การประมาณค่าตัวแบบการถดถอยแบบไร้พารามิเตอร์ด้วยชุดคำสั่ง SemiPar ในโปรแกรมอาร์

The Estimation of Nonparametric Regression Model with SemiPar Package in R Program

อัชฌา อระวีพร

Autcha Araveeporn

สาขาวิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง กรุงเทพมหานคร

บทคัดย่อ

วิธีการปรับให้เรียบที่เรียกว่า Penalized Spline เป็นวิธีการหนึ่งในการประมาณค่าฟังก์ชันจากตัวแบบการถดถอยแบบไร้พารามิเตอร์ ซึ่งวิธีการนี้ในการคำนวณค่อนข้างยุ่งยากเพราะต้องมีการปรับค่าที่เรียกว่า ค่าพารามิเตอร์ปรับให้เรียบ (smoothing parameter) ให้เหมาะสมกับฟังก์ชันที่ต้องการประมาณ บทความนี้ได้นำเสนอกลุ่มคำสั่ง SemiPar ในโปรแกรมอาร์ที่ช่วยในการประมาณค่าจากวิธี Penalized Spline ทำให้สะดวกรวดเร็วมากยิ่งขึ้น นอกจากนี้ยังแสดงตัวอย่างเพื่อเปรียบเทียบประสิทธิภาพและการประมาณค่าจากวิธีของ Penalized Spline สมการถดถอยเชิงเส้นอย่างง่าย และสมการถดถอยกำลังสอง

คำสำคัญ : พารามิเตอร์ปรับให้เรียบ, วิธีการปรับให้เรียบ, สมการถดถอยแบบไร้พารามิเตอร์

Abstract

The Penalized Spline method is one method of smoothing method for estimating function on nonparametric regression model. The Penalized Spline method has a smoothing parameter to adjust the desired flexibility function, so it is very complicated to evaluate function. This paper provides a command of SemiPar package in Program R that can help the user to estimate unknown function, while there is the example to compare the efficiency and estimation of Penalized Spline method, linear regression method , and polynomial regression method.

Keywords : Nonparametric Regression Method, Smoothing method, Smoothing Parameter

---

E-mail address : kaautcha@kmitl.ac.th

1. บทนำ

การวิเคราะห์การถดถอย (Regression Analysis) เป็นวิธีการวิเคราะห์ข้อมูลเพื่อหารูปแบบความสัมพันธ์ของตัวแปรตาม (y) และตัวแปรอิสระ (x) ตั้งแต่หนึ่งตัวขึ้นไป เป็นวิธีการทางสถิติที่สร้างสมการถดถอยเพื่อใช้พยากรณ์ตัวแปรตาม ซึ่งวิธีการนี้มีข้อสมมติเบื้องต้น คือประชากรต้องมีการแจกแจงแบบปกติ ตัวแปรตามและตัวแปรอิสระมีความสัมพันธ์เชิงเส้นตรง ความแปรปรวนของความคลาดเคลื่อนมีความคงที่ และตัวแปรอิสระต้องมีความสัมพันธ์ซึ่งกันและกัน (Multicollinearity) ซึ่งบางครั้งข้อมูลที่ต้องการใช้ไม่เป็นไปตามข้อสมมติเบื้องต้นทำให้อาจเกิดความความคลาดเคลื่อนในการพยากรณ์ได้ เพื่อแก้ปัญหาการวิเคราะห์การถดถอยที่มีข้อสมมติเบื้องต้นมากจึงได้มีการสร้างตัวแบบการถดถอยแบบไร้พารามิเตอร์ (Nonparametric Regression Model)

ตัวแบบการถดถอยแบบไร้พารามิเตอร์ หรือเรียกว่าวิธีการปรับให้เรียบ (Smoothing Methods) มีหลายวิธีที่นิยมใช้ เช่น วิธี local polynomial regression [1,2] สมการถดถอยแบบสไปลน์ (regression splines) [3,4] smoothing splines [5,6] และ Penalized Spline [7] วิธีเหล่านี้ไม่มีข้อสมมติเบื้องต้นของข้อมูลแต่เป็นสร้างฟังก์ชันของตัวแปรตาม กับตัวแปรอิสระเพื่อพิจารณาปรับค่าของตัวแปรตามให้เรียบไปตามค่าของตัวแปรอิสระซึ่งขึ้นอยู่กับค่าคงที่ที่เรียกว่า แบนวิทช์ (Bandwidth) หรือ ค่าพารามิเตอร์ปรับให้เรียบ (smoothing parameter)

