานการณ์จำลองที่มีความแตกต่างกันทั้งหมด 9 กรณี เพื่อใช้ในการตรวจสอบ ซึ่งทั้ง9 กรณีศึกษาจะมีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง โดยมีเงื่อนไขในการจำลองดังนี้

1. จำลองเวลาการอยู่รอดที่มีการแจกแจงแบบไวบูลล์ 9 กรณีศึกษาโดยให้ขนาดตัวอย่างมีขนาดเท่ากับ 100 และตัวแปรอิสระที่มีจำนวนเท่ากับ 300, 500และ1,000 และร้อยละของข้อมูลเซ็นเซอร์คือ 10% 50%และ70%
2. คัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ด้วยวิธีแบบเบย์เชิงประจักษ์ที่ทำงานร่วมกับวิธีICM/M
3. คำนวณอัตราความผิดพลาดในการตรวจจับเชิงบวกและอัตราความผิดพลาดในการตรวจจับเชิงลบ
4. วิเคราะห์ผลลัพธ์

3.2. ขั้นตอนการทำงาน

1.จำลองเวลาการอยู่รอด มีขั้นตอนดังนี้

1.1 ให้ตัวแปรอิสระที่นำมาศึกษามีการแจกแจงแบบปกติโดยแต่ละตัวเป็นอิสระต่อกัน

$$x_i \overset{iid}{\sim} N(0,1)$$

1.2 กำหนดค่าพารามิเตอร์ β ดังนี้

β ตัวที่ 1 ถึง 10 มีค่าเท่ากับ 5

β ตัวที่ 101 ถึง 110 มีค่าเท่ากับ 2 นอกนั้นให้มีค่าเท่ากับศูนย์

1.3 จำลองเวลาการอยู่รอดที่มีการแจกแจงแบบไวบูลล์ ทั้ง 9 กรณีศึกษา

จากตัวแบบ Cox's proportional hazard $h(T|X) = h_0(T)\exp\{\beta' X\}$ เมื่อ T คือเวลา, X คือเวกเตอร์ของตัวแปรต้น, β คือเวกเตอร์ของสัมประสิทธิ์ถดถอยและ $h_0(T)$ คือฟังก์ชันhazard baseline หรือฟังก์ชันhazardเมื่อตัวแปรต้นเป็นศูนย์ $(X = 0)$ อย่างไรก็ตามผลกระทบจากตัวแปรตามนั้นส่งผลต่อเวลาการอยู่รอด(survival time) เนื่องจากในการใช้ซอร์ฟแวร์แพ็คเกจสำหรับตัวแบบCox จะต้องระบุค่าเวลาการอยู่รอดที่มีค่าเฉพาะที่แตกต่างกันเมื่อค่าตัวแปรต้นแตกต่างกัน การแปลงสัมประสิทธิ์ถดถอยจากฟังก์ชันความเสี่ยง(hazard)เป็นฟังก์ชันเวลาการอยู่รอดนั้นทำได้ง่ายเมื่อค่า $h_0(T)$ เป็นค่าคงที่ ในการศึกษาครั้งนี้เราพิจารณาที่เวลาการอยู่รอดมีการแจกแจงแบบไวบูล์(Weibull) การแจกแจงไวบูล์สามารถจำแนกความแตกต่างโดยการสร้างจากพารามิเตอร์ที่แตกต่างกันในแต่ละเซตของแต่ละกลุ่ม พารามิเตอร์ที่กล่าวถึงนั้นสามารถเลือกได้จากค่า proportional hazard และค่าอัตราส่วนhazardที่แท้จริง(true hazard ratio(HR)) ในการเปรียบเทียบ2กลุ่มสามารถคำนวณได้จากพารามิเตอร์ของไวบูล์ จากนั้นการหาค่าสัมประสิทธิ์ถดถอยที่แท้จริงสำหรับตัวแบบ Cox ก็สามารถหาได้จากlog(HR)

เวลาการอยู่รอดที่มีการแจกแจงไวบูล์ในตัวแบบ Cox's proportional hazard (Ralf Benderและคณะ (2005)) เขียนได้ในรูป

$$T = \left(-\frac{\log(U)}{\lambda\exp(\beta' X)}\right)^{1/\nu}$$

hazard function อยู่ในรูป

$$h(t|X) = \lambda\exp(\beta' X)\nu t^{\nu-1}$$

$$= \left(\lambda\nu t^{\nu-1}\right)\left(\exp(\beta' X)\right)$$

$$= h_0(t)\left(\exp(\beta' X)\right)$$

ฟังก์ชันสะสมของ hazard สำหรับการแจกแจงแบบไวบูลล์ คือ

$$\begin{aligned} H(t,x,\varsigma) &= \int_0^t \exp(\beta' x) h_0(u)\, du \\ &= \lambda \exp(\beta' x) \int_o^t \nu u^{\nu-1} du \\ &= \lambda \exp(\beta' x)\left[u^{\nu}\right]_0^t \\ &= \lambda \exp(\beta' x) t^{\nu} \end{aligned}$$

เนื่องจากฟังก์ชัน hazard baseline: $h_0(t) = \lambda \nu t^{\nu-1}$ ดังนั้นฟังก์ชันสะสมของ hazard baselines จะเขียนได้ในรูป

$$\begin{aligned} H_0(t) &= \int_0^t h_0(u)\, du \\ &= \int_0^t \lambda \nu u^{\nu-1} du \\ &= \lambda \nu \left(u^{\nu} \Big|_{u=0}^{u=t} \right) = \lambda t^{\nu} \end{aligned}$$

ฟังก์ชันผกผันของฟังก์ชันสะสม hazard baseline: $H_0^{-1}(t) = \left(\lambda^{-1} t\right)^{1/\nu}$

จากฟังก์ชันการอยู่รอด

$$S(t \mid x) = \exp\left(-H_0(t)\exp(\beta' x)\right)$$

ดังนั้น

$$\begin{aligned} S(t \mid x = 0) &= S_0(t) \\ &= \exp\left(-H_0(t)\right) \\ &= \exp\left(-\lambda t^{\nu}\right) \end{aligned}$$

ฟังก์ชันความหนาแน่น

$$\begin{aligned} f_0(t) &= h_0(t)\exp\left(-H_0(t)\right) \\ &= \lambda \nu t^{\nu-1} \exp\left(-\lambda t^{\nu}\right) \end{aligned}$$

ค่าเฉลี่ย

$$E(T)=\frac{1}{\sqrt[\nu]{\lambda}}\Gamma\left(\frac{1}{\nu}+1\right)$$

ความแปรปรวน

$$Var(T)=\frac{1}{\sqrt[\nu]{\lambda^2}}\left[\Gamma\left(\frac{2}{\nu}+1\right)-\Gamma^2\left(\frac{1}{\nu}+1\right)\right]$$

เมื่อ Scale parameter $\nu>0$, Shape parameter $\lambda>0$ และ และ$U\sim U[0,1]$

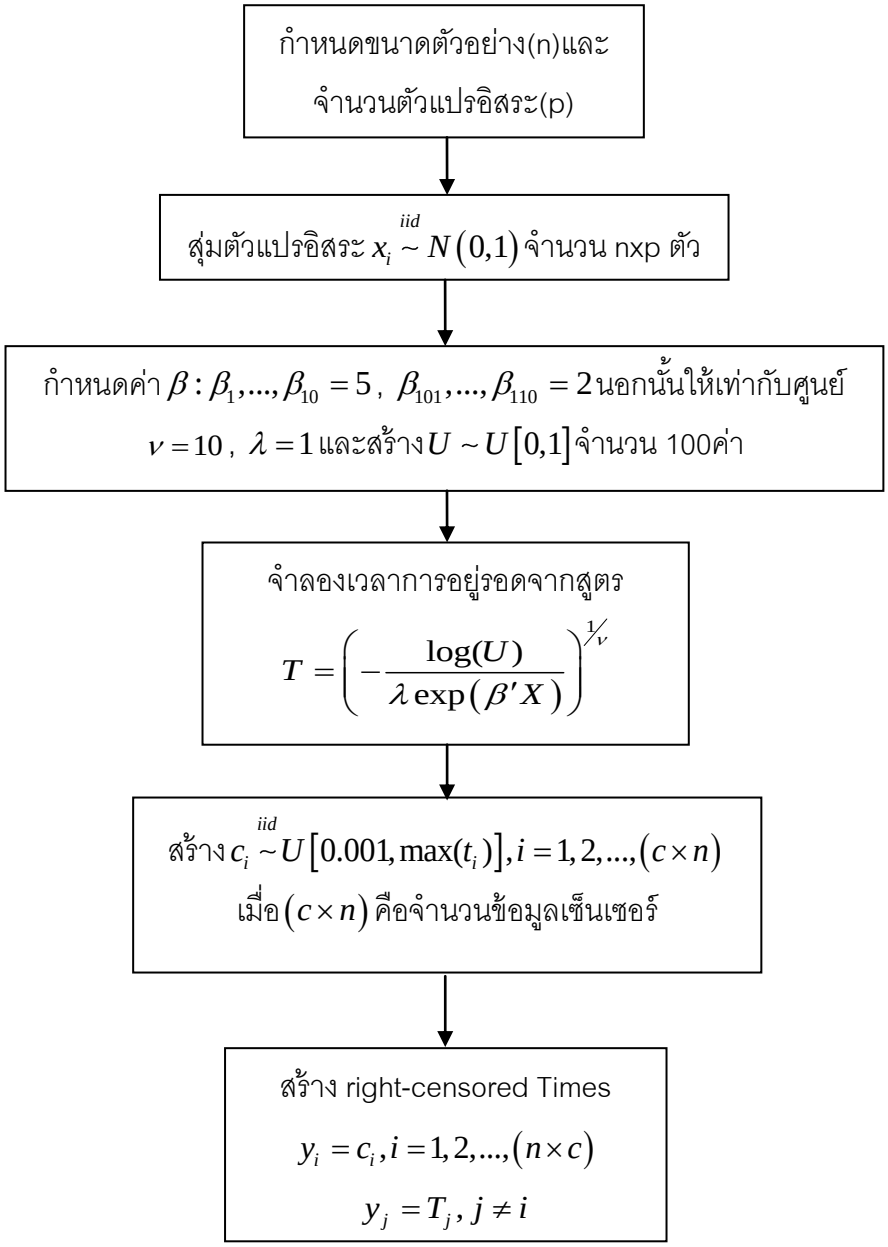
ในการศึกษาครั้งนี้ให้ $\nu=10$, $\lambda=1$

โดยทั้ง 9กรณีศึกษามีรายละเอียดดังนี้

ในการทดลองครั้งนี้ให้ค่า n=100, p=300, 500และ1000 ที่ร้อยละของข้อมูลสูญหายที่ 10% 50% และ 70% ดังนั้นข้อมูลที่ได้จะแบ่งเป็น 9 กรณีคือ

- กรณีที่ 1: n=100, p=300 ร้อยละของข้อมูลเซ็นเซอร์คือ 10%
- กรณีที่ 2: n=100, p=300 ร้อยละของข้อมูลเซ็นเซอร์คือ 50%
- กรณีที่ 3: n=100, p=300 ร้อยละของข้อมูลเซ็นเซอร์คือ 70%
- กรณีที่ 4: n=100, p=500 ร้อยละของข้อมูลเซ็นเซอร์คือ 10%
- กรณีที่ 5: n=100, p=500 ร้อยละของข้อมูลเซ็นเซอร์คือ 50%
- กรณีที่ 6: n=100, p=500 ร้อยละของข้อมูลเซ็นเซอร์คือ 70%
- กรณีที่ 7: n=100, p=1,000 ร้อยละของข้อมูลเซ็นเซอร์คือ 10%
- กรณีที่ 8: n=100, p=1,000 ร้อยละของข้อมูลเซ็นเซอร์คือ 50%
- กรณีที่ 9: n=100, p=1,000 ร้อยละของข้อมูลเซ็นเซอร์คือ 70%

<u>ภาพที่ 3.1</u> แผนผังการเขียนโปรแกรมจำลองเวลาการอยู่รอด



หมายเหตุ: ที่ร้อยละของข้อมูลเซ็นเซอร์10%, 50%และ 70%คือ $c = 0.1, 0.5$ และ 0.7 ตามลำดับ

2. คัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์แบบเบส์เชิงประจักษ์ ที่ทำงานร่วมกับวิธี ICM/M มีขั้นตอนดังนี้

2.1 กำหนดให้ Likelihood และ prior คือ ของตัวแบบ Cox's proportional hazard

Likelihood: $$L=\prod_{i=1}^{n}\left[\left\{h_0\left(t_i\right)e^{x_i^T\beta}\right\}^{\zeta_i}\exp\left\{-H_0\left(t_i\right)e^{x_i^T\beta}\right\}\right]$$

เมื่อ $H_0\left(t\right)=\sum_{j:y_j\le t_j}\Delta h_0\left(y_j\right)$ โดยที่ $y_1<y_2<\cdots<y_D$ เป็นค่า t_i ที่แตกต่างกัน

และ $\Delta\hat{h}_0\left(y_j\right)=\dfrac{d_j}{\sum_{i:t_j\ge y_j}e^{x_i^T\beta}}, d_j=\sum_{i:t_i=y_i}\zeta_i$

Prior: $$\beta_j\sim\left(1-\omega\right)\delta_0\left(\beta_j\right)+\omega\gamma\left(\beta_j\right)$$

เมื่อ $$\gamma\left(\beta\,|\,\alpha\right)=\frac{1}{2}\alpha\exp\left\{-\alpha|\beta|\right\},\alpha>0$$

ในการศึกษาครั้งนี้เราใช้ค่า $\alpha=0.5$ (แนะนำโดย Johnstone and Silverman (2004, 2005))

2.2 การคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์

ให้ $T=\left(t_1,t_2,...,t_n\right)$ คือ เวกเตอร์ของRight-censored timeของตัวอย่างขนาด n ตัวอย่าง $X=\left(x_1^T,x_2^T,...,x_n^T\right)^T$ เป็นเมทริกซ์ขนาด $n\times p$ ของตัวแปรอิสระจำนวน p ตัว

