응용통계연구 제 **11**권 *1*호, *1998*년, *pp. 193–204*

베지에 곡선을 이용한 함수의 미분에 대한 비모수적 추정 *

김충락[†] 정미선[‡] 김형순[§]

요 약

주어진 자료를 회귀모형에 적합시켜 적합된 함수의 미분을 구해야 하는 경우가 흔히 있다. 본 논문에서는 베지에 곡선을 이용하여 비모수적으로 추정하는 방법을 소개하고, 실제 자료에 적용시킨다. 이 방법의 장점은 원하는 차수의 미분이 가능할 뿐만 아니라, 비모수 추정에 따르는 커널의 선택과정이 필요없고 단지 평활모수만 선택하면 된다.

1. 서론

회귀모형

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \cdots, n, \quad x \in [a, b] \tag{1.1}$$

이 주어져 있을 때 f의 적합치를 f_λ라고 하자. 단, λ는 평활모수(smoothing parameter)이고 λ를 추정하는 방법으로 교차확인(cross-validation) 또는 일반화 교차확인(generalized cross-validation)등이 있다. 또한, f_λ를 구하기 위한 비모수적 방법으로 급수추정(series estimation), 커널추정(kernel estimation), 평활 스플라인(smoothing spline)등이 있다(Eubank 1988). 그런데 f의 추정치 f_λ보다 때로는 f의 미분인 f'의 추정치가 더 필요하다. 예를 들어, 어린이들의 성장치 그 자체도 중요하지만 남녀별 성장속도도 매우 중요하다. 생물학적으로 키는 사춘기 근처에서 많이 커지는데 여자의 사춘기가 12세 전후로 남자의 사춘기인 14세 전후보다 빠르기 때문이다. 이외에도 통화량의 증가율, 인구의 증가율등이 좋은 예라 할 수 있다. f'_λ를 구하기 위한 비모수적 방법의 하나로서 Gasser and Müller(1979)가 제안한 커널추정치는

$$f'_\lambda(x) = \lambda^{-2} \sum_{j=1}^{n} y_j \int_{s_{j-1}}^{s_j} K(\lambda^{-1}(x-s))ds \tag{1.2}$$

으로 주어지고 $S_j = (x_j + x_{j+1})/2$로 가정하는 경우가 많으며 K는 커널로서 흔히 0에 대해 대칭인 밀도함수가 자주 사용된다. 그러나, f'_λ는 커널 K뿐만 아니라 λ에 선택에 크게 좌우

*이 논문은 1996년도 부산대학교 기성회 학술연구조성비 지원에 의하여 부산대학교 기초과학연구소에서 연구수행되었음 (RIBS-PNU-96-104)

† (609-735) 부산시 금정구 장전동 산 30, 부산대학교 통계학과 부교수
‡ (609-735) 부산시 금정구 장전동 산 30, 부산대학교 통계학과
§ (609-735) 부산시 금정구 장전동 산 30, 부산대학교 통계학과 박사과정

되며, 그 정도가 f_λ보다 커지고, S_j의 선택도 중요한 문제로 인식되고 있다(Müller(1984), Gasser, Müller and Mammitzsch(1985)).

본 연구에서는 베지에 곡선(Bezier curve)을 이용하여 회귀함수의 비모수적 추정과 회귀함수의 1차 미분에 대한 추정을 소개하고자 한다. 베지에 곡선의 정의와 몇가지 성질이 2절에서 소개되고, 3절에서는 베지에 곡선을 이용한 회귀함수 및 그 미분의 추정법이 제안되고, 4절에서는 실제자료(성장자료, Eubank 1989)를 이용하여 회귀함수 및 그 미분의 추정치를 예로 제시한다.

