مقدمه‌ای بر هموارسازهای اسپلاین

مهدیه میرزایی باغینی، الهه آخوندزاده کاشانی
دانشگاه آزاد کرمان

چکیده

دو راهکار پارامتری و ناپارامتری برای برآورد مدل‌های غیر خطی رگرسیون به کار برده می‌شود و راهکار ناپارامتری که برازش دقیق‌تری به داده‌ها انجام می‌دهد از چند جمله‌ای‌های قطعه‌ای با درجه پیوستگی معینی که اسپلاین نامیده می‌شوند، بهره می‌گیرد. اسپلاین‌ها یک خانواده محبوب از هموارسازها هستند. یکی از گیج کننده‌ترین جنبه‌های اسپلاین–ها انواع بسیار زیاد آنهاست که در این مقاله به معرفی توابع اسپلاین پرداخته و نحوه استفاده از آنها در برازش منحنی ناپارامتری هموار به داده‌ها بیان می‌شود.

واژه‌های کلیدی: توابع اسپلاین، هموارسازی، رگرسیون ناپارامتری، جریمه، پارامتر هموارسازی.

۱ مقدمه

رگرسیون خطی یکی از قدیمی‌ترین و پراستفاده‌ترین تکنیک‌های آماری است که در ساده‌ترین حالت مقدار مورد انتظار یک متغیر پاسخ تک متغیره را برحسب یک پیش بین خطی مدل‌سازی می‌کند. یک پیش بین خطی تابعی پارامتری از متغیرهای توضیحی است که روی مقدار متغیر پاسخ اثر گذار است. مجموعه داده‌های زیادی وجود دارند که برازاندن یک مدل خطی به آن‌ها مناسب نیست. می‌توان پیش بینی خطی را با اضافه کردن توابع چندجمله‌ای از متغیرهای توضیحی تعمیم داد، اما مدل‌های چندجمله‌ای در بسیاری از مواقع داده‌ها را به درستی برازش نمی‌دهند که در این صورت می‌توان برای تلخیص روند متغیر پاسخ به عنوان تابعی از متغیرهای توضیحی از هموارسازها استفاده کرد. ویژگی مهم یک هموارساز ماهیت ناپارامتری آن است بدین صورت که یک فرم تابعی از پیش تعیین شده برای بیان رابطه بین این رابطه به طور مستقیم از داده‌ها حاصل می‌شود، به همین دلیل هموارسازی ابزاری برای رگرسیون ناپارامتری تلقی می‌شود. میانگین متحرک، خط متحرک، هموار سازهای هسته و هموار سازهای اسپلاین چند نمونه از تکنیک‌های

هموارسازی هستند.

در ابتدا تکنیک‌های برآورد مدل‌های رگرسیون با استفاده از توابع اسپلاین بر اساس دو شیوه‌ی اصلی بوده است: اسپلاین‌های رگرسیونی (هیستی و همکاران [۹]) و اسپلاین‌های هموارسازی (گرین و سیلورمن [۷]).

اسپلاین‌های رگرسیونی با استفاده از یک تعداد کوچک گره تعریف می‌شوند که برای تضمین هموار بودن منحنی به دقت انتخاب شده‌اند. با این فرض که f یک تابع نامعلوم هموار [۱] در یک مدل رگرسیون ناپارامتری باشد، اسپلاین‌های هموارسازی بهترین جواب برای برآورد f هستند اما تعداد پارامترهایی که باید برآورد شود به بزرگی تعداد مشاهدات است و این باعث می‌شود اسپلاین‌های هموارسازی یک شیوه شدیداً محاسباتی برای مدل‌سازی توابع نامعلوم f باشند. تقریباً به صورت هم زمان آیلرز و مارکس[۲] [۵] و راپرت و کارول[۳] [۱۱] استفاده از اسپلاین‌های جریمه شده را برای برای برآورد توابع هموار موجود در مدل‌های رگرسیونی پیشنهاد کردند. اسپلاین‌های جریمه شده را می‌توان به عنوان حد وسط برای اسپلاین‌های رگرسیونی و هموارسازی در نظر گرفت.

در این مقاله نخست به رگرسیون خطی و چندجمله‌ای به عنوان نمونه‌هایی از مدل‌های رگرسیون پارامتری اشاره خواهیم کرد تا ناتوانی این مدل‌ها را در برآورد رابطه واقعی بین دو متغیر نشان دهیم در بخش سوم به معرفی هموارسازهای اسپلاین می‌پردازیم، در بخش چهارم در رابطه با مسئله انتخاب گره بحث می‌کنیم، در بخش پنجم اسپلاین‌های هموارسازی را معرفی می‌کنیم و در بخش ششم به معرفی اسپلاین‌های جریمه شده می‌پردازیم.

[۱] تابعی که مشتق اول پیوسته داشته باشد.

[۲] Eilers and Marx

[۳] Ruppert and Carroll

در این مقاله از مجموعه داده‌های فسیل موجود در نرم‌افزار R استفاده کرده‌ایم. در این مجموعه داده که دارای ۱۰۶ مشاهده تصادفی است دو متغیر وجود دارد یک متغیر توضیحی که سن فسیل برحسب میلیون سال است و یک متغیر پاسخ که میزان ایزوتوپ استرانسیوم موجود در فسیل‌ها است. کلیه شکل هایی که در ادامه خواهد آمد در نرم افزار R با استفاده از پکیج‌های semi- mgcv، splines par، و pspline رسم شده‌اند [۱].

