修勻學(Graduation)－平滑接點修勻法(Spline)

授課教師：余清祥教授

課程日期：2024年10月23日

資料下載：

http://csyue.nccu.edu.tw

平滑接點修勻(Spline)法

- 實務上，編算生命表時經常會以五齡組為單位計算死亡率，再使用內插法(Interpolation)找出單一年齡死亡率的修勻值。這些內插法必須滿足某些條件(如平滑)，避免內插法在端點或接點產生不協調的現象。

→知名的多項式Spline即屬於內插法的一種，將編算範圍分成幾個區段，每個區段使用可不同的曲線或多項式。

Everett公式

- Everett公式較常用的內插公式，滿足

$$v_{x+s} = F(1-s)u_x + F(s)u_{x+1}, 0 \le s \le 1$$

其中 $F(s) = A(s) + B(s)\delta^2 + C(s)\delta^4 + \cdots$

$$\delta^2 f(x) = f(x+1) - 2f(x) + f(x-1)$$

→δ^2的作用類似MWA，為各加入前後一個死亡率的加權平均，因此Everett公式也會變成幾個原始值的加權平均。

- 常見的非整數年齡假設可視為特例：

→線性內插(Linear Interpolation)

$$S(x+t) = (1-t)\cdot S(x) + t\cdot S(x+1)$$

→指數內插(Exponential Interpolation)

$$S(x+t) = S(x)^{1-t} \times S(x+1)^{t}$$

$$\log(S(x+t)) = (1-t)\log(S(x)) + t\log(S(x+1)).$$

→調和內插(Harmonic Interpolation)

$$\frac{1}{S(x+t)} = \frac{1-t}{S(x)} + \frac{t}{S(x+1)}.$$

Everett的限制條件

- 對Everett公式的要求一般分成以下五點：

→連續性

→一階可微分

→二階可微分

→還原性(接點與原始值相同)

→精確性(真實值為*n*次多項式)

Everett四點公式

- 要求接點一階可微分、一階精確性

$$v_{x+s} = F(1-s)u_x + F(s)u_{x+1}, \text{其中 } F(s) = A(s) + B(s)\delta^2$$

且滿足

$$A(s) = s, B(s) = \frac{1}{2}s^2(s-1)$$

→這個公式又稱為Karup-King公式，滿足二階的精確性。

Everett六點公式

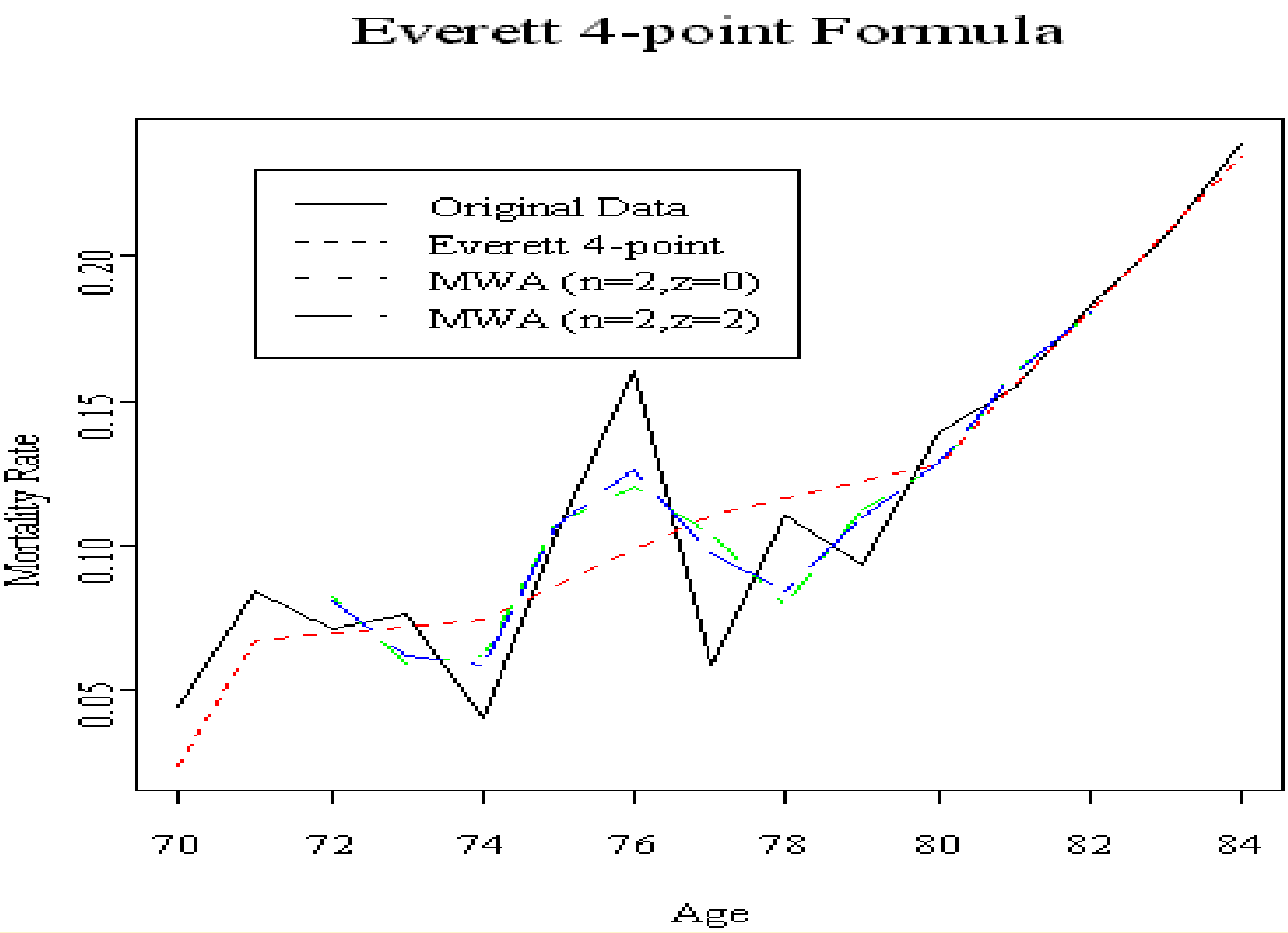
- 要求接點二階可微分、三階精確性

$$F(s) = A(s) + B(s)\delta^2 + C(s)\delta^4$$

且滿足

$$A(s) = s, B(s) = \frac{1}{2}s^2(s-1)$$

$$C(s) = \frac{1}{48}s^4$$



Everett 四點公式修勻範例

多項式Spline

- 多項式Spline或是平滑(Smoothness) Spline稱為片斷多項式函數(Piecewise Polynomial Function)，基本的想法是將原始值分成數段，再對每一段的數值作多項式修勻。

→若以一個多項式來表示死亡率特性，至少需要三次或三次以上的多項式。

→實證上Spline的近似效果不錯。

多項式Spline(續)

