บทที่ 9
การวิเคราะห์การถดถอย

1. แนวคิดเกี่ยวกับการวิเคราะห์การถดถอย

การวิเคราะห์การถดถอยเป็นเครื่องมือที่ใช้สำหรับประเมินความสัมพันธ์ของตัวแปรอิสระตัวเดียวหรือหลายตัวคือ $X_1, X_2, \ldots, X_k$ กับตัวแปรตามชนิดต่อเนื่องที่มีตัวเดียว Y ใช้มากทั้งในสถานการณ์ที่ไม่สามารถควบคุมตัวแปรอิสระ และในการทดลองที่สามารถควบคุมตัวแปรอิสระได้ ตัวอย่างเช่น ความสัมพันธ์ระหว่างความดันเลือดกับอายุ ความสูงกับน้ำหนัก ความเข้มข้นของยากับอัตราการเต้นของหัวใจ ความสัมพันธ์ของตัวแปรสามารถหาได้จากการวิเคราะห์การถดถอยและการวิเคราะห์ความสัมพันธ์

การวิเคราะห์การถดถอยใช้สำหรับการหารูปแบบของความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ เพื่อใช้ในการทำนายหรือประมาณค่าตัวแปรตามที่สนใจศึกษา เมื่อกำหนดตัวแปรอิสระตัวอื่น ๆ มาให้ ส่วนการวิเคราะห์ความสัมพันธ์ใช้สำหรับวัดความสัมพันธ์ระหว่างตัวแปร 2 ตัวดังอธิบายไว้แล้วในบทที่ 10

2. การวิเคราะห์การถดถอยเชิงเส้นตรง (Straight-Line Regression Analysis)

รูปแบบที่ง่ายที่สุดของปัญหาการถดถอยทั่ว ๆ ไป มีตัวแปรตาม Y และตัวแปรอิสระ X เพียงตัวเดียว จากกลุ่มตัวอย่างขนาด n เราสังเกตค่า X และ Y ของแต่ละตัวอย่าง ดังนั้นเราจะมีค่าสังเกตทั้งหมด n คู่ ที่สามารถแทนได้ด้วย $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ ซึ่งเราสามารถนำไปพล็อตกราฟได้ และคำนวณหาสมการถดถอยของกลุ่มตัวอย่าง $y = a + bx$ เพื่อเป็นพื้นฐานในการสรุปอ้างถึงสมการถดถอยของประชากร

$$y_j = \alpha + \beta x_j + e_j$$

เมื่อ y_j คือ ค่าตัวหนึ่งที่ได้จากประชากรย่อยกลุ่มหนึ่งของ Y

α, β คือ สัมประสิทธิ์การถดถอยของประชากร

e_j คือ ความคลาดเคลื่อน

โดยที่ $e_j = y_j - (\alpha + \beta x_j)$

$= y_j - \mu_{y/x}$

e คือ ปริมาณที่ y เบี่ยงเบนไปจากค่าเฉลี่ยของประชากรย่อยของ Y ซึ่งได้จากการสุ่ม

ข้อตกลงเบื้องต้นคือ ประชากรย่อยของ Y แต่ละกลุ่มจะมีการแจกแจงแบบปกติ โดยที่มีความแปรปรวนเท่ากันทุกกลุ่ม จึงทำให้ความคลาดเคลื่อนทั้งหลายของแต่ละประชากรย่อยมีการแจกแจงแบบปกติด้วย และมีความแปรปรวนคงที่เท่ากับ σ^2 และความคลาดเคลื่อนทุกตัวเป็นอิสระกัน

3. ขั้นตอนการวิเคราะห์การถดถอย

ขั้นตอนการวิเคราะห์การถดถอยคือ

(1) เริ่มจากการตรวจสอบว่าตัวแบบเส้นตรงเป็นตัวแบบที่เหมาะสมและเป็นไปตามข้อตกลงเบื้องต้นเกี่ยวกับความเป็นปกติ โดยใช้คำสั่ง Explore เพื่อดูว่าข้อมูลมีการแจกแจงแบบปกติหรือไม่และใช้คำสั่ง Scatter พล็อตกราฟตัวแปรอิสระ X แต่ละตัวกับตัวแปรตาม Y เพื่อดูว่ามีรูปแบบความสัมพันธ์เป็นเชิงเส้นหรือไม่

(2) คำนวณหาสมการถดถอยเชิงเส้นตรงที่เหมาะสมที่สุดกับข้อมูล โดยอาศัยวิธีกำลังสองน้อยที่สุด

(3) การทดสอบความเหมาะสมของตัวแบบที่ได้จากข้อ 2 (Model fit) โดยทดสอบว่าตัวแบบที่ได้สามารถอธิบาย Y ได้ดีอย่างไร โดยพิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์พหุ (R) และค่าสัมประสิทธิ์การกำหนด (R Square) เพื่อดูว่าตัวแปรอิสระทั้งหมดในสมการนั้นมีความสัมพันธ์กับตัวแปรตามมากน้อยเพียงใดและสามารถอธิบายความผันแปรของตัวแปรตาม Y ได้ร้อยละเท่าไร

(4) ตรวจสอบว่าเส้นตรงที่หามาได้เป็นไปตามข้อตกลงเบื้องต้นของการวิเคราะห์การถดถอยที่ว่าตัวแปรอิสระ X_i ทุกตัวเป็นอิสระกันโดยใช้คำสั่ง Covariance Matrix ซึ่งให้ผลลัพธ์แสดงค่า Correlation และค่า Covariance ระหว่างตัวแปรอิสระ X_i และคำสั่ง Collinearity Diagnostics เพื่อทดสอบเกี่ยวกับความเป็นอิสระของตัวแปรอิสระ X_i แต่ละตัว แต่ในการวิเคราะห์การถดถอยอย่างง่ายที่มีตัวแปรอิสระ X เพียงตัวเดียวก็ไม่ต้องตรวจสอบข้อตกลงเบื้องต้นข้อนี้

(5) ตรวจสอบข้อตกลงเบื้องต้นของตัวแบบการถดถอยเชิงเส้นตรงว่าถูกต้องหรือไม่ โดยการวิเคราะห์เศษตกค้าง (residual analysis) และอีกวิธีการหนึ่งที่สามารถทำได้คือการทดสอบที่เรียกว่า lack of fit

