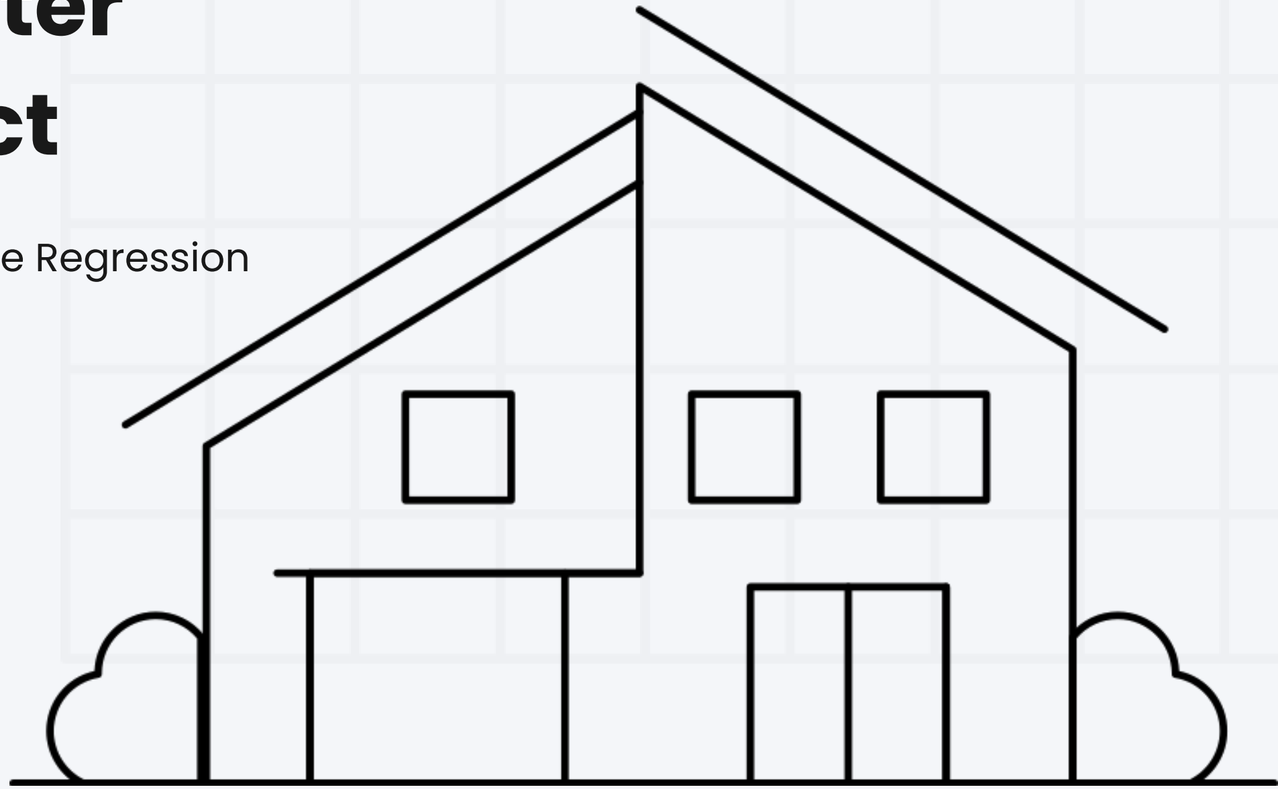
Ames, Iowa : Alternative to the Boston Housing Data as an End of Semester Regression Project

การประยุกต์ Multiple Linear Regression และ Ridge Regression



Presented by Chutimon Nunthasena 6640201080

อ้างอิงแนวคิดจาก De Cock (2011)

