Short Story of Regression Analysis

Linear and Logistic

Handbook for Beginner ebook

SmartResearchThai

revised: March 2023

Short Story of Regression Analysis:

Linear and Logistic

Handbook for Beginner

Panik Senariddhikrai

086-555-5949
Line: @SmartResearchThai
FB: SmartResearchThai
Youtube: @SmartResearchThai
www. SmartResearchThai.com

คำนำ

คู่มือการวิเคราะห์สมการถดถอย หรือ Regression Analysis เล่มนี้ มีจุดมุ่งหวังเพื่อให้ผู้อ่านทุกท่าน ที่ต้องวิเคราะห์ข้อมูลสถิติได้มีคู่มือ เป็นเสมือนติวเตอร์ ผู้ช่วย ให้สามารถเข้าใจ ทำตาม วิเคราะห์งานของตนเองได้อย่างถูกต้องและมีประสิทธิภาพ

โดยที่คู่มือฉบับนี้ เหมาะสำหรับผู้เริ่มต้นที่ต้องวิเคราะห์ Regression โดยเน้นความรู้พื้นฐาน เกริ่นนำเข้าสู่โลกของ Regression ซึ่งจะเริ่มต้นด้วยการวิเคราะห์สมการถดถอยเชิงเส้น (linear regression) และการวิเคราะห์สมการถอถอยโลจีสติกส์ (logistic regression)

คู่มือฉบับนี้ จึงเกิดขึ้นมาเพื่อให้ผู้วิจัยทุกท่านได้เข้าใจและสามารถนำไปประยุกต์ใช้ ไปวิเคราะห์สถิติที่เกี่ยวข้องได้อย่างมีประสิทธิภาพ

ขอให้ทุกท่านมีความสุขกับการอ่านคู่มือฉบับนี้

Panik Senariddhikrai

สารบัญ

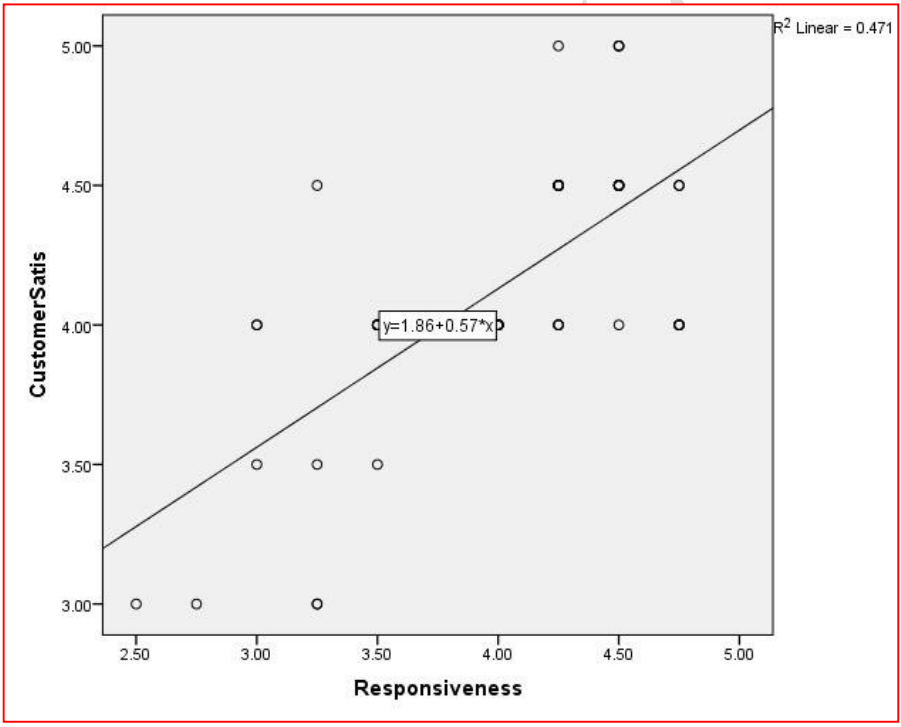
SmartResearchThai.com ให้บริการด้านสถิติ ทั้งการวิเคราะห์ข้อมูลสถิติสำหรับการวิจัย การให้คำปรึกษา การสอน การโค้ช ด้านสถิติสำหรับการวิจัย

