บทที่ 3
การถดถอยเชิงเส้นอย่างง่ายและสหสัมพันธ์

การศึกษาความสัมพันธ์ระหว่างตัวแปร 2 ตัวแปร เช่น ความสัมพันธ์ระหว่างความดันเลือดกับอายุ ความสูงกับน้ำหนัก ความเข้มข้นของยากับอัตราการเต้นของหัวใจ หรือรายได้ของครอบครัวกับค่ารักษาพยาบาล ความสัมพันธ์ของตัวแปรสามารถหาได้จากการวิเคราะห์การถดถอย และการวิเคราะห์ความสัมพันธ์

การวิเคราะห์การถดถอย ใช้สำหรับการหารูปแบบความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการทำนาย หรือประมาณค่าตัวแปรตัวหนึ่ง เมื่อกำหนดตัวแปรตัวอื่น ๆ มาให้

การวิเคราะห์สหสัมพันธ์ ใช้สำหรับวัดความสัมพันธ์ระหว่างตัวแปร

สำหรับบทนี้มีขอบเขตในการวิเคราะห์การถดถอยและสหสัมพันธ์ของ 2 ตัวแปรเท่านั้น

1. ตัวแบบการถดถอย (The Regression Model)

ตัวแบบการถดถอยจะแสดงรูปแบบความสัมพันธ์ของตัวแปรที่สนใจศึกษาซึ่งเก็บข้อมูลมาจากกลุ่มตัวอย่าง เพื่อใช้ในการตัดสินใจเกี่ยวกับประชากร

ข้อตกลงเบื้องต้นสำหรับตัวแบบการถดถอยเชิงเส้นอย่างง่ายที่มี 2 ตัวแปร คือตัวแปร X และตัวแปร Y โดยทั่วไปตัวแปร X จะหมายถึงตัวแปรอิสระ และตัวแปร Y จะหมายถึงตัวแปรตาม เราจะเรียกการถดถอยของ Y บน X ซึ่งมีข้อตกลงเบื้องต้น ดังนี้

ข้อตกลงเบื้องต้นคือ

1. ขอบเขตการศึกษาตัวแบบการถดถอยของบทนี้ จะศึกษาเฉพาะกรณีตัวแปรอิสระ X ที่ถูกกำหนดโดยผู้วิจัย เรียกว่าตัวแปรกำหนด ไม่ได้ศึกษากรณีที่ตัวแปรอิสระ X เป็น ตัวแปรสุ่ม
2. การวัดตัวแปร X ถือว่าไม่มีความคลาดเคลื่อน
3. ค่า Y มีการแจกแจงแบบปกติ (normal) สำหรับแต่ละค่าของ X มีประชากรย่อยของค่า Y ซึ่งมีการแจกแจงแบบปกติ
4. ความแปรปรวนของประชากรย่อยของค่า Y ทุกกลุ่ม มีค่าเท่ากัน (equal variance)
5. ค่าเฉลี่ยของประชากรย่อยของ Y ทุกกลุ่ม อยู่บนเส้นตรงเส้นเดียวกันเป็นข้อตกลงเบื้องต้นของเชิงเส้นตรง (linear) เขียนเป็นสัญลักษณ์ได้ คือ

$$\mu_{y|x} = \alpha + \beta x$$

เมื่อ $\mu_{y|x}$ คือ ค่าเฉลี่ยของประชากรย่อยของค่า Y สำหรับค่า X ค่าหนึ่ง

α, β คือ สัมประสิทธิ์การถดถอยของประชากร โดยที่ α คือ จุดตัดแกน Y ของเส้นตรง และ β คือความชันของเส้นตรงที่ลากผ่านค่าเฉลี่ยทุกค่าสำหรับค่า X ค่าต่าง ๆ

6. ค่าของ Y มีความเป็นอิสระ (independent) หมายถึง ค่าของ Y สำหรับ X ค่าหนึ่งจะไม่ขึ้นกับ ค่า Y สำหรับค่า X ค่าอื่น ๆ