۲ رگرسیون پارامتری

هدف از انجام رگرسیون برآورد رابطه آماری بین متغیر توضیحی X و متغیر پاسخ Y می‌باشد. برای برآورد یک مدل رگرسیونی باید تابع f را به گونه‌ای به دست آورد که $E(Y-f(X))^2$ مینیمم شود، با مینیمم‌سازی این عبارت میانگین شرطی Y به شرط X به صورت زیر حاصل می‌شود:

$$f(X)=E(Y\mid X),$$

که این تابع رگرسیون Y روی X نامیده می‌شود. از تعریف f داریم:

$$Y_i=f(X_i)+\epsilon_i,\qquad i=1,\ldots,n.$$

در عبارت بالا ϵ_i خطای تصادفی است. هنگامی که یک مدل خطی به داده‌ها برازش دهیم مدل رگرسیون به صورت زیر خواهد بود:

$$y=\beta_0+\beta_1 x+\epsilon_i,\qquad i=1,\ldots,n. \qquad (1)$$

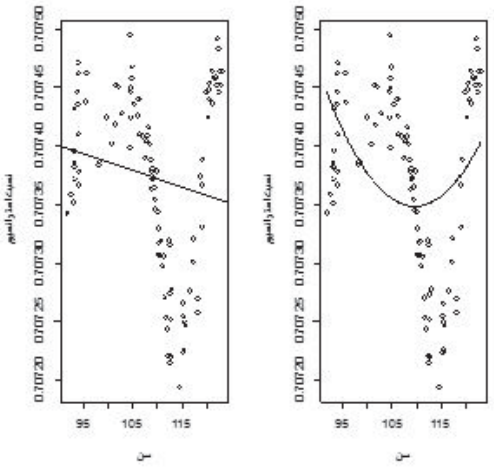
در مدل بالا فرض می‌شود x و y معلومند و فقط پارامترهای β_0 و β_1 نامعلوم هستند. از آن جا که با برآورد این دو پارامتر مدل رگرسیون برآورد می‌شود، مدل

رگرسیون خطی یک مدل پارامتری نامیده می‌شود.

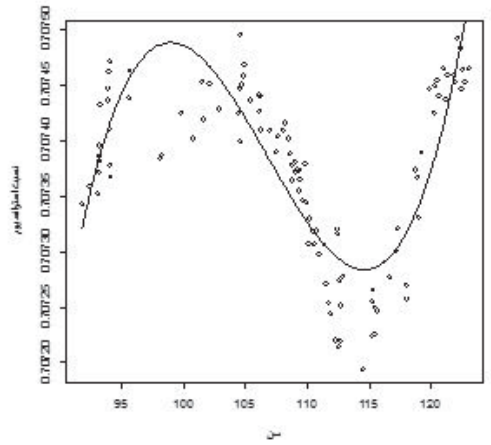
نمودار پراکنش مربوط به داده‌های فسیل و همچنین خط حاصل از برازش مدل خطی به داده‌ها با استفاده از فرمان ()lm در شکل ۱ نمایش داده‌ایم. با توجه به شکل ۱ کاملاً واضح است که رابطه بین سن فسیل و میزان ایزوتوپ استرانسیوم موجود در آن یک رابطه خطی نیست. با اضافه کردن درجات بالاتر X به مدل (۱) می‌توان مدل‌های چندجمله‌ای به صورت زیر به داده‌ها برازش داد.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n. \qquad (۲)$$

اگر در مدل (۲) قرار دهیم $n = ۲$ یک مدل دوجمله‌ای به داده‌ها برازش داده می‌شود. اما مدل دوجمله‌ای نیز توصیف‌کننده رابطه صحیح بین متغیر پاسخ و توضیحی نیست.

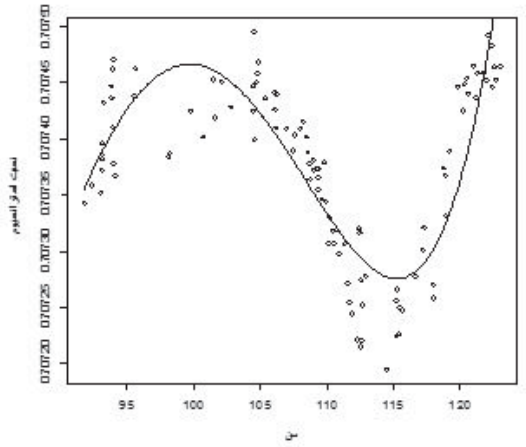


شکل ۱: نمودار پراکنش مربوط به داده‌های فسیل و برازش مدل خطی و دو جمله‌ای به داده‌ها



شکل ۲: منحنی حاصل از برازش رگرسیون درجه ۳ به داده‌ها

با افزایش درجه چندجمله‌ای در شکل های ۲ و ۳ به ترتیب یک مدل درجه سه و درجه چهار به داده‌ها برازش داده شده است اما باز هم منحنی‌های حاصل شده بیانگر رابطه واقعی بین سن فسیل و میزان ایزوتوپ استرانسیوم موجود در آن‌ها نمی‌باشند.



