12. ความถดถอยเชิงซ้อน

Multiple Regression

1. ความถดถอยและสหสัมพันธ์แบบเชิงซ้อน
2. การสร้างสมการถดถอยเชิงซ้อน
3. การใช้คอมพิวเตอร์ช่วยวิเคราะห์ความถดถอยเชิงซ้อน
4. การอนุมานเกี่ยวกับค่าพารามิเตอร์
5. ปัญหา Multicolinearity
6. แบบฝึกหัด

1. ความถดถอยและสหสัมพันธ์แบบเชิงซ้อน

เราได้ศึกษาวิธีวิเคราะห์ความถดถอยแบบเชิงเดี่ยว คือเมื่อมีตัวแปรอิสระเพียง 1 ตัว ในบทที่ 11 ในบางครั้งตัวแปรอิสระเพียงตัวเดียวอาจอธิบายความผันแปรของ Y หรือใช้พยากรณ์ค่า Y ได้ไม่ดีพอ เพราะยังมีตัวแปรอื่น ๆ อีกที่มีอิทธิพลต่อตัวแปรพึ่งพา อาทิเช่น ในธุรกิจที่ดินและบ้านจัดสรร ซึ่งจำนวนขายขึ้นอยู่กับค่าใช้จ่ายในการโฆษณาแล้ว ยังขึ้นอยู่กับจำนวนพนักงานขายด้วย ถ้าเราจะใช้ตัวแปรทั้ง 2 อย่างนี้ช่วยประมาณจำนวนขาย เราจะต้องสร้างสมการถดถอยแบบเชิงซ้อน ซึ่งจะมีขั้นตอนเหมือนการวิเคราะห์แบบเชิงเดี่ยว ดังนี้

1. สร้างสมการถดถอยเชิงซ้อน

2. หาความคลาดเคลื่อนมาตรฐานของค่าประมาณ

3. ใช้การวิเคราะห์สหสัมพันธ์แบบเชิงซ้อนตรวจสอบว่า สมการถดถอยเชิงซ้อนนั้นปรับเข้ากับข้อมูลได้ดีเพียงไร

ถ้ามีตัวแปรอิสระหลายตัว บางตัวอาจอธิบายความผันแปรของ Y ได้ไม่ดีนัก ดังนั้นต้องทดสอบนัยสำคัญของพารามิเตอร์

2. การสร้างสมการถดถอยเชิงซ้อน

สมมุติกรมสรรพากรต้องการประมาณภาษีค้างจ่ายต่อเดือนซึ่งตรวจพบโดยฝ่ายตรวจสอบ จำนวนรายการที่ตรวจพบจะขึ้นอยู่กับชั่วโมงการทำงานของฝ่ายตรวจสอบและชั่วโมงการทำงานของเครื่องคอมพิวเตอร์ที่ใช้ตรวจสอบภาษีค้างชำระ ให้ X แทนตัวแปรอิสระ เรามีตัวแปรอิสระ 2 ตัว คือ

X_1 = ชั่วโมงการทำงานของฝ่ายตรวจสอบ

X_2 = ชั่วโมงการทำงานของเครื่องคอมพิวเตอร์

Y = มูลค่าภาษีค้างชำระที่ตรวจพบ

ดังนั้น สมการแสดงความสัมพันธ์ของตัวแปรเหล่านี้ ของตัวอย่าง คือ

$$\hat{Y} = a + b_1X_1 + b_2X_2 \qquad (12.1)$$

a คือจุดตัดที่แกน Y

b_1 และ b_2 คือ ความลาดชันซึ่งเกี่ยวเนื่องกับ X_1, X_2 ตามลำดับ

สมการ (12.1) คือสมการของระนาบซึ่งมีแกน 3 ด้าน คือ X_1, X_2 และ Y เราจึงต้องการสมการ 3 อันเพื่อหาค่าของ a, b_1 และ b_2 คือ

$$\Sigma Y = na + b_1\Sigma X_1 + b_2\Sigma X_2 \quad (12.2)$$

$$\Sigma X_1 Y = a\Sigma X_1 + b_1\Sigma X_1^2 + b_2\Sigma X_1 X_2 \quad (12.3)$$

$$\Sigma X_2 Y = a\Sigma X_2 + b_1\Sigma X_1 X_2 + b_2\Sigma X_2^2 \quad (12.4)$$

สมมุติข้อมูลของ X_1, X_2, Y ในเวลา 10 เดือน คือ ตารางที่ 12.1

ตารางที่ 12.1 จำนวนภาษีค้างชำระที่ตรวจพบใน 10 เดือน

เดือน	X_1 จำนวนชั่วโมงตรวจสอบ (100 ชั่วโมง)	X_2 จำนวนชั่วโมงการทำงานของคอมพิวเตอร์ (100 ชั่วโมง)	Y ภาษีค้างชำระ (ล้านบาท)
ม.ค.	45	16	29
ก.พ.	42	14	24
มี.ค.	44	15	27
เม.ย.	45	13	25
พ.ค.	43	13	26
มิ.ย.	46	14	28
ก.ค.	44	16	30
ส.ค.	45	16	28
ก.ย.	44	15	28
ต.ค.	43	15	27

ตารางที่ 12.2 แสดงการคำนวณสำหรับสร้างสมการกำลังสองน้อยที่สุด

Y	X_1	X_2	X_1Y	X_2Y	X_1X_2	X_1^2	X_2^2	Y^2
(1)	(2)	(3)	(2)×(1)	(3)×(1)	(2)×(3)	$(2)^2$	$(3)^2$	$(1)^2$
29	45	16	1,305	464	720	2,025	256	841
24	42	14	1,008	336	588	1,764	196	576
27	44	15	1,188	405	660	1,936	225	729
25	45	13	1,125	325	585	2,025	169	625
26	43	13	1,118	338	559	1,849	169	676
28	46	14	1,288	392	644	2,116	196	784
30	44	16	1,320	480	704	1,936	256	900
28	45	16	1,260	448	720	2,025	256	784
28	44	15	1,232	420	660	1,936	225	784
27	43	15	1,161	405	645	1,849	225	729
272	441	147	12,005	4,013	6,485	19,461	2,173	7,428
↑	↑	↑	↑	↑	↑	↑	↑	↑
ΣY	ΣX_1	ΣX_2	ΣX_1Y	ΣX_2Y	ΣX_1X_2	ΣX_1^2	ΣX_2^2	ΣY^2

