

บทที่ 6

การวิเคราะห์การถดถอยเมื่อมีตัวแปรดัมมี่

โดยทั่วไปตัวแปรอิสระที่ใช้ในการวิเคราะห์การถดถอยมักเป็นตัวแปรเชิงปริมาณ (Quantitative variables) ที่มีสเกลการวัดกำหนดไว้อย่างชัดเจน เช่น อุณหภูมิ ความดัน ระยะเวลา รายได้ และราคา เป็นต้น แต่บางครั้งการวิเคราะห์อาจจำเป็นต้องใช้ตัวแปรอิสระที่เป็นตัวแปรเชิงคุณภาพ เช่น เพศ แบ่งเป็นเพศ หญิงและเพศชาย กรู๊ปเลือด แบ่งเป็นกรู๊ป A, B, AB, และ O และระดับการศึกษา อาจแบ่งเป็นต่ำกว่าปริญญาตรี และตั้งแต่ปริญญาตรีขึ้นไป เป็นต้น ในบทนี้จะกล่าวถึงการวิเคราะห์การถดถอยเมื่อตัวแปรอิสระเป็นตัวแปรเชิงคุณภาพ (Qualitative or categorical variables) ทั้งในกรณีที่มีตัวแปรเชิงคุณภาพหนึ่งตัวและมากกว่าหนึ่งตัว โดยที่ตัวแปรแต่ละตัวยังสามารถแบ่งเป็นระดับย่อย ๆ ที่มีเพียง 2 ระดับหรือมากกว่า 2 ระดับก็ได้ และในหัวข้อสุดท้ายจะกล่าวถึงการวิเคราะห์การถดถอยเชิงเส้นตรงแบบแบ่งเป็นช่วง (Piecewise linear regression) โดยที่ความสัมพันธ์ระหว่างตัวแปรมีลักษณะที่เปลี่ยนแปลงไปเมื่อผ่านจุด ๆ หนึ่ง

6.1 การวิเคราะห์การถดถอยเมื่อมีตัวแปรดัมมี่ 1 ตัว

เมื่อตัวแปรอิสระที่ใช้ในการวิเคราะห์การถดถอยเป็นตัวแปรเชิงคุณภาพ ซึ่งตัวแปรเชิงคุณภาพเหล่านี้ไม่มีสเกลการวัดที่แน่นอน และในการวิเคราะห์จะต้องกำหนดตัวเลขเพื่อแทนระดับย่อย ๆ ของตัวแปรเชิงคุณภาพ เพื่อแสดงผลของตัวแปรที่อาจมีอิทธิพลต่อตัวแปรตาม โดยสร้าง *ตัวแปรดัมมี่* (Dummy or indicator variables) ขึ้นมา และตัวเลขที่กำหนดเหล่านี้ไม่มีความหมายเชิงปริมาณ เพียงแต่แสดงถึงการมีหรือไม่มีคุณสมบัติดังกล่าว เช่น

ให้ 1 แทน เพศชาย

และ 0 แทน เพศหญิง

การใช้เลข 0 และ 1 เพื่อแทนระดับของตัวแปรเชิงคุณภาพไม่ได้เป็นการเจาะจง ตัวเลขอื่นที่แตกต่างกัน 2 ค่าก็สามารถใช้ได้เช่นเดียวกัน แต่อย่างไรก็ตาม 0 และ 1 มักเป็นทางเลือกที่ดีที่สุด นอกจากนี้จะเห็นได้ว่าการ

สร้างตัวแปรต้นมีสามารถทำได้ง่าย จึงเป็นที่นิยมใช้กันอย่างแพร่หลายในการวิเคราะห์การถดถอยเมื่อตัวแปรอิสระเป็นตัวแปรเชิงคุณภาพ

พิจารณาสถานการณ์ต่อไปนี้ นักวิจัยตลาดคนหนึ่งต้องการศึกษาความสัมพันธ์ระหว่างราคาขายบ้าน (Y) กับระยะทางระหว่างบ้านกับตัวเมือง (X_1) และสถานที่ตั้งของบ้าน ซึ่งแบ่งเป็นเมือง A และเมือง B เนื่องจากเขตที่อยู่อาศัยเป็นตัวแปรเชิงคุณภาพ ดังนั้นจะสร้างตัวแปรต้นมี X_2 เพื่อแทนเขตที่อยู่อาศัยได้ดังนี้

$$X_2 = \begin{cases} 0 & \text{ถ้าสถานที่ตั้งของบ้านอยู่ในเมือง A} \\ 1 & \text{ถ้าสถานที่ตั้งของบ้านอยู่ในเมือง B} \end{cases}$$

สมมติรูปแบบโมเดลกำลังหนึ่งเหมาะสม จะได้ว่า

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \quad (6.1)$$

การแปลงผลพารามิเตอร์ทำได้โดยแทนค่า X_2 เข้าไปในสมการ (6.1) ซึ่งหากสถานที่ตั้งของบ้านอยู่ในเมือง A แทนค่า $X_2 = 0$ โมเดลถดถอยจะกลายเป็น

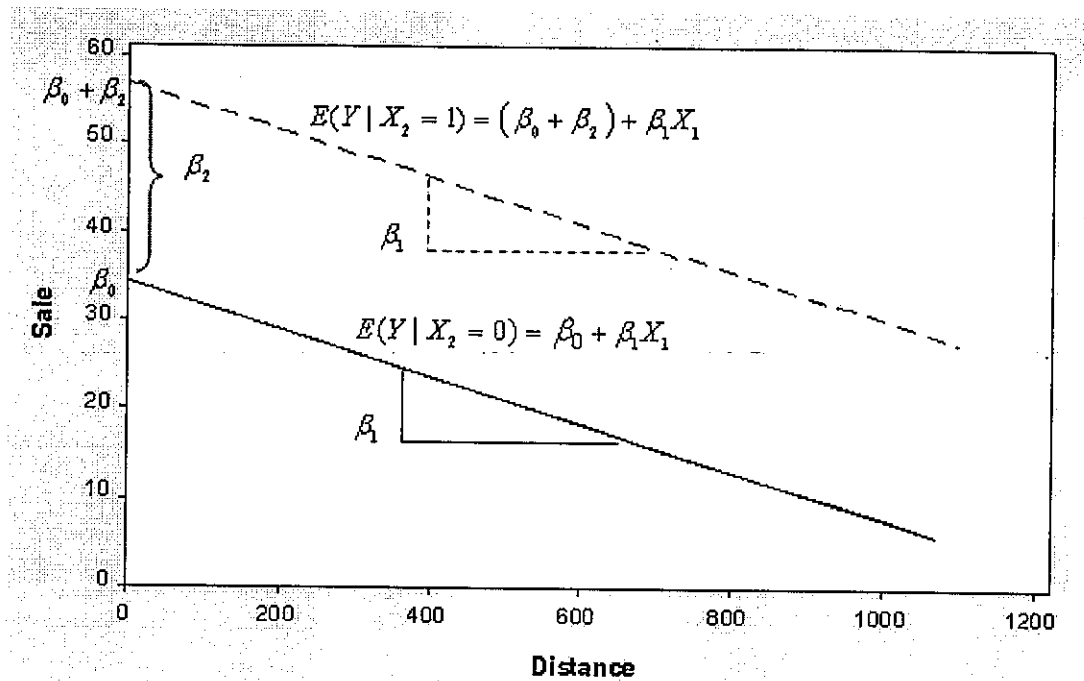
$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2(0) + \epsilon \\ &= \beta_0 + \beta_1 X_1 + \epsilon \end{aligned} \quad (6.2)$$

และถ้าสถานที่ตั้งของบ้านอยู่ในเมือง B แทนค่า $X_2 = 1$ จะได้ว่า

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2(1) + \epsilon \\ &= (\beta_0 + \beta_2) + \beta_1 X_1 + \epsilon \end{aligned} \quad (6.3)$$

นั่นคือ ราคาขายบ้านมีความสัมพันธ์เชิงเส้นตรงกับระยะทาง โดยมีระยะตัดแกน Y เป็น β_0 และความชันเป็น β_1 สำหรับบ้านที่ตั้งอยู่ในเมือง A และมีระยะตัดแกน Y เป็น $\beta_0 + \beta_2$ และความชันเป็น β_1 สำหรับบ้านที่ตั้งอยู่ในเมือง B

รูปที่ 6.1 แสดงสมการถดถอยที่ขนานกันสองเส้นของโมเดล (6.2) และ (6.3) โดยที่สมการทั้งสองมีความชันเหมือนกัน (Common slope) แต่มีระยะตัดแกน Y ที่แตกต่างกัน และสมมติให้ความแปรปรวนของ ความคลาดเคลื่อน ϵ ของทั้งสองสมการมีค่าเท่ากัน จากรูปจะเห็นได้ว่าพารามิเตอร์ β_2 แทนผลต่างของความสูงของทั้งสองเส้น กล่าวคือ β_2 เป็นค่าที่วัดความแตกต่างระหว่างราคาขายเฉลี่ยเนื่องมาจากการเปลี่ยนสถานที่



รูปที่ 6.1: แสดงสมการถดถอยสำหรับบ้านที่ตั้งอยู่ในเมือง A และ B

ตั้งของบ้านจากเมือง A ไปเป็นเมือง B นั้นเอง

นอกจากนี้การสร้างตัวแปรดัมมีในกรณีที่ตัวแปรเชิงคุณภาพประกอบด้วยระดับย่อย ๆ มากกว่าสองระดับ สามารถทำได้เช่นเดียวกัน เช่น ผู้วิจัยอาจสนใจสถานที่ตั้งของบ้านที่อยู่ในเมือง 3 แห่ง คือ A, B และ C ในที่นี้จะต้องสร้างตัวแปรดัมมี 2 ตัว เพื่อแทนเมืองทั้ง 3 แห่ง ดังนี้

X_2	X_3	ความหมาย
0	0	ถ้าสถานที่ตั้งของบ้านอยู่ในเมือง A
1	0	ถ้าสถานที่ตั้งของบ้านอยู่ในเมือง B
0	1	ถ้าสถานที่ตั้งของบ้านอยู่ในเมือง C

ตารางที่ 6.1: การสร้างตัวแปรดัมมีสำหรับตัวแปรเชิงคุณภาพที่มี 3 ระดับ

และได้โมเดลถดถอยเป็น

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \quad (6.4)$$

โดยทั่วไปแล้วถ้าตัวแปรเชิงคุณภาพประกอบด้วยระดับย่อย ๆ c ระดับ จะต้องสร้างตัวแปรดัมมีทั้งหมด $c-1$ ตัว และแต่ละตัวมีค่าเป็น 0 และ 1

