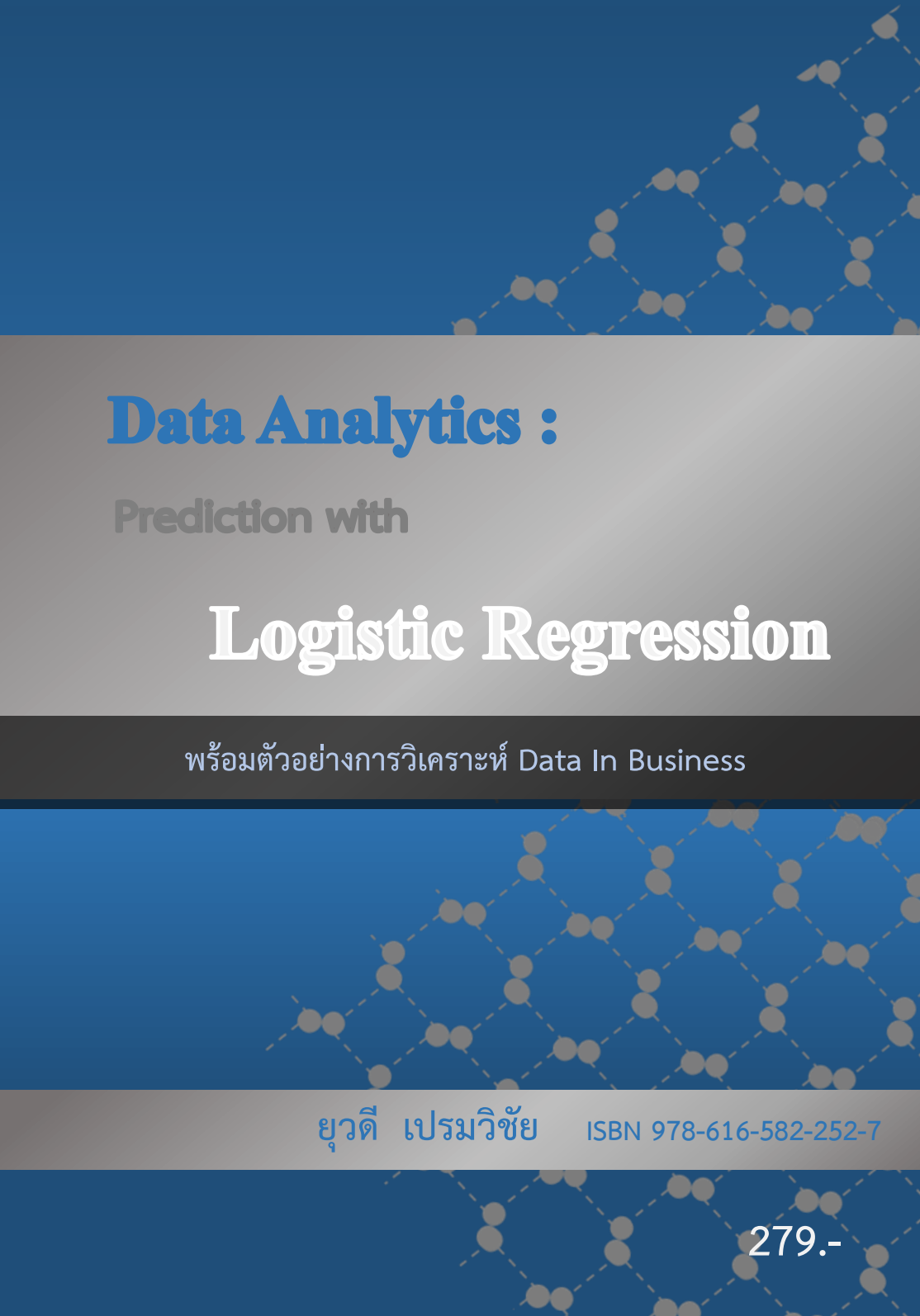




Data Analytics :

Prediction with

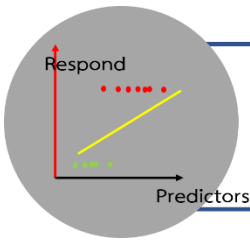
Logistic Regression

พร้อมตัวอย่างการวิเคราะห์ Data In Business

ยุวดี เปรมวิชัย

ISBN 978-616-582-252-7

279.-



Data Analytics :