(6) ถ้าพบว่าไม่เป็นไปตามข้อตกลงเบื้องต้นของการวิเคราะห์การถดถอยเชิงเส้นตรงต้องหาตัวแบบใหม่ที่เหมาะสมกับข้อมูลอีกครั้ง เช่น เอ็กซโพเนนเชียล แล้วทำซ้ำในขั้นที่ 3 คือหาว่าตัวแบบสามารถอธิบาย Y ได้ดีอย่างไร และทำซ้ำในขั้นที่ 4 เพื่อตัดสินใจว่าต้องหาตัวแบบใหม่ที่เหมาะสมกับข้อมูลอีกครั้งหรือไม่

4. การทดสอบความเหมาะสมของตัวแบบ

เพื่อประเมินสมการถดถอยว่าสามารถอธิบายความสัมพันธ์ระหว่างตัวแปร 2 ตัวได้ดีเพียงใด และสามารถใช้สมการถดถอยในการทำนาย และประมาณค่า Y ได้อย่างมีประสิทธิภาพหรือไม่

สมมติฐานทางสถิติที่ต้องการทดสอบคือ $H_0 : \beta = 0$ คู่กับ $H_1 : \beta \neq 0$

ถ้าในประชากรความสัมพันธ์ระหว่างตัวแปร X และ Y เป็นแบบเส้นตรง β คือความชันของเส้นตรงที่อธิบายความสัมพันธ์นั้น ซึ่งอาจมีค่าเป็นบวก ลบ หรือศูนย์ ถ้า β มีค่าศูนย์หมายความว่า กลุ่มตัวอย่างที่สุ่มมาจากประชากรให้สมการถดถอยที่ไม่สามารถทำนายและประมาณค่า Y ได้ นอกจากนี้ความสัมพันธ์ระหว่าง X และ Y อาจไม่ใช่แบบเส้นตรง

สำหรับกรณีที่สรุปว่าปฏิเสธ $H_0 : \beta = 0$ เราสรุปว่า 1) ความสัมพันธ์ระหว่าง X และ Y เป็นแบบเส้นตรง และสามารถใช้สมการถดถอยของกลุ่มตัวอย่างในการทำนายค่า Y โดยกำหนดค่า X ให้ได้ 2) ตัวแบบเชิงเส้นตรงเหมาะสมกับข้อมูล แต่อาจมีตัวแบบอื่นที่ไม่ใช่เส้นตรงเหมาะสมกับข้อมูลมากกว่าก็ได้

5. การใช้คำสั่ง Regression ตัวแบบเชิงเส้น (Linear Model)

ตัวอย่างเช่น ผู้วิจัยต้องการสมการที่ใช้ทำนายตัวแปรตาม Y (Deep Abdominal AT) ด้วยตัวแปรอิสระ X (Waist Circumference) หน่วยทดลองคือผู้ชายอายุระหว่าง 18 ถึง 42 ปี ที่ไม่มีเชื้อเมทาโบลิกเป็น ทรีทเมนต์ ทำการวัด deep abdominal AT และ waist circumference ของผู้ชาย 109 คน ได้ข้อมูลดังตาราง ผู้วิจัยต้องการทราบว่าตัวแปร waist circumference สามารถทำนายและ

ประมาณค่าตัวแปรตาม deep abdominal AT ได้ดีเพียงใด ซึ่งหาคำตอบได้โดยการวิเคราะห์การถดถอย ข้อมูลอยู่ในแฟ้มข้อมูล Regress1.sav. กำหนดให้ตัวแปร id คือ รหัสของตัวอย่าง ตัวแปร x แทน waist circumference ตัวแปร y แทน deep abdominal AT มีรูปแบบข้อมูลดังนี้

ตารางที่ 9.1 รูปแบบข้อมูลสำหรับการวิเคราะห์การถดถอย

id	x	y	id	x	y	id	x	y	id	x	y
1	74.75	25.72	31	83.5	72.56	61	77.6	57.05	91	97.5	165
2	72.6	25.89	32	88.1	89.31	62	84.9	99.73	92	105.5	152
3	81.8	42.6	33	90.8	78.94	63	79.8	27.96	93	98	181
4	83.95	42.8	34	89.4	83.55	64	108.3	123	94	94.5	80.95
5	74.65	29.84	35	102	127	65	119.6	90.41	95	97	137
6	71.85	21.68	36	94.5	121	66	119.9	106	96	105	125
7	80.9	29.08	37	91	107	67	96.5	144	97	106	241
8	83.4	32.98	38	103	129	68	105.5	121	98	99	134
9	63.5	11.44	39	80	74.02	69	105	97.13	99	91	150
10	73.2	32.22	40	79	55.48	70	107	166	100	102.5	198
11	71.9	28.32	41	83.5	73.13	71	107	87.99	101	106	151
12	75	43.86	42	76	50.5	72	101	154	102	109.1	229
13	73.1	38.21	43	80.5	50.88	73	97	100	103	115	253
14	79	42.48	44	86.5	140	74	100	123	104	101	188
15	77	30.96	45	83	96.54	75	108	217	105	100.1	124
16	68.85	55.78	46	107.1	118	76	100	140	106	93.3	62.2
17	75.95	43.78	47	94.3	107	77	103	109	107	101.8	133
18	74.15	33.41	48	94.5	123	78	104	127	108	107.9	208
19	73.8	43.35	49	79.7	65.92	79	106	112	109	108.5	208
20	75.9	29.31	50	79.3	81.29	80	109	192			
21	76.85	36.6	51	89.8	111	81	103.5	132			
22	80.9	40.25	52	83.8	90.73	82	110	126			
23	79.9	35.43	53	85.2	133	83	110	153			
24	89.2	60.09	54	75.5	41.9	84	112	158			
25	82	45.84	55	78.4	41.71	85	108.5	183			
26	92	70.4	56	78.6	58.16	86	104	184			
27	86.6	83.45	57	87.8	88.85	87	111	121			
28	80.5	84.3	58	86.3	155	88	108.5	159			
29	86	78.89	59	85.5	70.77	89	121	245			
30	82.5	64.75	60	83.7	75.08	90	109	137			

แหล่งที่มา : Jean-Pierre Despres, Ph.D. (อ้างถึงใน Daniel, W.W., 1995)