- Spline有兩種用途：

(1) 求取某個函數的近似值，通常計算內插(Interpolation)值

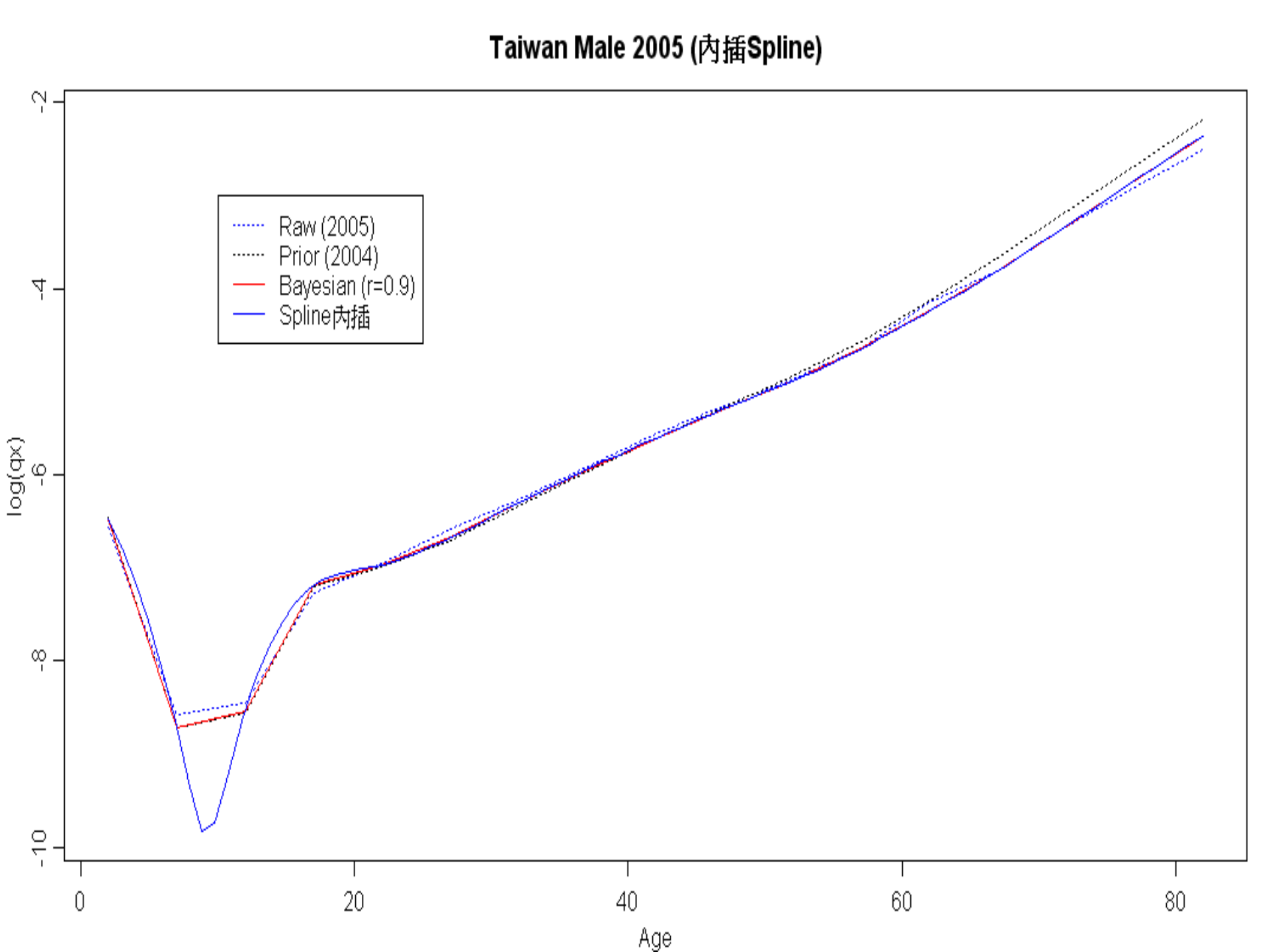
(2) 去除因為誤差而產生的震盪，使調整後的函數更為平滑

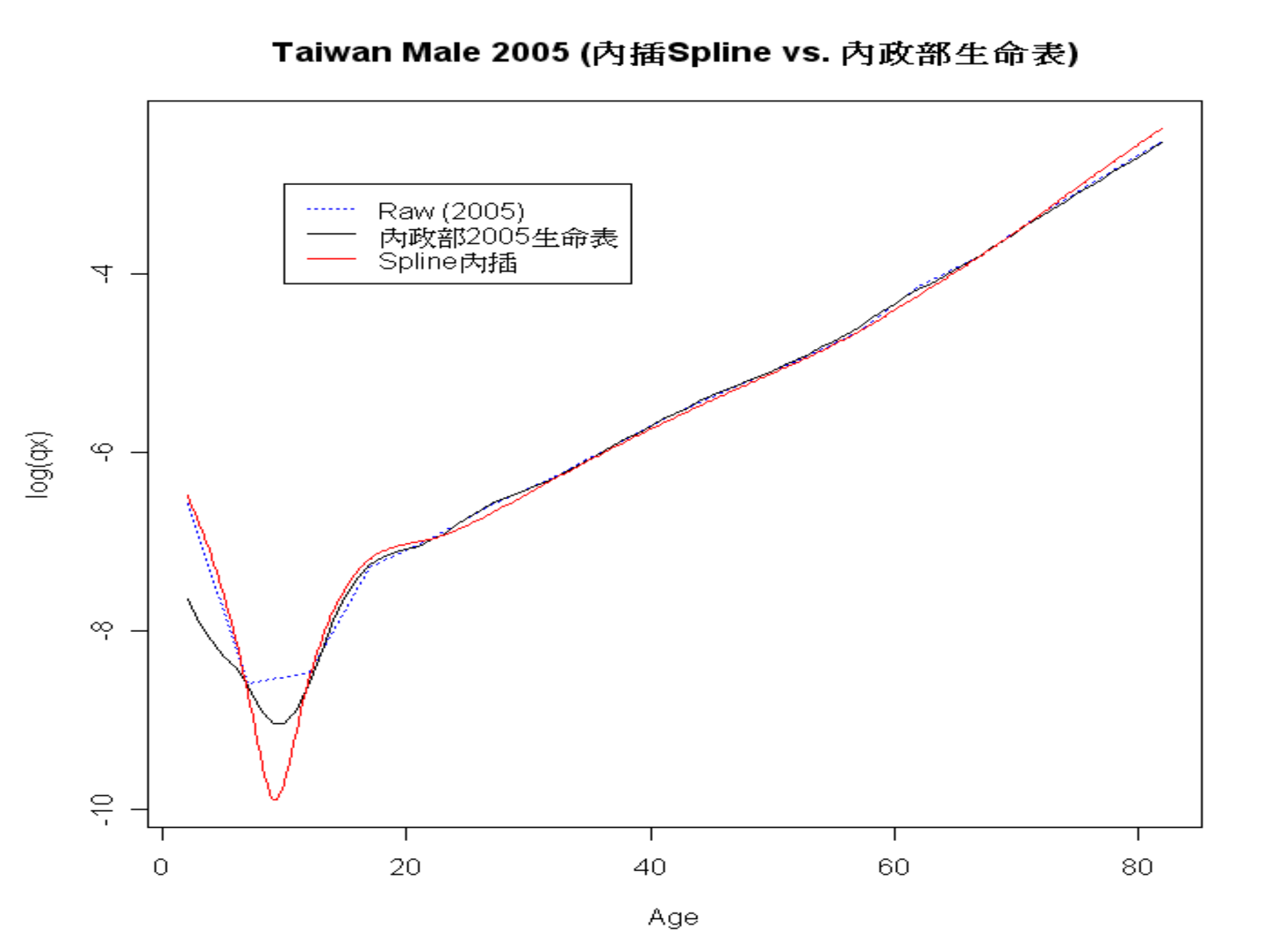
註：有時以內插Spline (Interpolating spline) 及平滑Spline (Smoothing spline)區隔。

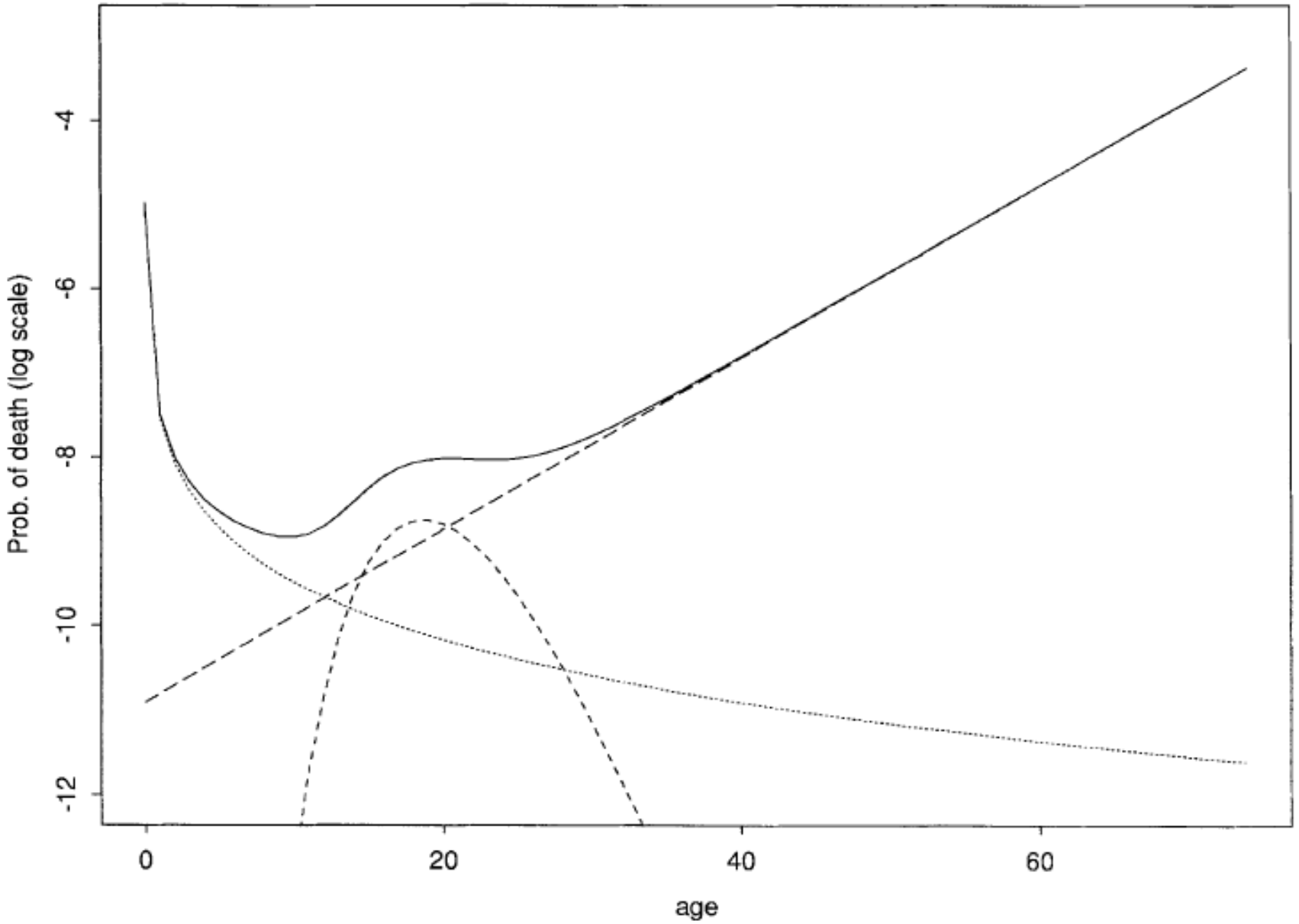
範例一、以2004年台灣地區單齡的簡易生命表為先驗資訊，原始死亡率則參照2005年的實際數值。（貝氏修勻）

→貝氏修勻之後，再以Spline內插求取單齡的死亡率，使用 R 軟體中的「spline」指令，可選擇內插的年齡，在本範例中為2至82歲的單齡死亡率。

→Spline內插後死亡率曲線較為平滑，在10歲的死亡率較低，其值約為萬分之一，與內政部公佈的生命表數值非常接近。


Taiwan Male 2005 (內插Spline)
Raw (2005)
Prior (2004)
Bayesian (r=0.9)
Spline內插
log(qx)
-2
-4
-6
-8
-10
0
20
40
60
80
Age


Taiwan Male 2005 (內插Spline vs. 內政部生命表)
Raw (2005)
內政部2005生命表
Spline內插
log(qx)
-4
-6
-8
-10
0
20
40
60
80
Age



Heligman-Pollard 模型中各年齡組的死亡曲線

(A,B,C,D,E,F,G,H) = (.000544,.0170,.101,.000158.10.72,18.67,.0000183,1.11)

多項式Spline(續)

- 分段式修勻也有幾項要求必須滿足:

→兩個時間分段的連接點必須要連續；

→相鄰的接點必須是可微分，以期滿足修勻中一段的平滑性要求；

→避免使用過於複雜的多項式，每個區段內所選取的多項式大多不大於四次，以便於計算及詮釋。

- 若將原始觀察值分成k+1個區間，其中$k_1, k_2, \ldots, k_k$為這k+1個區段內多項式的交接點；令每個區段內的多項式為三次，通常會假設：

$$\begin{cases} p_i(k_i) = p_{i+1}(k_i), \\ p_i'(k_i) = p_{i+1}'(k_i), \quad i = 1, 2, \ldots, k \\ p_i''(k_i) = p_{i+1}''(k_i), \end{cases}$$