2. 베지에 곡선

베지에 곡선은 1959년 de Casteljau에 의해 처음으로 고안되었으며 Boehm (1977)과 Bezier (1977)에 의해 체계화되었다. 베지에 곡선은 그래픽 분야에서 오래전부터 사용되어 왔고 특히 CAD(Computer Aided Design) 분야에서는 매우 중요한 도구로 사용되고 있다. 그러나 통계학 분야에서는 용어자체가 생소하고, 베지에 곡선을 이용한 통계적 방법론은 전무하여 참고문헌도 전혀 없는 실정이다. 저자가 아는 한 베지에 곡선을 밀도함수의 추정에 이용한 논문(Kim, 1996)이 처음인 것 같다.

평면상에 $N+1$개의 점 $b_0 = (z_0, w_0)', b_1 = (z_1, w_1)', \cdots, b_N = (z_n, w_N)'$이 주어져 있을 때 $b_0, b_1, \cdots, b_N$에 근거한 베지에 곡선의 좌표는 t를 매개변수로 하여

$$b(t) \equiv \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \sum_{j=0}^{N} b_j B_{N,j}(t), \quad t \in [0,1] \tag{2.1}$$

로 정의되며

$$B_{N,j}(t) = \begin{pmatrix} N \\ j \end{pmatrix} t^j (1-t)^{N-j}, \quad j = 0, 1, \cdots, N$$

으로 Bernstein 다항식이라 한다. 또한, $b_0, b_1, \cdots, b_N$을 베지에 점(Bezier points)이라 부른다.

예를 들어, b_0, b_1, b_2, b_3에 근거한 베지에 곡선의 좌표는

$$\begin{aligned}(1-t)[(1-t)\{(1-t)b_0 + tb_1\} + t\{(1-t)b_1 + tb_2\}] \\ +t[(1-t)\{(1-t)b_1 + tb_2\} + t\{(1-t)b_2 + tb_3\}]\end{aligned}$$

이므로, $t = 1/3$일 때 $(x(1/3), y(1/3))$은 다음과 같은 과정을 통하여 구해진다 (그림 1 참조).

step1 c_1: b_0와 b_1을 1 : 2 로 내분하는 점
c_2: b_1와 b_2을 1 : 2 로 내분하는 점
c_3: b_2와 b_3을 1 : 2 로 내분하는 점

step2 d_1: c_1와 c_2을 1 : 2 로 내분하는 점
d_2: c_2와 c_3을 1 : 2 로 내분하는 점

step3 $(x(1/3), y(1/3))$: d_1와 d_2을 1 : 2 로 내분하는 점

베지에 곡선은 여러 가지 성질을 가지고 있으나 몇 가지 중요한 성질만 언급하면 다음과 같다. 첫째, 베지에 곡선은 항상 첫 점 b_0와 끝 점 b_N을 지난다. 이는 베지에 곡선의 정의로 부터 $t = 0$과 $t = 1$을 대입하면 당연하다. 둘째는 대칭성이다. 즉, $b_0, b_1, \cdots, b_N$에 근거한 베지에 곡선은 $b_N, b_{N-1}, \cdots, b_0$에 근거한 베지에 곡선과 일치한다. 셋째, $\sum_{j=0}^{N}(j/N)B_{N,j}(t) = t$로부터 알 수 있듯이 베지에 점들이 일직선상에 있으면 베지에 곡선은 이 점들을 연결한 직선이 된다. 넷째, 베지에 곡선 $b(t)$를 t에 대해 1차 미분하면

$$\frac{d}{dt}b(t) = N\sum_{j=0}^{N-1}(b_{j+1} - b_j)B_{N-1,j}(t) \tag{2.2}$$

가 됨을 쉽게 보일 수 있다.

