รายวิชา 568 351 สถิติและการประยุกต์ทางเภสัชศาสตร์

สมการถดถอยเชิงเส้นอย่างง่าย (SIMPLE LINEAR REGRESSION)

รศ.ดร.ลาวัลย์ ศรัทธาพุทธ

ภาควิชาสารสนเทศศาสตร์ทางสุขภาพ คณะเภสัชศาสตร์ มหาวิทยาลัยศิลปากร

ความสัมพันธ์เชิงเส้น

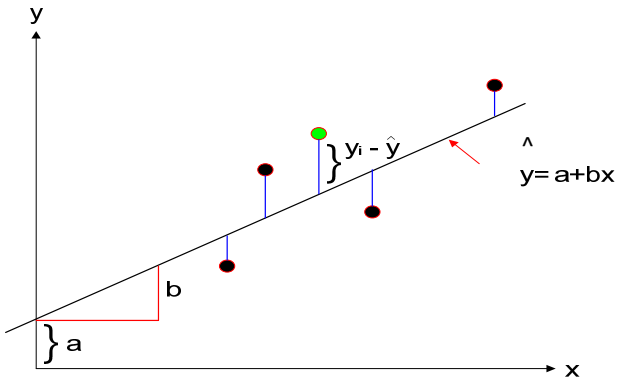
- สัมประสิทธิ์สหสัมพันธ์เชิงเส้น (r) เป็นค่าที่ใช้วัดองศาแห่งความสัมพันธ์เชิงเส้น**ระหว่างตัวแปร 2 ตัว**
- ในกรณีที่ 2 ตัวแปรมี**ความสัมพันธ์เชิงเส้นอย่างสมบูรณ์**ในทางบวกหรือในทางลบ ทุกหน่วยของข้อมูล**จะสามารถเขียนให้อยู่ในรูป** $Y_i = a+bX_i$ คือทราบอัตราการเปลี่ยนแปลง (ค่า X เปลี่ยนไป 1 หน่วย ค่า Y จะเปลี่ยนไป b หน่วย)
- ในกรณีที่ 2 ตัวแปรมี**ความสัมพันธ์เชิงเส้นที่ไม่สมบูรณ์**จะบอกไม่ได้ว่าค่า X เปลี่ยนไป 1 หน่วย ค่า Y จะเปลี่ยนไปกี่หน่วย และ**ไม่สามารถเขียนให้อยู่ในรูป** $Y_i = a+bX_i$

สมการถดถอยเชิงเส้น

- ถ้าต้องการทราบอัตราการเปลี่ยนแปลง ต้องหา**สมการเส้นตรงที่ลากผ่านจุด** (X_i, Y_i) **ต่างๆ โดยให้ห่างจากจุดเหล่านี้น้อยที่สุด** (นั่นคือลากเส้นตรงด้วยหลักการ least square) เรียกเส้นตรงนี้ว่า **สมการถดถอยเชิงเส้น (regression line)** ซึ่งแทนด้วย $\hat{Y}_i = a + bX_i$

"Least Squares" Concept

$$y_i = a + x_i b + e_i \quad \Leftarrow \quad Min\left[\sum_{i=1}^{n}(y_i - (a + x_i b))^2\right]$$



ระบบสมการเชิงเส้นที่สามารถหาผลลัพธ์ได้

- **ระบบสมการเชิงเส้น (Simultaneous linear equations)** ที่สามารถหาผลลัพธ์ได้ หมายถึงสมการทุกสมการเป็นจริงพร้อมๆ กันในเวลาเดียวกันเมื่อแทนค่าผลลัพธ์ในสมการเหล่านั้น เช่น

2b = 1
3b = 3/2
4b = 2

$$Y_i = a + bX_i$$

- ผลลัพธ์ของระบบสมการเชิงเส้นนี้คือ b = 1/2 และเขียนสมการทั้งสามในรูปเวกเตอร์สดมภ์ได้ดังนี้

$$\begin{bmatrix} 1 \\ 3/2 \\ 2 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}$$

ระบบสมการเชิงเส้นที่ไม่สามารถหาผลลัพธ์ได้

- ระบบสมการเชิงเส้นที่ไม่สามารถหาผลลัพธ์ได้ หมายถึง สมการทุกสมการไม่สามารถเป็นจริงพร้อมๆ กันในเวลาเดียวกันเมื่อแทนค่าผลลัพธ์ในสมการเหล่านั้น เช่น

error

2b = 1 → E1 = $Y_i - \hat{Y}_i$ = 1-2b
3b = 4 → E2 = 4-3b
4b = 8 → E3 = 8-4b

- กรณีนี้ไม่มีผลลัพธ์ b ที่ทำให้ทุกสมการเป็นจริงได้พร้อมกันในเวลาเดียวกัน ดังนั้นต้องใช้**ค่า b ประนีประนอม** คือ**ค่า b ที่ทำให้ผลบวกความคลาดเคลื่อนกำลังสองมีค่าน้อยที่สุด**
- เรียก**วิธีเลือก b** ที่ทำให้ผลบวกความคลาดเคลื่อนกำลังสองมีค่าน้อยที่สุดว่าวิธี Least squares

$$\hat{Y}_i = a + bX_i$$

ระบบสมการเชิงเส้นที่ไม่สามารถหาผลลัพธ์ได้

- ให้ E^2 แทนผลบวกความคลาดเคลื่อนกำลังสอง

$E^2 = (1-2b)^2 + (4-3b)^2 + (8-4b)^2$

- ในกรณีไม่มีค่า b ที่ทำให้ทุกสมการเป็นจริงพร้อมกันในเวลาเดียวกัน ค่า E^2 เป็นฟังก์ชันพาราโบลาหงายที่มีจุดต่ำที่สุดที่ $dE^2/db = 0$
- ดังนั้น b = 46/29
- ผลลัพธ์ประนีประนอม คือ

$$\hat{Y} = \frac{46}{29}\begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 3.1724 \\ 4.7586 \\ 6.3448 \end{bmatrix}$$

- จะเห็นว่า $\hat{Y} \neq Y$ โดย $Y = \begin{bmatrix} 1 \\ 4 \\ 8 \end{bmatrix}$
- จะเห็นว่าไม่สามารถเขียน (X_i, Y_i) อยู่ในรูป $Y_i = a+bX_i$ แต่เขียนอยู่ในรูป $\hat{Y} = a + bX$ ได้

สมการถดถอยเชิงเส้นอย่างง่าย

- สมการถดถอยเชิงเส้นอย่างง่าย (Simple linear regression) มี 2 ตัวแปรคือ **ตัวแปรตาม**และ**ตัวแปรอิสระ** และเป็นกรณีที่**ตัวแปรอิสระมีเพียงตัวเดียว** รูปแบบดังนี้
- สมการถดถอยเชิงเส้นอย่างง่ายของ**ประชากร**คือ $Y = \alpha+\beta X+\varepsilon$
- สมการถดถอยเชิงเส้นอย่างง่ายของ**ตัวอย่างสุ่ม**คือ Y = a+bX+e
- Y คือ ตัวแปรตาม X คือ ตัวแปรอิสระ
- α และ a คือ จุดตัดแกน Y
- β และ b คือ ความชันของเส้นตรง ก็คือค่าสัมประสิทธิ์ของสมการ หรือเรียกว่า **สัมประสิทธิ์สหสัมพันธ์ถดถอยเชิงเส้น**
- ε (epsilon) และ e คือ ค่าความคลาดเคลื่อนของ Y เมื่อกำหนดค่า X

สมการถดถอยเชิงเส้นอย่างง่าย

- สมการถดถอยเชิงเส้นของ Y บน X : $\hat{Y}_i = a + bX_i$

(regression of y on x)

- Slope b

เอสใหญ่

$$b = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{S_{XY}}{S_X^2}$$

$$b = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})/(n-1)}{\sum(X_i - \bar{X})^2/(n-1)} = \frac{Cov(X,Y)}{s_x^2}$$

