บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

แนวคิดและทฤษฎี

ข้อมูลที่ใช้ในการวิจัยครั้งนี้มาจาก 2 การแจกแจง คือการแจกแจงแบบปกติ และการแจกแจงแบบโลจิสติก ที่มีค่าเฉลี่ย 0 และความแปรปรวน 1 ในส่วนของการเปรียบเทียบประสิทธิภาพตัวสถิติที่ใช้ในการคัดกรองการแจกแจงแบบโลจิสติกจากการแจกแจงแบบปกติใช้ตัวสถิติ 6 ตัวสถิติ คือ ตัวสถิติ Kolmogorov-Smirnov, ตัวสถิติ Shapiro Wilk, ตัวสถิติ Anderson Darling, ตัวสถิติ Lilliefors, ตัวสถิติ Cramer Von Mises และตัวสถิติ Chi-square ส่วนของการทดสอบคุณสมบัติทั่วไปของการแจกแจงแบบปกติ และการแจกแจงแบบโลจิสติก ทำการทดสอบความสัมพันธ์ของ 3 การแจกแจง คือ การแจกแจงแบบไคสแควร์ การแจกแจงแบบที และการแจกแจงแบบเอฟ และในส่วนของการศึกษาผลกระทบของการประมาณค่าสัมประสิทธิ์การถดถอย เมื่อความคลาดเคลื่อนมีการแจกแจงแบบโลจิสติก ใช้การถดถอยอย่างง่าย และการถดถอยเชิงพหุ

2.1 การแจกแจงที่ใช้ในการวิจัย

การแจกแจงที่ใช้ในงานวิจัยครั้งนี้ มีดังนี้

1. การแจกแจงแบบ Location – Scale family

ให้ U เป็นตัวแปรสุ่ม และ a เป็นค่าคงที่ ที่บวกเพิ่มไปในตัวแปรสุ่ม U , $-\infty < a < \infty$

$$X = U + a$$

ถ้า $P(X \le x) = F(x-a)$ จะเรียกว่าเป็น location family.

และ ให้ b เป็นค่าคงที่ , b > 0

$$X = bU$$

ถ้า $P(X \le x) = F(x/b)$ จะเรียกว่าเป็น scale family.

เมื่อนำทั้ง 2 ประเภทมารวมกันเป็น

$$X = a + bU$$

ถ้า

$$P(X \le x) = F(\frac{x-a}{b})$$

$$= \frac{1}{b} f\left(\frac{x-a}{b}\right)$$

โดยที่ f เป็นฟังก์ชันความหนาแน่นความน่าจะเป็นบน $(-\infty,\infty)$ จะเรียกว่า Location – scale family มีการแจกแจงมากมายที่ถูกจัดรวมอยู่ในกลุ่ม Location – Scale family เช่น การแจกแจงแบบปกติ $N(\mu,\sigma^2)$, การแจกแจงแบบโลจิสติก $L(\mu^*,\sigma^*)$, การแจกแจงแบบยูนิฟอร์ม $U\left(a-\frac{b}{2},a+\frac{b}{2}\right)$ เป็นต้น

คุณสมบัติ Equivariant - Invariant

ตัวประมาณ f จะเรียกว่า location equivariant เมื่อกำหนด ดังนี้

$$f\left(\underset{\sim}{X}+\underset{\sim}{a}\right)=f\left(\underset{\sim}{X}\right)+\underset{\sim}{a}$$

ตัวอย่างเช่น ให้ X_i มีการแจกแจงแบบ $N(\mu,\sigma^2)$ และ iid $;i=1,2,...,n$

$\hat{\mu}(\underset{\sim}{X})=\frac{1}{n}\sum_{i=1}^{n}X_i$ เรียกว่า เป็นตัวประมาณ location equivariant เนื่องจาก

$$\begin{aligned}\hat{\mu}\left(\underset{\sim}{X}+\underset{\sim}{a}\right)&=\frac{1}{n}\sum_{i=1}^{n}(X_i+a)\\&=\frac{1}{n}\sum_{i=1}^{n}X_i+\frac{1}{n}\sum_{i=1}^{n}a\\&=\hat{\mu}(\underset{\sim}{X})+\underset{\sim}{a}\end{aligned}$$