Prediction with Logistic Regression

ชื่อเรื่อง Data Analytics: Prediction with Logistic Regression

ปีที่จัดทำ พฤษภาคม 2564

ชื่อผู้แต่ง ยุวดี เปรมวิชัย e-mail : yuwatisit@gmail.com

ราคา e-book 279.- บาท

ISBN (e-book) 978-616-582-252-7

จำนวนหน้า 114 หน้า

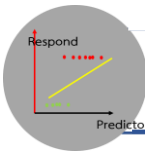
พิสูจน์อักษร/ออกแบบปก/ผู้จัดพิมพ์/ผู้จัดทำ ยุวดี เปรมวิชัย

จัดทำโดย ยุวดี เปรมวิชัย

สงวนลิขสิทธิ์ตามพระราชบัญญัติลิขสิทธิ์ 2537

ห้ามผู้ใด คัดลอก ทำซ้ำ ดัดแปลง หรือนำส่วนใดส่วนหนึ่งไปเผยแพร่โดย

ไม่ได้รับอนุญาตจากเจ้าของผลงาน



Data Analytics :
Prediction with Logistic Regression

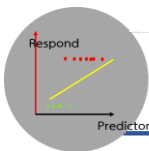
คำนำ

หนังสือเล่มนี้ สำหรับ Data Analytics ในขั้น Predictive Analytics โดยวิธี Logistic Regression Analysis จุดประสงค์เพื่อให้ ผู้ปฏิบัติเข้าใจพื้นฐานแนวคิด ไม่ลงรายละเอียดของทฤษฎีสถิติ ทำให้ หนังสือเล่มนี้ต่างจากหนังสืออื่น เพราะไม่มีรายการอ้างอิง(Reference) เนื่องจากเขียนโดยการยกตัวอย่างแทบทุกแนวทางการปฏิบัติที่เคยเป็น ปัญหาของตัวเอง และคาดว่าจะยังคงเป็นปัญหากับผู้ศึกษามาตลอด ได้แก่ จัดการกับ Input Data อย่างไร? Runโปรแกรมได้ทุกคำสั่ง แต่ ใช้ค่าอะไร? เพราะอะไร? เป็นต้น ซึ่งสรุปได้ว่าถึงจะมี Tools อำนวยความสะดวกมากมาย สถิติยังมีความสำคัญเสมอสำหรับให้เข้าใจ ผลลัพธ์ของโปรแกรม และมีเหตุผลทางวิชาการในการนำไปใช้ใน กิจกรรมจริง สามารถต่อยอดสู่โมเดลอื่นๆต่อไป

หวังในความเป็นประโยชน์สำหรับผู้อ่านทุกท่าน

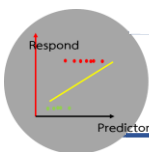
ยุวดี เปรมวิชัย

พฤษภาคม 2564

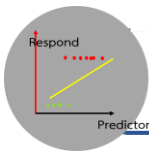


สารบัญ

บทที่		หน้า
1	บทนำ (Introduction)	1
	1.1 การวิเคราะห์การถดถอยแบบเดิม (Traditional Regression)	3
	1.2 การวิเคราะห์การถดถอยแบบลอจิสติกส์ (Logistic Regression)	4
	1.3 ประเภทของ Logistic Regression (Type of Logistic Regression)	8
	1.4 ข้อกำหนดที่จำเป็น ของ Logistic Regression (Assumptions of Logistic Regression)	10
	1.5 คำศัพท์ใน Logistic Regression	11
2	แนวคิดของ Logistic Regression	19
	2.1 หลักการวิเคราะห์ของ Logistic Regression	20
	2.2 การประมาณค่า β_0 และ β_j	32
	2.3 การทำนายและการตีความหมาย (Interpretation)	34



บทที่	สารบัญ (ต่อ)	หน้า
3	Binary Logistic Regression	37
	3.1 Binary Logistic Regression แบบ Simple Regression	38
	3.2 Binary Logistic Regression แบบ Multiple Regression	50
4	Multinomial Logistic Regression	76
	4.1 แนวคิดของ Multinomial Logistic Regression	77
	4.2 Nominal Multinomial Logistic Regression	79
	4.2.1 กรณี ตัวแปร Predictors (X) เป็น Categories ทั้งหมด	79
	4.2.2 กรณีตัวแปร Predictors(X) เป็น Continuous ทั้งหมด	89
	4.2.3 กรณี ตัวแปร Predictors (X) มีทั้ง Category และ Covariate	98
	4.3 Ordinal Multinomial Logistic Regression	106



