บทที่ 4
การถดถอยพหุและความสัมพันธ์

การทำนายตัวแปรตาม Y โดยใช้สมการถดถอยเชิงเส้นอย่างง่ายที่มีเพียง 2 ตัวแปรนั้นเราได้กล่าวถึงแล้วในบทก่อน นอกจากนี้เรายังสามารถใช้สมการถดถอยเชิงเส้นที่มีตัวแปรหลาย ๆ ตัวแปรได้ ตัวอย่างเช่น การตอบสนองของสัตว์ทดลองต่อยาบางชนิดอาจขึ้นอยู่กับปริมาณยา อายุและน้ำหนักของสัตว์ทดลอง ซึ่งในบทนี้จะกล่าวถึงตัวแบบการถดถอยพหุ วิธีการหาสมการถดถอย การประเมินสมการถดถอย และการใช้สมการถดถอย และตัวแบบความสัมพันธ์

1. ตัวแบบการถดถอยพหุเชิงเส้น

ในตัวแบบการถดถอย เราเชื่อว่ามีความสัมพันธ์ระหว่างตัวแปรตาม Y และตัวแปรอิสระ X จำนวน k ตัว คือ X_1, X_2, . . . , X_k ตัวแปรอิสระนี้อาจเรียกว่าตัวแปรอธิบาย (explanatory variables) หรือตัวแปรทำนาย (predictor variables) เพราะว่าตัวแปร X ถูกใช้ในการอธิบายความแปรปรวนในตัวแปร Y และใช้ในการทำนายตัวแปร Y

ข้อตกลงเบื้องต้น

ข้อตกลงเบื้องต้นของการวิเคราะห์การถดถอยพหุ คือ

1. ตัวแปร X_i ใด ๆ เป็นตัวแปรกำหนด (fixed) ไม่ใช่ตัวแปรเชิงสุ่ม
2. สำหรับแต่ละเซตของค่า X_i มีประชากรย่อยของค่า Y ซึ่งมีการแจกแจงแบบปกติ
3. ประชากรย่อยของ Y ทุกประชากรมีความแปรปรวนเท่ากัน
4. ค่า Y ทุกค่า มีความเป็นอิสระ หมายความว่า ค่าของ Y ที่วัดได้จากเซตหนึ่งของค่า X ไม่ขึ้นอยู่กับค่าของ Y ที่วัดได้จากเซตอื่น ๆ ของค่า X

ตัวแบบการถดถอยพหุเชิงเส้น

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_k x_{kj} + e_j$$

เมื่อ y_j คือ ค่าตัวหนึ่งจากประชากรย่อยของค่า Y ประชากรหนึ่งเป็นตัวแปรตาม

β_i , i = 1, 2, . . . , k คือสัมประสิทธิ์การถดถอยบางส่วน (partial - regression coeficient)

β_i แทนค่าการเปลี่ยนแปลงของ y สำหรับ x_i เปลี่ยนไป 1 หน่วยเมื่อตัวแปร x ตัวอื่น ๆ ทั้งหมดคงที่

$x_{1j}, x_{2j}, \ldots, x_{kj}$ คือ ค่าเฉพาะของตัวแปรอิสระ $X_1, X_2, \ldots, X_k$

e_j คือ ความคลาดเคลื่อนเป็นตัวแปรสุ่มที่มีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนเท่ากับ σ^2

ตัวแบบการถดถอยพหุเชิงเส้น ที่ประกอบด้วยตัวแปรตาม 1 ตัว และตัวแปรอิสระ 2 ตัวคือ

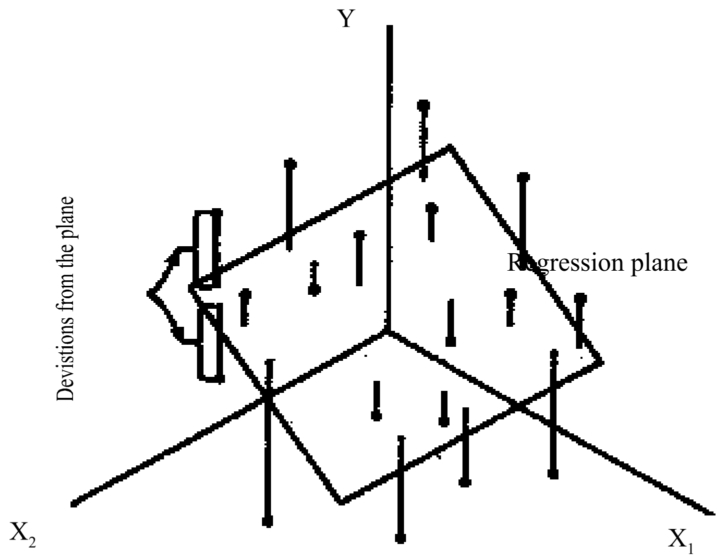
$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + e_j$$

เมื่อ β_0 คือจุดที่ระนาบตัดแกน Y

β_1 คือค่าการเปลี่ยนแปลงของ Y สำหรับ X_1 เปลี่ยนไป 1 หน่วย เมื่อ X_2 คงที่

β_2 คือค่าการเปลี่ยนแปลงของ Y สำหรับ X_2 ที่เปลี่ยนไป 1 หน่วย เมื่อ X_1 คงที่

ตัวแบบนี้แสดงได้ด้วยระนาบในภาพ 3 มิติต่อไปนี้



ภาพ 1 ระนาบการถดถอยพหุและการกระจายของจุด

ส่วนเบี่ยงเบนของจุด ๆ หนึ่งจากระนาบของการถดถอยพหุเชิงเส้นที่มีตัวแปรอิสระ 2 ตัว คือ

$$e_j = y_j - \beta_0 - \beta_1 x_{1j} - \beta_2 x_{2j}$$

2. **การหาสมการถดถอยพหุ**

การประมาณค่าพารามิเตอร์ β_0 , β_1 , ... , β_k ของตัวแบบให้ได้ตัวประมาณค่าที่ไม่เอนเอียงโดยใช้วิธีกำลังสองน้อยที่สุด คือให้ค่าผลบวกกำลังสองของส่วนเบี่ยงเบนของค่าสังเกต Y จากค่า Y ที่คำนวณได้จากสมการถดถอยมีค่าน้อยที่สุด คือ

$$\sum e_j^2 = \sum (y_j - \beta_0 - \beta_1 x_j - \ldots - \beta_k x_{kj})^2$$

หรือ

$$\sum e_j^2 = \sum \left(y_j - \hat{y} \right)^2$$

สมการปกติ (The Normal Equations)

$$nb_0 + b_1 \sum x_{1j} + b_2 \sum x_{2j} + \ldots + b_k \sum x_{kj} = \sum y_j$$

$$b_0 \sum x_{1j} + b_1 \sum x_{1j}^2 + b_2 \sum x_{1j} x_{2j} + \ldots + b_k \sum x_{1j} x_{kj} = \sum x_{1j} y_j$$

$$b_0 \sum x_{2j} + b_1 \sum x_{2j} x_{1j} + b_2 \sum x_{2j}^2 + \ldots + b_k \sum x_{2j} x_{kj} = \sum x_{2j} y_j$$

.