ตัวแบบCox's hazardคือ

$$h(T|\,X)=\exp\left\{\beta_0\left(T\right)+\beta'X\right\}$$

$$=h_0\left(T\right)\exp\left\{\beta'X\right\}$$

โดยที่

$$\hat{w}_i^{(k)}=\hat{H}_0^{(k)}\left(t_i\right)e^{x_i^T\hat{\beta}^{(k)}}\text{ และ }\hat{z}_i^{(k)}=x_i^T\hat{\beta}^{(k)}+\frac{1}{\hat{w}_i^{(k)}}\left(\zeta_i-\hat{w}_i^{(k)}\right)$$

เมื่อ $H_0\left(t\right)=\sum_{j:y_j\le t_j}\Delta h_0\left(y_j\right)$

$y_1 < y_2 < \cdots < y_D$ เป็นค่า t_i ที่แตกต่างกันและ $\Delta\hat{h}_0\left(y_j\right) = \dfrac{d_j}{\sum_{i:t_j \geq y_j} e^{x_i^T \beta}}, d_j = \sum_{i:t_i = y_i} \zeta_i$

ดังนั้น $z_i \approx N\left(x_i^T \hat{\beta}, w_i^{-1}\right)$ พิจารณา β_j, $j = 1, 2, ..., p$ โดยสมมติว่าค่าพารามิเตอร์ที่เหลือทราบค่า เราจะได้ค่าสถิติที่เพียงพอ(sufficient statistic) สำหรับ β_j คือ

$$\frac{\sum_{i=1}^{n} w_i x_{ij} \tilde{z}_i}{\sum_{i=1}^{n} w_i x_{ij}^2} \sim N\left(\beta_j, \frac{1}{\sum_{i=1}^{n} w_i x_{ij}^2}\right)$$

เมื่อ $\tilde{z}_i = z_i - x_i^T \beta + x_{ij}\beta_j$ ให้ $\tilde{\beta}_j = \left(\sqrt{\sum_{i=1}^{n} w_i x_{ij}^2}\right)\beta_j$ จาก prior แบบผสมสำหรับแต่ละ $\tilde{\beta}_j$ ที่อิสระกันเพื่อใช้ในการตรวจจับค่าประมาณเมื่อค่าส่วนใหญ่เป็นศูนย์สามารถเขียนได้ในรูป

$$\tilde{\beta}_j \sim (1-\omega)\delta_0\left(\tilde{\beta}_j\right) + \omega\gamma\left(\tilde{\beta}_j \middle| \alpha\right)$$

ในการศึกษาครั้งนี้เราจะใช้ค่า $\alpha = 0.5$ และใช้ $\gamma(.)$ ในรูปความหนาแน่น Laplace ซึ่งเป็นฟังก์ชันในส่วนที่ค่าประมาณมีค่าไม่เท่ากับศูนย์จากคำแนะนำของ Johnstone and Silverman (2004) ดังนั้น

$$\gamma\left(\tilde{\beta}_j \middle| \alpha\right) = \frac{1}{2}\alpha \exp\left(-\alpha\left|\tilde{\beta}_j\right|\right)$$

แทนค่า $\alpha = 0.5$ จะได้

$$\gamma\left(\tilde{\beta}_j\right) = \frac{1}{4}\exp\left(-\frac{1}{2}\left|\tilde{\beta}_j\right|\right)$$

จากขบวนการกำลังสองน้อยสุดถ่วงน้ำหนักแบบย้อนซ้ำ (IRLS)ดังนี้เราทราบว่าตัวสถิติที่เพียงพอ (sufficient statistic) สำหรับ β_j คือ $\left(\sum_{i=1}^{n} w_i x_{ij} \tilde{z}_i \Big/ \sum_{i=1}^{n} w_i x_{ij}^2\right)$ แต่จากเงื่อนไขของ Johnstone and Silverman (2004,2005) ตัวสถิติที่เพียงพอของตัวที่จะประมาณค่าต้องมีการแจกแจงปกติที่ความแปรปรวนเท่ากับ 1 ดังนั้นจึงให้ $\tilde{\beta}_j = \left(\sqrt{\sum_{i=1}^{n} w_i x_{ij}^2}\right)\beta_j$ จากความสัมพันธ์ดังกล่าวจะได้ตัวสถิติที่เพียงพอสำหรับ $\tilde{\beta}_j$ คือ $\left(\sum_{i=1}^{n} w_i x_{ij} \tilde{z}_i \Big/ \sqrt{\sum_{i=1}^{n} w_i x_{ij}^2}\right)$ ตามเงื่อนไขของ Johnstone

and Silverman (2004,2005) โดยค่าประมาณของ $\tilde{\beta}_j$ หาได้จากposterior medianโดยใช้เทคนิคเพื่อช่วยในการคำนวณหาค่าจากวิธี ICM/M(Pungpapongและคณะ(2012)) เราสามารถสร้างโครงสร้างให้มีรูปแบบเป็นมัธยฐานของposterior ของ β_j ตามการBayesian analysisได้ดังนี้

$$\begin{cases} \left| \dfrac{\sum_{i=1}^{n} w_i x_{ij} \tilde{z}_i}{\sqrt{\sum_{i=1}^{n} w_i x_{ij}^2}} \right| \beta_j \sim N\left(\left(\sqrt{\sum_{i=1}^{n} w_i x_{ij}^2} \right) \beta_j, 1 \right) \\ \beta_j \sim (1-\omega)\delta_0(\beta_j) + \dfrac{1}{4}\omega \left(\sqrt{\sum_{i=1}^{n} w_i x_{ij}^2} \right) \exp\left\{ -\dfrac{1}{2} \left(\sqrt{\sum_{i=1}^{n} w_i x_{ij}^2} \right) |\beta_j| \right\} \end{cases}$$

จากแนวคิดข้างต้นสามารถเขียนเป็นขั้นตอนการทำงานได้ดังนี้

1. กำหนดค่าเริ่มต้นของ β และ ω โดยให้ชื่อว่า $\hat{\beta}_j^{(0)}$ และ $\hat{\omega}$ เมื่อ $j = 1, 2, ..., p$

2. รับค่าถ่วงน้ำหนัก $\hat{W}^{(k)}$ และค่า สังเกตเทียม $\hat{Z}^{(k)}$ ซึ่งคำนวณได้จากสูตร

$$\hat{W}^{(k)} = diag\left\{ \hat{w}_i^{(k)} \right\}, \hat{w}_i^{(k)} = \hat{H}_0^{(k)}(t_i) e^{x_i^T \hat{\beta}^{(k)}}$$

เมื่อ $$\hat{w}^{(k)} = \left(\hat{w}_1^{(k)}, ..., \hat{w}_n^{(k)} \right)^T$$

และ $$\hat{Z}^{(k)} = X\hat{\beta}^{(k)} + \left(\hat{W}^{(k)} \right)^{-1} \left(\zeta - \hat{w}^{(k)} \right)$$

3. คำนวณค่า $\hat{\beta}_j^{(k+1)}$ เมื่อ $j = 1, 2, ..., p$ จากposterior median โดยให้

$$u_j = \left(\frac{\sum_{i=1}^{n} w_i x_{ij} \tilde{z}_i}{\sqrt{\sum_{i=1}^{n} w_i x_{ij}^2}} \right)$$

ดังนั้น

$$\hat{\beta}_j^{(k+1)}(u_j) = \text{median} P\left(\beta_j \middle| X, \hat{Z}^{(k)}, \hat{W}^{(k)}, \hat{\beta}_{1:(j-1)}^{(k+1)}, \hat{\beta}_{(j+1):p}^{(k)}, \hat{\omega}^{(k)} \right)$$

$$= \begin{cases} 0 & ;\omega_{post,j}\tilde{F}_1\left(0|u_j\right) \le \frac{1}{2} \\ \frac{1}{\sqrt{\sum_{i=1}^{n} w_i^k x_{ij}^2}}\left[\Phi^{-1}\left(1-\frac{\left(1-\Phi\left(u_j+1/2\right)\right)e^{u_j}+\Phi\left(u_j-1/2\right)}{2\omega_{post,j}}\right)+\left(u_j-\frac{1}{2}\right)\right] & ;\text{otherwise} \end{cases}$$

เมื่อ

$$\omega_{\text{post,j}} = P\left(\beta_j \neq 0 \middle| X\right)$$

$$= \frac{\hat{\omega} g\left(x\right)}{\left(1-\omega\right)\phi\left(x\right)+\hat{\omega} g\left(x\right)}$$

$$= \frac{\frac{1}{4}\hat{\omega}\left(\sqrt{\sum_{i=1}^{n} w_i x_{ij}^2}\right)\exp\left\{-\frac{1}{2}\left(\sqrt{\sum_{i=1}^{n} w_i x_{ij}^2}\right)\left|\beta_j\right|\right\}}{\left(1-\omega\right)\phi\left(x\right)+\frac{1}{4}\hat{\omega}\left(\sqrt{\sum_{i=1}^{n} w_i x_{ij}^2}\right)\exp\left\{-\frac{1}{2}\left(\sqrt{\sum_{i=1}^{n} w_i x_{ij}^2}\right)\left|\beta_j\right|\right\}}$$

และ

$$\tilde{F}_1\left(\beta|u_j\right) = \int_{\beta}^{\infty} f_1\left(\beta \mid u_j\right) d\beta$$

โดยที่

$$f_1\left(\beta \mid u_j\right) = \frac{\prod_{i=1}^{n}\left[\left\{h_0\left(t_i\right)e^{x_i^T\beta}\right\}^{\zeta_i}\exp\left\{-H_0\left(t_i\right)e^{x_i^T\beta}\right\}\cdot\frac{1}{4}\omega\left(\sqrt{\sum_{i=1}^{n} w_i x_{ij}^2}\right)\exp\left\{-\frac{1}{2}\left(\sqrt{\sum_{i=1}^{n} w_i x_{ij}^2}\right)|\beta|\right\}\right]}{\left(\frac{\exp\left\{-\frac{1}{2}\left(u_j-\beta_j\right)\right\}}{\sqrt{2\pi}\left(\sqrt{\sum_{i=1}^{n} w_i x_{ij}^2}\right)}\right)}$$

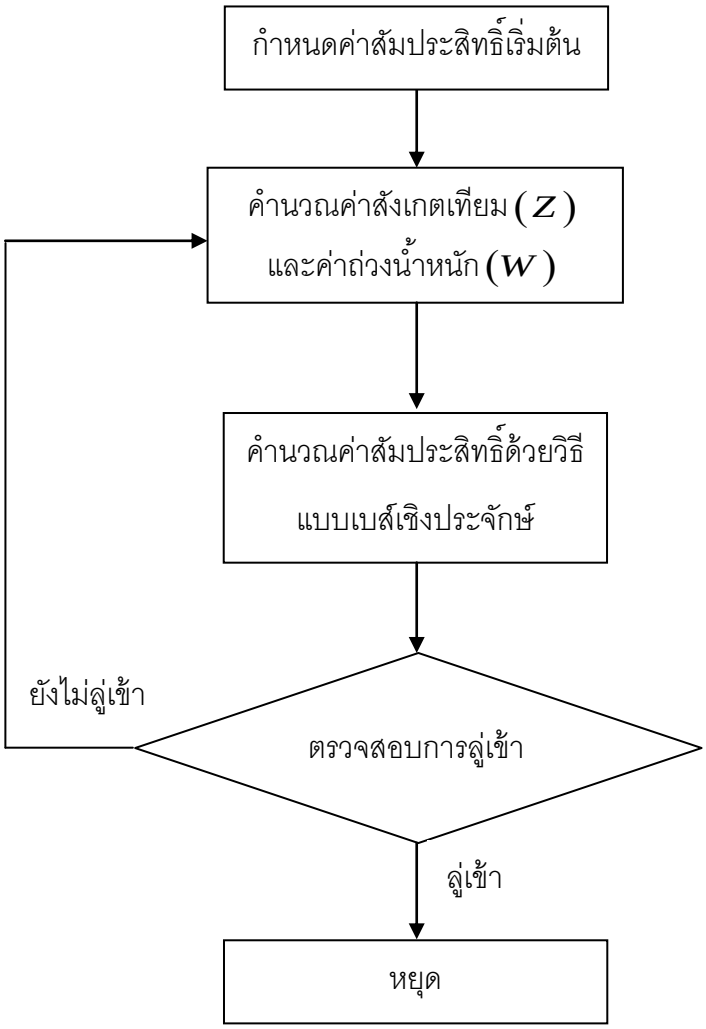
4. อัพเดทค่า $\hat{\omega}^{(k+1)}$ จาก

$$\hat{\omega}^{(k+1)} = \text{mode}\left(\omega \mid X, \hat{Z}^{(k)}, \hat{W}^{(k)}, \hat{\beta}^{(k+1)}\right) = \frac{\left\|\hat{\beta}^{(k+1)}\right\|^{\zeta_i=1}}{p}$$

5. ทำซ้ำขั้นที่1-4จนกว่า $\{\beta, \omega\}$ จะมีค่าความผันแปรเฉลี่ยมีลักษณะคงที่ตลอดช่วงเวลา

เนื่องจากการทำIRLS มักเกิดปัญหาที่ข้อมูลจะมีการเปลี่ยนแปลงแทบทุกรอบของการทำซ้ำไม่มากก็น้อย ซึ่งการตรวจสอบโดยการรอให้ข้อมูลลู่เข้าโดยการวัดระยะห่างของค่าประมาณรอบที่ k และ k+1 จึงไม่ค่อยเหมาะกับวิธีIRLSเท่าไรนัก Zaiwen Wen et al.,(2012)ได้เสนอวิธีการที่เรียกว่า active set algorithm โดยวิธีการดังกล่าวเป็นวิธีที่ใช้สำหรับข้อมูลที่มีค่าส่วนใหญ่เป็นศูนย์บนพื้นฐานของการหดตัว(shrinkage) หรือการเพิ่มประสิทธิภาพของสเปซและความต่อเนื่อง (SIAM J. Sci. Comput. 32 (2010), pp. 1832–1857) สำหรับแก้ปัญหา $l_1 - regularized$ เช่น ผลรวมถ่วงน้ำหนักของ $l_1 - norm$ และฟังก์ชันเรียบ(smooth function) $f(x)$ โดยในการศึกษาครั้งนี้จะเป็นการตรวจสอบการประมาณค่าของวิธีIRLS เมื่อเซตค่าประมาณของชุดที่ k และ k+1 มีลักษณะเดียวกันคือสำหรับตำแหน่งใดๆ ถ้าข้อมูลมีค่าไม่เป็นศูนย์และเป็นศูนย์เหมือนกันจะถือว่าข้อมูลมีการลู่เข้าแล้ว การประมาณค่าจะหยุดลง และถือเอาคำตอบของรอบสุดท้ายเป็นค่าประมาณที่ต้องการ

