บทที่ 4

การวิเคราะห์การถดถอย (Regression Analysis)

วัตถุประสงค์การเรียนรู้

1) เข้าใจวิธีการการวิเคราะห์การถดถอยทุกรูปแบบและแปรความหมายผลลัพธ์ได้อย่างถูกต้อง
2) สามารถวิเคราะห์การถดถอยได้ทั้ง การวิเคราะห์การถดถอยเชิงเส้นเดี่ยว การวิเคราะห์การถดถอยเชิงพหุคูณ การวิเคราะห์การถดถอยโพลีโนเมียล และ การวิเคราะห์การถดถอยโลจิสติกส์
3) เข้าใจความแตกต่างระหว่าง การวิเคราะห์การถดถอยเชิงเส้น และการวิเคราะห์การถดถอยโลจิสติกส์
4) ทราบถึงวิธีการต่าง ๆ ในการวัดประสิทธิภาพแบบจำลอง

บทที่ 4
การวิเคราะห์การถดถอย (Regression Analysis)

หลังจากจัดเตรียมข้อมูลเสร็จเรียบร้อยแล้ว ขั้นตอนต่อไป คือ การนำข้อมูลที่จัดเตรียมไว้มาวิเคราะห์สร้างแบบจำลองข้อมูลด้วยเทคนิคต่างๆ ตามความเหมาะสม งานด้าน Data Science โดยมากจะใช้เทคนิคมากกว่าหนึ่งแบบ เพื่อประเมินหาเทคนิคที่ให้ค่าคำตอบดีที่สุด ในบทนี้เทคนิคแรกที่จะกล่าวถึงได้แก่ เทคนิคชื่อว่าการวิเคราะห์การถดถอย (Regression Analysis)

4.1 การวิเคราะห์การถดถอย

การวิเคราะห์การถดถอย เป็นวิธีการทางสถิติใช้เพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรตั้งแต่ 2 ตัวขึ้นไป ประกอบด้วย ตัวแปรที่เราทราบค่า เรียกว่าตัวประมาณการหรือตัวแปรต้น (Predictor, Independent variable, X) และตัวแปรที่เราต้องการทราบค่า เรียกว่าตัวตอบสนองหรือตัวแปรตาม (Response , Dependent variable , Y) ว่าเป็นตัวแปรที่เป็นปัจจัยหรือเป็นเหตุผลของกันและกันหรือไม่ ตัวอย่างเช่นในบทที่ 3 ตัวอย่างที่ 1 เราได้มีการคำนวณหาว่าจำนวนของนกเป็ดน้ำและจำนวนของทารกที่เกิดมีความสัมพันธ์กันหรือไม่ด้วยวิธีการหา Correlation coefficient ซึ่งได้ค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.9602 หมายถึงตัวแปรทั้งสองมีความสัมพันธ์กันโดยมีแนวโน้มขึ้นลงไปในทางเดียวกัน แต่ถ้าต้องการทราบว่าตัวแปรทั้งสองเป็นปัจจัยหรือเป็นเหตุผลของกันและกันอย่างไร เช่น นกเป็ดน้ำมีจำนวน 100 ตัว จะได้ทารกที่เกิดจำนวนของเท่าไหร่ เป็นต้น โดยความสัมพันธ์ที่ได้จากการวิเคราะห์ด้วยวิธี Regression แทนด้วยสมการหรือฟังก์ชันคณิตศาสตร์ดังนี้

$$y = f(x)$$

หรือ

$$y = ax + b$$

โดยที่ x แทนข้อมูลนำเข้า (input)

y แทนข้อมูลผลลัพธ์ที่ได้ (output)

a แทนค่าคงที่ของสมการถดถอย ซึ่งเป็นค่าจุดตัด (Intercept) แกน y ของสมการ

b ค่าสัมประสิทธิ์การถดถอย (Regression Coefficient) ของตัวตอบสนอง x

4.2 ประเภทของการวิเคราะห์การถดถอย

การวิเคราะห์การถดถอยที่ใช้ในการวิเคราะห์ข้อมูล (Freeman, 2009) แบ่งออกได้เป็น 3 แบบ ได้แก่ การวิเคราะห์การถดถอยเชิงเส้น (Linear Regression) การวิเคราะห์การถดถอยโพลีโนเมียล (Polynomial regression) และการวิเคราะห์การถดถอยโลจิสติกส์ (Logistic regression) ดังนี้