.
.
$$b_0\sum x_{kj} + b_1\sum x_{kj}x_{1j} + b_2\sum x_{kj}x_{2j} + \ldots + b_k\sum x_{kj}^2 = \sum x_{kj}y_j$$

ให้ $b_0, b_1, b_2, \ldots, b_k$ เป็นค่าประมาณของพารามิเตอร์ $\beta_0, \beta_1, \beta_2, \ldots \beta_k$ สามารถคำนวณหาได้โดยแก้สมการปกติข้างต้นซึ่งยุ่งยากมาก แต่สามารถลดรูปให้ง่ายขึ้นโดยการทรานส์ฟอร์ม (transforming) แต่ละค่าให้เป็นส่วนเบี่ยงเบนจากค่าเฉลี่ยของตัวมันเอง ตัวอย่างเช่น สมการถดถอยพหุที่มีตัวแปรอิสระ 2 ตัว เราต้องการประมาณค่าพารามิเตอร์ $\beta_0, \beta_1, \beta_2$ โดยใช้วิธีกำลังสองน้อยที่สุด เซทของสมการปกติคือ

$$nb_0 + b_1\sum x_{1j} + b_2\sum x_{2j} = \sum y_j$$

$$b_0\sum x_{1j} + b_1\sum x_{1j}^2 + b_2\sum x_{1j}x_{2j} = \sum x_{1j}y_j$$

$$b_0\sum x_{2j} + b_1\sum x_{1j}x_{2j} + b_2\sum x_{2j}^2 = \sum x_{2j}y_j$$

การทรานส์ฟอร์ม

$$y'_j = y_j - \overline{y}$$

$$x'_{1j} = x_{1j} - \overline{x}_1$$

$$x'_{2j} = x_{2j} - \overline{x}_2$$

จะได้สมการถดถอยที่มีการทรานส์ฟอร์มแล้ว คือ

$$\hat{y}'_j = b'_0 + b'_1x'_{1j} + b'_2x'_{2j}$$

เมื่อ $b'_0, b'_1,$ และ b'_2 คือ สัมประสิทธิ์ของตัวแปรที่ทรานส์ฟอร์มแล้ว ซึ่งสามารถเขียนสมการถดถอยได้อีกอย่างหนึ่งคือ

$$\hat{y}_j - \overline{y} = b'_0 + b'_1\left(x_{1j} - \overline{x}_1\right) + b'_2\left(x_{2j} - \overline{x}_2\right)$$

$$\hat{y}_j = \overline{y} + b'_0 + b'_1x_{1j} - b'_1\overline{x}_1 + b'_2x_{2j} - b'_2\overline{x}_2$$

$$\hat{y}_j = b'_0 + \overline{y} - b'_1\overline{x}_1 - b'_2\overline{x}_2 + b'_1x_{1j} + b'_2x_{2j}$$

เมื่อเทียบให้เท่ากับสมการถดถอยอย่างง่าย

$$\hat{y}_j = b_0 + b_1x_{1j} + b_2x_{2j}$$

จะได้ว่า

$$b_1 = b'_1$$

$$b_2 = b'_2$$

$$b_0 = b'_0 + \overline{y} - b'_1\overline{x}_1 - b'_2\overline{x}_2$$

$$= b'_0 + \overline{y} - b_1\overline{x}_1 - b_2\overline{x}_2$$

เซทของสมการปกติสำหรับสมการถดถอยที่มีการทรานส์ฟอร์มแล้ว คือ

$$nb'_0 + b'_1\sum x'_{1j} + b'_2\sum x'_{2j} = \sum y'_j$$

$$b'_0\sum x'_{1j} + b'_1\sum x'^2_{1j} + b'_2x'_{1j}x'_{2j} = \sum x'_{1j}y'_j$$

$$b_0'\sum x_{2j}' + b_1'\sum x_{1j}'x_{2j}' + b_2'\sum x_{2j}'^2 = \sum x_{2j}'y_j'$$

จากคุณสมบัติ $\sum\left(x_i - \overline{x}\right) = 0$ จะได้ว่า

$$nb_0' + b_1'(0) + b_2'(0) = 0$$

$$b_0'(0) + b_1'\sum\left(x_{1j} - \overline{x}_1\right)^2 + b_2'\sum\left(x_{1j} - \overline{x}_1\right)\left(x_{2j} - \overline{x}_2\right) = \sum\left(x_{1j} - \overline{x}_1\right)\left(y_j - \overline{y}\right)$$

$$b_0'(0) + b_1'\sum\left(x_{1j} - \overline{x}_1\right)\left(x_{2j} - \overline{x}_2\right) + b_2'\sum\left(x_{2j} - \overline{x}_2\right)^2 = \sum\left(x_{2j} - \overline{x}_2\right)\left(y_j - \overline{y}\right)$$

โดยที่ $b_0' = 0$ ดังนั้น

$$b_0 = \overline{y} - b_1\overline{x}_1 - b_2\overline{x}_2$$

แทนค่า b_1' , b_2' ด้วย b_1, b_2 ตามลำดับในสมการปกติสำหรับสมการถดถอยที่มีการทรานส์ฟอร์มแล้ว

$$b_1\sum x_{1j}'^2 + b_2\sum x_{1j}'x_{2j}' = \sum x_{1j}'y_j'$$

$$b_1\sum x_{1j}'x_{2j}' + b_2\sum x_{2j}'^2 = \sum x_{2j}'y_j'$$

แล้วแก้สมการปกติหาค่า b_1 และ b_2

หรือใช้สูตร

$$b_1 = \frac{\sum x_{2j}'^2 \bullet \sum x_{1j}'y_j' - \sum x_{1j}'x_{2j}' \bullet \sum x_{2j}'y_j'}{\sum x_{1j}'^2 \bullet \sum x_{2j}'^2 - \left(\sum x_{1j}'x_{2j}'\right)^2}$$

$$b_2 = \frac{\sum x_{1j}'^2 \bullet \sum x_{2j}'y_j' - \sum x_{1j}'x_{2j}' \bullet \sum x_{1j}'y_j'}{\sum x_{1j}'^2 \bullet \sum x_{2j}'^2 - \left(\sum x_{1j}'x_{2j}'\right)^2}$$

ตัวอย่างที่ 1 Kalow and Tang ศึกษาความแปรปรวนของไซโตโครม P − 450IA2 แอคติวิตี โดยวัดค่าเฉลี่ยของคาเฟอีนจากประชากรที่เป็นอาสาสมัครที่มีสุขภาพดี วัตถุประสงค์เพื่อศึกษาเปรียบเทียบระหว่าง ผู้สูบบุหรี่กับผู้ไม่สูบบุหรี่ หน่วยทดลองคืออาสาสมัคร 19 คน เก็บข้อมูล 3 อย่างคือ (1) P − 450IA2 index (IA2Index) (2) จำนวนบุหรี่ที่สูบต่อวัน (บุหรี่ / วัน) และ (3) urinary cotinine level (cot) ได้ข้อมูลดังตาราง ต้องการหาสมการถดถอยพหุของตัวอย่าง

ตารางที่ 4.1 จำนวนบุหรี่ที่สูบต่อวัน ระดับ Urine Cotinine และ P − 450IA2 index ของอาสาสมัคร 19 คน

x_1 Cig / Day	x_2 Cot	y IA2 Index
1	.0000	4.1648
1	.0000	3.7314
1	.0000	5.7481
1	.0000	4.4370
1	.0000	6.4687
3	.0000	3.8923
8	10.5950	5.2952
8	4.6154	4.6031
8	27.1902	5.8112
8	5.5319	3.6890
8	2.7778	3.3722
10	19.7856	8.0213
10	22.8045	10.8367

15	.0000	4.1148
15	14.5193	5.5429
15	36.7113	11.3531
20	21.2267	7.5637
20	21.1273	7.2158
24	63.2125	13.5000

แหล่งที่มา : Werner Kalow.