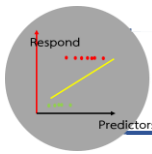
บทที่

1

บทนำ

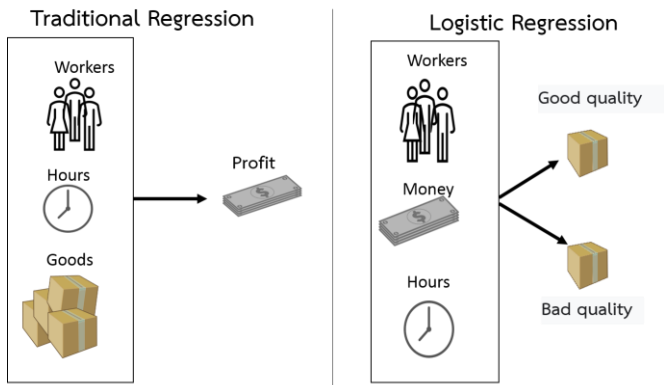
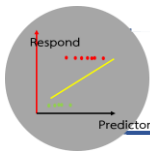
(Introduction)

วิธีการหนึ่งที่สำคัญของ Data Analytics คือ **Predictive analytics** เป็นการวิเคราะห์เพื่อทำนายหรือพยากรณ์ (Prediction) ลักษณะของ Data ในอนาคต เกิดจากการศึกษาความสัมพันธ์ของ Data นำมาสร้าง Model เพื่อใช้ทำนายเหตุการณ์ที่จะเกิดในอนาคต เช่น ศึกษาพบว่ายอดขายเบียร์มีความสัมพันธ์ทางบวกกับอัตราการว่างงาน และมีรายงานจากหน่วยงานที่เชื่อถือได้ว่าปีหน้าอัตราการว่างงานจะสูงขึ้น จึงสร้าง Model ทำนายยอดขายเบียร์ปีหน้าไว้ล่วงหน้าได้ เป็นต้น วิธีการแบบนี้ทางสถิติวิเคราะห์ เรียกว่า การวิเคราะห์การถดถอย (Regression Analysis) ซึ่ง **การวิเคราะห์การถดถอยแบบเดิม** (Traditional Regression) เป็นการอธิบายความสัมพันธ์ของตัวแปร Predictors หรือตัวแปร Independent (X) กับ ตัวแปร Respond หรือตัวแปร Dependent (Y) ที่เป็นตัวแปรแบบต่อเนื่อง (Continuous) ที่มีมาตรวัดระดับ Interval หรือระดับ



Ratio เช่น มีจำนวนพนักงาน(Workers; คน) จำนวนชั่วโมงการทำงาน (Hours; ชั่วโมง) จำนวนสินค้าที่ผลิต (Goods; ชิ้น) ต้องการทำนายกำไรสุทธิ(Profit;บาท) โดยกำไรสุทธิเป็นตัวแปรContinuous ที่มีมาตรวัดระดับ ระดับ Ratio เป็นต้น

ปัจจุบันการทำนายใช้กว้างขวางแทบทุกวงการ เช่น ทำนายการกู้เงินผ่านหรือไม่ผ่านในการธนาคาร ทำนายการรักษาโรคว่าหายหรือไม่หายในวงการแพทย์ ทำนายว่าควรศึกษาคณะวิศวกรรมศาสตร์ คณะแพทยศาสตร์ หรือคณะเกษตรศาสตร์ในการศึกษา เป็นต้น ตัวแปร Respond ในกรณีเหล่านี้เป็นตัวแปรกลุ่ม(Categories) มีมาตรวัดระดับ Nominal หรือ Ordinal การศึกษาความสัมพันธ์ของตัวแปร Predictors กับตัวแปร Respond ประเด็นนี้ต้องใช้การวิเคราะห์การถดถอยแบบลอจิสติกส์ (Logistic Regression) เช่น ต้องการทำนายคุณภาพสินค้าว่ามีคุณภาพ หรือไม่มีคุณภาพ จากจำนวน Workers จำนวนเงินทุน(Money) จำนวนเวลาทำงาน เป็นต้น



รูปที่ 1.1 Traditional Regression และ Logistic Regression

1.1 การวิเคราะห์การถดถอยแบบเดิม (Traditional Regression)

1) Model เป็น

$$\hat{Y} = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k$$

โดย $\hat{Y}$ = ตัวแปร Response แบบ Continuous Variable
(เป็น Interval Scale หรือ Ratio Scale)

$X_1, X_2, \dots, X_k$ = ตัวแปร Predictors

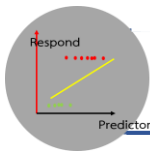
b_0 = ค่าคงที่ หรือ Intercept หมายถึงค่า Y เมื่อ X_j ทุก