其中p_i為第i個區段內的多項式，p_i'及p_i''為p_i的一次及二次微分。

- 引用數值分析的方法，讀者可找出滿足上述要求的多項式應具有以下的特性：

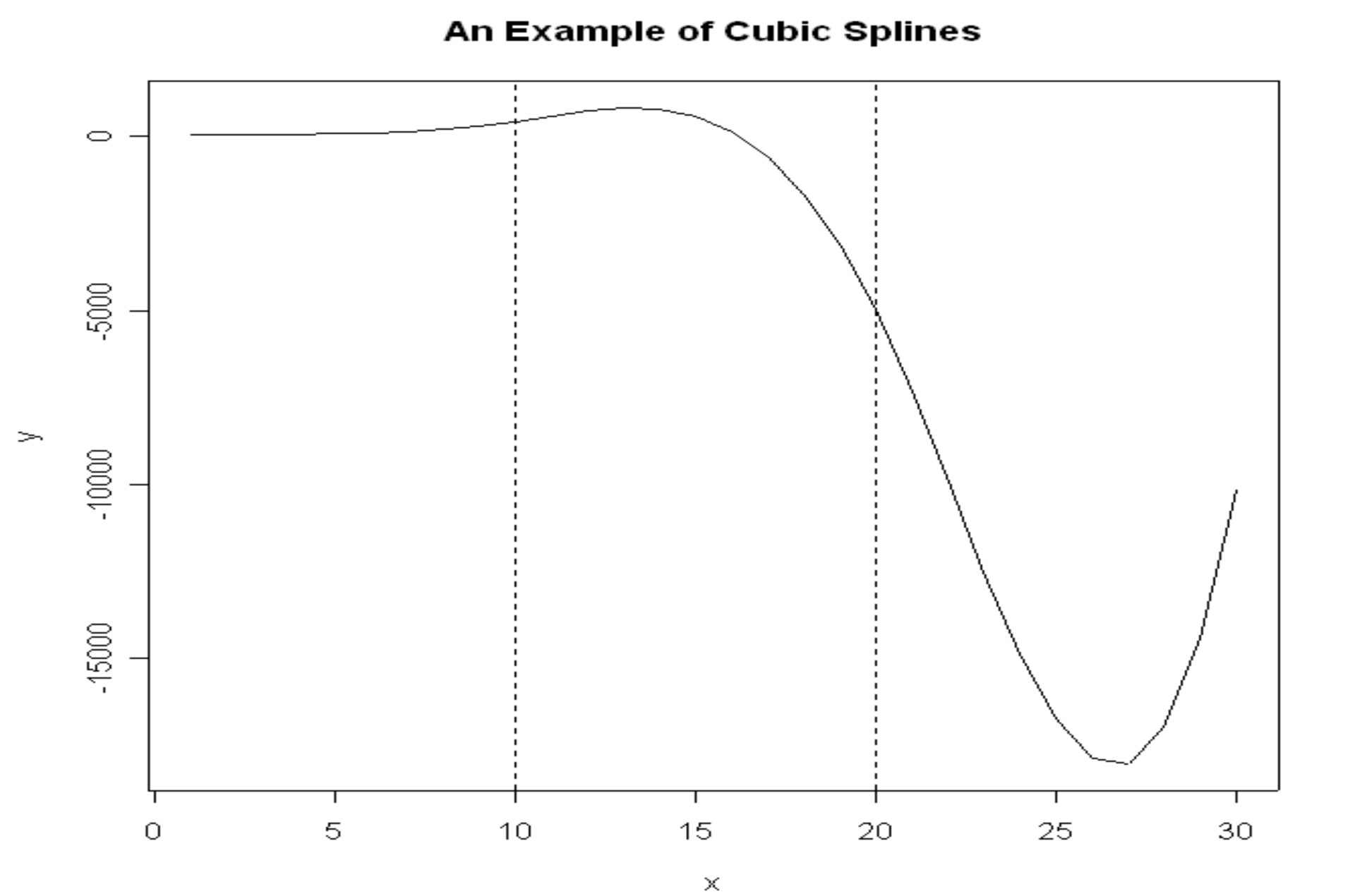
$$P(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + I(x \geq k_1)\beta_4 (x - k_1)^3 + I(x \geq k_2)\beta_5 (x - k_2)^3 + \cdots\cdots$$

→這個多項式可由一般的迴歸分析，即最小平方法或是加權最小平方法求出參數值 β's。

- 以矩陣表達，可表示為 $\underset{\sim}{t} = P = A\underset{\sim}{\phi}$，其中：

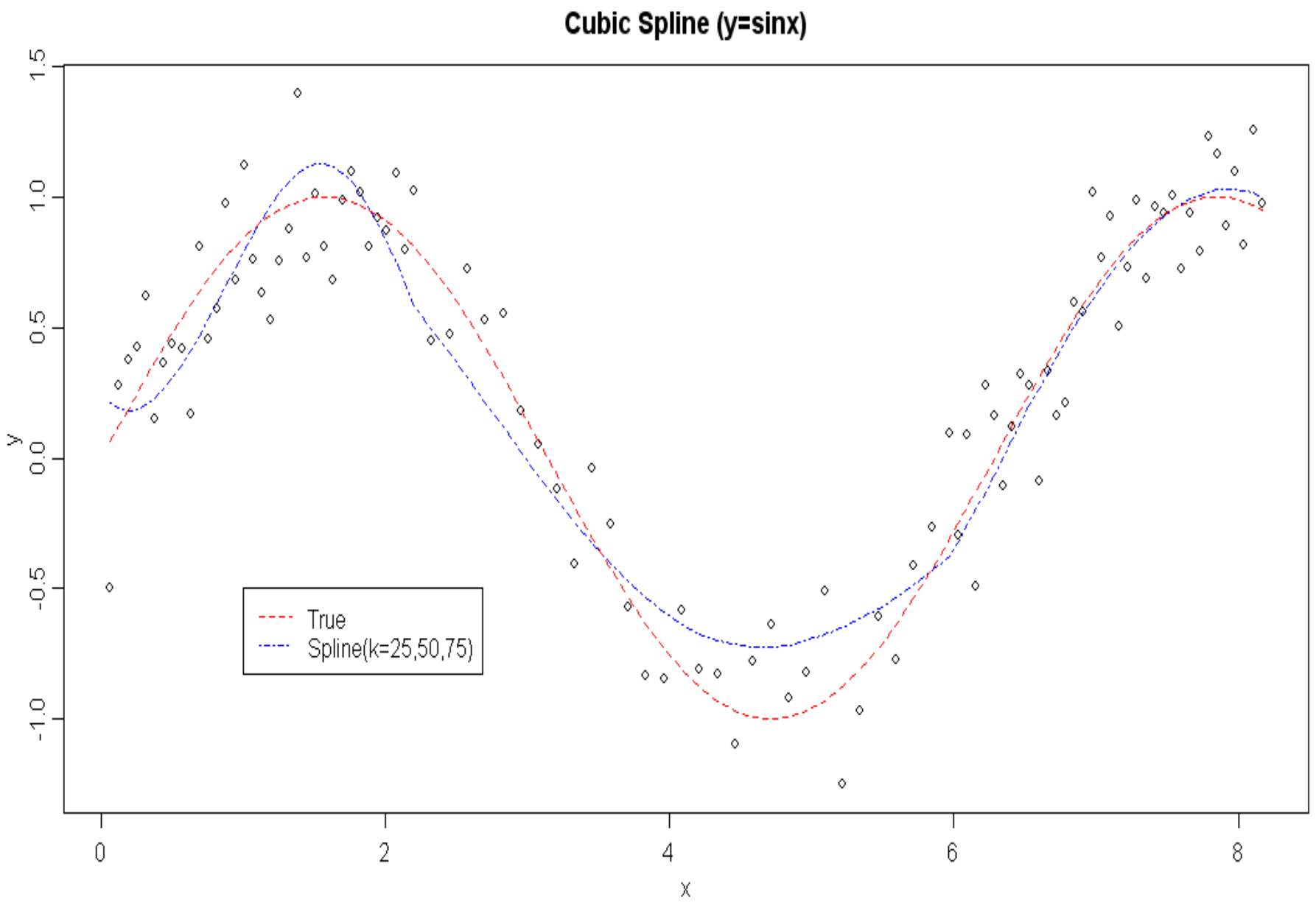
$$A = \begin{pmatrix} 1 & 1 & 1^2 & 1^3 & \left.\begin{matrix}0\\0\\\cdot\\0\end{matrix}\right\}k_1 & \\ 1 & 2 & 2^2 & 2^3 & & \\ 1 & 3 & 3^2 & 3^3 & & \\ \cdot & \cdot & \cdot & \cdot & 1^3 & \cdots\cdots \\ \cdot & \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \\ 1 & n & n^2 & n^3 & (n-k_1)^3 & \end{pmatrix}, \qquad \underset{\sim}{\phi} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \beta_{k+2} \end{pmatrix}$$