บทนำ

การวิเคราะห์ราคาบ้านนิยมใช้ Multiple Linear Regression เพราะช่วยอธิบายความสัมพันธ์ระหว่างราคาบ้านกับปจจัยต่าง ๆ ได้ชัดเจน เดิมนิยมใช้ Boston Housing Data Set แต่ข้อมูลมีขนาดเล็กและล้าสมัย จึงเปลี่ยนมาใช้ Ames Housing Data Set (De Cock, 2011) ทีมีข้อมูลมากและสมเหมาะกว่า อย่างไรก็ตาม ข้อมูลทีมีตัวแปรจำนวนมากทำให้เกิดปญหาความสัมพันธ์กันสูง (multicollinearity) งานนีจึงใช้ Multiple Linear Regression ใช้ Ordinary Least Squares (OLS) เปนโมเดลพืนฐาน และเปรียบเทียบกับ Ridge Regression ทีใช้ Regularization techniques เพือเพิมเสถียรภาพของโมเดล โดยประเมินผลด้วยวิธีการตรวจสอบความถูกต้อง (RMSE, MAE, R^2)

วัตถุประสงค์หลักของงาน

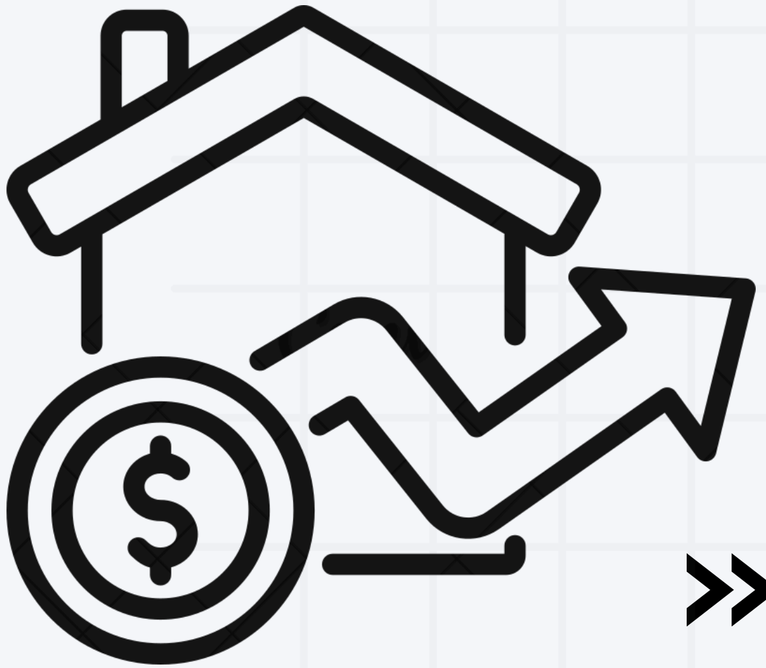
- เพื่อศึกษาการใช้ Multiple Linear Regression กับข้อมูลราคาบ้านที่มีตัวแปรจำนวนมาก
- เพื่อแสดงข้อจำกัดของ Ordinary Least Squares (OLS) เมื่อข้อมูลมีปัญหาตัวแปรอิสระที่มีความสัมพันธ์กันสูง (multicollinearity)
- เพื่อสาธิตการใช้ Ridge Regression เป็นแนวทางแก้ไขปัญหาสัมประสิทธิ์ไม่เสถียร
- เพื่อเปรียบเทียบประสิทธิภาพของโมเดลโดยใช้ตัวชี้วัดมาตรฐาน(RMSE, MAE, R^2)
- เพื่อใช้เป็น case study ทางการศึกษาแทน Boston Housing Data

"งานศึกษานี้มุ่งเน้นการทำความเข้าใจแนวคิดของแบบจำลองถดถอยเชิงเส้น ความเสถียรของค่าสัมประสิทธิ์ และการประยุกต์ใช้ regularization techniques"

เปเปอร์หลัก (De Cock, 2011) ทำอะไร?