ราคาขาย (Y)	ระยะทาง (X ₁)	สถานที่ตั้ง	X ₂
17.595	515	A	0
11.280	1025	A	0
15.645	680	A	0
11.310	860	A	0
9.660	1070	A	0
26.085	395	A	0
9.510	620	A	0
23.565	410	A	0
8.520	935	A	0
18.480	695	A	0
34.740	605	B	1
30.135	755	B	1
27.600	920	B	1
28.575	1100	B	1
39.735	740	B	1
42.930	485	B	1
28.605	965	B	1
44.670	575	B	1
41.925	815	B	1
55.005	350	B	1

ตารางที่ 6.2: ข้อมูลแสดงราคาขายบ้าน ระยะทางระหว่างบ้านและตัวเมือง และสถานที่ตั้งของบ้าน

ตัวอย่างที่ 6.1 ข้อมูลในตารางที่ 6.2 แสดงราคาขายบ้าน (หน่วย: แสนบาท) ระยะทางระหว่างบ้านและตัวเมือง (หน่วย: เมตร) และสถานที่ตั้งของบ้าน แบ่งเป็นเมือง A และ B จงสร้างสมการถดถอยแสดงความสัมพันธ์ของข้อมูลชุดนี้

วิธีทำ จากแผนภาพการกระจายในรูปที่ 6.2 ซึ่งว่าสมการถดถอยเชิงเส้นตรงที่แตกต่างกันสองสมการน่าจะแสดงความสัมพันธ์ของข้อมูลชุดนี้ได้อย่างเหมาะสม โดยที่ระยะตัดแกน Y ของสมการจะขึ้นอยู่กับสถานที่ตั้งของบ้าน ในที่นี้ให้ X₂ เป็นตัวแปรดัมมี่ ซึ่งแบ่งเป็น 2 ระดับ คือ 0 และ 1 แทน สถานที่ตั้งของบ้านที่อยู่ในเมือง A และ B ตามลำดับ

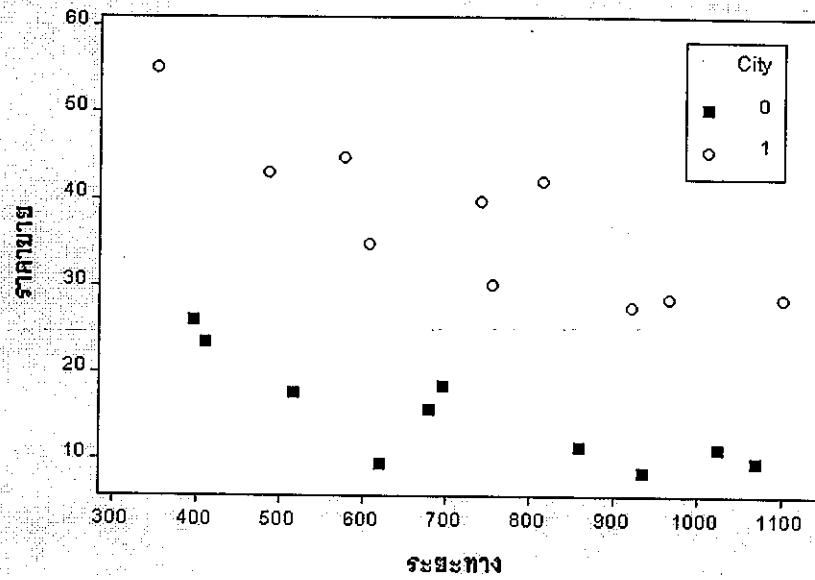
รูปแบบของโมเดลถดถอย คือ

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

หรือเขียนให้อยู่ในรูปของเมทริกซ์ได้เป็น

$$Y = X\beta + \epsilon$$

เมื่อ



รูปที่ 6.2: แผนภาพการกระจายระหว่างราคาขายบ้าน (Y) กับระยะทาง (X_1) สำหรับบ้านที่ตั้งอยู่ในเมือง A และ B

$$\mathbf{X} = \begin{bmatrix} 1 & 515 & 0 \\ 1 & 1025 & 0 \\ 1 & 680 & 0 \\ 1 & 860 & 0 \\ 1 & 1070 & 0 \\ 1 & 395 & 0 \\ 1 & 620 & 0 \\ 1 & 410 & 0 \\ 1 & 935 & 0 \\ 1 & 695 & 0 \\ 1 & 605 & 1 \\ 1 & 755 & 1 \\ 1 & 920 & 1 \\ 1 & 1100 & 1 \\ 1 & 740 & 1 \\ 1 & 485 & 1 \\ 1 & 965 & 1 \\ 1 & 575 & 1 \\ 1 & 815 & 1 \\ 1 & 350 & 1 \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} 17.595 \\ 11.280 \\ 15.645 \\ 11.310 \\ 9.660 \\ 26.085 \\ 9.510 \\ 23.565 \\ 8.520 \\ 18.480 \\ 34.740 \\ 30.135 \\ 27.600 \\ 28.575 \\ 39.735 \\ 42.930 \\ 28.605 \\ 44.670 \\ 41.925 \\ 55.005 \end{bmatrix}$$

สร้างสมการถดถอยด้วยวิธีกำลังสองน้อยที่สุดได้ผลดังนี้

$$\hat{Y} = 34.336 - 0.0267X_1 + 22.506X_2$$

โดยที่ $R^2 = 0.90$ จากนั้นทดสอบนัยสำคัญของสมการ

$$H_0 : \beta_1 = \beta_2 = 0$$

$$H_1 : \beta_j \neq 0 \text{ อย่างน้อย 1 ค่า, } j = 1, 2$$

ANOVA				
Source of variation	df	SS	MS	F
Regression	2	3190.576	1595.288	76.746
Error	17	353.373	20.787	
Total	19	3543.948		

เนื่องจาก $F_c = 76.746 > F_{2,17(0.05)} = 3.59$ จึงปฏิเสธ H_0 นั่นคือ อย่างน้อยมีตัวแปรอิสระหนึ่งตัวที่มีอิทธิพลต่อราคาขายบ้าน ที่ระดับนัยสำคัญ 0.05

พารามิเตอร์	ค่าประมาณ	ความคลาดเคลื่อนมาตรฐาน	สถิติทดสอบ t
β_0	34.3360	3.5610	9.64
β_1	-0.0267	0.0045	-5.89
β_2	22.5060	2.0400	11.04

ตารางที่ 6.3: การทดสอบนัยสำคัญของค่าสัมประสิทธิ์ถดถอย เมื่อ $t_{0.025, 17} = 2.110$

ตารางที่ 6.3 แสดงการทดสอบนัยสำคัญของสัมประสิทธิ์ถดถอยแต่ละตัว และสามารถสรุปได้ว่าตัวแปรอิสระ X_1 และ X_2 มีอิทธิพลต่อโมเดลอย่างมีนัยสำคัญ ที่ระดับ 0.05 โดยถ้าให้สถานที่ตั้งของบ้านมีค่าคงที่ หากระยะทางระหว่างบ้านไปยังตัวเมืองเพิ่มขึ้น 1 เมตร ราคาขายจะลดลง 2,670 บาท นอกจากนี้ยังสามารถแยกสมการถดถอยออกเป็น 2 สมการ ตามสถานที่ตั้งของบ้านได้ดังนี้

เมือง A

$$\hat{Y} = 34.336 - 0.0267X_1$$

เมือง B

$$\hat{Y} = 56.842 - 0.0267X_1$$

จะเห็นว่าราคาขายเฉลี่ยจะสูงขึ้น 22.5060 หน่วย เมื่อเปลี่ยนสถานที่ตั้งของบ้านจากเมือง A ไปเป็นเมือง B โดยที่ระยะทางมีค่าคงที่

การสร้างช่วงความเชื่อมั่นของพารามิเตอร์ β_2 ซึ่งวัดผลกระทบต่อราคาขายเฉลี่ยของบ้านที่ตั้งอยู่ในเมือง A เปรียบเทียบกับเมือง B ได้ดังนี้

ช่วงความเชื่อมั่น 95% ของ β_2 คือ

$$\begin{aligned}b_2 \pm t_{\alpha/2, df} \cdot S_{b_2} &= 22.5060 \pm (2.110)(2.0400) \\&= 22.5060 \pm 4.3044 \\&\text{หรือ } (18.2016, 26.8104)\end{aligned}$$

นั่นคือ เมื่อให้ระยะทางคงที่ การเปลี่ยนสถานที่ตั้งของบ้านจากเมือง A ไปอยู่เมือง B จะทำให้ราคาขายเฉลี่ยเพิ่มขึ้นอยู่ระหว่าง 18.2016 หน่วย ถึง 26.8104 หน่วย ด้วยความเชื่อมั่น 95%

หมายเหตุ

1. การทดสอบสมมติฐาน $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ ($j = 1, 2, \dots, k$) สามารถใช้ t -test หรือ Partial F -test ก็ได้
2. การสร้างสมการถดถอยเมื่อมีตัวแปรต้นมีเป็นการสร้างสมการเพียงสมการเดียว แต่สามารถแยกเป็นสมการย่อย ๆ ได้ตามระดับของตัวแปรต้นมี ซึ่งเหตุผลที่ไม่สร้างสมการถดถอยแยกกัน เนื่องจากในการวิเคราะห์จะสมมติว่าค่าความชันและความแปรปรวนของความคลาดเคลื่อนของทั้งสองสมการมีค่าเท่ากัน ดังนั้นการสร้างสมการถดถอยเพียงสมการเดียวจะประมาณค่าความชันร่วม (Common slope) โดยรวมเอาความชันของทั้งสองสมการเข้าด้วยกัน นอกจากนี้การอนุมานทางสถิติเกี่ยวกับพารามิเตอร์ β_0 และ β_2 จะมีความถูกต้องมากขึ้น เนื่องจากจำนวนองศาความเป็นอิสระของ MSE เพิ่มขึ้น
3. ในการวิเคราะห์การถดถอยเมื่อมีตัวแปรอิสระหนึ่งตัวเป็นตัวแปรเชิงคุณภาพ แบ่งเป็น c ระดับ หากสร้างตัวแปรต้นมี c ตัวเท่ากับจำนวนระดับของตัวแปรเชิงคุณภาพ จะพบว่าตัวแปรต้นมีทั้ง c ตัวมีความสัมพันธ์เชิงเส้นตรงอย่างสมบูรณ์ ทำให้เกิดปัญหา Multicollinearity ดังนั้นในการวิเคราะห์จึงต้องสร้างตัวแปรต้นมีให้น้อยกว่าจำนวนระดับอยู่ 1 ค่า

