บทความวิชาการ (Scholarly Article)

บทวิเคราะห์วิธีวิเคราะห์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูง

A Brief Review on High-dimensional Linear Regression

วิฐรา พึ่งพาพงศ์*

ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี

จุฬาลงกรณ์มหาวิทยาลัย แขวงวังใหม่ เขตปทุมวัน กรุงเทพมหานคร 10330

Vitara Pungpapong*

Department of Statistics, Faculty of Commerce and Accountancy,

Chulalongkorn University, Wangmai, Pathumwan, Bangkok 10330

บทคัดย่อ

การวิเคราะห์การถดถอยเชิงเส้นเป็นเครื่องมือที่ไม่สลับซับซ้อน แต่มีประโยชน์อย่างมากในการวิเคราะห์ข้อมูล ดังเห็นได้จากการประยุกต์ใช้งานที่แพร่หลายในปัจจุบัน ความก้าวหน้าที่รวดเร็วทางเทคโนโลยีช่วยให้เราสามารถจัดเก็บข้อมูลขนาดใหญ่ซึ่งมีจำนวนตัวแปรมหาศาลได้อย่างมีประสิทธิภาพมากขึ้น การวิเคราะห์การถดถอยแบบดั้งเดิมมีข้อจำกัดหนึ่งที่สำคัญคือสามารถประยุกต์ใช้งานได้เฉพาะกับข้อมูลที่มีขนาดตัวอย่างมากกว่าจำนวนตัวแปรอิสระ ในกรณีที่จำนวนตัวแปรอิสระมากกว่าจำนวนตัวอย่าง เราจะเรียกข้อมูลประเภทนี้ว่าข้อมูลที่มีมิติสูง บทความนี้มีวัตถุประสงค์ที่จะนำเสนอเครื่องมือเพื่อเป็นทางเลือกในการจัดการกับข้อมูลที่มีมิติสูงซึ่งรวมถึงวิธี penalized regression และวิธีเบส์ ตลอดจนข้อควรคำนึงถึงในการเลือกใช้เครื่องมือสำหรับการวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูง

คำสำคัญ : ข้อมูลที่มีมิติสูง; การวิเคราะห์การถดถอยเชิงเส้น; วิธี penalized regression; วิธีเบส์

Abstract

Linear regression is a simple yet powerful tool in data analysis as seen in many modern applications. Rapid technological advancements enable us to collect big data with massive number of variables more effectively. The traditional regression model has a major limitation in which it can only deal with datasets whose sample size is larger than number of independent variables. When number of independent variables is beyond the number of observations, such data is referred as high-dimensional data. The aim of this article is to introduce the alternative tools that can handle high-dimensional data including penalized regression and Bayesian approach. The concerns in choosing high-dimensional regression tools are also discussed here.

***ผู้รับผิดชอบบทความ :** vitara@cbs.chula.ac.th

Keywords: high-dimensional data; linear regression analysis; penalized regression; Bayesian approach

1. บทนำ

ด้วยเทคโนโลยีคอมพิวเตอร์ที่ก้าวหน้าในปัจจุบัน ข้อมูลมีการจัดเก็บในรูปแบบอิเล็กทรอนิกส์ที่ทันสมัยและมีประสิทธิภาพมากขึ้น นักวิเคราะห์ข้อมูลต้องเผชิญกับข้อมูลขนาดใหญ่และสลับซับซ้อนมากขึ้น บ่อยครั้งที่ข้อมูลเหล่านี้มีจำนวนตัวแปรมากกว่าขนาดตัวอย่างซึ่งเครื่องมือทางสถิติแบบดั้งเดิมหลาย ๆ เครื่องมือไม่สามารถนำมาใช้วิเคราะห์ได้ รวมถึงการวิเคราะห์การถดถอยเชิงเส้นแบบดั้งเดิมที่มีเงื่อนไขว่าขนาดตัวอย่างมากกว่าจำนวนตัวแปรอิสระ ในการวิเคราะห์การถดถอยเชิงเส้นแบบดั้งเดิม กรณีที่ตัวแปรอิสระมีจำนวนมากกว่าขนาดตัวอย่างหรือที่เรียกว่าข้อมูลที่มีมิติสูง ผู้วิเคราะห์จะต้องเผชิญความท้าทาย 3 ประการ คือ (1) ไม่สามารถประมาณค่าสัมประสิทธิ์การถดถอยได้ด้วยวิธีกำลังสองน้อยสุด (2) ปัญหาตัวแปรอิสระที่มีความสัมพันธ์เชิงเส้นสูงซึ่งจะส่งผลต่อการประมาณค่าที่ไม่เสถียรและไม่มีประสิทธิภาพ และ (3) การแปลผลลัพธ์ของตัวแบบมีความยากและสลับซับซ้อน

บทความนี้มีวัตถุประสงค์ในการนำเสนอทางเลือกในการวิเคราะห์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูง ซึ่งสามารถแก้ไขหรือลดปัญหา 3 ประการ ข้างต้น รวมถึงนำเสนอบทวิเคราะห์ในการเลือกใช้เครื่องมือสำหรับการวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูงให้มีความเหมาะสมในการประยุกต์ใช้งาน ผู้เขียนหวังว่าบทความนี้คงเป็นประโยชน์ต่อการประยุกต์ใช้ในงานของท่าน หรือเพื่อศึกษาค้นคว้าต่อไป

2. ปัญหาในการวิเคราะห์การถดถอยเชิงเส้นแบบดั้งเดิมสำหรับข้อมูลที่มีมิติสูง

ในการวิเคราะห์การถดถอยเชิงเส้น กรณีที่ขนาดตัวอย่างเท่ากับ n และตัวแปรอิสระ p ตัว สามารถเขียนสมการตัวแบบได้ ดังนี้

$$Y = \boldsymbol{X}\beta + \varepsilon \tag{1}$$

เมื่อ Y คือ คอลัมน์เวกเตอร์ของตัวแปรตามขนาด n

$\boldsymbol{X}$ คือ เมทริกซ์ของตัวแปรอิสระขนาด $n \times p$

β คือ คอลัมน์เวกเตอร์ของสัมประสิทธิ์การถดถอยขนาด p

และ ε คือ คอลัมน์เวกเตอร์ของค่าคลาดเคลื่อนขนาด n โดยที่ $\varepsilon \sim N(0, \sigma^2 \boldsymbol{I}_n)$

โดยทั่วไป ค่าประมาณของสัมประสิทธิ์การถดถอยจะหาได้โดยวิธีกำลังสองน้อยสุด ซึ่งจะหาได้จาก

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|^2 = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T Y \tag{2}$$

สำหรับกรณีที่ข้อมูลมีมิติสูงหรือข้อมูลที่ $p > n$ การวิเคราะห์การถดถอยเชิงเส้นแบบดั้งเดิมนั้นมีข้อจำกัด 3 ประการ ดังที่ได้กล่าวในบทนำ ดังนี้