วิธีทำ

1. คำนวณผลบวกกำลังสองและผลคูณของ y'_j , x'_{1j} และ x'_{2j} ดังต่อไปนี้

	x_{1j} Cig/Day	x_{2j} Cot	y_j IA2Index	x_{1j}^2	x_{2j}^2	y_j^2	$x_{1j}x_{2j}$	$x_{1j}y_j$	$x_{2j}y_j$
	1	.0000	4.1648	1	.00	17.346	.00	4.165	.000
	1	.0000	3.7314	1	.00	13.923	.00	3.731	.000
	1	.0000	5.7484	1	.00	33.044	.00	5.748	.000
	1	.0000	4.4370	1	.00	19.687	.00	4.437	.000
	1	.0000	6.4687	1	.00	41.844	.00	6.469	.000
	3	.0000	3.8923	9	.00	15.150	.00	11.677	.000
	8	10.5950	5.2952	64	112.26	28.039	84.76	42.361	56.103
	8	4.6154	4.6031	64	21.30	21.189	36.92	36.825	21.245
	8	27.1902	5.8112	64	739.31	33.770	217.52	46.489	158.007
	8	5.5319	3.6890	64	30.60	13.609	44.26	29.512	20.408
	8	2.7778	3.3722	64	7.72	11.372	22.22	26.978	9.367
	10	19.7856	8.0213	100	391.47	64.341	197.86	80.213	158.705
	10	22.8045	10.8367	100	520.05	117.435	228.05	108.367	247.126
	15	.0000	4.1148	225	.00	16.931	.00	61.721	.000
	15	14.5193	5.5429	225	210.81	30.724	217.79	83.144	80.479
	15	36.7113	11.3531	225	1347.72	128.892	550.67	170.296	416.785
	20	21.2267	7.5637	400	450.57	57.210	424.53	151.275	160.554
	20	21.1273	7.2157	400	446.36	52.067	422.55	144.315	152.449
	24	63.2125	13.5000	576	3995.82	182.250	1517.10	324.000	853.369
ผลรวม	177	250.098	119.362	2585	8273.98	898.822	3964.22	1341.72	2334.60
ค่าเฉลี่ย	9.3158	13.1630	6.2822						

2. การทรานส์ฟอร์ม

$$y'_j = \left(y_j - \bar{y}\right) \quad , \quad x'_{1j} = \left(x_{1j} - \bar{x}_1\right) \quad , \quad x'_{2j} = \left(x_{2j} - \bar{x}_2\right)$$

3. ใช้ข้อมูลในตารางคำนวณค่าต่าง ๆ ต่อไปนี้

$$\sum x'^2_{1j} = \sum\left(x_{1j} - \bar{x}_1\right)^2$$

$$= \sum x'^2_{1j} - \frac{\left(\sum x_{1j}\right)^2}{n}$$

$$= 2585 - \frac{(177)^2}{19}$$

$$= 936.105263$$

$$\sum x'^2_{2j} = \sum\left(x_{2j} - \bar{x}_2\right)^2$$

$$= \sum x^2_{2j} - \frac{\left(\sum x_{2j}\right)^2}{n}$$

$$= \quad 8273.98 - \frac{(250.098)^2}{19}$$

$$= \quad 4981.92686$$

$$\sum x'_{1j}x'_{2j} = \sum(x_{1j}-\bar{x}_1)(x_{2j}-\bar{x}_2)$$

$$= \quad \sum x_{1j}x_{2j} - \frac{\sum x_{1j}\sum x_{2j}}{n}$$

$$= \quad 3964.22 - \frac{(177.000)(250.098)}{19}$$

$$= \quad 1634.35968$$

$$\sum x'_{1j}y'_j = \sum(x_{1j}-\bar{x}_1)(y_j-\bar{y})$$

$$= \quad \sum x_{1j}y_j - \frac{\sum x_{1j}\sum y_j}{n}$$

$$= \quad 1341.72 - \frac{(177.000)(119.362)}{19}$$

$$= \quad 229.76874$$

$$\sum x'_{2j}y'_j = \sum(x_{2j}-\bar{x}2)(y_j-\bar{y})$$

$$= \quad \sum x_{2j}y_j - \frac{\sum x_{2j}\sum y_j}{n}$$

$$= \quad 2334.60 - \frac{(250.098)(119.362)}{19}$$

$$= \quad 763.43171$$

4. ทำให้ได้สมการปกติ

$$b_1\sum x'^2_{1j} + b_2\sum x'_{1j}x'_{2j} = \sum x'_{1j}y'_j$$

$$b_1\sum x'_{1j}x'_{2j} + b_2\sum x'^2_{2j} = \sum x'_{2j}y'_j$$

5. แทนค่าในสมการปกติ

$$936.105263\ b_1 + 1634.35968\ b_2 = 229.76874$$

$$1634.35968\ b_1 + 4981.92686\ b_2 = 763.43171$$

6. แก้สมการปกติเพื่อหาค่า b_1 , b_2 และ b_0 หรือใช้สูตรดังนี้

$$b_1 = \frac{\sum x'^{2}_{2j} \bullet \sum x'_{1j}y'_j - \sum x'_{1j}x'_{2j} \bullet \sum x'_{2j}y'_j}{\sum x'^{2}_{1j} \bullet \sum x'^{2}_{2j} - \left(\sum x'_{1j}x'_{2j}\right)^2}$$

$$= \frac{4981.92686(229.76874) - 1634.35968(763.4317)}{936.105263(4981.92686) - (1634.35968)^2}$$

$$= -0.05171$$

$$b_2 = \frac{\sum x'^{2}_{1j} \bullet \sum x'_{2j}y'_j - \sum x'_{1j}x'_{2j} \bullet \sum x'_{1j}y'_j}{\sum x'^{2}_{1j} \bullet \sum x'^{2}_{2j} - \left(\sum x'_{1j}x'_{2j}\right)^2}$$

$$= \frac{936.105263(763.43171) - 1634.35968(229.76874)}{936.105263(4981.92686) - (1634.35968)^2}$$

$$= 0.170204$$

และจากสูตร $b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2$

จะได้ $b_0 = 6.2822 - (-0.05171)(9.3158) - (0.170204)(13.1630)$

$= 4.5235$

7. ได้สมการถดถอยพหุของตัวอย่างคือ

$$= 4.5235 - 0.05171x_{1j} + 0.170204x_{2j}$$

8. การแปลความหมาย

$b_1 = -0.05171$ หมายความว่า IA2Index จะลดลง 0.05171 หน่วย ต่อจำนวนบุหรี่ที่สูบเพิ่มขึ้น 1 มวน สำหรับผู้ที่มีระดับ cot ระดับเดียวกัน

$b_2 = 0.170204$ หมายความว่า IA2Index จะเพิ่มขึ้น 0.170204 หน่วย ต่อระดับ cot ที่เพิ่มขึ้น 1 หน่วย สำหรับผู้ที่สูบบุหรี่ต่อวันจำนวนมวนเท่ากัน

3. **สัมประสิทธิ์การถดถอยมาตรฐาน** (Standardized regression coefficiont)

เราสนใจจัดลำดับความสำคัญของตัวแปรอิสระที่สามารถทำนายตัวแปรตาม y แต่ทำได้ยากเพราะตัวแปรอิสระแต่ละตัวมีหน่วยแตกต่างกัน โดยเฉพาะในตัวแบบการถดถอยพหุ สัมประสิทธิ์การถดถอย b ใช้ประมาณการเปลี่ยนแปลงของ y เมื่อ x เปลี่ยนไป 1 หน่วย ขณะที่ตัวแปรตัวอื่น ๆ ทั้งหมดในตัวแบบคงที่ ถ้า x เพิ่มขึ้น 1 ส่วนเบี่ยงเบนมาตรฐานหน่วยคือ $x + s_x$ แล้ว y จะเพิ่มขึ้น $b \times s_x$ หน่วยหรือ ส่วนเบี่ยงเบนมาตรฐานหน่วยของ y (s_y)

สัมประสิทธิ์การถดถอยมาตรฐาน (b_s) คิดได้จาก $b \times$ ใช้ทำนายการเพิ่มขึ้นของ y เมื่อ x เพิ่มขึ้น 1 ส่วนเบี่ยงเบนมาตรฐาน

ดังนั้นสัมประสิทธิ์การถดถอยมาตรฐานสามารถใช้เปรียบเทียบระหว่างตัวแปรอิสระหลายตัวได้เพราะมันแสดงการเพิ่มขึ้นเป็นส่วนเบี่ยงเบนมาตรฐานของ y เมื่อ x เพิ่มขึ้น 1 ส่วนเบี่ยงเบนมาตรฐาน จึงทำให้เห็นความแตกต่างของตัวแปรอิสระแต่ละตัวที่วัดด้วยมาตรฐานเดียวกัน

