บทที่ 2
ทฤษฎีสถิติที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงการแจกแจงที่เกี่ยวข้อง การผลิตตัวแปรสุ่มต่างๆ ที่ใช้ในการวิจัย ตัวสถิติทดสอบที่ใช้ในการทดสอบเทียบความกลมกลืนสำหรับตัวแบบการถดถอย การวิเคราะห์การถดถอย ความผิดพลาดในการทดสอบสมมติฐานทางสถิติ การจำลองโดยใช้เทคนิคมอนติคาร์โลในการค้นหาคำตอบที่สนใจ และการสุ่มตัวอย่างแบบบูตสแตรปเพื่อการประยุกต์ใช้ในการหาค่าวิกฤติสำหรับตัวสถิติทดสอบ KS และตัวสถิติทดสอบ CvM ซึ่งมีรายละเอียดต่างๆ ดังต่อไปนี้

2.1 การแจกแจงที่ใช้ในการวิจัย

ในการวิจัยครั้งนี้จะทำการศึกษาการทดสอบสมมติฐาน เพื่อทำการเปรียบเทียบและคัดเลือกวิธีการทดสอบที่เหมาะสมในการทดสอบเทียบความกลมกลืนสำหรับตัวแบบการถดถอย โดยพิจารณาตัวแบบการถดถอย 4 ตัวแบบดังนี้

ตัวแบบที่ 1 : $Y_i = 2 + 5X_i + \varepsilon_i$ เมื่อ $1 \le i \le n$

ตัวแบบที่ 2 : $Y_i = 2 + 5X_i + \beta_2 X_i^2 + \varepsilon_i$ เมื่อ $1 \le i \le n$

กำหนดให้ β_2 เท่ากับ 1, 3 และ 5

ตัวแบบที่ 3 : $Y_i = 2 + 5X_{1i} - 1X_{2i} + \varepsilon_i$ เมื่อ $1 \le i \le n$

ตัวแบบที่ 4 : $Y_i = 2 + 5X_{1i} - 1X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i$ เมื่อ $1 \le i \le n$

กำหนดให้ β_3 เท่ากับ 1, 3 และ 5

จากตัวแบบต่างๆ ข้างต้นจะเห็นได้ว่า ในการวิจัยครั้งนี้ต้องทำการผลิตตัวแปรสุ่มที่ใช้ในการวิจัยได้แก่ ตัวแปรอิสระ ความคลาดเคลื่อนสุ่ม ซึ่งตัวแปรอิสระนั้นจะผลิตจากการแจกแจงแบบปกติ ความคลาดเคลื่อนสุ่มจะมาจากการแจกแจงแบบปกติและการแจกแจงแบบลอกนอร์มอล นอกจากนี้การหาค่าวิกฤติสำหรับการทดสอบสมมติฐานด้วยตัวสถิติทดสอบเอฟนั้นผลิตมาจากการแจกแจงแบบเอฟ โดยในการผลิตตัวแปรสุ่มจากการแจกแจงดังกล่าวข้างต้น จะต้องทำการผลิตเลขสุ่มจากการแจกแจงแบบสม่ำเสมอในช่วง (0,1) เป็นพื้นฐาน รายละเอียดต่างๆ จากการแจกแจงที่เกี่ยวข้องนั้นมีดังต่อไปนี้

2.1.1 การแจกแจงแบบสม่ำเสมอ(Uniform Distribution)

การแจกแจงแบบสม่ำเสมอในช่วง (a,b) มีฟังก์ชันความหนาแน่นอยู่ในรูปแบบ

$$f(x) = \frac{1}{b-a} \quad ; a \le x \le b$$

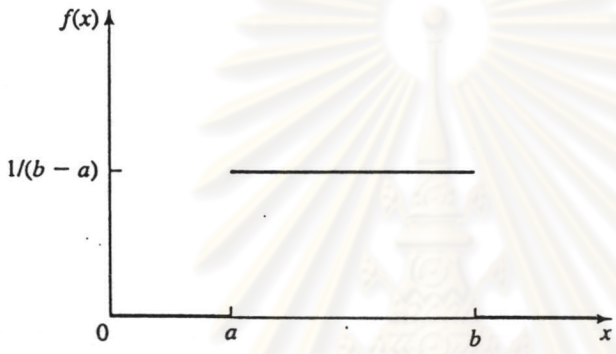
เมื่อ x เป็นตัวแปรสุ่มที่มีการแจกแจงแบบสม่ำเสมอ

a และ b เป็นจำนวนจริงใดๆ ที่ a<b

a คือ พารามิเตอร์ตำแหน่ง(location parameter)

b คือ พารามิเตอร์สเกล(scale parameter)

การแจกแจงแบบสม่ำเสมอในช่วง (0,1) จะนำมาใช้ในการผลิตเลขสุ่มเพื่อเป็นพื้นฐานในการผลิตข้อมูลที่มีการแจกแจงอื่นๆ และใช้ในการสุ่มตัวอย่างซ้ำแบบใส่คืนในการหาค่าวิกฤติสำหรับการทดสอบสมมติฐานด้วยตัวสถิติทดสอบ KS และตัวสถิติทดสอบ CvM ซึ่งแสดงรูปกราฟได้ดังนี้