2.1 ปัญหาในการประมาณค่าสัมประสิทธิ์

จากสมการที่ (2) ค่าประมาณของสัมประสิทธิ์จะหาค่าได้ก็ต่อเมื่อข้อมูลมีขนาดตัวอย่างมากกว่าจำนวนตัวแปรอิสระเท่านั้น ($p \leq n$) เนื่องจากหากข้อมูลมีมิติสูง ($p > n$) แล้ว เมทริกซ์ $\boldsymbol{X}^T \boldsymbol{X}$ จะเป็นเมทริกซ์เอกฐาน (singular matrix) ซึ่งไม่สามารถหาเมทริกซ์ผกผัน (inverse matrix) ได้ จึงไม่สามารถหาค่า $\hat{\beta}$ โดยวิธีกำลังสองน้อยสุดได้

2.2 ปัญหาตัวแปรอิสระที่มีความสัมพันธ์เชิงเส้นสูง

นอกจากปัญหาในการประมาณค่าสัมประสิทธิ์การถดถอยแล้ว โดยทั่วไปการที่จำนวนตัวแปรอิสระมีเป็นจำนวนมาก ผู้วิเคราะห์มักเผชิญกับปัญหาที่ตัวแปรอิสระมีความสัมพันธ์เชิงเส้นสูง (multicollinearity) ซึ่งจะส่งผลให้ความแปรปรวนและส่วนเบี่ยงเบนมาตรฐานของตัวประมาณสัมประสิทธิ์ที่ได้จากวิธีกำลังสองน้อยสุดมีค่ามากหรือตัวประมาณที่ได้มีความไม่เสถียรนั่นเอง

2.3 ปัญหาในการแปลผลลัพธ์ของตัวแบบ

กรณีที่ตัวแปรอิสระมีจำนวนมาก ตัวแบบที่ดีควรมีเฉพาะตัวแปรอิสระที่มีความสำคัญกับตัวแปรตามในตัวแบบเท่านั้น ทั้งนี้เพื่อให้ได้ตัวแบบที่สามารถแปลผลได้โดยง่ายและไม่สลับซับซ้อน แนวคิดในการคัดเลือกตัวแปรเข้าตัวแบบจึงมีความสำคัญเป็นอย่างมาก ในการวิเคราะห์การถดถอยเชิงเส้นแบบดั้งเดิมการคัดเลือกตัวแปรสามารถทำได้โดยการทดสอบสมมติฐานว่าสัมประสิทธิ์การถดถอยเท่ากับศูนย์หรือไม่โดยใช้ตัวสถิติทดสอบทีและตัวสถิติทดสอบเอฟ อย่างไรก็ตาม การทดสอบสมมติฐานเหล่านี้ไม่สามารถใช้งานได้กับข้อมูลที่มีมิติสูง ทั้งนี้เนื่องมากจากตัวสถิติทดสอบดังกล่าวถูกพัฒนาขึ้นจากแนวคิดการทดสอบด้วยอัตราส่วนภาวะน่าจะเป็น (likelihood ratio test) ซึ่งอ้างอิงตัวประมาณจากวิธีกำลังสองน้อยสุด

3. การวิเคราะห์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูง

บทความนี้ ผู้เขียนต้องการนำเสนอวิธีการ 2 วิธี ในการวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูง คือ (1) วิธี penalized regression ซึ่งเป็นวิธีที่มีแนวคิดแบบ frequentist หรือ non-Bayesian และ (2) วิธีเบส์ (Bayesian approach) แม้ว่าวิธีทั้งสองจะมีแนวคิดแตกต่างกันสิ้นเชิง แต่ก็มีจุดมุ่งหมายหลักเดียวกัน คือ การประมาณค่าสัมประสิทธิ์การถดถอยซึ่งสามารถใช้ได้กับข้อมูลที่มีมิติสูง และสามารถแก้ไขหรือลดปัญหา 3 ประการ ของการวิเคราะห์การถดถอยเชิงเส้นแบบดั้งเดิมได้

3.1 วิธี Penalized Regression

วิธี penalized regression เป็นวิธีที่นิยมใช้กันอย่างแพร่หลายในการประมาณค่า β เมื่อข้อมูลมีมิติสูง ซึ่งตัวประมาณดังกล่าวจะหาได้จากการหาค่า β ที่ทำให้ฟังก์ชันเป้าหมาย (objective function) ดังสมการที่ (3) มีค่าต่ำสุด

$$\hat{\beta} = \operatorname{argmin}_{\beta}\left\|Y - \sum_{j=1}^{p} X_j\beta_j\right\|^2 + P_\lambda(\beta) \qquad (3)$$

จากสมการที่ (3) จะมีความคล้ายคลึงกับวิธีกำลังสองน้อยสุดที่ใช้กันโดยทั่วไปในการหาประมาณค่า β แต่จะมีฟังก์ชัน $P_\lambda(\beta)$ เพิ่มขึ้นมาซึ่งเรียกว่า penalty function โดยฟังก์ชันนี้จะอยู่ในรูปของ β และมีพารามิเตอร์ λ ซึ่งมีค่ามากกว่าหรือเท่ากับศูนย์ เพื่อใช้ในการให้น้ำหนักของ penalty function ดังกล่า

สำหรับ penalty function นั้นมีอยู่ด้วยกันหลายรูปแบบ ซึ่งจะให้ค่าประมาณที่แตกต่างกัน ส่วนพารามิเตอร์ λ โดยทั่วไปจะใช้วิธี cross-validation ในการหาค่า λ ที่เหมาะสมสำหรับข้อมูลที่ต้องการวิเคราะห์ ในบทความนี้จะขอกล่าวถึง penalized regression ที่เป็นที่รู้จักกันดีและใช้กันอย่างแพร่หลาย 3 รูปแบบ ดังนี้

3.1.1 วิธี ridge regression หากเรามีความเชื่อว่าตัวแปรอิสระทุกตัวมีความสัมพันธ์กับตัวแปรตาม เราสามารถเลือกใช้วิธี ridge regression ซึ่งมี penalty function ดังสมการที่ (4)

$$P_\lambda(\beta) = \lambda\sum_{j=1}^{p}\beta_j^2 \qquad (4)$$

การใช้ penalty function ดังกล่าวจะเป็นการดึงค่าประมาณเข้าสู่ศูนย์ ทำให้ได้ค่า $\hat{\beta}$ ทุกตัวมีขนาดเล็กแต่ไม่เท่ากับศูนย์ (dense estimator) ดังนั้นวิธี ridge regression จึงมีความเหมาะสมในการวิเคราะห์ข้อมูลที่มีสัมประสิทธิ์ขนาดเล็กที่ไม่เท่ากับศูนย์จำนวนมาก [7] ตัวประมาณ ridge regression ที่ได้อยู่ในรูปดังสมการที่ (5)

$$\hat{\beta} = \left(\boldsymbol{X^T X} + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{X^T} Y \qquad (5)$$

สมบัติเด่นของ ridge regression คือตัวประมาณที่ได้จะมีความเสถียร จึงเป็นอีกวิธีที่ใช้กันมากเพื่อแก้ไขปัญหาตัวแปรอิสระมีความสัมพันธ์กันสูง แม้ว่าข้อมูลจะไม่มีมิติสูงก็ตาม สำหรับข้อเสียของวิธีนี้ คือ การขาดสมบัติในการเลือกตัวแปร ทำให้การแปลผลตัวแปรในตัวแบบทำได้ยาก