ตัวแบบการถดถอย คือ

$$y = \alpha + \beta x + e$$

เมื่อ y คือ ค่าหนึ่งที่ได้จากประชากรย่อยกลุ่มหนึ่งของ Y

α, β คือ สัมประสิทธิ์การถดถอยของประชากร

e คือ ความคลาดเคลื่อน

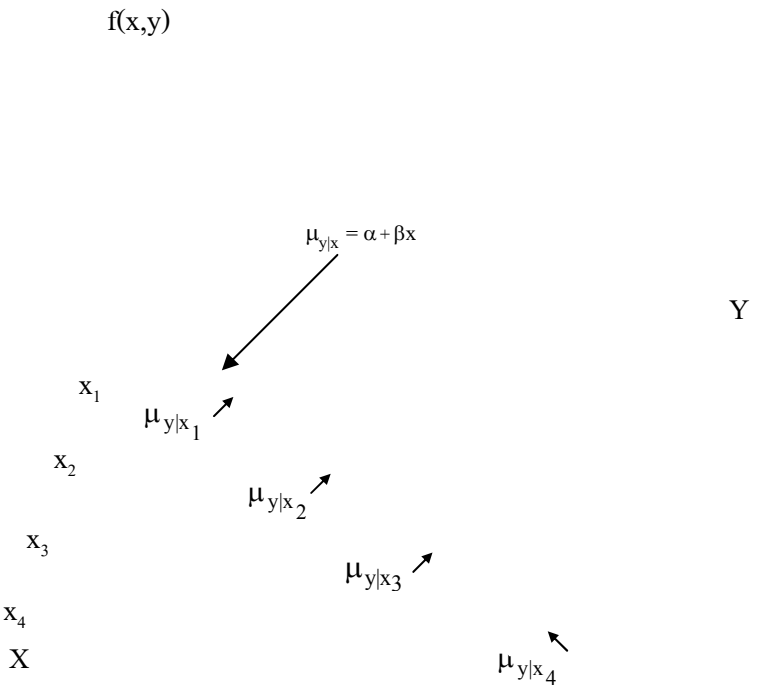
โดยที่

$$e = y - (\alpha + \beta x)$$

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

e คือปริมาณที่ y เบี่ยงเบนไปจากค่าเฉลี่ยของประชากรย่อยของ Y ซึ่งได้จากการสุ่ม ถ้าเป็นไปตามข้อตกลงเบื้องต้นจะได้ว่าประชากรย่อยของ Y แต่ละกลุ่มจะมีการแจกแจงแบบปกติ โดยที่มีความแปรปรวนเท่ากันทุกกลุ่ม จึงทำให้ความคลาดเคลื่อนทั้งหลายของแต่ละประชากรย่อยมีการแจกแจงแบบปกติด้วย และมีความแปรปรวนเท่ากันทุกกลุ่มเท่ากับ σ^2

ตัวแบบการถดถอยสามารถแสดงเป็นกราฟได้ดังนี้


f(x,y)
$\mu_{y|x} = \alpha + \beta x$
Y
x_1
$\mu_{y|x_1}$
x_2
$\mu_{y|x_2}$
x_3
$\mu_{y|x_3}$
x_4
X
$\mu_{y|x_4}$

ภาพ 1 ตัวแบบการถดถอยเชิงเส้นอย่างง่าย

2. สมการถดถอยอย่างง่าย

สำหรับสมการถดถอยอย่างง่าย มีวัตถุประสงค์เพื่อหาสมการถดถอยของประชากร ซึ่งอธิบายความสัมพันธ์ระหว่างตัวแปรตาม Y และตัวแปรต้น X สามารถทำได้โดยการสุ่มตัวอย่างจากประชากรที่สนใจ และคำนวณหาสมการการถดถอยของกลุ่มตัวอย่าง เพื่อเป็นพื้นฐานในการสรุปอ้างถึงสมการถดถอยของประชากร

3. ขั้นตอนการวิเคราะห์การถดถอย

1. พิจารณาข้อมูลที่จะนำมาวิเคราะห์ เป็นไปตามข้อตกลงเบื้องต้นหรือไม่
2. หาสมการเส้นตรงของข้อมูล $\mu_{y|x_3}$
3. ประเมินสมการถดถอยที่ได้จากข้อ 2
4. ถ้าข้อมูลเหมาะสมกับตัวแบบการถดถอยเชิงเส้นตรง จึงใช้สมการถดถอยนี้ในการประมาณค่าเฉลี่ยของ Y และใช้ทำนายค่า Y เมื่อกำหนดค่า X