วิธีการหาค่าฟังก์ชันของตัวแปรตามในแต่ละวิธีมีการคำนวณที่ค่อนข้างยุ่งยากจึงได้มีการนำเอาโปรแกรมคอมพิวเตอร์มาช่วยในการสร้างฟังก์ชัน โปรแกรมอาร์ (R Program) เป็นโปรแกรมหนึ่งที่มีการสร้างชุดคำสั่งสำหรับวิธี Penalized Spline เรียกว่า ชุดคำสั่ง SemiPar เป็นชุดคำสั่งที่สร้างขึ้นเพื่อสร้างฟังก์ชันปรับให้เรียบ สำหรับตัวแปรอิสระตัวเดียวและมากกว่า นอกจากนี้ยังสามารถพยากรณ์ข้อมูลของตัวแปรตามในอนาคตได้อีกด้วย

โปรแกรมอาร์ เป็นโปรแกรมที่เขียนเป็นภาษาเพื่อใช้ในการวิเคราะห์ข้อมูลทางสถิติ ผู้ที่ต้องการใช้สามารถดาวน์โหลดโดยไม่เสียค่าใช้จ่ายที่เว็บไซต์ http://www.R-project.org ซึ่งโปรแกรมนี้ได้พัฒนามาจากภาษาเอส (S Language) โดย Becker และ Chambers [8] Becker และคณะ [9] Chambers และ Hastie [10] และ Chambers [11] ในโปรแกรมอาร์มีชุดคำสั่งสำหรับวิเคราะห์ข้อมูลในแต่ละประเภทอยู่มากมายแล้วขึ้นอยู่กับผู้ใช้ว่าต้องการใช้เรื่องใดก็สามารถดาวน์โหลดได้โดยไม่เสียค่าใช้จ่าย

ในบทความนี้ได้นำเสนอตัวแบบของสถิติแบบไร้พารามิเตอร์ วิธี Penalized Spline เพื่อใช้สำหรับสร้างฟังก์ชันการปรับให้เรียบ และการเขียนชุดคำสั่ง SemiPar ในโปรแกรมอาร์

2.วิธี Penalized Spline

การสร้างตัวแบบการถดถอยแบบไร้พารามิเตอร์สามารถเขียนฟังก์ชันอย่างง่ายได้ดังนี้

$$y_t = f(x_t) + \varepsilon_t, \quad t = 1,2,...,n \tag{1}$$

เมื่อ $x_t, t = 1,2,...,n$ คือตัวแปรอิสระ $y_t, t = 1,2,...,n$ คือตัวแปรตาม $f(x_t)$ เป็นฟังก์ชันการถดถอยแบบไร้พารามิเตอร์ที่ต้องการประมาณ และ $\varepsilon_t, t = 1,2,...,n$ แทนค่าความคลาดเคลื่อน

Eubank [3] และ Eubank [4] ได้เสนอวิธีการสร้างสมการถดถอยแบบสไปลน์ (regression spline) เป็นวิธีเฉพาะโดยพิจารณาขอบเขตในกลุ่มของค่า

$$\tau_0, \tau_1, \tau_2, ..., \tau_K, \tau_{K+1} \tag{2}$$

ซึ่งจะอยู่ในช่วง $[a,b]$ ที่ขอบเขต $a = \tau_0 < \tau_1 < .... < \tau_K < \tau_{K+1} = b$ โดยกำหนดค่าขอบเขตเรียกว่า ค่าน็อต (knot) และ $\tau_r, r = 1,2,...,K$ เรียกว่าค่าน็อตภายในหรือค่าน็อตเริ่มต้น