$\bar{Y} = 27.1$, $\bar{X}_1 = 44.1$, $\bar{X}_2 = 14.7$

แทนค่าต่าง ๆ ในสมการ (12.2), (12.3) และ (12.4) ดังนี้

$$272 = 10\,a + 441\,b_1 + 147\,b_2 \quad (1)$$

$$12{,}005 = 441\,a + 19{,}461\,b_1 + 6{,}485\,b_2 \quad (2)$$

$$4{,}013 = 147\,a + 6{,}485\,b_1 + 2{,}173\,b_2 \quad (3)$$

$$1 \times (-441): \quad -119{,}941 = -4410\,a - 194{,}481\,b_1 - 64{,}827\,b_2$$

$$2 \times (10): \quad 120{,}050 = 4410\,a + 194{,}610\,b_1 + 64{,}850\,b_2$$

$$(4): \quad 98 = 129\,b_1 + 23\,b_2$$

$$
\begin{aligned}
\text{①}\times(-147) &: \quad -39{,}984 = -1470\,a - 64{,}827\,b_1 - 21{,}609\,b_2 \\
\text{③}\times(10) &: \quad 40{,}130 = 1470\,a + 64{,}850\,b_1 + 21{,}730\,b_2 \\
\text{⑤} &: \quad 146 = 23\,b_1 + 121\,b_2
\end{aligned}
$$

$$
\begin{aligned}
\text{④}\times(-23) &: \quad -2{,}254 = -2967\,b_1 - 529\,b_2 \\
\text{⑤}\times(129) &: \quad 18{,}834 = 2{,}967\,b_1 + 15{,}609\,b_2 \\
\text{⑥} &: \quad 16{,}580 = 15{,}080\,b_2
\end{aligned}
$$

$$b_2 = \frac{16{,}580}{15{,}080}$$

$$\boxed{b_2 = 1.099}$$

แทนค่า b_2 ใน④

$$
\begin{aligned}
\text{④} \quad 98 &= 129\,b_1 + 23\,b_2 \\
98 &= 129\,b_1 = 23(1.099) \\
98 &= 129\,b_1 + 25.277 \\
72.723 &= 129\,b_1
\end{aligned}
$$

$$\boxed{b_1 = .564}$$

แทนค่า b_1, b_2 ใน①

$$
\begin{aligned}
\text{①} \quad 272 &= 10\,a + 441\,b_1 + 147\,b_2 \\
272 &= 10\,a + (441)(.564) + (147)(1.099) \\
272 &= 10\,a + 248.724 + 161.553 \\
-138.277 &= 10\,a
\end{aligned}
$$

$$\boxed{a = -13.828}$$

แทนค่า a, b_1, b_2 ใน (12.1)

$$\hat{Y} = a + b_1X_1 + b_2X_2 \qquad (12.1)$$

$$\text{⑦} \quad \hat{Y} = -13.828 + .564X_1 + 1.099X_2$$

สมมุติกรมสรรพากรจะเพิ่มชั่วโมงทำงานของคอมพิวเตอร์เป็น 1600 ชั่วโมง แต่ให้มีชั่วโมงตรวจสอบเท่าเดิมในเดือนพฤศจิกายน จะประมาณมูลค่าภาษีค้างชำระได้ โดยการแทนค่า

$X_1 = 43$ (4300 ชั่วโมง ในเดือนตุลาคม)

$X_2 = 16$ (1600 ชั่วโมง)

แทนค่าในสมการที่ (7) ดังนี้

$$\begin{aligned}\hat{Y} &= -13.828 + .564X_1 + 1.099X_2 \\ &= -13.828 + (.564)(43) + 1.099(16) \\ &= -13.828 + 24.252 + 17.584 \\ &= 28.008\end{aligned}$$

นั่นคือประมาณว่า จะพบรายการภาษีค้างชำระมูลค่า 28,008,000 บาท

ความหมายของ $a, b_1,$ **และ** b_2

เรายังคงเรียก a, b_1, b_2 ว่า สัมประสิทธิ์ความถดถอย และอธิบายได้ดังนี้

a คือค่าประมาณของ Y เมื่อ X_1, X_2 เป็น 0

b_1, b_2 แสดงการเปลี่ยนแปลงของ Y ซึ่งเป็นผลจากการเปลี่ยนแปลงของตัวแปร X_1, X_2

ในสมการที่ 7 ถ้าให้ X_1 คือชั่วโมงการตรวจสอบคงที่ และเปลี่ยนแปลงแต่ชั่วโมงทำงานของคอมพิวเตอร์ (X_2) จะมีผลให้ Y เพิ่มขึ้น 1,099,000 บาท ทุก ๆ 100 ชั่วโมงการทำงานของคอมพิวเตอร์ที่เพิ่มขึ้น และถ้าเราให้ X_2 คือชั่วโมงการทำงานของคอมพิวเตอร์คงที่ และเปลี่ยนแปลงแต่ X_1 มีผลให้ $\hat{Y}$ เพิ่มขึ้น 564,0000 บาท สำหรับทุก 100 ชั่วโมงการตรวจสอบที่เพิ่มขึ้น

แบบฝึกหัด

12.1 เหตุใดจึงใช้ความถดถอยเชิงซ้อนแทนความถดถอยเชิงเดียวในการประมาณค่าตัวแปรพึ่งพา?

12.2 ในการศึกษาความถดถอยเชิงซ้อน จะต้องใช้ตัวแปรอิสระกี่ตัว?

12.3 ต้องการศึกษาอิทธิพลต่าง ๆ ที่มีผลกระทบต่อจำนวนขาย ถ้าใช้ระดับรายได้ของผู้ซื้อเป็นตัวแปรตัวหนึ่ง และจะใช้อายุของผู้ซื้อ (ลูกค้า) เป็นตัวแปรอีกตัวหนึ่ง ท่านคิดว่าอายุของลูกค้าจะช่วยอธิบายความผันแปรของจำนวนขายได้เพิ่มขึ้นไหม?

12.4 จงสร้างสมการถดถอยเชิงซ้อน และประมาณค่า Y เมื่อ $X_1 = 4.8, X_2 = 4.0$

Y	X_1	X_2
34	5.0	5.0
29	4.2	4.5
43	8.5	10.0
12	1.4	2.5
35	3.6	5.0
27	1.3	3.0

($Y = 14.5 + 1.118X_1 + 2.201X_2$) (เมื่อ $X_1 = 4.8, X_2 = 4.0$ $\hat{Y} = 28.693$)

12.5 จงสร้างสมการถดถอยเชิงซ้อน และประมาณค่า Y เมื่อ $X_1 = 36, X_2 = 16$

Y	X_1	X_2
8	10	8
36	37	21
23	18	14
27	29	11
14	14	9
12	21	4

($Y = -2.995 + .5996X_1 + .905X_2$ และ $\hat{Y} = 33.067$ เมื่อ $X_1 = 36, X_2 = 16$)

12.6 ผู้ผลิตอาหารสุกรต้องการทราบความสัมพันธ์ระหว่างอายุสุกรเมื่อเริ่มต้นให้อาหารตามสูตรใหม่ น้ำหนักสุกรเมื่อเริ่มให้อาหารสูตรใหม่ และน้ำหนักที่เพิ่มขึ้นเมื่อสิ้นสัปดาห์ เขาได้ข้อมูลจากสุกร 8 ตัว ดังนี้

สุกร	X_1 น้ำหนักเริ่มต้น (ปอนด์)	X_2 อายุ (สัปดาห์)	Y น้ำหนักเพิ่ม (ปอนด์)
1	39	8	7
2	52	6	6
3	48	7	7
4	46	12	10
5	61	9	9
6	34	6	4
7	25	10	3
8	55	4	4