ตัวอย่างที่ 2 การคำนวณสัมประสิทธิ์การถดถอยมาตรฐานสำหรับจำนวนบุหรี่ที่สูบต่อวัน (x_1) และระดับ cot (x_2) โดยใช้ข้อมูลจากตัวอย่างที่ 1

วิธีทำ

1. หาค่าประมาณของความแปรปรวนของ y, x_1 และ x_2

$$s_y^2 = \frac{\sum y_j^2 - \frac{\left(\sum y_j\right)^2}{n}}{n-1}$$

$$= \frac{989.822 - \frac{(119.362)^2}{19}}{18}$$

$$= 8.28$$

$$s_{x_1}^2 = \frac{\sum x_1^2 - \frac{\left(\sum x_1\right)^2}{n}}{n-1}$$

$$= \frac{2585 - \frac{(177)^2}{19}}{18}$$

$$= 52.01$$

$$s_{x_2}^2 = \frac{\sum x_2^2 - \frac{\left(\sum x_2\right)^2}{n}}{n-1}$$

$$= \frac{8273.98 - \frac{(250.098)^2}{19}}{18}$$

$$= 276.77$$

ดังนั้นจึงได้ส่วนเบี่ยงเบนมาตรฐานของ y, x_1 และ x_2 ดังนี้

s_y = 2.87 s_{x_1} = 7.21 s_{x_2} = 16.64

2. คำนวณหาสัมประสิทธิ์การถดถอยมาตรฐาน

$$b_s(x_1) = \frac{-0.05171 \times 7.21}{2.87}$$

$$= -0.1299$$

$$b_s(x_2) = \frac{0.170204 \times 16.64}{2.87}$$

$$= 0.9868$$

3. การแปลความหมาย

IA2Index คาดว่าจะลดลง 0.1299 หน่วยส่วนเบี่ยงเบนมาตรฐานของ IA2Index ต่อการเพิ่มขึ้น 1 ส่วนเบี่ยงเบนมาตรฐานของจำนวนบุหรี่ที่สูบต่อวัน เมื่อระดับ cot คงที่ และ IA2Index คาดว่าจะเพิ่มขึ้น 0.9868 ส่วนเบี่ยงเบนมาตรฐานของ IA2Index ต่อการเพิ่มขึ้น 1 ส่วนเบี่ยงเบนมาตรฐานของระดับ cot เมื่อจำนวนบุหรี่ที่สูบต่อวันคงที่ แสดงว่าระดับ cot เป็นตัวแปรที่มีความสำคัญมากกว่า

4. การประเมินสมการถดถอยพหุของตัวอย่าง

ก่อนที่จะใช้สมการถดถอยพหุในการทำนายและประมาณค่า Y นั้นต้องประเมินสมการถดถอยพหุ เพื่อพิจารณาว่าสมการอธิบายความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระได้ดีเพียงใด สำหรับการประเมินสมการถดถอยอย่างง่ายที่กล่าวมาแล้ว เราจะพิจารณาจาก R^2 (coefficient of determination) และค่าประมาณของความชัน แต่สำหรับการประเมินสมการถดถอยพหุ เราจะเน้นที่ R^2 (coefficient of multiple determination) และ สัมประสิทธิ์การถดถอยบางส่วน (partial regression coefficients)

การคำนวณ R^2 อาศัยหลักการเกี่ยวกับความแปรปรวนทั้งหมดของค่า Y สามารถแบ่งออกได้เป็น 2 ส่วน คือ ส่วนที่อธิบายได้ ซึ่งวัดได้จากความแปรปรวนที่อธิบายได้โดยสมการถดถอย และส่วนที่อธิบายไม่ได้ ซึ่งอธิบายไม่ได้โดยสมการถดถอย ความแปรปรวนทั้งหมดคือ ผลบวกกำลังสองของส่วนเบี่ยงเบนของค่าสังเกต Y แต่ละค่าจากค่าเฉลี่ยของค่าสังเกต เท่ากับ หรือ SST ความแปรปรวนส่วนที่อธิบายได้ คือ ผลบวกกำลังสองของส่วนเบี่ยงเบนของค่า y ที่คำนวณได้จากสมการถดถอย เบี่ยงเบนจากค่าเฉลี่ยของค่าสังเกต เท่ากับ $\sum(\hat{y}-\bar{y})^2$ เรียกว่าผลบวกกำลังสองของการถดถอย หรือ SSR ความ แปรปรวนส่วนที่อธิบายไม่ได้ คือ ผลบวกกำลังสองของส่วนเบี่ยงเบนของค่าสังเกตจากค่าที่คำนวณได้ เท่ากับ $\sum(y_j-\hat{y})^2$ หรือ SSE เราสามารถสรุปความสัมพันธ์ของผลบวกกำลังสองทั้ง 3 เทอมได้ดังนี้

$$\sum(y_j-\bar{y})^2 = \sum(\hat{y}-\bar{y})^2 + \sum(y_j-\hat{y})^2$$

เขียนได้อีกแบบหนึ่งคือ SST = SSR + SSE

สูตรการคำนวณของผลบวกกำลังสองคือ

$$\text{SST} = \Sigma(y_j - \bar{y})^2 = \sum y_j^2 - \frac{\left(\sum y_j\right)^2}{n}$$

$$\text{SSR} = \Sigma(\hat{y} - \bar{y})^2$$

$$= b_1 \sum x'_{1j} y'_j + b_2 \sum x'_{2j} y'_j + \ldots + b_k \sum x'_{kj} y'_j$$

$$\text{SSE} = \text{SST} - \text{SSR}$$

การคำนวณ $R^2_{y \bullet 12 \ldots k}$

$$R^2_{y \bullet 12 \ldots k} = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y_j - \bar{y})^2}$$

เมื่อ ตัวห้อย y.12…k หมายความว่า ในการวิเคราะห์การถดถอยให้ Y เป็นตัวแปรตาม และตัวแปร X จาก X_1 ถึง X_k เป็นตัวแปรอิสระ

$R^2_{y \bullet 12 \ldots k}$ แสดงสัดส่วนของความแปรปรวนทั้งหมดในค่าสังเกต Y ที่อธิบายได้ด้วยการถดถอยของ Y บน X_1, X_2, . . . , X_k

ตัวอย่างที่ 3 การคำนวณหา จากข้อมูลตัวอย่างที่ 1

วิธีทำ

$$\text{SST} = 898.822 - \frac{(119.362)^2}{19}$$

$$= 148.9648$$

$$\text{SSR} = (-0.05171)(229.76874) + (0.170204)(763.43171)$$

$$= 118.0578$$

$$\text{SSE} = 148.9648 - 118.0578$$

$$= 30.907$$

$$R^2_{y \bullet 12} = \frac{118.0578}{148.9648}$$

$$= 0.7925$$

การแปลความหมายคือ 79.25 เปอร์เซนต์ของความแปรปรวนทั้งหมดในค่า Y อธิบายได้ด้วยระนาบของการถดถอย หรือโดยความสัมพันธ์เชิงเส้นกับ X_1 และ X_2

5. **การทดสอบสมมติฐานของการถดถอย**

เพื่อทดสอบการถดถอย ว่ามีนัยสำคัญหรือไม่ ดำเนินตามขั้นตอนดังนี้

1. ข้อมูล เหมาะสมที่จะใช้การวิเคราะห์การถดถอย
2. ข้อตกลงเบื้องต้นของการวิเคราะห์การถดถอยเป็นจริง
3. สมมติฐานทางสถิติที่ต้องการทดสอบ คือ

   $H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$ VS $H_1: \beta_i \neq 0$ อย่างน้อย 1 ตัว

4. สถิติทดสอบ คือ อัตราส่วนของความแปรปรวน F_0 เท่ากับ $\frac{\text{MSR}}{\text{MSE}}$ คำนวณเหมือนการวิเคราะห์ความแปรปรวนในตาราง