3.1.2 วิธี Lasso Tibshirani คิดค้นวิธี Lasso โดยมีวัตถุประสงค์เพื่อให้เป็นวิธีที่สามารถเลือกตัวแปรเข้าสู่ตัวแบบและประมาณค่า β ในการวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูงได้ในคราวเดียวกัน [7] วิธี Lasso มี penalty function ดังสมการที่ (6)

$$P_\lambda(\beta) = \lambda \sum_{j=1}^{p} |\beta_j| \qquad (6)$$

ตัวประมาณที่ได้จากวิธี Lasso จะมีค่า $\hat{\beta}$ ส่วนใหญ่เป็นศูนย์และค่า $\hat{\beta}$ บางส่วนไม่เท่ากับศูนย์ (sparse estimator) ดังนั้นวิธี Lasso จึงเป็นการเลือกตัวแปรได้โดยอัตโนมัติ ซึ่งวิธีนี้ถูกนำไปใช้งานจริงอย่างแพร่หลาย เนื่องจากเป็นวิธี penalized regression วิธีแรกที่มีความสามารถในการเลือกตัวแปร การคำนวณไม่ยุ่งยาก และปัจจุบันมีโปรแกรมสถิติหลายโปรแกรมที่มีฟังก์ชันในการคำนวณตัวประมาณ Lasso อย่างไรก็ตาม ข้อจำกัดของวิธี Lasso ในการวิเคราะห์ข้อมูลที่มีมิติสูง [8] คือ (1) Lasso สามารถเลือกตัวแปรเข้าได้มากที่สุดจำนวน n ตัว ดังนั้นหากข้อมูลมีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่างเป็นจำนวนมาก ตัวแบบที่ได้จากวิธี Lasso อาจจะไม่มีความเหมาะสม และ (2) ในกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันสูง วิธี Lasso มีแนวโน้มที่จะเลือกตัวแปรเพียงตัวแปรเดียวจากกลุ่มตัวแปรอิสระที่มีความสัมพันธ์กันสูงเข้าตัวแบบ โดยไม่สนใจว่าจะเป็นตัวแปรใดในกลุ่ม

วิธี Lasso จะมีประสิทธิภาพสูงในการพยากรณ์ เมื่อตัวแบบจริงมีจำนวนตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามมีจำนวนไม่มากนัก และขนาดของ β ที่ไม่เท่ากับศูนย์มีขนาดใหญ่ [7]

3.1.3 วิธี elastic net เป็นวิธีที่คิดค้นโดย Zou [8] เพื่อแก้ไขข้อจำกัดของ Lasso ดังที่ได้กล่าวไว้ในหัวข้อ 1.2 ซึ่งมี Penalty Function อยู่ในรูป

$$P_\lambda(\beta) = \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j| \quad (7)$$

จาก penalty function ดังสมการที่ (7) จะเห็นว่าวิธี elastic net เป็นวิธีที่ผสมผสานระหว่าง ridge regression และวิธี Lasso เข้าด้วยกัน ดังนั้นจึงได้ตัวประมาณที่มีค่า $\hat{\beta}$ บางค่าที่ไม่เท่ากับศูนย์ แต่จะมีความเสถียรมากกว่าวิธี Lasso จากสมการที่ (7)

จากการที่วิธี elastic net เป็นวิธีที่สามารถเลือกตัวแปรและประมาณค่าได้ในคราวเดียวกันเฉกเช่นเดียวกับวิธี Lasso และยังสามารถแก้ไขข้อจำกัดของวิธี Lasso ได้ ดังนั้น วิธี elastic net จึงมีความเหมาะสมในการวิเคราะห์ข้อมูลที่มีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่างเป็นจำนวนมากและตัวแปรอิสระมีความสัมพันธ์กันสูง

3.2 วิธีเบส์ (Bayesian approach)

วิธีเบส์เป็นอีกวิธีหนึ่งที่สามารถใช้ในการวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูงได้ ในบทความนี้จะกล่าวถึงการเลือกใช้การแจกแจงก่อน (prior distribution) 4 รูปแบบ ซึ่ง 3 รูปแบบแรก มี

ความสัมพันธ์กับ penalized regression ที่ได้กล่าวถึงก่อนหน้านี้ คือ วิธี ridge regression วิธี Lasso และวิธี elastic net ตามลำดับ ซึ่งให้ค่าประมาณที่ใกล้เคียงกับค่าประมาณที่ได้จากวิธี penalized regression แต่ขาดสมบัติการเลือกตัวแปรเข้า ดังนั้นในบทความนี้จึงได้เสนอการแจกแจงก่อนในรูปแบบ spike and slab เพื่อแก้ไขปัญหาดังกล่าว

3.2.1 วิธี Bayesian ridge ตัวประมาณ ridge regression ที่ได้จากวิธี penalized regression จะหาได้จากสูตรดังสมการที่ (5) ซึ่งมีค่าเท่ากับค่าเฉลี่ยของการแจกแจงภายหลัง (posterior mean) และยังเท่ากับค่าฐานนิยมของการแจกแจงภายหลัง (posterior mode) เมื่อใช้การแจกแจกแจงก่อนของพารามิเตอร์ β แบบปกติที่ค่า σ^2 คงที่ใด ๆ

$$\beta|\sigma^2 \sim N(\mathbf{0}_p, \frac{\sigma^2}{\lambda}\boldsymbol{I}_p) \qquad (8)$$

จากการแจกแจงก่อนดังสมการที่ (8) จะได้การแจกแจงภายหลัง (posterior distribution) เป็นการแจกแจงแบบปกติเช่นกัน ดังนั้นค่าเฉลี่ยของการแจกแจงภายหลังจึงเท่ากับฐานนิยมของการแจงแจงหลัง เราสามารถเลือกใช้ตัวแบบเบส์เป็นลำดับชั้น (Bayesian hierarchical model) โดยให้การแจกแจงก่อนของ σ^2 มีการแจกแจงแบบ Jeffrey (Jeffrey's prior) ดังสมการที่ (9)

$$\begin{aligned} &Y|\boldsymbol{X},\beta,\sigma^2 \sim N(\boldsymbol{X}\beta, \sigma^2\boldsymbol{I}_n) \\ &\beta|\sigma^2 \sim N(\mathbf{0}_p, \frac{\sigma^2}{\lambda}\boldsymbol{I}_p) \\ &P(\sigma^2) \propto 1/\sigma^2 \end{aligned} \qquad (9)$$