ก. จงสร้างสมการกำลังสองน้อยที่สุดที่อธิบายความสัมพันธ์ของตัวแปรทั้ง 3 นี้ ได้ดีที่

ข. เราคาดว่าสุกรที่มีอายุเริ่มต้น 9 สัปดาห์ และหนัก 48 ปอนด์ เมื่อให้กินอาหารสูตรใ
จะมีน้ำหนักเพิ่มใน 1 สัปดาห์ กี่ปอนด์

(ก) $Y = -6.36 + 0.155X_1 - 0.726X_2$, (ข) $\hat{Y} = 7.62$ ปอนด์)

12.7 จงสร้างสมการถดถอยเชิงซ้อน และพยากรณ์ค่า Y เมื่อ

$X_1 = 5$ และ $X_2 = 7$ ($Y = -0.08 + 2.29X_1 + .1089X_2$, $\hat{Y} = 12.14$)

Y	2	8	5	6	12	19
X_1	1	3	2	3	5	8
X_2	0	4	1	3	3	8

3. การใช้คอมพิวเตอร์ช่วยวิเคราะห์ความถดถอยเชิงซ้อน

ในปัจจุบันเครื่องคอมพิวเตอร์ขนาดเล็กมีราคาไม่แพงมาก และช่วยประหยัดเวลาและแรงงานในการคำนวณได้มหาศาล เครื่องคอมพิวเตอร์ขนาดเล็กจะมีโปรแกรมของความถดถอยทั้งแบบเชิงเดี่ยวและเชิงซ้อน โดยที่ผู้ใช้ไม่จำเป็นต้องมีความรู้ทางคอมพิวเตอร์มากนัก เพราะไม่ต้องสร้างโปรแกรมใหม่ ผู้ใช้เพียงแต่ป้อนข้อมูลให้กับเครื่องซึ่งมีวิธีการคล้ายกับการพิมพ์ดีด แต่จะต้องมีความรู้ในทางสถิติเพื่อจะได้วิเคราะห์ผลการคำนวณของคอมพิวเตอร์ ดังตัวอย่างเดิมเรื่องการตรวจสอบภาษีค้างชำระ ถ้าเราใช้ตัวแปรอิสระ 3 ตัว คือ

X_1 = จำนวนชั่วโมงแรงงานที่ใช้ตรวจสอบ มีหน่วยเป็น 100 ชั่วโมง

X_2 = จำนวนชั่วโมงการทำงานของเครื่องคอมพิวเตอร์เพื่อการตรวจสอบภาษีค้างชำระ มีหน่วยเป็น 100 ชั่วโมง

X_3 = มูลค่าของรางวัลจูงใจที่จ่ายให้ผู้ให้ข่าวสารที่มีประโยชน์ มีหน่วยเป็น 1000 บาท

Y = มูลค่าภาษีค้างชำระที่ตรวจพบ มีหน่วยเป็นล้านบาท

มีตัวแปร X_3 ของเดือนต่าง ๆ 10 เดือน ที่เพิ่มจากเดิม ค่าของ X_3 ของเดือนมกราคมถึงตุลาคม มีดังนี้

X_3 = รางวัล : 71 70 72 71 75 74 76 69 74 73

เมื่อใส่ข้อมูลของ X_1, X_2, X_3 และ Y ของ 10 เดือนเข้าโปรแกรมความถดถอยเชิงซ้อน เครื่องจะพิมพ์ output ดังนี้

ตารางที่ 12.3 แสดง output ของโปรแกรมความถดถอยเชิงซ้อนจากเครื่องคอมพิวเตอร์

VARIABLE	REG. COFF.	STD. ERROR COEF.	COMPUTED T
1	0.59697	0.08113	7.35866
2	1.17684	0.08407	13.99748
3	0.40511	0.04223	9.59200

INTERCEPT	−45.79634
MULTIPLE CORRELATION	0.99167
STD. ERROR OF ESTIMATE	0.28613

เราทราบอะไรบ้างจากตาราง 12.3

1. ทราบว่า**สมการถดถอยเชิงซ้อน** คือ

$$\hat{Y} = -45.796 + .597X_1 + 1.177X_2 + .405X_3$$

และอธิบายความหมายของสัมประสิทธิ์ความถดถอยได้ดังนี้

ก. ถ้าให้ X_1 (ชั่วโมงตรวจสอบ) และ X_2 (ชั่วโมงทำงานของคอมพิวเตอร์) **คงที่** แต่ให้ X_3 (รางวัลจูงใจ) เปลี่ยนแปลง จะมีผลให้ Y เพิ่มขึ้น 405,000 บาท สำหรับเงินรางวัลทุก ๆ 1,000 บาท ที่จ่ายเป็นค่าจูงใจ

ข. ถ้าให้ X_1 และ X_3 คงที่ และให้ X_2 เปลี่ยนแปลง จะมีผลให้ $\hat{Y}$ เพิ่มขึ้น 1,177,000 บาท สำหรับทุก ๆ 100 ชั่วโมง การทำงานที่เพิ่มขึ้นของเครื่องคอมพิวเตอร์

ค. สุดท้ายคือ ถ้าให้ X_2, X_3 คงที่ และให้ X_1 เปลี่ยนแปลง จะมีผลให้ $\hat{Y}$ เพิ่มขึ้น 597,000 บาท สำหรับทุก 100 ชั่วโมง การตรวจสอบที่เพิ่มขึ้น

สมมุติว่า ในเดือนพฤศจิกายน จะให้ชั่วโมงตรวจสอบ และชั่วโมงการทำงานของคอมพิวเตอร์เท่าเดิมของเดือนตุลาคม คือ 4300 และ 1500 ชั่วโมง แต่จะเพิ่มเงินจูงใจเป็น 75,000 คาดว่าจะพบภาษีค้างชำระเท่าใด การหาค่าพยากรณ์ $\hat{Y}$ (พฤศจิกายน) ได้จากการแทนค่า $X_1 = 43, X_2 = 15$ และ $X_3 = 75$ เข้าในสมการถดถอย ดังนี้

$$\begin{aligned}\hat{Y} &= -45.796 + .597X_1 + 1.177X_2 + .405X_3 \\ &= -45.796 + .597(43) + 1.177(15) + .405(75) \\ &= -45.796 + 25.671 + 17.655 + 30.375 \\ &= 27.905\end{aligned}$$

นั่นคือ ในเดือนพฤศจิกายน คาดว่าจะตรวจพบภาษีค้างชำระมูลค่าประมาณ 28 ล้านบาท

2. ทราบมาตราวัดการกระจาย คาดว่าจะตรวจพบภาษีค้างชำระมูลค่าประมาณ

ในการวัดการกระจายของข้อมูลโดยรอบระนาบความถดถอย จะใช้วิธีการเหมือนการวิเคราะห์แบบเชิงเดี่ยว ซึ่งเราทราบว่า ค่าประมาณจะมีความแม่นยำขึ้น เมื่อการกระจายโดยรอบเส้นถดถอยต่ำ ซึ่งเราวัดโดย S_e หรือความคลาดเคลื่อนมาตรฐานของค่าประมาณ

$$S_e = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - k - 1}} \qquad (12.2)$$

