บทที่ 7

การเลือกสมการถดถอยที่เหมาะสม

การวิเคราะห์การถดถอยในบทที่แล้ว ๆ มาจะสมมติให้ตัวแปรอิสระที่อยู่ในโมเดลมีอิทธิพลต่อตัวแปรตาม รูปแบบของโมเดลถูกต้อง และมีลักษณะที่สอดคล้องกับข้อตกลงเบื้องต้นของการวิเคราะห์ บางครั้งพบว่าความรู้ทางทฤษฎีและประสบการณ์ในอดีตจะช่วยในการตัดสินใจเลือกตัวแปรอิสระเข้าในโมเดลได้เช่นกัน ซึ่งการค้นหาสับเซ็ตของตัวแปรอิสระที่เหมาะสมสำหรับโมเดลจะเรียกว่า *การเลือกตัวแปร (Variable selection)* ในบทนี้จะกล่าวถึงเกณฑ์ที่ใช้ในการพิจารณาคัดเลือกตัวแปร รวมทั้งวิธีในการคัดเลือกตัวแปร แต่อย่างไรก็ตามไม่มีวิธีคัดเลือกตัวแปรวิธีใดที่สามารถรับประกันได้ว่าให้สมการถดถอยที่ดีที่สุด (Best regression equation) สำหรับข้อมูลที่พิจารณา นอกจากนี้วิธีในการคัดเลือกตัวแปรต้องใช้คอมพิวเตอร์เป็นส่วนใหญ่ ซึ่งนักวิเคราะห์ไม่ควรเชื่อผลที่ได้จากคอมพิวเตอร์เพียงอย่างเดียว ควรคำนึงถึงประสบการณ์และความสมเหตุสมผลประกอบการตัดสินใจด้วย ดังนั้นวิธีการเลือกตัวแปรควรที่จะพิจารณาว่าเป็นวิธีการสำรวจโครงสร้างของข้อมูลเบื้องต้นมากกว่าที่จะเป็นผลสรุปในขั้นสุดท้าย

7.1 ปัญหาในการเลือกโมเดลถดถอย

การสร้างโมเดลถดถอยโดยทั่วไปพบว่านักวิเคราะห์ต้องการให้โมเดลประกอบด้วยตัวแปรอิสระเพียงไม่กี่ตัว แต่สามารถอธิบายอิทธิพลที่มีต่อค่าพยากรณ์ Y ได้อย่างเพียงพอ ซึ่งถ้าจำนวนตัวแปรอิสระเพิ่มขึ้น ความแปรปรวนของค่าพยากรณ์จะสูงขึ้นตาม อีกทั้งยังทำให้ค่าใช้จ่ายในการเก็บรวบรวมข้อมูลเพิ่มขึ้นอีกด้วย นอกจากนี้วิธีในการคัดเลือกตัวแปรเข้าในโมเดลมีหลายวิธี และบ่อยครั้งที่วิธีในการคัดเลือกตัวแปรเหล่านี้ใช้สับเซ็ตของตัวแปรอิสระที่แตกต่างกันในการสร้างสมการถดถอยที่ดีที่สุดของแต่ละวิธี อีกทั้งไม่มีนิยามที่ชัดเจนเกี่ยวกับสมการถดถอยที่ดีที่สุดอีกด้วย ซึ่งปัญหาในการเลือกตัวแปรมักจะนำเสนอในเชิงอุดมคติ โดยสมมติว่าทราบรูปแบบฟังก์ชันของตัวแปรอิสระที่ถูกต้อง เช่น $\ln X$, X^2 เป็นต้น และไม่มีค่าสังเกตที่ผิดปกติหรือค่าสังเกตที่มีอิทธิพลอยู่ในข้อมูล ซึ่งในทางปฏิบัติมักไม่ค่อยเกิดขึ้น การวิเคราะห์ความคลาดเคลื่อนที่ได้อธิบายไว้แล้วในบทที่ 4 จะช่วยในการค้นหาฟังก์ชันของตัวแปรอิสระที่เหมาะสม หรืออาจชี้ถึงตัวแปรอิสระตัวใหม่ที่ควรเพิ่มเข้าไป

ในโมเดล และข้อบกพร่องของข้อมูลเกี่ยวกับค่าสังเกตที่ผิดปกติและค่าสังเกตที่มีอิทธิพล ดังนั้นจะเห็นได้ว่าการตรวจสอบความเหมาะสมของโมเดลมีความสัมพันธ์อย่างใกล้ชิดกับปัญหาในการเลือกตัวแปร และควรพิจารณาควบคู่กันไป กล่าวคือ สร้างสมการถดถอยโดยใช้วิธีในการคัดเลือกตัวแปรเข้าในโมเดล หลังจากนั้นตรวจสอบความเหมาะสมของโมเดลที่ได้ในขั้นสุดท้ายเพื่อให้ได้สมการที่เหมาะสม

การตัดตัวแปรอิสระบางตัวออกจากโมเดล จะช่วยเพิ่มความแม่นยำในการพยากรณ์และการประมาณพารามิเตอร์ของตัวแปรอิสระที่ยังคงอยู่เหลืออยู่ในโมเดลได้ ถึงแม้ว่าตัวแปรที่ถูกตัดทิ้งบางตัวจะยังคงมีอิทธิพลต่อตัวแปรตามก็ตาม นอกจากนี้การตัดตัวแปรอิสระออกไปจะทำให้ตัวประมาณค่าสัมประสิทธิ์ถดถอยของตัวแปรอิสระที่ยังคงอยู่เหลืออยู่ในโมเดลเป็นตัวประมาณที่เอนเอียง แต่อย่างไรก็ตามถ้าตัวแปรที่ถูกตัดออกไปมีอิทธิพลน้อยมาก จะทำให้ MSE ของตัวประมาณที่เอนเอียงนั้นมีค่าน้อยกว่าความแปรปรวนของตัวประมาณที่ไม่เอนเอียง ซึ่งแสดงว่าขนาดของความเอนเอียงที่เกิดขึ้นมีค่าน้อยกว่าความผันแปรที่ลดลง ขณะเดียวกันการเก็บตัวแปรที่ไม่มีอิทธิพลไว้ในโมเดล นั่นคือ ตัวแปรที่มีค่าสัมประสิทธิ์ถดถอยเข้าใกล้ศูนย์หรือมีค่าน้อยกว่าค่าความคลาดเคลื่อนมาตรฐานของการประมาณที่ได้จาก Full model จะทำให้ความแปรปรวนของการประมาณค่าสัมประสิทธิ์ถดถอยและค่าพยากรณ์สูงขึ้น

7.2 เกณฑ์ที่ใช้ในการเลือกโมเดลถดถอย

การเลือกตัวแปรอิสระที่ควรอยู่ในโมเดลสามารถแบ่งได้เป็น 2 ขั้นตอนคือ การสร้างสับเซ็ตของสมการถดถอยที่เป็นไปได้ และการพิจารณาว่าสับเซ็ตของสมการใดดีกว่าสับเซ็ตอื่น ซึ่งเกณฑ์ที่ใช้ในการประเมินและเปรียบเทียบสับเซ็ตของสมการถดถอยมีดังนี้

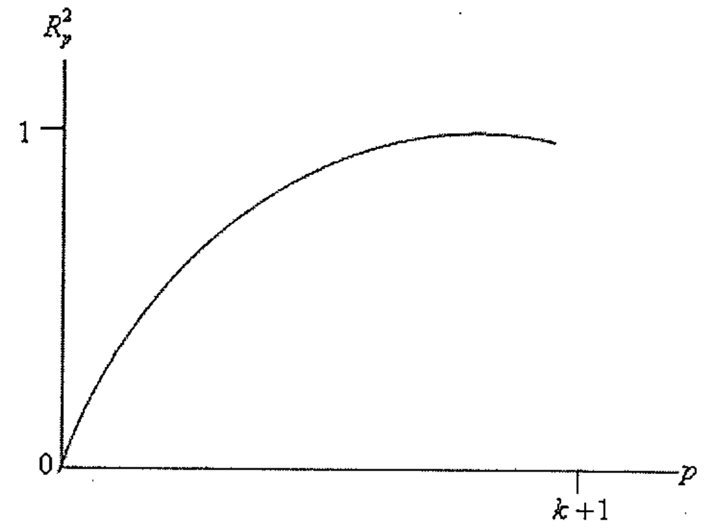
7.2.1 ค่าสัมประสิทธิ์ตัวกำหนดแบบพหุ

เป็นที่ทราบแล้วว่าค่าสัมประสิทธิ์ตัวกำหนดเชิงพหุ หรือ R^2 เป็นค่าที่ใช้กันอย่างแพร่หลายในการวัดความเหมาะสมของโมเดล ให้ R_p^2 แทน ค่าสัมประสิทธิ์ตัวกำหนดเชิงพหุสำหรับสับเซ็ตของโมเดลถดถอยที่มี p เทอม (ตัวแปรอิสระ $p-1$ ตัว และระยะตัดแกน β_0) นั่นคือ

$$R_p^2 = \frac{SSR_p}{SST} = 1 - \frac{SSE_p}{SST} \tag{7.1}$$

เมื่อ SSR_p และ SSE_p แทน ผลรวมกำลังสองเนื่องมาจากสมการถดถอยและความคลาดเคลื่อนของโมเดลที่มี p เทอม ตามลำดับ

ซึ่งในแต่ละสับเซ็ตของโมเดลขนาด p จะมี R_p^2 ทั้งหมด $\binom{k}{p-1}$ ตัว เมื่อ k แทน จำนวนตัวแปรอิสระ จะเห็นได้ว่า R_p^2 มีค่าสูงขึ้นเมื่อ p เพิ่มขึ้น และมีค่าสูงสุดเมื่อ $p = k + 1$ ดังนั้นการเพิ่มตัวแปรอิสระ



รูปที่ 7.1: กราฟระหว่าง R_p^2 และ p

เข้าไปในโมเดลจะหยุดกระทำก็ต่อเมื่อตัวแปรอิสระที่เพิ่มเข้าไปนั้นไม่มีประโยชน์หรือมีประโยชน์น้อยมาก โดยทำให้ค่า R_p^2 เพิ่มขึ้นเพียงเล็กน้อยเท่านั้น รูปที่ 7.1 แสดงลักษณะของกราฟระหว่างค่าสูงสุดของ R_p^2 ในแต่ละสับเซ็ตของโมเดลขนาด p กับจำนวนพารามิเตอร์ p โดยจุดที่กราฟเริ่มหักโค้งจะเป็นจุดที่กำหนดจำนวนตัวแปรอิสระที่ควรอยู่ในโมเดลสุดท้าย ซึ่งขึ้นอยู่กับการตัดสินใจของนักวิเคราะห์แต่ละคน