สำหรับพารามิเตอร์ λ มีวิธีการหาสองวิธีหลักๆ ในวิธีแบบเบส์ คือ (i) วิธีเบส์เชิงประจักษ์ (Empirical Bayes) หรือ การหาค่า λ ที่ทำให้ความน่าจะเป็นขอบ (Marginal Likelihood) มีค่าสูงสุด และ (ii) วิธีเบส์แบบลำดับชั้น โดยให้การแจกแจงก่อนกับพารามิเตอร์ λ^2 เป็นการแจกแจงแบบแกมมาผกผัน (Inverse Gamma) จาก (9) จะได้ว่า การแจกแจงภายหลังของพารามิเตอร์ β จะอยู่ในรูปแบบการแจกแจงปกติที่มีค่าเฉลี่ย $(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T Y$ ซึ่งเท่ากับตัวประมาณ Ridge Regression ที่ได้จากวิธี Penalized Likelihood ดังนั้นเมื่อทำการเลือกตัวอย่างแบบกิบส์ (Gibbs Sampling) จากการแจงแจงหลังแล้ว สามารถเลือกใช้ค่าเฉลี่ยจากการแจกแจงภายหลังเป็นค่าประมาณ β ได้

3.2.2 วิธี Bayesian Lasso ตัวประมาณ Lasso จากวิธี penalized regression คือ ฐานนิยมของการแจกแจงภายหลังเมื่อให้การแจกแจงก่อนของพารามิเตอร์ β เป็นแบบ double-exponential ที่ค่า σ^2 คงที่ใด ๆ [7]

$$P(\beta) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} exp(-\lambda\,|\beta_j|/\sqrt{\sigma^2}) \qquad (10)$$

Park และ Casella เสนอตัวแบบเบส์เป็นลำดับชั้น โดยให้การแจกแจงก่อนของ β แบบมีเงื่อนไขที่ σ^2 เป็นการแจกแจงแบบ double-exponential และให้การแจกแจงก่อนของ σ^2 อยู่ในรูปการแจกแจงก่อนแบบ Jeffreys [6] ดังนี้

$$\begin{aligned} &Y|\boldsymbol{X},\beta,\sigma^2 \sim N(\boldsymbol{X}\beta, \sigma^2\boldsymbol{I}_n) \\ &P(\beta|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} exp\left(-\lambda\frac{|\beta_j|}{\sqrt{\sigma^2}}\right) \\ &P(\sigma^2) \propto 1/\sigma^2 \end{aligned} \qquad (10)$$

การแจกแจงแบบ double-exponential สามารถเขียนแทนได้อยู่ในรูปแบบการแจกแจงผสมของการแจกแจงปกติ [1] ซึ่งการแจกแจงผสมนี้ถูกเลือกนำมาใช้ในวิธี Bayesian Lasso ที่นำเสนอโดย Park และ Casella เพื่อง่ายต่อการจำลองข้อมูลจากการแจกแจงภายหลังโดยวิธีกิบส์ [6] ค่าพารามิเตอร์ λ สามารถหาได้จากวิธีเบส์เชิงประจักษ์ หรือจะให้การแจกแจงก่อนกับพารามิเตอร์ λ^2 เป็นการแจกแจงแบบแกมมาก็ได้ เมื่อได้ข้อมูลที่จำลองจากวิธีการเลือกตัวอย่างแบบกิบส์เป็นที่เรียบร้อยแล้ว Park และ Casella เสนอให้ใช้ค่ามัธยฐานของการแจกแจงภายหลัง (posterior median) เป็นค่าประมาณพารามิเตอร์ β ซึ่งค่าที่ได้มีความใกล้เคียงกับ

ค่าประมาณ Lasso จากวิธี penalized regression [6]

3.2.3 วิธี Bayesian elastic net เช่นเดียวกับตัวประมาณ Lasso ตัวประมาณ elastic net จากวิธี penalized regression สามารถหาได้จากฐานนิยมของการแจกแจงภายหลังเมื่อใช้การแจกแจงก่อนของพารามิเตอร์ β ดังสมการที่ (11) ที่ค่า σ^2 คงที่ใด ๆ [9]

$$P(\beta) \propto exp\{-\lambda_1 \sum_{j=1}^{p} \beta_j^2 - \lambda_2 \sum_{j=1}^{p} |\beta_j|\} \quad (12)$$

จากแนวคิดข้างต้น Li และ Lin ได้เสนอตัวแบบเบส์เป็นลำดับชั้น โดยเลือกใช้การแจกแจงก่อนของ β และ σ^2 [4] ดังนี้

$$Y|\boldsymbol{X}, \beta, \sigma^2 \sim N(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}_n)$$

$$P(\beta|\sigma^2) \propto exp\left\{-\frac{1}{2\sigma^2}\left(\lambda_1 \sum_{j=1}^{p} \beta_j^2 - \lambda_2 \sum_{j=1}^{p} |\beta_j|\right)\right\} \quad (13)$$

$$P(\sigma^2) \propto 1/\sigma^2$$

Li และ Lin เสนอตัวแปรช่วย (auxiliary parameter) เพื่อให้การจำลองข้อมูลจากการแจกแจงภายหลังโดยวิธีการเลือกตัวอย่างแบบกิบส์ทำได้ง่ายขึ้น [4] เนื่องจากตัวแบบลำดับชั้นใน (13) ให้การแจกแจงแบบมีเงื่อนไขที่ไม่อยู่ในรูปการแจกแจงที่เป็นที่รู้จักโดยทั่วไป ค่าพารามิเตอร์ λ_1 และ λ_2 สามารถหาได้จากทั้งวิธีเบส์เชิงประจักษ์และวิธีเบส์แบบลำดับชั้น โดยให้การแจกแจงก่อนของ λ_1^2 เป็นการแจกแจงแบบเกาส์ผกผันทั่วไป (generalized inverse gaussian) และการแจกแจงก่อนของ λ_2^2 เป็นการแจกแจงแบบแกมมา ค่าประมาณแบบจุดที่ได้ของพารามิเตอร์ β ในวิธี Bayesian elastic net คือค่ามัธยฐานของการแจกแจงภายหลัง ซึ่งคำนวณจากข้อมูลที่จำลองจากวิธีการเลือกตัวอย่างแบบกิบส์ ค่าประมาณจากวิธี Bayesian elastic net จะมีค่าใกล้เคียงกับค่าประมาณที่ได้จากวิธี elastic net ที่มาจากวิธี penalized regression [4]

3.2.4 การใช้การแจกแจงก่อนแบบ spike and slab จากวิธีเบส์ทั้งสามวิธีข้างต้น การแจกแจงก่อนของ β เป็นการแจกแจงแบบต่อเนื่องทั้งสิ้น ซึ่งจะทำให้ได้การแจกแจงภายหลังเป็นการแจกแจงแบบต่อเนื่องด้วยทำให้ค่าประมาณแบบจุดของพารามิเตอร์ β ที่ได้จากค่าเฉลี่ยหรือค่ามัธยฐานของการแจกแจงภายหลังส่วนใหญ่จะมีขนาดเล็กเข้าใกล้ค่าศูนย์แต่จะไม่เท่ากับศูนย์ ดังนั้นวิธีทั้งสามข้างต้นจึงไม่มีสมบัติในการเลือกตัวแปรเข้า เพื่อให้ได้ตัวแบบที่มีสมบัติในการเลือกตัวแปรเข้า เราสามารถเลือกใช้การแจกแจงก่อนในรูปแบบที่เรียกว่า spike and slab [2, 3,5] ให้กับพารามิเตอร์ β โดยการแจกแจงแบบ spike and slab เป็นการแจกแจงที่ประกอบไปด้วยการแจกแจง 2 ส่วน คือ ในส่วนของ spike จะเป็นการให้มวลที่จุดศูนย์สำหรับพารามิเตอร์ β ตัวที่เท่ากับศูนย์ และในส่วนของ slab จะระบุการแจกแจงของพารามิเตอร์ β ตัวที่ไม่เท่ากับศูนย์ ซึ่งเขียนการแจกแจงก่อนได้ ดังนี้

$$P(\beta|\gamma) = P_{Slab}(\beta_\gamma) \prod_{j:\gamma_j=0} \delta(\beta_j) \quad (14)$$