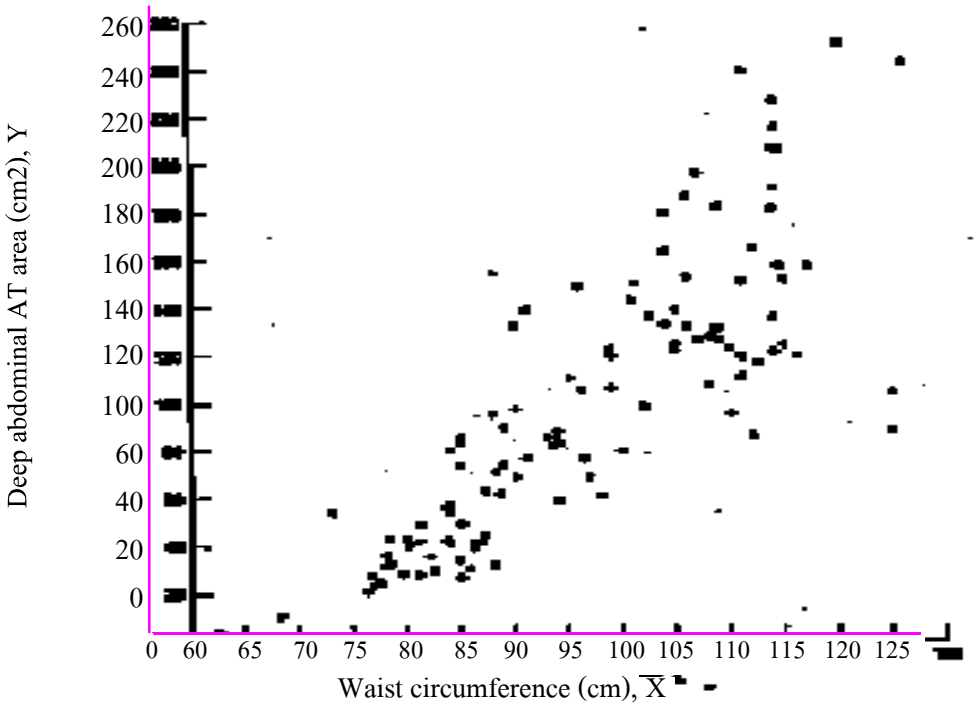
ตัวอย่างที่ 1 ผู้วิจัยต้องการหาสมการที่ใช้ทำนายตัวแปร Y (Deep Abdominal AT) ด้วยตัวแปร X (Waist Circumference) หน่วยทดลองคือผู้ชายอายุระหว่าง 18 ถึง 42 ปี ที่ไม่มี metabolic disease เป็นทรีทเมนต์ ทำการวัด deep abdominal AT และ waist circumference ของผู้ชาย 109 คน ได้ข้อมูลดังตารางผู้วิจัยต้องการทราบว่า ตัวแปร waist circumference สามารถทำนายและประมาณค่าตัวแปร deep abdominal AT ได้ดีเพียงใด ซึ่งหาคำตอบได้โดยการวิเคราะห์การถดถอย

ตาราง 3.1 ข้อมูล Waist Circumference (cm), X, and Deep Abdominal AT, Y, ของผู้ชาย 109 คน