SmartResearchThai.com มีบริการด้านสถิติสำหรับการวิจัยหลากหลายบริการ ดังนี้

- ☑ **Data Analysis การวิเคราะห์ข้อมูลสถิติสำหรับการวิจัย**
- ☑ **Stat Coaching ให้คำปรึกษาและสอนการใช้โปรแกรม**
- ☑ **Full Training สอนสถิติและโปรแกรมแบบเต็มรูปแบบ**
- ☑ **Handbook คู่มือการเรียนรู้ด้านสถิติและการใช้โปรแกรม**
- ☑ **นอกจากนี้ ยังมีบล็อกให้ความรู้ต่างๆ ติดตามได้โดยตรงจากเว็บไซต์**

Story 1 Linear Regression

การวิเคราะห์สมการถดถอยเชิงเส้น

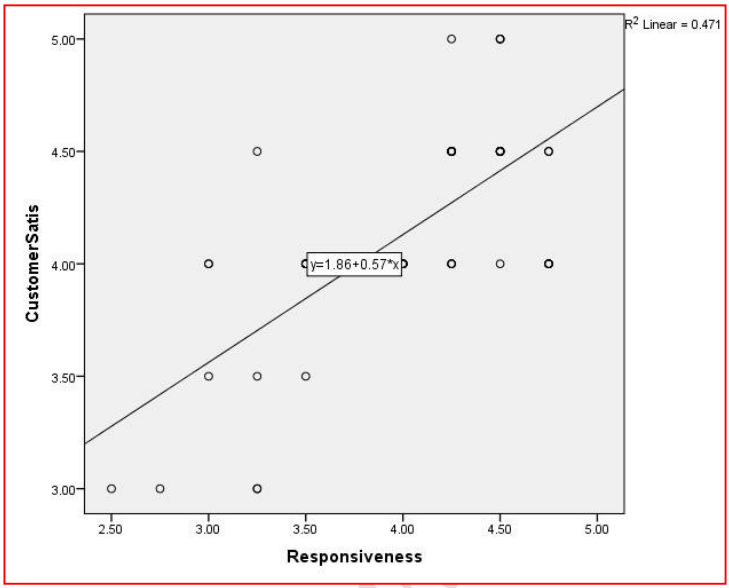


Story 1.1 Purpose จุดมุ่งหมาย

การวิเคราะห์สมการถดถอยเชิงเส้นมีจุดประสงค์สำคัญเพื่อ

- อธิบายความสัมพันธ์ระหว่างตัวอิสระและตัวแปรตาม
  - ทำให้รู้ว่าจะมีแนวโน้มที่ตัวแปรอิสระจะมีผลต่อตัวแปรตามหรือไม่

- พยากรณ์การเปลี่ยนแปลงของตัวแปรตามด้วยความเปลี่ยนแปลงในตัวแปรอิสระ
  - เป็นสิ่งสำคัญของการวิเคราะห์นี้ ก็คือ ตัวแปรตามที่เกิดขึ้นนั้น เกิดขึ้นจากตัวแปรอิสระที่นำเข้ามาทดสอบ ใช่หรือไม่