5.1 การทดสอบว่าตัวแปรอิสระ x มีความสัมพันธ์เชิงเส้นกับตัวแปรตาม y หรือไม่

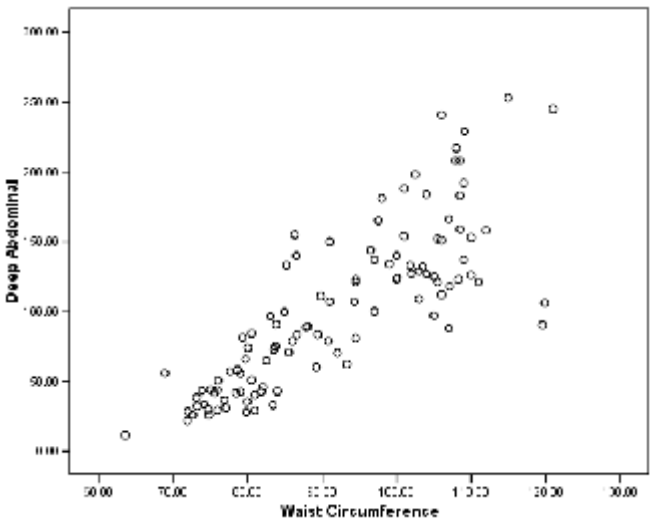
โดยการพล็อต กราฟการกระจายของข้อมูล มีวิธีการดังนี้

1. ไปที่เมนูบาร์ คลิกที่ Graphs , Scatter… จะได้หน้าต่าง Scatterplot

2. ในหน้าต่าง Scatterplot คลิกเลือกที่ Simple แล้วคลิกที่ปุ่ม Define จะได้หน้าต่าง Simple Scatterplot

3. ในหน้าต่าง Simple Scatterplot คลิกเลือกตัวแปร y ให้ย้ายเข้าไปอยู่ในช่อง Y Axis : แล้วคลิกเลือกตัวแปร x ให้ย้ายเข้าไปอยู่ในช่อง X Axis : แล้วคลิกปุ่ม OK จะได้ผลลัพธ์ดังภาพที่ 9.1

Graph



ภาพที่ 9.1

จากภาพการกระจายของข้อมูลจะเห็นว่าตัวแปรอิสระ x มีความสัมพันธ์เชิงเส้นกับตัวแปรตาม y คือ เมื่อ x เพิ่มขึ้นทำให้ y เพิ่มขึ้นด้วย ดังนั้นตัวแบบเชิงเส้น (linear model) เหมาะสมกับข้อมูลชุดนี้

5.2 คำนวณหาสมการถดถอยเชิงเส้นตรง และประเมินคุณภาพของเส้นถดถอยที่ได้

โดยทดสอบว่าตัวแปร x มีความสัมพันธ์กับตัวแปร y หรือไม่ โดยใช้คำสั่ง **Model fit** ทดสอบสมมติฐาน $H_0 : \beta = 0$

1. ไปที่เมนูบาร์ คลิกที่ Analyze , Regression , Linear จะได้หน้าต่าง Linear Regression

2. ในหน้าต่าง Linear Regression ในช่องซ้ายมือ คลิกเลือกตัวแปร y ให้ย้ายเข้าไปอยู่ในช่อง Dependent : แล้วคลิกเลือกตัวแปร x ให้ย้ายเข้าไปอยู่ในช่อง Independent(s) :

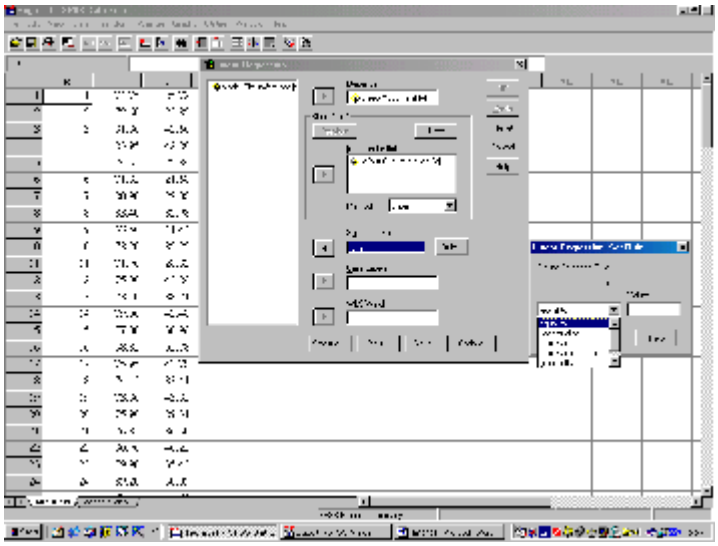
- ที่คำสั่ง Method : เป็นวิธีการเลือกตัวแปรอิสระเข้าสมการถดถอย ให้เลือกวิธีการ Enter ซึ่งโดยปกติโปรแกรมจะกำหนดให้อยู่แล้ว

- แล้วคลิกเลือกตัวแปร id ให้ย้ายเข้าไปอยู่ในช่อง Selection Variable : ถ้าต้องการเลือกเฉพาะบางตัวอย่างมาวิเคราะห์ แล้วกดปุ่ม Rule… จะได้หน้าต่าง Linear Regression : Set Rule เพื่อเลือกกฎสำหรับการเลือกตัวอย่าง ด้วยตัวแปร id ยกตัวอย่างเช่น ในที่นี้เราจะเลือกกฎ less than or equal to โดยการคลิกเลือกที่หัวลูกศร ▼ ดังภาพที่ 9.2 แล้วคลิกเลือกกฎดังกล่าว ต่อจากนั้นพิมพ์เลข 109 ในช่อง Value : แล้วคลิกปุ่ม Continue หน้าต่างนี้จะถูกปิดไป ซึ่งไม่ต้องทำก็ได้เพราะนี่เป็นการเลือกข้อมูลทุกตัวอย่าง