7.2.2 ค่าสัมประสิทธิ์ตัวกำหนดแบบพหุที่มีการปรับค่าแล้ว

เพื่อที่จะหลีกเลี่ยงความยุ่งยากในการแปลผลค่า R^2 บางครั้งจึงนิยมใช้ค่าสัมประสิทธิ์ตัวกำหนดแบบพหุที่มีการปรับค่าแล้ว หรือ R^2_{adj} เป็นเกณฑ์ในการเปรียบเทียบสับเซ็ตของโมเดล โดยที่

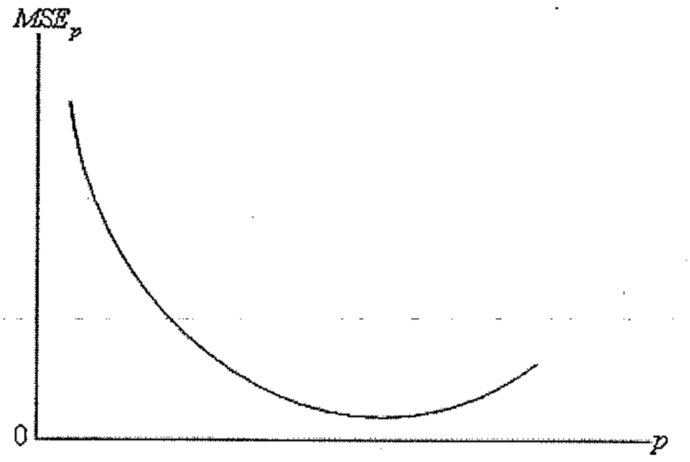
$$R^2_{adj\,(p)} = 1 - \frac{n-1}{n-p}(1 - R_p^2) \tag{7.2}$$

เมื่อ $R^2_{adj\,(p)}$ แทน ค่าสัมประสิทธิ์ตัวกำหนดแบบพหุที่มีการปรับค่าแล้วสำหรับสับเซ็ตของโมเดลถดถอยที่มี p เทอม โดยที่ $R^2_{adj\,(p)}$ ไม่จำเป็นต้องมีค่ามากขึ้นทุกครั้งที่เพิ่มตัวแปรอิสระเข้าไปในโมเดล ดังนั้นเกณฑ์ในการเลือกสับเซ็ตของโมเดลที่เหมาะสมทำได้โดยเลือกโมเดลที่มีค่า $R^2_{adj\,(p)}$ สูงสุดนั่นเอง

7.2.3 ความคลาดเคลื่อนกำลังสองเฉลี่ย

ความคลาดเคลื่อนกำลังสองเฉลี่ยสำหรับสับเซ็ตของโมเดลที่มี p เทอม แทนด้วย MSE_p มีสูตรดังนี้

$$MSE_p = \frac{SSE_p}{n-p} \tag{7.3}$$



รูปที่ 7.2: กราฟระหว่าง MSE_p และ p

เมื่อ p เพิ่มขึ้น พบว่า SSE_p มีค่าลดลงเสมอ รูปที่ 7.2 แสดงลักษณะของ MSE_p เมื่อ p เพิ่มขึ้น ซึ่งจะเห็นได้ว่า MSE_p มีค่าลดลงในช่วงแรก แล้วคงที่ และเพิ่มขึ้นในช่วงท้าย ๆ อาจเนื่องมาจากตัวแปรอิสระที่เพิ่มเข้าไปในโมเดลไม่เพียงพอที่จะชดเชยจำนวนองศาความเป็นอิสระของตัวหารในสมการที่ (7.3) ที่สูญเสียไป ดังนั้นเกณฑ์ในการเลือกจำนวนตัวแปรอิสระเมื่อพิจารณาจาก MSE_p ทำได้ดังนี้

- เลือก p ที่ทำให้ MSE_p มีค่าต่ำสุด
- เลือก p ที่ทำให้ MSE_p มีค่าใกล้เคียงกับ MSE ของ Full model
- เลือก p ที่อยู่ใกล้จุดที่ทำให้ MSE_p เพิ่มขึ้น

และสับเซ็ตของโมเดลถดถอยที่ทำให้ MSE_p มีค่าต่ำสุด จะทำให้ $R^2_{adj\,(p)}$ มีค่าสูงสุดด้วย โดยจะเห็นได้จาก

$$
\begin{aligned}
R^2_{adj\,(p)} &= 1 - \frac{n-1}{n-p}\left(1 - R_p^2\right) \\
&= 1 - \frac{n-1}{n-p} \cdot \frac{SSE_p}{SST} \\
&= 1 - \frac{n-1}{SST} \cdot \frac{SSE_p}{n-p} \\
&= 1 - \frac{n-1}{SST} \cdot MSE_p
\end{aligned}
$$

ดังนั้นการเลือกจำนวนตัวแปรอิสระเข้าในโมเดลโดยใช้เกณฑ์ที่ทำให้ MSE_p มีค่าต่ำสุด เทียบเท่ากับเกณฑ์ที่ทำให้ $R^2_{adj\,(p)}$ มีค่าสูงสุดนั่นเอง

7.2.4 สถิติ Mallows' C_p

ในปีค.ศ. 1964 Mallows ได้เสนอเกณฑ์ในการเลือกตัวแปรอิสระที่สัมพันธ์กับความคลาดเคลื่อนกำลังสอง

เฉลี่ยของค่าพยากรณ์ ซึ่งมีสูตรดังนี้

$$
\begin{aligned}
E\left[\hat{Y}_i - E(Y_i)\right]^2 &= V(\hat{Y}_i) + \left[\text{Bias}(\hat{Y}_i)\right]^2 \\
&= V(\hat{Y}_i) + \left[E(Y_i) - E(\hat{Y}_i)\right]^2 \qquad (7.4)
\end{aligned}
$$

เมื่อ

$E(Y_i)$ แทน ค่าคาดหวังที่ได้จากสมการถดถอยที่แท้จริง

$E(\hat{Y}_i)$ แทน ค่าคาดหวังที่ได้จากสับเซ็ตของสมการถดถอยที่มี p เทอม

ให้ Γ_p แทน ผลรวมของความคลาดเคลื่อนกำลังสองเฉลี่ยของค่าพยากรณ์ในรูปคะแนนมาตรฐาน (Standardized total mean square error) โดยที่

$$
\begin{aligned}
\Gamma_p &= \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n} V(\hat{Y}_i) + \sum_{i=1}^{n}\left[E(Y_i) - E(\hat{Y}_i)\right]^2\right\} \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n} V(\hat{Y}_i) + \frac{SSB_p}{\sigma^2} \qquad (7.5)
\end{aligned}
$$

เมื่อ $SSB_p = \sum_{i=1}^{n}\left[E(Y_i) - E(\hat{Y}_i)\right]^2$ แทน ผลรวมกำลังสองของความเอนเอียง (Total squared bias) สำหรับสมการถดถอยที่มี p เทอม

และสามารถแสดงได้ว่า

$$
\begin{aligned}
\sum_{i=1}^{n} V(\hat{Y}_i) &= p\sigma^2 \qquad (7.6) \\
E[SSE_p] &= SSB_p + (n-p)\sigma^2 \qquad (7.7)
\end{aligned}
$$

แทนค่า $\sum_{i=1}^{n} V(\hat{Y}_i)$ และ SSB_p เข้าไปในสมการที่ (7.5) จะได้ว่า

$$
\begin{aligned}
\Gamma_p &= \frac{p\sigma^2}{\sigma^2} + \frac{\left\{E[SSE_p] - (n-p)\sigma^2\right\}}{\sigma^2} \\
&= \frac{E[SSE_p]}{\sigma^2} - n + 2p
\end{aligned}
$$

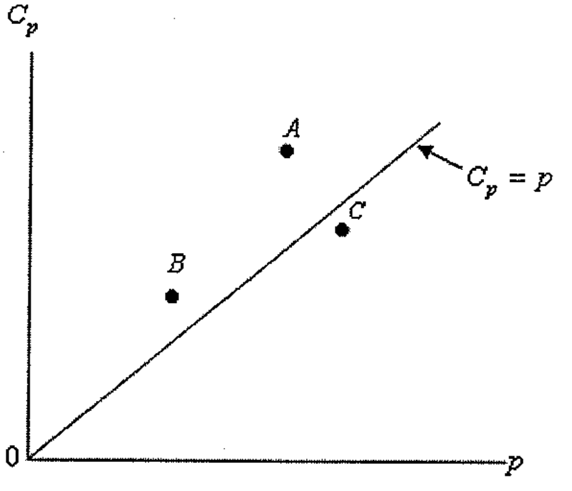
แทนค่า σ^2 และ $E[SSE_p]$ ด้วยค่าสังเกตที่ได้จากข้อมูลตัวอย่าง คือ $\hat{\sigma}^2$ และ SSE_p ตามลำดับ จะได้ว่า

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} - n + 2p \tag{7.8}$$

เมื่อ C_p เป็นตัวประมาณของ Γ_p

ถ้า $SSB_p = 0$ จะส่งผลให้ $E[SSE_p] = (n-p)\sigma^2$ ดังนั้น

$$E\big[C_p \mid \text{Bias} = 0\big] = \frac{(n-p)\sigma^2}{\sigma^2} - n + 2p = p \tag{7.9}$$



รูปที่ 7.3: กราฟระหว่าง C_p และ p

รูปที่ 7.3 แสดงกราฟของ C_p ซึ่งเป็นฟังก์ชันของ p สำหรับสมการถดถอยแต่ละสมการ โดยถ้าสมการถดถอยมีความเอนเอียงเล็กน้อย ค่า C_p จะตกใกล้กับเส้นตรง $C_p = p$ (จุด C) แต่ถ้าสมการถดถอยมีความเอนเอียงมาก ค่า C_p จะตกอยู่เหนือเส้นตรง $C_p = p$ (จุด A) โดยทั่วไปแล้วต้องการ C_p ที่มีค่าน้อย ๆ ถึงแม้ว่าจุด B จะตกอยู่เหนือเส้น $C_p = p$ แต่อยู่ต่ำกว่าจุด C แสดงว่าโมเดล B มีความคลาดเคลื่อนน้อยกว่าโมเดล C ซึ่งจะเห็นได้ว่าในบางครั้งอาจต้องรับเอาสมการที่มีความเอนเอียง เพื่อทำให้ความคลาดเคลื่อนเฉลี่ยของการพยากรณ์ลดลง

7.2.5 เกณฑ์การเปรียบเทียบโมเดลถดถอย จำแนกตามจุดมุ่งหมาย

เป็นที่ทราบแล้วว่าเกณฑ์ในการเลือกและเปรียบเทียบโมเดลมีความสัมพันธ์กับจุดมุ่งหมายของการสร้างโมเดล แสดงความสัมพันธ์ระหว่างข้อมูลดังนี้

