บทที่ 3

วิธีดำเนินการศึกษา

ในงานวิจัยครั้งนี้เป็นการศึกษาการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์แบบเบย์เชิงประจักษ์ที่ทำงานร่วมกับวิธีการทำซ้ำแบบมีเงื่อนไขฐานนิยม/มัธยฐานสำหรับตัวแบบCox's proportional hazard ในกรณีที่ข้อมูลที่ต้องการศึกษามีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง พร้อมทั้งเปรียบเทียบประสิทธิภาพการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ด้วยวิธีข้างต้น เมื่อจำลองข้อมูลที่มีลักษณะแตกต่างกัน 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

$$H(t,x,\varsigma) = \int_0^t \exp(\beta' x) h_0(u)\,du$$

$$= \lambda \exp(\beta' x)\int_o^t \nu u^{\nu-1} du$$

$$= \lambda \exp(\beta' x)\left[u^{\nu}\right]_0^t$$

$$= \lambda \exp(\beta' x) t^{\nu}$$

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

$$H_0(t) = \int_0^t h_0(u)\,du$$

$$= \int_0^t \lambda \nu u^{\nu-1} du$$

$$= \lambda \nu \left(u^{\nu}\Big|_{u=0}^{u=t}\right) = \lambda t^{\nu}$$

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

$$S(t \mid x = 0) = S_0(t)$$

$$= \exp\left(-H_0(t)\right)$$

$$= \exp\left(-\lambda t^{\nu}\right)$$

ฟังก์ชันความหนาแน่น

$$f_0(t) = h_0(t)\exp\left(-H_0(t)\right)$$

$$= \lambda \nu t^{\nu-1} \exp\left(-\lambda t^{\nu}\right)$$

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 $U \sim U[0,1]$

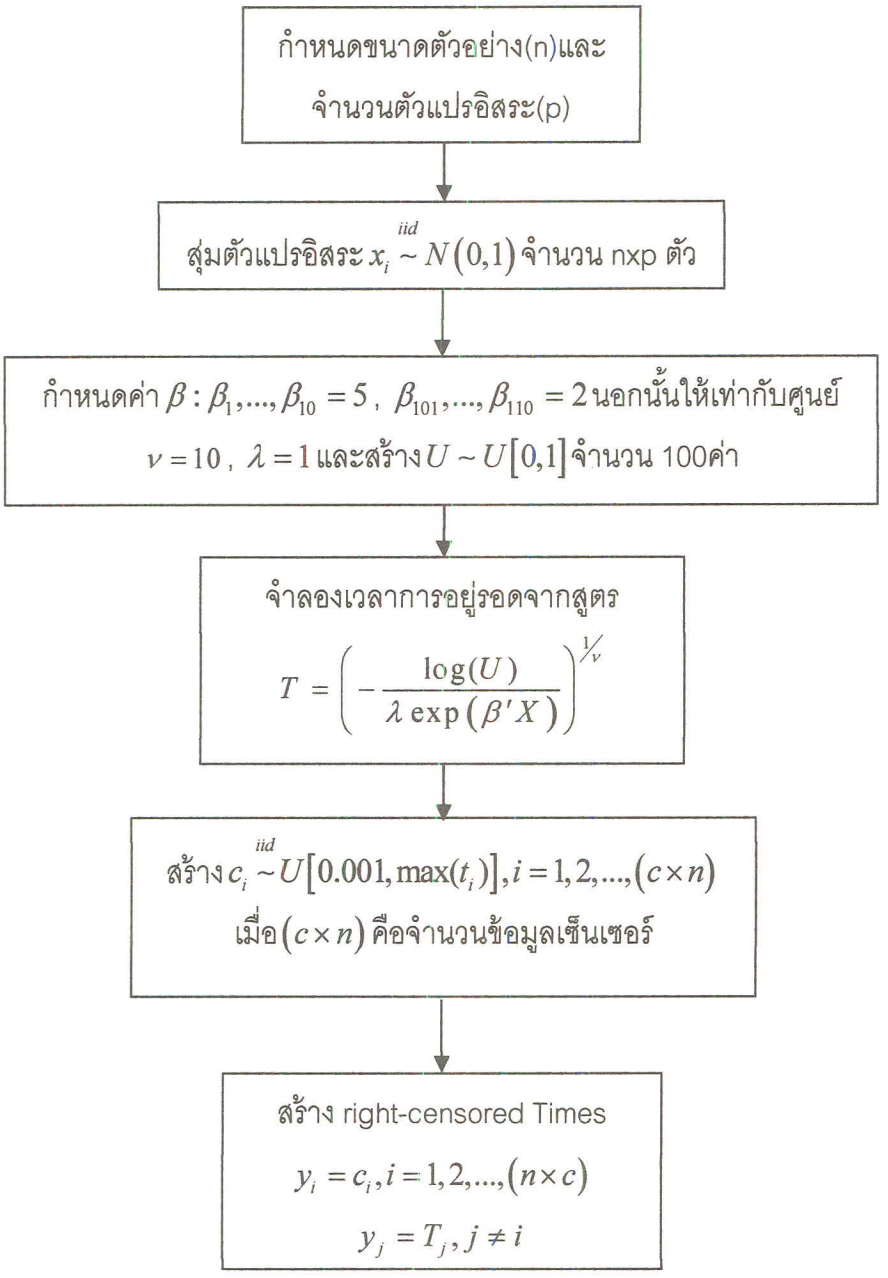
ในการศึกษาครั้งนี้ให้ $\nu = 10$, $\lambda = 1$