สมการถดถอยแบบสไปลน์สามารถสร้างโดยใช้การยกกำลัง l ค่าของ l กำหนดจากลักษณะของข้อมูล เมื่อข้อมูลมีลักษณะเป็นค่าคงที่ ค่า $l = 0$ ลักษณะเชิงเส้นตรง $l = 1$ ลักษณะกำลังสอง $l = 2$ และลักษณะยกกำลังสาม $l = 3$ กับ ค่าน็อต K ขึ้นอยู่กับฟังก์ชันสูญเสียและค่าขอบเขตภายใน $(\tau_1, \tau_2, ..., \tau_K)$

$$1, x, ..., x^l, (x-\tau_1)_+^l, ..., (x-\tau_K)_+^l \tag{3}$$

เมื่อ w_+^l คือ การยกกำลัง l ของค่าที่เป็นบวก w โดยที่ $w_+ = \max(0, w)$

ในสมการที่ (3) เป็นการใช้การยกกำลังของการแบ่งค่า ซึ่งสามารถนำไปใช้เขียนฟังก์ชันการถดถอยแบบสไปลน์ได้ดังนี้

$$f(x) = \sum_{s=0}^{l} \beta_s x^s + \sum_{r=1}^{K} \beta_{l+r} (x-\tau_r)_+^l \tag{4}$$

เมื่อ $\beta_0, \beta_1, ..., \beta_{l+K}$ เป็นค่าสัมประสิทธิ์ที่ต้องการประมาณเพื่อให้ได้ตัวประมาณที่เหมาะสมโดยใช้วิธีให้ทำให้ฟังก์ชันสูญเสียมีค่าต่ำสุด

จากงานวิจัยข้างต้น Ruppert และ Carroll [12] ได้พัฒนาการถดถอยแบบสไปลน์สู่วิธี Penalized Spline เป็นวิธีการประมาณค่าฟังก์ชันปรับให้เรียบโดยใช้ฟังก์ชันการแบ่งค่าแบบยกกำลัง (Truncated Power Function) ซึ่งสามารถเขียนฟังก์ชันได้ดังนี้

$$f(x_t) = \sum_{j=0}^{m-1} \alpha_j x_t^j + \sum_{l=1}^{K} \beta_l \left| x_t - \tau_l \right|^{2m-1} \tag{5}$$

เมื่อ $\beta = [\beta_1, ..., \beta_K]^T \sim N(0, \sigma_\beta^2 \Omega^{-1/2} (\Omega^{1/2})^{\mathrm{T}})$ กำหนดให้ค่า (a,b) ของ Ω หมายถึง $\left|\tau_a - \tau_b\right|^{2m-1}$ และค่าสัมประสิทธิ์ของ $\left|x_t - \tau_l\right|^{2m-1}$ เป็นค่า penalized ซึ่งนำค่าน็อตที่ลำดับ K มาใช้

ส่วนค่า m ที่ปรากฏในสมการที่ (5) โดยค่า m สามารถกำหนดได้ 4 ค่าคือ 0 1 2 3 แต่ละค่าจะแสดงลักษณะของข้อมูลเป็น ค่าคงที่ (m=0) เชิงเส้นตรง (m=1) รูปแบบกำลังสอง (m=2) และรูปแบบยกกำลังสาม (m=3) ในการกำหนดค่าเหล่านี้ขึ้นอยู่กับลักษณะของข้อมูล

ในบทความนี้สนใจข้อมูลของตัวแปรตามที่มีลักษณะกำลังสอง (quadratic) หรือ m=2 ซึ่งข้อมูลส่วนใหญ่มีลักษณะไม่อยู่ในรูปแบบเชิงเส้นสามารถเขียนฟังก์ชันได้ ดังนี้

$$f(x_t,\theta)=\alpha_0+\alpha_1 x_t+\sum_{l=1}^{K}\beta_l\left|x_t-\tau_l\right|^3 \qquad (6)$$

เมื่อ $\theta=(\alpha_0,\alpha_1,\beta_1,...,\beta_K)^T$ คือค่าสัมประสิทธิ์ของตัวแบบการถดถอยแบบไร้พารามิเตอร์ และค่าน็อต คือ $\tau_1<\tau_2<...<\tau_K$ ซึ่งค่า K สามารถเลือกใช้ได้จากวิธีพิจารณาความแตกต่างกันระหว่างข้อมูลจริงและค่าที่ประมาณได้ (cross validation method) หรือเกณฑ์การพิจารณาตัวแบบ เช่น BIC หรือ AIC