<u>ภาพที่ 3.2</u> แผนผังการเขียนโปรแกรมการคัดเลือกแบบเบส์เชิงประจักษ์



หมายเหตุ: ตรวจสอบการลู่เข้าหมายถึง active set convergenceหรือ จำนวนการทำซ้ำเท่ากับ 1000 รอบ

3. คำนวณอัตราความผิดพลาดในการตรวจจับเชิงบวก, อัตราความผิดพลาดในการตรวจจับเชิงลบ และพื้นที่ใต้กราฟ ROC Curve

3.1. อัตราความผิดพลาดในการตรวจจับเชิงบวก (False positive rate) สามารถคำนวณได้จากสูตร

$$\frac{\sum_{j=1}^{p} 1_{\{\hat{\beta}_j \neq 0 \text{ and } \beta_j = 0\}}}{\sum_{j=1}^{p} 1_{\{\hat{\beta}_j \neq 0\}}}$$

3.2. อัตราความผิดพลาดในการตรวจจับเชิงลบ (False negative rate) สามารถคำนวณได้จากสูตร

$$\frac{\sum_{j=1}^{p} 1_{\{\hat{\beta}_j = 0 \text{ and } \beta_j \neq 0\}}}{\sum_{j=1}^{p} 1_{\{\hat{\beta}_j = 0\}}}$$

เมื่อ p คือจำนวนตัวแปรอิสระ

เนื่องจากการทดลองนี้พิจารณาที่จำนวนตัวแปรอิสระที่ 3 ขนาดดังนั้นจำนวนตัวแปรอิสระ p ทั้งหมดคือ300, 500 และ 1000

3.3 การพิจารณาเส้นโค้ง ROC คือการสร้างกราฟความสัมพันธ์ระหว่าง true positive rate (Sensitivity) กับ false positive rate (1 – Specificity) เพื่อเทียบพื้นที่ใต้เส้นโค้งของแต่ละวิธี

โดยที่

$$\text{True positive rate (Sensitivity)} = \frac{\sum_{j=1}^{p} 1_{\{\hat{\beta}_j = 0 \text{ and } \beta_j = 0\}}}{\sum_{j=1}^{p} 1_{\{\beta_j = 0\}}} = 1 - \text{False negative rate}$$

และ

$$\text{False positive rate (1 – Specificity)} = 1 - \left(\frac{\sum_{j=1}^{p} 1_{\{\hat{\beta}_j \neq 0 \text{ and } \beta_j \neq 0\}}}{\sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}}} \right)$$

4. การวิเคราะห์ผลลัพธ์

เกณฑ์ที่ใช้ในการตัดสินว่าข้อมูลชุดใดเหมาะสมกับวิธีการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์แบบเบส์เชิงประจักษ์ที่ทำงานร่วมกับวิธีการทำซ้ำแบบมีเงื่อนไขฐานนิยม/มัธยฐานสำหรับตัวแบบ Cox's propostional hazard ที่ข้อมูลมีมิติสูง คือ การตรวจสอบอัตราความผิดพลาดในการตรวจจับเชิงบวก (False positive rate), อัตราความผิดพลาดในการตรวจจับเชิงลบ (False negative rate) และ การพิจารณาพื้นที่ใต้เส้นโค้ง ROC โดยข้อมูลที่ให้อัตราความผิดพลาดในการตรวจจับเชิงบวกและลบต่ำ และพื้นที่ใต้เส้นโค้งมากคือว่าเป็นวิธีที่มีประสิทธิภาพ

บทที่ 4

ผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อทดสอบการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์โดยวิธีแบบเบส์เชิงประจักษ์ พร้อมทั้งเปรียบเทียบประสิทธิภาพวิธีดังกล่าวโดยการแยกพิจารณาจากปัจจัยของอัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระและร้อยละของข้อมูลเซ็นเซอร์ เกณฑ์ที่ใช้ในการพิจารณาคืออัตราความผิดพลาดในการตรวจจับเชิงบวกและอัตราความผิดพลาดในการตรวจจับเชิงลบ โดยที่กรณีศึกษาใดที่ให้ค่าเฉลี่ยของอัตราความผิดพลาดต่ำถือว่าเป็นกรณีที่เหมาะกับการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ด้วยวิธีดังกล่าว

การนำเสนอผลการวิจัยจะแสดงในรูปของตารางแสดงค่าและกราฟเปรียบเทียบโดยมีตัวย่อหรือสัญลักษณ์ที่ใช้โดยตัวย่อหรือสัญลักษณ์ต่างๆ แทนความหมายดังนี้

c	แทน ร้อยละของข้อมูลเซ็นเซอร์
n	แทน ขนาดตัวอย่าง
p	แทน จำนวนตัวแปรอิสระ
n:p	แทน ขนาดตัวอย่างต่อตัวแปรอิสระ
Lasso	แทน การหาค่าสัมประสิทธิ์ด้วยวิธี Lasso
FNR	แทน อัตราความผิดพลาดในการตรวจจับเชิงลบ
FPR	แทน อัตราความผิดพลาดในการตรวจจับเชิงบวก
EBVS _true	แทน การคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ด้วยวิธีการแบบเบส์เชิงประจักษ์ ที่ค่าสัมประสิทธิ์เริ่มต้นเป็นค่าสัมประสิทธิ์ที่แท้จริง
EBVS _lasso	แทน การคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ด้วยวิธีการแบบเบส์เชิงประจักษ์ ที่ค่าสัมประสิทธิ์เริ่มต้นเป็นค่าที่หาจากวิธี Lasso

ในการนำเสนอผลการเปรียบเทียบจะนำเสนอโดยแบ่งออกเป็น 2 ส่วน โดยที่ส่วนแรกจะเปรียบเทียบอัตราความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยระหว่างการคัดเลือกแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso และค่าประมาณสัมประสิทธิ์ ที่ได้จากวิธี Lasso ส่วนที่ 2 จะเปรียบเทียบอัตราความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยระหว่างการคัดเลือกแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso และค่าประมาณสัมประสิทธิ์ ที่ได้จากวิธี Lasso

<u>ส่วนที่ 1</u> ผลการเปรียบเทียบอัตราความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยของข้อมูลจำลองขนาด100 ค่าระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso กับวิธี Lasso เมื่อพิจารณาในกรณีที่

1.1 เมื่อให้อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคงที่เมื่อขนาดตัวอย่างต่อตัวแปรอิสระคือ 100:300, 100:500 และ100:1,000

1.2 เมื่อให้ร้อยละของข้อมูลเซ็นเซอร์คงที่ที่ระดับ10%, 50%และ70%

<u>ส่วนที่ 2</u> ผลการเปรียบเทียบอัตราความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยของข้อมูลจำลองขนาด100 ค่าระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จาก Lasso กับวิธี Lasso เมื่อพิจารณาในกรณีที่

2.1 เมื่อให้อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคงที่เมื่อขนาดตัวอย่างต่อตัวแปรอิสระคือ 100:300, 100:500 และ100:1,000

2.2 เมื่อให้ร้อยละของข้อมูลเซ็นเซอร์คงที่ที่ระดับ10%, 50%และ70%

<u>ส่วนที่ 3</u> แสดงผลเปรียบเทียบ ROC Curveและพื้นที่ใต้กราฟระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จาก Lasso กับวิธี Lasso

3.1 เมื่อให้อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคงที่

3.2 เมื่อให้ร้อยละของข้อมูลเซ็นเซอร์คงที่

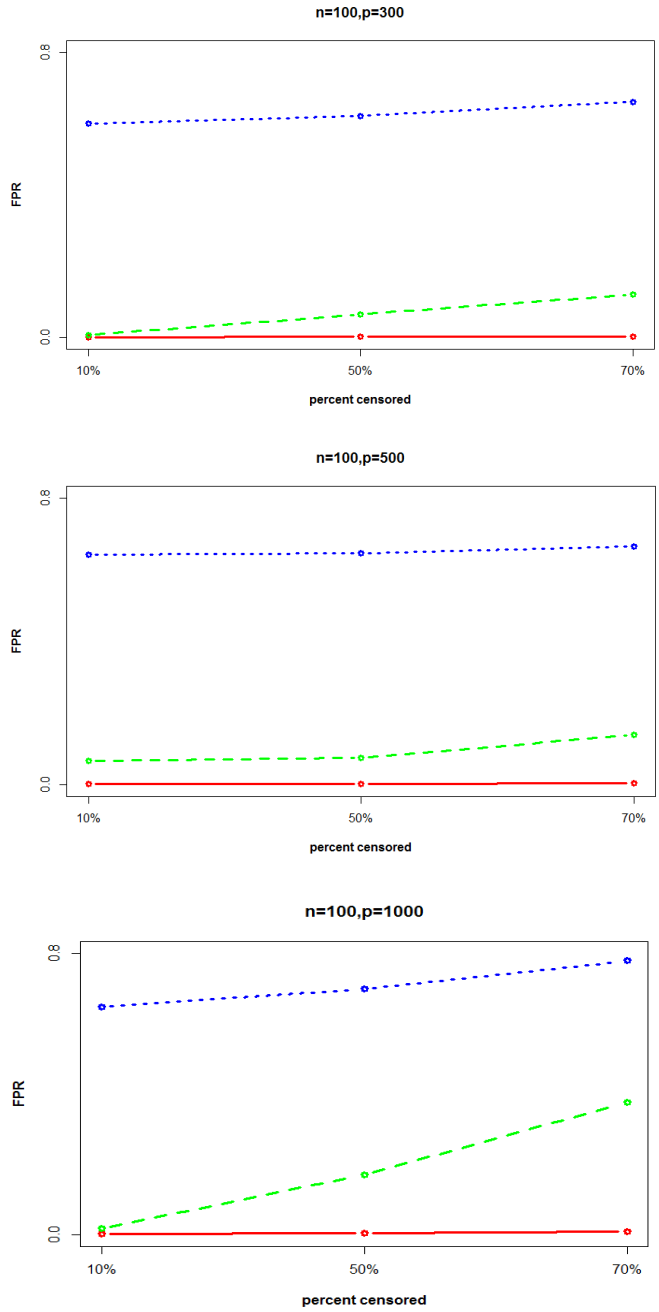
4.1 ผลการเปรียบเทียบอัตราความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยของข้อมูลจำลองขนาด100 ค่าระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso กับวิธี Lasso

ในส่วนแรกผู้วิจัยต้องการศึกษาว่าอัตราความผิดพลาดในการตรวจจับเชิงบวกของวิธีแบบเบส์เชิงประจักษ์เพื่อตรวจสอบว่าปัจจัยใดที่ส่งผลต่อประสิทธิภาพการทำงาน และเปรียบเทียบระหว่างวิธีแบบเบส์เชิงประจักษ์และวิธี Lasso ว่าให้ผลออกมาในลักษณะใด มีความแตกต่างกันหรือไม่และวิธีใดดีกว่า โดยแยกพิจารณาเป็น 2 กรณี กรณีแรกคือกรณีที่อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคงที่เมื่อขนาดตัวอย่างต่อตัวแปรอิสระคือ 100:300, 100:500 และ 100:1,000 และกรณีที่ 2 คือกรณีที่ให้ร้อยละของข้อมูลเซ็นเซอร์คงที่ที่ระดับ10%, 50%และ70% โดยแสดงในรูปของตารางที่ 4.1.1 และ4.1.2 โดยแต่ละตารางมีรายละเอียดคือ

ชนิดของอัตราความผิดพลาด	ตารางที่	ประเภทการพิจารณา	ประเภทข้อมูลที่ใช้เปรียบเทียบ
FPR	4.1.1	ขนาดตัวอย่างต่อตัวแปรอิสระคงที่	EBVS_true
			EBVS_lasso
			Lasso
	4.1.2	ร้อยละของข้อมูลเซ็นเซอร์คงที่	EBVS_true
			EBVS_lasso
			Lasso

ตารางที่ 4.1.1 แสดงอัตราความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยของข้อมูลจำลองขนาด100 ค่าระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso และวิธี Lasso เมื่อให้ อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคงที่ โดยที่ค่าเบี่ยงเบนมาตรฐานแสดงไว้ในวงเล็บ

n:p	c	EBVS_true	EBVS_lasso	Lasso
100:300	10%	0.0013(0.0054)	0.0061(0.0207)	0.6002(0.1084)
	50%	0.0019(0.0094)	0.0646(0.1630)	0.6216(0.0763)
	70%	0.0033(0.0122)	0.1200(0.2080)	0.6603(0.0895)
100:500	10%	0.0014(0.0082)	0.0661(0.1389)	0.6406(0.1062)
	50%	0.0029(0.0114)	0.0746(0.1334)	0.6451(0.0881)
	70%	0.0038(0.0130)	0.1387(0.2105)	0.6651(0.1014)
100:1,000	10%	0.0034(0.0123)	0.0186 (0.0709)	0.6480(0.0676)
	50%	0.0043(0.0137)	0.1704(0.2572)	0.6979(0.2070)
	70%	0.0091(0.0227)	0.3969(0.4410)	0.7780(0.1584)



ภาพที่ 4.1.1 แสดงการเปรียบเทียบอัตราความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso และวิธี Lassoเมื่อให้อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคงที่

จากตาราง4.1.1 ซึ่งแสดงผลของFPRโดยเฉลี่ยของข้อมูลจำลองขนาด100 ค่า ระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso และค่าประมาณสัมประสิทธิ์ที่ได้จากวิธี Lasso เมื่อ ให้อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคงที่พบว่า

i) ที่จำนวนตัวแปรอิสระเท่ากับ 300 การคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ทั้งในกรณีที่ค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงและค่าประมาณที่ได้จากวิธีLasso และวิธีLassoต่างให้ผลการทดลองสอดคล้องกันว่า ที่ร้อยละของข้อมูลเซ็นเซอร์ต่ำจะให้FPRต่ำกว่ากรณีที่ร้อยละของข้อมูลเซ็นเซอร์สูงโดยที่อัตราการเพิ่มของวิธีแบบเบส์เชิงประจักษ์ค่อนข้างคงที่ เมื่อเปรียบเทียบกับวิธีLassoพบว่าวิธีแบบเบส์เชิงประจักษ์ให้FPRต่ำกว่าวิธีLassoมากอย่างเห็นได้ชัด