ตัวอย่างที่ 6.2 ในการจำลองสถานการณ์ของการขับรถ เพื่อทดสอบประสิทธิภาพของยางรถยนต์ที่ผลิต โดยแบ่งเป็น 4 เกรด คือ A, B, C และ D และทำการศึกษาความสัมพันธ์ระหว่างค่าใช้จ่ายต่อระยะทาง (Y) กับความเร็วของรถ (X_1) จงเขียนโมเดลถดถอยแยกตามเกรดของยางรถยนต์

X_2	X_3	X_4	เกรดของยางรถยนต์
1	0	0	A
0	1	0	B
0	0	1	C
0	0	0	D

ตารางที่ 6.4: การสร้างตัวแปรดัมมี่สำหรับตัวแปรเชิงคุณภาพที่มี 4 ระดับ

วิธีทำ เนื่องจากยางรถยนต์แบ่งเป็น 4 เกรด จึงสร้างตัวแปรดัมมี่ได้ 3 ตัว ดังแสดงในตารางที่ 6.4 สร้างโมเดลถดถอยเชิงเส้นตรงได้เป็น

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon \quad (6.5)$$

จากตารางที่ 6.4 สามารถแยกสมการ (6.5) ออกเป็นสมการย่อย ๆ ตามระดับของตัวแปรดัมมี่ได้ดังนี้

$$\begin{aligned} \text{เกรด } A : \quad Y &= (\beta_0 + \beta_2) + \beta_1 X_1 + \epsilon \\ \text{เกรด } B : \quad Y &= (\beta_0 + \beta_3) + \beta_1 X_1 + \epsilon \\ \text{เกรด } C : \quad Y &= (\beta_0 + \beta_4) + \beta_1 X_1 + \epsilon \\ \text{เกรด } D : \quad Y &= \beta_0 + \beta_1 X_1 + \epsilon \end{aligned}$$

จากโมเดลข้างต้นจะเห็นได้ว่า ค่าใช้จ่ายต่อระยะทางมีความสัมพันธ์เชิงเส้นตรงกับความเร็ว และความชันของสมการไม่ขึ้นอยู่กับเกรดของยางรถยนต์ สำหรับพารามิเตอร์ β_2 , β_3 , และ β_4 แสดงความแตกต่างของเส้นถดถอยเมื่อใช้ยางรถยนต์ที่มีเกรดที่แตกต่างกัน กล่าวคือ β_2 , β_3 , และ β_4 จะวัดอิทธิพลของยางรถยนต์เกรด A, B และ C เมื่อเทียบกับเกรด D ตามลำดับ นอกจากนี้ยังสามารถศึกษาอิทธิพลอื่น ๆ โดยเปรียบเทียบค่าสัมประสิทธิ์ถดถอยที่เกี่ยวข้องได้โดยตรง เช่น $\beta_2 - \beta_3$ จะวัดประสิทธิภาพสัมพัทธ์ (Relative efficiency) ของยางรถยนต์เกรด A เทียบกับเกรด B

จากตัวอย่างข้างต้นดูเหมือนว่าไม่สมเหตุสมผลที่สมมติให้ความชันของสมการถดถอยที่แสดงความสัมพันธ์ระหว่างค่าใช้จ่ายเฉลี่ยและความเร็วของรถไม่ขึ้นอยู่กับเกรดของยางรถยนต์ที่ใช้ ทั้งนี้ค่าใช้จ่ายเฉลี่ยอาจเพิ่มตามความเร็วของรถ แต่อัตราการเพิ่มอาจมีค่าแตกต่างกันตามเกรดของยางรถยนต์ที่ใช้ ซึ่งแสดงถึงการเกิดอิทธิพลร่วมกัน (Interaction effect) ระหว่างความเร็วของรถและเกรดของยางรถยนต์ที่ใช้ ดังนั้นในการ

วิเคราะห์ปัจจัยที่จะรวมเอาอิทธิพลรวมไว้ในโมเดลด้วย ได้โมเดลถดถอยดังนี้

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_1 X_2 + \beta_6 X_1 X_3 + \beta_7 X_1 X_4 + \epsilon \quad (6.6)$$

โดยสามารถแยกเป็นโมเดลย่อย ๆ ได้ 4 โมเดล ตามเกรดของยางรถยนต์

$$\text{เกรด A :} \quad Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_5) X_1 + \epsilon$$

$$\text{เกรด B :} \quad Y = (\beta_0 + \beta_3) + (\beta_1 + \beta_6) X_1 + \epsilon$$

$$\text{เกรด C :} \quad Y = (\beta_0 + \beta_4) + (\beta_1 + \beta_7) X_1 + \epsilon$$

$$\text{เกรด D :} \quad Y = \beta_0 + \beta_1 X_1 + \epsilon$$

เมื่อรวมเอาอิทธิพลรวมเข้าไว้ในโมเดล จะเห็นได้ว่ายางรถยนต์แต่ละเกรดมีสมการถดถอยเฉพาะแยกจากกัน โดยมีระยะตัดแกนและความชันที่แตกต่างกัน

ตัวอย่างที่ 6.3 จากข้อมูลในตัวอย่างที่ 6.1 หากเชื่อว่าอัตราการเปลี่ยนแปลงของราคาขายบ้านต่อระยะทางของเมือง A และ B แตกต่างกัน จงสร้างสมการถดถอยที่มีอิทธิพลร่วมระหว่างระยะทางและสถานที่ตั้งของบ้าน

วิธีทำ พิจารณารูปแบบของโมเดลถดถอยที่มีอิทธิพลร่วมระหว่างระยะทางและสถานที่ตั้งของบ้าน

$$\text{Full model:} \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon \quad (6.7)$$

ซึ่งสามารถแยกเป็นสมการย่อย 2 สมการได้ดังนี้

$$\text{เมือง A :} \quad Y = \beta_0 + \beta_1 X_1 + \epsilon$$

$$\text{เมือง B :} \quad Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 + \epsilon$$

เขียนสมการที่ (6.7) ให้อยู่ในรูปของเมทริกซ์ได้เป็น

$$Y = X\beta + \epsilon$$

เมื่อ

$$X = \begin{bmatrix} 1 & 515 & 0 & 0 \\ 1 & 1025 & 0 & 0 \\ 1 & 680 & 0 & 0 \\ 1 & 860 & 0 & 0 \\ 1 & 1070 & 0 & 0 \\ 1 & 395 & 0 & 0 \\ 1 & 620 & 0 & 0 \\ 1 & 410 & 0 & 0 \\ 1 & 935 & 0 & 0 \\ 1 & 695 & 0 & 0 \\ 1 & 605 & 1 & 605 \\ 1 & 755 & 1 & 755 \\ 1 & 920 & 1 & 920 \\ 1 & 1100 & 1 & 1100 \\ 1 & 740 & 1 & 740 \\ 1 & 485 & 1 & 485 \\ 1 & 965 & 1 & 965 \\ 1 & 575 & 1 & 575 \\ 1 & 815 & 1 & 815 \\ 1 & 350 & 1 & 350 \end{bmatrix}, \quad Y = \begin{bmatrix} 17.595 \\ 11.280 \\ 15.645 \\ 11.310 \\ 9.660 \\ 26.085 \\ 9.510 \\ 23.565 \\ 8.520 \\ 18.480 \\ 34.740 \\ 30.135 \\ 27.600 \\ 28.575 \\ 39.735 \\ 42.930 \\ 28.605 \\ 44.670 \\ 41.925 \\ 55.005 \end{bmatrix}$$

สร้างสมการถดถอยด้วยวิธีกำลังสองน้อยที่สุดได้ผลดังนี้

$$\hat{Y} = 30.274 - 0.021X_1 + 31,178X_2 - 0.012X_1X_2$$

โดยที่ $R^2 = 0.91$ จากนั้นทดสอบนัยสำคัญของสมการ

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_1: \beta_j \neq 0 \text{ อย่างน้อย 1 ค่า, } j = 1, 2, 3$$

ANOVA				
Source of variation	df	SS	MS	F
Regression	3	3226.753	1075.584	54.255
Error	16	317.196	19.825	
Total	19	3543.948		

เนื่องจาก $F_c = 54.255 > F_{3,16}(0.05) = 3.24$ จึงปฏิเสธ H_0 นั่นคือ อย่างน้อยมีตัวแปรอิสระหนึ่งตัวที่มีอิทธิพลต่อราคาขายบ้าน ที่ระดับนัยสำคัญ 0.05

พารามิเตอร์	ค่าประมาณ	ความคลาดเคลื่อนมาตรฐาน	สถิติทดสอบ t
β_0	30.274	4.597	6.59
β_1	-0.021	0.006	-3.45
β_2	31.178	6.721	4.64
β_3	-0.012	0.009	-1.35

ตารางที่ 6.5: การทดสอบนัยสำคัญของค่าสัมประสิทธิ์ถดถอย เมื่อ $t_{0.025,16} = 2.120$

ตารางที่ 6.5 แสดงการทดสอบนัยสำคัญของสัมประสิทธิ์ถดถอยแต่ละตัว และสามารถสรุปได้ว่าตัวแปรอิสระ X_1 และ X_2 มีอิทธิพลต่อโมเดลอย่างมีนัยสำคัญ ที่ระดับ 0.05 แต่อิทธิพลร่วมระหว่าง X_1 และ X_2 ไม่มีนัยสำคัญทางสถิติ ที่ระดับ 0.05