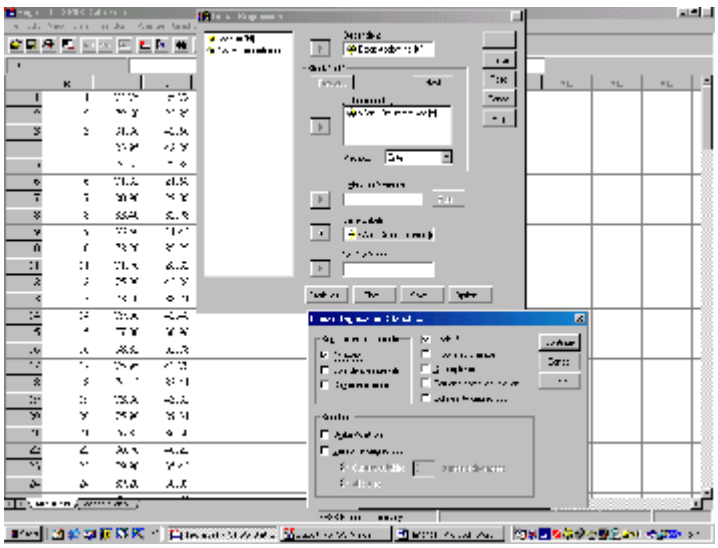
- คลิกที่ตัวแปร x ให้ย้ายเข้าไปอยู่ในช่อง Case Labels : เพื่อต้องการให้ผลลัพธ์แสดง Label ของตัวแปร x ลงในกราฟที่สั่งพล็อต

- คลิกที่ปุ่ม Statistics …จะได้หน้าต่าง Linear Regression : Statistics ดังภาพที่ 9.3

- คลิกที่ปุ่ม Plots… จะได้หน้าต่าง linear Regression : Plots ดังภาพที่ 9.7



ภาพที่ 9.2



ภาพที่ 9.3

3. ในหน้าต่าง Linear Regression : Statistics

- ในกรอบ Regression Coefficient

คลิกเลือกที่ ☐ Estimates เพื่อให้ผลลัพธ์ทำการประมาณค่าสัมประสิทธิ์ถดถอย คือ α และ β และค่าสถิติ t ที่ใช้ในการทดสอบสมมติฐานเกี่ยวกับตัวแปรอิสระ x แต่ละตัวในตัวแบบการถดถอยว่ามีความสัมพันธ์กับตัวแปรตาม y หรือไม่ แต่ในตัวอย่างนี้มีตัวแปรอิสระ x เพียงตัวเดียว ดังนั้นสมมติฐานคือ $H_0 : \beta = 0$ โดยปกติโปรแกรมจะเลือกให้อยู่แล้ว

คลิกเลือกที่ ☐ Confidence intervals เพื่อให้ผลลัพธ์คำนวณหาช่วงความเชื่อมั่นของค่าประมาณของสัมประสิทธิ์การถดถอย

- คลิกเลือกที่ ☐ Model fit เพื่อทำการทดสอบความเหมาะสมของตัวแบบเชิงเส้นที่ได้คือ $y = a + bx$ เหมาะสมกับข้อมูลของตัวอย่าง ด้วยการวิเคราะห์ความแปรปรวน (ANOVA) สถิติทดสอบคือ สถิติ F สมมติฐานที่ต้องการทดสอบคือ $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ คู่กับ $H_1 : \beta_i$ อย่างน้อย 1 ตัวที่ไม่เท่ากับ 0 แต่ในตัวอย่างนี้มีตัวแปรอิสระ x เพียงตัวเดียวดังนั้นสมมติฐานที่ต้องการทดสอบคือ $H_0 : \beta = 0$

การทดสอบความเหมาะสมของตัวแบบการถดถอยที่ประมาณได้คือ การพิจารณาว่าสมการถดถอยที่หามาได้สามารถอธิบายความสัมพันธ์ระหว่าง 2 ตัวแปรได้ดีเพียงใด พิจารณาจากสัดส่วนของ SSR/SST ของกลุ่มตัวอย่าง ผลลัพธ์ที่ได้เรียกว่า สัมประสิทธิ์การกำหนด (coefficient of determination) หรือ R^2 ซึ่งแสดงอยู่ในผลลัพธ์นี้ด้วย

นอกจากนี้ในผลลัพธ์ยังแสดงค่าความคลาดเคลื่อนมาตรฐานจากการประมาณค่า (standard error of the estimate) หรือรากที่สองของ MSE

แล้วคลิกปุ่ม Continue หน้าต่างนี้จะถูกปิดไป

4. ในหน้าต่าง Linear Regression คลิกที่ปุ่ม OK จะได้ผลลัพธ์ดังภาพที่ 9.4

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Waist Circumference[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Deep Abdominal

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.819[a]	.670	.667	33.06493

a. Predictors: (Constant), Waist Circumference

b. Dependent Variable: Deep Abdominal

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	237548.516	1	237548.516	217.279	.000[a]
	Residual	116981.986	107	1093.290		
	Total	354530.502	108			

a. Predictors: (Constant), Waist Circumference

b. Dependent Variable: Deep Abdominal

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-215.981	21.796		-9.909	.000	-259.190	-172.773
	Waist Circumference	3.459	.235	.819	14.74	.000	2.994	3.924

a. Dependent Variable: Deep Abdominal

Casewise Diagnostics [a]

Case Number	Waist Circumference	Std. Residual	Deep Abdominal
65	119.60	-3.245	90.41

a. Dependent Variable: Deep Abdominal

Residuals Statistics [a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	3.6561	202.5405	101.8940	46.89908	109
Residual	-107.28809	90.34239	.00000	32.91150	109
Std. Predicted Value	-2.095	2.146	.000	1.000	109
Std. Residual	-3.245	2.732	.000	.995	109

a. Dependent Variable: Deep Abdominal

ภาพที่ 9.4

จากภาพผลลัพธ์การคำนวณหาสมการถดถอยเชิงเส้นตรง ดูจากในตาราง Coefficients ในช่อง Unstandardized Coefficients ดูที่ค่า B และ Std.Error ของ Constant และตัวแปร Waist circumference (x) คือค่าประมาณของสัมประสิทธิ์การถดถอย ได้ค่าประมาณ a = -215.981 และ b = 3.459 มีค่าความคลาดเคลื่อนมาตรฐานเท่ากับ 21.796 และ .235 ตามลำดับ

95% ช่วงความเชื่อมั่นของค่าประมาณ α คือ (-259.190 , -172.773)

95% ช่วงความเชื่อมั่นของค่าประมาณ β คือ (2.994 , 3.924)