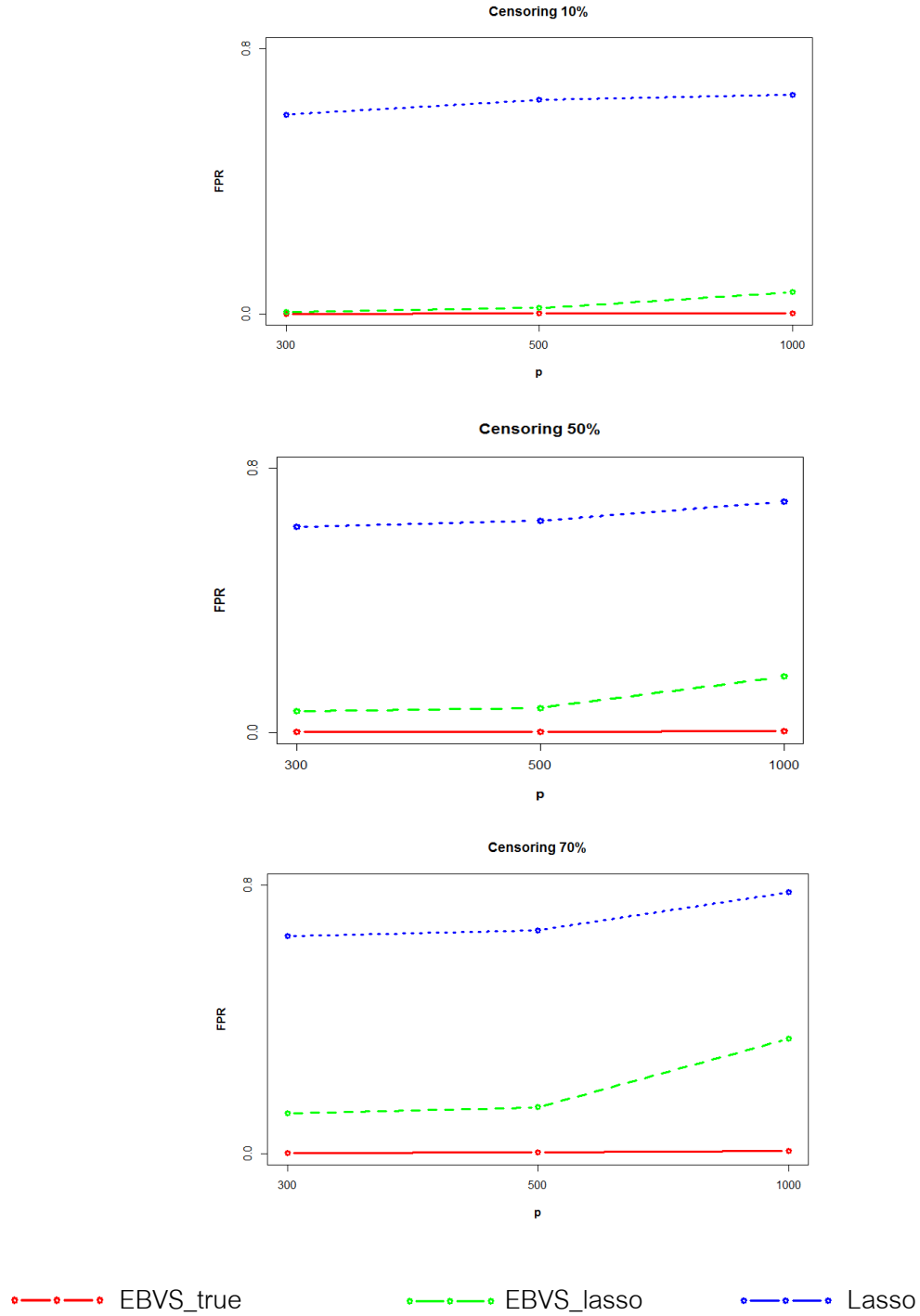
ii) ที่จำนวนตัวแปรอิสระเท่ากับ 500 แนวโน้มของFPRมีลักษณะเช่นเดียวกับกรณีที่จำนวนตัวแปรอิสระเท่ากับ 300 คือที่ร้อยละของข้อมูลเซ็นเซอร์ต่ำจะให้FPRต่ำกว่ากรณีที่ร้อยละของข้อมูลเซ็นเซอร์สูง แต่จากภาพที่ 4.1.1จะเห็นว่าFPRของวิธีแบบเบส์เชิงประจักษ์เมื่อค่าเริ่มต้นเป็นค่าประมาณจากวิธี Lasso จะมีความชันในช่วงระหว่าง10%กับ50%น้อยกว่าช่วงระว่าง50%กับ70% เมื่อเปรียบเทียบระหว่างวิธีแบบเบส์เชิงประจักษ์กับวิธีLassoพบว่าวิธีแบบเบส์เชิงประจักษ์ยังคงให้FPRต่ำกว่าวิธีLassoมากเช่นเดียวกรณีที่จำนวนตัวแปรอิสระเป็น300

iii) ที่จำนวนตัวแปรอิสระเท่ากับ1000 แนวโน้มของFPRยังเหมือนกับกรณีที่จำนวนตัวแปรอิสระเป็น 300และ 500 แต่ร้อยละของข้อมูลเซ็นเซอร์จะส่งผลกระทบต่อFPRของวิธีแบบเบส์เชิงประจักษ์มากยิ่งขึ้น

ซึ่งจะเห็นได้จากภาพ 4.1.1 คือที่ความชันในช่วงระหว่าง10%กับ50%มีค่าใกล้เคียงกับช่วงระหว่าง50%กับ70% และFPRที่ร้อยละของข้อมูลเซ็นเซอร์สูงก็ยังเพิ่มขึ้นอย่างรวดเร็ว อย่างไรก็ตามเมื่อเทียบกับวิธีLassoแล้วก็ยังถือว่าให้FPRต่ำกว่ามาก

ตารางที่ 4.1.2 แสดงอัตราความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยของข้อมูลจำลองขนาด100 ค่าระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso และวิธี Lasso เมื่อให้ร้อยละของข้อมูลเซ็นเซอร์คงที่ โดยที่ค่าเบี่ยงเบนมาตรฐานแสดงไว้ในวงเล็บ

c	n:p	EBVS_true	EBVS_lasso	Lasso
10%	100:300	0.0013(0.0054)	0.0061(0.0207)	0.6002(0.1084)
	100:500	0.0014(0.0082)	0.0186 (0.0709)	0.6451(0.0881)
	100:1,000	0.0034(0.0123)	0.0661(0.1389)	0.6603(0.0895)
50%	100:300	0.0019(0.0094)	0.0646(0.1630)	0.6216(0.0763)
	100:500	0.0029(0.0114)	0.0746(0.1334)	0.6406(0.1062)
	100:1,000	0.0043(0.0137)	0.1704(0.2572)	0.6979(0.2070)
70%	100:300	0.0033(0.0122)	0.1200(0.2080)	0.6480(0.0676)
	100:500	0.0038(0.0130)	0.1387(0.2105)	0.6651(0.1014)
	100:1,000	0.0091(0.0227)	0.3969(0.4410)	0.7780(0.1584)



ภาพที่ 4.1.2 แสดงการเปรียบเทียบอัตราความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso และวิธี Lasso เมื่อให้ร้อยละของข้อมูลเซ็นเซอร์คงที่

จากตาราง4.1.2 ซึ่งแสดงผลของFPRโดยเฉลี่ยของข้อมูลจำลองขนาด100 ค่า ระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso และค่าประมาณสัมประสิทธิ์ที่ได้จากวิธี Lasso เมื่อให้ร้อยละของข้อมูลเซ็นเซอร์คงที่

พบว่า

i)ที่ร้อยละของข้อมูลเซ็นเซอร์เท่ากับ 10% การคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ทั้งในกรณีที่ค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงและค่าประมาณที่ได้จากวิธีLasso และวิธีLassoต่างให้ผลการทดลองสอดคล้องกันว่า ที่อัตราส่วนระหว่างขนาดตัวอย่างตัวจำนวนแปรอิสระต่อสูงหรือเมื่อจำนวนตัวแปรอิสระน้อยจะให้FPRต่ำกว่ากรณีที่อัตราส่วนระหว่างขนาดตัวอย่างตัวจำนวนแปรอิสระต่อสูงหรือเมื่อจำนวนตัวแปรอิสระมาก จากภาพ 4.1.2 ที่อัตราการเพิ่มของFPRของวิธีแบบเบส์เชิงประจักษ์ช่วงระหว่างจำนวนตัวแปรอิสระเท่ากับ300กับ500จะมีความค่ากว่าช่วงระหว่างจำนวนตัวแปรอิสระเท่ากับ500กับ1000เมื่อเปรียบเทียบกับวิธีLasso พบว่าวิธีแบบเบส์เชิงประจักษ์ให้FPRต่ำกว่าวิธีLassoมากอย่างเห็นได้ชัด

ii) ที่ร้อยละของข้อมูลเซ็นเซอร์เท่ากับ 50% แนวโน้มของFPRมีลักษณะเช่นเดียวกับกรณีที่ร้อยละของข้อมูลเซ็นเซอร์เท่ากับ 10%คือกรณีที่จำนวนตัวแปรอิสระน้อย จะให้FPRต่ำกว่ากรณีที่จำนวนตัวแปรอิสระมาก และจากภาพที่ 4.1.2 ที่อัตราการเพิ่มของFPRของวิธีแบบเบส์เชิงประจักษ์ช่วงระหว่างจำนวนตัวแปรอิสระเท่ากับ300กับ500จะมีความค่ากว่าช่วงระหว่างจำนวนตัวแปรอิสระเท่ากับ500กับ1000 แต่จะแตกต่างจากกรณีที่ร้อยละของข้อมูลเซ็นเซอร์เท่ากับ10% ตรงที่อัตราการเพิ่มของความชันจะเพิ่มขึ้นเร็วกว่า เมื่อเปรียบเทียบกับวิธีLassoพบว่าวิธีแบบเบส์เชิงประจักษ์ให้FPRต่ำกว่าวิธีLassoมากอย่างเห็นได้ชัด

iii) ที่ร้อยละของข้อมูลเซ็นเซอร์เท่ากับ 70% แนวโน้มของFPRมีลักษณะเช่นเดียวกับกรณีที่ร้อยละของข้อมูลเซ็นเซอร์เท่ากับ 10% และ 50%คือกรณีที่จำนวนตัวแปรอิสระน้อย จะให้FPRต่ำกว่ากรณีที่จำนวนตัวแปรอิสระมาก และจากภาพที่ 4.1.2 ที่อัตราการเพิ่มของFPRของวิธีแบบเบส์เชิงประจักษ์ช่วงระหว่างจำนวนตัวแปรอิสระเท่ากับ300กับ500จะมีความแตกต่างต่างน้อยกว่าช่วงระหว่างจำนวนตัวแปรอิสระเท่ากับ500กับ1000 แต่จะแตกต่างจากกรณีที่ร้อยละของข้อมูลเซ็นเซอร์เท่ากับ10%และ 50%ตรงที่อัตราการเพิ่มของความชันจะเพิ่มขึ้นเร็วกว่า เมื่อเปรียบเทียบกับวิธีLassoพบว่าวิธีแบบเบส์เชิงประจักษ์ให้FPRต่ำกว่าวิธีLassoมากอย่างเห็นได้ชัด

<u>สรุปผลส่วนที่ 4.1</u> ผลการเปรียบเทียบอัตราความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso และค่าประมาณสัมประสิทธิ์ ที่ได้จากวิธี Lasso

พิจารณาตารางที่ 4.1.1 ที่เปรียบเทียบอัตราความผิดพลาดเชิงบวกโดยเฉลี่ยที่แต่ละระดับของจำนวนตัวแปรอิสระ พบว่าอัตราความผิดพลาดมีแนวโน้มเพิ่มขึ้นเมื่อร้อยละของข้อมูลเซ็นเซอร์เพิ่มขึ้น ร้อยละของข้อมูลเซ็นเซอร์ที่10%กับ50%จะมีความชันน้อยกว่าช่วงระหว่างร้อยละของข้อมูลเซ็นเซอร์ที่50%กับ 70% ซึ่งแสดงว่ายิ่งข้อมูลเซ็นเซอร์เพิ่มจำนวนขึ้น จะส่งผลต่ออัตราความผิดพลาดในการตรวจจับเชิงบวกแบบก้าวกระโดด อีกทั้งในกรณีที่จำนวนตัวแปรอิสระเพิ่มสูงขึ้น อัตราความผิดพลาดก็จะทวีคูณกว่าที่ระดับข้อมูลเซ็นเซอร์เท่ากันแต่จำนวนตัวแปรอิสระต่ำกว่าอย่างเห็นได้ชัด เมื่อเปรียบเทียบวิธีแบบเบส์เชิงประจักษ์กับวิธี Lasso พบว่าไม่ว่าจะกำหนดค่าสัมประสิทธิ์เร่มต้นเป็นค่าจริงหรือค่าประมาณที่ได้จากวิธี Lasso วิธีแบบเบส์เชิงประจักษ์ก็ให้อัตราความผิดพลาดในการตรวจจับเชิงบวกต่ำกว่าวิธี Lasso ซึ่งสามารถกล่าวได้ว่าวิธีแบบเบส์เชิงประจักษ์มีประสิทธิภาพในเชิงการตรวจจับเชิงบวก แม้ค่าเริ่มต้นจะมีความผิดพลาดสูง แต่เมื่อเข้าสู่ขั้นตอนการประมาณแบบเบส์เชิงประจักษ์แล้ว ค่าที่ได้มีความถูกต้องมากยิ่งขึ้น