รูปที่ 2.1 กราฟแสดงการแจกแจงแบบสม่ำเสมอในช่วง (a,b)

2.1.2 การแจกแจงแบบปกติ(Normal Distribution : $N(\mu,\sigma^2)$)

การแจกแจงแบบปกติ ที่มีค่าเฉลี่ยเท่ากับ μ และความแปรปรวนเท่ากับ σ^2 มีฟังก์ชันความหนาแน่นอยู่ในรูปแบบ

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

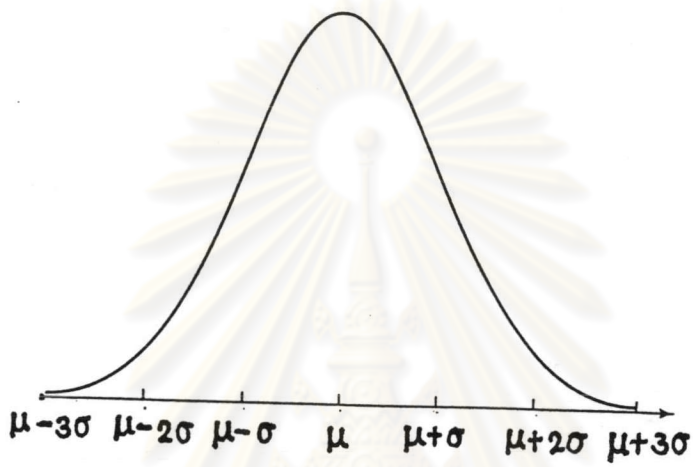
เมื่อ x เป็นตัวแปรสุ่มที่มีการแจกแจงแบบปกติ

μ คือ พารามิเตอร์ตำแหน่ง(location parameter) โดยที่ $\mu \in (-\infty, \infty)$

σ คือ พารามิเตอร์สเกล(scale parameter) โดยที่ $\sigma > 0$

การแจกแจงแบบปกตินี้จะใช้ในการผลิตค่าของตัวแปรอิสระโดยกำหนดในตัวแบบที่มีตัวแปรอิสระเพียงตัวเดียวจะใช้ ค่าเฉลี่ยเท่ากับ 20 และค่าความแปรปรวนเท่ากับ 100 สำหรับตัวแบบที่มีตัวแปรอิสระ 2 ตัว ตัวแปรอิสระตัวที่ 1 จะใช้ค่าเฉลี่ยและค่าความแปรปรวนข้างต้น ส่วนตัวแปรอิสระตัวที่ 2 จะใช้ค่าเฉลี่ยเท่ากับ 10 และค่าความแปรปรวนเท่ากับ 4 และ

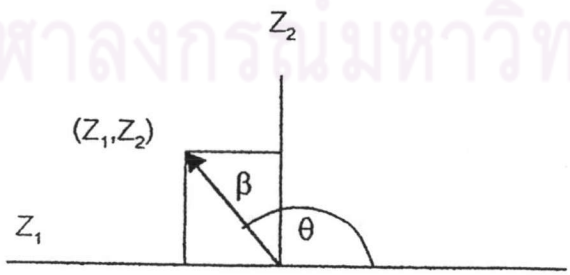
ใช้ในการผลิตค่าความคลาดเคลื่อนในตัวแบบการถดถอยด้วยค่าเฉลี่ยเท่ากับ 0 ความแปรปรวนเท่ากับ 1, 2 และ 3 ตามลำดับ
ซึ่งแสดงรูปกราฟได้ดังนี้



รูปที่ 2.2 กราฟแสดงการแจกแจงแบบปกติ ที่ค่าเฉลี่ย μ และความแปรปรวน σ^2

การผลิตเลขสุ่มที่มีการแจกแจงแบบปกติ

การผลิตเลขสุ่มที่การแจกแจงแบบปกติใช้วิธีการของ Box และ Muller (ค.ศ. 1958) โดยผลิตเลขสุ่มที่มีการแจกแจงแบบปกติมาตรฐาน ที่มีค่าเฉลี่ยเป็น 0 และค่าความแปรปรวนเป็น 1 พร้อมๆ กัน 2 ค่า โดยพิจารณาจากรูปต่อไปนี้



พิจารณาจากรูปจะได้

$$Z_1=\beta\cos(\theta) \quad \text{.....(1)}$$

$$Z_2=\beta\sin(\theta) \quad \text{.....(2)}$$

เมื่อ Z_1 และ Z_2 คือตัวแปรสุ่มที่มีการแจกแจงแบบปกติมาตรฐาน

จาก (1) และ (2) พิสูจน์ได้ว่า β และ θ เป็นอิสระกัน และ $\beta^2 = Z_1^2 + Z_2^2$ มีการแจกแจงแบบไคกำลังสอง(Chi-square) ด้วยระดับความเป็นอิสระเท่ากับ 2 ซึ่งเทียบเท่ากับการแจกแจงแบบเลขชี้กำลัง ค่าเฉลี่ยเท่ากับ 2 และ θ มีการแจกแจงแบบสม่ำเสมอในช่วง 0 ถึง 2π เรเดียน ดังนั้นจึงสามารถใช้วิธีการแปลงผกผัน(Inverse Transformation) สร้างเลขสุ่มของ β และ θ ได้ดังนี้