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

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

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

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

<u>Prior:</u> $$\beta_j \sim (1-\omega)\delta_0(\beta_j) + \omega\gamma(\beta_j)$$

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

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

2.2 การคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์

ให้ $T = (t_1, t_2, \ldots, t_n)$ คือ เวกเตอร์ของRight-censored timeของตัวอย่างขนาด n ตัวอย่าง $X = (x_1^T, x_2^T, \ldots, x_n^T)^T$ เป็นเมทริกซ์ขนาด $n \times p$ ของตัวแปรอิสระจำนวน p ตัว ตัวแบบCox's hazardคือ

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

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

โดยที่

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

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

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

ดังนั้น $z_i \approx N\left(x_i^T \hat{\beta}, w_i^{-1}\right)$ พิจารณา $\beta_j, j = 1, 2, \ldots, p$ 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

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

จากแนวคิดข้างต้นสามารถเขียนเป็นขั้นตอนการทำงานได้ดังนี้

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

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

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 $j = 1,2,\ldots,p$ 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

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

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

เมื่อ

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

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

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

และ

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

โดยที่

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

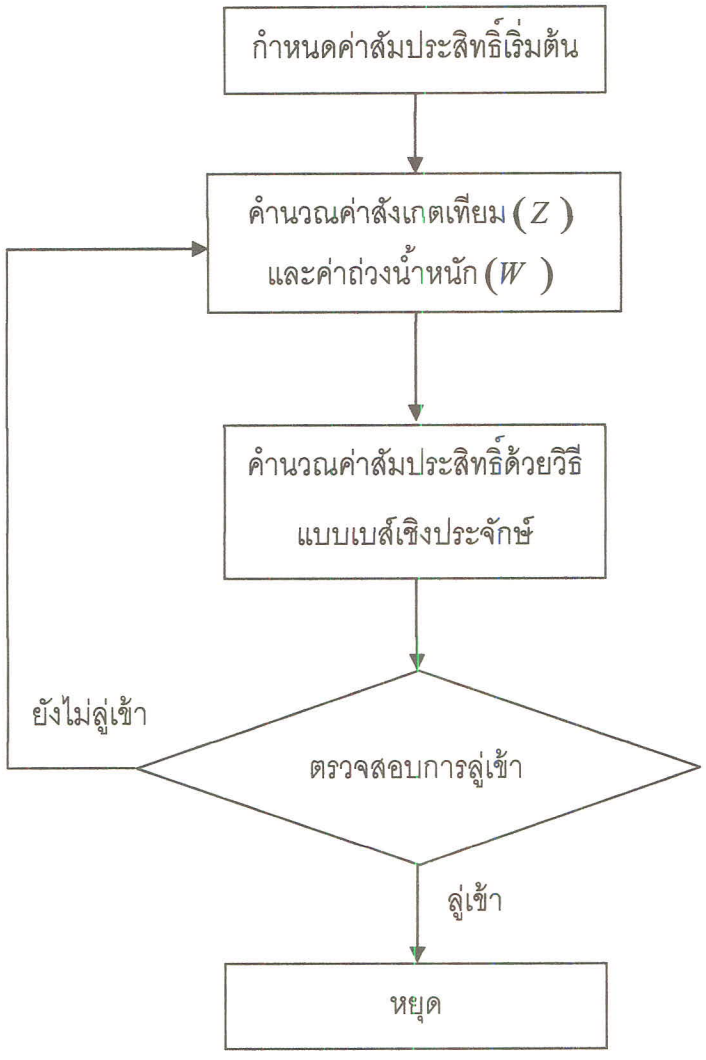
4. อัพเดทค่า $\hat{\omega}^{(k+1)}$ จาก

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

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

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

3.2. อัตราความผิดพลาดในการตรวจจับเชิงลบ (False negative rate) สามารถคำนวณได้จากสูตร

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

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

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

และ

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

4. การวิเคราะห์ผลลัพธ์

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