คนที่	x	y	คนที่	x	y	คนที่	x	y
1.	74.75	25.72	38.	103.00	129.00	75.	108.00	217.00
2.	72.60	25.89	39.	80.00	74.02	76.	100.00	140.00
3.	81.80	42.60	40.	79.00	55.48	77.	103.00	109.00
4.	83.95	42.80	41.	83.50	73.13	78.	104.00	127.00
5.	74.65	29.84	42.	76.00	50.50	79.	106.00	112.00
6.	71.85	21.68	43.	80.50	50.88	80.	109.00	192.00
7.	80.90	29.08	44.	86.50	140.00	81.	103.50	132.00
8.	83.40	32.98	45.	83.00	96.54	82.	110.00	126.00
9.	63.50	11.44	46.	107.10	118.00	83.	110.00	153.00
10.	73.20	32.22	47.	94.30	107.00	84.	112.00	158.00
11.	71.90	28.32	48.	94.50	123.00	85.	108.50	183.00
12.	75.00	43.86	49.	79.70	65.92	86.	104.00	184.00
13.	73.10	38.21	50.	79.30	81.29	87.	111.00	121.00
14.	79.00	42.48	51.	89.80	111.00	88.	108.50	159.00
15.	77.00	30.96	52.	83.80	90.73	89.	121.00	245.00
16.	68.85	55.78	53.	85.20	133.00	90.	109.00	137.00
17.	75.95	43.78	54.	75.50	41.90	91.	97.50	165.00
18.	74.15	33.41	55.	78.40	41.71	92.	105.50	152.00
19.	73.80	43.35	56.	78.60	58.16	93.	98.00	181.00
20.	75.90	29.31	57.	87.80	88.85	94.	94.50	80.95
21.	76.85	36.60	58.	86.30	155.00	95.	97.00	137.00
22.	80.90	40.25	59.	85.50	70.77	96.	105.00	125.00
23.	79.90	35.43	60.	83.70	75.08	97.	106.00	241.00
24.	89.20	60.09	61.	77.60	57.05	98.	99.00	134.00
25.	82.00	45.84	62.	84.90	99.73	99.	91.00	150.00
26.	92.00	70.40	63.	79.80	27.96	100.	102.50	198.00
27.	86.60	83.45	64.	108.30	123.00	101.	106.00	151.00
28.	80.50	84.30	65.	119.60	90.41	102.	109.10	229.00
29.	86.00	78.89	66.	119.90	106.00	103.	115.00	253.00
30.	82.50	64.75	67.	96.50	144.00	104.	101.00	188.00
31.	83.50	72.56	68.	105.50	121.00	105.	100.10	124.00
32.	88.10	89.31	69.	105.00	97.13	106.	93.30	62.20
33.	90.80	78.94	70.	107.00	166.00	107.	101.80	133.00
34.	89.40	83.55	71.	107.00	87.99	108.	107.90	208.00
35.	102.00	127.00	72.	101.00	154.00	109.	108.50	208.00
36.	94.50	121.00	73.	97.00	100.00			
37.	91.00	107.00	74.	100.00	123.00			

แหล่งที่มา : Jean - Pierre Despres, Ph.D.

ขั้นตอนแรก การศึกษาความสัมพันธ์ระหว่างตัวแปร 2 ตัวแปร สามารถทำได้โดยการพล็อตกราฟระหว่าง 2 ตัวแปรนั้น โดยให้ตัวแปร X ซึ่งเป็นตัวแปรอิสระอยู่ในแกนนอน และตัวแปร Y ซึ่งเป็นตัวแปรตามอยู่ในแกนตั้งดังภาพ



ภาพ 2 แผนภาพการกระจายของข้อมูล

จากภาพ ถ้าลากเส้นผ่านจุดต่าง ๆ เส้นตรงนั้นจะอธิบายความสัมพันธ์ระหว่าง X และ Y

วิธีกำลังสองน้อยที่สุด

เป็นวิธีที่ใช้หาเส้นตรงวิธีหนึ่ง เส้นตรงที่หาได้จากวิธีนี้เรียกว่าเส้นตรงที่มีกำลังสองน้อยที่สุด เส้นตรงโดยทั่วไปหาได้จากสูตร

$$y = a + bx$$

เมื่อ y คือ ค่าบนแกนตั้ง

x คือ ค่าบนแกนนอน

a คือ จุดซึ่งเส้นตรงตัดกับแกนตั้ง

b คือ ความชันของเส้นตรง แสดงปริมาณการเปลี่ยนของ y เมื่อ x เปลี่ยนไป 1 หน่วย

ดังนั้นการลากเส้นตรงจะต้องทราบค่า a และ b ซึ่งสามารถหาได้จาก normal equations ของข้อมูลดังนี้