$$\beta = (-2 \ln R)^{1/2} \qquad \text{.....(3)}$$

เมื่อ R เป็นเลขสุ่มที่มีการแจกแจงแบบสม่ำเสมอในช่วง [0,1] เพราะฉะนั้นแทนค่า β และ θ ใน (1) และ (2) จะได้เลขสุ่มที่มีการแจกแจงแบบปกติมาตรฐาน Z_1 , Z_2 อิสระกัน

$$Z_1 = (-2 \ln R_1)^{1/2} \cos(2\pi R_1)$$

$$Z_2 = (-2 \ln R_2)^{1/2} \sin(2\pi R_2)$$

จากนี้เมื่อต้องการเลขสุ่มที่มีการแจกแจงแบบปกติที่มีค่าเฉลี่ย μ และความแปรปรวน σ^2 ทำได้โดยการแปลงเลขสุ่ม Z_1 , Z_2 โดยอาศัยฟังก์ชันต่อไปนี้

$$X_1 = \mu + \sigma Z_1$$

$$X_2 = \mu + \sigma Z_2$$

ซึ่งจะได้ X_1 และ X_2 เป็นอิสระกันและต่างก็มีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเท่ากับ μ และความแปรปรวนเท่ากับ σ^2 โปรแกรมย่อยที่ใช้ในการจำลองเลขสุ่มที่มีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเท่ากับ μ และความแปรปรวนเท่ากับ σ^2 นี้คือ FUNCTION NORMAL(DMEAN,SIGMA) ดังแสดงในภาคผนวก ก

2.1.3 การแจกแจงแบบลอกนอร์มอล(Lognormal Distribution)

การแจกแจงแบบลอกนอร์มอล มีฟังก์ชันความหนาแน่นอยู่ในรูปแบบ

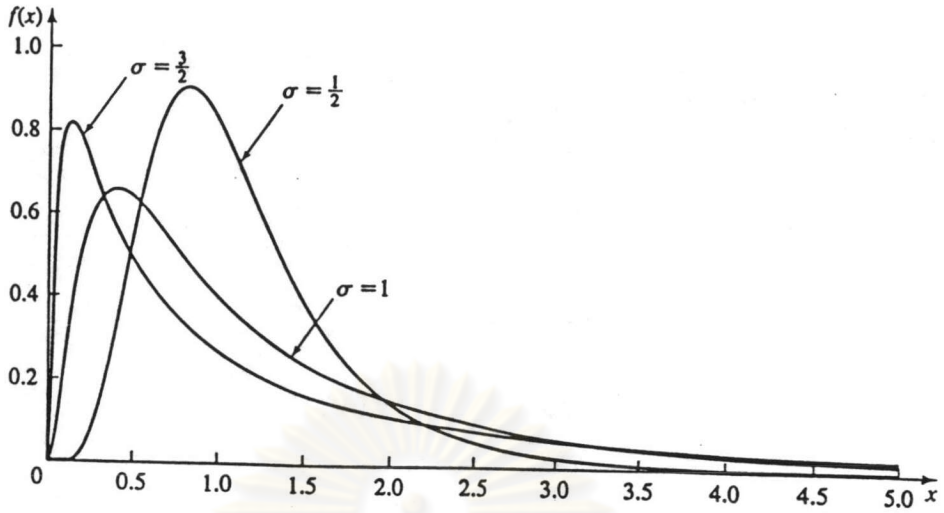
$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{\frac{-(\ln x - \mu)^2}{2\sigma^2}} \quad ; x > 0$$

เมื่อ x เป็นตัวแปรสุ่มที่มีการแจกแจงแบบลอกนอร์มอล

μ คือ พารามิเตอร์สเกล(scale parameter) โดย $\mu \in (-\infty, \infty)$

σ คือ พารามิเตอร์รูปทรง(shape parameter) โดย $\sigma > 0$

ค่าเฉลี่ยเท่ากับ $\exp(\mu + \sigma^2/2)$ และความแปรปรวนเท่ากับ $\exp(2\mu + \sigma^2) \times [\exp(\sigma^2) - 1]$ การแจกแจงแบบลอกนอร์มอลนี้จะใช้ในการผลิตค่าความคลาดเคลื่อนในตัวแบบการถดถอย โดยกำหนดให้ $\mu = 0$ และ $\sigma^2 = 0.25, 1.0$ และ 2.25 ซึ่งแสดงรูปกราฟได้ดังนี้



รูปที่ 2.3 กราฟแสดงการแจกแจงแบบลอกนอร์มอลที่มี $\mu = 0$ และ $\sigma = 0.5, 1.0, 1.5$

การผลิตเลขสุ่มที่มีการแจกแจงแบบลอกนอร์มอล

เนื่องจาการแจกแจงแบบลอกนอร์มอลมีความสัมพันธ์กับการแจกแจงแบบปกติ คือถ้า Y มีการแจกแจงแบบปกติที่มี μ และ σ^2 เป็นค่าเฉลี่ยและความแปรปรวนตามลำดับแล้ว เมื่อกำหนดให้ X = exp(Y) ก็จะได้ว่า X มีการแจกแจงแบบลอกนอร์มอลที่มีลักษณะตามที่ระบุข้างต้น ดังนั้นการผลิตตัวแปรสุ่มที่มีการแจกแจงแบบลอกนอร์มอลที่มีพารามิเตอร์ μ และ σ^2 สามารถทำได้จาการหา exponential ของตัวแปรสุ่มที่มีการแจกแจงแบบปกติหรือหา exponential ของฟังก์ชัน NORMAL(DMEAN,SIGMA) ดังแสดงในภาคผนวก ก