ในเมื่อ

Y = ค่าของตัวแปรพึ่งพาจากตัวอย่าง

$\hat{Y}$ = ค่าประมาณของ Y จากสมการถดถอย

n = จำนวนจุด (ขนาดตัวอย่าง)

k = จำนวนตัวแปรอิสระ (ในตัวอย่าง k = 3)

การที่ตัวหารหรือ df ของ S_e เหลือ $n - k - 1$ เพราะเราต้องประมาณพารามิเตอร์ $k + 1$ ตัว คือ $\alpha, \beta_1, \beta_2, \ldots \beta_k$ ด้วยค่าจากตัวอย่างคือ $a, b_1, b_2, \ldots, b_k$

จากตัวอย่าง $S_e = .286$ เราจะใช้ S_e และค่าตาราง t ประมาณช่วงเชื่อมั่นของ $\mu_{y/x}$ ของเดือนพฤศจิกายน ซึ่งได้ค่าประมาณแบบจุดคือ 27,905,000 บาท

จากตาราง t ที่ $df = n - k - 1 = 10 - 3 - 1 = 6$

$t_{.025, 6} = 2.447$

ดังนั้น จะประมาณ 95% ช่วงเชื่อมั่นของ $\mu_{y/x}$ คือ

$\hat{Y} \pm t(S_e) = 27,905,000 \pm (2.447)(286,000)$

$= 28,604,800, 27,205,200$

นั่นคือ กล่าวด้วยความเชื่อมั่น 95% ได้ว่า จะพบมูลค่าภาษีค้างชำระในเดือนพฤศจิกายนระหว่าง 27,205,200 - 28,604,800 บาท

มีข้อน่าสังเกตข้อหนึ่งคือ การที่เราเพิ่มตัวแปร X_3 เข้าในสมการอีกตัวหนึ่งนั้น มีผลทำให้การประมาณค่าดีขึ้นไหม? อันนี้ต้องดู S_e ถ้า S_e เล็กลงก็แสดงว่า ตัวแปรที่เพิ่มเข้าไปช่วยในการประมาณค่า จากตัวอย่างเมื่อมีตัวแปร 2 ตัว คือ X_1, X_2 มีค่า $S_e = 1.076$ เมื่อเพิ่ม X_3 เข้าไป $S_e = .286$ ซึ่งเล็กลงกว่าเดิมมาก จึงแสดงว่าการเพิ่ม X_3 ช่วยปรับปรุงสมการถดถอยให้ "fit" กับข้อมูลได้ดีขึ้น อย่างไรก็ตาม พึงระลึกว่า การเพิ่มตัวแปรไม่จำเป็นต้องลดค่า S_e เสมอไป

3. ทำให้ทราบ**สัมประสิทธิ์การตัดสินใจเชิงซ้อน** (coefficient of multiple determination) ในการวิเคราะห์ความถดถอยแบบเชิงเดี่ยว เราวัดความแรงของความสัมพันธ์ระหว่าง X และ Y ด้วย r^2 ในเมื่อ r^2 คือส่วนหนึ่งของความผันแปรของ Y ทั้งหมดที่ "อธิบาย" โดยสมการถดถอย

ในทำนองเดียวกันในสหสัมพันธ์แบบแชิงซ้อน (multiple correlation) เมื่อเราต้องการวัดความแรงของความสัมพันธ์ระหว่างตัวแปรต่าง ๆ เราจะวัดโดยการใช้สัมประสิทธิ์การตัดสินใจเชิงซ้อน (R^2) หรือกรณีที่สอง คือ R ซึ่งคือสัมประสิทธิ์สหสัมพันธ์เชิงซ้อน ในเมื่อ R^2 คือ สัดส่วนของความผันแปรทั้งหมดของ Y ซึ่ง "อธิบาย" โดยระนาบถดถอย

จาก out put คอมพิวเตอร์ในตาราง 12.3 ให้ค่า R = .99167 ดังนั้น $R^2 = (.99167)^2$ = .983 นั่นคือ 98.3% ของความผันแปรทั้งหมดของมูลค่าภาษีค้างชำระที่ตรวจพบสามารถอธิบายได้ โดยตัวแปรอิสระ 3 ตัวนั้น ถ้าใช้ตัวแปรอิสระเพียง 2 ตัว คือ X_1 และ X_2 จะได้ R^2 = .726 หรือเพียง 72.6% ของความผันแปรที่สามารถอธิบายได้โดยชั่วโมงการตรวจสอบและชั่วโมงการทำงานของเครื่องคอมพิวเตอร์ การเพิ่ม X_3 คือรางวัลจูงใจช่วยให้อธิบายความผันแปรได้เพิ่มขึ้น 25.7%

สำหรับช่อง "computed t" จะกล่าวถึงในหัวข้อต่อไปเรื่องการอนุมานเกี่ยวกับค่าประชากรของความถดถอยเชิงซ้อน

แบบฝึกหัด

12.8 จงใช้โปรแกรมคอมพิวเตอร์ (computor package) แบบใดก็ได้ที่ท่านหาได้ สร้างสมการถดถอยที่ปรับเข้ากับข้อมูลที่กำหนดให้ได้ดีที่สุด

ก) หาสมการถดถอย ($Y = -19.47 - 3.92X_1 = 17.33X_2 = 6.6768X_3$)

ข) หาความคลาดเคลื่อนมาตรฐานของค่าประมาณ

ค) หาค่า R^2 (.980558)

ง) จงประมาณค่า Y เมื่อ $X_1 = 5.8, X_2 = 4.2, X_3 = 6.1$ ($\hat{Y} = 71.3202$)

Y	X_1	X_2	X_3
74.6	5.3	4.0	6.8
90.8	6.1	4.7	7.2
64.2	5.0	3.6	6.4
103.4	7.7	5.4	9.1
77.7	5.5	4.1	7.0
60.2	3.8	2.9	6.0
96.6	7.2	5.2	8.6
34.3	2.1	2.2	4.1

12.9 ฝ่ายสินเชื่อของธนาคารแห่งหนึ่งต้องการประมาณยอดเงินค้างชำระของเดือนต่าง ๆ โดยอาศัยความสัมพันธ์จากตัวแปรจากเดือนต่าง ๆ ที่ผ่านมา 6 เดือน

Y = ยอดเงินกู้ค้างชำระ

X_1 = ยอดเงินกู้โดยถัวเฉลี่ยต่อคน

X_2 = จำนวนผู้กู้เงิน

X_3 = อัตราเงินเฟ้อของเดือนที่ผ่านมาเป็นเปอร์เซ็นต์

Y	X_1	X_2	X_3
2,033	1,722	697	1.2%
1,908	2,100	528	0.7
1,541	2,694	466	0.9
3,406	1,229	806	1.1
926	3,661	512	0.8
806	2,944	405	1.0

