مجموعه مقالات سی‌امین کنفرانس ریاضی ایران
۱۳ـ۱۰ مرداد ۱۳۷۸، دانشگاه محقق‌اردبیلی، صص ۱۳۰ـ۱۳۳

مقایسه اسپلاین همواری با رگرسیون چندجمله‌ای

محسن محمدزاده، روشنک علی‌محمدی
گروه آمار، دانشگاه تربیت مدرس، تهران

اسپلاین همواری روشی ناپارامتری برای برازاندن یک منحنی به داده‌ها است که در آن فرض همواری منحنی در نظر گرفته می‌شود. در این مقاله با مطالعهٔ شبیه‌سازی، رفتار و دقت دو روش رگرسیون چندجمله‌ای و اسپلاین همواری برای مقادیر مختلف حجم نمونه و انحراف معیار خطاها، مورد مقایسه عددی قرار گرفته‌اند. براساس معیار میانگین مجذور خطاها، نشان داده شده‌است بازای مقادیر مختلف انحراف معیار این دو روش برای نمونه‌های بزرگ تقریباً یکسان عمل می‌کنند و برای نمونه‌های کوچک، روش اسپلاین همواری مدل بهتری ارائه می‌دهد.

کلمات کلیدی: اسپلاین همواری، رگرسیون چند جمله‌ای و اعتبارمتقابل تعمیم‌یافته

۱ مقدمه

در بسیاری از مسائل کاربردی، تعیین ارتباط بین عوامل مختلف یا بعبارت دیگر ارتباط بین دو یا چند متغیر به منظور اینکه بتوان مقدار یک متغیر را با استفاده از مقادیر متغیرهای دیگر تخمین زد، از اهمیت خاصی برخوردار است. در آمار روشهای مختلفی برای تعیین نوع ارتباط بین متغیرها و ارائهٔ مدلی ریاضی که نشان‌دهنده رابطه بین آنها باشد وجود دارد. از جمله می‌توان روشهای آنالیز رگرسیون و روشهای هموارسازی را مورد استفاده قرار داد.

فرض کنید n مشاهده برای متغیرهای Y و X در دست باشند و مدل

$$y_i = g(x_i) + e_i \quad ; i = ۱, \ldots, n \qquad (۱)$$

را بتوان به‌عنوان رابطه بین دو متغیر درنظر گرفت که در آن e_i ها خطاهای تصادفی مستقل و g تابعی نامعلوم است. هدف برآورد تابع نامعلوم g بر اساس مشاهدات می‌باشد. درحالت کلی برای تعیین برآوردی منحصر بفرد از تابع g ضروری است شرایطی یا محدودیتهایی را در نظر گرفت. به عنوان مثال، در روش آنالیز رگرسیون خطی فرض می شود g تابعی خطی است و با استفاده از روش کمترین توانهای دوم، ضرایب مدل خطی به‌طور منحصر بفرد برآورد می‌گردند. در مواردی که برازش مدل خطی به داده‌ها مناسب نباشد، ممکن است روش رگرسیون چند جمله‌ای برآورد بهتری را برای g فراهم نماید، یا روش ناپارامتری برای برآورد g مورد استفاده قرار گیرد. در روش اسپلاین همواری با فرض اینکه g تابعی هموار از x است اقدام به برآورد آن می‌گردد. دراین مقاله دو روش رگرسیون چندجمله‌ای و اسپلاین همواری برای برآورد تابع g بکارگرفته شده و نکویی برازش آنها با استفاده از تکنیک شبیه‌سازی و بر اساس مجموع مجذورات خطاها مورد مقایسهٔ عددی قرار گرفته‌اند.

۲ رگرسیون چندجمله‌ای

در مواردی که رابطه بین متغیر پاسخ و متغیرهای مستقل به‌صورت منحنی باشد، از رگرسیون چندجمله‌ای استفاده می‌شود. حتی هنگامی که رابطهٔ غیرخطی پیچیده‌ای برقرار باشد مدل چندجمله‌ای روی دامنه‌های کوچک متغیر X قابل بکارگیری است. در حالت کلی مدل چندجمله‌ای یک متغیره از درجه k به صورت

