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

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

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 $P_{\lambda}(\beta) = \lambda\sum_{j=1}^{p}|\beta_j|$ สำหรับ $l_1 - norm$ Penalty function นี้ เปรียบเสมือนการหาค่าภาวะน่าจะเป็นสูงสุด โดยมีข้อจำกัด คือ $\sum_{j=1}^{p}|\beta_j| \leq s$ 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 $|\beta_j| < \lambda$ ค่า Penalty function จะเขียนได้ในรูป

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

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

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

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

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

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 $P_\lambda(\beta) = \lambda\sum_{j=1}^{p} \hat{w}_j |\beta_j|$ 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 โดยมีแนวคิดคือสัมประสิทธิ์ถดถอยสามารถคำนวณได้จากค่ามัธยฐานของการแจกแจงของฟังก์ชันความน่าจะเป็นภายใต้เงื่อนไขที่ว่าค่าพารามิเตอร์และค่าสัมประสิทธิ์ตัวอื่นๆที่ไม่ใช่ตัวที่ต้องการหาค่าประมาณเป็นค่าคงที่ (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

$X = \left(x_1^T, x_2^T, ..., x_n^T\right)^T$ 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

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

และ

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

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