ก) จงสร้างสมการถดถอยโดยใช้โปรแกรมของเครื่องคอมพิวเตอร์

($Y = 2275.88 - 0.605X_1 + 3.6X_2 - 1168.6486X_3$)

ข) จงหาค่า R^2 (.93995)

ค) จงหาความคลาดเคลื่อนมาตรฐานของค่าประมาณ (366,195.98)

ง) ถ้าผู้จัดการธนาคารประมาณว่า ในเดือนตุลาคม จะมีจำนวนเงินกู้โดยเฉลี่ยคนละ 1,995 บาท และมีผู้กู้ทั้งหมด 516 ราย และอัตราเงินเฟ้อของเดือนกันยายนคือ 1.2% จงประมาณช่วงเชื่อมั่น 95% ของยอดเงินกู้ค้างชำระประจำเดือนตุลาคม ($\hat{Y} = 1967.84$)

12.10 อาจารย์ภาควิชาสถิติศาสตร์และคอมพิวเตอร์ผู้หนึ่ง ได้ศึกษาผลสอบของนักศึกษาคณะบริหารธุรกิจ โดยใช้ความสัมพันธ์ระหว่างตัวแปรต่าง ๆ และใช้เครื่องคอมพิวเตอร์ช่วยคำนวณ โดยให้

คอลัมน์ 1 คะแนนสอบ

คอลัมน์ 2 จำนวนชั่วโมงที่ใช้ศึกษาสถิติ

คอลัมน์ 3 ระดับสติปัญญาของนักศึกษา

คอลัมน์ 4 จำนวนหนังสือที่ใช้ประกอบเพื่อเตรียมการสอบ

คอลัมน์ 5 อายุของนักศึกษา

เขาได้ output ดังนี้

VARIABLE	REG. COEF.	STD. ERROR COEF.	COMPUTED T
2	1.13119	0.92689	1.22042
3	1.28289	0.37918	3.38334
4	1.59373	1.65459	0.96322
5	-1.88029	0.77399	-2.42934

INTERCEPT	-39.62912
MULTIPLE CORRELATION	0.85693
STD. ERROR OF ESTIMATE	12.51129

ก. จงหาสมการถดถอย

ข. ความผันแปรของคะแนนกี่เปอร์เซ็นต์ที่อธิบายได้โดยสมการถดถอย (73.43%)

ค. ถ้านักศึกษาผู้หนึ่งอายุ 21 ปี ซึ่งมีระดับสติปัญญา 113 หน่วย และใช้เวลาศึกษา 5 ชั่วโมง จากหนังสือ 3 เล่ม จงประมาณคะแนนสอบของนักศึกษาผู้นี้ (76.2885 คะแนน)

12.11 บริษัทเงินทุนแห่งหนึ่งมีกิจการหลักคือการซื้อ-ขายหุ้นโดยใช้พนักงานที่จบสาขาบริหารธุรกิจ แต่พนักงานที่พึ่งจบใหม่ยังมีประสพการณ์ในการซื้อ-ขายหุ้นไม่พอ บริษัทจึงให้พนักงานเข้าหลักสูตรฝึกอบรมซึ่งมีอยู่หลายวิชา ในการใช้เครื่องคอมพิวเตอร์ ได้กำหนดให้

คอลัมน์ 1 จำนวนขายในเดือนนั้น (1,000 บาท)
คอลัมน์ 2 จำนวนหลักสูตรที่เข้าอบรมในปีนี้
คอลัมน์ 3 จำนวนปีที่ขายหุ้น (ประสพการณ์)
คอลัมน์ 4 จำนวนลูกค้าที่มาติดต่อในระหว่างเดือน

ผลการคำนวณมีดังนี้

VARIABLE	REG. COEF.	STD. ERROR COEF.	COMPUTED T
2	4.77515	1.40751	3.39263
3	-0.25491	0.52873	-0.48212
4	0.01561	0.33620	0.04643

INTERCEPT	2.54139
MULTIPLE CORRELATION	0.97937
STD. ERROR OF ESTIMATE	4.88820

ก) จงหาสมการถดถอย ($Y = 2.54 + 4.775X_1 - 0.255X_2 + 0.01561X_3$)

ข) จงหา S_e (4.88820)

ค) จงหาค่า R^2 (.9591655)

ง) ถ้าพนักงานเข้าฝึกอบรม 2 วิชา มีประสพการณ์ในการขาย 3 ปี และมีลูกค้ามาติดต่อภายในเดือนนั้น 61 คน จงประมาณยอดขายของพนักงานผู้นี้ (11.0839)

---

4. การอนุมานเกี่ยวกับค่าพารามิเตอร์

เมื่อเราศึกษาความถดถอยเชิงเดี่ยว เราทราบว่า $\hat{Y} = a + bX$ เป็นค่าประมาณของเส้นถดถอยประชากร $Y = \alpha + \beta X$ ในการศึกษาความถดถอยเชิงซ้อนก็เช่นกัน เราจะมีระนาบถดถอยจากตัวอย่างคือ

$$\hat{Y} = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k \qquad (12.3)$$

เป็นค่าประมาณของระนาบถดถอยของประชากร ซึ่งเราไม่ทราบค่าแท้จริง คือ

$$Y = \alpha + \beta_1X_1 + \beta_2X_2 + \ldots + \beta_kX_k \qquad (12.4)$$

ดังนั้น ในการอนุมาณระนาบถดถอยของประชากร เราจะต้องอนุมาณค่าความลาดชัน คือ $\beta_1, \beta_2, \ldots, \beta_k$ ซึ่งมี $b_1, b_2, \ldots, b_k$ เป็นค่าประมาณเราจะแบ่งหัวข้อการอนุมาณเป็น 4 ตอน คือ

1. การทดสอบ Ho : $\beta_i = 0, i = 1, 2, \ldots, k$
2. การทดสอบ Ho : $\beta_i = \beta_i'$, $\beta_i' \neq 0$

3. การสร้างช่วงเชื่อมั่นของ β_i

4. การทดสอบ Ho : $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

Ha : อย่างน้อยมี β_i ตัวหนึ่ง $\neq 0$

1. **การทดสอบ** Ho : $\beta_i = 0$, Ha : $\beta_i \neq 0$ ก็คือการทดสอบว่า X_i มีความจำเป็นสำหรับสมการถดถอยเชิงซ้อนหรือไม่ ตัวสถิติที่ใช้ทดสอบคือ

$$T = \frac{b_i}{S_{b_i}}$$

และจะมีค่าที่คำนวณได้ในคอลัมน์สุดท้ายของ output คือคอลัมน์ COMPUTED T ค่าที่คำนวณได้นี้ ต้องเทียบกับค่าในตาราง Student t ที่ df = n – k – 1 เช่นจากตัวอย่างเรื่องภาษีค้างชำระ มี out put ในตารางที่ 12.3 ดังนี้