เอสเล็ก

- Intercept a

$$a = \bar{Y} - b\bar{X}$$

สมการถดถอยเชิงเส้นอย่างง่าย

- **สมการถดถอยเชิงเส้น** $\hat{Y}_i = a + bX_i$ ถูกเรียกอีกอย่างว่า “**สมการพยากรณ์**” ซึ่งสามารถนำมาทำนายค่าตัวแปรตาม Y หรือเรียกว่าค่าพยากรณ์ ($\hat{Y}_i$) เมื่อกำหนดค่าตัวแปรอิสระ X ได้
- โดย $b = \frac{S_{XY}}{S_X^2} = \frac{cov(X,Y)}{s_x^2} = \frac{r_{xy}s_y}{s_x}$

$$a = \bar{Y} - b\bar{X}$$

สมการถดถอยเชิงเส้นอย่างง่าย

- ในการหา**ความสัมพันธ์เชิงเส้น**ระหว่าง 2 ตัวแปรจาก**สมการถดถอยเชิงเส้น**จะต้องมีการกำหนดว่าจะให้ตัวแปรตัวใดเป็น**ตัวแปรตาม**หรือเป็น**ตัวแปรอิสระ**
- **ไม่มีทฤษฎีที่กล่าวถึงกฎเกณฑ์ตายตัว**ว่าจะให้ตัวแปรใดเป็นตัวแปรตามหรือเป็นตัวแปรอิสระ และบางกรณีแยกได้ยากว่าจะให้ตัวแปรไหนเป็นตัวแปรอิสระหรือตัวแปรตาม
- สมการถดถอยเชิงเส้นที่มี Y เป็นตัวแปรตามและมี X เป็นตัวแปรอิสระ (regression of y on x) จะไม่เหมือนกับสมการถดถอยเชิงเส้นที่มี X เป็นตัวแปรตามและมี Y เป็นตัวแปรอิสระ (regression of x on y)

สรุปสูตร

- Slope

$$b = \frac{S_{XY}}{S_X^2}$$

$$S_{XY} = \sum(X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_iY_i - \frac{(\sum X_i)(\sum Y_i)}{n}$$

$$S_X^2 = \sum(X_i - \bar{X})^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{n}$$

- Intercept

$$a = \bar{Y} - b\bar{X}$$

ตัวอย่าง 1

- จงหาสมการถดถอยเชิงเส้นของ Y บน X และค่าพยากรณ์ของข้อมูลดังนี้

X	-2	-1	0	1	2
Y	-3	-1	1	3	5

$$b = \frac{\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}} = \frac{20}{10} = 2$$

$$a = \bar{Y} - b\bar{X} = 1 - (2)(0) = 1$$

- สรุปสมการถดถอยเชิงเส้นของ Y บน X คือ

$$\hat{Y} = 1 + 2X$$

ตัวอย่าง 1

- จากสมการถดถอยเชิงเส้นที่ได้นำมาคำนวณค่าพยากรณ์ได้ดังนี้

$$\hat{Y}_i = 1 + 2X_i$$

X	-2	-1	0	1	2
Y	-3	-1	1	3	5
$\hat{Y}$	-3	-1	1	3	5

- จะเห็นว่า ถ้าสองตัวแปรมีความสัมพันธ์กันอย่างสมบูรณ์การพยากรณ์ค่า Y จะถูกต้อง 100%

ตัวอย่าง 2

- จงหาสมการถดถอยเชิงเส้นของ Y บน X และค่าพยากรณ์ของข้อมูลดังนี้

X	5	6	9	9	10
Y	9	11	0	15	1

$$b = \frac{\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}} = \frac{-24.8}{18.8} = -1.3191$$

$$a = \bar{Y} - b\bar{X} = 7.2 - (-1.3191)(7.8) = 17.4894$$

- สรุปสมการถดถอยเชิงเส้นของ Y บน X คือ

$$\hat{Y} = 17.4894 - 1.3191X$$

ตัวอย่าง 2

- จากสมการถดถอยเชิงเส้นที่ได้นำมาคำนวณค่าพยากรณ์ได้ดังนี้

$$\hat{Y} = 17.48936 - 1.31915X$$

X	5	6	9	9	10
Y	9	11	0	15	1
$\hat{Y}$	10.89	9.57	5.62	5.62	4.30

- จะเห็นว่า ถ้าสองตัวแปรมีความสัมพันธ์กันไม่สมบูรณ์การพยากรณ์ค่า Y จะไม่ถูกต้อง 100%

ความสัมพันธ์ระหว่างสัมประสิทธิ์สหสัมพันธ์และสัมประสิทธิ์สหสัมพันธ์ถดถอยเชิงเส้น

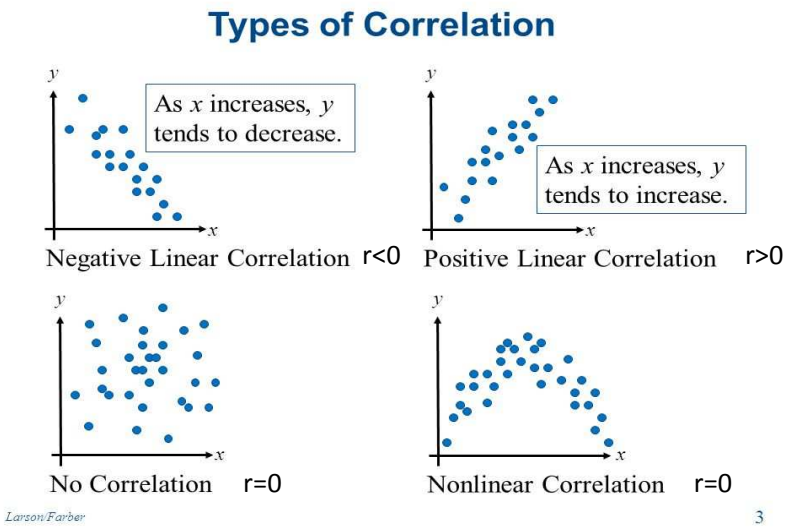
- จากสมการ $b_{YX} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{Cov(X,Y)}{s_x^2}$
- และ $r_{XY} = \frac{Cov(X,Y)}{s_x s_y} = r_{YX}$
- แทนค่า Cov(X,Y) หรือ S_{XY} จะได้ว่า

$$b_{YX} = r_{XY}\frac{s_y}{s_x}$$

ข้อแนะนำ

- ก่อนที่จะเริ่มทำการหาสมการถดถอยเชิงเส้นโดยวิธี least squares จะต้อง plot graph ก่อนเสมอ เพื่อศึกษาสิ่งต่อไปนี้
  - ดูความสัมพันธ์ว่าเป็นชนิด**เชิงเส้น**หรือไม่
    - ถ้าความสัมพันธ์ไม่เป็นเชิงเส้น เราอาจจะสามารถแปลงให้อยู่ในรูปความสัมพันธ์เชิงเส้นได้ เช่นแปลง Y ให้อยู่ในรูป ln Y และ/หรือแปลง X เป็น ln X
    - ถ้าแปลงแล้วความสัมพันธ์ไม่เป็นเชิงเส้นจริงๆ จะต้องดำเนินการหาสมการถดถอยโดยวิธี Nonlinear Regression เช่น Polynomial Regression
  - ดูการกระจายของจุดเป็น**เอกรูป**หรือไม่ (Homogeneous scatter) คือมีการกระจายของจุดสม่ำเสมอในกรอบเส้นขนาน

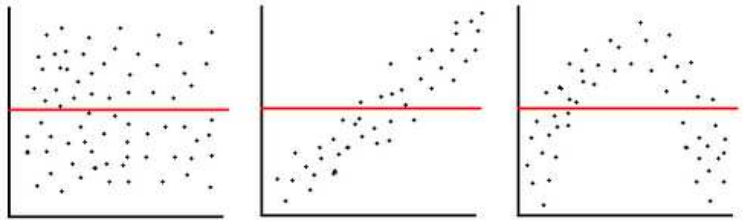
Linear Correlation (เชิงเส้น?)



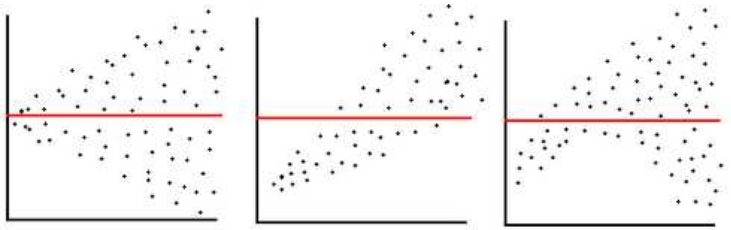
http://slideplayer.com/slide/5256836/

Homogeneous Scatter (เอกรูป?)