ตัวประมาณ f จะเรียกว่า scale equivariant เมื่อกำหนด ดังนี้

$$f\left(b\underset{\sim}{X}\right)=bf\left(\underset{\sim}{X}\right)$$

ตัวอย่างเช่น ให้ X_i มีการแจกแจงแบบ $N(\mu,\sigma^2)$ และ iid $;i=1,2,...,n$

$\hat{\sigma}(\underset{\sim}{X})=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$ เรียกว่า เป็นตัวประมาณ scale equivariant เนื่องจาก

$$\begin{aligned}\hat{\sigma}(b\underset{\sim}{X})&=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(bX_i-b\bar{X})^2}\\&=\sqrt{\frac{b^2}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}\\&=b\hat{\sigma}(\underset{\sim}{X})\end{aligned}$$

ตัวประมาณ f จะเรียกว่า invariant เมื่อกำหนด ดังนี้

$$f\left(\underset{\sim}{X}+\underset{\sim}{a}\right)=f\left(\underset{\sim}{X}\right)$$

ตัวอย่างเช่น ให้ X_i มีการแจกแจงแบบ $N(\mu,\sigma^2)$ และ iid $;i=1,2,...,n$

$\hat{\sigma}(\underset{\sim}{X})=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$ เรียกว่า เป็นตัวประมาณ invariant เนื่องจาก

สำนักงานคณะกรรมการวิจัยแห่งชาติ ห้องสมุดงานวิจัย

$$\hat{\sigma}(\underset{\sim}{X}+\underset{\sim}{a}) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left((X_i+a)-(\overline{X}+a)\right)^2}$$
$$= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}$$
$$= \hat{\sigma}(\underset{\sim}{X})$$

โดยที่ $(X_i + a)$ มีการแจกแจงแบบ $N(\mu + a, \sigma^2)$

ดังนั้น $(X_i - \mu)/\sigma$ มีการแจกแจงแบบ $N(0,1)$

2. การแจกแจงแบบปกติ

การแจกแจงแบบปกติ เป็นการแจกแจงความน่าจะเป็นที่สำคัญในการวิเคราะห์ทางสถิติ และนำไปใช้ประโยชน์อย่างกว้างขวางสำหรับตัวแปรสุ่มชนิดต่อเนื่องทั้งนี้เพราะเหตุการณ์ที่เกิดขึ้นส่วนใหญ่มีลักษณะใกล้เคียงการแจกแจงชนิดนี้ ฟังก์ชันความหนาแน่นความน่าจะเป็นของการแจกแจงแบบปกติที่มีค่าเฉลี่ย μ และความแปรปรวน σ^2 คือ

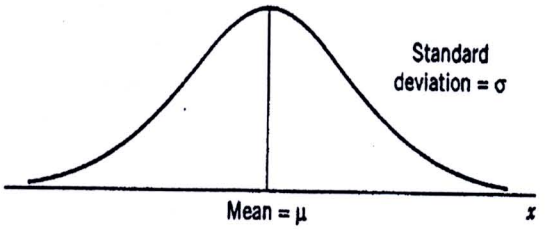
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \quad , \quad -\infty < x < \infty ; -\infty < \mu < \infty , \sigma^2 > 0$$

โดยที่ π เป็นค่าคงที่ (ค่าโดยประมาณเท่ากับ 3.1416)

e เป็นค่าคงที่ (ค่าโดยประมาณเท่ากับ 2.7183)