- ถ้าจุดมุ่งหมายของการสร้างสมการถดถอยเพื่ออธิบายกระบวนการหรือระบบที่ซับซ้อน พบว่าการค้นหาสมการถดถอยที่มีค่า SSE น้อยที่สุดจะเป็นที่ต้องการ เนื่องจาก SSE มีค่าน้อยที่สุด เมื่อสมการถด-

ถดถอยมีตัวแปรอิสระทั้ง k ตัว แต่นักวิเคราะห์มักต้องการใช้ตัวแปรอิสระเพียงไม่กี่ตัว ขณะเดียวกันสามารถอธิบายความผันแปรในข้อมูลได้ในสัดส่วนที่มากพอ ดังนั้นอาจต้องตัดตัวแปรอิสระบางตัวที่อธิบายความสัมพันธ์ในข้อมูลได้น้อยออกจากสมการ ซึ่งอาจส่งผลให้ SSE เพิ่มขึ้นเล็กน้อย

- ถ้าจุดมุ่งหมายของการสร้างสมการถดถอยเพื่อพยากรณ์ค่าสังเกตค่าใหม่หรือประมาณค่าเฉลี่ยของตัวแปรตาม โดยทั่วไปมักเลือกสมการที่ทำให้ผลรวมกำลังสองของความคลาดเคลื่อนของค่าพยากรณ์ (Predictive error sum of squares) สำหรับสับเซ็ตของโมเดลถดถอยที่มี p เทอม แทนด้วย PRESS_p มีค่าน้อยที่สุด มีสูตรดังนี้

$$\begin{aligned} \text{PRESS}_p &= \sum_{i=1}^{n} \left[Y_i - \hat{Y}_{(i)}\right]^2 \\ &= \sum_{i=1}^{n} \left(\frac{e_i}{1-h_{ii}}\right)^2 \end{aligned} \tag{7.10}$$

- ถ้าจุดมุ่งหมายของการสร้างสมการถดถอยเพื่อประมาณค่าพารามิเตอร์ ควรพิจารณาเลือกสมการโดยพิจารณาทั้งค่าความเอนเอียงและความแปรปรวนของการประมาณค่าสัมประสิทธิ์ถดถอยที่ได้จากโมเดลที่มีตัวแปรอิสระบางตัวถูกตัดออกไป

- ถ้าจุดมุ่งหมายของการสร้างสมการถดถอยเพื่อควบคุมค่าของตัวแปรตาม พบว่าการประมาณค่าพารามิเตอร์ที่ถูกต้องเป็นสิ่งสำคัญ จึงควรเลือกสมการโดยพิจารณาจากค่าความคลาดเคลื่อนมาตรฐานของการประมาณสัมประสิทธิ์ถดถอย ซึ่งควรมีค่าน้อย ๆ

7.3 วิธีเลือกตัวแปรอิสระสำหรับโมเดลถดถอย

จะเห็นได้ว่าการเลือกโมเดลถดถอยเป็นการค้นหาสับเซ็ตของตัวแปรอิสระที่เหมาะสมสำหรับข้อมูลที่พิจารณานั่นเอง ซึ่งวิธีเลือกตัวแปรอิสระใส่เข้าในโมเดลมีหลายวิธีด้วยกัน และปัจจุบันนี้โปรแกรมวิเคราะห์ข้อมูลทางสถิติส่วนใหญ่ได้รวมเอาวิธีการเลือกตัวแปรเหล่านี้ไว้แล้ว ทำให้สะดวกต่อการใช้งานมากขึ้น แต่อย่างไรก็ตามผู้วิเคราะห์ไม่ควรขึ้นอยู่กับผลที่ได้จากคอมพิวเตอร์เพียงอย่างเดียว แต่ควรคำนึงถึงความสมเหตุสมผลประกอบด้วย ในบทนี้จะกล่าวถึงวิธีการเลือกตัวแปรอิสระเข้าในโมเดลที่นิยมใช้ 4 วิธีดังนี้

7.3.1 การเลือกตัวแปรด้วยวิธี All Possible Regressions

วิธีนี้จะเลือกสมการถดถอยที่ดีที่สุดจากรูปแบบของสมการถดถอยที่เป็นไปได้ทั้งหมด ซึ่งได้จากการสร้างสมการถดถอยที่มีตัวแปรอิสระหนึ่งตัว สองตัว ไปเรื่อย ๆ จนครบทั้ง k ตัว แล้วพิจารณาจากเกณฑ์ที่ใช้ในการเลือกสมการ เพื่อค้นหาสมการถดถอยที่ดีที่สุด หากสมการที่เป็นไปได้ทั้งหมดมีระยะตัดแกน Y หรือ b_0 อยู่ด้วย

และมีตัวแปรอิสระที่ต้องพิจารณาทั้งหมด k ตัว พบว่าจะต้องสร้างและตรวจสอบสมการทั้งหมดจำนวน 2^k สมการ เช่น ถ้า $k = 5$ จำนวนสมการที่เป็นไปได้ทั้งหมด คือ $2^5 = 32$ สมการ และเมื่อ $k = 10$ จำนวนสมการที่เป็นไปได้ทั้งหมด คือ $2^{10} = 1,024$ จะเห็นได้ว่าจำนวนสมการที่ต้องตรวจสอบจะเพิ่มขึ้นอย่างรวดเร็ว เมื่อจำนวนตัวแปรอิสระเพิ่มขึ้น ซึ่งในอดีตการสร้างสมการถดถอยที่เป็นไปได้ทั้งหมดจะจำกัดอยู่เฉพาะกรณีที่มีตัวแปรอิสระเพียง $2 - 3$ ตัวเท่านั้น เมื่อคอมพิวเตอร์ความเร็วสูงได้พัฒนาขึ้น อัลกอริทึมที่ใช้ในการสร้างและตรวจสอบสมการที่เป็นไปได้ทั้งหมดมีประสิทธิภาพมากขึ้น ทำให้ข้อจำกัดเหล่านี้ลดน้อยลง

ตัวอย่างที่ 7.1 การศึกษากำลังในการขับเคลื่อนเครื่องบิน โดยมีตัวแปรอิสระที่พิจารณา 4 ตัว คือ

X_1 แทน กำลังต่อน้ำหนักหนึ่งหน่วย

X_2 แทน พิสัยในการบิน

X_3 แทน สัดส่วนของน้ำหนักผู้โดยสารต่อน้ำหนักสุทธิของเครื่องบิน

X_4 แทน ระวางน้ำหนัก

โดย Y แทน จำนวนวันที่ขึ้นบินนับจากเดือนมกราคม หน่วยเป็นเดือน ได้ข้อมูลดังนี้

Y	X_1	X_2	X_3	X_4
85	1.258	3.07	0.177	0.25
92	1.395	3.41	0.165	0.25
104	1.958	4.64	0.188	3.05
110	1.844	4.49	0.286	1.25
118	2.257	3.88	0.309	1.15
125	1.084	3.52	0.161	1.05
130	1.973	3.74	0.011	2.55
140	2.216	4.42	0.128	1.95
150	2.397	3.61	0.166	2.45
169	4.357	4.69	0.149	3.35
177	4.378	3.59	0.260	3.65
178	3.408	4.09	0.154	2.95
180	5.645	4.30	0.183	2.65
187	2.688	4.25	0.189	3.15
190	3.670	5.16	0.112	3.35

ตารางที่ 7.1: ข้อมูลของตัวอย่างที่ 7.1

จงเลือกตัวแปรเข้าในสมการถดถอยโดยใช้วิธี All possible regressions

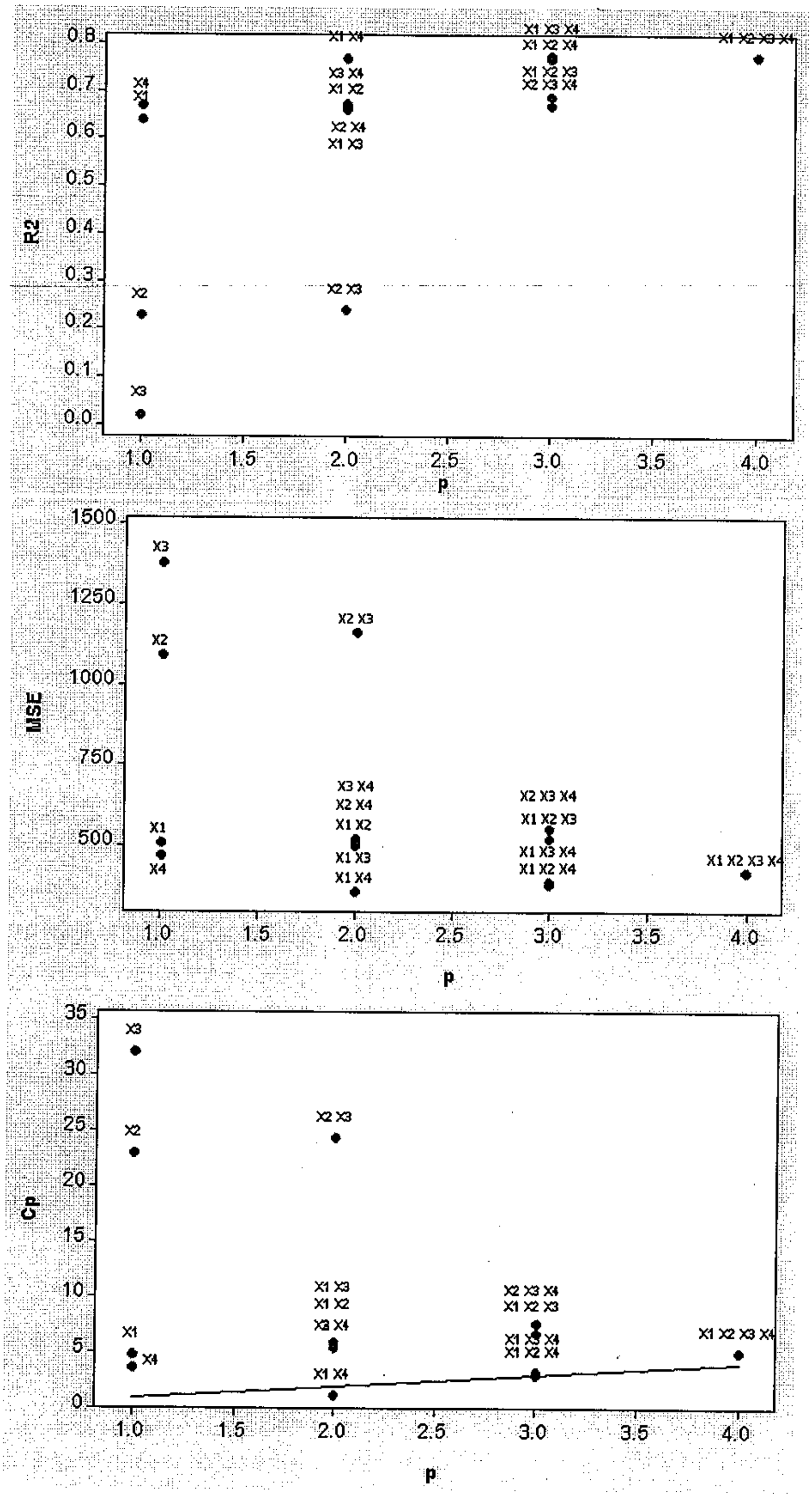
จำนวนตัวแปรที่อยู่ในโมเดล	R^2	R^2_{adj}	C_p	MSE	ตัวแปรที่อยู่ในโมเดล
0	0	0	30.9248	1308.2381	ไม่มี
1	0.6687	0.6432	3.5542	466.8110	X_4
1	0.6410	0.6134	4.7680	505.7431	X_1
1	0.2267	0.1672	22.9674	1089.4734	X_2
1	0.0206	0.0548	32.0222	1379.8986	X_3
2	0.7676	0.7288	1.2096	354.7533	$X_1\,X_4$
2	0.6734	0.6190	5.3455	498.4634	$X_1\,X_3$
2	0.6691	0.6139	5.5371	505.1195	$X_3\,X_4$
2	0.6687	0.6135	5.5528	505.6666	$X_2\,X_4$
2	0.6599	0.6032	5.9401	519.1231	$X_1\,X_2$
2	0.2392	0.1124	24.4196	1161.2276	$X_2\,X_3$
3	0.7721	0.7099	3.0109	379.4716	$X_1\,X_3\,X_4$
3	0.7679	0.7047	3.1933	386.3825	$X_1\,X_2\,X_4$
3	0.6879	0.6028	6.7093	519.6591	$X_1\,X_2\,X_3$
3	0.6691	0.5788	7.5351	550.9647	$X_2\,X_3\,X_4$
4	0.7723	0.6813	5.0000	416.9630	$X_1\,X_2\,X_3\,X_4$

