Lecture Notes

# ความน่าจะเป็นและสถิติ

# Probability and Statistics

**ศาสตราจารย์ ดร. สุขสันต์ พาณิชพาพิบูล**

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

# บทที่ 13

# การถดถอย

# Regression

ในบทนี้เราจะศึกษาการถดถอยอย่างง่าย ซึ่งประกอบไปด้วยหัวข้อต่างๆ ดังนี้

- สหสัมพันธ์ (Correlation)
- การถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression)
- การถดถอยเชิงซ้อน (Multiple Linear Regression)

## 13.1 สหสัมพันธ์ (Correlation)

บ่อยครั้งที่เราจำเป็นต้องหาความสัมพันธ์ระหว่างตัวแปร 2 ตัวในชุดข้อมูล เช่น เราอาจอยากทราบว่าอายุของผู้ใช้งานเฟซบุ๊คกับจำนวนของเพื่อนมีความสัมพันธ์กันหรือไม่ และในลักษณะใด เช่น จำนวนเพื่อนจะแปรผันตรงกับอายุของผู้ใช้งานหรือไม่ มีความสัมพันธ์กันในลักษณะเชิงเส้นหรือไม่ สามารถเขียนสมการขึ้นมาเพื่ออธิบายความสัมพันธ์ได้หรือไม่ หรืออีกตัวอย่างเช่น เราอาจอยากทราบว่าส่วนสูงของคนกับความยาวของแขนมีความสัมพันธ์กันหรือไม่ และในลักษณะใด เช่น ความยาวของแขนแปรผันตรงกับส่วนสูงของคนใช่หรือไม่ เมื่อทราบว่าชายคนหนึ่งสูง 165 ซม. จะสามารถประมาณความยาวแขนของเขาได้หรือไม่ มีวิธีการประมาณอย่างไร เป็นต้น

---

เนื้อหาในบทนี้อ้างอิงจาก [4]

**ปุจฉา 13.1** สัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) คืออะไร?

*วิสัชนา* สัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) คือ ค่าที่ใช้วัดว่าตัวแปร 2 ชุด เช่น $x$ และ $y$ มีความสัมพันธ์กันในลักษณะเชิงเส้นมากน้อยเพียงใด เรามักแทนค่าสัมประสิทธิ์สหสัมพันธ์ด้วยตัวแปร $r$

**ปุจฉา 13.2** ค่าสัมประสิทธิ์สหสัมพันธ์ มีค่าในช่วงใด?

*วิสัชนา* ค่าสัมประสิทธิ์สหสัมพันธ์ หรือ $r$ มีค่าตั้งแต่ -1 ถึง 1 กล่าวคือ $-1 \leq r \leq 1$ ถ้าค่า $r$ เป็นบวกแสดงว่าตัวแปร 2 ตัวมีความสัมพันธ์เชิงบวก เช่น ความสูงกับความยาวแขนของคนมีความสัมพันธ์ในเชิงบวก กล่าวคือ คนที่มีส่วนสูงมากก็มีแนวโน้มที่แขนจะยาวมากขึ้นตามสัดส่วน ถ้าค่า $r$ เป็นลบแสดงว่าตัวแปร 2 ตัวมีความสัมพันธ์เชิงลบหรือสวนทางกัน เช่น อุณหภูมิของอากาศกับยอดขายกาแฟร้อนมีความสัมพันธ์ในเชิงลบ กล่าวคือ ยิ่งอุณหภูมิสูง (อากาศร้อน) ยอดขายกาแฟร้อนก็ยิ่งลดลง เป็นต้น

ยิ่งค่า $r$ เข้าใกล้ 1 มากก็แสดงว่าตัวแปร 2 ตัวมีความสัมพันธ์เชิงบวกยิ่งเด่นชัดมาก ยิ่งค่า $r$ เข้าใกล้ -1 มากก็แสดงว่าตัวแปร 2 ตัวมีความสัมพันธ์เชิงลบยิ่งเด่นชัดมาก

**ปุจฉา 13.3** เมื่อค่า $r = 0$ หมายความว่าอย่างไร?

*วิสัชนา* เมื่อค่า $r = 0$ หมายความว่า ตัวแปรสองตัวไม่มีสหสัมพันธ์กัน (uncorrelated) กล่าวคือไม่มีความสัมพันธ์กันในลักษณะเชิงเส้นนั่นเอง

**ปุจฉา 13.4** ค่าสัมประสิทธิ์สหสัมพันธ์ มีวิธีคำนวณอย่างไร?

*วิสัชนา* กำหนด $(x_1, y_1), \ldots, (x_n, y_n)$ เป็นข้อมูลลำดับที่ 1 ถึง $n$ ของตัวแปร $x$ และ $y$ เราสามารถคำนวณค่าสัมประสิทธิ์สหสัมพันธ์ $r$ ดังนี้

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)$$

โดยที่ $\overline{x}$ คือค่าเฉลี่ยของ $x$, $\overline{y}$ คือค่าเฉลี่ยของ $y$, $s_x$ คือค่าเบี่ยงเบนมาตรฐานของ $x$, และ $s_y$ คือ ค่าเบี่ยงเบนมาตรฐานของ $y$