ตัวประมาณของ Penalized Spline ($\hat{f}(.)$) สามารถเขียนให้อยู่ในรูปเมตริกซ์ได้ดังนี้

$$\hat{f}=C(C^TC+\lambda^3D)^{-1}C^Ty \qquad (7)$$

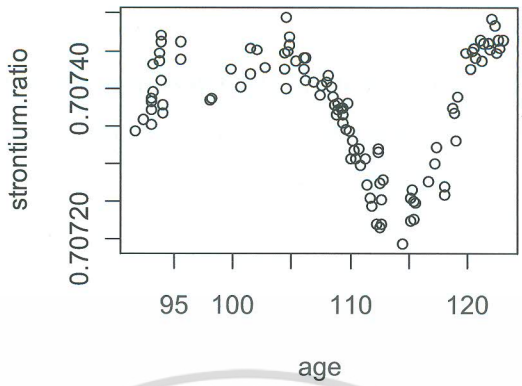
เมื่อกำหนดให้ λ คือค่าพารามิเตอร์ปรับให้เรียบ (Smoothing Parameter) และ $\hat{f}(x)$ เป็นตัวประมาณไม่เอนเอียงเชิงเส้นที่ดีที่สุด (EBLUP) ([13]) และเมตริกซ์ C และ D คือ

$$C=\begin{bmatrix}1 & x_t & \left|x_t-\tau_l\right|^3_{1\le l\le K}\end{bmatrix}_{1\le t\le n} \qquad D=\begin{bmatrix}0_{2\times 2} & 0_{2\times K}\\ 0_{K\times 2} & (\Omega_K^{1/2})^T\Omega_K^{1/2}\end{bmatrix}$$

ชุดคำสั่ง SemiPar ในโปรแกรมอาร์ จะช่วยในการประมาณค่าตัวประมาณของ Penalized Spline ($\hat{f}(.)$)

3. ตัวอย่าง

ข้อมูลที่นำมาวิเคราะห์นี้เก็บโดย Bralower และคณะ [14] เป็นการวัดข้อมูลอัตราส่วนของไอโซโทปสโตรเทียม (strontium isotope) ซึ่งสามารถพบได้ในซากดึกดำบรรพ์ของหอยให้เป็นตัวแปรตาม และอายุของซากหอยดึกดำบรรพ์มีหน่วยเป็นล้านปี ให้เป็นตัวแปรอิสระ ทั้งหมด 160 ค่า ลักษณะการกระจายของข้อมูลดังรูปที่ 1



รูปที่ 1 แสดงแผนภาพการกระจายของอัตราส่วนไอโซโทปสโตรเทียมและอายุของซากหอยดึกดำบรรพ์

ในการประมาณค่าค่าตัวประมาณของ Penalized Spline ($\hat{f}(.)$) สามารถเรียกชุดคำสั่ง SemiPar จากโปรแกรมอาร์ด้วยคำสั่ง (ในโปรแกรมอาร์จะแสดงเครื่องหมาย > ก่อนเขียนคำสั่ง)

```
> library(SemiPar)
```

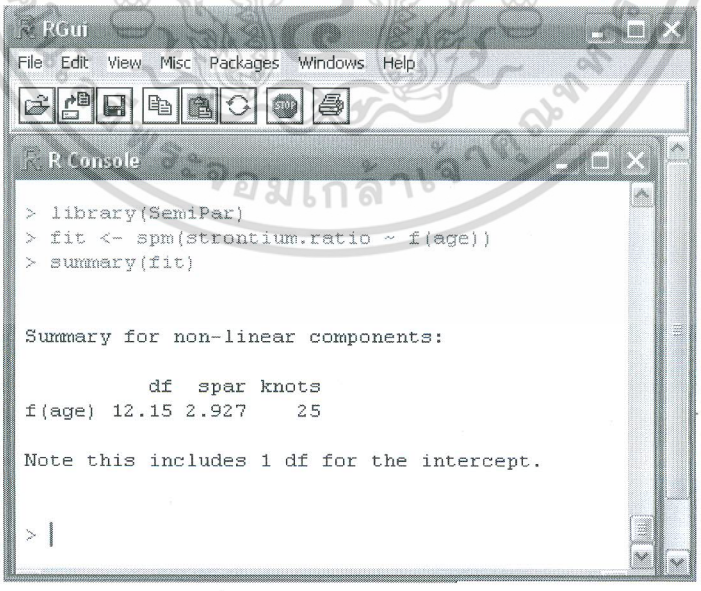
หลังจากนั้นประมาณโดยกำหนดให้ตัวแปรตามชื่อ strontium.ratio และตัวแปรอิสระชื่อ age ด้วยคำสั่ง

```
> fit1 <- spm(strontium.ratio ~ f(age))
```