ตารางที่ 7.2: แสดงการเลือกตัวแปรด้วยวิธี All possible regressions

วิธีทำ เนื่องจากมีตัวแปรอิสระ 4 ตัว ดังนั้นสมการถดถอยที่เป็นไปได้ทั้งหมดที่ต้องพิจารณาคือ $2^4 = 16$ สมการ ตารางที่ 7.2 สรุปการเลือกตัวแปรเข้าในโมเดลด้วยวิธี All possible regressions

จากการพิจารณาสับเซ็ตของโมเดลถดถอยในตารางที่ 7.2 จะเห็นได้ว่าโมเดลที่มีตัวแปรอิสระอยู่แล้ว 2 ตัว เมื่อเพิ่มตัวแปรอิสระเข้าไปอีก 1 ตัว ทำให้ค่า R^2 เพิ่มขึ้นเล็กน้อยเป็นส่วนใหญ่ ซึ่งสอดคล้องกับรูปที่ 7.4 (บน) และส่งผลให้การเลือกโมเดลสุดท้ายยังไม่ชัดเจน แต่เมื่อพิจารณาจากค่า R^2_{adj} พบว่าโมเดลที่ประกอบด้วยตัว-แปรอิสระ (X_1, X_4) มีค่า R^2_{adj} สูงสุด เมื่อพิจารณาค่า MSE ดังแสดงในรูปที่ 7.4 (กลาง) พบว่าโมเดลที่มี (X_1, X_4) มีค่า MSE ต่ำที่สุด โดยที่ $MSE = 354.7533$ ซึ่งสอดคล้องกับที่คาดไว้ นั่นคือ โมเดลที่มี MSE ต่ำที่สุด จะให้ R^2_{adj} สูงสุด และเมื่อพิจารณาจากกราฟระหว่าง C_p กับ p ในรูปที่ 7.4 (ล่าง) พบว่ามีโมเดลที่สามารถยอมรับได้ 3 โมเดล คือ โมเดลที่ประกอบด้วยตัวแปรอิสระ (X_1, X_4), (X_1, X_2, X_4) และ (X_1, X_3, X_4) หากไม่มีข้อมูลเพิ่มเติมเกี่ยวกับตัวแปรอิสระหรือค่าใช้จ่ายในการเก็บรวบรวมข้อมูล การเลือกโมเดลที่ง่ายที่สุดที่ประกอบด้วยตัวแปรอิสระ 2 ตัว คือ (X_1, X_4) อาจมีความเหมาะสม เนื่องจากมีค่า C_p น้อยที่สุด

หมายเหตุ บางครั้งการเลือกโมเดลถดถอยที่ดีที่สุดโดยใช้วิธี All possible regressions อาจให้ผลที่ไม่ชัดเจน และบ่อยครั้งอาจพบว่าการพิจารณาเกณฑ์ที่แตกแต่างกัน อาจนำไปสู่โมเดลที่แตกต่างกันด้วย แต่อย่างไรก็ตาม โมเดลที่จะรับไปใช้ในขั้นสุดท้าย ควรที่จะตรวจสอบความเหมาะสมของโมเดลเสมอ ซึ่งรวมถึงการตรวจสอบค่าสังเกตที่ผิดปกติ ค่าสังเกตที่มีอิทธิพล และปัญหา Multicollinearity ด้วย



รูปที่ 7.4: กราฟระหว่าง R^2 และ p (บน) กราฟระหว่าง MSE และ p (กลาง) และ กราฟระหว่าง C_p และ p (ล่าง)

7.3.2 การเลือกตัวแปรด้วยวิธี Forward Selection

จะเห็นได้ว่าการเลือกตัวแปรอิสระโดยพิจารณาจากโมเดลถดถอยที่เป็นไปได้ทั้งหมดย่อมก่อให้เกิดความยุ่งยากในการคำนวณ จึงได้มีการพัฒนาวิธีในการเลือกตัวแปรที่อยู่ในโมเดล ซึ่งจะพิจารณาเฉพาะสับเซ็ตของโมเดลถดถอยเพียงไม่กี่โมเดล การเลือกตัวแปรด้วยวิธี Forward Selection จะเริ่มจากโมเดลที่ไม่มีตัวแปรอิสระเลย การค้นหาสับเซ็ตของโมเดลที่เหมาะสมได้จากการเพิ่มตัวแปรอิสระเข้าไปในโมเดลครั้งละ 1 ตัว และจะค้นหาจนกระทั่งไม่มีตัวแปรอิสระตัวใดที่สามารถอธิบายความผันแปรของ Y ได้อีก เมื่อตัวแปรอิสระตัวใดถูกเลือกเข้ามาอยู่ในโมเดลแล้ว จะไม่มีการพิจารณาตัดออกภายหลัง ซึ่งขั้นตอนในการเลือกตัวแปรด้วยวิธี Forward Selection มีดังนี้

1. เลือกตัวแปรอิสระตัวแรกเข้าในโมเดล โดยเลือกตัวที่มีค่าสัมประสิทธิ์สหสัมพันธ์อย่างง่ายกับตัวแปรตาม Y สูงสุดเข้าเป็นตัวแรก ซึ่งจะเป็นตัวแปรที่ให้ค่าสถิติ F ที่ใช้ทดสอบนัยสำคัญของตัวแปรอิสระสูงสุดด้วยเช่นกัน สมมติตัวแปรอิสระที่ถูกเลือกเข้าเป็นตัวแรก คือ X_j

2. ทดสอบอิทธิพลของตัวแปรอิสระที่เพิ่มเข้ามา โดยกำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \beta_j = 0 \quad vs. \quad H_1 : \beta_j \neq 0$$

   - หากยอมรับ H_0 นั่นคือ F_c ที่คำนวณได้มีค่าน้อยกว่าค่า F ที่กำหนด ซึ่งแทนด้วย F_{IN} จะหยุดพิจารณาเลือกตัวแปรอิสระเข้าในโมเดล
   - หากปฏิเสธ H_0 นั่นคือ $F_c \geq F_{IN}$ จะรับเอา X_j เข้าในโมเดลถดถอย และค้นหาตัวแปรอิสระตัวถัดไป โดยทำต่อในขั้นที่ 3

3. เลือกตัวแปรอิสระตัวใหม่เข้าในสมการ โดยเลือกตัวแปรที่ให้ค่าสัมประสิทธิ์สหสัมพันธ์บางส่วน หรือค่าสัมประสิทธิ์ตัวกำหนดบางส่วนกับตัวแปรตาม Y เมื่อกำหนดให้ตัวแปรที่ถูกเลือกก่อนหน้านี้อยู่ในโมเดลด้วย มีค่าสูงสุด สมมติให้ตัวแปรที่พิจารณาเข้าในโมเดลตัวถัดไป คือ X_ℓ

4. ทดสอบอิทธิพลของตัวแปรอิสระที่เพิ่มเข้ามา เมื่อกำหนดให้ตัวแปรที่ถูกเลือกก่อนหน้านี้อยู่ในโมเดลด้วย โดยใช้ Partial F-test นั่นคือ

$$F_c = \frac{MSR(X_\ell \mid X_j)}{MSE(X_j, X_\ell)}$$

   - หากยอมรับ H_0 จะรับเอาโมเดลถดถอยเดิมที่มีเฉพาะ X_j อยู่ในโมเดลเป็นโมเดลสุดท้าย และหยุดพิจารณาเลือกตัวแปรอิสระเข้าในโมเดล

- หากปฏิเสธ H_0 จะรับเอา X_ℓ เข้าในโมเดลถดถอย และค้นหาตัวแปรอิสระตัวถัดไป การพิจารณาเลือกตัวแปรอิสระเข้าในโมเดลจะทำไปเรื่อย ๆ จนกระทั่งไม่สามารถหาตัวแปรอิสระเพิ่มเข้ามาในโมเดลได้อีก นั่นคือ $F_c < F_{IN}$ หรือจนกระทั่งตัวแปรอิสระตัวสุดท้ายถูกเพิ่มเข้ามาในโมเดล

ตัวอย่างที่ 7.2 จากข้อมูลในตัวอย่างที่ 7.1 จงเลือกตัวแปรอิสระเข้าในโมเดลโดยใช้วิธี Forward selection

วิธีทำ

	Y	X_1	X_2	X_3	X_4
Y	1.0000	0.8006	0.4761	-0.1434	0.8177
X_1	0.8006	1.0000	0.4408	0.0454	0.7072
X_2	0.4761	0.4408	1.0000	-0.0672	0.5877
X_3	-0.1434	0.0454	-0.0672	1.0000	-0.1989
X_4	0.8177	0.7072	0.5877	-0.1989	1.0000