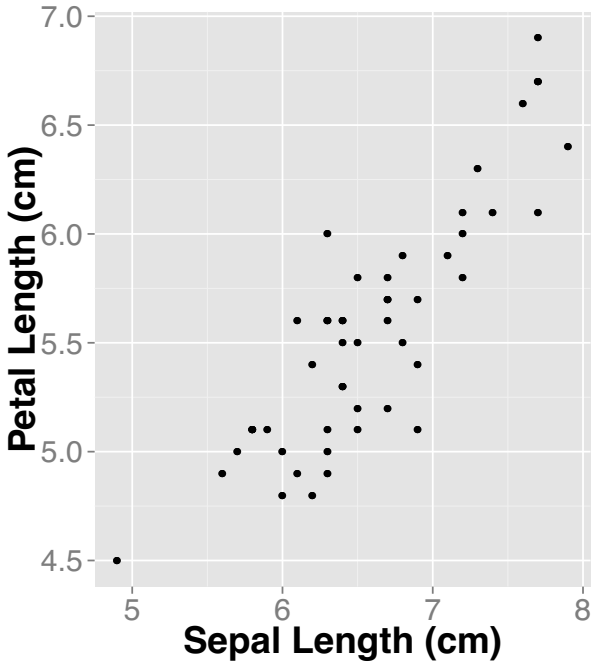
เนื่องจาก $s_x = \sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2/(n-1)}$ และ $s_y = \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2/(n-1)}$ เราสามารถแทนเข้าไปในสมการข้างต้นซึ่งจะได้

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

หรือจะเขียนอีกแบบหนึ่งก็ได้คือ

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}}$$

**ตัวอย่าง 13.1** จากรูป 13.1 ซึ่งแสดงความสัมพันธ์ระหว่างความยาวกลีบเลี้ยงและความยาวกลีบดอก ของดอกไอริส (Iris) โดยมีความยาวกลีบเลี้ยง (Sepal length) บนแกน $x$ และ ความยาวกลีบดอก (Petal length) บนแกน $y$ ค่าสัมประสิทธิ์สหสัมพันธ์ควรเป็นอย่างไร




รูป 13.1: แผนภาพแสดงความสัมพันธ์ระหว่างความยาวกลีบเลี้ยงและความยาวกลีบดอกไอริส สายพันธุ์ Virginica

*วิธีทำ* จากรูป 13.1 จะเห็นว่าเมื่อความยาวกลีบเลี้ยง ($x$) มากขึ้น ความยาวกลีบดอก ($y$) ก็มีแนวโน้มเพิ่มขึ้นตาม ดังนั้นค่าสัมประสิทธิ์สหสัมพันธ์ $r$ ควรมีเครื่องหมายเป็นบวก ซึ่งจากข้อมูลตัวอย่างนี้เมื่อคำนวณมาจะได้ค่า $r = 0.864$

**ปุจฉา 13.5** สหสัมพันธ์ระหว่างตัวแปร 2 ตัว $x, y$ เป็นความสัมพันธ์ในเชิงสาเหตุและผลที่เกิด กล่าวคือ $x$ ส่งผลให้เกิด $y$ ใช่หรือไม่?

*วิสัชนา* ไม่ใช่ สหสัมพันธ์ระหว่างตัวแปร 2 ตัว $x, y$ ไม่เป็นความสัมพันธ์ในเชิงสาเหตุและผลที่เกิด (Correlation is not causation) การที่ $x$ และ $y$ มีความสัมพันธ์กันไม่ได้หมายความว่า $x$ ส่งผลให้เกิด $y$ เสมอไป

เช่น เราอาจทราบว่าขนาดเท้าของเด็กมีความสัมพันธ์ในเชิงบวกกับจำนวนคำศัพท์ที่เด็กจำได้ กล่าวคือ ยิ่งเด็กมีขนาดเท้าใหญ่ขึ้นก็จะมีจำนวนคำศัพท์ที่จำได้มากขึ้น เหตุที่ความสัมพันธ์เป็นไปในลักษณะนี้เพราะเมื่อเด็กอายุมากขึ้นเท้าก็จะใหญ่ขึ้น รวมถึงมีความสามารถในการจำคำศัพท์ได้มากขึ้นด้วย แต่ไม่ได้แปลว่าขนาดเท้าที่ใหญ่เป็นสาเหตุให้เด็กจำคำศัพท์ได้มากขึ้น

**ตัวอย่าง 13.2** จากค่าของตัวแปร $x, y$ ที่กำหนดให้ในตาราง 13.1 จงคำนวณค่าสัมประสิทธิ์สหสัมพันธ์

ตาราง 13.1: แสดงค่า $x, y$

| x | y |
|---|---|
| 2 | 4 |
| 3 | 8 |
| 7 | 12 |
| 4 | 9 |
| 9 | 15 |

*วิธีทำ* คำนวณค่าต่างๆ ดังนี้

$$
\begin{aligned}
\overline{x} &= \frac{1}{5}(2+3+7+4+9) = 5 \\
\overline{y} &= \frac{1}{5}(4+8+12+9+15) = 9.6 \\
s_x &= \sqrt{\sum_{i=1}^{5} \frac{(x_i-5)^2}{4}} = 2.9154 \\
s_y &= \sqrt{\sum_{i=1}^{5} \frac{(y_i-9.6)^2}{4}} = 4.1593
\end{aligned}
$$