$$\sum y_i = na + b\sum x_i$$

$$\sum x_i y_i = a\sum x_i + b\sum x_i^2$$

จากตัวอย่างสามารถคำนวณหา normal equation ได้ดังนี้

ตาราง 3.2 การคำนวณหาสมการปกติของตัวอย่าง

x	y	x^2	y^2	xy
74.75	25.72	5587.6	661.5	1922.6
72.60	25.89	5270.8	670.3	1879.6
81.80	42.60	6691.2	1814.8	3484.7
83.95	42.80	7047.6	1831.8	3593.1
74.65	29.84	5572.6	890.4	2227.6
71.85	21.68	5162.4	470.0	1557.7
80.90	29.08	6544.8	845.6	2352.6
83.40	32.98	6955.6	1087.7	2750.5
63.50	11.44	4032.2	130.9	726.4
73.20	32.22	5358.2	1038.1	2358.5
71.90	28.32	5169.6	802.0	2036.2
75.00	43.86	5625.0	1923.7	3289.5
73.10	38.21	5343.6	1460.0	2793.2
79.00	42.48	6241.0	1804.6	3355.9
77.00	30.96	5929.0	958.5	2383.9
68.85	55.78	4740.3	3111.4	3840.5
75.95	43.78	5768.4	1916.7	3325.1
74.15	33.41	5498.2	1116.2	2477.4
73.80	43.35	5446.4	1879.2	3199.2
75.90	29.31	5760.8	859.1	2224.6
76.85	36.60	5905.9	1339.6	2812.7
80.90	40.25	6544.8	1620.1	3256.2
79.90	35.43	6384.0	1255.3	2830.9
89.20	60.09	7956.6	3610.8	5360.0
82.00	45.84	6724.0	2101.3	3758.9
92.00	70.40	8464.0	4956.2	6476.8
86.60	83.45	7499.6	6963.9	7226.8
80.50	84.30	6480.2	7106.5	6786.2
86.00	78.89	7396.0	6223.6	6784.5
82.50	64.75	6806.2	4192.6	5341.9
83.50	72.56	6972.2	5265.0	6058.8
88.10	89.31	7761.6	7976.3	7868.2
90.80	78.94	8244.6	6231.5	7167.8
89.40	83.55	7992.4	6980.6	7469.4
102.00	127.00	10404.0	16129.0	12954.0
94.50	121.00	8930.3	14641.0	11434.5
91.00	107.00	8281.0	11449.0	9737.0
103.00	129.00	10609.0	16641.0	13287.0
80.00	74.02	6400.0	5479.0	5921.6
79.00	55.48	6241.0	3078.0	4382.9
83.50	73.13	6972.2	5348.0	6106.4
76.00	50.50	5776.0	2550.3	3838.0
80.50	50.88	6480.2	2588.8	4095.8
86.50	140.00	7482.2	19600.0	12110.0
83.00	96.54	6889.0	9320.0	8012.8
107.10	118.00	11470.4	13924.0	12637.8
94.30	107.00	8892.5	11449.0	10090.1
94.50	123.00	8930.3	15129.0	11623.5
79.70	65.92	6352.1	4345.4	5253.8
79.30	81.29	6288.5	6608.1	6446.3
89.80	111.00	8064.0	12321.0	9967.8
83.80	90.73	7022.4	8231.9	7603.2
85.20	133.00	7259.0	17689.0	11331.6
75.50	41.90	5700.3	1755.6	3163.5
78.40	41.71	6146.6	1739.7	3270.1
78.60	58.16	6178.0	3382.6	4571.4
87.80	88.85	7708.8	7894.3	7801.0
86.30	155.00	7447.7	24025.0	13376.5
85.50	70.77	7310.3	5008.4	6050.8
83.70	75.08	7005.7	5637.0	6284.2
77.60	57.05	6021.8	3254.7	4427.1
84.90	99.73	7208.0	9946.1	8467.1
79.80	27.96	6368.0	781.8	2231.2
108.30	123.00	11728.9	15129.0	13320.9
119.60	90.41	14304.2	8174.0	10813.0