ตารางที่ 7.3: แสดงค่าสัมประสิทธิ์สหสัมพันธ์ของข้อมูลในตัวอย่างที่ 7.1

ขั้นที่ 1 พิจารณาค่าสัมประสิทธิ์สหสัมพันธ์อย่างง่ายระหว่างตัวแปรตาม Y และตัวแปรอิสระ X_j เมื่อ $j = 1, 2, 3, 4$ ดังแสดงในตารางที่ 7.3 พบว่า X_4 มีค่าสัมประสิทธิ์สหสัมพันธ์อย่างง่ายกับ Y สูงที่สุด นั่นคือ $r_{Y4} = 0.8177$ ดังนั้น X_4 จะถูกพิจารณาเลือกเข้าในโมเดลเป็นตัวแรก
สร้างสมการถดถอยระหว่าง Y กับ X_4 ได้เป็น

$$\hat{Y} = \underset{(12.444)}{85.3615} + \underset{(5.0482)}{25.8571}X_4$$

โดยที่ $SSR(X_4) = 12,247$, $SSE(X_4) = 6,068.5434$ และ $SST = 18315$

ขั้นที่ 2 ทดสอบอิทธิพลของ X_4 ที่มีต่อตัวแปรตาม Y โดยกำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \beta_4 = 0 \quad vs. \quad H_1 : \beta_4 \neq 0$$

คำนวณสถิติทดสอบ

$$\begin{aligned} F_c &= \frac{MSR(X_4)}{MSE(X4)} \\ &= \frac{12,247}{466.8110} \\ &= 26.24 \end{aligned}$$

ค่าวิกฤติ คือ $F_{IN} = F_{1,13\,(0.05)} = 4.67$

สรุปผล ปฏิเสธ H_0 นั่นคือ X_4 มีอิทธิพลต่อ Y อย่างมีนัยสำคัญ ที่ $\alpha = 0.05$ ดังนั้นจะรับเอา X_4 เข้าในสมการถดถอย

ขั้นที่ 3 ค้นหาตัวแปรอิสระตัวใหม่ใส่เข้าในโมเดลที่มี X_4 อยู่ในโมเดลแล้ว

ตัวแปรอิสระที่อยู่ในโมเดล	SSR	SSE	สมการถดถอย
(X_1, X_4)	14,058	4,257.0398	$\hat{Y} = \underset{(11.8911)}{74.3541} - \underset{(5.3926)}{12.1859}X_1 + \underset{(6.2245)}{15.9098}X_4$
(X_2, X_4)	12,247	6,069.9990	$\hat{Y} = \underset{(46.0019)}{86.8097} - \underset{(12.9517)}{0.4249}X_2 + \underset{(6.4398)}{25.9823}X_4$
(X_3, X_4)	12,254	6,061.4341	$\hat{Y} = \underset{(21.6298)}{83.3056} - \underset{(85.1776)}{10.1051}X_3 + \underset{(5.3584)}{25.9836}X_4$

ตารางที่ 7.4: แสดงการคำนวณค่าต่าง ๆ ของโมเดลถดถอยที่มี X_4 อยู่ในโมเดลแล้ว

จากตารางที่ 7.4 คำนวณค่าสัมประสิทธิ์ตัวกำหนดบางส่วนระหว่างตัวแปรอิสระแต่ละตัว เมื่อมี X_4 อยู่ในโมเดลแล้ว ดังนี้

$$
\begin{aligned}
R^2_{Y1\cdot 4} &= \frac{SSR(X_1 \mid X_4)}{SSE(X_4)} = \frac{SSR(X_1, X_4) - SSR(X_4)}{SSE(X_4)} = \frac{14,058 - 12,247}{6,068.5434} = 0.2984 \\
R^2_{Y2\cdot 4} &= \frac{SSR(X_2 \mid X_4)}{SSE(X_4)} = \frac{SSR(X_2, X_4) - SSR(X_4)}{SSE(X_4)} = \frac{12,247 - 12,247}{6,068.5434} = 0 \\
R^2_{Y3\cdot 4} &= \frac{SSR(X_3 \mid X_4)}{SSE(X_4)} = \frac{SSR(X_3, X_4) - SSR(X_4)}{SSE(X_4)} = \frac{12,254 - 12,247}{6,068.5434} = 0.0012
\end{aligned}
$$

เนื่องจาก $R^2_{Y1\cdot 4}$ มีค่าสูงสุด ดังนั้น X_1 เป็นตัวแปรอิสระตัวถัดไปที่ถูกพิจารณาเข้าในโมเดลที่มี X_4 อยู่ในโมเดลแล้ว

ขั้นที่ 4 ทดสอบอิทธิพลของ X_1 ที่มีต่อตัวแปรตาม Y เมื่อให้ X_4 อยู่ในโมเดลแล้ว โดยกำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0$$

คำนวณ Partial F-test

$$\begin{aligned} F_c &= \frac{MSR(X_1 \mid X_4)}{MSE(X_1,\, X4)} \\ &= \frac{1,811/1}{4,257.0398/12} \\ &= 5.1049 \end{aligned}$$

ค่าวิกฤติ คือ $F_{IN} = F_{1,12\,(0.05)} = 4.75$

สรุปผล ปฏิเสธ H_0 นั่นคือ X_1 มีอิทธิพลต่อ Y อย่างมีนัยสำคัญ ที่ $\alpha = 0.05$ เมื่อให้ X_4 อยู่ในโมเดลแล้ว ดังนั้นจะรับเอา X_1 เข้าในโมเดลถดถอยเป็นตัวถัดไป

ขั้นที่ 5 ค้นหาตัวแปรอิสระตัวใหม่ใส่เข้าในโมเดลที่มี $(X_1,\, X_4)$ อยู่ในโมเดลแล้ว

ตัวแปรอิสระที่อยู่ในโมเดล	SSR	SSE	สมการถดถอย
$(X_1,\, X_2,\, X_4)$	14,065	4,250.2077	$\hat{Y} = \underset{(40.3542)}{79.4602} + \underset{(5.6334)}{12.2189}X_1 - \underset{(11.3325)}{1.5069}\,X_2 + \underset{(7.2137)}{16.3269}X_4$
$(X_1,\, X_3,\, X_4)$	14,141	4,174.1876	$\hat{Y} = \underset{(18.7765)}{80.9837} + \underset{(5.7902)}{12.9127}X_1 - \underset{(76.6450)}{35.8135}X_3 + \underset{(14.8682)}{6.8128}\,X_4$

ตารางที่ 7.5: แสดงการคำนวณค่าต่าง ๆ ของโมเดลถดถอยที่มี $(X_1,\, X_4)$ อยู่ในโมเดลแล้ว

จากตารางที่ 7.5 คำนวณค่าสัมประสิทธิ์ตัวกำหนดบางส่วนระหว่างตัวแปรอิสระแต่ละตัว เมื่อมี $(X_1,\, X_4)$ อยู่ในโมเดลแล้ว ดังนี้

$$\begin{aligned} R^2_{Y2\cdot14} &= \frac{SSR(X_2 \mid X_1,\, X_4)}{SSE(X_1,\, X_4)} = \frac{SSR(X_1, X_2, X_4) - SSR(X_1, X_4)}{SSE(X_1, X_4)} = \frac{14,065 - 14,058}{4,257.0398} \\ &= 0.0016 \\ R^2_{Y3\cdot14} &= \frac{SSR(X_3 \mid X_1,\, X_4)}{SSE(X_1,\, X_4)} = \frac{SSR(X_1, X_3, X_4) - SSR(X_1, X_4)}{SSE(X_1, X_4)} = \frac{14,141 - 14,058}{4,257.0398} \\ &= 0.0195 \end{aligned}$$

เนื่องจาก $R^2_{Y3\cdot14}$ มีค่าสูงสุด ดังนั้น X_3 เป็นตัวแปรอิสระตัวถัดไปที่ถูกพิจารณาเข้าในโมเดลที่มี $(X_1,\, X_4)$ อยู่ในโมเดลแล้ว

ขั้นที่ 6 ทดสอบอิทธิพลของ X_3 ที่มีต่อตัวแปรตาม Y เมื่อให้ $(X_1,\, X_4)$ อยู่ในโมเดลแล้ว โดยกำหนดสมมติฐาน

ของการทดสอบเป็น

$$H_0 : \beta_3 = 0 \quad vs. \quad H_1 : \beta_3 \neq 0$$

คำนวณ Partial F-test

$$\begin{aligned} F_c &= \frac{MSR(X_3 \mid X_1, X_4)}{MSE(X_1, X_3, X4)} \\ &= \frac{83/1}{4,174.1876/11} \\ &= 0.2187 \end{aligned}$$

ค่าวิกฤติ คือ $F_{IN} = F_{1,11\,(0.05)} = 4.84$

สรุปผล ไม่สามารถปฏิเสธ H_0 ได้ นั่นคือ X_3 ไม่มีอิทธิพลต่อ Y อย่างมีนัยสำคัญ ที่ $\alpha = 0.05$ เมื่อให้ $(X_1,\, X_4)$ อยู่ในโมเดลแล้ว ดังนั้นจะหยุดการค้นหาตัวแปรอิสระเข้าในโมเดล และรับเอาโมเดลที่มี $(X_1,\, X_4)$ เป็นโมเดลสุดท้าย ได้สมการพยากรณ์ดังนี้

$$\hat{Y} = \underset{(11.8911)}{74.3541} - \underset{(5.3926)}{12.1859}X_1 + \underset{(6.2245)}{15.9098}X_4$$

โดยมีค่า $R^2 = 0.7676$ และ $R^2_{adj} = 0.7288$

7.3.3 การเลือกตัวแปรด้วยวิธี Backward Elimination

การเลือกตัวแปรอิสระด้วยวิธี Forward selection เริ่มด้วยโมเดลที่ไม่มีตัวแปรอิสระเลย แล้วจึงเพิ่มตัวแปรอิสระเข้าไปครั้งละตัว จนกระทั่งได้โมเดลที่เหมาะสม การเลือกตัวแปรด้วยวิธี Backward Elimination จะทำงานในลักษณะตรงกันข้าม นั่นคือ เริ่มต้นด้วยโมเดลถดถอยที่ประกอบด้วยตัวแปรอิสระทั้ง k ตัว แล้วจึงตัดตัวแปรอิสระที่อธิบายความสัมพันธ์กับตัวแปรตาม Y ได้น้อยออกไปครั้งละ 1 ตัว จนกระทั่งตัวแปรที่เหลืออยู่ในโมเดลสามารถอธิบายความสัมพันธ์กับ Y ได้ทั้งหมด ก็จะหยุดการทำงาน เมื่อตัวแปรอิสระตัวใดถูกตัดออกจากโมเดลแล้ว จะไม่มีการพิจารณาเพิ่มเข้ามาอยู่ในโมเดลอีกในภายหลัง ซึ่งการเลือกตัวแปรด้วยวิธี Backward elimination มีขั้นตอนดังนี้