2.1.4 การแจกแจงแบบเอฟ(F Distribution)

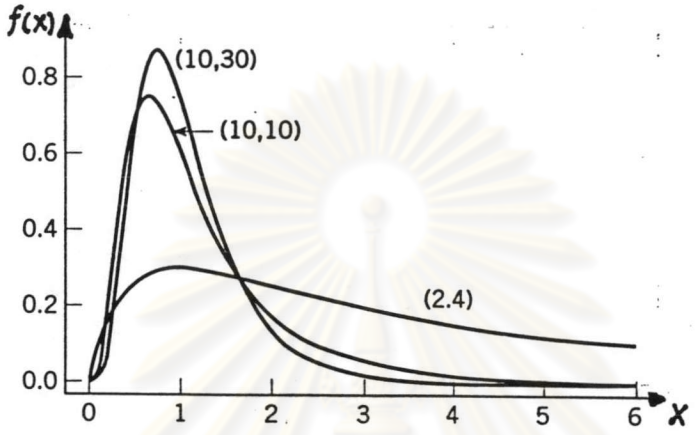
การแจกแจงแบบเอฟที่มีระดับขั้นความเป็นอิสระ m และ n มีฟังก์ชันความหนาแน่นอยู่ในรูปแบบ

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

เมื่อ x เป็นตัวแปรสุ่มที่มีการแจกแจงแบบเอฟ

m และ n เป็นเลขจำนวนเต็มบวก แทนระดับขั้นความเป็นอิสระ

การแจกแจงแบบเอฟนี้จะใช้ในการหาค่าวิกฤติสำหรับการทดสอบสมมติฐานด้วยตัวสถิติทดสอบเอฟ ซึ่งระดับขั้นความเป็นอิสระจะขึ้นอยู่กับลักษณะและขนาดตัวอย่างที่ใช้ในการวิเคราะห์สถานการณ์นั้นๆ แสดงรูปกราฟได้ดังนี้



รูปที่ 2.4 กราฟแสดงการแจกแจงแบบเอฟที่ระดับขั้นความเป็นอิสระ m และ n

2.1.5 ความสัมพันธ์ระหว่างการแจกแจงแบบเอฟและการแจกแจงแบบบีตา

การศึกษาถึงความสัมพันธ์ระหว่างการแจกแจงแบบเอฟและการแจกแจงแบบบีตานั้นก็เพื่อใช้ในการสร้างโปรแกรมคำนวณหาค่าวิกฤติจากการแจกแจงแบบเอฟ ซึ่ง Hald(1952) ได้อธิบายถึงความสัมพันธ์ระหว่างฟังก์ชันบีตาไม่สมบูรณ์กับการแจกแจงแบบเอฟดังนี้

– ถ้า B คือ ตัวแปรสุ่มบีตาที่มีพารามิเตอร์ (ν_1,ν_2) เราสามารถสร้างตัวแปรสุ่มเอฟ(f) ที่มีระดับขั้นความเป็นอิสระ $2\nu_1$ และ $2\nu_2$ ได้จากความสัมพันธ์ดังต่อไปนี้

$$f = \frac{\upsilon_1 x}{\upsilon_2(1-x)} \quad ;0 \le x \le 1$$

โดย x คือค่าที่สอดคล้องกับ $I_x(\nu_1,\nu_2)$ ซึ่งก็คือฟังก์ชันความน่าจะเป็นสะสมของตัวแปรสุ่มบีตา หรือที่เรียกว่าอัตราส่วนของฟังก์ชันบีตาที่ไม่สมบูรณ์(Incomplete Beta function ratio) ซึ่งมีรูปแบบดังต่อไปนี้

$$F(x) = \frac{1}{B(\upsilon_1,\upsilon_2)}\int_0^x w^{\upsilon_1-1}(1-w)^{\upsilon_2-1}\,dw$$

$$= \frac{B_x(\upsilon_1,\upsilon_2)}{B(\upsilon_1,\upsilon_2)}$$

$$= I_x(\nu_1,\nu_2)$$

โดยที่ $B(\upsilon_1,\upsilon_2)=\int_0^1 w^{\upsilon_1-1}(1-w)^{\upsilon_2-1}dw$ คือ ฟังก์ชันบีตา(Beta function)

$B_x(\upsilon_1,\upsilon_2)=\int_0^x w^{\upsilon_1-1}(1-w)^{\upsilon_2-1}dw$ คือ ฟังก์ชันบีตาที่ไม่สมบูรณ์(Incomplete Beta function)