شکل ۳: منحنی حاصل از برازش رگرسیون درجه ۴ به داده‌ها

۳ هموارسازهای اسپلاین

اسپلاین‌ها چندجمله‌ای‌های تکه‌ای هستند که در نقاطی به نام گره به هم متصل می‌شوند. اسپلاین‌ها در ادبیات آماری به عنوان درون‌یاب معرفی شده‌اند اما بیشتر مدل‌های آماری با یک خطای اندازه‌گیری داده‌ها را برازش می‌دهند. بنابراین باید یک نوع اسپلاین ایجاد کنیم که بتواند از نزدیک داده‌ها عبور کند اما نه فقط مشروط به آن که آن‌ها را درون‌یابی کند. به این عمل هموارسازی اسپلاین گفته می‌شود.

۱.۳ توابع اسپلاین

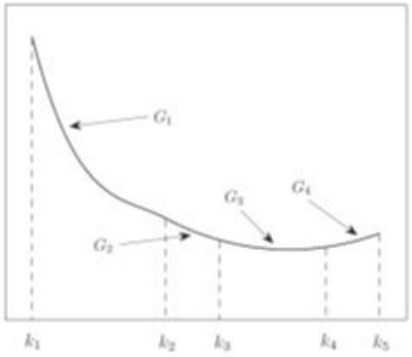
یک تابع اسپلاین $g(x)$ از درجه s یک چندجمله‌ای تکه‌ای است که چندجمله‌ای‌های تکه‌ای (همه از درجه s) برای ساختن یک منحنی هموار در گره‌های k_m، $m = ۱, ۲, \ldots, K$ به هم متصل می‌شوند. مجموعه گره‌های $\{k_m\}_{m=۱}^{K}$ همیشه یک دنباله اکیداً صعودی را نمایش می‌دهند بنابراین می‌توان نوشت:

$$g(x) = G_m(x) = c_{۰m} + c_{۱m}x + c_{۲m}x^۲ + \cdots + c_{sm}x^s, \quad k_m < x < k_{m+۱}.$$

تکه‌های چندجمله‌ای $G_m(x)$ به صورت هموار در گره‌ها به هم متصل می‌شوند یعنی در گره‌ها پیوسته و نیز دارای $s-۱$ مشتق پیوسته هستند به عبارت دیگر

$$G_m^{(d)}(k_{m+۱}) = G_{m+۱}^{(d)}(k_{m+۱}).$$

شکل ۳ قطعه‌های چندجمله‌ای را نمایش می‌دهد که یک اسپلاین درجه ۳ با ۵ گره $k_۱, \ldots, k_۵$ را تعریف می‌کند. گره‌های $\{k_m\}_{m=۱}^{K}$ دامنه متغیر دلخواه X را می‌پوشانند. بنابراین اگر $x \in [l, r]$ باشد آن‌گاه، $k_۱ = l$ ، $k_K = r$ و هر تابع روی بازه $[l, r]$ را می‌توان به وسیله اسپلاین‌های چندجمله‌ای با درجه ثابت s که با یک تعداد کافی از گره‌ها مشخص شده‌اند، تقریب زد. اسپلاین‌های با تعداد گره ثابت را اسپلاین‌های رگرسیونی می‌نامند.



شکل ۴: برازش اسپلاین درجه ۳ با ۵ گره

۲.۳ پارامتر سازی‌های اسپلاین

تعریف یک تابع اسپلاین برحسب چندجمله‌ای‌ها هنگامی مناسب است که ضرایب چندجمله‌ای معلوم باشد. اما از نظر محاسبه‌ای بهتر است یک اسپلاین از درجه s با گره‌های $\{k_m\}_{m=۱}^{K}$ را به عنوان ترکیب خطی از توابع پایه اسپلاین تعریف کنیم. توابع پایه اسپلاین مستقل خطی از درجه s با گره‌های $\{k_m\}_{m=۱}^{K}$ هستند که فضای توابع اسپلاین را تولید می‌کنند.

۱.۲.۳ پایه

در جبر خطی پایه یک مجموعه از بردارهای مستقل خطی است که هر بردار در یک فضای برداری داده شده را

می‌تواند در یک ترکیب خطی نشان دهد. برای مثال مدل

$$y = \beta_0 + \beta_1 x_i + \epsilon_i,$$

یک ترکیب خطی از توابع ۱ و x است بنابراین $[1, x]$یک پایه برای فضای برداری همه چند جمله‌ای های خطی در x است.

بنابراین می‌توان گفت توابع پایه اسپلاین یک مجموعه از توابع اسپلاین مستقل خطی از درجه با گره‌های $\{k_m\}_{m=1}^{k}$ هستند که فضای توابع اسپلاین را تولید می‌کنند.