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k + \epsilon$$

در نظر گرفته می‌شود و پارامترهای آن به روش کمترین توانهای دوم برآورد می‌گردند. در این روش باید درجه مدل چندجمله‌ای مناسب با ساختار داده‌ها انتخاب شود و ضروری است از برازش مدلی با درجه بالاتر از حد مطلوب اجتناب شود. یک روش انتخاب درجه چندجمله‌ای افزایش درجه تا زمانی است که آزمون t برای جمله دارای بالاترین درجه معنی‌دار نباشد، که روش پیشرو نامیده می‌شود. دیگری که روش پسرو نام دارد، برازش مدل از بالاترین درجه شروع و سپس با حذف عبارتهای بالاترین مرتبه ادامه می‌یابد تا هنگامی که جملات باقیمانده و با بزرگترین درجه در مدل با آزمون t معنی‌دار شود. جزئیات این دو روش که لزوماً به یک جواب منتهی نمی‌شوند، در مونت گومری (۱۹۹۱) بطور مفصل مورد بررسی قرار گرفته است.

۳ اسپلاین هموار

اگر مدل (۱) با روش کمترین مربعات باقیمانده‌ها برازش داده شود و برای تابع نامعلوم g شرطی منظور نشود و نقاط با خطوط راست به هم وصل شوند ، مجموع مربعات باقیمانده‌ها در نقاط مشاهده شده صفر می‌شود، اما منحنی برازنده شده به داده‌ها مواج و پرنوسان خواهد بود که از نظر کاربردی ممکن است چندان مطلوب نباشد. لذا منطقی است اگر خواسته شود g توسط یک منحنی هموار برآورد گردد. بدین منظور فرض می‌شود مشتق اول و دوم تابع g موجود و مجذور مشتق دوم آن نیز انتگرال پذیر است. در اینصورت میزان هموار تابع g را می توان توسط عبارت $\int_a^b \{g^m(x)\}^2 dx$ اندازه گیری نمود. از طرفی میزان نکویی برازش g به داده‌ها توسط مجموع مجذورات خطاهای $\sum(y_i - g(x_i))^2$ تعیین می‌شود. لذا برای برآورد منحنی g، می‌توان علاوه بر عامل مجموع مربعات باقیمانده‌ها، اندازه هموار g را نیز درنظر گرفته و ملاک مجموع مربعات جریمه‌ای را بصورت

$$S(g,\lambda) = \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int_a^b \{g^m(x)\}^2 dx$$

تعریف نمود، که درآن $\lambda > 0$ پارامتر هموار نامیده می‌شود و تقابل بین نکویی برازش منحنی به داده‌ها و میزان هموار تابع g را کنترل می‌کند. برای یک مقدار داده شده λ، مینیمم کردن $S(g,\lambda)$ بهترین راه توافق بین هموار و نیکویی برازش است.

فرض کنید $x_1, \ldots x_n$ روی بازه $[a,b]$ در شرط $x_1 < x_2 < \ldots < x_n$ صدق کنند و تابع g بر روی $[a,b]$ تعریف شود. تابعی مانند $\hat{g}$ با شرایط فوق‌الذکر که مجموع مربعات جریمه‌ای $S(g,\lambda)$ را کمینه نماید، اسپلاین هموار از درجه $2m-1$ نام دارد و بصورت

$$g(x) = \sum_{j=1}^{m} a_j x^{j-1} + \sum_{i=1}^{n} b_i |x - x_i|^{2m-1}$$

قابل حصول می‌باشد. بطوریکه ضرایب $b_1, \ldots, b_n$ در شرایط $\sum_{i=1}^{n} b_i = 0$ و $\sum_{i=1}^{n} b_i x_i = 0$ صدق نماید. به‌عنوان مثال، اگر $m = 2$ در نظر گرفته شود، تابع اسپلاین هموار به‌صورت زیر خواهد بود.

$$g(x) = a_1 + a_2 x + \sum_{i=1}^{n} b_i |x - x_i|^3$$

برای توضیح بیشتر دراین زمینه می‌توان به واهبا (۱۹۹۰)، هاردل (۱۹۹۰)، هیستی و تیبشیرانی (۱۹۹۰) و گرین وسیلورمن (۱۹۹۴) مراجعه کرد.