σ เป็นพารามิเตอร์ ซึ่งเท่ากับส่วนเบี่ยงเบนมาตรฐานของการแจกแจง

μ เป็นพารามิเตอร์ ซึ่งเท่ากับค่ากลางของการแจกแจง



ภาพที่ 2.1 แสดงโค้งของการแจกแจงแบบปกติ $N(\mu, \sigma^2)$

ถ้า ตัวแปรสุ่ม X มีการแจกแจง $N(\mu, \sigma^2)$ ได้ว่า

- ค่าเฉลี่ยของตัวแปรสุ่ม $E(x) = \mu$
- ความแปรปรวนของตัวแปรสุ่ม $Var(x) = \sigma^2$
- สัมประสิทธิ์ความเบ้ = 0, สัมประสิทธิ์ความโด่ง = 3

สำนักงานคณะกรรมการวิจัยแห่งชาติ
ห้องสมุดงานวิจัย
วันที่ 10 ก.ค. 2555
เลขทะเบียน 248397
เลขเรียกหนังสือ

ลักษณะของการแจกแจงแบบปกติ

- เส้นโค้งมีลักษณะสมมาตร รูปร่างคล้ายระฆังคว่ำ มียอดเดียวอยู่ที่กึ่งกลางของเส้นโค้ง
- ค่าเฉลี่ย มัธยฐาน และฐานนิยม มีค่าเท่ากัน อยู่ที่จุดกึ่งกลางจึงแบ่งพื้นที่ใต้เส้นโค้งปกติออกเป็น 2 ส่วนเท่าๆกัน
- ปลายทั้งสองข้างของเส้นโค้งจะค่อยๆ ลาดลงสู่แกน X และยื่นออกไปทั้งสองข้าง โดยไม่มีที่สิ้นสุดและไม่แตะแกน X และปลายทั้งสองข้างของเส้นโค้งปกติจะมีค่าตั้งแต่ $-\infty$ ถึง ∞ พื้นที่ใต้เส้นโค้งที่อยู่เหนือแกน X จะเท่ากับ 1
- μ และ σ^2 เป็นค่าพารามิเตอร์โดยเป็นตัวกำหนดตำแหน่งของเส้นโค้ง และลักษณะของเส้นโค้งว่าจะแบนหรือโด่งอย่างไร

3. การแจกแจงแบบโลจิสติก

การแจกแจงแบบโลจิสติกมีรูปร่างลักษณะใกล้เคียงกับการแจกแจงแบบปกติ โดยต่างกันเพียงส่วนหางที่หนักกว่าเล็กน้อย ฟังก์ชันความหนาแน่นความน่าจะเป็น คือ

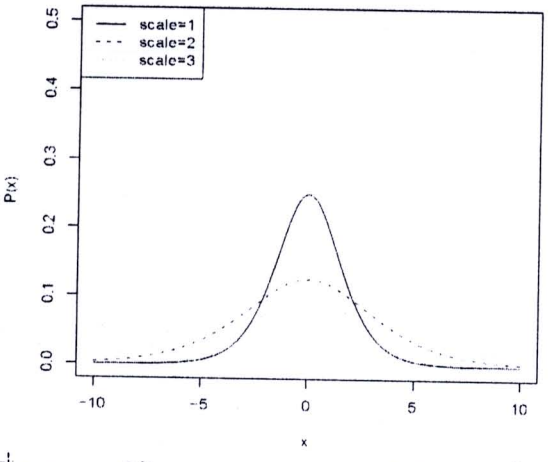
$$f(x) = \frac{e^{-(x-\mu^*)/\sigma^*}}{\sigma^*(1+e^{-(x-\mu^*)/\sigma^*})^2} \quad , \quad -\infty < x < \infty ; -\infty < \mu^* < \infty$$

โดยที่ e เป็นค่าคงที่ (ค่าโดยประมาณเท่ากับ 2.7183)