จากขั้นตอนดังกล่าวข้างต้นเมื่อกำหนดระดับนัยสำคัญ α และระดับขั้นความเป็นอิสระ m และ n สามารถหาค่าวิกฤติจากการแจกแจงแบบเอฟได้จากโปรแกรมย่อย SUBROUTINE MDBTI(XF,AF,BF,PF,IER) ดังแสดงในภาคผนวก ก โดย XF คือค่าที่สอดคล้องอัตราส่วนของฟังก์ชันบีตาที่ไม่สมบูรณ์ดังที่กล่าวข้างต้น PF = 1-α AF = m/2 และ BF = n/2 ดังนั้นค่าวิกฤติ $F_{\alpha,(m,n)}$ สามารถหาได้จากความสัมพันธ์ต่อไปนี้

$$F_{\alpha(m,n)}=\frac{n(XF)}{m(1-XF)}$$

2.2 สถิติทดสอบในการวิจัย

ในการวิจัยครั้งนี้จะทำการศึกษาเปรียบเทียบวิธีการทดสอบเทียบความกลมกลืนสำหรับตัวแบบการถดถอยด้วยตัวสถิติต่างๆ ดังนี้

2.2.1 ตัวสถิติทดสอบเอฟ

ตัวสถิติทดสอบเอฟ หรือการทดสอบเทียบความกลมกลืน

$$F=\frac{MSLF}{MSPE}$$

เมื่อ MSLF คือ ค่ากำลังสองเฉลี่ยของการเทียบความกลมกลืน (lack of fit mean square)

MSPE คือ ค่ากำลังสองเฉลี่ยของความคลาดเคลื่อนโดยตรง (pure error mean square) ซึ่งสถิติทดสอบเอฟนี้มีข้อจำกัดในการคำนวณค่าสถิติ คือตัวแปรอิสระจะต้องมีค่าที่ซ้ำกัน หรือสามารถแบ่งออกเป็นระดับได้ k ระดับ ในแต่ละระดับนั้นจะบรรจุค่าของตัวแปรตาม n_i หน่วย คือ $y_1^{(i)},y_2^{(i)},\ldots,y_{n_i}^{(i)}$ เมื่อ $1\le i\le k$ และ $n_1+n_2+\ldots+n_k=n$ จะได้

$$MSPE=\frac{SSPE}{n-k}=\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_j^{(i)}-\bar{y}^{(i)}\right)^2 \quad ;\bar{y}^{(i)}=\frac{1}{n_i}\sum_{j=1}^{n_i}y_j^{(i)}$$

และ $$MSLF=\frac{1}{k-p-1}(SSE-SSPE)$$

เมื่อ SSE คือ ผลรวมกำลังสองของความคลาดเคลื่อน (sum of square due to error) จากตารางการวิเคราะห์ความแปรปรวน

เราจะปฏิเสธสมมติฐานว่างเมื่อ $F>F_{\alpha,(k-p-1,n-k)}$ เมื่อ $F_{\alpha\,(k-p-1,n-k)}$ แทนค่าจากตารางการแจกแจงแบบเอฟด้วยระดับนัยสำคัญ α ระดับขั้นความเป็นอิสระคือ k-p-1 และ n-k

2.2.2 ตัวสถิติทดสอบ KS[1]

ในการคำนวณค่าของตัวสถิติทดสอบ Kolmogorov-Smirnov และตัวสถิติทดสอบ Cramer-von Mises นั้นจะคำนวณบนพื้นฐานของค่าประมาณ $R_n(x)$ ซึ่งเป็นฟังก์ชันของค่าความคลาดเคลื่อนจากการประมาณตัวแบบถดถอย มีขั้นตอนการคำนวณดังนี้

$$R_n(x) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} 1_{\{X_i \le x\}}\left[Y_i - \underset{\sim}{X_i}'\underset{\sim}{\hat{\beta}}\right]$$

เมื่อ $\underset{\sim}{X_i}$ เป็นเวกเตอร์ที่ i ของตัวแปรอิสระ

$\underset{\sim}{\hat{\beta}}$ เป็นเวกเตอร์ของค่าประมาณสัมประสิทธิ์การถดถอย

$1_{\{X_i \le x\}}$ เป็นฟังก์ชันดัชนี โดยที่ $1_{\{X_i \le x\}} = \begin{cases} 1 & ; X_i \le x \\ \text{อื่นๆ} \end{cases}$

และ $1 \le i \le n$

จากนิยามของการทดสอบเทียบความกลมกลืนด้วยตัวสถิติทดสอบ Kolmogorov-Smirnov คือ

$$D = \sup_x |F_n(x) - F(x)|$$

เมื่อ F(x) คือ ฟังก์ชันการแจกแจงสะสมของ x (cumulative distribution function)

$F_n(x)$ คือ ฟังก์ชันการแจกแจงเชิงตัวอย่างของ x (empirical distribution function)

ดังนั้นเมื่อประยุกต์ใช้ค่าประมาณ $R_n(x)$ จะสามารถเขียนสูตรการคำนวณตัวสถิติทดสอบ KS ในรูปแบบของตัวสถิติทดสอบตัวใหม่ ได้ดังนี้

$$D_n = \sup_x |R_n(x)|$$

เราจะปฏิเสธสมมติฐานว่างเมื่อ $D_n > D_{B\times\gamma}$ เมื่อ $D_{B\times\gamma}$ คือค่าวิกฤติจากการสุ่มตัวอย่างแบบบูตสแตรปจำนวน B รอบในการวิจัยนี้กำหนดให้เท่ากับ 500 รอบที่เปอร์เซ็นไทล์ที่ γ ซึ่งรายละเอียดการคำนวณได้แสดงไว้ในบทที่ 3

2.2.3 ตัวสถิติทดสอบ CvM[2]

พิจารณาเช่นเดียวกับการคำนวณตัวสถิติทดสอบ KS กล่าวคือ คำนวณบนพื้นฐานของค่าประมาณ $R_n(x)$ ซึ่งจากนิยามของการทดสอบเทียบความกลมกลืนด้วยตัวสถิติทดสอบ CvM คือ

---

[1] W. Stute, W. Gonzalez Manteiga and M. Presedo Quindimil. 'Bootstrap Approximations in Model Checks for Regression' . Journal of the American Statistical Association, Vol. 93 (1998) : 141-149.