การอนุมานเกี่ยวกับสมการถดถอยเชิงเส้นตรงทั้งสองเส้น ทำได้โดยทดสอบว่าสมการถดถอยทั้งสองสมการเป็นสมการเดียวกัน (Identical equation) หรือไม่ โดยกำหนดสมมติฐานดังนี้

$$H_0 : \beta_2 = \beta_3 = 0$$

$$H_1 : \beta_j \neq 0 \text{ อย่างน้อย 1 ค่า, } j = 2, 3$$

เมื่อ H_0 เป็นจริง Full model จะลดรูปเป็น

$$\text{Reduced model: } Y = \beta_0 + \beta_1 X_1 + \epsilon \quad (6.8)$$

ประมาณพารามิเตอร์ใน Reduced model ได้สมการถดถอยเป็น

$$\hat{Y} = 44.746 - 0.025X_1$$

เมื่อ $SSR(X_1) = 659.262$ และ $SSE(X_1) = 160.260$

การทดสอบจะใช้หลักการของ Extra-sum-of-squares ในการคำนวณ Partial F -test

$$F_c = \frac{SSR(X_2, X_1X_2 | X_1)/2}{SSE(X_1, X_2, X_1X_2)/(n-4)} \quad (6.9)$$

โดยที่

$$\begin{aligned} SSR(X_2, X_1X_2 | X_1) &= SSR_{Full} - SSR_{Reduced} \\ &= SSR(X_1, X_2, X_1X_2) - SSR(X_1) \\ &= 3226.753 - 659.262 \\ &= 2567.491 \end{aligned}$$

ดังนั้น

$$F_c = \frac{2567.491/2}{317.196/(20-4)} = 64.755$$

เนื่องจาก $F_c = 64.755 > F_{2,16}(0.05) = 3.63$ จึงปฏิเสธ H_0 นั่นคือ สมการถดถอยแสดงความสัมพันธ์ระหว่างราคาขายกับระยะทางของบ้านที่ตั้งอยู่ในเมือง A และ B ไม่เป็นสมการเดียวกัน ที่ระดับนัยสำคัญ 0.05

หลังจากนั้นหากต้องการตรวจสอบว่าความชันของทั้งสองสมการมีค่าเท่ากันหรือไม่ สามารถกำหนดสมมติฐานได้เป็น

$$H_0 : \beta_3 = 0 \quad \text{vs.} \quad H_1 : \beta_3 \neq 0$$

เมื่อ H_0 เป็นจริง Full model จะลดรูปเป็น

$$\text{Reduced model:} \quad Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \epsilon \quad (6.10)$$

ซึ่งก็คือโมเดลในตัวอย่างที่ 6.1 นั้นเอง และการทดสอบอาจใช้ t -test หรือ Partial F -test ก็ได้

$$t_c = \frac{b_3}{S_{b_3}} = \frac{-0.012}{0.009} = -1.35$$

ค่าวิกฤติคือ $t_{0.025,16} = 2.120$

เนื่องจาก $|t_c| < 2.120$ ดังนั้นไม่สามารถปฏิเสธ H_0 ได้

หรือ

$$F_c = \frac{SSR(X_1X_2 | X_1, X_2)/1}{SSE(X_1, X_2, X_1X_2)/(n-4)}$$

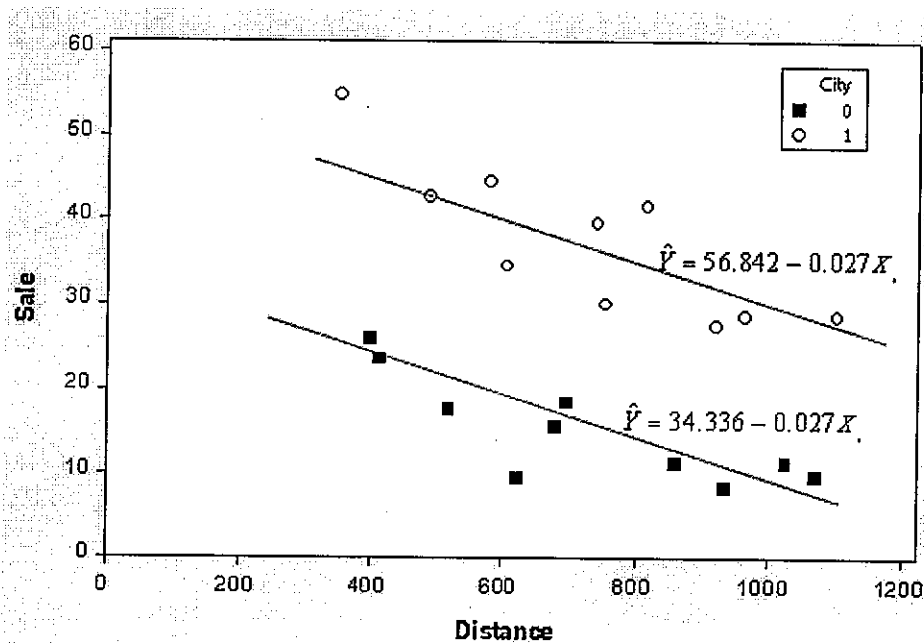
โดยที่

$$\begin{aligned} SSR(X_1X_2 | X_1, X_2) &= SSR_{Full} - SSR_{Reduced} \\ &= SSR(X_1, X_2, X_1X_2) - SSR(X_1, X_2) \\ &= 3226.753 - 3190.576 \\ &= 36.177 \end{aligned}$$

ดังนั้น

$$F_c = \frac{36.177/1}{317.196/(20-4)} = 1.825$$

ค่าวิกฤตคือ $F_{1,16}(0.05) = 4.49$



รูปที่ 6.3: แผนภาพแสดงสมการถดถอยสองสมการที่มีความชันเท่ากัน

เนื่องจาก $F_c < 4.49$ ดังนั้นไม่สามารถปฏิเสธ H_0 ได้ ซึ่งให้ผลการทดสอบที่สอดคล้องกับ t -test นั่นคือไม่มีอิทธิพลร่วมระหว่างระยะทางและสถานที่ตั้งของบ้าน หรือสมการถดถอยแสดงความสัมพันธ์ระหว่างราคาขายกับระยะทางของบ้านที่ตั้งอยู่ในเมือง A และ B มีความชันเท่ากัน ดังแสดงในรูปที่ 6.3

6.2 การวิเคราะห์การถดถอยเมื่อมีตัวแปรต้นมีมากกว่า 1 ตัว

ในทางปฏิบัติบ่อยครั้งที่ต้องรวมเอาตัวแปรเชิงคุณภาพหลาย ๆ ตัวเข้าไปในโมเดล โดยที่การวิเคราะห์ยังคงสามารถทำได้ในลักษณะที่คล้ายคลึงกัน

ตัวอย่างที่ 6.4 จากข้อมูลในตัวอย่างที่ 6.1 หากมีตัวแปรเชิงคุณภาพเพิ่มขึ้นมาอีก 1 ตัว คือ ขนาดของตัวบ้าน แบ่งเป็นปานกลางและใหญ่ เนื่องจากขนาดของตัวบ้านมีเพียง 2 ระดับ ดังนั้นสร้างตัวแปรดัมมี่เพิ่มขึ้น 1 ตัว คือ X_3 โดยที่

$$X_3 = \begin{cases} 0 & \text{ถ้าตัวบ้านมีขนาดปานกลาง} \\ 1 & \text{ถ้าตัวบ้านมีขนาดใหญ่} \end{cases}$$

จงเขียนโมเดลถดถอยของข้อมูลที่กำหนดให้

วิธีทำ โมเดลถดถอยแสดงความสัมพันธ์ระหว่างราคาขาย (Y) กับความเร็วของรถ (X_1) สถานที่ตั้งของตัวบ้าน (X_2) และขนาดของตัวบ้าน (X_3) สามารถแสดงได้ดังนี้

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

จากสมการข้างต้นจะเห็นได้ว่าค่าความชัน β_1 ของโมเดลถดถอยที่แสดงความสัมพันธ์ระหว่างราคาขายเฉลี่ยกับความเร็วของรถไม่ขึ้นอยู่กับสถานที่ตั้งหรือขนาดของตัวบ้าน แต่ระยะตัดแกนจะขึ้นอยู่กับปัจจัยเหล่านี้ในลักษณะเชิงบวก (Additive)

การเพิ่มอิทธิพลร่วมเข้าไปในโมเดลสามารถทำได้หลายลักษณะ เช่น หากต้องการศึกษาอิทธิพลร่วมระหว่างความเร็วรถกับตัวแปรเชิงคุณภาพทั้ง 2 ตัว จะได้โมเดลเป็น

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \epsilon \quad (6.11)$$

ซึ่งสามารถแยกเป็นโมเดลย่อย ๆ ได้ 4 โมเดลดังนี้

สถานที่ตั้ง	ขนาดของตัวบ้าน	โมเดลถดถอย
A	ปานกลาง	$Y = \beta_0 + \beta_1 X_1 + \epsilon$
B	ปานกลาง	$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) X_1 + \epsilon$
A	ใหญ่	$Y = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) X_1 + \epsilon$
B	ใหญ่	$Y = (\beta_0 + \beta_2 + \beta_3) + (\beta_1 + \beta_4 + \beta_5) X_1 + \epsilon$

ตารางที่ 6.6: โมเดลย่อยที่ได้จากโมเดลที่ (6.11)

จากตารางที่ 6.6 จะเห็นได้ว่าในแต่ละคอมบินेशनของสถานที่ตั้งและขนาดของตัวบ้าน จะได้โมเดลถดถอยเฉพาะแยกจากกัน โดยมีระยะตัดแกนและความชันที่แตกต่างกัน แต่อย่างไรก็ตามโมเดลยังคงมีลักษณะเชิงบวกเมื่อเทียบกับระดับของตัวแปรเชิงคุณภาพ นั่นคือ การเปลี่ยนจากบ้านที่มีขนาดปานกลางไปเป็นขนาดใหญ่ ทำให้ระยะตัดแกนและความชันเปลี่ยนแปลงไป β_3 และ β_5 ตามลำดับ โดยไม่คำนึงถึงสถานที่ตั้งของตัวบ้าน

สมมติเพิ่มเติม Cross-product ระหว่างตัวแปรตั้งมีทั้ง 2 ตัวเข้าไปในโมเดล (6.11) จะได้ว่า

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3 + \epsilon \quad (6.12)$$

และแยกเป็นโมเดลย่อย ๆ ได้ดังนี้

สถานที่ตั้ง	ขนาดของตัวบ้าน	โมเดลถดถอย
A	ปานกลาง	$Y = \beta_0 + \beta_1 X_1 + \epsilon$
B	ปานกลาง	$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) X_1 + \epsilon$
A	ใหญ่	$Y = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) X_1 + \epsilon$
B	ใหญ่	$Y = (\beta_0 + \beta_2 + \beta_3 + \beta_6) + (\beta_1 + \beta_4 + \beta_5) X_1 + \epsilon$

