6609655359 ศิรดา บุญหมื่น

Unit9 มีการพูดถึงหัวข้อหลักที่สำคัญเกี่ยวกับการวิเคราะห์ข้อมูลที่มีการ "collinear" หรือลักษณะความสัมพันธ์เชิงเส้นที่สูงในตัวแปรอิสระในการวิเคราะห์การถดถอย (Regression Analysis)

- ความหมายของ "Collinearity" หมายถึง Predictor Variables ใน multiple regression มีความสัมพันธ์เชิงเส้นที่สูงมากระหว่างกัน มีผลให้ค่าของตัวแปรอิสระบางตัวในแบบจำลองไม่สามารถอธิบายได้อย่างชัดเจน เมื่อตัวแปรมีความสัมพันธ์กันมากเกินไป จะทำให้การตีความค่าความชันของสมการถดถอยทำได้ยาก เพราะการเปลี่ยนแปลงในตัวแปรหนึ่งอาจทำให้เกิดการเปลี่ยนแปลงในตัวแปรอื่นโดยไม่สามารถบ่งชี้ตัวแปรเหล่านั้นเป็นอิสระต่อกันได้ ในกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันอย่างสมบูรณ์ ตัวแปรพยากรณ์ดังกล่าวจะเรียกว่า "Orthogonal" ซึ่งหมายถึงไม่มีความสัมพันธ์เชิงเส้นกันเลย แต่ในสถานการณ์ส่วนใหญ่ในชีวิตจริง การที่ตัวแปรอิสระเป็นอิสระจากกันอย่างสมบูรณ์นั้นหาได้ยาก

- ผลกระทบของ Collinearity

- ค่าสัมประสิทธิ์การถดถอย: ไม่เสถียร ค่าสัมประสิทธิ์การถดถอยอาจเปลี่ยนแปลงอย่างมาก เมื่อมีการเพิ่มหรือลบตัวแปรใดๆ ออกจากแบบจำลอง แม้ว่าตัวแปรที่เพิ่มหรือลบจะไม่มีความสำคัญมากก็ตาม

- ข้อสรุปจากแบบจำลองที่ใช้ค่าสัมประสิทธิ์ที่ได้จากข้อมูลที่มี collinearity อาจทำให้การพยากรณ์ผิดได้

- การตรวจพบ Collinearity

ในการตรวจหาปัญหานี้ ผู้เขียนเสนอการใช้ตัวชี้วัดต่างๆ เช่น : - ค่าสัมประสิทธิ์สหสัมพันธ์ : หากพบว่าค่าสหสัมพันธ์ระหว่างตัวแปรอิสระคู่ใดสูงเกินไป แสดงว่ามีความเสี่ยงที่จะเกิด collinearity

- Variance Inflation Factor (VIF) : ถ้าค่า VIF สูงเท่า 10 จะเป็นสัญญาณว่าข้อมูลอาจมีปัญหา collinearity

- Eigenvalues : การตรวจหาค่าที่มีค่าน้อยมากหรือใกล้ 0 ช่วยบ่งชี้มีปัญหา Collinearity

• วิธีแก้ปัญหา Collinearity มีการเสนอวิธีแก้ไขปัญหา collinearity ดังนี้

① "เพิ่มข้อมูล" การเติมข้อมูลเพิ่มเติมอาจช่วยลดปัญหานี้ได้ แต่ก็ไม่สามารถทำได้เสมอไป เนื่องจากข้อจำกัดด้านงบ

② "ลดจำนวนตัวแปร" สัดส่วนที่มีความสัมพันธ์สูงต่อกันออกจากโมเดล ก็สามารถช่วยแก้ปัญหานี้ได้ในบางกรณี

③ "การใช้วิธีทางสถิติอื่นๆ" เช่น PCA วิธีนี้จะช่วยแก้ปัญหา collinearity โดยการปรับแต่งในเสียร

④ "Ridge Regression" เป็นอีกวิธีหนึ่งที่ช่วยลดปัญหา Collinearity โดยการเพิ่มอคติในโมเดล ซึ่งทำต้องแปรที่มีความสัมพันธ์สูงมีผลกระทบต่อการคำนวณน้อยลง

★ ผู้เขียนเน้นว่าปัญหา collinearity ไม่ใช่ข้อผิดพลาดในการสร้างแบบจำลอง แต่เป็นข้อจำกัดของข้อมูลที่นำมาใช้ ดังนั้นวิธีแก้ไขที่ดีที่สุดควรเน้นไปที่การปรับแต่งข้อมูลเพิ่มเติม เข้ามา

