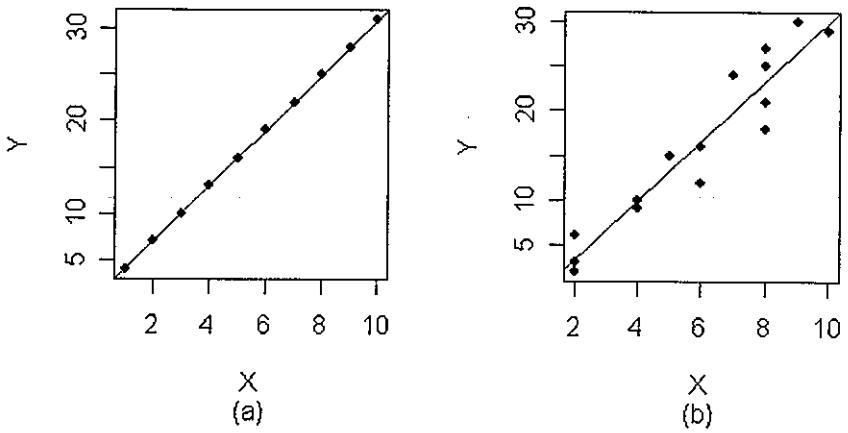
บทที่ 1

บทนำ

การวิเคราะห์การถดถอย (Regression Analysis) เป็นวิธีการทางสถิติที่ใช้ในการตรวจสอบและหาความสัมพันธ์ระหว่างตัวแปร พัฒนาโดย Sir Francis Galton ในปลายศตวรรษที่ 19 เพื่อศึกษาลักษณะความสัมพันธ์ระหว่างความสูงของพ่อกับความสูงของลูก ปัจจุบันมีการประยุกต์ใช้ปรากฏอยู่ทั่วไปในหลายสาขา ทั้งในด้านวิศวกรรมศาสตร์ วิทยาศาสตร์กายภาพ วิทยาศาสตร์สุขภาพ เศรษฐศาสตร์ การจัดการ และสังคมศาสตร์ เป็นต้น ส่วนประกอบที่สำคัญของการวิเคราะห์การถดถอยก็คือ การเก็บรวบรวมข้อมูล เนื่องจากผลสรุปที่ได้จากการวิเคราะห์ ขึ้นอยู่กับข้อมูลที่รวบรวมได้ ดังนั้นข้อมูลที่ใช้ในการวิเคราะห์การถดถอยควรจะเป็นตัวแทนที่ดีของเรื่องที่สนใจศึกษา เพื่อนำไปสู่ผลสรุปที่เชื่อถือได้

การวิเคราะห์การถดถอยเป็นการประยุกต์ของโมเดลเชิงเส้นตรง (Linear model) เพื่อหาความสัมพันธ์ระหว่างตัวแปร 2 ประเภท คือ ตัวแปรตาม (Dependent variable or response variable) นิยมแทนด้วย Y และตัวแปรอิสระ หรือตัวแปรพยากรณ์ (Independent variable or predictor or regressor) นิยมแทนด้วย X โดยมีจุดมุ่งหมายที่จะพยากรณ์ค่าของตัวแปรตาม ซึ่งเป็นตัวแปรสุ่ม (Random variable) จากตัวแปรอิสระซึ่งเป็นตัวแปรที่ทราบค่า (Known value) โดยใช้รูปแบบความสัมพันธ์ที่เรียกว่า *ฟังก์ชันการถดถอย* (Regression function) ลักษณะข้อมูลที่ใช้ในการวิเคราะห์การถดถอย สามารถแบ่งได้เป็น 2 ลักษณะใหญ่ ๆ คือ ข้อมูลที่ได้จากการสังเกต (Observational data) และข้อมูลที่ได้จากการทดลอง (Experimental data) หากค่าของตัวแปรอิสระไม่ถูกควบคุม หรือกำหนดไว้ล่วงหน้า ก่อนที่จะเก็บค่าของตัวแปรตาม Y เรียกข้อมูลลักษณะนี้ว่า ข้อมูลที่ได้จากการสังเกต แต่หากค่าของตัวแปรอิสระถูกกำหนดไว้ล่วงหน้า ก่อนที่ค่าของตัวแปรตาม Y จะถูกสังเกต เรียกข้อมูลลักษณะนี้ว่า ข้อมูลที่ได้จากการทดลอง ในการศึกษาหลาย ๆ สาขาพบว่า บ่อยครั้งที่ไม่สามารถควบคุมค่าของตัวแปรอิสระได้ ดังนั้นข้อมูลที่เก็บรวบรวมเพื่อประยุกต์ใช้กับการวิเคราะห์การถดถอยส่วนใหญ่จึงเป็นข้อมูลที่ได้จากการสังเกต และการอนุมานที่เกิดจากการใช้ข้อมูลที่ได้จากการสังเกตจะมีข้อจำกัดมากกว่าการใช้ข้อมูลที่ได้จากการทดลอง โดยข้อมูลที่ได้จากการสังเกตมักไม่แสดงถึงลักษณะความสัมพันธ์ที่เป็นเหตุเป็นผลต่อกันอย่างชัดเจน