ตารางที่ 6.7: โมเดลย่อยที่ได้จากโมเดลที่ (6.12)

จะเห็นได้ว่าการเพิ่มเติม $\beta_6 X_2 X_3$ เข้าไว้ในโมเดลที่ (6.12) จะทำให้อธิพจน์ของตัวแปรเชิงคุณภาพตัวหนึ่งที่มีต่อระยะตัดแกนขึ้นอยู่กับระดับของตัวแปรเชิงคุณภาพอีกตัวหนึ่ง นั่นคือ การเปลี่ยนขนาดของตัวบ้านจากปานกลางไปเป็นขนาดใหญ่ จะทำให้ระยะตัดแกนเปลี่ยนแปลงไป β_3 และ $\beta_3 + \beta_6$ หน่วย ถ้าบ้านตั้งอยู่ในเมือง A และเมือง B ตามลำดับ

หากเพิ่มเติม $\beta_7 X_1 X_2 X_3$ เข้าไปในโมเดลที่ (6.12) จะได้ว่า

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3 + \beta_7 X_1 X_2 X_3 + \epsilon \quad (6.13)$$

ได้โมเดลย่อย ๆ ดังนี้

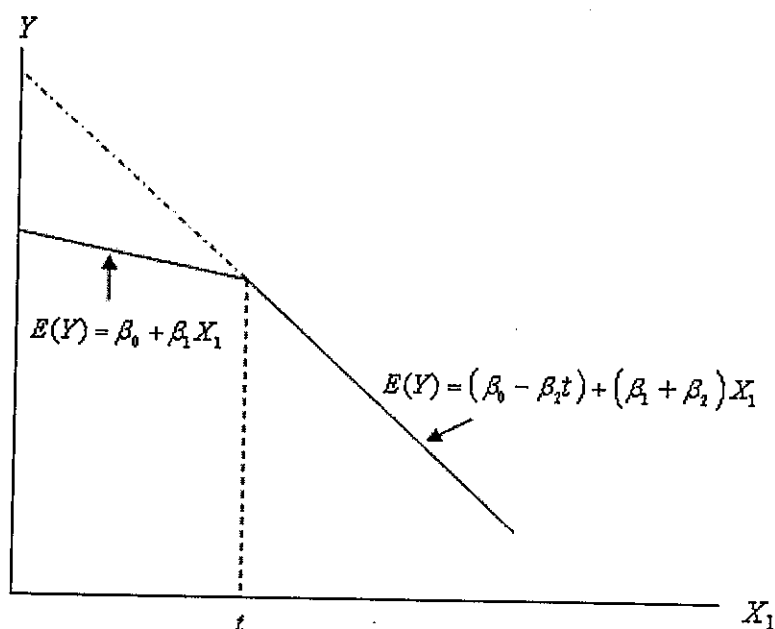
สถานที่ตั้ง	ขนาดของตัวบ้าน	โมเดลถดถอย
A	ปานกลาง	$Y = \beta_0 + \beta_1 X_1 + \epsilon$
B	ปานกลาง	$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) X_1 + \epsilon$
A	ใหญ่	$Y = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) X_1 + \epsilon$
B	ใหญ่	$Y = (\beta_0 + \beta_2 + \beta_3 + \beta_6) + (\beta_1 + \beta_4 + \beta_5 + \beta_7) X_1 + \epsilon$

ตารางที่ 6.8: โมเดลย่อยที่ได้จากโมเดลที่ (6.13)

นั่นคือ การเปลี่ยนขนาดของตัวบ้านจากปานกลางเป็นขนาดใหญ่ จะมีผลกระทบต่อทั้งระยะตัดแกนและความชัน ซึ่งขึ้นอยู่กับสถานที่ตั้งของตัวบ้าน

6.3 การวิเคราะห์การถดถอยแบบแบ่งเป็นช่วง (Piecewise Linear Regression)

ในบางครั้งตัวแปรตาม Y และตัวแปรอิสระ X อาจมีความสัมพันธ์เชิงเส้นตรงที่มีรูปแบบเฉพาะในพิสัยของ X ช่วงหนึ่ง ๆ แต่เมื่อพิสัยของ X เปลี่ยนไป รูปแบบความสัมพันธ์ระหว่าง Y และ X อาจแตกต่างกันไปจากเดิม เช่น การศึกษาความสัมพันธ์ระหว่างค่าใช้จ่ายต่อหน่วย (Y) กับปริมาณที่ผลิต (X) ซึ่งพบว่าความสัมพันธ์มีลักษณะเป็นเส้นตรงที่มีรูปแบบเฉพาะ จนกระทั่งเมื่อปริมาณการผลิตถึงจุด ๆ หนึ่ง นั่นคือ $X = t$ ความชันของเส้นถดถอยอาจเปลี่ยนแปลงไป เนื่องจากประสิทธิภาพของการผลิตเมื่อผลิตมากกว่า t หน่วย



รูปที่ 6.4: แผนภาพแสดงสมการถดถอยแบบแบ่งเป็นช่วง

รูปที่ 6.4 แสดงสมการถดถอยแบบแบ่งเป็นช่วง ซึ่งการสร้างสมการถดถอยแบบแบ่งเป็นช่วงนี้ สามารถใช้หลักของการสร้างตัวแปรดัมมี่เข้ามาช่วยในการวิเคราะห์ได้ ในที่นี้จะแสดงเฉพาะสมการถดถอยที่แบ่งเป็น 2 ช่วง

เท่านั้น และค่าของ $X = t$ ที่ทำให้สมการแบ่งเป็น 2 ช่วงเป็นค่าที่ทราบค่า ซึ่งมักจะถูกเรียกว่า *Knot* หรือ *Threshold*

รูปแบบทั่วไปของสมการถดถอยแบบแบ่งเป็นช่วงคือ

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 (X_1 - t) + \epsilon \quad (6.14)$$

โดยที่

$$X_2 = \begin{cases} 0 & \text{เมื่อ } X_1 \leq t \\ 1 & \text{เมื่อ } X_1 > t \end{cases}$$

โมเดลที่ (6.14) สามารถแยกได้เป็น 2 โมเดลย่อย ดังนี้

$$Y = \beta_0 + \beta_1 X_1 + \epsilon \quad \text{เมื่อ } X_1 \leq t \quad (6.15)$$

$$Y = (\beta_0 - \beta_2 t) + (\beta_1 + \beta_2) X_1 + \epsilon \quad \text{เมื่อ } X_1 > t \quad (6.16)$$

โดยที่ β_1 และ $\beta_1 + \beta_2$ แทน ความชันของสมการถดถอยเมื่อ $X_1 \leq t$ และ $X_1 > t$ ตามลำดับ

β_0 และ $\beta_0 - \beta_2 t$ แทน ระยะตัดแกนของสมการถดถอยเมื่อ $X_1 \leq t$ และ $X_1 > t$ ตามลำดับ

การทดสอบว่ามีการเปลี่ยนแปลงของสมการถดถอยที่จุด $X_1 = t$ หรือไม่ สามารถทำได้โดยการตรวจสอบนัยสำคัญของพารามิเตอร์ β_2

ตัวอย่างที่ 6.5 ข้อมูลต่อไปนี้แสดงต้นทุนของการผลิตสินค้าชนิดหนึ่ง (พันบาท) กับจำนวนสินค้าที่ผลิต (หน่วย)

Y	256	326	414	550	634	778	861	1003	1440
X	1000	1500	2000	2500	3000	4000	4500	5000	5500
Y	1590	2020	2210	2320	2655	2870			
X	6000	7000	7500	8000	9000	10000			

ตารางที่ 6.9: ข้อมูลแสดงต้นทุน (Y) และจำนวนสินค้าที่ผลิต (X)

1. จงสร้างสมการถดถอยแบบแบ่งเป็นช่วงแสดงความสัมพันธ์ระหว่างต้นทุนและจำนวนสินค้าที่ผลิต เมื่อจำนวนสินค้าไม่เกิน 5500 หน่วย และมากกว่า 5500 หน่วย
2. ที่ระดับนัยสำคัญ 0.05 จงทดสอบว่าสมการถดถอยแบบแบ่งเป็นช่วงเหมาะสมกับข้อมูลชุดนี้หรือไม่

วิธีทำ

1. ในที่นี้โมเดลของสมการถดถอยแบบแบ่งเป็นช่วงคือ

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 (X_1 - 5500) + \epsilon$$

โดยที่

$$X_2 = \begin{cases} 0 & \text{เมื่อ } X_1 \leq 5500 \\ 1 & \text{เมื่อ } X_1 > 5500 \end{cases}$$

สร้างตารางแสดงค่าของตัวแปรต่าง ๆ ได้ดังนี้

Y	X_1	X_2	$X_2(X_1 - 5500)$
256	1000	0	0
326	1500	0	0
414	2000	0	0
550	2500	0	0
634	3000	0	0
778	4000	0	0
861	4500	0	0
1003	5000	0	0
1440	5500	0	0
1590	6000	1	500
2020	7000	1	1500
2210	7500	1	2000
2320	8000	1	2500
2655	9000	1	3500
2870	10000	1	4500

ตารางที่ 6.10: แสดงข้อมูลที่ใช้ในการสร้างสมการแบบช่วงของตัวอย่างที่ 6.5

ประมาณพารามิเตอร์โดยสร้างเมตริกซ์ X และ Y ได้ดังนี้

$$X = \begin{bmatrix} 1 & 1000 & 0 \\ 1 & 1500 & 0 \\ 1 & 2000 & 0 \\ 1 & 2500 & 0 \\ 1 & 3000 & 0 \\ 1 & 4000 & 0 \\ 1 & 4500 & 0 \\ 1 & 5000 & 0 \\ 1 & 5500 & 0 \\ 1 & 6000 & 500 \\ 1 & 7000 & 1500 \\ 1 & 7500 & 2000 \\ 1 & 8000 & 2500 \\ 1 & 9000 & 3500 \\ 1 & 10000 & 4500 \end{bmatrix}, \quad Y = \begin{bmatrix} 256 \\ 326 \\ 414 \\ 550 \\ 634 \\ 778 \\ 861 \\ 1003 \\ 1440 \\ 1590 \\ 2020 \\ 2210 \\ 2320 \\ 2655 \\ 2870 \end{bmatrix}$$