→修勻值可表為 $\underset{\sim}{v} = A(A'A)^{-1}A'\underset{\sim}{u}$ 。

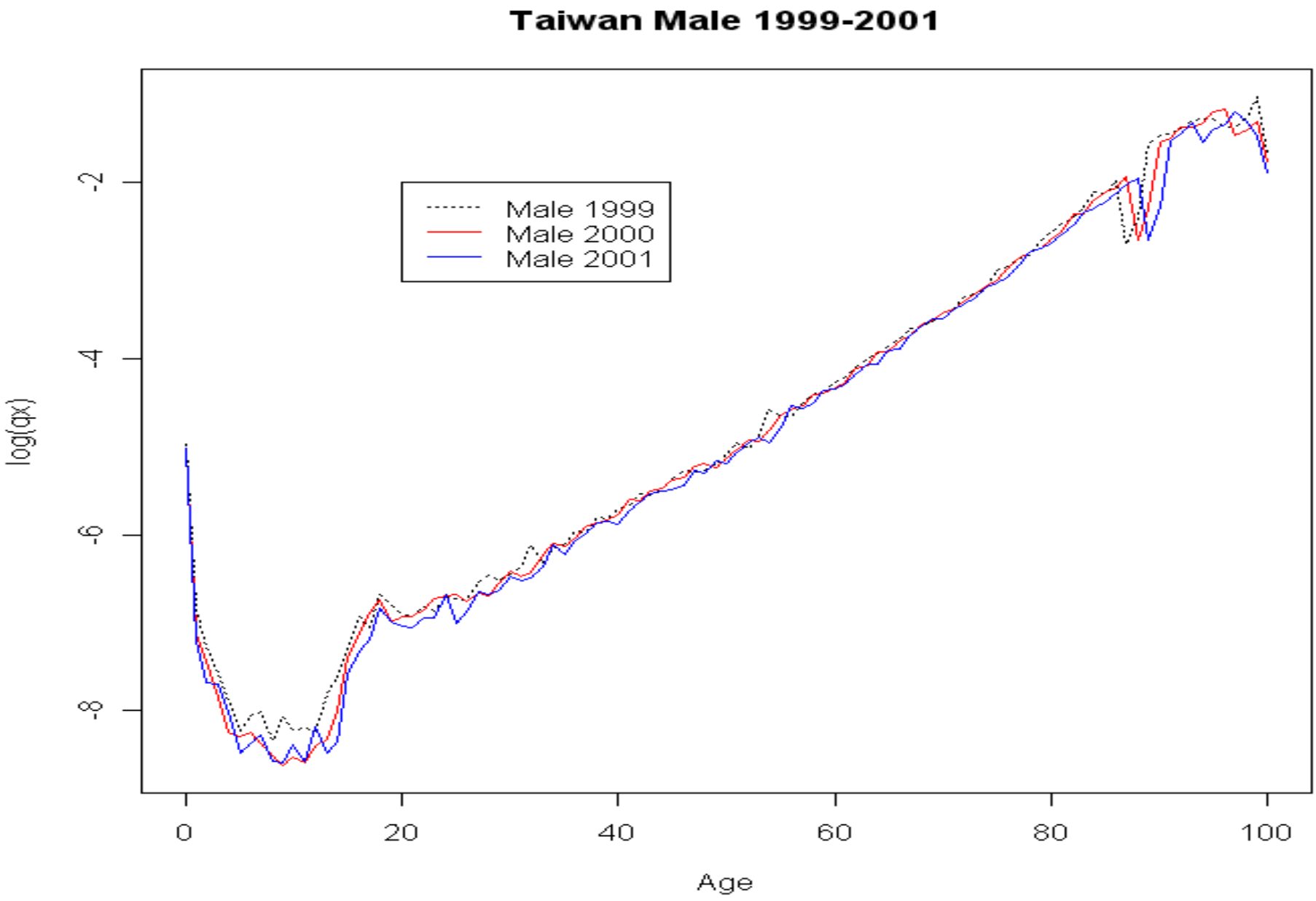


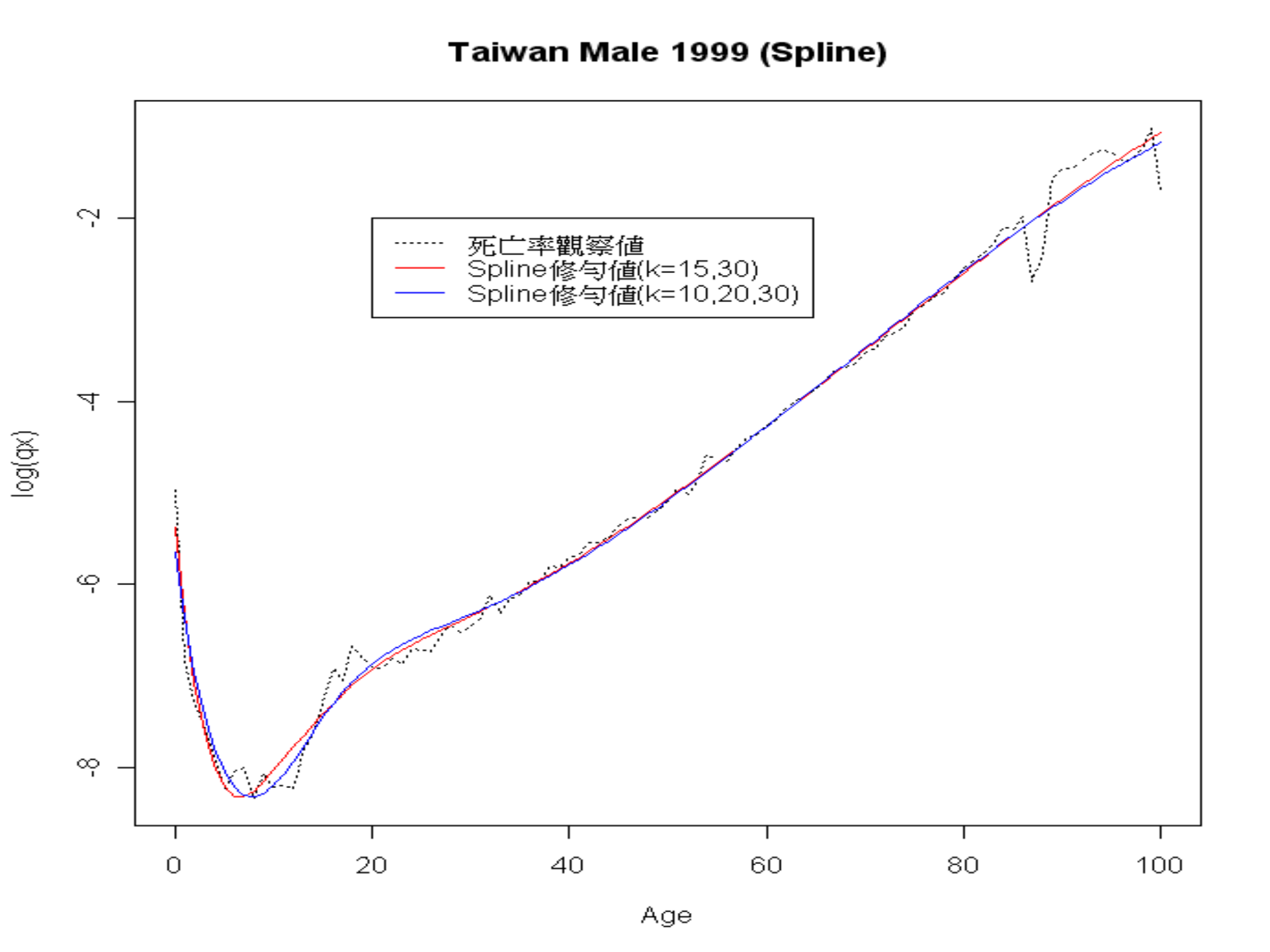
Q: 可能由一個多項式達到上述效果嗎？

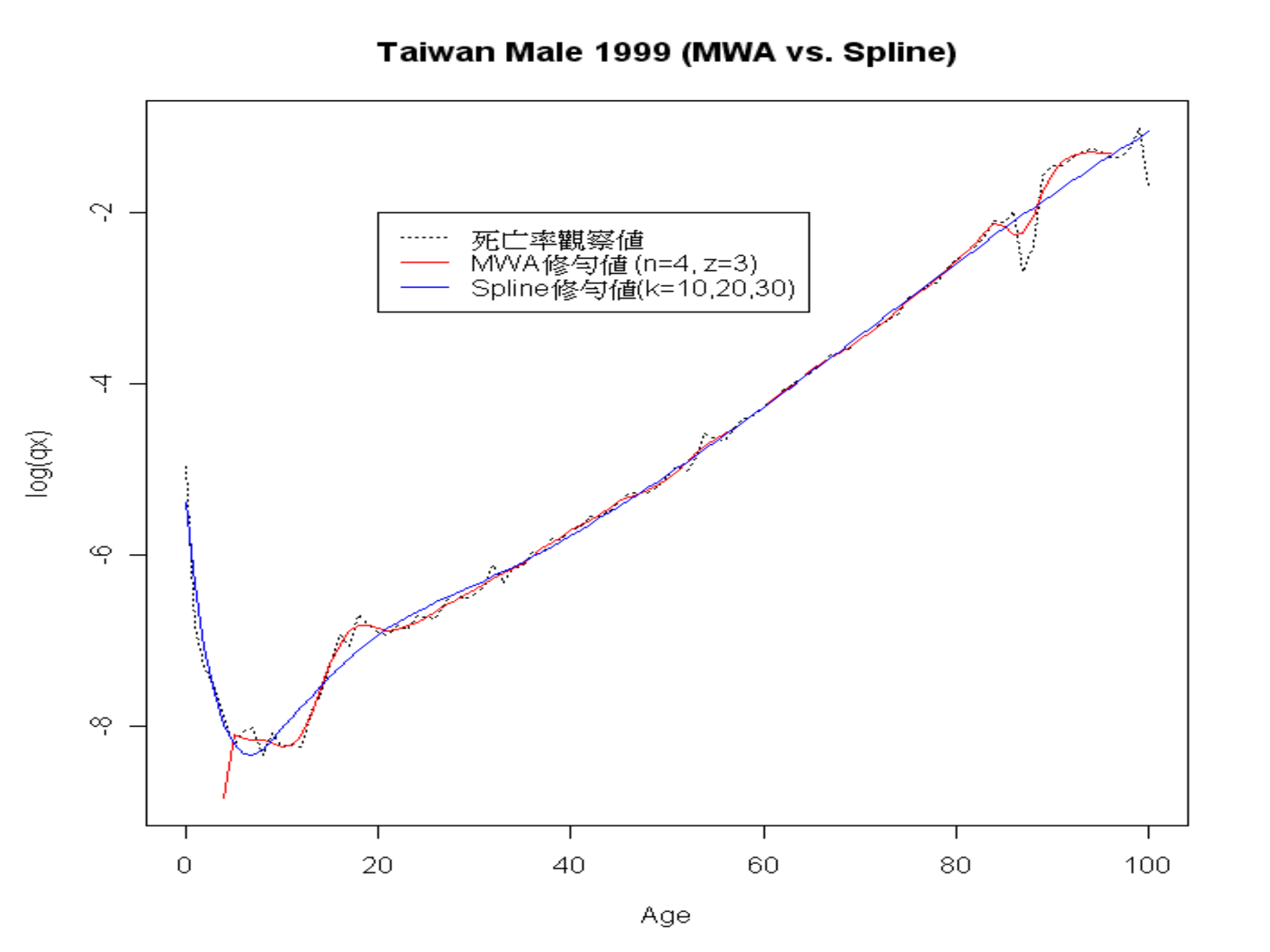
範例二：$Y = \sin(x) + \varepsilon,\ x \in [0, 2\pi]$

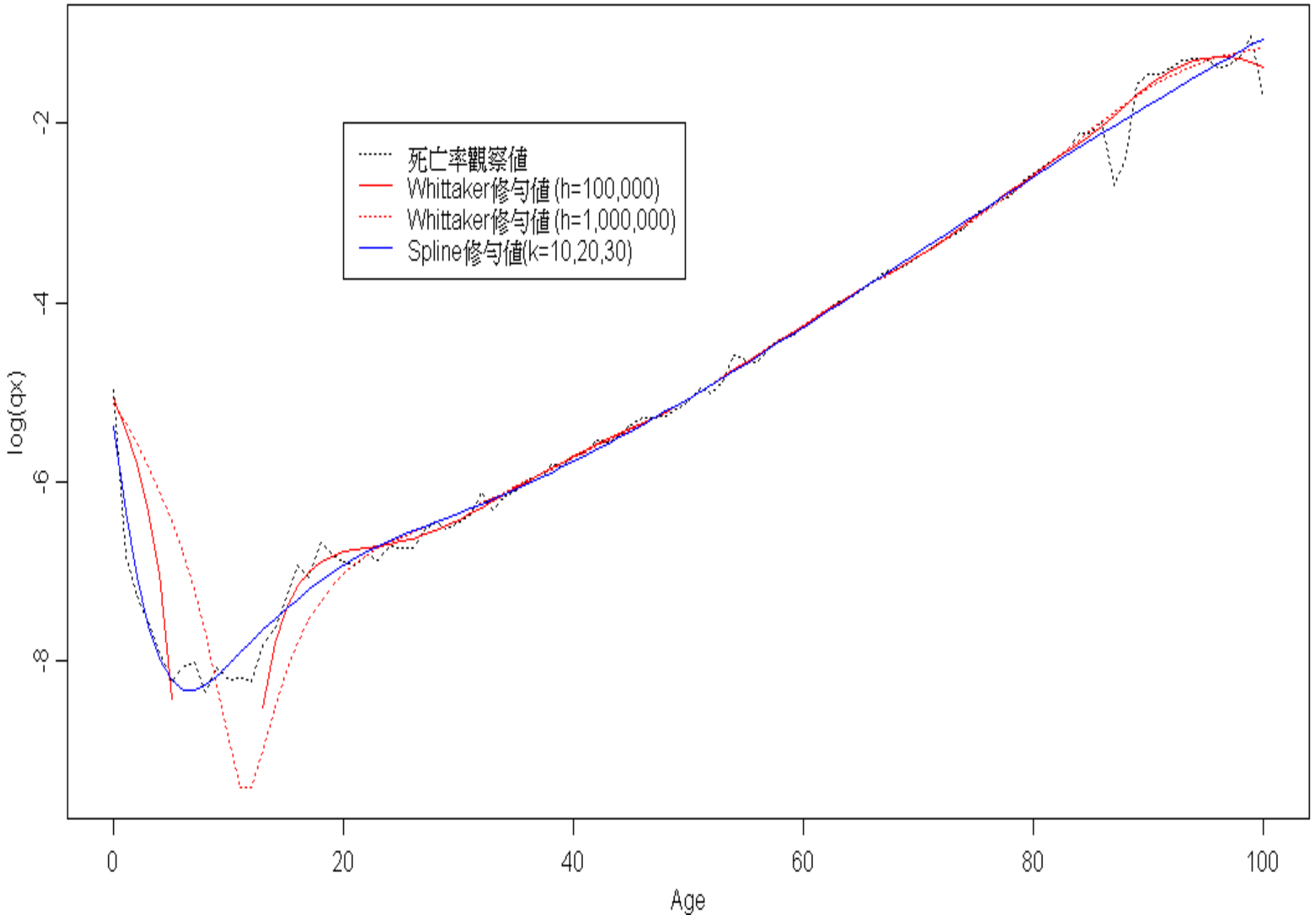


範例三：台灣1999-2001年男性生命表




Taiwan Male 1999 (Spline)
死亡率觀察值
Spline修勻值(k=15,30)
Spline修勻值(k=10,20,30)
log(qx)
-2
-4
-6
-8
0
20
40
60
80
100
Age