เมื่อพิจารณาผลการเปรียบเทียบในตารางที่4.1.2 ที่เปรียบเทียบอัตราความผิดพลาดเชิงบวกโดยเฉลี่ยที่แต่ละระดับของข้อมูลเซ็นเซอร์จะเห็นว่าที่จำนวนตัวแปรอิสระเท่ากันพบว่าทั้งวิธีการคัดเลือกแบบเบส์เชิงประจักษ์และวิธี Lasso ต่างมีแนวโน้มเหมือนกัน คือ เมื่อจำนวนตัวแปรอิสระเพิ่มมากขึ้น อัตราความผิดพลาดในการตรวจจับเชิงบวกก็จะสูงขึ้นด้วย เมื่อพิจารณาภาพที่ 4.1.2 ที่ระดับข้อมูลเซ็นเซอร์ 10% การกำหนดค่าสัมประสิทธิ์เริ่มต้นจะไม่ส่งผลต่ออัตราความผิดพลาดมากนัก แต่เมื่อ ร้อยละของข้อมูลเซ็นเซอร์เพิ่มระดับสูงขึ้นการกำหนดค่าสัมประสิทธิ์เริ่มต้นที่แตกต่างกัน ส่งผลให้อัตราความผิดพลาดแตกต่างกันมากขึ้น และจะยิ่งแตกต่างกันสูงเมื่อมีปัจจัยของจำนวนตัวแปรอิสระเข้ามาเกี่ยวข้องด้วย

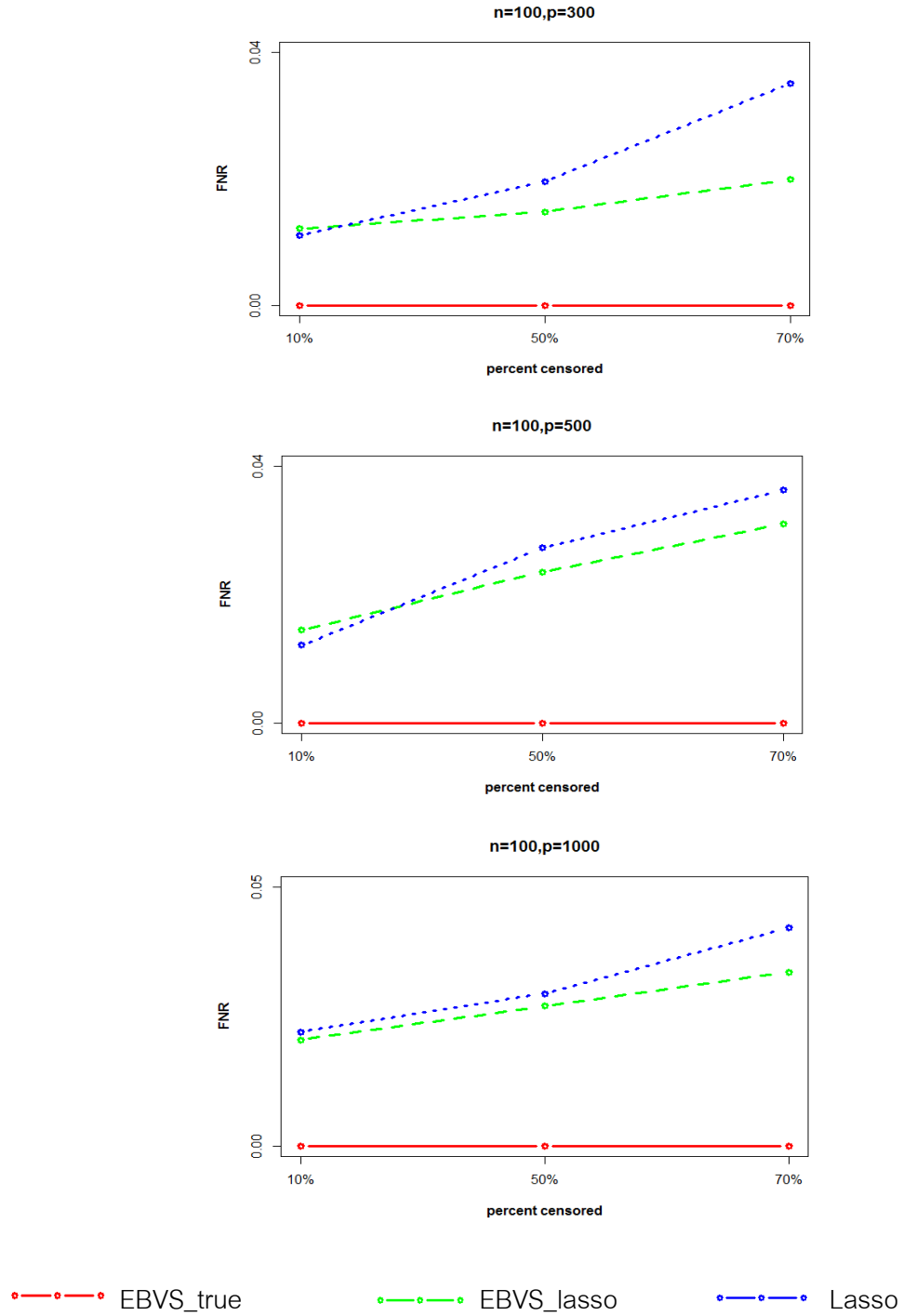
4.2 ผลการเปรียบเทียบอัตราความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยของข้อมูลจำลองขนาด100 ค่าระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso กับวิธี Lasso เมื่อพิจารณาในกรณีที่

ในส่วนที่ 2 ผู้วิจัยต้องการศึกษาว่าอัตราความผิดพลาดในการตรวจจับเชิงลบระหว่างวิธีแบบเบส์และวิธี Lasso ให้ผลออกมาในลักษณะใด มีความแตกต่างกันหรือไม่และวิธีใดดีกว่า โดยแยกพิจารณาเป็น 2 กรณี กรณีแรกคือกรณีที่อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคงที่ เมื่อขนาดตัวอย่างต่อตัวแปรอิสระคือ 100:300, 100:500 และ100:1,000 และกรณีที่ 2 คือกรณีที่ให้ร้อยละของข้อมูลเซ็นเซอร์คงที่ที่ระดับ10%, 50%และ70% โดยแสดงในรูปของตารางที่ 4.2.1 และ 4.2.2 โดยแต่ละตารางมีรายละเอียดคือ

ชนิดของอัตราความผิดพลาด	ตารางที่	ประเภทการพิจารณา	ประเภทข้อมูลที่ใช้เปรียบเทียบ
FNR	4.2.1	ขนาดตัวอย่างต่อตัวแปรอิสระคงที่	EBVS_true
			EBVS_lasso
			Lasso
	4.2.2	ร้อยละของข้อมูลเซ็นเซอร์คงที่	EBVS_true
			EBVS_lasso
			Lasso

ตารางที่ 4.2.1 แสดงอัตราความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยของข้อมูลจำลองขนาด 100 ค่าระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso และวิธี Lasso เมื่อให้อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคงที่ โดยที่ค่าเบี่ยงเบนมาตรฐานแสดงไว้ในวงเล็บ

n:p	c	EBVS_true	EBVS_lasso	Lasso
100:300	10%	0(0)	0.0121(0.0035)	0.0110(0.0159)
	50%	0(0)	0.0148(0.0204)	0.0196(0.0007)
	70%	0(0)	0.0199(0.0043)	0.0351(0.0140)
100:500	10%	0(0)	0.0130(0.0064)	0.0121(0.0341)
	50%	0(0)	0.0235(0.0074)	0.0273(0.0303)
	70%	0(0)	0.0310(0.0039)	0.0363(0.0123)
100:1,000	10%	0(0)	0.0205(0.0157)	0.0220(0.0313)
	50%	0(0)	0.0270(0.0133)	0.0294(0.0471)
	70%	0(0)	0.0334(0.0139)	0.0422(0.0105)



ภาพที่ 4.2.1 แสดงการเปรียบเทียบอัตราความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยระหว่างการคัดเลือกตัวแปร ด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริง กับค่าประมาณที่ได้จากวิธี Lasso และวิธี Lasso เมื่อให้อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคงที่

จากตาราง 4.2.1 แสดงผลของFNRโดยเฉลี่ยของข้อมูลจำลองขนาด100 ค่าระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso และค่าประมาณสัมประสิทธิ์ที่ได้จากวิธี Lasso เมื่อให้อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคง พบว่า

i) ที่จำนวนตัวแปรอิสระเท่ากับ 300 วิธีแบบเบส์เชิงประจักษ์เมื่อให้ค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงFNRจะเท่ากับศูนย์ทุกระดับของข้อมูลเซ็นเซอร์ ส่วนกรณีที่ค่าสัมประสิทธิ์เริ่มต้นเป็นค่าประมาณจากวิธีLasso กับวิธีLasso ต่างให้แนวโน้มของFNRเหมือนกันคือ ที่ระดับของข้อมูลเซ็นเซอร์ต่ำจะให้FNRต่ำ จากภาพ4.2.1 อัตราการเพิ่มของFNRช่วงระหว่างข้อมูลเซ็นเซอร์ 10% กับ50%จะต่ำกว่าช่วงระหว่าง 50%กับ70% โดยที่ระดับข้อมูลเซ็นเซอร์ 10%วิธีEBVS_lassoจะมีค่าFNRสูงกว่าวิธี Lassoเล็กน้อย แต่เนื่องจากอัตราการเพิ่มของFNRของวิธีLassoสูงกว่าอัตราการเพิ่มFNRของวิธีแบบเบส์เชิงประจักษ์ทำให้ที่ระดับข้อมูลเซ็นเซอร์50%และ70% FNRของวิธีEBVS_lassoจึงต่ำกว่าของวิธีLasso

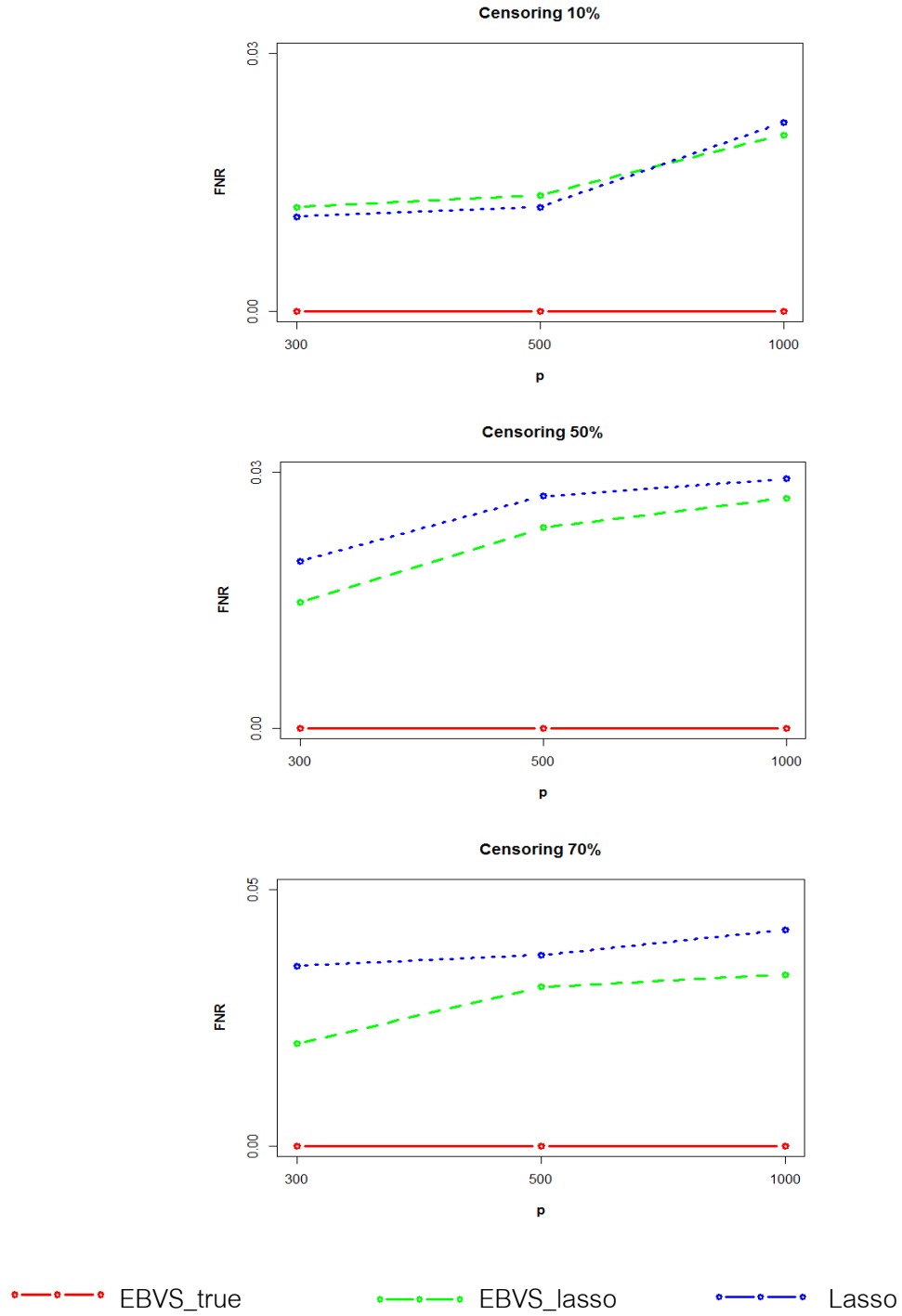
ii) ที่จำนวนตัวแปรอิสระเท่ากับ 500 มีลักษณะเช่นเดียวกับกรณีจำนวนตัวแปรอิสระเท่ากับ300 คือ วิธีแบบเบส์เชิงประจักษ์ ที่ค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงจะให้ค่าFNRเป็นศูนย์ทุกระดับของข้อมูลเซ็นเซอร์และ ส่วนกรณีที่ค่าสัมประสิทธิ์เริ่มต้นเป็นค่าประมาณจากวิธีLasso กับวิธีLasso ต่างให้แนวโน้มของFNRเหมือนกันคือ ที่ระดับของข้อมูลเซ็นเซอร์ต่ำจะให้FNRต่ำ จากภาพ4.2.1 อัตราการเพิ่มของFNRช่วงระหว่างข้อมูลเซ็นเซอร์ 10%กับ50%จะต่ำกว่าช่วงระหว่าง 50%กับ70% โดยที่ระดับข้อมูลเซ็นเซอร์ 10%วิธีEBVS_lassoจะมีค่าFNRสูงกว่าวิธีLassoเล็กน้อย แต่เนื่องจากอัตราการเพิ่มของFNRของวิธีLassoสูงกว่าอัตราการเพิ่มFNRของวิธีEBVS_lassoทำให้ที่ระดับข้อมูลเซ็นเซอร์50%และ70% FNRของวิธีEBVS_lassoจึงต่ำกว่าของวิธีLasso

iii) ที่จำนวนตัวแปรอิสระเท่ากับ1000 ก็มีลักษณะการเพิ่มของFNRเช่นเดียวกับกรณีที่จำนวนตัวแปรอิสระเท่ากับ 300และ500