เปเปอร์หลัก:
Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project
แนวคิดหลัก:

- เสนอ Ames Housing Data เป็น dataset เพื่อการเรียนการสอนแทน Boston
- เน้นการสาธิตแนวคิด Multiple Linear Regression ในบริบทของข้อมูลจริง
- มุ่งเน้นอธิบายความสัมพันธ์ของตัวแปร
- เหตุผลที่ไม่ใช้ Boston:
  - ข้อมูลเก่า (Boston ใช้ข้อมูลช่วงปี 1970s)
  - ตัวแปรน้อย
  - ไม่เหมาะกับข้อมูลมิติสูง
- เปลี่ยนใช้ Ames Housing แทน:
  - ข้อมูลจากธุรกรรมจริง
  - ตัวแปรมเยอะ
  - เหมาะสำหรับสอน regression ขั้นสูง



เปเปอร์นี้เพิ่มอะไร?

สิ่งที่เพิ่มเข้าไป:

- ใช้กรอบแนวคิดเดียวกับ De Cock (2011)
  - สร้าง baseline model (OLS) อย่างเป็นระบบ
  - เพิ่ม Ridge Regression + 10-fold Cross-Validation
  - ใช้ train–test split (70/30) เพื่อประเมินโมเดล
- รายงานตัวชี้วัดเชิงตัวเลข:
  - RMSE (Root Mean Squared Error)
  - MAE (Mean Absolute Error)
  - R^2 (R-Squared)
- แสดงกราฟ:
  - Cross-validation curve
  - Predicted vs Actual

แหล่งข้อมูล

ชุดข้อมูลที่ใช้

- Ames Housing Data Set
- ข้อมูลราคาบ้านจริงในเมือง Ames, Iowa
- ประกอบด้วย:
  - ตัวแปรเชิงปริมาณ (พื้นที่, ปีที่สร้าง, จำนวนโรงรถ ฯลฯ)
  - ตัวแปรเชิงหมวดหมู่ (ย่านที่อยู่อาศัย, การแบ่งเขตพื้นที่)

เหตุผลที่เลือกชุดข้อมูลนี้ :

- ข้อมูลเริ่มตั้งแต่ ปี ค.ศ. 2006 – 2010
- ตัวแปร 82 ตัว
- ข้อมูล 2,930 แถว
- เหมาะกับการศึกษาปัญหา ตัวแปรสัมพันธ์กันสูง

การเตรียมข้อมูล

ขั้นตอนการเตรียมข้อมูล

1.ลบตัวแปรบ้างตัวที่ไม่จำเป็น

2.จัดการ missing values

- ตัวแปรเชิงตัวอักษร → แทนด้วย None
- ตัวแปรเชิงตัวเลข → แทนด้วย median

3.แปลงตัวแปรเชิงหมวดหมู่

4.กำหนดตัวแปรเป้าหมาย: “price”

เหตุผล :

- ลด bias จาก outliers
- รองรับ regression ที่ต้องใช้ตัวเลข

โมเดลที่ใช้

Multiple Linear Regression

Ordinary Least Squares (OLS) เป็นวิธีการประมาณค่าสัมประสิทธิ์ของสมการถดถอยเชิงเส้นพหุคูณ โดยมีวัตถุประสงค์เพื่อหาค่าพารามิเตอร์ที่ ทำให้ผลรวมกำลังสองของความคลาดเคลื่อนระหว่างค่าจริงและค่าที่โมเดลทำนายได้มีค่าน้อยที่สุด

สมการ

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$$

- yi = ราคาบ้านหลังที่ i
- β0 = ค่าคงที่
- βj = ค่าสัมประสิทธิ์ของตัวแปรอิสระตัวที่ j
- xij = ค่าของตัวแปรอิสระตัวที่ j สำหรับบ้านหลังที่ i
- εi = ค่าความคลาดเคลื่อน
- p = จำนวนตัวแปรอิสระ

OLS ทำการประมาณค่าสัมประสิทธิ์โดยแก้ปัญหาการหาค่าต่ำสุดของฟังก์ชันสูญเสีย

ปัญหาของ OLS ในข้อมูลมิติสูง

- ตัวแปรจำนวนมาก → ตัวแปรสัมพันธ์กันสูง
- ผลกระทบ:
  → ค่าสัมประสิทธิ์จะไม่เสถียร, ความแม่นยำในการทำนายลดลง, overfitting ได้ง่าย