หาค่าสัมประสิทธิ์สหสัมพันธ์ ดังนี้

$$
\begin{aligned}
r &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{s_x}\right)\left(\frac{y_i-\overline{y}}{s_y}\right) \\
&= \frac{1}{4}\sum_{i=1}^{5}\left(\frac{x_i-5}{2.9154}\right)\left(\frac{y_i-9.6}{4.1593}\right) \\
&= 0.969
\end{aligned}
$$

ดังนั้นได้ค่า $r = 0.969$

## 13.2 การถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression)

บ่อยครั้งที่เรามีข้อมูลแสดงความสัมพันธ์ของตัวแปร 2 ตัว เช่น ส่วนสูงและความยาวแขนของคน หรือความยาวกลีบเลี้ยงและความยาวกลีบดอก ซึ่งเมื่อเรานำข้อมูลเหล่านี้มาแสดงผลในแผนภาพการกระจาย หรือที่เรียกว่าสแก็ตเตอร์พล็อต (Scatter Plot) เราก็จะเห็นแนวโน้มของความสัมพันธ์ อย่างเช่นที่แสดงในรูป 13.1 เราจะเห็นแนวโน้มของความสัมพันธ์ว่า หากความยาวกลีบเลี้ยงยาวขึ้น ความยาวกลีบดอกก็จะยาวขึ้นเป็นสัดส่วนที่สัมพันธ์กันด้วย และจะเป็นการดียิ่งขึ้น ถ้าหากเราสามารถหาสมการเชิงเส้นมาเชื่อมโยงความสัมพันธ์ระหว่าง 2 ตัวแปร ซึ่งการสร้างสมการเชิงเส้นมาอธิบายเชื่อมโยงความสัมพันธ์ระหว่าง 2 ตัวแปรก็คือการวิเคราะห์การถดถอยแบบเชิงเส้นนั่นเอง (Linear Regression)

**ปุจฉา 13.6** การวิเคราะห์การถดถอยแบบเชิงเส้นคืออะไร? มีวิธีการทำอย่างไร?

*วิสัชนา* การวิเคราะห์การถดถอยแบบเชิงเส้น คือ การหาสมการเชิงเส้นมาอธิบายความสัมพันธ์ระหว่างตัวแปร 2 ตัว $(x, y)$ สมการเชิงเส้นนี้หาได้โดยดูจากข้อมูล $x, y$ ที่มี ยกตัวอย่างเช่น เราอาจต้องการสมการเชิงเส้นมาอธิบายความสัมพันธ์ระหว่างความยาวกลีบเลี้ยงและความยาวกลีบดอกของดอกไอริส ตามที่แสดงในรูป 13.2

จากรูป 13.2 จะเห็นได้ว่า เส้นตรงที่สามารถแสดงแนวโน้มของความสัมพันธ์ระหว่างตัวแปร $x$ และ $y$ นั้นอาจเป็นไปได้ได้หลายเส้น เช่น เราอาจลากเส้นตรงอีกเส้นที่มีความชันต่างไปจากเส้นที่แสดงในรูป 13.2 เล็กน้อยก็ได้ แต่ทั้งนี้เราควรเลือก "**เส้นที่ดีที่สุด**" มาเป็นเส้นถดถอย ซึ่งวิธีการเลือกเส้นที่ดีที่สุดจะกล่าวถึงต่อไป

**ปุจฉา 13.7** กำหนด $x$ และ $y$ เป็นตัวแปรในชุดข้อมูล เรามีวิธีเขียนสมการเชิงเส้นแสดงความสัมพันธ์ระหว่าง $x$ และ $y$ อย่างไร?

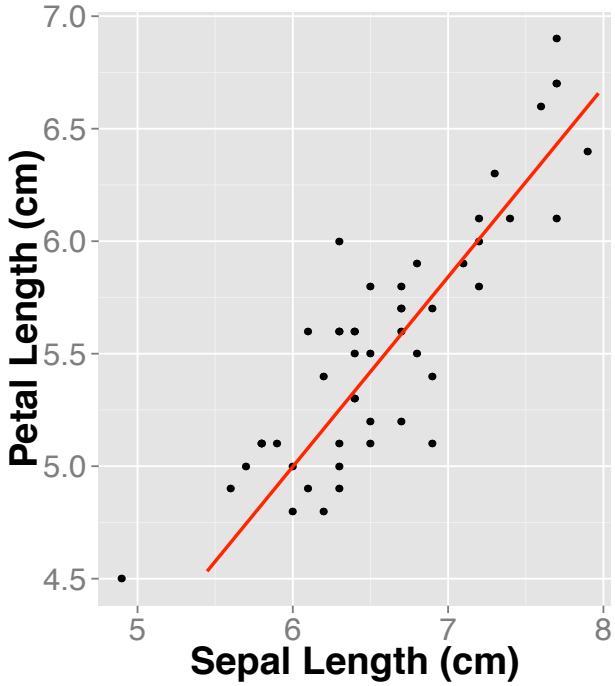
*วิสัชนา* เขียนเป็นสมการได้ดังนี้

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

เราเรียก $y_i$ ว่าตัวแปรตาม (Dependent Variable), เรียก $x_i$ ว่า ตัวแปรอิสระ (Independent Variable), เรียก $\beta_0$, $\beta_1$ ว่า สัมประสิทธิ์การถดถอย (Regression Coefficients) และเรียก $\epsilon_i$ ว่าความผิดพลาด (Error)