ถ้าต้องการดูผลลัพธ์ให้กำหนดคำสั่ง

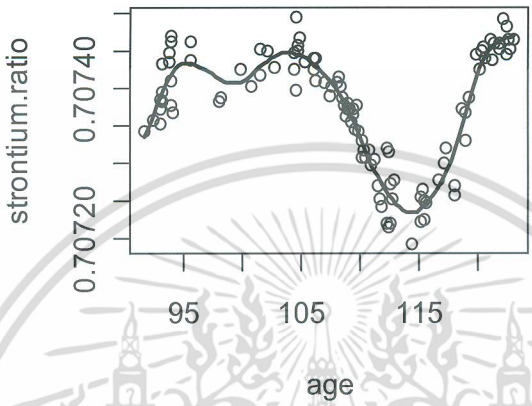
```
> summary(fit1)
```

จะปรากฏผลลัพธ์ดังรูปที่2



รูปที่ 2 แสดงผลลัพธ์จากโปรแกรมอาร์

จากผลลัพธ์ที่ได้ค่า df คือค่าผลรวมเส้นทแยงมุมของ เมตริกซ์ $C(C^TC+\lambda^3 D)^{-1}C^T$ ค่า spar คือค่า λ หรือเรียกว่าค่า พารามิเตอร์ปรับให้เรียบ และค่า knots แทนค่าน็อต ซึ่งผลลัพธ์ที่จากการประมาณค่าสามารถเขียนเป็นกราฟเส้นแสดงดังรูปที่ 3



รูปที่ 3 แสดงกราฟที่ได้จากการประมาณด้วยวิธี Penalized Spline

จากตัวอย่างข้อมูลชุดเดิมสามารถนำมาวิเคราะห์ด้วยวิธีการวิเคราะห์การถดถอยเชิงเส้นอย่างง่ายจากคำสั่ง

```
> fit2 <- lm(strontium.ratio ~ age)
```

ได้สมการถดถอยจากวิธีกำลังสองน้อยที่สุด ดังนี้ $\hat{y} = 0.7075 - 1.43\times10^{-6}x$ ดังรูปที่ 4 และถ้าลองปรับสมการให้อยู่ในรูปสมการถดถอยกำลังสองด้วยคำสั่ง

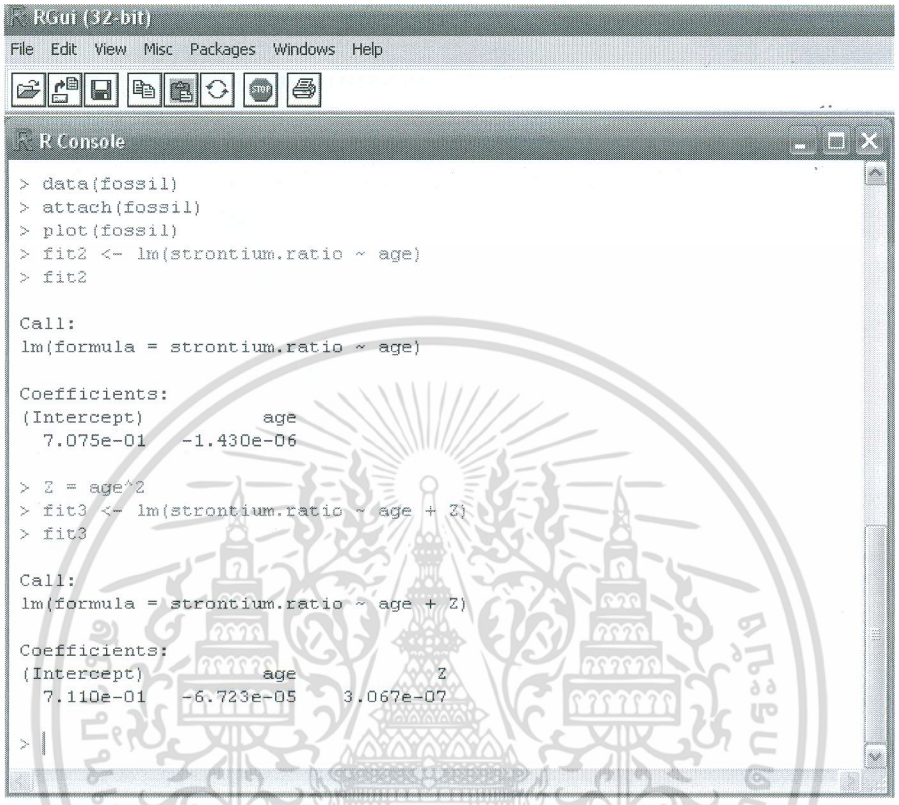
```
> Z = age^2
> fit3 <- lm(strontium.ratio ~ age + Z)
```