ตาราง 4.3 การวิเคราะห์ความแปรปรวนสำหรับการถดถอยพหุ

Source of Variation	Sum of Square	Degree of Freedom	Mean Square	F_0
Regression	SSR	k	MSR $= \frac{SSR}{k}$	$\frac{\text{MSR}}{\text{MSE}}$
Error	SSE	n − k − 1	MSE $= \frac{SSE}{n-k-1}$	
Total	SST	n − 1		

5. การแจกแจงของสถิติทดสอบ เมื่อ H_0 เป็นจริง และเป็นไปตามข้อตกลงเบื้องต้นของการวิเคราะห์การถดถอย สถิติทดสอบจะมีการแจกแจงแบบ F ที่มีจำนวนชั้นอิสระ k และ n − k − 1
6. กฎการตัดสินใจ คือ ถ้าค่าสถิติ F_0 ที่คำนวณได้ มากกว่าหรือเท่ากับค่าวิกฤติ $F_{\alpha,k,n-k-1}$ จะตัดสินใจปฏิเสธ H_0
7. คำนวณสถิติทดสอบ สรุปในตารางการวิเคราะห์ความแปรปรวน
8. ตัดสินใจปฏิเสธหรือยอมรับ H_0
9. สรุปผล ถ้าปฏิเสธ H_0 สรุปว่าในประชากรที่สุ่มตัวอย่างมา ตัวแปรตามมีความสัมพันธ์เชิงเส้นกับตัวแปรอิสระทุกตัว ถ้ายอมรับ H_0 สรุปว่าไม่มีความสัมพันธ์เชิงเส้นระหว่างตัวแปรตามและตัวแปรอิสระทุกตัว

ตัวอย่าง ที่ 4 อาศัยข้อมูลจากตัวอย่างที่ 1 ต้องการทดสอบสมมติฐานทางสถิติเกี่ยวกับไม่มีความสัมพันธ์เชิงเส้นระหว่างตัวแปรทั้ง 3 ตัว คือ P − 450IA2 index จำนวนบุหรี่ที่สูบต่อวัน และระดับ urinary cotinine

วิธีทำ

1. ข้อมูลอาศัยข้อมูลจากตัวอย่างที่ 1
2. ข้อตกลงเบื้องต้นของการวิเคราะห์การถดถอยเป็นจริง
3. สมมติฐานทางสถิติที่ต้องการทดสอบคือ

   $H_0: \beta_1 = \beta_2 = 0$ VS $H_1: \beta_i \neq 0$ อย่างน้อย 1 ค่า

ตาราง 4.4 วิเคราะห์ความแปรปรวนสำหรับการถดถอยพหุของตัวอย่างที่ 1

Source of Variation	Sum of Square	Degree of Freedom	Mean Saquare	F_0
Regression	118.0578	2	59.0289	30.55
Error	30.907	16	1.9317	
Total	148.9648	18		

4. สถิติทดสอบที่ใช้ทดสอบสมมติฐานทางสถิติ คือ อัตราส่วนของความแปรปรวน (F_0)

5. การแจกแจงของสถิติทดสอบ ถ้า H_0 เป็นจริง และเป็นไปตามข้อตกลงเบื้องต้นของการวิเคราะห์การถดถอย สถิติทดสอบจะมีการแจกแจงแบบ F ที่มีจำนวนชั้นอิสระ 2 และ 16

6. กฎการตัดสินใจ ที่ระดับนัยสำคัญ α = .01 ตัดสินใจปฏิเสธ H_0 ถ้า F_0 ที่คำนวณได้มากกว่าหรือเท่ากับค่าวิกฤติที่เปิดจากตาราง $F_{.01,2,16}$ เท่ากับ 6.23

7. คำนวณสถิติทดสอบ จากตารางวิเคราะห์ความแปรปรวน สถิติทดสอบที่คำนวณได้เท่ากับ 30.55

8. การตัดสินใจทางสถิติ เนื่องจาก F_0 ที่คำนวณได้มากกว่าค่าวิกฤติ F ที่เปิดจากตารางจึงตัดสินใจปฏิเสธ H_0

9. สรุปผลการทดสอบในประชากรที่ตัวอย่างถูกสุ่มมา มีความสัมพันธ์เชิงเส้นระหว่างตัวแปรทั้ง 3 ตัว

6. การประมาณค่าความคลาดเคลื่อนของการถดถอย (σ)

ค่าประมาณของ σ คิดได้จาก

$$\hat{\sigma} = \sqrt{\frac{\sum_{j=1}^{n}\hat{e}_j^2}{(n-k-1)}} = \sqrt{\frac{\sum_{j=1}^{n}\left(y_j - \hat{\beta}_0 - \hat{\beta}_1 x_{1j} - \ldots - \hat{\beta}_k x_{kj}\right)^2}{(n-k-1)}}$$

หรือเขียนเป็นสัญลักษณ์

$$\text{SSE}(X_1, \ldots, X_k) = \sum_{j=1}^{n}\left(y_j - \hat{\beta}_0 - \hat{\beta}_1 x_{1j} - \ldots - \hat{\beta}_k x_{kj}\right)^2$$

$$= \sum_{j=1}^{n}\hat{e}_j^2$$

เมื่อ SSE คือ sum of squared errors ของการประมาณการถดถอยของ Y บน $X_1, \ldots, X_k$

และ $$\text{MSE}(X_1, \ldots, X_k) = \frac{\text{SSE}(X_1, \ldots, X_k)}{(n-k-1)}$$

เมื่อ MSE คือ mean squared error ของการประมาณการถดถอยของ Y บน $X_1, \ldots, X_k$

ดังนั้นค่าประมาณของ σ คือ

$$\hat{\sigma}^2 = \text{MSE} = \frac{\text{SSE}}{n-k-1}$$

$$\hat{\sigma} = \sqrt{\text{MSE}}$$

และค่าประมาณของความแปรปรวนของ Y คือ

$$\hat{\sigma}_y^2 = \frac{\sum_{j=1}^{n}\left(y_j - \hat{\mu}_Y\right)^2}{n-1}$$

7. การใช้เมตริกในการประมาณพารามิเตอร์

การใช้เมตริกในการประมาณพารามิเตอร์ $\beta_0, \beta_1, \ldots, \beta_k$ โดยวิธีกำลังสองน้อยที่สุดเพื่อให้ได้ค่าประมาณ $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ หาได้จากการแก้สมการเมตริกของ $\hat{\beta}$ ได้สมการปกติคือ

$$X^T X\hat{\beta} = X^T y$$

เมื่อ y คือเวคเตอร์ขนาด $n \times 1$

X คือเมตริกร์ขนาด $n \times (k + 1)$

$\hat{\beta}$ คือเวคเตอร์ขนาด $(k + 1) \times 1$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{i1} & x_{i2} & \ldots & x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_i \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

ค่าประมาณ $\hat{\beta}$ หาได้จาก

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

การประมาณฟังก์ชันการถดถอย $\mu_Y(X_1, \ldots, X_k)$ คือ

$$\hat{\mu}_Y(X_1, \ldots, X_k) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$$

การทำนายค่า Y จากข้อมูลที่สุ่มมาจากประชากร โดยที่ $X_1 = x_1, \ldots, X_k = x_k$ คือ

$$\hat{Y}(x_1, \ldots, x_k) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$$

ตัวอย่างที่ 5 การคำนวณหา $\hat{\beta}$ โดยใช้เมตริกซ์ จากข้อมูลตัวอย่าง (Franklin A.Graybill และ Hariharan K.Iyer ในหนังสือ Regression Analysis)