[2] W. Stute, W. Gonzalez Manteiga and M. Presedo Quindimil. 'Bootstrap Approximations in Model Checks for Regression' . Journal of the American Statistical Association, Vol. 93 (1998) : 141-149.

$$W^2 = n\int_{-\infty}^{+\infty}[F_n(x) - F(x)]^2 dF(x)$$

เมื่อ F(x) คือ ฟังก์ชันการแจกแจงสะสมของ x (cumulative distribution function)

$F_n(x)$ คือ ฟังก์ชันการแจกแจงเชิงตัวอย่างของ x (empirical distribution function)

ดังนั้นเมื่อประยุกต์ใช้ค่าประมาณ $R_n(x)$ จะสามารถเขียนสูตรการคำนวณตัวสถิติทดสอบ CvM ในรูปแบบของตัวสถิติทดสอบตัวใหม่ ได้ดังนี้

$$W_n^2 = \sum_{i=1}^{k}[R_n(x_i)]^2 F_n(x_i)$$

เมื่อ $F_n(x_i)$ คือ ฟังก์ชันการแจกแจงเชิงตัวอย่างของ x เมื่อ $x = x_i$

k คือ จำนวนระดับของตัวแปรอิสระ และ k = n ในกรณีที่ตัวแปรอิสระมีค่าไม่ซ้ำกัน

เราจะปฏิเสธสมมติฐานว่างเมื่อ $W_n^2 > W_{B\times\gamma}^2$ เมื่อ $W_{B\times\gamma}^2$ คือค่าวิกฤติจากการสุ่มตัวอย่างแบบบูตสแตรปจำนวน B รอบในการวิจัยนี้กำหนดให้เท่ากับ 500 รอบที่เปอร์เซ็นไทล์ที่ γ ซึ่งรายละเอียดการคำนวณได้แสดงไว้ในบทที่ 3

2.3 การวิเคราะห์การถดถอย

การวิเคราะห์การถดถอย เป็นเทคนิคการพยากรณ์ทางสถิติหนึ่งที่เกี่ยวข้องกับการสร้างตัวแบบทางคณิตศาสตร์ เพื่อแสดงความสัมพันธ์ระหว่างตัวแปรสองกลุ่ม การสร้างตัวแบบแสดงความสัมพันธ์ดังกล่าวมีวัตถุประสงค์เพื่อการพยากรณ์และการอนุมานอื่นๆ รูปแบบความสัมพันธ์เชิงคณิตศาสตร์ที่ได้ เรียกว่า "ตัวแบบการถดถอย" หรือ "สมการการถดถอย"

ตัวแบบการถดถอยสามารถเขียนได้ดังนี้

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

เมื่อ $\underset{\sim}{Y}$ เป็นเวกเตอร์ของตัวแปรตามขนาด n×1

X เป็นเมตริกซ์ของตัวแปรอิสระขนาด n×(p+1) โดย X_0 มีค่าเท่ากับ 1

$\underset{\sim}{\beta}$ เป็นเวกเตอร์ของค่าพารามิเตอร์สัมประสิทธิ์การถดถอยขนาด (p+1)×1

$\underset{\sim}{\varepsilon}$ เป็นเวกเตอร์ของความคลาดเคลื่อนขนาด n×1 จากข้อตกลงเบื้องต้นจะได้ว่า

$$E(\underset{\sim}{\varepsilon}) = 0 \quad \text{และ} \quad V(\underset{\sim}{\varepsilon}) = \sigma^2\mathbf{I}_n$$

การประมาณค่าสัมประสิทธิ์การถดถอยโดยวิธีกำลังสองน้อยที่สุด เป็นการหาค่าประมาณของ $\beta_0, \beta_1, \ldots, \beta_p$ ในเทอมของตัวแปรอิสระและตัวแปรตาม เพื่อที่จะทำให้ผลบวกของความคลาดเคลื่อนกำลังสองมีค่าต่ำที่สุด ให้ $b_0, b_1, \ldots, b_p$ เป็นค่าประมาณของ $\beta_0, \beta_1, \ldots, \beta_p$ ตามลำดับ ซึ่งจะเขียนในรูปเวกเตอร์ $\underset{\sim}{b}$ ความคลาดเคลื่อนของสมการการถดถอยจึงเขียนได้เป็น