4.2.1 การวิเคราะห์การถดถอยเชิงเส้น (Linear Regression)

เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรที่มีความสัมพันธ์เชิงเส้นตรง (Xin, 2009) โดยถ้าเป็นการศึกษาตัวแปรต้น (X) หนึ่งตัวกับตัวแปรตาม (Y) หนึ่งตัว เรียกว่า การวิเคราะห์การถดถอยเชิงเส้นเชิงเดี่ยว หรือการวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression Analysis) แต่ถ้าเป็นการศึกษาตัวแปรต้น (X) สองตัวขึ้นไปกับตัวแปรตาม (Y) หนึ่งตัว เรียกว่า การวิเคราะห์การถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression)

1. การวิเคราะห์การถดถอยเชิงเส้นเดี่ยว (Simple Linear Regression)

เป็นวิธีการวิเคราะห์การถดถอยที่ความสัมพันธ์ของตัวแปรเป็นเชิงเส้นตรง และมีตัวแปประมาณการ (X) หนึ่งตัวและตัวแปรตอบสนอง (Y) หนึ่งตัว ซึ่งความสัมพันธ์แทนด้วยสมการทางคณิตศาสตร์ (4.1)

$$y = \beta_0 + \beta_1 x \tag{4.1}$$

โดยที่ y แทนข้อมูลผลลัพธ์ที่ได้ (output)

β_0 แทนค่าคงที่ของสมการถดถอย ซึ่งเป็นค่าจุดตัด (Intercept) แกน y ของสมการ

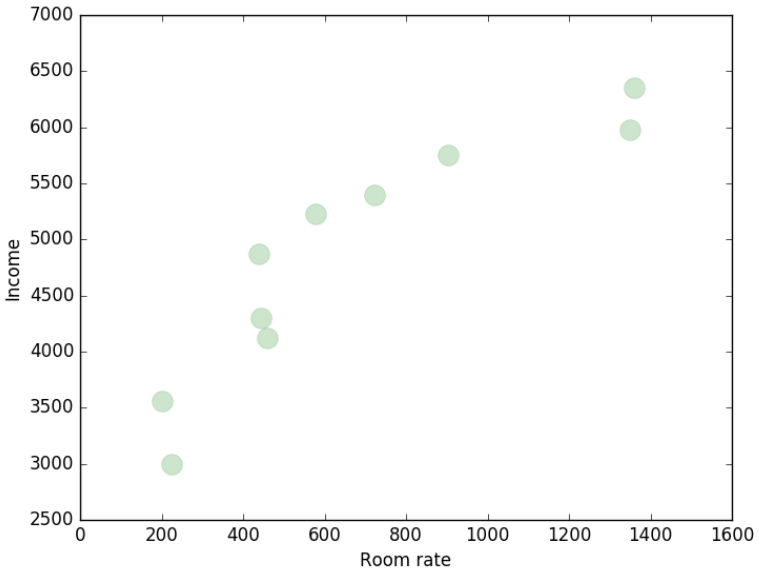
$\beta_1 x$ ค่าสัมประสิทธิ์การถดถอย (Regression Coefficient) ของตัวตอบสนอง x

ตัวอย่างการวิเคราะห์ Simple Linear Regression เช่น ข้อมูลในตารางที่ 4.1 เป็นข้อมูลที่ได้จากการสำรวจของโรงแรมแห่งหนึ่งในประเทศไทยเกี่ยวกับรายได้ต่อเดือนและราคาห้องพักที่นักท่องเที่ยวเข้าพักจำนวน 10 คน