در روش اسپلاین هموارى مقدار λ نقش مهمی دربرآورد منحنی ایفا می‌نماید. اگر λ بزرگ باشد، مولفه اصلی در $S(g,\lambda)$ عبارت جریمه ناهموارى خواهد بود و کمینه کننده آن یعنی $\hat{g}$، خیلی کم انحنا خواهد شد (شکل ۳). در حد وقتی λ به بی نهایت میل کند مقدار $\int_a^b\{g''(x)\}^2dx$ بایستی به سمت صفر میل کند و در نتیجه برآورد منحنی همان رگرسیون خطی خواهد بود. از طرف دیگر، اگر λ نسبتا کوچک باشد مجموع مربعات باقیمانده‌ها سهم اصلی را در $S(g,\lambda)$ خواهد داشت و برآورد g تا حد زیادی روند داده‌ها را دنبال می‌کند، حتی اگر به بهای تغییرات سریع آن تمام شود. در وضعیت حدی اگر λ به صفر نزدیک شود، برآورد حاصل به اسپلاین درونیاب میل می‌کند (شکل ۲). بنابراین محاسبه مقدار مناسب برای λ در اسپلاین هموارى مساله مهمی است. دو روش اعتبار متقابل و اعتبار متقابل تعمیم یافته برای محاسبه مقدار بهینه λ بر اساس داده‌ها در گرین و سیلورمن (۱۹۹۴) مطرح گردیده است و محمدزاده (۱۹۹۸b و۱۹۹۸a) الگوریتمی برای محاسبهٔ سریع آن ارائه نموده است.

امروزه از روشهای اسپلاین برای برآورد منحنی پاسخ در زمینه های گوناگون از جمله کشاورزی، اقتصاد، داروسازی، ... استفاده‌های فراوانی می‌شود. از جمله کاربردهای روش اسپلاین هموارى در مدل بندی نیمه پارامتری و همچنین مدلهای خطی تعمیم‌یافته که اولین بار توسط نلدر و ودربرن (۱۹۷۲) ارائه شده است، می‌توان نام برد.

۴ مقایسه و نتیجه‌گیری

برای مقایسه دو روش رگرسیون چند جمله‌ای و اسپلاین هموارى اقدام به تولید داده‌های تصادفی و برازش منحنی به آنها بر اساس دو روش مذکور گردیده است. بدین منظور ابتدا n مقدار تصادفی از توزیع یکنواخت $\mathbf{u}(0,1)$ تولید شده تا مشاهدات $x_1,\ldots,x_n$ بدست آیند.و سپس n مقدار تصادفی $e_1,\ldots,e_n$ از توزیع $N(0,\sigma^2)$ تولید گردیده و بااستفاده از رابطه $y=x+x^2+x^3+e$ مشاهدات $y_1,\ldots,y_n$ بدست آمده‌اند. مقادیر $y_1,\ldots,y_n$ بهمراه $x_1,\ldots,x_n$ یک مجموعه داده را تشکیل داده‌اند و به دو روش رگرسیون چندجمله‌ای و اسپلاین هموارى یک منحنی به داده‌ها برازانده می‌شوند. حال برای اینکه میزان تاثیر حجم نمونه (n) و انحراف معیار (σ) را در میزان دقت مدلهای برازانده شده مورد بررسی قرار داده شود، چهار حجم نمونه ۳۵ ، ۵۰ ، ۷۵ و ۱۰۰ و سه انحراف معیار ۰/۱ ، ۰/۵ و ۱ در نظر گرفته شده است. برای هر ترکیب از n و σ یک نمونه تصادفی n تایی به روشی که ذکر شد تولید شده است.

برای مقایسه دو روش اسپلاین هموارى و رگرسیون چند جمله‌ای ،ابتدا n را ثابت در نظر گرفته و اثر افزایش σ بر مجموع مجذورات خطاها(SSE) مورد بررسی قرار داده شده است. سپس برای مقدار ثابت σ تاثیر تغییرات n بر SSE مورد مطالعه قرار گرفته است.

مقایسه مقادیر مختلف SSE در شکل‌های (۱) الی (۷) نشان می‌دهد برای مقادیر کوچک n (کمتر از ۷۵) و برای تمام مقادیر σ اسپلاین هموارى بهتر از رگرسیون چندجمله‌ای منحنی را برآورد می‌کند، ولی برای نمونه‌های بزرگ (بزرگتر از ۷۵) تفاوت معنی‌داری بین دو روش وجود ندارد و براساس ملاک SSE دو روش یکسان عمل می‌کنند. چون با افزایش حجم نمونه تعداد پارامترهایی که در روش اسپلاین باید برآورد شوند افزایش می‌یابد و محاسبه آنها طولانی می‌گردد، برازاندن رگرسیون چندجمله‌ای سریعتر از اسپلاین هموارى صورت می‌پذیرد و بکاربردن آن مرقوم به صرفه‌تر از روش اسپلاین هموارى می‌باشد.