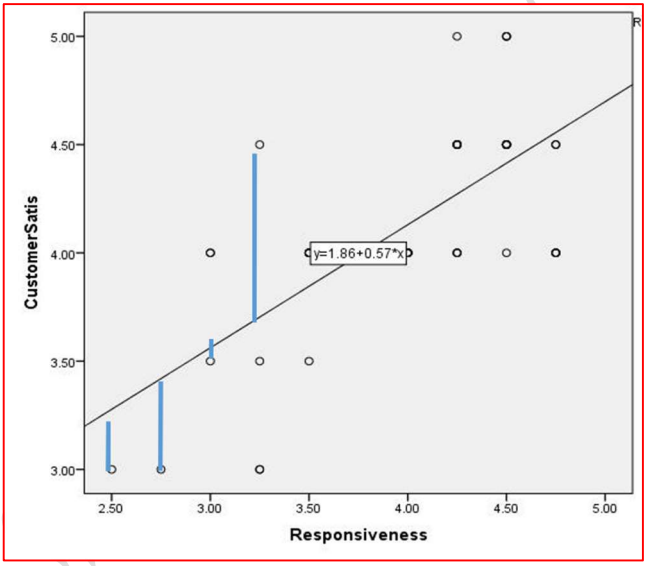


ภาพตัวอย่างที่ 1

จากภาพตัวอย่าง 1 แสดงจุดความสัมพันธ์ระหว่าง CustomerSatisfaction กับ Responsiveness และเส้นการพยากรณ์ว่า มีความสามารถในการพยากรณ์หรือไม่

- กราฟเป็นเพียงการแสดงการ plot จุดของสองตัวแปร เช่น Responsive มีค่า 2.50 ในขณะที่ Customer-Satisfaction มีค่า 3.00

- Plot จุดที่เกิดขึ้นในทุกๆ คู่ที่เกิดขึ้น จากนั้นลากเส้นพยากรณ์ regression เพื่อดูว่าทับเส้นหรือไม่
- หากทับเส้น หรือ จุดอยู่ใกล้เส้นแสดงว่าสมการพยากรณ์ได้ดี หรือ พูดอีกนัยนึงได้ว่าตัวแปรอิสระสามารถทำนายการเกิดตัวแปรตามได้เป็นอย่างดี



ภาพตัวอย่างที่ 2

จากภาพตัวอย่างที่ 2 แสดงระยะห่างระหว่างจุดความสัมพันธ์ที่เกิดขึ้นกับ เส้นพยากรณ์

- หากมีระยะห่างมากแสดงว่า สมการพยากรณ์มีประสิทธิภาพน้อย แต่หากเส้นสามารถลากผ่านจุดได้เยอะ แสดงว่าสมการพยากรณ์ทำงานได้เป็นอย่างดี

Story 1.2 Type of Linear Regression
ประเภทของการถดถอยเชิงเส้น

รูปแบบของ Linear Regression จะมี 2 รูปแบบหลักๆ ขึ้นอยู่กับจำนวนตัวแปรอิสระ

- Simple Linear Regression การวิเคราะห์ถดถอยอย่างง่าย
  - จะมีตัวแปรอิสระ 1 ตัว และมีตัวแปรตาม 1 ตัว
  - เช่น คะแนนวิชาคณิตศาสตร์ มีผลต่อ ผลสัมฤทธิ์ทางการเรียน
- Multiple Linear Regression การวิเคราะห์ถดถอยพหุ
  - จะมีตัวแปรอิสระ 2 ตัวขึ้นไป และมีตัวแปรตาม 1 ตัว
  - เช่น คะแนนวิชาคณิตศาสตร์ และคะแนนวิชาภาษาอังกฤษ มีผลต่อ ผลสัมฤทธิ์ทางการเรียน

Story 1.3 Sample Size ขนาดตัวอย่าง

Hair และคณะ (2019) p.279-280. ได้แนะนำว่า Simple Regression ควรมีอย่างน้อย 20 ตัวอย่าง ในขณะที่ Multiple Regression ควรมีอย่างน้อย 50 ตัวอย่าง หรือ 100 ตัวอย่างขึ้นไปจะเหมาะสม

หรือพิจารณาตามอัตราส่วน “ตัวอย่าง:ตัวแปร” ที่อย่างน้อย 5:1 และที่เหมาะสม 15:1 หรือ 20:1 เช่น ถ้ามีตัวแปรอิสระ 5 ตัว ควรมีจำนวนตัวอย่าง อย่างน้อย 25 ตัวอย่าง (5:1) หรือถ้าใช้อัตราส่วน 20:1 ณ ตัวแปรอิสระ 5 ตัว จะต้องมีจำนวนตัวอย่าง 100 ตัวอย่าง เป็นต้น