Ridge Regression

Regularization techniques เป็นการขยายแนวคิดของ OLS โดยการเพิ่ม penalty term เข้าไปในฟังก์ชันสูญเสีย เพื่อควบคุมขนาดของค่าสัมประสิทธิ์ และลดปัญหาที่เกิดจากตัวแปรสัมพันธ์กันสูง

สมการ

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

- min_β คือ หาค่าสัมประสิทธิ์ของโมเดลที่ทำให้ฟังก์ชัน loss มีค่าน้อยที่สุด
- yi = ราคาบ้านหลังที่ i
- y^i = ค่าที่โมเดลทำนายได้
- βj = ค่าสัมประสิทธิ์ของตัวแปรอิสระตัวที่ j
- λ = regularization parameter ใช้ควบคุมความแรงของ penalty
- p = จำนวนตัวแปรอิสระ
- n = จำนวนข้อมูล

หน้าที่

ลดผลกระทบจากตัวแปรสัมพันธ์กันสูง ทำให้ค่าสัมประสิทธิ์ไม่เหวี่ยง หรือผันผวนรุนแรง เพิ่มความเสถียรของโมเดล และเหมาะกับข้อมูลที่มีตัวแปรจำนวนมาก

ข้อดี

- Ridge ไม่ตัดตัวแปรทิ้ง แต่จะลดขนาดค่าสัมประสิทธิ์ให้เหมาะสมคงโครงสร้างโมเดล ไว้ทั้งหมด