ได้สมการถดถอยเป็น

$$\hat{Y} = -93.140 + 0.254X_1 + 0.128X_2(X_1 - 5500)$$

ซึ่งแยกได้ดังนี้

$$\text{เมื่อ } X_1 \leq 5500 : \quad \hat{Y} = -93.140 + 0.254X_1$$

$$\begin{aligned} \text{เมื่อ } X_1 > 5500 : \quad \hat{Y} &= (-93.140 - 704) + (0.254 + 0.128)X_1 \\ &= -797.14 + 0.382X_1 \end{aligned}$$

จะเห็นว่าเมื่อจำนวนสินค้าที่ผลิตเพิ่มขึ้น 1 หน่วย ต้นทุนเฉลี่ยจะสูงขึ้น 0.254 และ 0.382 หน่วย สำหรับ สินค้าที่ผลิตไม่เกิน 5500 หน่วย และมากกว่า 5500 หน่วย ตามลำดับ

2. การคำนวณค่าสัมประสิทธิ์ถดถอย ความคลาดเคลื่อนมาตรฐาน และสถิติทดสอบ t ได้ผลดังนี้

พารามิเตอร์	ค่าประมาณ	ความคลาดเคลื่อนมาตรฐาน	สถิติทดสอบ t
β_0	-93.140	95.300	-0.98
β_1	0.254	0.025	10.33
β_2	0.128	0.046	2.77

ตารางที่ 6.11: การทดสอบนัยสำคัญของค่าสัมประสิทธิ์ถดถอย เมื่อ $t_{0.025, 12} = 2.179$

ตารางที่ 6.11 แสดงการทดสอบนัยสำคัญของสัมประสิทธิ์ถดถอยแต่ละตัว โดยกำหนดสมมติฐานเป็น

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0 \quad (j = 0, 1, 2)$$

และสามารถสรุปได้ว่า X_1 และ $X_2(X_1 - 5500)$ มีอิทธิพลต่อโมเดลอย่างมีนัยสำคัญ ที่ระดับ 0.05 นั่นคือสมการถดถอยแบบแบ่งเป็นช่วง เมื่อจำนวนการผลิตไม่เกิน 5500 หน่วย และมากกว่า 5500 หน่วย มีความเหมาะสม

หมายเหตุ การสร้างสมการถดถอยแบบแบ่งเป็นช่วงสามารถขยายไปสู่กรณีที่มีมากกว่า 2 ช่วงก็ได้ เช่น ข้อมูลในตัวอย่างที่ 6.5 หากพบว่าความชันของสมการถดถอยเปลี่ยนแปลงเมื่อ $X_1 = 2500$ และ $X_1 = 5500$ สามารถสร้างสมการถดถอยแบบแบ่งช่วงได้ดังนี้

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2(X_1 - 2500) + \beta_3 X_3(X_1 - 5500) + \epsilon \quad (6.17)$$

โดยที่

$$X_2 = \begin{cases} 0 & \text{เมื่อ } X_1 \leq 2500 \\ 1 & \text{กรณีอื่น} \end{cases}$$

$$X_3 = \begin{cases} 0 & \text{เมื่อ } X_1 \leq 5500 \\ 1 & \text{กรณีอื่น} \end{cases}$$

6.4 ความไม่ต่อเนื่องของฟังก์ชันถดถอย

(Discontinuity in Regression Function)

ในบางกรณีที่ฟังก์ชันถดถอยเชิงเส้นตรงไม่เพียงแต่เปลี่ยนความชันที่จุด $X = t$ แต่ยังขาดความต่อเนื่องที่จุดดังกล่าวด้วย ซึ่งการวิเคราะห์ยังคงสามารถทำได้โดยสร้างตัวแปรดัมมี่ขึ้นมามากอีกหนึ่งตัวแทนลักษณะที่ไม่ต่อเนื่อง

ของฟังก์ชัน รูปแบบทั่วไปของโมเดลถดถอยที่มีลักษณะไม่ต่อเนื่องและเปลี่ยนค่าความชันที่จุด $X_1 = t$ คือ

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 (X_1 - t) X_2 + \beta_3 X_3 + \epsilon \quad (6.18)$$

โดยที่

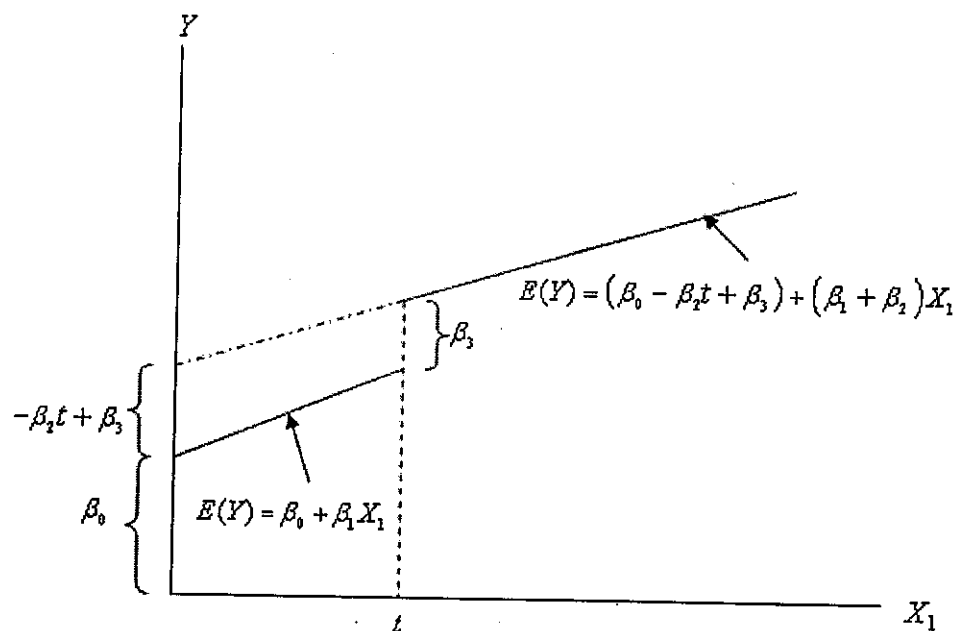
$$X_2 = \begin{cases} 1 & \text{เมื่อ } X_1 > t \\ 0 & \text{กรณีอื่น} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{เมื่อ } X_1 > t \\ 0 & \text{กรณีอื่น} \end{cases}$$

โมเดลที่ (6.18) สามารถแยกได้เป็น 2 โมเดลย่อย ดังนี้

$$Y = \beta_0 + \beta_1 X_1 + \epsilon \quad \text{เมื่อ } X_1 < t \quad (X_2 = X_3 = 0)$$

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 (X_1 - t)(1) + \beta_3(1) + \epsilon \quad \text{เมื่อ } X_1 \geq t \quad (X_2 = X_3 = 1) \\ &= (\beta_0 - \beta_2 t + \beta_3) + (\beta_1 + \beta_2) X_1 \end{aligned}$$



รูปที่ 6.5: แผนภาพแสดงลักษณะที่ไม่ต่อเนื่องของสมการถดถอย

รูปที่ 6.5 แสดงลักษณะที่ไม่ต่อเนื่องของสมการถดถอย โดย β_2 แทนความแตกต่างของความชันในสมการถดถอยทั้งสองสมการ และ β_3 แทนความแตกต่างของค่าเฉลี่ยของ Y ที่จุด $X_1 = t$ ซึ่งได้จากสมการทั้งสอง นอกจากนี้ยังสามารถทดสอบสมมติฐานเกี่ยวกับ $\beta_3 = 0$ ได้ หากผลการทดสอบยอมรับว่า $\beta_3 = 0$ แล้วแสดงว่าสมการถดถอยต่อเนื่องที่จุด $X_1 = t$ จากนั้นสร้างสมการถดถอยแบบแบ่งเป็นช่วงที่มีลักษณะต่อเนื่องแต่เปลี่ยนความชันที่จุด $X_1 = t$

ตัวอย่างที่ 6.6 จากข้อมูลในตัวอย่างที่ 6.5 หากผู้วิจัยเชื่อว่าความสัมพันธ์ระหว่างต้นทุนและจำนวนสินค้าที่ผลิตมีลักษณะไม่ต่อเนื่อง

1. จงสร้างสมการถดถอยแบบแบ่งเป็นช่วง (เมื่อ $X_1 \leq 5500$ และ $X_1 > 5500$) โดยให้สมการมีค่าไม่ต่อเนื่องที่จุด $X_1 = 5500$
2. ที่ระดับนัยสำคัญ 0.05 จงทดสอบว่าสมการในข้อ 1 มีลักษณะไม่ต่อเนื่องที่จุด $X_1 = 5500$ หรือไม่
3. ที่ระดับนัยสำคัญ 0.05 จงทดสอบว่าสมการถดถอยทั้งสองสมการที่ได้จากข้อ 1 ขนานกันหรือไม่

วิธีทำ

1. สร้างตารางแสดงการคำนวณของข้อมูลได้ดังนี้

Y	X_1	X_2	$X_2(X_1 - 5500)$	X_3
256	1000	0	0	0
326	1500	0	0	0
414	2000	0	0	0
550	2500	0	0	0
634	3000	0	0	0
778	4000	0	0	0
861	4500	0	0	0
1003	5000	0	0	0
1440	5500	0	0	0
1590	6000	1	500	1
2020	7000	1	1500	1
2210	7500	1	2000	1
2320	8000	1	2500	1
2655	9000	1	3500	1
2870	10000	1	4500	1

ตารางที่ 6.12: แสดงข้อมูลที่ใช้ในการสร้างสมการถดถอยแบบช่วงเมื่อมีลักษณะไม่ต่อเนื่อง

ประมาณพารามิเตอร์โดยสร้างเมตริกซ์ X และ Y ได้ดังนี้

$$\mathbf{X} = \begin{bmatrix} 1 & 1000 & 0 & 0 \\ 1 & 1500 & 0 & 0 \\ 1 & 2000 & 0 & 0 \\ 1 & 2500 & 0 & 0 \\ 1 & 3000 & 0 & 0 \\ 1 & 4000 & 0 & 0 \\ 1 & 4500 & 0 & 0 \\ 1 & 5000 & 0 & 0 \\ 1 & 5500 & 0 & 0 \\ 1 & 6000 & 500 & 1 \\ 1 & 7000 & 1500 & 1 \\ 1 & 7500 & 2000 & 1 \\ 1 & 8000 & 2500 & 1 \\ 1 & 9000 & 3500 & 1 \\ 1 & 10000 & 4500 & 1 \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} 256 \\ 326 \\ 414 \\ 550 \\ 634 \\ 778 \\ 861 \\ 1003 \\ 1440 \\ 1590 \\ 2020 \\ 2210 \\ 2320 \\ 2655 \\ 2870 \end{bmatrix}$$