- Homogeneous scatter



- Heterogeneous scatter



https://www.r-bloggers.com/model-validation-interpreting-residual-plots/

คุณสมบัติสมการถดถอยเชิงเส้น

1. มาตราวัดของ**ตัวแปรตาม** Y จะต้องเป็นชนิดอันตรภาค (interval) หรือชนิดอัตราส่วน (ratio) ส่วนมาตราวัดของ**ตัวแปรอิสระ** X จะ**เป็นชนิดใดก็ได้**

—ถ้าหากมาตราวัดของตัวแปรตาม Y ไม่เป็นชนิดอันตรภาค หรือชนิดอัตราส่วน จะใช้วิธี least squares หาสมการถดถอยเชิงเส้นไม่ได้ จะต้องใช้วิธีอื่น

คุณสมบัติสมการถดถอยเชิงเส้น

2. สมการถดถอยเชิงเส้นจากตัวอย่างสุ่มผ่านจุด $(\bar{X}, \bar{Y})$ เสมอ เนื่องจาก

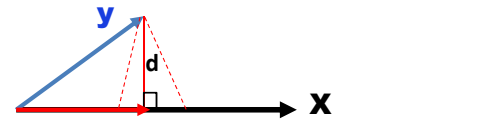
$$a = \bar{Y} - b\bar{X}$$
$$\bar{Y} = a + b\bar{X}$$

—คุณสมบัติข้อนี้ใช้ประโยชน์ในด้านต่างๆ ดังนี้

- ใช้พิสูจน์คุณสมบัติของเวคเตอร์ส่วนเหลือ (Residual vector)
- ใช้ในการวิเคราะห์ความแปรปรวน (ANOVA)

เวกเตอร์ส่วนเหลือ (Residual vector)

- การใช้วิธีฉาย (project) เวกเตอร์สดมภ์ **y** ลงในสเปซของเวกเตอร์สดมภ์ **X** เพื่อหาผลลัพธ์ b ประนีประนอม จะต้องฉายให้ระยะห่างจากเวกเตอร์สดมภ์ **y** มายังสเปซของเวกเตอร์สดมภ์ **X** น้อยที่สุด ตามแนวคิดของ least squares ดังนั้น จาก**ทฤษฎีทางเรขาคณิต** ระยะห่างจากจุดๆ มายังเส้นตรง ระยะห่างที่น้อยที่สุดคือ ระยะตั้งฉากกับสเปซของเวกเตอร์สดมภ์ **X**



Projection of **y** = $\hat{Y}$ = **X** $[\mathbf{X}^T\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{y}$

- **ระยะห่าง**คือ**เวกเตอร์ส่วนเหลือ d** (Residual vector) ซึ่งมีค่าเท่ากับ $Y_i - \hat{Y}_i$ จำนวน n ค่า

คุณสมบัติของเวกเตอร์ส่วนเหลือ

- จากทฤษฎีเวกเตอร์ การที่เวกเตอร์ส่วนเหลือ **d** ตั้งฉากกับสเปซของเวกเตอร์สดมภ์ **X** ทำให้ได้ว่าแต่ละเวกเตอร์สดมภ์ในสเปซของ **X** เป็นอิสระกับเวกเตอร์ส่วนเหลือ ดังนั้น**คุณสมบัติของเวกเตอร์ส่วนเหลือ** ได้แก่

1. ผลบวกของสมาชิกของเวกเตอร์ส่วนเหลือเป็นศูนย์
2. เวกเตอร์ของตัวแปรอิสระ (**X**) เป็นอิสระกับเวกเตอร์ส่วนเหลือ (**d**)

คุณสมบัติของเวกเตอร์ส่วนเหลือ

1. ผลบวกของสมาชิกของเวกเตอร์ส่วนเหลือ (**d**) เป็นศูนย์

$$\sum(Y_i - \hat{Y}_i) = 0 \qquad \sum \mathbf{d}_i = 0$$

- ค่า 0 หมายความว่า **0 เป็นศูนย์กลางของการกระจายของค่าความคลาดเคลื่อน**ที่เกิดจากประมาณ Y_i ด้วย $\hat{Y}_i$
  - ถ้า $\hat{Y}_i$ มีค่ามากกว่า Y_i ค่า $Y_i - \hat{Y}_i$ จะมีค่าเป็นลบ
  - ถ้า $\hat{Y}_i$ มีค่าน้อยกว่า Y_i ค่า $Y_i - \hat{Y}_i$ จะมีค่าเป็นบวก

คุณสมบัติของเวกเตอร์ส่วนเหลือ

2. เวกเตอร์ของตัวแปรอิสระ (**X**) เป็นอิสระกับเวกเตอร์ส่วนเหลือ (**d**)

$$\sum X_i(Y_i - \hat{Y}_i) = 0 \qquad \sum X_i \mathbf{d}_i = 0$$

- เวกเตอร์ส่วนเหลือตั้งฉากกับสเปซของเวกเตอร์สดมภ์ **X** จากทฤษฎีของเวกเตอร์ถ้า 2 เวกเตอร์ตั้งฉากกันจะได้ว่า**ผลบวกของผลคูณของแต่ละสมาชิกเป็นศูนย์**

ตัวอย่าง 3

- จงหาเวกเตอร์ส่วนเหลือ และ**ทดสอบคุณสมบัติของเวกเตอร์ส่วนเหลือ**ของข้อมูลดังนี้

X	-2	-1	0	1	2
Y	-3	-1	1	3	5

$r_{XY} = 1$

- จากตัวอย่างก่อนหน้าหาสมการถดถอยเชิงเส้นของ Y บน X ได้คือ

$$\hat{Y} = 1 + 2X$$

- ดังนั้นหา $\hat{Y}$ ได้คือ

X	-2	-1	0	1	2
Y	-3	-1	1	3	5
$\hat{Y}$	-3	-1	1	3	5

ตัวอย่าง 3

- หาเวกเตอร์ส่วนเหลือได้ดังนี้

$$\mathbf{d} = Y_i - \hat{Y}_i$$

X	-2	-1	0	1	2
Y	-3	-1	1	3	5
$\hat{Y}$	-3	-1	1	3	5
d	0	0	0	0	0
X*d	0	0	0	0	0

- ผลบวกของสมาชิกของเวกเตอร์ส่วนเหลือเป็นศูนย์ $\sum \mathbf{d}_i = 0$
- เวกเตอร์ของตัวแปรอิสระเป็นอิสระกับเวกเตอร์ส่วนเหลือ $\sum X_i \mathbf{d}_i = 0$

ตัวอย่าง 4

- จงหาเวกเตอร์ส่วนเหลือ และทดสอบคุณสมบัติของเวกเตอร์ส่วนเหลือของข้อมูลดังนี้

X	5	6	9	9	10
Y	9	11	0	15	1

$r_{XY} = -0.44$

- จากตัวอย่างก่อนหน้าหาสมการถดถอยเชิงเส้นของ Y บน X ได้คือ

$$\hat{Y} = 17.4894 - 1.3191X$$

- ดังนั้นหา $\hat{Y}$ ได้คือ

X	5	6	9	9	10
Y	9	11	0	15	1
$\hat{Y}$	10.89	9.57	5.62	5.62	4.30

ตัวอย่าง 4

- หาเวกเตอร์ส่วนเหลือได้ดังนี้

$$\mathbf{d} = Y_i - \hat{Y}_i$$

X	5	6	9	9	10
Y	9	11	0	15	1
$\hat{Y}$	10.89	9.57	5.62	5.62	4.30
d	-1.89	1.43	-5.62	9.38	-3.298
X*d	-9.47	8.58	-50.55	84.44	-32.98

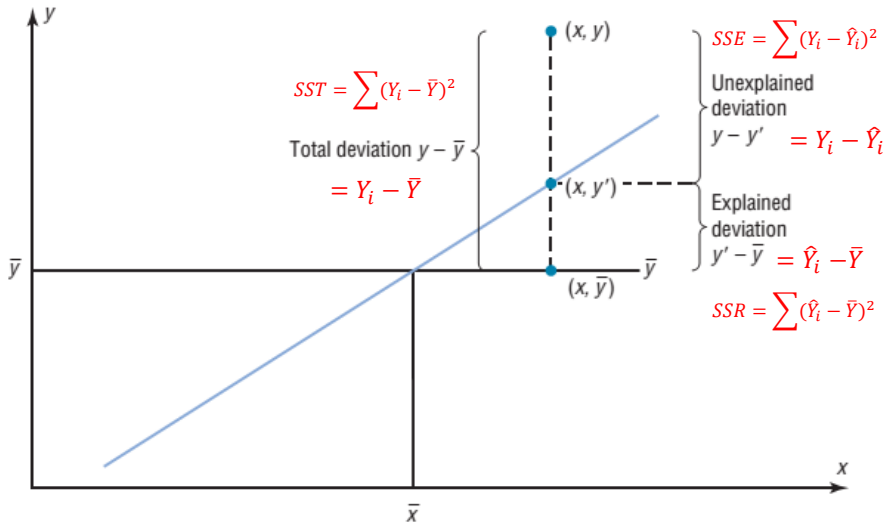
- ผลบวกของสมาชิกของเวกเตอร์ส่วนเหลือเป็นศูนย์ $\sum \mathbf{d}_i = 0.002$
- เวกเตอร์ของตัวแปรอิสระเป็นอิสระกับเวกเตอร์ส่วนเหลือ $\sum X_i \mathbf{d}_i = 0.02$