ตัวอย่างที่ผู้เขียนใช้ในหนังสือเป็นการทำนายข้อมูลเศรษฐกิจของประเทศฝรั่งเศสโดยใช้ตัวแปร เช่น การผลิตในประเทศ domestic production การนำเข้าในประเทศ และสินค้าคงคลัง ในช่วงปี 1949–1966 เมื่อตัวแปรเหล่านี้มี "collinearity" กันอย่างรุนแรง การคำนวณพยากรณ์จะได้ค่าที่ผิดเพี้ยนไปจากที่ควรจะเป็น แม้จะมีค่า "R^2" สูงถึง 0.99 แต่การพยากรณ์ก็ไม่น่าเชื่อถือ

หลักการใช้ Principal Component Analysis (PCA) และ Ridge Regression ในการแก้ไขปัญหาความสัมพันธ์เชิงเส้น (Collinearity) รวมถึงข้อจำกัดและข้อควรระวังในการใช้วิธีเหล่านี้

★ การวิเคราะห์องค์ประกอบหลัก (PCA)

PCA เป็นวิธีการที่ใช้ในการเปลี่ยนตัวแปรต้นฉบับที่มีความสัมพันธ์กัน (collinear) ให้กลายเป็นชุดของตัวแปรใหม่ที่ไม่มีความสัมพันธ์เชิงเส้นต่อกัน ซึ่งเราเรียกตัวแปรเหล่านี้ว่า "องค์ประกอบหลัก" (PC) โดยใช้ตัวแปรใหม่นี้จะเป็นการรวมเชิงเส้นของตัวแปรต้นฉบับ

ขั้นตอนการทำงานของ PCA :

① คำนวณ Eigenvalues และ Eigenvectors : จากเมทริกซ์ความสัมพันธ์ (correlation matrix) ของตัวแปรต้นฉบับ โดยตัวแปรใหม่แต่ละตัวจะถูกสร้างขึ้นมาจาก Eigenvectors เหล่านี้

② เลือกจำนวนองค์ประกอบหลัก (PC) : ตัวแปรใหม่เหล่านี้จะถูกจัดเรียงตามลำดับความสำคัญ โดยองค์ประกอบหลักแรกๆ มีความแปรปรวนมากที่สุด ดังนั้นเราสามารถเลือกใช้เฉพาะองค์ประกอบที่สำคัญ เนื่องจากที่ตัวแปรต้นฉบับที่มีปัญหา Collinearity

③ นำองค์ประกอบหลักมาวิเคราะห์ : หลังจากที่ได้ชุดขององค์ประกอบหลักที่ไม่มีความสัมพันธ์กันแล้ว จะนำชุดข้อมูลตัวแปรนี้ไปใช้ในการสร้างแบบจำลองใหม่

★ ข้อดีของ PCA

- ช่วยแก้ปัญหา collinearity ได้อย่างมีประสิทธิภาพ
- ลดจำนวนตัวแปรในแบบจำลองลงได้ ทำให้แบบจำลองง่ายต่อการตีความ
- ปรับปรุงความเสถียรของค่าความสัมพันธ์ในแบบจำลองการถดถอย

★ ข้อจำกัดของ PCA :

- การตีความที่ยากขึ้น
- ข้อมูลที่สูญหาย

★ Ridge Regression → เป็นอีกหนึ่งวิธีที่นิยมใช้ในการจัดการกับปัญหา Collinearity โดยการเพิ่มพารามิเตอร์ควบคุมในแบบจำลองถดถอย (regression model) ทำให้ตัวแปรอิสระมีความสัมพันธ์กันส่งผลกระทบต่อการคำนวณค่าถดถอยน้อยลง

หลักการทำงานของ Ridge Regression :

① เพิ่มพารามิเตอร์ควบคุม ② เลือกค่าพารามิเตอร์ k ที่เหมาะสม

การเลือกใช้วิธี Ridge Regression และ PCA ขึ้นอยู่กับลักษณะของข้อมูลและวัตถุประสงค์ของการวิเคราะห์ : ถ้าเน้นการแก้ปัญหา Collinearity และยังต้องการรักษาตัวแปรอิสระไว้หมดในแบบจำลอง Ridge Regression จะเป็นทางเลือกที่ดีกว่า

: ถ้าต้องการลดจำนวนตัวแปรและเน้นการวิเคราะห์ที่ง่ายต่อการจัดการ PCA เหมาะกว่า

Summary

การแก้ปัญหา Collinearity เป็นส่วนสำคัญในการวิเคราะห์การถดถอย โดย PCA และ Ridge Regression เป็นวิธีการที่ใช้ได้ผลดีในการจัดการกับปัญหานี้ ขึ้นอยู่กับลักษณะข้อมูลและเป้าหมายในการวิเคราะห์ ผู้ใช้งานต้องเลือกใช้วิธีที่เหมาะสมเพื่อให้ได้แบบจำลองที่มีความเสถียรและแม่นยำมากที่สุด