σ^* เป็นพารามิเตอร์สเกล ซึ่งส่งผลต่อรูปร่างของการแจกแจง

μ^* เป็นพารามิเตอร์ตำแหน่ง ซึ่งเท่ากับค่ากลางของการแจกแจง

f (x) เป็นความสูงของเส้นโค้งเมื่อพล็อตค่าบนแกน



ภาพที่ 2.2 แสดงโค้งของการแจกแจงแบบโลจิสติก $L(\mu^*,\sigma^*)$

สำนักงานคณะกรรมการวิจัยแห่งชาติ ห้องสมุดงานวิจัย

ถ้า ตัวแปรสุ่ม X มีการแจกแจง $L(\mu^*,\sigma^*)$ ได้ว่า

- ค่าเฉลี่ยของตัวแปรสุ่ม $E(x)=\mu^*$
- ความแปรปรวนของตัวแปรสุ่ม $Var(x)=(\pi\sigma^*)^2/3$
- สัมประสิทธิ์ความเบ้ = 0, สัมประสิทธิ์ความโด่ง = 4.2

<u>ลักษณะของการแจกแจงแบบโลจิสติก</u>

- เส้นโค้งมีลักษณะสมมาตร รูปร่างคล้ายระฆังคว่ำ มียอดเดียวอยู่ที่กึ่งกลางของเส้นโค้ง
- ค่าเฉลี่ย มัธยฐาน และฐานนิยม มีค่าเท่ากัน อยู่ที่จุดกึ่งกลางจึงแบ่งพื้นที่ใต้เส้นโค้งปกติออกเป็น 2 ส่วนเท่าๆกัน
- ปลายทั้งสองข้างของเส้นโค้งจะค่อยๆ ลาดลงสู่แกน X และยื่นออกไปทั้งสองข้าง โดยไม่มีที่สิ้นสุดและไม่แตะแกน X และปลายทั้งสองข้างของเส้นโค้งปกติจะมีค่าตั้งแต่ $-\infty$ ถึง ∞ พื้นที่ใต้เส้นโค้งที่อยู่เหนือแกน X จะเท่ากับ 1
- μ^* และ σ^* เป็นค่าพารามิเตอร์โดยเป็นตัวกำหนดตำแหน่งของเส้นโค้ง และลักษณะของเส้นโค้งว่าจะแบนหรือโด่งอย่างไร

4. การแจกแจงไคสแควร์

ฟังก์ชันความหนาแน่นความน่าจะเป็นของการแจกแจง เป็นดังนี้

$$f(x)=\frac{x^{(n-2)/2}e^{(-x/2)}}{2^{n/2}\Gamma(n/2)} \quad ;x\geq 0$$

โดยที่ n แทน ระดับความเป็นอิสระ

ตัวแปรสุ่มไคสแควร์ คือตัวแปรสุ่มที่พัฒนามาจากฟังก์ชันของตัวแปรสุ่มที่มีการแจกแจงแบบปกติ ถ้า $x_i\,;i=1,2,...,n$ เป็นตัวแปรสุ่มที่มีการแจกแจงแบบปกติ มีค่าเฉลี่ย μ ความแปรปรวน σ^2 ได้ว่า $Z^2=\left(\frac{x-\mu}{\sigma}\right)^2$ โดย z^2 มีการแจกแจง $\chi^2_{(1)}$ และ $Z_1^2+Z_2^2+Z_3^2+...+Z_n^2$ มีการแจกแจง $\chi^2_{(n)}$

5. การแจกแจงที

ฟังก์ชันความหนาแน่นความน่าจะเป็นของการแจกแจง เป็นดังนี้

$$f(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)\left(1+\frac{x^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}}{\sqrt{n\pi}\ \Gamma\left(\frac{n}{2}\right)}$$

โดยที่ n แทน ระดับความเป็นอิสระ

การแจกแจงทีมีลักษณะที่คล้ายกันกับการแจกแจงแบบปกติมาตรฐาน คือเป็นการแจกแจงแบบสมมาตร และมีค่าสูงสุดที่ 0 โดยตัวสถิติที่ใช้ทดสอบเกี่ยวกับค่าเฉลี่ยของประชากร ในกรณีที่ไม่ทราบค่าความแปรปรวนของประชากร $\left(\frac{\bar{x}-\mu}{s/\sqrt{n}}\right)$ จะมีการแจกแจง $t_{(n-1)}$