จาก 2 ตัวอย่าง

- จะเห็นว่า**ถ้า X และ Y ไม่มีความสัมพันธ์กันอย่างสมบูรณ์** ข้อสนเทศ X_i จะถูกเก็บไว้ใน $\hat{Y}_i$ ไม่หมด 100% **ข้อสนเทศของ X_i ส่วนที่เหลือจะถูกเก็บไว้ในเวกเตอร์ส่วนเหลือ**

- ยิ่งค่า r เล็ก ข้อสนเทศของ X_i จะถูกเก็บไว้ใน $\hat{Y}_i$ น้อยลง และส่วนเหลือจาก 100% จะมีมากขึ้น ข้อสนเทศของ X_i จะถูกเก็บไว้ในเวกเตอร์ส่วนเหลือมากขึ้น **เป็นผลให้คุณสมบัติของเวกเตอร์ส่วนเหลือคลาดเคลื่อนจากทฤษฎีได้**

การวิเคราะห์ความแปรปรวน



การวิเคราะห์ความแปรปรวน

- จากกราฟเขียนสมการได้ดังนี้ $Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$ โดย
  - $Y_i - \bar{Y}$ = ค่าเบี่ยงเบนของ Y_i จากค่าเฉลี่ยเลขคณิต $\bar{Y}$
  - $\sum(Y_i - \bar{Y})^2$ = ค่า**ความแปรปรวนรวม**ของ Y_i จำนวน n ค่า = Total sum of squares = SST
  - $\hat{Y}_i - \bar{Y}$ = ค่าเบี่ยงเบนของค่าประมาณ (ค่าพยากรณ์) ที่ i ($\hat{Y}_i$) จากค่าเฉลี่ยเลขคณิต $\bar{Y}$
  - $\sum(\hat{Y}_i - \bar{Y})^2$ = ค่า**ความแปรปรวนถดถอย**ของ $\hat{Y}_i$ จำนวน n ค่า = Regression sum of squares = SSR
  - $Y_i - \hat{Y}_i$ = ค่าส่วนเหลือจากการประมาณ Y_i ด้วย $\hat{Y}_i$
  - $\sum(Y_i - \hat{Y}_i)^2$ = ค่า**ส่วนเหลือกำลังสอง** = Residual (Error) sum of squares = SSE

การวิเคราะห์ความแปรปรวน

- การวิเคราะห์ความแปรปรวน (Analysis of Variance: ANOVA) อาศัยฐานความรู้เกี่ยวกับคุณสมบัติของสมการถดถอยเชิงเส้นที่ผ่านจุด $(\bar{X}, \bar{Y})$ เสมอ และคุณสมบัติของความเป็นอิสระระหว่างเวกเตอร์ส่วนเหลือกับเวกเตอร์สดมภ์ X
- สำหรับสมการถดถอยเชิงเส้นที่สร้างขึ้นมาจากแนวความคิด least squares จะได้ว่า ค่าความแปรปรวนรวม (SST) จะเท่ากับผลบวกของค่าความแปรปรวนถดถอย (SSR) และค่าส่วนเหลือกำลังสอง (SSE) เสมอ

$$\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2$$

SST = SSR + SSE

- หมายเหตุ: จะเห็นว่า SSE = SST-SSR

ตารางวิเคราะห์ความแปรปรวน (ANOVA)

- ตารางวิเคราะห์ความแปรปรวน (ANOVA)

แหล่งความแปรปรวน	องศาอิสระ	SS	MS	F
ถดถอย	1	**SSR**	**MSR** =SSR/1	**F**=MSR/MSE
ส่วนเหลือ	n-2	**SSE** =SST-SSR	**MSE** =SSE/(n-2)	
รวม	n-1	SST		

MSE = Mean square error MSR = Mean square regression

- F เป็นค่าสถิติจากการคำนวณที่ใช้สำหรับทดสอบสมมติฐานสัมประสิทธิ์สหสัมพันธ์ถดถอยเชิงเส้นของประชากร $H_0: \beta = 0$, $H_1: \beta \neq 0$

การหาค่าความแปรปรวน SST และ SSR

- สูตรหาความแปรปรวนรวม SST ได้แก่

$$SST = S_Y^2$$

- สูตรหาความแปรปรวนถดถอย SSR หาได้ 3 สูตร ได้แก่

1. หาจากค่าสัมประสิทธิ์สหสัมพันธ์ถดถอยเชิงเส้น (Regression coefficient, b)

$$SSR = b^2 S_X^2$$

2. หาจากค่าสัมประสิทธิ์สหสัมพันธ์ถดถอยเชิงเส้น (Regression coefficient, b)

$$SSR = bS_{XY}$$

3. หาจากสัมประสิทธิ์สหสัมพันธ์เชิงเส้น (Correlation coefficient, r)

$$SSR = r^2 S_Y^2 = r^2 SST$$

Recall

- สูตรคำนวณ

$$S_X^2 = \sum (X_i - \bar{X})^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{n}$$

$$S_Y^2 = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n}$$

$$S_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}$$

สัมประสิทธิ์แห่งการกำหนด, r^2

- จากสมการ $SSR = r^2 SST$
- จะได้สมการ $\frac{SSR}{SST} = r^2$
- ให้ข้อมูลดังนี้
  - ค่า r^2 เป็นร้อยละ (สัดส่วน) ที่ข้อสนเทศ (information) ของตัวแปรอิสระ X ที่ถูกเก็บไว้ในค่าประมาณ $\hat{Y}_i$ หาก r ≠ +1 หรือ -1 ย่อมมีความแปรปรวนในการประมาณ Y_i ดังนั้น r^2 จึงเป็น**ร้อยละที่ข้อสนเทศของตัวแปรอิสระ X อธิบายความแปรปรวนของค่าของตัวแปรตาม Y** เรียก r^2 ว่า**สัมประสิทธิ์แห่งการกำหนด** (Coefficient of Determination)

สัมประสิทธิ์แห่งการกำหนดกับส่วนเหลือกำลังสอง

- จากสมการ $SSE = SST - SSR$
- จะได้สมการ $\frac{SSE}{SST} = 1 - r^2$
- ให้ข้อมูลดังนี้
  - $(1-r^2)$ บอกถึงร้อยละ (สัดส่วน) ของ**ข้อสนเทศของตัวแปรอิสระ** X **ที่ไม่ถูกเก็บไว้ในค่าประมาณ** $\hat{Y}$ หรือ
  - $(1-r^2)$ บอกถึงร้อยละ (สัดส่วน) ของส่วนเหลือของข้อสนเทศของตัวแปรอิสระ X ที่**ถูกเก็บไว้ในเวคเตอร์ส่วนเหลือ d**
  - $(1-r^2)$ จึงเป็นร้อยละ (สัดส่วน) ของข้อสนเทศของตัวแปรอิสระ X **ที่ไม่สามารถอธิบายความแปรปรวนของค่าตัวแปรตาม** Y

สัมประสิทธิ์แห่งการกำหนดกับส่วนเหลือกำลังสอง

- ให้ข้อมูลดังนี้ (ต่อ)
  - เนื่องจาก SSE มีค่าเป็นบวกเสมอและมีค่าต่ำสุดเป็นศูนย์ ซึ่งถ้า SSE=0 จะเป็นกรณีที่ตัวแปรอิสระ X และตัวแปรตาม Y มีความสัมพันธ์เชิงเส้นอย่างสมบูรณ์ ดังนั้น
    - ขอบเขตของ r^2 จึงอยู่ในช่วงปิด [0,1]
    - ขอบเขตของ r จึงอยู่ในช่วงปิด [-1,1]
  - ถ้า r^2 มีค่าสูง $(1-r^2)$ จะมีค่าต่ำ ซึ่งทำให้ SSE มีค่าต่ำด้วย การที่ SSE มีค่าต่ำหมายความว่า ค่าประมาณ (ค่าพยากรณ์) $\hat{Y}_i$ มีความแม่นยำสูงในการประมาณ (พยากรณ์) Y_i