Taiwan Male 1999 (MWA vs. Spline)
死亡率觀察值
MWA修勻值 (n=4, z=3)
Spline修勻值(k=10,20,30)
log(qx)
-2
-4
-6
-8
0
20
40
60
80
100
Age


Taiwan Male 1999 (Whittaker vs. Spline)
死亡率觀察值
Whittaker修勻值(h=100,000)
Whittaker修勻值(h=1,000,000)
Spline修勻值(k=10,20,30)
log(qx)
-2
-4
-6
-8
0
20
40
60
80
100
Age

懲罰概似估計

- 懲罰概似估計(Penalized Likelihood Estimation)

→Whittaker修匀法是懲罰概似估計(PLE)的特例，將目標函數設為

$$L(\theta \mid data) + \frac{\lambda}{2} J(\theta)$$

其中L(θ|data)為對數概似函數取負號，J(θ)常為二次粗略的懲罰函數(Quadratic Roughness Penalty)。

懲罰概似估計(續)

- 懲罰函數J(θ)一般選為

$$J(\theta)=\int \ddot{\theta}(x)^2\,dx \quad \text{或} \quad J(\theta)=\sum_x (\Delta^z\theta(x))^2$$

→若觀察值滿足 $Y_i=\theta(x_i)+\varepsilon_i, \varepsilon_i \sim N(0,\sigma_i^2)$，上述範例一可視為這種模型的特例，則PLE修勻即是將下列目標函數最小化：

$$\sum_{i=1}^{n}\left(\frac{Y_i-\theta(x_i)}{\sigma_i}\right)^2+\frac{\lambda}{2}\int \ddot{\theta}(x)^2\,dx$$

Note: Two terms of the right-hand side of the objective function usually represent constraints opposite to each other.

→ The first term measures how far the smoothers differ from the original observations.

→ The second term, also known as *roughness penalty*, measures the smoothness of the smoothers.

Note: Methods which minimize the objective function are called *penalized LS methods*.

範例四、1988~2002年England 及Welsh的女性死亡資料。

→使用 R 軟體的套裝程式「gss」模組，操作手冊可從www.r-project.org下載，指令非常簡單，程式如下，圖形在下一頁。

```
t<-sqrt((0:74));
pois.fit <- gssanova((d/e)~t,family="poisson",weights=e);
est <- predict(pois.fit,data.frame(t=t),se=TRUE);
plot((0:74),log(d/e),type="l",xlab="Age", ylab="Log Mortality");
lines((0:74),(est$fit),col=2);
lines((0:74),(est$fit+1.96*est$se),col=3);
lines((0:74),(est$fit-1.96*est$se),col=4);
```

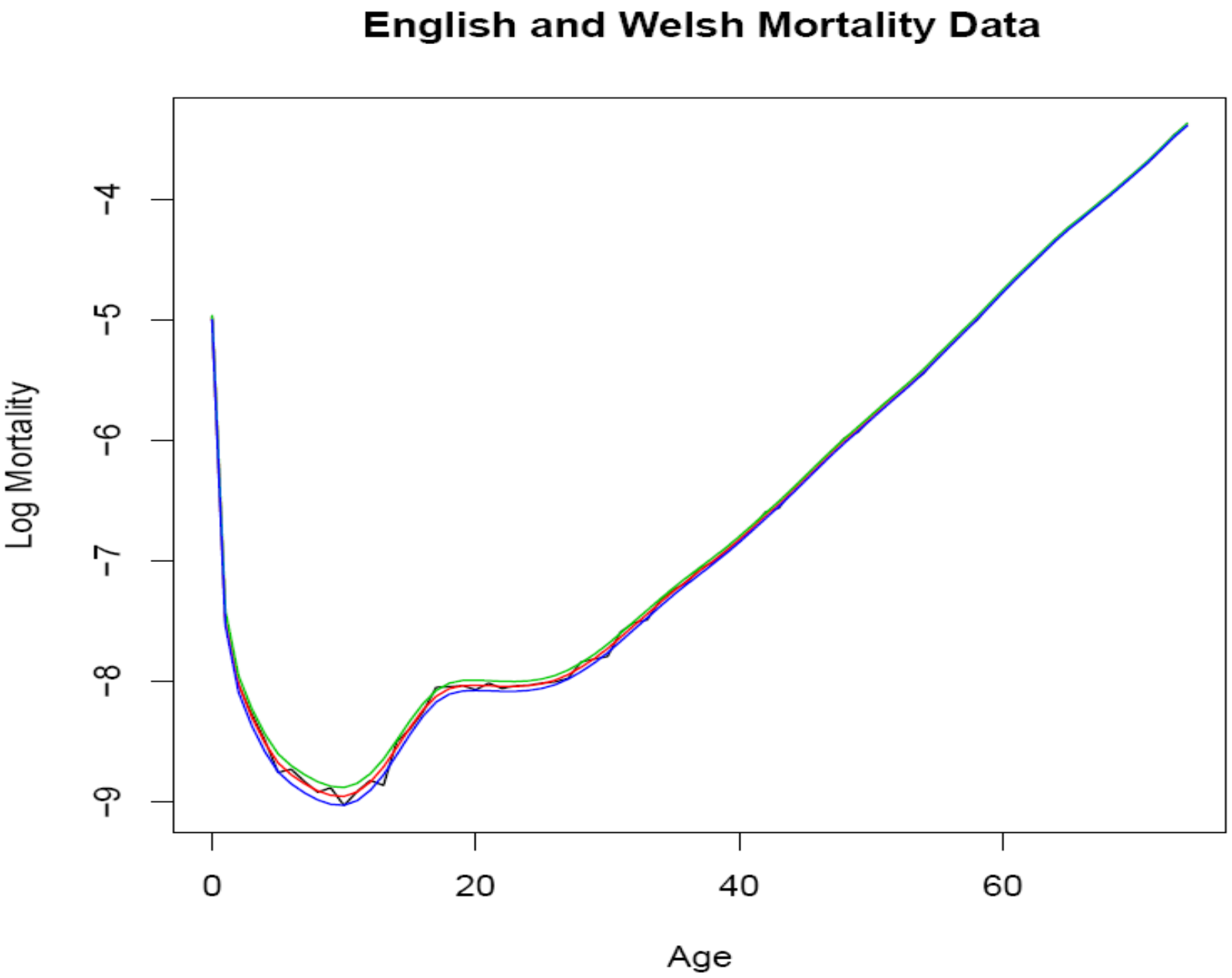


Figure 1 Raw Data (Black), Upper 95%, Lower 95% and Bayes Estimate.