1. สร้างโมเดลถดถอยที่ประกอบด้วยตัวแปรอิสระทุกตัว ซึ่งในที่นี้ก็คือ k ตัว จากนั้นคำนวณ Partial F-test หรือ t test ของตัวแปรอิสระแต่ละตัว แล้วพิจารณาตัวแปรอิสระที่มีส่วนอธิบายความสัมพันธ์กับ

Y ได้น้อยที่สุด นั่นคือ ตัวที่มีค่าสถิติ F_c หรือ $|t_c|$ ต่ำที่สุด สมมติให้เป็น X_i และทดสอบอิทธิพลของ X_i โดยกำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \beta_i = 0 \quad vs. \quad H_1 : \beta_i \neq 0$$

- หากปฏิเสธ H_0 นั่นคือ ค่า F_c ที่คำนวณได้มีค่ามากกว่าหรือเท่ากับค่า F ที่กำหนด ซึ่งเรียกว่า F_{OUT} หรือ $|t_c| \geq t_{OUT}$ แสดงว่าไม่สามารถตัด X_i ออกจากโมเดลได้ จะหยุดการทำงาน และรับเอาโมเดลดังกล่าวเป็นโมเดลสุดท้าย
- หากยอมรับ H_0 นั่นคือ $F_c < F_{OUT}$ หรือ $|t_c| < t_{OUT}$ จะตัดตัวแปรดังกล่าวออกจากโมเดล และค้นหาตัวแปรอิสระตัวใหม่ที่จะถูกตัดออกจากโมเดลเป็นลำดับต่อไป

2. สร้างสมการถดถอยใหม่ที่ประกอบด้วยตัวแปรอิสระ $k-1$ ตัว (ไม่รวม X_i) จากนั้นคำนวณ Partial F-test หรือ t test ของตัวแปรอิสระแต่ละตัวทั้ง $k-1$ ตัว พิจารณาตัวแปรที่มีค่า F_c หรือ $|t_c|$ ต่ำที่สุด สมมติให้เป็น X_ℓ แล้วกำหนดสมมติฐานเพื่อทดสอบอิทธิพลของ X_ℓ เมื่อตัวแปรอิสระอื่นอยู่ในโมเดลด้วย กำหนดสมมติฐานของการทดสอบได้เป็น

$$H_0 : \beta_\ell = 0 \quad vs. \quad H_1 : \beta_\ell \neq 0$$

- หากปฏิเสธ H_0 แสดงว่าไม่สามารถตัด X_ℓ ออกจากโมเดลได้ จะหยุดการทำงาน และรับเอาโมเดลที่ประกอบด้วยตัวแปรอิสระ $k-1$ ตัวเป็นโมเดลสุดท้าย
- หากยอมรับ H_0 จะตัด X_ℓ ออกจากโมเดล และค้นหาตัวแปรอิสระตัวใหม่ที่จะถูกตัดออกจากโมเดลเป็นลำดับต่อไป ซึ่งการพิจารณาตัดตัวแปรอิสระออกไปครั้งละตัวจะทำไปเรื่อย ๆ จนกระทั่งไม่สามารถตัดตัวแปรตัวใดออกจากโมเดลได้อีก นั่นคือ ตัวแปรทุกตัวมีค่า $F_c \geq F_{OUT}$ หรือ $|t_c| \geq t_{OUT}$ นั่นเอง

ตัวอย่างที่ 7.3 จากข้อมูลในตัวอย่างที่ 7.1 จงเลือกตัวแปรอิสระเข้าในโมเดลโดยใช้วิธี Backward elimination

วิธีทำ

ขั้นที่ 1 สร้างสมการถดถอยที่มีตัวแปรอิสระทั้ง 4 ตัว ได้ดังนี้

$$\hat{Y} = 85.0334 + 12.9307X_1 - 1.2326X_2 - 35.3682X_3 + 15.2223X_4 \qquad (7.11)$$

ตัวแปรอิสระที่อยู่ในโมเดล	ค่าสัมประสิทธิ์ถดถอย	ค่าความคลาดเคลื่อนมาตรฐาน	t
X_1	12.9307	6.0719	2.13
X_2	-1.2326	11.7889	-0.10
X_3	-35.3682	80.4549	-0.44
X_4	15.2223	7.9038	1.93

ตารางที่ 7.6: แสดงการทดสอบอิทธิพลของตัวแปรอิสระแต่ละตัวของสมการที่ (7.11)

จากนั้นคำนวณสถิติทดสอบ t_c ดังแสดงในตารางที่ 7.6 จะเห็นได้ว่า $|t_c|$ ของตัวแปรอิสระ X_2 มีค่าต่ำสุด ทำการทดสอบอิทธิพลของ X_2 โดยกำหนดสมมติฐานของการทดสอบดังนี้

$$H_0 : \beta_2 = 0 \quad vs. \quad H_2 : \beta_2 \neq 0$$

ได้ค่าสถิติทดสอบ $|t_c| = 0.10$

ค่าวิกฤติ คือ $t_{.025,10} = 2.228$

สรุปผล ไม่สามารถปฏิเสธ H_0 ดังนั้น สามารถตัด X_2 ออกจากโมเดลได้

ขั้นที่ 2 สร้างสมการถดถอยที่มีตัวแปรอิสระ 3 ตัว (ตัด X_2 ออก) ได้ดังนี้

$$\hat{Y} = 80.9837 + 12.9127X_1 - 35.8135X_3 + 6.8128X_4 \tag{7.12}$$

ตัวแปรอิสระที่อยู่ในโมเดล	ค่าสัมประสิทธิ์ถดถอย	ค่าความคลาดเคลื่อนมาตรฐาน	t
X_1	12.9127	5.7902	2.23
X_3	-35.8135	76.6450	-0.47
X_4	6.8128	14.8682	2.18

ตารางที่ 7.7: แสดงการทดสอบอิทธิพลของตัวแปรอิสระแต่ละตัวของสมการที่ (7.12)

จากตารางที่ 7.7 จะเห็นได้ว่า $|t_c|$ ของตัวแปรอิสระ X_3 มีค่าต่ำสุด ทำการทดสอบอิทธิพลของ X_3 โดยกำหนดสมมติฐานของการทดสอบดังนี้

$$H_0 : \beta_3 = 0 \quad vs. \quad H_2 : \beta_3 \neq 0$$

ได้ค่าสถิติทดสอบ $|t_c| = 0.47$

ค่าวิกฤติ คือ $t_{.025,11} = 2.201$

สรุปผล ไม่สามารถปฏิเสธ H_0 ดังนั้นสามารถตัด X_3 ออกจากโมเดลได้

ขั้นที่ 3 สร้างสมการถดถอยที่มีตัวแปรอิสระ 2 ตัว (ตัด X_2 และ X_3 ออก) ได้ดังนี้

$$\hat{Y} = 74.3541 - 12.1859X_1 + 15.9098X_4 \tag{7.13}$$

ตัวแปรอิสระที่อยู่ในโมเดล	ค่าสัมประสิทธิ์ถดถอย	ค่าความคลาดเคลื่อนมาตรฐาน	t
X_1	12.1859	5.3926	2.26
X_4	15.9098	6.2245	5.56

ตารางที่ 7.8: แสดงการทดสอบอิทธิพลของตัวแปรอิสระแต่ละตัวของสมการที่ (7.13)

จากตารางที่ 7.8 จะเห็นได้ว่า $|t_c|$ ของตัวแปรอิสระ X_1 มีค่าต่ำสุด ทำการทดสอบอิทธิพลของ X_1 โดยกำหนดสมมติฐานของการทดสอบดังนี้

$$H_0 : \beta_1 = 0 \quad vs. \quad H_2 : \beta_1 \neq 0$$

ได้ค่าสถิติทดสอบ $|t_c| = 2.26$
ค่าวิกฤติ คือ $t_{.025,12} = 2.179$
สรุปผล ปฏิเสธ H_0 นั่นคือ X_1 มีอิทธิพลต่อ Y อย่างมีนัยสำคัญที่ $\alpha = 0.05$ จึงไม่สามารถตัด X_1 ออกจากโมเดลได้ และจะหยุดการตัดตัวแปรอิสระออกจากโมเดล ดังนั้นจึงรับเอาสมการที่มี X_1 และ X_4 เป็นสมการสุดท้าย ซึ่งจะเห็นได้ว่าผลที่ได้สอดคล้องกับการเลือกตัวแปรด้วยวิธี Forward selection

หมายเหตุ บ่อยครั้งการเลือกตัวแปรด้วยวิธี Backward elimination จัดเป็นกระบวนการที่ดีในการคัดเลือกตัวแปรเข้าในโมเดล และเป็นที่นิยมของนักวิเคราะห์ที่ต้องการศึกษาอิทธิพลของตัวแปรอิสระที่เป็นไปได้ทั้งหมด โดยไม่ได้ละเลยหรือมองข้ามตัวแปรตัวใดตัวหนึ่ง

7.3.4 การเลือกตัวแปรด้วยวิธี Stepwise Regression

นอกจากการเลือกตัวแปรด้วยวิธี Forward selection และ Backward elimination แล้ว ยังมีการเลือกตัวแปรอีกวิธีหนึ่งที่เป็นที่นิยมใช้กันมาก ก็คือ วิธี Stepwise regression ซึ่งประยุกต์มาจากวิธี Forward selection โดยเลือกตัวแปรอิสระเข้าในโมเดลครั้งละ 1 ตัวเช่นกัน แต่เพิ่มการทดสอบอิทธิพลของตัวแปรอิสระที่อยู่ในโมเดลก่อนหน้านี้แล้วทุกตัว เมื่อกำหนดให้ตัวแปรอิสระอื่นยังคงอยู่ในโมเดลด้วย โดยใช้ Partial F-test หรือ t-test ดังนั้นตัวแปรอิสระที่เพิ่มเข้ามาในโมเดลถดถอยในขั้นแรก ๆ อาจถูกตัดออกไปได้ในภายหลัง นอกจากนี้การเลือกตัวแปรด้วยวิธี Stepwise regression ต้องการค่าวิกฤติ 2 ค่า คือ F_{IN} และ F_{OUT} (หรือ t_{IN} และ

t_{OUT}) โดยอาจให้ $F_{IN} = F_{OUT}$ ก็ได้ แต่โดยทั่วไปนิยมให้ $F_{IN} > F_{OUT}$ เพื่อให้การเพิ่มตัวแปรเข้าในโมเดลทำได้ยากกว่าที่จะตัดตัวแปรออกจากโมเดล ซึ่งการเลือกตัวแปรด้วยวิธี Stepwise regression มีขั้นตอนดังนี้