ตาราง 3.2 (ต่อ) การคำนวณหาสมการปกติ ของตัวอย่าง

x	y	x^2	y^2	xy
119.90	106.00	14376.0	11236.0	12709.4

	96.50	144.00	9312.3	20736.0	13896.0
	105.50	121.00	11130.2	14641.0	12765.5
	105.00	97.13	11025.0	9434.2	10198.6
	107.00	166.00	11449.0	27556.0	17762.0
	107.00	87.99	11449.0	7742.2	9414.9
	101.00	154.00	10201.0	23716.0	15554.0
	97.00	100.00	9409.0	10000.0	9700.0
	100.00	123.00	10000.0	15129.0	12300.0
	108.00	217.00	11664.0	47089.0	23436.0
	100.00	140.00	10000.0	19600.0	14000.0
	103.00	109.00	10609.0	11881.0	11227.0
	104.00	127.00	10816.0	16129.0	13208.0
	106.00	112.00	11236.0	12544.0	11872.0
	109.00	192.00	11881.0	36864.0	20928.0
	103.50	132.00	10712.2	17424.0	13662.0
	110.00	126.00	12100.0	15876.0	13860.0
	110.00	153.00	12100.0	23409.0	16830.0
	112.00	158.00	12544.0	24964.0	17696.0
	108.50	183.00	11772.2	33489.0	19855.5
	104.00	184.00	10816.0	33856.0	19136.0
	111.00	121.00	12321.0	14641.0	13431.0
	108.50	159.00	11772.2	25281.0	17251.5
	121.00	245.00	14641.0	60025.0	29645.0
	109.00	137.00	11881.0	18769.0	14933.0
	97.50	165.00	9506.3	27225.0	16087.5
	105.50	152.00	11130.2	23104.0	16036.0
	98.00	181.00	9604.0	32761.0	17738.0
	94.50	80.95	8930.3	6552.9	7649.8
	97.00	137.00	9409.0	18769.0	13289.0
	105.00	125.00	11025.0	15625.0	13125.0
	106.00	241.00	11236.0	58081.0	25546.0
	99.00	134.00	9801.0	17956.0	13266.0
	91.00	150.00	8281.0	22500.0	13650.0
	102.50	198.00	10506.2	39204.0	20295.0
	106.00	151.00	11236.0	22801.0	16006.0
	109.10	229.00	11902.8	52441.0	24983.9
	115.00	253.00	13225.0	64009.0	29095.0
	101.00	188.00	10201.0	35344.0	18988.0
	100.10	124.00	10020.0	15376.0	12412.4
	93.30	62.20	8704.9	3868.8	5803.3
	101.80	133.00	10363.2	17689.0	13539.4
	107.90	208.00	11642.4	43264.0	22443.2
	108.50	208.00	11772.2	43264.0	22568.0
Total	10017.30	11106.50	940464.0	1486212.0	1089381.0

แทนค่าจากตารางลงในสมการปกติได้ดังนี้

$$11106.50 = 109\ a + 10017.30\ b$$

$$1089381.0 = 10017.30\ a + 940464.00\ b$$

เราสามารถแก้สมการหาค่า a และ b ได้ หรือประมาณจากเส้นตรงที่มีกำลังสองน้อยที่สุด $y = a + bx$ ได้คือ

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) / n}{\sum_{i=1}^{n} (x_i - \bar{x})^2 / n}$$

$$= \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a = \bar{y} - b\bar{x} = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n}$$

ตัวอย่างการคำนวณ

$$b = \frac{109\left(1089381\right) - \left(10017.3\right)\left(11106.5\right)}{109\left(940464\right) - \left(10017.3\right)^2}$$

$$= 3.46$$

$$a = \frac{11106.5 - 3.46\left(10017.3\right)}{109}$$

$$= -216.08$$

ดังนั้นสมการเส้นตรงที่ใช้อธิบายความสัมพันธ์ระหว่าง waist circumference และ deep abdominal AT สามารถเขียนได้เป็น

$$\hat{y} = -216.08 + 3.46\,x$$

a เป็นค่าลบ หมายความว่า เส้นตรงตัดแกน Y ที่ต่ำกว่าจุดกำเนิด (origin)

b เป็นค่าบวก หมายความว่า เส้นตรงลากจากมุมซ้ายด้านล่างของกราฟไปมุมขวาของด้านบน และเมื่อ x เพิ่มขึ้น 1 หน่วย ทำให้ y เพิ่มขึ้น 3.46 หน่วย