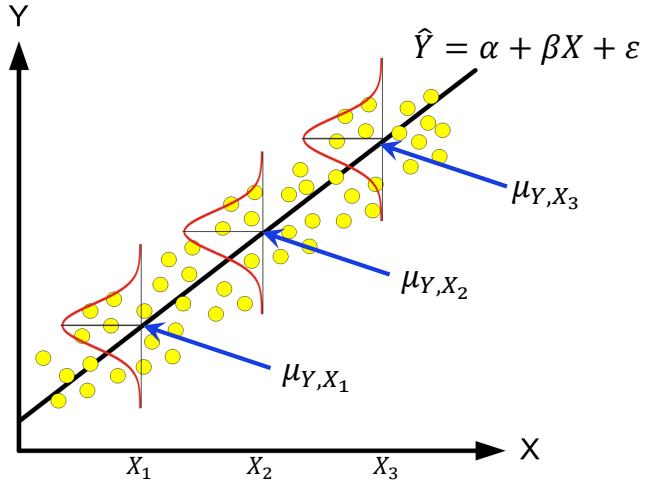
ข้อสมมติของรูปแบบสมการถดถอยเชิงเส้น

- ในที่นี้ศึกษาเฉพาะกรณีที่การเปลี่ยนแปลงค่าของตัวแปรอิสระมีช่วงเท่ากัน (equally spacing) กล่าวคือ**ไม่มีความคลาดเคลื่อนในการวัดตัวแปรอิสระ** X

  — หากเป็นกรณีที่ต้องพิจารณาความคลาดเคลื่อนในการวัดค่าของตัวแปรอิสระรูปแบบของสมการถดถอยจะเป็นอีกอย่างหนึ่ง รายละเอียดอ่านจาก: Snedecor & Cochran ใน “Statistical Method” และ Berkoon ใน “J.Am.Stat.Assoc.” 45(1950): p.164

ข้อสมมติของรูปแบบสมการถดถอยเชิงเส้น

- กราฟแจกแจงแบบปกติของตัวแปรตาม Y สำหรับแต่ละค่าของตัวแปรอิสระ X ซึ่งมีช่วงห่างเท่าๆกัน



ข้อสมมติที่ 1

$$Y \approx N(\mu_{Y,X}, \sigma^2_{Y,X})$$

- สำหรับแต่ละค่าของตัวแปรอิสระ X จะ**มีค่าของตัวแปรตาม Y ได้หลายค่า** ดังนั้นค่าต่างๆ ของตัวแปรตาม Y (ซึ่งก็คือ Y_i) จะกระจายเป็นรูปโค้งระฆังคว่ำคือ มี**การกระจายแบบปกติ** โดยมี**ค่าเฉลี่ย**อยู่ที่จุด $\mu_{Y,X}$ ซึ่งค่าเฉลี่ยนี้ตั้งอยู่บนสมการถดถอยเชิงเส้น $\hat{Y} = \alpha + \beta X + \varepsilon$
- การแจกแจงค่าต่างๆ ของ Y_i เป็นอิสระกัน คือ Y_i เป็นอิสระกับ Y_j : $\forall i \neq j$ ที่เกิดขึ้นจากแต่ละค่าของ X
- **ความแปรปรวน**ของตัวแปรตาม Y เท่ากันสำหรับแต่ละระดับของตัวแปรอิสระ X นั่นคือ $\sigma^2_{Y,X_1} = \sigma^2_{Y,X_2} = \cdots = \sigma^2_{Y,X_n} = \sigma^2_{Y,X}$
- ดังนั้นคุณสมบัตินี้จึง**ทำให้การกระจายจุดเป็นเอกรูป**ในกรอบเส้นคู่ขนาน (Homogeneous scatter)
- หมายเหตุ: ตัวแปรอิสระ X ต้องมีช่วงห่างเท่ากัน หากการเปลี่ยนค่าของตัวแปรอิสระ X มีช่วงห่างไม่เท่ากัน วิธีสร้างสมการถดถอยเชิงเส้นที่กล่าวในบทนี้ไม่เหมาะสม ต้องใช้วิธีอื่น

ข้อสมมติที่ 2

$$\varepsilon \approx N(0, \sigma^2_{Y,X})$$

- ε_i เป็น**ค่าความคลาดเคลื่อน**ในการประมาณค่า Y_i เนื่องจาก ε มีความสัมพันธ์เชิงเส้นกับ Y จากรูปแบบสมการถดถอยเชิงเส้นของประชากร $\hat{Y} = \alpha + \beta X + \varepsilon$ ดังนั้น การแจกแจงของ ε เหมือนกับของ Y กล่าวคือมี**การแจกแจงแบบปกติที่มีค่าเฉลี่ยศูนย์** (จากคุณสมบัติของเวคเตอร์ส่วนเหลือ) และมี**ค่าความแปรปรวนเหมือนกับของ Y** คือ ความแปรปรวนของ ε เท่ากันสำหรับแต่ละระดับของตัวแปรอิสระ X
- คุณสมบัติการแจกแจงปกติของ ε มีความสำคัญในการทดสอบสมมติฐาน และในการ**หาช่วงความเชื่อมั่นของ α (intercept) และ β (slope)** เนื่องจากค่าเฉลี่ยกำลังสองของ ε เป็นตัวส่วน (ตัวหาร, denominator) ของตัวสถิติ F (นั่นคือ F = MSR/MSE) และที่มาของตัวสถิติ F มาจากตัวแปรเชิงสุ่มที่มีการแจกแจงแบบปกติ

ข้อสมมติที่ 3

- การแจกแจงค่าต่างๆ ของ ε_i เป็นอิสระกัน คือ ε_i เป็นอิสระกับ ε_j $\forall i \neq j$ ที่เกิดขึ้นจากแต่ละค่าของ X เนื่องจาก การแจกแจงของ Yi เป็นอิสระกัน และ ε มีความสัมพันธ์เชิงเส้นกับ Y จากรูปแบบสมการถดถอยเชิงเส้นของประชากร ดังนั้นจึงทำให้ ε_i มีการแจกแจงที่ป็นอิสระไปด้วย
- คุณสมบัติข้อนี้ทำให้สามารถวิเคราะห์ข้อมูล**หาการแจกแจงความน่าจะเป็นของตัวสถิติ b (slope)** จากตัวแปรที่มีความสัมพันธ์เชิงเส้นและมีการกระจายสม่ำเสมอในกรอบเส้นขนาน (เอกรูป) โดยใช้สมการถดถอยเชิงเส้นได้
- หมายเหตุ: ถ้า ε_i ไม่เป็นอิสระกับ ε_j กล่าวคือมี Auto-Correlation จะวิเคราะห์ความสัมพันธ์โดยวิธีที่เรียกว่า การวิเคราะห์อนุกรมเวลา (Time series analysis)

ข้อสมมติที่ 4

- เวคเตอร์สดมภ์ X **ตั้งฉาก**และ**เป็นอิสระ**กับเวคเตอร์ส่วนเหลือ ε ซึ่ง
  - คุณสมบัติการ**ตั้งฉาก**มาจากแนวคิดของ least squares ที่ว่า เวคเตอร์ส่วนเหลือมีระยะห่างจากสเปซของเวคเตอร์สดมภ์ X น้อยที่สุด (จากทฤษฎีของเวคเตอร์ ระยะห่างน้อยที่สุดคือ ระยะตั้งฉาก)
  - คุณสมบัติการ**เป็นอิสระ**มาจากผลของการตั้งฉากกันของ 2 เวคเตอร์
- คุณสมบัติข้อนี้ทำให้ได้ SST = SSR+SSE ซึ่งนำไปสู่การสร้าง**ตารางวิเคราะห์ความแปรปรวน**เพื่อใช้**หาค่าสถิติ F** ที่ใช้ทดสอบสมมติฐานว่าตัวแปรทั้งสองมีความสัมพันธ์หรือไม่โดยผ่านสมการถดถอยเชิงเส้น (นั่นคือทดสอบค่า β)
- **สรุป** ตั้งฉาก => อิสระ => ไม่มีความสัมพันธ์เชิงเส้น

การแจกแจงความน่าจะเป็นของตัวสถิติจากตัวอย่างสุ่ม

- สมการถดถอยเชิงเส้นของประชากรเป็นดังนี้

$$Y = \alpha + \beta X + \varepsilon$$

- การแจกแจงความน่าจะเป็นของ b

$$b \sim N(\beta, \frac{\sigma_{yx}^2}{\sum(X_i - \bar{X})^2})$$

- การแจกแจงความน่าจะเป็นของ a

$$a \sim N(\alpha, \sigma_{yx}^2(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2})$$