3. 회귀함수의 추정

베지에 곡선을 이용하여 (1.1)에 주어진 회귀함수를 추정하기 위한 가장 중요하고 우선적으로 고려해야 할 점은 베지에 점을 어떻게 결정하는 가이다. 왜냐하면 베지에 곡선은 전적으로 주어진 베지에 점에 의존하기 때문이다. 먼저, f의 빈 추정치(bin estimator)를 구한다. 즉, 주어진 자료를 평면상에 산점도를 그린 다음 x가 존재하는 구간 $[a, b]$를 m등분하여 각 구간에 존재하는 y들의 산술평균을 구하여 그 높이를 추정치로 사용한다(그림 2 참조). 이러한 빈 추정치에서 각 구간의 막대의 양 끝점을 베지에 점으로 선택한다(그림 3 참조). 단, 첫 막대의 왼쪽 점과 마지막 막대의 오른쪽 점은 두 점씩 부여한다. 그 이유는 각 구간의 경계에서 두 점씩 부여되었으므로 일치성을 유지시키기 위한 것이다. 따라서, 베지에 점의 개수는 $2m + 2$개가 되며, 이 점들에 의한 베지에 곡선 $\hat{f}_B$를 회귀함수 f의 추정치로 제안하고자 한다.

이상의 논의를 수식으로 표현해 보면 베지에 점 $(z(j), w(j)), j = 0, 1, \cdots, 2m + 1$은

$$\begin{aligned} z(j) &= a + (b-a)\left\{\frac{j}{2m} + \frac{(-1)^j - 1}{4m}\right\} \\ w(j) &= \bar{y}_{k(j)} \end{aligned}$$

이며, 여기서

$$k(j) = \frac{j}{2} + \frac{1-(-1)^j}{4} + I_{\{0\}}(j) - I_{\{2m+1\}}(j)$$

이고 $\bar{y}_i, i = 1, \cdots, m$은 각 구간에서 y값들의 표본평균이다.

$\hat{f}_B$의 근사적 성질은 Kim, Kim, Hong, and Park(1997)에 의해 $x \in [0,1]$이고 x값들이 등간격으로 주어져 있는 상황에서 논의되었다. 즉, $\hat{f}_B$의 근사적 평균오차자승적분 (Asymptotic Mean Integrated Square Error)이

$$
\begin{aligned}
AMISE(\hat{f}_B) &= \{c_1(4c_1/c_2)^{-4/5} + c_2(4c_1/c_2)^{1/5}\}n^{-4/5} \\
c_1 &= \frac{1}{16}\int_0^1 \{(3-4x)f'(x) + x(1-x)f''\}^2 dx \\
c_2 &= \sigma^2 \int_0^1 \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{x(1-x)}} dx = \sqrt{\frac{\pi}{2}}\sigma^2
\end{aligned}
$$

이다. 이에 근거하여 $AMISE$를 최소화하는 최적의 m값은

$$m_{opt} = (4c_1/c_2)^{2/5} n^{2/5}$$

임을 쉽게 보일 수 있다. 한편, 커널 추정치 f_λ의 $AMISE$는

$$1.25\left\{\int_0^1 |f''(x)|^2 dx\right\}^{1/5}\left\{\sigma^4\left(\int_{-1}^1 K(t)^2 dt\right)^2\left(\int_{-1}^1 t^2 K(t) dt\right)\right\}^{2/5} n^{-4/5}$$

이므로 $\hat{f}_B$와 f_λ의 근사적 차수(asymptotic order)는 $n^{14/5}$으로 같음을 알 수 있다.

회귀함수 f의 미분인 f'의 베지에 추정치 $\hat{f}'_B$는 (2.2)에 의해 쉽게 계산된다.

4. 예제

$\hat{f}_B$와 $\hat{f}'_B$를 실제의 자료에 적용하기 위해 성장자료 (독일인 두 남녀에 대해 생후부터 20세 까지의 키를 조사한 것임. 단위는 millimeter)를 이용한다. 이 자료는 표 4.1에 주어져 있다. $mopt$은 미지의 f에 의존하므로 교차확인법(cross validation)으로 구한 값은 $m \simeq$로 주어진다. 이 값에 근거한 $\hat{f}_B$와 $\hat{f}'_B$이 그림 4와 5에 주어져 있다. $\hat{f}'_B$로부터 알 수 있듯이 성장속도가 남녀 모두 2 3세 근처에서 가장 빠르고, 그 이후로는 비슷한 양상을 보이다가 여자는 12세 근처, 남자는 14세 근처에서 최대가 된다.