เมื่อ $\gamma = (\gamma_1, \ldots, \gamma_p)$ โดยที่ $\gamma_j = \begin{cases} 1, & \beta_j \neq 0 \\ 0, & \beta_j = 0 \end{cases}$ คือตัวแปรบ่งชี้ว่าตัวสัมประสิทธิ์มีค่าเท่ากับศูนย์หรือไม่

$\delta(\beta_j)$ คือ ดิแรกเดลต้าฟังก์ชัน (dirac delta function) ซึ่งมีสมบัติ ดังนี้

$$\delta(\beta_j) = \begin{cases} 0, & \beta_j \neq 0 \\ \infty, & \beta_j = 0 \end{cases} \text{ และ } \int_{-\infty}^{\infty} \delta(\beta_j)\, \partial\beta_j = 1$$

และ $P_{Slab}(\beta_\gamma)$ คือ การแจกแจงร่วมสำหรับพารามิเตอร์ β ที่ไม่เท่ากับศูนย์

หากเราให้การแจกแจงของ β แต่ละตัวเป็นอิสระกัน เราสามารถเขียนสมการ (14) อยู่ในรูปสมการ (15) ได้

$$P(\beta|\gamma) = \prod_{j=1}^{p} \{(1-\gamma_j)\delta(\beta_j) + \gamma_j P_{Slab}(\beta_j)\} \quad (15)$$

บทความจำนวนมากเสนอให้ใช้การแจกแจงแบบปกติเป็นการแจกแจงในส่วน slab [2,3,

5] สำหรับการแจกแจงก่อนของ γ หากเราสันนิษฐานว่าตัวแปรอิสระแต่ละตัวมีความน่าจะเป็นในการเข้ามาอยู่ในตัวแบบเท่า ๆ กัน และเป็นอิสระต่อกัน เราอาจให้การแจกแจกแจงก่อนของ γ แต่ละตัวมีการแจกแจงแบบแบร์นูลลี ซึ่งตัวแบบเบส์เป็นลำดับชั้น สามารถเขียนได้ดัง (16)

$$
\begin{aligned}
&Y|\boldsymbol{X},\beta,\sigma^2 \sim N(\boldsymbol{X}\beta,\sigma^2\boldsymbol{I}_n)\\
&P(\beta|\gamma,\tau^2) = \textstyle\prod_{j=1}^{p}\{(1-\gamma_j)\delta(\beta_j) + \\
&\qquad \gamma_j N(0,\tau^2\boldsymbol{I}_n)\} \qquad (16)\\
&\gamma_j|\omega \sim \mathrm{Bernoulli}(\omega)\\
&\omega \sim \mathrm{Beta}(\mathrm{a},\mathrm{b})\\
&P(\sigma^2) \propto 1/\sigma^2
\end{aligned}
$$

เนื่องจากการแจกแจงก่อนของพารามิเตอร์ β แบบ spike and slab ไม่ใช่การแจกแจงที่ต่อเนื่องดังเช่นวิธีเบส์สามวิธีข้างต้น แต่เป็นการแจกแจงแบบผสมระหว่างมวลที่ศูนย์และการแจกแจงแบบต่อเนื่อง ด้วยเหตุนี้การแจกแจงภายหลังที่ได้จะเป็นการแจกแจงแบบผสมระหว่างมวลที่ศูนย์และการแจกแจงแบบต่อเนื่องเช่นกัน ดังนั้นหากเราเลือกใช้ค่ามัธยฐานของการแจกแจงภายหลังในการประมาณค่าพารามิเตอร์ β แล้วจะได้ว่าค่าประมาณสัมประสิทธิ์บางตัวมีค่าเท่ากับศูนย์ ซึ่งคือสมบัติในการเลือกตัวแปรเข้าไปยังตัวแบบนั่นเอง นอกจากนี้

จากตัวแบบลำดับชั้น (16) เราสามารถใช้การจำลองข้อมูลจากการแจกแจงภายหลังโดยวิธีการเลือกตัวอย่างแบบกิบส์ได้ สำหรับพารามิเตอร์ τ^2 เราอาจจะกำหนดค่า τ^2 ที่เหมาะสมให้เป็นค่าที่ใหญ่เพื่อสามารถจัดการกับค่าสัมประสิทธิ์ที่มีขนาดใหญ่ได้ หรืออาจใช้วิธีเบส์เชิงประจักษ์ หรือวิธีเบส์แบบลำดับชั้น โดยให้การแจกแจงก่อนกับพารามิเตอร์ τ^2 เป็นการแจกแจงแบบแกมมาผกผัน

4. ข้อควรคำนึงในการเลือกใช้วิธีวิเคราะห์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูง

4.1 ข้อดี ข้อเสีย และความเหมาะสมในการใช้งานแต่ละวิธี

แม้ว่าทั้งวิธี penalized regression และวิธีเบส์จะสามารถแก้ไขปัญหาของวิธีกำลังสองน้อยสุดในการวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูงได้เหมือนกัน อย่างไรก็ตาม วิธีแต่ละวิธีก็มีข้อดีข้อเสียและบริบทการใช้งานที่แตกต่างกัน ซึ่งสรุปได้ดังตารางที่ 1

หนึ่งในข้อที่ควรพิจารณาที่สำคัญในการเลือกใช้เครื่องมือทางสถิติ คือ ระยะเวลาในการคำนวณหรือทำงาน ซึ่งเมื่อเราคำนึงถึงจุดนี้ จากตารางที่ 1 จะเห็นว่าวิธีเบส์มีความเหมาะสมกับข้อมูลที่มีมิติไม่สูงมาก ในขณะที่วิธี penalized regression สามารถจัดการกับข้อมูลขนาดใหญ่ได้ดีกว่า อย่างไรก็ดี หากมีข้อมูลที่มีมิติไม่สูงจนเกินไป การใช้วิธีเบส์จะเป็นประโยชน์อย่างมาก เนื่องจากวิธีเบส์ไม่ได้ให้เฉพาะค่าประมาณของ β แต่ให้การแจกแจงภายหลังของทุกพารามิเตอร์ที่เกี่ยวข้องในตัวแบบ ทำให้เห็นภาพได้ชัดเจนและนำไปสู่การวิเคราะห์เชิงอนุมานที่หลากหลายตามความต้องการของผู้วิจัย เช่น สามารถเลือกใช้ค่าประมาณแบบจุดที่หลากหลายได้ ถึงกระนั้นวิธีเบส์ยังเป็นวิธีที่ใช้กันไม่แพร่หลายในประเทศไทย ซึ่งอาจมีสาเหตุหลัก คือ ข้อจำกัดทางความรู้ เนื่องจากวิธีเบส์ไม่มีสอนในหลักสูตรการศึกษาทั่วไป และหนังสือตำราภาษาไทยทางด้านนี้ยังมีไม่มากพอ อย่างไรก็ตามในต่างประเทศ มีการประยุกต์ใช้วิธีเบส์ในหลากหลายแขนง และเป็นที่ยอมรับในวงกว้าง