ตาราง 4.5 ข้อมูลตัวแปรตาม Y และตัวแปร X จำนวน 3 ตัว

Y	X_1	X_2	X_3
6	3	9	16
9	6	13	13
12	4	3	17
5	8	2	10
13	3	4	9
2	2	4	7

วิธีทำ

$$y = \begin{bmatrix} 6 \\ 9 \\ 12 \\ 5 \\ 13 \\ 2 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 3 & 9 & 16 \\ 1 & 6 & 13 & 13 \\ 1 & 4 & 3 & 17 \\ 1 & 8 & 2 & 10 \\ 1 & 3 & 4 & 9 \\ 1 & 2 & 4 & 7 \end{bmatrix} \qquad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 3 & 6 & 4 & 8 & 3 & 2 \\ 9 & 13 & 3 & 2 & 4 & 4 \\ 16 & 13 & 17 & 10 & 9 & 7 \end{bmatrix} ; \; X^TX = \begin{bmatrix} 6 & 26 & 35 & 72 \\ 26 & 138 & 153 & 315 \\ 35 & 153 & 295 & 448 \\ 72 & 315 & 448 & 944 \end{bmatrix} ; \; X^Ty = \begin{bmatrix} 47 \\ 203 \\ 277 \\ 598 \end{bmatrix}$$

$$\hat{\beta} = (X^TX)^{-1}X^TY$$

$$= \begin{bmatrix} 2.59578 & -0.15375 & -0.01962 & -0.13737 \\ -0.15375 & 0.03965 & -0.00014 & -0.00144 \\ -0.01962 & -0.00014 & 0.01234 & -0.00431 \\ -0.13737 & -0.00144 & -0.00431 & 0.01406 \end{bmatrix} \begin{bmatrix} 47 \\ 203 \\ 277 \\ 598 \end{bmatrix}$$

$$= \begin{bmatrix} 3.20975 \\ -0.07573 \\ -0.11162 \\ 0.46691 \end{bmatrix}$$

จะได้ค่าประมาณของสัมประสิทธิ์การถดถอยดังนี้

$$\hat{\beta}_0 = 3.20975 \qquad \hat{\beta}_1 = -0.07573 \qquad \hat{\beta}_2 = -0.11162 \qquad \hat{\beta}_3 = 0.46691$$

การประมาณฟังก์ชันการถดถอย คือ

$$\hat{\mu}_Y(x_1, x_2, x_3) = 3.20975 - 0.07573\,x_1 - 0.11162\,x_2 + 0.46691\,x_3$$

การทำนายค่า Y จากข้อมูลที่สุ่มมาจากประชากรได้ที่ $X_1 = x_1$, $X_2 = x_2$, $X_3 = x_3$ คือ

$$\hat{Y}(x_1, x_2, x_3) = \hat{\mu}_Y(x_1, x_2, x_3) = 3.20975 - 0.07573\,x_1 - 0.11162\,x_2 + 0.46691\,x_3$$

7.1 การคำนวณความคลาดเคลื่อนของการถดถอย

ถ้าเขียนความคลาดเคลื่อนในรูปของเวคเตอร์ดังนี้

$$\hat{e} = \begin{bmatrix} \hat{e}_1 \\ \hat{e}_2 \\ \vdots \\ \hat{e}_n \end{bmatrix}$$

เวคเตอร์ของความคลาดเคลื่อน คือ

$$\hat{e} = y - X\hat{\beta}$$

7.1.1 การคำนวณหาผลบวกกำลังสองของความคลาดเคลื่อนหรือ SSE

ผลบวกกำลังสองของความคลาดเคลื่อนคำนวณได้จาก

$$SSE = \hat{e}^T\hat{e}$$

$$SSE = \left(y - X\hat{\beta}\right)^T\left(y - X\hat{\beta}\right)$$

$$SSE = y^T y - \hat{\beta}^T X^T y$$

$$SSE = y^T\left[1 - X\left(X^T X\right)^{-1} X^T\right] y$$

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-k-1}$$

$$\hat{\sigma} = \sqrt{MSE}$$

7.1.2 การหาความคลาดเคลื่อนมาตรฐานจากเมตริก C

เนื่องจาก $C = (X^T X)^{-1}$

เมื่อ c_{ij} คือ สมาชิกที่ (i, j) ของเมตริกซ์ C

ความคลาดเคลื่อนมาตรฐานของพารามิเตอร์แต่ละตัวคำนวณได้จากสูตร

$$se\left(\hat{\beta}_i\right) = \hat{\sigma}\sqrt{c_{ii}} \quad ; \quad i = 1, \ldots, (k+1)$$

ดังนั้นความคลาดเคลื่อนมาตรฐานของพารามิเตอร์แต่ละตัวคือ

$$se(b_0) = \hat{\sigma}\sqrt{c_{00}}$$

$$se(b_1) = \hat{\sigma}\sqrt{c_{11}}$$

$$se(b_2) = \hat{\sigma}\sqrt{c_{22}}$$

ความคลาดเคลื่อนของฟังก์ชันการถดถอย คือ

$$se\left(\hat{\mu}_Y\left(x_1,\ldots,x_k\right)\right) = \hat{\sigma}\sqrt{x^TCx}$$

ความคลาดเคลื่อนของการทำนายค่า Y จากข้อมูลที่สุ่มมาจากประชากรโดยที่ $X_1 = x_1 , \ldots , X_k = x_k$ คือ

$$se\left(\hat{Y}\left(x_1,\ldots,x_k\right)\right) = \hat{\sigma}\sqrt{1+\left(x^TCx\right)}$$

$$= \sqrt{\hat{\sigma}^2+\left[se\left(\hat{\mu}_Y\left(x_1,\ldots,x_k\right)\right)\right]^2}$$

ตัวอย่างที่ 6 การคำนวณหาความคลาดเลื่อนมาตรฐานของพารามิเตอร์แต่ละตัว เมื่อ $\hat{\sigma}^2 = 37.612$

วิธีทำ

$$\hat{\sigma} = \sqrt{37.612}$$

$$= 6.1329$$

$$se(b_0) = 6.1329\sqrt{2.59578}$$

$$= 9.881$$

$$se(b_1) = 6.1329\sqrt{0.03965}$$

$$= 1.221$$

$$se(b_2) = 6.1329\sqrt{0.01234}$$

$$= 0.681$$

$$se(b_3) = 6.1329\sqrt{0.01406}$$

$$= 0.727$$

8. **การทดสอบแบบ F บางส่วน** (Partial F - Test)

การทดสอบแบบ F บางส่วนใช้ในการตัดสินว่าจะเพิ่มหรือตัดตัวแปร X_i ในตัวแบบการ ถดถอย ซึ่งมีจำนวนชั้นอิสระ 1,(n - k - 1) เพื่อทดสอบสมมติฐาน $H_0 : \beta_i = 0$ VS $H_A : \beta_i \neq 0$ ซึ่งเท่ากับสถิติ t^2 จำนวนชั้นอิสระ = (n − k − 1)

สถิติทดสอบ F บางส่วนคือ

$$F(X_1 | X_2X_3) = \frac{SSR\left(X_1 | X_2X_3\right)}{MSE\left(X_1, X_2, X_3\right)}$$

$$F(X_2 | X_3X_1) = \frac{SSR(X_2|X_1X_3)}{MSE(X_1, X_2, X_3)}$$

$$F(X_3 | X_1X_2) = \frac{SSR(X_3|X_1X_2)}{MSE(X_1, X_2, X_3)}$$

สถิติทดสอบ t คือ

$$t = \frac{b_i}{se(b_i)}$$

ตัวอย่างที่ 7 การทดสอบสมมติฐาน $H_O : \beta_i = 0$ ด้วยสถิติทดสอบ t โดยใช้ข้อมูลจากตัวอย่างที่ 1