ได้สมการถดถอยจากวิธีกำลังสองน้อยที่สุด ดังนี้

$\hat{y} = 0.71103 - 6.723\times10^{-5}x + 3.067\times10^{-7}x^2$ ดังรูปที่ 4

จากรูปแบบสมการถดถอยทั้ง 3 วิธีได้แก่

1. วิธี Penalized Spline จากรูปแบบสมการถดถอยแบบไร้พารามิเตอร์ (จากรูปที่ 2)
2. วิธีการวิเคราะห์ด้วยสมการถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression) (จากรูปที่ 4)
3. วิธีการวิเคราะห์ด้วยสมการถดถอยกำลังสอง (Polynomial Regression) (จากรูปที่ 4)



รูปที่ 4 แสดงผลลัพธ์จากโปรแกรมอาร์ จากวิธีสมการถดถอยเชิงเส้นอย่างง่าย และวิธีการถดถอยยกกำลังสอง

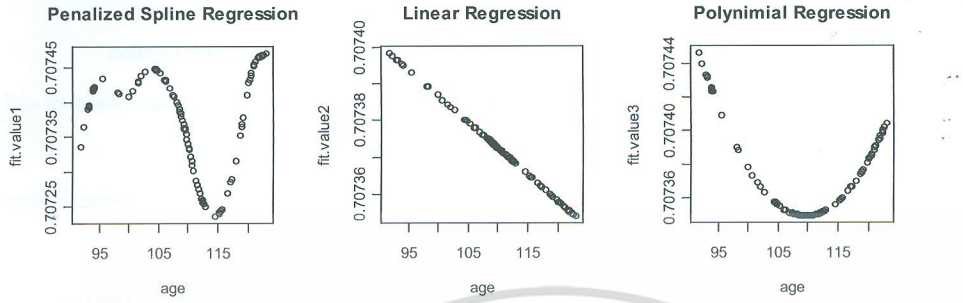
พิจารณาประสิทธิภาพการประมาณด้วยวิธีหาค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (MSE) จากสูตร

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

เมื่อ y_i คือค่าสังเกตจากข้อมูลจริง และ $\hat{y}_i$ คือค่าที่ได้จากการประมาณ ได้ผลดังนี้

วิธี	Penalized Spline	Linear Regression	Polynomial Regression
MSE	5.6059×10^{-10}	5.5612×10^{-9}	4.9040×10^{-9}

จากตารางพบว่าค่า MSE ของวิธี Penalized Spline มีค่าต่ำสุดรองลงมาคือ สมการถดถอยกำลังสองและสมการถดถอยเชิงเส้น หรือสามารถดูการกระจายของค่าประมาณได้จากรูปที่ 5



รูปที่ 5 แสดงการประมาณค่าจากวิธี Penalized Spline วิธีสมการถดถอยเชิงเส้นอย่างง่าย และวิธีการถดถอยยกกำลังสอง