ในความเป็นจริงแล้วเราไม่สามารถหาค่า $\beta_0$ และ $\beta_1$ ได้อย่างแม่นยำ 100% เนื่องจากความผิดพลาด $\epsilon_i$ ที่เกิดขึ้น แต่เราสามารถประมาณหาค่า $\beta_0$ และ $\beta_1$ ได้ ซึ่งเราจะเขียนเป็นสมการดังนี้

$$y = \widehat{\beta_0} + \widehat{\beta_1} x$$



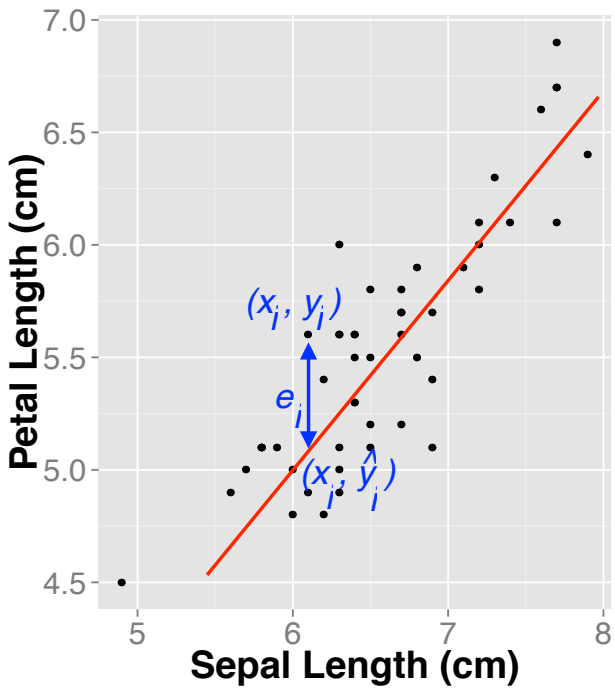

รูป 13.2: แผนภาพแสดงความสัมพันธ์ระหว่างความยาวกลีบเลี้ยงและความยาวกลีบดอกไอริส สายพันธุ์ Virginica และเส้นถดถอย

สมการเส้นตรง $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$ นี้จะถูกใช้เป็นสมการถดถอยสำหรับฟิต (Fit) ข้อมูล $x, y$

**ปุจฉา 13.8** เส้นตรงที่ดีที่สุด ในการแสดงแนวโน้มของความสัมพันธ์ระหว่างตัวแปร $x$ และตัวแปร $y$ ควรเป็นอย่างไร?

*วิสัชนา* ควรเป็นเส้นที่ทำให้ส่วนต่าง (Residual) ระหว่าง "ค่าจริง" กับ "ค่าบนเส้น" น้อยที่สุด กำหนด $e_i = y_i - \widehat{y}_i$ เป็นส่วนต่างระหว่าง $y_i$ ซึ่งเป็นค่าจริงของตัวแปร $y$ ในลำดับที่ $i$ และ $\widehat{y}_i$ ซึ่งเป็นค่าที่ได้จากสมการเส้นตรง $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ ส่วนต่าง $e_i$ นี้แสดงให้เห็นในรูป 13.3 ซึ่งก็คือระยะห่างในแนวตั้งระหว่างค่า $y_i$ (จุดข้อมูล) กับค่า $\widehat{y}_i$ บนเส้นตรงที่เราสร้างขึ้นนั่นเอง เส้นตรงที่ดีที่สุดคือเส้นที่ทำให้ผลรวมของส่วนต่างน้อยที่สุด หรือเส้นที่ทำให้ $\sum_{i=1}^{n} e_i^2$ มีค่าต่ำที่สุดนั่นเอง

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$




รูป 13.3: แผนภาพแสดงความสัมพันธ์ระหว่างความยาวกลีบเลี้ยงและความยาวกลีบดอกไอริส สายพันธุ์ Virginica และเส้นถดถอย ค่า $e_i$ คือส่วนต่างระหว่างค่าจริงของข้อมูล $(y_i)$ กับค่าที่อยู่บนเส้นถดถอยที่สร้างขึ้น $(\widehat{y}_i)$

**ปุจฉา 13.9** สมการเส้นตรงที่ดีที่สุดสำหรับใช้เป็นเส้นถดถอย มีสมการอย่างไร?

*วิสัชนา* มีสมการซึ่งหาได้ดังนี้ เนื่องจากเราต้องการทำให้ $\sum_{i=1}^{n} e_i^2$ มีค่าต่ำสุด จึงต้องมาดู $e_i$ ก่อน เขียนสมการของ $e_i$ ได้ดังนี้

$$e_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$$

ดังนั้น $\widehat{\beta}_0, \widehat{\beta}_1$ เป็นตัวที่จะทำให้ผลรวม $\sum_{i=1}^{n} e_i^2$ ต่ำที่สุด กำหนดค่า $S$ แทนผลรวมดังนี้