สรุปได้สมการถดถอยของตัวอย่างคือ

$$\hat{y} = -215.981 + 3.459x$$

a เป็นค่าลบ หมายความว่า เส้นตรงตัดแกน Y ที่ต่ำกว่าจุดกำเนิด (origin)

b เป็นค่าบวก หมายความว่า เส้นตรงลากจากมุมซ้ายด้านล่างของกราฟไปมุมขวาของด้านบน และ

เมื่อ x เพิ่มขึ้น 1 หน่วย ทำให้ y เพิ่มขึ้น 3.459 หน่วย

$\hat{y}$ คือ ค่าของ y ที่คำนวณได้จากสมการถดถอย

สำหรับการทดสอบความเหมาะสมของเส้นถดถอยของตัวอย่างนี้ ดูจากตาราง ANOVA ดูที่ค่าสถิติ F เท่ากับ 217.279 และ Sig. เท่ากับ .000 สรุปว่าปฏิเสธ $H_0 : \beta = 0$ นั่นคือตัวแปรอิสระ x สามารถอธิบายความผันแปรของตัวแปรตาม y ได้อย่างมีนัยสำคัญทางสถิติ (Sig < .05) จากตาราง Model Summary ดูที่ค่า R Square เท่ากับ .670 หมายความว่า x สามารถอธิบาย

ความผันแปรของ y ได้ร้อยละ 67.0 นอกจากนี้ความคลาดเคลื่อนมาตรฐานของการประมาณค่าดูที่ค่า Std. Error of the Estimate เท่ากับ 33.06493

ในตาราง Residuals Statistics แสดงค่าน้อยที่สุด (Minimum) ค่ามากที่สุด (Maximum) ค่าเฉลี่ย (Mean) ค่าเบี่ยงเบนมาตรฐาน (Std. Deviation) และจำนวนข้อมูล (N) ของค่าทำนายตัวแปร y (Predicted Value) ซึ่งคำนวณได้จากตัวแบบการถดถอย ค่าความแตกต่างระหว่างค่าทำนายตัวแปร y กับค่าสังเกตที่ได้จากกลุ่มตัวอย่างเรียกว่า เศษตกค้าง (Residual) ตลอดจนค่าเบี่ยงเบนมาตรฐานของค่าทำนายตัวแปร y (Std. Predicted Value) และค่าเบี่ยงเบนมาตรฐานของค่าเศษตกค้าง (Std. Residual)

นอกจากนี้ในตาราง Casewise Diagnostics ได้แสดงข้อมูลตัวที่เป็น outliers ที่มีค่าเศษตกค้างมาตรฐาน (Std. Residual) เท่ากับ -3.245 คือข้อมูลตัวที่ 65 ที่มีค่าของตัวแปร x คือ 119.6 และมีค่าของตัวแปร y คือ 90.41 ซึ่งข้อมูลตัวนี้มีผลทำให้สมการถดถอยเบี่ยงเบนไป และทำให้การทำนายค่า y มีความผิดพลาดได้ ดังนั้นถ้าตัดข้อมูลตัวนี้ออกไปแล้วทำการวิเคราะห์การถดถอยใหม่อีกครั้งจะได้ผลลัพธ์ใหม่ที่แตกต่างไปจากเดิมดังภาพที่ 9.5

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Waist Circumference[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Deep Abdominal

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.839[a]	.704	.701	31.45726

a. Predictors: (Constant), Waist Circumference

b. Dependent Variable: Deep Abdominal

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	249504.138	1	249504.138	252.137	.000[a]
	Residual	104893.260	106	989.559		
	Total	354397.398	107			

a. Predictors: (Constant), Waist Circumference

b. Dependent Variable: Deep Abdominal

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-229.393	21.089		-10.878	.000	-271.203	-187.583
	Waist Circumference	3.616	.228	.839	15.879	.000	3.165	4.068

a. Dependent Variable: Deep Abdominal

Casewise Diagnostics [a]

Case Number	Waist Circumference	Std. Residual	Deep Abdominal
65	119.90	-3.121	106.00

a. Dependent Variable: Deep Abdominal

Residuals Statistics [a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	.2256	208.1478	102.0004	48.28886	108
Residual	-98.17019	87.09274	.00000	31.30992	108
Std. Predicted Value	-2.108	2.198	.000	1.000	108
Std. Residual	-3.121	2.769	.000	.995	108

a. Dependent Variable: Deep Abdominal

ภาพที่ 9.5

ผลการวิเคราะห์การถดถอยที่ตัด outlier คือ ข้อมูลตัวที่ 65 ออกไป ทำให้ได้สมการถดถอยของตัวอย่างคือ

$$\hat{y} = -229.39 + 3.616x$$

และผลการทดสอบความเหมาะสมของเส้นถดถอยของตัวอย่างนี้ ดูที่ค่า R Square เท่ากับ .704 หมายความว่า x สามารถอธิบายความผันแปรของ y ได้ร้อยละ 70.4 ซึ่งสูงขึ้น และค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่าเท่ากับ 31.45726 ซึ่งลดลง

และยังมี outlier คือข้อมูลตัวที่ 65 ที่มีค่าของตัวแปร x คือ 119.9 และมีค่าของตัวแปร y คือ 106.0 โดยมีค่าเศษตกค้างมาตรฐาน เท่ากับ -3.121 ถ้าตัดข้อมูลตัวนี้ออกไปแล้วทำการวิเคราะห์การถดถอยใหม่อีกครั้งจะได้ผลลัพธ์ที่แตกต่างไปอีกดังภาพที่ 9.6

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Waist Circumference[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Deep Abdominal

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.856[a]	.733	.730	30.03758

a. Predictors: (Constant), Waist Circumference

b. Dependent Variable: Deep Abdominal

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	259644.335	1	259644.335	287.772	.000[a]
	Residual	94736.917	105	902.256		
	Total	354381.251	106			

a. Predictors: (Constant), Waist Circumference

b. Dependent Variable: Deep Abdominal

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-242.474	20.511		-11.822	.000	-283.143	-201.805
	Waist Circumference	3.769	.222	.856	16.964	.000	3.329	4.210

a. Dependent Variable: Deep Abdominal

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-3.1281	213.6026	101.963	49.49217	107
Residual	-72.84343	83.93580	.00000	29.89556	107
Std. Predicted Value	-2.123	2.256	.000	1.000	107
Std. Residual	-2.425	2.794	.000	.995	107

a. Dependent Variable: Deep Abdominal

ภาพที่ 9.6

ผลการวิเคราะห์การถดถอยที่ตัด outlier คือข้อมูลตัวที่ 65 ออกไปทำให้ได้สมการถดถอยของตัวอย่างคือ