۲.۲.۳ پایه توانی بریده

پایه توانی بریده (راپرت و همکاران[۱۲]) بر مبنای تابع توانی بریده ساخته می‌شود. یک تابع توانی بریده از درجه s به صورت زیر تعریف می‌شود:

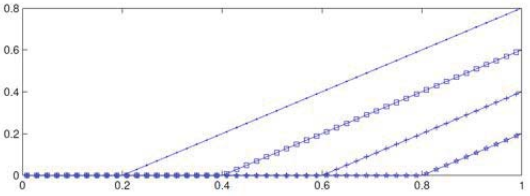
$$(x)_+^s = max\{x^s, 0\}.$$

به طور متناظر با انتخاب گره k تابع $p(x) = (x-k)_+^s$ یک چندجمله‌ای تکه‌ای از درجه s با یک نقطه توقف در k است که برای نقاط سمت چپ گره k مقدار صفر را می‌پذیرد. برای $s > 0$ تابع $p(x)$ در k پیوسته است و دارای $s-1$ مشتق پیوسته نیز می‌باشد. در شکل ۵ یک مجموعه از توابع توانی بریده برای $s = 1$ با گره‌های ۰/۲، ۰/۴، ۰/۶ و ۰/۸ نمایش داده شده است. هر تابع اسپلاین g از درجه s با گره‌های $\{k_m\}_{m=1}^{K}$ برحسب پایه توانی بریده به صورت زیر تعریف می‌شود:

$$\begin{aligned} g(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_s x^s \\ &+ \sum_{m=2}^{K-1} \gamma_m (x-k_m)_+^s , \end{aligned}$$

که یک ترکیب خطی از توابع به فرم زیر است:

$$(x-k_m)^s + x^s, \cdots, x^2, x^1, 1.$$



شکل ۵: پایه تابع توانی بریده درجه ۱ با گره‌های ۰/۲ تا ۰/۸

۳.۲.۳ اسپلاین‌های طبیعی

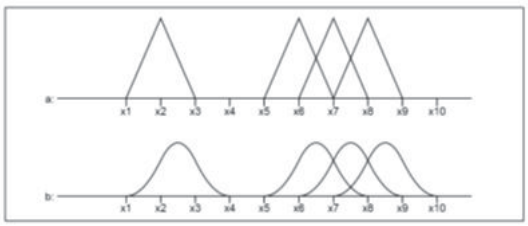
یک اسپلاین از درجه فرد $s = 2v+1$ یک اسپلاین طبیعی نامیده می‌شود اگر خارج از گره‌های مرزی (قسمت چپ گره k_1 و راست گره k_K) یک چند جمله از درجه $v-1$ باشد. برای مثال یک اسپلاین طبیعی درجه ۳ با گره‌های $\{k_m\}_{m=1}^{k}$ خارج از گره‌های مرزی خطی است و به‌صورت زیر تعریف می‌شود:

$$g(x) = \beta_0 + \beta_1 x + \beta_3 x^3 + = \sum_{m=2}^{k-1} Y_m (x-k_m)_+^3,$$

که توابع پایه آن به صورت $(x-k_m)^3 + x^3, x^2, x, 1$ هستند.

پایه توانی بریده برای فهم ساختار رگرسیون براساس اسپلاین‌ها مفید است، اما در عمل هنگامی مناسب خواهد بود که گره‌ها به دقت انتخاب شده باشند زیرا برای یک شبکه غیریکنواخت از گره‌ها بعضی از توابع پایه $(x-k_m)_+^s$ به توابع پایه دیگر وابسته خطی می‌شوند و در نتیجه تغییرات کوچک در ضرایب γ_m، می‌تواند تغییرات بزرگی در تابع $g(x)$ ایجاد کند، به علاوه اگرچه یک تابع

توانی بریده خاص برای یک بازه از گره‌های خاص تعریف شده است، در همه نقاط سمت راست گره تعیین کننده‌اش نیز مقدار می‌پذیرد و این باعث بروز مشکلاتی هنگام برآورد ضرایب می‌شود.



شکل ۶: نمایش پایه‌هایB- اسپلاین درجه ۱ و درجه ۲

۴.۲.۳ پایه B- اسپلاین

پایه B-اسپلاین‌ها از تکه‌های چندجمله‌ای تشکیل می‌شوند که به صورت خاصی به هم متصل شده‌اند. در شکل ۶a یک B-اسپلاین از درجه ۱ نمایش داده شده است. یک B-اسپلاین از درجه ۱ بر اساس گره‌های x_1 و x_2 و x_3 از دو تابع خطی یکی از x_1 تا x_2 و یکی از x_2 تا x_3 تشکیل شده است که برای مقادیر سمت چپ x_1 و سمت راست x_3 مقدارش صفر است. در شکل ۶b B- اسپلاین‌های درجه ۲ نمایش داده شده است. B-اسپلاین‌ها را می‌توان به عنوان تعمیمی از توابع توانی بریده در نظر گرفت که با استفاده از مقیاس سازی مناسب $s+1$-امین تفاضلات تقسیم شده توابع توانی بریده حاصل می‌شوند [۶]. فرض کنید

$$\psi_l(x) = \frac{(k_l - x)_+^s}{\prod_{v\neq l}^{m+s+1}(k_l - k_v)}.$$

یک تابع B-اسپلاین از درجه s به صورت زیر تعریف می‌شود:

$$B_m(x) = (k_{m+s+1} - k_m)\sum_{l=m}^{m+s+1}\psi_l(x).$$

برخی از ویژگی‌های مهم B-اسپلاین‌ها به شرح زیر است:

- از تعریف مشخص می‌شود که تابع $B_m(x)$ دارای یک تکیه‌گاه کوچک است به عبارت دیگر
$$B_m(x) = 0, \quad x\notin[k_m, k_{m+s+1}].$$

از این ویژگی نتیجه می‌شود که در هر بازه خاص $[k_u, k_{u+1}]$ فقط $s+1$ ، B- اسپلاین به صورت $B_{u-s},\ldots,B_u$ وجود دارد. برای تولید یک پایه B-اسپلاین کامل، یک مجموعه از گره‌های اضافی در دو انتهای دامنه g لازم داریم بنابراین با داشتن دنباله گره $\{k_m\}_{m=1}^K$ ، $2s$ گره اضافی به صورت زیر انتخاب می شوند:

$$k_{-(s-1)} < \cdots < k_0 < k_1 < \cdots < k_K < k_{K+1} < \cdots < k_{K+s}.$$

- برای هر $x\in[l,r]$ داریم $\sum_m B_m(x) = 1$.
- توابع $B_m(x)$ روی تکیه‌گاهشان مثبت هستند یعنی:

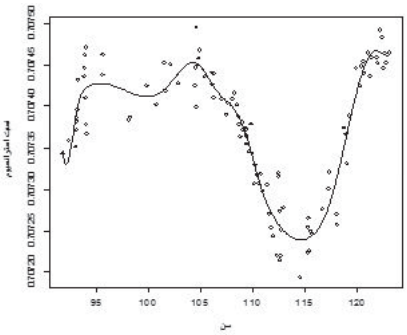
$$B_m(x) > 0, \qquad x\in[k_m, k_{m+s+1}].$$

تابع مورد نظر g را می‌توان با استفاده از B-اسپلاین‌ها به فرم زیر نمایش داد:

$$g(x) = \sum_{m=1}^{K_B}\gamma_m\beta_m(x), \quad x\in[r,l], \qquad (۳)$$

که در آن $K_B = K+s+1$ بعد پایه B-اسپلاین است. طبیعت موضعی توابع B-اسپلاین فرآیند برازشی را نتیجه می‌دهد که از لحاظ عددی پایدارتر است از آنچه به وسیله

توابع توانی بریده به دست می‌آید. (به مارکس و آیلرز [۶] مراجعه شود.)



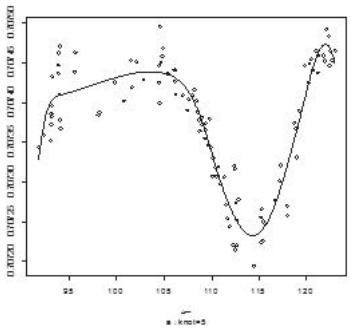
شکل ۷: برازش یک B- اسپلاین به داده‌ها

در شکل ۷ نمودار پراکنش مربوط به داده‌های فسیل را نشان می‌دهد که با استفاده از پایه B-اسپلاین و ده گره با استفاده از فرمان ()bs برازش داده شده است.

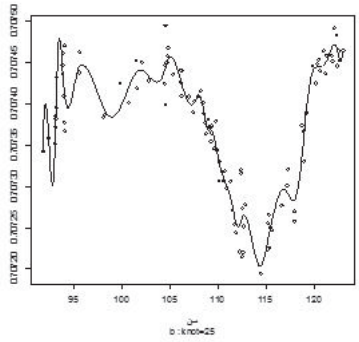
۴ انتخاب تعداد و مکان گره‌ها

مشکل اصلی در رابطه با اسپلاین‌های رگرسیونی انتخاب تعداد و مکان گره‌ها است. یک روش استاندارد قرار دادن گره‌ها در بازه‌هایی ست که به صورت یکسان فاصله گذاری شده‌اند. اگر داده‌ها دارای شکل واضح باشند می‌توان گره‌ها را در نقاط تغییر شیب داده‌ها قرار داد. تعداد گره‌ها اثر زیادی روی برازش اسپلاین دارد. با افزایش تعداد گره‌ها تعداد چندجمله‌ای‌های تکه‌ای برازش داده شده به داده‌ها افزایش می‌یابد و منحنی برازش داده شده ناهموار خواهد بود. در شکل های ۸ و ۹ داده‌های فسیل با استفاده از اسپلاین‌ها با تعداد گره‌های متفاوت هموارسازی شده است. در شکل ۸ با در نظر گرفتن تعداد کوچکی گره داده‌ها به اصطلاح کم برازش شده‌اند و در شکل ۹ با انتخاب تعداد زیادی گره داده‌ها بیش برازش شده‌اند.

با توجه به این‌که با اضافه کردن هر گره یک پارامتر به مدل اضافه می‌شود، می‌توان تعداد گره‌ها را با استفاده از محک‌های انتخاب مدل مانند آکائیک، اعتبارسنجی متقابل و c_p نیز تعیین کرد.