ตัว = 0

$b_1, b_2, \dots, b_k$ = สัมประสิทธิ์การถดถอยของ $X_1, X_2, \dots, X_k$

ตามลำดับ

สร้างโมเดลโดยวิธี Least Square ของ Y และ $\hat{Y}$



2) การตรวจสอบ Model

2.1) Fit of Model ใช้ F-test สมมติฐานคือ

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0 \quad (\text{สมการไม่}$$

เหมาะสม)

$$H_a : \beta_j \neq 0 \quad \text{บางตัว} \quad \text{เมื่อ } j = 1, 2, 3, \dots, k \quad (\text{สมการ}$$

เหมาะสม)

2.1) การทดสอบว่าตัวแปร X_j แต่ละตัวเหมาะสมที่จะเป็น Predictor หรือไม่ โดย t-test สมมติฐานคือ

$$H_0 : \beta_j = 0 \quad \text{เมื่อ } j = 1, 2, 3, \dots, k$$

(ตัวแปรอิสระนี้ไม่เหมาะสม)

$$H_a : \beta_j \neq 0 \quad (\text{ตัวแปรอิสระนี้เหมาะสม})$$

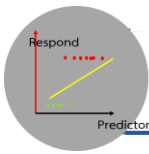
3) การทำนายเมื่อมีตัวแปร Predictor เข้ามาใหม่ ใช้การแทนค่า X ใน Model แล้วอ่านค่า Y ได้เลย

1.2 การวิเคราะห์การถดถอยแบบลอจิสติกส์

(Logistic Regression)

Logistic Regression เป็นวิธีการอธิบายความสัมพันธ์ของตัวแปร Predictors กับตัวแปร Respond ที่เป็น Categories มีมาตรวัดแบบ Nominal Scale หรือ Ordinal Scale แตกต่างจาก Traditional Regression ในเบื้องต้น ดังนี้

1) Model Logistic Regression คือ Logit Model ในรูป



$$\text{Log}\left(\frac{p}{1-p}\right) = w = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_j X_j$$

โดย Log ให้เป็น Natural Log หรือเขียน ln

หรือเขียนเป็น Model Logistic Regression คือ

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_j X_j)}}$$

หรือ

$$p = \frac{1}{1 + e^{-(\beta_0 + \sum \beta_j X_j)}}$$

เมื่อ ค่า p เป็นค่าความน่าจะเป็น (Probability) ซึ่ง $0 \leq p \leq 1$ ซึ่งไม่สามารถใช้วิธี Least Square ได้

2) ไม่สามารถใช้ F-test และ t-test เพราะหา Variance ไม่ได้เนื่องจาก ไม่เป็น Linear

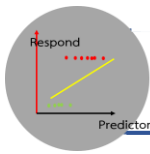
3) การทำนายตัวแปร Respond (Y) เมื่อมีค่าใหม่ของ ตัวแปร Predictor (X) ต้องใช้ค่า p เป็นเกณฑ์เพื่อจัดเข้า Categories ของ ตัวแปร Respond (Y)

เช่น การวิเคราะห์เพื่อจัดลูกค้า เป็น กลุ่ม 2 กลุ่ม คือ

$$\left. \begin{array}{l} \text{กลุ่มซื้อ แทนด้วย 1} \\ \text{และ} \\ \text{กลุ่มไม่ซื้อ แทนด้วย 0} \end{array} \right\} Y = \text{ตัวแปร Respond}$$

สมมติตัวแปร Predictors คือ $X_1 = \text{รายได้ (บาท)}$, $X_2 = \text{อายุ (ปี)}$

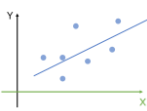
ผลลัพธ์ของการวิเคราะห์ Logistic Regression ได้



$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}}$$

เมื่อ แทนค่า X_1, X_2 ค่าใหม่ เพื่อ ทำนาย Y จะแทนค่าได้ p
โดยอาจตั้งเกณฑ์ว่า ถ้า $0 \leq p \leq 0.5$ ได้ $Y = 0$ หรือเป็น กลุ่มไม่ซื้อ
ถ้า $0.5 < p \leq 1.0$ ได้ $Y = 1$ หรือเป็น กลุ่มซื้อ เป็นต้น

ตารางที่ 1.1 ความแตกต่างเบื้องต้นของ Traditional Regression กับ Logistic Regression

ที่	Traditional Regression	Logistic Regression
1	ตัวแปร Respond เป็น Continuous มีมาตรวัดระดับ Interval Scale และ Ratio Scale	ตัวแปร Respond เป็น Categories มีมาตรวัดแบบ Nominal Scale หรือ Ordinal Scale
2	ตัวแปร Predictors อธิบายตัวแปร Respond ได้ชัดเจนแบบเชิงเส้น 	ตัวแปร Predictors ไม่เห็นการอธิบายถึงตัวแปร Respond ได้ 