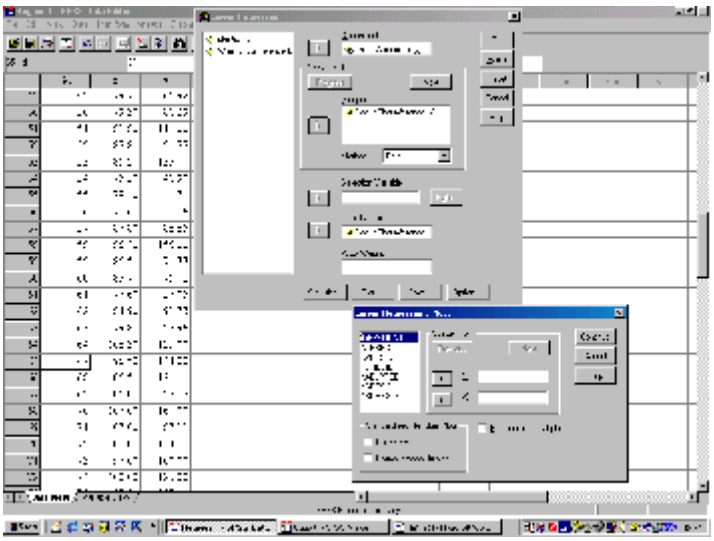
$$\hat{y} = -242.474 + 3.769x$$

และผลการทดสอบความเหมาะสมของเส้นถดถอยของตัวอย่างนี้ ดูที่ค่า R Square เท่ากับ .733 ซึ่งสูงขึ้น และค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่าเท่ากับ 30.03758 ซึ่งลดลง หมายความว่า เมื่อตัดข้อมูลตัวที่เป็น outlier แล้วทำให้สมการถดถอยมีความเหมาะสมกับข้อมูลมากขึ้น

5.3 ตรวจสอบข้อตกลงเบื้องต้นของตัวแบบการถดถอยเชิงเส้นตรง

เพื่อตรวจสอบว่าสมการเส้นตรงที่หามาได้เป็นไปตามข้อตกลงเบื้องต้นหรือไม่ ข้อตกลงเบื้องต้นเกี่ยวกับความเป็นปกติ และความคงที่ของความแปรปรวนของความคลาดเคลื่อน โดยใช้วิธีการพล็อตกราฟของความคลาดเคลื่อน มีวิธีการตรวจสอบดังนี้คือ ทำต่อจากขั้นตอนที่ 2 ในหัวข้อ 5.2

1. ในหน้าต่าง Linear Regression คลิกที่ปุ่ม Plots… จะได้หน้าต่าง Linear Regression : Plots ดังภาพที่ 9.7



ดังภาพที่ 9.7

2. ในหน้าต่าง Linear Regression : Plots

คลิกที่ตัวแปร ZRESID (Standardized Residuals) ให้ย้ายไปอยู่ในช่อง Y : แล้วคลิกที่ตัวแปร ZPRED (Standardized Predicted Values) ให้ย้ายไปอยู่ในช่อง X :

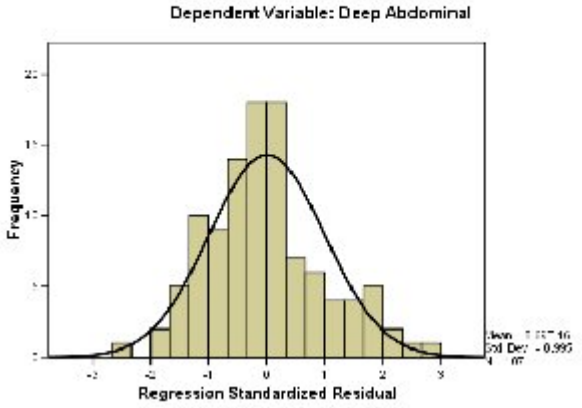
ในกรอบ Standardized Residual Plots คลิกที่ ☐ Histogram และ ☐ Normal probability plot เพื่อตรวจสอบว่าค่าความคลาดเคลื่อนมีการแจกแจงแบบปกติหรือไม่

แล้วคลิกปุ่ม Continue หน้าต่างนี้จะถูกปิดไป

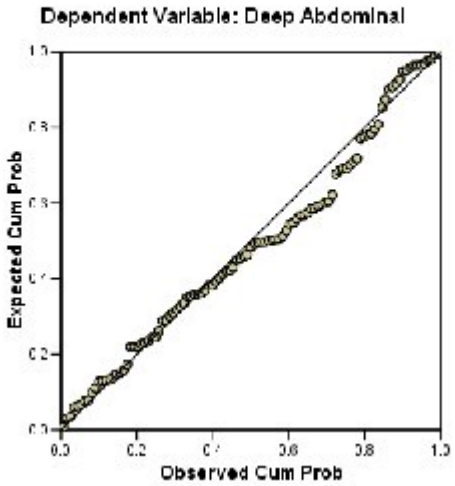
3. ในหน้าต่าง Linear Regression คลิกปุ่ม OK จะได้ผลลัพธ์ดังภาพที่ 9.8

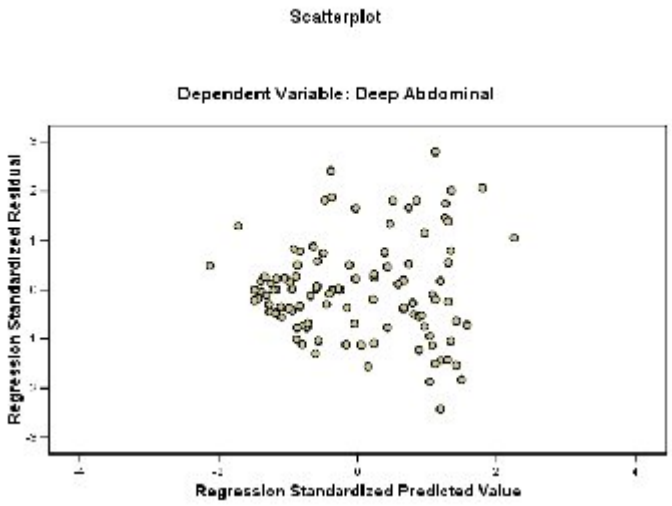
Charts

Histogram



Normal P-P Plot of Regression Standardized Residual





ภาพที่ 9.8

จากภาพผลการพล็อตกราฟ Histogram และ Normal P-P Plot ของความคลาดเคลื่อนมาตรฐาน มีแนวโน้มเป็นเส้นตรง แสดงว่าเข้าใกล้การแจกแจงปกติ และกราฟการกระจาย Scatterplot ระหว่างค่า Regression Standardized Residual กับค่า Regression Standardized Predicted Value มีการกระจายของความคลาดเคลื่อนแบบไม่มีรูปแบบแสดงว่าความแปรปรวนของความคลาดเคลื่อนคงที่