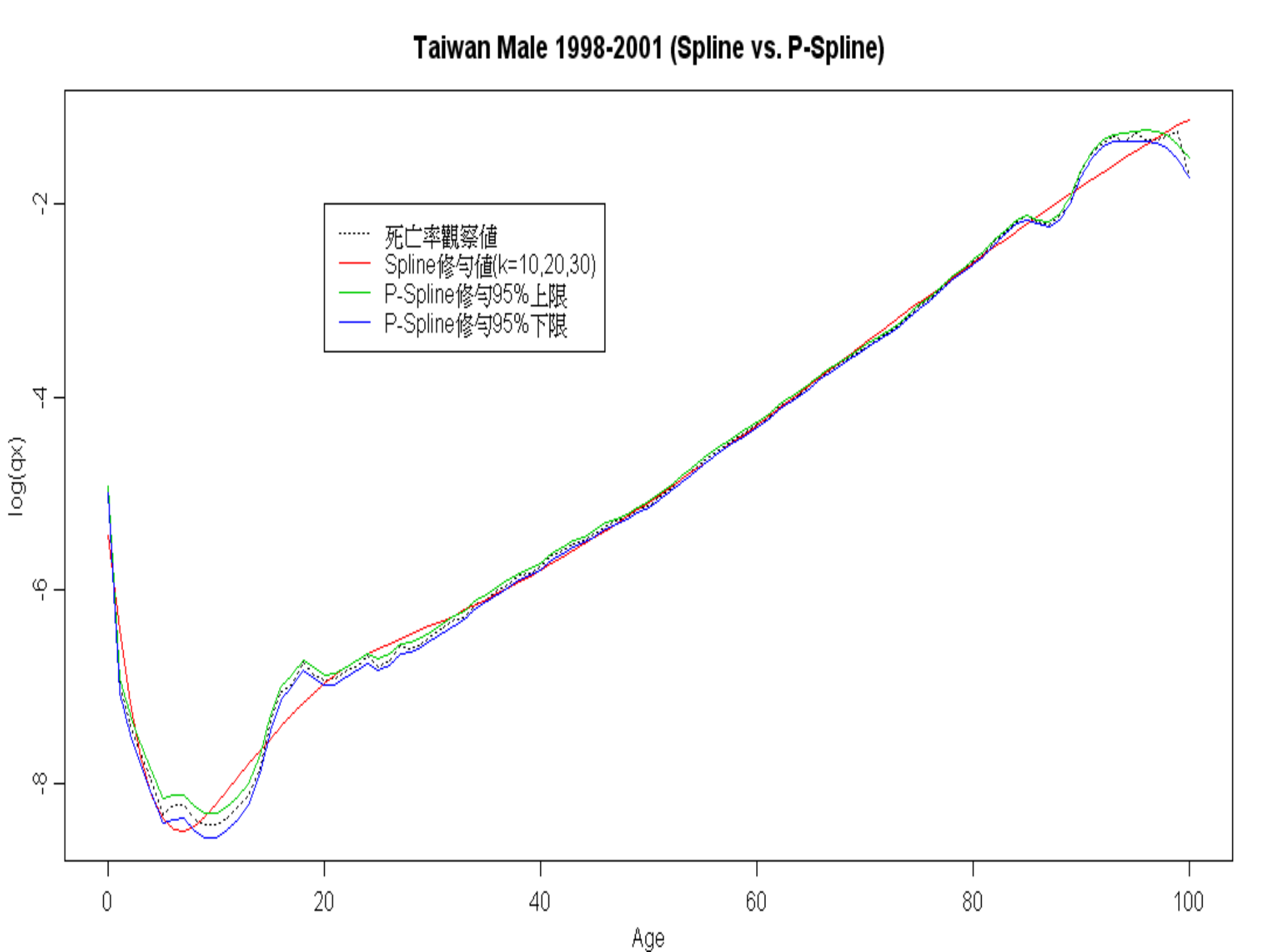
範例五：台灣1998-2001年男性生命表

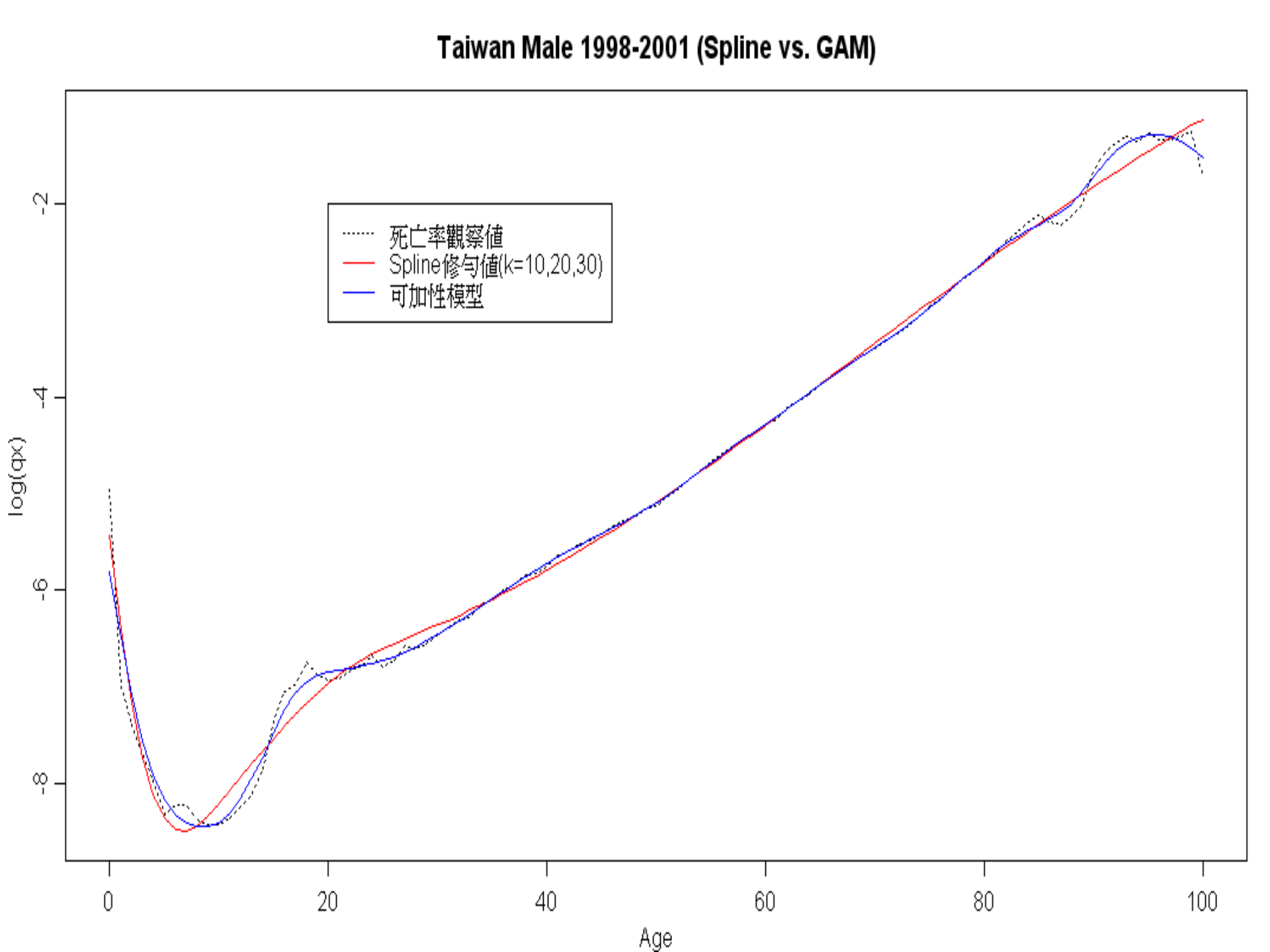
- 仿造範例三，考慮臺灣地區1998-2001年男性死亡率的修匀，首先將四個年度的死亡人數、總人數分別加總。

→多項式Spline有三個分段點；

→P-Spline可提供死亡率與年齡的迴歸方程式，提供 95%信賴區間的上、下限，也可提供迴歸方程式的估計值。

→也可使用可加性模型(Generalized Additive Model)，在 R中使用「gam」模組。


Taiwan Male 1998-2001 (Spline vs. P-Spline)
死亡率觀察值
Spline修匀值(k=10,20,30)
P-Spline修匀95%上限
P-Spline修匀95%下限
log(qx)
-2
-4
-6
-8
0
20
40
60
80
100
Age


Taiwan Male 1998-2001 (Spline vs. GAM)
死亡率觀察值
Spline修勻值(k=10,20,30)
可加性模型
log(qx)
-2
-4
-6
-8
0
20
40
60
80
100
Age

- Linear Smoother:

→ The goal is the smooth estimates $\hat{M}$ of a regression function $M(x) = E(Y \mid X = x)$. A well-known example is the ordinary linear regression, where the fitted values are

$$\hat{y} = Hy, \ \textit{where} \ \ H = X(X'X)^{-1}X'.$$

→ *A Linear Smoother* is the one which the smooth estimate satisfies the following form:

$$\hat{y} = S\,y,$$

where S is an $n \times n$ matrix depending on X.

- Running Means:

→ The simplest case is the running-mean smoother which computes $\hat{y}_i$ by averaging y_j's for which x_j falls in a neighborhood of x_i.

→ One possible choice of the neighborhood N_i is to adapt the idea in *Nearest-neighbor* where N_i is the one with points x_j for which $|i-j| \leq k$. Such a neighborhood contains k points to the left and k points to the right. (Note: The two tails have fewer points and could be less smooth.) s