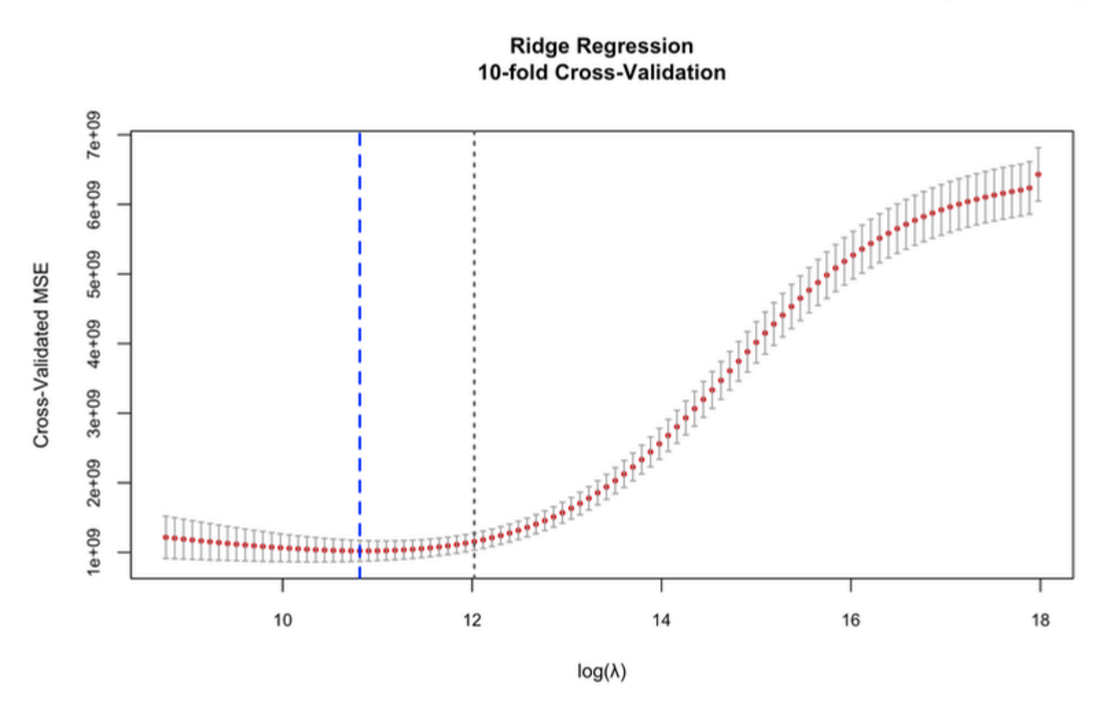
การเลือกค่า λ ด้วย Cross-Validation

วิธีเลือก λ

- ใช้ 10-fold Cross-Validation
- แบ่ง training set เป็น 10 ส่วน ใช้ 9 ส่วนฝึก + 1 ส่วนทดสอบ ทำซ้ำจนครบ 10 ครั้ง
- วัดค่า Mean Squared Error (MSE)
    - การเลือกค่า λ ด้วย Cross-Validation ช่วยควบคุมความซับซ้อนของโมเดลและลดความเสี่ยงของ overfitting ซึ่งเป็นปัญหาที่พบบ่อยใน regression ที่มีตัวแปรจำนวนมาก

ค่าที่พิจารณา:

- lambda.min → MSE ต่ำสุด ใช้เมื่อเน้นความถูกต้อง
- lambda.1se → โมเดลเรียบง่ายกว่าแต่เสถียร



ข้อมูลที่ได้จากกราฟ

พบว่า Cross-validation curve แสดงความสัมพันธ์ระหว่าง log(λ) กับ MSE จุดต่ำสุดที่ได้ คือ

- log(λ_min) = 10.81
- log(λ_1se) = 12.02

ตารางผลลัพธ์

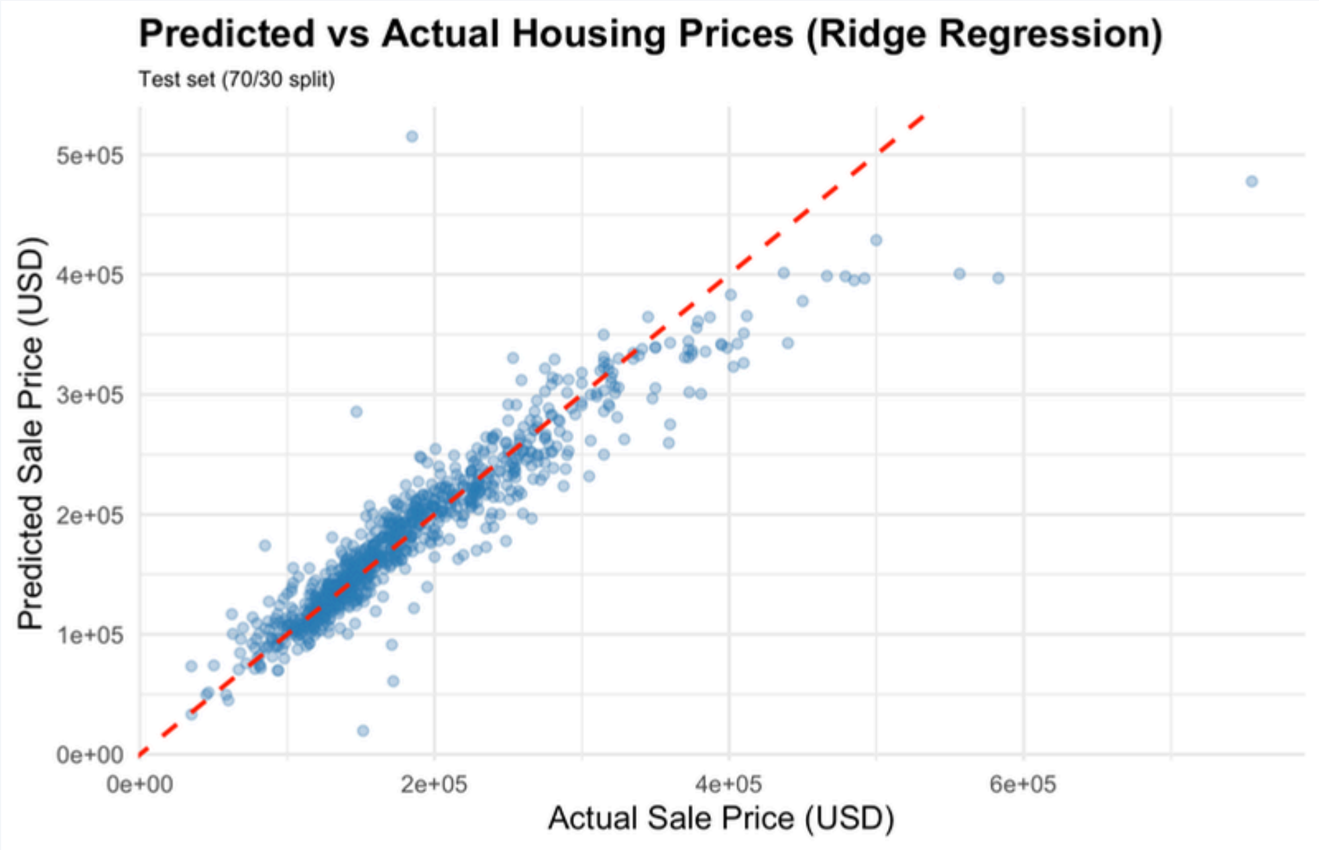
ตัวชี้วัด	De Cock (2011)	Multiple Linear Regression (OLS)	Ridge Regression
R^2	0.71	0.83	0.87
RMSE	ไม่รายงาน	32,160	28,986
MAE	ไม่รายงาน	21,426	17,238
การประเมิน	In-sample	Test set	Test set