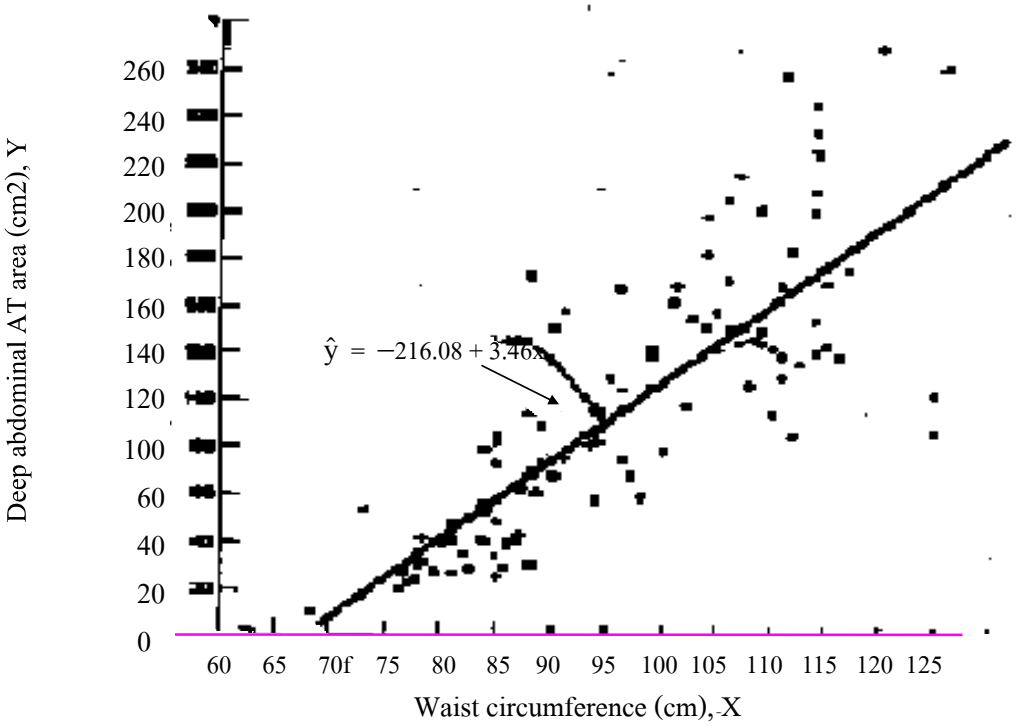
$\hat{y}$ คือ ค่าของ y ที่คำนวณได้จากสมการ

การลากเส้นตรงตามสมการที่ได้ข้างต้นทำได้โดยกำหนดจุด 2 จุด เช่นที่ x = 70 และ x = 110 จะได้

$$\hat{y} = -216.08 + 3.46\,(70) = 26.12$$

$$\hat{y} = -216.08 + 3.46\,(110) = 164.52$$

ลากเส้นตรงของข้อมูลจากตัวอย่างได้ดังภาพ



ภาพ 3 แผนภาพการกระจายของข้อมูลและเส้นกำลังสองน้อยที่สุดของตัวอย่างที่ 1

4. การประเมินสมการถดถอย

เพื่อประเมินสมการถดถอยว่าสามารถอธิบายความสัมพันธ์ระหว่างตัวแปร 2 ตัวแปรได้ดีเพียงใด และสามารถใช้สมการถดถอยในการทำนายและประมาณค่า Y ได้อย่างมีประสิทธิภาพหรือไม่

สมมุติฐานทางสถิติที่ต้องการทดสอบคือ $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$

ถ้าในประชากรความสัมพันธ์ระหว่าง X และ Y เป็นแบบเส้นตรง β คือความชันของเส้นตรงที่อธิบายความสัมพันธ์นั้น ซึ่งอาจมีค่าเป็นบวก ลบ หรือศูนย์ ถ้า β มีค่าศูนย์ หมายความว่ากลุ่มตัวอย่างที่สุ่มมาจากประชากรให้สมการถดถอยที่ไม่สามารถทำนายและประมาณค่า Y ได้ นอกจากนี้ความสัมพันธ์ระหว่าง X และ Y อาจไม่ใช่แบบเส้นตรง ดังนั้นถ้าทดสอบสมมติฐานได้ว่ายอมรับ H_0 เราจะสรุปว่า