ตารางที่ 12.3

VARIABLE	REG. COEF.	STD. ERROR COEF.	COMPUTED T
1	0.59697	0.08113	7.35866
2	1.176884	0.08407	13.99748
3	0.40511	0.04223	9.59200
INTERCEPT		-45.79634	
MULTIPLE CORRELATION		0.99167	
STD. ERROR OF ESTIMATE		0.28613	

ถ้าเราต้องการทดสอบว่า X_1 มีความสำคัญสำหรับการประมาณค่า Y หรือไม่ มีวิธีดังนี้

Ho : $\beta_1 = 0$ (X_1 ไม่จำเป็นในการประมาณค่า Y)

Ha : $\beta_2 \neq 0$ (X_1 จำเป็นในการประมาณค่า Y)

$$T = \frac{b_1}{S_{b_1}} = \frac{0.59697}{0.08113} = 7.35866$$

ค่า 7.35866 อยู่คอลัมน์สุดท้ายใน output เราจึงไม่ต้องคำนวณ แต่เราต้องสรุปผลการทดสอบโดยเทียบกับค่าจากตาราง t สมมุติใช้ระดับนัยสำคัย 10% ต้องเปิดที่ $\alpha = .05$, df = n − k − 1 = 10 − 3 − 1 = 6 จากตาราง $t_{.05, 6} = 1.943$ ค่าที่คำนวณได้ใหญ่กว่าค่าเปิดตาราง จึงปฏิเสธ Ho และสรุปว่าตัวแปร X_1 มีความสำคัญในการประมาณค่า Y สำหรับการทดสอบตัวแปร X_2, X_3 ก็เช่นกัน จะเห็นว่า ค่า T ที่คำนวณได้สำหรับทดสอบ X_2 = 13.997 และสำหรับทดสอบ X_3 ค่า T คำนวณได้ 9.59 ซึ่งใหญ่กว่าค่าเปิดตาราง (1.943) ทั้ง 2 ค่า จึงสรุปว่าทั้ง X_2, X_3 มีประโยชน์สำหรับใช้ประมาณค่า Y

2. สำหรับการทดสอบ $Ho : \beta_i = \beta_i'$ $\quad (\beta_i' = 0)$

$Ha : \beta_i \neq \beta_i'$

ก็มีวิธีการเหมือนข้อ (1) แต่คอมพิวเตอร์ไม่ได้คำนวณให้ ต้องทดสอบเอง เช่น สมมุติว่าแต่เดิม Y และ X_1 มีความสัมพันธ์กันโดยมีความลาดชัน .400 เราอยากทราบว่า ยังมีความสัมพันธ์เหมือนเดิมหรือไม่ นั่นคือ การทดสอบว่า

$Ho : \beta_1 = .400 \quad (B_1' = .400)$

$Ha : \beta_1 \neq .400$

$$T = \frac{b_1 - \beta_1'}{S_{b_1}} = \frac{.597 - .400}{0.0811} = 2.43$$

และเทียบกับค่า $t_{.05, 6} = 1.943$ ค่าที่คำนวณได้มีนัยสำคัญจึงปฏิเสธ Ho และยอมรับ Ha ว่าความลาดชันได้เปลี่ยนไปจากเดิมด้วยความเชื่อมั่น 90% นั่นคือ ทุก ๆ 100 ชั่วโมง การตรวจสอบที่เพิ่มขึ้น จะมีผลให้พบยอดภาษีค้างชำระเพิ่มขึ้นต่างไปจาก 400,000 บาท ซึ่งเป็นยอดเดิม (พบมากขึ้นกว่า 400,000 บาท) สำหรับการตรวจสอบพารามิเตอร์ β_2, β_3 ก็ทำเช่นเดียวกัน

3. การสร้างช่วงเชื่อมั่นของ β_i มีสูตรดังนี้

$$b_i \pm t(S_{b_i})$$

t คือค่าจากตาราง t ที่ n − k − 1 df

ค่า b_i และ S_{b_i} ได้จาก output ในตารางที่ 12.3

เช่นต้องการสร้างช่วงเชื่อมั่น 95% ของ β_3

$$b_3 = 0.405 \text{ , } S_{b_3} = 0.0422$$

เปิดค่า t จากตาราง ได้ $t_{.025,\ 6} = 2.447$

ดังนั้น ช่วงเชื่อมั่น 95 % ของ β_3 คือ

$$0.405 \pm 2.447(0.0422)$$
$$= .302, .508$$

นั่นคือ เราจะกล่าวด้วยความเชื่อมั่น 95% ได้ว่าทุก ๆ 1,000 บาทที่เพิ่มขึ้นของเงินจูงใจ จะมีผลทำให้การตรวจพบยอดภาษีค้างชำระเพิ่มขึ้นระหว่าง 302,000 ถึง 508,000 บาท

หรือ ช่วงเชื่อมั่น 90% ของ β_1 คือ

$$0.597 \pm (1.943)(0.0811)$$
$$= 0.439, 0.755$$

นั่นคือ กล่าวด้วยความเชื่อมั่น 95% ได้ว่า ทุก ๆ 100 ชั่วโมง การตรวจสอบที่เพิ่มขึ้น จะมีผลทำให้การตรวจพบยอดภาษีค้างชำระเพิ่มขึ้นระหว่าง 439,000 ถึง 755,000 บาท

พึงสังเกตว่าการหาช่วงเชื่อมั่นและการทดสอบสมมติฐาน เมื่อใช้ระดับนัยสำคัญเท่ากัน จะได้ข้อสรุปเหมือนกัน ดังเช่น การทดสอบ Ho : $\beta_1 = 0$ โดยใช้ $\alpha = .05$ ถ้าเราดูจากช่วงเชื่อมั่น 95% ของ β_1 จะไม่ยอมรับ Ho และจะยอมรับ Ha : $\beta_1 \neq 0$ แสดงว่า X_1 มีความสำคัญในการประมาณค่า Y ซึ่งเป็นข้อสรุปเหมือนการทดสอบในขั้นตอนที่ 1 หรือในการทดสอบ

Ho : $\beta_1 = .400$, Ha : $\beta_1 \neq .400$ เมื่อตรวจดูในช่วงเชื่อมั่น คือ $.439 < \beta_1 < .755$ จะเห็นว่า .400 ไม่อยู่ช่วงเชื่อมั่น จึงต้องปฏิเสธ Ho : $\beta_1 = .400$ มารับ Ha : $\beta_1 \neq .400$ จะเห็นว่า .400 ซึ่งเป็นข้อสรุปเหมือนกับการทดสอบสมมุติฐานในขั้นตอนที่ 2 นั่นเอง

4. **การทดสอบ** Ho : $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

Ha : มีอย่างน้อย β_i ตัวหนึ่งที่ $\neq 0$

ความหมายของ Ho คือตัวแปร X ทุกตัวไม่มีประโยชน์ในการประมาณค่า Y ส่วน Ha หมายความว่า Y ขึ้นอยู่กับตัวแปร X_i อย่างน้อย 1 ตัว ตัวสถิติที่ใช้ทดสอบคือ

$$F = \begin{matrix} MSR & \longrightarrow & \text{มี } k \text{ df} \\ MSE & \longrightarrow & \text{มี } n-k-1 \text{ df} \end{matrix}$$