表 4.1: 성장자료 : 독일인 두 남녀에 대한 생후부터 20세 까지의 키 (단위: millimeter)

나이	남	여	나이	남	여
1개월	534	525	11.5 년	1540	1452
3개월	601	584	12.0 년	1569	1493
6개월	685	652	12.5 년	1599	1524
9개월	729	696	13.0 년	1627	1552
1.1 년	762	736	13.5 년	1669	1584
1.5 년	822	798	14.0 년	1710	1600
2.0 년	885	843	14.5 년	1762	1607
3.0 년	984	930	15.0 년	1809	1620
4.0 년	1064	1014	15.5 년	1845	1626
5.0 년	1144	1085	16.0 년	1868	1631
6.0 년	1205	1135	16.5 년	1883	1632
7.0 년	1273	1200	17.0 년	1894	1636
8.0 년	1346	1266	17.5 년	1900	1639
9.0 년	1406	1316	18.0 년	1909	1639
9.5 년	-	1344	18.5 년	1913	-
10.0 년	1461	1365	19.0 년	1913	1639
10.5 년	-	1387	20.0 년	1915	1639
11.0 년	1511	1423			

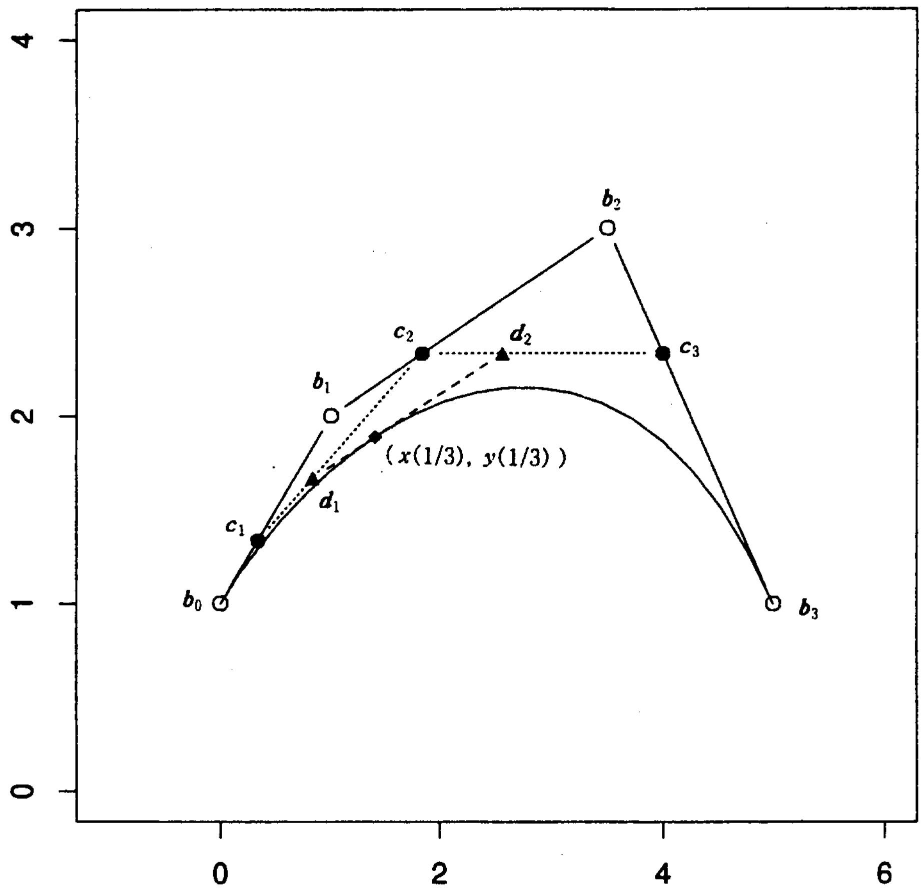

그림 1: 4개의 점 b_0, b_1, b_2, b_3에 근거한 베지에 곡선의 형성과정

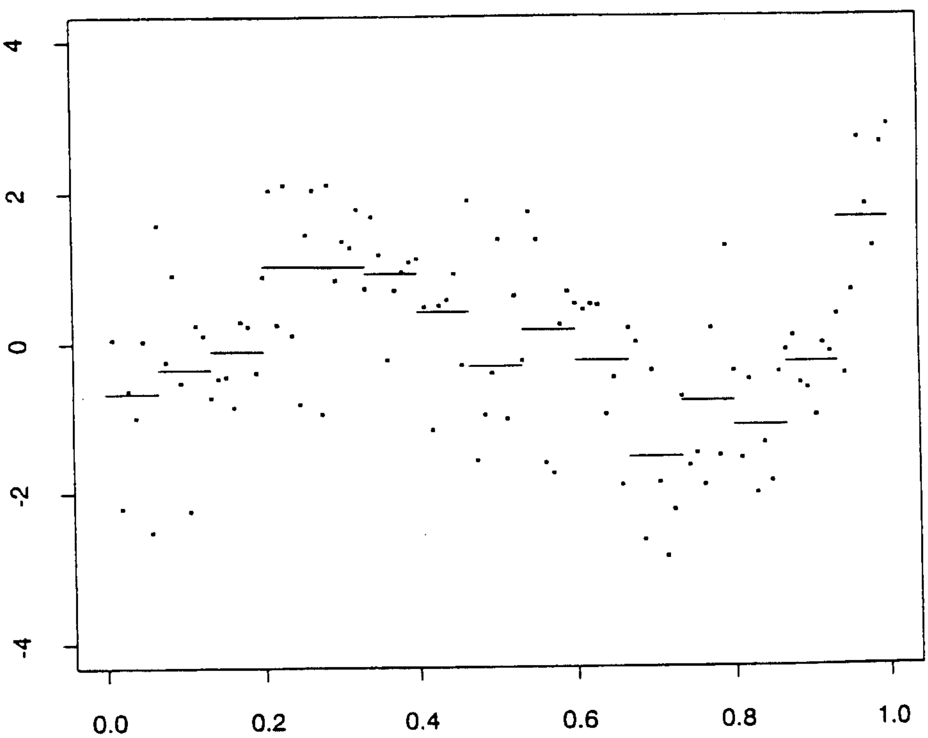

그림 2: 가상적 자료에 근거한 리그레쏘그램(regressoram)