ผลการวิเคราะห์

- Ridge ให้ RMSE และ MAE ต่ำกว่า OLS หมายความว่า ความคลาดเคลื่อนเฉลี่ยลดลง
- Ridge ให้ R^2 สูงกว่า OLS หมายความว่า อธิบายความแปรปรวนของราคาได้ดีขึ้น

→ จึงสรุปว่า Ridge “ดีกว่า OLS” ในแง่ความแม่นยำบนชุดทดสอบ ความเสถียร ภายใต้ข้อมูลที่มีตัวแปรจำนวนมากและสัมพันธ์กันสูง

กราฟ



Predicted vs Actual

- จุดใกล้เส้น 45° → โมเดล fit ดี
- เป็นการใช้เพื่อประเมินความสอดคล้องเชิงโครงสร้างของโมเดลไม่ได้ใช้เพื่อการพยากรณ์ค่าจริงในอนาคต

สรุปผล

งานศึกษานี้มุ่งเน้นการเปรียบเทียบเชิงวิธีการระหว่าง Ordinary Least Squares (OLS) และ regularization techniques ภายใต้ข้อมูลที่มีตัวแปรจำนวนมากและมีความสัมพันธ์กันสูง โดยใช้ Ames Housing Data Set เป็นกรณีศึกษา OLS ถูกนำมาใช้เป็นแบบจำลองพื้นฐาน เนื่องจากมีโครงสร้างที่เรียบง่ายและเข้าใจได้ง่าย อย่างไรก็ตาม วิธีการดังกล่าวมีข้อจำกัดด้านความเสถียรของค่าสัมประสิทธิ์เมื่อเผชิญกับปัญหาตัวแปรสัมพันธ์กันสูง (multicollinearity)

ดังนั้น การประยุกต์ใช้ Ridge Regression ร่วมกับ Cross-Validation ช่วยลดความแปรปรวนของค่าสัมประสิทธิ์ และเพิ่มความเสถียรของโมเดลได้อย่างชัดเจนขึ้น และผลการศึกษานี้ชี้ให้เห็นว่า Ames Housing Data Set มีความเหมาะสมในการใช้เป็นกรณีศึกษาแทน Boston Housing Data สำหรับการเรียนการสอน regression ขั้นสูง

ข้อเสนอแนะเพิ่มเติม

ในงานวิจัยต่อไป อาจพิจารณาศึกษาการประยุกต์ใช้ Lasso Regression เพื่อเปรียบเทียบผลของการใช้ตัวแปรทั้งหมดกับการคัดเลือกตัวแปรบางส่วนออกไป และประเมินผลกระทบต่อความเรียบง่ายและความเสถียรของโมเดล

Thank You