ได้สมการถดถอยเป็น

$$\hat{Y} = -24.350 + 0.223X_1 + 0.092X_2(X_1 - 5500) + 308.300X_3$$

ซึ่งแยกได้ดังนี้

$$\text{เมื่อ } X_1 \leq 5500 : \quad \hat{Y} = -24.350 + 0.223X_1$$

$$\begin{aligned}
 \text{เมื่อ } X_1 > 5500 : \quad \hat{Y} &= (-24.350 - 506 + 308.3) + (0.223 + 0.092)X_1 \\
 &= -222.050 + 0.315X_1
 \end{aligned}$$

2. การคำนวณค่าสัมประสิทธิ์ถดถอย ความคลาดเคลื่อนมาตรฐาน และสถิติทดสอบ t ได้ผลดังนี้

พารามิเตอร์	ค่าประมาณ	ความคลาดเคลื่อนมาตรฐาน	สถิติทดสอบ t
β_0	-24.350	77.75	-0.31
β_1	0.223	0.022	10.23
β_2	0.093	0.038	2.45
β_3	308.300	104.000	2.96

ตารางที่ 6.13: การทดสอบนัยสำคัญของค่าสัมประสิทธิ์ถดถอย เมื่อ $t_{0.025, 11} = 2.201$

กำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \beta_3 = 0 \quad \text{vs.} \quad H_1 : \beta_3 \neq 0$$

สถิติทดสอบ:

$$t_c = \frac{b_3}{S_{b_3}} = \frac{308.300}{104.000} = 2.96$$

ค่าวิกฤติคือ $t_{0.025, 11} = 2.201$

สรุปผล เนื่องจาก $t_c > 2.201$ จึงปฏิเสธ H_0 นั่นคือ สมการถดถอยมีลักษณะไม่ต่อเนื่องที่จุด $X_1 = 5500$ ที่ระดับนัยสำคัญ 0.05

3. กำหนดสมมติฐานของการทดสอบได้ดังนี้

$$H_0 : \beta_2 = 0 \quad \text{vs.} \quad H_1 : \beta_2 \neq 0$$

สถิติทดสอบ:

$$t_c = \frac{b_2}{S_{b_2}} = \frac{0.093}{0.038} = 2.45$$

ค่าวิกฤติคือ $t_{0.025, 11} = 2.201$

สรุปผล เนื่องจาก $t_c > 2.201$ จึงปฏิเสธ H_0 นั่นคือ สมการถดถอยทั้งสองสมการมีความชันเท่ากันหรือขนานกัน ที่ระดับนัยสำคัญ 0.05

นอกจากนี้การสร้างตัวแปรดัมนั้นยังสามารถประยุกต์ใช้กับข้อมูลที่เก็บรวบรวมตามลำดับเวลา (Time series data) ได้อีกด้วย เช่น การศึกษาความสัมพันธ์ระหว่างยอดขายรถยนต์และค่าใช้จ่ายในการโฆษณา ซึ่งรูปแบบความสัมพันธ์อาจขึ้นอยู่กับสภาพเศรษฐกิจในแต่ละปีด้วย ดังนั้นในการวิเคราะห์จะสร้างตัวแปรดัมนี่แทน

สภาพเศรษฐกิจ และเขียนโมเดลถดถอยได้ดังนี้

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \epsilon, \quad t = 1, 2, \dots, T$$

โดยที่

X_{t1} แทน ค่าใช้จ่ายในการโฆษณา

$$X_{t2} \text{ แทน สภาพเศรษฐกิจ เมื่อ } X_{t2} = \begin{cases} 1 & \text{ถ้าสภาพเศรษฐกิจปกติ} \\ 0 & \text{กรณีอื่น} \end{cases}$$

กรณีที่ข้อมูลอนุกรมเวลาเก็บรวบรวมตามรายเดือนหรือรายไตรมาส เช่น ข้อมูลเกี่ยวกับยอดขายของห้างสรรพสินค้าและค่าใช้จ่ายในการโฆษณา หากมีอิทธิพลของฤดูกาลเข้ามาเกี่ยวข้องแล้ว การวิเคราะห์อาจรวมเอาอิทธิพลของฤดูกาลเข้ากับสมการถดถอยเชิงเส้นตรงได้ดังนี้

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \beta_3 X_{t3} + \beta_4 X_{t4} + \epsilon, \quad t = 1, 2, \dots, T$$

โดยที่

X_{t1} แทน ค่าใช้จ่ายในการโฆษณารายไตรมาส

$$X_{tj} = \begin{cases} 1 & \text{ถ้าเป็นไตรมาสที่ } j \text{ เมื่อ } j = 2, 3, 4 \\ 0 & \text{ถ้าเป็นไตรมาสอื่น} \end{cases}$$

แต่อย่างไรก็ตามปัญหาของการสร้างสมการถดถอยสำหรับข้อมูลอนุกรมเวลาก็คือ ความคลาดเคลื่อนมักจะมีความสัมพันธ์ต่อกัน ดังนั้นจึงควรที่จะตรวจสอบเสมอว่าการสร้างสมการถดถอยสำหรับข้อมูลอนุกรมเวลาโดยใช้ตัวแปรต้นมีเพียงพอที่จะทำให้ความคลาดเคลื่อนไม่มีความสัมพันธ์ต่อกันได้หรือไม่

หมายเหตุ

1. การกำหนดตัวเลขให้กับตัวแปรอิสระที่เป็นตัวแปรเชิงคุณภาพ นอกจากการสร้างตัวแปรต้นมีแล้ว ยังสามารถกำหนดโดยใช้ตัวเลขในลักษณะอื่นที่ไม่ใช่ 0 และ 1 เพื่อแทนระดับย่อย ๆ ของตัวแปรอิสระได้อีกด้วย ซึ่งเรียกว่า *Allocated codes* เช่น พิจารณาความรุนแรงของโรค ซึ่งแบ่งเป็น 3 ระดับ คือ น้อย ปานกลาง และมาก การสร้าง *Allocated codes* เพื่อแทนระดับความรุนแรงของโรค ทำได้โดยสร้างตัวแปรอิสระ X_1 ขึ้นมาเพียงตัวเดียว และกำหนดค่าดังนี้

ระดับความรุนแรงของโรค	X_1
รุนแรงมาก	3
รุนแรงปานกลาง	2
รุนแรงน้อย	1

ตารางที่ 6.14: การใช้ Allocated codes ในการกำหนดตัวเลขให้กับตัวแปรเชิงคุณภาพ

การกำหนดตัวเลขใน Allocated codes ไม่เฉพาะเจาะจง อาจใช้ตัวเลขชุดอื่นที่แตกต่างไปจากที่แสดงในตารางที่ 6.14 เพื่อแทนระดับย่อย ๆ ของตัวแปรอิสระได้เช่นกัน หากรูปแบบของโมเดลกำลังหนึ่งเหมาะสม และไม่มีตัวแปรอิสระอื่น นั่นคือ

$$Y = \beta_0 + \beta_1 X_1 + \epsilon \quad (6.19)$$

และสามารถแยกเป็นโมเดลย่อย ๆ ได้ดังนี้

ระดับความรุนแรงของโรค	โมเดล
รุนแรงมาก	$Y = \beta_0 + 3\beta_1 + \epsilon$
รุนแรงปานกลาง	$Y = \beta_0 + 2\beta_1 + \epsilon$
รุนแรงน้อย	$Y = \beta_0 + \beta_1 + \epsilon$

ตารางที่ 6.15: โมเดลที่ได้จากการใช้ Allocated codes

จะเห็นได้ว่าการใช้ Allocated codes เป็น 1, 2, 3 ทำให้ตัวแปรตามเฉลี่ยเปลี่ยนแปลงไป β_1 หน่วยเท่ากัน เมื่อความรุนแรงของโรคเปลี่ยนจากระดับน้อยไปปานกลาง และปานกลางไปมาก ทั้งนี้เพราะการกำหนดตัวเลขเป็น 1, 2, 3 เป็นการกำหนดระยะห่างระหว่างความรุนแรงของโรคทั้ง 3 ระดับเท่ากัน ซึ่งอาจไม่สมเหตุสมผลในทางปฏิบัติ นอกจากนี้การใช้ Allocated codes ที่แตกต่างไป ก็จะทำให้ผลที่แตกต่างกัน ในทางตรงข้ามการใช้ตัวแปรดัมมี่ไม่มีข้อกำหนดเกี่ยวกับระยะห่างระหว่างระดับย่อย ๆ ของตัวแปรอิสระและขึ้นอยู่กับข้อมูล โดยจะแสดงถึงผลของความแตกต่าง (Differential effects) ที่เกิดขึ้น หากสร้างตัวแปรดัมมี่ X_1 และ X_2 แทนระดับความรุนแรงของโรค จะได้โมเดลกำลังหนึ่งเป็น

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \quad (6.20)$$

ตารางที่ 6.16 แสดงค่าของตัวแปรดัมมี่และโมเดลย่อยที่ได้จากโมเดลที่ (6.20) โดย β_1 และ β_2 วัดผลต่างของตัวแปรตามเฉลี่ยเมื่อเปลี่ยนระดับความรุนแรงของโรคจากน้อยไปมากและน้อยไปปานกลางตามลำดับ และ $\beta_1 - \beta_2$ วัดผลต่างของตัวแปรตามเฉลี่ยเมื่อเปลี่ยนระดับความรุนแรงของโรคจากปานกลาง

ระดับความรุนแรงของโรค	X_1	X_2	โมเดล
รุนแรงมาก	1	0	$Y = \beta_0 + \beta_1 X_1 + \epsilon$
รุนแรงปานกลาง	0	1	$Y = \beta_0 + \beta_2 X_2 + \epsilon$
รุนแรงน้อย	0	0	$Y = \beta_0 + \epsilon$

ตารางที่ 6.16: การสร้างตัวแปรดัมมี่ให้กับระดับความรุนแรงของโรค และโมเดลย่อย

ไปมาก จะเห็นได้ว่าการใช้ตัวแปรดัมมี่ไม่มีข้อจำกัดเกี่ยวกับระยะห่างของผลต่างเหล่านี้ นอกจากนี้การสร้างสมการถดถอยเมื่อใช้ตัวแปรดัมมี่จะให้ค่า R^2 สูงกว่าการใช้ Allocated codes เสมอ