شکل ۸: برازش با استفاده از هموارسازهای اسپلاین با تعداد گره کم



شکل ۹: برازش با استفاده از هموارسازهای اسپلاین با تعداد گره زیاد

۵ اسپلاین‌های هموارسازی

برای پرهیز از مسئله انتخاب گره می‌توان از روش‌های رگرسیون جریمه شده استفاده کرد. ایده‌ی اصلی روش‌های رگرسیون جریمه شده توجه به ناهمواری یک منحنی به واسطه یک تابع جریمه است (ویتاکر [۱۴]). در

این حالت منحنی به صورتی برآورد می‌شود که مصالحه لازم بین اریبی و تغییرپذیری ایجاد شود. روش‌های درست‌نمایی جریمه شده g را با به کار بردن ماکسیم سازی محک لگاریتم درست‌نمایی جریمه شده زیر برآورد می‌کنند:

$$j\left(g\right)=l\left(g;D\right)-P\left(g;\lambda\right),$$

که در آن $l\left(g;D\right)$ درست‌نمایی g و $P\left(g;\lambda\right)$ عبارت جریمه است، پارامتر λ در $P\left(g;\lambda\right)$ پارامتر هموار سازی نامیده می‌شود و موازنه نزدیکی منحنی برازش داده شده به داده‌ها که به وسیله $l\left(g;D\right)$ اندازه‌گیری می‌شود و همواری آن را که با $P\left(g;\lambda\right)$ اندازه‌گیری می‌شود، کنترل می‌کند. شانبرگ [۱۳] ناهمواری یک منحنی را با مقدار انتگرال d-امین مشتق اندازه گیری کرد که به صورت

$$\lambda\int\left[g^{d}(x)\right]^{۲}dx,\qquad\lambda\geq۰,$$

تعریف می‌شود. مدل رگرسیونی

$$y=g\left(x\right)+\varepsilon,$$

که در آن $\epsilon\sim\mathrm{N}(0,\sigma^{2})$ را در نظر بگیرید. در صورتی که g را با استفاده از مینیمم کردن محک کم‌ترین مربعات جریمه شده زیر به دست آوریم:

$$\sum_{i=۱}^{n}\left[y_{i}-g(x_{i})\right]^{۲}+\lambda\int\left[g^{(d)}(x)\right]^{۲}dx.\qquad(۴)$$

می‌توان نشان داد که تابع جریمه در عبارت (۴) به ازای یک اسپلاین طبیعی از درجه d مینیمم می‌شود با گره‌هایی که روی هر مقدار مشاهده شده x_i قرار گرفته‌اند. هنگامی که $\lambda\to\infty$ ، یک چندجمله‌ای از درجه $d-۱$ است که برای آن عبارت انتگرال موجود در محک (۴) صفر است. و در حالی که $\lambda\to۰$ منحنی برازش داده شده g یک اسپلاین برازش شده دقیق است یعنی اسپلاینی که از تمام نقاط طرح بگذرد.

۶ اسپلاین‌های جریمه شده

اسپلاین‌های جریمه شده را می‌توان به عنوان یک حد وسط برای اسپلاین‌های رگرسیونی و هموارسازی در نظر گرفت. ایده اصلی اسپلاین‌های جریمه شده برای برازش منحنی استفاده از تعداد نسبتاً زیادی گره (بین ۳۰ تا ۴۰) روی دامنه متغیر توضیحی و یک جریمه ناهمواری است. تعداد گره‌ها در اسپلاین‌های جریمه شده کمتر از تعداد مشاهدات است بنابراین از دیدگاه محاسبه‌ای بسیار کاراتر از اسپلاین‌های هموارسازی هستند. اسپلاین‌های جریمه شده براساس نوع توابع پایه و جریمه به دو نوع PB-اسپلاین‌ها (برای B-اسپلاین‌های جریمه شده) و PT-اسپلاین‌ها (برای توابع توانی بریده جریمه شده) تقسیم می‌شوند.

۱.۶ PB-اسپلاین‌ها

آیلرز و مارکس [۵] اسپلاین‌ها را به عنوان ترکیبات خطی از توابع B-اسپلاین روی یک شبکه از گره‌های یکسان فاصله‌گذاری شده نشان دادند. طبق تعریف مارکس مدل رگرسیونی به صورت زیر تعریف می‌شود:

$$E\left(y\right)=\mu=B\alpha,$$

که در آن B بردار توابع B-اسپلاین می‌باشد و پارامترهای مدل با مینیمم‌سازی تابع هدف زیر برآورد می‌شوند:

$$Q_{B}=\|y-B\alpha^{۲}\|+\lambda\|D_{d}\,\alpha|^{۲}.$$

در عبارت فوق D_d یک ماتریس است به طوری‌که $D_d\alpha = \Delta^d\alpha$ بردار d امین تفاضلات را می‌سازد و

$$\Delta^d\alpha_j = \sum_{t=0}^{d}(-1)^t\binom{t}{d}\alpha_{j-t}. \qquad (۵)$$

در این شیوه بعد مسئله به جای n (تعداد مشاهدات) روی m (تعداد B-اسپلاین‌ها) قرار می‌گیرد. با مینیمم‌سازی عبارت Q_B یک مجموعه از معادلات زیر حاصل می‌شود:

$$(B^TB + \lambda D_d^TD_d)\widehat{\alpha} = B^Ty.$$

در صورتی که $\lambda = 0$ باشد، این معادلات به معادلات نرمال برای رگرسیون خطی y روی B تبدیل می‌شوند و از آن‌جا که تعداد توابع پایه در B زیاد است برای $\lambda = 0$ منحنی برازش داده شده داده‌ها را بیش برازش خواهد کرد و یک منحنی با نوسانات زیاد تولید می‌شود. با افزایش λ هموارى منحنی تعدیل خواهد شد.