วิธีทำ

เนื่องจาก $\hat{\sigma}^2 = 1.9317$

ดังนั้น $\hat{\sigma} = \sqrt{1.9317}$

$= 1.3898$

และจากเมตริก c ได้ว่า $c_{11} = 0.0025$

$c_{22} = 0.0004698$

ดังนั้นความคลาดเคลื่อนมาตรฐานของพารามิเตอร์แต่ละตัว คือ

$se(b_1) = 1.3898\sqrt{0.0025}$

$= 0.0695$

$se(b_2) = 1.3898\sqrt{0.0004698}$

$= 0.0301$

1. สมมติฐานทางสถิติที่ต้องการทดสอบคือ

$H_0 : \beta_1 = 0$ VS $H_1 : \beta_1 \neq 0$

ที่ระดับนัยสำคัญ $\alpha = .05$

คำนวณค่าสถิติ

$$t = \frac{b_1}{se(b_1)} = \frac{-0.0517}{.0695} = -0.7439$$

สำหรับการทดสอบแบบสองทางเปิดตารางหาค่าวิกฤติ $= t_{.025,(19-2-1)} = t_{.025,16} = 2.1199$

เขตวิกฤติคือ $|t| \geq t_{.025,16}$ เนื่องจาก | t | ที่คำนวณได้น้อยกว่า 2.1199 จึงตัดสินใจยอมรับ H_0

สรุปว่าตัวแปรจำนวนบุหรี่ต่อวัน กับตัวแปรตาม IA2Index มีความสัมพันธ์เชิงเส้นอย่างไม่มีนัยสำคัญ นั่นคือไม่สามารถใช้ตัวแปรจำนวนบุหรี่ต่อวันในการทำนายตัวแปรตาม IA2Index

2. สมมติฐานทางสถิติที่ต้องการทดสอบคือ

H_0 : $\beta_2 = 0$ VS H_A : $\beta_2 \neq 0$

ที่ระดับนัยสำคัญ α = .05

คำนวณค่าสถิติทดสอบ $t = \dfrac{b_2}{se(b_2)}$

$$= \frac{0.1702}{.0301}$$

$$= 5.6545$$

สำหรับการทดสอบแบบสองทางเปิดตารางหาค่าวิกฤติ $t_{.025,16} = 2.1199$ เขตวิกฤติคือ $|t| \geq t_{.025,16}$ เนื่องจากค่า $|t|$ ที่คำนวณได้มากกว่า 2.1199 จึงตัดสินใจปฏิเสธ H_0

สรุปว่ามีความสัมพันธ์เชิงเส้นระหว่าง urinary cotinine level กับ IA2Index นั่นคือ สามารถใช้ตัวแปร urinary cotinine level ในการทำนายตัวแปรตาม IA2Index

9. การประมาณช่วงความเชื่อมั่นของพารามิเตอร์สำหรับการถดถอยพหุ

100 (1 - α) % ช่วงความเชื่อมั่นของ β_i คำนวณได้จากสูตร

$$b_i \pm t_{\frac{\alpha}{z},n-k-1} se(b_i)$$

ตัวอย่างที่ 8 การประมาณช่วงความเชื่อมั่นของ β_1 และ β_2 ของฟังก์ชันการถดถอยสำหรับตัวอย่างที่ 1

วิธีทำ

$$b_1 \pm t_{\frac{\alpha}{2},n-2-1} se(b_1) = (-0.0517 \pm t_{.025,16}\,(0.0695))$$

$$= (-0.0517 \pm 2.12\,(0.0695))$$

$$= (-0.2001 , 0.0967)$$

$$b_2 \pm t_{\frac{\alpha}{2},n-2-1} se(b_2) = (0.170 \pm 2.12\,(.0310))$$

$$= (0.1064 , 0.2336)$$

10. **สัมประสิทธิสหสัมพันธ์พหุคูณ** (The Multiple Correlation coefficient)

ขั้นแรกของการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร เราดูที่สัมประสิทธิสหสัมพันธ์พหุคูณ สัมประสิทธิสหสัมพันธ์พหุคูณ คือ รากที่สองของ R^2

$$R_{y.12\ldots k} = \sqrt{R^2_{y\bullet 12\bullet\bullet\bullet k}} = \sqrt{\frac{\sum(\hat{y}-\bar{y})^2}{\sum(y_i-\bar{y})^2}}$$

$R_{y.12}$ คือการวัดความสัมพันธ์ระหว่างตัวแปร เราสามารถใช้ $R_{y.12}$ ประมาณ $\rho_{y.12}$ ซึ่งเป็น สัมประสิทธิสหสัมพันธ์พหุคูณของประชากร และยังสามารถใช้แทนสัมประสิทธิสหสัมพันธ์ อย่างง่ายระหว่าง y_j และ $\hat{y}$ (ค่าสังเกตและค่าคำนวณของตัวแปรตาม)

11. **สหสัมพันธ์บางส่วน** (Partial Correlation)

สหสัมพันธ์บางส่วนคือค่าความสัมพันธ์ระหว่าง 2 ตัวแปร เมื่อตัวแปรอื่น ๆ ที่เหลือคงที่ ตัวอย่างเช่น $r_{y1.2}$ คือค่าความสัมพันธ์ระหว่าง Y และ X_1 เมื่อให้ X_2 คงที่

11.1 **การคำนวณหาสัมประสิทธิสหสัมพันธ์บางส่วน** (partial correlation coefficient)

กรณีของการหาสัมประสิทธิสหสัมพันธ์บางส่วนของตัวแปร 3 ตัว สามารถหาสัมประสิทธิสห-สัมพันธ์บางส่วน 3 ตัวคือ

1. สหสัมพันธ์บางส่วนระหว่าง Y และ X_1 เมื่อให้ X_2 คงที่

$$r_{y1.2} = \frac{(r_{y1} - r_{y2}r_{12})}{\sqrt{(1-r_{y2}^2)(1-r_{12}^2)}}$$

2. สหสัมพันธ์บางส่วนระหว่าง Y และ X_2 เมื่อให้ X_1 คงที่

$$r_{y2.1} = \frac{(r_{y2} - r_{y1}r_{12})}{\sqrt{(1-r_{y1}^2)(1-r_{12}^2)}}$$

3. สหสัมพันธ์บางส่วนระหว่าง X_1 และ X_2 เมื่อให้ Y คงที่

$$r_{12.y} = \frac{(r_{12} - r_{y1}r_{y2})}{\sqrt{(1-r_{y1}^2)(1-r_{y2}^2)}}$$

เมื่อ r_{y1} คือ สหสัมพันธ์อย่างง่ายระหว่าง Y และ X_1

r_{y2} คือ สหสัมพันธ์อย่างง่ายระหว่าง Y และ X_2

r_{12} คือ สหสัมพันธ์อย่างง่ายระหว่าง X_1 และ X_2

สูตรการคำนวณ

$$r_{y1} = \frac{\sum x'_{1j} y'_j}{\sqrt{\sum x'^2_{1j} \sum y'^2_j}}$$

$$r_{y2} = \frac{\sum x'_{2j} y'_j}{\sqrt{\sum x'^2_{2j} \sum y'^2_j}}$$

$$r_{12} = \frac{\sum x'_{1j} x'_{2j}}{\sqrt{\sum x'^2_{1j} \sum x'^2_{2j}}}$$

ตัวอย่างที่ 9 การศึกษาเกี่ยวกับนิโคติน หน่วยทดลองคือผู้ชายที่สุขภาพดีอายุ 24 ถึง 48 ปี ที่สูบบุหรี่ประจำจำนวน 9 คน เก็บข้อมูล 3 อย่าง คือ (1) puffs per cigarette (2) total particulate matter per cigarette (3) nicotine intake per cigarette ผู้วิจัยต้องการวิเคราะห์หาความสัมพันธ์ระหว่าง 3 ตัวแปรนี้

ตาราง 4.6 ข้อมูลผู้สูบบุหรี่ 9 คน

X_1	X_2	Y
7.5	21.9	1.38
9.0	46.4	1.78
8.5	24.0	1.68
10.0	28.8	2.12
14.5	43.8	3.26
11.0	48.1	2.98
9.0	50.8	2.56
12.0	47.8	3.47
14.0	49.1	3.22

แหล่งที่มา : Neal L.Benowitz, Peyton Jacob III, Charles Denaro, and Roger Jenkins, “Stable Isotope Studies of Nicotine Kinetics and Bioavailability,” Clinical Phamacology & Therapeutics, 49 (1991) , 270-277.