การทดสอบสมมติฐาน

- หลังจากที่ได้สมการถดถอยเชิงเส้นแล้ว ก่อนที่จะนำสมการถดถอยเชิงเส้นไปใช้ประมาณ (พยากรณ์/ทำนาย) ค่าตัวแปรตาม Y ควรมีการตรวจสอบทางสถิติโดยการทดสอบสมมติฐาน 2 ข้อ ได้แก่

  1. สมการถดถอยเชิงเส้นมีความเหมาะสมสามารถนำไปประมาณค่าของตัวแปรตามได้หรือไม่ โดย**การอนุมานเกี่ยวกับ β**
  2. สมการถดถอยเชิงเส้นผ่านจุดกำเนิดหรือไม่ โดย**การอนุมานเกี่ยวกับ α**

การอนุมานเกี่ยวกับ β

- เป็นการทดสอบสมมติฐานว่า**สมการถดถอยเชิงเส้นที่ได้มีความเหมาะสมที่จะนำไปประมาณค่าของตัวแปรตามหรือไม่**

- การอนุมานเกี่ยวกับ β ตั้งสมมติฐานดังนี้

  H_0: $\beta = 0$ (สมการไม่เหมาะสมสำหรับทำนาย y)

  H_1: $\beta \neq 0$ (สมการเหมาะสมสำหรับทำนาย y)

การอนุมานเกี่ยวกับ β

- การทดสอบสมมติฐานเกี่ยวกับ β
- ค่าสถิติที่ใช้ทดสอบสมมติฐานมี 2 ตัวคือ
  - ค่าสถิติ F
  - ค่าสถิติ t

การอนุมานเกี่ยวกับ β ด้วยค่าสถิติ F

- ตารางวิเคราะห์ความแปรปรวน (ANOVA)

แหล่งความแปรปรวน	องศาอิสระ	SS	MS	F
ถดถอย	1	SSR	MSR =SSR/1	F= MSR/MSE
ส่วนเหลือ	n-2	SSE	MSE =SSE/(n-2)	
รวม	n-1	SST		

- F เป็นค่าสถิติจากการคำนวณที่ใช้สำหรับทดสอบสมมติฐานสัมประสิทธิ์สหสัมพันธ์ถดถอยเชิงเส้นของประชากร H_0: $\beta = 0$, H_1: $\beta \neq 0$

การอนุมานเกี่ยวกับ β ด้วยค่าสถิติ F

- จากตารางวิเคราะห์ความแปรปรวน (ANOVA)
- ค่า SSR หาจาก
  - จากค่า b

    $$SSR = bS_{XY} = b^2S_X^2$$

  - จากค่า r

    $$SSR = r^2S_Y^2 = r^2SST$$

การอนุมานเกี่ยวกับ β ด้วยค่าสถิติ F

- ค่า SST หาจาก

$$SST = S_Y^2$$

- ค่า SSE หาจาก

$$SSE = SST - SSR = (1-r)^2 SST$$

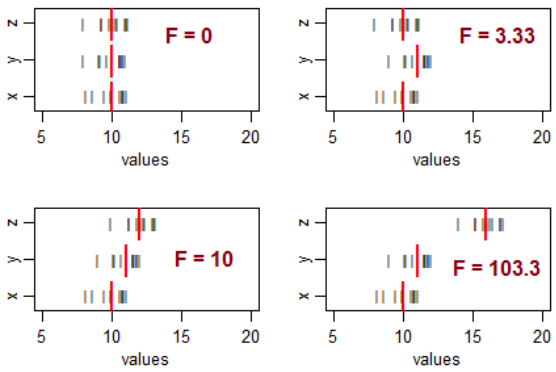
การอนุมานเกี่ยวกับ β ด้วยค่าสถิติ F

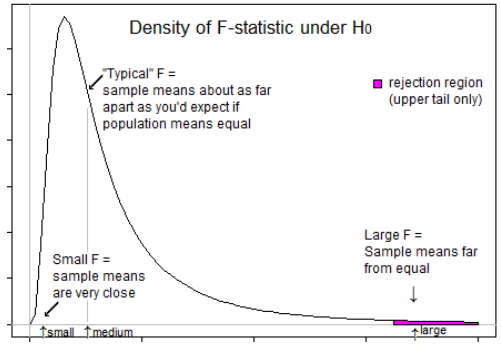
- สำหรับค่าสถิติ F:
- Decision rule คือ

  ปฏิเสธ H_0 ที่ระดับนัยสำคัญ α% ถ้า $|F| \geq F_{(\alpha,1,n-2)}$

หมายเหตุ

- ในการทดสอบ ANOVA (ทดสอบ equality of means) จะปฏิเสธ H_0 ที่ระดับนัยสำคัญ α% ถ้า $|F| \geq F_{(\alpha,n1,n2)}$ คือตรวจสอบแค่เพียงข้างเดียวคือข้างค่ามาก ถึงแม้สมมติฐานจะเป็นแบบ two-tailed
- ในการทดสอบ equality of variances จะปฏิเสธ H_0 ที่ระดับนัยสำคัญ α% เมื่อ $|F| \geq F_{(\alpha/2,n1,n2)}$ คือต้องตรวจสอบทั้งสองข้างในกรณีสมมติฐานเป็นแบบ two-tailed
- ในการทดสอบ β (ก็คือการทดสอบความแปรปรวนจากค่า mean ของสมการเชิงเส้นซึ่งก็คือทดสอบ equality of means นั่นเอง) จุดอ่อนของ F-test คือ ผู้วิจัยจะไม่ทราบเครื่องหมายของ b ว่าเป็นบวกหรือลบ ดังนั้น F-test ใช้ได้เฉพาะกรณีสมมติฐานทางเลือกเป็นชนิดสองทางเท่านั้น หากเป็นชนิดทางเดียวจะต้องใช้ t-test





การอนุมานเกี่ยวกับ β ด้วยค่าสถิติ t

- ค่าสถิติที่ใช้ทดสอบสมมติฐานเป็นดังนี้

$$t_{\left(\frac{\alpha}{2},n-2\right)} = \frac{b}{\sqrt{\frac{MSE}{S_X^2}}}$$

$$df = n - 2$$

$$MSE = \frac{S_Y^2 - bS_{XY}}{n-2}$$

$$S_X^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{n}$$

$$S_Y^2 = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n}$$

$$S_{XY} = \sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}$$

การอนุมานเกี่ยวกับ β ด้วยค่าสถิติ t

- ช่วงความเชื่อมั่น (1- α)100% ของ β เป็นดังนี้

$$b - t_{\left(\frac{\alpha}{2},n-2\right)}\sqrt{\frac{MSE}{S_X^2}} < \beta < b + t_{\left(\frac{\alpha}{2},n-2\right)}\sqrt{\frac{MSE}{S_X^2}}$$

ความสัมพันธ์ระหว่าง $F_{(1,n-2)}$ และ $t_{(n-2)}$

$$F_{(\alpha,1,n-2)} = t^2_{\left(\frac{\alpha}{2},n-2\right)} = \frac{MSR}{MSE} = \frac{r^2(n-2)}{1-r^2}$$

- เนื่องจาก

$$t = \frac{b}{\sqrt{MSE/S_X^2}} \qquad t^2 = \frac{b^2 S_X^2}{MSE} = \frac{SSR}{MSE} = \frac{MSR}{MSE}$$

$$MSR = \frac{SSR}{1} = b^2 S_X^2 = r^2 S_Y^2 \qquad MSE = \frac{SSE}{n-2} = \frac{(1-r^2)S_Y^2}{n-2}$$

ความสัมพันธ์ระหว่าง $F_{(1,n-2)}$ และ $t_{(n-2)}$

- จะเห็นว่า **ในกรณีที่ไม่มีตาราง F ก็สามารถใช้ตาราง t แทนได้** โดยการยกกำลังสองของค่า t ความสัมพันธ์ระหว่างค่าสถิติ F และค่าสถิติ t ตามสมการดังกล่าว**เป็นจริงเฉพาะกรณีที่องศาอิสระตัวแรกของค่าสถิติ F เป็น 1** เท่านั้นกล่าวคือ เป็น**กรณีที่มีตัวแปรอิสระ 1 ตัว**

- สมการยังแสดงความสอดคล้องกันของการทดสอบสมมติฐาน H_0: β = 0 และ H_0: ρ = 0 กล่าวคือ
  - ถ้าไม่ปฏิเสธ H_0: β = 0 ก็จะไม่ปฏิเสธ H_0: ρ = 0
  - ถ้าปฏิเสธ H_0: β = 0 ก็จะปฏิเสธ H_0: ρ = 0