จากภาพ 4.2.1 จะเห็นว่าจากทั้ง 3 กรณี วิธีEBVS_lassoและวิธี Lasso มีFNRใกล้เคียงกันมาก โดยวิธีแบบเบส์เชิงประจักษ์จะมีค่าFNRต่ำกว่าวิธีLassoเพียงเล็กน้อยเท่านั้น

ตารางที่ 4.2.2 แสดงอัตราความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยของข้อมูลจำลองขนาด 100 ค่าระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso และวิธี Lasso เมื่อให้ร้อยละของข้อมูลเซ็นเซอร์คงที่ โดยที่ค่าบี่ยงเบนมาตรฐานแสดงไว้ในวงเล็บ

c	n:p	EBVS_true	EBVS_lasso	Lasso
10%	100:300	0(0)	0.0121(0.0035)	0.0110(0.0159)
	100:500	0(0)	0.0135(0.0064)	0.0121(0.0341)
	100:1000	0(0)	0.0205(0.0157)	0.0220(0.0313)
50%	100:300	0(0)	0.0148 (0.0204)	0.0196(0.0007)
	100:500	0(0)	0.0235(0.0074)	0.0272(0.0471)
	100:1000	0(0)	0.0270(0.0133)	0.0293(0.0303)
70%	100:300	0(0)	0.0199(0.0043)	0.0351(0.0140)
	100:500	0(0)	0.0310(0.0039)	0.0363 (0.0123)
	100:1000	0(0)	0.0334(0.0139)	0.0422(0.0105)



ภาพที่ 4.2.2 แสดงการเปรียบเทียบอัตราความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริง กับค่าประมาณที่ได้จากวิธี Lasso และวิธี Lasso เมื่อให้ร้อยละของข้อมูลเซ็นเซอร์คงที่

จากตาราง4.2.1 แสดงผลของFNRโดยเฉลี่ยของข้อมูลจำลองขนาด100 ค่าระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso และค่าประมาณสัมประสิทธิ์ที่ได้จากวิธี Lasso เมื่อให้อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคง พบว่า

i) ที่ข้อมูลเซ็นเซอร์เท่ากับ10% วิธีEBVS_trueจะให้FNRเท่ากับศูนย์ ไม่ว่าจะมีตัวแปรอิสระเท่ากับ300, 500หรือ1000 เมื่อเปรียบเทียบวิธีEBVS_lassoกับวิธีLassoพบว่าที่จำนวนตัวแปรอิสระเท่ากับ300และ500ค่าFNRของวิธีEBVS_lassoมีค่าสูงกว่าวิธีLasso แต่ที่จำนวนตัวแปรอิสระเท่ากับ1000 FNRของวิธีEBVS_lasso เนื่องจากอัตราการเพิ่มช่วงระหว่างจำนวนตัวแปรอิสระเท่ากับ500และ1000ของวิธีLassoเพิ่มขึ้นมากกว่าวิธีแบบเบส์เชิงประจักษ์ และจากภาพ 4.2.1 ทั้งวิธีEBVS_lassoและวิธีLassoต่างมีแนวโน้มแบบเดียวกันคือที่จำนวนตัวแปรอิสระมาก FNRก็จะมากขึ้นด้วย

ii) ที่ข้อมูลเซ็นเซอร์เท่ากับ50% มีFNRของวิธีEBVS_true มีแนวโน้มเหมือนกรณีที่ข้อมูลเซ็นเซอร์เท่ากับ10%คือ จะให้FNRเท่ากับศูนย์ ไม่ว่าจะมีตัวแปรอิสระเท่ากับ300, 500หรือ1000 เมื่อเปรียบเทียบวิธีEBVS_lasso กับวิธีLassoพบว่าที่จำนวนตัวแปรอิสระเท่ากับ300, 500และ1000วิธีEBVS_lassoให้FNRต่ำกว่าวิธีLasso และจากภาพ 4.2.1 ทั้งวิธีEBVS_lassoและวิธีLassoต่างมีแนวโน้มแบบเดียวกันคือที่จำนวนตัวแปรอิสระมากFNRก็จะมากขึ้นด้วย

iii) ที่ข้อมูลเซ็นเซอร์เท่ากับ70% มีFNRของวิธีEBVS_true มีแนวโน้มเหมือนกรณีที่ข้อมูลเซ็นเซอร์เท่ากับ10%และ50%คือ จะให้FNRเท่ากับศูนย์ ไม่ว่าจะมีตัวแปรอิสระเท่ากับ300, 500 หรือ1000 เมื่อเทียบวิธีEBVS_lasso กับวิธีLassoพบว่ามีแนวโน้มเหมือนกันกับกรณีที่ข้อมูลเซ็นเซอร์เท่ากับ 50% คือทุกระดับของจำนวนตัวแปรอิสระที่พิจารณาของวิธีEBVS_lasso ให้FNRต่ำกว่าวิธีLasso มีความแตกต่างกันตรงอัตราการเพิ่มสูงสุดของวิธีEBVS_lassoคือช่วงระหว่างจำนวนตัวแปรอิสระเท่ากับ300กับ500 ส่วนวิธีLasso มีอัตราการเพิ่มสูงสุดคือช่วงระหว่างจำนวนตัวแปรอิสระเท่ากับ500กับ1000 อย่างไรก็ตามทั้ง2 วิธีต่างมีแนวโน้มเพิ่มขึ้นเมื่อจำนวนตัวแปรอิสระเพิ่มสูงขึ้น

ซึ่งเมื่อพิจารณาภาพ 4.2.2 พบว่า ทั้งวิธีแบบเบส์เชิงประจักษ์ที่ค่าสัมประสิทธิ์เป็นค่าประมาณจากวิธีLassoกับวิธีLassoให้FNRที่ใกล้เคียงกันมาก

<u>สรุปส่วนที่ 2</u> ผลการเปรียบเทียบอัตราความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จากวิธี Lasso และค่าประมาณสัมประสิทธิ์ ที่ได้จากวิธี Lasso

จากตารางที่ 4.2.1และ4.2.2 จะเห็นว่า การกำหนดค่าสัมประสิทธิ์เริ่มต้นด้วยค่าจริง ไม่ว่าจำนวนตัวแปรอิสระจะมากแค่ไหนหรือข้อมูลจะถูกเซ็นเซอร์ที่ระดับเท่าไรก็ไม่ส่งผลกระทบต่อวิธีแบบเบส์เชิงประจักษ์ แสดงว่าการกำหนดค่าเริ่มต้นมความสำคัญกว่าปัจจัยอื่นสำหรับการตรวจสอบด้วยFNR เมื่อพิจารณาวิธีEBVS_lassoเทียบกับวิธี Lasso พบว่าอัตราความผิดพลาดเชิงลบของทั้งสองวิธีใกล้เคียงกันมาก และเมื่อพิจารณาจากภาพที่ 4.2.1จะเห็นว่าถึงแม้วิธีแบบเบส์จะมีอัตราความผิดพลาดในการตรวจจับเชิงลบสูงกว่าวิธีLassoในช่วงแรก แต่เนื่องจากอัตราการเพิ่มของวิธีLasso มีการเปลี่ยนแปลงมากกว่าวิธีEBVS_lassoดังนั้นที่ระดับของร้อยละของข้อมูลเซ็นเซอร์เพิ่มขึ้นจึงมีค่าFNRสูงกว่าวิธีLasso แม้อัตราการเพิ่มของFNRของวิธีEBVS_lassoก็เพิ่มขึ้นตามผลกระทบของร้อยละของข้อมูลเซ็นเซอร์เช่นกัน แต่ผลกระทบก็ไม่มากเท่า แสดงว่าระดับของข้อมูลเซ็นเซอร์ส่งผลกระทบต่อการทำงานของวิธี Lasso มากกว่าวิธีแบบเบส์เชิงประจักษ์ ซึ่งภาพรวมที่เกิดขึ้น จึงกล่าวได้ว่าวิธีแบบเบส์เชิงประจักษ์มีความเสถียรในการทำงานที่ข้อมูลบางส่วนถูกเซ็นเซอร์มากกว่าวิธี Lasso และสำหรับผลจากตารางที่ 4.2.2 จะเห็นว่าปัจจัยของจำนวนตัวแปรอิสระส่งผลกระทบต่อวิธีแบบเบส์เชิงประจักษ์และวิธี Lasso ในอัตราที่ใกล้เคียงกัน คือเมื่อจำนวนตัวแปรอิสระเพิ่มมากขึ้นFNRก็จะมากขึ้นตามไปด้วย ซึ่งจากผลการทดลองสามารถกล่าวได้ว่าสิ่งที่ส่งผลต่อFNRมากที่สุดก็คือการกำหนดค่าสัมประสิทธิ์เริ่มต้นที่เหมาะสม และจากภาพ 4.2.2 จะเห็นว่าที่อัตราการซ็นเซอร์ต่ำที่จำนวนตัวแปรอิสระระดับน้อยถึงระดับกลางวิธี EBVS_lassoให้ FNR มากกว่าวิธี Lasso และวิธี EBVS_lasso จะเริ่มมีประสิทธภาพการทำงานในแง่FNRที่ดีกว่าเมื่อจำนวนตัวแปรอิสระระดับมาก แต่เมื่อร้อยละของข้อมูลเซ็นเซอร์มากขึ้นประสิทธิภาพของวิธี EBVS_lasso จะดีกว่า วิธี Lasso ทุกจำนวนตัวแปรอิสระกรณีที่พิจารณาค่าการตรวจสอบ FNR

4.3 แสดงผลเปรียบเทียบ ROC Curve และพื้นที่ใต้กราฟระหว่างการคัดเลือกตัวแปรด้วยวิธีแบบเบส์เชิงประจักษ์ที่มีค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริงกับค่าประมาณที่ได้จาก Lasso กับวิธี Lasso

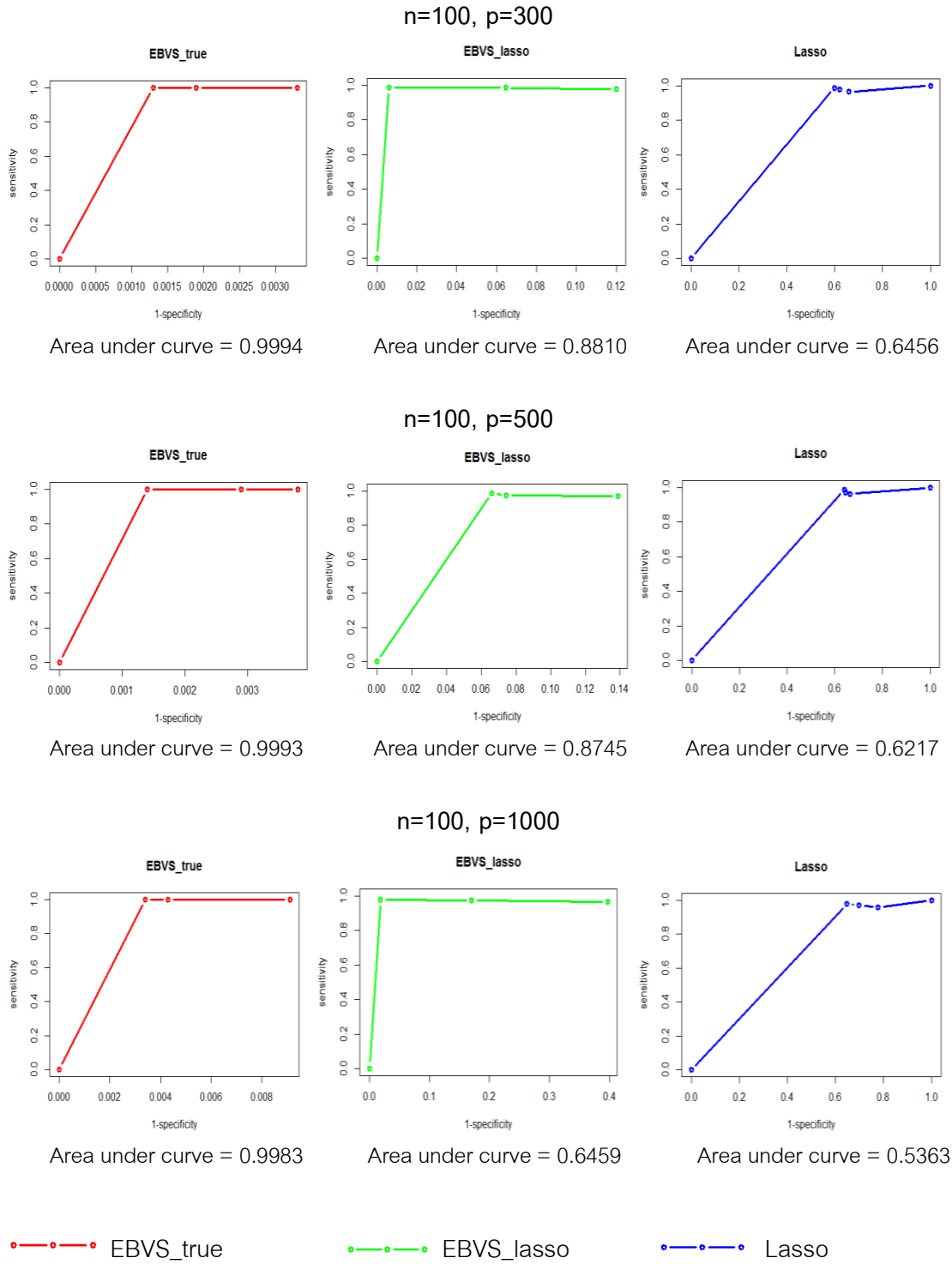
จากส่วนที่ 1และส่วนที่ 2 ที่แสดงผลปัจจัยร้อยละของข้อมูลเซ็นเซอร์และอัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระ โดยการแยกพิจารณาของอัตราความผิดพลาดในการตรวจจับเชิงบวกและลบ จากการตรวจสอบของ 2 เกณฑ์นี้ เราสามารถแสดงผลร่วมกันโดยการพิจารณาReceiver Operator Characteristic (ROC) curve คือการสร้างกราฟความสัมพันธ์ระหว่าง true positive rate (Sensitivity) กับ false positive rate (1 – Specificity) เพื่อเลือกจุดตัด (cut - off point) ที่เหมาะสม นอกจากนี้การสร้าง ROC curve ยังช่วยในการเปรียบเทียบประสิทธิภาพของการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ได้ด้วย โดยเปรียบ เทียบพื้นที่ใต้เส้นโค้งของแต่ละวิธี พื้นที่ใต้โค้งที่มากกว่าแสดงถึงประสิทธิภาพที่สูงกว่า โดยที่แกน x ของกราฟคือ1-specificity และ แกน y คือ sensitivity ซึ่งหาจากความสัมพันธ์ True positive rate (Sensitivity) = 1-false negative rateและFalse positive rate= (1 – Specificity) โดยแสดงผลตามตารางดังนี้