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$$

เราสามารถหาค่า $\widehat{\beta}_0$ และ $\widehat{\beta}_1$ ที่ทำให้สมการข้างต้นมีค่าต่ำสุดได้โดยการหาอนุพันธ์เพื่อดูจุดต่ำสุดของสมการข้างต้น กล่าวคือ ต้องหาค่า $\widehat{\beta}_0$ และ $\widehat{\beta}_1$ ที่เป็นคำตอบของระบบสมการต่อไปนี้

$$\begin{aligned} \frac{\partial S}{\partial \widehat{\beta}_0} &= -\sum_{i=1}^{n} 2(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0 \\ \frac{\partial S}{\partial \widehat{\beta}_1} &= -\sum_{i=1}^{n} 2x_i(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0 \end{aligned}$$

ซึ่งเมื่อแก้ระบบสมการนี้แล้วจะได้คำตอบคือ

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ \widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1 \overline{x} \end{aligned}$$

สำหรับผู้สนใจสามารถดูรายละเอียดวิธีการหาคำตอบของสมการเพิ่มเติมได้จาก [4]

**ตัวอย่าง 13.3** จากค่าของตัวแปร $x, y$ ที่กำหนดให้ในตาราง 13.2 จงหาสมการถดถอยสำหรับข้อมูลชุดนี้

*วิธีทำ* หาตามขั้นตอนดังนี้

ตาราง 13.2: แสดงค่า $x, y$

| x | y |
|---|---|
| 2 | 4 |
| 3 | 8 |
| 7 | 12 |
| 4 | 9 |
| 9 | 15 |

1. หาค่าที่จำเป็นต่างๆ ดังนี้

$$\begin{aligned}
\overline{x} &= \frac{1}{5}(2+3+7+4+9) = 5 \\
\overline{y} &= \frac{1}{5}(4+8+12+9+15) = 9.6 \\
\sum_{i=1}^{5}(x_i-\overline{x})^2 &= 34 \\
\sum_{i=1}^{5}(x_i-\overline{x})(y_i-\overline{y}) &= 47
\end{aligned}$$

2. หาค่า $\widehat{\beta}_1$

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^{5}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{5}(x_i-\overline{x})^2} \\
&= \frac{47}{34} = 1.3824
\end{aligned}$$

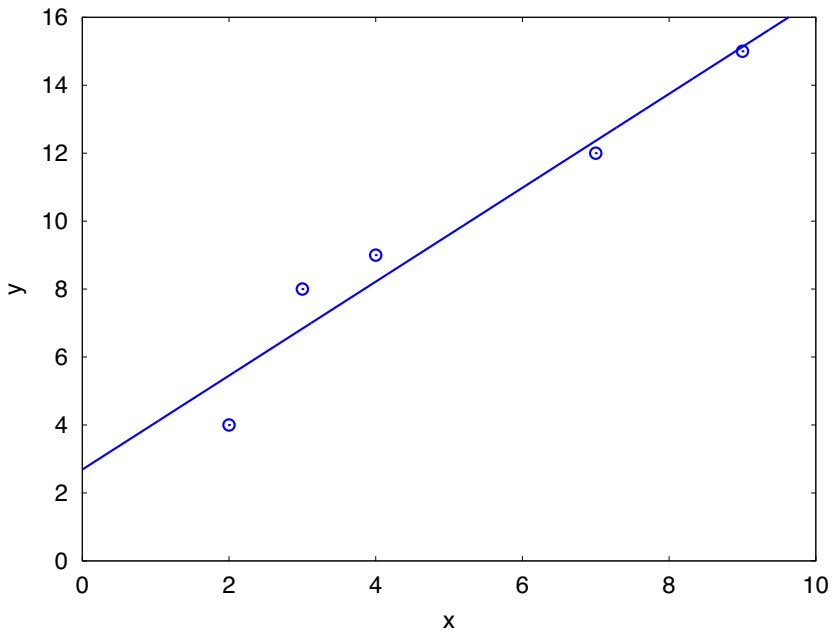
3. หาค่า $\widehat{\beta}_0$

$$\begin{aligned}
\widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1\overline{x} \\
&= 9.6 - (1.3824 \times 5) \\
&= 2.6882
\end{aligned}$$

4. เขียนสมการถดถอย คือ

$$
\begin{aligned}
y &= \widehat{\beta}_0 + \widehat{\beta}_1 x \\
y &= 2.6882 + 1.3824\,x
\end{aligned}
$$

เมื่อวาดสมการถดถอยและจุดข้อมูล $(x, y)$ จะได้ดังรูป 13.4




รูป 13.4: จุดข้อมูล $(x, y)$ และเส้นถดถอย

## 13.3 การถดถอยเชิงซ้อน (Multiple Linear Regression)

ในส่วนที่แล้ว เราได้ศึกษาวิธีการสร้างตัวแบบการถดถอยเชิงเส้นอย่างง่าย กล่าวคือ ตัวแปรตามหรือ $y$ นั้นเป็นฟังก์ชันของตัวแปรอิสระหรือ $x$ เพียงตัวเดียว ในการประยุกต์ใช้งานจริง เราพบว่าบ่อยครั้งที่ตัวแปรตามนั้นจะเป็นฟังก์ชันของตัวแปรอิสระมากกว่าหนึ่งตัว ยกตัวอย่างเช่น ราคาบ้านอาจขึ้นอยู่กับตัวแปรอิสระหลายๆตัว เช่น ขนาดพื้นที่ของบ้าน จำนวนชั้น และทำเลที่ตั้งของบ้าน เป็นต้น การสร้างตัวแบบการถดถอยเชิงซ้อน (Multiple Linear Regression Model) ก็คือการนำข้อมูลจากหลายๆปัจจัยที่ส่งผลต่อตัวแปรตามมาพิจารณาร่วมกันนั่นเอง

**ปุจฉา 13.10** กำหนด $x_1, x_2, \ldots, x_p$ เป็นตัวแปรอิสระจำนวน $p$ ตัวในชุดข้อมูล และ $y$ เป็นตัวแปรตาม สมการเชิงเส้นสำหรับแสดงความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระมีลักษณะอย่างไร?