۲.۶ PT-اسپلاین‌ها

راپرت و همکاران [۱۲] برای برازش منحنی به داده‌ها از پایه F توابع توانی بریده (TPF) استفاده کردند. برای TPF از درجه pام ستون jام F به صورت زیر است:

$$F_{ij} = (x_i - t_j)^pI(x_i > t_j),$$

که در آن $I(u)$ یک تابع نشانگر است (برای $u \geq 0$ مقدار یک و برای $u < 0$ مقدار صفر را می‌پذیرد). بردار t یک بردار از گره‌هایی است که می‌توانند در چندک‌های متغیر توضیحی (x) قرار داده شوند. مدل رگرسیونی $E(y) = \mu$ را با استفاده از TPF می‌توان به صورت زیر نوشت:

$$\mu_i = \sum_{k=0}^{p}\beta_kx_i^p + \sum_{j=1}^{n-1}b_jF_{ij},$$

و یا به صورت ماتریسی $\mu = X\beta + Fb$ بیان نمود. که X یک ماتریس $n \times (p+1)$ با عناصر x_i^0 تا x_i^p در سطر iام است. پارامترهای مدل با استفاده از مینیمم‌سازی تابع هدف زیر برآورد می‌شوند:

$$Q_F = \| y - X\beta - Fb\|^2 + \lambda\| b\|^2,$$

که درآن یک جریمه ریج روی b در نظر گرفته شده است. مینیمم‌سازی عبارت بالا یک مجموعه از معادلات زیر را نتیجه می‌دهد.

$$\begin{bmatrix} X^TX & X^TF \\ F^TX & F^TF + \lambda I \end{bmatrix}\begin{bmatrix}\beta \\ b\end{bmatrix} = \begin{bmatrix}X^Ty \\ F^Ty\end{bmatrix}.$$

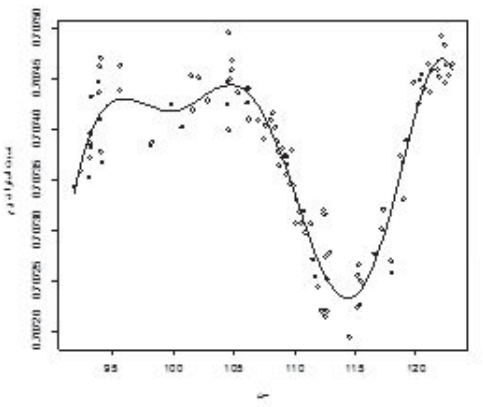
اغلب در استفاده از PT-اسپلاین‌ها از $p = 1$ استفاده می‌کنند. تعداد گره‌ها به گونه‌ای انتخاب می‌شود که منحنی برازش برای مقادیر کوچک λ داده را بیش برازش کند و با افزایش λ ($\lambda \to \infty$) g یک چندجمله‌ای درجه p خواهد بود.

قابل توجه است در PT-اسپلاین‌ها و PB-اسپلاین‌ها وجود عبارت جریمه ضروری است، پس λ ممکن است کوچک باشد ولی همیشه باید مثبت باشد. در شکل های ۱۰ و ۱۱ مجموعه داده‌های فسیل را با استفاده از PT-اسپلاین‌ها و PB-اسپلاین‌ها با فرمان‌های spm() و gam() برازش داده‌ایم.

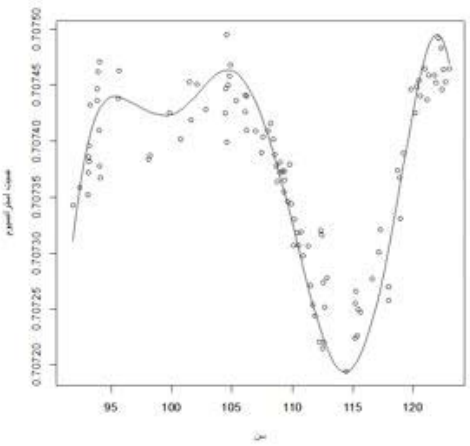
۷ گزینش پارامتر هموارسازی

انتخاب پارامتر هموارسازی λ به شدت روی برازش اسپلاین‌های جریمه شده اثرگذار است سه شیوه اصلی برای انتخاب پارامتر λ وجود دارد.

۱. استفاده از محک‌های نیکویی برازش مانند محک اطلاع آکائیک، محک اعتبارسنجی متقابل و محک



شکل ۱۰: برازش منحنی با استفاده از PT- اسپلاین ها



شکل ۱۱: برازش منحنی با استفاده از PB- اسپلاین ها

اعتبارسنجی متقابل تعمیم یافته (مارکس و آیلرز [۱۰]).

۲. بیان اسپلاین‌های جریمه شده به صورت یک مدل آمیخته و برآورد λ با استفاده از روش ماکسیمم درستنمایی محدود شده (راپرت و کارل [۱۱]).