และเปรียบกับค่า F ในตารางที่ $\nu_1 = k - 1$, $\nu_2 = n - k - 1$

ค่า F นี้ เราสามารถให้คอมพิวเตอร์พิมพ์ output คือตารางวิเคราะห์ความแปรปรวนของความถดถอยเชิงซ้อน ดังตัวอย่างในตารางที่ 12.4 ดังนี้

ตารางที่ 12.4 ตารางวิเคราะห์ความแปรปรวน

SOURCE OF VARIATION	D.F	SUM OF SQ.	MEAN SQ.	F VALUE
ATTRIBUTABLE TO REG.	3	29.109	9.703	118.515
DEVIATION FROM REG.	6	0.491	0.082	
TOTAL	9	29.600		

ค่าในตารางคือ SSR = 29.109 และมี df = k = 3

SSE = 0.491 และมี df = n − k − 1 = 10 − 3 − 1 = 6

ดังนั้น $$F = \frac{29\ 109/3}{0.491/6} = \frac{9.703}{0.082} = 118.515$$

เมื่อเทียบกับค่าจากตาราง $f_{3,\,6,\,.01} = 9.78$ ค่า F ที่คำนวณได้โตกว่าค่าเปิดตาราง จึงปฏิเสธ Ho มารับ Ha นั่นคือ สรุปว่า สมการถดถอยทั้งสมการมีประโยชน์ในการประมาณค่า Y

5. ปัญหา **MUTICOLINEARITY**

เมื่อตัวแปรอิสระมีสหสัมพันธ์ระหว่างกันค่อนข้างสูง จะเกิดปัญหาเรียกว่า multicoli--nearity เช่น ถ้าเราต้องการประมาณยอดขายรายเดือนของสินค้าชนิดหนึ่งโดยให้ X_1 = จำนวนหน้าโฆษณาในเดือนนั้น X_2 = จำนวนเงินค่าโฆษณาซึ่งต้องจ่ายล่วงหน้า 1 เดือน และมีหน่วยเป็น 100 บาท และ Y คือยอดขายสินค้าเดือนนั้น (มีหน่วยเป็น 1,000 บาท) และเก็บข้อมูลจำแนกรายเดือนรวม 12 เดือน มีตารางวิเคราะห์ความแปรปรวนแบบเชิงเดี่ยวของตัวแปร X_1 และ X_2 ในตาราง 12.5 และ 12.6 และการวิเคราะห์แบบเชิงซ้อนในตารางที่ 12.7 ดังนี้

ตารางที่ 12.5 ความถดถอยของยอดขายบนจำนวนหน้าโฆษณา

VARIABLE	REG. COEF.	STD. ERROR COEF.	COMPUTED T
1	2.08324	0.52709	3.95238

INTERCEPT	16.93689
MULTIPLE CORRELATION	0.78083
STD. ERROR OF ESTIMATE	4.20569

ANOVA

SOV	D.F	SS	MS	F VALUE
REGRESSION	1	276.307	276.307	15.621
DEV. FROM REG.	10	176.878	17.688	
TOTAL	11	453.186		

ตารางที่ 12.6 ความถดถอยของยอดขายบนจำนวนเงินค่าโฆษณา

VARIABLE	REG. COEF.	STD. ERROR COEF.	COMPUTED T
2	2.87247	0.63304	4.53760

INTERCEPT	4.17282
MUTTIPLE CORRELATION	0.82042
STD. ERROR OF ESTIMATE	3.84901

ANOVA

SOV	D.F.	SS	MS	F VALUE
REGRESSION	1	305.037	305.037	20.590
DEV. FROM REG.	10	148.149	14.815	
TOTAL	11	453.186		

จากตาราง 12.5 และ 12.6 จะเห็นว่าค่า T ที่คำนวณได้ของ X_1 = 3.952 และของ X_2 = 4.53760 ทั้งคู่โตกว่า $t_{.005, 10}$ = 3.169 จึงสรุปได้ว่า ทั้งจำนวนหน้าโฆษณา และจำนวนเงินค่าโฆษณาเป็นตัวแปรที่สำคัญสำหรับใช้อธิบายความผันแปรของยอดขาย r^2 ของ X_1 = $(.78083)^2$ = .6097 นั่นคือ จำนวนหน้าโฆษณาอธิบายความผันแปรของจำนวนขายได้ประมาณ 61% ส่วน r^2 ของ X_2 = $(.82042)^2$ = .6731 นั่นคือ จำนวนเงินที่ใช้โฆษณาอธิบายความผันแปรของยอดขายได้ประมาณ 67% จะเห็นว่าทั้ง X_1 และ X_2 ต่างก็เป็นตัวแปรที่มีนัยสำคัญเมื่อเราวิเคราะห์แบบเชิงเดี่ยว

แต่เมื่อนำ X_1 และ X_2 มาพิจารณาพร้อม ๆ กัน จะได้ตาราง 12.7 ดังนี้

ตารางที่ 12.7 ความถดถอยของจำนวนขายบนจำนวนหน้า และจำนวนเงินค่าโฆษณา

VARIABLE	REG. COEF.	STD. ERROR COEF.	COMPUTED T
1	0.62470	1.12028	0.55763
2	2.13882	1.47015	1.45483
INTERCEPT		6.58379	
MULTIPLE CORRELATION		0.82705	
STD. ERROR OF ESTIMATE		3.98890	

ANOVA

SOV	DF	SS	MS	F VALUE
REGRESSION	2	309.984	154.992	9.741
DEV. FROM REG.	9	143.202	15.911	
TOTAL	11	453.186		

จากตารางที่ 12.7 ค่า F = 9.741 เทียบกับ $f_{2, 9; .01}$ = 8.02 จึงมีนัยสำคัญ นั่นคือสมการถดถอยเชิงซ้อนใช้อธิบาย Y ได้ และมีค่าสัมประสิทธิ์การตัดสินใจเชิงซ้อน คือ R^2 = $(.82705)^2$ = .6840 นั่นคือ ทั้ง X_1 และ X_2 ช่วยอธิบายความผันแปรของ Y ได้ประมาณ 68%

แต่เมื่อพิจารณาค่า T ที่คำนวณได้ จะเห็นว่าเป็นค่าที่เล็กและไม่มีนัยสำคัญทั้งคู่ แม้ว่าจะให้ α โตถึง .10 , $t_{.05, 9}$ = 1.833 เหตุใดค่า T ในตาราง 12.7 จึงไม่มีนัยสำคัญ ในขณะที่ค่า T