ตัวอย่าง 5

- จากข้อมูลดังนี้ จงทดสอบสมมติฐานว่าสมการถดถอยเชิงเส้นเหมาะสมหรือไม่ ที่ระดับนัยสำคัญ 0.05

X	-2	-1	0	1	2
Y	-3	-1	1	3	5

- สมมติฐานคือ

  H_0: $\beta = 0$ (สมการไม่เหมาะสม)

  H_1: $\beta \neq 0$ (สมการเหมาะสม)

ตัวอย่าง 5

- **กรณีใช้ค่าสถิติ F ทดสอบสมมติฐาน** H_0: $\beta = 0$
- จะใช้ตาราง ANOVA คำนวณ ดังนี้

แหล่งความแปรปรวน	องศาอิสระ	SS	MS	F
ถดถอย	1	**SSR**	**MSR** =SSR/1	**F**= MSR/MSE
ส่วนเหลือ	n-2	**SSE**	**MSE** =SSE/(n-2)	$F = \frac{r^2(n-2)}{1-r^2}$
รวม	n-1	SST		

ตัวอย่าง 5

- ตาราง ANOVA กรณี SSR คำนวณจากค่า r

แหล่งความแปรปรวน	องศาอิสระ	SS	MS	F
ถดถอย	1	$SSR = r^2 S_Y^2 = 1(40) = 40$	$MSR = \frac{40}{1} = 40$	$F = \frac{40}{0} = \infty$
ส่วนเหลือ	5-2=3	$SSE = 40 - 40 = 0$	$MSE = \frac{0}{5-2} = 0$	$F = \frac{1(5-2)}{1-1} = \infty$
รวม	5-1=4	$SST = S_Y^2 = 40$		

- จากตารางสถิติ $F_{[0.05,1,3]}$ = 10.13
- พบว่า $|F| \geq F_{[0.05,1,3]}$
- สรุปว่า ปฏิเสธ H_0 ที่ α=0.05
- แปลว่า จากข้อมูลที่ได้จาก ต.ย.ขนาด 5 อนุมานได้ว่า สมการถดถอยเชิงเส้นมีความเหมาะสมที่จะนำไปประมาณค่าตัวแปรตาม Y

ตัวอย่าง 5

- **กรณีใช้ค่าสถิติ t ทดสอบสมมติฐาน** H_0: $\beta = 0$

$$t = \frac{b}{\sqrt{\frac{MSE}{S_X^2}}} = \frac{2}{\sqrt{0/10}} = \infty$$

- จากตารางสถิติ $t_{[0.025,3]}$ = 3.182
- พบว่า $|t| \geq t_{[0.025,3]}$
- สรุปว่า ปฏิเสธ H_0 ที่ α=0.05
- แปลว่า จากข้อมูลที่ได้จาก ต.ย.ขนาด 5 อนุมานได้ว่า สมการถดถอยเชิงเส้นมีความเหมาะสมที่จะนำไปประมาณค่าตัวแปรตาม Y

ตัวอย่าง 5

- **กรณีใช้ค่าสถิติ t ทดสอบสมมติฐาน** $H_0 : \rho = 0$

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 1\sqrt{\frac{5-2}{1-1}} = \infty$$

- จากตารางสถิติ $t_{[0.025,3]} = 3.182$
- พบว่า $|t| \geq t_{[0.025,3]}$
- สรุปว่า ปฏิเสธ H_0 ที่ α=0.05
- แปลว่า จากข้อมูลที่ได้จาก ต.ย.ขนาด 5 อนุมานได้ว่า ตัวแปรทั้งสองตัวมีความสัมพันธ์เชิงเส้น

ตัวอย่าง 6

- จากข้อมูลดังนี้ จงทดสอบสมมติฐานว่าสมการถดถอยเชิงเส้นเหมาะสมหรือไม่ ที่ระดับนัยสำคัญ 0.05

X	5	6	9	9	10
Y	9	11	0	15	1

- สมมติฐานคือ

  H_0: $\beta = 0$ (สมการไม่เหมาะสม)

  H_1: $\beta \neq 0$ (สมการเหมาะสม)

ตัวอย่าง 6

- **กรณีใช้ค่าสถิติ F ทดสอบสมมติฐาน** H_0: $\beta = 0$
- จะใช้ตาราง ANOVA คำนวณ ดังนี้

แหล่งความแปรปรวน	องศาอิสระ	SS	MS	F
ถดถอย	1	SSR	MSR=SSR/1	MSR/MSE
ส่วนเหลือ	n-2	SSE	MSE=SSE/(n-2)	
รวม	n-1	SST		

ตัวอย่าง 6

- ตาราง ANOVA กรณีค่า SSR คำนวณจากค่า r

แหล่งความแปรปรวน	องศาอิสระ	SS	MS	F
ถดถอย	1	$SSR = r^2 S_Y^2 = 32.7149$	$MSR = 32.7149$	$F = \frac{32.7149}{45.3617} = 0.7212$
ส่วนเหลือ	5-2=3	$SSE = 136.0851$	$MSE = 45.3617$	
รวม	5-1=4	$SST = S_Y^2 = 168.8$		

- จากตารางสถิติ $F_{[0.05,1,3]} = 10.13$
- พบว่า $|F| < F_{[0.05,1,3]}$
- สรุปว่า ไม่ปฏิเสธ H_0 ที่ α=0.05
- แปลว่า จากข้อมูลที่ได้จาก ต.ย.ขนาด 5 อนุมานได้ว่า สมการถดถอยเชิงเส้นไม่เหมาะสมที่จะนำไปประมาณค่าตัวแปรตาม Y

ตัวอย่าง 6

- ตาราง ANOVA กรณีค่า SSR คำนวณจากค่า b

แหล่งความแปรปรวน	องศาอิสระ	SS	MS	F
ถดถอย	1	$SSR = b^2 S_X^2 = 32.7149$	$MSR = 32.7149$	$F = \frac{32.7149}{45.3617}$
ส่วนเหลือ	5-2=3	$SSE = 136.0851$	$MSE = 45.3617$	$= 0.7212$
รวม	5-1=4	$SST = S_Y^2 = 168.8$		

- จากตารางสถิติ $F_{[0.05,1,3]} = 10.13$
- พบว่า $|F| < F_{[0.05,1,3]}$
- สรุปว่า ไม่ปฏิเสธ H_0 ที่ α=0.05
- แปลว่า จากข้อมูลที่ได้จาก ต.ย.ขนาด 5 อนุมานได้ว่า สมการถดถอยเชิงเส้นไม่เหมาะสมที่จะนำไปประมาณค่าตัวแปรตาม Y

ตัวอย่าง 6

- **กรณีใช้ค่าสถิติ t ทดสอบสมมติฐาน** H_0: $\beta = 0$

$$t = \frac{b}{\sqrt{\frac{MSE}{S_X^2}}} = \frac{-1.31915}{\sqrt{45.36/18.8}} = -0.849$$

- จากตารางสถิติ $t_{[0.025,3]} = 3.182$
- พบว่า $|t| < t_{[0.025,3]}$
- สรุปว่า ไม่ปฏิเสธ H_0 ที่ α=0.05
- แปลว่า จากข้อมูลที่ได้จาก ต.ย.ขนาด 5 อนุมานได้ว่า สมการถดถอยเชิงเส้นไม่เหมาะสมที่จะนำไปประมาณค่าตัวแปรตาม Y

ตัวอย่าง 6

- **กรณีใช้ค่าสถิติ t ทดสอบสมมติฐาน** H_0 : $\rho = 0$

$$t = -0.44\sqrt{\frac{5-2}{1-(-0.44)^2}} = -0.849$$

- จากตารางสถิติ $t_{[0.025,3]} = 3.182$
- พบว่า $|t| < t_{[0.025,3]}$
- สรุปว่า ไม่ปฏิเสธ H_0 ที่ α=0.05
- แปลว่า จากข้อมูลที่ได้จาก ต.ย.ขนาด 5 อนุมานได้ว่า ตัวแปรทั้งสองตัวไม่มีความสัมพันธ์เชิงเส้น

จากตัวอย่างที่ 6

- จะเห็นว่าผลการทดสอบสมมติฐาน H_0: $\beta = 0$ ไม่ปฏิเสธ H_0 ที่ระดับนัยสำคัญ 5% ซึ่งสอดคล้องกับผลการทดสอบสมมติฐาน H_0: $\rho = 0$ ไม่ปฏิเสธ H_0 ที่ระดับนัยสำคัญ 5% เพราะเหตุที่ตัวแปรทั้งสองไม่มีความสัมพันธ์เชิงเส้น จึงไม่เหมาะสมที่จะนำสมการถดถอยเชิงเส้นมาพยากรณ์ค่าตัวแปรตาม Y