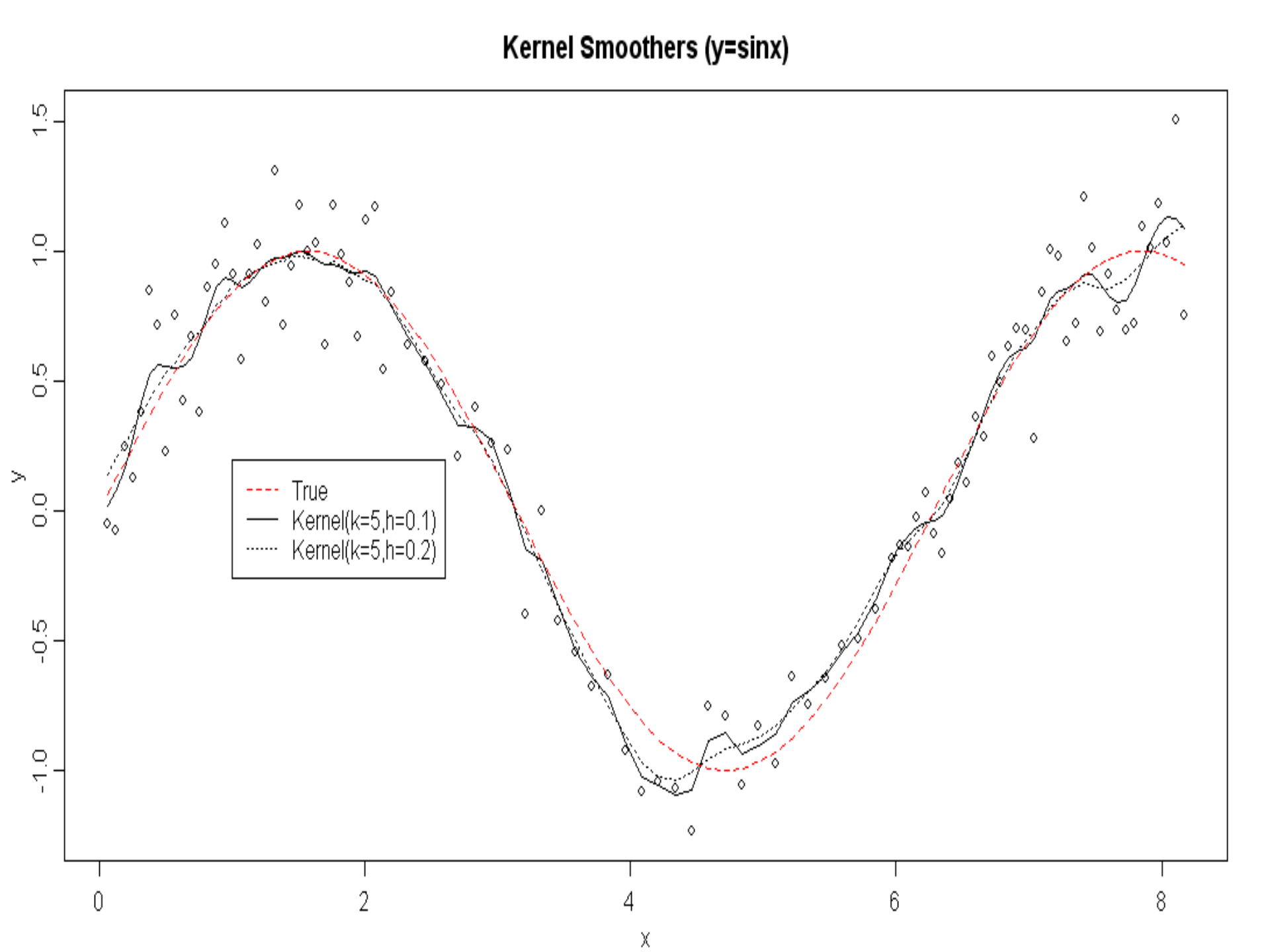
- Example. Suppose that

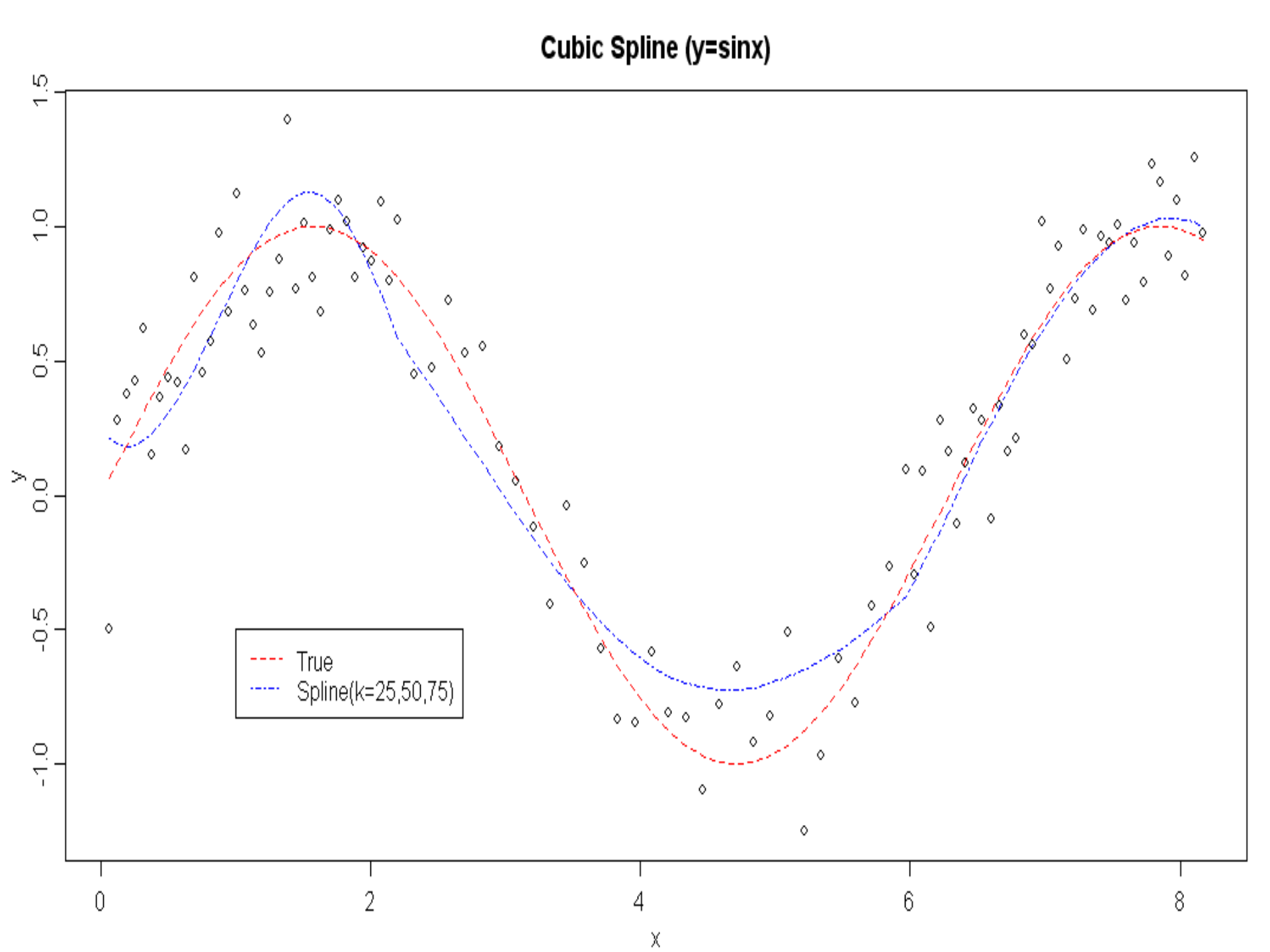
$$Y_i = \sin X_i + \varepsilon_i, \quad 0 \le X_i \le \pi,$$

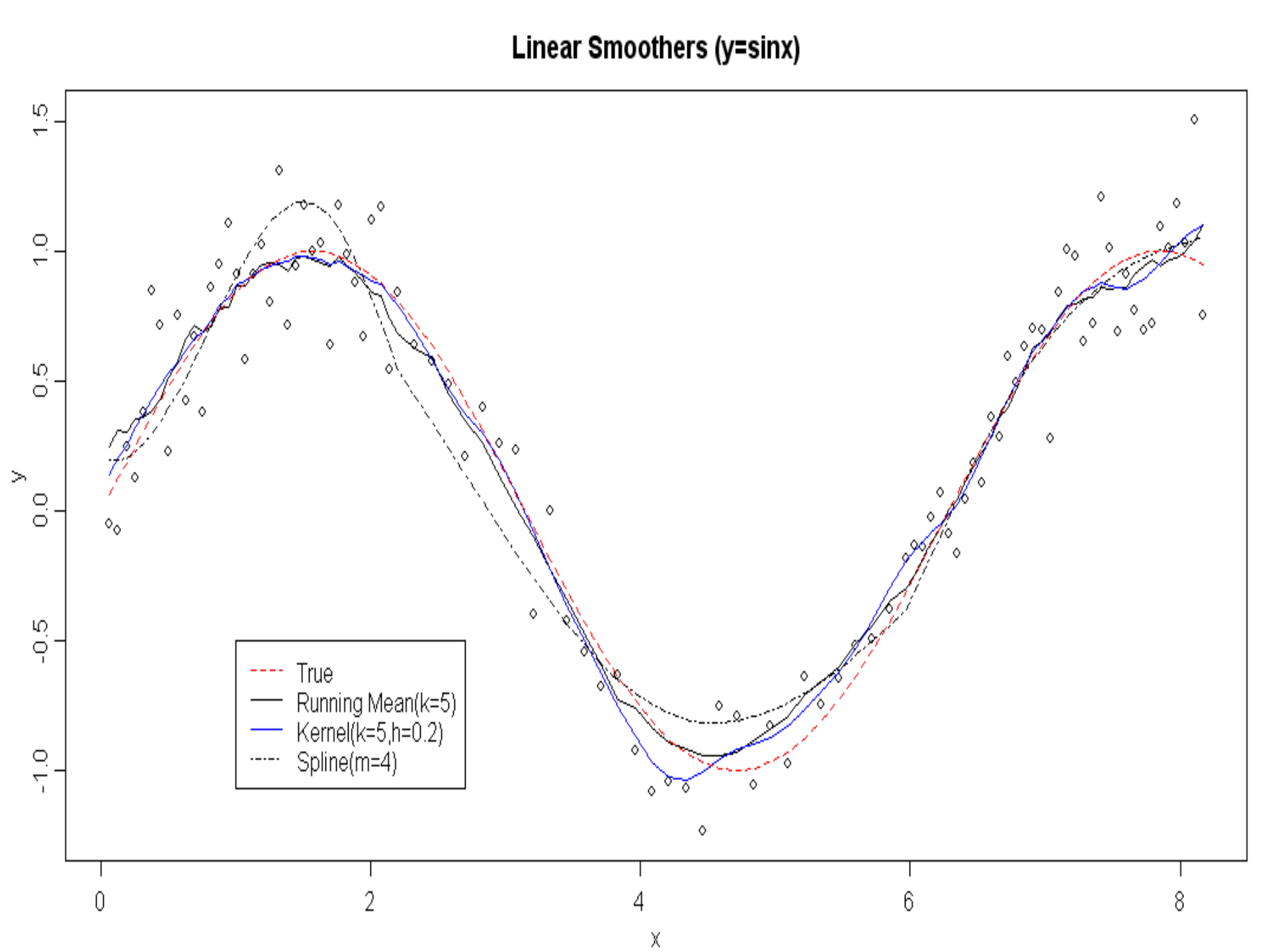
where the noise ε_i is normally distributed with mean 0 and variance 0.04. Also, the setting of X is 15 points on [0,0.3π], 10 points on [0.3π,0.7π] and 15 points on [0.7π,π].

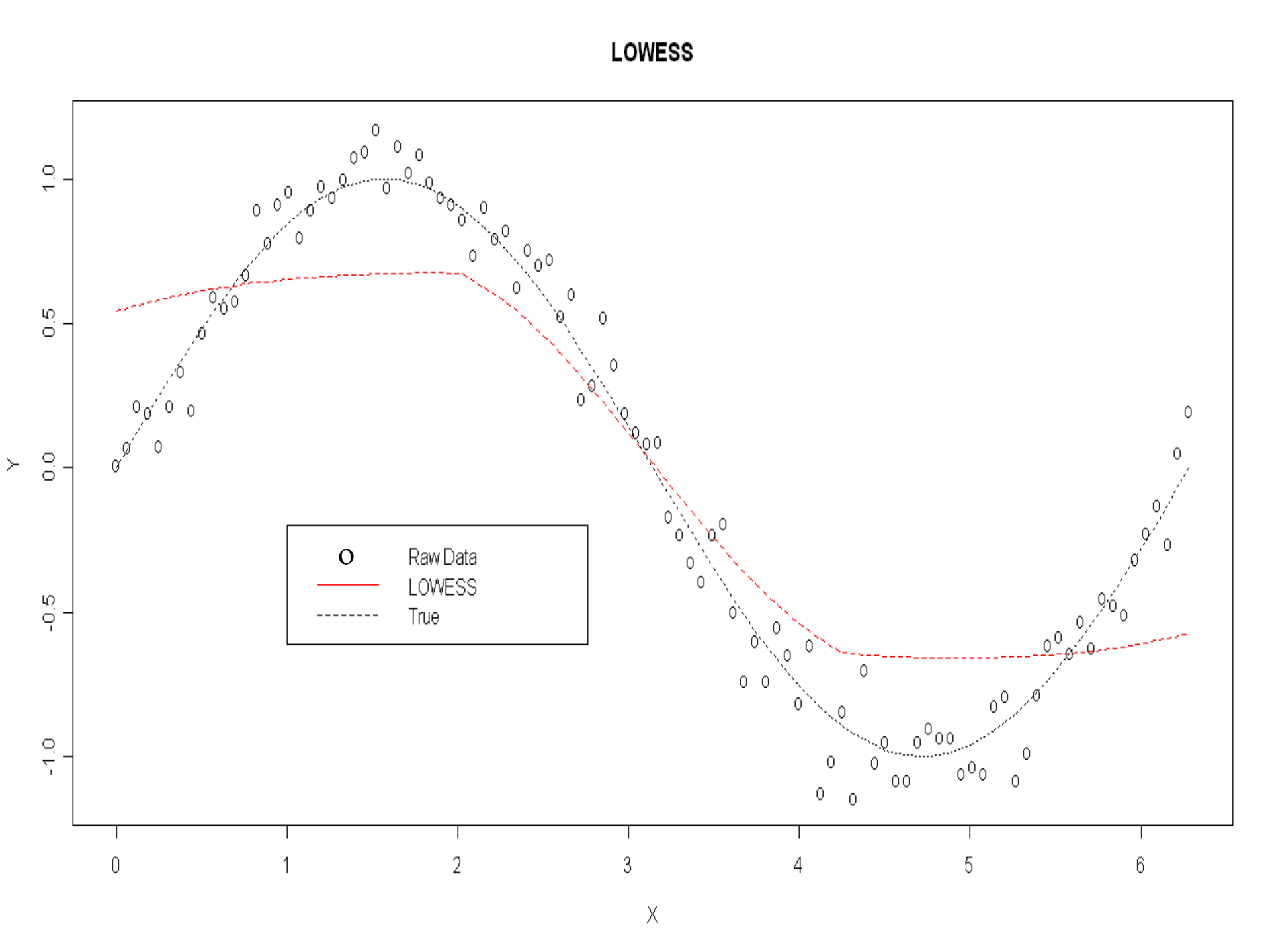
→Cubic splines with knots at {0,2π/3,4π/3,2π}.

Note: There are also other smoothing methods available, such as LOWESS (LOESS for an updated version) and running median (i.e., nonlinear smoothers), but we won't cover these topics in this class.