مراجع

[1] Green, P. J. and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Model*, London, Chapman and Hall.

[2] Hardle, W. (1990), *Applied Nonparametric Regression*, Cambridge: Cambridge University Press.

[3] Hastie, T. J. and Tibshirani, R. J. (1990),

tational Statistics, 81 -82, Bristol, UK.

[6] Montgomery, D. C. and Peak, E. A. (1992), *Introduction to Linear Regression Analysis*, 2nd, New York, John Wiley and Sons.

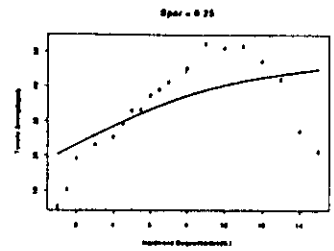
[7] Nelder, J. A. and Wedderburn, R. W. M. (1972), Generalized Linear Models. *J. Roy. Statist. Soc A* 135 370-384

[8] Wahba, G. (1990), *Spline Models for Observational Data*,Philodelphia: SIAM.

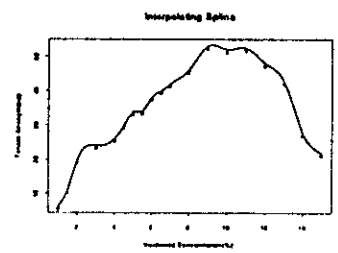
Generalized Additive Models, London, Chapman and Hall.

[4] Mohammadzadeh, M. (1998a), An Algorithm to Find the Smoothing Parameter in Smoothing Splines, *Proceeding of the 4th Iranian Statistical Conference, Shahid Beheshti University, Tehran, Iran.*

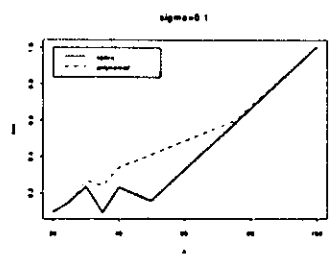
[5] Mohammadzadeh, M. (1998b), Estimating the Smoothing Parameter in Smoothing Splines, *COMPSTAT' Proceedings in Compu-*



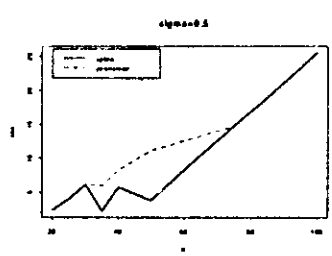
شکل ۱: اسپلاین همواری با $\lambda = ۰/۲۵$



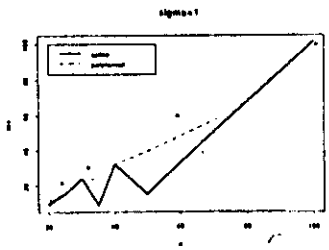
شکل ۲: اسپلاین درونیاب با $\lambda = ۰$



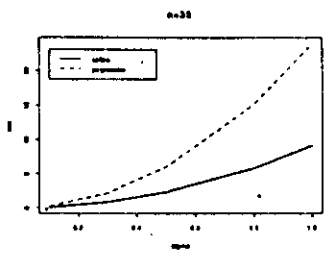
شکل ۳: SSE حاصل از دو روش، برای $\sigma = ۰/۱$



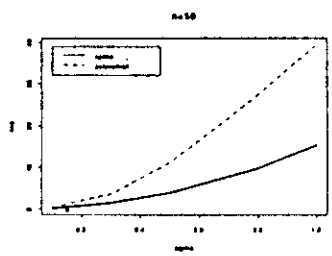
شکل ۴: SSE حاصل از دو روش، برای $\sigma = ۰/۵$



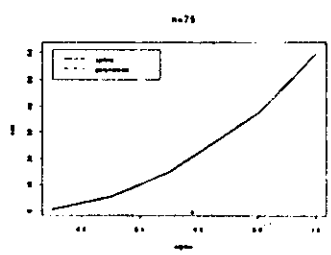
شکل ۵: SSE حاصل از دو روش، برای $\sigma = ۱$



شکل ۶: SSE حاصل از دو روش، برای $n = ۲۵$



شکل ۷: SSE حاصل از دو روش، برای $n = ۵۰$



شکل ۸: SSE حاصل از دو روش، برای $n = ۷۵$