6. การแจกแจงเอฟ

ฟังก์ชันความหนาแน่นความน่าจะเป็นของการแจกแจง เป็นดังนี้

$$f(x)=\frac{\Gamma\left(\frac{n_1+n_2}{2}\right)\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}x^{\frac{n_1}{2}-1}}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)\left(1+\frac{n_1}{n_2}x\right)^{\frac{n_1+n_2}{2}}}\quad ;x>0$$

โดยที่ n_1 , n_2 แทน ระดับความเป็นอิสระ

ความสำคัญของการแจกแจงเอฟ คือ สามารถเขียนอยู่ในรูปอัตราส่วนของตัวประมาณความแปรปรวนที่อิสระกัน 2 ตัว ถ้ามีข้อมูลที่มีการแจกแจงแบบปกติ และอิสระกัน 2 ชุด มีความแปรปรวน σ_1^2 , σ_2^2 ตามลำดับ ได้ว่า

$\frac{s_1^2}{s_2^2}$ มีการแจกแจง $F_{(n_1-1,n_2-1)}$

2.2 ตัวสถิติที่ใช้ในการเปรียบเทียบประสิทธิภาพการคัดกรอง

1. ตัวสถิติ Kolmogorov – Smirnov

ตัวสถิติที่ใช้ในการทดสอบเป็นดังนี้

$$D=\max\left(D^+,D^-\right)$$

$$D^+=\max_{1\le i\le n}\left\{\frac{i}{n}-Z_i\right\}$$

$$D^{-} = \max_{1 \le i \le n} \left\{ Z_i - \frac{i-1}{n} \right\}$$

โดย Z_i แทน ความน่าจะเป็นสะสมของการแจกแจงแบบปกติมาตรฐาน

$$= \phi\left(\left(x_{(i)} - \bar{x} \right) / s \right)^{2}$$

n แทน ขนาดของตัวอย่าง

สำหรับค่าวิกฤติของตัวสถิติ จะปฏิเสธสมมติฐานเมื่อ D ที่คำนวณได้มีค่ามากกว่า D_n ณ ระดับนัยสำคัญที่กำหนด

2. ตัวสถิติ Shapiro – Wilk

ข้อมูลที่นำมาใช้ อย่างน้อยที่สุดต้องวัดด้วยมาตรวัดอันตรภาค ตัวสถิติที่ใช้ในการทดสอบ W หรือตัวสถิติของ Shapiro-Wilk คือ

$$W = \frac{\left\{ \sum_{i=1}^{k} a_{n-i+1} \left(X_{(n-i+1)} - X_{(i)} \right) \right\}^{2}}{\sum_{i=1}^{n} \left(X_{(i)} - \bar{X} \right)^{2}}$$

โดยที่ n แทน ขนาดตัวอย่าง

k แทน จำนวนเต็มที่เล็กที่สุดที่มากกว่าหรือเท่ากับ n/2

a_{n-i+1} แทน ค่าสัมประสิทธิ์ที่ได้จากการเปิดตาราง

$\bar{X}$ แทน ค่าเฉลี่ยตัวอย่าง

$X_{(i)}$ แทน สถิติลำดับของตัวอย่างสุ่ม ลำดับที่ i

การสรุปผลการทดสอบ จะปฏิเสธสมมติฐาน เมื่อค่า W ที่คำนวณได้มีค่าน้อยกว่าค่า W ที่ได้จากตาราง ณ ระดับนัยสำคัญที่กำหนด

3. ตัวสถิติ Anderson Darling

ตัวสถิตินี้คิดขึ้นโดย Anderson Darling (1953) ซึ่งตัวสถิติที่ใช้ในการทดสอบเป็นดังนี้

$$A = -\frac{1}{n} \left\{ \sum_{i=1}^{n} (2i-1) \left[\ln Z_i + \ln\left(1 - Z_{n+1-i}\right) \right] \right\} - n$$