โดยทั่วไปมักจะแสดงความสัมพันธ์ระหว่างตัวแปรอยู่ในรูป $Y = f(X)$ ในที่นี้จะแบ่งความสัมพันธ์ระหว่างตัวแปรออกเป็น 2 ลักษณะใหญ่ ๆ คือ



รูปที่ 1.1: แผนภาพแสดง (a) ความสัมพันธ์ในเชิงฟังก์ชัน (b) ความสัมพันธ์ในเชิงสถิติ

- ความสัมพันธ์ในเชิงฟังก์ชัน (Funtional relation) มีลักษณะที่สำคัญ คือ ค่าสังเกตทุกค่าจะตกอยู่บนเส้นที่แสดงความสัมพันธ์โดยตรง ดังแสดงในรูปที่ 1.1(a)
- ความสัมพันธ์ในเชิงสถิติ (Statistical relation) จะแตกต่างจากความสัมพันธ์ในเชิงฟังก์ชันในแง่ที่ค่าสังเกตทุกค่าไม่จำเป็นที่จะต้องตกอยู่บนเส้นที่แสดงความสัมพันธ์ แต่ค่าสังเกตเหล่านี้มีการกระจายอยู่รอบ ๆ เส้นที่แสดงความสัมพันธ์ ดังแสดงในรูปที่ 1.1(b) ซึ่งการวิเคราะห์การถดถอยจัดเป็นการศึกษาความสัมพันธ์ในเชิงสถิติ มีลักษณะที่สำคัญ คือ แนวโน้มที่ตัวแปรตาม Y ผันแปรตามตัวแปรอิสระ X เกิดขึ้นอย่างมีระบบ และค่าสังเกตมีการกระจายอยู่รอบเส้นแสดงความสัมพันธ์ระหว่างตัวแปรคู่ดังกล่าว

ตัวอย่างของการวิเคราะห์การถดถอย

- ความสัมพันธ์ระหว่างยอดขายและค่าใช้จ่ายในการโฆษณาของบริษัท
- ความสัมพันธ์ระหว่างรายจ่ายและรายได้
- ความสัมพันธ์ระหว่างความสูงของลูกและความสูงของพ่อ
- ความสัมพันธ์ระหว่างน้ำหนักที่เพิ่มขึ้นกับปริมาณอาหารเสริมที่ให้แก่เด็กวัยก่อนเรียน
- ความสัมพันธ์ระหว่างเกรดเฉลี่ยของนักศึกษาชั้นปีที่ 1 และคะแนนเอ็นทรานซ์

การศึกษาความสัมพันธ์ระหว่างตัวแปรในโมเดลเชิงเส้นตรงเมื่อมีตัวแปรอิสระเพียง 1 ตัว ดังตัวอย่างข้างต้น เรียกว่า *โมเดลถดถอยเชิงเส้นตรงอย่างง่าย* (Simple linear regression model) ซึ่งสามารถแสดงได้ดังนี้

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1.1}$$

เมื่อ β_0 และ β_1 แทน พารามิเตอร์ที่ไม่ทราบค่า และ ϵ แทน ค่าความคลาดเคลื่อน (Error) ที่เกิดจากการใช้สมการถดถอยอธิบายความสัมพันธ์ในข้อมูล โดยทั่วไปแล้วตัวแปรตามมักจะมีความสัมพันธ์กับตัวแปรอิสระมากกว่า 1 ตัว เช่น การศึกษาความสัมพันธ์ระหว่างยอดขายของบริษัทกับค่าใช้จ่ายในการโฆษณา และต้นทุนการผลิต เป็นต้น เรียกโมเดลที่แสดงความสัมพันธ์ดังกล่าวนี้ว่า *โมเดลถดถอยเชิงเส้นตรงแบบพหุ* (Multiple linear regression model) มีรูปแบบดังนี้