ตารางที่ 4.3.1 เมื่อให้อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคงที่ เมื่อขนาดตัวอย่างต่อตัวแปรอิสระคือ 100:300, 100:500 และ100:1,000

ตารางที่ 4.3.2 เมื่อให้ร้อยละของข้อมูลเซ็นเซอร์คงที่ที่ระดับ10%, 50%และ70%

ตารางที่ 4.3.1 แสดงค่า sensitivity และ 1-specificity ที่ใช้สำหรับสร้างเส้นโค้งROC เมื่อให้อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคงที่ เมื่อขนาดตัวอย่างต่อตัวแปรอิสระคือ 100:300, 100:500 และ100:1,000

	n:p		10%	50%	70%
EBVS_true	100:300	sensitivity	1	1	1
		1-specificity	0.0013	0.0019	0.0033
	100:500	sensitivity	1	1	1
		1-specificity	0.0014	0.0029	0.0038
	100:1000	sensitivity	1	1	1
		1-specificity	0.0034	0.0043	0.0091
EBVS_lasso	100:300	sensitivity	0.9879	0.9852	0.9801
		1-specificity	0.0061	0.0646	0.1200
	100:500	sensitivity	0.9870	0.9765	0.9690
		1-specificity	0.0661	0.0746	0.1387
	100:1000	sensitivity	0.9795	0.9730	0.9666
		1-specificity	0.0186	0.1704	0.3969
Lasso	100:300	sensitivity	0.9890	0.9804	0.9649
		1-specificity	0.6002	0.6216	0.6603
	100:500	sensitivity	0.9879	0.9727	0.9637
		1-specificity	0.6406	0.6451	0.6651
	100:1000	sensitivity	0.9780	0.9706	0.9578
		1-specificity	0.6480	0.6979	0.7780

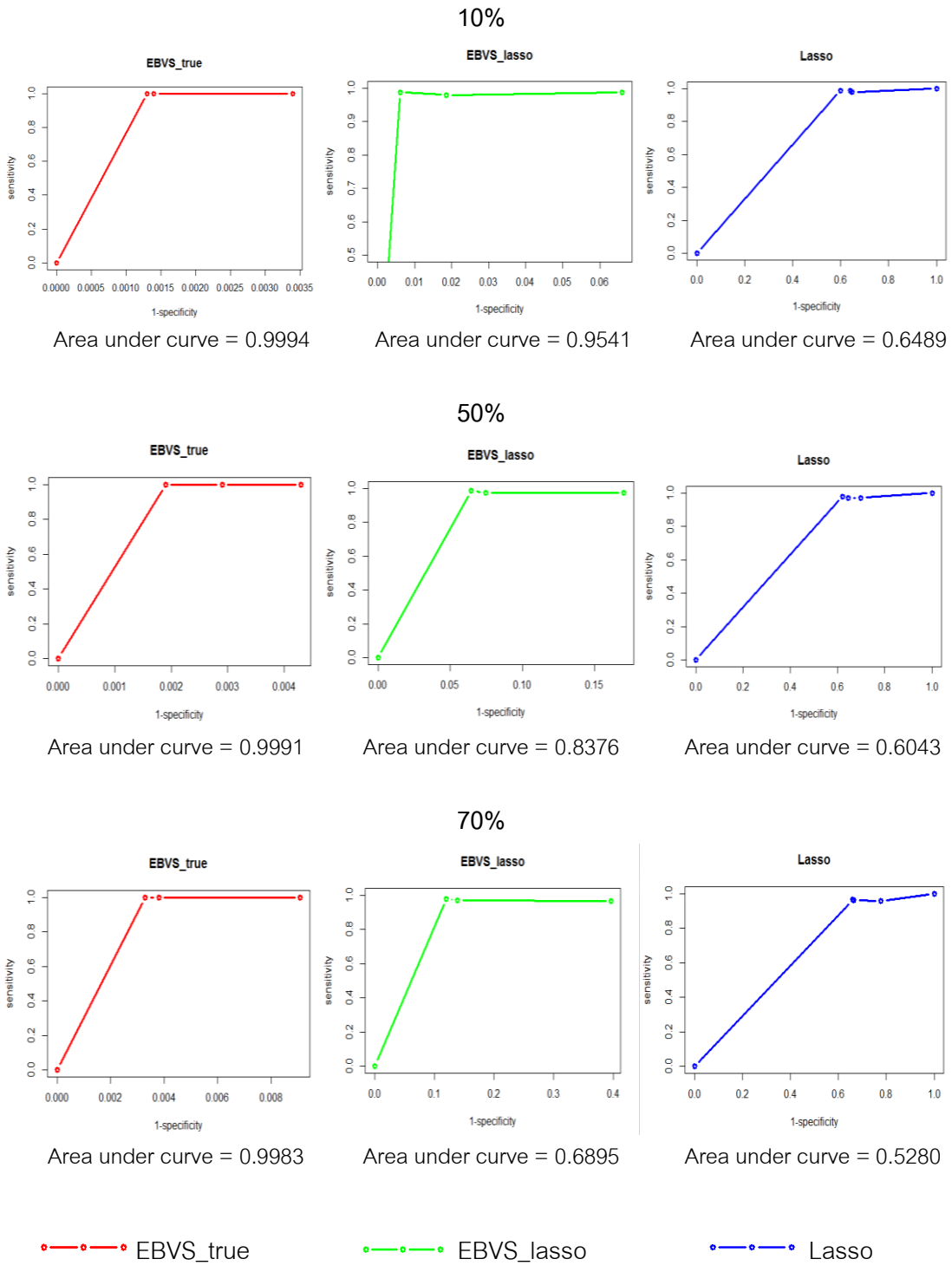


ภาพที่ 4.3.1 แสดงเส้นโค้ง ROC และพื้นที่ใต้เส้นโค้ง เมื่อให้อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคงที่ ที่ระดับ 100:300, 100:500 และ 100:1000

จากตาราง 4.3.1 แสดงค่า sensitivity และ 1-specificity ที่ใช้สำหรับสร้างเส้นโค้งROC เพื่อตรวจสอบพื้นที่ใต้เส้นโค้ง โดยจากผลของพื้นที่ใต้กราฟของวิธี EBVS_true, วิธี EBVS_lasso และวิธี Lasso เมื่อพิจารณาที่ร้อยละของข้อมูลเซ็นเซอร์คงที่ ที่ระดับ 10%, 50% และ 70% พบว่าจาก 3 วิธี วิธีที่ให้พื้นที่ใต้กราฟสูงสุดคือวิธี EBVS_true, วิธี EBVS_lasso และวิธี Lasso ตามลำดับ ซึ่งแสดงว่าวิธี EBVS_true เป็นวิธีที่ดีที่สุด รองลงมาคือวิธี EBVS_lasso และวิธีที่ให้ผลการทำงานไม่ค่อยดีคือวิธี Lasso เมื่อเทียบพื้นที่ใต้เส้นโค้งดดยพิจารณาปัจจัยร้อยละของข้อมูลเซ็นเซอร์ ทั้ง 3 กรณีคือ วิธี EBVS_true, วิธี EBVS_lasso และวิธี Lasso พบว่าต่างให้แนวโน้มของพื้นที่ใต้กราฟในทำนองเดียวกันคือ เมื่อร้อยละของข้อมูลเซ็นเซอร์สูงขึ้น พื้นที่ใต้เส้นโค้งจะลดลง

ตารางที่ 4.3.2 แสดงค่า sensitivity และ 1-specificity ที่ใช้สำหรับสร้างเส้นโค้ง ROC เมื่อให้ร้อยละของข้อมูลเซ็นเซอร์คงที่ที่ระดับ10%, 50%และ70%

	c		100:300	100:500	100:1000
EBVS_true	10%	Sensitivity	1	1	1
		1-specificity	0.0013	0.0014	0.0034
	50%	Sensitivity	1	1	1
		1-specificity	0.0019	0.0029	0.0043
	70%	Sensitivity	1	1	1
		1-specificity	0.0033	0.0038	0.0091
EBVS_lasso	10%	Sensitivity	0.9879	0.9870	0.9795
		1-specificity	0.0061	0.0661	0.0186
	50%	Sensitivity	0.9852	0.9765	0.9730
		1-specificity	0.0646	0.0746	0.1704
	70%	Sensitivity	0.9801	0.9690	0.9666
		1-specificity	0.1200	0.1387	0.3969
Lasso	10%	Sensitivity	0.9890	0.9879	0.9780
		1-specificity	0.6002	0.6406	0.6480
	50%	Sensitivity	0.9804	0.9727	0.9706
		1-specificity	0.6216	0.6451	0.6979
	70%	Sensitivity	0.9649	0.9637	0.9578
		1-specificity	0.6603	0.6651	0.7780



ภาพที่ 4.3.2 แสดงเส้นโค้ง ROC และพื้นที่ใต้เส้นโค้ง เมื่อให้ร้อยละของข้อมูลเซ็นเซอร์คงที่ ที่ระดับ 10%, 50% และ 70%

จากตาราง 4.3.2 แสดงค่า sensitivity และ 1-specificity ที่ใช้สำหรับสร้างเส้นโค้ง ROC เพื่อตรวจสอบพื้นที่ใต้เส้นโค้ง โดยจากผลของพื้นที่ใต้กราฟของวิธี EBVS_true, วิธี EBVS_lasso และวิธี Lasso เมื่อพิจารณาอัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคงที่ ที่ระดับ 100:300, 100:500 และ 100:1000 พบว่าจาก 3 วิธี วิธีที่ให้พื้นที่ใต้กราฟสูงสุดคือวิธี EBVS_true, วิธี EBVS_lasso และวิธี Lasso ตามลำดับ ซึ่งแสดงว่าวิธี EBVS_true เป็นวิธีที่ดีที่สุด รองลงมาคือวิธี EBVS_lasso และวิธีที่ให้ผลการทำงานไม่ค่อยดีคือวิธี Lasso เมื่อเทียบพื้นที่ใต้เส้นโค้งดดยพิจารณาปัจจัยร้อยละของข้อมูลเซ็นเซอร์ ทั้ง 3 กรณีคือ วิธี EBVS_true, วิธี EBVS_lasso และวิธี Lasso พบว่าต่างให้แนวโน้มของพื้นที่ใต้กราฟในทำนองเดียวกันคือ ที่ขนาดตัวอย่างเท่ากัน เมื่อจำนวนตัวแปรอิสระมากขึ้น พื้นที่ใต้เส้นโค้งจะลดลง

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

ในงานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาวิธีการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ด้วยวิธีแบบเบส์เชิงประจักษ์สำหรับตัวแบบCox' proportional hazardในกรณีที่ข้อมูลมีมิติสูง และทดสอบประสิทธิภาพของวิธีการดังกล่าวโดยสร้างกรณีศึกษาขึ้นมา แต่ละมีสถานการณ์ที่แตกต่างกัน โดยลักษณะข้อมูลที่สร้างขึ้นเป็นกรณีที่มักเกิดปัญหาในการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ที่พบได้ทั่วไป เพื่อจะตรวจสอบว่าสถานการณ์ของข้อมูลลักษณะใดที่ส่งผลต่อประสิทธิภาพของวิธีแบบเบส์เชิงประจักษ์ โดยในการศึกษาครั้งนี้จะทดสอบเฉพาะเมื่อกำหนดค่าสัมประสิทธิ์เริ่มต้นด้วยค่าจริงและค่าประมาณที่ได้จากวิธี Lasso เท่านั้น พร้อมทั้งเปรียบเทียบประสิทธิภาพกับวิธี Lasso เมื่อกำหนดให้ทดสอบในสถานการณ์ที่เหมือนกันและข้อมูลชุดเดียวกัน ผลการทดลองมีข้อสรุปดังนี้

5.1. สรุปผลการทดลอง

จากการศึกษาเพื่อทดสอบประสิทธิภาพการทำงานของวิธีแบบเบส์เชิงประจักษ์ โดยแบ่งการทดสอบเป็น 2 ส่วนคือ

5.1.1 ทดสอบหาปัจจัยที่ส่งผลต่อประสิทธิภาพในการทำงานของวิธีแบบเบส์เชิงประจักษ์

5.1.2 เปรียบเทียบประสิทธิภาพของวิธีแบบเบส์เชิงประจักษ์กับวิธี Lasso โดยผลการทดสอบทั้งสองกรณีมีรายละเอียดดังนี้