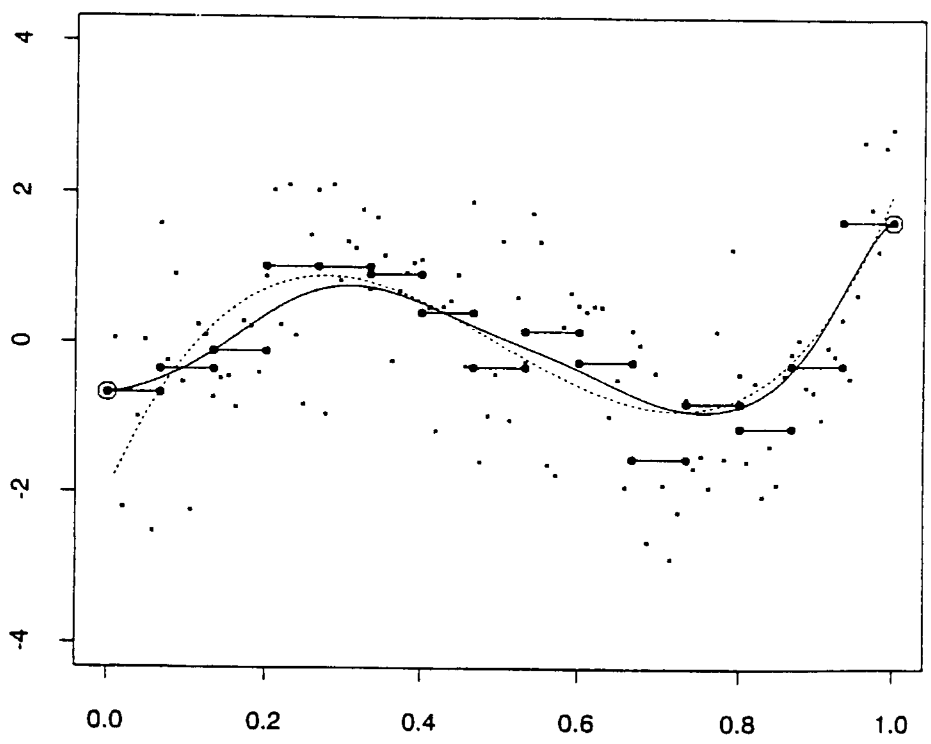

그림 3: 리그레쏘그램(regressoram)에 근거한 베지에 점과 베지에 곡선(—). 자료를 생성한 함수(···)

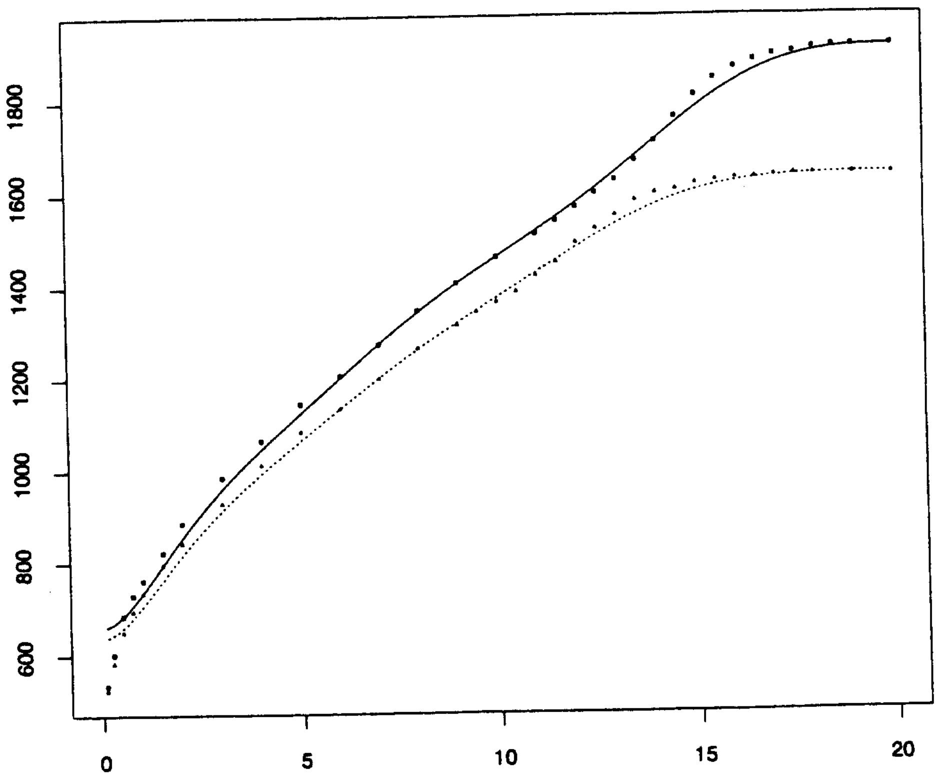

그림 4: 독일인 남녀에 대한 키의 베지에 곡선 추정치. 남(—), 여(···)