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon \tag{1.2}$$

ซึ่งลักษณะความสัมพันธ์เชิงเส้นตรง สามารถแบ่งได้เป็น 2 ประเภท ดังนี้

- ความสัมพันธ์เชิงเส้นตรงในตัวแปรอิสระ (Linear in the predictors) โดยตัวแปรอิสระทุกตัวมีกำลังสูงสุดเป็นหนึ่ง บางครั้งอาจเรียกโมเดลประเภทนี้ว่า โมเดลกำลังหนึ่ง (First-order model)
- ความสัมพันธ์เชิงเส้นตรงในพารามิเตอร์ (Linear in the parameters) โดยพารามิเตอร์ทุกตัวมีกำลังสูงสุดเป็นหนึ่ง และไม่มีพารามิเตอร์ตัวใดที่ติดอยู่ในรูป ผลคูณ ผลหาร หรือเอกซ์โปเนนท์ ของพารามิเตอร์ตัวอื่น

คำว่า *เชิงเส้นตรง* ที่ใช้ในการวิเคราะห์การถดถอยในที่นี้ หมายถึง โมเดลเชิงเส้นตรงในเทอมของพารามิเตอร์ $\beta_1, \beta_2, \ldots, \beta_k$ ไม่ใช่ตัวแปรตาม Y เป็นฟังก์ชันเชิงเส้นตรงของตัวแปรอิสระ X อย่างที่มักจะเข้าใจกัน

ตัวอย่างของโมเดลถดถอยเชิงเส้นตรง

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$
$$Y = \beta_0 + \beta_1 e^X + \epsilon$$
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \epsilon$$

สิ่งสำคัญประการหนึ่งของการวิเคราะห์การถดถอยก็คือ การประมาณพารามิเตอร์ที่ไม่ทราบค่าในโมเดลถดถอย โดยสมการที่ได้เป็นเพียงการประมาณความสัมพันธ์ที่แท้จริงระหว่างตัวแปร หลังจากที่ได้สมการแสดงความสัมพันธ์ในข้อมูลแล้ว ควรมีการตรวจสอบความเหมาะสมและคุณภาพของสมการดังกล่าว ซึ่งอาจนำไปสู่การรับเอาสมการที่สร้างขึ้น หรือปรับปรุงสมการเดิม เพื่อหารูปแบบที่เหมาะสมมากกว่า

นอกจากนี้การวิเคราะห์การถดถอยไม่ได้เป็นการหาความสัมพันธ์ในลักษณะที่เป็นเหตุเป็นผล (Cause-effect relationship) ระหว่างตัวแปรโดยตรง เนื่องจากลักษณะความสัมพันธ์ที่เป็นเหตุเป็นผลจะต้องถูกกำหนดโดยที่ไม่คำนึงถึงตัวข้อมูล ซึ่งอาจต้องอาศัยทฤษฎีที่เกี่ยวข้องช่วยในการกำหนดความเป็นเหตุเป็นผล เช่น ในการศึกษาความสัมพันธ์ระหว่างความเร็วในการเขียน (Y) และจำนวนคำศัพท์ที่เด็กอายุระหว่าง 5-10 ปีจำได้ แม้ว่าสมการถดถอยจะแสดงความสัมพันธ์เชิงบวกระหว่างตัวแปร แต่ไม่ได้หมายความว่า การเพิ่มจำนวนคำศัพท์เพียงอย่างเดียวจะทำให้เด็กเขียนเร็วขึ้น ในที่นี้อาจมีตัวแปรอิสระอื่น ๆ อาทิเช่น อายุและระดับการศึกษา ซึ่งล้วนแล้วแต่มีอิทธิพลต่อทั้งจำนวนคำศัพท์ที่เด็กจำได้และความเร็วในการเขียนทั้งสิ้น ดังนั้นเราอาจจะพิจารณาการวิเคราะห์การถดถอยว่าเป็นวิธีที่ช่วยยืนยันความสัมพันธ์ที่เป็นเหตุเป็นผลได้ แต่ไม่ใช่เป็นการแสดงความสัมพันธ์ดังกล่าวโดยตรง