۳. برآورد همزمان ضرایب رگرسیون و λ با به کارگیری تکنیک‌های شبیه‌سازی مونت کارلوی زنجیر مارکف (برزگر و لنگ [۳]).

تفصیل روش‌های مذکور از حوصله بحث این مقاله خارج است.

۸ بحث و نتیجه‌گیری

زمینه بحث و تحلیل رگرسیون پارامتری پیرامون الگوهای پارامتری است، الگوهای پارامتری می‌توانند به صورت خطی یا غیرخطی به پارامترها وابسته باشند. نتیجه یک تحلیل رگرسیون ناپارامتری برازش یک منحنی به داده‌ها است و از آنجا که این منحنی بدون فرض کردن شکلی پارامتری برای تابع رگرسیون به دست می‌آید، تفسیر برآوردگری که به این روش به دست می‌آید مشکل است. فنون رگرسیون ناپارامتری در مواقعی که درباره شکل منحنی رگرسیون اطلاعات ناچیز در دسترس است و از نمودار پراکنش نیز نمی‌توان برای استنباط درباره منحنی رگرسیون ایده مطمئنی گرفت، بهترین راه حل هستند. اما هنگامی که الگوهای پارامتری مناسب داده‌ها هستند کارایی برآوردگرهای ناپارامتری کمتر است.

در این مقاله ابتدا درباره ناتوانی روش‌های رگرسیون پارامتری در برآورد رابطه واقعی بین دو متغیر صحبت کردیم و سپس هموارسازهای اسپلاین را به عنوان یک روش ناپارامتری در برآورد یک مدل رگرسیونی معرفی کردیم. پس از آن به معرفی یکی از انواع اسپلاین‌ها، اسپلاین‌های رگرسیونی، پرداختیم و برای پرهیز از مسئله انتخاب گره اسپلاین‌های هموارسازی را معرفی کردیم. با توجه به جنبه‌های محاسباتی اسپلاین‌های هموارسازی اسپلاین‌های جریمه شده را به عنوان حد وسط برای اسپلاین‌های رگرسیونی و هموارسازی براساس دو نوع تابع پایه، پایه توانی بریده و پایه B-اسپلاین، مطرح کردیم.

از آنجا که برای استفاده از اسپلاین های هموارسازی و جریمه شده نیاز به برآورد پارامتر هموارسازی است خلاصه ای از روش های انتخاب پارامتر هموار سازی موجود را بیان کردیم که علاقه مندان برای اطلاعات بیشتر می توانند به برزگر و لنگ [۴] و بلیتز و لنگ [۲] مراجعه کنند.

قابل ذکر است در حالتی که بیش از یک متغیر توضیحی در مدل وجود داشته باشد می‌توان با استفاده از مدل‌های جمعی تعمیم‌یافته رابطه بین متغیر پاسخ و تک‌تک متغیرهای توضیحی را با استفاده از هموارسازهای اسپلاین برآورد کرد و نمودارهای مربوطه را به صورت جداگانه رسم کرد. که برای توضیحات بیشتر در این زمینه می توان به هیستی و تیبشیرانی [۸] مراجعه کرد.

مراجع

[۱] میرزایی، م. (۱۳۹۰). مقایسه اسپلاین‌های جریمه شده و چند جمله‌ای‌های کسری در برآورد مدل‌های جمعی تعمیم‌یافته. پایان‌نامه کارشناسی ارشد، دانشگاه فردوسی مشهد.

[2] Blitz, C. and Lang, S. (2008). Simultaneous of variables and smoothing parametersin structured additive regression models. Computational Statistics and Data Analysis, 53, 61-81.

[3] Brezger, A. and Lang, S. (2006). Generalized additive regression based on bayesian p-splines. Computational Statistics and Data Analysis, 50, 967-991.

[4] Brezger, A. and Lang, S. (2008). Simultaneous probability statements for bayzian p-splines. Statistical Modeling, 8, 141-168.

[5] Eilers, P. and Marx, B. (1996). Flexible smoothing using B-splines and penalized likelihood. Statistical Science, 11, 89-121.

[6] Eilers, P. and Marx, B. (2010). Splines, Knots, and Penalties. Computational Statistics, 2, 6, 637-653.

[7] Green, P.J. and Silverman, B.W. (1994). Nonparametric Regression and Generalzed Linear Models: a roughness Penalty Approach. Chapman & Hall, London.

[8] Hastie, T.J. and Tibshirani, R.J. (1990). Generalized Additive Models. Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

[9] Hastie, T.J., Tibshirani, R.J. and Friedman, J. (2003). The Elements of Statistical Learning. Spring, New York.

[10] Marx, B. and Eilers, P. (1999). Generalized linear regression for sampled signals or curves: A P-spline approach. Technometrics, 41, 1, 1-13.

[11] Ruppert, D. and Carroll, R.J. (1997). Penalized Regression Splines. Technical Report, Cornell University, Ithaca, USA.

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

[13] Schoenberg, l. (1964). Interpolation by spline functions and its minimum properties. International series of Numerical Analysis, 5, 109-129.

[14] Whittaker, E.T. (1923). On a new method of graduation. Proceedings of the Edinburgh Mathematical Society, 41, 63-75.