4.2 ความพร้อมใช้งานในโปรแกรมสำเร็จรูปทางสถิติ

ปัจจุบันคอมพิวเตอร์เข้ามามีบทบาทสำคัญอย่างมากในการดำเนินชีวิต การเก็บข้อมูลในคอมพิวเตอร์สามารถทำได้ง่าย รวดเร็ว และค่าใช้จ่ายไม่สูง ซึ่งความก้าวหน้าทางเทคโนโลยีคอมพิวเตอร์นี้ยัง

ตารางที่ 1 เปรียบเทียบข้อดี ข้อเสีย และความเหมาะสมในการใช้งานของแต่ละวิธี

วิธี	ข้อดี	ข้อเสีย	ความเหมาะสมในการใช้งาน
1. Penalized regression	- มีการประยุกต์ใช้งานอย่างแพร่หลาย และพร้อมใช้งานในโปรแกรมสำเร็จรูปทางสถิติต่าง ๆ - ใช้เวลาทำงานสั้นกว่าการวิเคราะห์แบบเบส์	- ให้ค่าประมาณแบบจุดของ β เพียงอย่างเดียว สำหรับค่าประมาณแบบช่วงของ β จะหาได้เฉพาะตัวที่ไม่เท่ากับศูนย์	- ใช้ได้ตั้งแต่ข้อมูลที่มีมิติไม่สูงจนถึงมิติสูงมาก
1.2 Ridge	- มีความเสถียรและแก้ไขปัญหากรณีตัวแปรอิสระมีความสัมพันธ์เชิงเส้นสูงได้	- ไม่มีสมบัติในการเลือกตัวแปรเข้าตัวแบบ	- ข้อมูลที่มีสัมประสิทธิ์ขนาดเล็กที่ไม่เท่ากับศูนย์จำนวนมาก
1.3 Lasso	- สามารถเลือกตัวแปรเข้าสู่ตัวแบบและประมาณค่า β ในคราวเดียวกัน	- สามารถเลือกตัวแปรเข้าได้มากที่สุดจำนวน n ตัว - ในกรณีที่ตัวแปรอิสระมีความสัมพันธ์เชิงเส้นสูง วิธี Lasso จะเลือกตัวแปรเข้าเพียงตัวแปรเดียวจากกลุ่มตัวแปรอิสระที่มีความสัมพันธ์เชิงเส้นสูงโดยไม่สนใจว่าจะเป็นตัวแปรใดในกลุ่ม	- ข้อมูลที่มีสัมประสิทธิ์ขนาดกลางที่ไม่เท่ากับศูนย์จำนวนน้อยถึงปานกลาง
1.4 Elastic net	- สามารถเลือกตัวแปรเข้าสู่ตัวแบบและประมาณค่า β ในคราวเดียวกัน - สามารถแก้ไขข้อเสียทั้งสองข้อของวิธี Lasso ได้	- ใช้เวลาในทำงานนานกว่าวิธี Ridge และ Lasso เนื่องจากมีพารามิเตอร์ใน penalty function 2 ตัว	- ข้อมูลที่มีสัมประสิทธิ์ขนาดเล็กถึงปานกลางที่ไม่เท่ากับศูนย์จำนวนปานกลาง หรือข้อมูลที่มีมิติสูงมากหรือกรณีที่ตัวแปรอิสระมีความสัมพันธ์เชิงเส้นสูง
2. วิธีเบส์	- ให้การแจกแจงภายหลังของทุกพารามิเตอร์ที่เกี่ยวข้องในตัวแบบ ทำให้มีความคล่องตัวในการวิเคราะห์มากกว่า - สามารถหาค่าประมาณแบบจุดและแบบช่วงของทุกพารามิเตอร์ได้	- เมื่อข้อมูลมีมิติสูงมาก การตรวจสอบการลู่เข้าในลูกโซ่มาร์คอฟทำได้ยากและใช้เวลานาน - ส่วนใหญ่ผู้วิเคราะห์จะต้องเขียนโปรแกรมในการทำงานเอง	- ข้อมูลที่มีมิติไม่สูงมาก
2.1 Bayesian Ridge	- เหมือนวิธี Ridge		
2.2 Bayesian Lasso	- เหมือนวิธี Lasso แต่ไม่มีสมบัติในการเลือกตัวแปรเข้าตัวแบบ		
2.3 Bayesian elastic net	- เหมือนวิธี elastic net แต่ไม่มีสมบัติในการเลือกตัวแปรเข้าตัวแบบ		
2.4 การแจกแจงก่อนแบบ spike and slab	- มีสมบัติในการเลือกตัวแปรเข้าตัวแบบ - การแจกแจงก่อนมีความหมายสมเหตุสมผล และยืดหยุ่นต่อการเปลี่ยนแปลง	- การลู่เข้าในลูกโซ่มาร์คอฟจะช้ากว่าวิธีเบส์สามวิธีข้างต้น เนื่องจากการแจกแจงก่อนเป็นการแจกแจงแบบไม่ต่อเนื่อง	- ข้อมูลที่มีสัมประสิทธิ์ที่เท่ากับศูนย์จำนวนมาก

ส่งผลไปยังศาสตร์ด้านสถิติอีกด้วย ดังจะเห็นได้จากการที่นักสถิติคิดค้นเครื่องมือทางสถิติที่มีความซับซ้อนมากขึ้น ซึ่งจะต้องอาศัยคอมพิวเตอร์ในการคำนวณ ไม่สามารถคำนวณได้ด้วยมืออีกต่อไป วิธีที่ได้นำเสนอทั้งหมดในบทความนี้ก็เช่นกัน จำเป็นต้องอาศัยโปรแกรมคอมพิวเตอร์ในการทำงาน ดังนั้นความพร้อมใช้งานในโปรแกรมสำเร็จรูปทางสถิติก็เป็นปัจจัยหลักอีกปัจจัยหนึ่งที่ผู้วิเคราะห์ควรคำนึงถึงในการเลือกใช้เครื่องมือ ในบทความนี้จะขอกล่าวถึงโปรแกรมคอมพิวเตอร์ทางสถิติ 3 โปรแกรม ที่ใช้กันอย่างแพร่หลายทั้งในวงการการศึกษาและวงการอุตสาหกรรมและพาณิชย์ ได้แก่ โปรแกรม SPSS โปรแกรม SAS และโปรแกรม R