*วิสัชนา* มีลักษณะดังนี้

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon_i$$

และในทำนองเดียวกันกับกรณีสมการถดถอยอย่างง่าย เราจะประมาณค่า $y$ ตามสมการดังนี้

$$\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + \cdots + \widehat{\beta_p} x_p$$

โดยที่ $\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_p}$ คือค่าประมาณของสัมประสิทธิ์การถดถอย

**ปุจฉา 13.11** สมการเชิงเส้นที่ดีที่สุดสำหรับใช้เป็นสมการถดถอยในกรณีที่ตัวแปรอิสระมีหลายตัว หาได้อย่างไร?

*วิสัชนา* เช่นเดียวกันกับกรณีสมการถดถอยอย่างง่าย เราต้องทำให้ส่วนต่างระหว่างค่าจริง ($y_i$) กับค่าที่ได้จากสมการ ($\widehat{y_i}$) ต่างกันน้อยที่สุด ซึ่งส่วนต่างเขียนได้ตามสมการ

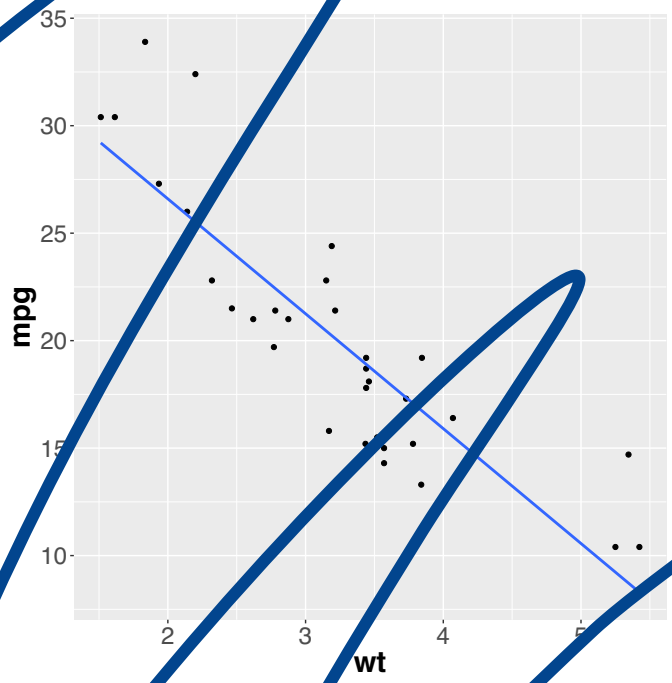
$$\begin{aligned} e &= y_i - \widehat{y_i} \\ &= y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_{1i} + \widehat{\beta_2} x_{2i} + \cdots + \widehat{\beta_p} x_{pi}) \end{aligned}$$

ผลรวมกำลังสองของส่วนต่างคือ

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left( y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_{1i} + \widehat{\beta_2} x_{2i} + \cdots + \widehat{\beta_p} x_{pi}) \right)^2$$

สมการเชิงเส้นที่ดีที่สุดที่ควรใช้เป็นสมการถดถอยก็คือ สมการเชิงเส้นที่มีสัมประสิทธิ์ $\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_p}$ ที่ทำให้ $\sum_{i=1}^{n} e_i^2$ ในสมการข้างต้นมีค่าน้อยที่สุด ซึ่งสามารถหาได้โดยหาอนุพันธ์ต่างๆของสมการข้างต้นซึ่งก็คือ $\frac{\partial S}{\partial \widehat{\beta_0}}, \frac{\partial S}{\partial \widehat{\beta_1}}, \ldots, \frac{\partial S}{\partial \widehat{\beta_p}}$ แล้วตั้งอนุพันธ์เหล่านี้ให้มีค่าเท่ากับ 0 และแก้ระบบสมการเพื่อหาจุดต่ำสุด คำตอบของระบบสมการที่ได้ก็คือ ค่า $\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_p}$ ที่จะทำให้ $\sum_{i=1}^{n} e_i^2$ มีค่าน้อยที่สุด

อย่างไรก็ตาม ในทางปฏิบัติเราจะใช้เครื่องมือ เช่น โปรแกรมคอมพิวเตอร์ ในการคำนวณหาค่าสัมประสิทธิ์เหล่านี้ ดังนั้นเราไม่จำเป็นต้องคำนวณเองด้วยมือ