จากรูปที่ 5 จะเห็นได้ว่าลักษณะของข้อมูลมีความสำคัญต่อการเลือกวิธีในการประมาณค่าฟังก์ชัน จากรูปแรกวิธี Penalized Spline เหมาะสำหรับข้อมูลที่มีลักษณะไม่เชิงเส้นและตัวแปรอิสระไม่จำเป็นต้องเป็นข้อมูลเชิงปริมาณก็สามารถใช้เพื่อประมาณค่าฟังก์ชันได้ ส่วนวิธีสมการถดถอยเชิงเส้น (Linear Regression) เหมาะสำหรับตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในเชิงเส้น และวิธีสุดท้ายคือสมการถดถอยเชิงเส้นยกกำลังสอง (Polynomial Regression) เหมาะสำหรับข้อมูลที่มีรูปแบบเป็นเส้นโค้งกำลังสอง ดังนั้นก่อนที่จะเลือกวิธีใดควรสร้างแผนภาพการกระจายเพื่อดูลักษณะการกระจายข้อมูลก่อนที่จะเลือกวิธีที่เหมาะสมในการประมาณค่า

4. บทสรุป

การใช้โปรแกรมสำเร็จรูปที่ไม่มีปัญหาทางลิขสิทธิ์ เช่น โปรแกรมอาร์ ช่วยให้การวิเคราะห์ข้อมูลทางสถิติทำงานได้หลากหลายมากยิ่งขึ้น ในโปรแกรมอาร์มีชุดคำสั่งอยู่มากมาย ชุดคำสั่ง SemiPar ผู้ใช้สามารถดาวน์โหลดได้ฟรีผ่านเครือข่ายอินเตอร์เน็ตโดยเรียกจากโปรแกรมอาร์ ชุดคำสั่งนี้ช่วยในการประมาณค่าฟังก์ชันจากวิธีของ Penalized Spline ได้สะดวกมากขึ้น โดยเฉพาะข้อมูลที่มีลักษณะไม่เชิงเส้น (Non-linear) และข้อมูลไม่เป็นไปตามเงื่อนไขของการประมาณค่าจากการวิเคราะห์การถดถอย เพื่อหลีกเลี่ยงปัญหาดังกล่าว การใช้ตัวแบบการถดถอยแบบไร้พารามิเตอร์ช่วยในการประมาณค่าฟังก์ชันได้ดีกว่า

ในบทความนี้จึงได้อธิบายคำสั่งของการทำงานที่จำเป็นในการประมาณค่าฟังก์ชันจากวิธี Penalized Spline อย่างละเอียดเหมาะสำหรับผู้สนใจที่จะวิเคราะห์ข้อมูลและแปลผลด้วยวิธีนี้

เอกสารอ้างอิง

[1] Wand, M.P. and Jones, M.C., 1995. Kernel Smoothing. Chapman and Hall, London.

[2] Fan, J. and Gijbels, I., 1996. Local Polynomial Modelling and Its Applications. Chapman and Hall, London.

[3] Eubank, R.L., 1988. Spline Smoothing and Nonparametric Regression. Marcel Dekker, New York.

[4] Eubank, R.L., 1999. Nonparametric Regression and Spline Smoothing. Marcel Dekker, New York.

[5] Wahba, G., 1990. Spline Models for Observational Data. , SIAM, Philadelphia, PA.

[6] Green, P.J. and Silverman, B. W., 1994. Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. Chapman and Hall, London.

[7] Ruppert, D., Wand, M.P. and Carroll. R.J., 2003. Semiparametric Regression. Cambridge University Press.

[8] Becker, R.A. and Chambers, J.M., 1984. S. An Interactive Environment for Data Analysis and Graphics. Wadsworth and Brooks/Cole, Monterey.

[9] Becker, R. A., Chambers, J. M., and Wilk, A. R., 1988. The NEW S Language-A Programming Environment for Data Analysis and Graphics. Chapman & Hall, New York.

[10] Chambers, J. M. and Hastie, T. J., 1992. Statistical Models in *S.* Chapman & Hall, NewYork.

[11] Chambers, J. M., 1998. Programming with Data. A Guide to the S Language. Springer-Verlag, New York.

[12] Ruppert, D. and Carroll, R.J., 2000. Spatial-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics*, 42, 205-224.

[13] Robinson, G. K., 1991. That BLUP is a good thing : the estimation of random effects. *Statistical Science*, 6 , 15-51.

[14] Bralower, T.J., Fullagar, P.D., Paull, C.K., Dwyer, G.S. and Leckie, R.M., 1997. Mid-cretaceous strontium-isotope stratigraphy of deep-sea sections. *Geological Society of America Bulletin*, 109, 1421-1442.