โดย Z_i แทน ความน่าจะเป็นสะสมของการแจกแจงแบบปกติมาตรฐาน

$$= \phi\left(\left(x_{(i)} - \bar{x} \right) / s \right)$$

---

[2] ให้ ϕ คือ การแจกแจงปกติมาตรฐาน (N(0,1))

n แทน ขนาดของตัวอย่าง

สำหรับค่าวิกฤติของตัวสถิติ จะปฏิเสธสมมติฐานเมื่อ A ที่คำนวณได้มีค่ามากกว่าค่าจากตาราง ณ ระดับนัยสำคัญที่กำหนด

4. ตัวสถิติ Lilliefors

ตัวสถิติที่ใช้ในการทดสอบเป็นดังนี้

$$D = \max\left(D^{+}, D^{-}\right)$$

$$D^{+} = \max_{1 \le i \le n}\left\{\frac{i}{n} - Z_i\right\}$$

$$D^{-} = \max_{1 \le i \le n}\left\{Z_i - \frac{i-1}{n}\right\}$$

โดย Z_i แทน ความน่าจะเป็นสะสมของการแจกแจงแบบปกติมาตรฐาน

$$= \phi\left(\left(x_{(i)} - \bar{x}\right)/s\right)$$

n แทน ขนาดของตัวอย่าง

โดยจะปฏิเสธสมมติฐานว่างเมื่อ D ที่คำนวณได้มีค่ามากกว่าค่าจากตาราง Lilliefors ณ ระดับนัยสำคัญที่กำหนด

5. ตัวสถิติ Cramer Von Mises

ตัวสถิติที่ใช้ในการทดสอบเป็นดังนี้

$$W^2 = \sum_{i=1}^{n}\left\{Z_i - \frac{(2i-1)}{2n}\right\}^2 + \frac{1}{12n}$$

โดย Z_i แทน ความน่าจะเป็นสะสมของการแจกแจงแบบปกติมาตรฐาน

$$= \phi\left(\left(x_{(i)} - \bar{x}\right)/s\right)$$

n แทน ขนาดของตัวอย่าง

สำหรับค่าวิกฤติของตัวสถิติ จะปฏิเสธสมมติฐานเมื่อ W^2 ที่คำนวณได้มีค่ามากกว่าค่าจากตาราง ณ ระดับนัยสำคัญที่กำหนด

6. ตัวสถิติ Chi-square

ตัวสถิติที่ใช้ในการทดสอบเป็นดังนี้

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

โดย O_i แทน จำนวนความถี่ของข้อมูลในช่วงที่ i

E_i แทน ค่าคาดหวังของจำนวนความถี่ของข้อมูลในช่วงที่ i

n แทน ขนาดของตัวอย่าง

k แทน จำนวนช่วง

สำหรับค่าวิกฤติของตัวสถิติ จะปฏิเสธสมมติฐานเมื่อค่า χ^2 ที่คำนวณได้มีค่ามากกว่า $\chi^2_{(k-3)}$ [3]

2.3 การวิเคราะห์การถดถอย

การวิเคราะห์การถดถอย เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรตั้งแต่ 2 ตัวขึ้นไป การวิเคราะห์การถดถอยมีวัตถุประสงค์ในการอธิบายความสัมพันธ์ระหว่างตัวแปร และทำนายค่าของตัวแปรตาม การวิจัยครั้งนี้ทำการศึกษาผลกระทบของการประมาณค่าสัมประสิทธิ์การถดถอยในการถดถอยอย่างง่าย และการถดถอยเชิงพหุ

1. ตัวแบบการถดถอย

- การถดถอยอย่างง่าย

การถดถอยอย่างง่าย(Simple regression) เป็นการศึกษาถึงความสัมพันธ์ระหว่างตัวแปร 2 ตัว ตัวแบบการถดถอยอย่างง่ายสามารถเขียนได้ดังนี้