ตารางที่ 4.1 ข้อมูลแสดงรายได้ต่อเดือนและราคาห้องพักของนักท่องเที่ยวที่ได้จากการสำรวจ

นักท่องเที่ยว	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
รายได้/เดือน (USD)	3000	3560	4120	4300	4870	5230	5400	5750	5980	6350
ราคาห้องพัก (USD)	225	200	459	445	439	577	722	903	1350	1360

จากข้อมูลเราสามารถนำข้อมูลรายได้ต่อเดือนของนักท่องเที่ยวและราคาห้องที่เข้าพักมาศึกษาความสัมพันธ์ด้วยวิธีวิเคราะห์การถดถอยเชิงเส้นอย่างง่ายได้ ตามขั้นตอนที่ได้กล่าวไปแล้วในข้างต้น ในที่นี้กำหนดตัวแปรต้น (X) ได้แก่ ความรายได้ต่อเดือน และตัวแปรตาม (Y) ได้แก่ ราคาห้องพัก ตัวอย่างผลลัพธ์จากการทำ Regression แสดงได้รูปที่ 4.1 และ 4.2



รูปที่ 4.1 ตัวอย่างการสร้าง Scatter plot จากข้อมูล

SUMMARY OUTPUT								
Regression Statistics								
Multiple F	0.904288							
R Square	0.817737							
Adjusted I	0.794954							
Standard I	189.2196							
Observati	10							
ANOVA								
	df	*SS*	*MS*	*F*	*gnificance F*			
Regressio	1	1285101	1285101	35.89261	0.000327			
Residual	8	286432.6	35804.07					
Total	9	1571534						
	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-1014.43	287.1282	-3.53302	0.007698	-1676.55	-352.311	-1676.55	-352.311
Income	0.346464	0.05783	5.991044	0.000327	0.213107	0.479821	0.213107	0.479821

รูปที่ 4.2 ผลลัพธ์ของวิธี Simple Linear Regression จากโปรแกรม Excel

การประเมินและสรุปผลแบบจำลอง สามารถสรุปได้โดยพิจารณาจากค่าการทดสอบทางสถิติต่างๆ ที่ได้ในรูปที่ 4.2 แบบจำลองข้อมูลที่ได้มีความเหมาะสมหรือไม่ พิจารณาจากค่า R-square และ R-adjusted ที่ได้ (รายละเอียดอธิบายบทก่อนหน้า) และในส่วนของความสัมพันธ์ระหว่าง X และ Y แสดงได้โดยการนำค่าจุดตัด (Intercept) แกน y และค่าสัมประสิทธิ์การถดถอย (Regression Coefficient) ของตัวแปรตาม มาสร้างอยู่รูปของสมการคณิตศาสตร์ได้ดังนี้

$$Y = -1014.43 + 0.346(X)$$

หรือ เขียนเป็นสมการความสัมพันธ์ตามข้อมูลในตารางที่ 4.1 แรก ได้ดังนี้

ราคาห้องพัก (โดยประมาณ) = -1014.43 + 0.346(รายได้ต่อเดือน)

2. การวิเคราะห์การถดถอยเชิงพหุคูณ (Multiple Linear Regression Analysis)

เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรต้น (X) และตัวแปรตาม (Y) ที่มีลักษณะเหมือนกันกับวิธี Simple Linear Regression คือ ความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามเป็นแบบเชิงเส้นตรง แต่ที่แตกต่าง คือ Multiple Linear Regression Analysis (Freeman, 2009) จะเป็นการศึกษาความสัมพันธ์ที่มีตัวแปรต้นมากกว่า 1 ตัว โดยความสัมพันธ์แทนด้วยสมการทางคณิตศาสตร์ (4.2)