แบบฝึกหัดบทที่ 9

1. นักสัตววิทยาได้ทำการศึกษาเกี่ยวกับพฤติกรรมการกินอาหารของปลาน้ำจืดชนิดหนึ่ง โดยสังเกตเกี่ยวกับจำนวนครั้งของการแสดงพฤติกรรมก้าวร้าวของปลา 2 ตัว ขณะกินอาหารที่ก้นตู้กระจกเลี้ยงปลา เป็นเวลา 10 นาที หลังจากการให้อาหารผู้วิจัยสังเกตพฤติกรรมปลาเป็นเวลา 9 สัปดาห์ ในแต่ละสัปดาห์จะเก็บข้อมูลเกี่ยวกับจำนวนครั้งของพฤติกรรมที่ก้าวร้าวของปลาและอายุของปลาเป็นวัน ได้ข้อมูลดังตาราง อยากทราบว่าอายุของปลาสามารถทำนายพฤติกรรมก้าวร้าวของปลาได้ดีเพียงใด

ตาราง ข้อมูลจำนวนครั้งของพฤติกรรมและอายุของปลา

สัปดาห์	จำนวนครั้งของพฤติกรรม (y)	อายุปลา (วัน) (x)
1	85	120
2	63	136
3	34	150
4	39	155
5	58	162
6	35	169
7	57	178
8	12	184
9	15	190

แหล่งที่มา : Shand, J., et al. “Variability in the location of the retinal ganglion cell area centralis is correlated with ontogenetic changes in feeding behavior in the Blackbream, Acanthopagrus ‘butcher’“ Brain and Behavior, Vol. 55, No.4, Apr. 2000 (Figure H). อ้างถึงใน Mendenhall, W. and Sincich, T., 2003.

ก. จงตรวจสอบว่าตัวแบบเส้นตรงเป็นตัวแบบที่เหมาะสมกับข้อมูลและเป็นไปตามข้อตกลงเบื้องต้นของตัวแบบการถดถอยเชิงเส้นตรงหรือไม่

ข. จงใช้ข้อมูลในตารางเพื่อหาสมการถดถอยของกลุ่มตัวอย่างปลาโดยอาศัยวิธีกำลังสองน้อยที่สุด และแปลความหมายเกี่ยวกับสัมประสิทธิ์การถดถอยของสมการที่ได้จากข้อ ก.

ค. จงทดสอบความเหมาะสมของเส้นถดถอยเชิงเส้นตรงที่ได้จากข้อ ข.

2. ในการศึกษาเกี่ยวกับคุณภาพด้านรสชาดของน้ำส้มคั้นในกล่องสำเร็จรูป โดยพิจารณาจากดัชนีความหวาน ผู้วิจัยอยากทราบว่ามีความสัมพันธ์ระหว่างดัชนีความหวานกับปริมาณเพคติน หน่วยเป็น ppm (parts per million) ในน้ำส้มคั้นหรือไม่ จึงทำการเก็บข้อมูลจากผลิตภัณฑ์ของโรงงานผลิตน้ำส้มคั้นออกจำหน่าย 24 โรงงาน เพื่อใช้การวิเคราะห์การถดถอยอย่างง่ายในการทำนายความหวานจากปริมาณของเพคติน ข้อมูลอยู่ในตารางต่อไปนี้

ตาราง ข้อมูลดัชนีความหวาน และปริมาณของเพคติน (ppm) ในน้ำส้มคั้นในกล่องสำเร็จรูป

ผลิตภัณฑ์	ดัชนีความหวาน	เพคติน (ppm)
1	5.2	220
2	5.5	227
3	6.0	259
4	5.9	210
5	5.8	224
6	6.0	215
7	5.8	231
8	5.6	268
9	5.6	239
10	5.9	212
11	5.4	410
12	5.6	256
13	5.8	306
14	5.5	259
15	5.3	284
16	5.3	383
17	5.7	271
18	5.5	264
19	5.7	227
20	5.3	263
21	5.9	232
22	5.8	220
23	5.8	246
24	5.9	241

แหล่งที่มา : Mendenhall, W. and Sincich, T. (2003)

ก. จงหาสมการถดถอยของดัชนีความหวานที่ถดถอยบนปริมาณของเพคติน

ข. จงแปลความหมายของสัมประสิทธิ์การถดถอยของสมการที่ได้จากข้อ ก.

ค. จงทำนายดัชนีความหวานของน้ำส้มคั้น เมื่อปริมาณเพคตินเท่ากับ 350 ppm

3. การออกแบบโรงเก็บสินค้าแบบทันสมัยที่ใช้คอมพิวเตอร์และรถขับเคลื่อนอัตโนมัติในการจัดวางสินค้า ต้องคำนึงถึงการป้องกันเกี่ยวกับความแออัดของรถขับเคลื่อนอัตโนมัติและเวลาในการจัดวางสินค้าที่ดีที่สุด จากการศึกษาเกี่ยวกับการออกแบบโรงเก็บสินค้าแบบอัตโนมัติใน Journal of Engineering for Industry (Aug. 1993 อ้างถึงใน Mendenhall, W. and Sincich, T., 2003) โดยมีเงื่อนไขของการออกแบบคือ รถขับเคลื่อนอัตโนมัติไม่มีการขวางทางกันขณะที่วิ่งในโรงเก็บสินค้าคือไม่มีการแออัด วิธีการศึกษาทำโดยใช้คอมพิวเตอร์ในการสร้างสถานการณ์จำลองในโรงเก็บสินค้าที่ออกแบบ โดยให้มีตัวแปรคือ จำนวนรถขับเคลื่อนอัตโนมัติ และเก็บข้อมูลเวลาที่เกิดความแออัดของรถขับเคลื่อนอัตโนมัติที่ขวางทางกันและกัน ได้ข้อมูลดังตาราง ผู้วิจัยอยากทราบรูปแบบของความสัมพันธ์ระหว่างเวลาที่เกิดความแออัดกับจำนวนรถขับเคลื่อนอัตโนมัติ