5.1.1 ทดสอบหาปัจจัยที่ส่งผลต่อประสิทธิภาพในการทำงานของวิธีแบบเบส์เชิงประจักษ์

ในการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์มักเกิดปัญหาเมื่อข้อมูลที่พิจารณามีสถานการณ์ที่ไม่เอื้อต่อการทำงานของวิธีที่ใช้ ในการศึกษาครั้งนี้ได้แยกสถานการณ์ที่คาดว่าจะส่งผลกระทบต่อการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ด้วยวิธีแบบเบส์เอาไว้ 3 กรณี คือ กรณีที่ข้อมูลถูกเซ็นเซอร์ที่ระดับต่างๆ, กรณีที่จำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่างโดยแบ่งเป็น 3 ระดับคือจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่างเล็กน้อย (100:300) มากกว่า

ปานกลาง (100:500) และมากกว่ามาก (100:1000) และผลกระทบสุดท้ายที่คาดว่าจะส่งผลต่อการทำงานคือ การกำหนดค่าสัมประสิทธิ์เริ่มต้น จากการทดลองสามารถสรุปผลได้ดังนี้

ความผิดพลาดในการตรวจจับเชิงบวก

สิ่งที่ส่งผลต่ออัตราความผิดพลาดในการตรวจจับเชิงบวกคือ ระดับข้อมูลเซ็นเซอร์ ดังที่ปรากฏในผลการทดลอง ที่ร้อยละของข้อมูลเซ็นเซอร์ต่ำ (10%) จะให้ค่าความผิดพลาดเชิงบวกต่ำกว่ากรณีร้อยละของข้อมูลเซ็นเซอร์กลาง (50%) และร้อยละของข้อมูลเซ็นเซอร์สูง (70%) ตามลำดับ กล่าวคือ เมื่อจำนวนข้อมูลที่ถูกเซ็นเซอร์เพิ่มมากขึ้น อัตราความผิดพลาดในการตรวจจับเชิงบวกก็จะสูงขึ้นด้วย ปัจจัยต่อมาคือ อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระ คือ เมื่อพิจารณาที่ขนาดตัวอย่างเท่ากัน ถ้าตัวแปรอิสระเพิ่มมากขึ้น อัตราความผิดพลาดในการตรวจจับเชิงบวกก็จะมีแนวโน้มสูงขึ้น และปัจจัยสุดท้ายคือ การกำหนดค่าสัมประสิทธิ์เริ่มต้น โดยเมื่อกำหนดค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริง จะทำให้วิธีแบบเบส์เชิงประจักษ์มีประสิทธิภาพมากกว่าการกำหนดสัมประสิทธิ์เริ่มต้นด้วยค่าประมาณที่ได้จากวิธี Lasso

ความผิดพลาดในการตรวจจับเชิงลบ

สิ่งที่ส่งผลต่ออัตราความผิดพลาดในการตรวจจับเชิงลบคือ ระดับข้อมูลเซ็นเซอร์ ดังที่ปรากฏในผลการทดลอง ที่ร้อยละของข้อมูลเซ็นเซอร์ต่ำ (10%) จะให้ค่าความผิดพลาดเชิงลบต่ำกว่ากรณีร้อยละของข้อมูลเซ็นเซอร์กลาง (50%) และร้อยละของข้อมูลเซ็นเซอร์สูง (70%) ตามลำดับ กล่าวคือ เมื่อจำนวนข้อมูลที่ถูกเซ็นเซอร์เพิ่มมากขึ้น อัตราความผิดพลาดในการตรวจจับเชิงลบก็จะสูงขึ้นด้วย ปัจจัยต่อมาคือ อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระ คือ ยิ่งอัตราส่วนระหว่างขนาดต่อตัวแปรอิสระต่ำ (100:1000) อัตราผิดพลาดในการตรวจจับเชิงลบจะสูงตาม และปัจจัยสุดท้ายคือ การกำหนดค่าสัมประสิทธิ์เริ่มต้น โดยเมื่อกำหนดค่าสัมประสิทธิ์เริ่มต้นเป็นค่าจริง จะทำให้วิธีแบบเบส์เชิงประจักษ์มีประสิทธิภาพมากกว่าการกำหนดสัมประสิทธิ์เริ่มต้นด้วยค่าประมาณที่ได้จากวิธี Lasso

จากผลการทดลองที่เกิดขึ้น ยังสามารถสรุปได้ว่าสิ่งที่ส่งผลต่ออัตราความผิดพลาดในการตรวจจับเชิงลบของการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ด้วยวิธีแบบเบส์เชิงประจักษ์มากที่สุดคือ การกำหนดค่าสัมประสิทธิ์เริ่มต้น เนื่องจากถ้ากำหนดให้ค่าสัมประสิทธิ์เริ่มต้นคือค่าจริงแล้วไม่ว่าอัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระหรือข้อมูลเซ็นเซอร์จะเป็นเท่าไร

อัตราความผิดพลาดในการตรวจจับเชิงลบจะเท่ากับศูนย์เสมอ ส่วนอีกสองปัจจัยแม้จะส่งผลต่อความผิดพลาดในการตรวจจับเชิงลบ แต่ผลก็ไม่เห็นชัดเท่าปัจจัยการกำหนดค่าเริ่มต้น

สรุป

จากค่าทดสอบประสิทธิภาพวิธีแบบเบส์เชิงประจักษ์ ทั้งสองคืออัตราความผิดพลาดในการตรวจจับเชิงบวกและอัตราในการตรวจจับเชิงลบ ต่างให้แนวโน้มของผลการทดสอบเหมือนกัน คือ ทั้ง 3 ปัจจัย ได้แก่ ร้อยละของข้อมูลเซ็นเซอร์, อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระ และ การกำหนดสัมประสิทธิ์เริ่มต้น ต่างส่งผลต่อประสิทธิภาพการทำงานของวิธีแบบเบส์เชิงประจักษ์ ซึ่งเมื่อเปรียบเทียบผลที่เกิดขึ้นจะเห็นว่าทั้ง 3 ปัจจัย ส่งผลกระทบต่ออัตราความผิดพลาดในการตรวจจับเชิงบวก สูงกว่าอัตราในการตรวจจับเชิงลบ โดยเฉพาะอย่างยิ่งร้อยละของข้อมูลเซ็นเซอร์ที่ยิ่งมีค่าสูงเท่าไร อัตราความผิดพลาดก็จะเพิ่มความแตกต่างจากเดิมอย่างเห็นได้ชัด

5.1.2 เปรียบเทียบประสิทธิภาพของวิธีแบบเบส์เชิงประจักษ์กับวิธี Lasso

ผลการเปรียบเทียบวิธีคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์สำหรับตัวแบบ Cox's proportional hazard ที่ข้อมูลมีมิติสูงระหว่างวิธีแบบเบส์เชิงประจักษ์และวิธี Lasso พบว่าเมื่อพิจารณาอัตราความผิดพลาดในการตรวจจับเชิงลบของทั้ง 2 วิธี มีผลใกล้เคียงกัน โดยภาพรวมถือว่าวิธีแบบเบส์เชิงประจักษ์ให้อัตราความผิดพลาดเชิงลบต่ำกว่า สำหรับอัตราความผิดพลาดในการตรวจจับเชิงบวก ผลที่ได้จะมีความแตกต่างกันมาก โดยวิธีแบบเบส์เชิงประจักษ์จะมีอัตราความผิดพลาดในการตรวจจับเชิงลบต่ำกว่าวิธี Lasso แต่ในกรณีศึกษาที่ 8เมื่ออัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระ เป็น 100:1000ข้อมูลเซ็นเซอร์ 50% และกรณีที่ 9 เมื่ออัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระ เป็น 100:1000ข้อมูลเซ็นเซอร์ 70% อัตราความผิดพลาดในการตรวจจับเพิ่มสูงขึ้นมาก ดังนั้น มีแนวโน้มว่า เมื่อจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่างมากขึ้น และร้อยละข้อมูลเซ็นเซอร์เพิ่มขึ้น มีแนวโน้มว่าจะส่งผลต่อประสิทธิภาพของวิธีแบบเบส์

5.2. ข้อเสนอแนะ

ผลการวิจัยครั้งนี้มีข้อเสนอแนะแบ่งเป็น 2 ด้าน คือ

5.2.1 ข้อเสนอแนะด้านการนำไปใช้งาน

จากผลการศึกษาพบว่าการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ด้วยวิธีแบบเบส์เชิงประจักษ์สำหรับตัวแบบ Cox's proportional hazard มีประสิทธิภาพกว่าวิธี Lasso ทั้งในด้านของความถูกต้อง แม่นยำ แม้ในสถานการณ์ที่มีจำนวนตัวแปรอิสระมากและระดับของร้อยละของข้อมูลเซ็นเซอร์สูง และในด้านระยะเวลาของกระบวนการทำงาน พบว่าวิธีแบบเบส์เชิงประจักษ์ยังใช้เวลาในการทำงานน้อยกว่าวิธี Lasso มาก ดังนั้น วิธีแบบเบส์เชิงประจักษ์ถือเป็นทางเลือกหนึ่งที่เหมาะสำหรับการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ในสถานการณ์ที่ข้อมูลบางส่วนเป็นข้อมูลเซ็นเซอร์ หรือเมื่อจำนวนตัวแปรอิสระมีมากกว่าขนาดตัวอย่าง (ข้อมูลมิติสูง)

5.2.1 ข้อเสนอแนะด้านการศึกษาวิจัย

เพื่อเป็นแนวทางให้ผู้ที่สนใจได้ศึกษาเพิ่มเติม ในการศึกษาครั้งต่อๆไป อาจทำการศึกษาในกรณีต่างๆดังนี้

1. เนื่องจากการทดสอบการศึกษาในครั้งนี้ กำหนดให้ตัวแปรอิสระแต่ละตัวไม่มีความสัมพันธ์กัน ในความเป็นจริง เหตุการณ์เช่นนี้มักเกิดได้น้อย หรือแทบไม่เกิดขึ้นเลย ดังนั้นในการศึกษาครั้งต่อไป อาจเพิ่มให้ตัวแปรอิสระมีความสัมพันธ์กันในลักษณะต่างๆ

2. เพื่อประโยชน์ของการนำไปใช้งาน อาจเพิ่มขนาดตัวอย่างต่อตัวแปรอิสระให้มีความแตกต่างกันมากขึ้น เนื่องจากการศึกษาครั้งนี้อาจพิจารณาปัจจัยระหว่างอัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระที่ระดับไม่แตกต่างกันมากนักจึงไม่สามารถบอกขีดจำกัดของวิธีแบบเบส์เชิงประจักษ์ว่าจะมีประสิทธิภาพมากกว่าวิธี Lasso ไปจนถึงอัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระเป็นเท่าไร

3. ทดสอบวิธีแบบเบส์เชิงประจักษ์กรณีที่ข้อมูลมีมิติสูงกับตัวแบบอื่นๆ

รายการอ้างอิง

Andrew, D. F. and Mallows, C. L. Scale mixtures of normal distributions. Journal of the Royal StatisticalSociety Series B V.36 no.1 (April 1974) : 99-102

Brian S. Everitt. Mulitivariable Modeling And Mulitivariate Analysis For The Behavioral Sciences. Taylor & Francis Group, LLC, 2010.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association V.96 (December 2001) : 348-1360.

Fan, J. and Li, R. Variable selection for Cox's proportional hazards model and frailty model, The Annals of Statistics V.30 no.1 (August 2001) : 74-99.

Johnstone, I. M. and Silverman, B. W. Needles and straw in haystacks:empirical Bayes estimates of possibly sparse sequence. The Annals of Statistics V.32 (November, 4 2004) : 1594-1649.

Lee, K.H., Chakraborty, S., and Sun, J. Bayesian variable selection in semiparametric proportional hazards model for high dimensional survival data. The International Journal of Biostatistics V.7 no.1(April 2011) : 1-32

Pungpapong, V. Empirical Bayes variable selection for high dimensional regression. Doctoral dissertation, Department of Philosophy, Purdue University, 2012

Ralf Bender, Thomas Augustin and Maria BlettnerGenerating survival times to simulate Cox proportional hazard models. Statistic in medicine 27(2004) : 1713-1723.

Stale Nygard, Ornulf Borgan, Ole Christian Lingjaerde and Hege Leite Storvold. Partial least squares Cox regression for genome-wide data. Lifetime data Anal V.14 (December, 7 2007): 179-195.

Tibshirani, R. The Lasso method for variable selection in the Cox model. Statistic in medicine Vol.16(December 1997) : 385-395.

Zhang, H. H. and Lu, W. Adaptive lasso for Cox's proportional hazards model. Biometrika 94 (March 2007) : 1-13.

Zou, H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101(2006) : 1418-1429.

บรรณานุกรม

ภาษาไทย

วีรานันท์ พงศาภักดี. การวิเคราะห์ข้อมูลเชิงกลุ่ม: ทฤษฎีและการประยุกต์. พิมพ์ครั้งที่ 1. นครปฐม: โรงพิมพ์มหาวิทยาลัยศิลปากร, 2544.

ภาษาอังกฤษ

Hastie, T. J. and Tibshirani, R. J. Generalized Adaptive Model. Monograpgs on Statistics and Applied Probability 43. London: Chapman & Hall, 1990.

McCullagh, P. and Nelder, J. A. Generalized Linear Model. Second edition. Monograpgs on Statistics and Applied Probability 37. London: Chapman & Hall, 1990.

Mitchell, T. J. and Beauchamp, J. J. Bayesian variable selection in linear regression (with Discussion). Journal of the American Statistical Association V.83 no.1 (December 1988): 1023-1036.

Park, T. and Casella, G. The Bayesian lasso. Journal of the American Statistical Association V.103 no.482 (June 2008) : 681-686.

Peter C. Austin Generating survival times to simulate Cox proportional hazards models with time-varying covariates. Statistics in Medicine V.31 issue 29 (July, 4 2012) : 3946-3958.

Tibshirani, R. Regression shrinkage and selection via the lasso.Journal of the Royal Statistical Society. Series B 58 no.1 (January 1996) : 267-288.

Searle, S. R.Linear model. Canada:John Wiley & Sons, 1971.

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวอรณิชา ห่อนบุญเหิม เกิดวันที่ 16 มิถุนายน พ.ศ. 2531 สำเร็จการศึกษาปริญญาวิทยาศาสตร์บัณฑิต (วท.บ.) สาขาวิชาคณิตศาสตร์ ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2553 และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตร์มหาบัณฑิต (วท.ม.) สาขาสถิติ ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2555