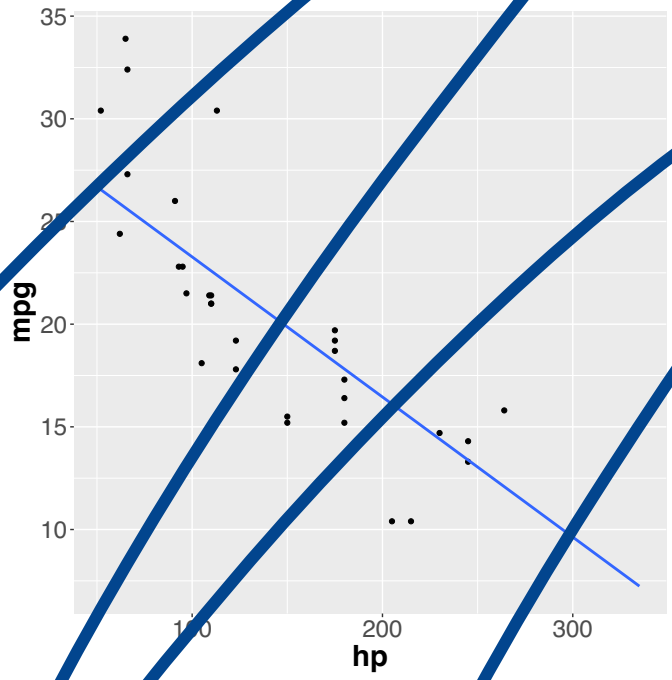

รูป 13.5: แผนภาพแสดงความสัมพันธ์ระหว่างประสิทธิภาพการใช้พลังงาน (mpg) และน้ำหนักของรถ (wt) และเส้นถดถอย

**ตัวอย่าง 13.4** ในตัวอย่างนี้ เราจะลองดูตัวอย่างความสัมพันธ์ระหว่างประสิทธิภาพด้านการใช้พลังงานของรถ ซึ่งวัดในหน่วย Miles/gallon (mpg) กับตัวแปรอิสระอื่นๆ เช่น น้ำหนักของรถในหน่วยตัน และกำลังของรถในหน่วยแรงม้า

รูป 13.5 แสดงแผนภาพการกระจายระหว่าง ประสิทธิภาพการใช้พลังงาน (mpg) กับ ตัวแปรอิสระ 1 ตัว คือ น้ำหนักของรถ (wt) ในหน่วยตัน จากรูป จะสังเกตได้ว่าความชันของเส้นถดถอยเป็นลบ ซึ่งกล่าวได้ว่าประสิทธิภาพการใช้พลังงานกับน้ำหนักของรถ มีความสัมพันธ์กันในเชิงลบ กล่าวคือ เมื่อน้ำหนักรถเพิ่มมากขึ้น ประสิทธิภาพการใช้พลังงานลดลง หรือระยะทางที่รถวิ่งได้ต่อน้ำมันหนึ่งแกลลอนลดลงนั่นเอง เมื่อใช้โปรแกรมคำนวณจะได้สมการถดถอยดังนี้

$$y = 37.2851 - 5.3444x$$

เราสามารถนำสมการถดถอยนี้ไปประเมินหาประสิทธิภาพการใช้พลังงานได้ เช่น เราอาจจะอยากประเมินว่ารถที่มีน้ำหนัก 4 ตัน น่าจะมีค่าประสิทธิภาพการใช้พลังงานเท่าใด ซึ่งทำได้โดยแทนค่า $x = 4$ เข้าไปในสมการถดถอย จะได้ $y = 15.9075$ ดังนั้นสามารถประเมินได้ว่า รถที่มีน้ำหนัก 4 ตัน น่าจะมีค่าประสิทธิภาพการใช้พลังงานเท่ากับ 15.9075 miles/gallon



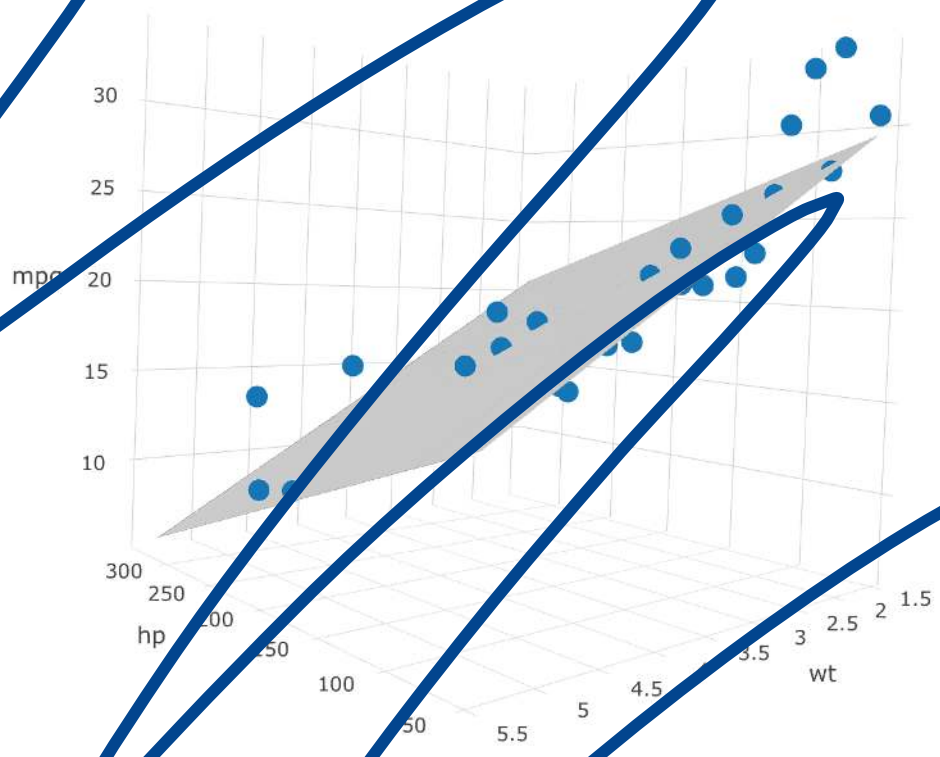

รูป 13.6: แผนภาพแสดงความสัมพันธ์ระหว่างประสิทธิภาพการใช้พลังงาน (mpg) และกำลังของรถ (hp) และเส้นถดถอย