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \quad (4.2)$$

โดยที่ y แทนข้อมูลผลลัพธ์ที่ได้ (output)

β_0 แทนค่าคงที่ของสมการถดถอย ซึ่งเป็นค่าจุดตัด (Intercept) แกน y ของสมการ

$\beta_1, \beta_2, ...\beta_k$ คือ ค่าสัมประสิทธิ์การถดถอย (Regression Coefficient) ของตัวแปรต้น $x_1, x_2, .., x_k$

ตัวอย่างเช่น การศึกษาเพื่อพยากรณ์จำนวนวันเฉลี่ยที่นักท่องเที่ยวใช้ในการพักผ่อน ซึ่งมีความสัมพันธ์กับจำนวนปีที่ทำงาน (Years of working) และรายได้ต่อปี (Yearly income) ในที่นี้ตัวแปรต้นและตัวแปรกำหนด ดังนี้

ตัวแปรตาม (Y) ได้แก่ จำนวนวันเฉลี่ยที่นักท่องเที่ยวใช้ในการท่องเที่ยว

ตัวแปรต้น (X) ได้แก่ จำนวนปีที่ทำงาน แทนด้วย X1 และรายได้ต่อปี แทนด้วย X2

ตัวอย่างผลลัพธ์จากการทำ Regression แสดงได้รูปที่ 4.3

SUMMARY OUTPUT									
Regression Statistics									
Multiple R	0.995								
R Square	0.989								
Adjusted R Square	0.986								
Standard Error	1.044								
Observations	10								
ANOVA									
	df	*SS*	*MS*	*F*	*Significance F*				
Regression	2	690.65	345.33	317.05	0.00				
Residual	7	7.62	1.09						
Total	9	698.27							
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*	
Intercept	4.54	0.77	5.87	0.00	2.71	6.38	2.71	6.38	
X1	2.22	0.10	21.74	0.00	1.97	2.46	1.97	2.46	
X2	0.01	0.00	6.64	0.00	0.01	0.02	0.01	0.02	

รูปที่ 4.3 ผลลัพธ์ของวิธี Multiple Linear Regression Analysis จากโปรแกรม Excel

ผลลัพธ์จากการทำ Regression ในรูปที่ 4.3 นำมาเขียนสมการ Regression ได้ดังสมการต่อไปนี้

$$Y = 4.54 + 2.22(X_1) + 0.01(X_2)$$

4.2.2. การวิเคราะห์การถดถอยโพลีโนเมียล (Polynomial regression)

การวิเคราะห์การถดถอยโพลีโนเมียล (Jiangqing, 1996) เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรที่มีความสัมพันธ์ไม่เป็นเส้นตรง (ความสัมพันธ์เป็นแบบเส้นโค้ง) ซึ่งขั้นตอนการวิเคราะห์จะมีความยากและซับซ้อนยิ่งขึ้น โดยรูปแบบความสัมพันธ์เขียนแสดงในรูปแบบสมการ (4.3)

$$y = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_n X^n \quad (4.3)$$

ตัวอย่างเช่น ต้องการหาความสัมพันธ์ระหว่างผลผลิตและความหนาแน่นของปริมาณต้นพืช ว่าความหนาแน่นของปริมาณต้นพืชจะส่งผลต่อปริมาณผลผลิตอย่างไร (ซึ่งคาดว่าความสัมพันธ์ของสองตัวแปรจะมีลักษณะเป็นโพลีโนเมียล) ความสัมพันธ์แบบโพลีโนเมียลแสดงอยู่ในรูปสมการคณิตศาตร์ ดังนี้

ปริมาณผลผลิต = 16.4 + 1.2(ความหนาแน่น) +-1.0 (ความหนาแน่นของต้นพืช)^2

4.2.3 การวิเคราะห์การถดถอยโลจิสติกส์ (Logistic regression)

เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปร โดยที่ตัวแปรตอบสนอง (Y) มีค่าได้ 2 สถานะ เช่น ใช่/ไม่ใช่ เป็นต้น และตัวประมาณการ (X) มีค่าเป็นแบบต่อเนื่องปกติ (Continuous value) โดยส่วนมาก Logistic Regression Analysis จะถูกนำมาใช้เพื่อทำนายว่า จะมีเหตุการณ์หนึ่งเกิดขึ้นหรือไม่ หรือมีโอกาสเกิดขึ้นมากน้อยเพียงใด โดยตัวแปรต้น (X) ที่คาดว่าจะส่งผลต่อการเกิดเหตุการณ์อาจมีได้มากกว่า 1 ตัว และสามารถเป็นได้ทั้งค่าต่อเนื่องและไม่ต่อเนื่อง ตัวอย่างเช่น การคาดการณ์เกี่ยวกับการอนุมัติหรือไม่อนุมัติสินเชื่อให้กับลูกค้าแต่ละคน ภายใต้เงื่อนไขต่างๆ เช่น เพศ อายุ เงินเดือน หรืออื่นๆ เป็นต้น ซึ่งเหล่านี้ถือเป็นตัวแปรต้นที่อาจส่งผลต่อการพิจารณาอนุมัติหรือไม่อนุมัติสินเชื่อ

4.3 การวัดประสิทธิภาพของแบบจำลอง

ในเทคนิคของการทำ Machine Learning มี Metrics ที่ใช้วัดความถูกต้องของแบบจำลองหลายตัว สำหรับ Metrics ที่ใช้ในแบบจำลอง regression มีดังนี้

4.3.1 Mean Squared Error (MSE) หรือค่าเฉลี่ยความผิดพลาดกำลังสอง

เป็นการวัดค่าความคลาดเคลื่อนของแบบจำลอง โดยค่าที่ได้ยิ่งน้อยจะยิ่งแสดงให้เห็นว่าแบบจำลองที่ได้มีความแม่นยำมาก โดยค่า MSE คำนวณได้จากการนำค่าความคลาดเคลื่อนมายกกำลัง แล้วนำไปหาค่าเฉลี่ย ดังสมการ (4.4)

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 \tag{4.4}$$

โดยที่ n คือ จำนวนข้อมูลที่ใช้

y_t คือ ค่าจริงที่ t ใด ๆ

$\hat{y}_t$ คือ ค่าที่ได้จากการทำนายที่ t ใด ๆ

4.3.2 Root Mean Square Error (RMSE) หรือค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย

เป็นวิธีการวัดค่าความคลาดเคลื่อนแบบมาตรฐาน ที่นิยมใช้กันอย่างแพร่หลาย โดยค่าที่ได้ยิ่งน้อยจะยิ่งแสดงว่าแบบจำลองที่ได้มีความแม่นยำมาก วิธีการคำนวณคือจะเป็นการนำค่า MSE ที่คำนวณได้มาหารากที่สอง สมการ(4.5)

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{4.5}$$

4.3.3 Coefficient of determination หรือค่า R-square (R2) เป็นค่าที่ใช้พิสูจน์ว่าแบบจำลองที่ได้นั้นเหมาะสมหรือไม่ โดยมีค่า 0-1 ซึ่งยิ่งเข้าใกล้ 1 ยิ่งดี โดยทั่วไปควรมีค่ามากกว่า 0.6 จะถือว่าแบบจำลองที่ได้เป็นแบบจำลองที่ดี

4.4 เอกสารอ้างอิง

Freedman, David A., (2009). *Statistical Models: Theory and Practice*. Cambridge University Press. ISBN 978-1-139-47731-4.

Jianqing, Fan. (1996). *Local Polynomial Modelling and Its Applications: From linear regression to nonlinear regression*. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC. ISBN 978-0-412-98321-4.

Xin, Yan. (2009). *Linear Regression Analysis: Theory and Computing*. World Scientific. pp. 1–2, ISBN 978-9-81283411-9.

-หน้าว่าง-