ในขณะที่ หากเป็นการวิเคราะห์ด้วย Stepwise ควรมีอัตราส่วน 50:1

อีกประเด็นที่สำคัญ คือ การอ้างอิงประชากร หรือ Generalizability. โดย Hair แนะนำว่า เมื่อมีจำนวนตัวอย่างที่เหมาะสมตามอัตราส่วนข้างต้นแล้ว จะสามารถอ้างอิงหรือ Generalize ได้

	Significance Level (α) = .01				Significance Level (α) = .05			
	No. of Independent Variables				No. of Independent Variables			
Sample Size	2	5	10	20	2	5	10	20
20	45	56	71	NA	39	48	64	NA
50	23	29	36	49	19	23	29	42
100	13	16	20	26	10	12	15	21
250	5	7	8	11	4	5	6	8
500	3	3	4	6	3	4	5	9
1,000	1	2	2	3	1	1	2	2

Note: Values represent percentage of variance explained.
NA = not applicable.

ที่มา Hair 2019 p.279

ภาพนี้แสดงถึงค่าอำนาจการพยากรณ์ (R-square) ที่จะเกิดขึ้น โดยพิจารณาตามจำนวนตัวอย่างขั้นต่ำ กับ จำนวนตัวแปรอิสระ ณ ระดับนัยสำคัญที่ 0.01 และ 0.05 เช่น ถ้ามีจำนวนตัวอย่าง 20 ตัวอย่าง ด้วยตัวแปรอิสระ 2 ตัว จะเกิดค่าอำนาจการพยากรณ์ที่ 39% ซึ่งไม่ผ่านความต้องการขั้นต่ำ (minimum requirement) ที่แนะนำก่อนหน้า

และถ้าเพิ่มจำนวนตัวอย่าง ก็จะทำให้ค่าอำนาจการพยากรณ์ลดลง (ณ ตัวแปรอิสระ 2 ตัว)

ในขณะที่ถ้าเพิ่มจำนวนตัวแปรอิสระจะมีโอกาสที่ค่าอำนาจการพยากรณ์จะเพิ่มขึ้นได้

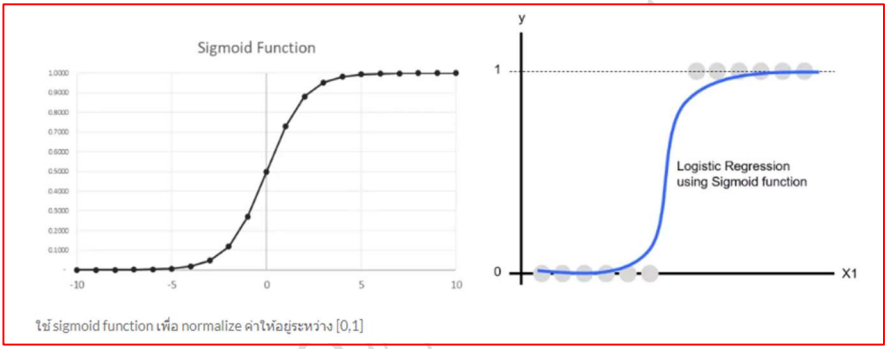
**ดังนั้น จึงควรมีกลุ่มตัวอย่างที่เหมาะสม *ไม่มากหรือน้อยเกินไป* รวมถึงจำนวนตัวแปรอิสระที่ไม่น้อยหรือมากเกินด้วยเช่นกัน ถึงจะอ้างอิงกลับไปยังประชากร (Generalizability) ได้