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

และผลบวกกำลังสองของความคลาดเคลื่อนจึงเขียนได้เท่ากับ

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

ซึ่งจะมีค่าต่ำสุดเมื่อ

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

แทนค่าเวกเตอร์ $\underset{\sim}{\beta}$ ด้วยเวกเตอร์ $\underset{\sim}{b}$ จะได้สมการปกติเป็น

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

เนื่องจากแรงค์(rank) ของ X เท่ากับ p+1 ดังนั้น X'X จึงมีแรงค์เท่ากับ p+1 ด้วย เมตริกซ์ X'X จึงมีเมตริกซ์ผกผัน การแก้สมการข้างต้นจะได้ว่า

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

2.4 ความผิดพลาดในการทดสอบสมมติฐานทางสถิติ

ในการทดสอบสมมติฐานทางสถิติโดยทั่วไปแล้วผลการทดสอบจะเกิดความผิดพลาดได้ ซึ่งความผิดพลาดดังกล่าวนี้แบ่งออกได้เป็น 2 ประเภทคือ ความผิดพลาดประเภทที่ 1 (type I error) และความผิดพลาดประเภทที่ 2 (type II error) ลักษณะของความผิดพลาดทั้ง 2 ประเภทนี้แสดงได้ดังตารางต่อไปนี้

ตารางที่ 2.1 แสดงความผิดพลาดในการทดสอบสมมติฐานทางสถิติ

สมมติฐานว่าง H_0	การตัดสินใจ	
	ปฏิเสธ H_0	ยอมรับ H_0
เป็นจริง	ความผิดพลาดประเภทที่ 1(α)	ตัดสินใจถูกต้อง
เป็นเท็จ	ตัดสินใจถูกต้อง	ความผิดพลาดประเภทที่ 2(β)

ในการทดสอบสมมติฐานทางสถิติ ผู้ทำการทดสอบไม่ต้องการให้เกิดความผิดพลาดทั้ง 2 ประเภทคือทั้ง α และ β แต่ในทางปฏิบัติผู้ทดลองไม่สามารถหลีกเลี่ยงความผิดพลาดดังกล่าวได้ ดังนั้นก็จะให้เกิดค่าความผิดพลาดทั้งสองน้อยที่สุด การเปรียบเทียบอำนาจการทดสอบซึ่งมีค่าเท่ากับ $1-\beta$ จะมีความน่าเชื่อถือได้มากน้อยเพียงใดจะต้องพิจารณาถึงความสามารถในการควบคุมความผิดพลาดประเภทที่ 1 ด้วย เพราะหากว่าไม่สามารถควบคุมความผิดพลาดประเภทที่ 1 ได้

กล่าวคือความผิดพลาดประเภทที่ 1 มีค่าสูงกว่าค่าเกณฑ์ที่ใช้ในการพิจารณา หรือระดับนัยสำคัญที่กำหนด จะส่งผลให้ค่าความผิดพลาดประเภทที่ 2 มีค่าต่ำไปด้วย ซึ่งก็หมายถึงว่าค่าอำนาจการทดสอบนั้นมีค่าสูงตามไปด้วย จึงไม่สามารถเชื่อถือได้ว่าการทดสอบนั้นๆ เหมาะสมหากพิจารณาจากค่าอำนาจการทดสอบดังกล่าว

2.5 การจำลองโดยใช้เทคนิคมอนติคาร์โล

เทคนิคในการจำลองตัวแบบทางคณิตศาสตร์มีอยู่หลายวิธี เทคนิคมอนติคาร์โลเป็นวิธีหนึ่งที่นิยมใช้กันอย่างแพร่หลายในปัจจุบัน ซึ่งหลักการของเทคนิคมอนติคาร์โลนั้นเป็นการจำลองหรือผลิตตัวเลขสุ่ม(Random number) มาช่วยในการหาคำตอบที่ต้องการศึกษา ในการวิจัยครั้งนี้จะใช้เทคนิคมอนติคาร์โลดังกล่าวในการผลิตข้อมูลที่มีลักษณะการแจกแจงตามที่ต้องการ ซึ่งขั้นตอนของเทคนิคมอนติคาร์โลที่ใช้กันอยู่ในปัจจุบันแบ่งออกเป็น 3 ขั้นตอนดังนี้

2.5.1 การผลิตตัวเลขสุ่ม

การใช้ตัวเลขสุ่มเป็นสิ่งที่สำคัญมากในเทคนิคมอนติคาร์โล เพราะว่าหลักการของเทคนิคมอนติคาร์โลนั้น จะใช้ตัวเลขสุ่มมาช่วยในการหาคำตอบของปัญหา ซึ่งลักษณะของตัวเลขสุ่มนั้นจะมีการแจกแจงแบบสม่ำเสมอ(Uniform distribution) ในช่วง(0,1) ที่มีคุณสมบัติดังนี้

ก) ตัวเลขที่ได้มีการกระจายของความน่าจะเป็นแบบสม่ำเสมอและเป็นอิสระซึ่งกันและกัน

ข) อนุกรมของตัวเลขที่ได้สามารถสร้างซ้ำเดิมได้(reproducible)

ค) อนุกรมของตัวเลขไม่ซ้ำค่าเดิมในช่วงที่ต้องการใช้ตัวเลขสุ่ม หรือได้ตัวเลขที่มีขนาดของอนุกรมยาวพอสำหรับการใช้งาน