1. เลือกตัวแปรอิสระตัวแรกเข้าในโมเดล โดยเลือกตัวที่มีค่าสัมประสิทธิ์สหสัมพันธ์อย่างง่ายกับตัวแปรตามสูงสุด สมมติให้เป็น X_j

2. สร้างโมเดลถดถอยที่มี X_j อยู่ด้วย และทดสอบอิทธิพลของ X_j โดยกำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \beta_j = 0 \quad vs. \quad H_2 : \beta_j \neq 0$$

   - หากยอมรับ H_0 นั่นคือ $F_c < F_{IN}$ จะหยุดพิจารณาเลือกตัวแปรอิสระเข้าในโมเดล
   - หากปฏิเสธ H_0 นั่นคือ $F_c \geq F_{IN}$ จะรับเอา X_j เข้าในโมเดลถดถอย และค้นหาตัวแปรอิสระตัวถัดไป โดยทำต่อในขั้นที่ 3

3. เลือกตัวแปรอิสระตัวใหม่เข้าในสมการ โดยเลือกตัวแปรที่ให้ค่าสัมประสิทธิ์สหสัมพันธ์บางส่วน หรือค่าสัมประสิทธิ์ตัวกำหนดบางส่วนกับตัวแปรตาม Y เมื่อกำหนดให้ตัวแปรที่ถูกเลือกก่อนหน้านี้อยู่ในโมเดลด้วย มีค่าสูงสุด สมมติให้ตัวแปรที่พิจารณาเข้าในโมเดลตัวถัดไป คือ X_ℓ

4. สร้างโมเดลถดถอยที่มีทั้ง X_j และ X_ℓ รวมอยู่ด้วย และทดสอบอิทธิพลของตัวแปรอิสระที่อยู่ในโมเดลทีละตัว เมื่อกำหนดให้ตัวแปรอื่นยังคงอยู่ในโมเดลด้วย นั่นคือ ทดสอบอิทธิพลของ X_ℓ เมื่อมี X_j รวมอยู่ในโมเดลด้วย โดยกำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \beta_\ell = 0 \quad vs. \quad H_2 : \beta_\ell \neq 0$$

   คำนวณ Partial F-test นั่นคือ

$$F_c = \frac{MSR(X_\ell \mid X_j)}{MSE(X_j, X_\ell)}$$

   - หากยอมรับ H_0 แสดงว่า X_ℓ ไม่มีอิทธิพลต่อ Y เมื่อมี X_j อยู่ด้วย จะตัด X_ℓ ออกไป
   - หากปฏิเสธ H_0 จะรับเอา X_ℓ เข้าในโมเดลถดถอย และทดสอบอิทธิพลของ X_j เมื่อมี X_ℓ รวมอยู่ในโมเดลด้วย โดยกำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \beta_j = 0 \quad vs. \quad H_2 : \beta_j \neq 0$$

คำนวณ Partial F-test นั่นคือ

$$F_c = \frac{MSR(X_j \mid X_\ell)}{MSE(X_j, X_\ell)}$$

– หากยอมรับ H_0 แสดงว่า X_j ไม่มีอิทธิพลต่อ Y เมื่อมี X_ℓ อยู่ด้วย จะตัด X_j ออกไป

– หากปฏิเสธ H_0 จะเก็บ X_j ไว้ในโมเดลถดถอยต่อไป

5. การพิจารณาเลือกตัวแปรดังกล่าวจะทำไปเรื่อย ๆ จนกระทั่งไม่สามารถหาตัวแปรอิสระอื่นเพิ่มเข้ามาในโมเดลได้อีก และขณะเดียวกันก็ไม่สามารถตัดตัวแปรอิสระที่อยู่ในโมเดลออกได้ นั่นคือ ตัวแปรอิสระทุกตัวที่อยู่ในโมเดลมีค่า $F_c \geq F_{OUT}$

ตัวอย่างที่ 7.4 จากข้อมูลในตัวอย่างที่ 7.1 จงเลือกตัวแปรอิสระเข้าในโมเดลโดยใช้วิธี Stepwise selection

วิธีทำ

ขั้นที่ 1 พิจารณาค่าสัมประสิทธิ์สหสัมพันธ์อย่างง่ายระหว่าง Y กับตัวแปรอิสระแต่ละตัว

$$r_{Y1} = 0.8066, \qquad r_{Y2} = 0.4761,$$
$$r_{Y3} = -0.1434, \qquad r_{Y4} = 0.8177$$

จะเห็นได้ว่า r_{Y4} มีค่าสูงสุด ดังนั้นตัวแปร X_4 จะถูกเลือกเข้าในโมเดลเป็นตัวแรก

ขั้นที่ 2 สร้างสมการถดถอยที่มี X_4 อยู่ในสมการได้ดังนี้

$$\hat{Y} = \underset{(12.444)}{85.3615} + \underset{(5.0482)}{25.8571}X_4$$

โดยที่ $SSR(X_4) = 12,247$, $SSE(X_4) = 6,068.5434$ และ $SST = 18315$

ทดสอบอิทธิพลของ X_4 ที่มีต่อตัวแปรตาม Y โดยกำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \beta_4 = 0 \quad vs. \quad H_1 : \beta_4 \neq 0$$

คำนวณสถิติทดสอบ

$$F_c = \frac{SSR(X_4)}{SSE(X4)} = \frac{12,247}{466.8110} = 26.24$$

ค่าวิกฤติ คือ $F_{IN} = F_{1,13\,(0.05)} = 4.67$

สรุปผล ปฏิเสธ H_0 นั่นคือ X_4 มีอิทธิพลต่อ Y อย่างมีนัยสำคัญ ที่ $\alpha = 0.05$ ดังนั้นจะรับเอา X_4 เข้าในสมการถดถอย

ขั้นที่ 3 เลือกตัวแปรอิสระตัวใหม่เข้าในสมการ โดยพิจารณาค่าสัมประสิทธิ์ตัวกำหนดบางส่วนระหว่างตัวแปรอิสระแต่ละตัวกับ Y เมื่อ X_4 อยู่ในโมเดลด้วย ซึ่งได้จากตัวอย่างที่ 7.2 ดังนี้

$$R^2_{Y1\cdot 4} = 0.2984, \qquad R^2_{Y2\cdot 4} = 0, \qquad R^2_{Y3\cdot 4} = 0.0012$$

เนื่องจาก $R^2_{Y1\cdot 4}$ มีค่าสูงสุด ดังนั้น X_1 เป็นตัวแปรอิสระตัวถัดไปที่ถูกพิจารณาเข้าในโมเดลที่มี X_4 อยู่ในโมเดลแล้ว

ขั้นที่ 4 สร้างสมการถดถอยที่มีตัวแปร (X_1, X_4) อยู่ในสมการได้ดังนี้

$$\hat{Y} = 74.3541 - 12.1859X_1 + 15.9098X_4 \tag{7.14}$$

การทดสอบอิทธิพลของ X_1 ที่มีต่อตัวแปรตาม Y เมื่อให้ X_4 อยู่ในโมเดลแล้ว ทำได้โดยกำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0$$

คำนวณ Partial F-test

$$\begin{aligned} F_c &= \frac{MSR(X_1 \mid X_4)}{MSE(X_1,\, X4)} \\ &= 5.1049 \end{aligned}$$

ค่าวิกฤติ คือ $F_{IN} = F_{1,12\,(0.05)} = 4.75$

สรุปผล ปฏิเสธ H_0 นั่นคือ X_1 มีอิทธิพลต่อ Y อย่างมีนัยสำคัญ ที่ $\alpha = 0.05$ เมื่อให้ X_4 อยู่ในโมเดลแล้ว

จากนั้นทดสอบอิทธิพลของ X_4 ที่มีต่อตัวแปรตาม Y เมื่อให้ X_1 อยู่ในโมเดลแล้ว โดยกำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \beta_4 = 0 \quad vs. \quad H_1 : \beta_4 \neq 0$$

คำนวณ Partial F-test

$$
\begin{aligned}
F_c &= \frac{MSR(X_4 \mid X_1)}{MSE(X_1, X4)} \\
&= \frac{[SSR(X_1, X_4) - SSR(X_1)]/1}{SSE(X_1, X4)/12} \\
&= \frac{(14,058 - 11,741)/1}{4,257.0398/12)} \\
&= 6.5313
\end{aligned}
$$

ค่าวิกฤติ คือ $F_{OUT} = F_{1,12\,(0.05)} = 4.75$

สรุปผล ปฏิเสธ H_0 นั่นคือ X_4 มีอิทธิพลต่อ Y อย่างมีนัยสำคัญ ที่ $\alpha = 0.05$ เมื่อให้ X_1 อยู่ในโมเดลแล้ว ดังนั้นจะรับเอาทั้ง X_1 และ X_4 เข้าในโมเดล

ขั้นที่ 5 ค้นหาตัวแปรอิสระตัวใหม่ใส่เข้าในโมเดลที่มี (X_1, X_4) อยู่ในโมเดลแล้ว โดยพิจารณาจากค่าสัมประสิทธิ์ตัวกำหนดบางส่วนที่ได้จากตัวอย่างที่ 7.2 ดังนี้

$$R^2_{Y2\cdot 14} = 0.0016, \qquad R^2_{Y3\cdot 14} = 0.0195$$

เนื่องจาก $R^2_{Y3\cdot 14}$ มีค่าสูงสุด ดังนั้น X_3 เป็นตัวแปรอิสระตัวถัดไปที่ถูกพิจารณาเข้าในโมเดลที่มี (X_1, X_4) อยู่ในโมเดลแล้ว

ขั้นที่ 6 ทดสอบอิทธิพลของ X_3 ที่มีต่อตัวแปรตาม Y เมื่อให้ (X_1, X_4) อยู่ในโมเดลแล้ว โดยกำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \beta_3 = 0 \quad vs. \quad H_1 : \beta_3 \neq 0$$

คำนวณ Partial F-test

$$
\begin{aligned}
F_c &= \frac{MSR(X_3 \mid X_1, X_4)}{MSE(X_1, X_3, X4)} \\
&= 0.2187
\end{aligned}
$$

ค่าวิกฤติ คือ $F_{IN} = F_{1,11\,(0.05)} = 4.84$

สรุปผล ไม่สามารถปฏิเสธ H_0 ได้ นั่นคือ X_3 ไม่มีอิทธิพลต่อ Y อย่างมีนัยสำคัญ ที่ $\alpha = 0.05$ เมื่อให้ (X_1, X_4) อยู่ในโมเดลแล้ว ดังนั้นจะหยุดการเลือกตัวแปรอิสระเข้าในโมเดล และรับเอาโมเดล