ตารางที่ 2 ความพร้อมใช้งานโปรแกรมสำเร็จรูปทางสถิติ

วิธี	โปรแกรมทางสถิติ		
	SPSS	SAS	R
1. penalized regression			
1.1 ridge	●*	●	●*
1.2 lasso	●*	●	●*
1.3 elastic net	●*	●	●*
2. วิธีเบส์			
2.1 Bayesian ridge	○	⦿	●*
2.2 Bayesian Lasso	○	⦿	●*
2.3 Bayesian elastic net	○	⦿	⦿
2.4 การแจกแจงก่อนแบบ spike and slab	○	⦿	⦿

● มีฟังก์ชันพร้อมใช้งาน (*หมายถึงเป็นฟังก์ชันเพิ่มเติมจากโปรแกรมมาตรฐาน)

⦿ มีฟังก์ชันสนับสนุนแต่จำเป็นต้องเขียนโปรแกรมเพิ่มเพื่อใช้งาน

○ ไม่สามารถใช้งานได้

จากตารางที่ 2 จะเห็นว่าวิธี penalized regression ทั้งสามวิธีสามารถทำงานได้โดยใช้ฟังก์ชันสำเร็จรูปในโปรแกรม SPSS, SAS และ R โดยที่หากผู้วิเคราะห์ต้องการใช้โปรแกรม SPSS จะต้องใช้ฟังก์ชันใน module เพิ่มเติมที่ชื่อว่า SPSS categories สำหรับในโปรแกรม SAS ผู้วิเคราะห์สามารถใช้คำสั่งมาตรฐานได้ทันที โดยใช้คำสั่ง PROC REG สำหรับวิธี ridge regression และ PROC GLMSELECT สำหรับวิธี Lasso และ elastic net แต่หากเลือกใช้โปรแกรม R ผู้วิเคราะห์จะต้องดาวน์โหลดและติดตั้งแพ็คเกจ MASS สำหรับวิธี ridge regression หรือสามารถใช้แพ็คเกจ penalized ในการวิเคราะห์ทั้งสามวิธีได้

สำหรับวิธีเบส์ โปรแกรม SPSS ไม่สามารถทำการวิเคราะห์ได้ เนื่องจาก โปรแกรมไม่มีฟังก์ชันสำเร็จรูปและมีข้อจำกัดในการรองรับการเขียนโปรแกรมเพิ่มเติม อย่างไรก็ตาม ผู้วิเคราะห์สามารถใช้

โปรแกรม SAS และ R ในการวิเคราะห์ข้อมูลแบบเบส์ได้ โดยที่ผู้วิเคราะห์จำเป็นต้องเขียนโปรแกรมเพิ่มเติมเองโดยใช้คำสั่ง PROC MCMC ในการสุ่มข้อมูลแบบกิบส์ในโปรแกรม SAS สำหรับโปรแกรม R ผู้วิเคราะห์สามารถเรียกใช้ฟังก์ชันสำเร็จรูปในการวิเคราะห์ข้อมูลด้วยวิธี Bayesian ridge และ Bayesian Lasso ได้ผ่านแพ็คเกจ BLR ส่วนวิธี Bayesian elastic net และการใช้การแจกแจงก่อนแบบ spike and slab prior ผู้วิเคราะห์จำเป็นที่จะต้องเขียนโปรแกรม โดย R มีฟังก์ชันสำเร็จรูปหลากหลายที่เอื้อต่อการเขียนโปรแกรมเพิ่มเติมได้โดยง่าย

นอกจากความพร้อมใช้งานวิธี penalized regression ในโปรแกรม SPSS, SAS และ R ดังตารางที่ 2 ข้างต้นแล้ว ผู้วิเคราะห์ควรทราบถึงจุดเด่นและจุดด้อยของแต่ละโปรแกรม เพื่อสามารถเลือกใช้โปรแกรมให้มีความเหมาะสมและมีประสิทธิภาพสูงสุดในการทำงาน เริ่มจากโปรแกรม SPSS ที่ใช้กันอย่างแพร่หลายในประเทศไทย จุดเด่นของโปรแกรมนี้คือการใช้งานที่ง่าย เนื่องจากผู้ใช้ไม่จำเป็นต้องจำคำสั่ง แต่เลือกใช้คำสั่งผ่านเมนูที่ต้องการได้ทันที สำหรับจุดด้อยคือข้อจำกัดในการเขียนโปรแกรมเพิ่มเติมและการรองรับข้อมูลขนาดใหญ่ โปรแกรม SAS มีข้อดีคือสามารถรองรับการเขียนโปรแกรมเพิ่มเติมและสามารถจัดการกับข้อมูลขนาดใหญ่ที่ซับซ้อนได้ดี ผลการวิเคราะห์มีความน่าเชื่อถือสูง ส่วนข้อเสียคือผู้ใช้จำเป็นต้องมีการฝึกฝนและเข้าใจในการเขียนคำสั่งการทำงาน นอกจากนี้ โปรแกรม SAS อาจไม่มีฟังก์ชันการทำงานของวิธีการวิเคราะห์ใหม่ ๆ ในขณะที่โปรแกรม R มีแพ็คเกจหรือฟังก์ชันการวิเคราะห์ข้อมูลที่ทันสมัยตลอดเวลา เนื่องจาก R เป็นซอฟต์แวร์โอเพนซอร์ซ (open source software) ยิ่งไปกว่านั้น โปรแกรม R และแพ็คเกจต่าง ๆ สามารถดาวน์โหลดได้ฟรี ในขณะที่โปรแกรม SPSS และ SAS ผู้ใช้จำเป็นต้องซื้อโปรแกรมเพื่อติดตั้ง อย่างไรก็ตาม ข้อเสียของซอฟต์แวร์โอเพนซอร์ซที่ทุกคนสามารถอัปโหลดแพ็คเกจของตัวเองขึ้นไปไว้บนเซิร์ฟเวอร์เพื่อให้คนอื่นสามารถใช้งานได้ ก็คือความถูกต้องและความน่าเชื่อถือของแพ็คเกจดังกล่าว ดังนั้นในการเลือกใช้แพ็คเกจ ผู้วิเคราะห์จำเป็นต้องตรวจสอบเอกสารประกอบของแพ็คเกจนั้น ๆ ว่าผู้เขียนเป็นใคร รวมถึงแหล่งอ้างอิงของวิธีการว่ามีความน่าเชื่อถือมากน้อยเพียงใด ส่วนข้อเด่นและข้อด้อยของโปรแกรม R ในจุดอื่น ๆ จะมีความคล้ายคลึงกับโปรแกรม SAS แต่โปรแกรม R จะใช้เวลาในการทำงานมากกว่าโปรแกรม SAS สำหรับข้อมูลที่มีขนาดใหญ่มาก