กำหนดให้ X_1 คือ puffsper cigarette

X_2 คือ total particulate matter (mg / cigarette)

Y คือ nicotine intake / cigarette (mg)

วิธีทำ

$\sum x_{1j} = 95.50$ $\sum x_{2j} = 360.70$ $\sum y_j = 22.45$

$\sum x_{1j}^2 = 1061.75$ $\sum x_{2j}^2 = 15546.3$ $\sum y_j^2 = 60.8605$

$\sum x_{1j} x_{2j} = 3956.25$ $\sum x_{1j} y_j = 251.66$ $\sum x_{2j} y_j = 954.332$

1. การทรานส์ฟอร์ม

$$y'_j = (y_j - \bar{y}) \qquad x'_{1j} = (x_{1j} - \bar{x}_1) \qquad x'_{2j} = (x_{2j} - \bar{x}_2)$$

2. คำนวณค่าต่าง ๆ ต่อไปนี้

$$\sum x'^2_{1j} = 1061.75 - \frac{(95.5)^2}{9} = 48.38889$$

$$\sum x'^2_{2j} = 15546.3 - \frac{(360.7)^2}{9} = 1090.24556$$

$$\sum y'^2_j = 60.8605 - \frac{(22.45)^2}{9} = 4.86022$$

$$\sum x'_{1j} x'_{2j} = 3956.25 - \frac{(95.5)(360.7)}{9} = 128.82222$$

$$\sum x'_{1j} y'_j = 251.66 - \frac{(95.5)(22.45)}{9} = 13.44056$$

$$\sum x'_{2j} y'_j = 954.332 - \frac{(360.7)(22.45)}{9} = 54.58589$$

3. จากข้อ 2 ทำให้ได้สมการปกติ คือ

$$b_1 \sum x'^2_{1j} + b_2 \sum x'_{1j} x'_{2j} = \sum x'_{1j} y'_j$$

$$b_1 \sum x'_{1j} x'_{2j} + b_2 \sum x'^2_{2j} = \sum x'_{2j} y'_j$$

4. แทนค่าสมการปกติในข้อ 3

$$48.38889\, b_1 + 128.82222\, b_2 = 13.44056$$

$$128.82222\, b_1 + 1090.24556\, b_2 = 54.58589$$

5. แก้สมการปกติหาค่า b_1 b_2 และ b_0 หรือใช้สูตรคำนวณได้ผลลัพธ์ดังนี้

$$b_1 = 0.21077$$

$$b_2 = 0.02516$$

หาค่า b_0 ได้จาก

$$\begin{aligned} b_0 &= \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 \\ &= 2.494 - (0.21077)(10.611) - (0.02516)(40.08) \\ &= -0.75089 \end{aligned}$$

6. จะได้สมการถดถอยจากวิธีกำลังสองน้อยที่สุดคือ

$$\hat{y}_j = -0.75089 + 0.21077\, x_{1j} + 0.02516\, x_{2j}$$

7. ประเมินสมการถดถอยที่ได้โดยการคำนวณสัมประสิทธิของ multiple determination ($R^2_{y\bullet 12}$)

$$SSR = \sum(\hat{y}-\bar{y})^2 = b_1 \sum x'_{1j} y'_j + b_2 \sum x'_{2j} y'_j$$

$$= (0.21077)(13.44056) + (0.02516)(54.58589)$$

$$= 4.20625$$

$$SST = \sum(y_j-\bar{y})^2 = \sum y_j^2 - \frac{\left(\sum y_j\right)^2}{n}$$

$$= 60.8605 - \frac{(22.45)^2}{9}$$

$$= 4.86022$$

สัมประสิทธิของ multiple determination คือ

$$R^2_{y\bullet 12} = \frac{SSR}{SST} = \frac{4.20625}{4.86022} = 0.865444$$

และสัมประสิทธิสหสัมพันธ์พหุคูณคือ

$$R_{y.12} = \sqrt{0.865444} = 0.93029$$

8. การแปลความหมาย $R^2_{y.12}$ คือ 93 เปอร์เซ็นต์ของความแปรปรวนทั้งหมดในค่า Y อธิบายได้ด้วยระนาบของการถดถอยหรือโดยความสัมพันธ์เชิงเส้นกับ x_1 และ X_2 และการแปลความหมาย $R_{y.12}$ คือการวัดความสัมพันธ์ระหว่างตัวแปร nicotine intake per cigarette, number of puffs per cigarette และ total particulate matter per cigarette ในตัวอย่างผู้ชายที่สุขภาพดี 9 คน อายุระหว่าง 24 ถึง 48 ปี ถ้าข้อมูลของเราสุ่มมาจากประชากร เราสามารถใช้ $R_{y.12}$ เป็นค่าประมาณของ $\rho_{y.12}$ ซึ่งเป็นสัมประสิทธิสหสัมพันธ์ของประชากร นอกจากนี้ $R_{y.12}$ ยังแสดงสัมประสิทธิสหสัมพันธ์ระหว่าง y_j และ $\hat{y}$

9. คำนวณหาสัมประสิทธิสหสัมพันธ์อย่างง่ายต่อไปนี้

$$r_{y1} = \frac{13.44056}{\sqrt{(48.38889)(4.86022)}}$$

$$= 0.8764$$

$$r_{y2} = \frac{54.58589}{\sqrt{(1090.24556)(4.86022)}}$$

$$= 0.7499$$

$$r_{12} = \frac{128.82222}{\sqrt{(48.38889)(1090.24556)}}$$

$$= 0.5609$$

10. คำนวณหาสัมประสิทธิ์สหสัมพันธ์บางส่วน

$$r_{y1.2} = \frac{[0.8764-(0.7499)(0.5609)]}{\sqrt{(1-0.7499^2)(1-0.5609^2)}}$$

$$= 0.8322$$

$$r_{y2.1} = \frac{[0.7499-(0.8764)(0.5609)]}{\sqrt{(1-0.8764^2)(1-0.5609^2)}}$$

$$= 0.6479$$

$$r_{12.y} = \frac{[0.5609-(0.8764)(0.7499)]}{\sqrt{(1-0.8764^2)(1-0.7499^2)}}$$

$$= -0.3023$$

11.2 การทดสอบสมมติฐานเกี่ยวกับสัมประสิทธิสหสัมพันธ์บางส่วน

สมมติฐานทางสถิติที่ต้องการทดสอบคือ $H_0 : \rho_{y1.2...k} = 0$

สถิติที่ใช้ทดสอบ คือ $t = r_{y1.2...k}\sqrt{\frac{n-k-1}{1-r^2_{y1.2...k}}}$

ซึ่งมีการแจกแจงแบบ Student's t มีจำนวนชั้นอิสระ เท่ากับ $n-k-1$

ตัวอย่างที่ 10 ต้องการทดสอบสมมติฐาน $H_0 : \rho_{y1.2} = 0$ VS $H_1 : \rho_{y1.2} \neq 0$ โดยใช้ข้อมูลจากตัวอย่างที่ 9

วิธีทำ

คำนวณสถิติทดสอบ t

$$t = 0.8322\sqrt{\frac{9-2-1}{1-0.8322^2}}$$

$$= 3.6764$$

สำหรับการทดสอบแบบสองทางเปิดตารางหาค่าวิกฤติ $t_{\frac{\alpha}{z}, n-k-1}$ เขตวิกฤติคือ $|t| \geq t_{.025,\ 9-2-1}$ เนื่องจาก $|t|$ ที่คำนวณได้มากกว่าค่า $t_{.025,6}$ เท่ากับ 2.4469 จึงตัดสินใจปฏิเสธ H_0 สรุปว่ามีความสัมพันธ์อย่างมีนัยสำคัญระหว่าง nicotine intake และ puff per cigarette เมื่อให้ total particulate matter คงที่ การทดสอบสัมประสิทธิสหสัมพันธ์บางส่วนตัวอื่น ๆ ก็ทำเช่นเดียวกัน