การอนุมานเกี่ยวกับ α

- เป็นการทดสอบสมมติฐานว่า**สมการถดถอยเชิงเส้นของประชากรผ่านจุดกำเนิดหรือไม่**
- การอนุมานเกี่ยวกับ α ตั้งสมมติฐานดังนี้

H_0: $\alpha = 0$

H_1: $\alpha \neq 0$

การอนุมานเกี่ยวกับ α ด้วยค่า t

- ใช้ตัวสถิติ t ในการทดสอบสมมติฐาน

$$t = \frac{a}{\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_X^2}}}$$

- มีองศาอิสระ n-2
- สำหรับค่าสถิติ t: Decision rule คือ

ปฏิเสธ H_0 ที่ระดับนัยสำคัญ α% ถ้า $|t| \geq t_{(\alpha/2,n-2)}$

การอนุมานเกี่ยวกับ α ด้วยค่า t

- ช่วงความเชื่อมั่น (1- α)100% ของ α เป็นดังนี้

$$a - t_{\left(\frac{\alpha}{2},n-2\right)}S_a < \alpha < a + t_{\left(\frac{\alpha}{2},n-2\right)}S_a$$

$$S_a = S_{YX}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_X^2}}$$

$$S_{YX} = \sqrt{MSE}$$

$$S_X^2 = \sum (X_i - \bar{X})^2$$

ตัวอย่าง 7

- จากตารางข้อมูลขนาดบ้าน (X) และราคาบ้าน (Y) 20 หลัง ดังนี้

บ้านที่	ขนาด	ราคา	บ้านที่	ขนาด	ราคา	บ้านที่	ขนาด	ราคา	บ้านที่	ขนาด	ราคา
1	1.8	32	6	0.8	17	11	2.3	44	16	2.5	43
2	1.0	24	7	3.6	52	12	0.9	19	17	1.4	27
3	1.7	27	8	1.1	20	13	1.2	25	18	3.3	50
4	2.8	47	9	2.0	38	14	3.4	50	19	2.2	37
5	1.2	35	10	2.6	45	15	1.7	30	20	1.5	28

1) จงหาเวกเตอร์ส่วนเหลือ และจงทดสอบคุณสมบัติของเวกเตอร์ส่วนเหลือ

2) จงสร้างสมการถดถอยเชิงเส้นพร้อมทั้งทดสอบสมมติฐานว่าสมการถดถอยเชิงเส้นที่สร้างขึ้นเหมาะสมที่จะนำไปทำนายราคาบ้านหรือไม่ ที่ระดับนัยสำคัญ 0.05

ตัวอย่าง 7 – สมการถดถอย

- หาสมการถดถอยเชิงเส้นของ Y บน X

$$S_{XY} = \sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n} = 1554.9 - \frac{(40)(690)}{20} = 174.9$$

$$S_X^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{n} = 93.56 - \frac{(40)^2}{20} = 13.56$$

$$b = \frac{S_{XY}}{S_X^2} = \frac{174.9}{13.56} = 12.8982$$

$$a = \bar{Y} - b\bar{X} = (690/20) - (12.8982)\left(\frac{40}{20}\right) = 8.7035$$

- สรุปสมการถดถอยเชิงเส้นของ Y บน X คือ

$$\hat{Y} = 8.7035 + 12.8982X$$

ตัวอย่าง 7 – สมการถดถอย

- จากสมการถดถอยเชิงเส้นที่ได้นำมาคำนวณค่าพยากรณ์ได้ดังนี้

$$\hat{Y}_i = 8.7035 + 12.8982X_i$$

บ้านที่	Y_i	$\hat{Y}_i$	บ้านที่	Y_i	$\hat{Y}_i$	บ้านที่	Y_i	$\hat{Y}_i$	บ้านที่	Y_i	$\hat{Y}_i$
1	32	31.9204	6	17	19.0222	11	44	30.8695	16	43	40.9491
2	24	21.6018	7	52	55.1371	12	19	30.8120	17	27	26.7611
3	27	30.6305	8	20	22.8916	13	25	24.1814	18	50	51.2677
4	47	44.8186	9	38	34.5000	14	50	52.5575	19	37	37.0796
5	35	37.0704	10	45	42.2389	15	30	30.6305	20	28	28.0509

- จะเห็นว่า ถ้าสองตัวแปรมีความสัมพันธ์กันไม่สมบูรณ์การพยากรณ์ค่า Y จะไม่ถูกต้อง 100%

ตัวอย่าง 7 - เวกเตอร์ส่วนเหลือ

- หาเวคเตอร์ส่วนเหลือและทดสอบคุณสมบัติของเวคเตอร์ส่วนเหลือได้ดังนี้

$$\mathbf{d} = Y_i - \hat{Y}_i$$

บ้านที่	d	Xd	บ้านที่	d	Xd	บ้านที่	d	Xd	บ้านที่	d	Xd
1	0.0796	0.1434	6	-2.0221	-1.6177	11	5.6305	12.9502	16	2.0509	5.1272
2	2.3982	2.3982	7	-3.1372	-11.2938	12	-1.3119	-1.1808	17	0.2389	0.3345
3	-3.6309	-6.1719	8	-2.8916	-3.1808	13	0.8186	0.9823	18	-1.2677	-4.1834
4	2.1814	6.1080	9	3.5000	7.0000	14	-2.5575	-8.6956	19	-0.0796	-0.1752
5	-2.0794	-4.5752	10	2.7611	7.1788	15	-0.6305	-1.0719	20	-0.0509	-0.0763

- ทดสอบผลบวกของสมาชิกของเวกเตอร์ส่วนเหลือเป็นศูนย์ $\sum \mathbf{d}_i = -7.11E-15$
- ทดสอบเวกเตอร์ของตัวแปรอิสระเป็นอิสระกับเวคเตอร์ส่วนเหลือ $\sum X_i \mathbf{d}_i = -4.95E-13$

ตัวอย่าง 7 - อนุมานเกี่ยวกับ β

- **กรณีใช้ค่าสถิติ F ทดสอบ** จะใช้ตาราง ANOVA คำนวณ

แหล่งความแปรปรวน	องศาอิสระ	SS	MS	F
ถดถอย	1	SSR	MSR =SSR/1	F= MSR/MSE
ส่วนเหลือ	n-2	SSE	MSE =SSE/(n-2)	
รวม	n-1	SST		

- สมมติฐานคือ

H_0: $\beta = 0$ (สมการไม่เหมาะสม)

H_1: $\beta \neq 0$ (สมการเหมาะสม)

ตัวอย่าง 7 - อนุมานเกี่ยวกับ β

- ตาราง ANOVA คำนวณจากค่า b

แหล่งความแปรปรวน	องศาอิสระ	SS	MS	F
ถดถอย	1	$SSR = b^2 S_X^2 = 2255.9004$	$MSR = 2255.9004$	$F = \frac{MSR}{MSE}$
ส่วนเหลือ	20-2=18	$SSE = SST - SSR = 117.0996$	$MSE = 6.5055$	$= 346.7665$
รวม	20-1=19	$SST = S_Y^2 = 2373$		

- จากตารางสถิติ $F_{[0.05,1,18]} = 4.41$
- พบว่า $|F| > F_{[0.05,1,18]}$
- สรุปว่า ปฏิเสธ H_0 ที่ α=0.05
- แปลว่า จากข้อมูลที่ได้จาก ต.ย.ขนาด 20 อนุมานได้ว่า สมการถดถอยเชิงเส้นเหมาะสมที่จะนำไปประมาณค่าตัวแปรตาม Y

ตัวอย่าง 7 – อนุมานเกี่ยวกับ α

- **กรณีใช้ค่าสถิติ t ทดสอบสมมติฐาน** H_0: $\alpha = 0$

$$t_{(\frac{\alpha}{2},n-2)} = \frac{a}{\sqrt{MSE}\sqrt{\frac{1}{n}+\frac{\bar{X}^2}{S_X^2}}} = \frac{8.7035}{\sqrt{6.5055}\sqrt{\frac{1}{20}+\frac{4}{13.56}}} = 5.810$$

- จากตารางสถิติ $t_{[0.025,18]} = 2.101$
- พบว่า $|t| \geq t_{[0.025,18]}$
- สรุปว่า ปฏิเสธ H_0 ที่ α=0.05
- แปลว่า จากข้อมูลที่ได้จาก ต.ย.ขนาด 5 อนุมานได้ว่า สมการถดถอยเชิงเส้นของประชากรไม่ผ่านจุดกำเนิด

คำถาม