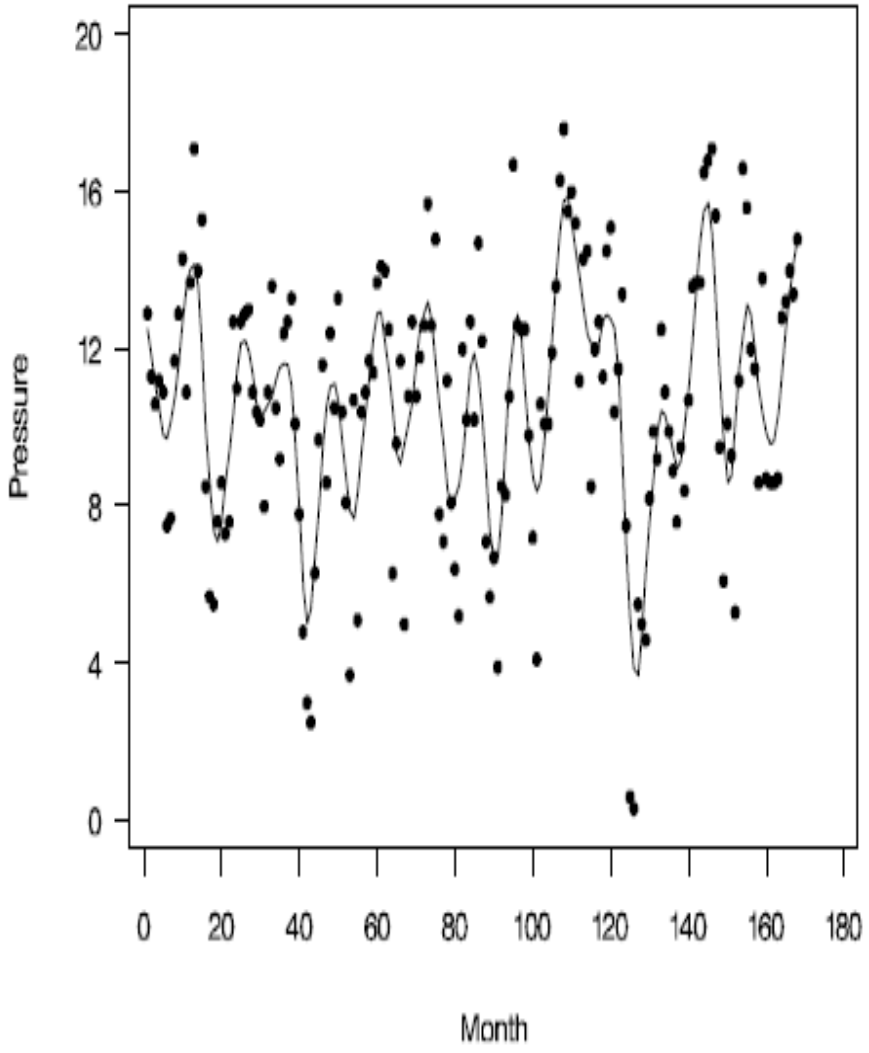

Kernel Smoothers (y=sinx)
True
Kernel(k=5,h=0.1)
Kernel(k=5,h=0.2)
y
x
1.5
1.0
0.5
0.0
-0.5
-1.0
0
2
4
6
8

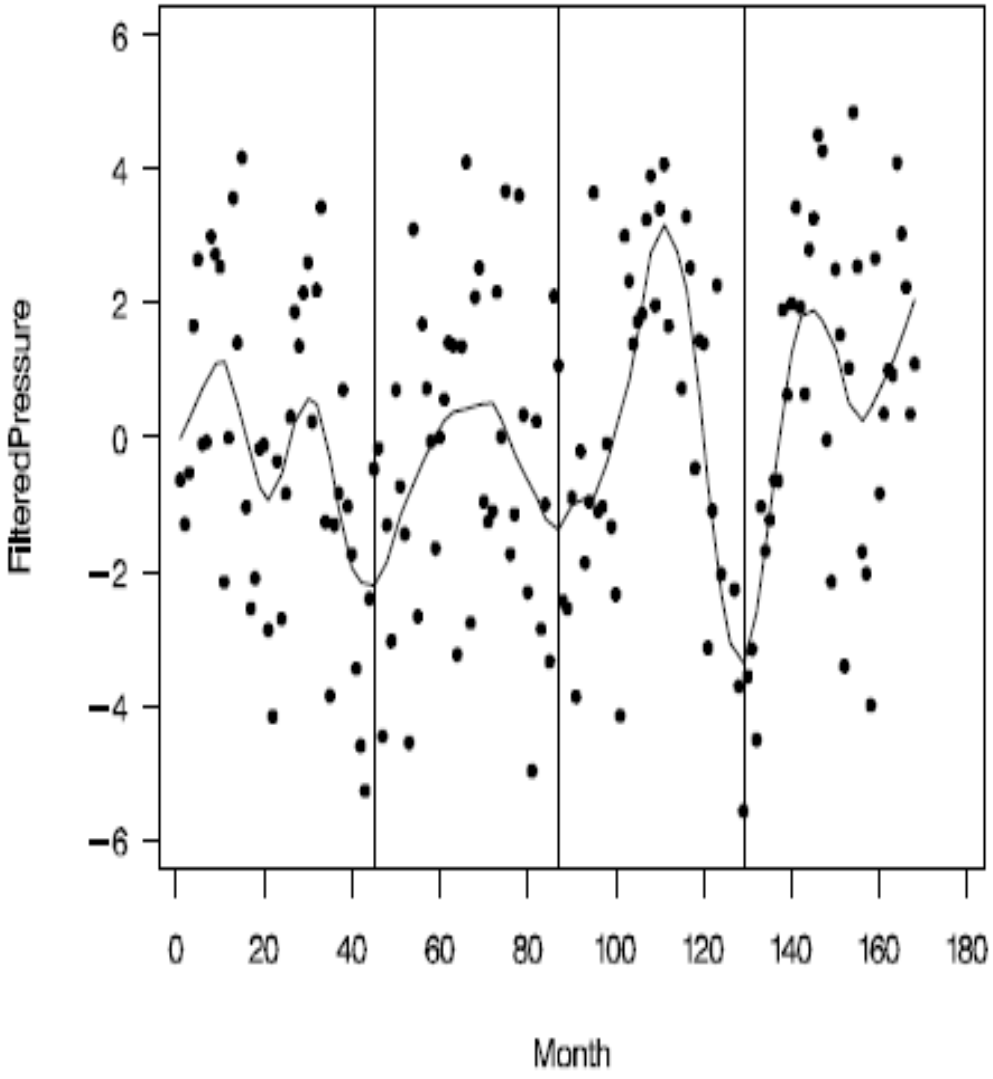

Cubic Spline (y=sinx)
1.5
1.0
0.5
0.0
-0.5
-1.0
y
0
2
4
6
8
x
True
Spline(k=25,50,75)


Linear Smoothers (y=sinx)
1.5
1.0
0.5
0.0
-0.5
-1.0
y
0
2
4
6
8
x
True
Running Mean(k=5)
Kernel(k=5,h=0.2)
Spline(m=4)

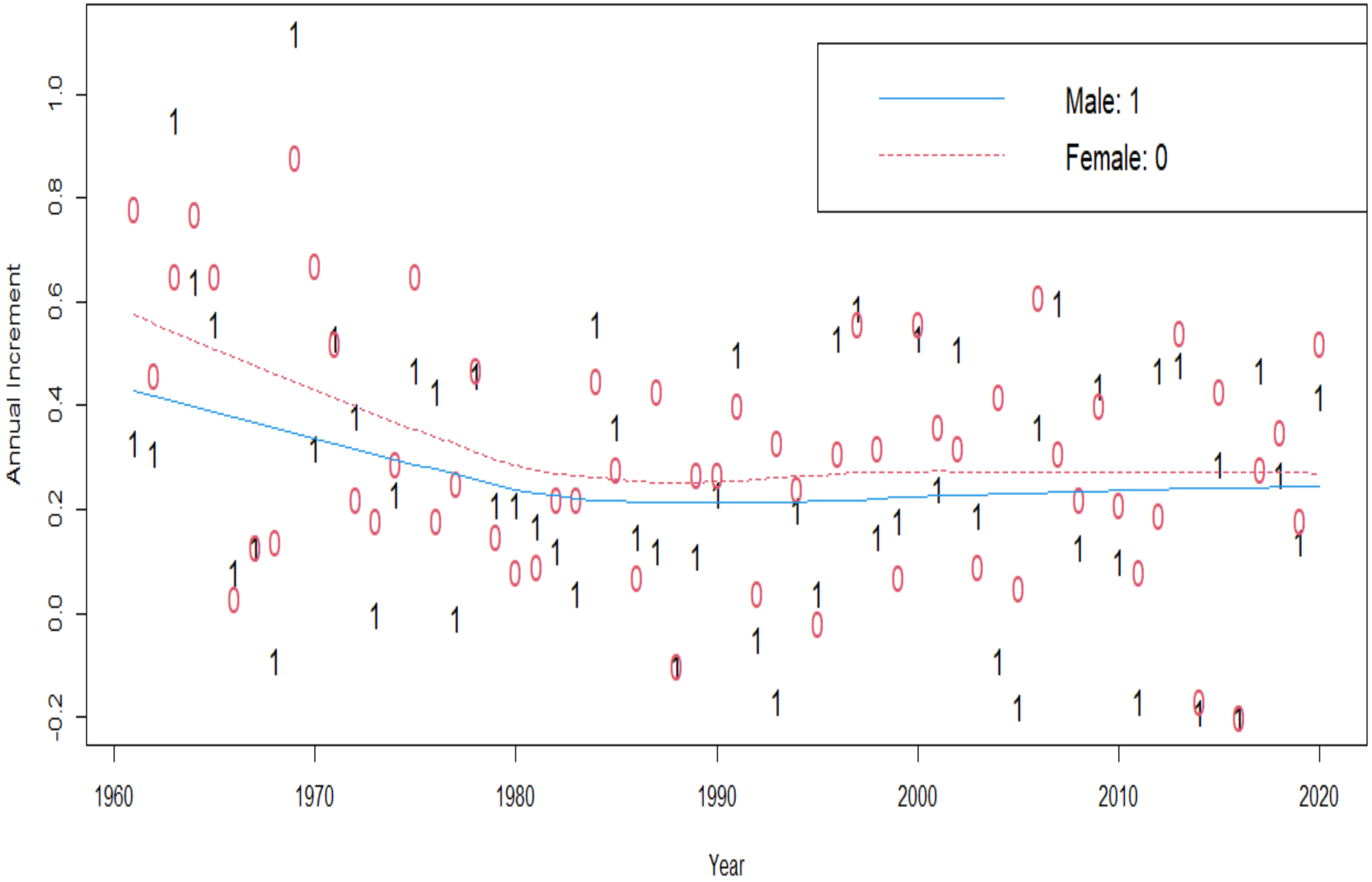

LOWESS
Raw Data
LOWESS
True
Y
X
1.0
0.5
0.0
-0.5
-1.0
0
1
2
3
4
5
6

LOWESS (Locally-weighted Polynomial Regression) or *LOESS* (Local Polynomial Regression Fitting)





國人壽命穩定增長 (簡易生命表、LOWESS)



Spline的使用建議

- Spline可以用於內插，也可用於修勻，與Whittaker類似，可以同時考量適度性與平滑性，而且在樣本數較少時不像Whittaker在兒童死亡率產生負值，實證上是不錯的選擇。

→但參數、分段點無法事先決定，需要以試誤法(Trial and Error)反覆嘗試。

→也可試試「可加性模型」等方法。