ในตารางที่ 12.5 และ 12.6 ซึ่งเป็นการวิเคราะห์แบบเชิงเดี่ยวมีนัยสำคัญ และในขณะเดียวกัน ค่า F ซึ่งแสดงการวิเคราะห์โดยรวม X_1, X_2 เข้าด้วยกันก็มีนัยสำคัญ คำตอบก็คือ เพราะ X_1 และ X_2 มีสหสัมพันธ์ค่อนข้างสูง คือ r = .8949 จึงเป็นปัญหาของ multicolinearity อันที่จริงตัวแปรคู่นี้น่าจะมีความสัมพันธ์แบบสมบูรณ์ แต่เหตุที่ $r \neq 1.0$ เพราะหน้าโฆษณามีราคาแตกต่างกัน เมื่อ X_1 และ X_2 มีสหสัมพันธ์กันอย่างใกล้ชิด นั่นคือทั้ง X_1 และ X_2 ต่างก็อธิบายความผันแปรส่วนเดียวกัน จะเห็นว่า r^2 ของ X_1 = .6097 และ r^2 ของ X_2 = .6731 และในความถดถอยเชิงซ้อน R^2 = .6830 นั่นคือเมื่อรวม X_1 เข้ากับ X_2, X_1 ช่วยเพิ่มการอธิบายได้เพียง 1 % ผลจากที่ X_1, X_2 มีสหสัมพันธ์สูง จึงทำให้ S_{b_1} และ S_{b_2} มีค่าสูงและในที่สุดค่าสถิติ T จึงมีค่าต่ำ เพราะ $T = b_i/S_{b_i}$ ค่า T ที่คำนวณได้จึงไม่มีนัยสำคัญเมื่อเกิดปัญหาเช่นนี้ จึงควรใช้การวิเคราะห์แบบเชิงเดี่ยว คือเลือกตัวแปรที่ดีที่สุดเพียงตัวเดียว ในกรณีนี้ควรเลือก X_2 (จำนวนเงินค่าโฆษณา) เพราะให้ค่า r^2 สูงกว่า

แบบฝึกหัด

12.12 กำหนด output ของคอมพิวเตอร์ ซึ่งมี n = 8 ดังนี้

VARIABLE	REG. COEF.	STD. ERROR COEF.	COMPUTED T
1	0.23249	0.08337	2.78863
2	1.15057	0.09967	11.54347

INTERCEPT	18.01212
MULTIPLE CORRELATION	0.99568
STD. ERROR OF ESTIMATE	0.94551

ก. จงทดสอบว่า β_1 เป็น .150 หรือไม่โดยใช้ α = .05 (T = .99, ยอมรับ Ho)

ข. จงประมาณช่วงเชื่อมั่น 90% ของ β_1 $(.064 < \beta_1 < .400)$

12.13 จากข้อ 12.10 ให้ใช้ α = .05 ทดสอบว่า ระดับสติปัญญาเป็นตัวแปรที่ใช้อธิบายคะแนนสอบได้หรือไม่ (มีนักศึกษาตัวอย่าง 12 คน) (T = 3.338 > 2.365, ปฏิเสธ Ho, IQ ใช้อธิบายได้)

12.14 จากข้อ 12.10 ถ้ามี Output เพิ่มเติม ดังนี้

SOV	df	SS	MS
Regression	4	3,028.523	757.131
Dev. from reg.	7	1,095.727	156.532
Total	11	4,124.250	

ก) จงคำนวณค่า F (F = 4.84)

ข) จงใช้ α = .01 ทดสอบนัยสำคัญของ F และสรุปผลการทดสอบ

(ยอมรับ Ho : $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ ตัวแปรทั้ง 4 ใช้อธิบายความผันแปรของคะแนนไม่ได้ นั่นคือ สมการตัวแบบที่สร้างใช้พยากรณ์คะแนนไม่ได้)

12.15 จากข้อ 12.11 จงทดสอบว่า จำนวนหลักสูตรที่เข้าอบรมเป็นตัวแปรที่ใช้อธิบายยอดขายรายเดือนที่ระดับนัยสำคัญ .05 ไหม? (n = 10)

(T = 3,392, $t_{8, .025}$ = 2.326, ปฏิเสธ Ho, X_1 เป็นตัวแปรที่อธิบายยอดขายได้)

12.16 จากข้อ 12.11 มี output เพิ่มเติม ดังนี้

SOV	df	SS	MS	F Value
Regression	3	3367.269	1122.423	46.974
Dev. from reg.	6	143.367	23.894	
Total	9	3510.636		

จงทดสอบว่า สมการถดถอยมีนัยสำคัญที่ระดับนัยสำคัญ 5% หรือไม่?

12.17 บริษัทรับเหมาทาสีต้องการทราบความสัมพันธ์ของต้นทุนค่าแรงงาน ($) กับตัวแปรต่าง ๆ ดังนี้ (n = 15)

X_1 พื้นที่เป็นตารางฟุต ที่จะทาสี

X_2 = ความยาวเป็นฟุตของริมขอบพื้นที่

X_3 = จำนวนสี

X_4 = 1 ถ้าต้องใช้นั่งร้าน

= 0 ถ้าไม่ต้องใช้นั่งร้าน

บริษัทได้คอมพิวเตอร์ output ดังนี้

VARIABLE	REG. COEF.	STD. ERROR COEF.	COMPUTED T
1	0.05877	0.00756	7.77188
2	0.05896	0.01196	4.93002
3	-2.12239	7.04050	-0.30145
4	105.89273	25.33568	4.17959
INTERCEPT			
MULTIPLE DETERMINATION		39.04846	
STD. ERROR OF ESTIMATE		0.97296	
ADJUSTED R SQ = 0.96213		47.04616	

บริษัทได้ตรวจดูผลคอมพิวเตอร์ output แล้วได้ตัด X_3 ทิ้ง และได้คอมพิวเตอร์ output ใหม่ ดังนี้

VARABLE	REG. COEF.	STD. ERROR COEF	COMPUTED T
1	0.05866	0.00723	8.10899
2	0.05968	0.01122	5.31887
4	104.23616	23.68846	4.40029
INTERCEPT		32.02061	
MULTIPLE DETERMINATION		0.97271	
STD. ERROR OF ESTIMATE		45.06011	
ADJUSTED R SQ. = 0.96526			

ก. เหตุใดบริษัทจึงต้องทำโปรแกรมใหม่

ข. จงอธิบายผลของ output ครั้งที่ 2 ให้กับผู้จัดการบริษัทซึ่งไม่มีความรู้ทางสถิติ หัวข้อที่ผู้จัดการสนใจ คือ

1. ผลจาก output จะช่วยในการประกวดราคาจ้างเหมาอย่างไร?
2. ผลจาก output จะช่วยจัดตารางแจกจ่ายงานแก่ช่างทาสีได้อย่างไร?
3. ต้องเสียเงินเท่าไรสำหรับการสร้างนั่งร้าน?
4. ต้องเสียค่าใช้จ่ายเท่าใดสำหรับเตรียมงาน 1 โครงการ?
5. ต้องเสียค่าใช้จ่ายเท่าใดสำหรับการทาสี 100 ตารางฟุต
6. ต้องเสียค่าใช้จ่ายเท่าใดสำหรับการแต่งริมขอบพื้นที่ยาว 100 ฟุต