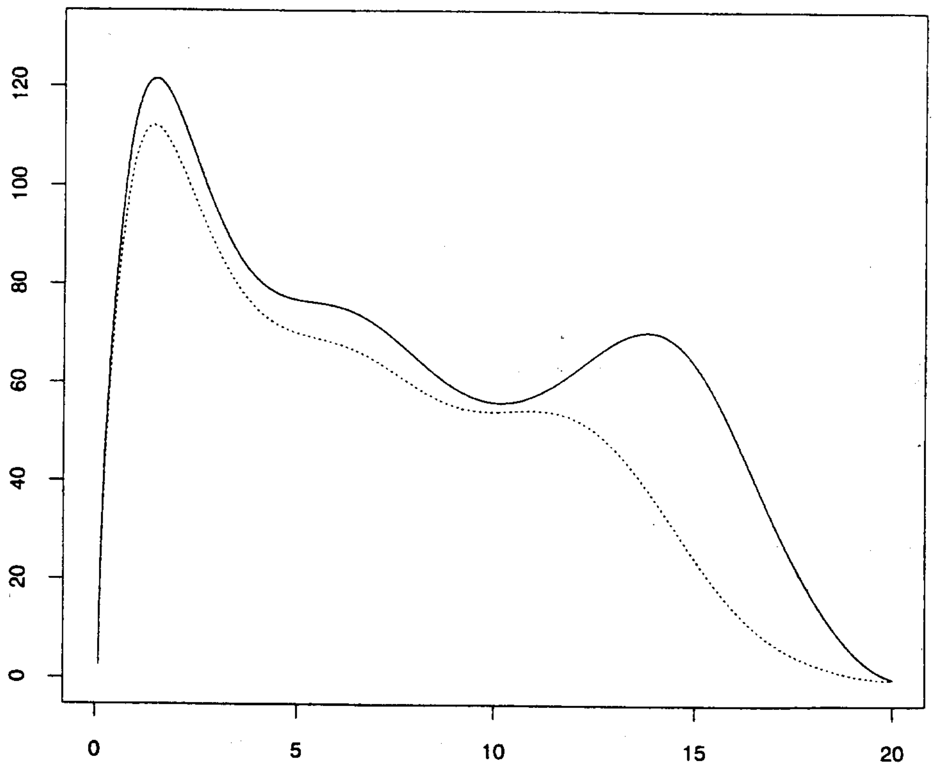

그림 5: 독일인 남녀에 대한 키의 성장속도에 대한 베지에지에 곡선 추정치.남(—),여(···)

참고문헌

[1] Bezier, P. (1977). Essay de definition numerique des courbes et des surfaces experimentals. ph. D. Thesis, University of Paris.

[2] Boehm, W. (1977). Cubic B-spline curves and surfaces in computer aided geometric design. *Computing*, 19, 29-34.

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

[4] Gasser, Th. and Müller, H. G. (1979). Kernel estimation of regression functions. *In Smoothing Techniques for Curve Estimation* (Th. Gasser and M. Rosenblatt, eds.), 23-68. Heidelberg : Springer.

[5] Gasser, Th., Müller, H. G. and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society.* Ser. B. 47, 238-252.

[6] Kim, C. (1996). Nonparametric density estimation via the Bezier curve. *Proceedings of 96 ASA Conference.* Chicago.

[7] Kim, C., Kim, W., Hong, C. and Park, B. (1997). Smoothing techniques via the Bezier curve. *Manuscript.*

[8] Müller, H. G. (1984). Smooth optimal kernel estimators of densities, regression curves and modes. *Annals of Statistics*, 12, 766-774.

[1997년 6월 접수, 1997년 12월 최종수정]

Nonparametric Estimation of the Derivative of Function via the Bezier Curve *

Choongrak Kim † Meeseon Jeong ‡ Hyoungsoon Kim §

ABSTRACT

It is quite that we have to estimate the derivative of the regression function. The Bezier curve, rarely known to statisticians, is very popular in computer graphics area. In this paper, we use nonparametric method via the Bezier curve, and apply this method to real data set. This method seems to be very easy to compute and can be easily applied to other smoothing techniques.

*The present studies were supported by the matching fund programs of research institute for basic sciences, Pusan National University, Korea, 1996, Project No. RIBS-PNU-96-104.

† Associate Professor, Department of Statistics, Pusan National University, Pusan 609-735, Korea.

‡ Department of Statistics, Pusan National University, Pusan 609-735, Korea.

§ Graduate student, Department of Statistics, Pusan National University, Pusan 609-735, Korea.