Regression Analysis by example จะมุ่งเน้นที่การทำงานกับข้อมูลที่มีปัญหา

Unit 10

* Collinearity * และขยายความเกี่ยวกับการใช้วิธีการต่างๆ ในการจัดการปัญหานี้ โดยในบทที่ 10 จะเจาะลึกในวิธีการสองแบบที่นิยมใช้ ได้แก่ Principal Component Regression (PCR) และ Ridge Regression ซึ่งเป็นเทคนิคหลักในการจัดการกับปัญหา collinearity

การทำงานกับข้อมูลที่มี Collinearity (Unit 10) ผู้เขียนได้เสนอวิธีการเพิ่มเติมเกี่ยวกับการทำงานกับข้อมูลที่มีปัญหา * collinearity * โดยเฉพาะการใช้วิธี <PCR> และ Ridge Regression

ความสำคัญของการจัดการกับปัญหา Collinearity เป็นปัญหาสำคัญที่พบได้บ่อยในแบบทดลองการถดถอยเชิงพหุ (Multiple Regression) เนื่องจากตัวแปรอิสระหลายตัวในข้อมูลมักมีความสัมพันธ์กัน ซึ่งอาจให้ค่าสัมประสิทธิ์ของตัวแปรเหล่านั้นไม่มีความเสถียร นำไปสู่การตีความที่ผิดพลาด และการทำนายที่ไม่แม่นยำ การแก้ไขปัญหานี้จึงมีความสำคัญในการปรับปรุงความถูกต้องของแบบจำลอง โดยวิธีการต่างๆ เช่น Principal Component Regression และ Ridge Regression จะช่วยลดปัญหา Collinearity และเพิ่มความเสถียรของค่าสัมประสิทธิ์ในแบบจำลอง

สรุป การวิเคราะห์ข้อมูลที่มีปัญหา Collinearity เป็นกระบวนการที่ต้องใช้เครื่องมือที่เหมาะสมเพื่อให้ได้ผลลัพธ์ที่ถูกต้อง และเชื่อถือได้ (PCA) และ Ridge Regression เป็นสองวิธีที่สามารถจัดการกับปัญหานี้ได้อย่างมีประสิทธิภาพ การเลือกใช้วิธีใดวิธีหนึ่งควรขึ้นอยู่กับเป้าหมายของการวิเคราะห์ ความซับซ้อนของข้อมูล และความสามารถในการตีความผลลัพธ์

* ทางเลือกอื่นในการจัดการกับ Collinearity นอกจาก Principal Componant Regression (PCR) และ Ridge Regression และยังมีวิธีอื่นที่สามารถใช้ในการจัดการกับปัญหา Collinearity เช่น ① Lasso Regression ② Elastic Net Regression ③ Partial Least Squares Regression

สรุปจากหนังสือ PCR และ Ridge Regression เป็นสองวิธีที่นิยมใช้ในการแก้ปัญหา Collinearity

- PCR เหมาะสำหรับกรณีที่ต้องลดจำนวนตัวแปรและลดผลกระทบของ Collinearity โดยใช้ตัวแปรใหม่ที่สร้างขึ้นจากองค์ประกอบหลัก

- Ridge Regression เหมาะสำหรับกรณีที่ต้องการรักษาตัวแปรอิสระทั้งหมดไว้ในแบบจำลองและลดผลกระทบจาก Collinearity โดยใช้ตัวแปรใหม่ที่สร้างขึ้นจากองค์ประกอบหลัก

* Ridge Regression เหมาะสำหรับกรณีที่ต้องการรักษาตัวแปรอิสระทั้งหมดไว้ในแบบจำลองและลดผลกระทบจาก Collinearity โดยการเพิ่มพารามิเตอร์ควบคุม

• นอกจากนี้ยังมีวิธีการทางเลือกอื่นๆ เช่น Lasso Regression, Elastic Net และ PLS ที่สามารถใช้ในการแก้ปัญหา collinearity ได้เช่นกัน

การประเมินผลลัพธ์ของโมเดลมีความสำคัญอย่างมาก ซึ่งสามารถทำได้โดยการแบ่งชุดข้อมูลทดสอบการใช้ * Cross-validation และการวิเคราะห์ Residual

การใช้เทคนิคที่เหมาะสมในการแก้ปัญหา Collinearity จะช่วยทำให้โมเดลมีความแม่นยำและเสถียรมากขึ้น รวมถึงการนำไปใช้งานในการพยากรณ์หรือการวิเคราะห์ข้อมูลเชิงธุรกิจ