ที่มี (X_1, X_4) เป็นโมเดลสุดท้าย ได้สมการพยากรณ์ดังนี้

$$\hat{Y} = 74.3541 - 12.1859X_1 + 15.9098X_4$$

โดยมีค่า $R^2 = 0.7676$ และ $R^2_{adj} = 0.7288$

ซึ่งจะเห็นได้ว่าสอดคล้องกับโมเดลที่ได้จากวิธี Forward selection และ Backward elimination

หมายเหตุ วิธีการเลือกตัวแปรทั้ง 4 วิธี อาจให้โมเดลสุดท้ายที่แตกต่างกัน เนื่องจากตัวแปรอิสระอาจมีความสัมพันธ์ต่อกันและมีอิทธิพลต่อลำดับของการเพิ่มและตัดออกจากโมเดล ซึ่งนักวิเคราะห์บางท่านแนะนำให้ใช้วิธีในการเลือกตัวแปรทั้ง 4 วิธี แล้วพิจารณาผลที่ได้ว่าสอดคล้องกันหรือไม่ หรือพิจารณาโครงสร้างของข้อมูลที่อาจถูกมองข้ามโดยการใช้วิธีใดวิธีหนึ่งเพียงอย่างเดียวในการเลือกตัวแปร

ข้อดีของการเลือกตัวแปรด้วยวิธี Forward selection, Backward elimination และ Stepwise regression ก็คือ สะดวก รวดเร็ว และง่ายต่อการใช้งาน แต่ข้อเสียก็คือ สมการสุดท้ายที่ได้อาจไม่ใช่สมการที่ดีที่สุดที่สอดคล้องกับเกณฑ์มาตรฐานของการพิจารณาเลือกตัวแปร กรณีที่มีจำนวนตัวแปรอิสระมากเกินกว่าที่ใช้วิธี All possible regressions ได้ Montgomery และ Peck (1992) แนะนำให้ทำการเลือกตัวแปร 2 ขั้นตอน คือ คัดตัวแปรอิสระ (Screen) ในเบื้องต้นก่อน โดยใช้วิธี Forward selection, Backward elimination หรือ Stepwise regression แล้วจึงตรวจสอบตัวแปรอิสระที่ได้ด้วยวิธี All possible regressions ซึ่งนักวิเคราะห์จะต้องมีความรู้เกี่ยวกับตัวแปรที่กำลังพิจารณาประกอบด้วย นอกจากเกณฑ์ที่ต้องพิจารณาในการคัดเลือกตัวแปรแล้ว นักวิเคราะห์ควรที่จะตอบคำถามต่อไปนี้ได้

- สมการที่ได้สมเหตุสมผลหรือไม่
- สมการที่ได้มีประโยชน์ตามวัตถุประสงค์ของการศึกษาหรือไม่
- ค่าสัมประสิทธิ์ถดถอยมีขนาดและทิศทางสอดคล้องกับข้อเท็จจริงหรือไม่ และความคลาดเคลื่อนมาตรฐานของการประมาณมีค่าน้อยผิดปกติหรือไม่
- มีการตรวจสอบความเหมาะสมของโมเดลหรือไม่

ดังนั้นการเลือกตัวแปรเข้าในโมเดลควรที่จะต้องอาศัยประสบการณ์และการพิจารณาอย่างมีเหตุผลประกอบการตัดสินใจเลือกตัวแปรเสมอ

แบบฝึกหัดบทที่ 7

1. ในการศึกษาความสัมพันธ์ระหว่างตัวแปรต่อไปนี้

   Y = ราคาขายบ้าน / 1000 หน่วย

   X_1 = ภาษี / 1000 หน่วย

   X_2 = ขนาดพื้นที่ (sq ft x 1000)

   X_3 = จำนวนห้องทั้งหมด

   X_4 = จำนวนห้องนอน

   X_5 = อายุของบ้าน (ปี)

Y	X_1	X_2	X_3	X_4	X_5
25.9	4.9176	3.472	7	4	42
29.5	5.0208	3.531	7	4	62
27.9	4.5429	2.275	6	3	40
25.9	4.5573	4.050	6	3	54
29.9	5.0597	4.455	6	3	42
29.9	3.8910	4.455	6	3	56
30.9	5.8980	5.850	7	3	51
28.9	5.6039	9.520	6	3	32
35.9	5.8282	6.435	6	3	32
31.5	5.3003	4.988	6	3	30
31.0	6.2712	5.520	5	2	30
30.9	5.9592	6.666	6	3	32
30.0	5.0500	5.000	5	2	46
36.9	8.2464	5.150	8	4	50
41.9	6.6969	6.902	7	2	22
40.5	7.7841	7.102	6	2	17

1.1 จงเลือกตัวแปรอิสระด้วยวิธี All possible regressions พร้อมทั้งเขียนสมการที่ได้ในขั้น สุดท้าย

1.2 จงเลือกตัวแปรอิสระด้วยวิธี Forward selection ให้แสดงวิธีทำเป็นขั้น ๆ เมื่อกำหนดให้ค่า $\alpha = 0.10$ เป็นค่า α ที่ใช้ในการคัดเลือกตัวแปรใส่เข้าไปในสมการ พร้อมทั้งเขียนสมการที่ได้ในขั้นสุดท้าย

1.3 จงเลือกตัวแปรอิสระด้วยวิธี Backward elimination ให้แสดงวิธีทำเป็นขั้น ๆ เมื่อกำหนดให้ค่า $\alpha = 0.10$ เป็นค่า α ที่ใช้ในการตัดตัวแปรออกจากสมการ พร้อมทั้งเขียนสมการที่ได้ในขั้นสุดท้าย

1.4 จงเลือกตัวแปรอิสระด้วยวิธี Stepwise regression ให้แสดงวิธีทำเป็นขั้น ๆ เมื่อกำหนดให้ค่า $\alpha = 0.10$ เป็นค่า α ที่ใช้ในการคัดเลือกตัวแปรใส่เข้าไปในสมการ และตัดตัวแปรออกจากสมการ พร้อมทั้งเขียนสมการที่ได้ในขั้นสุดท้าย

1.5 จงเปรียบเทียบสมการที่ได้จากข้อ 1.1, 1.2, 1.3 และ 1.4 และสรุปผล

2. ข้อมูลต่อไปนี้ได้จากการศึกษาผลของธาตุเหล็ก (X_1) อลูมิเนียม (X_2) และระดับ pH ในดิน (X_3) ที่มีต่อการดูดซึมฟอสเฟตในดิน (Y)

X_1	X_2	X_3	Y
57	15	7.5	8
171	23	7.5	22
107	26	6.6	18
120	25	7.1	22
126	66	4.9	30
169	40	5.5	30
165	35	5.6	25
165	63	5.0	34
12	41	6.1	32
240	73	5.5	40
253	114	4.2	69
329	90	4.3	66
195	56	6.0	44

2.1 จงเลือกตัวแปรอิสระด้วยวิธี All possible regressions พร้อมทั้งเขียนสมการที่ได้ในขั้น สุดท้าย

2.2 จงเลือกตัวแปรอิสระด้วยวิธี Forward selection ให้แสดงวิธีทำเป็นขั้น ๆ เมื่อกำหนดให้ค่า $\alpha = 0.10$ เป็นค่า α ที่ใช้ในการคัดเลือกตัวแปรใส่เข้าไปในสมการ พร้อมทั้งเขียนสมการที่ได้ในขั้นสุดท้าย

2.3 จงเลือกตัวแปรอิสระด้วยวิธี Backward elimination ให้แสดงวิธีทำเป็นขั้น ๆ เมื่อกำหนดให้ค่า $\alpha = 0.10$ เป็นค่า α ที่ใช้ในการตัดตัวแปรออกจากสมการ พร้อมทั้งเขียนสมการที่ได้ในขั้นสุดท้าย

2.4 จงเลือกตัวแปรอิสระด้วยวิธี Stepwise regression ให้แสดงวิธีทำเป็นขั้น ๆ เมื่อกำหนดให้ค่า $\alpha = 0.10$ เป็นค่า α ที่ใช้ในการคัดเลือกตัวแปรใส่เข้าไปในสมการ และตัดตัวแปรออกจากสมการ พร้อมทั้งเขียนสมการที่ได้ในขั้นสุดท้าย

2.5 จงเปรียบเทียบสมการที่ได้จากข้อ 2.1, 2.2, 2.3 และ 2.4 และสรุปผล

3. ข้อมูลต่อไปนี้ได้จากการศึกษาผลกระทบของอุณหภูมิ (X_1) และระดับความเข้มข้น (X_2) ที่มีต่อความบริสุทธิ์ของกระบวนการทางเคมี (Y)

3.1 จงเลือกตัวแปรอิสระด้วยวิธี Forward selection ให้แสดงวิธีทำเป็นขั้น ๆ เมื่อกำหนดให้ค่า $\alpha = 0.10$ เป็นค่า α ที่ใช้ในการคัดเลือกตัวแปรใส่เข้าไปในสมการ พร้อมทั้งเขียนสมการที่ได้ในขั้นสุดท้าย

3.2 จงเลือกตัวแปรอิสระด้วยวิธี Backward elimination ให้แสดงวิธีทำเป็นขั้น ๆ เมื่อกำหนดให้ค่า $\alpha = 0.10$ เป็นค่า α ที่ใช้ในการตัดตัวแปรออกจากสมการ พร้อมทั้งเขียนสมการที่ได้ในขั้นสุดท้าย

X_1	X_2	Y
85.68	43.7	16.0
83.68	44.8	18.0
84.38	44.1	18.5
86.18	45.0	18.0
85.08	44.9	18.0
83.68	45.1	17.8
85.98	44.7	18.2
85.78	44.8	18.0
85.58	44.7	17.8
86.18	44.0	18.0
83.38	45.4	17.8
85.68	44.2	18.2
85.78	44.5	18.7
84.08	44.9	18.0

3.3 จงเลือกตัวแปรอิสระด้วยวิธี Stepwise regression ให้แสดงวิธีทำเป็นขั้น ๆ เมื่อกำหนดให้ค่า $\alpha = 0.10$ เป็นค่า α ที่ใช้ในการคัดเลือกตัวแปรใส่เข้าไปในสมการ และตัดตัวแปรออกจากสมการ พร้อมทั้งเขียนสมการที่ได้ในขั้นสุดท้าย

3.4 จงเปรียบเทียบสมการที่ได้จากข้อ 3.1, 3.2 และ 3.3 และสรุปผล