ง) ใช้เวลาสั้นๆ ในการสร้างตัวเลขแบบสุ่ม

จ) ใช้หน่วยความจำของเครื่องคอมพิวเตอร์น้อย

สำหรับโปรแกรมย่อยที่ใช้ในการสร้างเลขสุ่มนี้คือ FUNCTION RAND(IX) ดังที่แสดงไว้ในภาคผนวก ก

2.5.2 การประยุกต์ปัญหาที่ต้องการศึกษาเพื่อใช้กับตัวเลขสุ่ม

ในการวิจัยครั้งนี้ต้องการศึกษาเปรียบเทียบวิธีการทดสอบเทียบความกลมกลืนโดยพิจารณาจากค่าความผิดพลาดประเภทที่ 1 และค่าอำนาจการทดสอบ ซึ่งในขั้นตอนการดำเนินงานจะใช้ตัวเลขสุ่มเป็นพื้นฐานในการผลิตข้อมูลให้มีการแจกแจงแบบต่างๆ ภายใต้เงื่อนไขและสถานการณ์ที่กำหนด นอกจากนี้ในขั้นตอนการหาค่าวิกฤติเพื่อการทดสอบสมมติฐานด้วยตัวสถิติทดสอบ KS และตัวสถิติทดสอบ CvM จะใช้ตัวเลขสุ่มสำหรับทำการสุ่มตัวอย่างซ้ำแบบใส่คืนในการสุ่มตัวอย่างแบบบูตสแตรป

2.5.3 การทดลองกระทำซ้ำ(Replication)

จากการประยุกต์ปัญหาเพื่อใช้กับตัวเลขสุ่มแล้ว จะทำการทดลองโดยใช้กระบวนการสุ่ม(Random process) มากระทำในลักษณะที่ซ้ำๆ กัน เพื่อค้นหาคำตอบของปัญหาที่เราต้องการศึกษา ซึ่งในที่นี้คือการวนรอบตั้งแต่การผลิตเลขสุ่มและผลิตข้อมูล จนถึงการคำนวณค่าสถิติทดสอบและค่าวิกฤติเพื่อพิจารณาเปรียบเทียบหาผลสรุปของการทดสอบสมมติฐานเป็นจำนวน 1,000 รอบ แล้วคำนวณหาค่าความผิดพลาดประเภทที่ 1 และค่าอำนาจการทดสอบเพื่อใช้พิจารณาเปรียบเทียบตัวสถิติทดสอบเอฟ KS และ CvM ต่อไป

2.6 การสุ่มตัวอย่างแบบบูตสแตรป

วิธีการการสุ่มตัวอย่างแบบบูตสแตรปนี้เป็นวิธีที่นำเสนอโดย แบรดเลย์ เอฟรอน (Bradley Efron) ในปี ค.ศ. 1979 โดยมีหลักเกณฑ์ดังนี้คือ ข้อมูลที่เก็บรวบรวมมาจะทำการสุ่มตัวอย่างแบบใส่คืน (with replacement) ขนาดเท่ากับจำนวนตัวอย่างหรือข้อมูลที่มีอยู่ เพื่อสร้างข้อมูลชุดใหม่แล้วนำมาใช้ในการประมาณค่าพารามิเตอร์ที่สนใจ

ในการวิจัยครั้งนี้จะใช้เทคนิคการสุ่มตัวอย่างแบบบูตสแตรปในการหาค่าวิกฤติ สำหรับการทดสอบเทียบความกลมกลืนด้วยตัวสถิติทดสอบ KS และตัวสถิติทดสอบ CvM มีขั้นตอนในการทำงานดังนี้

2.6.1 สร้างเลขสุ่มจากการแจกแจงแบบสม่ำเสมอในช่วง(0,1) เพื่อนำไปใช้ในการสุ่มตัวอย่างแบบใส่คืน ซึ่งจะทำการสุ่มตัวอย่างจำนวน 500 ชุด

2.6.2 จากตัวอย่างแต่ละชุดที่ได้จากการสุ่มตัวอย่างแบบใส่คืนข้างต้น นำมาคำนวณหาค่าประมาณของพารามิเตอร์หรือในที่นี้คือตัวสถิติทดสอบ KS และตัวสถิติทดสอบ CvM

2.6.3 นำค่าสถิติทดสอบที่ได้มาเรียงลำดับจากน้อยไปหามาก เพื่อหาค่าวิกฤติในตำแหน่งเปอร์เซ็นไทล์ที่ต้องการ กล่าวคือ

ค่าสถิติทดสอบอันดับที่ 495 คือค่าวิกฤติที่เปอร์เซ็นไทล์ที่ 99 หรือที่ระดับนัยสำคัญ 0.01

ค่าสถิติทดสอบอันดับที่ 475 คือค่าวิกฤติที่เปอร์เซ็นไทล์ที่ 95 หรือที่ระดับนัยสำคัญ 0.05

ค่าสถิติทดสอบอันดับที่ 450 คือค่าวิกฤติที่เปอร์เซ็นไทล์ที่ 90 หรือที่ระดับนัยสำคัญ 0.10

ตัวอย่างของวิธีการสุ่มตัวอย่างแบบใส่คืนและโปรแกรมการทำงานนั้นจะเสนอไว้ในภาคผนวก