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

โดย β_0 , β_1 คือ พารามิเตอร์ โดยเรียกพารามิเตอร์ทั้งสองว่า สัมประสิทธิ์การถดถอย

Y คือ ตัวแปรตาม

X คือ ตัวแปรอิสระ

ε คือ ความคลาดเคลื่อน

หรือเขียนตัวแบบในรูปเมตริกซ์ได้เป็น

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

---

[3] χ^2 มี Degree of Freedom เท่ากับ k-m-1 เมื่อ m คือ จำนวนพารามิเตอร์ในตัวแบบที่ถูกประมาณจากข้อมูล คือ μ และ σ หรือ σ^2 และที่ต้องลบ 1 เนื่องจากข้อจำกัดที่ว่า $\sum_{i=1}^{k} O_i = n = \sum_{i=1}^{k} E_i$ ในที่นี้จึงเทียบกับ $\chi^2_{(k-3)}$

โดย

$$\underset{\sim}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}, \quad \underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \underset{\sim}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- การถดถอยเชิงพหุ

ในที่นี้จะกล่าวถึงการถดถอยเชิงพหุ กรณีมีตัวแปรอิสระ 2 ตัว ตัวแบบการถดถอยอยู่ในรูป

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

โดย β_0 , β_1, β_2 คือ พารามิเตอร์ โดยเรียกว่า สัมประสิทธิ์การถดถอย

Y คือ ตัวแปรตาม

X คือ ตัวแปรอิสระ

ε คือ ความคลาดเคลื่อน

หรือเขียนตัวแบบในรูปเมตริกซ์ได้เป็น

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

โดย

$$\underset{\sim}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \end{bmatrix}, \quad \underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad \underset{\sim}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

2. เงื่อนไขของการถดถอย

- ε_i มีการแจกแจงแบบปกติ
- ε_i มีค่าเฉลี่ย 0 ความแปรปรวนคงที่ σ^2
- cov (ε_i, ε_j) = 0, $i \neq j$

3. การประมาณค่าสัมประสิทธิ์การถดถอย

การวิจัยครั้งนี้ใช้วิธีกำลังสองน้อยที่สุด ซึ่งเป็นการหาค่าประมาณค่าสัมประสิทธิ์การถดถอย จากข้อมูลที่ทำให้ผลบวกกำลังสองของความคลาดเคลื่อน ($\sum_{i=1}^{n} e_i^2$) มีค่าต่ำสุด

จากตัวแบบในรูปเมตริกซ์ ได้ว่า

$$\underset{\sim}{\varepsilon} = \underset{\sim}{Y} - X\underset{\sim}{\beta}$$

ผลบวกกำลังสองของความคลาดเคลื่อนเขียนได้เป็น

$$\underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon} = \underset{\sim}{Y}'\underset{\sim}{Y} - 2\underset{\sim}{\beta}'X'\underset{\sim}{Y} + \underset{\sim}{\beta}'X'X\underset{\sim}{\beta}$$

โดยจะมีค่าต่ำสุด เมื่อ

$$\frac{\partial}{\partial \underset{\sim}{\beta}} \underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon} = -2X'\underset{\sim}{Y} + 2X'X\underset{\sim}{\beta} = 0$$

แทนค่า $\underset{\sim}{\beta}$ ด้วย $\underset{\sim}{b}$ จะได้สมการปกติ

$$X'X\underset{\sim}{b} = X'\underset{\sim}{Y}$$

$$\underset{\sim}{b} = (X'X)^{-1}X'\underset{\sim}{Y}$$

<u>หมายเหตุ</u>

$\underset{\sim}{b}$ หาค่าได้ ก็ต่อเมื่อมี $(X'X)^{-1}$ โดยที่ X ต้องเป็น full column rank ถึงจะสามารถหา $(X'X)^{-1}$ ได้