ประสิทธิภาพด้านการใช้พลังงานของรถ ยังมีสหสัมพันธ์กับตัวแปรอื่นๆ อีก เช่น กำลังของรถในหน่วยแรงม้า รูป 13.6 แสดงแผนภาพการกระจายระหว่าง ประสิทธิภาพการใช้พลังงาน (mpg) กับ ตัวแปรอิสระ 1 ตัว คือ กำลังของรถ (hp) ในหน่วยแรงม้า จากรูป จะสังเกตได้ว่าความชันของเส้นถดถอยเป็นลบ ซึ่งกล่าวได้ว่าประสิทธิภาพการใช้พลังงานกับกำลังของรถ มีความสัมพันธ์กันในเชิงลบ กล่าวคือ เมื่อกำลังของรถเพิ่มมากขึ้น ประสิทธิภาพการใช้พลังงานลดลง หรือ ระยะทางที่รถวิ่งได้ต่อน้ำมันหนึ่งแกลลอนลดลงนั่นเอง เมื่อใช้โปรแกรมคำนวณจะได้สมการถดถอยดังนี้

$$y = 30.0988 - 0.0682x$$

เราสามารถนำสมการถดถอยนี้ไปประเมินหาประสิทธิภาพการใช้พลังงานได้ เช่น เราอาจจะอยากประเมินว่ารถที่มีกำลัง 100 แรงม้า น่าจะมีค่าประสิทธิภาพการใช้พลังงานเท่าใด ซึ่งทำได้โดยแทนค่า $x = 100$ เข้าไปในสมการถดถอย จะได้ $y = 23.2788$ ดังนั้นเราสามารถประเมินได้ว่ารถที่มีกำลัง 100 แรงม้า น่าจะมีค่าประสิทธิภาพการใช้พลังงานเท่ากับ 23.2788 miles/gallon

นอกจากนี้ เรายังสามารถใช้การถดถอยเชิงซ้อน มาประเมินความสัมพันธ์ระหว่างประสิทธิภาพการใช้พลังงานกับตัวแปรอิสระ 2 ตัวร่วมกัน รูป 13.7 แสดงแผนภาพการกระจายระหว่างประสิทธิภาพการใช้




รูป 13.7: แผนภาพการกระจายระหว่างประสิทธิภาพการใช้พลังงาน (mpg) กับ ตัวแปรอิสระ 2 ตัว คือ น้ำหนักของรถ (wt) ในหน่วยตัน และกำลังของรถ (hp) ในหน่วยแรงม้า และระนาบถดถอย

พลังงาน (mpg) กับ ตัวแปรอิสระ 2 ตัว คือ น้ำหนักของรถ (wt) ในหน่วยตัน และกำลังของรถ (hp) ในหน่วยแรงม้า เนื่องจากมีตัวแปรอิสระ 2 ตัว แผนภาพการกระจายจึงมี 3 มิติ เราสามารถสร้างระนาบถดถอยสำหรับความสัมพันธ์นี้ เมื่อใช้โปรแกรมคำนวณจะได้สมการถดถอยดังนี้

$$y = 37.2272 - 3.8778x_1 - 0.0317x_2$$

โดยที่ $x_1$ คือน้ำหนักรถในหน่วยตัน และ $x_2$ คือกำลังของรถในหน่วยแรงม้า เราสามารถนำสมการถดถอยนี้ไปประเมินหาประสิทธิภาพการใช้พลังงานได้ เช่น เราอาจจะอยากประเมินว่า รถที่มีน้ำหนัก 3 ตัน มีกำลัง 200 แรงม้า น่าจะมีค่าประสิทธิภาพการใช้พลังงานเท่าใด ซึ่งทำได้โดยแทนค่า $x_1 = 3$ และ $x_2 = 200$ เข้าไปในสมการถดถอย จะได้ $y = 19.2538$ ดังนั้นสามารถประเมินได้ว่ารถที่มีน้ำหนัก 3 ตัน มีกำลัง 200 แรงม้า น่าจะมีค่าประสิทธิภาพการใช้พลังงานเท่ากับ 19.2538 miles/gallon

# บรรณานุกรม


[1] S. Ross, *A First Course in Probability*, 9th ed. Pearson, 2012.

[2] D. P. Bertsekas and J. N. Tsitsiklis, *Introduction to Probability*, 2nd ed. Athena Scientific, 2008.

[3] A. Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*, 3rd ed. Pearson, 2008.

[4] W. Navidi, *Statistics for Engineers and Scientists*, 3rd ed. McGraw-Hill, 2010.