END of

Story 1 Multiple Linear Regression

Story 2 Logistic Regression

การวิเคราะห์สมการถดถอยโลจีสติกส์



ที่มา: https://datarockie.com/2019/05/12/logistic-regression-r/

Story 2.1 Purpose จุดมุ่งหมาย

การวิเคราะห์สมการถดถอยโลจีสติกส์มีจุดประสงค์สำคัญเพื่อ

- อธิบายความสัมพันธ์ระหว่างตัวอิสระและตัวแปรตาม ในกรณีตัวแปรตามเป็นแบบตัวเลือก (ช้อย)
    - ให้รู้ว่าจะมีแนวโน้มที่ตัวแปรอิสระจะมีผลต่อตัวแปรตามหรือไม่

- พยากรณ์โอกาสที่จะเกิดเหตุการณ์ที่สนใจ (ตัวแปรเป็นแบบตัวเลือก-ช้อย เช่น เกิด/ไม่เกิด ซื้อ/ไม่ซื้อ) ด้วยชุดของตัวแปรอิสระ
    - เป็นสิ่งสำคัญของการวิเคราะห์นี้ ก็คือ โอกาสของตัวแปรตามที่เกิดขึ้นนั้น เกิดขึ้นจากตัวแปรอิสระที่นำเข้ามาทดสอบ ใช่หรือไม่

Story 2.2 Type of Logistic
ประเภทของการถดถอยโลจีสติกส์

Binary เป็นประเภทของ Logistic ที่ตัวแปรตามมี 2 ตัวเลือก (ช้อย) เช่น โอกาสที่ลูกค้าจะซื้อ/ ไม่ซื้อสินค้า, โอกาสที่จะเกิดโรค/ ไม่เกิดโรค

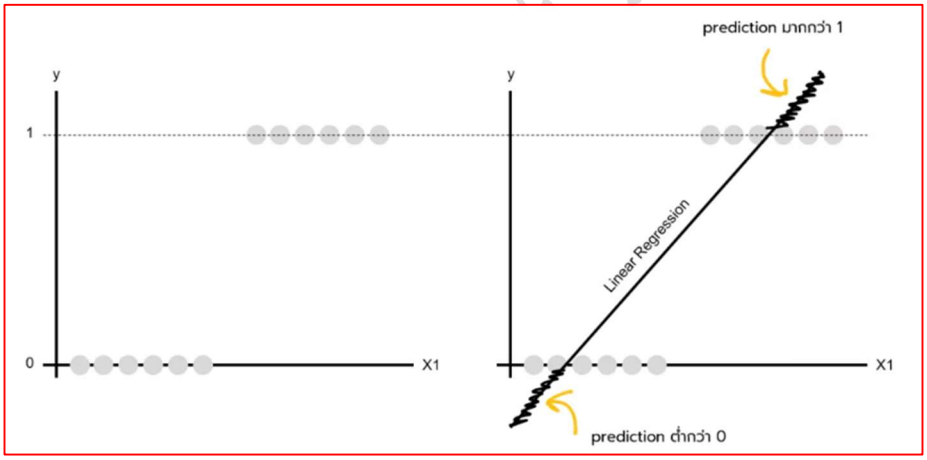
Multinomial เป็นประเภทของ Logistic ที่ตัวแปรตามมี 3 ตัวเลือก (ช้อย) ขึ้นไป เช่น โอกาสที่เปลี่ยนสถานภาพเป็น โสด/ สมรส/ หย่าร้าง

ลักษณะตัวแปรที่ใช้

- ตัวแปรอิสระ ต้องเป็น Continuous หรือถ้าเป็น Categorical ต้องแปลงเป็นตัวแปรหุ่น (Dummy)
- ตัวแปรตาม ต้องเป็น Categorical (ตัวเลือก-ช้อย) เท่านั้น (2 ตัวเลือก หรือ 3 ตัวเลือก) เช่น เกิดโรค/ ไม่เกิดโรค หรือ ซื้อสินค้า/ ไม่ซื้อสินค้า เป็นต้น