จำนวนรถ	เวลาความแออัด (นาที)	จำนวนรถ	เวลาความแออัด (นาที)
1	0	9	.02
2	0	10	.04
3	.02	11	.04
4	.01	12	.04
5	.01	13	.03
6	.01	14	.04
7	.03	15	.05
8	.03		

แหล่งที่มา : Pandit, R., and U.S. Paleker. “Response time considerations for optimal warehouse layout design.” Jounal of Engineering for Industry, Transactions of the ASME, Vol. 115, Aug. 1993, p.326 (Table 2). อ้างถึงใน Mendenhall, W. and Sincich, T., 2003.

ก. จงสร้างกราฟการกระจายของข้อมูลในตาราง

ข. จงหาสมการถดถอยของเวลาความแออัดที่ถดถอยบนจำนวนรถและแปลความหมาย

ค. จงทดสอบความเหมาะสมของเส้นถดถอยเชิงเส้นตรงที่ได้จากข้อ ข.

4. การศึกษานก 2 ชนิดที่กินสัตว์อื่นเป็นอาหาร คือ ชนิด A และ B ซึ่งมีการต่อสู้แย่งชิงโพรงต้นไม้ในการทำรัง ในระหว่างช่วงผสมพันธ์บนเกาะหนึ่งของสวีเดน มักจะพบนกชนิด A ตายในรังที่ถูกยึดครองโดยนกชนิด B ผู้วิจัยอยากทราบว่าจำนวนการตายของนกชนิด A มีความสัมพันธ์กับความรุนแรงของการต่อสู้แย่งชิงโพรงต้นไม้ในการทำรังระหว่างนก 2 ชนิดนี้หรือไม่ (The Conder , May 1995 อ้างถึงใน Mendenhall, W. and Sincich, T., 2003) ผู้วิจัยทำการเก็บข้อมูลในพื้นที่ 14 แห่งบนเกาะ เก็บข้อมูลเกี่ยวกับจำนวนนกชนิด A ที่ถูกฆ่าซึ่งพบในรังที่นกชนิด B เข้ายึดครองคิดเป็นเปอร์เซ็นต์ในแต่ละพื้นที่ทั้ง 14 แห่ง ได้ข้อมูลดังตาราง

ตาราง ข้อมูลจำนวนนกชนิด A ที่ถูกฆ่าและจำนวนรังที่ถูกยึดครองโดยนกชนิด B ในพื้นที่ 14 แห่ง

พื้นที่	จำนวนนกชนิด A ที่ถูกฆ่า	จำนวนรังที่ถูกยึดครองโดยนกชนิด B (%)
1	0	24
2	0	33
3	0	34
4	0	43
5	0	50
6	1	35
7	1	35
8	1	38
9	1	40
10	2	31
11	2	43
12	3	55
13	4	57
14	5	64

แหล่งที่มา : merila, J., and Wiggins, D.A. “Interspecific competition for nest holes causes adult mortality in the collard flycatcher.” The Condor, Vol. 97, No.2, May 1995, p.449 (Figure 2). Cooper Ornithological Society. อ้างถึงใน Mendenhall, W. and Sincich, T., 2003.

ก. จงสร้างกราฟการกระจายของข้อมูลในตารางและดูว่ามีความสัมพันธ์เชิงเส้นตรงระหว่างจำนวนนกชนิด A ที่ถูกฆ่า กับจำนวนรังที่ถูกยึดครองโดยนกชนิด B หรือไม่

ข. จงหาสมการถดถอยของจำนวนนกชนิด A ที่ถูกฆ่า ที่ถดถอยบนจำนวนรังที่ถูกยึดครองโดยนกชนิด B และแปลความหมาย

5. จากข้อมูลสถิติ (ตุลาคม – ธันวาคม 1999) ที่รายงานเกี่ยวกับตัวเลขค่ารักษาพยาบาลเฉลี่ยของโรงพยาบาลและจำนวนวันเฉลี่ยที่คนไข้อยู่ในโรงพยาบาลในรัฐต่าง ๆ ของสหรัฐอเมริกาที่เป็นกลุ่มตัวอย่างจำนวน 12 รัฐ ได้ข้อมูลดังตาราง

ตาราง ข้อมูลค่ารักษาพยาบาลเฉลี่ยและจำนวนวันเฉลี่ยที่คนไข้อยู่ในโรงพยาบาลในรัฐต่าง ๆ ของสหรัฐอเมริกา 12 รัฐ

รัฐ	ค่ารักษาพยาบาลเฉลี่ย (ดอลลาร์)	จำนวนวันเฉลี่ยที่คนไข้อยู่ในโรงพยาบาล (วัน)
1	11,680	3.64
2	11,630	4.20
3	9,850	3.84
4	9,950	3.11
5	8,490	3.86
6	9,020	3.54
7	13,820	4.08
8	8,440	3.57
9	8,790	3.80
10	10,400	3.52
11	12,860	3.77
12	16,740	3.78

แหล่งที่มา : Statistical Bulletin, Vol. 80, No. 4, Oct – Dec. 1999. p.13. อ้างถึงใน Mendenhall, W. and Sincich, T., 2003

ก. จงตรวจสอบว่าตัวแบบเส้นตรงเป็นตัวแบบที่เหมาะสมกับข้อมูล และเป็นไปตามข้อตกลงเบื้องต้นของตัวแบบการถดถอยเชิงเส้นตรงหรือไม่

ข. จงสร้างกราฟการกระจายของข้อมูล

ค. จงหาสมการถดถอยของค่ารักษาพยาบาลเฉลี่ยที่ถดถอยบนจำนวนวันเฉลี่ยที่คนไข้อยู่ในโรงพยาบาล และแปลความหมาย

ง. จงทดสอบความเหมาะสมของเส้นถดถอยเชิงเส้นตรงที่ได้จากข้อ ค.

จ. จงทำนายค่ารักษาพยาบาลเฉลี่ยเมื่อคนไข้อยู่ในโรงพยาบาล 4 วัน