4.3 จุดมุ่งหมายในการวิเคราะห์การถดถอย

จุดมุ่งหมายหลักในการวิเคราะห์การถดถอยมีอยู่ 2 ประการด้วยกัน คือ (1) เพื่อใช้ตัวแบบในการพยากรณ์ เช่น การพยากรณ์ยอดขายจากข้อมูลในอดีตและตัวบ่งชี้ทางเศรษฐกิจต่าง ๆ และ (2) เพื่อศึกษาว่าตัวแปรอิสระใดบ้างที่มีความสัมพันธ์กับตัวแปรตาม เช่น ในการสำรวจตลาด เราอาจต้องการทราบว่าปัจจัยอะไรบ้างที่ส่งผลกับพฤติกรรมการซื้อสินค้าของผู้บริโภค การหาตัวแบบที่จะบรรลุเป้าหมายทั้งสองข้ออาจทำไม่ได้ในคราวเดียวกัน ตัวอย่าง เช่น ในการที่จะบรรลุจุดมุ่งหมายแรก เราเพียงต้องการตัวแบบที่สามารถพยากรณ์ข้อมูลในอนาคตได้แม่นยำที่สุด โดยไม่คำนึงถึงว่ามีตัวแปรอิสระใดบ้างอยู่ในตัวแบบหรืออาจมีทั้งตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามมากและน้อยอยู่ในตัวแบบ ในขณะที่หากวัตถุประสงค์หลักในการวิเคราะห์การถดถอยของเราคือจุดประสงค์ประการที่สอง ตัวแบบที่เหมาะสมควรเป็นตัวแบบที่ประกอบไปด้วยตัวแปรอิสระที่สำคัญหรือมีความสัมพันธ์สูงกับตัวแปรตามเท่านั้น ดังนั้น หากเราทราบจุดมุ่งหมายที่แน่ชัด เราก็จะสามารถเลือกวิธีที่เหมาะสมในการวิเคราะห์การถดถอยได้

อนึ่ง ข้อแนะนำในการเลือกใช้วิธีที่เหมาะสมในทางปฏิบัติ คือ หากจุดมุ่งหมายหลักคือการพยากรณ์ ผู้วิเคราะห์อาจเลือกใช้วิธี penalized likelihood กับข้อมูลที่มีขนาดใหญ่หรือข้อมูลที่มีมิติสูงมาก เนื่องจากวิธีเบส์จะใช้เวลาทำงานนานในการเลือกตัวอย่างแบบกิบส์ในขณะที่ให้ค่าประมาณแบบจุดที่ใกล้เคียงกับวิธี penalized regression แต่หากข้อมูลมีขนาดหรือมีจำนวนตัวแปรอิสระไม่สูงมาก ผู้วิเคราะห์อาจพิจารณาวิธีเบส์ควบคู่ไปด้วย นอกจากนี้ แม้ว่าเราจะทราบว่าวิธีการแต่ละวิธีมีความเหมาะสมกับข้อมูลที่มีจำนวน β ที่ไม่เท่ากับศูนย์และขนาดของ β ที่ไม่เท่ากับศูนย์แตกต่างกันดังแสดงไว้ในตารางที่ 1 แต่ในความเป็นจริงเราไม่สามารถทราบได้ว่า β ที่แท้จริงมีค่าเป็นเช่นไร ดังนั้นในทางปฏิบัติ ผู้วิเคราะห์อาจแบ่งข้อมูลออกเป็นสองส่วน โดยส่วนแรกเรียกว่าชุดข้อมูลฝึกฝน (training set) และอีกส่วนเรียกว่าชุดข้อมูลทดสอบ (test set) จากนั้นผู้วิเคราะห์ใช้วิธีแต่ละวิธีในการหาค่าประมาณแบบจุดของ β โดยใช้ชุดข้อมูลฝึกฝนเท่านั้น และนำค่าประมาณที่ได้มาคำนวณค่าความคลาดเคลื่อนที่เกิดขึ้นโดยคำนวณจากชุดข้อมูลทดสอบ วิธีที่เหมาะสมคือวิธีที่ให้ค่าคลาดเคลื่อนที่น้อยที่สุด

แต่หากจุดมุ่งหมายหลักคือการศึกษาความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม ผู้วิเคราะห์ควรเลือกใช้วิธี Lasso วิธี elastic net หรือวิธีเบส์ที่ใช้การแจกแจงก่อนแบบ spike and slab เนื่องจากวิธีอื่นนอกเหนือจากนี้ไม่มีความสามารถในการคัดเลือกตัวแปรเข้าตัวแบบโดยตรง

5. สรุป

การวิเคราะห์การถดถอยเชิงเส้นเป็นเทคนิคที่ใช้กันอย่างแพร่หลายในการวิเคราะห์ข้อมูลในศาสตร์ต่าง ๆ ไม่ว่าจะเป็นด้านธุรกิจ ด้านวิทยาศาสตร์ ด้านการแพทย์ ฯลฯ ด้วยข้อจำกัดของวิธีกำลังสองน้อยสุดในการวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูงซึ่งพบได้ทั่วไปในปัจจุบัน บทความนี้ได้นำเสนอแนวทางแก้ไขข้อจำกัดดังกล่าวด้วยวิธี penalized regression และวิธีเบส์ แม้ว่าทั้งสองวิธีจะสามารถนำไปใช้ในการวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูงได้เหมือนกัน แต่มีข้อที่ควรคำนึงถึงในการเลือกใช้วิธีที่เหมาะสมในการวิเคราะห์ ได้แก่ ข้อดีข้อเสียของแต่ละวิธี ความพร้อมใช้งานในโปรแกรมสำเร็จรูปทางสถิติ และจุดมุ่งหมายในการวิเคราะห์การถดถอยว่าต้องการสร้างตัวแบบเพื่อพยากรณ์หรือศึกษาถึงความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม ดังนั้นก่อนการเลือกใช้เครื่องมือใดก็ตาม ผู้วิเคราะห์ควรมีความรู้ความเข้าใจถึงข้อมูลที่ตนเองมีอย่างถ่องแท้ และศึกษาความเป็นไปได้ในแต่ละวิธีว่าตอบโจทย์วัตถุประสงค์ที่เราต้องการหรือไม่ เพื่อให้ได้ผลลัพธ์ที่เป็นประโยชน์ในการทำงานของท่านต่อไป

6. รายการอ้างอิง

[1] Andrews, D.R. and Mallows, C.L., 1974, Scale mixtures of normal distributions, J. Royal Stat. Soc. Ser. B 36: 99-102.

[2] George, E.I. and McCulloch, R.E., 1993, Variable selection via Gibb sampling, J. Am. Stat. Assoc. 88: 881-889.

[3] Ishwaran, H. and Rao, J.S., 2005, Spike and slab variable selection: Frequentist and Bayesian strategies, Ann. Stat. 33: 730-733.

[4] Li, Q. and Lin, N., 2010, The Bayesian elastic net, Bayesian Anal. 5: 151-170.

[5] Malsiner-Walli, G. and Wagner, H., 2011, Comparing spike and slab priors for

Bayesian variable selection, Aus. J. Stat. 40: 241-264.

[6] Park, T. and Casella, G., 2008, The Bayesian Lasso, J. Am. Stat. Assoc. 103: 681-686.

[7] Tibshirani, R., 1996, Regression shrinkage and selection via the Lasso, J. Royal Stat. Soc. Ser. B 58: 267-288.

Zou, H., 2006, The adaptive Lasso and its oracle properties, J. Am. Stat. Assoc. 101: 1418-1429.

Zou, H. and Hastie, T., 2005, Regularization and variable selection via the elastic net, J. Royal Stat. Soc. Ser. B 67: 301-320.