จุดมุ่งหมายของการใช้โมเดลถดถอย

- อธิบายลักษณะของข้อมูล ซึ่งจะเห็นได้ว่าบ่อยครั้งที่มีการใช้สมการเพื่ออธิบายและสรุปลักษณะของข้อมูลที่มีอยู่ และมักให้ผลที่ชัดเจนกว่าการใช้ตารางหรือกราฟเพียงอย่างเดียว
- ประมาณค่าพารามิเตอร์ที่ไม่ทราบค่า เพื่อสร้างสมการแสดงความสัมพันธ์ของข้อมูล
- การพยากรณ์และการประมาณค่า ซึ่งการประยุกต์ใช้สมการถดถอยส่วนใหญ่จะเกี่ยวข้องกับการพยากรณ์ค่าของตัวแปรตาม โดยที่ค่าพยากรณ์เหล่านี้อาจจะเป็นประโยชน์ต่อการวางแผนของหน่วยงานในอนาคต แต่อย่างไรก็ตามควรระลึกไว้เสมอว่า การใช้สมการถดถอยเพื่อการพยากรณ์ จะทำได้เฉพาะในกรณีที่ค่าของตัวแปรอิสระอยู่ในขอบเขตของข้อมูลที่ใช้ในการสร้างสมการ บางครั้งถึงแม้ว่ารูปแบบของสมการถูกต้อง หากการประมาณค่าพารามิเตอร์ในโมเดลไม่เหมาะสมแล้ว ก็อาจส่งผลต่อประสิทธิภาพของการพยากรณ์ได้เช่นกัน
- การควบคุม เป็นการประยุกต์ใช้สมการถดถอยเพื่อหาระดับที่เหมาะสมของค่าของตัวแปรที่จะทำให้ได้ผลผลิตหรือมีประสิทธิภาพสูงสุด เมื่อสมการถดถอยถูกใช้เพื่อจุดมุ่งหมายของการควบคุม สิ่งสำคัญที่ควรพิจารณา ก็คือ ตัวแปรเหล่านี้ควรที่จะเกี่ยวข้องกันในลักษณะที่เป็นเหตุเป็นผลต่อกัน

แบบฝึกหัดบทที่ 1

1. ในการศึกษาความสัมพันธ์ระหว่างยอดขายและจำนวนสินค้าที่ขาย พบว่าสามารถเก็บข้อมูลของจำนวนสินค้าได้อย่างถูกต้อง ในขณะที่การจดบันทึกยอดขายโดยพนักงานมักจะมีความผิดพลาด ถ้าท่านต้องการสร้างสมการแสดงความสัมพันธ์ระหว่างตัวแปรทั้งสอง

   1.1 จงระบุอะไรคือตัวแปรตามและตัวแปรอิสระ

   1.2 ลักษณะความสัมพันธ์ที่ได้จัดเป็นความสัมพันธ์ในเชิงฟังก์ชันหรือความสัมพันธ์ในเชิงสถิติ เพราะเหตุใด

2. ในแต่ละข้อที่กำหนดให้ต่อไปนี้ จัดเป็นโมเดลถดถอยเชิงเส้นตรงหรือไม่

   2.1 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1^2 X_2^2 + \epsilon$

   2.2 $Y = \beta_0 + e^{\beta_1 X_1 + \beta_2 X_2} + \epsilon$

   2.3 $Y = \beta_1 \log X_1 + \beta_2 \log X_2 + \epsilon$

   2.4 $Y = \dfrac{e^{\beta_0 + \beta_1 X + \epsilon}}{1 + e^{\beta_0 + \beta_1 X + \epsilon}}$

   2.5 $\ln Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

3. นิสิตคนหนึ่งเขียนโมเดลถดถอยเชิงเส้นตรงอย่างง่าย ดังนี้ $E(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i$ ท่านคิดว่าโมเดลดังกล่าวถูกต้องหรือไม่ เพราะเหตุใด

4. จงยกตัวอย่างเหตุการณ์ที่ใช้ในการวิเคราะห์การถดถอย โดยระบุตัวแปรตามและตัวแปรอิสระให้ชัดเจน