Story 2.3 Logistic Function ฟังก์ชันโลจีสติกส์

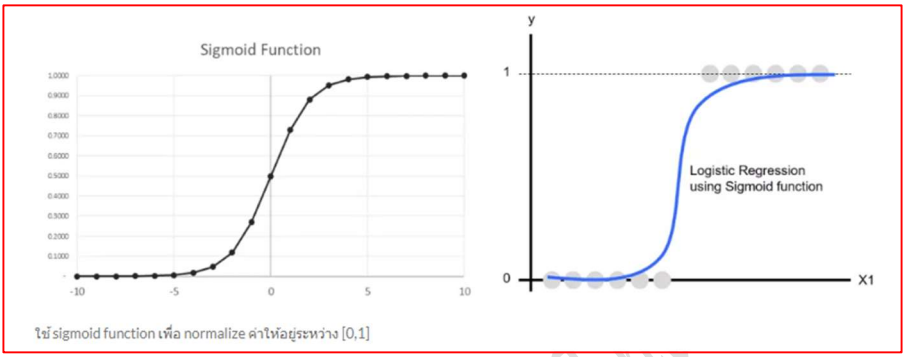
สมการถดถอยเชิงเส้นสามารถใช้เส้นตรงในการนำเสนอความสัมพันธ์ระหว่างตัวแปรได้ เพราะเป็นตัวเลขต่อเนื่อง (continuous) ในขณะที่สมการถดถอยโลจีสติกส์ มีค่าเพียง 2 ค่า คือ 0 กับ 1 ดังนั้น กราฟที่เกิดขึ้นไม่สามารถสร้างด้วยเส้นตรงได้ จึงต้องมีการแปลงรูปแบบเพื่อให้เหมาะสม



ซ้าย:การพล็อตกราฟของตัวแปรอิสระกับตัวแปรตามที่เป็นค่า 0,1

ขวา:การแปลงกราฟให้เป็นลักษณะเชิงเส้น

จะเห็นว่าการพยากรณ์จะเกิดขึ้นแค่ 2 จุด คือ 0 กับ 1 เท่านั้น



การแปลงค่าแบบ sigmoid เรียกว่า **S-Curve**

ทำการแปลงค่า 0,1 ให้กลายเป็นสัดส่วนตั้งแต่ 0-1 โดยการตัดค่าที่ 0.5 ว่า ถ้ามากกว่าถือว่าทำนายเป็น 1 และถ้าน้อยกว่าให้ทำนายว่าเป็น 0 เป็นความน่าจะเป็นที่จะเกิดโอกาสระหว่าง 0 กับ 1 ไม่เป็นเส้นตรง ซึ่งเหมาะสมกับข้อมูลมากกว่าแบบเชิงเส้น

END of

Story 2 Multiple Logistic Regression

Story 3 Extra Issue: Outlier

Multivariate Normality test

- **Mahalanobis** เป็นการทดสอบระยะห่างของข้อมูล หากมีระยะห่างมากก็จะมีแนวโน้มเป็นตัวแปร Outlier

$$D_i^2 = (y_i - \bar{y})' \mathbf{S}^{-1} (y_i - \bar{y})$$

- ยิ่งค่า D^2 สูง ก็มีแนวโน้มสูงว่าจะเป็น Outliers
- สามารถตรวจสอบนัยสำคัญได้โดย D^2/df โดยที่ df = จำนวนตัวแปร
- ทดสอบนัยสำคัญด้วยค่าสถิติ t และตั้งนัยสำคัญไว้ต่ำกว่าทั่วไป คือ .005 หรือ .001

ที่มา: ดร.นำชัย ศุภฤกษ์ชัยสกุล. บรรยายในสำนักงานวิจัยแห่งชาติ
http://www.priv.nrct.go.th/ewt_dl.php?nid=820

ทดสอบนัยสำคัญทางสถิติโดยการสร้างตัวแปรใหม่ โดยที่ทำการ compute โดยใช้สูตร 1-CDF.CHISQ(Mah,df)

Mah คือตัวแปรที่ได้จากการออกค่า mahalanobis ใน regression

df คือ จำนวนตัวแปรอิสระที่มีในโมเดล

END of

Story 3 Extra Issue: Outlier

END

Short Story of Regression Analysis:

Linear and Logistic

Handbook for Beginner