- บางครั้งการเก็บรวบรวมข้อมูลเชิงปริมาณเพื่อให้ได้ข้อมูลที่ถูกต้องอย่างแท้จริงอาจทำได้ยาก เช่น รายได้ครัวเรือน หรือค่าใช้จ่ายต่อเดือน เป็นต้น ซึ่งการนำเอาข้อมูลดังกล่าวไปวิเคราะห์ อาจทำได้โดยแบ่งกลุ่มให้กับข้อมูลเชิงปริมาณเหล่านี้ แล้วสร้างตัวแปรดัมมี่เพื่อแทนกลุ่มของข้อมูล เช่น การศึกษาค่าใช้จ่ายต่อเดือนของนิสิตในมหาวิทยาลัย อาจแบ่งกลุ่มของค่าใช้จ่ายได้ดังนี้

0	ถึง	2,999 บาท
3,000	ถึง	5,999 บาท
6,000	ถึง	8,999 บาท
9,000	ถึง	11,999 บาท
ตั้งแต่	12,000 บาท	

การวิเคราะห์ทำได้โดยสร้างตัวแปรดัมมี่ 4 ตัว เพื่อแทนกลุ่มของค่าใช้จ่าย แต่ข้อเสียของการเปลี่ยนข้อมูลเชิงปริมาณให้เป็นข้อมูลเชิงคุณภาพก็คือ ทำให้โมเดลซับซ้อนมากขึ้น เนื่องจากมีพารามิเตอร์ที่ต้องประมาณเพิ่มขึ้น และจำนวนองศาความเป็นอิสระของความคลาดเคลื่อนมีค่าลดลง ส่วนข้อดีของการใช้ตัวแปรดัมมี่ก็คือ ไม่ต้องกำหนดรูปแบบความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระล่วงหน้า

- การวิเคราะห์การถดถอยเชิงเส้นตรงในกรณีที่มีตัวแปรอิสระทุกตัวเป็นตัวแปรเชิงคุณภาพ จะเรียกการวิเคราะห์ดังกล่าวนี้ว่า การวิเคราะห์ความแปรปรวน (Analysis of variance)

แบบฝึกหัดบทที่ 6

1. สุ่มนักศึกษาจากมหาวิทยาลัยแห่งหนึ่งมา 20 คน เพื่อศึกษาว่า GPA ของนักศึกษาเมื่อสิ้นปีการศึกษา ที่หนึ่ง (Y) สามารถพยากรณ์จากคะแนนสอบคัดเลือกเข้ามหาวิทยาลัย (X_1) ได้หรือไม่ ได้ข้อมูลดังนี้

Y	3.1	2.3	3.0	1.9	2.5	3.7	3.4	2.6	2.8	1.6	2.0	2.9	2.3	3.2	1.8
X_1	5.5	4.8	4.7	3.9	4.5	6.2	6.0	5.2	4.7	4.3	4.9	5.4	5.0	6.3	4.6
Y	1.4	2.0	3.8	2.2	1.5										
X_1	4.3	5.0	5.9	4.1	4.7										

ต่อมาทางมหาวิทยาลัยคาดว่ารูปแบบความสัมพันธ์อาจเหมาะสมมากขึ้นถ้ามีการพิจารณาตัวแปรอื่นเพิ่มเข้ามาในสมการ จึงได้พิจารณาจากสาขาวิชาหลักที่นักศึกษาระบุไว้ในใบสมัครสอบ โดยให้ $X_2 = 1$ ถ้านักศึกษาระบุสาขาวิชาหลักไว้ในใบสมัคร และ $X_2 = 0$ ถ้านักศึกษาไม่ได้ระบุสาขาวิชาหลักไว้ จากการเก็บข้อมูลพบว่า นักศึกษาคนที่ 3, 4, 5, 6, 9 และ 17 ได้ระบุสาขาวิชาหลักไว้ในใบสมัคร

- 1.1 จงสร้างสมการถดถอยแสดงอิทธิพลร่วม (Interaction effect) ระหว่างตัวแปร X_1 และ X_2 พร้อมทั้งอธิบายความหมายของอิทธิพลร่วมดังกล่าว
 - 1.2 ที่ระดับนัยสำคัญ 0.05 จงทดสอบนัยสำคัญของอิทธิพลร่วม และแปลผลที่ได้ ท่านคิดว่าสามารถตัดเทอมดังกล่าวออกจากสมการได้หรือไม่
 - 1.3 จงสร้างสมการถดถอยที่มีเฉพาะอิทธิพลเชิงบวก (Additive effect) และทดสอบนัยสำคัญของ X_1 และ X_2 ที่ระดับ 0.05 พร้อมทั้งแปลผลที่ได้
 - 1.4 จงใช้สมการที่ได้ในข้อ 1.3 สร้างสมการถดถอยสำหรับนักศึกษาที่ระบุสาขาวิชาหลัก และไม่ระบุสาขาวิชาหลักไว้ในใบสมัคร ท่านคิดว่าสมการทั้งสองมีค่าใดที่เหมือนกัน และค่าใดที่ต่างกัน
 - 1.5 จงสร้างกราฟของความคลาดเคลื่อน (Residual plot) และ Normal probability plot เพื่อตรวจสอบว่ารูปแบบสมการในข้อ 1.3 เหมาะสมหรือไม่ อธิบายผลที่ได้
2. ในการศึกษาความสัมพันธ์ระหว่างราคาขายบ้าน (Y) หน่วยเป็นหมื่นบาท กับขนาดของบ้าน (X_1) หน่วยเป็น 100 ตารางฟุต และสภาพของบ้าน (Rating) แบ่งเป็น 3 ระดับคือ แย่ พอใช้ และดีมาก ได้ข้อมูลดังนี้

Y	X ₁	สภาพของบ้าน
60.0	23	พอใช้
32.7	11	แย่
57.7	20	ดีมาก
45.5	17	แย่
47.0	15	ดีมาก
55.3	21	พอใช้
64.5	24	พอใช้
42.6	13	พอใช้
54.5	19	พอใช้
57.5	25	แย่

- 2.1 จงสร้างตัวแปรดัมมี่ X_2 และ X_3 โดยให้ $X_2 = 1$ ถ้าสภาพของบ้านดีมาก และ $X_2 = 0$ ถ้าสภาพบ้านเป็นแบบอื่น $X_3 = 1$ ถ้าสภาพของบ้านพอใช้ และ $X_3 = 0$ ถ้าสภาพบ้านเป็นแบบอื่น โดยสร้างตารางแสดง Combination ของค่าที่เป็นไปได้ทั้งหมดของตัวแปรดัมมี่ X_2 และ X_3 พร้อมทั้งระบุความหมายของแต่ละ Combination
 - 2.2 จงสร้างสมการถดถอยที่มีอิทธิพลร่วม (Interaction effect) ระหว่างตัวแปร X_1 , X_2 และ X_1 , X_3
 - 2.3 ที่ระดับนัยสำคัญ 0.05 จงทดสอบนัยสำคัญของอิทธิพลร่วมในโมเดลและแปลผลที่ได้ ท่านคิดว่าสามารถตัดเทอมดังกล่าวออกจากโมเดลได้หรือไม่ เพราะเหตุใด
 - 2.4 จงสร้างสมการถดถอยที่มีเฉพาะอิทธิพลเชิงบวก (Additive effect) และทดสอบนัยสำคัญของ X_1 , X_2 และ X_3 ที่ระดับ 0.05 พร้อมทั้งแปลผลที่ได้ และคำนวณค่าสัมประสิทธิ์ของตัวกำหนดพร้อมอธิบายความหมาย
 - 2.5 ท่านคิดว่ารูปแบบความสัมพันธ์ในข้อ 2.2 และ 2.4 ข้อใดเหมาะสมกว่ากัน เพราะเหตุใด และใช้รูปแบบที่ท่านเลือกสร้างสมการสำหรับสภาพบ้านแต่ละประเภท ท่านคิดว่าสมการทั้ง 3 สมการมีค่าใดที่เหมือนกันและแตกต่างกัน อธิบายให้ชัดเจน
3. ในการศึกษาระยะทางที่รถเคลื่อนที่ได้ (Y) หน่วยเป็นไมล์ต่อแกลลอน กับน้ำหนักรถ (X_1) หน่วยเป็น 100 ปอนด์ ได้ข้อมูลดังนี้

Y	18.90	17.00	20.00	18.25	20.07	11.20	22.12	21.47	34.70	30.40
X ₁	3.91	3.86	3.51	3.89	3.36	4.21	3.02	3.18	1.90	2.32
Y	16.50	36.50	21.50	19.70	20.30	17.80	14.39	14.89	17.80	16.41
X ₁	3.88	2.00	2.65	3.37	2.70	3.89	5.29	5.18	3.91	3.66
Y	23.54	21.47								
X ₁	3.05	4.25								

- 3.1 จงสร้างสมการถดถอยแบบแบ่งเป็นช่วงที่มีค่าไม่ต่อเนื่องที่จุด $X_1 = 3.5$ โดยสร้างตัวแปรดัมมี่ X_2, X_3 นั่นคือ ให้ $X_2 = 0$ เมื่อ $X_1 \leq 3.5$ และ $X_2 = 1$ เมื่อ $X_1 > 3.5$ และสร้างตัวแปร X_3 ในลักษณะเดียวกัน
- 3.2 ที่ระดับนัยสำคัญ 0.01 จงทดสอบว่ารูปแบบความสัมพันธ์ในข้อ 3.1 ที่มีค่าไม่ต่อเนื่องที่จุด $X_1 = 3.5$ เหมาะสมหรือไม่
- 3.3 จงสร้างสมการถดถอยแบบแบ่งเป็นช่วงที่มีค่าต่อเนื่อง โดยความชันของสมการเปลี่ยนค่าที่จุด $X_1 = 3.5$ พร้อมทั้งเขียนสมการถดถอยเชิงเส้นเมื่อ $X_1 \leq 3.5$ และ $X_1 > 3.5$
- 3.4 ที่ระดับนัยสำคัญ 0.01 จงทดสอบว่ารูปแบบความสัมพันธ์ในข้อ 3.3 ที่ความชันเปลี่ยนค่าที่จุด $X_1 = 3.5$ เหมาะสมหรือไม่ สรุปผลการทดสอบที่ได้